UNIVERSIDADE FEDERAL FLUMINENSE

ALTOBELLI DE BRITO MANTUAN

CONTEXTUALIZAÇÃO ESPACIAL PARA MINERAÇÃO DE ITEMSETS

NITERÓI

2021

UNIVERSIDADE FEDERAL FLUMINENSE

ALTOBELLI DE BRITO MANTUAN

CONTEXTUALIZAÇÃO ESPACIAL PARA MINERAÇÃO DE ITEMSETS

Tese de doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação

Orientador:
Prof. Dr. Leandro Augusto Frata Fernandes

NITERÓI

2021

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

M291c Mantuan, Altobelli de Brito
CONTEXTUALIZAÇÃO ESPACIAL PARA MINERAÇÃO DE ITEMSETS / Altobelli de Brito Mantuan ; Leandro Augusto Frata Fernandes, orientador. Niterói, 2021.
323 f. : il.

Tese (doutorado)-Universidade Federal Fluminense, Niterói, 2021.

DOI: http://dx.doi.org/10.22409/PGC.2021.d.11055850775

1. Mineração de dados (Computação). 2. Produção intelectual. I. Fernandes, Leandro Augusto Frata, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

ALTOBELLI DE BRITO MANTUAN

CONTEXTUALIZAÇÃO ESPACIAL PARA MINERAÇÃO DE ITEMSETS

Tese de doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação

Aprovada em Setembro de 2021.

BANCA EXAMINADORA

Prof. Dr. Leandro Augusto Frata Fernandes, IC-UFF (Orientador)

Profa. Dra. Adriana Bechara Prado, Dell Technologies

Prof. Dr. Alexandre Plastino de Carvalho, IC-UFF

Profa. Dra. Aline Marins Paes Carvalho, IC-UFF

Profa. Dra. Ana Cristina Bicharra Garcia, UNIRIO

Prof. Dr. Wagner Meira Júnior, DCC-UFMG

Niterói

2021

Dedico este trabalho aos meus pais Antônio (in memoriam) e Marlene Maria, ao meu querido irmão Aslen (in memoriam), a minha sobrinha/afilhada Ana Carolina e aos meus afilhados Eduardo e Miguel.

Agradecimentos

À minha mãe, pelo amor, carinho, dedicação e conselhos.

A meu irmão, pelo amor e amizade. Nossa família seguira feliz, unida e honrando tudo aquilo que você representou pra gente.

A meu orientador, pelas oportunidades de aprendizado.

Aos meus amigos de UFF, pelos momentos de estudo, conversas e descontração.

À CAPES pelo apoio financeiro.

E à todas as outras pessoas que ajudaram de alguma maneira nesta caminhada e nas anteriores.

Resumo

Identificação de itemsets desempenha um papel importante no processo descritivo da base de dados. Porém, encontrar os itemsets é uma tarefa computacionalmente custosa que requer a definição de limiares de corte, tal como o suporte mínimo. Além do custo inerente ao processo de criação de itemsets, existe também a dificuldade em definir os valores dos limiares, dado que, para alcançar bons resultados, o especialista da técnica de mineração deve ter conhecimento da base a ser minerada. Esses problemas são fatores motivadores para a investigação de novos algoritmos para identificação de itemsets. Algumas pesquisas são: novas métricas para definir relevância de itemsets; novas estruturas que garantem melhor desempenho de processamento; e algoritmos que sejam menos vulneráveis aos parâmetros de entrada. Neste trabalho, é apresentado o uso da contextualização espacial nos algoritmos de mineração. O procedimento de mineração cria um espaço métrico multidimensional para determinada base de dados de tal forma que itemsets relevantes possam ser recuperados em relação à localização espacial relativa de seus itens. A representação da base de dados no espaço multidimensional auxilia na interpretação e definição de clusters com sobreposição de itens relacionados. A distância do item ao centro do cluster é usada como critério para gerar itemsets. Essa tese apresenta três vertentes de estudo, onde a primeira vertente trata o caso de recuperar itemsets fechados para base de dados estática. Para essa vertente é apresentado o algoritmo sequencial *SCIM*, que introduz um novo procedimento que usa as estruturas FP-tree e CFI-*tree*. Na segunda vertente, é proposto uma solução paralela para recuperar itemsets fechados em bases de dados estáticas, o algoritmo proposto PSCIM utiliza a estrutura LCM no processo de busca de itemsets fechados, além disso, propõe alterações no cálculo do mapeamento espacial, para melhorar os resultados para bases de dados esparsas. Embora os algoritmos SCIM e PSCIM difiram, é utilizada as mesmas métricas para os dois algoritmos, onde se mostra que em várias bases de dados que a média de *all-confidence* e de *cross-support* dos itemsets recuperados por nossas técnicas superam os resultados dos algoritmos do estado da arte. Para o algoritmo paralelo PSCIM observamos um *speedup* médio de 3,3× para 4 *threads* em um ambiente de 4 núcleos reais. Na última vertente, é apresentado um resultado teórico para o cálculo do *Dual Scaling* para bases de dados de fluxo contínuo. Neste cenário, é apresentado duas formulações para os dois tipos de atualização de novas informações, sendo elas *landmark* e *sliding*. Outra formulação proposta com base no resultado teórico proposto para base de dados de fluxo contínuo foi um algoritmo capaz de retirar informações de determinado usuário do procedimento de *Dual Scaling*, sem ter que recalcular todo o processo visto que a base de dados seja atualizada.

Palavras-chave: processamento de dados, itemset, redução de dimensão, clusterização, *Dual Scaling*, base de dados de fluxo contínuo.

Abstract

Itemset pattern identification plays an essential role in the database description process. However, finding these patterns is a computationally expensive task that requires the definition of cut-off thresholds, like the minimal support. Additionally, there is also the complexity in defining a threshold value since, to achieve good results, the specialist in the mining technique must know the database behavior. These problems are motivating factors for the investigation of new algorithms for the itemset mining process. A few research topics arise in this area, like new metrics to define the relevance of itemsets, new structures that ensure better processing performance, and algorithms that are less vulnerable to input parameters. In this work, we present the use of spatial contextualization in the itemset mining algorithms. The mining procedure creates a metric space for a given database so that relevant itemsets can be retrieved to the relative spatial location of their items. Our approach uses Dual Scaling to map the items to a multidimensional metric space called solution space. The database representation in the solution space helps interpret and define clusters with overlapping related items. The distance of the item from the center of the cluster is used as a criterion to generate itemsets. This thesis presents three study topics. The first one deals with the case of retrieving closed itemsets for the static database. For this strand the sequential algorithm *SCIM* is presented, which proposes a new procedure that uses the FP-tree structures and CFI-Tree. In the second topic, we propose a parallel solution to recover closed itemsets in static databases. The proposed parallel algorithm PSCIM uses the LCM structure in the process of searching for closed itemsets. In addition, it proposes changes in the calculation of the spatial contextualization mapping to improve the results for sparse databases. The same metrics were used in both algorithms, SCIM and PSCIM, where we demonstrated that in several databases that the average all-confidence and cross-support of the itemsets retrieved by our techniques surpasses the results of state-of-the-art algorithms. In addition, we use the minimum description length (MDL) as a metric to check how descriptive the collection of itemsets retrieved by each compared approach is. For the parallel algorithm PSCIM, we observe an average speedup of 3.3× for 4 threads in a real 4 cores environment. Finally, as the last topic, we propose a theoretical solution to calculate the Dual Scaling for streaming databases. In this scenario, we propose formulation for both types of updating new information, namely landmark and sliding. Another formulation proposed based on theoretical solution for streaming database was the algorithm capable of removing information from a given user from the Dual Scaling procedure without recalculating the entire process given the new database.

Keywords: data mining, streaming database, itemset, pruning, dimension reduction, clustering, Dual Scaling.

Lista de Figuras

1.1 Ideia central do processo de mineração proposta na Tese. 3

2.1 Representa a compressão de uma base de dados usando o conjunto CT de itemsets minerados. Comparando com os itemsets de tamanho 1 ST. . . . 10

3.1 Representação da estrutura FP-*tree*. 15

4.2 Exemplo da aplicação do *Dual Scaling* em uma base sintética $\mathcal{D}$. 26

4.3 Resultado da distância e ângulo entre os atributos com o atributo de referência Esposa - Sem formação. As distâncias entre pares são representadas pelo eixo y e as cores do ponto reflete a angulação. 28

4.4 Resultado da distância e ângulo entre os atributos com o atributo de referência Contraceptivo - Uso contínuo. As distâncias entre pares são representadas pelo eixo y e as cores do ponto reflete a angulação. 29

5.1 Algoritmo SCIM. 32

5.2 Estrutura de formação do cluster. 35

5.3 Single 53

6.1 Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM_{v1}, PSCIM_{v2}, PSCIM_{v3} e PSCIM_{v4} sobre as bases de dados densas da Tabela 6.2. Neste estudo foi usado o melhor valor de parâmetro para cada variedade de PSCIM. 73

6.2 Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM_{v1}, PSCIM_{v2}, PSCIM_{v3} e PSCIM_{v4} sobre as bases de dados esparsas da Tabela 6.3. Neste estudo foi usado o melhor valor de parâmetro para cada variedade de PSCIM. 79

6.3 Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM, TopPI, Slim e LAM sobre as bases de dados densas da Tabela 6.4. Neste estudo foi usado o melhor valor de parâmetro para cada técnica. 86

6.4 Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM, TopPI, Slim e LAM sobre as bases de dados esparsa da Tabela 6.5. Neste estudo foi usado o melhor valor de parâmetro para cada técnica. 92

6.5 Proporção de tempo de processamento para cada etapa do algoritmo PSCIM. 99

7.1 Três tipos de janelas que define o modo de atualização para bases de dados de fluxo contínuo. 102

Lista de Tabelas

5.1 Matriz de distância entre itens para a base sintética $\mathcal{D}$. 34

5.2 Todos os itemsets fechados I contidos na base sintética $\mathcal{D}$. 40

5.3 Base de dados usada em nossos experimentos. 44

5.4 Distribuição de valores médios de *all-confidence* e de *cross-support* dos itemsets fechados recuperados por FPClose, Slim, TopPI e SCIM sobre as bases de dados de múltiplas escolhas da Tabela 5.3. 46

5.5 Resumo de significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo SCIM com os algoritmos Slim e TopPI. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas na Tabela 5.4. 55

6.1 Informações sobre as bases de dados 67

6.2 Desempenho do algoritmo/parametrização para as variações $\text{PLSCIM}_{\text{v1}}$, $\text{PLSCIM}_{\text{v2}}$, $\text{PLSCIM}_{\text{v3}}$, e $\text{PLSCIM}_{\text{v4}}$ sobre as bases de dados densas da Tabela 6.1. 74

6.3 Desempenho do algoritmo/parametrização para as variações $\text{PLSCIM}_{\text{v1}}$, $\text{PLSCIM}_{\text{v2}}$, $\text{PLSCIM}_{\text{v3}}$, e $\text{PLSCIM}_{\text{v4}}$ sobre as bases de dados densas da Tabela 6.1. 80

6.4 Desempenho do algoritmo/parametrização para os algoritmos PLSCIM, TopPI, LAM e Slim sobre as bases de dados densas da Tabela 6.1. 87

6.5 Desempenho do algoritmo/parametrização para os algoritmos PLSCIM, TopPI, LAM e Slim sobre as bases de dados esparsas da Tabela 6.1. 93

6.6 Resumo de significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo PSCIM com os algoritmos Slim, LAM e TopPI. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas nas Tabelas 6.4 e 6.5. 97

6.7 *Speedup* do algoritmo PSCIM. 100

Sumário

1 Introdução **1**

1.1 Ideia Central 4

1.2 Desafios 4

1.3 Demonstração e Análises 6

1.4 Contribuições 6

2 Conceitos e Definições **8**

2.1 Definições Gerais 8

2.2 Definições para Bases de Dados de Fluxo Contínuo 11

3 Trabalhos Relacionados **13**

3.1 Itemsets em Bases de Dados Estáticas 13

3.2 Itemsets em Bases de Dados de Fluxo Contínuo 18

3.3 Discussão 20

4 Dual Scaling **22**

4.1 Dados Categóricos 22

4.2 Algoritmo *Dual Scaling* 24

4.3 Exemplo de Aplicação do Cálculo 26

4.4 Aplicação em Bases de Dados Reais 27

4.5 Limitações 29

5 Mineração de Itemsets Fechados: SCIM **31**

5.1 Visão Geral do Algoritmo Desenvolvido 31

5.2 Procedimento de Clusterização com Sobreposição 33

5.3 Geração de Itemsets Fechados 37

5.4 Resultados 41

5.4.1 Definição das Métricas Utilizadas 41

5.4.1.1 Primeira Métrica: *All-confidence* dos Itemsets Fechados Selecionados 42

5.4.1.2 Segunda Métrica: *Cross-support* dos Itemsets Fechados Selecionados 42

5.4.1.3 Terceira Métrica: Tempo de Execução 43

5.4.2 Experimentos 43

5.4.3 Discussão Sobre Qualidade e Tempo de Execução 44

5.4.4 Discussão Sobre a Técnica de Clusterização 56

6 Mineração Paralela de Itemsets Fechados: PSCIM 57

6.1 SCIM Revisitado 58

6.2 Mineração de Itemset Fechado em Tempo Linear 61

6.3 Procedimento de Clusterização 62

6.4 Geração de Itemset Fechado 64

6.5 Resultados 66

6.5.1 Definição de Métricas 68

6.5.1.1 Primeira Métrica: *All-confidence* dos Itemsets Fechados Selecionados 68

6.5.1.2 Segunda Métrica: *Cross-support* dos Itemsets Fechados Selecionados 68

6.5.1.3 Terceira Métrica: Tempo de Execução 69

6.5.1.4 Quarta Métrica: *Speedup* 69

6.5.2 Discussão Sobre SCIM Revisitado 69

6.5.2.1 Base de Dados Densa 77
6.5.2.2 Base de Dados Esparsa 77
6.5.2.3 Discussão 83
6.5.3 Discussão Sobre Qualidade e Tempo de Execução 83
6.5.4 Análise do Custo Relativo de Processamento das Etapas do PSCIM 98
6.5.5 Escalabilidade do PSCIM por Número de Threads 99

7 Dual Scaling Processado em Blocos de Transações 102

7.1 *Dual Scaling* em Bases de Dados de Fluxo Contínuo 104
7.1.1 Fluxo de Dados do Tipo *Landmark* 105
7.1.2 Fluxo de Dados do Tipo *Sliding* 107
7.2 Regulamentação Geral de Segurança de Dados 108
7.3 Discussão 109

8 Conclusões 111

8.1 Trabalhos Futuros 113

Referências 115

Apêndice A – SCIM: EXPERIMENTOS COM DIFERENTES CONFIGURAÇÕES DE PARÂMETROS 120

A.1 Escolha dos Parâmetro 134
A.1.1 Letter recognition 134
A.1.2 mFeat 136
A.1.3 Wine 138
A.1.4 Page Blocks 140
A.1.5 Pen digits 143
A.1.6 Waveform 145
A.1.7 Ecoli 148

A.1.8 Connect-4 150

A.1.9 Tic-tac-toe 152

A.1.10 Led7 155

A.1.11 Pima 157

Apêndice B – PSCIM: EXPERIMENTOS COM QUATRO VARIAÇÕES DO ALGORITMO 160

B.1 Escolha dos Parâmetros 173

B.1.1 Bases de Dados Densa 173

B.1.1.1 Chess 173

B.1.1.2 Kddcup99 177

B.1.1.3 Mushrooms 182

B.1.1.4 PowerC 186

B.1.1.5 Pumsb 191

B.1.1.6 RecordLink 195

B.1.1.7 Skin 200

B.1.1.8 Susy 204

B.1.2 Bases de Dados esparsa 209

B.1.2.1 Accidents 209

B.1.2.2 BMSWebView2 209

B.1.2.3 BMS1 218

B.1.2.4 FoodmartFIM 222

B.1.2.5 Fruithut 227

B.1.2.6 OnlineRetail 231

B.1.2.7 PAMP 235

B.1.2.8 Retail 240

Apêndice C – PSCIM: EXPERIMENTOS COM DIFERENTES CONFIGURAÇÕES

DE PARÂMETROS **245**

C.1 Escolha dos Parâmetros 265

C.1.1 Bases de Dados Densa 266

C.1.1.1 Chess 266

C.1.1.2 Kddcup99 268

C.1.1.3 Mushrooms 271

C.1.1.4 PowerC 273

C.1.1.5 Pumsb 276

C.1.1.6 RecordLink 278

C.1.1.7 Skin 281

C.1.1.8 Susy 283

C.1.2 Bases de Dados esparsa 286

C.1.2.1 Accidents 286

C.1.2.2 BMSWebView2 288

C.1.2.3 BMS1 288

C.1.2.4 FoodmartFIM 294

C.1.2.5 Fruithut 296

C.1.2.6 OnlineRetail 299

C.1.2.7 PAMP 301

C.1.2.8 Retail 304

Capítulo 1

Introdução

O objetivo da mineração de dados é descobrir em base de dados padrões frequentes que são interessantes, úteis e, muitas vezes, inesperados para o cientista de dados. A mineração de padrões frequentes é um tipo de aprendizado não supervisionado, pois não requer dados rotulados. O primeiro algoritmo para extração de padrões (itemsets) frequentes foi introduzido por Agrawal et al. [3]. O algoritmo Apriori extrai itemsets frequentes dada uma base transacional. A descoberta de itemsets frequentes é considerado como uma atividade importante para o processo de mineração de dados e tem aplicação prática em diferentes áreas, tais como, biologia, telecomunicações, medicina, etc.

O algoritmo Apriori tem como um dos parâmetros de entrada o limiar de suporte mínimo, ou seja, um itemset só é considerado frequente se o mesmo possuir o valor de frequência maior ou igual ao suporte mínimo definido. O algoritmo pode ser dividido em duas etapas principais: (i) a geração de itemsets candidatos e a identificação de itemsets frequentes; (ii) a extração de regras de associação. Vale ressaltar que a etapa (ii) não será abordada neste trabalho. A mineração de itemsets traz grandes desafios devido ao custo computacional inerente ao processo [9]. Na primeira etapa para a geração de itemsets candidatos, o problema está relacionado com o processo combinatório entre os itens, que cresce de forma exponencial dada a quantidade de itens que compõem a base de dados. Para exemplificar, m itens distintos permitem a criação de $2^m - 1$ itemsets distintos. Outro problema da primeira parte está relacionado com a identificação de itemsets frequentes, visto que esse processo possui custo ao percorrer a base de dados, para contabilizar a frequência de cada itemset e, assim, definir se determinado itemset é frequente, dado o suporte mínimo especificado pelo especialista.

Tendo em evidência o custo computacional do Apriori, é de se esperar que para bases grandes, ou seja, aquelas que possuem um conjunto grande de itens e de transações, a

execução pode se tornar impraticável. A importância desse algoritmo é notória e, como consequência, podemos observar várias vertentes de pesquisa na literatura [22]. Uma das vertentes tem como foco reduzir o número de passadas pela base de dados, onde são propostas estruturas que conseguem condensar as informações pertinentes para cálculos de frequência. Em alguns estudos é necessário percorrer a base de dados apenas duas vezes [27]. Outra solução para essa vertente é o uso de amostra da base de dados, onde é selecionado um conjunto de transações, onde o tamanho da amostra depende do valor de erro esperado que o usuário aceita tolerar [67]. Pensando no desempenho, existem trabalhos que usam projeções da base de dados, subconjuntos de transações da base de dados original, onde a cada projeção da base de dados o número de transações e itens é reduzido ao longo do processo de mineração [58].

Do ponto de vista prático do especialista usuário de técnicas de mineração de itemsets, outro problema é a definição dos parâmetros de entrada dos algoritmos de mineração, como exemplo o suporte mínimo. Este limiar é usado para controlar a quantidade de itemsets frequentes extraídos da base de dados. Por fim, o analista de dados observa os itemsets frequentes selecionados e define interpretações da base de dados. Fica evidente que o sucesso deste processo depende muito do especialista de dados, visto que uma definição errada para parâmetro de suporte mínimo pode conduzir a uma interpretação errônea da informação. Ao identificar os itemsets frequentes, estamos dizendo, implicitamente, que apenas estes itemsets são interessantes, no entanto, essa afirmação não se aplica para itemsets raros, ou seja, itemsets não frequentes e interessantes. Em resumo, o especialista tem a seguinte decisão a tomar: (i) definir um valor alto de suporte mínimo podendo acarretar perda de informação ou, por outro lado, (ii) definir um valor baixo de suporte mínimo podendo minerar uma quantidade extrapolada de itemsets frequentes.

Todas as vertentes de pesquisa mencionadas têm como premissa algoritmos que processam base de dados estáticas. Em outras palavras, a base de dados não sofre alteração do conteúdo durante o processo de mineração. Todavia, existem bases de dados onde o processo de obtenção de novas informações é contínua (e.g., transações de cartão de crédito, controle de tráfico urbano, informações de sensores, entre outros). A base de dados de fluxo contínuo possui características bem diferentes de uma base de dados estática, dentre elas temos: (i) novas informações chegam continuamente ao longo do tempo; (ii) não é possível armazenar essas informações, portanto as informações podem ser lidas apenas uma vez; e (iii) a quantidade total é ilimitada, não podendo ser definida para esse tipo de bases de dados [26]. Manku e Motwani [38] propuseram o algoritmo *lossy counting* para calcular o conjunto aproximado de itemsets frequentes de uma base de dados de fluxo

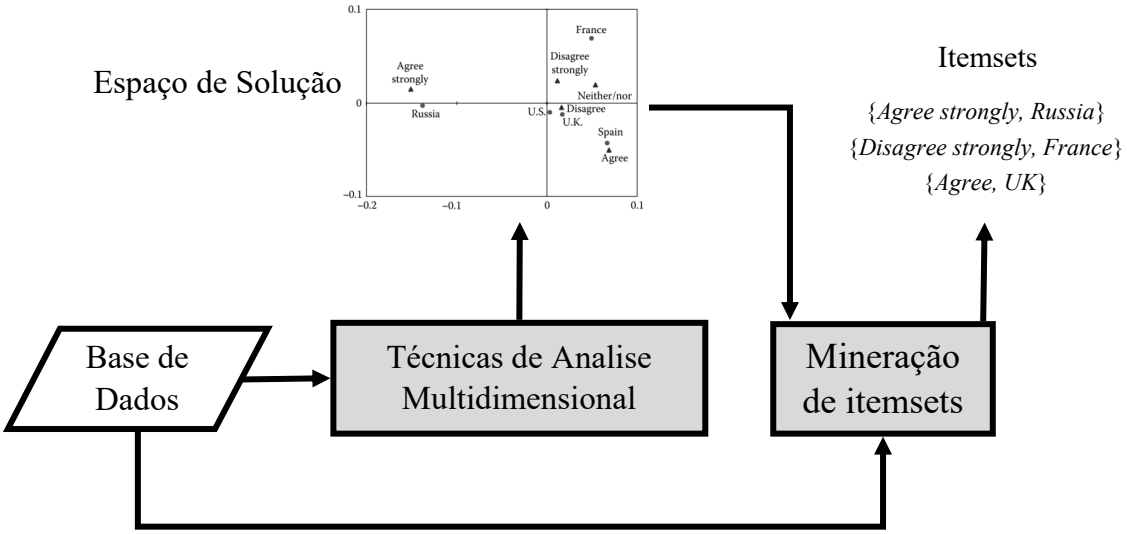


Figura 1.1: Ideia central do processo de mineração proposta na Tese.

contínuo. Essa vertente de pesquisa possui desafios já conhecidos, tais como definição de parâmetro e complexidade no processo de busca de itemsets frequentes, problemas com limitações de leitura do dado, que no caso pode acontecer apenas uma vez, e tempo de resposta de processamento deve ser o mais rápido possível para atender o problema.

Mantuan [39], em seu trabalho de mestrado, demonstrou que a abordagem baseada em técnicas de análise multidimensional podem ser usada para gerar itemsets, pois é possível definir clusters no espaço multidimensional de tal modo a auxiliar o processo de busca de itemsets relevantes. A métrica *all-confidence* [48] foi utilizada para definir se um itemset é relevante. A interpretação dessa métrica define o quão correlacionados são os itens que compõem o itemset em questão. Dada essa observação, este trabalho propõe usar o mapeamento da base de dados no espaço multidimensional em diferentes vertentes de mineração de itemsets (Figura 1.1).

Nesta tese são propostos dois algoritmos para mineração de itemsets. O primeiro lida com base de dados transacionais estáticas, tendo como foco recuperar itemsets fechados relevantes representados dentro dos clusters. O segundo também estuda o caso da base de dados transacionais estáticas, visando recuperar itemsets fechados representados dentro dos clusters, no entanto, a solução proposta define uma alteração na definição na formação dos clusters e, define também, uma solução paralela para o processo de mineração. Além dos dois algoritmos, também é proposto um estudo teórico que permite atualizar o mapeamento do espaço multidimensional e, além disso, atualizar a formação dos clusters a cada novo bloco de transações para uma base de dados transacionais de fluxo contínuo.

1.1 Ideia Central

Dentre as diversas vertentes de trabalhos relacionados às técnicas de mineração de itemsets, este trabalho propõe uma forma menos custosa de minerar os itemsets. Sua ideia central pode ser descrita como o estudo da seguinte hipótese:

"Se a relação de distância entre itens no espaço multidimensional é proporcional a coocorrência desses itens na base de dados então itemsets formados por itens próximos devem ter o all-confidence maior do que itemsets formados por itens distantes. Logo, é possível estimar a contextualização espacial de itens de uma base de dados e utilizar tal contextualização na formação de clusters de itens relacionados, de modo a promover a mineração de itemsets fechados, sem o uso de limiares como suporte mínimo."

1.2 Desafios

A proposta dessa tese é usar a contextualização espacial da base de dados com o propósito de obter informações importantes que podem ser usadas para reduzir o espaço de busca no processo de mineração de itemsets. ***O primeiro desafio*** é escolher o procedimento para o mapeamento dos itens da base de dados. Neste trabalho será usado, como prova de conceito, o *Dual Scaling* [45], que recebe como entrada uma base de dados transacionais ou bases de dados de múltiplas escolhas, e como saída, promove relações de itens representadas por distância (Capítulo 4).

O mapeamento de itens da base de dados no espaço multidimensional, chamado espaço de soluções, traz consigo uma contextualização semântica, onde a organização espacial dos itens define características de comportamento da base de dados. ***O segundo desafio*** é desenvolver um algoritmo não supervisionado capaz de criar clusters de itens com sobreposição no espaço de soluções. Vale ressaltar que seria um equívoco propor ou usar uma técnica de clusterização por particionamento, visto que itens da base de dados se relacionam em diferentes níveis. Logo, é possível prever que itens podem pertencer a mais de um cluster, dependendo apenas do nível de correlação com o cluster em questão. Sendo assim, é criado um algoritmo de clusterização que lide com o espaço de soluções (Capítulos 5 e 6).

Dado que temos os clusters já definidos, é importante definir algoritmos de mineração de itemsets capazes de tomar decisões utilizando estes clusters. ***O terceiro desafio*** é desenvolver um algoritmo de mineração de itemsets fechados para bases de dados transa-

cionais estáticas usando a informação dos clusters como tomada de decisão. Os desafios secundários nesse algoritmo são escolher uma estrutura compacta capaz de evitar um número elevado de passadas na base de dados e, além disso, a estrutura deve conseguir evitar a geração de conjunto a itemset candidato, o que resulta em um melhor aproveitamento na desempenho do processamento. Uma vez definida a estrutura, utilizam-se as informações dos clusters durante o processo de busca a fim de reduzir o custo, pois itens que não são considerados relevantes pelos clusters não devem ser visitados (Capítulo 5).

A demanda por algoritmos mais eficientes, dado o crescimento das bases de dados tanto em números de itens quanto em números de transações, faz com que novas estratégias de processamento sejam empregadas, uma delas é a paralelização do processo de mineração de itemsets fechados. Neste contexto, ***o quarto desafio*** é melhorar o desempenho no processo de mineração, para isso foi criado um algoritmo paralelo de mineração de itemsets de fechados para bases de dados transacionais estáticas usando a contextualização espacial como tomada de decisão para podar a árvore de busca. O algoritmo deve usar uma estrutura capaz de criar projeções da base de dados, dessa forma ter consumo de memória linear, percorrer a base de dados em tempo linear e permitir o processamento paralelo sem colisão de leitura. O mapeamento do espaço de soluções deve lidar com bases de dados esparsas, além disso, promover uma melhora na formação dos clusters. Por fim, usar os clusters na tomada de decisão na construção das possíveis projeções, assim reduzindo o espaço de busca (Capítulo 6).

O quinto desafio é conseguir atualizar, de forma eficiente, o espaço de soluções a cada novo bloco de informações dado uma base de dados de fluxo contínuo. Neste sentindo é apresentado um estudo teórico capaz de atualizar o mapeamento do espaço de soluções a cada novo bloco de transações. Adicionalmente, é preciso manter as informações dos clusters atualizado, ou seja, clusters podem ser alterados em função de novos dados. O estudo proposto deve abordar dois cenários. No primeiro cenário, novas informações de blocos de transações vão chegando sendo concatenados com as transações da base de dados atual. No segundo cenário, a base de dados em memória, chamado janela de informação, tem um tamanho limitado de espaço, logo, uma vez cheia, deve-se remover desta janela o bloco de transações mais antigas antes de inserir o novo bloco (Capítulo 7).

1.3 Demonstração e Análises

A implementação das técnicas propostas para mineração de itemsets fechados frequentes foram feitas utilizando C++. Quanto aos experimentos, para o primeiro algoritmo proposto (Capítulo 5) foram selecionados os algoritmos FPClose, Krimp, Slim e TopPI para serem confrontados com a técnica proposta (SCIM). Foram usados 11 bases de dados no formato de múltiplas escolhas. Para o segundo algoritmo proposto (Capítulo 6) foram selecionados para comparação com a técnica proposta (PSCIM) os algoritmos Slim, TopPI e LAM. Nesta comparação, foram selecionados 16 bases de dados transacionais. Para os experimentos envolvendo tanto o SCIM quanto o PSCIM, foram realizadas execuções dos algoritmos com diferentes valores de parâmetros, de modo a escolher o melhor resultado para uma comparação justa entre as técnicas. Detalhes sobre as escolhas dos melhores parâmetros tanto para o SCIM e quanto para o PSCIM, são apresentados nos Apêndices A e C. Adicionalmente são apresentados nos apêndices a significância estatística dos itemsets fechados recuperados e o comprimento mínimo de descrição (MDL) que define o quanto o conjunto de itemsets minerados conseguem comprimir a base de dados transacional.

Para que uma comparação possa ser feita, é preciso que se defina métricas. O primeiro cenário visa a análise qualitativa dos itemsets selecionados. Para isso, são utilizadas duas métricas, o *all-confidence* e o *cross-support* [63] que definem a relevância do itemset fechados recuperados. Por fim, o último cenário analisa os tempos de processamentos. Essa análise tem como finalidade verificar o custo computacional de cada técnica. Para o estudo do algoritmo PSCIM foi adicionado a métrica de *speedup*, com o intuito de observar a escalabilidade com o uso de mais núcleos de processamento. Os Capítulo 5 e 6 definem com um maior rigor as métricas utilizadas e abordam de forma detalhada os experimentos executados.

1.4 Contribuições

As contribuições desta tese são:

- Um procedimento de clusterização de itens no espaço de soluções proveniente do mapeamento do *Dual Scaling*.
- Um procedimento para geração de itemsets fechados baseado na contextualização da base de dados de múltipla escolha estática.

- Um procedimento paralelo para geração de itemsets fechados baseado na contextualização da base de dados transacional estática.
- Um estudo teórico para atualização do mapeamento do *Dual Scaling* em blocos de transações.

Uma informação adicional, o estudo e a definição do algoritmo SCIM proposto gerou uma publicação na conferência IEEE ICDM 2018 [40].

Capítulo 2

Conceitos e Definições

Neste capítulo introduzimos algumas definições usadas no processo de mineração de itemsets. Além disso, são feitos comentários sobre as principais propriedades do itemset.

2.1 Definições Gerais

O problema de mineração de itemsets tem como finalidade a extração de padrões de uma base de dados transacional. Nesta base de dados cada registro (i.e., transação) é um conjunto de itens. Formalmente, uma base de dados transacional $\mathcal{D}$ é definida da seguinte forma:

Definição 1 (Base de dados) *Seja $\mathcal{I} = \{i_1, i_2, \cdots, i_m\}$ o conjunto universal dos itens. Uma base de dados transacional $\mathcal{D} = T_1, T_2, \cdots, T_n$ é um conjunto de transações, onde $\mathcal{T}_k$ é uma transação, i.e., um conjunto de itens dado $\mathcal{T}_k \subseteq \mathcal{I}$.*

Os algoritmos de mineração visam retornar um conjunto de padrões, sendo usualmente chamados itemsets, e tem a seguinte definição:

Definição 2 (Itemset) *P é definido como itemset se P é um subconjunto de $\mathcal{I}$. Nós dizemos que a transação T_k contém P, se somente se, $P \subseteq T_k$.*

Cada itemset possui um valor que define a quantidade de ocorrência do mesmo na base de dados (i.e., suporte conjuntivo). Outra definição equivalente é dada como o percentual do número total de transações, chamado suporte absoluto ou suporte.

Definição 3 (Suporte) *Para um itemset P, todas as transações que inclui P definem uma coleção $\mathcal{T}(P) = \{\mathcal{T}_k \mid \mathcal{T}_k \in \mathcal{D}, P \subseteq \mathcal{T}_k\}$ chamada ocorrências de P. O suporte conjuntivo, i.e., o número de transações em $\mathcal{D}$ que contem P, é representado por $|\mathcal{T}(P)|$. O suporte é a probabilidade de um itemset P ocorrer na base $\mathcal{D}$. Sua definição é dada por:*

$$sup(P) = \frac{|\mathcal{T}(P)|}{n}, \tag{2.1}$$

onde $n = |\mathcal{D}|$ é o número de transações em $\mathcal{D}$.

O itemset fechado evita a extração de itemsets redundantes, ou seja, itemsets que não agregam nenhum conhecimento novo. Por exemplo, dado os itemsets fechados $I = \{b, c\}$ e $I' = \{a, b, c\}$, onde $sup(I) = sup(I')$, logo podemos afirmar, neste caso, que I' é itemset fechado e o itemset I é redundante pois não agrega informação nova. Sua definição formal é dada por:

Definição 4 (Itemset fechado) *Um itemset P em $\mathcal{D}$ é um itemset fechado apenas se não existir um itemset $P' \supset P$ em $\mathcal{D}$, tal que $sup(P') = sup(P)$ [50]. O fechamento de um itemset Q em $\mathcal{T}$ é definido por $clo(Q) = \bigcap_{t \in \mathcal{T}(Q)} t$.*

O valor de suporte do itemset tem a propriedade anti-monotônica, isto é, o suporte de um itemset é sempre menor ou igual ao suporte de qualquer dos seus subconjuntos:

Definição 5 (Anti-monotônico) *O suporte de um subconjunto X sempre será maior ou igual que o suporte de seu superconjunto Y, ou seja:*

$$X \subseteq Y \Rightarrow sup(X) \geq sup(Y). \tag{2.2}$$

O valor de suporte define a frequência do itemset na base de dados. No entanto, a métrica de suporte não fornece a característica de correlação entre os itens que compõem um itemset importante para identificar itemsets que vão gerar regras de associação interessantes. O *all-confidence* [48] define qual a confiança mínima de todas as regras de associação possíveis geradas a partir do itemset, segue a definição formal:

Definição 6 (*All-confidence*) *é calculada por:*

$$allconf(P) = \frac{sup(P)}{\max\limits_{a}(sup(a))}, \tag{2.3}$$

onde $a \in P$. Altos valores de all-confidence evidenciam que o itemset gera regras de associação relevantes.

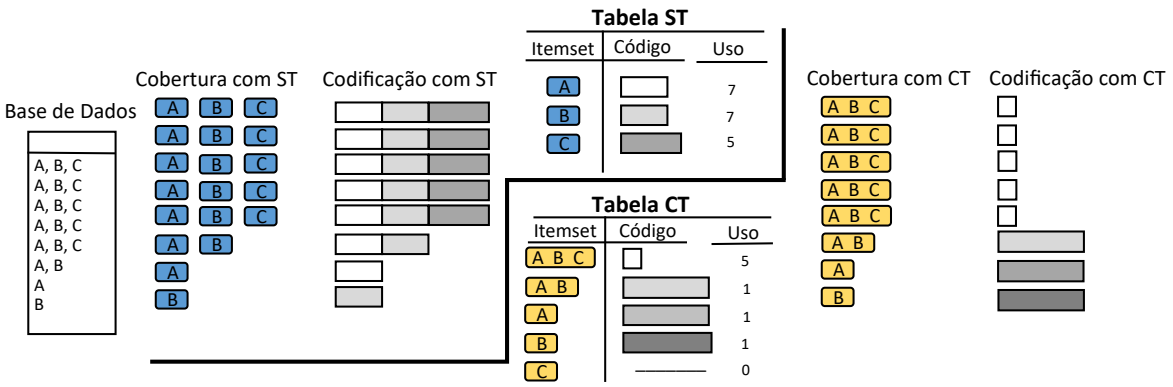


Figura 2.1: Representa a compressão de uma base de dados usando o conjunto CT de itemsets minerados. Comparando com os itemsets de tamanho 1 ST.

A métrica *cross-support* [63] define a proporção do suporte do item menos frequente para o suporte do item mais frequente dado um itemset, ou seja:

Definição 7 (*cross-support*) *é calculada por:*

$$crossSup(P) = \frac{\min_a(sup(a))}{\max_a(sup(a))}, \tag{2.4}$$

onde $a \in P$. Baixos valores de cross-support evidenciam itemsets contendo itens com níveis de suporte substancialmente diferentes.

De forma global, a métrica MDL [52] calcula a relevância de um conjunto de itemsets, onde a relevância desse conjunto é traduzida para termos de valor de compressão da base de dados. Segue a definição formal do problema:

Definição 8 (Comprimento Mínimo de Descrição) *O MDL (minimum description length) indica a compressão da base de dados $\mathcal{D}$ em bits tendo como referência para compressão a tabela de códigos CT. Esta tabela é gerada para atender:*

$$L(\mathcal{D}, CT) = L(CT) + L(\mathcal{D} \mid CT), \tag{2.5}$$

onde CT é uma coleção de itemsets, $L(CT)$ é o tamanho da descrição de CT, e $L(\mathcal{D} \mid CT)$ é o tamanho da descrição da base de dados quando codificado por CT. A ideia principal do MDL é que a tabela de código que gera a menor codificação de bits é a escolhida como sendo a melhor que descreve a base de dados.

O tamanho total relativo comprimido é dado como a relação entre o MDL de $\mathcal{D}$ comprimido por CT e pela tabela de código padrão ST composta por conjuntos de itens com

cardinalidade 1:

$$L\% = \frac{L(\mathcal{D}, CT)}{L(\mathcal{D}, ST)}. \tag{2.6}$$

De forma objetiva, na Figura 2.1 temos o exemplo de como são usados os itemsets minerados para fazer o cálculo da compressão relativa para determinada base de dados. Inicialmente são usados os itemsets de cardinalidade 1 (tabela ST) para calcular o valor de bit de cada itemset, no exemplo o número de bits é representado dado o tamanho do retângulo na coluna de código, para mais detalhes de como calcular os bits veja [52]. Para cada transação da base de dados, são selecionados os itemsets usados para representar os itens na transação. No caso da tabela ST, o número de itemsets usados é igual ao número de itens de cada transação. Por fim, para calcular o número final de bits de uma base de dados dado os itemsets, basta somar os valores de bits de cada itemset usado, como representado na codificação com ST. Por fim, o mesmo processo é feito para os itemsets selecionados por qualquer algoritmo de mineração. Perceba que na tabela CT, os itemsets selecionados tiveram um valor de bit menor, e o mesmo quando usado na cobertura com CT das transações. No somatório é notório ver que a tabela CT comprimiu mais a base de dados quando comparado com a tabela ST.

2.2 Definições para Bases de Dados de Fluxo Contínuo

A mineração de base de dados de fluxo contínuo difere da mineração tradicional em base de dados estática, visto que novos dados aparecem a cada momento. Essa característica traz novas formas de modelar o problema. Segue algumas definições formais nesse contexto:

Definição 9 (Base de dados de fluxo contínuo) *Seja $I = \{i_1, i_2, \cdots, i_m\}$ o conjunto universal dos itens e T_l a transação um conjunto de itens dado $T_l \subseteq I$. Uma janela W_c consiste em k transações, onde c é o identificador da janela que começa com o valor 1.*

Seja $D = [W_1, W_2, \cdots, W_N)$ a base de dados de fluxo contínuo onde N é o identificador da última janela.

Definição 10 (Tamanho corrente base de fluxo contínuo) *O tamanho corrente TC é dado por $k\,N$, ou seja $|W_1|+|W_2|+\cdots+|W_N|$. As janelas chegam em ordem cronológica e devem ser vistas apenas uma vez.*

Definição 11 (Tamanho corrente Itemset) *O tamanho corrente itemset TCI é dado por $(N - j + 1)$, ou seja $|W_j| + |W_{j+1}| + \cdots + |W_N|$, onde W_j é a primeira janela que contém o itemset I.*

Um desafio no processo de mineração de itemset frequentes vem do fato que itemset não frequentes devem ser armazenados, pois, esses itemsets podem ser tornar frequentes mais tarde. Logo, para garantir a integridade dos itemsets frequentes selecionados para determinada base de dados em fluxos contínuo, é necessário armazenar não apenas as informações relacionadas aos itemsets frequentes, mas também relacionadas aos pouco frequentes. Visto que se esses itemsets em algum momento se tornassem frequentes, seria impossível descobrir seu suporte, visto que as informações sobre esses itemsets não frequentes não foram armazenadas, consequentemente essa informação estaria perdida. Dado este cenário, os algoritmos de mineração em base de dados de fluxo contínuo possuem dois parâmetros de entrada, o primeiro chamado de suporte mínimo $S \in (0,1)$ e o segundo de suporte máximo de erro $\varepsilon \in (0,S)$.

Definição 12 (Frequência) *A frequência verdadeira $tfreq(X)$ de um itemset X é dado pelo número de transações na base de dados DS que contém o itemset X como subconjunto. Segue definição formal:*

$$tfreq(I) = \frac{|\{\mathcal{W}_c \in D \mid \mathcal{T}_l \in \mathcal{W}_c, I \subseteq \mathcal{T}_l\}|}{TC}, \tag{2.7}$$

A frequência estimada $efreq(X)$ de um itemset X é dado pelo valor estimado de X contido na estrutura em memória, onde $0 < efreq(X) \leq tfreq(X)$.

$$efreq(I) = \frac{|\{\mathcal{W}_c \in DI \mid \mathcal{T}_l \in \mathcal{W}_c, I \subseteq \mathcal{T}_l\}|}{TCI}, \tag{2.8}$$

onde a base começa com a ocorrência do itemset I, ou seja, $DI = [W_j, W_{j+1}, \cdots, W_N)$.

Em base de dados de fluxo contínuo o itemset é categorizado em três tipos: itemset frequente, itemset semi-frequente e itemset pouco frequente. Veja a definição formal:

Definição 13 (Tipo de frequência) *Sendo assim, um itemset X é frequente se $tfreq(X) \geq S\,TC$. Um itemset X é um itemset semi-frequente se $S\,TC > tfreq(X) \geq \varepsilon$. Um itemset X é infrequente se $\varepsilon\,TC > tfreq(X)$.*

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta a evolução de algoritmos de mineração de itemsets em bases de dados estáticas com abordagens de processamento sequencial e paralela, para entender as vantagens e desvantagens de cada técnica. O escopo deste estudo foca em algumas vertentes de pesquisa importantes para o entendimento desta tese. Os focos de interesses encontrados na literatura são: a escalabilidade do algoritmo, e controle da quantidade de itemsets reportados. Este capítulo propõe discutir as diferentes abordagens e entender melhor as estratégia para a mineração de itemsets frequentes, dado o uso da métrica de suporte mínimo ou, também, por outras métricas de interesse (Seção 3.1).

Outro cenário de estudo é a mineração de itemset em base de dados de fluxo contínuo. Ao contrário da base de dados estática, esta base de dados recebe novas informações (transações) a cada faixa de tempo, de forma contínua. Os principais desafios desse tipo de base de dados são: a estrutura usada para processamento, o tempo de processamento, e a base de dados só pode ser percorrida apenas uma vez. Este estudo mostra alguns algoritmos para este tipo de base de dados (Seção 3.2).

3.1 Itemsets em Bases de Dados Estáticas

O algoritmo sequencial Apriori [3] emprega a propriedade anti-monotônica (Definição 5) da medida de suporte (Equação 2.1) para localizar itemsets frequentes. Dados uma base de dados e o valor de suporte mínimo, o processo de mineração começa com um conjunto de candidatos a itemsets de tamanho um (todos os itens da base de dados), em seguida é necessário percorrer a base de dados para obter a frequência dos mesmos. São selecionados apenas os itemsets que possuem o suporte maior que o mínimo de suporte especificado. Logo após, estes itemsets frequentes de tamanho 1 são usados na geração dos candidatos

a itemset frequente de tamanho 2. O processo continua até que não possam ser incluídos candidatos a itemsets de tamanho $k+1$ dado os itemsets frequentes de tamanho k. A ideia desse algoritmo traz, inicialmente, dois grandes desafios para o processo de mineração. O primeiro é a necessidade de percorrer a base de dados repetidamente. Outro aspecto é a geração de um grande conjunto de candidatos a itemset frequente durante o processo de mineração.

O algoritmo sequencial *AprioriTid*, proposto por Agrawal et al. [4], tem como ideia reduzir o custo ao percorrer a base de dados repetidamente. Para isso, a base de dados D não é utilizada para a contagem do suporte, e sim a base $D_k^{'}$. A base $D_k^{'}$ possui o seguinte formato $< TID, X_k >$, onde k representa o tamanho do itemset, e X_k é o k-itemset de itens presentes na transação representada com o identificador único chamado TID. A base $D_k^{'}$ pode conter menos transações que a base original D, especialmente para grandes valores de k, pois estes itemsets, normalmente, não estão representados em muitas transações. Assim como Apriori, este algoritmo também usa o processo de geração de candidatos a itemset. Durante o estudo, o autor identificou que para bases muito grandes o algoritmo é ineficiente quando comparado com Apriori. Isso acontecia porque a base $< TID, X_k >$, para valores menores de k, nem sempre pode ser alocada na memória principal. Neste mesmo estudo, o autor propõe o algoritmo *AprioriHybrid*, que utiliza inicialmente o Apriori e depois troca para o *AprioriTid* no momento que $D_k^{'}$ consegue ser alocado na memória principal.

Com o intuito de reduzir a quantidade de itemsets candidatos, foi proposto o algoritmo sequencial DHP (*standing for direct hashing and pruning*) [49]. Durante o processo de percorrer a base de dados para obter o valor de suporte para os itemsets de tamanho k, o DHP já reúne informações sobre os itemsets de tamanho $k + 1$ inserindo-os em uma tabela de *hash* contendo a informação de contador de ocorrência de determinada *hash*, que pode ser traduzido para o valor de suporte. Desta forma, os itemsets candidatos de tamanho $k+1$ serão aqueles que foram mapeados para *hash* que possuírem suporte maior que o mínimo especificado. Fournier-Viger et al. [23] fizeram um estudo que contém mais informações de variações do algoritmo Apriori com diferentes heurísticas.

Outra vertente de algoritmo sequencial para mineração de itemsets frequentes foi proposta por Han et al. [28], onde os autores definem uma estrutura chamada FP-*tree* (árvore de padrões frequentes), veja Figura 3.1. Nesse algoritmo só é necessário percorrer a base duas vezes. No primeiro momento a base é percorrida para obter valores de frequência de cada item da base de dados, logo após é construída a tabela *header* que contém os

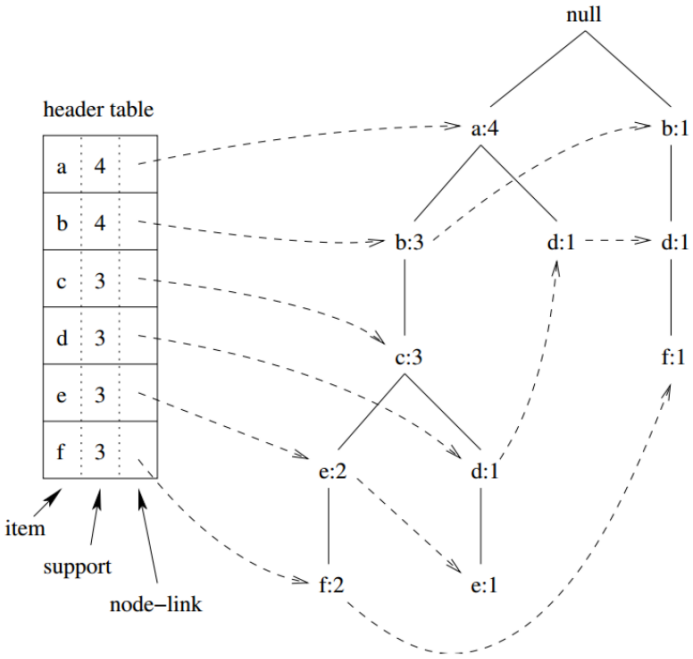


Figura 3.1: FP-*tree* representa a forma compacta da base de dados. A tabela *header* contém todos os itens da base de dados, o valor do suporte e o link para os nós que representa a ocorrência do item na base. Exemplo do artigo [28].

itens ordenados de forma decrescente por frequência e seus respectivos valores de suporte para cada item. Nesse momento os links para os nós da árvore não estão definidos. No segundo momento, antes de percorrer a base de dados, a estrutura é inicializada com o nó raiz, chamado *null*. Depois, para cada transação, os itens são ordenados segundo a tabela *header* pelo valor de suporte. Consequentemente, ao inserir os itens da transação na estrutura FP-*tree*, os nós perto da raiz serão os mais frequentes. Cada nó contém a informação do item e o contador que define o número de transações que contém este nó. Um novo nó a ser adicionado por uma transação gera o incremento, contador mais 1, de cada nó ao longo do seu prefixo em comum, e os nós dos sufixos que não existem são criados e conectados. Para o processo de busca de itemsets frequentes, o procedimento recursivo FP-*Growth* percorre os nós da FP-*tree* condicional de baixo para cima. Para cada item i na tabela *header* são recuperados todos os ramos que contém o item i até a raiz. Com estes ramos é criado a FP-*tree* condicional ao prefixo i.

A estrutura FP-*Tree* foi considerada uma grande evolução para o algoritmo de mineração de itemset, pois com apenas duas passadas pela base torna-se possível gerar uma informação compacta da base de dados e, além disso, eliminar a necessidade do processo de geração de candidatos a itemsets. Ao passo que, para bases grandes e esparsas, a estrutura torna-se pouco compacta e lenta, consequentemente, impossibilitando a aloca-

ção da mesma na memória principal. Grahme e Zhu [27] em seu estudo perceberam que em média 80% do tempo gasto da CPU é gasto percorrendo as FP-*trees*, por isso em seu trabalho os autores definiram uma estrutura auxiliar, i.e., FP-*array*, que elimina a necessidade de percorrer a FP-*Tree* para o processo de achar os ramos necessários para construir as FP-*tree* condicionais. Outros algoritmos sequenciais que tem como base a estrutura FP-*tree* podem ser encontrados no livro de Aggarwal et .al [2].

Outro avanço na recuperação de itemsets foi proposto por Uno et al. [58], onde é definido o algoritmo sequencial LCM (*linear time closed pattern miner*) para identificação de itemsets fechados frequentes. Neste algoritmo, o custo da identificação de itemsets fechados frequentes é em tempo linear, dada a quantidade de itemsets fechados reportados. A função PPC (*prefix preservinf closure*) garante uma busca direcionada para a base de dados, consequentemente, conseguindo enumerar todos os itemsets fechados frequentes sem reportar itemsets fechados repetidos. Com essa função não é preciso guardar em memória os itemsets reportados. Desta forma, o uso de memória para o algoritmo LCM depende exclusivamente do tamanho da base de dados. Para reduzir o custo de verificação do itemset na base de dados a função *anytime database reduction* define uma estrutura de array T que contém informações de transações definidas dado o prefixo itemset da recursão corrente, criando projeções reduzidas da base de dados a cada recursão. Além disso, a função *occurrence deliver* tem como finalidade recuperar o valor de suporte real usando a estrutura T atual construída pela função *anytime database reduction*.

Diferente da estrutura FP-*tree*, a estrutura LCM tem um custo computacional menor e uma estrutura mais simples de atualização. Benjamin e Takeaki [43] propuseram uma solução paralela com memória compartilhada para o algoritmo LCM. O algoritmo utiliza uma interface para implementação paralela chamada MELINDA. As paralelizações acontecem nas seguintes etapas: na redução da base de dados, na criação da tabela de entrega de ocorrência e no processo de recursão ao percorrer a árvore de busca da base de dados.

Com relação ao uso de restrições para controlar o número de itemsets detectados, o limiar de mínimo suporte min_{sup} [3, 28, 27, 58] pressupõem, implicitamente, que todos os itens da base de dados são da mesma natureza ou têm uma frequência semelhante. No entanto, alguns grupos de itens relacionados podem aparecer com muita frequência na base, enquanto outros raramente aparecem. Este último grupo levaria à formação de itemsets infrequentes interessantes. Quando as frequências dos itens variam muito, é preciso lidar com a escolha entre gerar um grande número de itemsets, definindo um valor pequeno de min_{sup}, ou gerar menos itemsets, aumentando o valor de min_{sup}, e assumindo

o risco de perder itemsets interessantes pouco frequentes.

Liu et al. [35] propôs o algoritmo sequencial *MSApriori*. Essa técnica atribui um valor de MIS (*minimum item supports*) para cada item da base. Durante o processo de mineração de itemsets frequentes, o itemset só é selecionado se o seu valor de suporte for maior que o menor valor de MIS dentre os itemsets que o compõem. Os parâmetros para esse algoritmo são o valor de suporte mínimo e o valor proporcional de suporte β. Em seu estudo, os autores mostram que a técnica cumpre com a finalidade de retornar itemsets interessantes em diferentes faixas de suporte. No entanto, para bases onde o suporte dos itens tem grandes variações esse comportamento não se manteve. Para resolver este problema foi proposto por Uday et al. [56] uma alteração na equação MIS onde o valor de β é alterado por valores de média e desvio padrão de frequência dos itens da base.

Omiecinski [48] definiu o cálculo de três métricas: *any-confidence*, *all-confidence* (Equação 2.3) e *bond*. Essas métricas são usadas para o processo de seleção de itemsets interessantes, embora o foco do artigo seja para o processo de criação de regras de associação interessantes. O algoritmo paralelo TopPI [30] é considerado uma evolução do algoritmo Top-k [62], cuja característica é encontrar os top k itemsets para cada item da base de dados. O TopPI usa o algoritmo LCM para acelerar o processo de mineração e requer a definição de dois valores de parâmetro: k e suporte conjuntivo mínimo (min_{csup}, Definição 3).

Os limiares complementares pode impedir a formação de itemsets desnecessários quando um valor mínimo de suporte é escolhido. A desvantagem inerente ao processo é a dificuldade de escolher valores apropriados para estes parâmetros. O algoritmo sequencial Krimp [61] usa a métrica MDL (Equação 2.5) para identificar o conjunto de itemsets que compactam melhor a base de dados. Apesar de Krimp conter o parâmetro mínimo suporte, o próprio autor define que o melhor parâmetro será o valor de suporte conjuntivo igual a 1. O aspecto negativo desta técnica diz respeito a sua escalabilidade para bases muito grandes. O algoritmo paralelo Slim [55] tem uma heurística que utiliza o espaço de busca da base de dados, no processo de seleção dos conjuntos de itemsets candidatos, usados para o cálculo do MDL. Como consequência, há uma melhora no tempo para convergir e em melhores valores de compressão do conjunto de itemsets selecionado. O algoritmo paralelo LAM (*localized approximate miner*) proposto por Gregory et al. [11], propõe o uso de *min-hash* para aproximar porções da base de dados e, com isso, ser possível propor um conjunto de itemsets fechados que melhor comprime a base de dados.

3.2 Itemsets em Bases de Dados de Fluxo Contínuo

Uma base de dados transacional de fluxo contínuo é caracterizada como sendo uma sequência de transações de entrada onde cada trecho deste fluxo é chamado janela. Os algoritmos nessa área são divididos em três modelos de mineração de itemset, onde cada um define a regra de como o novo dado deve ser processado [1]. *Landmark* considera todos os dados que chegam na base a partir do início até a janela atual e cada janela é tratada com a mesma importância. *Damped* ou *fading* também considera todo o dado desde o início, porém neste caso as janelas são divididas em tempos que possuem diferentes pesos. Neste modelo, janelas mais recentes tem mais importância. *Sliding model* processa apenas parte dos dados, apenas aqueles que estão na janela corrente. Desta forma, dados de transações antigas são excluídas quando uma nova transação chega.

Manku e Motwani [38] propuseram o algoritmo *lossy counting* para recuperar conjuntos de itemsets frequentes dado o suporte mínimo Min_{sup} e o suporte máximo de erro ε em uma base de dados em fluxo contínua dividida em janelas (Definição 9). O algoritmo mantém uma estrutura *hashing* $\mathcal{D}$, que mantém a informações dos itemsets que chegam da base de dados em janelas. Essa estrutura é representada por uma tupla no seguinte formato (X,$efreq(X), err(X)$), onde X é um itemset, $efreq(X)$ (Equação 2.8) a frequência estimada de X dada sua primeira ocorrência e $err(X)$ (Definição 13) o limite superior da frequência de X. Existem três operações nessa estrutura: atualizar, adicionar e remover tupla. A implementação deste algoritmo é dividida em três módulos: (i) *Buffer* responsável em lidar com os dados que chegam em janela, fazendo a remoção de itens que não são frequentes dado o suporte máximo de erro especificado, depois ordena os itens em ordem lexicográfica; (ii) *Trie* mantém a estrutura $\mathcal{D}$ como uma estrutura floresta de prefixo, onde o nó v de um prefixo representa uma tupla $(X, efreq(X), err(X))$; (iii) *setGen* gera os itemsets representados em cada janela. Os autores empregam uma estrutura de fila chamada *Heap* onde é possível podar em forma recursiva itemsets dado os limiares informados pelos parâmetros de entrada.

Li et al. [33] definiram uma estrutura chamado *Item-suffix Frequent Itemset forest* (*IsFI-Forest*) que dá suporte ao algoritmo DSM-FI proposto pelo autor. Assim como *lossy counting*, este algoritmo usar os parâmetros, suporte mínimo Min_{sup} e suporte máximo de erro ε. As informações chegam em janela contendo um conjunto de transações, para cada transação y o algoritmo proposto ordena de forma lexicográfica e depois aplica a função *transation projection* (TP) que cria sub-transações x, por exemplo, dado a transação $< a,c,d,f >$ aplicando a função $TP(< a,c,d,f >)$ temos as sub-transações $< a,c,d,f >$,

$< c,d,f >$, $< d,f >$, $< f >$. Cada sub-transação é associado a um *header table* (HT) e SFI-Tree. A inserção nessa estrutura tem os seguintes parâmetros (x,$efreq(x)$,$id - janela, link$), onde é inserido a sub-transação x em na estrutura HT ou incrementado a $efreq(x)$ caso exista na estrutura, $id - janela$ é o id da janela corrente e o $link$ aponta para a primeira ocorrência de determinado item na estrutura. Essa estrutura é idêntica à estrutura FP-*Tree*. No final do processamento de cada janela é aplicada a função de remoção de itemsets não frequentes onde $efreq(X) < \varepsilon TC$. Nesses casos a estrutura SFI-Tree de prefixo X é removida e em alguns casos necessita de atualizar as outras estruturas que contém o prefixo X.

Para lidar com a quantidade massiva de itemsets frequentes, Liu et al. [36] propuseram o algoritmo FP-CDS que retorna itemsets fechados. Dado uma base de dados de fluxo contínuo, o algoritmo mantém um conjunto global de itens, chamado F_{list}, e para cada janela que chega o algoritmo atualiza essa estrutura. Para cada nova janela é criada uma estrutura de árvore *FP-CDS Tree*. Com uma condição, para a janela corrente $i > 1$ cria uma árvore *FP-CDS Tree* com os dados da *FP-CDS Tree* da janela anterior $i - 1$, sendo que essas informações são inseridas na nova árvore obedecendo à organização da F_{list}. Os itens que não estão incluídos em f_{list} (a.k.a., lista global com informação de frequência de cada item da base de dados) são ignorados e restante é inserido na nova árvore. Logo após a árvore da janela $i - 1$ é excluída. O algoritmo FP-CDS define uma nova métrica para remover itens infrequentes do processo de mineração. No final de cada janela, ele percorre cada elemento da F_{list} e decrementa um atributo chamado *del* de um item. No final de cada janela, se algum item tiver valor *del* igual a zero, este elemento é eliminado do processo de mineração. Já no caso do algoritmo FP-CDS quando são requisitados os itemsets fechados para determinada janela corrente é chamada a função *CDSgrowth*, que percorre a estrutura *FP-CDS Tree* em busca de itemsets fechados que respeitam o valor de mínimo suporte.

Um dos grandes desafios para um especialista de dados é definir o limiar de suporte mínimo ideal, pois isso requer algum conhecimento da base a ser estudada. Yang e Huang [64] propuseram um algoritmo *TOPSIL-Miner* que retorna os top-K itemsets significativos para determinada janela corrente. Foi definida uma estrutura chamada *TOPSIL-Tree* que armazena possíveis itemsets significativos de cada janela recebida, uma estruturas de lista de suporte máximo (*MaxSL*) com tamanho $L = \lceil log_2(k + 1) \rceil$, uma estrutura ordenada de forma decrescente por suporte dos itens (*OIL*), a estrutura *TOPSET* que contém os k itemsets significantes e a lista de suporte mínimo (*MinSL*) com tamanho n que representa o número de janelas recebidas. O algoritmo *TOPSIL-Miner* usa a lista

MinSL como limite de erro para podar itens que não são frequentes. O procedimento *tree incremet* cria a estrutura *TOPSIL-Tree* usando a estrutura da janela anterior mais as informações de OIL e MaxSL para definição do limiar de corte para as novas informações que serão inseridas na *TOPSIL-Tree* .

3.3 Discussão

Durante este capítulo foram elencados alguns trabalhos sobre mineração de padrões. Este trabalho evidencia apenas algumas vertentes que são de interesse desta tese. A primeira vertente trabalha com algoritmos sequenciais para a identificação de itemsets, o algoritmo foi inicialmente proposto por Agrawal et al. [3] que utiliza o suporte mínimo para reportar itemsets frequentes, e desde então são propostas melhorias para este problema de mineração. Os algoritmos nessa vertente propõem soluções para lidar com o custo de percorrer a base de dados [34, 27, 58], outras propostas que visam reduzir o número de candidatos a itemsets [49], casos onde é retirada a etapa de criação de candidatos a itemset [27], trabalhos com a finalidade de reduzir a quantidade de itemsets reportados, retornando aqueles sendo considerados mais relevantes dada uma métrica [58, 61] e por último trabalhos que têm como foco usar outros limiares sem ser o suporte ou em alguns casos nem usar limiares [35, 61].

Não foi mencionado neste capítulo, por não se tratar de um algoritmo de itemset, ainda assim é importante ressaltar, que existem dois trabalhos que introduziram a ideia de usar contextualização espacial na área de mineração de regras de associação. Bruzzese e Davino [10] criaram um visualizador que ajuda o especialista a selecionar regras que são mais relevantes das mineradas pelo algoritmo de regra de associação, usando representação gráfica em duas dimensões dos pontos dos itens da base. Já no trabalho proposto por Fernandes e Garcia [20] é proposto um processo similar ao representar graficamente as duas dimensões os itens da base de dados, porém eles usam clusters reportados pelo algoritmo de clusterização para eliminar itens da base no processo de mineração. O usuário participa na tomada de decisão na hora de selecionar os itens a serem eliminados. O problema dessas duas abordagens está no fato selecionar apenas as duas dimensões do espaço de múltiplas dimensões, traz perda de informação na projeção, consequentemente podendo gerar uma interpretação errônea do comportamento da base. No trabalho de mestrado, Mantuan [39] propôs o estudo onde utiliza técnicas de análise multidimensionais como critério de decisão para recuperação de itemsets fechados da bases de dados. O algoritmo de geração de itemset proposto gera todas as possíveis combinações de itemset dado

um cluster e, por fim, percorre a base de dados para obter as informações de suporte (Equação 2.1) e all-confidence (Equação 2.3) de cada itemset selecionado. Na parte de comparação com outras técnicas foi selecionado apenas o algoritmo base Apriori [3]. As métricas de comparação utilizadas foram: (*i*) a quantitativa, ou seja, o quanto o algoritmo reportou de itemsets em cada faixa de suporte definida por base; (*ii*) *all-confidence* médio dos itemsets reportados em cada faixa de suporte; e a (*iii*) verificação de identificação de itemsets raros, ou seja, itemsets com pouca frequência na base de dados, porém com alto valor de *all-confidence*, comparado com o algoritmo CORI [8].

Outra vertente apresentada neste capítulo foram os algoritmos paralelos, mais especificamente, propostas para arquiteturas de multinúcleo. Com demanda de grandes bases de dados, os algoritmos paralelos ganham cada vez mais interesse. A primeira proposta abordada foi o uso da estrutura LCM [43]. Nesse estudo o algoritmo retorna os itemsets fechados dado o valor de suporte mínimo. Esse algoritmo resolve o problema de geração de candidatos de itemset fechados, visto que ele não tem essa etapa no processo de mineração. Além disso, LCM reduz o processo de passadas pela base de dados. No segundo momento, foi apresentado o TopPI [30], que utiliza a métrica top-k para retornar itemsets fechados relevantes. Por fim, foram apresentadas propostas que se consideram como livre de parâmetro, ou seja, dada uma base de dados, o algoritmo consegue retornar um conjunto que melhor comprime a base de dados [11, 55].

A última vertente escolhida para este estudo foi a mineração de itemsets frequentes em base de dados de fluxo contínuo. Os desafios observados nessa área são: (*i*) transações só podem ser obtidas apenas uma vez, e essas transações vêm em conjuntos chamados janela; (*ii*) o uso de memória para armazenar as informações, visto que novas informações de elementos aparecem a cada momento; (*iii*) o tempo de resposta para a atualização das estruturas em memória deve ser rápido; e (*iv*) os itemset frequentes devem ser retornados de forma rápida, com uma segurança de no mínimo de erro ε. Manku e Motwani [38] apresentaram o primeiro algoritmo nesta área, o *lossy counting*, que garante a identificação dos itemsets frequentes, não existe falso negativo, ou seja, ser frequente e não ser reportado. Outro algoritmo proposto, o DSM-FI [33], usa uma estrutura mais compacta que o FP-Tree, criando uma SFI-Tree para cada sub-transação. O problema dessa estrutura é o seu custo de atualização. FP-CDS [36] é um algoritmo para retornar itemset fechados, reduzindo assim o número de itemsets redundantes. A proposta TOP-SIL [64] propõe retornar os k itemset mais importantes sem a necessidade de passar o mínimo suporte.

Capítulo 4

Dual Scaling

Para fins analíticos, nesta tese iremos usar como técnica de análise multidimensional o algoritmo *Dual Scaling*, proposto por Nishisato como uma ferramenta para inspeção visual de indivíduos e suas preferências a estímulos coletados através de pesquisas de opinião [46]. Com o mapeamento do *Dual Scaling*, cada sujeito e estímulo pesquisados são representados por um ponto no espaço de soluções (também conhecido como Espaço de Estilo de Resposta). Preferências e comportamentos de grupos de sujeitos com opiniões semelhantes emergem da distribuição dos pontos no espaço de soluções, visto que os estímulos e os sujeitos relacionados são mapeados próximos uns dos outros, enquanto os não relacionados são mantidos distantes no espaço da solução.

Apesar de o algoritmo ter sido proposto originalmente para análise de preferências de indivíduos, Nishisato afirma que *Dual Scaling* pode ser empregado para descobrir estilos de respostas em praticamente todas as naturezas de dados, não se restringindo apenas a dados de julgamento humano [46]. Para validar sua afirmação, Nishisato aplicou o *Dual Scaling* em dados de diferentes naturezas, tais como: distribuições ecológicas de espécies animais em diferentes terrenos, dados sobre poluição do ar e condições climáticas.

4.1 Dados Categóricos

A base de dados categórica é caracterizada por ter finitos valores para determinada variável e seus valores podem não ter uma ordem lógica (e.g., tipo gênero e tipo material). *Dual Scaling* é um método versátil que analisa uma variedade de tipos de dados categóricos, incluindo dados por ordem de classificação (*incidence data*), comparações pareadas (*paired-comparasion*), categorias sucessivas (*sucessive categories data*), tabelas de contingência (*contigency tables*), dados de múltipla escolha (*multiple-choice data*) e dados de

Base de dados de múltiplas escolhas

Pressão			Enxaqueca			Idade			Ansiedade 1		
Baixa (1)	Normal (2)	Alta (3)	Raro (4)	As vezes (5)	Sempre (6)	20-34 (7)	35-49 (8)	50-65 (9)	Baixa (10)	Normal (11)	Alta (12)
1	0	0	1	0	0	1	0	0	1	0	0
0	1	0	0	1	0	1	0	0	0	0	1

Matriz padrão de resposta *Dual Scaling*

1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	1	0	0	1	0	0	1	0	0
0	1	0	0	1	0	1	0	0	0	0	1

Base de dados transacional

Produtos
Carne (1)
Queijo (2)
Açúcar (3)
Leite (4)
Pão (5)
Manteiga (6)
Sabão (7)
Amaciante (8)

	Carne, leite, manteiga
	Queijo, Pão
	Sabão, Amaciante

Matriz padrão de resposta *Dual Scaling*

1	2	3	4	5	6	7	8
1	0	0	1	0	1	0	0
0	1	0	0	1	0	0	0
0	0	0	0	0	0	1	1

Figura 4.1: Exemplo de dois formatos bases de dados: múltiplas escolhas e transacional. Para cada formato de base de dados é apresentado a conversão para a matriz de resposta padrão usada pelo algoritmo *Dual Scaling*. Os números entre parênteses na frente de cada atributo representam o indexador único para as colunas da matriz padrão de resposta.

ordenação (*sorting data*). Veja [46] para mais informação sobre o formato de cada tipo de base de dados.

Nishisato classificou os dados categóricos em dois grupos distintos: (*i*) os dados de incidência (*incidence data*), grupo que abrange as tabelas de contingência (*contingency tables*), os dados de múltipla escolha (*multiplechoice data*) e os dados ordenados (*sorting data*); e (*ii*) os dados de dominância (*dominance data*), grupo este que abrange os dados por ordem de classificação (*rankorder data*), os dados de comparação pareada (*paired-comparison data*) e os dados de categorias sucessivas (*successive categories*). Dado o escopo deste trabalho, são apresentados apenas os dados de múltipla escolha e transacional (Figura 4.1). Os dados do tipo múltipla escolha consistem na apresentação de uma série de alternativas, onde apenas uma será escolhida, atribuindo à alternativa escolhida o valor 1, enquanto as demais alternativas recebem o valor 0. A Figura 4.1 apresenta um exemplo onde a questão pressão tem três alternativas (Baixa, Normal e Alta) e apenas uma pode ser escolhida a cada transação, por consequência, a quantidade de itens por transações é fixa. Já no caso de base de dados transacional, a quantidade de itens por transação pode ser variada. O resultado então é apresentado em uma tabela chamada tabela de padrão de respostas (*response-pattern table*). Caso seja de interesse, os demais tipos de dados podem ser consultados no livro de Nishisato [46].

4.2 Algoritmo *Dual Scaling*

O algoritmo trata a base de dados $\mathcal{D}$ como dados de múltipla escolha com q questões com m_i opções de resposta cada ou como dados transacionais. A abordagem representa as entradas da base de dados em uma matriz F de padrão de resposta $(1, 0)$ de tamanho $n \times m$, onde cada transação é um sujeito (linhas da matriz) e os itens são organizados como possíveis estímulos (respostas de múltipla escolha, i.e., colunas da matriz ou ocorrência de itens). Assim, o número total de colunas/itens, no caso de base de dados múltiplas escolhas, é $m = \sum_{i=1}^{q} m_i$. As seguintes equações aplicam o *Dual Scaling* à matriz F. No espaço de solução mapeado por *Dual Scaling* pode-se citar duas propriedades importantes [45, 46, 47]:

Propriedade 1 (Distância do item a origem) *A distância dos pontos dos itens até a origem do espaço de solução é inversamente proporcional à sua frequência na base de dados. Em outras palavras, os itens mais frequentes estão mais próximos da origem do que os itens menos frequentes.*

Propriedade 2 (Distância entre items) *A distância entre os itens diminui à medida que a frequência relativa de coocorrência de itens aumenta.*

Seja D_r e D_c matrizes diagonais de, respectivamente, frequência da marginal de linhas e colunas de F. Dado $\lambda = [\lambda_1, \cdots, \lambda_m]$ e V que são, respectivamente, o conjunto de autovalores e uma matriz cujas colunas são os autovetores direitos da matriz correspondente

$$M = F^T D_r^{-1} F D_c^{-1}, \tag{4.1}$$

de tal modo que $1 = \lambda_1 \geq \cdots \geq \lambda_m$. Na Equação 4.1, dada uma matriz A temos que A^T e A^{-1} denota, respectivamente, a transposição e inversão da matriz A.

Dual Scaling calcula as coordenadas dos pontos (a.k.a., itens) no espaço de solução e é representada pelas linhas da matriz x com tamanho $m \times n_s$. A matriz x é chamada pesos projetados para as colunas de F. As coordenadas dos pontos das transações no espaço de solução são representadas pelas linhas na matriz y com tamanho $n \times n_s$. A matriz y contém os pesos normalizados das linhas de F, onde n_s é o número máximo de dimensões do espaço de soluções. Para o caso de base de dados de múltiplas escolhas, temos que $n_s = m - q$, e para base de dados transacional n_s é a quantidade total de autovalores maiores que 10^{-8}. Neste trabalho, não é apresentado como calcular a matriz y porque

o projeto não utiliza essas coordenadas em nenhum processo dos algoritmos propostos. No entanto, é importante enfatizar que o cálculo dos pontos x_i e y_i no espaço de solução ocorre no conjunto de estímulos (coluna) dado a coocorrência entre sujeitos (linhas) [46]. As entradas dos valores para o cálculo da matriz de peso projetado x é dado por:

$$x_{i,j} = \sqrt{\frac{\lambda_{j+1}}{c_{j+1}}} V_{i,j+1}, \tag{4.2}$$

onde $c_{j+1} = \frac{1}{t}\sum_{k=1}^{m} T_{k,j+1}$, $T = D_c\,(V \circ V)$, $A \circ B$ denota o produto de Hadamard da matriz A e B, e t é o número total de 1 repostas em F.

A métrica χ-quadrado define a distância ao quadrado entre pares de pontos:

$$d^2_{i,i'} = \sum_{k=1}^{n_s} \sqrt{\lambda_{k+1}} \left(\frac{x_{i,k}}{\sqrt{p_i}} - \frac{x_{i',k}}{\sqrt{p_{i'}}} \right)^2, \tag{4.3}$$

onde λ_{k+1} é o $(k+1)$-ésimo autovalor de M (Equação 4.1), $x_{i,k}$ e $x_{i',k}$ são a k-ésima coordenadas dos pontos de estímulos indexados por i e i' (Equação 4.2), respectivamente, p_i e $p_{i'}$ são proporções da frequência das colunas, *i.e.* $p_i = f_{c_i}/n$, onde f_{c_i} é a frequência marginal de respostas para o i-th item de F.

A distância ao quadrado $d^2_{i,i'}$ (Equação 4.3) está relacionada à distância entre pontos da teoria da quantificação [31]. Ela modula o quadrado da distância euclidiana entre pontos (ou itens) por pesos computados em função de suas frequências. Por exemplo, dado que o ponto x_i seja mais frequente do que o ponto $x_{i'}$, usando a métrica χ-quadrado, espera-se que o ponto médio esteja mais próximo de x_i do que $x_{i'}$. Isso porque a frequência age como massa. Assim, um ponto com uma massa maior tem uma atração maior do que um ponto com uma massa menor. O termo $\sqrt{\lambda_{k+1}}$ na Equação 4.3 é o índice de correlação associado à k-ésima dimensão do espaço de soluções. Ele codifica a importância dessa dimensão para a representação dos dados, pois os autovalores estão relacionados à variabilidade na direção de seus respectivos autovetores.

A angulação $a_{i,i'}$ (Equação 4.4) define valores de cosseno no intervalo $[-1, 1]$. A interpretação deste valor diz que quanto menor for a angulação entre pares de itens maior é a coocorrência entre estes itens quando comparado com itens com angulação maiores.

$$a_{i,i'} = \|x_i\|\|x_{i'}\| \cos\left(\angle x_i x_{i'}\right), \tag{4.4}$$

onde denota o produto escalar dos vetores x_i e $x_{i'}$, e cos é a função cosseno.

$\mathcal{D}$	F						fr
$\{1,4\}$	1	0	0	1	0	0	2
$\{1,4\}$	1	0	0	1	0	0	2
$\{3,4\}$	0	0	1	1	0	0	2
$\{3,5\}$	0	0	1	0	1	0	2
$\{3,5\}$	0	0	1	0	1	0	2
$\{3,5\}$	0	0	1	0	1	0	2
$\{3,5\}$	0	0	1	0	1	0	2
$\{3,5\}$	0	0	1	0	1	0	2
$\{3,6\}$	0	0	1	0	0	1	2
$\{2,5\}$	0	1	0	0	1	0	2
$\{2,6\}$	0	1	0	0	0	1	2
$\{2,6\}$	0	1	0	0	0	1	2
$\{2,6\}$	0	1	0	0	0	1	2
$\{2,4\}$	0	1	0	1	0	0	2
$\{2,6\}$	0	1	0	0	0	1	2
fc	2	6	7	4	6	5	

M

						λ
0,5000	0,000	0,000	0,2500	0,000	0,000	
0,000	0,5000	0,000	0,1250	0,0833	0,4000	0,8264
0,000	0,000	0,5000	0,1250	0,4167	0,1000	0,7899
0,5000	0,0833	0,0714	0,5000	0,000	0,000	0,2101
0,000	0,0833	0,3571	0,000	0,5000	0,000	0,1736
0,000	0,3333	0,0714	0,000	0,000	0,5000	

V				T			
0,4782	0,0710	0,0710	0,4782	0,4573	0,0101	0,0101	0,4573
-0,0986	-0,5513	-0,5513	-0,0986	0,0583	1,8239	1,8239	0,0583
-0,3796	0,4803	0,4803	-0,3796	1,0086	1,6148	1,6148	1,0086
0,6243	0,0824	-0,0824	-0,6243	1,5589	0,0271	0,0271	1,5589
-0,4405	0,4332	-0,4332	0,4405	1,1644	1,1260	1,1260	1,1644
-0,1838	-0,5156	0,5156	0,1838	0,1688	1,3292	1,3292	0,1688

x				d^2					
1,1329	0,1420	0,0732	0,5193	0	16,6052	20,5072	3,2338	21,9519	18,4428
-0,2336	-1,1021	-0,5684	-0,1071	16,6052	0	10,9090	14,0740	10,4710	1,7637
-0,8994	0,9600	0,4951	-0,4122	20,5072	10,9090	0	17,5109	1,8114	9,7269
1,4791	0,1647	-0,0849	-0,6780	3,2338	14,0740	17,5109	0	21,4269	17,5219
-1,0437	0,8659	-0,4466	0,4784	21,9519	10,4710	1,8114	21,4269	0	10,8555
-0,4354	-1,0306	0,5315	0,1996	18,4428	1,7637	9,7269	17,5219	10,8555	0

Figura 4.2: Exemplo da aplicação do *Dual Scaling* em uma base sintética $\mathcal{D}$.

4.3 Exemplo de Aplicação do Cálculo

Para exemplificar o cálculo do *Dual Scaling* é definida uma base de dados sintética. Na Figura 4.2, a base de dados sintética $\mathcal{D}$ é composta por $n = 15$ transações e $q = 2$ perguntas com $m_i = 3$ opções cada, levando a m $= 3 + 3 = 6$ itens. Após converter essa base de dados em uma matriz $(1, 0)$ de padrão de resposta $F_{15\times6}$, as frequências marginais de respostas para linhas ($f_{r_{15\times1}}$) e colunas ($f_{c_{1\times6}}$) definem os valores na diagonal principal das matrizes $D_{r_{15\times15}}$ e $D_{c_{6\times6}}$, respectivamente. Na próxima etapa, a matriz $M_{6\times6}$ é calculada usando a Equação 4.1.

Depois de calcular os autovalores (λ) e autovetores (V) de $M_{6\times6}$, é possível usar a Equação 4.2 para calcular a matriz $x_{6\times4}$ de pesos projetados. Segundo Nishisato [46], o número máximo de dimensões do espaço de solução será $n_s = 6 - 2 = 4$ neste exemplo. É importante notar que o cálculo de $x_{6\times4}$ depende de λ e do cálculo de $T_{6\times4}$ de V (veja Equação 4.2).

Implicitamente, este procedimento elimina o maior autovalor (i.e., $\lambda_1 = 1$) e seu respectivo autovetor (a primeira coluna de V) porque eles correspondem à solução trivial do sistema. O procedimento também elimina todos os componentes principais cujos ín-

dices são maiores que $n_s + 1$. Como resultado, neste exemplo, apenas as partes de λ e V apresentadas na Figura 4.2 são realmente usadas para calcular $T_{6\times4}$ e $x_{6\times4}$. Finalmente, a matriz de distância $d^2_{6\times6}$ é obtida de $x_{6\times4}$, $f_{c_{1\times6}}$ e λ, usando a Equação 4.3.

4.4 Aplicação em Bases de Dados Reais

Nesta seção é selecionada uma base de dados de múltiplas escolhas sobre métodos contraceptivos, para compreender como o mapeamento do *Dual Scaling* pode ajudar especialistas no processo de entendimento do comportamento da base de dados. A base é uma pesquisa nacional de prevalência de contraceptivos realizado na Indonésia no ano de 1987 [17]. A base de dados possui 1473 transações, com 37 itens, divididos em 10 questões, segue as informações sobre os atributos:

- **Idade da Esposa** - faixa etária da esposa: 15 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 45 a 49 anos;
- **Escolaridade da Esposa** - nível de escolaridade da esposa: sem formação / primeiro grau incompleto, primeiro grau completo, segundo grau completo, Superior / Mestrado / Doutorado;
- **Escolaridade do Marido** – nível de escolaridade do marido: sem formação / primeiro grau incompleto, primeiro grau completo, segundo grau completo, Superior / Mestrado / Doutorado;
- **Número de Filhos Nascidos** – faixa do número de filhos já nascidos do relacionamento: 0 a 2 filhos, 3 a 5 filhos, 6 a 8 filhos, 9 a 11 filhos, mais do que 11 filhos;
- **Religião da Esposa**: não islâmica, islâmica;
- **Esposa Empregada** – contém a informação se a esposa trabalha fora: empregada, desempregada ou não trabalha;
- **Ocupação do Marido** – classificação de ocupações feita pela Indonésia para dividir de forma macro os empregos: trabalho rural, produção, vendas e serviços, serviços administrativos e empregadores, professores e educadores;
- **Padrão de Vida** – padrão de vida da família: extremamente pobre, classe baixa, classe média, classe alta;

- **Acesso à Mídia** – acesso da família a mídias em geral, como rádio, jornais e televisão: muito acesso, pouco acesso;
- **Método Contraceptivo** – classificação do método contraceptivo utilizado pela família: não utiliza, uso contínuo, uso esporádico ou pontual.

Para esta base de estudo é aplicado o mapeamento do *Dual Scaling*. Neste exemplo temos $n_s = 37 - 10 = 27$ dimensões, algo impossível de ser visualmente analisado. Porém, ao utilizamos dados como distância e angulação entre pares de itens conseguimos identificar, sem conhecimento prévio das informações, algumas relações sobre o comportamento da base de dados.

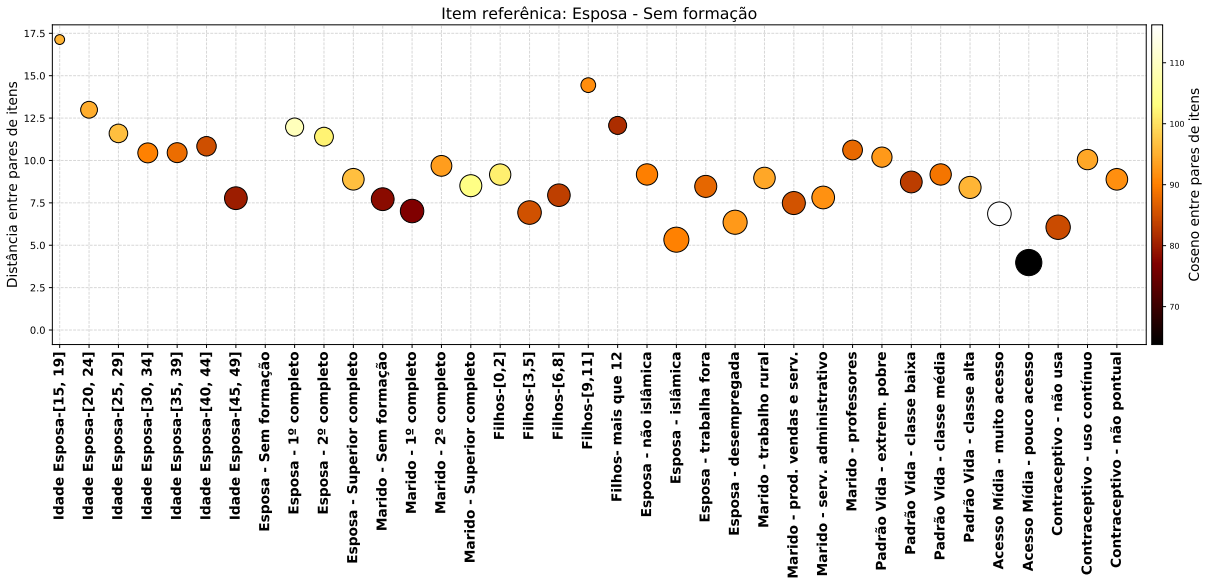


Figura 4.3: Resultado da distância e ângulo entre os atributos com o atributo de referência Esposa - Sem formação. As distâncias entre pares são representadas pelo eixo y e as cores do ponto reflete a angulação.

No primeiro exemplo, Figura 4.3, contém o item referência, chamado "esposa - sem formação". Quando observamos a relação de distância dos outros itens da base de dados nota-se que os itens mais próximos do item referência são: "esposa - islâmica", "esposa - desempregada", "mídia - pouco acesso"e "contraceptivo - não usa". No caso contrário, os atributos mais distantes são: "idade da esposa - [15-19]", "filhos - [9-11]"e "marido - professor". No entanto, olhar apenas para a distância pode trazer interpretações errôneas. Um exemplo, é o caso do atributo "mídia - muito acesso", nota-se que a angulação do mesmo é bem maior quando comparado com o atributo "mídia - pouco acesso". Como interpretação temos que a quantidade de transações que contém a relação "esposa - sem formação"e "mídia - muito acesso"é bem menor, quase inexistente, que a relação "esposa -

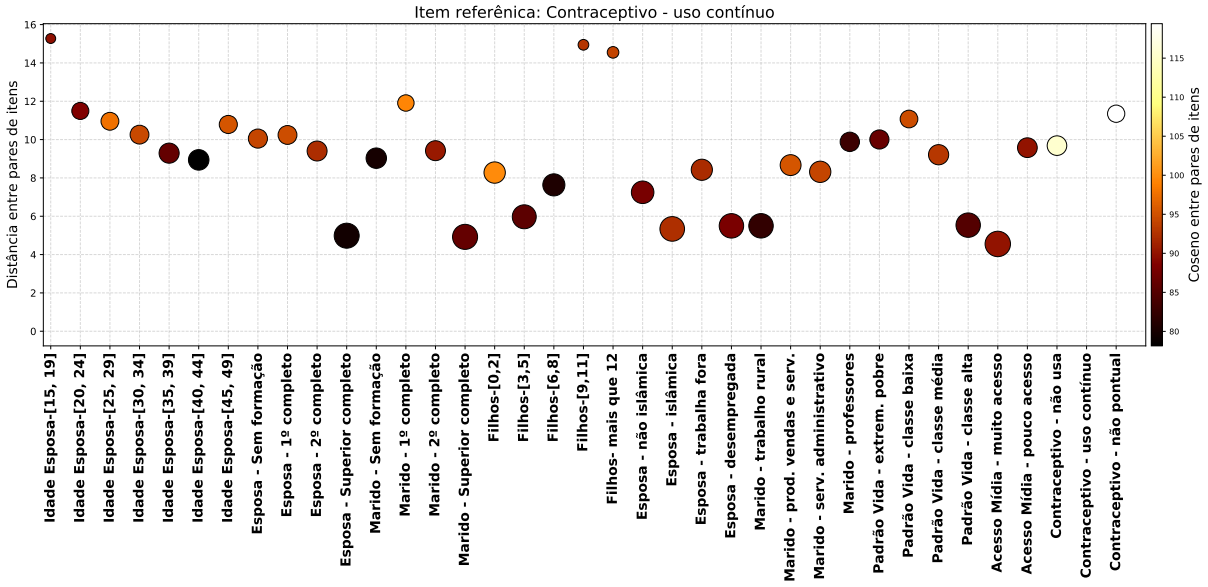


Figura 4.4: Resultado da distância e ângulo entre os atributos com o atributo de referência Contraceptivo - Uso contínuo. As distâncias entre pares são representadas pelo eixo y e as cores do ponto reflete a angulação.

sem formação"e "mídia - pouco acesso". No espaço de soluções "esposa - sem formação"e "mídia - muito acesso"estão em quadrantes opostos.

Na Figura 4.4 tem, como exemplo, o atributo "contraceptivo - uso contínuo". Da mesma forma observado o primeiro exemplo, os itens mais próximos são: "esposa - superior completo", "marido - superior completo"e "mídia - muito acesso". No caso contrário, nota-se os atributos mais distantes tais como: "idade esposa - [15-19]", "filhos - [9,11]"e "filhos - acima de 12". A ideia do *Dual Scaling* é gerar relações sobre o comportamento entre itens, dada a coocorrência apresentadas nas transações. Esse tipo de informação de proximidade nos fornece informações sobre o comportamento da base de dados.

4.5 Limitações

No procedimento original proposto por Nishisato, a matriz de respostas F deve respeitar as seguintes condições: (i) não ter itens de contexto, *i.e.,* itens com 100% de frequência na base de dados; (ii) não ter dados faltantes; e (iii) não ter categorias com tamanho muito divergente de itens. Um exemplo do caso (iii) seria ter uma base de dados que tem categorias tais com o sexo (masculino, feminino), com tamanho 2, e emprego (analista de sistemas, pedreiro, advogado, médico, motorista, taxista, jornalista, segurança), que tem tamanho 8. Na literatura, esse tipo de categoria é chamado suplementar ou passiva.

Vale ressaltar que Nishisato propôs o *Dual Scaling* como uma ferramenta para inspeção visual, e que as limitações citadas tem como preocupação as projeções gráficas de duas dimensões. Logo, não afeta no cálculo do mapeamento do espaço de soluções. Os algoritmos propostos na tese utilizam todas as dimensões, i.e., sem perdas de informações sobre o comportamento da base de dados. No entanto, vale ressaltar a discussão sobre possíveis distorções ao realizar o mapeamento dos itens para o espaço de soluções. Nishisato não menciona distorções no cenário de uso de todas as dimensões, mas possíveis distorções são encontradas em outras técnicas de análise multidimensional [19].

Nishisato não menciona limitações a respeito da não linearidade dos dados. No entanto, como uma técnica da família de análise multidimensional, supõe-se que tal limitação exista. Uma possível solução é aplicar o método *kernel trick* [29] para mapear os dados não lineares, e depois achar uma cobertura afim ou outro mapeamento linear e, logo após, nesses dados aplicar o mapeamento *Dual Scaling*.

Capítulo 5

Mineração de Itemsets Fechados: SCIM

Em linhas gerais, os algoritmos de mineração de itemsets possuem limiares usados no processo de busca de itemsets. Estes limiares, por exemplo, o suporte, têm dentre uma das finalidades reduzir o espaço de busca para a mineração de possíveis itemsets e, consequentemente, reduzir o custo computacional inerente ao processo. Outro aspecto importante do limiar de corte é reduzir a quantidade de itemsets gerados, pois dependendo da quantidade de itemsets retornados fica impraticável para um analista de dados fazer uma inspeção manual. Apesar da importância do limiar, a sua definição não é uma tarefa fácil, por exigir conhecimento aprofundado da base a ser estudada e do algoritmo utilizado. Para um especialista é normal ter que executar o algoritmo de mineração de itemsets várias vezes, com parâmetros de limiar diferentes, de modo a encontrar itemsets relevantes.

Este trabalho propõe o algoritmo SCIM(*spatial contextualization for closed itemset mining*) que utiliza técnica de análise multidimensional, como prova de conceito, neste estudo é usado o mapeamento do *Dual Scaling* na mineração de itemsets fechados. Para este fim, o algoritmo necessita de algumas etapas de pré-processamento. A Figura 5.1 ilustra as etapas do algoritmo e as seções desse capítulo descrevem cada uma destas etapas.

5.1 Visão Geral do Algoritmo Desenvolvido

As etapas são ilustradas na Figura 5.1, onde na primeira etapa, o usuário da técnica passa como parâmetro de entrada a base de dados $\mathcal{D}$ e a cobertura dos clusters dr. Logo após, a técnica constrói uma matriz F de padrão de resposta que representa a base de dados de entrada e, em seguida, é usado o *Dual Scaling* para mapear cada um dos m itens da base

de dados para um ponto $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n_s})$ (Equação 4.2) no espaço de soluções n_s-dimensional, onde $i \in \{1, 2, \cdots, m\}$ (Capítulo 4).

A próxima etapa consiste em aplicar um algoritmo de clusterização com sobreposição, desenvolvido neste trabalho, adotando uma forma conservadora onde cada item da base de dados terá um cluster associado, onde cada cluster tem seu raio de cobertura em proporção modulado pelo parâmetro dr. O cluster representa um conjunto de itens que tem grandes chances de gerar itemsets fechados relevantes e, consequentemente, itens que estão fora do cluster são pouco prováveis de compor itemsets fechados (Seção 5.2).

Na última etapa, é gerado os itemsets fechados utilizando os clusters como guia no processo de busca e assim reduzindo drasticamente a quantidade de caminhos possíveis a serem percorridos no processo de mineração de itemsets fechados. Além disso, os clusters também são usados para extrair os itemsets fechados, logo não é usado suporte ou qualquer outra métrica nesse processo de formação de itemset. Para a estrutura é utilizada a representação vertical FP-*tree* [28] para evitar a redundância de combinações no processo de mineração (Seção 5.3).

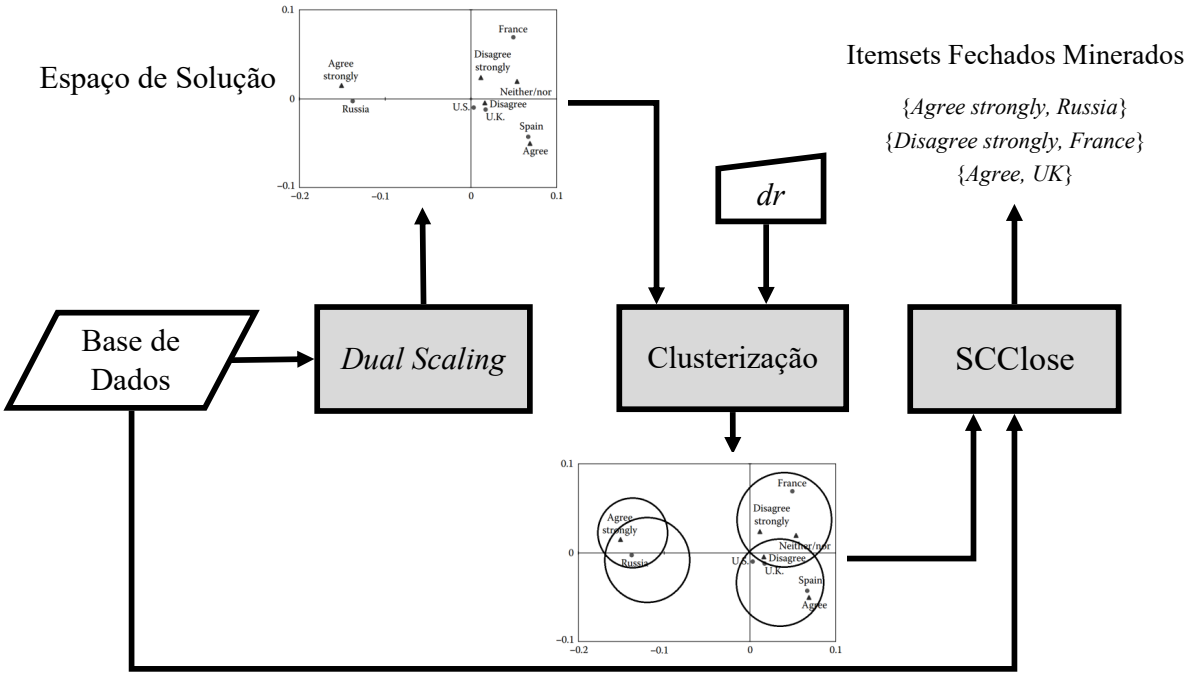


Figura 5.1: Algoritmo SCIM.

Vale ressaltar que o esforço deste trabalho, além de minerar itemsets fechados relevantes, está em definir um algoritmo que consiga reduzir o espaço de busca, inerente do processo combinatório, e também definir um limiar onde oferece um ponto de partida para o processo de busca de itemsets fechados. Diferente do parâmetro suporte, embora

seja intuitivo, o usuário não tem um ponto de partida para buscar os itemsets fechados.

5.2 Procedimento de Clusterização com Sobreposição

A distância $d_{i,i'}$ (Equação 4.3) entre os pontos de itens x_i e $x_{i'}$ é interpretada como os itens que se relacionam na base de dados. Assim, é natural esperar que, para um determinado ponto de item x_i, os pontos mais próximos são aqueles prováveis para a criação de itemsets fechados relevantes. Portanto, a definição de clusters para esse cenário é muito importante, visto que o cluster é usado para a redução do espaço de busca e na formação de itemsets fechados. Duas questões centrais devem ser abordadas para explorar essa propriedade: (i) como escolher o número de clusters; e (ii) como definir a cobertura de cada cluster.

Para o primeiro problema abordada, é adotado uma estratégia conservadora. Inicialmente o procedimento de clusterização assume que cada ponto x_i é o centro do cluster C_i. Para o segundo problema, deve-se considerar que cada base de dados pode conter conjuntos não-disjuntos de itens que podem ser combinados na formação de itemsets fechados. Por isso, o procedimento de clusterização deve permitir a sobreposição de clusters. Além disso, intuitivamente, a distância entre x_i e a origem do espaço de soluções define como o i-ésimo item se relaciona com os outros itens da base de dados. Sendo assim, para calcular a distância entre o ponto de item x_i para a origem, basta substituir o ponto de item $x_{i'}$ pela origem. Simplificando a Equação 4.3, temos:

$$ref_i^2 = \sum_{k=1}^{n_s} \sqrt{\lambda_{k+1}} \left(\frac{x_{i,k}}{\sqrt{p_i}} \right)^2, \tag{5.1}$$

onde λ_{k+1} é o $(k+1)$-ésimo autovalor de M (Equação 4.1), $x_{i,k}$ é calculado pela Equação 4.2, e $p_i = \frac{f_{c_i}}{n}$ é a proporção da frequência da i-ésima coluna, onde f_{c_i} é a frequência marginal de respostas para a i-ésima coluna de F.

De fato, Leroux e Rouanet [53] apontam uma propriedade interessante do mapeamento *Dual Scaling*, que mostra que a distância do item para a origem do espaço de soluções é inversamente proporcional à sua frequência na base de dados. Em outras palavras, os itens mais frequentes estão mais próximos da origem do que itens menos frequentes. A Tabela 5.1 contém os cálculos de $ref_{1\times 6}^2$. Para o exemplo, perceba que os itens menos frequentes e mais frequentes são, respectivamente, o mais distante e o ponto menos distante da origem. Na Tabela 5.1, a matriz $fc_{1\times 6}$ mostra a frequência de cada item na base de dados.

Item	d^2						ref^2	fc
	1	2	3	4	5	6		
1	0	16,6052	20,5072	3,2338	21,9519	18,4428	9,7467	2
2	16,6052	0	10,9090	14,0740	10,4710	1,7637	3,2046	6
3	20,5072	10,9090	0	17,5109	1,8114	9,7269	3,7234	7
4	3,2338	14,0740	17,5109	0	21,4269	17,5219	8,2790	4
5	21,9519	10,4710	1,8114	21,4269	0	10,8555	4,6087	6
6	18,4428	1,7637	9,7269	17,5219	10,8555	0	3,7871	5

Tabela 5.1: Matriz de distância entre itens para a base sintética $\mathcal{D}$. As últimas colunas incluem as informações da distância de referência ref_i^2 e a frequência fc_i de cada item da base. Para mais detalhes do processo de mapeamento do *Dual Scaling* veja a Figura 4.2.

A função de clusterização explora a propriedade que relaciona a distância com a origem e a frequência ao tentar garantir a inclusão de itens no cluster C_i que são tão importantes para x_i quanto x_i é importante para a base de dados. A ideia define uma cobertura mínima de C_i como a distância de referência ref_i^2 tomada da origem do espaço para x_i, isto é, um determinado item $x_{i'}$ será incluído no i-ésimo cluster se $d_{i,i'}^2 \leq ref_i^2$.

Nos experimentos observou-se a distância de referência ref_i^2 leva à formação de clusters que conseguem identificar itemsets fechados interessantes. No entanto, nota-se que essa cobertura pode não ser suficiente para recuperar itemsets fechados que também podem ser importantes para determinadas base de dados. A experiência durante o estudo dos casos mostra que esse problema surge quando a base de dados não inclui transações suficientes para definir um contexto com baixa incerteza para os itens da base de dados. Para superar esse problema é definido o parâmetro dr, ou seja, o limite máximo de razão de distância, para estender a cobertura dos clusters proporcionalmente à distância do item mais distante que pode ser considerado significativo a formação de itemsets fechados.

Conforme a literatura de *Dual Scaling* [46], a inspeção visual do espaço de solução mostra que cada eixo do espaço expressa uma característica particular (ou combinação de características) da população (e.g., grupos etários em alguma ordem particular). Portanto, para um dado ponto x_i, os únicos itens que podem ser considerados significativos para ele é aqueles cujos vetores de suporte definem ângulos na faixa $[-1, 1]$. Usando o parâmetro dr, a cobertura do cluster C_i é calculada como:

$$cvr_i = \left(\max_l (d_{i,l}) - ref_i\right) dr + ref_i, \tag{5.2}$$

onde ref_i é definido na Equação 5.1. Na Equação 5.2, $\max_l (d_{i,l})$ é a distância do item mais distante indexado por $l \in \{i' \mid a_{i,i'} \geq 0\}$, para $i, i' \in \{1, 2, \cdots, m\}$, sendo $a_{i,i'}$ o arco-

coseno (Equação 4.4). Fica evidente a partir da Equação 5.2 que $ref_i \leq cvr_i \leq \max_l (d_{i,l})$.

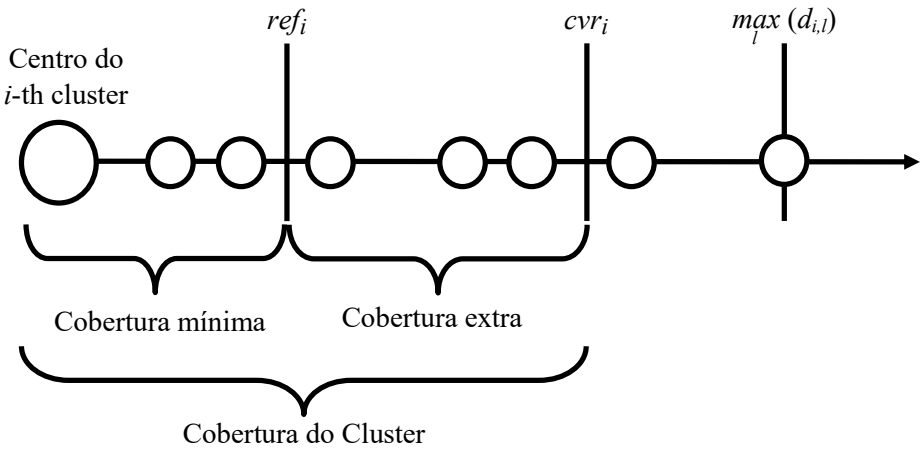


Figura 5.2: Estrutura do cluster. O círculo do lado esquerdo representa o ponto x_i, *i.e.,* o centro do cluster C_i. Os outros círculos representam pontos de itens x_l, para $l \in \{i' \mid x_i \cdot x_{i'} \geq 0\}$, ordenados da esquerda para a direita pela sua distância $d_{i,l}$ (Equação 4.3) para x_i. A linha horizontal representa o eixo das distâncias. O raio da cobertura mínima é a distância de referência ref_i (Equação 5.1), enquanto a cobertura total do cluster é dada por cvr_i (Equação 5.2).

A Figura 5.2 ilustra a estrutura de um cluster. O par de proposições a seguir resume os nossos critérios de clusterização:

Definição 14 (Centro do cluster) *Cada ponto de item x_i define o centro de um cluster C_i, para $i \in \{1, 2, \cdots, m\}$.*

Definição 15 (Cobertura do cluster) *O cluster C_i inclui todos os pontos de itens que satisfazem $d_{i,l} \leq cvr_i$, onde $d_{i,l}$ e cvr_i são calculados por, respectivamente, Equação 4.3, Equação 4.4 e Equação 5.2, para $l \in \{i' \mid a_{i,i'} \geq 0\}$ e $i, i' \in \{1, 2, \cdots, m\}$. A cobertura do cluster é parametrizada pelo limite definido pelo parâmetro $dr \in [0, 1]$.*

Para compreender melhor o procedimento de clusterização é apresentado um exemplo onde $dr = 0.00$ para o exemplo da Tabela 5.1, ou seja, assumindo a cobertura mínima como a cobertura $cvr_i = ref_i$ de C_i. Embora existam inicialmente $m = 6$ clusters neste exemplo, como mencionado antes, esses clusters podem se sobrepor e, em alguns casos, serem redundantes. No exemplo, o conjunto de clusters $\{C_1, C_4\}$ produzem clusters similares, da mesma forma que os conjuntos de clusters $\{C_2, C_6\}$ e $\{C_3, C_5\}$.

O Algoritmo 1 apresenta a função de clusterização, tendo como entrada o espaço de soluções e, definido pelo usuário, o parâmetro dr de cobertura dos clusters. Nas linhas 2

Algoritmo 1: Clusterização de itens no espaço de soluções

1 **Função** `SCCluster`(X, dr)
  Entrada: X espaço de soluções; dr parâmetro de cobertura do cluster
  Resultado: Cluster para cada item da base de dados
2 $D \leftarrow$ matriz de distância de pares de itens de X (Equação 4.3)
3 $A \leftarrow$ matriz de arcocoseno de pares de itens de X (Equação 4.4)
4 **para cada** $i \in \mathcal{I}$ **faça**
5 $\mathcal{C}_i \leftarrow \{\}$
6 Calcular cobertura do cluster corrente cvr_i usando D, A e dr (Equação 5.2)
7 **para cada** $i' \in \mathcal{I}$ **faça**
8 **se** $D_{i,i'} \leq cvr_i$ ***e*** $A_{i,i'} \geq 0$ **então**
9 $\mathcal{C}_i \leftarrow \mathcal{C}_i \cup \{i'\}$
10 **fim**
11 **fim**
12 **fim**
13 **retorna** $\mathcal{C}$
14 **fim**

e 3, respectivamente, é calculado a matriz de distância (Equação 4.3), e a matriz de arcocoseno entre os pares de itens dado o espaço de soluções X (Equação 5.2). Na linha 4 é feita uma interação em cada item da base de dados. Segundo a Definição 14, cada item tem seu cluster associado. Por isso, para cada item i da base de dados é definido o cluster $\mathcal{C}_i$ (linha 5). A cobertura total do cluster, cvr_i, é definida pela Equação 5.2 (linha 6), que calcula o raio do cluster corrente $\mathcal{C}_i$. A linha 7 percorre todos os itens da base, no intuito de identificar itens que pertence ao cluster corrente. Para ser considerado item pertencente ao cluster, o par de itens i e i' deve respeitar a condição de distância, onde a distância do par de itens não pode ser maior que a cobertura total do cluster corrente, e terem arcocoseno maior que 0 (linha 8). Desta forma, o procedimento de clusterização atualiza o cluster corrente com o item i' (linha 9).

O algoritmo SCIM utiliza apenas o dr como parâmetro de entrada da técnica. O dr tem como valor padrão 0, que tem como interpretação a cobertura mínima do cluster (Equação 5.2). O entendimento deste parâmetro, ou seja, valores maiores que $dr = 0$, está relacionado com o aumento da cobertura mínima dos clusters e consequentemente a agregação de mais itens para os clusters. Agregar mais itens no cluster significar reportar mais itemsets fechados nos clusters. No entanto, como já mencionado anteriormente, observou-se os itemsets fechados reportados nos clusters com cobertura mínima tem maiores valores de *all-confidence* e a partir que é acrescentado a cobertura extra nos clusters os valores de *all-confidence* vão diminuindo. Desta forma, podemos dizer que o parâmetro dr é menos subjetivo quando comparado ao parâmetro suporte.

De forma prática, digamos que existam duas bases de dados, onde a primeira base de dados T possui itemsets fechados com no máximo 70% de suporte, e na outra base de dados G possui itemsets fechados com no máximo 30% de suporte. Dado um especialista que não conheça as informações das duas bases de dados, temos dois cenários de uso de parâmetros: para o primeiro cenário, com o uso do parâmetro suporte como limiar de corte para seleção de itemset fechado, temos que para o usuário começar a identificar itemsets fechados tem que definir como mínimo suporte valores menores que 70% para base de dados T e no caso da base de dados G, o especialista tem que definir como mínimo suporte valores menores que 30%; no segundo cenário, com o uso do parâmetro dr, tanto para base de dados T ou G, o especialista pode começar com o valor $dr = 0$, desta forma independente da natureza dos dados na base, os clusters iniciais tendem a reportar itemsets fechados e caso o especialista da técnica queira aumentar a quantidade do mesmo, basta aumentando o valor do dr.

5.3 Geração de Itemsets Fechados

Dados os clusters gerados com o procedimento descrito na Seção 5.2, uma solução simplória para a geração de itemsets fechados seria a geração de todas as combinações, por força bruta, dos itens que estão dentro de cada cluster. Esta solução, no entanto, teria um custo computacional, agregado à necessidade de gerar todas as possíveis combinações e adicionalmente as verificações na base de dados. Outro aspecto seria o controle de itemset fechados repetidos, pois os clusters reportados podem conter sobreposições, consequentemente, mais de um cluster poder reportar o mesmo padrão de itemsets fechados.

De modo a resolver estes problemas, a etapa de geração de itemsets fechados do algoritmo SCIM (Figure 5.1) usa a representação vertical da base de dados fornecida pela estrutura de dados FP-*tree* para evitar repetidas varreduras da base de dados, e CFI-*tree* para reportar apenas uma instância de cada itemset fechado recuperado.

Esta tese apresenta a definição formal das estruturas de dados FP-*tree* e CFI-*tree* no Capítulo 3. Adicionalmente, detalhes e implementações podem ser encontrados na literatura para FP-*tree* [28] e CFI-*tree* [27]. Visto que o algoritmo não altera nenhum aspecto dessas estruturas, mas sim define melhores alternativas ao percorrer estas estruturas a fim de obter itemsets fechados relevantes, mesmo assim é necessário lembrar que as FP-*trees* são condicionais a um itemset $\mathcal{T}.base$ e têm um cabeçalho $\mathcal{T}.header$ com itens ordenados por suporte conjuntivo (Definição 3), em ordem decrescente, onde cada item é atribuído

Algoritmo 2: Busca na FP-*tree* para identificação de itemsets fechados

```
1  Procedimento SCClose(𝒯, ℋ, 𝒞)
       Entrada: 𝒯 a estrutura FP-tree que representa a base de dados; ℋ uma
                CFI-tree usada para validação de itemset fechado; 𝒞 um conjunto de
                clusters
       Resultado: Atualiza ℋ e reporta itemsets fechados
2      se 𝒯 possui um ramo B então
3          K ← ∅
4          para cada itemset fechado I ⊂ ℬ faça
5              para cada item i ∈ ℐ faça
6                  Q ← I ∩ C_i
7                  se 𝒯.base ∈ Q então
8                      K ← K ∪ {Q}
9                  fim
10             fim
11         fim
12         Ordernar K por cardinalidade de forma decrescente
13         para cada itemset I ∈ K faça
14             se I respeita a regra de formação de itemset então
15                 se I não é um itemset fechado em ℋ então
16                     Reportar itemset fechado I
17                     Inserir I em ℋ
18                 fim
19             fim
20         fim
21     senão
22         para cada item i ∈ 𝒯.header faça
23             I ← 𝒯.base ∪ {i}
24             se algum cluster em C contém I então
25                 𝒯_i ← A FP-tree condicional para i com base I
26                 SCClose(𝒯_i,ℋ, 𝒞)
27                 se I respeita a regra de formação de itemset então
28                     se I não é um itemset fechado em ℋ então
29                         Reportar itemset fechado I
30                         Inserir I em ℋ
31                     fim
32                 fim
33             fim
34         fim
35     fim
36 fim
```

a uma lista de links dos nós da estrura FP-*tree* que contêm esse item. Além disso, os nós da FP-*tree* possuem valores de ocorrência, usados para o cálculo da frequência do itemset.

O Algoritmo 2 apresenta o procedimento SCClose para geração de itemsets fechados com base na contextualização espacial fornecida pelo *Dual Scaling*. Na primeira chamada do procedimento, o algoritmo recursivo recebe como entrada uma FP-*tree* $\mathcal{T}$ construída sobre a base de dados $\mathcal{D}$ assumindo suporte conjuntivo maior do que zero, uma CFI-*tree* $\mathcal{H}$ vazia e o conjunto de clusters $\mathcal{C} = \{C_1, C_2, \cdots, C_m\}$ calculado conforme a Seção 5.2.

Nas chamadas recursivas (veja a linha 26), o primeiro argumento é uma FP-*tree* condicional construída para um dado item i com itemset base I e suporte conjuntivo maior que zero. Normalmente, os algoritmos que percorrem as FP-*trees* requerem a definição do limiar min_{sup}, pois as FP-*trees* são contraídas usando o respectivo limite inferior para o suporte conjuntivo. O procedimento SCClose, por outro lado, constrói as FP-*trees* independentemente do suporte. Isso porque sua estratégia de poda é baseada na seguinte regra de formação de itemsets:

Definição 16 (Regra de formação de itemset) *I é um itemset bem formado se e somente se houver pelo menos um cluster C_i com centro x_i onde seu respectivo item i satisfaz: (i) $i \in I$; (ii) cada item $j \in I$ é mapeado para um ponto x_j com $d_{i,j} \leq cvr_i$; e (iii) exista pelo menos um item $k \in I$ mapeado para um ponto x_k com $d_{i,k} \leq ref_i$ para $i \neq k$. A última condição é ignorada se a cobertura mínima de C_i inclui apenas x_i.*

A ideia principal por trás da Definição 16 é que só faz sentido minerar um itemset de um cluster se o mesmo possui o item representado pelo centro do cluster e pelo menos um item da região de cobertura mínima. Conforme os estudos realizados, as combinações exclusivas do item central com itens da região de cobertura extra (veja Figura 5.2) tendem a não produzirem itemsets interessantes, exceto nos casos onde a região de cobertura mínima inclui apenas o ponto central.

No Algoritmo 2, as linhas 3 a 20 lidam com o caso onde a FP-*tree* $\mathcal{T}$ contém apenas uma única ramificação B. Nesse caso, o laço de repetição na linha 4 itera sobre todos os candidatos a itemsets fechados de B. Por exemplo, seja $(i_1 : c_1, i_2 : c_2, \cdots, i_p : c_p)$ o caminho de B, onde p é o comprimento da ramificação, e $i_j : c_j$ denota o item i_j com a contagem c_j. A partir de $i_1 : c_1$, comparando as contagens de cada dois elementos adjacentes $i_j : c_j$ e $i_{j+1} : c_{j+1}$, é criado um candidato a itemset fechado frequente com a contagem c_j como $I = \{i_1, i_2, \cdots, i_j\} \cup \mathcal{T}.header$ sempre que $c_j \neq c_{j+1}$.

As linhas 5 a 10 geram para cada candidato a itemset fechado $I = \{i_1, i_2, \cdots, i_k\}$ o maior subconjunto incluído em cada cluster $C_{i_1}, C_{i_2}, \cdots, C_{i_k}$. Observe que o procedimento examina apenas os clusters representados pelos itens no itemset fechado candidato I.

Tabela 5.2: Todos os itemsets fechados I contidos na base sintética $\mathcal{D}$. Veja a Figura 4.2 para mais detalhes da base de dados.

I	**{3 , 5}**	**{2 , 6}**	**{1 , 4}**	{2 , 4}	{2 , 5}	{3 , 4}	{3 , 6}
$sup(I)$	**0,3333**	**0,2666**	**0,1333**	0,0666	0,0666	0,0666	0,0666
$allconf(I)$	**0,7142**	**0,6666**	**0,5000**	0,1666	0,1666	0,1428	0,1428

Isso porque a Definição 16 afirma que itemsets bem formados dado o cluster C_i devem incluir i. Além disso, apenas são selecionados os conjuntos de itemsets fechados que incluem o conjunto de itens de prefixo $\mathcal{T}.base$ (consulte a linha 7). O prefixo é importante por o algoritmo ser baseado em uma estrutura de árvore de prefixo. Após selecionar e inserir todos os itens fechados do candidato em K, são ordenados os elementos em K por cardinalidade do itemset em ordem decrescente (linha 12) para garantir que os supersets sejam inseridos primeiro na CFI-*tree* $\mathcal{H}$ (linha17).

Quando a FP-*tree* $\mathcal{T}$ contém mais de uma ramificação, cada iteração do laço de repetição definida pelas linhas 22 a 34 verifica se é possível construir um itemset I a partir do prefixo $\mathcal{T}.header$ e do item atual i, de modo que todos os itens em I sejam incluídos pelo menos em um dos clusters de C (linha 24). Se assim for, é construída a FP-*tree* condicional $\mathcal{T}_i$ para i com base I e é executada a chamada recursiva do procedimento SCClose passando $\mathcal{T}_i$, a CFI-*tree* $\mathcal{H}$ e C como argumentos (linha 26) . A função recursiva procura pelo superset do prefixo I. Após a chamada recursiva, reportamos I se o mesmo obedece à regra de formação de itemset (linha 27) e se não está incluído na CFI-*tree* $\mathcal{H}$ (linha 28). Neste laço de repetição (linha 13), os itens i são selecionados de forma que a estrutura FP-*tree* é percorrida de baixo para cima, ou seja, em ordem crescente de contagem de suporte conjuntivo.

É importante observar que, embora não seja utilizada a métrica de suporte para a remoção de itemsets no procedimento SCClose, a FP-*tree* preserva o suporte conjuntivo dos itemsets. Portanto, o suporte pode ser facilmente recuperado para a criação de regras de associação ou cálculos de outras métricas de mineração (e.g., *all-confidence*, *bond*, *lift*, etc.).

A primeira linha da Tabela 5.2 relata todos os itemsets fechados I que podem ser recuperados da base de dados $\mathcal{D}$ do exemplo da Figura 4.2. A segunda e terceira linhas mostram o suporte (Equação 2.1) e o *all-confidence* (Equação 2.3) de cada conjunto de itens. Os itemsets fechados recuperados quando $dr = 0{,}00$ são destacados em negrito. Eles são os mais significativos, ou seja, têm alta confiança.

5.4 Resultados

Nesta sessão é conduzido um conjunto de experimentos e análises no algoritmo proposto, SCIM. O algoritmo proposto foi desenvolvido em C++. Os experimentos foram executados em uma máquina com CPU I7 4.0GHz e 16GB de memória RAM rodando o sistema operacional Windows 8 64-bits.

Foi avaliado o algoritmo realizando experimentos e comparando a desempenho do SCIM com os algoritmos FPClose [27], Krimp [61], Slim [55] e TopPI [30]. Foi usado a implementação fornecida pelos autores[1,2] ou pela biblioteca de mineração de dados de código aberto SPMF[3]. A implementação do algoritmo SCIM se encontra disponível publicamente[4].

O algoritmo FPClose foi escolhido para se ter o conhecimento geral do comportamento dos possíveis itemsets fechados reportados para uma determinada base de dados. O algoritmo TopPI é considerado estado da arte em top-k algoritmos. Onde a mineração é centrada nos top-k de cada item da base de dados, seguindo a mesma lógica do SCIM onde é gerado um cluster para cada item da base de dados. O algoritmo Slim é o estado da arte nos algoritmos de utilizam a ideia que o conjunto de itemsets fechados reportados podem ser usados para comprimir a base de dados original, e quanto melhor o conjunto de dados reportados melhor é a compressão da base de dados. Neste sentido, seria interessante comparar o comportamento de seleção do SCIM, visto que a técnica propõe reportar itemses relevantes.

5.4.1 Definição das Métricas Utilizadas

Inicialmente é importante estabelecer algumas métricas que permitam a comparação quantitativa e qualitativa dos resultados da execução de determinado algoritmo. Neste trabalho foram adotadas três métricas: *all-confidence* (Seção 5.4.1.1), *cross-support* (Seção 5.4.1.2) e tempo de execução(Seção 5.4.1.3).

[1]Slim: `https://people.mmci.uni-saarland.de/~jilles/prj/slim`
[2]TopPI: `https://github.com/slide-lig/TopPI`
[3]FPClose: https://www.philippe-fournier-viger.com/spmf
[4]SCIM: `https://github.com/Prograf-UFF/SCIM`

5.4.1.1 Primeira Métrica: *All-confidence* dos Itemsets Fechados Selecionados

O objetivo do uso desta métrica é entender o comportamento de geração de itemsets fechados. A métrica *all-confidence* é usada para qualificar se este conjunto contém itemsets fechados relevantes. Esta medida tem variação de [0,1], onde valores próximos de 1 indicam que os itemsets fechados geram regras de associações com altos valores de confiança, enquanto valores próximos de 0 indicam que os itemsets fechados geram regras de associação com baixo valores de confiança.

Nos experimentos são comparados as distribuições de medidas de *all-confidence* médio dos itemsets fechados recuperados pelo FPClose, Krimp, Slim, TopPI e a abordagem proposta. Para definir a média de *all-confidence*, primeiro são criados sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte dado uma determinada base de dados. Em seguida, calculamos a média de *all-confidence* a partir dos itemsets fechados recuperados em cada partição. Ao fazê-lo, é possível inspecionar a distribuição dos valores médios de *all-confidence* dos itemsets fechados em todas as frequências de suporte e assim verificar a capacidade de cada técnica de recuperar itemsets fechados em cada partição de suporte.

5.4.1.2 Segunda Métrica: *Cross-support* dos Itemsets Fechados Selecionados

O objetivo do uso desta métrica é entender o comportamento de seleção de itemsets fechados. A métrica *cross-support* é usada para qualificar se este conjunto contém itemsets fechados relevantes. Esta medida tem variação de [0,1], onde valores próximos de 1 indicam que os itens que compõem o itemset fechado são correlacionados, enquanto valores próximos de 0 indicam que os itens que compõem o itemset fechado são pouco correlacionados.

Da mesma forma que a métrica *all-confidence*, nos experimentos são comparados as distribuições de medidas de *cross-support* médio dos itemsets fechados recuperados pelo FPClose, Krimp, Slim, TopPI e a abordagem proposta. Para definir a média de *cross-support*, primeiro são criados sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte dado uma determinada base de dados. Em seguida, calculamos a média de *cross-support* a partir dos itemsets fechados recuperados em cada partição. Ao fazê-lo, é possível inspecionar a distribuição dos valores médios de *cross-support* dos itemsets fechados em todas as frequências de suporte e assim verificar a capacidade de cada técnica de recuperar itemsets fechados em cada partição de suporte.

5.4.1.3 Terceira Métrica: Tempo de Execução

Está métrica tem como foco identificar o quão custoso é determinado algoritmo dado as variações de parâmetros definidas durante o teste. Os valores dos tempos são definidos em segundos. Foi utilizada a implementação fornecida pelos autores de cada técnica comparada. Nos experimentos foi contabilizado todo o processo dos algoritmos, deste a leitura da base de dados até o resultado dos itemsets fechados selecionados.

5.4.2 Experimentos

No decorrer dos experimentos foi realizada uma análise detalhada da escolha dos parâmetros para a comparação não tendenciosa dos algoritmos. Foi definido um conjunto de variações de parâmetros para cada base de dados, e posteriormente, este estudo contém as métricas calculadas para o FPClose, Krimp, Slim, TopPI e SCIM em todas as bases de dados usadas nos experimentos. No Apêndice A é apresentado este estudo de forma detalhada onde é possível observar o comportamento de seleção de itemsets dado um parâmetro de interesse com os restantes dos parâmetros definidos pelo estudo.

O algoritmo proposto foi testado em onze bases de dados públicas de LUCS-KDD [14]. Das bases de dados disponíveis, foram selecionamos aquelas que não têm dados faltantes, ou seja, alguma transação aonde alguma questão não tem ocorrência de resposta. O número de transações (n), items (m) e questões (q) em cada base de dados estão descritas na Tabela 5.3. Foi retirado o item *Class* de cada base de dados porque os experimentos não estão relacionados com procedimentos de classificação.

Para uma comparação justa no quesito tempo de processamento, os algoritmos sequenciais SCIM e FPClose foram comparados com implementações em *multithreading* TopPI e Slim usando apenas uma thread. Vale ressaltar que o Slim foi configurado para encontrar apenas itemsets fechados. Na Tabela 5.3, o símbolo ✓ indica se foi possível a execução do algoritmo na base de dados em questão.

FPClose falhou ao processar as bases de dados *Letter recognition*, *mFeat*, *Pen digits*, *Waveform* e *Connect-4* dado que o mesmo esgotou a memória disponível do sistema. A execução da implementação do Krimp foi abortada para as bases de dados *mFeat* e *Connect-4* após passar mais de um dia em execução, mesmo no modo multithreading.

É importante enfatizar que no restante desta seção não se discute os resultados alcançados pelo Krimp, visto que nossa análise mostra que ele é superado pelo seu concorrente

Tabela 5.3: Base de dados usada em nossos experimentos.

Base de dados	n	m	q	FPClose	Krimp	Slim	TopPI	SCIM
Letter recognition	20,000	80	16		✓	✓	✓	✓
mFeat	2,000	1,648	240			✓	✓	✓
Wine	178	65	13	✓	✓	✓	✓	✓
Page blocks	5,473	41	10	✓	✓	✓	✓	✓
Pen digits	10,992	79	16		✓	✓	✓	✓
Waveform	5,000	98	21		✓	✓	✓	✓
Ecoli	336	26	7	✓	✓	✓	✓	✓
Connect-4	67,557	126	42			✓	✓	✓
Tic-tac-toe	958	27	9	✓	✓	✓	✓	✓
Led7	3,200	14	7	✓	✓	✓	✓	✓
Pima	768	36	8	✓	✓	✓	✓	✓

mais próximo, o algoritmo Slim.

Como o algoritmo SCIM pode recuperar itemsets fechados com qualquer valor de suporte, configuramos $min_{csup} = 1$ para FPClose, TopPI e Slim. Os parâmetros variados nestes experimentos foram k e dr, respectivamente, para os algoritmos TopPI e SCIM.

Para escolher os melhores valores de parâmetro para os algoritmos TopPI e SCIM em cada base de dados, foi definido um gráfico onde o eixo horizontal corresponde a valores médios de *all-confidence* variando de 0 a 1, enquanto o eixo vertical distribui os valores de suporte de 0 até o limite superior de cada base de dados. Espera-se que quanto melhor for o conjunto de itemsets fechados recuperados, mais à direita será a curva que representa o desempenho de uma técnica/parametrização.

5.4.3 Discussão Sobre Qualidade e Tempo de Execução

Dado uma base de dados, o algoritmo SCIM tem como proposta recuperar itemsets fechados que sejam interessantes. Itemsets fechados com valores altos de *all-confidence* (Equação 2.3) são interessantes porque geram regras de associação com altos valores de confiança. A Tabela 5.4 apresenta resultados para todas as bases de dados usadas em nossos experimentos. A primeira coluna contém a informação da base de estudo com as informações de parâmetros escolhidos para cada algoritmo do estudo. Já na segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. As colunas três a nove mostram o número de conjuntos de itens fechados recuperados (#), os valores médios da métrica corrente (μ) e os valores de desvio padrão (σ) calculados por intervalo de suporte por todas as técnicas comparadas. Esses intervalos de suporte têm o mesmo tamanho por base de dados e estes tamanhos foram definidos sobre o intervalo de suporte da base de dados. Na Figura 5.3, os eixos hori-

zontais correspondem aos valores médios de *all-confidence* ou *cross-support* variando de 0 a 1, enquanto os eixos verticais distribuem os valores de suporte de 0 ao limite superior de cada base de dados. Espera-se que quanto melhor o conjunto de itemsets fechados recuperados, mais à direita a curva representando o desempenho de uma técnica será. As duas últimas colunas da Tabela 5.4 apresentam o número total de itemsets fechados detectados e os tempos de processamento (em segundos). A técnica vencedora possui a maior quantidade de faixas de suporte com melhores valores médios de *all-confidence*. Essa mesma interpretação pode ser usado para a métrica *cross-support*.

FPClose e Slim não possuem parâmetros adicionais para serem definidos visto que min_{csup} seja escolhido. O TopPI também requer o número $k \in \{1, 2, \cdots, (m-1)!\}$ de primeiros itemsets fechados esperados para cada item. O SCIM não depende de min_{csup}, mas requer a definição de $dr \in [0, 1]$. São apresentados nas tabelas e gráficos dessa seção os resultados para os valores k e dr que produzem os melhores valores de *all-confidence* para cada base de dados. Observe na Tabela 5.4 como a parametrização do SCIM é estável: os melhores resultados foram obtidos com $dr \leq 0{,}04$ em 9 das 11 bases de dados (4 casos com $dr = 0{,}00$, 3 casos com $dr = 0{,}03$ e 1 caso com dr com valores 0,02, 0,04, 0,10 ou 0,37). Em contraste, valores diferentes de k variando de 1 a 30 são as melhores escolhas para o TopPI em praticamente todas as bases de dados.

A Figura 5.3 apresenta distribuições de valores médios (μ) de *all-confidence* calculados para a base de dados do estudo. Dentre elas temos cinco bases de dados onde o SCIM supera outros algoritmos. Nesses exemplos, o SCIM tem valores médios mais altos *all-confidence* em quase todas as partições de suporte. As bases de dados *Page blocks*, *Led7*, *Letter recognition* e *Wine* (Figuras 5.3a, 5.3e, 5.3g e 5.3a) promoveram melhores separações entre μ calculados para itemsets fechados encontrados pelos algoritmos comparados e os itemsets fechados encontrados por nossa técnica. Já as bases de dados *Connect-4* e *Waveform* (Figuras 5.3c e 5.3c) apresentaram empate, pois tiveram alternâncias entre os melhores valores de média de *all-confidence*.

Os únicos casos onde os valores μ e outra técnica são superiores aos valores calculados para SCIM são as bases de dados *mFeat* e *Ecoli* (Figura 5.3). Na base de dados *mFeat*, o SCIM supera o TopPI. O desempenho da abordagem proposta é semelhante ao FPClose e TopPI na base de dados *Ecoli*. Em ambos os casos, Slim apresenta melhor desempenho na faixa de suporte [0,1 , 0,6).

Tabela 5.4: Distribuição de valores médios de *all-confidence* e de *cross-support* dos itemsets fechados recuperados por FPClose, Slim, TopPI e SCIM sobre as bases de dados de múltiplas escolhas da Tabela 5.3.

Connect	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	1.499	0,009	0,030	4	0,343	0,168	3	0,362	0,053	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1.506	88,79
	cross-support		0,097	0,099		0,624	0,253		0,382	0,054		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
TopPI $k = 8$	all-confidence	496	0,038	0,039	123	0,221	0,050	55	0,335	0,042	35	0,486	0,047	40	0,651	0,044	44	0,801	0,029	172	0,951	0,040	965	2,87
	cross-support		0,038	0,039		0,221	0,051		0,336	0,043		0,488	0,047		0,654	0,044		0,804	0,029		0,953	0,040		
SCIM $dr = 0{,}00$	all-confidence	273	0,053	0,051	577	0,223	0,052	67	0,302	0,024	54	0,472	0,027	7	0,707	0,060	6	0,803	0,032	18	0,949	0,037	1.002	9,98
	cross-support		0,062	0,064		0,263	0,053		0,307	0,023		0,480	0,025		0,707	0,060		0,804	0,032		0,949	0,037		

Ecoli	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,55]			(0,55 , 0,69]			(0,69 , 0,83]			(0,83 , 0,97]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	436	0,029	0,030	35	0,194	0,041	21	0,359	0,029	13	0,457	0,039	22	0,634	0,081	1	0,833	0,000	2	0,913	0,084	530	0,14
	cross-support		0,102	0,084		0,280	0,032		0,552	0,105		0,604	0,009		0,674	0,081		0,857	0,000		0,915	0,082		
Slim	all-confidence	17	0,051	0,045	2	0,498	0,493	1	0,346	0,000	2	0,641	0,155	2	0,746	0,266	1	0,833	0,000	0	0,000	0,000	25	0,17
	cross-support		0,162	0,220		0,591	0,443		0,594	0,000		0,782	0,266		0,778	0,234		0,857	0,000		0,000	0,000		
TopPI $k = 3$	all-confidence	37	0,041	0,033	7	0,230	0,054	3	0,298	0,015	1	0,546	0,000	9	0,654	0,113	1	0,833	0,000	2	0,913	0,084	60	0,25
	cross-support		0,049	0,038		0,246	0,056		0,300	0,017		0,594	0,000		0,665	0,110		0,857	0,000		0,915	0,082		
SCIM $dr = 0{,}02$	all-confidence	34	0,040	0,035	10	0,221	0,044	2	0,306	0,006	3	0,480	0,059	12	0,643	0,098	1	0,833	0,000	2	0,913	0,084	64	0,03
	cross-support		0,063	0,051		0,256	0,048		0,310	0,000		0,594	0,000		0,680	0,092		0,857	0,000		0,915	0,082		

Led7	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,09]			(0,09 , 0,17]			(0,17 , 0,26]			(0,26 , 0,34]			(0,34 , 0,43]			(0,43 , 0,51]			(0,51 , 0,60]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		

Continua na próxima página.

FPClose	all-confidence	1.445	0,032	0,032	313	0,163	0,045	100	0,285	0,062	42	0,390	0,055	18	0,486	0,042	9	0,626	0,060	9	0,681	0,056	1.936	0,20
	cross-support		0,347	0,105		0,465	0,143		0,576	0,161		0,670	0,155		0,746	0,094		0,861	0,070		0,838	0,100		
Slim	all-confidence	73	0,028	0,032	4	0,137	0,009	1	0,247	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	78	0,17
	cross-support		0,348	0,096		0,333	0,094		0,512	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
TopPI $k = 7$	all-confidence	0	0,000	0,000	6	0,188	0,020	18	0,321	0,074	16	0,394	0,056	9	0,509	0,030	9	0,626	0,060	9	0,681	0,056	67	0,30
	cross-support		0,000	0,000		0,245	0,016		0,444	0,144		0,534	0,102		0,723	0,117		0,861	0,070		0,838	0,100		
SCIM $dr = 0{,}10$	all-confidence	0	0,000	0,000	1	0,278	0,000	3	0,410	0,097	3	0,444	0,050	3	0,516	0,032	2	0,680	0,113	3	0,674	0,089	15	0,04
	cross-support		0,000	0,000		0,449	0,000		0,551	0,171		0,544	0,072		0,598	0,042		0,938	0,079		0,759	0,108		

		Partição de suporte																						
Letter recognition	***Métrica***	[0,00 , 0,09]			(0,09 , 0,18]			(0,18 , 0,28]			(0,28 , 0,37]			(0,37 , 0,46]			(0,46 , 0,55]			(0,55 , 0,64]			Itemset #	Tempo (s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	1.210	0,014	0,032	11	0,199	0,074	4	0,339	0,050	0	0,000	0,000	4	0,615	0,083	0	0,000	0,000	2	0,760	0,059	1.231	34,31
	cross-support		0,149	0,144		0,425	0,184		0,729	0,200		0,000	0,000		0,833	0,083		0,000	0,000		0,852	0,070		
TopPI $k = 7$	all-confidence	260	0,077	0,061	87	0,162	0,046	19	0,304	0,049	15	0,424	0,070	35	0,574	0,078	23	0,644	0,060	8	0,747	0,035	447	0,62
	cross-support		0,100	0,083		0,212	0,070		0,397	0,089		0,558	0,144		0,738	0,136		0,812	0,064		0,855	0,094		
SCIM $dr = 0{,}03$	all-confidence	29	0,192	0,197	4	0,235	0,183	12	0,333	0,051	21	0,451	0,068	16	0,571	0,085	2	0,762	0,144	5	0,735	0,040	89	0,48
	cross-support		0,282	0,296		0,341	0,385		0,510	0,126		0,599	0,094		0,681	0,133		0,831	0,149		0,798	0,064		

		Partição de suporte																						
mfeat	***Métrica***	[0,00 , 0,11]			(0,11 , 0,23]			(0,23 , 0,34]			(0,34 , 0,46]			(0,46 , 0,57]			(0,57 , 0,69]			(0,69 , 0,80]			Itemset #	Tempo (s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	4.438	0,057	0,079	242	0,404	0,171	194	0,584	0,145	131	0,703	0,123	75	0,773	0,098	33	0,862	0,050	8	0,866	0,040	5.121	10.053,94
	cross-support		0,252	0,276		0,621	0,195		0,730	0,172		0,818	0,127		0,850	0,124		0,939	0,056		0,904	0,062		
TopPI $k = 3$	all-confidence	2.631	0,052	0,045	131	0,329	0,182	237	0,427	0,136	244	0,551	0,125	190	0,675	0,111	105	0,772	0,101	29	0,835	0,063	3.567	1,35
	cross-support		0,055	0,045		0,348	0,189		0,449	0,143		0,583	0,135		0,712	0,120		0,825	0,110		0,885	0,066		
SCIM $dr = 0{,}00$	all-confidence	449	0,296	0,142	4.560	0,394	0,104	3.903	0,492	0,109	2.150	0,606	0,093	761	0,737	0,064	105	0,856	0,050	15	0,892	0,034	11.943	3.351,34
	cross-support		0,446	0,213		0,550	0,163		0,630	0,152		0,732	0,120		0,870	0,083		0,931	0,064		0,938	0,054		

Continua na próxima página.

Page blocks	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,29]			(0,29 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,86]			(0,86 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	179	0,003	0,004	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	535	0,955	0,024	714	0,20
	cross-support		0,004	0,005		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,970	0,014		
Slim	all-confidence	30	0,029	0,133	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	10	0,955	0,036	40	0,20
	cross-support		0,031	0,133		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,975	0,019		
TopPI $k = 1$	all-confidence	29	0,004	0,005	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2	0,997	0,002	31	0,29
	cross-support		0,004	0,005		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,997	0,002		
SCIM $dr = 0{,}00$	all-confidence	32	0,039	0,139	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	22	0,988	0,012	54	0,13
	cross-support		0,047	0,149		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,991	0,009		

Pen digits	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,04]			(0,04 , 0,08]			(0,08 , 0,13]			(0,13 , 0,17]			(0,17 , 0,21]			(0,21 , 0,25]			(0,25 , 0,29]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	1.125	0,016	0,023	52	0,175	0,055	22	0,265	0,067	10	0,434	0,092	7	0,487	0,130	3	0,570	0,078	1	0,587	0,000	1.220	45,24
	cross-support		0,349	0,186		0,582	0,182		0,629	0,182		0,697	0,121		0,718	0,155		0,667	0,053		0,729	0,000		
TopPI $k = 7$	all-confidence	95	0,040	0,027	46	0,151	0,043	84	0,275	0,069	98	0,366	0,086	59	0,431	0,065	14	0,521	0,050	5	0,549	0,027	401	0,55
	cross-support		0,082	0,059		0,265	0,084		0,527	0,148		0,611	0,153		0,718	0,138		0,818	0,132		0,776	0,129		
SCIM $dr = 0{,}04$	all-confidence	2	0,106	0,031	15	0,175	0,044	57	0,263	0,067	46	0,397	0,100	21	0,469	0,082	5	0,558	0,062	2	0,563	0,034	148	0,37
	cross-support		0,417	0,404		0,447	0,135		0,501	0,129		0,572	0,156		0,648	0,136		0,662	0,038		0,651	0,111		

pima	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,13]			(0,13 , 0,26]			(0,26 , 0,40]			(0,40 , 0,53]			(0,53 , 0,66]			(0,66 , 0,79]			(0,79 , 0,93]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	1.345	0,027	0,029	16	0,153	0,011	0	0,000	0,000	8	0,526	0,017	112	0,626	0,041	96	0,757	0,047	31	0,884	0,034	1.608	0,17
	cross-support		0,050	0,043		0,184	0,006		0,000	0,000		0,791	0,015		0,805	0,039		0,888	0,053		0,935	0,030		

Continua na próxima página.

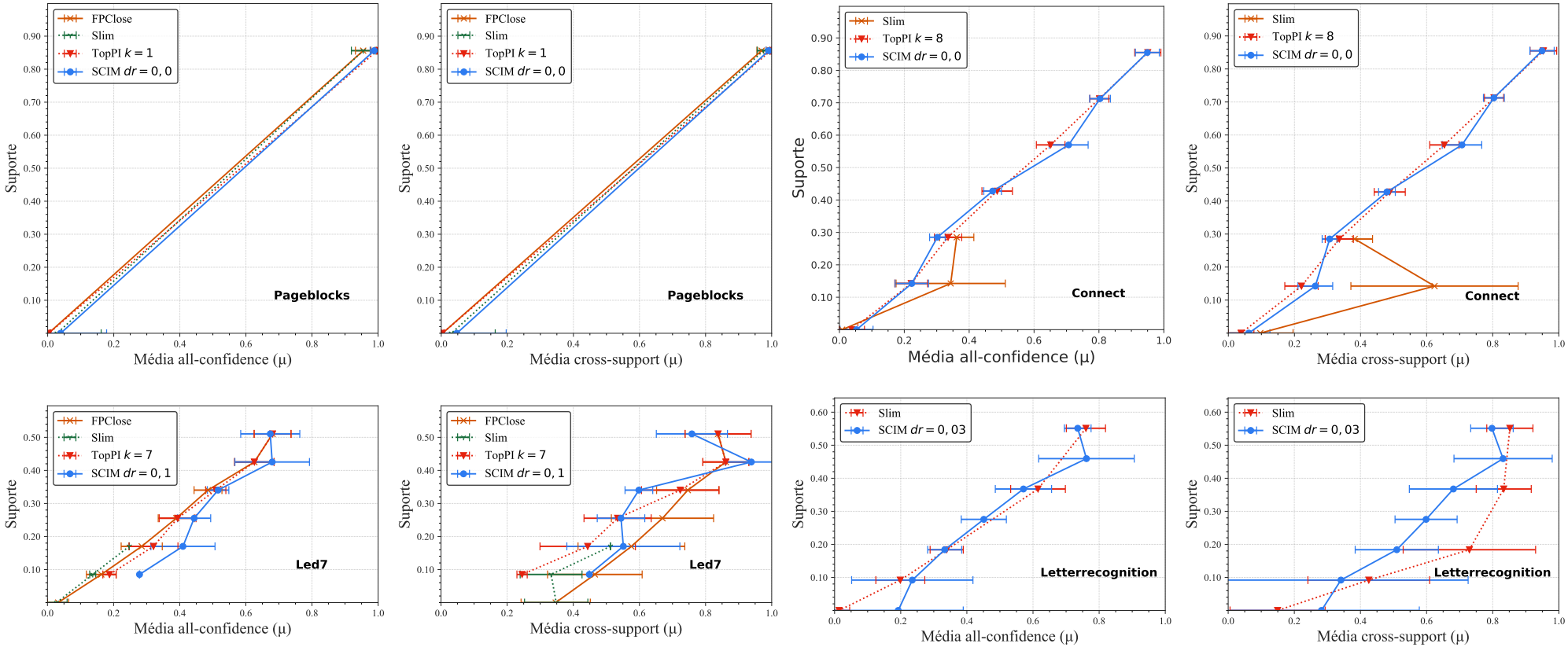
Slim	all-confidence	47	0,029	0,029	0	0,000	0,000	0	0,000	0,000	1	0,511	0,000	3	0,618	0,051	2	0,758	0,042	2	0,911	0,015	55	0,21
	cross-support		0,066	0,131		0,000	0,000		0,000	0,000		0,783	0,000		0,798	0,026		0,864	0,114		0,956	0,043		
TopPI $k = 30$	all-confidence	593	0,026	0,024	16	0,153	0,011	0	0,000	0,000	0	0,000	0,000	8	0,683	0,004	69	0,773	0,044	31	0,884	0,034	717	0,31
	cross-support		0,035	0,029		0,184	0,006		0,000	0,000		0,000	0,000		0,787	0,004		0,880	0,060		0,935	0,030		
SCIM $dr = 0{,}03$	all-confidence	302	0,029	0,021	4	0,154	0,012	0	0,000	0,000	8	0,526	0,017	90	0,622	0,041	87	0,759	0,048	31	0,884	0,034	522	0,06
	cross-support		0,040	0,026		0,182	0,000		0,000	0,000		0,791	0,015		0,806	0,043		0,896	0,048		0,935	0,030		

Tic-tac-toe	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,03]			(0,03 , 0,06]			(0,06 , 0,08]			(0,08 , 0,11]			(0,11 , 0,14]			(0,14 , 0,17]			(0,17 , 0,20]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	40.316	0,016	0,013	1.698	0,090	0,024	396	0,166	0,027	138	0,239	0,041	44	0,280	0,031	32	0,369	0,018	60	0,415	0,031	42.684	2,98
	cross-support		0,580	0,130		0,682	0,142		0,765	0,125		0,759	0,141		0,733	0,195		0,860	0,069		0,846	0,074		
Slim	all-confidence	68	0,033	0,014	20	0,104	0,036	16	0,185	0,030	7	0,258	0,021	6	0,267	0,022	0	0,000	0,000	8	0,442	0,031	125	0,23
	cross-support		0,551	0,146		0,745	0,108		0,712	0,183		0,693	0,164		0,726	0,203		0,000	0,000		0,821	0,058		
TopPI $k = 15$	all-confidence	0	0,000	0,000	0	0,000	0,000	46	0,204	0,022	76	0,267	0,028	36	0,284	0,032	32	0,369	0,018	60	0,415	0,031	250	0,31
	cross-support		0,000	0,000		0,000	0,000		0,549	0,083		0,735	0,177		0,695	0,197		0,860	0,069		0,846	0,074		
SCIM $dr = 0{,}37$	all-confidence	142	0,034	0,022	40	0,091	0,018	54	0,174	0,022	48	0,206	0,022	6	0,284	0,012	2	0,364	0,000	38	0,429	0,029	330	0,06
	cross-support		0,709	0,112		0,660	0,142		0,817	0,057		0,692	0,128		0,581	0,027		0,801	0,000		0,835	0,065		

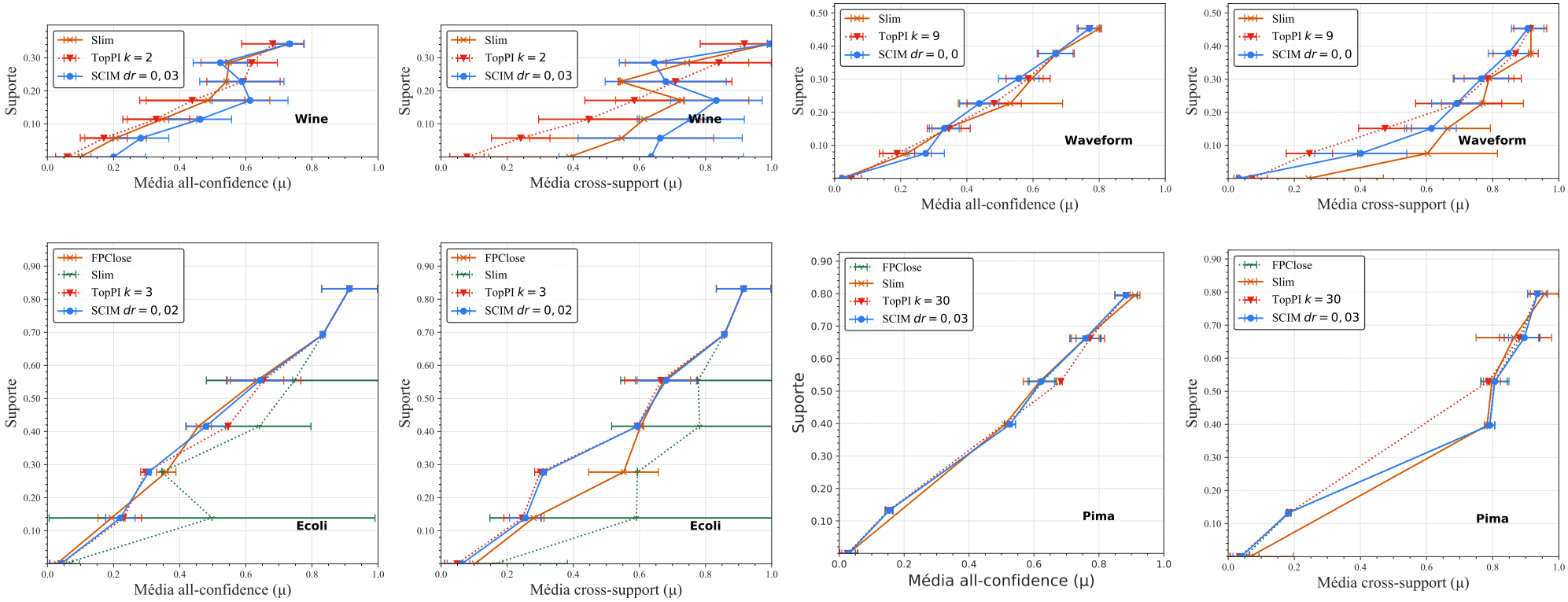
Waveform	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,08]			(0,08 , 0,15]			(0,15 , 0,23]			(0,23 , 0,30]			(0,30 , 0,38]			(0,38 , 0,45]			(0,45 , 0,53]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	675	0,024	0,032	23	0,219	0,073	10	0,339	0,044	3	0,532	0,157	4	0,597	0,034	1	0,671	0,000	1	0,800	0,000	717	8,50
	cross-support		0,243	0,226		0,603	0,211		0,663	0,130		0,769	0,124		0,789	0,075		0,916	0,000		0,916	0,000		
TopPI $k = 9$	all-confidence	413	0,049	0,032	54	0,189	0,053	74	0,345	0,065	52	0,483	0,082	70	0,585	0,067	57	0,669	0,052	14	0,771	0,034	734	0,44
	cross-support		0,072	0,047		0,246	0,070		0,474	0,080		0,697	0,130		0,784	0,103		0,869	0,068		0,914	0,050		
SCIM $dr = 0{,}00$	all-confidence	1	0,023	0,000	13	0,276	0,056	289	0,332	0,045	255	0,437	0,059	110	0,557	0,061	44	0,669	0,056	12	0,770	0,037	724	0,39
	cross-support		0,032	0,000		0,401	0,139		0,615	0,075		0,691	0,076		0,766	0,082		0,848	0,061		0,906	0,049		

Continua na próxima página.

Wine	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,06]			(0,06 , 0,11]			(0,11 , 0,17]			(0,17 , 0,23]			(0,23 , 0,28]			(0,28 , 0,34]			(0,34 , 0,40]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	10.944	0,052	0,031	1.680	0,161	0,052	371	0,280	0,077	120	0,399	0,096	36	0,495	0,077	13	0,603	0,061	5	0,633	0,094	13.169	0,42
	cross-support		0,361	0,161		0,508	0,160		0,614	0,160		0,723	0,155		0,792	0,137		0,851	0,134		0,875	0,111		
Slim	all-confidence	18	0,100	0,032	14	0,206	0,093	10	0,347	0,105	7	0,486	0,187	1	0,541	0,000	3	0,550	0,085	2	0,732	0,044	55	0,23
	cross-support		0,388	0,244		0,546	0,278		0,614	0,170		0,729	0,201		0,541	0,000		0,744	0,187		0,995	0,008		
TopPI $k = 2$	all-confidence	25	0,060	0,035	12	0,171	0,071	10	0,329	0,101	7	0,438	0,159	5	0,593	0,111	6	0,617	0,078	3	0,681	0,094	68	0,25
	cross-support		0,078	0,052		0,241	0,088		0,447	0,152		0,585	0,149		0,709	0,171		0,840	0,159		0,917	0,134		
SCIM $dr = 0{,}03$	all-confidence	6	0,200	0,096	18	0,282	0,084	18	0,462	0,095	10	0,613	0,115	4	0,588	0,127	2	0,523	0,082	2	0,732	0,044	60	0,04
	cross-support		0,636	0,279		0,663	0,248		0,755	0,162		0,833	0,139		0,680	0,183		0,645	0,106		0,995	0,008		



Continua na próxima página.



Continua na próxima página.

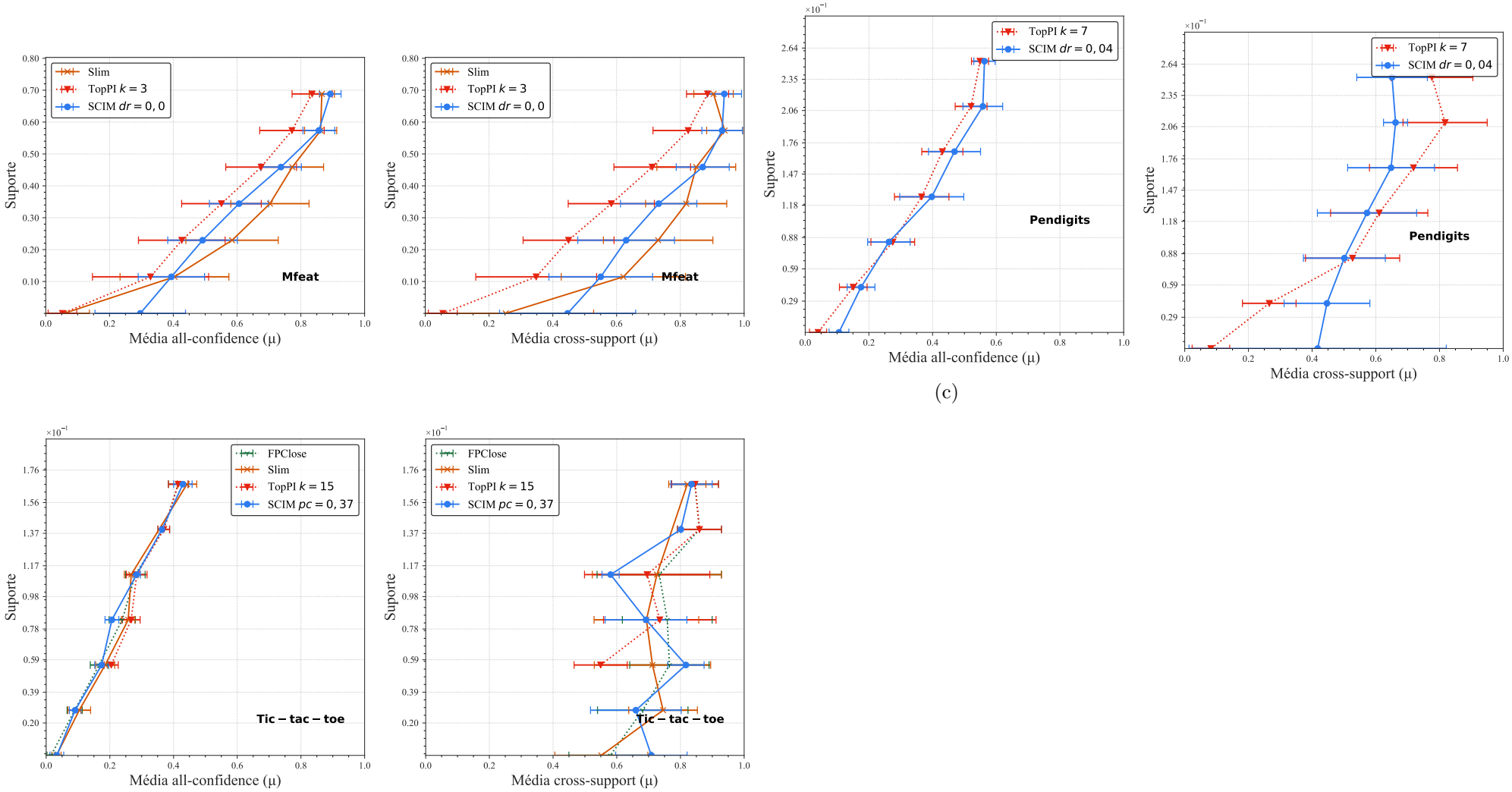


Figura 5.3: Distribuições dos valores médios de *all-confidence* e de *cross-support* dos itemsets fechados recuperados por FPClose, Slim, TopPI e SCIM sobre as bases de dados da Tabela 5.4. Neste estudo foi usado o melhor valor de parâmetro para cada técnica.

Na Figura 5.3 também é possível observar o caso onde a técnica proposta não obteve a maioria das partições de suporte com média alta de *all-confidence*. Porém, vale enfatizar alguns comportamentos observados neste cenário. A separação das curvas não é tão clara nas bases de dados *Pen digits* e *Tic-tac-toe* apresentadas nas Figuras 5.3c e 5.3e. No entanto, o desempenho do algoritmo SCIM é um pouco melhor em *Pen digits* nas faixas de suporte que varia entre [0,00 , 0,04). As bases de dados *Ecoli*, *Pima* e *mFeat* (Figuras 5.3e, 5.3g e 5.3a) apresentam alternância entre as técnicas com melhor valor médio de *all-confidence* dentre as partições de suporte. Quando ao tempo de processamento também são considerados, nossa abordagem geralmente se destaca das outras.

Para validar a significância estatística das médias apresentadas no estudo, foi definido o teste *t-student* [66]. No nosso estudo cada algoritmo é considerado uma variável independente categórica e a variável dependente qualitativa são as amostras de *all-confidence* ou *cross-support* de cada faixa de suporte. No entanto, foi necessário respeitar algumas características para ser possível rodar o teste *t-student*. As amostras devem seguir uma distribuição normal e terem homogeneidade das variâncias. Caso se comprove não haver homogeneidade das variâncias, neste caso é aplicado o teste *Welch's t-student* [65]. E por fim, caso não se comprove que exista uma distribuição normal das amostras, nesse cenário é aplicado o teste não paramétrico *Mann-Whitney* [44]. Ao contrário do teste *t-student*, que testa a igualdade das médias, o teste de *Mann-Whitney* testa a igualdade das medianas. Nosso objetivo é descobrir se dada as hipóteses nulas $H_0 : \mu_{SCIM} \leq \mu$, $H_0 : \mu_{SCIM} = \mu$ e $H_0 : \mu_{SCIM} \geq \mu$, onde μ_{SCIM} sendo as amostras de valores de métricas de *all-confidence* ou *cross-support* dos itemsets fechados minerados pela técnica SCIM e μ ser as amostras de valores de métricas de *all-confidence* ou *cross-support* dos itemsets fechados minerados pelos algoritmos concorrentes. Os testes de significância estatística tem como resultado o *p-value* que é uma medida de quanta evidência você tem contra a hipótese nula. Quanto menor o *p-value*, mais evidência você tem contra a hipótese nula. No caso contrário, quanto maior o valor de *p-value* menos evidência existe para rejeitar a hipótese nula. Partições que não reportaram itemsets fechados ou aqueles que tem no máximo dois itemsets fechados não podem ser usados para realizar o teste. No Apêndice A é apresentado os valores de *p-value* das três hipóteses nulas para cada partição de suporte da base de dados.

A Tabela 5.5 resume os valores de significância estatística obtido por cada técnica comparada em cada base de dados. A primeira coluna contém a informação da base de dados. Já na segunda coluna (#) mostra duas informações separados por / referentes ao algoritmo SCIM. A primeira informação mostra a quantidade de partições que não tiveram

Tabela 5.5: Resumo de significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo SCIM com os algoritmos Slim e TopPI. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas na Tabela 5.4.

Base de dados	**SCIM**	**TopPI**			**Slim**		
	#	#	***All-confidence***	***Cross-support***	#	***All-confidence***	***Cross-support***
Connect	0/0	0/0	2/3/2	3/1/3	4/0	2/0/1	3/0/0
Ecoli	0/3	0/3	1/2/0	0/2/1	1/5	1/0/0	1/0/0
Led7	1/2	1/0	1/1/2	2/1/1	4/1	0/0/0	0/0/0
Letter recognition	0/1	0/0	1/1/4	2/0/4	2/1	1/2/1	3/0/1
mfeat	0/0	0/0	0/0/7	0/0/7	0/0	5/0/2	4/0/3
Page blocks	5/0	5/1	0/0/1	0/0/1	5/0	0/1/1	0/0/2
Pen digits	0/2	0/0	1/0/4	4/0/1	0/1	1/4/0	4/1/0
pima	1/0	2/0	2/1/2	1/1/3	2/3	0/2/0	1/1/0
Tic-tac-toe	0/1	2/0	2/1/1	3/0/1	1/0	4/0/2	2/2/2
Waveform	0/1	0/0	3/2/1	3/1/2	0/2	3/0/1	4/0/0
Wine	0/2	0/0	0/1/4	0/1/4	0/2	0/0/4	0/0/4

itemsets fechados reportados pela técnica e o segundo valor representa a quantidade de partições de suporte que reportaram apenas um ou dois itemsets fechados, essa última informação é importante, visto que não é possível realizar o teste de significância estatística para esse tamanho de amostra. As colunas três e quatro são as técnicas comparadas e cada uma delas contém as seguintes informações: as quantidades de partições sem itemset fechados / quantidades de partições com apenas um ou dois itemsets na partição; quantidade de partições que têm hipóteses nulas não rejeitadas para $H_0 : \mu_{SCIM} \leq \mu$ / $H_0 : \mu_{SCIM} = \mu$ / $H_0 : \mu_{SCIM} \geq \mu$ dado a métrica *all-confidence*; e quantidades de hipóteses nulas não rejeitadas para $H_0 : \mu_{SCIM} \leq \mu$ / $H_0 : \mu_{SCIM} = \mu$ / $H_0 : \mu_{SCIM} \geq \mu$ dado a métrica *cross-support*. Na tabela o texto é marcado com a cor verde quando a quantidade de partições com hipótese nula $H_0 : \mu_{SCIM} \geq \mu$ for maior quando comparado com as outras hipóteses. Outra situação, o texto é marcado com cor vermelha quando a quantidade de partições com hipótese nula $H_0 : \mu_{SCIM} \leq \mu$ for maior quando comparado com as outras hipóteses. E no caso de empate não é alterado a cor na tabela.

Analisando os resultados da Tabela 5.5 podemos notar que o algoritmo SCIM teve melhores resultados de média de *all-confidence* quando confrontado com a técnica TopPI. No entanto, quando usamos a métrica *cross-support* observa-se uma perda de desempenho, chegando a igualar o número de ganhos e perdas. O SCIM quando comparado com o Slim perdeu em desempenho, no entanto, é preciso observar os dois casos *Connect* e *Ecoli*. Neste dois exemplos, embora a técnica Slim tenha ganhado ela reportou itemsets em apenas 3 faixas de suporte, no caso da base *Connect* e 1 para a base de dados *Ecoli*. Vale lembrar que a representatividade de itemsets nas diferentes faixas de suporte da base de dados também conta como critério no momento de selecionar a técnica vencedora.

Quando o tempo de processamento é considerado, o algoritmo SCIM supera outras abordagens em quase todas as bases de dados. As únicas exceções são a base de dados *Connect-4*, onde os tempos de processamento do TopPI e do SCIM foram de 2,87 e 9,98 segundos, respectivamente, e da base de dados *mFeat*, onde a abordagem SCIM passou a maior parte do tempo gerando FP-*tree* condicionais. Em bases de dados esparsas, a FP-*tree* pode se tornar muito complexa e grande [28], para resolver este problema o algoritmo paralelo apresentado no Capítulo 6 lida com este problema substituindo a estrutura FP-*tree* por LCM. Os algoritmos FPClose e Slim são geralmente os mais lentos.

5.4.4 Discussão Sobre a Técnica de Clusterização

Este documento não compara a estratégia de clusterização proposta (Seção 5.2) com outras abordagens de clusterização disponíveis na literatura. Contudo, é importante tecer algumas observações baseados em experimentos realizados durante o projeto. Se adotássemos as estratégias de clusterização com clusters disjuntos, o procedimento proposto de geração de itemsets fechados seria impedido de combinar itens de diferentes clusters. Por outro lado, espera-se que os conjuntos de itens não disjuntos definam contextos para a formação de padrões mais interessantes. Portanto, acredita-se que técnicas como *k-Mean* [37], *Mean-Shift* [15], *Affinity Propagation* [25], e DBSCAN [18] não sejam adequadas para compor o algoritmo proposto.

Nos experimentos iniciais foram testadas as técnicas de clusterização com sobreposições existentes na literatura. É bem conhecido que os clusters identificados por essas técnicas são propensos a ter um nível alto de sobreposições (veja [51] para discussão). Na prática, observou-se que tal comportamento resulta na formação de muitos itemsets fechados irrelevantes. Outros problemas são a necessidade de ajustar vários parâmetros para o processo de clusterização e as técnicas existentes não são adaptáveis para considerar as propriedades do *Dual Scaling*.

Capítulo 6

Mineração Paralela de Itemsets Fechados: PSCIM

O algoritmo SCIM foi definido como uma solução sequencial e, por consequência, não consegue utilizar processamento multinúcleo. A velocidade de um processador é medida pelo número de ciclos por segundo executados pela CPU, em GHz. Por volta de 2005, a velocidade do processador parou de crescer devido à limitação física do *hardware* [32] que, por consequência, resultou na limitação da desempenho dos algoritmos com solução sequenciais. Por causa desta limitação do *hardware*, foi necessário para a indústria intensificar o desenvolvimento de soluções com dois ou mais núcleos em um mesmo *chip*.

Neste trabalho é proposto o algoritmo paralelo PSCIM (*parallel solution to spatial contextualization for closed itemset mining*). O PSCIM é escalável dado o número de *threads*. Diferente da solução sequencial do algoritmo SCIM, que utiliza a estrutura FP-*tree* durante o processo de mineração, o algoritmo paralelo proposto usa a estrutura LCM, que tem complexidade linear de tempo e memória dada a quantidade de itemsets fechados recuperados. PSCIM e SCIM reportam diferentes resultados, o algoritmo paralelo proposto também faz alterações no mapeamento dos itens no espaço de soluções e também faz alteração no algoritmo de clusterização usado no processo de tomada de decisão de inclusão de itemsets ao percorrer a árvore de pesquisa.

Este capítulo cobrirá todas as decisões para definição do algoritmo. No primeiro momento, na Seção 6.1, são mostradas algumas modificações propostas que afetam diretamente o mapeamento do espaço de soluções e a formação dos clusters. Após, na Seção 6.3, são mostradas as soluções paralelas no processo de formação dos clusters e, por fim, na Seção 6.4, é mostrada a solução paralela para o procedimento de mineração de itemsets fechados. O paralelismo apresentado considera um sistema de memória

compartilhada.

O programa principal apresentado no Algoritmo 3, tem como argumentos de entrada a base de dados $\mathcal{D}$ e a distância percentil *dp*. O programa começa com um prefixo vazio na linha 1. As linhas 2 e 3, respectivamente, reduzem a base de dados e criam a estrutura chamada entrega de ocorrência (Subseção 6.2). Essas duas etapas levam uma grande melhoria no custo computacional do *Dual Scaling* e no processo recursivo para identificação de itemsets fechados. As linhas 4 e 5 são responsáveis pela definição dos clusters. PSCIM reordena e indexa os itens da base de dados em ordem decrescente por sua frequência. Finalmente, o processo recursivo é chamado na linha 6.

Algoritmo 3: Função principal do PSCIM

Entrada: Base de dados $\mathcal{D}$, distância percentil *dp*.
Resultado: Itemset fechado em $\mathcal{D}$.

1: $\mathcal{P} \leftarrow \{\}$
2: $\mathcal{G} \leftarrow$ Base de dados $\mathcal{D}$ reduzida
3: $\mathcal{O} \leftarrow$ Estrutura de entrega de ocorrência da base de dados $\mathcal{G}$
4: $\mathcal{S} \leftarrow$ Espaço de soluções da bases de dados $\mathcal{G}$
5: $\mathcal{C} \leftarrow$ SCClusterP($\mathcal{S}$, *dp*)
6: SCCloseLCM($\mathcal{G}$, $\mathcal{O}$, $\mathcal{C}$, $\mathcal{P}$)

Vale ressaltar que a contextualização espacial usada pelo algoritmo SCIM não usa suporte ou qualquer outra métrica estatística como tomada de decisão durante o processo de recuperação de itemsets fechado. O algoritmo segue uma nova vertente, onde a base de dados é mapeada para o espaço de soluções, calculada pelo algoritmo *Dual Scaling*. Os itens são projetados no espaço de soluções, onde as distâncias entre os pares de itens podem inferir algumas informações relevantes. Dada sua grande relevância, é proposta uma solução paralela, o algoritmo PSCIM, para possibilitar o uso dessa nova vertente em bases de dados que demandam cada vez mais poder computacional.

6.1 SCIM Revisitado

O algoritmo SCIM tem um valor do parâmetro de fácil definição, por parte do usuário, porque os experimentos mostram que, para as bases de dados consideradas no estudo, a tendência é que o mesmo seja definido como $dr = 0{,}0$. Este aspecto é uma vantagem para bases de dados onde não se sabe o comportamento da informação. Porém, para algumas bases de dados, observou-se que a cobertura mínima do raio no cluster não foi suficiente para agregar itens relevantes ao mesmo e, consequentemente, não foi possível

reportar alguns dos itemsets no processo de mineração de itemsets fechados. O SCIM define a limite máximo de razão de distância (dr) que resulta em um algoritmo mais flexível (Equação 5.2), e este parâmetro define os raios dos clusters no processo de clusterização (Algoritmo 1).

Os testes realizados com o algoritmo SCIM (Capítulo 5) consistiam no total de 11 bases de dados densas de múltiplas escolhas. Portanto, não foi realizada uma análise do comportamento da técnica sobre bases esparsas, embora não haja limitação do algoritmo *Dual Scaling* para base de dados deste tipo.

Durante os testes iniciais em bases de dados esparsas, observam-se algumas características sobre o comportamento da técnica SCIM. A técnica não conseguiu relatar itemsets fechados em alguns intervalos de suporte dada uma base de dados esparsa. Para obter itemsets fechados nesses intervalos de suporte, foi necessário definir valores altos do parâmetro dr. Porém, neste cenário, a técnica retornou muitos itemsets espúrios, dificultando a leitura dos itemsets recuperados. Outro ponto analisado, muito comum em bases de dados esparsas, são as transações com apenas 1 item. Em termos práticos, essas transações são eliminadas durante o processo de mineração, visto que os itemsets com um item não são considerados interessantes para o entendimento das relações entre itens das bases de dados. Porém, isso levantou uma questão sobre o impacto dessas transações no cálculo do *Dual Scaling* e, consequentemente, na formação dos clusters.

De forma geral, para determinada base de dados com m itens há também m clusters, onde cada cluster tem seu valor de cobertura mínimo e máximo, e o tamanho da região de cobertura extra pode variar muito para cada cluster. Percebe-se que a distância entre os itens dentro do cluster não segue uma distribuição homogênea entre as distâncias dos itens. Por exemplo, todos os itens de um determinado cluster podem estar próximos da cobertura mínima, enquanto em outro cluster todos os itens podem estar próximos da cobertura máxima. É evidente que usar a razão de distância (dr) para calcular a distância de cobertura extra não garante a adição de novos itens aos clusters igualmente. O cluster possui um conjunto de itens ordenados pela distância entre cada item e a origem. Assim, ao revisar o algoritmo, é proposto usar o percentil de distância (dp) para controlar o raio do cluster para, consequentemente, ter uma agregação mais homogenia de novos itens nos clusters. Agora, usando o parâmetro dp, pode-se definir que a cobertura do cluster C_i é calculada como:

$$cvp_i = d'_{i,\,\lfloor t+(dp\times q)\rfloor}. \tag{6.1}$$

Para o i-th item da base de dados temos $d'_i = \left\{d'_{i,1}, d'_{i,2}, \cdots, d'_{i,m}\right\}$, que são valores orde-

nados de forma crescente dada as distâncias d_i. O valor t é a posição do último item que se encontra a cobertura mínima, e q é a quantidade de itens que não estão dentro da cobertura mínima do cluster. Logo pode-se redefinir a cobertura do cluster como:

Definição 17 (Cobertura do cluster com percentil) *O cluster C_i inclui todos os pontos de itens que satisfazem $d_{i,l} \leq cvp_i$, onde $d_{i,l}$ e cvp_i são calculados por, respectivamente, Equação 4.3, Equação 6.1 e Equação 4.4, para $l \in \{i' \mid a_{i,i'} \geq 0\}$ e $i, i' \in \{1, 2, \cdots, m\}$. A cobertura do cluster é parametrizada pelo limite definido pelo parâmetro $dp \in [0, 1]$.*

Outra observação interessante que pode ser feita ao analisar as bases de dados esparsas são os casos onde ocorrerem transações com apenas um item. Esse tipo de transação não agrega nenhum conhecimento em itemsets fechados, pois não há correlação deste item com outro na transação na base de dados. Além disso, esse tipo de transação também impacta em algumas métricas usadas para qualificar se um itemset fechado é relevante. Pode-se citar, como exemplo, dado os itens a e b, totalmente correlacionados em uma base de dados $\mathcal{D}$, ou seja, o $allconf(\{a, b\}) = 1$ (Equação 2.3). Quando se insere novas transações com apenas o item a na base de dados $\mathcal{D}$, o *all-confidence* diminui. Além disso, esse comportamento pode ser observado no espaço de soluções. Itens altamente correlacionados e, consequentemente, próximos no espaço de solução, tendem a ser projetados mais distantes. Objetivamente, esse cenário dificulta a identificação desses itemsets no cluster, dado o raio mínimo de cobertura. Em muitos casos, há necessidade de usar a cobertura extra para identificar a correlação entre eles, embora altamente correlacionados.

São propostas duas mudanças diretamente relacionadas à formação dos clusters com relação ao que é apresentado no SCIM. A primeira proposta é sobre o parâmetro dr, que atualmente é uma razão linear entre os valores de cobertura mínimo e máximo. É proposta a modificação do parâmetro dr para se tornar uma razão de percentil dos itens que estão fora da cobertura mínima e, estes, ordenados por uma distância dentro de cada cluster. Ou seja, percentil de distância (dp). A segunda mudança proposta está relacionada às transações de tamanho um. Neste caso, são removidas essas transações da matriz de entrada F do *Dual Scaling* (Equação4.1), mas não são removidas essas transações da base de dados original, porque essas transações fazem parte do cálculo da métrica frequência que será usada para ordenar os itens da base de dados por frequência.

6.2 Mineração de Itemset Fechado em Tempo Linear

Antes de definir o algoritmo PSCIM, é necessário entender as características da estrutura usada para percorrer o espaço de busca no processo de mineração de itemsets fechados. A principal característica do algoritmo LCM é percorrer o espaço de busca apenas por itemsets fechados. O algoritmo LCM pode encontrar todos os itemsets fechados com uma complexidade linear de tempo e espaço, dada a quantidade de itemsets fechados reportados. Nesta seção, descreve-se a versão 2 deste algoritmo [59] utilizado pelo PSCIM no processo de mineração de itemsets fechados.

LCM é um algoritmo de *backtracking* que, por um processo recursivo, percorre o espaço de busca entre as combinações de itens da base de dados. No algoritmo LCM, cada recursão recebe um itemset fechado P e, o mesmo, usa uma enumeração dos itemsets para evitar a duplicação no processo de busca.

No algoritmo LCM, os itens são renomeados por sua frequência, em ordem decrescente. Para o itemset P e item $i \in P$, seja $P(i) = P \cap \{1, \cdots, i\}$ o subconjunto de P, com itens não maiores que i. O índice central de P, denominado por $core_i(P)$, é o índice mínimo i tal que $T(P(i)) = T(P)$, sendo $core_i(\{\}) = 0$. O itemset fechado P' é um ppc-*extension* (*prefix preserving closure extension*) de P se $P' = clo(P \cup \{i\})$ e $P'(i-1) = P(i-1)$ para o item $i \notin P$ e $i > core_i(P)$. Esta propriedade fornece a relação única entre o prefixo P e o ppc-*extension* P'.

Durante o processo, é necessário fazer vários acessos à base de dados. O LCM descreve algumas etapas para reduzir o custo desta operação. Essas etapas são: remover itens pouco frequentes; remover itens em todas as transações e alocar em outra estrutura; e mesclar transações duplicadas. No final das etapas apresentadas, existe uma base de dados reduzida R, ou seja, $R = \{T_1, \cdots, T_k\}$, onde $T_i \neq T_j$ para todos os $i \neq j$ e W_k é o peso que corresponde ao número de ocorrências de T_i na base de dados original. Essa estrutura de peso mantém as informações de frequência originais da base de dados. Antes de juntar as transações repetidas, o algoritmo radix [59] foi usado para ordenar as transações.

Uma das etapas mais caras do algoritmo é contar a frequência de itemsets fechados. Por ser necessário acessar a base de dados várias vezes, o LCM utiliza uma estrutura capaz de indexar as transações por item dado o prefixo P. Com essa abordagem, é possível acessar todas as frequências de $P \cup \{e\}$ com apenas uma passada na base de dados reduzida corrente.

Devido ao escopo desta teste, não são mostradas as provas que comprovam o funciona-

mento da estrutura LCM. Consulte [57, 59] para detalhes sobre as definições mencionadas acima.

6.3 Procedimento de Clusterização

Para deixar o algoritmo mais flexível, são propostas modificações no processo do *Dual Scaling*. A primeira modificação que se definiu foi utilizar a base de dados reduzida (Subseção 6.2). Dada uma base de dados reduzida R onde uma transação T_i tem o peso W_i. Para manter a integridade da Equação 4.1, foi necessário reescrevê-la:

$$M = R^T \, D_w \, D_r^{-1} \, R \, D_c^{-1}, \tag{6.2}$$

onde D_w é a matriz diagonal de peso das linhas. É simplificada a multiplicação da Equação 6.2 onde só é necessário calcular onde $R_{i,j} \neq 0$. Não houve problemas em paralelizar o processo de multiplicação, pois é definido que a matriz resultante é densa. Cada *thread* usa o índice do item para processar parte da matriz de pesos projetados (Equação 4.2).

Este trabalho vai além de apenas tornar o algoritmo SCIM escalonável dado o número de *threads*. Também visa reduzir o custo computacional inerente ao processo. Investigando a matriz M, pode-se identificar a seguinte propriedade.

Definição 18 (Coocorrência entre itens em M) *Para qualquer par de itens i e j em uma base de dados $\mathcal{D}$, $M_{i,j} = M_{j,i}$ se mantém para $|T(i,j)| = 0$.*

A simetria em M é evidente para valores iguais a zero. Para qualquer matriz $A^T \, A$ é uma matriz quadrada e simétrica. No entanto, não se pode afirmar isso para toda a matriz M, pois existem matrizes diagonais D_c, D_r e D_w multiplicando-se como um valor escalar quando $M_{i,j} \neq 0$

A Definição 18 pode ser usada para evitar o custo computacional do cálculo da matriz de distância. Não é necessário calcular a distância para todos os pares de itens na base de dados. Sendo que para alguns pares, $M_{i,j} = 0$, não há coocorrência na base de dados. Consequentemente, não há itemsets fechados a serem relatados neste cenário. Para o Algoritmo 1, cada *thread* usa o índice de itens na base de dados.

O Algoritmo 4 apresenta a nova função de clusterização, tendo como entrada o espaço de soluções e, definido pelo usuário, o parâmetro *dp* de cobertura dos clusters. Nas linhas 2 e 3, respectivamente, é calculada a matriz de distância (Equação 4.3), e a matriz

Algoritmo 4: Clusterização de itens no espaço de soluções

1 **Função** `SCClusterP`(X, dp)
  Entrada: X espaço de soluções; distância percentil dp.
  Resultado: Cluster para cada item da base de dados
2 $D \leftarrow$ matriz de distância de pares de itens de X (Equação 4.3)
3 $A \leftarrow$ matriz de arcocoseno de pares de itens de X (Equação 4.4)
4 **para cada** $i \in \mathcal{I}$ **faça**
5 $\mathcal{C}_i \leftarrow \{\}$
6 Calcular cobertura do cluster corrente cvp_i usando D, A e dp (Equação 5.2)
7 **para cada** $i' \in \mathcal{I}$ **faça**
8 **se** $D_{i,i'} \leq cvp_i$ ***e*** $A_{i,i'} \geq 0$ **então**
9 $\mathcal{C}_i \leftarrow \mathcal{C}_i \cup \{i'\}$
10 **fim**
11 **fim**
12 **fim**
13 **retorna** $\mathcal{C}$
14 **fim**

de arcocoseno entre os pares de itens dado o espaço de soluções X (Equação 5.2). Na linha 4 é feita uma interação em cada item da base de dados. Segundo a Definição 17, cada item tem seu cluster associado. Por isso, para cada item i da base de dados é definido o cluster $\mathcal{C}_i$ (linha 5). A cobertura total do cluster, cvp_i, é definida pela Equação 6.1 (linha 6), que calcula o raio do cluster corrente $\mathcal{C}_i$. A linha 7 percorre todos os itens da base, no intuito de identificar itens que pertence ao cluster corrente. Para ser considerado item pertencente ao cluster, o par de itens i e i' devem respeitar a condição de distância, onde a distância do par de itens não pode ser maior que a cobertura total do cluster corrente, e terem arcocoseno maior que 0 (linha 8). Desta forma, o procedimento de clusterização atualiza o cluster corrente com o item i' (linha 9).

As modificações propostas não alteram a complexidade para o algoritmo de clusterização. No entanto, tornou o custo do processo mais baixo para bases de dados esparsas. A complexidade do tempo é definida pelas Equações 6.2, 4.2, 4.3, e a decomposição de autovalores e vetores. Seja m número de itens e n número de transações, $O((m^2\,n) + (m^2 - m) + (m^3 - m^2) + (m^3))$, limitado pelo sistema de decomposição, de complexidade $O(m^3)$. A complexidade espacial tem o limite superior dada pela matriz M de $O(m^2)$.

6.4 Geração de Itemset Fechado

O algoritmo SCIM usa durante o processo de mineração a estrutura FP-*tree* e, através de uma busca em profundidade, consegue enumerar todos os itemsets possíveis. Além disso, na solução é preciso usar a estrutura CFI-*tree* para garantir que apenas itemsets fechados sejam reportados. Este capítulo tem como finalidade apresentar uma solução paralela para o algoritmo SCIM. Durante o estudo perceberam-se alguns desafios nas etapas de busca de itemsets fechados. A primeira dificuldade é gerenciar o compartilhamento paralelo da FP-*tree*, uma solução para o problema é proposta por Dehao [12], porém vale salientar que essa solução percorre todos os itemsets frequentes dado o mínimo suporte. O segundo problema é gerenciar a atualização da árvore CFI-*tree*. Neste caso, a solução proposta por Dehao, foi definir *blocks* durante a atualização da estrutura para manter a integridade das informações.

Portanto, dado as dificuldades expostas, optou-se por usar a estrutura LCM para pesquisar itemsets fechados. Ao contrário da FP-*tree* com CFI-*tree*, a estrutura LCM pode enumerar todos os itemsets fechados dado o limiar no tempo de complexidade linear e sem estruturas extras, além de ser mais fácil de gerenciar em uma solução paralela.

O Algoritmo 5 apresenta o novo procedimento recursivo `SCCloseLCM`, que recupera itemsets fechados usando a contextualização espacial. O algoritmo LCM navega por todos os itemsets fechados dado o suporte mínimo. Por outro lado, o novo algoritmo paralelo proposto navega por todos os itemsets fechados representados nos clusters, seguindo a Definição 19.

Definição 19 (Itemset fechado em clusters) *Dado qualquer itemset P e seu ppc-extension Q. Seja C uma coleção de clusters contendo P, e seja $L = \{c \cap Q \mid c \in C\}$ é um conjunto de todos os itemsets fechados definidos em cada cluster. O itemset fechado $S = \bigcap_{l \in L} l$ é reportado se $S \in L$ e S é o único itemset fechado que obedece à regra de formação do itemset em L (Definição 16).*

A Definição 19 atende a três cenários onde o itemset fechado não é reportado. No primeiro cenário, L possui apenas o itemset fechado Q, portanto $S = Q$. No entanto, S não respeita a formação de um itemset no cluster. No segundo cenário, L tem subconjuntos de Q. Nesta condição, S não pertence à lista L, logo S não será reportado porque é um subconjunto de todos os itemsets fechados em L. Por isso S ainda não é o itemset fechado para os clusters $\mathcal{J}$ e irá ser relatado posteriormente na recursão como possível

Algoritmo 5: Busca na estrutura LCM para identificar itemsets fechados

Entrada: Base de dados reduzida $\mathcal{G}$, entrega de ocorrência $\mathcal{O}$, clusters C, prefixo P
Resultado: Reportar itemsets fechados

```
1  Procedimento SCCloseLCM(𝒢, 𝒪, 𝒞, 𝒫)
2      para cada e > core_i(𝒫) faça
3          𝒥 = {c | c ∈ 𝒞 and c ⊇ (𝒫 ∪ {e})}
4          se |𝒥| ≠ 0 então
5              𝒬 = clo(𝒫 ∪ {e})
6              ℒ = {𝒬 ∩ j | j ∈ 𝒥}
7              𝒮 = ⋂_{l∈ℒ} l
8              se 𝒫(e − 1) = 𝒮(e − 1) então
9                  se 𝒮 é um itemset fechado válido (Definição 19) então
10                     reportar itemset fechado 𝒮
11                 fim
12                 ℋ ← Base de dados 𝒢 reduzida
13                 ℛ ← Estrutura de entrega de ocorrência da base de dados ℋ
14                 SCCloseLCM(ℋ, ℛ, 𝒥, 𝒮)
15             fim
16         fim
17     fim
18 fim
```

itemset fechado. No último cenário, $S \in L$ e há mais de um itemset fechado além de S que obedecem às regras de formação do itemset em L. Neste caso S será relatado posteriormente na recursão, visto que S é um subconjunto de todos os itens fechados definidos em L.

Na primeira chamada do procedimento, `SCCloseLCM` recebe como parâmetro a base de dados reduzida $\mathcal{G}$; a estrutura entrega da ocorrência $\mathcal{O}$; os clusters $\mathcal{C}$, calculado conforme a Seção 6.1; e o prefixo vazio P. Vale ressaltar que a redução da base de dados não exclui itens pela sua frequência. O algoritmo PSCIM não usa suporte mínimo. Neste laço, linha 2, o item $e \in \mathcal{I}$ é visitado no modo de busca em profundidade, se $e > core_i(P)$. Na linha 3, cria-se um novo conjunto de clusters correntes $\mathcal{J}$ com clusters em $\mathcal{C}$ que têm o itemset $(P \cup \{e\})$. Se $\mathcal{J}$ estiver vazio então não há necessidade de descer na árvore de busca, visto que não há clusters com itemsets fechados gerados a partir dessas chamadas de recursão.

Na linha 5, o itemset fechado Q é gerado pelo itemset $(P \cup \{e\})$ (Definição 4). Ao contrário do LCM, que usa suporte mínimo para selecionar itemsets interessantes, o PSCIM usa as informações dos clusters como tomada de decisão. Na linha 6, para cada cluster em $\mathcal{J}$ é encontrado o itemset fechado $\mathcal{L}_k \subseteq \mathcal{Q}$. No entanto, para cada recursão, apenas um itemset fechado pode ser reportado. Por isso, a linha 7, relata o itemset comum S

em todos os itemsets fechados reportados no conjunto $\mathcal{L}$. Na linha 8, há a verificação que garante a propriedade de ppc-*extension* (Seção 6.2) evitando, desta forma, reportar itemsets fechados duplicados. Na linha 9 verifica-se o itemset fechado S pertence à lista L. Caso contrário, S não é reportado, pois S é um subconjunto de um possível itemset fechado reportado posteriormente nas recursões da árvore de busca. Se $S \in \mathcal{L}$ e se houver mais de um itemset fechado válido, significa que S será relatado posteriormente na recursão. A linha 12 aplica a redução na base de dados para o itemset S e a linha 13 define a estrutura de entrega da ocorrência dada a base de dados reduzida $\mathcal{H}$. A chamada recursiva (linha 14) recebe como parâmetros a base de dados reduzido $\mathcal{H}$, a entrega de ocorrência $\mathcal{R}$, os clusters atuais $\mathcal{J}$ e o itemset fechado S.

A paralelização do procedimento `SCCloseLCM` foi aplicada na linha 2. No entanto, criar uma *thread* para cada chamada recursiva pode gerar muitas *threads*, e muitas delas teriam pouco trabalho associado. Vale ressaltar que criar e gerenciar *threads* vem com algumas sobrecargas, por isso que o número de *threads* não deve ser muito alto. Por fim, o procedimento `SCCloseLCM` restringi a criação de *threads* até o segundo nível de profundidade das chamadas recursivas. A solução paralela proposta utiliza memória compartilhada.

A complexidade de tempo do algoritmo pode ser definida pela busca na estrutura LCM, $O(|\mathcal{T}(P)| + \sum_{e>core_i(P)} |\mathcal{T}(P \cup \{e\})|)$, e pela contextualização espacial do clusters usado como tomada de decisão, $O(|\mathcal{C}(P)| + \sum_{e>core_i(P)} |\mathcal{C}(P \cup \{e\})|)$, para cada itemset fechado P. Aqui, $|\mathcal{A}| = \sum_{A \in \mathcal{A}} A$, e $\mathcal{C}(A)$ sendo todos os clusters com itemset A e $\mathcal{T}(A)$ sendo todas as transações com itemset A. A complexidade espacial é definida em função do número de itens da base de dados $\mathcal{D}$, i.e., $O(|\mathcal{I}|^2)$, e o tamanho da base de dados $\mathcal{D}$, i.e., $O(|\mathcal{D}|)$.

6.5 Resultados

Os experimentos apresentados nesta seção tiveram como objetivo compreender melhor o comportamento do algoritmo PSCIM tanto num ambiente de processamento paralelo quanto na qualidade dos itemsets fechados. Na Seção 6.5.2 é mostrado o comportamento de seleção dos itemsets fechados, dada as mudanças propostas sobre a formação dos clusters. Na Seção 6.5.3 é comparada a desempenho do algoritmo proposto com algoritmos da literatura. Na Seção 6.5.4 são mostradas as proporções de tempo de processamento para a formação do cluster. Por fim, na Seção 6.5.5 é mostrado o *speedup* do algoritmo

paralelo proposto. O algoritmo proposto é implementado usando C++. A solução paralela foi implementada com a biblioteca OpenMP [16]. Neste projeto, usa-se a biblioteca StarNeig [42] porque é considerado o estado da arte no cálculo paralelo de autovalores e autovetores para matrizes densas e não simétricas. Os experimentos foram realizados em um PC com CPU Intel I7 4.0GHz e 16 GB de RAM rodando Windows 8 64-*bits*.

Foi avaliado o algoritmo proposto realizando experimentos e comparando a performance do PSCIM com os algoritmos Slim [55], TopPI [30] e LAM [11]. Foi usado a implementação fornecida pelos autores publicamente[1,2] ou por e-mail disponibilizado pelo autor.

O algoritmo PSCIM define para cada item da base de dados um cluster para o processo de mineração, seguindo a mesma ideia, o TopPI é centrado nos top-k de cada item da base de dados. O algoritmo Slim é estado da arte nos algoritmos de utilizam a ideia que o conjunto de itemsets fechados relevantes reportados podem ser usados para comprimir a base de dados original. O Algoritmo LAM, também utiliza a ideia de reportar conjuntos de itemsets fechados relevantes que melhor comprime a base de dados, porém agrupa transações semelhantes no seu processo de mineração. Neste sentido para essas duas técnicas, seria interessante comparar o comportamento de seleção do PSCIM, visto que a técnica propõe reportar itemses fechados relevantes.

Tabela 6.1: Informações sobre as bases de dados

Base de dados	n	m	**Tamanho médio das transações**	**Esparsa**
accidents	340,183	468	33.8	✓
BMS_ WebView_ 2	77,512	3,340	4.62	✓
BMS1	59,602	497	2.51	✓
chess	3,196	75	37	
foodmartFIM	4,141	1,559	4.42	✓
Fruithut	181,970	1,265	3.58	✓
kddcup99	1,000,000	135	16	
mushrooms	8,416	119	23	
OnlineRetail	541,909	2,603	4.37	✓
PAMP	1,000,000	141	23.93	✓
PowerC	1,040,000	140	7	
pumsb	49,046	2,113	74	
RecordLink	574,913	29	10	
retail	8,162	16,470	10.30	✓
Skin	245,057	11	4	
Susy	5,000,000	190	19	

Foram utilizadas 16 bases de dados de SPMF [21]. A Tabela 6.1 mostra dados gerais

[1]Slim: `https://people.mmci.uni-saarland.de/~jilles/prj/slim`
[2]TopPI: `https://github.com/slide-lig/TopPI`

sobre as bases de dados. A primeira coluna contém o nome da base de dados, a segunda o número de transações, a terceira o número de itens, a quarta contém o tamanho médio das transações e, a última coluna, indica se é uma base de dados esparsa. Em algumas bases de dados, foram removidas transações com zero itens e itens duplicados na mesma transação sempre que esses foram observados. Considera-se esses dois casos como um erro para uma base de dados transacional. As bases de dados foram convertidas em matrizes de resposta padrão $(1, 0)$ antes dos experimentos.

6.5.1 Definição de Métricas

As métricas utilizada nesse estudo são iguais às do estudo do Capítulo 5, com o adicional de mais uma métrica, conforme descrito nas subseções que seguem.

6.5.1.1 Primeira Métrica: *All-confidence* dos Itemsets Fechados Selecionados

O objetivo do uso desta métrica é entender o comportamento de geração de itemsets fechados. A métrica *all-confidence* é usada para qualificar se este conjunto contém itemsets fechados relevantes. Esta medida tem variação de [0,1], onde valores próximos de 1 indicam que os itemsets fechados geram regras de associações com altos valores de confiança, enquanto valores próximos de 0 indicam que os itemsets fechados geram regras de associação com baixo valores de confiança.

Nos experimentos, é comparado as distribuições de medidas de *all-confidence* médio dos itemsets fechados recuperados pelo Slim, TopPI, LAM e PSCIM. Para calcular a média de *all-confidence*, primeiro cria-se sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte, dado uma determinada base de dados. Em seguida, calcula-se a média de *all-confidence* a partir dos conjuntos de itens fechados recuperados em cada partição. Ao fazê-lo, é possível inspecionar a distribuição dos valores médios de *all-confidence* dos itemsets fechados em todas as frequências de suporte e assim verificar a capacidade de cada técnica de recuperar itemsets fechados em cada partição de suporte.

6.5.1.2 Segunda Métrica: *Cross-support* dos Itemsets Fechados Selecionados

O objetivo do uso desta métrica é entender o comportamento de geração de itemsets fechados. A métrica *cross-support* é usada para qualificar se este conjunto contém itemsets fechados relevantes. Esta medida tem variação de [0,1], onde valores próximos de 1 indicam que os itens deste itemset fechado são bastante correlacionados, enquanto valores próximos

de 0 indicam que os itens deste itemset fechado são pouco correlacionados.

Da mesma forma que a métrica *all-confidence*, nos experimentos são comparados as distribuições de medidas de *cross-support* médio dos itemsets fechados recuperados pelo LAM, Slim, TopPI e a abordagem proposta. Para definir a média de *cross-support*, primeiro são criados sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte dada uma determinada base de dados. Em seguida, calculamos a média de *cross-support* a partir dos itemsets fechados recuperados em cada partição. Ao fazê-lo, é possível inspecionar a distribuição dos valores médios de *cross-support* dos itemsets fechados em todas as frequências de suporte e assim verificar a capacidade de cada técnica de recuperar itemsets fechados em cada partição de suporte.

6.5.1.3 Terceira Métrica: Tempo de Execução

Esta métrica tem como foco identificar o quão custoso é determinado algoritmo dada as variações de parâmetros definidas durante o teste. Os valores dos tempos são definidos em segundos. Foi utilizada a implementação fornecida pelos autores de cada técnica comparada. Nos experimentos foi contabilizado todo o processo dos algoritmos, deste a leitura da base de dados até o resultado dos itemsets fechados selecionados.

6.5.1.4 Quarta Métrica: *Speedup*

Esta métrica tem como foco identificar a relação de tempo gasto para execução do problema com apenas um processador e o tempo gasto com N processadores, i.e., o *speedup* de ganho de tempo. Este pode ser calculado como:

$$S = \frac{T(1)}{T(N)}, \tag{6.3}$$

onde S é o *speedup* e $T(N)$ o tempo gasto com N processadores. No trabalho proposto é tirado do cálculo o tempo de leitura da base de dados e na escrita em memória secundária.

6.5.2 Discussão Sobre SCIM Revisitado

Para observar o efeito das mudanças propostas na formação dos clusters (Seção 6.1), consideradas as combinações das duas mudanças, ou seja, o percentil para modular o raio dos clusters e a remoção das transações de tamanho um. Para o cenário de estudo cada variação tem o prefixo PSCIM, como resultado é definido: a implementação do PSCIM, mas considerando o tratamento da base de dados e o *dr* do SCIM original (**PSCIM$_{\mathbf{v1}}$**);

a variação de PSCIM_{v1} com modificação no procedimento de clusterização, usando percentil para modular o raio do cluster, com o novo parâmetro *dp*, usados para estender a cobertura dos clusters (**PSCIM_{v2}**); a variação PSCIM_{v1} com modificação na base de dados da entrada retirando transações de tamanho 1, ou seja, aquelas que contém apenas um item na transação (**PSCIM_{v3}**); e a combinação das variações PSCIM_{v2} e PSCIM_{v3} (**PSCIM_{v4}**).

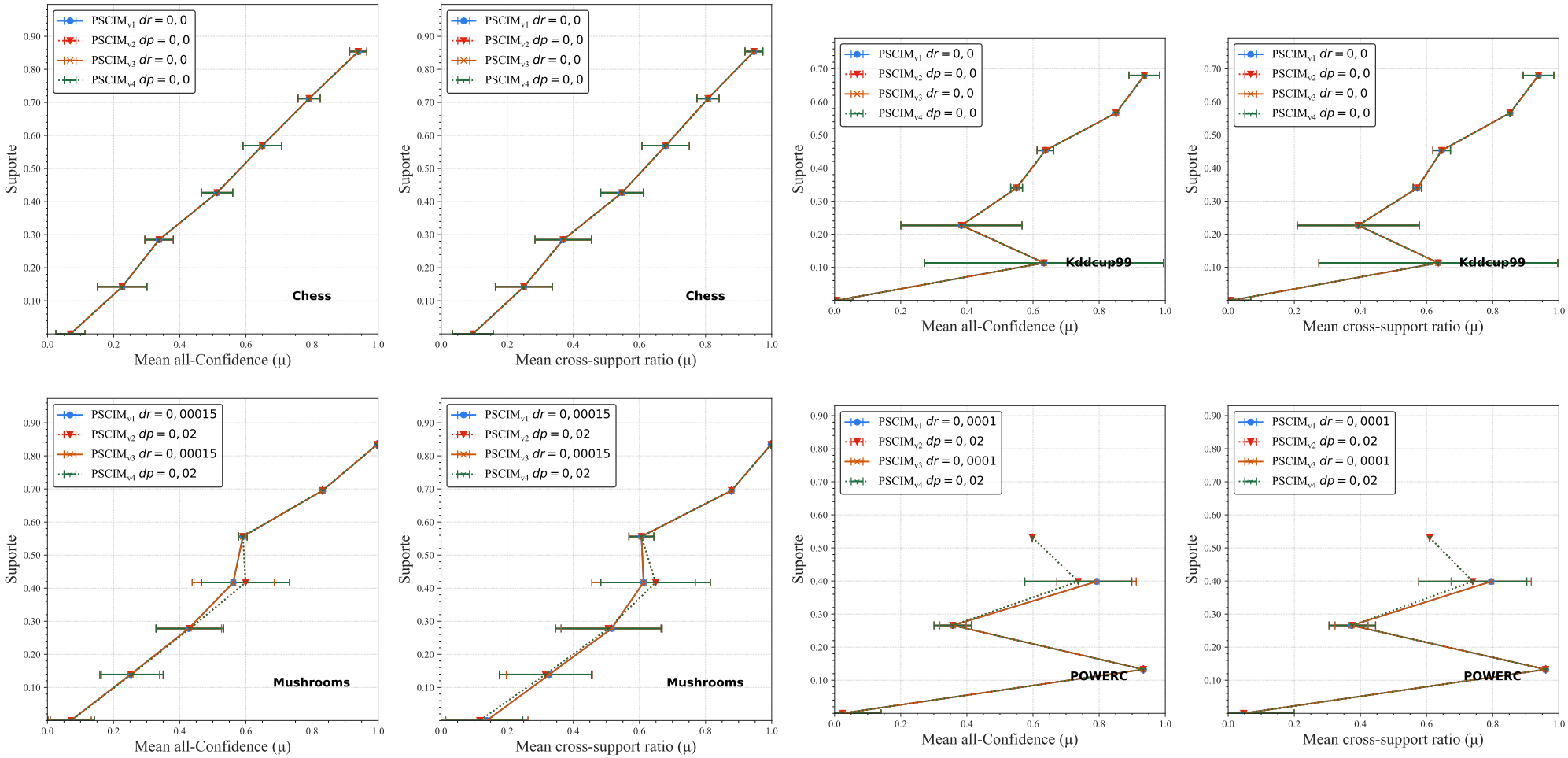
Esta seção não compara o tempo de processamento para cada variação PSCIM, visto que esta informação não está relacionada à discussão. Os detalhes sobre a variação do algoritmo PSCIM sob diferentes parâmetros são descritos no Apêndices B. Nesse Apêndice é possível ver a evolução de seleção de itemsets fechados do caso automático, ou seja, $dr = 0$ ou $dp = 0$, até o último valor de parâmetro considerado.

Como pode ser visto nas Figuras 6.1 e 6.2 os eixos horizontais correspondem à média de *all-confidence* variando de 0 a 1, enquanto os eixos verticais correspondem aos valores de suporte de 0 até o limite superior de cada base de dados. Espera-se que quanto melhor for o conjunto de itemsets fechados recuperados, mais à direita estará a curva que representa o desempenho de uma técnica. A métrica *cross-support*, definida na Seção 6.5.1.2, não é usado como critério de escolha.

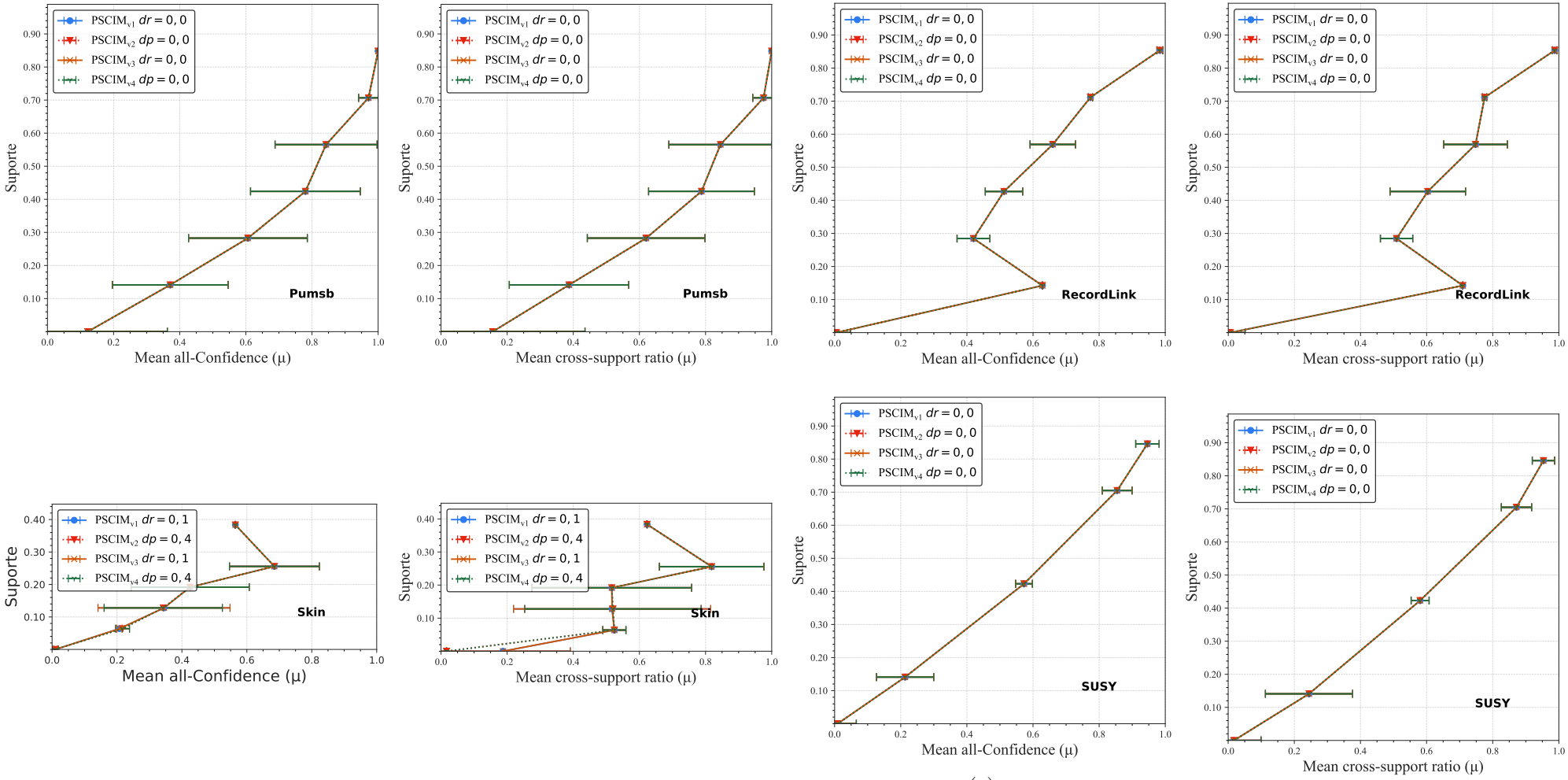
As Tabelas 6.2 e 6.3 resumem o melhor parâmetro alcançado por cada variação do algoritmo PSCIM para uma determinada base de dados. A primeira coluna contém a informação da base de estudo com as informações de parâmetros escolhidos para cada variação do algoritmo PSCIM. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. As colunas três a nove mostram o número de conjuntos de itens fechados recuperados (#), os valores médios da métrica corrente (μ) e os valores de desvio padrão (σ) calculados por intervalo de suporte por todas as técnicas comparadas. Esses intervalos de suporte têm o mesmo tamanho por base de dados e estes tamanhos foram definidos sobre o intervalo de suporte da base de dados. A última coluna mostra o número total de itemsets fechados recuperados. O entendimento utilizado para escolher a melhor variação do PSCIM é igual ao processo de escolher o melhor parâmetro *dr* ou *dp* em cada variação do PSCIM. Ou seja, quanto melhor for o conjunto de itemsets fechados recuperados, maior será a média de *all-confidence* em cada partição de suporte. Destaca-se em negrito a melhor variação do algoritmo PSCIM para cada base de dados. Em um cenário de empate, escolhe-se aquela com mais combinações de variações em sua definição.

Na Tabela 6.2, na primeira coluna, todos os nomes das bases de dados, exceto a

base de dados *Accidents*, possuem o símbolo (τ) na frente do nome. Isso significa que a base de dados atual possui transações com tamanho 1, ou seja, existem transações que contém apenas um item. O símbolo (†) pode aparecer nas colunas três a nove em bases de dados onde o intervalo de suporte é inferior a 0,10. Nesses casos, usa-se a notação científica para representar os valores de suporte. A Figura 6.3 ilustra a distribuição de μ das bases de dados esparsas. Se as faixas de suporte forem inferiores a 0,10, a proporção da imagem original não é mantida porque, nesses casos, o eixo y é muito menor que o eixo x, dificultando a leitura do gráfico.



Continua na próxima página.



(g)

Figura 6.1: Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM$_{v1}$, PSCIM$_{v2}$, PSCIM$_{v3}$ e PSCIM$_{v4}$ sobre as bases de dados densas da Tabela 6.2. Neste estudo foi usado o melhor valor de parâmetro para cada variedade de PSCIM.

Tabela 6.2: Desempenho do algoritmo/parametrização para as variações $\text{PLSCIM}_{\text{v1}}$, $\text{PLSCIM}_{\text{v2}}$, $\text{PLSCIM}_{\text{v3}}$, e $\text{PLSCIM}_{\text{v4}}$ sobre as bases de dados densas da Tabela 6.1.

Chess	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	
V2 $dr = 0{,}0$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	
V3 $dr = 0{,}0$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	
V4 $dr = 0{,}0$	**all-confidence**	**42**	**0,069**	**0,044**	**29**	**0,226**	**0,075**	**42**	**0,338**	**0,043**	**66**	**0,513**	**0,047**	**26**	**0,650**	**0,058**	**8**	**0,791**	**0,034**	**12**	**0,940**	**0,026**	**225**
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	

Kddcup99	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,11]			(0,11 , 0,23]			(0,23 , 0,34]			(0,34 , 0,45]			(0,45 , 0,57]			(0,57 , 0,68]			(0,68 , 0,79]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	
V2 $dr = 0{,}0$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	
V3 $dr = 0{,}0$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	
V4 $dr = 0{,}0$	**all-confidence**	**1.472**	**0,007**	**0,049**	**113**	**0,633**	**0,361**	**10**	**0,384**	**0,183**	**17**	**0,550**	**0,018**	**37**	**0,637**	**0,025**	**1**	**0,851**	**0,000**	**113**	**0,937**	**0,047**	**1.763**
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	

Mushrooms	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,83]			(0,83 , 0,97]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}00015$	all-confidence	289	0,071	0,062	108	0,251	0,088	57	0,427	0,099	19	0,562	0,124	3	0,591	0,013	1	0,832	0,000	1	0,997	0,000	478
	cross-support		0,139	0,124		0,328	0,130		0,516	0,153		0,613	0,157		0,606	0,038		0,878	0,000		0,998	0,000	
V2 $dr = 0{,}02$	all-confidence	197	0,072	0,070	86	0,254	0,096	43	0,431	0,102	13	0,599	0,133	3	0,591	0,013	1	0,832	0,000	1	0,997	0,000	344
	cross-support		0,117	0,130		0,316	0,139		0,506	0,160		0,649	0,166		0,606	0,038		0,878	0,000		0,998	0,000	

Continua na próxima página.

V3 $dr = 0{,}00015$	all-confidence cross-support	289	0,071 0,139	0,062 0,124	108	0,251 0,328	0,088 0,130	57	0,427 0,516	0,099 0,153	19	0,562 0,613	0,124 0,157	3	0,591 0,606	0,013 0,038	1	0,832 0,878	0,000 0,000	1	0,997 0,998	0,000 0,000	478
V4 $dr = 0{,}02$	**all-confidence** cross-support	**197**	**0,072** 0,117	**0,070** 0,130	**86**	**0,254** 0,316	**0,096** 0,139	**43**	**0,431** 0,506	**0,102** 0,160	**13**	**0,599** 0,649	**0,133** 0,166	**3**	**0,591** 0,606	**0,013** 0,038	**1**	**0,832** 0,878	**0,000** 0,000	**1**	**0,997** 0,998	**0,000** 0,000	**344**

PowerC	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,13]			(0,13 , 0,27]			(0,27 , 0,40]			(0,40 , 0,53]			(0,53 , 0,66]			(0,66 , 0,80]			(0,80 , 0,93]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0001$	all-confidence cross-support	779	0,023 0,046	0,115 0,151	1	0,934 0,960	0,000 0,000	3	0,358 0,373	0,040 0,050	4	0,792 0,796	0,120 0,121	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	787
V2 $dr = 0{,}02$	all-confidence cross-support	758	0,024 0,047	0,117 0,153	1	0,934 0,960	0,000 0,000	2	0,357 0,375	0,056 0,070	5	0,737 0,739	0,162 0,163	1	0,598 0,609	0,000 0,000	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	767
V3 $dr = 0{,}0001$	all-confidence cross-support	779	0,023 0,046	0,115 0,151	1	0,934 0,960	0,000 0,000	3	0,358 0,373	0,040 0,050	4	0,792 0,796	0,120 0,121	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	787
V4 $dr = 0{,}02$	**all-confidence** cross-support	**758**	**0,024** 0,047	**0,117** 0,153	**1**	**0,934** 0,960	**0,000** 0,000	**2**	**0,357** 0,375	**0,056** 0,070	**5**	**0,737** 0,739	**0,162** 0,163	**1**	**0,598** 0,609	**0,000** 0,000	**0**	**0,000** 0,000	**0,000** 0,000	**0**	**0,000** 0,000	**0,000** 0,000	**767**

Pumsb	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 0,99]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence cross-support	1.135	0,122 0,156	0,240 0,279	69	0,371 0,387	0,175 0,181	33	0,606 0,620	0,180 0,178	31	0,780 0,788	0,166 0,160	4	0,843 0,845	0,154 0,156	5	0,971 0,975	0,030 0,033	4	1,000 1,000	0,000 0,000	1.281
V2 $dr = 0{,}0$	all-confidence cross-support	1.135	0,122 0,156	0,240 0,279	69	0,371 0,387	0,175 0,181	33	0,606 0,620	0,180 0,178	31	0,780 0,788	0,166 0,160	4	0,843 0,845	0,154 0,156	5	0,971 0,975	0,030 0,033	4	1,000 1,000	0,000 0,000	1.281
V3 $dr = 0{,}0$	all-confidence cross-support	1.135	0,122 0,156	0,240 0,279	69	0,371 0,387	0,175 0,181	33	0,606 0,620	0,180 0,178	31	0,780 0,788	0,166 0,160	4	0,843 0,845	0,154 0,156	5	0,971 0,975	0,030 0,033	4	1,000 1,000	0,000 0,000	1.281
V4 $dr = 0{,}0$	**all-confidence** cross-support	**1.135**	**0,122** 0,156	**0,240** 0,279	**69**	**0,371** 0,387	**0,175** 0,181	**33**	**0,606** 0,620	**0,180** 0,178	**31**	**0,780** 0,788	**0,166** 0,160	**4**	**0,843** 0,845	**0,154** 0,156	**5**	**0,971** 0,975	**0,030** 0,033	**4**	**1,000** 1,000	**0,000** 0,000	**1.281**

RecordLink	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence cross-support	220	0,005 0,006	0,043 0,045	1	0,628 0,709	0,000 0,000	15	0,420 0,509	0,050 0,049	13	0,512 0,603	0,056 0,114	14	0,660 0,748	0,069 0,096	10	0,774 0,775	0,006 0,006	4	0,984 0,988	0,009 0,006	277
V2 $dr = 0{,}0$	all-confidence cross-support	220	0,005 0,006	0,043 0,045	1	0,628 0,709	0,000 0,000	15	0,420 0,509	0,050 0,049	13	0,512 0,603	0,056 0,114	14	0,660 0,748	0,069 0,096	10	0,774 0,775	0,006 0,006	4	0,984 0,988	0,009 0,006	277

Continua na próxima página.

V3 $dr = 0{,}0$	all-confidence cross-support	220	0,005 0,006	0,043 0,045	1	0,628 0,709	0,000 0,000	15	0,420 0,509	0,050 0,049	13	0,512 0,603	0,056 0,114	14	0,660 0,748	0,069 0,096	10	0,774 0,775	0,006 0,006	4	0,984 0,988	0,009 0,006	277
V4 $dr = 0{,}0$	**all-confidence** cross-support	**220**	**0,005** 0,006	**0,043** 0,045	**1**	**0,628** 0,709	**0,000** 0,000	**15**	**0,420** 0,509	**0,050** 0,049	**13**	**0,512** 0,603	**0,056** 0,114	**14**	**0,660** 0,748	**0,069** 0,096	**10**	**0,774** 0,775	**0,006** 0,006	**4**	**0,984** 0,988	**0,009** 0,006	**277**

Skin	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,06]			(0,06 , 0,13]			(0,13 , 0,19]			(0,19 , 0,26]			(0,26 , 0,32]			(0,32 , 0,38]			(0,38 , 0,45]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}1$	all-confidence cross-support	4	0,010 0,187	0,006 0,204	3	0,208 0,524	0,008 0,035	5	0,345 0,517	0,203 0,298	6	0,425 0,516	0,182 0,242	3	0,685 0,818	0,138 0,158	0	0,000 0,000	0,000 0,000	1	0,565 0,623	0,000 0,000	22
V2 $dr = 0{,}4$	all-confidence cross-support	2	0,010 0,017	0,010 0,000	3	0,217 0,524	0,021 0,035	6	0,342 0,520	0,182 0,267	6	0,425 0,516	0,182 0,242	3	0,685 0,818	0,138 0,158	0	0,000 0,000	0,000 0,000	1	0,565 0,623	0,000 0,000	21
V3 $dr = 0{,}1$	all-confidence cross-support	4	0,010 0,187	0,006 0,204	3	0,208 0,524	0,008 0,035	5	0,345 0,517	0,203 0,298	6	0,425 0,516	0,182 0,242	3	0,685 0,818	0,138 0,158	0	0,000 0,000	0,000 0,000	1	0,565 0,623	0,000 0,000	22
V4 $dr = 0{,}4$	**all-confidence** cross-support	**2**	**0,010** 0,017	**0,010** 0,000	**3**	**0,217** 0,524	**0,021** 0,035	**6**	**0,342** 0,520	**0,182** 0,267	**6**	**0,425** 0,516	**0,182** 0,242	**3**	**0,685** 0,818	**0,138** 0,158	**0**	**0,000** 0,000	**0,000** 0,000	**1**	**0,565** 0,623	**0,000** 0,000	**21**

Susy	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,85]			(0,85 , 0,99]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence cross-support	2.702	0,009 0,017	0,057 0,082	24	0,213 0,245	0,086 0,132	0	0,000 0,000	0,000 0,000	8	0,573 0,580	0,025 0,027	0	0,000 0,000	0,000 0,000	9	0,855 0,872	0,045 0,046	18	0,946 0,953	0,035 0,034	2.761
V2 $dr = 0{,}0$	all-confidence cross-support	2.702	0,009 0,017	0,057 0,082	24	0,213 0,245	0,086 0,132	0	0,000 0,000	0,000 0,000	8	0,573 0,580	0,025 0,027	0	0,000 0,000	0,000 0,000	9	0,855 0,872	0,045 0,046	18	0,946 0,953	0,035 0,034	2.761
V3 $dr = 0{,}0$	all-confidence cross-support	2.702	0,009 0,017	0,057 0,082	24	0,213 0,245	0,086 0,132	0	0,000 0,000	0,000 0,000	8	0,573 0,580	0,025 0,027	0	0,000 0,000	0,000 0,000	9	0,855 0,872	0,045 0,046	18	0,946 0,953	0,035 0,034	2.761
V4 $dr = 0{,}0$	**all-confidence** cross-support	**2.702**	**0,009** 0,017	**0,057** 0,082	**24**	**0,213** 0,245	**0,086** 0,132	**0**	**0,000** 0,000	**0,000** 0,000	**8**	**0,573** 0,580	**0,025** 0,027	**0**	**0,000** 0,000	**0,000** 0,000	**9**	**0,855** 0,872	**0,045** 0,046	**18**	**0,946** 0,953	**0,035** 0,034	**2.761**

6.5.2.1 Base de Dados Densa

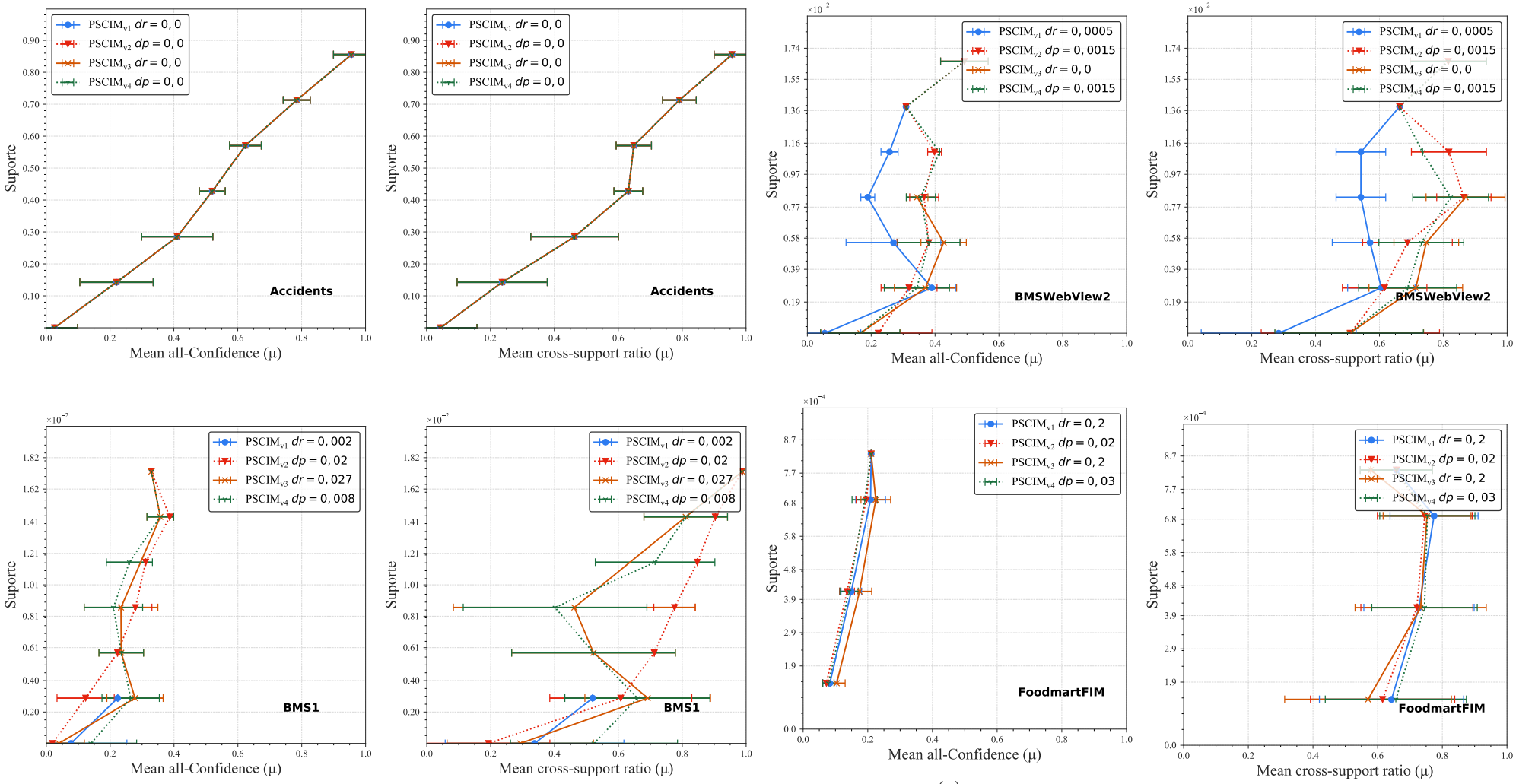
No primeiro experimento, na Figura 6.1, nota-se que a variação PSCIM_{v3} não alterou os itemsets fechados recuperados quando comparados à variação PSCIM_{v1}. Isso porque, nesse tipo de base de dados, não foram encontradas transações com tamanho um. Na base de dados *chess*, *kddcup99*, *recordlink*, *pumbsp*, *skin* e *susy*, é observado um empate na média de *all-confidence* em todas as variações de PSCIM.

Já na base de dados *mushrooms*, as variações PSCIM_{v2} e PSCIM_{v4} tiveram os melhores valores de *all-confidence* em comparação com a variação PSCIM_{v1} e PSCIM_{v3}. Na base de dados *PowerC* houve um empate em *all-confidence* entre as quatro variações do PSCIM.

De forma geral, observa-se que o parâmetro de entrada *dp* das variações PSCIM_{v2} e PSCIM_{v4} é menos sensível na definição de valores. Isso acontece dado o uso do percentil para escolher a distância extra do cluster, diferente de quando é usado a proporção de cobertura extra e cobertura mínima do cluster. Das oito bases de dados densas do estudo, cinco tiveram como melhor valor de parâmetro o, caso automático, ou seja, $dr = 0$ ou $dp = 0$. A base de dados *Skin* foi a única base de dados que teve valores maiores de parâmetro de entrada, de $dr = 0{,}1$ para PSCIM_{v1} e PSCIM_{v3}, e $dp = 0{,}4$ para PSCIM_{v2} e PSCIM_{v4}.

6.5.2.2 Base de Dados Esparsa

Neste cenário, na Figura 6.2 é observado um comportamento diferente nas variações do PSCIM proposto, visto que sete bases de dados esparsas possuem transações com um único item. Nas bases de dados *Accidents* e *PAMP*, todas as variações recuperaram os mesmos resultados de itemsets fechados. Nas bases de dados *fruithut*, *BMSWebview2* e *retail*, a variação PSCIM_{v4} venceu quando comparado o *all-confidence*. Na base de dados *retail*, observa-se que PSCIM_{v2} e PSCIM_{v4} ganharam nos valores de *all-confidence*. Nas bases de dados *fruithut* e *BMSWebview2*, observou-se que PSCIM_{v2} e PSCIM_{v4} foram muito mais representativo nas diferentes faixas de suporte.



(g)

Continua na próxima página.

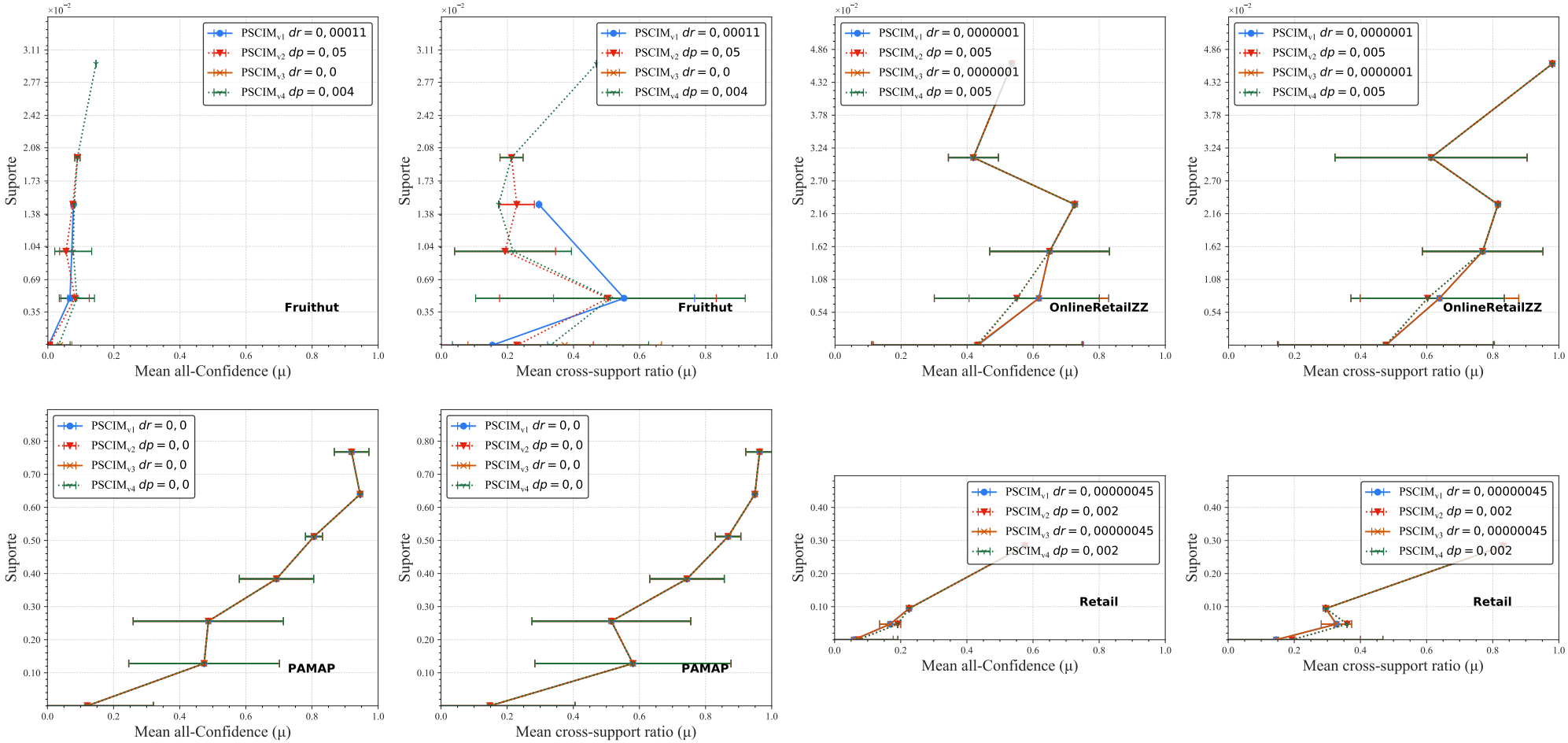


Figura 6.2: Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM$_{v1}$, PSCIM$_{v2}$, PSCIM$_{v3}$ e PSCIM$_{v4}$ sobre as bases de dados esparsas da Tabela 6.3. Neste estudo foi usado o melhor valor de parâmetro para cada variedade de PSCIM.

Tabela 6.3: Desempenho do algoritmo/parametrização para as variações $\text{PLSCIM}_{\text{v1}}$, $\text{PLSCIM}_{\text{v2}}$, $\text{PLSCIM}_{\text{v3}}$, e $\text{PLSCIM}_{\text{v4}}$ sobre as bases de dados densas da Tabela 6.1.

Accidents	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,29]			(0,29 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,86]			(0,86 , 1,00]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	
V2 $dr = 0{,}0$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	
V3 $dr = 0{,}0$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	
V4 $dr = 0{,}0$	**all-confidence**	**2.222**	**0,026**	**0,073**	**104**	**0,221**	**0,115**	**2**	**0,411**	**0,111**	**151**	**0,520**	**0,041**	**77**	**0,624**	**0,050**	**12**	**0,785**	**0,042**	**8**	**0,956**	**0,056**	**2.576**
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	

BMSWebView2 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset #
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,08] †			(0,08 , 0,11] †			(0,11 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,19] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0005$	all-confidence	2.524	0,055	0,113	3	0,390	0,074	4	0,270	0,148	2	0,189	0,022	2	0,257	0,027	1	0,309	0,000	0	0,000	0,000	2.536
	cross-support		0,285	0,243		0,606	0,106		0,571	0,118		0,542	0,078		0,542	0,078		0,663	0,000		0,000	0,000	
V2 $dr = 0{,}0015$	all-confidence	482	0,222	0,168	21	0,319	0,088	9	0,380	0,100	4	0,366	0,045	2	0,398	0,022	1	0,309	0,000	2	0,492	0,075	521
	cross-support		0,509	0,279		0,616	0,132		0,688	0,140		0,864	0,085		0,818	0,118		0,663	0,000		0,816	0,120	
V3 $dr = 0{,}0$	all-confidence	4.119	0,166	0,124	32	0,370	0,097	6	0,427	0,071	2	0,344	0,034	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	4.159
	cross-support		0,507	0,232		0,714	0,147		0,747	0,101		0,870	0,124		0,000	0,000		0,000	0,000		0,000	0,000	
V4 $dr = 0{,}0015$	**all-confidence**	**4.163**	**0,165**	**0,124**	**47**	**0,343**	**0,102**	**12**	**0,380**	**0,097**	**5**	**0,356**	**0,046**	**1**	**0,414**	**0,000**	**1**	**0,309**	**0,000**	**2**	**0,492**	**0,075**	**4.231**
	cross-support		0,505	0,233		0,688	0,153		0,731	0,133		0,823	0,119		0,734	0,000		0,663	0,000		0,816	0,120	

BMS1 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset #
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,09] †			(0,09 , 0,12] †			(0,12 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,20] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}002$	all-confidence	160	0,077	0,175	1	0,224	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	161
	cross-support		0,337	0,280		0,520	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000	
V2 $dr = 0{,}02$	**all-confidence**	**1.670**	**0,019**	**0,061**	**25**	**0,123**	**0,090**	**1**	**0,223**	**0,000**	**2**	**0,279**	**0,051**	**1**	**0,311**	**0,000**	**1**	**0,387**	**0,000**	**1**	**0,329**	**0,000**	**1.701**
	cross-support		0,193	0,192		0,608	0,222		0,713	0,000		0,776	0,065		0,847	0,000		0,903	0,000		0,987	0,000	

Continua na próxima página.

V3 $dr = 0{,}027$	all-confidence	1.954	0,040	0,080	23	0,278	0,088	3	0,235	0,070	2	0,234	0,116	0	0,000	0,000	2	0,357	0,042	1	0,329	0,000	1.985
	cross-support		0,293	0,229		0,692	0,197		0,522	0,256		0,462	0,379		0,000	0,000		0,811	0,131		0,987	0,000	
V4 $dr = 0{,}008$	all-confidence	363	0,140	0,143	24	0,264	0,090	3	0,235	0,070	3	0,210	0,091	2	0,260	0,072	2	0,357	0,042	1	0,329	0,000	398
	cross-support		0,524	0,261		0,659	0,227		0,522	0,256		0,402	0,288		0,715	0,187		0,811	0,131		0,987	0,000	

FoodmartFIM τ	***Métrica***	**Partição de suporte † $\times 10^{-3}$**																					Itemset #
		[0,00 , 0,14] †			(0,14 , 0,28] †			(0,28 , 0,41] †			(0,41 , 0,55] †			(0,55 , 0,69] †			(0,69 , 0,83] †			(0,83 , 0,97] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}2$	all-confidence	0	0,000	0,000	532	0,083	0,022	0	0,000	0,000	248	0,149	0,032	0	0,000	0,000	29	0,209	0,046	2	0,211	0,000	811
	cross-support		0,000	0,000		0,643	0,223		0,000	0,000		0,728	0,171		0,000	0,000		0,774	0,136		0,658	0,112	
V2 $dr = 0{,}02$	all-confidence	0	0,000	0,000	465	0,072	0,012	0	0,000	0,000	230	0,136	0,023	0	0,000	0,000	28	0,195	0,031	2	0,211	0,000	725
	cross-support		0,000	0,000		0,615	0,223		0,000	0,000		0,721	0,173		0,000	0,000		0,746	0,147		0,658	0,112	
V3 $dr = 0{,}2$	**all-confidence**	**0**	**0,000**	**0,000**	**63**	**0,104**	**0,026**	**0**	**0,000**	**0,000**	**51**	**0,175**	**0,037**	**0**	**0,000**	**0,000**	**17**	**0,225**	**0,046**	**1**	**0,211**	**0,000**	**132**
	cross-support		0,000	0,000		0,570	0,258		0,000	0,000		0,733	0,203		0,000	0,000		0,753	0,136		0,579	0,000	
V4 $dr = 0{,}03$	all-confidence	0	0,000	0,000	951	0,078	0,017	0	0,000	0,000	452	0,142	0,028	0	0,000	0,000	50	0,190	0,039	2	0,211	0,000	1.455
	cross-support		0,000	0,000		0,656	0,218		0,000	0,000		0,745	0,163		0,000	0,000		0,754	0,147		0,658	0,112	

Fruithut τ	***Métrica***	**Partição de suporte † $\times 10^{-1}$**																					Itemset #
		[0,00 , 0,05] †			(0,05 , 0,10] †			(0,10 , 0,15] †			(0,15 , 0,20] †			(0,20 , 0,25] †			(0,25 , 0,30] †			(0,30 , 0,35] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}00011$	all-confidence	6.881	0,002	0,010	3	0,067	0,007	0	0,000	0,000	1	0,078	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	6.885
	cross-support		0,154	0,167		0,553	0,214		0,000	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000	
V2 $dr = 0{,}05$	all-confidence	18.504	0,003	0,010	47	0,082	0,043	8	0,055	0,019	4	0,076	0,006	2	0,089	0,008	0	0,000	0,000	0	0,000	0,000	18.565
	cross-support		0,229	0,231		0,504	0,328		0,193	0,153		0,228	0,053		0,212	0,035		0,000	0,000		0,000	0,000	
V3 $dr = 0{,}0$	all-confidence	298	0,037	0,035	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	298
	cross-support		0,373	0,293		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000	
V4 $dr = 0{,}004$	**all-confidence**	**398**	**0,033**	**0,035**	**15**	**0,088**	**0,053**	**4**	**0,077**	**0,056**	**1**	**0,080**	**0,000**	**2**	**0,089**	**0,008**	**0**	**0,000**	**0,000**	**1**	**0,146**	**0,000**	**421**
	cross-support		0,330	0,297		0,512	0,408		0,217	0,177		0,173	0,000		0,212	0,035		0,000	0,000		0,470	0,000	

OnlineRetail τ	***Métrica***	**Partição de suporte † $\times 10^{-1}$**																					Itemset #
		[0,00 , 0,08] †			(0,08 , 0,15] †			(0,15 , 0,23] †			(0,23 , 0,31] †			(0,31 , 0,39] †			(0,39 , 0,46] †			(0,46 , 0,54] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	

Continua na próxima página.

V1 $dr = 0{,}0000001$	all-confidence	1.061	0,433	0,318	5	0,617	0,212	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.073
	cross-support		0,476	0,327		0,639	0,240		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	
V2 $dr = 0{,}005$	all-confidence	1.059	0,434	0,318	6	0,551	0,250	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.072
	cross-support		0,478	0,327		0,603	0,232		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	
V3 $dr = 0{,}0000001$	**all-confidence**	**1.064**	**0,429**	**0,319**	**5**	**0,617**	**0,212**	**3**	**0,650**	**0,181**	**1**	**0,726**	**0,000**	**2**	**0,419**	**0,076**	**0**	**0,000**	**0,000**	**1**	**0,536**	**0,000**	**1.076**
	cross-support		0,475	0,327		0,639	0,240		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	
V4 $dr = 0{,}005$	all-confidence	1.065	0,429	0,319	6	0,551	0,250	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.078
	cross-support		0,475	0,327		0,603	0,232		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	

PAMP τ	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,13]			(0,13 , 0,26]			(0,26 , 0,38]			(0,38 , 0,51]			(0,51 , 0,64]			(0,64 , 0,77]			(0,77 , 0,90]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	
V2 $dr = 0{,}0$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	
V3 $dr = 0{,}0$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	
V4 $dr = 0{,}0$	**all-confidence**	**74**	**0,120**	**0,200**	**6**	**0,474**	**0,228**	**11**	**0,487**	**0,227**	**6**	**0,693**	**0,112**	**5**	**0,806**	**0,025**	**1**	**0,946**	**0,000**	**2**	**0,920**	**0,052**	**105**
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	

Retail τ	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,05]			(0,05 , 0,09]			(0,09 , 0,14]			(0,14 , 0,19]			(0,19 , 0,24]			(0,24 , 0,28]			(0,28 , 0,33]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}00000045$	all-confidence	9.830	0,059	0,119	2	0,168	0,032	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.834
	cross-support		0,145	0,255		0,327	0,046		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	
V2 $dr = 0{,}002$	all-confidence	9.729	0,072	0,119	1	0,191	0,000	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.732
	cross-support		0,193	0,275		0,360	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	
V3 $dr = 0{,}00000045$	all-confidence	9.830	0,059	0,119	2	0,168	0,032	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.834
	cross-support		0,145	0,255		0,327	0,046		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	
V4 $dr = 0{,}002$	**all-confidence**	**9.729**	**0,072**	**0,119**	**1**	**0,191**	**0,000**	**1**	**0,225**	**0,000**	**0**	**0,000**	**0,000**	**0**	**0,000**	**0,000**	**0**	**0,000**	**0,000**	**1**	**0,575**	**0,000**	**9.732**
	cross-support		0,193	0,275		0,360	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	

A variação PSCIM_{v3} venceu para as bases de dados *OnlineRetail* e *foodmartFIM*. Observa-se que a variação apresentou os melhores valores de *all-confidence*. O PSCIM_{v2} venceu na base de dados *BMS1*, onde foram observados os melhores valores de *all-confidence*.

6.5.2.3 Discussão

Algumas observações podem ser feitas considerando os dois tipos de bases de dados. Em geral, na maioria dos resultados é notado que as variações propostas não afetam o comportamento original do algoritmo SCIM no cenário de base de dados densa. Houve apenas dois casos nas bases de dados *mushrooms* e *POWERC* onde o comportamento de seleção de itemsets fechados sofreram alterações. No entanto, no cenário de base de dados densa, a variação PSCIM_{v4} ganhou em todas as bases de dados.

Por outro lado, no cenário de base de dados esparsa as variações propostas apresentaram uma melhoria mais expressiva na qualidade dos itemsets fechados recuperados. A variação PSCIM_{v4} venceu na maioria das bases de dados. Mais especificamente, em seis casos, observa-se melhoras nos valores de *all-confidence*. Por outro lado, houve um caso onde o PSCIM_{v2} mostrou melhores resultados apenas de *all-confidence* e houve dois casos onde o PSCIM_{v3} mostrou uma melhoria no *all-confidence* e na quantidade de itemsets fechados recuperados.

Esses resultados observados reforçam a ideia de que combinar as duas alterações propostas, ou seja, PSCIM_{v4}, proporciona um impacto positivo na qualidade dos resultados, principalmente para o caso de base de dados esparsa, onde a variação mostrou resultados de itemsets mais representativos nas faixas de suporte. Logo, o PSCIM apresentados nos próximos estudos se refere à variação PSCIM_{v4}.

6.5.3 Discussão Sobre Qualidade e Tempo de Execução

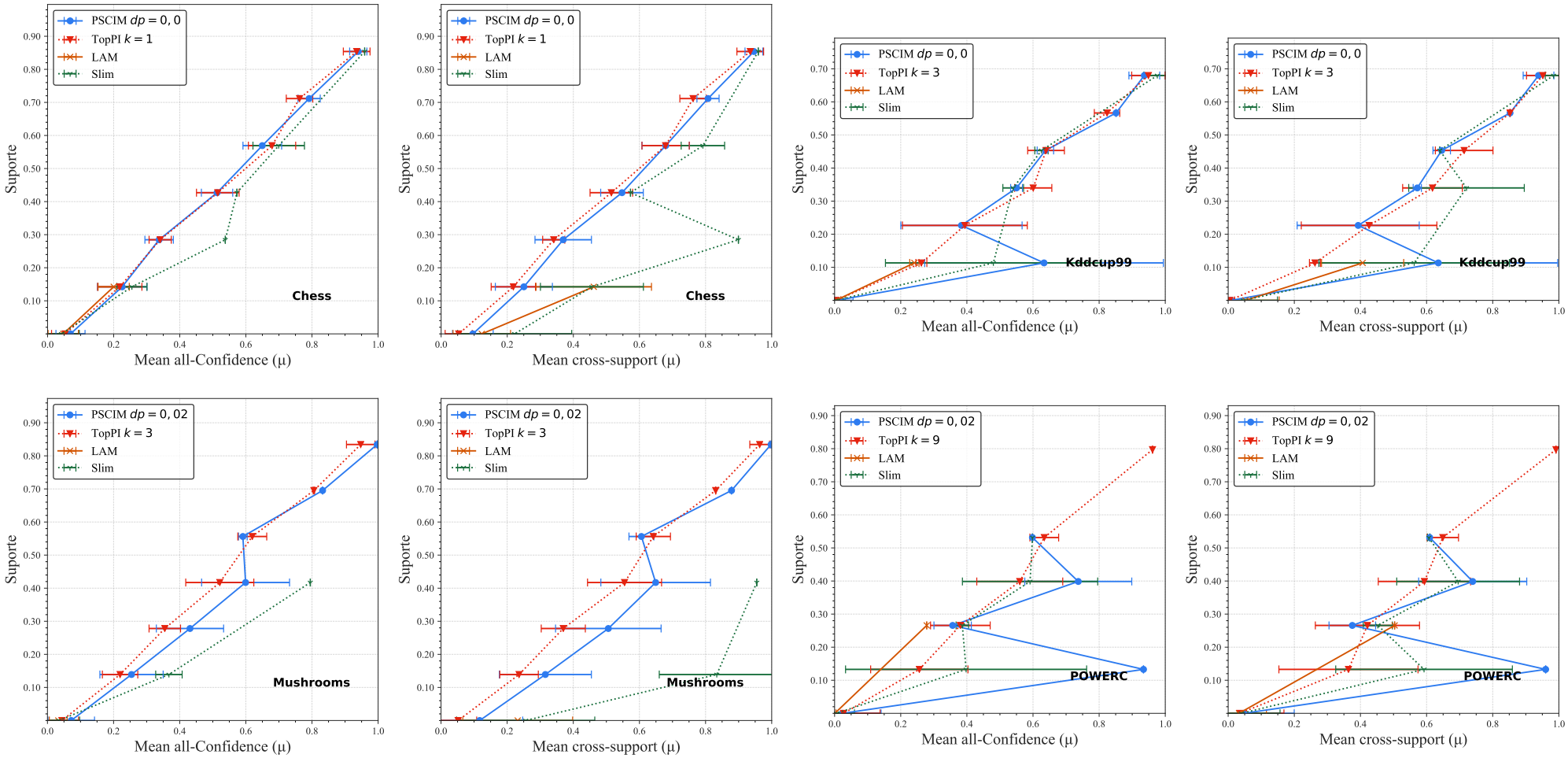
Neste estudo são comparadas as distribuições de *all-confidence* dos itemsets fechados recuperados pelas técnicas PSCIM, TopPI, Slim e LAM. Nos algoritmos TopPI, Slim e LAM, é necessário definir o suporte mínimo igual a 1 para fins de comparação, já que o algoritmo PSCIM pode retornar itemsets fechados com qualquer suporte. Portanto, é importante observar o comportamento de seleção de itemsets fechados em todas as faixas de suporte da base de dados. Os experimentos realizados tiveram como objetivo medir a capacidade de cada técnica em identificar itemsets fechados em todas as faixas de suporte,

além de observar os casos de itemsets raros, ou seja, com baixos valores de suporte. As bases de dados foram separadas em dois grupos: bases de dados densas e base de dados esparsas. Essa decisão ajuda a entender melhor o comportamento do algoritmo para esses dois cenários de tipo de base de dados.

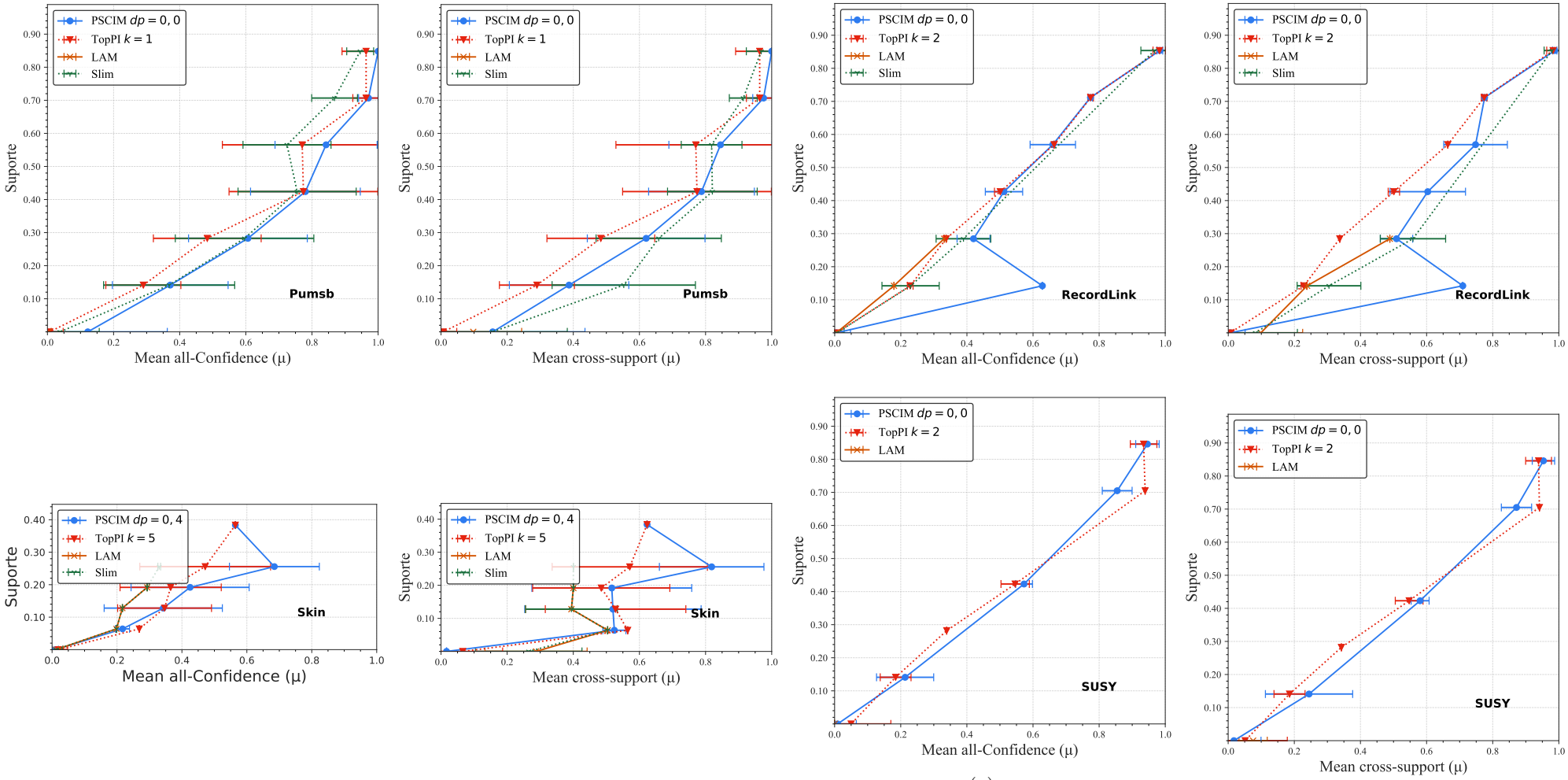
Neste estudo foram comparadas as distribuições de medidas das médias de *all-confidence* e *cross-support* com os algoritmos TopPI, PSCIM, LAM e Slim em cada base de dados. Devido à quantidade expressiva de relatório gerado neste estudo, não é mostrada a análise de dados da escolha dos parâmetros para PSCIM e TopPI, vale lembrar ser usado apenas a métrica *all-confidence* para a escolha do melhor parâmetro. Para obter mais detalhes, consulte o Apêndice C. As Figuras 6.3 e 6.4 mostram que os eixos horizontais correspondem a μ de *all-confidence* ou *cross-support*, variando de 0 a 1, enquanto os eixos verticais distribuem os valores de suporte de 0 até o limite superior de cada base de dados. Espera-se que quanto melhor os itemsets fechados recuperados, mais à direita estará a curva que representa o desempenho de uma técnica/parametrização.

As Tabelas 6.4 e 6.5 mostram na primeira coluna o nome da base de dados, seguido do parâmetro escolhido por cada algoritmo. O algoritmo LAM e Slim não tem parâmetros. A primeira coluna contém a informação da base de estudo com as informações de parâmetros escolhidos para cada algoritmo do estudo. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. As colunas três a nove mostram o número de conjuntos de itens fechados recuperados (#), os valores médios da métrica corrente (μ) e os valores de desvio padrão (σ) calculados por intervalo de suporte por todas as técnicas comparadas. Esses intervalos de suporte têm o mesmo tamanho por base de dados e estes tamanhos foram definidos sobre o intervalo de suporte da base de dados. As últimas duas colunas mostram o número total de itemsets fechados recuperados (#) e os tempos médios de processamento (em segundos) usando a média de tempo obtida em 10 execuções. As implementações de LAM e Slim têm um parâmetro de entrada denominado suporte conjuntivo mínimo (Definição 3), e para o estudo foi definido o valor 1. Nesse cenário, os algoritmos se tornam livres de parâmetros sem nenhuma restrição de suporte para os itemsets fechados recuperados.

Na Tabela 6.5, todos os nomes da base de dados, exceto a base *Accidents*, possuem, na primeira coluna, o símbolo τ na frente do nome. Este símbolo significa que a base de dados possui transações com tamanho 1. O símbolo † pode aparecer nas colunas 1 a 7 em bases de dados onde o intervalo de suporte é inferior a 0,10. Nestes casos, usa-se notação científica para representar a faixa de valores de suporte.



Continua na próxima página.



(g)

Figura 6.3: Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM, TopPI, Slim e LAM sobre as bases de dados densas da Tabela 6.4. Neste estudo foi usado o melhor valor de parâmetro para cada técnica.

Tabela 6.4: Desempenho do algoritmo/parametrização para os algoritmos PLSCIM, TopPI, LAM e Slim sobre as bases de dados densas da Tabela 6.1.

Chess	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225	0,70
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027		
TopPI $k = 1$	all-confidence	18	0,053	0,041	7	0,218	0,067	9	0,341	0,033	5	0,515	0,064	4	0,679	0,072	2	0,762	0,040	3	0,935	0,040	48	0,49
	cross-support		0,053	0,041		0,218	0,067		0,341	0,033		0,515	0,064		0,679	0,072		0,762	0,040		0,935	0,040		
LAM	all-confidence	54	0,047	0,046	22	0,200	0,047	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	76	1,38
	cross-support		0,123	0,087		0,462	0,175		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	224	0,042	0,054	10	0,255	0,046	1	0,538	0,000	1	0,573	0,000	4	0,699	0,078	0	0,000	0,000	1	0,959	0,000	241	0,64
	cross-support		0,222	0,173		0,456	0,156		0,899	0,000		0,573	0,000		0,792	0,066		0,000	0,000		0,959	0,000		

Kddcup99	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,11]			(0,11 , 0,23]			(0,23 , 0,34]			(0,34 , 0,45]			(0,45 , 0,57]			(0,57 , 0,68]			(0,68 , 0,79]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763	5,35
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046		
TopPI $k = 3$	all-confidence	287	0,005	0,015	41	0,262	0,017	9	0,394	0,189	4	0,601	0,056	7	0,639	0,055	2	0,824	0,038	17	0,949	0,050	367	2,73
	cross-support		0,005	0,015		0,263	0,017		0,426	0,205		0,618	0,090		0,714	0,087		0,852	0,000		0,952	0,049		
LAM	all-confidence	454	0,004	0,015	17	0,241	0,014	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	471	25.453,68
	cross-support		0,052	0,103		0,407	0,124		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	821	0,004	0,028	18	0,480	0,326	0	0,000	0,000	4	0,540	0,031	2	0,626	0,020	0	0,000	0,000	5	0,976	0,041	850	29,63
	cross-support		0,046	0,103		0,563	0,288		0,000	0,000		0,721	0,175		0,640	0,000		0,000	0,000		0,984	0,025		

Mushrooms	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,83]			(0,83 , 0,97]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 20{,}00$	all-confidence	197	0,072	0,070	86	0,254	0,096	43	0,431	0,102	13	0,599	0,133	3	0,591	0,013	1	0,832	0,000	1	0,997	0,000	344	0,75
	cross-support		0,117	0,130		0,316	0,139		0,506	0,160		0,649	0,166		0,606	0,038		0,878	0,000		0,998	0,000		
TopPI $k = 3$	all-confidence	190	0,044	0,050	52	0,220	0,054	26	0,355	0,048	21	0,521	0,103	9	0,619	0,044	2	0,806	0,002	3	0,947	0,043	303	0,37
	cross-support		0,052	0,056		0,236	0,058		0,370	0,067		0,555	0,112		0,642	0,052		0,831	0,000		0,964	0,030		

Continua na próxima página.

LAM	all-confidence	86	0,033	0,027	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	86	1,39
	cross-support		0,232	0,166		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	412	0,035	0,063	3	0,367	0,040	0	0,000	0,000	1	0,795	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	416	1,85
	cross-support		0,256	0,209		0,834	0,174		0,000	0,000		0,955	0,000		0,000	0,000		0,000	0,000		0,000	0,000		

PowerC	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,13]			(0,13 , 0,27]			(0,27 , 0,40]			(0,40 , 0,53]			(0,53 , 0,66]			(0,66 , 0,80]			(0,80 , 0,93]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 2{,}00$	all-confidence	758	0,024	0,117	1	0,934	0,000	2	0,357	0,056	5	0,737	0,162	1	0,598	0,000	0	0,000	0,000	0	0,000	0,000	767	3,45
	cross-support		0,047	0,153		0,960	0,000		0,375	0,070		0,739	0,163		0,609	0,000		0,000	0,000		0,000	0,000		
TopPI $k = 9$	all-confidence	832	0,027	0,110	28	0,256	0,147	10	0,380	0,090	19	0,560	0,130	6	0,634	0,043	0	0,000	0,000	1	0,961	0,000	896	2,25
	cross-support		0,035	0,133		0,364	0,210		0,421	0,158		0,593	0,140		0,649	0,047		0,000	0,000		0,991	0,000		
LAM	all-confidence	1.964	0,001	0,004	0	0,000	0,000	1	0,279	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1.965	4.517,61
	cross-support		0,033	0,083		0,000	0,000		0,503	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	3.710	0,006	0,054	4	0,398	0,365	3	0,386	0,019	6	0,591	0,205	2	0,598	0,000	0	0,000	0,000	0	0,000	0,000	3.725	813,39
	cross-support		0,039	0,117		0,592	0,267		0,453	0,044		0,695	0,186		0,607	0,004		0,000	0,000		0,000	0,000		

Pumsb	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 0,99]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	1.135	0,122	0,240	69	0,371	0,175	33	0,606	0,180	31	0,780	0,166	4	0,843	0,154	5	0,971	0,030	4	1,000	0,000	1.281	190,07
	cross-support		0,156	0,279		0,387	0,181		0,620	0,178		0,788	0,160		0,845	0,156		0,975	0,033		1,000	0,000		
TopPI $k = 1$	all-confidence	1.855	0,007	0,041	18	0,290	0,113	7	0,483	0,163	12	0,774	0,225	2	0,770	0,241	3	0,964	0,040	4	0,964	0,073	1.901	1,20
	cross-support		0,007	0,041		0,290	0,113		0,483	0,163		0,774	0,225		0,770	0,241		0,964	0,040		0,964	0,073		
LAM	all-confidence	2.322	0,004	0,007	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2.322	91,15
	cross-support		0,098	0,146		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	6.206	0,038	0,119	49	0,368	0,198	21	0,596	0,209	18	0,755	0,179	9	0,724	0,133	14	0,868	0,069	8	0,946	0,041	6.325	12.073,94
	cross-support		0,156	0,226		0,553	0,217		0,658	0,189		0,821	0,136		0,818	0,092		0,914	0,042		0,968	0,045		

RecordLink	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	220	0,005	0,043	1	0,628	0,000	15	0,420	0,050	13	0,512	0,056	14	0,660	0,069	10	0,774	0,006	4	0,984	0,009	277	2,35
	cross-support		0,006	0,045		0,709	0,000		0,509	0,049		0,603	0,114		0,748	0,096		0,775	0,006		0,988	0,006		

Continua na próxima página.

TopPI $k = 2$	all-confidence	16	0,006	0,012	3	0,229	0,009	1	0,337	0,000	2	0,501	0,017	1	0,664	0,000	3	0,774	0,009	4	0,982	0,019	30	0,95
	cross-support		0,006	0,012		0,229	0,009		0,337	0,000		0,501	0,017		0,664	0,000		0,774	0,009		0,983	0,019		
LAM	all-confidence	179	0,005	0,017	2	0,179	0,000	2	0,331	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	183	3.309,18
	cross-support		0,098	0,127		0,238	0,000		0,489	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	217	0,006	0,022	5	0,230	0,087	2	0,390	0,083	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2	0,967	0,041	226	3,34
	cross-support		0,082	0,127		0,305	0,096		0,559	0,099		0,000	0,000		0,000	0,000		0,000	0,000		0,980	0,024		

Skin	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,06]			(0,06 , 0,13]			(0,13 , 0,19]			(0,19 , 0,26]			(0,26 , 0,32]			(0,32 , 0,38]			(0,38 , 0,45]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}40$	all-confidence	2	0,010	0,010	3	0,217	0,021	6	0,342	0,182	6	0,425	0,182	3	0,685	0,138	0	0,000	0,000	1	0,565	0,000	21	0,56
	cross-support		0,017	0,000		0,524	0,035		0,520	0,267		0,516	0,242		0,818	0,158		0,000	0,000		0,623	0,000		
TopPI $k = 5$	all-confidence	6	0,019	0,028	1	0,269	0,000	9	0,347	0,145	9	0,365	0,156	7	0,472	0,201	0	0,000	0,000	1	0,565	0,000	33	0,79
	cross-support		0,066	0,120		0,565	0,000		0,528	0,212		0,485	0,207		0,571	0,235		0,000	0,000		0,623	0,000		
LAM	all-confidence	53	0,012	0,022	1	0,198	0,000	2	0,217	0,000	1	0,293	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	57	36,86
	cross-support		0,287	0,154		0,504	0,000		0,394	0,139		0,400	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	58	0,011	0,021	1	0,198	0,000	2	0,217	0,000	1	0,293	0,000	2	0,330	0,004	0	0,000	0,000	0	0,000	0,000	64	0,42
	cross-support		0,265	0,161		0,504	0,000		0,394	0,139		0,400	0,000		0,400	0,000		0,000	0,000		0,000	0,000		

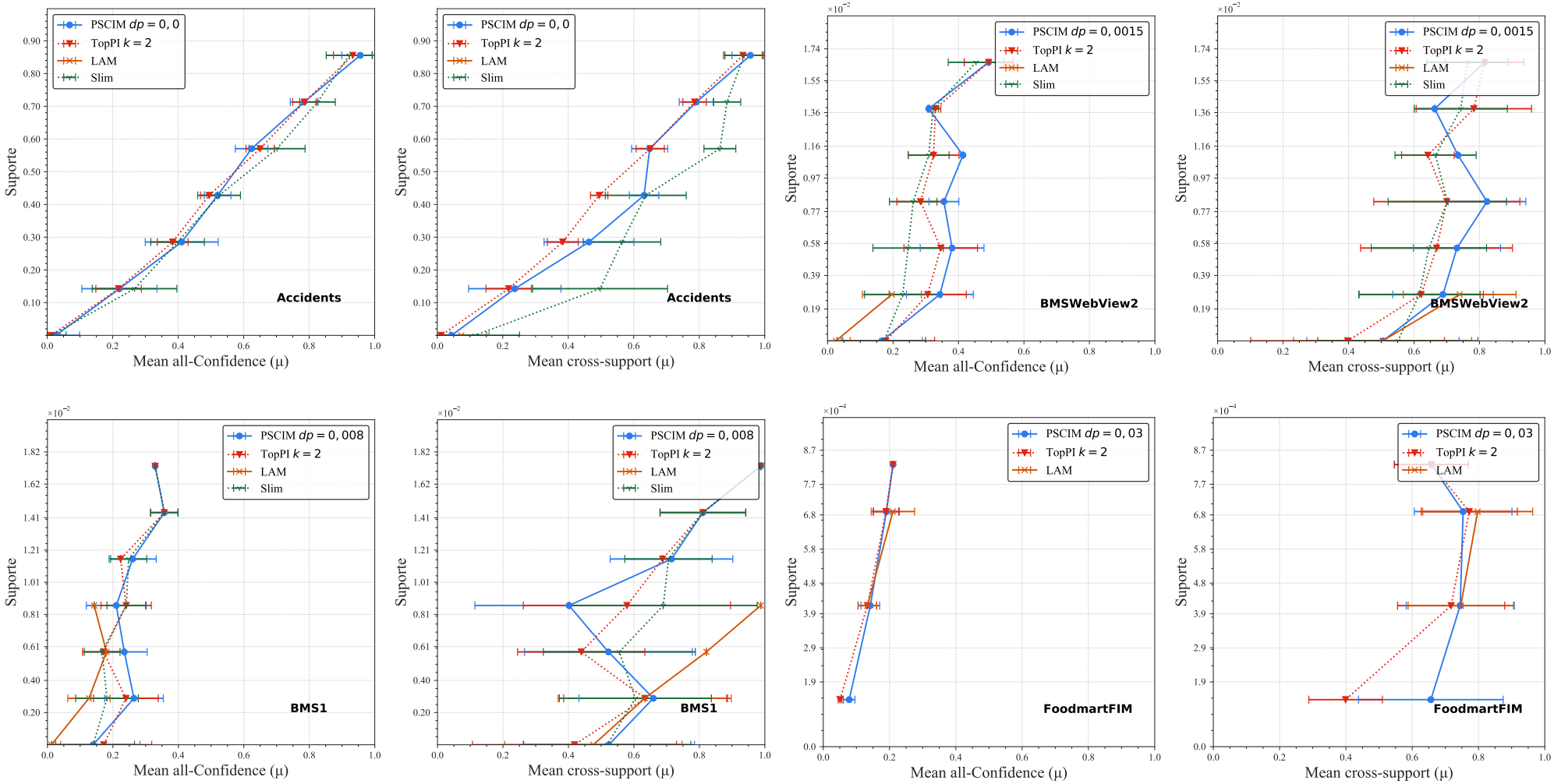
Susy	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,85]			(0,85 , 0,99]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	2.702	0,009	0,057	24	0,213	0,086	0	0,000	0,000	8	0,573	0,025	0	0,000	0,000	9	0,855	0,045	18	0,946	0,035	2.761	252,10
	cross-support		0,017	0,082		0,245	0,132		0,000	0,000		0,580	0,027		0,000	0,000		0,872	0,046		0,953	0,034		
TopPI $k = 2$	all-confidence	190	0,050	0,120	9	0,185	0,047	1	0,339	0,000	2	0,546	0,043	0	0,000	0,000	1	0,939	0,000	8	0,934	0,040	211	12,03
	cross-support		0,051	0,127		0,186	0,047		0,342	0,000		0,547	0,042		0,000	0,000		0,941	0,000		0,939	0,039		
LAM	all-confidence	41.532	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	41.532	986,11
	cross-support		0,075	0,043		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		

A Figura 6.4 ilustra as distribuições de μ das bases de dados esparsas. Se as faixas de suporte forem inferiores a 0,10, a proporção não é mantida, pois, nesses casos, o eixo y é muito menor que o eixo x, dificultando a leitura da figura.

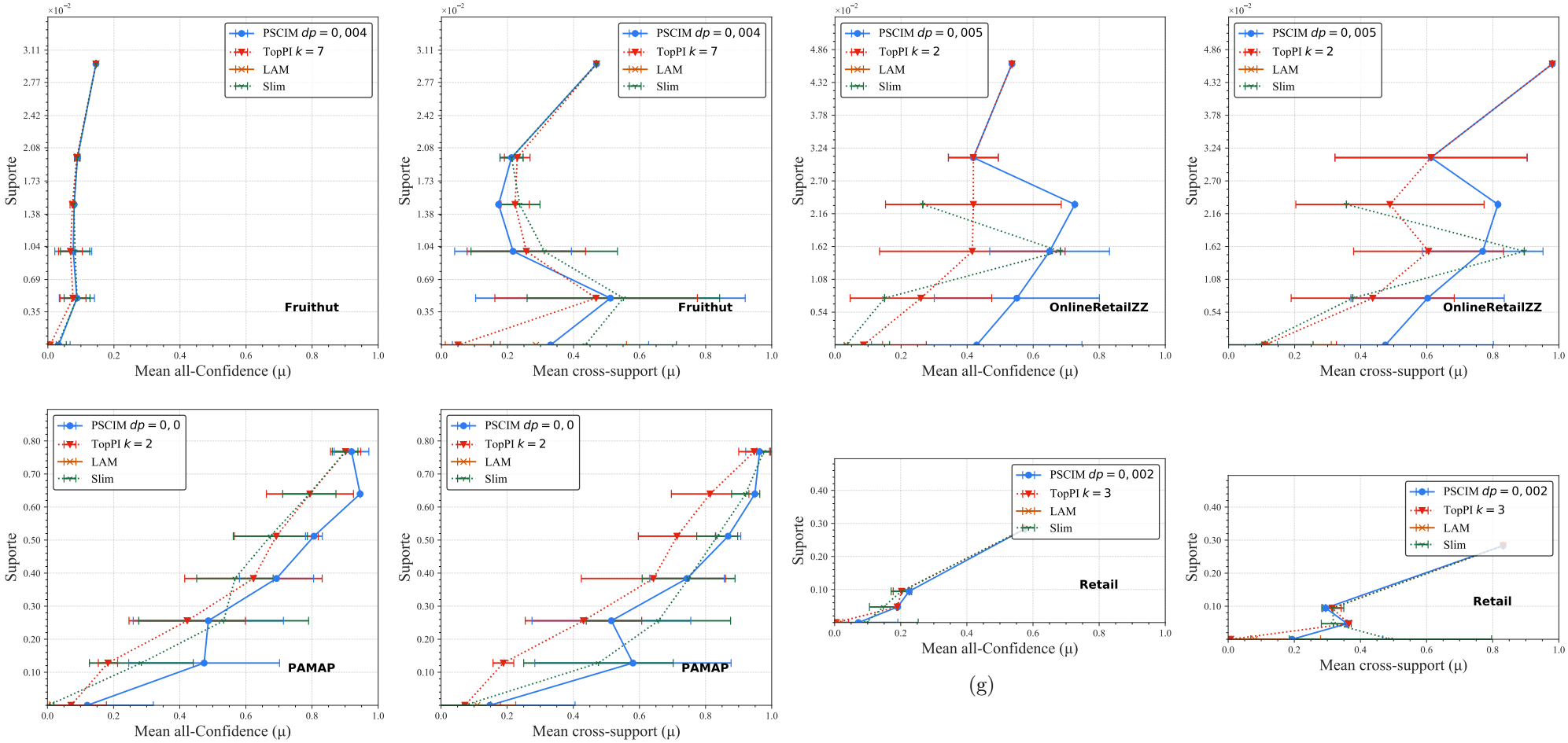
O algoritmo Slim não funcionou nas bases de dados: *Susy* e *Fruithut*. As execuções que levaram mais de um dia para serem finalizadas foram abortadas.

Todos os algoritmos mencionados são executados no modo *multi-threading*. Neste estudo, foram definidas 8 *threads* para todas as execuções, visto que este estudo visa compreender o desempenho de uma técnica/parametrização e não o comportamento do modo *multi-threading*.

A Figura 6.3 contém os resultados das distribuições de *all-confidence* para as bases de dados densas. Esse tipo de análise não é nova, pois o algoritmo SCIM já testa para esse tipo de bases de dados. Porém, é fundamental verificar se o algoritmo manterá o comportamento de seleção mesmo com as duas mudanças propostas pelo algoritmo PSCIM. Na primeira observação, o algoritmo PSCIM vence nas bases de dados *Pumsb*, *Recordlink*, *Skin*, *Kddcup99* e *mushrooms*. Nas bases de dados *Skin* e *Pumsb*, o ganho de desempenho foi mais evidente quando comparado às outras técnicas. O algoritmo TopPI apresentou os melhores valores de tempo, perdendo apenas para Slim na base de dados *Skin*. Neste cenário, os valores de tempo são: PSCIM com 0,56 segundos, TopPI com 0,79 segundos, LAM com 36,86 segundos e Slim com 0,42 segundos. Na base de dados *Pumsb*, Slim teve o pior tempo de 12.073,92 segundos. Nas bases de dados *Recordlink*, *Kddcup99* e *Skin*, o algoritmo LAM teve os piores resultados de tempo, desviando-se muito dos valores de tempo de outras técnicas.



Continua na próxima página.



(g)

Figura 6.4: Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados por PSCIM, TopPI, Slim e LAM sobre as bases de dados esparsa da Tabela 6.5. Neste estudo foi usado o melhor valor de parâmetro para cada técnica.

Tabela 6.5: Desempenho do algoritmo/parametrização para os algoritmos PLSCIM, TopPI, LAM e Slim sobre as bases de dados esparsas da Tabela 6.1.

Accidents	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,29]			(0,29 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,86]			(0,86 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576	111,12
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056		
TopPI $k = 2$	all-confidence	634	0,011	0,027	33	0,218	0,069	4	0,383	0,047	6	0,494	0,026	7	0,650	0,043	7	0,786	0,036	7	0,933	0,059	698	1,85
	cross-support		0,012	0,027		0,218	0,069		0,383	0,047		0,494	0,026		0,650	0,043		0,786	0,036		0,933	0,059		
LAM	all-confidence	9.662	0,001	0,001	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	9.662	59,63
	cross-support		0,073	0,078		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	13.013	0,012	0,045	86	0,267	0,129	41	0,398	0,082	10	0,525	0,065	14	0,702	0,086	8	0,825	0,054	3	0,923	0,071	13.175	128.565,67
	cross-support		0,115	0,136		0,497	0,206		0,564	0,118		0,637	0,124		0,863	0,049		0,885	0,042		0,937	0,059		

BMSWebView2 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset #	Tempo (s)
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,08] †			(0,08 , 0,11] †			(0,11 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,19] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}15$	all-confidence	4.163	0,165	0,124	47	0,343	0,102	12	0,380	0,097	5	0,356	0,046	1	0,414	0,000	1	0,309	0,000	2	0,492	0,075	4.231	241,63
	cross-support		0,505	0,233		0,688	0,153		0,731	0,133		0,823	0,119		0,734	0,000		0,663	0,000		0,816	0,120		
TopPI $k = 2$	all-confidence	2.778	0,178	0,149	50	0,306	0,118	17	0,346	0,112	4	0,284	0,072	3	0,324	0,078	2	0,330	0,017	2	0,492	0,075	2.856	0,79
	cross-support		0,397	0,296		0,622	0,190		0,669	0,232		0,700	0,223		0,642	0,080		0,783	0,176		0,816	0,120		
LAM	all-confidence	6.203	0,025	0,044	10	0,196	0,090	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	6.213	2,28
	cross-support		0,504	0,272		0,740	0,172		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	4.814	0,173	0,126	235	0,230	0,117	66	0,246	0,107	26	0,262	0,073	12	0,309	0,063	3	0,323	0,017	3	0,454	0,085	5.159	9.119,22
	cross-support		0,551	0,245		0,616	0,185		0,645	0,176		0,702	0,181		0,666	0,124		0,743	0,142		0,763	0,124		

BMS1 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset #	Tempo (s)
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,09] †			(0,09 , 0,12] †			(0,12 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,20] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}01$	all-confidence	363	0,140	0,143	24	0,264	0,090	3	0,235	0,070	3	0,210	0,091	2	0,260	0,072	2	0,357	0,042	1	0,329	0,000	398	3,34
	cross-support		0,524	0,261		0,659	0,227		0,522	0,256		0,402	0,288		0,715	0,187		0,811	0,131		0,987	0,000		
TopPI $k = 2$	all-confidence	356	0,172	0,147	39	0,240	0,099	9	0,171	0,064	5	0,240	0,077	1	0,224	0,000	2	0,357	0,042	1	0,329	0,000	413	0,49
	cross-support		0,418	0,312		0,634	0,249		0,439	0,194		0,579	0,317		0,687	0,000		0,811	0,131		0,987	0,000		

Continua na próxima página.

LAM	all-confidence	2.685	0,012	0,028	8	0,127	0,065	1	0,181	0,000	1	0,143	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2.695	1,77
	cross-support		0,476	0,271		0,633	0,264		0,822	0,000		0,987	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	836	0,140	0,127	86	0,182	0,096	17	0,167	0,055	8	0,241	0,059	3	0,248	0,055	2	0,357	0,042	1	0,329	0,000	953	45,57
	cross-support		0,519	0,255		0,605	0,232		0,556	0,232		0,690	0,287		0,706	0,133		0,811	0,131		0,987	0,000		

FoodmartFIM τ	***Métrica***	**Partição de suporte † $\times 10^{-3}$**																					Itemset #	Tempo (s)
		[0,00 , 0,14] †			(0,14 , 0,28] †			(0,28 , 0,41] †			(0,41 , 0,55] †			(0,55 , 0,69] †			(0,69 , 0,83] †			(0,83 , 0,97] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}03$	all-confidence	0	0,000	0,000	951	0,078	0,017	0	0,000	0,000	452	0,142	0,028	0	0,000	0,000	50	0,190	0,039	2	0,211	0,000	1.455	12,53
	cross-support		0,000	0,000		0,656	0,218		0,000	0,000		0,745	0,163		0,000	0,000		0,754	0,147		0,658	0,112		
TopPI $k = 2$	all-confidence	0	0,000	0,000	247	0,051	0,006	0	0,000	0,000	709	0,133	0,028	0	0,000	0,000	69	0,189	0,038	2	0,211	0,000	1.027	0,51
	cross-support		0,000	0,000		0,399	0,111		0,000	0,000		0,717	0,162		0,000	0,000		0,772	0,145		0,658	0,112		
LAM	all-confidence	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	266	0,133	0,028	0	0,000	0,000	6	0,210	0,065	0	0,000	0,000	272	0,58
	cross-support		0,000	0,000		0,000	0,000		0,000	0,000		0,747	0,159		0,000	0,000		0,798	0,166		0,000	0,000		

Fruithut τ	***Métrica***	**Partição de suporte † $\times 10^{-1}$**																					Itemset #	Tempo (s)
		[0,00 , 0,05] †			(0,05 , 0,10] †			(0,10 , 0,15] †			(0,15 , 0,20] †			(0,20 , 0,25] †			(0,25 , 0,30] †			(0,30 , 0,35] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	398	0,033	0,035	15	0,088	0,053	4	0,077	0,056	1	0,080	0,000	2	0,089	0,008	0	0,000	0,000	1	0,146	0,000	421	22,08
	cross-support		0,330	0,297		0,512	0,408		0,217	0,177		0,173	0,000		0,212	0,035		0,000	0,000		0,470	0,000		
TopPI $k = 7$	all-confidence	6.957	0,007	0,018	99	0,076	0,039	23	0,068	0,036	6	0,074	0,007	3	0,087	0,007	0	0,000	0,000	1	0,146	0,000	7.089	0,83
	cross-support		0,050	0,128		0,468	0,307		0,257	0,179		0,223	0,043		0,230	0,039		0,000	0,000		0,470	0,000		
LAM	all-confidence	11.171	0,002	0,005	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	11.171	3,50
	cross-support		0,286	0,274		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	2.032	0,030	0,026	67	0,088	0,039	12	0,082	0,044	3	0,079	0,001	2	0,089	0,008	0	0,000	0,000	1	0,146	0,000	2.117	658,36
	cross-support		0,435	0,277		0,551	0,292		0,311	0,222		0,237	0,061		0,212	0,035		0,000	0,000		0,470	0,000		

OnlineRetail τ	***Métrica***	**Partição de suporte † $\times 10^{-1}$**																					Itemset #	Tempo (s)
		[0,00 , 0,08] †			(0,08 , 0,15] †			(0,15 , 0,23] †			(0,23 , 0,31] †			(0,31 , 0,39] †			(0,39 , 0,46] †			(0,46 , 0,54] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}01$	all-confidence	1.065	0,429	0,319	6	0,551	0,250	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.078	15,69
	cross-support		0,475	0,327		0,603	0,232		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000		
TopPI $k = 2$	all-confidence	2.335	0,087	0,190	24	0,260	0,214	6	0,416	0,281	3	0,419	0,266	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	2.371	1,08
	cross-support		0,109	0,217		0,436	0,247		0,605	0,227		0,489	0,285		0,613	0,291		0,000	0,000		0,980	0,000		
LAM	all-confidence	3.019	0,035	0,109	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	3.019	8,13

Continua na próxima página.

	cross-support		0,119	0,192		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000			
Slim	all-confidence	3.962	0,033	0,132	1	0,150	0,000	1	0,682	0,000	1	0,266	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	3.965	8.484,03
	cross-support		0,090	0,166		0,376	0,000		0,895	0,000		0,357	0,000		0,000	0,000		0,000	0,000		0,000	0,000		

PAMP τ	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,13]			(0,13 , 0,26]			(0,26 , 0,38]			(0,38 , 0,51]			(0,51 , 0,64]			(0,64 , 0,77]			(0,77 , 0,90]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105	8,99
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042		
TopPI $k = 2$	all-confidence	49	0,071	0,107	4	0,183	0,029	9	0,423	0,176	3	0,623	0,208	5	0,692	0,127	3	0,794	0,132	11	0,902	0,046	84	2,84
	cross-support		0,072	0,107		0,189	0,031		0,431	0,176		0,642	0,218		0,714	0,117		0,814	0,117		0,947	0,047		
LAM	all-confidence	3.610	0,002	0,004	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	3.610	2.333,96
	cross-support		0,109	0,116		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	8.717	0,003	0,017	25	0,284	0,157	8	0,533	0,257	7	0,567	0,116	11	0,674	0,112	9	0,792	0,081	8	0,900	0,039	8.785	14.237,68
	cross-support		0,071	0,079		0,476	0,226		0,657	0,218		0,749	0,140		0,835	0,062		0,921	0,043		0,976	0,021		

Retail τ	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,05]			(0,05 , 0,09]			(0,09 , 0,14]			(0,14 , 0,19]			(0,19 , 0,24]			(0,24 , 0,28]			(0,28 , 0,33]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	9.729	0,072	0,119	1	0,191	0,000	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.732	8.353,96
	cross-support		0,193	0,275		0,360	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000		
TopPI $k = 3$	all-confidence	30.563	0,005	0,038	2	0,190	0,002	4	0,203	0,025	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	30.570	2,49
	cross-support		0,007	0,061		0,365	0,007		0,314	0,028		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000		
LAM	all-confidence	7.960	0,003	0,021	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	7.960	3,13
	cross-support		0,080	0,199		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	6.557	0,096	0,157	3	0,148	0,042	3	0,202	0,031	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	6.564	103.297,54
	cross-support		0,491	0,306		0,318	0,036		0,316	0,033		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000		

Ainda na Figura 6.3 é apresentado o segundo caso onde o Slim ganhou o PSCIM na distribuição de valores de *all-confidence* na base de dados *Chess*. Nas bases de dados *Susy* e *PowerC*, o terceiro caso onde o algoritmo TopPI ganhou PSCIM nos valores de distribuição de *all-confidence*. No entanto, em alguns intervalos de all-confidence, o algoritmo proposto teve um resultado superior.

Como mencionado anteriormente, o algoritmo SCIM apresentou apenas uma análise em bases de dados densas. Nesta seção, é mostrado também o comportamento da seleção de itemsets fechados em bases de dados esparsas, dado um algoritmo que utiliza contextualização espacial no processo de mineração. Conforme mostra Figura 6.4, o algoritmo PSCIM venceu em todas as bases de dados quando comparado à distribuição de *all-confidence*. Em algumas bases de dados, como *PAMAP*, *OnlineRetaillZZ* e *Retail*, o algoritmo PSCIM teve um valor de *all-confidence* muito mais alto para os valores de intervalo de suporte mais baixos. Essas faixas de suporte baixas indicam que a técnica conseguiu retornar itemsets fechados raros. No entanto, o TopPI ganhou o PSCIM nas bases de dados *BMSWebview2*, *BMS1*, *foodmart* e *OnlineRetailZZ*. O Slim não funcionou nas bases de dados *Retail*, *FoodmartFIM* e *Accidents*.

No PSCIM definimos o mesmo teste estatístico para validar a significância estatística das médias apresentadas no estudo, foi definido o teste *t-student* [66]. No nosso estudo cada algoritmo é considerado uma variável independente categórica e a variável dependente qualitativa são as amostras de *all-confidence* ou *cross-support* de cada faixa de suporte. No entanto, foi necessário respeitar algumas características para ser possível rodar o teste *t-student*. As amostras devem seguir uma distribuição normal e terem homogeneidade das variâncias. Caso se comprove não houve homogeneidade das variâncias, neste caso é aplicado o teste *Welch's t-student* [65]. E por fim, caso não se comprove que exista uma distribuição normal das amostras, nesse cenário é aplicado o teste não paramétrico *Mann-Whitney* [44]. Ao contrário do teste *t-student*, que testa a igualdade das médias, o teste de *Mann-Whitney* testa a igualdade das medianas. Nosso objetivo é descobrir se dada as hipóteses nulas $H_0 : \mu_{PSCIM} \leq \mu$, $H_0 : \mu_{PSCIM} = \mu$ e $H_0 : \mu_{PSCIM} \geq \mu$, onde μ_{PSCIM} sendo as amostras de valores de métricas de *all-confidence* ou *cross-support* dos itemsets fechados minerados pela técnica PSCIM e μ ser as amostras de valores de métricas de *all-confidence* ou *cross-support* dos itemsets fechados minerados pelos algoritmos concorrentes. Os testes de significância estatística tem como resultado o *p-value* que é uma medida de quanta evidência você tem contra a hipótese nula. Quanto menor o *p-value*, mais evidência você tem contra a hipótese nula. No caso contrário, quanto maior o valor de *p-value* menos evidência existe para rejeitar a hipótese nula. Partições que

Tabela 6.6: Resumo de significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo PSCIM com os algoritmos Slim, LAM e TopPI. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas nas Tabelas 6.4 e 6.5.

Base de dados	**PSCIM**	**TopPI**			**Slim**			**LAM**		
Densa	#	#	***All-confidence***	***Cross-support***	#	***All-confidence***	***Cross-support***	#	***All-confidence***	***Cross-support***
Chess	0/0	0/1	1/4/1	0/2/4	1/3	2/0/1	3/0/0	5/0	0/0/2	2/0/0
Kddcup99	0/1	0/1	3/1/2	3/1/2	2/1	1/0/3	3/0/1	5/0	0/0/2	1/0/1
Mushrooms	0/2	0/1	1/0/4	1/0/4	4/1	1/0/1	2/0/0	6/0	0/0/1	1/0/0
PowerC	2/3	1/1	1/0/1	0/0/2	2/1	0/0/2	0/0/2	5/1	0/0/1	0/0/1
Pumsb	0/0	0/1	0/2/4	0/1/5	0/0	0/2/5	2/2/3	6/0	0/0/1	0/0/1
RecordLink	0/1	0/3	1/2/0	0/3/0	3/2	1/0/0	1/0/0	4/2	0/1/0	1/0/0
Skin	1/2	1/2	0/2/1	0/2/1	2/4	0/0/0	0/0/0	3/3	0/0/0	0/0/0
Susy	2/0	1/3	1/0/2	1/0/2	-	-	-	6/0	0/0/1	1/0/0
Esparsa										
Accidents	0/1	0/0	2/1/3	0/2/4	0/0	3/1/2	4/0/2	6/0	0/0/1	1/0/0
BMSWebView2	0/3	0/2	1/0/3	0/0/4	0/0	1/0/3	1/0/3	5/0	0/0/2	1/1/0
BMS1	0/3	0/3	1/1/2	1/1/2	0/2	1/1/2	1/2/1	3/2	0/0/2	0/1/1
FoodmartFIM	3/1	3/1	0/1/2	1/0/2	-	-	-	5/0	1/0/1	1/1/0
Fruithut	1/3	1/1	0/1/2	0/2/1	1/2	0/2/1	2/1/0	6/0	0/0/1	0/0/1
OnlineRetail	1/3	1/2	0/0/3	0/0/3	3/3	0/0/1	0/0/1	6/0	0/0/1	0/0/1
PAMP	0/2	0/0	0/0/5	0/0/5	0/0	0/1/4	1/1/3	6/0	0/0/1	0/0/1
Retail	3/3	3/2	0/0/1	0/0/1	3/1	1/0/0	1/0/0	6/0	0/0/1	0/0/1

não reportaram itemsets fechados ou aqueles que tem no máximo dois itemsets fechados não podem ser usados para realizar o teste. No Apêndice C é apresentado os valores de *p-value* das três hipóteses nulas para cada partição de suporte da base de dados.

A Tabela 6.6 resume os valores de significância estatística obtidos com as técnicas comparadas. A primeira coluna contém as informações de tipo e nome da base de dado. A segunda coluna (#) mostra duas informações separados por / referentes ao algoritmo PSCIM. A primeira informação mostra a quantidade de partições que não tiveram itemsets fechados reportados pela técnica e o segundo valor representa a quantidade de partições de suporte que reportaram apenas um ou dois itemsets fechados, essa última informação é importante visto que não é possível realizar o teste de significância estatística para esse tamanho de amostra. As colunas três a cinco são as técnicas comparadas e cada uma delas contém as seguintes informações: as quantidades de partições sem itemset fechados / quantidades de partições com apenas um ou dois itemsets na partição (#); quantidades de hipóteses nulas não rejeitadas para $H_0 : \mu_{PSCIM} \leq \mu$ / $H_0 : \mu_{PSCIM} = \mu$ / $H_0 : \mu_{PSCIM} \geq \mu$ dada a técnica corrente e a métrica *all-confidence*; e quantidades de hipóteses nulas não rejeitadas para $H_0 : \mu_{PSCIM} \leq \mu$ / $H_0 : \mu_{PSCIM} = \mu$ / $H_0 : \mu_{PSCIM} \geq \mu$ dada a técnica corrente e a métrica *cross-support*. Na tabela o texto é marcado com a cor verde quando a quantidade de partições com hipótese nula $H_0 : \mu_{SCIM} \geq \mu$ for maior que outras hipóteses. Outra situação, o texto é marcado com cor vermelha quando a quantidade de partições com hipótese nula $H_0 : \mu_{SCIM} \leq \mu$ for maior que outras hipóteses. E no caso de empate

não é alterado a cor na tabela.

Observando os resultados da Tabela 5.5 nota-se que para a base de dados densas, o algoritmo PSCIM teve melhores resultados de média de *all-confidence* quando confrontado com a Técnica TopPI. O mesmo comportamento aconteceu, quando usamos a métrica *cross-support*. O PSCIM quando comparado com o Slim também teve um desempenho melhor, inclusive na base *Susy* não foi possível comparar visto que o Slim não executou nesta base de dados. No entanto, usando a métrica *cross-support* o algoritmo TopPI teve melhores resultados, embora a técnica tenha na maioria das vezes não reportado itemsets fechados em algumas faixas de suporte. Vale lembrar que a representatividade de itemsets nas diferentes faixas de suporte da base de dados também conta como critério no momento de selecionar a técnica vencedora. Com o algoritmo LAM, usando a métrica *all-confidence* o algoritmo PSCIM teve melhores resultados, enquanto usando *cross-support* a técnica LAM deve melhores resultados, e também neste caso, observa-se muitas faixas de suporte sem itemsets fechados. No caso da base de dados esparsa, o algoritmo PSCIM, usando a métrica *all-confidence*, ganhou quando comparado com todos os algoritmos. No caso da métrica *cross-support*, quando PSCIM é confrontado com o Slim não se observa um ganho de desempenho superior.

6.5.4 Análise do Custo Relativo de Processamento das Etapas do PSCIM

Este estudo visa observar o comportamento do tempo de processamento das etapas individuais do algoritmo PSCIM, para isso é selecionado as etapas de cálculo da matriz M (6.2), seguido pelo cálculo dos pesos projetados x (4.2), a matriz de ângulo (4.4), a matriz de distância (4.3), e o processo de busca de itemsets fechados na estrutura LCM. Na Figura 6.5 encontram-se as informações da proporção dos tempos consumidos por cada parte do algoritmo. Para obter o tempo de processamento do algoritmo, é usado a média de 10 execuções.

Algumas observações podem ser notadas nos resultados. Nas bases de dados *SUSY*, *Mushrooms*, *chess*, *accidents* e *pumsb*, são observadas as maiores proporções de tempo de processamento na parte M do processo *Dual Scaling*. No pior cenário, há a base de dados *SUSY* com um tamanho de 5.000.000 transações. O *Dual Scaling* neste cenário gastou 55,3% do processamento total. Isso se deve basicamente ao cálculo da matriz M, pois esta matriz possui apenas 190 itens. Além disso, a base de dados do *pumbs* tem 49.046 transações, entretanto, tem 2.113 itens. O custo de processamento, neste caso,

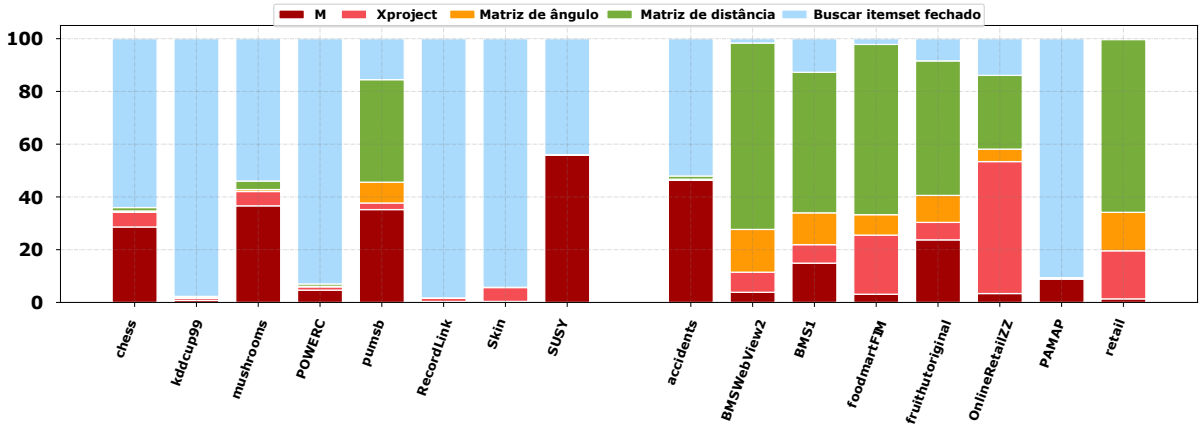


Figura 6.5: Proporção de tempo de processamento para cada etapa do algoritmo PSCIM.

está relacionado ao cálculo da matriz de distâncias.

Por outro lado, para bases de dados esparsas, como *retail*, *frutihut*, *foodmartFMI* e *BMSweb2*, observa-se que a maior proporção do tempo de processamento foi gasto para calcular a matriz de distância. Essas bases de dados têm mais de 1000 itens.

Outra observação que pode ser feita nas bases de dados *kddcup99*, *POWERC* e *PAMP*, é que essas possuem configurações semelhantes para o número de transações e o número de itens. No entanto, são vistas algumas diferenças nas proporções de tempo na parte do processamento da matriz M. Isso pode ser explicado pela redução da base de dados, aplicada a base de dados original antes do início dos cálculos. Desta forma, é eliminada a repetição de as transações repetidas na base de dados e, como consequência, há a redução do custo computacional.

Em geral, observa-se que o custo computacional está mais relacionado ao cálculo da matriz de distância dado o aumento de itens. Por outro lado, nos casos onde o número de transações é grande e o número de itens é baixo, o cálculo do custo está mais relacionado ao cálculo da matriz M.

6.5.5 Escalabilidade do PSCIM por Número de Threads

Este experimento teve como objetivo identificar o *speedup* do algoritmo PSCIM em relação ao número de *threads*. Portanto, executa-se o algoritmo PSCIM para diferentes quantidades de *threads*. O tempo de processamento é o tempo médio de 10 execuções. Neste estudo, não é considerado o carregamento da base de dados de entrada e a escrita na memória secundaria dos itemsets fechados reportados. Para cada base de dados, foi escolhido o valor vencedor para o parâmetro *dp*, ou seja, o valor de parâmetro usado no

Tabela 6.7: *Speedup* do algoritmo PSCIM.

Base de dados	Tempo (s)	*Speedup* por quantidade de threads (#)						
		# 2	# 3	**# 4**	# 5	# 6	# 7	# 8
Retail *	38.777,79	1,9	2,7	3,4	3,7	3,9	4,2	**4,4**
BMSWebView2 *	1.154,02	2,0	2,9	3,7	4,0	4,3	4,6	**4,9**
Pumsb	843,23	2,0	2,9	3,6	3,9	4,1	4,3	**4,5**
SUSY	838,58	1,9	2,6	3,1	3,3	3,4	3,5	**3,6**
Accidents *	437,87	1,9	2,8	3,4	3,6	3,7	3,9	**4,0**
Fruithutoriginal *	94,71	2,0	2,8	3,4	3,7	3,9	4,2	**4,4**
FoodmartFIM *	54,96	1,9	2,8	3,4	3,7	4,0	4,2	**4,4**
OnlineRetailZZ *	46,85	1,8	2,4	2,8	2,9	3,0	3,1	**3,2**
BMS1 *	13,37	1,9	2,8	3,3	3,6	3,8	4,0	**4,2**
PAMAP *	8,40	1,6	1,9	2,1	2,2	2,2	2,3	**2,4**
kddcup99	2,89	1,5	1,5	1,8	1,9	1,9	**2,0**	**2,0**
Mushrooms	2,34	1,8	2,5	3,1	3,1	3,2	3,5	**3,6**
Chess	2,14	1,8	2,6	3,0	3,3	3,4	3,6	**3,7**
POWERC	1,74	1,4	1,6	1,7	1,8	1,8	1,8	**1,9**
RecordLink	0,61	**1,0**	**1,0**	**1,0**	**1,0**	**1,0**	**1,0**	**1,0**
Skin	0,12	**1,0**	**1,0**	**1,0**	**1,0**	0,9	0,9	0,9

estudo da Seção 6.5.3. A Tabela 6.7 mostra o *speedup* dado o número de *threads*. A primeira coluna mostra o nome da base de dados, para as bases de dados esparsa aparece o símbolo *. A segunda coluna mostra o tempo gasto em segundos para rodar com uma *thread*. As colunas três a nove mostram o valor do *speedup* dada a quantidade de *threads*. Vale ressaltar que as bases de dados foram ordenadas de forma decrescente dado o tempo de processamento de uma *thread*.

Diante dos resultados apresentados, pode-se citar algumas observações. Não é notado um comportamento diferente devido ao tipo de base de dados. Para ilustrar esses casos, houve uma redução no *speedup* nas bases de dados *SUSY* e *Retail* de, respectivamente, 3,1× e 3,4×. Para outras bases de dados, por exemplo, *pumsb* e *BMSWebView2*, também de diferentes tipos, houve uma melhoria no *speedup* de 3,6× e 3,7×, respectivamente. Como é de se esperar no caso de mais de uma *thread*, à medida que o custo computacional de um problema diminui menor fica o *speedup*, visto que, nesses casos, o custo extra das threads acaba sendo maior que no problema original. O *speedup* médio observado nas cinco primeiras bases de dados foi de 3,4×. Vale ressaltar que apesar de utilizar até oito *threads* nos testes, define-se como *speedup* apenas o exemplo de 4 *threads*, como mostrado em negrito na Tabela 6.7, principalmente pelo fato de que o processador usado nos testes possuir apenas quatro núcleos reais.

Com os dados apresentados, observam-se um comportamento de um algoritmo escalável por número de *threads*. Porém, não podemos afirmar isso, visto que os testes apresentados foram feitos em um cenário de até 4 núcleos reais, é necessário rodar os experimentos em computadores com maiores números de núcleos de processadores para poder afirmar tal característica ao algoritmo proposto.

Capítulo 7

Dual Scaling Processado em Blocos de Transações

Com o avanço da tecnologia e o baixo custo, por exemplo, para dispositivos móveis, os dados gerados tendem a ter um fluxo quase contínuo. Nessas aplicações, não é viável guardar os dados que chegam em uma base de dados tradicional. Neste cenário, é chamado de fluxo de dados aquelas informações que apenas ficam um instante no dispositivo e depois são removidas para receber uma nova informação, sendo nominalmente chamada bloco de informação [13]. Na área de processamento de dados é definido o domínio de bases de dados de fluxo contínuo. Onde o modo de como as novas informações chegam continuamente e como são mantidas as mesmas em memória principal definem o tipo de janela. Os três tipos de janelas mais relevantes são ilustrados na Figura 7.1.

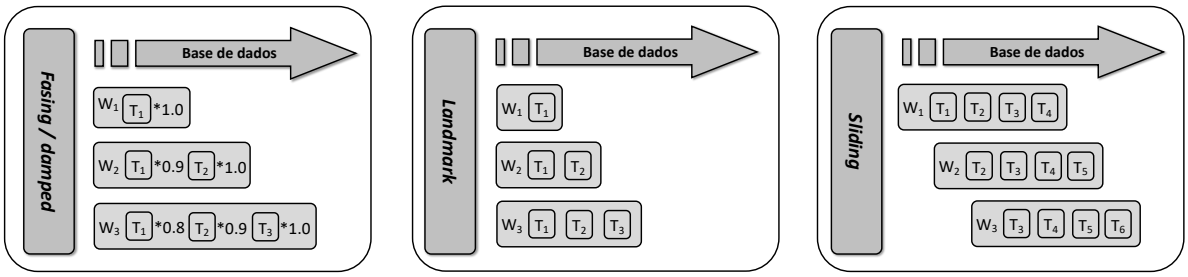


Figura 7.1: Três tipos de janelas que define o modo de atualização para bases de dados de fluxo contínuo.

O primeiro tipo de janela, *landmark* [68], considera todos os elementos (transações) do fluxo de leitura, do início até o momento atual T_k, onde k representa o bloco de transações corrente. Neste tipo, a janela fica maior à medida que mais blocos de elementos chegam no fluxo de dados. Cada elemento é atribuído com o mesmo peso de relevância, como consequência, a influência de cada transação desaparece gradualmente à medida que a

janela vai ficando maior. O segundo tipo de janela, *sliding* [6], considera apenas a faixa que compreende o bloco de transações $T_{\max(k-q+1,1)}$ até o bloco de transações mais recentes T_k, onde q é o tamanho da janela. Neste tipo, a janela tem um tamanho fixo q e caso a janela esteja cheia, a cada novo bloco de transações, o bloco mais antigo deve ser excluído da janela. O último tipo de janela, *fasing/damped* [68], considera os blocos de transações com pesos diferentes. Neste tipo, os elementos mais recentes têm um peso maior do que os mais antigos. Com cada novo bloco que chega, o peso de todos os blocos anteriores decaem por um fator tal que blocos vistos ha muito tempo serão esquecidos eventualmente dado o peso pequeno definido.

Na base de dados estática o dado reside em memória secundária e pode ser acessado várias vezes. No entanto, no cenário de base de dados de fluxo contínuo o domínio possui crescimento contínuo de informações, sendo muitas vezes impossível manter as mesmas salvas para futuro acesso [7]. Podemos citar alguns exemplos, tais como: análise de registro de detalhes de chamadas telefônicas, detecção de fraude, monitoramento de sensores, recomendações ou segurança de rede. Como mencionado antes, nas bases de dados transacional de fluxo contínuo, os elementos (transações) chegam continuamente, não existe um controle sobre a ordem de chegada dos elementos. O tamanho da base de dados é ilimitado e, por consequência, as informações de transações residem por um período na memória principal. A distribuição de frequência dos itens de um novo bloco de transações pode mudar com o tempo, ou seja, o comportamento das coocorrência dos itens da base de dados pode mudar a cada nova atualização.

A ideia central desta tese é usar a contextualização espacial no processo de mineração de itemsets. No entanto, o cálculo do *Dual Scaling* apresentado no Capítulo 4 e utilizado nos Capítulos 5 e 6 não permite trabalhar com bases de dados de fluxo contínuo. Isso porque, para usar a solução atual do algoritmo *Dual Scaling*, teríamos que a cada novo bloco de transações recalcular todo o processo de mapeamento espacial fornecido pelo algoritmo *Dual Scaling*. Por isso, esse estudo propõe uma solução que consegue manter a atualização do *Dual Scaling* a cada novo bloco (Seção 7.1), sem a necessidade de reprocessamento completo dos dados. Apenas a parte afetada (i.e., dados novos e dados a serem esquecidos) são consideradas. Adicionalmente, será demostrando neste capítulo que a solução aplicada para fluxo de dados pode também ser empregada para lidar com segurança da informação em bases estáticas ou de fluxo contínuo, tendo como respaldo a regulamentação geral de segurança de dados (Seção 7.2).

7.1 *Dual Scaling* em Bases de Dados de Fluxo Contínuo

O cálculo do *Dual Scaling* está dividido em três partes, sendo a primeira parte o cálculo da matriz de resíduo M (Equação 4.1), depois o cálculo de autovalores e vetores da matriz M, e por último o cálculo da matriz do peso projetado (Equação 4.2). A solução proposta para atualização em blocos tem como foco a primeira parte do algoritmo:

$$M = F^T D_r^{-1} F D_c^{-1} \equiv F^T D_r^{-1} F D_c^{+}, \tag{7.1}$$

onde D_c^+ é a matriz pseudo-inversa da marginal de frequência das colunas de F. O uso da pseudo-inversa é necessário, pois não é possível garantir que todos os itens da base de dados estarão presentes no momento atual do fluxo contínuo. Portanto, em alguns casos, possa haver marginais das colunas com valores zero, levando a matrize D_c não inversível.

Dada a matriz F de entrada com dimensões $n \times m$, considere que esta é dividida em dois blocos, A e B, com dimensões $l \times m$ e $p \times m$, respectivamente, sendo que $l + p = n$ é o número de transações e m o número de itens da base de dados. Desta forma, é possível reescrever o cálculo da matriz M do seguinte modo:

$$\begin{aligned}
M &= F^T D_r^{-1} F D_c^{+} \\
&= \left[\begin{array}{c|c} A^T & B^T \end{array}\right] \left[\begin{array}{c|c} D_{r_A}^{-1} & 0 \\ \hline 0 & D_{r_B}^{-1} \end{array}\right] \left[\begin{array}{c} A \\ \hline B \end{array}\right] D_c^{+} \\
&= \left[\begin{array}{c|c} A^T D_{r_A}^{-1} & B^T D_{r_B}^{-1} \end{array}\right] \left[\begin{array}{c} A \\ \hline B \end{array}\right] D_c^{+} \\
&= \left(A^T D_{r_A}^{-1} A \; + \; B^T D_{r_B}^{-1} B \right) D_c^{+} \\
&= \left(A^T D_{r_A}^{-1} A \underbrace{D_{c_A}^{+} D_{c_A}}_{I} + B^T D_{r_B}^{-1} B \underbrace{D_{c_B}^{+} D_{c_B}}_{I} \right) D_c^{+} \\
&= \left(\underbrace{A^T D_{r_A}^{-1} A D_{c_A}^{+}}_{M_A} D_{c_A} \; + \; \underbrace{B^T D_{r_B}^{-1} B D_{c_B}^{+}}_{M_B} D_{c_B} \right) D_c^{+} \\
&= \left(M_A D_{c_A} \; + \; M_B D_{c_B} \right) D_c^{+},
\end{aligned} \tag{7.2}$$

onde seja D_{r_A} e D_{r_B} são matrizes diagonais de frequência da marginal de linhas, respectivamente, dos blocos A e B. A matriz diagonal de frequência da marginal de colunas da matriz F atual é representada por D_c. Seja D_{c_A} e D_{c_B} matrizes diagonais de frequência da marginal de colunas, respectivamente, dos blocos A e B. Dada uma matriz Q, temos que

Q^T, Q^{-1}, Q^+ e $Q^T\,Q$ denotam, respectivamente, a transposição, inversão, pseudo-inversa da matriz Q e a matriz identidade.

No cenário de bases de dados de fluxo contínuo novas informações de blocos chegam a cada momento. Por isso, a necessidade de evitar recálculos de toda a matriz M a cada novo bloco. Na Equação 7.2 a matriz F foi dividida em dois blocos A e B. O primeiro passo foi separar a manipulação em blocos de informações de modo que sejam independentes, desta forma foi possível definir a manipulação $T\,Dc^+$, onde $T = \ A^T\,D_{r_A}^{-1}\,A \ + \ B^T\,D_{r_B}^{-1}\,B$. No primeiro momento nota-se que para ter a propriedade distributiva é necessário conhecer o pseudo-inversa $D_c^+ = D_{c_A}^+ + D_{c_B}^+$ pois ela representa a marginal de coluna da matriz F atual, portanto, a cada novo bloco a matriz D_c^+ é alterada. Diferente quando comparado com as matrizes $D_{r_A}^{-1}$ e $D_{r_B}^{-1}$ que não precisam da informação do F atual, visto que são marginais de linhas de cada bloco. Para calcular a Equação 7.1 para cada bloco de forma independente aplicamos a propriedade de matriz identidade, desta forma foi possível isolar a Equação 7.1 por blocos.

Para exemplificar a atualização da matriz M é definido um novo bloco C para ser inserido na matriz F atual. No primeiro caso, para conseguir atualizar a matriz M é necessário guardar, também, a informação da matriz T. Já no segundo caso, apenas é preciso usar a matriz M, já calculada, e a matriz D_c^+. Com essa manipulação do cálculo da matriz M conseguimos definir soluções para o cálculo de *Dual Scaling* em blocos de transações para os dois tipos de janela: *landmark* e *sliding*. O *fasing/damped* não pôde ser considerado, visto que o *Dual Scaling* lida apenas com matrizes no formato de padrão de resposta (1,0) e o uso de peso sobre cada transação iria descaracterizar o tipo de dados permitido.

É importante salientar que o resultado da atualização da matriz M proposto pela Equação 7.2 não se trata de uma aproximação e ela pode ser usada nos Algoritmos SCIM e PSCIM. Logo, o resultado esperado é igual ao apresentado nos Capítulos 5 e 6, se assumirmos que as bases de dados utilizadas nesses capítulos correspondem aos dados contínuos na janela atual de um fluxo de dados.

7.1.1 Fluxo de Dados do Tipo *Landmark*

Neste tipo de janela, a base de dados de fluxo contínuo mantém os dados dês o início dos registros de blocos de transação até o bloco corrente atual. Dado o resultado da manipulação algébrica apresentada na Equação 7.2, podemos definir uma solução para atualização da matriz de resíduo M. No Algoritmo 6 define-se a atualização da matriz M

para o tipo de janela *landmark*. O algoritmo tem como entrada a base de dados de fluxo contínuo onde os blocos de transações F_k são de dimensões $j_k \times m$, onde j_k é o número de transações no k-ésimo bloco, para $j_k \geq 1$, e m número de itens.

Algoritmo 6: Atualização da matriz M para incremento do tipo *landmark*

Entrada: Blocos F_1, F_2, ... para a base de dados de fluxo contínuo

1 $D_c \leftarrow D_{c_1}$ ▷ Matriz diagonal de frequência das colunas em F atual
2 $M \leftarrow F_1^T Dr_1^{-1} F_1 Dc_1^+$ ▷ Matriz M atualizada, Equação 7.1
3 **para cada** *novo bloco F_k, com $k > 1$* **faça**
4 Calcular *Dual Scaling* dada a matriz M ▷ Subseção 4.2
5 $M_B \leftarrow F_k^T D_{r_k}^{-1} F_k D_{c_k}^+$ ▷ Matriz da Equação 7.1, dado o bloco F_k
6 $S \leftarrow D_c$ ▷ Matriz diagonal D_c
7 $D_c \leftarrow D_c + D_{c_k}$ ▷ Atualizar D_c dado F atual
8 $M \leftarrow (M\,S + M_B\,D_{c_k})\,D_c^+$ ▷ Matriz M atualizada, Equação 7.2
9 **fim**

No Algoritmo 6, linha 1, é definida a matriz diagonal D_c contendo a marginal de frequência das colunas da matriz F atual, a matriz D_c vai ser usado para acumular os valores de frequência a cada novo bloco. Logo após, na linha 2 é calculada a matriz de resíduo para o bloco F_1, vez é a matriz M atualizada, pois nesse momento apenas existe o bloco F_1 para a matriz F atual. No laço, na linha 3, temos a condição para entrada de novo bloco de transações. Na linha 4 temos a chamada do resto do cálculo para o mapeamento do *Dual Scaling*, onde é calculado as projeções dos itens da base de dados no espaço de soluções. Na linha 5, temos a matriz diagonal S que recebe o valor da diagonal D_c antes da atualização do novo bloco F_k. A linha 6 atualiza as informações de marginal de frequência das colunas da matriz F atual. Perceba que é sempre acumulado com as frequências anteriores, pois o tipo de janela concatena cada novo bloco na matriz F. E por fim, é atualizado a matriz M dada a nova atualização da matriz F. E esse processo continua a cada novo bloco de transações no laço da linha 3.

O tipo de janela *landmark* não é muito utilizado, visto que é um desafio manter um dado ilimitado em uma memória limitada. No entanto, observando o comportamento do algoritmo proposto, percebe-se uma característica bem interessante, pois a memória necessária para manter a matriz de resíduo não muda ao longo do tempo, visto que a complexidade de tamanho é dada pelo número de itens, $\mathcal{O}(m^2)$. Deste modo, a matriz atualizada M poderia ser usada como solução para este problema.

Algoritmo 7: Atualização da matriz M para incremento do tipo *Sliding*

Entrada: - Blocos F_1, F_2, ..., F_k para a base de dados de fluxo contínuo
- Janela com no máximo q blocos

1 $D_c \leftarrow D_{c_1}$ ▷ Matriz diagonal de frequência das colunas em W atual
2 $M \leftarrow F_1^T D_{r_1}^{-1} F_1 D_{c_1}^+$ ▷ Matriz M atualizada, Equação 7.1
3 Inicializar a fila $\mathcal{G}$ com D_{c_1}
4 Inicializar a fila $\mathcal{M}$ com M
5 **para cada** *novo bloco* F_k, *com* $k > 1$ **faça**
6 Calcular *Dual Scaling* dada a Matriz M ▷ Subseção 4.2
7 $M_B = F_k^T D_{r_k}^{-1} F_k D_{c_k}^+$ ▷ Matriz da Equação 7.1 dado o bloco F_k
8 Inserir D_{c_k} no final da fila $\mathcal{G}$
9 Inserir M_B no final da fila $\mathcal{M}$
10 $S \leftarrow D_c$ ▷ Matriz diagonal D_c
11 **se** $k > q$ **então**
12 $R \leftarrow$ primeiro elemento da fila $\mathcal{G}$
13 Remover primeiro elemento fila $\mathcal{G}$
14 $T \leftarrow$ primeiro elemento da fila $\mathcal{M}$
15 Remover primeiro elemento fila $\mathcal{M}$
16 $M \leftarrow M - T$
17 $D_c \leftarrow D_c - R$
18 **fim**
19 $D_c \leftarrow D_c + D_{c_k}$ ▷ Atualizar D_c dado o W atual
20 $M \leftarrow (M\,S + M_B\,D_{c_k})\,D_c^+$ ▷ Matriz M atualizada, Equação 7.2
21 **fim**

7.1.2 Fluxo de Dados do Tipo *Sliding*

O próximo algoritmo proposto tem como característica atualizar a matriz de resíduo M dado um fluxo de dados do tipo *sliding*. Neste cenário, blocos de transações antigos são eliminados dando espaço para novos blocos de transações. Com a Equação 7.2 pode-se fazer o processo de retirar as informações de blocos no processo de atualização da matriz de resíduo M. O Algoritmo 7 tem como entrada a base de dados de fluxo contínuo onde os blocos de transações F_k são de dimensões $j_k \times m$, onde j_k é o número de transações no k-ésimo bloco, para $j_k \geq 1$, e m número de itens. O segundo parâmetro é o tamanho máximo da janela, podendo conter q blocos de transações.

Na linha 1 define a matriz diagonal B contendo as marginais de coluna da janela. Sendo que no caso, apenas existe o bloco F_1 na janela. Na linha 2 é calculada a matriz de resíduo para o primeiro bloco de transações, isso significa que temos a matriz M atualizada dada a janela W atual. Nas linhas 3 e 4 são criadas as listas $\mathcal{G}$ e $\mathcal{M}$ que serão usadas para armazenar as matrizes diagonais dos blocos de transações e as matrizes de

resíduo calculadas por cada bloco, respectivamente. Armazenar esses dados em listas são importantes, porque no futuro, quando a janela estiver cheia as matrizes armazenadas são utilizadas para remover informações antigas que já não são mais encontradas na janela W atual.

No laço, linha 5, é representada a condição para entrada de novo bloco de transações F_k. É importante salientar que $k > 1$, pois o primeiro bloco já foi computado anteriormente. A linha 6 chama o resto do cálculo para o mapeamento do *Dual Scaling*. Nos próximos passos a janela é atualizada com novo bloco F_k, onde para o caso onde $k < q$ temos que a janela ainda não está cheia. Na linha 7 é feito o cálculo da matriz de resíduo para o bloco corrente F_k salvo na matriz M_B. Nas linhas 8 e 9 são inseridos na lista as marginais de coluna e matriz de resíduo para o bloco corrente. Se a janela estiver cheia (linha 11) é necessário remover o bloco mais antigo, por isso são obtidos S e T das listas (linhas 12 e 14). Logo após os dados são removidos das listas (linhas 13 e 15). O processo de remover informações antigas da matriz de resíduo consiste em subtrair o dado antigo com o atual (linhas 16 e 17). Na linha 19 é atualizada a matriz marginal de colunas da janela atual, e por fim, na linha 20, é atualizado a matriz M dada a janela atual. Como mencionado anteriormente, dado o cenário limitado de blocos de transações em memória principal, onde é definido q blocos.

7.2 Regulamentação Geral de Segurança de Dados

A GDPR (*General Data Protection Regulation*) foi definida pelo parlamento europeu estabelecendo regras comuns para a legislação nacional de proteção de dados dos países membros da união europeia (UE) [60]. A GDPR tem como meta proteger a privacidade de todos os indivíduos dentro da UE e, como consequência, evitar o uso indiscriminado de informações.

As normas da GDPR são compostas por vários artigos, dentre eles podemos mencionar Artigo 17 que tem como garantia o direito ao esquecimento, ou seja, empresas tem obrigação de remover qualquer informação do usuário. Os indivíduos que requerem esse direito devem ter seus dados pessoais removidos dos registros da empresa. Desta forma, a etapa de remoção dos dados pessoais deve abranger não apenas os dados brutos, mas também dados processados com essas informações. Um exemplo que podemos citar, é o caso de ataque de invasores aos modelos neurais treinados com dados de usuários [24]. Nesse sentindo, se um invasor pode atacar um modelo treinado e aprender informações

Algoritmo 8: Remover informações de usuário da Matriz de resíduo M

Entrada: - U: base de dados contendo informações do usuário
- M: matriz de resíduo existente
- D_c: matriz diagonal de marginal de colunas da base atual

Resultado: Matriz M atualizada sem os dados do usuário

1 $M_B \leftarrow U^T D_{r_u}^{-1} U D_{c_u}^{+}$ ▷ Matriz da Equação 7.1 dado o bloco U
2 $M \leftarrow M - M_B$
3 $D_c \leftarrow D_c - D_{c_u}$
4 $M \leftarrow M D_c^{+}$ ▷ Matriz da Equação 7.2

privadas sobre um indivíduo que invocou o direito ao esquecimento, então o proprietário do modelo pode ser responsabilizado.

Neste trabalho é proposto o Algoritmo 8, cujo objetivo é remover as informações prévias de um usuário e foram utilizadas no cálculo matriz de resíduo M. Nesse procedimento não é preciso recalcular M para a nova base de dados obtida com a exclusão dos registros do usuário em questão. O diferencial desse algoritmo é que apenas seja preciso calcular a matriz de resíduo do conjunto de dados do usuário em questão e por fim subtrair a informação da matriz M. Na linha 1 é calculada a matriz de resíduo M da base de dados U que contém todas as transações, incluindo as do usuário, e onde D_{r_u} e D_{c_u} são matrizes diagonais de marginais de linha e coluna da matriz U de entrada. As linhas 2 a 4 são responsáveis pela atualização da matriz M sem os dados do usuário.

7.3 Discussão

Durante este capítulo, foi proposto o cálculo do mapeamento dos itens das bases de dados no espaço de soluções para bases de dados de fluxo continuo. Infelizmente o tipo de janela *fasing/damped* não pode ser utilizada, porque neste tipo de janela são aplicados pesos nas transações, inviabilizando o uso de tipo de base de dados para o cálculo do *Dual Scaling*.

A manipulação proposta (Equação 7.2) garante a atualização da matriz M para evitar recalculá-la explicitamente a partir de toda a base de dados atualizada dado um novo bloco de transações. Vale salientar que a manipulação afeta apenas a matriz M. O resto do cálculo do *Dual Scaling* segue o mesmo fluxo de processamento como descrito no Capítulo 4.

Durante este capítulo não foi preciso demonstrar experimentação e nem validação da técnica proposta, porque as manipulações apresentadas aqui, sobre a matriz de resíduo

M, não são aproximações. Ou seja, existe a igualdade da solução. A solução proposta resolve um dos desafios encontrados num cenário de bases de dados de fluxo contínuo. Porém, é preciso enfatizar que apresentamos uma solução teórica do problema em base de dados de fluxo contínuo, por tanto é preciso realizar experimentos para observar a solução proposta.

Um fator importante que a manipulação da matriz M trouxe foi o seu uso prático para aplicado no ambiente sujeito a regulamentação geral de segurança de dados. Nesse contexto, com a manipulação algébrica proposta, foi possível definir uma funcionalidade que respeita o Artigo 17 da GDPR. Não foi mostrado neste capítulo um exemplo prático do uso da função, no entanto, acredita-se que possa ser útil para diferentes contextos.

Capítulo 8

Conclusões

Os algoritmos de mineração de itemsets fechados têm como característica comum o uso de limiares para controlar a quantidade de padrões reportados. Vimos que esses tipos de limiares têm um papel importante no processo de mineração. Porém, é observado na literatura esforços para tornar o algoritmo de mineração escalável, menos sensível aos parâmetros e mais aptos à seleção de itemsets que sejam relevantes para o usuário da técnica. Para que esses objetivos sejam alcançados, o problema tem sido atacado sob diversas perspectivas, seja através de definição de estruturas que visam fazer indexações para otimizar a geração de itemset, seja pela definição de novos limiares, em alguns casos múltiplos limiares, ou com o intuito de melhorar a extração, agregando itemsets raros.

Nesta tese, foram apresentados três vertentes de pesquisa. Na primeira, foi proposto um algoritmo sequencial para identificação de itemset fechado para base de dados estáticas. Já na segunda vertente, foi proposta uma solução paralela para identificação de itemsets fechados também em base de dados estáticas. E por último, foi proposto um estudo teórico para o cálculo do mapeamento do espaço de soluções para bases de dados de fluxo contínuo.

Essa tese propõe uma solução de algoritmo que aplica a técnica de mapeamento de análise multidimensional, como prova de conceito foi utilizado o *Dual Scaling* para o processo de mineração. Durante os estudos observou-se que a técnica de análise multidimensional deve ter as seguintes propriedades: A distância do item a origem é inversamente proporcional à sua frequência na base de dados (Propriedade 1); A distância entre itens diminui à medida que a coocorrência de itens aumenta (Propriedade 2). Experimentos mostram que com a definição de clusteres no espaço contextualizado é possível usar informação de distância entre pares de itens durante o processo de mineração, de modo a reduzir o espaço de busca e melhorar a seleção de itemsets fechados relevantes.

Os estudos realizados para a primeira vertente, SCIM, encontram-se concluídos e publicados [40]. Na abordagem desenvolvida, a contextualização espacial permite a formação de clusters de itens relacionados. A abordagem explora a visão vertical do banco de dados para combinar itens eficientemente em itemsets fechados que respeitam a regra de formação definida conforme a distribuição espacial dos dados nos clusters. Durante o estudo, foi demonstrada a eficácia da técnica proposta usando o procedimento em várias bases de dados públicas. A validação experimental mostrou que a média de *all-confidence* dos itemsets fechados recuperados pelo algoritmo *SCIM* supera os resultados de outras soluções. Os experimentos mostram que os itemsets fechados do *SCIM* são naturalmente recuperados em todas as faixas de suporte, mesmo quando o limite máximo da relação de distância é mais restritivo. Até onde sabemos, o algoritmo *SCIM* é o primeiro algoritmo que usa a contextualização espacial como critério para formação de itemsets e para estratégia de poda para a mineração de itemsets.

O algoritmo PSCIM propõe uma solução paralela de mineração de itemsets fechados. Utilizando a contextualização espacial, foi possível controlar o espaço de busca na estrutura LCM usada para o processo de recuperação dos itemsets fechados. No intuito de tornar o algoritmo mais flexível para base de dados esparsas, algumas alterações foram propostas no cálculo do mapeamento espacial, tanto para melhorar a performance dos cálculos quanto para melhorar a qualidade da formação dos clusters. Foi apresentada 16 bases de dados, onde metade era densa e a outra metade esparsa. A validação experimental demostrou que a média de *all-confidence* dos itemsets fechados recuperados pelo algoritmo PSCIM supera o resultado de outras soluções. Um fato importante, com a alteração proposta no mapeamento espacial é notado que a técnica proposta ganhou em todas as bases de dados esparsas testadas. O comportamento de selecionar itemsets fechados em todas as faixas de suporte se manteve tanto para base de dados densas quanto para bases de dados esparsas.

O primeiro desafio, no cenário de bases de dados de fluxo contínuo, é o mapeamento do espaço de soluções, visto que o cálculo atual era definido para bases de dados estáticas. Nesse sentindo foi apresentado uma solução teórica para o cálculo do *Dual Scaling* em blocos de transações e desta forma evitar o recalculo da matriz de resíduo M. As soluções foram propostas para dois tipos de atualização de blocos, o *landmark* e o *sliding*. Além disso, notamos que o cálculo do mapeamento para base de dados contínuo poderia também ser usado para retirar informações de usuários da matriz de resíduo calculada, propondo assim uma solução teórica para as normas de proteção de dados (GDPR), como por exemplo o Artigo 17.

8.1 Trabalhos Futuros

O algoritmo paralelo PSCIM possui otimizações que flexibilizaram o custo de processamento nos procedimentos até a definição de clusters. No entanto, um dos gargalos encontrados no algoritmo está relacionado ao custo de memória, mais especificamente na Equação 4.1 onde temos a complexidade de memória $\mathcal{O}(m^2)$, sendo m o número de itens da base de dados. Embora tenha proposto uma implementação otimizada no cálculo da matriz M, no final a matriz resultante deve ser densa, pois a biblioteca StarNEig, responsável pelo cálculo dos autovalores e autovetores, foi desenvolvida para matrizes densas e não simétricas. Na literatura são encontradas soluções para matrizes esparsas [54, 41], no entanto, não achamos implementações da mesma. Neste sentido, seria interessante implementar tais soluções, para tratar o gargalo de memória da matriz M.

Nesta tese foi definida uma solução de recálculo do *Dual Scaling*, sem a necessidade de recalcular toda a matriz M dada a base de dados de fluxo contínuo atualizada a cada novo bloco de transações. Um trabalho futuro seria implementar um algoritmo de mineração de itemsets fechados no cenário de base de dados de fluxo contínuo. Alguns fatores têm que ser estudados, sobre o custo de recalcular sempre a matriz de distância e da matriz de angulação para todos os itens, a visto que as atualizações de blocos não apresentam a ocorrência de todos os itens da base de dados. Outro desafio é usar estruturas que auxiliem na recuperação de itemsets fechados para bases de fluxo contínuo. Uma estrutura que pode ser citada é a SFI-*florest* [33]. Assim como foi feito com outras estruturas, é necessário verificar como pode-se usar as informações dos clusters para auxiliar na atualização da estrutura SFI-*florest*, de tal modo que seja possível recuperar itemsets fechados representados pelos clusters.

Uma das vertentes apresentadas nesse projeto, durante a qualificação, foi a utilização do mapeamento espacial para auxiliar no processo de mineração de itemsets de utilidade alta. Nesse cenário, o itemset de utilidade alta pode ser visto como uma generalização do itemset frequente, onde cada item em vez de ocorrência agora possui o seu valor de utilidade. Os algoritmos nessa área também usam limiares, por exemplo, mínima utilidade, para retornar apenas itemsets de utilidade de interesse do especialista. Durante os estudos iniciais não foram observados um comportamento de seleção de itemsets de utilidade alta relevantes, neste caso foi usado a métrica afinidade de frequência [5] para medir relevância do itemset de utilidade alta selecionado. No entanto, os testes não foram aprofundados no sentido de poder refutar essa vertente. Por isso, como trabalho futuro é sugerido utilizar outras técnicas da família de análise multidimensional que possa auxiliar

como tomada de decisão na seleção de itemsets de utilidade alta.

A análise de escalabilidade do algoritmo pode ser feita através de um cenário onde é fixo o número de itens da base de dados e em contrapartida é feito um aumento gradual de transações na base de dados. Outro cenário seria fixar as transações e variar os itens. Como trabalho futuro, fazer tal análise de modo a observar o comportamento de tempo de processamento e uso de memória neste cenários.

Durante a tese, foi utilizado o *Dual Scaling* como prova de conceito para validar a ideia central. Foram identificadas duas características importantes que devem estar presentes no mapeamento dessas técnicas de análise multidimensional (Propriedades 1 e 2. Como trabalho futuro, seria interessante explorar outras técnicas além do *Dual Scaling*.

Em resumo, este trabalho visa dar um passo adiante em uma área estudada tão exaustivamente, através da definição de novos algoritmos que usam a contextualização da base de dados como parte do processamento, tendo como finalidade reduzir o espaço de busca para a geração de itemsets relevantes. Esperamos com os resultados apresentados encorajar o uso e estudo dos algoritmos propostos em aplicações futuras, visando um dia atingir outras vertentes de pesquisa além das expostas nesta tese.

Referências

[1] *Proceedings of the Second International Conference on Advanced Data Mining and Applications* (Berlin, Heidelberg, 2006), Springer-Verlag.

[2] Aggarwal, Charu C.; Bhuiyan, M. A., Hasan, M. A. *Frequent Pattern Mining Algorithms: A Survey.* Springer International Publishing, 2014, p. 19–64.

[3] Agrawal, R., Imieliński, T., Swami, A. Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*, 2 (1993), 207–216.

[4] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB* (1994), p. 487–499.

[5] Ahmed, C. F., Tanbeer, S. K., Jeong, B.-S., Choi, H.-J. A framework for mining interesting high utility patterns with a strong frequency affinity. *Information Sciences 181*, 21 (2011), 4878 – 4894.

[6] Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J. Models and issues in data stream systems. In *In PODS* (2002), p. 1–16.

[7] Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J. Models and issues in data stream systems. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (New York, NY, USA, 2002), PODS '02, Association for Computing Machinery, p. 1–16.

[8] Bouasker, S., Ben Yahia, S. Key correlation mining by simultaneous monotone and anti-monotone constraints checking. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (2015), p. 851–856.

[9] Brin, S., Motwani, R., Ullman, J. D., Tsur, S. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec. 26*, 2 (1997), 255–264.

[10] Bruzzese, D., Davino, C. Visual mining of association rules. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, S. J. Simoff, M. H. Böhlen, and A. Mazeika, Eds. Springer Berlin Heidelberg, 2008, p. 103–122.

[11] Buehrer, G., de Oliveira, R. L., Fuhry, D., Parthasarathy, S. Towards a parameter-free and parallel itemset mining algorithm in linearithmic time. In *2015 IEEE 31st International Conference on Data Engineering* (2015), p. 1071–1082.

[12] Chen, D., Lai, C., Hu, W., Chen, W., Zhang, Y., Zheng, W. Tree partition based parallel frequent pattern mining on shared memory systems. In *Proceedings 20th IEEE International Parallel Distributed Processing Symposium* (2006), p. 8 pp.–.

[13] Cheng, J., Ke, Y., Ng, W. A survey on algorithms for mining frequent itemsets over data streams. *Knowl. Inf. Syst. 16*, 1 (2008), 1–27.

[14] Coenen, F. The lucs-kdd discretised/normalised arm and carm data library. `http://www.csc.liv.ac.uk/\textasciitildefrans/KDD/Software/LUCS\_KDD\_DN`, junho de 2021.

[15] Comaniciu, D., Meer, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 5 (2002), 603–619.

[16] Dagum, L., Menon, R. Openmp: an industry standard api for shared-memory programming. *Computational Science & Engineering, IEEE 5*, 1 (1998), 46–55.

[17] DHEERU, D.; KARRA TANISKIDOU, E. Contraceptive method choice data set. uci machine learning repository. `https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`, junho de 2021.

[18] Ester, M., Kriegel, H.-P., Sander, J., Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Data base analysis* (1996), AAAI Press, p. 226–231.

[19] Fasham, M. J. R. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. *Ecology 58*, 3 (1977), 551–561.

[20] Fernandes, L. A. F., García, A. C. B. Association rule visualization and pruning through response-style data organization and clustering. In *Advances in Artificial Intelligence – IBERAMIA 2012*, J. Pavón, N. D. Duque-Méndez, and R. Fuentes-Fernández, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, p. 71–80.

[21] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., Tseng, V. S. Spmf: A java open-source pattern mining library. *J. Mach. Learn. Res. 15* (2014).

[22] Fournier-Viger, P., Lin, J. C.-W., Vo, B., Chi, T. T., Zhang, J., Le, H. B. A survey of itemset mining. *Data Mining and Knowledge Discovery 7*, 4 (2017), e1207.

[23] Fournier-Viger, P., Lin, J. C.-W., Vo, B., Chi, T. T., Zhang, J., Le, H. B. A survey of itemset mining. In *WIREs Data Mining and Knowledge Discovery* (2017), p. 1207.

[24] Fredrikson, M., Jha, S., Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2015), Association for Computing Machinery, p. 1322–1333.

[25] Frey, B. J., Dueck, D. Clustering by passing messages between data points. *Science 315*, 5814 (2007), 972–976.

[26] Garofalakis, M., Gehrke, J., Rastogi, R. Querying and mining data streams: You only get one look a tutorial. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data* (2002), p. 635–635.

[27] Grahne, G., Zhu, J. Fast algorithms for frequent itemset mining using fp-trees. *IEEE Transactions on Knowledge and Data Engineering 17*, 10 (2005), 1347–1362.

[28] Han, J., Pei, J., Yin, Y. Mining frequent patterns without candidate generation. *SIGMOD Rec. 29*, 2 (2000), 1–12.

[29] Hofmann, T., Schölkopf, B., Smola, A. J. Kernel methods in machine learning. *The annals of statistics* (2008), 1171–1220.

[30] Kirchgessner, M., Leroy, V., Kirchgessner, M., Termier, A., Amer-Yahia, S. TopPI: an efficient algorithm for item-centric mining. *Inf. Syst. 64* (2017), 104–118.

[31] Lebart, L., Morineau, A., Warwick, K. M. *Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices.* Appl. Stochastic Models Data Anal., 1984.

[32] Ledin, J. *Modern Computer Architecture and Organization.* Packt Publishing, 2020.

[33] Li, H.-F., Shan, M.-K., Lee, S.-Y. Dsm-fi: an efficient algorithm for mining frequent itemsets in data streams. *Knowledge and Information Systems 17*, 1 (2008), 79–97.

[34] Li, Z.-C., He, P.-L., Lei, M. A high efficient aprioritid algorithm for mining association rule. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on* (2005), vol. 3, p. 1812–1815.

[35] Liu, B., Hsu, W., Ma, Y. Mining association rules with multiple minimum supports. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999), p. 337–341.

[36] Liu, X., Guan, J., Hu, P. Mining frequent closed itemsets from a landmark window over online data streams. *Computers and Mathematics with Applications 57*, 6 (2009), 927–936.

[37] Macqueen, J. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability* (1967), p. 281–297.

[38] Manku, G. S., Motwani, R. Approximate frequency counts over data streams. In *Proceedings of the 28th International Conference on Very Large Data Bases* (2002), p. 346–357.

[39] Mantuan, A. Contextualização espacial para mineração de itemsets raros ou frequentes não-redundantes em bases de dados. Dissertacao de mestrado, Universidade Federal Fluminense (UFF) LaTeX, 2016.

[40] Mantuan, A., Fernandes, L. Spatial contextualization for closed itemset mining. In *in Proc. IEEE International Conference on Data Mining (ICDM)* (2018), p. 1176–1181.

[41] MEERBERGEN, K., ROOSE, D. Matrix transformations for computing rightmost eigenvalues of large sparse non-symmetric eigenvalue problems. *IMA Journal of Numerical Analysis 16*, 3 (1996), 297–346.

[42] Myllykoski, M. A task-based multi-shift qr/qz algorithm with aggressive early deflation, 2020.

[43] Negrevergne, B., Termier, A., Méhaut, J.-F., Uno, T. Discovering closed frequent itemsets on multicore: Parallelizing computations and optimizing memory accesses. In *2010 International Conference on High Performance Computing Simulation* (2010), p. 521–528.

[44] Neuhäuser, M. *Wilcoxon–Mann–Whitney Test*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, p. 1656–1658.

[45] Nishisato, S. On quantifying different types of categorical data. *Psychometrika 58*, 4 (1993), 617–629.

[46] Nishisato, S. *Elements of Dual Scaling: An introduction to practical data analysis*. Psychology Press, 1994.

[47] Nishisato, S. Gleaning in the field of dual scaling. *Psychometrika 61*, 4 (1996), 559–599.

[48] Omiecinski, E. R. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering 15*, 1 (2003), 57–69.

[49] Park, J. S., Chen, M.-S., Yu, P. S. An effective hash-based algorithm for mining association rules. *SIGMOD Rec. 24*, 2 (1995), 175–186.

[50] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory* (1999), p. 398–416.

[51] Pérez-Suárez, A., Martńez-Trinidad, J. F., Carrasco-Ochoa, J. A., Medina-Pagola, J. E. *Advances in Artificial Intelligence: MICAI Part I*. Springer Berlin Heidelberg, 2013, cap. A New Overlapping Clustering Algorithm Based on Graph Theory, p. 61–72.

[52] Rissanen, J. Modeling by shortest data description. *Automatica 14* (1978), 465–471.

[53] Roux, B., Rouanet, H. *Multiple Correspondence Analysis*. SAGE Publications, 2010.

[54] Saad, Y. Numerical solution of large nonsymmetric eigenvalue problems. *Computer Physics Communications 53*, 1 (1989), 71–90.

[55] Smets, K., Vreeken, J. Slim: directly mining descriptive patterns. In *Proc. SDM* (2012), p. 236–247.

[56] Uday Kiran, R., Krishna Re, P. An improved multiple minimum support based approach to mine rare association rules. In *Computational Intelligence and Data Mining, IEEE Symposium* (2009), p. 340–347.

[57] Uno, T., Asai, T., Uchida, Y., Arimura, H. An efficient algorithm for enumerating closed patterns in transaction databases. In *Discovery Science* (2004), E. Suzuki and S. Arikawa, Eds., p. 16–31.

[58] Uno, T., Kiyomi, M., Arimura, H. Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *in Proc. FIMI* (2004).

[59] Uno, T., Kiyomi, M., Arimura, H. Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *FIMI* (2004), vol. 126 of *CEUR Workshop Proceedings*.

[60] Voigt, P., Bussche, A. v. d. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[61] Vreeken, J., van Leeuwen, M., Siebes, A. Krimp: mining itemsets that compress. *Data Min. Knowl. Discov. 23* (2011), 169–214.

[62] Wang, J., Han, J., Lu, Y., Tzvetkov, P. TFP: An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans. Knowl. Data Eng. 17*, 5 (2005), 652–664.

[63] Xiong, H., Tan, P.-N., Kumar, V. Mining strong affinity association patterns in data sets with skewed support distribution. In *Proceedings of the IEEE International Conference on Data Mining* (2003), p. 387–394.

[64] Yang, B., Huang, H. Topsil-miner: an efficient algorithm for mining top-k significant itemsets over data streams. *Knowledge and Information Systems 23*, 2 (2010), 225–242.

[65] YUEN, K. K. The two-sample trimmed t for unequal population variances. *Biometrika 61*, 1 (04 1974), 165–170.

[66] YUEN, K. K., DIXON, W. J. The approximate behaviour and performance of the two-sample trimmed t. *Biometrika 60*, 2 (08 1973), 369–374.

[67] Zaki, M. J., Parthasarathy, S., , Ogihara, M. Evaluation of sampling for data mining of association rules. In *Proceedings Seventh International Workshop on Research Issues in Data Engineering. High Performance Database Management for Large-Scale Applications* (1997), p. 42–50.

[68] Zhu, Y., Shasha, D. Statstream: Statistical monitoring of thousands of data streams in real time. In *VLDB* (2002).

APÊNDICE A – SCIM: EXPERIMENTOS COM DIFERENTES CONFIGURAÇÕES DE PARÂMETROS

Este documento apresenta a análise detalhada da escolha dos parâmetros para a comparação não tendenciosa dos algoritmos. Ele inclui as métricas calculadas por FPClose [27], Krimp [61], Slim [55], TopPI [30] e SCIM [40] em todas bases de dados usadas nos experimentos. São usadas onze bases de dados do ARM Discretizado/Normalizado LUCS-KDD [14]. Das bases de dados disponíveis, foi selecionado os que não possuem itens faltantes. No projeto houve a necessidade de remover o item *Class* de todas as bases de dados pois os experimentos não estão relacionados a procedimentos de classificação. FPClose falhou ao processar as bases de dados *Letter recognition*, *mFeat*, *Pen digits*, *Waveform* e *Connect-4* dado que o mesmo esgotou a memória disponível do sistema. A execução da implementação do Krimp foi abortada para as bases de dados *mFeat* e *Connect-4* depois de passar mais de um dia em execução, mesmo no modo *multithreading*.

É importante enfatizar que na tese não se discute os resultados alcançados pelo Krimp uma vez que a análise mostra que ela é superada pelo seu concorrente mais próximo, o algoritmo Slim.

A Tabela A.1 resume a organização geral deste apêndice em relação à análise do comportamento dos algoritmos TopPI e SCIM sob diferentes parametrizações. Mais especificamente, as tabelas e figuras referenciadas pelas colunas de dois a cinco da Tabela A.1 mostram a distribuição dos valores médios de *all-confidence* (Equação 6) calculados para os conjuntos de itens fechados recuperados pelo TopPI e SCIM, usando diferentes valores de parâmetros para, respectivamente, k e dr. Não foi necessário o estudo de parâmetros para os algoritmos FPClose, Krimp e Slim. Para estes algoritmos apenas foi definido o parâmetro de suporte conjuntivo mínimo igual a 1 (Equação 3). Para calcular a média de *all-confidence*, primeiro é criado sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte de uma determinada base de dados. Em seguida,

Tabela A.1: Conjunto de tabelas e figuras mostrando o comportamento dos algoritmos TopPI e SCIM sob diferentes parâmetros.

Base de dados	**Média *All-Confidence***				**MDL**
	TopPI		SCIM		
	Tabela	Figura	Tabela	Figura	Figura
Letter recognition	A.4	A.2	A.5	A.3	A.4
mFeat	A.6	A.5	A.7	A.6	A.7
Wine	A.8	A.8	A.9	A.9	A.10
Page blocks	A.10	A.11	A.11	A.12	A.13
Pen digits	A.12	A.14	A.13	A.15	A.16
Waveform	A.14	A.17	A.15	A.18	A.19
Ecoli	A.16	A.20	A.17	A.21	A.22
Connect-4	A.18	A.23	A.19	A.24	A.25
Tic-tac-toe	A.20	A.26	A.21	A.27	A.28
Led7	A.22	A.29	A.23	A.30	A.31
Pima	A.24	A.32	A.25	A.33	A.34

calculamos o *all-confidence* a partir dos conjuntos de itens fechados recuperados em cada partição.

Como pode ser visto nas figuras referenciadas pelas colunas três e cinco da Tabela A.1, foi comparado as distribuições das médias de *all-confidence* para a escolha dos melhores valores de parâmetro para os algoritmos TopPI e SCIM para cada base de dados. Nessas figuras, os eixos horizontais correspondem o *all-confidence* variando de 0 a 1, enquanto os eixos verticais distribuem os valores de suporte de 0 para o limite superior de cada base de dados. Espera-se que quanto melhor for o conjunto de itens fechados recuperados, mais à direita será a curva que representa o desempenho de uma técnica/parametrização. Os valores dos parâmetros escolhidos são destacados em negrito nas tabelas referenciadas pelas colunas dois e quatro da Tabela A.1.

Os números referenciados pela coluna seis da Tabela A.1 mostram o *Minimum Description Length* (MDL) calculado para os conjuntos de itens fechados recuperados por cada abordagem, sob diferentes parametrizações. Os valores de MDL são usados para calcular $L\%$. Veja Equação 2.6 para detalhes. Essa métrica não é usada para critério de escolha de parâmetros dos algoritmos TopPI e SCIM. Ela não é apresentada na discussão da Tese.

Embora não seja usada como critério de escolha para os parâmetros dos algoritmos TopPI e SCIM, é apresentado neste estudo a média de *cross-support* (Equação 7) a partir dos conjuntos de itens fechados recuperados em cada partição.

A Tabela A.2 resume os melhores resultados obtidos por cada técnica comparada em cada base de dados. A primeira coluna contém a informação da base de estudo junto com as informações de parâmetros escolhidos para cada algoritmo do estudo. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. As colunas três a nove mostram o número de conjuntos de itens fechados recuperados (#), os valores médios da métrica corrente (μ) e os valores de desvio padrão (σ) calculados por intervalo de suporte por todas as técnicas comparadas. A Figura A.1 ilustra as distribuições de μ para as métricas *all-confidence* e *cross-support*. As duas últimas colunas da Tabela A.2 apresentam o número total de padrões detectados, os tempos de processamento (em segundos).

Neste estudo é apresentado a significância estatística entre duas amostras independentes qualitativas/categóricas nominal, i.e., os itemsets fechados minerados por SCIM, TopPI e Slim de determinada partição de suporte, com variáveis dependentes quantitativas que são as médias das duas métricas utilizadas, *all-confidence* e *cross-support*. O teste de significância estatística não é usado como critério de escolha para os parâmetros das técnicas TopPI e SCIM. Para cada partição de suporte de uma base de dados, foi definido três testes de hipóteses nulas para cada algoritmo comparado, i.e., TopPI ou Slim, e o tipo de métrica utilizada, i.e., *all-confidence* e *cross-support*. Na primeira hipótese, a média de distribuição da métrica utilizada do algoritmo SCIM é menor ou igual a média de distribuição da métrica dos algoritmos comparados. Na segunda hipótese, a média de distribuição da métrica utilizada do algoritmo SCIM é igual a média de distribuição da métrica dos algoritmos comparados. E por último, a média de distribuição da métrica utilizada do algoritmo SCIM é maior ou igual a média de distribuição da métrica dos algoritmos comparados.

A Tabela A.3 resume os valores de significância estatística obtido por cada técnica comparada em cada base de dados. A primeira coluna contém a informação da base de dados junto com as informações de cada algoritmo a ser comparado com o algoritmo SCIM. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. A terceira coluna mostra a hipótese nula (H_0) usado no teste de siginificância estatística. As colunas quatro a dez mostram os *p-value* calculados por intervalo de suporte por todas as técnicas comparadas. Os testes apresentados são referentes a Tabela A.2.

Tabela A.2: Métricas calculadas por FPClose, Krimp, Slim, TopPI, and SCIM sobre as bases de dados da Tabela A.1.

Connect	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	1.499	0,009	0,030	4	0,343	0,168	3	0,362	0,053	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1.506	88,79
	cross-support		0,097	0,099		0,624	0,253		0,382	0,054		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
TopPI $k = 8$	all-confidence	496	0,038	0,039	123	0,221	0,050	55	0,335	0,042	35	0,486	0,047	40	0,651	0,044	44	0,801	0,029	172	0,951	0,040	965	2,87
	cross-support		0,038	0,039		0,221	0,051		0,336	0,043		0,488	0,047		0,654	0,044		0,804	0,029		0,953	0,040		
SCIM $dr = 0{,}00$	all-confidence	273	0,053	0,051	577	0,223	0,052	67	0,302	0,024	54	0,472	0,027	7	0,707	0,060	6	0,803	0,032	18	0,949	0,037	1.002	9,98
	cross-support		0,062	0,064		0,263	0,053		0,307	0,023		0,480	0,025		0,707	0,060		0,804	0,032		0,949	0,037		

Ecoli	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,55]			(0,55 , 0,69]			(0,69 , 0,83]			(0,83 , 0,97]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	436	0,029	0,030	35	0,194	0,041	21	0,359	0,029	13	0,457	0,039	22	0,634	0,081	1	0,833	0,000	2	0,913	0,084	530	0,14
	cross-support		0,102	0,084		0,280	0,032		0,552	0,105		0,604	0,009		0,674	0,081		0,857	0,000		0,915	0,082		
Krimp	all-confidence	15	0,046	0,041	3	0,189	0,057	2	0,324	0,032	2	0,485	0,065	2	0,619	0,070	0	0,000	0,000	0	0,000	0,000	24	0,13
	cross-support		0,112	0,097		0,279	0,002		0,452	0,201		0,603	0,013		0,667	0,027		0,000	0,000		0,000	0,000		
Slim	all-confidence	17	0,051	0,045	2	0,498	0,493	1	0,346	0,000	2	0,641	0,155	2	0,746	0,266	1	0,833	0,000	0	0,000	0,000	25	0,17
	cross-support		0,162	0,220		0,591	0,443		0,594	0,000		0,782	0,266		0,778	0,234		0,857	0,000		0,000	0,000		
TopPI $k = 3$	all-confidence	37	0,041	0,033	7	0,230	0,054	3	0,298	0,015	1	0,546	0,000	9	0,654	0,113	1	0,833	0,000	2	0,913	0,084	60	0,25
	cross-support		0,049	0,038		0,246	0,056		0,300	0,017		0,594	0,000		0,665	0,110		0,857	0,000		0,915	0,082		
SCIM $dr = 0{,}02$	all-confidence	34	0,040	0,035	10	0,221	0,044	2	0,306	0,006	3	0,480	0,059	12	0,643	0,098	1	0,833	0,000	2	0,913	0,084	64	0,03
	cross-support		0,063	0,051		0,256	0,048		0,310	0,000		0,594	0,000		0,680	0,092		0,857	0,000		0,915	0,082		

Led7	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,09]			(0,09 , 0,17]			(0,17 , 0,26]			(0,26 , 0,34]			(0,34 , 0,43]			(0,43 , 0,51]			(0,51 , 0,60]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	1.445	0,032	0,032	313	0,163	0,045	100	0,285	0,062	42	0,390	0,055	18	0,486	0,042	9	0,626	0,060	9	0,681	0,056	1.936	0,20
	cross-support		0,347	0,105		0,465	0,143		0,576	0,161		0,670	0,155		0,746	0,094		0,861	0,070		0,838	0,100		
Slim	all-confidence	73	0,028	0,032	4	0,137	0,009	1	0,247	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	78	0,17
	cross-support		0,348	0,096		0,333	0,094		0,512	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		

Continua na próxima página.

TopPI $k = 7$	all-confidence	0	0,000	0,000	6	0,188	0,020	18	0,321	0,074	16	0,394	0,056	9	0,509	0,030	9	0,626	0,060	9	0,681	0,056	67	0,30
	cross-support		0,000	0,000		0,245	0,016		0,444	0,144		0,534	0,102		0,723	0,117		0,861	0,070		0,838	0,100		
SCIM $dr = 0{,}10$	all-confidence	0	0,000	0,000	1	0,278	0,000	3	0,410	0,097	3	0,444	0,050	3	0,516	0,032	2	0,680	0,113	3	0,674	0,089	15	0,04
	cross-support		0,000	0,000		0,449	0,000		0,551	0,171		0,544	0,072		0,598	0,042		0,938	0,079		0,759	0,108		

Letter recognition	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,09]			(0,09 , 0,18]			(0,18 , 0,28]			(0,28 , 0,37]			(0,37 , 0,46]			(0,46 , 0,55]			(0,55 , 0,64]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	1.210	0,014	0,032	11	0,199	0,074	4	0,339	0,050	0	0,000	0,000	4	0,615	0,083	0	0,000	0,000	2	0,760	0,059	1.231	34,31
	cross-support		0,149	0,144		0,425	0,184		0,729	0,200		0,000	0,000		0,833	0,083		0,000	0,000		0,852	0,070		
TopPI $k = 7$	all-confidence	260	0,077	0,061	87	0,162	0,046	19	0,304	0,049	15	0,424	0,070	35	0,574	0,078	23	0,644	0,060	8	0,747	0,035	447	0,62
	cross-support		0,100	0,083		0,212	0,070		0,397	0,089		0,558	0,144		0,738	0,136		0,812	0,064		0,855	0,094		
SCIM $dr = 0{,}03$	all-confidence	29	0,192	0,197	4	0,235	0,183	12	0,333	0,051	21	0,451	0,068	16	0,571	0,085	2	0,762	0,144	5	0,735	0,040	89	0,48
	cross-support		0,282	0,296		0,341	0,385		0,510	0,126		0,599	0,094		0,681	0,133		0,831	0,149		0,798	0,064		

mfeat	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,11]			(0,11 , 0,23]			(0,23 , 0,34]			(0,34 , 0,46]			(0,46 , 0,57]			(0,57 , 0,69]			(0,69 , 0,80]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	4.438	0,057	0,079	242	0,404	0,171	194	0,584	0,145	131	0,703	0,123	75	0,773	0,098	33	0,862	0,050	8	0,866	0,040	5.121	10.053,94
	cross-support		0,252	0,276		0,621	0,195		0,730	0,172		0,818	0,127		0,850	0,124		0,939	0,056		0,904	0,062		
TopPI $k = 3$	all-confidence	2.631	0,052	0,045	131	0,329	0,182	237	0,427	0,136	244	0,551	0,125	190	0,675	0,111	105	0,772	0,101	29	0,835	0,063	3.567	1,35
	cross-support		0,055	0,045		0,348	0,189		0,449	0,143		0,583	0,135		0,712	0,120		0,825	0,110		0,885	0,066		
SCIM $dr = 0{,}00$	all-confidence	449	0,296	0,142	4.560	0,394	0,104	3.903	0,492	0,109	2.150	0,606	0,093	761	0,737	0,064	105	0,856	0,050	15	0,892	0,034	11.943	3.351,34
	cross-support		0,446	0,213		0,550	0,163		0,630	0,152		0,732	0,120		0,870	0,083		0,931	0,064		0,938	0,054		

Page blocks	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,29]			(0,29 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,86]			(0,86 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	179	0,003	0,004	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	535	0,955	0,024	714	0,20
	cross-support		0,004	0,005		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,970	0,014		
Slim	all-confidence	30	0,029	0,133	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	10	0,955	0,036	40	0,20
	cross-support		0,031	0,133		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,975	0,019		

Continua na próxima página.

TopPI $k = 1$	all-confidence	29	0,004	0,005	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2	0,997	0,002	31	0,29
	cross-support		0,004	0,005		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,997	0,002		
SCIM $dr = 0{,}00$	all-confidence	32	0,039	0,139	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	22	0,988	0,012	54	0,13
	cross-support		0,047	0,149		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,991	0,009		

Pen digits	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,04]			(0,04 , 0,08]			(0,08 , 0,13]			(0,13 , 0,17]			(0,17 , 0,21]			(0,21 , 0,25]			(0,25 , 0,29]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	1.125	0,016	0,023	52	0,175	0,055	22	0,265	0,067	10	0,434	0,092	7	0,487	0,130	3	0,570	0,078	1	0,587	0,000	1.220	45,24
	cross-support		0,349	0,186		0,582	0,182		0,629	0,182		0,697	0,121		0,718	0,155		0,667	0,053		0,729	0,000		
TopPI $k = 7$	all-confidence	95	0,040	0,027	46	0,151	0,043	84	0,275	0,069	98	0,366	0,086	59	0,431	0,065	14	0,521	0,050	5	0,549	0,027	401	0,55
	cross-support		0,082	0,059		0,265	0,084		0,527	0,148		0,611	0,153		0,718	0,138		0,818	0,132		0,776	0,129		
SCIM $dr = 0{,}04$	all-confidence	2	0,106	0,031	15	0,175	0,044	57	0,263	0,067	46	0,397	0,100	21	0,469	0,082	5	0,558	0,062	2	0,563	0,034	148	0,37
	cross-support		0,417	0,404		0,447	0,135		0,501	0,129		0,572	0,156		0,648	0,136		0,662	0,038		0,651	0,111		

pima	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,13]			(0,13 , 0,26]			(0,26 , 0,40]			(0,40 , 0,53]			(0,53 , 0,66]			(0,66 , 0,79]			(0,79 , 0,93]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	1.345	0,027	0,029	16	0,153	0,011	0	0,000	0,000	8	0,526	0,017	112	0,626	0,041	96	0,757	0,047	31	0,884	0,034	1.608	0,17
	cross-support		0,050	0,043		0,184	0,006		0,000	0,000		0,791	0,015		0,805	0,039		0,888	0,053		0,935	0,030		
Slim	all-confidence	47	0,029	0,029	0	0,000	0,000	0	0,000	0,000	1	0,511	0,000	3	0,618	0,051	2	0,758	0,042	2	0,911	0,015	55	0,21
	cross-support		0,066	0,131		0,000	0,000		0,000	0,000		0,783	0,000		0,798	0,026		0,864	0,114		0,956	0,043		
TopPI $k = 30$	all-confidence	593	0,026	0,024	16	0,153	0,011	0	0,000	0,000	0	0,000	0,000	8	0,683	0,004	69	0,773	0,044	31	0,884	0,034	717	0,31
	cross-support		0,035	0,029		0,184	0,006		0,000	0,000		0,000	0,000		0,787	0,004		0,880	0,060		0,935	0,030		
SCIM $dr = 0{,}03$	all-confidence	302	0,029	0,021	4	0,154	0,012	0	0,000	0,000	8	0,526	0,017	90	0,622	0,041	87	0,759	0,048	31	0,884	0,034	522	0,06
	cross-support		0,040	0,026		0,182	0,000		0,000	0,000		0,791	0,015		0,806	0,043		0,896	0,048		0,935	0,030		

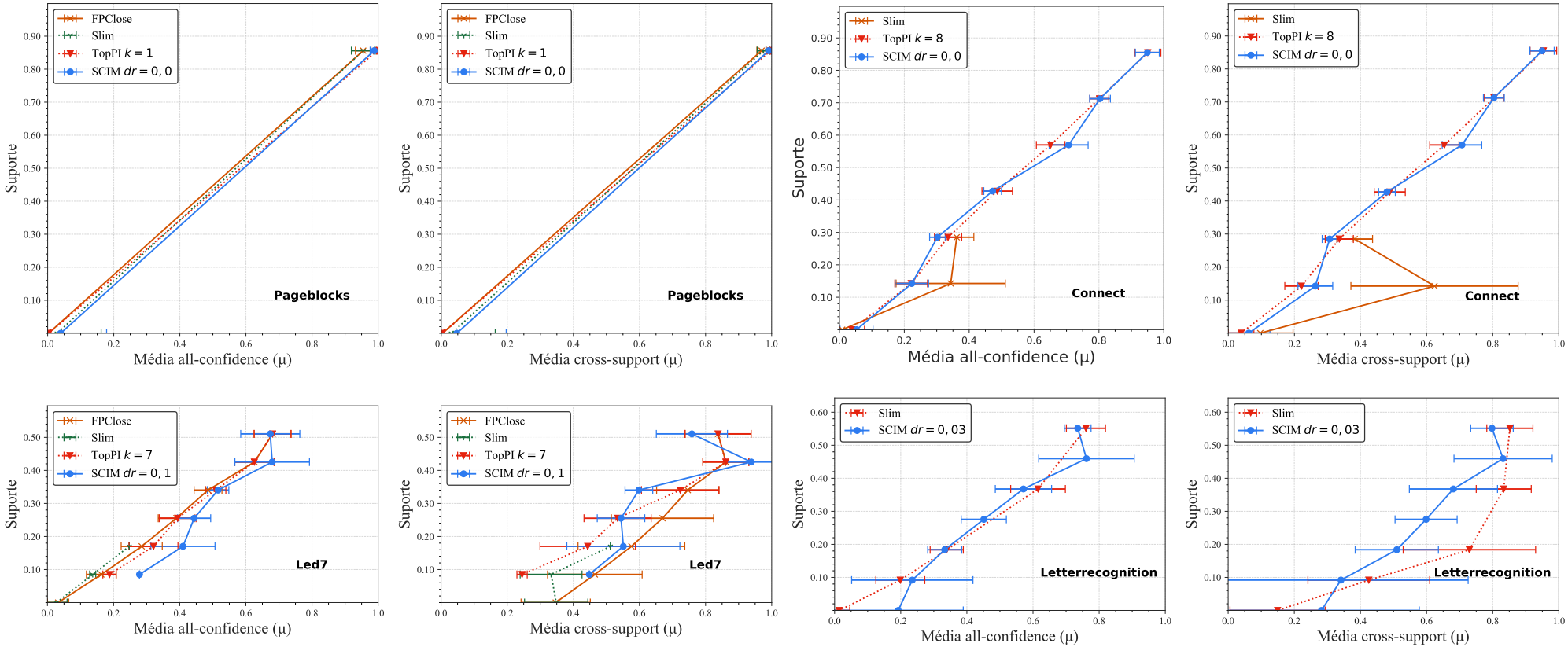
Tic-tac-toe	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,03]			(0,03 , 0,06]			(0,06 , 0,08]			(0,08 , 0,11]			(0,11 , 0,14]			(0,14 , 0,17]			(0,17 , 0,20]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	40.316	0,016	0,013	1.698	0,090	0,024	396	0,166	0,027	138	0,239	0,041	44	0,280	0,031	32	0,369	0,018	60	0,415	0,031	42.684	2,98
	cross-support		0,580	0,130		0,682	0,142		0,765	0,125		0,759	0,141		0,733	0,195		0,860	0,069		0,846	0,074		

Continua na próxima página.

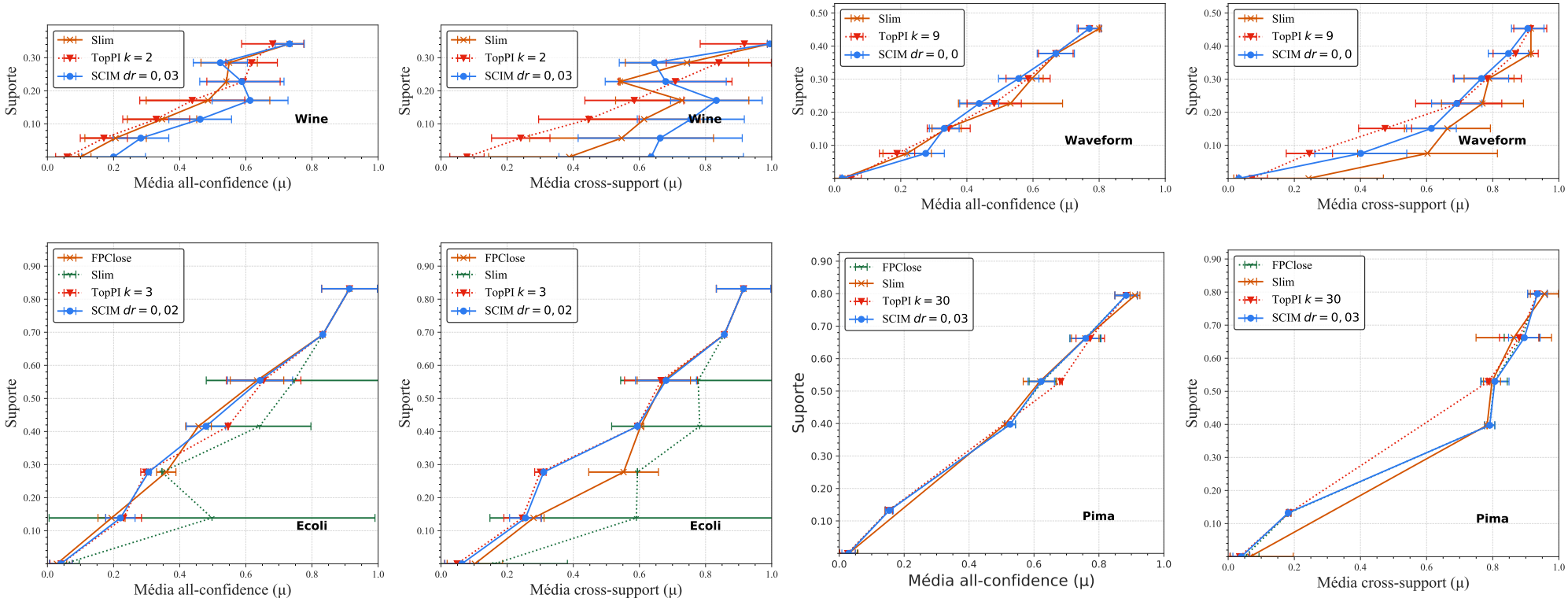
Slim	all-confidence	68	0,033	0,014	20	0,104	0,036	16	0,185	0,030	7	0,258	0,021	6	0,267	0,022	0	0,000	0,000	8	0,442	0,031	125	0,23
	cross-support		0,551	0,146		0,745	0,108		0,712	0,183		0,693	0,164		0,726	0,203		0,000	0,000		0,821	0,058		
TopPI $k = 15$	all-confidence	0	0,000	0,000	0	0,000	0,000	46	0,204	0,022	76	0,267	0,028	36	0,284	0,032	32	0,369	0,018	60	0,415	0,031	250	0,31
	cross-support		0,000	0,000		0,000	0,000		0,549	0,083		0,735	0,177		0,695	0,197		0,860	0,069		0,846	0,074		
SCIM $dr = 0{,}37$	all-confidence	142	0,034	0,022	40	0,091	0,018	54	0,174	0,022	48	0,206	0,022	6	0,284	0,012	2	0,364	0,000	38	0,429	0,029	330	0,06
	cross-support		0,709	0,112		0,660	0,142		0,817	0,057		0,692	0,128		0,581	0,027		0,801	0,000		0,835	0,065		

Waveform	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,08]			(0,08 , 0,15]			(0,15 , 0,23]			(0,23 , 0,30]			(0,30 , 0,38]			(0,38 , 0,45]			(0,45 , 0,53]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
Slim	all-confidence	675	0,024	0,032	23	0,219	0,073	10	0,339	0,044	3	0,532	0,157	4	0,597	0,034	1	0,671	0,000	1	0,800	0,000	717	8,50
	cross-support		0,243	0,226		0,603	0,211		0,663	0,130		0,769	0,124		0,789	0,075		0,916	0,000		0,916	0,000		
TopPI $k = 9$	all-confidence	413	0,049	0,032	54	0,189	0,053	74	0,345	0,065	52	0,483	0,082	70	0,585	0,067	57	0,669	0,052	14	0,771	0,034	734	0,44
	cross-support		0,072	0,047		0,246	0,070		0,474	0,080		0,697	0,130		0,784	0,103		0,869	0,068		0,914	0,050		
SCIM $dr = 0{,}00$	all-confidence	1	0,023	0,000	13	0,276	0,056	289	0,332	0,045	255	0,437	0,059	110	0,557	0,061	44	0,669	0,056	12	0,770	0,037	724	0,39
	cross-support		0,032	0,000		0,401	0,139		0,615	0,075		0,691	0,076		0,766	0,082		0,848	0,061		0,906	0,049		

Wine	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,06]			(0,06 , 0,11]			(0,11 , 0,17]			(0,17 , 0,23]			(0,23 , 0,28]			(0,28 , 0,34]			(0,34 , 0,40]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
FPClose	all-confidence	10.944	0,052	0,031	1.680	0,161	0,052	371	0,280	0,077	120	0,399	0,096	36	0,495	0,077	13	0,603	0,061	5	0,633	0,094	13.169	0,42
	cross-support		0,361	0,161		0,508	0,160		0,614	0,160		0,723	0,155		0,792	0,137		0,851	0,134		0,875	0,111		
Slim	all-confidence	18	0,100	0,032	14	0,206	0,093	10	0,347	0,105	7	0,486	0,187	1	0,541	0,000	3	0,550	0,085	2	0,732	0,044	55	0,23
	cross-support		0,388	0,244		0,546	0,278		0,614	0,170		0,729	0,201		0,541	0,000		0,744	0,187		0,995	0,008		
TopPI $k = 2$	all-confidence	25	0,060	0,035	12	0,171	0,071	10	0,329	0,101	7	0,438	0,159	5	0,593	0,111	6	0,617	0,078	3	0,681	0,094	68	0,25
	cross-support		0,078	0,052		0,241	0,088		0,447	0,152		0,585	0,149		0,709	0,171		0,840	0,159		0,917	0,134		
SCIM $dr = 0{,}03$	all-confidence	6	0,200	0,096	18	0,282	0,084	18	0,462	0,095	10	0,613	0,115	4	0,588	0,127	2	0,523	0,082	2	0,732	0,044	60	0,04
	cross-support		0,636	0,279		0,663	0,248		0,755	0,162		0,833	0,139		0,680	0,183		0,645	0,106		0,995	0,008		



Continua na próxima página.



Continua na próxima página.

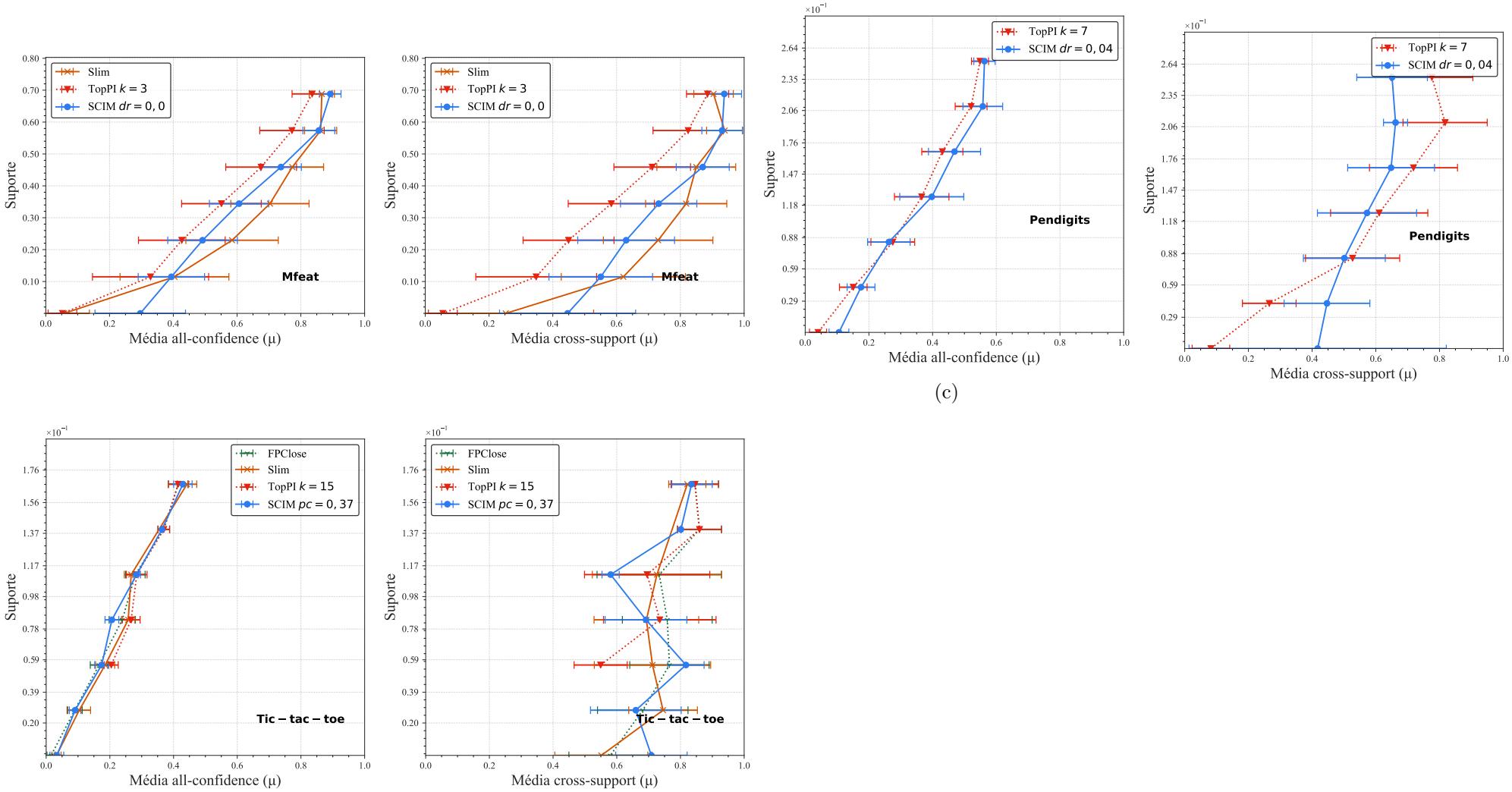


Figura A.1: Distribuições dos valores médios de *all-confidence* e de *cross-support* dos itemsets fechados recuperados por FPClose, Slim, TopPI e SCIM sobre as bases de dados da Tabela A.2. Neste estudo foi usado o melhor valor de parâmetro para cada técnica.

Tabela A.3: Significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo SCIM com os algoritmos Slim e TopPI para as bases de dados da Tabela A.1. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas na Tabela A.2.

			Partição de suporte						
Connect	***Métrica***	H_0	[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,43]	(0,43 , 0,57]	(0,57 , 0,71]	(0,71 , 0,85]	(0,85 , 1,00]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,000	0,961	0,998	-	-	-	-
		$\mu_{SCIM} = \mu$	0,000	0,079	0,004	-	-	-	-
		$\mu_{SCIM} \geq \mu$	1,000	0,040	0,002	-	-	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	1,000	1,000	0,997	-	-	-	-
		$\mu_{SCIM} = \mu$	0,000	0,000	0,006	-	-	-	-
		$\mu_{SCIM} \geq \mu$	0,000	0,000	0,003	-	-	-	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,000	0,340	1,000	0,942	0,009	0,406	0,624
		$\mu_{SCIM} = \mu$	0,000	0,680	0,000	0,116	0,019	0,811	0,756
		$\mu_{SCIM} \geq \mu$	1,000	0,660	0,000	0,058	0,991	0,606	0,378
	cross-support	$\mu_{SCIM} \leq \mu$	0,000	0,000	1,000	0,831	0,010	0,382	0,718
		$\mu_{SCIM} = \mu$	0,000	0,000	0,000	0,338	0,020	0,764	0,567
		$\mu_{SCIM} \geq \mu$	1,000	1,000	0,000	0,169	0,991	0,629	0,284

			Partição de suporte						
Ecoli	***Métrica***	H_0	[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,42]	(0,42 , 0,55]	(0,55 , 0,69]	(0,69 , 0,83]	(0,83 , 0,97]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,819	-	-	-	-	-	-
		$\mu_{SCIM} = \mu$	0,361	-	-	-	-	-	-
		$\mu_{SCIM} \geq \mu$	0,181	-	-	-	-	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,958	-	-	-	-	-	-
		$\mu_{SCIM} = \mu$	0,085	-	-	-	-	-	-
		$\mu_{SCIM} \geq \mu$	0,042	-	-	-	-	-	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,526	0,769	-	-	0,592	-	-
		$\mu_{SCIM} = \mu$	0,948	0,524	-	-	0,817	-	-
		$\mu_{SCIM} \geq \mu$	0,474	0,262	-	-	0,408	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,097	0,357	-	-	0,368	-	-
		$\mu_{SCIM} = \mu$	0,194	0,714	-	-	0,737	-	-
		$\mu_{SCIM} \geq \mu$	0,903	0,643	-	-	0,632	-	-

			Partição de suporte						
Led7	***Métrica***	H_0	[0,00 , 0,09]	(0,09 , 0,17]	(0,17 , 0,26]	(0,26 , 0,34]	(0,34 , 0,43]	(0,43 , 0,51]	(0,51 , 0,60]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	-	-	-	-	-	-
		$\mu_{SCIM} = \mu$	-	-	-	-	-	-	-
		$\mu_{SCIM} \geq \mu$	-	-	-	-	-	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	-	-	-	-	-	-	-
		$\mu_{SCIM} = \mu$	-	-	-	-	-	-	-
		$\mu_{SCIM} \geq \mu$	-	-	-	-	-	-	-

Continua na próxima página.

TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	-	0,035	0,089	0,390	-	0,742
		$\mu_{SCIM} = \mu$	-	-	0,070	0,179	0,780	-	0,642
		$\mu_{SCIM} \geq \mu$	-	-	0,972	0,927	0,679	-	0,321
	cross-support	$\mu_{SCIM} \leq \mu$	-	-	0,127	0,411	0,961	-	0,918
		$\mu_{SCIM} = \mu$	-	-	0,255	0,822	0,114	-	0,227
		$\mu_{SCIM} \geq \mu$	-	-	0,873	0,632	0,057	-	0,113

			Partição de suporte						
Letter recognition	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,09]	(0,09 , 0,18]	(0,18 , 0,28]	(0,28 , 0,37]	(0,37 , 0,46]	(0,46 , 0,55]	(0,55 , 0,64]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,000	0,362	0,404	-	0,828	-	-
		$\mu_{SCIM} = \mu$	0,000	0,724	0,808	-	0,395	-	-
		$\mu_{SCIM} \geq \mu$	1,000	0,638	0,642	-	0,197	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,011	0,931	0,979	-	0,982	-	-
		$\mu_{SCIM} = \mu$	0,023	0,177	0,055	-	0,046	-	-
		$\mu_{SCIM} \geq \mu$	0,989	0,089	0,027	-	0,023	-	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,002	0,243	0,064	0,124	0,627	-	0,769
		$\mu_{SCIM} = \mu$	0,004	0,485	0,128	0,249	0,761	-	0,555
		$\mu_{SCIM} \geq \mu$	0,998	0,757	0,941	0,876	0,380	-	0,278
	cross-support	$\mu_{SCIM} \leq \mu$	0,001	0,275	0,003	0,155	0,920	-	0,881
		$\mu_{SCIM} = \mu$	0,003	0,550	0,006	0,311	0,167	-	0,302
		$\mu_{SCIM} \geq \mu$	0,999	0,725	0,997	0,845	0,083	-	0,151

			Partição de suporte						
mfeat	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,11]	(0,11 , 0,23]	(0,23 , 0,34]	(0,34 , 0,46]	(0,46 , 0,57]	(0,57 , 0,69]	(0,69 , 0,80]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,000	0,809	1,000	1,000	0,999	0,722	0,043
		$\mu_{SCIM} = \mu$	0,000	0,383	0,000	0,000	0,003	0,559	0,087
		$\mu_{SCIM} \geq \mu$	1,000	0,191	0,000	0,000	0,001	0,280	0,962
	cross-support	$\mu_{SCIM} \leq \mu$	0,000	1,000	1,000	1,000	0,084	0,738	0,098
		$\mu_{SCIM} = \mu$	0,000	0,000	0,000	0,000	0,169	0,523	0,196
		$\mu_{SCIM} \geq \mu$	1,000	0,000	0,000	0,000	0,916	0,262	0,913
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,000	0,000	0,000	0,000	0,000	0,000	0,002
		$\mu_{SCIM} = \mu$	0,000	0,000	0,000	0,000	0,000	0,000	0,004
		$\mu_{SCIM} \geq \mu$	1,000	1,000	1,000	1,000	1,000	1,000	0,998
	cross-support	$\mu_{SCIM} \leq \mu$	0,000	0,000	0,000	0,000	0,000	0,000	0,006
		$\mu_{SCIM} = \mu$	0,000	0,000	0,000	0,000	0,000	0,000	0,011
		$\mu_{SCIM} \geq \mu$	1,000	1,000	1,000	1,000	1,000	1,000	0,994

			Partição de suporte						
Page blocks	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,14]	(0,14 , 0,29]	(0,29 , 0,43]	(0,43 , 0,57]	(0,57 , 0,71]	(0,71 , 0,86]	(0,86 , 1,00]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*

Continua na próxima página.

Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,381	-	-	-	-	-	0,009
		$\mu_{SCIM} = \mu$	0,763	-	-	-	-	-	0,018
		$\mu_{SCIM} \geq \mu$	0,619	-	-	-	-	-	0,991
	cross-support	$\mu_{SCIM} \leq \mu$	0,324	-	-	-	-	-	0,014
		$\mu_{SCIM} = \mu$	0,647	-	-	-	-	-	0,028
		$\mu_{SCIM} \geq \mu$	0,676	-	-	-	-	-	0,986
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,079	-	-	-	-	-	-
		$\mu_{SCIM} = \mu$	0,157	-	-	-	-	-	-
		$\mu_{SCIM} \geq \mu$	0,921	-	-	-	-	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,055	-	-	-	-	-	-
		$\mu_{SCIM} = \mu$	0,110	-	-	-	-	-	-
		$\mu_{SCIM} \geq \mu$	0,945	-	-	-	-	-	-

			Partição de suporte						
Pen digits	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,04]	(0,04 , 0,08]	(0,08 , 0,13]	(0,13 , 0,17]	(0,17 , 0,21]	(0,21 , 0,25]	(0,25 , 0,29]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	0,502	0,544	0,947	0,521	0,560	-
		$\mu_{SCIM} = \mu$	-	0,995	0,912	0,111	1,000	1,000	-
		$\mu_{SCIM} \geq \mu$	-	0,498	0,456	0,055	0,500	0,560	-
	cross-support	$\mu_{SCIM} \leq \mu$	-	0,997	0,998	0,994	0,879	0,676	-
		$\mu_{SCIM} = \mu$	-	0,007	0,004	0,012	0,265	0,879	-
		$\mu_{SCIM} \geq \mu$	-	0,004	0,002	0,006	0,132	0,440	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	0,032	0,853	0,026	0,017	0,177	-
		$\mu_{SCIM} = \mu$	-	0,064	0,295	0,052	0,034	0,353	-
		$\mu_{SCIM} \geq \mu$	-	0,968	0,147	0,974	0,983	0,846	-
	cross-support	$\mu_{SCIM} \leq \mu$	-	0,000	0,857	0,917	0,981	0,984	-
		$\mu_{SCIM} = \mu$	-	0,000	0,285	0,165	0,040	0,041	-
		$\mu_{SCIM} \geq \mu$	-	1,000	0,143	0,083	0,020	0,021	-

			Partição de suporte						
pima	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,13]	(0,13 , 0,26]	(0,26 , 0,40]	(0,40 , 0,53]	(0,53 , 0,66]	(0,66 , 0,79]	(0,79 , 0,93]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,516	-	-	-	0,465	-	-
		$\mu_{SCIM} = \mu$	0,968	-	-	-	0,931	-	-
		$\mu_{SCIM} \geq \mu$	0,484	-	-	-	0,543	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,913	-	-	-	0,368	-	-
		$\mu_{SCIM} = \mu$	0,174	-	-	-	0,736	-	-
		$\mu_{SCIM} \geq \mu$	0,087	-	-	-	0,632	-	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,024	0,335	-	-	1,000	0,972	0,503
		$\mu_{SCIM} = \mu$	0,047	0,670	-	-	0,000	0,057	1,000
		$\mu_{SCIM} \geq \mu$	0,976	0,699	-	-	0,000	0,028	0,503
	cross-support	$\mu_{SCIM} \leq \mu$	0,003	0,742	-	-	0,000	0,039	0,500
		$\mu_{SCIM} = \mu$	0,007	0,583	-	-	0,000	0,078	1,000
		$\mu_{SCIM} \geq \mu$	0,997	0,291	-	-	1,000	0,961	0,500

Continua na próxima página.

Tic-tac-toe	***Métrica***	H_0	**Partição de suporte**						
			[0,00 , 0,03]	(0,03 , 0,06]	(0,06 , 0,08]	(0,08 , 0,11]	(0,11 , 0,14]	(0,14 , 0,17]	(0,17 , 0,20]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,333	0,929	0,905	1,000	0,050	-	0,868
		$\mu_{SCIM} = \mu$	0,666	0,142	0,190	0,000	0,101	-	0,263
		$\mu_{SCIM} \geq \mu$	0,667	0,071	0,095	0,000	0,965	-	0,132
	cross-support	$\mu_{SCIM} \leq \mu$	0,000	0,981	0,005	0,384	0,929	-	0,389
		$\mu_{SCIM} = \mu$	0,000	0,040	0,010	0,768	0,195	-	0,779
		$\mu_{SCIM} \geq \mu$	1,000	0,020	0,995	0,626	0,098	-	0,622
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	-	1,000	1,000	0,513	-	0,011
		$\mu_{SCIM} = \mu$	-	-	0,000	0,000	0,975	-	0,022
		$\mu_{SCIM} \geq \mu$	-	-	0,000	0,000	0,487	-	0,989
	cross-support	$\mu_{SCIM} \leq \mu$	-	-	0,000	0,928	0,999	-	0,771
		$\mu_{SCIM} = \mu$	-	-	0,000	0,144	0,002	-	0,458
		$\mu_{SCIM} \geq \mu$	-	-	1,000	0,072	0,001	-	0,229

Waveform	***Métrica***	H_0	**Partição de suporte**						
			[0,00 , 0,08]	(0,08 , 0,15]	(0,15 , 0,23]	(0,23 , 0,30]	(0,30 , 0,38]	(0,38 , 0,45]	(0,45 , 0,53]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	0,009	0,780	0,888	0,958	-	-
		$\mu_{SCIM} = \mu$	-	0,018	0,441	0,228	0,086	-	-
		$\mu_{SCIM} \geq \mu$	-	0,992	0,221	0,114	0,043	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	-	0,998	0,923	0,924	0,773	-	-
		$\mu_{SCIM} = \mu$	-	0,004	0,154	0,153	0,464	-	-
		$\mu_{SCIM} \geq \mu$	-	0,002	0,077	0,077	0,232	-	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	-	0,000	0,953	1,000	0,999	0,552	0,510
		$\mu_{SCIM} = \mu$	-	0,000	0,094	0,000	0,001	0,902	1,000
		$\mu_{SCIM} \geq \mu$	-	1,000	0,047	0,000	0,001	0,451	0,510
	cross-support	$\mu_{SCIM} \leq \mu$	-	0,000	0,000	0,602	0,885	0,950	0,688
		$\mu_{SCIM} = \mu$	-	0,000	0,000	0,797	0,229	0,101	0,661
		$\mu_{SCIM} \geq \mu$	-	1,000	1,000	0,398	0,115	0,051	0,330

Wine	***Métrica***	H_0	**Partição de suporte**						
			[0,00 , 0,06]	(0,06 , 0,11]	(0,11 , 0,17]	(0,17 , 0,23]	(0,23 , 0,28]	(0,28 , 0,34]	(0,34 , 0,40]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
Slim	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,009	0,011	0,006	0,071	-	-	-
		$\mu_{SCIM} = \mu$	0,018	0,021	0,013	0,142	-	-	-
		$\mu_{SCIM} \geq \mu$	0,993	0,989	0,995	0,941	-	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,029	0,161	0,018	0,130	-	-	-
		$\mu_{SCIM} = \mu$	0,057	0,322	0,036	0,259	-	-	-
		$\mu_{SCIM} \geq \mu$	0,975	0,848	0,984	0,890	-	-	-
TopPI	*all-confidence*	$\mu_{SCIM} \leq \mu$	0,001	0,000	0,002	0,012	0,646	-	-
		$\mu_{SCIM} = \mu$	0,001	0,001	0,004	0,025	0,901	-	-
		$\mu_{SCIM} \geq \mu$	0,999	1,000	0,998	0,990	0,450	-	-
	cross-support	$\mu_{SCIM} \leq \mu$	0,002	0,000	0,000	0,006	0,646	-	-
		$\mu_{SCIM} = \mu$	0,004	0,000	0,000	0,012	0,901	-	-
		$\mu_{SCIM} \geq \mu$	0,998	1,000	1,000	0,995	0,450	-	-

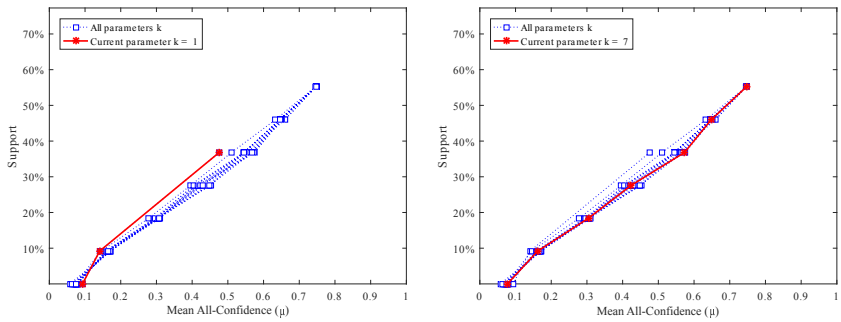
A.1 Escolha dos Parâmetro

Nesta seção, nós mostramos as distribuições de média de *all-confidence* (μ) e comprimento mínimo de descrição (MDL) dos conjuntos de itens fechados recuperados pelos algoritmos TopPI e SCIM usando valores de parâmetros diferentes.

A.1.1 Letter recognition

Tabela A.4: *Letter recognition*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,09]		(0,09 , 0,18]		(0,18 , 0,28]		(0,28 , 0,37]		(0,37 , 0,46]		(0,46 , 0,55]		(0,55 , 0,64]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	9	0,093	2	0,141	0	0,000	0	0,000	1	0,476	0	0,000	0	0,000	12	0,41
2	47	0,072	19	0,146	3	0,278	3	0,396	4	0,510	4	0,633	7	0,745	87	0,41
3	86	0,074	36	0,158	6	0,294	5	0,406	8	0,553	11	0,661	8	0,747	160	0,50
4	127	0,080	51	0,159	9	0,300	7	0,404	12	0,545	18	0,659	8	0,747	232	0,51
5	169	0,079	65	0,162	12	0,298	10	0,415	20	0,571	21	0,651	8	0,747	305	0,58
6	214	0,077	76	0,160	16	0,297	12	0,422	29	0,575	21	0,651	8	0,747	376	0,57
7	**260**	**0,077**	**87**	**0,162**	**19**	**0,304**	**15**	**0,424**	**36**	**0,573**	**22**	**0,648**	**8**	**0,747**	**447**	**0,62**
8	307	0,076	97	0,162	23	0,307	17	0,429	43	0,563	23	0,646	8	0,747	518	0,62
9	355	0,076	106	0,164	27	0,307	20	0,448	52	0,568	23	0,646	8	0,747	591	0,61
10	403	0,075	115	0,164	31	0,304	23	0,449	59	0,566	23	0,646	8	0,747	662	0,64
20	909	0,070	189	0,169	59	0,301	67	0,452	100	0,545	23	0,646	8	0,747	1.355	0,81
30	1.437	0,068	249	0,172	89	0,302	124	0,443	100	0,545	23	0,646	8	0,747	2.030	1,02
40	1.965	0,066	309	0,170	124	0,304	189	0,430	100	0,545	23	0,646	8	0,747	2.718	1,14
50	2.497	0,063	369	0,170	163	0,304	234	0,422	100	0,545	23	0,646	8	0,747	3.394	1,34
100	5.255	0,058	599	0,171	424	0,310	319	0,410	100	0,545	23	0,646	8	0,747	6.728	1,93



(a) (b)

Figura A.2: *Letter recognition*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 7$, onde essas imagens representam, por similaridade, o comportamento do $k \in \{2, 3, 4, 5, 6, 8, 9, 10, 20, 30, 40, 50, 100\}$. Veja Tabela A.4 para detalhes.

Tabela A.5: *Letter recognition*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,09]		(0,09 , 0,18]		(0,18 , 0,28]		(0,28 , 0,37]		(0,37 , 0,46]		(0,46 , 0,55]		(0,55 , 0,64]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	23	0,293	3	0,268	0	0,000	11	0,467	11	0,590	2	0,762	3	0,753	53	0,42
0,01	23	0,293	3	0,268	0	0,000	11	0,467	11	0,590	2	0,762	4	0,739	54	0,41
0,02	25	0,284	3	0,268	0	0,000	12	0,463	13	0,586	2	0,762	4	0,739	59	0,43
0,03	**29**	**0,192**	**4**	**0,235**	**12**	**0,333**	**21**	**0,451**	**16**	**0,571**	**2**	**0,762**	**5**	**0,735**	**89**	**0,48**
0,04	31	0,138	8	0,201	13	0,331	21	0,451	16	0,571	3	0,739	5	0,735	97	0,51
0,05	46	0,124	10	0,188	18	0,319	23	0,449	18	0,576	3	0,739	5	0,735	123	0,58
0,06	63	0,110	18	0,178	18	0,319	24	0,448	19	0,571	4	0,713	5	0,735	151	0,67
0,07	79	0,096	22	0,177	33	0,298	28	0,441	22	0,568	5	0,688	6	0,739	195	0,72
0,08	119	0,084	28	0,174	44	0,290	37	0,429	29	0,560	9	0,672	6	0,739	272	0,84
0,09	232	0,064	70	0,175	87	0,289	59	0,419	38	0,551	11	0,669	7	0,745	504	0,91
0,10	398	0,056	149	0,178	143	0,291	82	0,416	44	0,553	16	0,654	8	0,747	840	1,03

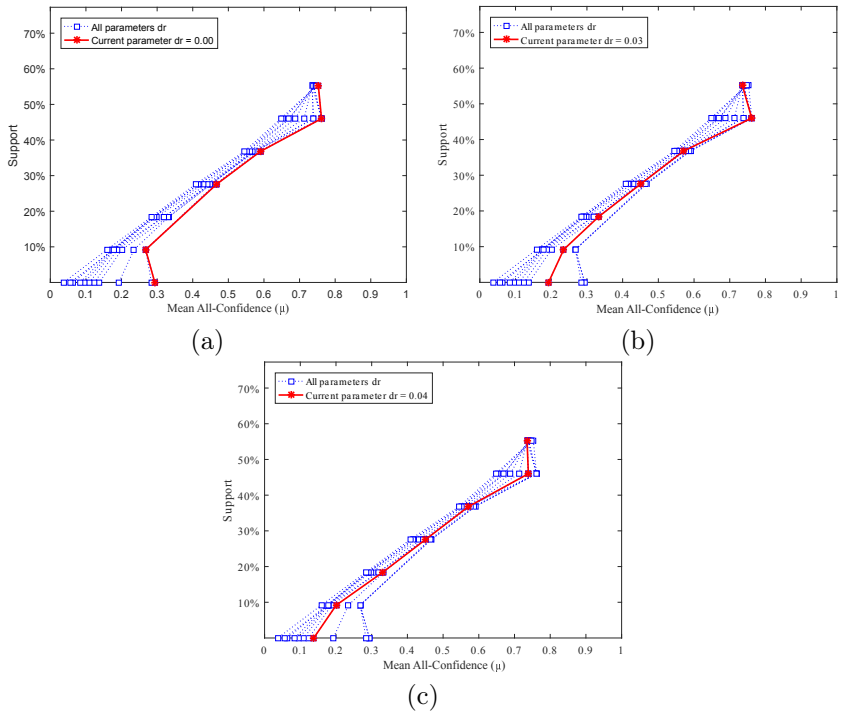


Figura A.3: *Letter recognition*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$, onde essa imagem representa, por similaridade, o comportamento do $dr \in \{0{,}01\,,0{,}02\}$, (b) com $dr = 0{,}03$, e (c) com $dr = 0{,}04$, onde estas imagens representam, por similaridade, o comportamento de $dr \in \{0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\}$. Veja Tabela A.5 para detalhes.

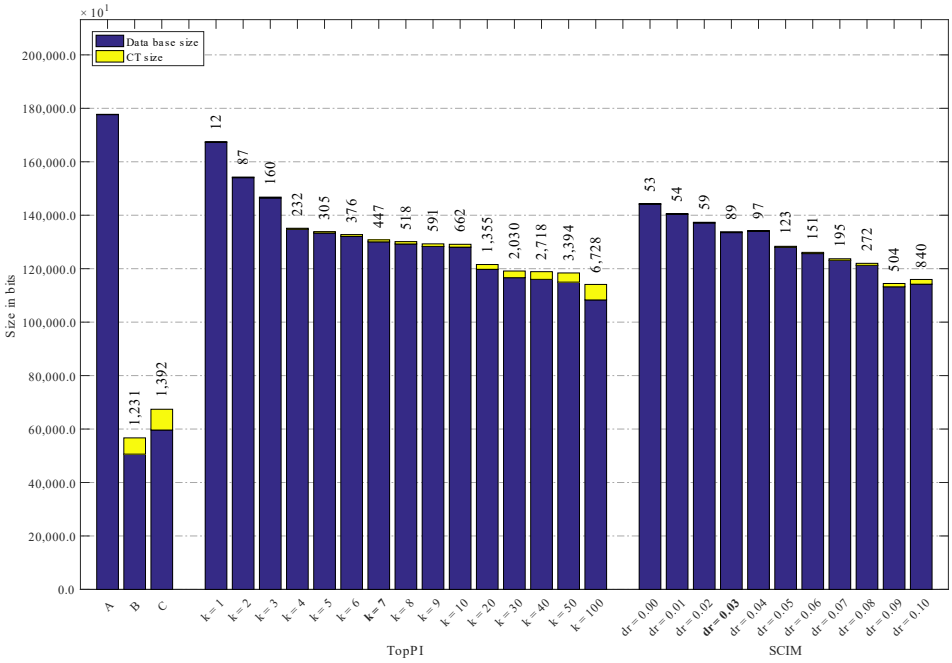
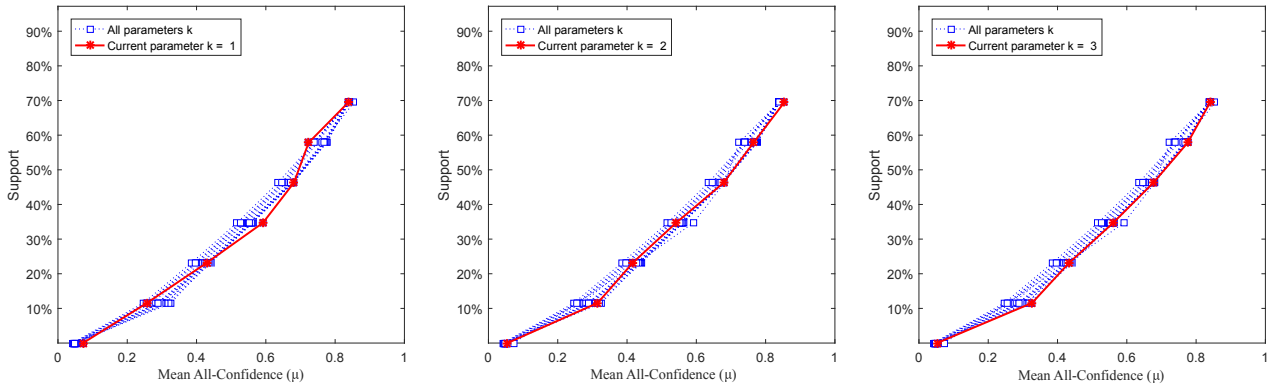


Figura A.4: *Letter recognition*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto as barras B e C representam os tamanhos relacionados aos itemsets fechados recuperados pelos algoritmos Slim e Krimp, respectivamente.

A.1.2 mFeat

Tabela A.6: *mFeat*: distribuições dos valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de Suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,12]		(0,12 , 0,23]		(0,23 , 0,35]		(0,35 , 0,46]		(0,46 , 0,58]		(0,58 , 0,69]		(0,69 , 0,81]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	234	0,072	2	0,257	11	0,429	17	0,592	17	0,681	4	0,724	2	0,838	287	0,75
2	1.433	0,055	68	0,315	125	0,416	128	0,542	102	0,679	60	0,765	15	0,854	1.931	0,89
3	**2.632**	**0,053**	**138**	**0,327**	**241**	**0,433**	**246**	**0,561**	**179**	**0,677**	**110**	**0,776**	**21**	**0,840**	**3.567**	**1,35**
4	3.831	0,052	209	0,319	360	0,441	365	0,559	252	0,679	145	0,776	22	0,840	5.184	1,59
5	5.031	0,051	281	0,309	481	0,440	485	0,562	326	0,680	183	0,773	26	0,839	6.813	1,86
6	6.231	0,050	353	0,299	607	0,434	604	0,561	396	0,677	223	0,772	28	0,840	8.442	2,22
7	7.431	0,049	426	0,297	734	0,435	722	0,558	464	0,677	254	0,769	29	0,837	10.060	2,48
8	8.631	0,049	500	0,293	862	0,433	839	0,553	535	0,674	288	0,768	29	0,837	11.684	2,83
9	9.833	0,049	576	0,289	989	0,428	954	0,553	605	0,672	314	0,767	29	0,837	13.300	3,06
10	11.035	0,048	653	0,287	1.116	0,427	1.063	0,552	673	0,669	341	0,765	29	0,837	14.910	3,39
11	12.237	0,048	730	0,285	1.246	0,426	1.183	0,550	739	0,667	363	0,764	29	0,837	16.527	3,77
20	23.077	0,046	1.425	0,272	2.463	0,413	2.230	0,539	1.335	0,660	538	0,749	29	0,837	31.097	7,01
30	35.137	0,045	2.196	0,263	3.886	0,405	3.353	0,532	1.968	0,655	633	0,743	29	0,837	47.202	11,79
40	47.197	0,044	2.987	0,258	5.360	0,399	4.424	0,529	2.603	0,650	663	0,740	29	0,837	63.263	16,75
50	59.264	0,044	3.822	0,255	6.838	0,396	5.490	0,526	3.222	0,646	673	0,739	29	0,837	79.338	22,45
100	119.670	0,042	8.323	0,247	14.278	0,386	10.806	0,517	5.678	0,635	679	0,738	29	0,837	159.463	57,39

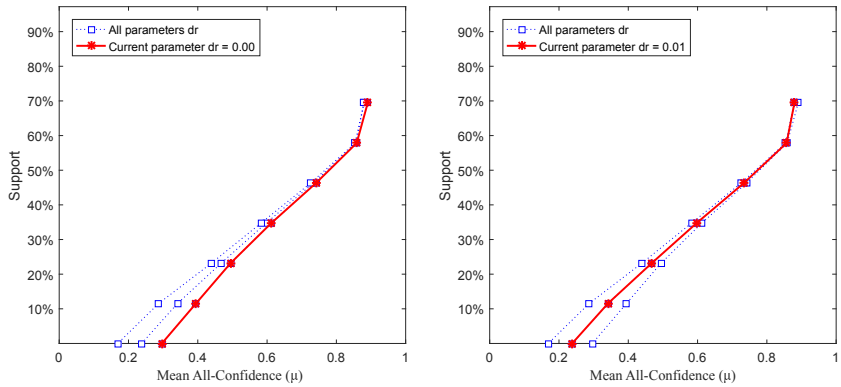


(a) (b) (c)

Figura A.5: *mFeat*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$ e (c) com $k = 3$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{4, 5, 6, 7, 8, 9, 10, 11, 20, 30, 40, 50, 100\}$. Veja Tabela A.6 para detalhes.

Tabela A.7: *mFeat*: distribuições dos valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,12]		(0,12 , 0,23]		(0,23 , 0,35]		(0,35 , 0,46]		(0,46 , 0,58]		(0,58 , 0,69]		(0,69 , 0,81]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**462**	**0,296**	**4.624**	**0,395**	**3.939**	**0,494**	**2.088**	**0,612**	**717**	**0,741**	**102**	**0,859**	**11**	**0,890**	**11.943**	**3.351,34**
0,01	1.097	0,238	9.223	0,342	6.087	0,467	2.927	0,597	845	0,734	105	0,857	11	0,880	20.295	5.730,18
0,02	3.405	0,169	31.905	0,287	12.139	0,439	4.323	0,582	993	0,726	108	0,854	11	0,880	52.884	12.943,20



(a) (b)

Figura A.6: *mFeat*: distribuições dos valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$ e (b) com $dr = 0{,}01$, onde estas imagens representam, por similaridade, o comportamento do $dr = 0{,}02$. Veja Tabela A.7 para detalhes.

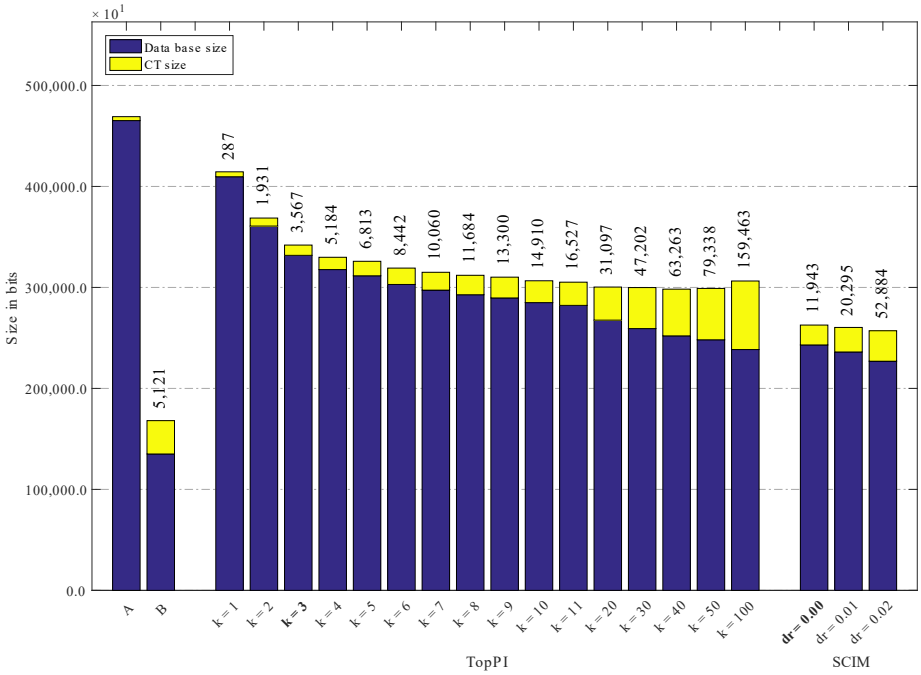
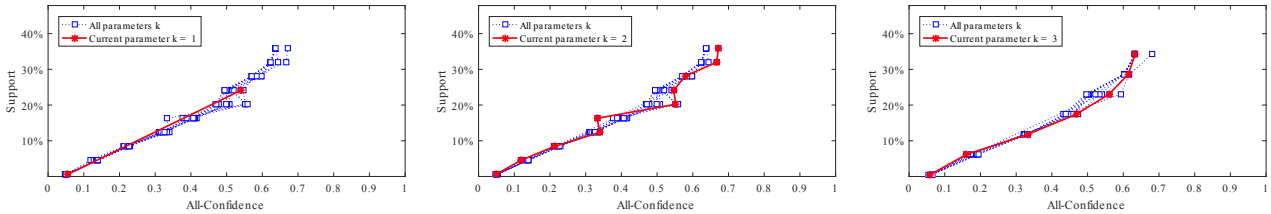


Figura A.7: *mFeat*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pela CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto que B representa o tamanho relacionado aos itemsets fechados recuperados pelo algoritmo Slim. Nessa base não foi possível executar o Krimp.

A.1.3 Wine

Tabela A.8: *Wine*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempos (s)
	[0,01 , 0,06]		(0,06 , 0,12]		(0,12 , 0,17]		(0,17 , 0,23]		(0,23 , 0,29]		(0,29 , 0,34]		(0,34 , 0,40]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	5	0,055	0	0,000	0	0,000	0	0,000	1	0,541	0	0,000	0	0,000	6	0,24
2	**25**	**0,060**	**12**	**0,171**	**10**	**0,329**	**7**	**0,438**	**5**	**0,593**	**7**	**0,629**	**2**	**0,671**	**68**	**0,25**
3	46	0,060	23	0,161	21	0,332	14	0,468	10	0,564	9	0,625	4	0,616	127	0,26
4	67	0,065	35	0,166	31	0,321	20	0,472	16	0,539	11	0,616	4	0,616	184	0,26
5	91	0,066	44	0,173	40	0,324	30	0,454	19	0,533	12	0,616	4	0,616	240	0,27
6	116	0,065	53	0,180	49	0,322	37	0,462	23	0,515	12	0,616	4	0,616	294	0,28
7	141	0,064	64	0,187	57	0,324	44	0,457	27	0,512	12	0,616	4	0,616	349	0,29
8	164	0,064	75	0,190	68	0,330	50	0,446	29	0,509	12	0,616	4	0,616	402	0,28
9	187	0,065	86	0,194	77	0,331	57	0,439	31	0,504	12	0,616	4	0,616	454	0,28
10	212	0,065	99	0,193	84	0,331	63	0,433	33	0,504	12	0,616	4	0,616	507	0,29
20	449	0,067	225	0,199	158	0,317	106	0,413	34	0,504	12	0,616	4	0,616	988	0,32
30	669	0,067	359	0,194	220	0,307	120	0,405	34	0,504	12	0,616	4	0,616	1.418	0,35

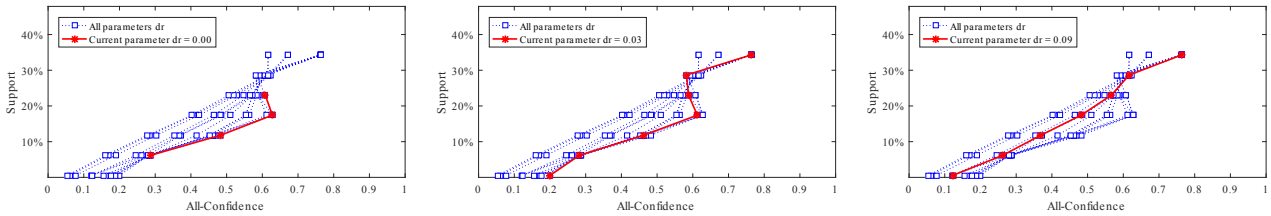


(a) (b) (c)

Figura A.8: *Wine*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$ e (c) com $k = 3$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{4\,,5\,,6\,,7\,,8\,,9\,,10\,,20\,,30\}$. Veja Tabela A.8 para detalhes.

Tabela A.9: *Wine*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,01 , 0,06]		(0,06 , 0,12]		(0,12 , 0,17]		(0,17 , 0,23]		(0,23 , 0,29]		(0,29 , 0,34]		(0,34 , 0,40]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	0	0,000	10	0,287	13	0,483	9	0,628	3	0,608	0	0,000	0	0,000	35	0,04
0,01	0	0,000	10	0,287	14	0,469	9	0,628	3	0,608	2	0,583	1	0,763	39	0,04
0,02	3	0,173	18	0,282	18	0,462	9	0,628	4	0,588	2	0,583	1	0,763	55	0,04
0,03	**6**	**0,200**	**18**	**0,282**	**18**	**0,462**	**10**	**0,613**	**4**	**0,588**	**3**	**0,582**	**1**	**0,763**	**60**	**0,04**
0,04	7	0,188	18	0,282	18	0,462	14	0,563	6	0,584	4	0,596	1	0,763	68	0,05
0,05	7	0,188	18	0,282	20	0,452	15	0,553	8	0,573	6	0,608	1	0,763	75	0,05
0,06	9	0,176	18	0,282	20	0,452	15	0,553	8	0,573	6	0,608	1	0,763	77	0,08
0,07	12	0,156	22	0,287	30	0,416	22	0,511	8	0,573	6	0,608	1	0,763	101	0,06
0,08	13	0,121	34	0,281	47	0,371	27	0,486	8	0,573	6	0,608	1	0,763	136	0,06
0,09	18	0,122	45	0,262	50	0,369	28	0,482	9	0,567	7	0,617	1	0,763	158	0,07
0,10	18	0,122	57	0,246	60	0,354	32	0,463	14	0,547	7	0,617	2	0,671	190	0,08
0,20	243	0,077	282	0,189	169	0,303	74	0,423	20	0,529	10	0,624	2	0,671	800	0,18
0,30	1.426	0,061	711	0,168	244	0,286	100	0,407	27	0,519	12	0,616	4	0,616	2.524	0,51



(a) (b) (c)

Figura A.9: *Wine*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento do $dr = 0{,}01$, (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento do $dr \in \{0{,}02\,,0{,}04\,,0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\}$, e (c) com $dr = 0{,}09$, onde esta imagem representa, por similaridade, o comportamento do $dr \in \{0{,}10\,,0{,}20\,,0{,}30\}$. Veja tabela A.9 para detalhes.

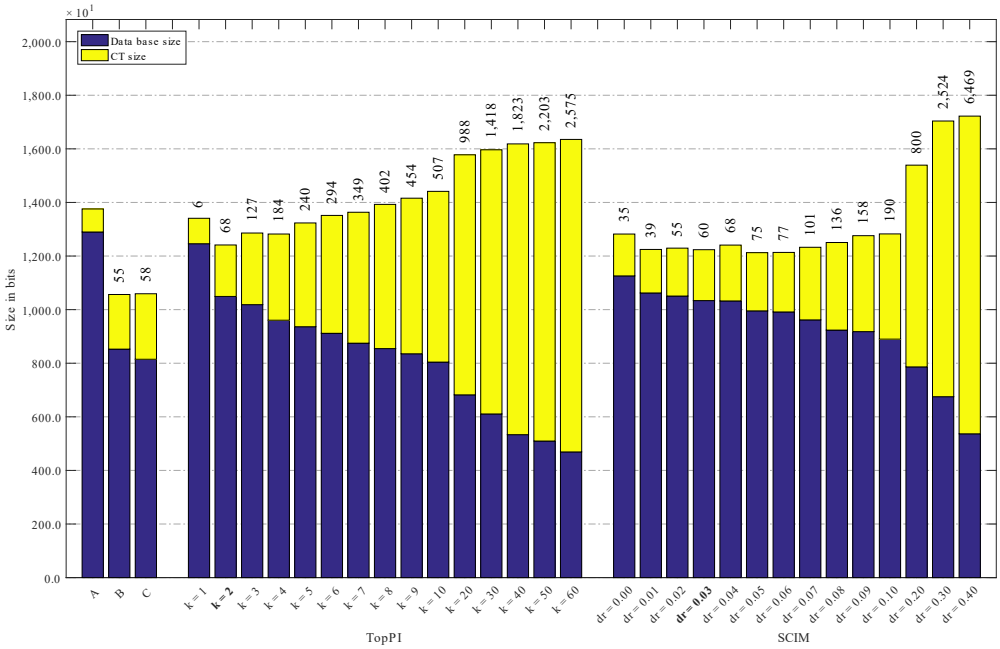


Figura A.10: *Wine*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto as barras B e C representam os tamanhos relacionados aos itemsets fechados recuperados pelos algoritmos Slim e Krimp, respectivamente.

A.1.4 Page Blocks

Tabela A.10: *Page Blocks*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1.00]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	**29**	**0,004**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**2**	**0,997**	**31**	**0,29**
2	52	0,004	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	11	0,990	63	0,30
3	73	0,004	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	19	0,989	92	0,34
4	90	0,004	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	27	0,988	117	0,32
5	103	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	34	0,988	137	0,34
6	116	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	40	0,987	156	0,37
7	122	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	46	0,986	168	0,35
8	128	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	53	0,986	181	0,36
9	133	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	58	0,986	191	0,34
10	137	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	63	0,985	200	0,35
20	170	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	112	0,982	282	0,36
30	176	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	149	0,979	325	0,41
40	179	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	183	0,977	362	0,42
50	179	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	225	0,975	404	0,47

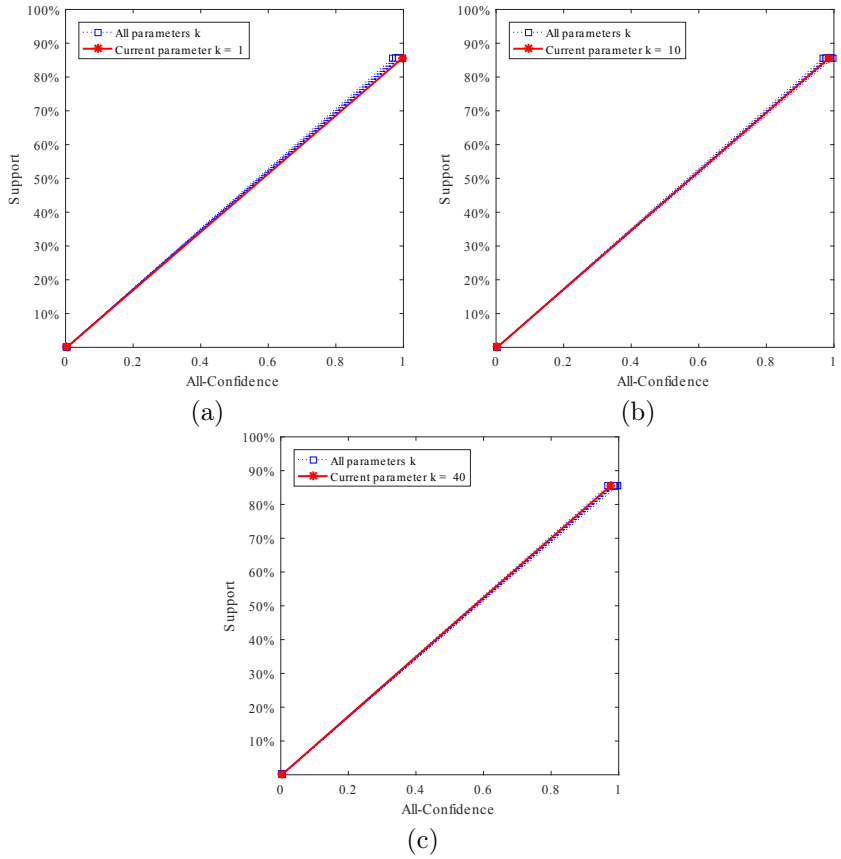


Figura A.11: *Page Blocks*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, onde esta imagem representa, por similaridade, o comportamento do $k \in \{2, 3, 4, 5, 6, 7, 8, 9\}$, (b) com $k = 10$, onde esta imagem representa, por similaridade, o comportamento do $k \in \{20, 30\}$, e (c) com $k = 40$, onde esta imagem representa, por similaridade, o comportamento do $k \in \{50, 100\}$. Veja tabela A.10 para detalhes.

Tabela A.11: *Page Blocks*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**32**	**0,039**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**22**	**0,988**	**54**	**0,13**
0,01	83	0,004	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	616	0,16
0,02	100	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	633	0,16
0,03	101	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	634	0,16
0,04	112	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	645	0,17
0,05	133	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	666	0,17
0,06	135	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	668	0,20
0,07	135	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	668	0,16
0,08	135	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	668	0,16
0,09	136	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	669	0,17
0,10	137	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	670	0,16
0,20	142	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	675	0,16
0,30	154	0,003	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	533	0,955	687	0,16

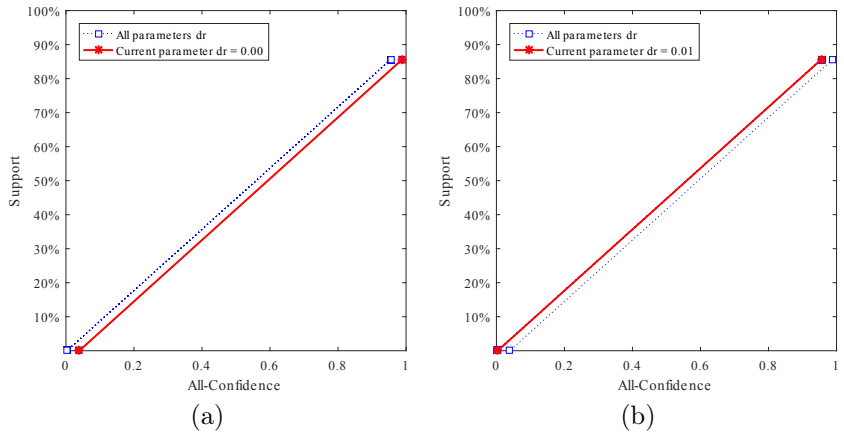


Figura A.12: *Page Blocks*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$ e (b) com $dr = 0{,}01$, onde estas imagens representam, por similaridade, o comportamento do $dr \in \{0{,}02\,, 0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}10\,, 0{,}20\,, 0{,}30\}$. Veja Tabela A.11 para detalhes.

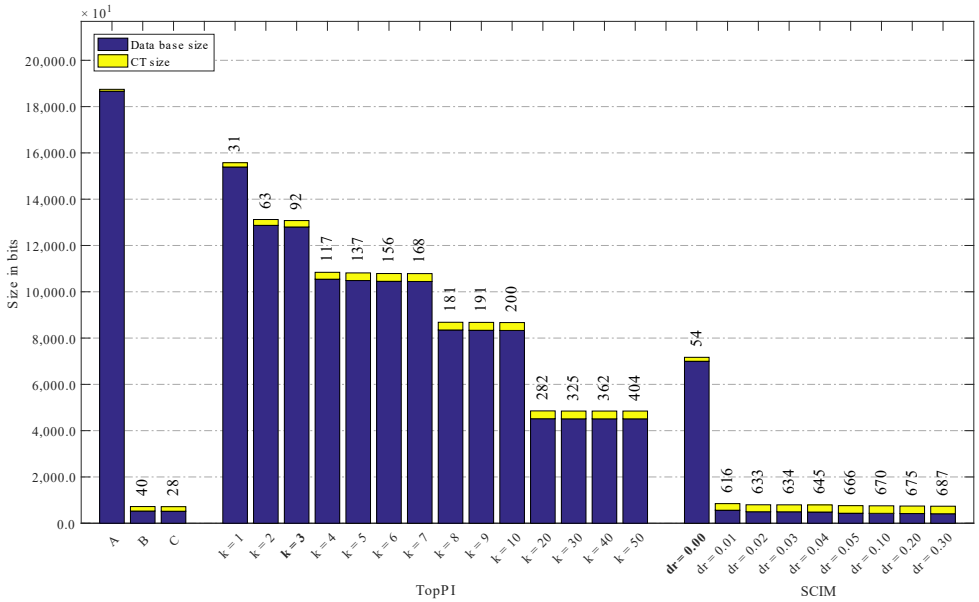


Figura A.13: *Page Blocks*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto as barras B e C representam os tamanhos relacionados aos itemsets fechados recuperados pelos algoritmos Slim e Krimp, respectivamente.

A.1.5 Pen digits

Tabela A.12: *Pen digits*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,04]		(0,04 , 0,08]		(0,08 , 0,13]		(0,13 , 0,17]		(0,17 , 0,21]		(0,21 , 0,25]		(0,25 , 0,29]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
2	14	0,036	7	0,154	8	0,250	16	0,353	14	0,451	9	0,541	5	0,549	73	0,34
3	29	0,036	15	0,148	19	0,256	33	0,369	28	0,441	12	0,529	5	0,549	141	0,41
4	44	0,035	23	0,144	32	0,263	50	0,371	40	0,440	13	0,526	5	0,549	207	0,48
5	61	0,037	30	0,151	47	0,267	67	0,368	49	0,438	13	0,526	5	0,549	272	0,47
6	78	0,040	38	0,152	65	0,273	82	0,365	55	0,435	14	0,521	5	0,549	337	0,50
7	**95**	**0,040**	**46**	**0,151**	**84**	**0,275**	**98**	**0,366**	**59**	**0,431**	**14**	**0,521**	**5**	**0,549**	**401**	**0,55**
8	112	0,040	55	0,153	107	0,282	111	0,365	61	0,430	14	0,521	5	0,549	465	0,56
9	129	0,041	65	0,153	129	0,279	120	0,364	62	0,429	14	0,521	5	0,549	524	0,57
10	146	0,041	75	0,158	153	0,280	132	0,361	63	0,428	14	0,521	5	0,549	588	0,55
11	165	0,041	85	0,163	174	0,277	140	0,360	64	0,427	14	0,521	5	0,549	647	0,57
12	184	0,041	96	0,164	198	0,277	147	0,358	64	0,427	14	0,521	5	0,549	708	0,57
13	203	0,042	107	0,167	221	0,278	151	0,356	64	0,427	14	0,521	5	0,549	765	0,56
14	222	0,042	119	0,169	239	0,277	155	0,355	64	0,427	14	0,521	5	0,549	818	0,60
15	241	0,042	132	0,170	262	0,275	160	0,354	64	0,427	14	0,521	5	0,549	878	0,57
20	336	0,042	215	0,175	363	0,267	174	0,348	64	0,427	14	0,521	5	0,549	1.171	0,62
30	546	0,042	415	0,177	502	0,260	180	0,347	64	0,427	14	0,521	5	0,549	1.726	0,80
40	766	0,043	638	0,177	611	0,254	180	0,347	64	0,427	14	0,521	5	0,549	2.278	0,94
50	986	0,042	906	0,174	668	0,250	180	0,347	64	0,427	14	0,521	5	0,549	2.823	0,93
100	2.193	0,041	2.278	0,160	745	0,245	180	0,347	64	0,427	14	0,521	5	0,549	5.479	1,20
150	3.650	0,043	3.335	0,151	748	0,245	180	0,347	64	0,427	14	0,521	5	0,549	7.996	1,31
200	5.201	0,044	4.238	0,145	748	0,245	180	0,347	64	0,427	14	0,521	5	0,549	10.450	1,45

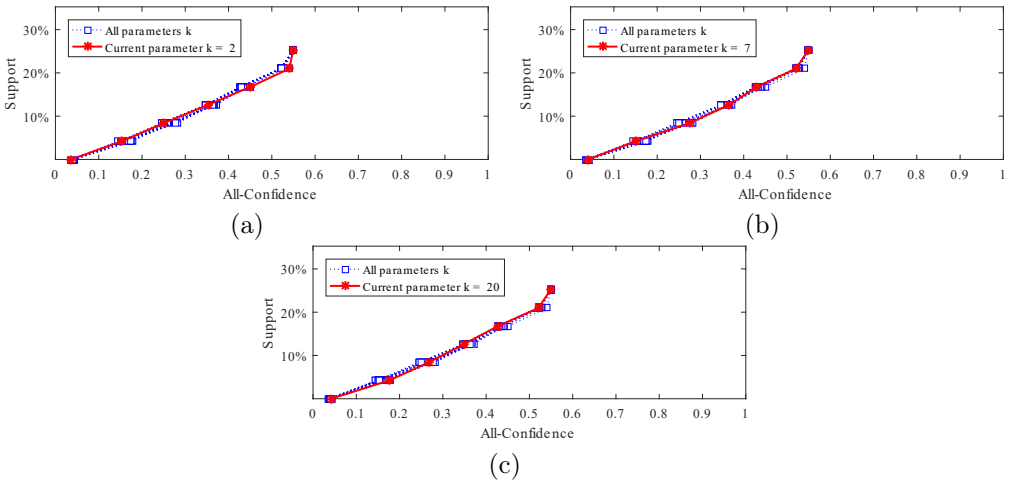


Figura A.14: *Pen digits*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 2$, onde esta imagem representa, por similaridade, o comportamento do $k \in \{3, 4, 5, 6\}$, (b) com $k = 7$ e (c) com $k = 20$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{8, 9, 10, 11, 12, 13, 14, 15, 30, 40, 50, 100, 150, 200\}$. Veja Tabela A.12 para detalhes.

Tabela A.13: *Pen digits*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,04]		(0,04 , 0,08]		(0,08 , 0,13]		(0,13 , 0,17]		(0,17 , 0,21]		(0,21 , 0,25]		(0,25 , 0,29]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	0	0,000	6	0,187	36	0,261	39	0,404	15	0,477	5	0,558	2	0,563	103	0,29
0,01	0	0,000	6	0,187	40	0,265	41	0,402	15	0,477	5	0,558	2	0,563	109	0,31
0,02	1	0,084	6	0,187	40	0,265	41	0,402	18	0,467	5	0,558	2	0,563	113	0,31
0,03	2	0,106	7	0,187	48	0,262	43	0,400	19	0,466	5	0,558	2	0,563	126	0,34
0,04	**2**	**0,106**	**15**	**0,175**	**57**	**0,263**	**46**	**0,397**	**21**	**0,469**	**5**	**0,558**	**2**	**0,563**	**148**	**0,37**
0,05	5	0,094	29	0,166	66	0,265	49	0,394	25	0,455	5	0,558	2	0,563	181	0,41
0,06	15	0,073	51	0,161	79	0,261	56	0,387	25	0,455	5	0,558	4	0,558	235	0,48
0,07	79	0,056	122	0,150	102	0,256	60	0,383	26	0,454	6	0,549	4	0,558	399	0,59
0,08	208	0,053	211	0,146	130	0,253	68	0,382	29	0,447	6	0,549	4	0,558	656	0,72
0,09	548	0,051	416	0,141	178	0,251	74	0,380	31	0,445	8	0,540	4	0,558	1.259	1,07
0,10	1.749	0,043	751	0,136	230	0,248	83	0,372	35	0,440	10	0,531	4	0,558	2.862	1,76
0,11	3.898	0,036	983	0,135	264	0,247	89	0,371	38	0,440	12	0,529	4	0,558	5.288	2,78
0,12	6.523	0,032	1.258	0,133	309	0,244	100	0,364	39	0,441	13	0,526	4	0,558	8.246	4,04
0,15	44.647	0,025	2.840	0,131	431	0,246	117	0,359	46	0,435	14	0,521	4	0,558	48.099	26,99
0,20	528.030	0,011	5.106	0,129	599	0,242	145	0,351	60	0,426	14	0,521	4	0,558	533.958	1.384,26

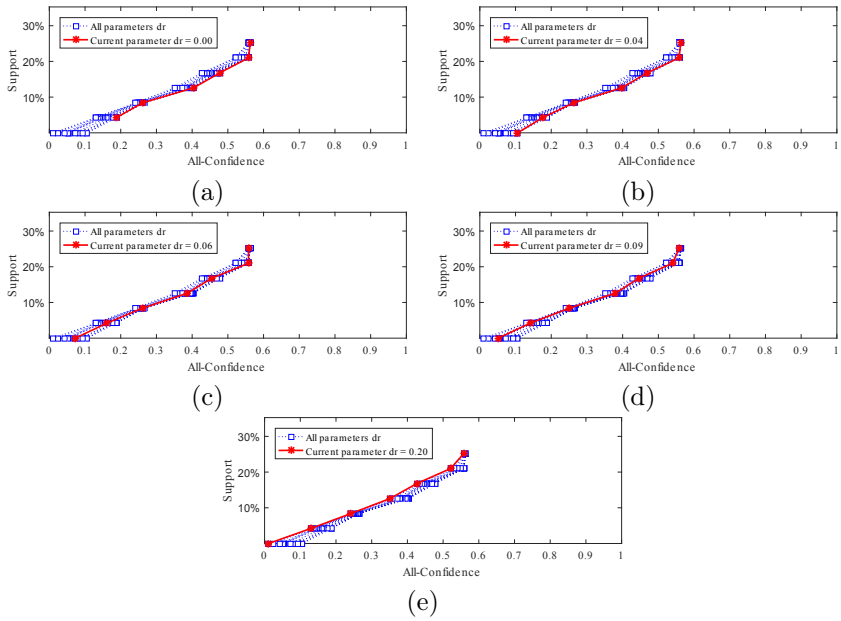


Figura A.15: *Pen digits*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento do $dr = 0{,}01$, (b) com $dr = 0{,}04$, onde esta imagem representa, por similaridade, o comportamento do $dr \in \{0{,}02\,, 0{,}03\,, 0{,}05\}$, (c) com $dr = 0{,}06$, onde esta imagem representa, por similaridade, o comportamento do $dr \in \{0{,}07\,, 0{,}08\}$, (d) com $dr = 0{,}09$ e (e) com $dr = 0{,}20$ onde esta imagem representa, por similaridade, o comportamento do $dr \in \{0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}15\}$. Veja Tabela A.13 para detalhes.

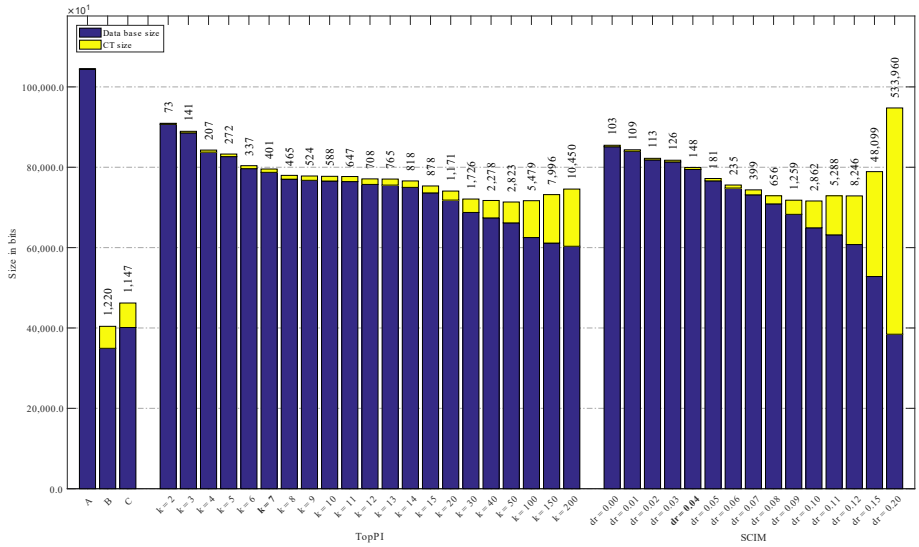


Figura A.16: *Pen digits*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto as barras B e C representam os tamanhos relacionados aos itemsets fechados recuperados pelos algoritmos Slim e Krimp, respectivamente.

A.1.6 Waveform

Tabela A.14: *Waveform*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,08]		(0,08 , 0,15]		(0,15 , 0,23]		(0,23 , 0,30]		(0,30 , 0,38]		(0,38 , 0,45]		(0,45 , 0,53]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,008	1	0,34
2	7	0,772	10	0,662	8	0,561	5	0,412	9	0,341	7	0,176	51	0,047	97	0,34
3	11	0,779	20	0,674	15	0,557	11	0,434	17	0,329	14	0,186	102	0,048	190	0,41
4	13	0,777	28	0,683	23	0,576	16	0,439	27	0,332	21	0,182	153	0,048	281	0,41
5	13	0,777	36	0,686	32	0,588	21	0,456	37	0,335	27	0,185	205	0,049	371	0,41
6	14	0,771	43	0,682	40	0,581	28	0,472	47	0,334	33	0,187	257	0,049	462	0,41
7	14	0,771	49	0,676	49	0,588	36	0,483	56	0,333	40	0,188	309	0,049	553	0,41
8	14	0,771	53	0,672	60	0,584	44	0,489	65	0,339	47	0,189	361	0,049	644	0,43
9	**14**	**0,771**	**57**	**0,669**	**70**	**0,585**	**52**	**0,483**	**74**	**0,345**	**54**	**0,189**	**413**	**0,049**	**734**	**0,44**
10	14	0,771	60	0,667	80	0,580	60	0,481	82	0,346	62	0,191	465	0,049	823	0,45
20	14	0,771	73	0,659	166	0,555	149	0,463	165	0,337	140	0,199	995	0,048	1.702	0,63
30	14	0,771	73	0,659	213	0,547	244	0,459	262	0,337	220	0,204	1.535	0,046	2.561	0,76
40	14	0,771	73	0,659	250	0,539	346	0,451	362	0,331	300	0,204	2.075	0,045	3.420	0,91
50	14	0,771	73	0,659	263	0,536	458	0,445	459	0,325	386	0,203	2.615	0,044	4.268	0,99
100	14	0,771	73	0,659	264	0,536	849	0,427	996	0,319	913	0,200	5.315	0,040	8.424	1,69
200	14	0,771	73	0,659	264	0,536	1.116	0,414	2.203	0,312	2.185	0,195	10.739	0,036	16.594	2,56
300	14	0,771	73	0,659	264	0,536	1.132	0,413	3.184	0,309	3.720	0,194	16.239	0,034	24.626	3,03
400	14	0,771	73	0,659	264	0,536	1.132	0,413	4.052	0,304	5.314	0,191	21.739	0,033	32.588	3,82

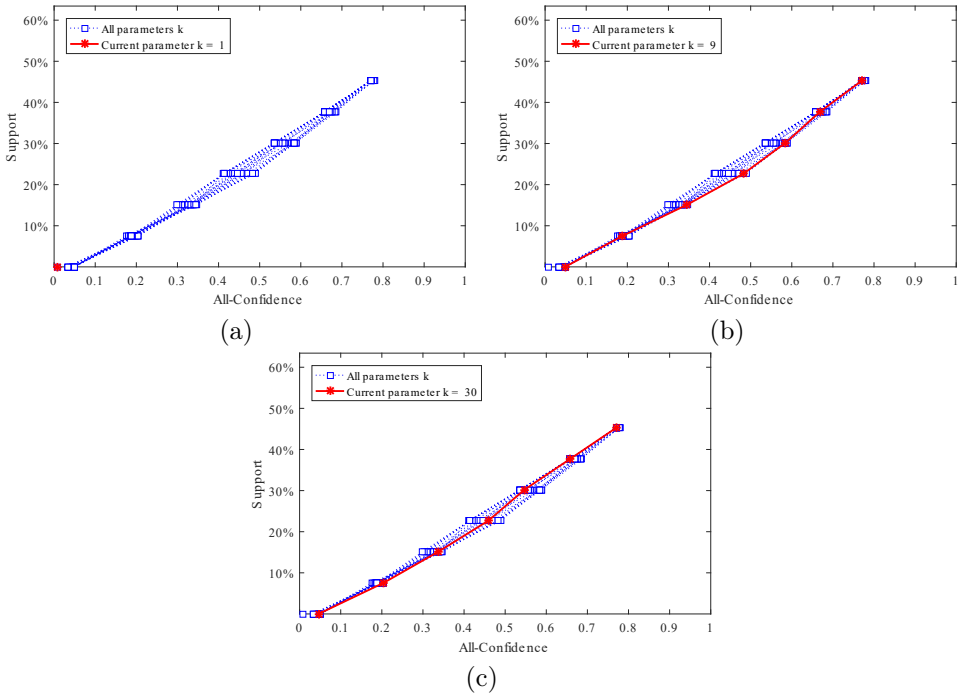


Figura A.17: *Waveform*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 9$ e (c) com $k = 30$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{2, 3, 4, 5, 6, 7, 8, 10, 20, 40, 50, 100, 200, 300, 400\}$. Veja Tabela A.14 para detalhes.

Tabela A.15: *Waveform*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,08]		(0,08 , 0,15]		(0,15 , 0,23]		(0,23 , 0,30]		(0,30 , 0,38]		(0,38 , 0,45]		(0,45 , 0,53]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**1**	**0,023**	**13**	**0,276**	**289**	**0,332**	**255**	**0,437**	**110**	**0,557**	**44**	**0,669**	**12**	**0,770**	**724**	**0,39**
0,01	4	0,051	106	0,227	482	0,317	308	0,434	118	0,554	50	0,668	12	0,770	1.080	0,48
0,02	10	0,073	268	0,209	630	0,314	326	0,433	122	0,552	50	0,668	12	0,770	1.418	0,61
0,03	38	0,097	1.334	0,195	966	0,307	350	0,432	132	0,552	54	0,667	12	0,770	2.886	1,01
0,04	996	0,105	4.743	0,178	1.487	0,301	400	0,432	144	0,550	57	0,667	12	0,770	7.839	1,87
0,05	2.109	0,101	6.884	0,175	1.913	0,298	466	0,429	155	0,548	59	0,668	12	0,770	11.598	2,97
0,06	5.623	0,092	10.200	0,173	2.449	0,296	546	0,424	161	0,548	59	0,668	12	0,770	19.050	4,44
0,07	11.623	0,085	13.809	0,171	2.811	0,294	573	0,424	163	0,549	61	0,667	12	0,770	29.052	10,53
0,08	18.029	0,082	17.259	0,169	3.092	0,293	612	0,422	170	0,547	64	0,666	12	0,770	39.238	14,02
0,09	29.957	0,077	19.143	0,168	3.323	0,293	661	0,421	186	0,545	65	0,665	12	0,770	53.347	26,26
0,10	48.357	0,064	20.782	0,168	3.616	0,292	747	0,418	207	0,540	70	0,661	12	0,770	73.791	35,62
0,11	94.723	0,048	22.897	0,167	3.902	0,291	790	0,416	217	0,537	71	0,660	12	0,770	122.612	87,22
0,12	216.440	0,036	26.695	0,165	4.174	0,290	818	0,416	224	0,536	71	0,660	12	0,770	248.434	194,74

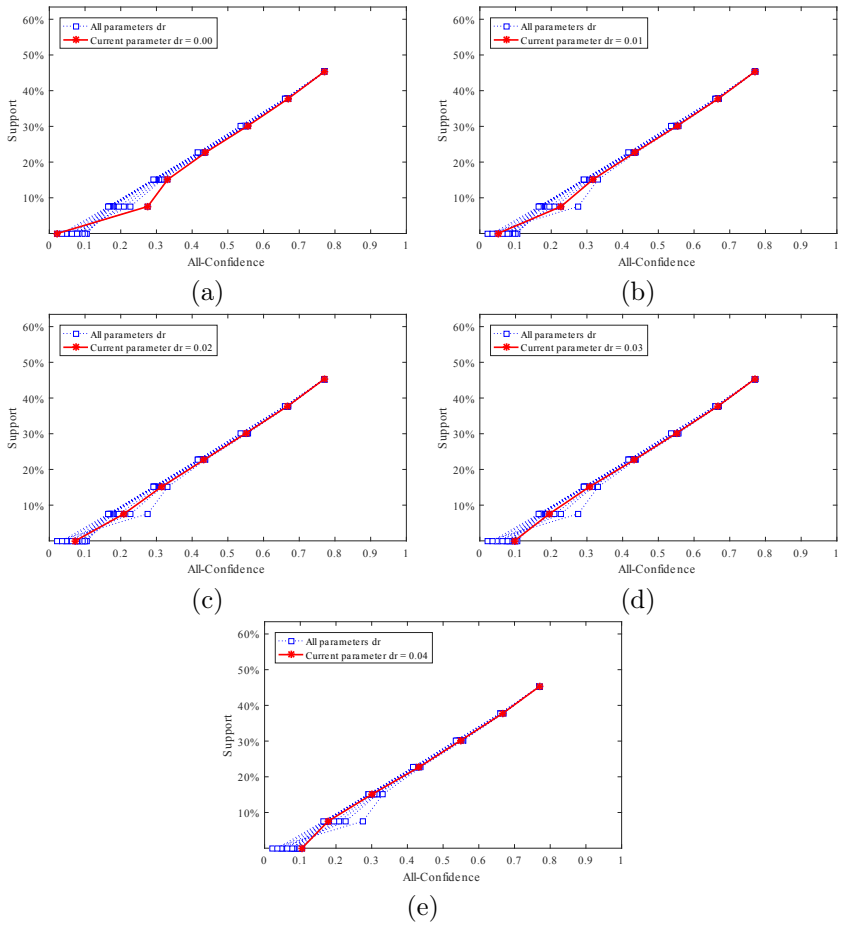


Figura A.18: *Waveform*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}01$, (c) com $dr = 0{,}02$, (d) com $dr = 0{,}03$, e (e) com $dr = 0{,}04$, onde estas imagens representam, por similaridade, o comportamento do $dr \in \{0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\}$. Veja Tabela A.15 para detalhes.

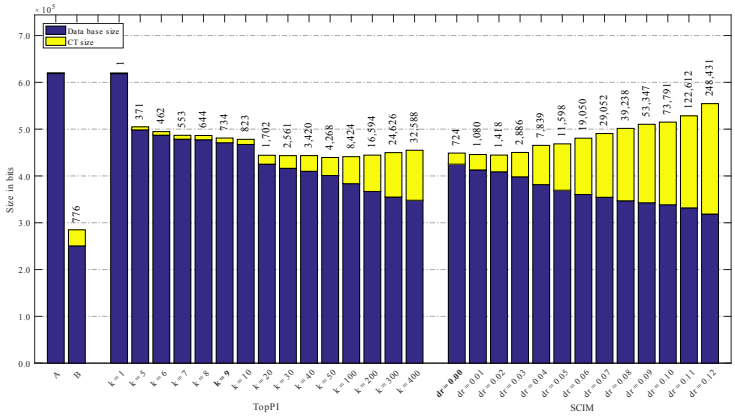


Figura A.19: *Waveform*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pela CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto que B representa o tamanho relacionado aos itemsets fechados recuperados pelo algoritmo Slim. Nessa base não foi possível executar o Krimp.

A.1.7 Ecoli

Tabela A.16: *Ecoli*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,84]		(0,84 , 0,98]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	12	0,035	3	0,241	1	0,310	0	0,000	2	0,603	0	0,000	1	0,973	19	0,26
2	24	0,040	6	0,235	2	0,306	0	0,000	6	0,619	1	0,833	2	0,913	41	0,25
3	**37**	**0,041**	**8**	**0,237**	**2**	**0,306**	**1**	**0,546**	**9**	**0,654**	**1**	**0,833**	**2**	**0,913**	**60**	**0,25**
4	50	0,041	10	0,236	2	0,306	3	0,545	12	0,656	1	0,833	2	0,913	80	0,25
5	63	0,040	13	0,233	2	0,306	4	0,524	14	0,652	1	0,833	2	0,913	99	0,29
6	76	0,039	16	0,230	2	0,306	6	0,502	15	0,648	1	0,833	2	0,913	118	0,26
7	89	0,038	18	0,227	2	0,306	7	0,492	16	0,650	1	0,833	2	0,913	135	0,26
8	102	0,036	20	0,224	2	0,306	9	0,481	17	0,646	1	0,833	2	0,913	153	0,27
9	115	0,035	23	0,218	3	0,335	10	0,477	19	0,644	1	0,833	2	0,913	173	0,26
10	127	0,035	26	0,213	4	0,348	11	0,474	21	0,637	1	0,833	2	0,913	192	0,26
20	234	0,033	36	0,196	20	0,367	13	0,467	21	0,637	1	0,833	2	0,913	327	0,30
30	320	0,033	36	0,196	21	0,366	13	0,467	21	0,637	1	0,833	2	0,913	414	0,31

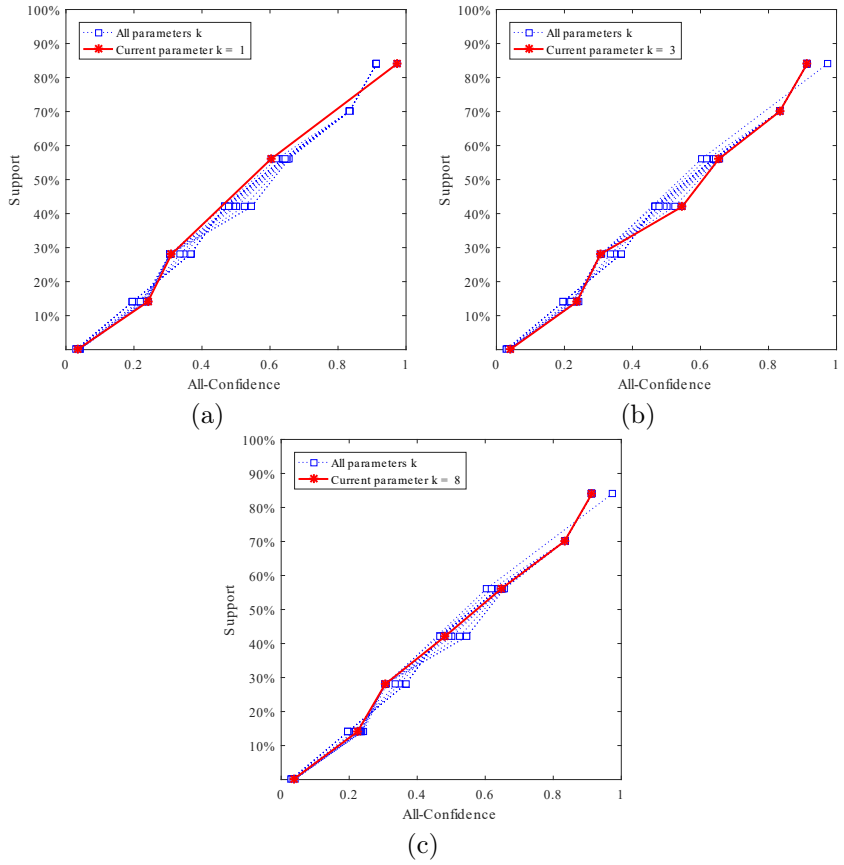


Figura A.20: *Ecoli*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 3$ e (c) com $k = 8$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{2, 4, 5, 6, 7, 9, 10, 20, 30\}$. Veja Tabela A.16 para detalhes.

Tabela A.17: *Ecoli*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,84]		(0,84 , 0,98]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	16	0,036	4	0,243	1	0,310	1	0,460	2	0,764	0	0,000	1	0,854	25	0,05
0,01	27	0,034	5	0,224	1	0,310	1	0,460	5	0,682	0	0,000	2	0,913	41	0,03
0,02	**36**	**0,041**	**10**	**0,221**	**2**	**0,306**	**3**	**0,480**	**12**	**0,643**	**1**	**0,833**	**2**	**0,913**	**66**	**0,03**
0,03	64	0,045	15	0,206	2	0,306	3	0,480	12	0,643	1	0,833	2	0,913	99	0,03
0,04	100	0,041	17	0,203	3	0,344	9	0,466	13	0,640	1	0,833	2	0,913	145	0,03
0,05	162	0,039	26	0,199	15	0,360	12	0,468	15	0,631	1	0,833	2	0,913	233	0,03
0,06	199	0,036	26	0,199	17	0,362	12	0,468	15	0,631	1	0,833	2	0,913	272	0,04
0,07	220	0,035	29	0,199	17	0,362	12	0,468	15	0,631	1	0,833	2	0,913	296	0,03
0,08	241	0,032	29	0,199	17	0,362	12	0,468	16	0,630	1	0,833	2	0,913	318	0,09
0,09	270	0,031	29	0,199	17	0,362	13	0,467	17	0,633	1	0,833	2	0,913	349	0,09
0,10	289	0,031	29	0,199	17	0,362	13	0,467	19	0,631	1	0,833	2	0,913	370	0,09
0,11	296	0,031	29	0,199	17	0,362	13	0,467	19	0,631	1	0,833	2	0,913	377	0,12
0,12	315	0,029	31	0,196	17	0,362	13	0,467	19	0,631	1	0,833	2	0,913	398	0,13
0,20	394	0,028	35	0,194	21	0,366	13	0,467	19	0,631	1	0,833	2	0,913	485	0,14

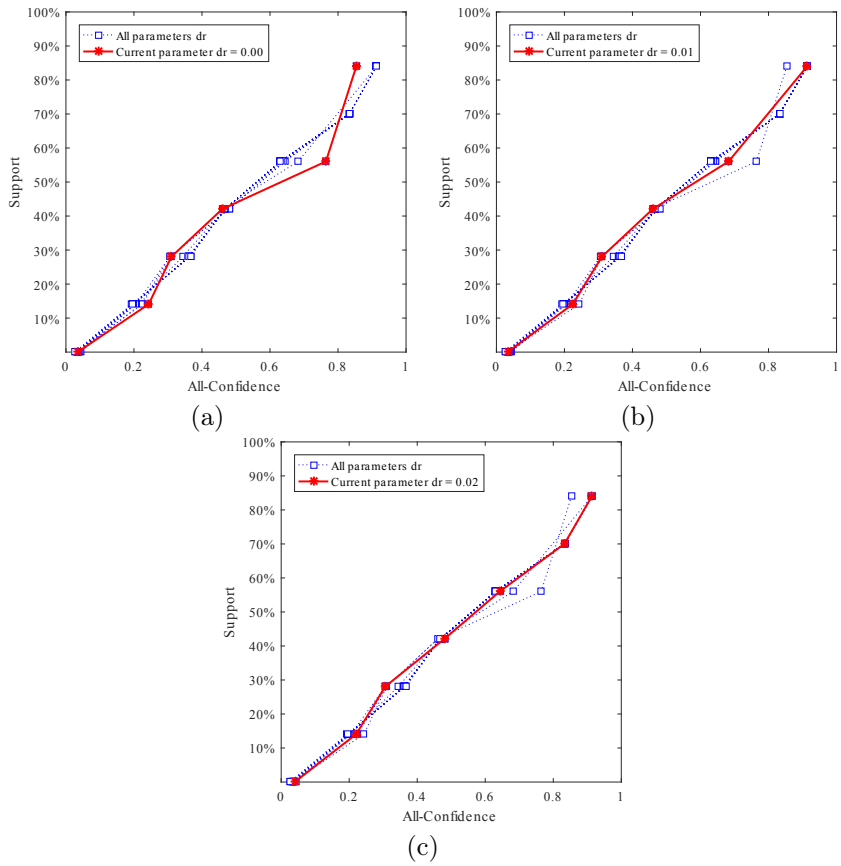


Figura A.21: *Ecoli*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM (a) com $dr = 0{,}00$, (b) com $dr = 0{,}01$, e (c) com $dr = 0{,}02$, onde estas imagens representam, por similaridade, o comportamento do $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}20\}$. Veja Tabela A.17 para detalhes.

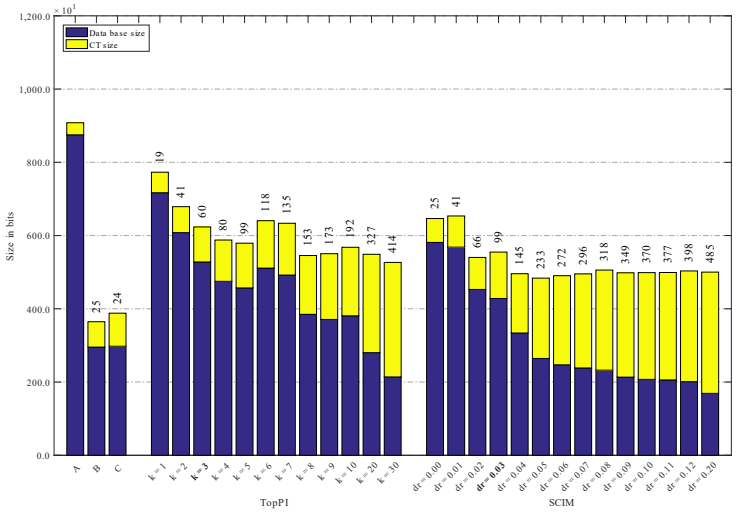
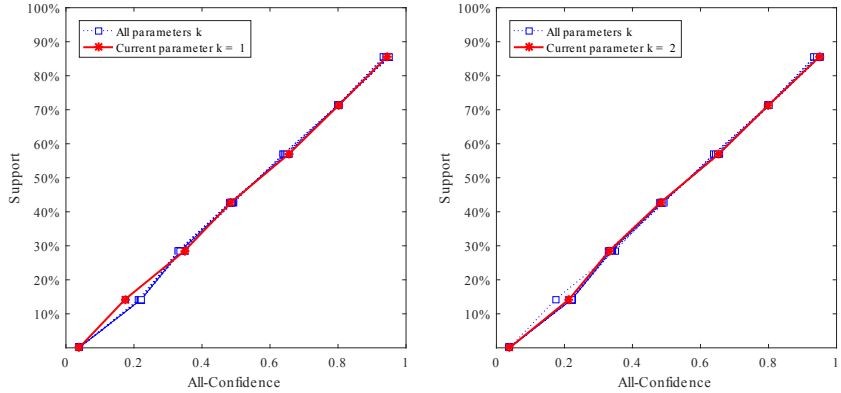


Figura A.22: *Ecoli*: valores métricos de MDL para as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão, e B e C representam os tamanhos dos itemsets fechados recuperados pelos algoritmos Slim e Krimp

A.1.8 Connect-4

Tabela A.18: *Connect-4*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 1,00]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	62	0,038	4	0,174	2	0,349	4	0,485	5	0,656	5	0,801	18	0,944	100	0,88
2	124	0,038	21	0,214	9	0,332	9	0,483	10	0,654	10	0,799	42	0,949	225	1,08
3	186	0,038	38	0,218	16	0,330	14	0,482	15	0,653	15	0,798	65	0,950	349	1,66
4	248	0,038	55	0,219	23	0,329	19	0,482	20	0,653	20	0,798	88	0,950	473	1,93
5	310	0,038	72	0,220	31	0,332	23	0,484	25	0,652	26	0,799	109	0,951	596	2,24
6	372	0,038	89	0,221	39	0,333	27	0,485	30	0,652	32	0,800	130	0,951	719	2,41
7	434	0,038	106	0,221	47	0,334	31	0,486	35	0,652	38	0,801	151	0,951	842	2,63
8	**496**	**0,038**	**123**	**0,221**	**55**	**0,335**	**35**	**0,486**	**40**	**0,651**	**44**	**0,801**	**172**	**0,951**	**965**	**2,87**
9	558	0,038	140	0,221	63	0,336	39	0,487	45	0,651	50	0,802	192	0,950	1.087	3,06
10	620	0,038	157	0,222	71	0,336	43	0,487	50	0,651	56	0,802	212	0,950	1.209	3,29
11	682	0,038	174	0,222	79	0,336	47	0,488	55	0,650	62	0,802	232	0,950	1.331	3,72
12	744	0,038	191	0,222	87	0,336	51	0,488	60	0,650	68	0,802	252	0,949	1.453	4,10
13	806	0,038	208	0,222	95	0,336	55	0,488	65	0,650	74	0,801	272	0,949	1.575	4,37
14	868	0,038	225	0,222	103	0,337	59	0,488	70	0,649	80	0,801	292	0,949	1.697	4,63
15	930	0,038	242	0,222	111	0,337	63	0,488	75	0,649	86	0,801	312	0,949	1.819	5,06
16	992	0,038	259	0,222	119	0,337	67	0,488	80	0,649	92	0,801	331	0,948	1.940	5,38
17	1.054	0,038	276	0,222	127	0,337	71	0,488	85	0,649	98	0,801	349	0,948	2.060	5,86
18	1.116	0,038	293	0,222	135	0,337	75	0,488	90	0,649	104	0,801	369	0,948	2.182	6,28
19	1.178	0,038	310	0,222	143	0,337	79	0,488	95	0,649	110	0,801	388	0,947	2.303	6,57
20	1.240	0,038	327	0,222	151	0,337	83	0,488	100	0,648	116	0,801	406	0,947	2.423	6,67
30	1.860	0,038	497	0,222	231	0,337	123	0,488	150	0,647	176	0,800	586	0,944	3.623	10,21
40	2.480	0,038	667	0,222	311	0,337	163	0,488	200	0,646	236	0,799	760	0,942	4.817	14,37
50	3.100	0,038	837	0,222	391	0,337	203	0,487	250	0,644	304	0,799	929	0,941	6.014	18,74

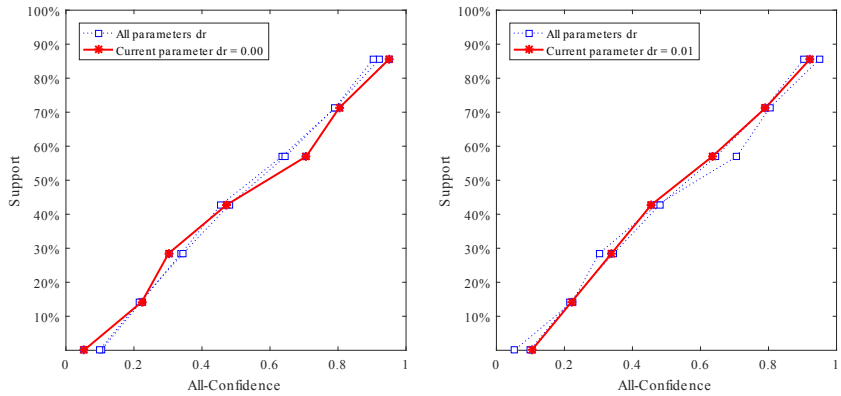


(a) (b)

Figura A.23: *Connect-4*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50\}$. Veja Tabela A.18 para detalhes.

Tabela A.19: *Connect-4*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 1,00]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**273**	**0,053**	**577**	**0,223**	**67**	**0,302**	**54**	**0,472**	**7**	**0,707**	**6**	**0,803**	**18**	**0,949**	**1.002**	**9,98**
0,01	7.754	0,104	40.946	0,221	5.585	0,339	3.470	0,455	955	0,635	1.904	0,791	3.785	0,921	64.399	78,29
0,02	104.200	0,098	199.370	0,216	20.743	0,345	9.116	0,480	7.605	0,645	14.726	0,790	7.642	0,903	363.402	3.089,55



(a) (b)

Figura A.24: *Connect-4*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}00$ e (b) com $dr = 0{,}01$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{0{,}02\}$. Veja Tabela A.19 para detalhes.

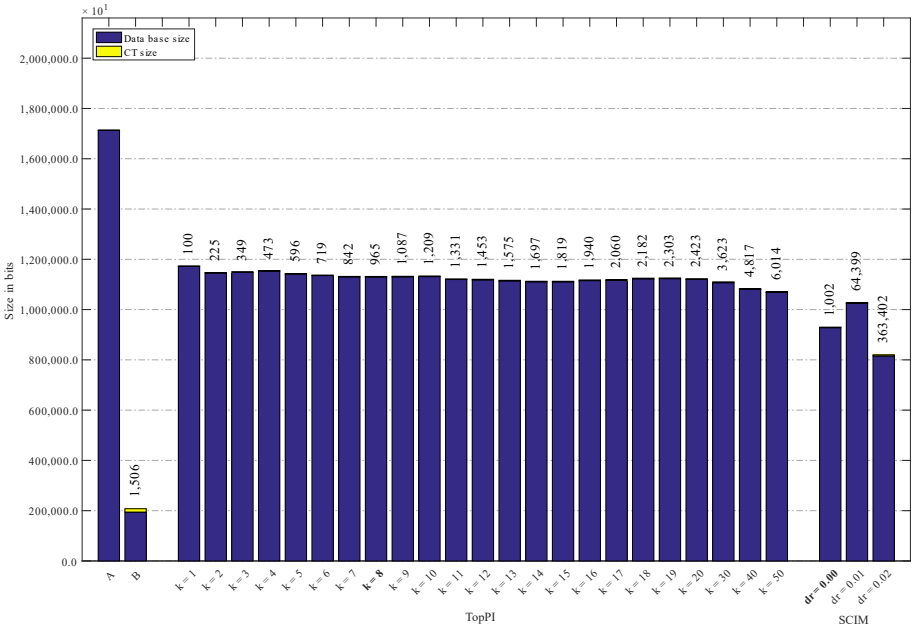


Figura A.25: *Connect-4*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pela CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto que B representa o tamanho dos itemsets fechados recuperados pelo algoritmo Slim. Nessa base não foi possível executar o Krimp.

A.1.9 Tic-tac-toe

Tabela A.20: *Tic-tac-toe*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,03]		(0,03 , 0,06]		(0,06 , 0,08]		(0,08 , 0,11]		(0,11 , 0,14]		(0,14 , 0,17]		(0,17 , 0,20]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
2	0	0,000	0	0,000	1	0,187	0	0,000	8	0,252	0	0,000	15	0,434	24	0,26
3	0	0,000	0	0,000	2	0,187	4	0,222	12	0,265	0	0,000	22	0,440	40	0,26
4	0	0,000	0	0,000	3	0,187	8	0,222	16	0,272	0	0,000	31	0,429	58	0,27
5	0	0,000	0	0,000	4	0,187	16	0,230	16	0,272	0	0,000	48	0,418	84	0,27
6	0	0,000	0	0,000	5	0,183	24	0,237	16	0,272	1	0,410	60	0,415	106	0,28
7	0	0,000	0	0,000	6	0,180	32	0,243	16	0,272	10	0,370	60	0,415	124	0,28
8	0	0,000	0	0,000	11	0,187	36	0,247	16	0,272	19	0,369	60	0,415	142	0,29
9	0	0,000	0	0,000	16	0,190	40	0,251	20	0,283	28	0,367	60	0,415	164	0,28
10	0	0,000	0	0,000	21	0,197	45	0,251	24	0,292	28	0,367	60	0,415	178	0,28
15	**0**	**0,000**	**2**	**0,161**	**44**	**0,206**	**84**	**0,266**	**28**	**0,292**	**32**	**0,369**	**60**	**0,415**	**250**	**0,31**
20	0	0,000	15	0,152	66	0,197	112	0,253	36	0,285	32	0,369	60	0,415	321	0,30
30	0	0,000	85	0,129	127	0,188	146	0,240	36	0,285	32	0,369	60	0,415	486	0,33
40	0	0,000	167	0,122	199	0,184	146	0,240	36	0,285	32	0,369	60	0,415	640	0,34
50	0	0,000	257	0,117	248	0,179	146	0,240	36	0,285	32	0,369	60	0,415	779	0,34
100	35	0,064	764	0,109	344	0,170	146	0,240	36	0,285	32	0,369	60	0,415	1.417	0,37

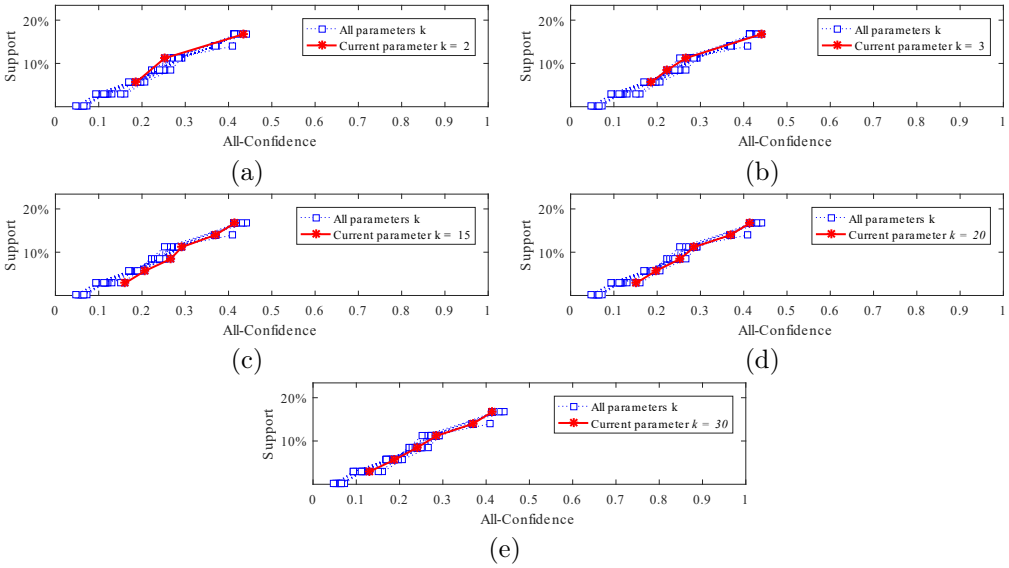


Figura A.26: *Tic-tac-toe*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 2$, (b) com $k = 3$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{4, 5, 6, 7, 8, 9, 10\}$, (c) com $k = 15$, (d) com $k = 20$ e (e) com $k = 30$, onde estas imagens representam, por similaridade, o comportamento do $k \in \{40, 50, 100\}$. Veja Tabela A.20 para detalhes.

Tabela A.21: *Tic-tac-toe*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,03]		(0,03 , 0,06]		(0,06 , 0,08]		(0,08 , 0,11]		(0,11 , 0,14]		(0,14 , 0,17]		(0,17 , 0,20]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,30	4	0,057	4	0,110	8	0,169	16	0,213	0	0,000	0	0,000	21	0,451	53	0,07
0,31	20	0,049	4	0,110	16	0,175	20	0,210	0	0,000	0	0,000	26	0,446	86	0,04
0,32	28	0,047	6	0,108	18	0,179	30	0,203	0	0,000	0	0,000	30	0,438	112	0,05
0,33	32	0,046	10	0,101	22	0,180	32	0,206	0	0,000	0	0,000	30	0,438	126	0,04
0,34	32	0,046	10	0,101	22	0,180	32	0,206	0	0,000	0	0,000	30	0,438	126	0,04
0,35	52	0,039	14	0,102	24	0,182	36	0,210	0	0,000	0	0,000	30	0,438	156	0,05
0,36	66	0,038	26	0,096	36	0,179	40	0,209	6	0,284	0	0,000	34	0,433	208	0,05
0,37	**142**	**0,034**	**40**	**0,091**	**54**	**0,174**	**48**	**0,206**	**6**	**0,284**	**2**	**0,364**	**38**	**0,429**	**330**	**0,06**
0,38	353	0,028	96	0,085	96	0,173	60	0,203	8	0,286	8	0,360	46	0,423	667	0,07
0,39	353	0,028	98	0,085	96	0,173	60	0,203	10	0,287	8	0,360	46	0,423	671	0,08
0,40	591	0,026	144	0,087	118	0,173	64	0,203	10	0,287	14	0,360	46	0,423	987	0,09
0,41	845	0,026	192	0,089	132	0,172	64	0,203	16	0,272	16	0,361	50	0,420	1.315	0,09
0,42	855	0,026	200	0,088	134	0,172	64	0,203	16	0,272	16	0,361	54	0,418	1.339	0,09
0,43	1.405	0,023	248	0,091	152	0,171	64	0,203	16	0,272	20	0,362	54	0,418	1.959	0,10
0,44	1.641	0,023	293	0,090	156	0,171	68	0,204	16	0,272	20	0,362	54	0,418	2.248	0,12

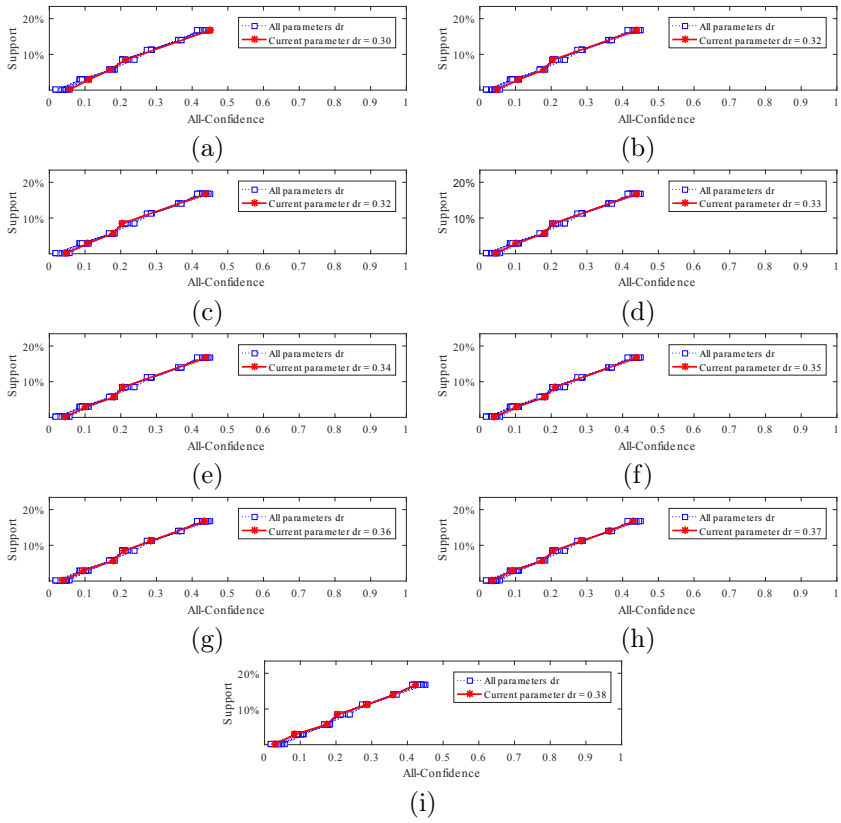


Figura A.27: *Tic-tac-toe*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM. (a) com $dr = 0{,}30$, (b) com $dr = 0{,}31$, (c) com $dr = 0{,}32$, (d) com $dr = 0{,}33$, (e) com $dr = 0{,}34$, (f) com $dr = 0{,}35$, (g) com $dr = 0{,}36$, (h) com $dr = 0{,}37$, e (i) com $dr = 0{,}38$, onde estas imagens representam, por similaridade, o comportamento do $dr \in \{0{,}39\,, 0{,}40\}$. Veja Tabela A.21 para detalhes.

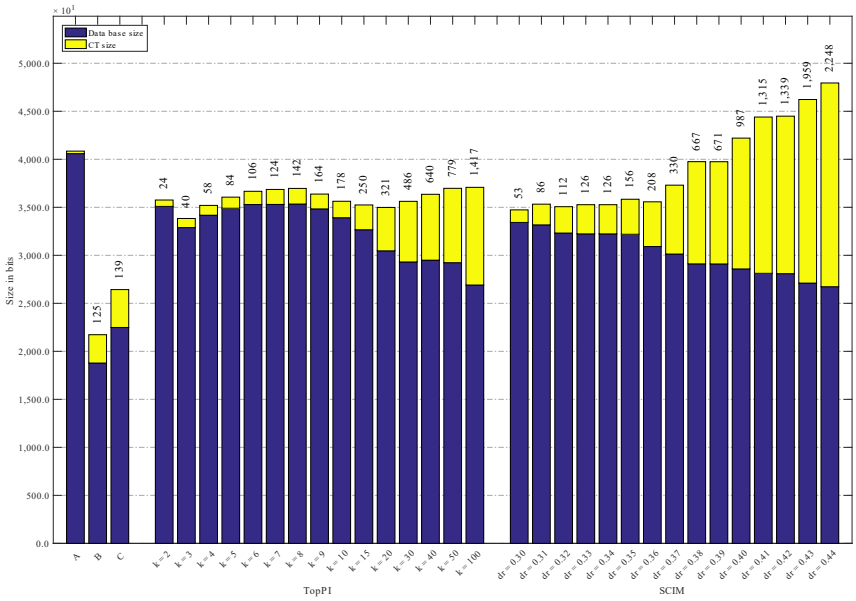


Figura A.28: *Tic-tac-toe*: valores métricos de MDL para as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão, e B e C representam os tamanhos dos itemsets fechados recuperados pelos algoritmos Slim e Krimp

A.1.10 Led7

Tabela A.22: *Led7*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,09]		(0,09 , 0,17]		(0,17 , 0,26]		(0,26 , 0,34]		(0,34 , 0,43]		(0,43 , 0,51]		(0,51 , 0,60]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
2	0	0,000	1	0,198	2	0,279	2	0,368	2	0,497	1	0,622	5	0,696	13	0,28
3	0	0,000	2	0,195	4	0,286	5	0,380	3	0,516	3	0,608	8	0,688	25	0,31
4	0	0,000	3	0,200	7	0,304	8	0,404	4	0,505	6	0,642	8	0,688	36	0,31
5	0	0,000	4	0,200	10	0,298	11	0,402	6	0,528	8	0,626	8	0,688	47	0,31
6	0	0,000	5	0,193	13	0,302	14	0,390	8	0,527	9	0,623	8	0,688	57	0,30
7	**0**	**0,000**	**6**	**0,188**	**18**	**0,321**	**16**	**0,394**	**10**	**0,523**	**9**	**0,623**	**8**	**0,688**	**67**	**0,30**
8	0	0,000	7	0,183	23	0,309	18	0,400	12	0,515	9	0,623	8	0,688	77	0,29
9	0	0,000	9	0,195	27	0,311	20	0,398	15	0,509	9	0,623	8	0,688	88	0,33
10	0	0,000	11	0,200	30	0,317	21	0,394	16	0,509	9	0,623	8	0,688	95	0,33
20	0	0,000	45	0,196	57	0,301	39	0,394	18	0,499	9	0,623	8	0,688	176	0,34
30	4	0,117	87	0,191	80	0,296	41	0,392	18	0,499	9	0,623	8	0,688	247	0,34
40	14	0,115	136	0,188	94	0,290	41	0,392	18	0,499	9	0,623	8	0,688	320	0,36
50	24	0,107	185	0,181	100	0,287	41	0,392	18	0,499	9	0,623	8	0,688	385	0,37
100	177	0,092	300	0,166	100	0,287	41	0,392	18	0,499	9	0,623	8	0,688	653	0,40
200	595	0,062	310	0,164	100	0,287	41	0,392	18	0,499	9	0,623	8	0,688	1.081	0,40
300	900	0,048	310	0,164	100	0,287	41	0,392	18	0,499	9	0,623	8	0,688	1.386	0,40
400	1.146	0,040	310	0,164	100	0,287	41	0,392	18	0,499	9	0,623	8	0,688	1.632	0,41
500	1.360	0,035	310	0,164	100	0,287	41	0,392	18	0,499	9	0,623	8	0,688	1.846	0,42

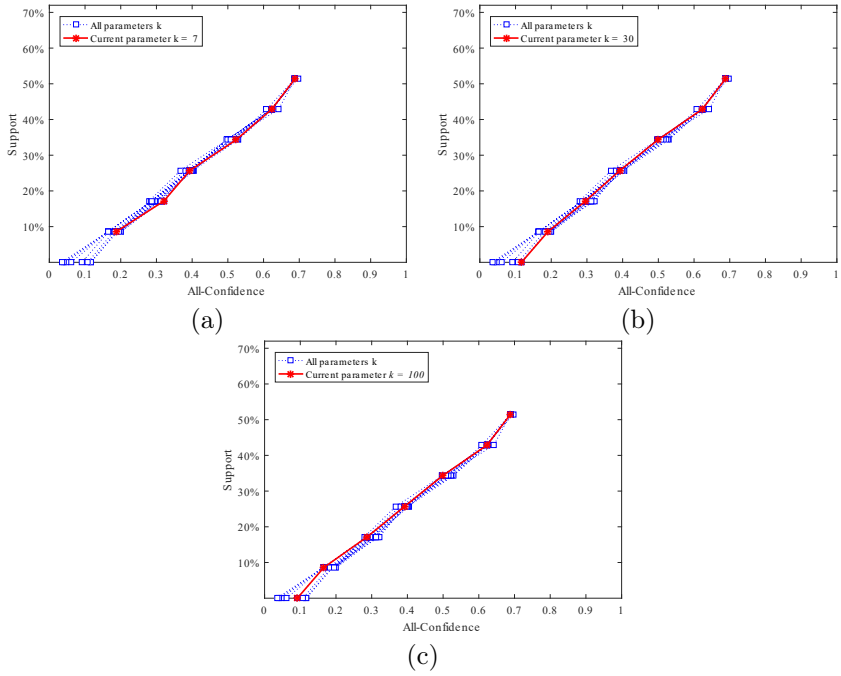


Figura A.29: *Led7*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI (a) com $k = 7$, (b) com $k = 30$ e (c) com $k = 100$ onde estas imagens representam, por similaridade, o comportamento do $k \in \{2, 3, 4, 5, 6, 8, 9, 10, 20, 40, 50, 200, 300, 400, 500\}$. Veja tabela A.22.

Tabela A.23: *Led7*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,09]		(0,09 , 0,17]		(0,17 , 0,26]		(0,26 , 0,34]		(0,34 , 0,43]		(0,43 , 0,51]		(0,51 , 0,60]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	0	0,000	1	0,278	2	0,439	2	0,469	3	0,516	0	0,000	1	0,777	9	0,07
0,10	**0**	**0,000**	**1**	**0,278**	**3**	**0,410**	**3**	**0,444**	**3**	**0,516**	**3**	**0,660**	**2**	**0,700**	**15**	**0,04**
0,20	0	0,000	10	0,212	12	0,302	3	0,444	3	0,516	3	0,660	3	0,698	34	0,06
0,30	0	0,000	30	0,196	23	0,299	9	0,407	7	0,522	4	0,661	3	0,698	76	0,05
0,40	7	0,103	38	0,201	30	0,309	10	0,401	11	0,499	6	0,631	5	0,691	107	0,05
0,50	7	0,103	49	0,200	34	0,308	12	0,387	13	0,492	7	0,628	5	0,691	127	0,06
0,60	14	0,098	74	0,192	53	0,296	21	0,390	15	0,494	8	0,625	7	0,697	192	0,06
0,70	102	0,082	139	0,176	70	0,289	29	0,382	16	0,498	8	0,625	7	0,697	371	0,06
0,80	145	0,073	165	0,174	79	0,289	31	0,387	16	0,498	9	0,623	8	0,688	453	0,14
0,90	214	0,065	187	0,170	81	0,288	32	0,386	16	0,498	9	0,623	8	0,688	547	0,07
1,00	750	0,039	249	0,164	87	0,286	36	0,389	17	0,498	9	0,623	8	0,688	1.156	0,09

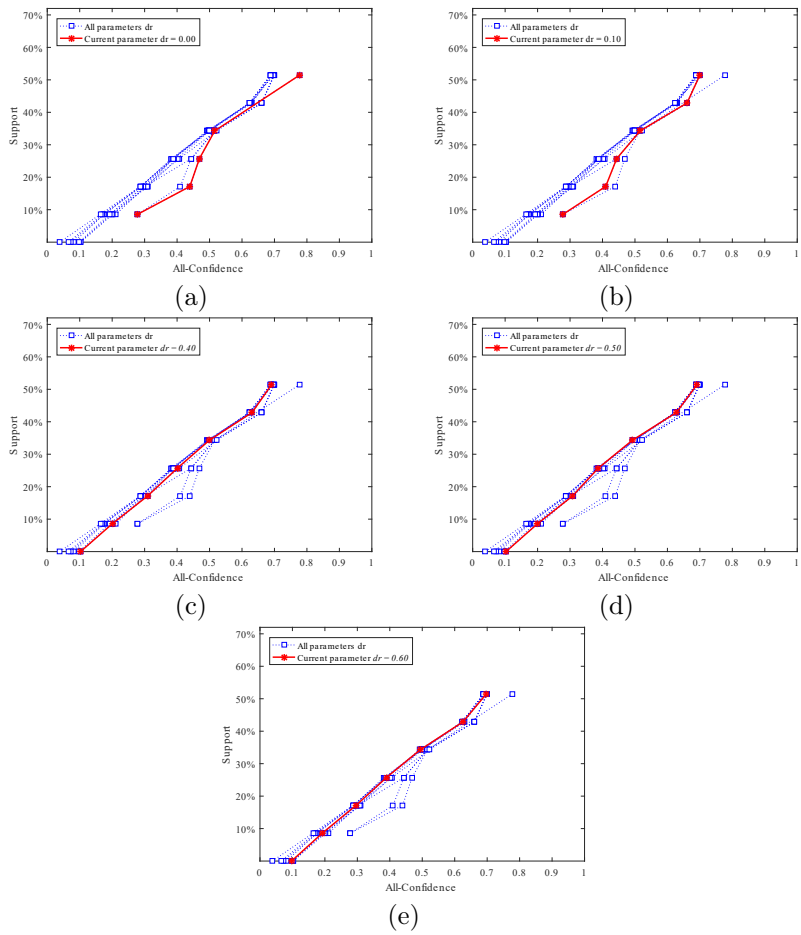


Figura A.30: *Led7*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM (a) com $dr = 0{,}00$, (b) com $dr = 0{,}10$,(c) com $dr = 0{,}40$, (d) com $dr = 0{,}50$ e (e) com $dr = 0{,}60$ onde estas imagens representam, por similaridade, o comportamento do $dr \in \{0{,}20\,, 0{,}30\,, 0{,}70\,, 0{,}80\,, 0{,}90\,, 1{,}00\}$. Veja Tabela A.23 para detalhes.

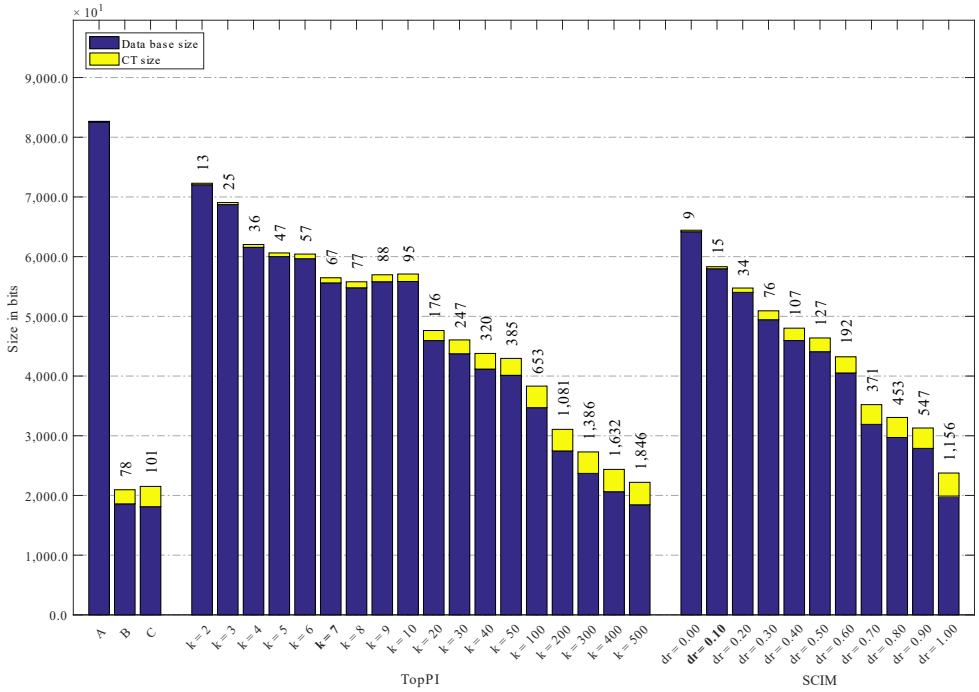


Figura A.31: *Led7*: valores métricos de MDL para as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão, e B e C representam os tamanhos dos itemsets fechados recuperados pelos algoritmos Slim e Krimp

A.1.11 Pima

Tabela A.24: *Pima*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

k	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,13]		(0,13 , 0,27]		(0,27 , 0,40]		(0,40 , 0,53]		(0,53 , 0,66]		(0,66 , 0,79]		(0,79 , 0,93]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	19	0,016	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	19	0,26
2	44	0,021	1	0,171	0	0,000	0	0,000	0	0,000	1	0,776	6	0,913	52	0,26
3	69	0,022	2	0,169	0	0,000	0	0,000	0	0,000	2	0,768	11	0,910	84	0,30
4	94	0,022	3	0,165	0	0,000	0	0,000	0	0,000	3	0,767	16	0,905	116	0,28
5	119	0,022	4	0,165	0	0,000	0	0,000	0	0,000	4	0,763	20	0,896	147	0,29
10	236	0,022	9	0,160	0	0,000	0	0,000	0	0,000	16	0,786	31	0,884	292	0,28
20	423	0,024	16	0,153	0	0,000	0	0,000	0	0,000	48	0,780	31	0,884	518	0,30
30	**593**	**0,026**	**16**	**0,153**	**0**	**0,000**	**0**	**0,000**	**8**	**0,683**	**69**	**0,773**	**31**	**0,884**	**717**	**0,31**
40	754	0,026	16	0,153	0	0,000	0	0,000	18	0,675	83	0,766	31	0,884	902	0,31
50	896	0,026	16	0,153	0	0,000	0	0,000	28	0,671	94	0,759	31	0,884	1.065	0,32
100	1.274	0,026	16	0,153	0	0,000	0	0,000	92	0,639	96	0,757	31	0,884	1.509	0,35
200	1.345	0,027	16	0,153	0	0,000	8	0,526	112	0,626	96	0,757	31	0,884	1.608	0,37

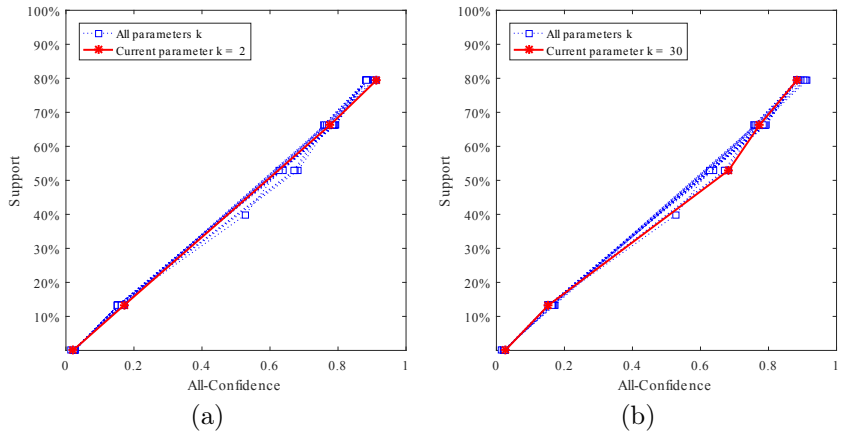


Figura A.32: *Pima*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI (a) com $k = 2$, onde esta imagem representa, por similaridade, o comportamento do $k \in \{1, 3, 4, 5, 6, 7, 8, 9, 10, 20\}$, e (b) com $k = 30$, onde esta imagem representa, por similaridade, o comportamento do $k \in \{40, 50, 100, 200\}$. Veja tabela A.24 para detalhes.

Tabela A.25: *Pima*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo SCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,13]		(0,13 , 0,27]		(0,27 , 0,40]		(0,40 , 0,53]		(0,53 , 0,66]		(0,66 , 0,79]		(0,79 , 0,93]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	24	0,019	1	0,171	0	0,000	0	0,000	0	0,000	1	0,776	0	0,000	26	0,03
0,01	62	0,034	3	0,157	0	0,000	0	0,000	0	0,000	11	0,778	19	0,885	95	0,04
0,02	205	0,032	3	0,157	0	0,000	0	0,000	58	0,635	79	0,759	31	0,884	376	0,05
0,03	**302**	**0,029**	**4**	**0,154**	**0**	**0,000**	**8**	**0,526**	**90**	**0,622**	**87**	**0,759**	**31**	**0,884**	**522**	**0,06**
0,04	411	0,028	5	0,153	0	0,000	8	0,526	111	0,626	94	0,757	31	0,884	660	0,06
0,05	455	0,030	5	0,153	0	0,000	8	0,526	111	0,626	94	0,757	31	0,884	704	0,06
0,10	754	0,036	13	0,153	0	0,000	8	0,526	112	0,626	96	0,757	31	0,884	1.014	0,07
0,20	1.237	0,027	16	0,153	0	0,000	8	0,526	112	0,626	96	0,757	31	0,884	1.500	0,10
0,30	1.334	0,027	16	0,153	0	0,000	8	0,526	112	0,626	96	0,757	31	0,884	1.597	0,08
0,40	1.337	0,027	16	0,153	0	0,000	8	0,526	112	0,626	96	0,757	31	0,884	1.600	0,10
0,50	1.345	0,027	16	0,153	0	0,000	8	0,526	112	0,626	96	0,757	31	0,884	1.608	0,08

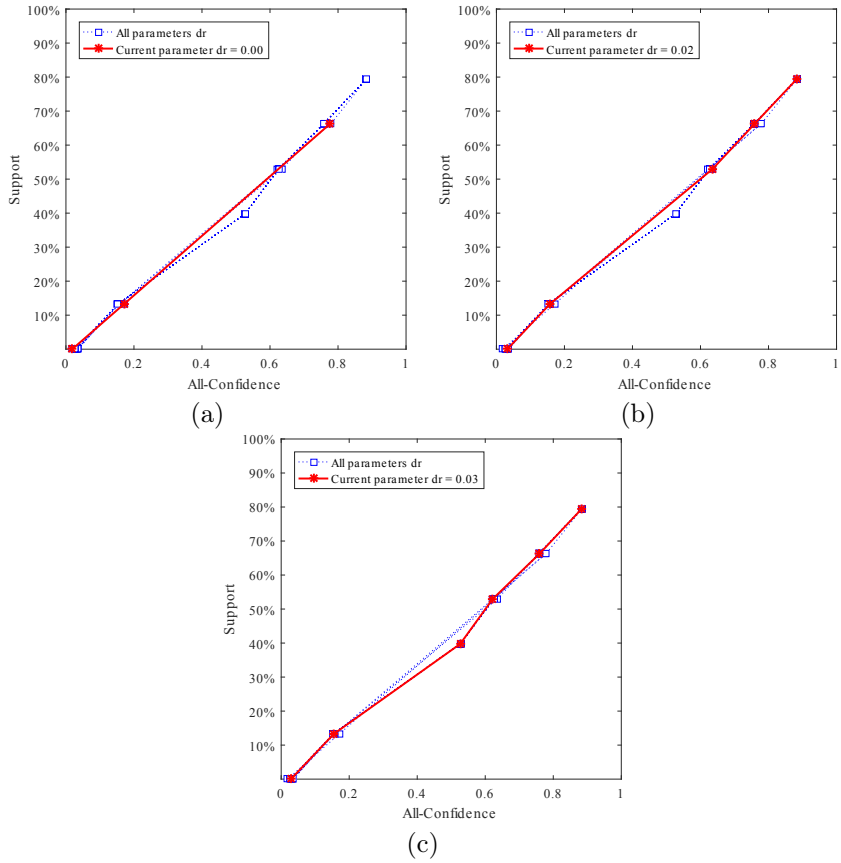


Figura A.33: *Pima*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo SCIM (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento do $dr = 0{,}01$, (b) com $dr = 0{,}02$, e (c) com $dr = 0{,}03$, onde estas imagens representam, por similaridade, o comportamento do $dr \in \{0{,}04\,, 0{,}05\,, 0{,}10\,, 0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja tabela A.25 para detalhes.

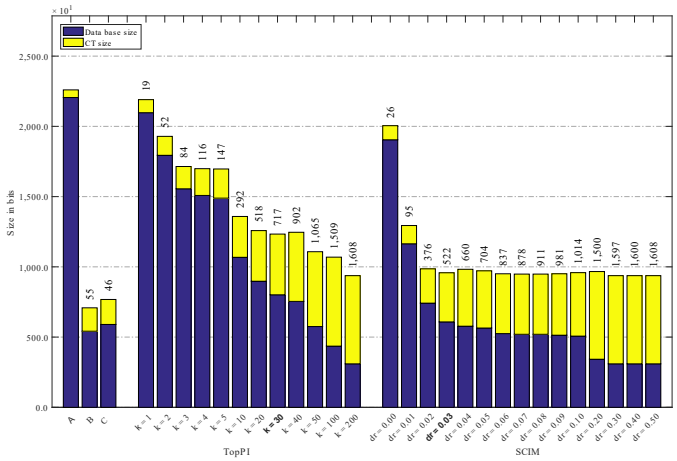


Figura A.34: *Pima*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido), enquanto as barras B e C representam os tamanhos relacionados aos itemsets fechados recuperados pelos algoritmos Slim e Krimp, respectivamente.

APÊNDICE B – PSCIM: EXPERIMENTOS COM QUATRO VARIAÇÕES DO ALGORITMO

Este documento contém detalhes dos itemsets fechados minerados pelas quatro variações do algoritmo PSCIM. Apresentamos as escolhas dos parâmetros para uma comparação não tendenciosa das variações propostas. São apresentadas quatro variações neste estudo, para mais detalhes veja a Subseção 6.1, de forma objetiva temos:

- PSCIM_{v1}: procedimento paralelo do algoritmo SCIM [40];
- PSCIM_{v2}: variação PSCIM_{v1} com clusterização utilizando percentil para modular o parâmetro dr, usados para estender a cobertura dos *clusters*;
- PSCIM_{v3}: variação PSCIM_{v1} com remoção das transações de tamanho 1, ou seja, apenas contém um item na transação;
- PSCIM_{v4}: Combinação das variações PSCIM_{v2} e PSCIM_{v3};

Usamos 16 bases de dados do SMPF [21]. Em algumas bases de dados, nós removemos as transações que tinham zero item e removemos a repetição de determinado item na mesma transação, consideramos esses dois casos como um erro para uma base de dados transacional. Neste estudo as bases de dados são separadas em dois grupos: bases de dados esparsas e densas. Essa decisão facilita a compreensão do comportamento dos algoritmos para esses dois cenários de tipo da bases de dados.

As Tabelas B.1 e B.2 resumem as organizações gerais deste apêndice em relação à análise do comportamento das variações do algoritmo PSCIM sob diferentes parametrizações. Mais especificamente, as tabelas e figuras referenciadas pelas colunas de dois a oito das Tabelas B.1 e B.2 mostram a distribuição dos valores médios de *All-confidence* (μ) calculados para os conjuntos de itens fechados recuperados por PSCIM_{v1}, PSCIM_{v2}, PSCIM_{v3}

e PSCIM_{v4}, usando diferentes valores de parâmetros *dr* ou *dp*. Os números referenciados pela coluna nove das Tabelas B.1 e B.2 mostram o MDL (Equação 2.6) calculado para os itemsets fechados recuperados por cada abordagem, sob diferentes parametrizações.

Para calcular μ, primeiro criamos sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte, dado uma determinada base de dados. Em seguida, calculamos μ a partir dos conjuntos de itens fechados recuperados em cada partição. Como pode ser visto nas figuras referenciadas pelas colunas três, cinco, sete e nove das Tabelas B.1 e B.2, comparamos as distribuições de medidas da média *All-confidence* para escolher os melhores valores de parâmetro para as variações do algoritmo PSCIM em cada base de dados. Nessas figuras, os eixos horizontais correspondem a μ variando de 0 a 1, enquanto os eixos verticais distribuem os valores de suporte de 0 para o limite superior de cada base de dados. Espera-se que quanto melhor for o conjunto de itens fechados recuperados, mais à direita será a curva que representa o desempenho de uma técnica/parametrização. Os valores dos parâmetros escolhidos são destacados em negrito nas tabelas referenciadas pelas colunas dois, quatro, seis e oito das Tabelas B.1 e B.2. Embora não seja usada como critério de escolha de parâmetros das variações do algoritmo PSCIM, é apresentado neste estudo a média de *cross-support* (Equação 7) a partir dos conjuntos de itens fechados recuperados em cada partição de suporte.

Tabela B.1: Bases de dados densa: Conjunto de tabelas e figuras apresentando o comportamento das variações do PSCIM sob diferentes parâmetros.

Bases de dados densa	**Média *All-confidence***								**MDL**
	PSCIM_{v1}		**PSCIM_{v2}**		**PSCIM_{v3}**		**PSCIM_{v4}**		
	Tabela	Figura	Tabela	Figura	Tabela	Figura	Tabela	Figura	Figura
Chess	B.5	B.3	B.6	B.4	B.7	B.5	B.8	B.6	B.7
Kddcup99	B.9	B.8	B.10	B.9	B.11	B.10	B.12	B.11	B.12
Mushrooms	B.13	B.13	B.14	B.14	B.15	B.15	B.16	B.16	B.17
PowerC	B.17	B.18	B.18	B.19	B.19	B.20	B.20	B.21	B.22
Pumsb	B.21	B.23	B.22	B.24	B.23	B.25	B.24	B.26	B.27
RecordLink	B.25	B.28	B.26	B.29	B.27	B.30	B.28	B.31	B.32
Skin	B.29	B.33	B.30	B.34	B.31	B.35	B.32	B.36	B.37
Susy	B.33	B.38	B.34	B.39	B.35	B.40	B.36	B.41	B.42

Os números referenciados pela coluna seis das Tabelas B.1 e B.2 mostram o *Minimum Description Length* (MDL) calculado para os conjuntos de itens fechados recuperados por cada abordagem, sob diferentes parametrizações. Os valores de MDL são usados para calcular $L\%$. Veja Equação 2.6 para detalhes. Essa métrica não é usada para critério de escolha de parâmetros das variações do algoritmo PSCIM. Ela não é apresentada na

Tabela B.2: Base de dados esparsa: Conjunto de tabelas e figuras apresentando o comportamento das variações do PSCIM sob diferentes parâmetros.

Bases de dados esparsa	**Média *All-confidence***								**MDL**
	PSCIM$_{v1}$		**PSCIM$_{v2}$**		**PSCIM$_{v3}$**		**PSCIM$_{v4}$**		
	Tabela	Figura	Tabela	Figura	Tabela	Figura	Tabela	Figura	Figura
Accidents	B.37	B.43	B.38	B.44	B.39	B.45	B.40	B.46	B.47
BMSWebView2	B.41	B.48	B.42	B.49	B.43	B.50	B.44	B.51	B.52
BMS1	B.45	B.53	B.46	B.54	B.47	B.55	B.48	B.56	B.57
FoodmartFIM	B.49	B.58	B.50	B.59	B.51	B.60	B.52	B.61	B.62
Fruithut	B.53	B.63	B.54	B.64	B.55	B.65	B.56	B.66	B.67
OnlineRetail	B.57	B.68	B.58	B.69	B.59	B.70	B.60	B.71	B.72
PAMP	B.61	B.73	B.62	B.74	B.63	B.75	B.64	B.76	B.77
Retail	B.65	B.78	B.66	B.79	B.67	B.80	B.68	B.81	B.82

discussão da Tese.

As Tabelas B.3 e B.4 resumem para uma determinada base de dados, o parâmetro vencedor obtido por cada variação do algoritmo PSCIM. A primeira coluna contém a informação da base de estudo junto com as informações de parâmetros escolhidos para cada algoritmo do estudo. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. As colunas três a nove mostram o número de conjuntos de itens fechados recuperados (#), os valores médios da métrica corrente (μ) e os valores de desvio padrão (σ) calculados por intervalo de suporte por todas as variações comparadas. A última coluna apresenta o número total de padrões detectados. A variação do PSCIM vencedora é destacado em negrito, usando como métrica de comparação a média de *all-confidence*, para cada base de dados. Em caso de empate entre as variações, escolhemos aquela que possui maior combinações de variações em sua definição.

Na Tabela B.4, na primeira coluna, todos os nomes de base de dados, exceto a base de dados *Accidents*, apresentam o símbolo (τ) na frente do nome, isso significa que a base corrente possui transações com tamanho 1, ou seja, a transação contém apenas um item. O símbolo (†) pode aparecer nas colunas 1 a 7, nas bases de dados onde a faixa de suporte é menor que 0,10 de suporte, nestes casos usamos notação científica para representar a faixa de valores de suporte. A Figura B.2 ilustram as distribuições de μ das bases de dados esparsas, se as faixas de suporte forem menor que 0,10, a proporção de largura e altura não é mantido, visto que nesses casos o eixo y é muito menor que o eixo x, dificultando a leitura do gráfico.

Tabela B.3: Desempenho do algoritmo/parametrização para as variações $PSCIM_{v1}$ (v1), $PSCIM_{v2}$ (v2), $PSCIM_{v3}$ (v3), e $PSCIM_{v4}$ (v4) sobre as bases de dados densas da Tabela B.1.

Chess	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	
V2 $dr = 0{,}0$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	
V3 $dr = 0{,}0$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	
V4 $dr = 0{,}0$	**all-confidence**	**42**	**0,069**	**0,044**	**29**	**0,226**	**0,075**	**42**	**0,338**	**0,043**	**66**	**0,513**	**0,047**	**26**	**0,650**	**0,058**	**8**	**0,791**	**0,034**	**12**	**0,940**	**0,026**	**225**
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027	

Kddcup99	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,11]			(0,11 , 0,23]			(0,23 , 0,34]			(0,34 , 0,45]			(0,45 , 0,57]			(0,57 , 0,68]			(0,68 , 0,79]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	
V2 $dr = 0{,}0$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	
V3 $dr = 0{,}0$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	
V4 $dr = 0{,}0$	**all-confidence**	**1.472**	**0,007**	**0,049**	**113**	**0,633**	**0,361**	**10**	**0,384**	**0,183**	**17**	**0,550**	**0,018**	**37**	**0,637**	**0,025**	**1**	**0,851**	**0,000**	**113**	**0,937**	**0,047**	**1.763**
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046	

Mushrooms	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,83]			(0,83 , 0,97]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}00015$	all-confidence	289	0,071	0,062	108	0,251	0,088	57	0,427	0,099	19	0,562	0,124	3	0,591	0,013	1	0,832	0,000	1	0,997	0,000	478
	cross-support		0,139	0,124		0,328	0,130		0,516	0,153		0,613	0,157		0,606	0,038		0,878	0,000		0,998	0,000	
V2 $dr = 0{,}02$	all-confidence	197	0,072	0,070	86	0,254	0,096	43	0,431	0,102	13	0,599	0,133	3	0,591	0,013	1	0,832	0,000	1	0,997	0,000	344
	cross-support		0,117	0,130		0,316	0,139		0,506	0,160		0,649	0,166		0,606	0,038		0,878	0,000		0,998	0,000	

Continua na próxima página.

V3 $dr = 0{,}00015$	all-confidence cross-support	289	0,071 0,139	0,062 0,124	108	0,251 0,328	0,088 0,130	57	0,427 0,516	0,099 0,153	19	0,562 0,613	0,124 0,157	3	0,591 0,606	0,013 0,038	1	0,832 0,878	0,000 0,000	1	0,997 0,998	0,000 0,000	478
V4 $dr = 0{,}02$	**all-confidence** cross-support	**197**	**0,072** 0,117	**0,070** 0,130	**86**	**0,254** 0,316	**0,096** 0,139	**43**	**0,431** 0,506	**0,102** 0,160	**13**	**0,599** 0,649	**0,133** 0,166	**3**	**0,591** 0,606	**0,013** 0,038	**1**	**0,832** 0,878	**0,000** 0,000	**1**	**0,997** 0,998	**0,000** 0,000	**344**

PowerC	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,13]			(0,13 , 0,27]			(0,27 , 0,40]			(0,40 , 0,53]			(0,53 , 0,66]			(0,66 , 0,80]			(0,80 , 0,93]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0001$	all-confidence cross-support	779	0,023 0,046	0,115 0,151	1	0,934 0,960	0,000 0,000	3	0,358 0,373	0,040 0,050	4	0,792 0,796	0,120 0,121	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	787
V2 $dr = 0{,}02$	all-confidence cross-support	758	0,024 0,047	0,117 0,153	1	0,934 0,960	0,000 0,000	2	0,357 0,375	0,056 0,070	5	0,737 0,739	0,162 0,163	1	0,598 0,609	0,000 0,000	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	767
V3 $dr = 0{,}0001$	all-confidence cross-support	779	0,023 0,046	0,115 0,151	1	0,934 0,960	0,000 0,000	3	0,358 0,373	0,040 0,050	4	0,792 0,796	0,120 0,121	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	0	0,000 0,000	0,000 0,000	787
V4 $dr = 0{,}02$	**all-confidence** cross-support	**758**	**0,024** 0,047	**0,117** 0,153	**1**	**0,934** 0,960	**0,000** 0,000	**2**	**0,357** 0,375	**0,056** 0,070	**5**	**0,737** 0,739	**0,162** 0,163	**1**	**0,598** 0,609	**0,000** 0,000	**0**	**0,000** 0,000	**0,000** 0,000	**0**	**0,000** 0,000	**0,000** 0,000	**767**

Pumsb	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 0,99]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence cross-support	1.135	0,122 0,156	0,240 0,279	69	0,371 0,387	0,175 0,181	33	0,606 0,620	0,180 0,178	31	0,780 0,788	0,166 0,160	4	0,843 0,845	0,154 0,156	5	0,971 0,975	0,030 0,033	4	1,000 1,000	0,000 0,000	1.281
V2 $dr = 0{,}0$	all-confidence cross-support	1.135	0,122 0,156	0,240 0,279	69	0,371 0,387	0,175 0,181	33	0,606 0,620	0,180 0,178	31	0,780 0,788	0,166 0,160	4	0,843 0,845	0,154 0,156	5	0,971 0,975	0,030 0,033	4	1,000 1,000	0,000 0,000	1.281
V3 $dr = 0{,}0$	all-confidence cross-support	1.135	0,122 0,156	0,240 0,279	69	0,371 0,387	0,175 0,181	33	0,606 0,620	0,180 0,178	31	0,780 0,788	0,166 0,160	4	0,843 0,845	0,154 0,156	5	0,971 0,975	0,030 0,033	4	1,000 1,000	0,000 0,000	1.281
V4 $dr = 0{,}0$	**all-confidence** cross-support	**1.135**	**0,122** 0,156	**0,240** 0,279	**69**	**0,371** 0,387	**0,175** 0,181	**33**	**0,606** 0,620	**0,180** 0,178	**31**	**0,780** 0,788	**0,166** 0,160	**4**	**0,843** 0,845	**0,154** 0,156	**5**	**0,971** 0,975	**0,030** 0,033	**4**	**1,000** 1,000	**0,000** 0,000	**1.281**

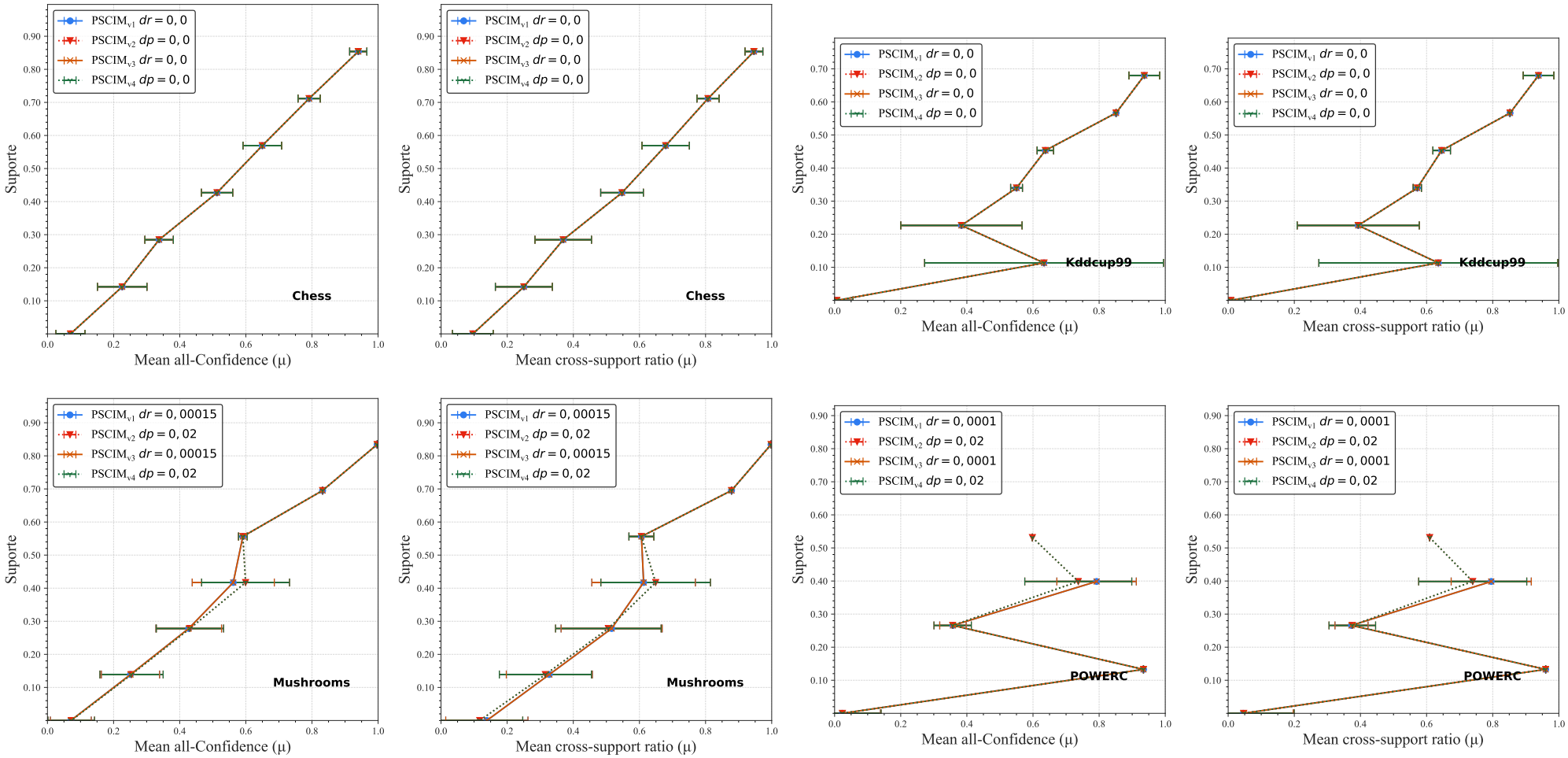
RecordLink	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence cross-support	220	0,005 0,006	0,043 0,045	1	0,628 0,709	0,000 0,000	15	0,420 0,509	0,050 0,049	13	0,512 0,603	0,056 0,114	14	0,660 0,748	0,069 0,096	10	0,774 0,775	0,006 0,006	4	0,984 0,988	0,009 0,006	277
V2 $dr = 0{,}0$	all-confidence cross-support	220	0,005 0,006	0,043 0,045	1	0,628 0,709	0,000 0,000	15	0,420 0,509	0,050 0,049	13	0,512 0,603	0,056 0,114	14	0,660 0,748	0,069 0,096	10	0,774 0,775	0,006 0,006	4	0,984 0,988	0,009 0,006	277

Continua na próxima página.

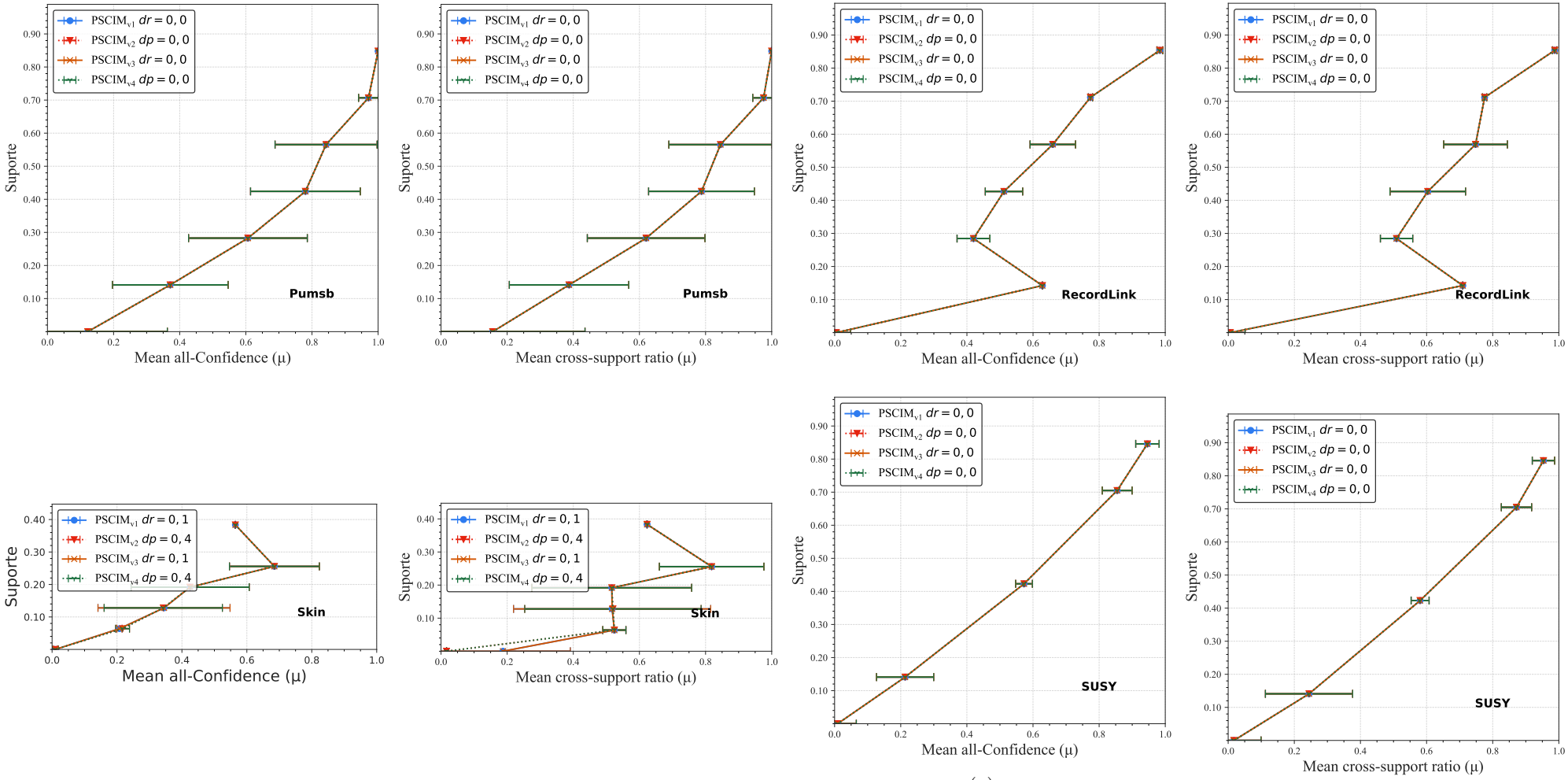
V3 $dr = 0{,}0$	all-confidence cross-support	220	0,005 0,006	0,043 0,045	1	0,628 0,709	0,000 0,000	15	0,420 0,509	0,050 0,049	13	0,512 0,603	0,056 0,114	14	0,660 0,748	0,069 0,096	10	0,774 0,775	0,006 0,006	4	0,984 0,988	0,009 0,006	277
V4 $dr = 0{,}0$	**all-confidence** cross-support	**220**	**0,005** 0,006	**0,043** 0,045	**1**	**0,628** 0,709	**0,000** 0,000	**15**	**0,420** 0,509	**0,050** 0,049	**13**	**0,512** 0,603	**0,056** 0,114	**14**	**0,660** 0,748	**0,069** 0,096	**10**	**0,774** 0,775	**0,006** 0,006	**4**	**0,984** 0,988	**0,009** 0,006	**277**

Skin	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,06]			(0,06 , 0,13]			(0,13 , 0,19]			(0,19 , 0,26]			(0,26 , 0,32]			(0,32 , 0,38]			(0,38 , 0,45]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}1$	all-confidence cross-support	4	0,010 0,187	0,006 0,204	3	0,208 0,524	0,008 0,035	5	0,345 0,517	0,203 0,298	6	0,425 0,516	0,182 0,242	3	0,685 0,818	0,138 0,158	0	0,000 0,000	0,000 0,000	1	0,565 0,623	0,000 0,000	22
V2 $dr = 0{,}4$	all-confidence cross-support	2	0,010 0,017	0,010 0,000	3	0,217 0,524	0,021 0,035	6	0,342 0,520	0,182 0,267	6	0,425 0,516	0,182 0,242	3	0,685 0,818	0,138 0,158	0	0,000 0,000	0,000 0,000	1	0,565 0,623	0,000 0,000	21
V3 $dr = 0{,}1$	all-confidence cross-support	4	0,010 0,187	0,006 0,204	3	0,208 0,524	0,008 0,035	5	0,345 0,517	0,203 0,298	6	0,425 0,516	0,182 0,242	3	0,685 0,818	0,138 0,158	0	0,000 0,000	0,000 0,000	1	0,565 0,623	0,000 0,000	22
V4 $dr = 0{,}4$	**all-confidence** cross-support	**2**	**0,010** 0,017	**0,010** 0,000	**3**	**0,217** 0,524	**0,021** 0,035	**6**	**0,342** 0,520	**0,182** 0,267	**6**	**0,425** 0,516	**0,182** 0,242	**3**	**0,685** 0,818	**0,138** 0,158	**0**	**0,000** 0,000	**0,000** 0,000	**1**	**0,565** 0,623	**0,000** 0,000	**21**

Susy	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,85]			(0,85 , 0,99]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence cross-support	2.702	0,009 0,017	0,057 0,082	24	0,213 0,245	0,086 0,132	0	0,000 0,000	0,000 0,000	8	0,573 0,580	0,025 0,027	0	0,000 0,000	0,000 0,000	9	0,855 0,872	0,045 0,046	18	0,946 0,953	0,035 0,034	2.761
V2 $dr = 0{,}0$	all-confidence cross-support	2.702	0,009 0,017	0,057 0,082	24	0,213 0,245	0,086 0,132	0	0,000 0,000	0,000 0,000	8	0,573 0,580	0,025 0,027	0	0,000 0,000	0,000 0,000	9	0,855 0,872	0,045 0,046	18	0,946 0,953	0,035 0,034	2.761
V3 $dr = 0{,}0$	all-confidence cross-support	2.702	0,009 0,017	0,057 0,082	24	0,213 0,245	0,086 0,132	0	0,000 0,000	0,000 0,000	8	0,573 0,580	0,025 0,027	0	0,000 0,000	0,000 0,000	9	0,855 0,872	0,045 0,046	18	0,946 0,953	0,035 0,034	2.761
V4 $dr = 0{,}0$	**all-confidence** cross-support	**2.702**	**0,009** 0,017	**0,057** 0,082	**24**	**0,213** 0,245	**0,086** 0,132	**0**	**0,000** 0,000	**0,000** 0,000	**8**	**0,573** 0,580	**0,025** 0,027	**0**	**0,000** 0,000	**0,000** 0,000	**9**	**0,855** 0,872	**0,045** 0,046	**18**	**0,946** 0,953	**0,035** 0,034	**2.761**



Continua na próxima página.



(g)

Figura B.1: Distribuições dos valores médios de *All-confidence* e *cross-support* dos itemsets fechados recuperados pelo PSCIM$_{v1}$, PSCIM$_{v2}$, PSCIM$_{v3}$, e PSCIM$_{v4}$ sobre as bases de dados densa da Tabela B.3, neste estudo é usando o melhor valor de parâmetro para cada variação do PSCIM.

Tabela B.4: Desempenho do algoritmo/parametrização para as variações $PSCIM_{v1}$ (v1), $PSCIM_{v2}$ (v2), $PSCIM_{v3}$ (v3), e $PSCIM_{v4}$ (v4) sobre as bases de dados esparsas da Tabela B.2

Accidents	***Métrica***	**Partição de suporte**																					Itemset
		[0,00 , 0,14]			(0,14 , 0,29]			(0,29 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,86]			(0,86 , 1,00]			#
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	
V2 $dr = 0{,}0$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	
V3 $dr = 0{,}0$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	
V4 $dr = 0{,}0$	**all-confidence**	**2.222**	**0,026**	**0,073**	**104**	**0,221**	**0,115**	**2**	**0,411**	**0,111**	**151**	**0,520**	**0,041**	**77**	**0,624**	**0,050**	**12**	**0,785**	**0,042**	**8**	**0,956**	**0,056**	**2.576**
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056	

BMSWebView2 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,08] †			(0,08 , 0,11] †			(0,11 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,19] †			#
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0005$	all-confidence	2.524	0,055	0,113	3	0,390	0,074	4	0,270	0,148	2	0,189	0,022	2	0,257	0,027	1	0,309	0,000	0	0,000	0,000	2.536
	cross-support		0,285	0,243		0,606	0,106		0,571	0,118		0,542	0,078		0,542	0,078		0,663	0,000		0,000	0,000	
V2 $dr = 0{,}0015$	all-confidence	482	0,222	0,168	21	0,319	0,088	9	0,380	0,100	4	0,366	0,045	2	0,398	0,022	1	0,309	0,000	2	0,492	0,075	521
	cross-support		0,509	0,279		0,616	0,132		0,688	0,140		0,864	0,085		0,818	0,118		0,663	0,000		0,816	0,120	
V3 $dr = 0{,}0$	all-confidence	4.119	0,166	0,124	32	0,370	0,097	6	0,427	0,071	2	0,344	0,034	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	4.159
	cross-support		0,507	0,232		0,714	0,147		0,747	0,101		0,870	0,124		0,000	0,000		0,000	0,000		0,000	0,000	
V4 $dr = 0{,}0015$	**all-confidence**	**4.163**	**0,165**	**0,124**	**47**	**0,343**	**0,102**	**12**	**0,380**	**0,097**	**5**	**0,356**	**0,046**	**1**	**0,414**	**0,000**	**1**	**0,309**	**0,000**	**2**	**0,492**	**0,075**	**4.231**
	cross-support		0,505	0,233		0,688	0,153		0,731	0,133		0,823	0,119		0,734	0,000		0,663	0,000		0,816	0,120	

BMS1 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,09] †			(0,09 , 0,12] †			(0,12 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,20] †			#
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}002$	all-confidence	160	0,077	0,175	1	0,224	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	161
	cross-support		0,337	0,280		0,520	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000	
V2 $dr = 0{,}02$	**all-confidence**	**1.670**	**0,019**	**0,061**	**25**	**0,123**	**0,090**	**1**	**0,223**	**0,000**	**2**	**0,279**	**0,051**	**1**	**0,311**	**0,000**	**1**	**0,387**	**0,000**	**1**	**0,329**	**0,000**	**1.701**
	cross-support		0,193	0,192		0,608	0,222		0,713	0,000		0,776	0,065		0,847	0,000		0,903	0,000		0,987	0,000	

Continua na próxima página.

V3 $dr = 0{,}027$	all-confidence	1.954	0,040	0,080	23	0,278	0,088	3	0,235	0,070	2	0,234	0,116	0	0,000	0,000	2	0,357	0,042	1	0,329	0,000	1.985
	cross-support		0,293	0,229		0,692	0,197		0,522	0,256		0,462	0,379		0,000	0,000		0,811	0,131		0,987	0,000	
V4 $dr = 0{,}008$	all-confidence	363	0,140	0,143	24	0,264	0,090	3	0,235	0,070	3	0,210	0,091	2	0,260	0,072	2	0,357	0,042	1	0,329	0,000	398
	cross-support		0,524	0,261		0,659	0,227		0,522	0,256		0,402	0,288		0,715	0,187		0,811	0,131		0,987	0,000	

FoodmartFIM τ	***Métrica***	**Partição de suporte † $\times 10^{-3}$**																					Itemset #
		[0,00 , 0,14] †			(0,14 , 0,28] †			(0,28 , 0,41] †			(0,41 , 0,55] †			(0,55 , 0,69] †			(0,69 , 0,83] †			(0,83 , 0,97] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}2$	all-confidence	0	0,000	0,000	532	0,083	0,022	0	0,000	0,000	248	0,149	0,032	0	0,000	0,000	29	0,209	0,046	2	0,211	0,000	811
	cross-support		0,000	0,000		0,643	0,223		0,000	0,000		0,728	0,171		0,000	0,000		0,774	0,136		0,658	0,112	
V2 $dr = 0{,}02$	all-confidence	0	0,000	0,000	465	0,072	0,012	0	0,000	0,000	230	0,136	0,023	0	0,000	0,000	28	0,195	0,031	2	0,211	0,000	725
	cross-support		0,000	0,000		0,615	0,223		0,000	0,000		0,721	0,173		0,000	0,000		0,746	0,147		0,658	0,112	
V3 $dr = 0{,}2$	**all-confidence**	**0**	**0,000**	**0,000**	**63**	**0,104**	**0,026**	**0**	**0,000**	**0,000**	**51**	**0,175**	**0,037**	**0**	**0,000**	**0,000**	**17**	**0,225**	**0,046**	**1**	**0,211**	**0,000**	**132**
	cross-support		0,000	0,000		0,570	0,258		0,000	0,000		0,733	0,203		0,000	0,000		0,753	0,136		0,579	0,000	
V4 $dr = 0{,}03$	all-confidence	0	0,000	0,000	951	0,078	0,017	0	0,000	0,000	452	0,142	0,028	0	0,000	0,000	50	0,190	0,039	2	0,211	0,000	1.455
	cross-support		0,000	0,000		0,656	0,218		0,000	0,000		0,745	0,163		0,000	0,000		0,754	0,147		0,658	0,112	

Fruithut τ	***Métrica***	**Partição de suporte † $\times 10^{-1}$**																					Itemset #
		[0,00 , 0,05] †			(0,05 , 0,10] †			(0,10 , 0,15] †			(0,15 , 0,20] †			(0,20 , 0,25] †			(0,25 , 0,30] †			(0,30 , 0,35] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}00011$	all-confidence	6.881	0,002	0,010	3	0,067	0,007	0	0,000	0,000	1	0,078	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	6.885
	cross-support		0,154	0,167		0,553	0,214		0,000	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000	
V2 $dr = 0{,}05$	all-confidence	18.504	0,003	0,010	47	0,082	0,043	8	0,055	0,019	4	0,076	0,006	2	0,089	0,008	0	0,000	0,000	0	0,000	0,000	18.565
	cross-support		0,229	0,231		0,504	0,328		0,193	0,153		0,228	0,053		0,212	0,035		0,000	0,000		0,000	0,000	
V3 $dr = 0{,}0$	all-confidence	298	0,037	0,035	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	298
	cross-support		0,373	0,293		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000	
V4 $dr = 0{,}004$	**all-confidence**	**398**	**0,033**	**0,035**	**15**	**0,088**	**0,053**	**4**	**0,077**	**0,056**	**1**	**0,080**	**0,000**	**2**	**0,089**	**0,008**	**0**	**0,000**	**0,000**	**1**	**0,146**	**0,000**	**421**
	cross-support		0,330	0,297		0,512	0,408		0,217	0,177		0,173	0,000		0,212	0,035		0,000	0,000		0,470	0,000	

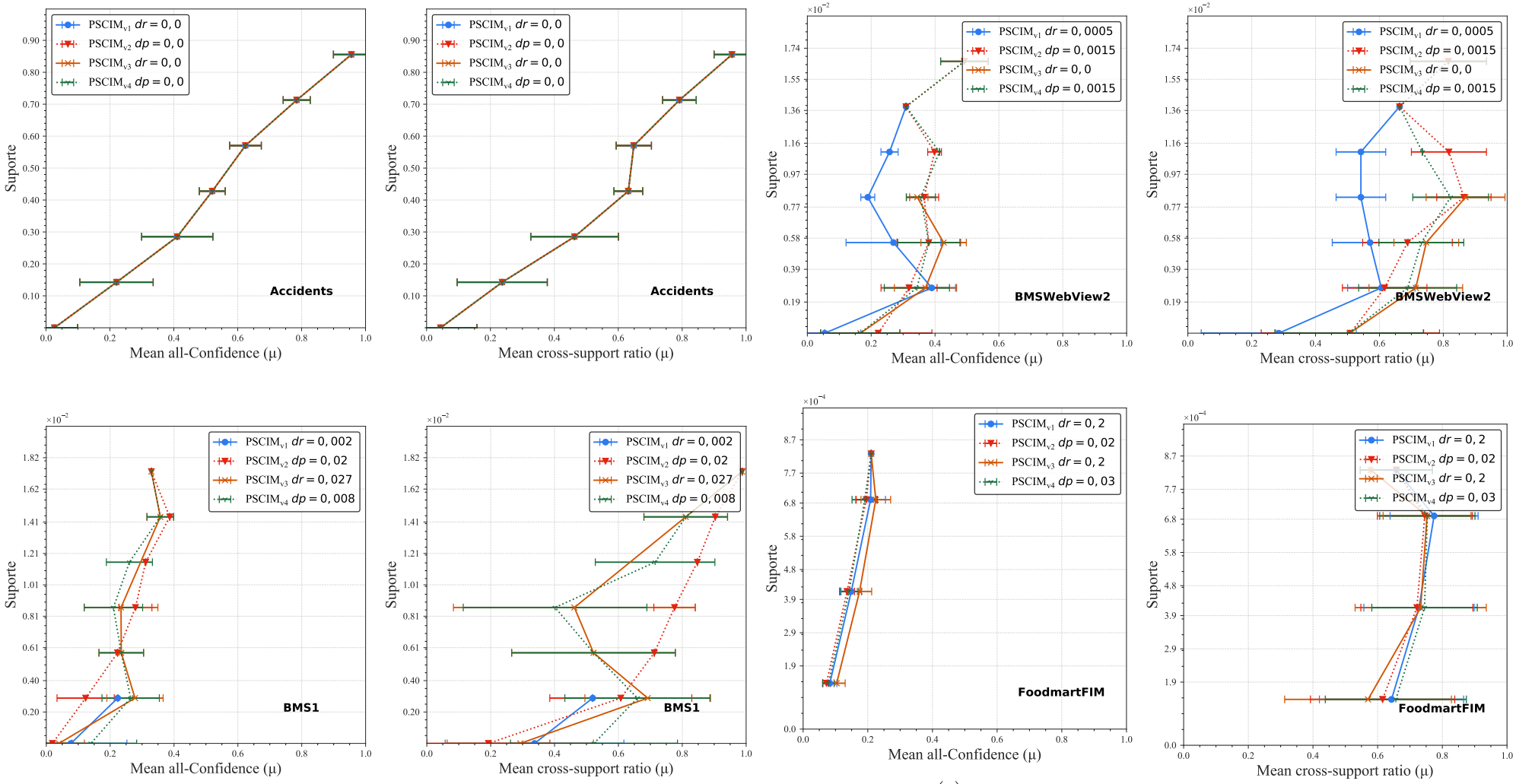
OnlineRetail τ	***Métrica***	**Partição de suporte † $\times 10^{-1}$**																					Itemset #
		[0,00 , 0,08] †			(0,08 , 0,15] †			(0,15 , 0,23] †			(0,23 , 0,31] †			(0,31 , 0,39] †			(0,39 , 0,46] †			(0,46 , 0,54] †			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	

Continua na próxima página.

V1 $dr = 0{,}0000001$	all-confidence	1.061	0,433	0,318	5	0,617	0,212	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.073
	cross-support		0,476	0,327		0,639	0,240		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	
V2 $dr = 0{,}005$	all-confidence	1.059	0,434	0,318	6	0,551	0,250	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.072
	cross-support		0,478	0,327		0,603	0,232		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	
V3 $dr = 0{,}0000001$	**all-confidence**	**1.064**	**0,429**	**0,319**	**5**	**0,617**	**0,212**	**3**	**0,650**	**0,181**	**1**	**0,726**	**0,000**	**2**	**0,419**	**0,076**	**0**	**0,000**	**0,000**	**1**	**0,536**	**0,000**	**1.076**
	cross-support		0,475	0,327		0,639	0,240		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	
V4 $dr = 0{,}005$	all-confidence	1.065	0,429	0,319	6	0,551	0,250	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.078
	cross-support		0,475	0,327		0,603	0,232		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000	

PAMP τ	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,13]			(0,13 , 0,26]			(0,26 , 0,38]			(0,38 , 0,51]			(0,51 , 0,64]			(0,64 , 0,77]			(0,77 , 0,90]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}0$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	
V2 $dr = 0{,}0$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	
V3 $dr = 0{,}0$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	
V4 $dr = 0{,}0$	**all-confidence**	**74**	**0,120**	**0,200**	**6**	**0,474**	**0,228**	**11**	**0,487**	**0,227**	**6**	**0,693**	**0,112**	**5**	**0,806**	**0,025**	**1**	**0,946**	**0,000**	**2**	**0,920**	**0,052**	**105**
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042	

Retail τ	***Métrica***	**Partição de suporte**																					Itemset #
		[0,00 , 0,05]			(0,05 , 0,09]			(0,09 , 0,14]			(0,14 , 0,19]			(0,19 , 0,24]			(0,24 , 0,28]			(0,28 , 0,33]			
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	
V1 $dr = 0{,}00000045$	all-confidence	9.830	0,059	0,119	2	0,168	0,032	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.834
	cross-support		0,145	0,255		0,327	0,046		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	
V2 $dr = 0{,}002$	all-confidence	9.729	0,072	0,119	1	0,191	0,000	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.732
	cross-support		0,193	0,275		0,360	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	
V3 $dr = 0{,}00000045$	all-confidence	9.830	0,059	0,119	2	0,168	0,032	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.834
	cross-support		0,145	0,255		0,327	0,046		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	
V4 $dr = 0{,}002$	**all-confidence**	**9.729**	**0,072**	**0,119**	**1**	**0,191**	**0,000**	**1**	**0,225**	**0,000**	**0**	**0,000**	**0,000**	**0**	**0,000**	**0,000**	**0**	**0,000**	**0,000**	**1**	**0,575**	**0,000**	**9.732**
	cross-support		0,193	0,275		0,360	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000	



(g)

Continua na próxima página.

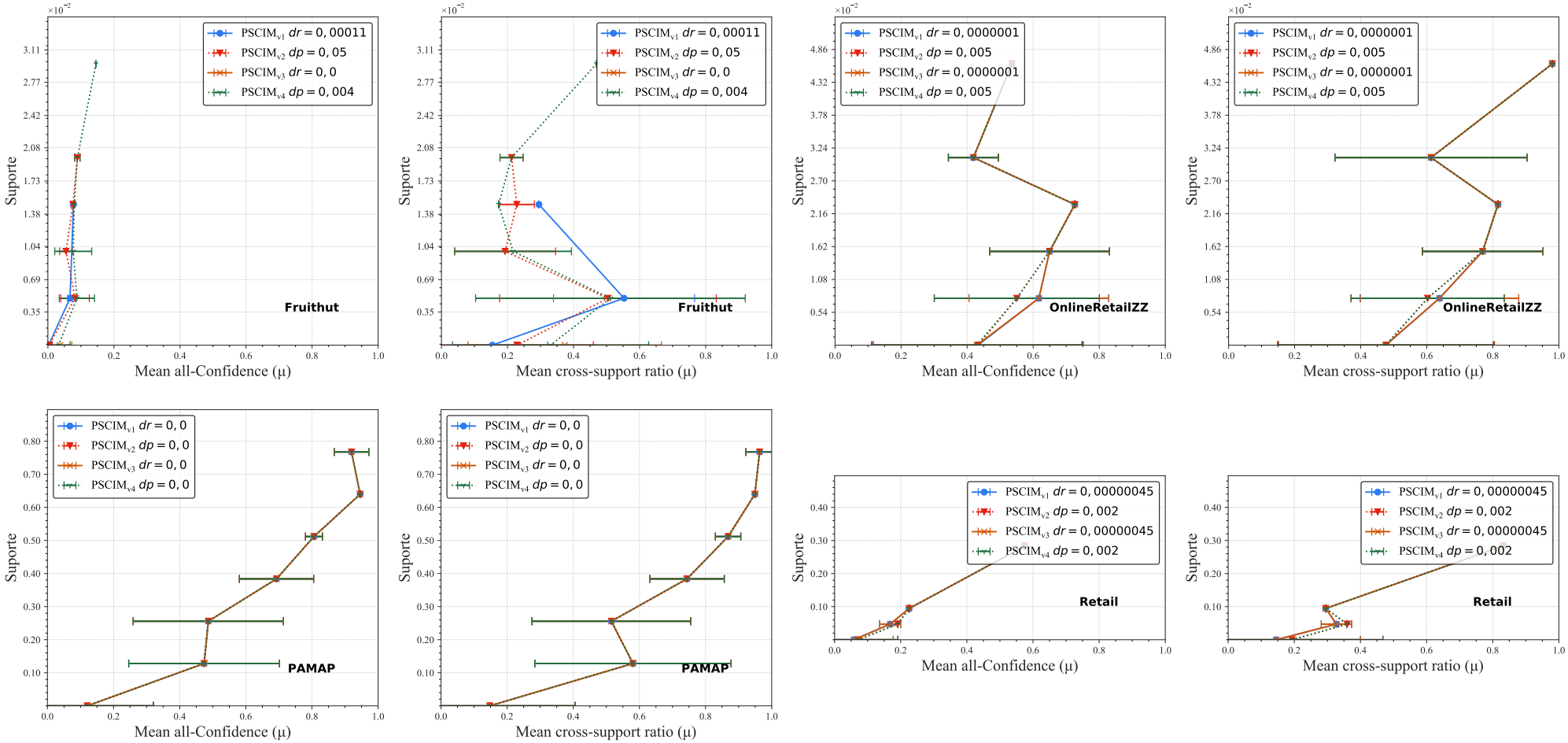


Figura B.2: Distribuições dos valores médios de *all-confidence* e *cross-support* dos itemsets fechados recuperados pelo PSCIM$_{v1}$, PSCIM$_{v2}$, PSCIM$_{v3}$, e PSCIM$_{v4}$ sobre as bases de dados esparsa da Tabela B.4, neste estudo é usando o melhor valor de parâmetro para cada variação do PSCIM.

B.1 Escolha dos Parâmetros

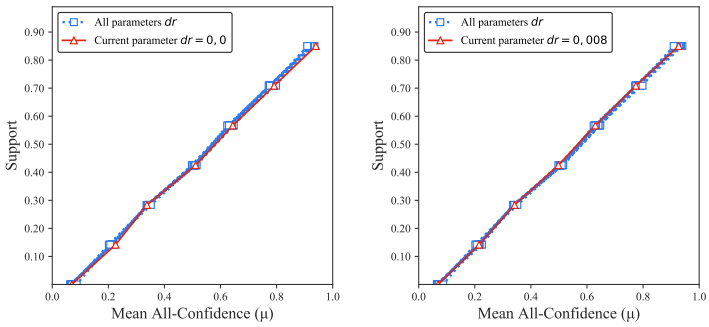
Nesta seção, nós mostramos as distribuições de média de *All-confidence* (μ) e comprimento mínimo de descrição (MDL) dos conjuntos de itens fechados recuperados pelas variações do algoritmo PSCIM usando valores de parâmetros diferentes. Separamos os resultados do estudo em duas subseções: bases de dados Densa (Subseção B.1.1) e esparsa (Subseção B.1.2).

B.1.1 Bases de Dados Densa

B.1.1.1 Chess

Tabela B.5: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
$\ddagger \times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**42**	**0,069**	**29**	**0,226**	**42**	**0,338**	**64**	**0,511**	**28**	**0,644**	**8**	**0,791**	**12**	**0,940**	**225**	**0,69**
0,01 ‡	42	0,065	38	0,216	52	0,341	65	0,511	32	0,645	10	0,787	26	0,933	265	0,71
0,02 ‡	49	0,066	45	0,211	56	0,339	67	0,512	35	0,647	18	0,772	32	0,927	302	0,77
0,03 ‡	49	0,066	49	0,209	56	0,339	74	0,515	47	0,640	23	0,773	34	0,925	332	0,74
0,04 ‡	55	0,066	53	0,215	75	0,338	98	0,500	51	0,641	26	0,774	41	0,935	399	0,75
0,05 ‡	55	0,066	57	0,216	81	0,337	113	0,504	64	0,631	26	0,774	45	0,931	441	0,79
0,06 ‡	60	0,065	57	0,216	81	0,337	114	0,503	66	0,630	26	0,774	45	0,931	449	0,77
0,07 ‡	63	0,068	60	0,216	85	0,339	115	0,503	68	0,630	26	0,774	53	0,927	470	0,78
0,08 ‡	65	0,069	82	0,215	117	0,342	131	0,498	68	0,630	26	0,774	59	0,927	548	0,82
0,09 ‡	65	0,069	93	0,214	125	0,340	133	0,498	72	0,629	26	0,774	60	0,927	574	0,83
0,10 ‡	67	0,069	99	0,213	138	0,340	160	0,503	74	0,630	31	0,773	67	0,921	636	0,89
0,15 ‡	80	0,069	167	0,204	162	0,341	168	0,505	104	0,633	38	0,781	95	0,911	814	0,90
0,20 ‡	101	0,071	290	0,202	277	0,341	243	0,508	141	0,624	43	0,776	103	0,913	1.198	0,98
0,30 ‡	188	0,087	503	0,207	473	0,352	455	0,510	271	0,630	107	0,798	139	0,909	2.136	1,10



(a) (b)

Figura B.3: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-1}$, onde essa imagem representa, por similaridade, o comportamento do $dr \in \{0{,}01\,, 0{,}02\,, 0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\}$, e (b) com $dr = 0{,}08 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}15\,, 0{,}20\,, 0{,}30\}$. Veja Tabela B.5 para detalhes.

Tabela B.6: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**42**	**0,069**	**29**	**0,226**	**42**	**0,338**	**64**	**0,511**	**28**	**0,644**	**8**	**0,791**	**12**	**0,940**	**225**	**0,71**
0,03	47	0,068	33	0,228	43	0,335	72	0,510	38	0,644	16	0,786	28	0,938	277	0,73
0,04	67	0,065	40	0,210	68	0,336	134	0,505	49	0,636	17	0,783	28	0,938	403	0,78
0,05	67	0,065	43	0,208	69	0,336	134	0,505	49	0,636	17	0,783	28	0,938	407	0,82
0,06	76	0,069	45	0,210	72	0,336	152	0,506	71	0,635	36	0,776	61	0,932	513	0,90
0,07	93	0,067	74	0,212	116	0,348	199	0,503	97	0,626	36	0,776	61	0,932	676	0,89
0,08	93	0,067	78	0,212	118	0,347	253	0,505	97	0,626	36	0,776	61	0,932	736	0,89
0,09	107	0,071	89	0,212	147	0,353	272	0,507	142	0,629	52	0,771	120	0,916	929	0,93
0,10	120	0,068	154	0,214	180	0,350	335	0,510	171	0,622	52	0,771	120	0,916	1.132	0,99
0,11	151	0,073	164	0,211	274	0,364	371	0,506	183	0,623	64	0,770	131	0,917	1.338	1,03
0,12	191	0,078	191	0,215	376	0,373	474	0,501	248	0,627	116	0,780	192	0,907	1.788	1,11
0,13	212	0,075	302	0,217	454	0,368	625	0,504	258	0,625	116	0,780	192	0,907	2.159	1,13

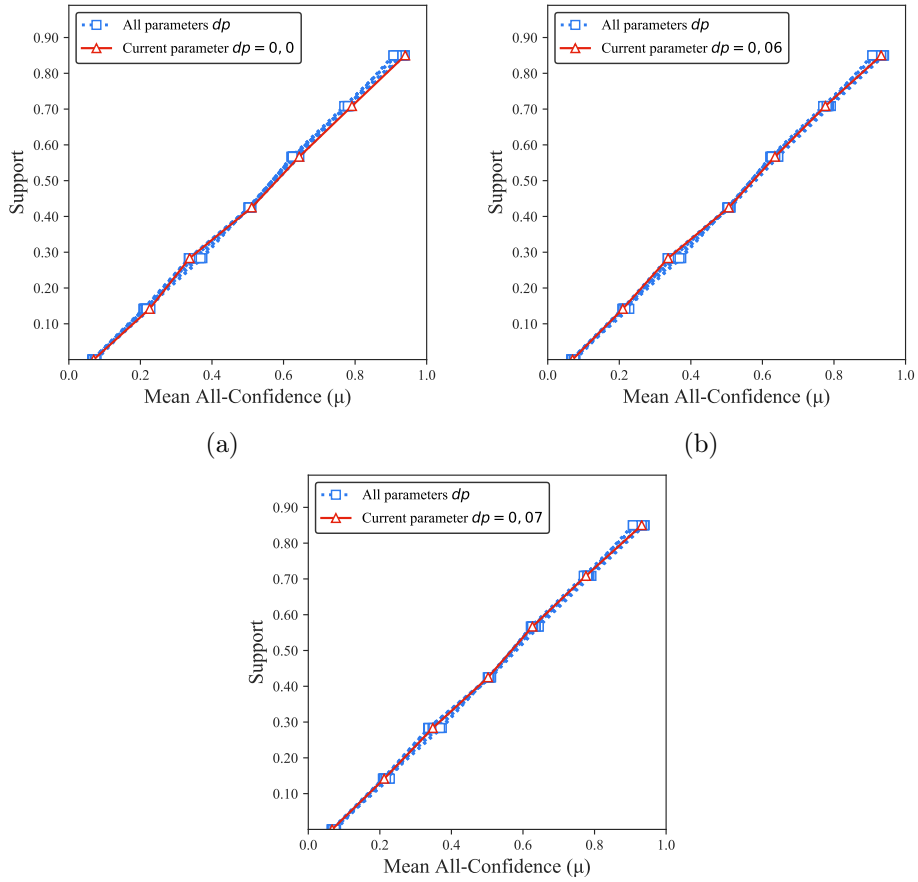


Figura B.4: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v2}. (a) com $dr = 0{,}00$, onde essa imagem representa, por similaridade, o comportamento do $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\}$, (b) com $dr = 0{,}06$, e (c) com $dr = 0{,}07$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\}$. Veja Tabela B.6 para detalhes.

Tabela B.7: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
$\ddagger \times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**42**	**0,069**	**29**	**0,226**	**42**	**0,338**	**64**	**0,511**	**28**	**0,644**	**8**	**0,791**	**12**	**0,940**	**225**	**0,66**
0,01 ‡	42	0,065	38	0,216	52	0,341	65	0,511	32	0,645	10	0,787	26	0,933	265	0,68
0,02 ‡	49	0,066	45	0,211	56	0,339	67	0,512	35	0,647	18	0,772	32	0,927	302	0,70
0,03 ‡	49	0,066	49	0,209	56	0,339	74	0,515	47	0,640	23	0,773	34	0,925	332	0,71
0,04 ‡	55	0,066	53	0,215	75	0,338	98	0,500	51	0,641	26	0,774	41	0,935	399	0,72
0,05 ‡	55	0,066	57	0,216	81	0,337	113	0,504	64	0,631	26	0,774	45	0,931	441	0,73
0,06 ‡	60	0,065	57	0,216	81	0,337	114	0,503	66	0,630	26	0,774	45	0,931	449	0,76
0,07 ‡	63	0,068	60	0,216	85	0,339	115	0,503	68	0,630	26	0,774	53	0,927	470	0,77
0,08 ‡	65	0,069	82	0,215	117	0,342	131	0,498	68	0,630	26	0,774	59	0,927	548	0,78
0,09 ‡	65	0,069	93	0,214	125	0,340	133	0,498	72	0,629	26	0,774	60	0,927	574	0,80
0,10 ‡	67	0,069	99	0,213	138	0,340	160	0,503	74	0,630	31	0,773	67	0,921	636	0,82
0,15 ‡	80	0,069	167	0,204	162	0,341	168	0,505	104	0,633	38	0,781	95	0,911	814	0,87
0,20 ‡	101	0,071	290	0,202	277	0,341	243	0,508	141	0,624	43	0,776	103	0,913	1.198	0,94
0,30 ‡	188	0,087	503	0,207	473	0,352	455	0,510	271	0,630	107	0,798	139	0,909	2.136	1,07

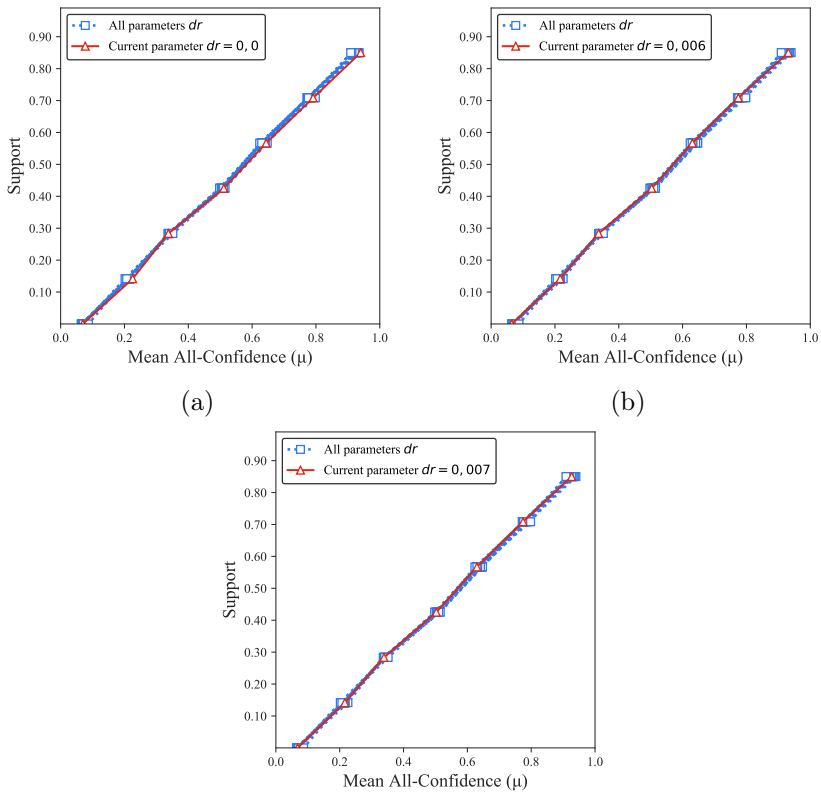


Figura B.5: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00 \times 10^{-1}$, onde essa imagem representa, por similaridade, o comportamento do $dr \in \{0{,}01\,, 0{,}02\,, 0{,}03\,, 0{,}04\,, 0{,}05\}$, (b) com $dr = 0{,}06 \times 10^{-1}$, e (c) com $dr = 0{,}07 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, ,10\,, 0{,}15\,, 0{,}20\,, 0{,}30\}$. Veja Tabela B.7 para detalhes.

Tabela B.8: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**42**	**0,069**	**29**	**0,226**	**42**	**0,338**	**64**	**0,511**	**28**	**0,644**	**8**	**0,791**	**12**	**0,940**	**225**	**0,70**
0,03	47	0,068	33	0,228	43	0,335	72	0,510	38	0,644	16	0,786	28	0,938	277	0,68
0,04	67	0,065	40	0,210	68	0,336	134	0,505	49	0,636	17	0,783	28	0,938	403	0,73
0,05	67	0,065	43	0,208	69	0,336	134	0,505	49	0,636	17	0,783	28	0,938	407	0,75
0,06	76	0,069	45	0,210	72	0,336	152	0,506	71	0,635	36	0,776	61	0,932	513	0,78
0,07	93	0,067	74	0,212	116	0,348	199	0,503	97	0,626	36	0,776	61	0,932	676	0,85
0,08	93	0,067	78	0,212	118	0,347	253	0,505	97	0,626	36	0,776	61	0,932	736	0,84
0,09	107	0,071	89	0,212	147	0,353	272	0,507	142	0,629	52	0,771	120	0,916	929	0,88
0,10	120	0,068	154	0,214	180	0,350	335	0,510	171	0,622	52	0,771	120	0,916	1.132	0,96
0,11	151	0,073	164	0,211	274	0,364	371	0,506	183	0,623	64	0,770	131	0,917	1.338	0,97
0,12	191	0,078	191	0,215	376	0,373	474	0,501	248	0,627	116	0,780	192	0,907	1.788	1,02
0,13	212	0,075	302	0,217	454	0,368	625	0,504	258	0,625	116	0,780	192	0,907	2.159	1,07

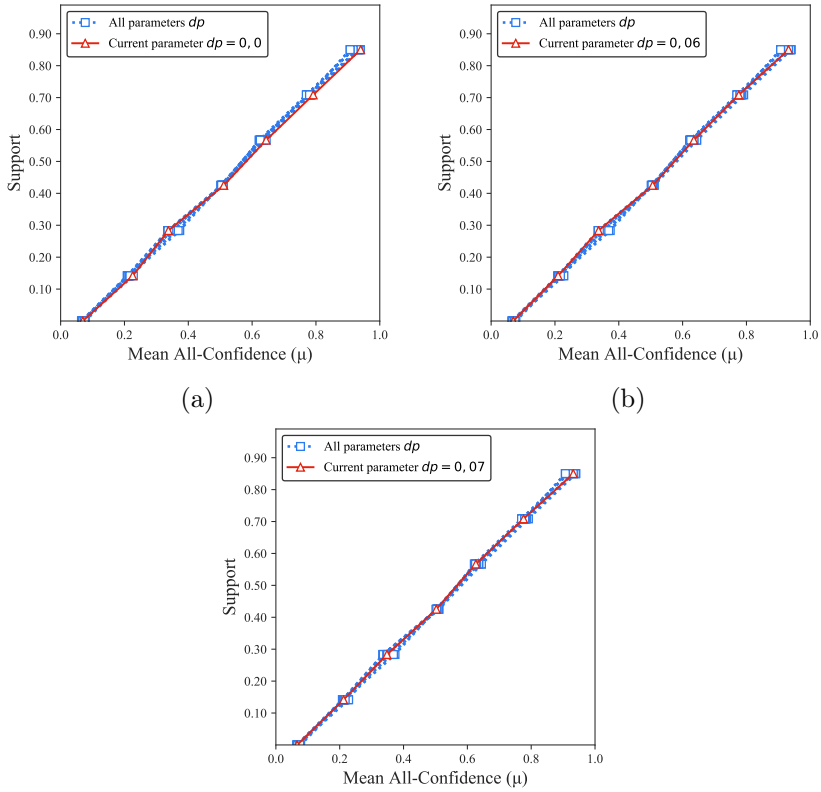


Figura B.6: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, onde essa imagem representa, por similaridade, o comportamento do $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\}$, (b) com $dr = 0{,}06$, e (c) com $dr = 0{,}07$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\}$. Veja Tabela B.8 para detalhes.

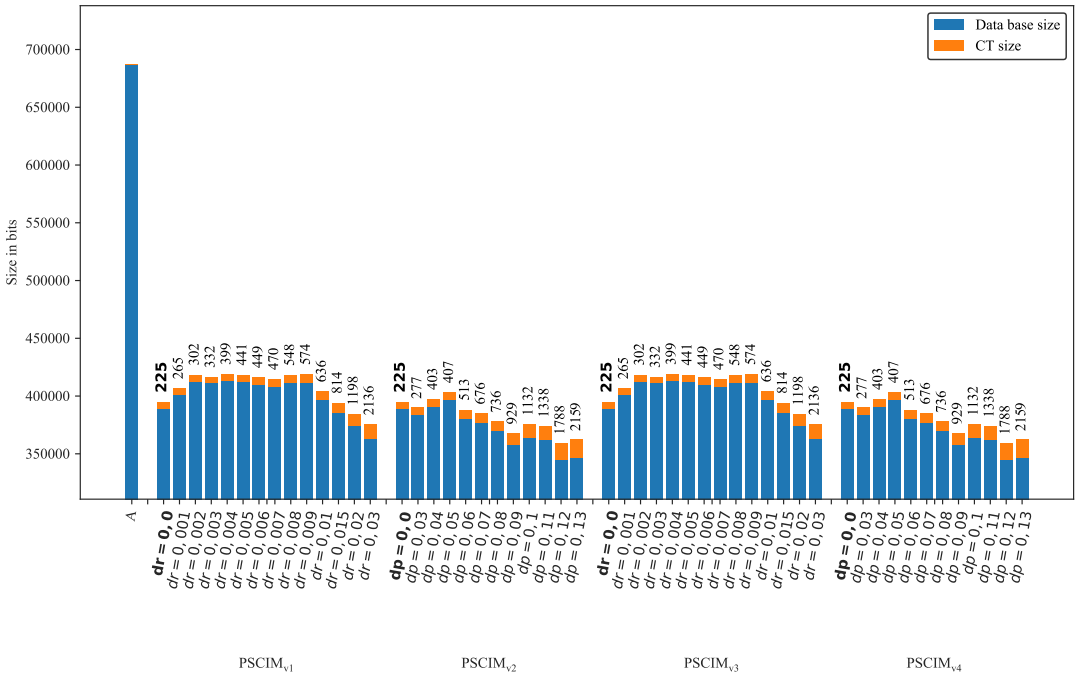


Figura B.7: *Chess*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.2 Kddcup99

Tabela B.9: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v1}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,11]		(0,11 , 0,23]		(0,23 , 0,34]		(0,34 , 0,45]		(0,45 , 0,57]		(0,57 , 0,68]		(0,68 , 0,79]		#	(s)
‡ × 10^{-7}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**1.472**	**0,007**	**113**	**0,633**	**10**	**0,384**	**17**	**0,550**	**37**	**0,637**	**1**	**0,851**	**113**	**0,937**	**1.763**	**5,37**
0,01 ‡	1.472	0,007	139	0,559	10	0,384	17	0,550	37	0,637	1	0,851	113	0,937	1.789	5,35
0,02 ‡	1.472	0,007	139	0,436	10	0,384	17	0,550	37	0,637	1	0,851	113	0,937	1.789	5,34
0,03 ‡	1.472	0,007	141	0,434	15	0,356	17	0,550	37	0,637	1	0,851	113	0,937	1.796	5,35
0,04 ‡	1.472	0,007	182	0,395	20	0,342	17	0,550	37	0,637	1	0,851	135	0,930	1.864	5,35
0,05 ‡	1.472	0,007	182	0,395	20	0,342	17	0,550	37	0,637	2	0,819	135	0,930	1.865	5,36
0,06 ‡	1.472	0,007	184	0,386	20	0,342	17	0,550	37	0,637	2	0,819	135	0,930	1.867	5,31
0,07 ‡	1.472	0,007	184	0,386	20	0,342	17	0,550	37	0,637	2	0,819	135	0,930	1.867	5,26
0,08 ‡	1.472	0,007	246	0,338	20	0,342	17	0,550	37	0,637	32	0,847	143	0,926	1.967	5,27
0,09 ‡	1.472	0,007	246	0,336	20	0,342	21	0,554	40	0,633	32	0,847	143	0,926	1.974	5,27
0,10 ‡	1.472	0,007	253	0,334	20	0,342	22	0,554	166	0,626	54	0,796	143	0,926	2.130	5,28
0,15 ‡	1.768	0,022	433	0,303	51	0,325	251	0,551	711	0,621	54	0,796	143	0,926	3.411	5,30
0,20 ‡	2.237	0,023	537	0,262	108	0,310	339	0,551	929	0,623	54	0,796	143	0,926	4.347	5,33

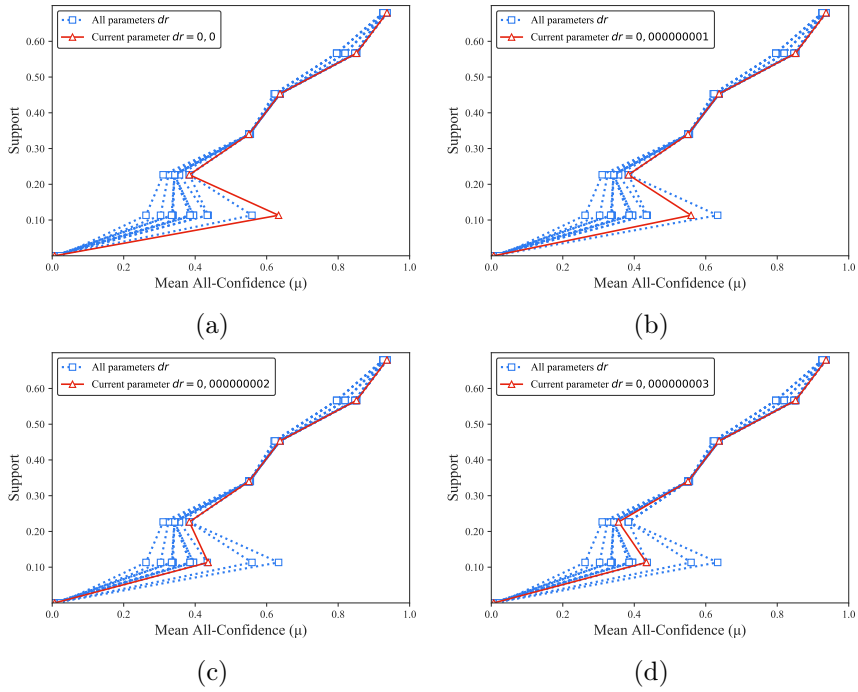


Figura B.8: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v1}}$. (a) com $dr = 0{,}00 \times 10^{-8}$, (b) com $dr = 0{,}01 \times 10^{-7}$, (c) com $dr = 0{,}02 \times 10^{-7}$ e (d) com $dr = 0{,}03 \times 10^{-7}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}15\,, 0{,}20\}$. Veja Tabela B.9 para detalhes.

Tabela B.10: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v2}}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,11]		(0,11 , 0,23]		(0,23 , 0,34]		(0,34 , 0,45]		(0,45 , 0,57]		(0,57 , 0,68]		(0,68 , 0,79]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**1.472**	**0,007**	**113**	**0,633**	**10**	**0,384**	**17**	**0,550**	**37**	**0,637**	**1**	**0,851**	**113**	**0,937**	**1.763**	**5,35**
0,02	1.472	0,007	141	0,549	14	0,365	17	0,550	38	0,636	2	0,788	114	0,936	1.798	5,36
0,03	1.499	0,005	146	0,287	14	0,365	17	0,550	74	0,632	3	0,788	145	0,927	1.898	5,37
0,04	1.516	0,005	148	0,287	22	0,342	17	0,550	80	0,634	21	0,827	145	0,927	1.949	5,35
0,05	1.566	0,007	187	0,262	22	0,342	31	0,549	174	0,628	57	0,798	153	0,924	2.190	5,36
0,06	1.591	0,007	256	0,262	33	0,329	32	0,550	393	0,622	67	0,786	153	0,924	2.525	5,38
0,07	1.832	0,007	301	0,261	33	0,329	107	0,560	717	0,623	76	0,779	153	0,924	3.219	5,41
0,08	1.872	0,007	360	0,262	40	0,323	235	0,556	909	0,623	76	0,779	153	0,924	3.645	5,44
0,09	1.940	0,008	396	0,262	41	0,323	313	0,553	980	0,622	92	0,768	153	0,924	3.915	5,45
0,10	2.206	0,017	400	0,260	60	0,316	331	0,553	982	0,623	92	0,768	153	0,924	4.224	5,44

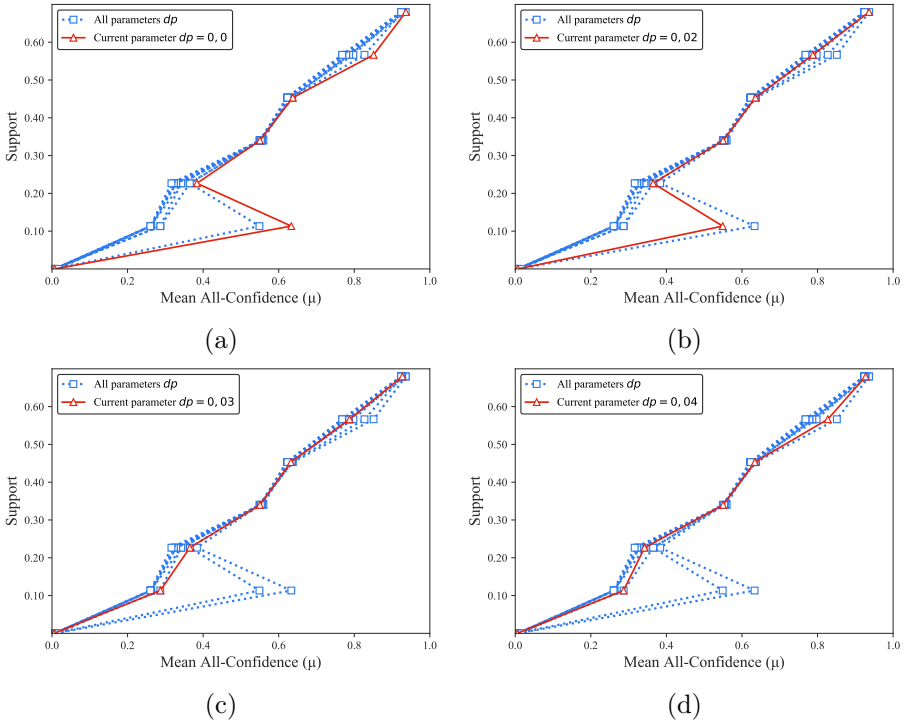


Figura B.9: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}02$, (c) com $dr = 0{,}03$ e (b) com $dr = 0{,}04$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\}$. Veja Tabela B.10 para detalhes.

Tabela B.11: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,11]		(0,11 , 0,23]		(0,23 , 0,34]		(0,34 , 0,45]		(0,45 , 0,57]		(0,57 , 0,68]		(0,68 , 0,79]		#	(s)
‡ $\times 10^{-7}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**1.472**	**0,007**	**113**	**0,633**	**10**	**0,384**	**17**	**0,550**	**37**	**0,637**	**1**	**0,851**	**113**	**0,937**	**1.763**	**5,35**
0,01 ‡	1.472	0,007	139	0,559	10	0,384	17	0,550	37	0,637	1	0,851	113	0,937	1.789	5,29
0,02 ‡	1.472	0,007	139	0,436	10	0,384	17	0,550	37	0,637	1	0,851	113	0,937	1.789	5,34
0,03 ‡	1.472	0,007	141	0,434	15	0,356	17	0,550	37	0,637	1	0,851	113	0,937	1.796	5,36
0,04 ‡	1.472	0,007	182	0,395	20	0,342	17	0,550	37	0,637	1	0,851	135	0,930	1.864	5,36
0,05 ‡	1.472	0,007	182	0,395	20	0,342	17	0,550	37	0,637	2	0,819	135	0,930	1.865	5,35
0,06 ‡	1.472	0,007	184	0,386	20	0,342	17	0,550	37	0,637	2	0,819	135	0,930	1.867	5,35
0,07 ‡	1.472	0,007	184	0,386	20	0,342	17	0,550	37	0,637	2	0,819	135	0,930	1.867	5,30
0,08 ‡	1.472	0,007	246	0,338	20	0,342	17	0,550	37	0,637	32	0,847	143	0,926	1.967	5,34
0,09 ‡	1.472	0,007	246	0,336	20	0,342	21	0,554	40	0,633	32	0,847	143	0,926	1.974	5,34
0,10 ‡	1.472	0,007	253	0,334	20	0,342	22	0,554	166	0,626	54	0,796	143	0,926	2.130	5,35
0,15 ‡	1.768	0,022	433	0,303	51	0,325	251	0,551	711	0,621	54	0,796	143	0,926	3.411	5,38
0,20 ‡	2.237	0,023	537	0,262	108	0,310	339	0,551	929	0,623	54	0,796	143	0,926	4.347	5,35

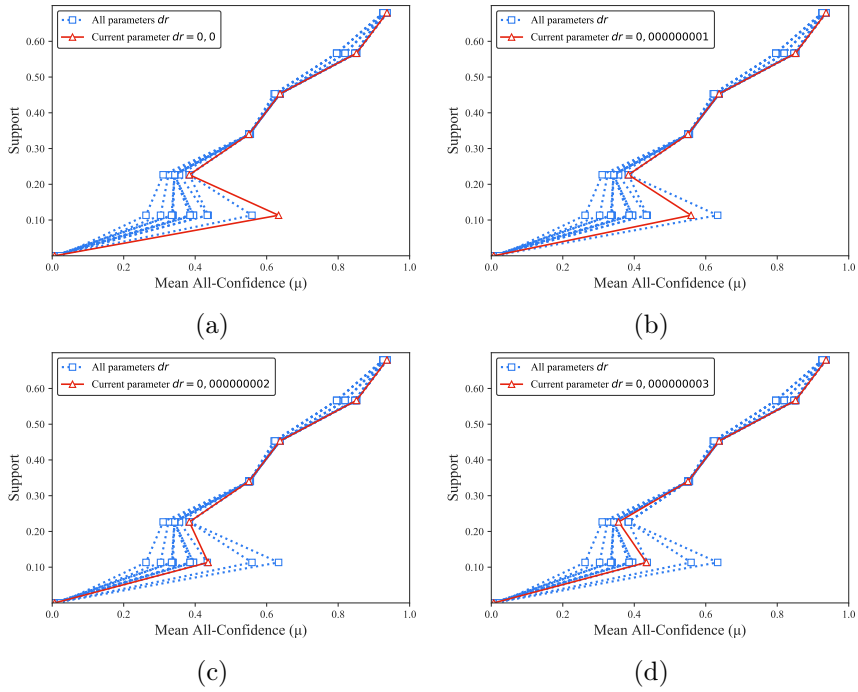


Figura B.10: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v3}}$. (a) com $dr = 0{,}00 \times 10^{-7}$, (b) com $dr = 0{,}01 \times 10^{-7}$, (c) com $dr = 0{,}02 \times 10^{-7}$ e (d) com $dr = 0{,}03 \times 10^{-7}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}15\,, 0{,}20\}$. Veja Tabela B.11 para detalhes.

Tabela B.12: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v4}}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,11]		(0,11 , 0,23]		(0,23 , 0,34]		(0,34 , 0,45]		(0,45 , 0,57]		(0,57 , 0,68]		(0,68 , 0,79]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**1.472**	**0,007**	**113**	**0,633**	**10**	**0,384**	**17**	**0,550**	**37**	**0,637**	**1**	**0,851**	**113**	**0,937**	**1.763**	**5,35**
0,02	1.472	0,007	141	0,549	14	0,365	17	0,550	38	0,636	2	0,788	114	0,936	1.798	5,45
0,03	1.499	0,005	146	0,287	14	0,365	17	0,550	74	0,632	3	0,788	145	0,927	1.898	5,56
0,04	1.516	0,005	148	0,287	22	0,342	17	0,550	80	0,634	21	0,827	145	0,927	1.949	5,40
0,05	1.566	0,007	187	0,262	22	0,342	31	0,549	174	0,628	57	0,798	153	0,924	2.190	5,40
0,06	1.591	0,007	256	0,262	33	0,329	32	0,550	393	0,622	67	0,786	153	0,924	2.525	5,37
0,07	1.832	0,007	301	0,261	33	0,329	107	0,560	717	0,623	76	0,779	153	0,924	3.219	5,44
0,08	1.872	0,007	360	0,262	40	0,323	235	0,556	909	0,623	76	0,779	153	0,924	3.645	5,45
0,09	1.940	0,008	396	0,262	41	0,323	313	0,553	980	0,622	92	0,768	153	0,924	3.915	5,47
0,10	2.206	0,017	400	0,260	60	0,316	331	0,553	982	0,623	92	0,768	153	0,924	4.224	5,51

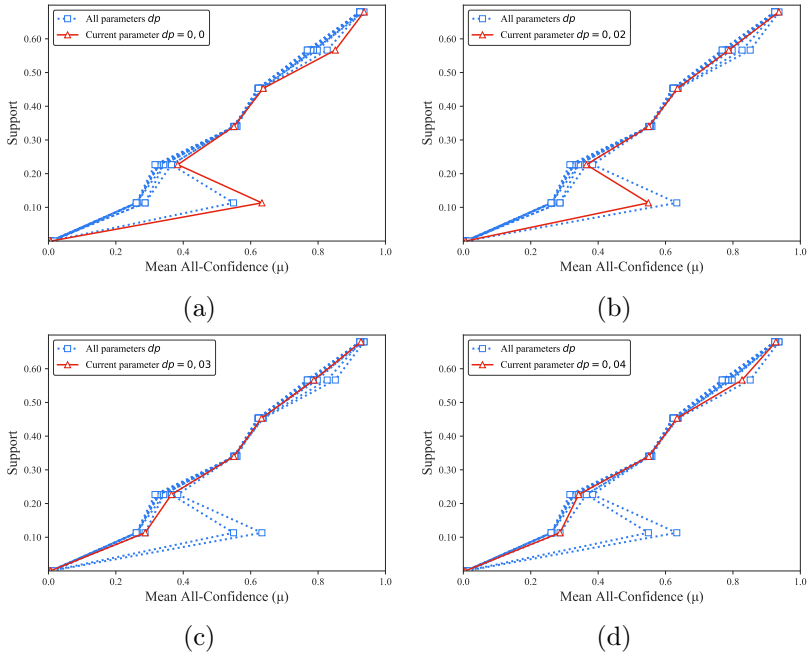


Figura B.11: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}02$, (c) com $dr = 0{,}03$ e (d) com $dr = 0{,}04 \times 10^{-7}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\}$. Veja Tabela B.12 para detalhes.

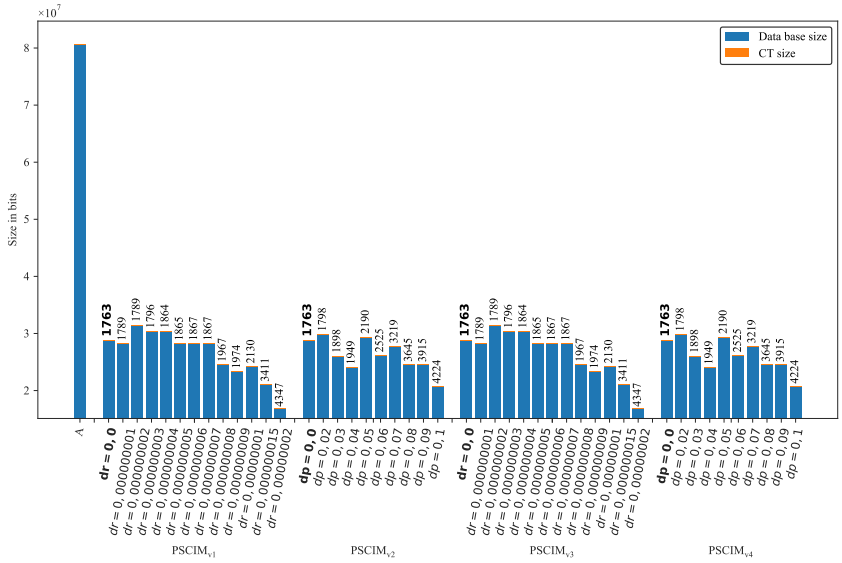
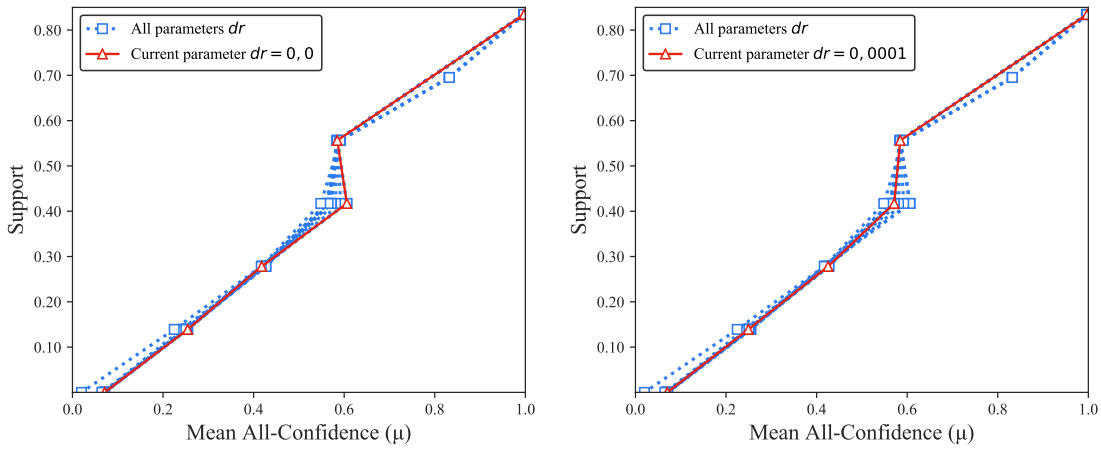


Figura B.12: *Kddcup99*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.3 Mushrooms

Tabela B.13: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,83]		(0,83 , 0,97]		#	(s)
$\ddagger \times 10^{-3}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	194	0,071	84	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	319	0,80
0,01 ‡	194	0,071	85	0,253	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	320	0,80
0,02 ‡	194	0,071	86	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	321	0,80
0,03 ‡	197	0,070	86	0,254	31	0,424	11	0,592	2	0,585	0	0,000	1	0,997	328	0,80
0,04 ‡	197	0,070	86	0,254	32	0,423	12	0,581	2	0,585	0	0,000	1	0,997	330	0,80
0,05 ‡	207	0,070	92	0,252	32	0,423	14	0,562	2	0,585	0	0,000	1	0,997	348	0,81
0,06 ‡	213	0,068	92	0,252	36	0,418	15	0,562	2	0,585	0	0,000	1	0,997	359	0,81
0,07 ‡	224	0,070	96	0,249	36	0,418	16	0,571	2	0,585	0	0,000	1	0,997	375	0,81
0,08 ‡	224	0,070	96	0,249	36	0,418	16	0,571	2	0,585	0	0,000	1	0,997	375	0,82
0,09 ‡	227	0,070	96	0,249	38	0,425	16	0,571	2	0,585	0	0,000	1	0,997	380	0,82
0,10 ‡	227	0,070	97	0,249	39	0,425	16	0,571	2	0,585	0	0,000	1	0,997	382	0,82
0,15 ‡	**289**	**0,071**	**108**	**0,251**	**57**	**0,427**	**19**	**0,562**	**3**	**0,591**	**1**	**0,832**	**1**	**0,997**	**478**	**0,85**
0,20 ‡	538	0,067	140	0,241	87	0,425	19	0,562	3	0,591	1	0,832	1	0,997	789	0,89
0,30 ‡	1.072	0,065	239	0,250	109	0,427	21	0,565	3	0,591	1	0,832	1	0,997	1.446	0,95
1,00 ‡	44.913	0,020	984	0,225	156	0,417	28	0,548	3	0,591	1	0,832	1	0,997	46.086	6,15



(a) (b)

Figura B.13: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}02\,, 0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\}$ e (b) com $dr = 0{,}10 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}11\,, 0{,}12\}$. Veja Tabela B.13 para detalhes.

Tabela B.14: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,83]		(0,83 , 0,97]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	194	0,071	84	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	319	0,80
0,02	**197**	**0,072**	**86**	**0,254**	**43**	**0,431**	**13**	**0,599**	**3**	**0,591**	**1**	**0,832**	**1**	**0,997**	**344**	**0,81**
0,03	255	0,071	111	0,242	48	0,425	17	0,583	3	0,591	1	0,832	1	0,997	436	0,82
0,04	268	0,072	135	0,244	67	0,420	20	0,568	3	0,591	1	0,832	1	0,997	495	0,82
0,05	357	0,067	168	0,237	79	0,412	21	0,564	3	0,591	1	0,832	1	0,997	630	0,83
0,06	390	0,067	207	0,235	103	0,413	24	0,553	3	0,591	1	0,832	1	0,997	729	0,84
0,07	455	0,068	242	0,233	115	0,406	24	0,552	3	0,591	1	0,832	1	0,997	841	0,89
0,08	528	0,068	300	0,234	125	0,403	26	0,552	3	0,591	1	0,832	1	0,997	984	0,88
0,09	618	0,071	334	0,234	134	0,394	26	0,552	3	0,591	1	0,832	1	0,997	1.117	0,89
0,10	844	0,067	402	0,231	141	0,392	26	0,552	3	0,591	1	0,832	1	0,997	1.418	0,92
0,11	1.009	0,068	436	0,228	146	0,391	26	0,552	3	0,591	1	0,832	1	0,997	1.622	0,93
0,12	1.329	0,070	530	0,227	149	0,391	27	0,548	3	0,591	1	0,832	1	0,997	2.040	0,96

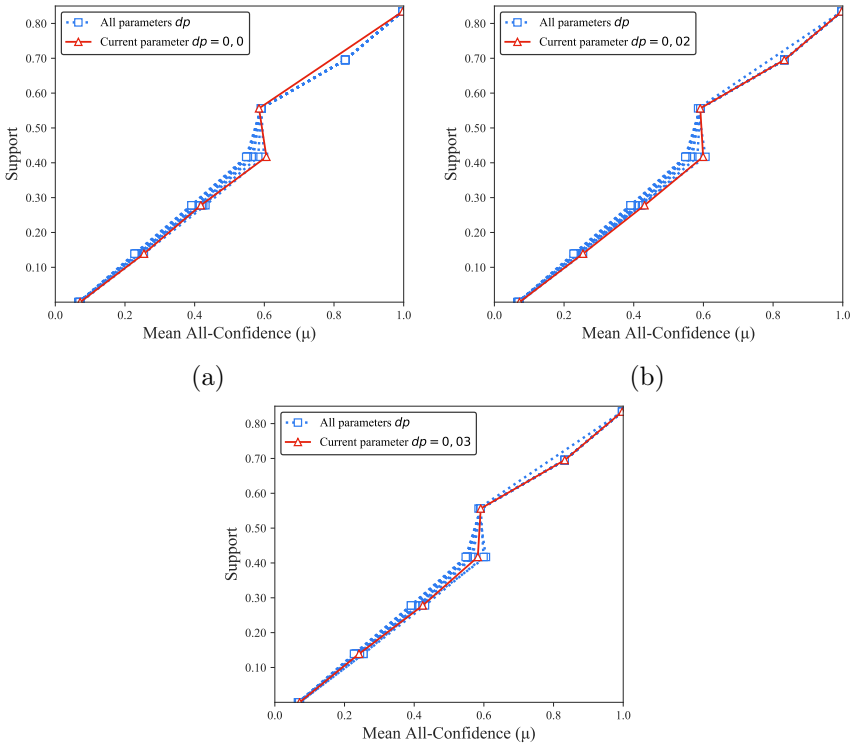
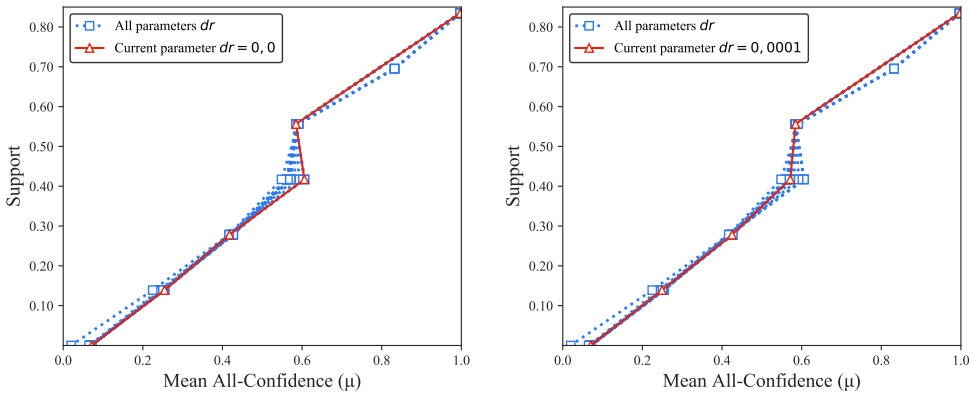


Figura B.14: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}02$ e (c) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\}$. Veja Tabela B.14 para detalhes.

Tabela B.15: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,83]		(0,83 , 0,97]		#	(s)
$\ddagger \times 10^{-3}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	194	0,071	84	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	319	0,74
0,01 ‡	194	0,071	85	0,253	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	320	0,75
0,02 ‡	194	0,071	86	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	321	0,79
0,03 ‡	197	0,070	86	0,254	31	0,424	11	0,592	2	0,585	0	0,000	1	0,997	328	0,75
0,04 ‡	197	0,070	86	0,254	32	0,423	12	0,581	2	0,585	0	0,000	1	0,997	330	0,75
0,05 ‡	207	0,070	92	0,252	32	0,423	14	0,562	2	0,585	0	0,000	1	0,997	348	0,75
0,06 ‡	213	0,068	92	0,252	36	0,418	15	0,562	2	0,585	0	0,000	1	0,997	359	0,75
0,07 ‡	224	0,070	96	0,249	36	0,418	16	0,571	2	0,585	0	0,000	1	0,997	375	0,76
0,08 ‡	224	0,070	96	0,249	36	0,418	16	0,571	2	0,585	0	0,000	1	0,997	375	0,76
0,09 ‡	227	0,070	96	0,249	38	0,425	16	0,571	2	0,585	0	0,000	1	0,997	380	0,76
0,10 ‡	227	0,070	97	0,249	39	0,425	16	0,571	2	0,585	0	0,000	1	0,997	382	0,78
0,15 ‡	**289**	**0,071**	**108**	**0,251**	**57**	**0,427**	**19**	**0,562**	**3**	**0,591**	**1**	**0,832**	**1**	**0,997**	**478**	**0,81**
0,20 ‡	538	0,067	140	0,241	87	0,425	19	0,562	3	0,591	1	0,832	1	0,997	789	0,83
0,30 ‡	1.072	0,065	239	0,250	109	0,427	21	0,565	3	0,591	1	0,832	1	0,997	1.446	0,90
1,00 ‡	44.913	0,020	984	0,225	156	0,417	28	0,548	3	0,591	1	0,832	1	0,997	46.086	6,11

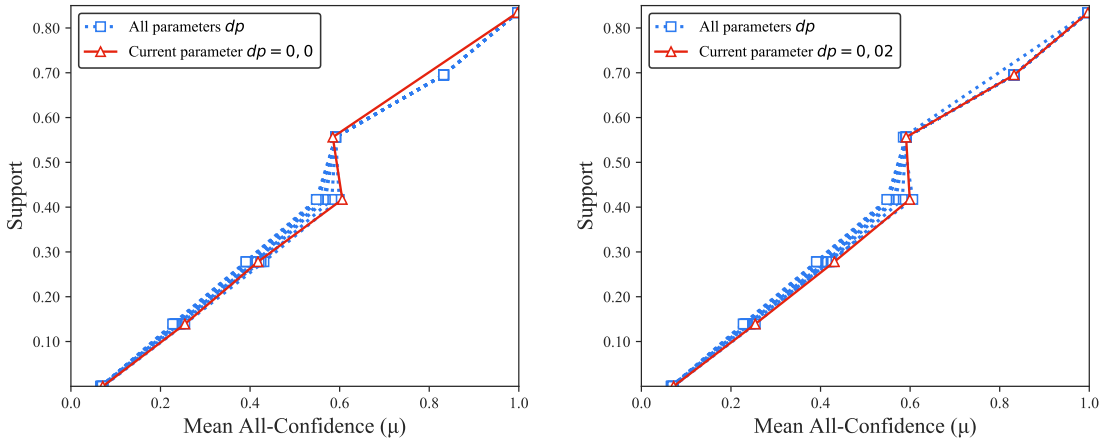


(a) (b)

Figura B.15: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,, 0{,}02\,, 0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\}$ e (b) com $dr = 0{,}10 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}15\,, 0{,}20\,, 0{,}30\,, 1{,}00\}$. Veja Tabela B.15 para detalhes.

Tabela B.16: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,83]		(0,83 , 0,97]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	194	0,071	84	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	319	0,74
0,02	**197**	**0,072**	**86**	**0,254**	**43**	**0,431**	**13**	**0,599**	**3**	**0,591**	**1**	**0,832**	**1**	**0,997**	**344**	**0,75**
0,03	255	0,071	111	0,242	48	0,425	17	0,583	3	0,591	1	0,832	1	0,997	436	0,75
0,04	268	0,072	135	0,244	67	0,420	20	0,568	3	0,591	1	0,832	1	0,997	495	0,76
0,05	357	0,067	168	0,237	79	0,412	21	0,564	3	0,591	1	0,832	1	0,997	630	0,77
0,06	390	0,067	207	0,235	103	0,413	24	0,553	3	0,591	1	0,832	1	0,997	729	0,81
0,07	455	0,068	242	0,233	115	0,406	24	0,552	3	0,591	1	0,832	1	0,997	841	0,79
0,08	528	0,068	300	0,234	125	0,403	26	0,552	3	0,591	1	0,832	1	0,997	984	0,80
0,09	618	0,071	334	0,234	134	0,394	26	0,552	3	0,591	1	0,832	1	0,997	1.117	0,81
0,10	844	0,067	402	0,231	141	0,392	26	0,552	3	0,591	1	0,832	1	0,997	1.418	0,84
0,11	1.009	0,068	436	0,228	146	0,391	26	0,552	3	0,591	1	0,832	1	0,997	1.622	0,85
0,12	1.329	0,070	530	0,227	149	0,391	27	0,548	3	0,591	1	0,832	1	0,997	2.040	0,88



(a) (b)

Figura B.16: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}02$, e (c) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\}$. Veja Tabela B.16 para detalhes.

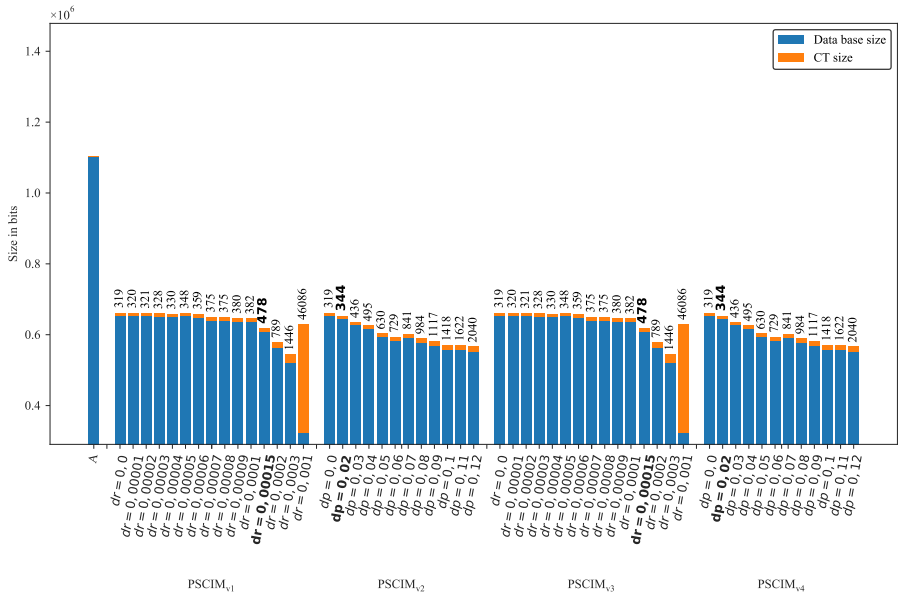
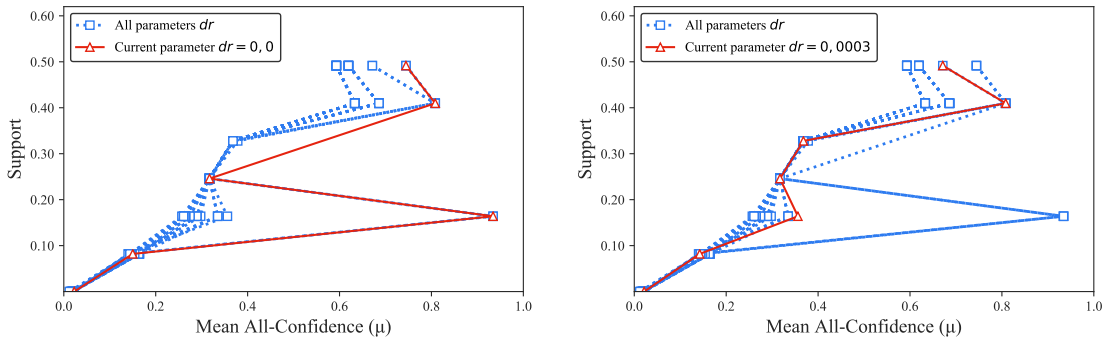


Figura B.17: *Mushrooms*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.4 PowerC

Tabela B.17: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v1}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,08]		(0,08 , 0,16]		(0,16 , 0,25]		(0,25 , 0,33]		(0,33 , 0,41]		(0,41 , 0,49]		(0,49 , 0,57]		#	(s)
‡ $\times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	744	0,023	7	0,150	1	0,934	1	0,317	0	0,000	3	0,808	1	0,745	757	3,39
0,01 ‡	**771**	**0,022**	**8**	**0,142**	**1**	**0,934**	**1**	**0,317**	**2**	**0,378**	**3**	**0,808**	**1**	**0,745**	**787**	**3,38**
0,02 ‡	797	0,022	8	0,142	1	0,934	1	0,317	4	0,368	3	0,808	1	0,745	815	3,38
0,03 ‡	808	0,021	8	0,142	7	0,356	1	0,317	4	0,368	3	0,808	2	0,671	833	3,40
0,04 ‡	847	0,021	8	0,142	8	0,335	1	0,317	4	0,368	3	0,808	2	0,671	873	3,36
0,05 ‡	870	0,021	11	0,140	8	0,335	1	0,317	4	0,368	5	0,686	3	0,619	902	3,41
0,06 ‡	880	0,021	20	0,165	8	0,335	1	0,317	4	0,368	5	0,686	3	0,619	921	3,36
0,07 ‡	891	0,021	21	0,164	11	0,298	1	0,317	4	0,368	5	0,686	3	0,619	936	3,41
0,08 ‡	928	0,020	21	0,164	12	0,289	1	0,317	4	0,368	5	0,686	3	0,619	974	3,37
0,09 ‡	938	0,020	21	0,164	12	0,289	1	0,317	4	0,368	5	0,686	3	0,619	984	3,41
0,10 ‡	954	0,020	21	0,164	12	0,289	1	0,317	4	0,368	5	0,686	3	0,619	1.000	3,36
0,15 ‡	1.025	0,019	21	0,164	14	0,275	1	0,317	4	0,368	5	0,686	3	0,619	1.073	3,41
0,20 ‡	1.178	0,019	21	0,164	14	0,275	1	0,317	4	0,368	7	0,633	4	0,593	1.229	3,36
0,50 ‡	1.518	0,018	32	0,154	16	0,265	1	0,317	4	0,368	7	0,633	4	0,593	1.582	3,43
1,00 ‡	2.541	0,012	41	0,144	19	0,261	1	0,317	4	0,368	7	0,633	4	0,593	2.617	3,38

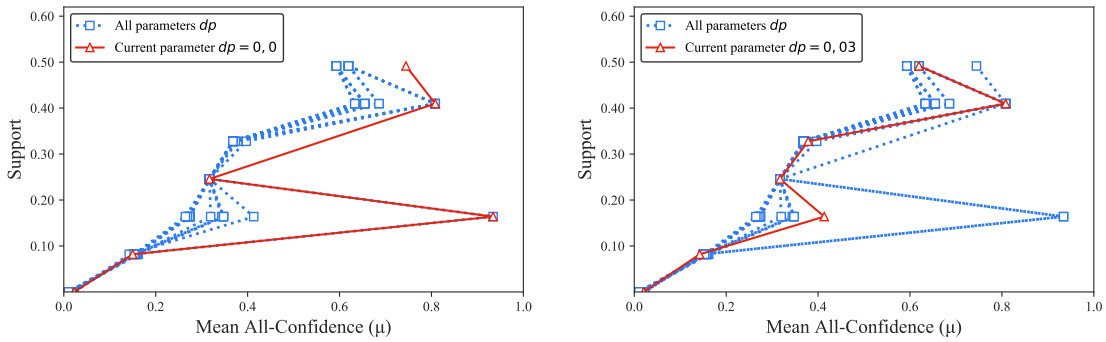


(a) (b)

Figura B.18: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v1}$. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,,0{,}02\}$ e (b) com $dr = 0{,}03 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,,0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\,,0{,}15\,,0{,}20\,,0{,}50\,,1{,}00\}$. Veja Tabela B.17 para detalhes.

Tabela B.18: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v2}$.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,08]		(0,08 , 0,16]		(0,16 , 0,25]		(0,25 , 0,33]		(0,33 , 0,41]		(0,41 , 0,49]		(0,49 , 0,57]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	744	0,023	7	0,150	1	0,934	1	0,317	0	0,000	3	0,808	1	0,745	757	3,41
0,02	**751**	**0,022**	**7**	**0,150**	**1**	**0,934**	**1**	**0,317**	**1**	**0,397**	**3**	**0,808**	**3**	**0,619**	**767**	**3,44**
0,03	843	0,021	8	0,142	4	0,413	1	0,317	2	0,378	3	0,808	3	0,619	864	3,38
0,04	1.037	0,017	12	0,156	6	0,344	1	0,317	3	0,373	3	0,808	3	0,619	1.065	3,44
0,05	1.111	0,017	12	0,156	7	0,348	1	0,317	4	0,368	3	0,808	3	0,619	1.141	3,39
0,06	1.276	0,015	21	0,162	7	0,348	1	0,317	4	0,368	5	0,686	3	0,619	1.317	3,46
0,07	1.347	0,015	23	0,159	9	0,320	1	0,317	4	0,368	5	0,686	3	0,619	1.392	3,39
0,08	1.572	0,014	24	0,158	14	0,275	1	0,317	4	0,368	6	0,655	4	0,593	1.625	3,45
0,09	1.715	0,013	24	0,158	14	0,275	1	0,317	4	0,368	6	0,655	4	0,593	1.768	3,41
0,10	1.861	0,012	24	0,158	16	0,273	1	0,317	4	0,368	6	0,655	4	0,593	1.916	3,47
0,11	2.126	0,011	24	0,158	16	0,273	1	0,317	4	0,368	6	0,655	4	0,593	2.181	3,44
0,12	2.407	0,011	27	0,156	16	0,273	1	0,317	4	0,368	7	0,633	4	0,593	2.466	3,46
0,13	2.570	0,010	27	0,156	17	0,269	1	0,317	4	0,368	7	0,633	4	0,593	2.630	3,42
0,14	2.777	0,010	33	0,153	18	0,265	1	0,317	4	0,368	7	0,633	4	0,593	2.844	3,47

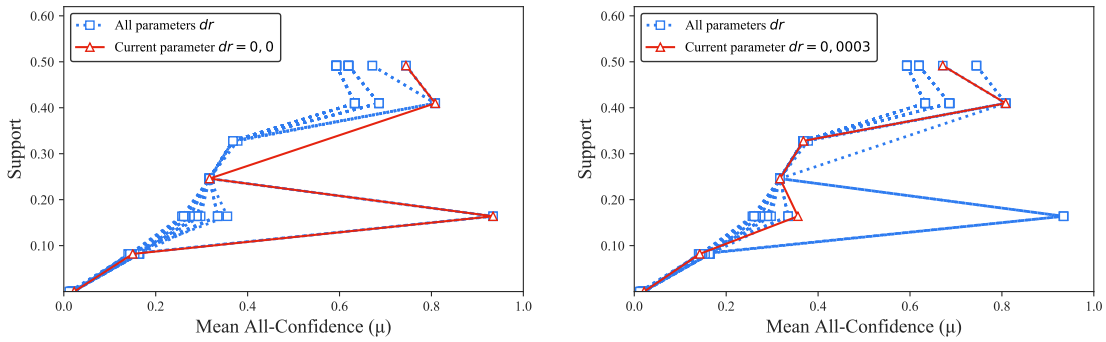


(a) (b)

Figura B.19: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,,0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,,0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\,,0{,}11\,,0{,}12\,,0{,}13\,,0{,}14\}$. Veja Tabela B.18 para detalhes.

Tabela B.19: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,08]		(0,08 , 0,16]		(0,16 , 0,25]		(0,25 , 0,33]		(0,33 , 0,41]		(0,41 , 0,49]		(0,49 , 0,57]		#	(s)
‡ × 10^{-2}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	744	0,023	7	0,150	1	0,934	1	0,317	0	0,000	3	0,808	1	0,745	757	3,37
0,01 ‡	**771**	**0,022**	**8**	**0,142**	**1**	**0,934**	**1**	**0,317**	**2**	**0,378**	**3**	**0,808**	**1**	**0,745**	**787**	**3,38**
0,02 ‡	797	0,022	8	0,142	1	0,934	1	0,317	4	0,368	3	0,808	1	0,745	815	3,37
0,03 ‡	808	0,021	8	0,142	7	0,356	1	0,317	4	0,368	3	0,808	2	0,671	833	3,36
0,04 ‡	847	0,021	8	0,142	8	0,335	1	0,317	4	0,368	3	0,808	2	0,671	873	3,37
0,05 ‡	870	0,021	11	0,140	8	0,335	1	0,317	4	0,368	5	0,686	3	0,619	902	3,37
0,06 ‡	880	0,021	20	0,165	8	0,335	1	0,317	4	0,368	5	0,686	3	0,619	921	3,37
0,07 ‡	891	0,021	21	0,164	11	0,298	1	0,317	4	0,368	5	0,686	3	0,619	936	3,37
0,08 ‡	928	0,020	21	0,164	12	0,289	1	0,317	4	0,368	5	0,686	3	0,619	974	3,36
0,09 ‡	938	0,020	21	0,164	12	0,289	1	0,317	4	0,368	5	0,686	3	0,619	984	3,37
0,10 ‡	954	0,020	21	0,164	12	0,289	1	0,317	4	0,368	5	0,686	3	0,619	1.000	3,37
0,15 ‡	1.025	0,019	21	0,164	14	0,275	1	0,317	4	0,368	5	0,686	3	0,619	1.073	3,38
0,20 ‡	1.178	0,019	21	0,164	14	0,275	1	0,317	4	0,368	7	0,633	4	0,593	1.229	3,38
0,50 ‡	1.518	0,018	32	0,154	16	0,265	1	0,317	4	0,368	7	0,633	4	0,593	1.582	3,38
1,00 ‡	2.541	0,012	41	0,144	19	0,261	1	0,317	4	0,368	7	0,633	4	0,593	2.617	3,40

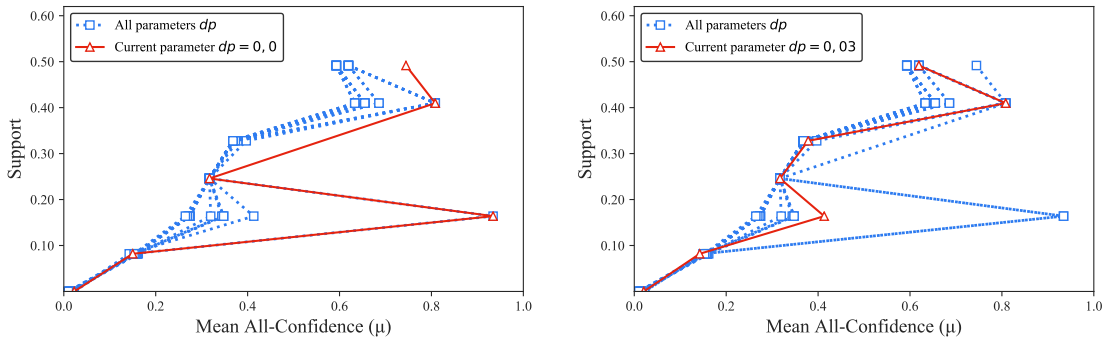


(a) (b)

Figura B.20: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v3}$. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,,0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,,0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\,,0{,}15\,,0{,}20\,,0{,}50\,,1{,}00\}$. Veja Tabela B.19 para detalhes.

Tabela B.20: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v4}$.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,08]		(0,08 , 0,16]		(0,16 , 0,25]		(0,25 , 0,33]		(0,33 , 0,41]		(0,41 , 0,49]		(0,49 , 0,57]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	744	0,023	7	0,150	1	0,934	1	0,317	0	0,000	3	0,808	1	0,745	757	3,38
0,02	**751**	**0,022**	**7**	**0,150**	**1**	**0,934**	**1**	**0,317**	**1**	**0,397**	**3**	**0,808**	**3**	**0,619**	**767**	**3,45**
0,03	843	0,021	8	0,142	4	0,413	1	0,317	2	0,378	3	0,808	3	0,619	864	3,38
0,04	1.037	0,017	12	0,156	6	0,344	1	0,317	3	0,373	3	0,808	3	0,619	1.065	3,43
0,05	1.111	0,017	12	0,156	7	0,348	1	0,317	4	0,368	3	0,808	3	0,619	1.141	3,39
0,06	1.276	0,015	21	0,162	7	0,348	1	0,317	4	0,368	5	0,686	3	0,619	1.317	3,44
0,07	1.347	0,015	23	0,159	9	0,320	1	0,317	4	0,368	5	0,686	3	0,619	1.392	3,40
0,08	1.572	0,014	24	0,158	14	0,275	1	0,317	4	0,368	6	0,655	4	0,593	1.625	3,43
0,09	1.715	0,013	24	0,158	14	0,275	1	0,317	4	0,368	6	0,655	4	0,593	1.768	3,39
0,10	1.861	0,012	24	0,158	16	0,273	1	0,317	4	0,368	6	0,655	4	0,593	1.916	3,44
0,11	2.126	0,011	24	0,158	16	0,273	1	0,317	4	0,368	6	0,655	4	0,593	2.181	3,41
0,12	2.407	0,011	27	0,156	16	0,273	1	0,317	4	0,368	7	0,633	4	0,593	2.466	3,45
0,13	2.570	0,010	27	0,156	17	0,269	1	0,317	4	0,368	7	0,633	4	0,593	2.630	3,46
0,14	2.777	0,010	33	0,153	18	0,265	1	0,317	4	0,368	7	0,633	4	0,593	2.844	3,41



(a) (b)

Figura B.21: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,, 0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\,, 0{,}14\}$. Veja Tabela B.20 para detalhes.

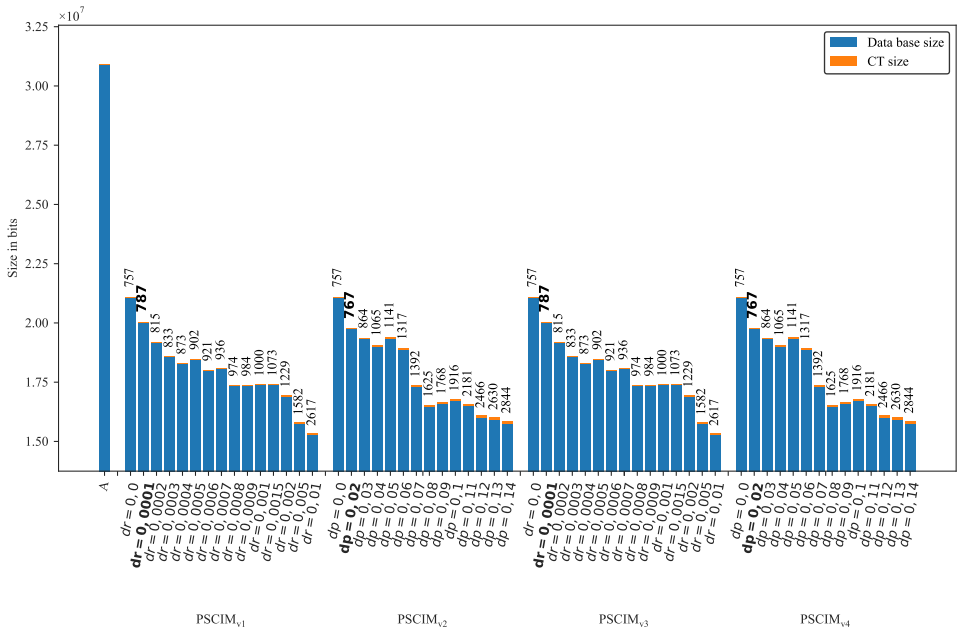
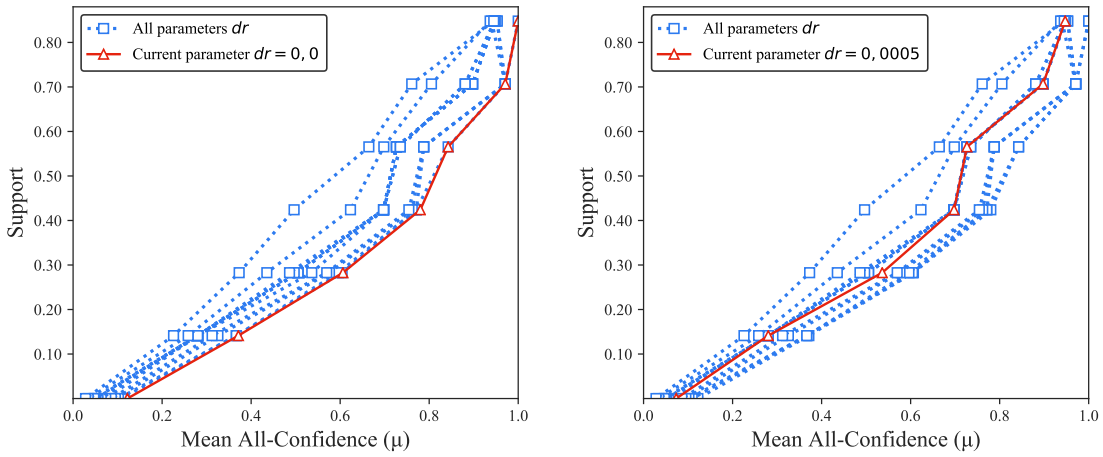


Figura B.22: *PowerC*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.5 Pumsb

Tabela B.21: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**1.135**	**0,122**	**69**	**0,371**	**33**	**0,606**	**31**	**0,780**	**4**	**0,843**	**5**	**0,971**	**4**	**1,000**	**1.281**	**189,09**
0,01 ‡	1.282	0,109	71	0,367	33	0,604	32	0,771	4	0,843	5	0,971	4	1,000	1.431	190,07
0,02 ‡	1.330	0,102	71	0,367	34	0,596	32	0,771	5	0,788	5	0,971	5	0,953	1.482	190,31
0,03 ‡	1.397	0,095	105	0,325	37	0,576	34	0,762	5	0,788	5	0,971	5	0,953	1.588	190,45
0,04 ‡	1.487	0,086	110	0,312	38	0,570	35	0,754	5	0,788	5	0,971	9	0,940	1.689	190,34
0,05 ‡	1.637	0,073	126	0,280	45	0,536	46	0,698	9	0,726	8	0,898	13	0,947	1.884	190,98
0,06 ‡	1.787	0,063	137	0,279	54	0,505	46	0,698	9	0,726	8	0,898	13	0,947	2.054	191,08
0,07 ‡	1.937	0,056	139	0,279	54	0,505	46	0,698	9	0,726	8	0,898	13	0,947	2.206	191,01
0,08 ‡	2.036	0,051	140	0,280	54	0,505	46	0,698	9	0,726	8	0,898	13	0,947	2.306	190,97
0,09 ‡	2.113	0,048	140	0,280	58	0,497	46	0,698	11	0,735	9	0,881	13	0,947	2.390	191,06
0,10 ‡	2.188	0,047	141	0,280	68	0,486	46	0,698	11	0,735	9	0,881	13	0,947	2.476	191,16
0,20 ‡	3.267	0,034	204	0,259	114	0,435	70	0,623	47	0,698	23	0,805	17	0,937	3.742	197,45
0,30 ‡	7.288	0,028	548	0,226	590	0,373	622	0,497	158	0,664	75	0,761	69	0,945	9.350	205,43

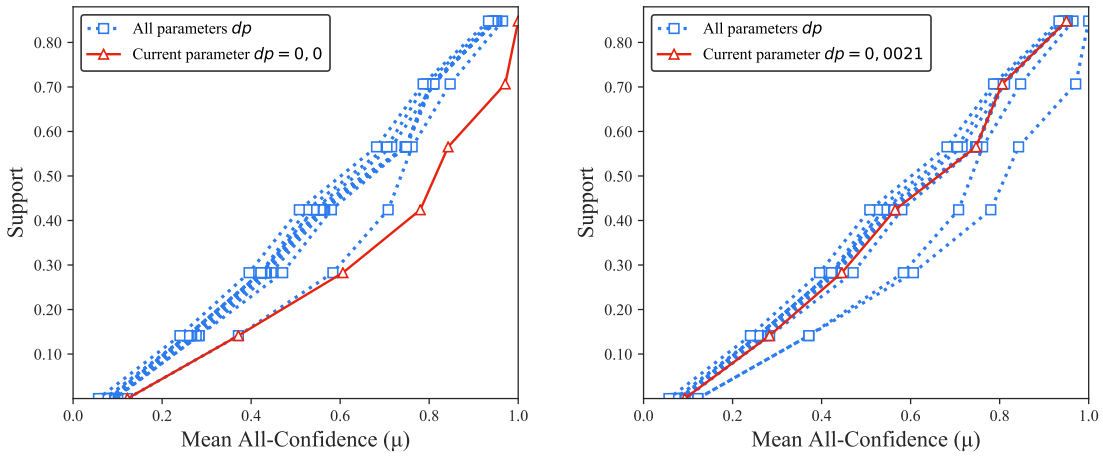


(a) (b)

Figura B.23: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representam, por similaridade, o comportamento de $dr \in \{0{,}01\,, 0{,}02\,, 0{,}03\,, 0{,}04\}$ e (b) com $dr = 0{,}05 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, ,10\,, 0{,}20\,, 0{,}30\}$. Veja Tabela B.21 para detalhes.

Tabela B.22: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**1.135**	**0,122**	**69**	**0,371**	**33**	**0,606**	**31**	**0,780**	**4**	**0,843**	**5**	**0,971**	**4**	**1,000**	**1.281**	**190,46**
0,10 ‡	1.135	0,122	69	0,371	40	0,584	45	0,708	12	0,762	12	0,847	20	0,965	1.333	191,41
0,20 ‡	1.591	0,094	136	0,283	81	0,471	95	0,580	22	0,747	23	0,803	39	0,952	1.987	193,68
0,21 ‡	1.606	0,093	141	0,283	106	0,445	115	0,565	22	0,747	25	0,806	45	0,950	2.060	193,92
0,22 ‡	1.656	0,090	152	0,277	118	0,434	119	0,562	22	0,747	25	0,806	45	0,950	2.137	194,12
0,23 ‡	1.686	0,089	169	0,266	120	0,433	123	0,559	22	0,747	25	0,806	45	0,950	2.190	195,15
0,24 ‡	1.722	0,088	178	0,264	125	0,431	133	0,557	22	0,747	25	0,806	45	0,950	2.250	194,89
0,25 ‡	1.780	0,085	213	0,258	125	0,431	135	0,555	22	0,747	25	0,806	45	0,950	2.345	195,36
0,26 ‡	1.807	0,084	213	0,258	133	0,424	135	0,555	22	0,747	25	0,806	45	0,950	2.380	196,64
0,27 ‡	1.846	0,082	223	0,257	133	0,424	135	0,555	22	0,747	25	0,806	45	0,950	2.429	196,66
0,28 ‡	1.911	0,080	223	0,257	143	0,423	153	0,542	24	0,749	30	0,808	52	0,950	2.536	197,19
0,29 ‡	2.145	0,073	223	0,257	147	0,422	159	0,540	39	0,716	32	0,811	62	0,946	2.807	198,91
0,30 ‡	2.187	0,072	223	0,257	160	0,422	188	0,525	46	0,705	42	0,793	76	0,942	2.922	199,87
0,40 ‡	2.973	0,057	313	0,240	303	0,395	319	0,508	86	0,682	81	0,786	134	0,933	4.209	199,76

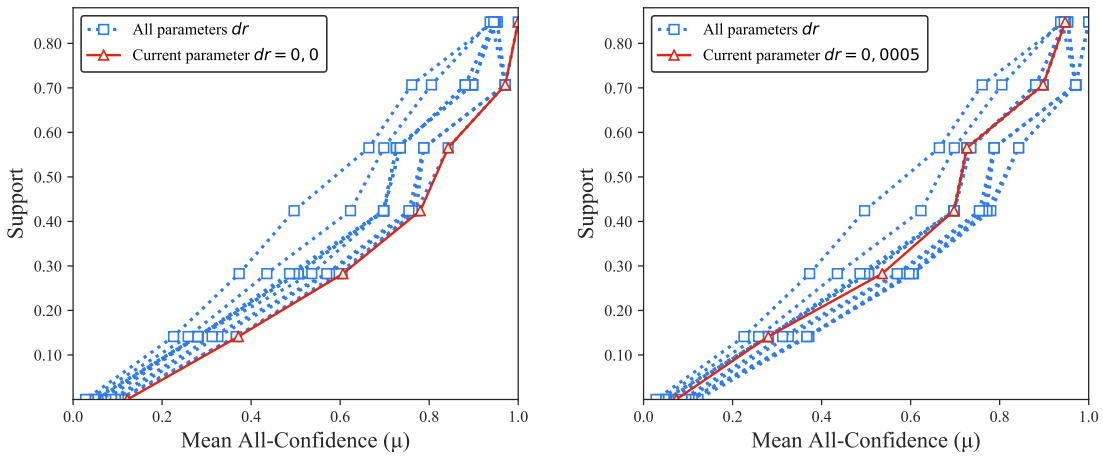


(a) (b)

Figura B.24: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v2}. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,,0{,}20\}$ e (b) com $dr = 0{,}21 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}22\,,0{,}23\,,0{,}24\,,0{,}25\,,0{,}26\,,0{,}27\,,0{,}28\,,0{,}29\,,0{,}30\,,0{,}40\}$. Veja Tabela B.22 para detalhes.

Tabela B.23: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**1.135**	**0,122**	**69**	**0,371**	**33**	**0,606**	**31**	**0,780**	**4**	**0,843**	**5**	**0,971**	**4**	**1,000**	**1.281**	**192,19**
0,01 ‡	1.282	0,109	71	0,367	33	0,604	32	0,771	4	0,843	5	0,971	4	1,000	1.431	192,71
0,02 ‡	1.330	0,102	71	0,367	34	0,596	32	0,771	5	0,788	5	0,971	5	0,953	1.482	192,48
0,03 ‡	1.397	0,095	105	0,325	37	0,576	34	0,762	5	0,788	5	0,971	5	0,953	1.588	192,66
0,04 ‡	1.487	0,086	110	0,312	38	0,570	35	0,754	5	0,788	5	0,971	9	0,940	1.689	192,78
0,05 ‡	1.637	0,073	126	0,280	45	0,536	46	0,698	9	0,726	8	0,898	13	0,947	1.884	193,08
0,06 ‡	1.787	0,063	137	0,279	54	0,505	46	0,698	9	0,726	8	0,898	13	0,947	2.054	192,99
0,07 ‡	1.937	0,056	139	0,279	54	0,505	46	0,698	9	0,726	8	0,898	13	0,947	2.206	193,60
0,08 ‡	2.036	0,051	140	0,280	54	0,505	46	0,698	9	0,726	8	0,898	13	0,947	2.306	193,44
0,09 ‡	2.113	0,048	140	0,280	58	0,497	46	0,698	11	0,735	9	0,881	13	0,947	2.390	193,55
0,10 ‡	2.188	0,047	141	0,280	68	0,486	46	0,698	11	0,735	9	0,881	13	0,947	2.476	193,80
0,20 ‡	3.267	0,034	204	0,259	114	0,435	70	0,623	47	0,698	23	0,805	17	0,937	3.742	199,60
0,30 ‡	7.288	0,028	548	0,226	590	0,373	622	0,497	158	0,664	75	0,761	69	0,945	9.350	206,14

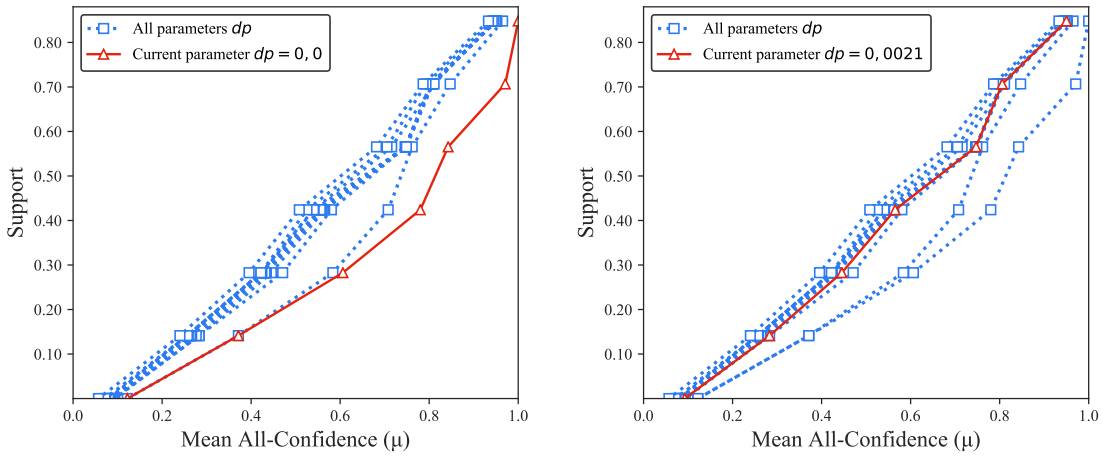


(a) (b)

Figura B.25: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,, 0{,}02\,, 0{,}03\,, 0{,}04\}$ e (b) com $dr = 0{,}05 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}20\,, 0{,}30\}$. Veja Tabela B.23 para detalhes.

Tabela B.24: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v4}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
‡ $\times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**1.135**	**0,122**	**69**	**0,371**	**33**	**0,606**	**31**	**0,780**	**4**	**0,843**	**5**	**0,971**	**4**	**1,000**	**1.281**	**190,07**
0,10 ‡	1.135	0,122	69	0,371	40	0,584	45	0,708	12	0,762	12	0,847	20	0,965	1.333	190,34
0,20 ‡	1.589	0,094	136	0,283	81	0,471	95	0,580	22	0,747	23	0,803	39	0,952	1.985	193,05
0,21 ‡	1.604	0,093	141	0,283	106	0,445	115	0,565	22	0,747	25	0,806	45	0,950	2.058	193,58
0,22 ‡	1.654	0,090	152	0,277	118	0,434	119	0,562	22	0,747	25	0,806	45	0,950	2.135	193,58
0,23 ‡	1.684	0,089	169	0,266	120	0,433	123	0,559	22	0,747	25	0,806	45	0,950	2.188	193,73
0,24 ‡	1.720	0,088	178	0,264	125	0,431	133	0,557	22	0,747	25	0,806	45	0,950	2.248	194,15
0,25 ‡	1.778	0,085	213	0,258	125	0,431	135	0,555	22	0,747	25	0,806	45	0,950	2.343	194,41
0,26 ‡	1.805	0,084	213	0,258	133	0,424	135	0,555	22	0,747	25	0,806	45	0,950	2.378	195,37
0,27 ‡	1.845	0,082	223	0,257	133	0,424	135	0,555	22	0,747	25	0,806	45	0,950	2.428	195,50
0,28 ‡	1.910	0,080	223	0,257	143	0,423	153	0,542	24	0,749	30	0,808	52	0,950	2.535	195,90
0,29 ‡	2.144	0,073	223	0,257	147	0,422	159	0,540	39	0,716	32	0,811	62	0,946	2.806	198,40
0,30 ‡	2.186	0,072	223	0,257	160	0,422	188	0,525	46	0,705	42	0,793	76	0,942	2.921	198,88
0,40 ‡	2.967	0,057	313	0,240	303	0,395	319	0,508	86	0,682	81	0,786	134	0,933	4.203	198,44



(a) (b)

Figura B.26: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v4}$. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}20\}$ e (b) com $dr = 0{,}21 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}22\,, 0{,}23\,, 0{,}24\,, 0{,}25\,, 0{,}26\,, 0{,}27\,, 0{,}28\,, 0{,}28\,, 0{,}29\,, 0{,}30\,, 0{,}40\}$. Veja Tabela B.24 para detalhes.

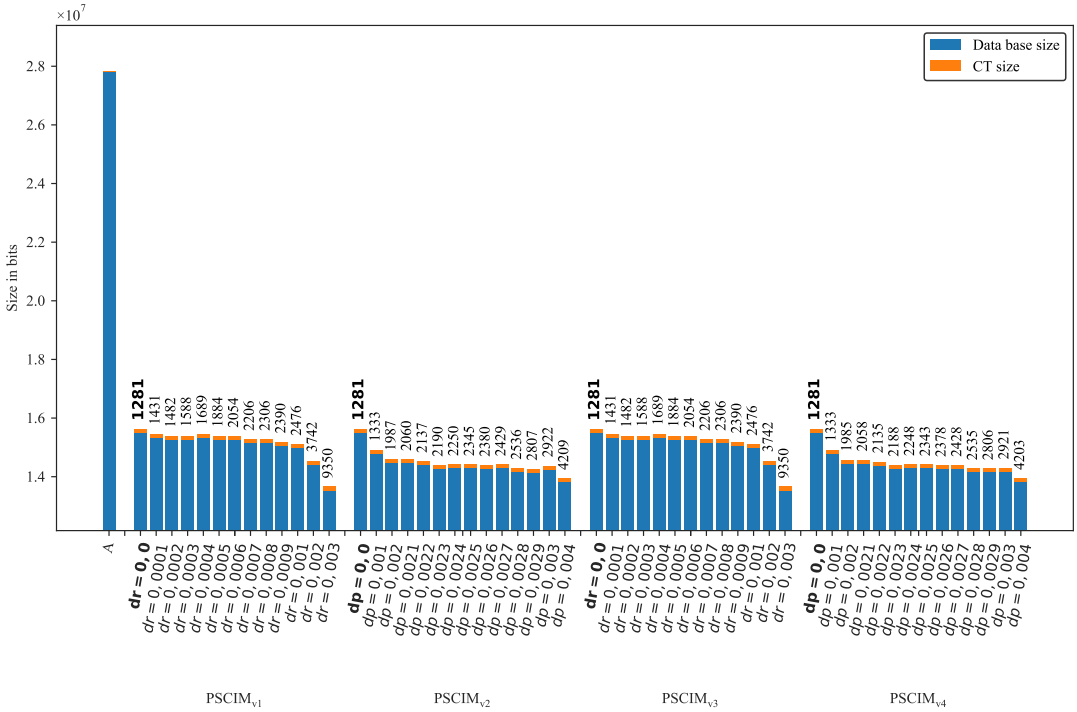
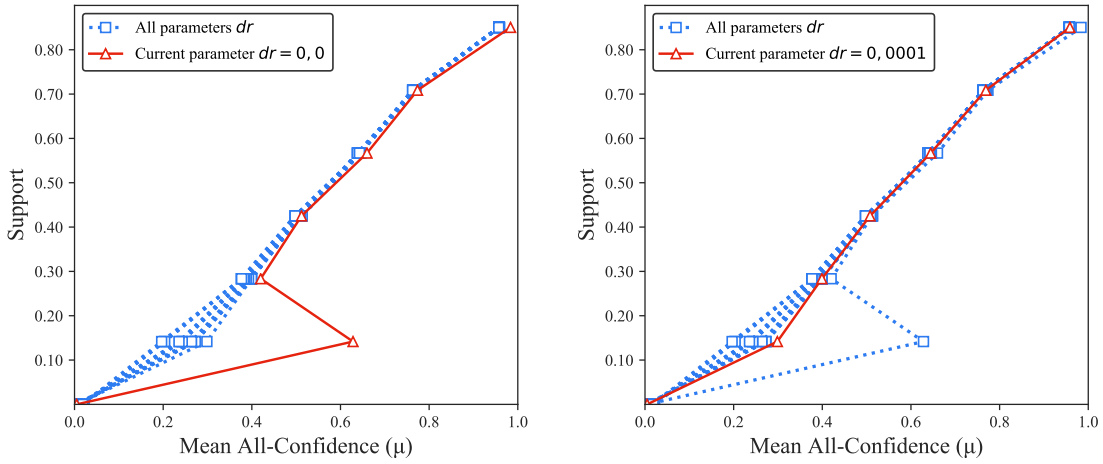


Figura B.27: *Pumsb*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.6 RecordLink

Tabela B.25: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v1}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
‡ × 10^{-2}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**220**	**0,005**	**1**	**0,628**	**15**	**0,420**	**13**	**0,512**	**14**	**0,660**	**10**	**0,774**	**4**	**0,984**	**277**	**2,33**
0,01 ‡	220	0,005	5	0,299	25	0,399	29	0,508	24	0,645	18	0,769	15	0,959	336	2,33
0,02 ‡	235	0,005	9	0,273	25	0,399	29	0,508	24	0,645	18	0,769	15	0,959	355	2,34
0,03 ‡	238	0,006	9	0,273	27	0,399	31	0,511	25	0,643	23	0,771	15	0,959	368	2,34
0,04 ‡	247	0,007	9	0,273	27	0,399	31	0,511	28	0,638	29	0,767	17	0,957	388	2,33
0,05 ‡	249	0,007	11	0,264	27	0,399	31	0,511	28	0,638	29	0,767	17	0,957	392	2,33
0,06 ‡	249	0,007	11	0,264	48	0,393	33	0,512	28	0,638	30	0,766	17	0,957	416	2,34
0,07 ‡	253	0,008	19	0,237	58	0,388	40	0,514	32	0,646	34	0,763	17	0,957	453	2,37
0,08 ‡	265	0,012	23	0,243	64	0,384	52	0,507	38	0,644	34	0,763	17	0,957	493	2,34
0,09 ‡	268	0,011	23	0,243	71	0,384	59	0,510	38	0,644	34	0,763	17	0,957	510	2,33
0,10 ‡	270	0,011	25	0,236	85	0,382	61	0,509	38	0,644	34	0,763	17	0,957	530	2,33
0,15 ‡	412	0,022	83	0,203	172	0,376	86	0,498	38	0,644	34	0,763	17	0,957	842	2,34
0,20 ‡	522	0,021	86	0,203	195	0,377	94	0,497	40	0,643	35	0,763	17	0,957	989	2,35
10,00 ‡	1.597	0,018	189	0,197	207	0,377	94	0,497	40	0,643	35	0,763	17	0,957	2.179	2,38

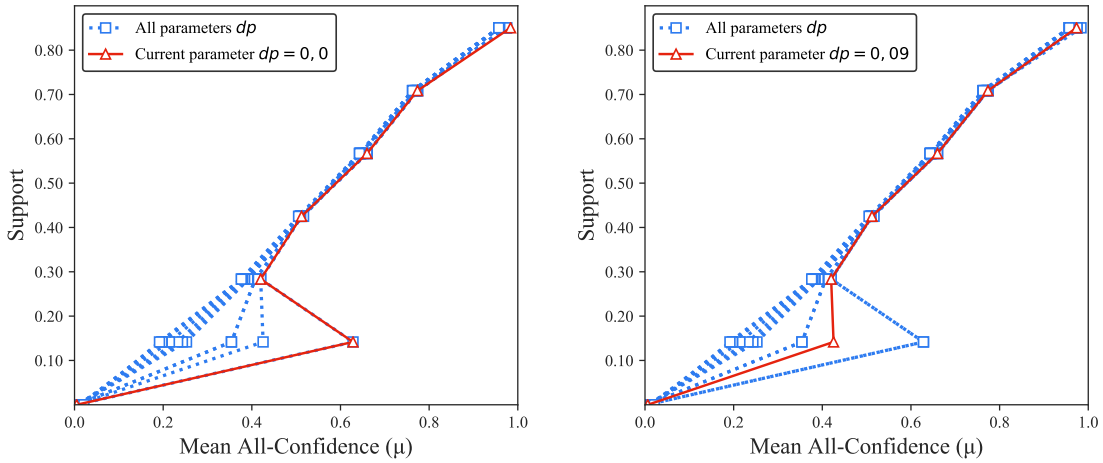


(a) (b)

Figura B.28: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v1}$. (a) com $dr = 0{,}00 \times 10_{-2}$, (b) com $dr = 0{,}01 \times 10_{-2}$ e (c) com $dr = 0{,}02 \times 10_{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}15\,, 0{,}20\,, 10{,}00\}$. Veja Tabela B.25 para detalhes.

Tabela B.26: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v2}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**220**	**0,005**	**1**	**0,628**	**15**	**0,420**	**13**	**0,512**	**14**	**0,660**	**10**	**0,774**	**4**	**0,984**	**277**	**2,37**
0,08	220	0,005	1	0,628	15	0,420	13	0,512	14	0,660	10	0,774	6	0,974	279	2,30
0,09	222	0,006	2	0,425	15	0,420	13	0,512	14	0,660	10	0,774	6	0,974	282	2,41
0,10	222	0,006	3	0,354	21	0,408	22	0,512	14	0,660	10	0,774	9	0,975	301	2,45
0,15	247	0,006	3	0,354	21	0,408	22	0,512	20	0,645	15	0,770	9	0,975	337	2,39
0,20	262	0,006	8	0,253	47	0,393	31	0,511	24	0,645	18	0,770	15	0,959	405	2,39
0,25	271	0,007	12	0,243	49	0,391	33	0,511	25	0,643	23	0,771	15	0,959	428	2,49
0,30	287	0,007	18	0,233	76	0,383	39	0,515	29	0,651	27	0,767	15	0,959	491	2,40
0,35	302	0,009	29	0,214	77	0,383	41	0,517	32	0,646	33	0,764	17	0,957	531	2,35
0,40	369	0,013	68	0,198	124	0,379	60	0,510	35	0,645	35	0,763	17	0,957	708	2,35
0,45	410	0,016	77	0,197	135	0,377	72	0,508	36	0,645	35	0,763	17	0,957	782	2,34
0,50	620	0,021	140	0,192	139	0,376	80	0,505	40	0,643	35	0,763	17	0,957	1.071	2,37



(a) (b)

Figura B.29: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\}$ e (b) com $dr = 0{,}09$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}15\,, 0{,}20\,, 0{,}25\,, 0{,}30\,, 0{,}40\,, 0{,}45\,, 0{,}50\}$. Veja Tabela B.26 para detalhes.

Tabela B.27: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
‡ × 10^{-2}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**220**	**0,005**	**1**	**0,628**	**15**	**0,420**	**13**	**0,512**	**14**	**0,660**	**10**	**0,774**	**4**	**0,984**	**277**	**2,34**
0,01 ‡	220	0,005	5	0,299	25	0,399	29	0,508	24	0,645	18	0,769	15	0,959	336	2,34
0,02 ‡	235	0,005	9	0,273	25	0,399	29	0,508	24	0,645	18	0,769	15	0,959	355	2,31
0,03 ‡	238	0,006	9	0,273	27	0,399	31	0,511	25	0,643	23	0,771	15	0,959	368	2,36
0,04 ‡	247	0,007	9	0,273	27	0,399	31	0,511	28	0,638	29	0,767	17	0,957	388	2,31
0,05 ‡	249	0,007	11	0,264	27	0,399	31	0,511	28	0,638	29	0,767	17	0,957	392	2,32
0,06 ‡	249	0,007	11	0,264	48	0,393	33	0,512	28	0,638	30	0,766	17	0,957	416	2,37
0,07 ‡	253	0,008	19	0,237	58	0,388	40	0,514	32	0,646	34	0,763	17	0,957	453	2,33
0,08 ‡	265	0,012	23	0,243	64	0,384	52	0,507	38	0,644	34	0,763	17	0,957	493	2,32
0,09 ‡	268	0,011	23	0,243	71	0,384	59	0,510	38	0,644	34	0,763	17	0,957	510	2,36
0,10 ‡	270	0,011	25	0,236	85	0,382	61	0,509	38	0,644	34	0,763	17	0,957	530	2,32
0,15 ‡	412	0,022	83	0,203	172	0,376	86	0,498	38	0,644	34	0,763	17	0,957	842	2,32
0,20 ‡	522	0,021	86	0,203	195	0,377	94	0,497	40	0,643	35	0,763	17	0,957	989	2,38
10,00 ‡	1.597	0,018	189	0,197	207	0,377	94	0,497	40	0,643	35	0,763	17	0,957	2.179	2,36

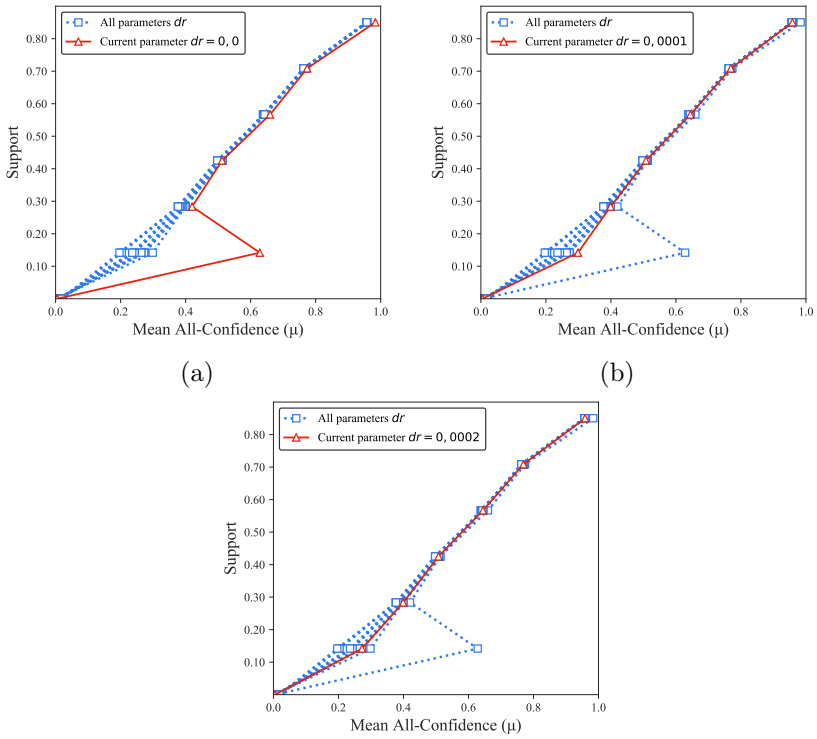
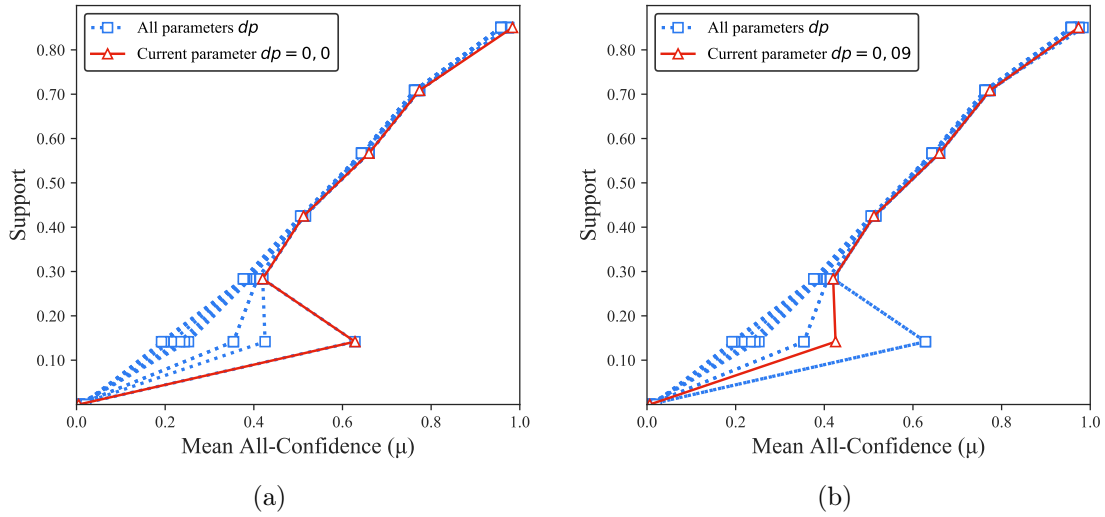


Figura B.30: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v3}$. (a) com $dr = 0{,}00 \times 10_{-2}$, (b) com $dr = 0{,}01 \times 10_{-2}$ e (c) com $dr = 0{,}02 \times 10_{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}20\,, 10{,}00\}$. Veja Tabela B.27 para detalhes.

Tabela B.28: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v4}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**220**	**0,005**	**1**	**0,628**	**15**	**0,420**	**13**	**0,512**	**14**	**0,660**	**10**	**0,774**	**4**	**0,984**	**277**	**2,35**
0,08	220	0,005	1	0,628	15	0,420	13	0,512	14	0,660	10	0,774	6	0,974	279	2,33
0,09	222	0,006	2	0,425	15	0,420	13	0,512	14	0,660	10	0,774	6	0,974	282	2,39
0,10	222	0,006	3	0,354	21	0,408	22	0,512	14	0,660	10	0,774	9	0,975	301	2,34
0,15	247	0,006	3	0,354	21	0,408	22	0,512	20	0,645	15	0,770	9	0,975	337	2,34
0,20	262	0,006	8	0,253	47	0,393	31	0,511	24	0,645	18	0,770	15	0,959	405	2,38
0,25	271	0,007	12	0,243	49	0,391	33	0,511	25	0,643	23	0,771	15	0,959	428	2,34
0,30	287	0,007	18	0,233	76	0,383	39	0,515	29	0,651	27	0,767	15	0,959	491	2,34
0,35	302	0,009	29	0,214	77	0,383	41	0,517	32	0,646	33	0,764	17	0,957	531	2,39
0,40	369	0,013	68	0,198	124	0,379	60	0,510	35	0,645	35	0,763	17	0,957	708	2,35
0,45	410	0,016	77	0,197	135	0,377	72	0,508	36	0,645	35	0,763	17	0,957	782	2,38
0,50	620	0,021	140	0,192	139	0,376	80	0,505	40	0,643	35	0,763	17	0,957	1.071	2,37



(a) (b)

Figura B.31: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\}$ e (b) com $dr = 0{,}09$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}15\,, 0{,}20\,, 0{,}25\,, 0{,}30\,, 0{,}35\,, 0{,}40\,, 0{,}45\,, 0{,}50\}$. Veja Tabela B.28 para detalhes.

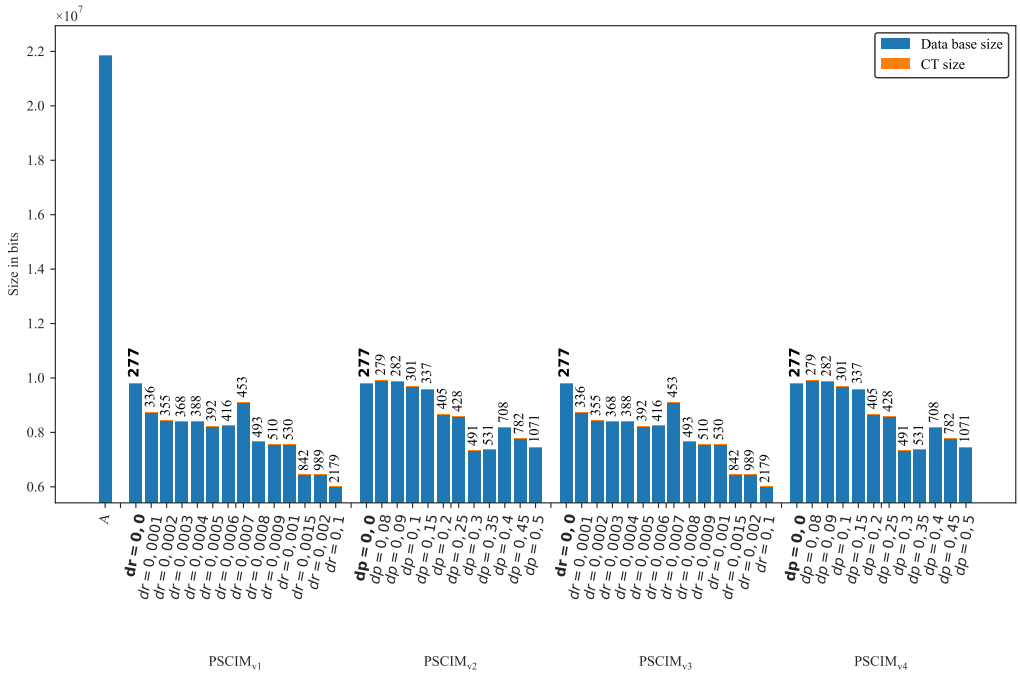


Figura B.32: *RecordLink*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.7 Skin

Tabela B.29: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,06]		(0,06 , 0,13]		(0,13 , 0,19]		(0,19 , 0,26]		(0,26 , 0,32]		(0,32 , 0,38]		(0,38 , 0,45]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,57
0,10	**4**	**0,010**	**3**	**0,208**	**5**	**0,345**	**6**	**0,425**	**3**	**0,685**	**0**	**0,000**	**1**	**0,565**	**22**	**0,56**
0,20	4	0,010	3	0,208	6	0,342	6	0,425	3	0,685	0	0,000	1	0,565	23	0,56
0,30	4	0,010	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,56
0,40	4	0,010	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,56
0,50	6	0,055	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	26	0,56
0,60	6	0,055	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	26	0,56
0,70	6	0,055	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	26	0,56
0,80	8	0,061	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	28	0,56
0,90	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
1,00	18	0,033	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	38	0,56

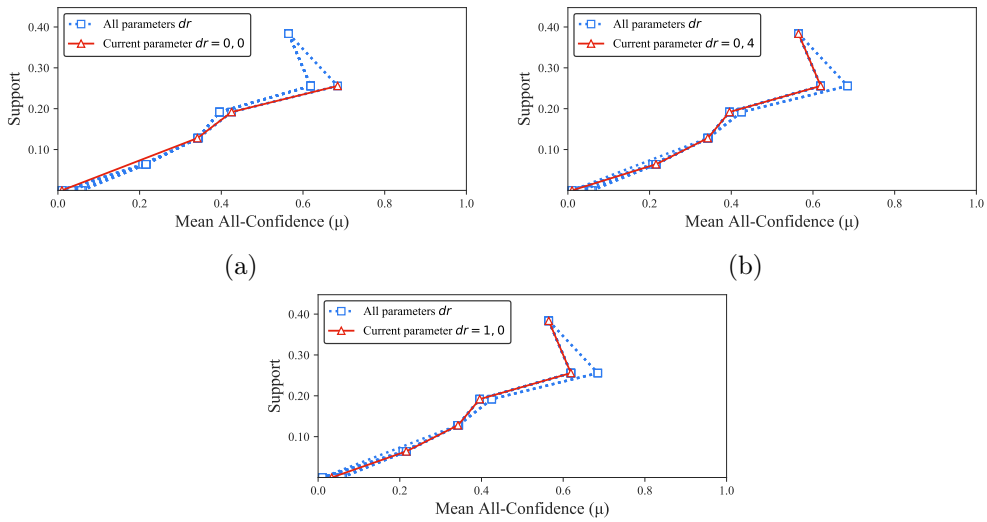


Figura B.33: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}20\,, 0{,}30\}$, (b) com $dr = 0{,}40$, e (c) com $dr = 0{,}50$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}60\,, 0{,}70\,, 0{,}80\,, 0{,}90\,, 1{,}00\}$. Veja Tabela B.29 para detalhes.

Tabela B.30: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v2}$.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,06]		(0,06 , 0,13]		(0,13 , 0,19]		(0,19 , 0,26]		(0,26 , 0,32]		(0,32 , 0,38]		(0,38 , 0,45]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,56
0,10	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,56
0,20	2	0,010	0	0,000	5	0,339	6	0,425	3	0,685	0	0,000	0	0,000	16	0,58
0,30	2	0,010	0	0,000	6	0,342	6	0,425	3	0,685	0	0,000	1	0,565	18	0,62
0,40	**2**	**0,010**	**3**	**0,217**	**6**	**0,342**	**6**	**0,425**	**3**	**0,685**	**0**	**0,000**	**1**	**0,565**	**21**	**0,55**
0,50	3	0,044	3	0,217	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	22	0,55
0,60	5	0,085	3	0,217	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,56
0,70	5	0,085	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	25	0,56
0,80	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
0,90	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
1,00	18	0,033	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	38	0,56

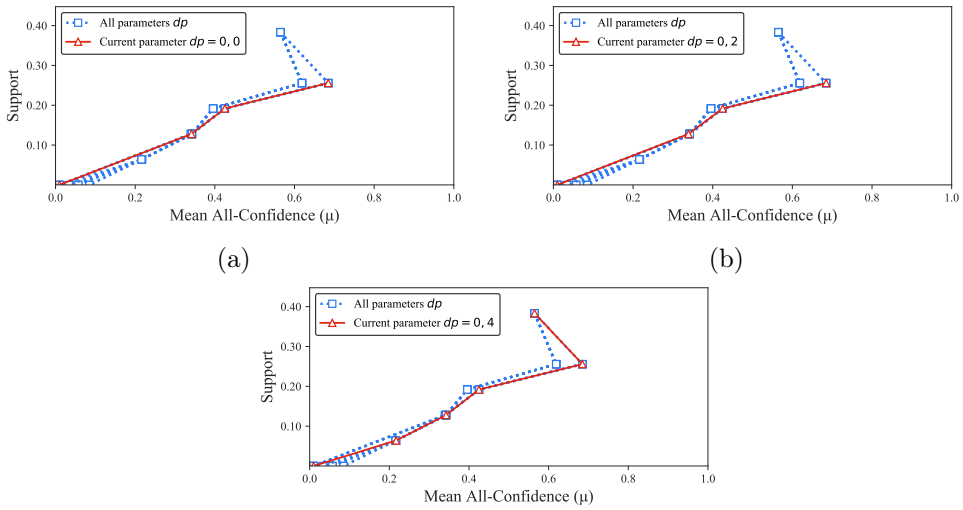


Figura B.34: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\}$, (b) com $dr = 0{,}20$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}30\}$ e (c) com $dr = 0{,}40$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\,, 0{,}60\,, 0{,}70\,, 0{,}80\,, 0{,}90\,, 1{,}00\}$. Veja Tabela B.30 para detalhes.

Tabela B.31: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,06]		(0,06 , 0,13]		(0,13 , 0,19]		(0,19 , 0,26]		(0,26 , 0,32]		(0,32 , 0,38]		(0,38 , 0,45]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,56
0,10	**4**	**0,010**	**3**	**0,208**	**5**	**0,345**	**6**	**0,425**	**3**	**0,685**	**0**	**0,000**	**1**	**0,565**	**22**	**0,56**
0,20	4	0,010	3	0,208	6	0,342	6	0,425	3	0,685	0	0,000	1	0,565	23	0,56
0,30	4	0,010	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,59
0,40	4	0,010	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,57
0,50	6	0,055	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	26	0,56
0,60	6	0,055	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	26	0,56
0,70	6	0,055	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	26	0,56
0,80	8	0,061	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	28	0,56
0,90	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
1,00	18	0,033	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	38	0,56

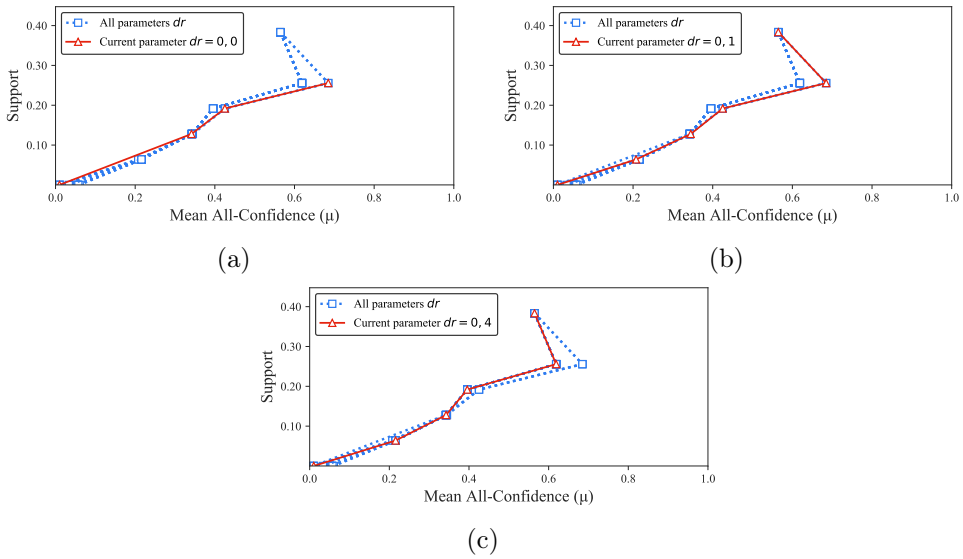


Figura B.35: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}10$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,,0{,}30\}$ e (c) com $dr = 0{,}40$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\,,0{,}60\,,0{,}70\,,0{,}80\,,0{,}90\,,1{,}00\}$. Veja Tabela B.31 para detalhes.

Tabela B.32: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,06]		(0,06 , 0,13]		(0,13 , 0,19]		(0,19 , 0,26]		(0,26 , 0,32]		(0,32 , 0,38]		(0,38 , 0,45]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,57
0,10	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,56
0,20	2	0,010	0	0,000	5	0,339	6	0,425	3	0,685	0	0,000	0	0,000	16	0,56
0,30	2	0,010	0	0,000	6	0,342	6	0,425	3	0,685	0	0,000	1	0,565	18	0,56
0,40	**2**	**0,010**	**3**	**0,217**	**6**	**0,342**	**6**	**0,425**	**3**	**0,685**	**0**	**0,000**	**1**	**0,565**	**21**	**0,56**
0,50	3	0,044	3	0,217	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	22	0,56
0,60	5	0,085	3	0,217	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,56
0,70	5	0,085	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	25	0,56
0,80	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
0,90	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
1,00	18	0,033	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	38	0,61

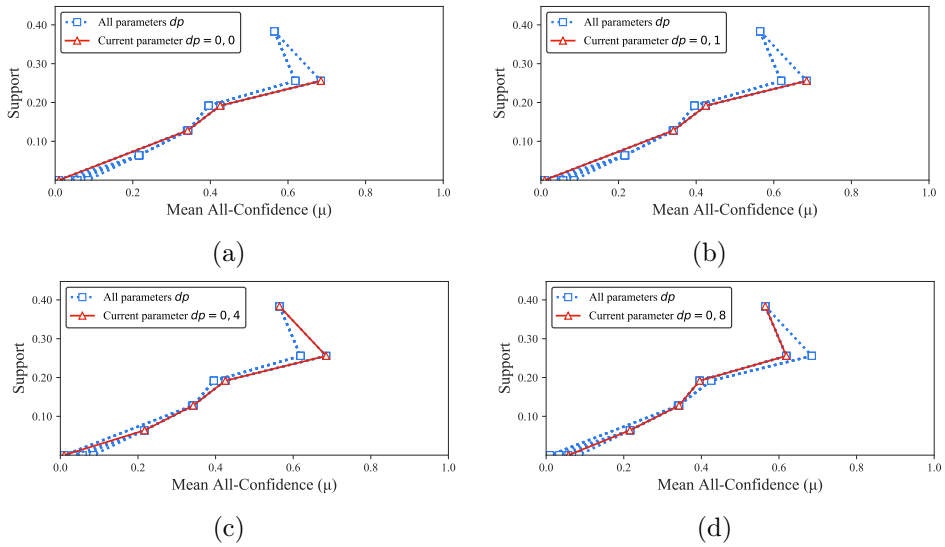


Figura B.36: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}10$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}30\}$, (c) com $dr = 0{,}40$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\,, 0{,}60\,, 0{,}70\}$ e (d) com $dr = 0{,}80$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}90\,, 1{,}00\}$. Veja Tabela B.32 para detalhes.

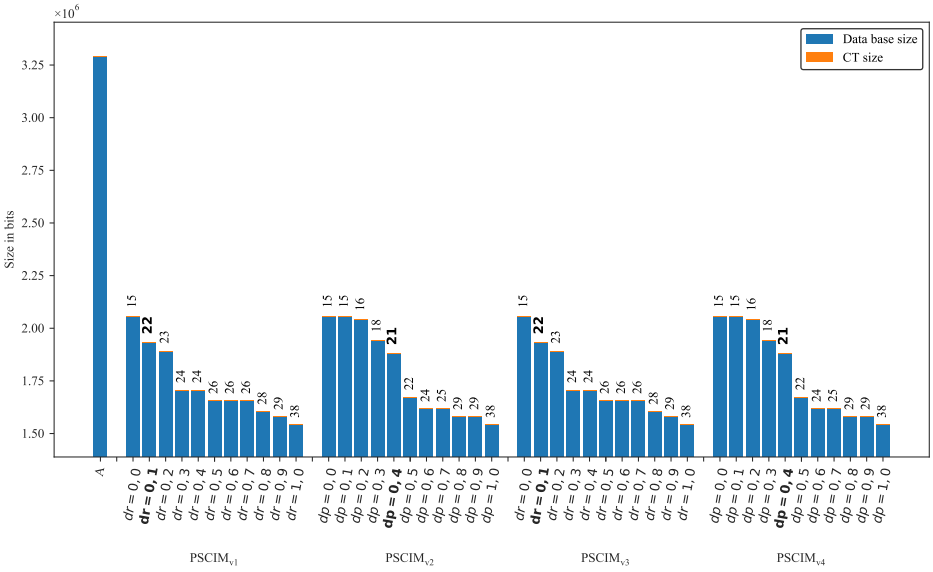


Figura B.37: *Skin*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.1.8 Susy

Tabela B.33: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v1}}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,85]		(0,85 , 0,99]		#	(s)
‡ × 10^{-3}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.702**	**0,009**	**24**	**0,213**	**0**	**0,000**	**8**	**0,573**	**0**	**0,000**	**9**	**0,855**	**18**	**0,946**	**2.761**	**276,49**
0,10 ‡	2.758	0,010	32	0,216	0	0,000	9	0,566	0	0,000	26	0,841	28	0,920	2.853	277,11
0,20 ‡	2.776	0,010	35	0,212	0	0,000	9	0,566	0	0,000	26	0,841	28	0,920	2.874	277,76
0,30 ‡	2.813	0,010	36	0,210	0	0,000	9	0,566	0	0,000	26	0,841	28	0,920	2.912	280,34
0,40 ‡	2.917	0,011	36	0,210	0	0,000	16	0,567	0	0,000	26	0,841	28	0,920	3.023	278,73
0,50 ‡	3.158	0,011	37	0,210	0	0,000	18	0,560	0	0,000	26	0,841	28	0,920	3.267	280,39
0,60 ‡	3.209	0,011	54	0,207	0	0,000	32	0,554	0	0,000	44	0,840	30	0,915	3.369	280,88
0,70 ‡	3.274	0,011	54	0,207	0	0,000	36	0,546	0	0,000	44	0,840	30	0,915	3.438	281,18
0,80 ‡	3.331	0,012	62	0,201	0	0,000	64	0,538	0	0,000	44	0,840	30	0,915	3.531	282,56
1,00 ‡	3.405	0,013	75	0,194	0	0,000	106	0,530	0	0,000	44	0,840	30	0,915	3.660	282,37

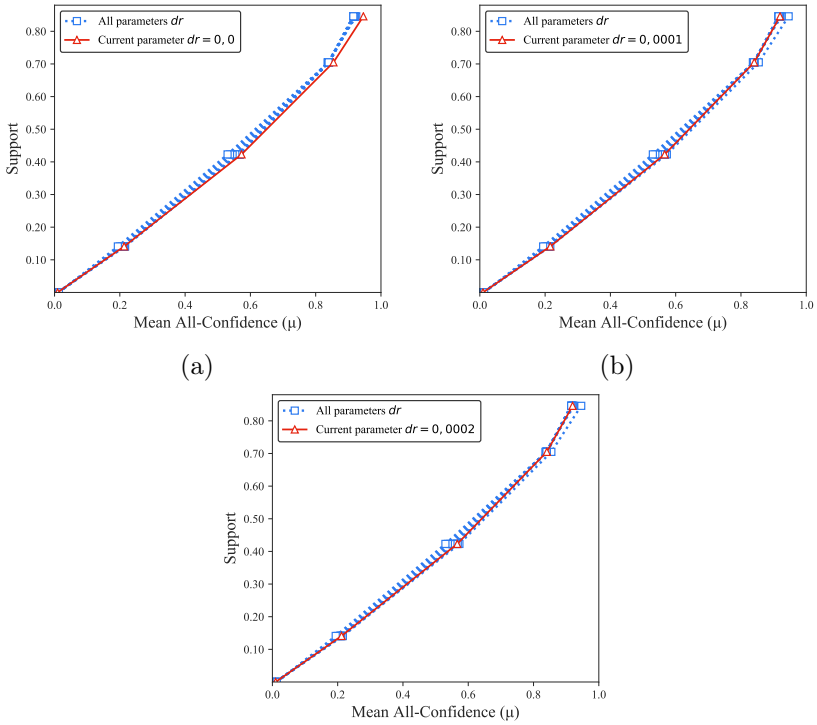


Figura B.38: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v1}}$. (a) com $dr = 0{,}00 \times 10^{-3}$, (b) com $dr = 0{,}10 \times 10^{-3}$, e (c) com $dr = 0{,}20 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}30\,, 0{,}40\,, 0{,}50\,, 0{,}60\,, 0{,}70\,, 0{,}80\,, 1{,}00\}$. Veja Tabela B.33 para detalhes.

Tabela B.34: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v2}}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,85]		(0,85 , 0,99]		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.702**	**0,009**	**24**	**0,213**	**0**	**0,000**	**8**	**0,573**	**0**	**0,000**	**9**	**0,855**	**18**	**0,946**	**2.761**	**272,64**
0,12 ‡	2.704	0,009	24	0,213	0	0,000	8	0,573	0	0,000	9	0,855	25	0,926	2.770	272,55
0,13 ‡	2.719	0,010	28	0,206	0	0,000	8	0,573	0	0,000	15	0,841	26	0,924	2.796	272,74
0,14 ‡	2.791	0,011	30	0,202	0	0,000	9	0,566	0	0,000	15	0,841	27	0,922	2.872	278,30
0,15 ‡	2.874	0,012	39	0,208	0	0,000	16	0,567	0	0,000	18	0,837	27	0,922	2.974	280,10
0,16 ‡	3.199	0,011	48	0,198	0	0,000	16	0,567	0	0,000	18	0,837	27	0,922	3.308	282,27
0,17 ‡	3.274	0,010	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.386	280,93
0,18 ‡	3.346	0,010	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.458	285,77
0,21 ‡	3.729	0,009	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.841	280,94
0,22 ‡	3.928	0,009	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	4.040	281,47
0,23 ‡	3.935	0,009	48	0,198	0	0,000	26	0,553	0	0,000	25	0,842	35	0,919	4.069	282,60
0,24 ‡	3.956	0,009	48	0,198	0	0,000	26	0,553	0	0,000	25	0,842	35	0,919	4.090	282,20

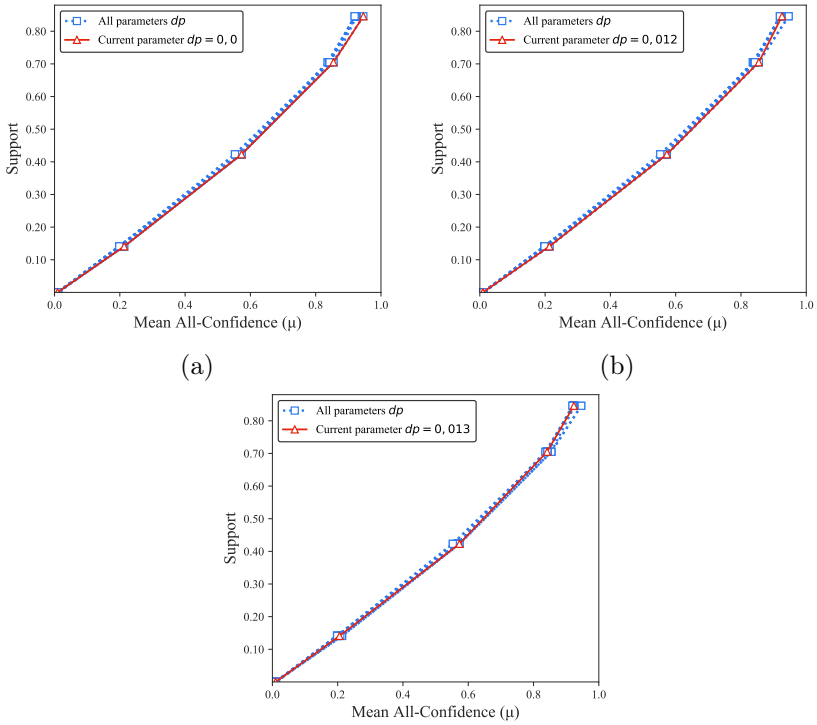
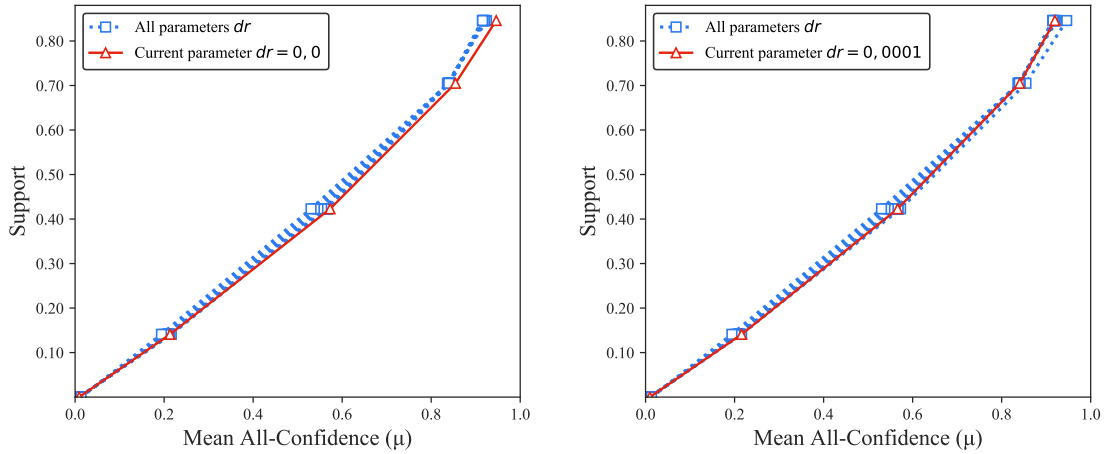


Figura B.39: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}12$ e (b) com $dr = 0{,}13$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}14\,, 0{,}15\,, 0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}21\,, 0{,}22\,, 0{,}23\,, 0{,}24\}$. Veja Tabela B.34 para detalhes.

Tabela B.35: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,85]		(0,85 , 0,99]		#	(s)
‡ × 10^{-3}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.702**	**0,009**	**24**	**0,213**	**0**	**0,000**	**8**	**0,573**	**0**	**0,000**	**9**	**0,855**	**18**	**0,946**	**2.761**	**264,28**
0,10 ‡	2.758	0,010	32	0,216	0	0,000	9	0,566	0	0,000	26	0,841	28	0,920	2.853	264,73
0,20 ‡	2.776	0,010	35	0,212	0	0,000	9	0,566	0	0,000	26	0,841	28	0,920	2.874	264,05
0,30 ‡	2.813	0,010	36	0,210	0	0,000	9	0,566	0	0,000	26	0,841	28	0,920	2.912	256,31
0,40 ‡	2.917	0,011	36	0,210	0	0,000	16	0,567	0	0,000	26	0,841	28	0,920	3.023	258,82
0,50 ‡	3.158	0,011	37	0,210	0	0,000	18	0,560	0	0,000	26	0,841	28	0,920	3.267	259,12
0,60 ‡	3.209	0,011	54	0,207	0	0,000	32	0,554	0	0,000	44	0,840	30	0,915	3.369	258,84
0,70 ‡	3.274	0,011	54	0,207	0	0,000	36	0,546	0	0,000	44	0,840	30	0,915	3.438	259,88
0,80 ‡	3.331	0,012	62	0,201	0	0,000	64	0,538	0	0,000	44	0,840	30	0,915	3.531	260,44
1,00 ‡	3.405	0,013	75	0,194	0	0,000	106	0,530	0	0,000	44	0,840	30	0,915	3.660	260,27

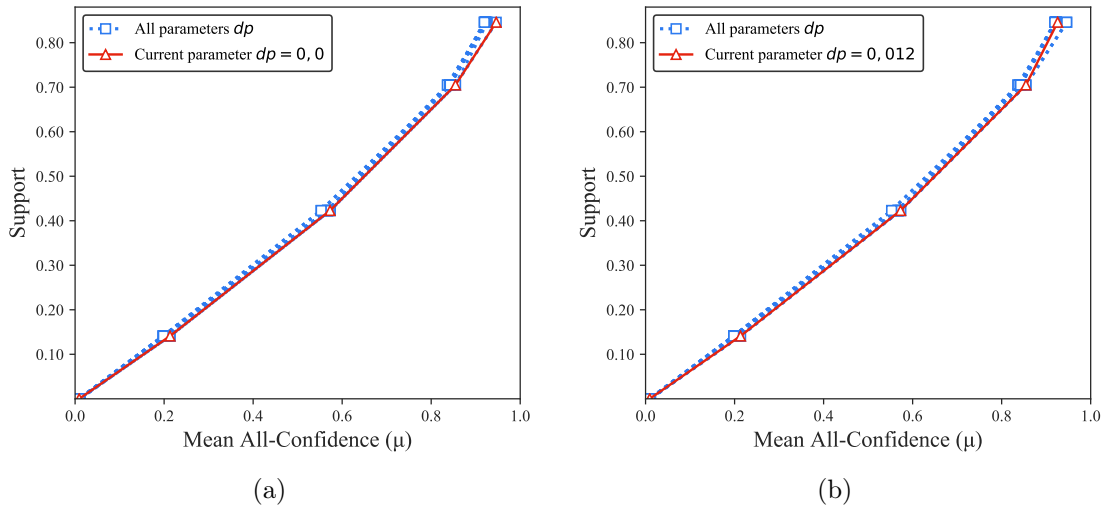


(a) (b)

Figura B.40: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v3}}$. (a) com $dr = 0{,}00 \times 10^{-3}$ e (b) com $dr = 0{,}10 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\,, 0{,}60\,, 0{,}70\,, 0{,}80\,, 0{,}90\,, 1{,}00\}$. Veja Tabela B.35 para detalhes.

Tabela B.36: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v4}}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,85]		(0,85 , 0,99]		#	(s)
$\ddagger \times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.702**	**0,009**	**24**	**0,213**	**0**	**0,000**	**8**	**0,573**	**0**	**0,000**	**9**	**0,855**	**18**	**0,946**	**2.761**	**252,10**
0,02 ‡	2.702	0,009	24	0,213	0	0,000	8	0,573	0	0,000	9	0,855	18	0,946	2.761	252,25
0,12 ‡	2.704	0,009	24	0,213	0	0,000	8	0,573	0	0,000	9	0,855	25	0,926	2.770	251,93
0,13 ‡	2.719	0,010	28	0,206	0	0,000	8	0,573	0	0,000	15	0,841	26	0,924	2.796	251,98
0,14 ‡	2.791	0,011	30	0,202	0	0,000	9	0,566	0	0,000	15	0,841	27	0,922	2.872	253,79
0,15 ‡	2.874	0,012	39	0,208	0	0,000	16	0,567	0	0,000	18	0,837	27	0,922	2.974	255,64
0,16 ‡	3.199	0,011	48	0,198	0	0,000	16	0,567	0	0,000	18	0,837	27	0,922	3.308	255,63
0,17 ‡	3.274	0,010	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.386	256,11
0,18 ‡	3.346	0,010	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.458	255,66
0,21 ‡	3.729	0,009	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.841	256,92
0,22 ‡	3.928	0,009	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	4.040	257,49
0,23 ‡	3.935	0,009	48	0,198	0	0,000	26	0,553	0	0,000	25	0,842	35	0,919	4.069	258,15
0,24 ‡	3.956	0,009	48	0,198	0	0,000	26	0,553	0	0,000	25	0,842	35	0,919	4.090	258,49



(a) (b)

Figura B.41: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v4}$. (a) com $dr = 0{,}00$ e (b) com $dr = 0{,}19$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}12\,, 0{,}13\,, 0{,}14\,, 015\,, 0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}21\,, 0{,}22\,, 0{,}23\,, 0{,}24\}$. Veja Tabela B.36 para detalhes.

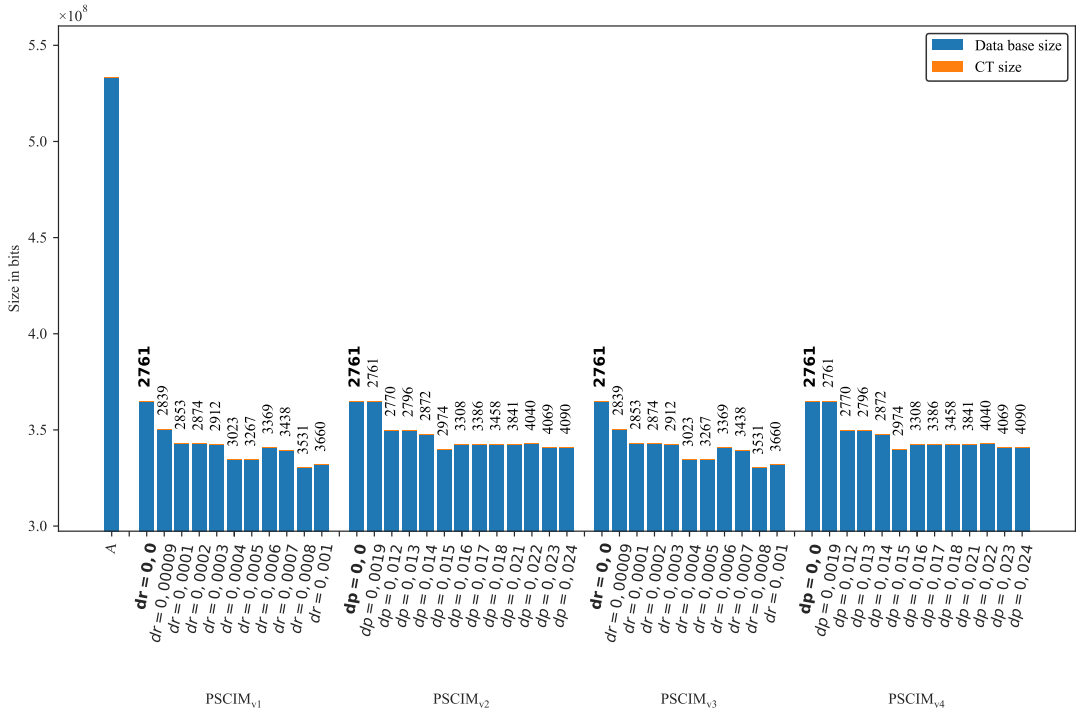


Figura B.42: *Susy*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2 Bases de Dados esparsa

B.1.2.1 Accidents

Tabela B.37: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.222**	**0,026**	**104**	**0,221**	**2**	**0,411**	**151**	**0,520**	**77**	**0,624**	**12**	**0,785**	**8**	**0,956**	**2.576**	**115,74**
0,01 ‡	2.388	0,029	108	0,222	3	0,371	152	0,520	83	0,627	15	0,782	16	0,945	2.765	116,29
0,02 ‡	2.699	0,033	139	0,217	3	0,371	294	0,508	94	0,632	25	0,781	17	0,944	3.271	117,17
0,03 ‡	3.294	0,040	199	0,207	4	0,389	332	0,507	107	0,635	33	0,785	23	0,934	3.992	118,06
0,04 ‡	4.619	0,046	210	0,207	6	0,356	357	0,506	160	0,637	68	0,788	29	0,923	5.449	119,85
0,05 ‡	4.898	0,047	268	0,202	81	0,350	378	0,505	167	0,640	76	0,787	29	0,923	5.897	120,50
0,06 ‡	5.967	0,045	339	0,201	87	0,351	382	0,504	191	0,640	79	0,788	34	0,919	7.079	121,17
0,07 ‡	6.199	0,046	403	0,198	87	0,351	454	0,501	207	0,639	93	0,788	34	0,919	7.477	121,56
0,08 ‡	6.429	0,048	488	0,194	96	0,356	524	0,501	239	0,638	105	0,785	36	0,915	7.917	122,85
0,09 ‡	6.571	0,048	575	0,201	281	0,349	724	0,502	362	0,637	157	0,778	42	0,909	8.712	123,74
0,10 ‡	6.948	0,049	691	0,199	281	0,349	724	0,502	370	0,638	157	0,778	42	0,909	9.213	124,56

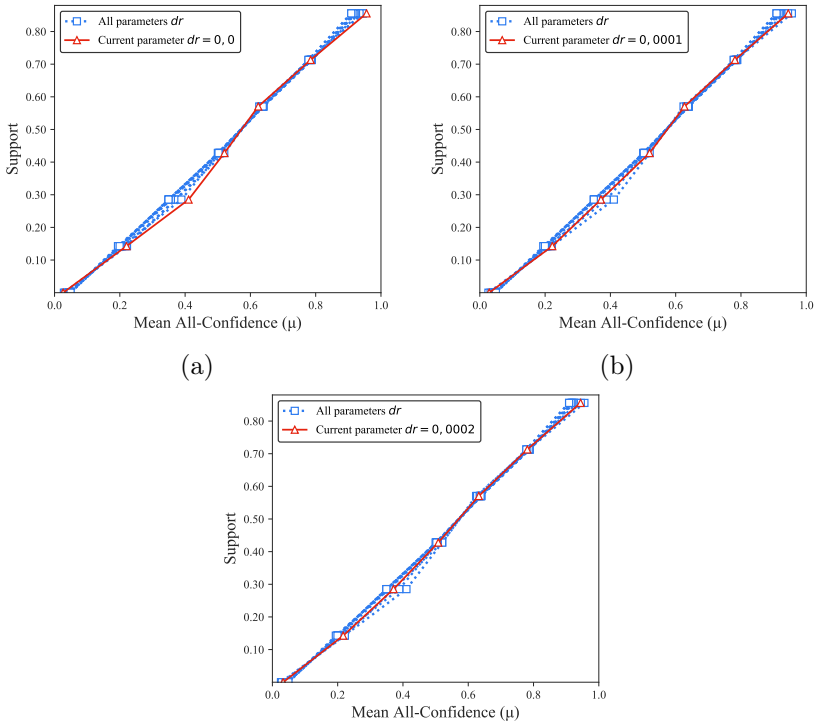


Figura B.43: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}01 \times 10^{-2}$ e (c) com $dr = 0{,}02 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\}$. Veja Tabela B.37 para detalhes.

B.1.2.2 BMSWebView2

Tabela B.38: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]		#	(s)
$\ddagger \times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.222**	**0,026**	**104**	**0,221**	**2**	**0,411**	**151**	**0,520**	**77**	**0,624**	**12**	**0,785**	**8**	**0,956**	**2.576**	**117,45**
0,05 ‡	2.222	0,026	106	0,221	2	0,411	175	0,514	83	0,627	15	0,788	15	0,941	2.618	117,65
0,06 ‡	2.396	0,030	127	0,218	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.027	120,07
0,07 ‡	2.581	0,032	180	0,206	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.265	122,48
0,08 ‡	2.732	0,033	196	0,202	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.432	122,54
0,09 ‡	2.844	0,032	196	0,202	69	0,350	309	0,507	96	0,636	19	0,791	17	0,934	3.550	122,73
0,10 ‡	3.657	0,025	212	0,203	70	0,351	346	0,505	104	0,637	27	0,797	22	0,925	4.438	124,56
0,11 ‡	3.913	0,025	253	0,203	195	0,368	598	0,498	126	0,636	32	0,793	22	0,925	5.139	125,70
0,12 ‡	4.224	0,028	257	0,203	195	0,368	598	0,498	133	0,639	39	0,784	22	0,925	5.468	126,70
0,13 ‡	4.491	0,029	306	0,201	195	0,368	598	0,498	135	0,640	48	0,785	22	0,925	5.795	127,88

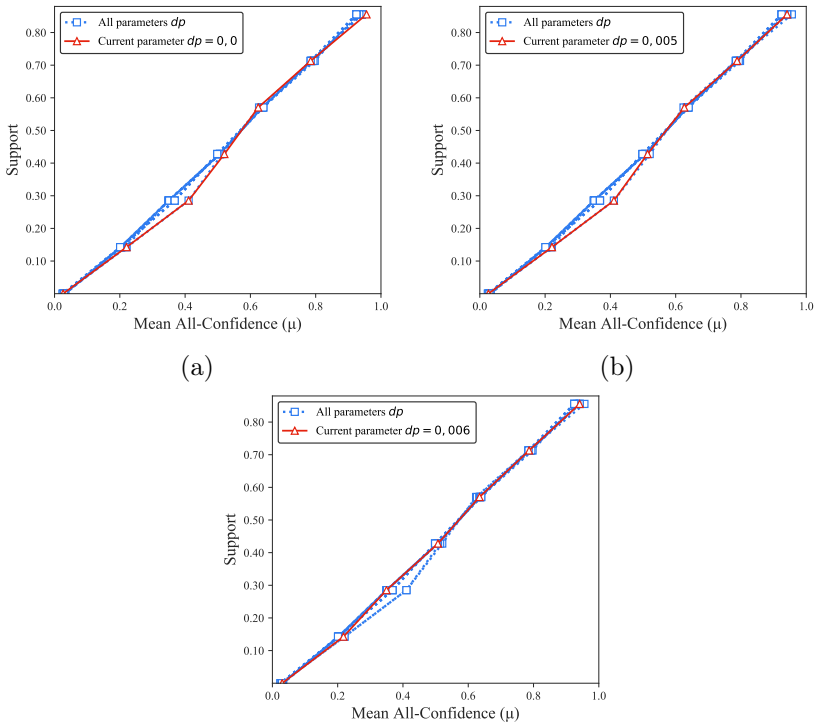


Figura B.44: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v2}. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$ e (c) com $dr = 0{,}06 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\}$. Veja Tabela B.38 para detalhes.

Tabela B.39: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.222**	**0,026**	**104**	**0,221**	**2**	**0,411**	**151**	**0,520**	**77**	**0,624**	**12**	**0,785**	**8**	**0,956**	**2.576**	**112,27**
0,01 ‡	2.388	0,029	108	0,222	3	0,371	152	0,520	83	0,627	15	0,782	16	0,945	2.765	112,77
0,02 ‡	2.699	0,033	139	0,217	3	0,371	294	0,508	94	0,632	25	0,781	17	0,944	3.271	113,56
0,03 ‡	3.294	0,040	199	0,207	4	0,389	332	0,507	107	0,635	33	0,785	23	0,934	3.992	114,62
0,04 ‡	4.619	0,046	210	0,207	6	0,356	357	0,506	160	0,637	68	0,788	29	0,923	5.449	115,38
0,05 ‡	4.898	0,047	268	0,202	81	0,350	378	0,505	167	0,640	76	0,787	29	0,923	5.897	116,21
0,06 ‡	5.967	0,045	339	0,201	87	0,351	382	0,504	191	0,640	79	0,788	34	0,919	7.079	117,03
0,07 ‡	6.199	0,046	403	0,198	87	0,351	454	0,501	207	0,639	93	0,788	34	0,919	7.477	117,40
0,08 ‡	6.429	0,048	488	0,194	96	0,356	524	0,501	239	0,638	105	0,785	36	0,915	7.917	118,87
0,09 ‡	6.571	0,048	575	0,201	281	0,349	724	0,502	362	0,637	157	0,778	42	0,909	8.712	119,46
0,10 ‡	6.948	0,049	691	0,199	281	0,349	724	0,502	370	0,638	157	0,778	42	0,909	9.213	121,12

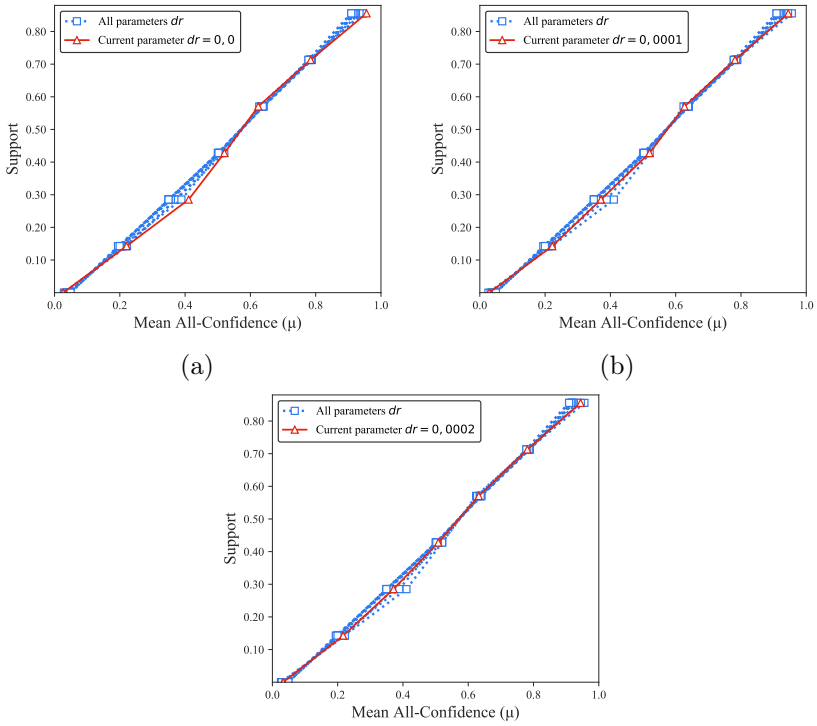
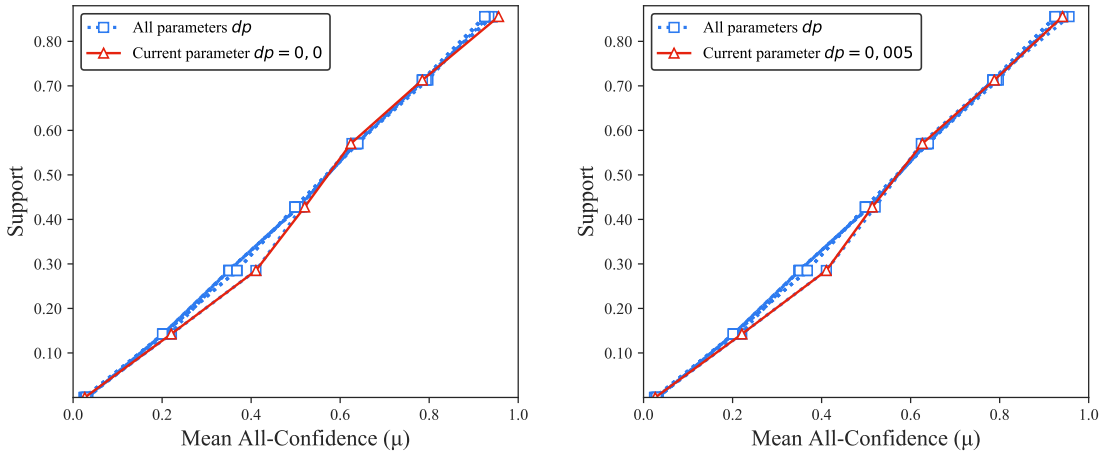


Figura B.45: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}01 \times 10^{-2}$ e (c) com $dr = 0{,}02 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\}$. Veja Tabela B.39 para detalhes.

Tabela B.40: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]		#	(s)
$\ddagger \times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**2.222**	**0,026**	**104**	**0,221**	**2**	**0,411**	**151**	**0,520**	**77**	**0,624**	**12**	**0,785**	**8**	**0,956**	**2.576**	**111,12**
0,05 ‡	2.222	0,026	106	0,221	2	0,411	175	0,514	83	0,627	15	0,788	15	0,941	2.618	111,64
0,06 ‡	2.396	0,030	127	0,218	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.027	113,82
0,07 ‡	2.581	0,032	180	0,206	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.265	115,46
0,08 ‡	2.732	0,033	196	0,202	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.432	116,62
0,09 ‡	2.844	0,032	196	0,202	69	0,350	309	0,507	96	0,636	19	0,791	17	0,934	3.550	116,61
0,10 ‡	3.657	0,025	212	0,203	70	0,351	346	0,505	104	0,637	27	0,797	22	0,925	4.438	118,45
0,11 ‡	3.913	0,025	253	0,203	195	0,368	598	0,498	126	0,636	32	0,793	22	0,925	5.139	120,26
0,12 ‡	4.224	0,028	257	0,203	195	0,368	598	0,498	133	0,639	39	0,784	22	0,925	5.468	123,24
0,13 ‡	4.491	0,029	306	0,201	195	0,368	598	0,498	135	0,640	48	0,785	22	0,925	5.795	121,78



(a) (b)

Figura B.46: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00 \times 10^{-1}$, e (b) com $dr = 0{,}05 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\}$. Veja Tabela B.40 para detalhes.

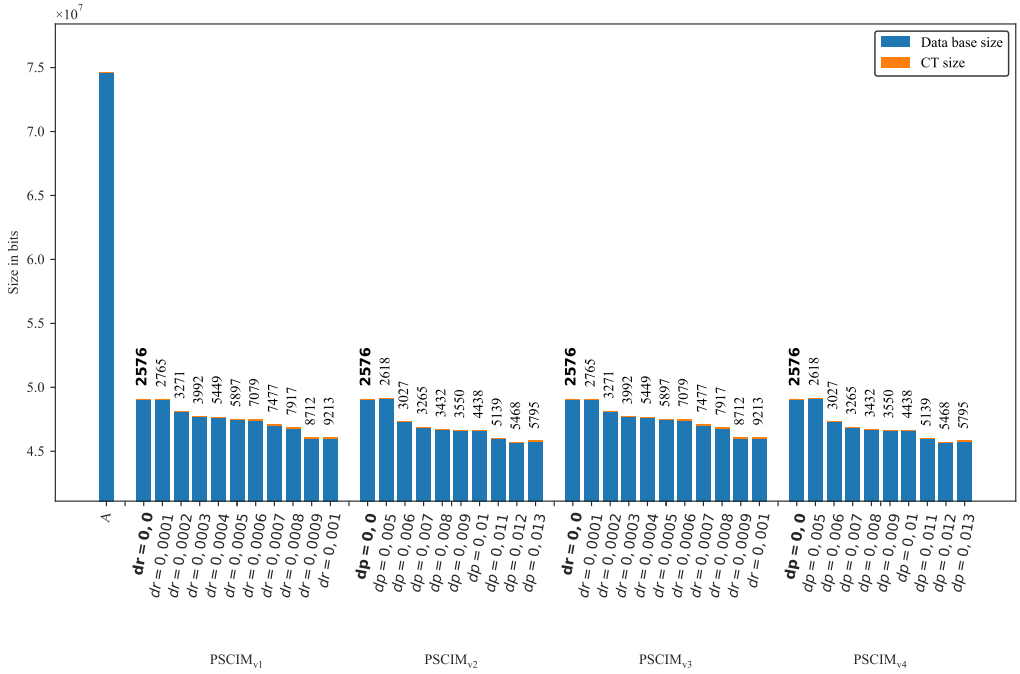


Figura B.47: *Accidents*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

Tabela B.41: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte † × 10^{-2}														Itemset	Tempo
	[0,00 , 0,28] †		(0,28 , 0,55] †		(0,55 , 0,83] †		(0,83 , 1,11] †		(1,11 , 1,38] †		(1,38 , 1,66] †		(1,66 , 1,94] †		#	(s)
‡ × 10^{-2}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	458	0,228	3	0,390	1	0,372	0	0,000	0	0,000	0	0,000	0	0,000	462	226,41
0,01 ‡	469	0,225	3	0,390	1	0,372	0	0,000	0	0,000	0	0,000	0	0,000	473	226,06
0,02 ‡	479	0,223	3	0,390	1	0,372	0	0,000	0	0,000	0	0,000	0	0,000	483	225,94
0,03 ‡	528	0,206	3	0,390	2	0,397	0	0,000	0	0,000	0	0,000	0	0,000	533	226,45
0,04 ‡	1.154	0,104	3	0,390	2	0,397	0	0,000	0	0,000	1	0,309	0	0,000	1.160	227,34
0,05 ‡	**2.524**	**0,055**	**3**	**0,390**	**4**	**0,270**	**2**	**0,189**	**2**	**0,257**	**1**	**0,309**	**0**	**0,000**	**2.536**	**244,28**
0,06 ‡	5.151	0,033	39	0,113	19	0,165	5	0,200	3	0,264	2	0,325	0	0,000	5.219	268,61
0,07 ‡	8.296	0,026	68	0,100	23	0,159	5	0,200	4	0,257	2	0,325	0	0,000	8.398	318,48
0,08 ‡	10.655	0,022	70	0,099	23	0,159	6	0,197	4	0,257	2	0,325	0	0,000	10.760	368,63
0,09 ‡	13.933	0,019	71	0,099	23	0,159	6	0,197	4	0,257	2	0,325	0	0,000	14.039	404,25
0,10 ‡	17.261	0,016	71	0,099	23	0,159	6	0,197	4	0,257	2	0,325	0	0,000	17.367	432,78

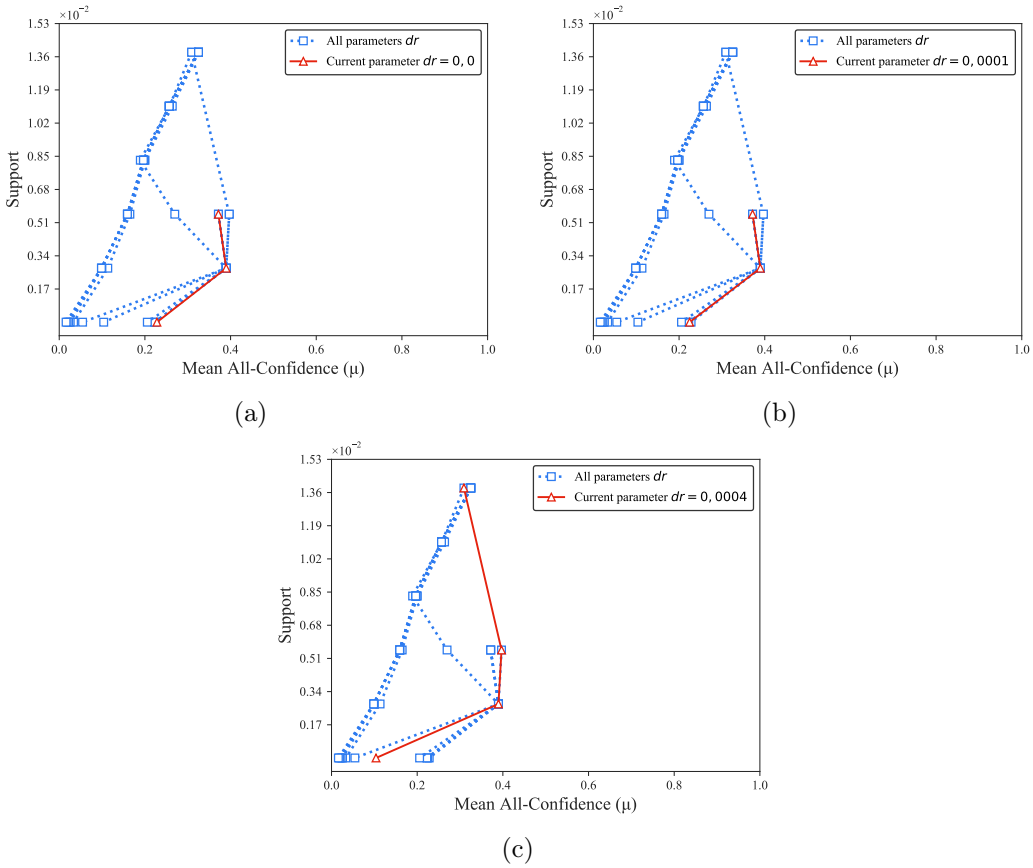


Figura B.48: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v1}}$. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}01 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}02\,,0{,}03\}$ e (c) com $dr = 0{,}04 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\}$. Veja Tabela B.41 para detalhes.

Tabela B.42: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v2}}$.

dr	Partição de suporte † $\times 10^{-2}$														Itemset	Tempo
	[0,00 , 0,28] †		(0,28 , 0,55] †		(0,55 , 0,83] †		(0,83 , 1,11] †		(1,11 , 1,38] †		(1,38 , 1,66] †		(1,66 , 1,94] †		#	(s)
‡ $\times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	458	0,228	3	0,390	1	0,372	0	0,000	0	0,000	0	0,000	0	0,000	462	223,49
0,15 ‡	**482**	**0,222**	**21**	**0,319**	**9**	**0,380**	**4**	**0,366**	**2**	**0,398**	**1**	**0,309**	**2**	**0,492**	**521**	**223,61**
0,16 ‡	502	0,217	24	0,331	11	0,357	4	0,366	2	0,398	1	0,309	2	0,492	546	224,29
0,17 ‡	521	0,213	28	0,343	11	0,357	4	0,366	2	0,398	1	0,309	2	0,492	569	223,87
0,18 ‡	543	0,209	33	0,337	11	0,357	4	0,366	2	0,398	1	0,309	2	0,492	596	224,05
0,19 ‡	563	0,206	34	0,339	11	0,357	4	0,366	2	0,398	1	0,309	2	0,492	617	224,02
0,20 ‡	602	0,201	37	0,336	11	0,357	4	0,366	2	0,398	1	0,309	2	0,492	659	224,03
0,30 ‡	1.132	0,154	60	0,308	19	0,328	9	0,319	6	0,335	1	0,309	3	0,454	1.230	224,01
0,40 ‡	1.868	0,129	87	0,285	30	0,279	13	0,292	8	0,323	1	0,309	3	0,454	2.010	224,11
0,50 ‡	2.748	0,111	132	0,248	40	0,263	14	0,287	8	0,323	1	0,309	3	0,454	2.946	224,49

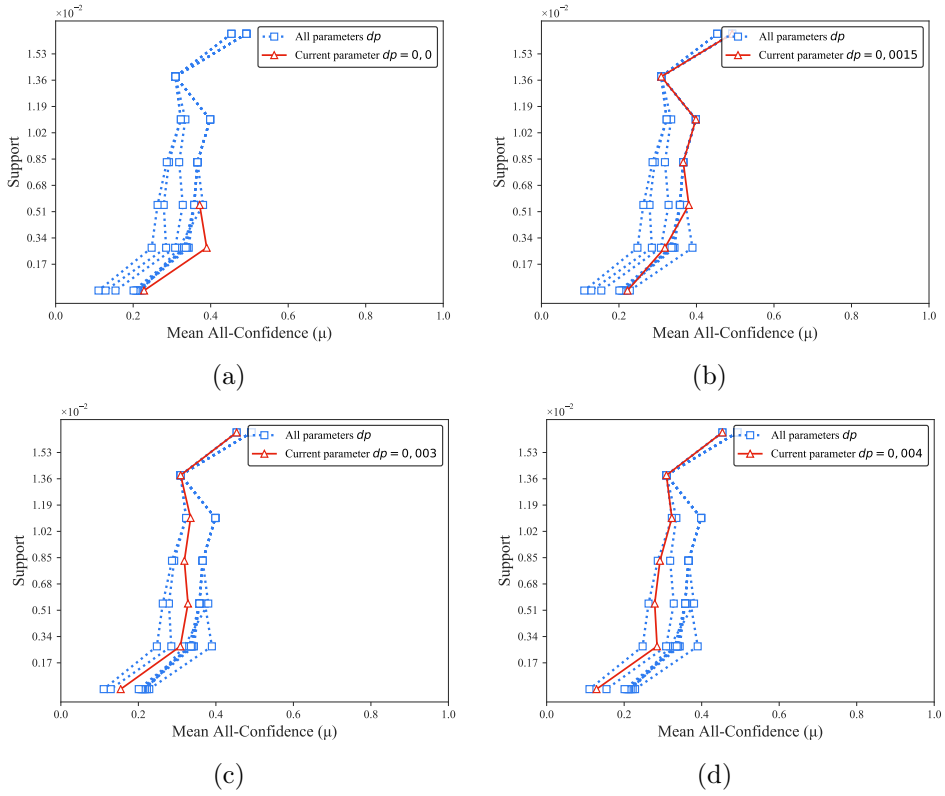


Figura B.49: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}15 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}19\,, 0{,}20\}$ (c) com $dr = 0{,}30 \times 10^{-2}$, e (d) com $dr = 0{,}40 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\}$. Veja Tabela B.42 para detalhes.

Tabela B.43: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte † $\times 10^{-2}$														Itemset	Tempo
	[0,00 , 0,28] †		(0,28 , 0,55] †		(0,55 , 0,83] †		(0,83 , 1,11] †		(1,11 , 1,38] †		(1,38 , 1,66] †		(1,66 , 1,94] †		#	(s)
‡ $\times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**4.119**	**0,166**	**32**	**0,370**	**6**	**0,427**	**2**	**0,344**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**4.159**	**241,21**
0,01 ‡	4.126	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.166	241,25
0,02 ‡	4.143	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.183	241,42
0,03 ‡	4.284	0,162	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.324	241,57
0,04 ‡	4.716	0,149	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.756	241,70
0,05 ‡	5.314	0,134	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	5.354	244,04
0,06 ‡	6.637	0,110	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	6.677	256,29
0,07 ‡	8.349	0,091	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	8.389	268,24
0,08 ‡	9.651	0,081	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	9.691	286,57
0,09 ‡	10.816	0,073	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	10.856	313,41
0,10 ‡	11.496	0,070	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	11.536	342,48

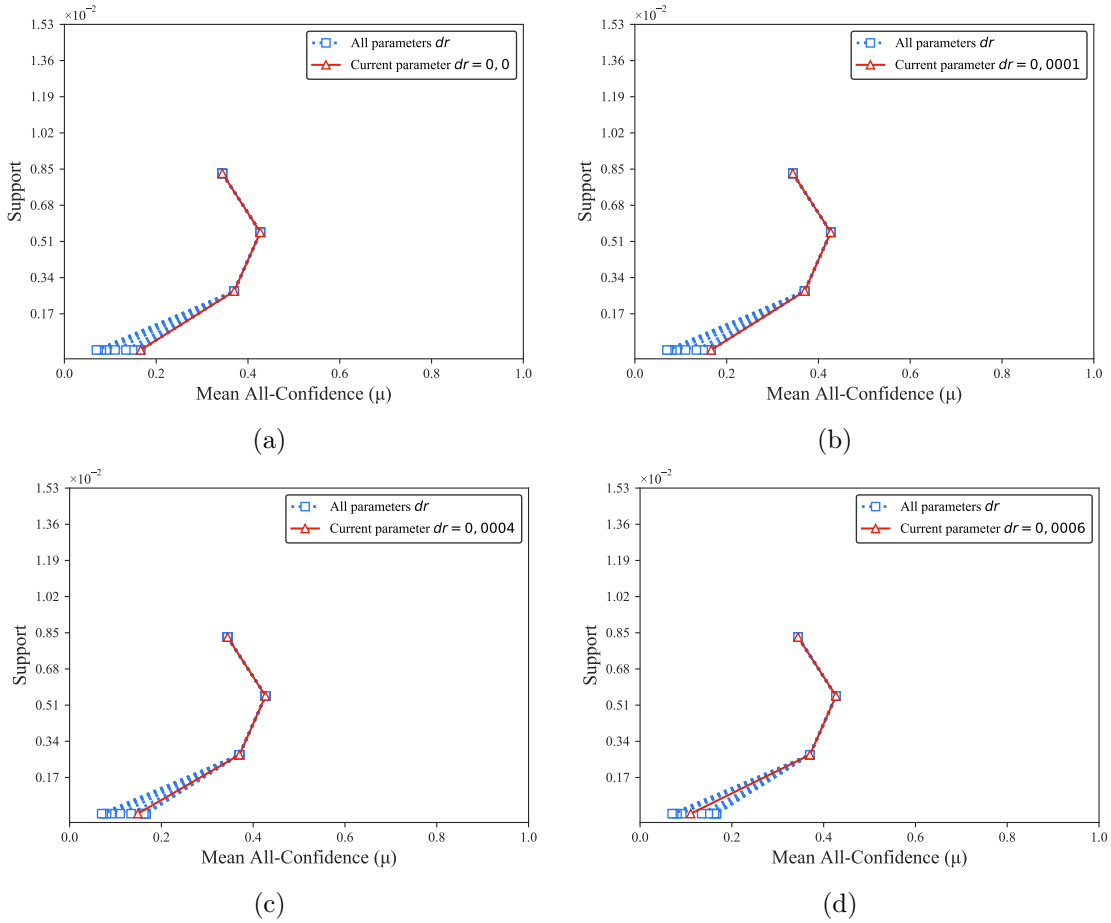


Figura B.50: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v3}$. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}01 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}02\,, 0{,}03\}$, (c) com $dr = 0{,}04 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\}$ e (d) com $dr = 0{,}06 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\}$. Veja Tabela B.43 para detalhes.

Tabela B.44: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v4}$.

dr	Partição de suporte † × 10^{-2}														Itemset	Tempo
	[0,00 , 0,28] †		(0,28 , 0,55] †		(0,55 , 0,83] †		(0,83 , 1,11] †		(1,11 , 1,38] †		(1,38 , 1,66] †		(1,66 , 1,94] †		#	(s)
‡ × 10^{-2}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	4.119	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.159	238,99
0,05 ‡	4.119	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.159	238,99
0,10 ‡	4.119	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.159	240,32
0,15 ‡	**4.163**	**0,165**	**47**	**0,343**	**12**	**0,380**	**5**	**0,356**	**1**	**0,414**	**1**	**0,309**	**2**	**0,492**	**4.231**	**241,63**
0,16 ‡	4.191	0,165	54	0,344	12	0,380	5	0,356	1	0,414	1	0,309	2	0,492	4.266	239,07
0,17 ‡	4.207	0,165	57	0,341	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.286	238,72
0,18 ‡	4.271	0,163	58	0,340	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.351	239,02
0,19 ‡	4.319	0,162	60	0,338	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.401	239,10
0,20 ‡	4.387	0,161	60	0,338	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.469	238,82
0,30 ‡	5.278	0,146	73	0,328	24	0,329	9	0,309	5	0,338	1	0,309	3	0,454	5.393	239,21
0,40 ‡	6.237	0,136	102	0,303	36	0,294	14	0,279	9	0,326	2	0,313	3	0,454	6.403	239,31
0,50 ‡	7.522	0,123	169	0,256	50	0,264	18	0,287	9	0,326	2	0,313	3	0,454	7.773	239,52

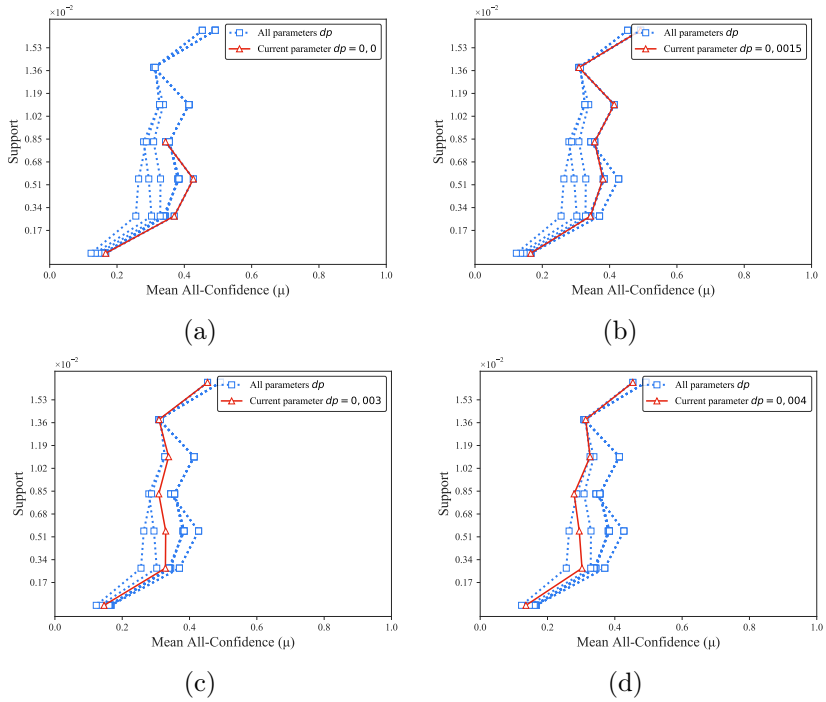


Figura B.51: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v4}$. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}15 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}19\,, 0{,}20\}$ (c) com $dr = 0{,}30 \times 10^{-2}$, e (d) com $dr = 0{,}40 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\}$. Veja Tabela B.44 para detalhes.

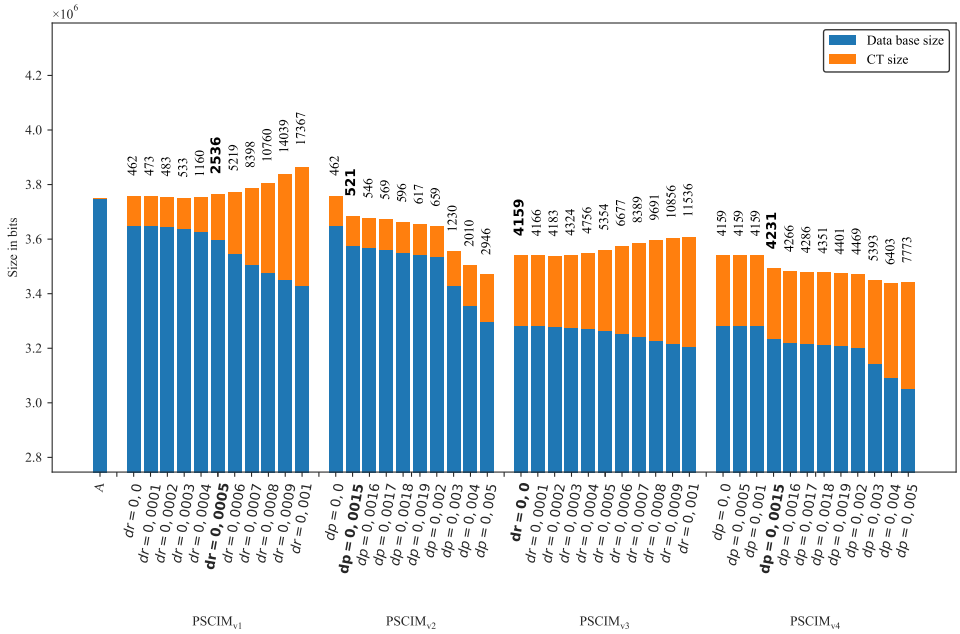


Figura B.52: *BMSWebView2*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2.3 BMS1

Tabela B.45: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,09] †		(0,09 , 0,12] †		(0,12 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,20] †		#	(s)
‡ $\times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	79	0,145	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	79	3,22
0,10 ‡	81	0,143	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	81	3,29
0,20 ‡	**160**	**0,077**	**1**	**0,224**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**161**	**3,33**
0,21 ‡	244	0,053	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	245	3,35
0,22 ‡	362	0,038	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	363	3,37
0,23 ‡	497	0,029	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	498	3,40
0,24 ‡	732	0,021	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	733	3,56
0,25 ‡	1.233	0,014	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1.234	3,87
0,26 ‡	2.030	0,010	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	2.031	4,61
0,27 ‡	2.253	0,010	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	2.254	5,52
0,28 ‡	3.335	0,008	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	3.336	8,02
0,29 ‡	3.940	0,007	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	3.941	11,00
0,30 ‡	4.522	0,007	1	0,224	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4.523	14,23

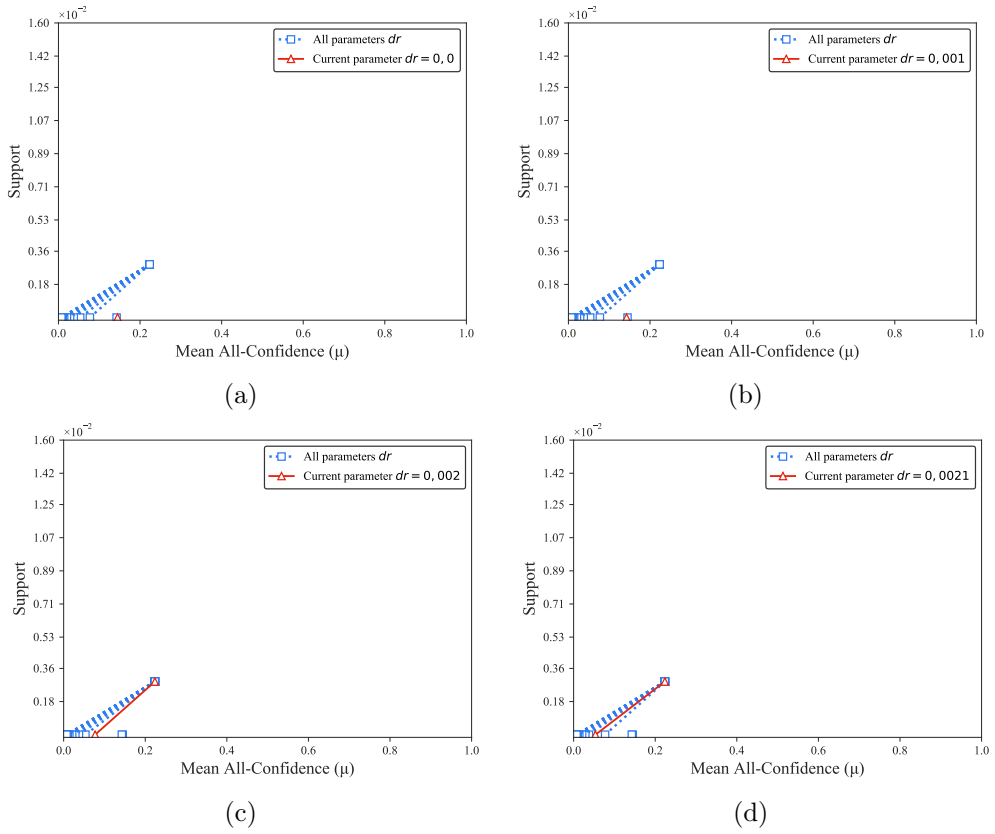


Figura B.53: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}10 \times 10^{-2}$, (c) com $dr = 0{,}20 \times 10^{-2}$, e (d) com $dr = 0{,}21 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}22\,, 0{,}23\,, 0{,}24\,, 0{,}25\,, 0{,}26\,, 0{,}27\,, 0{,}28\,, 0{,}29\,, 0{,}30\}$. Veja Tabela B.45 para detalhes.

Tabela B.46: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v2}$.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,09] †		(0,09 , 0,12] †		(0,12 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,20] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	79	0,145	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	79	3,23
0,05 ‡	79	0,145	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	79	3,28
0,06 ‡	167	0,078	5	0,098	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	173	3,31
0,07 ‡	345	0,045	10	0,097	0	0,000	0	0,000	0	0,000	1	0,387	1	0,329	357	3,32
0,08 ‡	377	0,042	10	0,097	0	0,000	0	0,000	0	0,000	1	0,387	1	0,329	389	3,32
0,09 ‡	396	0,040	10	0,097	0	0,000	0	0,000	0	0,000	1	0,387	1	0,329	408	3,33
0,10 ‡	409	0,039	10	0,097	0	0,000	0	0,000	0	0,000	1	0,387	1	0,329	421	3,35
0,15 ‡	933	0,025	17	0,108	0	0,000	1	0,316	0	0,000	1	0,387	1	0,329	953	3,37
0,20 ‡	**1.670**	**0,019**	**25**	**0,123**	**1**	**0,223**	**2**	**0,279**	**1**	**0,311**	**1**	**0,387**	**1**	**0,329**	**1.701**	**3,42**
0,25 ‡	2.600	0,016	31	0,124	2	0,189	3	0,273	1	0,311	1	0,387	1	0,329	2.639	3,48
0,30 ‡	3.880	0,014	37	0,130	4	0,169	3	0,273	2	0,267	1	0,387	1	0,329	3.928	3,55

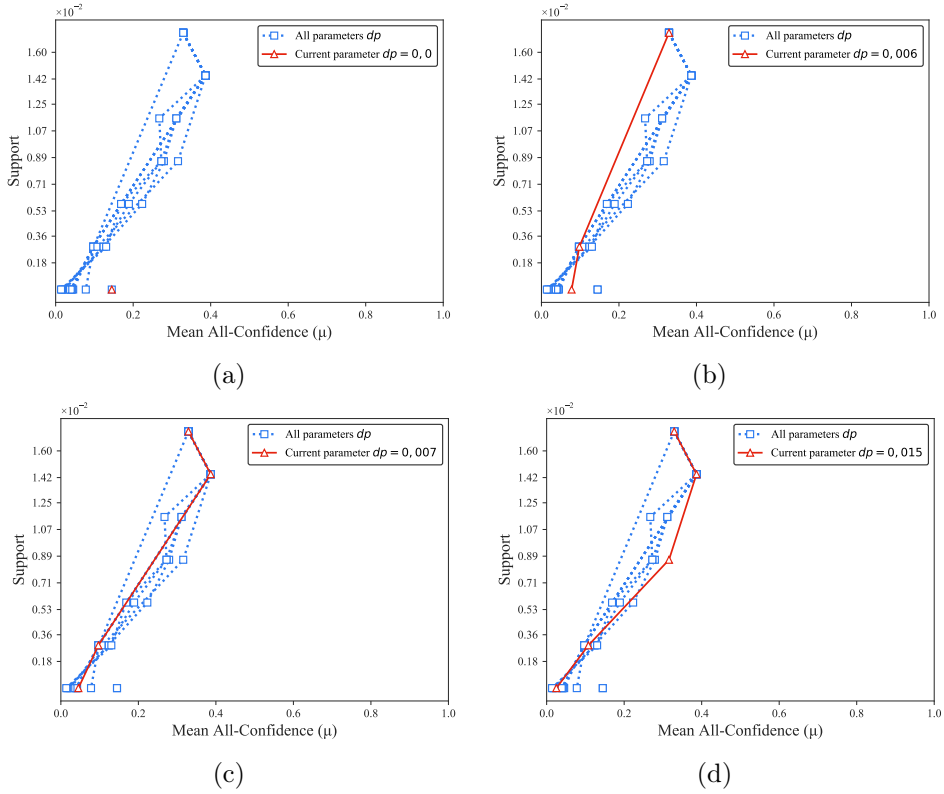


Figura B.54: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}06\}$, (c) com $dr = 0{,}07 \times 10^{-2}$, , onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\}$, e (d) com $dr = 0{,}15 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}25\,, 0{,}30\}$. Veja Tabela B.46 para detalhes.

Tabela B.47: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,09] †		(0,09 , 0,12] †		(0,12 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,20] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,19
0,10 ‡	139	0,190	7	0,353	1	0,304	1	0,316	0	0,000	1	0,387	1	0,329	150	3,27
0,20 ‡	388	0,106	15	0,311	2	0,270	2	0,234	0	0,000	2	0,357	1	0,329	410	3,68
0,21 ‡	474	0,092	15	0,311	2	0,270	2	0,234	0	0,000	2	0,357	1	0,329	496	3,85
0,22 ‡	685	0,069	15	0,311	2	0,270	2	0,234	0	0,000	2	0,357	1	0,329	707	4,81
0,23 ‡	824	0,064	16	0,299	3	0,235	2	0,234	0	0,000	2	0,357	1	0,329	848	5,85
0,24 ‡	1.039	0,053	18	0,292	3	0,235	2	0,234	0	0,000	2	0,357	1	0,329	1.065	6,57
0,25 ‡	1.304	0,048	20	0,288	3	0,235	2	0,234	0	0,000	2	0,357	1	0,329	1.332	7,72
0,26 ‡	1.588	0,045	21	0,287	3	0,235	2	0,234	0	0,000	2	0,357	1	0,329	1.617	10,10
0,27 ‡	**1.954**	**0,040**	**23**	**0,278**	**3**	**0,235**	**2**	**0,234**	**0**	**0,000**	**2**	**0,357**	**1**	**0,329**	**1.985**	**14,18**
0,28 ‡	2.387	0,036	23	0,278	3	0,235	2	0,234	0	0,000	2	0,357	1	0,329	2.418	17,46
0,29 ‡	2.637	0,034	23	0,278	4	0,204	2	0,234	0	0,000	2	0,357	1	0,329	2.669	22,99

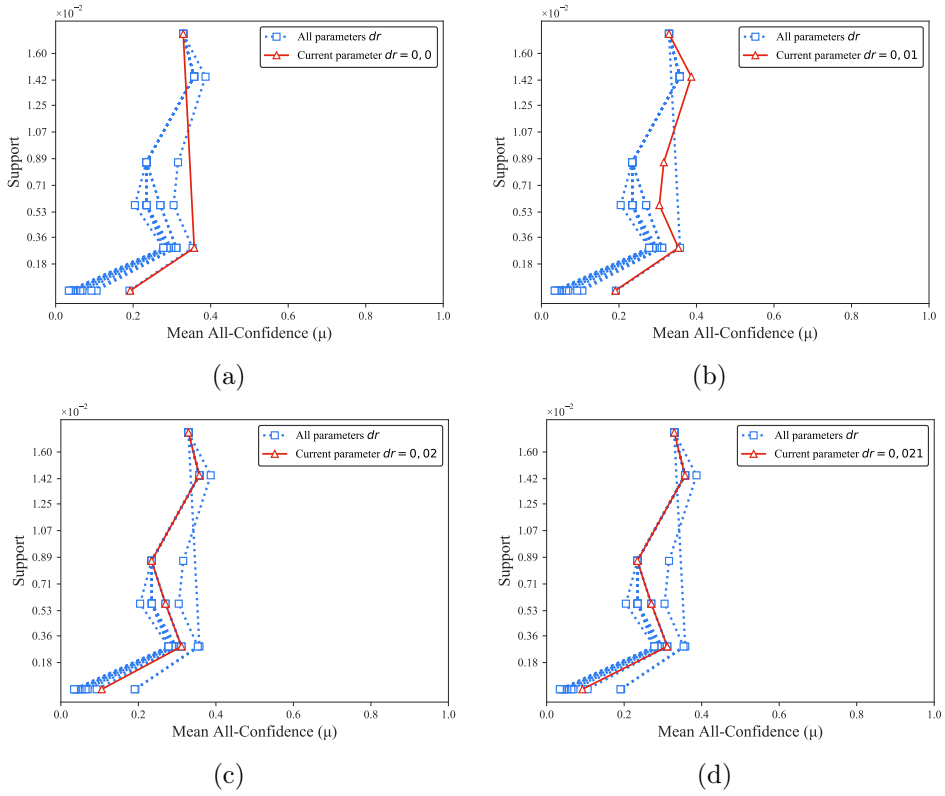


Figura B.55: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}10 \times 10^{-1}$, (c) com $dr = 0{,}20 \times 10^{-1}$ e (d) com $dr = 0{,}21 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}22\,, 0{,}23\,, 0{,}24\,, 0{,}25\,, 0{,}26\,, 0{,}27\,, 0{,}28\,, 0{,}29\}$. Veja Tabela B.47 para detalhes.

Tabela B.48: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,09] †		(0,09 , 0,12] †		(0,12 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,20] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,22
0,05	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,28
0,06	144	0,164	9	0,341	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	154	3,31
0,07	310	0,147	22	0,270	2	0,200	3	0,210	0	0,000	1	0,387	1	0,329	339	3,35
0,08 ‡	**363**	**0,140**	**24**	**0,264**	**3**	**0,235**	**3**	**0,210**	**2**	**0,260**	**2**	**0,357**	**1**	**0,329**	**398**	**3,37**
0,09	372	0,139	24	0,264	3	0,235	3	0,210	2	0,260	2	0,357	1	0,329	407	3,37
0,10	379	0,138	24	0,264	3	0,235	3	0,210	2	0,260	2	0,357	1	0,329	414	3,37
0,15	886	0,102	31	0,246	3	0,235	6	0,241	2	0,260	2	0,357	1	0,329	931	3,50
0,20	1.390	0,086	36	0,241	6	0,178	6	0,241	3	0,248	2	0,357	1	0,329	1.444	3,64
0,25	1.972	0,074	45	0,224	6	0,178	7	0,244	3	0,248	2	0,357	1	0,329	2.036	3,83
0,30	2.668	0,066	49	0,214	9	0,173	8	0,241	3	0,248	2	0,357	1	0,329	2.740	4,05

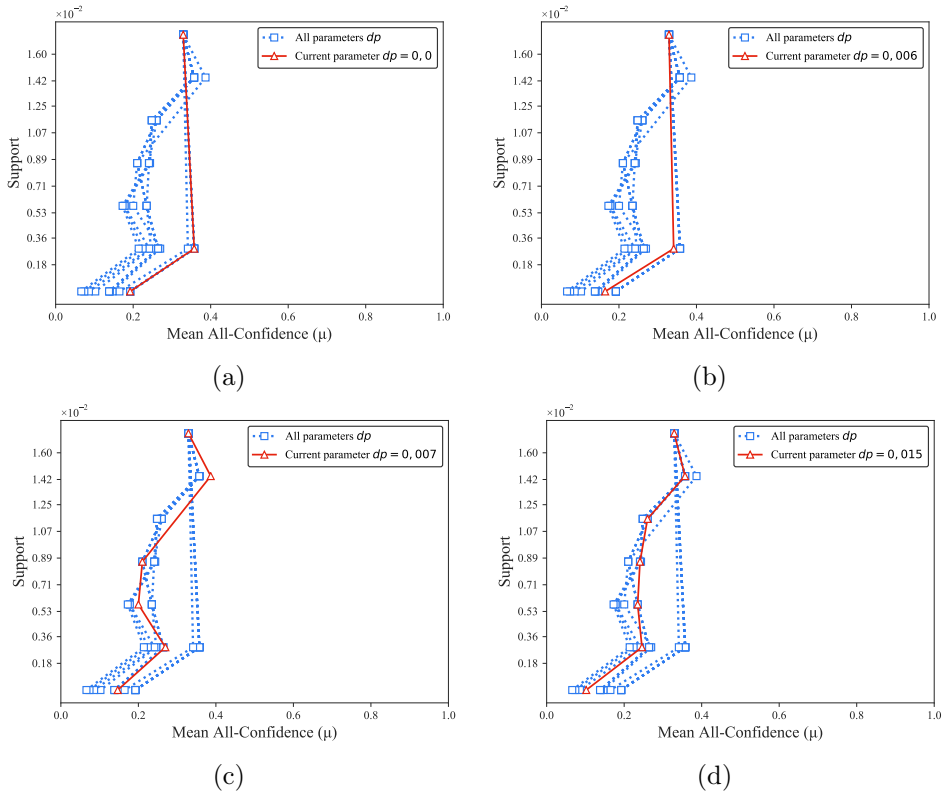


Figura B.56: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\}$, (b) com $dr = 0{,}06 \times 10^{-1}$, (c) com $dr = 0{,}07 \times 10^{-1}$,onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\}$, e (d) com $dr = 0{,}15 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}25\,, 0{,}30\}$. Veja Tabela B.48 para detalhes.

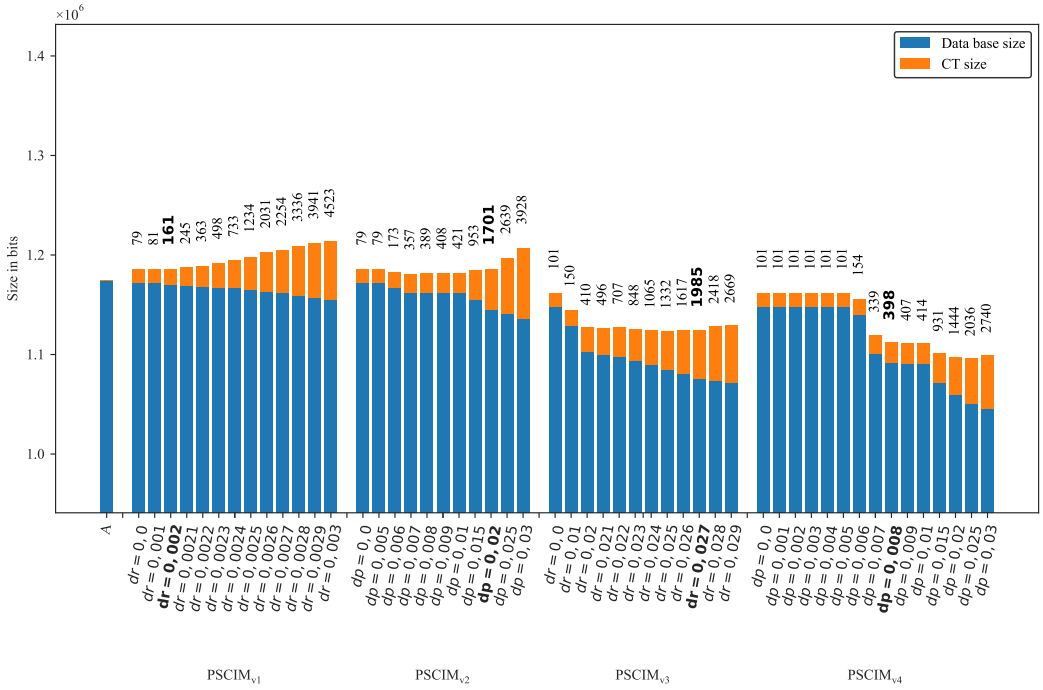


Figura B.57: *BMS1*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2.4 FoodmartFIM

Tabela B.49: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte † × 10^{-3}														Itemset #	Tempo (s)
	[0,00 , 0,14] †		(0,14 , 0,28] †		(0,28 , 0,41] †		(0,41 , 0,55] †		(0,55 , 0,69] †		(0,69 , 0,83] †		(0,83 , 0,97] †			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	0	0,000	3	0,083	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	7	12,30
0,10	0	0,000	76	0,073	0	0,000	30	0,162	0	0,000	10	0,231	0	0,000	116	12,29
0,20	**0**	**0,000**	**532**	**0,083**	**0**	**0,000**	**248**	**0,149**	**0**	**0,000**	**29**	**0,209**	**2**	**0,211**	**811**	**12,33**
0,21	0	0,000	683	0,082	0	0,000	295	0,147	0	0,000	31	0,206	2	0,211	1.011	12,37
0,22	0	0,000	905	0,081	0	0,000	352	0,147	0	0,000	33	0,205	2	0,211	1.292	12,41
0,23	0	0,000	1.175	0,080	0	0,000	398	0,145	0	0,000	37	0,202	2	0,211	1.612	12,42
0,24	0	0,000	1.477	0,079	0	0,000	457	0,144	0	0,000	41	0,200	2	0,211	1.977	12,40
0,25	0	0,000	1.842	0,078	0	0,000	520	0,143	0	0,000	45	0,196	2	0,211	2.409	12,43
0,26	0	0,000	2.209	0,077	0	0,000	577	0,141	0	0,000	48	0,195	2	0,211	2.836	12,42
0,27	0	0,000	2.613	0,076	0	0,000	635	0,141	0	0,000	52	0,195	2	0,211	3.302	12,44
0,28	0	0,000	3.028	0,075	0	0,000	700	0,140	0	0,000	53	0,194	2	0,211	3.783	12,48
0,29	0	0,000	3.446	0,074	0	0,000	769	0,138	0	0,000	57	0,193	2	0,211	4.274	12,50
0,30	0	0,000	3.849	0,073	0	0,000	813	0,138	0	0,000	58	0,192	2	0,211	4.722	12,51
0,40	0	0,000	6.009	0,068	0	0,000	1.194	0,134	0	0,000	67	0,190	2	0,211	7.272	12,92
0,50	0	0,000	5.224	0,065	0	0,000	1.306	0,133	0	0,000	70	0,189	2	0,211	6.602	13,50

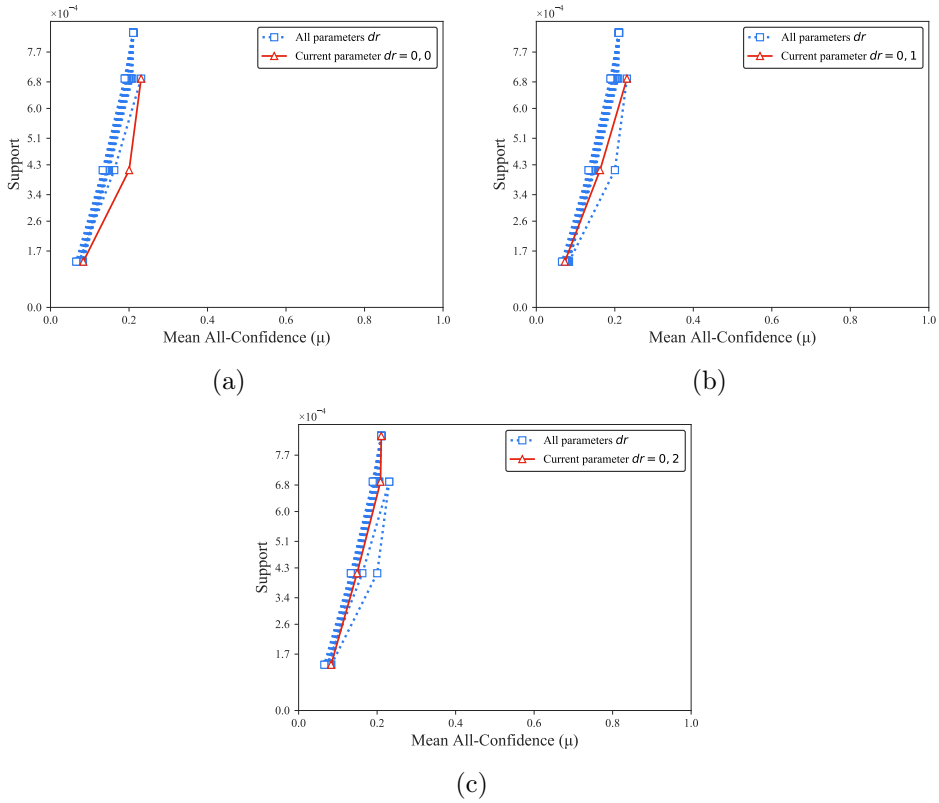


Figura B.58: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v1}}$. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}10$, e (c) com $dr = 0{,}20$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}21\,, 0{,}22\,, 0{,}23\,, 0{,}24\,, 0{,}25\,, 0{,}26\,, 0{,}27\,, 0{,}28\,, 0{,}29\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela B.49 para detalhes.

Tabela B.50: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v2}}$.

dr	Partição de suporte † $\times 10^{-3}$														Itemset	Tempo
	[0,00 , 0,14] †		(0,14 , 0,28] †		(0,28 , 0,41] †		(0,41 , 0,55] †		(0,55 , 0,69] †		(0,69 , 0,83] †		(0,83 , 0,97] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	0	0,000	3	0,083	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	7	12,13
0,10 ‡	0	0,000	4	0,074	0	0,000	4	0,170	0	0,000	1	0,231	0	0,000	9	12,16
0,11 ‡	0	0,000	6	0,064	0	0,000	6	0,145	0	0,000	1	0,231	0	0,000	13	12,17
0,12 ‡	0	0,000	15	0,058	0	0,000	13	0,122	0	0,000	3	0,180	0	0,000	31	12,18
0,13 ‡	0	0,000	48	0,058	0	0,000	28	0,117	0	0,000	4	0,188	1	0,211	81	12,17
0,14 ‡	0	0,000	84	0,059	0	0,000	45	0,119	0	0,000	9	0,183	1	0,211	139	12,20
0,15 ‡	0	0,000	130	0,061	0	0,000	78	0,123	0	0,000	13	0,182	2	0,211	223	12,14
0,16 ‡	0	0,000	183	0,063	0	0,000	111	0,126	0	0,000	20	0,187	2	0,211	316	12,18
0,17 ‡	0	0,000	256	0,066	0	0,000	149	0,129	0	0,000	21	0,188	2	0,211	428	12,24
0,18 ‡	0	0,000	322	0,068	0	0,000	174	0,131	0	0,000	24	0,192	2	0,211	522	12,23
0,19 ‡	0	0,000	383	0,070	0	0,000	196	0,133	0	0,000	27	0,192	2	0,211	608	12,20
0,20 ‡	**0**	**0,000**	**465**	**0,072**	**0**	**0,000**	**230**	**0,136**	**0**	**0,000**	**28**	**0,195**	**2**	**0,211**	**725**	**12,17**
0,30 ‡	0	0,000	1.014	0,078	0	0,000	410	0,143	0	0,000	41	0,196	2	0,211	1.467	12,21
0,40 ‡	0	0,000	1.539	0,078	0	0,000	522	0,142	0	0,000	48	0,195	2	0,211	2.111	12,24
0,50 ‡	0	0,000	2.099	0,078	0	0,000	612	0,141	0	0,000	52	0,194	2	0,211	2.765	12,27

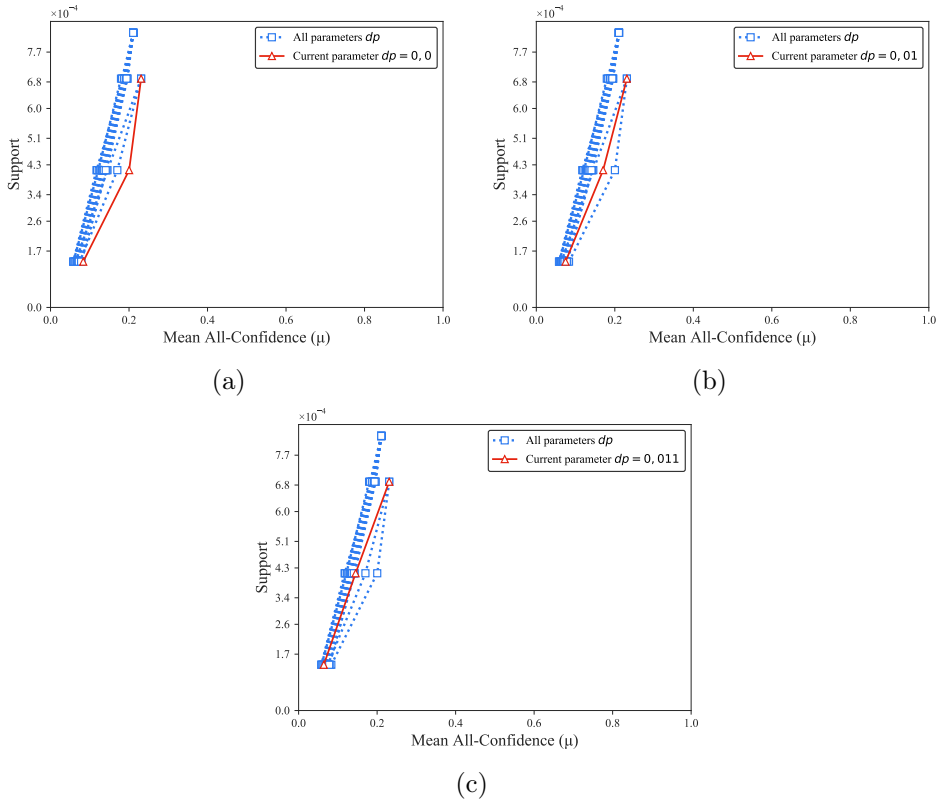


Figura B.59: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}10 \times 10^{-1}$ e (c) com $dr = 0{,}11 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}12\,, 0{,}13\,, 0{,}14\,, 0{,}15\,, 0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}19\,, 0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela B.50 para detalhes.

Tabela B.51: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte † $\times 10^{-3}$														Itemset #	Tempo (s)
	[0,00 , 0,14] †		(0,14 , 0,28] †		(0,28 , 0,41] †		(0,41 , 0,55] †		(0,55 , 0,69] †		(0,69 , 0,83] †		(0,83 , 0,97] †			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	0	0,000	1	0,111	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	5	12,61
0,10	0	0,000	10	0,117	0	0,000	11	0,189	0	0,000	6	0,255	0	0,000	27	12,61
0,20	**0**	**0,000**	**63**	**0,104**	**0**	**0,000**	**51**	**0,175**	**0**	**0,000**	**17**	**0,225**	**1**	**0,211**	**132**	**12,63**
0,21	0	0,000	73	0,102	0	0,000	59	0,170	0	0,000	17	0,225	1	0,211	150	12,62
0,22	0	0,000	87	0,101	0	0,000	85	0,165	0	0,000	17	0,225	1	0,211	190	12,63
0,23	0	0,000	109	0,098	0	0,000	103	0,163	0	0,000	19	0,222	1	0,211	232	12,62
0,24	0	0,000	140	0,095	0	0,000	130	0,161	0	0,000	21	0,222	1	0,211	292	12,60
0,25	0	0,000	174	0,092	0	0,000	159	0,158	0	0,000	24	0,220	1	0,211	358	12,63
0,26	0	0,000	209	0,091	0	0,000	190	0,156	0	0,000	27	0,215	1	0,211	427	12,62
0,27	0	0,000	276	0,091	0	0,000	214	0,155	0	0,000	28	0,214	2	0,211	520	12,63
0,28	0	0,000	341	0,090	0	0,000	239	0,154	0	0,000	30	0,212	2	0,211	612	12,66
0,29	0	0,000	426	0,089	0	0,000	274	0,152	0	0,000	33	0,210	2	0,211	735	12,67
0,30	0	0,000	509	0,089	0	0,000	313	0,151	0	0,000	37	0,205	2	0,211	861	12,67
0,40	0	0,000	3.354	0,077	0	0,000	795	0,139	0	0,000	59	0,191	2	0,211	4.210	12,74
0,50	0	0,000	6.926	0,070	0	0,000	1.181	0,134	0	0,000	67	0,189	2	0,211	8.176	13,03

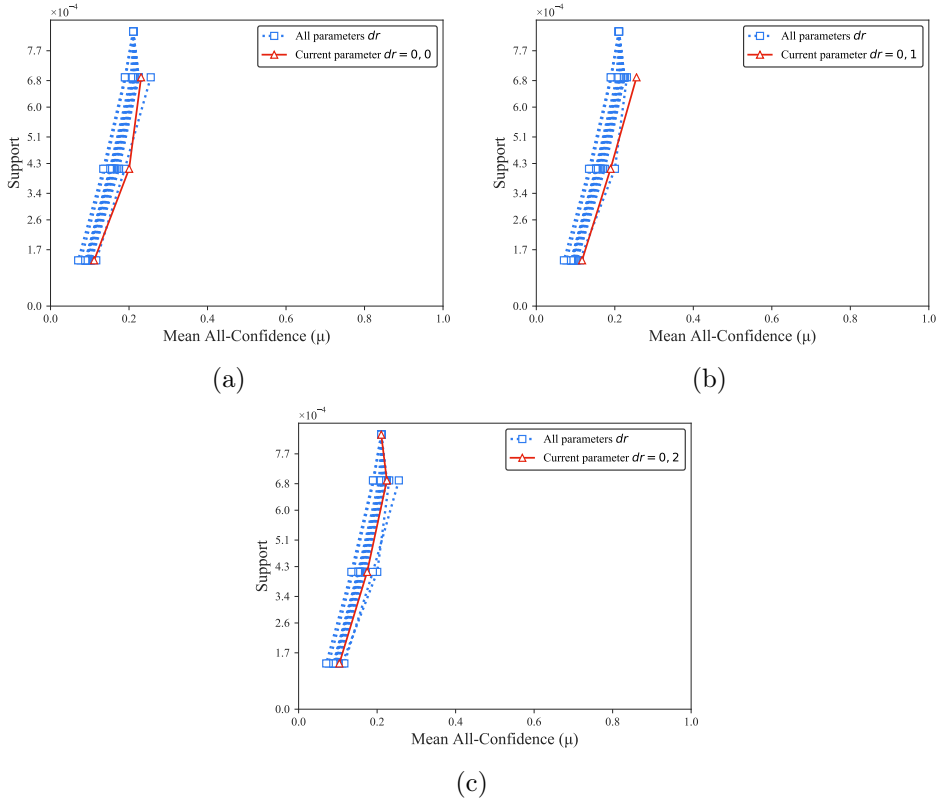
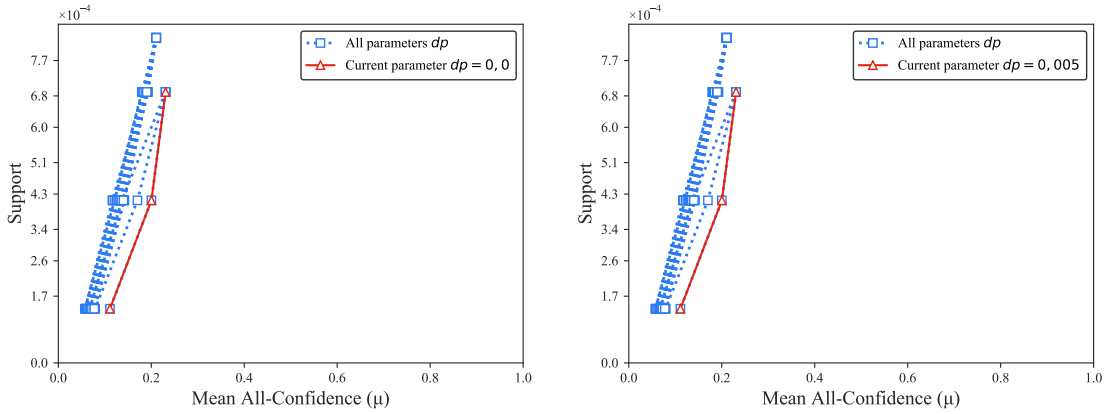


Figura B.60: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v3}}$. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}10$, e (c) com $dr = 0{,}20$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}21\,, 0{,}22\,, 0{,}23\,, 0{,}24\,, 0{,}25\,, 0{,}26\,, 0{,}27\,, 0{,}28\,, 0{,}29\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela B.51 para detalhes.

Tabela B.52: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v4}}$.

dr	Partição de suporte † × 10^{-3}														Itemset	Tempo
	[0,00 , 0,14] †		(0,14 , 0,28] †		(0,28 , 0,41] †		(0,41 , 0,55] †		(0,55 , 0,69] †		(0,69 , 0,83] †		(0,83 , 0,97] †		#	(s)
‡ × 10^{-1}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	0	0,000	1	0,111	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	5	12,54
0,05 ‡	0	0,000	1	0,111	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	5	12,54
0,10 ‡	0	0,000	2	0,079	0	0,000	4	0,170	0	0,000	1	0,231	0	0,000	7	12,60
0,11 ‡	0	0,000	3	0,067	0	0,000	7	0,137	0	0,000	1	0,231	0	0,000	11	12,58
0,12 ‡	0	0,000	14	0,057	0	0,000	12	0,121	0	0,000	3	0,180	0	0,000	29	12,56
0,13 ‡	0	0,000	43	0,058	0	0,000	30	0,116	0	0,000	5	0,186	1	0,211	79	12,55
0,14 ‡	0	0,000	74	0,059	0	0,000	51	0,117	0	0,000	11	0,181	1	0,211	137	12,56
0,15 ‡	0	0,000	116	0,061	0	0,000	88	0,122	0	0,000	14	0,180	2	0,211	220	12,55
0,16 ‡	0	0,000	166	0,063	0	0,000	126	0,126	0	0,000	21	0,182	2	0,211	315	12,56
0,17 ‡	0	0,000	234	0,066	0	0,000	169	0,129	0	0,000	23	0,184	2	0,211	428	12,56
0,18 ‡	0	0,000	297	0,068	0	0,000	196	0,131	0	0,000	26	0,189	2	0,211	521	12,57
0,19 ‡	0	0,000	357	0,070	0	0,000	218	0,133	0	0,000	29	0,189	2	0,211	606	12,61
0,20 ‡	0	0,000	437	0,073	0	0,000	253	0,136	0	0,000	31	0,193	2	0,211	723	12,60
0,30 ‡	**0**	**0,000**	**951**	**0,078**	**0**	**0,000**	**452**	**0,142**	**0**	**0,000**	**50**	**0,190**	**2**	**0,211**	**1.455**	**12,62**
0,40 ‡	0	0,000	1.469	0,078	0	0,000	559	0,142	0	0,000	57	0,191	2	0,211	2.087	12,65
0,50 ‡	0	0,000	2.026	0,077	0	0,000	642	0,140	0	0,000	60	0,190	2	0,211	2.730	12,64



(a) (b)

Figura B.61: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v4}$. (a) com $dr = 0{,}00 \times 10^{-1}$ e (b) com $dr = 0{,}05$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\,, 0{,}14\,, 0{,}15\,, 0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}19\,, 0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela B.52 para detalhes.

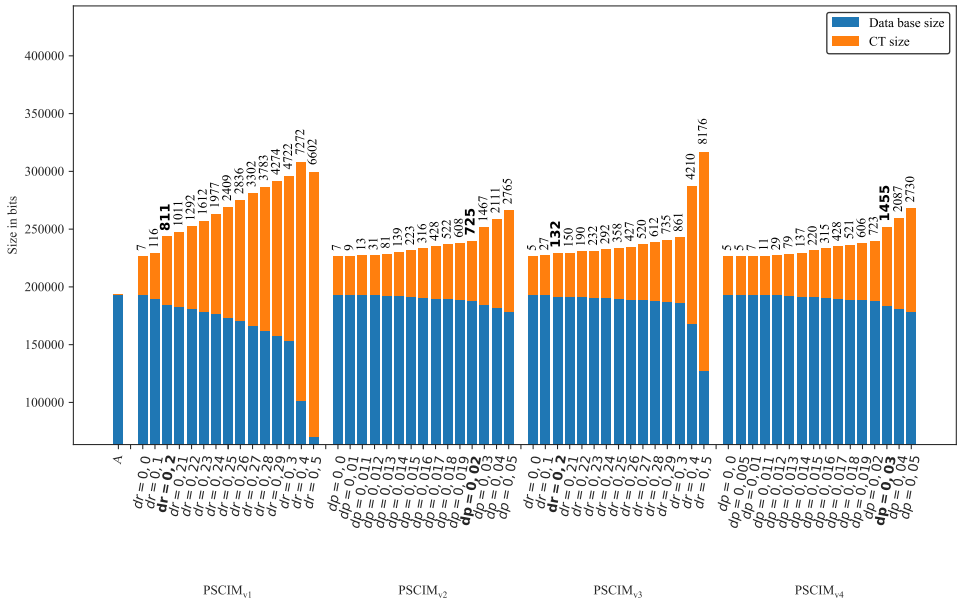


Figura B.62: *FoodmartFIM*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2.5 Fruithut

Tabela B.53: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,05] †		(0,05 , 0,10] †		(0,10 , 0,15] †		(0,15 , 0,20] †		(0,20 , 0,25] †		(0,25 , 0,30] †		(0,30 , 0,35] †		#	(s)
‡ $\times 10^{-3}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	423	0,021	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	423	20,96
0,05 ‡	423	0,021	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	423	20,96
0,06 ‡	424	0,021	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	424	20,95
0,07 ‡	426	0,021	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	426	20,95
0,08 ‡	501	0,018	0	0,000	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	502	20,94
0,09 ‡	876	0,011	0	0,000	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	877	21,00
0,10 ‡	2.923	0,004	2	0,065	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	2.926	21,21
0,11 ‡	**6.881**	**0,002**	**3**	**0,067**	**0**	**0,000**	**1**	**0,078**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**6.885**	**21,69**
0,12 ‡	11.848	0,001	3	0,067	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	11.852	22,35
0,13 ‡	18.447	0,001	3	0,067	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	18.451	23,46
0,14 ‡	21.884	0,001	3	0,067	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	21.888	24,00
0,15 ‡	23.894	0,001	3	0,067	0	0,000	1	0,078	0	0,000	0	0,000	0	0,000	23.898	24,34

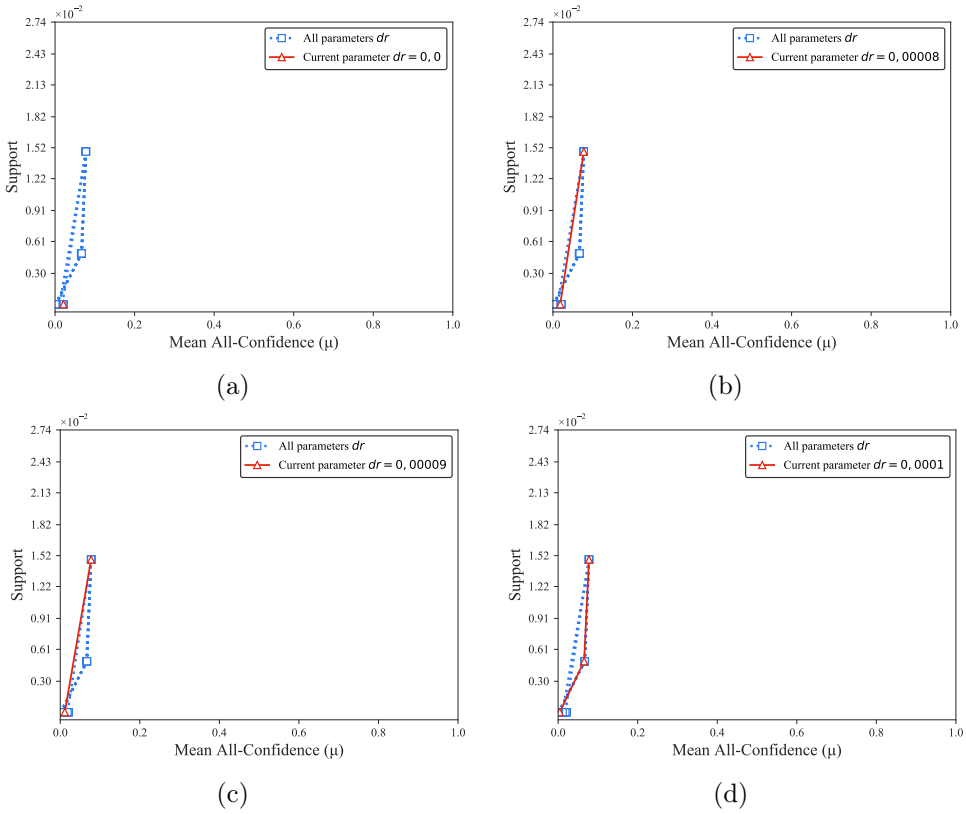
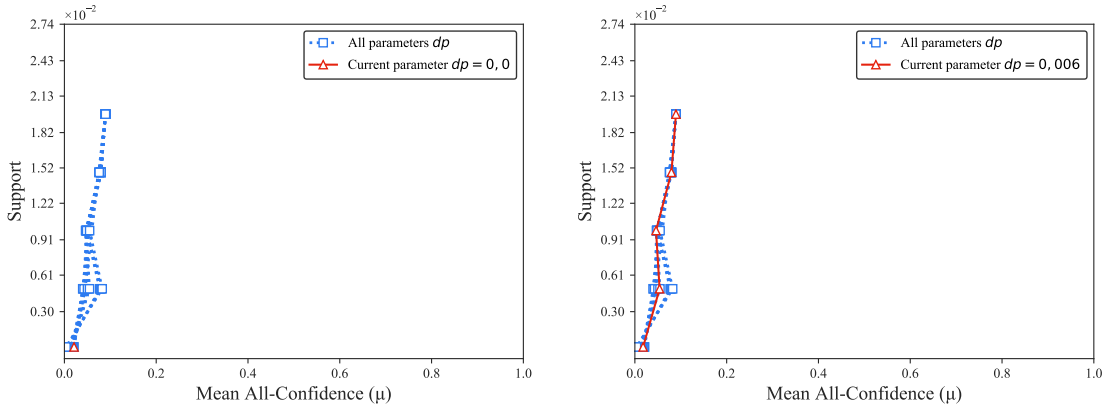


Figura B.63: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\,, 0{,}06\,, 0{,}07\}$, (b) com $dr = 0{,}08 \times 10^{-3}$, (c) com $dr = 0{,}09 \times 10^{-3}$, (d) com $dr = 0{,}10 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}11\,, 0{,}12\,, 0{,}13\,, 0{,}14\,, 0{,}15\}$. Veja Tabela B.53 para detalhes.

Tabela B.54: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,05] †		(0,05 , 0,10] †		(0,10 , 0,15] †		(0,15 , 0,20] †		(0,20 , 0,25] †		(0,25 , 0,30] †		(0,30 , 0,35] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	423	0,021	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	423	20,80
0,03	461	0,021	2	0,045	0	0,000	0	0,000	2	0,089	0	0,000	0	0,000	465	20,78
0,04	511	0,020	4	0,040	1	0,053	1	0,080	2	0,089	0	0,000	0	0,000	519	20,80
0,05	596	0,019	5	0,040	1	0,053	1	0,080	2	0,089	0	0,000	0	0,000	605	20,79
0,06	698	0,018	6	0,054	3	0,046	1	0,080	2	0,089	0	0,000	0	0,000	710	20,82
0,07	792	0,017	9	0,046	4	0,048	1	0,080	2	0,089	0	0,000	0	0,000	808	20,87
0,08	916	0,016	11	0,043	5	0,048	1	0,080	2	0,089	0	0,000	0	0,000	935	20,81
0,09	1.063	0,015	12	0,050	5	0,048	1	0,080	2	0,089	0	0,000	0	0,000	1.083	20,84
0,10	1.227	0,014	14	0,055	5	0,048	1	0,080	2	0,089	0	0,000	0	0,000	1.249	20,81
0,20	3.365	0,009	27	0,077	7	0,048	4	0,076	2	0,089	0	0,000	0	0,000	3.405	20,95
0,30	6.797	0,006	34	0,078	7	0,048	4	0,076	2	0,089	0	0,000	0	0,000	6.844	21,16
0,40	11.769	0,004	39	0,081	7	0,048	4	0,076	2	0,089	0	0,000	0	0,000	11.821	21,33
0,50 ‡	**18.504**	**0,003**	**47**	**0,082**	**8**	**0,055**	**4**	**0,076**	**2**	**0,089**	**0**	**0,000**	**0**	**0,000**	**18.565**	**21,60**



(a) (b)

Figura B.64: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v2}. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\}$ e (b) com $dr = 0{,}06 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela B.54 para detalhes.

Tabela B.55: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,05] †		(0,05 , 0,10] †		(0,10 , 0,15] †		(0,15 , 0,20] †		(0,20 , 0,25] †		(0,25 , 0,30] †		(0,30 , 0,35] †		#	(s)
‡ $\times 10^{-3}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**298**	**0,037**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**298**	**22,29**
0,30 ‡	305	0,037	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	305	22,28
0,40 ‡	325	0,035	4	0,119	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	329	22,33
0,45 ‡	389	0,030	6	0,088	0	0,000	1	0,080	0	0,000	0	0,000	0	0,000	396	22,32
0,50 ‡	499	0,025	6	0,088	0	0,000	1	0,080	0	0,000	0	0,000	0	0,000	506	22,33
0,55 ‡	882	0,015	7	0,095	0	0,000	1	0,080	1	0,084	0	0,000	0	0,000	891	22,36
0,60 ‡	1.627	0,009	11	0,095	0	0,000	1	0,080	1	0,084	0	0,000	0	0,000	1.640	22,42
0,70 ‡	8.020	0,003	14	0,098	1	0,043	1	0,080	2	0,089	0	0,000	1	0,146	8.039	22,89
0,80 ‡	16.198	0,002	16	0,090	3	0,044	1	0,080	2	0,089	0	0,000	1	0,146	16.221	23,61

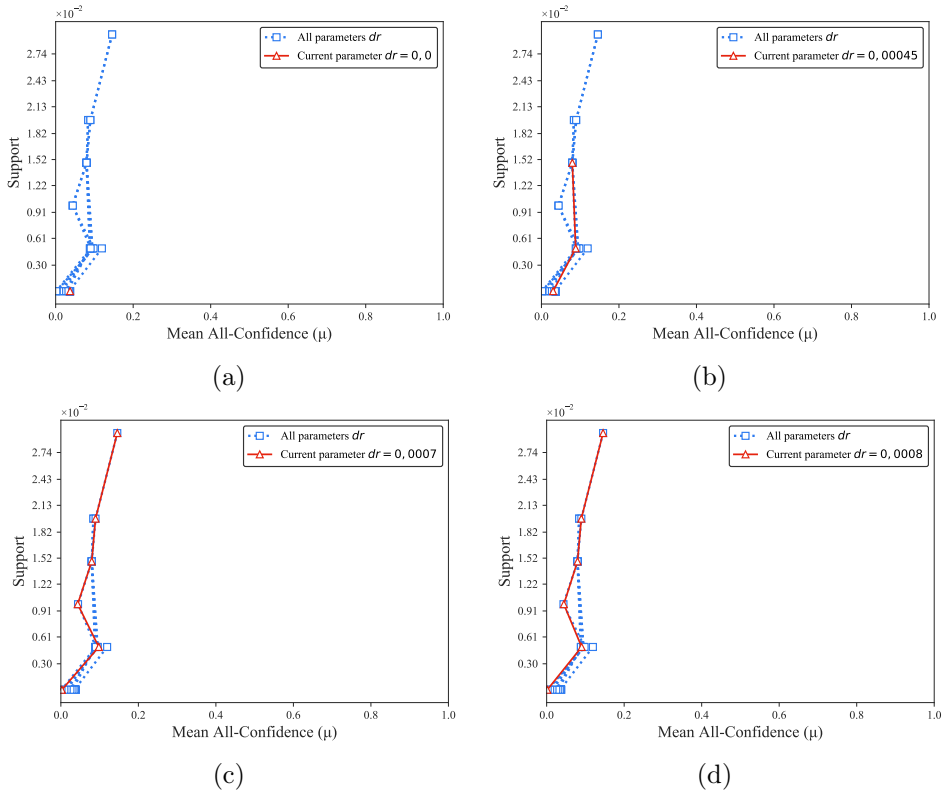
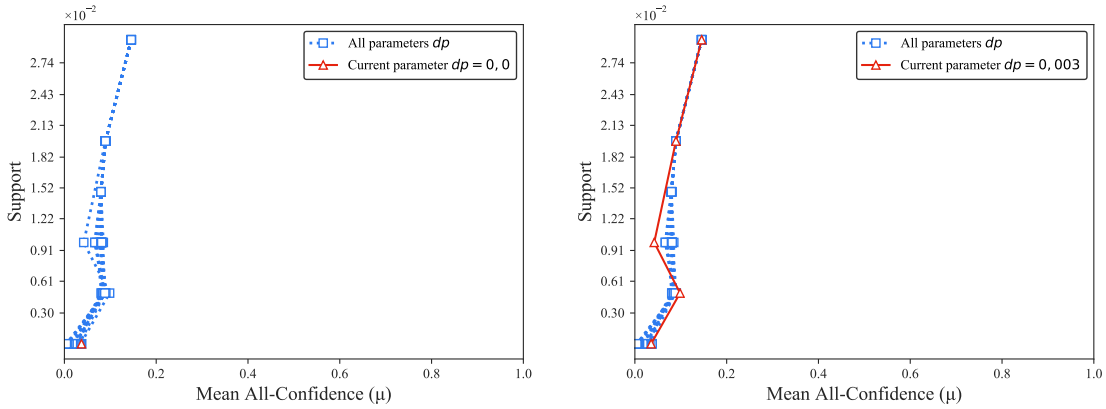


Figura B.65: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v3}}$. (a) com $dr = 0{,}00 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}30\,, 0{,}40\}$, (b) com $dr = 0{,}45 \times 10^{-3}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\,, 0{,}55\,, 0{,}60\}$, (c) com $dr = 0{,}70 \times 10^{-3}$ e (d) com $dr = 0{,}80 \times 10^{-3}$. Veja Tabela B.55 para detalhes.

Tabela B.56: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v4}}$.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,05] †		(0,05 , 0,10] †		(0,10 , 0,15] †		(0,15 , 0,20] †		(0,20 , 0,25] †		(0,25 , 0,30] †		(0,30 , 0,35] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,19
0,05 ‡	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,25
0,06 ‡	144	0,164	9	0,341	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	154	3,28
0,07 ‡	310	0,147	22	0,270	2	0,200	3	0,210	0	0,000	1	0,387	1	0,329	339	3,32
0,08 ‡	**363**	**0,140**	**24**	**0,264**	**3**	**0,235**	**3**	**0,210**	**2**	**0,260**	**2**	**0,357**	**1**	**0,329**	**398**	**3,34**
0,09 ‡	372	0,139	24	0,264	3	0,235	3	0,210	2	0,260	2	0,357	1	0,329	407	3,34
0,10 ‡	379	0,138	24	0,264	3	0,235	3	0,210	2	0,260	2	0,357	1	0,329	414	3,34
0,15 ‡	886	0,102	31	0,246	3	0,235	6	0,241	2	0,260	2	0,357	1	0,329	931	3,47
0,20 ‡	1.390	0,086	36	0,241	6	0,178	6	0,241	3	0,248	2	0,357	1	0,329	1.444	3,61
0,25 ‡	1.972	0,074	45	0,224	6	0,178	7	0,244	3	0,248	2	0,357	1	0,329	2.036	3,80
0,30 ‡	2.668	0,066	49	0,214	9	0,173	8	0,241	3	0,248	2	0,357	1	0,329	2.740	4,02



(a) (b)

Figura B.66: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,, 0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela B.56 para detalhes.

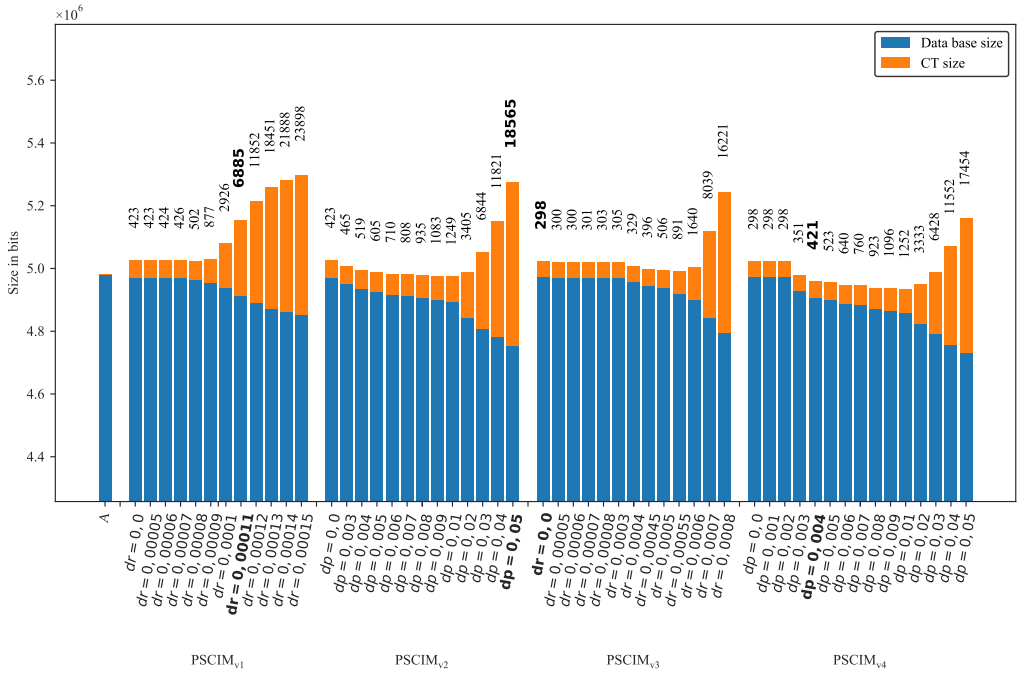


Figura B.67: *Fruithut*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2.6 OnlineRetail

Tabela B.57: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,08] †		(0,08 , 0,15] †		(0,15 , 0,23] †		(0,23 , 0,31] †		(0,31 , 0,39] †		(0,39 , 0,46] †		(0,46 , 0,54] †		#	(s)
‡ $\times 10^{-5}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	1.058	0,434	5	0,617	3	0,650	1	0,726	2	0,419	0	0,000	0	0,000	1.069	15,71
0,01 ‡	**1.061**	**0,433**	**5**	**0,617**	**3**	**0,650**	**1**	**0,726**	**2**	**0,419**	**0**	**0,000**	**1**	**0,536**	**1.073**	**16,17**
0,02 ‡	1.095	0,420	9	0,393	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.112	15,63
0,03 ‡	1.260	0,360	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.282	15,65
0,04 ‡	1.425	0,314	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.447	15,72
0,05 ‡	1.451	0,304	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.473	15,64
0,06 ‡	1.462	0,300	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.484	15,70
0,07 ‡	1.463	0,299	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.485	15,67
0,08 ‡	1.464	0,298	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.486	15,71
0,09 ‡	1.465	0,298	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.487	15,76
0,10 ‡	1.468	0,298	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.490	15,63
1,00 ‡	1.611	0,267	16	0,291	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.635	15,68
10,00 ‡	3.416	0,124	33	0,228	6	0,412	2	0,496	2	0,419	0	0,000	1	0,536	3.460	15,76
100,00 ‡	4.493	0,091	41	0,210	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	4.547	15,86

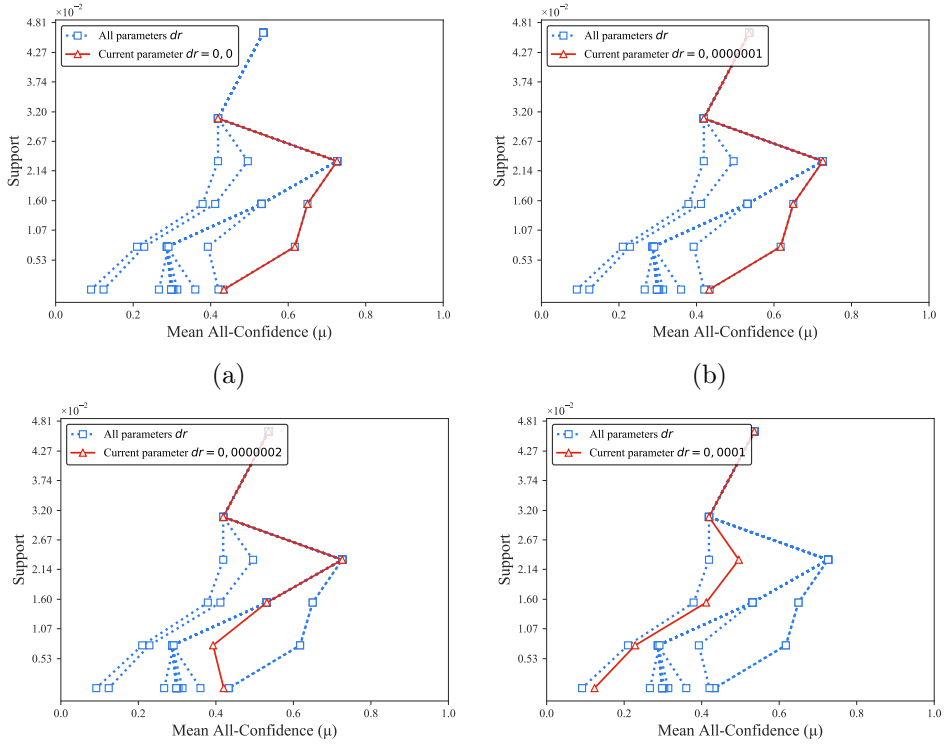


Figura B.68: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-5}$, (b) com $dr = 0{,}01 \times 10^{-5}$, (c) com $dr = 0{,}02 \times 10^{-5}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 1{,}00\}$ e (d) com $dr = 10{,}00 \times 10^{-5}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{100{,}00\}$. Veja Tabela B.57 para detalhes.

Tabela B.58: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,08] †		(0,08 , 0,15] †		(0,15 , 0,23] †		(0,23 , 0,31] †		(0,31 , 0,39] †		(0,39 , 0,46] †		(0,46 , 0,54] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	1.058	0,434	5	0,617	3	0,650	1	0,726	2	0,419	0	0,000	0	0,000	1.069	15,61
0,05 ‡	**1.059**	**0,434**	**6**	**0,551**	**3**	**0,650**	**1**	**0,726**	**2**	**0,419**	**0**	**0,000**	**1**	**0,536**	**1.072**	**15,51**
0,10 ‡	1.084	0,426	10	0,416	3	0,650	2	0,496	2	0,419	0	0,000	1	0,536	1.102	15,51
0,15 ‡	1.120	0,414	14	0,350	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.143	15,48
0,20 ‡	1.155	0,405	19	0,313	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.183	15,65
0,25 ‡	1.200	0,391	23	0,281	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.232	15,51
0,30 ‡	1.244	0,380	26	0,266	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.280	15,55
0,40 ‡	1.363	0,353	33	0,236	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.406	15,53
0,50 ‡	1.473	0,329	38	0,219	6	0,416	3	0,419	2	0,419	0	0,000	1	0,536	1.523	15,62
1,00 ‡	2.103	0,234	49	0,198	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	2.165	15,68
2,00 ‡	3.311	0,139	51	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	3.375	15,65

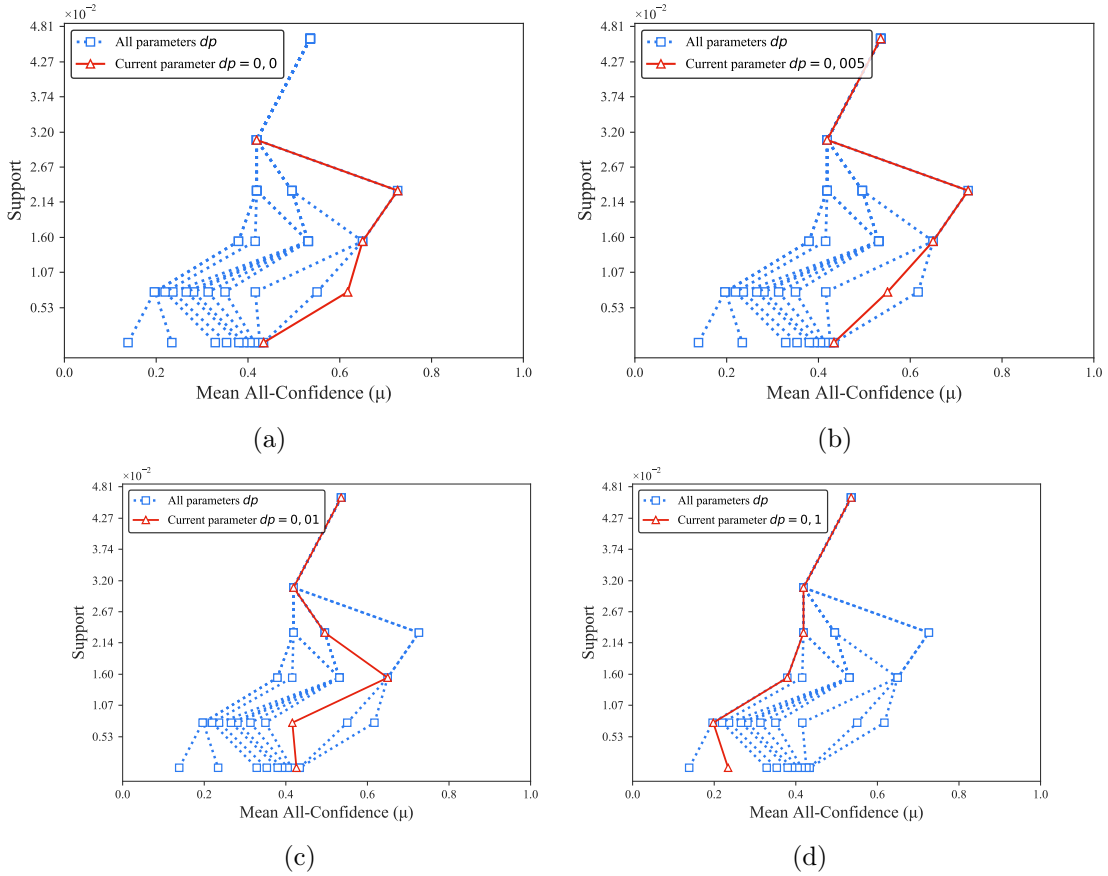


Figura B.69: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v2}. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$, (c) com $dr = 0{,}10 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}15\,, 0{,}20\,, 0{,}25\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$ e (d) com $dr = 1{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{2{,}00\}$. Veja Tabela B.58 para detalhes.

Tabela B.59: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,08] †		(0,08 , 0,15] †		(0,15 , 0,23] †		(0,23 , 0,31] †		(0,31 , 0,39] †		(0,39 , 0,46] †		(0,46 , 0,54] †		#	(s)
‡ $\times 10^{-5}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	1.064	0,429	5	0,617	3	0,650	1	0,726	2	0,419	0	0,000	0	0,000	1.075	15,81
0,01 ‡	**1.064**	**0,429**	**5**	**0,617**	**3**	**0,650**	**1**	**0,726**	**2**	**0,419**	**0**	**0,000**	**1**	**0,536**	**1.076**	**15,82**
0,02 ‡	1.080	0,424	7	0,476	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.095	15,94
0,03 ‡	1.182	0,385	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.204	15,78
0,04 ‡	1.355	0,329	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.377	15,82
0,05 ‡	1.479	0,298	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.501	17,09
0,06 ‡	1.505	0,290	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.527	16,79
0,07 ‡	1.507	0,289	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.529	17,65
0,08 ‡	1.508	0,287	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.530	15,79
0,09 ‡	1.508	0,287	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.530	15,84
0,10 ‡	1.508	0,287	14	0,287	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.530	15,92
1,00 ‡	1.711	0,249	15	0,298	4	0,531	1	0,726	2	0,419	0	0,000	1	0,536	1.734	15,86
10,00 ‡	3.339	0,126	32	0,232	6	0,412	2	0,496	2	0,419	0	0,000	1	0,536	3.382	15,90
100,00 ‡	4.539	0,090	46	0,200	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	4.598	15,96

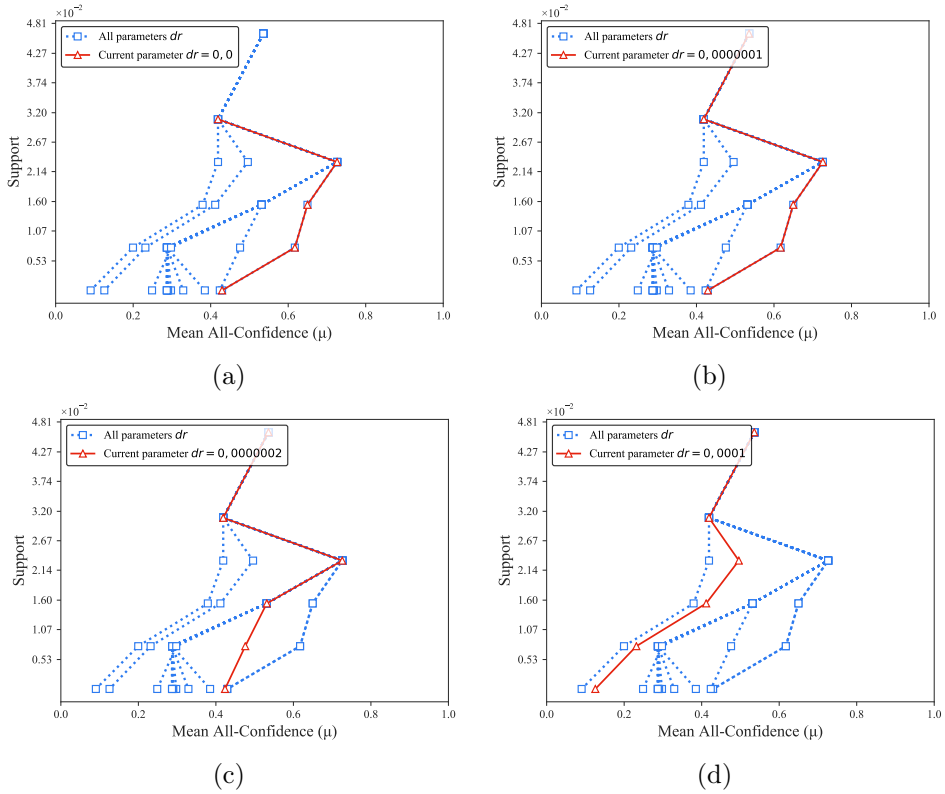


Figura B.70: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00 \times 10^{-5}$, (b) com $dr = 0{,}01 \times 10^{-5}$, (c) com $dr = 0{,}02 \times 10^{-5}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 1{,}00\}$ e (d) com $dr = 10{,}00 \times 10^{-5}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{100{,}00\}$. Veja Tabela B.59 para detalhes.

Tabela B.60: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,08] †		(0,08 , 0,15] †		(0,15 , 0,23] †		(0,23 , 0,31] †		(0,31 , 0,39] †		(0,39 , 0,46] †		(0,46 , 0,54] †		#	(s)
‡ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	1.064	0,429	5	0,617	3	0,650	1	0,726	2	0,419	0	0,000	0	0,000	1.075	15,71
0,05 ‡	**1.065**	**0,429**	**6**	**0,551**	**3**	**0,650**	**1**	**0,726**	**2**	**0,419**	**0**	**0,000**	**1**	**0,536**	**1.078**	**15,69**
0,10 ‡	1.091	0,421	10	0,416	3	0,650	2	0,496	2	0,419	0	0,000	1	0,536	1.109	15,77
0,15 ‡	1.126	0,410	13	0,369	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.148	15,67
0,20 ‡	1.158	0,401	17	0,333	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.184	15,71
0,25 ‡	1.207	0,388	21	0,293	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.238	15,64
0,30 ‡	1.255	0,376	25	0,271	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.290	15,66
0,40 ‡	1.357	0,351	31	0,242	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.398	15,74
0,50 ‡	1.487	0,324	40	0,213	6	0,416	3	0,419	2	0,419	0	0,000	1	0,536	1.539	15,66
1,00 ‡	2.114	0,232	49	0,198	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	2.176	15,79
2,00 ‡	3.340	0,137	51	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	3.404	15,80

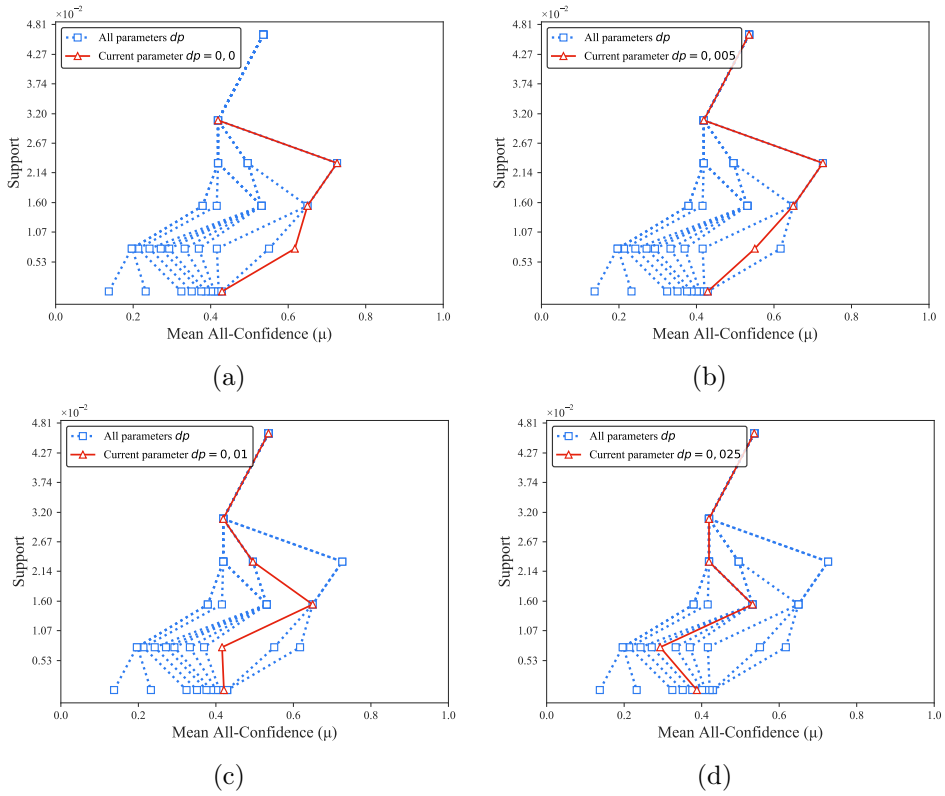


Figura B.71: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$, (c) com $dr = 0{,}10 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}15\,, 0{,}20\}$ e (d) com $dr = 0{,}25 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}30\,, 0{,}40\,, 0{,}50\,, 1{,}00\,, 2{,}00\}$. Veja Tabela B.60 para detalhes.

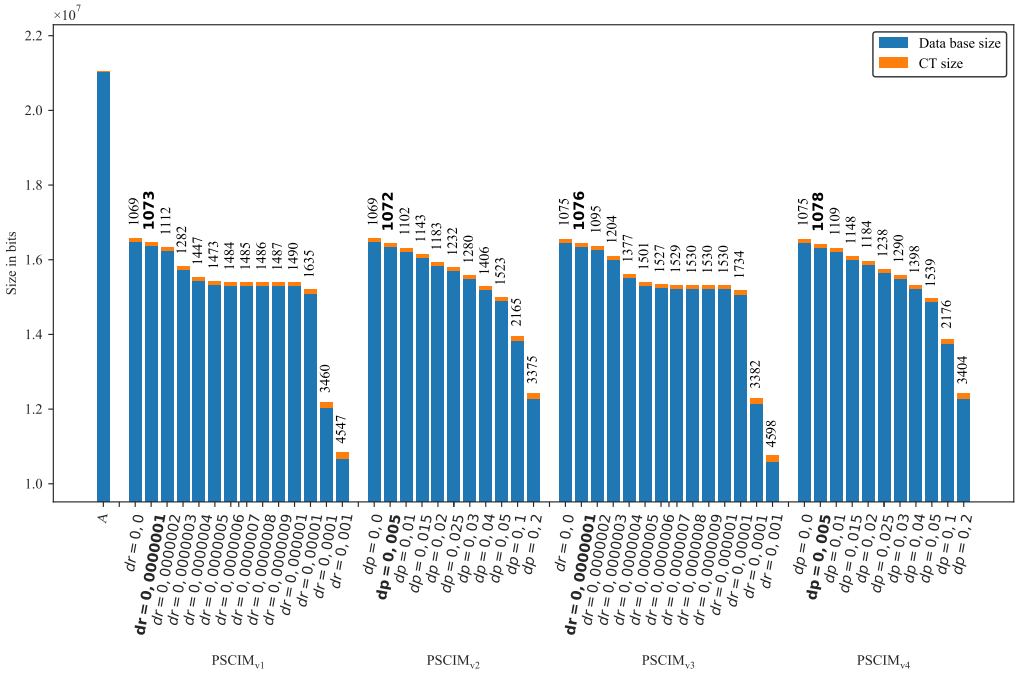


Figura B.72: *OnlineRetail*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e*, o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2.7 PAMP

Tabela B.61: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v1}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,13]		(0,13 , 0,26]		(0,26 , 0,38]		(0,38 , 0,51]		(0,51 , 0,64]		(0,64 , 0,77]		(0,77 , 0,90]		#	(s)
‡ × 10^{-1}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**74**	**0,120**	**6**	**0,474**	**11**	**0,487**	**6**	**0,693**	**5**	**0,806**	**1**	**0,946**	**2**	**0,920**	**105**	**9,05**
0,01 ‡	84	0,117	6	0,474	13	0,471	7	0,664	5	0,806	1	0,946	2	0,920	118	9,02
0,02 ‡	134	0,088	6	0,474	25	0,407	13	0,562	5	0,806	1	0,946	2	0,920	186	9,06
0,03 ‡	154	0,082	18	0,314	26	0,403	14	0,557	8	0,722	2	0,874	2	0,920	224	9,12
0,04 ‡	246	0,069	24	0,272	43	0,380	23	0,515	11	0,680	2	0,874	2	0,920	351	9,17
0,05 ‡	375	0,061	35	0,240	81	0,359	41	0,487	12	0,682	2	0,874	8	0,896	554	9,24
0,06 ‡	461	0,061	48	0,239	341	0,334	81	0,462	20	0,643	7	0,817	16	0,887	974	9,34
0,07 ‡	738	0,058	160	0,252	702	0,331	112	0,455	21	0,640	18	0,790	24	0,884	1.775	9,46
0,08 ‡	1.002	0,058	344	0,256	1.273	0,328	135	0,456	48	0,637	126	0,747	65	0,874	2.993	9,54
0,09 ‡	1.129	0,057	389	0,253	1.328	0,329	226	0,464	68	0,626	162	0,747	66	0,874	3.368	9,63
0,10 ‡	1.234	0,057	487	0,246	1.330	0,329	262	0,468	157	0,638	274	0,741	79	0,870	3.823	9,69
0,11 ‡	1.804	0,057	517	0,242	1.337	0,329	269	0,469	351	0,632	389	0,738	95	0,865	4.762	9,74
0,12 ‡	3.615	0,067	1.533	0,238	1.987	0,326	288	0,466	431	0,633	433	0,737	97	0,865	8.384	10,08

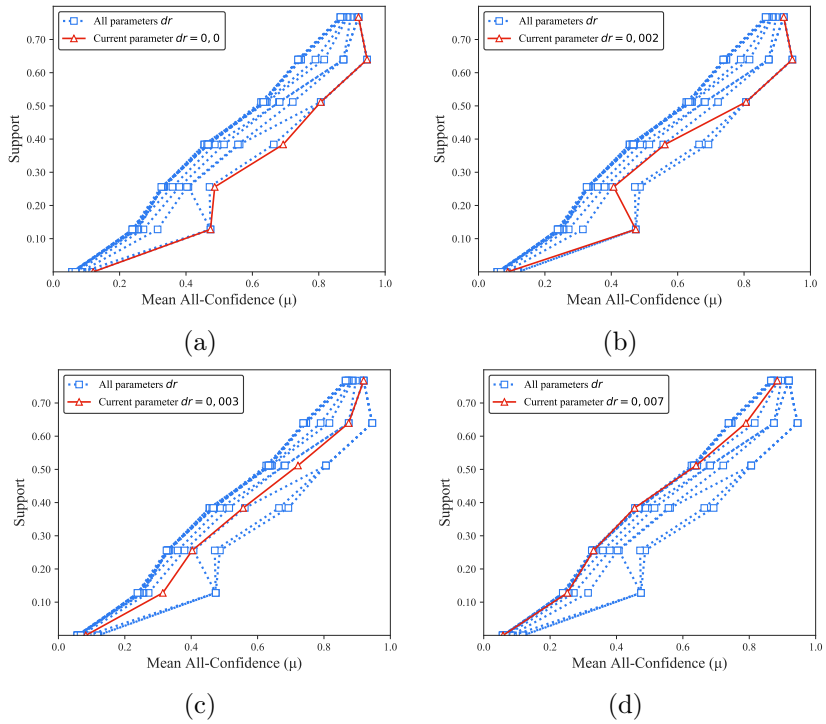
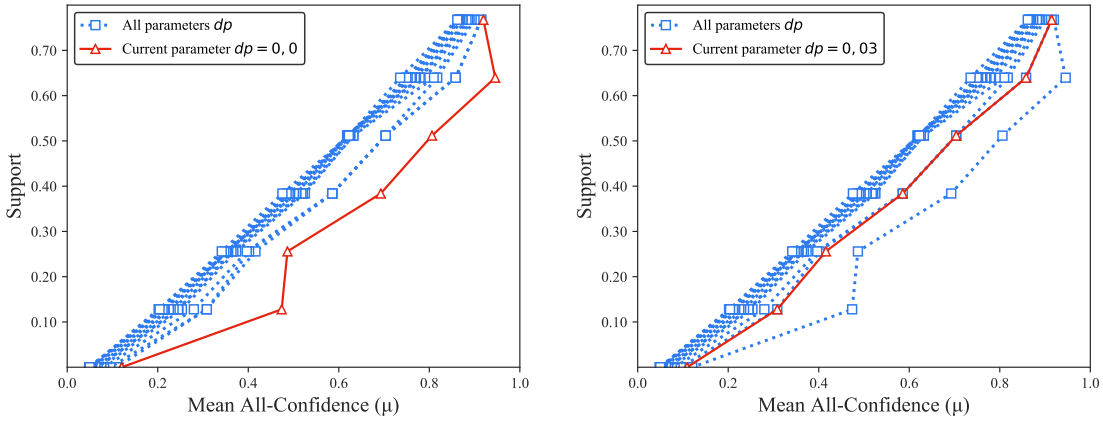


Figura B.73: *PAMAP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{\text{v1}}$. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\}$, (b) com $dr = 0{,}02 \times 10^{-1}$, (c) com $dr = 0{,}03 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\}$, e (b) com $dr = 0{,}07 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\}$. Veja Tabela B.61 para detalhes.

Tabela B.62: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{\text{v2}}$.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,13]		(0,13 , 0,26]		(0,26 , 0,38]		(0,38 , 0,51]		(0,51 , 0,64]		(0,64 , 0,77]		(0,77 , 0,90]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**74**	**0,120**	**6**	**0,474**	**11**	**0,487**	**6**	**0,693**	**5**	**0,806**	**1**	**0,946**	**2**	**0,920**	**105**	**9,12**
0,03	92	0,109	17	0,308	21	0,416	15	0,586	10	0,704	3	0,858	10	0,915	168	9,16
0,04	123	0,095	17	0,308	25	0,402	15	0,586	10	0,704	3	0,858	10	0,915	203	9,21
0,05	162	0,082	25	0,280	28	0,397	15	0,586	10	0,704	7	0,818	22	0,900	269	9,25
0,06	198	0,075	44	0,253	43	0,373	34	0,525	23	0,632	9	0,810	26	0,893	377	9,39
0,07	258	0,067	48	0,246	53	0,367	36	0,521	23	0,632	18	0,791	35	0,889	471	9,46
0,08	266	0,067	60	0,245	69	0,363	39	0,517	26	0,625	24	0,781	42	0,886	526	9,51
0,09	345	0,062	101	0,235	86	0,355	72	0,506	37	0,620	33	0,779	47	0,882	721	9,56
0,10	475	0,055	132	0,229	129	0,351	73	0,505	43	0,623	58	0,771	74	0,874	984	9,68
0,11	527	0,054	156	0,225	147	0,353	104	0,501	57	0,618	64	0,768	75	0,873	1.130	9,77
0,12	648	0,052	236	0,215	181	0,355	168	0,492	72	0,619	127	0,761	88	0,869	1.520	9,90
0,13	811	0,050	305	0,214	212	0,353	169	0,492	89	0,620	159	0,759	96	0,867	1.841	9,97
0,14	1.104	0,048	405	0,211	331	0,351	283	0,489	139	0,622	238	0,752	104	0,865	2.604	10,20
0,15	1.319	0,051	595	0,202	397	0,347	350	0,485	166	0,625	319	0,749	109	0,864	3.255	10,29
0,20	5.025	0,050	1.668	0,204	1.722	0,341	1.279	0,476	905	0,622	698	0,736	118	0,862	11.415	11,97

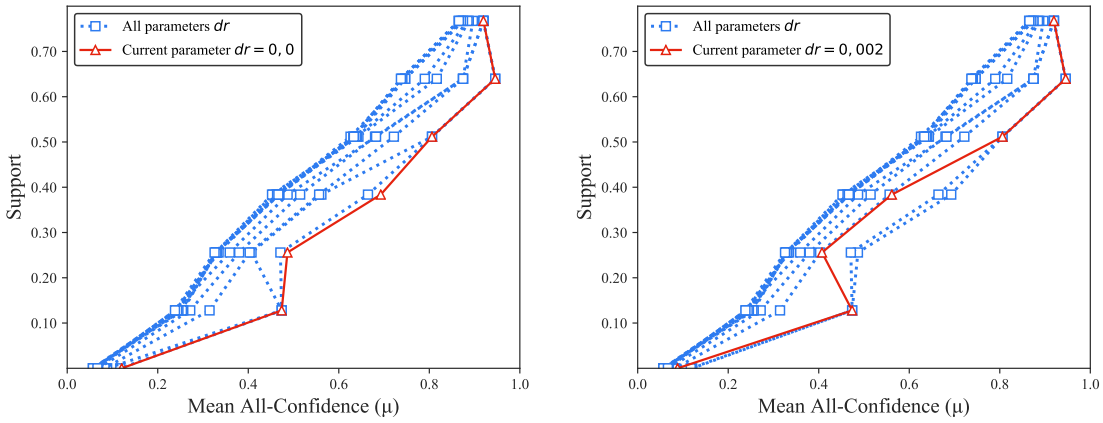


(a) (b)

Figura B.74: *PAMAP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v2}$. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{1{,}00\,, 2{,}00\}$ e (b) com $dr = 2{,}10$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{2{,}20\,, 2{,}30\,, 2{,}40\,, 2{,}50\,, 2{,}60\,, 2{,}70\,, 2{,}80\,, 2{,}90\,, 3{,}00\,, 4{,}00\}$. Veja Tabela B.62 para detalhes.

Tabela B.63: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v3}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,13]		(0,13 , 0,26]		(0,26 , 0,38]		(0,38 , 0,51]		(0,51 , 0,64]		(0,64 , 0,77]		(0,77 , 0,90]		#	(s)
‡ × 10^{-1}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	**74**	**0,120**	**6**	**0,474**	**11**	**0,487**	**6**	**0,693**	**5**	**0,806**	**1**	**0,946**	**2**	**0,920**	**105**	**8,80**
0,01 ‡	84	0,117	6	0,474	13	0,471	7	0,664	5	0,806	1	0,946	2	0,920	118	8,80
0,02 ‡	134	0,088	6	0,474	25	0,407	13	0,562	5	0,806	1	0,946	2	0,920	186	8,86
0,03 ‡	154	0,082	18	0,314	26	0,403	14	0,557	8	0,722	2	0,874	2	0,920	224	8,89
0,04 ‡	246	0,069	24	0,272	43	0,380	23	0,515	11	0,680	2	0,874	2	0,920	351	8,94
0,05 ‡	375	0,061	35	0,240	81	0,359	41	0,487	12	0,682	2	0,874	8	0,896	554	9,03
0,06 ‡	461	0,061	48	0,239	341	0,334	81	0,462	20	0,643	7	0,817	16	0,887	974	9,12
0,07 ‡	738	0,058	160	0,252	702	0,331	112	0,455	21	0,640	18	0,790	24	0,884	1.775	9,27
0,08 ‡	1.002	0,058	344	0,256	1.273	0,328	128	0,452	43	0,639	126	0,747	65	0,874	2.981	9,32
0,09 ‡	1.129	0,057	389	0,253	1.328	0,329	226	0,464	68	0,626	162	0,747	66	0,874	3.368	9,39
0,10 ‡	1.235	0,057	487	0,246	1.330	0,329	262	0,468	157	0,638	274	0,741	79	0,870	3.824	9,47
0,11 ‡	1.804	0,057	501	0,243	1.337	0,329	269	0,469	351	0,632	389	0,738	95	0,865	4.746	9,53
0,12 ‡	3.615	0,067	1.533	0,238	1.987	0,326	288	0,466	431	0,633	433	0,737	97	0,865	8.384	9,87

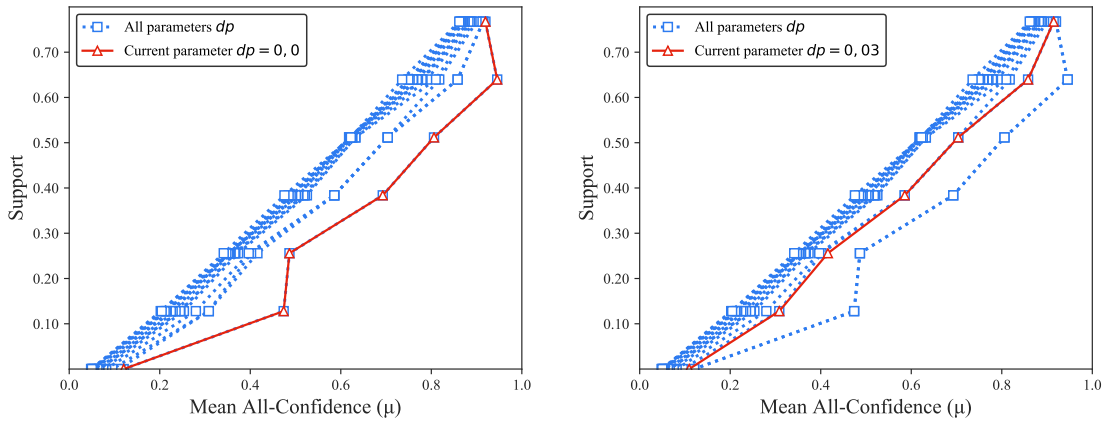


(a) (b)

Figura B.75: *PAMAP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM$_{v3}$. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\}$ e (b) com $dr = 0{,}02 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\}$. Veja Tabela B.63 para detalhes.

Tabela B.64: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM$_{v4}$.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,13]		(0,13 , 0,26]		(0,26 , 0,38]		(0,38 , 0,51]		(0,51 , 0,64]		(0,64 , 0,77]		(0,77 , 0,90]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**74**	**0,120**	**6**	**0,474**	**11**	**0,487**	**6**	**0,693**	**5**	**0,806**	**1**	**0,946**	**2**	**0,920**	**105**	**8,99**
0,02	74	0,120	6	0,474	11	0,487	6	0,693	5	0,806	1	0,946	2	0,920	105	8,98
0,03	92	0,109	17	0,308	21	0,416	15	0,586	10	0,704	3	0,858	10	0,915	168	9,11
0,04	123	0,095	17	0,308	25	0,402	15	0,586	10	0,704	3	0,858	10	0,915	203	9,16
0,05	162	0,082	25	0,280	28	0,397	15	0,586	10	0,704	7	0,818	22	0,900	269	9,20
0,06	198	0,075	44	0,253	43	0,373	34	0,525	23	0,632	9	0,810	26	0,893	377	9,32
0,07	258	0,067	48	0,246	53	0,367	36	0,521	23	0,632	18	0,791	35	0,889	471	9,40
0,08	270	0,066	60	0,245	65	0,365	39	0,517	26	0,625	24	0,781	42	0,886	526	9,43
0,09	345	0,062	101	0,235	82	0,357	72	0,506	37	0,620	33	0,779	47	0,882	717	9,53
0,10	475	0,055	132	0,229	125	0,352	73	0,505	43	0,623	58	0,771	74	0,874	980	9,61
0,11	527	0,054	156	0,225	147	0,353	104	0,501	57	0,618	64	0,768	75	0,873	1.130	9,74
0,12	648	0,052	236	0,215	181	0,355	168	0,492	72	0,619	118	0,760	83	0,870	1.506	9,87
0,13	811	0,050	305	0,214	216	0,353	169	0,492	89	0,620	159	0,759	96	0,867	1.845	9,97
0,14	1.104	0,048	411	0,211	329	0,351	283	0,489	139	0,622	238	0,752	104	0,865	2.608	10,21
0,15	1.319	0,051	601	0,202	395	0,347	350	0,485	166	0,625	319	0,749	109	0,864	3.259	10,28
0,20	5.038	0,050	1.652	0,204	1.722	0,341	1.279	0,476	905	0,622	698	0,736	118	0,862	11.412	11,95



(a) (b)

Figura B.76: *PAMAP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\,, 0{,}14\,, 0{,}15\,, 0{,}20\}$. Veja Tabela B.64 para detalhes.

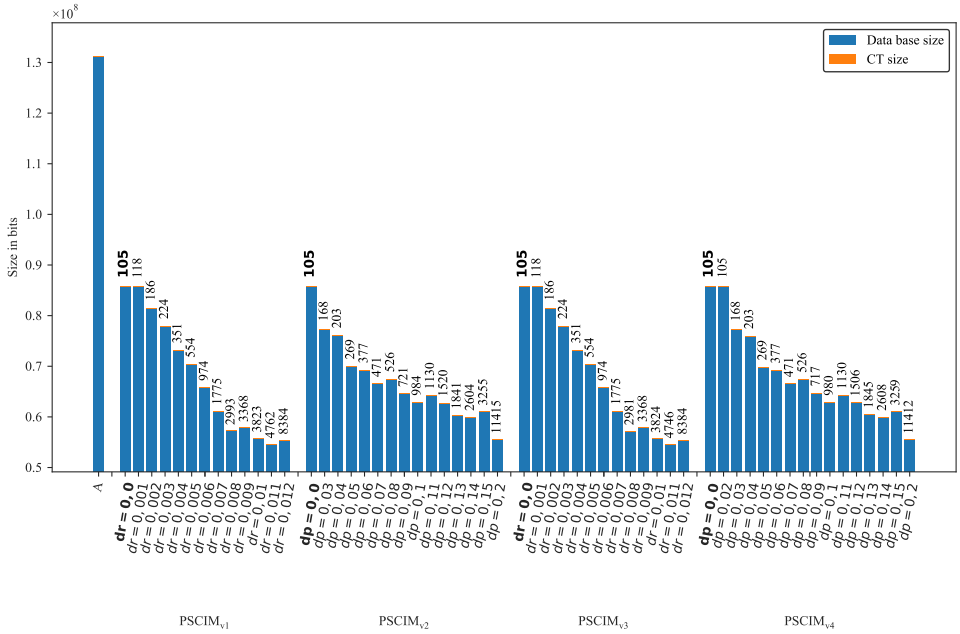


Figura B.77: *PAMAP*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

B.1.2.8 Retail

Tabela B.65: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v1}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,05]		(0,05 , 0,09]		(0,09 , 0,14]		(0,14 , 0,19]		(0,19 , 0,24]		(0,24 , 0,28]		(0,28 , 0,33]		#	(s)
$\ddagger \times 10^{-6}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4.436	8.352,13
0,10 ‡	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.437	8.352,16
0,20 ‡	4.439	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.440	8.352,38
0,30 ‡	4.609	0,126	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.610	8.352,18
0,31 ‡	4.645	0,125	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.646	8.352,28
0,32 ‡	4.703	0,124	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.704	8.352,20
0,33 ‡	4.773	0,122	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.774	8.352,25
0,34 ‡	4.860	0,119	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.862	8.352,22
0,35 ‡	4.948	0,117	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.950	8.352,26
0,36 ‡	5.051	0,115	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.053	8.352,27
0,37 ‡	5.152	0,112	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.154	8.352,24
0,38 ‡	5.280	0,109	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.282	8.352,47
0,39 ‡	5.530	0,104	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.532	8.352,48
0,40 ‡	5.786	0,100	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.788	8.352,49
0,45 ‡	**9.830**	**0,059**	**2**	**0,168**	**1**	**0,225**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**1**	**0,575**	**9.834**	**8.353,71**

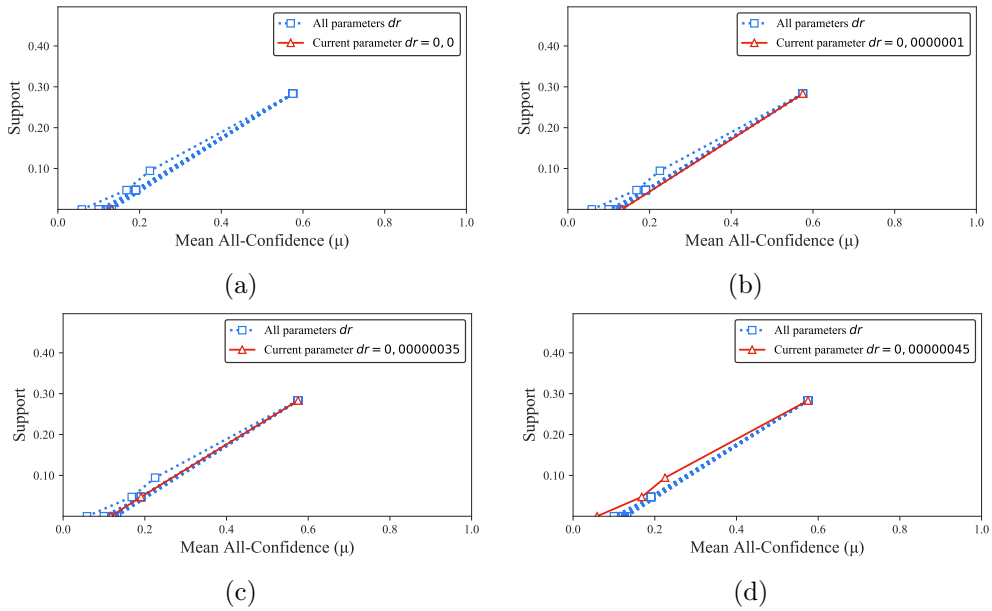


Figura B.78: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v1}. (a) com $dr = 0{,}00 \times 10^{-6}$, (b) com $dr = 0{,}10 \times 10^{-6}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}30\,, 0{,}31\,, 0{,}32\,, 0{,}33\,, 0{,}34\}$, (c) com $dr = 0{,}35 \times 10^{-6}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}36\,, 0{,}37\,, 0{,}38\,, 0{,}39\,, 0{,}40\}$ e (d) com $dr = 0{,}45 \times 10^{-6}$. Veja Tabela B.65 para detalhes.

Tabela B.66: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v2}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,05]		(0,05 , 0,09]		(0,09 , 0,14]		(0,14 , 0,19]		(0,19 , 0,24]		(0,24 , 0,28]		(0,28 , 0,33]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4.436	8.353,36
0,03 ‡	4.451	0,132	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.453	8.353,33
0,04 ‡	4.514	0,130	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.516	8.353,37
0,05 ‡	4.647	0,126	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.649	8.353,41
0,06 ‡	4.802	0,123	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.804	8.353,49
0,07 ‡	5.044	0,117	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.046	8.353,64
0,08 ‡	5.326	0,112	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.328	8.353,63
0,09 ‡	5.584	0,107	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.586	8.353,76
0,10 ‡	5.894	0,103	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.896	8.353,83
0,20 ‡	**9.729**	**0,072**	**1**	**0,191**	**1**	**0,225**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**1**	**0,575**	**9.732**	**8.355,00**
0,30 ‡	14.179	0,057	2	0,168	2	0,220	0	0,000	0	0,000	0	0,000	1	0,575	14.184	8.356,31

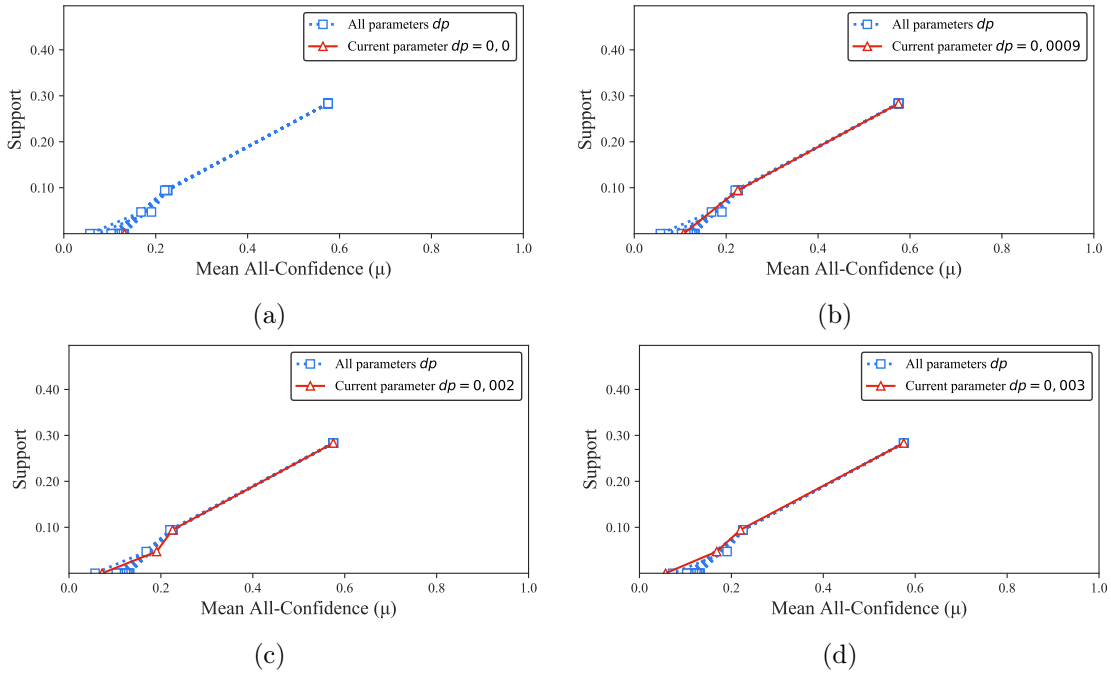


Figura B.79: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v2}. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\}$, (b) com $dr = 0{,}09 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\}$, (c) com $dr = 0{,}20 \times 10^{-2}$ e (d) com $dr = 0{,}30 \times 10^{-2}$. Veja Tabela B.66 para detalhes.

Tabela B.67: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v3}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,05]		(0,05 , 0,09]		(0,09 , 0,14]		(0,14 , 0,19]		(0,19 , 0,24]		(0,24 , 0,28]		(0,28 , 0,33]		#	(s)
$\ddagger \times 10^{-6}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4.436	8.352,13
0,10 ‡	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.437	8.352,16
0,20 ‡	4.439	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.440	8.352,14
0,30 ‡	4.609	0,126	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.610	8.352,15
0,31 ‡	4.645	0,125	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.646	8.352,18
0,32 ‡	4.703	0,124	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.704	8.352,24
0,33 ‡	4.773	0,122	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.774	8.352,14
0,34 ‡	4.860	0,119	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.862	8.352,21
0,35 ‡	4.948	0,117	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	4.950	8.352,19
0,36 ‡	5.051	0,115	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.053	8.352,32
0,37 ‡	5.152	0,112	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.154	8.352,23
0,38 ‡	5.280	0,109	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.282	8.352,28
0,39 ‡	5.530	0,104	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.532	8.352,31
0,40 ‡	5.786	0,100	1	0,191	0	0,000	0	0,000	0	0,000	0	0,000	1	0,575	5.788	8.352,40
0,45 ‡	**9.830**	**0,059**	**2**	**0,168**	**1**	**0,225**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**1**	**0,575**	**9.834**	**8.353,59**

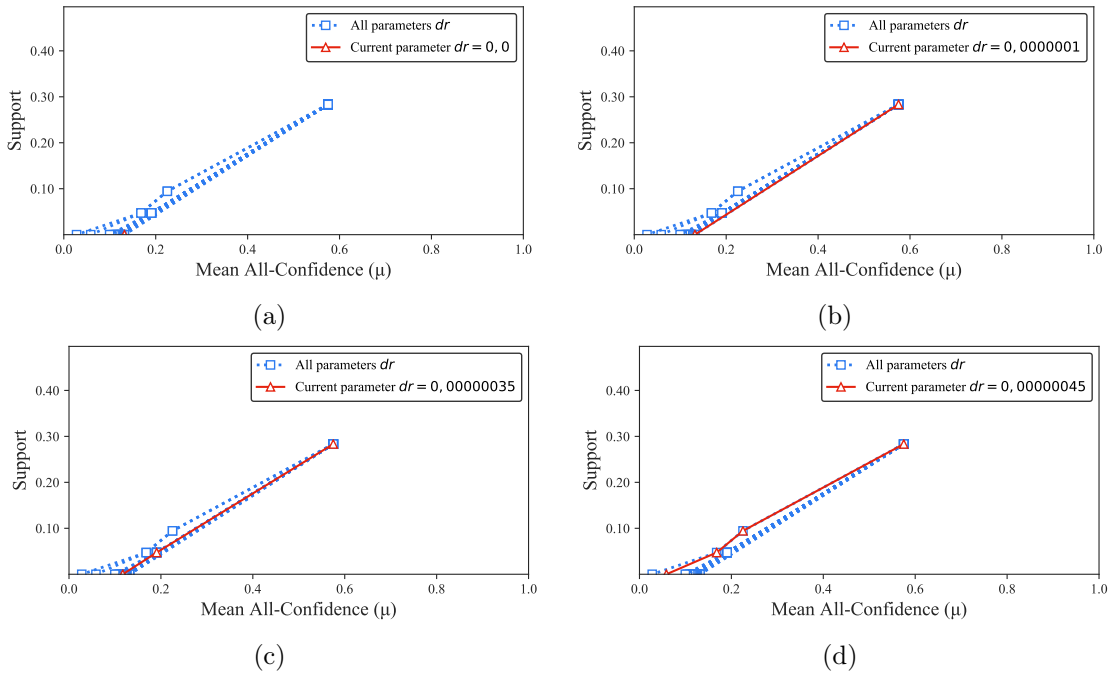


Figura B.80: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v3}. (a) com $dr = 0{,}00 \times 10^{-6}$, (b) com $dr = 0{,}10 \times 10^{-6}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}30\,, 0{,}31\,, 0{,}32\,, 0{,}33\,, 0{,}34\}$, (c) com $dr = 0{,}35 \times 10^{-6}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}36\,, 0{,}37\,, 0{,}38\,, 0{,}39\,, 0{,}40\}$ e (d) com $dr = 0{,}45 \times 10^{-6}$. Veja Tabela B.67 para detalhes.

Tabela B.68: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM_{v4}.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,05]		(0,05 , 0,09]		(0,09 , 0,14]		(0,14 , 0,19]		(0,19 , 0,24]		(0,24 , 0,28]		(0,28 , 0,33]		#	(s)
$\ddagger \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ‡	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4.436	8.352,25
0,03 ‡	4.451	0,132	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.453	8.352,19
0,04 ‡	4.514	0,130	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.516	8.352,25
0,05 ‡	4.647	0,126	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.649	8.352,28
0,06 ‡	4.802	0,123	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.804	8.352,39
0,07 ‡	5.044	0,117	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.046	8.352,45
0,08 ‡	5.326	0,112	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.328	8.352,57
0,09 ‡	5.584	0,107	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.586	8.352,66
0,10 ‡	5.894	0,103	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.896	8.352,75
0,20 ‡	**9.729**	**0,072**	**1**	**0,191**	**1**	**0,225**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**1**	**0,575**	**9.732**	**8.353,96**
0,30 ‡	14.179	0,057	2	0,168	2	0,220	0	0,000	0	0,000	0	0,000	1	0,575	14.184	8.355,32

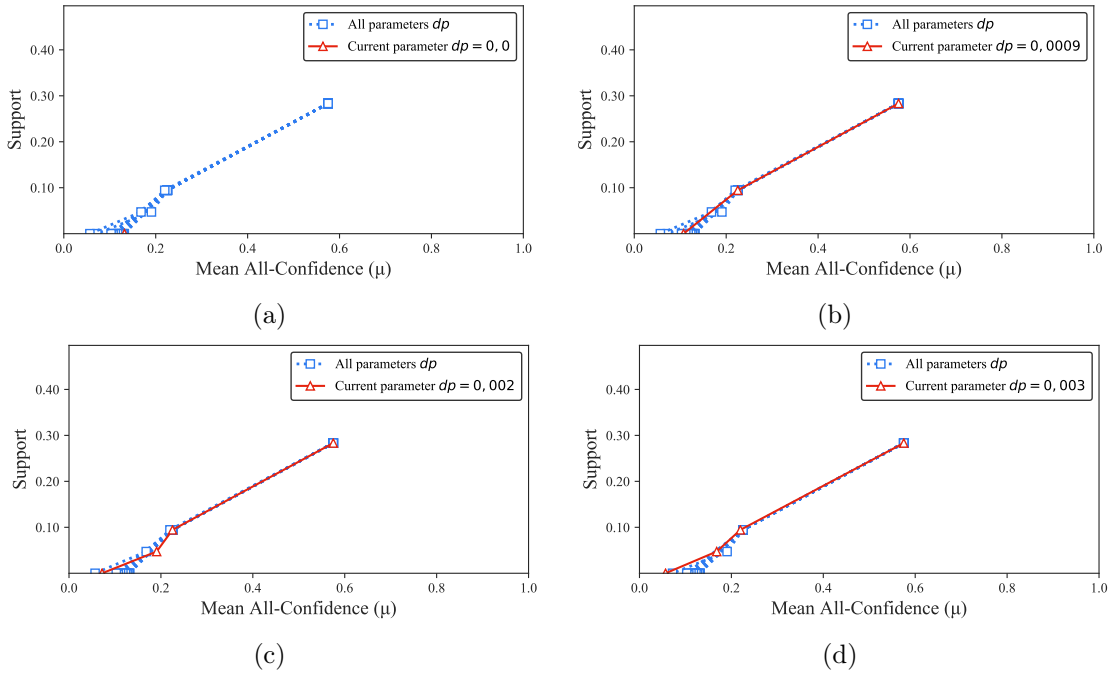


Figura B.81: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM_{v4}. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\}$, (b) com $dr = 0{,}09 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\}$, (c) com $dr = 0{,}20 \times 10^{-2}$ e (d) com $dr = 0{,}30 \times 10^{-2}$. Veja Tabela B.68 para detalhes.

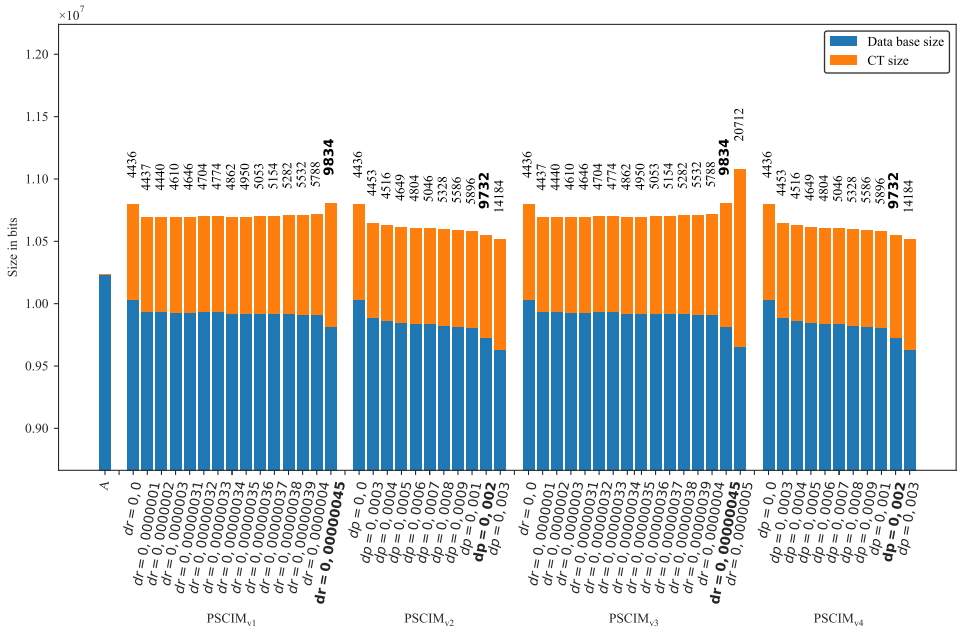


Figura B.82: *Retail*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

APÊNDICE C – PSCIM: EXPERIMENTOS COM DIFERENTES CONFIGURAÇÕES DE PARÂMETROS

Este documento apresenta a análise detalhada da escolha dos parâmetros para a comparação não tendenciosa dos algoritmos. Ele inclui as métricas calculadas por LAM [55], TopPI [30] e PSCIM [40] em todas as bases de dados usadas em nossos experimentos. Todos os algoritmos citados executam em *multithreading*, por isso, nesse estudo foi definido 8 threads. Usamos 16 bases de dados do SMPF [21]. Em algumas bases de dados, nós removemos as transações que tinham zero item e removemos a repetição de determinado item na mesma transação, consideramos esses dois casos como um erro para uma base de dados transacional. Neste estudo as bases de dados são separadas em dois grupos: bases de dados esparsas e densas. Essa decisão facilita a compreensão do comportamento dos algoritmos para esses dois cenários de tipo das bases de dados.

As Tabelas C.1 e C.2 resumem as organizações gerais deste apêndice em relação à análise do comportamento dos algoritmos TopPI e PSCIM sob diferentes parametrizações. Mais especificamente, as tabelas e figuras referenciadas pelas colunas de dois a cinco das Tabelas C.1 e C.2 mostram a distribuição dos valores médios de *all-confidence* (μ) calculados para os itemsets fechados recuperados por TopPI e PSCIM, usando diferentes valores de parâmetros. Os números referenciados pela coluna seis das Tabelas C.1 e C.2 mostram o MDL (Equação 2.6) calculado para os itemsets fechados recuperados por cada abordagem, sob diferentes parametrizações.

Para calcular μ, primeiro criamos sete partições de valores de suporte com o mesmo tamanho sobre o intervalo de suporte, dado uma determinada base de dados. Em seguida, calculamos μ a partir dos conjuntos de itens fechados recuperados em cada partição. Como pode ser visto nas figuras referenciadas pelas colunas três e cinco das Tabelas C.1 e C.2, comparamos as distribuições de medidas da média *all-confidence* para escolher os melhores valores de parâmetro para os algoritmo TopPI e PSCIM em cada base de dados.

Tabela C.1: Bases de dados densa: Conjunto de tabelas e figuras mostrando o comportamento dos algoritmos TopPI e SCIM sob diferentes parâmetros.

Bases de dados densa	**Média *All-Confidence***				**MDL**
	PSCIM		TopPI		
	Tabela	Figura	Tabela	Figura	Figura
Chess	C.7	C.3	C.8	C.4	C.5
Kddcup99	C.9	C.6	C.10	C.7	C.8
Mushrooms	C.11	C.9	C.12	C.10	C.11
PowerC	C.13	C.12	C.14	C.13	C.14
Pumsb	C.15	C.15	C.16	C.16	C.17
RecordLink	C.17	C.18	C.18	C.19	C.20
Skin	C.19	C.21	C.20	C.22	C.23
Susy	C.21	C.24	C.22	C.25	C.26

Nessas figuras, os eixos horizontais correspondem a μ variando de 0 a 1, enquanto os eixos verticais distribuem os valores de suporte de 0 para o limite superior de cada base de dados. Espera-se que quanto melhor for os itemsets fechados recuperados, mais à direita será a curva que representa o desempenho de uma técnica/parametrização. Os valores dos parâmetros escolhidos são destacados em negrito nas tabelas referenciadas pelas colunas dois e quatro das Tabelas C.1 e C.2. Embora não seja usada como critério de escolha de parâmetros das variações do algoritmo PSCIM, é apresentado neste estudo a média de *cross-support* (Equação 7) a partir dos conjuntos de itens fechados recuperados em cada partição de suporte. Os números referenciados pela coluna seis das Tabelas C.1 e C.2 mostram o MDL calculado para os conjuntos de itens fechados recuperados por cada abordagem, sob diferentes parametrizações. Os valores de MDL são usados para calcular $L\%$. Veja Equação 2.6 para detalhes. Essa métrica não é usada para critério de escolha de parâmetros das variações do algoritmo PSCIM.

As Tabelas C.3 e C.4 resumem para uma determinada base de dados, o parâmetro vencedor obtido para PSCIM e TopPI. A primeira coluna mostra o nome da base de dados, seguido do parâmetro escolhido por cada algoritmo. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. As colunas três a nove mostram o número de itemsets fechados recuperados (#), os valores médios (μ) e os valores de desvio padrão (σ) calculados por intervalo de suporte por todos os algoritmos dado a métrica utilizada. As duas últimos colunas apresentam o número total de itemsets fechados detectados (#) e a média de 10 execuções dos tempos de processamento (em segundos). Para os algoritmos LAM e Slim não foram necessários definir parâmetros. A implementação dos dois algoritmos possuem como entrada o

Tabela C.2: Bases de dados esparsa: Conjunto de tabelas e figuras mostrando o comportamento dos algoritmos TopPI e SCIM sob diferentes parâmetros.

Bases de dados esparsa	**Média *All-Confidence***				**MDL**
	PSCIM		TopPI		
	Tabela	Figura	Tabela	Figura	Figura
Accidents	C.23	C.27	C.24	C.28	C.29
BMSWebView2	C.25	C.30	C.26	C.31	C.32
BMS1	C.27	C.33	C.28	C.34	C.35
FoodmartFIM	C.29	C.36	C.30	C.37	C.38
Fruithut	C.31	C.39	C.32	C.40	C.41
OnlineRetail	C.33	C.42	C.34	C.43	C.44
PAMP	C.35	C.45	C.36	C.46	C.47
Retail	C.37	C.48	C.38	C.49	C.50

parâmetro mínimo de suporte conjuntivo, nesse caso, usamos o valor 1 de suporte conjuntivo. Dessa forma, o algoritmo se torna livre de parâmetro, sem ter nenhuma restrição de suporte conjuntivo para os itemsets fechados recuperados. Na Tabela C.4, na primeira coluna, todos os nomes de base de dados, exceto a base de dados *Accidents*, apresentam o símbolo (τ) na frente do nome, isso significa que a base corrente possui transações com tamanho 1, ou seja, a transação contém apenas um item. O símbolo (†) pode aparecer nas colunas 1 a 7, nas bases de dados onde a faixa de suporte é menor que 0,10 de suporte, nestes casos usamos notação científica para representar a faixa de valores de suporte. A Figura C.2 ilustram as distribuições de μ das bases de dados esparsas, se as faixas de suporte forem menor que 0,10, a proporção de largura e altura não é mantido, visto que nesses casos o eixo y é muito menor que o eixo x, dificultando a leitura do gráfico.

As Tabelas C.5 e C.6 resumem os valores de significância estatística obtido por cada técnica comparada para cada base de dados. A primeira coluna contém a informação da base de dados junto com as informações de cada algoritmo a ser comparado com o algoritmo PSCIM. A segunda coluna mostra qual a métrica corrente usada para calcular as médias e os desvios padrões das partições de suporte. A terceira coluna mostra a hipótese nula (H_0) usado no teste de significância estatística. As colunas quatro a dez mostram os *p-value* calculados por intervalo de suporte por todas as técnicas comparadas. Os testes apresentados são referentes as Tabelas C.3 e C.4.

Tabela C.3: Desempenho do algoritmo/parametrização das técnicas *PSCIM*, *TopPI* e *LAM* sobre as bases de dados densas da Tabela C.1.

Chess	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	42	0,069	0,044	29	0,226	0,075	42	0,338	0,043	66	0,513	0,047	26	0,650	0,058	8	0,791	0,034	12	0,940	0,026	225	0,70
	cross-support		0,096	0,062		0,250	0,086		0,369	0,085		0,547	0,064		0,679	0,072		0,807	0,033		0,946	0,027		
TopPI $k = 1$	all-confidence	18	0,053	0,041	7	0,218	0,067	9	0,341	0,033	5	0,515	0,064	4	0,679	0,072	2	0,762	0,040	3	0,935	0,040	48	0,49
	cross-support		0,053	0,041		0,218	0,067		0,341	0,033		0,515	0,064		0,679	0,072		0,762	0,040		0,935	0,040		
LAM	all-confidence	54	0,047	0,046	22	0,200	0,047	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	76	1,38
	cross-support		0,123	0,087		0,462	0,175		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	224	0,042	0,054	10	0,255	0,046	1	0,538	0,000	1	0,573	0,000	4	0,699	0,078	0	0,000	0,000	1	0,959	0,000	241	0,64
	cross-support		0,222	0,173		0,456	0,156		0,899	0,000		0,573	0,000		0,792	0,066		0,000	0,000		0,959	0,000		

Kddcup99	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,11]			(0,11 , 0,23]			(0,23 , 0,34]			(0,34 , 0,45]			(0,45 , 0,57]			(0,57 , 0,68]			(0,68 , 0,79]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	1.472	0,007	0,049	113	0,633	0,361	10	0,384	0,183	17	0,550	0,018	37	0,637	0,025	1	0,851	0,000	113	0,937	0,047	1.763	5,35
	cross-support		0,008	0,060		0,635	0,361		0,393	0,185		0,572	0,012		0,646	0,026		0,852	0,000		0,939	0,046		
TopPI $k = 3$	all-confidence	287	0,005	0,015	41	0,262	0,017	9	0,394	0,189	4	0,601	0,056	7	0,639	0,055	2	0,824	0,038	17	0,949	0,050	367	2,73
	cross-support		0,005	0,015		0,263	0,017		0,426	0,205		0,618	0,090		0,714	0,087		0,852	0,000		0,952	0,049		
LAM	all-confidence	454	0,004	0,015	17	0,241	0,014	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	471	25.453,68
	cross-support		0,052	0,103		0,407	0,124		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	821	0,004	0,028	18	0,480	0,326	0	0,000	0,000	4	0,540	0,031	2	0,626	0,020	0	0,000	0,000	5	0,976	0,041	850	29,63
	cross-support		0,046	0,103		0,563	0,288		0,000	0,000		0,721	0,175		0,640	0,000		0,000	0,000		0,984	0,025		

Mushrooms	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,83]			(0,83 , 0,97]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 20{,}00$	all-confidence	197	0,072	0,070	86	0,254	0,096	43	0,431	0,102	13	0,599	0,133	3	0,591	0,013	1	0,832	0,000	1	0,997	0,000	344	0,75
	cross-support		0,117	0,130		0,316	0,139		0,506	0,160		0,649	0,166		0,606	0,038		0,878	0,000		0,998	0,000		
TopPI $k = 3$	all-confidence	190	0,044	0,050	52	0,220	0,054	26	0,355	0,048	21	0,521	0,103	9	0,619	0,044	2	0,806	0,002	3	0,947	0,043	303	0,37
	cross-support		0,052	0,056		0,236	0,058		0,370	0,067		0,555	0,112		0,642	0,052		0,831	0,000		0,964	0,030		

Continua na próxima página.

LAM	all-confidence	86	0,033	0,027	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	86	1,39
	cross-support		0,232	0,166		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	412	0,035	0,063	3	0,367	0,040	0	0,000	0,000	1	0,795	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	416	1,85
	cross-support		0,256	0,209		0,834	0,174		0,000	0,000		0,955	0,000		0,000	0,000		0,000	0,000		0,000	0,000		

PowerC	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,13]			(0,13 , 0,27]			(0,27 , 0,40]			(0,40 , 0,53]			(0,53 , 0,66]			(0,66 , 0,80]			(0,80 , 0,93]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 2{,}00$	all-confidence	758	0,024	0,117	1	0,934	0,000	2	0,357	0,056	5	0,737	0,162	1	0,598	0,000	0	0,000	0,000	0	0,000	0,000	767	3,45
	cross-support		0,047	0,153		0,960	0,000		0,375	0,070		0,739	0,163		0,609	0,000		0,000	0,000		0,000	0,000		
TopPI $k = 9$	all-confidence	832	0,027	0,110	28	0,256	0,147	10	0,380	0,090	19	0,560	0,130	6	0,634	0,043	0	0,000	0,000	1	0,961	0,000	896	2,25
	cross-support		0,035	0,133		0,364	0,210		0,421	0,158		0,593	0,140		0,649	0,047		0,000	0,000		0,991	0,000		
LAM	all-confidence	1.964	0,001	0,004	0	0,000	0,000	1	0,279	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1.965	4.517,61
	cross-support		0,033	0,083		0,000	0,000		0,503	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	3.710	0,006	0,054	4	0,398	0,365	3	0,386	0,019	6	0,591	0,205	2	0,598	0,000	0	0,000	0,000	0	0,000	0,000	3.725	813,39
	cross-support		0,039	0,117		0,592	0,267		0,453	0,044		0,695	0,186		0,607	0,004		0,000	0,000		0,000	0,000		

Pumsb	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 0,99]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	1.135	0,122	0,240	69	0,371	0,175	33	0,606	0,180	31	0,780	0,166	4	0,843	0,154	5	0,971	0,030	4	1,000	0,000	1.281	190,07
	cross-support		0,156	0,279		0,387	0,181		0,620	0,178		0,788	0,160		0,845	0,156		0,975	0,033		1,000	0,000		
TopPI $k = 1$	all-confidence	1.855	0,007	0,041	18	0,290	0,113	7	0,483	0,163	12	0,774	0,225	2	0,770	0,241	3	0,964	0,040	4	0,964	0,073	1.901	1,20
	cross-support		0,007	0,041		0,290	0,113		0,483	0,163		0,774	0,225		0,770	0,241		0,964	0,040		0,964	0,073		
LAM	all-confidence	2.322	0,004	0,007	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2.322	91,15
	cross-support		0,098	0,146		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	6.206	0,038	0,119	49	0,368	0,198	21	0,596	0,209	18	0,755	0,179	9	0,724	0,133	14	0,868	0,069	8	0,946	0,041	6.325	12.073,94
	cross-support		0,156	0,226		0,553	0,217		0,658	0,189		0,821	0,136		0,818	0,092		0,914	0,042		0,968	0,045		

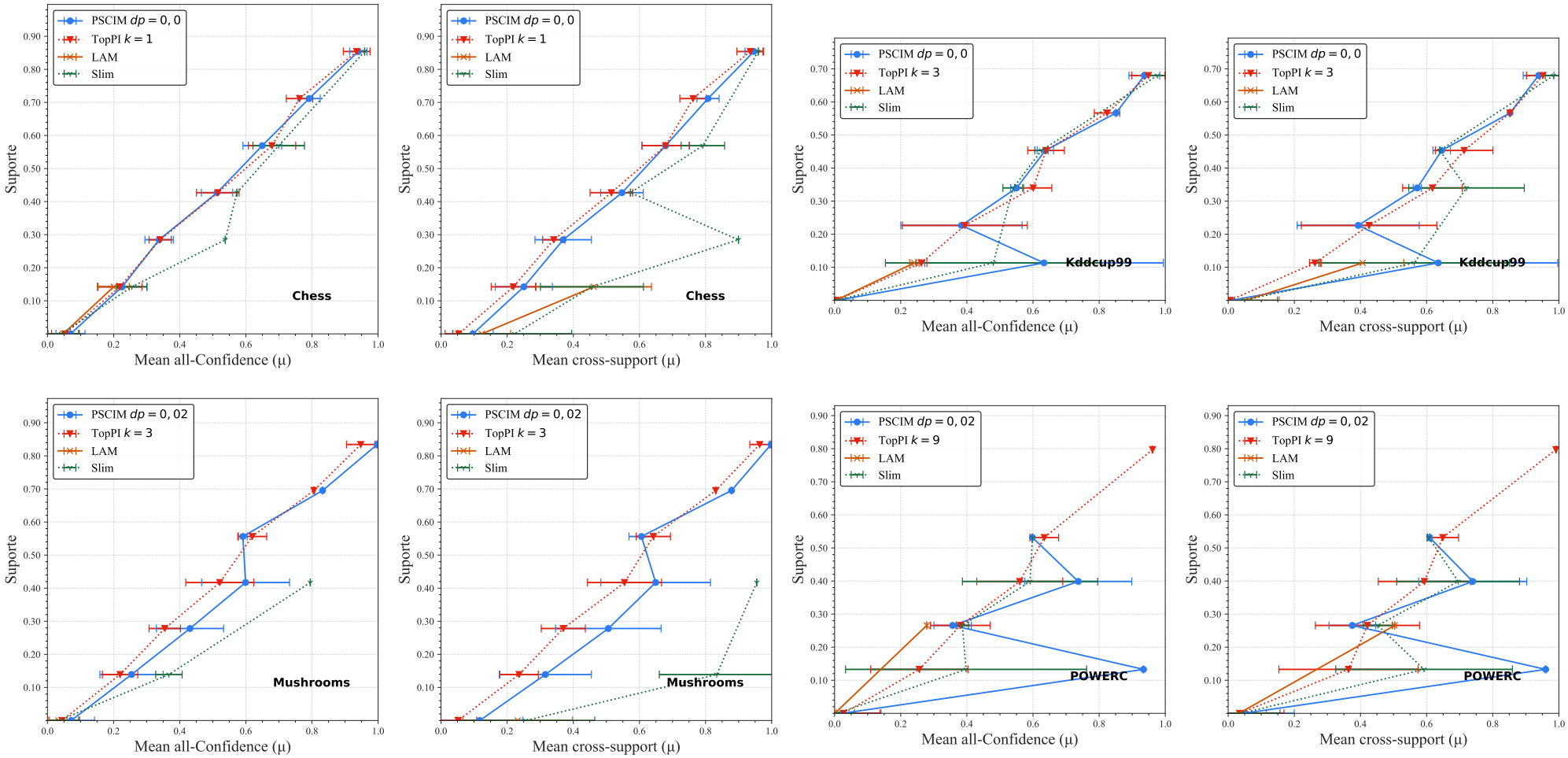
RecordLink	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,85]			(0,85 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	220	0,005	0,043	1	0,628	0,000	15	0,420	0,050	13	0,512	0,056	14	0,660	0,069	10	0,774	0,006	4	0,984	0,009	277	2,35
	cross-support		0,006	0,045		0,709	0,000		0,509	0,049		0,603	0,114		0,748	0,096		0,775	0,006		0,988	0,006		

Continua na próxima página.

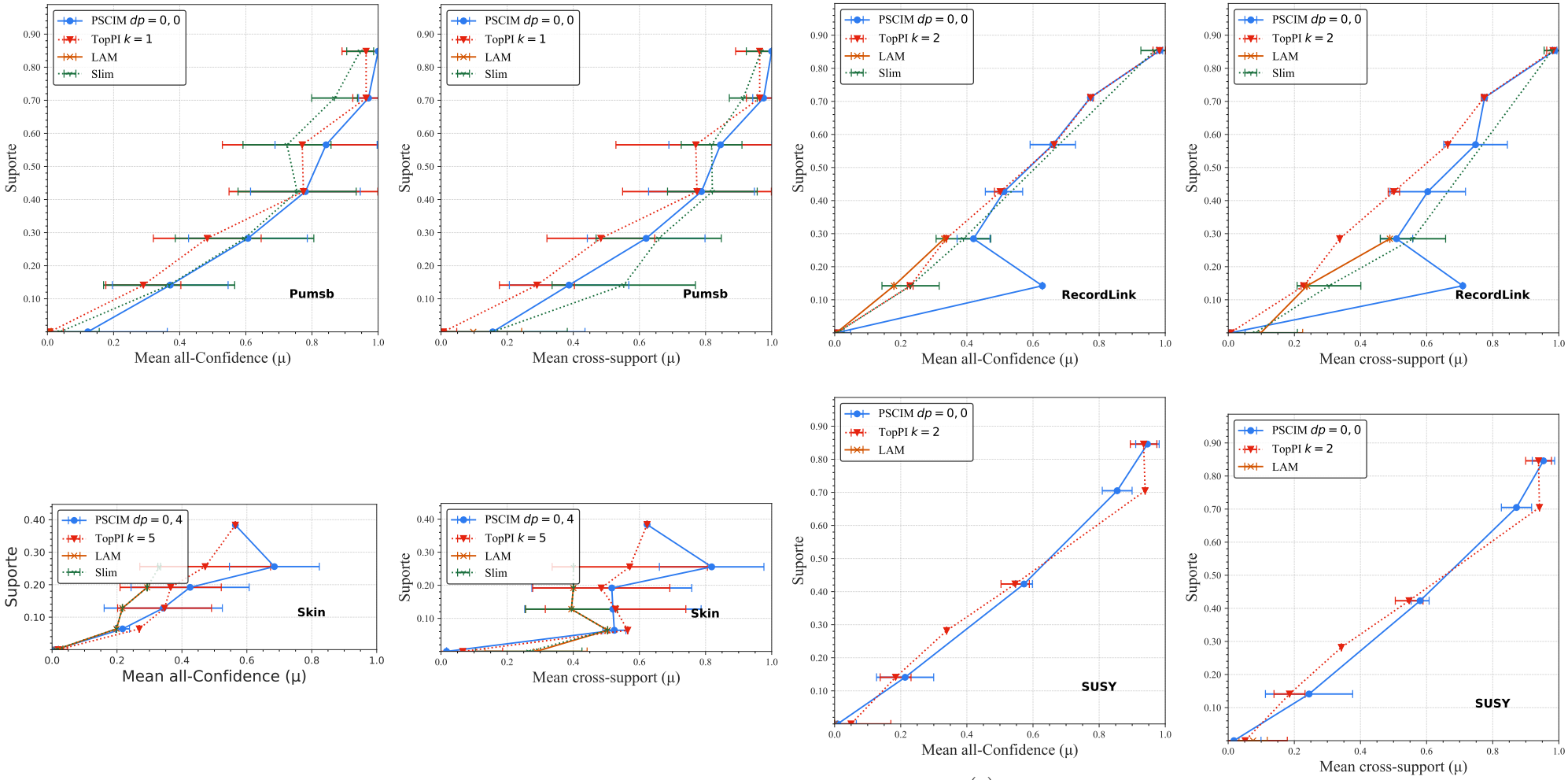
TopPI $k = 2$	all-confidence	16	0,006	0,012	3	0,229	0,009	1	0,337	0,000	2	0,501	0,017	1	0,664	0,000	3	0,774	0,009	4	0,982	0,019	30	0,95
	cross-support		0,006	0,012		0,229	0,009		0,337	0,000		0,501	0,017		0,664	0,000		0,774	0,009		0,983	0,019		
LAM	all-confidence	179	0,005	0,017	2	0,179	0,000	2	0,331	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	183	3.309,18
	cross-support		0,098	0,127		0,238	0,000		0,489	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	217	0,006	0,022	5	0,230	0,087	2	0,390	0,083	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2	0,967	0,041	226	3,34
	cross-support		0,082	0,127		0,305	0,096		0,559	0,099		0,000	0,000		0,000	0,000		0,000	0,000		0,980	0,024		

Skin	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,06]			(0,06 , 0,13]			(0,13 , 0,19]			(0,19 , 0,26]			(0,26 , 0,32]			(0,32 , 0,38]			(0,38 , 0,45]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}40$	all-confidence	2	0,010	0,010	3	0,217	0,021	6	0,342	0,182	6	0,425	0,182	3	0,685	0,138	0	0,000	0,000	1	0,565	0,000	21	0,56
	cross-support		0,017	0,000		0,524	0,035		0,520	0,267		0,516	0,242		0,818	0,158		0,000	0,000		0,623	0,000		
TopPI $k = 5$	all-confidence	6	0,019	0,028	1	0,269	0,000	9	0,347	0,145	9	0,365	0,156	7	0,472	0,201	0	0,000	0,000	1	0,565	0,000	33	0,79
	cross-support		0,066	0,120		0,565	0,000		0,528	0,212		0,485	0,207		0,571	0,235		0,000	0,000		0,623	0,000		
LAM	all-confidence	53	0,012	0,022	1	0,198	0,000	2	0,217	0,000	1	0,293	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	57	36,86
	cross-support		0,287	0,154		0,504	0,000		0,394	0,139		0,400	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	58	0,011	0,021	1	0,198	0,000	2	0,217	0,000	1	0,293	0,000	2	0,330	0,004	0	0,000	0,000	0	0,000	0,000	64	0,42
	cross-support		0,265	0,161		0,504	0,000		0,394	0,139		0,400	0,000		0,400	0,000		0,000	0,000		0,000	0,000		

Susy	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,28]			(0,28 , 0,42]			(0,42 , 0,56]			(0,56 , 0,70]			(0,70 , 0,85]			(0,85 , 0,99]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	2.702	0,009	0,057	24	0,213	0,086	0	0,000	0,000	8	0,573	0,025	0	0,000	0,000	9	0,855	0,045	18	0,946	0,035	2.761	252,10
	cross-support		0,017	0,082		0,245	0,132		0,000	0,000		0,580	0,027		0,000	0,000		0,872	0,046		0,953	0,034		
TopPI $k = 2$	all-confidence	190	0,050	0,120	9	0,185	0,047	1	0,339	0,000	2	0,546	0,043	0	0,000	0,000	1	0,939	0,000	8	0,934	0,040	211	12,03
	cross-support		0,051	0,127		0,186	0,047		0,342	0,000		0,547	0,042		0,000	0,000		0,941	0,000		0,939	0,039		
LAM	all-confidence	41.532	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	41.532	986,11
	cross-support		0,075	0,043		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		



Continua na próxima página.



(g)

Figura C.1: Distribuições dos valores médios de *All-confidence* dos itemsets fechados recuperados pelo *PSCIM*, *TopPI* e *LAM* sobre as bases de dados densa da Tabela C.3, neste estudo é usando o melhor valor de parâmetro para cada técnica.

Tabela C.4: Desempenho do algoritmo/parametrização das técnicas *PSCIM*, *TopPI* e *LAM* sobre as bases de dados esparsas da Tabela C.2.

Accidents	***Métrica***	**Partição de suporte**																					Itemset #	Tempo (s)
		[0,00 , 0,14]			(0,14 , 0,29]			(0,29 , 0,43]			(0,43 , 0,57]			(0,57 , 0,71]			(0,71 , 0,86]			(0,86 , 1,00]				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	2.222	0,026	0,073	104	0,221	0,115	2	0,411	0,111	151	0,520	0,041	77	0,624	0,050	12	0,785	0,042	8	0,956	0,056	2.576	111,12
	cross-support		0,044	0,114		0,237	0,141		0,463	0,137		0,632	0,045		0,649	0,055		0,792	0,053		0,956	0,056		
TopPI $k = 2$	all-confidence	634	0,011	0,027	33	0,218	0,069	4	0,383	0,047	6	0,494	0,026	7	0,650	0,043	7	0,786	0,036	7	0,933	0,059	698	1,85
	cross-support		0,012	0,027		0,218	0,069		0,383	0,047		0,494	0,026		0,650	0,043		0,786	0,036		0,933	0,059		
LAM	all-confidence	9.662	0,001	0,001	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	9.662	59,63
	cross-support		0,073	0,078		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	13.013	0,012	0,045	86	0,267	0,129	41	0,398	0,082	10	0,525	0,065	14	0,702	0,086	8	0,825	0,054	3	0,923	0,071	13.175	128.565,67
	cross-support		0,115	0,136		0,497	0,206		0,564	0,118		0,637	0,124		0,863	0,049		0,885	0,042		0,937	0,059		

BMSWebView2 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset #	Tempo (s)
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,08] †			(0,08 , 0,11] †			(0,11 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,19] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}15$	all-confidence	4.163	0,165	0,124	47	0,343	0,102	12	0,380	0,097	5	0,356	0,046	1	0,414	0,000	1	0,309	0,000	2	0,492	0,075	4.231	241,63
	cross-support		0,505	0,233		0,688	0,153		0,731	0,133		0,823	0,119		0,734	0,000		0,663	0,000		0,816	0,120		
TopPI $k = 2$	all-confidence	2.778	0,178	0,149	50	0,306	0,118	17	0,346	0,112	4	0,284	0,072	3	0,324	0,078	2	0,330	0,017	2	0,492	0,075	2.856	0,79
	cross-support		0,397	0,296		0,622	0,190		0,669	0,232		0,700	0,223		0,642	0,080		0,783	0,176		0,816	0,120		
LAM	all-confidence	6.203	0,025	0,044	10	0,196	0,090	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	6.213	2,28
	cross-support		0,504	0,272		0,740	0,172		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	4.814	0,173	0,126	235	0,230	0,117	66	0,246	0,107	26	0,262	0,073	12	0,309	0,063	3	0,323	0,017	3	0,454	0,085	5.159	9.119,22
	cross-support		0,551	0,245		0,616	0,185		0,645	0,176		0,702	0,181		0,666	0,124		0,743	0,142		0,763	0,124		

BMS1 τ	***Métrica***	**Partição de suporte $\dagger \times 10^{-1}$**																					Itemset #	Tempo (s)
		[0,00 , 0,03] †			(0,03 , 0,06] †			(0,06 , 0,09] †			(0,09 , 0,12] †			(0,12 , 0,14] †			(0,14 , 0,17] †			(0,17 , 0,20] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}01$	all-confidence	363	0,140	0,143	24	0,264	0,090	3	0,235	0,070	3	0,210	0,091	2	0,260	0,072	2	0,357	0,042	1	0,329	0,000	398	3,34
	cross-support		0,524	0,261		0,659	0,227		0,522	0,256		0,402	0,288		0,715	0,187		0,811	0,131		0,987	0,000		

Continua na próxima página.

TopPI $k = 2$	all-confidence	356	0,172	0,147	39	0,240	0,099	9	0,171	0,064	5	0,240	0,077	1	0,224	0,000	2	0,357	0,042	1	0,329	0,000	413	0,49
	cross-support		0,418	0,312		0,634	0,249		0,439	0,194		0,579	0,317		0,687	0,000		0,811	0,131		0,987	0,000		
LAM	all-confidence	2.685	0,012	0,028	8	0,127	0,065	1	0,181	0,000	1	0,143	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	2.695	1,77
	cross-support		0,476	0,271		0,633	0,264		0,822	0,000		0,987	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	836	0,140	0,127	86	0,182	0,096	17	0,167	0,055	8	0,241	0,059	3	0,248	0,055	2	0,357	0,042	1	0,329	0,000	953	45,57
	cross-support		0,519	0,255		0,605	0,232		0,556	0,232		0,690	0,287		0,706	0,133		0,811	0,131		0,987	0,000		

FoodmartFIM τ	***Métrica***	**Partição de suporte † × 10^{-3}**																					Itemset #	Tempo (s)
		[0,00 , 0,14] †			(0,14 , 0,28] †			(0,28 , 0,41] †			(0,41 , 0,55] †			(0,55 , 0,69] †			(0,69 , 0,83] †			(0,83 , 0,97] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}03$	all-confidence	0	0,000	0,000	951	0,078	0,017	0	0,000	0,000	452	0,142	0,028	0	0,000	0,000	50	0,190	0,039	2	0,211	0,000	1.455	12,53
	cross-support		0,000	0,000		0,656	0,218		0,000	0,000		0,745	0,163		0,000	0,000		0,754	0,147		0,658	0,112		
TopPI $k = 2$	all-confidence	0	0,000	0,000	247	0,051	0,006	0	0,000	0,000	709	0,133	0,028	0	0,000	0,000	69	0,189	0,038	2	0,211	0,000	1.027	0,51
	cross-support		0,000	0,000		0,399	0,111		0,000	0,000		0,717	0,162		0,000	0,000		0,772	0,145		0,658	0,112		
LAM	all-confidence	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	266	0,133	0,028	0	0,000	0,000	6	0,210	0,065	0	0,000	0,000	272	0,58
	cross-support		0,000	0,000		0,000	0,000		0,000	0,000		0,747	0,159		0,000	0,000		0,798	0,166		0,000	0,000		

Fruithut τ	***Métrica***	**Partição de suporte † × 10^{-1}**																					Itemset #	Tempo (s)
		[0,00 , 0,05] †			(0,05 , 0,10] †			(0,10 , 0,15] †			(0,15 , 0,20] †			(0,20 , 0,25] †			(0,25 , 0,30] †			(0,30 , 0,35] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	398	0,033	0,035	15	0,088	0,053	4	0,077	0,056	1	0,080	0,000	2	0,089	0,008	0	0,000	0,000	1	0,146	0,000	421	22,08
	cross-support		0,330	0,297		0,512	0,408		0,217	0,177		0,173	0,000		0,212	0,035		0,000	0,000		0,470	0,000		
TopPI $k = 7$	all-confidence	6.957	0,007	0,018	99	0,076	0,039	23	0,068	0,036	6	0,074	0,007	3	0,087	0,007	0	0,000	0,000	1	0,146	0,000	7.089	0,83
	cross-support		0,050	0,128		0,468	0,307		0,257	0,179		0,223	0,043		0,230	0,039		0,000	0,000		0,470	0,000		
LAM	all-confidence	11.171	0,002	0,005	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	11.171	3,50
	cross-support		0,286	0,274		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	2.032	0,030	0,026	67	0,088	0,039	12	0,082	0,044	3	0,079	0,001	2	0,089	0,008	0	0,000	0,000	1	0,146	0,000	2.117	658,36
	cross-support		0,435	0,277		0,551	0,292		0,311	0,222		0,237	0,061		0,212	0,035		0,000	0,000		0,470	0,000		

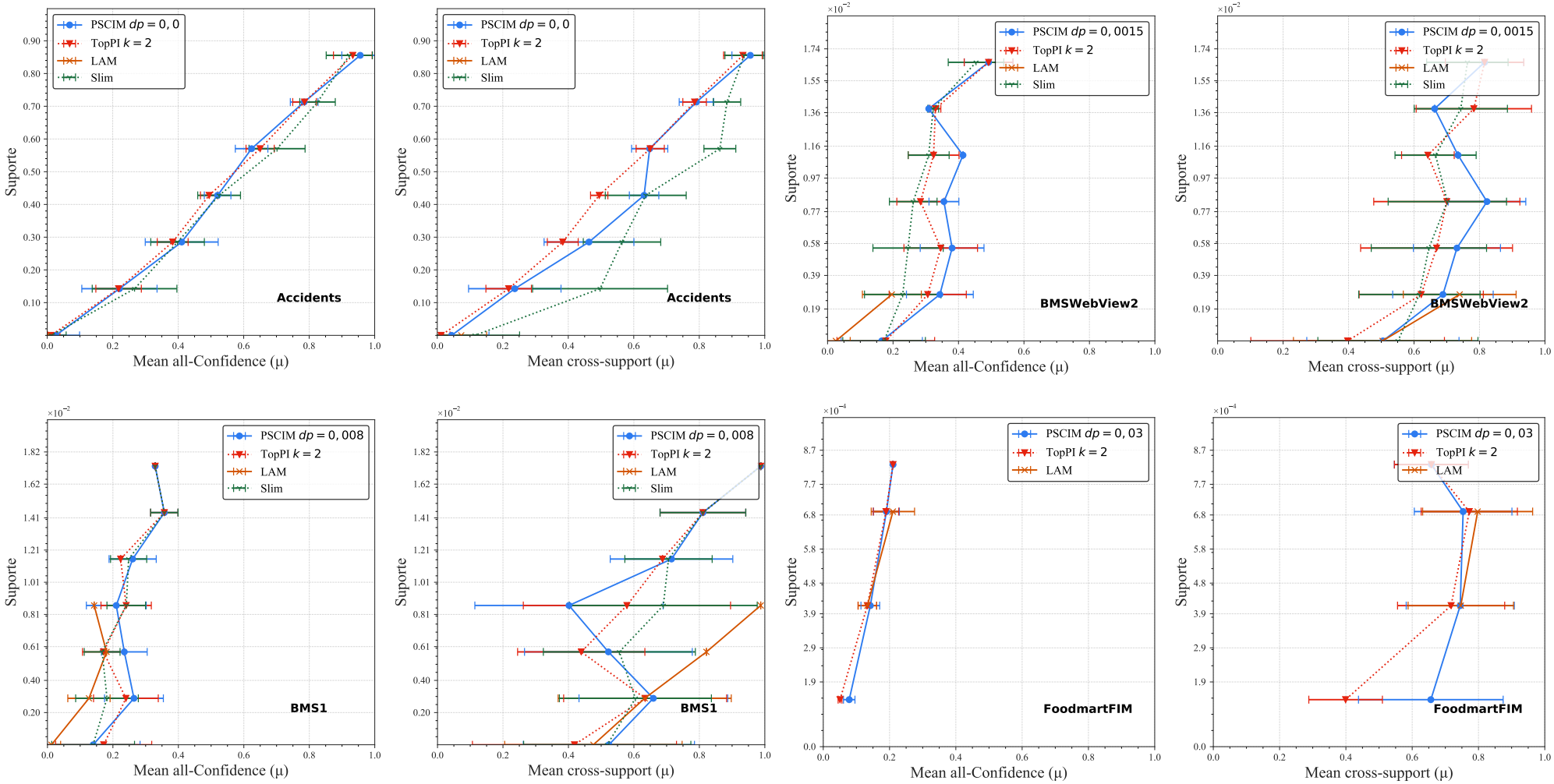
OnlineRetail τ	***Métrica***	**Partição de suporte † × 10^{-1}**																					Itemset #	Tempo (s)
		[0,00 , 0,08] †			(0,08 , 0,15] †			(0,15 , 0,23] †			(0,23 , 0,31] †			(0,31 , 0,39] †			(0,39 , 0,46] †			(0,46 , 0,54] †				
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}01$	all-confidence	1.065	0,429	0,319	6	0,551	0,250	3	0,650	0,181	1	0,726	0,000	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	1.078	15,69
	cross-support		0,475	0,327		0,603	0,232		0,769	0,183		0,816	0,000		0,613	0,291		0,000	0,000		0,980	0,000		

Continua na próxima página.

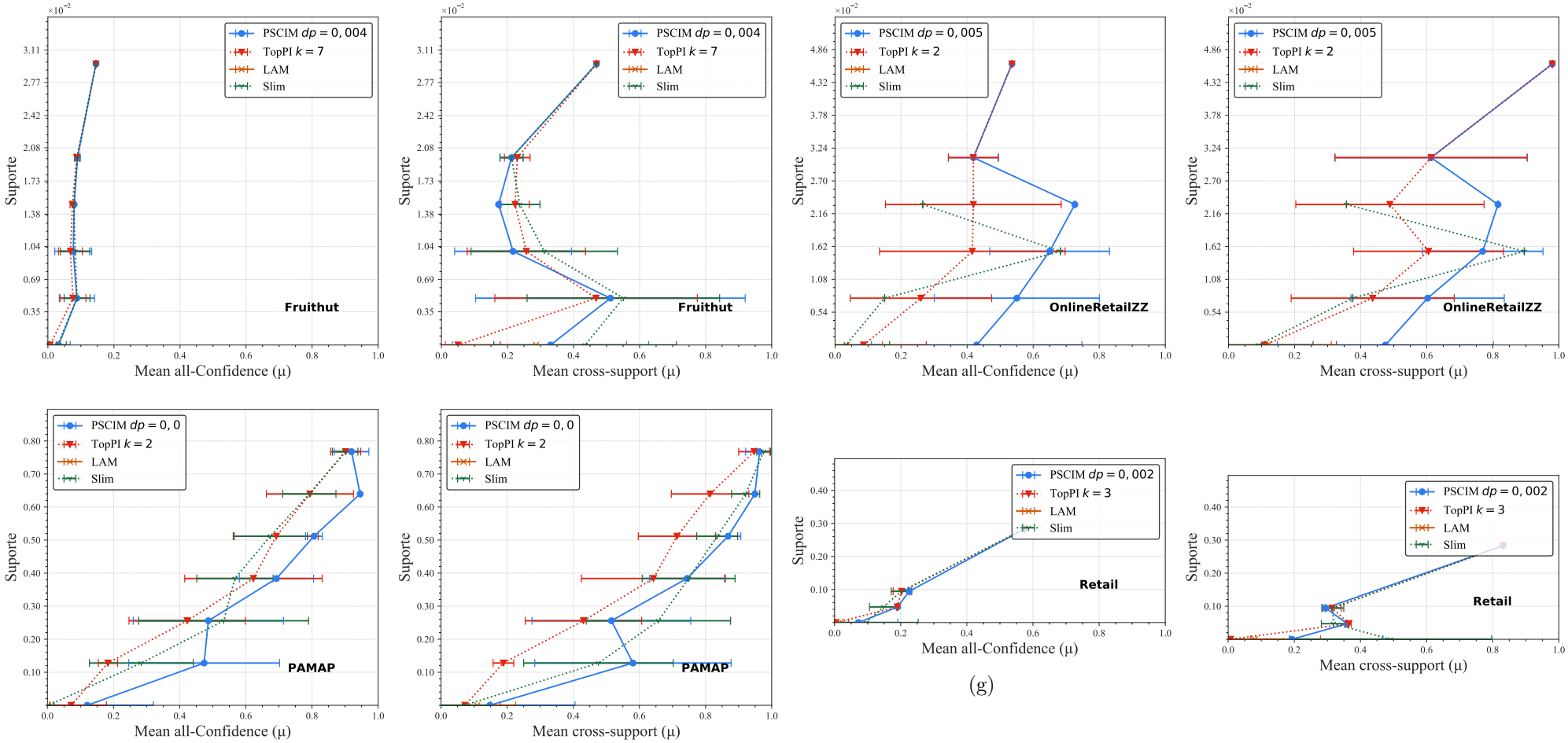
TopPI $k = 2$	all-confidence	2.335	0,087	0,190	24	0,260	0,214	6	0,416	0,281	3	0,419	0,266	2	0,419	0,076	0	0,000	0,000	1	0,536	0,000	2.371	1,08
	cross-support		0,109	0,217		0,436	0,247		0,605	0,227		0,489	0,285		0,613	0,291		0,000	0,000		0,980	0,000		
LAM	all-confidence	3.019	0,035	0,109	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	3.019	8,13
	cross-support		0,119	0,192		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	3.962	0,033	0,132	1	0,150	0,000	1	0,682	0,000	1	0,266	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	3.965	8.484,03
	cross-support		0,090	0,166		0,376	0,000		0,895	0,000		0,357	0,000		0,000	0,000		0,000	0,000		0,000	0,000		

PAMP τ	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,13]			(0,13 , 0,26]			(0,26 , 0,38]			(0,38 , 0,51]			(0,51 , 0,64]			(0,64 , 0,77]			(0,77 , 0,90]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	74	0,120	0,200	6	0,474	0,228	11	0,487	0,227	6	0,693	0,112	5	0,806	0,025	1	0,946	0,000	2	0,920	0,052	105	8,99
	cross-support		0,148	0,258		0,580	0,297		0,515	0,240		0,744	0,113		0,868	0,038		0,949	0,000		0,963	0,042		
TopPI $k = 2$	all-confidence	49	0,071	0,107	4	0,183	0,029	9	0,423	0,176	3	0,623	0,208	5	0,692	0,127	3	0,794	0,132	11	0,902	0,046	84	2,84
	cross-support		0,072	0,107		0,189	0,031		0,431	0,176		0,642	0,218		0,714	0,117		0,814	0,117		0,947	0,047		
LAM	all-confidence	3.610	0,002	0,004	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	3.610	2.333,96
	cross-support		0,109	0,116		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	8.717	0,003	0,017	25	0,284	0,157	8	0,533	0,257	7	0,567	0,116	11	0,674	0,112	9	0,792	0,081	8	0,900	0,039	8.785	14.237,68
	cross-support		0,071	0,079		0,476	0,226		0,657	0,218		0,749	0,140		0,835	0,062		0,921	0,043		0,976	0,021		

Retail τ	***Métrica***	**Partição de suporte**																					Itemset	Tempo
		[0,00 , 0,05]			(0,05 , 0,09]			(0,09 , 0,14]			(0,14 , 0,19]			(0,19 , 0,24]			(0,24 , 0,28]			(0,28 , 0,33]			#	(s)
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ		
PLSCIM $dr = 0{,}00$	all-confidence	9.729	0,072	0,119	1	0,191	0,000	1	0,225	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	9.732	8.353,96
	cross-support		0,193	0,275		0,360	0,000		0,295	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000		
TopPI $k = 3$	all-confidence	30.563	0,005	0,038	2	0,190	0,002	4	0,203	0,025	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	30.570	2,49
	cross-support		0,007	0,061		0,365	0,007		0,314	0,028		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000		
LAM	all-confidence	7.960	0,003	0,021	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	7.960	3,13
	cross-support		0,080	0,199		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		0,000	0,000		
Slim	all-confidence	6.557	0,096	0,157	3	0,148	0,042	3	0,202	0,031	0	0,000	0,000	0	0,000	0,000	0	0,000	0,000	1	0,575	0,000	6.564	103.297,54
	cross-support		0,491	0,306		0,318	0,036		0,316	0,033		0,000	0,000		0,000	0,000		0,000	0,000		0,831	0,000		



Continua na próxima página.



(g)

Figura C.2: Distribuições dos valores médios de *All-confidence* dos itemsets fechados recuperados pelo *PSCIM*, *TopPI* e *LAM* sobre as bases de dados esparsa da Tabela C.4, neste estudo é usando o melhor valor de parâmetro para cada algoritmo.

Tabela C.5: Significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo PSCIM com os algoritmos Slim, LAM e TopPI para as bases de dados densas da Tabela C.1. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas na Tabela C.3.

			Partição de suporte						
Chess	***Métrica***	H_0	[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,43]	(0,43 , 0,57]	(0,57 , 0,71]	(0,71 , 0,85]	(0,85 , 1,00]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,115	0,484	0,581	0,576	0,836	-	0,399
		$\mu_{PSCIM} = \mu$	0,229	0,968	0,839	0,866	0,360	-	0,798
		$\mu_{PSCIM} \geq \mu$	0,888	0,532	0,419	0,433	0,180	-	0,601
	cross-support	$\mu_{PSCIM} \leq \mu$	0,006	0,172	0,053	0,137	0,402	-	0,386
		$\mu_{PSCIM} = \mu$	0,011	0,345	0,106	0,274	0,804	-	0,772
		$\mu_{PSCIM} \geq \mu$	0,995	0,838	0,947	0,868	0,622	-	0,668
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,002	0,082	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,004	0,164	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,998	0,918	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,954	1,000	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,091	0,000	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,046	0,000	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,976	-	-	0,946	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,052	-	-	0,123	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,026	-	-	0,061	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	1,000	-	-	0,994	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,000	-	-	0,015	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	0,000	-	-	0,007	-	-

			Partição de suporte						
Kddcup99	***Métrica***	H_0	[0,00 , 0,11]	(0,11 , 0,23]	(0,23 , 0,34]	(0,34 , 0,45]	(0,45 , 0,57]	(0,57 , 0,68]	(0,68 , 0,79]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,244	0,000	0,548	0,915	0,882	-	0,843
		$\mu_{PSCIM} = \mu$	0,487	0,000	0,905	0,170	0,248	-	0,314
		$\mu_{PSCIM} \geq \mu$	0,756	1,000	0,452	0,085	0,124	-	0,157
	cross-support	$\mu_{PSCIM} \leq \mu$	0,021	0,000	0,641	0,808	0,998	-	0,858
		$\mu_{PSCIM} = \mu$	0,043	0,000	0,717	0,384	0,005	-	0,284
		$\mu_{PSCIM} \geq \mu$	0,979	1,000	0,359	0,192	0,002	-	0,142
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,013	0,000	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,027	0,000	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,987	1,000	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	0,000	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,000	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	1,000	-	-	-	-	-

Continua na próxima página.

Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,038	0,040	-	0,060	-	-	0,967
		$\mu_{PSCIM} = \mu$	0,075	0,081	-	0,120	-	-	0,066
		$\mu_{PSCIM} \geq \mu$	0,962	0,960	-	0,951	-	-	0,033
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	0,173	-	0,976	-	-	0,994
		$\mu_{PSCIM} = \mu$	0,000	0,345	-	0,061	-	-	0,012
		$\mu_{PSCIM} \geq \mu$	0,000	0,827	-	0,030	-	-	0,006

			Partição de suporte						
Mushrooms	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,42]	(0,42 , 0,56]	(0,56 , 0,70]	(0,70 , 0,83]	(0,83 , 0,97]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,009	0,000	0,031	0,823	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,018	0,000	0,063	0,458	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,991	1,000	0,971	0,229	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,000	0,000	0,036	0,868	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,000	0,000	0,073	0,353	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	1,000	1,000	0,966	0,176	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,990	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,021	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,011	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	0,998	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,005	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	0,002	-	-	-	-	-

			Partição de suporte						
PowerC	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,13]	(0,13 , 0,27]	(0,27 , 0,40]	(0,40 , 0,53]	(0,53 , 0,66]	(0,66 , 0,80]	(0,80 , 0,93]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,725	-	-	0,009	-	-	-
		$\mu_{PSCIM} = \mu$	0,550	-	-	0,019	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,275	-	-	0,992	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,053	-	-	0,031	-	-	-
		$\mu_{PSCIM} = \mu$	0,106	-	-	0,062	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,947	-	-	0,974	-	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,007	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,014	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,993	-	-	-	-	-	-

Continua na próxima página.

Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	0,050	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	0,100	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	0,966	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,099	-	-	0,291	-	-	-
		$\mu_{PSCIM} = \mu$	0,199	-	-	0,581	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,901	-	-	0,769	-	-	-

			Partição de suporte						
Pumsb	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,42]	(0,42 , 0,57]	(0,57 , 0,71]	(0,71 , 0,85]	(0,85 , 0,99]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,033	0,051	0,467	-	0,676	0,227
		$\mu_{PSCIM} = \mu$	0,000	0,065	0,102	0,934	-	0,879	0,453
		$\mu_{PSCIM} \geq \mu$	1,000	0,967	0,953	0,533	-	0,440	0,894
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,016	0,040	0,425	-	0,324	0,227
		$\mu_{PSCIM} = \mu$	0,000	0,033	0,080	0,851	-	0,649	0,453
		$\mu_{PSCIM} \geq \mu$	1,000	0,984	0,963	0,575	-	0,776	0,894
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,462	0,479	0,291	0,038	0,006	0,009
		$\mu_{PSCIM} = \mu$	0,000	0,923	0,958	0,582	0,076	0,012	0,017
		$\mu_{PSCIM} \geq \mu$	1,000	0,538	0,528	0,716	0,975	0,995	0,995
	cross-support	$\mu_{PSCIM} \leq \mu$	0,497	1,000	0,834	0,638	0,243	0,005	0,009
		$\mu_{PSCIM} = \mu$	0,993	0,000	0,342	0,739	0,485	0,010	0,017
		$\mu_{PSCIM} \geq \mu$	0,503	0,000	0,171	0,370	0,803	0,996	0,995

			Partição de suporte						
RecordLink	***Métrica***	$\boldsymbol{H_0}$	[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,43]	(0,43 , 0,57]	(0,57 , 0,71]	(0,71 , 0,85]	(0,85 , 1,00]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,535	-	-	-	-	0,751	0,669
		$\mu_{PSCIM} = \mu$	0,930	-	-	-	-	0,611	0,884
		$\mu_{PSCIM} \geq \mu$	0,465	-	-	-	-	0,305	0,442
	cross-support	$\mu_{PSCIM} \leq \mu$	0,489	-	-	-	-	0,534	0,617
		$\mu_{PSCIM} = \mu$	0,977	-	-	-	-	1,000	1,000
		$\mu_{PSCIM} \geq \mu$	0,511	-	-	-	-	0,534	0,500
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,505	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,990	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,495	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-

Continua na próxima página.

Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,689	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,623	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,311	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-

Skin	***Métrica***	H_0	**Partição de suporte**						
			[0,00 , 0,06]	(0,06 , 0,13]	(0,13 , 0,19]	(0,19 , 0,26]	(0,26 , 0,32]	(0,32 , 0,38]	(0,38 , 0,45]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	-	-	0,519	0,384	0,069	-	-
		$\mu_{PSCIM} = \mu$	-	-	0,962	0,767	0,138	-	-
		$\mu_{PSCIM} \geq \mu$	-	-	0,481	0,661	0,931	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	-	-	0,595	0,571	0,082	-	-
		$\mu_{PSCIM} = \mu$	-	-	0,905	0,952	0,164	-	-
		$\mu_{PSCIM} \geq \mu$	-	-	0,452	0,476	0,948	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	-	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	-	-	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	-	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	-	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	-	-	-	-	-	-	-

Susy	***Métrica***	H_0	**Partição de suporte**						
			[0,00 , 0,14]	(0,14 , 0,28]	(0,28 , 0,42]	(0,42 , 0,56]	(0,56 , 0,70]	(0,70 , 0,85]	(0,85 , 0,99]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	1,000	0,177	-	-	-	-	0,289
		$\mu_{PSCIM} = \mu$	0,000	0,354	-	-	-	-	0,578
		$\mu_{PSCIM} \geq \mu$	0,000	0,823	-	-	-	-	0,730
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	0,102	-	-	-	-	0,165
		$\mu_{PSCIM} = \mu$	0,000	0,203	-	-	-	-	0,329
		$\mu_{PSCIM} \geq \mu$	0,000	0,898	-	-	-	-	0,849
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-

Tabela C.6: Significâncias estatísticas das médias de distribuições de *all-confidence* e *cross-support* das partições de suporte comparando o algoritmo PSCIM com os algoritmos Slim, LAM e TopPI para as bases de dados esparsas da Tabela C.2. Os valores de média e desvio padrão das amostras de cada partição de suporte podem ser encontradas na Tabela C.4.

			Partição de suporte						
Accidents	***Métrica***	H_0	[0,00 , 0,14]	(0,14 , 0,29]	(0,29 , 0,43]	(0,43 , 0,57]	(0,57 , 0,71]	(0,71 , 0,86]	(0,86 , 1,00]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,452	-	0,028	0,971	0,789	0,261
		$\mu_{PSCIM} = \mu$	0,000	0,904	-	0,055	0,060	0,472	0,522
		$\mu_{PSCIM} \geq \mu$	1,000	0,548	-	0,973	0,030	0,236	0,775
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,232	-	0,000	0,659	0,634	0,207
		$\mu_{PSCIM} = \mu$	0,000	0,465	-	0,000	0,693	0,798	0,414
		$\mu_{PSCIM} \geq \mu$	1,000	0,768	-	1,000	0,347	0,399	0,825
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,995	-	0,357	1,000	0,965	0,153
		$\mu_{PSCIM} = \mu$	0,000	0,011	-	0,713	0,001	0,082	0,306
		$\mu_{PSCIM} \geq \mu$	1,000	0,005	-	0,646	0,000	0,041	0,890
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	1,000	-	0,072	1,000	0,999	0,268
		$\mu_{PSCIM} = \mu$	0,000	0,000	-	0,145	0,000	0,002	0,536
		$\mu_{PSCIM} \geq \mu$	0,000	0,000	-	0,929	0,000	0,001	0,796

			Partição de suporte † $\times 10^{-1}$						
BMSWebView2	***Métrica***	H_0	[0,00 , 0,03] †	(0,03 , 0,06] †	(0,06 , 0,08] †	(0,08 , 0,11] †	(0,11 , 0,14] †	(0,14 , 0,17] †	(0,17 , 0,19] †
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	1,000	0,113	0,213	0,133	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,225	0,425	0,266	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	0,889	0,800	0,913	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,078	0,200	0,194	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,156	0,400	0,387	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,923	0,812	0,867	-	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,000	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,000	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	1,000	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,414	0,769	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,827	0,476	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,586	0,238	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,996	0,000	0,000	0,009	-	-	-
		$\mu_{PSCIM} = \mu$	0,007	0,000	0,000	0,018	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,004	1,000	1,000	0,992	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	0,007	0,048	0,073	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,013	0,096	0,147	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	0,993	0,953	0,934	-	-	-

Continua na próxima página.

BMS1	***Métrica***	H_0	**Partição de suporte † × 10^{-1}**						
			[0,00 , 0,03] †	(0,03 , 0,06] †	(0,06 , 0,09] †	(0,09 , 0,12] †	(0,12 , 0,14] †	(0,14 , 0,17] †	(0,17 , 0,20] †
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,998	0,146	0,082	0,676	-	-	-
		$\mu_{PSCIM} = \mu$	0,003	0,292	0,163	0,879	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,002	0,857	0,943	0,440	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,402	0,258	0,856	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,804	0,515	0,448	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,603	0,799	0,224	-	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,000	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,001	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	1,000	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,001	0,517	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,002	1,000	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,999	0,500	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,490	0,000	0,050	0,731	-	-	-
		$\mu_{PSCIM} = \mu$	0,980	0,000	0,100	0,681	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,510	1,000	0,960	0,341	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,378	0,156	0,563	0,950	-	-	-
		$\mu_{PSCIM} = \mu$	0,757	0,313	0,958	0,150	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,622	0,845	0,479	0,075	-	-	-

FoodmartFIM	***Métrica***	H_0	**Partição de suporte † × 10^{-3}**						
			[0,00 , 0,14] †	(0,14 , 0,28] †	(0,28 , 0,41] †	(0,41 , 0,55] †	(0,55 , 0,69] †	(0,69 , 0,83] †	(0,83 , 0,97] †
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	-	0,000	-	0,000	-	0,438	-
		$\mu_{PSCIM} = \mu$	-	0,000	-	0,000	-	0,876	-
		$\mu_{PSCIM} \geq \mu$	-	1,000	-	1,000	-	0,562	-
	cross-support	$\mu_{PSCIM} \leq \mu$	-	0,000	-	0,002	-	0,752	-
		$\mu_{PSCIM} = \mu$	-	0,000	-	0,005	-	0,496	-
		$\mu_{PSCIM} \geq \mu$	-	1,000	-	0,998	-	0,248	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	-	-	-	0,000	-	0,739	-
		$\mu_{PSCIM} = \mu$	-	-	-	0,000	-	0,539	-
		$\mu_{PSCIM} \geq \mu$	-	-	-	1,000	-	0,270	-
	cross-support	$\mu_{PSCIM} \leq \mu$	-	-	-	0,553	-	0,751	-
		$\mu_{PSCIM} = \mu$	-	-	-	0,894	-	0,499	-
		$\mu_{PSCIM} \geq \mu$	-	-	-	0,447	-	0,249	-

Fruithut	***Métrica***	H_0	**Partição de suporte † × 10^{-1}**						
			[0,00 , 0,05] †	(0,05 , 0,10] †	(0,10 , 0,15] †	(0,15 , 0,20] †	(0,20 , 0,25] †	(0,25 , 0,30] †	(0,30 , 0,35] †
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,210	0,347	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,421	0,694	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,790	0,653	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,349	0,659	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,698	0,682	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,651	0,341	-	-	-	-

Continua na próxima página.

LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,002	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,004	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,998	-	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,049	0,516	0,580	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,098	0,968	0,841	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,951	0,484	0,420	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	1,000	0,636	0,772	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,729	0,455	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	0,364	0,228	-	-	-	-

OnlineRetail	***Métrica***	H_0	**Partição de suporte † $\times 10^{-1}$**						
			[0,00 , 0,08] †	(0,08 , 0,15] †	(0,15 , 0,23] †	(0,23 , 0,31] †	(0,31 , 0,39] †	(0,39 , 0,46] †	(0,46 , 0,54] †
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	0,004	0,148	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,008	0,296	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,997	0,904	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	0,070	0,216	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	0,139	0,433	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	0,937	0,852	-	-	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
Slim	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-

PAMP	***Métrica***	H_0	**Partição de suporte**						
			[0,00 , 0,13]	(0,13 , 0,26]	(0,26 , 0,38]	(0,38 , 0,51]	(0,51 , 0,64]	(0,64 , 0,77]	(0,77 , 0,90]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
TopPI	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,041	0,057	0,249	0,302	0,147	-	-
		$\mu_{PSCIM} = \mu$	0,081	0,114	0,499	0,604	0,293	-	-
		$\mu_{PSCIM} \geq \mu$	0,959	0,967	0,751	0,782	0,896	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,014	0,021	0,196	0,299	0,018	-	-
		$\mu_{PSCIM} = \mu$	0,027	0,042	0,392	0,598	0,035	-	-
		$\mu_{PSCIM} \geq \mu$	0,986	0,988	0,804	0,786	0,990	-	-
LAM	*all-confidence*	$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} \leq \mu$	0,104	-	-	-	-	-	-
		$\mu_{PSCIM} = \mu$	0,208	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,896	-	-	-	-	-	-

Continua na próxima página.

	$\mu_{PSCIM} \leq \mu$	0,000	0,011	0,659	0,031	0,035	-	-
	all-confidence $\mu_{PSCIM} = \mu$	0,000	0,021	0,683	0,063	0,069	-	-
Slim	$\mu_{PSCIM} \geq \mu$	1,000	0,989	0,341	0,977	0,973	-	-
	$\mu_{PSCIM} \leq \mu$	0,006	0,204	0,972	0,614	0,196	-	-
	cross-support $\mu_{PSCIM} = \mu$	0,012	0,409	0,069	0,885	0,392	-	-
	$\mu_{PSCIM} \geq \mu$	0,994	0,809	0,034	0,442	0,834	-	-

			Partição de suporte						
Retail	***Métrica***	H_0	[0,00 , 0,05]	(0,05 , 0,09]	(0,09 , 0,14]	(0,14 , 0,19]	(0,19 , 0,24]	(0,24 , 0,28]	(0,28 , 0,33]
			p-value	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*	*p-value*
		$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
	all-confidence	$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
TopPI		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
	all-confidence	$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
LAM		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} \leq \mu$	0,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	1,000	-	-	-	-	-	-
		$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
	all-confidence	$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
Slim		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \leq \mu$	1,000	-	-	-	-	-	-
	cross-support	$\mu_{PSCIM} = \mu$	0,000	-	-	-	-	-	-
		$\mu_{PSCIM} \geq \mu$	0,000	-	-	-	-	-	-

C.1 Escolha dos Parâmetros

Nesta seção, nós mostramos as distribuições de média de All-confidence (μ) e comprimento mínimo de descrição (MDL) dos conjuntos de itens fechados recuperados pelos algoritmos PSCIM e TopPI, usando valores de parâmetros diferentes. Separamos os resultados do estudo em duas subseções: bases de dados Densa (Subseção C.1.1) e esparsa (Subseção C.1.2). Essa seção tem o intuito de mostrar o processo de escolha dos parâmetros para as técnicas TopPI e PSCIM. Também é possível observar o comportamento de seleção dado um parâmetro quando comparado com os outros parâmetros usados no estudo para determinada base de dados. Para cada base de dados é mostrado um resumo com do comportamento da técnica dado a métrica MDL.

C.1.1 Bases de Dados Densa

C.1.1.1 Chess

Tabela C.7: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 1,00]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**42**	**0,069**	**29**	**0,226**	**42**	**0,338**	**66**	**0,513**	**26**	**0,650**	**8**	**0,791**	**12**	**0,940**	**225**	**0,70**
0,03	47	0,068	33	0,228	43	0,335	74	0,512	36	0,648	16	0,786	28	0,938	277	0,68
0,04	67	0,065	40	0,210	68	0,336	136	0,506	47	0,638	17	0,783	28	0,938	403	0,73
0,05	67	0,065	43	0,208	69	0,336	136	0,506	47	0,638	17	0,783	28	0,938	407	0,75
0,06	76	0,069	45	0,210	73	0,338	153	0,507	69	0,637	36	0,776	61	0,932	513	0,78
0,07	93	0,067	74	0,212	118	0,349	200	0,504	94	0,628	36	0,776	61	0,932	676	0,85
0,08	93	0,067	78	0,212	120	0,348	254	0,507	94	0,628	36	0,776	61	0,932	736	0,84
0,09	107	0,071	89	0,212	149	0,354	274	0,508	138	0,630	53	0,773	119	0,916	929	0,88
0,10	120	0,068	154	0,214	183	0,351	337	0,511	166	0,623	53	0,773	119	0,916	1.132	0,96
0,11	153	0,074	162	0,212	278	0,364	372	0,507	179	0,625	64	0,772	130	0,917	1.338	0,97
0,12	193	0,078	189	0,216	388	0,375	471	0,505	243	0,630	117	0,785	187	0,908	1.788	1,02
0,13	214	0,075	301	0,218	466	0,370	624	0,507	250	0,629	117	0,785	187	0,908	2.159	1,07

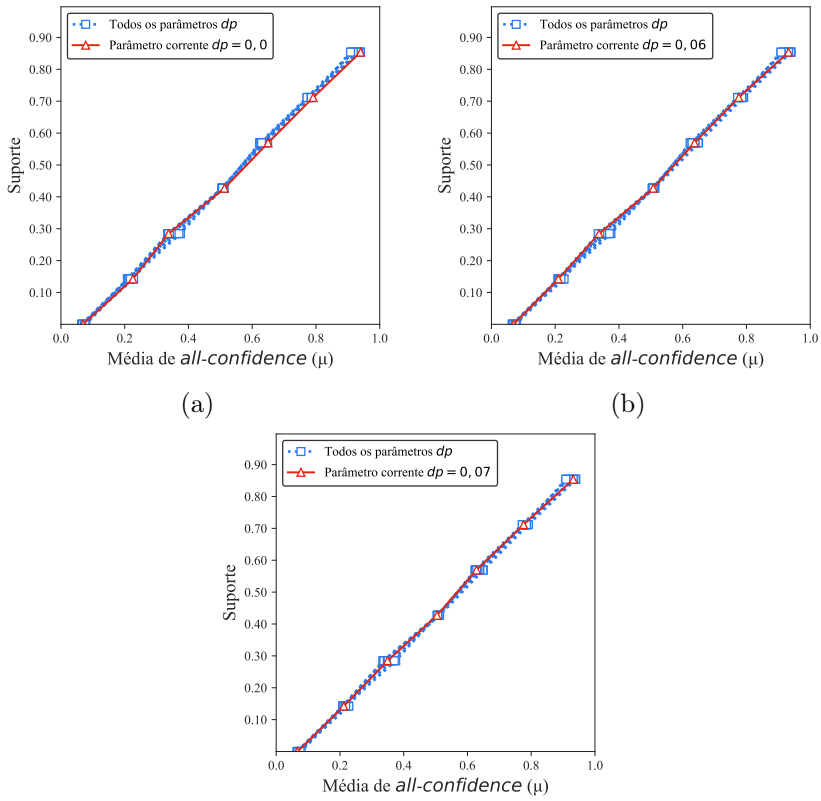
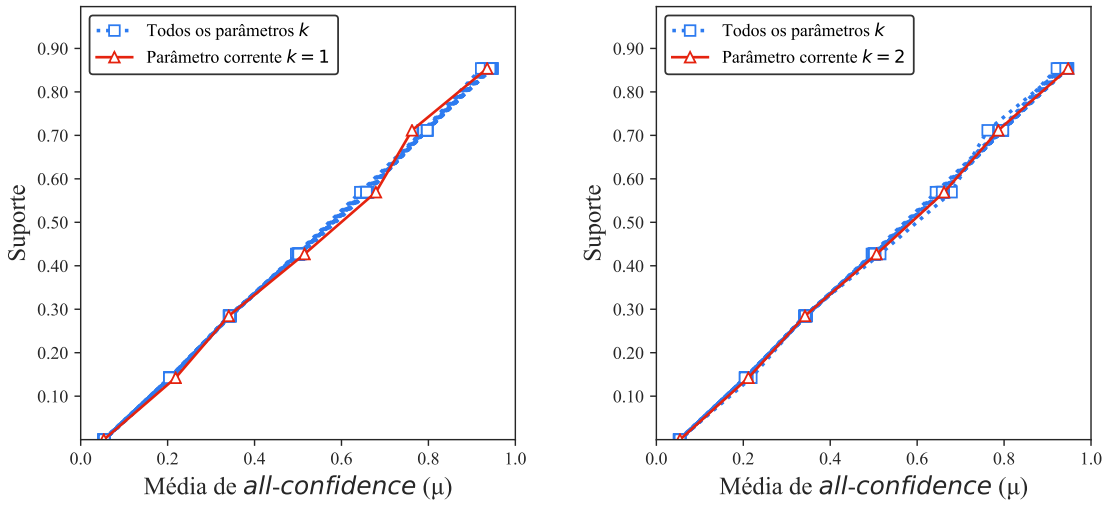


Figura C.3: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, onde essa imagem representa, por similaridade, o comportamento do $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\}$, (b) com $dr = 0{,}06$, e (c) com $dr = 0{,}07$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\}$. Veja Tabela C.7 para detalhes.

Tabela C.8: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

	Partição de suporte															
dr	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 1,00]		Itemset #	Tempo (s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	**18**	**0,053**	**7**	**0,218**	**9**	**0,341**	**5**	**0,515**	**4**	**0,679**	**2**	**0,762**	**3**	**0,935**	**48**	**0,49**
2	35	0,054	14	0,212	19	0,342	11	0,507	14	0,662	10	0,787	18	0,947	121	0,51
3	52	0,055	21	0,209	29	0,342	17	0,504	25	0,661	17	0,793	32	0,947	193	0,56
4	69	0,055	28	0,208	39	0,342	23	0,502	36	0,660	24	0,795	46	0,946	265	0,57
5	86	0,055	35	0,207	49	0,341	29	0,500	47	0,660	31	0,796	59	0,945	336	0,60
6	103	0,055	42	0,206	59	0,341	35	0,499	58	0,659	38	0,797	72	0,944	407	0,60
7	120	0,054	49	0,205	70	0,342	40	0,500	69	0,659	45	0,797	85	0,943	478	0,60
8	137	0,054	56	0,204	81	0,343	45	0,501	80	0,658	52	0,797	98	0,942	549	0,61
9	154	0,054	63	0,204	92	0,343	50	0,501	91	0,657	59	0,797	110	0,942	619	0,62
10	171	0,054	71	0,205	102	0,343	55	0,502	102	0,657	66	0,797	122	0,941	689	0,63
11	188	0,054	79	0,205	112	0,344	60	0,502	113	0,656	73	0,797	134	0,940	759	0,67
12	205	0,054	87	0,205	122	0,344	65	0,502	124	0,655	80	0,796	146	0,939	829	0,63
13	222	0,054	95	0,206	132	0,345	70	0,502	135	0,655	87	0,796	158	0,938	899	0,64
14	239	0,054	103	0,206	142	0,345	75	0,502	146	0,655	94	0,796	170	0,937	969	0,66
15	256	0,054	111	0,206	152	0,345	80	0,502	157	0,654	101	0,796	182	0,936	1.039	0,66
20	341	0,053	151	0,206	202	0,345	105	0,502	212	0,652	141	0,796	233	0,934	1.385	0,73
30	511	0,053	231	0,206	302	0,345	155	0,499	322	0,648	221	0,794	333	0,929	2.075	0,75
40	681	0,052	311	0,205	402	0,344	205	0,497	432	0,646	301	0,791	423	0,925	2.755	0,81
50	851	0,052	392	0,204	501	0,343	255	0,496	542	0,643	381	0,789	510	0,922	3.432	0,84



(a) (b)

Figura C.4: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50\}$. Veja Tabela C.8 para detalhes.

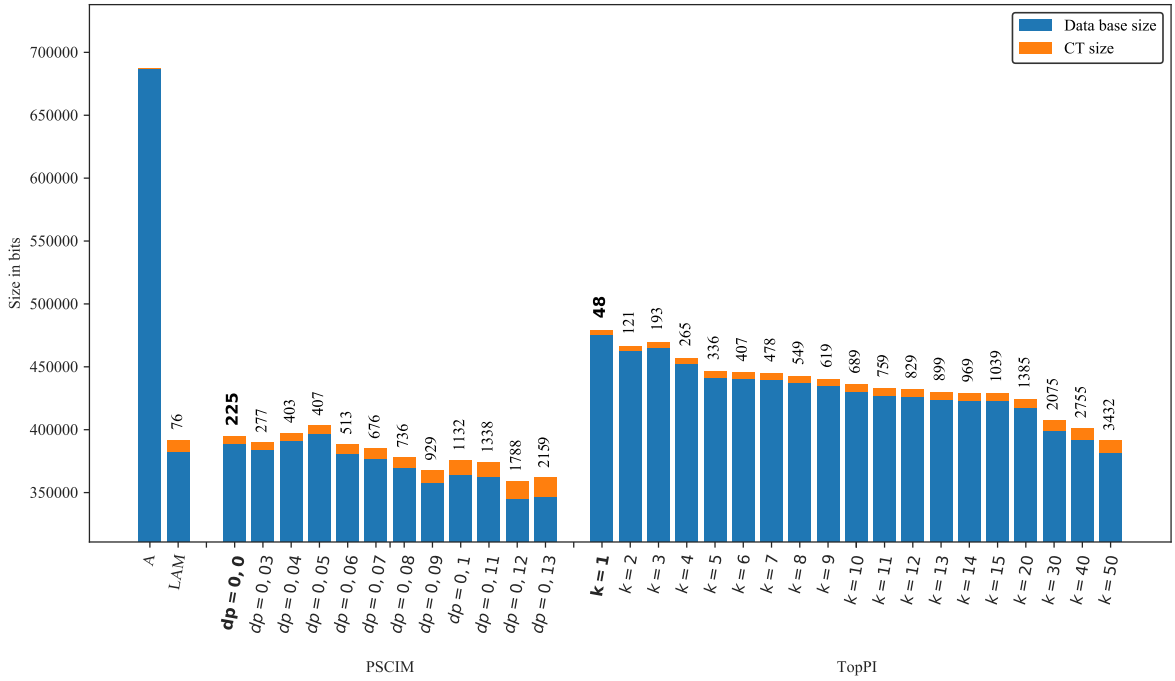


Figura C.5: *Chess*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.2 Kddcup99

Tabela C.9: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,11]		(0,11 , 0,23]		(0,23 , 0,34]		(0,34 , 0,45]		(0,45 , 0,57]		(0,57 , 0,68]		(0,68 , 0,79]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**1.472**	**0,007**	**113**	**0,633**	**10**	**0,384**	**17**	**0,550**	**37**	**0,637**	**1**	**0,851**	**113**	**0,937**	**1.763**	**5,35**
0,02	1.472	0,007	141	0,549	14	0,365	17	0,550	38	0,636	2	0,788	114	0,936	1.798	5,45
0,03	1.499	0,005	146	0,287	14	0,365	17	0,550	74	0,632	3	0,788	145	0,927	1.898	5,56
0,04	1.516	0,005	148	0,287	22	0,342	17	0,550	80	0,634	21	0,827	145	0,927	1.949	5,40
0,05	1.566	0,007	187	0,262	22	0,342	31	0,549	174	0,628	57	0,798	153	0,924	2.190	5,40
0,06	1.591	0,007	256	0,262	33	0,329	32	0,550	393	0,622	67	0,786	153	0,924	2.525	5,37
0,07	1.832	0,007	301	0,261	33	0,329	107	0,560	717	0,623	76	0,779	153	0,924	3.219	5,44
0,08	1.872	0,007	360	0,262	40	0,323	235	0,556	909	0,623	76	0,779	153	0,924	3.645	5,45
0,09	1.940	0,008	396	0,262	41	0,323	313	0,553	980	0,622	92	0,768	153	0,924	3.915	5,47
0,10	2.206	0,017	400	0,260	60	0,316	331	0,553	982	0,623	92	0,768	153	0,924	4.224	5,51

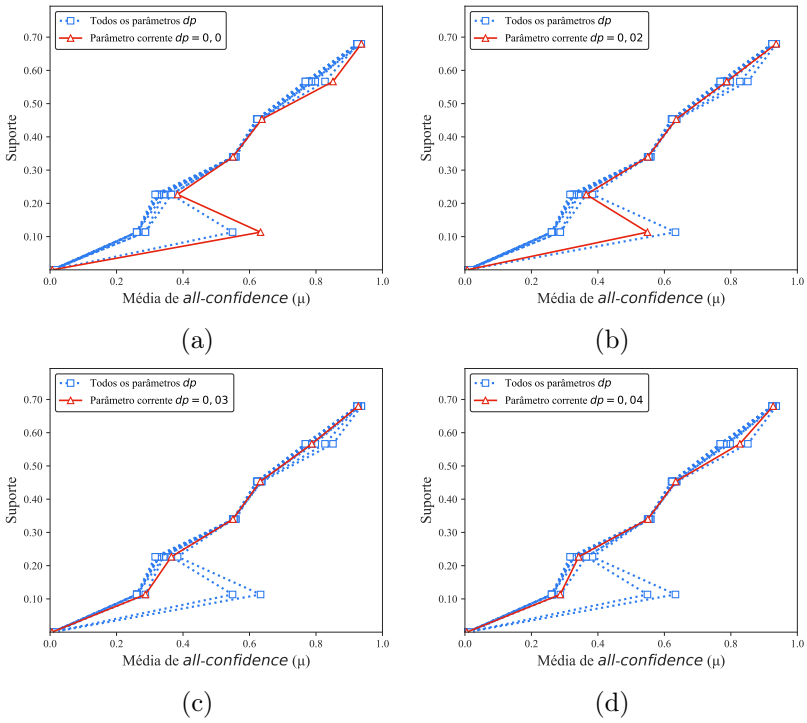
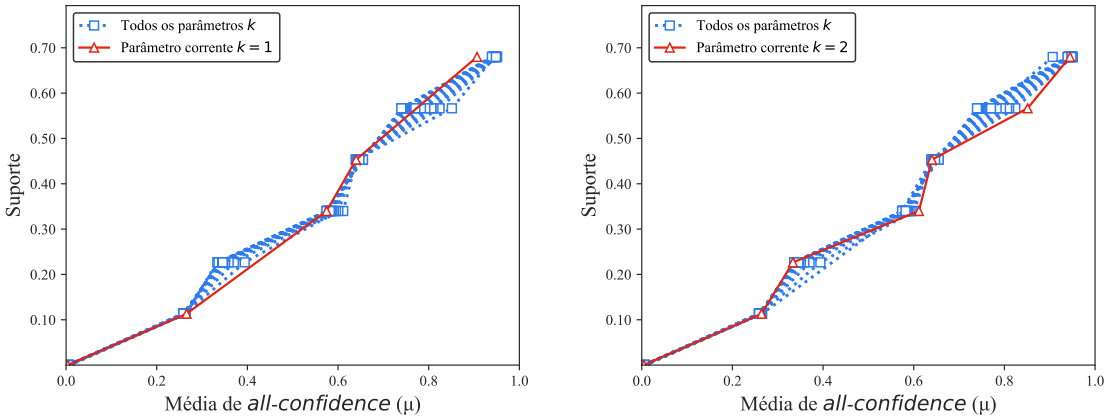


Figura C.6: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}02$, (c) com $dr = 0{,}03$ e (d) com $dr = 0{,}04 \times 10^{-7}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\}$. Veja Tabela C.9 para detalhes.

Tabela C.10: *Kddcup99*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,11]		(0,11 , 0,23]		(0,23 , 0,34]		(0,34 , 0,45]		(0,45 , 0,57]		(0,57 , 0,68]		(0,68 , 0,79]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	97	0,004	13	0,266	0	0,000	1	0,575	1	0,640	0	0,000	2	0,907	114	1,89
2	192	0,004	27	0,264	4	0,334	3	0,611	4	0,640	1	0,851	11	0,945	242	1,99
3	**287**	**0,005**	**41**	**0,262**	**9**	**0,394**	**4**	**0,601**	**7**	**0,639**	**2**	**0,824**	**17**	**0,949**	**367**	**2,73**
4	379	0,005	54	0,261	14	0,370	5	0,595	10	0,639	4	0,811	23	0,951	489	3,01
5	471	0,005	64	0,260	19	0,359	6	0,591	14	0,641	5	0,806	28	0,950	607	3,32
6	563	0,005	74	0,259	24	0,351	7	0,588	18	0,643	5	0,806	33	0,950	724	3,53
7	647	0,005	85	0,259	28	0,347	8	0,585	23	0,645	6	0,793	38	0,950	835	3,91
8	728	0,005	96	0,258	32	0,344	9	0,583	28	0,646	7	0,783	43	0,949	943	4,19
9	811	0,006	105	0,259	37	0,346	10	0,582	33	0,647	9	0,771	47	0,948	1.052	4,46
10	891	0,006	113	0,259	42	0,342	11	0,581	38	0,648	11	0,763	51	0,947	1.157	4,66
11	971	0,006	121	0,259	46	0,341	12	0,580	43	0,648	13	0,757	55	0,946	1.261	4,99
12	1.052	0,006	128	0,259	50	0,339	13	0,580	48	0,649	15	0,753	59	0,945	1.365	5,31
13	1.132	0,006	136	0,261	53	0,339	14	0,579	53	0,649	17	0,750	63	0,945	1.468	5,50
14	1.208	0,007	146	0,262	56	0,338	15	0,578	58	0,649	19	0,748	67	0,944	1.569	5,67
15	1.278	0,007	155	0,262	59	0,338	16	0,577	63	0,649	21	0,746	71	0,944	1.663	5,93
16	1.342	0,007	163	0,263	62	0,337	17	0,577	68	0,649	23	0,744	75	0,943	1.750	6,17
17	1.407	0,007	172	0,263	65	0,337	18	0,576	72	0,653	25	0,743	78	0,942	1.837	6,38
18	1.472	0,007	181	0,263	68	0,337	20	0,576	76	0,653	27	0,741	81	0,941	1.925	6,56
19	1.535	0,007	192	0,264	71	0,336	22	0,575	80	0,654	29	0,740	84	0,940	2.013	6,84
20	1.598	0,007	201	0,264	74	0,336	24	0,574	84	0,655	31	0,740	87	0,939	2.099	7,10



(a) (b)

Figura C.7: *Chess*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50\}$. Veja Tabela C.10 para detalhes.

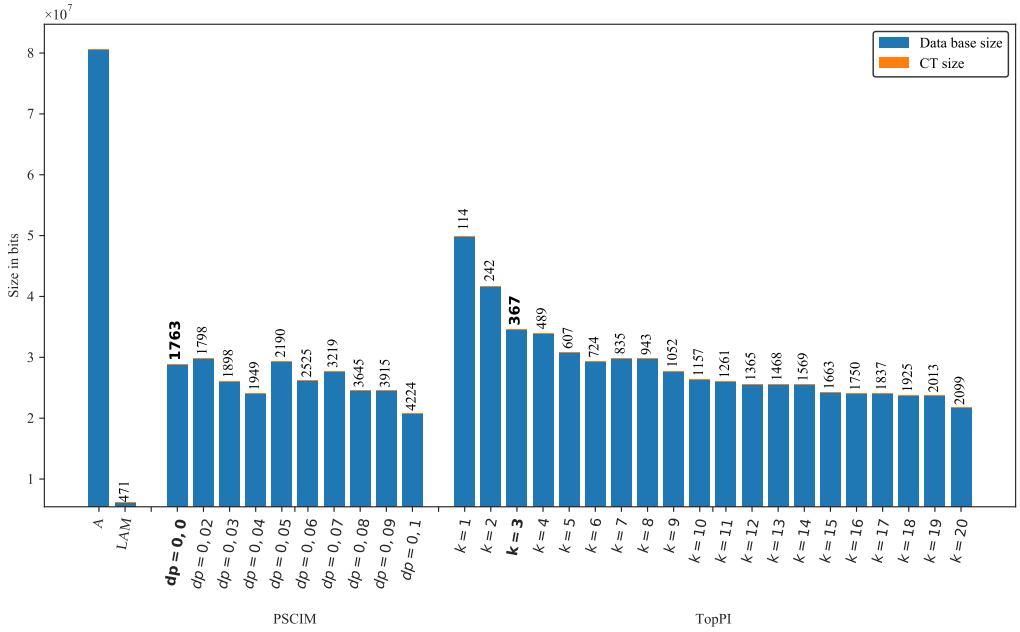


Figura C.8: *Kddcup99*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.3 Mushrooms

Tabela C.11: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,83]		(0,83 , 0,97]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	194	0,071	84	0,254	28	0,418	10	0,606	2	0,585	0	0,000	1	0,997	319	0,74
0,02	**197**	**0,072**	**86**	**0,254**	**43**	**0,431**	**13**	**0,599**	**3**	**0,591**	**1**	**0,832**	**1**	**0,997**	**344**	**0,75**
0,03	255	0,071	111	0,242	48	0,425	17	0,583	3	0,591	1	0,832	1	0,997	436	0,75
0,04	268	0,072	135	0,244	67	0,420	20	0,568	3	0,591	1	0,832	1	0,997	495	0,76
0,05	357	0,067	168	0,237	79	0,412	21	0,564	3	0,591	1	0,832	1	0,997	630	0,77
0,06	390	0,067	207	0,235	103	0,413	24	0,553	3	0,591	1	0,832	1	0,997	729	0,81
0,07	455	0,068	242	0,233	115	0,406	24	0,552	3	0,591	1	0,832	1	0,997	841	0,79
0,08	528	0,068	300	0,234	125	0,403	26	0,552	3	0,591	1	0,832	1	0,997	984	0,80
0,09	618	0,071	334	0,234	134	0,394	26	0,552	3	0,591	1	0,832	1	0,997	1.117	0,81
0,10	844	0,067	402	0,231	141	0,392	26	0,552	3	0,591	1	0,832	1	0,997	1.418	0,84
0,11	1.009	0,068	436	0,228	146	0,391	26	0,552	3	0,591	1	0,832	1	0,997	1.622	0,85
0,12	1.329	0,070	530	0,227	149	0,391	27	0,548	3	0,591	1	0,832	1	0,997	2.040	0,88

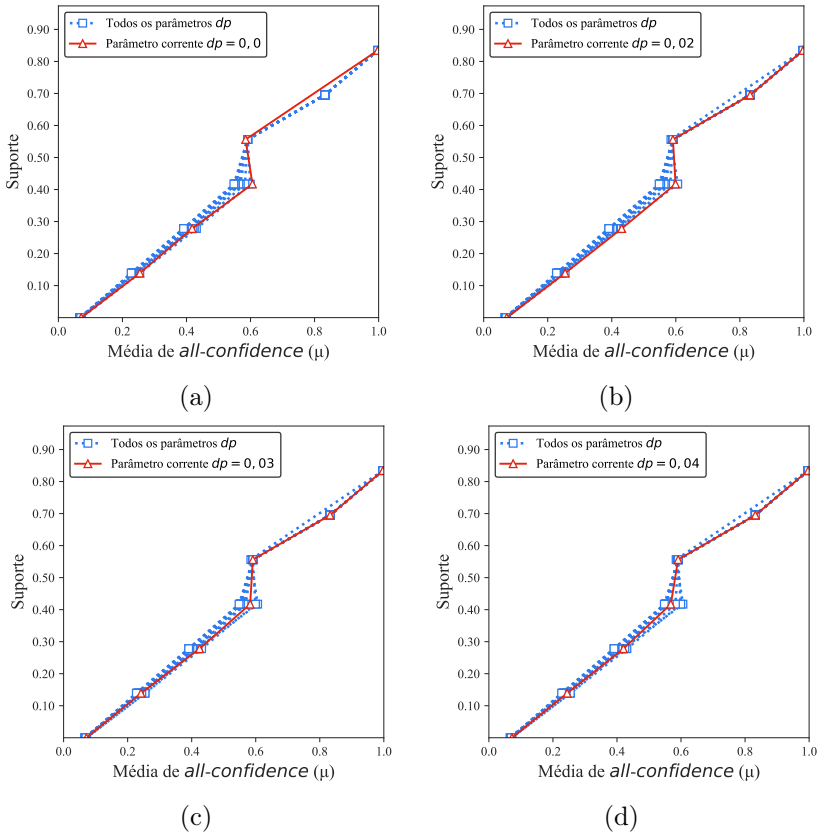
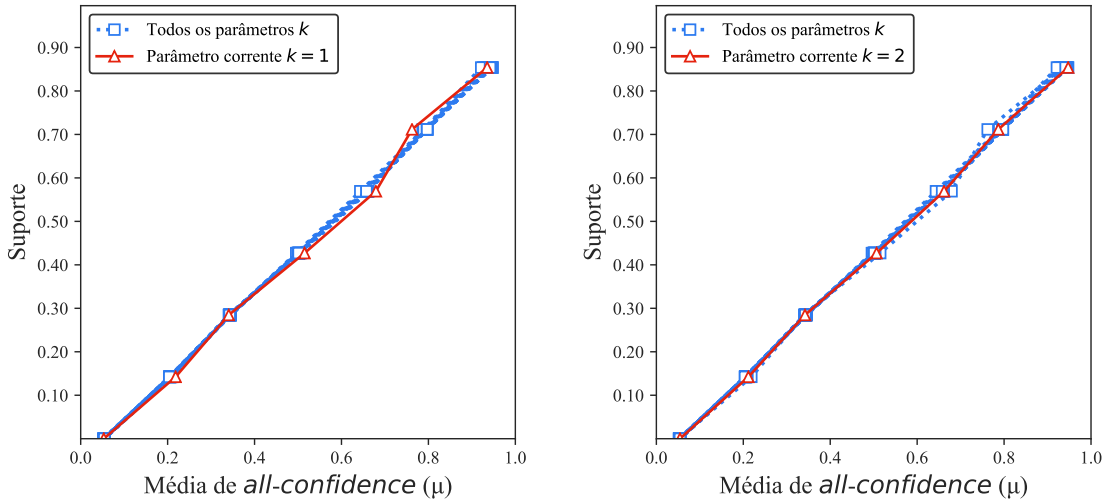


Figura C.9: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}02$, (c) com $dr = 0{,}03$ e (d) com $dr = 0{,}04$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\}$. Veja Tabela C.11 para detalhes.

Tabela C.12: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

	Partição de suporte															
dr	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,83]		(0,83 , 0,97]		Itemset #	Tempo (s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	62	0,045	16	0,215	7	0,337	2	0,514	2	0,585	0	0,000	0	0,000	89	0,37
2	126	0,044	34	0,217	17	0,347	12	0,517	6	0,611	1	0,807	2	0,960	198	0,37
3	**190**	**0,044**	**52**	**0,220**	**26**	**0,355**	**21**	**0,521**	**9**	**0,619**	**2**	**0,806**	**3**	**0,947**	**303**	**0,37**
4	256	0,042	70	0,219	37	0,364	28	0,510	12	0,617	3	0,815	3	0,947	409	0,37
5	328	0,042	83	0,220	47	0,372	35	0,523	14	0,616	4	0,802	3	0,947	514	0,37
6	393	0,043	98	0,222	59	0,376	42	0,531	14	0,616	4	0,802	3	0,947	613	0,37
7	459	0,043	112	0,222	70	0,381	48	0,526	14	0,616	4	0,802	3	0,947	710	0,37
8	525	0,043	124	0,222	81	0,380	54	0,523	14	0,616	4	0,802	3	0,947	805	0,37
9	591	0,043	139	0,221	95	0,379	58	0,520	14	0,616	4	0,802	3	0,947	904	0,37
10	661	0,042	154	0,221	106	0,380	61	0,518	14	0,616	4	0,802	3	0,947	1.003	0,38
11	730	0,042	168	0,221	118	0,382	64	0,519	14	0,616	4	0,802	3	0,947	1.101	0,38
12	803	0,041	183	0,221	129	0,384	65	0,520	14	0,616	4	0,802	3	0,947	1.201	0,37
13	871	0,041	198	0,220	139	0,383	67	0,520	14	0,616	4	0,802	3	0,947	1.296	0,37
14	942	0,041	212	0,220	149	0,382	67	0,520	14	0,616	4	0,802	3	0,947	1.391	0,37
15	1.006	0,040	225	0,219	161	0,380	68	0,521	14	0,616	4	0,802	3	0,947	1.481	0,37
16	1.073	0,040	240	0,218	171	0,381	68	0,521	14	0,616	4	0,802	3	0,947	1.573	0,37
17	1.134	0,040	255	0,218	182	0,379	69	0,520	14	0,616	4	0,802	3	0,947	1.661	0,37
18	1.199	0,039	272	0,218	192	0,379	69	0,520	14	0,616	4	0,802	3	0,947	1.753	0,37
19	1.261	0,039	289	0,219	198	0,379	69	0,520	14	0,616	4	0,802	3	0,947	1.838	0,37
20	1.326	0,039	305	0,220	204	0,378	69	0,520	14	0,616	4	0,802	3	0,947	1.925	0,37



(a) (b)

Figura C.10: *Mushrooms*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50\}$. Veja Tabela C.12 para detalhes.

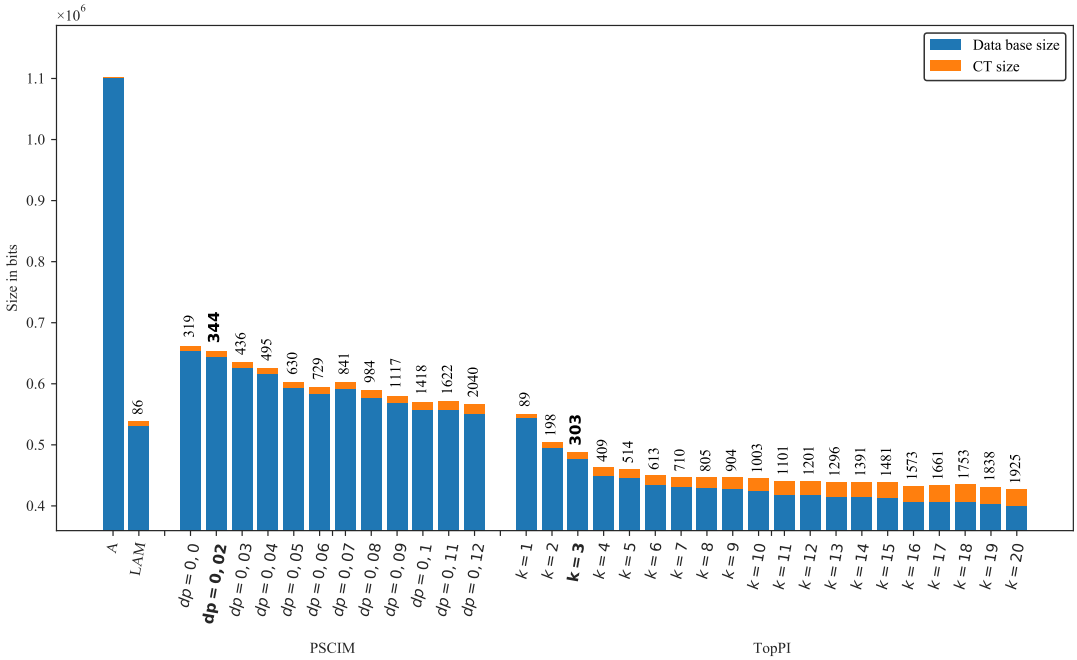
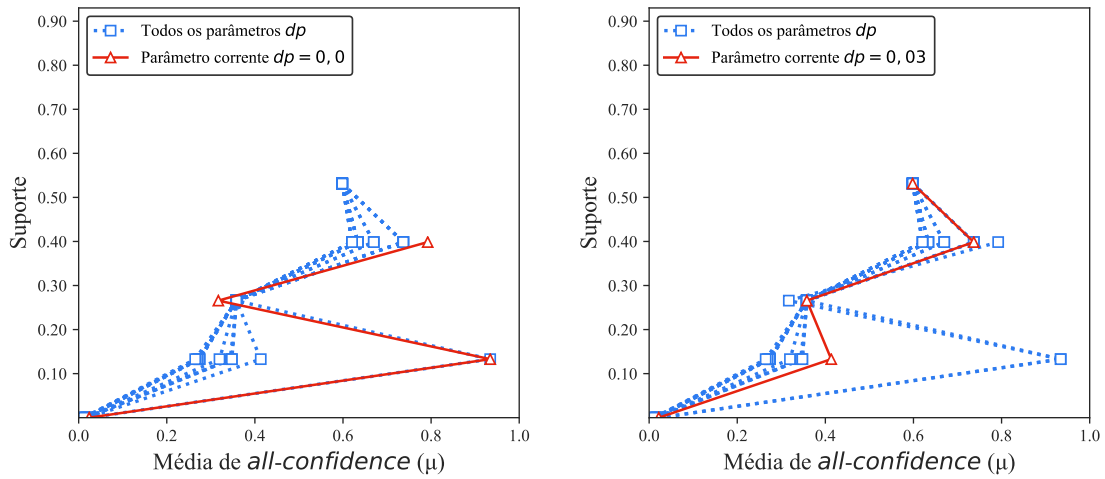


Figura C.11: *Mushrooms*: valores métricos de MDL para todas as configurações de parâmetros em cada variação do algoritmo PSCIM. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.4 PowerC

Tabela C.13: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,13]		(0,13 , 0,27]		(0,27 , 0,40]		(0,40 , 0,53]		(0,53 , 0,66]		(0,66 , 0,80]		(0,80 , 0,93]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	751	0,024	1	0,934	1	0,317	4	0,792	0	0,000	0	0,000	0	0,000	757	3,38
0,02	**758**	**0,024**	**1**	**0,934**	**2**	**0,357**	**5**	**0,737**	**1**	**0,598**	**0**	**0,000**	**0**	**0,000**	**767**	**3,45**
0,03	851	0,022	4	0,413	3	0,358	5	0,737	1	0,598	0	0,000	0	0,000	864	3,38
0,04	1.049	0,019	6	0,344	4	0,359	5	0,737	1	0,598	0	0,000	0	0,000	1.065	3,43
0,05	1.123	0,018	7	0,348	5	0,358	5	0,737	1	0,598	0	0,000	0	0,000	1.141	3,39
0,06	1.297	0,017	7	0,348	5	0,358	7	0,670	1	0,598	0	0,000	0	0,000	1.317	3,44
0,07	1.370	0,017	9	0,320	5	0,358	7	0,670	1	0,598	0	0,000	0	0,000	1.392	3,40
0,08	1.596	0,016	14	0,275	5	0,358	9	0,634	1	0,598	0	0,000	0	0,000	1.625	3,43
0,09	1.739	0,015	14	0,275	5	0,358	9	0,634	1	0,598	0	0,000	0	0,000	1.768	3,39
0,10	1.885	0,014	16	0,273	5	0,358	9	0,634	1	0,598	0	0,000	0	0,000	1.916	3,44
0,11	2.150	0,013	16	0,273	5	0,358	9	0,634	1	0,598	0	0,000	0	0,000	2.181	3,41
0,12	2.434	0,012	16	0,273	5	0,358	10	0,621	1	0,598	0	0,000	0	0,000	2.466	3,45
0,13	2.597	0,012	17	0,269	5	0,358	10	0,621	1	0,598	0	0,000	0	0,000	2.630	3,46
0,14	2.810	0,012	18	0,265	5	0,358	10	0,621	1	0,598	0	0,000	0	0,000	2.844	3,41



(a) (b)

Figura C.12: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,,0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,,0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\,,0{,}11\,,0{,}12\,,0{,}13\,,0{,}14\}$. Veja Tabela C.13 para detalhes.

Tabela C.14: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,13]		(0,13 , 0,27]		(0,27 , 0,40]		(0,40 , 0,53]		(0,53 , 0,66]		(0,66 , 0,80]		(0,80 , 0,93]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	12	0,061	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	12	1,16
2	121	0,024	3	0,182	3	0,358	2	0,509	2	0,638	0	0,000	1	0,961	132	1,14
3	226	0,048	6	0,182	6	0,359	4	0,511	4	0,639	0	0,000	1	0,961	247	1,67
4	327	0,039	9	0,180	9	0,354	6	0,511	6	0,634	0	0,000	1	0,961	358	1,88
5	430	0,041	13	0,274	9	0,354	9	0,580	6	0,634	0	0,000	1	0,961	468	1,98
6	528	0,035	17	0,261	9	0,354	12	0,558	6	0,634	0	0,000	1	0,961	573	2,05
7	628	0,031	21	0,252	9	0,354	15	0,546	6	0,634	0	0,000	1	0,961	680	2,12
8	725	0,028	25	0,246	9	0,354	18	0,537	6	0,634	0	0,000	1	0,961	784	2,19
9	**832**	**0,027**	**28**	**0,256**	**10**	**0,380**	**19**	**0,560**	**6**	**0,634**	**0**	**0,000**	**1**	**0,961**	**896**	**2,25**
10	936	0,025	30	0,253	11	0,380	20	0,557	6	0,634	0	0,000	1	0,961	1.004	2,25
11	1.040	0,024	32	0,250	12	0,380	21	0,555	6	0,634	0	0,000	1	0,961	1.112	2,32
12	1.143	0,023	34	0,248	13	0,379	22	0,552	6	0,634	0	0,000	1	0,961	1.219	2,37
13	1.247	0,023	36	0,255	14	0,391	23	0,560	6	0,634	0	0,000	1	0,961	1.327	2,34
14	1.349	0,023	38	0,251	15	0,390	24	0,557	6	0,634	0	0,000	1	0,961	1.433	2,32
15	1.449	0,022	40	0,248	16	0,389	25	0,555	6	0,634	0	0,000	1	0,961	1.537	2,38
16	1.547	0,021	42	0,246	17	0,388	26	0,553	6	0,634	0	0,000	1	0,961	1.639	2,45
17	1.651	0,021	44	0,249	18	0,387	26	0,553	6	0,634	0	0,000	1	0,961	1.746	2,36
18	1.753	0,021	46	0,247	19	0,386	26	0,553	6	0,634	0	0,000	1	0,961	1.851	2,54
19	1.855	0,020	48	0,244	20	0,385	26	0,553	6	0,634	0	0,000	1	0,961	1.956	2,50
20	1.955	0,020	50	0,242	21	0,384	26	0,553	6	0,634	0	0,000	1	0,961	2.059	2,52
30	2.936	0,017	70	0,238	25	0,381	26	0,553	6	0,634	0	0,000	1	0,961	3.064	2,87
40	3.926	0,015	90	0,229	25	0,381	26	0,553	6	0,634	0	0,000	1	0,961	4.074	2,77
50	4.896	0,014	100	0,222	25	0,381	26	0,553	6	0,634	0	0,000	1	0,961	5.054	2,87

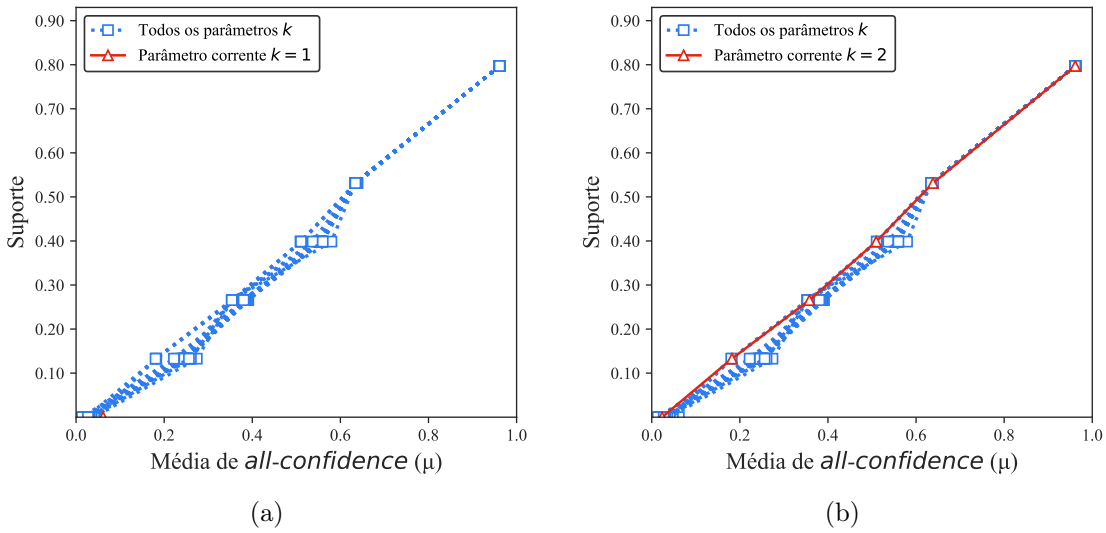


Figura C.13: *PowerC*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50\}$. Veja Tabela C.14 para detalhes.

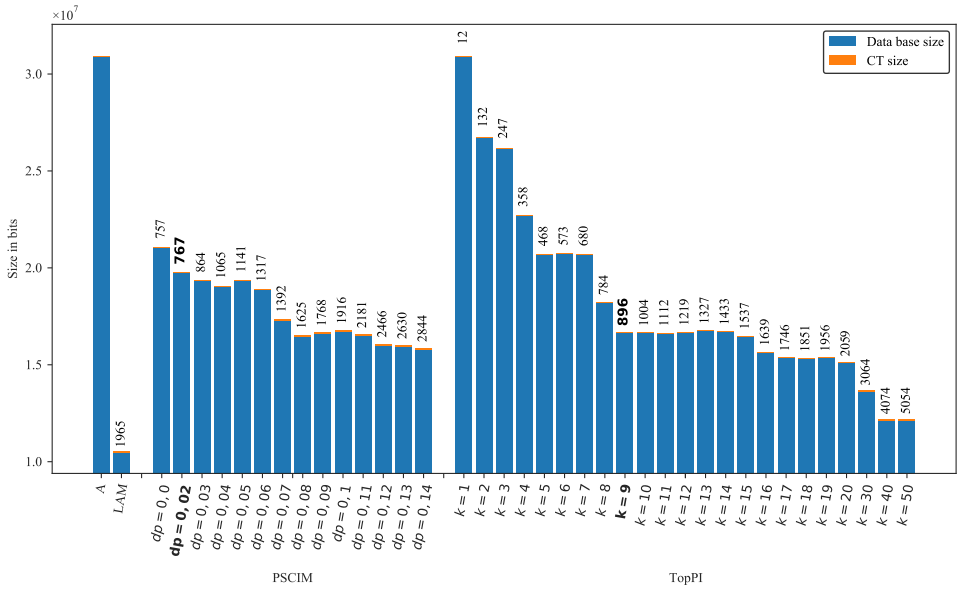
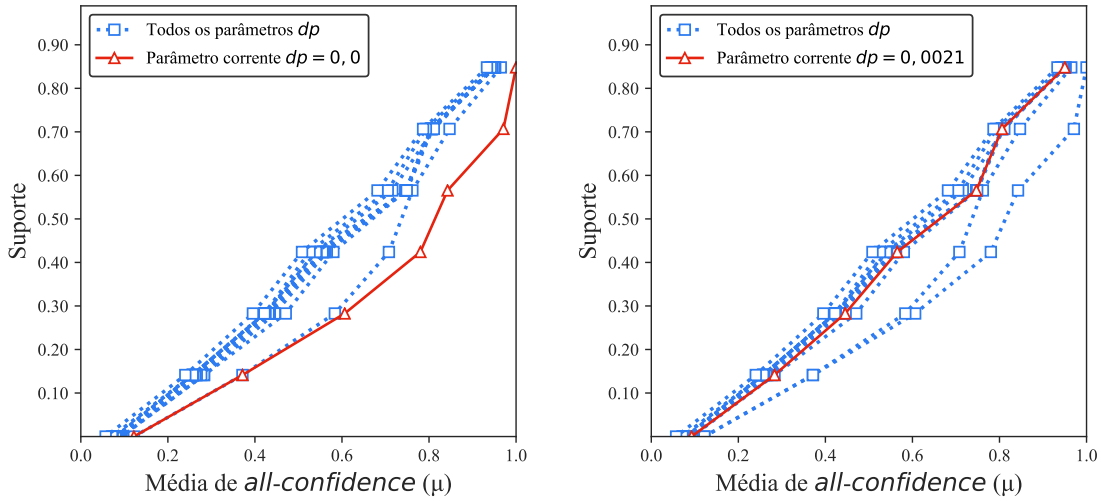


Figura C.14: *PowerC*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.5 Pumsb

Tabela C.15: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]			
$\clubsuit \times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	**1.135**	**0,122**	**69**	**0,371**	**33**	**0,606**	**31**	**0,780**	**4**	**0,843**	**5**	**0,971**	**4**	**1,000**	**1.281**	**190,07**
0,10 ♣	1.135	0,122	69	0,371	40	0,584	45	0,708	12	0,762	12	0,847	20	0,965	1.333	190,34
0,20 ♣	1.589	0,094	136	0,283	81	0,471	95	0,580	22	0,747	23	0,803	39	0,952	1.985	193,05
0,21 ♣	1.604	0,093	141	0,283	106	0,445	115	0,565	22	0,747	25	0,806	45	0,950	2.058	193,58
0,22 ♣	1.654	0,090	152	0,277	118	0,434	119	0,562	22	0,747	25	0,806	45	0,950	2.135	193,58
0,23 ♣	1.684	0,089	169	0,266	120	0,433	123	0,559	22	0,747	25	0,806	45	0,950	2.188	193,73
0,24 ♣	1.720	0,088	178	0,264	125	0,431	133	0,557	22	0,747	25	0,806	45	0,950	2.248	194,15
0,25 ♣	1.778	0,085	213	0,258	125	0,431	135	0,555	22	0,747	25	0,806	45	0,950	2.343	194,41
0,26 ♣	1.805	0,084	213	0,258	133	0,424	135	0,555	22	0,747	25	0,806	45	0,950	2.378	195,37
0,27 ♣	1.845	0,082	223	0,257	133	0,424	135	0,555	22	0,747	25	0,806	45	0,950	2.428	195,50
0,28 ♣	1.910	0,080	223	0,257	143	0,423	153	0,542	24	0,749	30	0,808	52	0,950	2.535	195,90
0,29 ♣	2.144	0,073	223	0,257	147	0,422	159	0,540	39	0,716	32	0,811	62	0,946	2.806	198,40
0,30 ♣	2.186	0,072	223	0,257	160	0,422	188	0,525	46	0,705	42	0,793	76	0,942	2.921	198,88
0,40 ♣	2.967	0,057	313	0,240	303	0,395	319	0,508	86	0,682	81	0,786	134	0,933	4.203	198,44

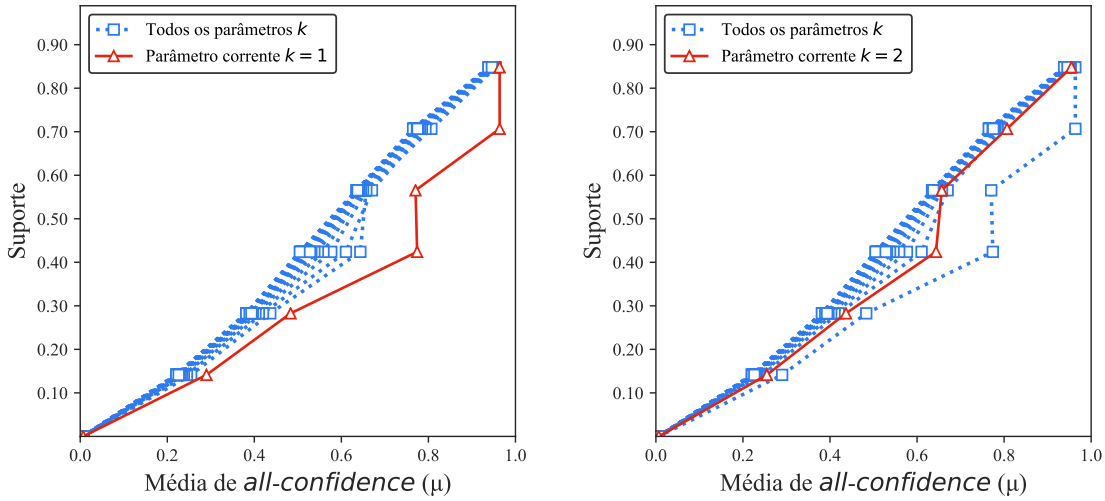


(a) (b)

Figura C.15: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}20\}$ e (b) com $dr = 0{,}21 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}22\,, 0{,}23\,, 0{,}24\,, 0{,}25\,, 0{,}26\,, 0{,}27\,, 0{,}28\,, 0{,}28\,, 0{,}29\,, 0{,}30\,, 0{,}40\}$. Veja Tabela C.15 para detalhes.

Tabela C.16: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 0,99]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	**1.855**	**0,007**	**18**	**0,290**	**7**	**0,483**	**12**	**0,774**	**2**	**0,770**	**3**	**0,964**	**4**	**0,964**	**1.901**	**1,20**
2	3.648	0,007	46	0,255	22	0,437	32	0,644	10	0,657	14	0,807	22	0,955	3.794	1,59
3	5.442	0,007	73	0,245	37	0,420	52	0,610	18	0,670	23	0,793	39	0,953	5.684	1,86
4	7.151	0,007	100	0,238	52	0,410	71	0,577	26	0,657	33	0,784	56	0,952	7.489	1,98
5	8.851	0,007	127	0,234	67	0,403	91	0,559	34	0,650	44	0,785	72	0,950	9.286	2,23
6	10.557	0,007	154	0,230	82	0,396	111	0,546	42	0,645	54	0,780	88	0,948	11.088	2,64
7	12.257	0,007	180	0,228	97	0,395	131	0,538	50	0,642	65	0,777	104	0,948	12.884	2,57
8	13.899	0,007	206	0,227	112	0,389	151	0,531	58	0,639	76	0,774	118	0,947	14.620	2,77
9	15.542	0,007	232	0,226	127	0,393	171	0,528	66	0,637	86	0,775	134	0,946	16.358	2,90
10	17.183	0,007	258	0,225	142	0,395	190	0,524	74	0,638	95	0,774	149	0,945	18.091	3,15
11	18.821	0,007	284	0,224	157	0,394	210	0,521	82	0,638	104	0,772	163	0,944	19.821	3,62
12	20.454	0,007	310	0,223	172	0,392	230	0,517	90	0,639	113	0,771	179	0,943	21.548	3,52
13	22.086	0,007	336	0,223	187	0,388	250	0,515	98	0,640	121	0,771	193	0,942	23.271	3,68
14	23.716	0,007	362	0,222	202	0,386	269	0,512	106	0,638	131	0,769	206	0,941	24.992	3,97
15	25.332	0,007	388	0,222	217	0,383	289	0,511	114	0,639	140	0,769	220	0,941	26.700	4,31
16	26.922	0,007	414	0,221	232	0,382	309	0,509	122	0,637	150	0,768	234	0,940	28.383	4,56
17	28.510	0,007	440	0,221	247	0,383	329	0,508	130	0,636	160	0,767	249	0,939	30.065	4,77
18	30.098	0,007	466	0,221	262	0,382	348	0,506	138	0,637	169	0,766	264	0,939	31.745	5,02
19	31.682	0,007	493	0,221	276	0,382	368	0,505	146	0,636	177	0,766	278	0,939	33.420	5,20
20	33.268	0,007	520	0,221	290	0,382	387	0,505	154	0,634	188	0,765	291	0,938	35.098	5,49



(a) (b)

Figura C.16: *Pumsb*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.16 para detalhes.

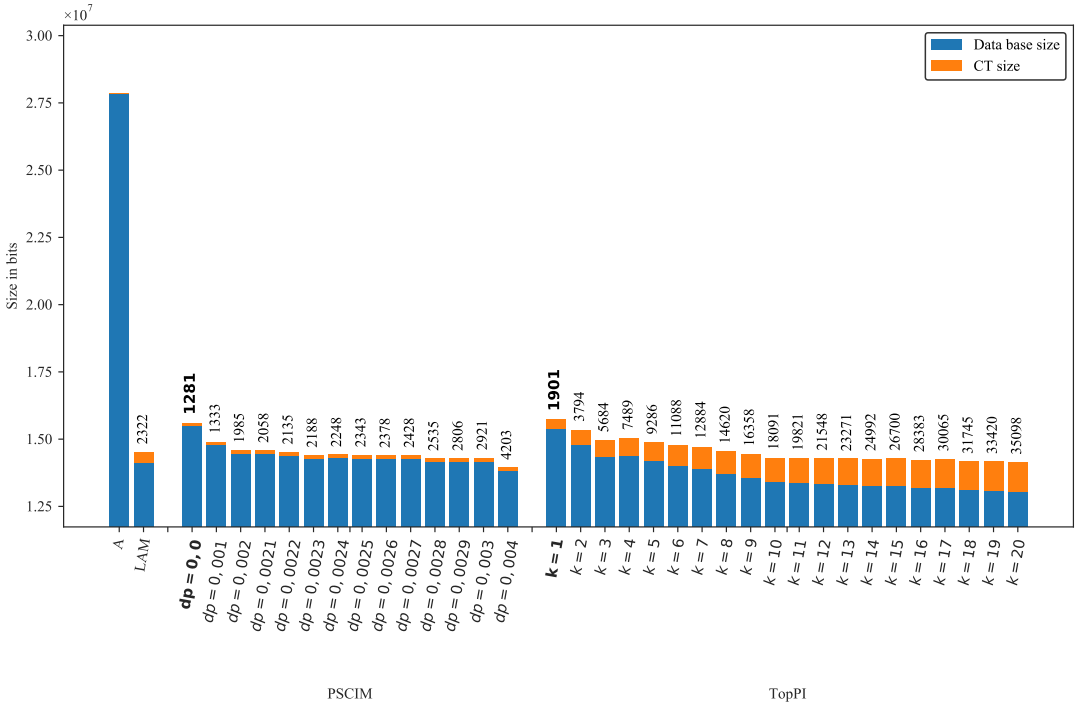
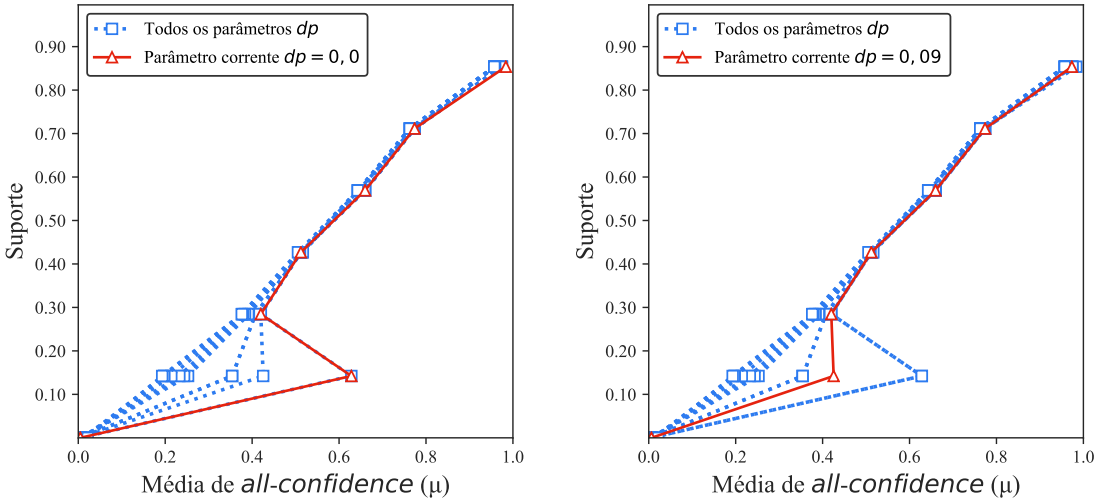


Figura C.17: *Pumsb*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.6 RecordLink

Tabela C.17: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 1,00]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**220**	**0,005**	**1**	**0,628**	**15**	**0,420**	**13**	**0,512**	**14**	**0,660**	**10**	**0,774**	**4**	**0,984**	**277**	**2,35**
0,08	220	0,005	1	0,628	15	0,420	13	0,512	14	0,660	10	0,774	6	0,974	279	2,33
0,09	222	0,006	2	0,425	15	0,420	13	0,512	14	0,660	10	0,774	6	0,974	282	2,39
0,10	222	0,006	3	0,354	21	0,408	22	0,512	14	0,660	10	0,774	9	0,975	301	2,34
0,15	247	0,006	3	0,354	21	0,408	22	0,512	20	0,645	15	0,770	9	0,975	337	2,34
0,20	262	0,006	8	0,253	47	0,393	31	0,511	24	0,645	18	0,770	15	0,959	405	2,38
0,25	271	0,007	12	0,243	49	0,391	33	0,511	25	0,643	23	0,771	15	0,959	428	2,34
0,30	287	0,007	18	0,233	76	0,383	39	0,515	29	0,651	27	0,767	15	0,959	491	2,34
0,35	302	0,009	29	0,214	77	0,383	41	0,517	32	0,646	33	0,764	17	0,957	531	2,39
0,40	369	0,013	68	0,198	124	0,379	60	0,510	35	0,645	35	0,763	17	0,957	708	2,35
0,45	410	0,016	77	0,197	135	0,377	72	0,508	36	0,645	35	0,763	17	0,957	782	2,38
0,50	620	0,021	140	0,192	139	0,376	80	0,505	40	0,643	35	0,763	17	0,957	1.071	2,37

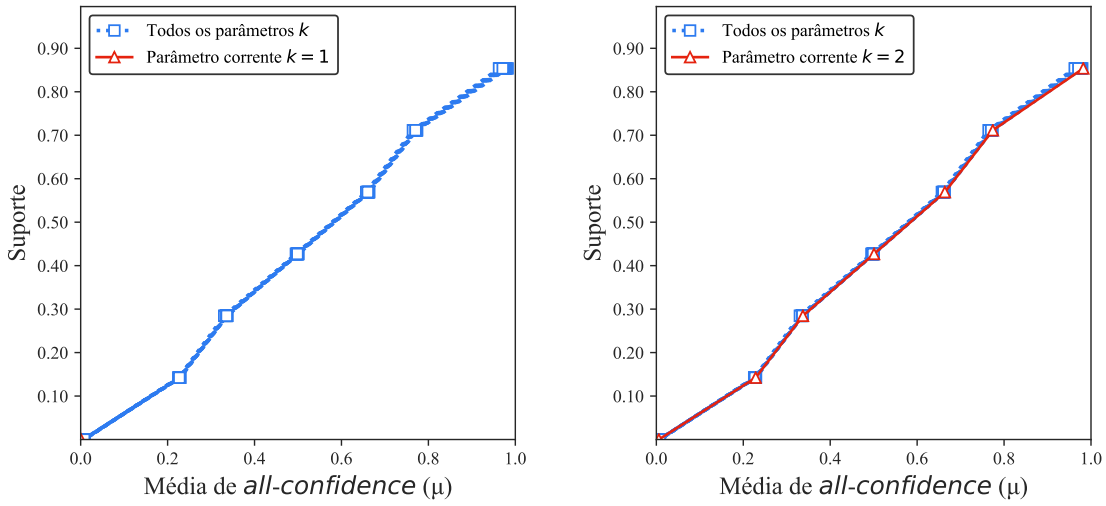


(a) (b)

Figura C.18: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\}$ e (b) com $dr = 0{,}09$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\,, 0{,}15\,, 0{,}20\,, 0{,}25\,, 0{,}30\,, 0{,}35\,, 0{,}40\,, 0{,}45\,, 0{,}50\}$. Veja Tabela C.17 para detalhes.

Tabela C.18: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

	Partição de suporte															
dr	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,85]		(0,85 , 1,00]		Itemset #	Tempo (s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	6	0,001	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	6	0,96
2	**16**	**0,006**	**3**	**0,229**	**1**	**0,337**	**2**	**0,501**	**1**	**0,664**	**3**	**0,774**	**4**	**0,982**	**30**	**0,95**
3	26	0,007	6	0,229	2	0,337	4	0,500	2	0,663	6	0,773	7	0,980	53	1,46
4	36	0,007	9	0,228	3	0,336	6	0,499	3	0,662	9	0,772	10	0,978	76	1,46
5	46	0,007	12	0,228	4	0,336	8	0,499	4	0,662	12	0,772	12	0,977	98	1,61
6	56	0,007	15	0,228	5	0,336	10	0,499	5	0,662	15	0,772	14	0,975	120	1,67
7	65	0,007	18	0,228	6	0,337	12	0,500	6	0,661	18	0,772	16	0,973	141	1,81
8	74	0,008	21	0,227	7	0,336	14	0,499	7	0,661	21	0,771	18	0,972	162	1,79
9	83	0,008	24	0,227	8	0,336	16	0,499	8	0,661	24	0,770	19	0,971	182	1,91
10	92	0,008	27	0,227	9	0,336	18	0,498	9	0,659	27	0,769	20	0,969	202	1,91
11	101	0,008	30	0,226	10	0,335	20	0,498	10	0,661	30	0,768	21	0,968	222	2,10
12	110	0,008	33	0,226	11	0,335	22	0,497	11	0,660	33	0,767	22	0,966	242	2,12
13	119	0,008	36	0,226	12	0,334	24	0,496	12	0,659	36	0,767	23	0,965	262	2,13
14	128	0,008	39	0,225	13	0,332	26	0,496	13	0,658	39	0,766	24	0,964	282	2,24
15	137	0,008	42	0,225	14	0,330	28	0,495	14	0,657	42	0,764	25	0,963	302	2,18



(a) (b)

Figura C.19: *RecordLink*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$. Veja Tabela C.18 para detalhes.

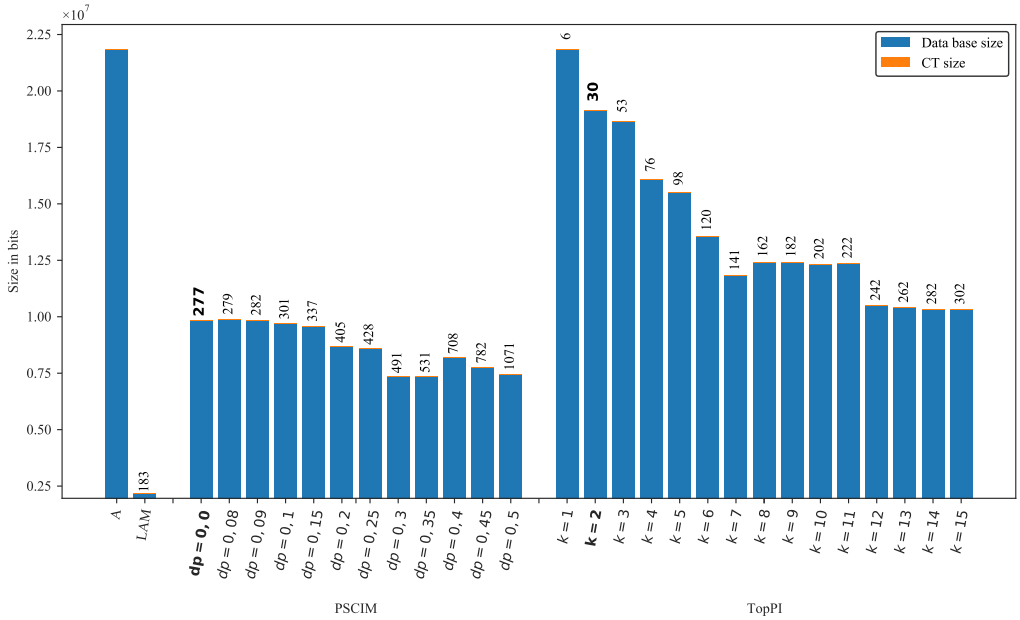


Figura C.20: *RecordLink*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.7 Skin

Tabela C.19: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,06]		(0,06 , 0,13]		(0,13 , 0,19]		(0,19 , 0,26]		(0,26 , 0,32]		(0,32 , 0,38]		(0,38 , 0,45]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,57
0,10	2	0,010	0	0,000	4	0,341	6	0,425	3	0,685	0	0,000	0	0,000	15	0,56
0,20	2	0,010	0	0,000	5	0,339	6	0,425	3	0,685	0	0,000	0	0,000	16	0,56
0,30	2	0,010	0	0,000	6	0,342	6	0,425	3	0,685	0	0,000	1	0,565	18	0,56
0,40	**2**	**0,010**	**3**	**0,217**	**6**	**0,342**	**6**	**0,425**	**3**	**0,685**	**0**	**0,000**	**1**	**0,565**	**21**	**0,56**
0,50	3	0,044	3	0,217	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	22	0,56
0,60	5	0,085	3	0,217	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	24	0,56
0,70	5	0,085	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	25	0,56
0,80	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
0,90	9	0,057	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	29	0,56
1,00	18	0,033	4	0,216	6	0,342	6	0,396	3	0,619	0	0,000	1	0,565	38	0,61

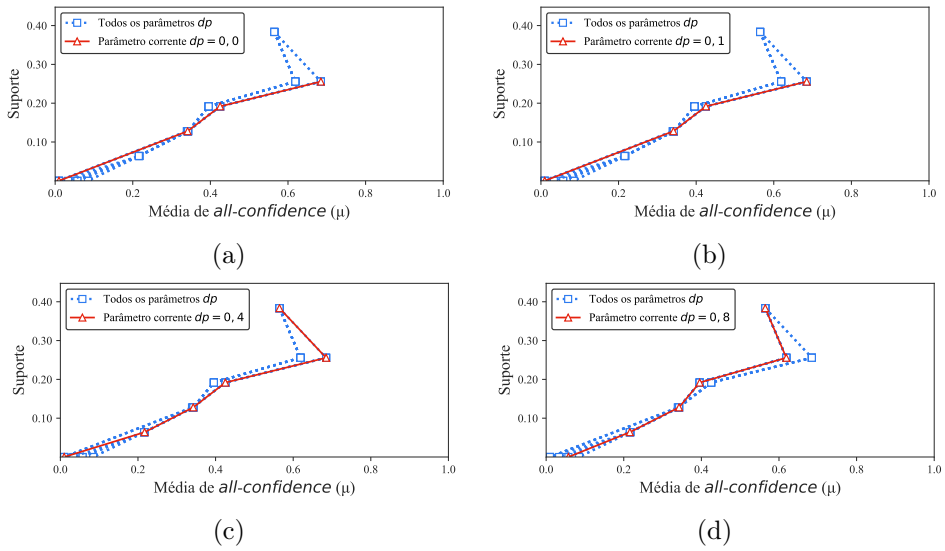


Figura C.21: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, (b) com $dr = 0{,}10$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}30\}$, (c) com $dr = 0{,}40$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\,, 0{,}60\,, 0{,}70\}$ e (d) com $dr = 0{,}80$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}90\,, 1{,}00\}$. Veja Tabela C.19 para detalhes.

Tabela C.20: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

	Partição de suporte														Itemset	Tempo
dr	[0,00 , 0,06]		(0,06 , 0,13]		(0,13 , 0,19]		(0,19 , 0,26]		(0,26 , 0,32]		(0,32 , 0,38]		(0,38 , 0,45]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	1	0,017	0	0,000	0	0,000	1	0,312	0	0,000	0	0,000	0	0,000	2	0,59
2	2	0,012	0	0,000	2	0,461	3	0,297	3	0,511	0	0,000	1	0,565	11	0,59
3	3	0,010	0	0,000	5	0,370	5	0,391	6	0,495	0	0,000	1	0,565	20	0,74
4	4	0,009	0	0,000	8	0,343	8	0,374	7	0,472	0	0,000	1	0,565	28	0,76
5	**6**	**0,019**	**1**	**0,269**	**9**	**0,347**	**9**	**0,365**	**7**	**0,472**	**0**	**0,000**	**1**	**0,565**	**33**	**0,79**
6	8	0,024	3	0,261	10	0,335	11	0,348	7	0,472	0	0,000	1	0,565	40	0,81
7	13	0,045	4	0,249	11	0,337	11	0,348	7	0,472	0	0,000	1	0,565	47	0,82
8	19	0,058	6	0,234	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	56	0,83
9	25	0,058	6	0,234	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	62	0,84
10	31	0,057	7	0,214	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	69	0,84
11	37	0,055	7	0,214	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	75	0,89
12	42	0,053	7	0,214	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	80	0,87
13	48	0,052	7	0,214	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	86	0,89
14	52	0,050	7	0,214	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	90	0,87
15	58	0,047	7	0,214	12	0,327	11	0,348	7	0,472	0	0,000	1	0,565	96	0,86

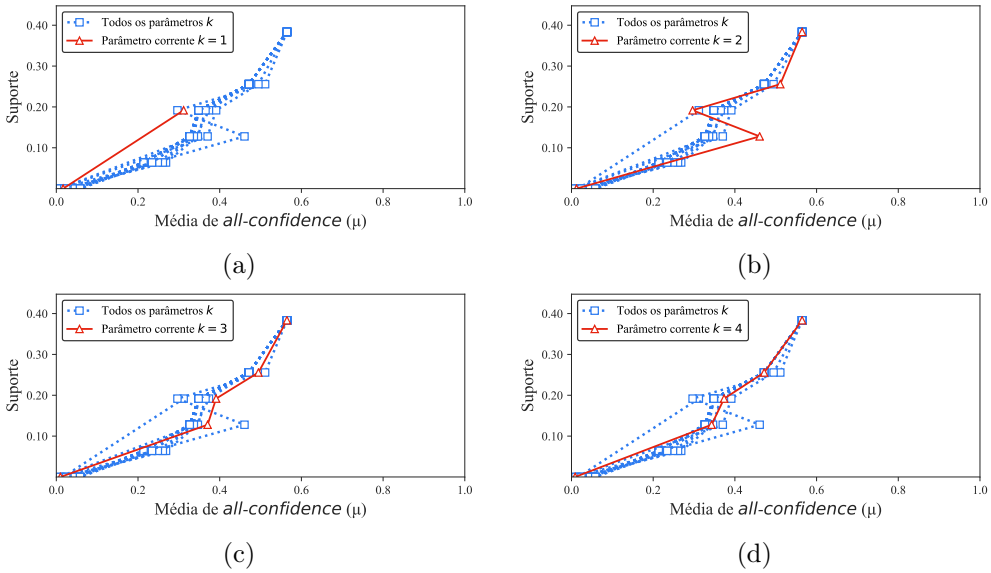


Figura C.22: *Skin*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$, (c) com $k = 3$ e (d) com $k = 4$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.20 para detalhes.

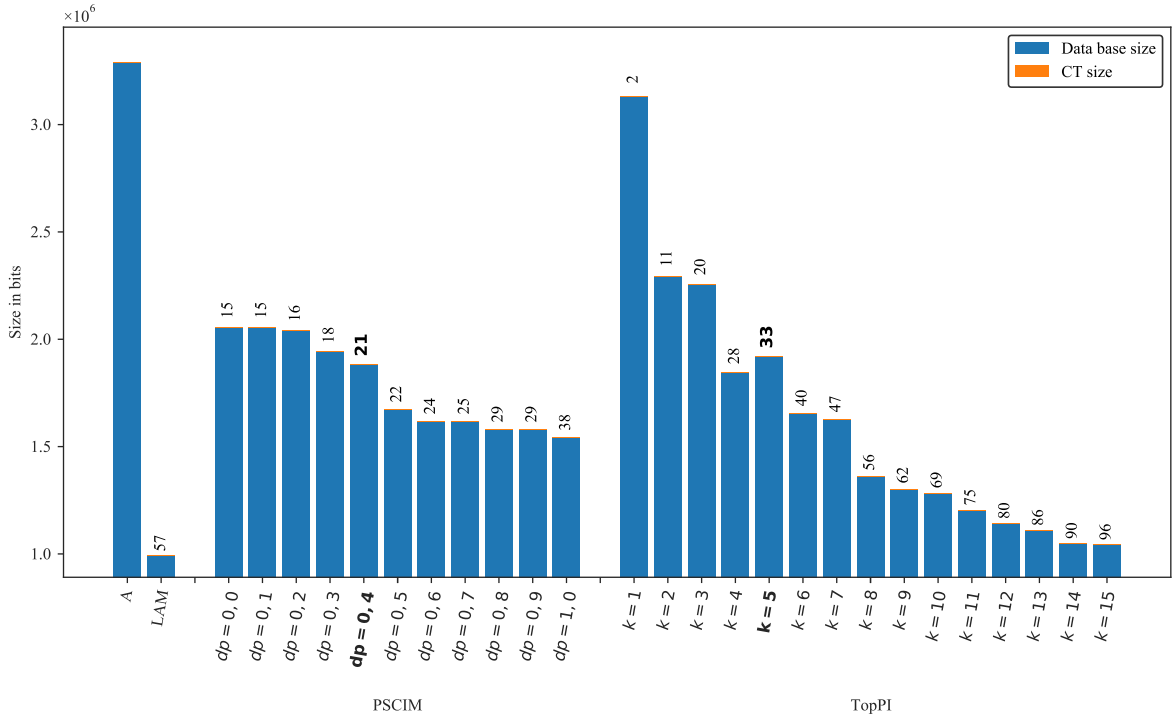
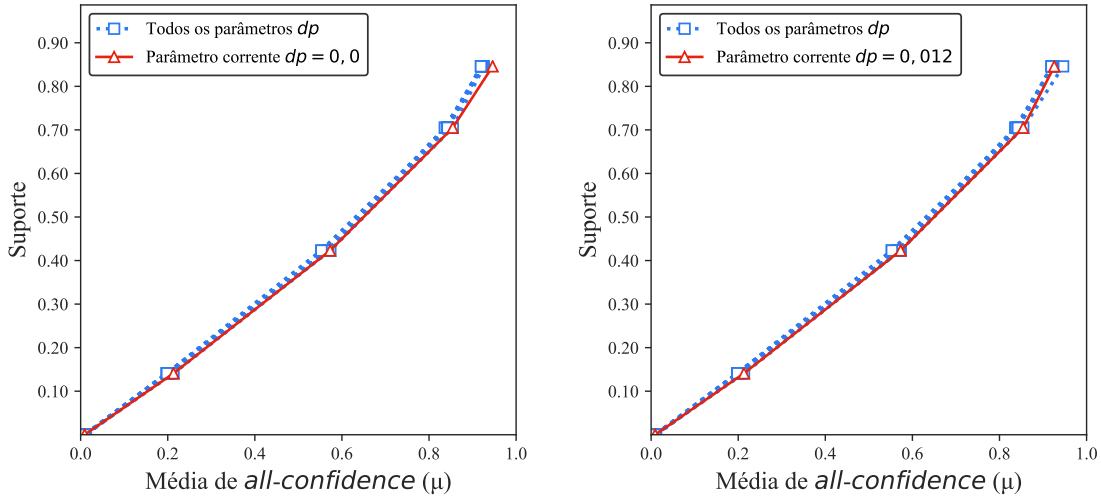


Figura C.23: *Skin*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.1.8 Susy

Tabela C.21: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,85]		(0,85 , 0,99]			
♣ × 10^{-1}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	**2.702**	**0,009**	**24**	**0,213**	**0**	**0,000**	**8**	**0,573**	**0**	**0,000**	**9**	**0,855**	**18**	**0,946**	**2.761**	**252,10**
0,12 ♣	2.704	0,009	24	0,213	0	0,000	8	0,573	0	0,000	9	0,855	25	0,926	2.770	251,93
0,13 ♣	2.719	0,010	28	0,206	0	0,000	8	0,573	0	0,000	15	0,841	26	0,924	2.796	251,98
0,14 ♣	2.791	0,011	30	0,202	0	0,000	9	0,566	0	0,000	15	0,841	27	0,922	2.872	253,79
0,15 ♣	2.874	0,012	39	0,208	0	0,000	16	0,567	0	0,000	18	0,837	27	0,922	2.974	255,64
0,16 ♣	3.199	0,011	48	0,198	0	0,000	16	0,567	0	0,000	18	0,837	27	0,922	3.308	255,63
0,17 ♣	3.274	0,010	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.386	256,11
0,18 ♣	3.346	0,010	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.458	255,66
0,21 ♣	3.729	0,009	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	3.841	256,92
0,22 ♣	3.928	0,009	48	0,198	0	0,000	16	0,567	0	0,000	20	0,842	28	0,922	4.040	257,49
0,23 ♣	3.935	0,009	48	0,198	0	0,000	26	0,553	0	0,000	25	0,842	35	0,919	4.069	258,15
0,24 ♣	3.956	0,009	48	0,198	0	0,000	26	0,553	0	0,000	25	0,842	35	0,919	4.090	258,49



(a) (b)

Figura C.24: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$ e (b) com $dr = 0{,}19$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{1{,}20\,, 1{,}30\,, 1{,}40\,, 1{,}50\,, 1{,}60\,, 1{,}70\,, 1{,}80\,, 2{,}10\,, 2{,}20\,, 2{,}30\,, 2{,}40\}$. Veja Tabela C.21 para detalhes.

Tabela C.22: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte														Itemset	Tempo
	[0,00 , 0,14]		(0,14 , 0,28]		(0,28 , 0,42]		(0,42 , 0,56]		(0,56 , 0,70]		(0,70 , 0,85]		(0,85 , 0,99]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	43	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	43	11,66
2	**190**	**0,050**	**9**	**0,185**	**1**	**0,339**	**2**	**0,546**	**0**	**0,000**	**1**	**0,939**	**8**	**0,934**	**211**	**12,03**
3	338	0,049	18	0,185	2	0,338	4	0,550	0	0,000	2	0,881	15	0,930	379	15,60
4	482	0,050	27	0,185	3	0,338	6	0,548	0	0,000	3	0,862	22	0,930	543	18,77
5	624	0,050	34	0,186	4	0,343	8	0,544	0	0,000	5	0,862	27	0,931	702	20,69
6	769	0,049	41	0,187	5	0,345	10	0,541	0	0,000	7	0,858	32	0,929	864	22,61
7	913	0,049	48	0,187	6	0,346	12	0,538	0	0,000	9	0,852	36	0,928	1.024	26,44
8	1.054	0,049	55	0,188	7	0,349	14	0,536	0	0,000	11	0,853	41	0,926	1.182	22,96
9	1.193	0,049	62	0,188	8	0,347	16	0,538	0	0,000	13	0,850	46	0,926	1.338	26,52
10	1.332	0,048	69	0,188	9	0,345	18	0,538	0	0,000	15	0,850	49	0,925	1.492	37,98
11	1.473	0,048	76	0,188	10	0,345	20	0,539	0	0,000	17	0,848	53	0,923	1.649	33,91
12	1.615	0,048	83	0,188	11	0,343	22	0,538	0	0,000	19	0,845	57	0,921	1.807	43,78
13	1.754	0,048	90	0,189	12	0,342	24	0,537	0	0,000	21	0,845	61	0,919	1.962	44,34
14	1.895	0,048	97	0,189	13	0,341	26	0,536	0	0,000	23	0,843	65	0,918	2.119	44,45
15	2.036	0,047	104	0,188	14	0,340	28	0,534	0	0,000	25	0,841	69	0,917	2.276	39,99
16	2.174	0,047	111	0,188	15	0,339	30	0,533	0	0,000	27	0,839	73	0,917	2.430	41,67
17	2.311	0,048	118	0,188	16	0,340	32	0,533	0	0,000	29	0,840	76	0,916	2.582	47,74
18	2.451	0,047	125	0,187	17	0,339	34	0,532	0	0,000	31	0,839	79	0,915	2.737	54,15
19	2.592	0,047	132	0,187	18	0,338	36	0,532	0	0,000	33	0,837	82	0,913	2.893	46,99
20	2.730	0,047	139	0,187	19	0,338	38	0,531	0	0,000	35	0,836	85	0,912	3.046	52,51
30	4.119	0,047	194	0,187	29	0,336	58	0,528	0	0,000	55	0,833	107	0,907	4.562	75,61
40	5.494	0,047	234	0,189	39	0,333	78	0,524	0	0,000	82	0,833	125	0,903	6.052	86,93
50	6.839	0,047	274	0,190	49	0,332	98	0,522	0	0,000	111	0,831	135	0,901	7.506	93,93

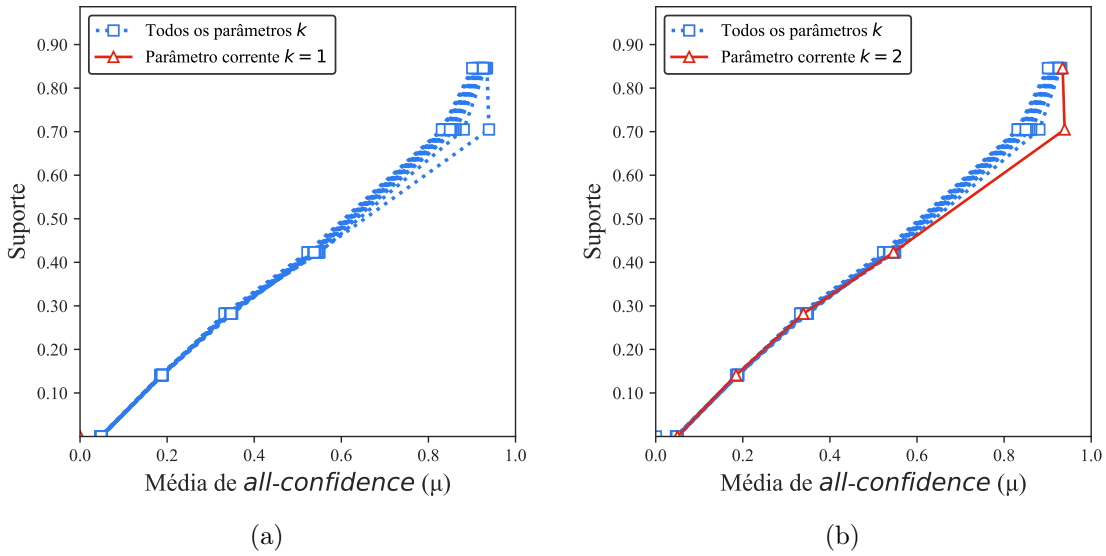


Figura C.25: *Susy*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50\}$. Veja Tabela C.22 para detalhes.

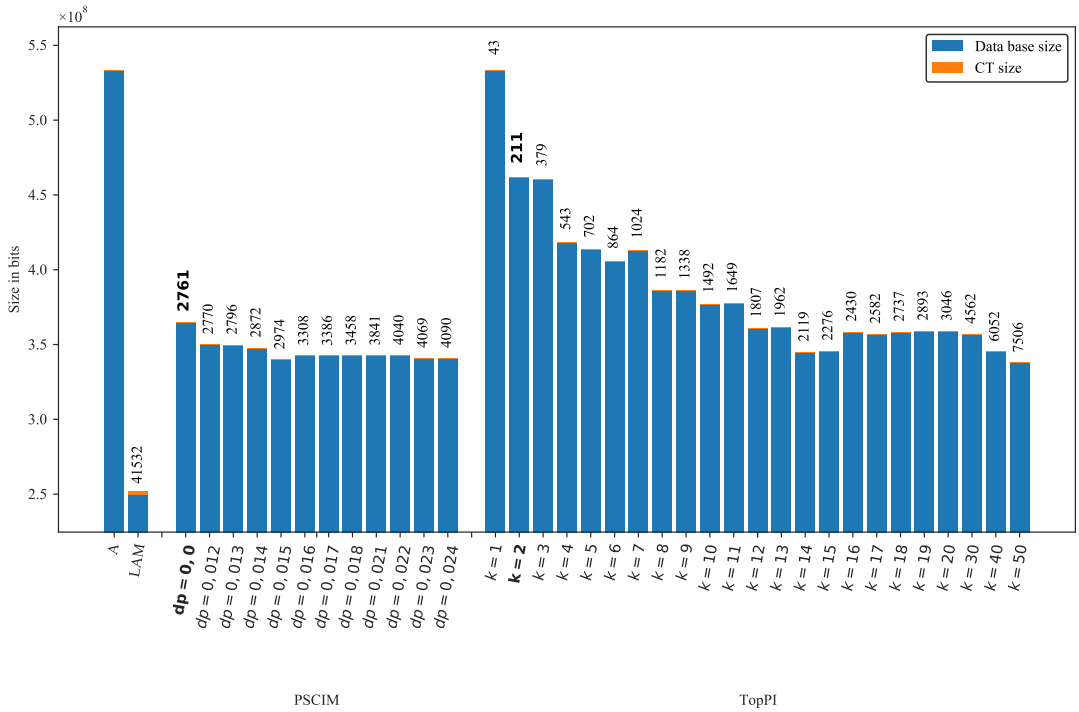


Figura C.26: *Susy*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2 Bases de Dados esparsa

C.1.2.1 Accidents

Tabela C.23: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]			
$\clubsuit \times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	**2.222**	**0,026**	**104**	**0,221**	**2**	**0,411**	**151**	**0,520**	**77**	**0,624**	**12**	**0,785**	**8**	**0,956**	**2.576**	**111,12**
0,05 ♣	2.222	0,026	106	0,221	2	0,411	175	0,514	83	0,627	15	0,788	15	0,941	2.618	111,64
0,06 ♣	2.396	0,030	127	0,218	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.027	113,82
0,07 ♣	2.581	0,032	180	0,206	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.265	115,46
0,08 ♣	2.732	0,033	196	0,202	69	0,350	309	0,507	94	0,635	17	0,786	15	0,941	3.432	116,62
0,09 ♣	2.844	0,032	196	0,202	69	0,350	309	0,507	96	0,636	19	0,791	17	0,934	3.550	116,61
0,10 ♣	3.657	0,025	212	0,203	70	0,351	346	0,505	104	0,637	27	0,797	22	0,925	4.438	118,45
0,11 ♣	3.913	0,025	253	0,203	195	0,368	598	0,498	126	0,636	32	0,793	22	0,925	5.139	120,26
0,12 ♣	4.224	0,028	257	0,203	195	0,368	598	0,498	133	0,639	39	0,784	22	0,925	5.468	123,24
0,13 ♣	4.491	0,029	306	0,201	195	0,368	598	0,498	135	0,640	48	0,785	22	0,925	5.795	121,78

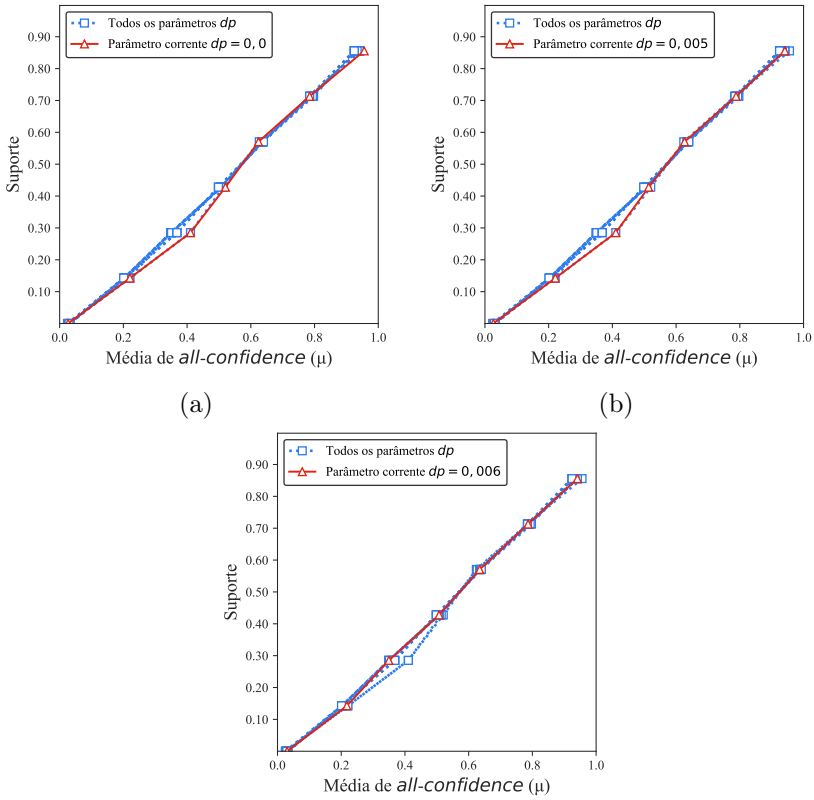
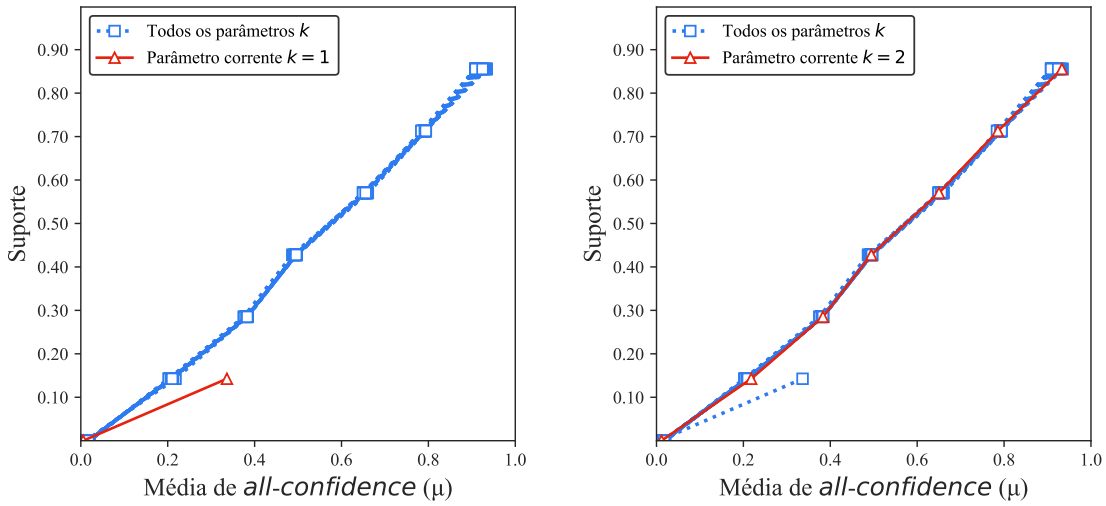


Figura C.27: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$, e (c) com $dr = 0{,}06 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\}$. Veja Tabela C.23 para detalhes.

Tabela C.24: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

	Partição de suporte														Itemset	Tempo
dr	[0,00 , 0,14]		(0,14 , 0,29]		(0,29 , 0,43]		(0,43 , 0,57]		(0,57 , 0,71]		(0,71 , 0,86]		(0,86 , 1,00]		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	321	0,005	3	0,336	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	324	1,91
2	**634**	**0,011**	**33**	**0,218**	**4**	**0,383**	**6**	**0,494**	**7**	**0,650**	**7**	**0,786**	**7**	**0,933**	**698**	**1,85**
3	947	0,014	63	0,212	8	0,383	12	0,494	14	0,650	15	0,791	12	0,934	1.071	2,75
4	1.244	0,015	93	0,210	12	0,383	18	0,494	21	0,650	23	0,792	17	0,934	1.428	3,01
5	1.538	0,016	123	0,209	16	0,383	24	0,494	28	0,650	31	0,792	21	0,932	1.781	3,24
6	1.834	0,016	153	0,209	20	0,383	30	0,494	35	0,650	39	0,792	25	0,929	2.136	3,51
7	2.128	0,016	183	0,208	24	0,383	36	0,494	42	0,650	47	0,792	29	0,928	2.489	4,00
8	2.411	0,017	213	0,208	28	0,383	42	0,494	49	0,650	55	0,792	33	0,927	2.831	3,92
9	2.694	0,017	243	0,210	32	0,383	49	0,496	56	0,655	62	0,794	36	0,925	3.172	4,54
10	2.977	0,017	273	0,209	36	0,382	56	0,496	63	0,656	69	0,793	39	0,921	3.513	4,42
11	3.259	0,018	303	0,208	40	0,382	63	0,496	70	0,656	76	0,793	42	0,918	3.853	4,75
12	3.541	0,018	333	0,208	44	0,381	70	0,496	77	0,657	83	0,792	45	0,916	4.193	4,95
13	3.823	0,018	363	0,207	48	0,381	77	0,496	84	0,657	90	0,792	48	0,914	4.533	5,22
14	4.102	0,018	393	0,207	52	0,381	84	0,496	91	0,657	97	0,791	51	0,912	4.870	5,46
15	4.379	0,018	423	0,206	56	0,380	91	0,496	98	0,657	104	0,791	54	0,910	5.205	5,75
16	4.649	0,018	453	0,206	60	0,380	98	0,495	105	0,657	111	0,791	57	0,908	5.533	6,12
17	4.922	0,019	481	0,207	64	0,381	105	0,496	113	0,661	119	0,794	57	0,908	5.861	6,33
18	5.195	0,019	509	0,206	68	0,380	112	0,495	121	0,660	127	0,794	57	0,908	6.189	6,57
19	5.468	0,019	537	0,206	72	0,380	119	0,494	129	0,659	135	0,793	57	0,908	6.517	6,91
20	5.741	0,019	565	0,206	76	0,379	126	0,493	137	0,659	143	0,793	57	0,908	6.845	7,17
30	8.459	0,019	843	0,204	116	0,375	196	0,487	217	0,654	223	0,787	57	0,908	10.111	10,44
40	11.181	0,020	1.105	0,202	165	0,376	273	0,488	309	0,657	267	0,785	57	0,908	13.357	13,49
50	13.902	0,020	1.365	0,202	215	0,381	345	0,491	407	0,658	295	0,784	57	0,908	16.586	16,87



(a) (b)

Figura C.28: *Accidents*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50\}$. Veja Tabela C.24 para detalhes.

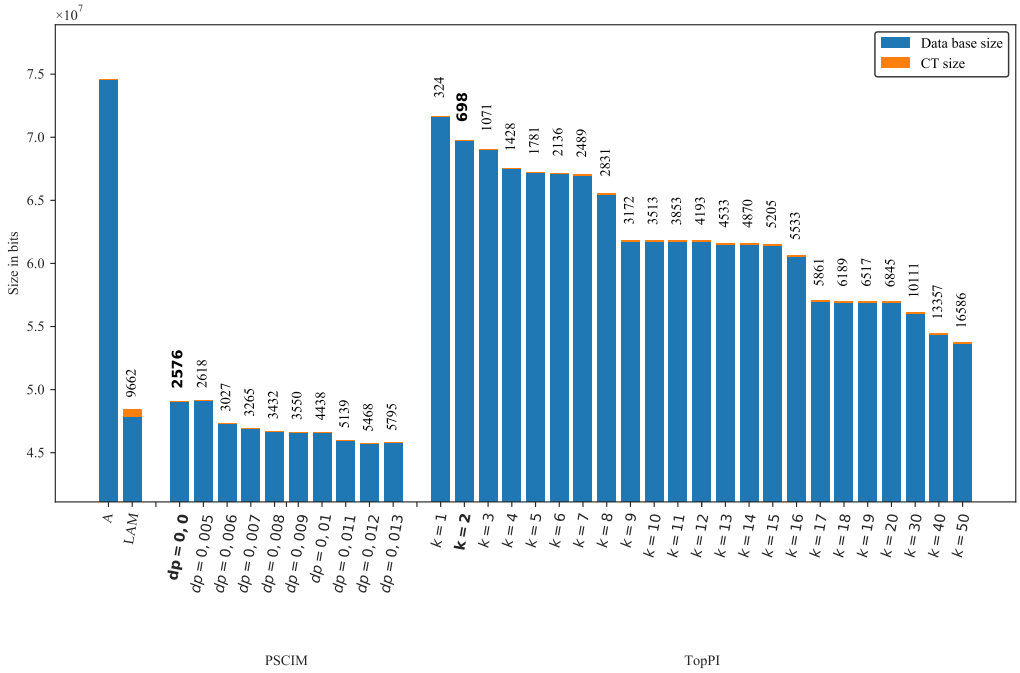


Figura C.29: *Accidents*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e*, o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2.2 BMSWebView2

Tabela C.25: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte † × 10^{-1}														Total de itemsets	Tempo (s)
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,08] †		(0,08 , 0,11] †		(0,11 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,19] †			
♣ × 10^{-2}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	4.119	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.159	238,99
0,05 ♣	4.119	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.159	238,99
0,10 ♣	4.119	0,166	32	0,370	6	0,427	2	0,344	0	0,000	0	0,000	0	0,000	4.159	240,32
0,15 ♣	**4.163**	**0,165**	**47**	**0,343**	**12**	**0,380**	**5**	**0,356**	**1**	**0,414**	**1**	**0,309**	**2**	**0,492**	**4.231**	**241,63**
0,16 ♣	4.191	0,165	54	0,344	12	0,380	5	0,356	1	0,414	1	0,309	2	0,492	4.266	239,07
0,17 ♣	4.207	0,165	57	0,341	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.286	238,72
0,18 ♣	4.271	0,163	58	0,340	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.351	239,02
0,19 ♣	4.319	0,162	60	0,338	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.401	239,10
0,20 ♣	4.387	0,161	60	0,338	13	0,384	5	0,356	1	0,414	1	0,309	2	0,492	4.469	238,82
0,30 ♣	5.278	0,146	73	0,328	24	0,329	9	0,309	5	0,338	1	0,309	3	0,454	5.393	239,21
0,40 ♣	6.237	0,136	102	0,303	36	0,294	14	0,279	9	0,326	2	0,313	3	0,454	6.403	239,31
0,50 ♣	7.522	0,123	169	0,256	50	0,264	18	0,287	9	0,326	2	0,313	3	0,454	7.773	239,52

C.1.2.3 BMS1

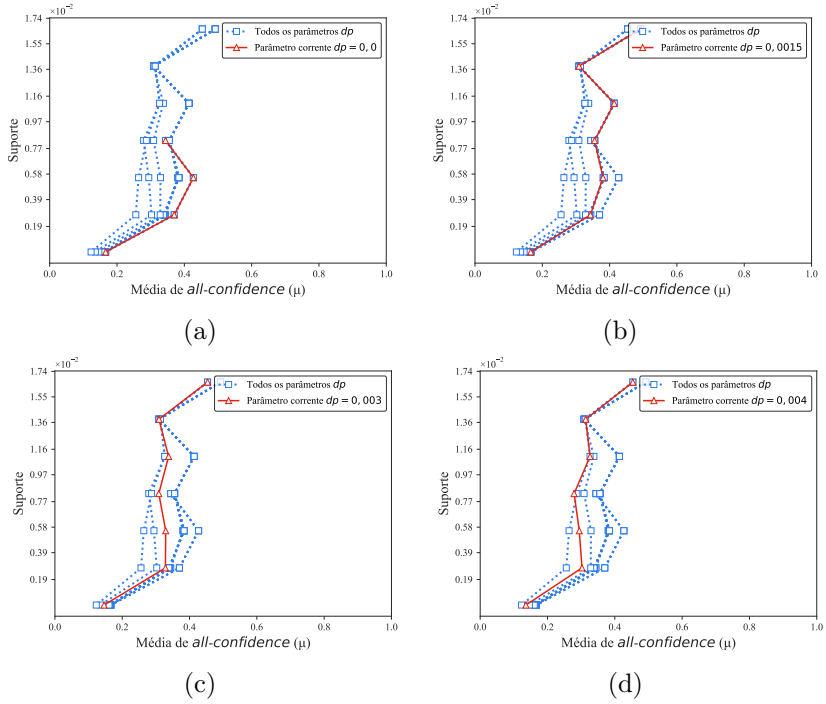


Figura C.30: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-2}$, (b) com $dr = 0{,}15 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}19\,, 0{,}20\}$ (c) com $dr = 0{,}30 \times 10^{-2}$, e (d) com $dr = 0{,}40 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}50\}$. Veja Tabela C.25 para detalhes.

Tabela C.26: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte † $\times 10^{-1}$														Itemset	Tempo
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,08] †		(0,08 , 0,11] †		(0,11 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,19] †		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	90	0,095	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	90	0,69
2	**2.778**	**0,178**	**50**	**0,306**	**17**	**0,346**	**4**	**0,284**	**3**	**0,324**	**2**	**0,330**	**2**	**0,492**	**2.856**	**0,79**
3	5.351	0,153	81	0,296	24	0,333	7	0,258	7	0,324	3	0,323	3	0,454	5.476	0,90
4	7.794	0,140	105	0,296	31	0,303	10	0,288	10	0,301	3	0,323	3	0,454	7.956	1,01
5	10.208	0,128	125	0,285	37	0,291	14	0,284	12	0,309	3	0,323	3	0,454	10.402	0,98
6	12.580	0,119	143	0,278	43	0,283	17	0,282	12	0,309	3	0,323	3	0,454	12.801	1,04
7	14.900	0,112	159	0,271	48	0,275	19	0,287	12	0,309	3	0,323	3	0,454	15.144	1,05
8	17.206	0,106	174	0,266	52	0,268	22	0,281	12	0,309	3	0,323	3	0,454	17.472	1,05
9	19.449	0,101	188	0,260	57	0,262	26	0,275	12	0,309	3	0,323	3	0,454	19.738	1,05
10	21.698	0,096	203	0,255	64	0,259	27	0,272	12	0,309	3	0,323	3	0,454	22.010	1,15
11	23.885	0,092	222	0,249	68	0,257	29	0,266	12	0,309	3	0,323	3	0,454	24.222	1,13
12	26.064	0,089	242	0,242	72	0,253	29	0,266	12	0,309	3	0,323	3	0,454	26.425	1,11
13	28.211	0,086	259	0,237	74	0,251	29	0,266	12	0,309	3	0,323	3	0,454	28.591	1,17
14	30.344	0,083	273	0,235	79	0,245	29	0,266	12	0,309	3	0,323	3	0,454	30.743	1,13
15	32.497	0,081	288	0,231	83	0,241	29	0,266	12	0,309	3	0,323	3	0,454	32.915	1,21
16	34.575	0,078	301	0,227	86	0,238	29	0,266	12	0,309	3	0,323	3	0,454	35.009	1,28
17	36.615	0,076	314	0,223	89	0,236	29	0,266	12	0,309	3	0,323	3	0,454	37.065	1,33
18	38.695	0,074	327	0,220	94	0,231	29	0,266	12	0,309	3	0,323	3	0,454	39.163	1,34
19	40.724	0,073	340	0,217	98	0,228	29	0,266	12	0,309	3	0,323	3	0,454	41.209	1,35
20	42.774	0,071	349	0,215	100	0,226	29	0,266	12	0,309	3	0,323	3	0,454	43.270	1,37

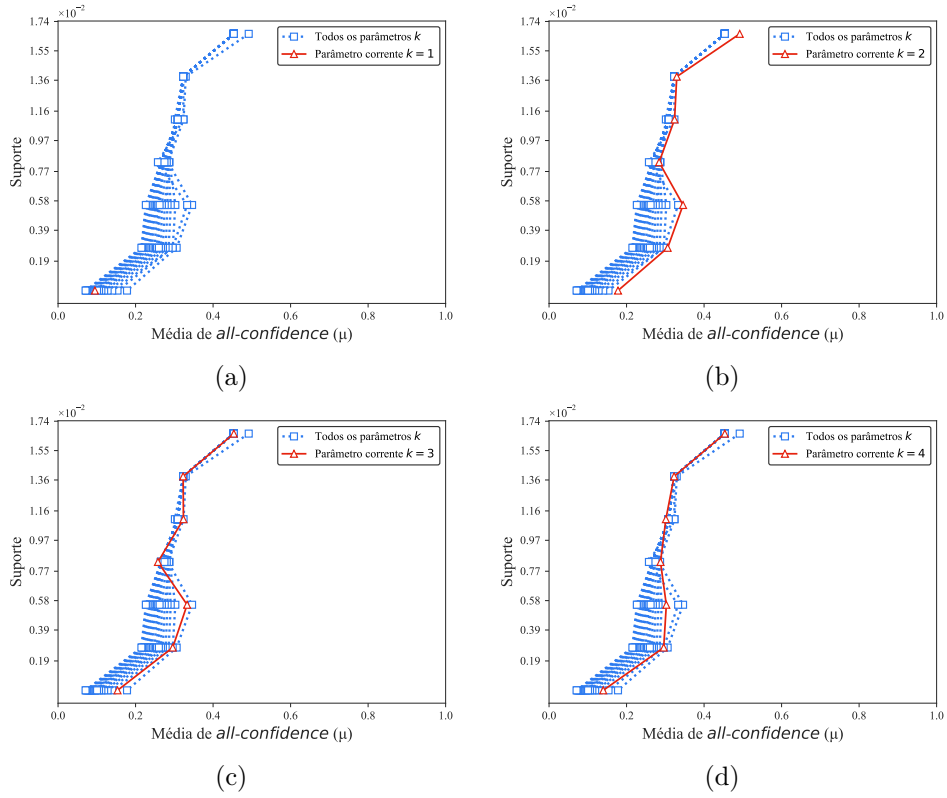


Figura C.31: *BMSWebView2*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$, (c) com $k = 3$ e (d) com $k = 4$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.26 para detalhes.

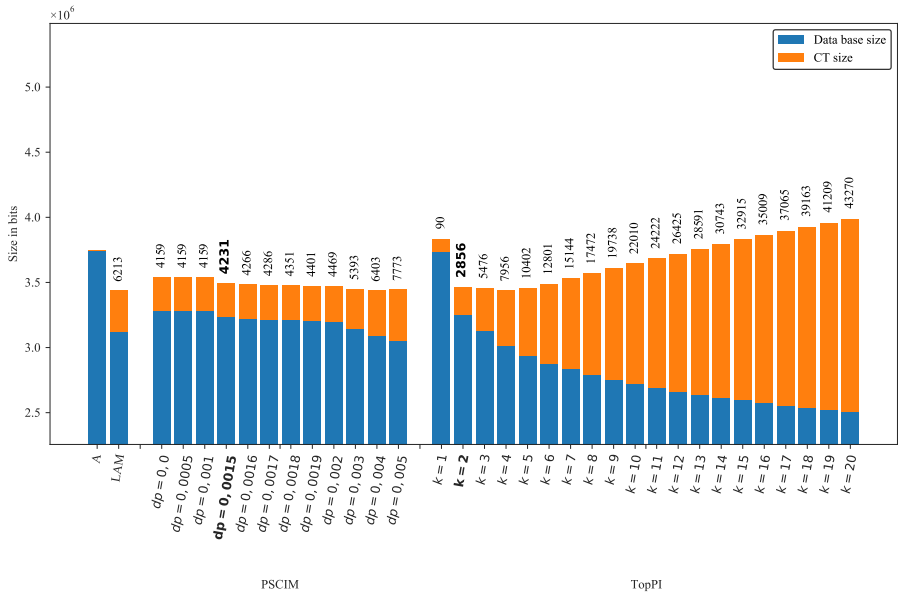


Figura C.32: *BMSWebView2*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

Tabela C.27: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte † $\times 10^{-1}$														Total de itemsets	Tempo (s)
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,09] †		(0,09 , 0,12] †		(0,12 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,20] †			
♣ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,19
0,05 ♣	94	0,192	6	0,358	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	101	3,25
0,06 ♣	144	0,164	9	0,341	0	0,000	0	0,000	0	0,000	0	0,000	1	0,329	154	3,28
0,07 ♣	310	0,147	22	0,270	2	0,200	3	0,210	0	0,000	1	0,387	1	0,329	339	3,32
0,08 ♣	**363**	**0,140**	**24**	**0,264**	**3**	**0,235**	**3**	**0,210**	**2**	**0,260**	**2**	**0,357**	**1**	**0,329**	**398**	**3,34**
0,09 ♣	372	0,139	24	0,264	3	0,235	3	0,210	2	0,260	2	0,357	1	0,329	407	3,34
0,10 ♣	379	0,138	24	0,264	3	0,235	3	0,210	2	0,260	2	0,357	1	0,329	414	3,34
0,15 ♣	886	0,102	31	0,246	3	0,235	6	0,241	2	0,260	2	0,357	1	0,329	931	3,47
0,20 ♣	1.390	0,086	36	0,241	6	0,178	6	0,241	3	0,248	2	0,357	1	0,329	1.444	3,61
0,25 ♣	1.972	0,074	45	0,224	6	0,178	7	0,244	3	0,248	2	0,357	1	0,329	2.036	3,80
0,30 ♣	2.668	0,066	49	0,214	9	0,173	8	0,241	3	0,248	2	0,357	1	0,329	2.740	4,02

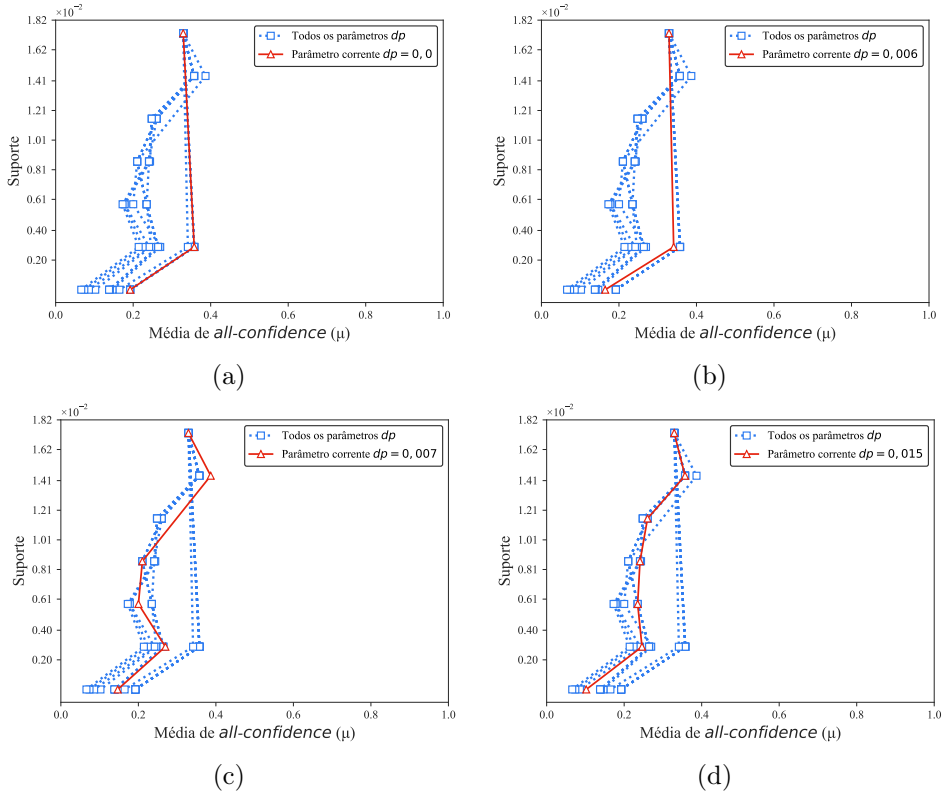


Figura C.33: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}05\}$, (b) com $dr = 0{,}06 \times 10^{-1}$, (c) com $dr = 0{,}07 \times 10^{-1}$,onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}08\,, 0{,}09\,, 0{,}10\}$, e (d) com $dr = 0{,}15 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}20\,, 0{,}25\,, 0{,}30\}$. Veja Tabela C.27 para detalhes.

Tabela C.28: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte † × 10^{-1}														Itemset #	Tempo (s)
	[0,00 , 0,03] †		(0,03 , 0,06] †		(0,06 , 0,09] †		(0,09 , 0,12] †		(0,12 , 0,14] †		(0,14 , 0,17] †		(0,17 , 0,20] †			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	26	0,152	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	26	0,49
2	**356**	**0,172**	**39**	**0,240**	**9**	**0,171**	**5**	**0,240**	**1**	**0,224**	**2**	**0,357**	**1**	**0,329**	**413**	**0,49**
3	691	0,153	64	0,207	13	0,175	6	0,237	3	0,248	2	0,357	1	0,329	780	0,53
4	1.007	0,144	81	0,190	18	0,163	8	0,241	3	0,248	2	0,357	1	0,329	1.120	0,54
5	1.331	0,136	95	0,178	19	0,162	8	0,241	3	0,248	2	0,357	1	0,329	1.459	0,55
6	1.646	0,127	110	0,166	21	0,157	8	0,241	3	0,248	2	0,357	1	0,329	1.791	0,55
7	1.975	0,120	118	0,159	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	2.130	0,57
8	2.266	0,115	128	0,154	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	2.431	0,59
9	2.576	0,110	137	0,150	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	2.750	0,62
10	2.883	0,106	144	0,147	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	3.064	0,67
11	3.181	0,103	151	0,145	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	3.369	0,66
12	3.478	0,100	157	0,142	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	3.672	0,70
13	3.780	0,097	163	0,140	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	3.980	0,70
14	4.072	0,095	167	0,139	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	4.276	0,70
15	4.378	0,092	171	0,137	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	4.586	0,72
16	4.686	0,090	172	0,137	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	4.895	0,79
17	4.972	0,088	173	0,137	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	5.182	0,78
18	5.249	0,086	175	0,136	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	5.461	0,80
19	5.534	0,084	178	0,135	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	5.749	0,84
20	5.825	0,082	179	0,134	23	0,152	8	0,241	3	0,248	2	0,357	1	0,329	6.041	0,81

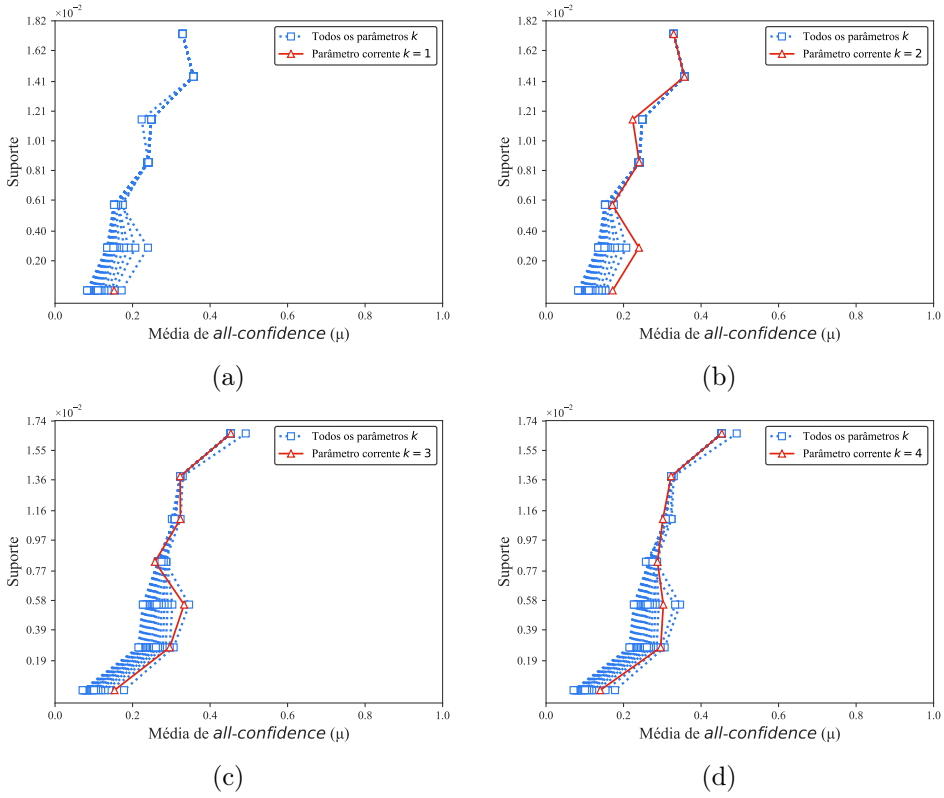


Figura C.34: *BMS1*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$, (c) com $k = 3$ e (d) com $k = 4$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.28 para detalhes.

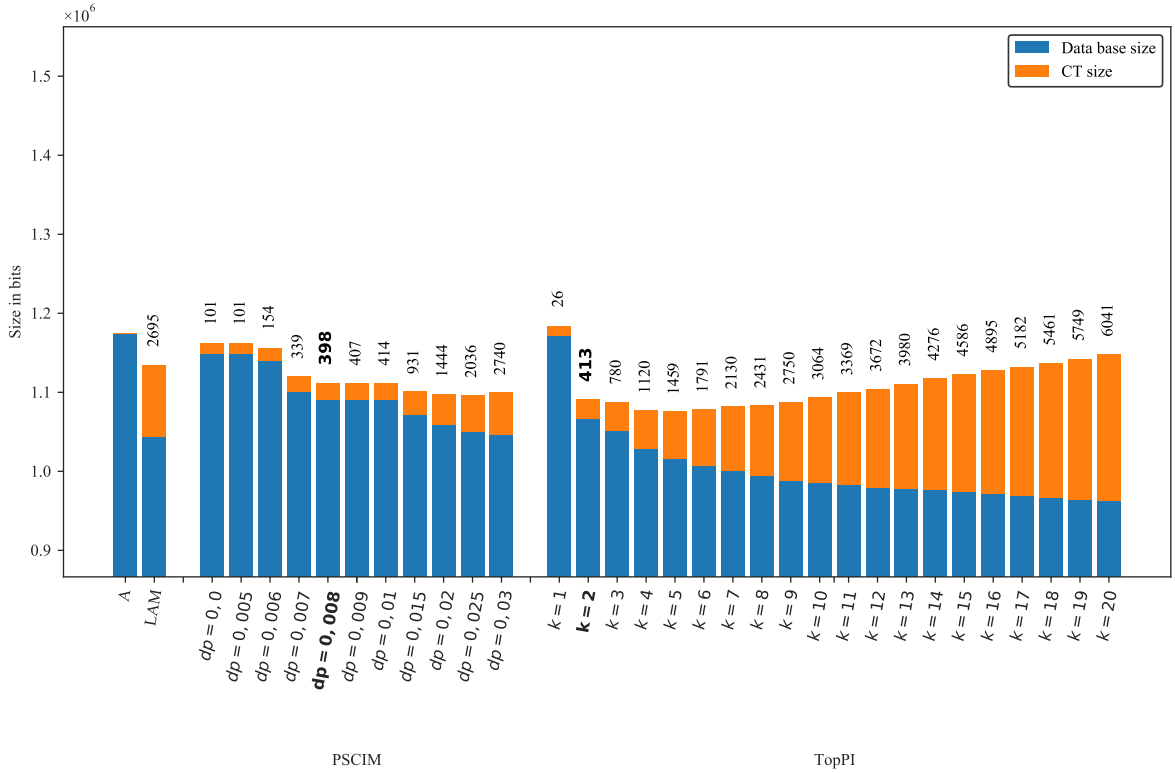


Figura C.35: *BMS1*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2.4 FoodmartFIM

Tabela C.29: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte † $\times 10^{-3}$														Total de itemsets	Tempo (s)
	[0,00 , 0,14] †		(0,14 , 0,28] †		(0,28 , 0,41] †		(0,41 , 0,55] †		(0,55 , 0,69] †		(0,69 , 0,83] †		(0,83 , 0,97] †			
♣ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	0	0,000	1	0,111	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	5	12,46
0,05 ♣	0	0,000	1	0,111	0	0,000	3	0,200	0	0,000	1	0,231	0	0,000	5	12,45
0,10 ♣	0	0,000	2	0,079	0	0,000	4	0,170	0	0,000	1	0,231	0	0,000	7	12,51
0,11 ♣	0	0,000	3	0,067	0	0,000	7	0,137	0	0,000	1	0,231	0	0,000	11	12,50
0,12 ♣	0	0,000	14	0,057	0	0,000	12	0,121	0	0,000	3	0,180	0	0,000	29	12,48
0,13 ♣	0	0,000	43	0,058	0	0,000	30	0,116	0	0,000	5	0,186	1	0,211	79	12,47
0,14 ♣	0	0,000	74	0,059	0	0,000	51	0,117	0	0,000	11	0,181	1	0,211	137	12,48
0,15 ♣	0	0,000	116	0,061	0	0,000	88	0,122	0	0,000	14	0,180	2	0,211	220	12,46
0,16 ♣	0	0,000	166	0,063	0	0,000	126	0,126	0	0,000	21	0,182	2	0,211	315	12,48
0,17 ♣	0	0,000	234	0,066	0	0,000	169	0,129	0	0,000	23	0,184	2	0,211	428	12,48
0,18 ♣	0	0,000	297	0,068	0	0,000	196	0,131	0	0,000	26	0,189	2	0,211	521	12,49
0,19 ♣	0	0,000	357	0,070	0	0,000	218	0,133	0	0,000	29	0,189	2	0,211	606	12,52
0,20 ♣	0	0,000	437	0,073	0	0,000	253	0,136	0	0,000	31	0,193	2	0,211	723	12,52
0,30 ♣	**0**	**0,000**	**951**	**0,078**	**0**	**0,000**	**452**	**0,142**	**0**	**0,000**	**50**	**0,190**	**2**	**0,211**	**1.455**	**12,53**
0,40 ♣	0	0,000	1.469	0,078	0	0,000	559	0,142	0	0,000	57	0,191	2	0,211	2.087	12,57
0,50 ♣	0	0,000	2.026	0,077	0	0,000	642	0,140	0	0,000	60	0,190	2	0,211	2.730	12,56

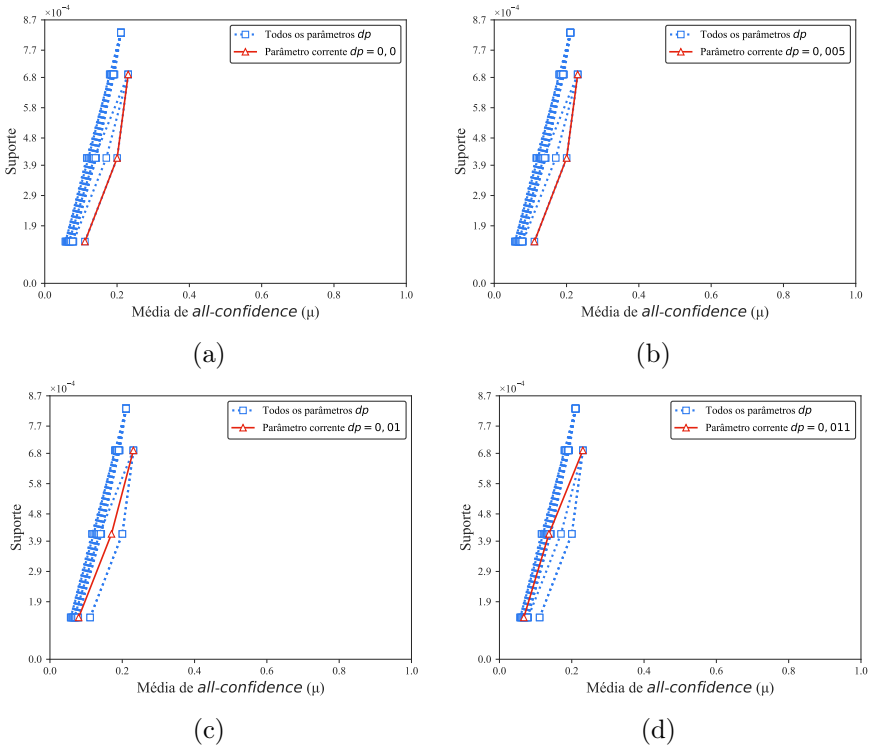
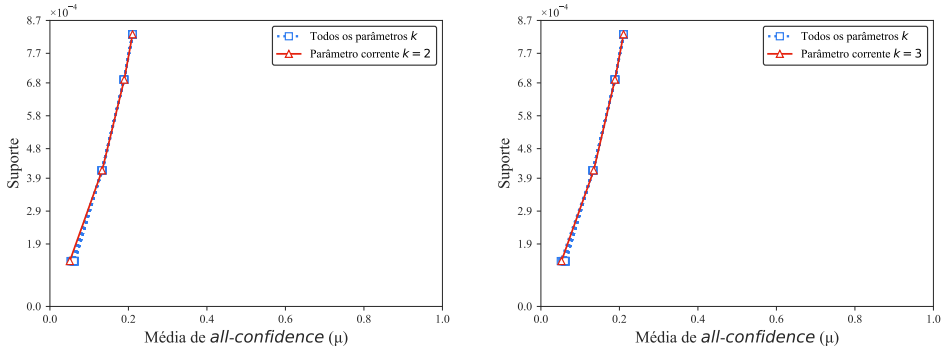


Figura C.36: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$, (c) com $dr = 0{,}10 \times 10^{-1}$ (d) com $dr = 0{,}11 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}12\,, 0{,}13\,, 0{,}14\,, 0{,}15\,, 0{,}16\,, 0{,}17\,, 0{,}18\,, 0{,}19\,, 0{,}20\,, 0{,}30\,, 0{,}40\,, 0{,}50\}$. Veja Tabela C.29 para detalhes.

Tabela C.30: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

	Partição de suporte † $\times 10^{-3}$														Itemset	Tempo
dr	[0,00 , 0,14] †		(0,14 , 0,28] †		(0,28 , 0,41] †		(0,41 , 0,55] †		(0,55 , 0,69] †		(0,69 , 0,83] †		(0,83 , 0,97] †		#	(s)
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
2	**0**	**0,000**	**247**	**0,051**	**0**	**0,000**	**709**	**0,133**	**0**	**0,000**	**69**	**0,189**	**2**	**0,211**	**1.027**	**0,51**
3	0	0,000	624	0,052	0	0,000	1.101	0,135	0	0,000	70	0,189	2	0,211	1.797	0,57
4	0	0,000	1.032	0,054	0	0,000	1.270	0,134	0	0,000	70	0,189	2	0,211	2.374	0,54
5	0	0,000	1.432	0,056	0	0,000	1.320	0,133	0	0,000	70	0,189	2	0,211	2.824	0,54
6	0	0,000	1.862	0,058	0	0,000	1.328	0,133	0	0,000	70	0,189	2	0,211	3.262	0,56
7	0	0,000	2.232	0,059	0	0,000	1.334	0,133	0	0,000	70	0,189	2	0,211	3.638	0,57
8	0	0,000	2.588	0,060	0	0,000	1.334	0,133	0	0,000	70	0,189	2	0,211	3.994	0,56
9	0	0,000	2.923	0,061	0	0,000	1.335	0,133	0	0,000	70	0,189	2	0,211	4.330	0,57
10	0	0,000	3.216	0,062	0	0,000	1.335	0,133	0	0,000	70	0,189	2	0,211	4.623	0,58
11	0	0,000	3.435	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	4.843	0,60
12	0	0,000	3.589	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	4.997	0,61
13	0	0,000	3.672	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.080	0,63
14	0	0,000	3.700	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.108	0,68
15	0	0,000	3.710	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.118	0,62
16	0	0,000	3.711	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.119	0,62
17	0	0,000	3.713	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.121	0,63
18	0	0,000	3.713	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.121	0,63
19	0	0,000	3.713	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.121	0,64
20	0	0,000	3.713	0,063	0	0,000	1.336	0,133	0	0,000	70	0,189	2	0,211	5.121	0,63



(a) (b)

Figura C.37: *FoodmartFIM*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 2$ e (b) com $k = 3$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.30 para detalhes.

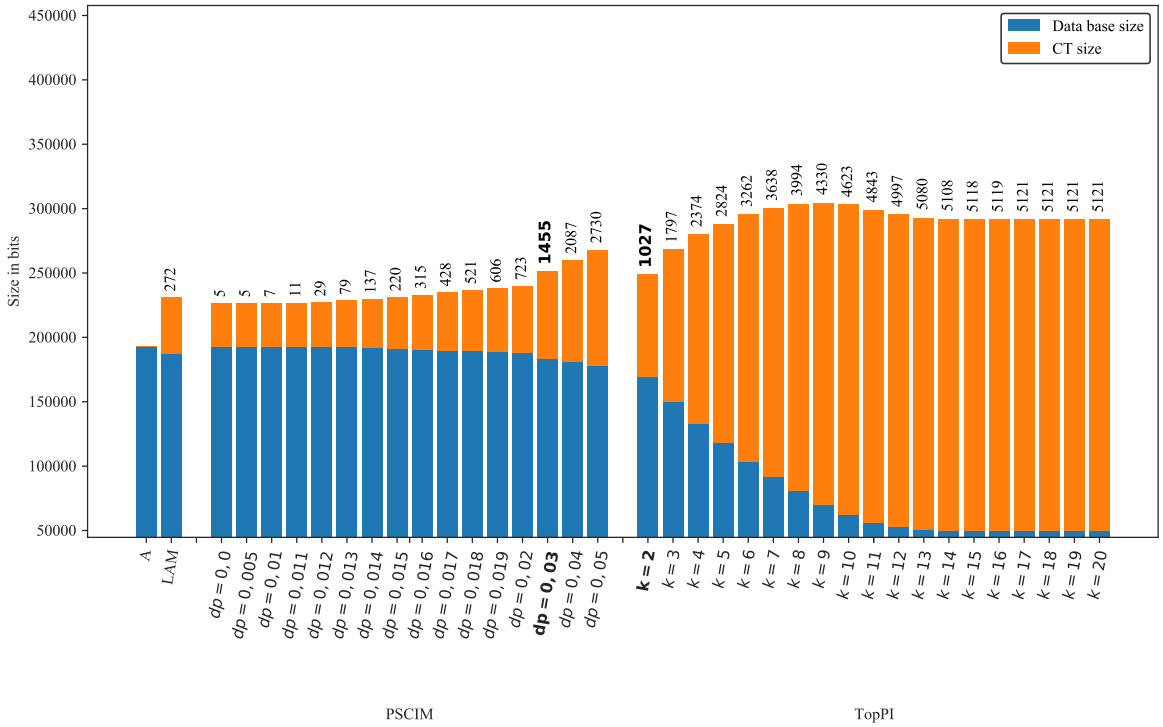
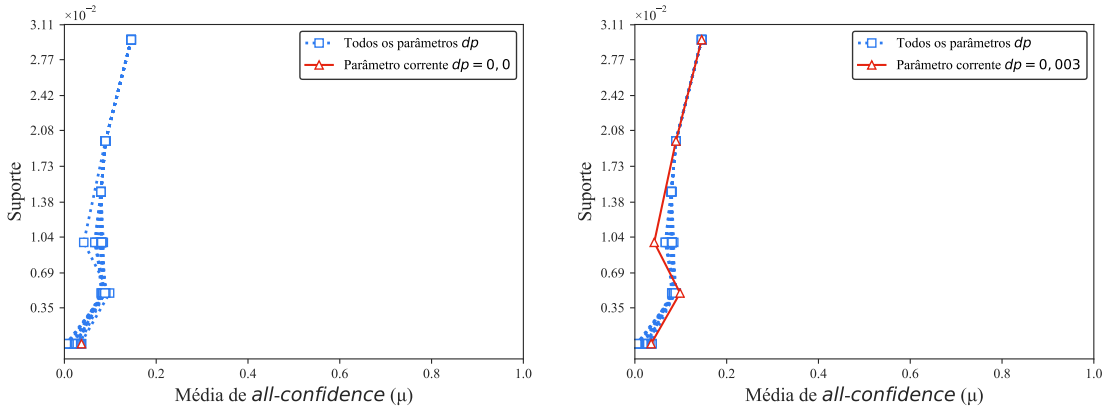


Figura C.38: *FoodmartFIM*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2.5 Fruithut

Tabela C.31: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte † × 10^{-1}														Total de itemsets	Tempo (s)
	[0,00 , 0,05] †		(0,05 , 0,10] †		(0,10 , 0,15] †		(0,15 , 0,20] †		(0,20 , 0,25] †		(0,25 , 0,30] †		(0,30 , 0,35] †			
♣ × 10^{-1}	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	298	0,037	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	298	21,93
0,01 ♣	298	0,037	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	298	22,04
0,02 ♣	298	0,037	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	298	22,34
0,03 ♣	336	0,035	11	0,098	1	0,042	0	0,000	2	0,089	0	0,000	1	0,146	351	22,09
0,04 ♣	**398**	**0,033**	**15**	**0,088**	**4**	**0,077**	**1**	**0,080**	**2**	**0,089**	**0**	**0,000**	**1**	**0,146**	**421**	**22,08**
0,05 ♣	497	0,029	17	0,080	5	0,071	1	0,080	2	0,089	0	0,000	1	0,146	523	22,11
0,06 ♣	612	0,026	18	0,085	6	0,066	1	0,080	2	0,089	0	0,000	1	0,146	640	22,06
0,07 ♣	731	0,023	19	0,088	6	0,066	1	0,080	2	0,089	0	0,000	1	0,146	760	22,06
0,08 ♣	891	0,021	20	0,085	8	0,085	1	0,080	2	0,089	0	0,000	1	0,146	923	22,07
0,09 ♣	1.059	0,019	24	0,081	9	0,080	1	0,080	2	0,089	0	0,000	1	0,146	1.096	22,07
0,10 ♣	1.213	0,018	26	0,084	9	0,080	1	0,080	2	0,089	0	0,000	1	0,146	1.252	22,05
0,20 ♣	3.288	0,010	31	0,085	10	0,081	1	0,080	2	0,089	0	0,000	1	0,146	3.333	22,16
0,30 ♣	6.377	0,007	37	0,089	10	0,081	1	0,080	2	0,089	0	0,000	1	0,146	6.428	22,37
0,40 ♣	11.498	0,005	40	0,090	10	0,081	1	0,080	2	0,089	0	0,000	1	0,146	11.552	22,63
0,50 ♣	17.394	0,004	46	0,088	10	0,081	1	0,080	2	0,089	0	0,000	1	0,146	17.454	22,93

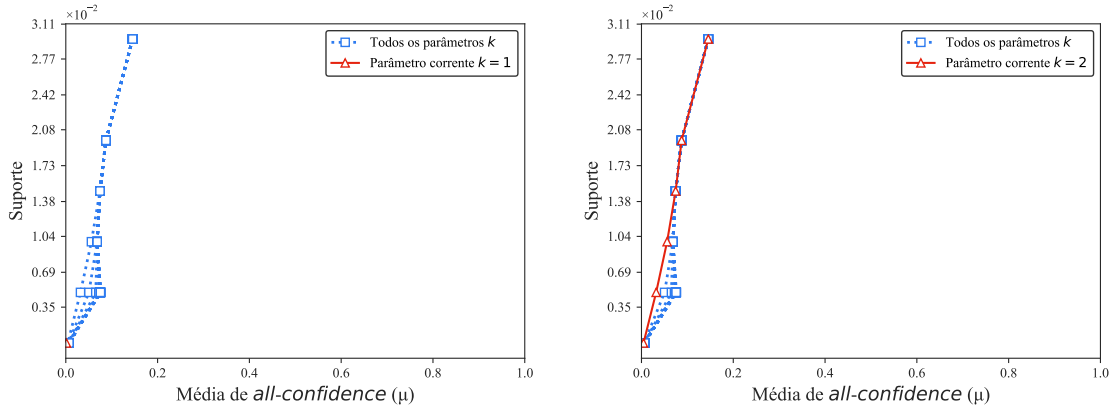


(a) (b)

Figura C.39: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}01\,,0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,,0{,}05\,,0{,}06\,,0{,}07\,,0{,}08\,,0{,}09\,,0{,}10\,,0{,}20\,,0{,}30\,,0{,}40\,,0{,}50\}$. Veja Tabela C.31 para detalhes.

Tabela C.32: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte † $\times 10^{-1}$														Itemset #	Tempo (s)
	[0,00 , 0,05] †		(0,05 , 0,10] †		(0,10 , 0,15] †		(0,15 , 0,20] †		(0,20 , 0,25] †		(0,25 , 0,30] †		(0,30 , 0,35] †			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	28	0,002	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	28	0,63
2	1.207	0,004	28	0,032	18	0,056	6	0,074	3	0,087	0	0,000	1	0,146	1.263	0,67
3	2.399	0,006	53	0,050	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	2.485	0,77
4	3.566	0,006	74	0,065	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	3.673	0,79
5	4.704	0,007	86	0,072	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	4.823	0,80
6	5.829	0,007	96	0,076	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	5.958	0,81
7	**6.957**	**0,007**	**99**	**0,076**	**23**	**0,068**	**6**	**0,074**	**3**	**0,087**	**0**	**0,000**	**1**	**0,146**	**7.089**	**0,83**
8	8.066	0,007	99	0,076	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	8.198	0,85
9	9.159	0,007	100	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	9.292	0,84
10	10.232	0,007	100	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	10.365	0,84
11	11.297	0,007	100	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	11.430	0,88
12	12.342	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	12.476	0,89
13	13.369	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	13.503	0,92
14	14.384	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	14.518	0,93
15	15.373	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	15.507	0,98
16	16.353	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	16.487	0,96
17	17.322	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	17.456	0,96
18	18.275	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	18.409	0,98
19	19.215	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	19.349	1,00
20	20.150	0,007	101	0,075	23	0,068	6	0,074	3	0,087	0	0,000	1	0,146	20.284	0,98



(a) (b)

Figura C.40: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.32 para detalhes.

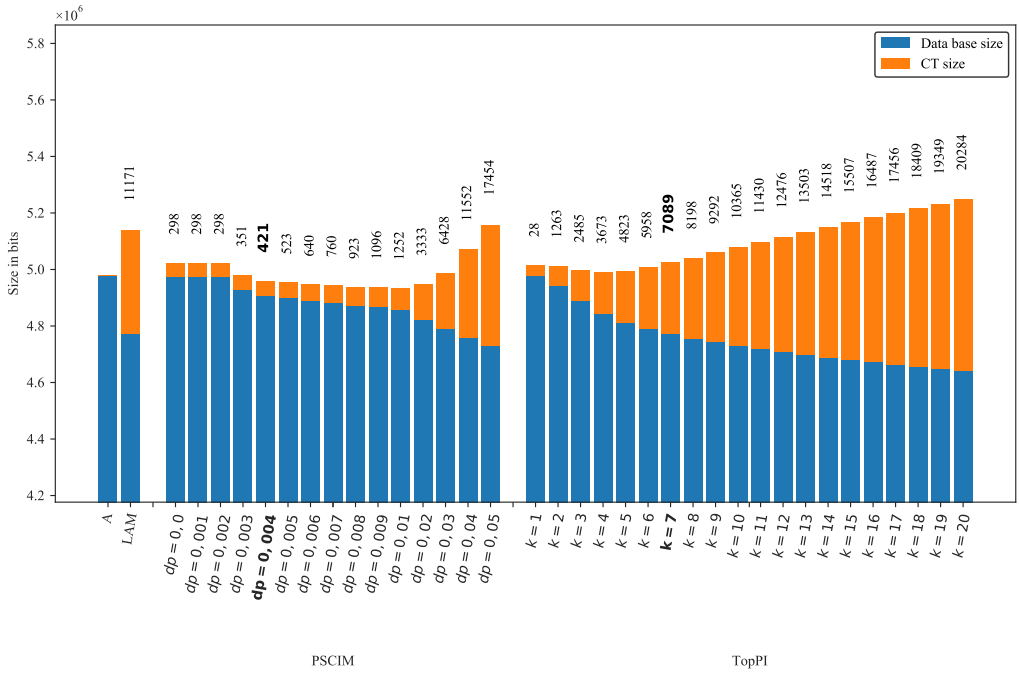


Figura C.41: *Fruithut*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2.6 OnlineRetail

Tabela C.33: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte † $\times 10^{-1}$														Total de itemsets	Tempo (s)
	[0,00 , 0,08] †		(0,08 , 0,15] †		(0,15 , 0,23] †		(0,23 , 0,31] †		(0,31 , 0,39] †		(0,39 , 0,46] †		(0,46 , 0,54] †			
♣ $\times 10^{-1}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	1.064	0,429	5	0,617	3	0,650	1	0,726	2	0,419	0	0,000	0	0,000	1.075	15,71
0,05 ♣	**1.065**	**0,429**	**6**	**0,551**	**3**	**0,650**	**1**	**0,726**	**2**	**0,419**	**0**	**0,000**	**1**	**0,536**	**1.078**	**15,69**
0,10 ♣	1.091	0,421	10	0,416	3	0,650	2	0,496	2	0,419	0	0,000	1	0,536	1.109	15,77
0,15 ♣	1.126	0,410	13	0,369	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.148	15,67
0,20 ♣	1.158	0,401	17	0,333	4	0,531	2	0,496	2	0,419	0	0,000	1	0,536	1.184	15,71
0,25 ♣	1.207	0,388	21	0,293	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.238	15,64
0,30 ♣	1.255	0,376	25	0,271	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.290	15,66
0,40 ♣	1.357	0,351	31	0,242	4	0,531	3	0,419	2	0,419	0	0,000	1	0,536	1.398	15,74
0,50 ♣	1.487	0,324	40	0,213	6	0,416	3	0,419	2	0,419	0	0,000	1	0,536	1.539	15,66
1,00 ♣	2.114	0,232	49	0,198	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	2.176	15,79
2,00 ♣	3.340	0,137	51	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	3.404	15,80

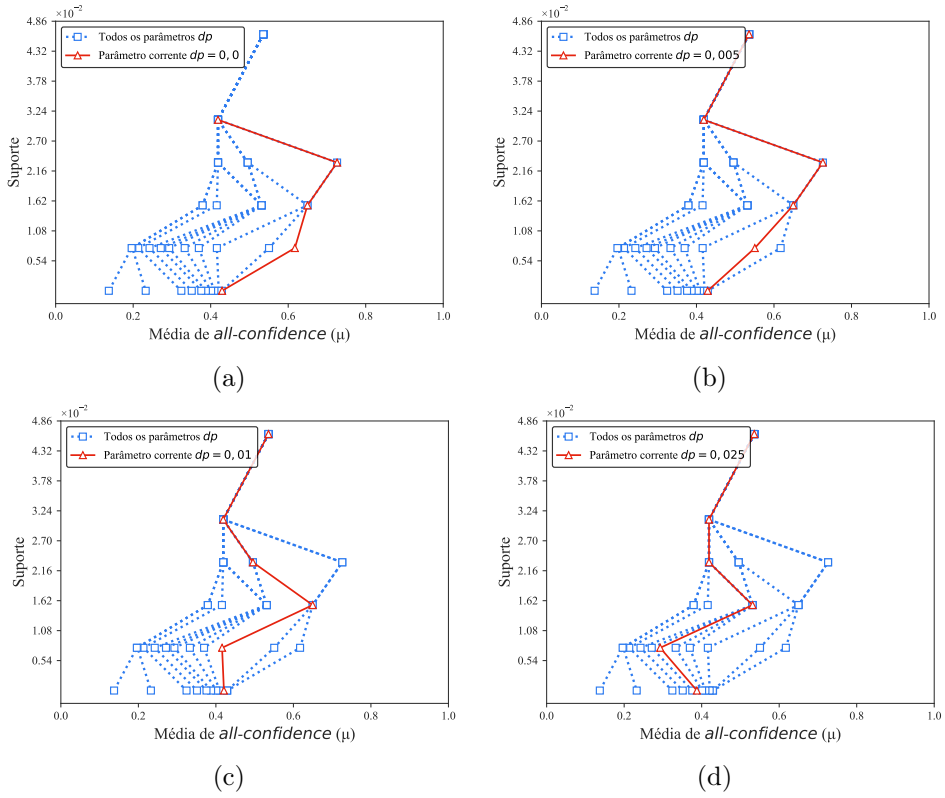
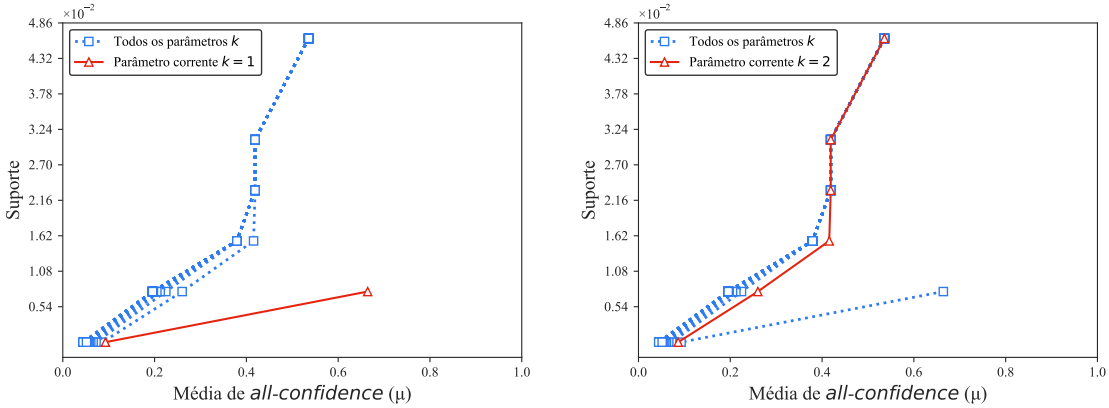


Figura C.42: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-1}$, (b) com $dr = 0{,}05 \times 10^{-1}$, (c) com $dr = 0{,}10 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}15\,, 0{,}20\}$ e (d) com $dr = 0{,}25 \times 10^{-1}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}30\,, 0{,}40\,, 0{,}50\,, 1{,}00\,, 2{,}00\}$. Veja Tabela C.33 para detalhes.

Tabela C.34: *OnlineRetail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte † × 10^{-1}														Itemset #	Tempo (s)
	[0,00 , 0,08] †		(0,08 , 0,15] †		(0,15 , 0,23] †		(0,23 , 0,31] †		(0,31 , 0,39] †		(0,39 , 0,46] †		(0,46 , 0,54] †			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	1.305	0,093	3	0,664	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1.308	0,97
2	**2.335**	**0,087**	**24**	**0,260**	**6**	**0,416**	**3**	**0,419**	**2**	**0,419**	**0**	**0,000**	**1**	**0,536**	**2.371**	**1,08**
3	3.242	0,072	36	0,224	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	3.291	1,12
4	3.805	0,066	46	0,212	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	3.864	1,17
5	4.326	0,061	49	0,205	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	4.388	1,19
6	4.732	0,058	53	0,197	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	4.798	1,22
7	5.054	0,055	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	5.121	1,27
8	5.343	0,053	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	5.410	1,25
9	5.604	0,052	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	5.671	1,22
10	5.814	0,050	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	5.881	1,24
11	6.007	0,049	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.074	1,24
12	6.164	0,048	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.231	1,29
13	6.317	0,047	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.384	1,27
14	6.460	0,047	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.527	1,28
15	6.581	0,046	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.648	1,25
16	6.678	0,046	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.745	1,25
17	6.779	0,045	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.846	1,28
18	6.857	0,045	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	6.924	1,35
19	6.940	0,044	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	7.007	1,34
20	6.999	0,044	54	0,196	7	0,379	3	0,419	2	0,419	0	0,000	1	0,536	7.066	1,36



(a) (b)

Figura C.43: *Fruithut*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.34 para detalhes.

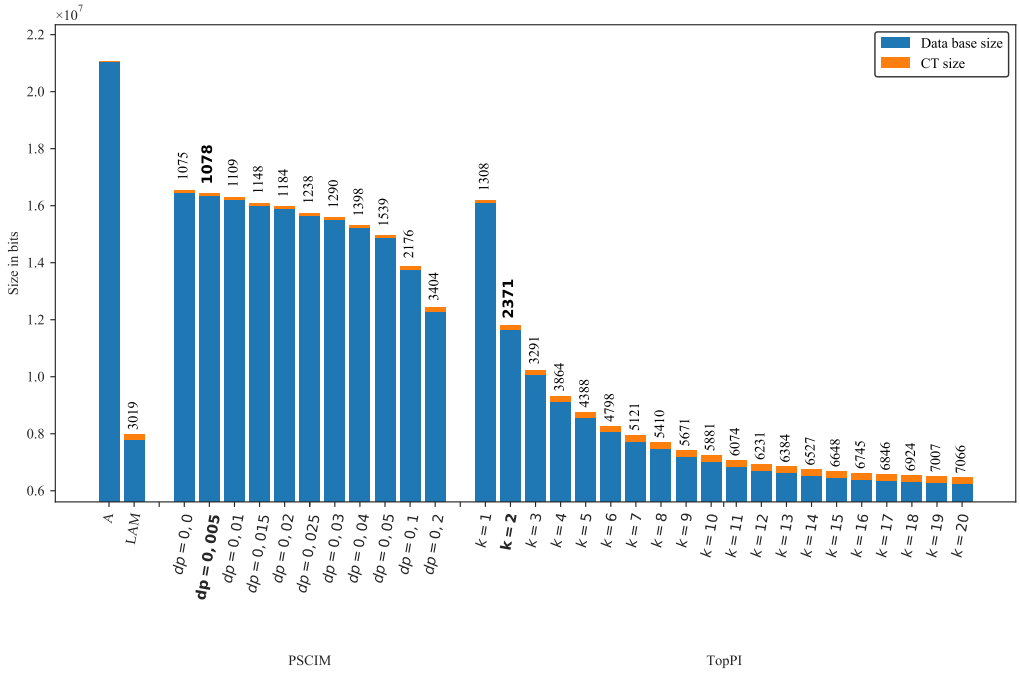
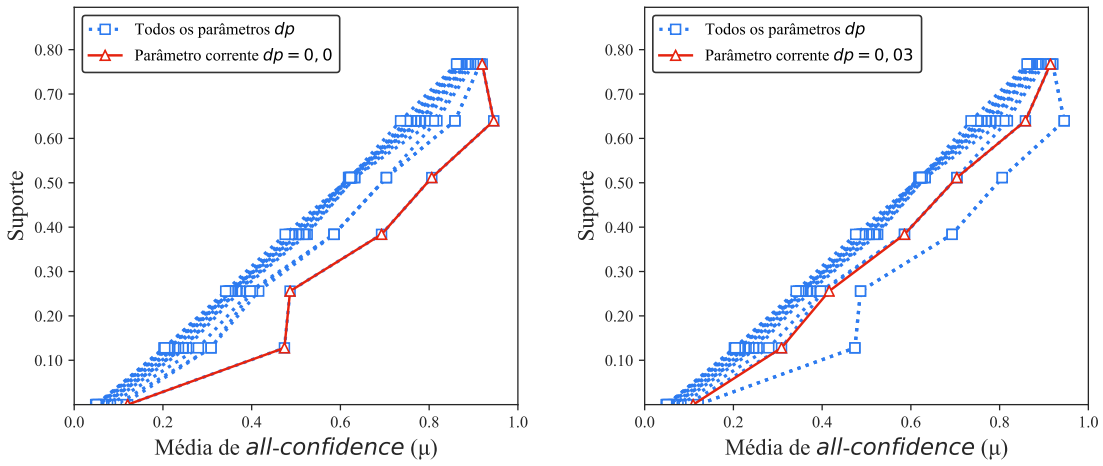


Figura C.44: *OnlineRetail*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2.7 PAMP

Tabela C.35: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,13]		(0,13 , 0,26]		(0,26 , 0,38]		(0,38 , 0,51]		(0,51 , 0,64]		(0,64 , 0,77]		(0,77 , 0,90]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00	**74**	**0,120**	**6**	**0,474**	**11**	**0,487**	**6**	**0,693**	**5**	**0,806**	**1**	**0,946**	**2**	**0,920**	**105**	**8,99**
0,02	74	0,120	6	0,474	11	0,487	6	0,693	5	0,806	1	0,946	2	0,920	105	8,98
0,03	92	0,109	17	0,308	21	0,416	15	0,586	10	0,704	3	0,858	10	0,915	168	9,11
0,04	123	0,095	17	0,308	25	0,402	15	0,586	10	0,704	3	0,858	10	0,915	203	9,16
0,05	162	0,082	25	0,280	28	0,397	15	0,586	10	0,704	7	0,818	22	0,900	269	9,20
0,06	198	0,075	44	0,253	43	0,373	34	0,525	23	0,632	9	0,810	26	0,893	377	9,32
0,07	258	0,067	48	0,246	53	0,367	36	0,521	23	0,632	18	0,791	35	0,889	471	9,40
0,08	270	0,066	60	0,245	65	0,365	39	0,517	26	0,625	24	0,781	42	0,886	526	9,43
0,09	345	0,062	101	0,235	82	0,357	72	0,506	37	0,620	33	0,779	47	0,882	717	9,53
0,10	475	0,055	132	0,229	125	0,352	73	0,505	43	0,623	58	0,771	74	0,874	980	9,61
0,11	527	0,054	156	0,225	147	0,353	104	0,501	57	0,618	64	0,768	75	0,873	1.130	9,74
0,12	648	0,052	236	0,215	181	0,355	168	0,492	72	0,619	118	0,760	83	0,870	1.506	9,87
0,13	811	0,050	305	0,214	216	0,353	169	0,492	89	0,620	159	0,759	96	0,867	1.845	9,97
0,14	1.104	0,048	411	0,211	329	0,351	283	0,489	139	0,622	238	0,752	104	0,865	2.608	10,21
0,15	1.319	0,051	601	0,202	395	0,347	350	0,485	166	0,625	319	0,749	109	0,864	3.259	10,28
0,20	5.038	0,050	1.652	0,204	1.722	0,341	1.279	0,476	905	0,622	698	0,736	118	0,862	11.412	11,95

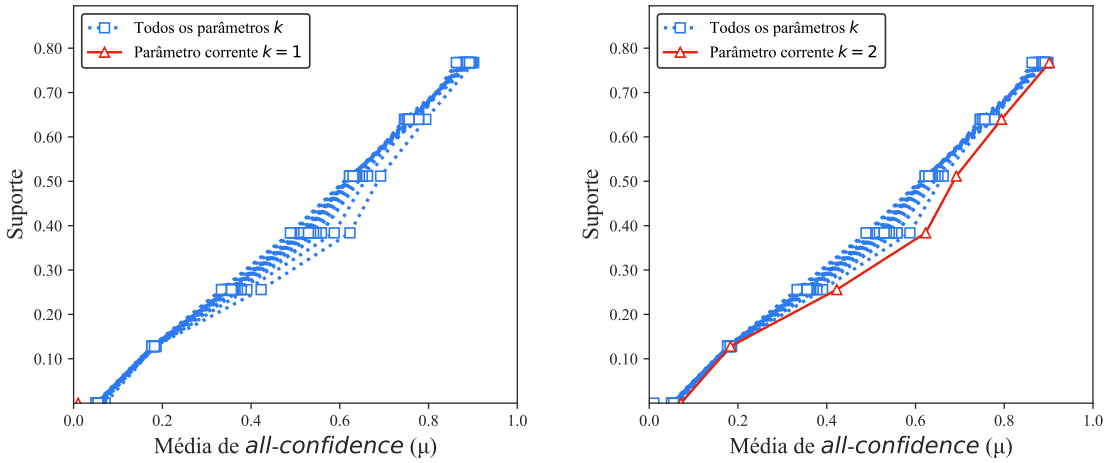


(a) (b)

Figura C.45: *PAMAP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}02\}$ e (b) com $dr = 0{,}03$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\,, 0{,}09\,, 0{,}10\,, 0{,}11\,, 0{,}12\,, 0{,}13\,, 0{,}14\,, 0{,}15\,, 0{,}20\}$. Veja Tabela C.35 para detalhes.

Tabela C.36: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,13]		(0,13 , 0,26]		(0,26 , 0,38]		(0,38 , 0,51]		(0,51 , 0,64]		(0,64 , 0,77]		(0,77 , 0,90]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	4	0,010	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4	2,87
2	**49**	**0,071**	**4**	**0,183**	**9**	**0,423**	**3**	**0,623**	**5**	**0,692**	**3**	**0,794**	**11**	**0,902**	**84**	**2,84**
3	94	0,063	8	0,185	18	0,391	8	0,587	9	0,663	5	0,764	22	0,898	164	4,29
4	138	0,060	12	0,186	27	0,378	13	0,558	13	0,651	6	0,756	33	0,898	242	4,62
5	182	0,058	16	0,184	36	0,369	18	0,559	17	0,644	8	0,746	41	0,896	318	5,02
6	226	0,057	20	0,183	45	0,364	23	0,547	21	0,638	11	0,747	48	0,894	394	5,36
7	270	0,056	24	0,184	54	0,361	28	0,541	25	0,635	15	0,751	54	0,895	470	5,68
8	314	0,055	28	0,183	63	0,358	33	0,534	29	0,632	19	0,756	58	0,894	544	6,06
9	358	0,055	32	0,182	72	0,356	38	0,529	33	0,630	23	0,756	66	0,892	622	6,49
10	402	0,056	36	0,181	81	0,354	43	0,526	37	0,628	27	0,755	72	0,891	698	6,99
11	446	0,055	40	0,181	90	0,352	48	0,523	41	0,627	32	0,755	79	0,889	776	7,53
12	490	0,055	44	0,180	99	0,350	53	0,520	46	0,626	36	0,756	86	0,887	854	8,07
13	534	0,054	48	0,180	108	0,349	58	0,520	51	0,626	40	0,758	91	0,886	930	8,62
14	578	0,054	52	0,179	117	0,348	63	0,518	56	0,628	44	0,759	96	0,884	1.006	9,47
15	622	0,054	56	0,179	126	0,352	68	0,515	61	0,628	48	0,759	100	0,883	1.081	9,76
16	666	0,053	60	0,178	135	0,351	73	0,513	67	0,628	51	0,761	103	0,882	1.155	12,08
17	710	0,053	64	0,178	144	0,350	78	0,513	72	0,627	54	0,763	108	0,881	1.230	11,39
18	754	0,053	68	0,178	153	0,348	83	0,511	77	0,625	57	0,765	113	0,879	1.305	12,09
19	798	0,053	72	0,178	162	0,347	88	0,509	83	0,627	60	0,766	116	0,879	1.379	12,64
20	842	0,053	76	0,178	171	0,346	93	0,507	88	0,626	63	0,767	118	0,878	1.451	13,54
30	1.282	0,051	116	0,180	260	0,340	142	0,497	145	0,624	100	0,775	151	0,871	2.196	20,69
40	1.722	0,050	156	0,177	350	0,336	200	0,491	194	0,622	147	0,778	174	0,866	2.943	25,99
50	2.152	0,050	200	0,176	435	0,333	263	0,489	240	0,622	197	0,778	186	0,863	3.673	31,89



(a) (b)

Figura C.46: *PAMP*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$ e (b) com $k = 2$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Veja Tabela C.36 para detalhes.

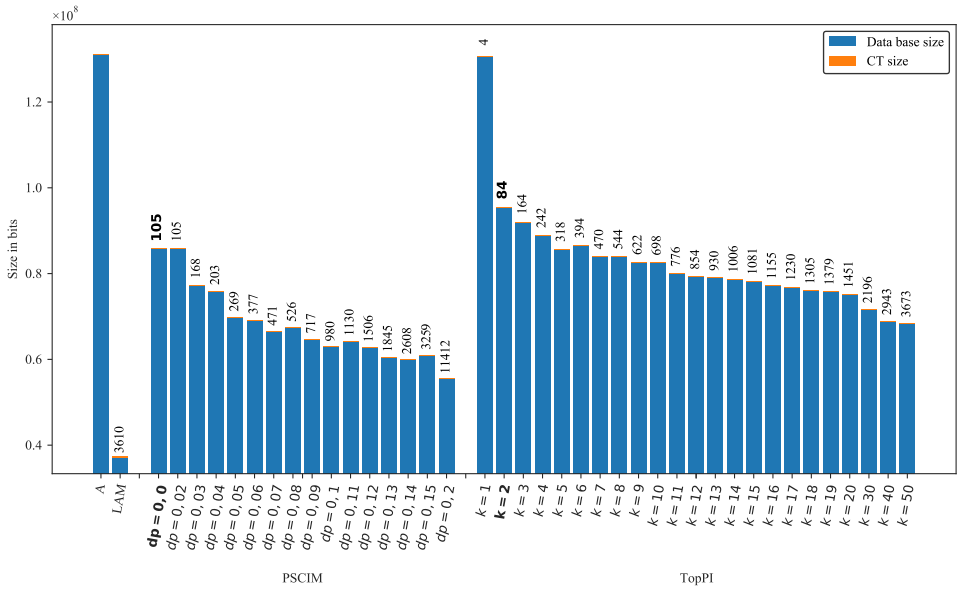


Figura C.47: *PAMAP*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).

C.1.2.8 Retail

Tabela C.37: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo PSCIM.

dr	Partição de suporte														Total de itemsets	Tempo (s)
	[0,00 , 0,05]		(0,05 , 0,09]		(0,09 , 0,14]		(0,14 , 0,19]		(0,19 , 0,24]		(0,24 , 0,28]		(0,28 , 0,33]			
♣ $\times 10^{-2}$	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
0,00 ♣	4.436	0,132	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	4.436	8.352,25
0,03 ♣	4.451	0,132	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.453	8.352,19
0,04 ♣	4.514	0,130	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.516	8.352,25
0,05 ♣	4.647	0,126	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.649	8.352,28
0,06 ♣	4.802	0,123	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	4.804	8.352,39
0,07 ♣	5.044	0,117	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.046	8.352,45
0,08 ♣	5.326	0,112	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.328	8.352,57
0,09 ♣	5.584	0,107	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.586	8.352,66
0,10 ♣	5.894	0,103	0	0,000	1	0,225	0	0,000	0	0,000	0	0,000	1	0,575	5.896	8.352,75
0,20 ♣	**9.729**	**0,072**	**1**	**0,191**	**1**	**0,225**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**1**	**0,575**	**9.732**	**8.353,96**
0,30 ♣	14.179	0,057	2	0,168	2	0,220	0	0,000	0	0,000	0	0,000	1	0,575	14.184	8.355,32

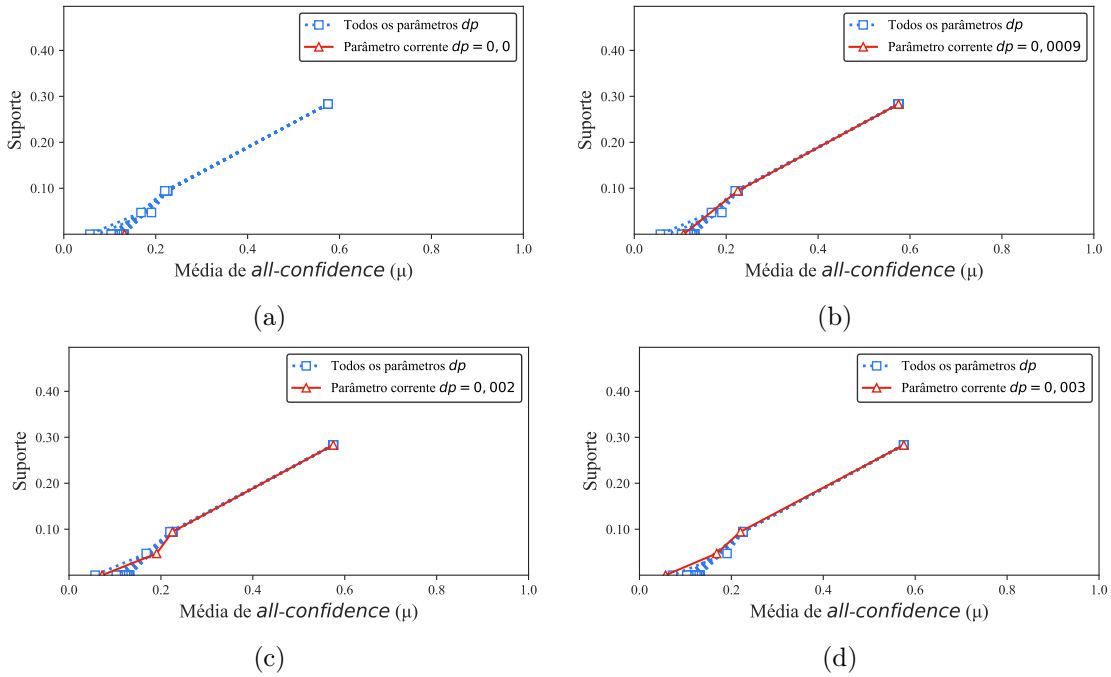


Figura C.48: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo PSCIM. (a) com $dr = 0{,}00 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}03\,, 0{,}04\,, 0{,}05\,, 0{,}06\,, 0{,}07\,, 0{,}08\}$, (b) com $dr = 0{,}09 \times 10^{-2}$, onde esta imagem representa, por similaridade, o comportamento de $dr \in \{0{,}10\}$, (c) com $dr = 0{,}20 \times 10^{-2}$ e (d) com $dr = 0{,}30 \times 10^{-2}$. Veja Tabela C.37 para detalhes.

Tabela C.38: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo algoritmo TopPI.

dr	Partição de suporte														Itemset #	Tempo (s)
	[0,00 , 0,05]		(0,05 , 0,09]		(0,09 , 0,14]		(0,14 , 0,19]		(0,19 , 0,24]		(0,24 , 0,28]		(0,28 , 0,33]			
	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ	#	μ		
1	3.241	0,006	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	3.241	1,73
2	17.040	0,004	0	0,000	3	0,199	0	0,000	0	0,000	0	0,000	1	0,575	17.044	2,03
3	**30.563**	**0,005**	**2**	**0,190**	**4**	**0,203**	**0**	**0,000**	**0**	**0,000**	**0**	**0,000**	**1**	**0,575**	**30.570**	**2,49**
4	42.711	0,005	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	42.721	2,62
5	54.365	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	54.375	2,65
6	65.339	0,007	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	65.349	2,68
7	75.818	0,007	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	75.828	2,75
8	85.931	0,007	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	85.941	2,74
9	95.705	0,007	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	95.715	2,86
10	105.091	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	105.101	2,95
11	114.223	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	114.233	2,92
12	123.063	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	123.073	2,91
13	131.638	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	131.648	3,02
14	140.020	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	140.030	3,04
15	148.211	0,006	5	0,150	4	0,203	0	0,000	0	0,000	0	0,000	1	0,575	148.221	3,12

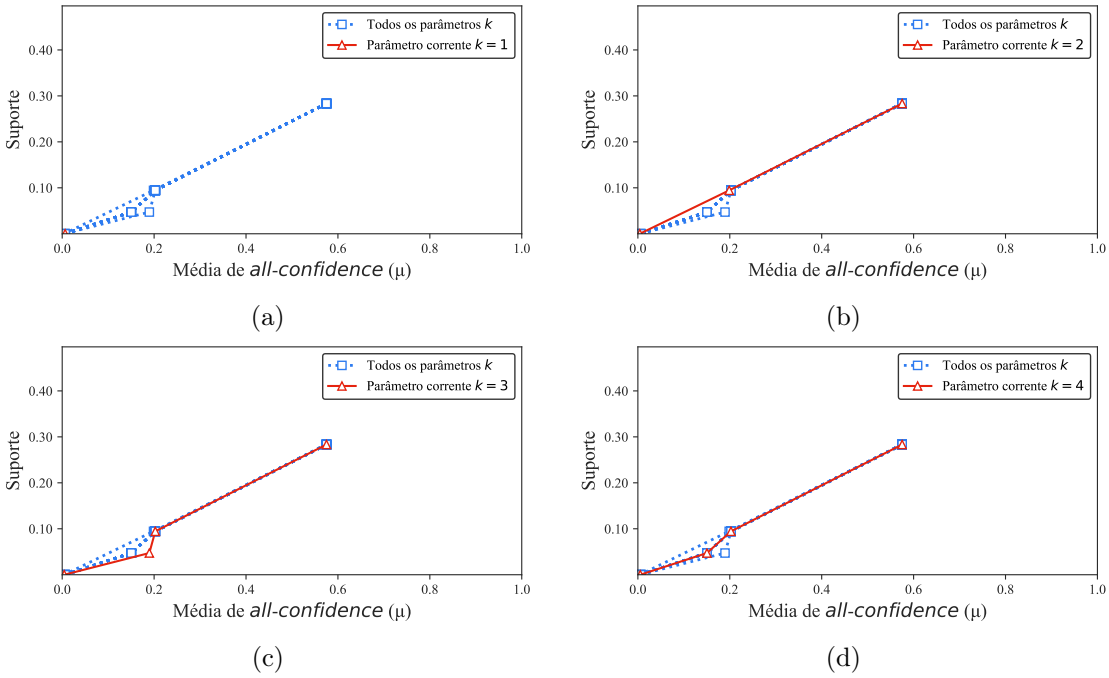


Figura C.49: *Retail*: distribuições de valores μ dos conjuntos de itens fechados recuperados pelo TopPI. (a) com $k = 1$, (b) com $k = 2$, (c) com $k = 3$ e (d) com $k = 4$, onde essa imagem representa, por similaridade, o comportamento de $k \in \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$. Veja Tabela C.38 para detalhes.

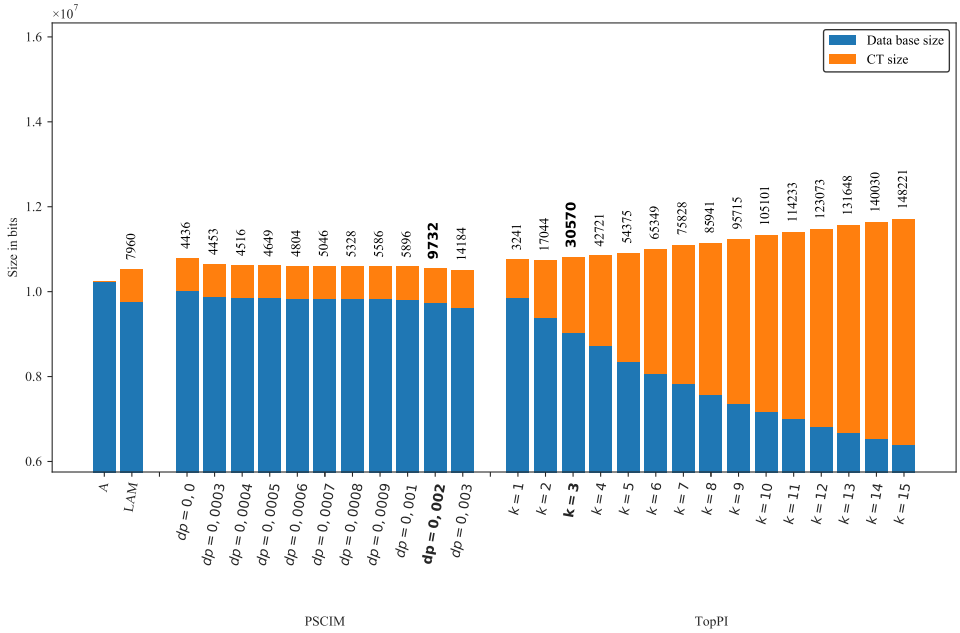


Figura C.50: *Retail*: valores métricos de MDL para todas as configurações de parâmetros em cada algoritmo. O número acima de cada barra é o número total de itemsets fechados recuperados. A barra A representa a compressão alcançada pelo CT padrão (*i.e,* o tamanho de referência usado para calcular o tamanho total comprimido).