UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Douglas Blanc Pereira

Avaliação de Medidas de Relevância para Seleção *Lazy* de Atributos

Niterói
2010

Douglas Blanc Pereira

Avaliação de Medidas de Relevância para Seleção *Lazy* de Atributos

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal Fluminense como parte dos requisitos para obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Alexandre Plastino
Co-orientadora: Bianca Zadrozny

Niterói
2010

Douglas Blanc Pereira

Avaliação de Medidas de Relevância para Seleção *Lazy* de Atributos

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal Fluminense como parte dos requisitos para obtenção do Grau de Bacharel em Ciência da Computação.

Aprovado em Julho de 2010

BANCA EXAMINADORA

Prof. Alexandre Plastino, D.Sc.
Orientador
UFF

Profa. Bianca Zadrozny, Ph.D.
Co-orientadora
UFF

Profa. Flavia Cristina Bernardini, D.Sc.
UFF

Prof. Leonardo Cruz da Costa, M.Sc.
UFF

Niterói

2010

RESUMO

Seleção de atributos é tradicionalmente uma etapa de pré-processamento que visa identificar atributos relevantes para a tarefa de classificação. O seu objetivo é remover atributos da base de dados que não contribuam para a classificação ou que possam prejudicar a capacidade preditiva do classificador.

Uma técnica *lazy* para seleção de atributos, proposta recentemente, adia a escolha dos atributos a serem utilizados ao momento em que a instância é submetida à classificação. Acredita-se que o conhecimento dos valores dos atributos da instância a ser classificada possa contribuir para a identificação dos melhores atributos para a classificação daquela instância em particular. Tal proposta utiliza uma medida baseada no conceito de entropia para avaliar a qualidade dos atributos.

Neste trabalho, são propostas medidas adicionais para seleção *lazy*, baseadas no teste chi-quadrado, no coeficiente de *Cramer*, índice *Gini* e *gain ratio*. Resultados experimentais utilizando o classificador k-NN mostram que a seleção *lazy* baseada nas medidas propostas permite obter um desempenho superior – melhores acurácias preditivas – para um conjunto significativo das bases de dados avaliadas, quando comparado com a seleção realizada como etapa de pré-processamento.

Palavras Chave:
medidas para seleção de atributos, seleção de atributos, classificação, mineração de dados.

ABSTRACT

Attribute selection is a data preprocessing step which aims at identifying relevant attributes for the classification task. It attempts to remove from the data set attributes which do not contribute to, or that can even reduce, the predictive ability of a classifier.

Recently, a lazy technique for attribute selection that postpones the choice of attributes to the moment an instance is submitted to classification was proposed. It is believed that knowledge about the attribute values of an instance to be classified may contribute in identifying the best attributes for the classification of that particular instance. This lazy approach uses a measure based on the entropy concept to evaluate the quality of the attributes.

In this work, additional measures are proposed for lazy selection, based on the chi-square test, Cramer's coefficient, Gini index and gain ratio. Experimental results using the k-NN classifier show that the lazy selection based on the proposed measures achieves superior performance – better predictive accuracy – for a significant number of the databases assessed, when compared with attribute selection performed as a preprocessing step.

Keywords:
attribute selection measures, attribute selection, classification, data mining.

LISTA DE ACRÔNIMOS

k-NN:	*k-Nearest Neighbors*
SVM:	*Suport Vector Machines*
CART:	*Classification and Regression Trees*
WEKA:	*Waikato Environment for Knowledge Analysis*
UCI:	*University of California, Irvine*
ENT:	Entropia
CHI:	Chi-quadrado
CRV:	Coeficiente de *Cramer*
GIN:	Índice *Gini*
GRT:	*Gain ratio*

SUMÁRIO

LISTA DE TABELAS

CAPÍTULO 1 - INTRODUÇÃO

Mineração de dados é o processo de descoberta de informações relevantes a partir de um grande conjunto de dados [10, 28]. A evolução dos computadores e dos meios de transmissão e de armazenamento de informações digitais possibilitou o acúmulo de grandes quantidades de dados que, por sua vez, motivou o desenvolvimento de técnicas automatizadas de mineração de dados.

Dentre as tarefas da área de mineração de dados, destaca-se a de classificação, que tem sido alvo de grande esforço de estudo e pesquisa devido à sua aplicabilidade em diversos domínios. Representam importantes aplicações da tarefa de classificação: detecção de fraudes, diagnóstico médico, estimativa de desempenho, entre outras com objetivos preditivos nas áreas de educação, finanças e biologia molecular [10, 28].

A partir de uma base de dados na qual cada instância é caracterizada por um conjunto de atributos e pela classe à qual pertence, o problema de classificação consiste em estimar a classe à qual pertence uma nova instância a partir dos valores de seus atributos.

A construção de algoritmos e modelos de classificação precisos e eficientes, para bases de dados de altas cardinalidades e muitos atributos, continua sendo um tema importante de investigação na área de mineração de dados. Dentre as técnicas de classificação mais importantes, destacam-se: indução de árvores de decisão [24, 25], o algoritmo k-NN (k-*Nearest Neighbors*) [4, 5], classificadores Bayesianos [6], redes neurais [26], a técnica SVM (*Suport Vector Machines*) [3], classificadores associativos [15], entre outras.

As técnicas de classificação são tradicionalmente categorizadas como *eager* ou *lazy*. As estratégias de classificação do tipo *eager* utilizam um conjunto de dados de treinamento para construir um modelo de classificação que define um mapeamento de instâncias para classes. No momento da classificação, esse modelo é utilizado para predizer a qual classe pertence uma nova instância.

Estratégias *lazy*, por outro lado, não constroem um modelo de classificação explícito e adiam a maior parte do processamento dos dados de treinamento até o momento em que se conheça a instância a ser classificada.

Sabe-se que o desempenho de técnicas de classificação está diretamente relacionado, entre

outros fatores, à qualidade dos dados da base de treinamento. Atributos redundantes e irrelevantes podem não somente prejudicar a acurácia de um classificador, mas também tornar o processo de construção do modelo ou a execução do algoritmo de classificação mais lento.

A fim de tentar evitar esses problemas, são utilizadas técnicas de seleção de atributos, que visam eliminar da base de treinamento atributos que não contribuem ou mesmo prejudicam o desempenho das estratégias de classificação [8, 16].

Tradicionalmente, técnicas de seleção são executadas na fase de pré-processamento – ou preparação – dos dados e suas decisões são definitivas para a fase de construção do modelo ou classificação propriamente dita. Porém, em [22], foi proposta uma técnica de seleção de atributos cuja característica principal é adiar a seleção dos atributos relevantes – de forma *lazy* – ao momento em que uma instância for submetida ao processo de classificação, em vez de se fazer a seleção previamente – de forma *eager*. Essa proposta tem como hipótese que o conhecimento dos valores dos atributos da instância a ser classificada pode contribuir para a identificação dos melhores atributos para aquela instância em particular. Dessa forma, para diferentes instâncias a serem classificadas, subconjuntos distintos de atributos, e customizados para cada instância, poderão ser selecionados. Na primeira versão da proposta *lazy*, foi utilizado, como medida de qualidade dos valores dos atributos, o conceito de entropia da distribuição de classes.

Em [17], o teste estatístico chi-quadrado, que pode ser adotado para se avaliar a correlação entre dois atributos, foi utilizado como medida de qualidade dos atributos em uma estratégia de seleção do tipo *eager*. Em [20], resultados preliminares foram obtidos com a adaptação da medida chi-quadrado para o contexto de seleção *lazy*. Neste trabalho, serão propostas e avaliadas adaptações de medidas baseadas no teste chi-quadrado, no coeficiente de *Cramer*, no índice *Gini* e no *gain ratio* para serem utilizadas como medidas de qualidade dos atributos de uma instância na estratégia de seleção *lazy* proposta em [22]. Um dos principais objetivos dessas propostas é verificar se, além do conceito de entropia, outras medidas também podem ser utilizadas na seleção *lazy* de atributos.

O restante deste trabalho está organizado conforme descrito a seguir. No Capítulo 2, realiza-se uma revisão sobre seleção de atributos, assim como uma revisão sobre medidas de relevância de atributos comumente utilizadas. A seleção *lazy* de atributos baseada na medida entropia é revisada no Capítulo 3. Novas medidas para seleção *lazy* propostas neste trabalho são apresentadas no Capítulo 4. No Capítulo 5, são analisados os resultados obtidos experimentalmente com as novas medidas e, finalmente, no Capítulo 6, são apresentadas as conclusões.

CAPÍTULO 2 - SELEÇÃO DE ATRIBUTOS

De maneira geral, nem todos os atributos de uma base de dados são necessários para discriminar a classe de maneira precisa e incluí-los no modelo de classificação pode até mesmo gerar resultados inferiores do que seriam obtidos se eles fossem removidos da base [11].

De acordo com [8], as técnicas de seleção de atributos são, a princípio, empregadas para identificar atributos relevantes e com informações essenciais. Em geral, além desse objetivo principal, existem outras importantes motivações: o aperfeiçoamento da acurácia preditiva do classificador, a redução e simplificação do conjunto de dados, a maior agilidade na tarefa de classificação e a simplificação do modelo de classificação gerado.

Estratégias de seleção de atributos são categorizadas como embutidas, *wrapper* ou do tipo filtro [16]. Técnicas embutidas aparecem incorporadas ao algoritmo de indução do modelo de classificação e realizam a seleção de atributos no seu processo de treinamento. Um exemplo típico são os algoritmos de indução de árvores de decisão, pois realizam a seleção dos atributos da base que serão colocados nos nós das árvores geradas. Técnicas *wrapper* e filtro procuram pelo subconjunto de atributos mais adequado que será utilizado pelo algoritmo de classificação ou pelo indutor do modelo. No caso da seleção *wrapper*, o próprio algoritmo de classificação adotado é utilizado para avaliar a qualidade dos subconjuntos de atributos.

Técnicas filtro são independentes do algoritmo de classificação adotado e medem a qualidade dos atributos por meio de medidas específicas, avaliando as distribuições de valores dos atributos e da classe, como exemplificado pelas técnicas *Information Gain Attribute Ranking* [29] e *Relief* [12, 13]. Ou podem ainda avaliar subconjuntos de atributos, buscando de forma heurística o melhor subconjunto. As técnicas mais conhecidas desse último grupo são: *Correlation-based Feature Selection* [9] e *Consistency-based Feature Selection* [18].

Em geral, técnicas *wrapper* resultam em uma acurácia preditiva melhor do que a obtida por técnicas do tipo filtro, uma vez que avaliam os subconjuntos de atributos utilizando o próprio algoritmo de classificação que será usado posteriormente. Porém, por terem que executar o algoritmo de classificação previamente e inúmeras vezes, seu custo computacional é superior ao das técnicas do tipo filtro. Existem ainda técnicas híbridas que tentam se beneficiar das características de ambas as abordagens [19].

Uma forma comum de se realizar a seleção de atributos de uma base de dados é utilizar uma métrica para avaliar a qualidade de cada um dos atributos individualmente. Dessa maneira é possível ordenar o conjunto de atributos e obter um ranqueamento que pode então ser utilizado para selecionar todos os atributos cujo valor da métrica esteja acima de um determinado limiar (*threshold*), ou mesmo um número fixo de melhores atributos, em termos absolutos ou percentuais.

Uma maneira de se medir a qualidade de um atributo para classificação é avaliar o seu grau de associação com a classe. Para determinar esse tipo de associação, foram propostas diversas métricas, que serão descritas a seguir. Essas métricas são empregadas em uma base de dados $D(A_1,A_2,...,A_n,C), n \geq 1$, com $n+1$ atributos, onde C é o atributo classe e o seu domínio é $\{c_1,c_2,...,c_m\}, m \geq 2$. Assume-se que os valores dos atributos e da classe são discretos.

2.1 ENTROPIA

O conceito de entropia tem origem na área de teoria da informação. Dado um atributo A, cujo domínio é $\{a_1,a_2,...,a_k\}$, $k \geq 1$, define-se a probabilidade $p_i, 1 \leq i \leq k$, de cada valor a_i do atributo como a razão entre o número de instâncias da base em que ocorre o valor a_i para o atributo A e o número total de instâncias. A entropia desse atributo, representada por $Ent(A)$ é dada por [10]:

$$Ent(A) = -\sum_{i=1}^{k}[p_i \times log_2(p_i)] \tag{2.1}$$

A entropia da classe, representada por $Ent(C)$, pode ser calculada da mesma forma, considerando p_j a razão entre o número de instâncias em que o valor c_j da classe, $1 \leq j \leq m$, ocorre na base e o número total de instâncias.

Seja a probabilidade $p_{j|i}$ a razão entre o número de instâncias da base que pertencem à classe c_j em que ocorre o valor a_i do atributo A, e o número total de instâncias da base. A entropia condicional da classe C, dado o atributo A, é calculada usando a fórmula:

$$Ent(C|A) = -\sum_{i=1}^{k}\sum_{j=1}^{m}\left[p_{j|i} \times log_2\left(\frac{p_{j|i}}{p_i}\right)\right] \tag{2.2}$$

Quanto mais informativo um atributo A for em relação à classe C, menor será a sua entropia condicional $Ent(C|A)$.

2.2 CHI-QUADRADO

A métrica Chi-quadrado (ou χ^2) avalia a qualidade de um atributo de acordo com a sua correlação com a classe, por meio de um teste estatístico χ^2. Para cada valor a_i do atributo A ($1 \leq i \leq k$) e para cada valor c_j da classe C ($1 \leq j \leq m$), existe uma frequência esperada quando $(A = a_i)$ e $(C = c_j)$, que pode ser calculada pela fórmula [10]:

$$e_{ij} = \frac{count(A = a_i) \times count(C = c_j)}{N}, \tag{2.3}$$

onde N é o número total de instâncias, $count(A = a_i)$ é o número de instâncias em que ocorre o valor a_i do atributo A, e $count(C = c_j)$ é o número de instâncias que pertencem à classe c_j. A partir da frequência esperada de todas as combinações de valores i e j, pode-se calcular a métrica χ^2 pela fórmula:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \left[\frac{(o_{ij} - e_{ij})}{e_{ij}} \right], \tag{2.4}$$

onde o_{ij} é a frequência observada da combinação $(A = a_i)$ & $(C = c_j)$, ou seja, a razão entre o número de instâncias em que ocorre o valor a_i do atributo A simultaneamente com o valor c_j da classe C, e o número total de instâncias da base. O teste estatístico pode ser aplicado para determinar se o atributo A e a classe C são independentes, usando o número de graus de liberdade dado por $(k-1) \times (m-1)$, que é proporcional ao número de ocorrências distintas de valores de A multiplicado pela cardinalidade da classe C. Se essa hipótese puder ser rejeitada, de acordo com um nível de significância estatística pré-determinado e uma distribuição χ^2, significa que o atributo é fortemente relacionado à classe [10].

2.3 COEFICIENTE DE CRAMER

O valor calculado pela métrica Chi-quadrado está diretamente relacionado ao número de instâncias de uma base e à cardinalidade dos atributos. Por outro lado, o coeficiente de *Cramer* não é afetado pelo tamanho da amostra, sendo muito útil quando se suspeita que um aumento significativo do Chi-quadrado é resultante do grande tamanho da amostra, em vez de uma relação entre os atributos. Este coeficiente é interpretado como uma medida da correlação entre dois atributos e varia de 0 a 1. Quanto mais próximo de 0, menor a correlação e, quanto mais próximo de 1 maior a correlação. O valor do coeficiente de *Cramer* (V) é calculado pela fórmula [27]:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}}, \tag{2.5}$$

onde χ^2 é o valor de Chi-quadrado do atributo em questão, n o é total de instâncias da base e q é o mínimo entre o número de valores distintos do atributo e a cardinalidade da classe.

2.4 ÍNDICE GINI

O índice *Gini* é uma medida estatística de desigualdade (ou impureza), comumente utilizada para calcular a desigualdade de distribuição de renda, embora seja aplicável a qualquer distribuição. Ele é utilizado para a tarefa de seleção de atributos no algoritmo CART [2] de árvores de decisão, com o objetivo de encontrar o atributo que contém a melhor partição de valores, considerando o índice *Gini*. Esse índice é máximo (pior) quando as instâncias relativas a uma determinada partição de valores estão igualmente distribuídas entre todas as classes, e o índice é mínimo (melhor) quando todas as instâncias pertencem a uma única classe. O índice *Gini* de cada valor do atributo A é dado por:

$$Gini(D,A,a_i) = 1 - \sum_{j=1}^{m} (p_{j|i})^2, \tag{2.6}$$

onde m é o número de valores distintos do atributo classe e $p_{j|i}$ é a probabilidade da instância pertencer à classe c_j $(1 \leq j \leq m)$ quando ocorre o valor a_i do atributo A $(1 \leq i \leq k)$.

O índice *Gini* do atributo A é dado por [14]:

$$Gini(D,A) = \sum_{i=1}^{k} p_i \times Gini(D,A,a_i), \tag{2.7}$$

onde k é o número de valores distintos do atributo A, p_i $(1 \leq i \leq k)$ a razão entre o número de instâncias da base em que ocorre o valor a_i para o atributo e o número total de instâncias e $Gini(D,A,a_i)$ é dado pela Equação 2.6.

2.5 GAIN RATIO

A medida *Gain ratio* [24] é baseada no conceito de entropia que permite calcular a taxa de ganho de informação de determinado atributo. Na tentativa de não priorizar atributos que têm muitos valores diferentes, a medida faz um tipo de normalização, que é a divisão do ganho de informação pelo $SplitInfo(D,A_j)$, definido por [10]:

$$SplitInfo(D,A_j) = -\sum_{j=1}^{v} w_j \times log_2(w_j), \tag{2.8}$$

onde w_j pode ser descrito como o peso da partição que contém os valores v_j do atributo A_j, dado pela razão entre número de ocorrências de v_j e o total de instâncias da base.

O *Gain ratio* pode ser obtido através da fórmula [10]:

$$GainRatio(D,A_j) = \frac{Ent(C) - Ent(C|A)}{SplitInfo(D,A_j)}, \tag{2.9}$$

onde $Ent(C)$ é a entropia da distribuição de classes e $Ent(C|A)$ é a entropia da classe C, dado o atributo A definida pela Equação 2.2.

CAPÍTULO 3 - SELEÇÃO *LAZY* DE ATRIBUTOS

Em estratégias convencionais de seleção de atributos – que, de agora em diante, serão referenciadas como estratégias de seleção *eager* – os atributos são selecionados na etapa de pré-processamento e aqueles não selecionados são descartados da base de dados, não participando do processo de classificação.

Proposta em [22], a estratégia de seleção de atributos *lazy* baseia-se na hipótese de que o conhecimento dos valores dos atributos da instância a ser classificada pode contribuir para o processo de escolha dos atributos da base mais adequados para a classificação dessa instância em particular. A seleção de atributos, para cada instância, é, portanto, realizada apenas no momento da sua classificação.

A seguir é apresentado um exemplo para ilustrar o fato de que a classificação de certas instâncias pode se beneficiar de atributos que provavelmente seriam descartados por estratégias convencionais de seleção de atributos. Percebe-se então a importância de se conhecer os valores dos atributos da instância a ser classificada antes de se selecionar os atributos.

Na Tabela 3.1, é apresentada uma base de dados composta por três atributos, X, Y e a classe C, representada de duas formas. A representação da esquerda está ordenada pelos valores de X e a da direita, ordenada pelos valores de Y. Observa-se que os valores do atributo X são fortemente correlacionados com os valores da classe, tornando-o um atributo importante para a classificação. Apenas o valor 4 não é um bom determinante dos valores do atributo classe. Nota-se que o atributo Y seria um forte candidato a ser eliminado por uma técnica de seleção de atributos uma vez que seus valores não discriminam bem as classes. Contudo, existe uma forte correlação entre o valor 4 do atributo Y e a classe B, que seria perdida se o atributo Y fosse eliminado. A classificação de um elemento com valor 4 no atributo Y claramente se beneficiaria com a presença desse atributo.

Em [22], trabalho em que se propôs a estratégia de seleção *lazy* de atributos, a capacidade do valor de um atributo discriminar bem o atributo classe é medida por meio de conceito de entropia da distribuição de classes [24]. Nesse trabalho, a seleção *lazy* foi avaliada quando utilizada em conjunto com o algoritmo de classificação *k-NN* [4, 5], pela sua simplicidade e eficiência.

Tabela 3.1: Base de dados de exemplo (Motivação)

Base ordenada por X			Base ordenada por Y		
– X –	– Y –	– C –	– X –	– Y –	– C –
1	2	B	2	1	A
1	3	B	3	1	B
1	4	B	4	1	A
2	1	A	1	2	B
2	2	A	2	2	A
2	3	A	3	2	B
3	1	B	1	3	B
3	2	B	2	3	A
3	4	B	4	3	B
4	1	A	1	4	B
4	3	B	3	4	B
4	4	B	4	4	B

O conceito de entropia pode ser usado para medir o quão bem os valores dos atributos de uma instância determinam a sua classe. Comumente, é utilizado para medir a relevância de um atributo em estratégias *eager* do tipo filtro que analisam atributos individualmente [29].

A seguir, define-se a estratégia de seleção *lazy* de atributos, proposta em [22], que utiliza o conceito de entropia na identificação de atributos relevantes.

Seja $D(A_1,A_2,...,A_n,C), n \geq 1$, uma base de dados com $n+1$ atributos, onde C é o atributo classe, $\{c_1,c_2,...,c_m\}, m \geq 2$, o domínio do atributo classe C, e $\{a_{j1},a_{j2},...,a_{jk}\}, k \geq 1$, o domínio do atributo A_j, a entropia da distribuição de classes em D, representada por $Ent(D)$, é definida por:

$$Ent(D) = -\sum_{i=1}^{m}[p_i \times log_2(p_i)], \tag{3.1}$$

onde p_i é a probabilidade de uma instância arbitrária em D pertencer à classe c_i.

Seja $\{a_{j1},a_{j2},...,a_{ji}\}, i \geq 1$, o domínio do atributo $A_j, 1 \leq j \leq n$, e seja $D_{ji}, 1 \leq j \leq n$, a partição de D composta por todas as instâncias cujo valor de A_j é igual a a_{ji}. A entropia da distribuição de classes em D, restrita aos valores do atributo $A_j, 1 \leq j \leq n$, representada por $Ent(D,A_j)$, é definida por:

$$Ent(D,A_j) = \sum_{i=1}^{k_j}\left[\left(\frac{|D_{ji}|}{|D|}\right) \times Ent(D_{ji})\right]. \tag{3.2}$$

Define-se, então, a entropia da distribuição de classes em D, restrita ao valor de a_{ji} do

atributo A_j, representada por $Ent(D,A_j,a_{ji})$, da seguinte forma:

$$Ent(D,A_j,a_{ij}) = Ent(D_{ji}). \tag{3.3}$$

O conceito definido na Equação 3.2 é usado pela estratégia *eager* de seleção de atributos conhecida como *Information Gain Attribute Ranking* [29] para medir a capacidade de um atributo discriminar os valores da classe. A Equação 3.3 é usada pela estratégia de seleção *lazy*, proposta em [22], para medir a capacidade de um valor especifico a_{ji}, de um atributo particular A_j, de discriminar as classes. Quanto mais próxima a entropia $Ent(D,A_j,a_{ji})$ é de zero, maior é a chance de o valor de a_{ji} do atributo A_j determinar bem alguma classe.

Os parâmetros de entrada da estratégia *lazy* proposta em [22] são: uma base de dados D com n atributos, uma instância $I = (v_1, v_2, ..., v_n)$ e um número $r, 1 \leq r \leq n$, que representa o número de atributos a serem selecionados. Para selecionar os r melhores atributos para classificar a instância I, os n atributos são avaliados baseados na medida *lazy* proposta em [22] ($Ent_{Lazy}(D,A_j,v_j)$), definida na Equação 3.4, que estabelece que, para cada atributo A_j, se a capacidade discriminatória do valor específico v_j de A_j $(Ent(D,A_j,v_j))$ é melhor que (menor que) a capacidade de discriminação geral do atributo A_j $(Ent(D,A_j))$, então a primeira será considerada como medida *lazy* do atributo A_j.

Dessa forma, a medida proposta para avaliar a qualidade de cada atributo A_j é definida por:

$$Ent_{Lazy}(D,A_j,v_j) = Min(Ent(D,A_j,v_j), Ent(D,A_j)), \tag{3.4}$$

onde Min() retorna o menor entre seus parâmetros.

Após calcular o valor $Ent_{Lazy}(D,A_j,v_j)$ para cada atributo A_j, a estratégia *lazy* selecionará os r atributos que apresentem os r menores valores dessa medida.

O algoritmo *k-NN*, a partir de uma base de dados D, uma instância I e um valor de k, sendo $k \geq 1$, basicamente, atribui a I a classe majoritária entre os seus k elementos mais próximos. A distância entre I e as instâncias de D é calculada a partir de uma função definida sobre os valores dos atributos das respectivas instâncias. Dessa forma, a utilização da seleção *lazy* de atributos, em conjunto com o algoritmo *k-NN*, por exemplo, implica que o cálculo da distância entre instâncias poderá ser realizado utilizando-se diferentes subconjuntos de atributos, para diferentes instâncias de entrada.

A seleção de atributos *lazy* baseada no conceito de entropia [22] apresentou resultados interessantes, demonstrando ser uma técnica bastante promissora. Para permitir uma avaliação

mais abrangente, tentando dar evidências de que a técnica *lazy* pode ter um bom desempenho independentemente do tipo de medida utilizada, neste trabalho, propõem-se novas medidas para avaliação *lazy* dos valores dos atributos, baseadas no teste chi-quadrado, no coeficiente de *Cramer*, no índice *Gini* e no *gain ratio*, que serão apresentadas no próximo capítulo.

CAPÍTULO 4 - NOVAS MEDIDAS PARA SELEÇÃO LAZY

Neste trabalho, propõem-se quatro novas medidas para seleção *lazy* de atributos, que tomam como base o teste estatístico Chi-quadrado [17], o coeficiente de *Cramer* [27], o índice *Gini* [10] e por fim a medida *gain ratio* [10].

A proposta *lazy* apresentada em [22] utiliza como parâmetros de entrada uma base de dados D com n atributos, uma instância $I = (v_1, v_2, ..., v_n)$ a ser classificada de acordo com os valores dos seus atributos e um número r, $1 \leq r \leq n$, que representa os atributos a serem selecionados. Os mesmos parâmetros são utilizados para avaliação das medidas propostas neste trabalho.

4.1 CHI-QUADRADO LAZY

O teste chi-quadrado permite medir o grau de correlação entre dois atributos que, neste caso, é utilizado para medir a correlação entre o atributo a ser selecionado e o atributo classe da base de dados.

Considerando A um atributo da base e C o atributo classe, o valor χ^2 calculado parece uma boa medida da capacidade de o atributo A determinar a classe, pois quanto maior o valor de χ^2, maior a chance de A e C serem correlacionados. No caso da seleção *lazy* de atributos, deve-se levar em consideração a capacidade de cada valor da instância a ser classificada de determinar o atributo classe. Portanto faz-se necessária uma adaptação das Equações 2.3 e 2.4 para que o valor χ^2 possa ser utilizado também na abordagem *lazy* de seleção de atributos.

Neste trabalho, propõe-se a Equação 4.1 que avalia o quanto o valor a_{ji} do atributo A_j está correlacionado com o atributo classe C:

$$\chi^2(D, A_j, v_j) = \sum_{k=1}^{m} \frac{(o_k - e_k)}{e_k}, \tag{4.1}$$

onde m é o número de valores distintos do atributo classe C, o_k é a frequência observada do par $\{A_j = v_j, C = c_k\}$ e e_k é a frequência esperada do mesmo par.

Quanto maior for o valor de χ^2, maior a chance de a_{ji} ser um bom determinante de uma classe.

A frequência esperada do par $\{A_j = v_j, C = c_k\}$ é definida na Equação

$$e_k = \frac{count(A_j = v_j) \times count(C = c_k)}{N}, \tag{4.2}$$

onde $count(A_j = v_j)$ é o número de instâncias em que ocorre o valor v_j do atributo A_j, $count(C = c_k)$ é o número de instâncias que pertencem à classe c_k e N é o número total de instâncias da base.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos pela medida χ^2_{Lazy} definida pela Equação 4.3, onde χ^2 é o valor de χ^2 calculado para o atributo A_j (Equação 2.4) e Max() é a função que retorna o maior entre seus parâmetros.

$$\chi^2_{Lazy}(D, A_j, v_j) = Max\left(\frac{\chi^2}{n_j}, \chi^2(D, A_j, vj)\right) \tag{4.3}$$

Quanto maior o valor de χ^2_{Lazy}, maior correlação entre o atributo A_j e o atributo classe C.

Como o valor de χ^2 é obtido a partir do acumulo dos n_j valores do domínio do atributo A_j e valor de $\chi^2(D, A_j, v_j)$ é obtido por uma única parcela desta soma (referente ao valor v_j), houve a necessidade de torná-los comparáveis. Para tal foi introduzido o denominador n_j que é o número de valores distintos do atributo A_j.

Após calcular o valor χ^2_{Lazy} para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r maiores valores de χ^2_{Lazy}.

4.2 COEFICIENTE DE CRAMER LAZY

Assim como o χ^2, o coeficiente de *Cramer* também pode ser utilizado para medir a correlação entre dois atributos de uma determinada base de dados.

Novamente, considerando A um atributo da base e C o atributo classe, o valor V calculado parece ser outra boa medida da capacidade de o atributo A determinar a classe, já que quanto maior o valor de V, maior a chance do atributo A ter uma correlação com a classe C. Para utilizar V na abordagem *lazy*, que leva em consideração a capacidade de cada valor da instância a ser classificada de determinar o atributo classe, há necessidade de uma adaptação da Equação 2.5.

A proposta de adaptação sugerida é representada pela Equação 4.4 que avalia o quanto o valor v_j do atributo A_j está correlacionado com o atributo classe C:

$$V(D, A_j, v_j) = \sqrt{\frac{\chi^2(D, A_j, v_j)}{N(q-1)}}, \tag{4.4}$$

onde $\chi^2(D,A_j,v_j)$ é o valor calculado de Chi-quadrado para o valor v_j do atributo A_j dado pela Equação 4.1, N o é total de instâncias e q é o mínimo entre o número de valores distintos do atributo A_j e a cardinalidade da classe C.

Quanto maior for o valor de $V(D,A_j,v_j)$, maior a chance de v_j ser um bom determinante de uma classe.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos de acordo com a medida V_{Lazy} definida pela Equação 4.5, onde χ^2 é o valor de χ^2 calculado para o atributo A_j (Equação 2.4) e Max() é a função que retorna o maior entre seus parâmetros.

$$V_{Lazy}(D,A_j,v_j) = Max\left(\sqrt{\frac{\chi^2/n_j}{N(q-1)}}, V(D,A_j,v_j)\right) \tag{4.5}$$

Quanto maior o valor de V_{Lazy}, maior correlação entre o atributo A_j e o atributo classe C.

Houve a necessidade de introduzir o denominador n_j na parcela referente ao valor de V obtido pela técnica *eager* pelo mesmo motivo apresentado no caso da medida Chi-quadrado *lazy*.

Após calcular o valor V_{Lazy} para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r maiores valores de V_{Lazy}.

4.3 ÍNDICE GINI LAZY

O índice *Gini* pode ser utilizado para medir a desigualdade da distribuição de valores de um atributo com relação às classes. O coeficiente é mínimo (melhor) quando todas as instâncias pertencem a uma única classe e máximo (pior) quando as instâncias relativas a uma determinada partição de valores estão igualmente distribuídas entre todas as classes.

A adaptação *lazy* proposta neste trabalho é descrita pela seguinte Equação:

$$Gini(D,A_j,v_j) = 1 - \sum_{k=1}^{m} (p_k)^2, \tag{4.6}$$

onde m é o número de valores distintos do atributo classe, p_j é a razão entre o número de instâncias da base que pertencem à classe c_k em que ocorre o valor v_j, e o número de ocorrências do valor v_j do atributo A_j.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da

avaliação dos n atributos de acordo com a medida $Gini_{Lazy}$ definida pela Equação 4.7, onde $Gini(D,A_j)$ é o índice *Gini* calculado para o atributo A_j (Equação 2.7) e Min() é a função que retorna o menor valor entre o índice *Gini* calculado a partir da estratégia *eager* e o considerando os valores da instância I.

$$Gini_{Lazy}(D,A_j,v_j) = Min(Gini(D,A_j), Gini(D,A_j,v_j)) \tag{4.7}$$

Quanto menor o valor de $Gini_{Lazy}$ menor a desigualdade na relação dos valores do atributo A_j com os da classe C.

Após calcular o valor $Gini_{Lazy}$ para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r menores valores de $Gini_{Lazy}$.

4.4 GAIN RATIO LAZY

Conforme visto na Seção 2.5, a medida entropia tende a superestimar a qualidade de atributos com muitos valores. Para evitar esse viés, a medida *gain ratio* faz um tipo de normalização, que é a divisão do ganho de informação pelo $SplitInfo(D,A_j)$ dado pela Equação 2.8.

No caso da estratégia *lazy*, a adaptação proposta para o cálculo do $SplitInfo(D,A,v_j)$ é definida por:

$$SplitInfo(D,A_j,v_j) = -log_2(w_j). \tag{4.8}$$

A entropia do valor do atributo adaptada para a forma *lazy* é dada por:

$$Ent(D,A_j,v_j) = -\sum_{j=1}^{m} p_j \times log_2(p_j), \tag{4.9}$$

onde m é o número de classes e p_j a probabilidade de uma instância ser da classe c_j.

A partir das definições, pode-se calcular o valor de $GainRaito(D,A_j,v_j)$ a partir da Equação:

$$GainRatio(D,A_j,v_j) = \frac{Ent(D) - Ent(D,A,v_j)}{Split(D,A_j,v_j)}, \tag{4.10}$$

onde $Ent(D)$ é a entropia calculada para a base D.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos pela medida $GainRatio_{Lazy}$ definida pela Equação 4.11, com Max() retornando o maior entre seus parâmetros, sendo $Gain(D,A_j)$ a taxa de ganho calculada a partir

da Equação 2.9 por $GainRatio(C,A_j)$ onde C é o atributo classe de D.

$$GainRatio_{Lazy}(D,A_j,v_j) = Max(GainRatio(D,A_j), GainRatio(D,A_j,v_j)) \tag{4.11}$$

Um maior valor de $GainRatio_{Lazy}$ indica um aumento no ganho de informação da base obtido pela escolha do atributo A_j considerando o valor v_j.

Após calcular o valor $Gain_{Lazy}$ para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r maiores valores de $Gain_{Lazy}$.

CAPÍTULO 5 - RESULTADOS EXPERIMENTAIS

Para realizar os experimentos computacionais, as estratégias de seleção *lazy* descritas no Capítulo 4 foram implementadas, integradas à ferramenta Weka (Waikato Environment for Knowledge Analysis) [21] e avaliadas com o classificador *k-NN*. A seleção ocorre quando o classificador recebe uma nova instância a ser classificada. Uma vez que, para cada nova instância, um subconjunto distinto de atributos pode ser considerado pelo classificador, os atributos não selecionados pela estratégia *lazy* para uma dada instância não são removidos do conjunto de dados, mas somente desconsiderados pelo classificador.

Conforme visto anteriormente, o objetivo principal deste trabalho é evidenciar que a estratégia *lazy* de seleção de atributos pode apresentar desempenho superior ao da estratégia *eager*, não somente com a medida baseada no conceito de entropia, mas também com medidas que avaliam a qualidade de um atributo a partir de outros conceitos, como os descritos no Capítulo 2, adotados neste trabalho. Além disso, neste trabalho, é realizada uma análise comparativa de desempenho entre as diferentes medidas quando utilizadas de forma *eager* e depois comparadas entre si quando adotadas de forma *lazy*.

Para realizar as comparações com as implementações das medidas *lazy* Chi-quadrado e *gain ratio*, foram utilizadas as técnicas de seleção de atributos *eager* correspondentes disponíveis na ferramenta WEKA que têm os nomes "*ChiSquaredAttributeEval*" e "*GainRatioAttributeEval*" respectivamente. Já para as medidas coeficiente de *Cramer* e índice *Gini*, além da implementação dos avaliadores *lazy*, houve necessidade da criação dos avaliadores de atributos para executar os experimentos para seleção *eager*, o "*CramerAttributeEval*" e o "*GiniIndexAttributeEval*", ambos baseados no pré-existente "*ChiSquaredAttributeEval*", pois a WEKA não dispunha das implementações *eager*.

As estratégias foram testadas em um grande número de bases de dados retiradas do repositório de base de dados da UCI [1]. Foram utilizadas 40 bases de dados, que possuem uma ampla variedade de tamanho, complexidade e áreas de aplicação. A Tabela 5.1 apresenta algumas informações sobre essas bases de dados: nome, número de instâncias, número de atributos e número de classes.

As medidas utilizadas para avaliar a qualidade dos atributos necessitam de valores de atribu-

tos discretos. Logo foi adotado um método supervisionado de discretização baseado em ganho de informação, proposto em [7], para discretizar os atributos contínuos dessas bases.

Tabela 5.1: Bases de dados do repositório da UCI

Base de dados	**Atributos**	**Classes**	**Instâncias**
anneal	38	5	898
audiology	69	24	226
autos	25	6	205
breast-cancer	9	2	286
breast-w	9	2	699
chess-Kr-vs-Kp	36	2	3196
credit-a	15	2	690
diabetes	8	2	768
flags	29	8	194
glass	9	6	214
heart-cleveland	13	2	303
heart-hungarian	13	2	294
hepatitis	19	2	155
horse-colic	27	2	368
hypo-thyroid	29	4	3772
ionosphere	34	2	351
labor	16	2	57
letter-recog	16	26	20000
lymph	18	4	148
mol-bio-promot	57	2	106
mol-bio-splice	60	3	3190
mushroom	22	2	8124
optdigits	64	10	5620
pendigits	16	10	10992
postoperative	8	3	90
primary-tumor	17	21	339
solar-flare1	12	6	323
solar-flare2	12	6	1066
sonar	60	2	208
soybean-large	35	19	683
spambase	57	2	4601
statlog-heart	13	2	270
statlog-segment	19	7	2310
statlog-vehicle	18	4	846
thyroid-sick	29	2	3772
vote	16	2	465
vowel	13	11	990
waveform-5000	40	3	5000
wine	13	3	178
zoo	17	7	101

O algoritmo de classificação utilizado, tanto para a seleção *eager* como para a *lazy*, foi o *k-NN*, especificamente, a implementação da ferramenta WEKA conhecida como *IBk*. Cabe ressaltar que o código do *k-NN* teve que ser adaptado para considerar, no cálculo da distância entre elementos, apenas os atributos selecionados de forma *lazy*.

As estratégias de seleção *lazy* e *eager* serão analisadas de duas maneiras diferentes para cada medida utilizada neste trabalho:

Na primeira abordagem – denominada "quantitativa" – serão comparados os valores médios de acurácia resultantes da utilização do classificador *k-NN*. Esses valores foram obtidos utilizando-se a técnica de validação cruzada com dez partições [10], na qual cada partição foi obtida de maneira aleatória. A acurácia correspondente a cada execução do classificador é uma média dos valores obtidos em cada partição. As mesmas partições foram utilizadas nos experimentos de cada técnica empregada.

Nas Tabelas 5.2, 5.4, 5.6, 5.8 e 5.10, estão sumarizados os resultados obtidos pelas estratégias *lazy* e *eager* usando os classificadores *1-NN*, *3-NN* e *5-NN*, para as 40 bases de dados e para cada uma das cinco medidas utilizadas. Para cada classificador e base de dados, foram comparadas as acurácias das estratégias ao se variar o percentual de atributos selecionados de 10% a 90% do total de atributos, com um incremento regular de 10%, considerando as nove execuções distintas das duas estratégias. As colunas *"Lazy"* e *"Eager"* indicam o número de vezes em que cada estratégia obteve a maior acurácia preditiva, considerando as nove comparações. O melhor comportamento é indicado por valores em negrito. A coluna "Empate" representa o número de vezes em que as estratégias obtiveram a mesma acurácia, ou seja, o mesmo número de acertos. A penúltima linha mostra a soma total de melhores resultados obtidos por cada estratégia. Os totais apresentados na última linha indicam o número de vezes em que cada estratégia foi melhor do que a outra com significância estatística (Total com sig. est.). Para realizar tal avaliação, foi empregado o teste estatístico de comparação de médias *t–Student*, bi-caudal e pareado, adotando o nível de significância $p = 0,05$, que indica que a probabilidade da diferença de desempenho das estratégias ter sido causada por fatores aleatórios é menor que 5%.

Na segunda análise – denominada "qualitativa" –, os resultados serão avaliados considerando que a melhor opção de número de atributos seria adotada por cada estratégia, dentro das faixas apresentadas anteriormente. Na prática, esse percentual de atributos poderia ser estimada de maneira independente para cada estratégia, utilizando-se, por exemplo, um procedimento de validação cruzada.

Nas Tabelas 5.3, 5.5, 5.7, 5.9 e 5.11, são relatadas, para cada base de dados, as melhores

acurácias obtidas com cada estratégia de seleção de atributos, dentre as nove diferentes quantidades percentuais de atributos. Para os classificadores *1-NN*, *3-NN* e *5-NN*, são apresentadas nas colunas *"Lazy"* e *"Eager"* as melhores acurácias obtidas por cada estratégia. A coluna "Sem sel." representa a acurácia obtida quando nenhuma seleção de atributos é realizada. Para cada base de dados e execução do *k-NN*, os valores em negrito indicam os melhores resultados obtidos para cada valor de *k* enquanto os valores sublinhados representam as melhores acurácias obtidas para a base independente do valor de *k*. Na última linha da tabela, estão sumarizados os totais de bases em que cada estratégia de seleção obteve o melhor resultado para as 40 bases de dados.

No caso da medida entropia, como pode ser observado nas duas últimas linhas da Tabela 5.2, para a maior parte das execuções, a estratégia *lazy* alcançou um número maior de resultados superiores do que a estratégia *eager*, independente do valor de *k*. Quando *k* é igual a 1, a estratégia *lazy* obteve o melhor resultado 220 vezes contra 87 vezes para a estratégia *eager*, com 53 empates. O comportamento foi semelhante para *k* igual a 3 e 5. Esses resultados já haviam sido evidenciados em [22].

Ao levar em consideração a significância estatística, a vantagem fica mais evidente, pois para *k* igual a 1, a estratégia *lazy* obteve o melhor resultado 87 vezes contra apenas 7 vezes para a estratégia *eager*.

Para as bases *Letter*, *Mol-Bio-Splice*, *Pendigits* e *Wine*, a estratégia *lazy* obteve os melhores resultados em todos os testes, independente do número de atributos escolhidos ou do parâmetro *k* do classificador. Para as bases *Chess-kr-vs-kp*, *Ionosphere*, *Labor*, *Primary-Tumor* e *Spambase*, para pelo menos um dos parâmetros de *k* a estratégia *lazy* alcançou o melhor resultado em todos os nove testes. O comportamento oposto não ocorreu, ou seja, uma prevalência da estratégia *eager* em todos os nove testes para alguma base de dados.

Utilizando como medida para seleção de atributos a entropia, é possível observar, através da Tabela 5.3, que, no cenário qualitativo, a estratégia *lazy* tende a alcançar acurácias superiores às da estratégia *eager*, independente do parâmetro *k*. Para o valor de *k* igual a 1, a estratégia *lazy* obteve a melhor acurácia 28 vezes, enquanto a *eager* obteve a melhor acurácia 21 vezes. Para *k* igual a 3, os resultados apontaram 29 melhores resultados da estratégia *lazy* contra 15 da *eager* e, para *k* igual a 5, o resultado foi 26 a favor da estratégia *lazy* contra 17 da estratégia *eager*.

Na Tabela 5.4, estão os resultados obtidos utilizando a medida Chi-quadrado. Novamente, como pode ser observado na penúltima linha, para a maior parte das execuções, a estratégia *lazy* alcançou um número maior de resultados superiores do que a estratégia *eager*. Quando

Tabela 5.2: Número de execuções em que cada estratégia obteve o melhor resultado utilizando a medida **Entropia**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**
anneal	**5**	1	3	**6**	0	3	**5**	1	3
audiology	**5**	0	4	**5**	3	1	4	4	1
autos	1	**6**	2	2	**6**	1	2	**6**	1
breast-cancer	**5**	4	0	4	**5**	0	**6**	3	0
breast-w	**7**	1	1	**5**	3	1	4	4	1
chess-Kr-vs-Kp	**9**	0	0	**8**	1	0	**7**	2	0
credit-a	**5**	4	0	**5**	4	0	**6**	3	0
diabetes	3	3	3	3	3	3	2	**4**	3
flags	2	**7**	0	2	**7**	0	2	**6**	1
glass	**7**	0	2	**7**	0	2	**7**	0	2
heart-cleveland	**4**	2	3	2	**4**	3	3	3	3
heart-hungarian	2	**4**	3	**5**	2	2	1	**5**	3
hepatitis	**8**	1	0	**7**	2	0	**7**	2	0
horse-colic	**5**	4	0	**5**	4	0	**5**	4	0
hypo-thyroid	**8**	1	0	**8**	1	0	**8**	1	0
ionosphere	**9**	0	0	**6**	3	0	**5**	4	0
labor	**9**	0	0	**7**	1	1	**6**	2	1
letter-recog	**9**	0	0	**9**	0	0	**9**	0	0
lymph	**6**	3	0	**8**	1	0	**7**	2	0
mol-bio-promot	**5**	4	0	**5**	4	0	**5**	4	0
mol-bio-splice	**9**	0	0	**9**	0	0	**9**	0	0
mushroom	3	0	6	3	0	6	3	0	6
optdigits	**6**	1	2	**6**	2	1	**5**	3	1
pendigits	**9**	0	0	**9**	0	0	**9**	0	0
postoperative	**4**	2	3	**6**	0	3	**5**	0	4
primary-tumor	**6**	3	0	**6**	3	0	**9**	0	0
solar-flare1	2	**7**	0	**5**	4	0	**5**	3	1
solar-flare2	3	**4**	2	**6**	2	1	**6**	1	2
sonar	2	**7**	0	**6**	3	0	**7**	2	0
soybean-large	**7**	1	1	**7**	1	1	**8**	0	1
spambase	**9**	0	0	**9**	0	0	**6**	3	0
statlog-heart	3	3	3	3	3	3	2	**4**	3
statlog-segment	**8**	1	0	**8**	1	0	**8**	1	0
statlog-vehicle	**7**	2	0	**8**	1	0	**7**	2	0
thyroid-sick	2	**5**	2	3	**5**	1	3	**5**	1
vote	**6**	2	1	**7**	1	1	**7**	1	1
vowel	**4**	2	3	**5**	2	2	**4**	3	2
waveform-5000	3	1	5	3	1	5	3	1	5
wine	**9**	0	0	**9**	0	0	**9**	0	0
zoo	4	1	4	4	4	1	3	**5**	1
Total	**220**	87	53	**231**	87	42	**219**	94	47
Total com sig. est.	**87**	7	–	**84**	10	–	**80**	25	–

k é igual a 1, a estratégia *lazy* obteve o melhor resultado 192 vezes contra 113 vezes para a estratégia *eager*, com 55 empates. Comportamento semelhante para *k* igual a 3 e 5, com uma ligeira redução na diferença entre os totais das duas estratégias.

A última linha da Tabela 5.4 mostra que, com significância estatística, a estratégia *lazy*

Tabela 5.3: As melhores acurácias preditivas alcançadas por cada estratégia utilizando a medida **Entropia**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**
anneal	**99,55**	99,33	99,22	**98,44**	98,00	98,00	**97,77**	97,22	97,22
audiology	**76,99**	75,66	76,11	**72,12**	68,58	65,93	**73,01**	69,47	60,62
autos	87,32	**88,29**	85,85	**81,46**	**81,46**	**81,46**	**77,07**	**77,07**	**77,07**
breast-cancer	**74,13**	72,73	69,93	**73,43**	72,73	70,28	**75,17**	74,13	74,13
breast-w	97,14	**97,28**	97,14	**97,00**	**97,00**	96,85	**97,14**	**97,14**	97,00
chess-Kr-vs-Kp	**96,81**	96,43	96,56	96,15	96,09	**96,59**	95,62	95,65	**96,06**
credit-a	**85,51**	**85,51**	82,32	**85,94**	85,51	84,20	85,80	**85,94**	84,64
diabetes	**77,99**	77,34	76,43	**78,26**	77,99	76,82	77,99	**78,13**	77,21
flags	60,82	**63,40**	59,79	61,34	**63,40**	55,15	**63,40**	62,37	57,73
glass	**77,10**	**77,10**	**77,10**	**75,70**	**75,70**	**75,70**	**73,83**	**73,83**	**73,83**
heart-cleveland	82,84	**84,49**	80,53	82,84	**84,16**	82,51	82,84	**83,50**	82,84
heart-hungarian	81,63	**82,65**	80,27	**83,67**	82,99	82,99	83,33	**84,01**	82,31
hepatitis	**86,45**	83,87	83,87	**86,45**	**86,45**	83,87	**85,16**	84,52	84,52
horse-colic	**83,70**	83,42	78,53	82,88	**83,15**	77,17	82,07	**83,15**	77,45
hypo-thyroid	**96,98**	94,59	91,52	**97,53**	94,80	93,21	**97,30**	94,62	93,27
ionosphere	**94,59**	93,45	92,59	**92,31**	91,45	90,60	**92,31**	90,88	89,74
labor	**100,00**	96,49	96,49	**98,25**	96,49	96,49	**96,49**	**96,49**	91,23
letter-recog	**91,93**	91,86	91,87	**90,59**	90,43	90,57	89,83	89,67	**89,84**
lymph	**85,14**	**85,14**	82,43	**83,78**	83,11	**83,78**	**85,14**	84,46	83,78
mol-bio-promot	89,62	**90,57**	80,19	88,68	**91,51**	80,19	**89,62**	87,74	79,25
mol-bio-splice	**90,72**	89,72	73,32	**91,22**	88,46	77,37	**91,16**	88,15	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	**94,84**	94,75	94,25	**95,50**	95,41	95,28	95,53	**95,55**	95,50
pendigits	96,79	96,31	**97,05**	96,54	95,95	**96,74**	96,10	95,54	**96,45**
postoperative	**71,11**	**71,11**	63,33	**71,11**	**71,11**	68,89	**71,11**	**71,11**	**71,11**
primary-tumor	42,48	**42,77**	38,35	**44,54**	44,25	43,66	**46,61**	45,72	46,31
solar-flare1	71,52	**72,76**	65,94	72,14	**73,07**	65,33	70,90	**71,21**	66,56
solar-flare2	**76,27**	75,89	73,45	**75,98**	75,70	74,02	**76,17**	75,70	73,83
sonar	86,54	**88,46**	86,54	**88,46**	87,98	86,06	**85,58**	83,65	84,62
soybean-large	**93,41**	**93,41**	92,24	**92,39**	**92,39**	91,51	**91,95**	**91,95**	90,78
spambase	**93,68**	93,13	92,98	**93,59**	93,35	93,31	**93,28**	93,24	93,20
statlog-heart	**85,19**	84,81	84,07	84,44	**84,81**	80,74	**85,93**	85,56	82,22
statlog-segment	94,68	**94,76**	94,68	93,90	93,94	**93,98**	92,90	**93,07**	**93,07**
statlog-vehicle	**71,39**	**71,39**	70,92	**71,51**	71,39	71,28	**71,04**	70,57	70,69
thyroid-sick	97,48	**97,72**	97,45	97,32	**97,48**	97,08	97,48	**97,61**	96,58
vote	**96,09**	95,17	92,18	**95,63**	95,17	92,41	**95,40**	95,17	92,87
vowel	**89,80**	**89,80**	**89,80**	84,14	84,14	**84,65**	78,08	78,28	**78,59**
waveform-5000	**75,52**	74,62	73,82	**79,06**	79,04	78,58	**80,84**	80,82	79,74
wine	**98,88**	**98,88**	98,31	**98,88**	97,19	96,63	**100,00**	96,63	96,07
zoo	**97,03**	**97,03**	96,04	**97,03**	**97,03**	93,07	92,08	**93,07**	**93,07**
Total	**28**	21	4	**29**	15	8	**26**	17	10

continua em vantagem, porém, esta é menor que a observada para o caso da medida entropia. Para k igual a 1, a estratégia *lazy* obteve o melhor resultado 55 vezes contra 22 vezes para a estratégia *eager*. No caso da entropia, os totais foram 87 e 7, para as estratégias *lazy* e *eager* respectivamente.

Para a base *Pendigits*, a estratégia *lazy* obteve os melhores resultados em todos os testes, independente do número de atributos escolhidos ou do parâmetro k do classificador. Para as bases *Breast-w* e *Spambase*, para pelo menos um dos valores de k a estratégia *lazy* alcançou o melhor resultado em todos os nove testes. Mais uma vez, o comportamento oposto não ocorreu, ou seja, uma prevalência da estratégia *eager* em todos os nove testes, para alguma base de dados.

Tabela 5.4: Número de execuções em que cada estratégia obteve o melhor resultado utilizando a medida **Chi-quadrado**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**
anneal	1	**5**	3	2	3	4	1	**5**	3
audiology	**6**	0	3	**7**	2	0	**5**	3	1
autos	3	**5**	1	1	**7**	1	3	**5**	1
breast-cancer	**5**	4	0	2	**7**	0	4	**5**	0
breast-w	**8**	1	0	**9**	0	0	**9**	0	0
chess-Kr-vs-Kp	**8**	1	0	**7**	2	0	**7**	2	0
credit-a	**4**	3	2	**7**	1	1	**6**	2	1
diabetes	**8**	1	0	**7**	2	0	**5**	4	0
flags	**8**	1	0	4	**5**	0	4	4	1
glass	**4**	3	2	3	**4**	2	**4**	3	2
heart-cleveland	2	3	4	2	**4**	3	1	**5**	3
heart-hungarian	3	3	3	**4**	3	2	3	**4**	2
hepatitis	3	**6**	0	2	**7**	0	4	**5**	0
horse-colic	**5**	4	0	4	**5**	0	4	**5**	0
hypo-thyroid	**5**	1	3	**4**	2	3	**5**	1	3
ionosphere	**8**	1	0	**8**	1	0	**8**	1	0
labor	**6**	1	2	**6**	1	2	**7**	1	1
letter-recog	**7**	2	0	**7**	2	0	**7**	2	0
lymph	3	**6**	0	**8**	1	0	**6**	3	0
mol-bio-promot	3	**6**	0	1	**8**	0	1	**8**	0
mol-bio-splice	2	**7**	0	3	**6**	0	4	**5**	0
mushroom	0	3	6	0	3	6	0	3	6
optdigits	**7**	1	1	**5**	3	1	**5**	3	1
pendigits	**9**	0	0	**9**	0	0	**9**	0	0
postoperative	2	**5**	2	3	2	4	2	2	5
primary-tumor	**8**	1	0	**8**	1	0	**6**	3	0
solar-flare1	4	4	1	3	**5**	1	2	**6**	1
solar-flare2	**6**	1	2	**6**	1	2	**5**	2	2
sonar	**7**	2	0	**6**	3	0	**5**	4	0
soybean-large	3	**6**	0	**5**	4	0	**5**	4	0
spambase	**9**	0	0	**9**	0	0	**6**	3	0
statlog-heart	1	4	4	0	4	5	1	4	4
statlog-segment	4	4	1	4	4	1	4	4	1
statlog-vehicle	4	**5**	0	4	**5**	0	4	**5**	0
thyroid-sick	**5**	1	3	**5**	1	3	**4**	2	3
vote	**5**	3	1	**5**	3	1	**4**	3	2
vowel	**4**	2	3	**4**	3	2	3	**4**	2
waveform-5000	4	0	5	2	2	5	2	2	5
wine	3	**5**	1	4	**5**	0	4	**5**	0
zoo	**5**	2	2	**6**	2	1	**8**	1	0
Total	**192**	113	55	**186**	124	50	**177**	133	50
Total com sig. est.	**55**	22	–	**48**	24	–	**44**	29	–

Dessa vez, utilizando a medida chi-quadrado para seleção de atributos, é possível observar através da Tabela 5.5 que, qualitativamente, a estratégia *lazy* tende a alcançar acurácias superiores às da estratégia *eager*, porém, agora dependendo do parâmetro k. Para o valor de k igual a 1, a estratégia *lazy* obteve a melhor acurácia 24 vezes, enquanto a *eager* obteve a melhor acurá-

cia 20 vezes. Para *k* igual a 3, os resultados apontaram 22 melhores resultados da estratégia *lazy* contra 18 da *eager* e, para *k* igual a 5, o resultado sofreu uma inversão com 22 a favor da estratégia *lazy* contra 23 da estratégia *eager*.

Apesar de ser possível observar uma provável tendência de vitórias da estratégia *eager* em relação a *lazy* para valores de *k* maiores, ela não é acompanhada de uma melhora nos valores de acurácia. Observe pois, na Tabela 5.5, que há uma maior ocorrência de valores sublinhados nas execuções do *1-NN* do que nas outras execuções com *k* igual a 3 ou 5.

Tabela 5.5: As melhores acurácias preditivas alcançadas por cada estratégia utilizando a medida **Chi-quadrado**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**
anneal	99,44	**99,55**	99,22	98,44	**98,66**	98,00	**98,00**	97,88	97,22
audiology	**78,32**	76,11	76,11	**72,12**	69,91	65,93	**69,91**	68,14	60,62
autos	**89,27**	88,29	85,85	81,46	**81,95**	81,46	**77,07**	**77,07**	**77,07**
breast-cancer	**75,52**	74,13	69,93	**75,17**	74,13	70,28	**75,52**	**75,52**	74,13
breast-w	97,14	**97,28**	97,14	**97,28**	97,00	96,85	**97,42**	97,14	97,00
chess-Kr-vs-Kp	96,50	96,43	**96,56**	95,96	96,09	**96,59**	95,56	95,65	**96,06**
credit-a	**86,38**	85,51	82,32	**86,67**	85,51	84,20	**86,67**	85,94	84,64
diabetes	74,61	74,35	**76,43**	74,74	74,09	**76,82**	74,35	74,09	**77,21**
flags	**60,82**	**60,82**	59,79	61,86	**62,37**	55,15	61,86	**62,37**	57,73
glass	**77,10**	**77,10**	**77,10**	**75,70**	**75,70**	**75,70**	**73,83**	**73,83**	**73,83**
heart-cleveland	82,84	**84,49**	80,53	83,17	**84,16**	82,51	83,17	**83,83**	82,84
heart-hungarian	**82,65**	**82,65**	80,27	**83,33**	82,99	82,99	82,99	**83,33**	82,31
hepatitis	**85,81**	84,52	83,87	83,87	**85,81**	83,87	83,87	**84,52**	**84,52**
horse-colic	**84,78**	83,42	78,53	**85,05**	83,15	77,17	**85,60**	83,70	77,45
hypo-thyroid	**93,61**	**93,61**	91,52	**93,58**	**93,58**	93,21	**93,56**	**93,56**	93,27
ionosphere	**93,45**	93,16	92,59	**92,31**	91,45	90,60	**92,31**	90,88	89,74
labor	**98,25**	**98,25**	96,49	96,49	**98,25**	96,49	**96,49**	**96,49**	91,23
letter-recog	**92,09**	91,86	91,87	**90,61**	90,43	90,57	89,79	89,67	**89,84**
lymph	83,78	**85,14**	82,43	**83,78**	83,11	**83,78**	**85,14**	84,46	83,78
mol-bio-promot	87,74	**90,57**	80,19	89,62	**91,51**	80,19	86,79	**87,74**	79,25
mol-bio-splice	88,21	**89,72**	73,32	86,46	**88,46**	77,37	86,24	**88,15**	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	94,73	**94,75**	94,25	**95,46**	95,37	95,28	95,53	**95,60**	95,50
pendigits	96,63	96,31	**97,05**	96,28	95,95	**96,74**	95,80	95,54	**96,45**
postoperative	70,00	**71,11**	63,33	**71,11**	**71,11**	68,89	**71,11**	**71,11**	**71,11**
primary-tumor	**43,07**	41,00	38,35	**44,54**	43,95	43,66	**46,90**	46,02	46,31
solar-flare1	70,90	**72,76**	65,94	71,21	**73,07**	65,33	68,73	**71,21**	66,56
solar-flare2	**75,89**	**75,89**	73,45	**75,70**	**75,70**	74,02	**75,70**	**75,70**	73,83
sonar	82,21	79,81	**86,54**	79,81	77,88	**86,06**	77,40	78,37	**84,62**
soybean-large	91,80	92,09	**92,24**	90,19	91,36	**91,51**	89,46	**91,22**	90,78
spambase	**93,61**	93,13	92,98	**93,41**	93,35	93,31	**93,26**	93,24	93,20
statlog-heart	84,07	**84,81**	84,07	83,70	**84,81**	80,74	**85,56**	**85,56**	82,22
statlog-segment	94,68	**94,76**	94,68	93,90	93,94	**93,98**	92,90	**93,07**	**93,07**
statlog-vehicle	**71,87**	70,45	70,92	**72,22**	71,39	71,28	**71,63**	70,45	70,69
thyroid-sick	**97,83**	97,77	97,45	**97,59**	97,51	97,08	97,59	**97,64**	96,58
vote	**95,63**	95,17	92,18	**95,40**	94,94	92,41	**95,17**	94,94	92,87
vowel	**89,80**	**89,80**	**89,80**	84,14	84,14	**84,65**	78,08	78,28	**78,59**
waveform-5000	**75,54**	74,62	73,82	78,96	**79,04**	78,58	80,32	**80,82**	79,74
wine	**98,88**	**98,88**	98,31	**97,19**	**97,19**	96,63	**96,63**	**96,63**	96,07
zoo	**98,02**	97,03	96,04	**97,03**	**97,03**	93,07	**93,07**	**93,07**	**93,07**
Total	**24**	20	8	**22**	18	10	22	**23**	13

Pode-se fazer a análise quantitativa em relação a medida coeficiente de *Cramer* através da Tabela 5.6. A penúltima linha da tabela nos mostra que, mais uma vez, a estratégia *lazy* alcançou um número maior de resultados superiores em relação à estratégia *eager*. Para *k* igual a 1, a

estratégia *lazy* obteve o melhor resultado 173 vezes contra 136 vezes para a estratégia *eager*, com 51 empates. Para k igual a 3 e 5, mais uma vez o comportamento foi semelhante. Como essa medida é baseada no Chi-quadrado, observamos novamente uma melhora da estratégia *eager* com o aumento do valor de k.

Tabela 5.6: Número de execuções em que cada estratégia obteve o melhor resultado utilizando a medida **coeficiente de Cramer**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**
anneal	2	**4**	3	**4**	3	2	2	**4**	3
audiology	**4**	3	2	4	**5**	0	2	**6**	1
autos	0	**8**	1	2	**6**	1	2	**7**	0
breast-cancer	**5**	4	0	2	**7**	0	4	**5**	0
breast-w	**8**	1	0	**9**	0	0	**9**	0	0
chess-Kr-vs-Kp	**8**	1	0	**7**	2	0	**7**	2	0
credit-a	**4**	3	2	**7**	1	1	**6**	2	1
diabetes	**8**	1	0	**7**	2	0	**5**	4	0
flags	**5**	4	0	**5**	4	0	**6**	3	0
glass	3	3	3	3	3	3	3	3	3
heart-cleveland	2	3	4	2	**4**	3	1	**5**	3
heart-hungarian	3	3	3	**4**	3	2	3	**4**	2
hepatitis	3	**6**	0	2	**7**	0	4	**5**	0
horse-colic	**5**	4	0	4	**5**	0	4	**5**	0
hypo-thyroid	2	**4**	3	3	2	4	2	**5**	2
ionosphere	**8**	1	0	**8**	1	0	**8**	1	0
labor	**6**	1	2	**6**	1	2	**7**	1	1
letter-recog	1	**8**	0	1	**8**	0	1	**8**	0
lymph	**5**	4	0	**7**	2	0	**8**	1	0
mol-bio-promot	3	**6**	0	1	**8**	0	1	**8**	0
mol-bio-splice	2	**7**	0	3	**6**	0	4	**5**	0
mushroom	0	3	6	0	3	6	0	3	6
optdigits	1	**7**	1	1	**7**	1	1	**7**	1
pendigits	**9**	0	0	**9**	0	0	**9**	0	0
postoperative	2	**6**	1	3	3	3	1	1	7
primary-tumor	**5**	3	1	**5**	3	1	**5**	3	1
solar-flare1	**6**	3	0	**5**	4	0	4	**5**	0
solar-flare2	3	**4**	2	**4**	3	2	3	3	3
sonar	**7**	2	0	**6**	3	0	**5**	4	0
soybean-large	**8**	0	1	**8**	0	1	**7**	1	1
spambase	**9**	0	0	**9**	0	0	**6**	3	0
statlog-heart	1	4	4	0	4	5	1	4	4
statlog-segment	**5**	4	0	**5**	4	0	4	**5**	0
statlog-vehicle	2	**7**	0	2	**7**	0	2	**7**	0
thyroid-sick	**5**	1	3	**5**	1	3	**4**	2	3
vote	**5**	3	1	**5**	3	1	**4**	3	2
vowel	**4**	3	2	3	**4**	2	3	**4**	2
waveform-5000	**4**	0	5	2	2	5	2	2	5
wine	4	**5**	0	2	**7**	0	1	**8**	0
zoo	**6**	2	1	**7**	2	0	**9**	0	0
Total	**173**	136	51	**172**	140	48	**160**	149	51
Total com sig. est.	**46**	34	–	**39**	31	–	**39**	34	–

A última linha da Tabela 5.6 mostra que, com significância estatística, a estratégia *lazy* ainda está em vantagem, porém, esta é menor que a observada para os casos das medidas entropia e chi-quadrado. Para k igual a 1, a estratégia *lazy* obteve o melhor resultado 46 vezes contra 34 vezes para a estratégia *eager*.

Como observado para a medida chi-quadrado, para a base *Pendigits*, a estratégia *lazy* obteve os melhores resultados em todos os testes, independente do número de atributos escolhidos ou do parâmetro k do classificador. Para as bases *Breast-w* e *Spambase*, para pelo menos um dos valores de k a estratégia *lazy* alcançou o melhor resultado em todos os nove testes. E, novamente não foi observada uma prevalência da estratégia *eager* em todos os nove testes, para alguma base de dados.

A Tabela 5.7 mostra qualitativamente o desempenho da medida coeficiente de *Cramer*, onde pode-se observar que a estratégia *lazy* passa a obter um número menor de acurácias superiores às da estratégia *eager*, apesar de ter conseguido um maior número de vitórias na análise quantitativa independente do valor de k. Para o valor de k igual a 1, a estratégia *lazy* obteve a melhor acurácia 19 vezes, enquanto a *eager* obteve a melhor acurácia 23 vezes. Para k igual a 3, os resultados apontaram 16 melhores resultados da estratégia *lazy* contra 21 da *eager* e, para k igual a 5, foram 17 a favor da estratégia *lazy* contra 24 da estratégia *eager*.

Novamente, como observado para as medidas entropia e chi-quadrado, os melhores valores de acurácia são obtidos quando o valor de k é igual a 1.

Tem-se na Tabela 5.8, a análise quantitativa dos experimentos realizados com a medida índice *Gini*. Pode-se observar resultados semelhantes aos obtidos pela medida entropia, onde para a maior parte das execuções, a estratégia *lazy* alcançou mais uma vez, um número maior de resultados superiores do que a estratégia *eager*. Com k é igual a 1, a estratégia *lazy* obteve o melhor resultado 231 vezes contra 98 vezes para a estratégia *eager*, com 31 empates. Mais uma vez, para k igual a 3 e 5, os resultados se assemelham. Porém, como o observado para as medidas chi-quadrado e coeficiente de *Cramer*, temos uma tendência de melhora da estratégia *eager* quando os valores de k aumentam.

Cabe destacar que ao levar em consideração a significância estatística, a vantagem em utilizar a estratégia *lazy* fica bastante evidente. Vemos que para k igual a 1, a estratégia *lazy* obteve o melhor resultado 76 vezes contra apenas 11 vezes para a estratégia *eager*.

Para as bases *Letter* e *Pendigits*, a estratégia *lazy* obteve os melhores resultados em todos os testes, independente do número de atributos escolhidos ou do parâmetro k do classificador. Para as bases *Hepatitis*, *Ionosphere*, *Labor*, *Mol-Bio-Splice*, *Spambase* e *Wine*, para pelo menos

Tabela 5.7: As melhores acurácias preditivas alcançadas por cada estratégia utilizando a medida **coeficiente de Cramer**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**
anneal	99,22	**99,55**	99,22	98,11	**98,33**	98,00	97,22	**97,44**	97,22
audiology	77,88	**78,32**	76,11	68,58	**70,35**	65,93	**67,26**	66,37	60,62
autos	86,83	**89,27**	85,85	**81,46**	**81,46**	**81,46**	**77,56**	77,07	77,07
breast-cancer	**75,52**	74,13	69,93	**75,17**	74,13	70,28	**75,52**	**75,52**	74,13
breast-w	97,14	**97,28**	97,14	**97,28**	97,00	96,85	**97,42**	97,14	97,00
chess-Kr-vs-Kp	96,50	96,43	**96,56**	95,96	96,09	**96,59**	95,56	95,65	**96,06**
credit-a	**86,38**	85,51	82,32	**86,67**	85,51	84,20	**86,67**	85,94	84,64
diabetes	74,61	74,35	**76,43**	74,74	74,09	**76,82**	74,35	74,09	**77,21**
flags	**61,34**	60,82	59,79	61,34	**63,40**	55,15	62,89	**63,40**	57,73
glass	**77,10**	**77,10**	**77,10**	**75,70**	**75,70**	**75,70**	**73,83**	**73,83**	**73,83**
heart-cleveland	82,84	**84,49**	80,53	83,17	**84,16**	82,51	83,17	**83,83**	82,84
heart-hungarian	**82,65**	**82,65**	80,27	**83,33**	82,99	82,99	82,99	**83,33**	82,31
hepatitis	**85,81**	84,52	83,87	83,87	**85,81**	83,87	83,87	**84,52**	**84,52**
horse-colic	**84,78**	83,42	78,53	**85,05**	83,15	77,17	**85,60**	83,70	77,45
hypo-thyroid	**93,61**	**93,61**	91,52	**93,58**	**93,58**	93,21	**93,56**	**93,56**	93,27
ionosphere	**93,45**	93,16	92,59	**92,31**	91,45	90,60	**92,31**	90,88	89,74
labor	**98,25**	**98,25**	96,49	96,49	**98,25**	96,49	**96,49**	**96,49**	91,23
letter-recog	90,06	91,80	**91,87**	88,39	90,43	**90,57**	87,54	89,67	**89,84**
lymph	**85,14**	**85,14**	82,43	**84,46**	83,11	83,78	81,08	**84,46**	83,78
mol-bio-promot	87,74	**90,57**	80,19	89,62	**91,51**	80,19	86,79	**87,74**	79,25
mol-bio-splice	88,21	**89,72**	73,32	86,46	**88,46**	77,37	86,24	**88,15**	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	94,73	**94,82**	94,25	95,32	**95,59**	95,28	95,50	**95,52**	95,50
pendigits	96,55	96,31	**97,05**	96,20	95,95	**96,74**	95,94	95,54	**96,45**
postoperative	**70,00**	**70,00**	63,33	**71,11**	**71,11**	68,89	**71,11**	**71,11**	**71,11**
primary-tumor	40,71	**41,89**	38,35	**44,25**	43,66	43,66	**46,61**	46,02	46,31
solar-flare1	**70,28**	**70,28**	65,94	69,35	**71,21**	65,33	68,11	**68,73**	66,56
solar-flare2	**74,95**	**74,95**	73,45	74,48	**74,86**	74,02	74,20	**74,77**	73,83
sonar	82,21	79,81	**86,54**	79,81	77,88	**86,06**	77,40	78,37	**84,62**
soybean-large	92,09	90,92	**92,24**	90,48	90,48	**91,51**	89,90	89,90	**90,78**
spambase	**93,61**	93,13	92,98	**93,41**	93,35	93,31	**93,26**	93,24	93,20
statlog-heart	84,07	**84,81**	84,07	83,70	**84,81**	80,74	**85,56**	**85,56**	82,22
statlog-segment	94,33	**94,76**	94,68	93,59	93,94	**93,98**	92,51	**93,07**	**93,07**
statlog-vehicle	70,80	**71,99**	70,92	71,28	**72,22**	71,28	70,21	**71,16**	70,69
thyroid-sick	**97,83**	97,77	97,45	**97,59**	97,51	97,08	97,59	**97,64**	96,58
vote	**95,63**	95,17	92,18	**95,40**	94,94	92,41	**95,17**	94,94	92,87
vowel	88,18	**89,80**	**89,80**	82,42	84,14	**84,65**	76,77	78,28	**78,59**
waveform-5000	**75,54**	74,62	73,82	78,96	**79,04**	78,58	80,32	**80,82**	79,74
wine	97,75	97,75	**98,31**	96,07	**97,19**	96,63	93,82	**96,07**	**96,07**
zoo	96,04	**97,03**	96,04	95,05	**97,03**	93,07	**93,07**	**93,07**	**93,07**
Total	19	**23**	10	16	**21**	11	17	**24**	14

um dos parâmetros de *k* a estratégia *lazy* alcançou o melhor resultado em todos os nove testes. Dessa vez, o comportamento oposto ocorreu, ou seja, uma prevalência da estratégia *eager* em todos os nove testes para a base *Hypo-thyroid* com o valor de *k* igual a 3 e 5.

Para a análise das melhores acurácias obtidas com a medida índice *Gini* tem-se a Tabela 5.9 onde é observado que a estratégia *lazy* passa a obter um maior número de acurácias inferiores às da estratégia *eager* para *k* igual a 3, apesar de ter conseguido um maior número de vitórias na análise quantitativa. Para o valor de *k* igual a 1, a estratégia *lazy* obteve a melhor acurácia 22 vezes, enquanto a *eager* obteve a melhor acurácia 18 vezes. Para *k* igual a 3, os resultados apontaram 19 melhores resultados da estratégia *lazy* contra 23 da *eager* e, para *k* igual a 5, foram 22 a favor da estratégia *lazy* contra 21 da estratégia *eager*.

Tabela 5.8: Número de execuções em que cada estratégia obteve o melhor resultado utilizando a medida índice **Gini**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**
anneal	**4**	3	2	**4**	3	2	1	**6**	2
audiology	**8**	1	0	**8**	1	0	**5**	4	0
autos	**7**	2	0	**5**	4	0	**6**	3	0
breast-cancer	**5**	4	0	**6**	3	0	**6**	3	0
breast-w	**7**	1	1	**6**	2	1	**5**	3	1
chess-Kr-vs-Kp	**8**	1	0	**8**	1	0	**6**	3	0
credit-a	**7**	2	0	**7**	2	0	**7**	2	0
diabetes	**5**	4	0	3	**6**	0	3	**6**	0
flags	2	**7**	0	4	**5**	0	4	**5**	0
glass	**5**	2	2	**4**	2	3	**6**	1	2
heart-cleveland	2	**4**	3	2	**4**	3	2	**4**	3
heart-hungarian	3	**6**	0	**5**	4	0	**5**	4	0
hepatitis	**9**	0	0	**9**	0	0	**7**	2	0
horse-colic	4	**5**	0	3	**6**	0	4	**5**	0
hypo-thyroid	2	**7**	0	0	**9**	0	0	**9**	0
ionosphere	**9**	0	0	**6**	3	0	**7**	2	0
labor	**9**	0	0	**7**	2	0	**7**	2	0
letter-recog	**9**	0	0	**9**	0	0	**9**	0	0
lymph	4	**5**	0	**6**	3	0	**6**	3	0
mol-bio-promot	**5**	4	0	**5**	4	0	**5**	4	0
mol-bio-splice	**8**	1	0	**9**	0	0	**9**	0	0
mushroom	2	0	7	2	0	7	2	0	7
optdigits	**7**	1	1	3	**5**	1	**5**	3	1
pendigits	**9**	0	0	**9**	0	0	**9**	0	0
postoperative	**7**	2	0	**5**	1	3	4	0	5
primary-tumor	**6**	3	0	**7**	2	0	**7**	2	0
solar-flare1	3	**6**	0	4	**5**	0	**5**	4	0
solar-flare2	**5**	4	0	**5**	4	0	**6**	3	0
sonar	**6**	3	0	**7**	2	0	**7**	2	0
soybean-large	**8**	1	0	**7**	2	0	**7**	2	0
spambase	**9**	0	0	**8**	1	0	**8**	1	0
statlog-heart	3	3	3	3	3	3	2	**4**	3
statlog-segment	**8**	1	0	**8**	1	0	**8**	1	0
statlog-vehicle	**8**	1	0	**6**	3	0	**6**	3	0
thyroid-sick	3	**5**	1	3	**5**	1	4	**5**	0
vote	**6**	2	1	**7**	1	1	**7**	1	1
vowel	**5**	2	2	**5**	2	2	3	**4**	2
waveform-5000	2	2	5	2	2	5	2	2	5
wine	**8**	1	0	**9**	0	0	**9**	0	0
zoo	**4**	2	3	**6**	3	0	**6**	3	0
Total	**231**	98	31	**222**	106	32	**217**	111	32
Total com sig. est.	**76**	11	–	**71**	13	–	**65**	18	–

Pode-se observar novamente o que havia ocorrido para as medidas já apresentadas: na maioria das vezes, as melhores acurácias são alcançadas com o valor de *k* igual a 1.

Finalmente, com a Tabela 5.10, pode-se fazer a análise quantitativa para a medida *gain ratio*. Dessa vez observa-se um comportamento diferente: para a maior parte das execuções

Tabela 5.9: As melhores acurácias preditivas alcançadas por cada estratégia utilizando a medida índice **Gini**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**
anneal	99,44	**99,55**	99,22	98,44	**98,66**	98,00	97,55	**97,88**	97,22
audiology	**77,88**	75,22	76,11	**70,35**	67,70	65,93	**71,68**	68,58	60,62
autos	87,80	**88,29**	85,85	81,46	**81,95**	81,46	76,59	**77,07**	**77,07**
breast-cancer	73,08	**73,78**	69,93	**73,78**	73,08	70,28	**74,48**	**74,48**	74,13
breast-w	97,00	**97,28**	97,14	**97,00**	**97,00**	96,85	**97,14**	**97,14**	97,00
chess-Kr-vs-Kp	**96,75**	96,40	96,56	96,15	96,06	**96,59**	95,56	95,62	**96,06**
credit-a	**85,22**	84,35	82,32	**85,80**	**85,80**	84,20	85,80	**85,94**	84,64
diabetes	68,49	72,66	**76,43**	69,40	72,53	**76,82**	69,79	72,79	**77,21**
flags	**61,34**	60,31	59,79	62,89	**63,40**	55,15	**63,40**	**63,40**	57,73
glass	**77,10**	**77,10**	**77,10**	**75,70**	**75,70**	**75,70**	**73,83**	**73,83**	**73,83**
heart-cleveland	82,51	**84,49**	80,53	82,51	**84,16**	82,51	82,84	**83,83**	82,84
heart-hungarian	81,63	**81,97**	80,27	**83,67**	82,99	82,99	**83,67**	82,99	82,31
hepatitis	**86,45**	83,23	83,87	**85,81**	**85,81**	83,87	**85,16**	**85,16**	84,52
horse-colic	**84,24**	82,88	78,53	**83,97**	82,07	77,17	**83,70**	82,34	77,45
hypo-thyroid	93,19	**93,58**	91,52	93,19	**93,53**	93,21	93,03	**93,48**	93,27
ionosphere	**94,30**	93,16	92,59	**92,02**	91,45	90,60	**92,02**	90,88	89,74
labor	**100,00**	98,25	96,49	**98,25**	**98,25**	96,49	**98,25**	96,49	91,23
letter-recog	**91,95**	91,86	91,87	**90,59**	90,43	90,57	89,83	89,67	**89,84**
lymph	84,46	**85,81**	82,43	83,78	**85,14**	83,78	**84,46**	**84,46**	83,78
mol-bio-promot	88,68	**90,57**	80,19	88,68	**91,51**	80,19	**89,62**	87,74	79,25
mol-bio-splice	87,49	**87,55**	73,32	**87,46**	86,08	77,37	**87,24**	85,86	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	**94,86**	94,75	94,25	**95,37**	**95,37**	95,28	95,57	**95,60**	95,50
pendigits	96,71	96,31	**97,05**	96,49	95,95	**96,74**	96,02	95,54	**96,45**
postoperative	**71,11**	70,00	63,33	**71,11**	**71,11**	68,89	**71,11**	**71,11**	**71,11**
primary-tumor	41,30	**42,18**	38,35	43,95	**44,84**	43,66	45,72	**46,61**	46,31
solar-flare1	72,14	**72,76**	65,94	72,45	**73,07**	65,33	70,59	**71,21**	66,56
solar-flare2	**76,27**	74,48	73,45	**76,17**	74,48	74,02	**76,17**	74,77	73,83
sonar	82,21	80,29	**86,54**	79,33	78,37	**86,06**	78,37	77,88	**84,62**
soybean-large	**93,27**	92,83	92,24	91,51	**92,39**	91,51	91,07	**91,95**	90,78
spambase	**93,72**	93,13	92,98	**93,57**	93,35	93,31	**93,26**	93,24	93,20
statlog-heart	**85,19**	84,81	84,07	84,44	**84,81**	80,74	**85,93**	85,56	82,22
statlog-segment	94,68	**94,76**	94,68	93,90	93,94	**93,98**	92,90	**93,07**	**93,07**
statlog-vehicle	**71,28**	70,45	70,92	71,28	**71,39**	71,28	**71,04**	70,45	70,69
thyroid-sick	97,59	**97,77**	97,45	97,40	**97,51**	97,08	97,38	**97,64**	96,58
vote	**96,09**	95,17	92,18	**95,40**	94,94	92,41	**95,40**	94,94	92,87
vowel	**89,90**	89,80	89,80	84,14	84,14	**84,65**	77,88	78,28	**78,59**
waveform-5000	**76,34**	74,62	73,82	78,98	**79,04**	78,58	80,70	**80,82**	79,74
wine	**98,88**	**98,88**	98,31	**98,88**	97,19	96,63	**100,00**	96,63	96,07
zoo	96,04	**97,03**	96,04	93,07	**97,03**	93,07	**93,07**	**93,07**	**93,07**
Total	**22**	18	5	19	**23**	8	**22**	21	12

a estratégia *eager* alcança um número maior de resultados superiores ao da estratégia *lazy*. A tendência da estratégia *eager* se beneficiar de valores de k maiores continua. Com k igual a 1, a estratégia *lazy* obteve o melhor resultado 160 vezes contra 156 vezes para a estratégia *eager*, com 44 empates. Os resultados para k igual a 3, mostram a estratégia *eager* com o melhor resultado 163 vezes contra 152 vezes da *lazy*. Para k igual a 5, podemos verificar que a estratégia *eager* fica com o melhor resultado 167 vezes contra 148 vezes da *lazy*.

Ao observar os resultados com significância estatística, fica claro que a estratégia *lazy* adotada apresenta pela primeira vez resultados piores que a *eager* independente do valor de k.

Apesar do resultado limitado, para as bases *Mol-Bio-Splice* e *Pendigits*, a estratégia *lazy*

obteve os melhores resultados em todos os testes, independente do número de atributos escolhidos e do valor de *k* do classificador. Por outro lado, nas bases *Breast-W* e *Ionosphere*, para pelo menos um dos parâmetros de *k* a estratégia *eager* alcançou o melhor resultado em todos os nove testes.

Tabela 5.10: Número de execuções em que cada estratégia obteve o melhor resultado utilizando a medida **gain ratio**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**	**Lazy**	**Eager**	**Empate**
anneal	1	**6**	2	0	**7**	2	0	**7**	2
audiology	**5**	4	0	**5**	4	0	**5**	4	0
autos	1	**8**	0	3	**6**	0	2	**7**	0
breast-cancer	4	**5**	0	**6**	3	0	**7**	2	0
breast-w	1	**8**	0	0	**9**	0	0	**9**	0
chess-Kr-vs-Kp	**6**	1	2	**5**	4	0	4	**5**	0
credit-a	2	**5**	2	3	3	3	**5**	1	3
diabetes	**6**	3	0	**5**	4	0	3	**6**	0
flags	4	**5**	0	3	**6**	0	4	**5**	0
glass	**6**	1	2	**6**	1	2	**6**	1	2
heart-cleveland	3	3	3	2	**4**	3	3	3	3
heart-hungarian	1	**6**	2	1	**6**	2	2	**5**	2
hepatitis	**7**	2	0	4	**5**	0	4	**5**	0
horse-colic	**6**	3	0	**5**	4	0	**5**	4	0
hypo-thyroid	1	**8**	0	3	**6**	0	2	**7**	0
ionosphere	0	**9**	0	1	**8**	0	1	**8**	0
labor	**7**	2	0	3	**4**	2	4	**5**	0
letter-recog	**8**	1	0	**8**	1	0	**8**	1	0
lymph	3	**6**	0	1	**8**	0	1	**8**	0
mol-bio-promot	3	**6**	0	4	**5**	0	3	**6**	0
mol-bio-splice	**9**	0	0	**9**	0	0	**9**	0	0
mushroom	0	1	8	0	1	8	0	1	8
optdigits	**8**	0	1	**7**	1	1	**7**	1	1
pendigits	**9**	0	0	**9**	0	0	**9**	0	0
postoperative	**5**	3	1	**6**	0	3	0	2	7
primary-tumor	**6**	3	0	**5**	4	0	**7**	2	0
solar-flare1	**6**	3	0	**6**	2	1	**6**	3	0
solar-flare2	2	**5**	2	2	**5**	2	3	**4**	2
sonar	**7**	2	0	**5**	4	0	**5**	4	0
soybean-large	1	**8**	0	1	**8**	0	1	**8**	0
spambase	3	**6**	0	2	**7**	0	2	**7**	0
statlog-heart	2	**4**	3	2	**4**	3	0	**6**	3
statlog-segment	3	**4**	2	3	**4**	2	2	**5**	2
statlog-vehicle	**5**	4	0	**5**	4	0	**5**	4	0
thyroid-sick	0	**8**	1	2	**7**	0	2	**7**	0
vote	3	**4**	2	3	**5**	1	4	4	1
vowel	**4**	2	3	**5**	2	2	**4**	3	2
waveform-5000	**4**	0	5	1	3	5	0	**4**	5
wine	**7**	0	2	**8**	0	1	**8**	0	1
zoo	1	**7**	1	3	**4**	2	**5**	3	1
Total	**160**	156	44	152	**163**	45	148	**167**	45
Total com sig. est.	51	**54**	–	50	**58**	–	42	**60**	–

Na Tabela 5.11, observa-se qualitativamente os resultados obtidos com a utilização da medida *gain ratio*. Percebe-se que as melhores acurácias novamente são obtidas quando o valor de *k* é 1. Para o valor de *k* igual a 1, a estratégia *lazy* obteve a melhor acurácia 19 vezes, enquanto a *eager* obteve a melhor acurácia 22 vezes. Para *k* igual a 3, os resultados apontaram 16 melhores resultados da estratégia *lazy* contra 25 da *eager* e, para *k* igual a 5, foram 15 a favor da estratégia *lazy* contra 25 da estratégia *eager*.

Os resultados obtidos com a utilização da medida *gain ratio* evidenciam a necessidade de se tentar uma melhor adaptação da forma *lazy*, já que a proposta neste trabalho obteve resultados inferiores quando comparada a outras medidas.

Tabela 5.11: As melhores acurácias preditivas alcançadas por cada estratégia utilizando a medida **gain ratio**

Base	**1-NN**			**3-NN**			**5-NN**		
	Lazy	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**	**Lazy**	**Eager**	**Sem sel.**
anneal	**99,55**	**99,55**	99,22	98,33	**98,89**	98,00	97,44	**98,55**	97,22
audiology	**78,76**	76,55	76,11	**70,35**	69,03	65,93	**68,14**	65,93	60,62
autos	88,78	**89,76**	85,85	80,00	80,00	**81,46**	76,10	75,61	**77,07**
breast-cancer	73,43	**74,48**	69,93	73,08	**73,78**	70,28	**74,48**	74,13	74,13
breast-w	95,99	**97,28**	97,14	96,71	**97,00**	96,85	96,42	**97,00**	**97,00**
chess-Kr-vs-Kp	96,21	96,53	**96,56**	95,71	95,53	**96,59**	95,18	95,24	**96,06**
credit-a	**86,23**	86,09	82,32	**86,67**	86,52	84,20	**87,10**	86,96	84,64
diabetes	73,05	73,83	**76,43**	72,92	73,57	**76,82**	72,66	73,70	**77,21**
flags	**65,98**	64,43	59,79	**65,98**	62,89	55,15	**64,43**	63,92	57,73
glass	**77,10**	**77,10**	**77,10**	**75,70**	**75,70**	**75,70**	**73,83**	**73,83**	**73,83**
heart-cleveland	**81,52**	81,19	80,53	**82,51**	**82,51**	**82,51**	82,84	**83,17**	82,84
heart-hungarian	83,67	**84,01**	80,27	**82,99**	**82,99**	**82,99**	**83,67**	**83,67**	82,31
hepatitis	**87,10**	84,52	83,87	84,52	**86,45**	83,87	84,52	**85,16**	84,52
horse-colic	84,78	**85,33**	78,53	84,51	**85,05**	77,17	84,78	**85,60**	77,45
hypo-thyroid	93,45	**93,61**	91,52	93,48	**93,58**	93,21	93,45	**93,56**	93,27
ionosphere	93,16	**93,45**	92,59	90,31	**91,17**	90,60	89,74	**90,03**	89,74
labor	**100,00**	98,25	96,49	**96,49**	**96,49**	**96,49**	**94,74**	92,98	91,23
letter-recog	91,86	**91,87**	**91,87**	90,38	90,40	**90,57**	89,56	89,56	**89,84**
lymph	84,46	**87,16**	82,43	81,76	**88,51**	83,78	81,08	**86,49**	83,78
mol-bio-promot	86,79	**90,57**	80,19	88,68	**91,51**	80,19	86,79	**87,74**	79,25
mol-bio-splice	**90,88**	89,72	73,32	**91,16**	88,46	77,37	**91,16**	88,15	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	**94,52**	94,32	94,25	**95,41**	95,28	95,28	**95,50**	**95,50**	**95,50**
pendigits	96,74	96,31	**97,05**	96,49	95,95	**96,74**	95,95	95,54	**96,45**
postoperative	**71,11**	70,00	63,33	**71,11**	**71,11**	68,89	**71,11**	**71,11**	**71,11**
primary-tumor	**41,89**	40,41	38,35	43,66	**44,54**	43,66	46,02	44,25	**46,31**
solar-flare1	**70,59**	70,28	65,94	**71,21**	**71,21**	65,33	69,35	**69,66**	66,56
solar-flare2	75,05	**75,98**	73,45	75,14	**75,52**	74,02	75,05	**75,52**	73,83
sonar	83,17	82,69	**86,54**	79,33	79,33	**86,06**	77,88	78,85	**84,62**
soybean-large	91,51	91,22	**92,24**	91,36	91,22	**91,51**	**90,78**	**90,78**	**90,78**
spambase	**93,72**	93,54	92,98	93,39	**93,72**	93,31	93,33	**93,44**	93,20
statlog-heart	**84,07**	**84,07**	**84,07**	**84,44**	**84,44**	80,74	84,44	**85,56**	82,22
statlog-segment	94,76	**95,19**	94,68	93,94	93,94	**93,98**	**93,07**	**93,07**	**93,07**
statlog-vehicle	70,33	**70,92**	**70,92**	71,04	**71,63**	71,28	70,69	**71,63**	70,69
thyroid-sick	97,56	**97,77**	97,45	97,51	**97,59**	97,08	97,48	**97,64**	96,58
vote	95,17	**95,63**	92,18	**95,40**	**95,40**	92,41	94,48	**95,17**	92,87
vowel	**89,80**	**89,80**	**89,80**	84,14	84,14	**84,65**	78,08	78,28	**78,59**
waveform-5000	**75,86**	75,36	73,82	79,10	**79,24**	78,58	80,50	**80,72**	79,74
wine	**98,88**	**98,88**	98,31	**98,31**	97,75	96,63	**97,75**	96,63	96,07
zoo	97,03	**98,02**	96,04	**97,03**	**97,03**	93,07	**93,07**	**93,07**	**93,07**
Total	19	**22**	11	16	**25**	14	15	**25**	16

Conforme foi possível observar, através dos resultados já apresentados, para a grande maioria das bases de dados, as melhores acurácias foram obtidas para o valor de *k* igual a 1. A fim de comparar as cinco medidas *lazy* para a seleção de atributos descritas neste trabalho, a Tabela 5.12 apresenta as acurácias obtidas por cada medida fixando valor de *k* em 1. As tabelas contendo os resultados obtidos com os outros valores de *k* assim como as tabelas comparativas para as medidas *eager* encontram-se nos apêndices A e B.

Nessas tabelas, apresenta-se a sumarização das análises qualitativas de cada medida, ou seja, a melhor acurácia que cada medida obteve para um determinado valor de *k*, independente do percentual de atributos selecionados. Em negrito, a melhor acurácia conseguida para a base de dados em questão. Na última linha da tabela, encontra-se o total de vezes que a medida obteve a melhor acurácia para uma das bases. A coluna "ENT" mostra as acurácias obtidas com o uso da entropia, a coluna "CHI" representa as acurácias da medida chi-quadrado, a coluna "CRV" representa os valores obtidos pela medida coeficiente de *Cramer*, a coluna "GIN" contém os valores da medida índice *Gini*, a coluna "GRT" com os valores da medida *gain ratio* e por fim a coluna "Sem sel." mostra as acurácias obtidas sem a seleção de atributos.

A Tabela 5.12 deixa claro que a seleção *lazy* de atributos é bastante útil, já que apenas para cinco bases de dados ela não melhorou a acurácia obtida sem seleção. Sendo que, em quatro dessas, pelo menos para uma medida a seleção *lazy* atingiu o mesmo resultado que a execução sem seleção. Fica claro também o desempenho superior da medida entropia que conseguiu a melhor acurácia em 19 das 40 bases utilizadas.

A Tabela 5.13 destaca, na segunda coluna, a melhor acurácia obtida em cada base de dados, indicando a medida (*eager* ou *lazy*) que a alcança, assim como o valor de *k*. Por exemplo, para a base *breast-cancer*, a maior acurácia, de 75,52%, foi obtida pela estratégia *lazy* com as medidas Chi-quadrado (com *k* igual a 1 e 5) e coeficiente de *Cramer* (com *k* igual a 1 e 5) e também pela estratégia *eager* com a medida Chi-quadrado (com *k* igual a 5) e coeficiente de *Cramer* (com *k* igual a 5).

Essa tabela mostra que a medida entropia se destaca das outras independente do valor de *k*, quando utilizada em conjunto com a estratégia *lazy*, já que é responsável por 17 das melhores acurácias obtidas. Por outro lado, ao se utilizar a estratégia *eager* a vantagem fica com a medida *gain ratio* que passa a ser responsável por 13 das melhores acurácias. Também fica fácil observar que o valor de *k* igual a 1 proporciona os melhores resultados.

Tabela 5.12: Comparativo entre medidas *lazy usando 1-NN*

Base	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**	**Sem sel.**
anneal	**99,55**	99,44	99,22	99,44	**99,55**	99,22
audiology	76,99	78,32	77,88	77,88	**78,76**	76,11
autos	87,32	**89,27**	86,83	87,80	88,78	85,85
breast-cancer	74,13	**75,52**	**75,52**	73,08	73,43	69,93
breast-w	**97,14**	**97,14**	**97,14**	97,00	95,99	**97,14**
chess-Kr-vs-Kp	**96,81**	96,50	96,50	96,75	96,21	96,56
credit-a	85,51	**86,38**	**86,38**	85,22	86,23	82,32
diabetes	**77,99**	74,61	74,61	68,49	73,05	76,43
flags	60,82	60,82	61,34	61,34	**65,98**	59,79
glass	**77,10**	**77,10**	**77,10**	**77,10**	**77,10**	**77,10**
heart-cleveland	**82,84**	**82,84**	**82,84**	82,51	81,52	80,53
heart-hungarian	81,63	82,65	82,65	81,63	**83,67**	80,27
hepatitis	86,45	85,81	85,81	86,45	**87,10**	83,87
horse-colic	83,70	**84,78**	**84,78**	84,24	**84,78**	78,53
hypo-thyroid	**96,98**	93,61	93,61	93,19	93,45	91,52
ionosphere	**94,59**	93,45	93,45	94,30	93,16	92,59
labor	**100,00**	98,25	98,25	**100,00**	**100,00**	96,49
letter-recog	91,93	**92,09**	90,06	91,95	91,86	91,87
lymph	**85,14**	83,78	**85,14**	84,46	84,46	82,43
mol-bio-promot	**89,62**	87,74	87,74	88,68	86,79	80,19
mol-bio-splice	90,72	88,21	88,21	87,49	**90,88**	73,32
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	94,84	94,73	94,73	**94,86**	94,52	94,25
pendigits	96,79	96,63	96,55	96,71	96,74	**97,05**
postoperative	**71,11**	70,00	70,00	**71,11**	**71,11**	63,33
primary-tumor	42,48	**43,07**	40,71	41,30	41,89	38,35
solar-flare1	71,52	70,90	70,28	**72,14**	70,59	65,94
solar-flare2	**76,27**	75,89	74,95	**76,27**	75,05	73,45
sonar	**86,54**	82,21	82,21	82,21	83,17	**86,54**
soybean-large	**93,41**	91,80	92,09	93,27	91,51	92,24
spambase	93,68	93,61	93,61	**93,72**	**93,72**	92,98
statlog-heart	**85,19**	84,07	84,07	**85,19**	84,07	84,07
statlog-segment	94,68	94,68	94,33	94,68	**94,76**	94,68
statlog-vehicle	71,39	**71,87**	70,80	71,28	70,33	70,92
thyroid-sick	97,48	**97,83**	**97,83**	97,59	97,56	97,45
vote	**96,09**	95,63	95,63	**96,09**	95,17	92,18
vowel	89,80	89,80	88,18	**89,90**	89,80	89,80
waveform-5000	75,52	75,54	75,54	**76,34**	75,86	73,82
wine	**98,88**	**98,88**	97,75	**98,88**	**98,88**	98,31
zoo	97,03	**98,02**	96,04	96,04	97,03	96,04
Total	**19**	14	9	13	14	5

Tabela 5.13: Análise geral das melhores acurácias obtidas

Base	**Acurácia**	*Lazy*					*Eager*				
		ENT	**CHI**	**CRV**	**GIN**	**GRT**	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**
anneal	99,55	1				1		1	1	1	1
audiology	78,76					1					
autos	89,76										1
breast-cancer	75,52		1,5	1,5				5	5		
breast-w	97,42		5	5							
chess-Kr-vs-Kp	96,81	1									
credit-a	87,10					5					
diabetes	78,26	3									
flags	65,98					1,3					
glass	77,10	1	1	1	1	1	1	1	1	1	1
heart-cleveland	84,49						1	1	1	1	
heart-hungarian	84,01						5				1
hepatitis	87,10					1					
horse-colic	85,60		5	5							5
hypo-thyroid	97,53	3									
ionosphere	94,59	1									
labor	100,00	1			1	1					
letter-recog	92,09		1								
lymph	88,51										3
mol-bio-promot	91,51						3	3	3	3	3
mol-bio-splice	91,22	3									
mushroom	100,00	1,3,5	1,3,5	1,3,5	1,3,5	1,3,5	1,3,5	1,3,5	1,3,5	1,3,5	1,3,5
optdigits	95,60							5		5	
pendigits	97,05										
postoperative	71,11	1,3,5	3,5	3,5	1,3,5	1,3,5	1,3,5	1,3,5	3,5	3,5	3,5
primary-tumor	46,90		5								
solar-flare1	73,07						3	3		3	
solar-flare2	76,27	1			1						
sonar	88,46	3					1				
soybean-large	93,41	1					1				
spambase	93,72				1	1					3
statlog-heart	85,93	5			5						
statlog-segment	95,19										1
statlog-vehicle	72,22		3						3		
thyroid-sick	97,77							1	1	1	1
vote	96,09	1			1						
vowel	89,90				1						
waveform-5000	80,84	5									
wine	100,00	5			5						
zoo	98,02		1								1
Total		17	10	6	10	10	9	10	9	9	13

CAPÍTULO 6 - CONCLUSÕES

Neste trabalho, foram propostas novas medidas para seleção *lazy* de atributos, baseadas no teste Chi-quadrado, no coeficiente de *Cramer*, no índice *Gini* e no *gain ratio* que foram comparadas com suas respectivas implementações *eager*. Foi feita também, uma análise comparativa de desempenho entre as diferentes medidas quando utilizadas de forma *lazy* e depois comparadas entre si quando adotadas de forma *eager*.

As avaliações das novas medidas propostas foram realizadas a partir de 40 bases de domínio público adotadas frequentemente em experimentos de mineração de dados, obtidas do repositório de dados *UCI Machine Learning Repository*.

Os resultados experimentais evidenciaram que outras medidas, além da entropia, podem se beneficiar do uso da estratégia *lazy* para a seleção de atributos, sendo elas baseadas no chi-quadrado, no coeficiente de *Cramer* e no índice *Gini*. No caso da medida *gain ratio* o cenário se inverteu e, além disso, ela obteve os melhores resultados quando comparada com as outras medidas *eager*. Por isso, uma sugestão de trabalho futuro é que seja feito um estudo sobre outras formas de adaptação da medida *gain ratio* proposta para a estratégia *lazy*.

Propõe-se também como continuação deste trabalho, a criação de uma métrica para cada medida avaliada, que seja capaz de estimar se uma base de dados específica pode se favorecer da estratégia de seleção de atributos *lazy* conforme proposto em [22] e, mais adiante, adaptar, para o contexto de seleção lazy, medidas que avaliam subconjuntos de atributos, tais como o *Correlation-based Feature Selection*[9] e o *Consistency-based Feature Selection* [18], que medem a capacidade de um determinado conjunto de atributos discriminar os valores do atributo classe.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] A. Asuncion e J. Newman. Uci machine learning repository. http://www.ics.uci.edu/ mlearn/mlrepository.html, 2007.

[2] L. Breiman, J. Friedman, R. Olshen e C. Stone. Classification and regression trees. 1984.

[3] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, (2):121–168, 1998.

[4] T. Cover e P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, (13):21–27, 1967.

[5] B. Dasarathy. Nearest neighbor (nn) norms: Nn pattern classification techniques. *IEEE Computer Society Press*, 1991.

[6] R. Duda, P. Hart e D. Stork. Pattern classification. *John Wiley & Sons*, 2001.

[7] U. M. Fayyad e K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. Em *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1029, 1993.

[8] I. Guyon, S. Gunn, M. Nikravesh e L. Zadeh. *Feature Extraction, Foundations and Applications*. Springer, 2006.

[9] M. A. Hall. A correlation-based feature selection for discrete and numeric class machine learning. Em *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[10] J. Han e M. Kamber. *Data Mining: Concepts and Techniques (2nd edition)*. Morgan Kaufmann, 2006.

[11] D. J. Hand, H. Mannila e P. Smyth. *Principles of Data Mining*. Bradford Book, Cambridge, 2000.

[12] K. Kira e L. Rendell. A practical approach to feature selection. Em *Proceedings of the 9th International Conference on Machine Learning*, pp. 249–256, 1992.

[13] I. Kononenko. Estimating attributes: Analysis and extensions of relief. Em *Proceedings of the 7th European Conference on Machine Learning*, pp. 171–182, 1994.

[14] I. Kononenko. On biases in estimating multi-valued attributes: Analysis and extensions of relieff. Em *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 1034–1040, 1995.

[15] B. Liu, W. Hsu e Y. Ma. Integrating classification and association rule mining. Em *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 80–86, 1998.

[16] H. Liu e H. Motoda. Computational methods of feature selection. *Chapman & Hall/CRC*, 2008.

[17] H. Liu e R. Setiono. Chi2: Feature selection and discretization of numeric attributes. Em *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pp. 388–391, 1995.

[18] H. Liu e R. Setiono. A probabilistic approach to feature selection: A filter solution. Em *Proceedings of the 13th International Conference on Machine Learning*, pp. 319–327, 1996.

[19] H. Liu e L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, (17):491–502, 2005.

[20] R. Menezes, A. Plastino, B. Zadrozny, R. Pereira, L. H. Merschmann e A. Freitas. Avaliação de uma nova medida para seleção *lazy* de atributos baseada no teste chi-quadrado. *Anais do V Workshop em Algoritmos e Aplicações de Mineração de Dados*, pp. 58–65, 2009.

[21] The University of Waikato. Weka (waikato environment for knowledge analysis) machine learning project http://www.cs.waikato.ac.nz/ml/weka/, junho de 2009.

[22] R. Pereira. Seleção *lazy* de atributos para a tarefa de classificação. Dissertação de Mestrado, 2009.

[23] R. Pereira, A. Plastino, B. Zadrozny, L. Merschmann e A. Freitas. Seleção *lazy* de atributos – uma nova perspectiva. *Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados*, pp. 1–9, 2008.

[24] J. Quinlan. Induction of decision trees. *Machine Learning*, (1):81–106, 1986.

[25] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[26] B. Ripley. Pattern recognition and neural networks. *Cambridge University Press*, 1996.

[27] M. R. Spiegel. *Estatística*. Makron Books, 1993.

[28] I. Witten e E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (2nd edition)*. Morgan Kaufmann, 2005.

[29] Y. Yang e J. O. Pedersen. A comparative study on feature selection in text categorization. Em *Proceedings of the 14th International Conference on Machine Learning*, pp. 412–420, 1997.

APÊNDICE

Apêndice A

Comparação entre Medidas *Lazy*

Apêndice B

Comparação entre Medidas *Eager*

APÊNDICE A – COMPARAÇÃO ENTRE MEDIDAS *LAZY*

Neste apêndice, são apresentadas as tabelas que comparam as melhores acurácias preditivas obtidas por cada uma das medidas *lazy* utilizando *3-NN* e *5-NN*.

Tabela A.1: Comparativo entre medidas *lazy usando 3-NN*

Base	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**	**Sem sel.**
anneal	**98,44**	**98,44**	98,11	**98,44**	98,33	98,00
audiology	**72,12**	**72,12**	68,58	70,35	70,35	65,93
autos	**81,46**	**81,46**	**81,46**	**81,46**	80,00	**81,46**
breast-cancer	73,43	**75,17**	**75,17**	73,78	73,08	70,28
breast-w	97,00	**97,28**	**97,28**	97,00	96,71	96,85
chess-Kr-vs-Kp	96,15	95,96	95,96	96,15	95,71	**96,59**
credit-a	85,94	**86,67**	**86,67**	85,80	**86,67**	84,20
diabetes	**78,26**	74,74	74,74	69,40	72,92	76,82
flags	61,34	61,86	61,34	62,89	**65,98**	55,15
glass	**75,70**	**75,70**	**75,70**	**75,70**	**75,70**	**75,70**
heart-cleveland	82,84	**83,17**	**83,17**	82,51	82,51	82,51
heart-hungarian	**83,67**	83,33	83,33	**83,67**	82,99	82,99
hepatitis	**86,45**	83,87	83,87	85,81	84,52	83,87
horse-colic	82,88	**85,05**	**85,05**	83,97	84,51	77,17
hypo-thyroid	**97,53**	93,58	93,58	93,19	93,48	93,21
ionosphere	**92,31**	**92,31**	**92,31**	92,02	90,31	90,60
labor	**98,25**	96,49	96,49	**98,25**	96,49	96,49
letter-recog	90,59	**90,61**	88,39	90,59	90,38	90,57
lymph	83,78	83,78	**84,46**	83,78	81,76	83,78
mol-bio-promot	88,68	**89,62**	**89,62**	88,68	88,68	80,19
mol-bio-splice	**91,22**	86,46	86,46	87,46	91,16	77,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	**95,50**	95,46	95,32	95,37	95,41	95,28
pendigits	96,54	96,28	96,20	96,49	96,49	**96,74**
postoperative	**71,11**	**71,11**	**71,11**	**71,11**	**71,11**	68,89
primary-tumor	**44,54**	**44,54**	44,25	43,95	43,66	43,66
solar-flare1	72,14	71,21	69,35	**72,45**	71,21	65,33
solar-flare2	75,98	75,70	74,48	**76,17**	75,14	74,02
sonar	**88,46**	79,81	79,81	79,33	79,33	86,06
soybean-large	**92,39**	90,19	90,48	91,51	91,36	91,51
spambase	**93,59**	93,41	93,41	93,57	93,39	93,31
statlog-heart	**84,44**	83,70	83,70	**84,44**	**84,44**	80,74
statlog-segment	93,90	93,90	93,59	93,90	93,94	**93,98**
statlog-vehicle	71,51	**72,22**	71,28	71,28	71,04	71,28
thyroid-sick	97,32	**97,59**	**97,59**	97,40	97,51	97,08
vote	**95,63**	95,40	95,40	95,40	95,40	92,41
vowel	84,14	84,14	82,42	84,14	84,14	**84,65**
waveform-5000	79,06	78,96	78,96	78,98	**79,10**	78,58
wine	**98,88**	97,19	96,07	**98,88**	98,31	96,63
zoo	**97,03**	**97,03**	95,05	93,07	**97,03**	93,07
Total	**26**	24	24	8	9	8

Tabela A.2: Comparativo entre medidas usando *lazy 5-NN*

Base	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**	**Sem sel.**
anneal	97,77	**98,00**	97,22	97,55	97,44	97,22
audiology	**73,01**	69,91	67,26	71,68	68,14	60,62
autos	77,07	77,07	**77,56**	76,59	76,10	77,07
breast-cancer	75,17	**75,52**	**75,52**	74,48	74,48	74,13
breast-w	97,14	**97,42**	**97,42**	97,14	96,42	97,00
chess-Kr-vs-Kp	95,62	95,56	95,56	95,56	95,18	**96,06**
credit-a	85,80	86,67	86,67	85,80	**87,10**	84,64
diabetes	**77,99**	74,35	74,35	69,79	72,66	77,21
flags	63,40	61,86	62,89	63,40	**64,43**	57,73
glass	**73,83**	**73,83**	**73,83**	**73,83**	**73,83**	**73,83**
heart-cleveland	82,84	**83,17**	**83,17**	82,84	82,84	82,84
heart-hungarian	83,33	82,99	82,99	**83,67**	**83,67**	82,31
hepatitis	**85,16**	83,87	83,87	**85,16**	84,52	84,52
horse-colic	82,07	**85,60**	**85,60**	83,70	84,78	77,45
hypo-thyroid	**97,30**	93,56	93,56	93,03	93,45	93,27
ionosphere	**92,31**	**92,31**	**92,31**	92,02	89,74	89,74
labor	96,49	96,49	96,49	**98,25**	94,74	91,23
letter-recog	89,83	89,79	87,54	89,83	89,56	**89,84**
lymph	**85,14**	**85,14**	81,08	84,46	81,08	83,78
mol-bio-promot	**89,62**	86,79	86,79	**89,62**	86,79	79,25
mol-bio-splice	**91,16**	86,24	86,24	87,24	**91,16**	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	95,53	95,53	95,50	**95,57**	95,50	95,50
pendigits	96,10	95,80	95,94	96,02	95,95	**96,45**
postoperative	**71,11**	**71,11**	**71,11**	**71,11**	**71,11**	**71,11**
primary-tumor	46,61	**46,90**	46,61	45,72	46,02	46,31
solar-flare1	**70,90**	68,73	68,11	70,59	69,35	66,56
solar-flare2	**76,17**	75,70	74,20	**76,17**	75,05	73,83
sonar	**85,58**	77,40	77,40	78,37	77,88	84,62
soybean-large	**91,95**	89,46	89,90	91,07	90,78	90,78
spambase	93,28	93,26	93,26	93,26	**93,33**	93,20
statlog-heart	**85,93**	85,56	85,56	**85,93**	84,44	82,22
statlog-segment	92,90	92,90	92,51	92,90	**93,07**	**93,07**
statlog-vehicle	71,04	**71,63**	70,21	71,04	70,69	70,69
thyroid-sick	97,48	**97,59**	**97,59**	97,38	97,48	96,58
vote	**95,40**	95,17	95,17	**95,40**	94,48	92,87
vowel	78,08	78,08	76,77	77,88	78,08	**78,59**
waveform-5000	**80,84**	80,32	80,32	80,70	80,50	79,74
wine	**100,00**	96,63	93,82	**100,00**	97,75	96,07
zoo	92,08	**93,07**	**93,07**	**93,07**	**93,07**	**93,07**
Total	**26**	**26**	23	23	11	13

APÊNDICE B – COMPARAÇÃO ENTRE MEDIDAS *EAGER*

Neste apêndice, são apresentadas as tabelas que comparam as melhores acurácias preditivas obtidas por cada uma das medidas *eager* utilizando *1-NN*, *3-NN* e *5-NN*.

Tabela B.1: Comparativo entre medidas usando *eager 1-NN*

Base	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**	**Sem sel.**
anneal	99,33	**99,55**	**99,55**	**99,55**	**99,55**	99,22
audiology	75,66	76,11	**78,32**	75,22	76,55	76,11
autos	88,29	88,29	89,27	88,29	**89,76**	85,85
breast-cancer	72,73	74,13	74,13	73,78	**74,48**	69,93
breast-w	**97,28**	**97,28**	**97,28**	**97,28**	**97,28**	97,14
chess-Kr-vs-Kp	96,43	96,43	96,43	96,40	96,53	**96,56**
credit-a	85,51	85,51	85,51	84,35	**86,09**	82,32
diabetes	**77,34**	74,35	74,35	72,66	73,83	76,43
flags	63,40	60,82	60,82	60,31	**64,43**	59,79
glass	**77,10**	**77,10**	**77,10**	**77,10**	**77,10**	**77,10**
heart-cleveland	**84,49**	**84,49**	**84,49**	**84,49**	81,19	80,53
heart-hungarian	82,65	82,65	82,65	81,97	**84,01**	80,27
hepatitis	83,87	**84,52**	**84,52**	83,23	**84,52**	83,87
horse-colic	83,42	83,42	83,42	82,88	**85,33**	78,53
hypo-thyroid	**94,59**	93,61	93,61	93,58	93,61	91,52
ionosphere	**93,45**	93,16	93,16	93,16	**93,45**	92,59
labor	96,49	**98,25**	**98,25**	**98,25**	**98,25**	96,49
letter-recog	91,86	91,86	91,80	91,86	**91,87**	**91,87**
lymph	85,14	85,14	85,14	85,81	**87,16**	82,43
mol-bio-promot	**90,57**	**90,57**	**90,57**	**90,57**	**90,57**	80,19
mol-bio-splice	**89,72**	**89,72**	**89,72**	87,55	**89,72**	73,32
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	94,75	94,75	**94,82**	94,75	94,32	94,25
pendigits	96,31	96,31	96,31	96,31	96,31	**97,05**
postoperative	**71,11**	**71,11**	70,00	70,00	70,00	63,33
primary-tumor	**42,77**	41,00	41,89	42,18	40,41	38,35
solar-flare1	**72,76**	**72,76**	70,28	**72,76**	70,28	65,94
solar-flare2	75,89	75,89	74,95	74,48	**75,98**	73,45
sonar	**88,46**	79,81	79,81	80,29	82,69	86,54
soybean-large	**93,41**	92,09	90,92	92,83	91,22	92,24
spambase	93,13	93,13	93,13	93,13	**93,54**	92,98
statlog-heart	**84,81**	**84,81**	**84,81**	**84,81**	84,07	84,07
statlog-segment	94,76	94,76	94,76	94,76	**95,19**	94,68
statlog-vehicle	71,39	70,45	**71,99**	70,45	70,92	70,92
thyroid-sick	97,72	**97,77**	**97,77**	**97,77**	**97,77**	97,45
vote	95,17	95,17	95,17	95,17	**95,63**	92,18
vowel	**89,80**	**89,80**	**89,80**	**89,80**	**89,80**	**89,80**
waveform-5000	74,62	74,62	74,62	74,62	**75,36**	73,82
wine	**98,88**	**98,88**	97,75	**98,88**	**98,88**	98,31
zoo	97,03	97,03	97,03	97,03	**98,02**	96,04
Total	17	15	15	12	**26**	6

Tabela B.2: Comparativo entre medidas usando *eager 3-NN*

Base	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**	**Sem sel.**
anneal	98,00	98,66	98,33	98,66	**98,89**	98,00
audiology	68,58	69,91	**70,35**	67,70	69,03	65,93
autos	81,46	**81,95**	81,46	**81,95**	80,00	81,46
breast-cancer	72,73	**74,13**	**74,13**	73,08	73,78	70,28
breast-w	**97,00**	**97,00**	**97,00**	**97,00**	**97,00**	96,85
chess-Kr-vs-Kp	96,09	96,09	96,09	96,06	95,53	**96,59**
credit-a	85,51	85,51	85,51	85,80	**86,52**	84,20
diabetes	**77,99**	74,09	74,09	72,53	73,57	76,82
flags	**63,40**	62,37	**63,40**	**63,40**	62,89	55,15
glass	**75,70**	**75,70**	**75,70**	**75,70**	**75,70**	**75,70**
heart-cleveland	**84,16**	**84,16**	**84,16**	**84,16**	82,51	82,51
heart-hungarian	**82,99**	**82,99**	**82,99**	**82,99**	**82,99**	**82,99**
hepatitis	**86,45**	85,81	85,81	85,81	**86,45**	83,87
horse-colic	83,15	83,15	83,15	82,07	**85,05**	77,17
hypo-thyroid	**94,80**	93,58	93,58	93,53	93,58	93,21
ionosphere	**91,45**	**91,45**	**91,45**	**91,45**	91,17	90,60
labor	96,49	**98,25**	**98,25**	**98,25**	96,49	96,49
letter-recog	90,43	90,43	90,43	90,43	90,40	**90,57**
lymph	83,11	83,11	83,11	85,14	**88,51**	83,78
mol-bio-promot	**91,51**	**91,51**	**91,51**	**91,51**	**91,51**	80,19
mol-bio-splice	**88,46**	**88,46**	**88,46**	86,08	**88,46**	77,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	95,41	95,37	**95,59**	95,37	95,28	95,28
pendigits	95,95	95,95	95,95	95,95	95,95	**96,74**
postoperative	**71,11**	**71,11**	**71,11**	**71,11**	**71,11**	68,89
primary-tumor	44,25	43,95	43,66	**44,84**	44,54	43,66
solar-flare1	**73,07**	**73,07**	71,21	**73,07**	71,21	65,33
solar-flare2	**75,70**	**75,70**	74,86	74,48	75,52	74,02
sonar	**87,98**	77,88	77,88	78,37	79,33	86,06
soybean-large	**92,39**	91,36	90,48	**92,39**	91,22	91,51
spambase	93,35	93,35	93,35	93,35	**93,72**	93,31
statlog-heart	**84,81**	**84,81**	**84,81**	**84,81**	84,44	80,74
statlog-segment	93,94	93,94	93,94	93,94	93,94	**93,98**
statlog-vehicle	71,39	71,39	**72,22**	71,39	71,63	71,28
thyroid-sick	97,48	97,51	97,51	97,51	**97,59**	97,08
vote	95,17	94,94	94,94	94,94	**95,40**	92,41
vowel	84,14	84,14	84,14	84,14	84,14	**84,65**
waveform-5000	79,04	79,04	79,04	79,04	**79,24**	78,58
wine	97,19	97,19	97,19	97,19	**97,75**	96,63
zoo	**97,03**	**97,03**	**97,03**	**97,03**	**97,03**	93,07
Total	**19**	16	17	16	18	8

Tabela B.3: Comparativo entre medidas usando *eager 5-NN*

Base	**ENT**	**CHI**	**CRV**	**GIN**	**GRT**	**Sem sel.**
anneal	97,22	97,88	97,44	97,88	**98,55**	97,22
audiology	**69,47**	68,14	66,37	68,58	65,93	60,62
autos	**77,07**	**77,07**	**77,07**	**77,07**	75,61	**77,07**
breast-cancer	74,13	**75,52**	**75,52**	74,48	74,13	74,13
breast-w	**97,14**	**97,14**	**97,14**	**97,14**	97,00	97,00
chess-Kr-vs-Kp	95,65	95,65	95,65	95,62	95,24	**96,06**
credit-a	85,94	85,94	85,94	85,94	**86,96**	84,64
diabetes	**78,13**	74,09	74,09	72,79	73,70	77,21
flags	62,37	62,37	63,40	63,40	**63,92**	57,73
glass	**73,83**	**73,83**	**73,83**	**73,83**	**73,83**	**73,83**
heart-cleveland	83,50	**83,83**	**83,83**	**83,83**	83,17	82,84
heart-hungarian	**84,01**	83,33	83,33	82,99	83,67	82,31
hepatitis	84,52	84,52	84,52	**85,16**	**85,16**	84,52
horse-colic	83,15	83,70	83,70	82,34	**85,60**	77,45
hypo-thyroid	**94,62**	93,56	93,56	93,48	93,56	93,27
ionosphere	**90,88**	**90,88**	**90,88**	**90,88**	90,03	89,74
labor	**96,49**	**96,49**	**96,49**	**96,49**	92,98	91,23
letter-recog	89,67	89,67	89,67	89,67	89,56	**89,84**
lymph	84,46	84,46	84,46	84,46	**86,49**	83,78
mol-bio-promot	**87,74**	**87,74**	**87,74**	**87,74**	**87,74**	79,25
mol-bio-splice	**88,15**	**88,15**	**88,15**	85,86	**88,15**	79,37
mushroom	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits	95,55	**95,60**	95,52	**95,60**	95,50	95,50
pendigits	95,54	95,54	95,54	95,54	95,54	**96,45**
postoperative	**71,11**	**71,11**	**71,11**	**71,11**	**71,11**	**71,11**
primary-tumor	45,72	46,02	46,02	**46,61**	44,25	46,31
solar-flare1	**71,21**	**71,21**	68,73	**71,21**	69,66	66,56
solar-flare2	**75,70**	**75,70**	74,77	74,77	75,52	73,83
sonar	83,65	78,37	78,37	77,88	78,85	**84,62**
soybean-large	**91,95**	91,22	89,90	**91,95**	90,78	90,78
spambase	93,24	93,24	93,24	93,24	**93,44**	93,20
statlog-heart	**85,56**	**85,56**	**85,56**	**85,56**	**85,56**	82,22
statlog-segment	**93,07**	**93,07**	**93,07**	**93,07**	**93,07**	**93,07**
statlog-vehicle	70,57	70,45	71,16	70,45	**71,63**	70,69
thyroid-sick	97,61	**97,64**	**97,64**	**97,64**	**97,64**	96,58
vote	**95,17**	94,94	94,94	94,94	**95,17**	92,87
vowel	78,28	78,28	78,28	78,28	78,28	**78,59**
waveform-5000	**80,82**	**80,82**	**80,82**	**80,82**	80,72	79,74
wine	**96,63**	**96,63**	96,07	**96,63**	**96,63**	96,07
zoo	**93,07**	**93,07**	**93,07**	**93,07**	**93,07**	**93,07**
Total	**26**	20	16	20	19	11