IV Workshop do Curso de Tecnologia em Sistemas de Computação

Pesquisa e Aplicação em Mineração de Dados

Alexandre Plastino
plastino@ic.uff.br

Instituto de computação uff

Mineração de Dados (Data Mining):

- Processo de descoberta de novas informações e conhecimento, no formato de regras e padrões, a partir de grandes bases de dados.

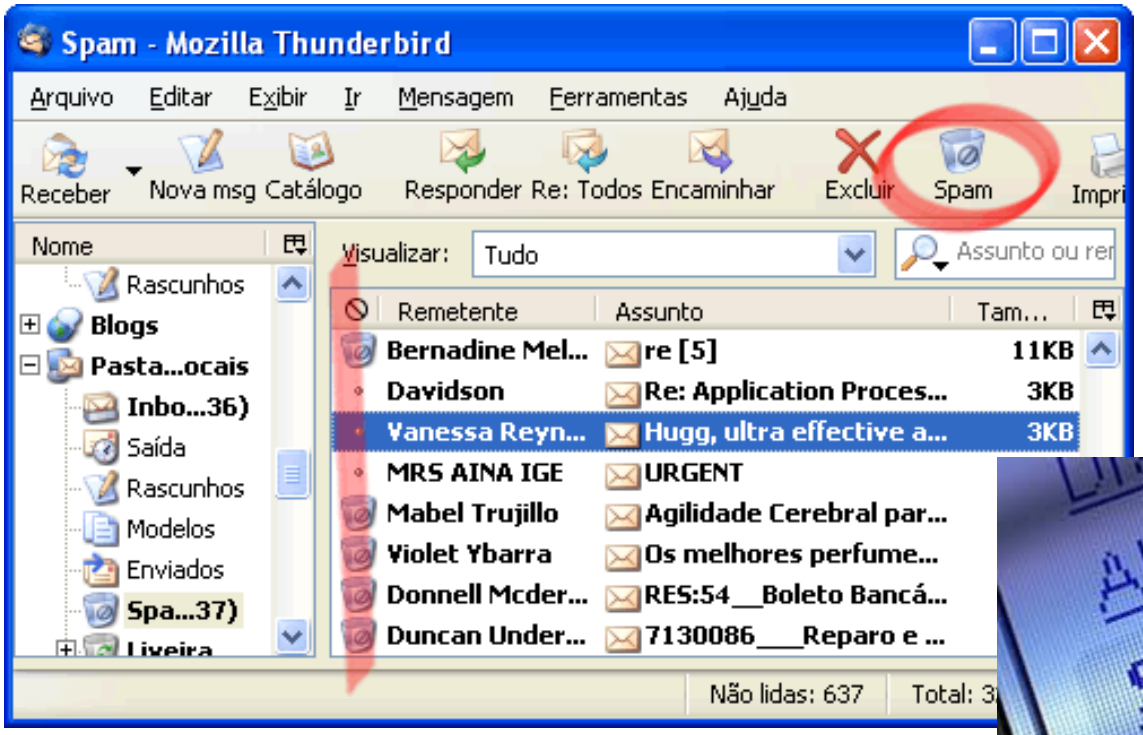
Mineração descritiva:

- compreender os dados;

Mineração preditiva:

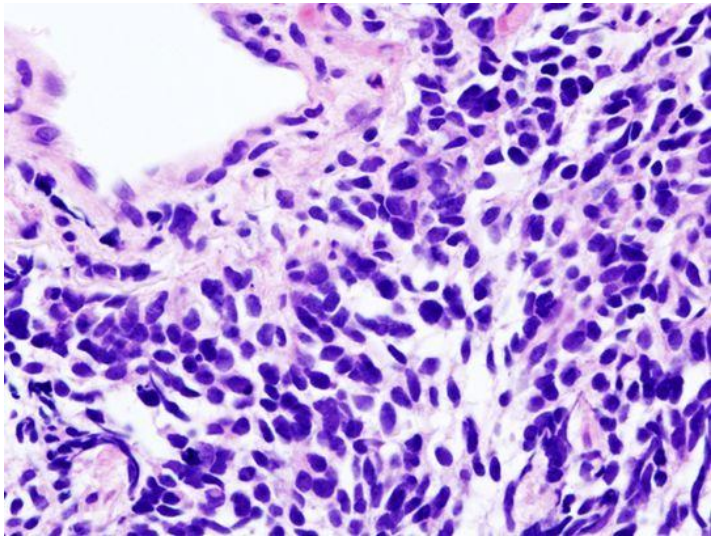
- prever/estimar valores ainda não conhecidos.

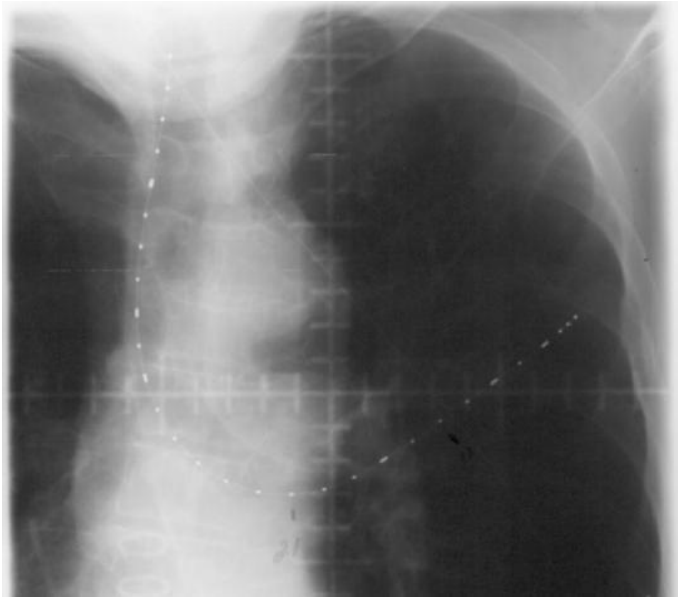
Aplicações de Mineração de Dados



- Detecção de SPAM (classificação)

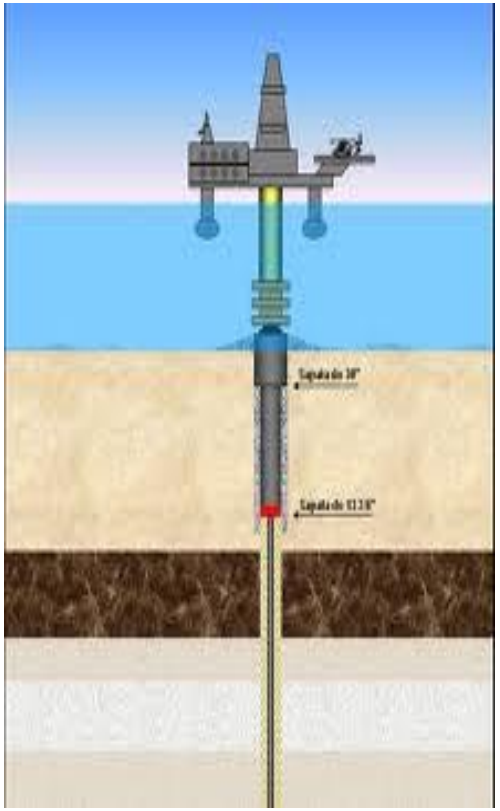
Aplicações de Mineração de Dados





- Detecção de patologias por análise de imagens (classificação)

Aplicações de Mineração de Dados



- Previsão/Estimativa da permeabilidade de rochas por ressonância magnética (classificação)

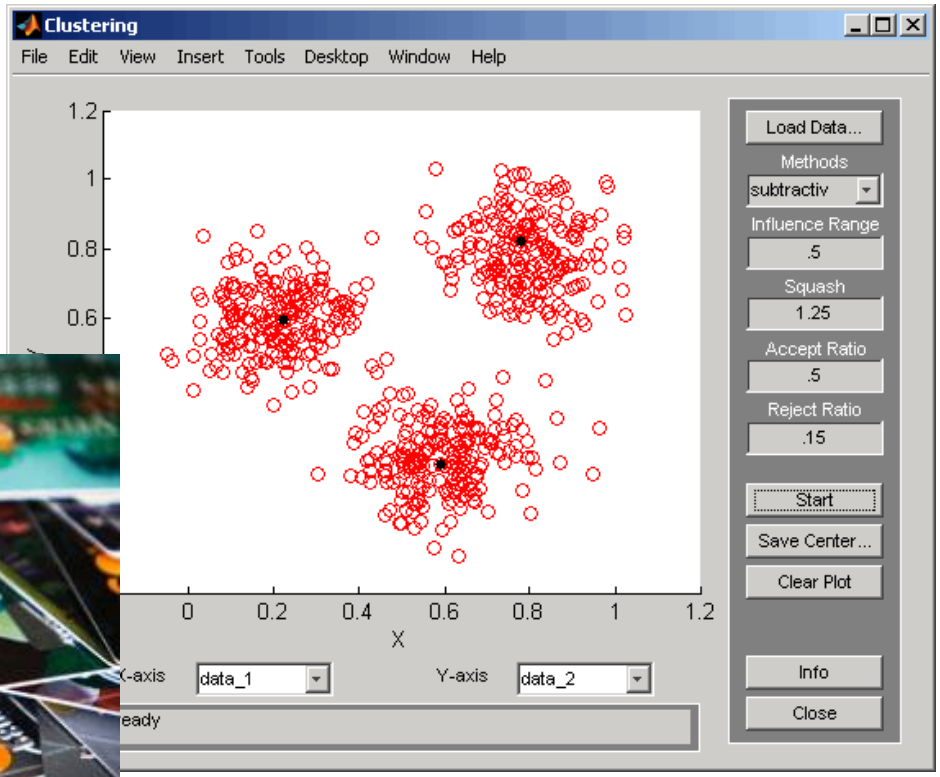
Aplicações de Mineração de Dados

- Market Basket Analysis
  (regras de associação)

Fralda $\rightarrow$ Cerveja

Aplicações de Mineração de Dados

- Detecção de Fraudes (clusterização/ classificação)



Aplicações de Mineração de Dados

- Mineração de opiniões e sentimentos (classificação)

Mineração Preditiva:

- Deseja-se prever o valor desconhecido de um determinado atributo, a partir da análise dos dados armazenados na base (base de treinamento).

Mineração Descritiva:

- Padrões e regras descrevem características importantes dos dados com os quais se está trabalhando.

Processos de Mineração de Dados

- Regras de Associação
- Padrões Sequenciais
- Classificação
- Clusterização

Regras de Associação

Uma regra de associação representa um padrão de relacionamento entre itens de dados do domínio da aplicação que ocorre com uma determinada frequência na base.

Id-Transação (TID)	Itens Comprados
1	leite, pão, refrigerante
2	cerveja, carne
3	cerveja, fralda, leite, refrigerante
4	cerveja, fralda, leite, pão
5	fralda, leite, refrigerante

{fralda} $\Rightarrow$ {cerveja}	confiança de 66%	suporte de 40%
{fralda} $\Rightarrow$ {leite}	confiança de 100%	suporte de 60%
{leite} $\Rightarrow$ {fralda}	confiança de 75%	suporte de 60%
{carne} $\Rightarrow$ {cerveja}	confiança de 100%	suporte de 20%

Pesquisa em Regras de Associação

- Desenvolvimento de algoritmos eficientes:
  Apriori (1993/94), *FPGrowth* (2000), etc.

- Criação de medidas de interesse:
  suporte, confiança, *lift*, *rule interest*, etc.

- Desenvolvimento de algoritmos para domínios específicos:
  bioinformática, mineração de texto, etc.

- Novas variações:
  regras negativas, exceções, etc.

Padrões de Seqüências

Padrões de seqüências representam sequências de conjuntos de itens que ocorrem nas transações de diferentes consumidores, com determinada frequência (na ordem especificada).

Consumidor	Data/Hora	Produtos
João	01.08.2001/17:01	leite, pão
João	03.08.2001/14:25	carne, cerveja
João	10.08.2001/21:15	queijo, manteiga, sal
Marcos	05.08.2001/10:16	leite, ovos
Marcos	08.08.2001/18:30	queijo, manteiga

Padrão de seqüência: {(leite) (queijo, manteiga)}

→ Cada transação deve ser definida por um consumidor, pelo instante (tempo) em que ocorreu e por um conjunto de itens.

Pesquisa em Padrões Sequenciais

- Em geral, “a reboque” do desenvolvimentos de novos algoritmos e ideias no contexto de regras de associação:

Apriori (1993/94) $\rightarrow$ GSP (1995)

Classificação

Um classificador identifica, entre um conjunto pré-definido de classes, aquela a qual pertence um elemento, a partir de seus atributos.

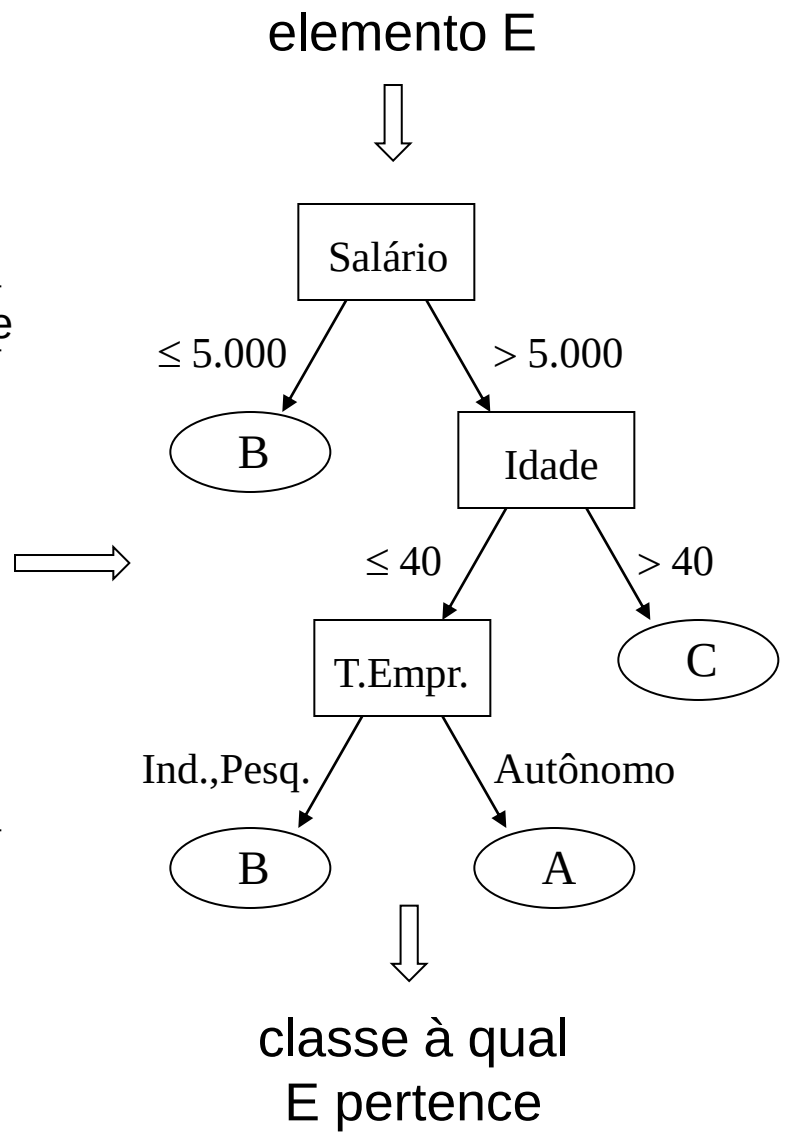
→ Implementar/minerar um classificador significa gerar/descobrir a função que realiza tal mapeamento.

→ O processo de classificação necessita de uma base de treinamento.

ID	Salário	Idade	Tipo Emprego	Classe
1	3.000	30	Autônomo	B
2	4.000	35	Indústria	B
3	7.000	50	Pesquisa	C
4	6.000	45	Autônomo	C
5	7.000	30	Pesquisa	B
6	6.000	35	Indústria	B
7	6.000	35	Autônomo	A
8	7.000	30	Autônomo	A
9	4.000	45	Indústria	B

Classificação

ID	Salário	Idade	Tipo Emprego	Classe
1	3.000	30	Autônomo	B
2	4.000	35	Indústria	B
3	7.000	50	Pesquisa	C
4	6.000	45	Autônomo	C
5	7.000	30	Pesquisa	B
6	6.000	35	Indústria	B
7	6.000	35	Autônomo	A
8	7.000	30	Autônomo	A
9	4.000	45	Indústria	B



Pesquisa em Classificação

- Desenvolvimento de algoritmos eficientes:
  em termos de tempo e acurácia preditiva.

- Desenvolvimento de algoritmos para domínios específicos:
  detecção de SPAM, análise de sentimento, etc.

- Novas variações:
  classificação hierárquica, classificação multirrótulo, etc.

Agrupamento (Clusterização)

Agrupamento é o resultado da identificação de um conjunto finito de categorias (ou grupos - clusters) que contêm objetos similares.
→ Grupos/classes não são previamente definidos.

Exemplo: Deseja-se separar os clientes em grupos de forma que aqueles que apresentam o mesmo comportamento de consumo fiquem no mesmo grupo.

Cada tupla deste exemplo indica a quantidade total de produtos consumidos e o preço médio destes produtos relativos a cada consumidor.

Consumidor	Qtd.Méd.Tot.Prods.	Preç.Méd.Prods.
1	2	1.700
2	10	1.800
3	2	100
4	3	2.000
5	12	2.100
6	3	200
7	4	2.300
8	11	2.040
9	3	150

Agrupamento (Clusterização)

Consumidor	Qtd.Méd.	Preç.Méd.
1	2	1.700
2	10	1.800
3	2	100
4	3	2.000
5	12	2.100
6	3	200
7	4	2.300
8	11	2.040
9	3	150

Grupo	Consumidor	Qtd.Méd.	Preç.Méd.
1	1	2	1.700
	4	3	2.000
	7	4	2.300
2	2	10	1.800
	5	12	2.100
	8	11	2.040
3	3	2	100
	6	3	200
	9	3	150

Cada grupo identificado é caracterizado por consumidores semelhantes em relação à quantidade média total e ao preço médio dos produtos consumidos.

Pesquisa em Clusterização

- Desenvolvimento de algoritmos eficientes.

- Desenvolvimento de algoritmos que tratem ruídos.

- Identificação adequada do número de clusters.

- Visto também como problema de otimização combinatória.

Ferramenta Weka

(Waikato Environment for Knowledge Analysis)

http://www.cs.waikato.ac.nz/ml/weka/

Livros sobre Mineração de Dados