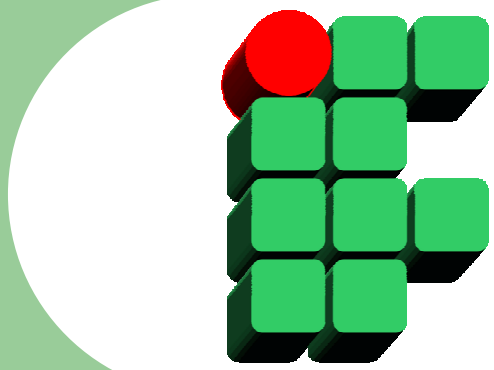


Mineração de Dados

Extraído dos trabalhos de

Liliane Santos, Menandro Santana, Sandoval Costa (UFBA)

Eduardo Massao Arakaki, Marcela Fontes Lima Guerra

(UFPE)

Motivação

- A informatização dos meios produtivos permitiu a geração de grandes volumes de dados:
  - Transações eletrônicas;
  - Novos equipamentos científicos e industriais para observação e controle;
  - Dispositivos de armazenamento em massa;
- Aproveitamento da informação permite ganho de competitividade: "*conhecimento é poder (e poder = $$!)*"

Motivação

- Os recursos de análise de dados tradicionais são inviáveis para acompanhar esta evolução
- *“Morrendo de sede por conhecimento em um oceano de dados”*

Motivação

- Solução:
  - ferramentas de automatização das tarefas repetitivas e sistemática de análise de dados
  - ferramentas de auxílio para as tarefas cognitivas da análise
  - integração das ferramentas em sistemas apoiando o processo completo de descoberta de conhecimento para tomada de decisão

Exemplo Preliminar

- Um problema do mundo dos negócios: entender o perfil dos clientes
  - desenvolvimento de novos produtos;
  - controle de estoque em postos de distribuição;
  - propaganda mal direcionada gera maiores gastos e desestimula o possível interessado a procurar as ofertas adequadas;
- Quais são meus clientes típicos?

Exemplo

Como Descubro Estes DADOS ????

Descoberta de Conhecimento em Bancos de Dados

- “O processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis de uma fonte de dados”;
- “Torture os dados até eles confessarem”;
- O que é um padrão interessante ? (válido, novo, útil e interpretável)

Conceitos

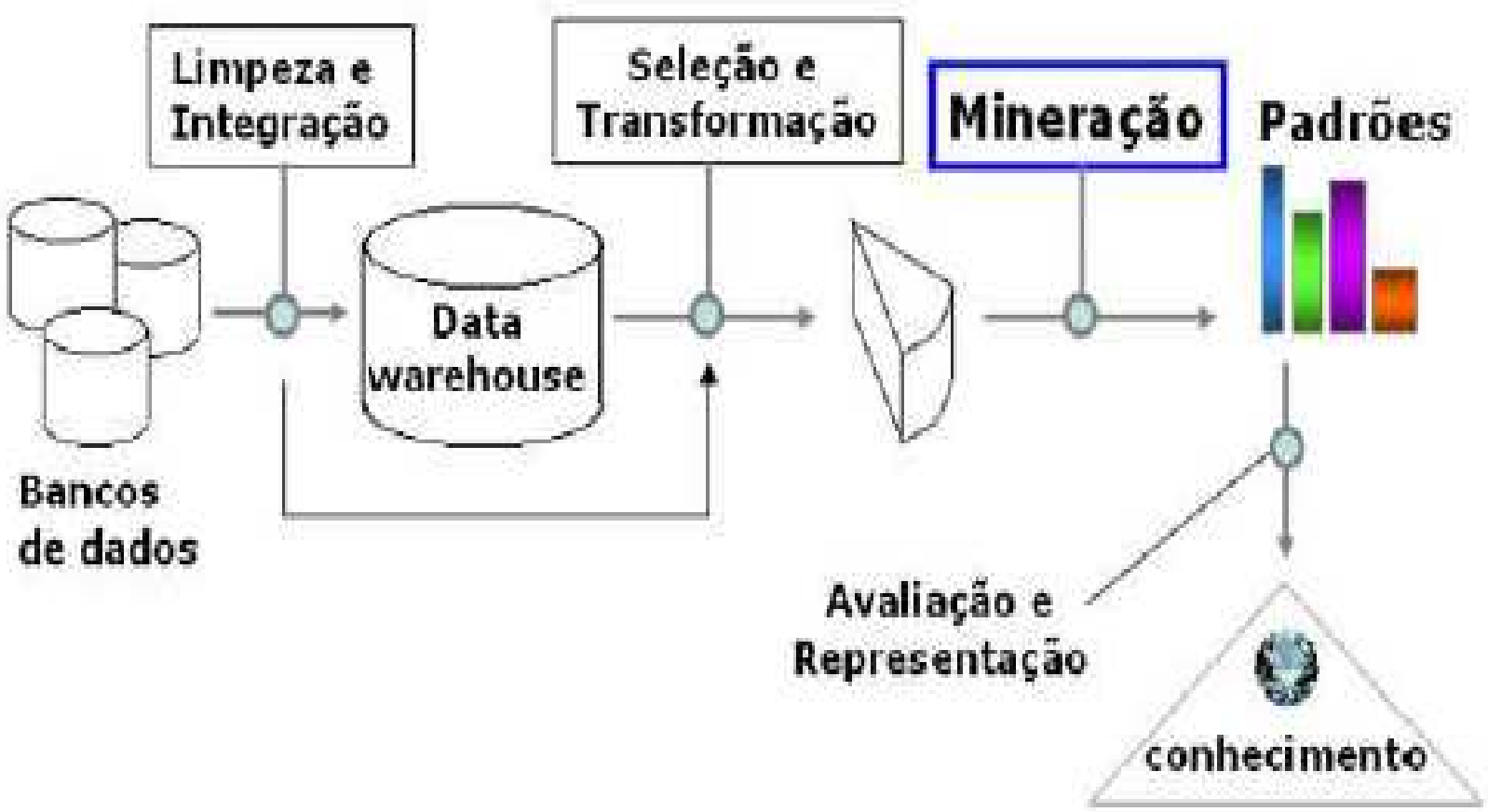
KDD x Data Mining

- Mineração de dados é o passo do processo de KDD que produz um conjunto de padrões sob um custo computacional aceitável;
- KDD utiliza algoritmos de *data mining* para extrair padrões classificados como "conhecimento". Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados;

Etapas do KDD

- Limpeza de dados
- Integração dos dados
  - Data Warehouse
- Seleção
- Transformação dos dados
- Mineração
- Avaliação ou pós-processamento
- Visualização dos resultados

Etapas do KDD



Áreas de Relação do KDD

Áreas de Relação do KDD

- Aprendizado de máquina
- Reconhecimento de padrões
- Base de dados
- Estatística e Matemática
- Sistemas Especialistas
- Visualização de dados

Aplicações da Mineração de dados

- Comércio
  - Real
  - Virtual
- Medicina
- Detecção de Fraudes
- Inteligência Competitiva
  - Concorrentes
  - Tendências do Mercado

Exemplos

- Áreas de aplicações potenciais:
  - Vendas e Marketing
    - *Identificar padrões de comportamento de consumidores*
    - *Associar comportamentos à características demográficas de consumidores*
    - *Campanhas de marketing direto (mailing campaigns)*
    - *Identificar consumidores “leais”*

Exemplos

- Áreas de aplicações potenciais:
  - Bancos
    - *Identificar padrões de fraudes (cartões de crédito)*
    - *Identificar características de correntistas*
    - *Mercado Financeiro ($$$)*

Exemplos

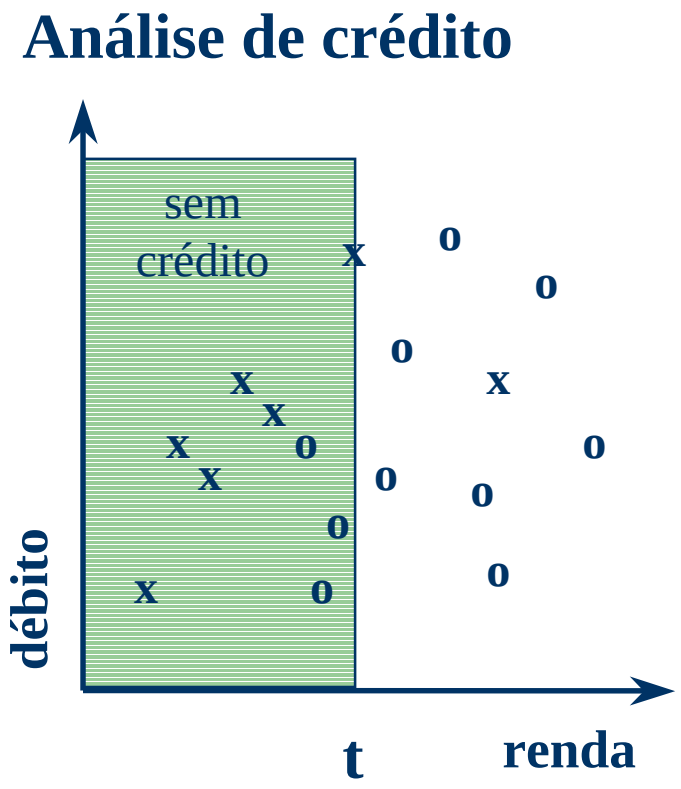
- Áreas de aplicações potenciais
  - Médica
    - *Comportamento de pacientes*
    - *Identificar terapias de sucessos para diferentes tratamentos*
    - *Fraudes em planos de saúdes*
    - *Comportamento de usuários de planos de saúde*

Quais Tarefas de Mineração são utilizadas?

Tarefas de Mineração de Dados

- Análise de Regras de Associação
- Análise de Padrões Sequenciais
- Classificação
- Análise de Clusters (agrupamentos) – Segmentação
- Análise de Outliers (exceções)
- Estimativa (ou regressão)
- Sumarização

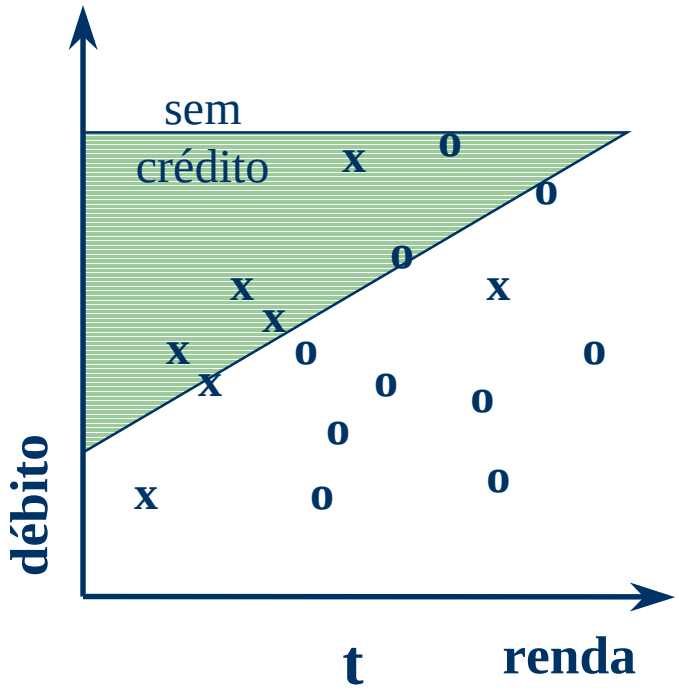
Exemplo de previsão (I)



- Um hiperplano paralelo de separação: pode ser interpretado diretamente como uma regra:
  - se a renda é menor que *t*, então o crédito não deve ser liberado
- Exemplo:
  - árvores de decisão;
  - indução de regras

x: exemplo recusado
o: exemplo aceito

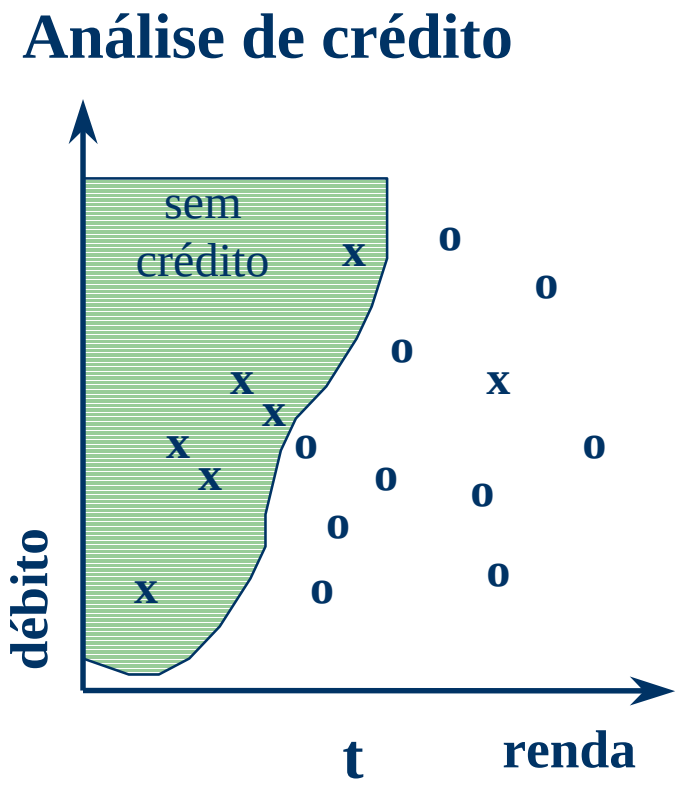
Exemplo de previsão (II)



- Hiperplano oblíquo: melhor separação:
- Exemplos:
  - regressão linear;
  - perceptron;

x: exemplo recusado
o: exemplo aceito

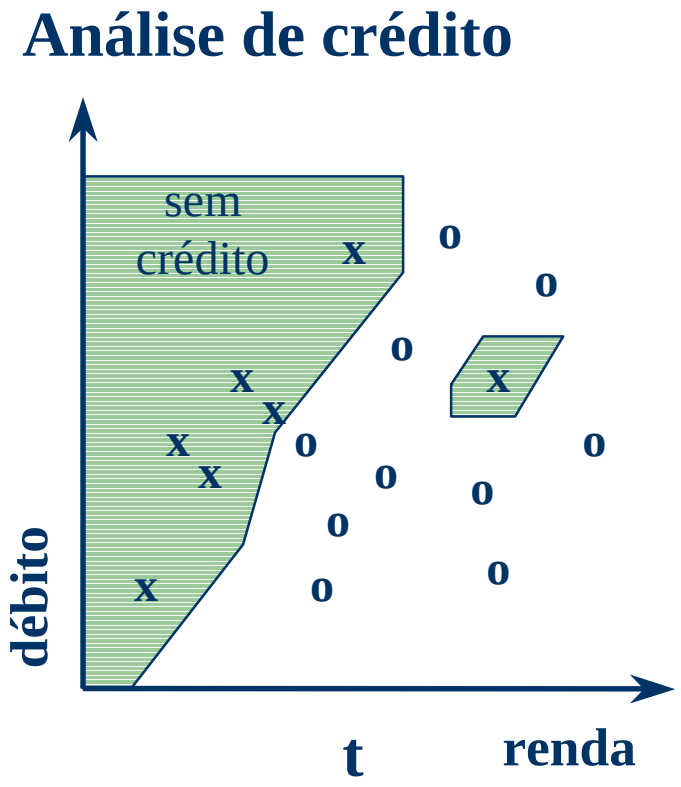
Exemplo de previsão (III)



x: exemplo recusado
o: exemplo aceito

- Superfície não linear: melhor poder de classificação, pior interpretação;
- Exemplos:
  - perceptrons multicamadas;
  - regressão não-linear;

Exemplo de previsão (IV)



- Métodos baseado em exemplos;
- Exemplos:
  - k-vizinhos mais próximos;
  - raciocínio baseado em casos;

x: exemplo recusado
o: exemplo aceito

Análise de Clusters (agrupamentos) – Segmentação

- Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos

Análise de Outliers (exceções)

- Identificação de dados que não apresentam o comportamento geral

Estimativa (ou regressão)

- Usada para definir um valor para alguma variável contínua desconhecida

Sumarização

- Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados

Análise de Regras de Associação

ID	Compras
1	Pão, **Leite**, **Manteiga**
2	**Leite**,Açucar
3	**Leite**, **Manteiga**
4	Manteiga, Açúcar

Leite → Manteiga

$$\text{Suporte} = \frac{\text{número de clientes que compraram Leite, Manteiga}}{\text{Total de clientes}} = 50\%$$

$$\text{Confiança} = \frac{\text{número de clientes que compraram Leite, Manteiga}}{\text{número de clientes que compraram Leite}} = 66{,}6\%$$

Análise de Padrões Sequenciais

Itens = { TV, Vídeo , DVD, FitaDVD, ... }

ITEMSET >> ITEMSET >> ITEMSET >> ... >>ITEMSET

Análise de Padrões Sequenciais

1	{TV , Rádio} >>{DVD}
2	{Computador}
3	{TV} >> {Rádio, DVD}
4	{Rádio} >>{Comp}
5	{Comp} >> {Impressora}

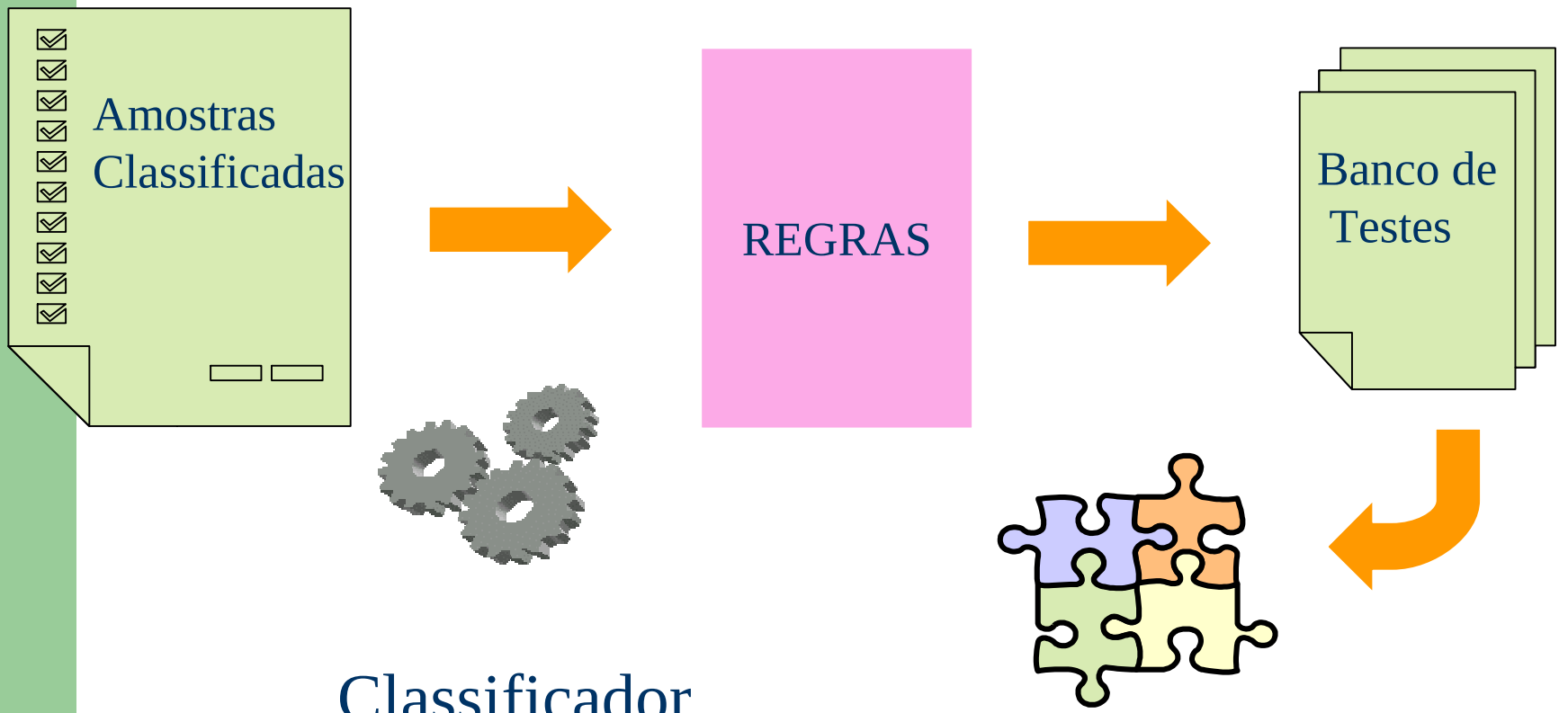
< {TV} , {DVD} >

$$\text{Suporte} = \frac{\text{número de clientes que compraram TV, DVD em seqüência}}{\text{Total de clientes}} = 40\%$$

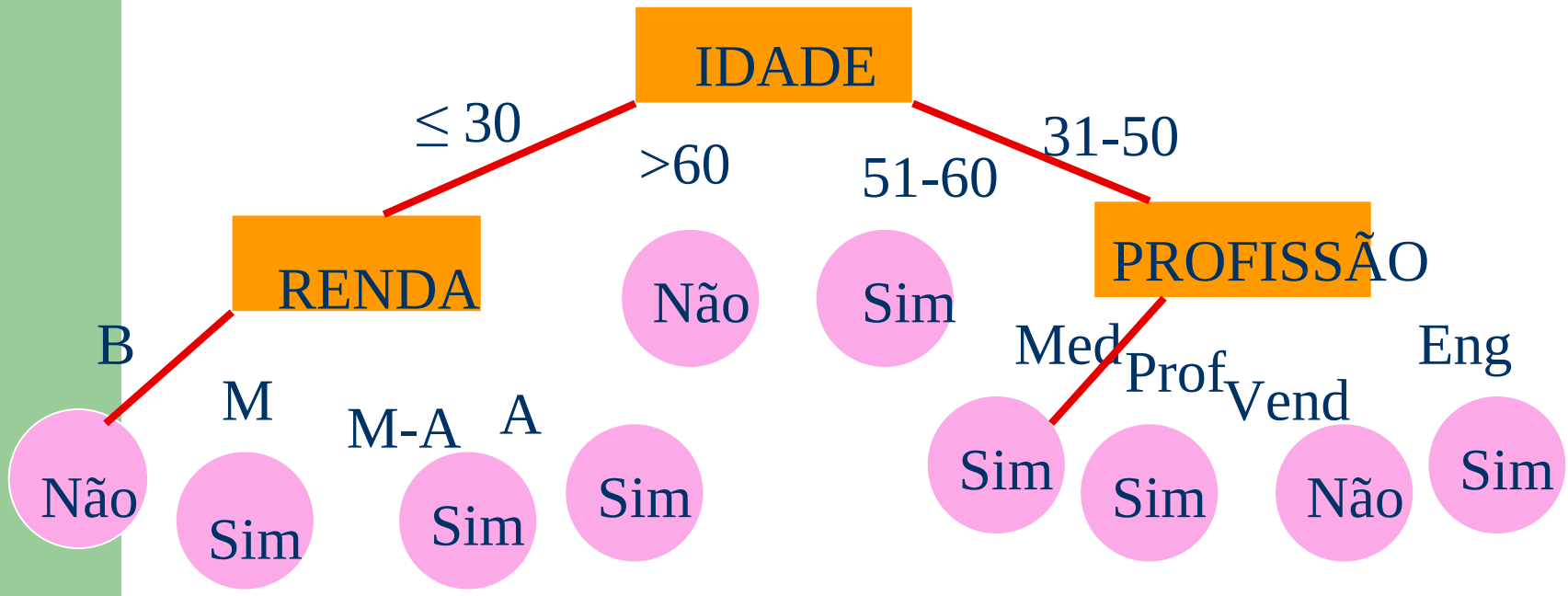
Classificação

Nome	Idade	Renda	Profissão	Classe
Daniel	≤ 30	Média	Estudante	Sim
João	31..50	Média-Alta	Professor	Sim
Carlos	31..50	Média-Alta	Engenheiro	Sim
Maria	31..50	Baixa	Vendedora	Não
Paulo	≤ 30	Baixa	Porteiro	Não
Otavio	> 60	Média-Alta	Aposentado	Não

Classificação



Classificação
Árvore de Decisão



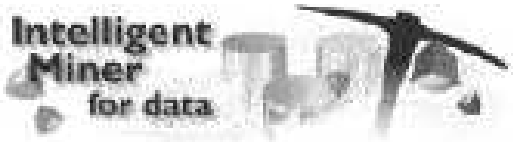
Se Idade ≤ 30 e Renda é Baixa então **Não compra Eletrônico**

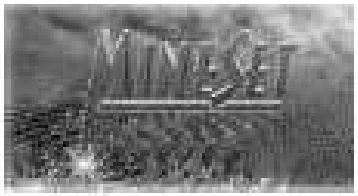
Se Idade = 31-50 e Prof é Médico então **compra Eletrônico**

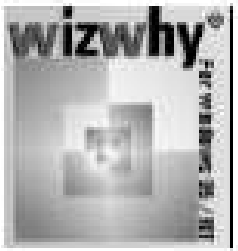
Técnicas de Mineração de Dados

Técnica	Tarefas	Exemplos
Descoberta de Regras de Associação	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen *et al.*., 1996).
Árvores de Decisão	Classificação Regressão	CART, CHAID, C5.0, Quest (Two Crows, 1999); ID-3 (Chen *et al.*., 1996); SLIQ (Metha *et al.*., 1996); SPRINT (Shafer *et al.*., 1996).
Raciocínio Baseado em Casos ou MBR	Classificação Segmentação	BIRCH (Zhang *et al.*., 1996); CLARANS (Chen *et al.*., 1996); CLIQUE (Agrawal *et al.*., 1998).
Algoritmos Genéticos	Classificação Segmentação	Algoritmo Genético Simples (Goldberg, 1989); Genitor, CHC (Whitley, 1993); Algoritmo de Hillis (Hillis, 1997); GA-Nuggets (Freitas, 1999); GA-PVMINER (Araújo *et al.*., 1999).
Redes Neurais Artificiais	Classificação Segmentação	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB (Azevedo, 2000), (Braga *et al.*, 2000), (Haykin, 2001)

Data Mining Products

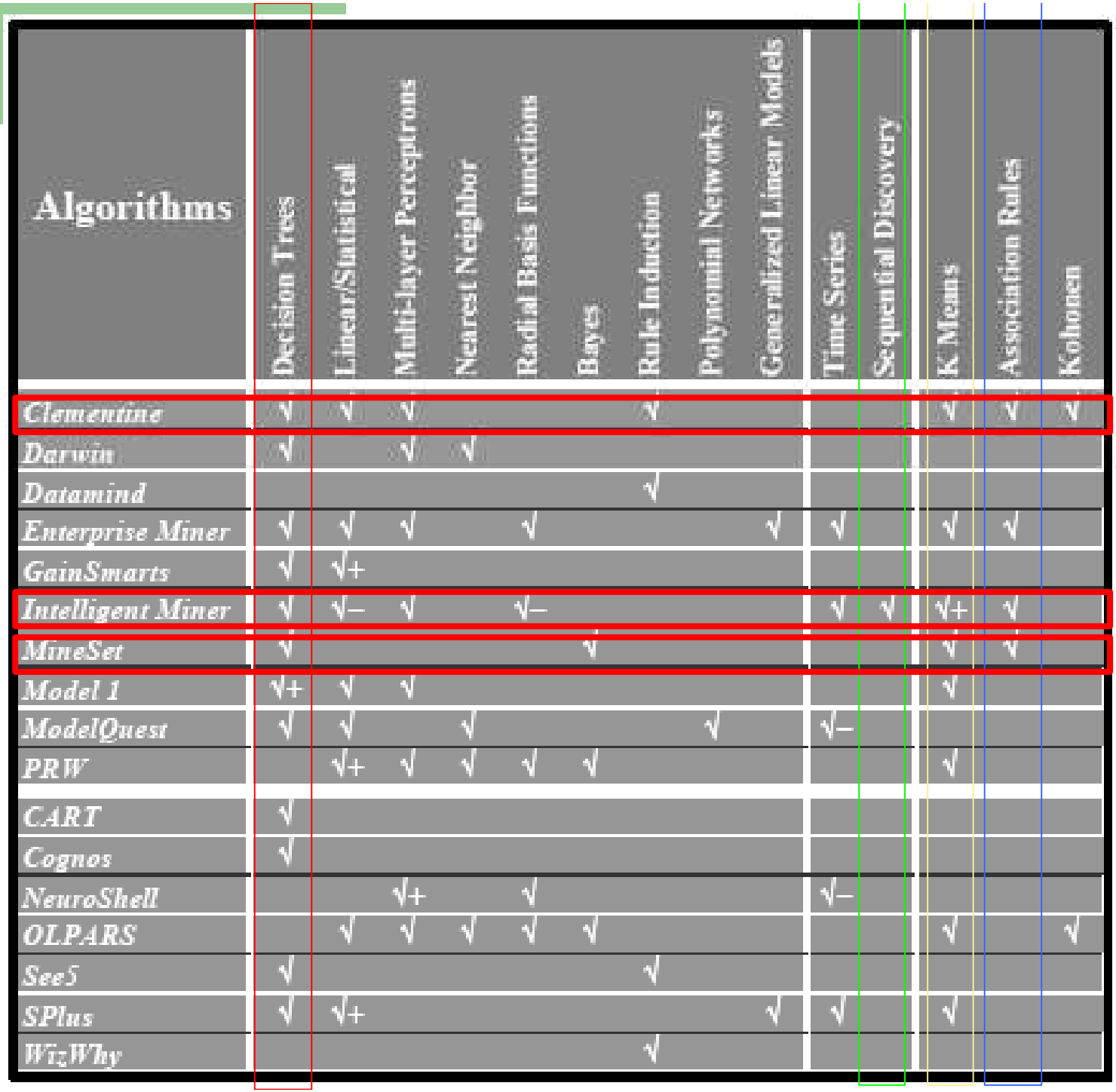






Exemplos

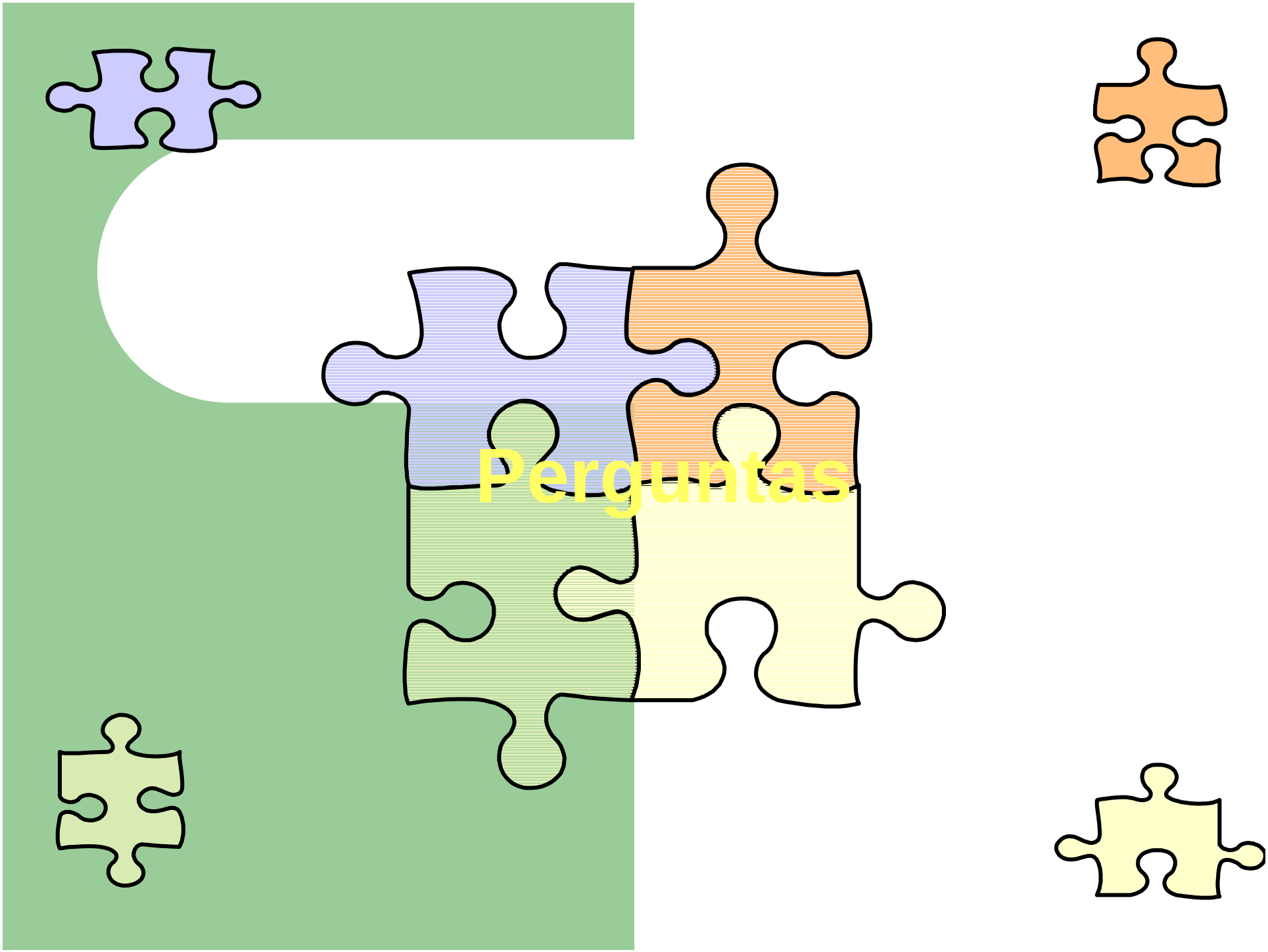
- Empresas de software para Data mining:
  - SAS http://www.sas.com
  - Information Havesting http://www.convex.com
  - Red Brick http://www.redbrick.com
  - Oracle http://www.oracle.com
  - Sybase http://www.sybase.com
  - Informix http://www.informix.com
  - IBM http://www.ibm.com



Algorithms	Decision Trees	Linear/Statistical	Multi-layer Perceptrons	Nearest Neighbor	Radial Basis Functions	Bayes	Rule Induction	Polynomial Networks	Generalized Linear Models	Time Series	Sequential Discovery	K Means	Association Rules	Kohonen
Clementine	√	√	√				√					√	√	√
Darwin	√		√	√										
Datamind							√							
Enterprise Miner	√	√	√		√				√	√		√	√	
GainSmarts	√	√+												
Intelligent Miner	√	√–	√		√–					√	√	√+	√	
MineSet	√					√						√	√	
Model 1	√+	√	√									√		
ModelQuest	√	√		√				√		√–				
PRW		√+	√	√	√	√						√		
CART	√													
Cognos	√													
NeuroShell			√+		√					√–				
OLPARS		√	√	√	√	√						√		√
See5	√						√							
SPlus	√	√+							√	√		√		
WizWhy							√							

Conclusões

- Data mining é um processo que permite compreender o comportamento dos dados.
- Data mining analisa os dados usando técnicas de aprendizagem para encontrar padrões e regulariedades nestes conjuntos de dados.
- É um problema pluridisciplinar, envolve Inteligência Artificial, Estatística, Computação Gráfica, Banco de Dados.
- Pode ser bem aplicado em diversas áreas de negócios


Perguntas

Referências Bibliográficas

- Técnicas de Mineração de Dados -JAI - SBC2004
  - http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf (Acesso 02/06/2005)
  - http://www.deamo.prof.ufu.br/arquivos/JAI-slides.ppt (Acesso 02/06/2005)
- Gimenes, Eduardo. “Data Mining – Data Warehouse” – Importância da Mineração de Dados em tomadas de decisão. Taquaritinga, 2000. Monografia sobre Mineração de Dados
  - http://geocities.yahoo.com.br/dugimenes/arquivos/data_mining.zip (Acesso 8/07/2005)
- Neto, Manoel Gomes de Mendonça. “Mineração de Dados”.
  - http://www.nuperc.unifacs.br/publicacoes.htm(Acesso 10/07/2005)
- Parâmetros na escolha de técnicas e ferramentas de mineração de dados
  - http://www.ppg.uem.br/Docs/ctf/Tecnologia/2002/18_279_02_Maria%20Dias_Parametros%20na%20escolha.pdf (Acesso 9/7/2005)

Referências Bibliográficas

- A Comparison of Leading Data Mining Tools (PDF format). A presentation by John F. Elder IV and Dean W.
  - http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf (Acesso 9/7/2005)
- Oliveira, Aracele G.; Garcia, Denise F. Mineração da Base de Dados de um Processo Seletivo Universitário. p.38-43.
  - http://www.dcc.ufla.br/infocomp/artigos/v3.2/art07.pdf (Acesso 31/05/2005)

Referências

- Fayyad et al. (1996). Advances in knowledge discovery and data mining, AAAI Press/MIT Press.
- Holsheimer, M. & Siebes, A.P.J.M. Data Mining: The Search for Knowledge in Databases, 1994.
- http://www-pcc.qub.ac.uk/tec/courses/datamining
- http://www.rio.com.br/~extended
- http://www.datamining.com
- http://www.santafe.edu/~kurt
- http://www.datamation.com
- http://www-dse.doc.ic.ac.uk/~kd
- http://www.cs.bham.ac.uk/~anp
- http://www.dbms.com
- http://www.infolink.com.br/~mpolito/mining/mining.html
- http://www.lci.ufrj.br/~labbd/semins/grupo1