Inteligência Artificial

Descoberta de Conhecimento em Bases de Dados

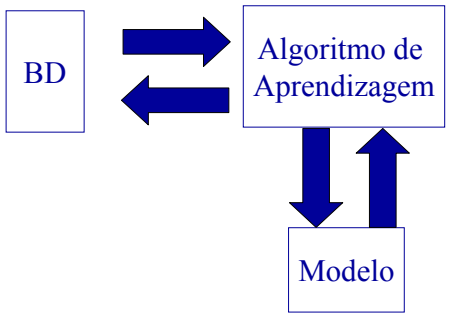
Prof. Paulo Martins Engel

Descoberta de Conhecimento em Bases de Dados

- DCBD (Descoberta de Conhecimento em Bases de Dados) ou KDD (Knowledge Discovery in Databases) é o *processo* de extração de conhecimento novo, útil e interessante a partir de bases de dados.
- A etapa mais importante deste processo, do ponto de vista tecnológico, é a *Mineração de Dados*, na qual um *Algoritmo de Aprendizagem* interage com a BD extraindo um *modelo* para ser utilizado numa determinada *tarefa* do processo DCBD.

Algoritmos de Aprendizagem

- Um Algoritmo de Aprendizagem (AA) é capaz de criar um *modelo* específico para os dados de entrada.
- Cada tipo de AA cria modelos para *tarefas* diferentes, por exemplo, para *prever* a classe de instâncias (*classificação*), ou prever atributos que ocorrem juntos (*associação*), ou ainda descobrir *perfis* de comportamento (*agrupamento*).



Exemplo 1 – Análise de Risco de uma proposta de empréstimo

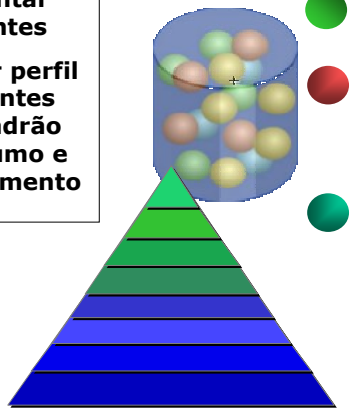
- A partir de dados históricos de clientes que obtiveram empréstimos e como os seus pagamentos ocorreram, criar um modelo de classificação bom/mau pagador, para determinar se deve ou não conceder crédito a novo cliente.

- **Tarefa: classificação**

Exemplo 2 - Loja virtual quer identificar seu cliente

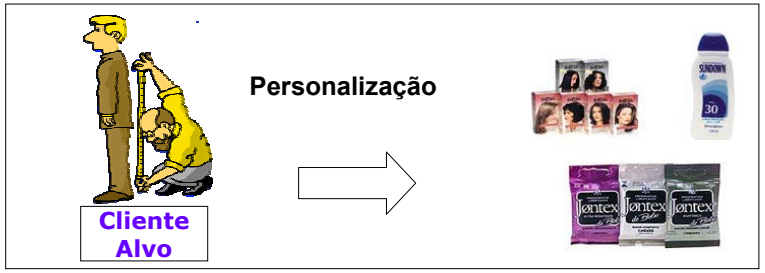
Segmentar os clientes

Descobrir perfil dos clientes e seu padrão de consumo e comportamento



- **perfil dos clientes lucrativos**
- **quais produtos são consumidos por cada perfil de cliente**
- **perfil de consumo por região**
  - **Tarefa: agrupamento**
- **quais produtos são consumidos de forma associada**
  - **Tarefa: associação**

Exemplo 3 - Marketing de precisão em lojas virtuais



A oferta personalizada de produtos e serviços:

- **Aumenta conversão de navegadores em compradores**
- **Aumenta nº itens por transação (cross-sales)**
- **Aumenta valor dos itens (up-sales)**

Introdução e Motivação

DESCOBERTA DE CONHECIMENTO:

➔Necessidade de ferramentas mais robustas para a indução de conhecimento.

➔Recuperação e análise das informações ocultas nas bases de dados, que serão utilizadas no processo de tomada de decisão.

➔Envolve várias etapas complexas, entre elas a etapa de Mineração de Dados.

Descoberta de Conhecimento em Bases de Dados

CONCEITO:

"Processo não trivial de identificar *padrões* válidos, não conhecidos, potencialmente úteis e interpretáveis" [Fayyad, 96].

Padrão no contexto de DCBD

CONCEITO:

Um padrão é uma descrição de um subconjunto de dados que têm características comuns.

Mineração de Dados

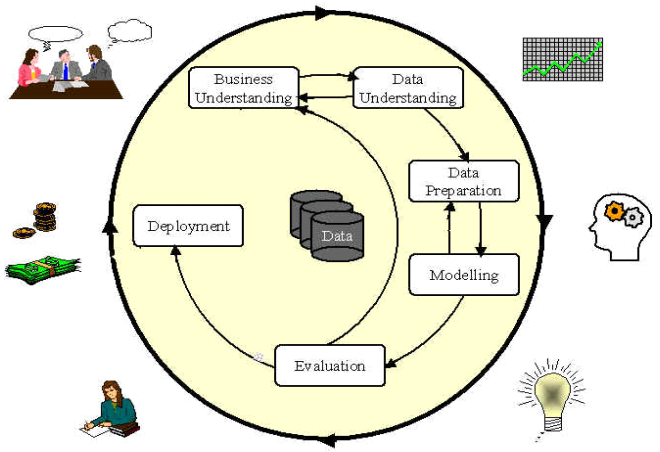
- Extração de informação implícita, previamente desconhecida e potencialmente útil
- Necessidades: programas que detectam padrões e regularidades nos dados
- Padrões fortes podem ser usados para fazer predições
  - Problema 1: a maioria dos padrões não são interessantes
  - Problema 2: padrões podem ser imprecisos (ou mesmo completamente espúrios) se houver dados deturpados ou faltantes

Aprendizado de Máquina

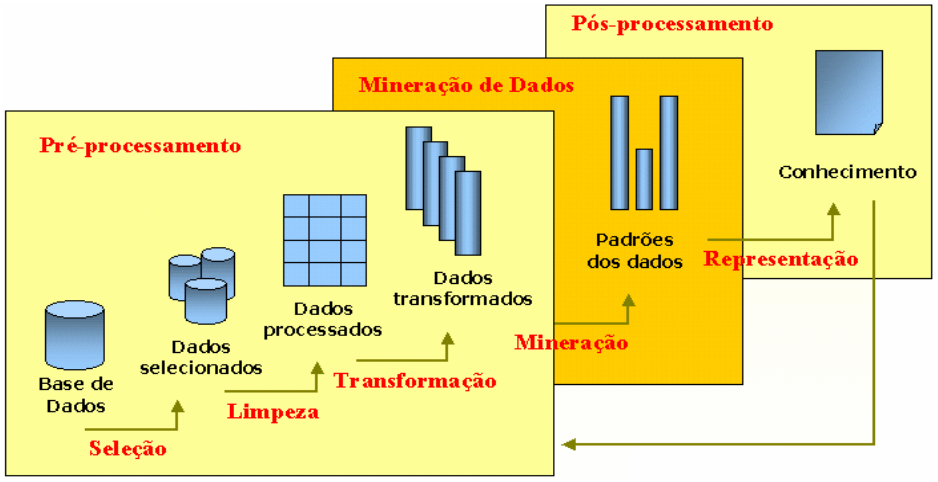
- Base técnica para mineração de dados: algoritmos para adquirir descrições estruturais a partir de exemplos
- Utiliza métodos de *raciocínio indutivo* para descrever relações lógicas encontradas num subconjunto de dados.
- Utiliza *exemplos* para construir um modelo.
- O modelo é representado simbolicamente.
- Exemplos de representações simbólicas:
  - Regras Associativas
  - Regras de Classificação
  - Árvores de Decisão

O modelo CRISP-DM

"*CRoss-Industry Standard Process for Data Mining*"



Etapas do Processo de DCBD

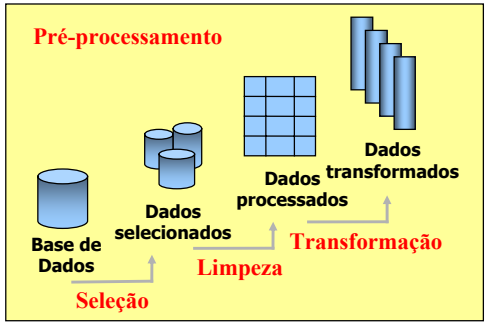


Fonte: Fayyad

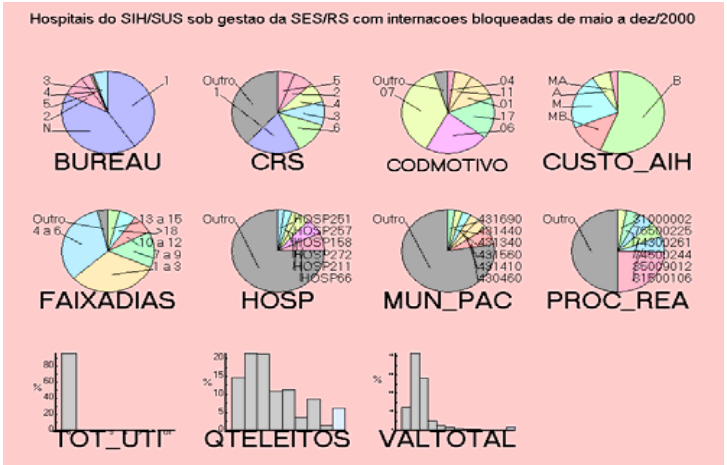
DCBD

Pré-processamento:

- ➔ Levantamento do domínio;
- ➔ Seleção;
- ➔ Limpeza;
- ➔ Transformação.



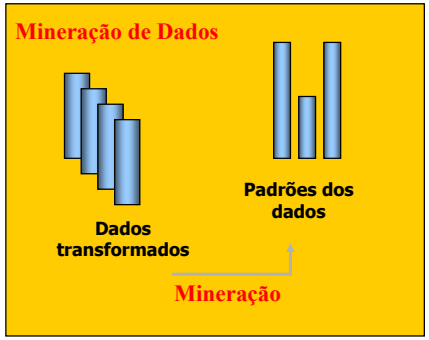
Exemplo de Estatísticas para Exploração dos Dados



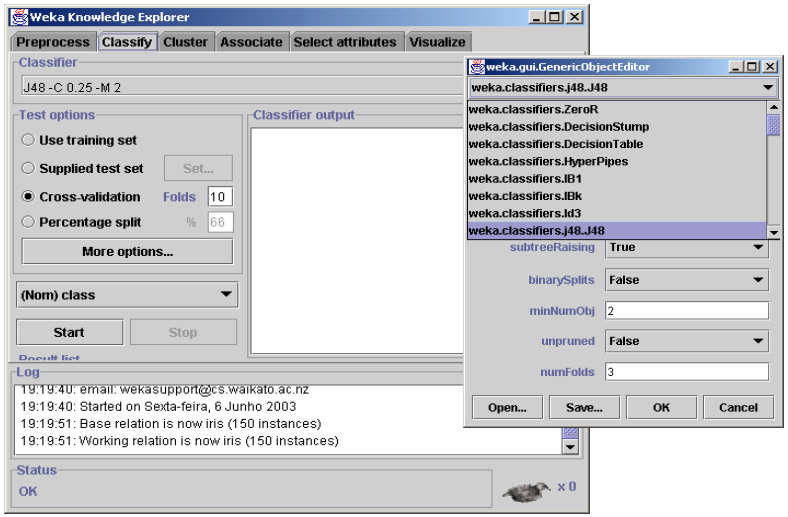
DCBD

Modelagem - Mineração de Dados:

- ➔Escolha da tarefa;
- ➔ Escolha da técnica;
- ➔ Aplicação do algoritmo.



Escolha da Tarefa e do Algoritmo

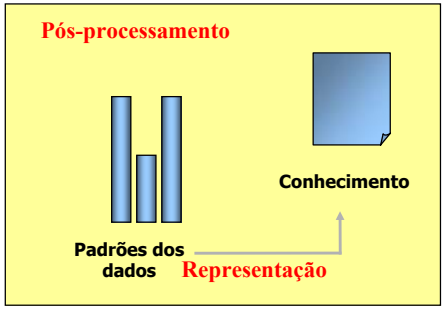


DCBD

Pós-processamento:

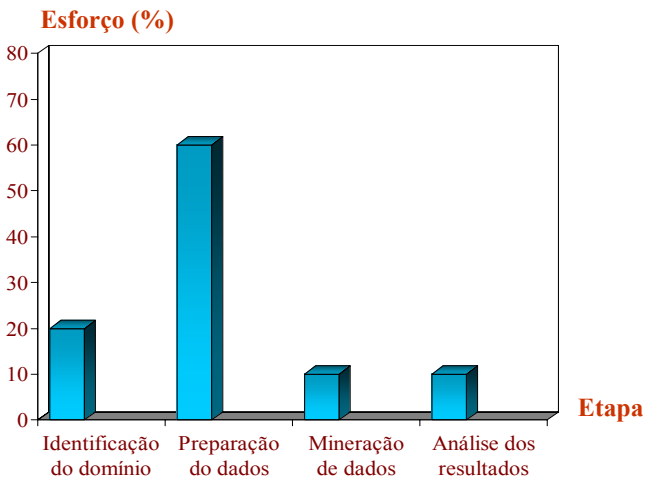
➔ Interpretação dos padrões;

➔ Consolidação da descoberta.



DCBD

DESAFIOS:



Fonte: Adriaans

Representação de Padrões

➔ Os padrões podem ser representados numa linguagem simbólica.

➔ Lógica de predicados, regras de produção, árvores de decisão, regras associativas, etc.

➔ Pode-se representar padrões também através de um (*elemento*) protótipo (eventualmente hipotético).

➔ Os padrões podem ser representados por modelos matemáticos (não simbólicos).

➔ Redes neurais, modelos estatísticos, etc.

Escolha da Linguagem de Representação de Padrões

- A escolha da linguagem de representação de padrões é um passo muito importante do processo de DCBD pois ela determina um *viés* para a descrição do conhecimento.
- Em geral, a representação *simbólica* enfatiza a compreensão (qualitativa) dos relacionamentos.
- A representação *sub-simbólica* normalmente foca na precisão do reconhecimento dos padrões.

Mineração de Dados

TÉCNICAS PRINCIPAIS:

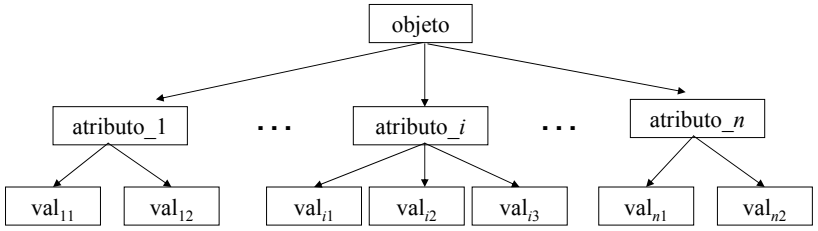
- Regras Associativas
- Árvores de Decisão
- Regras de Produção
- Redes Neurais
- Descoberta de Agrupamentos

FERRAMENTA: WEKA

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

Representação do Domínio

- Do ponto de vista do processo de Descoberta de Conhecimento, o domínio é representado por trincas do tipo: (objeto, atributo, valor).
- O Banco de Dados alvo do processo (e o seu modelo) fornece o conhecimento a priori do domínio.



Preparação para a aprendizagem

- Conceitos: tipos de noções que podem ser aprendidas
  - Objetivo: descrição inteligível e operacional de um conceito
- Amostras: os exemplos individuais e independentes de um conceito
- Atributos: medem aspectos de uma amostra
  - Podem ser de vários tipos, p. ex.: atributos nominais e numéricos

O que é um conceito?

- Conceito: algo a ser aprendido; um padrão que descreve um subconjunto dos dados e que depende do estilo de aprendizado (tarefa).
- Estilos de aprendizado:
  - Aprendizado classificatório: prever uma classe discreta
  - Aprendizado associativo: detectar associações entre características
  - Aprendizado aglomerativo: agrupar amostras similares
  - Previsão numérica: prever uma quantidade numérica
- Descrição de conceito: saída do esquema de aprendizado

O que é um exemplo?

- Amostra: tipo específico de exemplo
  - Objeto a ser classificado, associado ou agrupado
  - Exemplo individual e independente do conceito alvo
  - Caracterizado por um conjunto predeterminado de atributos
- Entrada para o esquema de aprendizagem: conjunto de amostras/ dados
  - Representado como uma única relação (*arquivo plano*)
- É uma forma restrita de entrada
  - Não pode haver relacionamentos entre objetos
- É a forma mais comum na prática de MD

O que é um atributo?

- Cada amostra é descrita por um conjunto pré-definido de características, os seus "atributos"
- Mas: na prática, número de atributos pode variar
  - Solução possível: *flag* "valor irrelevante" (p. ex. "?")
- Problema relacionado: existência de um atributo pode depender de valor de um outro atributo
- Tipos possíveis de atributos ("níveis de medidas"):
  - *Nominal*, *ordinal*, *intervalar* e *racional*

Dados de um problema (classificação) com incertezas: o problema do tempo

Atributos previsores | Atributo meta (a ser previsto)

Tempo	Temperatura	Umidade	Ventoso	Joga
ensolarado	quente	alta	falso	não
ensolarado	quente	alta	verdadeiro	não
nublado	quente	alta	falso	sim
chuvoso	amena	alta	falso	sim
chuvoso	fria	normal	falso	sim
chuvoso	fria	normal	verdadeiro	não
nublado	fria	normal	verdadeiro	sim
ensolarado	amena	alta	falso	não
ensolarado	fria	normal	falso	sim
chuvoso	amena	normal	falso	sim
ensolarado	amena	normal	verdadeiro	sim
nublado	amena	alta	verdadeiro	sim
nublado	quente	normal	falso	sim
chuvoso	amena	alta	verdadeiro	não

- Lista de dias, apresentando as condições climáticas e se o jogador foi jogar ou não.
- Arquivo lista *apenas* as combinações dos valores dos atributos que *realmente* apareceram no domínio.
- As combinações podem não ser exaustivas e podem ser contraditórias.
- Tem apenas 14 das 36 combinações possíveis (3×3×2×2).
- Situação muito comum.
- O domínio é não determinístico.
- Para um certo conjunto de valores de atributos, existe uma *probabilidade* de ocorrer o valor previsto.

Quantidades nominais

- Valores são símbolos distintos
  - Os valores servem apenas como rótulos ou nomes
- Exemplo: atributo "tempo" dos dados do tempo
  - Valores: "ensolarado", "nublado", "chuvoso"
- Não existe relação implícita entre valores nominais (ordenação ou distância)
- Pode-se realizar apenas testes de igualdade

Quantidades racionais

- Quantidades racionais são aquelas para as quais o esquema de medida define um ponto zero
- Exemplo: atributo "distância"
  - Distância entre um objeto e ele mesmo é zero
- Quantidades racionais são tratadas como números reais (valores numéricos)
  - Todas as operações matemáticas são permitidas

Dados do tempo com atributos nominais e numéricos

Tempo	Temperatura	Umidade	Ventoso	Joga
ensolarado	29	85	falso	não
ensolarado	26	90	verdadeiro	não
nublado	28	86	falso	sim
chuvoso	21	96	falso	sim
chuvoso	20	80	falso	sim
chuvoso	18	70	verdadeiro	não
nublado	17	65	verdadeiro	sim
ensolarado	23	95	falso	não
ensolarado	21	70	falso	sim
chuvoso	24	80	falso	sim
ensolarado	24	70	verdadeiro	sim
nublado	23	90	verdadeiro	sim
nublado	27	75	falso	sim
chuvoso	22	91	verdadeiro	não

Geração de um arquivo plano

- O processo de tornar um arquivo plano é chamado de "desnormalização"
  - Juntam-se várias relações numa única relação
- É possível de se fazer com qualquer conjunto finito de relações finitas
- Desnormalização pode produzir regularidades espúrias que refletem estrutura da base de dados
  - Ex.: "fornecedor" prevê "endereço do fornecedor"