Resumo IA - AB2

Mineração de Dados (Data Mining)

Introdução:

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.

Tipos de mineração de dados:

- Preditiva: deseja-se prever o valor desconhecido de um determinado atributo, a partir da análise histórica dos dados armazenados na base.
- Descritiva: padrões e regras descrevem caractéristicas importantes dos dados com os quais se está trabalhando.

Mineração de Dados:

Etapa principal do processo de KDD(Knowledge Discovery in Databases)(Descoberta de conhecimento em Bases de Dados), na qual é realizada a busca por novas informações e conhecimento.

O processo de **KDD** é composto por seis fases(Navathe):

1. Seleção dos dados;
2. Limpeza dos dados;
3. Enriquecimento dos dados;
4. Transformação dos dados;
5. **Mineração dos dados**;
6. Apresentação e análise dos resultados.

Fases:

1. Seleção (Selection): esta etapa consiste em selecionar um conjunto ou subconjunto de dados que farão parte da análise. As fontes de dados podem ser variadas (planilhas, sistemas gerenciais, data warehouses) e possuir dados com formatos diferentes (estruturados, semiestruturados e não-estruturados).
2. Processamento (Preprocessing): esta etapa consiste em fazer a verificação da qualidade dos dados armazenados. A base passa por um processo de limpar, corrigir ou remover dados inconsistentes, verificar dados ausentes ou incompletos, identificar anomalias (outliers).
3. Transformação (Transformation): esta etapa consiste em aplicar técnicas de transformação como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Aqui os dados ficam disponíveis agrupados em um mesmo local para a aplicação dos modelos de análise.
4. Mineração de Dados (Data Mining): esta etapa consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo (1) verificar uma hipótese, (2) descobrir novos padrões de forma autônoma. Além disso, a descoberta pode ser dividida em: preditiva e descritiva.

Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.

5. Interpretação e Avaliação (Interpretation / Evaluation): esta etapa consiste em avaliar o desempenho do modelo, aplicando em cima de dados que não foram utilizados na fase de treinamento ou mineração. A validação pode ser feita de diversas formas, algumas delas são: utilizar medidas estatísticas, passar pela avaliação dos profissionais de negócio.

Tarefas em Mineração de Dados:

- Regras de associação;
- 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 freqüência na base.

Exemplos:

- {fralda} => {cerveja}
  - parte significativa das compras de homens, às sextas à noite, que inclui fraldas, inclui também cerveja.
- {pão, manteiga} => {leite}
  - o cliente que compra pão e manteiga, 80% das vezes compra leite.
- {candidíase} => {pneumonia}
  - muitos pacientes aidéticos que contraem a doença candidíase também têm pneumonia

As **regras de associação** são extraidas da base de dados que contêm transações - formadas por conjuntos de itens do domínio da aplicação.

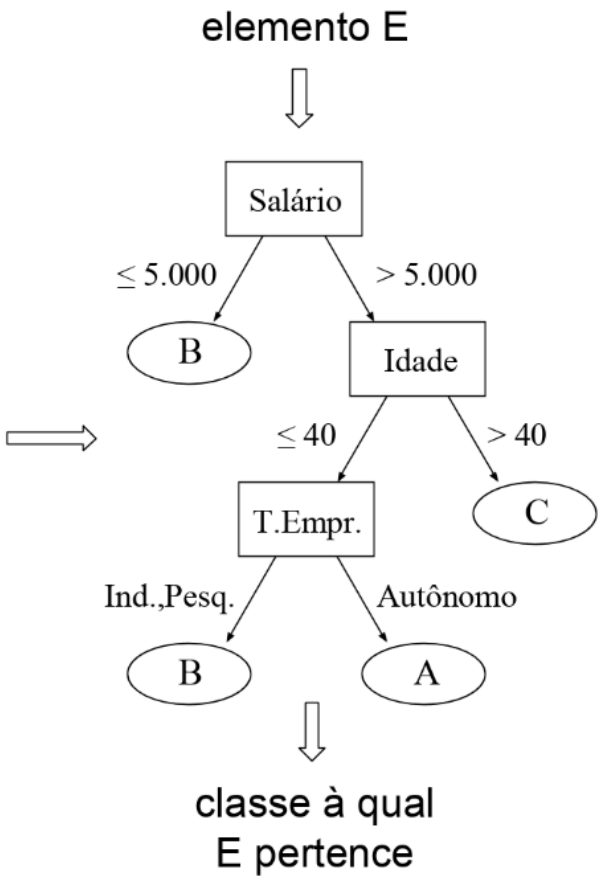
Padrões de sequências representam sequências de conjuntos de itens que ocorrem nas transações de diferentes consumidores, com determinada frequência na ordem específica.

Classificação:

- 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 precisa de uma base de treinamento.

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



Clusterização (Agrupamento)

- É o resultado da identificação de um conjunto finito de categorias (ou grupos - clusters) que contêm objetos similares.
    - Grupos 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.

Consumidor	Qtd.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

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

Consumidor	Qtd.Tot.	Preço.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.Tot.	Preço.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 total e ao preço médio dos produtos consumidos.

Técnicas de mineração de dados:

Tarefa	**Técnicas**
Classificação	Árvores de Decisão / K-NN / Classificador Bayesiano
Associação	Algoritmos de Extração de Regras de Associação
Clusterização	Algoritmos de Particionamento / Algoritmos Hierárquicos

Aplicações das Técnicas de MD:

- Marketing:
  - Análise do comportamento dos clientes baseada no padrão de compras.
- Finanças:
  - Análise do risco na concessão de empréstimos.
- Saúde:

    - Previsão dos resultados de determinados tratamentos.
- Educação:
    - Avaliação da evasão escolar e do desempenho de alunos.
- Segurança:
    - Identificação de roubo de cartão de crédito, detecção de SPAM.