Mineração de Dados e o Processo de KDD (Knowledge Discovery in Databases)

Daricélio Moreira Soares

Introdução à Mineração de Dados

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.

Este processo é executado sobre grandes quantidade de dados, estejam estes armazenados em bancos de dados tradicionais, em data warehouses ou em outra forma de repositório.

Introdução à Mineração de Dados

⇒ O acúmulo de grandes quantidades de dados históricos, nos anos 70 e 80, em empresas e centros de pesquisa, motivou, a partir do início dos anos 90, o desenvolvimento deste tipo de análise de dados que viabiliza a extração de informações novas, inesperadas e úteis.

⇒ Base para construção dos

Sistemas de Apoio à Decisão.

Mineração de Dados: etapa principal do processo de KDD (Knowledge Discovery in Databases), na qual é realizada a busca por novas informações e conhecimento.

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

- Seleção dos dados,
- Limpeza dos dados,
- Enriquecimento dos dados,
- Transformação dos dados,
- Mineração dos dados**,**
- Apresentação e análise dos resultados.

KDD (Knowledge Discovery in Databases)

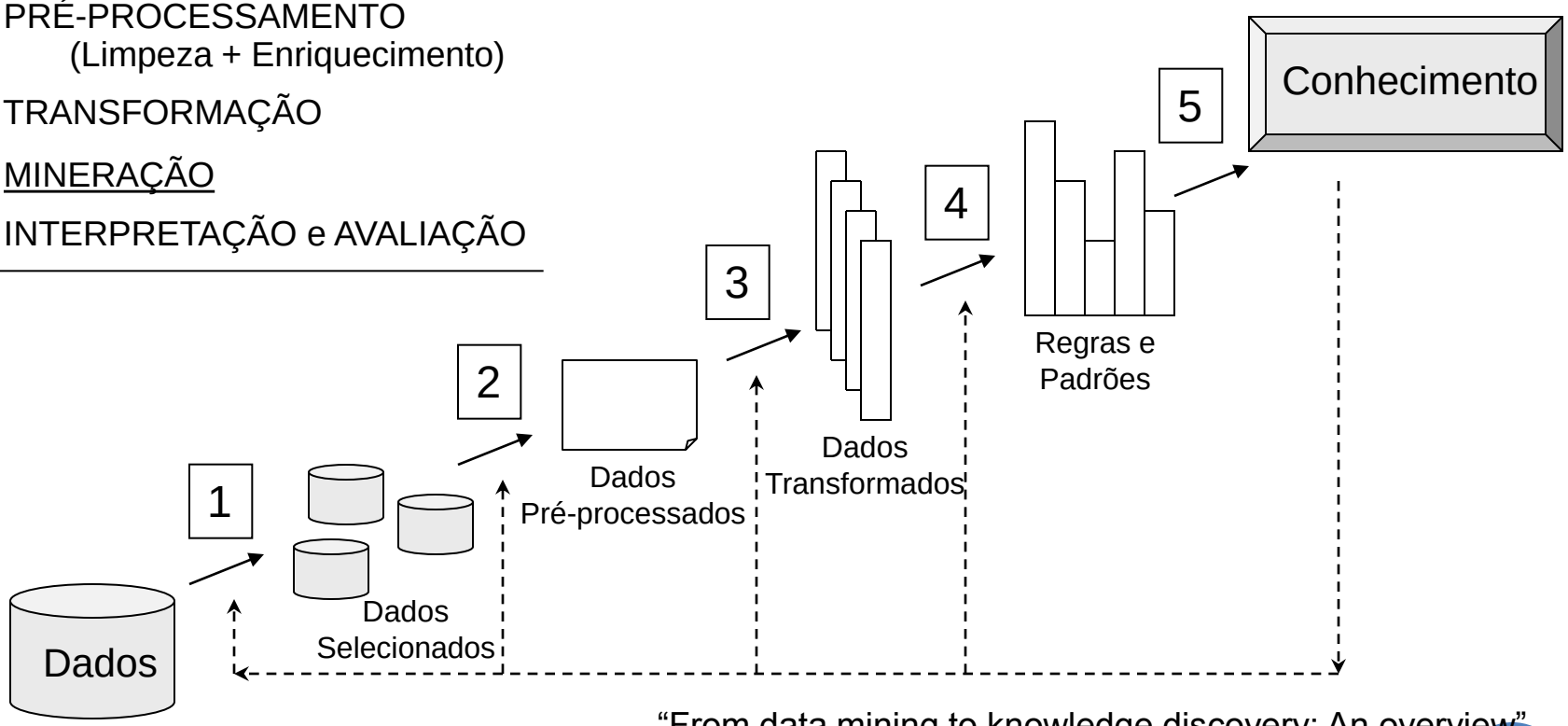
1 - SELEÇÃO

2 - PRÉ-PROCESSAMENTO
(Limpeza + Enriquecimento)

3 - TRANSFORMAÇÃO

4 - MINERAÇÃO

5 - INTERPRETAÇÃO e AVALIAÇÃO



"From data mining to knowledge discovery: An overview", U.M.Fayyad *et. al.*, 1996.

→O processo de mineração de dados ocorre, muitas vezes, a partir de *Data warehouses*.

→ De acordo com W.H.Inmon, “um DW é uma coleção de dados orientada a um assunto, integrada, temporal e não volátil, projetada para dar suporte a sistemas de apoio à decisão”.

Tarefas em Mineração de Dados

- Regras de Associação
- Padrões de Seqüências
- 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.

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

Regras de Associação

(*market basket analysis*)

Regras de associação são extraídas a partir de bases de dados que contêm transações - formadas por conjuntos de itens do domínio da aplicação.

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 médio)
{fralda} $\Rightarrow$ {leite}	confiança de 100%	(suporte alto)
{leite} $\Rightarrow$ {fralda}	confiança de 75%	(suporte alto)
{carne} $\Rightarrow$ {cerveja}	confiança de 100%	(suporte baixo)

Padrões de Seqüências

Padrões de seqüências representam seqüências de conjuntos de itens que ocorrem nas transações de diferentes consumidores, com determinada freqüência (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.

Classificação

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

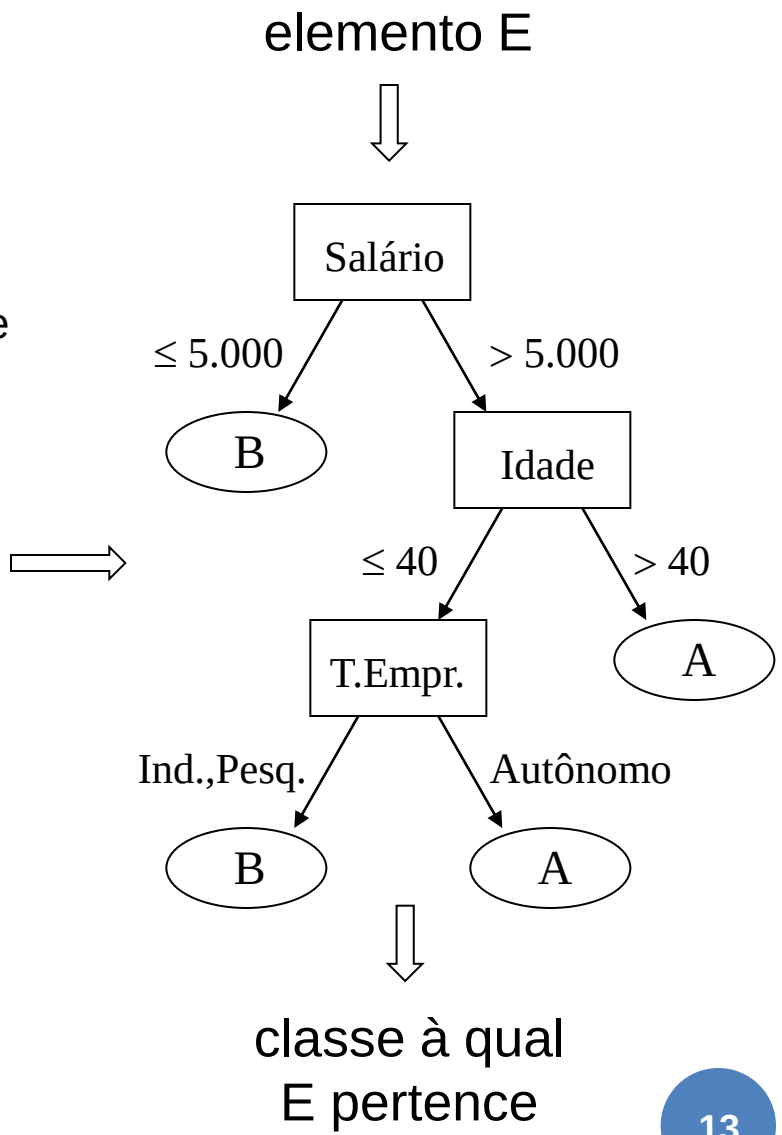
→ Implementar/minerar um classificador significa gerar/descobrir a função que realiza tal mapeamento.

→ 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	A
4	6.000	45	Autônomo	A
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	A
4	6.000	45	Autônomo	A
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



Agrupamento (*Clustering*)

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.

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 (*Clustering*)

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.

Técnicas de Mineração de Dados

Técnicas são utilizadas para realizar a tarefa de mineração de dados.

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

Mineração de Dados Direta

Através de uma técnica de mineração, extrai-se ou treina-se um modelo que será posteriormente utilizado.

Principais tarefas: classificação, regressão (previsão).

“Quem vai falir?”

Mineração caracterizada pela existências de um campo específico cujo valor deve ser estimado a partir dos valores dos demais atributos.

Mineração de Dados Indireta

Através de uma técnica de mineração, extraem-se padrões significativos que serão posteriormente avaliados.

Principais tarefas: análise de associações, clusters.

“Diga-me algo interessante?”

O resultado da mineração complementa o conhecimento do especialista e deverá ser examinado e avaliado por este.

Áreas de Aplicação das Técnicas de MD

Analisar tendências e padrões a partir de dados históricos, com o objetivo de prever ações futuras e apoiar decisões, pode ser um procedimento útil em diversas áreas:

- Marketing:
  → análise do comportamento dos clientes baseada no padrão de compras.
- Finanças:
  → análise do risco na concessão de empréstimos.
- Saúde:
  → detecção de tumores por análise imagem.
- Educação:
  → avaliação da evasão escolar e do desempenho de alunos.
- Segurança:
  → detecção de fraudes, detecção de SPAM.