

Mineração de Dados

CONCEITOS E PRINCIPAIS TÉCNICAS

Antonio Braz Finizola

Roteiro

- Introdução
- Mineração de Dados
- Descoberta de Conhecimento em Bases de Dados
- Principais tarefas
- Principais técnicas
- Como escolher a melhor técnica de MD
- Mineração de Dados Complexos
- Exemplo de Ferramenta de Mineração
- Referências

Introdução

- Surgimento dos Sistemas Computacionais
 - Um dos principais objetivos: Armazenamento de Dados
- Com o passar dos anos
 - Aquisição de Hardware: custo cada vez mais baixo
 - Estruturas de armazenamento robustas
 - Bancos de Dados
 - Data Warehouses
 - Bibliotecas Virtuais
 - Web

Introdução

- Volumes de dados gerados por algumas aplicações
 - NASA
 - Projeto Genoma
 - Grandes instituições bancárias



Motivação

- Diante disso, eis a questão:
 - O que fazer com a **-grande quantidade de dados gerados-**?
- Milhões gastos em armazenamento de dados
 - **Organizações rica em dados, pobre em informação**
- Necessidade de criação de novas tecnologias para análise automática e inteligente de bancos de dados [Fayyad et al. 1996].
 - Transformação de dados inúteis em conhecimentos úteis

Mineração de Dados

- Proposta na década de 80
- *“Mineração de dados é a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos.”*
(Berry e Linoff, 1997).
- Potencial
 - Análise matemática para derivar padrões e tendências;
 - Lidar com grandes volumes de dados complexos
- Aplicação de duas formas (Groth, 1998)
 - **Processo de Verificação**
 - **Processo de Descoberta**

Mineração de Dados

- Camilo (2009) apresenta algumas das áreas nas quais a Mineração de Dados é aplicada de forma satisfatória:
 - Retenção de clientes
 - Bancos
 - Medicina
 - Telemarketing
 - Eleitoral

Descoberta de Conhecimento em bases de dados

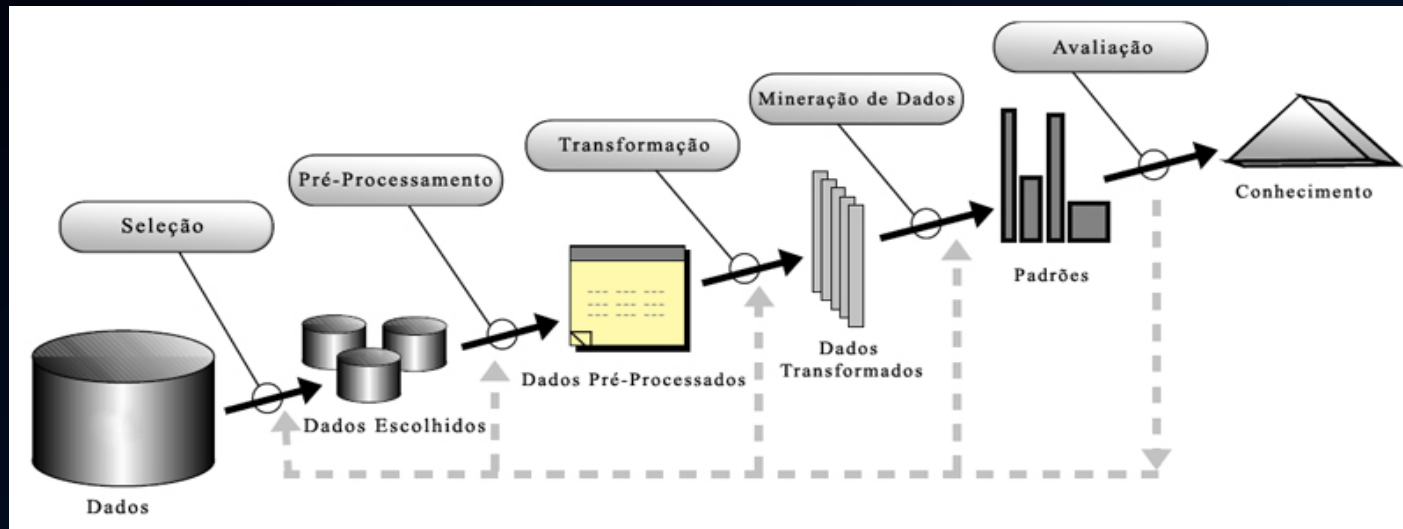
- Problema: Modelo tradicional para transformação dos dados em conhecimento (Fayyad, 1996)
 - Processamento manual
 - Produção relatórios para análise
 - Inviabilidade: volume de dados
- Solução: *KDD (Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados)
 - "era da informação": a sobrecarga de dados.

Descoberta de Conhecimento em bases de dados

- “Um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis”. [Fayyad, 1996]
- Área dinâmica e evolutiva
 - Estatística, Inteligência Artificial e Bancos de Dados
- Garantia
 - Confiabilidade, utilidade e compreensibilidade dos padrões

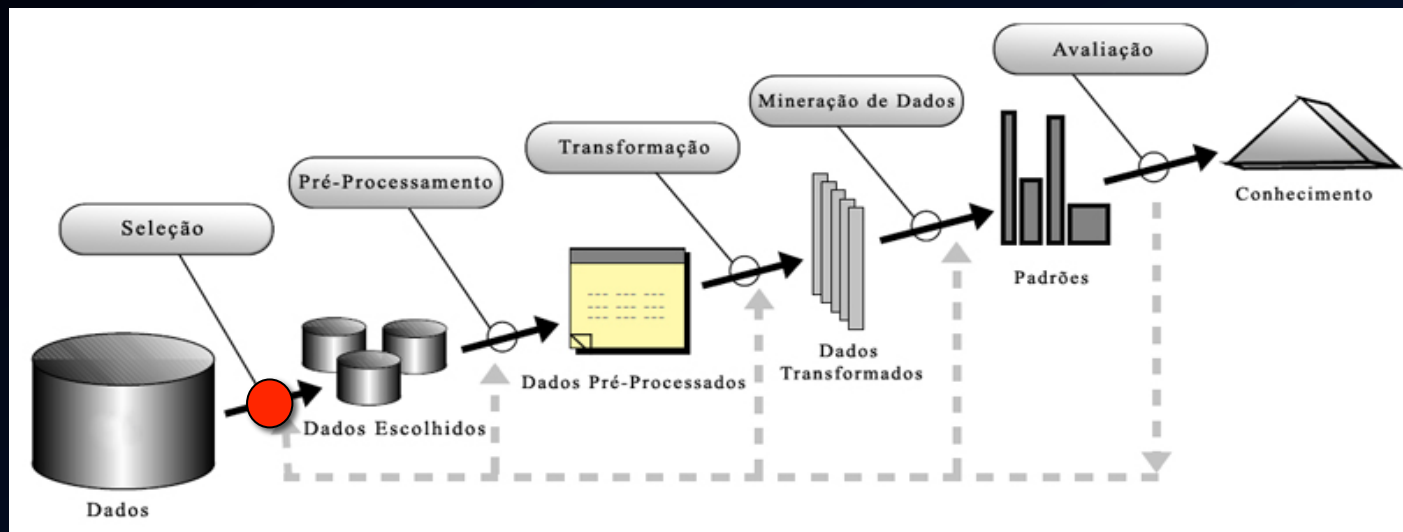
Descoberta de Conhecimento em bases de dados

As 5 Etapas do KDD



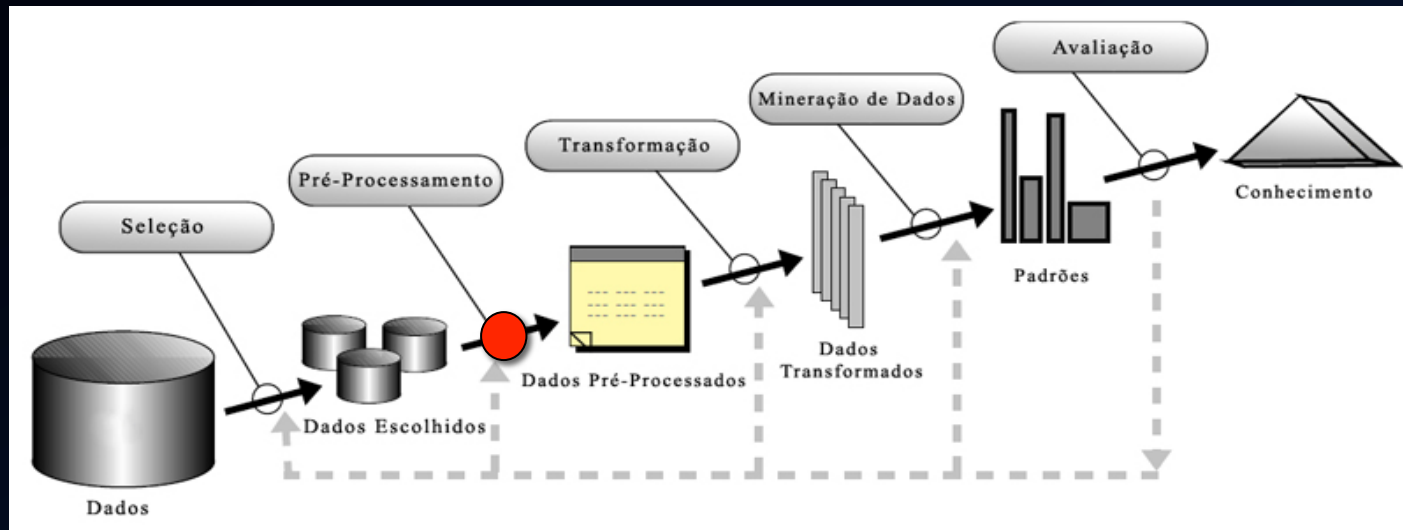
Descoberta de Conhecimento em bases de dados

- **Seleção:** escolher um conjunto de dados contendo todas as possíveis variáveis e registros que farão parte da análise.



Descoberta de Conhecimento em bases de dados

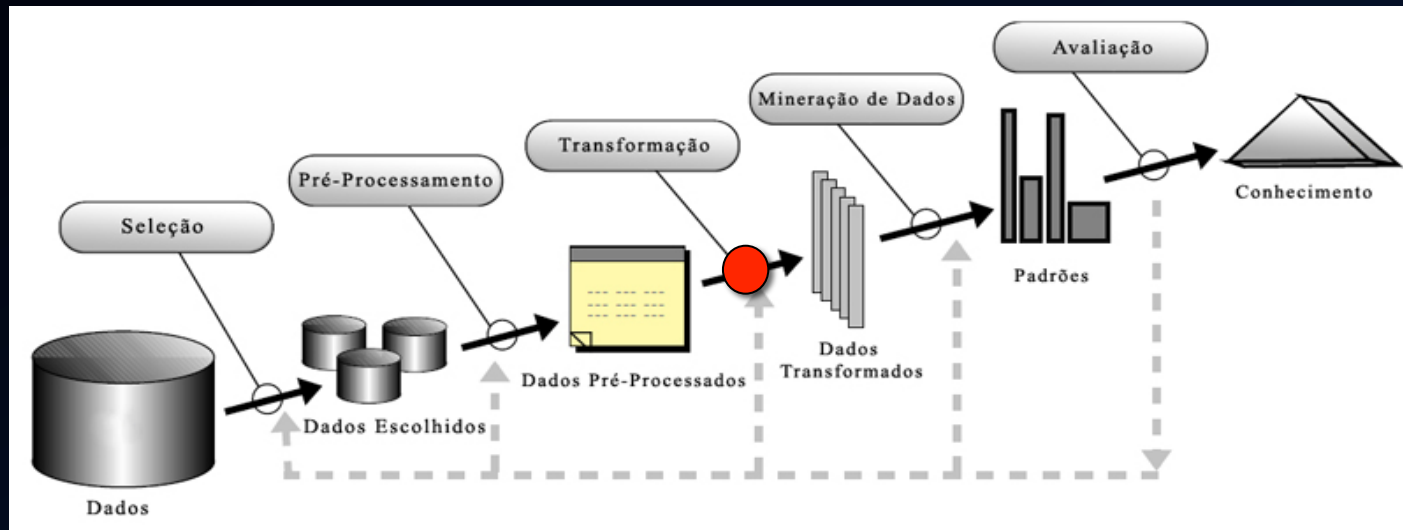
- **Pré-processamento:** eliminação de dados redundantes e inconsistentes



- Eventuais situações:
 - **Dados ausentes** e **Outliers**

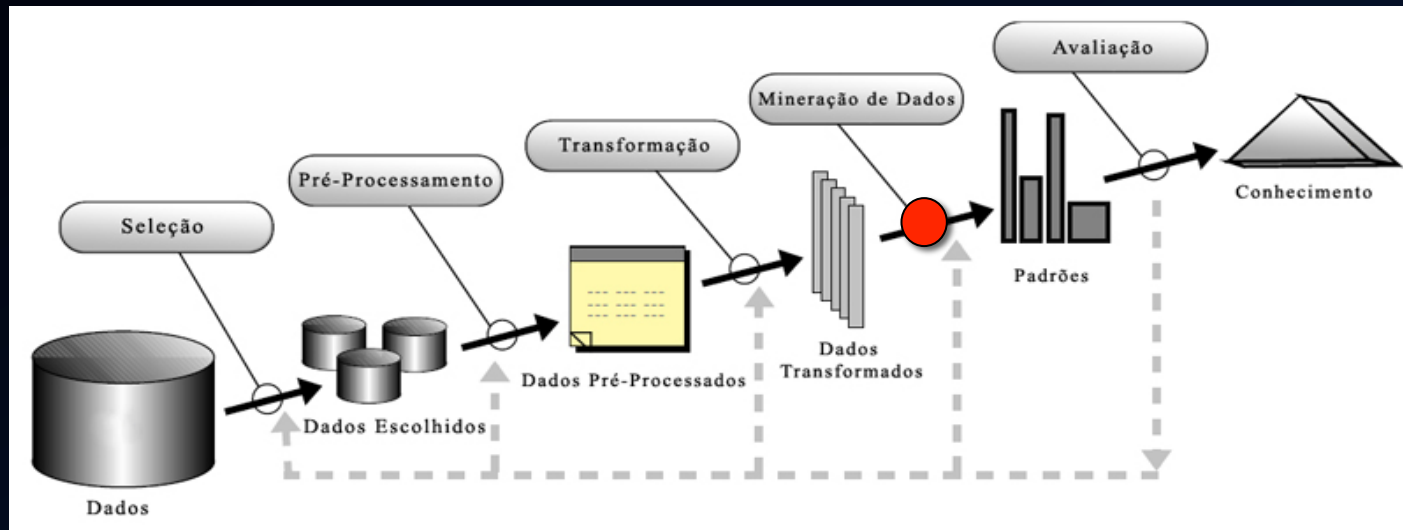
Descoberta de Conhecimento em bases de dados

Transformação: organização dos dados num formato apropriado para a aplicação de algoritmos de mineração.



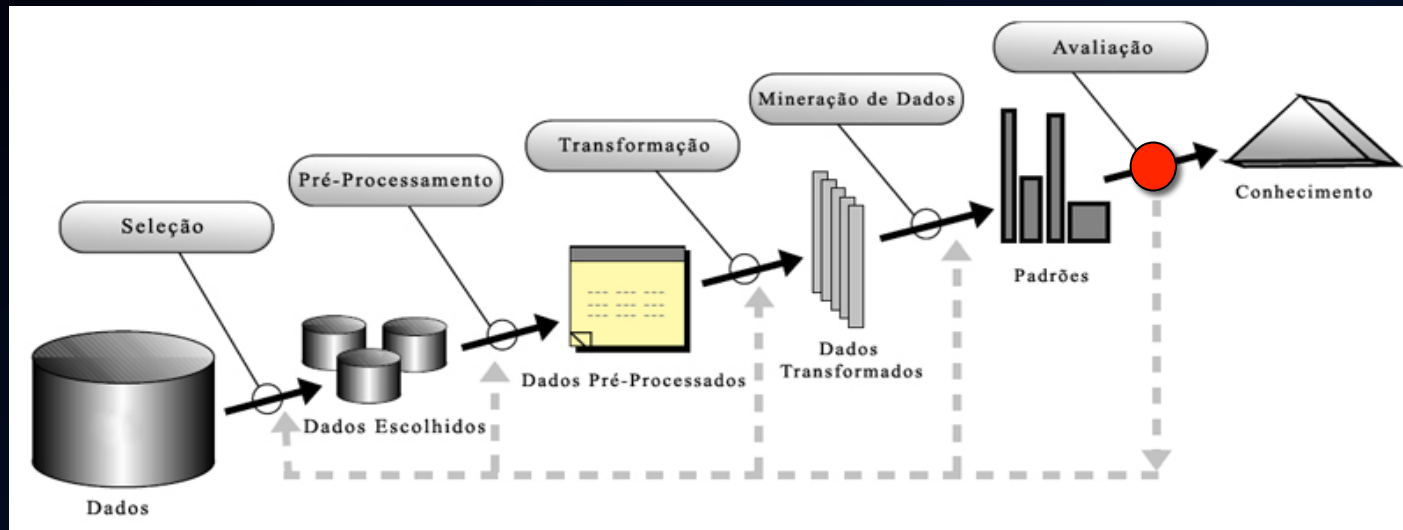
Descoberta de Conhecimento em bases de dados

Mineração: Aplicações de técnicas de reconhecimento e extração de padrões



Descoberta de Conhecimento em bases de dados

Avaliação: avaliação dos resultados pós-minerados



Principais tarefas

A Mineração de Dados é comumente classificada pela sua capacidade em realizar determinadas tarefas [Larose et al, 2005]. As tarefas mais comuns são:

- **Classificação:**
 - visa identificar a qual classe um determinado registro pertence.
- Exemplo
 - Conjunto de Dados: Colaboradores de uma empresa
 - Classes: Perfil Técnico, Perfil Negocial e Perfil Gerencial
 - Objetivo: definir em qual perfil se encaixa um novo colaborador
- Outros exemplos
 - Detecção de fraudes de compras com cartão de crédito
 - Diagnóstico de doenças

Principais tarefas

- **Regressão ou Estimação:**
 - estimar o valor de uma determinada variável analisando-se os valores das demais (atributos numéricos)
 - O objetivo é prever os valores de uma variável dependente com base em resultados da variável independente
 - Regressão Linear: $y = aX + B$
- Exemplo
 - Estimar a pressão ideal de um paciente, com base na altura, idade e peso

Principais tarefas

- **Agrupamento (*Clustering*)**
 - Identifica e aproxima registros similares, agrupando-os em seguida
 - Não necessita que os dados esteja previamente categorizados
 - Aprendizado não-supervisionado
 - Exemplos
 - Agrupar pessoas por altura, sexo e hábitos
 - Agrupar clientes com comportamento de compra similar

Principais tarefas

- **Associação**
 - Consiste em identificar quais atributos estão relacionados.
 - Apresentam a forma: SE *atributo X* ENTÃO *atributo Y*.
 - X é o antecedente
 - Y é o consequente
 - Encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.
- Exemplos
 - Identificar quais produtos são levados junto com outros pelos consumidores (Cesta de compras)

Principais técnicas

- **Aprendizado Supervisionado**
 - Necessita de dados pré-categorizados
 - Variável-alvo pré-definida
- **Não-Supervisionado**
 - Não necessita de dados pré-categorizados
 - Utilização de medidas de similaridade
- Combinação de diversas tarefas/técnicas
 - Comparações acerca da melhor técnica ou do melhor conjunto de técnicas para determinado problema



Principais técnicas

- Harrison (1998) afirma que não há uma técnica que resolva todos os problemas de mineração de dados.
 - Diferentes métodos para diferentes propósitos
 - Vantagens e desvantagens.
 - Importância da familiaridade com as técnicas
- Principais técnicas de Mineração de Dados
- Regras de Associação
 - estabelece uma correlação estatística entre certos itens de dados em um conjunto de dados (Goebel e Gruenwald, 1999).
 - Regra: $X_1 \wedge \dots \wedge X_n \Rightarrow Y [C, S]$, onde $X_1, \dots, X_n$ são itens que preveem a ocorrência de Y com um grau de confiança C e com um suporte mínimo de S e “ $\wedge$ ” denota um operador de conjunção (AND).

Principais técnicas

- Regras de Associação (Cont.)

Conclusões simples:

- Quem compra leite, quase sempre compra ovos;
- Quem compra ovos e açúcar sempre compra leite;
- Quem compra cerveja sempre compra fraldas.

Transação	Itens
1	leite, ovos, café, açúcar, fraldas, manteiga
2	leite, café, farinha
3	leite, ovos, açúcar
4	café, açúcar
5	fraldas
6	manteiga, ovos, leite
7	café, açúcar, leite, ovos
8	farinha, manteiga, ovos
9	manteiga, ovos, leite, café, açúcar
10	fraldas, café, cerveja

Principais técnicas

- Regras de Associação (Cont.)
 - Métricas (Garantia de resultados válidos)
 - Suporte
 - porcentagem de transações onde um determinado item/itemset aparece
 - Confiança
 - Dado $A \rightarrow B$, é porcentagem das transações que suportam " $A \cup B$ " dividida por todas as transações que suportam A
 - Algoritmos
 - **Apriori**, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen et al, 1996).

Principais técnicas

- Regras de Associação (Cont.)
 - *Apriori* (Projeto Quest, IBM)
 1. Entrada: coleção de dados associados, suporte mínimo, confiança mínima.
 2. Considerar $K = 1$ para criação de K-itemsets
 3. Analisar os dados associados e criar uma tabela de K-itemsets com suporte acima do suporte mínimo.
 4. Criar com os itemsets filtrados um conjunto de candidatos a $(K + 1)$ itemsets.
 5. Eliminar itemsets infrequentes.
 6. Repetir desde o passo 3 até que o conjunto gerado seja vazio.
 7. Listar regras de associação (com permutações) e aplicar limite de confiança.

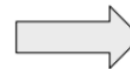
Principais técnicas

- Regras de Associação (Cont.)
 - Apriori* ($S = 25\%$, $C = 75\%$)

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1



1-itemsets	Suporte
leite	60%
ovos	60%
café	60%
açúcar	50%
fraldas	30%
manteiga	40%
farinha	20%
cerveja	10%



2-itemsets	Suporte
[leite,ovos]	50%
[leite,café]	40%
[leite,açúcar]	40%
[leite,fraldas]	10%
[leite,manteiga]	30%
[ovos,café]	30%
[ovos,açúcar]	40%
[ovos,fraldas]	10%
[ovos,manteiga]	40%
[café,açúcar]	40%
[café,fraldas]	20%
[café,manteiga]	20%
[açúcar,fraldas]	10%
[açúcar,manteiga]	20%
[fraldas,manteiga]	10%

Principais técnicas

- Regras de Associação (Cont.)
 - Apriori* ($S = 25\%$, $C = 75\%$)

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1

2-itemsets	Suporte
[leite,ovos]	50%
[leite,café]	40%
[leite,açúcar]	40%
[leite,fraldas]	10%
[leite,manteiga]	30%
[ovos,café]	30%
[ovos,açúcar]	40%
[ovos,fraldas]	10%
[ovos,manteiga]	40%
[café,açúcar]	40%
[café,fraldas]	20%
[café,manteiga]	20%
[açúcar,fraldas]	10%
[açúcar,manteiga]	20%
[fraldas,manteiga]	10%



Regra	Suporte	Confiança
[ovos → leite]	50%	83%
[leite → ovos]	50%	83%
[café → leite]	40%	66%
[leite → café]	40%	66%
[açúcar → leite]	40%	80%
[leite → açúcar]	40%	66%
[manteiga → leite]	30%	75%
[leite → manteiga]	30%	50%
[café → ovos]	30%	50%
[ovos → café]	30%	50%
[açúcar → ovos]	40%	80%
[ovos → açúcar]	40%	66%
[manteiga → ovos]	40%	100%
[ovos → manteiga]	40%	66%
[açúcar → café]	40%	80%
[café → açúcar]	40%	66%

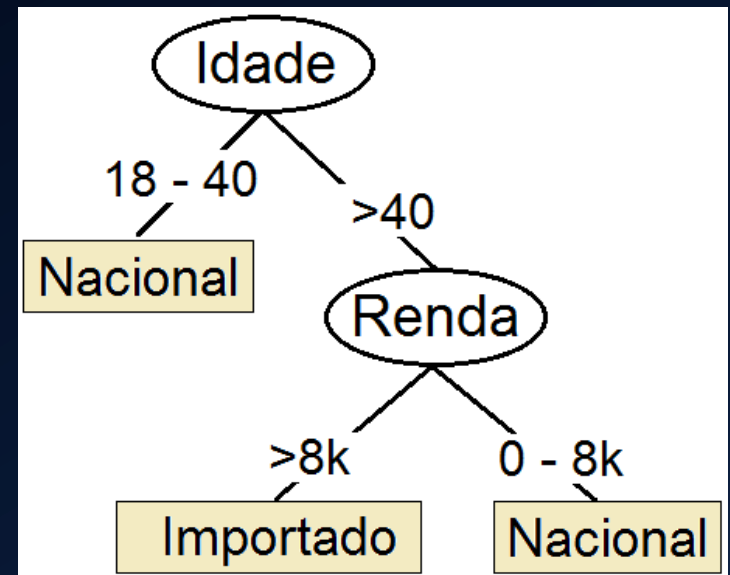
Principais técnicas

- Árvores de Decisão
 - Árvore onde cada nó não-terminal representa um teste ou decisão sobre o item de dado considerado (Goebel e Gruenwald, 1999).
 - Métricas utilizadas
 - Entropia (aleatoriedade da variável-objetivo)
 - Ganho (diminuição da entropia)
 - Fluxo
 - Nó raiz (uma dada entrada)
 - Análise de nós não-terminais (decisões)
 - Algum nó-folha (categoria ou classe)
- Técnica simples, bom grau de assertividade e não precisa de parâmetros de configuração

Principais técnicas

- **Árvores de Decisão**
 - **Entrada:** base de dados, com atributos preditivos e atributo-alvo.
 - **Saída:** classificação do atributo alvo em função dos preditivos.

Id	Idade	Renda	Carro
1	>40	> 8k	Importado
2	18 – 40	0-8k	Nacional
3	>40	0-8k	Nacional
4	18 – 40	> 8k	Nacional
5	18 – 40	0-8k	Nacional
6	18 – 40	0-8k	Nacional
7	18 – 40	0-8k	Nacional
8	>40	> 8k	Importado



Principais técnicas

- K-means
 - Aprendizagem não-supervisionada
 - Utilizada em tarefas de *Agrupamento*
 - Conceito de *Centróides*
 - Entrada:
 - Instâncias: conjunto de dados
 - Medida de distância (similaridade):
 - Distância Euclidiana
 - Cosseno
 - Número de grupos (clusters)
 - Saída:
 - Centróides dos grupos, pertinência das instâncias

Principais técnicas

- K-means
 - Algoritmo (iterativo)
 - 1. Definir K centróides iniciais, selecionando K objetos aleatórios;
 - 2. Alocar cada objeto para o cluster correspondente ao centróide mais similar;
 - Distância Euclidiana, Cosseno, Manhattan.
 - 3. Recalcular os centróides dos clusters;
 - 4. Voltar ao passo 2

Principais técnicas

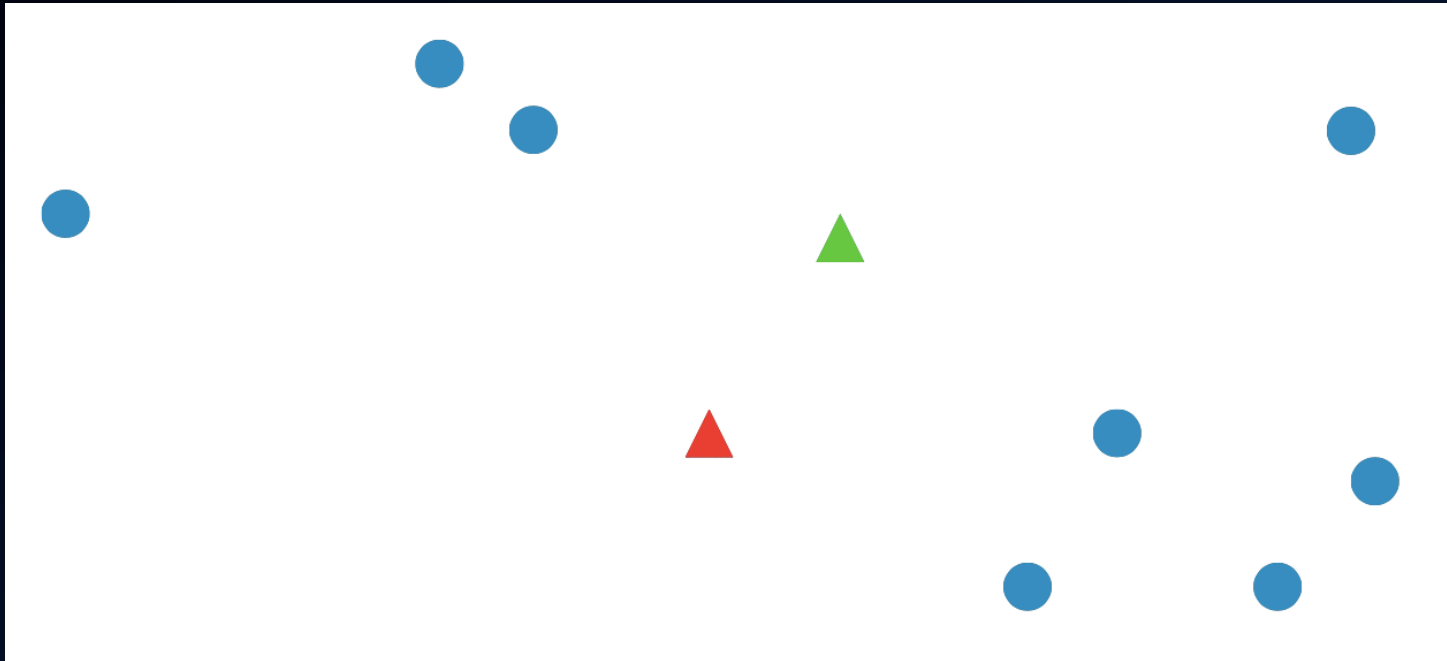
- K-means
 - Encontrar os centros ótimos
 - Minimizar a junção das distâncias quadráticas (média dos pontos)

$$e^2(D, C) = \sum_{j=1}^K \sum_{\vec{d}_i \in C_j} \left\| \vec{d}_i - c_j \right\|^2$$

- Critério de Parada
 - Não ocorrerem mais alterações nos centróides (convergência dos centros ótimos)
- Se um centróide não tiver registros associados no início, ele deve ser reiniciado para escolha de novo centróide

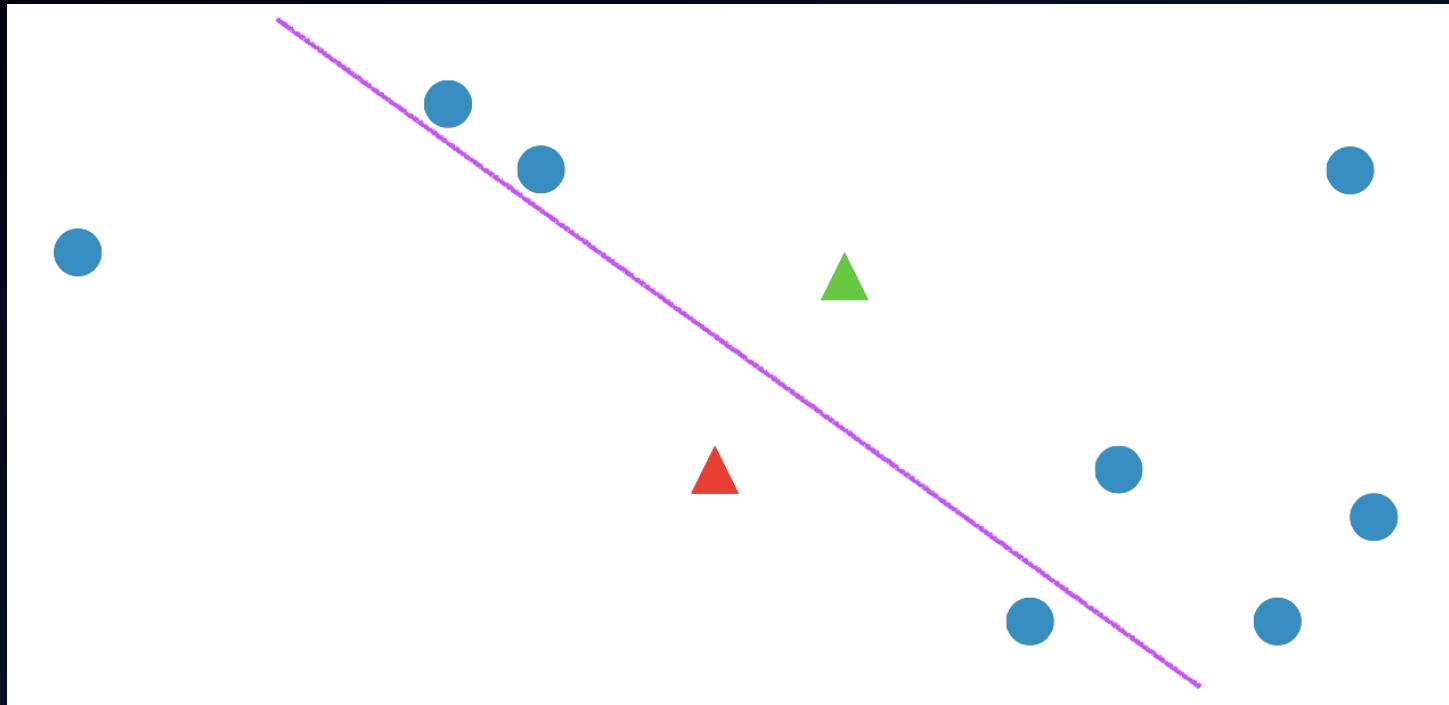
Principais técnicas

- K-means



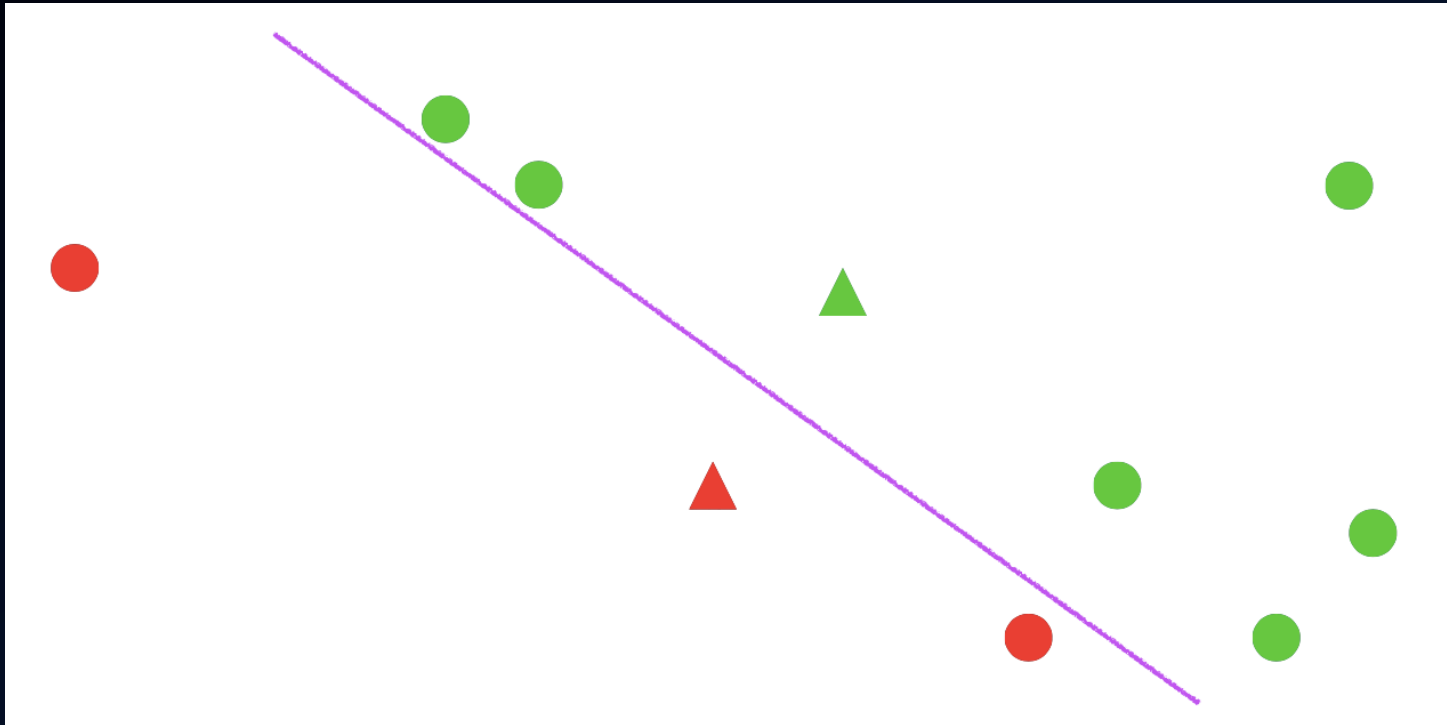
Principais técnicas

- K-means



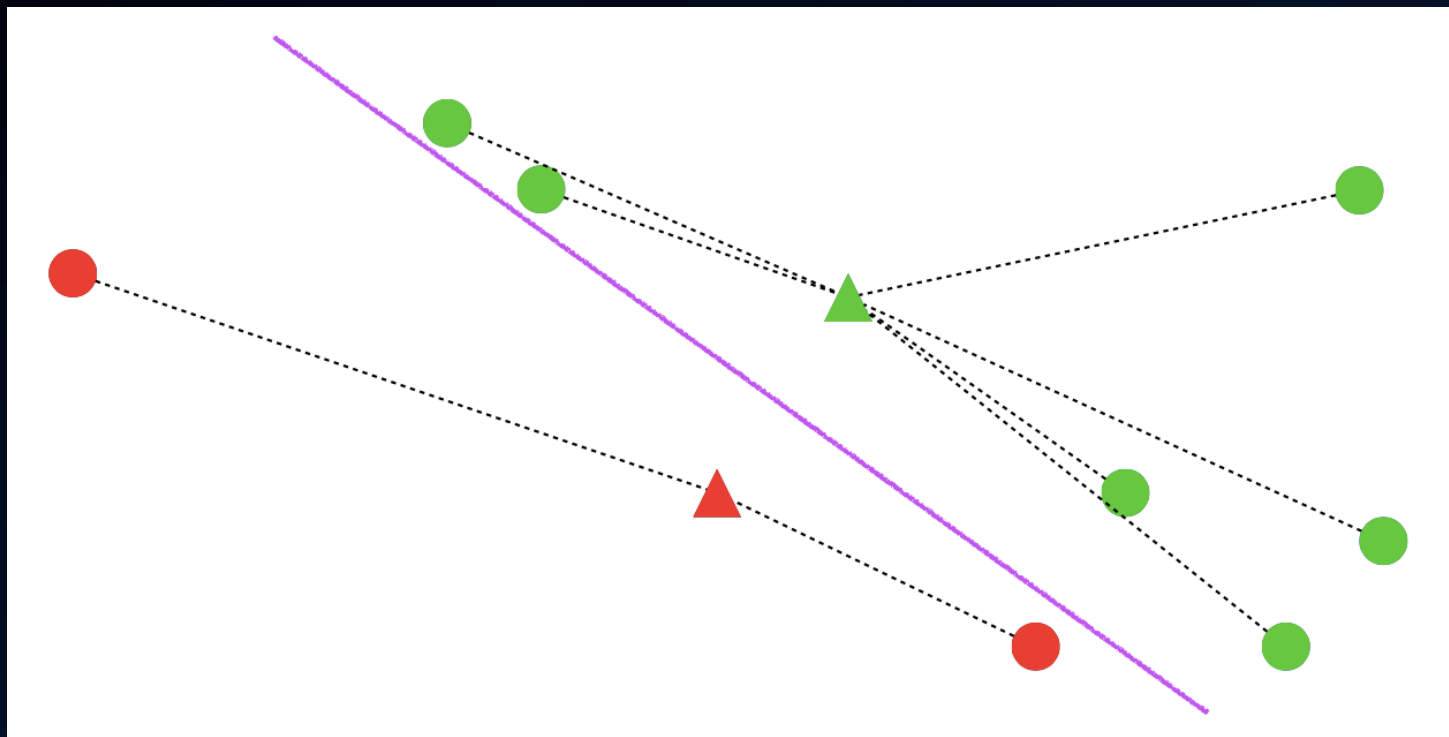
Principais técnicas

- K-means



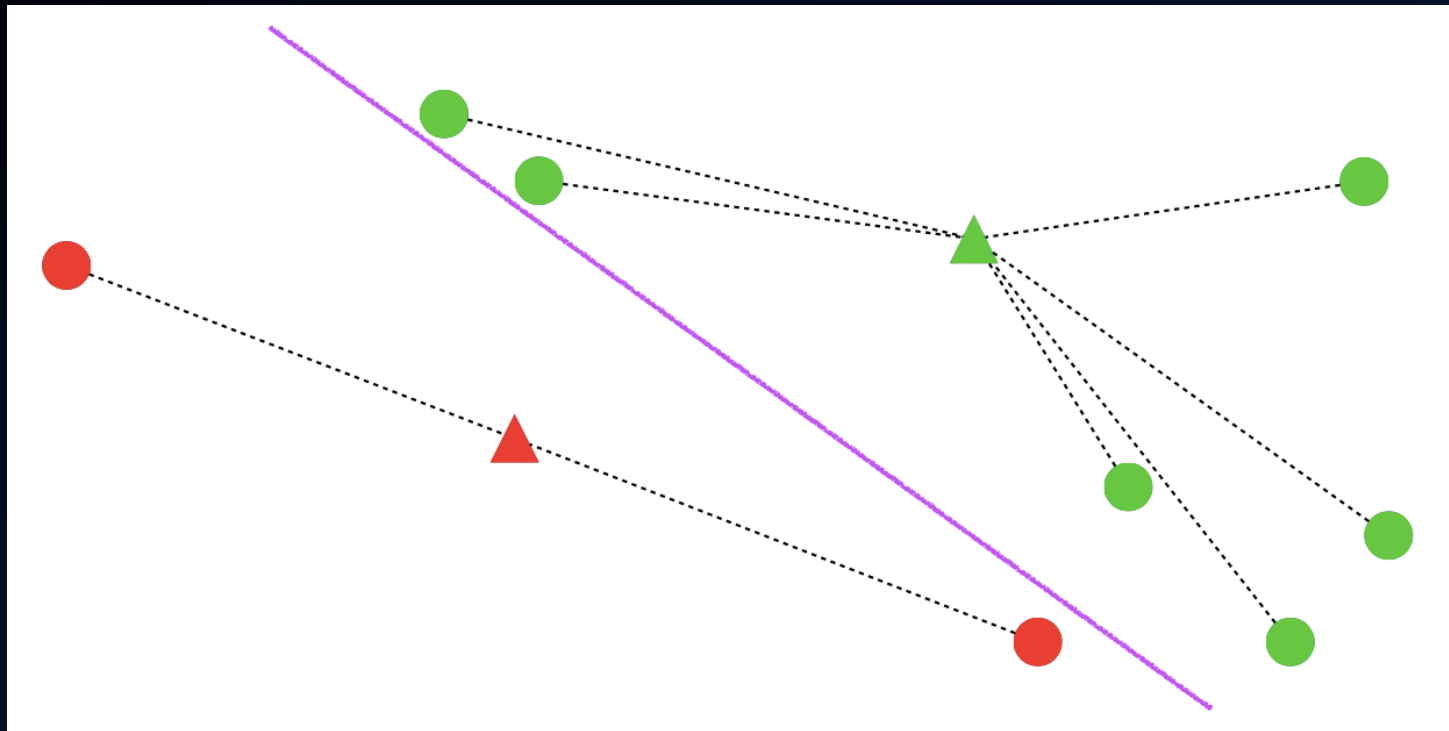
Principais técnicas

- K-means



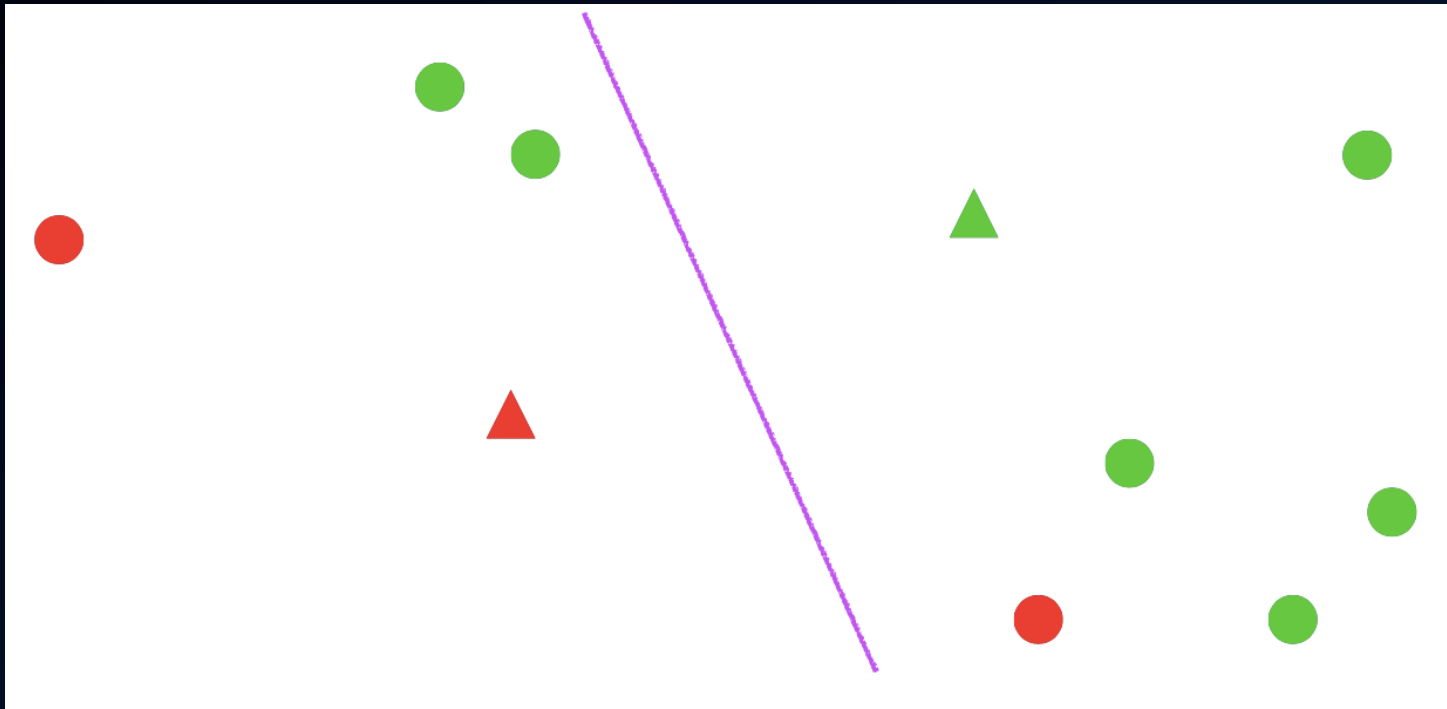
Principais técnicas

- K-means



Principais técnicas

- K-means



Como escolher a Técnica de Mineração de Dados Mais adequada

- Depende dos dados disponíveis para análise
- Tarefa expecífica a ser executada (Agrupamento, Classificação)
- 2 passos (Harrison, 1998)
 - 1. traduzir o problema de negócio a ser resolvido em séries de tarefas de mineração de dados;
 - 2. compreender a natureza dos dados (conteúdo, campos, relações).
- Segundo Chen et al,1996
 - Com que tipos de bancos de dados trabalhar
 - Qual o tipo de conhecimento a ser explorado
 - Qual tipo de técnica a ser utilizada

Mineração de dados complexos

- Mineração realizado sobre diversos tipos de dados
 - Inicialmente: Repositórios estruturados de dados
 - Bancos de Dados, Data Warehouses, etc.
 - Atualmente: Repositórios de dados em geral
 - Multimídia, Web, Temporal, Grafos
- Alguns Tipos
 - **Mineração de fluxo de dados:**
 - alto volume de dados (análise sob demanda)
 - Algoritmo *appearing-bit-sequence-based incremental mining*
 - **Mineração de séries temporais**
 - Informações de eventos num dado tempo
 - Mineração temporal por agrupamento: fatores de risco na área médica

Mineração de dados complexos

- Tipos (Cont)
 - Mineração de textos:
 - Recuperação de Informação (documentos)
 - Resolver problemas como documentos não relacionados, volume alto de documentos, informações dispensáveis
 - Mineração da Internet (ou *Web Mining*)
 - Consiste em minerar as estruturas de ligação, o conteúdo, os padrões de acesso, classificação de documentos, entre outras

Exemplo de ferramenta de Mineração

- Weka
 - Software livre open source para mineração de dados, desenvolvido em Java.
 - Desenvolvido por pesquisadores da Universidade de Waikato, Nova Zelândia.
 - Consolidou-se como a ferramenta de data mining mais utilizada em ambiente acadêmico.
- Potenciais: Classificação, Regras de Associação e Clusterização
- Modo de Uso: **Weka API Java**, Weka Explorer
- Arquivos .arff para expressar a base de conhecimento



Exemplo de ferramenta de Mineração

- Weka
 - Titanic (Exemplo de Aplicação)
 - Objetivo: classificação dos sobreviventes do naufrágio
 - Problema idealizado por Kaggle
<https://www.kaggle.com/c/titanic#evaluation>
 - Utilizar informações dos passageiros (treinamento)
 - Criar um classificador (Rede Bayesiana, Árvore de Decisão, Regressão) utilizando os dados de treinamento
 - Classificar um conjunto de testes
 - Atributo-alvo: **survived** {0,1}

Exemplo de ferramenta de Mineração

```
@relation train
```

```
@attribute PassengerId numeric
```

```
@attribute Survived {0,1}
```

```
@attribute Pclass numeric
```

```
@attribute Sex {male,female}
```

```
@attribute Age numeric
```

```
@attribute SibSp numeric
```

```
@attribute Parch numeric
```

```
@attribute Ticket {'A/5 21171','PC 17599', ...}
```

```
@attribute Fare {7.925,53.1,8.05,84.583,518.625, ...}
```

```
@attribute Embarked {S,C,Q}
```

Exemplo de ferramenta de Mineração

@data

```
1,0,3,male,22,1,0,'A/5 21171',7.25,S
2,1,1,female,38,1,0,'PC 17599',712.833,C
3,1,3,female,26,0,0,'STON/O2. 3101282',7.925,S
4,1,1,female,35,1,0,113803.0,53.1,S
5,0,3,male,35,0,0,373450.0,8.05,S
6,0,3,male,?,0,0,330877.0,84.583,Q
7,0,1,male,54,0,0,17463.0,518.625,S
8,0,3,male,2,3,1,349909.0,21.075,S
9,1,3,female,27,0,2,347742.0,111.333,S
10,1,2,female,14,1,0,237736.0,300.708,C
...
```

Exemplo de ferramenta de Mineração

```
@relation test
```

```
@attribute PassengerId numeric
```

```
@attribute Survived {0,1}
```

```
@attribute Pclass numeric
```

```
@attribute Sex {male,female}
```

```
@attribute Age numeric
```

```
@attribute SibSp numeric
```

```
@attribute Parch numeric
```

```
@attribute Ticket {'A/5 21171','PC 17599', ...}
```

```
@attribute Fare {7.925,53.1,8.05,84.583,518.625, ...}
```

```
@attribute Embarked {S,C,Q}
```

Exemplo de ferramenta de Mineração

@data

```
892,?,3,male,34.5,0,0,330911.0,78.292,Q
893,?,3,female,47,1,0,363272.0,7.0,S
894,?,2,male,62,0,0,240276.0,96.875,Q
895,?,3,male,27,0,0,315154.0,86.625,S
896,?,3,female,22,1,1,3101298.0,122.875,S
897,?,3,male,14,0,0,7538.0,9.225,S
898,?,3,female,30,0,0,330972.0,76.292,Q
899,?,2,male,26,1,1,248738.0,29.0,S
900,?,3,female,18,0,0,2657.0,72.292,C
```

...

Exemplo de ferramenta de Mineração

```
Instances trainDataSet = loadFileAndReturnDataSet("train.arff");
Instances testDataSet = loadFileAndReturnDataSet("test.arff");

StringToWordVector filter = new StringToWordVector();
filter.setInputFormat(trainDataSet);
trainDataSet = Filter.useFilter(trainDataSet, filter);

BayesNet classifier = new BayesNet();
classifier.buildClassifier(trainDataSet);

trainDataSet.setClass(trainDataSet.attribute(1));
testDataSet.setClass(testDataSet.attribute(1));

Enumeration enumDataSet = testDataSet.enumerateInstances();

while (enumDataSet.hasMoreElements()) {
    try {
        Instance test = (Instance) enumDataSet.nextElement();
        test.setClassValue(classifier.classifyInstance(test));
        System.out.println(test.toString());
    } catch (Exception ex) {
        continue;
    }
}
```

Exemplo de ferramenta de Mineração

Resultado

892,0,3,male,34.5,0,0,330911.0,78.292,Q
893,1,3,female,47,1,0,363272.0,7.0,S
894,1,2,male,62,0,0,240276.0,96.875,Q
895,0,3,male,27,0,0,315154.0,86.625,S
896,0,3,female,22,1,1,3101298.0,122.875,S
897,0,3,male,14,0,0,7538.0,9.225,S
898,0,3,female,30,0,0,330972.0,76.292,Q
899,0,2,male,26,1,1,248738.0,29.0,S
900,1,3,female,18,0,0,2657.0,72.292,C

...

Referências

- FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37, 1996.
- HAN, J; KAMBER, M. Data Mining: Concepts and Techniques. Elsevier, 2006.
- BERRY, Michael J. A.; LINOFF, Gordon. **Data Mining Techniques: For Marketing, Sales, and Customer Support**. New York: Wiley Computer Publishing, 1997.
- GROTH, R. Data Mining: a hands-on approach for business professionals. Prentice Hall, New Jersey, 1998.
- CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Goiânia: Universidade Federal de Goiás, 2009.
- HARRISON, T. H. INTRANET DATA WAREHOUSE. Berkley, 1998

Referências

- GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. In: SIGKDD Explorations, 1999.
- LAROSE, D. T. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons. Inc., 2005.