Aula 2 – Mineração de Dados
Introdução

Profa. Elaine Faria

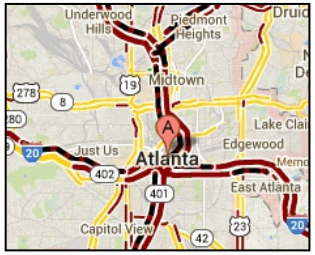
UFU

Motivação

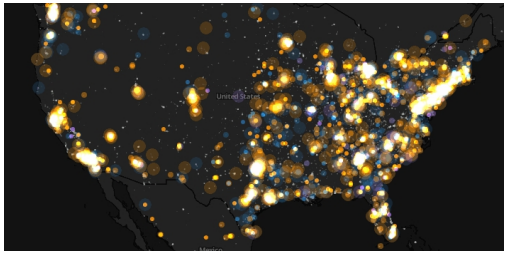
- Enorme crescimento das bases de dados comerciais e científicas
  - Avanços na geração de dados e tecnologias de coleta de dados
- Expectativa
  - Os dados coletados terão valor para a finalidade coletada ou para uma finalidade não prevista

Cyber Security

E-Commerce

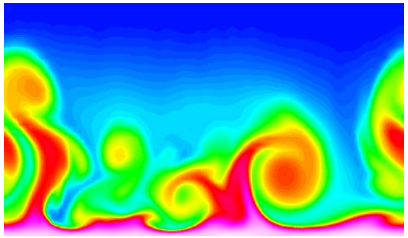


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

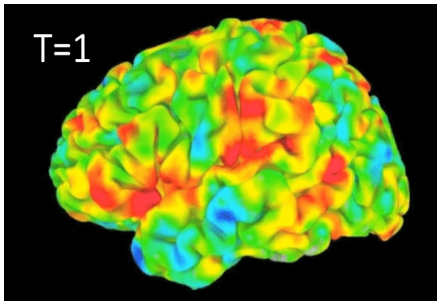
Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2a edição 2018

Por que mineração de dados? Ponto de Vista Comercial

- Muitos dados estão sendo coletados e armazenados
  - Dados da web
    - Ex: Google tem Peta Bytes de dados da web
    - Ex: Facebook tem bilhões de usuários ativos
  - Compras em lojas de departamento / supermercados, e-commerce
    - Amazon lida com milhões de visitas / dia
  - Transações bancárias / com cartão de crédito
- Os computadores se tornaram mais baratos e mais poderosos
- A pressão competitiva é forte
  - Fornecer serviços melhores e personalizados

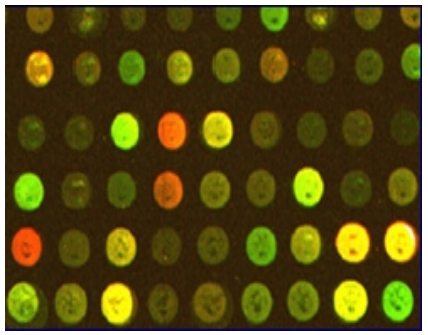
Por que mineração de dados? Ponto de Vista Científico

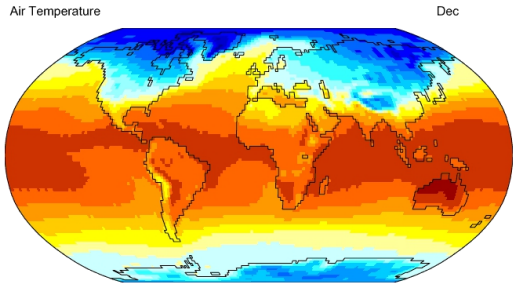
- Dados coletados e armazenados em velocidades enormes
  - Sensores remotos em um satélite
  - Telescópios explorando os céus
  - Dados biológicos
  - Simulações científicas
- A mineração de dados ajuda os cientistas
  - Na análise automatizada de grandes conjuntos de dados
  - Na formação de hipóteses



fMRI Data from Brain


Sky Survey Data


Gene Expression Data



Surface Temperature of Earth

Grande oportunidade para resolver problemas importantes!

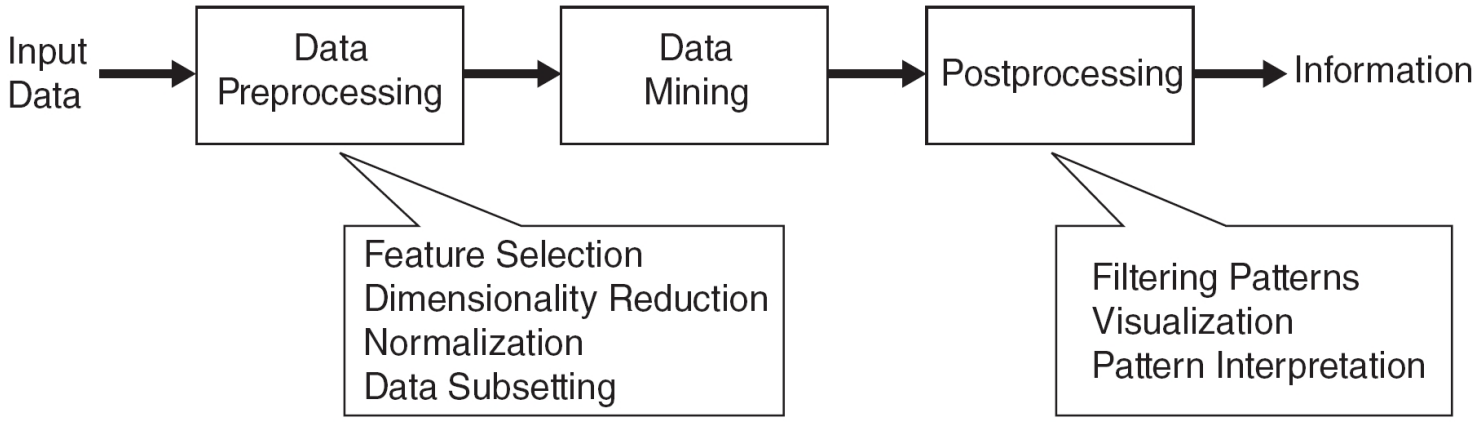
Mineração de Dados - Definição

Processo de automaticamente descobrir informação útil em grandes repositórios de dados

Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Mineração de Dados - Definição

- Extração não trivial de informações implícitas, anteriormente desconhecidas, e potencialmente úteis dos dados
- Exploração e análise, por meios automáticos ou semiautomáticos de grandes quantidades de dados para descobrir padrões significativos



Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2a edição 2018

Mineração de Dados - Motivação

- Algumas das Motivações
  - Grandes quantidades de dados geradas
  - Dados de alta dimensionalidade
  - Dados heterogêneos e complexos
  - Dados distribuídos
  - Análise não tradicional

Mineração de Dados - Origem

- Extrai ideias de aprendizado de máquina / IA, reconhecimento de padrões, estatísticas e sistemas de banco de dados
- As técnicas tradicionais podem ser inadequadas devido aos dados:
  - Grande escala
  - Alta dimensionalidade
  - Heterogêneo
  - Complexo
  - Distribuído
- É uma componente-chave dos campo emergente da ciência de dados e descoberta baseada em dados

Mineração de Dados e Aprendizado de Máquina

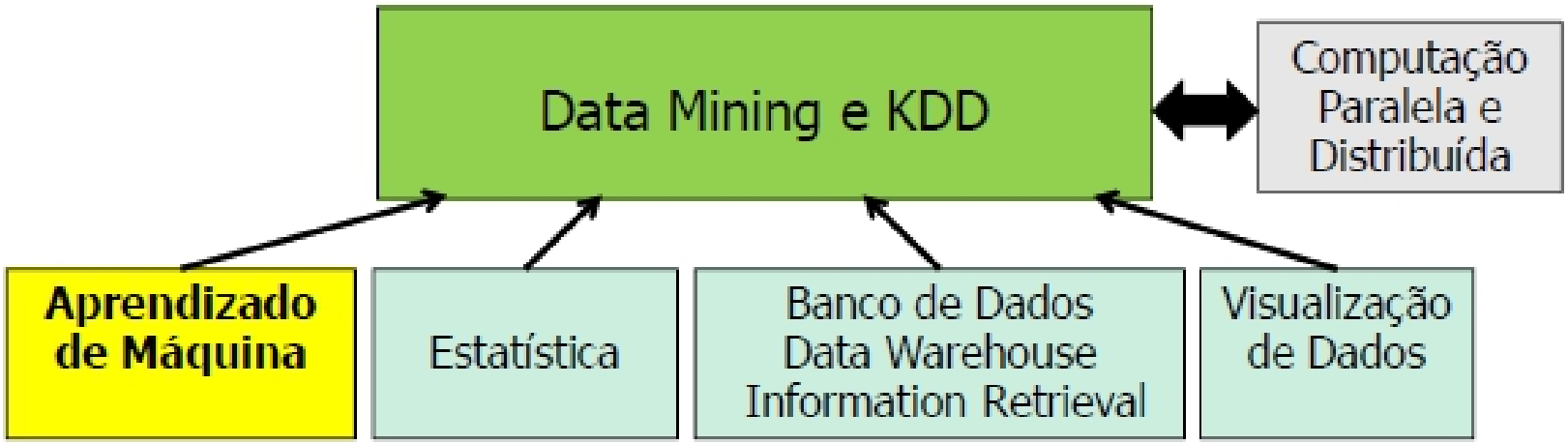
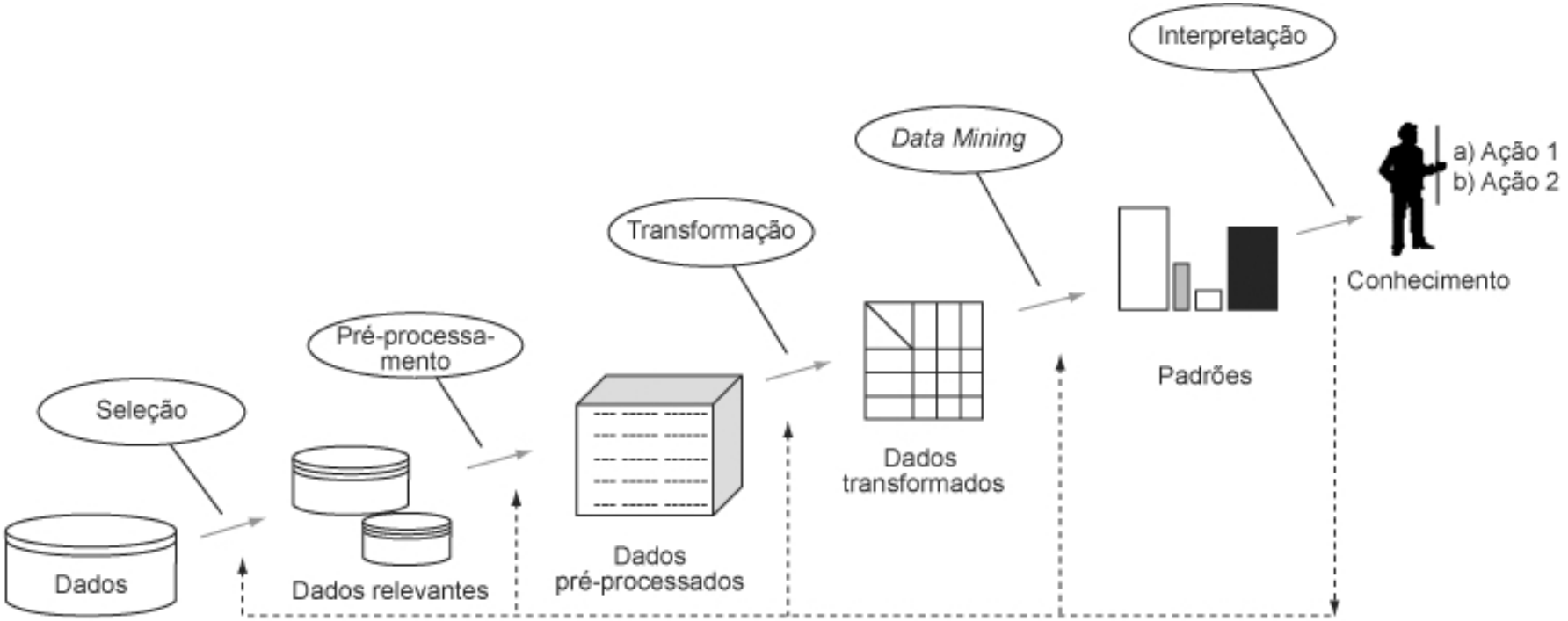


Figura retirada dos slides do Prof. Ricardo Campello – ICMC-USP

Mineração de Dados - Exemplos

- Exemplos de aplicação
  - Negócios
    - Padrões de compra/ligações → Marketing
    - *Logs* da Web → Melhor design de sites
    - Detecção de fraude
    - Agrupamento de clientes
  - Bioinformática
    - Genes importantes
    - Sintomas associados a doenças

Descoberta do conhecimento em bases de dados - KDD



Etapas do processo *KDD*

Fayyad U. et al. Knowledge Discovery and Data Mining: Towards a Unfiying Framework, KDD 96, 1996

Tarefas da Mineração de Dados

- Tarefas Preditivas
  - Usa algumas variáveis para prever valores desconhecidos ou futuros de outras variáveis
    - Classificação
    - Regressão
- Tarefas Descritivas
  - Encontra padrões interpretáveis por humanos, que descrevem os dados
    - Agrupamento
    - Regras de associação

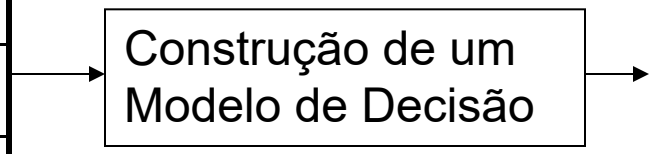
Classificação

- Objetivo: Encontrar um modelo para o atributo classe em função dos valores dos outros atributos
- Conhecido como aprendizado supervisionado
  - Indução do modelo a partir de um conjunto de dados rotulados

Exemplo de Classificação

Nome	Idade	Renda	Pagador
João	<30	Média	Bom
Ana	41..50	Alta	Bom
Pedro	41..50	Alta	Bom
Maria	41..50	Baixa	Ruim
Paulo	<30	Baixa	Ruim
Aldo	>60	Alta	Ruim

Base de Dados
Treinamento



Algoritmo

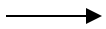
Se idade = 41..50 e
Renda = Alta então
Pagador = Bom

Se renda = baixa
então
Pagador = Ruim

Modelo

Exemplo de Classificação

Nome	Idade	Renda	Pagador
Ivo	31..40	Baixa	**????**



Se idade = 41..50 e
Renda = Alta então
Pagador = Bom

Se renda = baixa
então
Pagador = Ruim

Ruim

Novo Dado

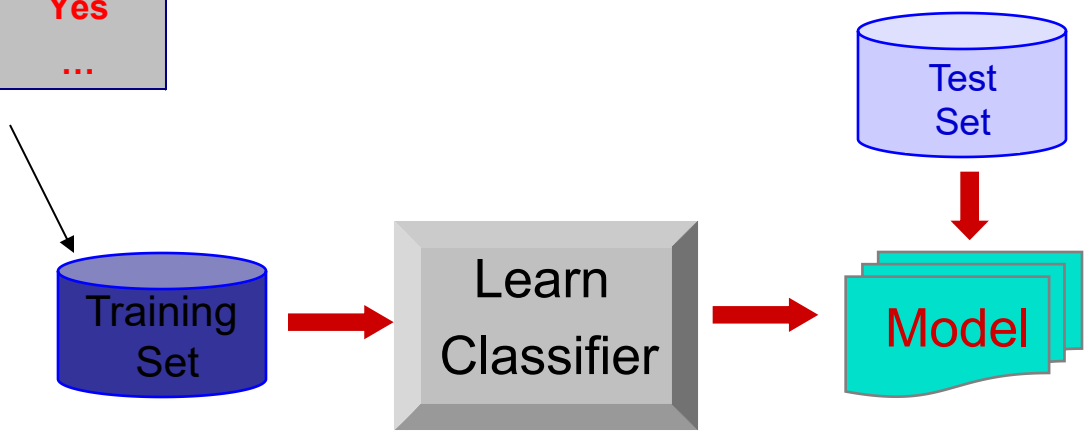
Classificador

Classificação

Classificação

Tid	Employed	Level of Education	# years at present address	Credit Worthy
	categorical	categorical	quantitative	class
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Exemplos de aplicações de classificação

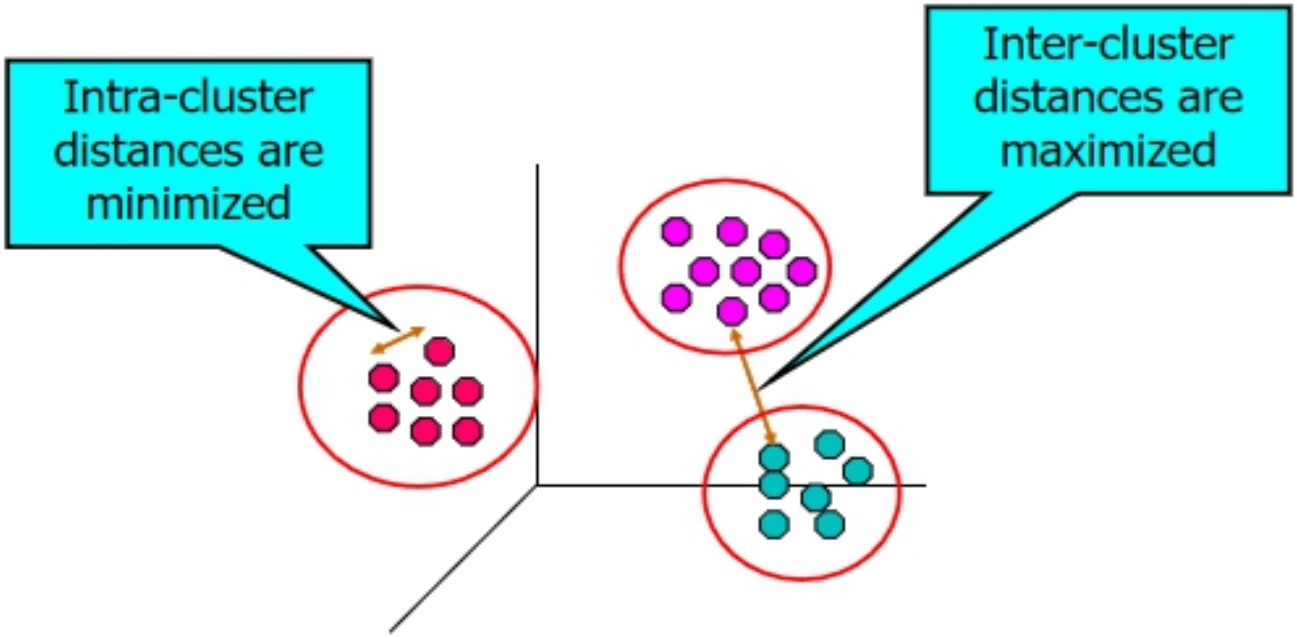
- Classificar transações de cartão de crédito como legítimas ou fraudulentas
- Classificar coberturas do solo (corpos d'água, áreas urbanas, florestas, etc.) usando dados de satélite
- Categorizar notícias como finanças, clima, entretenimento, esportes, etc.
- Identificar intrusão em redes de computadores
- Predizer células tumorais como benignas ou malignas
- Predizer risco de churn de clientes de telefonia

Regressão

- Objetivo: Predizer um valor de uma determinada variável de valor contínuo com base nos valores de outras variáveis, assumindo um modelo linear ou não linear de dependência
- Exemplos:
  - Predição do valor das vendas de um novo produto com base em despesas de consultoria
  - Predição das velocidades do vento em função da temperatura, umidade, pressão do ar, etc.
  - Predição de séries temporais de índices do mercado de ações.

Agrupamento

- Objetivo: Encontrar grupos de objetos de forma que os objetos em um grupo sejam semelhantes (ou relacionados) entre si e diferentes (ou não relacionados) aos objetos em outros grupos

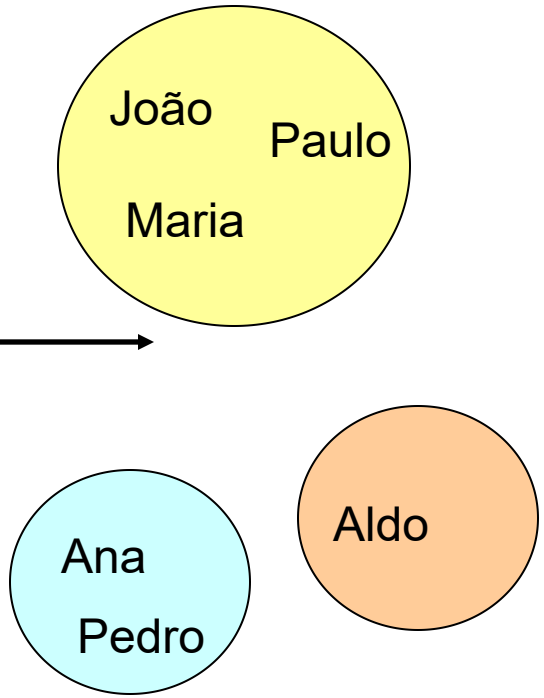


Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2a edição 2018

Exemplo de Agrupamento

Nome	Idade	Peso
João	25	60
Ana	50	75
Pedro	60	90
Maria	22	65
Paulo	18	68
Aldo	15	80

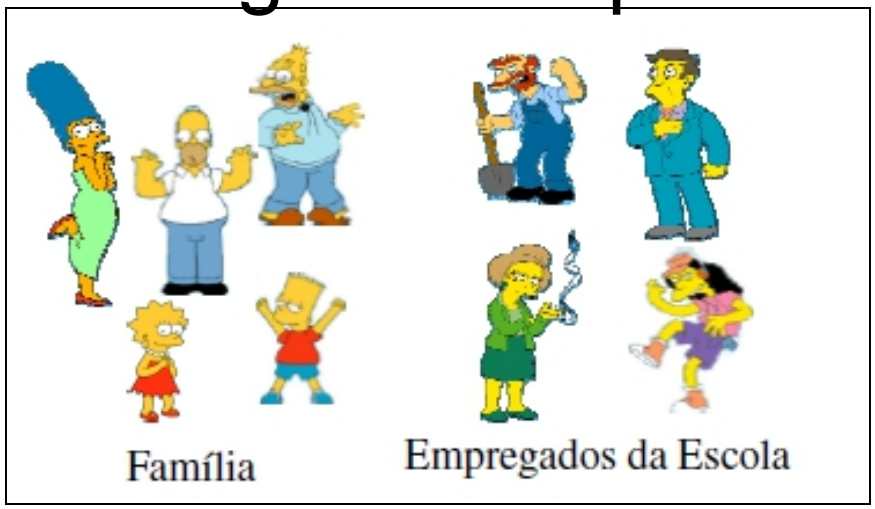
Aplicação de uma técnica agrupamento

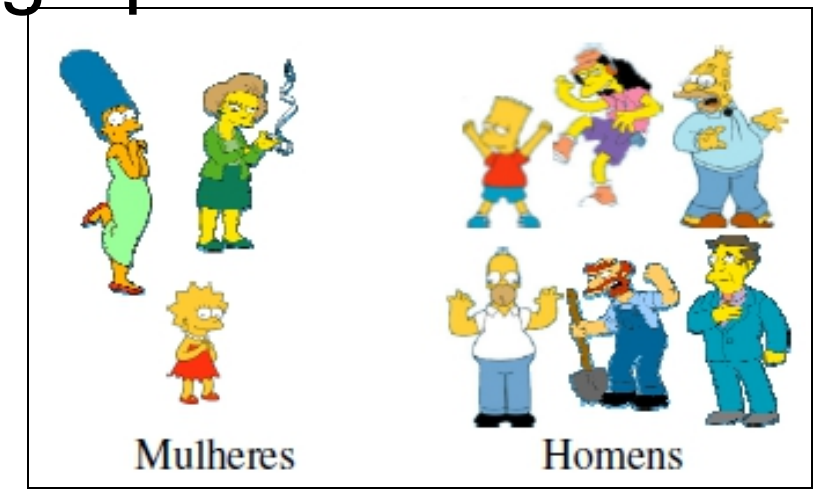


Agrupamento

- Como agrupar os objetos??

- Alguns dos possíveis agrupamento





Keogh, E. A g. Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Exemplos de aplicações de agrupamento de dados

- Marketing
  - Ex: Agrupamento de clientes para direcionar campanhas de marketing
- Documentos
  - Ex: agrupar documentos que tratam do mesmo assunto
- Imagens
  - Ex: Segmentar imagens usando técnicas de agrupamento
- Biologia
  - Ex: agrupar animais de acordo com o reino, ramo, classe, ordem, gênero, ...

Regras de Associação

- Objetivo: Dado um conjunto de instâncias, cada uma contendo algum número de itens de uma determinada coleção
  - Produzir regras de dependência que irão prever a ocorrência de um item com base nas ocorrências de outros itens.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

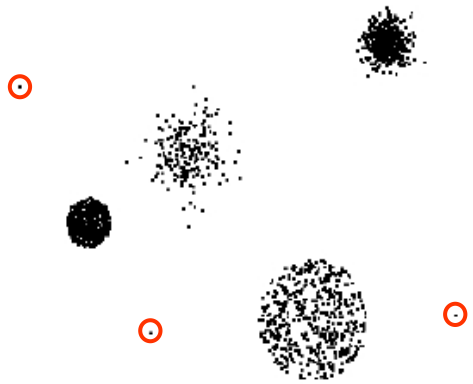
{Diaper, Milk} --> {Beer}

Exemplos de aplicações de regras de associação

- Análise de cesta de compras
  - As regras são usadas para promoção de vendas, gerenciamento de prateleiras e gerenciamento de estoque
- Diagnóstico de alarme de telecomunicação
  - As regras são usadas para encontrar a combinação de alarmes que ocorrem juntos frequentemente no mesmo período de tempo
- Informática Médica
  - As regras são usadas para encontrar a combinação de sintomas do paciente e resultados de testes associados a certas doenças

Detecção de Anomalias

- Detectar desvios significativos do comportamento normal
- Aplicações
  - Detecção de fraude de cartão de crédito
  - Detecção de intrusão em redes de computadores
  - Identificação de comportamento anômalo de redes de sensores para monitoramento e vigilância
  - Detecção de mudanças na cobertura florestal global



Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2a edição 2018

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
- Keogh, E. A g. Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.
- Aggarwal, C. Data Mining: The text book, Springer, 2015.