



Soluções em Mineração de dados: Pré-processamento

Prof. Leandro Almeida
lma3@cin.ufpe.br

Porque pré-processar os dados?

- As bases de dados hoje são extremamente grandes (da ordem de gibabytes e terabytes).
- As fontes de informação não são únicas e as vezes não são padronizadas.
- Toda base de dados está susceptível a conteúdo:
 - Ruidoso
 - Ex: contém erros, ou valores diferentes do esperado
 - Incompleto
 - Ex: atributos com valores faltosos, ou dados agregados
 - Inconsistente
 - Ex: discrepâncias nos códigos dos departamentos usados para categorizar itens

Sem dados de boa qualidade o resultado da mineração é pobre!

Porque pré-processar os dados?

Como podemos pré-processar os dados visando melhorar a qualidade dos dados e consequentemente os resultados da mineração?

Como podemos pré-processar os dados, de modo a melhorar a eficiência e a facilidade do processo de mineração?

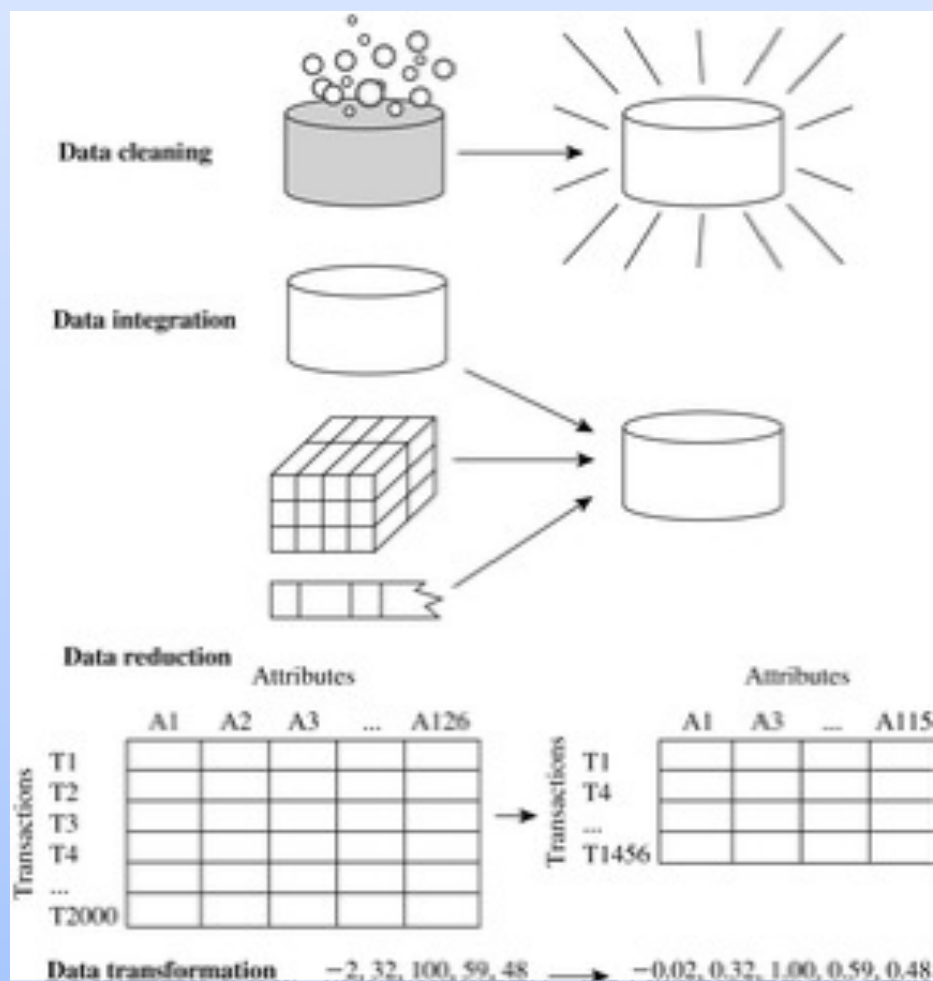
Qualidade dos dados: Por que pré-processar dados?

- Medidas de **qualidade de dados**: uma visão multidimensional
 - Um dado de qualidade deve ser...
 - Preciso
 - Ex: não pode existir valores incorretos, ou imprecisos, ...
 - Completo
 - Ex: não deve haver dados não registrados, ou indisponíveis, ...
 - Consistentes
 - Ex: não deve existir registros modificados, e outros não, dados pendentes, ...
 - Oportunos - Ocorrer no tempo certo
 - Ex: informações atrasadas não servem para tomar decisões
 - Confiáveis (Credibilidade)
 - Ex: não deve haver dúvidas quanto à correteude dos dados
 - Interpretáveis
 - Ex: os dados não podem ser difíceis de ser entendidos
-

Principais etapas do pré-processamento

- Limpeza dos Dados
 - preencher dados ausentes, “suavização” de ruído, identificar e/ou remover outliers, resolver inconsistências.
 - Integração dos Dados
 - Integração dados de múltiplas bases de dados, como data warehouse
 - Transformação dos Dados
 - normalização e agregação
 - Redução dos Dados
 - redução de dimensionalidade
 - redução no volume de dados
 - Compressão de dados a partir de características similares
-

Principais etapas do pré-processamento



Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- Dados não estão sempre disponíveis.
 - Ex: atributos com valores faltosos, ausência de atributos de interesse, ou existência de apenas dados agregados
 - Ex: renda do cliente em dados relativos a vendas.
 - A ausência de dados pode ser consequência:
 - mau funcionamento do equipamento.
 - inconsistência com outros dados gravados e consequente supressão.
 - não entrada de dados devido a enganos.
 - determinados dados podem não ser considerados importantes no momento do registro.
 - Pode ser necessário interferir nos dados.
-

Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- Tratamentos usuais:
 1. Ignorar a descrição do indivíduo ou mesmo eliminar o descritor
 - Ex: quando o rótulo da classe está faltando
 2. Preencher os valores ausentes manualmente - muitas vezes inviável
 3. Usar uma constante global para representar os valores ausentes (não recomendado, pois o sistema pode identificar esse valor como um conceito)
 4. Usar a média (ou a moda)
 - Ex: usar o valor médio de renda de uma amostra
 5. Usar a média (ou a moda) por classe
 - Ex: usar o valor médio de idade dos alunos do primeiro período de SI
 6. Usar o valor mais provável segundo um modelo (regressão, regra de Bayes, árvores de decisão)
-

Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- Análise dos métodos
 - Os métodos de 3 a 6 preenchem os dados faltosos de forma enviesada
 - Contudo, o método de regressão (6) é a estratégia mais usada em mineração de dados
 - O método 6 usa o máximo de informações dos dados atuais para prever valores em falta, aumentando as chances de acerto
- Em alguns casos, um valor em falta não pode implicar um erro nos dados
 - Por exemplo, candidatos a um cartão de crédito precisam fornecer o número de sua carteira de motorista, mas há pessoas que não têm carteira de motorista e podem deixar este campo em branco.
 - Os formulários devem permitir a especificação de valores como "não aplicável" e programas podem ser usados para descobrir outros valores nulos (ex: "Não sei", "?" ou "nenhum").
 - Idealmente, cada atributo deve ter um ou mais regras sobre a condição nula.

Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- Dado ruidoso ou outlier é um erro aleatório ou uma variabilidade em uma determinada variável.
 - Tratamentos usuais
 - Remoção de ruído
 - Alisamento (Suavização)
 - Regressão
 - Identificação de valores aberrantes
 - Clustering
-

Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- **Alisamento:**

- consiste em suavizar um valor de dados de acordo com seus vizinhos
- os dados ordenados são distribuídos em caixas tendo como referência os seus vizinhos.

Ordenação: 1, 1, 2, 3, 3, 3, 4, 5, 5, 7

Particionamento em “caixas”

Alisamento pela mediana

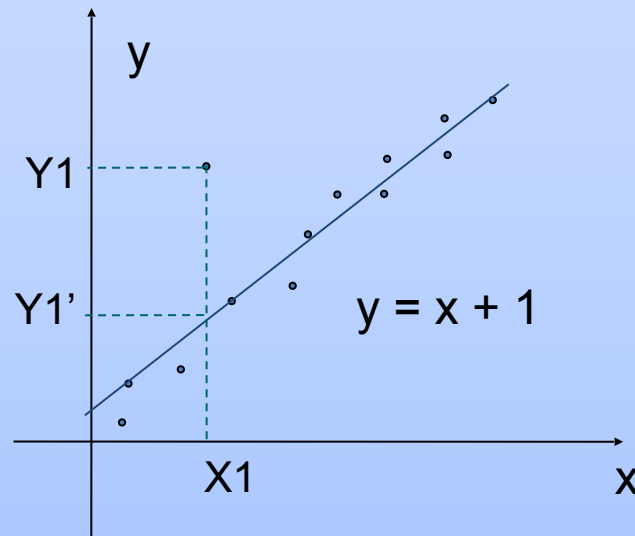
Outras alternativas: média, limites

1,1,2	3,3,3	4,5,5,7
caixa1	caixa2	caixa3
1,1,1	3,3,3	5,5,5,5
caixa1	caixa2	caixa3

Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

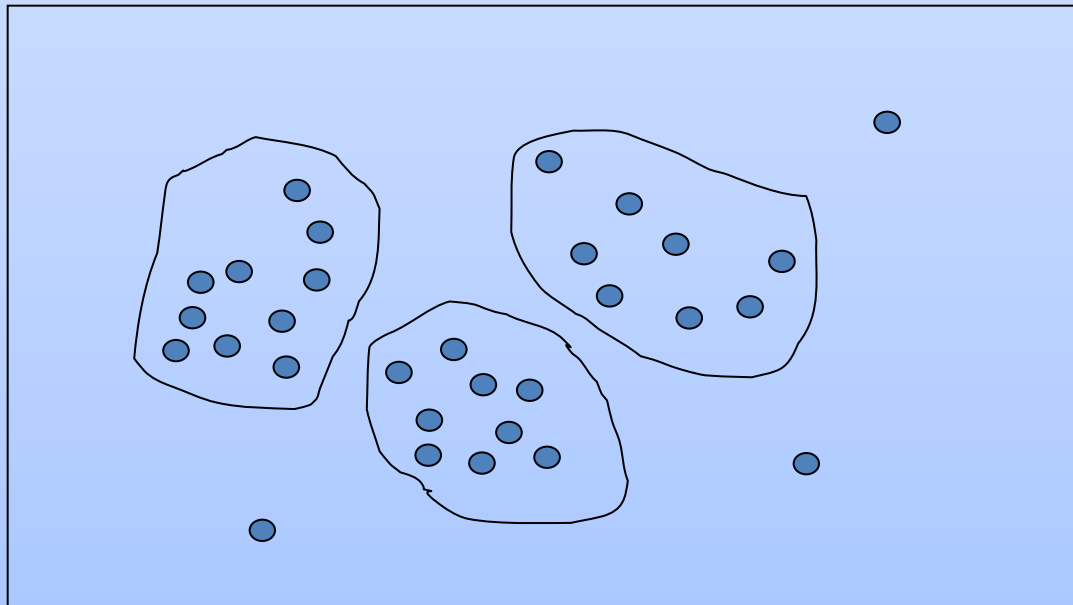
- **Regressão:** os dados podem ser alisados pelo ajustamento a uma função (regressão linear, por exemplo).



Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- **Clustering:** Os valores são organizados em grupos e os valores isolados podem ser considerados aberrantes.



Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- Muitos métodos de suavização podem ser utilizados para a discretização de dados (uma forma de transformação de dados) e redução de dados
 - Por exemplo, técnicas de suavização podem ser usadas para reduzir o número de valores para um atributo
 - Ex: um mapeamento pode ser feito entre preços reais e preços baratos, moderados e caros, reduzindo o número de valores possíveis a serem tratados pelo processo de mineração

Limpeza dos Dados

ausentes - ruidosos e/ou aberrantes - inconsistentes

- Erros no momento da introdução dos dados
 - Erros oriundos da integração de várias bases de dados
 - Mesmo atributo com diferentes atribuições
 - Masculino/Feminino - Homem/Mulher
 - Duplicação de objetos
 - Casa - Residência
 - Tratamento:
 - Correções manuais ou automática através de scripts.
-

Integração dos Dados

- Geralmente, a integração de dados é uma etapa necessária na Mineração de Dados
 - A fusão de dados a partir de diferentes fontes em uma única fonte coerente.
 - Visando evitar/reduzir redundâncias e inconsistências
 - Aumentar a acurácia e a velocidade dos passos subsequentes do processo de mineração
 - As fontes podem ser bases de dados, cubos ou arquivos texto.
 - Esquema em base de dados relacional
 - Identificação (correspondência) de entidade do mundo real a partir de múltiplas fontes de dados.
 - Integração dos metadados de diferentes fontes.
-

Integração dos Dados

- Problemas de integração:
 - Redundância:
 - Diferentes nomes para o mesmo atributo.
 - Ex: id_cliente e mat_cliente
 - Os metadados podem ser usados para evitar erros na integração
 - Atributo derivado de outro (Ex: receita anual)
 - Tratamento: Análise de correlação
 - Dado dois atributos, a ideia é saber como eles estão relacionadas

$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B}$$

Onde N é a quantidade de tuplas,
 a_i e b_i os respectivos valores de A e B,
 \bar{A} e \bar{B} são as respectivas médias e
 σ_A e σ_B são os desvios padrão

Integração dos Dados

- Problemas de integração:
 - Detecção e resolução de conflitos:
 - Os valores de um mesmo atributo pode diferir segundo as diversas fontes.
 - Isso pode acontecer devido a diferenças na representação, escala ou codificação.
 - Exemplos:
 - Peso (em libras ou em quilos)
 - Altura (valor numérico ou categórico (médio, pequeno...))
 - Preço ou dados de compra (pode indicar serviços diferentes)
 - Tratamento: Tabelas de conversão.
-

Transformação dos Dados

- Processo realizado para obter-se os dados em uma forma mais apropriada para a mineração.
 - Tratamentos:
 - **Normalização:** minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis .
 - As variáveis podem ser normalizadas segundo a amplitude ou segundo a distribuição.
 - Algumas ferramentas de modelização são beneficiadas com a Normalização (redes neurais, KNN, clustering).
-

Transformação dos Dados

- **Normalização min-max**

- Dados são escalados dentro de um intervalo [-1.0 - 0.0, ou 0.0 - 1.0]

$$v' = \frac{v - \min_A}{\max_A - \min_A} (new_max - new_min) + new_min_A$$

- Exemplo: Salário mínimo \$12,000 e máximo \$98,000, range [0.0, 1.0]. Normalizar um salário de \$73,600.

$$v' = \frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0.0) + 0.0 = 0.716$$

Transformação dos Dados

- **Normalização z-score**

- Normalização baseada na média e desvio padrão.

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

- Exemplo: Média dos salários \$54,000 e o desvio padrão \$16,000. Normalizar um salário de \$73,600.

$$v' = \frac{73,600 - 54,000}{16,000} = 1.225$$

Transformação dos Dados

- **Normalização por escala decimal**
 - Normaliza através do deslocamento de pontos decimais.
 - O número de pontos decimais depende do máximo valor absoluto dos dados.

$$v' = \frac{v}{10^j} \quad \text{onde } j \text{ é o menor inteiro tal que } \text{Max}(|v'|) < 1$$

- Exemplo: Suponha valores entre -986 e 917. O máximo valor absoluto é 986. Para normalizar dividimos as entradas por 1000 (ou seja, $j = 3$)

$$v' = \frac{-986}{1,000} = -0.986 \quad v' = \frac{917}{1,000} = 0.917$$

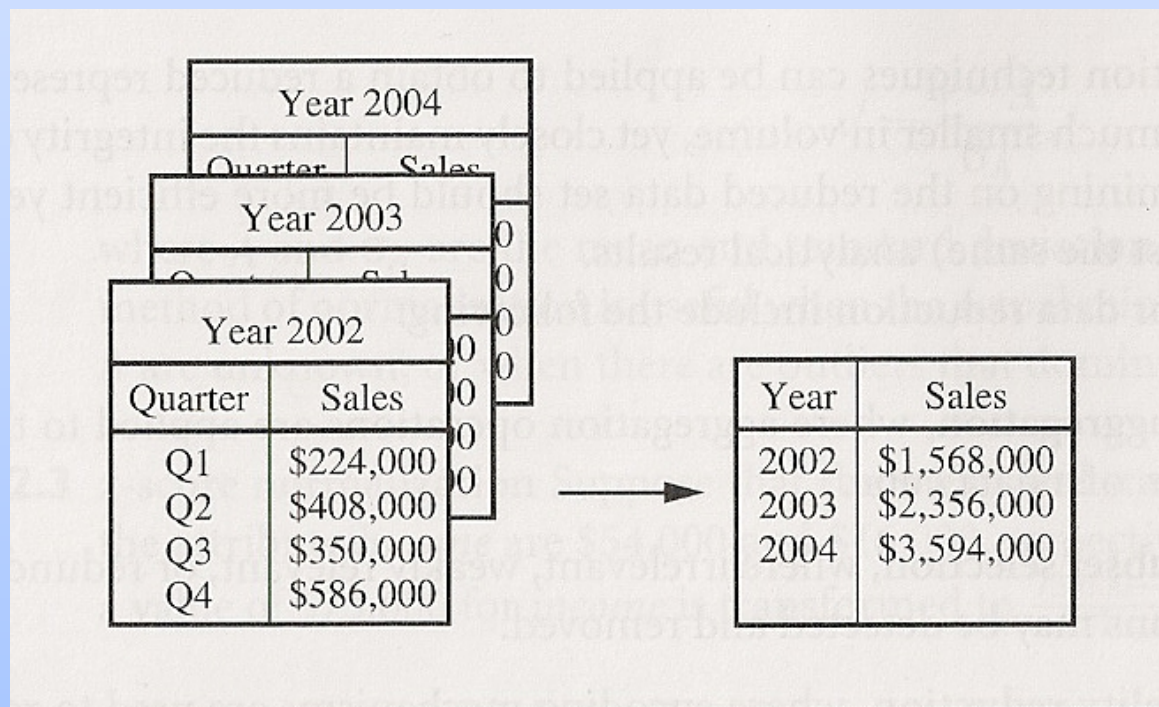
Redução dos Dados

- Obtém uma representação reduzida do conjunto de dados que é muito menor no volume, mas que produz os mesmos (ou quase) resultados analíticos.
 - Em muitos casos, *datasets* possuem um **número elevado** de atributos e de observações (**objetos**).
 - Para que reduzir dados?
 - a análise de dados complexos pode levar muito tempo para se obter uma solução (complexidade computacional muito alta)
 - algoritmos podem não rodar de forma satisfatória
 - Vantagens
 - redução do tempo de aprendizagem
 - interpretação mais fácil dos conceitos aprendidos
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

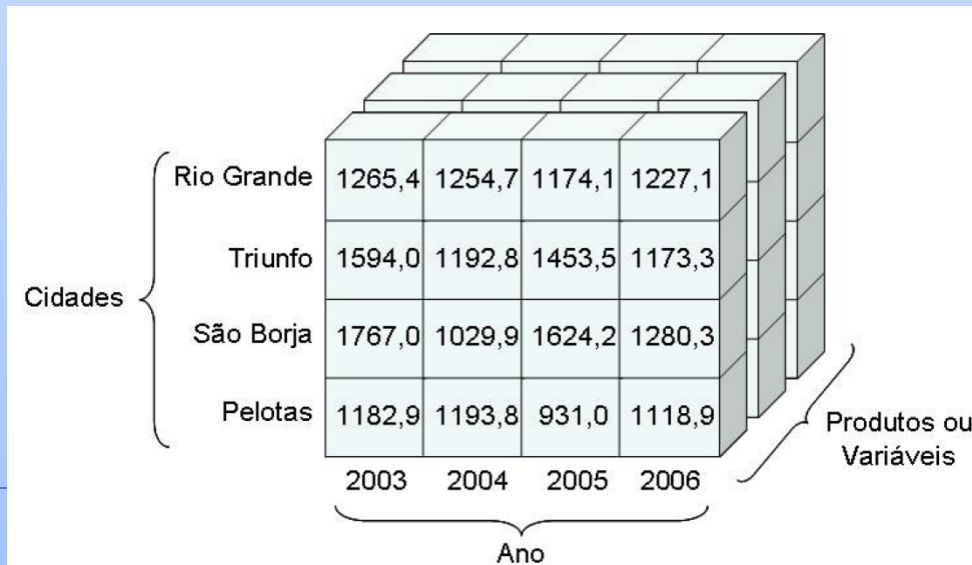
- Agregação via cubo:



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Agregação via cubo:**
 - forma de visualização e interpretação dos dados no modelo multidimensional para dados acumulados de chuva nos anos de 2003 a 2006, em algumas cidades do Rio Grande do Sul.



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Redução da dimensão:**
 - Em data mining a supressão de uma coluna (atributo) é muito mais delicada do que a supressão de uma linha (observação).
 - Retirar atributos relevantes ou permanecer com atributos irrelevantes pode implicar na descoberta de padrões de baixa qualidade.
 - Pode ser prejudicial para o algoritmo de mineração empregado
 - Daí a necessidade de um estágio de seleção de atributos.
 - Abordagens para a redução de dimensionalidade
 - Seleção manual baseada em conhecimento especialista.
 - Transformada Wavelet
 - Principal Componente Analysis (PCA)
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Busca heurística em Seleção de Atributos**
 - Problema: busca exaustiva - $2^n - 1$ possíveis combinações de n atributos
 - Heurística típica para métodos de seleção
 - Melhor atributo supondo independência entre atributos
 - Seleção dos melhores atributos passo a passo:
 - Escolha o melhor atributo primeiro
 - Após, escolha o melhor atributo condicionado a escolha do primeiro, ...
 - Eliminação de atributos passo a passo:
 - Repetidamente eliminar o pior atributo
 - Melhor seleção de atributos combinando seleção e eliminação
 - Uso de eliminação de atributos e backtracking
 - Ex: Algoritmo de construção de árvores de decisão
 - » Aplicar esse algoritmo nos dados completos e então selecionar apenas as variáveis presentes na árvore de decisão.

Redução dos Dados

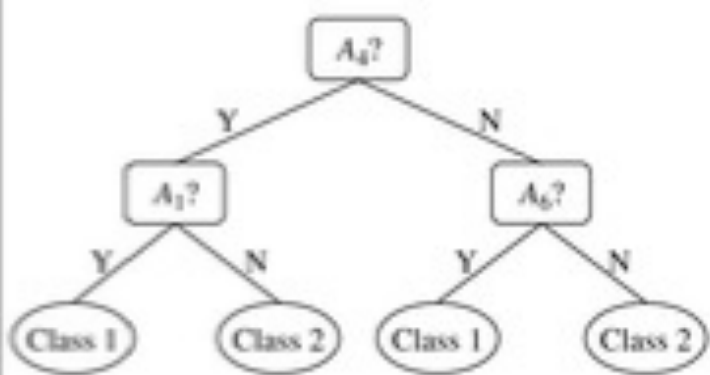
agregação via cubo - redução da dimensão - redução dos casos

- **Redução da dimensão:**
 - Seleção do menor (sub)conjunto de atributos:
Selecionar o menor conjunto de atributos suficiente para dividir o espaço das instâncias de tal maneira que a distribuição das classes no novo espaço é tão próxima quanto possível daquela do espaço original
 - A mineração em um conjunto reduzido de atributos possibilita reduzir o número de atributos em que os padrões aparecem
 - Facilitando a visualização dos padrões.
 - Em ML esse processo é conhecido como **seleção de subconjuntos de características**
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- Busca heurística em Seleção de Atributos

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1((Class 1)) A1 -- N --> C2_1((Class 2)) A6 -- Y --> C1_2((Class 1)) A6 -- N --> C2_2((Class 2)) </pre> \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- Redução (sintetização) do volume de dados via representação *econômica* dos mesmos.
 - Métodos Paramétricos:
 - Supõe que os dados ajustam um modelo, estimam os parâmetros do modelo, armazena apenas os parâmetros e descarrega os dados (exceto os aberrantes).
 - Principais modelos: regressão (simples e múltipla) e modelo log-linear
 - Métodos não Paramétricos:
 - Não assume modelos
 - Famílias principais: histogramas, clustering, amostragem
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

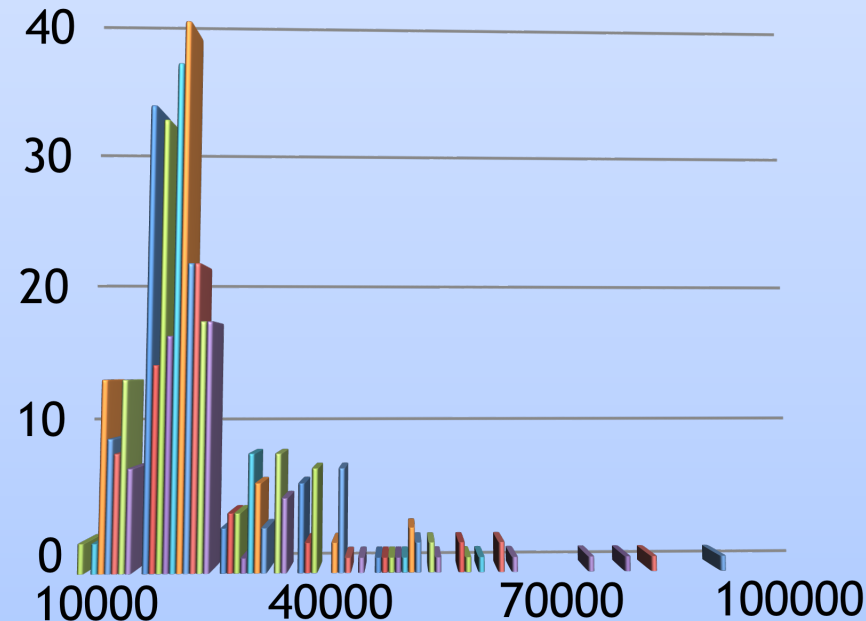
- **Regressão e Modelo Log-Linear**
 - Regressão linear: Os dados são modelados para estabelecer uma equação linear - relacionamento entre duas variáveis
 - Em geral usa-se o método dos mínimos quadrados para ajustar a linha
 - Regressão múltipla: permite que uma variável resposta (Y) seja modelada como uma função linear de um vetor de atributos
 - Modelo Log-linear : aproxima distribuições de probabilidade discretas multidimensionais
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Histogramas**

- Técnica popular para redução de dados
- Particiona os dados em caixas (classes) e armazena a frequência média dos valores.
- Em uma dimensão pode ser construído pela otimização de um critério via programação dinâmica.



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

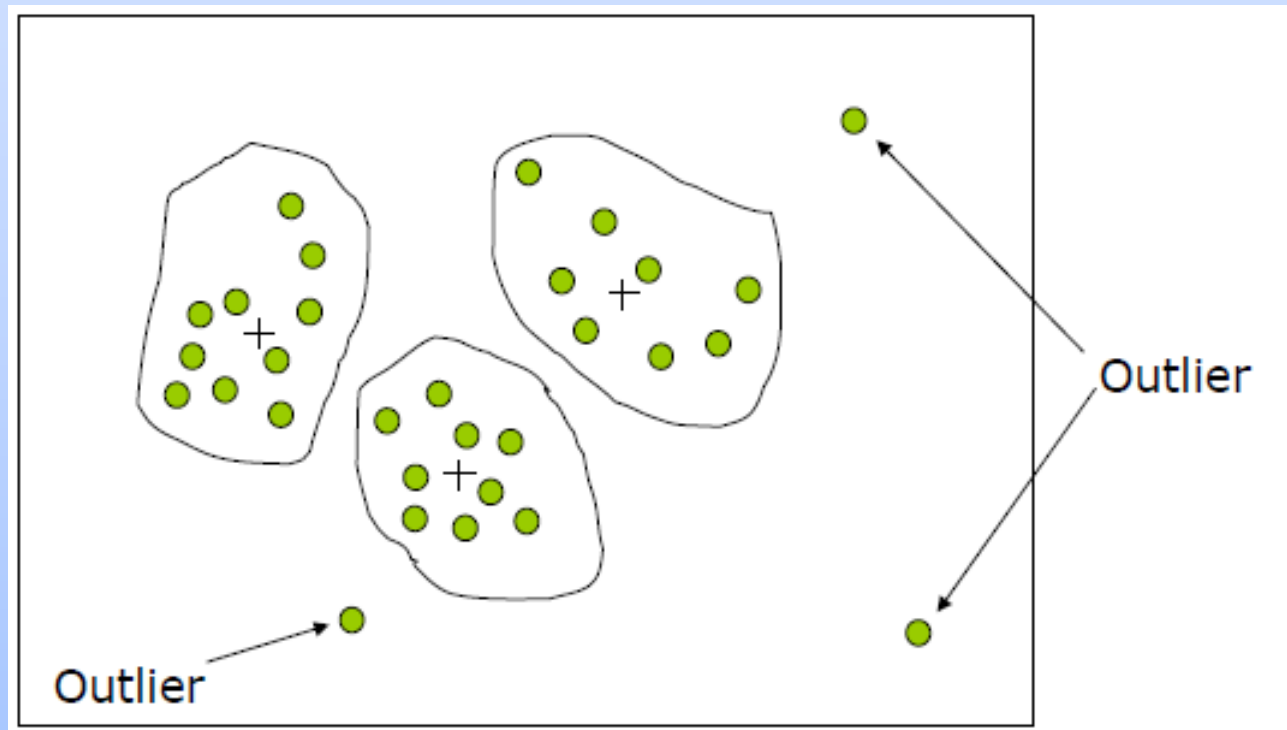
- **Clustering**

- Os dados são particionados em clusters e armazena-se apenas a representação (centróides) do mesmo.
 - Os representantes são os centróides e os outliers.
 - Pode ser muito eficaz se os dados são agrupados, mas não se estão apenas sujos.
 - A eficácia depende da distribuição dos dados
 - Existem muitas opções de métodos de e algoritmos de agrupamento.
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- Clustering



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Amostragem**

- Permite que os algoritmos de mineração tratem enormes bases de dados pela redução dos casos.
 - É geralmente usada em investigações preliminares de dados e também na análise final dos dados.
 - Estatísticos usam bastante as técnicas de amostragem porque trabalhar com o conjunto de dados completo é caro e computacionalmente custoso.
 - Amostragem pode ser usada em mineração de dados, quando o conjunto de dados, sob análise, é grande (em termos de objetos e atributos).
 - Tipos de Amostragem:
 - Amostragem aleatória simples com reposição
 - Amostragem aleatória simples sem reposição
 - Amostragem estratificada
 - Amostragem por conglomerado
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Princípio da Amostragem**
 - Uma amostra produzirá resultados de qualidade semelhantes aqueles produzidos pelo conjunto de dados completos (se a amostra for representativa).
 - Uma amostra é representativa se ela tem aproximadamente as mesmas propriedades (de interesse) do conjunto de dados original.
-

Redução dos Dados

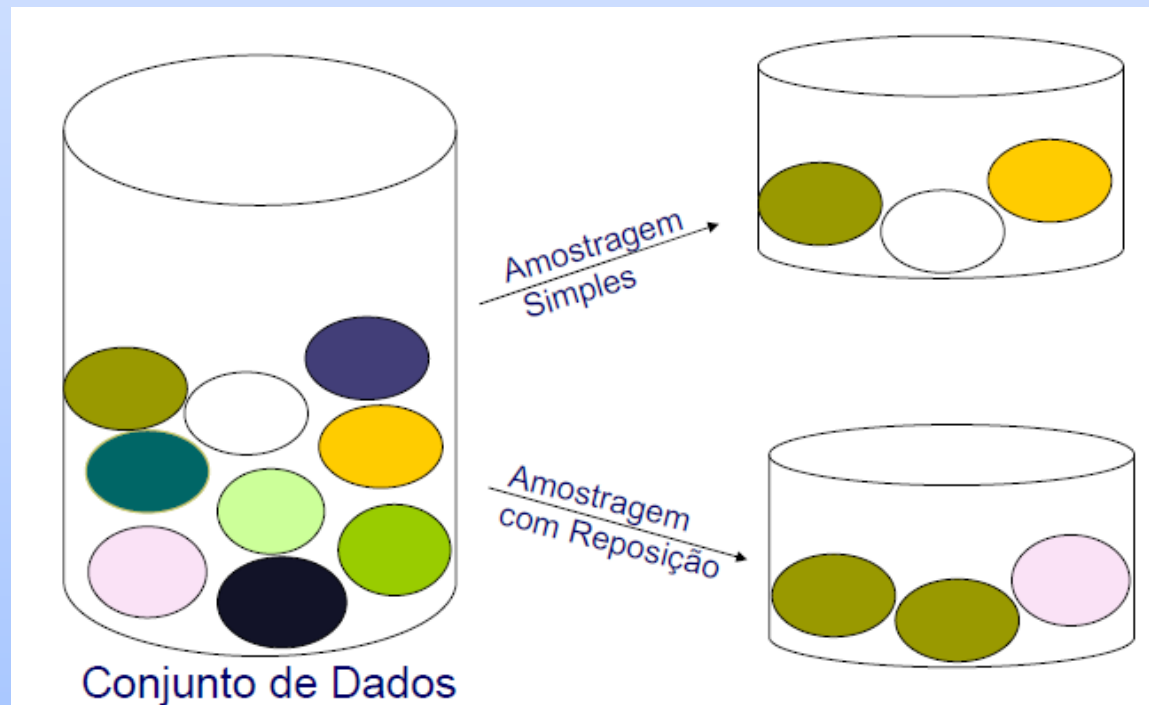
agregação via cubo - redução da dimensão - redução dos casos

- **Tipos de Amostragem**
 - Amostragem aleatória simples sem reposição
 - Existe uma probabilidade igual para a seleção de qualquer item.
 - Um item é selecionado e removido da população
 - Amostragem aleatória simples com reposição
 - Objetos não são removidos da população à medida em que são selecionados para a amostra.
 - O mesmo objeto pode ser selecionado mais de uma vez.
 - Amostragem estratificada (por conglomerado)
 - Separa os dados em diversas partições (estratos). Toma-se de cada partição uma amostra percentual igual a porcentagem do estrato em relação a população.
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

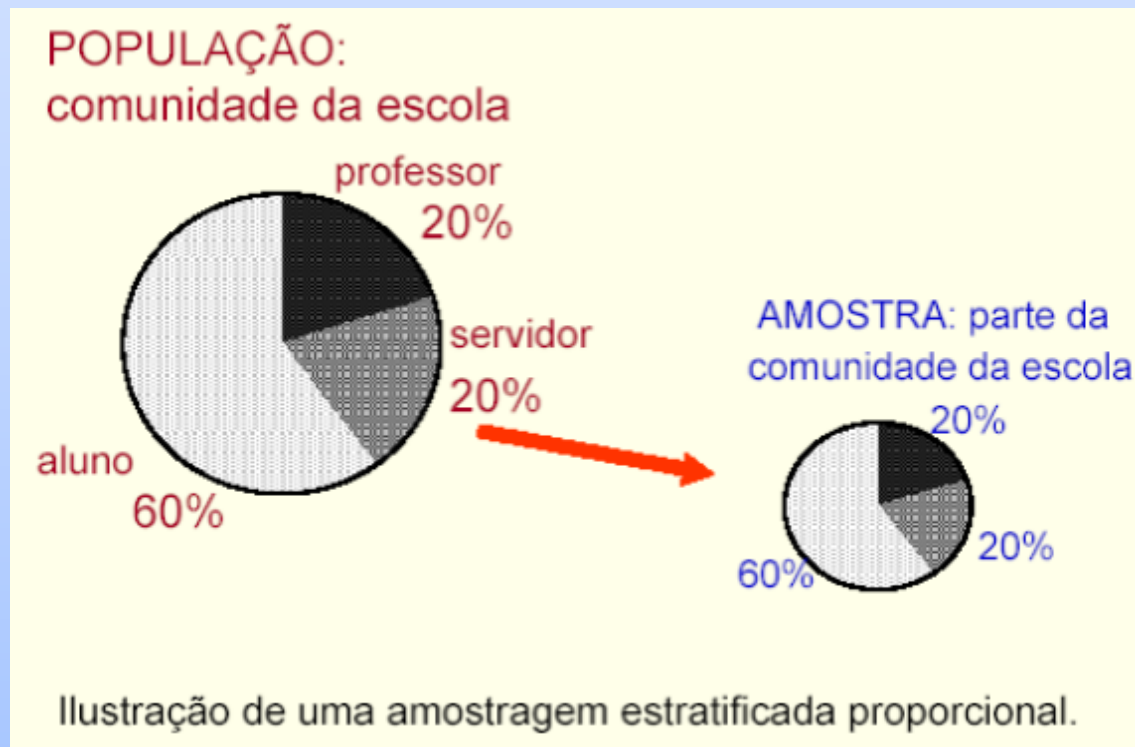
- Amostragem Simples e c/ Reposição



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

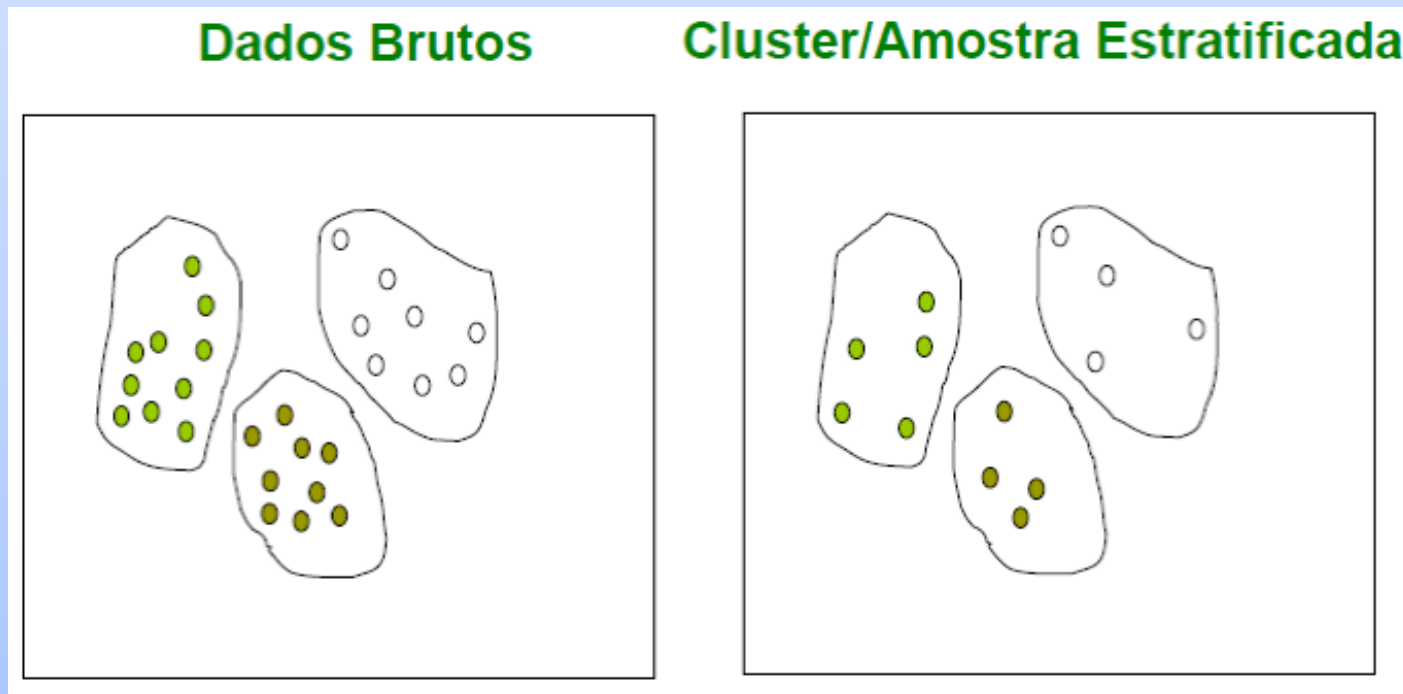
- Amostragem Estratificada



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- Amostragem Estratificada



Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Amostragem**
 - Duas formas básicas de amostragem são interessantes no contexto da mineração de dados:
 - Amostragens incrementais
 - Amostragens seguida de voto

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Amostragem incremental**

- O treinamento é realizado em amostras aleatórias cada vez maiores de casos, observar a tendência e parar quando não há mais progresso.

Um padrão típico de tamanhos de amostras pode ser 10%, 20%, 33%, 50%, 67% e 100%

- Critérios para passar para uma outra amostra
 - O erro diminuiu?
 - A complexidade do tratamento aumentou mais do que a queda da taxa de erro?
 - A complexidade da solução atual é aceitável para a interpretação?
-

Redução dos Dados

agregação via cubo - redução da dimensão - redução dos casos

- **Amostragem seguida de voto**
 - Interesse: quando o método de mineração suporta apenas N casos.
 - O mesmo método de mineração é aplicado para diferentes amostras de mesmo tamanho resultando em uma solução para cada amostra.
 - Quando um novo caso aparece, cada solução fornece uma resposta.
 - A resposta final é obtida por votação (classificação) ou pela média (regressão).
-

Pré-processamento usando o WEKA

- Preencher valores ausentes
 - `weka.filters.unsupervised.attribute.ReplaceMissingValues`
 - Remover instâncias com dados ausentes
 - `weka.filters.unsupervised.instances.RemoveWithValues`
 - Converter dados nominais para binário
 - `weka.filters.unsupervised.attribute.NominalToBinary`
 - Normalização de dados numéricos
 - `weka.filters.unsupervised.attribute.Normalize`
 - Renomear valores nominais de atributos (weka 3.8)
 - `weka.filters.unsupervised.attribute.RenameNominalValues`
-

Pré-processamento via Python

- Para realização de re-escala, padronização, normalização e binarização
 - Scripts: pre-processing1.py, pre-processing2.py, pre-processing3py e pre-processing4.py
 - Para tratamento de dados ausentes
 - pre-processing5.py
 - Convertendo colunas com valores nominais para valores binários
 - pre-processing6.py
-