KDD E MINERAÇÃO DE DADOS

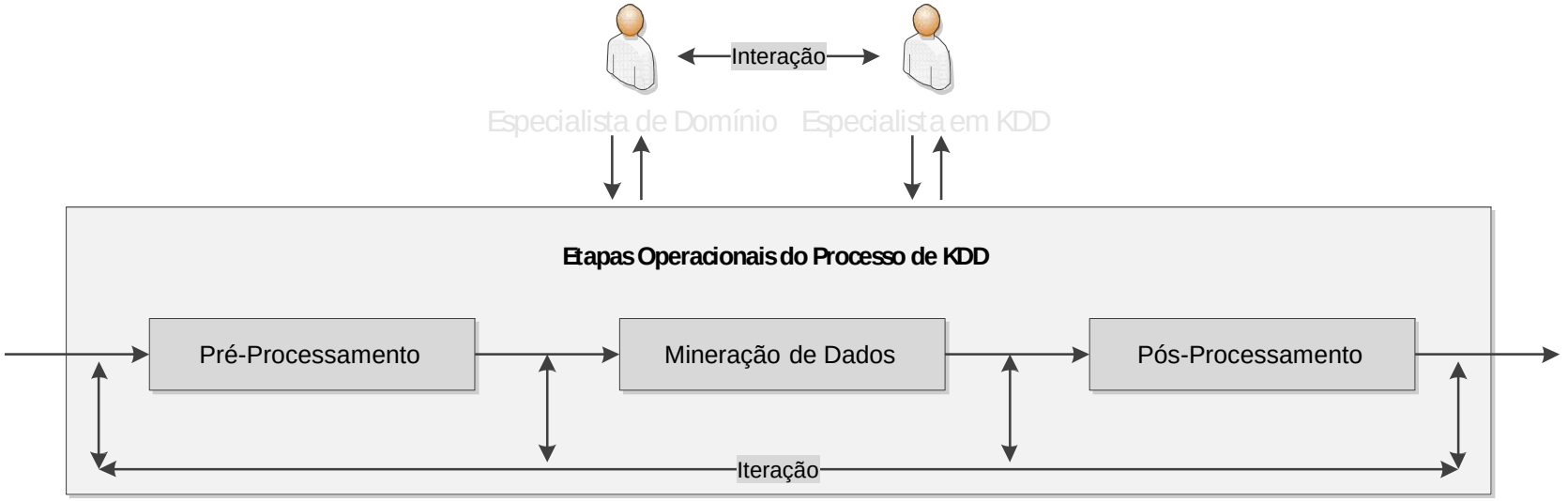
Etapas do Processo de KDD

Prof. Paulo Mello

ETAPAS DO PROCESSO DE KDD

Processo de Descoberta do Conhecimento em Bases de Dados

- **Visão Pragmática [Goldschmidt et al., 2002a]:**



- **Operações e Métodos de KDD**

ETAPAS DO PROCESSO DE KDD

Pré-Processamento dos Dados

Compreende as ações para adequar os dados aos algoritmos de mineração de dados a serem aplicados.

Tipos de Variáveis/Atributos/Características

- Nominais ou Categóricas
- Discretas
- Contínuas

ETAPAS DO PROCESSO DE KDD

Tipos de Variáveis/Atributos: Nominais ou Categóricas

- Utilizadas para nomear ou atribuir rótulos a objetos.
- Valores: conjunto finito e pequeno de estados.
- Não há ordem entre os valores.
- Podem ser representados por tipos de dados alfanuméricos.

Exemplo:

Estado Civil:
Solteiro, Casado, Viúvo, Divorciado, etc...

ETAPAS DO PROCESSO DE KDD

Tipos de Variáveis/Atributos: Discretas

- Similar às variáveis nominais.
- Há ordem (com significado) entre os valores.
- Podem ser representados por tipos alfanuméricos.

Exemplos:

Dia da Semana: Segunda-Feira, Terça-Feira, Quarta-Feira, etc...,
Faixa Etária,
Faixa de Renda, etc...

ETAPAS DO PROCESSO DE KDD

Tipos de Variáveis/Atributos: Contínuas

- Variáveis Quantitativas.
- Valores: conjuntos finitos ou infinitos, ordenados.
- Tipos de dados numéricos.

Exemplos:

Renda, Idade, Altura, Etc...

ETAPAS DO PROCESSO DE KDD

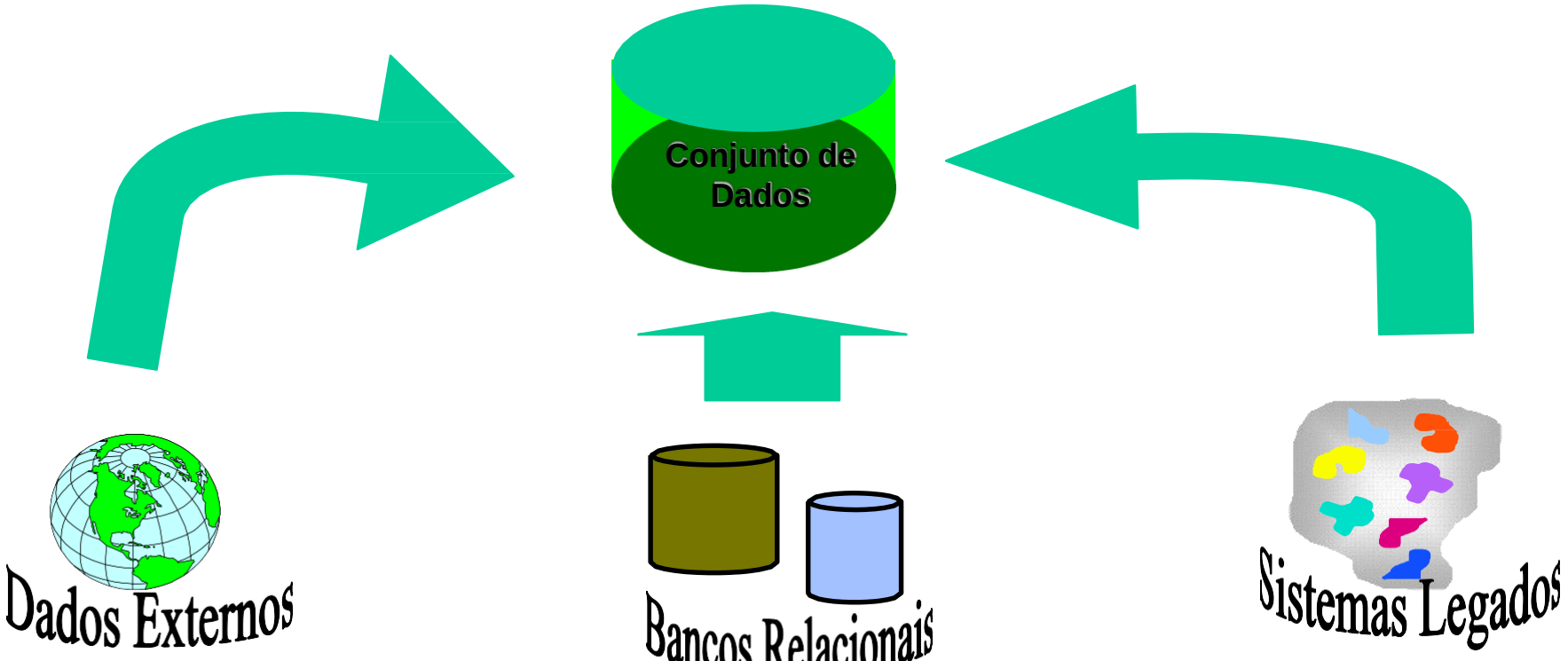
Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

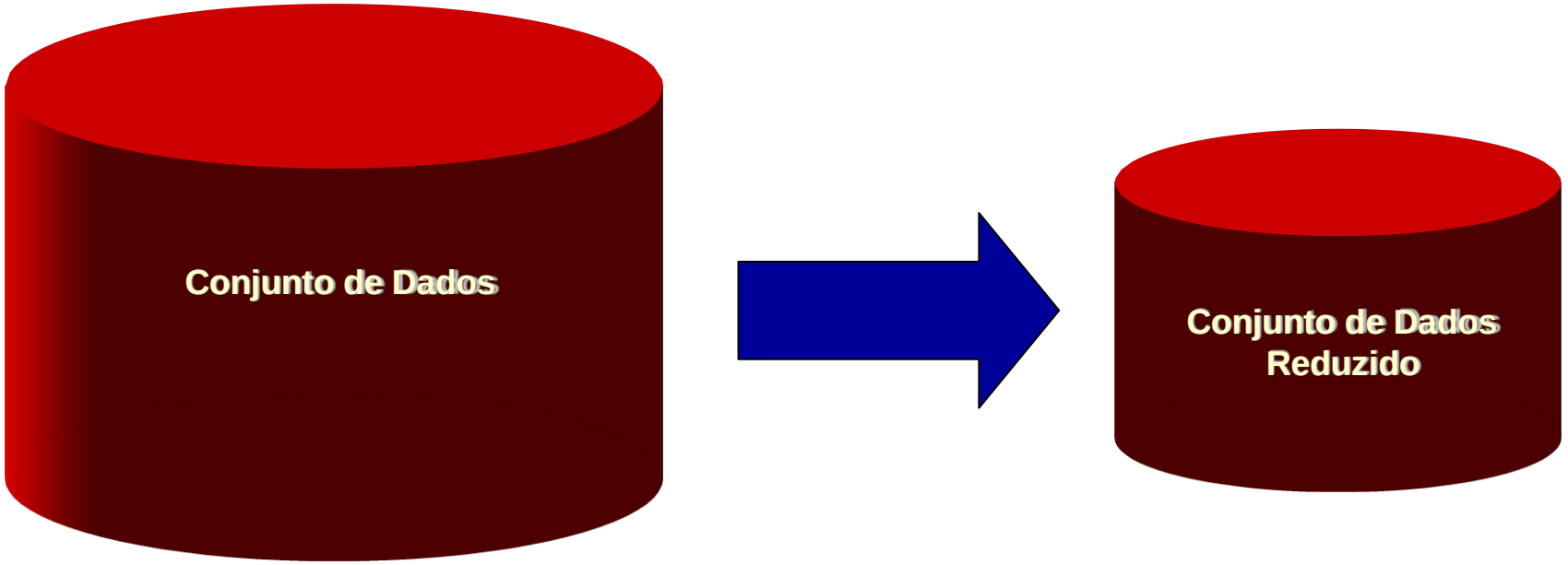
Definição 1: Extração dos dados de diversas fontes e carga no conjunto de dados a ser analisado.



ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

Definição 2: Escolha dentre os dados de um conjunto de dados, quais deverão ser efetivamente considerados na análise.



ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

Definição 3: Ocorre em dois momentos:

1º. Extração dos dados de diversas fontes e carga no conjunto de dados a ser analisado

2º. Escolha dentre os dados de um conjunto de dados, quais deverão ser efetivamente considerados na análise

Combinação das definições anteriores

Bastante comum na prática

Cópia do Conjunto de Dados em quaisquer das definições

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

A cópia dos dados em um conjunto separado dos dados transacionais → Evita interferência tanto no cotidiano quanto nas ações de mineração.

Consideremos que os dados estejam organizados em um mesmo conjunto: estrutura tabular bidimensional.
Assim, adotaremos a definição 2 para seleção de dados:

Definição 2: Escolha dentre os dados de um conjunto de dados, quais deverão ser efetivamente considerados na análise.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Segmentação do Banco de Dados**

Exemplo: Analisar somente clientes com residência própria

```
SELECT *
FROM CLIENTE
WHERE TP_RES=‘P’;
```

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Eliminação Direta de Casos (variação da anterior)**

Exemplo: Analisar somente clientes com residência própria

```
DELETE FROM CLIENTE
WHERE TP_RES<>'P';
```

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas (n < N).

- **Amostragem Aleatória Simples sem Reposição**
  - Probabilidade de seleção: *1/N, que aumenta ao longo do processo*.
  - Todos os registros selecionados são excluídos do conjunto de dados original para evitar nova seleção.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas (n < N).

– **Amostragem Aleatória Simples com Reposição**

- Probabilidade de seleção: *1/N*.
- Todos os registros selecionados são mantidos no conjunto de dados original podendo sofrer nova seleção.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas (n < N).

- **Amostragem de Clusters**
  - Registros devem estar agrupados em M clusters (grupos).
  - Cada repetição do processo de amostragem pressupõe a escolha do cluster.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas (n < N).

- **Amostragem Estratificada**
  - Registros devem estar separados em grupos distintos, previamente definidos segundo algum critério.
  - O processo de amostragem deve ocorrer dentro de cada grupo.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
  - **Agregação de Informações**

  - Dados em um maior nível de detalhe são agrupados em um nível mais consolidado.

  Exemplo: Somar compras por cliente por período.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos**

Sendo ***S*** um conjunto de dados com atributos A_1, A_2, A_3, ..., A_n, o problema da **redução de dados vertical** consiste em **identificar** qual das 2^n-1 combinações entre tais **atributos** deve ser considerada no processo de descoberta de conhecimento.

Tem como **objetivo** procurar **encontrar** um **conjunto mínimo de atributos** de tal forma que a **informação original** seja ao **máximo preservada**.

Quanto maior o valor de *n*, maior o desafio na escolha dos atributos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Motivações:**

✓ Um conjunto de atributos bem selecionado pode conduzir a modelos de conhecimento mais concisos e com maior precisão.

✓ A eliminação de um atributo é muito mais significativa em termos de redução do tamanho de um conjunto de dados do que a exclusão de um registro.

ETAPAS DO PROCESSO DE KDD

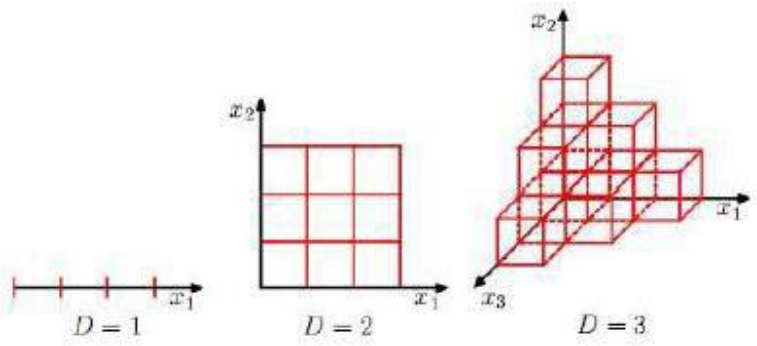
Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Motivações:**
  - ✓ Se o método de seleção for rápido, o tempo de processamento necessário para utilizá-lo e, em seguida, aplicar o algoritmo de mineração de dados em um subconjunto dos atributos, pode ser inferior ao tempo de processamento para aplicar o algoritmo de mineração sobre todo o conjunto de atributos;
  - ✓ Atributos altamente correlacionados podem trazer pouco ou nenhum ganho discriminatório (e, portanto, não precisam ser utilizados em conjunto para representar a informação original).

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Motivações:**
  - ✓ Permite lidar com a **maldição da (alta) dimensionalidade**: expressão que se refere ao problema causado pelo aumento exponencial no volume associado com a adição de dimensões extras a um espaço matemático;

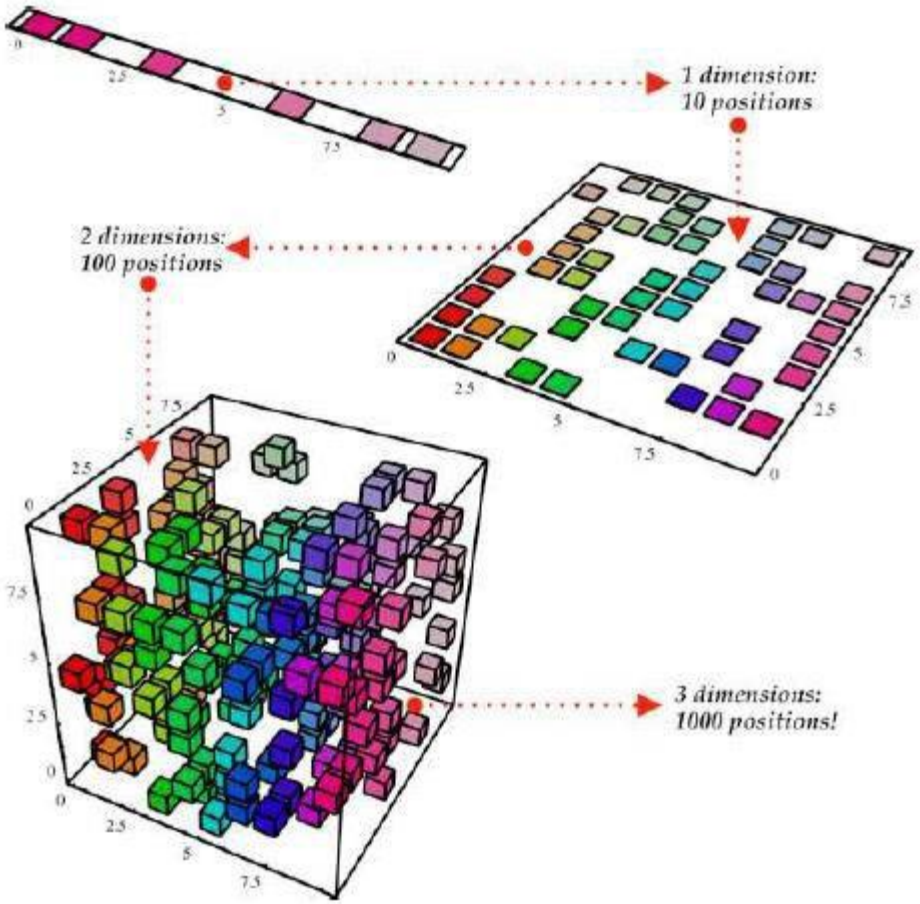


Se dividirmos o espaço em células regulares, o número de células cresce exponencialmente com a dimensão do espaço.

Assim, o número de amostras deve crescer para garantir que nenhuma célula fique vazia.

ETAPAS DO PROCESSO DE KDD

Maldição da (Alta) Dimensionalidade



Dado um tamanho de amostras, existe um número máximo de características a partir do qual o desempenho dos algoritmos de mineração de dados irão degradar ao invés de melhorar.

Solução: reduzir a dimensão do espaço através de métodos de redução de dimensionalidade (seleção ou projeção de características).

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Objetivos:**

✓ Eliminar atributos redundantes ou irrelevantes.

✓ Obter uma representação reduzida em volume mas que produz resultados de análise idênticos ou similares aos obtidos com o conjunto completo de atributos.

✓ Melhorar o desempenho dos modelos de aprendizado.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Manual x Automática:**

  ✓ Manual: requer entendimento profundo sobre o problema de aprendizado e sobre o significado de cada atributo.

  ✓ Automática: utiliza abordagens e estratégias computacionais voltadas à identificação do subconjunto de atributos a ser utilizado no problema.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Seleção x Projeção:**

  ✓ Seleção – Escolha dentre os atributos de um conjunto de dados. Não altera a natureza de cada atributo selecionado.

  ✓ Projeção – Transforma os dados originais, criando novos atributos para descrever o conjunto.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

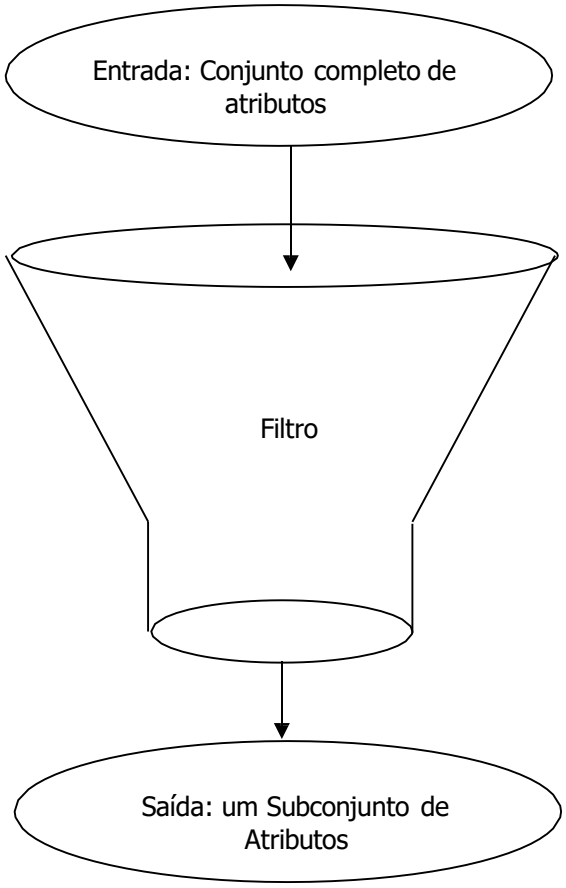
- **Escolha de características/atributos – Abordagens segundo Freitas (2002):**

✓ Independente de Modelo (***Filter***): Filtro

✓ Dependente de Modelo (***Wrapper***): Empacotamento

ETAPAS DO PROCESSO DE KDD

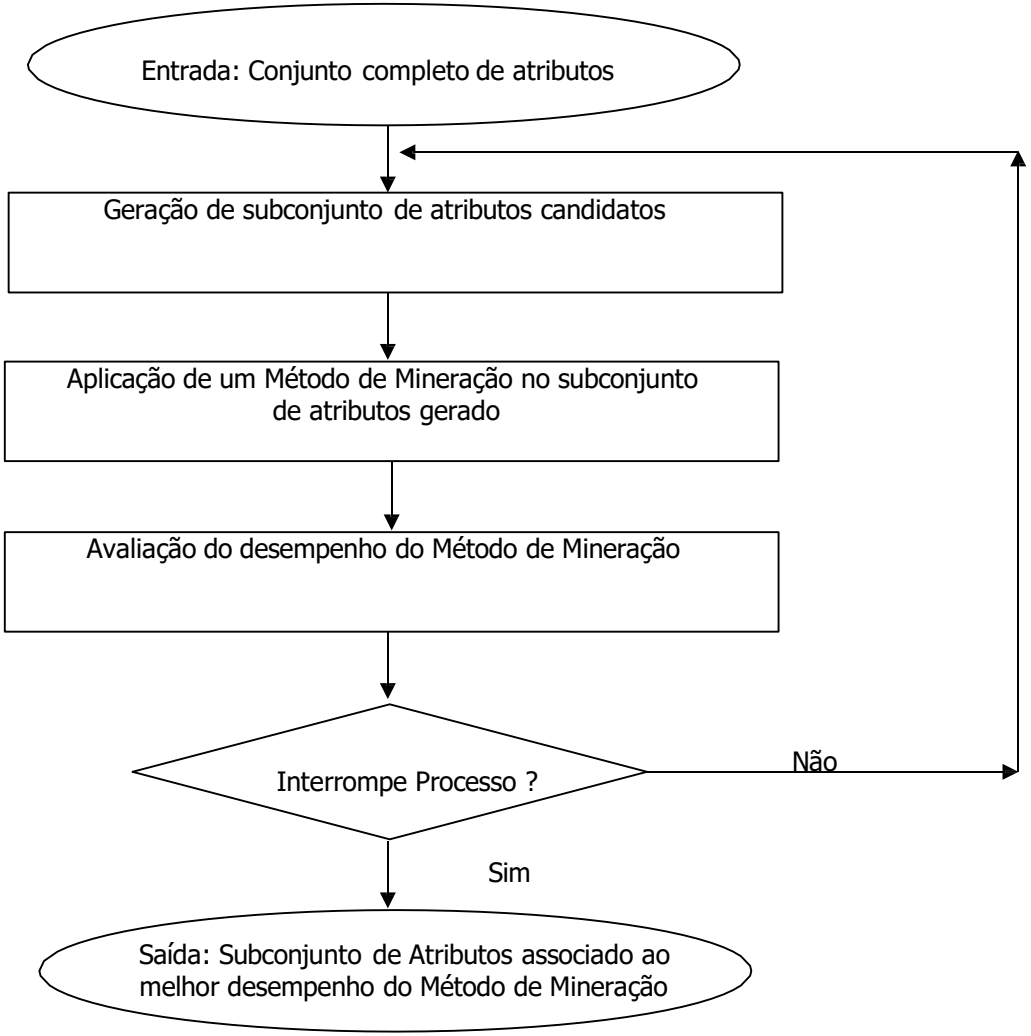
Seleção/Redução de Dados Vertical – Abordagens



Abordagem Independente de Modelo (***Filter***):
A seleção de atributos é realizada sem considerar o algoritmo de mineração de dados que será aplicado.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical – Abordagens



Abordagem Dependente de Modelo (***Wrapper***):
Consiste em experimentar o algoritmo de mineração de dados para cada conjunto de atributos e avaliar os resultados obtidos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical - Filtros

- **Vantagens:**

✓ Execução rápida.

✓ Por avaliarem características intrínsecas aos dados, seus resultados exibem maior generalidade.

- **Desvantagens:**

✓ Tende a selecionar subconjuntos de variáveis com muitos elementos.

✓ Força a intervenção do analista na escolha das variáveis.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical - Empacotadores

- **Vantagens:**

✓ Levam a modelos mais precisos.

✓ Levam a modelos com boa capacidade de generalização.

- **Desvantagens:**

✓ Execução lenta.

✓ Solução muito associada ao método de mineração utilizado.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Tipos de Estratégias:**

  ✓ Sequenciais

  ✓ Exponenciais

  ✓ Aleatórias / Estocásticas

Podem ser utilizadas em quaisquer das abordagens citadas

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Estratégias Sequenciais – Exemplos:**

✓ Seleção Sequencial para Frente (*Forward Selection*)

✓ Seleção Sequencial para Trás (*Backward Selection*)

✓ Combinação das Anteriores (*Mixed Selection*)

Estas estratégias adicionam e/ou removem características sequencialmente. Possuem tendência em se prender em mínimos locais.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Estratégias Sequenciais:**

✓ **Seleção Sequencial para Frente (*Forward Selection*)**

- Subconjunto de atributos candidatos começa vazio.
- O processo é iterativo.
- Cada atributo é adicionado ao subconjunto de candidatos
- O subconjunto de candidatos é avaliado segundo medida de qualidade.
- Ao final de cada iteração, é incluído no subconjunto de atributos candidatos, aquele atributo que tenha maximizado a medida de qualidade considerada

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Estratégias Sequenciais:**

✓ **Seleção Sequencial para Trás (*Backward Selection*)**

- Contrário da Seleção Sequencial para Frente.
- Subconjunto de candidatos começa completo.
- Cada atributo é retirado do subconjunto, que é avaliado segundo alguma medida de qualidade.
- Ao final de cada iteração, é excluído do subconjunto de candidatos, aquele atributo que tenha minimizado a medida de qualidade.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Estratégias Sequenciais:**

✓ **Combinação Anteriores (*Mixed Selection*)**

- A seleção para frente e a seleção para trás são combinadas.
- A cada iteração, o algoritmo seleciona o melhor atributo (incluindo-o no subconjunto de atributos candidatos) e remove o pior atributo dentre os remanescentes do conjunto de atributos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Estratégias Exponenciais – Exemplos:**

✓ Busca Exaustiva

✓ Branch and Bound

✓ Beam Search

Estas estratégias avaliam um número de subconjuntos que crescem exponencialmente com a dimensão do espaço de busca do conjunto de dados.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Estratégias Aleatórias / Estocásticas – Exemplos:**

✓ Simulated Annealing

✓ Algoritmos Genéticos

✓ Otimização por Colônia de Formigas

Estas estratégias incorporam aleatoriedade em seus procedimentos de busca a fim de escapar de mínimos locais.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

- **Observações:**

✓ Pré-processamento recomendado: remoção de *outliers* e normalização de cada atributo.

✓ Exemplo de Medida de Qualidade – Taxa de Inconsistência

Sexo	Est_Civil	Result	Count(*)
M	C	A	2
M	C	I	1
F	S	A	3
M	S	I	1
F	C	I	2
M	V	A	1

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Eliminação Direta de Atributos (Filtro)

Consiste da escolha dos atributos desejados, desprezando os demais.

Depende do conhecimento prévio sobre o problema.

Heurísticas Importantes para eliminação de atributos:

a) Atributos com valores constantes → não adicionam informação.

b) Atributos que sejam identificadores → particularizam objetos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses (Filtro)

Aplicável em problemas de classificação.

Procura descartar características com pouca capacidade discriminatória individual. As características restantes devem ser combinadas e analisadas em conjunto com as demais.

Objetivo: para cada característica individual, verificar se os valores apresentados em diferentes classes diferem significativamente.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses (Filtro)

O tipo do teste varia em função da:

- Natureza dos atributos: contínua ou discreta.
- Quantidade de classes envolvidas: duas ou mais de duas
- Quantidade de amostras disponíveis: até 30 ou mais que 30.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses (Filtro)

Entrada	Saída	Qtde Classes	Qtde Amostras	Tipo Teste
Quantitativa	Qualitativa	2	Até 30	Student
Qualitativa	Quantitativa	2	Até 30	Student
Quantitativa	Qualitativa	> 2	-	ANOVA
Qualitativa	Quantitativa	> 2	-	ANOVA
Qualitativa	Qualitativa	-	-	Qui-Quadrado

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

- Pressupõe que a variável qualitativa (entrada ou saída) divide os valores disponíveis da variável quantitativa (saída ou entrada) em dois grupos.

- Cada grupo contém os valores quantitativos que estão associados a um dos valores qualitativos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

Exemplo: conjunto de dados contendo indicações climáticas para a realização de jogos de tênis.

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Variáveis:

- Saída: Jogar?
- Entrada (1): Temperatura
- Entrada (2): Umidade

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

Exemplo: conjunto de dados contendo indicações climáticas para a realização de jogos de tênis.

Variável de Entrada Temperatura:
Saída Não: 85, 80, 65, 72, 71 (média 74,6)
Saída Sim: 83, 70, 68, 64, 69, 75, 75, 72, 81 (média 73,0)

Variável de Entrada Umidade:
Saída Não: 85, 90, 70, 95, 91 (média 86,2)
Saída Sim: 86, 96, 80, 65, 70, 80, 70, 90, 75 (média 79,1)

Pergunta a ser respondida pelo teste (feito independentemente para cada variável): a diferença entre as médias é significativa?

Se for, o atributo deve ser selecionado. Caso contrário, deve ser descartado.

ETAPAS DO PROCESSO DE KDD

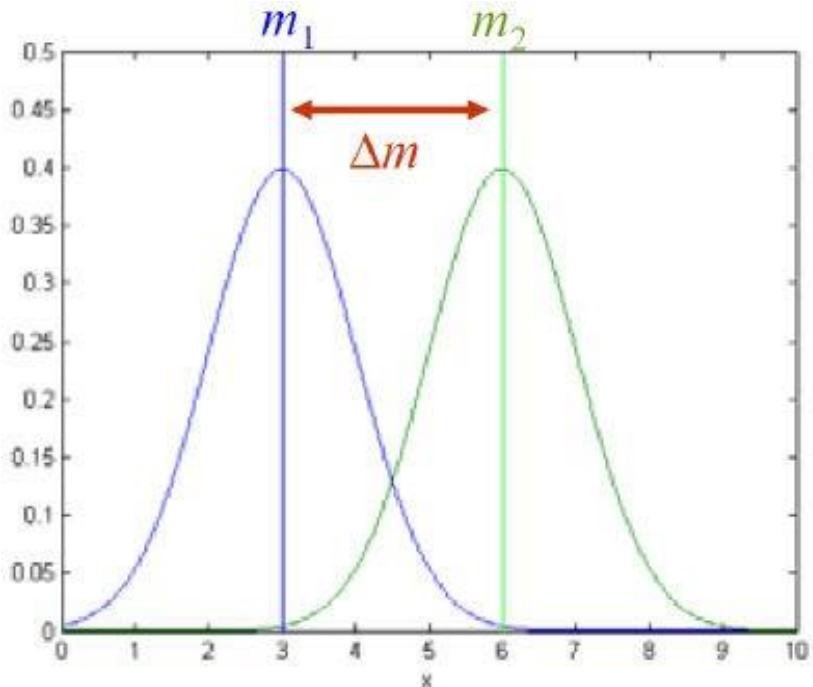
Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

Exemplo: conjunto de dados contendo indicações climáticas para a realização de jogos de tênis.

A diferença entre as médias é significativa?

O atributo influi na definição das classes?



ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

Premissas para aplicação do teste:

- Formação de duas V.A. distintas X_1 e X_2 (uma por classe)
- Sejam n_1 e n_2 a qtde de amostras de X_1 e X_2, respectivamente
- Sejam μ_1 e μ_2 as médias e σ_1^2 e σ_2^2 as variâncias respectivas
- X_1 e X_2 possuem distribuição normal

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

Hipóteses:

- H_0: $\mu_1 = \mu_2$ (hipótese nula)
- H_1: $\mu_1 \neq \mu_2$ (hipótese alternativa)
- Objetivo do teste: rejeitar a hipótese nula!

Estatística de Teste (Welch): $$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(Student): $$t = \frac{|\mu_1 - \mu_2|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Teste de Hipóteses de Student (Filtro)

Distribuição *t* de Student:

- Família de distribuições, definidas pelo número de graus de liberdade N

- MatLab:

```
x = -10:0.01:10;
p = tfdp(x, N);
plot(x,p)
```