

KDD E MINERAÇÃO DE DADOS

Tarefas de KDD

Prof. Ronaldo R. Goldschmidt

ronaldo@de9.ime.eb.br

rribeiro@univercidade.br

geocities.yahoo.com.br/ronaldo_goldschmidt

TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Deteção de Desvios

Sumarização

Mineração de Textos

TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Deteção de Desvios

Sumarização

Mineração de Textos

TAREFAS DE KDD

Regras de Associação – Caracterização Intuitiva:

“Consiste em encontrar **conjuntos de itens** que **ocorram simultaneamente de forma frequente** em um banco de dados.”

TAREFAS DE KDD

Regras de Associação – Exemplo de Aplicação:

“Encontrar produtos que sejam frequentemente vendidos de forma conjunta.”

N. Trans.	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	não	sim	não	sim	sim	não	não
2	sim	não	sim	sim	sim	não	não
3	não	sim	não	sim	sim	não	não
4	sim	sim	não	sim	sim	não	não
5	não	não	sim	não	não	não	não
6	não	não	não	não	sim	não	não
7	não	não	não	sim	não	não	não
8	não	não	não	não	não	não	sim
9	não	não	não	não	não	sim	sim
10	não	não	não	não	não	sim	não

TAREFAS DE KDD

Regras de Associação – Formato Basket:

Nº Transação	Item
1	Café
1	Pão
1	Manteiga
2	Leite
2	Cerveja
2	Pão
2	Manteiga
...	...

TAREFAS DE KDD

Regras de Associação – Algumas Definições:

Def: **Transação:** Elemento de ligação existente em cada ocorrência de itens no BD.

Def: **Regra de Associação:** $X \rightarrow Y$, onde X e Y são **itemsets** (conjuntos de itens) tais que $X \cap Y = \emptyset$.

Def: **Regra de Associação Frequente:** se $|X \wedge Y|/|D| \geq \text{minsup}$.

Def: **Regra de Associação Válida:** se $|X \wedge Y|/|X| \geq \text{minconf}$.

Def: **K-Itemset** é um itemset contendo exatamente k itens

TAREFAS DE KDD

Regras de Associação – Formalização:

“Consiste em encontrar **regras de associação frequentes e válidas** em um banco de dados, a partir da especificação dos parâmetros de suporte e confiança mínimos.”

Exemplos de Regras de Associação:

Leite $\rightarrow$ Açúcar

Pão $\wedge$ Manteiga $\rightarrow$ Café

TAREFAS DE KDD

ASSOCIAÇÃO

EXEMPLOS DE ALGORITMOS TRADICIONAIS:

- **APRIORI**
- **DHP – DIRECT HASHING AND PRUNING**
- **PARTITION**
- **DIC – DYNAMIC ITEMSET COUNTING**

TAREFAS DE KDD

ASSOCIAÇÃO

ALGORITMOS MAIS RECENTES [Zaki et al., 2002]:

- **CD – CANDIDATE DISTRIBUTION**
- **DD – DATA DISTRIBUTION**
- **ECLAT, CLIQUE**
- **MAX-ECLAT, MAX-CLIQUE**

TAREFAS DE KDD

Regras de Associação – Estrutura Comum:

- Identificação dos conjuntos de itens frequentes:

$$|X \wedge Y| / |D| \geq \text{MinSup (Suporte Mínimo)}$$

Maior custo computacional

- Identificação, dentre os conjuntos de itens frequentes, quais as regras válidas:

$$|X \wedge Y| / |X| \geq \text{MinConf (Confiança Mínima)}$$

TAREFAS DE KDD

Regras de Associação – Estrutura Comum:

Baseia-se na propriedade de anti-monotonicidade do suporte:

“Um k-itemset somente pode ser frequente se todos os seus (k-1)-subconjuntos forem frequentes”

TAREFAS DE KDD

Regras de Associação – Exemplo:

Considere o seguinte BD:

N. Trans.	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	não	sim	não	sim	sim	não	não
2	sim	não	sim	sim	sim	não	não
3	não	sim	não	sim	sim	não	não
4	sim	sim	não	sim	sim	não	não
5	não	não	sim	não	não	não	não
6	não	não	não	não	sim	não	não
7	não	não	não	sim	não	não	não
8	não	não	não	não	não	não	sim
9	não	não	não	não	não	sim	sim
10	não	não	não	não	não	sim	não

TAREFAS DE KDD

Regras de Associação – Exemplo:

Algumas Regras Descobertas:

- Regra: SE (café) ENTÃO (pão).
- Regra: SE (café) ENTÃO (manteiga).
- Regra: SE (pão) ENTÃO (manteiga).
- Regra: SE (manteiga) ENTÃO (pão).
- Regra: SE (café E pão) ENTÃO (manteiga).
- Regra: SE (café E manteiga) ENTÃO (pão).
- Regra: SE (café) ENTÃO (manteiga E pão).

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase I: Definir os valores de suporte e confiança mínimos:

MinSup = 0,3

MinConf = 0,8

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

1ª Iteração:

1 - Itemsets	Suportes
Leite	0,2
Café	0,3
Cerveja	0,2
Pão	0,5
Manteiga	0,5
Arroz	0,2
Feijão	0,2

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

1ª Iteração:

1 - Itemsets	Suportes
Leite	0,2
Café	0,3
Cerveja	0,2
Pão	0,5
Manteiga	0,5
Arroz	0,2
Feijão	0,2

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

2ª Iteração: Combinar os 1-itemsets identificados anteriormente

2 - Itemsets	Suportes
Café , Pão	0,3
Café , Manteiga	0,3
Pão , Manteiga	0,4

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

2ª Iteração: Combinar os 1-itemsets identificados anteriormente

2 - Itemsets	Suportes
Café , Pão	0,3
Café , Manteiga	0,3
Pão , Manteiga	0,4

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

3ª Iteração: Combinar os 2-itemsets identificados anteriormente

3 - Itemsets	Suportes
Café , Pão , Manteiga	0,3

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

3ª Iteração: Combinar os 2-itemsets identificados anteriormente

3 - Itemsets

Suportes

Café , Pão , Manteiga

0,3

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase II: Identificar os conjuntos de itens freqüentes:

Lista de **todos** os k-itemsets freqüentes obtidos ($K \geq 2$)

- Café e Pão,
- Café e Manteiga,
- Pão e Manteiga,
- Café e Pão e Manteiga

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase III: Identificação das Regras Válidas:

- **Conjunto de itens: {café, pão}.**
 - SE café ENTÃO pão. Conf = 1,0.
 - SE pão ENTÃO café. Conf = 0,6.
- **Conjunto de itens: {café, manteiga}.**
 - SE café ENTÃO manteiga. Conf = 1,0.
 - SE manteiga ENTÃO café. Conf = 0,6.
- **Conjunto de itens: {manteiga, pão}.**
 - SE manteiga ENTÃO pão. Conf = 0,8.
 - SE pão ENTÃO manteiga. Conf = 0,8.

TAREFAS DE KDD

Regras de Associação – Como obtê-las?

Fase III: Identificação das Regras Válidas:

- **Conjunto de itens: {café, manteiga, pão}.**
 - SE café, pão ENTÃO manteiga. Conf = 1,0.
 - SE café, manteiga ENTÃO pão. Conf = 1,0.
 - SE manteiga, pão ENTÃO café. Conf = 0,75.
 - SE café ENTÃO pão, manteiga. Conf = 1,0.
 - SE pão ENTÃO café, manteiga. Conf = 0,6.
 - SE manteiga ENTÃO café, pão. Conf = 0,6.

Finalmente, seleciona-se regras com Conf. maior ou igual ao valor mínimo especificado pelo usuário (**MinConf = 0,8**).

TAREFAS DE KDD

Regras de Associação – Regras Obtidas:

SE café ENTÃO pão.
SE café ENTÃO manteiga.
SE manteiga ENTÃO pão.
SE pão ENTÃO manteiga.
SE café,pão ENTÃO manteiga.
SE café, manteiga ENTÃO pão.
SE café ENTÃO pão, manteiga.

TAREFAS DE KDD

ASSOCIAÇÃO

EXEMPLOS DE APLICAÇÕES

- **MARKETING**
- **PESQUISAS CIENTÍFICAS – PADRÕES SIMULTÂNEOS**
- **CLASSIFICAÇÃO POR REGRAS DE ASSOCIAÇÃO**

TAREFAS DE KDD

Regras de Associação Generalizadas – Caracterização Intuitiva:

A descoberta de associações generalizadas é uma extensão da tarefa de descoberta de associações.

Sua compreensão depende da percepção de que é comum a existência de hierarquia e abstração entre conceitos.

Exemplo: Calça e camisa são tipos de roupa. Tênis e sapato são especializações do conceito calçado. Algumas regras:

camisa → sapato

roupa → sapato

camisa → calçado

roupa → calçado

TAREFAS DE KDD

Regras de Associação Generalizadas – Estratégias de Busca:

Independente do Nível de Abstração:

Consiste em percorrer todos os níveis da árvore de conceitos, sem utilizar conhecimento prévio acerca dos conjuntos de itens frequentes para eliminar alternativas de busca.

Esta estratégia demanda um maior volume de processamento.

TAREFAS DE KDD

Regras de Associação Generalizadas – Estratégias de Busca:

Máscara de Filtragem de um Item:

Um item do i -ésimo nível hierárquico de conceitos é analisado, se e somente se, o seu nó filho do $(i-1)$ -ésimo nível for freqüente.

Nesta abordagem, uma associação específica somente é analisada a partir de uma associação mais geral, que seja freqüente.

TAREFAS DE KDD

Regras de Associação Generalizadas – Estratégias de Busca:

Máscara de Filtragem de K-Itemsets:

Um K -Itemset do i -ésimo nível hierárquico de conceitos é analisado, se e somente se, seus nós filhos (K -Itemsets) do $(i-1)$ -ésimo nível forem freqüentes.

TAREFAS DE KDD

Associação

Descoberta de Seqüências

Classificação

Clusterização

Detecção de Desvios

Sumarização

Mineração de Textos

TAREFAS DE KDD

Descoberta de Seqüências – Caracterização Intuitiva:

- Extensão da Mineração de Associações: aspecto “temporal”.
- Regras de Associação: Padrões intra-transação
- Seqüências: Padrões inter-transação (mais complexa)

Exemplos de Aplicação:

- Histórico de itens comprados por consumidores ao longo de um período
- Histórico de acessos a páginas de um site pelos usuários da web.

TAREFAS DE KDD

Descoberta de Seqüências – Formalização:

“Consiste em encontrar **seqüências frequentes** em um banco de dados, a partir da especificação do parâmetro de suporte mínimo.”

Ex:

DATABASE		
CID	TID	Items
1	10	A B
	20	B
	30	A B
2	20	A C
	30	A B C
	50	B
3	10	A
	30	B
	40	A
4	30	A B
	40	A
	50	B

FREQUENT SET (75% Minimum Support)
(A, A->A, B->A, B, AB, A->B, B->B, AB->B)

TAREFAS DE KDD

Descoberta de Seqüências – Definições Relevantes:

Def: Seqüência: Lista ordenada de Itemsets. Caracterizada por objeto, rótulo temporal e eventos. Cada registro armazena ocorrências de eventos sobre um objeto em um instante de tempo particular. Notação: $\langle s_1 s_2 \dots s_n \rangle$, onde s_j é um itemset.

Exemplo:

Consumidores $\Leftrightarrow$ objetos

itens comprados $\Leftrightarrow$ eventos

Def: O itemset s_j é também chamado de **elemento da seqüência**. Cada elemento de uma seqüência é denotado por $(x_1, x_2, \dots, x_m)$, onde x_j é um item ou evento.

TAREFAS DE KDD

Descoberta de Sequências – Definições Relevantes:

Def: Uma sequência $\langle a_1 a_2 \dots a_n \rangle$ é uma **subsequência (ou especialização)** de outra sequência $\langle b_1 b_2 \dots b_n \rangle$ se existirem inteiros $i_1 < i_2 < \dots < i_n$ tais que $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ... e $a_n \subseteq b_{i_n}$.

Exemplo:

$\langle (3) (4, 5) (8) \rangle$ é uma subsequência de $\langle (7) (3, 8) (9) (4, 5, 6) (8) \rangle$, pois $(3) \subseteq (3, 8)$, $(4, 5) \subseteq (4, 5, 6)$ e $(8) \subseteq (8)$.

No entanto, a sequência $\langle (3) (5) \rangle$ não é uma subsequência de $\langle (3, 5) \rangle$ e vice versa.

TAREFAS DE KDD

Descoberta de Sequências – Definições Relevantes:

Def: O **suporte (ou frequência)** de uma sequência α refere-se ao número total de objetos que contêm α .

Def: Dado um limiar definido pelo usuário, denominado **suporte mínimo**, diz-se que uma **sequência** é **frequente** se esta ocorrer mais do que o suporte mínimo.

Def: Uma **k-sequência** é uma sequência com exatamente k elementos.

Def: Uma **sequência** é **maximal** se não for subsequência de nenhuma outra sequência.

TAREFAS DE KDD

Descoberta de Seqüências – Algoritmos Específicos:

- GSP – Generalized Sequential Patterns
- MSDD – Multi Stream Dependency Detection
- SPADE – Sequential Pattern Discovery using Equivalence Classes

Baseiam-se na propriedade de anti-monotonicidade do suporte:

“Uma k-seqüência somente pode ser freqüente se todas as suas (k-1)-subseqüências forem freqüentes”

TAREFAS DE KDD

Descoberta de Seqüências

ALGORITMOS MAIS RECENTES [Zaki et al., 2002]:

- **NPSPM – NON-PARTITIONED SEQUENTIAL PATTERN MINING**
 - **D-MSDD – DISTRIBUTED MSDD**
 - **pSPADE – PARALLEL SPADE**

TAREFAS DE KDD

Descoberta de Seqüências

EXEMPLOS DE APLICAÇÕES

- **MARKETING**
- **RE-ESTRUTURAÇÃO DE WEB SITES**

TAREFAS DE KDD

Seqüências Generalizadas – Caracterização Intuitiva:

A descoberta de seqüências generalizadas é uma extensão da tarefa de descoberta de seqüências.

Utiliza a hierarquia e a abstração entre conceitos eventualmente existentes em cada aplicação.

Exemplo: Calça e camisa são tipos de roupa. Tênis e sapato são especializações do conceito calçado. Exs. seqüências generalizadas:

<(roupa) (calçado)>
<(roupa) (sapato)>
<(camisa) (sapato)>
<(camisa, sapato)>
<(roupa, calçado)>

TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Detecção de Desvios

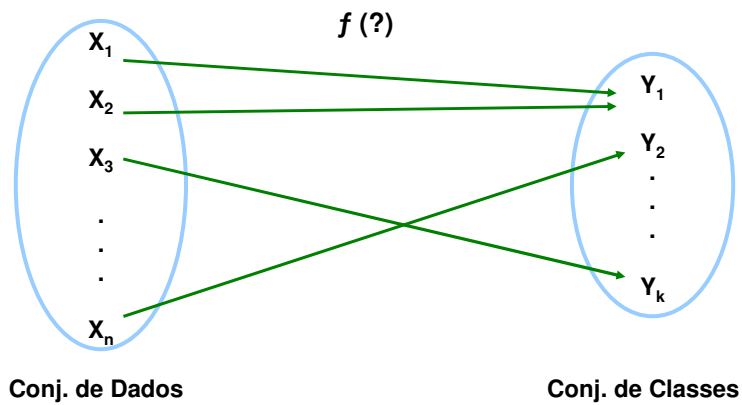
Sumarização

Mineração de Textos

TAREFAS DE KDD

Classificação – Formalização:

Caracterização do Problema:



TAREFAS DE KDD

Classificação – Formalização:

Objetivo:

$$\hat{f} \cong f$$

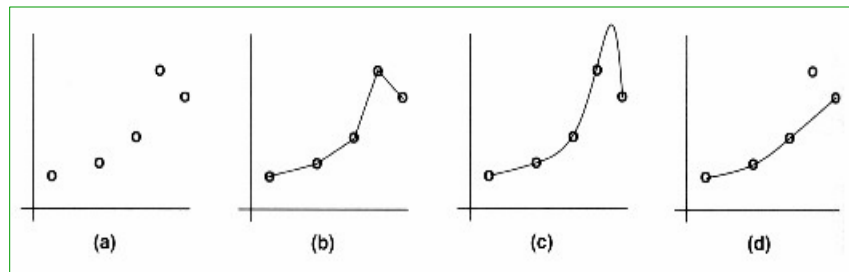


TAREFAS DE KDD

Classificação

EXEMPLOS DE HIPÓTESE

$$\hat{f} \cong f$$



TAREFAS DE KDD

Classificação – Formalização:

Nos casos em que a imagem de f é formada por rótulos de classes, a tarefa de inferência indutiva é denominada classificação e toda hipótese h chamada de classificador.

A identificação da função h consiste de um processo de busca no espaço de hipóteses H , pela função que mais se aproxime da função original f . Este processo é denominado *aprendizado* (Russell e Norvig, 1995).

Todo algoritmo que possa ser utilizado na execução do processo de aprendizado é chamado *algoritmo de aprendizado*.

TAREFAS DE KDD

Classificação – Formalização:

O conjunto de todas as hipóteses que podem ser obtidas por um algoritmo de aprendizado L é representado por H_L . Cada hipótese pertencente ao H_L é representada por h_L .

Acurácia da hipótese h : qualidade ou precisão de h em mapear corretamente cada vetor de entradas x em $f(x)$.

$$Acc(h) = 1 - Err(h)$$

$$Err(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{I} [y_i \neq h(i)]$$

TAREFAS DE KDD

Classificação – Formalização:

Conjunto de treinamento: $(x, f(x))$ utilizados na identificação da função h . Conjunto de testes: $(x, f(x))$ utilizados para avaliar a acurácia de h .

L é uma função $L: T \rightarrow H_L$, onde T é o espaço de todos os conjuntos de treinamento possíveis para L .

TAREFAS DE KDD

Classificação – Formalização:

Cada algoritmo possui um *bias* indutivo que direciona o processo de construção dos classificadores.

Bias indutivo: o conjunto de fatores que coletivamente influenciam na seleção de hipóteses [Utgoff, 1986].

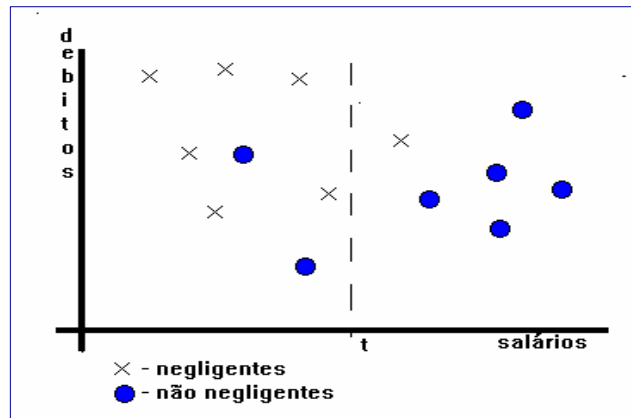
O *bias* de um algoritmo L afeta o processo de aprendizado de duas formas: restringe o tamanho do espaço de hipóteses H_L , e impõe uma ordem de preferência sobre as hipóteses em H_L .

Teorema *NFL* (*No Free Lunch Theorem*) [Wolpert, 1996].

TAREFAS DE KDD

Classificação

EXEMPLO DE HIPÓTESE



TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – UM EXEMPLO DE APLICAÇÃO

Sexo	Pafs	Idade	Comprar
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
M	França	55	Não

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – UM EXEMPLO DE APLICAÇÃO

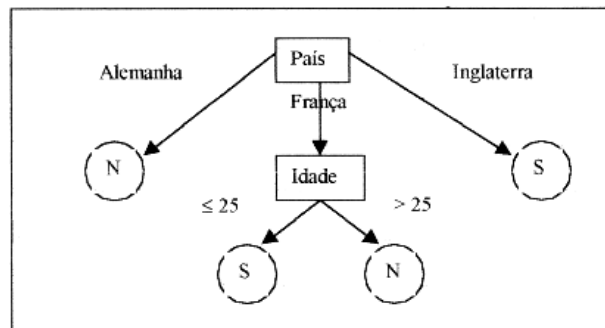
Algumas Regras:

- Se (País = Alemanha) Então Comprar = Não
- Se (País = Inglaterra) Então Comprar = Sim
- Se (País = França e Idade ≤ 25) Então Comprar = Sim
- Se (País = França e Idade > 25) Então Comprar = Não

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – UM EXEMPLO DE APLICAÇÃO

Uma Árvore de Decisão:



TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO

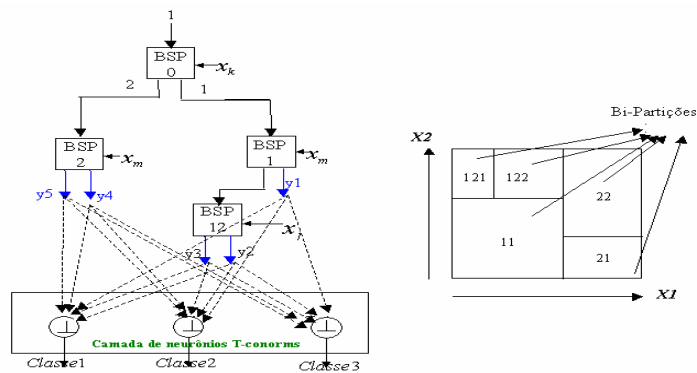
EXEMPLOS DE TÉCNICAS TRADICIONAIS:

- REDES NEURAIS → BACKPROPAGATION
- ÁRVORES DE DECISÃO → ID3, C4.5
- ALGORITMOS GENÉTICOS → RULE EVOLVER
- ESTATÍSTICA → CLASSIFICADORES BAYESIANOS
- BASEADAS EM INSTÂNCIA → K-NN

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO

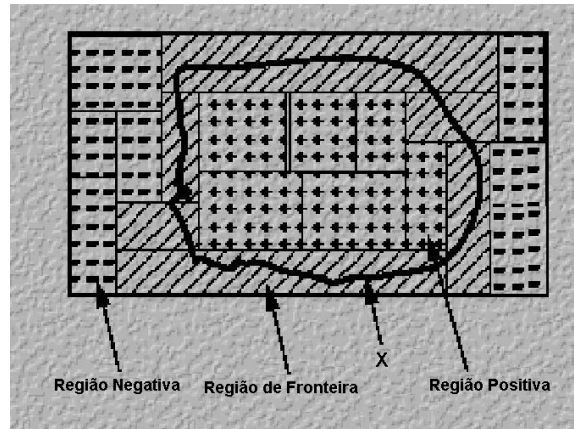
TÉCNICA: MODELOS NEURO-FUZZY HIERÁRQUICOS [Contreras, 2002]



TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO

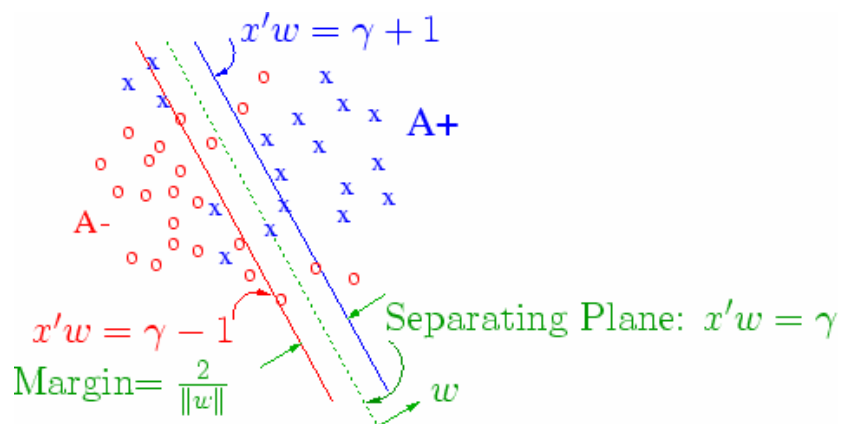
TÉCNICA: ROUGH SETS [Cid, 2002]



TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO

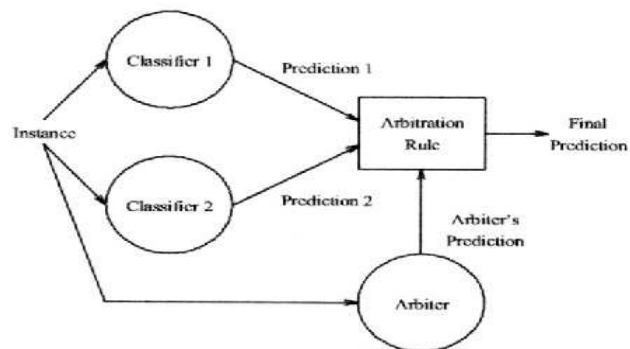
TÉCNICA: SVM – SUPPORT VECTOR MACHINES [Haykin, 2002]



TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO

TÉCNICA: COMITÊS DE CLASSIFICAÇÃO [Chan, 2001]



TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO

EXEMPLOS DE APLICAÇÕES

- FINANÇAS E INVESTIMENTOS
 - SEGUROS
- RECONHECIMENTO DE IMAGEM
- RECONHECIMENTO DE VOZ
 - ETC...

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – Observações Complementares

Uma hipótese pode ser muito específica para o conjunto de treinamento utilizado.

Caso este conjunto não seja suficientemente representativo, o classificador pode ter bom desempenho no conjunto de treinamento, mas não no conjunto de teste.

Diz-se, neste caso, que o classificador ajustou-se em excesso ao conjunto de treinamento, ocorrendo um fenômeno denominado *overfitting*.

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – Observações Complementares

Por outro lado, quando o classificador ajusta-se muito pouco ao conjunto de treinamento, diz-se que ocorre um *underfitting*.

Este fenômeno costuma ocorrer em função de parametrizações inadequadas do algoritmo de aprendizado.

Por exemplo, um número de neurônios insuficiente em uma rede neural, ou uma tolerância de erro excessivamente alta.

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – Observações Complementares

Matriz de Confusão de um Classificador – Mostra, para cada classe, o número de classificações corretas em relação ao número de classificações indicadas pelo modelo.

Classes	Predita C_1	Predita C_2	...	Predita C_k
Verdadeira C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	...	$M(C_1, C_k)$
Verdadeira C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	...	$M(C_2, C_k)$
...	...	...	...	...
Verdadeira C_k	$M(C_k, C_1)$	$M(C_k, C_2)$	...	$M(C_k, C_k)$

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – Observações Complementares

Matriz de Confusão de um Classificador – Mostra, para cada classe, o número de classificações corretas em relação ao número de classificações indicadas pelo modelo.

Classes	Predita C_+	Predita C_-
Verdadeira C_+	Verdadeiros Positivos	Falsos Negativos
Verdadeira C_-	Falsos Positivos	Verdadeiros Negativos

TAREFAS DE KDD

TAREFA: CLASSIFICAÇÃO – Observações Complementares

A matriz de custos pode ser utilizada em determinados algoritmos de aprendizado para compensar a prevalência.

O custo, $Cost(C_i, C_j)$, representa uma penalidade aplicada quando o classificador comete um erro ao rotular exemplos.

$Cost(C_i, C_j) = 0$ quando $i = j$

$Cost(C_i, C_j) > 0$ quando $i \neq j$

$$Err(h) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n M(C_i, C_j) * Cost(C_i, C_j)$$

TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Detecção de Desvios

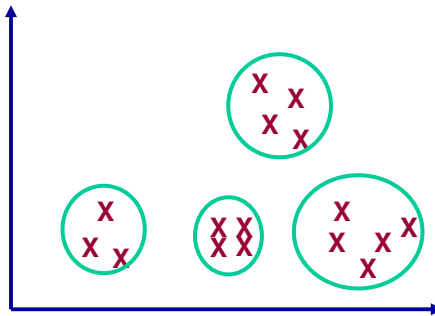
Sumarização

Mineração de Textos

TAREFAS DE KDD

Clusterização – Caracterização Intuitiva:

- Separação dos registros em n “clusters”
- Maximizar/Minimizar similaridade intra/inter cluster



TAREFAS DE KDD

Clusterização – Definições Relevantes:

Def: Cluster: Grupo de registros de um conjunto de dados que compartilham propriedades que os tornam similares entre si.

Def: Clusterização: Processo de particionamento de uma base de dados em conjuntos em que o objetivo é maximizar a similaridade intra-cluster e minimizar a similaridade inter-cluster.

Obs: Não envolve rótulos pré-definidos: **processo de indução não supervisionada.**

TAREFAS DE KDD

Clusterização – Formalização:

Sejam:

- n pontos de dados $x^1, x^2, \dots, x^n$ tais que cada ponto pertença a um espaço k dimensional R^k
- $d: R^k \times R^k \rightarrow R$, uma distância entre pontos de R^k

O processo de Clusterização consiste em encontrar m_j pontos (centróides dos clusters), $j=1, \dots, r$ que minimizem a função

$$\frac{1}{n} \sum_{i=1}^n (\min_j d^2(X_i, m_j))$$

TAREFAS DE KDD

Clusterização – Técnicas Tradicionais:

- Redes Neurais
- Algoritmos Genéticos
- Estatística

TAREFAS DE KDD

Clusterização – Algoritmos Específicos:

- K-Means
- Fuzzy K-Means
- K-Modes
- K-Medoids
- K-Prototypes

TAREFAS DE KDD

Clusterização – Estrutura Comum:

- **Inicialização:** Seleção de um conjunto com k centróides de clusters iniciais no espaço de dados. Esta seleção pode ser aleatória ou de acordo com alguma heurística.
- **Cálculo da Distância:** Calcula a distância euclidiana de cada ponto ou padrão ao centróide de cada cluster. Atribui cada ponto ao cluster cuja distancia do ponto ao centróide do cluster seja minima.

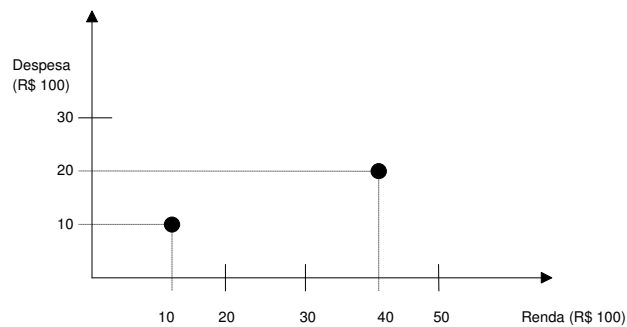
TAREFAS DE KDD

Clusterização – Estrutura Comum:

- **Recálculo dos Centróides:** Recalcula o centróide de cada cluster pela média dos pontos de dados atribuídos ao respectivo cluster.
- **Condição de Convergência:** Repete os passos 2 e 3 até que o critério de convergência tenha sido atingido. Em geral, considera-se um valor de tolerância do erro quadrado médio global abaixo do qual a distribuição dos pontos de dados pelos clusters é considerada satisfatória.

TAREFAS DE KDD

Clusterização – Exemplo de Aplicação:

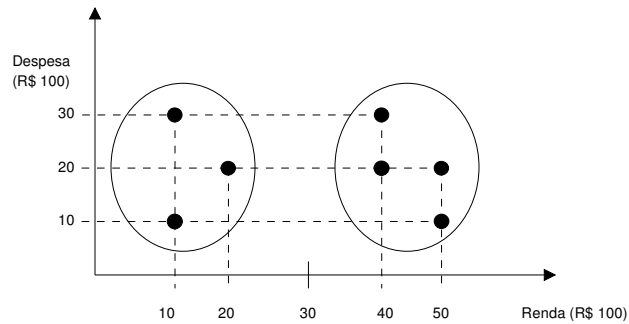


– 02 Clusters com Centróides: (10,10) e (40,20)

TAREFAS DE KDD

Clusterização – Exemplo de Aplicação:

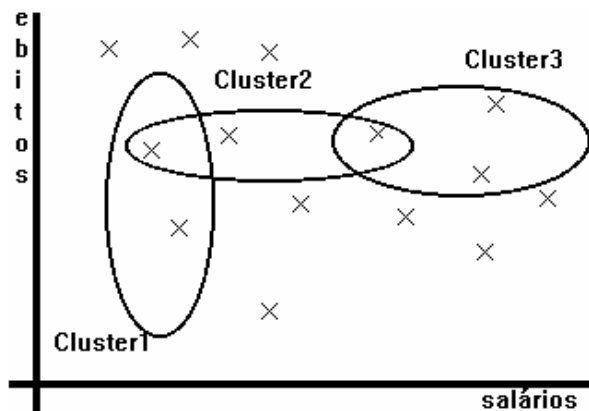
– Sup. os casos: (50,10), (20,20), (10,30), (40,30) e (50,20)



TAREFAS DE KDD

CLUSTERIZAÇÃO

TÉCNICA: FUZZY K-MEANS



TAREFAS DE KDD

CLUSTERIZAÇÃO

EXEMPLOS DE APLICAÇÕES

- **MARKETING DIRETO**
- **SEGMENTAÇÃO DE CLIENTES**
- **MINERAÇÃO DE SUB-ESTRUTURAS EM IMAGENS**

TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Detecção de Desvios

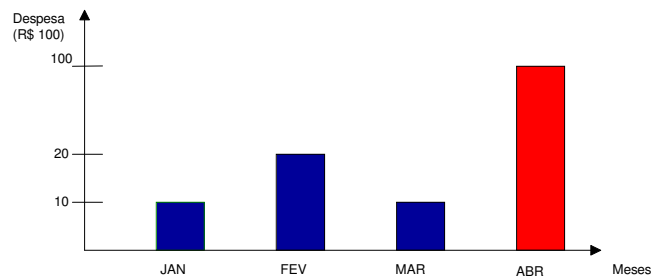
Sumarização

Mineração de Textos

TAREFAS DE KDD

Detecção de Desvios - Caracterização Intuitiva:

- Percepção de valores que vão se enquadram em:
 - Medidas Anteriores
 - Valores Normativos



TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Detecção de Desvios

Sumarização

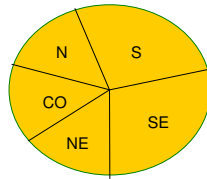
Mineração de Textos

TAREFAS DE KDD

TAREFA: SUMARIZAÇÃO

“Consiste em **descrever** as **características** de **subconjuntos** da base de **dados**.”

Ex: Distribuição dos Assinantes da Revista “X” por Regiões.



Ex.: Qual o perfil dos meninos de rua do Rio de Janeiro? Faixa Etária X, pais consomem drogas, possuem na faixa de Y irmãos, etc...

TAREFAS DE KDD

TAREFA: SUMARIZAÇÃO

EXEMPLOS DE ALGORITMOS TRADICIONAIS:

- **MODELOS ESTATÍSTICOS – VISUALIZAÇÃO**
- **CUBOS DE DADOS - VISUALIZAÇÃO**

TAREFAS DE KDD

TAREFA: SUMARIZAÇÃO

TÉCNICA: ALGORITMOS GENÉTICOS – RULE EVOLVER [LOPES, 2001]

Cromossoma $\equiv$ Regra

Genes $\equiv$ atributos do banco de dados

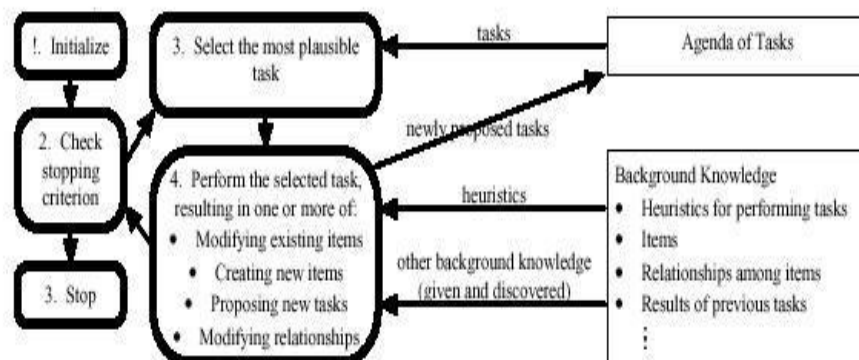
cruzamento

P ₁	Receita Serviço 1 1000<R\$<2000	Receita Serviço 2 4000<R\$<9000	COD_ATIV = 13	10<#_Filiais<50	Empregados>100
P ₂	Receita Serviço 1 5000<R\$<7000	Receita Serviço 2 7000<R\$<8000	COD_ATIV = 14	30<#_Filiais<60	Empregados>300
F ₁	Receita Serviço 1 1000<R\$<2000	Receita Serviço 2 4000<R\$<9000	COD_ATIV = 14	30<#_Filiais<60	Empregados>300
F ₂	Receita Serviço 1 5000<R\$<7000	Receita Serviço 2 7000<R\$<8000	COD_ATIV = 13	10<#_Filiais<50	Empregados>100

TAREFAS DE KDD

TAREFA: SUMARIZAÇÃO

ALGORITMO: HAWB – MINERAÇÃO DE DADOS AUTÔNOMA [Liv, 2002]



TAREFAS DE KDD

Associação

Descoberta de Sequências

Classificação

Clusterização

Deteção de Desvios

Sumarização

Mineração de Textos

TAREFAS DE KDD

MINERAÇÃO DE TEXTOS

- Descobrir conhecimento em textos em Linguagem Natural
- Utiliza recursos de PLN - Processamento de Linguagem Natural

Exemplos de Aplicações:

- Recuperação de Informação
- Extração de Dados
- Classificação de Documentos
- Resumo / Correção de Textos

TAREFAS DE KDD

MINERAÇÃO DE TEXTOS

TÉCNICA: ASSOCIAÇÃO DE TERMOS

Ex.: “Muitas aplicações de KDD utilizam Data Warehouses”.

Após PLN: Aplicações, KDD, Data Warehouses

Após MD: KDD e Data Warehouses são termos frequentes:
têm alguma relação entre si.

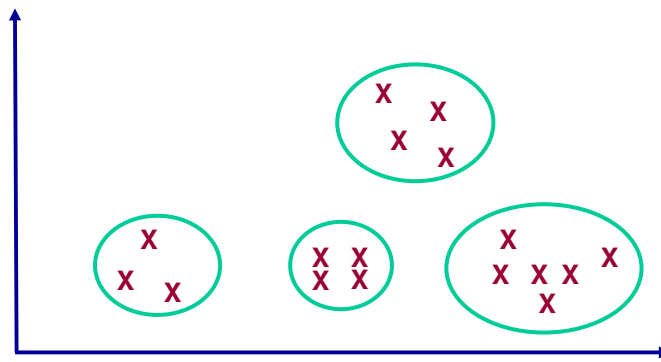
TAREFAS DE KDD

MINERAÇÃO DE TEXTOS

TÉCNICA: HISTOGRAMA - CLUSTERIZAÇÃO DE DOCUMENTOS

TEXTO $\leftrightarrow$ VETOR:

--	--	--	--	--	--	--	--	--



TAREFAS DE KDD

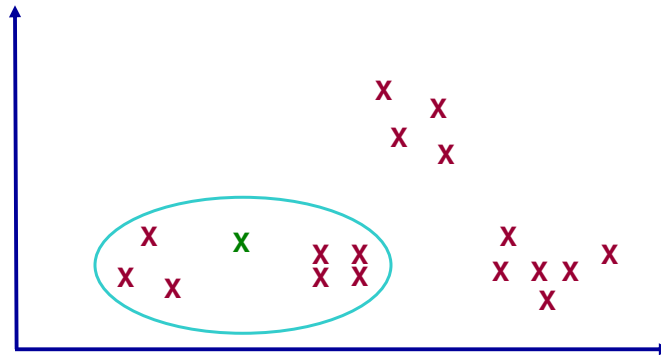
MINERAÇÃO DE TEXTOS

TÉCNICA: CLASSIFICAÇÃO DE DOCUMENTOS (K-NN)

TEXTO $\leftrightarrow$ VETOR:

--	--	--	--	--	--	--	--

 CLASSE



TAREFAS DE KDD

Outras Tarefas – Leitura Complementar:

Regressão

Previsão de Séries Temporais

Clusterização $\rightarrow$ Classificação

Clusterização $\rightarrow$ Sumarização