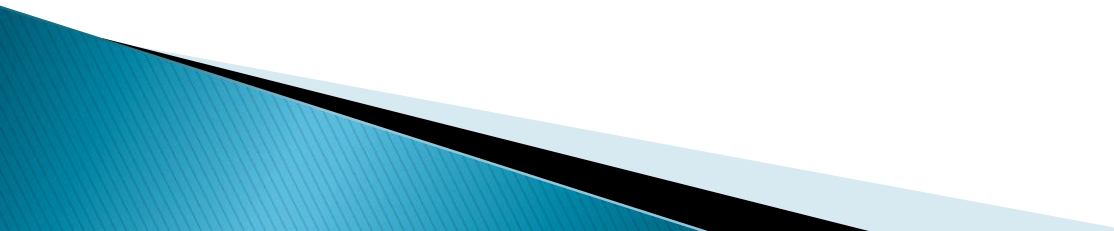


AVALIAÇÃO DE UMA ABORDAGEM LAZY DE SELEÇÃO DE ATRIBUTOS BASEADA NA MEDIDA DE CONSISTÊNCIA

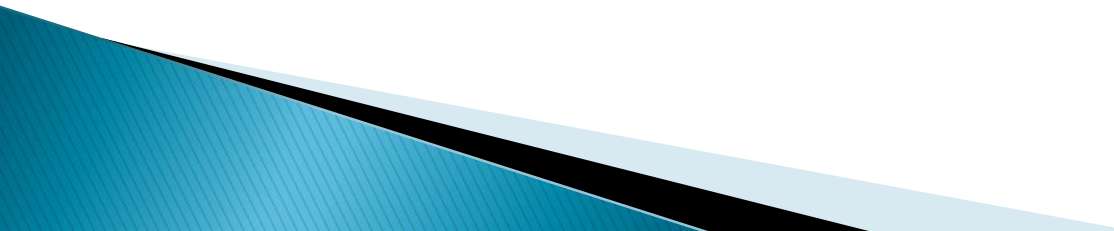
Marcus Vinicius Silva Soares
Orientador: Luiz Merschmann

Dezembro / 2010

Sumário

1. Introdução
 2. Objetivo
 3. Fundamentação Teórica
 4. Implementação
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

Sumário

1. **Introdução**
 2. Objetivo
 3. Fundamentação Teórica
 4. Implementação
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

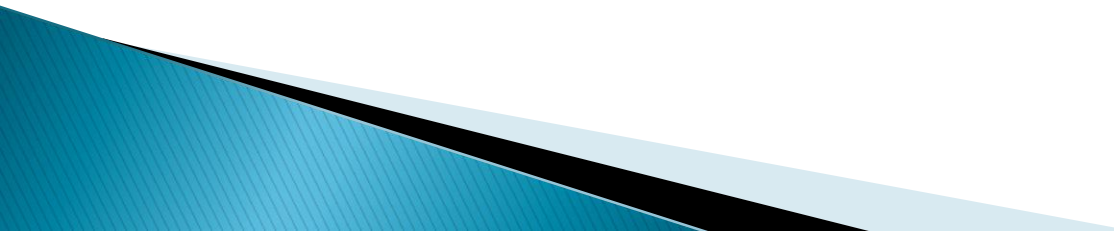
Introdução

- ▶ Grandes quantidades de dados armazenados
- ▶ Dados armazenados podem esconder diversos tipos de padrões e comportamentos relevantes
- ▶ Mineração de dados – (*Knowledge Discovery in Databases* – *KDD*)



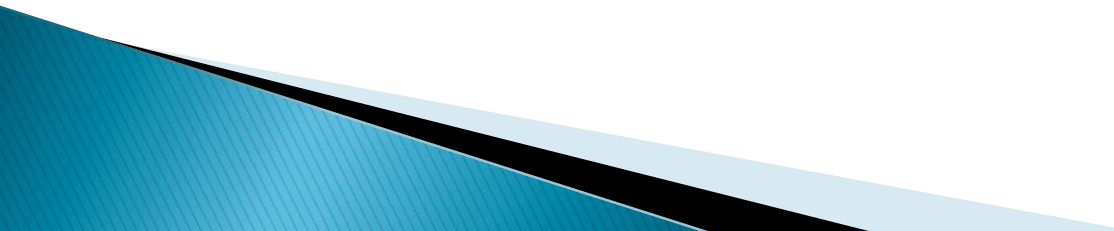
Introdução

Processo KDD

- ▶ Limpeza dos dados
 - ▶ Integração dos dados
 - ▶ Redução de Dados
 - ▶ Transformação dos dados
 - ▶ Mineração
 - ▶ Avaliação ou Pós-processamento
 - ▶ Visualização dos Resultados
- 

Introdução

Processo KDD

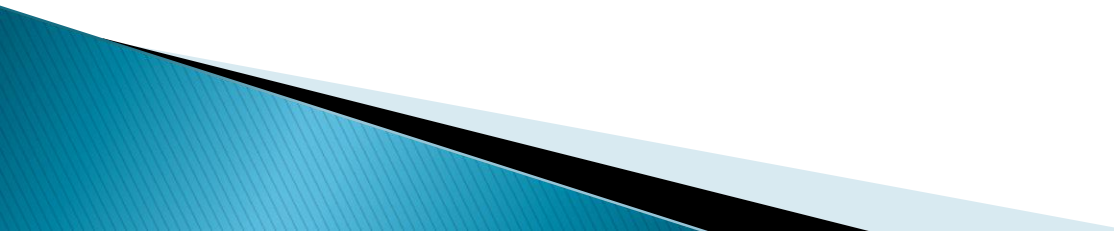
- ▶ Limpeza dos dados
 - ▶ Integração dos dados
 - ▶ Redução de Dados
 - ▶ Transformação dos dados
 - ▶ Mineração
 - ▶ Avaliação ou Pós-processamento
 - ▶ Visualização dos Resultados
- 

Introdução

Processo KDD

- ▶ Limpeza dos dados
- ▶ Integração dos dados
- ▶ Redução de Dados
- ▶ Transformação dos dados
- ▶ Mineração
- ▶ Avaliação ou Pós-processamento
- ▶ Visualização dos Resultados

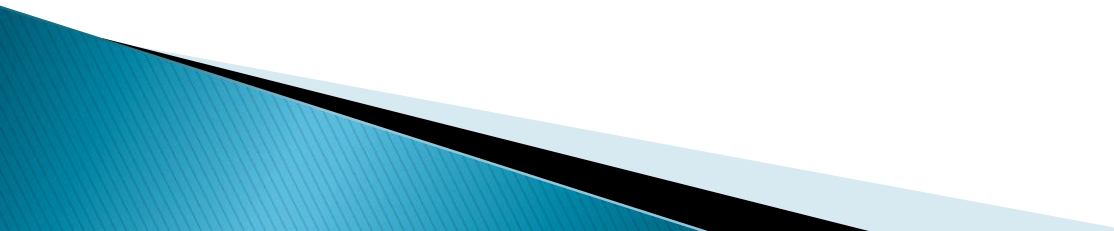
Sumário

1. Introdução
 2. **Objetivo**
 3. Fundamentação Teórica
 4. Implementação
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

Objetivo

- ▶ Avaliar uma abordagem *Lazy* de seleção de atributos baseada na medida de consistência

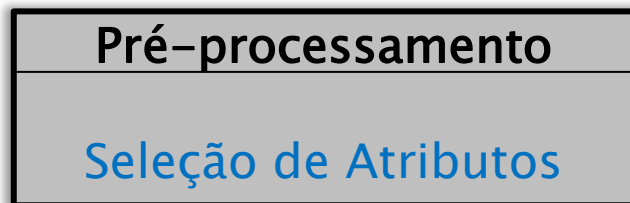
Sumário

1. Introdução
 2. Objetivo
 3. **Fundamentação Teórica**
 4. Implementação
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

Seleção de Atributos

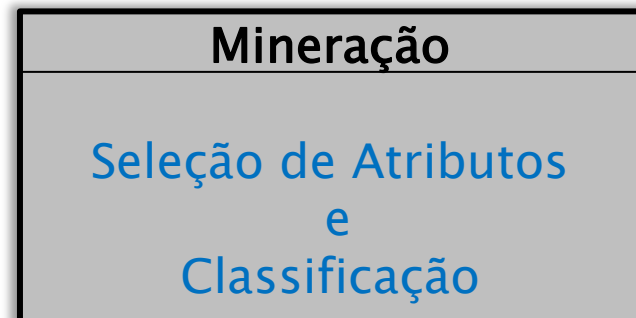
Abordagem Tradicional

Eager



Abordagem Proposta

Lazy



Seleção *Lazy* de Atributos

Base ordenada por X			Base ordenada por Y		
- X -	- Y -	- C -	- X -	- Y -	- C -
1	2	B	2	1	A
1	3	B	3	1	B
1	4	B	4	1	A
2	1	A	1	2	B
2	2	A	2	2	A
2	3	A	3	2	B
3	1	B	1	3	B
3	2	B	2	3	A
3	4	B	4	3	B
4	1	A	1	4	B
4	3	B	3	4	B
4	4	B	4	4	B

Exemplo de Base de Dados: Motivação

Consistency-based Feature Selection

- ▶ Avalia os subconjuntos de atributos e utiliza a consistência como medida de avaliação.
- ▶ Adota alguma heurística.
- ▶ A medida de consistência.

Exemplo: Base de dados D

$S = \{\text{Outlook, Temperature e Windy}\}$

$C(S) = 1 - \text{taxa de inconsistência}(S, D).$

Consistency-based Feature Selection

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	FALSE	yes
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Base de dados D

Consistency-based Feature Selection

i	Padrão S_i Outlook, Temperature, Windy	Classe (Freq.)	Taxa de inconsistência de S_i
1	sunny, hot, FALSE	no(1), yes(1)	$2 - 1 = 1$
2	overcast, hot, FALSE	yes(2)	$2 - 2 = 0$
3	rainy, mild, FALSE	yes(2)	$2 - 2 = 0$
4	rainy, cool, FALSE	yes(1)	$1 - 1 = 0$
5	rainy, cool, TRUE	no(1)	$1 - 1 = 0$
6	overcast, cool, TRUE	yes(1)	$1 - 1 = 0$
7	sunny, mild, FALSE	no(1)	$1 - 1 = 0$
8	sunny, cool, FALSE	yes(1)	$1 - 1 = 0$
9	sunny, mild, TRUE	yes(1)	$1 - 1 = 0$
10	overcast, mild, TRUE	yes(1)	$1 - 1 = 0$
11	rainy, mild, TRUE	no(1)	$1 - 1 = 0$

$$TI(S, D) = \frac{1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0}{14} = 0,0714.$$

$$C(S) = 1 - TI(S, D) = 1 - 0,0714 = 0,929.$$

Adaptação da medida de consistência

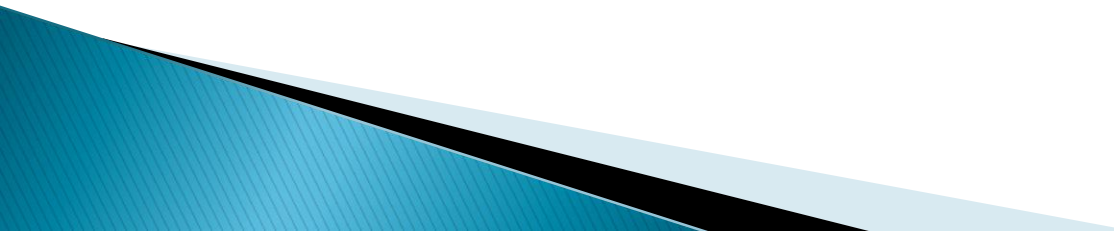
- ▶ Consistência do subconjunto de valores $S_1 = \{\text{Outlook} = \text{sunny}, \text{Temperature} = \text{hot} \text{ e } \text{Windy} = \text{FALSE}\}$ será:

$$C(S_1) = 1 - TI(S_1)$$

$$TI(S_1) = \frac{2 - 1}{2} = 0,5.$$

$$C(S_1) = 1 - TI(S_1) = 1 - 0,5 = 0,5$$

Sumário

1. Introdução
 2. Objetivo
 3. Fundamentação Teórica
 4. **Implementação**
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

Propostas Implementadas

- ▶ Seleção *Lazy* Pura
- ▶ Seleção Híbrida

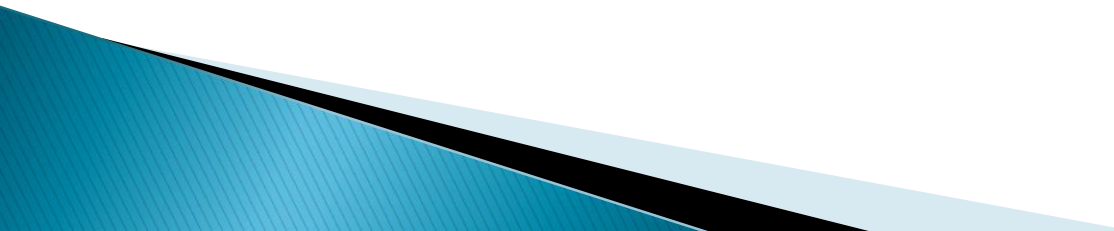
Seleção *Lazy* Pura

```
1 Procedimento SeleçãoLazyPura ( $D(A_1, A_2, \dots, A_n, C), I[v_1, v_2, \dots, v_n]$ )
2  $\text{melhorSubconjunto} \leftarrow \text{SEM SUBCONJUNTO}$ ;
3  $\text{maiorConsist\^encia} \leftarrow -1$ ;
4 for cada subconjunto de valores  $S_i \subseteq I$  gerado pelo Best-first do
5    $CL \leftarrow \text{ConsistencyLazy}(S_i)$ ;
6   if ( $\text{maiorConsist\^encia} < CL$ ) then
7      $\text{melhorSubconjunto} \leftarrow S_i$ ;
8      $\text{maiorConsist\^encia} \leftarrow CL$ ;
9 Retorne melhorSubconjunto;
```

Seleção Híbrida

```
1 Procedimento SeleçãoHíbrida ( $D(A_1, A_2, \dots, A_n, C), I[v_1, v_2, \dots, v_n]$ )
2 melhorSubconjuntoLazy  $\leftarrow$  SEM SUBCONJUNTO;
3 maiorConsistênciaLazy  $\leftarrow -1$ ;
4 maiorConsistênciaEager  $\leftarrow$  ConsistencyEager( $D$ );
5 melhorSubconjuntoEager  $\leftarrow$  SUBCONJUNTO SELECIONADO DE FORMA EAGER;
6 for cada subconjunto de valores  $S_i \subseteq I$ , gerado pelo Best-first do
7   |  $CL \leftarrow$  ConsistencyLazy( $S_i$ );
8   | if ( $\text{maiorConsistênciaLazy} < CL$ ) then
9   |   | melhorSubconjuntoLazy  $\leftarrow S_i$ ;
10  |   | maiorConsistênciaLazy  $\leftarrow CL$ ;
11 if
    ( $\text{maiorConsistênciaLazy} \geq \text{maiorConsistênciaEager}$ )  $\wedge$  ( $\text{Tamanho}(\text{melhorSubconjuntoLazy}) < \text{Tamanho}(\text{melhorSubconjuntoEager})$ )
    then
12   | Retorne melhorSubconjuntoLazy;
13 else
14   | Retorne melhorSubconjuntoEager;
```

Sumário

1. Introdução
 2. Objetivo
 3. Fundamentação Teórica
 4. Implementação
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

Experimentos

- ▶ Algoritmos integrados à implementação do k-NN (IBK).
- ▶ A acurácia preditiva foi o critério utilizado na avaliação
- ▶ Bases utilizadas: *Wine*, *Iris*, *Breast-cancer*.

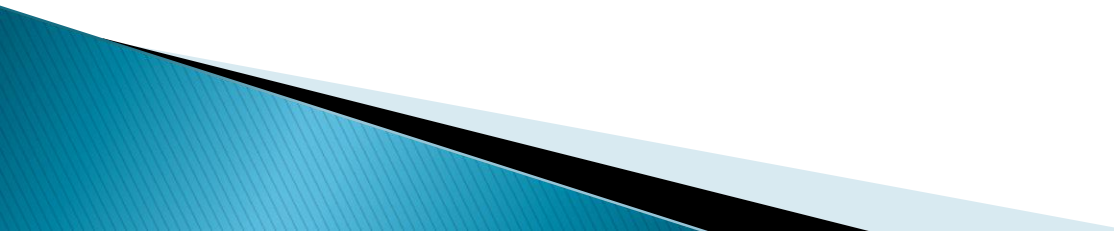
Resultados dos experimentos

<i>wine</i>	k-NN(K=1)	k-NN(K=3)	k-NN(K=5)
Eager	95,55554	94,96731	90,42481
Lazy	98,30065	98,30065	98,30065
Híbrido	98,30065	98,30065	98,30065

<i>iris</i>	k-NN(K=1)	k-NN(K=3)	k-NN(K=5)
Eager	93,99999	94,66666	95,99999
Lazy	92,66666	92,66666	92,66666
Híbrido	93,99999	94,66666	95,99999

<i>breast-cancer</i>	k-NN(K=1)	k-NN(K=3)	k-NN(K=5)
Eager	67,06896	71,24384	73,04187
Lazy	70,23397	70,92364	70,56650
Híbrido	69,85221	70,56650	70,56650

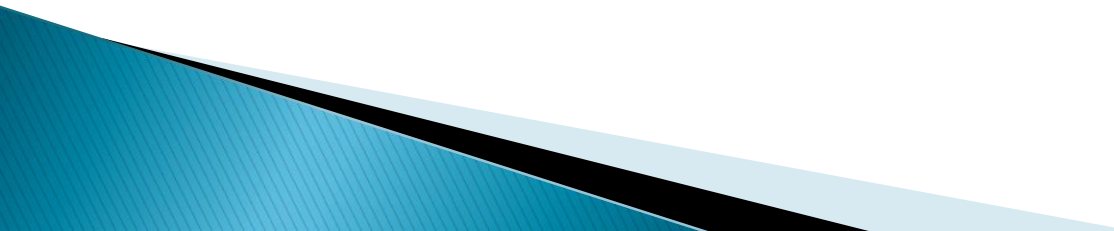
Sumário

1. Introdução
 2. Objetivo
 3. Fundamentação Teórica
 4. Implementação
 5. Experimentos e Análise de Resultados
 6. Conclusão e Trabalhos Futuros
- 

Conclusão

- ▶ Determinadas bases de dados podem se beneficiar da estratégia aqui proposta.

Trabalhos Futuros

- ▶ Avaliar para um número maior de bases de dados.
 - ▶ Avaliar e Compreender as limitações dessa abordagem.
- 

Perguntas?

