Técnicas de Seleção de Atributos utilizando Paradigmas de Algoritmos

Theo Silva Lins

Orientador: Luiz Henrique de Campos Merschmann

Programa de Pós-Graduação em Ciência da Computação
Universidade Federal de Ouro Preto

14 de julho de 2011

Mineração de Dados

- Mineração de Dados
- Importância
- Aplicações
- Classificação

Técnicas de Seleção de atributos

- Técnicas de Seleção de atributos
- Tipos
  - Embedded
  - Wrappers
  - Filtros
- Métricas
  - Ganho de Informação
  - Dependência
  - Correlação
  - Consistência

Técnicas de Seleção de atributos

- Técnicas de Seleção de atributos
- Tipos
  - Embedded
  - Wrappers
  - Filtros
- Métricas
  - Ganho de Informação
  - Dependência
  - Correlação
  - Consistência

Técnicas de Seleção de atributos

- Técnicas de Seleção de atributos
- Tipos
  - Embedded
  - Wrappers
  - Filtros
- Métricas
  - Ganho de Informação
  - Dependência
  - Correlação
  - Consistência

Motivação e Objetivos

- Motivação
- Objetivos

Desenvolvimento do Problema

- Fazer um estudo das técnicas de seleção de atributos.
- Encontrar quais técnicas utilizam os paradigmas de algoritmos.
- Selecionar e implementar uma dessas técnicas, utilizando paradigmas diferentes.
- Fazer experimentos e analisar os resultados.
- Obter um comparativo entre as abordagens utilizadas.

UFOP
Universidade Federal de Ouro Preto

Base de dados - Golf

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

Métrica de Consistência

- Duas instâncias são consideradas inconsistentes se elas possuirem os mesmos valores de atributos e pertencerem a diferentes classes.

Cálculo das taxas de inconsistência.

- A inconsitência é medida com a seguinte formula:
- $TI(S, D) = \frac{\sum_{i=1}^{m} TI(S_i)}{numerodeinstancias}$ onde m é o número de conjuntos de instâncias padrões.
- Logo a consistência será: $C(S) = 1 - TI(S, D)$

Exemplo de Cálculo de Consitência

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

Portanto, teremos a seguinte taxa de inconsitência ($TI(S, D)$):

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

Por fim, a consistência desse subconjunto S será:

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

Algoritmo

- Menor inconsistência e o menor tamanho.
- A base de dados com n atributos, tem-se 2^n subconjuntos de atributos possíveis.
- Métodos heurísticos.

Algoritmo

- Menor inconsistência e o menor tamanho.
- A base de dados com n atributos, tem-se 2^n subconjuntos de atributos possíveis.
- Métodos heurísticos.

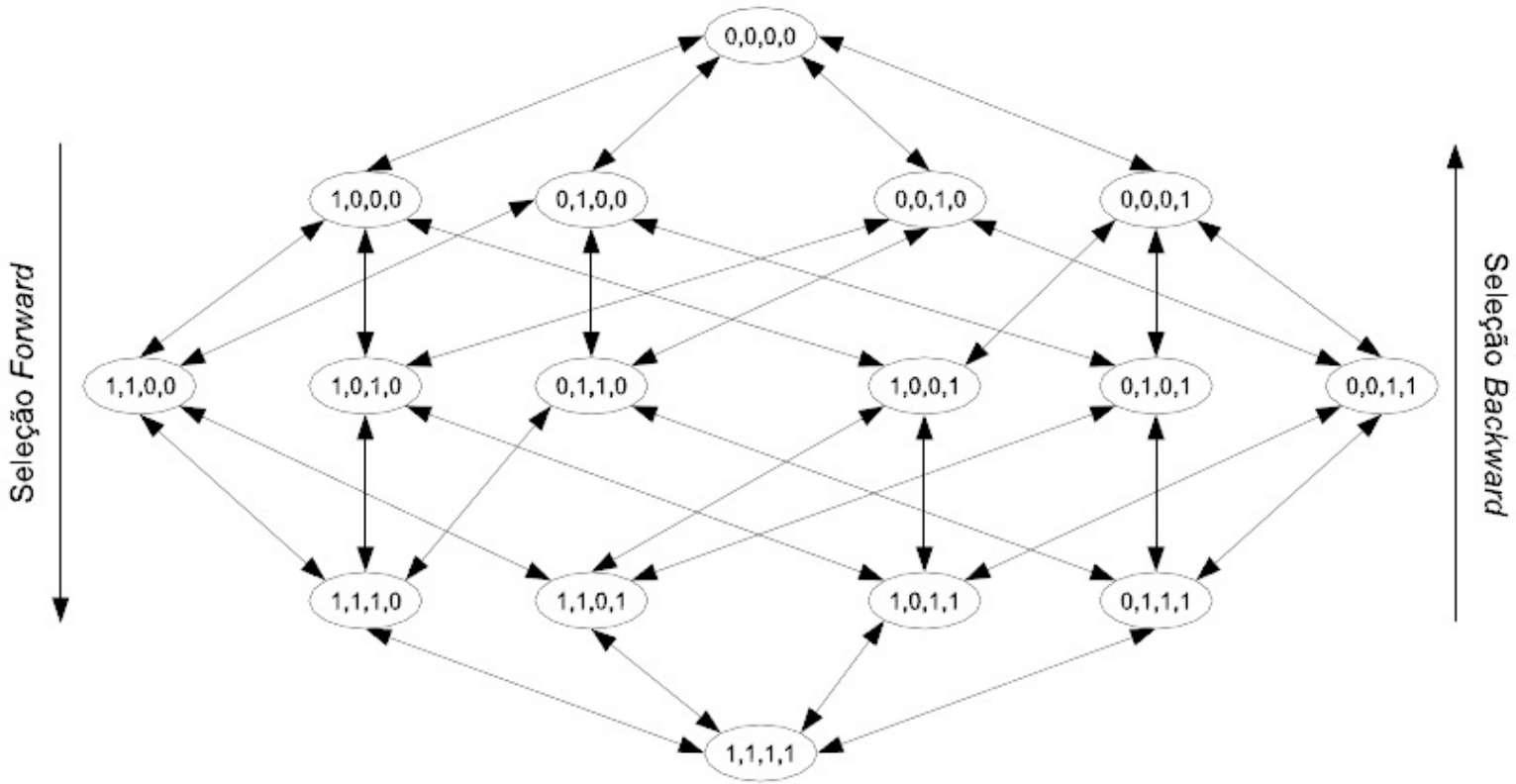
Algoritmo

- Menor inconsistência e o menor tamanho.
- A base de dados com n atributos, tem-se 2^n subconjuntos de atributos possíveis.
- Métodos heurísticos.

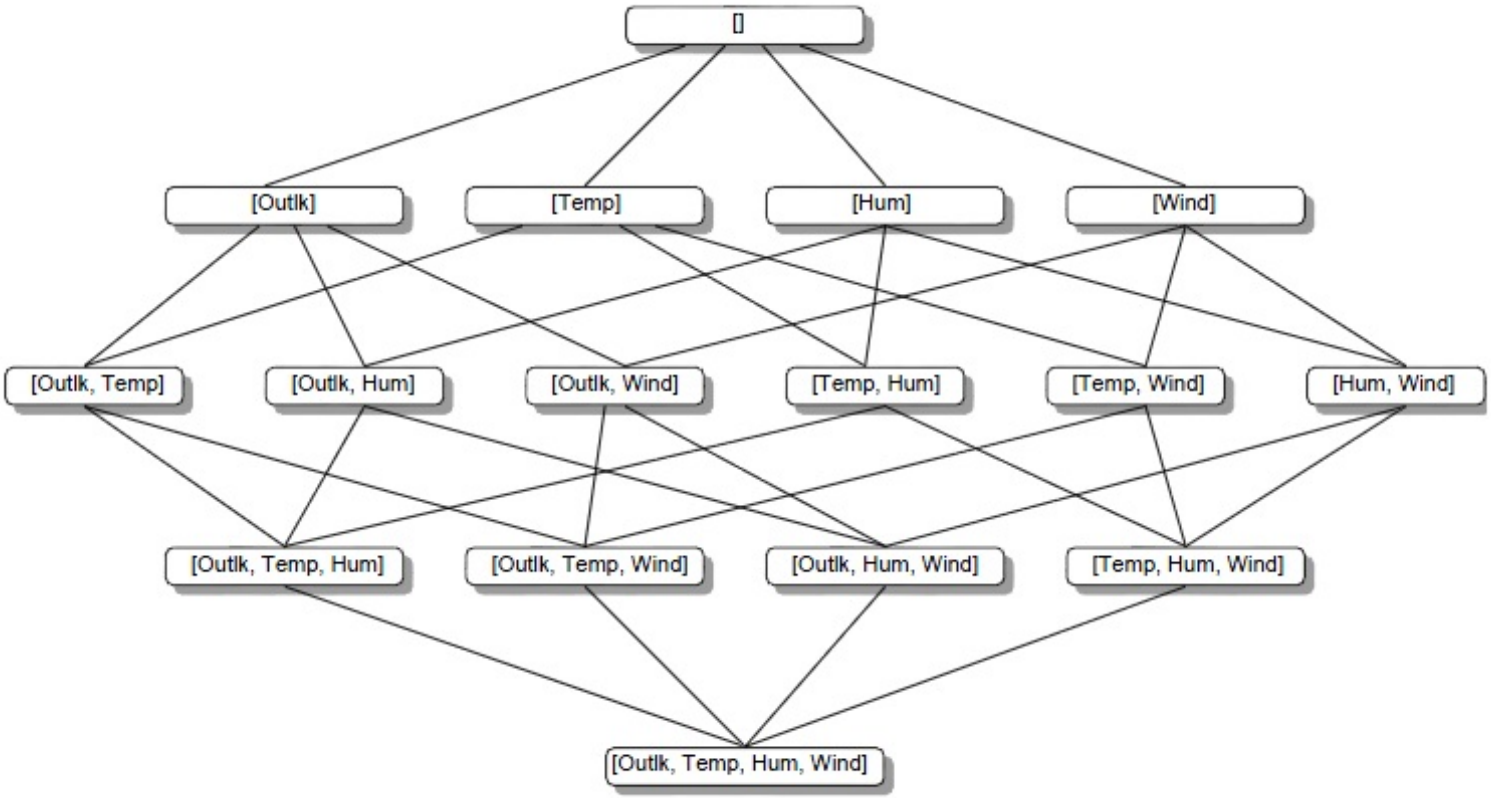
Heuristicas

- Ponto de partida.
- A estratégia que será utilizada para percorrer o espaço de soluções.
- A forma de avaliação dos subconjuntos de atributos.
- Critério de parada.

Espaço de Soluções - Exemplo 1



Espaço de Soluções - Exemplo 2



Backtracking

- Método impraticável para datasets com muitos atributos.
- Realiza uma busca completa.

Max Tries

- A idéia desse algoritmo é, por um número máximo de tentativas (max tries), e gerar subconjuntos de atributos.
- O algoritmo para quando o número de tentativas acabar
- Realiza uma busca não-determinística.

Hill Climbing

- Gera os subconjuntos de soluções vizinhas
- Escolher a melhor solução entre vizinhas
- O algoritmo para quando não encontrar nenhum subconjunto melhor entre os vizinhos.
- Realiza uma busca heurística

Hill Climbing

- Gera os subconjuntos de soluções vizinhas
- Escolher a melhor solução entre vizinhas
- O algoritmo para quando não encontrar nenhum subconjunto melhor entre os vizinhos.
- Realiza uma busca heurística

Hill Climbing

- Gera os subconjuntos de soluções vizinhas
- Escolher a melhor solução entre vizinhas
- O algoritmo para quando não encontrar nenhum subconjunto melhor entre os vizinhos.
- Realiza uma busca heurística

Hill Climbing

- Gera os subconjuntos de soluções vizinhas
- Escolher a melhor solução entre vizinhas
- O algoritmo para quando não encontrar nenhum subconjunto melhor entre os vizinhos.
- Realiza uma busca heurística

- Complexidade $f(n) = O(2^n)$
  - Onde n é o número de atributos.

- Complexidade $f(n) = O(2^n)$
  - Onde n é o número de atributos.

- Complexidade $f(n) = O(m)$
  - Onde m é o número tentativas.

- Complexidade $f(n) = O(m)$
  - Onde m é o número tentativas.

- Complexidade $f(n) = O(n^2)$
  - Onde n é o número de atributos.

- Complexidade $f(n) = O(n^2)$
  - Onde n é o número de atributos.

Informações

- Consistência.
- Tempo de Execução
- Atributos Selecionados.

Base de Dados

- Golf Data set
- Breast Cancer Wisconsin Data Set
- SPECTF Heart Data Set
- Congressional Voting Records Data Set

Algoritmos

- Max-Tries
  - Tentativas = 1000;
  - Consistência Mínima = 0.8;
- Backtracking

Tabelas de Resultado

Tabela I: Max Tries

Base de dados	Instancias	Atrib.	Atrib. Selecionados	Tempo	Consistência
Golf	14	4	3	0,001	0.928
Breast Cancer	699	10	10	19,4	0,992
SPECTF Heart	187	44	9	19,9	0,994
Congress. Voting	435	16	10	41,8	0,990

Tabela II: Backtracking

Base de dados	Instancias	Atrib.	Atrib. Selecionados	Tempo	Consistência
Golf	14	4	3	0,001	0.928
Breast Cancer	699	10	7	27,1	1
SPECTF Heart	187	44	X	X	X
Congress. Voting	435	16	9	611,7	0,990

Universidade Federal de Ouro Preto

Conclusão

- Importância das Técnicas de Seleção de atributos
- Experimentos realizados

Trabalhos Futuros

- Implementação de outros algoritmos.
- Estudo e implementação com outras Técnicas de Seleção de Atributos