Classificação e 1R

Prof. Eduardo Raul Hruschka

Visão Geral:

- Aquecimento: 1R;
- Naïve Bayes.

Inferindo regras rudimentares:

- 1R: aprende uma árvore de decisão de um nível.
  - Todas as regras usam somente um atributo.
- Versão Básica:
  - Um ramo para cada valor do atributo;
  - Para cada ramo, atribuir a classe mais freqüente;
  - Taxa de erro de classificação: proporção de exemplos que não pertencem à classe majoritária do ramo correspondente;
  - Escolher o atributo com a menor taxa de erro de classificação;

* Atributos nominais/categóricos;

- Há vários algoritmos de discretização para definir estratégias de corte nos valores dos atributos (<=, <, >, >=).

Algoritmo 1R em pseudo-código:

Para cada atributo:

Para cada valor do atributo gerar uma regra como segue:

Contar a freqüência de cada classe;

Encontrar a classe mais freqüente;

Formar uma regra que atribui à classe mais freqüente este atributo-valor;

Calcular a taxa de erro de classificação das regras;

Escolher as regras com a menor taxa de erro de classificação.

1R para o problema *weather* :

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
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

Attribute	Rules	Errors	Total errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No*	3/6	

* empate

Qual seria a capacidade de generalização do modelo?

Discussão para o 1R:

- 1R foi descrito por Holte (1993):
  - Contém uma avaliação experimental em 16 bases de dados;
  - Em muitos *benchmarks*, regras simples não são muito piores do que árvores de decisão mais complexas...
  - Complexidade de tempo?
- Implementado no Weka;
- Atualmente usado para análise exploratória de dados;
- Árvores de Decisão estendem essa ideia;
- Mas antes de abordá-las, abordaremos um algoritmo muito eficaz e (computacionalmente) eficiente: NB.

Holte, Robert C., Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, *Machine Learning* 11 (1), pp. 63-90, 1993.