Estudo Comparativo entre Métodos de Seleção de Atributos Baseados em Medidas de Precisão e Correlação Aplicados a Bases de Dados

Antonio Rafael Sabino Parmezan, Huei Diana Lee, Carlos Andrés Ferrero, Willian Zalewski, André Gustavo Maletzke, Feng Chung Wu
Laboratório de Bioinformática – LABI, UNIOESTE, Foz do Iguaçu, PR

Objetivos

A propagação de sistemas computacionais tem contribuído para a geração e o armazenamento de uma quantidade crescente de dados. A qualidade desses dados afeta diretamente a aplicação de processos que podem auxiliar na extração de conhecimento, como o de Mineração de Dados [1]. Nesse contexto, além da significância da amostra de dados, tarefas de pré-processamento, como a Seleção de Atributos (SA), tornam-se importantes para melhorar o desempenho de ferramentas de análise de dados. Neste trabalho, o objetivo consiste no estudo de medidas de avaliação de importância de atributos e algoritmos para a tarefa de SA.

Métodos/Procedimentos

A avaliação dos métodos de SA foi conduzida em três etapas. Na etapa (1) foi realizada a SA utilizando as medidas de importância de correlação e precisão por meio dos algoritmos CfsSubsetEval (CFS) e ClassifierSubsetEval (CS), respectivamente, sobre quatro conjuntos de dados naturais obtidos do repositório de dados UCI[1]. Na etapa (2) os conjuntos de dados foram avaliados, com e sem SA, por meio do desempenho de modelos construídos usando o algoritmo J48 para indução de árvores de decisão. Nesta etapa utilizou-se validação cruzada com 10 partições, sendo que os experimentos descritos foram executados no ambiente Weka [2]. Na etapa (3) foi aplicado o teste estatístico ANOVA, para verificar a existência de diferença estatisticamente significativa entre os valores médios de precisão dos modelos construídos.

Resultados

Os resultados dos experimentos realizados neste trabalho são apresentados na Tabela 1.

Tabela 1: Precisão, desvio padrão e percentagem de redução de atributos para cada conjunto de dados e cada algoritmo considerado

Conjunto de Dados	Original	CFS	CS
BreastCancer	74,28(6,05) 0%	72,94(5,47) 40%	73,40(5,43) 60%
Bupa	65,83(7,40) 0%	62,24(8,67) 71,43%	67,51(7,91) 28,58%
Hungarian	80,22(7,95) 0%	78,99(7,29) 50%	81,62(6,84) 64,29%
Pima	74,49(5,26) 0%	74,37(5,04) 44,45%	75,14(5,11) 22,23%
Média	73,70(13,33)	72,13(13,23)	74,41(12,64)

Estatisticamente não foi possível identificar diferenças significativas entre CFS e CS em relação ao Original, entretanto, observa-se uma expressiva redução de atributos após a SA. Nesse contexto, os resultados da aplicação de SA são motivadores, tendo em vista que a redução da dimensionalidade pode auxiliar na diminuição dos custos de coleta de dados e da complexidade dos modelos construídos.

Conclusões

Neste trabalho, constatou-se que, embora os resultados da SA não tenham apresentado diferenças significativas, a seleção de um subconjunto de atributos pode simplificar a linguagem de descrição de exemplos quando esta possui atributos redundantes. Como trabalhos futuros destacam-se o estudo comparativo entre outros tipos de medidas de avaliação de importância de atributos e de algoritmos para a tarefa de SA, com foco na abordagem filtro.

Referências Bibliográficas

[1] Lee HD. Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado – USP, 2005.
[2] Witten IH and Frank E. Data Mining: Practical Machine Learning Tools and Techniques. California: MK, 2005.

[1]Asuncion A and Newman D. UCI machine learning repository, 2007. Disponível em:<http://archive.ics.uci.edu/ml/>. Acesso em: 3 set. 2010.