Redução da Dimensionalidade em Bases de Dados Naturais através de Métodos de Filtro para Seleção de Atributos Importantes

Antonio Rafael Sabino Parmezan, Huei Diana Lee, Wu Feng Chung
Laboratório de Bioinformática – LABI, UNIOESTE, Foz do Iguaçu, PR

Objetivos

Os recentes avanços computacionais têm contribuído para o crescimento acelerado de grandes volumes de dados. A qualidade desses dados afeta diretamente processos que visam auxiliar em extração e análise inteligente de dados, como o de Mineração de Dados (MD) [1]. Nesse contexto, métodos de pré-processamento para redução da dimensionalidade, tal como a Seleção de Atributos (SA), permitem, além da remoção de atributos redundantes e irrelevantes, uma melhor compreensibilidade dos resultados gerados [2]. Neste trabalho, o objetivo consiste na avaliação de quatro medidas de importância, empregadas na abordagem filtro, para a SA.

Métodos/Procedimentos

A avaliação dos métodos de filtro foi conduzida em três etapas com o suporte do ambiente Weka[1]. Na etapa 1, foi realizada a SA utilizando as medidas de consistência, correlação, informação e distância por meio dos algoritmos *Consistency-Based Filter* (CBF), *Correlation-based Feature Selection* (CFS), InfoGain e ReliefF, respectivamente, sobre cinco conjuntos de dados (CD) naturais obtidos do repositório de dados UCI[2]. Para os algoritmos de avaliação individual (InfoGain e ReliefF), considerou-se a seleção de 30% do total de atributos. Na etapa 2, modelos foram construídos, com e sem SA, utilizando o algoritmo J48 para indução de árvores de decisão. Na etapa 3, os algoritmos foram comparados com o Original (sem SA), quanto à precisão dos modelos construídos (MC) estimada por meio de validação cruzada com dez partições, usando o teste não paramétrico *Kruskal-Wallis* (nível de significância de 5%) e pós-teste de *Dunn*.

Resultados

Os resultados dos experimentos descritos anteriormente são apresentados na Tabela 1.

Tabela 1 – Valor médio de precisão e desvio padrão entre parênteses; abaixo % de redução de atributos para cada CD e algoritmos considerados

Original	CBF	CFS	InfoGain	ReliefF
BreastCancer	72,51(6,32)	72,94(5,47)	**71,86(5,65)**	73,33(6,27)
74,28(6,05)	11,11%	44,44%	66,67%	66,67%
Bupa	62,24(8,67)	62,24(8,67)	**62,50(6,78)**	68,07(6,41)
65,84(7,40)	83,33%	83,33%	66,67%	66,67%
Hepatitis	79,74(8,96)	80,27(9,04)	**83,85(7,22)**	**84,22(8,19)**
79,22(9,57)	36,84%	47,37%	68,42%	68,42%
Hungarian	80,40(7,97)	79,00(7,30)	79,44(7,25)	78,91(5,75)
80,22(7,95)	23,08%	53,85%	69,23%	69,23%
Pima	74,49(5,27)	74,38(5,04)	74,65(5,02)	74,65(5,02)
74,49(5,27)	0,00%	50,00%	75,00%	75,00%
Média	73,88(10,00)	73,77(9,67)	74,46(9,69)	75,84(8,39)
74,81(8,96)	30,87%	55,80%	69,20%	69,20%

Em negrito, resultados com diferença estatisticamente significativa (d.e.s), (*BreastCancer*: $p = 0{,}0108$; *Bupa* e *Hepatitis*: $p < 0{,}0001$).

Como pode ser observado, o CBF selecionou os maiores subconjuntos de atributos, variando de um mínimo de 16,67% para *Bupa* até o máximo de 100,00% (todos os atributos) para *Pima*. Já o CFS proporcionou, em média, a redução de 55,80% da dimensionalidade dos CD. Considerando InfoGain houve deterioração da performance para *Bupa* e *BreastCancer*, enquanto para *Hepatitis* houve incremento da precisão para ambos os algoritmos de avaliação individual.

Conclusões

Embora não tenha sido possível identificar d.e.s em 80,00% das comparações, os métodos de filtro, além de promoverem a redução média de 56,28% da dimensionalidade dos CD, apresentaram desempenho semelhante aos Originais quanto à precisão dos MC. Como trabalhos futuros, relacionaremos a conduta desses algoritmos com as propriedades dos CD, visando, por meio de métodos para Meta-Aprendizado, a recomendação de algoritmos mais apropriados para SA.

Referências Bibliográficas

[1] Lee HD. Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado – USP, 2005.
[2] Liu H and Motoda H. Computational Methods of Feature Selection. C & Hall/CRC, 2008.

[1]Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, v. 11, 1 ed. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 10 ago. 2011.

[2]Frank A and Asuncion A. UCI Machine Learning Repository. Irvine: University of California, School of Information and Computer Science, 2010. Disponível em: <http://archive.ics.uci.edu/ml>. Acesso em: 10 ago. 2011.