ESTUDO DE MEDIDAS DE IMPORTÂNCIA E ALGORITMOS PARA SELEÇÃO DE ATRIBUTOS PARA MINERAÇÃO DE DADOS

Antonio Rafael Sabino Parmezan (PIBIC/CNPq-Unioeste), Wu Feng Chung, Huei Diana Lee (Orientadora), e-mail: hueidianalee@gmail.com.

Universidade Estadual do Oeste do Paraná/Centro de Engenharias e Ciências Exatas/Laboratório de Bioinformática/Foz do Iguaçu, PR.

Ciências Exatas e da Terra, Ciência da Computação.

Palavras-chave: pré-processamento de dados, aprendizado de máquina, extração de padrões.

Resumo:

Tarefas de pré-processamento em Mineração de Dados são de fundamental importância para assegurar a qualidade dos dados fornecidos como entrada para algoritmos de extração de padrões. Nesse contexto, a Seleção de Atributos é realizada, dentre outros motivos, para determinar os atributos mais significativos em um conjunto de dados. Neste trabalho é apresentado um estudo comparativo entre os diferentes tipos de medidas de importância de atributos, através de algoritmos para tarefa de Seleção de Atributos. Resultados experimentais mostram que a Seleção de Atributos, a partir da redução da dimensionalidade, pode auxiliar na diminuição dos custos de coleta de dados e da complexidade de modelos construídos.

Introdução

A difusão de sistemas computacionais tem contribuído para a geração e o armazenamento de uma quantidade crescente de dados. A qualidade dessa informação[1] afeta diretamente processos que podem auxiliar na extração de conhecimento a partir de bases de dados. Desse modo, é necessário que esses dados sejam representados de maneira apropriada, processados e o modelo construído (MC), avaliado e validado [4]. Uma das maneiras de se alcançar esse objetivo é por meio da realização do processo de Mineração de Dados (MD), o qual é divido em três fases: pré-processamento, extração de padrões e pós-processamento. Todas essas três fases são importantes para que esse processo seja realizado com sucesso. No entanto, a fase de pré-processamento que é considerada como uma das mais trabalhosas e demoradas, cerca de 80% de todo o processo, é de fundamental importância para certificar que os dados sejam de qualidade [1]. Nesse contexto, tarefas

[1] Neste trabalho os termos dados e informação serão usados indistintamente.

como a Seleção de Atributos (SA), podem auxiliar na melhoria e no desempenho de ferramentas de análise de dados, a partir da simplificação da linguagem de descrição de exemplos quando esta possuir mais atributos que os necessários [2].

Neste trabalho, é apresentado um estudo comparativo entre os diferentes tipos de medidas de importância de atributos pertencentes às categorias: clássica (informação, distância e correlação), consistência e precisão. Para tanto, foram investigados alguns dos principais algoritmos de SA que se baseiam nessas medidas. Resultados experimentais obtidos com diversos conjuntos de dados (CD) demonstraram que a SA, por meio da redução da dimensionalidade, auxilia na melhora da qualidade dos dados sob a perspectiva de desempenho preditivo (DP), o que contribui para a construção de modelos de indução com menor custo computacional.

Materiais e Métodos

Neste trabalho foram considerados cinco algoritmos para a tarefa de SA comumente empregados pela comunidade acadêmica. Quatro desses algoritmos são utilizados na abordagem filtro, sendo que dois deles (CBF e CFS) selecionam os atributos pelo critério de avaliação de subconjuntos e os outros dois (InfoGain e ReliefF) por meio da avaliação individual [1, 2, 3]. Diferentemente desses algoritmos, o CSE é um método *wrapper* que avalia subconjuntos de atributos a partir do CD de treinamento ou de teste, utilizando, para tanto, um classificador [2]. Em relação à medida usada para avaliar a importância dos atributos, esses algoritmos se baseiam em consistência (CBF), correlação (CFS), distância (ReliefF), informação (InfoGain) e precisão (CSE). A avaliação desses métodos de SA foi conduzida em três etapas, as quais são ilustradas na Figura 1.

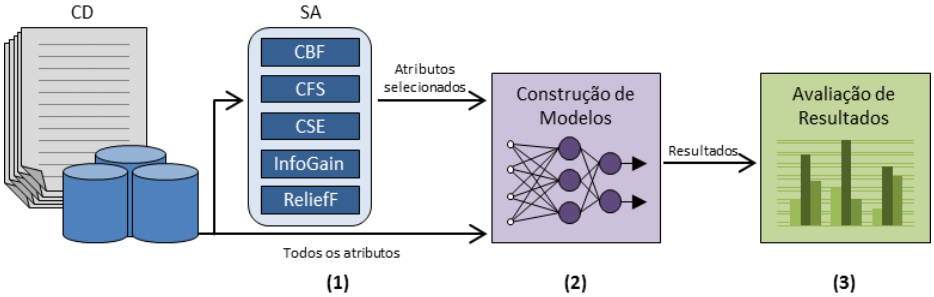

Figura 1 – Configuração dos experimentos.

Na etapa (1), foi realizada a SA utilizando os algoritmos descritos anteriormente sobre sete CD biológicos naturais de acesso público obtidos do repositório de dados UCI[2]. Todos esses algoritmos, à exceção do CSE, foram executados considerando seus parâmetros configurados com valores

[2] FRANK, A.; 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: 25 jul. 2011.

padrão. Para o CSE utilizou-se como classificador o algoritmo *Multilayer Perceptron* (MLP). Considerou-se, em média, a seleção de 30% do total de atributos para os algoritmos de avaliação individual. A variação nessa percentagem de seleção estabelecida está relacionada ao número original de atributos, que assume valores inteiros. Na etapa (2), modelos foram construídos, com e sem SA, usando o indutor MLP, totalizando 42 MC. Na etapa (3), os algoritmos foram comparados com o Original (sem SA), quanto ao DP dos MC estimado por meio de validação cruzada com dez partições, utilizando o teste estatístico não paramétrico *Kruskal-Wallis* para grupos não pareados, com nível de significância de 5%, seguido do pós-teste de *Dunn*[3]. Os experimentos foram realizados com o suporte do ambiente Weka [4].

Resultados e Discussão

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

Tabela 1 – Média do DP e desvio padrão entre parênteses; abaixo % de redução de atributos para cada CD e algoritmos considerados.

CD (ECM)	MLP						Média SA
	Original	CBF	CFS	CSE	InfoGain	ReliefF	
BreastCancer (29,72)	66,95(8,51) 0,00%	66,66(8,92) 11,11%	**71,97(7,34)** 44,44%	66,66(8,92) 11,11%	70,14(6,87) 66,67%	**72,25(6,78)** 66,67%	69,53(8,17) 40,00%
Bupa (42,03)	68,73(7,38) 0,00%	**60,07(7,32)** 83,33%	**60,07(7,32)** 83,33%	71,29(7,24) 16,67%	**60,70(7,74)** 66,67%	68,29(6,68) 66,67%	64,08(8,66) 63,33%
Haberman (26,47)	70,32(6,70) 0,00%	72,76(5,65) 33,33%	72,76(5,65) 33,33%	70,32(6,70) 0,00%	**73,62(5,90)** 66,67%	**73,01(2,16)** 66,67%	72,49(5,54) 40,00%
Hepatitis (20,65)	81,29(9,17) 0,00%	78,71(8,95) 36,84%	82,50(9,76) 47,37%	78,19(9,77) 36,84%	83,63(7,92) 68,42%	**85,33(9,00)** 68,42%	81,67(9,49) 51,58%
Hungarian (36,05)	80,32(6,36) 0,00%	82,45(6,58) 23,08%	81,43(7,06) 53,85%	80,32(6,36) 0,00%	79,39(6,62) 69,23%	78,88(6,90) 69,23%	80,49(6,81) 43,08%
LungCancer (59,38)	68,75(22,83) 0,00%	**84,75(19,03)** 92,86%	**83,25(20,74)** 85,71%	**84,75(19,03)** 92,86%	69,58(22,11) 69,64%	70,50(22,14) 69,64%	78,57(21,73) 82,14%
Pima (34,98)	74,75(4,90) 0,00%	74,75(4,90) 0,00%	75,97(4,47) 50,00%	74,75(4,90) 0,00%	75,50(4,70) 75,00%	75,50(4,70) 75,00%	75,29(4,74) 40,00%
Média	73,02(12,19) 0,00%	74,31(12,69) 40,08%	**75,42(12,71)** 56,86%	75,18(11,54) 22,50%	73,22(12,45) 68,90%	74,82(11,52) 68,90%	

Erro da Classe Majoritária (ECM): erro cometido no caso de novos exemplos serem classificados como sendo pertencentes à classe majoritária. Em negrito resultados estatisticamente significativos.

Como pode ser observado na Tabela 1, foi possível identificar diferenças estatisticamente significativas (d.e.s) quando comparados individualmente, para cada CD, os algoritmos de SA em relação ao Original. Para *BreastCancer*, os algoritmos CFS e ReliefF apresentaram melhora na performance dos MC a partir dos subconjuntos de atributos selecionados, com $p < 0,0001$. Diferentemente, para *Bupa* os algoritmos CBF, CFS e InfoGain apresentaram degradação da performance dos classificadores induzidos, com $p < 0,0001$. Esse fato pode ter sido causado pela expressiva

[3] Testes estatísticos realizados utilizando GraphPad InStat versão 3.05 para Windows, GraphPad Software. Disponível em: <http://www.graphpad.com>. Acesso em: 25 jul. 2011.

redução do número de atributos. Já para *Haberman*, InfoGain e ReliefF aprimoraram o desempenho dos MC, igualmente, para *Hepatitis*, em relação a ReliefF. *LungCancer* foi o único CD no qual todos os algoritmos promoveram expressivas reduções de atributos e apresentaram melhora do desempenho preditivo, com d.e.s, para três do total de cinco MC a partir dos subconjuntos de atributos selecionados. É importante ressaltar que os MC utilizando os CD originais *BreastCancer* e *Haberman* apresentaram média do erro superior ao ECM. Ainda, na última linha da Tabela 1, é apresentada a média referente ao desempenho geral de cada um dos algoritmos de SA sobre os sete CD. Assim sendo, foi possível identificar d.e.s entre o CFS em relação ao Original, com p = 0,0073. Esse resultado indica que, para os CD utilizados, o CFS, baseado em correlação, mostrou-se adequado para a tarefa de SA.

Conclusões

Neste trabalho, constatou-se que, do ponto de vista de DP dos MC usando o indutor MLP, não foi possível identificar d.e.s em 68,57% das comparações. Isso significa que o desempenho entre os MC a partir dos subconjuntos selecionados e os Originais foi similar, no entanto, utilizando um número menor de atributos. Desse modo, é possível constatar que os algoritmos de SA contribuíram de maneira significativa para o aperfeiçoamento da qualidade dos dados, e consequentemente para a construção de modelos de indução com menor custo computacional. Como trabalhos futuros, pretende-se relacionar o comportamento desses algoritmos com as propriedades dos CD, visando, por meio de métodos para Meta-Aprendizado, a recomendação de algoritmos mais apropriados para SA.

Agradecimentos

Ao Programa Institucional de Bolsas de IC (PIBIC/CNPq) e ao LABI/UNIOESTE.

Referências

1. LEE, H. D. Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado – USP, SC, 2005.
2. LIU, H.; MOTODA, H. Computational Methods of Feature Selection. Chapman & Hall/CRC data mining and knowledge discovery, 2008.
3. PARMEZAN, A. R. S. et al. Estudo Comparativo entre Métodos de Seleção de Atributos Baseados em Medidas de Precisão e Correlação Aplicados a Bases de Dados. In: XVIII SIICUSP, São Paulo, SP, 2010.
4. WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques. California, USA: Morgan Kaufmann, 2005.