Estudo e Avaliação da Seleção de Atributos para Pré-processamento no Processo de Mineração de Dados

Silvani Weber da Silva Borges[1] (PIBIC/CNPq/Unioeste), Renato B. Machado (Orientador), Newton Spolaôr[1], Huei Diana Lee[12], Wu Feng Chung[12],e-mail: silvani.borges2@gmail.com

[1]Universidade Estadual do Oeste do Paraná/Centro de Engenharias e Ciências Exatas/Laboratório de Bioinformática (LABI)/Foz do Iguaçu, PR
[2]Universidade Estadual de Campinas (UNICAMP)/
Faculdade de Ciências Médicas (FCM)/Campinas, SP

Área/subárea: Ciências Exatas e da Terra – Ciência da Computação.

Palavras-chave: descoberta de conhecimento, abordagem filtro, validação cruzada.

Resumo

Um dos benefícios que o avanço da computação tem oferecido consiste no apoio para o estudo, desenvolvimento e aplicação de técnicas inteligentes. O processo de mineração de dados utiliza algumas dessas técnicas com o intuito de descobrir conhecimento útil a algum domínio. Para aumentar as chances de sucesso do processo, os dados podem ser pré-processados por meio de algoritmos de seleção de atributos. O objetivo deste trabalho consiste no estudo e na avaliação da seleção de atributos no contexto de mineração de dados. Os experimentos realizados indicam resultados competitivos em relação a uma referência já empregada na literatura.

Introdução

Com o avanço da tecnologia da informação nas últimas décadas, é possível observar um rápido crescimento na quantidade de dados coletados nas organizações. A Descoberta de Conhecimento em Bases de Dados (DCBD) consiste na atividade de transformar grandes quantidades de dados em conhecimento útil (Galvão & Marin, 2009).

O processo de DCBD compreende as etapas de pré-processamento, mineração de dados e pós-processamento – Figura 1 (Rezende, 2003). As tarefas de pré-processamento se destacam por possibilitar a transformação dos dados para um formato mais apropriado para as demais fases. Em particular, a tarefa de Seleção de Atributos (SA) é útil para encontrar atributos – dimensões ou

características que descrevem os dados – que são não redundantes e ou relevantes para o domínio de aplicação (Liu & Motoda, 2007). Como resultado, o DCBD pode obter um melhor desempenho no aprendizado e gerar modelos (hipóteses) que representam o conhecimento adquirido com maior simplicidade para humanos.

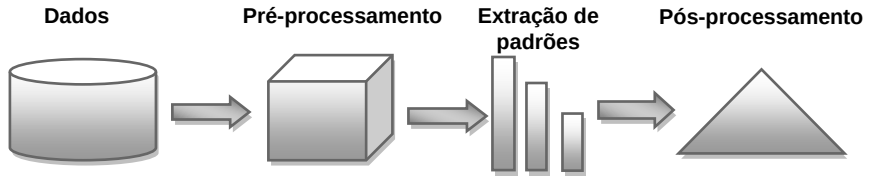

Figura 1 - Etapas do processo DCBD (adaptado de Rezende, 2003).

A avaliação de abordagens de AS em dados *benchmark* é essencial para, no futuro, aplicá-las na mineração de dados em diferentes domínios, como a área da saúde. Assim, o objetivo deste trabalho consiste no estudo e na avaliação da SA para pré-processamento de dados no DCBD.

Material e Métodos

Neste trabalho, as etapas do DCBD foram instanciadas da seguinte maneira:

- Pré-processamento: tarefa de seleção de atributos;
- Extração de padrões: tarefa de classificação;
- Pós-processamento: avaliação do desempenho da classificação.

Algoritmos de SA podem ser organizados de distintas maneiras. Uma categorização freqüente considera a interação com o algoritmo de extração de padrões, a qual ocorre conforme as abordagens filtro, *wrapper* e embutida (Liu & Motoda, 2007). A abordagem filtro se destaca pela capacidade em filtrar (remover) atributos irrelevantes e ou redundantes independentemente do algoritmo de aprendizado, o que potencialmente leva a custo computacional relativamente menor.

O algoritmo ReliefF representa neste trabalho a abordagem filtro. Sua estratégia para encontrar características relevantes envolve a comparação de valores de atributos dos vizinhos (exemplos) mais próximos de cada exemplo em uma amostra. Os atributos mais importantes são então os que melhor diferenciam vizinhos de classes distintas e aproximam vizinhos de mesma classe (Lee, 2005).

A classificação é uma das tarefas mais comumente empregadas na extração de padrões, possuindo como objetivo a predição da classe de um novo exemplo (Rezende, 2003). Um algoritmo de classificação comum em DCBD é a árvore de decisão, a qual busca gerar um modelo compreensível para humanos baseado em testes de valores de atributos organizados hierarquicamente (Galvão & Marin, 2009).

Em particular, uma árvore de decisão é constituída por: (1) nós que representam os atributos, (2) ramos que correspondem a possíveis valores desses

atributos, e (3) folhas que indicam classes (Han et al, 2011). Na Figura 2 é ilustrada parte de uma árvore, em que três atributos são testados no domínio de doenças hepáticas: Bilirrubina Direta (BD), Amilase (Ami) e Fosfatases Alcalinas (FA) (Steiner et al, 2004). Dependendo do valor de atributo, é possível predizer para um exemplo (exame de paciente) a classe cálculo no duto biliar ou câncer.

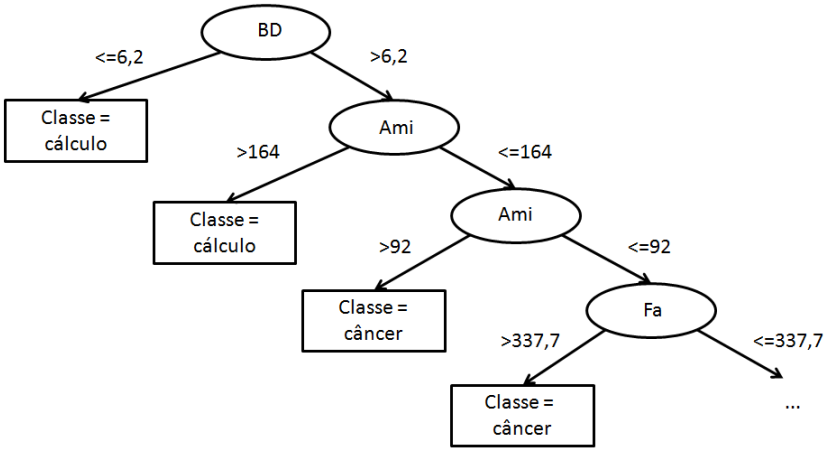

Figura 2. Parte de uma árvore de decisão para doenças hepáticas (Steiner et al, 2004).

Neste trabalho, é realizada a avaliação de árvore de decisão por meio da taxa de erro de classificação estimada via validação cruzada de 10 folds (Han et al., 2011). Quanto menor a taxa, melhor é o modelo.

Para estimar a qualidade da seleção de atributos, é realizada uma comparação do erro da árvore construída após SA (usando 25%, 50% e 75% dos atributos melhores ranqueados por ReliefF) contra o erro da árvore gerada sem SA, *i.e.*, incluindo 100% dos atributos – um *baseline* já usado na literatura (Lee, 2005).

Essa comparação é aplicada em cada um dos 10 conjuntos de dados *benchmark* obtidos do repositório UCI (https://archive.ics.uci.edu/ml/datasets.html) descritos na Tabela 1: Arrhythmia (A), Breast cancer (B), CMC (C), Dermatology (D), Heart-c (E), Hepatitis (F), Hypothyroid (G), Lung cancer (H), Post operative patient data (I) e Primary tumor (J). Em particular, para cada conjunto, são exibidos o número de exemplos (N), atributos (M), classes distintas (R) e a taxa de erro majoritário (T) – erro obtido em um cenário em que um classificador simplista sempre prediz a classe mais freqüente no conjunto de dados.

Tabela 1 – Descrição dos 10 conjuntos de dados empregados para a avaliação experimental.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
N	452	286	1473	366	303	155	3772	32	90	339
M	279	9	9	34	13	19	29	56	8	17
R	16	2	3	6	5	2	4	2	3	22
T (%)	45,79	29,72	57,30	69,40	45,54	20,65	7,71	59,38	28,89	75,22

Os algoritmos ReliefF e J48 (árvore de decisão) implementados no framework *open-source* Weka (Witten et al, 2011) foram empregados nos experimentos.

Resultados e Discussão

Na Tabela 2 são apresentados os valores médios e os desvios padrão correspondentes à taxa de erro calculada para cada porcentagem de atributos.

Tabela 2. Avaliação de árvores de decisão construídas com diferentes porcentagens de atributos.

		A	B	C	D	E	F	G	H	I	J
25%	Média	33,20	30,79	54,65	24,57	26,03	23,88	6,31	46,67	28,89	68,73
	Desvio	4,40	5,45	2,23	3,22	9,59	5,17	0,53	36,89	5,74	3,50
50%	Média	34,51	29,73	53,02	10,13	17,49	15,96	0,61	65,83	28,89	58,99
	Desvio	5,13	5,52	2,69	2,96	7,68	8,84	0,35	32,02	5,74	6,13
75%	Média	33,19	30,76	48,13	6,30	20,09	16,58	0,69	59,17	28,89	56,93
	Desvio	3,91	5,13	4,08	2,62	6,49	8,28	0,40	38,18	5,74	7,40
(***baseline***) **100%**	Média	33,40	24,45	49,70	4,66	25,70	19,25	0,45	59,17	28,89	57,22
	Desvio	4,16	3,81	3,33	2,63	7,24	9,28	0,22	38,18	5,74	5,81

É possível observar que o algoritmo de seleção de atributos contribuiu na melhoria da taxa de erro em vários casos. Nos conjuntos A, H e I, por exemplo, o uso de somente 25% dos atributos resultou em erro médio melhor ou similar ao estimado para o *baseline*. Além do desempenho melhor, o ReliefF contribuiu para, em geral, obter árvores de decisão menores e mais simples nesses conjuntos. Convém ressaltar que os fatos do conjunto A possuir uma quantidade relativamente alta de atributos e do conjunto I igualar o erro majoritário (Tabela 1) merecem ser estudados no futuro para melhor relacionar características dos dados com a qualidade de classificação.

Nos conjuntos E e F, a SA se destacou com o uso de 50% e 75% dos atributos, enquanto que bons resultados foram encontrados nos conjuntos C e J após a seleção de 75% dos atributos melhores ranqueados.

Entretanto, nos conjuntos de dados B, D e G, o desempenho médio das árvores geradas após aplicação do ReliefF foi pior do que o *baseline*. Esses resultados motivam o estudo e a avaliação futura de outras tarefas de pré-processamento de dados. Um exemplo consiste no tratamento de dados faltantes (Purwar & Singh, 2015), um problema que afeta G e outros conjuntos.

Conclusões

A avaliação do algoritmo ReliefF realizada neste trabalho sugere a competitividade da tarefa de seleção de atributos no pré-processamento de diferentes conjuntos de dados. De fato, em 7 dos 10 conjuntos avaliados, foram

obtidas taxas médias de erro de classificação melhores ou similares às taxas obtidas pelo *baseline*. Além dos benefícios na qualidade de aprendizado, convém destacar a contribuição da SA para reduzir a dimensionalidade dos dados, o que pode levar a árvores de decisão mais simples e compreensíveis por humanos.

Agradecimentos

A UNIOESTE/CNPq pela concessão de bolsa de iniciação científica.

Referências

Galvão, N.D. & Marin, H.F. (2009). Técnicas de mineração de dados: uma revisão da literatura. *Acta Paulista Enfermagem* **22**, 686 -690.

Han, J., Kamber, M. & Pei, J. (2011). *Data mining: concepts and techniques*. New York: Elsevier.

Lee, H.D. (2005). *Seleção de atributos importantes para extração de conhecimento de bases de dados*. Tese de doutorado. Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional, Universidade de São Paulo.

Liu, H. & Motoda, H. (2007). *Computational methods of feature selection*. Boca Raton: Chapman & Hall/CRC.

Purwar, A. & Singh, S.K. (2015). Hybrid prediction model with missing value imputation for medical data. *Expert Systems* with Applications **42**, 5621- 5631.

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. São Paulo: Manole.

Steiner, M.T.A., Soma, N.Y., Shimizu, T., Nievola, J.C., Smiderla, A. & Lopes, F.M. (2004). Data mining como suporte à tomada de decisões uma aplicação no diagnóstico médico. In Anais do 36º Simpósio Brasileiro de Pesquisa Operacional, São João Del Rei, Minas Gerais, Brasil.

Witten, I.H.; Frank, E. & Hall, M.A. (2011) *Data mining: practical machine learning tools and techniques*. São Francisco, EUA: Morgan Kaufmann Publishers.