Metodologias para a Seleção de Atributos Relevantes

José Augusto Baranauskas e **Maria Carolina Monard**
Departamento de Computação e Estatística
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo
Av. Carlos, Botelho, 1465 – Caixa Postal 668 – CEP 13560-970 – São Carlos - SP
e-mail: {jaugusto,mcmonard}@icmsc.sc.usp.br

Resumo

Tanto os algoritmos de Aprendizado de Máquina Simbólico por exemplos mais utilizados quanto Redes Neurais usando Backpropagation não possuem desempenho satisfatório quando aprendem utilizando exemplos com muitos atributos. Mesmo sob o ponto de vista estatístico, exemplos com muitos atributos irrelevantes e com ruídos fornecem pouca informação. Basicamente, na maioria dos casos, os algoritmos ficam confusos na presença de muitos atributos e constróem classificadores com pouca utilidade. Neste trabalho são descritas algumas formas para a extração de atributos relevantes, usando-se uma delas para comparação da precisão de classificação em alguns conjuntos de dados. O desempenho de tais algoritmos de Aprendizado de Máquina Simbólicos utilizando todos os atributos e somente os atributos relevantes é apresentada bem como trabalhos futuros nessa área.

I. Introdução

A metodologia tradicional de transformar dados em informação útil — conhecimento — baseia-se na análise e interpretação manual. Por exemplo, numa companhia de seguros de saúde, é comum os especialistas analisarem as tendências e alterações nos dados de saúde de seus clientes; a partir daí, eles geram um relatório que será usado para decisões futuras quanto à forma e custos de atendimento. Assim, o método clássico de análise de dados reside em um ou mais analistas humanos tornando-se intimamente familiar com os dados e atuando como uma interface entre dados e usuários. Esta forma manual de tratamento de dados é lenta, cara e altamente subjetiva. Na medida que as informações armazenadas e disponibilizadas pelos computadores atuais crescem cada vez mais, essa abordagem torna-se impraticável em vários domínios. Isso pode ser facilmente explicado sabendo-se que os dados (ou bases de dados ou *datasets*) atuais crescem, fundamentalmente, de duas formas:

1. o número de registros **N**, ou instâncias ou objetos no banco de dados e
2. o número de atributos **a,** ou campos de um objeto

Bancos contendo registros da ordem de $N = 10^9$ objetos tornam-se cada vez mais comuns, por exemplo, em Astronomia. De forma similar, o número de campos pode facilmente atingir a ordem de $a = 10^2$ ou mesmo $a = 10^3$ em aplicações, por exemplo, de diagnóstico médico [Teller 95]. Nos últimos anos a utilização de computadores para resolver tal situação levou ao desenvolvimento de uma nova área de pesquisa denominada *Data Mining* — *DM* — parte de um processo mais amplo denominado *Knowledge Discovery on Databases* — KDD [Fayyad 96].

Uma das formas de adquirir conhecimento em DM é através da utilização de algoritmos de Aprendizado de Máquina — AM. A metodologia geral utilizada nestes casos consiste em retirar várias amostras representativas da base de dados que são apresentadas ao(s) algoritmo(s) de AM. Posteriormente, é realizada uma combinação do conhecimento adquirido pelos algoritmos de AM usando essas amostras, tendo esta metodologia mostrado-se promissora.

Entretanto, em muitas situações não é simples escolher ou realizar o processo de redução da dimensão dos atributos pois não se sabe exatamente quais os atributos mais relevantes. Por exemplo, o aprendizado de regras de diagnóstico para diferentes doenças proveniente de vários registros médicos. Esses registros, freqüentemente, apresentam muito mais informação (atributos) daquela que é realmente necessária para descrever cada doença. No caso específico para aprender o diagnóstico individual entre uma doença cardíaca e uma doença hormonal, sabe-se que os atributos relevantes para uma não são, necessariamente, os mesmos para a outra. Com receio que algo seja perdido durante o processo de aprendizado, normalmente decide-se incluir todos os atributos e deixa-se que o algoritmo de AM selecione aqueles mais importantes.

Todavia, na maioria dos casos, os algoritmos de AM ficam *confusos* com muitos atributos e constróem classificadores com pouca utilidade. Assim, é importante pesquisar métodos para selecionar atributos relevantes. O objetivo deste trabalho é mostrar uma comparação da precisão do algoritmos C4.5 e CN2 com e sem seleção de atributos relevantes.

II. Seleção de Atributos

Os algoritmos de AM comumente utilizam *datasets* contendo poucos exemplos (N < 20.000) com número restrito de atributos (a < 30). Quando se trabalha com uma dimensão restrita é possível usar um algoritmo de seleção de atributos que simplesmente procura pelas possíveis combinações, selecionando aqueles que melhoram a taxa de classificação do algoritmo de AM.

Mesmo considerando o fato que os algoritmos de AM possam ser utilizados com muitos atributos, sabe-se que o desempenho dos indutores e redes neurais artificiais (RNA) usando *backpropagation* é prejudicado quando existem muitos atributos irrelevantes [Dietterich 97]. Além disso, estatisticamente, exemplos com muitos atributos e ruídos fornecem pouca informação.

Assim, é necessário pesquisar algoritmos para selecionar os atributos mais relevantes tal que a precisão da classificação utilizando este subconjunto de atributos melhora ou, a menos mantém-se no mesmo nível que utilizando todos os atributos do *dataset*. Basicamente, existem 3 métodos de seleção de atributos:

1. Filtro: efetua uma análise inicial dos dados e seleciona os atributos relevantes para alimentar um algoritmo de aprendizado
2. Embutido: integra seleção e ponderação dos atributos no próprio algoritmo de aprendizado
3. Wrapper [John 94]: seleciona e testa diferentes subconjuntos de atributos com o algoritmo de aprendizado, escolhendo o melhor subconjunto de todos

A fim de mostrar o desempenho de filtros na seleção de atributos relevantes, foi utilizada a ferramenta MineSet ™ contendo como base a biblioteca MLC++ [Kohavi 94] utilizando conjuntos de dados extraídos do projeto Statlog [Taylor 94] (dna e genetics) e de UCI [Merz 98] (sonar). A caracterização dos *datasets* utilizados é dada na Tabela 1.

Dataset	# Instâncias	# Atributos (contínuos,discretos)	Classe	% Classe	% Erro Majoritário
dna	3186	180 (0,180)	1	24.07%	48.09% no valor 3
			2	24.01%	
			3	51.91%	
genetics	3190	60 (0,60)	N	51.88%	48.12% no valor N
			EI	24.04%	
			IE	24.08%	
sonar	208	60 (60,0)	M	53.37%	46.63% no valor M
			R	46.63%	

Tabela 1: Caracterização dos *datasets* originais utilizados

Após obtido o subconjunto de atributos relevantes para cada *dataset*, eles foram submetidos aos algoritmos de AM Simbólico C4.5 e CN2. A Tabela 2 mostra o subconjuntos obtidos usando todos os atributos e os subconjuntos obtidos por FSS. Para medir a precisão foram utilizados *10-fold crossvalidation* (cv) e *10-fold stratified-crossvalidation* (strat-cv). A diferença entre cv e strat-cv consiste em que, no último, os subconjuntos de instâncias (folds) são estratificados de maneira a conter as mesmas proporções de cada classe que no *dataset* original. Na coluna *Atributos Selecionados* são também mostrados os atributos relevantes selecionados, numerados a partir de zero, seguindo a ordem original do *dataset*.

Dataset	Atributos Selecionados	Medida	Precisão Classificação C4.5	Precisão Classificação CN2
dna	todos	10-cv	92.50% +- 0.63%	88.20% +- 0.80%
		10-strat-cv	92.40% +- 0.46%	88.15% +- 0.62%
	68 81 83 84 89 92 93 94 95 96 99 104 128 140 7,78% do total de atributos	10-cv	94.63% +- 0.49%	94.13% +- 0.62%
		10-strat-cv	94.63% +- 0.44%	94.16% +- 0.54%
genetics	todos	10-cv	94.08% +- 0.52%	76.39% +- 1.65%
		10-strat-cv	94.17% +- 0.39%	79.53% +- 1.45%
	15 20 27 28 29 30 31 34 35 15.00% do total de atributos	10-cv	94.36% +- 0.45%	83.81% +- 1.74%
		10-strat-cv	94.42% +- 0.35%	83.41% +- 1.66%
sonar	todos	10-cv	67.83% +- 2.79%	73.16% +- 3.44%
		10-strat-cv	69.74% +- 1.97%	71.19% +- 3.30%
	3 10 30 35 45 50 51 11,67% do total de atributos	10-cv	82.24% +- 2.68%	70.67% +- 3.69%
		10-strat-cv	83.19% +- 2.30%	75.98% +- 3.00%

Tabela 2: Classificação com e sem seleção de atributos relevantes

Pode ser notada que a precisão da classificação é dependente do *dataset* e algoritmo de AM utilizado. Além disso, geralmente obtém-se uma boa precisão considerando apenas uma fração dos atributos originas (15.00% no *dataset* genetics).

Dos 3 métodos de seleção de atributos já descritos, o Embutido envolve modificação nos algoritmos de AM já existentes; *Wrapper* tem sua utilização dificultada quando se tem muitos atributos e muitos exemplos, pelo alto custo computacional. O método de filtrar aparenta ser o mais indicado para seleção de atributos com grandes *datasets*, além de ser independente do algoritmo de AM utilizado, como mostra a Figura 1, ao contrário das duas outras abordagens.

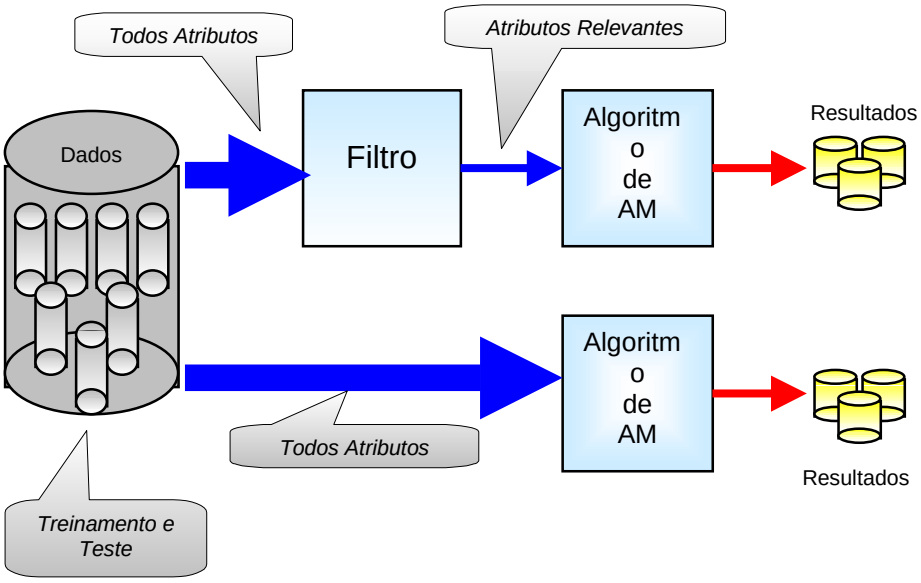

Figura 1: Filtrar atributos irrelevantes não requer alteração no algoritmo de AM a ser utilizado.

III. Conclusão

A utilização de seleção de atributos em conjunto de dados com muitos atributos demonstra ser benéfica ao processo de aprendizado de duas maneiras:

a) reduzindo a dimensão de um conjunto de dados — permitindo o uso de algoritmos de AM que não conseguiriam trabalhar com o conjunto inteiro de atributos e

b) aumentando, geralmente, a precisão da classificação

Um outro ponto a ser notado consiste na qualidade das regras obtidas. Estudos preliminares mostram que, geralmente, a qualidade das regras obtidas filtrando atributos é melhor do que quando todo o *dataset* original é usado.

Existem algoritmos que filtram atributos, por exemplo Relief-F [Kononenko 94] que seleciona atributos fracamente e fortemente relevantes; Focus-2 [Almuallim 97] seleciona somente os atributos fortemente relevantes. De toda forma, os filtros atuais têm seu tempo de execução polinomial ou exponencial, inadequados para *datasets* com muitos atributos, típicos em KDD.

Atualmente, como parte da pesquisa sendo realizada no programa de Doutorado, estão sendo projetados novos algoritmos para seleção de atributos que filtrem os atributos que não são relevantes, ou seja, aqueles totalmente irrelevantes. Esses algoritmos se aplicarão a grandes bases

dados e deverão ser rápidos, embora não sejam perfeitos. A idéia básica consiste remover aqueles atributos irrelevantes como passo inicial. Se necessário, pode-se então utilizar o subconjunto obtido nos algoritmos existentes.

IV. Referências

[Almuallim 97] Almuallim, H.; Dietterich, T.G.; *Efficient Algorithms for Identifying Relevant Features*, Oregon State University, ftp://ftp.cs.orst.edu/pub/tgd/papers.

[Dietterich 97] Dietterich, T.G.; *Machine Learning Research: Four Current Directions*, Draft of May 23, 1997, Oregon State University, ftp://ftp.cs.orst.edu/pub/tgd/papers.

[Fayyad 96] Fayyad, U. M.; Djorgovski, S.G.; Weir, N.; *From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey*, AI Magazine, 17(2), 1996, pp. 51-66.

[John 94] John, G.; Kohavi, R.; Pfleger, K., *Irrelevant Features and the Subset Selection Problem*, Em *Proceedings of the Tenth International Conference on Machine Learning*, pp 167-173, San Francisco, CA, Morgan Kaufmann.

[Kohavi 94] Kohavi, R. et all; *MLC++: A Machine Learning Library in C++*, in Tools with Artificial Intelligence, IEEE Computer Society Press, 1994, pp. 740-743.

[Kononenko 94] Kononenko, I; *Estimating Attributes: Analysis and Extensios of Relief*. Em *Proceedings of the 1994 European Conference on Machine Learning*, pp. 171-182, Amsterdam, 1994, Springer-Verlag.

[Merz 98] Merz, C.J.; Murphy, P.M.; *UCI Repository of Machine Learning Datasets*, University of California, Irvine, CA, 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html

[Taylor 94] Taylor, C.; Mitchie, D.; Spiegalhater, D.; *Machine Learning, Neural and Statistical Classification*, Paramount Publishing International, 1994.

[Teller 95] Teller, A.; Veloso, M.; *Program Evolution for Data Mining*, International Journal of Expert Systems, Vol. 8, No. 3, pp. 213-236.