Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais
pp. 127-130, June 2–5, 2003 – Centro Universitário da FEI, São Paulo - SP - Brazil

Seleção de Variáveis e Redes Neurais: Uma Aplicação de Classificação de Risco de Evento Adverso em Leucemia Infantil

L. Macrini[(1)] C.E.Pedreira[(1)] E.S.Costa[(2)] M. Land[(2)]
[1]Departamento de Engenharia elétrica; Universidade Católica do Rio de Janeiro PUC-RIO;
[2]Escola de Medicina; Universidade Federal de Rio de Janeiro UFRJ
E-mails: macrini@ele.puc-rio.br,pedreira@ele.puc-rio.br

Abstract

Neste artigo trabalhamos com dois métodos de seleção de variáveis a serem utilizadas como variáveis de entrada em uma Rede neural Feedforward, cujo objetivo é avaliar o risco de evento adverso (recaída ou morte) em um grupo de crianças portadoras de Leucemia Linfoblástica Aguda. Este tipo de problema na área médica é bastante importante visto que os casos disponíveis são sempre em número reduzido e a escolha de variáveis pertinentes é fundamental na performance do modelo. O desempenho obtido mostrou resultados excelentes para generalização (abordagem leave-one-out), alcançando uma taxa de acerto superior a 96% para todas as simulações.

1. Introdução

Seleção de variáveis tem um papel fundamental na classificação de sistemas como redes neurais. Quando trabalhamos com grande quantidade de variáveis é bastante útil quando podemos separar atributos irrelevantes ou redundantes dos pertinentes. Neste caso, selecionando somente atributos pertinentes, estaremos trabalhando com características quantitativas ou qualitativas que identificam membros de um conjunto de dados observados de forma parcimoniosa. Primeiramente desejamos utilizar as variáveis de entrada que forneçam a máxima quantidade de informação na saída, o que não ocorre quando essas variáveis são redundantes ou irrelevantes. Segundo porque o número de parâmetros associados a um modelo esta diretamente relacionado ao tamanho do conjunto dos dados. Trabalhar com um conjunto grande de atributos implica em um maior número de parâmetros do modelo a serem estimados. Este ponto é particularmente importante na maioria dos problemas médicos como, por exemplo, em diagnóstico de câncer e modelos de estimação de risco, onde, usualmente, o conjunto de dados é, em geral, limitado. Nestes casos, os dados são compostos por paciente que tenham sido parte de uma mesma experiência clínica, tenham sido tratados com o mesmo protocolo, etc.

Vários autores têm pesquisado sobre o problema de seleção de variáveis. Um dos métodos mais populares é a análise de componentes principais, PCA (Joliffe, 1986 [7]), que criam novas variáveis de entrada pelo processamento das variáveis físicas originais. Porém, quando precisamos preservar os dados originais este método não é satisfatório. Outros métodos importantes que podemos mencionar são: árvores de decisão (Setiono e Liu, 1997 [11]) e regressão stepwise (Breiman, et al., 1984 [3]; Draper e Smith, 1981 [5]), onde as variáveis pertinentes são encontradas de modo iterativo, e mais recentemente métodos baseados na teoria da informação (Battiti, 1994 [1]; Kwak e Choi, 2002 [8]).

Neste artigo, iremos trabalhar com uma aplicação na avaliação do risco de evento adverso (recaída ou morte) em um grupo de crianças portadoras de Leucemia Linfoblástica Aguda (LLA) através de um método que utiliza a informação mútua para seleção de variáveis de entrada. LLA é a neoplasia mais comum na infância e, atualmente, a taxa de cura alcança de 75-80% dos casos em 5 anos. Apesar disso, 20-25% dos pacientes terão um evento adverso num período de até 5 anos do diagnóstico. O índice de recaída em pacientes com LLA é maior que o número de todos os outros tipos de câncer na infância. Sendo assim, um sistema de avaliação do risco de evento adverso é considerado a característica mais importante na definição da estratégia a ser usada no tratamento. Dessa forma, a intensidade do tratamento a ser aplicada esta diretamente relacionada a este risco. Nos anos setenta, informações clínicas e de laboratório no diagnóstico da doença começaram a ser associadas a este risco. Embora o número de trabalhos publicados nesta área seja grande, existem discordâncias significantes entre os protocolos terapêuticos mais importantes. Além disso, o erro na estimação deste risco ainda é inferior ao nível desejável.

O objetivo principal deste artigo é avaliar o risco de evento adverso em um grupo de crianças brasileiras diagnosticadas com LLA. As variáveis de entrada utilizadas em nosso problema são compostas por um conjunto de dados clínicos e biológicos obtido na hora do diagnóstico. Este risco é um fator crítico na definição do nível da agressividade do tratamento.

2. Metodologia

Originalmente nosso banco de dados era composto por 78 possíveis variáveis de entrada num total de 128 casos. Aplicamos um filtro inicial ao banco de dados

para eliminar as variáveis não informativas e os fatores, isto é, variáveis criadas em função de uma ou mais variáveis e aquelas que estavam de uma certa forma diretamente relacionada a variável de saída e por fim os casos com dados faltantes. Após o filtro nós ainda contávamos com 41 possibilidades de variáveis de entrada num total de 116 pacientes. Como o número de variáveis ainda é bastante significativo em relação ao número de casos aplicamos dois métodos para executar a redução das variáveis de entrada excluindo variáveis irrelevantes ou redundantes dos dados. Os métodos aplicados foram de Seleção de Característica por Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U) (Kwak e Choi, 2002 [8]), e o método de Taguchi (Peterson *et al.*, 1985 [10]). O método MIFS-U é uma variante melhorada do MIFS proposto por (Battiti, 1994 [1]) e é baseado na idéia de se escolher as variáveis de entrada que maximizam a informação mútua (Cover e Thomas, 1991 [4]) entre estas variáveis e a saída.

A informação mútua pode ser definida como

$$I(X,Y) = \sum_{x \notin \chi} \sum_{y \in \gamma} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

sendo uma medida de dependência entre variáveis aleatórias. Note que
$I(X,Y) \geq 0$ e $I(X,Y) = 0$ se X e Y são estatisticamente independente.

Um valor alto (pequeno) de informação mútua significa que as variáveis são muito (pouco) relacionadas.

O método de Taguchi (Taguchi, 1993 [12]) está baseado no trabalho publicado por Fisher's nos anos trinta (Fisher, 1935 [6]) e diz respeito a métodos experimentais. Sua aplicação em redes neurais pode ser encontrada em (Peterson *et al*., 1995[10]).

Se um experimento é composto de N variáveis, para considerar todas as possibilidades de seleção de variáveis de entrada teríamos que verificar todas as possibilidades de combinação executando 2^N experiências. Em nosso caso nós teríamos que trabalhar com 2^{41} Redes neurais para selecionar a combinação de variáveis de entrada que fornecesse a melhor performance. Logicamente esse procedimento é impossível de ser realizado do ponto de vista prático.

O método de Taguchi só requer uma fração de todas as combinações possíveis das N variáveis e é usado para encontrar variáveis de entrada que influenciem o desempenho da variável de saída. Primeiramente são selecionadas algumas variáveis que podem influenciar o desempenho de uma saída utilizando, por exemplo, o algoritmo MIFS-U. O método de Taguchi trabalha com a informação de que a efetividade de uma variável depende da presença ou não de outras variáveis. Nesse sentido os candidatos de entrada são organizados em níveis e boa parcela dos 2^N experimentos não precisam ser realizados.

Após o processo de seleção das variáveis de entrada nós usamos um modelo de Rede neural Feedforward para classificar o risco de recaída.

3. Resultados Numericos

O banco de dados original é composto de 78 variáveis (41 após a remoção de correlações e de variáveis não informativas) de 128 crianças (116 após a remoção dos dados faltantes) com Leucemia Linfoblástica Aguda (LLA) diagnosticadas e tratadas nos hospitais IPPMG/UFRJ e HUPE/UERJ. Estas variáveis (candidatas à entrada) podem ser organizadas basicamente do seguinte modo: a) Identificação - idade em diagnóstico; sexo; raça; data do diagnóstico; data de nascimento; b) dados clínicos à diagnóstico - presença ou ausência de perda de peso; febre; anemia; artrite; hemorragia; aumento de linfonodos; aumento do fígado e baço; foco infeccioso; dor abdominal; dor óssea; massa mediastinal; massa abdominal; c) dados laboratoriais à diagnóstico - contagem de células vermelhas do sangue; concentração de hemoglobina; hemácias; contagem de células brancas do sangue; contagem de blastos do sangue periférico; contagem de plaquetas; soro LDH; porcentagem de blastos na medula óssea; imunofenotipos de blastos e anormalidades citogenética em blastos; d) protocolo de tratamento. Os dados de desfecho (saída), no caso, presença ou ausência de evento adverso (recaída ou morte), é usado como o resultado do modelo.

Levando-se em consideração que a incidência de LLA em crianças é de aproximadamente 1 caso por 100 mil habitantes por ano, um conjunto de dados com 128 casos pode ser considerado aproximadamente como o número de casos em uma população de 2.5 milhões durante 5 anos. Como nesses tipos de problema médico o número de casos serão sempre reduzidos, os métodos de seleção de variáveis tem papel fundamental na seleção das variáveis pertinentes.

A tabela 1 mostra a seleção das doze variáveis mais pertinentes em ordem de importância utilizando os dois métodos de seleção de variáveis. Blasto é a variável mais importante segundo o método MIFS-U e a variável idade pelo método de Taguchi.

Tabela 1. Doze variáveis selecionadas

MIFS_U	Taguchi
Contagem de blastos	**Idade**
Idade	**Raça**
Raça	**Sexo**
Tamanho do baço	**Contagem de blastos**
Morfologia L1	**Tamanho de baço**
Protocolo	**Morfologia L1**
Sexo	Protocolo
Dor torácica	Dor torácica
Artrite	Artrite
Dor articulação	Dor articulação
Febre	Febre
Palidez	Palidez

Podemos notar que a ordem de seleção das variáveis pertinentes pelo dois métodos é bem semelhante. Houve uma pequena inversão de posições na seleção inicial das variáveis, mas de uma certa forma a seleção pelo método de Taguchi preservou a escolha das variáveis realizada pelo método MIFSU.

Após esse processo de seleção das variáveis de entrada mais pertinentes nós usamos um modelo de Rede neural Feedforward para classificar os eventos adversos até cinco anos depois do diagnóstico. Dessa forma, a saída (desfecho) do modelo é zero ou um para evento adverso ou não respectivamente.

Consideramos como variáveis de entrada da rede as seis primeiras variáveis mais pertinentes selecionadas pelos dois métodos e utilizamos para o modelo interno da rede a Regularização Bayesiana (Mackay, 1992 [9]). Em função do limitado número de casos disponíveis aplicamos o método de leave-one-out (Bishop, 1995 [2]), em cem simulações, para testar a performance de generalização. A tabela 2 mostra o resultado médio das 100 simulações realizadas e o desvio padrão encontrado por ambos os métodos (MIFS-U e Taguchi). Como podemos observar os resultados foram bastante significativos. A média de acerto tanto para os casos 'tipo 0', evento adverso e para os casos 'tipo 1', eventos não-adverso, foi de 98.03% em ambos os métodos.

Tabela 2 - Taxas de estimação correta do risco de recaída por MIFS-U e Taguchi para as seis melhores entradas selecionadas (da Tabela 1)

	MIFS-U		Taguchi	
	Média	Desvio	Média	Desvio
Tipo0 correto (%)	96.18	4.06	96.64	2.61
Tipo1 correto (%)	99.59	1.09	99.18	1.41
Média (%)	98.03	2.08	98.03	1.86

Com o intuito de validar o método de seleção de variáveis realizamos mais duas experiências onde na primeira experiência selecionamos, de forma aleatória, seis variáveis, e na segunda experiência,acrescentamos duas variáveis as seis primeiras variáveis selecionadas pelo método de Taguchi.

A Tabela 3 mostra as seis variáveis selecionadas de forma aleatória e na Tabela 4 apresentamos os resultados encontrados empregando a mesma metodologia usada na experiência original (Tabela 2).

Como era de se esperar, os resultados são significativamente bem inferiores aos encontrados originalmente pelos métodos MIFS-U e Taguchi.

Tabela 3. Escolha aleatória de seis entradas

Idade
Tamanho do fígado
Tamanho do baço
Raça
Infecção pulmonar
Massa torácica

Tabela 4 - Taxas de estimação correta do risco de recaída

	Média	Desvio
Tipo 0 correto (%)	69.98	8.38
Tipo 1 correto (%)	69.68	13.88
Média (%)	69.83	10.56

Na segunda experiência acrescentamos mais duas variáveis escolhidas aleatoriamente as primeiras seis variáveis mais pertinentes selecionadas pelo método de Taguchi (Tabela 1), isto é, porcentagem de blastos e infecção. A tabela 5 mostra o novo conjunto de 8 variáveis de entrada e os resultados das 100 simulações são mostrados na tabela 6.

Tabela 5 - Duas variáveis selecionadas aleatoriamente adicionadas as seis selecionadas pelo método de Taguchi

Idade
Raça
Sexo
Contagem de blastos
Tamanho do baço
Morfologia L1
% Blastos
Infecção

Tabela 6 - Taxas de estimação correta do risco de recaída para a Tabela 5 de Taguchi com duas variáveis escolhidas aleatoriamente

	Média	Desvio
Tipo 0 correto (%)	53.03	0.40
Tipo 1 correto (%)	72.16	1.67
Média (%)	63.41	0.41

Da mesma forma podemos notar que os resultados são pobres comparados aos da Tabela 2, isto é, aos resultados encontrados pelos dois métodos de seleção de variáveis. Isso vem confirmar os pontos destacados acima na introdução e apontam à real necessidade de uma seleção de variáveis de entrada, especialmente no tipo de problema médico abordado neste artigo onde o número de padrões de entrada é sempre bastante reduzido.

4. Observações Finais

Nesse artigo abordamos uma aplicação dos métodos de seleção de variáveis chamada Informação Mutua para Seleção de Características sob Distribuição de Informação Uniforme e método de Taguchi para seleção de variáveis de entrada em Redes neurais com o objetivo de se avaliar o risco de evento adverso em um grupo de crianças com Leucemia Linfoblástica Aguda (LLA). A avaliação correta do risco desta doença é um fator decisivo na estratégia de tratamento a ser adotado. Os dois métodos mencionados acima foram usados para selecionar as seis variáveis mais pertinentes como variáveis de entrada em um modelo de Rede neural para avaliar o risco num grupo de 128 crianças com LLA. Em problemas médicos como mostrado aqui, onde os números disponíveis de padrões de entrada são sempre muitos reduzidos, a escolha do conjunto de variáveis de entrada se faz muito importante. Os resultados encontrados para a saída da rede utilizando os métodos de seleção foram comparados com o resultado de seleção aleatória de seis variáveis do conjunto total de variáveis disponíveis de variáveis de entrada e também com o acréscimo de duas variáveis as seis variáveis mais pertinentes previamente selecionadas pelo método de Taguchi. O desempenho foi, nas duas experiências, como esperado, significativamente mais pobre. Para finalizar, é importante ressaltar que o percentual de acerto obtido do risco de evento adverso mostrou resultados excelentes para generalização, alcançando mais de 96% de acerto em todas as simulações. Estes resultados, com dados reais indicam claramente uma real potencialidade para este tipo de aplicação.

Referências

[1] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Networks,* Vol. 5, pp. 537-550.

[2] Bishop, C. M. (1995). *Neural Networks for Pattern Recognittion*. Oxford. Clarendon Press.

[3] Breiman, L., J. H. Friedman, R. A. Olshen and C. Stone (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

[4] Cover, T. M. And J. A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

[5] Draper, N. R. And H. Smith (1981). *Applied Regression Analysis*. 2nd. New York: Wiley.

[6] Fisher, R. A. (1935). *The Desihn of Experiments*. Edinburgh, U. K.:Oiver and Boyd.

[7] Joliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

[8] Kwak, N. and C. Choi (2002). Input Feature Selection for Classification Problems. *IEEE Trans. Neural Networks,* Vol. 13, no.1, pp. 143-159.

[9] Mackay, D. (1992). Bayesian Interpolation. *Neural Computation,* Vol. 4, pp. 415-447.

[10] Peterson G. E. P. *et al*. (1995). Using Taguchi's method of experimental design to control errors in layered perceptrons. *IEEE Trans. Neural Networks,* Vol. 6., pp. 797-799.

[11] Setiono, R. And H. Liu (1997). Neural network feature selector. *IEEE Trans. Neural Networks,* Vol. 8, pp. 654-661.

[12]Taguchi, G. (1993). *Taguchi on Robust Technology Development.* New York Amer. Soc. Mech. Eng.