Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

Aplicação de Redes Neurais Artificiais em Dados Epidemiológicos de Hepatite A

Alcione Miranda dos Santos[1], Basílio de Bragança Pereira[1,2]
Roberto de Andrade Medronho[2], Mônica Rodrigues Campos[2]
José Manoel de Seixas[3], Luiz Pereira Calôba[3]
[1]COPPE-UFRJ
[2]Faculdade de Medicina-NESC-UFRJ
[3]Laboratório de Processamento de Sinais-COPPE-EE-UFRJ
E-mails: amiranda@pep.ufrj.br,seixas@lps.ufrj.br

Abstract

This work aims at developing a supporting system for medical diagnosis of type A hepatitis. For this, logistic regression and artificial neural network models are considered. Perfomance is measured from classification indexes on sample persons from an area in Rio de Janeiro Estate that suffers seriously from such disease. Results show that neural model out perfoms logistic regression and achieves an overrall classification efficiency above 80%.

1. Introdução

Na grande maioria dos problemas da área epidemiológica, várias características em um mesmo indivíduo (paciente) podem ser observadas. Além disso, existem alguns fatores que influenciam uma determinada doença, por exemplo, tempo e espaço.

Para estudarmos esses problemas, algumas técnicas estatísticas clássicas são utilizadas; entre elas, podemos citar a regressão logística[1]. Entretanto, essa técnica possui restrição, por exemplo, temos que supor que as variáveis explicativas não são correlacionadas[1]. Se essa suposição é violada, o modelo de regressão logística fornece resultados errôneos.

Ultimamente, redes neurais artificiais[2] (RNA) vêm sendo muito utilizadas no auxílio ao diagnóstico e terapêutica médica. Devido ao fato de não haver necessidade de independência e normalidade das variáveis em estudo, bem como a grande capacidade de aprendizado a partir do ambiente, a aplicação de redes neurais artificiais na análise estatística de dados epedimiológicos tem tido grande aceitação. Além do mais, o processamento neural é capaz de extrair correlações das variáveis de entrada diretamente sobre os epaços de dimensão elevada que tipicamente as caracterizam, tornando tal processamento uma ferramenta valiosa em problemas complexos de reconhecimento de padrões.

A região em estudo neste trabalho envolveu uma localidade denominada, pelo Programa de Despoluição da Baía de Guanabara, de *Setor Parque Fluminense*, que abrange uma parte do segundo distrito do município de Duque de Caxias. Este município está localizado na região metropolitana do Estado do Rio de Janeiro, às margens da Baía de Guanabara, fazendo divisa com os municípios do Rio de Janeiro, Belfort Roxo, São João de Meriti, Miguel Pereira, Nova Iguaçu, Magé, Vassouras e Petropólis. Este trabalho faz parte de um projeto do Núcleo de Saúde Coletiva da Universidade Federal do Rio de Janeiro e os dados aqui tratados foram gentilmente cedidos pelos coordenadores de tal projeto.

No setor Parque Fluminense, foram selecionados 19 setores censitários para realização de um inquérito sorológico, para anticorpos contra vírus da hepatite A (anti-HAV), e um inquérito domiciliar, para avaliação de condições sócio-econômica e sanitárias. A amostragem dos indivíduos foi estratificada por grupo etário, tendo sido selecionados 3.079 indivíduos com idade entre 1 e 83 anos, residentes em 2.291 domicílios.

O exame sorológico foi realizado após esclarecimento à população dos objetivos do trabalho e assinatura de um termo de consentimento. As amostras de sangue para a pesquisa de anti-HAV total no soro foram colhidas em todos os indivíduos amostrados. Os indivíduos são classificados como soropositivo ou soronegativo.

As variáveis relacionadas ao indivíduo, seu ambiente domiciliar e peridomiciliar, foram coletadas por meio de entrevistas domiciliares, através de um formulário précodificado. Foram observadas 66 variáveis, entre elas podemos citar: idade, sexo, renda mensal e escolaridade da dona de casa, a não utilização de filtro de água, densidade de moradores por cômodos, tempo de moradia na residência, condições sanitárias e outras.

Tivemos como primeiro objetivo desse estudo identificar a soroprevalência global para a anti-HAV (hepatite A) entres os indivíduos observados. Em seguida, identificar quais as variáveis que influenciam no diagnóstico da doença.

Para estabelecermos a influência conjunta das variáveis sobre a soroprevalência anti-HAV, utilizou-se a técnica de regressão logística multivariada [3].

Após termos adotado um modelo de regressão logística capaz de predizer e classificar os dados em estudo, projetamos uma rede neural artificial, com o objetivo

de comparar os resultados obtidos pelo modelo logístico com os resultados fornecidos pela rede neural.

2. Materiais e Métodos

2.1. Definição das Amostras

Um dos primeiros passos do processo de desenvolvimento de redes neurais artificiais é a definição da amostra de treinamento, de validação e de teste. A amostra de treinamento será utilizada no treinamento da rede, enquanto a amostra de validação será utilizada para verificar a eficiência da rede quanto a sua capacidade de generalização durante o treinamento, e podendo ser empregada como critério de parada do treinamento. Por fim, a amostra de teste, será utilizada para verificar a performance da rede neural sob condições reais de utilização.

Para definirmos a amostra de treinamento, de validação e de teste a serem utilizadas na rede neural artificial bem como, no modelo de regressão logística, foram retirados da amostra em estudo os indivíduos com dados incompletos. Assim, a amostra total a ser utilizada será composta de 2.815 indivíduos. Nessa amostra, através de diagnóstico médico, verificou-se uma soroprevalência global para hepatite A de 36,6%.

A amostra selecionada para treinamento é composta de 1.200 indivíduos, sendo que 400 indivíduos são soropositivos para a anti-HAV e os restantes são soronegativos para a anti-HAV. A amostra a ser utilizada na validação é composta de 762 indivíduos dos quais 290 são soropositivos para a anti-HAV e os restantes são soronegativos para a anti-HAV. Por fim, a amostra selecionada para teste é composta pelos indivíduos restantes que não participam das fases de treinamento e validação, podendo, assim, testar a generalização do sistema neural de classificação, bem como do modelo de regressão logística.

2.2. Regressão Logística

Como dito anteriomente, o nosso objetivo é determinar um modelo capaz de predizer e classificar os dados em estudo. O modelo de regressão logística, também conhecido por modelo logístico, é uma técnica bastante conhecida e utilizada para predizer uma variável resposta de caráter binário. Este modelo estabelece uma relação entre a probabilidade de ocorrência dos resultados de uma variável resposta binária (em geral chamada de variável dependente), que normalmente é representada pleos termos sucesso e fracasso e variáveis explicativas categóricas ou contínuas (conhecidas como variáveis independentes).

A idéia básica do modelo logístico é estabelecer uma relação linear entre as variáveis explicativas (ou alguma transformação dessas) e uma transformação, denominada logito (logit), da variável resposta. Este modelo é representando por

$$log\left(\frac{P\{Y_i(\mathbf{x})=1\}}{P\{Y_i(\mathbf{x})=0\}}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (1)$$

onde $P\{Y_i(\mathbf{x}) = 1\}$ representa a probabilidade de sucesso para a variável resposta, $P\{Y_i(\mathbf{x}) = 0\}$ representa a probabilidade de fracasso,$\mathbf{x}$=$(x_1, x_2, ..., x_p)$ é um vetor de variáveis explicativas e $\beta_0, \beta_1, \cdots, \beta_p$ são parâmetros desconhecidos, devendo assim serem estimados.

Dessa forma, de (1), a probabilidade de sucesso para a variável resposta é dada por

$$\pi_i(\mathbf{x}) = P\{Y_i(\mathbf{x}) = 1\} = \frac{1}{1 + exp(-Z_i)} \quad (2)$$

com $Z_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

Alguns aspectos importantes a serem considerados consistem no ajuste do modelo, na seleção de variáveis explicativas e na verificação do ajuste. O ajuste do modelo é feito através do método de máxima verossimilhança[1]. Em síntese, este método retorna, para um dado conjunto de observações, valores para os parâmetros desconhecidos que maximizam a probabilidade dos dados serem originados pelo modelo proposto. Para aplicação do método, faz-se necessário construir uma função, chamada de *função de máxima verossimilhança*, a qual expressa a probabilidade das observações como uma função dos parâmetros desconhecidos. Os estimadores de máxima verossimilhança para esse parâmetros são escolhidos como sendo os valores que maximizam essa função.

Inicialmente, para avaliarmos a variação do risco em relação à soroprevalência de hepatite A, criamos um modelo logístico com as 66 covariáveis observadas nos indivíduos residentes nos domicílios amostrados, em seguida, através deste modelo, selecionamos as variáveis independentes, que fossem estatististicamente significantes.

A seleção das variáveis do modelo logístico foi realizada por um procedimento ”passo a passo”(stepwise[4]), considerando-se 10% e 20% os níveis de significância para inclusão e exclusão de variáveis, respectivamente.

O modelo de regressão logística resultante, para a amostra em estudo, incluiu apenas sete variáveis independentes, consideradas estatisticamente significantes ao nível de 10% de significância, refletindo informações relativas ao indivíduo, ao ambiente domiciliar e peridomiciliar, além de variáveis sócio-econômicas. Através do modelo de regressão logística foi possível identificar, por exemplo, as variáveis relacionadas ao ambiente domiciliar e peri-domiciliar que estão associadas a um risco maior ou menor de adquirir a doença. As sete variáveis independentes resultantes foram: idade, proximidade do domicílio à vala negra, densidade de moradores/cômodos, número de pontos de água no domicílio, a não utilização de filtro de água, número de anos estudo da dona de casa e renda média mensal.

Após identificarmos as variáveis estatisticamente significantes, ajustamos um modelo logístico, contendo as sete variáveis independentes que fossem capaz de classificar os indivíduos pertecentes a amostra de treinamento. O modelo logístico é descrito abaixo.

Seja Y_i o resultado do exame sorológico para o in-

divíduo i. Temos que,

$$Y_i = \begin{cases} 1, & \text{se o indivíduo é soropositivo} \\ 0, & \text{se o indivíduo é soronegativo} \end{cases}$$

Seja $\pi_i(\mathbf{x}) = P\{Y_i = 1(\mathbf{x})\}$, onde $0 < \pi_i(\mathbf{x}) < 1$, a probabilidade do indivíduo i ser soropositivo. Assim, assumindo que os indivíduos são independentes, é natural modelar Y_i por uma distribuição de Bernoulli com probabilidade $\pi_i(\mathbf{x})$, denotada por,

$$Y_i \mid \pi_i(\mathbf{x}) \sim Ber(\pi_i(\mathbf{x})) \tag{3}$$

Assim, a probabilidade $\pi_i(\mathbf{x})$ do indivíduo i, na amostra de treinamento, ser soropositivo está relacionada com as covariáveis $x_1, x_2, ..., x_7$ através do seguinte modelo logístico:

$$log\left(\frac{\pi_i(\mathbf{x})}{1-\pi_i(\mathbf{x})}\right) = Z_i \tag{4}$$

onde $Z_i = 0,172x_{i1} + 0,192x_{i2} + 0,210x_{i3} - 0,048x_{i4} + 0,197x_{i5} - 0,049x_{i6} - 0,058x_{i7} - 1,708$ e x_{ij} é a j-ésima covariável observada para o indivíduo i.

Assim, a probabilidade do indivíduo i ser soropositivo é dada por:

$$\pi_i(\mathbf{x}) = \frac{1}{1 + exp(-Z_i)} \tag{5}$$

A qualidade de um determinado modelo ajustado a partir da regressão logística pode ser considerada ao compararmos os valores observados e os valores preditos para a variável resposta. A qualidade do modelo logístico ajustado, definido acima, foi verificada comparando-se os valores observados com os valores preditos pelo modelo. Visto que, o modelo logístico descrito acima apresentou uma boa capacidade de classificar os indivíduos pertecentes a amostra de treinamento, dadas as variáveis explicativas $x_1, x_2, ..., x_7$, o mesmo foi adotado para fazermos a comparação com o classificador neural.

Para compararmos os resultados do modelo de regressão logística com os da RNA, aplicamos a equação logística (eq.4) para a amostra de validação e de teste.

2.3. Rede Neural Artificial

Identificadas as variáveis que melhor explicam a doença em estudo, passamos a projetar uma RNA capaz de classificar os indivíduos em estudo.

Planejamos uma rede neural artificial sem realimentação e totalmente conectada. Tal rede possui sete variáveis de entradas, as quais correspondem as sete variáveis explicativas do modelo de regressão logística, uma camada escondida, contendo 3(três) neurônios, e uma única saída (soropositividade para a anti-HAV), a qual possui um caráter binário. A RNA foi treinada por backpropagation.

As variáveis contínuas foram reescaladas[5] para se adptarem ao intervalo de -1 e 1. A variável reescalada X_i^* foi obtida da seguinte forma:

$$X_i^* = \frac{(X_i - med_i)}{\text{máx}(X_i - med_i)} \tag{6}$$

onde med_i é a média da variável X_i. As variáveis dicotômicas foram codificadas em -1 e 1, onde -1 representa a ausência do atributo observado e 1 representa a presença do atributo.

A rede foi treinada por um número determinado de épocas, sendo que os pesos iniciais foram escolhidos aleatoriamente ente -0,1 e 0,1.

3. Resultados

Após o treinamento, a rede teve o seu desempenho avaliado, utilizando-se a amostra de teste. A proporção de concordância total, ou seja a proporção de indivíduos da amostra de teste que foram classificados como soropositivo (soronegativo), sendo realmente portadores da doença (não portador da doença), para a RNA foi igual a 88%. A regressão logística apresentou uma proporção de concordância total igual a 83%,para o mesmo conjunto de indivíduos. Sendo assim, observamos que a RNA consegue classificar melhor os indivíduos da amostra de teste.

Ao selecionar um modelo, é bastante importante para o especialista conhecer a sensibilidade e a especificidade[6] do modelo que o mesmo propõe-se a utilizar. A sensibilidade do modelo nos dirá o quanto o modelo está classificando corretamente os indivíduos portadores da hepatite A, enquanto a especificidade nos dirá o quanto o modelo está classificando corretamente os indivíduos não portadores da hepatite A. Assim, um modelo sensível é o modelo que classifica corretamante o indivíduo que possui a doença. Na prática, tal modelo é escolhido quando a penalidade, por deixar de diagnosticar uma doença, é grande. Isso é o caso, por exemplo, quando há suspeita de uma condição perigosa, mas tratável, como tuberculose.

Para obtermos mais informações sobre os modelos que estamos estudando (RNA e regressão logística), calculamos a sensibilidade e especificidade da RNA e do modelo logístico (ML). A Tabela 1 mostra os resultados obtidos.

	RNA	**ML**
Sensibilidade (%)	70	52
Especificidade (%)	99	99

Tabela 1: *Sensibilidade e Especificidade*

Quanto mais sensível for o modelo, ou seja, quanto maior for a probabilidade do modelo classificar o indivíduo como soropositivo dado que o paciente possui a doença, maior será a segurança do especialista de que um paciente com resultado negativo não tenha a doença(redução da probabilidade de perda do alvo[7]).

Em contra posição, quanto mais específico for o modelo, maior será a segurança do especialista de que um resultado positivo confirma ou mantém o diagnóstico procurado(diminuição da probabilidade de falso alarme).

De acordo com os resultados apresentados na Tabela 1, observamos que a RNA apresenta maior sensibilidade, ou seja, ela consegue classificar melhor os indivíduos portadores da hepatite A do que o modelo de regressão logística. Em relação a especificidade, ambos os modelos apresentaram a especificidade bastante elevada. Assim, podemos afirmar que o sistema neural de classificação é mais sensível em relação à doença em estudo.

4. Conclusões

Redes neurais artificiais têm sido bastante utilizadas como modelo de classificação na área epidemiológica. A computação envolvida nas etapas de aprendizado na RNA é facilitada se o especialista do problema é posto para trabalhar em conjunto com o processamento neural, criando um enfoque de processamento híbrido que ataca o problema.

Na área médica, onde se pretende utilizar as redes neurais como instrumento de apoio ao diagnóstico, esta metodologia é atraente e se tem mostrado eficiente. No caso da identificação de indivíduos portadores de hepatite A, a escolha das variáveis de entrada realiza um pré-processamento fundamental para a classificação. Com este enfoque, o classificador neural foi capaz de classificar corretamente 88% da amostra de generalização.

Como dito anteriormente, o objetivo principal deste trabalho é estabelecer um modelo que auxilie no diagnóstico do paciente. Nossos resultados mostram que a RNA é o melhor modelo de classificação para os dados em estudo, pois o mesmo possui maior capacidade de generalização. Além disso, tal modelo possui maior sensibilidade, ou seja, a RNA possui maior capacidade de identificar os indivíduos portadores da hepatite A do que o modelo de regressão logística.

5. Agradecimentos

Os autores são gratos ao CNPq, CAPES, FAPERJ pelo apoio ao projeto.

Referências

[1] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience, New York, 1998.

[2] S. Haykin. *Neural Networks: A comparative Foundation*. Prentice Hall, New Jersey, 1999.

[3] J. F. H. Jr, R. E. Anderson, R. L. Tatham, and W. C. Black. *Multivariate Data Analysis*. Prentice Hall, New Jersey, 1998.

[4] D. G. Kleinbaum. *Logistic Regression - A Self-Learning Text*. New York, 1994.

[5] M. Smith. *Neural Network for Statistical Modelling*. Van Nostrond Reinhold, New York, 1993.

[6] K. J. Rothaman and S. Greenland. *Modern Epidemiology*. Lippincott-Raven Publishers, United States of America, 1998.

[7] H. L. V. Trees. *Detection Estimation and Modulation, Part I*. John Wiley & Sons, 1968.