Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais
pp.427–432, June 2–5, 2003 - Centro Universitario da FEI, Sao Paulo - SP - Brazil

Árvore de Classificação e Redes Neurais Artificiais: Uma Aplicação à Predição de Tuberculose Pulmonar

Alcione Miranda dos Santos[1,2], Basílio de Bragança Pereira[2,3]
José Manoel de Seixas[4], Fernanda Carvalho de Q. Mello[5]
Afrânio L. Kritski[5]
[1]Departamento de Matemática-UFMA
[2]COPPE-UFRJ
[3] Faculdade de Medicina e NESC–UFRJ
[4]Laboratório de Processamento de Sinais-COPPE-EE-UFRJ
[5] Faculdade de Medicina e Unidade de Pesquisa em Tuberculose–UFRJ
E-mails: amiranda@pep.ufrj.br,seixas@lps.ufrj.br

Abstract

Smear negative pulmonary tuberculosis (SNPT) accounts of 30% of Pulmonary Tuberculosis (PT) cases reported yearly. Rapid and accurate diagnosis of SNPT could provide lower morbidity and mortality, and case detection at a less contagious status. The main objective this work is to evaluate a prediction model for diagnosing SNPT, useful for outpatients attended in settings with limited resources. One hundred thirty six patients from Health Care Units were included. They were referred to our Teaching Hospital, in Rio de Janeiro, Brazil, from March, 2001 to September, 2002, with clinical-radiological suspicion of SNPT. Only symptoms and physical signs were used for constructing the neural network (NN) and the classification tree. The NN classified correctly 73% of patients from the test sample, while the classification tree classified only 40% of those patients. The NN model suggests that mathematical modeling, for classifying SNPT cases, could be a useful tool for optimizing utilization of more expensive tests, and to avoid costs of unnecessary anti-PT treatment.

1. Introdução

A tuberculose (TB) permanece entre as principais enfermidades que acometem a humanidade. Com o surgimento, em 1981, da Síndrome de Imunodeficiência Adquirida (SIDA/AIDS), vem-se observando, tanto em países desenvolvidos como nos países em desenvolvimento, um crescente número de casos notificados de tuberculose em pessoas infectadas pelo vírus da imunodeficiência humana (HIV). A associação (HIV/TB) constitui, nos dias atuais, um sério problema de saúde pública, podendo levar ao aumento da morbidade e mortalidade pela tuberculose, em muitos países.

Estima-se que cerca de 1,7 bilhões de indivíduos em todo o mundo estejam infectados pelo *Mycobacterium tuberculosis*, correspondendo a 30% da população mundial. No Brasil, estima-se que do total da população, 35 a 45 milhões de pessoas estão infectadas pelo bacilo *Mycobacterium tuberculosis*, com aproximadamente 100 mil casos novos por ano. O número de mortes pela doença em nosso meio é de 4 a 5 mil, anualmente [1].

Nos grandes centros urbanos, a ocorrência da TB aumentou significativamente, como foi observado na cidade do Rio de Janeiro. Nessa cidade, em 1985, a incidência era de 80/100.000, mas em 1998, essa taxa atingiu o valor de 126/100.000. Desta forma, a TB revelou-se novamente como uma importante pandemia, considerada uma urgência mundial pela Organização Mundial de Saúde (OMS) desde 1993, assumindo o seu controle elevada prioridade, pelo grande problema de saúde pública que se tornou.

Vários eventos contribuíram para o atual panorama da TB no mundo: o advento da infecção pelo vírus HIV, a deterioração das condições sócio-econômicas de parte da população mundial, a elevação da taxa de abandono do tratamento antituberculose, o aparecimento da multiresistência e a falta de interesse da comunidade científica e dos formadores de políticas públicas em relação à TB, ao não incentivarem o desenvolvimento de novos instrumentos para o controle da TB. A concentração dos casos, nas áreas urbanas de países industrializados, e, em particular, nas Unidades Hospitalares, local de elevada prevalência de pacientes com co-morbidades, também propiciou um aumento do risco de transmissão da infecção e de adoecimento por TB. Co-morbidades como o câncer, a insuficiência renal crônica, e o diabetes mellitus, ao determinarem maior demora na suspeita e na confirmação diagnóstica, contribuíram sobremaneira para uma maior transmissão da TB nas Unidades de Saúde entre pacientes, mas também para profissionais de saúde [1].

Os Centros para Controle de Doenças (CDC) dos EUA consideram as técnicas atuais para o diagnóstico de TB lentas e sem sensibilidade e especificidade ideais. A baciloscopia do escarro é o método utilizado rotineiramente para a identificação do bacilo álcool-ácido resistente causador da TB. Entretanto, esta técnica ca-

rece de sensibilidade (capacidade de identificar os indivíduos portadores da doença), que oscila entre 30% a 80%, com média de 60%, em pacientes com cultura positiva. Além disso, não é capaz de discriminar a espécie da micobactéria [2, 3]. A cultura para o bacilo da TB é um método mais sensível, pois detecta 70% a 89% dos casos, em média 80%, e permite a identificação posterior da espécie, através de testes bioquímicos ou sondas genéticas [4]. A limitação deste método reside no tempo necessário, visto que o resultado da cultura fica disponível apenas 15 a 60 dias após a coleta do material respiratório. Portanto, nos pacientes com baciloscopia negativa no escarro, o diagnóstico micobacteriológico da TB é geralmente tardio.

No Brasil, cerca de 26,7% dos pacientes adultos são tratados sem confirmação para TB pulmonar (TBP), com base apenas no quadro clínico-radiológico. No Município do Rio de Janeiro, o índice de tratamentos anti-TB de prova atinge 46% dos casos de TBP[5].

Neste contexto, emerge a possibilidade de utilização de modelos estatísticos para o auxílio no diagnóstico da TBP. Os modelos podem servir para previsão, a longo prazo, da tendência da ocorrência da infecção ou da doença. Além disto,possibilitam simular situações epidemiológicas e intervenções preventivas ou curativas, além de estimativas do seu impacto teórico na redução do problema. Tais modelos preditivos (de previsão epidemiológica), se formulados de maneira adequada, e alimentados com dados que tenham qualidade e que sejam representativos de determinada realidade, podem auxiliar os médicos na sua prática clínica, como também os administradores da saúde pública [6].

Neste trabalho, utilizam-se modelos estatísticos para a elaboração de instrumentos de predição de tuberculose pulmonar para a população de suspeitos de desenvolver a doença atendida em diferentes Unidades de Saúde e encaminhados para uma Unidade de Saúde Terciária da Cidade do Rio de Janeiro.

2. Métodos

Dois métodos de análise são apresentados para comparação em termos do desempenho de classificação. Primeiramente, desenvolveremos um árvore de classificação. Em seguida, discutiremos o uso de uma rede neural artificial, treinada por supervisão.

2.1. Árvore de Classificação

As árvores de classificação foram popularizadas na comunidade estatística através do trabalho de Breiman et.al. [7]. Eles propuseram um método de decisão por árvore binária, conhecido como CART (*Classification and Regression Tree*). Tal modelo descreve a distribuição condicional da variável resposta y dado $\boldsymbol{x}$, onde $\boldsymbol{x}=(x_1, x_2, \cdots, x_p)'$ é um vetor de variáveis preditoras (ou explicativas). Quando a variável resposta assume valores categóricos, a árvore é tratada como *árvore de classificação*, caso contrário, a árvore é conhecida como *árvore de regressão*. As variáveis explicativas, em ambas as árvores, podem assumir valores contínuos ou categóricos. Neste trabalho, iremos apenas tratar com árvore de classificação, visto que a variável resposta observada possui caráter binário.

O principal objetivo das árvores de classificação é predizer ou explicar uma variável resposta. No nosso estudo, a árvore de classificação pode ser utilizada para predizer a probabilidade do paciente ter tuberculose, dado que o mesmo apresenta alguns sintomas sugestivos.

Ao se trabalhar com os modelos CART, costuma-se utilizar a seguinte terminologia. Cada posição da árvore é chamada de *nó*, sendo o primeiro nó chamado de *nó raiz*, e equivale ao conjunto de dados completo. Cada nó representa uma decisão ou teste sobre o valor de um atributo (variável). Os nós gerados pela divisão de um nó já existente recebem o nome de *descendentes* e o nó que os originou é chamado de *ascendente* ou *pai*. Quando o conjunto de dados contido em um determinado nó não é particionado entre dois nós descendentes, o nó é declarado *terminal* e a este é associada uma classe, a qual será atribuída a todos os casos encontrados neste nó.

A construção de um modelo CART consiste em uma seleção de divisões binárias em um específico nó, no qual a divisão é realizada de acordo com o valor de uma variável selecionada.

Para cada variável explicativa numérica serão incluídas questões da forma:

variável x é $\geq$ s?

havendo uma questão para cada valor de **s**, que pode assumir um número finito de valores, variando entre o mínimo e o máximo da variável em questão. Para cada variável explicativa categórica que assuma valores em $\{c_1, c_2, \cdots, c_k\}$ as questões serão da forma:

variável x $\in$ ao subconjunto B?

onde os subconjuntos **B** testados são todos os subconjuntos de $\{c_1, c_2, \cdots, c_k\}$. Cada divisão conduzirá um determinado caso ao nó descendente da direita ou da esquerda, conforme a resposta apresentada à mesma, negativa ou positiva.

Para cada nó, o algoritmo realiza essa pesquisa sobre todas as variáveis, uma por uma. Para cada variável, ele encontra a melhor divisão. Então, ele compara a melhor divisão de cada uma das variáveis e seleciona a melhor divisão. O algoritmo é encerrado quando o nó resultante for bastante homogêneo ou se ele possuir um número mínimo especificado de observações. Um nó será homogêneo quando a maioria dos indivíduos, neste nó, possui o mesmo padrão.

Para escolher entre as possíveis divisões, é necessário um parâmetro que indique o quanto uma divisão é melhor que a outra. Para isto, define-se um índice de qualidade da divisão, que será tanto melhor quanto mais homogêneos forem os nós descendentes resultantes das divisões. Para medir a homogeneidade, ou equivalentemente a impureza de determinado nó, algumas medidas

são utilizadas, por exemplo o *índice de diversidade de Gini* [7].

O processo inicial de construção gera, normalmente, árvores desnecessariamente grandes e fortemente influenciadas pelo conjunto de dados considerado (amostra de treinamento), e implica em previsões pobres para um novo conjunto de observações. Isto acontece porque uma árvore exageradamente grande é muito especializada na amostra de treinamento, perdendo assim o seu poder de generalização. Um caminho para evitarmos a superestimação é o princípio de poda (*prunning*) ou encolhimento (*shrinking*)[8]. Outro método utilizado para selecionar o tamanho correto da árvore, afim de evitar o risco de superestimação é o método de validação cruzada (*cross validation*) [9]. Este método é bastante utilizado quando dispomos de uma amostra relativamente pequena, tornando-se inviável obtermos uma amostra significativa de teste.

Após definido o tamanho correto da árvore, devemos analisar sua precisão. Um caminho para analisarmos a precisão da árvore de classificação é calculando a fração de observações na amostra que estão mal classificadas pela árvore de classificação, a qual é conhecida como *erro por resubstituição* [7].

O erro por resubstituição é fácil de calcular e tem sido bastante utilizado por vários pesquisadores, no estudo de classificação para medir a acurácia do classificador. Entretanto, esta estimativa é uma medida enviesada, quando calculada com os mesmos dados já utilizados para definir a árvore de classificação. Uma forma de resolver isto é dividir a amostra em estudo em dois subamostras: *amostra de treinamento* e *amostra de teste*. A amostra de treinamento é usada para estimar a árvore de classificação, enquanto que a amostra de teste serve para estimar o erro de classificação real.

Deve-se observar, entretanto, que dividido a amostra de dados em amostra de treinamento e teste, não se faz o melhor uso dela. É claro que seria melhor basearmos a construção da árvore de classificação utilizando a amostra de dados completa, especialmente quando sua cardinalidade não for muito grande.

Cuidados devem ser tomados, de tal forma que os casos contidos na amostra de treinamento possam ser considerados independentes dos casos da amostra de teste e sorteados da mesma população. Frequentemente, toma-se 1/3 dos casos da amostra de estudo para compor a amostra de teste, mas não se conhece nenhuma justificação teórica para este procedimento.

2.2. Rede Neural Artificial

Redes Neurais Artificiais (RNA) podem ser definidas como modelos não-lineares, podendo ser aplicadas em problemas de regressão, classificação e redução de dados. Além disso, são aplicadas freqüentemente em situações onde existem interações não-lineares entre as variáveis dependentes e as independentes.

Nas últimas décadas, redes neurais artificiais vêm sendo 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 epidemiológicos tem tido grande aceitação. Além do mais, o processamento neural é capaz de extrair correlações das variáveis de entrada diretamente sobre os espaços de dimensão elevada que tipicamente as caracterizam, tornando tal processamento uma ferramenta valiosa em problemas complexos de reconhecimento de padrões [10].

Atualmente, tem crescido o número de aplicação de RNA em áreas mais tradicionais da estatísticas, em particular para problemas de classificação e análise de dados de sobrevivência. Alguns estudos aplicados à problemas de classificação têm avaliado o potencial de utilização das RNA, mostrando que elas podem apresentar, em alguns casos, desempenho preditivo melhor que os métodos estatísticos convencionais.

Existem vários tipos de redes neurais [11]. Aqui, iremos nos restringir às de redes *multicamadas feedforward*, onde os neurônios de uma camada estão conectados apenas aos neurônios da camada imediatamente a seguir, não havendo nem realimentação (comunicação unidirecional), nem conexões entre neurônios da mesma camada.

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. A rede implementada neste trabalho utiliza dados inteligentes, ou seja, dados que representam o problema a partir do conhecimento especialista acumulado, podendo a rede neural servir como ferramenta de apoio ao diagnóstico de TBP ativa.

Na implementação de uma RNA, se faz necessário sabermos quantas camadas escondidas devemos utilizar na rede, além disso, quantos neurônios deverá conter a camada escondida. Outra dúvida que surge é saber quais variáveis explicativas devemos colocar na camada de entrada.

Influenciados por trabalhos teóricos [12, 13] que mostram que uma única camada escondida é suficiente para uma rede neural aproximar qualquer função não linear, vários autores usam apenas uma camada escondida na rede neural.

A tarefa de determinar um número ótimo de neurônios na camada escondida não é trivial. Em geral, redes neurais com poucos neurônios escondidos são preferidas, visto que elas possuem melhor poder de generalização e pouco problema de superestimação. Entretanto, redes com poucos neurônios escondidos não possuem a habilidade suficiente para modelar e aprender os dados.

Um caminho comumente seguido para determinar o número de neurônios escondidos é via experimentos. Uma proposta, quando dispomos de um conjunto de dados relativamente grande, consiste em treinar várias redes com um número distinto de neurônios na camada escon-

dida, a partir da amostra de treinamento e verificar o erro de generalização para cada uma delas a partir da amostra de teste.

Sabe-se que um modelo neural com um grande número de neurônios escondidos poderá apresentar um bom ajuste para a amostra de treinamento, entretanto, isto não garante que a rede terá um bom desempenho na amostra de teste. Neste sentido, é necessário estabelecer uma medida associada à complexidade dos modelos neurais para termos um critério que permita a escolha de determinados modelos.

Alguns autores utilizam critérios de ajustamento-penalidade (*complexity-regularization*) para selecionar o número de neurônios escondidos. Esses critérios são análogos aos critérios estatísticos AIC (*Akaike Information Criterion*) [14] e BIC (*Bayesian Information Criterion*) [15].

Em problemas de classificação, a medida mais importante para medir o desempenho da rede neural é avaliar o desempenho da rede a partir da classificação de novos casos (amostra de teste). O desempenho da rede, medida através da amostra de teste, é uma boa indicação de seu desempenho.

Quando estamos tratando com problemas de classificação binária, costuma-se medir o desempenho da rede através do seu poder de discriminação. Isto é feito obtendo-se a taxa de classificação correta para a amostra de teste, como também, calculando-se algumas estatísticas descritivas, tais como falso-positivos, falso-negativos, sensibilidade e especificidade que podem fornecer resultados mais significantes.

3. Base de Dados para Construção dos Modelos

A base de dados aqui tratada, refere-se aos dados de 136 pacientes que procuraram o ambulatório do Hospital Clementino Fraga Filho (HUCFF) da Universidade Federal do Rio de Janeiro e que consentiram em participar do estudo. Estes pacientes estavam sob suspeita de apresentarem TBP ativa, apresentando resultado de baciloscopia negativa.

Os dados dos pacientes foram obtidos por meio de fichas e prontuários médicos. As informações contidas nas fichas consistem das variáveis demográficas e dos fatores de risco para TBP.

As variáveis explicativas foram escolhidas, a princípio, de acordo com a dependência com a TBP. Podemos citar: idade, tosse, escarro, sudorese, febre, emagrecimento, dor torácica, calafrios, dispnéia, diabetes, alcoolismo e outras, num total de 23 variáveis.

Da base de dados original foram criados aleatoriamente duas amostras: uma amostra, a qual denominaremos de *amostra de treinamento*, para a construção dos modelos e uma amostra para validação dos modelos ajustados (*amostra de teste)*.

A amostra selecionada para treinamento é composta de 91 pacientes, sendo que 47 pacientes possuem diagnóstico positivo para a tuberculose e os restantes não possuem. Por fim, a amostra selecionada para teste é composta pelos pacientes que não participam das fases de treinamento, podendo assim testar a generalização do sistema neural de classificação, bem como da árvore de classificação.

4. Resultados

A árvore de classificação foi ajustada contendo as vinte e seis variáveis explicativas observadas nos pacientes. A variável resposta, y_i, em estudo é definida da seguinte forma:

$$y_i = \begin{cases} +1, & \text{se o paciente estiver com TBP ativa} \\ -1, & \text{se o paciente não estiver com TBP ativa} \end{cases}$$

A RNA foi desenvolvida contendo vinte e três nós de entradas, correspondentes às variáveis explicativas dos pacientes, uma camada escondida com três neurônios, com função de transferência tangente hiperbólica e uma camada de saída com um neurônio (contendo a variável resposta em estudo), cuja função de transferência adotada foi novamente a tangente hiperbólica.

Para facilitar o processo de convergência da rede, todas as variáveis foram normalizadas para a faixa -1 e 1. O algoritmo de treinamento utilizado foi o *Backpropagation*. O treinamento da rede foi interrompido após um determinado número de ciclos.

A taxa de classificação correta para a RNA foi de 100% para a amostra de treinamento, enquanto na amostra teste, obtivemos uma taxa de 69%. Para a árvore de classificação, obtivemos uma taxa igual a 85% na amostra de treinamento e uma taxa de 60% na amostra teste.

Como falamos anteriormente, em problemas de classificação binária costuma-se medir o desempenho da rede através da taxa de classificação correta para a amostra de teste, como também, calculando-se algumas estatísticas descritivas, tais como falso-positivos, falso-negativos, sensibilidade e especificidade. Estas medidas são bastante utilizadas na área médica.

A sensibilidade do modelo nos dirá o quanto o modelo está classificando corretamente aqueles pacientes que estão com TBP em atividade, enquanto a especificidade nos dirá o quanto o modelo está classificando corretamente os pacientes que não estão com TBP em atividade. Assim, um modelo sensível é o modelo que classifica corretamante o paciente que possui a doença. Na prática, tal modelo é escolhido quando a penalidade, por deixar de diagnosticar uma doença, é grande.

Para obtermos mais informações sobre os modelos que estamos estudando (RNA e árvore de classificação), calculamos a sensibilidade e especificidade da RNA e do árvore de classificação (CART). A Tabela 1 mostra os resultados obtidos.

De acordo com os resultados apresentados na Tabela 1, observamos que a RNA apresenta maior sensibilidade, ou seja, ela consegue classificar melhor os indivíduos que estão com TBP em atividade do que a árvore de classificação, enquanto que, em termos de especificidade,

	RNA	CART
Sensibilidade (%)	73	40
Especificidade (%)	67	70

Tabela 1: *Sensibilidade e Especificidade*

os métodos exibem um desempenho similar. Assim, podemos afirmar que o sistema neural de classificação é mais sensível em relação à doença em estudo.

5. Conclusões

Como descrevemos anteriormente, os métodos tradicionais para o diagnóstico da tuberculose pulmonar apresentam limitações. A baciloscopia com sensibilidade de 40% a 60% e a cultura, apesar de sensibilidade de 70% a 80%, demanda 4 a 8 semanas, período no qual poderá haver agravamento significativo da doença, além de transmissão inter humana.

Neste estudo, os pacientes apresentavam resultado da baciloscopia do escarro espontâneo negativo, sendo portanto um desafio o diagnóstico para os clínicos, demandando intervenções de maior complexidade e elevados custos para o indivíduo e o sistema de saúde.

Pelos resultados, observamos que no caso da identificação de indivíduos portadores de tuberculose pulmonar com baciloscopia negativa, o classificador neural foi capaz de classificar corretamente 73% da amostra de generalização, sinalizando o potencial deste método na condução dos casos de tuberculose pulmonar com baciloscopia negativa.

6. Agradecimentos

Os autores são gratos a CAPES, CNPq, FAPERJ pelo apoio a este trabalho e aos funcionários e alunos da UPT (Unidade de Pesquisa em Tuberculose) da Universidade Federal do Rio de Janeiro pela coleta e organização da base de dados.

Referências

[1] MELLO, F. C. Q. *Modelos Preditivos para Tuberculose Pulmonar Paucibacilar*. Tese de doutorado, Faculdade de Medicina - Universidade Federal do Rio de Janeiro, 2001.

[2] REIDER, H. L., CONDE, T. M., MYKING, H., and et. al. The public health service national tuberculosis reference laboratory and the national laboratory netowrk: minimum requirements, role and operation in a low-income country. *International Union Against Tuberculosis and Lung Disease*, 1998.

[3] GEBRE, N., KARISSON U., JONHSSON, G. , and et al. Improved microscopical diagnosis of pulmonary tuberculosis. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 89:191–193, 1995.

[4] SCHIRM, J., OOSTENDORP, L. A., and MULDR, J. G. Comparasion of amplicor, in house pcr and conventional culture for detection of mycobacterium in clinical samples. *Journal Clinical of Microbiology*, 33:3321–3324, 1995.

[5] Boletim Eletrônico Secretaria de Saúde do Rio de Janeiro. 2000.

[6] KRITSKI, A. L. and RUFFINO–NETO, A. Health sector reform in brazil: impact on tuberculosis control. *International Journal Tuberculosis Lung Disease*, 4(7):622–626, 2000.

[7] BREIMAN, L.L., FREIDMAN, J., OLSHEN, R., and STONE, C. *Classification and Regression Trees*. Wadsworth, Blemont, CA, 1984.

[8] CLARK, L. A. and PREGIBON, D. Tree based models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, pages 377–420. Wadsworth, 1992.

[9] STONE, C. J. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Serie B*, 36:111–147, 1974.

[10] SANTOS, A. M., PEREIRA, B. B, MEDRONHO, R. A., CAMPOS, M. R., SEIXAS, J. M, and CALÔBA, L. P. Aplicação de redes neurais artificiais em dados epidemiológicos de hepatite A. In *V Congresso Brasileiro de Redes Neurais*, pages 586–589, 2001.

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

[12] CYBENKO, G. Aproximations by superpositionn of sigmoidal function. *Mathematical Control Signals Systems*, 2:303–314, 1989.

[13] HORNIK, K., STINCHCOMBE, M., and WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[14] AKAIKE, H. . A new look at the statistical model identification. *IEEE Trans. On Automatic Control*, 19(6):716–723, 1974.

[15] SCHWARZ, G. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.