Usando Redes Neurais Artificiais e Regressão Logística na Predição da Hepatite A

Using Artificial Neural Networks and Logistic Regression in the Prediction of Hepatitis A

Resumo

Este trabalho desenvolve um sistema para predição da soroprevalência da Hepatite A. Para isto, são considerados os modelos de regressão de logística e redes neurais artificiais. O desempenho de tais modelos é medido através da taxa de classificação incorreta em uma amostra do município de Duque de Caxias, Rio de Janeiro, que possui elevada prevalência da doença. Resultados mostram que o modelo neural, aplicado sobre a informação relevante extraída do modelo de regressão logística, apresenta um bom desempenho, alcançando uma eficiência de classificação geral acima de 88%.

Palavras-chaves: Redes Neurais. Regressão Logística. Hepatite A. Sistemas de Apoio ao Diagnóstico Médico.

Alcione Miranda dos Santos[1]*

José Manoel de Seixas[2]

Basílio de Bragança Pereira[3]

Roberto de Andrade Medronho[4]

[1]Departamento de Matemática, Universidade Federal do Maranhão

[2]COPPE e Escola Politécnica, Universidade Federal do Rio de Janeiro

[3]Faculdade de Medicina - NESC e COPPE, Universidade Federal do Rio de Janeiro

[4]Faculdade de Medicina – NESC, Universidade Federal do Rio de Janeiro

***Correspondência:** Alcione Miranda dos Santos. Av. dos Portugueses, S/N, Campus do Bacanga, 65080-040 São Luís - MA. E-mail: amiranda@demat.ufma.br

Abstract

This paper aims to develop a support system for seroprevalence prediction of hepatitis A. Logistic regression and artificial neural network models were considered. The accuracy of these models was measured based on the misclassification rate in a sample from the city of Duque de Caxias, Rio de Janeiro, where there is a high incidence of this disease. The results of the evaluation show that the neural model achieves an overall classification efficiency of 88%, when it uses relevant information extracted from the logistic model.

Key Words: Neural Networks. Logistic Regression. Hepatitis A. Supporting Systems for Medical Diagnosis.

Introdução

A hepatite A é uma doença do fígado altamente contagiosa e às vezes fatal. A infecção pelo vírus da hepatite A constitui um dos maiores problemas de saúde pública no Brasil. A transmissão do vírus da hepatite A (HAV) ocorre principalmente pela via orofecal, deste modo a água é um importante veículo da disseminação do HAV [1].

O Brasil tem índices elevados de transmissão de Hepatite A, em razão de condições deficientes ou inexistentes de saneamento básico, nas quais é obrigada a viver grande parte da população, inclusive nos grandes centros urbanos.

A hepatite A apresenta alta proporção de casos assintomáticos, contribuindo para que o Sistema de Vigilância Epidemiológica, baseado em notificação passiva de casos, seja de baixa sensibilidade, representatividade, utilidade e oportunidade para o monitoramento, o que dificulta a estimativa da magnitude dos eventos e a decisão a respeito de medidas de intervenção em tempo hábil[2]. Diante deste contexto, torna-se necessária a utilização de ferramentas que auxiliem o diagnóstico da hepatite A.

O principal objetivo deste trabalho é a criação de um sistema de predição da soroprevalência da hepatite A, visando o apoio ao diagnóstico. O sistema, a partir de uma base de informações relevantes para o problema em estudo, contribuirá na identificação de indivíduos com alto risco de contrair a doença, atenuando o risco de disseminação da doença para a população, como também na identificação daqueles indivíduos que merecem posterior investigação em Unidades de Saúde.

Nesse sentido, para criação do sistema, tomou-se como amostra 3.079 indivíduos com idade entre 1 e 83 anos, residentes em uma região metropolitana do Estado do Rio de Janeiro. Em todos os indivíduos amostrados foram colhidas amostras de sangue para a pesquisa de anti-HAV (*hepatite A*) total no soro. Em seguida, identificou-se a soroprevalência global para o anti-HAV (*hepatite A*) entre os indivíduos observados.

Para estabelecermos a influência conjunta das variáveis sobre a soroprevalência anti-HAV, utilizamos a técnica de regressão logística multivariada[3] . Identificadas as variáveis explicativas relevantes ao problema em estudo, projetamos uma rede neural artificial, com o objetivo de avaliar a eficiência de classificação neural e sua potencialidade enquanto sistema de apoio ao diagnóstico médico para a hepatite A.

Materiais e Métodos

Nesta seção, especificamos a área de estudo, bem como o plano de amostragem utilizado. Em seguida, são apresentados os modelos estudados. Por fim, é mostrado como os modelos foram desenvolvidos.

Área de Estudo

Neste trabalho, a região em estudo 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 Estudos de Saúde Coletiva (NESC) da Universidade Federal do Rio de Janeiro e os dados aqui tratados foram gentilmente cedidos pelos coordenadores desse projeto[4].

No setor Parque Fluminense, foram selecionados 19 setores censitários para a realização de um inquérito sorológico para anticorpos contra vírus da hepatite A (anti-HAV), e um inquérito domiciliar para avaliação de condições socioeconômicas e sanitárias.

Plano de Amostragem

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 foram 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 pré-codificado. Foram observadas 66 variáveis, entre as quais 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[4].

Redes Neurais Artificiais

As redes neurais artificiais[5] (RNA) são vistas como modelos paramétricos não-lineares. Uma potencial desvantagem das redes neurais, para área médica, é que os parâmetros (pesos sinápticos) não têm uma interpretação imediata, exigindo análise adicional para se compreender a forma com que a informação é extraída[6]. Entretanto, esta metodologia possui a vantagem de detectar implicitamente qualquer relação não-linear entre a variável resposta e as variáveis explicativas. O modelo de regressão logística também pode ser utilizado para modelar uma relação não-linear entre a variável resposta e as variáveis explicativas; entretanto esta relação não-linear tem que ser explicitada pelo desenvolvedor do modelo[7].

Devido ao fato de não haver necessidade de independência e normalidade das variáveis em estudo, bem como a sua grande capacidade de aprendizado a partir do ambiente, a aplicação de redes neurais artificiais na análise estatística de dados epidemiológicos é atraente. Além do mais, o processamento neural é capaz de extrair relações 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. Por outro lado, redes neurais podem trabalhar em conjunto com outras técnicas de processamento, permitindo que se utilize o conhecimento acumulado em uma determinada área de aplicação[8]. Assim, os dados podem ser pré-processados, identificando-se a informação relevante à tarefa de processamento de interesse, e a rede neural irá operar sobre esta informação qualificada, ao invés de trabalhar com os dados brutos. Desta maneira, evitam-se modelos neurais de alta complexidade, que normalmente são poucos práticos[9].

Redes Neurais Artificiais (RNA) podem ser aplicadas em problemas de regressão, classificação e compactação de dados, como também em situações onde existem interações não-lineares entre as variáveis dependentes e as independentes. Ultimamente, as RNA também vêm sendo utilizadas na área de diagnóstico (ou prognóstico) médico[10-12] e análise de dados de sobrevivência[13]. Também são encontrados alguns trabalhos que apresentam o uso do modelo neural em estudos epidemiológicos[14,15].

A idéia básica subjacente ao paradigma das redes neurais é construir um modelo composto por um grande número de unidades de processamento muito simples, que são chamadas de neurônios, com um grande número de conexões entre eles. O processamento básico de informação da rede ocorre nos neurônios. A informação entre os neurônios é transmitida através de conexões denominadas *sinapses* ou *pesos sinápticos.*

A capacidade de *aprender* através de exemplos e de *generalizar* a informação aprendida representam, sem dúvida, atrativos importantes para a escolha de uma solução neural de problemas diversos. A generalização – associada à capacidade da rede de aprender através de um conjunto de exemplos, representativo do problema que se pretende estudar, e, posteriormente, fornecer respostas coerentes para dados não apresentados anteriormente – é uma demonstração de que a capacidade das RNA vai muito além de mapear relações de entrada e saída. As RNA são capazes de extrair informações não apresentadas de forma explícita através dos exemplos[16].

Diferentes topologias de redes neurais são encontradas na literatura. Neste trabalho, foram utilizadas redes neurais multicamadas *feedforward*[5], que serão aqui denominadas de redes neurais multicamadas.

Uma *rede neural multicamadas* é tipicamente composta de camadas alinhadas de neurônios. Neste tipo de rede, as entradas da rede são apresentadas na primeira camada, que é chamada *camada de entrada*. Esta camada distribui as informações de entrada para a(s) camada(s) escondida(s) da rede. A última camada é a *camada de saída*, onde a solução do problema é obtida. A camada de entrada e a camada de saída poderão ser separadas por uma ou mais camadas intermediárias, chamadas *camadas escondidas*. Na grande maioria das aplicações, considera-se apenas uma camada escondida. Além disso, os neurônios de uma camada estão conectados apenas aos neurônios da camada imediatamente posterior, não havendo realimentação (comunicação unidirecional) nem conexões entre neurônios da mesma camada. Além disso, caracteristicamente, as camadas são totalmente conectadas. Um exemplo de rede neural multicamadas está mostrado na Figura 1.

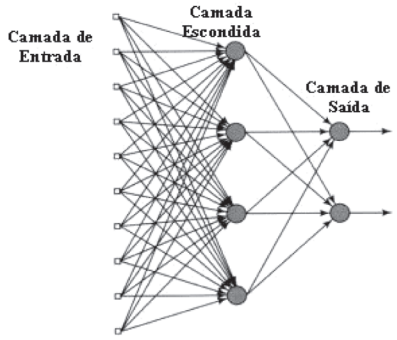

Figura 1 – Exemplo de Rede Neural Feedforward Multicamadas
Figure 1 – *Example of a Multilayer Feedforward Neural Network*

Em termos de topologia, para implementarmos uma rede neural devemos determinar as seguintes variáveis: (a) o número de nós na camada de entrada, (b) o número de camadas escondidas e o número de neurônios a serem colocados nessas camadas, (c) o número de neurônios na camada de saída. Estes parâmetros afetam o desempenho da RNA, devendo ser cuidadosamente escolhidos.

O número de nós na camada de entrada corresponde ao número de variáveis que serão usadas para alimentar a rede neural. Freqüentemente, são as variáveis mais relevantes para o problema em estudo.

Pelo teorema geral de Kolmogorov[17] sobre aproximações de funções, teoricamente basta uma camada escondida na rede. O número de neurônios escondidos é escolhido através de critérios de ajustamento-penalidade (*complexity-regularization*), que são análogos aos critérios estatísticos AIC (*Akaike Information Criterion*) e BIC (Bayesian *Information Criterion*), ou através da capacidade preditiva da rede[18].

Não existe um critério geral que permita definir o número de neurônios na camada escondida. Em geral, redes neurais com poucos neurônios escondidos são preferidas, visto que elas tendem a possuir um melhor poder de generalização, reduzindo o problema de sobreajuste (*overfitting*). Entretanto, redes com poucos neurônios escondidos podem não possuir a habilidade suficiente para modelar e aprender os dados em problemas complexos, podendo ocorrer *underfitting*, ou seja, a rede não converge durante o treinamento[19].

Em alguns casos em que o número de neurônios na única camada escondida torna-se elevado, o uso de duas ou três (muito raramente) camadas pode às vezes nos permitir a diminuição do número de neurônios na camada escondida[20].

Outras decisões a serem tomadas incluem: a seleção da função de ativação dos neurônios da camada escondida e da camada de saída, o algoritmo de treinamento e seus respectivos parâmetros, a transformação dos dados ou método de padronização, a seleção do conjunto de treinamento e conjunto de teste, o critério de parada do treinamento e a escolha de uma medida de desempenho da rede[21].

A *função de ativação*, também chamada de *função de transferência*, é uma função matemática que, aplicada à combinação linear entre as variáveis de entrada e pesos que chegam a determinado neurônio, retorna ao seu valor de saída. Existem diversas funções matemáticas que são utilizadas como função de ativação. As funções de ativação mais comumente usadas são: *função logística* e a *função tangente hiperbólica*.

Existem algumas regras heurísticas para a seleção da função de ativação. Por exemplo, Klimasaukas[22] sugere a função de ativação logística para problemas de classificação que envolvam o aprendizado de um determinado padrão. A função tangente hiperbólica também é bastante utilizada em problemas de classificação, devido ao fato de, em algumas situações práticas, a função tangente hiperbólica acelerar a convergência do algoritmo de treinamento da rede neural[19]. Entretanto, não é claro se diferentes funções de ativação têm maiores efeitos no desempenho da rede.

Ao implementarmos uma rede neural, normalmente o conjunto de dados é separado em dois conjuntos: *conjunto de treinamento* e *conjunto de teste*.

O conjunto de treinamento é utilizado para o treinamento da rede e ajuste dos parâmetros da rede, devendo conter um número estatisticamente significativo de casos em estudo, de modo a constituir uma amostra representativa do problema que se pretende estudar.

O conjunto de teste é utilizado para verificar a capacidade de generalização da rede sob condições reais de utilização. Os dados do conjunto de teste não devem ser usados para ajuste dos parâmetros da rede. A habilidade de generalização da rede se refere a seu desempenho ao classificar padrões do conjunto de teste. Deficiências na capacidade de generalização da rede podem ser atribuídas ao problema de sobreajuste (*overfitting*). Esse problema ocorre quando,

após um certo período de treinamento, a rede se especializa no conjunto de treinamento e perde a capacidade de generalização. Diz-se então que a rede memorizou os padrões de treinamento, gravando suas peculiaridades e ruídos, proporcionando perdas na capacidade de generalização quando essa é utilizada para classificar os padrões pertencentes ao conjunto de teste.

Para evitarmos o problema de sobreajuste, pode-se usar também uma subdivisão do conjunto de treinamento, criando um *conjunto de validação* cuja finalidade é verificar a eficiência da rede quanto a sua capacidade de generalização durante o processo de treinamento, também podendo ser empregado como critério de parada do treinamento da rede. Entretanto, em algumas situações, principalmente na área médica, é comum dispormos de uma pequena quantidade de dados, dificultando a divisão do conjunto de dados em três conjuntos (treinamento, validação e teste). Quando a conjunto de dados não é suficientemente grande para o dividirmos em três conjuntos, uma das alternativas para evitarmos o problema de sobreajuste é utilizarmos o método de validação cruzada (*cross validation*)[21,23].

Após especificarmos a arquitetura da rede neural, torna-se necessário definir o algoritmo de treinamento da rede. Basicamente, o treinamento da rede neural consiste em um problema de minimização não linear sem restrições, em que os pesos sinápticos da rede são iterativamente modificados para minimizar o erro médio quadrático entre a resposta desejada a partir dos dados de entrada, e a saída obtida no neurônio de saída. Do ponto de vista estatístico, o treinamento seria estimar os parâmetros do modelo considerando-se um conjunto de dados[20].

Vários métodos para treinamento supervisionado de RNA são propostos[24]. Entretanto, o algoritmo mais popularmente usado para esse tipo de treinamento é o algoritmo de retropropagação (*backpropagation*)[5].

A aplicação do algoritmo *backpropagation* requer a escolha de um conjunto de parâmetros (número de iterações do algoritmo, critério de parada, pesos iniciais, taxa de aprendizado), cuja influência pode ser decisiva para a capacidade de generalização da rede.

O critério de parada do treinamento exige considerar a capacidade de generalização da rede. Um treinamento prolongado demais pode levar a um sobreajuste da rede, especialmente no caso de dispormos de poucos pares de entrada e saída para o conjunto de treinamento, o que pode piorar o desempenho da rede quando o conjunto de teste lhe for apresentado[25].

A escolha da taxa de aprendizado α depende da função a aproximar. Valores muito pequenos de α tornam o treinamento lento, enquanto valores muito grandes podem provocar divergência do processo de treinamento[26].

A rede neural baseia-se nos dados a ela exibidos para extrair o modelo desejado. Portanto, a fase de treinamento deve ser rigorosa e verdadeira, a fim de serem evitados modelos espúrios.

Descrição dos Modelos Estudados

Primeiramente, para identificarmos as variáveis relevantes ao problema em estudo, foi construído um modelo de regressão logística. Identificadas as variáveis relevantes, implementamos uma rede neural artificial para processar as variáveis escolhidas, a partir de um treinamento supervisionado.

Nesta seção, são descritos ambos os modelos de regressão logística (pré-processamento) e neural (classificação).

Regressão Logística

Em nosso estudo, a variável de interesse é definida da seguinte forma:

$$y_i = \begin{cases} 1, \textit{ se o paciente é soropositivo} \\ 0, \textit{ se o paciente é soronegativo} \end{cases}$$

Seja $\pi_i = P(y_i=1)$, com $0 < \pi_i < 1$, 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 π_i, denotada por,

$$y_i \sim \text{Ber}(\pi_i) \quad (3)$$

Assim, a probabilidade π_i do indivíduo *i* ser soropositivo está relacionada com as variáveis explicativas $x_{i1}, x_{i2}, ..., x_{ip}$ através do seguinte modelo logístico:

$$log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i + ... + \beta_p x_{ip}, \quad (4)$$

sendo β_0, β_1, ..., β_p parâmetros desconhecidos, devendo, assim, ser estimados.

Inicialmente, para avaliarmos a variação do risco na probabilidade de ocorrência da hepatite A, criamos um modelo logístico com as 66 variáveis explicativas observadas nos 3.079 indivíduos residentes nos domicílios amostrados. Em seguida, através deste modelo, selecionamos as variáveis independentes que foram estatisticamente significantes.

A seleção das variáveis do modelo logístico foi realizada por um procedimento "passo a passo", conhecido como *stepwise*[26], 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 socioeconômicas. Através do modelo de regressão logística, foi possível identificar, por exemplo, as variáveis relacionadas ao ambiente domiciliar e peridomiciliar que estão associadas a uma maior ou menor probabilidade de se 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 de estudo e renda média mensal da dona de casa. Sobre esta informação relevante ao diagnóstico será desenvolvido o modelo neural, esperando-se, assim, obter uma modelagem parcimônia e com boa generalização.

Implementação do Modelo Neural

Para definirmos os conjuntos de treinamento, validação e teste a serem utilizados no projeto da 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 ficou reduzida a 2.815 indivíduos. Nessa amostra verificou-se uma soroprevalência global para hepatite A de 36,6%.

Na nossa aplicação, o conjunto de dados foi dividido aleatoriamente em três conjuntos. O conjunto de treinamento foi composto com 1.200 indivíduos, sendo que 33% dos indivíduos são soropositivos para o anti-HAV, enquanto o conjunto de validação incluiu 762 indivíduos, dos quais 38% são soropositivos para o anti-HAV. Por fim, o conjunto de teste foi composto pelos indivíduos restantes, que, portanto, não participaram do processo de treinamento. Os resultados de classificação serão dados para o conjunto de teste. A prevalência de indivíduos soropositivos nesse conjunto é igual a 34%.

O modelo neural proposto se apóia na seleção de variáveis produzida pelo modelo de regressão logística. Assim, a informação de cada paciente a ser processada pela rede neural compõe-se das sete variáveis explicativas identificadas como estatisticamente significantes pelo modelo de regressão logística. Isto permite uma redução significativa da dimensão do espaço de entrada e a conseqüente redução da complexidade da rede neural.

No conjunto de variáveis significativas, as variáveis contínuas foram re-escaladas, em cada conjunto, para se adaptarem ao intervalo (-1, 1) da função de ativação escolhida para cada neurônio, no caso, a tangente hiperbólica. A variável re-escalada x_i^* foi obtida da seguinte forma:

$$x_i^* = \frac{(x_i - \bar{x})}{\max(x_i - \bar{x})} \quad (5)$$

sendo $\bar{x}$ a média aritmética da variável explicativa x_i no conjunto considerado. As

variáveis dicotômicas foram codificadas em -1 e 1, onde -1 representa a ausência do atributo e 1 representa a sua presença.

A escolha do número de neurônios na camada escondida foi feita através de experimentos, sempre buscando-se redes com poucos neurônios escondidos e com um bom poder de generalização. Primeiramente, projetamos uma rede neural com três neurônios escondidos, que apresentou um bom poder de generalização. Também foram avaliadas redes neurais com quatro e cinco neurônios na camada escondida; entretanto, algumas redes ficaram bastante especializadas no conjunto de treinamento. Com isso, resolvemos considerar redes neurais com apenas três neurônios na camada escondida.

O único neurônio de saída da rede corresponde à soropositividade para o anti-HAV, que possui um caráter binário. Em ambas as camadas, a escondida e a de saída, a tangente hiperbólica foi a função de ativação utilizada, visto que as variáveis de entradas se encontram no intervalo entre -1 e 1. Para o treinamento da rede, utilizamos o algoritmo de retropropagação (*backpropagation)*.

Para evitarmos efeito de sobreajuste, utilizamos, como critério de parada, o erro de generalização obtido para o conjunto de validação ao longo do processo de treinamento. O treinamento da rede foi interrompido quando o erro do conjunto de validação passou a aumentar, enquanto o erro do conjunto de treinamento ainda decaía. O treinamento da rede foi realizado considerando-se a taxa de aprendizado fixa em 0,1.

Algumas técnicas são tipicamente utilizadas para acelerar o algoritmo *backpropagation* e evitar mínimos locais. A mais utilizada é a adição do termo momento[5]. Na prática, normalmente escolhe-se o termo momento na faixa de 0,9 ou 0,7. Aqui, considerou-se tal parâmetro igual a 0,7.

Os pesos iniciais foram selecionados aleatoriamente no intervalo de -0,1 a 0,1. Dado que o algoritmo *backpropagation* está sujeito a mínimos locais, foram realizadas diversas inicializações da rede, a fim de evitarmos uma estimativa tendenciosa do erro de classificação. Foram consideradas 10.000 iterações de treinamento.

O desempenho das redes treinadas, após o processo de treinamento, foi avaliado em relação à classificação dos indivíduos pertencentes ao conjunto de teste. Para isto, calculamos a taxa de classificação correta (acurácia sobre os indivíduos), a sensibilidade e a especificidade no conjunto de teste.

Para termos uma base de avaliação do erro de classificação obtido pela rede neural, ajustamos um modelo logístico contendo as sete variáveis independentes em questão. Os parâmetros do modelo logístico foram estimados utilizando-se apenas o conjunto de treinamento. Para a validação do modelo logístico, utilizamos o conjunto de teste.

Resultados

A proporção de concordância total, ou seja, a proporção de indivíduos no conjunto 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.

Ao selecionar um modelo, é bastante importante que o especialista conheça a sensibilidade e a especificidade[27] do modelo que se propõe a utilizar. Desta maneira, para avaliarmos mais detalhadamente o modelo proposto, calculamos a sensibilidade e especificidade da RNA e, para fins de comparação, do modelo logístico (ML). A Tabela 1 mostra os resultados obtidos.

De acordo com os resultados apresentados na Tabela 1, observamos que a RNA apresenta boa sensibilidade, ou seja, ela consegue classificar eficientemente os soropositivos para hepatite A. Em relação à especificidade, ambos os modelos apresentaram uma especificidade bastante elevada.

No caso da identificação de indivíduos soropositivos para hepatite A, a escolha das variáveis de entrada – no nosso caso, através do modelo de regressão logística – realiza um pré-processamento fundamental para a

Tabela 1 – Resultados para Rede Neural e Regressão Logística
Table 1 *– Result for the neural network and Logistic Regression*

	Rede Neural	Regressão Logística
Taxa de Erro (%)	12	17
Sensibilidade (%)	70	52
Especificidade (%)	99	99

classificação. Com este enfoque, o classificador neural proposto foi capaz de classificar corretamente 88% do conjunto de teste.

Conclusões

Redes neurais artificiais vêm sendo utilizadas como modelo de classificação no campo da epidemiologia. 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 utilizando todo o conhecimento acumulado.

O sistema proposto neste trabalho possui uma base de dados de desenvolvimento e a utiliza para prover um modelo neural para o diagnóstico de hepatite A. A rede neural assim projetada, quando alimentada com dados de um novo indivíduo, proveniente da região em estudo, fornece a classificação do indivíduo (soropositivo ou soronegativo) e a probabilidade de o indivíduo ser soropositivo, permitindo uma melhor identificação da condição específica do indivíduo. Além disso, o banco de dados de desenvolvimento do modelo neural pode ser atualizado, permitindo melhorar a caracterização estatística do diagnóstico da hepatite A ou incorporar nova informação, de acordo com a dinâmica que a transmissão da doença possa desenvolver. O sistema neural também poderá ser validado em outras populações com características socioeconômicas semelhantes à região em estudo.

Neste estudo, para identificarmos as variáveis relevantes utilizamos o modelo de regressão logística; entretanto, vários métodos para seleção das variáveis de entrada da rede neural têm sido propostos.

Seixas *et al.*[28] apresentam uma metodologia para a identificação das variáveis relevantes para o problema em estudo. A relevância de uma determinada variável é avaliada através de uma determinada estatística, que mede a variação produzida na saída da rede, quando o valor de uma variável específica de entrada é substituído pelo seu respectivo valor médio, para todos os eventos do conjunto de treinamento. Através dessa metodologia torna-se possível uma visualização das variáveis relevantes. Os métodos AIC[29] (Akaike Information Criterion) e BIC[30] (Bayesian Information Criterion), também são bastante utilizados como procedimento de seleção das variáveis de entrada. A relevância das variáveis explicativas que alimentaram a rede neural implementada neste estudo pode ser verificada utilizando-se algumas dessas técnicas.

A escolha das variáveis relevantes na implementação das redes neurais deve ser realizada cuidadosamente, visto que a inclusão de variáveis não relevantes ao problema em estudo poderá prejudicar o desempenho da rede neural, assim como o erro de classificação.

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. O modelo neural, aqui proposto, pode ser conduzido de forma inovadora como ferramenta de apoio para o diagnóstico da hepatite A, visando respostas úteis ao gerenciamento da hepatite A.

Agradecimentos

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

Referências

1. Villar LM, Paula VS, Gaspar AMC. Seasonal variation of hepatitis A virus infection in the city of Rio de Janeiro, Brazil. *Rev Inst Med Trop São Paulo* 2002; 44(5): 289-92.

2. Wakimoto MD, Arzochi KBF, Hartz ZM. A. Avaliação do sistema de vigilância epidemiológica no município do Rio de Janeiro. In: *Anais do 4º Congresso Brasileiro de Epidemiologia*; Rio de Janeiro: ABRASCO; 1998. p. 257.

3. Hosmer DG, Lemeshow S. *Applied Logistic Regression*. New York: Wiley-Interscience; 1998. p. 131.

4. Medronho RA. *Avaliação do Método Geoestatístico na Distribuição Espacial da Hepatite A* [tese de doutorado], Rio de Janeiro: Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz; 1999.

5. Haykin S. *Neural Networks: A comparative Foundation*. New Jersey: Prentice Hall; 1999.

6. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000; 19: 541-61.

7. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996; 49: 1225-31.

8. Littman E. Trends in neural networks research and an application to computer vision. In: Becks KH, Perret-Gallix D (eds). *New Computing Techniques in Physics Research III* 1994. World Scientific. p. 253-62.

9. Medeiros MC, Terasvita T, Rech G. Building neural networks models for time series: a statistical approach. *SSE/EFI Working Paper Series in Economics and Finance* 2002; 508: 1- 47.

10. Duh M-S, Walker AM, Pagano M, Krounlund K. Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorders as an example, *Am J Epidemiol* 1998; 147: 407-13.

11. Edwards DF, Hollingsworth H, Zazulia AR, Diringer MN. Artificial neural networks improve the prediction of mortality in intracerebral hemorrhage. *Neurology* 1999, 53: 351-6.

12. Lapper RJA, Dalton KJ, Prager RW, Forsstrom JJ, Selbmann HK, Derom R. Application of neural networks to the ranking of perinatal variables influencing birthweight, *Scand J Clin Lab Invest* 1995; 55: 83-93.

13. Ripley RM. *Neural Network Models for Breast Cancer Prognosis* [tese de doutorado], University of Oxford; 1998.

14. Hammad TA, Abdel-Wahab MF, El-Sahly A, El-Kady N, Strickland GT. Comparative evaluation of the use of artificial neural networks for modelling the epidemiology of schistosomiasis mansoni, *Trans R Soc Trop Med Hyg* 1996; 90: 372-6.

15. El-Solh AA, Hsiao C-B, Goodnough S, Serghani J, Grant BJB. Predicting active pulmonary tuberculosis using an artificial neural network. *Chest* 1999, 116: 968-73.

16. Braga AP, Carvalho APLF, Ludemir TB. *Fundamentos de Redes Neurais Artificiais*, Rio de Janeiro: 11ª Escola de Computação; 1998.

17. Murtgath, F. , Neural networks and related "massively parallel" methods for statistics: a short review, *Int Stat Rev* 1994; 62: 275-88.

18. Rocha EC, Pereira BB. Métodos automáticos de previsão para séries temporais multivariadas, *Rev Bras Estat* 1997; 58(209): 105-46.

19. Pereira BB., *Introduction to Neural Networks in Statistics*, Center of Multivariate Analysis, Technical Report; Penn. State University; 1999.

20. Santos AM. *Redes neurais e árvores de classificação aplicadas ao diagnóstico da tuberculose pulmonar paucibacilar* [tese de doutorado]. Rio de Janeiro: COPPE/ UFRJ; 2003.

21. Klimasaukas CC. Applying neural networks, Part 3: Trainning a neural network, *Proceedings in Artificial Intelligence* 1991; 20-24.

22. Stone M. Cross-Validatory choice and assessment of statistical predictions, *J R Stat Soc* 1974; 36: 111-47.

23. Hertz J, Krogh A, Palmer RG. *Introduction to the Theory of Neural Computation*, New York Addison-Wesley Publishing Company; 1991.

24. Smith M. *Neural Network for Statistical Modelling*, New York: Van Nostrond Reinhold; 1993.

25. Calôba LP. Introdução à computação neuronal. In: *Anais do 9º Congresso Brasileiro de Automação*; Vitória-ES, 1995. p 25-38.

26. Kleinbaum DG. *Logistic Regression - A Self-Learning Text*. New York: Spring Verlag;1994.

27. Rothaman KJ, Greenland S. *Modern Epidemiology*. United States of America: Lippincott-Raven Publishers HL; 1998.

28. Seixas JM, Calôba LP, Delpino I. Relevance criteria for variable selection in classifier designs, International Conference on Engineering Applications of Neural Networks; 1996: 451-4.

29. Akaike H. A new look at the statistical model identification, *IEEE Transaction on Automatic Control* 1974; 6(19): 716-23.

30. Schwarz G. Estimating the dimension of a model. Ann Stat 1978: 6; 461-4.

recebido em: 19/12/04
versão final reapresentada em: 05/05/05
aprovado em: 02/06/05