# MODELO NEURAL PARA O SUPORTE AO DIAGNÓSTICO DE TUBERCULOSE PULMONAR NA TRIAGEM DE PACIENTES

F. A. A. Andrade*, J. B. O. Souza Filho*, R. M. Galliez** e Afrânio Kritski**

*LAPSI/DEPEL/CEFET-RJ, Rio de Janeiro, Brasil
** PAT/FM/UFRJ, Rio de Janeiro, Brasil
e-mail: fabio@ieee.org, jbfilho@gmail.com, galliez77@gmail.com, kritskia@gmail.com

**Resumo:** Doenças infecciosas como a tuberculose necessitam de uma rápida identificação de pacientes infectados, portanto ferramentas para o apoio a triagem de pacientes são de elevada utilidade. Este trabalho propõe um Sistema de Apoio ao Diagnóstico de pacientes suspeitos de tuberculose pulmonar baseado em Redes Neurais Artificiais. Como o objetivo de produzir um sistema acurado, três tecnologias são avaliadas: o algoritmo de *Backpropagation* (BP), a Máquina de Aprendizado Extremo (ELM) e Máquina de Aprendizado Extremo Otimamente Podada (OP-ELM). Resultados apontam que as redes derivadas pela técnica OP-ELM apresentam uma generalização semelhante à técnica BP, alcançando uma sensibilidade de 90% e especificidade de 50%.
**Palavras-chave:** Suporte ao Diagnóstico, Tuberculose, Redes Neurais Artificiais, Retropropagação, Máquina de Aprendizado Extremo Otimamente Podada.

***Abstract:*** *Infectious diseases such as tuberculosis require a fast identification of infected patients, thus tools to support patient screening are extremely useful. This work proposes a Diagnostic Support System for patients suspected of pulmonary tuberculosis based on Artificial Neural Networks. Aiming at producing an accurate system, three developing methodologies are evaluated: Backpropagation algorithm (BP), Extreme Learning Machines (ELM) and Optimally Pruned Extreme Learning Machines (OP-ELM). Results show that networks derived by OP-ELM technique show a similar generalization capability than BP technique, achieving a sensitivity of 90% and specificity of 50%.*
***Keywords:*** *Diagnosis Support, Tuberculosis, Neural Networks, Backpropagation, Optimally Pruned Extreme Learning Machine.*

## Introdução

A tuberculose (TB) é uma das enfermidades infectocontagiosas que mais ocasiona mortes no mundo. Estima-se que aproximadamente 30% da população mundial está infectada, porém nem todos desenvolvem a doença [1]. O Brasil se encontra entre os 22 países de maior carga de TB a nível mundial, possuindo cerca de 70.000 casos de TB e 4.000 mortes por ano [2].

Os exames diagnósticos realizáveis (baciloscopia e cultura), quando disponíveis, não são apropriados para uma rápida e eficaz identificação de pacientes com TB positiva, propiciando a transmissão do seu bacilo causador na comunidade. Por isso, Sistemas de Apoio ao Diagnóstico (SAD) podem constituir uma importante ferramenta para agilizar a tomada de decisão clinica, caso sejam desenvolvidos de forma criteriosa e com base em dados clínicos confiáveis e representativos.

O acesso a dados clínicos e laboratoriais de pacientes suspeitos de TB permite o desenvolvimento de sistemas de apoio à decisão, em especial por meio da adoção de algoritmos eficazes de Inteligência Computacional. No entanto, a produção de modelos preditivos para a identificação de pacientes com TB ativa encontra dificuldades devido à complexidade do diagnóstico [3]. Entre fatores desafiadores, citamos: o elevado número de sintomas, a baixa qualidade dos dados coletados em condições de rotina, as restrições de cobertura comuns nas amostras disponíveis para a caracterização do problema, bem como a exigência de elevada acurácia do sistema final, visto influenciar a tomada de decisões com risco de vida.

Redes Neurais Artificiais (RNA), em especial a topologia *Multilayer Perceptron* (MLP), consistem numa abordagem atrativa ao desenvolvimento de SAD. RNAs são hábeis na solução de problemas complexos, obtida através de processamento não linear, definido por um processo de aprendizado baseado em exemplos. Este fato tem motivado sua aplicação em problemas de diversas áreas do conhecimento nos últimos anos [4]. Recentemente, as Máquinas de Aprendizado Extremo (*Extreme Learning Machines* - ELM) [5] têm sido propostas como uma importante alternativa no uso de redes MLP, face à simplicidade do processo de treinamento e ao bom desempenho comumente obtido.

Visando desenvolver um sistema preditivo acurado para a identificação de pacientes com tuberculose pulmonar, a ser utilizado no apoio à triagem em ambiente ambulatorial em Unidades de Saúde de nível secundário, buscou-se avaliar a técnica de ELM como alternativa ao processo usual de treinamento de redes MLP com uma única camada de neurônios ocultos (*Single Hidden Layer Feedforward Network* – SLFN). Esta comparação empregou a técnica de validação cruzada e o método de teste de hipóteses para uma validação estatística dos resultados.

Este trabalho é iniciado com uma discussão dos métodos de treinamento utilizados no desenvolvimento de SLFN's e demais materiais utilizados. Em sequência, os resultados obtidos são apresentados e discutidos. Por

fim, tem-se a conclusão e as perspectivas para os trabalhos futuros.

## Materiais e métodos

**Redes Neurais Artificiais –** Uma rede MLP possui como elemento principal o neurônio artificial, que basicamente realiza uma soma ponderada dos sinais de entrada que o alimentam, aplicando este resultado a uma função de ativação a fim de definir sua saída. Nesta rede, que é ilustrada na Figura 1, estes neurônios são dispostos em um número arbitrário de camadas, porém as SLFN's possuem apenas uma camada oculta, também referida como intermediária ou escondida [6].

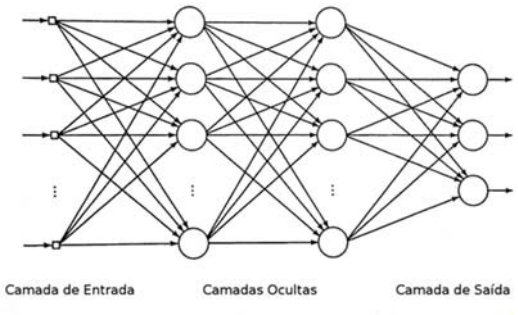


Figura 1: Estrutura de uma Rede Neural MLP

O algoritmo mais utilizado para o treinamento das SLFN's é o BP. A despeito de eficaz, o BP é um algoritmo iterativo que, para obter soluções acuradas, demanda o ajuste de uma série de parâmetros de treinamento, usualmente realizado por meio da produção e avaliação de um expressivo número de redes. Assim, o esforço computacional normalmente envolvido é elevado.

A técnica ELM consiste numa importante alternativa a BP, que visa reduzir os problemas apontados. Sua principal desvantagem, no entanto, reside no fato de usualmente empregar um expressivo número de neurônios na camada intermediária. Em muitos casos, esta maior complexidade implica num menor poder de generalização da ELM quando comparada com a BP. Assim, o principal diferencial do algoritmo OP-ELM [7] é realizar uma análise de relevância quanto à utilidade dos neurônios da camada intermediária, obtendo soluções mais compactas e de melhor desempenho que a ELM. A seguir serão destacadas algumas características importantes dos métodos citados.

**BP** – Neste algoritmo, os pesos e limiares associados aos neurônios das camadas internas e mais exteriores são modificados através da retropropagação do erro. Várias funções objetivo e algoritmos de otimização podem ser utilizados. Neste trabalho foi utilizado o valor médio do erro quadrático com função objetivo e o algoritmo de otimização de Levenberg-Marquardt [8]. Para se evitar *overtraining* [6] foi utilizada a técnica de parada antecipada [6], enquanto o quantitativo de neurônios na camada intermediária foi sintonizado pela técnica de validação cruzada [9].

**ELM** – Sua principal estratégia é estimar os valores dos pesos e limiares entre as camadas de entrada e intermediária aleatoriamente, permitindo que os pesos associados à camada de saída sejam determinados através da solução de um sistema linear sobredeterminado, realizado por meio do cálculo de uma matriz pseudo-inversa numa única iteração. Outro aspecto interessante diz respeito à flexibilidade relativa à escolha da função de ativação dos neurônios da camada intermediária, visto que a condição imposta pelo BP de que tais funções sejam diferenciáveis pode ser relaxada. Novamente o número de neurônios na camada intermediária foi determinado por validação cruzada.

**OP-ELM** – Esta técnica é baseada no algoritmo ELM tradicional e inclui passos adicionais visando à poda de neurônios na camada intermediária, o que torna o modelo mais compacto, robusto e com melhor capacidade de generalização [7]. O OP-ELM é baseado na ELM usual, porém pode empregar função de ativação gaussiana, sigmoide ou linear em cada um dos neurônios, bem como aplica a técnica de Regressão Multiresposta Esparsa (*Multiresponse Sparse Regression* - MRSR) para ranqueá-los conforme sua utilidade. Posteriormente, a técnica *Leave-One-Out* (LOO) é utilizada para determinar uma quantidade ótima destes neurônios por validação cruzada.

**Base de dados –** A base de dados utilizada foi coletada na Policlínica Augusto Amaral Peixoto (PAAP/SMS-RJ), responsável pela Área de Planejamento de Saúde (AP) 3.3. As entrevistas foram realizadas entre 2006 e 2009, com pacientes suspeitos de TB pulmonar que aceitaram participar do estudo.

A base é composta por 2.469 sujeitos, sendo 716 com TB ativa (TB Positivo) e 1.753 sem TB ativa (TB Negativo), e possui 280 variáveis, tais como: datas, sintomas, resultados de exames, entre outras informações.

**Comitê de ética –** Este estudo foi aprovado pelo Comitê de Ética Nacional (processo nº 67/06) e está em cumprimento aos princípios éticos contidos na Declaração de Helsinki.

**Seleção de variáveis –** Para a seleção das variáveis foi utilizada a técnica estatística de Regressão Logística, seguida de crítica especialista. Segundo esta técnica é possível produzir um modelo que permita a predição de valores representados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias [10]. De posse de um subconjunto de variáveis identificadas como chave pelos especialistas (variáveis independentes), associações individuais com o desfecho (variável dependente) foram derivadas, bem como o nível de significância foi inferido, de forma a balizar a seleção daquelas mais relevantes do modelo. Esta análise utilizou o programa *Statistical Package for the Social Sciences* (SPSS).

**Validação Cruzada –** Para a avaliação dos classificadores foi utilizado o método de validação cruzada *k-fold* [9], realizado em dois estágios. No

primeiro, foi considerado um valor de *k* igual a 4, portanto 3 *folds* foram destinados aos conjuntos de treinamento e validação, e 1 *fold* para o conjunto de teste. Em sequência, para a separação dos conjuntos de treino e validação, foi utilizado um valor de *k* igual a 5, tomando-se o conteúdo dos 3 *folds* anteriormente citados, o qual foi subdividido em 4 *folds* para o treinamento e 1 *fold* para a validação; o último utilizado para o dimensionamento das redes e pelo critério de parada antecipada.

**Índice de Desempenho –** A comparação empregou a área sob a curva ROC (*Area Under roc Curve* - AUC) [11] como índice de desempenho. A curva ROC (*Receiver Operation Curve*) permite observar, simultaneamente, como se comportam a sensibilidade e a especificidade de um classificador binário quando o seu limiar de decisão é modificado. Adicionalmente, o número de neurônios da camada intermediária foi sintonizado por validação cruzada, de forma a maximizar a AUC associada aos classificadores produzidos para o conjunto de validação.

## Resultados

Para a seleção das variáveis integrantes do modelo foram analisados os valores de significância (*p*) obtidos por meio da técnica de regressão logística, sendo eliminadas aquelas com valor de *p* inferior a 0,2. Os 15 sintomas identificados como relevantes foram: Idade, Cicatriz BCG, Sexo, Raça, Solteiro, Expectoração, Hemoptoicos, Hemoptise, Sudorese Noturna, Febre, Dispneia, Perda de Peso (10%), Contato TBP 2 anos, Fumante e Tosse há mais de uma semana.

Para cada um dos vinte subconjuntos de validação, foram treinadas cinco SLFN, considerando os métodos BP, ELM e OP-ELM, cada uma empregando parâmetros iniciais de rede sorteados aleatoriamente; e escolhido, para cada caso, o ensaio correspondente à média do valor de AUC, inferido através do conjunto de validação. Posteriormente, foi selecionada a SLFN correspondente a média do valor da AUC entre os cinco subconjuntos de validação. Todo este processo foi ainda executado dez vezes, resultando em dez SLFN's para cada um dos quatro conjuntos de teste. O diagrama de caixas [12] dos resultados obtidos é apresentado na Figura 2.

Através da Figura 2 é possível observar que a dispersão dos valores para o método BP é menor que a obtida para as outras técnicas. Ocorre, porém, que tanto a mediana quanto os quartis superiores (75% e 95%) associados ao método OP-ELM são superiores aos demais métodos.

Para a análise estatística dos resultados foram realizados o teste de comparação de médias ANOVA e a técnica de comparações múltiplas de Tukey [12], ambos considerando um nível de significância de 0,05 (5%). Os resultados obtidos para o teste de Tukey são apresentados na Tabela 1. Pode-se observar que a técnica OP-ELM possui uma diferença média positiva em relação ao ELM e ao BP. Esta diferença é bem mais significativa para o método ELM do que para o BP. Observa-se, no entanto, que a hipótese de igualdade entre as médias do OP-ELM e BP não deve ser descartada, devido ao valor de *p* alto (*p*=0,812), mostrando que o OP-ELM, técnica mais rápida e mais simples pode ser utilizada como uma alternativa eficiente ao BP no desenvolvimento do SAD em análise.

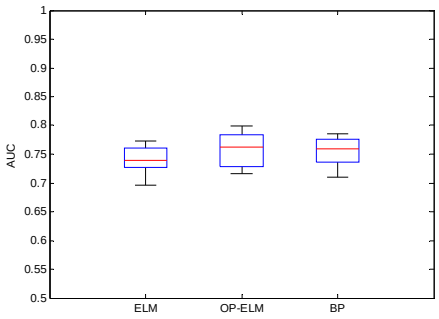


Figura 2: Diagrama de caixas (*boxplot*) para os métodos em análise (vide o texto).

Tabela 1: Resultados do teste de Tukey na comparação dos métodos

| Referência | Alternativa | Diferença Média | Sig. |
|---|---|---|---|
| ELM | OP-ELM | -0,017965 | 0,007 |
| | BP | -0,014398 | 0,038 |
| OP-ELM | ELM | 0,017965 | 0,007 |
| | BP | 0,003567 | 0,812 |
| BP | ELM | 0,014398 | 0,038 |
| | OP-ELM | -0,003567 | 0,812 |

Quanto às redes geradas, no que se refere à quantidade de neurônios, às suas funções de ativação, às conexões e aos pesos, o BP empregou redes com apenas dois neurônios na camada intermediária com função de ativação sigmoide para todos os neurônios. Para o OP-ELM, por sua vez, as redes possuíam, em média, dezesseis neurônios na camada intermediária com funções de ativação diversas: gaussiana, sigmoide ou linear. Já para o ELM, as redes empregaram, em média, 48 neurônios na camada intermediária, todas com função de ativação do tipo sigmoide.

A curva ROC associada ao ensaio correspondente à média do valor de AUC para a técnica OP-ELM é exibida na Figura 3. É possível observar que o escore proposto atinge uma sensibilidade 90% e especificidade de 50%, obtido para um valor de limiar igual a -0,30.

## Discussão

Para o desenvolvimento do SAD, o BP é uma técnica que envolve uma definição cuidadosa de uma série de parâmetros de projeto, entre eles: o número de neurônios, a função de ativação, o critério de parada, o algoritmo de treinamento, o processo de cálculo de gradiente (época, batelada, LMS) e a inicialização dos parâmetros. Tais parâmetros são usualmente definidos

por tentativa e erro, o que usualmente resulta na produção e avaliação de centenas de redes. O treinamento por BP é ainda um processo iterativo e comumente lento. Adicionalmente, por se basear em gradiente, as funções de ativação dos neurônios têm que ser diferenciáveis. Por outro lado, para o ELM, é apenas obrigatória a definição do número de neurônios e de uma inicialização apropriada dos parâmetros da camada escondida. O ajuste dos pesos e limiares remanescentes é realizado por um processo determinístico (pseudo-inversa) de passo único e há uma grande liberdade quanto à escolha das funções de ativação. Como contrapartida, o ELM, por se basear em *features* aleatoriamente extraídos, costuma empregar um maior número de neurônios, muitos deles irrelevantes, e apresentar uma menor capacidade de generalização. Assim, um interessante compromisso entre esforço computacional, desempenho e complexidade da rede pode ser obtido pela técnica OP-ELM, que produziu rede mais compactas que a técnica ELM, porém com desempenho estatisticamente similar à técnica BP para o problema em estudo.

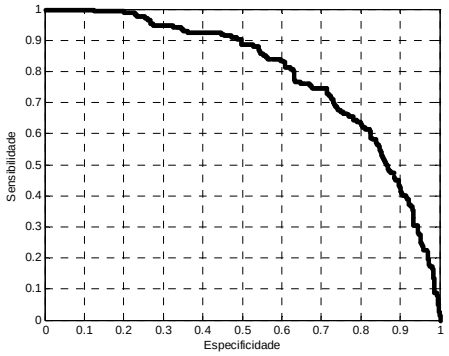


Figura 3: Curva ROC para o modelo OP-ELM selecionado (vide texto).

## Conclusão

Neste artigo foi discutido o desenvolvimento de um sistema preditivo acurado, baseado em redes neurais artificiais do tipo SLFN, para a identificação de pacientes com TB pulmonar, a ser utilizado no apoio à triagem de pacientes em ambiente ambulatorial. Três métodos de treinamento foram avaliados: o BP, o ELM e o OP-ELM. Através dos testes estatísticos se observou que a técnica OP-ELM permite a derivação de classificadores de acurácia similar a BP, porém através de um processo de treinamento significativamente mais simples. O modelo OP-ELM selecionado atingiu uma sensibilidade de 90% e especificidade de 50%, resultado que é expressivo.

Como trabalhos futuros se pretende utilizar mecanismos de seleção de instâncias para retirar as eventuais amostras que possam estar prejudicando o treinamento das redes. A formação dos conjuntos de treino, validação e teste com base em análise de agrupamentos também será investigada. Algoritmos de otimização, tais como a técnica de Otimização por Enxames de Partículas (PSO), também serão avaliados para a otimização dos parâmetros da rede neural, visto os excelentes resultados reportados em problemas de domínios variados [13].

## Agradecimentos

Agradecemos a CAPES, CNPQ e a FAPERJ pelo apoio financeiro concedido.

## Referências

[1] Ferreira, C.D.A. (2011), Integração de dados de expressão gênica global em tuberculose. Dissertação de Mestrado. Instituto Oswaldo Cruz.

[2] World Health Organization (2014), ‘World Health Statistics 2014’.

[3] El-Solh, A.A., Hsiao, C.B., Goodnough, S, Serghani, J., Grant, B.J. (1999), ‘Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network’. Chest, 116:968-973.

[4] Hush, D.R., Horne, N.G. (1993), ‘Progress in supervised neural networks’. IEEE Signal ProcessMag 10:8–39.

[5] Huang, G.-B., Zhu, Q.-Y., Siew, C.-K. (2006), ‘Extreme Learning Machine: Theory and Applications’. Neurocomputing 70, 489-501.

[6] Haykin, S. (2008), ‘Neural Networks: A comprehensive Foundation’, 2ed. New Jersey, Prentice-Hall.

[7] Miche Y, Sorjamaa A, Bas P, Simula O, Jutten C, Lendasse A (2010). ‘OP-ELM: optimally pruned extreme learning machine’. IEEE Trans Neural Netw 21(1):158–162

[8] Hagan, M.T., and M. Menhaj, "Training feed-forward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, No. 6, 1999, pp. 989–993, 1994.

[9] Tan , P., Steinbach, M., Kumar, V. (2009), Introdução ao Data Mining. Editora Ciência Moderna.

[10] Gelman, A., Hill, J. (2006). 'Data Analysis Regression Multilevel Hierarchical Models', 1ed, ISBN-10 052168689X

[11] Zhou, Xiao-Hua; Obuchowski, Nancy A.; McClish, Donna K. (2002), ‘Statistical Methods in Diagnostic Medicine’. New York, NY: Wiley & Sons. ISBN 978-0-471-34772-9.

[12] Lowry, R. (2013). 'Concepts and Applications of Inferential Statistics', Poughkeepsie, NY. Vassar College.

[13] Teixeira, L.A., Oliveira, F.T.G., Oliveira, A.L.I., Bastos Filho, C.J.A. (2008). ‘Adjusting Weights and Architecture of Neural Networks through PSO with Time-Varying Parameters and Early Stopping’. 10th Brazilian Symposium on Neural Networks.