ESCORE DE PONTOS PARA A IDENTIFICAÇÃO DE TUBERCULOSE PULMONAR DERIVADO POR INTELIGÊNCIA COMPUTACIONAL

Victor Hugo da Silva Muniz*, João Baptista de Oliveira e Souza Filho**,
Luciana Faletti Almeida**, Fabio Augusto de Alcantara Andrade**,
Afrânio Lineu Kritski*** e Rafael Mello Galliez***

*LAPSI / DEPEL – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, Brasil
**LAPSI / DEPEL / PPEEL – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, Brasil
***Programa Acadêmico de Tuberculose, Faculdade de Medicina – Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brasil

e-mail: vhsm.95@gmail.com

Resumo: A tuberculose é uma das doenças que mais acometem a humanidade devido a sua fácil transmissão através da inalação do agente causador. A ampla adoção de testes diagnósticos esbarra em limitações de custo, tempo para conclusão e acurácia. A proposta deste estudo é derivar um escore de pontos utilizando ferramentas de Inteligência Computacional. Assim, diferentes técnicas serão avaliadas, entre elas: Algoritmo Genético Multiobjetivo e Mono-objetivo, Otimização por Enxame de Partículas, *Simulated Annealing* e Regressão Linear Aproximada. Considerando o formato e a área sob a curva *ROC*, bem como o quantitativo de sintomas envolvidos, a técnica da Regressão Linear Aproximada obtém o melhor desempenho para a base de dados avaliada, alcançando uma sensibilidade de 83,0% e especificidade de 59,6%.
Palavras-chave: Suporte ao Diagnóstico, Tuberculose, Escore Clínico, Regressão Linear, Otimização Natural.

Abstract: *Tuberculosis is one of the diseases that most affects humanity due to its easy transmission through inhalation of the causative agent. A wide adoption of diagnosis tests is hampered by limitations of cost, time to completion and accuracy. The purpose of this study is to derive a point-score system using Computational Intelligence tools. Thus, different techniques will be evaluated, such as: Multiobjective and Mono-objective Genetic Algorithm, Particle Swarm Optimization, Simulated Annealing and Approximate Linear Regression. Considering the format and area under ROC curve as well as the quantity of involved symptoms, Approximate Linear Regression technique performed best for the evaluated data base, achieving a sensitivity of 83.0% and specificity of 59.6%.*
Keywords: *Diagnosis Support, Tuberculosis, Clinical Score, Linear Regression, Natural Optimization.*

Introdução

A tuberculose é uma doença infectocontagiosa causada por uma bactéria que afeta principalmente os pulmões, cuja transmissão ocorre de forma direta através das vias aéreas.

O diagnóstico da doença pode ser feito através de exames como a baciloscopia e a cultura. O primeiro apresenta um elevado falso-negativo. Já o segundo, que possui um melhor desempenho, leva de 4 a 6 semanas para ser concluído, e não é disponibilizado para todas as unidades de saúde.

Neste contexto, escores de pontos podem representar uma ferramenta rápida e eficaz para o diagnóstico, permitindo uma ampla utilização na comunidade médica por sua simplicidade intrínseca. Técnicas de Inteligência Computacional (IC), em especial de otimização natural, são normalmente hábeis na solução de problemas complexos e, portanto, especialmente indicadas para a produção destes escores, desde que derivados utilizando um conjunto de dados certificado por especialistas da área.

A proposta deste trabalho é obter um escore de pontos que auxilie à tomada de decisão médica através de técnicas de IC. A adoção deste método visa reduzir o número de pacientes suspeitos de tuberculose que são enviados ao isolamento respiratório desnecessariamente. Além de permitir uma melhor gestão dos recursos hospitalares, sua adoção pode levar a um decréscimo nas chances de transmissão da tuberculose nestes ambientes. A estrutura do trabalho é a seguinte: inicialmente, é realizada a formulação do escore e descritas, de forma sucinta, as técnicas de IC avaliadas. Em sequência, a base de dados e os principais resultados são apresentados. Encerrando, têm-se as conclusões e trabalhos futuros.

Materiais e métodos

O Escore de Pontos – O escore de pontos é uma expressão algébrica que, com base na presença, ausência ou desconhecimento de um dado conjunto de sinais e sintomas, produz um valor numérico indicativo da probabilidade do paciente apresentar uma determinada doença. Diferentes codificações do escore podem ser empregadas, sendo um total de quatro avaliadas neste trabalho, referidas como A, B, C e D.

Para as codificações A e B, a expressão do escore de pontos equivale a Equação (1), onde w_i representa a pontuação associada ao *i*-ésimo sinal ou sintoma, dentre os n disponíveis. Para a codificação A, a presença, ausência e os casos desconhecidos são indicados pelas variáveis s_i, as quais são atribuídos os valores +1, -1 e 0, respectivamente. No que tange a codificação B, esta se diferencia da A, exclusivamente, por atribuir o valor -1 a variável s_i quando o sinal ou sintoma for desconhecido.

$$y = w_1.s_1 + w_2.s_2 + \ldots + w_{n-1}.s_{n-1} + w_n.s_n + c \qquad (1)$$

Para as codificações C e D, o escore de pontos se resume a Equação (2). Na codificação C, utilizam-se 3 pesos independentes, de acordo com a presença ($w_{i,p}$), ausência ($w_{i,a}$) ou desconhecimento ($w_{i,d}$) de cada sinal ou sintoma ($1 \leq i \leq n$). Assim, as variáveis $s_{i,a}$, $s_{i,p}$ e $s_{i,d}$ assumem os valores 0 ou 1, onde o valor 1 sinaliza cada caso. Com respeito à codificação D, esta opera de forma similar a anterior, porém emprega apenas 2 pesos independentes para cada sinal ou sintoma, considerando que os casos desconhecidos são tratados como ausentes.

$$y = w_{1,p}.s_{1,p} + w_{1,a}.s_{1,a} + w_{1,d}.s_{1,d} + \ldots + c \qquad (2)$$

Para a decisão, se o valor resultante de (1) ou (2) for maior ou igual a um limiar apropriadamente escolhido, o paciente é classificado como tuberculoso, tornando-o apto a ser encaminhado ao isolamento respiratório.

Métodos – Para a obtenção das pontuações por sintomas, bem como para a seleção daqueles constituintes do escore de pontos, foram avaliados os seguintes métodos: Algoritmo Genético Mono-objetivo [1] e Multiobjetivo [2], *Simulated Annealing* [3], Regressão Linear [4] Aproximada e Otimização por Enxame de Partículas [5].

O Algoritmo Genético (AG) é um método de busca inspirado na teoria de seleção natural das espécies proposta por Darwin. Um conjunto de possíveis soluções é tratado como uma população de indivíduos que evolui ao longo do tempo, na qual os mais aptos sobrevivem, passando a fazer parte de uma nova geração. Cada variável define um gene que formará um indivíduo, o qual representa uma possível solução para o problema [6]. A cada geração, os indivíduos da população são avaliados através da função de *fitness*, que é o mecanismo determinante para se selecionar os melhores indivíduos a compor a próxima geração.

Duas modalidades de Algoritmos Genéticos foram utilizadas: o Multiobjetivo (AG-MU) e o Mono-objetivo (AG-MO). O primeiro procurou a maximização conjunta dos valores de sensibilidade (acerto entre pacientes positivos) e especificidade (acertos entre pacientes negativos). Já o segundo, buscou maximizar o valor da área sob a curva *ROC*, considerando a adição de condições restritivas à função de *fitness* (AG-MOR), de modo a reduzir o número de sintomas utilizados no escore.

O *Simulated Annealing* (SA) é um algoritmo de otimização inspirado em um fenômeno físico conhecido como recozimento, típico na têmpera de metais. Seu funcionamento envolve parâmetros como temperatura e energia. Para este trabalho se considerou otimizar o valor da área sob a curva *ROC*.

A Regressão Linear Aproximada (RLA) permite a derivação do escore pela resolução de um sistema de equações sobredeterminado utilizando a técnica de mínimos quadráticos (*ordinary least square*). Neste caso, os valores obtidos no processo foram arredondados para os inteiros mais próximos.

A Otimização por Enxame de Partículas (PSO) é inspirada no comportamento social de pássaros, peixes e enxames. Esta técnica utiliza uma população que explora o hiperespaço do problema a uma velocidade, que é definida de acordo com a melhor posição histórica individual e da vizinhança inferida para uma determinada função objetivo. Assim, o movimento de cada partícula evolui naturalmente para a solução ótima. Neste trabalho, as velocidades das partículas foram arredondadas para números inteiros a cada iteração, de modo que os vetores das posições das partículas fossem sempre compostos por números inteiros.

Base de Dados – A base de dados é composta por 960 pacientes e foi disponibilizada pelo Instituto de Doenças do Tórax, que funciona no Hospital Universitário Clementino Fraga Filho, situado na UFRJ. Deste total, 231 pacientes apresentam tuberculose.

As informações presentes nesta base correspondem aos seguintes 13 sinais e sintomas: Sexo Feminino, Emagrecimento, Tosse, Escarro, Hemoptoico, Sudorese, Dispneia, Febre, HIV+, Tuberculose Pulmonar Anterior, Fumo, Alcoolismo e Tuberculose Ativa no Raio-x.

Resultados

Com o intuito de se comparar a capacidade de generalização dos modelos, foi utilizada a técnica de reamostragem *5-fold* [7]. Assim, o conjunto de dados foi dividido em 5 subconjuntos para a avaliação dos modelos. Face ao comportamento estocástico das técnicas de otimização sob análise, as simulações foram repetidas de 5 a 10 vezes em cada caso.

O Algoritmo Genético Multiobjetivo foi executado 5 vezes para cada *fold*, utilizando uma população inicial de 100 indivíduos, onde 50% foi reservada para a construção da Curva de Pareto. O número de gerações foi definido pela multiplicação da quantidade de variáveis do cromossomo por 200, situando-se entre 2800 e 8000, conforme a codificação empregada.

No que se refere ao Algoritmo Genético Mono-objetivo, este foi repetido 10 vezes, efetuando-se 100 gerações com 20 indivíduos cada.

Em ambos os Algoritmos Genéticos, utilizou-se como taxa os valores: 0,8 para o crossover aritmético [6]; 0,07 para a mutação de dois pontos [6], e 0,2 no *steady-state* [6]. Ademais, arbitrou-se como domínio das variáveis o intervalo de -10 a 10.

A técnica de *Simulated Annealing* considerou a proposta de *Boltzmann* [8] e um total de 5 repetições. O intervalo de busca foi limitado entre -5 e 5, e escolheu-se um decréscimo exponencial da temperatura, de valor inicial 100, determinado por uma progressão geométrica de razão 0,95.

Para a Otimização por Enxame de Partículas, cada conjunto de desenvolvimento foi dividido em três subconjuntos (treino, validação e teste). Arbitrou-se uma população de 40 partículas e 500 épocas de treinamento. A partir da centésima época, a área sob a curva *ROC* do subconjunto de validação foi avaliada, e a iteração de maior valor de área definiu a solução ótima. Este processo foi repetido 5 vezes. Considerou-se o valor 2 para os parâmetros pessoal e social, e um fator de constrição unitário. A componente inercial foi modificada, linearmente, entre 0,4 e 0,9; e para o limitador de velocidade, aplicaram-se os valores -20 e 20. Assumiu-se o intervalo entre 0 e 10 como domínio das variáveis.

Os valores de área sob a curva *ROC* (*AUC*) para as diferentes técnicas são exibidos no *boxplot* [9] da Figura 1. É possível observar que a mediana associada à técnica de Regressão Linear Aproximada é superior às demais para as codificações A, B e D; enquanto, para a codificação C, fato similar ocorre para o Algoritmo Genético Mono-objetivo com Restrição.

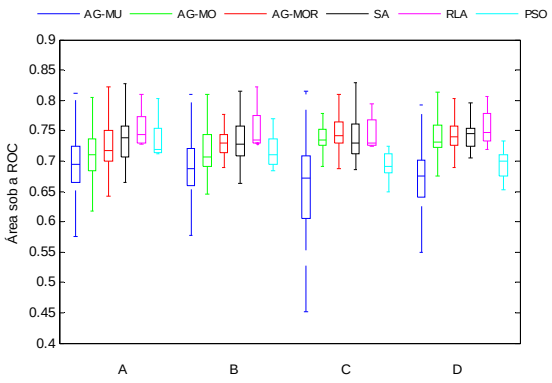

Figura 1: Valores de *AUC* por técnica e codificação.

A Tabela 1 resume os valores de *AUC* e a quantidade de sintomas associados às técnicas de melhor desempenho para cada codificação, enquanto na Figura 2 têm-se as respectivas curvas *ROC*. Pode-se observar que o formato das curvas é bastante similar, bem como os valores de *AUC*, em especial para as codificações A, B e D, que resultam em escores equivalentes do ponto de vista da figura de mérito considerada. Tais escores apresentam uma sensibilidade ligeiramente superior a 80% para uma especificidade em torno de 60%. Diante da equivalência, a opção pela codificação A que envolve um menor número de sintomas se faz mais atrativa.

Tabela 1: Valores por codificação.

Codificação	Técnica	*AUC*	Sintomas Considerados
A	RLA	0,7437	8
B	RLA	0,7344	9
C	AG-MOR	0,7411	12
D	RLA	0,7472	13

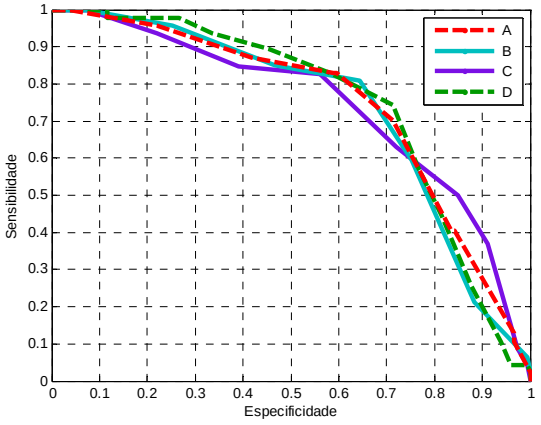

Figura 2: Curvas *ROC* associadas às técnicas de melhor desempenho para cada codificação (vide texto).

Na Tabela 2, os pesos w_i associados à codificação A são exibidos. Este escore de pontos alcança uma sensibilidade de 83,0% e especificidade de 59,6%, ao se considerar o tratamento de pacientes com pontuação igual ou superior a 2.

Tabela 2: Pesos w_i obtidos na técnica de Regressão Linear Aproximada para a codificação A.

Sintoma	Peso w_i
Sexo Feminino	1
Emagrecimento	1
Hemoptoico	-1
Sudorese	1
Dispneia	-1
HIV+	-1
TB Pulmonar Anterior	-1
TB Ativa no Raio-x	3

Discussão

Através da avaliação do formato da curva *ROC* e dos valores de sensibilidade e especificidade a ela associados, bem como o número de sintomas considerados, verifica-se para a base de dados em estudo, um melhor desempenho da técnica de Regressão Linear Aproximada empregando a codificação A.

Cumpre destacar que a simplicidade do escore, além de ser atrativa do ponto de vista estatístico, favorece sua adoção pela comunidade médica.

Critérios clínicos comumente utilizados são demais conservativos e resultam no isolamento desnecessário de cerca de 70% dos pacientes [10]. Ademais, nem sempre podem ser facilmente adaptados a populações com prevalências variadas, ao contrário dos escores que possibilitam ser configurados para diferentes cenários epidemiológicos.

Uma simples adoção deste escore permitiria a redução destes casos para em torno de 40%, quase metade do valor original. Como os leitos de isolamento respiratório são disponíveis em quantidade restrita face ao custo elevado de sua implementação, bem como são utilizados por pacientes portadores de outras doenças, o uso do escore pode levar a uma redução expressiva do quantitativo de falso-positivos. Dessa forma, ocorre uma melhoria na gestão dos leitos hospitalares, acarretando em um atendimento mais eficiente aos pacientes.

Destaca-se, no entanto, a necessidade de estudos adicionais de custo-efetividade, onde o escore proposto tenha seu uso avaliado e validado em condições de rotina.

Conclusão

A tuberculose é uma doença infectocontagiosa facilmente transmissível, cujos exames de detecção apresentam várias limitações de cunho prático que acarretam no isolamento respiratório desnecessário de um número expressivo de pacientes.

O escore de pontos é um método simples, rápido e de baixo custo para se diagnosticar a tuberculose. Derivado através de ferramentas de Inteligência Computacional, este pode prover relevante auxílio à tomada de decisão médica.

Dentre as técnicas avaliadas, a Regressão Linear Aproximada permitiu a construção de um modelo que envolve apenas 8 sintomas e atinge 83,0% de sensibilidade e 59,6% de especificidade, caso se proceda o isolamento para um valor de escore superior a 2.

Como trabalhos futuros, pretende-se testar o escore em outros conjuntos de dados, bem como explorar a produção de modelos especializados em subpopulações, entre elas de pacientes HIV+ e HIV-.

Agradecimentos

Ao Hospital Universitário Clementino Fraga Filho (IDT-HUCFF-UFRJ) pela disponibilização da base de dados utilizada neste trabalho e pela cooperação técnica. A CAPES, ao CNPq, a FAPERJ, bem como ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) do CEFET/RJ por financiar esta pesquisa.

Referências

[1] Azzaro-Pantel C, Domenech S, Gomez A, Pibouleau L. Teaching mono and multi-objective genetic algorithms in process systems engineering: an illustration with the MULTIGEN environment. In: Escape 18, Lyon, France. 2008 Jun. p. 1-4.

[2] Deb K. Multi-objective optimization using evolutionary algorithms. John Wiley & Sons; 2001.

[3] Jhankal NK, Adhyaru D. Comparative analysis of bacterial foraging optimization algorithm with simulated annealing. In: International Journal of Science and Research (IJSR); Volume 3, Issue 3; 2014 Mar.

[4] Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. 2ª ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 2003

[5] Kennedy J, Eberhart RC, Shi Y. Swarm Intelligence (The Morgan Kaufmann Series in Artificial Intelligence). Elsevier; 2001.

[6] Almeida LF. Otimização de alternativas para desenvolvimento de campo de petróleo utilizando computação evolucionária [dissertação de mestrado]. DEE-PUC/RJ; 2003.

[7] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

[8] Madić M, Radovanović M, Nedić B. Modeling and simulated annealing optimization of surface roughness in CO2 laser nitroge cutting of stainless steel. In: Tribology in Industry; Vol. 35, No. 3; 2013. p. 167-176.

[9] Triola MF. Introdução à estatística, 10ª. ed. Rio de Janeiro: LTC Editora; 2008.

[10] Souza Filho JBO, Vieira APP, Seixas JM, Aguiar FS, Mello FCQ, Kritski AL. An intelligent system for managing the isolation of patients suspected of pulmonary tuberculosis. In: 13th International Conference on Intelligent Data Engineering and Automated Learning, 2012, Natal, v. 1. p. 1-8.