Anais

CACSI

2015

Congresso Amazônico de Computação e Sistemas Inteligentes

Manaus - Amazonas - Brasil

ISSN: 2447-0414

Realização

UEA
UNIVERSIDADE DO ESTADO DO AMAZONAS

EST
Escola Superior de Tecnologia da UEA

NUCOMP
Núcleo de Computação

Apoio

GOVERNO DO ESTADO DO AMAZONAS

FAPEAM
Fundação de Amparo à Pesquisa do Estado do Amazonas

SBC
Sociedade Brasileira de Computação

ASSUEA
Associação dos Servidores da Universidade do Estado do Amazonas

TambaCode

LUSTRABITS

Mark design

Comparação de técnicas de aprendizado de máquina com pré-processamento para decisão de profilaxia da tuberculose

Marcos Filipe Alves Salame[1], Flavia Matos Salame[2]

[1]Núcleo de Tecnologia da Informação – Embrapa Amazônia Ocidental
Caixa Postal 319 – 69010-970 – Manaus, AM – Brasil

[2]Escola Superior de Ciências da Saúde – Universidade do Estado do Amazonas (UEA) – Manaus, AM – Brasil

marcos.salame@embrapa.br, fsalame@uea.edu.br

Abstract. *Tuberculosis is a global public health problem. There are patients with latent infection which may evolve over time with disease. The Ministry of Health of Brazil identified the illness risk groups and assigned some variables that assist the medical staff in the decision to treat these individuals before they get sick. Based on this context it was made the simulation of 305 clinical cases, which were later submitted to processing five machine learning algorithms using the WEKA tool, getting better accuracy the J48 algorithm with the decision tree technique.*

Resumo. *A Tuberculose é um problema de saúde pública mundial. Existem portadores de infecção latente, que podem evoluir futuramente com a doença. O Ministério da Saúde do Brasil identificou os grupos de risco de adoecimento e atribuiu algumas variáveis que auxiliam a equipe médica na decisão de tratar esses indivíduos antes que adoeçam. Baseando-se nesse contexto foi realizada a simulação de 305 casos clínicos, que posteriormente foram submetidos ao processamento de cinco algoritmos de aprendizado de máquina utilizando a ferramenta WEKA, obtendo melhor acurácia o algoritmo J48 com a técnica de árvore de decisão.*

1. Introdução

A tuberculose (TB) é uma das doenças transmissíveis com maior mortalidade no mundo. Estima-se que em 2013 cerca de 9 milhões de pessoas desenvolveram TB e 1,5 milhão morreram da doença (Zumla et al. 2015). O Brasil encontra-se entre os 22 países com maior prevalência de TB e registra cerca de 73 mil casos novos por ano, sendo o estado do Amazonas o de maior prevalência em 2014 (Ministério da Saúde 2015).

A Organização Mundial da Saúde (OMS) publicou um Plano Global denominado *STOP TB Partnership* para estabelecer novas atividades de controle de TB para o período de 2006 a 2015 (WHO 2011). A proposta objetiva zero mortes por TB e a eliminação da TB como problema de saúde pública. Um dos principais pilares dessa estratégia é a prevenção de casos novos através do tratamento da infecção latente por tuberculose (ILTB) (Minnery et al. 2013; Raviglione 2012).

A ILTB é a forma latente da infecção, ou seja, quando há a presença do agente (*Mycobacterium tuberculosis-M.tb*) sem que haja doença em atividade. Estima-se que cerca de um terço da população mundial tenha ILTB. A grande maioria das pessoas infectadas não apresenta sinais ou sintomas da doença e não são contagiosas, mas estão em risco de desenvolver TB ativa no futuro. O diagnóstico da infecção latente por tuberculose é realizado pela realização da Prova Tuberculínica (PT), um teste cutâneo simples e de baixo custo (WHO 2015).

A OMS vem enfatizando a importância da investigação de casos de ILTB (Hopewell et al. 2012). O tratamento reduz a ocorrência de doença em 70 a 90% dos que receberam tratamento profilático (Fox et al. 2013; Morrison, Pai, Hopewell 2008).

Existem critérios para indicar o tratamento da ILTB, de acordo com os resultados da PT, em associação com características individuais de risco maior para progressão para doença clínica ativa (Figura 1) (Ministério da Saúde 2011). Pela diversidade de critérios e variáveis envolvidas na indicação correta da profilaxia, a elaboração de ferramentas que facilitem esse processo pode contribuir para que as equipes de saúde na atenção básica melhorem o processo de detecção de casos de ILTB e indiquem corretamente o tratamento.

Risco	PT ≥ 5mm	PT ≥ 10mm	Conversão*
Maior (indicado tratamento em qualquer idade)	HIV/aids**	Silicose	Contatos de TB bacilífera
	Contatos adultos*** e contatos menores de dez anos não vacinados com BCG ou vacinados há mais de dois anos****	Contato com menos de 10 anos vacinados com BCG há menos de dois anos	Profissional de saúde
	Uso de inibidores do TNF-a	Neoplasia de cabeça e pescoço	Profissional de laboratório de micobactéria
	Alterações radiológicas fibróticas sugestivas de sequela de TB	Insuficiência renal em diálise	Trabalhador de sistema prisional
	Transplantados em terapia imunossupressora		Trabalhadores de instituições de longa permanência
Moderado (indicado tratamento em < 65 anos)	Uso de corticosteróides (> 15mg de prednisona por >1 mês)*	Diabetes *mellitus*	
MENOR***** (indicado tratamento em < 50 anos)		Baixo peso (< 85% do peso ideal)	
		Tabagistas (≥ 1 maço/dia)	
		Calcificação isolada (sem fibrose) na radiografia	

Figura 1. Indicações de tratamento ILTB de acordo com a idade, resultado da PT e risco de adoecimento (Ministério da Saúde, 2011**).**

 ISSN: 2447-0414

Nas últimas décadas, métodos computacionais tem sido utilizados para auxiliar diversas tarefas na área da saúde (Zhao et al. 2015), pois permitem a geração, coleta e armazenamento de uma grande quantidade de dados (Marins et al. 2012). O uso de algoritmos de aprendizado de máquina em áreas da saúde é comum, no entanto, são inúmeros os desafios a superar para obter resultados consideráveis. Um requisito importante para algoritmos de aprendizado de máquina é que eles sejam capazes de lidar com dados imperfeitos: presença de ruídos, dados inconsistentes, dados ausentes e dados redundantes. Entretanto dependendo de sua extensão, esses problemas podem prejudicar o processo indutivo, tornando-se necessária a aplicação de técnicas de pré-processamento para minimizar a ocorrência desses problemas (Facelli 2011).

O trabalho foi dividido em 5 sessões. Os materiais e métodos utilizados constam na seção 2, enquanto os experimentos e resultados obtidos são apresentados na seção 3. A discussão, baseada nos resultados observados é mostrada na seção 4, finalizando na seção 5 com as considerações finais e sugestões de trabalhos futuros.

2. Material e Métodos

A proposta para este trabalho foi classificar se o paciente deve ou não realizar o tratamento profilático para tuberculose. Os dados foram simulados por uma médica especializada em Pneumologia e Tisiologia, coautora desse artigo, objetivando aproximação máxima com dados reais.

O conjunto de dados continha 305 casos clínicos simulados. Cada caso clínico consistia nas respostas obtidas a partir das variáveis mostradas na Tabela 1.

Tabela 1. Variáveis utilizadas para realizar a simulação de casos clínicos

É portador de Silicose?	(Silicose)
É infectado pelo HIV(Virus da Imunodeficiência Humana)?	(HIV)
É Portador de Neoplasia de Cabeça e Pescoço?	(CA_cabeçapescoço)
Está em uso de inibidores do TNF-α (Fator de Necrose Tumoral)?	(Anti_TNF)
É Portador de Insuficência renal em diálise?	(IRC_dialitica)
Possui alterações radiológicas fibróticas sugestivas de sequela de tuberculose?	(raiox_sequelaTB)
É paciente transplantado em terapia imunossupressora?	(tx_imunossupressor)
Está em uso de corticosteróide em quantidade maior que 15 mg por período maior que 1 mês?	(corticoide)
É portador de diabetes mellitus?	(diabetes)
Possui peso inferior a 85% do ideal?	(baixopeso)
É tabagista com carga tabágica superior a 1 maço/dia?	(tabagista)
Possui calcificação isolada sem fibrose na radiografia de tórax?	(calcificacaoIsolada)
É Contato de doente e tem menos de 10 anos sem vacina ou com vacina há mais de 2 anos?	(menor10anosSemVacina);
É contato de doente e tem menos de 10 anos e foi vacinado há menos de 2 anos?	(menor10anosComVacina),
Qual a idade?	(idade)
Qual o resultado da Prova tuberculínica?	(PT)
Indicado o tratamento profilático?	(profilaxia)

 ISSN: 2447-0414

Pela existência de muitos dados ausentes no mundo real, a simulação teve 20% das instâncias criadas com dados ausentes, escolhidos de forma aleatória, de forma a aproximar-se de um ambiente não perfeito. Por ser um número considerável, que prejudicaria a predição, foi utilizada uma técnica de pré-processamento de dados do grupo de tarefas de limpeza de dados, a qual mostrou-se eficiente, visto que em comparação com a análise com dados completos não evidenciou diferença significativa entre os resultados.

2.2. Ambiente WEKA

Todas as simulações foram realizadas no WEKA (*Weikato Environment for Knowledge Analysis*), uma plataforma de aprendizado de máquina que fornece uma coleção de esquemas de aprendizado implementados, que pode ser utilizado para mineração de dados e trabalhos de aprendizado de máquina (Witten, Frank, Hall 2011).

Para realizar a classificação, foram selecionadas cinco técnicas de aprendizado de máquina: árvore de decisão, rede neural artificial, máquina de suporte de vetor, Bayes e do vizinho mais próximo.

Para avaliar o desempenho das técnicas, a amostra inteira foi aleatoriamente dividida em conjunto para treinamento e conjunto para teste. A amostra selecionada para treinamento foi de 70% e a amostra para teste foi de 30%.

2.3. Pré-processamento de limpeza de dados

O problema de base de dados contendo valores ausentes é bastante comum em ambientes médicos, gerando dificuldades nos processos de análise e tomada de decisão. O processo de tomada de decisão é altamente dependente desses dados, exigindo métodos de estimativa precisos e eficientes (Jayalskshmi e Santhakumaran 2010).

A abordagem de limpeza de dados foi escolhida com a técnica de substituição dos valores ausentes pela média. No WEKA, o filtro *ReplaceMissingvalues* substitui os valores ausentes, tanto numéricos quanto nominais, por um conjunto de dados com as modas e médias dos dados do treinamento.

2.4. Técnicas de classificação

2.4.1 Árvores de decisão

Árvore de decisão é um método baseado em procura, normalmente usado para problemas de classificação. Nesta técnica, o conjunto de dados é aprendido e modelado, portanto, sempre que um novo item de dados é avaliado, será classificado adequadamente. Um ponto positivo em árvores de decisão, é que ele funciona bem com conjuntos de dados grandes (Barros et al. 2012). Para os propósitos deste estudo, foi escolhido o algoritmo J48.

2.4.2. Rede Neural Artificial

Rede neural artificial é um modelo matemático ou computacional baseado em otimização que simula o aspecto da estrutura funcional de uma rede neural biológica. Ela processa informação através de uma abordagem conexionista e consiste de neurônios artificiais. A técnica das redes neurais segue as mesmas teorias de como o

 ISSN: 2447-0414

cérebro humano funciona. No cérebro humano, há uma grande coleção de neurônios interconectados que conectam os nervos sensoriais e motores (Ahmadi et al. 2013).

Para categorizar uma decisão prática foi utilizado o algoritmo *back propagation* da rede *Multi-Layer Perceptron* (MLP).

2.4.3. Bayes

É um método baseado em probabilidade (Ronquist et al. 2012). Para a classificação foi escolhido o algoritmo NaiveBayes.

2.4.4. Máquinas de Vetores de Suporte

Neste método, um conjunto de exemplos de treinamento é fornecido com cada exemplo marcado pertencendo a uma das duas categorias. Em seguida, através do uso do algoritmo, um modelo que pode predizer se um novo exemplo cai em uma das categorias ou outro é construído (Wang et al. 2014).

Máquina de vetor de suporte é um método de aprendizado supervisionado usado para classificação baseado em otimização. Este método realiza a classificação através da construção de um hiperplano n-dimensional que otimamente separa os dados em duas categorias (Orrù et al. 2012). Para a classificação foi selecionado o algoritmo SMO.

2.4.5. Vizinho mais Próximo

É um método baseado em distância que considera a proximidade entre os dados na realização de predições. A hipótese base é que os dados similares tendem a estar concentrados em uma mesma região no espaço de entrada (Facelli 2011).

Foi selecionado o algoritmo IBk para realizar a classificação dos dados.

3. Experimentos e Resultados

Os resultados da simulação são mostrados nas tabelas 2 e 3, com o intuito de comparar os resultados obtidos de cada algoritmo e identificar a aplicabilidade dos mesmos na área da saúde, mais especificamente na decisão de profilaxia da tuberculose.

A Tabela 2 apresenta principalmente a acurácia de cada algoritmo de aprendizado de máquina, juntamente com o tempo e estatística Kappa, que mede a concordância além do que seria esperado somente pelo acaso. Quanto mais próximo do 1, maior o grau de concordância. A Tabela 3 apresenta resultados baseados nas diferentes taxas de erro.

Tabela 2. Resultado da classificação para cada algoritmo

Algoritmos WEKA	Corretamente classificado	Incorretamente classificado	Tempo (seg)	Estatística Kappa
J48	96.7033% (88)	3.2967% (3)	0.01	0.9333
MLP	68.1319% (62)	31.8681% (29)	1.25	0.3676
SMO	74.7253% (68)	25.2747% (23)	0.04	0.4721
IBk	70.3297% (64)	29.6703% (27)	0	0.4112
NaiveBayes	68.1319% (62)	31.8681% (29)	0.01	0.3211

 ISSN: 2447-0414

Tabela 3. Erro de treinamento e simulação

Algoritmos WEKA	Erro absoluto médio	Erro quadrático da raiz média	Erro absoluto relativo (%)	Erro quadrático da raiz relativa (%)
J48	0.0486	0.1774	9.9743	35.8322
MLP	0.3092	0.5038	63.4272	101.7522
SMO	0.2527	0.5027	51.8526	101.5333
IBk	0.2986	0.5422	61.2567	109.502
NaiveBayes	0.4206	0.4575	86.2959	92.4062

A tomada de decisão para este problema envolve a combinação de muitas variáveis e na Figura 2 é apresentada a árvore de decisão criada automaticamente pelo algoritmo J48.

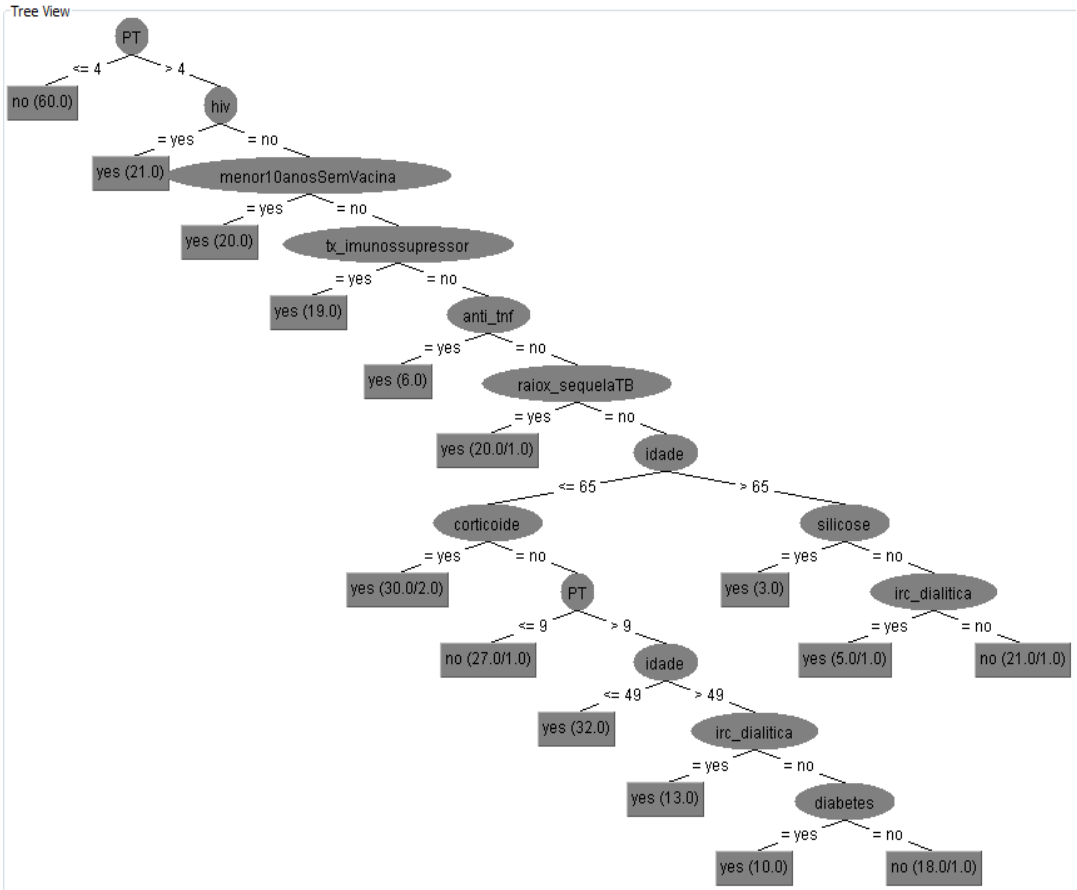

Figura 2. Árvore de decisão gerada pelo algoritmo J48

4. Discussão

Com base nos resultados obtidos, um número de conclusões úteis podem ser produzidos, com relação ao desempenho e as taxas de erros dos algoritmos escolhidos.

Conforme apresentado na Tabela 2, o algoritmo J48 obteve acurácia muito superior (96%) com um valor Kappa de 0.9333. A pior acurácia foi do MLP juntamente com o NaiveBayes, que obtiveram a mesma pontuação (68%), a diferença ficou no tempo, sendo o NaiveBayes mais rápido (0.01s). O algoritmo mais rápido foi o IBk (0s),

 ISSN: 2447-0414

que já era esperado devido a ser um algoritmo cujo processo de aprendizado consiste apenas em memorizar os objetos

Além disso, com base na Tabela 3, relativa aos erros de classificação, pode-se observar que o algoritmo J48 obteve o menor erro absoluto médio, menor erro absoluto relativo e erro quadrático da raiz relativa. Os valores de erros de NaiveBayes ficaram bem próximos com os do MLP, que foram os algoritmos que obtiveram somente 62 instâncias corretamente classificadas.

5. Considerações Finais

Ao longo deste trabalho foi apresentada uma comparação de cinco algoritmos de aprendizado de máquina com o objetivo de classificação de dados para tomada de decisão de profilaxia da tuberculose, que foram simulados por uma médica especialista em pneumologia e tisiologia. Resultados úteis foram obtidos com relação ao desempenho e às taxas de erro dos algoritmos.

Os experimentos realizados mostraram que o melhor algoritmo para classificar se determinado paciente deve ou não realizar profilaxia para prevenção de tuberculose é o J48, que utiliza árvore de decisão. Com acurácias bem inferiores vem o algoritmo SMO, seguido do IBk e por último MLP e NaiveBayes com a mesma pontuação.

O algoritmo J48 obteve ainda um bom desempenho no tempo e na pontuação Kappa, bem como taxas de erros baixas.

O trabalho corroborou no uso de inteligência computacional na área médica, podendo ser desenvolvido software para ajudar especialistas médicos. Uma sugestão de trabalho futuro seria coletar dados reais e experimentar aplicar técnicas de pré-processamento visando aumentar a acurácia do algoritmo J48 ou de outro algoritmo que utiliza árvores de decisão. Pode ser desenvolvido também um aplicativo para rodar em dispositivos móveis ou um software para computador pessoal de forma a auxiliar os médicos pneumologistas rapidamente no diagnóstico.

7. Referências

Ahmadi, Mohammad Ali, Mohammad Ebadi, Amin Shokrollahi, Seyed Mohammad Javad Majidi. (2013). “Evolving artificial neural network and imperialist competitive algorithm for prediction oil flow rate of the reservoir.” *Applied Soft Computing* 13(2): 1085–98.

Barros, Rodrigo Coelho, Márcio Porto Basgalupp, André C. P. L. F. de Carvalho, Alex A. Freitas. (2012). “A Survey of Evolutionary Algorithms for Decision-Tree Induction.” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(3): 291–312.

Facelli, Katti. (2011). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. 1ª ed Rio de Janeiro: Editira LTC.

Fox, Gregory J, Simone E Barry, Warwick J Britton, Guy B Marks. (2013). “Contact Investigation for Tuberculosis: A Systematic Review and Meta-Analysis.” *The European respiratory journal* 41(1): 140–56.

Hopewell, Philip C, Elizabeth Fair, Cecil Miller, World Health Organization. (2012). *Recommendations for Investigating Contacts of Persons with Infectious Tuberculosis*

in Low- and Middle-Income Countries. Geneva, Switzerland: World Health Organization.

Jayalskshmi, T., A. Santhakumaran. (2010). "Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks." In *2010 Second International Conference on Machine Learning and Computing*, IEEE, 109–12.

Marins, Oudival Luiz Fraccaro et al. (2012). "Aplicação de algoritmos de aprendizagem de máquina para mineração de dados sobre beneficiários de planos de saúde suplementar." *Journal of Health Informatics* 4(2).

Ministério da Saúde. (2011). *Manual de recomendações para o controle da tuberculose no Brasil*. Brasília, Distrito Federal, Brazil.

Ministério da Saúde (2015). 46, nº 9 *Boletim Epidemiológico: Detectar, tratar e curar: desafios e estratégias brasileiras frente à tuberculose*.

Minnery, Mark et al. (2013). "A cross sectional study of knowledge and attitudes towards tuberculosis amongst front-line tuberculosis personnel in high burden areas of Lima, Peru." *PloS one* 8(9): e75698.

Morrison, Janina, Madhukar Pai, Philip C Hopewell. (2008). "Tuberculosis and Latent Tuberculosis Infection in Close Contacts of People with Pulmonary Tuberculosis in Low-Income and Middle-Income Countries: A Systematic Review and Meta-Analysis." *The Lancet infectious diseases* 8(6): 359–68.

Orrù, Graziella et al. (2012). "Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review." *Neuroscience & Biobehavioral Reviews* 36(4): 1140–52.

Raviglione, Mario. (2012). *Developing the post-2015 TB Strategy and Targets : Vision and Process*. Kuala Lumpu, Malaysia.

Ronquist, Fredrik et al. (2012). "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space." *Systematic biology* 61(3): 539–42.

Wang, Kai et al. (2014). "Prediction of piRNAs using transposon interaction and a support vector machine." *BMC bioinformatics* 15(1): 419.

WHO. (2011). *The Global Blan to stop TB 2011-2015: Transforming the fight towards elimination of tuberculosis*.

WHO. (2015). *WHO | Guidelines on the management of latent tuberculosis infection*. 1ª ed World Health Organization.

Witten, Ian H., Eibe Frank, Mark A. Hall. (2011). Data Mining: Practical Machine Learning Tools and Techniques *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.

Zhao, Changbo, Guo-Zheng Li, Chengjun Wang, Jinling Niu. (2015). "Advances in Patient Classification for Traditional Chinese Medicine: A Machine Learning Perspective." *Evidence-based complementary and alternative medicine : eCAM* 2015: 376716.

Zumla, Alimuddin et al. (2015). "The WHO 2014 Global tuberculosis report — further to go." *The Lancet Global Health* 3(1): e10–12.