Applied Mathematical Sciences, Vol. 6, 2012, no. 50, 2481 – 2488

Modelling of Tuberculosis Patients Using Multilayer Feed forward Neural Network

Nor Azlida Aleng[1], Norizan Mohamed and Wan Muhamad Amir W Ahmad

Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia
Universiti Malaysia Terengganu (UMT)
21030 Kuala Terengganu, Terengganu Malaysia
[1] e-mail: azlida_aleng@umt.edu.my

Abstract

Tuberculosis (TB) is a highly infectious bacterial disease that has spread from the lungs to other parts of the body through the blood or lymph system. TB most commonly affects the lungs but also can involve almost any organ of the body. Prevention and treatment efforts have dramatically reduced its spread, but about one-third of the world's population is still infected with tuberculosis. The objective of the current study is to develop a multilayer feed-forward (MLFF) neural network model of tuberculosis by using the combinations of significant variables from multiple linear regression (MLR) model. The MLR of tuberculosis showed that gender, married status, diseases genital and descendant were significant. These four variables were used to develop the best (MLFF) neural network model of tuberculosis.

Keywords: Multilayer Feed-Forward, Multiple Linear Regression and Tuberculosis

INTRODUCTION

Tuberculosis or *tubercles bacillus* is one of the most critical illnesses in Malaysia. The burden of tuberculosis (TB) today is greatest in low-income countries. Over 90% of all cases arise and over 95% of deaths from the disease occur there. Moreover, a high and increasing proportion of cases in many industrialized countries occur in people who were born and became infected in low income countries, before they moved to the country where their TB has been detected [1].

TB is a common and often deadly infectious disease which is caused by various strains of *Mycobacterium tuberculosis* in humans. This bacterium usually attacks lungs, heart and other parts of the body which is known as miliary TB. Testing for miliary TB is conducted in the same manner as for other forms of tuberculosis. TB is spread through the air from one person to another (spread through the air with coughing or sneezing). Tuberculosis (TB) and human immunodeficiency virus (HIV) infections are the deadliest chronic infections globally. Although each is deadly alone, they are deadlier together, with TB causing one-quarter of AIDS-related deaths and HIV infecting at least 15% of patients with TB worldwide [2].

Approximately 7 million new cases of TB and 1.7 million deaths due to TB were reported in 2006 [2]. The HIV epidemic has fuelled the current TB epidemic worldwide and in particular in sub-Saharan Africa. HIV is the strongest factor in the development of active TB; it is estimated that only one out of ten immunocompetent persons infected with TB develops active TB in his/her lifetime; whereas, one out of ten HIV-infected persons infected with TB will develop active TB every year. Autopsy studies have shown that 30 to 40.0% of HIV-infected adults die from tuberculosis in Africa. On the other hand TB has been shown to accelerate HIV disease progression to AIDS and probably early death [3].

This article is an extension of previous study as reported in Wan Muhamad Amir W. Ahmad et al. [4]. In that article, they were used the path analysis to model the associated factors of TB among HIV positive patients. In this current study, we develop a multilayer feed-forward (MLFF) neural network model with the best combination of significant variables from multiple linear regression (MLR) model.

MATERIAL AND METHODS

Study Population

A total of 284 Miliary TB patients were studied for the present of Tuberculosis of other organs. We used MLR method and MLFF Neural Network to develop the best model which considered the significance variable from the MLR method. Dataset consists of 284 Miliary TB patients and they were collected directly from Medical Unit Records Department. According to the Table 1, material of this study is a hypothetical sample which is composed of twelve variables.

Table 1: Explanation of the Variables

Code	Variables	Explanation of the variables
Y	Age	Age of Patients
X1	Gender	Patient's Gender
X2	Married_Status	Married Status (0 = No and 1 = Yes)
X3	Tuber_Organ	Tuberculosis of other organs (0 = No and 1 = Yes)
X4	Infect	Infection with parasite diseases (0 = No and 1 = Yes)
X5	Endo	Endocrine, Nutritional And Metabolic Diseases (0 = No and 1 = Yes)
X6	Nervous	Diseases of The Nervous System (0 = No and 1 = Yes)
X7	Hypertensive	Hypertensive Disease (0 = No and 1 = Yes)
X8	Respiratory Infections	Influenza, Pneumonia and other Acute Lower Respiratory Infections (0 = No and 1 = Yes)
X9	Oesophagus	Diseases of Esophagus, Stomach and Duodenums, Stomach and Duodenum (0 = No and 1 = Yes)
X10	Dorsopathies	Spondylopathies (0 = No and 1 = Yes)
X11	Diseases_Genital	Having Diseases of The Genito-Urinary System (0 = No and 1 = yes)
X12	Decendent	Descendant (1=Malay, 2=Chinese, 3= Indian & 4=Others)

Sample size calculation for the current study as follows:

Anticipated population proportion (p) = 0.89

Level of significance = 5% (0.05)

Absolute precision (Δ) = $\pm$ 5%

$$= (1.96/0.05)^2\, 0.89\,(1 - 0.89)$$

=150 respondents.

According to the calculation of sample size, the minimum requirement sample size need is 150 patients. In this study, we already collect data from 284 patients. Hence it shows that the sample size of the current study is adequate.

Methods

Multiple linear regression (MLR)

MLR is a method used to model the linear relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the relationship between the explanatory and response variables. In this linear regression model, every value of the independent variable x is associated with a value of the dependent variable y [5].

The multiple linear regression model assumes a linear (in parameters) relationship between a dependent variable y_i and a set of explanatory variables $x_i' = (x_{i0}, x_{i1}, ..., x_{iK})$. x_{ik}'s are also called independent variables, covariates or regressors. The first regressor $x_{i0} = 1$ is a constant unless otherwise specified. The model for MLR, given *n* observations is follow:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \tag{1}$$

where β is a (K + 1)- dimensional column vector of parameters x_i' is a (K + 1) - dimensional row vector and ε_i is a scalar called the error term.The whole sample of *N* observations can be expressed in the following matrix notation,

$$y = X\beta + \varepsilon \tag{2}$$

where y is an *N*-dimensional column vector, X is an $N \times (K + 1)$ matrix and ε is an *N*-dimensional column vector of error terms, i.e.

$$\underset{N\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix}} = \underset{N\times(K+1)}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ 1 & x_{31} & \cdots & x_{3K} \\ 1 & \vdots & \cdots & \vdots \\ 1 & x_{N1} & \cdots & x_{NK} \end{bmatrix}} \underset{(K+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}} + \underset{N\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}}$$

Multilayer Feed-forward Neural Network (MLFF)

A well known neural model, which consists of an input layer, one or several hidden layers and an output layer. The neurons in the feed-forward neural network are generally grouped into layers. Signals flow in one direction from the input layer to the next, but not within the same layer [6]. An essential factor of successes of the neural networks depends on the training network. Among the several learning algorithms available, back-propagation has been the most popular and most widely implemented [7]. Basically, the BP training algorithm with three-layer feed-forward architecture means that, the network has an input layer, one hidden layer and an output layer. In this research the output node is fixed at one since there is only one independent variable. Thus, for the feed-forward network with *N* input nodes, *H* hidden nodes and one output node, the values $\hat{Y}$ are given by:

$$\hat{Y} = g_2\left(\sum_{j=1}^{H} w_j h_j + w_0\right) \tag{3}$$

where w_j is an output *weight* from hidden node j to output node, w_0 is the bias for output node, and g_1 is an activation function. The values of the hidden nodes h_j, $j = 1,\ldots,H$ are given by:

$$h_j = g_1\left(\sum_{i=1}^{N} v_{ji} X_i + v_{j0}\right),\ j = 1,\ldots,H \tag{4}$$

Here, v_{ji} is the input *weight* from input node i to hidden node j, v_{j0} is the bias for hidden node j, X_i are the independent variables where $i = 1,\ldots,N$ and g_1 is an activation function.

RESULTS AND DISCUSSION

MLR Analysis

In this analysis, we use MLR approach to determine which factors are strongly associated with Miliary TB diseases. The results are presented in Table 2. These variables had a direct relationship with age.

Table 2: Estimates of Parameters of MLR Model

Model	Unstandardized Coefficients Beta (β)	Std. Error	*t*	Sig.
(Constant)	27.300	2.618	10.427	0.000
Gender	4.614	1.881	2.452	0.015
Married Status	22.744	1.806	12.593	0.000
Diseases Genital	21.586	10.652	2.027	0.044
Descendant	-2.226	0.637	-3.493	0.001

Dependent Variable: AGE

As illustrated in Table 2, there are exist direct positive and negative relationships between Age and gender, married status, diseases genital and descendent respectively. Results in Table 2 also indicate–that only four variables are significant and were include in the regression equation model; Gender (β = -0.305 $p < 0.015$), Married Status (β = 22.744, $p < 0.000$), Diseases Genital (= 21.586, $p < 0.044$) and Descendant (β = -2.226, $p < 0.000$).

The regression equation model with four independent variables can be expressed as follows:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_{11} + \beta_{12} x_{12} + \varepsilon_i \\ \hat{y}_i &= 27.3 + 4.614 x_1 + 22.744 x_2 + 21.586 x_{11} - 2.2263 x_{12} + \varepsilon_i \end{aligned} \tag{5}$$

where x_1, x_2, x_{11} and x_{12} are gender, married status, diseases genital and descendant respectively.

MLFF Analysis

Previous study, Norizan et al. [8] used the significant variables from MLR model as input nodes of MLFF neural network model. In this current study, the input variables of MLFF model are also selected based on the significant variables from MLR model. To select the appropriate number of hidden nodes, we apply forward procedure as proposed by Norizan [9]. The output node in this study is one node since we have one dependent variable which is age of tuberculosis patients.

The best MLR of tuberculosis showed that four significant variables which are X_1 (gender), X_2 (married status), X_3 (diseases genital) and X_4 (descendant) respectively. By following the model selection strategies which proposed by Norizan [9], we found that the best number of hidden nodes is three nodes. Using difference combination of significant variables, result shown that the best combination is four input variables as illustrated in Table 3. Hence, the architecture of the best multilayer feed-forward neural network model with one hidden layer, four input variables, 3 hidden nodes and one output node is presented in Figure 1. It can be represented as follows:

$$\hat{Y} = g_2\left(\sum_{j=1}^{3} w_j h_j + w_0\right) \tag{6}$$

where w_j is an output *weight* from hidden node j to output node, w_0 is the bias for output node, and g_1 is an activation function. The values of the hidden nodes h_j, $j = 1, 2, 3$ are given by:

$$h_j = g_1\left(\sum_{i=1}^{4} v_{ji} X_i + v_{j0}\right),\ j = 1, 2, 3 \tag{7}$$

Here, v_{ji} is the input *weight* from input node i to hidden node j, v_{j0} is the bias for hidden node j, X_i are the independent variables where $i = 1,\ldots,4$ and g_1 is an activation function.

Table 3: The results of input variables, R^2 training. R^2 testing, MSE training and MSE testing.

Input Variables	R^2 Training	R^2 Testing	MSE Training	MSE Testing
X_1, X_2, X_3, X_4	0.8694	0.9222	224.6022	171.8232
X_1, X_3, X_4	0.8732	0.9198	218.0624	182.9936
X_1, X_4	0.8672	0.9196	228.3181	183.2568
X_1	0.8563	0.9129	247.224	191.1831

X_1, X_2, X_3 and X_4 are gender, married status, diseases genital and descendant respectively

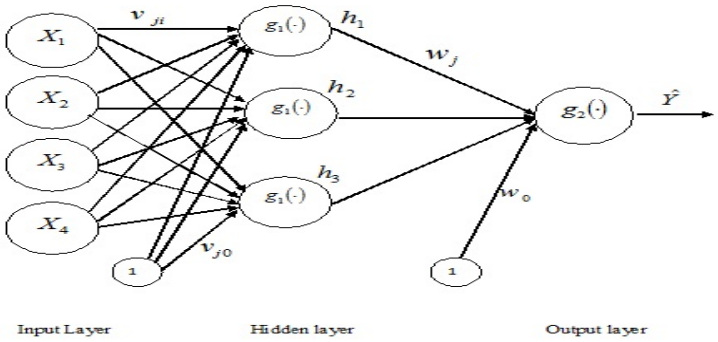

Figure 1: The architecture of the best multilayer feed-forward neural network model with one hidden layer, four input variables, three hidden nodes and one output node.

CONCLUSION

The purpose of the current study is to develop a MLFF neural network model of tuberculosis by using the combinations of significant variables from multiple linear regression (MLR) model. The MLR of tuberculosis showed that gender, married status, diseases genital and descendant were significant. Using the significant variables the performance of MLFF neural network model for difference combinations are tested. The performance of MLFF was evaluated using R^2 and MSE of testing/out-sample. The combination of four input variables outperformance other combinations. Hence, four variables were used to develop the best (MLFF) neural network model of tuberculosis.

REFERENCES

[1] P.J. Dolin, M.C. Raviglione and A. Kochi, Global tuberculosis incidence and mortality. Bull World Health Organ, 1994.

[2] W.D. Dupont and W.D. Plummer, Power and sample size calculation: a review and computer program, Controlled Clinical Trails, 1990.

[3] F.M. Mugusi, S. Mehta, E. Villamor, W. Urassa, E. Saathoff, R.J. Bosch and W.W. Fawzi, Factors associated with mortality in HIV-infected and uninfected patients with pulmonary Tuberculosis. BMC Public Health, 2009.

[4] Wan Muhamad Amir W. Ahmad, Nor Azlida Aleng, Zalila Ali and Arif Bin Awang. Modelling, Associated factors of HIV-infected Tuberculosis (TB) patients using path model analysis, World Applied Sciences Journal, 2011.

[5] N.R. Draper and H. Smith, Applied regression analysis, New York: John Wiley & Son, 1981.

[6] T.D. Pham and X. Liu, Neural networks for identification, prediction and control, Great Britain. 1995.

[7] G.A. Darbellay and M. Slama, Forecasting the short-term demand for electricity. Do neural networks stand a better chance?, International Journal of Forecasting, 2000.

[8] Norizan Mohamed, Wan Muhamad Amir W. Ahmad, Nor Azlida Aleng and Maizah Hura Ahmad, Assessing the efficiency of multilayer feed-forward neural network model: application to body mass index data, World Application. Sci. Journal, 2011.

[9] Norizan Mohamed, Parametric and artificial intelligence based methods for forecasting short term electricity load demand. Unpublished PhD thesis, 2011.

Received: December, 2011