World Applied Sciences Journal 15 (5): 677-682, 2011
ISSN 1818-4952

Assessing the Efficiency of Multilayer Feed-Forward Neural Network Model: Application to Body Mass Index Data

[1]Norizan Mohamed, [1]Wan Muhamad Amir W. Ahmad,
[1]Nor Azlida Aleng and [2]Maizah Hura Ahmad

[1]Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia,
Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu Malaysia
[4]Jabatan Matematik, Fakulti Sains, Universiti Teknologi Malaysia, 81030 UTM Skudai, Johor, Malaysia

Abstract: A variety of statistical approaches have been used to find the directed dependencies among a set of interest variables and to identify the associated important factors. Among the most popular methods are proportional hazard regression and logistic regression. The aim of the current study is to suggest another approach by using a multilayer feed-forward neural network model (MLFF). Using body mass index (BMI) as the dependent variable, we identify its related and appropriate independent variables. In this study we put forth two MLFF models. Model 1 is where all the independent variables as identified in the literature are included, while Model 2 is where only variables found significance as a result from a multiple linear regression (MLR) analysis are included in the model. Analyses were done by using SPSS and MATLAB packages. As a result of the study, we found that the best MLFF model was the model which considered the input variables based on selection criteria for regression.

Key words: Multiple linear regression and multilayer feed-forward neural network

INTRODUCTION

The body mass index (BMI) is a measurement that is calculated using a person's height and weight measurements. BMI is used by healthcare professionals all over the world as a reliable indicator to determine whether a person is overweight or clinically obese [1]. However, the BMI measurement can sometimes be misleading because for example, a muscleman with much less fat than an unfit person may have a higher BMI. Nevertheless, in general, the BMI measurement can be a useful screening tool to classify weight categories that can lead to serious health problems such as heart disease and diabetes [2]. Table 1 gives the body mass index categories.

BMI is usually related to body fat measurement. A person with high body fat will usually have a high BMI, while a high body fat percentage can put that person at risk for many serious diseases [1]. BMI of above 25 increases the risk of having high blood pressure, heart disease, stroke, diabetes, certain types of cancer, arthritis and breathing problems. Previous research shows that obesity reduces one's life expectancy [1, 2]. For overweight or obese people, losing their weight also means lowering their blood pressure, total cholesterol, LDL cholesterol, increasing their HDL cholesterol, improving their blood sugar levels and reducing their amount of abdominal fat [2].

According to the report from National Institutes of Health in 1998, the government has emphasized the importance of the treatment of overweight and obesity in adults. Previous research finding shows that, when an overweight patient loses weight his/her chronic diseases can be reduced and that patient's life span can be improved. Clinicians and the public should be informed of the findings and the importance of weight management [1]. In writing/preparing this report, more than 43,627 research articles were obtained from a search of the scientific literature and reviewed by a panel of researchers. Researchers have examined the importance

Table 1: Body mass index categories

BMI	Weight category
Less than 18.5	• Unhealthy.
18.5 – 24.9	• Healthy
25- 29.9	• Weight maintenance is appropriate or reduce weight
30 and above	• Weight loss is suggested

(*Source: McKinley Health Center, 2009)*

Corresponding Author: Norizan Mohamed, Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia.

Table 2: Explanation of the variables

Code	Variables	Explanation of the variables	Categorical
Y	BMI	Body Mass Index	
X1	SBP	Systolic Blood Pressure	
X2	DBP	Diastolic Blood Pressure	-
X3	AGE	Age In Years	-
X4	CHOLES	Serum Cholesterol (Mmol/L) Less than 200 mg/dL-Desirable 200-239 mg/dL -Borderline high 240 mg/dL and above -High	-
X5	GENDER	Gender	0 = Female 1= Male
X6	PULSE	Pulse beats per minute	-
X7	HEIGHT	Height of a patient in cm	-
X8	WEIGHT	Weight of a patient in kg	-
X9	ARM	The length of arm in cm	-
X10	LEG	The length of leg in cm	-
X11	WAIST	The length of waist in cm	-
X12	WRIST	The length of wrist in cm	-

of weight reduction in people with high blood cholesterol, high blood pressure, diabetes, cancer and osteoarthritis and reported that weight loss reduces the risks for these diseases [1, 2].

The objective of the current study is to show that by using the variable of the multiple linear regression model (MLR) more accurate forest is obtained. Hence we analyze the medical data by using two approaches. For the first approach, all variables are considered as input variables and for the second approach, the input variables are selected best on the best multiple linear regression model (MLR). Material of this study is a hypothetical sample which is composed of twelve variables. Namely variables are as in Table 2 and data were collected from Hospital in Kelantan.

Sample size calculation for the current study as follows:	
Anticipated population proportion (P)	= 0.96
Level of significance	= 5% (0.05)
Absolute precision (Δ)	= ±5%
	=$(1.96/0.05)^2$ 0.95 (1-0.95)
	=74 respondents.

The minimum sample size required at the stage is 74 respondents. For the current study, we used 4690 respondents. Hence it shows that the sample size of the current study is adequate.

MATERIAL AND METHODS

Details of Methodologies Are as Follows

Linear Regression: Attempts to model the relationship between independent and dependent variables by fitting a linear equation to observed data. It is a powerful statistical tool that allows us to account for (or predict) the variance in a continuous dependent variable based on a linear combination of multiple independent variables. In this linear regression model, every value of the independent variable x is associated with a value of the dependent variable y [3].

Multiple logistic regression is a technique for modeling and studying an association between several variables. It is observed that both of multiple linear regression and multiple logistic regression methods are frequently used in social sciences, medical sciences, education sciences and many more. Just like multiple linear regression cases, it's focused on the relationship between a dependent variable and one or more independent variables. However, to use this method it is depends on some circumstances [4].

Wan Muhamad Amir W. Ahmad *et al*. [5] used logistic regression method and multiple linear regression method to gain the path model for modeling associated factors of HIV-infected tuberculosis (TB) patients. Actually, path analysis technique is an idea of how we interpreting the extension results of the linear regression model. It goes beyond regression in that it allows for the analysis of more complicated models. In particular, it can check situations in which there are several final dependent variables and those in which there are "chains" of influence, in that variable X influences variable Y, which in turn affects variable Z and so on.

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. Consider a sample of N observations i = 1,..., N. Every single observation i follows.

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

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,

$$y = X\beta + \varepsilon \quad (2)$$

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

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

We used logistic regression method and multiple linear regression method.

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:

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

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,...,H are given by:

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

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,...,N and g_1 is an activation function. The architecture of the multilayer feed-forward neural network model is illustrated in Figure 1.

RESULTS AND DISCUSSION

Multiple Regression Analysis: Multiple regression analysis was used to analyze the direct relationship between all the predictor (independent) variables and body mass index (dependent variable). The results are presented in Table 3.

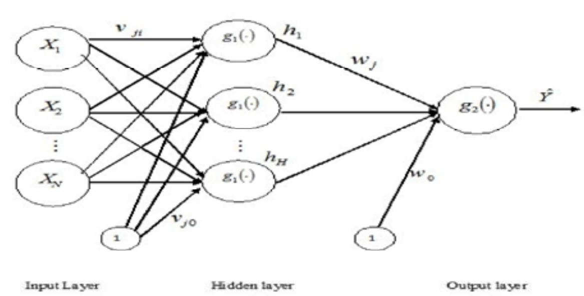

Fig. 1: The architecture of the multilayer feed-forward neural network model with one hidden layer, N input nodes, H hidden nodes and one output node.

Table 3: Regression Analysis of Pulse, Arm, Height, Weight with BMI

Dependent Variable	Independent Variable	Std. Coefficient Beta (β)	Sig.	R^2	Adjust R^2
Body Mass Index	(Constant)	48.408	0.000	0.99	0.995
	Pulse	0.011	0.022		
	Arm	0.063	0.004		
	Height	-0.305	0.000		
	Weight	0.329	0.000		

*Note: Significant levels: **p < 0.01, *p < 0.05*

As illustrated in Table 3, there exist direct positive and negative relationships between BMI and pulse, arm, height and weight respectively. Results in Table 3 also indicated that only four variables, which are pulse, arm, height and weight are significant and are to be included in the regression equation model. R^2 value is close to one (0.99) indicate that the current model has great ability to predict a future trend. The regression equation model with four independent variables can be expressed as follows:

$$y_i = \beta_0 + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \varepsilon_i$$
$$\hat{y}_i = 48.404 + 0.011 x_{i6} - 0.305 x_{i7} + 0.329 x_{i8} + 0.063 x_{i9} + \varepsilon_i \quad (5)$$

Where x_{i6}, x_{i7}, x_{i8} and x_{i9} are pulse, height, weight and arm respectively.

Results of MLFF: The neural network architecture is composed of the number of input, hidden and output nodes. The main focus of the current study is to compare between two approaches in the selection of input nodes. In the first approach we consider all the variables that have a relationship with body mass index as input variables. There are 12 selected variables, which are systolic blood pressure (SBP), diastolic blood pressure (DBP), age, serum cholesterol (CHOLES), gender, pulse, height, weight, arm, leg, waist and wrist. In the second approach we consider the independent variables from the best MLR model as input variables in the MLFF neural network model. There are four independent variables namely pulse, arm, height and weight. To select the appropriate number of hidden nodes, we apply forward procedure as proposed by Mohamed [8]. The output node in this study is one node since we have one dependent variable which is body mass index. The data was partitioned into two parts which are training and testing. 70 data are used for training and 10 data are used for testing.

The Levenberg-Marquardt back-propagation is used as the training algorithm since it was claimed by Mohamed [8] as the best training algorithm. The log-sigmoid function and the linear transfer function are applied in the hidden layer and output layer respectively. This combination is selected due to the justification put forth by Mohamed [8].

Model 1: Since we consider all independent variables as inputs for Model 1, then input nodes are 12 nodes and as body mass index is considered as the output, then the output node is one. We then apply the forward procedure to find the best number of hidden nodes. We found that the best number of hidden nodes for the Model 1 is three nodes. Hence, the appropriate neural network architecture which results in the best multilayer feed-forward neural network model for Model 1 is composed of 12 input nodes, three hidden nodes and one output node. It can be represented as follows:

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

Where w_j is an output *weight* from hidden node j to output node, w_0 is the bias for output node and g_2 is the linear function. h_j are the values of the hidden layer nodes which can be represented as:

$$h_j = g_1\left(\sum_{i=1}^{12} v_{ji} X_i + v_{j0}\right) \; j = 1,2,3. \quad (7)$$

Where v_{ji} is the input weight from input node i to hidden node j, v_{j0} is the bias for hidden node j and g_1 is an activation function. X_i are the independent variables where $X_1, X_2, \ldots, X_{12}$ are systolic blood pressure (SBP), diastolic blood pressure (DBP), age, serum cholesterol (CHOLES), gender, pulse, height, weight, arm, leg, waist and wrist respectively.

Equations 6 and 7 can also be represented as follows:

$$\hat{Y} = w_0 + w_1 h_1 + w_2 h_2 \quad (8)$$

$$h_j = \left[1 + \exp\left[-\left(v_{j0} + v_{j1}X_1 + v_{j2}X_2 + \ldots + v_{j13}X_{13}\right)\right]\right]^{-1} \; j = 1,2. \quad (9)$$

The architecture of model 1 is illustrated in Figure 2.

As a result of training this architecture using the Levenberg-Marquardt back-propagation training algorithm and using log-sigmoid transfer function in the hidden layer and linear function in the output layer, the MSE training and MSE testing are presented in Table 4.

Model 2: The input variables of Model 2 are selected based on the best MLR model. The best MLR showed that four independent variables which are pulse, arm, height and weight have strong relationship with body mass index. Hence the input nodes of Model 2 are four nodes. Applying the forward procedure to find the best number of hidden nodes, we found that the best

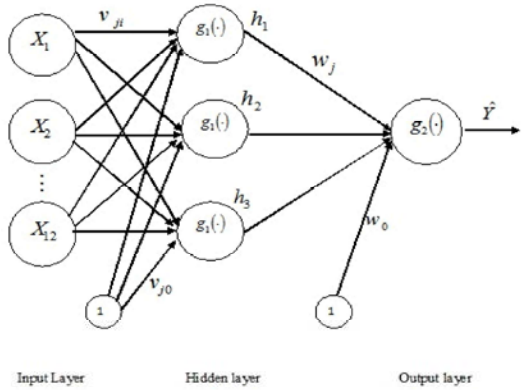

Fig. 2: The architecture of model 1 with one hidden layer, 12 input nodes, 3 hidden nodes, one output node, log-sigmoid transfer function in the hidden layer and linear transfer function in the output layer

Table 4: The MSE model 1

	MSE
Training/In-sample	0.0011
Testing/Out-sample	0.0059

Table 5: The MSE of model 2.

	MSE
Training/In-sample	0.0018
Testing/Out-sample	0.0017

number of hidden nodes for model 2 is also three nodes. Hence, the appropriate neural network architecture which results in the best multilayer feed-forward neural network model for Model 2 is composed of four input nodes, three hidden nodes and one output node. It can be represented as follows:

$$\hat{Y} = g_2\left(\sum_{j=1}^{3} w_j h_j + w_0\right) \qquad (10)$$

Where w_j is an output *weight* from hidden node j to output node, w_0 is the bias for output node and g_2 is the linear function. h_j are the values of the hidden layer nodes which can be represented as:

$$h_j = g_1\left(\sum_{i=1}^{4} v_{ji} X_i + v_{j0}\right) \quad j = 1,\ldots,3 \qquad (11)$$

Where v_{ji} is the input weight from input node i to hidden node j, v_{j0} is the bias for hidden node j and g_1 is an activation function. X_i are the independent variables where X_1, X_2, X_3 and X_4 are pulse, arm, height and weight respectively.

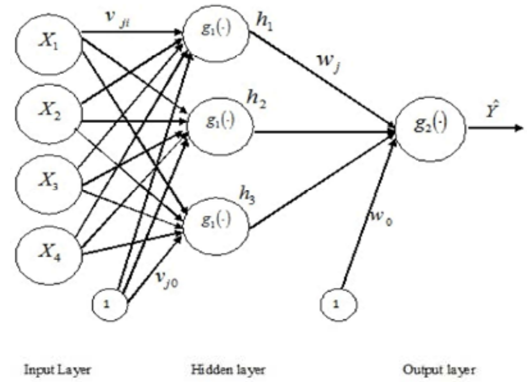

Fig. 3: The architecture model 2 with one hidden layer, 4 input nodes, 3 hidden nodes, one output node, log-sigmoid transfer function in the hidden layer and linear transfer function in the output layer

Equations 11 and 12 can also be represented as follows:

$$\hat{Y} = w_0 + w_1 h_1 + w_2 h_2 + w_3 h_3 \qquad (13)$$

$$h_j = \left[1 + \exp\left[-\begin{pmatrix} v_{j0} + v_{j1}X_1 + v_{j2}X_2 + \\ v_{j3}X_3 + v_{j4}X_4 \end{pmatrix}\right]\right]^{-1} \quad j = 1,\ldots,3 \qquad (14)$$

The architecture of model 2 is illustrated in Figure 3.

As a result of training this architecture using the Levenberg-Marquardt back-propagation training algorithm and using log-sigmoid transfer function in the hidden layer and linear function in the output layer, the MSE training and MSE testing are presented in Table 5.

CONCLUSION

The purpose of the current study is to develop a model that can be used to predict *y* (a dependent variable) based on some *x*'s (independent variables). For this purpose, we developed two MLFF neural network models. The data used were body mass index (BMI) as the dependent variable and 12 independent variables as suggested in the literature. Model 1 was developed based on all the independent variables. However, using multiple regression, we discovered only four independent variables were significant. Using only these four independent variables, we developed Model 2. The efficiency of MLFF was evaluated using MSE.

The MSE for testing/out-sample of Model 2 was lower than that of Model 1, implying that the architecture of Model 2 was more accurate than Model 1. We therefore can conclude that the best selection variables from MLR could be considered in selecting the input nodes of a MLFF model.

REFERENCES

1. Cosmetic Surgery Consultants. http://www.cosmeticsurgeryconsultants.co.uk/obesity-what-is-bmi.htm
2. Walker S.P., Rimm EB, Ascherio A, Kawachi I, Stampfer MJ, Willett WC (1996). Body size and fat distribution as predictors of stroke among US men. Am J. Epidemiol; 144; 1143-1150.
3. Draper, N.R. and H. Smith, 1981. Applied Regression Analysis. 2nd. New York: John WileyandSon.
4. Wan Muhamad Amir, W. Ahmad, Nor Azlida Aleng and Zalila Ali, 2010. Binary logistic regression analysis technique used in analyzing the categorical data in education science: A case Study of Terengganu state, Malaysia. World Appl. Sci. J., 9(9): 1062-1066.
5. Wan Muhamad Amir W. Ahmad, Nor Azlida Aleng, Zalila Ali and Arif Bin Awang Nawi, 2011. Modelling Associated Factors of HIV-Infected Tuberculosis (TB) Patients Using Path Model Analysis. World Appl. Sci. J., 12(9): 1580-1584.
6. Pham, T. D. and X. Liu, 1995. Neural Networks for Identification, Prediction and Control, Great Britain.
7. Darbellay G. A and M. Slama, 2000. Forecasting the short-term demand for electricity. Do neural networks stand a better chance? Int. J. Forecasting, 16: 71-83.
8. Mohamed, N., 2011. Parametric and artificial intelligence based methods for forecasting short term electricity load demand. Unpublished PhD thesis
9. McKinley Health Center http:// www.mckinley.illinois.edu/ handouts/ body_mass_index.htm