Multilayer Perceptron (MLP) Vs. Multiple Regression (MR) Model: The performance evaluation analysis using Breast Cancer Database

Zukime Mat Junoh[1]
School of Business Innovation and Technopreneurship,
Universiti Malaysia Perlis
Kangar, Malaysia
zukime@unimap.edu.my

Abdullah Osman[2]
School of Business Innovation and Technopreneurship
Universiti Malaysia Perlis
Kangar, Malaysia
abdullahosman@unimap.edu.my

Muhammad Safizal Abdullah[3]
School of Business Innovation and Technopreneurship
Universiti Malaysia Perlis (UniMAP)
Kangar, Malaysia
safizal@unimap.edu.my

Mohd Harith Amlus[4]
School of Business Innovation and Technopreneurship
Universiti Malaysia Perlis (UniMAP)
Kangar, Malaysia
harith@unimap.edu.my

Mohamad Sharif Ismail[5]
School of Business Innovation and Technopreneurship
Universiti Malaysia Perlis (UniMAP)
Kangar, Perlis
sheriff@unimap.edu.my

***Abstract*— Breast Cancer is the common cancer among the women in the world. The aim of this paper therefore, is not to inhabit on the prediction techniques themselves, but rather concentrate on comparison of results produced from these two alternatives (Neural Networks and Multiple Regression), yet complementary techniques. In order to gauge the success (or otherwise) of either techniques, a comparative analysis of prediction performance must be made [1]. The models use the conventional statistical technique multiple regression and artificial neural networks. Performance analyses using mean percentage error, mean absolute percentage error and percentage of correctness (generalisation). Results reveals that ANNs model perform well, having low mean absolute percentage error values indicating that predictor variables were reliable inputs for modelling breast cancer database. Overall, the neural network model performs slightly better as it was able to predict up to 97.14 % generalisation compare with multiple regression just only 82.10 % generalisation**

Keywords*— *Breast Cancer analysis, multilayer perceptron and multiple regression

I. INTRODUCTION

Breast Cancer is the common cancer among the women in the world. In a sense, breast cancer in easy to detect-just kill the patient, perform a thorough autopsy and likely to discover any existing cancer. Of course, this detection method defeats the purpose, but it serves to illustrate the fact that cancer detection is a spectrum from maximally invasive, expensive, and possibly less accurate methods. The aim of this paper therefore, is not to inhabit on the prediction techniques themselves, but rather concentrate on comparison of results produced from these two alternatives (Neural Networks and Multiple Regression), yet complementary techniques. In order to gauge the success (or otherwise) of either techniques, a comparative analysis of prediction performance must be made [1].

II. OBJECTIVE OF STUDY

A. This paper seeks to determine the most accurate and robust model for predicting Breast Cancer Database in Medicine Industry.

B. To state relevant predictor variables of breast cancer analysis in both multiple regression and neural network (Multilayer Perceptron) model.

C. To evaluate the forecast performance of both models, through the use diagnostic of R-square performance measures namely mean percentage error (MPE) and mean absolute percentage error (MAPE).

D. Maintaining the Integrity of the Specifications.

III. LITERATURE REVIEW

Initial research work, applied conventional statistical techniques, in particular multiple regression (MR) analysis [2]. The great advantage of regression is that it has an established record of use within the modeling of structural problems [3], [4]. More recent studies have utilised the artificial intelligent forecasting technique "Neural Network" (more specifically, the multilayer perceptron) [2]. This was similarly successful in the prediction of the Breast Cancer Database. However, in order to gauge the success (or otherwise) of either techniques, a comparative analysis of prediction performance must be made [1]. The use of artificial neural networks (ANNs) is a relatively new computational modelling techniques, yet it has an established record of applications and developments [5]. The technique is biologically inspired, being based upon the computational power of the human brain [6].

Similarly the biological brain, the neural networks consists of a network of interconnected processing elements (neurons) which adjust their "memory" via weights, which link neurons together [7]. This is achieved through repeated representation of training data (as under supervised training/learning) such that an input-output data map (pattern recognition) is obtained [2]. Learning therefore constitutes derivation of minimal error between actual and predicted observations [8]. Neural network and statistics are not competing methodologies for data analysis. There is considerable overlap between the two fields. Statistical methodology is directly applicable to neural networks in a variety of ways including criteria, optimization algorithm, confident intervals, diagnostic and graphical methods. Complementary between the fields of statistic and neural network would benefit both [9]. The great benefit of ANNs is contained within its inherent ability to generalise. Having been trained, the network is able to produce an optimum output on previously unseen data [10],[11]. Moreover, ANNs perform well on data which are noisy, missing observations and imprecise [5]. At this juncture it should be noted that while the mathematical content of ANNs may be complex, the underlying model is basic in comparison the massive computational power of the biological neuron [8].

IV METHODOLOGY

This study utilised of forecasting process to compare conventional methods with neural networks. Step of process in forecasting utilised methodology suggested by [12].

(see Appendix 1)

Forecasting result in multiple regression using the *Statistical Package for the Social Scientist* (SPSS-version 20.0) [13], several analytical trials was conducted on the sample data set. The multiple regression (MR) equation is expressed mathematically as;

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n \qquad \ldots\ldots(1)$$

$$\text{MPE} = \frac{\sum_{i=1}^{n} PE_i}{n} \quad \text{Where; } PE_i = (x_i - p_i / x_i)\ 100\ \% \qquad \ldots\ldots(2)$$

$$\text{MAPE} = \frac{\sum_{i=1}^{n} AE_i}{n} \quad \text{where; } AE_i = \sqrt{(x_i - p_i)}^2 \qquad \ldots\ldots(3)$$

Where; MPE = Mean Percentage Error
MAPE = Mean Average Percentage Error
PE_i = The Percentage Error of number i ; x_i = Actual Value for number of input i ; p_i = Predicted Value for number of input i.

The equation comprises the sum of partial regression coefficients multiplied by their respective predictor variables. To determined how well the model fitted the data, analysis of variance and the associated F test were conducted to test the two hypotesis;

Null Hypothesis : $B_i = B_2 = B_n = 0$ (Siegel and Morgan, 1996)
Alternative Hypothesis : $B_i \neq B_2 \neq B_n \neq 0$

The null hypothesis could be rejected and the conclusion drawn that a linear relationship exists between dependent variable (y = Class [Benign and malignant]) and the all independent variable in the equation.

V DATA AND SOURCE

This study were used 10 attributes and 699 instance for analysis. These attributes were selected as a major factor to contributing for breast cancer. Ten attributes included clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses and class (benign and malignant). This breast cancer database was obtained from the University of Wisconsin Hospital USA, Madison from Dr. William H. Wolberg.

VI RESULTS AND DISCUSSION

The predicting performance obtained by multiple regression and neural network model are exhibited in Table 1. The result

indicated that the Multilayer Perceptron model achieved higher performance (97.14 %) than Multiple Regression (82.10 %). [14] explained that a good statistics to evaluate the model fit is sum of square of error, R-square, Adjusted R-square, mean percentage error, mean absolute percentage error when comparing alternative parameter solutions in the training process and all of these statistics measurement are relevant.

Table 1: Forecasting Performance of Multiple Regression and Multilayer Perceptron Model

Type	RMSE	MPE	MAPE	%
Multiple Regression (MR)	**1.873**	**1.222**	**27.93 %**	**82.10 %**
Multilayer Perceptron (MLP)	**1.552**	**0.528**	**12.40 %**	**97.14 %**

In order to facilitate a comparison between MR and MLP result above, each of the measures of prediction performance was scrutinised and discussed below.

- *MPE*. Analysis of prediction performance using the MPE resulted in table 1 is 12.22 percent and 5.28 percent for MR and MLP models respectively.

- *MAPE*. Scrutiny of MAPE output at 27.93 percent and 12.40 percent for MR and MLP models respectively, revealed that MLP model tent to perform well, although the MLP model appears slightly more accurate.

To determined how well the model fitted the data, analysis of variance and the associated *F* test were conducted to test the two hypotesis;
First Hypothesis: Overall Model.

Null Hypothesis: $B_i = B_2 = B_n = 0$
Alternative Hypothesis: $B_i \neq B_2 \neq B_n \neq 0$

Result reveals that F test = 442.639 and the observed significance level (signif. F) is 0.0000, the null hypothesis could be rejected and the conclusion drawn that a linear relationship exists between dependent variable (y = Class [Benign and malignant]) and the independent variable in the equation.
Second Hypothesis: Individual Model (t-test).

Null Hypothesis: $B_i = B_2 = B_n = 0$
Alternative Hypothesis: $B_i \neq B_2 \neq B_n \neq 0$

In MR model, all independent variable show t-test predictor was significant contribution at 0.05 percent level. So, the null hypothesis could be rejected and the conclusion drawn that a relationship exists between dependent variable (y = Class [Benign and malignant]) and the all independent variable in the equation. The correlation between dependent and independent variable mostly very high correlation through Pearson Correlation test. The two independence's was remove from our model through using scatter plot and stepwise regression.

VI RESULTS AND DISCUSSION

The comparative analysis between MR and MLP models for predicting breast cancer database serves to enhance current work, which seeks to choose the optimum predictor model. Using identical predictor variables one can conclude that both models perform well, although the performance of the Artificial Neural Network (ANN) model is slightly superior. Furthermore, although the average mean residual value is lower for the MR model, suggesting that overall the MR model has slightly less predictive power. Performance analyses using mean percentage error, mean absolute percentage error and percentage of correctness (generalisation). Results reveals that ANNs model perform well, having low mean absolute percentage error values indicating that predictor variables were reliable inputs for modelling breast cancer database. Overall, the neural network model performs slightly better as it was able to predict up to 97.14 % generalisation compare with multiple regression just only 82.10 % generalisation.

References

[1] Hua, G.B., (1996), "Residential construction demand forecasting using economic indicators: a comparative study of artificial neural networks and multiple regression" *Construction Management and Economic*, 14, 25-34.

[2] Edward, D.J., Holt, G.D., & Harris F.C, (1999). " Application of the multilayer perceptron 'neural network' for predicting construction plant maintenance cost', *Computer Integrated Design and Construction.*

[3] Akintoye, A., & Skitmore, M. (1994). " Models of UK private sector quarterly construction demand", *Construction Management and Economic*, 12, 1, 3-13.

[4] Akintola, A.O, (1997). " An Intelligent model of Variations contingency on Construction projects", *Doctor of Philosophy thesis , University of Wolverhampton.*

[5] William H. W. & Mangasarian, O. L. (1990), "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", *Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December, pp 9193 – 9196.*

[6] Swingler, K., (1996) "Applying Neural Networks: A practical Guide, *Academic Press Limited, London.*

[7] Picton, P. (1994). " Introduction to Neural Networks, *Macmillan, London.* [12] Chin Kuo & Arthur Reitsch (1996). " Neural Networks VS. Conventional Methods of Forecasting", *The Journal of Business Forecasting* , Winter 1995-96 pp.17-22.

[8] Schalkoff, R.J. (1997) "Artificial Neural Networks, *McGraw-Hill, London*

[9] Siegel, A. F, & Morgan, C.J. (1996) " Statistics and Data Analysis : An Introduction, 2nd ed, *John Wiley & Son, Chichester*.

[10] Sarle W.S. (1994). "Neural Network and Statistical Models", *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, April .pp.1-13.

[11] Rees, D.G. (1996). " Essential Statistics, 3rd, *Chapman & Hall, London*

[12] Chin Kuo & Arthur Reitsch (1996). " Neural Networks VS. Conventional Methods of Forecasting", *The Journal of Business Forecasting* , Winter 1995-96 pp.17-22.

[13] Coakes, J. & Steed, G. (2001). " SPSS for Window Base System User Guides Analysis without Anguish Release version 10.0, SPSS Inc, *John Wiley & Sons Australia Ltd.*

[14] Stuart M. J. (1997). "A Primer on Neural Networks for Forecasting", *The Journal of Business Forecasting*. Pp. 17-22.