Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 5 (8): 1394-1398, 2005
ISSN 1812-5654

The Effects of Outliers Data on Neural Network Performance

[1]Azme Khamis, [2]Zuhaimy Ismail, [3]Khalid Haron and [3]Ahmad Tarmizi Mohammed
[1]Science Studies Centre, Kolej Universiti Teknologi Tun Hussein Onn, Malaysia
[2]Department of Mathematic, Technology University of Malaysia, Malaysia
[3]Malaysia Palm Oil Board, Malaysia

Abstract: The study was carried out to investigate the influence of outliers on neural network performance in two ways; by examining the percentage outliers and secondly the magnitude outliers. The results of two experiments, training and test data are reported. For training data set, shows that the percentage outliers (ranging from 5 to 30%) and the magnitude of outliers (ranging from $\mu \pm 2$ to $\pm 4\hat{\sigma}$) are statistically significant affected on the modeling accuracy. For test data set, the results show that percentage outliers and magnitude outliers in the used to build the model affect the neural network performance.

Key words: Neural network, percentage-outliers, magnitude-outliers

INTRODUCTION

Outliers in a set of data will influence the modelling accuracy as well as the estimated parameters especially in statistical analysis[1-6]. An outliers is a set of data to be an observation or subset of data which appears to be inconsistent with the remainder of that set of data[3,7]. Reviews show that no extensive study was conducted on the influence of outliers in neural network modelling. The effects of data errors in neural network modelling and found that neural network performance is influenced by errors in the data[8,9]. Observation is defined as outliers if its values are outside the range $\mu \pm 1.5\,\hat{\sigma}$ where, $\hat{\sigma}$ is the estimated variance from the data set[10]. This study examined the effect of outliers on the application of neural network models to the analysis of oil palm yield data.

This experiment was conducted to investigate the influence of outliers on neural network performance in two ways; by examining the percentage of outliers (percentage-outliers) and the magnitude of outliers (magnitude-outliers). In general, when claims about the predictive accuracy of neural networks are made, it is assumed that the data used to train the models and the data input to make modelling, are free of outliers.

NEURAL NETWORK MODEL

A neural network is an artificial intelligence model originally designed to replicate the human brain's learning process. A network consists of many elements or neurons that are connected by communications channels or connectors. These connectors carry numeric data arranged by a variety of means and organized into layers. The neural network can perform a particular function when certain values are assigned to the connections or weights between elements. To describe a system, there is no assumed structure of the model, instead the network are adjusted or trained so that a particular input leads to a specific target output[11-13].

The mathematical model of a neural network comprises of a set of simple functions linked together by weights. The network consists of a set of inputs x, output units y and hidden units z, which link the inputs to outputs (Fig. 1). The hidden units extract useful information from inputs and use them to predict the output. The type on neural network here is known the multilayer perceptron[11,13].

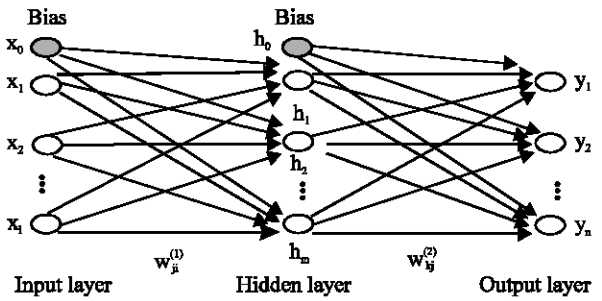

Fig. 1: Feed-forward neural network

Corresponding Author: Azme Khamis, Science Studies Centre, Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

A network with an input vector of elements x_l (l = 1, 2,.., N_i) is transmitted through a connection that is multiplied by weight, w_{jl}, to give the hidden unit z_l (j = 1, 2, 3, …, N_k):

$$z_j = \sum_{l=1}^{N_i} w_{jl}x_l + w_{j0} \quad (1)$$

Where, N_k is the number of hidden units and N_i is the number of input units. The hidden units consist of the weighted input and a bias (w_{j0}). A bias is simply a weight with constant input of 1 that serves as a constant added to the weight. These inputs are passed through a layer of activation function f which produces:

$$h_j = f\left[\sum_{l=1}^{N_i} w_{jl}x_l + W_{j0}\right] \quad (2)$$

The activation functions are designed to accommodate the nonlinearity in the input-output relationships. A common function is sigmoid or hyperbolic tangent:

$$f(z) = \tanh(z) = 1 - \frac{2}{[1+\exp(2z)]} \quad (3)$$

The outputs from hidden units pass another layer of filters:

$$v_k = \sum_{j=1}^{N_h} w_{kj}h_j + w_{k0} = \sum_{j=1}^{N_h} w_{kj} f\left[\sum_{l=1}^{N_i} w_{jl}x_l + w_{j0}\right] + w_{k0} \quad (4)$$

and fed into another activation function F to produce output y (k = 1, 2, 3, …, N_o)

$$y_k = F(v_k) = F\left[\sum_{j=1}^{N_h} w_{kj} f\left(\sum_{l=1}^{N_i} w_{jl}x_l + w_{j0}\right) + w_{k0}\right] \quad (5)$$

The weights adjustable parameters of the network and are determined from a set of data through the process of training[11,14-16]. The training of a network is accomplished using an optimization procedure (such as nonlinear least squares). The objective is to minimize the Sum of Squares of the Error (SSE) between the measured and predicted output. There are no assumptions about functional form, or about the distributions of the variables and errors of the model, NN model is more flexible than the standard statistical technique[17-20]. It allows for nonlinear relationship and complex classificatory equations. The users do not need to specify as much details about the functional form before estimating the classification equation but, instead, it lets the data determine the appropriate functional form[21].

In accordance to standard analytical practice, the sample size was divided on a random basis two sets, namely the training set and the testing set. The training set and the testing set contain 80 and 20 % of the total sample, respectively. To evaluate the modeling accuracy the correlation coefficient, r and MSE were calculated. The model with a higher r and lower MSE was considered to be a relatively superior model.

DATA AND SCOPE

The Malaysian Oil Palm Board (MPOB) provided us with a data set taken from one of the estates in Peninsular Malaysia. The factors included in the data set were foliar composition and Fresh Fruit Bunches (FFB) yield. The variables in foliar composition included percentage of nitrogen, phosphorus, potassium, calcium and magnesium concentration. The concentrations were considered as input variables and the FFB yield as an output variable.

Two factors are considers in this study: (i) the percentage-outliers and (ii) the magnitude-outliers. The percentage-outliers are the percentage of the data in the appropriate section of the data set, which are perturbed. The magnitude-outliers are the degree to which the data deviate from the estimated mean. This study is considered that five input variables and one output variable and 243 data for each variable. The total numbers of observations is 1458. This study considers six levels of percentage-outliers factors from the total numbers of observations; 5, 10, 15, 20, 25 and 30%. The 5% outliers' level means that the data set will contain 72 outliers. Therefore, the 10% level indicates 144 observations, the 15% level indicates 216 observations, the 20% level indicates 288 observations, the 25% level indicates 360 observations and the 30% level indicates 432 observations. This study suggests five levels of magnitude-outliers namely $\mu \pm 2.0\ \hat{\sigma}$, $\mu \pm 2.5\ \hat{\sigma}$, $\mu \pm 3.0\ \hat{\sigma}$, $\mu \pm 3.5\ \hat{\sigma}$ and $\mu \pm 4.0\ \hat{\sigma}$. The observations were selected randomly and replaced uniformly with outliers. For each level of percentage-outliers and magnitude-outliers, the number of hidden nodes increased from five to thirty and the MSE values were recorded.

MATERIALS AND METHODS

The results of the analysis of variance (ANOVA) tests and independent sample t-tests[22] were conducted to test the effects of percentage-outliers and magnitude-outliers on MSE. Tests are also performed to obtain which combinations of percentage-outliers and magnitude-outliers differ significantly from the base-case scenario with no data outliers and their findings are reported. For both experiments, actual and predicted values were compared using mean squares error (MSE) as a measure of modeling accuracy.

RESULTS AND DISCUSSION

Outliers in the training data: Without outliers observation, the MSE value was recorded as 0.0400. The results show that as percentage-outliers increases from 5 to 30%, MSE values also increases, indicating a decrease in modelling accuracy (Table 1). As magnitude-outliers increases from 2 to 4 $\hat{\sigma}$, MSE values also increase, again indicating a decrease in modelling accuracy in the training data.

A one-factor ANOVA test was conducted to investigate the individual effects of percentage-outliers and magnitude-outliers on the neural network's performance. The independent variables are the percentage-outliers (5, 10, 15, 20, 25 and 30%) and the magnitude-outliers $\mu \pm 3.5$, $\mu \pm 2.0$, $\mu \pm 2.5$, $\mu \pm 3.0$, and $\mu \pm 4.0\ \hat{\sigma}$. The F values were recorded as 18.481 ($p = 0.000$) and 3.988 ($p = 0.002$) for the percentage-outliers and magnitude-outliers, respectively, indicating that both factors produced a statistically significant effect on the modelling accuracy.

Following this, the two-factor ANOVA test was conducted to examine the effects of both independent variables on MSE simultaneously. Significant main effects for the percentage-outliers ($F = 28.246$) and the magnitude-outliers ($F = 3.332$) and their interaction ($F = 2.507$), were found as the p-values were less then 0.05. These results indicated that modelling accuracy in the training data could be affected by both the percentage-outliers and the magnitude-outliers.

When more than two levels of factor were conducted, the ANOVA results did not indicate where significant differences occurred. For example, while the percentage-outliers is a significant factor, this difference may be a result of the percentage-outliers changing from 10 to 15%, or 15 to 20%, or 25 to 30%. It could also have come from a larger jump, such as 5 to 25% or 10 to 30%.

The independent t-test was performed to test the MSE values between results with no outliers and the conjunction of percentage-outliers and magnitude-outliers. Independent sample t-tests were performed in order to determine exactly where significant differences

Table 1: The MSE values for different levels of the percentage-outliers and magnitude-outliers in the training data

Magnitude-Outliers ($\hat{\sigma}$)	Percentage-outliers (%)					
	5	10	15	20	25	30
2.0	0.0401	0.0469	0.0573	0.0600	0.0576	0.0595
2.5	0.0411	0.0460	0.0593	0.0617	0.0728	0.0757
3.0	0.0491	0.0545	0.0587	0.0579	0.0734	0.0724
3.5	0.0466	0.0487	0.0649	0.0585	0.0682	0.0799
4.0	0.0464	0.0519	0.0596	0.0629	0.0765	0.0778

Table 2: The t-statistic values in the training data

Magnitude-Outliers ($\hat{\sigma}$)	Percentage-outliers (%)					
	5	10	15	20	25	30
2.0	0.410	-0.918	-2.902*	-3.797*	-3.374*	-2.722*
2.5	0.208	-0.597	-2.857*	-3.266*	-3.687*	-3.517*
3.0	-1.348	-2.080	-3.301*	-3.218*	-3.979*	-3.503*
3.5	-0.897	-0.142	-3.048*	-3.178*	-4.805*	-6.867*
4.0	-0.861	-1.991	-2.831*	-3.990*	-5.147*	-6.211*

* p-value < 0.05

Table 3: The MSE values for different levels of the percentage-outliers and magnitude-outliers in test data

Magnitude-Outliers ($\hat{\sigma}$)	Percentage-outliers (%)					
	5	10	15	20	25	30
2.0	0.0518	0.0517	0.0574	0.0709	0.0762	0.0914
2.5	0.0561	0.0559	0.0691	0.0780	0.0721	0.0785
3.0	0.0460	0.0593	0.0697	0.0748	0.0738	0.0761
3.5	0.0468	0.0583	0.0678	0.0734	0.0913	0.0962
4.0	0.0472	0.0479	0.0619	0.1066	0.1041	0.1224

Table 4: The t-statistic values for the test data

Magnitude-Outliers ($\hat{\sigma}$)	Percentage-outliers (%)					
	5	10	15	20	25	30
2.0	-1.043	-1.196	-3.092*	-5.429*	-2.558*	-8.283*
2.5	-1.365	-0.982	-2.814*	-4.304*	-3.073*	-6.072*
3.0	-0.567	-1.442	-3.535*	-4.461*	-5.086*	-5.669*
3.5	-0.090	-0.523	-2.999*	-3.619*	-5.902*	-6.768*
4.0	-0.172	-0.346	-3.061*	-3.322*	-5.141*	-3.355*

* p-value < 0.05

occurred. For all the $\hat{\sigma}$'s of magnitude-outliers, significant differences ($p<0.05$) were found between the percentage-outliers of 15, 20, 25 and 30% and data sets with no outliers (Table 2). This means that the neural network was first influenced by the outliers in the training data when the percentage-outliers reached 15%. The neural network is unaffected by the outliers impact when the percentage-outliers in the training data is lower than 15%.

Outliers in the test data: Experiment conducted for outliers in test data, which used the same procedures of ANOVA and independent sample t-tests as the training data. Without outliers observation in the data set, the MSE value was recorded as 0.0405. They show that as the percentage-outliers increases from 5 to 30%, the MSE also increases, indicating a decrease in estimate accuracy (Table 3). As the magnitude-outliers increases from 2 to 4 $\hat{\sigma}$, the MSE also increases, which indicates a decrease in the modelling accuracy.

A one-factor ANOVA test was conducted to investigate the individual effects of percentage-outliers and the magnitude-outliers on the neural network's performance in the test data set. The independent variables used are percentage-outliers (6 levels) and magnitude-outliers (5 levels). The F values were recorded as 12.171 ($p = 0.000$) and 3.570 ($p = 0.004$) for the percentage-outliers and magnitude-outliers, respectively. Thus indicate that both factors are statistically significant therefore affecting the modelling accuracy.

Next, the two-factor ANOVA test was conducted to investigate for the effect of both independent variables on MSE simultaneously. Significant main effects for percentage-outliers ($F = 11.709$), magnitude-outliers ($F = 2.640$) and their interaction ($F = 2.273$) were found as the p-values were less then 0.05. These results indicated that the percentage-outliers and magnitude-outliers had an effect on modelling accuracy.

The independent t-tests were also performed to examine the MSE values between results with no outliers and the conjunction of percentage-outliers and magnitude-outliers. Independent sample t-tests were performed in order to determine exactly where significant differences occurred. For all the $\hat{\sigma}$'s of magnitude-outliers, significant differences ($p < 0.05$) were found between percentage-outliers of 15, 20, 25 and 30% and data sets with no outliers (Table 4). Therefore, the conclusion can be made that the neural network was first influenced by the outliers when the percentage-outliers reached 15%. The neural network is resilient to the outliers' impact when the percentage-outliers in the test data is lower than 15%. This result is consistent with the result from the training set data.

CONCLUSIONS

For outliers in the training data, it has been demonstrated that modelling accuracy decreases as the percentage-outliers and magnitude-outliers increases. It has also been shown that the magnitude-outliers affect on modelling accuracy and that the relationship between the percentage-outliers and model accuracy is linear. When the percentage-outliers is lower than 15% (even though the magnitude of outliers may increase), the effect on model accuracy is statistically insignificant as there are no outliers in the training data. The model's accuracy is statistically significant compared to having no outliers data, starting at the combination of 15% of percentage-outliers and magnitude-outliers at all $\hat{\sigma}$'s.

For outliers in the test data it has been demonstrated that modelling accuracy decreases as the percentage-outliers and magnitude-outliers increases. The finding that modelling accuracy decreased as the percentage of outliers increased is a departure from the study of Bansal *et al.*[23], who discussed a neural network application that is not affected by the error rate of test data. Results of this study confirm the findings of Klein and Rossin[9]. One difference between this study and the study of Bansal *et al.*[23] and Klein and Rossin[9] is that the magnitude of the outliers in this study is defined using variance from the data set and has five levels, while their study was based on percentage where only two levels were considered. Therefore, this study shows that variations in the percentage of outliers and magnitude of outliers in the test data may affect modeling accuracy at these higher levels.

REFERENCES

1. Hampel, F.R., 1974. The influence curve and its role in robust estimation. J. Am. Stat. Assoc., 69: 383-393.
2. Andrew, D.F., 1974. A robust method for multiple linear regression. Technometrics, 16: 523-551.
3. Rousseeuw, P.J. and A.M. Leroy, 1987. Robust Regression and Outlier Detection. Wiley, New York.
4. Birkes, D. and Y . Dodge, 1993. Alternative Methods of Regression. John Wiley and Sons, Inc. NY.
5. Mokhtar, A., 1994. Analisis Regresi. Dewan Bahasa dan Pustaka, Kuala Lumpur.
6. Azme, K. and A. Mokhtar, 2004. On robust environmental quality indices. Pertanika J. Sci. Technol., 12: 1-10.
7. Barnett, V. and T. Lewis, 1995. Outliers in Statistical Data. John Wiley and Sons, England.
8. Klein, B.D. and D.F. Rossin, 1999. Data errors in neural network and linear regression models: An experimental comparison. Data Quality J., 5: 1-19.

9. Klein, B.D. and D.F. Rossin, 1999b. Data quality in neural network: effect of error rate and magnitude of error on predictive accuracy. Omega Intl. J. Manage. Sci., 27: 569-582.
10. Huber, P.J., 1981. Robust Statistics. Wiley, New York.
11. Patterson, D.W., 1996. Artificial Neural Networks: Theory and Applications. Prentice Hall, Singapore.
12. Garshenfeld, N., 1999. The Nature of Mathematical Modeling. Cambridge University Press, Cambridge.
13. Hykin, S., 1999. Neural Networks: A Comprehensive Foundation. 2nd Edn., Prentice Hall, New Jersey.
14. Loi, L.L., 1998. Intelligent System Applications in Power Engineering: Evolutionary Programming and Neural Networks. John Wiley and Sons. West Sussex, England.
15. Wong, B.K., V.S. Lai and J. Lam, 2000. A bibliography of neural networks business applications research: 1994-1998. Computers and Operations Res., 27: 1045-1076.
16. Zhang, G.P., G.E. Patuwo and M.Y. Hu, 2001. A simulation study of artificial neural networks for nonlinear time-series forecasting. Computers and Operations Res., 28: 381-396.
17. Ripley, B.D., 1994. Neural networks and flexible regression and discrimination. Adv. Applied Stat., pp: 39-57.
18. Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.
19. Faraway, J. and C. Chatfield, 1998. Time series forecasting with neural networks: A case study. J. Royal Stat. Series C., 47: 231-250.
20. Sarles, W.S., 1994. Neural networks and statistical models. Proceedings of the Nineteenth Annual SAS Users Group International Conference, Aprtil 1994, 1-13.
21. Limsombunchai, V., C. Gan and M. Lee, 2004. House price prediction: Hedonic price model vs. artificial neural network. Am. J. Applied Sci., 1: 193-201.
22. Norusis, M.J., 1998. SPSS® 8.0. Guide to Data Analysis. Prentice Hall, New Jersey.
23. Bansal, A., R. Kauffman and R. Weitz, 1993. Comparing the modeling performance of regression and neural networks as data quality varies: A business value approach. J. Manage. Inform. Sys., 10: 11-32.