Investigating the Effect of Data Representation On Neural Network and Regression

Fadzilah Siraj[a], Ehab A. Omer El Fallah[b]

[a,b]College of Arts and Sciences
Universiti Utara Malaysia, UUM 06010, Sintok, Kedah, Malaysia
Tel: 04-9284672, Fax: 04-9284753
E-mail: [a]fad173@uum.edu.my, [b]ea_fallah@hotmail.com

ABSTRACT

In this research the impact of different data representation on the performance of neural network and regression was investigated on different datasets that has binary or Boolean class target. In addition, the performance of particular predictive data mining model could be affected with the change of data representation. The seven data representations that have been used in this research are As_Is, Min Max normalization, standard deviation normalization, sigmoidal normalization, thermometer representation, flag representation and simple binary representation. Moreover, all data representations have been applied on two datasets which are Wisconsin breast cancer and German credit dataset. As a result, the neural network performance is better than logistic regression on both datasets if we exclude the thermometer and flag representations. For datasets having a binary or Boolean target class, flag or thermometer binary representation is recommended to be used if logistic regression analysis is performed. Meanwhile, As_is representation, min max normalization, standard deviation normalization or sigmoidal normalization is recommended for neural network analysis on datasets having binary or Boolean target class.

Keywords: *Data Representation, Neural Network, Logistic Regression*

1.0 INTRODUCTION

Investigating the prediction performance on different datasets involves many uncertainties for different data type. In the task of prediction, one particular predictive model might give the best result for one dataset but gives poor result in another dataset although these two datasets contain the same data with different representations (Hashemi *et al.*, 2002; Jia & Chua, 1993; Altun *et al.*, 2000; O'Neal *et al.*, 2002; Wessels *et al.*, 2002).
The objective of this research is to investigate the impact of various prediction models on different datasets. This study focuses on two predictive data mining models which are commonly used for prediction purposes, namely neural network and regression model. One medical dataset (Wisconsin Breast Cancer) and one business data (German credit) that have Boolean targets have being used for experimental purposes.

This paper is organized into 5 sections. Section 2 discusses the related work on data representation. The methodology and the experiments for carrying out the investigation are covered in Section 3. The results are the subject of discussion in Section 4. Finally, the conclusion and future research are presented in Section 5.

2.0 RELATED WORKS

Neural networks (NN), decision trees, and logistic regression are three classification models that are commonly used for comparative studies (Delen & Patil, 2006). These models have been applied to a prostate cancer dataset obtained from SEER (the Surveillance, Epidemiology, and End Results) program of the National Cancer Institute. The results of the study show that NN performed best with the highest accuracy, sensitivity and specificity, followed by decision tree and then logistic regression. Similar models have been applied to detect credit card fraud. The results indicate that neural networks give better performance than logistic regression and decision tree (Shen *et al.*, 2007).

In data representation study, Hashemi *et al.* (2002) used NN to extrapolate the presence of mercury in human blood from animal data. The effect of different data representations such as ***As-is, Category, Simple binary, Thermometer***, and ***Flag*** on the prediction models. The study concludes that the ***Thermometer*** data representation using NN performs extremely well.

O'Neal *et al.* (2002) used five different data representations (***Maximum Value***, ***Maximum*** and ***Minimum Value***, ***Logarithm***, T***hermometer*** (powers of 10), and ***Binary*** (powers of 2)) on a set of data to predict maize yield at three scales in east-central Indiana of the

Midwest USA. The data used consist of weather data and yield data from farm, county and state levels from the year 1901 to 1996. The results indicate that data representation has a significant effect on NN performance.

In another study, Zhu *et al.* (2001) investigate the performance of data representation formats such as ***Binary*** and ***Integer*** on the classification accuracy of network intrusion detection system. Three data mining techniques such as rough sets, NN and inductive learning were applied on binary and integer representations. The experimental results show that different data representations did not cause significant difference in the classification accuracy. This may be due to the fact that the same phenomenon were captured and put into different representation formats (Zhu *et al.*, 2001). In addition, the data was primarily discrete values of qualitative variables (system class), and different results could be obtained if the values were continuous variables.

Data representation plays a crucial role in the performance of neural network, "especially for the applications of neural networks in real world" (Jia & Chua, 1993). Numerical encoding schemes (***Decimal Normalization*** and ***Split Decimal Digit representation***) and bit pattern encoding schemes (***Binary representation***, ***Binary Code Decimal representation***, ***Gray Code representation***, ***Temperature code representation***, and ***Gray Coded Decimal representation***) were applied on Fisher Iris data and the performance of the various encoding approaches were analyzed. The results indicate that encoding approaches affect the training errors (such as maximum error and root mean square error) and encoding methods that uses more input nodes that represent one single parameter resulted in lower training errors. Consequently, Jia and Chua's (1993) work laid an important foundation for later research on the effect of data representation on the classification performance using neural network.

Continuing the empirical study by Jia and Chua (1993), a theoretical background has been provided by Altun, Halcinoz and Tezekiei (2000) to support the fndings that input data manipulation could improve neural learning in NN. Altun *et al*. (2000) evaluated the impact of the modified training sets and how the learning process depends on data distribution within the training sets. NN training was performed on input datasets that has been arranged so that three different sets are produced with each set having different number of occurrence of 1's and 0's. The ***Temperature Encoding*** is then employed on the three datasets and then being used to train NN again. The results show that by employing ***Temperature Encoding*** on the datasets, the training process is improved by significantly reducing the number of epochs or iteration needed for training. Altun ***et al***. (2000)'s findings proved that by changing input data representation, the performance of a NN model is affected.

3.0 METHODOLOGY

The methodology in this research is being adapted from Hashemi *et al*. (2002)'s approach of using different data representations. The major steps involved in the methodology starts with data collection, followed by data preparation, analysis and experiment, and finally investigation and comparison stage.

The Wisconsin Breast Cancer dataset comprises of 699 cases with ten attributes including the target. The second dataset (German Credit) consists of 1000 credit card applicants with twenty attributes plus the target class.

Each dataset has been transformed into data representation identified in this study, namely ***As_Is***, ***Min Max Normalization***, ***Standard Deviation Normalization***, ***Sigmoidal Normalization***, ***Thermometer Representation***, ***Flag Representation*** and ***Simple Binary Representation***.

In ***As_Is*** representation, the data remain the same as the original data without any changes. The ***Min Max Normalization*** is used to transform all values into numbers between 0 and 1. The ***Min Max Normalization*** applies linear transformation on the raw data, keeping the relationship of the data values in the same range. This method does not deal with any possible outliers in the future value, and the min max formula (Kantardzic, 2003) is written as:

$$V'=(v-Min(v(i)))/(Max(v(i))-Min(v(i))) \qquad (1)$$

Where V' is the new value, Min(v(i)) is the minimum value in the particular attribute, Max(v(i)) the maximum value in particular attribute and v is the old value.

The ***Standard Deviation Normalization*** is a technique based on the mean value and standard deviation function for each attribute on the dataset. For a variable v, the mean value **Mean(v)** and the standard deviation **Std_dev(v)** are calculated from dataset itself. The standard deviation normalization formula (Kantardzic, The ***Sigmoidal Normalization*** transforms all nonlinear input data into the range between -1 and 1 using a sigmoid function. It calculates the mean value and standard deviation function value from the input data. Data points within a standard deviation of the mean are converted to the linear area of the sigmoid. In addition, outlier points in the data are compacted along the sigmoidal function tails. The sigmoidal normalization formula(Kantardzic, 2003) is given by:

$$V' = (1-e^{(-a)}) / (1+e^{(-a)}) \qquad (3)$$

where

$a = (v - mean(v))/std\_dev(v)$
$meanv(v) = Sum(v)/n$
$std\_dev(v) = sqr(sum(v^2) - (sum(v)^2/n)/(n-1))$

In the ***Thermometer*** representation, the categorical value was converted into binary form prior to performing analysis. For example, if the range of values for a category field is 1 to 6, then the value of 4 that needs to be represented in thermometer format will have the representation of "111100" (Hashemi *et al*., 2002).

In the ***Flag*** format, digit 1 is represented in the binary location for the value. Thus, following the same assumption that the range values for a category field is 1 to 6, if the value 4 needs to be represented in ***Flag*** format, the representation will be shown as "000100". The representation in ***Simple Binary*** is obtained by directly changing the categorical value into binary. Table 1 exhibits the different representations of Wisconsin Breast Cancer and German Credit dataset.

The training of multi layer perceptron (MLP) could be stated as a nonlinear optimization problem. The objective of multi layer perceptron is performing *learning* in order to identify the best weights that minimize the difference between the input and the output. The most popular training algorithm used in NN is back propagation, and it has been used in solving many problems in pattern recognition and classification.

Table 1: Various dataset representations

Representations	Wisconsin Breast Cancer	German Credit
As_Is representation	5 | 4 | 3 |	1 | 6.0 | 4 |
Min Max normalization	.0000 | .4444 | .3333 |	.0000 | .0294 | 1.000 |
Standard Deviation normalization	-1.637 | .2068 | .2836 |	-1.264 | -1.236 | 1.343 |
Sigmoidal normalization	-.675 | .1102 | .149 |	-.362 | .8103 | -.676 |
Thermometer representation	1111100000011110	10001000000111111
Flag representation	000010000000010000	10001000000001000
Simple Binary representation	01010100000110000	0001000010100000110

Logistic regression is a statistical regression model for binary dependent variables (Yun *et al.*, 2007), which is simpler in terms of computation during training while still giving a good classification performance (Ksantini, *et al.*, 2008). Figure 2 shows the general steps involve in performing NN experiments and regression analysis using different data representations in this study.

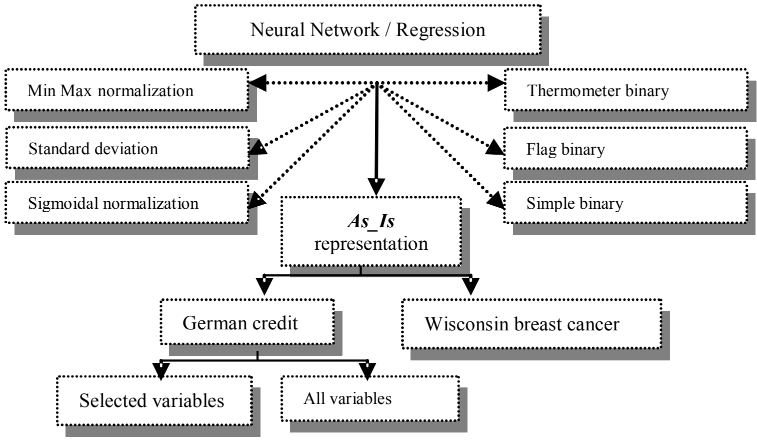

Figure 2: Illustration of Data Representation for NN/ Regression analysis experiments

4.0 RESULTS

Initial experimental results of correlation analysis on Wisconsin Breast Cancer indicate that all attributes (independent variables) have significant correlation with the dependent variable (target). However, German Credit dataset indicates otherwise. Therefore, for German Credit dataset, two different approaches (all dependent variables and selected variables) were performed in order to complete the investigation.

Based on the results exhibited in Table 2, although NN obtained the same percentage of accuracy, ***As_Is*** achieved the lowest training results (98.57%, 96.24%). On the other hand, regression exhibits highest percentage of accuracy for ***Thermometre*** and ***Flag*** representation (100%) followed by ***Simple Binary*** representation.

Table 2: *Percentage of accuracy for Wisconsin Breast Cancer Dataset*

	Wisconsin Breast Cancer		
	Neural Network		Regression
	Train	Test	Accuracy
As_Is representation	96.24%	98.57%	96.9%
Min Max normalization	96.42%	98.57%	96.9%
Standard Deviation normalization	96.42%	98.57%	96.9%
Sigmoidal normalization	96.60%	98.57%	96.9%
Thermometer representation	97.14%	98.57%	100.0%
Flag representation	97.67%	98.57%	100.0%
Simple Binary representation	97.14%	98.57%	97.6%

Referring to the result shown in Fig. 3, similar observation has been noted for German Credit dataset when **all variables** are considered in the experiments. ***As_Is*** representation obtained the highest percentage of accuracy (79%) for NN model. For regression analysis, ***Thermometre*** and ***Flag*** representation obtained the highest percentage of accuracy (80.1%). Similar to earlier observation on the Wisconsin Breast Cancer dataset, ***Simple Binary*** representation obtained the second highest percentage of accuracy (79.5%).

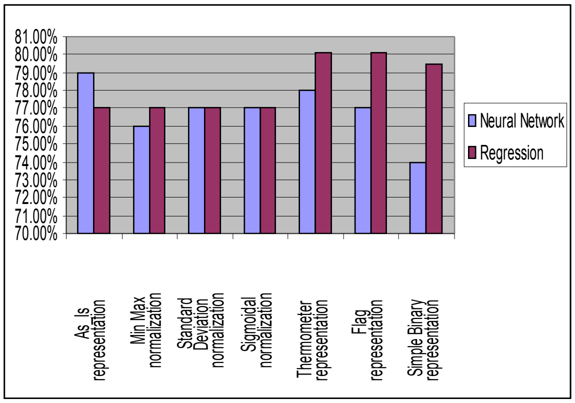

Figure 3: German Credit All Variables accuracy for Neural Network and Regression

When **selected variables** of German Credit dataset was tested with NN, the highest percentage accuracy was obtained using ***As_Is*** representation (80%), followed by ***Standard Deviation Normalization*** (79%) ***Min Max Normalization*** (78%) and **Thermometer** (78%) representation. The regression results show similar patterns with results illustrated in Figure 3. In other words, the data representation techniques, namely ***Thermometer*** (77.4%) and Flag(77.4%) representations produce the highest and second highest percentage of accuracy for selected variables of German Credit.

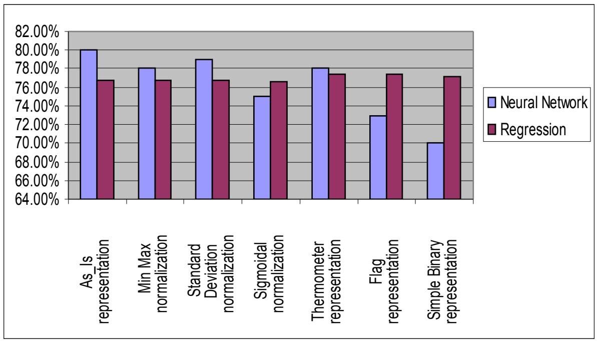

Figure 4: German Credit Selected Variables accuracy for Neural Network and Regression

For brevity, Table 3 exhibits NN parameters that produce the highest percentage of accuracy for Wisconsin Breast Cancer, and German Credit dataset using all variables as well as selected variables in the experiments.

Table 3: *The summary of NN experimental results using* ***As_Is*** *representation*

Neural Network	Wisconsin Breast lCancer	German credit using all variables	German credit using selected variables
Percentage of Accuracy	**98.57%**	80.00%	79.00%
Input units	9	20	12
Hidden units	2	6	20
Learning rate	0.1	0.6	0.6
Momentum rate	0.8	0.1	0.1
No. of epoch	100	100	100

The logistic regression and correlation results for Wisconsin Breast Cancer dataset are exhibited in Table 4. Note that based on Wald Statistics, variables such as ***CellSize***, ***Cellshape, EpiCells, NormNuc*** and ***Mito*** are not significant to the prediction model. However, these variables have significant correlation with Type of Breast Cancer. Thus, the logistic regression independent variables include all variables listed in Table 4.

Table 4*: List of variables included in logistic regression of Wisconsin breast cancer*

Logistic Regression			**Correlation**	
Variables	B	Sig.	r	p
CTHick	.531	.000		
CellSize	.006	.975	.818(**)	.000
CellShape	.333	.109	.819(**)	.000

MarAd	.240	.036		
EpiCells	.069	.645	.683(**)	.000
BareNuc	.400	.000		
BLChr	.411	.009		
NormNuc	.145	.157	.712(**)	.000
Mito	.551	.069	.423(**)	.000
Constant	-9.671	.000		

For German Credit dataset, NN obtained the highest percentage of accuracy when all variables are considered in the training (see Table 3). The appropriate parameters for this dataset are also listed in the same table.

The summary of logistic regression results is shown in Table 5. All shaded variables displayed in Table 5 are significant independent variables for determining whether a credit application is successful or not.

Note also that variable *Age* is not significant to German Credit target. However, its correlation with the target is significant therefore this variable is included in logistic regression equation that represents German credit application.

Table 5: List of variables included in logistic regression of German Credit dataset

Regression (Thermometer representation)	**German Credit using all variables (80%)**			
Variables	Logistic Regression		Correlation	
	B	Sig.	r	p
SECA	-.588	000	-.348(**)	.000
DurMo	.025	.005	.206(**)	.000
CreditH	-.384	.000	-.222(**)	.000
CreditA	-.384	.018	.087(**)	.003
SavingA	-.240	.000	-.175(**)	.000
EmploPe	-.156	.029	-.120(**)	.000
InstalRate	.300	.000	.074(**)	.010
PersonalS	-.267	.022	-.091(**)	.002
OtherDep	-.363	.041	-0.003	.460
Property	.182	.046	.141(**)	.000
Age	**-.010**	**.246**	**-.112(**)**	**.000**
OtherInst	-.322	.004	-.113(**)	.000
Forgn Work	-1.216	.047	-.082(**)	.005
Constant	4.391	.000		

5.0 CONCLUSION AND FUTURE RESEARCH

In this study, the effect of different data representations on the performance of neural network and regression was investigated on different datasets that have binary or Boolean class target. The results indicate that different data representation produces different percentage of accuracy.

Based on the empirical results, data representation ***As_Is*** is a better approach for NN with Boolean targets (see also Table 6). NN has shown consistent performance for both datasets. Further inspection on the results exhibited in Table 6 also indicate that for German Credit dataset, NN performance improves by 1%. This leads to suggestion that by considering correlation and regression analysis, both NN results using ***As_Is*** and ***Standard Deviation Normalization*** could be improved. For regression analysis, ***Thermometer, Flag*** and ***Simple Binary*** representations produce consistent regression performance. However, the performance decreases when the independent variables have been reduced through correlation and regression analysis.

As for future research, more datasets will be utilized to investigate further on the effect of data representation on the performance of both NN and regression. One possible area is to investigate which cases fail during training, and how to correct the representation of cases such that the cases will be correctly identified by the model. Studying the effect of different data representations on different predictive models enable future researchers or data mining models developer to present data correctly for binary or Boolean target in the prediction task.

Table 6: *Summary of NN and regression analysis of German Credit dataset*

REFERENCES

Altun, H., Talcinoz, T. & Tezekiei B. S. (2000). Improvement in the Learning Process as a Function of Distribution Characteristics of Binary Data Set. *Proceedings of the 10th Mediterranean Electrotechnical Conference*. Vol. 2. 567-569.

Delen, D. & Patil, N. (2006). Knowledge Extraction from Prostate Cancer Data. *Proceedings of the 39th Annual Hawaii International Conference, HICSS '06: System Sciences*. 04-07 Jan. Vol. 5 92b-92b.

	German Credit All Variables			German Credit Selected Variables		
	Neural Network		Regn	Neural Network		Regn
	Train	Test		Train	Test	
As_Is representation	77.25	79.00	77.0	75.00	80.00	76.8
Min Max normalization	76.50	76.00	77.0	75.25	78.00	76.8
Standard Deviation normalization	76.75	77.00	77.0	75.13	79.00	76.8
Sigmoidal normalization	76.75	77.00	77.0	74.00	75.00	76.6
Thermometer representation	78.38	78.00	80.1	77.00	78.00	77.4
Flag representation	76.75	77.00	80.1	75.13	73.00	77.4
Simple Binary representation	75.75	74.00	79.5	70.63	70.00	77.1

Hashemi R. R., Bahar, M., Tyler, A. A. & Young, J. (2002). The Investigation of Mercury Presence in Human Blood: An Extrapolation from Animal Data Using Neural Networks. *Proceedings of International Conference: Information Technology: Coding and Computin*g. 8-10 April. 512-517.

Jia, J. & Chua, H. C. (1993). Neural Network Encoding Approach Comparison: An Empirical Study. *Proceedings of First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems.* 24-26 November . 38-41.

Kantardzic, M. (2003). DATA MINING: Concepts, Models, Methods and Algorithms. *IEEE Transactions on Neural Networks*, 14(2), 464-464.

Ksantini, R., Ziou, D., Colin, B., & Dubeau, F. (2008). Weighted Pseudometric Discriminatory Power Improvement Using a Bayesian Logistic Regression Model Based on a Variational Method. *IEEE Transactions* on Pattern Analysis and Machine Intelligence. 30(2), 253-266.

O'Neal, M. R., Engek, B. A. Ess, D. R., & Frankernberger, J. R. (2002). J. R. (2002). Neural network prediction of maize yield using alternative data coding algorithms. *Biosystems Engineering*. 84. 31-45.

Shen, A., Tong, R., & Deng, Y. (20307). Application of Classification Models on Credit Card Fraud Detection. *International Conference: Service Systems and Service Management*, 9-11 June. 1-4.

UCI Machine Learning Repository. (n.d.). Retrieved May 15, 2008, from http://www.ics.uci.edu/~mlearn/MLRepository.html.

Wessels, L.F.A., Reinders, M.J.T., Welsem, T. V. & Nederlof, P. M. (2002). Representation and classification for high-throughput data sets. SPIE-BIOS2002, *Biomedial Nanotechnology Architectures and Applications*, 4626, 226-237, San Jose, USA.

Yun, W. H., Kim, D. H., Chi, S. Y. & Yoon, H. S. (2007). Two-dimensional Logistic Regression. *19th IEEE International Conference, ICTAI 2007: Tools with Artificial Intelligence*. 29-31 October , Vol. 2., 349-353.

Zhu, D., Premkumar, G., Zhang, X. & Chu, C.H. (2001). Data mining for Network Intrusion Detection: A Comparison of Alternative Methods. *Decision Sciences*, 32(4), 635-660.