Chapter 4

Data Mining and Neural Networks: The Impact of Data Representation

Fadzilah Siraj, Ehab A. Omer A. Omer and Md. Rajib Hasan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51594

1. Introduction

The extensive use of computers and information technology has led toward the creation of extensive data repositories from a very wide variety of application areas [1]. Such vast data repositories can contribute significantly towards future decision making provided appropriate knowledge discovery mechanisms are applied for extracting hidden, but potentially useful information embedded into the data [2].

Data mining (DM) is one of the phases in knowledge discovery in databases. It is the process of extracting the useful information and knowledge in which the data is abundant, incomplete, ambiguous and random [3], [4], [5]. DM is defined as an automated or semi-automated exploratory data analysis of large complex data sets that can be used to uncover patterns and relationships in data with an emphasis on large observational databases [6]. Modern statistical and computational technologies are applied to the problem in order to find useful patterns hidden withina large database [7], [8], [9]. To uncover hidden trends and patterns, DM uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge. In effect, the predictive models formed from the trends and patterns through DM enable analysts to produce new observations from existing data. DM methods can also be viewed as statistical computation, artificial intelligence (AI) and database approach[10]. However, these methods are not replacing the existing traditional statistics; in fact, it is an extension of traditional techniques. For example, its techniques have been applied to uncover hidden information and predict future trends in financial markets. Competitive advantages achieved by DM in business and finance include increased revenue, reduced cost, and improved market place responsiveness and awareness [11]. It has also been used to derive new information that could be integrated in decision support, forecasting and estimation to help business gain competitive advantage [9]. In higher educational institutions, DM can be used in the process of uncovering hidden trends and patterns that help them in forecasting the students' achievement. For instance, by using DM

approach, a university could predict the accuracy percentage of students' graduation status, whether students will or will not be graduated, the variety of outcomes, such as transferability, persistence, retention, and course success[12], [13].

The objective of this study is to investigate the impact of various data representations on predictive data mining models. In the task of prediction, one particular predictive model might give the best result for one data set but gives a poor results in another data set although these two datasets contain the same data with different representations [14],[15],[16], [17]. This study focuses on two predictive data mining models, which are commonly used for prediction purposes, namely neural network (NN) and regression model. A medical data set (known as Wisconsin Breast Cancer) and a business data (German credit) that has Boolean targets are used for experimental purposes to investigate the impact of various data representation on predictive DM model. Seven data representations are employed for this study; they are As_Is, Min Max normalization, standard deviation normalization, sigmoidal normalization, thermometer representation, flag representation and simple binary representation.

This chapter is organized as follows. The second section describes data mining, and data representation is described in the third section. The methodology and the experiments for carrying out the investigations are covered in Section 4. The results are the subject of discussion which is presented in Section 5. Finally, the conclusion and future research are presented in Section 6.

2. Data mining

It is well known that DM is capable of providing highly accurate information to support decision-making and forecasting for scientific, physiology, sociology, the military and business decision making [13]. DM is a powerful technology with great potential such that it helps users focus on the most important information stored in data warehouses or streamed through communication lines. DM has a potential to answer questions that were very time-consuming to resolve in the past. In addition, DM can predict future trends and behavior, allowing us to make proactive, knowledge-driven decisions [18].

NN, decision trees, and logistic regression are three classification models that are commonly used in comparative studies [19]. These models have been applied to a prostate cancer data set obtained from SEER (the Surveillance, Epidemiology), and results program of the National Cancer Institute. The results from 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 NN give better performance than logistic regression and decision tree [20].

3. Data representation

Data representation plays a crucial role on the performance of NN, "especially for the applications of NNs in a real world." In data representation study,[14] used NNs 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 are investigated. The study concludes that the *Thermometer* data representation using NN performs extremely well.

[16], [21] 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 [17]. The data used to 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, [21] 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 to the classification accuracy. This may be due to the fact that the same phenomenon were captured and put into different representation formats [21]. 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.

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, [22] work laid an important foundation for later research on the effect of data representation on the classification performance using NN.

[22] conducted an empirical study based on a theoretical provided by [15] to support the findings that input data manipulation could improve neural learning in NN. In addition, [15] 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 data set that has been arranged so that three different sets are produced with each set having a different number of occurrences of 1's and 0's. The *Temperature Encoding* is then employed on the three data sets and then being used to train NN again. The results show that by employing *Temperature Encoding* on the data sets, the training process is improved by significantly reducing the number of epochs or iteration needed for training. [15]'s findings proved that by changing input data representation, the performance in a NN model is affected.

4. Methodology

The methodology for this research is being adapted from [14] by using different data representations on the data set, and the steps involved in carrying out the studies are shown

in Figure 1 [14]. The study starts with data collection, followed by data preparation stage, analysis and experiment stage, and finally, investigation and comparison stage.

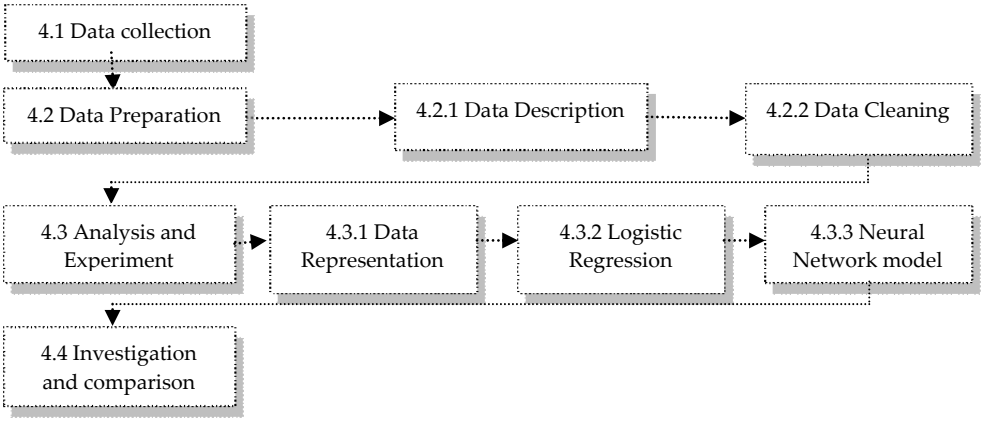


Figure 1. Steps in carrying out the study

4.1. Data collection

At this stage, data sets have been acquired through the UCI machine learning repository which can be accessed at http://archive.ics.uci.edu/ml/ datasets.html. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for conducting empirical studies on machine learning algorithms. Two types of data have been obtained from UCI; they are Wisconsin Breast Cancer data set and German credit data set.

4.2. Data preparation

After the data has been collected in the previous stage, data preparation would be performed to prepare the data for the experiment in the next stage. Each attribute is examined and missing values are treated prior to training.

4.2.1. Data description

In this study, two sets of data are used, namely Wisconsin Breast Cancer and German Credit. Each data set is described in details in the following subsections.

4.2.1.1. Wisconsin breast cancer data set

Wisconsin breast cancer data set is originated from University of Wisconsin Hospitals, Madison donated by Dr. William H. Wolberg. Each instance or data object from the data represents one patient record. Each record comprises of information about Breast Cancer patient whose cancer condition is either benign or malignant. A total of 699 cases in the data

set with nine attributes (excluding Sample Code Number) that represent independent variables and one attribute, i.e. Class represent the output or dependent variable.

Table 1 describes the attribute in the data set, code which represents the short form for this attribute, type, which shows the data type for particular attribute, domain, which represents the possible range in the value and the last column, shows the missing values in all attributes in the study. From Table 1, only one attribute has been missing values (a total of 16 instances), and this attribute is Bare Nuclei.

No	Attribute description	Code	Type	Domain	Missing value
1	Sample code number	CodeNum	Continues	Id number	0
2	Clump Thickness	CTHick	Discrete	1 – 10	0
3	Uniformity of Cell Size	CellSize	Discrete	1 – 10	0
4	Uniformity of Cell Shape	CellShape	Discrete	1 – 10	0
5	Marginal Adhesion	MarAd	Discrete	1 – 10	0
6	Single Epithelial Cell Size	EpiCells	Discrete	1 – 10	0
7	Bare Nuclei	BareNuc	Discrete	1 – 10	16
8	Bland Chromatin	BLChr	Discrete	1 – 10	0
9	Normal Nucleoli	NormNuc	Discrete	1 – 10	0
10	Mitoses	Mito	Discrete	1 – 10	0
11	Class:	Cl	Discrete	2 for benign 4 for malignant	0

Table 1. Attribute of Wisconsin Breast Cancer Dataset

Based on the condition of Breast Cancer patients, a total of 65.5% (458) of them has benign condition and the rest (34.5% or 241) is Malignant.

4.2.1.2. German credit dataset

German credit data set classifies applicants as good or bad credit risk based upon a set of attributes specified by financial institutions. The original data set is provided by Professor Hofmann contains categorical and symbolic attributes. A total of 1000 instances have been provided with 20 attributes, excluding the German Credit Class (Table 2). The applicants are classified as good credit risk (700) or bad (300) with no missing value in this data set.

No.	Attribute description	Code	Type	Domain	Missing value
1	Status of existing checking account	SECA	Discrete	1, 2, 3, 4	0
2	Duration in month	DurMo	Continuous	4 - 72	0
3	Credit history	CreditH	Discrete	0, 1, 2, 3, 4	0
4	Purpose	Purpose	Discrete	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	0

No.	Attribute description	Code	Type	Domain	Missing value
5	Credit amount	CreditA	Continuous	250 - 18424	0
6	Savings account/bonds	SavingA	Discrete	1, 2, 3, 4, 5	0
7	Present employment since	EmploPe	Discrete	1, 2, 3, 4, 5	0
8	Instalment rate in percentage of disposable income	InstalRate	Continuous	2 – 4	0
9	Personal status	PersonalS	Discrete	1, 2, 3, 4, 5	0
10	Other debtors / guarantors	OtherDep	Discrete	1, 2, 3	0
11	Present residence since	PresentRe	Discrete	1 – 4	0
12	Property	Property	Discrete	1, 2, 3, 4	0
13	Age in years	Age	Continuous	19 – 75	0
14	Other instalment plans	OtherInst	Discrete	1, 2, 3	0
15	Housing	Housing	Discrete	1, 2, 3	0
16	Number of existing credits at bank	NumCBnk	Discrete	1,2,3	0
17	Job	Job	Discrete	1, 2, 3, 4	0
18	Number of people being liable to provide maintenance for	Numppl	Discrete	1, 2	0
19	Telephone	Telephone	Discrete	1, 2	0
20	Foreign worker	ForgnWor	Discrete	1, 2	0
21	German Credit Class	GCL	Discrete	1 good 2 bad	0

Table 2. Attribute of German Credit Dataset

4.2.2. Data cleaning

Before using the data that has been collected in the previous stage, missing values should be identified. Several methods that could be performed to solve missing values on data, such as deleting the attributes or instances, replacing the missing values with the mean value of a particular attribute, or ignore the missing values. However, which action would be performed to handle the missing values depends upon the data that has been collected.

German credit application data set has no missing values (refer to Table 2); therefore, no action was taken on German credit data set. On the other hand, Wisconsin breast cancer data set has 16 missing values of an attribute Bare Nuclei (see Table 1). Therefore, these missing values have been resolved by replacing the mean value to this attribute. The mean value to this attribute is 3.54, since the data type for this attribute is categorical so the value was rounded to 4. Finally, all the missing values have been replaced by value 4.

4.3. Analysis and experiment

The data representations used for the experiments are described in the following subsections.

4.3.1. Data representation

Each data set has been transformed into data representation identified for 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 to 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 [25] is written in Eqn. (1).

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

Where V′ is the new value, Min(v(i)) is the minimum value in a particular attribute, Max(v(i)) the maximum value in a 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 data set. For a variable v, the mean value **Mean (v)** and the standard deviation Std_dev(v) is calculated from the data set itself. The standard deviation normalization formula [25] is written as in Eqn. (2).

$$V' = \frac{(v - mean(v))}{std\_dev(v)} \quad (2)$$

where

$$meanv(v) = \frac{Sum(v)}{n}$$

std_dev(v)= sqr(sum(v^2)-(sum(v)^2/n)/(n-1))

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 to the data are compacted along the sigmoidal function tails. The sigmoidal normalization formula [25] is given by Eq. (3).

$$V' = \frac{(v - mean(v))}{std\_dev(v)} \quad (3)$$

Where

$$a = \frac{(v - mean(v))}{std_{dev(v)}}$$

$$mean(v) = \frac{Sum(v)}{n}$$

std_dev(v)= sqr(sum(v^2)-(sum(v^2)/n)/(n-1))

In the *Thermometer* representation, the categorical value was converted into a binary form prior to performing analysis. For example, if the range of values for a category field is 1 to 6, thus value 4 can be represented in thermometer format as "111100" [15].

In the *Flag* format, digit 1 is represented in the binary location for the value. Thus, following the same assumption that the range values in 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 3 exhibits the different representations of Wisconsin Breast Cancer and German Credit data set.

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	-1.254 -1.236 1.343
Sigmoidal normalization	-.675 .1102 .149	-.362 .8103 -.576
Thermometer representation	1 1 1 1 1 0 0 0 0 0 1 1 1 1 0	1 0 0 0 1 0 0 0 0 0 1 1 1 1 1
Flag representation	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0	1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0
Simple Binary representation	0 1 0 1 0 1 0 0 0 0 1 1 0 0 0	0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 1 0

Table 3. Various dataset representations

4.3.2. Logistic regression

Logistic regression is one of the statistical methods used in DM for non-linear problems either to classify or for prediction. Logistic Regression is one of the parts of statistical models, which allows one to predict a discrete outcome (known as dependent variable), such as group membership, from a set of variables (also known as independent variables) that may be continuous, discrete, dichotomous, or a combination of any of these. The logistic regression aims to correctly predict the category of outcome for individual cases using the most parsimonious model. In order to achieve the goal, a model is created, which comprises of all predictor (independent) variables that are useful in predicting the desired target. The relationship between the predictor and the target is not linear instead; the logistic regression function is usedwhose equation can be written as Eqn. (4) [26].

$$\theta = \frac{\exp\left(\beta_0 + \beta_{1}X_1 + \ldots + \beta_{k}x_k\right)}{1 + \exp\left(\beta_0 + \beta_{1}X_1 + \ldots + \beta_{k}x_k\right)} \tag{4}$$

Where α = the constant from the equation andβ = the coefficient of the predictor variables. Alternatively, the logistic regression equation can be written as Eqn. (5).

$$\log it\left[\theta(x)\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \left(\beta_0 + \beta_{1}X_1 + \ldots + \beta_{k}x_k\right) \tag{5}$$

Anodd's ratio is formed from logistic regression that calculates the probability or success over the probability of failure. For example, logistic regression is often used for epidemiological studies where the analysis result shows the probability of developing cancer after controlling for other associated risks. In addition, logistic regression also provides knowledge about the relationships and strengths among the variables (e.g., smoking 10 packs a day increases the risk for developing cancer than working in asbestos mine)[27].

Logistic regression is a model which is simpler in terms of computation during training while still giving a good classification performance [28]. The simple logistic regression model has the form as in Eqn. (6), viz:

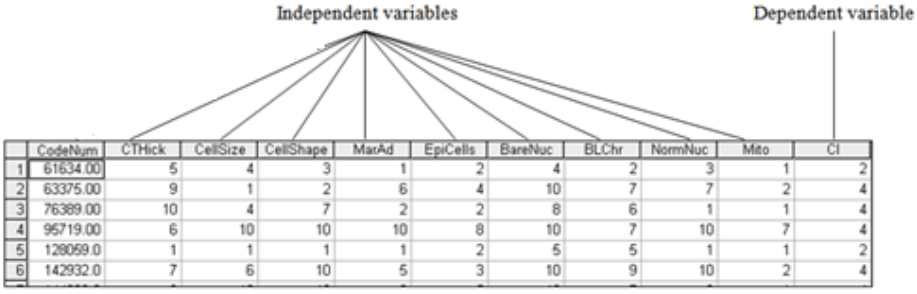
$$logit(Y) = natural\log(odds) = ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \tag{6}$$

Taking the antilog of Eqn. (1) on both sides, an equation to predict the probability to the occurrence of the outcome of interest is as follows:

$$\pi = Probability(Y = outcome\ of\ interest\ |X = x, a\ specific\ value\ of\ X) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta s}} \tag{7}$$

Where π is theprobability for the outcome of interest or "event," α is the intercept, ß is the regression coefficient, and e = 2.71828 is the base forthe system of natural logarithms X can be categorical or continuous, but Y is always categorical.

For the Wisconsin Breast Cancer dataset, there are ten independent variables and one dependent variable for logistic regression as shown in Figure 2. However, the CodeNum is not included for analysis.



	CodeNum	CThick	CellSize	CellShape	MarAd	EpiCells	BareNuc	BLChr	NormNuc	Mito	Cl
1	61634.00	5	4	3	1	2	4	2	3	1	2
2	63375.00	9	1	2	6	4	10	7	7	2	4
3	76389.00	10	4	7	2	2	8	6	1	1	4
4	95719.00	6	10	10	10	8	10	7	10	7	4
5	128059.0	1	1	1	1	2	5	5	1	1	2
6	142932.0	7	6	10	5	3	10	9	10	2	4

Figure 2. Independent and dependent variables of Wisconsin Breast Cancer dataset

Similar approach is applied to German Credit dataset.

4.3.3. Neural network

NN or artificial neural network (ANN) are one of the DM techniques; defined as an information-processing system which is inspired from the function of the human brain whose performance characteristics are somehow in common with biologicalNN [30]. It comprises of a large number of simple processing units, called artificial neurons or nodes. All nodes are interconnected by links known as connections.These nodes are linked together to perform parallel distributed processing in order to solve a desired computational taskby simulating the learning process [3].

There are weights associated with the links that represent the connection strengths between two processing units. These weights determine the behavioron the network. The connection strengths determine the relationship between the input and the output for the network, and in a way represent the knowledge stored on the network. The knowledge is acquired by NN through a process of training during which the connection strengths between the nodes are modified. Once trained, the NN keeps this knowledge, and it can be used for the particular task it was designed to do [29]. Through training, a network understands the relationship of the variables and establishes the weights between the nodes.Once the learning occurs, a new case can be loaded over the network to produce more accurate prediction or classification [31].

NN models can learn from experience, generalize and "see through" noise and distortion, and also abstract essential characteristics in the presence of irrelevant data [32]. NN model is also described as a 'black box' approach which has great capacity in predictive modelling. NN models provide a high degree of robustness and fault tolerance since each processing node has primarily local connections[33]. NNs techniques are also advocated as a replacement for statistical forecasting methods because of its capabilities and performance [33], [34], [33]. However, NNs are very much dependent upon the problem at hand.

The techniques of NNs have been extensively used in pattern recognition, speech recognition and synthesis, medical applications (diagnosis, drug design), fault detection, problem diagnosis, robot control, and computer vision [36], [37]. One major application areas of NNs is forecasting, and the NNs techniques have been used as to solve many forecasting problems ([33], [36], [39], [38].

There are two types of perceptron in NN, namely simple or linear perceptron and MLP. Simple perceptron consists of only two layers; the input layer and output layer. MLP consists of at least three layers input layer, hidden layer and output layer. Figure 3 illustrates the two types of perceptron.

The basic operation of NN involves summing its input weights and the activation function is applied to these layers to yield the output. Generally, there are three types of activation functions used in NN, which are threshold function, Piecewise-linear function and Sigmoid function (Figure 4). Among these sigmoid function is the most commonly used in NN.

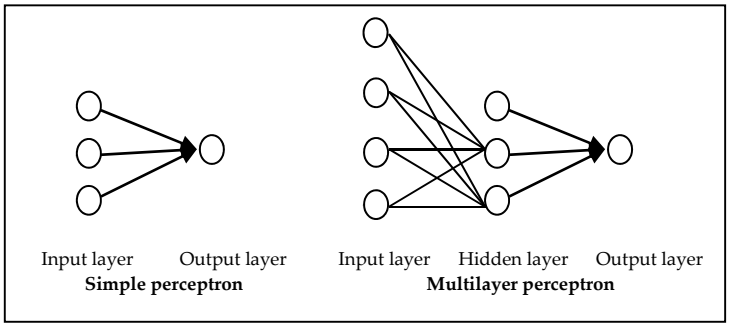


Figure 3. Simple and MLP architecture

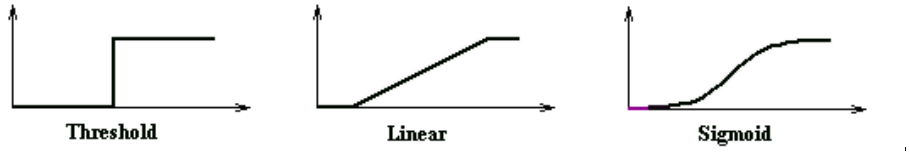


Figure 4. Activation function for BP learning

Multilayer Perceptron (MLP) is one of the most common NN architecture that has been used for diverse applications, particularly in forecasting problems [40]. The MLP network is normally composed of a number of nodes or processing units, and it is organized into a series of two or more layers. The first layer (or the lowest layer) is named as an input layer where it receives the external information while the last layer (or the highest layer) is an output layer where the solution to the problem is obtained. The hidden layer is the intermediate layer in between the input layer and the output layer, and may compose with one or more layers. The training of MLP could be stated as a nonlinear optimization problem. The objective of MLP learning is to find out the best weights that minimize the difference between the input and the output. The most popular training algorithm used in NN is Back propagation (BP), and it has been used in solving many problems in pattern recognition and classification. This algorithm depends upon several parameters such as a number of hidden nodes at the hidden layers 'learning rate, momentum rate, activation function and the number of training to take place. Furthermore, these parameters could change the performance on the learning from bad to good accuracy [23].

There are three stages involved when training the NN using BP algorithm[36]. The first step is the feed forward of the input training pattern, second is calculating the associated error from the output with the input. The last step is the adjustment to the weight. The learning process basically starts with feed forward stage when each of input units receives the input information and sends the information to each of the hidden units at the hidden layer. Each hidden unit computes the activation and sends its signal to each output unit, and applies the activation to form response of the net for given input pattern. The accuracy of NN is provided by a confusion matrix. In a confusion matrix, the information about actual values and the predictive values are illustrated in Table 4. Each row of the matrix represents the

actual accounts of a class of target for the actual data, while each column represents the predictive value from the actual data. To obtain the accuracy of NN, the summation of the correct instance will be divided by the summation for all instances. The accuracy of NN is calculated using Eqn. (7).

$$Percentage\ of\ Correct = \left(\frac{Total\ of\ correctly\ predicted\ pattern}{Total\ no.of\ pattern}\right) * 100\% \quad (7)$$

Based on Table 4, the Percentage of correct is calculated as:

Percentage of Correct = ((48 + 39) / (48 + 2 + 11 + 39)) * 100%

		Predicted	
	Actual	Benign	Malignant
Wrongly predicted	Benign	48	2
	Malignant	11	39

Correctly predicted

Correctly predicted

Table 4. Confusion matrix

Experiments are conducted to obtain a set of training parameters that gives the optimum accuracy for both data sets. Figure.5 shows general architecture of NN for the Wisconsin Breast Cancer data set. Note that the ID number is not including in the architecture.

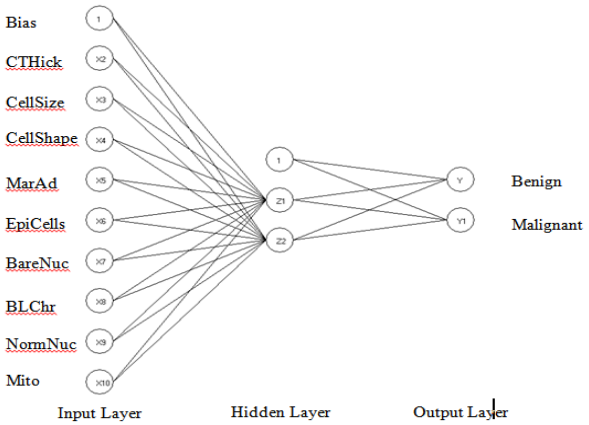


Figure 5. Neural Network architecture for Wisconsin Breast Cancer

Similar architecture can be drawn for German Credit dataset; however, the number of hidden units and output units will be different from the Wisconsin Breast Cancer.

4.4. Investigation and comparison

The accuracy results obtained from previous experiments are compared and investigated further. Two data sets are considered for this study, the Logistic regression and Neural Network. Logistic regression is a statistical regression model for binary dependent variables [24], which is simpler in terms of computation during training while still giving a good classification performance [27]. Figure 6 shows the general steps involve in performing logistic regression and NN experiments using different data representations in this study.

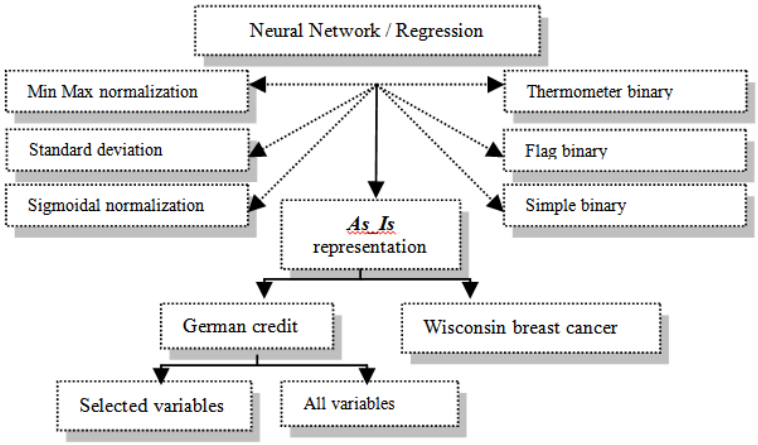


Figure 6. Illustration of Data Representation for NN/ Regression analysis experiments

5. Results

Investigating the prediction performance on different data sets involves many uncertainties for a different data type. In the task of prediction, one particular predictive model might give the best result for one data set but gives the poor results in another data set although these two data sets contain the same data with different representations [14],[15],[16], [17].

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

Based on the results exhibited in Table 5, 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 the highest percentage of accuracy for *Thermometre*and *Flag* representation (100%) followed by *Simple Binary* representation.

Referring to the result shown in Figure 7, similar observation has been noted for German Credit data set when **all variables** are considered for the experiments. *As_Is* representation obtained the highest percentage of accuracy (79%) for NN model. For regression analysis,

Thermometer 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%).

	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%

Table 5. Percentage of accuracy for Wisconsin Breast Cancer Dataset

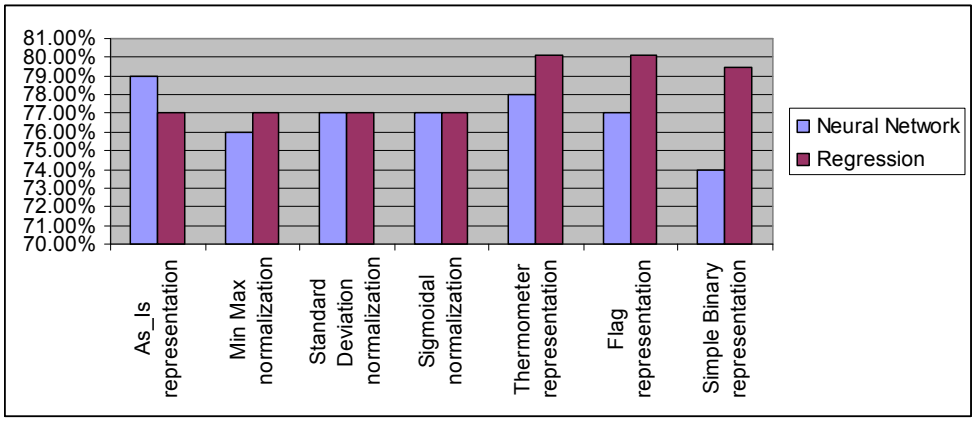


Figure 7. German Credit All Variables accuracy for Neural Network and Regression

When **selected variables** of German Credit data set 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. 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.

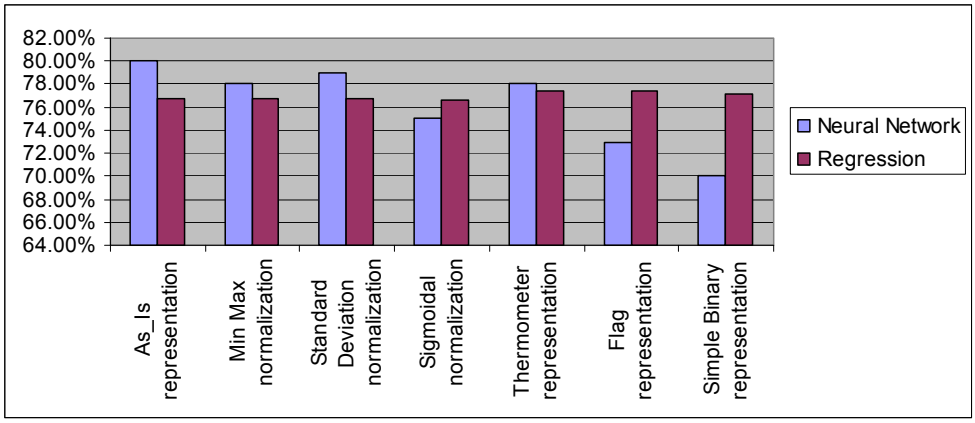


Figure 8. German Credit Selected Variables accuracy for Neural Network and Regression

For brevity, Table 6 exhibits NN parameters that produce the highest percentage of accuracy for Wisconsin Breast Cancer, and German Credit data set using all variables as well as selected variables in the experiments.

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
Number of epoch	100	100	100

Table 6. The summary of NN experimental results using *As_Is* representation

The logistic regression and correlation results for Wisconsin Breast Cancer data set are exhibited in Table 7. Note that based on Wald Statistics, variables such as *CellSize*, *Cellshape, EpiCells, NormNuc* and *Mito* are not significant in 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 7.

For German Credit data set, NN obtained the highest percentage of accuracy when all variables are considered for the training (see Table 6). The appropriate parameters for this data set are also listed in the same table. The summary of logistic regression results is shown in Table 8. All shaded variables displayed in Table 8 are significant independent variables for determining whether a credit application is successful or not.

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		

Table 7. List of variables included in logistic regression of Wisconsin breast cancer

Note also that variable *age* is not significant to German Credit target. However, its correlation with the target is significant. Therefore, these are variable included in logistic regression equation that represents German credit application.

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		

Table 8. List of variables included in logistic regression of German Credit dataset

6. Conclusion and future research

In this study, the effect of different data representations on the performance of NN and regression was investigated on different data sets that have a binary or boolean class target. The results indicate that different data representation produces a 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 9). NN has shown consistent performance for both data sets. Further inspection of the results exhibited in Table 6 also indicates that for German Credit data set, 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 data sets 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 model's developer to present data correctly for binary or Boolean target in the prediction task.

	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

Table 9. Summary of NN and regression analysis of German Credit dataset

Author details

Fadzilah Siraj, Ehab A. Omer A. Omer and Md. Rajib Hasan
School of Computing, College of Arts and Sciences, University Utara Malaysia, Sintok, Kedah, Malaysia

7. References

[1] C. Li, and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Transactions on Knowledgeand Data Engineering,* vol. 14, no. 4, pp. 673-690, 2002.

[2] A. Ahmad, and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data &KnowledgeEngineering,* vol. 63, no. 2, pp. 503-527, 2007.

[3] Li Kan, LuiYushu, "Agent Based Data Mining Framework for the High Dimensional Environment," Journal of Beijing institute of technology, vol. 14, pp. 113-116, Feb 2004.

[4] Pan Ding, ShenJunyi, "Incorporating Domain Knowledge into Data Mining Process: An Ontology Based Framework," Wuhan University Journal of Natural Sciences, vol. 11, pp. 165-169, Jan. 2006.

[5] XianyiQian; Xianjun Wang; , "A New Study of DSS Based on Neural Network and Data Mining," E-Business and Information System Security, 2009. EBISS '09. International Conference on , vol., no., pp.1-4, 23-24 May 2009 doi: 10.1109/EBISS.2009.5137883

[6] Zhihua, X. (1998) Statistics and Data Mining. *Department of Information System and computer Scince, National University of Singapore.*

[7] Tsantis, L &Castellani, J. (2001) Enhancing Learning Environment Solution-based knowledge Discovery Tools: Forecasting for Self-perpetuating Systematic Reform. JSET Journal 6

[8] Luan, J (2002). Data Mining Application in Higher education. SPSS Executive Report. Retrieved from http://www.crisp-dm.org/CRISPWP.pdf

[9] A. Ahmad, and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data &KnowledgeEngineering,* vol. 63, no. 2, pp. 503-527, 2007.

[10] Fernandez, G., (2003), Data Mining Using SAS Application. CRC press LLC. pp 1-12

[11] Dongsong Zhang; Lina Zhou; , "Discovering golden nuggets: data mining in financial application," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol.34, no.4, pp.513-522, Nov. 2004 doi: 10.1109/TSMCC.2004.829279

[12] Luan, J (2006). Data Mining and Knowledge Management in Higher education Potential Application.*Proceeding of Air Forum, Toronto, Canada*

[13] Siraj, F., &Abdoulha, M. A. (2009). Uncovering hidden information within university's student enrollment data using data mining. Paper presented at the Proceedings - 2009 3rd Asia International Conference on Modelling and Simulation, AMS 2009, 413-418. Retrieved from www.scopus.com

[14] 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.

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

[16] O'Neal, M.R., Engel, B.A., Ess, D.R. &Frankenberger, J.R. (2002). Neural Network prediction of maize yield using alternative data coding algorithms. Biosystems Engineering, 83, 31-45.

[17] 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, Jan 2002.

[18] Jovanovic, N. Milutinovic, V. Obradovic, Z. (2002). Neural Network Applications in Electrical Engineering. Neural Network Applications in Electrical Engineering,, pp. 53-58.

[19] 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.

[20] Shen, A., Tong, R., & Deng, Y. (2007). Application of Classification Models on Credit Card Fraud Detection. International Conference: Service Systems and Service Management, 9-11 June 2007 (pp. 1-4).

[21] 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.

[22] 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.

[23] Nawi, N. M., Ransing, M. R. and Ransing R. S. (2006). An Improved Learning Algorithm Based on The Broyden-Fletcher-Goldfarb-Shanno (BFGS) Method For Back Propagation Neural Networks. Sixth International Conference on Intelligent Systems Design and Applications, October 2006, Vol. 1, pp.152-157.

[24] 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 2007, Vol. 2 (pp. 349-353).

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

[26] O'Connor, M., Marquez, L., Hill, T., & Remus, W. (2002). Neural network models for forecast a review. *IEEE proceedings of the 25th Hawaii International Conference on System Sciences*, 4, pp. 494–498.

[27] Duarte, L. M., Luiz, R. R., Marcos, E. M. P. (2008). The cigarette burden (measured by the number of pack-years smoked) negatively impacts the response rate to platinum-based chemotherapy in lung cancer patients. Lung Cancer, 61(2), 244-254.

[28] 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.

[29] Chiang, L. & Wen, L. (2009). A neural network weight determination model designed uniquely for small data set learning. Expert Systems with Applications.36 (6). 9853-9858

[30] Fausett, L. (1994). Fundamentals Of Neural Networks Architectures, Algorithms, and Applications. Upper Saddle River, New Jersey07458: Prentice Hall.

[31] Lippmann, R.P. (1987). An introduction to Computing with neural neural network.*IEEE Transactions on nets, IEEE ASSP Magazine*, April, pp. 4–22.

[32] Wasserman, P. D. (1989). Neural Computing: Theory and Practice, Van Nostrand-Reinhold, New York.

[33] Marquez, L., Hill, T., O'Connor, M., & Remus, W. (1992). Neural network models for forecast a review. *IEEE proceedings of the 25th Hawaii International Conference on System Sciences*, 4, pp. 494–498.

[34] Siraj, F., &Asman, H. (2002). Predicting Information Technology Competency Using Neural Networks.*Proceedings of the 7th Asia Pacific Decision Sciences Institute Conference*, pp. 249 – 255

[35] Siraj, F. & Mohd Ali, A. (2004). *Web-Based Neuro Fuzzy Classification for Breast Cancer.* Proceedings of the Second International Conference on Artificial Intelligence in Engineering &*Technology, pp. 383 – 387.*

[36] Zhang, D. & Zhou, L. (2004). Discovering Golden Nuggets: Data Mining in Financial Application. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Review, 34(4), 513-522.

[37] Hung, C. & Tsai, C. F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. Expert Systems with Applications, 34, 780-787.

[38] Heravi, S., Osborn, D. R., &Brichernhall, C. R. (2004). Linear versus neural network forecasts for European industrial production series. *International Journal of Forecasting, 20 (3), 435-446*

[39] Lam, M. (2004). Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision Support System, 37 (4),567-581*

[40] De Andre, J., Landajo, M., & Lorca P. (2005). Forecasting business profitability by using classification techniques: A comparative analysis based on a spanish case. *Electric Power Engineering, PowerTech Budapest 99.*