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 within a 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 result 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, Thermometer (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.

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.

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

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.

<table>
<thead>
<tr>
<th>No</th>
<th>Attribute description</th>
<th>Code</th>
<th>Type</th>
<th>Domain</th>
<th>Missing value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Status of existing checking account</td>
<td>SECA</td>
<td>Discrete</td>
<td>1, 2, 3, 4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Duration in month</td>
<td>DurMo</td>
<td>Continuous</td>
<td>4 - 72</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Credit history</td>
<td>CreditH</td>
<td>Discrete</td>
<td>0, 1, 2, 3, 4</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Purpose</td>
<td>Purpose</td>
<td>Discrete</td>
<td>0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10</td>
<td>0</td>
</tr>
</tbody>
</table>
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).

\[ V' = \frac{(v - \text{Min}(v(i)))/(\text{Max}(v(i)) - \text{Min}(v(i)))}{\text{std}_{\text{dev}}(v)} \]

(1)

Where \( V' \) is the new value, \( \text{Min}(v(i)) \) is the minimum value in a particular attribute, \( \text{Max}(v(i)) \) 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 \( \text{Mean}(v) \) and the standard deviation \( \text{Std}_{\text{dev}}(v) \) is calculated from the data set itself. The standard deviation normalization formula [25] is written as in Eqn. (2).

\[ V' = \frac{(v - \text{mean}(v))}{\text{std}_{\text{dev}}(v)} \]

(2)

where

\[ \text{mean}(v) = \frac{\text{Sum}(v)}{n} \]

\[ \text{std}_{\text{dev}}(v) = \sqrt{(\text{sum}(v^2) - (\text{sum}(v)^2/n)/(n-1))} \]

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).

\[ V' = \frac{(v - \text{mean}(v))}{\text{std}_{\text{dev}}(v)} \]

(3)

Where

\[ a = \frac{(v - \text{mean}(v))}{\text{std}_{\text{dev}}(v)} \]
\[
\text{mean}(v) = \frac{\text{Sum}(v)}{n}
\]
\[
\text{std}_{\text{dev}}(v) = \sqrt{\text{sum}(v^2) - (\text{sum}(v)^2/n)/(n-1)}
\]

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.

<table>
<thead>
<tr>
<th>Representations</th>
<th>Wisconsin Breast Cancer</th>
<th>German Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>As_Is representation</td>
<td>5 4 3</td>
<td>1 6 0 4</td>
</tr>
<tr>
<td>Min Max normalization</td>
<td>0.000 4.444 1.333</td>
<td>0.000 0.284 1.000</td>
</tr>
<tr>
<td>Standard Deviation normalization</td>
<td>-1.637 2.068 2.836</td>
<td>-1.254 -1.236 1.342</td>
</tr>
<tr>
<td>Sigmoidal normalization</td>
<td>-2.675 -1.102 1.49</td>
<td>-1.362 0.813 -0.976</td>
</tr>
<tr>
<td>Thermometer representation</td>
<td>1 1 1 1 0 0 0 0 0 1 1 1 1 1 1</td>
<td>0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Flag representation</td>
<td>0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0</td>
<td>0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0</td>
</tr>
<tr>
<td>Simple Binary representation</td>
<td>0 1 0 1 0 0 1 0 0 0 1 0 0 0 0</td>
<td>0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0</td>
</tr>
</tbody>
</table>

**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 used whose equation can be written as Eqn. (4) [26].
\[ \theta = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)} \]  

(4)

Where \( \alpha \) = the constant from the equation and \( \beta \) = the coefficient of the predictor variables. Alternatively, the logistic regression equation can be written as Eqn. (5).

\[ \log \left[ \frac{\theta(x)}{1 - \theta(x)} \right] = \alpha + (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \]  

(5)

An odd'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:

\[ \text{logit}(Y) = \text{natural log(odds)} = \ln \left( \frac{\pi}{1-\pi} \right) = \alpha + \beta X \]  

(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:

\[ \pi = \text{Probability}(Y = \text{outcome of interest} \mid X = x, \text{a specific value of } X) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \]  

(7)

Where \( \pi \) is the probability for the outcome of interest or “event,” \( \alpha \) is the intercept, \( \beta \) is the regression coefficient, and \( e = 2.71828 \) is the base for the 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.

![Figure 2. Independent and dependent variables of Wisconsin Breast Cancer dataset](image-url)
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 task by 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 behavior on 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.
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).

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

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

$$\text{Percentage of Correct} = \frac{(48 + 39)}{(48 + 2 + 11 + 39)} \times 100\%$$

**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%).

<table>
<thead>
<tr>
<th>Representation</th>
<th>Neural Network</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>As_Is representation</td>
<td>96.24%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Min Max normalization</td>
<td>96.42%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>96.42%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Sigmoidal normalization</td>
<td>96.60%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Thermometer representation</td>
<td>97.14%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Flag representation</td>
<td>97.67%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Simple Binary representation</td>
<td>97.14%</td>
<td>98.57%</td>
</tr>
</tbody>
</table>

**Table 5.** Percentage of accuracy for Wisconsin Breast Cancer Dataset

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.
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.

<table>
<thead>
<tr>
<th>Neural Network</th>
<th>Wisconsin Breast Cancer</th>
<th>German credit using all variables</th>
<th>German credit using selected variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of Accuracy</td>
<td>98.57%</td>
<td>80.00%</td>
<td>79.00%</td>
</tr>
<tr>
<td>Input units</td>
<td>9</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>Hidden units</td>
<td>2</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.1</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Momentum rate</td>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Number of epoch</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

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.
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.

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.

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

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


