Effects of the neural network s-Sigmoid function on KDD in the presence of imprecise data

John F. Kros\textsuperscript{a,*}, Mike Lin\textsuperscript{b}, Marvin L. Brown\textsuperscript{c}

\textsuperscript{a}Department of Decision Sciences, East Carolina University, 3206B Bate Building, School of Business, Greenville, NC 27858-4353, USA
\textsuperscript{b}Department of Computer and Information Science, School of Business, Cleveland State University, Cleveland, OH 44114, USA
\textsuperscript{c}Department of Computer Information Systems, College of Business, Grambling State University, Grambling, LA 71245, USA

Available online 1 February 2005

Abstract

This research explores a specific step in the Knowledge Discovery of Databases (KDD) process, Data Mining. The actual data mining process deals significantly with prediction, estimation, classification, pattern recognition and the development of association rules. Therefore, this analysis depends heavily on the accuracy of the database and on the chosen sample data to be used for model training and testing. Data mining is based upon searching the concatenation of multiple databases that usually contain some amount of missing data along with a variable percentage of inaccurate data, pollution, outliers and noise. The issue of missing data must be addressed as ignoring this problem can introduce bias into the models being evaluated and lead to inaccurate data mining conclusions. The objective of this research is to address the Effects of the Neural Network s-Sigmoid Function on KDD in the Presence of Imprecise Data using a three factor ANOVA test and Tukey’s Honestly Significant Difference statistics.

This research investigates the accuracy and impact of Data Imputation Methodologies that are employed when a specific Data Mining algorithm is utilized within a Knowledge Discovery In Databases (KDD) process. This study will employ certain Knowledge Discovery processes that are widely accepted in both the academic and commercial worlds. This work includes testing the impact of missing data on the Neural Network s-Sigmoid Transfer Function type in the Data Mining process, by experimenting with three factors: imputation method, data set size, and level of data missingness.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: KDD; Data mining; Neural networks; Imputation

* Corresponding author. Tel.: +1 252 328 6364.
E-mail address: krosj@mail.ecu.edu (J.F. Kros).
1. Introduction

The purpose of this research is to investigate the accuracy and impact of Data Imputation Methodologies that are employed when a specific Data Mining algorithm is utilized within a Knowledge Discovery in Databases (KDD) process. This study employs certain Knowledge Discovery processes that are widely accepted in both the academic and commercial worlds. A Knowledge Discovery model is developed using secondary data containing known correct values. Tests are conducted on the secondary data both before and after storing data instances with known results and then identifying imprecise data values. Imprecise data values are deemed “missing”, along with several percentages of randomly selected missing data values. Differences between the results of Knowledge Discovery with known results and new results utilizing the same Knowledge Discovery process with different levels of data missingness are then measured. Further tests utilizing two imputation methods to deal with the missing data are conducted. The Knowledge Discovery process is conducted following the employment of each imputation method and the results recorded. These results will be compared to the original correct data instances, the instances following data imputation, as well as with each of the other data imputation methods employed. The differences between imputation methodologies and their effectiveness are also measured in this study.

The research conducted draws upon the existing body of knowledge and extends it in several directions. First, it will explore a specific step in the Knowledge Discovery process, Data Mining. Further, the investigation focuses on those knowledge discovery processes that utilize Neural Networks as the data-mining algorithm. Next, a specific type of Neural Network that utilizes the s-Sigmoid as its Transfer Function is tested. Finally, the base theory of data imputation on imprecise and missing data is extended by conducting tests on the s-Sigmoid transfer function by utilizing various accepted data imputation methodologies.

2. Neural network activation and transfer functions

An obstacle to the selection of a neural network as the data-mining algorithm to utilize in the KDD process is the internal configuration of the neural network itself. A “hidden layer” exists as a process between the input nodes and the output nodes within a neural network. By training a neural network internal weights are developed within the hidden layer. An activation function combines the inputs to the network into a single output using these weights.

The activation function of the basic unit of a neural network has two sub-functions: the combination function and the transfer function. The combination function commonly uses the “standard weighted sum” (the summation of the input attribute values multiplied by the weights that have been assigned to those attributes) to calculate a value to be passed on to the transfer function. The transfer function applies either a linear or non-linear transformation to the value passed to it by the combination function. The “hidden layer” then employs this transfer function in moving data to the output nodes.

2.1. Types of transfer functions

Various types of transfer functions exist, including Linear, s-Sigmoid, and Gaussian. Almost all types of neural nets use a transfer function. This function translates the activation of a node into the node’s output. The transfer function accepts inputs (the sum of the weighted outputs from an earlier layer) and
produces a standardized output. The most commonly used type of transfer function is one that limits the size of the output; it “squashes” the activation function. A popular example of such a “squashing” function is the s-Sigmoid transfer function [1].

2.2. The s-Sigmoid function

The s-Sigmoid function with an underlying equation of

\[ S(x) = \frac{1}{1 + e^{-x}} \]

takes on the shape as shown in Fig. 1. This function maps any value to a new value between 0.0 and 1.0. There are two advantages of this. First, a squashing function ensures that the values in a network remain within a reasonable range. Even for very high activation values, the node outputs can never grow larger than 1.0. Second, on a more theoretical note, it turns out that a non-linear transfer function is necessary to enable feed-forward networks to approximate any arbitrary function, that is, to map any input to a desired output.

In addition to these two advantages, the authors chose the s-Sigmoid function as it has a lengthy history in neural network applications [2–4]. The s-Sigmoid function appeared early in the neural network research. Its use was justified as in that it is a continuous function that is a solution to a particular differential equation. In later research, the function was used as one of the many possible smooth, monotonic “squashing” functions defined on a bounded interval. As of late, the s-Sigmoid function has received more emphasis as the function of choice for Neural Networks and further study due to its probabilistic properties related to classification. This emphasis and research has strengthened the link between neural networks and classical statistics [1]. The next section begins with a discussion of missing data theory and continues by defining and categorizing missing data types.
3. Base theory of missing data

The analysis of missing data is comparatively recent. With the advent of the mainframe computer in the 1960s, businesses were capable of collecting large amounts of data on their customer bases. As large amounts of data were collected, the issue of missing data began to appear. Afifi and Elashoff [5], Hartley and Hocking [6], Orchard and Woodbury [7], and Dempster et al. [8] provide early seminal works on the analysis of missing data.

Little [9] discussed models for non-response, while Little and Rubin [10] considered statistical analysis with missing data. These papers sparked numerous works in the area of missing data, which include: Dingle and Kenward [11], Graham et al. [12], Little [13], Little and Rubin [14], and Howell [15]. The problem of missing data is also very complex. Little and Rubin [10] and Schafer [16] provide conventional statistical methods for analyzing missing data. However, the statistical literature on missing data deals almost exclusively with the training of models rather than prediction [17].

Before an analyst can began to address the issue of missing data, it is important to understand the types of missing data that may be encountered. According to Little and Rubin [10], there are several categories of missing data:

- Data Missing At Random
- Data Missing Completely At Random
- Non-Ignorable Missing Data
- Outliers Treated As Missing Data

Each of these categories is discussed next.

3.1. Data Missing At Random (MAR)

It is obvious that cases containing incomplete data must be treated differently than cases with complete data. Rubin [18] defined missing data as MAR “when given the variables $X$ and $Y$, the probability of response depends on $X$ but not on $Y$”. Simply stated, some correlation exists between an attribute containing missing values and some other attribute(s) within the data structure. The pattern of the missing data may be traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing. Examples of MAR could include the following:

- if Democratic party identifiers are more likely to refuse the vote choice question, then the process is MAR so long as party identification is a question to which at least some people responded
- data on attitudinal measures may be dependent on one’s educational level, but within each educational level, the data are missing randomly.

MAR data differs from other missing data such as data missing completely at random or MCAR. MCAR data is discussed next.

3.2. Data Missing Completely At Random (MCAR)

MCAR data exhibits a higher level of randomness than does MAR. Rubin [18] and Kim [19] classified data as MCAR when “the probability of response [indicates that] independence exists between $X$ and $Y$".
In other words, $X$ is independent of $Y$, or probabilistically, $P(X|Y) = P(X)$. Missing values in processes that are MCAR cannot be predicted any better with information in $Y$, observed or not. Examples of MCAR could be as follows:

- respondents deciding whether to answer questions on the basis of coin flips

or

- a research associate shuffling raw data sheets and arbitrarily discarding some of the sheets.

The observed values of $Y$ are truly a random sample for all values of $Y$, and no other factors included in the study may bias the observed values of $Y$. It should be noted that MCAR rarely applies to survey-based data.

In practice, the MCAR assumption is seldom met. Most missing data methods are applied upon the assumption of MAR, although that is not always tenable. And in correspondence to Kim [19], “Non-Ignorable missing data is the hardest condition to deal with, but unfortunately, the most likely to occur as well.” In contrast to the MAR situation where data missingness is explainable by other measured variables in a study; non-ignorable missing data arise due to the data missingness pattern being explainable—and only explainable—by the very variable(s) on which the data are missing [20].

3.3. Non-Ignorable Missing Data

Given two variables $X$ and $Y$, data is deemed Non-Ignorable when the probability of response depends on variable $X$ and possibly on $Y$. Examples of Non-Ignorable Missing Data could include the following:

- if the likelihood of an individual providing his or her weight varied according to the weight values in each age category, the missing data is non-ignorable [19]

or

- if a participant in a weight-loss study does not attend a weigh-in due to concerns about his weight loss, his data are missing due to nonignorable factors.

In comparison to the MAR situation where data missingness is explainable by other variables, nonignorable missing data arise due to the data missingness pattern being explainable—and only explainable—by the very variable(s) on which the data are missing. Thus, the pattern of missing data is non-random and is not predictable from other variables in the database.

3.4. Outliers Treated As Missing Data

Data whose values fall outside of pre-defined ranges may skew test results. Many times it is necessary to classify these outliers as Missing Data. Pre-testing and calculating threshold boundaries are also necessary in the pre-processing of data in order to identify those values, which are to be classified as missing. Sampled values significantly different or inconsistent with the remaining set of data are called outliers. Outliers may be due to recording or measurement error.
Table 1  
Profit/(loss) values (\$)

<table>
<thead>
<tr>
<th>Period</th>
<th>Profit/(loss)</th>
<th>Period</th>
<th>Profit/(loss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>300</td>
<td>11</td>
<td>560</td>
</tr>
<tr>
<td>2</td>
<td>230</td>
<td>12</td>
<td>390</td>
</tr>
<tr>
<td>3</td>
<td>1560</td>
<td>13</td>
<td>520</td>
</tr>
<tr>
<td>4</td>
<td>410</td>
<td>14</td>
<td>220</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>15</td>
<td>280</td>
</tr>
<tr>
<td>6</td>
<td>1390</td>
<td>16</td>
<td>310</td>
</tr>
<tr>
<td>7</td>
<td>550</td>
<td>17</td>
<td>200</td>
</tr>
<tr>
<td>8</td>
<td>(670)</td>
<td>18</td>
<td>370</td>
</tr>
<tr>
<td>9</td>
<td>(110)</td>
<td>19</td>
<td>550</td>
</tr>
<tr>
<td>10</td>
<td>450</td>
<td>20</td>
<td>370</td>
</tr>
</tbody>
</table>

The confidence interval threshold approach can be used to detect outliers for one-dimensional samples. The general method is to identify all data points outside of a confidence interval threshold. A confidence interval threshold can be defined as follows: \( \text{mean}(\mu) \pm k \times \text{standard-deviation}(\sigma) \)—customarily \( k \) is defined on the interval \( 1.0 \leq k \leq 3.0 \) and for this application \( k = 2.0 \).

To illustrate the confidence interval threshold approach the following example is given. Table 1 contains a set of manually inputted profit/(loss) values, in dollars, over twenty periods for a small business. The business owner feels that some of the numbers may be outliers.

Statistics for the data set are: \( \mu = 385 \) and \( \sigma = 468.62 \). Therefore, if the threshold value for the data \( (\mu) \pm 2\sigma \), all data out of \([-552.24, 1322.24]\) are potential outliers and could be treated as missing data.

For even greater precision, various levels of data missingness for specific attributes can be calculated for their volume, magnitude, percentage and overall impact on other attributes in order to determine their effect on overall data mining performance. This “trigger” can be defined in the data mining procedure to identify which test samples may be polluted with an over abundance of missing data, thus skewing the sample taken.

Once missing data has been defined and categorized, an appropriate method can be chosen for the treatment of that missing data. The next section discusses commonly used methods of addressing missing data.

4. Commonly used methods of addressing missing data

Several methods have been developed for the treatment of missing data. The simplest of these methods can be broken down into the following categories:

- Use of Complete Data Only
- Deleting Selected Cases or Variables
- Data Imputation.
These categories are based on the randomness of the missing data, and how the missing data is estimated and used for replacement. While the use of complete data only is a common approach, the cost of lost data and information when cases containing missing value are simply deleted can be huge. Another alternative is deletion of selected cases or variables.

This research employs Data Imputation for addressing missing data. Therefore, use of complete data only and deletion of selected cases or variables are briefly discussed next and references provided for the reader to pursue. Following this discussion data imputation methods are elaborated upon in more detail.

4.1. Use of Complete Data Only

Use of complete data only is generally referred to as the “complete case approach” and is readily available in all statistical analysis packages. When the relationships within a data set are strong enough to not be significantly affected by missing data, large sample sizes may allow for the deletion of a pre-determined percentage of cases. This method tends to be much more successful when missing data is classified as MCAR. Overall, this method is best suited to situations where the amount of missing data is small (see [10,21] for further discussion of this method).

4.2. Delete Selected Cases or Variables

The simple deletion of data that contains missing values may be utilized when a non-random pattern of missing data is present. However, it may be ill advised to eliminate ALL of the samples taken from a test. Nie et al. [22] examined this strategy, but no firm guidelines exist for the deletion of offending cases.

4.3. Data Imputation methods

A number of researchers have begun to discuss specific imputation methods. Little [17], Rubin [23], Schafer [16,24], and Schafer and Olsen [25] have all published articles regarding imputation methodologies. In addition, a number of case studies have been published regarding the use of imputation in medicine [26,27] and in survey research [28]. Hot deck imputation and nearest neighbor methods are very popular in practice, but have received sparse coverage in the theoretical literature (for practical application papers see [29–32]).

Imputation methods are procedures resulting in the replacement of missing values by attributing them to other available data. A definition of imputation is as follows: “the process of estimating missing data of an observation based on valid values of other variables” [33]. Commonly used imputation methods include:

- Case Deletion
- Case Substitution
- Mean Substitution
- Hot Deck Imputation
- Cold Deck Imputation
- Regression Imputation
- Multiple Imputation.
This research investigates two methods, case deletion and mean substitution. These two methods were chosen mainly as they are very common in the literature and their ease and expediency of application [10]. As Dempster and Rubin [34] commented, “imputation is a general and flexible method for handling missing-data problems, but is not without its pitfalls. Caution should be used when employing imputation methods as they can generate substantial biases between real and imputed data.” Case deletion and mean substitution do have pitfalls, but those pitfalls can be overcome with carefully chosen classes for means and totals [35]. Nonetheless, imputation methods tend to be a popular method for addressing the issue of missing data. As this research investigates case deletion and mean substitution imputation methods, only these two methods are detailed (see [16,17,23–34] for in depth discussions on the other imputation methods listed).

4.3.1. Case deletion

The simple deletion of data that contains missing values may be utilized when a nonrandom pattern of missing data is present. Large sample sizes permit the deletion of a predetermined percentage of cases, and/or when the relationships within the data set are strong enough to not be significantly affected by missing data. Case deletion is not recommended for small sample sizes or when the user knows strong relationships within the data exist. This study contains a very large sample size and overall the data do not exhibit strong relationships. Therefore, in this study, case deletion can be employed as an acceptable imputation method.

4.3.2. Mean substitution

This type of imputation is accomplished by estimating missing values by using the mean of the recorded or available values. This is a popular imputation method for replacing missing data. However, it is important to calculate the mean only from responses that have been proven to be valid and are chosen from a population that has been verified as having a normal distribution. If the data distribution is skewed, the median of the available data can be used as a substitute.

Mean imputation is a widely used method for dealing with missing data. The main advantage is its ease of implementation and ability to provide all cases with complete information. Mean imputation can also be regarded as a special type of regression imputation. For data where the relationships between variables are sufficiently established, regression imputation is a very good method of imputing values for missing data.

5. Using neural networks in the KDD process with imprecise data

Neural Networks have been found to perform very well on classification tasks, and it has also been discovered that they are both reliable and effective when applied to applications involving prediction, classification, and clustering [36]. This work focuses on those KDD processes that utilize Neural Networks as the data-mining algorithm. The next section briefly describes neural networks and explains their application to the KDD process.

5.1. Neural networks

A neural network is a system loosely modeled after the human brain in an attempt to simulate the multiple layers of simple processing elements called neurons. Each neuron is linked to specific neighbors
with varying coefficients of connectivity that represent the strength of these connections. Learning is accomplished by adjusting these strengths (weights) by to cause the overall network to produce the best possible resulting output.

Neural networks can be used to build explanatory models by exploring datasets in search of relevant variables or groups of variables. Haykin [37], Masters [38], and Ripley [39] provide information on neural networks. Warner and Misra [40] provide a good overview of neural networks used as statistical tools. The neural net literature of late also contains some good papers covering prediction with missing data (see [41,42]).

5.2. The impact of missing data on neural networks

Neural Networks have been found to be both reliable and effective when applied to applications involving prediction, classification, and clustering [36]. Missing data has a similar impact on neural networks as it does on other types of classification algorithms, such as k-Nearest Neighbor. These similarities include variance understatement, distribution distortion, and correlation depression.

When neural networks are used in the presence of missing data, it may be necessary to “train” the initial network with missing data if the data to be tested and evaluated later is itself going to contain missing data. By training the network with cases containing complete data only, the internal weights developed with this type of training set cannot be accurately applied to a test set containing missing values later.

Missing data actually impacts the internal execution of the neural network in several ways. Since the internal weights used to calculate outputs are created and distributed within the network without providing the insight as to how a solution is created, missing or imprecise data can distort the weights that are assigned as the associations between nodes in a manner unknown to the research analyst.

The hidden layer is where the actual weights are developed for the network. The activation function combines the inputs to the network into a single output [43]. The output remains low until the combined inputs reach a predetermined threshold, and small changes to the input can have a dramatic effect on the output [44]. The activation function can also be very sensitive to missing data.

The activation function of the basic unit of a neural network has two sub-functions: the combination function and the transfer function. The combination function commonly uses the “standard weighted sum” (the summation of the input attribute values multiplied by the weights that have been assigned to those attributes) to calculate a value to be passed on to the transfer function. The transfer function applies either a linear or non-linear transformation to the value passed to it by the combination function. Even though a linear function used in a feed-forward neural network is simply performing a linear regression, missing values can distort the coefficients in the regression equation and therefore pass on invalid values as output [2].

6. Scope and methodology

The scope of this study is to perform an experimental design to explain the impact of imprecise and missing data on the KDD process. More precisely, the Data Mining phase of the Knowledge Discovery process is tested utilizing Neural Network software that employs the s-Sigmoid as its Transfer Function. Data sets containing various frequencies of cases are analyzed, altered and reevaluated by this research.
Secondary data is used for this experiment utilizing five data sets containing different frequencies of cases, obtained from various sources where the value of a dependent variable within each data set has previously been determined and verified. Methods of data imputation are employed within a data mining model to attempt to identify how various levels of data missingness within data sets of varying frequencies of cases may impact a Data Mining study. The data sets contained case levels of $N = 500, 1000, 3500, 5000$ and $7000$.

The Intelligent Data Analyzer (iDA) software product was selected to perform the data mining session [45]. A backpropagation Neural Network architecture employing an s-Sigmoid Transfer Function was chosen for study. The network is trained using the data set(s) with known values for the dependent variable(s). The Root Mean Square (RMS) error (comparison between desired output and computed output) was selected as the metric to be evaluated in determining the performance of each Neural Network model.

Each data set was initially mined with no missing data, without altering the standard parameters necessary for data mining utilizing a neural network (learning rate, number of input nodes, number of hidden layers, number of epochs) and obtaining an RMS value. Each data set was then injected with a particular level of data missingness (e.g., 10%, 20%, 30%, 40%, 50%, 60%, 70%) and mined again, using the same standard neural network parameters. Data imputation is performed on the missing values using the Case Deletion and Mean Imputation methods. The RMS results are then analyzed using a three factor ANOVA and Tukey’s Honestly Significant Difference (HSD) statistic to determine if original data set size, level of data missingness, and/or data imputation method are significant factors.

### 7. Results

The results of RMS values obtained from all data mining sessions are displayed in Table 2. A three factor ANOVA test was then conducted at the 0.05 significance level. The three factors include original data set size, level of data missingness and imputation method. The results of the ANOVA are displayed in Table 3.

The ANOVA results indicate that two of the three factors tested are significant at the 0.05 significance level. Original data set size and imputation method were found to be significant, while the percentage level of data missingness was found not to be significant. Therefore, we fail to accept the null hypothesis and conclude that original data set size does have an effect on the RMS values. In general, one could suggest that as the size of the original data set becomes larger there is more tolerance by the data mining algorithm for data that is missing.

Further, we fail to accept the null hypothesis regarding the imputation method employed and conclude that the imputation method employed also has an effect on the RMS values. However, for this study we do accept the null hypothesis that the level of data missingness within a data set is not significant and therefore does not have an effect on the RMS value.

Tukey’s HSD test was then performed upon the two factors (original data set size and imputation method) proven to be significant in the ANOVA. The test was performed at the .05 significance level and indicated the following results displayed in Table 4.

The results conclude that the data set containing 500 cases was found to be significantly different from all data sets (1000, 3500, 5000 and 7000 cases). In addition, the data set containing 1000 cases was found to be significantly different from all other data sets used in this study.
### Table 2
RMS statistics for data sets

<table>
<thead>
<tr>
<th>Original data set size</th>
<th>Imputation method</th>
<th>% missing data 10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>No Method</td>
<td>.449</td>
<td>.454</td>
<td>.454</td>
<td>.454</td>
<td>.454</td>
<td>.445</td>
<td>.453</td>
</tr>
<tr>
<td></td>
<td>Mean Sub.</td>
<td>.255</td>
<td>.200</td>
<td>.250</td>
<td>.290</td>
<td>.199</td>
<td>.240</td>
<td>.270</td>
</tr>
<tr>
<td></td>
<td>Case Del.</td>
<td>.249</td>
<td>.230</td>
<td>.210</td>
<td>.269</td>
<td>.270</td>
<td>.230</td>
<td>.200</td>
</tr>
<tr>
<td>1000</td>
<td>No Method</td>
<td>.144</td>
<td>.144</td>
<td>.144</td>
<td>.144</td>
<td>.144</td>
<td>.144</td>
<td>.144</td>
</tr>
<tr>
<td></td>
<td>Mean Sub.</td>
<td>.087</td>
<td>.087</td>
<td>.087</td>
<td>.087</td>
<td>.085</td>
<td>.087</td>
<td>.087</td>
</tr>
<tr>
<td></td>
<td>Case Del.</td>
<td>.083</td>
<td>.094</td>
<td>.098</td>
<td>.097</td>
<td>.096</td>
<td>.101</td>
<td>.107</td>
</tr>
<tr>
<td>3500</td>
<td>No Method</td>
<td>.105</td>
<td>.340</td>
<td>.020</td>
<td>.250</td>
<td>.096</td>
<td>.360</td>
<td>.420</td>
</tr>
<tr>
<td></td>
<td>Mean Sub.</td>
<td>.254</td>
<td>.196</td>
<td>.248</td>
<td>.287</td>
<td>.195</td>
<td>.234</td>
<td>.261</td>
</tr>
<tr>
<td></td>
<td>Case Del.</td>
<td>.243</td>
<td>.225</td>
<td>.208</td>
<td>.261</td>
<td>.267</td>
<td>.228</td>
<td>.198</td>
</tr>
<tr>
<td>5000</td>
<td>No Method</td>
<td>.125</td>
<td>.340</td>
<td>.020</td>
<td>.250</td>
<td>.430</td>
<td>.360</td>
<td>.420</td>
</tr>
<tr>
<td></td>
<td>Mean Sub.</td>
<td>.251</td>
<td>.189</td>
<td>.241</td>
<td>.280</td>
<td>.190</td>
<td>.230</td>
<td>.259</td>
</tr>
<tr>
<td></td>
<td>Case Del.</td>
<td>.240</td>
<td>.209</td>
<td>.200</td>
<td>.260</td>
<td>.259</td>
<td>.221</td>
<td>.193</td>
</tr>
<tr>
<td>7000</td>
<td>No Method</td>
<td>.105</td>
<td>.340</td>
<td>.200</td>
<td>.250</td>
<td>.430</td>
<td>.360</td>
<td>.420</td>
</tr>
<tr>
<td></td>
<td>Mean Sub.</td>
<td>.250</td>
<td>.190</td>
<td>.240</td>
<td>.280</td>
<td>.190</td>
<td>.230</td>
<td>.260</td>
</tr>
<tr>
<td></td>
<td>Case Del.</td>
<td>.240</td>
<td>.221</td>
<td>.200</td>
<td>.261</td>
<td>.260</td>
<td>.218</td>
<td>.191</td>
</tr>
</tbody>
</table>

### Table 3
ANOVA results

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected model</td>
<td>.672</td>
<td>12</td>
<td>.056</td>
<td>13.628</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>5.752</td>
<td>1</td>
<td>5.752</td>
<td>1399.048</td>
<td>.000</td>
</tr>
<tr>
<td>% Data missingness</td>
<td>.034</td>
<td>6</td>
<td>.006</td>
<td>1.391</td>
<td>.227</td>
</tr>
<tr>
<td>Original data set size</td>
<td>.470</td>
<td>4</td>
<td>.117</td>
<td>28.555</td>
<td>.000</td>
</tr>
<tr>
<td>Imputation method</td>
<td>.168</td>
<td>2</td>
<td>.084</td>
<td>20.482</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>.378</td>
<td>92</td>
<td>.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6.803</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected total</td>
<td>1.051</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results also indicate that the data set containing 3500 cases was not found to be significantly different from the data sets containing 5000 and 7000 cases while the data set containing 5000 cases was not significantly different than the data sets containing 7000 cases. These statistics lead the author's to propose for this study that the neural network algorithm and imputation methods tend to perform best when the original data set size is at least larger than 1000 cases and the closer to 3500 cases the better.
Concerning the imputation method employed, Tukey’s HSD test was again performed at the .05 significance level. The results contained in Table 5 indicate that both imputation methods tested in this study, mean substitution and case deletion, were found to be significantly different “from no imputation being performed” on the missing values in the data sets. However, no significant differences were discovered between the two imputation methods themselves.

This affirms what the research literature has concluded in the past that it behooves a user to perform some imputation method but in being the simplest methods of imputation the mean substitution and case deletion methods perform roughly in the same manner. In the conclusion section, the author’s speak to further research opportunities regarding experimentation with various imputation methods, which are more robust in their attribution power.

8. Conclusions

This research took to task the goal of determining how three factors impact a data mining process. The factors of original data set size, level of data missingness and imputation method were selected and a data mining study completed for five independent data sets. Data missingness was injected at various levels and compared to the original data mining results. ANOVA and Tukey’s HSD tests were performed to determine which factors were significant for a successful data mining processes.
From the ANOVA results, our research revealed that original data set size and imputation method to be significant factors, while the level of data missingness was not proven to be significant in this study. It was also discovered, via the Tukey’s HSD analysis, that while there is a significant difference between employing and not employing an imputation method there is no significant difference between the mean substitution and case deletion imputation methods themselves.

In the future, other dimensions such as the data mining algorithm employed and/or varying the parameters within a specific algorithm may be added/and tested for their impact of the Knowledge Discovery process. Data-mining algorithms such as Nearest Neighbor, Decision Trees, Association Rules, and Genetic Algorithms may be employed against the same data sets and a comparison of results conducted. Also, the parameters used within a particular data mining algorithm, such as a neural network, may be adjusted in an attempt to determine which combination of parameter settings perform most effectively when implemented upon various types (size and structure) of data sets. Concerning neural networks specifically, a comparison of different combination and/or transfer functions can be performed to determine the most effective type of function or combination of functions is most desirable for data sets of various dimensions.

The type of imputation method utilized when confronted with missing data is yet another area of research that may be explored. The imputation techniques of Hot Deck, Cold Deck, Regression and Multiple Imputation are just a few methods that may be tested in combination with the aforementioned algorithms for more effective data mining.

References