Effects of Data Imputation Methods on Data Missingness in Data Mining

Marvin L. Brown
Department of Computer Information Systems
College of Business
Grambling State University
Grambling, LA 71245
brownmarv@yahoo.com

Chien-Hua Mike Lin
Department of Computer and Information Science
School of Business
Cleveland State University
Cleveland, OH 44114
lin@cis.csuohio.edu

Abstract—**The purpose of this paper is to study the effectiveness of data imputation methods in dealing with data missingness in the data mining phase of knowledge discovery in Database (KDD). The application of data mining techniques without careful consideration of missing data can result into biased results and skewed conclusions. This research explores the impact of data missingness at various levels in KDD models employing neural networks as the primary data mining algorithm. Four of the most commonly utilized data imputation methods - Case Deletion, Mean Substitution, Regression Imputation, and Multiple Imputation were evalutated using Root Mean Square (RMS) Values, ANOVA Testing, T-tests, and Tukey's Honestly Significant Difference Test to assess the differences of performance levels between various Knowledge Discovery and Neural Network Models, both in the presence and absence of Missing Data.**

Keywords- KDD; Data mining; Data Imputation; Missing Data; Neural Networks Introduction (Heading 1)

I. INTRODUCTION

With the proliferation of low-cost hardware and software for the housing and support of high-volume historical data sets, a focus has emerged in the area of decision support on data warehousing and knowledge discovery in Databases (KDD). The KDD process has multiple steps and is interdisciplinary in nature. Several streams of research have emerged which include data mining, treatment of missing data and data imputation with the development of isolated methods and techniques in each area. However, there has not been research integrating these three areas.

In this paper, we outline the base theory for KDD and give an overview of the related disciplines, namely data mining algorithms, and provide a base theory of missing data and prominent data imputation methods. We execute an experimental design method to explain the impact of imprecise and missing data on the process of Knowledge Discovery in Databases. We test the Data Mining phase of the Knowledge Discovery process utilizing Neural Network software that employs the S-Sigmoid as its Transfer Function. Secondary data was used for the experiment Benefits derived from the research include a better understanding of the impact that data missingness has on select types of Knowledge Discovery, the value of data imputation when missing data is confronted, and the identification of critical levels at which data missingness affects the performance of data imputation.

II. BASE THEORY OF KNOWLEDGE DISCOVERY IN DATABASE

Knowledge Discovery in Database (KDD) is an excellent example of the interdisciplinary nature of the scientific endeavor. It has evolved, and continues to evolve, in and from a wide-range of research fields, such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition, visualization, and high performance computing. There are many successful KDD application. For example, in science, KDD has been successfully implemented in astronomy to process massive amount of image data from sky surveys. In business, it has been successfully applied to marketing, financial and stock market investments, credit card fraud detection, CRM (Customer Relations Management), and anti-terrorism initiatives [1].

Originally described simply as "data mining", Knowledge Discovery in Databases (KDD) is now understood as the application of scientific method to data mining methodology [2]. As the application of Data Mining to large data sets grew more widely practiced, the process was separated into a series of logical procedures. KDD is a process that can be utilized to identify and isolate previously undiscovered patterns and relationships within a large data warehouse containing historical data for an organization. Although various authors and researchers have broken down the KDD process into separate categories (some possessing as many as fourteen steps or procedures), the following are a widely accepted series of procedures identified for use in the Knowledge Discovery process:

1) Data Selection

2) Data Cleansing

3) Data Enrichment

4) Data Coding

5) Data Mining

6) Data Reporting and Interpretation

DOI: 10.5176_2010-2283_1.2.53

A growing number of knowledge engineers have added two additional steps to the aforementioned procedure to aid the efficiency of the entire process. A Goal Identification (Knowledge Requirements) Phase may be added to the front end of the process, and an Action Phase appended to the final step [3].

III. BASE THEORY OF DATA MINING ALGORITHMS

Although many data mining techniques have been developed and successfully implemented, five algorithms have been become widely accepted as standards in commercial data mining software packages, namely, k-Nearest Neighbor, Decision Trees, Association Rules, Neural Networks, and Genetic Algorithms.

A. k-Nearest Neighbor

The k-nearest neighbor concept is thus named due to the fact that each data record is located in a particular cluster or “neighborhood”, with records closest to each other referred to as neighbors. This method is used to predict the behavior of certain data elements [4]..

The k-Nearest Neighbors of an observation are first identified. The k stands for a predetermined constant representing the number of neighbors that contains no missing data and qualifies to be considered in the analysis. According to Witten and Frank [5], it is advised to keep the value for k small, say five or less, so that the impact of any noise present will be kept to a minimum. Since data sets with a large number of attributes or closely related cases may result in a high number of closely related neighborhoods, this algorithm is not recommended for large data sets.

B. Decision Trees

Decision Trees are analytical tools used to discover rules and relationships in data by systematically breaking down and subdividing information from a general view down into a detailed perspective. Contained in this tree structure are branches that represent the outcomes of a particular test, with leaf nodes standing in for resulting classes or class distributions [6]. The greatest benefit of decision trees is their ease of use and understandability [7]. It also scales up very well for large data sets [4].

C. Association Rules

Agrawal, Imielinski, and Swami introduced Association Rules for the first time in their seminal 1993 paper *"Mining Association Rules between Sets of Items in Large Databases"* [8]. A second paper by Agrawal and Srikant introduced the Apriori Algorithm, which is the reference algorithm for the problem of finding Association Rules in a database [9]. Association Rules help to identify how various attribute values are related within a data set. They are developed to predict the value of an attribute (or sets of attributes) in the same data set [10], or to discover correlations or co-occurrences of transactional events [11]. They are useful when performing exploratory analysis, or when searching for interesting relationships that may exist within a data set since Association Rules are often developed specifically to identify these various regularities (patterns) within a data set. Algorithms utilizing association rules have been found to work best with large data sets.

D. Neural Networks

An Artificial Neural Network (ANN) is a system loosely modeled after the human brain and its the multiple layers of simple processors called neurons. Each neuron is linked to specific neighboring neurons with varying coefficients of connectivity representing the strength of these connections. Learning is accomplished by adjusting these strengths (weights) so that the network can produce the best possible output.

Neural networks may be used to build explanatory models by searching datasets for relevant variables or groups of variables. A good overview of neural networks used as statistical tools can be found in [12].

Neural networks have also been found to perform very well on classification tasks. They are both reliable and effective when applied to applications involving prediction, classification, and clustering [4].

E. Genetic Algorithms

Genetic Algorithms integrate combinatorial optimization techniques based on processes that occur in natural biological evolution. The name of this method is derived from its similarity to the process of natural selection. Three natural processes mimicked by software packages using genetic algorithms, include selection, crossover, and mutation. Applying this concept to a knowledge discovery application involves the optimization of a data model along with a genetic method to obtain the fit model [7]. Presently, genetic algorithms are considered some of the most successful of the machine-learning techniques in use.

Genetic algorithms are also a learning-based data mining technique. Holland [13] is credited for being the first to apply genetic algorithms to search and optimization problems.

Each methodology has its own strengths and weaknesses for mining the data. The algorithm should only be selected following an analysis of the type of data to be mined and the relationships within the volume of cases in question. The implementation of an algorithm to an inappropriate environment may result in improper data categorization, incorrect classification of cases, and invalid test conclusions [3]. We chose to test the Neural Network algorithm, due to its proven ability to adapt to changing environments and for the ability of the Artificial Neural Network Model to be refreshed and re-trained with addition and more recent data instances. Various forms of Dependency Analysis, in which columns across rows are evaluated, may be employed to discover dependencies between data attributes.

IV. BASE THEORY OF MISSING DATA

Missing or inconsistent data has been a pervasive problem in data analysis since the origin of data collection. More historical data is being collected today than ever before due to the proliferation of computer software and the increased capacity of storage media. The management of missing data in organizations has recently seen more discussion as firms begin to implement large-scale enterprise resource planning systems [14, 15]. The issue of missing data becomes a serious and pervasive dilemma in the Knowledge Discovery process, in that as more data is collected, the probability of missing data grows. Data dependencies, data sparseness (especially within critical data clusters) and anomaly analysis are directly impacted by the issue of missing data [16].

Before an analyst can begin 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 [16], 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

A. 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". 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 predicted based on other variables in the database rather than the specific variable for which the data is missing [19].

B. Data Completely Missing at Random (MCAR)

MCAR data exhibits a higher level of randomness than does MAR. Rubin [18], and Kim [20] classified data as MCAR when "the probability of response [indicates that] independence exists between X and Y". In other words, 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. In contrast to the MAR situation where data missingness can be explained by other measured variables in a study; this type of missing data is non-ignorable due to the data missingess pattern being explainable --- and *only* explainable --- by the very variable for which the data is missing. Non-ignorable missing data emerges due to the data missingness pattern being explainable --- and only explainable --- by the very variable for which the data is missing [19].

In practice, the MCAR assumption is seldom met. Most missing data methods are applied upon the assumption of MAR, although such practice is not always feasible.

C. 4.3 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. For example, if the likelihood of an individual providing his or her weight varies according to the weight values in each age category, the missing data is non-ignorable. Thus, the pattern of missing data is non-random and cannot be predicted from other variables in the database. Again, as stated by Kim [20], "Non-Ignorable missing data is the hardest condition to deal with, but unfortunately the most likely to occur as well."

D. 4.4 Outliers Treated as Missing Data

Data whose values fall outside of predefined 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 what values are to be classified as missing.

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

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

V. COMMON 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 replaced. While the use of complete data only is a common approach, the cost of lost data and information when cases containing missing values are simply deleted can be tremendous. Another alternative is the deletion of select cases or variables. The third alternative is Data Imputation, the replacement of missing values with other known or derived values.

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" [21]. As Dempster and Rubin [22] have 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." Nonetheless, data imputation methods tend to be a popular method for addressing the issue of missing data [23-27]. In addition, a number of case studies have been published regarding the use of imputation in medicine [28, 29], and in survey research [30].

Some of the most commonly used imputation methods include:

- Case Substitution
- Mean Substitution
- Hot Deck Imputation
- Cold Deck Imputation
- Regression Imputation
- Multiple Imputation
- Model-Based Procedures

A. *Case Substitution*

This method is the most widely-used to replace observations with completely missing data. Cases are simply replaced by non-sampled observations. Only a researcher with complete knowledge of the data (and its history) should have the authority to impute missing data with values from previous research.

B. *Mean Substitution*

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

Mean imputation is a widely accepted method for dealing with missing data. The main advantage is its ease of implementation and ability to provide all cases with complete information.

C. *Cold Deck Imputation*

Cold deck imputation methods select values or employ relationships obtained from sources other than the current database [331-34]. With this method, the end-user substitutes a constant value derived from external sources or from previous research for the missing values. It must be ascertained by the end user that the replacement value used is more valid than any internally derived value. Pennell [35] contains an excellent example of using cold deck imputation to provide values for an ensuing hot deck imputation application.

Unfortunately, feasible values are not always provided using cold deck imputation methods. Many of the same disadvantages that apply to the mean substitution method here. Cold deck imputation methods are rarely used as the sole method of imputation and instead are generally used to provide starting values for hot deck imputation methods.

D. *Hot Deck Imputation*

Generally speaking, hot deck imputation replaces missing values with values drawn from the next most similar case(s). Once the most similar case(s) has been identified, hot deck imputation substitutes the most similar complete case's value for the missing value. The implementation of this imputation method results in the replacement of a missing value with a value selected from an estimated distribution of similar responding units for each missing value. In most instances, the empirical distribution consists of values from responding units. This method is very common in practice, but has received little attention in missing data literature [36]. Advantages of hot deck imputation include conceptual simplicity, maintenance and proper measurement level of variables, and the availability of a complete set of data at the end of the imputation process that can be analyzed accordingly. One of hot deck's disadvantages lies in the difficulty of defining what is "similar", as many different schemes for deciding on what is "similar" may arise.

E. *Regression Imputation*

Regression Analysis is used to predict missing values based on the variable's relationship to other variables in the data set. Simple and/or multiple regression techniques may be utilized to impute missing values. The first step consists of identifying the independent variables and the dependent variables. In turn, the dependent variable is regressed based on the independent variable(s). The resulting regression equation is then used to predict the missing values.

Although regression imputation is useful for simple estimates, it has several inherent disadvantages: it reinforces relationships that already exist within the data; the variance of the distribution is understated; it implies that the variable being estimated has a substantial correlation to other attributes within the data set; the estimated value is not constrained and therefore may fall outside predetermined boundaries for the given variable. An additional adjustment may necessary for over-prediction [37].

F. *Multiple Imputation*

Rubin [18] was the first to propose multiple imputation as a method for dealing with missing data. Multiple imputation combines a number of imputation methods into a single procedure. In most cases, expectation maximization is combined with maximum likelihood estimates and hot deck imputation to provide data for analysis [17]. The method works by generating a maximum likelihood covariance matrix and a mean vector. Statistical uncertainty is introduced into the model and is used to emulate the natural variability of the complete database. Hot deck imputation is then used to fill in missing data points to complete the data set.

G. *Model-Based Procedures*

Model-based procedures incorporate missing data into the analysis. These procedures are characterized as maximum likelihood estimation or missing data inclusion. Dempster, Little, and Rubin [38] give a general approach for computing maximum likelihood estimates from missing data. They call their technique the EM approach. The approach consists of two steps, "E" for the conditional expectation step and "M" for the maximum likelihood step. The first step makes best possible estimates of the missing data and the second step makes estimates of the parameters (e.g., means, variances, or correlations) assuming that the missing data will be replaced. Each of the stages is repeated until the change in the estimated values is negligible. The missing data is then replaced with these estimated values. This approach has become very popular and is included in commercial software packages such as SPSS. SPSS 7.5 includes a missing value module employing the EM procedure for treating missing data.

Cohen and Cohen [39] prescribe inclusion of missing data into the analysis. In general, the missing data is grouped as a subset of the entire data set. This subset of missing data is then analyzed using any standard statistical test. If the missing data occur on a non-metric variable, statistical method such as ANOVA, MANOVA, or discriminant analysis can be used. If the missing data occur on a metric variable in a dependence relationship, regression can be used as the analysis method.

VI. THE RESEARCH MODEL AND METHODOLOGY

Before we proceeded further, a few important notes about the S-Sigmoid Transfer Function.. The s-Sigmoid function with an underlying equation of $S(x) = 1/[1 + e-(x)]$ takes on the shape as shown in Fig. 1. This function maps any value to a new value between 0.0 and 1.0. It also enables 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 [40]. 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 [41]. Kros, Lin, and Brown studied the effects of the of the neural network s-Sigmoid function on KDD in the presence of imprecise data [42].

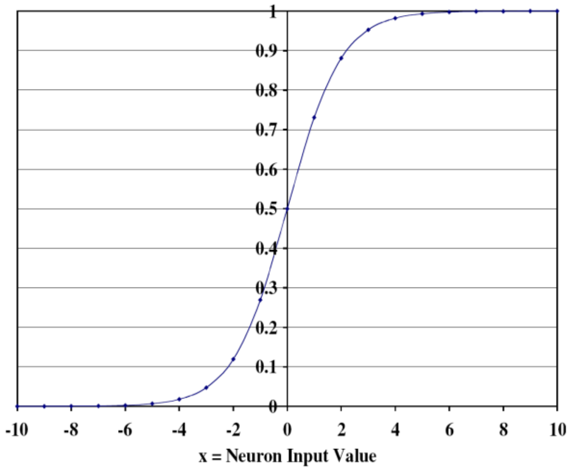


Figure 1. s-Sigmoid Transfer Function

The scope of this study is to perform an experimental design to explain the impact of data imputation methods on data missingness in the process of Knowledge Discovery in Databases. 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. Secondary data will be used for the experiment, utilizing data obtained from a collaborative project of the Center for Disease Control (CDC). The Behavioral Surveillance Branch (BSB) of the CDC has verified the data and made it available to researchers for statistical analysis purposes. This data will be used in conjunction with software for Knowledge Discovery involving a Neural Network employing an S-Sigmoid Transfer Function. The network will be trained using the complete data set(s) with known values for the dependent variable(s). Using a verified SQL algorithm for randomization, various percentages of data elements within the data set(s) will be randomly identified and recorded, and then replaced with missing data. The model will be retested using the same software and parameters. Missing values will then be imputed from several accepted imputation methods. The data set will be retested by the knowledge discovery process following the use of each imputation method. The percentage of data missingness will be increased, retested, with results recorded. Percentages of missing data to be injected into the data sets under investigation will be set at 10%, 20%, 30%, 40%, 50%, 60% and 70% prior to imputation.

This conceptual research model which builds on traditional methodology by extending the constructs of previous research [43, 44] is shown in Figure 2. The innovation proposed in this research model is to simultaneously alter the case volume of the secondary data and the level of data missingness within the data set, and to test the impact of Data Imputation and data Re-Sequencing on model performance.

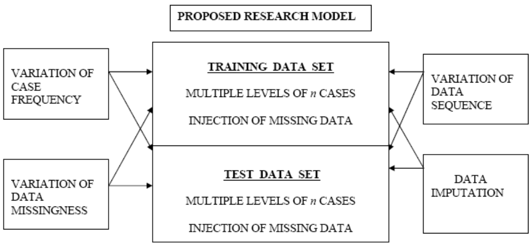

PROPOSED RESEARCH MODEL
VARIATION OF CASE FREQUENCY
TRAINING DATA SET
MULTIPLE LEVELS OF n CASES
INJECTION OF MISSING DATA
VARIATION OF DATA SEQUENCE
VARIATION OF DATA MISSINGNESS
TEST DATA SET
MULTIPLE LEVELS OF n CASES
INJECTION OF MISSING DATA
DATA IMPUTATION

Figure 2. Figure 2 The Conceptual Research Model

Ten hypotheses with progressive complexity were developed to test the impact of data missingness and the effectiveness of major data imputation methods in data mining. They are presented below. Root Mean Square (RMS) values, ANOVA testing, T-tests, and Tukey's Honestly Significant Difference Test were used to evaluate the differences in performance levels between various Knowledge Discovery and Neural Network Models, both in the presence and absence of missing data in the Training and Testing data sets under study.

H Description

H_1 Data set case frequency is not a significant factor in the performance of KDD models that utilize a Neural Network as the Data Mining Algorithm and employ an S-Sigmoid Transfer Function, as measured by the Root Mean Square Value calculated for the model.

H_2 The Level of data missingness in the Training and Testing data sets of a KDD model is not a significant factor in the calculation of the Root Mean Square Value for a KDD model in small models (Case Frequency N=500).

H_3 The Level of data missingness in the Training and Testing data sets of a KDD model is not a significant factor in the calculation of the Root Mean Square Value for KDD models containing various levels of Case Frequency.

H_4 Data Imputation (Mean Substitution and Case Deletion) is not a significant factor in the performance of KDD models containing various volumes of Case Frequency and various levels of missing data.

H_5 There is no difference between the Data Imputation Methods of Mean Substitution and Case Deletion when performed on Missing Values in KDD models containing various levels of Case Frequency and various levels of data missingness.

H_6 The re-sequencing of data cases in the training and test data sets of KDD models containing various volumes of Case Frequency is not a significant factor in the performance of those models.

H_7 The re-sequencing of cases in the training and test data sets of KDD models containing various volumes of Case Frequency and various volumes of data missingness is not significant in the performance of those models.

H_8 There is no significant difference between the Imputation Methods of Mean Substitution and Case Deletion in the performance of KDD models when the level of data missingness in those models is increasingly varied and the Case Frequency is held constant, N=500.

H_9 There is no significant difference between the imputation methods of Regression Imputation, Mean Substitution and Case Deletion when performed on KDD models containing increasing levels of data missingness, and when the Case Frequency of the models are constant, N=1000.

H_{10} There is no difference in the performance of the Imputation Methods: Multiple Imputation, Regression Imputation, Mean Substitution and Case Deletion with various levels of data missingness while the Case Frequency is held constant, e.g., N=1000.

Since our goal is to present a superior approach for effective Knowledge Discovery in the presence of various levels of missing data to practitioners when confronted with the real-world problem of data pollution, with an emphasis on data missingness. The scope of this study will also include the study of 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 with different frequencies of cases. Each dataset contains historical data from various scientific/medical/business disciplines, containing proven data values previously used in the prediction of dependent variables. Therefore, the content validity of the secondary data was met by the selection of data sets already validated in the prediction of a selected dependent variable in each KDD model. Adequate coverage for the investigation of the impact of missing data on the performance of a KDD application using a Neural Network as its Data Mining algorithm was addressed by selecting KDD models containing various levels of case frequency. The data sets contained case levels of *N* = 500, 1000, 3500, 5000 and 7000.

In order to test the impact of data missingness, a Visual Basic module was developed for the random selection of cases and for random injection of specific levels of data missingness into the data sets.

Data imputation methods were employed within a data mining model in an attempt to identify how various levels of data missingness within data sets of varying frequencies of cases might impact a Data Mining study.

The Intelligent Data Analyzer (iDA) software product was selected to perform the data mining session [45]. A back propagation Neural Network architecture employing an s-Sigmoid Transfer Function was chosen for the study. The network was 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, and 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%, and 70%) and mined again, using the same standard neural network parameters. Data imputation was performed on the missing values using Case Deletion, Mean Substitution, Multiple Regression and Multiple Imputation methods. The RMS results were then analyzed using a three factor ANOVA and Tukey's Honestly Significant Difference (HSD) statistic to determine if original data set sizes, level of data missingness, and/or data imputation method were significant factors.

VII. RESULTS AND DISCUSSION

Hypothesis 1 tested the significance of the volume of data set case frequency in our KDD models. The two largest of the original five KDD models with complete data (N=5000, N=7000) were selected for testing. This hypothesis tested the ten new models that were constructed for each of the two original KDD models. Random instances were selected to construct these ten new models with Case Frequencies of N=500, N=1000, N=1500, N=2000, N=2500, N=3000, N=3500, N=4000, N=4500 and N=5000 instances, respectively, for a total of twenty new KDD models.

These twenty new KDD models were trained and tested using the iDA data mining software. 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 and was normalized for scaling purposes. They were calculated employing iDA's Neural Network Algorithm utilizing an S-Sigmoid Transfer Function within the Activation Function. The standard default parameters of the ANN architecture were left intact prior to training and testing each KDD model.

Figure 3 illustrates the resulting Root Mean Square (RMS) values for the twenty KDD models that were analyzed when Complete Data was used for testing by the iDA software:

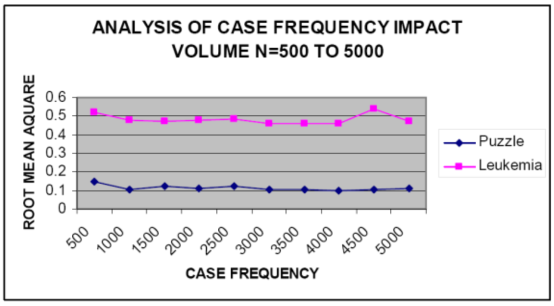
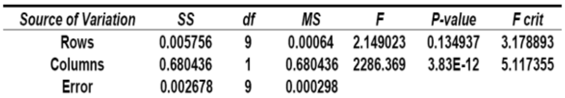


Figure 3. Root Mean Square Values For Two Models with Complete Data

It can be seen from the above graph that a slight positive variation (lower RMS value) was observed when the Case Frequency Volume was increased from 500 to 1000 instances in both of the original KDD models. As the Case Frequency volume was increased in increments of five hundred, only slight positive or negative variation was observed.

An ANOVA test was then performed on the new Root Mean Square (RMS) values calculated by the iDA software. This test supported the conclusion that data set size (case frequency volume) of a KDD does not significantly impact the Root Mean Square values calculated by the proposed KDD's that utilize an S-Sigmoid transfer function employing a Neural Network as its data mining algorithm. Table 1 illustrates the results of the ANOVA test.

TABLE I. ANOVA Results for Root Mean Square Value



Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.005756	9	0.00064	2.149023	0.134937	3.178893
Columns	0.680436	1	0.680436	2286.369	3.83E-12	5.117355
Error	0.002678	9	0.000298			

Therefore, at the 0.05 level of significance, we accept the null hypothesis that data set case frequency is not a significant factor in the calculation of Root Mean Square Values when a Neural Network utilizing an S-Sigmoid Transfer Function within the Activation Function is employed as the data mining algorithm by a KDD model.

Hypothesis 2 tested the significance of data missingness on KDD models from various scientific/medicine/business disciplines that contain case frequencies of less than 1000 instances (N = 500). That is, no significant difference exists in the RMS values calculated in a KDD model containing less than 1000 instances using complete data, when data missingness is injected into the model at levels of 10%, 20%, 30%, 40%, 50%, 60% and 70%.

Five new KDD models were constructed by randomly selecting 500 cases from the original KDD models (N=500, N=1000, N=3500, N=5000, N=7000) selected for this study. Using complete data for these new KDD models containing 500 cases, the models were trained and tested, and the resulting Root Mean Square (RMS) values were recorded.

Data missingness was then injected into each of these five new KDD models (N=500) at the aforementioned levels (10%-70%). The KDD models were re-trained and re-tested and new Root Mean Square (RMS) values were calculated.

The following graphs illustrate a comparison of all Root Mean Square (RMS) values that were calculated when the datasets used for training and testing the ANN in the KDD model contained N=500 instances only, and when noise was randomly injected into each model, at 10%-70%

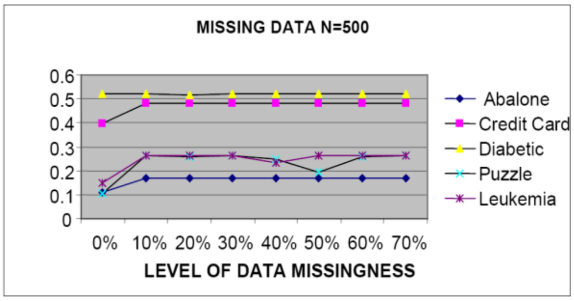


Figure 4. 500 Instances – Missing Data Plot

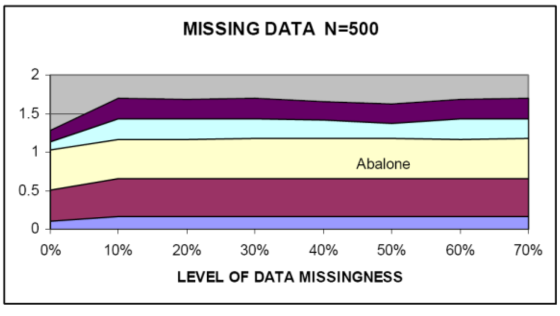


Figure 5. 500 Instances – Missing Data Area Chart

Figures 4 and 5 illustrate that in four out of the five models tested (80%), significant degradation occurred immediately upon the original injection of data missingness into the model at the 10% level. It can also be seen that all five of the KDD models failed to significantly degrade (or improve) with increased levels of data missingness (20%-70%).

The only model that did not spike (show a significant increase in the calculated Root Mean Square value) at the 10% level of data missingness was the model that originally contained N=7000 cases. This may be attributed to inheritance factors contained in the original data set prior to case selection to create the smaller data subset of N=500 cases.

An ANOVA test was executed to test if the level of data missingness injected into the KDD models using only N=500 cases for ANN training and testing was a significant factor. The resulting ANOVA testing the level of data missingness is shown in Table II.

TABLE I. AVONA TEST FOR LEVEL OF DATA MISSINGNESS

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.813433	4	0.203358	162.7526	4.28E-22	2.641465
Within Groups	0.043732	35	0.001249			
Total	0.857165	39				

The ANOVA test indicates that we must accept the null hypothesis that the level of data missingness injected into a KDD model containing less than 1000 cases for training and testing is not a significant factor.

However, it should be noted that the Root Mean Square value (RMS) value spiked significantly at the original introduction of missing data (at the 10% level) in four of the five models, with the exception being the model that had the largest original case frequency (N=7000).

Hypothesis 3, Hypothesis 4 and Hypothesis 5 were tested simultaneously. Hypothesis 3 tested the significance of the level of data missingness in the training and testing of KDD models; Hypothesis 4 tested the impact of performing Data Imputation on KDD models when missing data is encountered; Hypothesis 5 tested the performance of Mean Imputation versus Case Deletion performance as the method of Data Imputation in the replacement of missing data in KDD models.

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, and number of epochs) to obtain an RMS value. Each data set was then injected with a particular level of data missingness (e.g., 10%, 20%, 30%, 40%, 50%, 60%, and 70%) and mined again, using the same standard neural network parameters. Data imputation was performed on the missing values using the Case Deletion and Mean Imputation methods. The RMS results were then analyzed using a two-factor ANOVA test conducted at the 0.05 significance level. The two factors included level of data missingness and imputation method. The results of the two-factor ANOVA test are displayed in Table 3.

Concerning Hypothesis 3, the ANOVA results at the .05 level of significance indicate that we must accept the null hypothesis, and conclude that the percentage level of data missingness is not a significant factor in the performance of KDD models from multiple disciplines containing a large frequency of cases (N=3500).

Concerning Hypothesis 4, the two-factor ANOVA test indicates that one of the two factors tested (Data Imputation) is significant at the 0.05 significance level. Therefore, the null hypothesis H4 was rejected.

Concerning Hypothesis 5, the two-factor ANOVA test displayed above resulted in a rejection of the null hypothesis (that performing Data Imputation was not significant). A test of multiple comparisons was then performed on type of Data Imputation method (i.e., Case Deletion and Mean Substitution). Tukey's Honestly Significant Test was chosen for this test at the 0.05 significance level. The results from Tukey's test are illustrated below in Table III:

TABLE I. RESULTS OF TUKEY'S HSD MULTIPLE COMPARISONS FOR IMPUTATION MEHTOD

Imputation Method	|Tukey's HSD Mean Substitution	Case Deletion
No Imputation	.0931 (*)	.0964 (*)
Mean Substitution	---	.0033

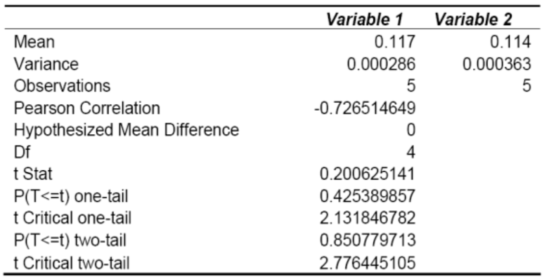
*The mean difference is significant at the .05 level.

The results of Turkey's Honestly Significant Difference Test indicate that there is no significant difference between the type of Data Imputation Method employed (Case Deletion vs. Mean Substitution) on missing data in KDD models of various case frequencies (N=500, N=1000, N=3500, N=5000, N=7000) and over multiple disciplines of study.

Concerning Hypothesis 6, Due to factors such as the timeliness of data entry, implementation of indexes, data sorting, original data sequence, data deletion, and data aging—data clusters may be formed within the training data for the neural network in KDD. Even if the data is presumed smooth and no clustering is known to exist, this test will determine if data re-sequencing has an effect on the calculated Root Mean Square values in KDD models of various case frequencies.

A T-Test was conducted to test the means of the Original Order vs. Re-sequenced data sets. The results shown in Table IV indicate that the difference in means is not statistically significant at the .05 level.

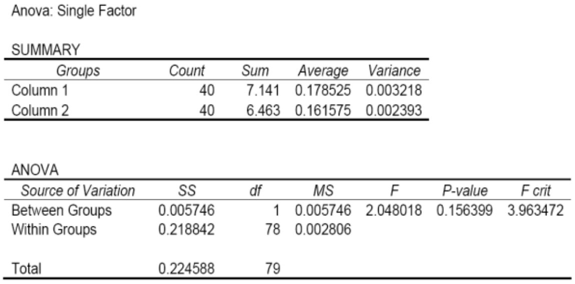
TABLE II. PAIRED MEANS OF ORIGINAL ORDER VS. RE-SEQUENCED DATA SET



	Variable 1	*Variable 2*
Mean	0.117	0.114
Variance	0.000286	0.000363
Observations	5	5
Pearson Correlation	-0.726514649	
Hypothesized Mean Difference	0	
Df	4	
t Stat	0.200625141	
P(T<=t) one-tail	0.425389857	
t Critical one-tail	2.131846782	
P(T<=t) two-tail	0.850779713	
t Critical two-tail	2.776445105	

Concerning Hypothesis 7, Data sets with case frequencies of 500, 1000, 3500, 5000 and 7000 with different levels of Data missingness (10%-70%) were tested and Root Mean Square (RMS) values computed. The data sets were then re-sequenced, re-trained, and re-tested to compute their Root Mean Square (RMS) values. Table V gives the results of the ANOVA test. It indicates that the re-sequencing of data instances containing missing data in KDD models had no statistically significant impact on the calculation of the Root Mean Square (RMS) Value for those models.

TABLE I. PAIRED MEANS OF ORIGINAL ORDER VS. RE-SEQUENCED DATA SET WITH MISSING DATA



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	40	7.141	0.178525	0.003218
Column 2	40	6.463	0.161575	0.002393

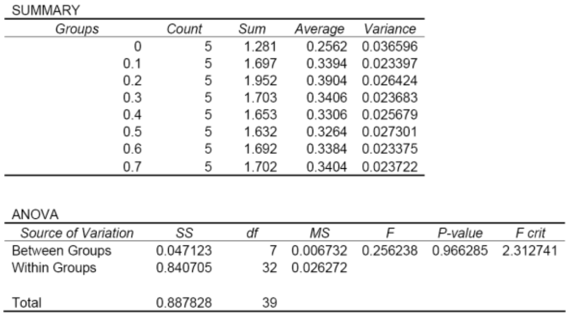
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.005746	1	0.005746	2.048018	0.156399	3.963472
Within Groups	0.218842	78	0.002806			
Total	0.224588	79				

Concerning Hypothesis 8, Root Mean Square Values (RMS) were first calculated for the five KDD models from various disciplines when tested with N=500 randomly selected data instances. Data missingness was then injected into the model at the

10%, 20%, 30%, 40%, 50%, 60% and 70% levels and the KDD models re-trained and re-tested. Table VI with ANOVA results indicates that there is no significant difference in the Root Mean Square Values (RMS) calculated for the five KDD models from various disciplines when tested with N=500 randomly selected data instances containing missing data injected at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels.

TABLE I. ANOVA RESULTS FOR DATA MISSINGNESS



SUMMARY

Groups	Count	Sum	Average	Variance
0	5	1.281	0.2562	0.036596
0.1	5	1.697	0.3394	0.023397
0.2	5	1.952	0.3904	0.026424
0.3	5	1.703	0.3406	0.023683
0.4	5	1.653	0.3306	0.025679
0.5	5	1.632	0.3264	0.027301
0.6	5	1.692	0.3384	0.023375
0.7	5	1.702	0.3404	0.023722

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.047123	7	0.006732	0.256238	0.966285	2.312741
Within Groups	0.840705	32	0.026272			
Total	0.887828	39				

Hypothesis 9 dealt with the comparison of of Multiple Regression Imputation to Mean Substitution and Case Deletion, and concluded that the Mutiple Regression Inputation Method was not significantly different from Mean Substitution nor Case Deletion.

A KDD model that originally contained 1000 data instances was selected for this test, and a Root Mean Square (RMS) value was calculated for the model. Data missingness was then randomly injected into the independent variables in the model at increasing levels of data missingness, i.e., 10%, 20%, 30%, 40%, 50%, 60% and 70%. Root mean square values were calculated for the new KDD models at each of the specified levels of data missingness.

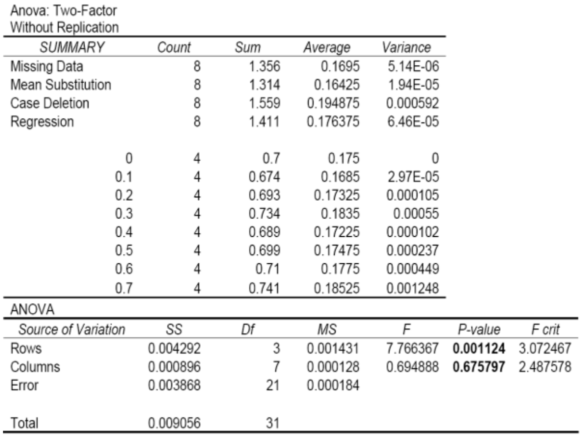
Regression Imputation, Mean Substitution and Case Deletion were then performed on the missing data in each of these new KDD models. All KDD models were re-trained, re-tested, and new Root Mean Square (RMS) values were calculated.

A two-factor Analysis of Variance (ANOVA) test was then performed at the .05 level of significance, testing the level of Data missingness and the type of Imputation Method employed (prior to training and testing the KDD models) when calculating Root mean Square (RMS) values for the KDD models.

Table VII displays the ANOVA results of Root Mean Square values for Data missingness and Data Imputation Method in KDD models with 1000 data instances.

The two-factor ANOVA test shows that the employment of some type of data imputation method is significant and all but one of the imputation methods is significant at 0.5 level, with all imputation methods being significant at .10 level. When KDD models tested with a Case Frequency Volume of N=1000, the imputation method of Regression Imputation did not result in better (lower) Root Mean Square (RMS) values than the imputation methods of Mean Substitution and Case Deletion.

TABLE I. ANOVA RESULTS FOR LEVEL OF DATA MISSINGNESS AND IMPUTATION METHOD N= 1000



Anova: Two-Factor
Without Replication

SUMMARY	Count	Sum	Average	Variance
Missing Data	8	1.356	0.1695	5.14E-06
Mean Substitution	8	1.314	0.16425	1.94E-05
Case Deletion	8	1.559	0.194875	0.000592
Regression	8	1.411	0.176375	6.46E-05
0	4	0.7	0.175	0
0.1	4	0.674	0.1685	2.97E-05
0.2	4	0.693	0.17325	0.000105
0.3	4	0.734	0.1835	0.00055
0.4	4	0.689	0.17225	0.000102
0.5	4	0.699	0.17475	0.000237
0.6	4	0.71	0.1775	0.000449
0.7	4	0.741	0.18525	0.001248

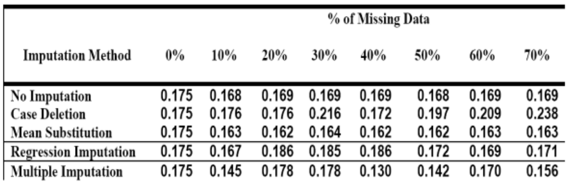
ANOVA

Source of Variation	SS	Df	MS	F	P-value	F crit
Rows	0.004292	3	0.001431	7.766367	**0.001124**	3.072467
Columns	0.000896	7	0.000128	0.694888	**0.675797**	2.487578
Error	0.003868	21	0.000184			
Total	0.009056	31				

Hypothesis 10 Hypothesis 10 compared the results of using Multiple Inputation (a Hybrid of Regression Imputation and Mean Substitution) to Mean Substitution, Case Deletion and Mutiple Regression. It was found that the Multiple Imputation Method did not result in significantly better results (a lower Root Mean Square value) than any of the other methods of Data Imputation.

A KDD model containing 1000 cases was selected for training and testing. New data sets were created by injecting data missingness into the original model at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels. The new KDD models were constructed by employing data imputation methods, namely regression imputation, case deletion, mean substitution, and multiple imputation. The multiple imputation method used is a hybrid method combining Regression Imputation and Mean Substitution with 50% of these Missing Values imputed utilizing Regression Imputation, and the remaining 50% imputed using Mean Substitution. All of the KDD models were re-trained and re-tested, and new Root Mean Square (RMS) values were calculated. The results of all tests are displayed in Table VIII:

TABLE I. ROOT MEAN SQUARE STATISITCS



Imputation Method	% of Missing Data 0%	10%	20%	30%	40%	50%	60%	70%
No Imputation	0.175	0.168	0.169	0.169	0.169	0.168	0.169	0.169
Case Deletion	0.175	0.176	0.176	0.216	0.172	0.197	0.209	0.238
Mean Substitution	0.175	0.163	0.162	0.164	0.162	0.162	0.163	0.163
Regression Imputation	0.175	0.167	0.186	0.185	0.186	0.172	0.169	0.171
Multiple Imputation	0.175	0.145	0.178	0.178	0.130	0.142	0.170	0.156

Figure 5 illustrates a comparison of Multiple Imputation performance in the calculation of Root Mean Square (RMS) values in KDD models against the same models when no Data Imputation method is conducted and when three data imputation methods (Regression Imputation, Mean Substitution and Case Deletion) are utilized.

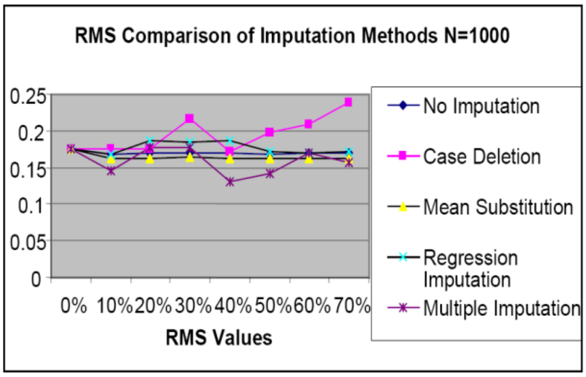
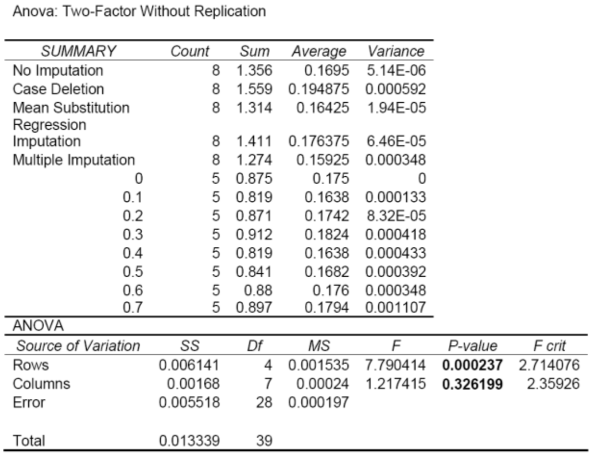


Figure 6. Multiple Imputation Method Comparison

A two-factor ANOVA test was performed on the calculated Root Mean Square (RMS) values for KDD models containing 1000 data instances when Data missingness was injected at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels, and when no Data Imputation was performed on the Missing Values and when the Imputation Methods of Multiple Imputation, Regression Imputation, Mean Substitution and Case Deletion were performed on the models prior to re-training, re-testing and calculating new Root Mean Square (RMS) values for the KDD models. The results of the two-factor ANOVA test are displayed in Table 5.94:

TABLE I. TABLE 9 ANOVA TEST FOR LEVEL OF DATA MISSINGNESS AND AND IMPUTATION METHODS (N=1000)



Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
No Imputation	8	1.356	0.1695	5.14E-06
Case Deletion	8	1.559	0.194875	0.000592
Mean Substitution	8	1.314	0.16425	1.94E-05
Regression Imputation	8	1.411	0.176375	6.46E-05
Multiple Imputation	8	1.274	0.15925	0.000348
0	5	0.875	0.175	0
0.1	5	0.819	0.1638	0.000133
0.2	5	0.871	0.1742	8.32E-05
0.3	5	0.912	0.1824	0.000418
0.4	5	0.819	0.1638	0.000433
0.5	5	0.841	0.1682	0.000392
0.6	5	0.88	0.176	0.000348
0.7	5	0.897	0.1794	0.001107

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.006141	4	0.001535	7.790414	**0.000237**	2.714076
Columns	0.00168	7	0.00024	1.217415	**0.326199**	2.35926
Error	0.005518	28	0.000197			
Total	0.013339	39				

The two-factor ANOVA and the Tukey’s Honestly Significant Difference test (HSD) discovered that only one of the Data Imputation methods tested in this research is not statistically significant at the .05 level, but all Imputation Methods are statistically significant at the .10 level. In summation, the Multiple Imputation method does not result in better (lower) Root Mean Square (RMS) values than the imputation methods of Regression Imputation, Mean Substitution and Case Deletion.

VIII. CONCLUSIONS AND IMPLICATIONS

This research took to task the goal of determining the possible impact of data set size (case frequency volume), level of data missingness, and type of data imputation method on data mining processes. Five independent datasets were selected to perform the analysis. 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 T-tests, ANOVA results, and Tukey’s Honestly Significant Difference (HSD) Tests, this research revealed that original the KDD Case Frequency and the type of Imputation Method employed are significant factors in the performance of KDD models that utilize a Neural Network as its Data Mining Algorithm and employ an S-Sigmoid Transfer Function.

However, while the level of data missingness in a KDD model was found to promote a higher (i.e., worse) Root Mean Square (RMS) Value when missing data was first introduced to a KDD model, increased levels of data missingness were not proven to be significant in this study.

It was also discovered, via that Tukey’s Honestly Significant Difference Test (HSD) analysis, that while there is a significant difference between employing and not employing a Data Imputation Method, there is no significant difference between the Imputation Methods of Multiple Imputation (utilizing a hybrid of Regression Imputation and Mean

Substitution), Regression Imputation, Mean Substitution and Case Deletion.

Currently, Data Mining is viewed as an evolving, but not yet mature field (KDNuggets, 2007). Also, as the focus on other technical dimensions (such as Data Warehousing and Data Shaping) continue to evolve concurrently, KDD and Data Mining software is likely to adapt to those advances. In future research, more complex methodology in the design of hybrid data mining algorithms (including merging concepts from Nearest Neighbor, Decision Trees, Association Rules, Genetic Algorithms and newly developed hybrids) can be employed. Parameters used within a particular data mining algorithm may be adjusted to determine which combination of parameter settings perform most effectively when implemented in various types (size and structure) of data sets.

REFERENCES

[1] Fayyad,U., Piatetsky-Shapiro,G. and Smyth P. “From data mining to knowledge discovery in databases”, AI Magazine 17(3): Fall 1996, 37-54

[2] Roiger R., and Geatz, M., Data Mining, Addison-Wesley, 2003.

[3] Miller, H. J., “Geographic representation in spatial analysis”, Journal of Geographical Systems, 2:55-60, 2000.

[4] Adriaans, P. and Zantinge, D. 1997, Data Mining, Addison-Wesley, New York, 2000.

[5] Witten, I., and Frank, E., Data Mining, Academic Press, San Francisco, 2000.

[6] Han, J., and Kamber, M. 1, Data Mining: Concepts and Techniques, Academic Press, San Francisco, 2000.

[7] Groth, R., Data Mining: Building Competitive Advantage, Prentice-Hall, Upper Saddle River, NJ, 2000.

[8] Agrawal, R., Imielinski, T., and Swami, A., “Mining Associations between Sets of Items in Large Databases”. Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, 1993, pp. 207-216.

[9] Agrawal, R. and Srikant, R., “Fast algorithms for mining association rules in large databases”, Proceedings of the 20th International Conference on Very Large Databases, 1994, Santiago de Chile, Chile.

[10] Darling, Charles B., “Datamining for the masses”, Datamation, Vol. 52, 1997, p. 5.

[11] StatSoft , “Pre-empting user questions through anticipation: data mining FAQ lists”. 2002, http://www.statsoft.com/faq.

[12] Warner, B., and Misra, M., “Understanding neural networks as statistical tools”, The American Statistician, 50, 1996, pp. 284-293.

[13] Holland, J., “Adaptation in natural and artificial systems”, University of Michigan Press, 1975, Ann Arbor, MI.

[14] Vosburg; J. and Kumar, A., “Managing dirty data in organizations using ERP: Lessons From A Case Study”, Industrial Management & Data Systems, Vol. 101, No 1, 2001, pp. 21-31.

[15] Xu, H., Horn Nord, J., Brown, N. and Nord, G. D., “data quality issues in implementing an erp”, industrial management & data systems, Vol. 102, No 1, 2002, pp. 47-58.

[16] Loshin, D., “Knowledge integrity: data standards and data models”, DM Review, January, 2004 pp. 2-3.

[17] Little, R., and Rubin, D., , Statistical Analysis With Missing Data, Wiley, New York, 1987.

[18] Rubin, D., “Multiple Imputations In Sample Surveys - A Phenomenological Bayesian Approach To Nonresponse”, Imputation and Editing of Faulty or Missing Survey Data, U.S. Department of Commerce, 1978, pp. 1-23.

[19] Statistical Services of University of Texas, “General FAQ #25: missing or incomplete Data”. 2000, http://www.utexas.edu/cc/faqs/stat/general/gen25.html/

[20] Kim, Y., “The curse of the missing data”, Y. Kim personal website. 2001. http://209.68.240.11:8080/2ndMoment/978476655/addPostingForm

[21] Hair, J., Anderson, R., Tatham, R., and Black, W., Multivariate Data Analysis, Prentice-Hall, Upper Saddle River, NJ, 1998.

[22] Dempster, A., and Rubin, D., “Incomplete data in sample surveys”, in Madow, W., G., Olkin, I., and Rubin, D. (Eds.), Sample Surveys Vol. II: Theory and Annotated Bibliography, Academic Press, New York, 1983, pp. 3-10.

[23] Schafer, J., “Multiple imputation: a primer”, Statistical Methods in Medical Research, Vol. 8, 1999, pp. 3-15.

[24] Schafer, J., and Olsen, M., “Multiple imputation for multivariate missing-data problems: a data analyst's perspective”, Multivariate Behavioral Research, Vol. 33, 1998, pp. 545-571.

[25] Rubin, D., “Multiple imputation after 18+ years (with discussion)”, Journal of the American Statistical Association, Vol. 91, 1996, pp. 473-489.

[26] Schafer, J., Analysis Of Incomplete Multivariate Data, Chapman and Hall, London, 1997.

[27] Little, R., “Regression with missing x's: a review”, Journal of the American Statistical Association, Vol. 87, 1992, pp. 1227-1237.

[28] Barnard, J. and Meng, X., “Applications of multiple imputation in medical studies: from AIDS to NHANES”, Statistical Methods in Medical Research, Vol. 8, 1999, pp. 17-36.

[29] Van Buren, S., Boshuizen, H., and Knook, D., “Multiple imputation of missing blood pressure covariates in survival analysis”, Statistics in Medicine, Vol. 18, 1999, pp. 681-694.

[30] Clogg, C., Rubin, D., Schenker, N., Schultz, B., and Weidman, L., “Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression”, Journal of the American Statistical Association, Vol. 86, No 413, 1991, pp. 68-78.

[31] Kalton, G., and Kasprzyk, D., “Imputing for missing survey responses”, American Statistical Association,

Proceedings of the Section on Survey Research Methods, 1982, pp. 22-31.

[32] Kalton, G., and Kasprzyk, D., “The treatment of missing survey data”, American Statistical Association, Proceedings of the Section on Survey ResearchMethods, 1986, pp. 22-31.

[33] Sande, L., “Imputation in surveys: coping with reality”, The American Statistician, Vol. 36, 1982, pp. 145-152.

[34] Sande, L., “Hot-Deck imputation procedures”, in Madow, W. G. and Olkin, I. (Eds.), “Incomplete Data In Sample Surveys”, Vol. 3, 1986, Proceedings of the Symposium, Academic Press, New York, pp. 339-349.

[35] Pennell, S., “Cross-Sectional imputation and longitudinal editing procedures in the survey of income and program participation”, Technical report, 1993, Institute of Social Research, University of Michigan, Ann Arbor, MI.

[36] Iannacchione, V., “Weighted sequential hot deck imputation macros”, Proceedings of the SAS Users Group International Conference, Vol. 7, 1982 pp. 759-763.

[37] Graham, J., Hofer, S., and Piccinin, A., “Analysis with missing data in drug prevention research”, in Collins, L. M. and Seitz, L. (Eds.), Advances in Data Analysis for Prevention Intervention Research NIDA Research Monograph, Series (#142), 1994, National Institute on Drub Abuse, Washington, D.C.

[38] Dempster, A., Laird, N., and Rubin, D., “Maximum likelihood from incomplete data via the em algorithm (with discussion)”, Journal of the Royal Statistical Society, Vol. B39, 1977, pp. 1-38.

[39] Cohen, J., and Cohen, P., Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences, 2nd ed. Hillsdale, NJ, Lawrence Erlbaum Associates, 1983.

[40] Veelenturf LPJ. 1995, Analysis and applications of artificial neural networks. New York: Prentice-Hall..

[41] Jordan MI.,, “Why the logistic function? A tutorial discussion on probabilities and neural networks’, . Computational Cognitive Science Technical Report 9503. Cambridge, MA, 1995, Massachusetts Institute of Technology.

[42] Kros, JF.,. Lin, M.. and Brown, M L., , “Effects of the neural network s-Sigmoid function on KDD in the presence of imprecise data.” Computers and Research, Vol. 33 No. 11, 2006, pp. 3136-3149.

[43] Howell, D.C., “Treatment of missing data”, D. C. Howell personal website, 1998. http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html/

[44] Little, R., and Rubin, D., “The analysis of social science data with missing values”, Sociological Methods and Research, Vol. 18, 1989, pp. 292-326.

Marvin L. Brown, DBA

Dr. Brown is an assistant Professor of Computer and Information Systems at the Grambling State University. He received his Doctor in Business Administration in July, 2008. His research interests include Decision Support Systems, Data Mining, Data Imputation Methodology.

Chien_Hua Mike Lin, Ph.D.

Dr. Lin is a professor of Computer and Information Science at the Cleveland State University. He received his Ph.D. in Operations Research from Case Western Reserve University. His background covers Business Administration, Computer Science, and Information Systems which enabled him to teach and research Information Technology widely and deeply. His research interests include Optimization, Information Management, Data Mining, Software Engineering and Information Security and Management. He is currently serving as a Fulbright Scholar at the Chung Gang University, Taiwan.