Data mining and the impact of missing data

Marvin L. Brown
School of Business, Hawaii Pacific University, Honolulu, Hawaii, USA
John F. Kros
Department of Decision Sciences, East Carolina University, Greenville, North Carolina, USA

Keywords
Data handling,
Database management systems,
Information gathering,
Information retrieval

Abstract
The actual data mining process deals significantly with prediction, estimation, classification, pattern recognition and the development of association rules. Therefore, the significance of the 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 since 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 impact of missing data on the data mining process.

Introduction

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 due to the proliferation of computer software and the high capacity of storage media. The management of missing data in organizations has recently been addressed as more firms implement large-scale enterprise resource planning systems (see Vosburg and Kumar, 2001; Xu *et al.*, 2002). Missing data and data quality regarding data warehousing and customer relationship management is also an area of recent research (see Ma *et al.*, 2000; Berry and Linoff, 2000). The issue of missing data becomes an even more pervasive dilemma in the Knowledge Discovery process, in that as more data is collected, the higher the likelihood of missing data becomes.

The objective of this research is to address the impact of missing data on the data mining operation of the Knowledge Discovery process. The paper begins with a background analysis, including a review of both seminal and current literature, followed by reasons for data inconsistency along with definitions of various types of missing data. The main body of the research focuses on methods of addressing missing data and the impact that missing data has on the Knowledge Discovery process. For a review of data mining techniques within the Knowledge Discovery process see Lee and Siau (2001).

Background

The analysis of missing data is a comparatively recent discipline. However, the literature holds a number of works that provide perspective on missing data and data mining. Afifi and Elashoff (1966) provide an early seminal paper reviewing the missing data and data mining literature. Hartley and Hocking (1971) presented one of the first discussions on dealing with skewed and categorical data, especially maximum likelihood (ML) algorithms such as those used in Amos. An early approach for approaching missing data was proposed by Orchard and Woodbury (1972) using what is commonly referred to as an expectation maximization (EM) algorithm to produce unbiased estimates when the data are missing at random (MAR). ML and EM algorithms were also discussed in Dempster *et al.*'s (1977) work.

Models for nonresponse were discussed by Little (1982) while Little and Rubin (1987) considered statistical analysis with missing data. Little and Rubin's (1987) work defined three unique types of missing data mechanisms and provided parametric methods for handling these types of missing data. These papers sparked numerous works in the area of missing data. Diggle and Kenward (1994) addressed issues regarding data missing completely at random, data missing at random, and likelihood based inference. Graham *et al.* (1997) discussed using the EM algorithm to estimate means and covariance matrices from incomplete data. Papers from Little (1995) and Little and Rubin (1989) extended the concept of ML estimation in data mining but they also tended to concentrate on data that have a few distinct patterns of missing data. A good overview of basic statistical calculations to handle missing data is provided by Howell (1998).

Industrial Management & Data Systems
103/8 [2003] 611-621

[ISSN 0263-5577]
[DOI 10.1108/02635570310497657]

Industrial Management &
Data Systems
103/8 [2003] 611-621

Imputation methodology

In theory, statistical methods such as Bayesian techniques can be used to ameliorate this issue. Bayesian methods have strong assumptions associated with them that are not always met. A valuable alternative is Imputation. A number of articles have been published since the early 1990s regarding imputation methodology. Seminal articles by Schafer and Olsen (1998) and Schafer (Schafer, 1999) provided an excellent starting point for investigating multiple imputation. A detailed discussion is provided by Rubin (1996) on the interrelationship between the model used for imputation and the model used for analysis. Schafer's (1997) text has been considered a follow-up to Rubin's 1987 text. A number of conceptual issues associated with imputation methods are clarified in Little (1992). In addition, a number of case studies have been published regarding the use of imputation in medicine (see Barnard and Meng, 1999; van Buren *et al.*, 1999) and in survey research (Clogg *et al.*, 1991).

The least statistical and/or mathematically complex imputation method is the case deletion method. In this method, records with missing data are simply deleted from the database altogether. Although easy to implement, this method has obvious drawbacks regarding statistical significance issues from the deletion of information and henceforth smaller sample sizes. Improved and/or more mainstream imputation methods include case substitution, mean substitution, cold deck, hot deck, regression, multiple, and nearest neighbor imputation. Even though being very popular in practice, hot deck imputation and nearest neighbor methods have received little overall coverage with regard to Data Mining (for a brief discussion see Ernst, 1980; Kalton and Kish, 1981; Ford, 1981; and David *et al.*, 1986).

The impact of incomplete or missing data on the Knowledge Discovery (data mining) process has more recently been approached in association with these individual methodologies. The next section discusses the impact on data mining with inconsistent data or missing data.

Data mining with inconsistent data/missing data

Methods of addressing missing data and the impact that missing data has on the Knowledge Discovery process (depending on the data mining algorithm being utilized) is the focus of the following sections. Reasons for data inconsistency are discussed followed by types of missing data.

Reasons for data inconsistency

Data inconsistency may arise for a number of reasons, including:

- procedural factors;
- refusal of response;
- inapplicable responses.

These three reasons tend to cover the largest areas of missing data in the data mining process. The reasons are discussed next.

Procedural factors

Data entry errors are common. In fact, errors in databases are a fact of life but their impact on the Knowledge Discovery process and data mining can generate serious problems. Dillman (1999) provided an excellent text for designing and collecting data. He also promoted discussion for the reduction of survey error including coverage, sampling, measurement, and nonresponse.

Inaccurate classifications of new data can occur, resulting in classification error or omission, whenever invalid codes are allowed to slip into a database. Erroneous estimates, predictions, and invalid pattern recognition conclusions may also take place. Correlation between attributes can also become skewed which will result in erroneous association rules.

In situations where databases are being refreshed with new data, blank responses from questionnaires further complicate the data mining process. If a large number of similar respondents fail to complete similar questions, the deletion or misclassification of these observations can take the researcher down the wrong path of investigation or lead to inaccurate decision-making by end-users. Methods for prevention of procedural data inconsistency are presented in Jenkins and Dillman (1997). Included are topics such as questionnaire design with regard to typography and layout in order to avoid data inconsistency. Brick and Kalton (1996) also provide an excellent work in which they discuss the handling of missing data in survey research.

Refusal of response

Some respondents may find certain survey questions offensive or they may be personally sensitive to certain questions. For example, some respondents may have no opinion regarding certain questions such as political or religious affiliation. In addition, questions that refer to one's education level, income, age or weight may be deemed too private for some respondents to answer.

Industrial Management &
Data Systems
103/8 [2003] 611-621

Furthermore, respondents may simply have insufficient knowledge to accurately answer particular questions (Hair *et al.*, 1998). Students or inexperienced individuals may have insufficient knowledge to answer certain questions. When polled for data concerning future goals and/or career choices, they may not have had the time to investigate certain aspects of their career choice (such as salaries in various regions of the country, retirement options, insurance choices, etc).

Inapplicable responses
Sometimes questions are left blank simply because the questions apply to a more general population rather than to an individual respondent. If a subset of questions on a questionnaire does not apply to the individual respondent, data may be missing for a particular expected group within a data set.

For example, many graduate students may choose to leave questions blank that concern social activities that they simply do not have time for. Likewise, adults who have never been married or who are widowed or divorced are likely to not answer a question regarding years of marriage.

Types of missing data

The following is a list of the standard types of missing data:

- data missing at random;
- data missing completely at random;
- non-ignorable missing data;
- outliers treated as missing data.

It is important for an analyst to understand the different types of missing data before they can address the issue. Each type of missing data is defined in the following sections.

[Data] Missing At Random (MAR)
Rubin (1978) defined missing data as MAR "when given the variables X and Y, the probability of response depends on X but not on Y". Cases containing incomplete data must be treated differently than cases with complete data. 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 (Statistical Services of University of Texas, 2000).

Consider the situation of reading comprehension. Investigators may administer a reading comprehension test at the beginning of a survey administration session in order to find participants with lower reading comprehension scores. These individuals may be less likely to complete questions that are located at the end of the survey. Similarly, if the likelihood that a respondent will provide his or her weight depends on the probability that the respondent will not provide his or her age, then the missing data is considered to be Missing At Random (MAR) (Kim, 2001).

[Data] Missing Completely At Random (MCAR)
Rubin (1978) and Kim (2001) classified data as MCAR when "the probability of response [shows that] independence exists between X and Y". MCAR data exhibits a higher level of randomness than does MAR. 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.

Consider the case of a laboratory providing the results of a chemical compound decomposition test in which a significant level of iron is being sought. If certain levels of iron are met or missing entirely and no other elements in the compound are identified to correlate then it can be determined that the identified or missing data for iron is MCAR.

Non-ignorable missing data
In contrast to the MAR situation where data missingness is explained 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 (Statistical Services of University of Texas, 2000).

For example, given two variables, X and Y, data is deemed Non-Ignorable when the probability of response depends on variable X and possibly on variable Y. For example, if the likelihood of an individual providing his or her weight varied within various age categories, the missing data is non-ignorable (Kim, 2001). Thus, the pattern of missing data is non-random and possibly predictable from other variables in the database.

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 (2001), "Non-Ignorable missing data is the hardest condition to deal with, but unfortunately, the most likely to occur as well."

Outliers treated as missing data
Pre-testing and calculating threshold boundaries are necessary in the pre-processing of data in order to identify those values which are to be classified as missing. Data whose values fall outside of predefined ranges may skew test results. Many times it

Industrial Management &
Data Systems
103/8 [2003] 611-621

is necessary to classify these outliers as missing data.

Consider the case of a laboratory providing the results of a chemical compound decomposition test. If it has been predetermined that the maximum amount of iron that can be contained in a particular compound is 500 parts/million, then the value for the variable "iron" should never exceed that amount. If, for some reason, the value does exceed 500 parts/million, then some visualization technique should be implemented to identify that value. Those offending cases are then presented to the end users.

For even greater precision, various levels of a specific attribute can be calculated according to its volume, magnitude, percentage and overall impact on other attributes and subsequently used to help determine their impact on overall data mining performance.

Methods of addressing missing data

Methods for dealing with missing data can be broken down into the following categories:

- use of complete data only;
- deleting selected cases or variables;
- data imputation;
- model-based approaches.

These categories are based on the randomness of the missing data and how the missing data is estimated and used for replacement. The next section describes each of these categories.

Use of complete data only

One of the most direct and simple methods of addressing missing data is to include only those values with complete data. Only when missing data is classified as MCAR can this method be used successfully. If missing data are not classified as MCAR, bias will be introduced and make the results non-generalizable to the overall population. This method 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 predetermined percentage of cases. Overall, this method is best suited to situations where the amount of missing data is small.

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. If the deletion of a particular subset (cluster) significantly detracts from the usefulness of the data, case deletion may not be effective. Furthermore, it may not be cost effective simply to delete cases from a sample. Nie *et al.* (1975) examined this strategy, however, no firm guidelines exist for the deletion of offending cases.

For illustration purposes, let us assume that new automobiles costing $20,000 each have been selected and used to test new oil additives. During a 100,000-mile test procedure, the drivers of the automobiles found it necessary to add an oil-additive to the engine while driving. If the chemicals in the oil-additive significantly polluted the oil samples taken throughout the 100,000-mile test, it would be ill advised to eliminate *all* of the samples taken from a $20,000 test automobile. The researchers may determine other methods to gain new knowledge from the test without dropping all sample cases from the test.

In following, if the deletion of an attribute (containing missing data) that is to be used as an independent variable in a statistical regression procedure has a significant impact on the dependent variable, various imputation methods may be applied to replace the missing data (rather than altering the significance of the independent variable on the dependent variable).

Imputation methods for missing data

Imputation methods are literally methods of filling in 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" (Hair *et al.*, 1998). As Dempster and Rubin (1983) 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, imputation methods tend to be a popular method for addressing missing data.

Commonly used imputation methods include:

- case substitution;
- mean substitution;
- hot deck imputation;
- cold deck imputation;
- regression imputation;
- multiple imputation.

Case substitution

This method is most widely used to replace observations with completely missing data.

Industrial Management &
Data Systems
103/8 [2003] 611-621

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 replace missing data with values from previous research.

For example, if the records were lost for an automobile test sample, an authorized researcher could review similar previous test results and determine if they could be substituted for the lost sample values. If it were found that all automobiles had nearly identical sample results for the first 10,000 miles of the test then these results could easily be used in place of the lost sample values.

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 been proven to be valid and are chosen from a population that has been verified to have a normal distribution. If the data is proven to be skewed, the median of the available data can also be used as a substitute.

For example, suppose that respondents to a survey are asked to provide their income levels and choose not to respond. If the mean income from an availably normal and verified distribution is determined to be $48,250, then any missing income values are assigned that value. The rationale for using the mean for missing data is that, without any additional knowledge, the mean provides the best estimate. Otherwise, another measure of central tendency, such as the median, should be considered as an alternative replacement value. There are three main disadvantages to mean substitution:

1. Variance estimates derived using this new mean are invalid by the understatement of the true variance.
2. The actual distribution of values is distorted. It would appear that more observations fall into the category containing the calculated mean than may actually exist.
3. Observed correlations are depressed due to the repetition of a single constant value.

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. Obviously, a researcher must weigh the advantages against the disadvantages before implementation.

Cold deck imputation

Cold deck imputation methods select values or use relationships obtained from sources other than the current database (see Kalton and Kasprzyk, 1982, 1986; Sande, 1982, 1983). 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 (1993) 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 apply to cold deck imputation. 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.

Hot deck imputation

Generally speaking, hot deck imputation replaces missing values with values drawn from the next most similar case. 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. One paper using SAS to perform hot deck imputation is Iannacchione (1982).

From the data set in Table I, it is noted that case three is missing data for item four. In this example, case one, two, and four are examined. Using hot deck imputation, each of the other cases with complete data is examined and the value for the most similar case is substituted for the missing data value. Case four is easily eliminated, as it has nothing in common with case three. Case one and two both have similarities with case three. Case one has one item in common whereas case two has two items in common. Therefore, case two is the most similar to case three.

Once the most similar case has been identified, hot deck imputation substitutes the most similar complete case's value for the missing value. Table II provides the revised data set and displays the hot deck imputation results. Since case two contains the value of 13 for item four, a value of 13 replaces the missing data point for case three.

Industrial Management &
Data Systems
103/8 [2003] 611-621

The 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 like any complete set of data. One of hot deck's disadvantages is the difficulty in defining what is "similar". Hence, many different schemes for deciding on what is "similar" may evolve.

Regression imputation

Regression analysis is used to predict missing values based on the variable's relationship to other variables in the data set. Single and/or multiple regression can be used to impute missing values. The first step consists of identifying the independent variables and the dependent variables. In turn, the dependent variable is regressed on the independent variables. The resulting regression equation is then used to predict the missing values. Table III displays an example of regression imputation.

From the Table, 20 cases with three variables (income, age, and years of college education) are listed. Income contains missing data and is identified as the dependent variable while age and years of college education are identified as the independent variables.

The following regression equation is produced for the example

$$\dot{y} = 33912.14 + 300.87(\text{age}) + 1554.25(\text{years of college education})$$

Predictions of income can be made using the regression equation and the right-most column of the table displays these predictions. For cases 18, 19, and 20, income is predicted to be \$59,785.56, \$50,659.64, and \$53,417.37, respectfully. An advantage to regression imputation is that it preserves the variance and covariance structures of variables with missing data.

Although regression imputation is useful for simple estimates, it has several inherent disadvantages:

- This method reinforces relationships that already exist within the data. As this method is utilized more often, the resulting data becomes more reflective of the sample and becomes less generalizable to the universe it represents.
- The variance of the distribution is understated.
- The assumption is implied 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.

In addition to these points, there is also the problem of over-prediction. Regression imputation may lead to over-prediction of the model's explanatory power. For example, if the regression R^2 is too strong, multicollinearity most likely exists. Otherwise, if the R^2 value is modest, errors in the regression prediction equation will be substantial (see Graham *et al.*, 1994).

Mean imputation can 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.

Overall, regression imputation not only estimates the missing values but also derives inferences for the population (see discussion of variance and covariance above). For discussions on regression imputation see, Royall and Herson (1973) or Hansen *et al.*, (1983).

Multiple imputation

Rubin (1978) 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 (see Little and Rubin, 1987) is combined with maximum likelihood estimates and hot deck imputation to provide data for analysis. 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.

Table I
Illustration of hot deck imputation: incomplete data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	???
4	2	5	10	2

Table II
Illustration of hot deck imputation: imputed data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	13
4	2	5	10	2

Table III
Illustration of regression imputation

Case	Income ($)	Age	Years of college education	Regression prediction ($)
1	45,251.25	26	4	47,951.79
2	62,498.27	45	6	56,776.85
3	49,350.32	28	5	50,107.78
4	46,424.92	28	4	48,553.54
5	56,077.27	46	4	53,969.22
6	51,776.24	38	4	51,562.25
7	51,410.97	35	4	50,659.64
8	64,102.33	50	6	58,281.20
9	45,953.96	45	3	52,114.10
10	50,818.87	52	5	57,328.70
11	49,078.98	30	0	42,938.29
12	61,657.42	50	6	58,281.20
13	54,479.90	46	6	57,077.72
14	64,035.71	48	6	57,679.46
15	51,651.50	50	6	58,281.20
16	46,326.93	31	3	47,901.90
17	53,742.71	50	4	55,172.71
18	???	55	6	59,785.56
19	???	35	4	50,659.64
20	???	39	5	53,417.37

Multiple imputation differs from hot deck imputation in the number of imputed data sets generated. Whereas hot deck imputation generates one imputed data set to draw values from, multiple imputation creates multiple imputed data sets. Multiple imputation creates a summary data set for imputing missing values from these multiple imputed data sets.

Multiple imputation has a distinct advantage in that it is robust to the normalcy conditions of the variables used in the analysis and it outputs complete data matrices. The method is time intensive as the researcher must create the multiple data sets, test the models for each data set separately, and then combine the data sets into one summary set. The process is simplified if the researcher is using basic regression analysis as the modeling technique. It is much more complex when models such as factor analysis, structural equation modeling, or high order regression analysis are used.

A comprehensive handling of multiple imputation is given in Rubin (1987) and Schafer (1997). Other seminal works include Rubin (1986), Herzog and Rubin (1983), Li (1985), and Rubin and Schenker (1986). Other model-based procedures incorporate missing data into the analysis. These procedures are characterized in one of two ways: maximum likelihood estimation or missing data inclusion. Dempster *et al.* (1977) 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 conditional expectation step and "M" for the maximum likelihood step.

The impact of missing data on data mining algorithms

Missing data impacts the Knowledge Discovery process in various ways depending on which data-mining algorithm is being utilized. We will now address the impact of missing data on various types of data mining algorithms.

The impact of missing data on the *k*-nearest neighbor algorithm

The very nature of the k-nearest neighbor algorithm is based on the accuracy of the data. Missing and inaccurate data have a severe impact on the performance of this type of algorithm. If data is missing entirely, misrepresented clusters (data distributions) can occur depending upon the frequency and categorization of the cases containing the missing data. One method to help solve this problem is to use the k-nearest neighbor data mining algorithm itself to approach the missing data problem. The imputed values obtained can be used to enhance the performance of the nearest neighbor algorithm itself.

First, the k-nearest neighbors (those containing no missing data) to the observation that does contain missing data

Industrial Management &
Data Systems
103/8 [2003] 611-621

are identified. The k stands for a predetermined constant representing the number of neighbors containing no missing data to be considered in the analysis. According to Witten and Frank (2000), it is advised to keep the value for k small, say five, so that the impact of any noise present will be kept to a minimum.

Hence, this algorithm is not recommended for large data sets (Adriaans and Zantinge, 1997). Once these "neighbors" have been identified, the majority class for the attribute in question can be assigned to the case containing the missing value. Berson *et al.* (2000) maintained that a historical database containing attributes containing similar predictor values to those in the offending case can also be utilized to aid in the classification of unclassified records.

Of course, the three main disadvantages mentioned in the imputation section (variance understatement, distribution distortion and correlation depression) should be addressed whenever a constant value is used to replace missing data. The proportion of values replaced should be calculated and compared to all clusters and category identification that existed prior to the replacement of the missing data.

The impact of missing data on decision trees

Decision trees are a good methodology for dealing with missing data when it occurs frequently (Berry and Linoff, 1997). Decision trees also scale up very well for large data sets (Adriaans and Zantinge, 1997). It is sometimes useful to prune the tree whenever there is an overabundance of missing data in certain branches (Berry and Linoff, 1997). Eliminating particular paths may be necessary to ensure that the overall success of the decision-making process is not inhibited by the inclusion of cases containing missing data. Witten and Frank (2000) advise the use of prepruning during the tree-building process to determine when to stop developing subtrees. Postpruning can be utilized after a tree is completely built. If one chooses postpruning, decisions for pruning rules can then be made after the tree has been built and analyzed.

The impact of missing data on association rules

Association rules help to identify how various attribute values are related within a data set. Since association rules are many times developed to help identify various regularities (patterns) within a data set, algorithms that utilize association rules have been found to work best with large data sets. They are developed to predict the value of an attribute (or sets of attributes) in the same data set (Darling, 1997). The main focus of association rule discovery is to identify rules that apply to large numbers of cases that the rules can directly relate to, missing data may overstate both the support and the confidence of any newly discovered rules sets (Witten and Frank, 2000).

Attributes containing missing or corrupted data values may easily result in the creation of invalid rule sets or in the failure of identifying valid patterns that normally exist within the data. However, if the data set used to train the algorithm contains only "pristine" data, overfitting the model based on the patterns included in the training set typically results.

Therefore, rules need to be developed for the "exceptions-to-rule-sets" that have been constructed in violation of correct or "clean" data. It is then necessary to populate the training set for algorithms that utilize association rules with a sufficient percentage of "noisy data", representing all possible types of exceptions to existing rules.

In this way, exception rules can be developed to handle all patterns of noise that may be associated with a given data set rather than redesigning rule sets that deal with "clean" data or attempting to force cases that do not belong to existing rule sets into those sets. As exceptions are discovered for initial exceptions, a type of tree structure is created, forming a decision list for the treatment of missing and noisy data for the data set. It becomes necessary to utilize both propositional rules and relational rules in the rule set for the treatment of missing or noisy data.

Propositional rules test an attribute's value against a constant value thereby developing very concise limits to delineate between "clean" and "noisy" data. In extreme instances, the constants, breakpoints and values from associated attributes are used to grow a regression tree in order to estimate missing data values under various conditions.

Incorporating an additional rule or rule set to deal with exceptions (such as missing data) can easily be incorporated since some rules may be developed to predict multiple outcomes. Failure to allow for the missing data exception may easily misrepresent some of the associations between attributes.

Although a rule may have both high support and confidence, a subjective evaluation by the end-user may determine how interesting a newly discovered rule is (Groth, 2000). Some association rule software packages may be trained to automatically

Industrial Management &
Data Systems
103/8 [2003] 611-621

prune "uninteresting rules". Therefore, minimum values (breakpoints) must be established for both the confidence and support of newly discovered rules.

In some instances, a hierarchy of rules can be developed so that some rules may imply other rules. In some cases, only the strongest rule is presented as a newly discovered rule and rules of "lesser strength" (support and confidence) are linked to the stronger rule for use at a later time (Han and Kamber, 2001).

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 (Adriaans and Zantinge, 1997). 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 using neural networks on missing data in the data mining process 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 only, "clean" data, the internal weights developed using the training set cannot be accurately applied to the test set later.

A common question that is asked is "How does missing data actually impact the internal execution of the neural network?" 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 dirty data can distort the weights that are assigned as the associations between nodes in a manner unknown to the research analyst.

While 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 (Westphal and Blaxton, 1998). 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 (Groth, 2000). The activation function can 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 function 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 (Berry and Linoff, 1997).

Conclusions

The issues concerning the impact of inconsistent data and missing data are a fact of life in the world of knowledge discovery and data mining. They must be faced with rigor by developers of new data mining applications before viable decisions can be developed by the end-users of these systems. A review of existing methods for addressing the problem of missing data was conducted for the deletion of cases or variables and various imputation methods. Imputation methods discussed were case substitution, mean substitution, cold deck imputation, hot deck imputation, regression imputation and multiple imputation. The impact of missing data on various data mining algorithms was also addressed, including k-nearest neighbor, decision trees, association rules and neural networks algorithms.

It is the goal of the authors that the issues of inconsistent data and missing data be exposed to individuals new to the venues of knowledge discovery and data mining. It is a topic worthy of research and investigation by developers of fresh data mining applications as well as a method of review for systems that have already been developed or for those that are currently under construction.

References

Adriaans, P. and Zantinge, D. (1997), *Data Mining*, Addison-Wesley, New York, NY.

Afifi, A. and Elashoff, R. (1966), "Missing observations in multivariate statistics I: review of the literature", *Journal of the American Statistical Association*, Vol. 61, pp. 595-604.

Barnard, J. and Meng, X. (1999), "Applications of multiple imputation in medical studies: from AIDS to NHANES", *Statistical Methods in Medical Research*, Vol. 8, pp. 17-36.

Berry, M. and Linoff, G. (1997), *Data Mining Techniques*, Wiley, New York, NY.

Berry, M. and Linoff, G. (2000), "The art and science of customer relationship", *Industrial Management & Data Systems*, Vol. 100 No. 5, pp. 245-6.

Berson, A., Smith, S. and Thearling, K. (2000), *Building Data Mining Applications for CRM*, McGraw-Hill, New York, NY.

Industrial Management &
Data Systems
103/8 [2003] 611-621

Brick, J.M. and Kalton, G. (1996), "Handling missing data in survey research", *Statistical Methods in Medical Research*, Vol. 5, pp. 215-38.

Clogg, C., Rubin, D., Schenker, N., Schultz, B. and Weidman, L. (1991), "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, pp. 68-78.

Darling, C.B. (1997), "Datamining for the masses", *Datamation*, Vol. 52, pp. 5.

David, M., Little, R., Samuhel, M. and Triest, R. (1986), "Alternative methods for CPS income imputation", *Journal of the American Statistical Association*, Vol. 81, pp. 29-41.

Dempster, A. and Rubin, D. (1983), "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, NY, pp. 3-10.

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

Diggle, P. and Kenward, M. (1994), "Informative dropout in longitudinal data analysis (with discussion)", *Applied Statistics*, Vol. 43, pp. 49-94.

Dillman, D.A. (1999), *Mail and Internet Surveys: The Tailored Design Method*, John Wiley Company, New York, NY.

Ernst, L. (1980), "Variance of the estimated mean for several imputation procedures", *American Statistical Association 1980, Proceedings of the Survey Research Methods Section*, pp. 716-20.

Ford, B. (1981), "An overview of hot deck procedures", in Madow, W.G., Olkin, I. and Rubin, D. (Eds), *Sample Surveys Vol. II: Theory and Annotated Bibliography*, Academic Press, New York, NY, pp. 3-10.

Graham, J., Hofer, S. and Piccinin, A. (1994), "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)*, National Institute on Drub Abuse, Washington, DC.

Graham, J., Hofer, S., Donaldson, S., MacKinnon, D. and Schafer, J. (1997), "Analysis with missing data in prevention research", in Bryant, K., Windle, W. and West, S. (Eds), *New Methodological Approaches to Alcohol Prevention Research*, American Psychological Association, Washington, DC.

Groth, R. (2000), *Data Mining: Building Competitive Advantage*, Prentice-Hall, Upper Saddle River, NJ.

Hair, J., Anderson, R., Tatham, R. and Black, W. (1998), *Multivariate Data Analysis*, Prentice-Hall, Upper Saddle River, NJ.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Academic Press, San Francisco, CA.

Hansen, M., Madow, W. and Tepping, J. (1983), "An evaluation of model-dependent and probability-sampling inferences in sample surveys", *Journal of the American Statistical Association*, Vol. 78, pp. 776-807.

Hartley, H. and Hocking, R. (1971), "The analysis of incomplete data", *Biometrics*, Vol. 27, pp. 783-808.

Herzog, T. and Rubin, D. (1983), "Using multiple imputations to handle nonresponse in sample surveys", in Madow, W. G., Olkin, I. and Rubin, D. (Eds), *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, Academic Press, New York, NY, pp. 209-45.

Howell, D.C. (1998), "Treatment of missing data", D.C. Howell personal Web site, available at: www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html/

Iannacchione, V. (1982), "Weighted sequential hot deck imputation macros", *Proceedings of the SAS Users Group International Conference*, Vol. 7, pp. 759-63.

Jenkins, C.R. and Dillman, D.A. (1997), "Towards a theory of self-administered questionnaire design", in Lyberg, L. *et al.* (Eds), *Survey Measurement and Process Quality*, John Wiley Company, New York, NY.

Kalton, G. and Kasprzyk, D. (1982), "Imputing for missing survey responses", *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pp. 22-31.

Kalton, G. and Kasprzyk, D. (1986), "The treatment of missing survey data", *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pp. 22-31.

Kalton, G. and Kish, L. (1981) "Two efficient random imputation procedures", *American Statistical Association 1981, Proceedings of the Survey Research Methods Section*, pp. 146-51.

Kim, Y. (2001), "The curse of the missing data", Y. Kim personal Web site, available at: http://209.68.240.11:8080/2ndMoment/978476655/addPostingForm/

Lee, S. and Siau, K. (2001), "A review of data mining techniques", *Industrial Management & Data Systems*, Vol. 101 No. 1, pp. 41-6.

Li, K. (1985), "Hypothesis testing in multiple imputation – with emphasis on mixed-up frequencies in contingency tables", PhD thesis, The University of Chicago, Chicago, IL.

Little, R. (1982), "Models for nonresponse in sample surveys", *Journal of the American Statistical Association*, Vol. 77, pp. 237-50.

Little, R. (1992), "Regression with missing X's: a review", *Journal of the American Statistical Association*, Vol. 87, pp. 1227-37.

Little, R. (1995), "Modeling the drop-out mechanism in repeated-measures studies", *Journal of the American Statistical Association*, Vol. 90, pp. 1112-21.

Industrial Management &
Data Systems
103/8 [2003] 611-621

Little, R. and Rubin, D. (1987), *Statistical Analysis with Missing Data*, Wiley, New York, NY.

Little, R. and Rubin, D. (1989), "The analysis of social science data with missing values", *Sociological Methods and Research*, Vol. 18, pp. 292-26.

Ma, C., Chou, D. and Yen, D. (2000), "Data warehousing, technology assessment and management", *Industrial Management & Data Systems*, Vol. 100 No. 3, pp. 125-35.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K. and Bent, D. (1975), *SPSS*, 2nd ed., McGraw-Hill, New York, NY.

Orchard, T. and Woodbury, M. (1972), "A missing information principle: theory and applications", *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 697-715.

Pennell, S. (1993), "Cross-sectional imputation and longitudinal editing procedures in the survey of income and program participation", Technical report, Institute of Social Research, University of Michigan, Ann Arbor, MI.

Royall, R. and Herson, J. (1973), "Robust estimation from finite populations", *Journal of the American Statistical Association*, Vol. 68, pp. 883-9.

Rubin, D. (1978) "Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse", *Imputation and Editing of Faulty or Missing Survey Data*, US Department of Commerce, Washington, DC, pp. 1-23.

Rubin, D. (1986), "Statistical matching using file concatenation with adjusted weights and multiple imputations", *Journal of Business and Economic Statistics*, Vol. 4, pp. 87-94.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, NY.

Rubin, D. (1996), "Multiple imputation after 18+ years (with discussion)", *Journal of the American Statistical Association*, Vol. 91, pp. 473-89.

Rubin, D. and Schenker, N. (1986), "Multiple imputation for interval estimation from simple random sample with ignorable nonresponse", *Journal of the American Statistical Association*, Vol. 81, pp. 366-74.

Sande, L. (1982), "Imputation in surveys: coping with reality", *The American Statistician*, Vol. 36, pp. 145-52.

Sande, L. (1983), "Hot-deck imputation procedures", in Madow, W.G. and Olkin, I. (Eds), *Incomplete Data in Sample Surveys, Vol. 3, Proceedings of the Symposium*, Academic Press, New York, NY, pp. 339-49.

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

Schafer, J. (1999), "Multiple imputation: a primer", *Statistical Methods in Medical Research*, Vol. 8, pp. 3-15.

Schafer, J. and Olsen, M. (1998), "Multiple imputation for multivariate missing-data problems: a data analyst's perspective", *Multivariate Behavioral Research*, Vol. 33, pp. 545-71.

Statistical Services of University of Texas (2000), "General FAQ #25: handling missing or incomplete data", available at: www.utexas.edu/cc/faqs/stat/general/gen25.html

van Buren, S., Boshuizen, H. and Knook, D. (1999), "Multiple imputation of missing blood pressure covariates in survival analysis", *Statistics in Medicine*, Vol. 18, pp. 681-94.

Vosburg, J. and Kumar, A. (2001), "Managing dirty data in organizations using ERP: lessons from a case study", *Industrial Management & Data Systems*, Vol. 101 No. 1, pp. 21-31.

Westphal, C. and Blaxton, T. (1998), *Data Mining Solutions*, Wiley, New York, NY.

Witten, I. and Frank, E. (2000), *Data Mining*, Academic Press, San Francisco, CA.

Xu, H., Horn Nord, J., Brown, N. and Nord, G.D. (2002), "Data quality issues in implementing an ERP", *Industrial Management & Data Systems*, Vol. 102 No. 1, pp. 47-58.

Further reading

Heitjan, D.F. (1997), "Annotation: What can be done about missing data? Approaches to imputation", *American Journal of Public Health*, Vol. 87 No. 4, pp. 548-50.

Roth, P. (1994), "Missing data: a conceptual review for applied psychologists", *Personnel Psychology*, Vol. 47, pp. 537-60.

Wothke, W. (1998), "Longitudinal and multi-group modeling with missing data", in Little, T.D., Schnabel, K. U. and Baumert, J. (Eds), *Modelling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples*, Lawrence Erlbaum Associates, Mahwah, NJ.