DOI: 10.15415/mjis.2013.12015

Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods

Shamsher Singh

Dept. Computer Science, A.B College
Pathankot (Punjab)

Prof. Jagdish Prasad

HOD. Dept. Of Stat. Univ. of Raj. Jaipur

Abstract

Many existing, industrial, and research data sets contain missing values (MVs). There are various reasons for their existence, such as manual data entry procedures, equipment errors, and incorrect measurements. The presence of such imperfections usually requires a preprocessing stage in which the data are prepared and cleaned , in order to be useful to and sufficiently clear for the knowledge extraction process. MVs make the performance of data analysis difficult. The presence of MVs can also pose serious problems for researchers. In fact, in appropriate handling of the MVs in the analysis may introduce bias and can result in misleading conclusions being drawn from a research study and can also limit the generalize ability of the research findings. The various types of problem are usually associated with MVs in data mining are (1) loss of efficiency;(2) complications in handling and analyzing the data; and(3) bias resulting from differences between missing and complete data. We will focus our attention on the use of imputation methods. A fundamental advantage of this approach is that the MV treatment is independent of the learning algorithm used. For this reason, the user can select the most appropriate method for each situation he faces. In this paper different methods of estimation of missing values are discussed. The comparison of different imputation methods are given by using non parametric methods.

Keywords: Missing values, imputation methods, non parametric, data mining.

1 INTRODUCTION

Information quality is important to organizations. People use information attributes as a tool for assessing information quality. Information quality is measured based on users' as well as experts' opinions on the information attributes. The commonly known information attributes for information quality including accuracy, objectivity, believability, reputation, access, security, relevancy, value-added, timeliness, completeness, amount of data, interpretability, ease of understanding, concise representation, and consistent representation. As the boundary between data and information can never be unambiguous, these attributes can also be applicable to data quality. Commonly, one can rarely find a data set that contains complete entries. According to "10 potholes in the road to information quality", the common causes of incompleteness include:

Mathematical Journal of
Interdisciplinary Sciences
Vol. 1, No. 2,
March 2013
pp. 75–90

- data are produced using subjective judgments, leading to omission;
- systemic errors in information production lead to lost data;
- access to data may conflict with requirements for security, privacy, and confidentiality; and
- lack of sufficient computing resources limits access.

The extent of damage of missing data is unknown when it is virtually impossible to return the data source for completion. The incompleteness of data is vital to data quality.

There have been several traditional approaches to handling missing values in data analysis including eliminating from the data set those records that have missing values and imputations. However, these traditional approaches can lead to biased analysis results and invalid conclusions. Research has pointed out that the properties of missing values must be taken into account in assessing the data quality of a data set.

2 PREVIOUS RESEARCH

• **Hruschka et al.** propose two imputation methods based on Bayesian networks. They compare them with 4 classical imputation methods: EM, Data Augmentation, C4.5, and the CMCmethod, using 4 nominal data sets from the UCI repository with natural MVs (but inducingMVs in them as well). In their analysis, they employ 4 classifiers as follows: one-rule, Naïve-Bayes, C4.5, and PART.

• **Farhangfar et al.** take as the objective of their paper to develop a unified framework supporting a host of imputation methods. Their study inserts some imputation methods into their framework (Naïve-Bayes and Hot Deck) and compares this with other basic methods: mean, Linear Discriminant Analysis, Logreg, etc. All their experimentation is based on discrete data, so they use the "accuracy" of imputed values against randomly generated MVs.

• **Song et al.** study the relationship between the use of the KNNI method and the C4.5 performance (counting with its proper MV technique) over 6 data sets of software projects. They emphasize the different MVs' mechanisms (MCAR, MAR, and NMAR)and the amount of MVs introduced. From their analysis, they found results that agree with Batista and Monard : KNNI can improve the C4.5 accuracy. They ran a Mann–Whitney statistical test to obtain significant differences in this statement. They also show that the missingness mechanism and pattern affect the classifier and imputation method performance.

• **Twala** empirically analyzes 7 different procedures to treat artificial MVs for decision trees over 21 real data sets. From the study, it can be concluded that listwise deletion is the worst choice, while the multiple imputation strategy performs better than the rest of the imputation methods (particularly those with high amounts of MVs), although there is no outstanding procedure.

3 PATTERNS OF MISSING VALUES

- **MCAR –** Missing completely at random This type is defined by the Equation

$$P(p_l \mid X, Y_{O,l,} Y_{m,l}) = f(l, X),$$

where f is a function, i.e., only the covariate variables X have an effect on the missing data patterns. Note here that if there are no covariates in the model then MARX is equivalent to MCAR.

- **MARX –** Missing at random with respect to X This type is defined by the equation

P(pljX;Yo;l;Ym;l) = f (l;X);

where f is a function, i.e., only the covariate variables X have an effect on the missing data patterns. Note here that if there are no covariates in the model then MARX is equivalent to MCAR.

- **MAR –** Missing at random This type is defined by the equation

$$P(p_l | X, Y_{\mathrm{O},l}, Y_{m,l}) = f(l, X, Y_{\mathrm{o},l}),$$

where f is a function, i.e., only the covariate variables X and the observed dependent variables Yo;lhave an effect on the missing data patterns. Note here that if there is only one dependent variable Y then there is only one incomplete pattern which has no observed dependent variables in it. Therefore MAR is equivalent to MARX for models with one dependent variable.

Missing values often randomly distributed throughout the sample space. There is no particular assumption on the reason of value missing. Few correlations among the missing values can be observed if the missing values have the MAR pattern. Since values are missing at random, the missing values distribute almost equally towards each attribute. The quality of the entire data set is homogeneous. Accordingly, complete data can be considered representative for the entire data set.

- **NMAR –** Not Missing at random This type is defined by the equation

$$P(p_l \mid X, Y_{o,l}, Y_{m,l}) = f(l, X, Y_{o,l}, Y_{m,l}),$$

where f is a function, i.e., all three types of variables have an effect on the missing data patterns. It is well know how FIML (full information maximum-likelihood) estimation performs under all of these conditions.

Missing in Cluster (MIC)

Data are often missing in some attributes more than in others. Also, missing values in those attributes can be correlated. It is difficult to use statistical techniques to detect multi-attribute correlations of missing values. The quality of data with this pattern of missing values is less homogeneous than that with MAR. Applications of any analytical results based on the complete data set should be cautious, since the sample data are biased in the attributes with a large number of missing values.

Systematic Irregular Missing (SIM)

Data can be missing highly irregularly, but systematically. There might be too many missing correlations between the attributes, but these correlations are too tedious to analyze. An implication of SIM is that the data with complete entities are unpredictably under-representative.

4 ROLE OF CORRELATION TO DETECT THE PATTERN OF MISSING VALUE FOR DATA MINING:

Introduction:

Databases contain data with different characteristics and it is usual to classify data into one of two types. Real-valued data contains real numbers and commonly arises from measurements, e.g. engine capacity, days-worked, temperature. With real-valued data, the values are ordered by $\leq$ and meaning can be inferred from the differences between values. In many cases, values can also be multiplied and divided and meaning inferred from the results. However, it is not always the case that this is possible. For example, it is dangerous to infer that 80 F is twice as hot as 40 F; after all, using the centigrade scale, one is approximately 27° and the other is about 4°. The problem is that with either scale, the zero value of the scale is chosen differently. If real valued data is such that differences can be formed but not products or quotients, the data is called *interval* data.

However, in many cases, the data is not real-valued but is *categorical*; values are drawn from a domain comprising a finite set of possible values. Examples include sex, colour-of-car, and degree-class. Categorical data can either be *nominal* or *ordinal*. It is called nominal iff there is no assumed ordering between the elements of the domain.

Thus, sex with domain {*male, female*} and colour-of-car, perhaps with domain {*red, white, blue, black, green*}, are examples of nominal data whilst degree-class, perhaps with domain {*pass,* 3*rd,* 2(ii)*,* 2(i)*,* 1*st*}, is ordinal because there is a clear ordering of the elements of the domain.

When data mining, databases are handled where the fields can be real valued, nominal-valued or ordinal-valued. It is always important for the data miner to recognise associations between fields. Correlation is a powerful technique and is used to measure the way that the values of one field tend to vary with respect to the values of another. Given two fields, X and Y, the database of n records defines a sequence of values, $\mathbf{x} = (x1, x2, \dots, xn)$, for the first field and a sequence of values, $\mathbf{y} = (y1, y2, \dots, yn)$ for the second field. The correlation, $corr(\mathbf{x}, \mathbf{y})$ produces a value, generally between −1 and 1, which measures how they change together.

The detailed calculation of correlation for different categories of data and its application for data mining is discussed by Rayard-Smith(2007).

5 METHODS FOR IMPUTATION OF MISSING VALUES IN DATABASES FOR DATA MINING:

Now we will describe the various methods used to treat missing values in the supervised classification context.

A. Case Deletion (CD). Also is known as complete case analysis. It is available in all statistical packages and is the default method in many programs. This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute, and delete the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis.

B. Mean Imputation (MI). This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing

Let us consider that the value x_{ij} of the k-th class, Ck, is missing then it will be replaced by

$$\hat{x}_{ij} = \sum_{i:x_{ij} \in c_k} \frac{x_{ij}}{n_k}, \qquad (1)$$

where nk represents the number of non-missing values in the j-th feature of the k-th class. In some studies the overall mean is used but we considered that this does not take in account the sample size of the class where the instance with the missing values belongs to.

C. Median Imputation (MDI). Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given feature is replaced by the median of all known values of that attribute in the class where the instance with the missing feature belongs. This method is also a recommended choice when the distribution of the values of a given feature is skewed. Let us consider that the value xij of the k-th class, Ck, is missing.

It will be replaced by

$$\hat{x}_{ij} = median_{\{i:x_{ij} \in C_k\}}\{x_{ij}\}. \qquad (2)$$

In case of a missing value in a categorical feature we can use mode imputation instead of either mean or median imputation. These imputation methods are applied separately in each feature containing missing values. Notice that the correlation structure of the data is not being considered in the above methods. The existence of others features with similar information (high correlation), or similar predicting power can make the missing data imputation useless, or even harmful.

D. KNN Imputation (KNNI). This method the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function.

The algorithm is as follows:

1. Divide the data set D into two parts. Let Dm be the set containing the instances in which at least one of the features is missing. The remaining instances will complete feature information form a set called Dc.
2. For each vector x in Dm:

a) Divide the instance vector into observed and missing parts as x = [xo; xm].

b) Calculate the distance between the xo and all the instance vectors from the set Dc. Use only those features in the instance vectors from the complete set Dc, which are observed in the vector x.

c) Use the K closest instances vectors (K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the k-nearest neighborhood. The median could be used instead of the median.

E. Hot deck Imputation. In this method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data. In Random Hot deck, a missing value (the recipient) of a attribute is replaced by a observed value (the donor) of the attribute chosen randomly. There are also cold deck imputation methods that are similar to hot deck but in this case the data source to choose the imputed value must be different from the current data source. For more details see Kalton and Kasprzyk (1986).

F. Cold Deck Imputation. Cold deck imputation replaces a missing value of an item by a constant value from an external source, such as a value from a previous realization of the same survey. As with substitution, current practice usually treats the resulting data as a complete sample, that is ignores the consequences of imputation. Satisfactory theory for the analysis of data obtained by cold deck imputation is lacking.

G. Imputation using a prediction model. These methods consist of creating a predictive model to estimate values that will substitute the missing data. The attribute with missing data is used as the response attribute, and the remaining attributes are used as input for the predictive model. The disadvantages of this approach are (i) the model estimated values are usually more well-behaved than the true values would be; (ii) If there are no relationships among attributes in the data set and the attribute with missing data, then the model will not be precise for estimating missing values; (iii) the computational cost since we have to build a large amount of models to predict missing values. Imputation using decision trees algorithms. All the decision trees classifiers handle missing values by using built in approaches. For instance, CART replaces a missing value of a given attribute using the corresponding value of a surrogate attribute, which has the highest correlation with the original attribute. C4.5 uses a probabilistic approach to handle missing data in both the training and the test sample.

H. Multiple imputation. In this method the missing values in a feature are filled in with values drawn randomly (with replacement) from a fitted distribution for that The treatment of missing values and its effect in the classifier accuracy 5 feature. Repeat this a number of times, say M=5 times. After that we can apply the classifier to each" complete" dataset and compute the misclassification error for each dataset. Average the misclassification error rates to obtain a single estimation and also estimate variances of the error rate.

I. CLIP4 CLIP4 is a rule-based algorithm that works in three phases. During the first phase a decision tree is grown and pruned to divide the data into subsets. During the second phase the set covering method is used to generate production rules. Finally, during the third phase goodness of each of the generated rules is evaluated, and only the best rules are kept while the remaining (weaker) rules are discarded. A specific feature of CLIP4 is use of the integer programming model to perform crucial operations, such as splitting the data into subsets during the first phase, selecting the data subsets that generate the least overlapping and the most general rules, and generating the rules from the data subsets in the second phase. The CLIP4 generates data model that consists of production rules, which use inequalities in all selectors, i.e. IF NUMBER_OF_WHEELS ≠ 4 AND ENGINE ≠ yes THEN CLASS=bicycle. It works only with discrete data.

J. Naïve-Bayes Naïve-Bayes is a classification technique based on computing probabilities. It analyzes relationship between each independent variable and the dependent variable to derive a conditional probability for each relationship. When a new example is analyzed, a prediction is made by combining the effects of the independent variables on the dependent variable, i.e. the outcome that is predicted. Naïve-Bayes requires only one pass through the training set to generate a classification model, which makes it very efficient. The Naïve Bayes generates data model that consists of set of conditional probabilities, and works only with discrete data.

K. C4.5 C4.5 is a decision tree algorithm. It uses an entropy based measure, which is called gain ratio, as a splitting criterion to generate decision trees. Each tree level is generated by dividing the data at a given node into a number of subsets, which are represented by branches. For each division, gain ratio is used to select the best attribute, which values are used to divide the data into subsets. Each subset contains data that takes on one of the values of the selected attribute. C4.5 generates data model that consists of a decision tree, which can be translated into a set of production rules that use equalities in all selectors. It can work with both discrete and continuous data.

L. Estimation of Missing Value Through Analysis of Covariance Ancova offers an alternative to the missing data formula technique. It is applicable to any number of missing data. One covariate is assigned to each missing observation. The technique prescribes an appropriate set of values for each covariate.

The only difference, between the use of ANCOVA for error control and that for estimation of missing data, is the manner in which the values of the covariate are assigned.

But when covariance analysis is used to estimate missing data, the covariate is not measured but is assigned, one each, to a missing observation. The case of only one missing observation is discussed here.

The rules for the application of ANCOVA to a data set with one missing observation are:

For the missing observation, set $Y = 0$.

Assign the values of covariate as $X = 1$ for the experimental unit with the missing observation, and $X = 0$ otherwise.

With the complete set of data for the Y variable and the X variable as assigned above, perform the ANCOVA following the standard procedures.

Following is the example of two way (rows vs columns) in which second column of Y there is only one missing observation denoted by zero will be obtained by using ANCOVA procedure

X	Y	X	Y	X	Y	X	Y	X	Y
0	5113	0	5398	0	5307	0	4678	0	20496
0	5346	0	5952	0	4719	0	4264	0	20281
0	5272	0	5713	0	5483	0	4749	0	21217
0	5164	1	0	0	4986	0	4410	1	14560
0	4804	0	4848	0	4432	0	4748	0	18832
0	5254	0	4542	0	4919	0	4098	0	18813

- The various sum of squares for the X variable are:

$$\text{Total SS} \quad = 1 - \frac{1}{rt} = 1 - \frac{1}{24} = 0.9583$$

$$\text{colunm SS} = \frac{1}{t} - \frac{1}{rt} = \frac{1}{6} - \frac{1}{24} = 0.1250$$

$$\text{Row SS} \quad = \frac{1}{t} - \frac{1}{rt} = \frac{1}{4} - \frac{1}{24} = 0.2083$$

$$\text{Error SS} \quad = \text{Total SS} - \text{Column SS} - \text{Treatment SS}$$

$$= 09583 - 01250 - 2083$$

$$= 0.6250$$

- The various sum of cross products are computed as:

$$\text{C.F} = \frac{G_y}{(r)(t)} = \frac{114199}{(4)(6)} = 4758.2917$$

$$\text{Total SCP} = -\,(\text{C.F.}) = -4758.2917$$

$$\text{Column SCP} = \frac{B_y}{t} - \text{C.F.} = \frac{26543}{6} - 4758.2917 = -349.4584$$

$$\text{Row SCP} \quad = \frac{T_y}{r} - \text{C.F} = \frac{14560}{4} - 4758.2917 = 1.118.2917$$

$$\text{Error SCP} \quad = \text{Total SCP} - 1 \ \text{Column SCP} - \text{ROW SCP}$$

$$= -4758.2917 - (-349.4584) - (-1118.2917)$$

$$= -3290.5416$$

Where By is the column total for the Y variable, of the column in which the missing data occurred, and Ty is the treatment total, for the Y variable, corresponding to the row with the missing data.

The estimate of the missin data is computed as:

$$= -\frac{\text{ErrorSCP}}{\text{ErrorSSX}}$$

$$\text{Estimate of missing data Regular } -b_{yx} = \frac{-(-3290.5416)}{0.6250}$$

Estimation of missing observation of fourth row in column II is 5265

M. Estimating Missing Values through Method of Least Squares

The procedure will be explained by considering the dataset arranged in two way (rows vs columns) as follows

1	2	I	C
Y_{11}	Y_{21}	Y_{i1}	Y_{c1}
Y_{12}	Y_{22}	Y_{i2}	Y_{c2}
Y_{1j}	Y_{2j}	X	Y_{cj}
Y_{1r}	Y_{ir}	Y_{ir}	Y_{cr}

Let $y_{ij=}$ x in the jth row in the ith column be missing let $y_{i.}'$ is total of known observations corresponding to ith column . $y_{.j}'$ is the total of known observation in jth row and y..' is total of all known observations. The estimated value of missing observation x is obtained by method of least squares is

$$X=(r\ y_{.j}'+c\ y_{i.}'-\ y..')/\ (r-1)(c-1)$$

Similarly more than one missing observation can also be obtained by the method of least square .

6 COMPARISON OF IMPUTATION METHODS FOR ESTIMATING MISSING VALUES.

In the particular case of classification, learning from incomplete data becomes even more important. Incomplete data in either the training set or test set or in both sets affect the prediction accuracy of learned classifiers. The seriousness of this problem depends in part on the proportion of MVs. Most classification algorithms cannot work directly with incomplete data sets, and due to the high dimensionality of real problems (i.e. large number of cases), it is possible that no valid (complete) cases would be present in the data set. Therefore, it is important to analyze which is the best technique or preprocessing considered in order to treat the present MVs before applying the classification methods as no other option is possible. Usually, the treatment of MVs in data mining can be handled in three different ways:

- The first approach is to discard the examples with MVs in their attributes. Therefore, deleting attributes with elevated levels of MVs is included in this category too.
- Another approach is the use of maximum likelihood procedures, where the parameters of a model for the complete data are estimated, and later used for imputation by means of sampling.
- Finally, the imputation of MVs is a class of procedures that aims to fill in the MVs with estimated ones. In most cases, a data set's attributes are not independent of each other .Thus, through the identification of relationships among attributes, MVs can be determined .

We will focus our attention on the use of imputation methods. A fundamental advantage of this approach is that the MV treatment is independent of the learning algorithm used. For this reason, the user can select the most appropriate method for each situation he faces. There is a wide family of imputation methods, from simple imputation techniques like mean substitution, K-nearest neighbor, etc. to those which analyze the relationships between attributes such as support vector machine-based, clustering-based, logistic regressions, maximum-likelihood procedures, and multiple imputation.

The literature on imputation methods in data mining employs well-known machine learning methods for their studies, in which the authors show the convenience of imputing the MVs for the mentioned algorithms, particularly for classification. The vast majority of MVs studies in classification usually analyze and compare one imputation method against a few others under controlled amounts of MVs and induce them artificially with known mechanisms and probability distributions. We want to analyze the effect of the use of a large set of imputation methods on all the considered classifiers.

In order to perform the analysis, we can use any of the data sets . All the data sets have their proper MVs, and we do not induce them, as we want to stay as close to the real-world data as possible. First, we have to analyze the use of the different imputation strategies , for each data set. All the imputation and classification algorithms are publicly available in the KEEL software1. These results are to be compared using the non parametric test namely krushkal wallis test for testing the effectivenrss of all the imputation methods. If this hypothesis is rejected then Wilcoxon Signed Rank test has to be applied in order to obtain the best method(s) for each classifier. With this information, we can extract the best imputation method. We can also analyzed two metrics related to the data characteristics, formerly known as Wilson's noise ratio and mutual information. Using these measures, we can observe the influence of the imputation procedures on the noise and on the relationship of the attributes with the class label as well. This procedure tries to quantify the quality of each imputation method independently of the classification algorithm and to observe the theoretical advantages of the imputation methods a priori. The obtained results will help us to explain how imputation may be a useful tool to overcome the negative impact of MVs and the most suitable imputation method for each classifier, each group and all the classification methods together.

Comparison methodology

In order to appropriately analyze the imputation and classification methods, we use the kruskal wallis test for testing the performance of K classifiers. In this test we rank the all the values of the K classifiers. The sum of all the N ranks in N(n+1)/2.If Ith classifier has n_i observations . the observed value of the rank sum ofr the Ith classifier is denoted by Ri . Then kruskal wallis statistic H is $H=12/N(N+1)\Sigma R_i^2 /n_i-3(N+1)$. If the value of H is greater then the tabulated value of χ^2 at 5% level of significance wih (K-1) degrees of freedom then we reject the hypothesis that the performance of all K classifiers are different . If the calculated value of H is less then the tabulated value of χ^2 at 5% level of significance wih (K-1) degrees of freedom then we accept the hypothesis that the performance of all K classifiers are same.

CONDUCTING PAIRWISE COMPARISONS AFTER OBTAINING A SIGNIFICANT KRUSKAL-WALLIS TEST.

The pariwise comparisons will be conducted using the Wilcoxon Signed Rank test, which yields identical results with the Kruskal-Wallis test for two independent samples.

We can use Wilcoxon tables directly from the web page. These tables provide us with an average ranking for each imputation method. The content of the tables and its interpretation are as follows:

1. We create an n × n table for each classification method. In each cell, the outcome of the Wilcoxon Signed Rank test is shown.

2. In the aforementioned tables, if the p-value obtained by the Wilcoxon tests for a pair of imputation methods is higher than our α level, formerly 0.05, then we establish that there is a tie in the comparison (no significant difference was found), represented by a D.
3. If the p-value obtained by the Wilcoxon tests is lower than our α level, formerly 0.1, then we establish that there is a win (represented by a W) or a loss (represented by an L) in the comparison. If the method presented in the row has a better ranking than the method presented in the column in the Wilcoxon test, then there is a win, otherwise there is a loss. With these columns, we have produced an average ranking for each classifier. We have computed the number of times that an imputation methods wins and the number of times that an imputation method wins and ties. Then, we obtain the average ranking by putting those imputation methods which have a higher "wins + ties" sum first among the rest of the imputation methods. If a draw is found for "wins + ties", we use the "wins" to establish the rank. If some methods obtain a draw for both "wins + ties" and "wins", then an average ranking is assigned to all of them.

In order to compare the imputation methods for the classification methods considered in each situation (global or family case), we have added two more final columns in the tables contained in the next subsections. In the first new column, we compute the mean of the rankings for each imputation method across all the classifiers of the correspondent group (column "Avg."), that is, the mean of every row. By doing so, we can obtain a new rank (final column RANKS), in which we propose a new ordering for the imputation methods for a given classifier's group, using the values of the column "Avg." to sort the imputation methods.

7 DATA MINING TECHNIQUE'S BASED PACKAGES FOR DETECTION AND ESTIMATION OF MISSING VALUES:

1. SOLAS: This is from Statistical Solutions, Ireland. Available are Group Means, Last Value Carried Forward (for longitudinal data), Hot Deck imputation, and Multiple Imputation (based on propensity scores)

2. SPSS MVA Module: Available here are Listwise analysis, All Value analysis, Regression Imputation (with a random stochastic component), EM (single) Imputation.

3. S-PLUS: supports the Norm, Cat, Mix and Pan libraries, which use the MCMC multiple imputation. Also under the S-PLUS platform is MICE (multiple imputation by chainedequations). S-PLUS 6 has a missing data routines built in, using MCMC methods for MI.

4. **SAS:** PROC MI, and PROC MIANALYSE, are beta versions in SAS 8.2. These use MCMC MI methods, and when a full version is released will form a very powerful tool, as it will integrate with all available SAS analysis.
5. **BMDP:** Have routines available to impute data, using both single and multiple imputation.
6. **BUGS:** MCMC Multiple Imputation is a natural extension of Bayesian analysis.

8 CONCLUSION

In this paper, we have discussed the patterns of missing values in the context of data quality in completeness. Completeness is one of the important attributes of data quality. The ultimate objective of data quality assessment is to fully understand the characteristics of the data set and determine strategies for the data analyses.

Correlation is an important concept in data mining and its determination is not always straightforward. Different techniques must be used dependent upon the nature of the data.

Missing data is a common problem in data mining. An empty entry in the database sometimes indicates the value is zero or, in some cases, cannot possibly exist (e.g. an entry for an individual in a field SALARY when the individual is a baby). However, in many cases, a blank field represents an unknown quantity and techniques based on correlation can then be used to complete that entry. Where mapping functions are developed within the techniques used to compute correlations, this should enable the replacement of missing values to be achieved more effectively.

REFERENCES

Acuna E, Rodriguez C (2004) Classification, clustering and data mining applications. Springer, Berlin, pp 639–648. http://dx.doi.org/10.1007/978-3-642-17103-1_60

Asuncion A, Newman D (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/

Batista G, MonardM (2003) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17(5):519–533. http://dx.doi.org/10.1080/713827181

C.E. Shannon, A Mathematical Theory of Communication, *Bell Systems Technical Journal*, vol.27, pp.379-423, 1948

Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* B39(1), 1997, 1-38.

Ding Y, Simonoff JS (2010) An investigation of missing data methods for classification trees applied to binary response data. J Mach Learn Res 11:131–170

English, L. P., "Help for data quality problems -- A number of automated tools can ease data cleansing and help improve data quality," *InformationWeek*, Oct 7, 1996, 53.

English, L. P., Information quality for business intelligence and data mining: Assuring quality for strategic information uses, 2005. <http://support.sas.com/news/users/LarryEnglish_0206.pdf> [retrieved April 1, 2007].

Farhangfar A, Kurgan LA, Pedrycz W (2007) A novel framework for imputation of missing values in databases. IEEE Trans Syst Man Cybern Part A 37(5):692–709. http://dx.doi.org/10.1109/TSMCA.2007.902631

Garvin, D. A., *Managing Quality*, The Free Press, New York, 1988.

Hruschka ER Jr., Hruschka ER, Ebecken NF (2007) Bayesian networks for imputation in classification problems. J Intell Inf Syst 29(3):231–252. http://dx.doi.org/10.1007/s10844-006-0016-x

Huang, K. T., Lee, Y. W., Wang, R. Y., *Quality Information and Knowledge*, Prentice-Hall, New York, 1999.

J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993 http://dx.doi.org/10.1023/A:1022645310020

J.R. Quinlan, Induction of Decision Trees, *Machine Learning*, vol.1, pp.81-106, 1986 http://dx.doi.org/10.1023/A:1022643204877

K. J. Cios, L.A. Kurgan, Hybrid Inductive Machine Learning: An Overview of CLIP Algorithms, In: L.C. Jain, and J. Kacprzyk, (Eds.), *New Learning Paradigms in Soft Computing*, pp. 276-322, Physica-Verlag (Springer), 2001 http://dx.doi.org/10.1007/978-3-7908-1803-1_10

K. Y. TAM and M. Y. KIANG (1992) Managerial applications of neural networks: The case of bank failure predictions. Mgmt Sci. 38, 936-947. http://dx.doi.org/10.1287/mnsc.38.7.926

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. Survey Methodology 12, 1-16.

Kim H, Golub GH, Park H (2005) Missing value estimation for dna microarray gene expression data: local least squares imputation. Bioinformatics 21(2):187–198 http://dx.doi.org/10.1093/bioinformatics/bth499

L. M. SALCHENBERGERE,. M. CINAR and N. A. LASH (1992) Neural networks: A new tool for predicting thrift failures. Decis. Sci. 23, 899-916. http://dx.doi.org/10.1111/j.1540-5915.1992.tb00425.x

Little, R. J. A., and Rubin, D. B., *Statistical Analysis with Missing Data, 2nd Ed*. New York: John Wiley and Sons, 2002.

N. CAPON (1982) Credit scoring systems: A critical analysis. J. Marketing 41, 82-91. http://dx.doi.org/10.2307/3203343

R. A. WALKING (1985) Predicting tender offer success: A logistic analysis. J. Finance and Quantitative Analysis 20, 461-478. http://dx.doi.org/10.2307/2330762

Rayward-Smith V.J Statistics to measure correlation for data mining applications Computational Statistics & Data Analysis 51 (2007) 3968 – 3982 http://dx.doi.org/10.1016/j.csda.2006.05.025

R. Y. AWH and D. WALTERS (1974) A discriminant analysis of economic, demographic, and attitudinal characteristics of bank charge-card holders: A case study. J. Finance. 29, 973-980. http://dx.doi.org/10.1111/j.1540-6261.1974.tb01495.x

R.O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, 1977

Salaun, Y. and Flores, K., “Information quality: meeting the needs of the consumer,” *International Journal of Information Management*, 21(1), 2001, 21-37. http://dx.doi.org/10.1016/S0268-4012(00)00048-7

Salmela, H., “From information systems quality to sustainable business quality,” *Information and Software Technology*, 39(12), 1997, 819-825. http://dx.doi.org/10.1016/S0950-5849(97)00040-2

Song Q, Shepperd M, Chen X, Liu J (2008) Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation. J Syst Softw 81(12):2361–2370 http://dx.doi.org/10.1016/j.jss.2008.05.008

Strong, D., Lee, Y. W., and Wang, R. Y., 10 potholes in the road to information quality, *IEEE Computer*, 30(8), 1997, 38-46. http://dx.doi.org/10.1109/2.607057

Tozer, G., *Metadata Management for Information Control and Business Success*, Artech House, Norwood, MA, 1999.

Twala B (2009) An empirical comparison of techniques for handling incomplete data using decision trees. Appl Artif Intell 23:373–405 http://dx.doi.org/10.1080/08839510902872223

Wang, H, and Wang, S., Data mining with incomplete data, in *Encyclopedia of Data Warehousing and Mining*, John Wang (Ed.), Idea Group Inc.: Hershey, PA, 2005, pp.293-296.

Wang, R. Y., Lee, Y. W., Pipino, L. L., and Strong, D. M., "Manage your information as a product," *Sloan Management Review*, 39(4), 1998, 95-105.

Shamsher Singh is System Analyst in Department of Computer Science, S.R.P.A A.B College, Pathankot. He has been working in the field of Data Mining along with his great interest in computer languages, he has published various books in Data Mining and C language and also presented various papers in National and international Conferences and in Journal of repute. Email-Id- samsingh94@yahoo.com

Jagdish Prasad is Professor in Department of Statistics, University Of Rajasthan, Jaipur. He is also Head of Department in Department of Statistics. He has been working in the field of Design and experiment, Applied Statistics, Neural network, data mining and other related field in the subject of statistics. He has published various research articles in national and international journal. Email-id-jagdish55_singh@yahoo.co.in