MISSING DATA IMPUTATION IN CARDIAC DATA SET (SURVIVAL PROGNOSIS)

R.KAVITHA KUMAR

Department of Computer Science and Engineering
Pondicherry Engineering College,
Pudhucherry, India

DR. R.M.CHADRASEKAR

Professor, University,
Annamalai University,
Chidambaram, India.

ABSTRACT

Treating missing value is very big task in the data preprocessing methods. Missing data are a potential source of bias when analyzing clinical trials. In this paper we analyze the performance of different data imputation methods in a task where the aim is to predict the probability of survival of cardiac patient. In this paper, comparison of handling missing data in cardiac dataset. Mean Imputation, KNN imputation method, two correlation based methods known as EMImputed _ columns, LSImputed _ Rows and multiple imputation method referred as NORM (which is based on Expectation Maximization algorithm) method were used to replace missing values found in a dataset containing 3500 records of patients. The results were analyzed in terms of the calibration of the results. Nevertheless, k-NN methods may be useful to provide relatively accurate estimations with lower error variability

KEYWORDS:

Missing data, multiple imputations, MAR, MCAR.

I. *INTRODUCTION*

Missing values are a common problem in real datasets. There are many possible explanations for why a data value may be unavailable: the measurements were simply not made, human or machine error in processing a sample, and error in transmitting or storing data values into their respective records and thus different methods for handling this problem have been developed. Several authors ([3],[4],[5]) have demonstrated the dangers of simply removing cases ('list wise deletion') from the original data set, as deletion can introduce substantial biases in the study, specially when missing data is distributed in a not random way. However, this method is practical only when the data contain relatively small number of examples with missing values and when analysis of the complete examples will not lead to serious bias during the inference. In this case predicting missing values is a special data mining prediction problem.

II. *PATTERNS OF MISSINGNESS*

Missing values in the data set is fallen in these two types

a. *Missing Completely At Random (MCAR)*

This is the highest level of randomness. It occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data missing values are randomly distributes across all observations. This is not a realistic assumption for many real time data's.

b. *Missing at Random (MAR)*

When missingness does not depend on the true value of the missing variable, but it might depend on the value of other variables that are observed. This method occur when missing values are not randomly distributed across all observations, rather they are randomly distributed within one or more sub samples

c. *Non-Ignorable* (NI)

NI exists when missing values are not randomly distributed across observations. if the probability that a cell is missing depends on the unobserved value of the missing response, then the process is non-ignorable

III MISSING VALUE TREATING RULE

Any method should satisfy the following points or Rule (i) Estimation without bias. Any missing data treatment method should not change the data distribution. (ii) The relationship among the attributes should be retained (iii) Cost. Minimize the cost.

IV *HANDLING MISSING VALUES*

4.1 Theoretical Framework

The framework in the literature for the applicability of the different methods to handle missing ness is based on a classification according to the following missing ness mechanisms:

• If the probability of an observation being missing does not depend on observed or

unobserved measurements then the observation is Missing Completely At Random (MCAR). A typical example is a patient moving to another city for non-health reasons. Patients who drop-out of a study for this reason could be considered a random sample from the total study population and their characteristics are similar.

• If the probability of an observation being missing depends only on observed measurements then the observation is Missing At Random (MAR). This assumption implies that the behavior of the post drop-out observations can be predicted from the observed variables, and therefore that response can be estimated without bias using exclusively the observed data. For example, when a patient drops out due to lack of efficacy reflected by a series of poor efficacy outcomes that have been observed, the appropriate value to assign to the subsequent efficacy endpoint for this patient can be calculated using the observed data.

• When observations are neither MCAR nor MAR, they are classified as Missing Not At Random (MNAR) or non-ignorable i.e. the probability of an observation being missing depends on unobserved measurements. In this scenario, the value of the unobserved responses depends on information not available for the analysis (i.e. not the values observed previously on the analysis variable or the covariates being used), and thus, future observations cannot be predicted without bias by the model. For example, it may happen that after a series of visits with good outcome, a patient drops-out due to lack of efficacy. In this situation the analysis model based on the observed data, including relevant covariates, is likely to continue to predict a good outcome, but it is usually unreasonable to expect the patient to continue to derive benefit from treatment., it is impossible to be certain whether there is a relationship between missing values and the unobserved outcome variable or to judge whether that missing data can be adequately predicted from the observed data. It is not possible to know whether the MAR, never mind MCAR, assumptions are appropriate in any practical situation. A proposition that no data in a confirmatory clinical trial are MNAR seems implausible. Because it is considered that some data are MNAR, the properties (e.g. bias) of any methods based on MCAR or MAR assumptions cannot be reliably determined for any given dataset.

Therefore the method chosen should not depend primarily on the properties of the method under the MAR or MCAR assumptions but on whether it is considered to provide an appropriately conservative estimate in the circumstances of the trial under consideration.

V. *MISSING VALUES TREATMENT METHODS*

5.1 Types of Methods

Missing value treating method play an important role in the data preprocessing. Methods divided into three categories as proposed by Larid and et al ([1],[3]).

[i] Ignoring Discarding Data. In this method also there two ways to discard the data with missing values

a. Complete case analysis, this method discarding the entire Instance with missing values.

b. Discarding Instances and/or attributes this method determining the level of missing value on each instance and attributes. It deletes the Instance with high level of missing data.

[ii] ***Parameter estimation*** Maximum like hood procedure is used to estimate the parameters of a model defined for the complete data. Maximum like hood procedures that use variants of the Expectation–Maximization algorithm can handle parameter estimation in the presence of missing data [1,2]

[iii] ***Imputation techniques imputation*** is the substitution of some value for a missing data point or a missing component of a data point. Once all missing values have been imputed, the dataset can then be analyzed using standard techniques for complete data. The analysis should ideally take into account that there is a greater degree of uncertainty than if the imputed values had actually been observed, however, and this generally requires some modification of the standard complete-data analysis methods. While many imputation techniques are available Imputation of missing data on a variable is replacing that missing by a value that is drawn from an estimate of the distribution of this variable.

VI . *MISSING VALUE ESTIMATION METHODS*

Randomly simulated missing values were estimated by five data imputation methods:

a) ***k-Nearest Neighbors (k-NN)***

This method the missing values of an instance are imputer considering a given number of instances that are most similar to the instance of interest. The distance is calculate using distance function.

The advantage of this method is

(i) It predicts both quantitative and qualitative attributes (ii) It can easily treat the records with multiple missing values. The disadvantage of this method is searches through all the dataset looking for the most similar instances. This is very time consumable one. (ii) choice of distance function.

b) ***Mean based imputation (Single Imputation)****[7]*

In the *mean imputation*, the mean of the values of an attribute that contains missing data is used to fill in the missing values. In the case of a categorical attribute, the mode, which is the most frequent value, is used instead of the mean. The algorithm imputes missing values for each attribute separately. Mean imputation can be conditional or unconditional, i.e., not conditioned on the values of other variables in the record. Conditional mean method imputes a mean value, that depends on the values of the complete attributes for the incomplete record .

c) ***NORM*** *t*hat implements missing value estimation based on the expectation maximization algorithm [6];

Multiple imputation inference involves three distinct phases:

• The missing data are filled in m times to generate m complete data sets.
• The m complete data sets are analyzed by using standard procedures.
• The results from the m complete data sets are Combined for the inference

d)***LSImpute_Rows,***

LSImpute_Rows method estimates missing values based on the least square error principle and correlation between cases (rows in the input matrix) [7,8].

e)***EMImpute_Columns***.

The EMImpute_Columns estimates missing values using the same imputation model but based on the correlation between features [9] (columns in the input matrix). LSImpute_Rows and EMImpute_Columns involve multiple regressions to make their Predictions

In each dataset missing values were simulated by randomly labeling feature values as missing values. Datasets with different amounts of missing values (from 5% to 35% of the total available data) were generated. For each percentage of missing data 20 random simulations were conducted. The data were standardised using the max difference normalisation procedure which mapped the data into the interval [0..1]. The estimated values were compared to those in the original data set. The average estimation error E was calculated as follows:

$$E = \left[\sum_{k=1}^{m} \left[\left[\sum_{i=i}^{n} \left(| oij - Iij | / \left(\max_j - \min_j \right) \right) \right] / n \right] \right] / m$$

where n is the number of imputed values, m is the number of random simulations for each missing value, 0_{ij} is the original value to be imputed, I_{ij} is the imputed value, j is the corresponding feature to which O_i and I_i belong.

VII *RESULTS*

Figures 1 present the estimation error results obtained from different methods for the databases respectively.
Different *k*-NN estimators were implemented, but only the most accurate model is shown. The *10*-NN models produced an average estimation error that is consistently more accurate than those obtained using the *Mean imputation*, *NORM* and *LSImpute_Rows* methods. Tables 1 and 2 show the average estimated errors

TABLE -2 Average estimated error

Method	5	10	15	20	25	30	35
10-NN	10.2	10.9	11.5	12.4	13.2	14.5	15
NORM	12.4	13.3	12.7	14	14.2	14.7	15.3
EMImpute_Columns	8.1	9	9.5	9.2	9.3	8.2	7.5
LSImpute_Rows	12.3	12.5	13.5	14.3	14.6	13.1	12.9
Mean Imputation	13.6	14	13.5	13.7	13.4	13.7	13.8

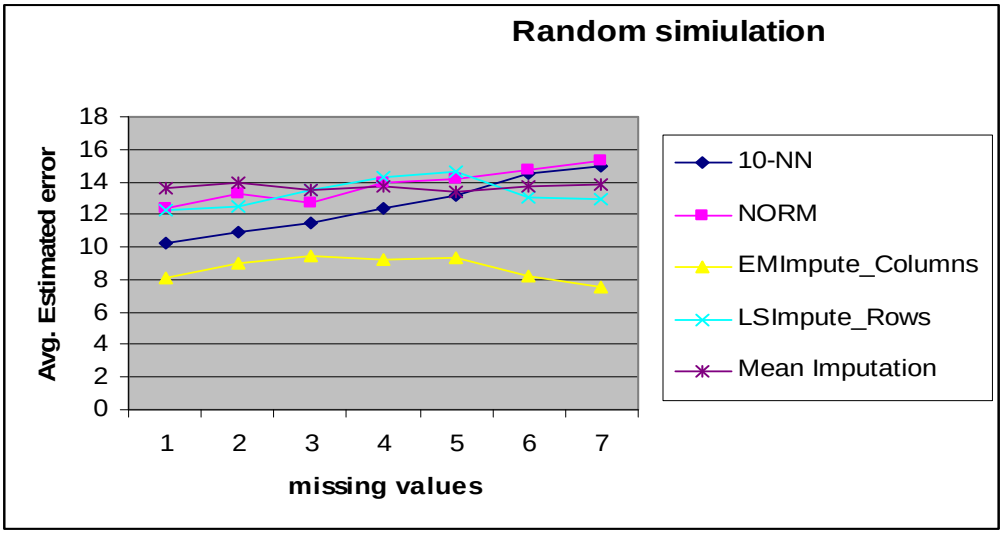

Figure 1: Comparison of different methods using different fractions of missing values in imputation process

Mean imputation surprisingly did not produce the highest estimation errors. Figure 1 show that the EMImpute_Columns produced the best average estimation errors. In comparison to the k-NN, the EMImpute_Columns shows more variable results. The NORM method produced the least accurate estimation results However, Figures 1 do not clearly display a trend about the minimum amount of missing data required for obtaining reliable estimations. The inaccuracy of such estimations is further highlighted in Tables 2 and 3, which indicate that these estimations are highly variable for each method and amount of missing data.

VIII CONCLUSION AND DISCUSSION

In this paper the result of comparison of missing value estimation methods, which is a problem that as received relatively little attention from the medical informatics community. This study is part of the pre-processing phase in the development of systems for assessing risk factor of heart disease patients. This facilities an understanding of the limitations of the data available and possible solutions to address the problem of missing data. In datasets a Feature-based correlation method known as EMImpute_Columns produced the most promising results. The k-NN was able to generate relatively accurate and less variable results for different amounts of missing data, which were assessed using 20 missing value random simulations. However, it is important to remark that, while on the one and this study allowed us to assess the potential of different missing data estimation methods, on the other hand it did not offer significant evidence to describe a relationship between the amount of missing data and the accuracy of the predictions. IN this paper it was constrained by two important factors: the relatively small amounts of data included, and the low number of simulation experiments. Further studies should include more simulations for each amount of missing data. The k-NN method should be adapted to consider incomplete cases. A weighted adaptation may be implemented to reflect relevant correlations between the features. In the case of the EMImpute_Columns method, a control threshold could be set to guide the optimization process, which may facilitate a faster learning convergence. In conclusion, clinical databases often contain a substantial amount of missing data, due to the lack of test results for certain interventions or administrative inaccuracies. When using such a database for classification, it is important to have as complete a data set as possible. If data are imputed, the validity of these values should be assessed. In this paper has compared techniques in the cardiac domain by quantifying the error for the percentage of missing data. As a final caveat, it is important to stress that, due to the relative small amounts of data analyzed and the low number of simulation experiments, this initial study did not offer conclusive evidence to define relationships between the amount of missing data and the accuracy of the predictions. The imputation of missing data is of particular importance when, classification will address a subset of the database (i.e. a reduced case base)

REFERENCES

[1] A.P.Dempster, R.J Larid, D.B Rubin,"Maximum likelihood from imcomplete data via the Em Algorithm (with discussion)" Journal of royal Statistical society vol.B39, pp. 1-38, 1977.
[2]. R.Mehala, P.Ranjit Jeba Thangaiah and K.vivekanandan., "Selecting Scalable Algorithms to Deal with Missing Values", Poster Paper. In International Journal of Recent Trends in Engineering, Vol 1, No.2, May 2009
[3] R. J. Little and D. B. Rubin. Statistical Analysis with Missing Data. John Wiley and Sons, New York, 1997.
[4] D. Rubin. Multiple Imputation for Nonresponse Surveys. Wiley & Sons Inc., 2004.

[5] R.Kavitha Kumar & R.M.Chadrasekar," Data Preprocessing methods and unified Framework for a cardiac Database", in CIIT international journal on Knowledge engineering. ISSN 0974 – 9683, 2009
[6] J.L. Schafer, NORM: multiple Imputations of incomplete multivariate data under a normal model, version 2.03, software for Windows 95, 98, NT, Website: http;//www.stat.psu.edu/~jls/misoftwa.html, 1999.
[7] Liu P., Lei L. and Zhang X.F., "A Comparison Study of Missing Value Processing Methods" , Computer Science, 31(10): 155-156, 2004.
[8] Jos´ e M. Jerez, Ignacio Molina, Jos´ e L. Subirats, Leonardo Franco. "Missing Data Imputation in breast Cancer Prognosis" , IASTED,2006
[9] Marisol Giardina,Yongyang Huo1, Francisco Azuaje, Paul McCullagh , Roy Harper. "A Missing Data Estimation Analysis in Type II Diabetes Databases", CBMS'05, 1063/05.
[10] Liu Peng and Lei Lei , "A Review of missing Data Treatment Methods"
[11] Bhekisipho Twala, Michelle Cartwright and Martin Shepperd, "Comparison of Various Methods for Handling Incomplete Data in Software Engineering Databases" ,in 0-7803-9508-5, 2005 IEEE.

AUTHORS PROFILE

R. Kavitha Kumar Working as a programmer in Department of Computer Science and Engineering , Pondicherry Engineering College., Puducherry, India. Has 11 years teaching experience and 3 years Inducstry Experience as a Programmer. Did M.Phil (Computer Science) Degree in Mother Tersa Women's University ., Kodaikanal, India. M.Sc (Computer Science) in Bharathidasan University, Trichy, India. B.Sc (Computer Science) in Madras University, India. At present pursing Ph.D (Compter Science) in Mother Tersa Women's University ., Kodaikanal, India.

Dr.R.M.Chandrasekar Working as Professor in Annamalai University , Chidambaram, India., has 17 years teaching experience, 3 years worked as a Registrar ,in Trichy Anna University, Trichy . 2 years Worked as software consultant in USA. Area of interest and guiding in Data Mining, Image Mining, Software Slicing, Document Image Segmentation