Classification of Incomplete Data Handling Techniques – An Overview

N.C. Vinod,
Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Dr. M. Punithavalli,
Research Supervisor, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Abstract-The task of classification with incomplete data is a complex phenomena and its performance depends upon the method selected for handling the missing data. Missing data occur in datasets when no data value is stored for an attribute / feature in the dataset. This paper provides a brief overview to the problem of incomplete data handling techniques and discusses the various methods used with classification and missing data.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a technology with great potential that predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. It is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing and data visualization [11].

Fayyad and Simoudis [6] have defines data mining process as a five step procedure. The first step is selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined. The second step is preprocessing. This is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries. In this step, storage of unnecessary values (Example : gender details of a patient when studying pregnancy), out-of-range values (Example : Salary 100), missing values, and data values which in general lead to misleading errors, are identified and attempts to correct these problematic data are made. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may recorded as f or m and also as 1 or 0. The third step transforms the cleaned data to a format which is readily usable and navigable by the data mining techniques. The fourth stage is concerned with using data mining techniques for the extraction of patterns from the transformed dataset. The discovered knowledge is then interpreted and evaluated for human decision-making in the last step.

Out of this, the preprocessing stage, in particular, handling of missing values in datasets is considered very critical. Missing values occur when no data value is stored for an attribute / feature in the dataset. Missing values are a common occurrence and it can severely disturb the conclusions drawn from the data if handled inappropriately in empirical research. Missing values may occur because of non-availability of information for several items or whole unit. Non-availability of data is sensitive in various applications like database storing information about private subjects items such as income. Dropout is a type of missingness that occurs mostly when studying development over time. In this type of study the measurement is repeated after a certain period of time. Missingness occurs when participants drop out before the test ends and one or more measurements are missing (http://www.wikipedia.org). Sometimes missing values are caused by the researchers themselves. For example, when data collection is not done properly or when mistakes were made in data entry [1]. And a great deal of missing data arise in cross-national research in economics, sociology, and political science because governments choose not to, or fail to, report critical statistics for one or more years [12].

Generally, missing values can occur in datasets in different forms. They can be classified into three categories and a clear knowledge on which category the missing values likes is a clear step towards a positive solution.

(i) Missing values occur in several attributes (columns)
(ii) Missing values occur in a number of instances (rows)
(iii) Missing values occur randomly in attributes and instances

Methods used for each of these category differs, therefore selection of correct algorithm is significant. Normally, missing rates less than one per cent are considered trivial, 1-5 per cent are considered manageable. But databases with 5-15% missing data values rate needs sophisticated methods to handle them correctly and more than 15% requires careful handling as they affect interpretation. It is in the last category most of the solutions have been proposed and it is understood that several alternative ways of dealing with missing data exists.

Cautious handling of missing or incomplete values is also considered important in the field of classification. Classification is the process of dividing a dataset into mutually exclusive groups such that the members of each group are as “close” as possible to one another, and different groups are as “far” as possible from one another, where distance is measured with respect to specific variable(s) that are used for prediction. Mishandling missing values during classification produces may produce erroneous classification results. This paper is an attempt to outline these approaches.

The paper is organized as follows : Section II provides an overview of incomplete data. Section III studies the various techniques that are offered as solutions to classification. Section IV provides a conclusion with future research directions.

II. OVERVIEW OF MISSING VALUES

Efficient treatment of missing values requires a complete understanding behind it. This section outlines some fundamental aspects of incomplete or missing values.

A. *Types of Incomplete Data*

Little and Rubin [10] define a list of missing mechanisms, which are widely accepted by the community.

1) Missing completely at random (MCAR) : MCAR is the probability that an observation (X_i) is missing, is unrelated to the value of X_i or to the value of any other variables and the reason for missing is completely random. Typical examples of MCAR are when a tube containing a blood sample of a study subject is broken by accident (such that the blood parameters can not be measured) or when a questionnaire of a study subject is accidentally lost [5]. This situation is rare in real world and is usually discussed in statistical theory.

2) Missing at random (MAR) : MAR is the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the unobserved data. An example of this is accidentally or deliberately skipping an answer on a questionnaire by the participant. This mechanism is common in practice and is generally considered as the default type of missing data.

3) Not missing at random (NMAR). If the probability that an observation is missing depends on information that is not observed, this type of missing data is called NMAR. For example, high incomers may be more reluctant to provide their income information [5]. This situation is relatively complicated and there is no universal solution.

B. *Statistical Framework of Incomplete Data*

The statistical framework of incomplete or missing data is presented in this section and is based on [10]. In this framework, the dataset is denoted as X have N items ($x_1, x_2, ..., x_N$), which is composed of two components, namely, observed components (x_o) and missing component (x_m). The framework considers a random process for both data generation and missing data mechanism with joint probability distribution as given in Equation (1).

$$P(X, R|\theta, \phi) = P(X|\theta)\, P(R|X, \phi) \qquad (1)$$

where θ is data generation process and ϕ for missing data mechanism. The notion of missing data mechanism can be formalized using a missing data indicator matrix R (Equation 2)

$$R_{ij} = \begin{cases} 1, & \text{Observed} \\ 0, & \text{Mis sin g} \end{cases} \qquad (2)$$

Once the data probability model is decomposed using Equation (1), the next step is to identify the type of missing data. That is, whether the missing data is MCAR or MAR or NMAR. This, as mentioned previously, is critical while evaluating algorithms for handling incomplete data. A fact that has to be stressed at this point is that methods that work for MCAR need not necessarily work for MAR.

III. TECHNIQUES USED FOR MISSING VALUES

Dealing with missing values means to find an approach that can fill them and maintain (or approximate) as closely as possible the original distribution of the data. In this section, the various methods used are discussed.

A. *Place of Implementation during Mining*

Generally, the methods that deal with missing values can be implemented at two stages [7]. They are,

(1) Before mining (Pre-replacing methods) and
(2) During mining (Embedded methods).

Pre-replacing methods replace missing values before the data mining process, while embedded methods deal with missing values during or along with the data mining process. Pre-replacing methods are either statistics-based or machine-based. Examples to statistic-based methods include linear regression, replacement under same standard deviation [8] and mean-mode method [17]. Nearest-Neighbour estimator, auto-associative neural network [8] and decision tree imputation [4] are some examples for machine-based methods. Case-wise deletion [15], lazy decision tree [16], dynamic path generation [9], C4.5 and CART are some examples for Embedded methods. Table I presents a summarization of these techniques while evalauting them in terms of their group, computation cost, attribute types and missing value cases applicable [7]. In the table, case 1 represents missing values that occur in several attributes (columns) and case 2 represents missing values that occur in a number of instances (rows).

B. Machine Learning Classification Methods with Missing Values

The pattern of missing values is an important characteristic that plays an vital role in the performance of a classifier. The problem of classification with missing data generally involves two steps. They are

(i) Handling missing values and
(ii) Classification.

TABLE I : Comparative Evaluation

Method	Cost	Attribute Type*	Case
Pre-replacing Methods			
Mean-and-mode	Low	No. and Char	Case 2
Linear regression	Low	No.	Case 2
Standard deviation	Low	No.	Case 2
Nearest Neighbour Estimator	High	No. and Char	Case 1
Decision tree Imputation	Middle	Char	Case 1
Auto-associative neural network	High	No. and Char	Case 1
Embedded Methods			
Casewise deletion	Low	No. and Char	Case 2
Lazy decision tree	High	No. and Char	Case 1
Dynamic path generation	High	No. and Char	Case 1
C4.5	Middle	No. and Char	Case 1
Surrogate split	Middle	No. and Char	Case 1

* No. – Number; Char – Multiple Attribute types

Depending on the method used for both these steps, the techniques can be grouped into four main categories as given below (Figure 1).

A1) Deletion of missing values (complete cases and available data analysis), and classifier design using only the complete instances,
A2) Imputation (estimation and replacement) of missing input values, and after that, another machine learns the classification task using the edited complete set, i.e., complete instances and incomplete patterns with imputed values,
A3) Use of Maximum Likelihood (ML) approaches, where the input data distribution is modeled by the Expectation-Maximization (EM) algorithm, and the classification is performed by means of the Bayes rule,
A4) Use of machine learning procedures able to handle missing data without an explicit imputation.

In Figure 1, in the two first types of approaches (data deletion and imputation), the two steps, handling missing values and classification, are solved separately. In contrast, the third type of approaches model the Probability Density Function (PDF) of the input data (complete and incomplete cases), which is used to classify using the Bayes-decision theory. Finally, in the last kind of approaches, the classifier has been designed for handling missing values without a previous missing data imputation. However, a drawback of these learning machines is that they cannot handle input vectors that present missing data on its features [13].

C. Other Methods

This section discusses three techniques that are less frequently used. The reason behind their infrequent usage is its poor classification performance when presented with a datasets with missing data. They are,

(i) Hot deck imputation

(ii) Mean substitution
(iii) Regression substitution

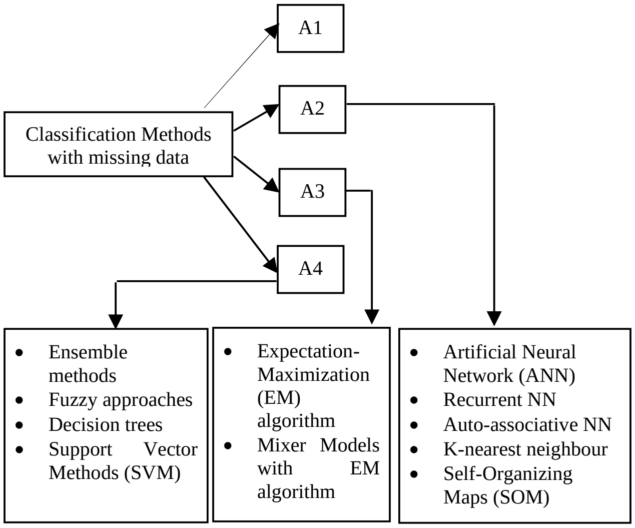
A1
A2
Classification Methods with missing data
A3
A4
Ensemble methods
Fuzzy approaches
Decision trees
Support Vector Methods (SVM)
Expectation-Maximization (EM) algorithm
Mixer Models with EM algorithm
Artificial Neural Network (ANN)
Recurrent NN
Auto-associative NN
K-nearest neighbour
Self-Organizing Maps (SOM)

Figure 1 : Machine Learning Classification methods with missing data

Hot deck imputation was used successfully by the Census Bureau during 1940's and '50's, at which time, encountering missing data was an infrequent situation. In hot-deck imputation, a missing value was imputed from a randomly selected similar record. The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed. Cold-deck imputation, by contrast, selects donors from another dataset. Since computer power has advanced rapidly and punched cards are no longer used, more sophisticated methods of imputation have generally superseded the original random and sorted hot deck imputation techniques, such as nearest neighbour hot deck imputation and the approximate Bayesian bootstrap. Scheuren [14] has an interesting discussion of how the process was developed within the U. S. Census Bureau. Similarly, [2] discusses the use of Hot-deck imputation for processing census information.

Mean substitution method substitute the missing data with its mean value. For example, while maintaining systolic blood pressure of persons, if a value of a person is missing, then the mean value of all the available systolic blood pressure will be calculated and will be used to fill the missing value. The disadvantage of this approach is that the overall mean, with or without replacing missing data, will be the same. In addition, such a process leads to an underestimate of error as the number of samples is increased, without adding new information. Cohen et al. (2003) gave an interesting example of a data set on university faculty. The data consisted of data on salary and citation level of publications. There were 62 cases with complete data and 7 cases for which the citation index was missing. Cohen gives the following table.

Regression substitution uses linear regression to predict missing value on the basis of other values present in a dataset. This approach was quite famous in early 1980’s and has one advantage over mean substitution. The missing value computed depends on the other values of the missing data. For example, with mean substitution, if a male person’s weight is missing, then it is replaced by the average weight. But in regression substitution, the average weight of male persons would be calculated and will be used for substitution. By substituting a value that is perfectly predictable from other variables, more information is added, thus increasing the sample size and reducing the standard error.

IV. CONCLUSION

This paper discussed the various approaches used in classification with missing values. It could be seen that both statistical approaches and machine learning approaches have been successful to a certain extent in the problem domain under discussion. While considering the missing data imputation approaches based on machine learning, artificial neural network algorithms, K-Nearest Neighbour algorithm and Self-Organizing Maps (SOM) are more frequently used. Several variants of SOM like tree-structured SOM are also in existence. EM algorithm is frequently used while considering maximum likelihood based approaches. Decision trees and fuzzy approaches have also been studied. In spite of these studies, it is understood that hundred per cent success is still seen only as a distant possibility because of the numerous factors influencing the relative success of the competing techniques. Currently, no one method can be used for handling all types of missing data problem and the only right answer, as opined by [3] for missing data procedures, “we return to the old precept that still holds true: The only real cure for missing data is to not have any”. However, with the growing database size and complexity in the data attributes, missing value handling procedures is a mandatory process. Different approaches suit different datasets, which should be selected according to the property of the dataset at hand as well as the requirement on algorithm complexity and efficiency. Moreover, a previous analysis of the classification problem to be solved is very important in order to select the most suitable missing data treatment. The various techniques identified in this study, in future, can be compared with respect to their performance in classification accuracy while provided with incomplete datasets.

REFERENCES

[1] Adèr, H.J. and Mellenbergh, G.J. (Eds.) (2008) Chapter 13: Missing data, Advising on Research Methods: A consultant's companion, Huizen, The Netherlands: Johannes van Kessel Publishing, Pp. 305-332.

[2] Altmayer, L. (2010) Hot-Deck Imputation: A simple data stepapproach, http://analytics.ncsu.edu/sesug/1999/075.pdf

[3] Anderson, A.B., Basilevsky, A. and Hum, D.P.J. (1983) Missing data: A review of the literature, P.H. Rossi, Wright, J.D. and A.B. Anderson (Eds.), Handbook of survey research, San Diego: Academic Press, Pp.415-494.

[4] Chen, J. and Shao, J., (2001) Jackknife variance estimation for nearest-neighbor imputation. J. Amer. Statist. Assoc., Vol.96, Pp. 260-269.

[5] Donders, A., van der Heijden, G., Stijnen, T. and Moons, K. (2006) Review: a gentle introduction to imputation of missing values, Journal of Clinical Epidemiology, Vol. 59, Pp. 1087-1091.

[6] Fayyad, U.M., Shapiro, G.P. and Smyth, P. (1996) Data Mining and Knowledge Discovery in Databases: An overview, Communications of ACM, Vol. 39, No. 11, P. 27-34.

[7] Fujikawa, Y. and Ho, T. (2002) Cluster-based algorithms for dealing with missing values, M.S. Chen, Yu, P.S. and Liu, B. (Eds.), PAKDD 2002, LNAI 2336, Springer-Verlag, Pp. 549-554.

[8] Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2nd edition.

[9] Lall, U. and Sharma, A., (1996) A nearest-neighbor bootstrap for resampling hydrologic time series, Water Resource. Res., Vol.32, Pp.679–693.

[10] Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data, 2nd Edition, John Wiley and Sons, New York.

[11] Martin, T. (2003) A day in the life of a Data Miner, Bulletin of the International Statistical Institute, 54th Session, Vol. LX, Invited Papers, August 2003, Berlin, Germany. Pp. 298-301.

[12] Messner, S.F. (1992) Exploring the Consequences of Erratic Data Reporting for Cross-National Research on Homicide. Journal of Quantitative Criminology, Vol.8, No.2, Pp. 155-173.

[13] Sancho-Gomez, J., Garcia-Laencina, P.J. and Figueiras-Vidal, A.R. (2009) Combining missing data imputation and pattern classification in a multi-layer perceptron, Inteligent Automation and Soft Computing, Vol. 15, No. 4, Pp. 539-553.

[14] Scheuren, F. (2005) Multiple imputation: How it began and continues, The American Statistician, Vol. 59, Pp. 315-319.

[15] Zhang, C.Q., et al., (2007) An Imputation Method for Missing Values. PAKDD, LNAI 4426, Pp. 1080–1087.

[16] Zhang, S.C., et al., (2004) Information Enhancement for Data Mining, IEEE Intelligent Systems, Vol. 19, No.2,Pp. 12-13

[17] Zhang, S.C., et al., (2005) Missing is useful: Missing values in cost-sensitive decision trees, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No.12, Pp. 1689-1693.