International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), August 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SURVEY ON CLASSIFICATION OF INCOMPLETE DATA HANDLING TECHNIQUES

Dr.C.Yamini[1], M.Kowsalya[2]

[1]Associate Professor, [2]Research Scholar, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore, (India)

ABSTRACT

Data is often incomplete. Classification with incomplete data is a new subject. This study proposes a classification for incomplete survey data. 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. It proposes a various techniques of classification use of incomplete data.

Keywords: Classification, Incomplete Data, Missing Values.

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, and it can generate results which come out to be significant and which cannot actually predict future behavior and cannot be reproduced on a new sample of data and allow small use. It is 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. An incomplete data may severely affect the quality of learned patterns and the performance of algorithms. As a result, how to properly handle incomplete data is an important and challenging problem in the practice of machine learning and data mining.

There are many approaches to handling incomplete data as far as classification is concerned, from simply removing samples or features with missing values to completing the original data set by filling in specific values. In this paper, focused on the classification of incomplete data. However, the deletion of samples or features may result in the loss of useful information especially when a large portion of samples or features have missing values. In this paper, a novel scheme is developed for conducting classification on incomplete data with applying various techniques.

Pattern classification was developed starting from the 1960s. It progressed to a great extent in parallel with the growth of research on knowledge-based systems and artificial neural networks. Increasing computational resources, while enabling faster processing of huge data sets, have also facilitated the research on pattern classification, providing new developments of methodology and applications. This interdisciplinary field has been successfully applied in several scientific areas such as computer science, engineering, statistics, biology,

and medicine, among others. These applications include biometrics (personal identification based on several physical attributes such as fingerprints and iris), medical diagnosis (CAD, computer aided diagnosis), financial index prediction, and industrial automation (fault detection in industrial process).

The complete goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The final part of this paper is organized as follows. An overview of the previous research on incomplete data is given in Section II. Section III introduces the details of the proposed method for handling incomplete data. Section IV provides a conclusion with future research directions.

II. AN OVERVIEW OF INCOMPLETE DATA

Missing values are a common occurrence, and need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in.

Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. In incomplete datasets, some values are missing for one or more features. When choosing the right techniques for dealing with this issue, it is necessary to have a good understanding of different reasons that lead to incomplete data. Efficient treatment of missing values requires a complete understanding behind it.

2.1 Types of Incomplete Data

Little and Rubin [10] define a list of missing mechanisms, which are widely accepted by the community. There are three mechanisms under which missing data can occur:

1) Missing completely at random (MCAR): MCAR is the probability that an observation (Xi) is missing, is unrelated to the value of Xi 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 cannot 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.

2.2 Review of Work on Incomplete Data

In the past decades, significant efforts have been devoted to this area from the point of view of statistical theory, machine learning and so on. Various methods for handling incomplete data have been introduced and these methods can be summarized as follows: samples or features deletion, missing values imputation and learning with missing data.

2.2.1 Samples or Features Deletion

Samples or features with missing values are simply removed from the dataset. This method is easy to implement and usually performs well when the missing rate is low. However, it is obvious that it may ignore some potentially valuable information and create bias in the dataset.

2.2.2 Imputation of Missing Values

Most studies on incomplete data focus on imputation. Imputation techniques are based on the idea that any subject in a study sample can be replaced by a new randomly chosen subject from the same source population. The imputation of missing data of a feature is to generate values drawn from an estimate of the distribution of this variable. Common imputation schemes include completing missing data with specific values such as the unconditional mean or the conditional mean (if one has an estimate for the distribution of missing features given the observed features).

2.2.3 Learning with Missing Data

Some classifiers can be customized in order to handle incomplete data directly, such as Artificial Neural Network (ANN), C4.5 decision trees, Bayesian Networks (BN) , Rough sets and Logistic regression algorithm. Generally speaking, 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.

2.3 Statistical Framework of Incomplete Data

The statistical framework of incomplete or missing data is present based on Little and Rubin (1987).. In this framework, the dataset is denoted as X have N items (x1, x2, ..., xN), which is composed of two components, namely, observed components (xo) and missing component (xm). The framework considers a random process for both data generation and missing data mechanism with joint probability distribution as given in Equation (1).

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

III. METHODOLOGY

Data mining process as a five step procedure. The first step declares the selecting or segmenting the data according to some criteria e.g. all people who own vehicle, in this way subsets of the data can be determined.

The second step is preprocessing. This is the data cleaning stage where certain information is removed which is judged 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.

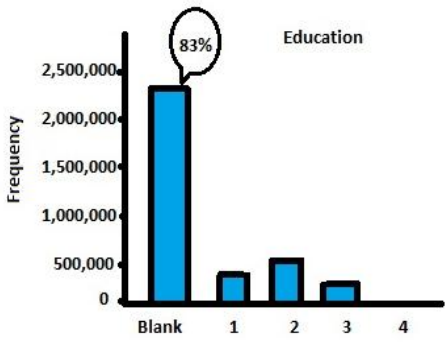
The fourth stage is concerned with using data mining techniques for the extraction of patterns from the transformed dataset. The discovered knowledge is then interpreted and evaluated for human decision-making in the last step.

Incomplete data problems usually occur in areas such as social sciences, bank or shop surveys and medical research. Suppose a set of diagnostic data of patients and normal people among different hospitals in different areas has been collected. The dataset is likely to be incomplete due to several reasons.

Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values.

In general, pattern classification with missing data concerns two different problems, handling missing values and pattern classification. Most of the approaches in the literature can be grouped into four different types depending on how both problems are solved,

- Deletion of incomplete cases and classifier design using only the complete data portion.
- Imputation or estimation of missing data and learning of the classification problem using the edited set, i.e.,
- Complete data portion and incomplete patterns with imputed values.
- Use of model-based procedures, where the data distribution is modeled by means of some procedures, e.g., by expectation–maximization (EM) algorithm.
- Use of machine learning procedures, where missing values are incorporated to the classifier.



In the two-first types of approaches, the two problems, handling missing values (data deletion and imputation, in each case, respectively) and pattern classification, are solved separately; in contrast, the third type of approach 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 incomplete input data without a previous estimation of missing data. Our main goal is to show the most representative and useful procedures for handling missing data in classification problems, with a special emphasis on solutions based on machine learning. Due to space constraints, are not able to provide a complete and detailed study of proposed solutions for incomplete data classification. Thus, this review paper provides a wide and general overview of the state-of the- art in this field. The remainder of this paper is structured as follows.

This work reviews the most important missing data techniques in pattern classification, trying to highlight their advantages and disadvantages. An excellent reference for missing data is the book written by little and Rubin [16], which gives an accurate mathematical and statistical background in this field.

Missing value replacement policies:

i. Ignore the records with missing values.
ii. Replace them with a global constant.
iii. Fill the missing value based on domain knowledge.
iv. Replace them with a mean or frequent value.
v. Use modeling techniques such as nearest neibours, Bay's rule, and Decision tree.

IV. 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.

4.1 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.

4.2 Machine Learning Classification Methods with Missing Values

The pattern of missing values is an important characteristic that plays a 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 Pre-Replacing Methods to Deal with Missing Values

Method	Computation Cost	Attributes
Mean-mode method	Low	Num & Cat
Linear Regression	Low	Num
Standard Deviation	Low	Num
Nearest Neighbor Estimator	High	Num & Cat
Decision Tree imputation	Middle	Char
Auto Associative Neural Network	High	Num & Cat

Table II: Comparative Evaluation Embedded Methods to Deal with Missing Values

Method	Computation Cost	Attributes
Case wise Deletion	Low	Num & Cat
Lazy Decision Tree	High	Num & Cat
Dynamic Path Generation	High	Num & Cat
C4.5	Middle	Num & Cat
Surrogate split	Middle	Num & Cat

In the above tables describes large dataset with missing values, complicated methods are not suitable because of their high computational cost .Use simple methods that can reach performance as good as complicated ones.

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

[1]. Deletion of missing values (complete cases and available data analysis), and classifier design using only the complete instances,

[2]. 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,

[3]. 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,

[4]. Use of machine learning procedures able to handle missing data without an explicit imputation.

4.3 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

Machine Learning Classification Methods

- Ensemble methods
- Fuzzy approaches
- Decision trees
- Support Vector
- Methods (SVM)
- Expectation-Maximization(EM) algorithm
- Mixer Models with EM algorithm

V. CONCLUSION

The major purpose of this paper discussed the various approaches used in classification with incomplete data values.

Missing or incomplete data is a usual drawback in many real-world applications of pattern classification. Data may contain unknown features due to different reasons, e.g., sensor failures producing a distorted or immeasurable value, data occlusion by noise, non-response in surveys.

Handling missing data has become a fundamental requirement for pattern classification because an inappropriate missing data treatment may cause large errors or false classification results. 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 Neighbor 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, "Rreturn 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.

Researchers and practitioners often face missing values when applying learned models. This study provides a valuable step toward understanding how best to deal with them, and why.

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

[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 imputations: 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]. Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New Jersey

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

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

[19]. 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.

[20]. Duda RO, Hart PE, Stork DG (2000) Pattern classification. Wiley, New York