Imputation Methods for Incomplete Data

Vaishali H. Umathe
Department of Computer Technology
Y.C.C.E, Nagpur (M.S.)- 441110, India
vaishaliumathe@gmail.com

Prof. Gauri Chaudhary
Department of Computer Technology
Y.C.C.E, Nagpur (M.S.)- 441110, India
chaudhary_gauri@yahoo.com

Abstract - **Although sometimes encounter data sets that contain one or more missing feature values (incomplete data). Many existing industrial and research data sets contain missing values due to various reasons, such as manual data entry procedures, equipment errors and incorrect measurements. The important factor for selection of approach to missing values is missing data mechanism. There are various strategies for dealing with missing values. Some analytical methods have their own approach to handle missing values. Finally missing values problem can be handled by missing values imputation. This paper presents simple methods for missing values imputation like EMSI and MMSI.**

***Index Terms* — Incomplete data, missing data, missing values imputation, missing data mechanisms.**

I. INTRODUCTION

Any datum with some (but not all) missing feature values is referred to as an incomplete datum. An example of an incomplete datum is $a_5=(1.23,3.8,?,5.6,?)^T$, where a_{53} and a_{55} are missing.

A data set contain at least one incomplete datum is called as an incomplete data set; otherwise, it is called complete data set. Missing data are an often unavoidable problem when analyzing data[11]. In many situations standard analyses of the data are affected by the problem of missing values. Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns. Missing values increases the difficulty of extracting useful information from that data set. Missing data is the absence of data items that hide some information that may be important. Most of the real world databases are characterized by an unavoidable problem of incompleteness, in terms of missing or erroneous values.

Imputation methods involve replacing missing values with estimated or calculated ones based on nonmissing values available in the dataset. Because missing data element can create problems for analyzing data, imputation is seen as a way to avoid drawback involved in listwise deletion of cases that have missing values. When one or more data elements are missing in data set, most statistical packages default to delete any datum that has a missing data element,which may introduce bias or affect the representation of the results. Imputation preserves all datum by replacing missing data elements with a probable value based on other available information in data set. Once all missing data elements have been imputed, the data set can then be analyzed using standard techniques for complete data. Imputation methods can be divided into single and multiple imputation methods[6]. In single imputation the missing value is replaced with one imputed value and in multiple imputation, several values are used.

Multiple imputation method (Rubin 1987) replaces each missing data element with a set of possible values that represent the probability about the right value to impute. The completed data sets are analyzed by using standard methods for complete data and combining the results from these analysis. This process results in valid statistical inferences that properly reflect the uncertainty due to missing data element.
Multiple imputation involves following distinct phases:

- The missing data elements are filled in *x* times to generate *x* complete data sets.
- The complete data set *x* is analyzed by using standard techniques.
- The results from the *x* complete data sets are combined for the analysis.

There are two common solutions to the problem of incomplete data or missing value that are currently applied by software engineering researchers. The first solution includes discarding the instances having missing values (i.e. listwise deletion),which not only does it reduce the sample sizes available for analysis it also ignores the mechanism causing the missingness. A smaller sample size gives greater possibility of a insignificant result, i.e., the larger the data set the greater the statistical power of the test. The second solution is imputes or estimate missing values from the available data. There are three different mechanism of missing data induction

1. Missing completely at random (MCAR)
   When distribution of dataset having a missing value for feature attribute does not depend on observed data as well as on the missing data.
2. Missing at random (MAR)
   When the distribution of dataset having a missing value for feature attribute depends on observed data, but does not depend on the missing data.
3. Not missing at random(NMAR)
   When the distribution of dataset having a missing value for feature attribute depends upon the missing values.

978-1-4799-6818-3/15/$31.00 © 2015 IEEE

Three types of problems are usually associated with missing values

1. Efficiency losses
2. Difficulties in handling and analyzing the data.
3. Bias differences between missing and complete data from results.

II. Literature Review

In this paper various imputation techniques are studied, these techniques are used for the replacing missing values with estimated real value. The need of addressing the problem incomplete data or missing values in dataset is because of its adverse effect on result of clustering.

Richard J. Hathaway and James C. Bezdek introduced four strategies for clustering incomplete data sets. Three of these consist of new adaptations of the fuzzy -means (FCM) algorithm and all four provide estimates of the locations of cluster centers and fuzzy partitions of the data. The four strategies are Whole Data Strategy (WDS), Partial Distance Strategy (PDS), Optimal Completion Strategy (OCS), Nearest Prototype Strategy (NPS). If the proportion of incomplete data is small, then whole data strategy may be useful to simply delete all incomplete data. But proportion of incomplete data is large then other three methods are used. The PDS FCM uses an approach recommended by Dixon to do the calculations in such a way that uses all available information. Comparison is done between these four methods on basis of % missing and mean number of iteration to termination [11].

The comparison between statistical representation of missing attributes (sr-fcm) technique and wds-fcm, pds-fcm, ocs-fcm, nps-fcm. The problem of missing data handling for fuzzy clustering is considered, and a statistical representation of missing attributes is proposed. A statistical analysis of missing attributes is given with the aim of imputation, which reduces the statistical analysis. The performance of fuzzy clustering is improved based on the recovered data. In this introduced the drawbacks of four missing value calculation methods, these are wds,pds,ocs and nps.[3]

The Fuzzy C-Means (FCM) algorithm is one of algorithm used for clustering. The FCM algorithm performance depends on the selection of the initial cluster center as well as the initial membership value. The problems of the unknown number of clusters and initialization of prototypes in the FCM clustering algorithm for symbolic interval-values data. To overcome the above problems, the concepts of competitive agglomeration clustering algorithm is incorporated into FCM clustering algorithm for symbolic interval-values data. The proposed approach is called as FCMwUNC clustering algorithm [2]. A FCM clustering algorithm that handles mixed data containing missing values, the mixed data is combination of numerical and categorical. The First applied the imputation method to missing categorical data before clustering, and then used the FCM clustering algorithm. When encountered numerical missing data, the PDS distance is used for numerical missing data. Most missing data imputations are restricted to numerical data but Takashi Furukawa, Shin-ichi Ohnishi and Takahiro Yamanoi are stated the imputation method for numerical as well categorical data.[7]

A very important issue faced by researchers who use industrial and research datasets is incompleteness of data. Handling incomplete data is an important issue for classification since incomplete data is present in either the training data or test (unknown) data affect the prediction of accuracy of classification. Incomplete data could be caused by unit nonresponse (where no data could be collected from the sampled unit) or item nonresponse (where data is collected for the unit, but some items have missing values). The *k*-NN approach to determine the imputed data, where nearest is usually defined in terms of a distance function based on the auxiliary variable. Bhekisipho Twala and Michelle Cartwright is stated that BAMINNSI consistently takes more time to train and test. The robustness of two current imputation methods and further introduce a new ensemble method based on the two imputation methods [4]

Bhekisipho Twala, Michelle Cartwright and Martin Shepperd represented the imputation methods involve replacing missing values with estimated or calculated ones based on available values available in the dataset. Various imputation methods are explained on the basis of single imputation and multiple imputation like dtsi, knnsi,mmsi,emsi,emmi,fc,svs. One other common method to avoid losing data due to LD is the mean or mode substitution of missing data. With this procedure, whenever a value is missing for one instance on a particular attribute, the mean (for an ordered attribute) or mode (for a nominal attribute), based on all nonmissing instances, is used in place of the missing value. Multiple imputation (MI) represents superior approach to handling missing data. EMMI is the best-performance method [6].

III. Single Imputation techniques

In this section we describe two methods for missing values calculation.

A. Mean or mode single imputation(MMSI)

MMSI is one of the most frequently used method for imputation. It consists of replacing the missing data element for a given feature (attribute) by the mean of all known available values of that attribute in the class where the instance with missing data element attribute belongs. If assume the existence of true value for each unknown one, can try to estimate this true value based on the known information. The simple approach for utilizing data-dependent information is to replace unknown value of continuous attribute by their average values(the mean value). Mean substitution has the advantage of returning a complete data set, so estimates are based off of the same cases included in each analysis. The advantage of mean imputation is easy to implement.

B. Expectation maximization single imputation(EMSI)

In short, EM is an iterative methodology. In the E-step, one peruses in the information, one occurrence at once. As every

case is perused in, one adds to the figuring of the sufficient measurements (wholes, aggregates of squares, entireties of cross items). In the event that missing qualities are accessible for the occurrence, they add to these wholes specifically. On the off chance that a variable is lost for the occurrence, then the best figure is utilized as a part of spot of the missing worth. In the M-step, once all the aggregates have been gathered, the covariance network can just be computed. This two stage procedure proceeds until the change in covariance network from one cycle to the next gets to be inconsequentially little. Subtle elements of the EM calculation for covariance networks are given in. EM obliges that information are absent at irregular.

IV. Experimental Analysis

In this section we compare MMSI and EMSI imputation methods using artificially generated incomplete data sets. The scheme for artificially generating an incomplete data set is described.

A. IRIS Data Set

IRIS dataset includes 150 4-dimension pattern which belong to three physical categories (Setosa, Versicolor, Virginica), that's to clustering analysis is to delete all data which have missing say, one category has 50 samples. There is 4 numeric, predictive attributes and the class. The attribute information is sepal length, sepal width, petal length, petal width.

B. Incomplete dataset

The IRIS dataset containing one or more missing values which are generated artificially. In fig2, NULL indicates the missing value in particular attribute. The system read each and every data elements row wise for missing values in data set. If any data element has missing values is indicated with Null and after that give to the imputation method for calculation.

id	attribute1	attribute2	attribute3	attribute4
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	NULL	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	NULL	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3	1.4	0.1
14	4.3	3	1.1	0.1
15	5.8	4	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3

Fig2: Dataset with missing value

C. Complete dataset

The fig3 shows complete dataset. The missing values are replaced with estimated calculated value. The mean or mode single imputation method is used for missing value calculation.

id	attribute1	attribute2	attribute3	attribute4
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	2.0458136	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	2.7404826	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3	1.4	0.1

Fig 3:Complete dataset

V. Conclusion and Future Work

Missing data is a critical issue for researchers. Selecting methods for handling missing data is difficult, since the same procedure can give higher predictive accuracy rates in certain circumstances and not in others. It has been found that EMSI is the best-performance method. The poor performance of LD was expected as this technique drastically reduces the sample size by sacrificing a large amount of data. Several techniques of imputation are used to complete data is depend on quantity of missing values in dataset, sometimes missing values row is deleted. But this not very useful so that to calculate each and every value which is missing in dataset.

Our results show replacing missing value with estimated one. Most of the clustering methods have been developed to perform clusters of only complete data. Clustering methods cannot be used for incomplete data. The future work is developing clustering technique for complete data.The problems of the unknown clusters number and the initialization of prototypes in the FCM clustering algorithm for symbolic interval-values data are overcome with the help of IFCMwUNC clustering algorithm. IFCMwUNC clustering algorithm is giving fast performance in a few iterations regardless of the initial number of clusters.

References

[1] Zhiping Jia and Zhiqiang Yu Chenghui Zhang,"Fuzzy C-Means Clustering Algorithm Based on Incomplete Data", Proceedings of the 2006 IEEE International Conference on Information Acquisition August 20 - 23, 2006,pp. 600-604..

[2] Chen-Chia Chuang, Jin-Tsong Jeng and Chih-Wen Li , "Fuzzy C-Means Clustering Algorithm with Unknown Number of Clusters for Symbolic Interval Data ", SICE Annual Conference August 20-22, 2008, pp. 358-363.

[3] Dan Li, Chongquan Zhong, Liyong Zhang, "Fuzzy *c*-means Clustering of Partially Missing Data Sets Based on Statistical Representation", Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), 2010,pp. 460-464.

[4] Bhekisipho Twala and Michelle Cartwright, "Ensemble Imputation Methods for Missing Software Engineering Data", 11th IEEE International Software Metrics Symposium (METRICS 2005) ,2005.

[5] Hidetomo Ichihashi, Katsuhiro Honda, Akira Notsu, and Takafumi Yagi," Fuzzy *c*-Means Classifier with Deterministic Initialization and Missing Value Imputation", Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007), 2007, p. 214-221.

[6] Bhekisipho Twala, Michelle Cartwright and Martin Shepperd, "Comparison of Various Methods for Handling Incomplete Data in Software Engineering Databases", 2005, pp. 105-114.

[7] Takashi Furukawa, Shin-ichi Ohnishi and Takahiro Yamanoi, "A study on a fuzzy clustering for mixed numerical and categorical incomplete data", Proceedings of 2013 International Conference on Fuzzy Theory and Its Application National Taiwan University of Science and Technology, Taipei, Taiwan, Dec. 6-8, 2013, p. 425-428.

[8] Jianhua Wu, Qinbao Song and Junyi Shen, "An Novel Association Rule Mining Based Missing Nominal Data Imputation Method", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing ,2007, pp.244-249.

[9] Xinqing Geng and Fengmei Tao, "GNRFCM: A new fuzzy clustering algorithm and its application",International Conference on Information Management, Innovation Management and Industrial Engineering, 2012,pp. 446-448.

[10] Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang, "The Global Fuzzy C-Means Clustering Algorithm", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, pp. 3604-3607.

[11] Richard J. Hathaway and James C. Bezdek, "Fuzzy c-Means Clustering of Incomplete Data", ieee transactions on systems, man, and cybernetics—part b: cybernetics, vol. 31, no. 5, october 2001, p. 735-744

[12] Ming-Chuan Hung and Don-Lin Yang, "An Efficient Fuzzy C-Means Clustering Algorithm", 2001, pp.225-232

[13] Katsuhiro Honda, Ryoichi Nonoguchi, Akira Notsu, Hidetomo Ichihashi, "PCA-guided *k*-Means Clustering With Incomplete Data", 2011 IEEE International Conference on Fuzzy Systems June 27-30, 2011, pp. 1710-1714

[14] Hidetomo Ichihashi, Katsuhiro Honda, Akira Notsu, and Takafumi Yagi," Fuzzy *c*-Means Classifier with Deterministic Initialization and Missing Value Imputation", Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007), 2007, p. 214-221