ISSN:0975-9646

A Review on Incomplete Data And Clustering

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

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

Abstract **- Clustering methods have been developed to analyze only complete data. Although sometimes encounter data sets that contain one or more missing feature values (incomplete data), traditional clustering methods cannot be used for such data. Missing value handling is an important preparation step for clustering of partially missing data sets, and inappropriate treatment of missing data in clustering may cause large errors or false results. The imputation approach is used for incomplete data. IFCMwUNC clustering algorithm, the problems of the unknown clusters number and the initialization of prototypes in the FCM clustering algorithm for symbolic interval-values data are overcome. IFCMwUNC clustering algorithm can be fast converges in a few iterations regardless of the initial number of clusters.**

Keywords— Incomplete data, missing data, imputation, Fuzzy c-means clustering algorithm.

I. INTRODUCTION

Clustering is a process of partitioning or grouping a given set of unlabeled patterns into a number of clusters such that similar patterns are assigned to one cluster. A cluster is a collection of data elements that are similar to one another within the same cluster and are dissimilar to data elements in other clusters. Clustering is an example of unsupervised learning. Cluster analysis has been widely used in various applications such as market research, pattern recognition, data analysis, and image processing.[7] Clustering partitions large data sets into groups according to their similarity, so due to this property clustering is also called data segmentation in some applications.

There are two main approaches to clustering. One method is crisp clustering or hard clustering : in this the data are divided into distinct clusters, where each data element belongs to exactly one cluster and the other one is fuzzy clustering: in this data elements belong to more than one cluster, and associated with each element is a set of membership levels.[10] Any datum with some (but not all) missing feature values is referred to as an incomplete datum. A data set with at least one incomplete datum is referred to as an incomplete data set; otherwise, it is called complete. Missing data are an often unavoidable problem when analyzing data. In many situations standard analyses of the data are affected by the problem of missing values.

Traditional clustering methods cannot be directly applied to data sets that contain incomplete data, so need to treat such data. A common approach to analyzing data with missing values is to remove attributes and/or instances with large fractions of missing values. Imputation methods involve replacing missing values with estimated ones based on information available in the dataset. 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.

There are two common solutions to the problem of incomplete data that are currently applied by software engineering researchers. The first includes omitting the instances having missing values (i.e. listwise deletion),which not only does it seriously reduce the sample sizes available for analysis it also ignores the mechanism causing the missingness. A smaller sample size gives greater possibility of a non-significant result, i.e., the larger the sample the greater the statistical power of the test. The second solution imputes (or estimate) missing values from the existing 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 an attribute does not depend on either the observed data or the missing data.
2. Missing at random (MAR)
   When the distribution of dataset having a missing value for an attribute depends on observed data, but does not depend on the missing data.
3. Not missing at random(NMAR)
   When distribution of dataset having a missing value for an attribute depends upon the missing values.

Three types of problems are usually associated with missing values

1. Loss of efficiency
2. Complications in handling and analyzing the data.
3. Bias resulting from differences between missing and complete data.

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 alter the calculations in 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) and those of 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 commonly used for clustering. The performance of the FCM algorithm depends on the selection of the initial cluster center and/or the initial membership value. The problems of the unknown clusters number and the initialization of prototypes in the FCM clustering algorithm for symbolic interval-values data are overcome. 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 classifier learning since incomplete data in either the training data or test (unknown) data affect the prediction accuracy of learned classifiers. Incomplete data could be caused by unit nonresponse (where no data could be collected from the sampled unit) or item nonresponse (where partial data is collected for the unit, but some items are missing). 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]. Xinqing Geng and Fengmei Tao stated that the main defect of the traditional fuzzy partitional clustering algorithm is to know the number of clustering in advance and the limitation of FCM algorithm is only applicable to spheroid and sensitive to the isolated data. GNRFCM algorithm adds a weight to the membership of the data, which is to decrease the effect of the isolated data on the initial cluster center. Since the number of clustering and the initial cluster centers affect cluster result. GNRFCM algorithm overcomes the defect that FCM is only suitable for spheroid space, sensitive to the isolated data, and partitional fuzzy clustering algorithm is to know in advance. GNRFCM algorithm is more suitable for text mining than FCM algorithm and partitional fuzzy clustering algorithm[9].

According to the percentage of incomplete data in fuzzy c-means(FCM) clustering algorithm can not directly act on the data set and its effect on clustering analysis, a modified fuzzy c-means clustering algorithm is proposed by Zhiping Jia and Zhiqiang Yu Chenghui Zhang.The algorithm is IDFCM(Incomplete Data fuzzy c-means) clustering algorithm based on FCM algorithm. In this algorithm, consider two cases on the basis of percentage of incomplete data. If percentage is low of incomplete data then simply deletes all data which have missing feature values otherwise replaced that missing values with estimated one[1].

Bhekisipho Twala, Michelle Cartwright and Martin Shepperd represented the imputation methods involve replacing missing values with estimated ones based on information 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. PROPOSED PLAN

The MMSI and EMSI is designed to calculate missing value in dataset. Firstly find out missing values in dataset and replace with estimated one successfully. After completion of incomplete dataset, apply clustering on it. Comparison between these two imputation methods on the basis calculated values. Last step is to find the accuracy of the clustering algorithm.

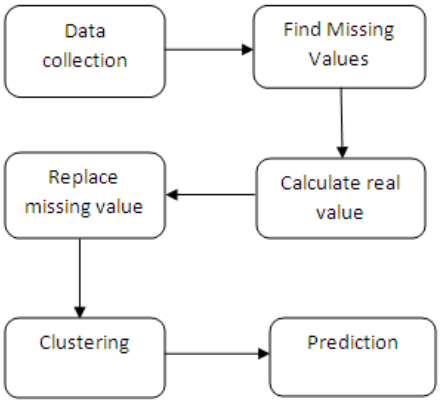

Fig. 1. Phases of implementation of proposed plan

IV. CONCLUSION

In literature survey section study of various techniques of incomplete data handling is done. Different researches have focused on one or more problem in imputation techniques as well as in clustering algorithm. 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. Imputation methods involve replacing missing values with estimated ones based on information available in the dataset. inappropriate treatment of missing data in clustering may cause large errors or false results.

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