RESEARCH ARTICLE OPEN ACCESS

Clustering Large Databases Using Gmm

Dr. C. Chandrasekar*, M. Srisankar**
* (Assistant Professor, Department of Computer Applications, Sree Narayana Guru College, Coimbatore – 105)
** (Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore - 46)

ABSTRACT
Innovation of such clusters of data is essential in illuminating main links in categorical data regulatory networks. There are lot of problems exists in the previous clustering methods especially while grouping the data with mixed data types. This experiment analyzes those existing methods and comes with the new approach for clustering the mixed data items. There are many methods occur for clustering the parallel kind of data, whereas only very few methods exists for clustering mixed data items and it leads to the need of better clustering technique for classification of mixed data. For investigative data, the clustering with the help of Gaussian mixture models is widely used.

Keywords **-** Clustering methods, High dimensional data, K-means clustering, GMM, Adult data set, Categorical data

I. INTRODUCTION

In knowledge discovery, a basis data mining technique used is data clustering. Most widely used methods in data mining is clustering and its core is grouping the whole data based on its parallel measures that based on some distance measure. The problem of clustering has become increasingly important in recent years. Data mining offers a suite of algorithms, each addressing a different task and in the process elucidating a unique facet of the data [1]. Of the many facets of data mining, we are particularly interested in clustering problems, i.e. the process of finding similarities in the data and then grouping similar data into identifiable clusters. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Most techniques of clustering comprise document grouping, scientific data analysis and customer/market segmentation. Usually, clustering involves the classification of the granting data that includes n points in m dimension into k clusters. The clustering must be such that the data points in the corresponding cluster that should be highly similar to one another.The problems exists in the clustering techniques are: Identifying the similar measure to find the similarity among various data, it is complex to find out the suitable methods for identifying the identical data in unsupervised way and originate a description that can differentiate the data of a cluster in an efficient manner. It is a vital task of experimental data mining, and a usual method for mathematical data study, used in many areas, such as machine learning, image analysis, information retrieval and pattern recognition

II. BACKGROUND STUDY

Efficient collaborative method for mixed numeric and categorical data is recommended by earlier researchers[4]. Most of the existing clustering algorithms focus on numerical data whose inbuilt geometric characteristics can be put-upon obviously to outline distance functions between data points and a massive or enormous data exist in the databases is categorical, where attribute values will not be reasonably arranged as numerical values. Since the differences in the features of the categories of data, force to build up standard functions for mixed data[2] was not successful. In previous method novel divide-and-conquer method to overcome this issue. In the beginning, the real mixed dataset is segmented with different sub-datasets and next, accessible well predictable clustering method developed for various types of datasets is applied to produce equivalent clusters. Finally, the clustering results on the categorical and numeric dataset are combined as a categorical dataset, on which that clustering method is used to produce the final result. The vital contribution in this experimental study is to present an algorithm for the mixed features clustering complications, in which previous clustering algorithms like k-means[5] as shown in fig 1. can be effortlessly incorporated. The majority of existing clustering algorithms encounter serious scalability and/or accuracy related problems when used on databases with a large number of records and/or attributes. Only few methods can handle numeric, nominal, and mixed data.

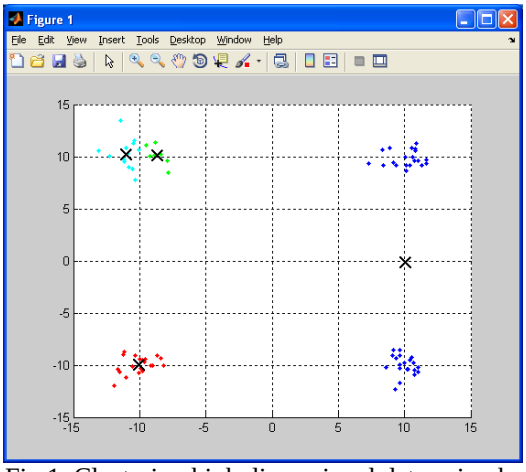

Fig 1. Clustering high dimensional data using k-means

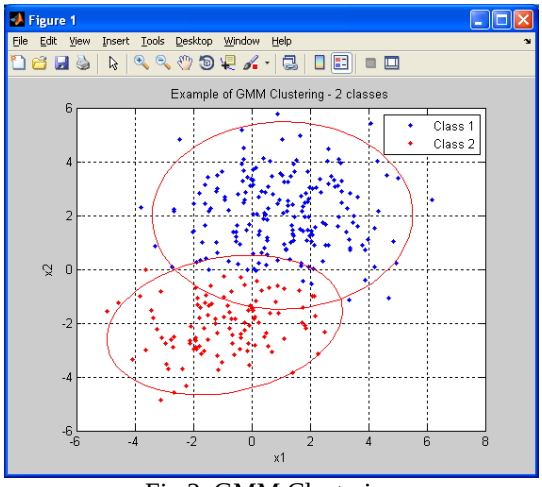

Fig 2. GMM Clustering

III. METHODOLOGY

Clustering in high dimensional spaces is frequently challenging as theoretical results [3] doubted the meaning of closest matching in high dimensional spaces. When the data items are definite or mixed, the similarity is difficult to find by Euclidean distance method. With the enormous data collected from different fields such as banks, or health sector, universities, colleges, schools web-log data and biological sequence data which are definite data need good clustering method. It is highly complicated to group the unconditional data into important category with good distance measure, acquiring sufficient data similarity and to use combination with an effective clustering algorithm and in with mixed numeric and categorical data, only very few methods available. There are many methods occur for clustering the parallel kind of data, whereas only very few methods exists for clustering mixed data items and it leads to the need of better clustering technique for classification of mixed data. The clusters identified in the present study differ somewhat from those from earlier studies. Previous methods often cluster adult dataset into groups of persons by age, salary, married, un-married and country cognitive performance. Even when it compared adult dataset test performance to that of male and female category, the category in the present study produced few performances that would be considered clinically impaired. For investigative data, the clustering with the help of Gaussian mixture models is widely used.

IV. EXPERIMENTAL RESULT

The investigation study of the proposed technique was accepted with the support of Adult Data Set. The number of clusters answered for the proposed method is lesser when compared to the other approaches and the outliers were not found by using the proposed method. The Proposed algorithm utilizes an active sampling mechanism and compels at most a single examine through the data. The capability to generalize our identifications to other instances is reduced in many approaches. One possible merit identified in the operational definition of age, salary, married, unmarried and the country used in this study as in fig 4. We determine the high quality of the clustering solutions that was found, their explanatory power, and proposed model's good scalability. GMM method has good accuracy, uses a single tunable parameter, and can successfully function with partial memory resources.

4.1 GMM Method

Gaussian Mixture Models[6] are one among the significant statistically developed approaches for clustering. The concept of clustering, and see how one form of clustering in which it assumed that individual data points are produced by choosing one of a set of multivariate Gaussians and then sampling from them. This can be a distinct related to operation and examine how to learn such a thing from data, and we discover that an optimization. This optimization method is called Expectation Maximization (EM).

4.2 Adult data set

The adult data set from the UCI repository contains massive data with both numeric and nominal values. There are a total of 18,322 records, split in a 2:1 ratio to form a train and test set. The target class indicates whether a person by age, income, married and unmarried. After challenging the nominal

attributes into binary values, the training data were clustered into 14 clusters. Then these clusters were labeled with the predominant class. On the test data, k-Meansaccuracy was 77.2% whereas Gaussian Mixture Model results with accuracy rate of 91%.The distance-based k- Means algorithm produced clusters that did not separate well the rare class from the dominant class.

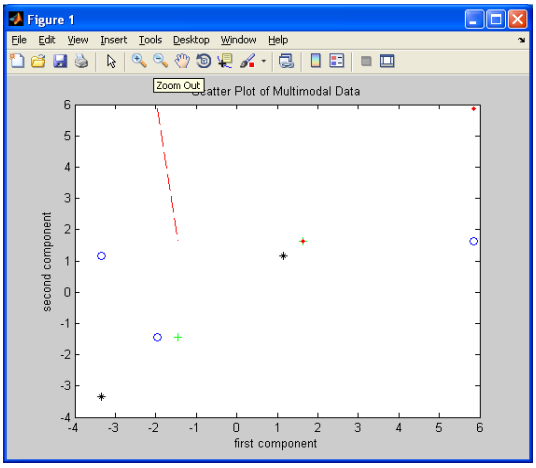

Fig 3. Scatter Plot of Multimodal Data

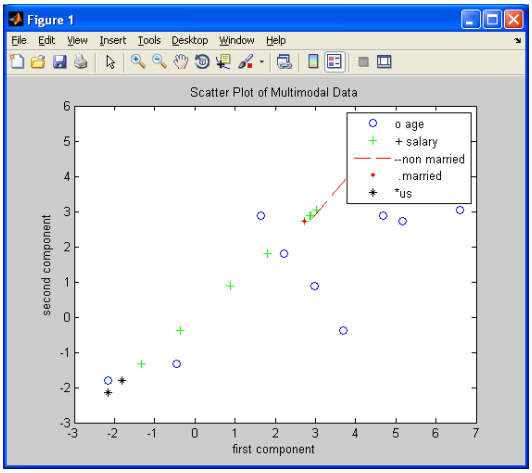

Fig 4. Plotting Multimodal Data with age, salary, married,

V. CONCLUSION

This research focus on effective clustering method for mixed category data. There are various methods available for clustering categorical, but those techniques resulted with some drawbacks. To overcome this issue, the clustering in mixed data can be performed based on many approaches. But these approaches took more time consumption for classification. To this issue, a new modified Gaussian Mixture Model technique is used in this study based on adult dataset. The experiment is performed with the help of Adult Data Set and it can be examine that the improved classification result is achieved by the proposed technique when compared to the previous experimental methods. The study through the paper suggests that the proposed approach can be used to cluster the mixed data items with better accuracy of classification.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro,Psmthy… - citeulike.org. Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park:CA, 1996.

[2] Hari Prasad, D and M. Punithavilli, 2010a. A review on data clustering algorithms for mixed data. Global J. Comput. Sci. Technol., 10: 43-48.

[3] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful. In *Proc. of the Int. Conf. Database Theories*, pages 217–235, 1999.

[4] Reddy, M.V.J. and B. Kavitha, 2010. Efficient ensemble algorithm for mixed numeric and categorical data. Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, Dec. 28-29, IEEE Xplore Press, Coimbatore, pp: 1-4. DOI: 10.1109/ICCIC.2010.5705738.

[5] Huang, Z. (1998). Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values. Data Mining and Knowledge Discovery, 2, p. 283-304 (Publitemid 128695480)

[6] K. Zhang and J. T. Kwok. Simplifying mixture models through function approximation. In B. Scholkopf, J. Platt, and T. Hoffman, editors, Advances in Neural Information Processing Systems 17, pages 1577–1584, Cambridge, MA, 2007. MIT Press.