ISSN:0975-9646

Review of Algorithms for Clustering Random Data

Bhumika Ingale, Priyanka Fulare

Department CSE,
Nagpur University
India

***Abstract*— Clustering of Random data is one of the essential task in data mining. Uptill now most of the research is done on methods which deals with clustering of certain data. Clustering is nothing but grouping data objects. Data objects in one clusters are same where as data object in different clusters are not. Less research is done on methods that can be used for clustering of random data. Random data is data which is unstructured or data which does not have any proper arrangement. Today, there is huge amount of data available but to mine useful information from such data is difficult task. This paper dicuss some algorithm which can be used for clustering of Random data.**

***Keywords*— Uncertain data, Probabilistic Clustering, Partitioning method.**

I. INTRODUCTION

Data mining can be applied to any meaningful data set for particular application. Data mining concept can be used for huge data object. In this information age, where there is a huge amount of data saturation. For a particular task such a huge amount of data may not be required only a small amount of data may be required for particular tasks .Data mining can be used in many other applications like in financial data analysis, banking, Telecommunication industries, business and also in web technologies. For example, in medical field there are number of diseases and each disease have different cure. In such a case a database is created. If the database is large then management of that database becomes tricky. Data mining can be used in such a case where database is huge and when the classification of such a data is difficult.

Clustering is the process of grouping data files. Data files which are similar belongs to one cluster and data objects which are dissimilar belongs to another cluster. Clustering is unsupervised learning that is learning from raw data. Clustering of random data is a new concept less research is done on clustering of random data. Many algorithms exist for clustering of certain data. Each algorithm has its merits and demerits. Very few algorithm exists for clustering of uncertain data. For clustering of random data hybrid algorithm is required.

This paper discuss some methods that can be used for clustering of random data. Clustering of random data can be termed as arrangement of data into certain format. For example , if the database of the system contains some data which is random in nature that is the database may contain text files or graphical files or some other files. So the algorithm should be able to arrange the data in proper format.

The process of managing uncertain data is much more complicated than that for traditional databases. This is because the uncertainty information needs to be represented in a form which is easy to process and query. Different models for uncertain data provide different tradeoffs between usability and expressiveness [7]. All most every field of human life has become data-intensive, which makes data mining as an essential component[8].

Traditionally, clustering algorithms deal with a set of objects whose positions are accurately known. The objective is to find a way to divide objects into clusters so that the total distance of the objects to their assigned cluster centres is minimised[10].

Clustering of Random data, is one of the essential tasks in data mining, posts significant challenges on both modelling similarity between Random objects and developing efficient computational methods for the same is been done. But still very less influence is given to Random data mining approach. Randomness suggests a order-less or non-coherence in a sequence of data, text such that there is no specific pattern or combination. Data which is not organized in a proper way is known as Random data. It can be unstructured data or semi structured data. Randomness means lack of pattern.

II. LITERATURE SURVEY

[1] introduces a novel density-based network clustering method, called graph-skeleton-based clustering (gSkeletonClu). [2] discusses Kullback-Leibler divergence method. KL divergence is very costly and even infeasible to implement. To tackle the problem, kernel density estimation and the fast Gauss transform technique is used to further speed up the computation.

[3] surveys the broad areas of work in Data Mining field . It explore the various models utilized for uncertain data representation. In uncertain data mining, it examine traditional mining problems such as frequent pattern mining, outlier detection, and clustering. It discuss different methodologies to process and mine uncertain data in a variety of forms. In this paper, it provide a survey of uncertain data mining and management applications. In [9] presented the interesting problem of evaluating spatial queries for existentially uncertain data objects. Variants of common spatial queries, like range and NN search, have probabilistic versions for this data model.

III. ALGORITHMS FOR CLUSTERING RANDOM DATA

There are many algorithms that can be used for clustering data objects. Following are the basic algorithms which can be used for clustering of Random data.

A) Partitioning Algorithm

Partitioning method is the simplest and most fundamental type of cluster analysis. In partitioning algorithm a data set D is given. The data set contain N objects and K is the number of clusters to be formed. The data objects is organized into K partitions where each partition is a cluster. This method finds mutually exclusive clusters mostly of spherical shape. It can find cluster of spherical shape which limits the use of partitioning algorithm. This method can use mean or medoid method to represent the center of cluster. But the method is effective for small to medium size. The partitioning process continues until each cluster at the lowest level is coherent enough either containing only one objector the objects with a cluster is sufficiently similar to each other[4].

[6] discuss about PAM (Partitioning Around Medoids) in which partitioning of data is done and CLARA (Clustering LARge Applications) which relies on sampling. CLARAN (Clustering Large Application Based upon RANdomized Search) which has many advantages over CLARA. When compared with CLARA, CLARANS has the advantage that the search space is not localized to a specific subgraph chosen a priori, as in the case of CLARA[6].

B) Hierarchical Algorithm

Partitioning method meets the basic requirement of clustering i.e. it organizes a set of objects into a number of exclusive groups. There may be a need to partition data at different levels so this algorithm fails or does not work properly . In such a situation where data has to be arranged into a hierarchy or tree like structure hierarchical algorithm is used. It overcomes some of the limitation of partitioning algorithm.

Hierarchical method can be classified as agglomerative or divisive method. In agglomerative method merges cluster to form cluster of large size i.e. it starts by each object form its own cluster and iteratively merges clusters into larger and larger clusters until all the objects are in a same cluster two clusters are merged in each iteration where each cluster contains atleast one object so this method requires at most N iteration. Agglomerative approach is also known as bottom up approach. Divisive Hierarchical clustering method divides the root cluster into several number of smaller subclusters and then it recursively partition cluster. Divisive Hierarchical method is also known as top down method.

C) Probabilistic clustering

Data is selected from a mixture of probability distribution. It then uses the mean, variance of each distribution as parameters for clustering e**.** Probabilistic Graphical Models are a popular and versatile class of models which have significantly greater expressive power because of their graphical structure. They allow us to intuitively capture and reason about complex interactions between the uncertainties of different data items[7].

Hierarchical method performs clustering process and presents the result in a tree like structure. Partitioning method divides the data into n number of objects.

IV. CONCLUSIONS

Clustering of random data requires a hybrid algorithm. We can use Hierarchical method along with other algorithm for clustering purpose. Data uncertainty brings new challenges for cluster formation, since clustering of uncertain data demands a measurement of similarity between uncertain data objects[2]. In [5], it propose to use fuzzy distance functions to measure the similarity between uncertain object representations. Contrary to the traditional approaches, it do not extract aggregated values from these fuzzy distance functions but propose to enhance data mining algorithms so that they can exploit the full information provided by these functions.

There are many advantages of using Hierarchical method for clustering purpose. But partition method is a simple and performs clustering at very basic level. So this algorithm can be used along with Hierarchical method or Probabilistic method. As per to condition for clustering algorithm can be selected and used.

ACKNOWLEDGMENT

The author would like to thank the guide for there guidance. The author is grateful to the guide for reviewing the paper as well.

REFERENCES

[1] Jianbin Huang, Heli Sun , Qinbao Song, Hongbo Deng, Jiawei Han " Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network" IEEE Trans on knowledge and data engineering, Vol. 25, No. 8. August 2013

[2] Bin Jiang, Jian Pei, Yufei Tao , Xuemin Lin " Clustering Uncertain Data Based on Probability Distribution Similarity" IEEE Trans on knowledge and data engineering, Vol. 25, No. 4. pp.753 April 2013.

[3] Charu C. Aggarwal and Philip S. Yu " A Survey of Uncertain Data Algorithms and Applications" IEEE Trans on knowledge and data engineering, Vol. 25, No. 5. May 2009.

[4] Jiawei Han , Micheline Kamber , Jian Pei " Data Mining: Concepts and Techniques Third Edition" pp.459 2011.

[5] Hans-Peter Kriegel and Martin Pfeifle "Hierarchical Density-Based Clustering of Uncertain Data" Proc. IEEE Int'l Conf. Data Mining (ICDM) 2005.

[6] Raymond T. Ng and Jiawei Han "CLARANS: A Method for Clustering Objects for Spatial Data Mining" IEEE Trans on knowledge and data engineering, Vol. 14, No.5 pp.1003-1015. . September/October 2002.

[7] Charu C. Aggarwal "Managing and Mining Uncertain Data" IBM T. J. Watson Research Center Hawthorne, NY 10532 pp.02-03.

[8] Venkatadri.M, Dr. Lokanatha C. Reddy " A Review on Data mining from Past to the Future " International Journal of Computer Applications (0975 – 8887) Volume 15– No.7. pp.19 February 2011.

[9] Man Lung Yiu, Nikos Mamoulis, Xiangyuan Dai, Yufei Tao, and Michail Vaitis "Efficient Evaluation of Probabilistic Advanced Spatial Queries on Existentially Uncertain Data" IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 1, January 2009.

[10] Ben Kao, Sau Dan Lee, David W. Cheung, Wai-Shing Ho, K. F. Chan " Clustering Uncertain Data Using Voronoi Diagrams" Eighth IEEE International Conference on Data Mining. pp.333 2008.