RESEARCH ARTICLE OPEN ACCESS

Study of Clusterization Methods for Random Data

Bhumika Ingale*, Ms.Priyanka Fulare**
*(Department of Computer Science, RTMNU University,Nagpur-India

ABSTRACT-
Clustering is the process of grouping data object. Data sets within a cluster should be similar and data sets outside the cluster should be dissimilar. It is the commonest form of unsupervised learning i.e learning from raw data. Data mining is a knowledge discovering process. Clusters are formed as a result of data mining process. In the recent years much work is been done on mining certain as well as uncertain data. Many algorithms exist for mining of certain data. Data mining process can also be used for random data. Data which is random in nature does not contain a particular pattern. These papers discuss some algorithms that can be used for clustering of Random Data.

Keywords - Uncertain Data, knowledge discovery process, certain data, probabilistic graphs, clustering algorithms.

I. INTRODUCTION

Clustering is the process of grouping a set of data objects such that objects with in a cluster have high similarity and data objects outside the cluster have dissimilarity. Different clustering algorithms may come out with different clustering results. Applying different clustering algorithms on same data may prove to be useful in the sense that it may come out with the previously unknown groupings with in the data. The problem of clustering Random objects according to their probability distributions happens in many scenarios.

For example, in marketing research, users are asked to evaluate digital cameras by scoring on various aspects, such as image quality, battery performance, shotting performance, and user friendliness. Each camera may be scored by many users. Thus, the user satisfaction to a camera can be modeled as an uncertain object on the user score space. There are often a good number of cameras under a user study. A frequent analysis task is to cluster the digital cameras under study according to user satisfaction data. One challenge in this clustering task is that we need to consider the similarity between cameras not only in terms of their score values, but also their score distributions. One camera receiving high scores is different from one receiving low scores. At the same time, two cameras, though with the same mean score, are substantially different if their score variances are very different[2]. Association rule mining can also be used for mining purpose. Association rule mining is an extremely popular data mining technique that can discover relationships between data[6]. In the data stream environment, the patterns generated at different time instances are different due to data evolution. The frequency of change for rules and patterns is different for different applications[7].

II. LITERATURE REVIEW

[4] is based on, clustering of probabilistic graphs using the edit distance metric. It focused on the problem of finding the cluster graph that minimizes the expected edit distance from the input probabilistic graph. The paper adheres to the possible- worlds semantics. Also, its objective function does not require the number of clusters as input; the optimal number of clusters is determined algorithmically. In addition, it proposed various intuitive heuristics to address it. Further, it established a framework to compute deviations of a random world to the proposed clustering and to test the significance of the resulting clusterings to randomized ones. Also, it addressed versions of the problem where the output clustering is itself noisy.

[2]discusses Kullback-Leibler divergence method. KL divergence is very difficult to implement. To solve the problem, kernel density estimation and the fast Gauss transform technique is used to further speed up the computation.

III. RANDOM DATA

Random means lack of pattern or predictability in Data. Randomness suggests a non-order or non-coherence in a sequence of data such that there is no exact pattern or combination between them.

In computer science, uncertain data is the notion of data that contains some uncertainty. Uncertain data is usually found in the area of sensor networks. Fundamentally, uncertain data has been explored in many applications. A traditional field where uncertainty is evident, such as sensor data, is joined by many topics from computer science and data analytics[3].

IV. METHODS FOR CLUSTERING RANDOM DATA

There are number of algorithms used for clusterization of certain data. The following algorithms can be used for clustering of Random data.

A) Clustering high dimensional data**:** Clustering high-dimensional data is a particularly important task in cluster analysis because many applications require the analysis of objects containing a large number of features or dimensions[5].CLUIQE algorithm can be used for such a high dimensional data. CLIQUE method searches for clusters in subspaces of data rather than over the entire data space.

B) Partitioning method: In this method a database of n objects or data tuples is given , a partitioning method constructs k partitions of the data set, where each partition is a cluster and where n is greater then or equal to k. It classifies the data into k groups, which together must satisfy some conditions . In the first condition each group must contain minimum one data object, and each data object must belong to exactly one group. The second requirement can be relaxed in some fuzzy partitioning techniques. Heuristic methods like k-means and k-medoids algorithms**.** Each cluster is represented by the center of the cluster in K-means whereas in k-medoids each cluster is represented by one of the objects located near the center of the cluster. Heuristic clustering methods work well for finding spherical-shaped clusters from small to medium-sized databases.

PAM is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean.PAM cannot be used for large data set. It can be used for small data sets. It takes $O(k(n-k)^2)$ for each iteration .

CLARA(Clustering Large Applications) , a small portion of actual data set is chosen as a representative of data , instead of taking the whole set of data into consideration .It draws multiple samples of the data objects, PAM is applied on each sample, and it gives the best clustering as the output. It deals with larger data sets as compared to PAM .But its efficiency depends on the size of sample, clustering is based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased. The scalability and quality of CLARA can be improved. CLARAN is a K-medoid type of algorithm.

CLARAN (Clustering Large Applications based upon RANdomized Search) combines the sampling techniques with PAM. CLARAN does not confines itselt to any sample at any given time. While CLARA has a fixed sample at each stage of the search, CLARANS draws a sample with some randomness in each step of the search[5].

C) Hierarchical cluster: Hierarchical clustering method creates a decomposition of the given set of data objects. Hierarchical clustering can be classified as Agglomerative and divisive Clustering. In Agglomerative method merging of the most similar pairs of data points are done until one big cluster left. This is called a bottom-up approach. In divisive method splitting of large data is done. It is top- down approach. In contrast to the majority of algorithms for clustering uncertain objects which are based on partitional or density based schemes, it should be noted that there is relatively poor research on hierarchical clustering of uncertain data[1]. In [8], it shows a systematically study of the usage of hierarchical co-clustering methods for organizing different types of music data. Hierarchical co-clustering aims at simultaneously constructing hierarchical structures for two or more data types, that is, it attempts to achieve the function of both hierarchial clustering and co-clustering.

V. CONCLUSIONS

There are many algorithms which can be used for clustering of data. But for clusterization of Random data a hybrid algorithm must be used .Considering the merits and demerits the algorithm must be chosen. For clusterization of Random data combination of any two or three algorithms can be used .

Using partition method with hierarchical algorithm can form a cluster with great accuracy. But as per to the requirement of clustering output the algorithms must be selected.

ACKNOWLEDGMENT

The author would like to thank our guide for there guidance, foresight, and assurance in difficult time. The author is grateful to the guide for reviewing the paper as well.

REFERENCES

[1] Francesco Gullo, Giovanni Ponti , Andrea Tagarelli Sergio Greco "A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach" Eighth IEEE International Conference on Data Mining 2008 pp. 822.

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

[3] Peter Benjamin Volk, Frank Rosenthal, Martin Hahmann, Dirk Habich, Wolfgang Lehner "Clustering Uncertain Data With Possible Worlds " IEEE International Conference on Data Engineering.pg.1625

[4] George Kollios, Michalis Potamias, and Evimaria Terzi "Clustering Large Probabilistic Graphs". IEEE Trans on knowledge and data engineering, Vol. 25, NO. 1, April 2006.

[5] Jiawei Han Micheline Kamber "Data Mining: Concepts and Techniques Second Edition" pg.400,pg.401,pg. 408.

[6] Cheng-Hsiung Weng "A study of mining certain itemsets from uncertain data" International Conference on Fuzzy Theory and Its Applications National Chung Hsing University, Taichung, Taiwan, Nov.16-18, 2012.pg.352.

[7] Bi-Ru Dai, Jen-Wei Huang, Mi-Yen Yeh, and Ming-Syan Chen "Adaptive Clustering for Multiple Evolving Streams" IEEE Trans on knowledge and data engineering, vol. 18, no. 9, pp.1166 september 2006.

[8] Jingxuan Li, Bo Shao, Tao Li, and Mitsunori Ogihara "Hierarchical Co-Clustering: A New Way to Organize the Music Data" IEEE Trans on Multimedia, Vol. 14, No. 2, April 2012