International Journal of Research in Advent Technology, Vol.2, No.2, February 2014

E-ISSN: 2321-9637

Review of Clusterization Techniques for Random Data

Bhumika Ingale[1], Minal Kamble[2], Dr.S.P.Khandait[3]
Department Computer Science and Engineering, RTMNU, Nagpur India
Email: ingale.bhumika0@gmail.com

Abstract-Clustering is the process of grouping a set of objects into classes of similar objects. Clustering of Random data is one of the essential task in data mining. Uptill now most of the research is been done on techniques which deals with clusterization of certain data. This paper describes clusterization techniques which can be used for clustering of Random data. The previous methods extend traditional partitioning clustering methods like k-means and density-based clustering methods like DBSCAN, which rely on geometric distances between objects. But these methods are still in used as they gives clusters of higher accuracy. In this paper, various methods for clustering of Random data have been considered.

Index Terms- Uncertain data, Density based Clustering, Probabilistic Clustering.

1. INTRODUCTION

Clustering is the process of grouping a set of objects into classes of similar objects. Data Objects with in a cluster should be similar. Data Objects from different clusters should be dissimilar. It is the commonest form of Unsupervised learning i.e learning from raw data. Clustering is one solution to the case of unsupervised learning, where the class labeling information of the data is not available[9]. Data mining can be applied to any meaningful data set for particular application.

Data mining concept can be used for large data sets. In this information age, where there is a huge amount of data saturation. Such a huge amount of data may not be required for a particular task. Only a small amount of data may be required for particular tasks. 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 large and when the classification of such a data is difficult. Data mining can be used in many other applications like in banking, business and also in web technologies. But there are many issues related to data mining. The issues may be related to efficiency and scalability of the data mining algorithms, it can be related to mining methodology or user interaction. Cluster Analysis is a branch of statistics that, in the past three decades, has been intensely studied and successfully applied to many applications[7].

Clusterization of Random data, is one of the essential tasks in data mining, posts significant challenges on both modeling 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. The process of managing uncertain data is much more complicated than that for traditional databases.

2. LITERATURE SURVEY

[1] 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. [2] introduces a novel density-based network clustering method, called graph-skeleton-based clustering (gSkeletonClu). [3] surveys the broad areas of work in Data Mining field. In this paper, it provide a survey of uncertain data mining and management applications. 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.

3. RANDOM DATA

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. Randomness suggests a order-less or non-coherence in a sequence of data, text such that there is no specific pattern or combination.

4. PROPOSED METHOD

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

4.1 Density based method: Density-based clustering methods was develop to build a cluster of arbitrary shape. In density based method, dense regions of objects forms a cluster which is separated by regions of low density data objects. This method can handle noise. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. It is very sensitive to input parameters. In DBSCAN algorithm the input parameter is hard to determine. Run time complexity of DBSCAN

algorithm is $O(n^2)$ for each point it has to be determined. In a spatial index, the computational complexity of DBSCAN algorithm is O(n log n), where n is the number of database object. The performance of DBSCAN is affected by high-dimensional datasets. It does not work well in high dimensional datasets. DBSCAN helps in detecting outliers.

OPTICS extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings. In DENCLUE clusters objects are based on a set of density distribution functions[5]. OPTICS is based on DBSCAN. It does not produce clusters explicitly. Instead of that, it generates an ordering of data objects representing density-based clustering structure. OPTICS gives good results if the parameters are large.

4.2 Grid-Based Methods*:* The grid-based clustering approach uses a multi-resolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The advantage of the approach is its fast processing time, which is independent of the number of data objects, but dependent on only the number of cells in each dimension in the quantized space[5].

STING (A Statistical Information Grid Approach) is an example of grid based approach. It is a grid based multi-resolution techniques in which the spatial area is divided into rectangular cells. There are many levels of cells corresponding to different levels of resolution. It removes the irrelevant cells. It uses top down approach to answer the spatial queries. This approach is query independent and it can easily parallelize. But it has a disadvantage that there is no diagonal boundary in the resulting cluster i.e every cluster boundaries are either horizontal or vertical but no diagonal boundaries are detected. This may lower the quality and accuracy of the cluster even if the processing time of the technique is fast. The quality of the STING clustering is based on the granularity of the lowest level of the grid structure as STING uses multi-resolution approach to cluster analysis.

4.3 Probabilistic clustering : Data are picked from mixture of probability distribution. Use the mean, variance of each distribution as parameters for clustering.

Algorithmic Hierarchical Clustering method uses linkage measures that is easy to understand and efficient in clustering. This algorithm is used in many application. But it has some shortcoming, that is choosing a good distance measure for clustering is very difficult. Probabilistic Hierarchical Clustering method is used for data clustering it overcomes that drawback of Algorithmic Hierarchical Clustering. It uses the probabilistic model for measuring the distance between the clusters.

Clustering is a technique that has been studied and implemented to number of real-life applications. Many efficient algorithms, have been devised to solve the clustering problem efficiently. Traditionally, clustering algorithms deal with a set of objects whose positions are accurately known. The aim 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 [4].

[7] 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[7]. PAM works well if the data sets are small. There are many other algorithms which can be used for clustering of random data. The data may contain some errors and noise. Data obtained from measurements by physical devices are often imprecise due to measurement error[8]. The algorithm used for clustering of random data should give resulting cluster with minimum error and less outliers. But there are some issues related to clustering of random data. The data may have different attributes or it may be multidimensional data depending upon those conditions the algorithm can be selected and implemented.

5. CONCLUSION

In [6], 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 number of algorithms which can be used for clusterization of data. But clusterization of random data requires hybrid algorithm clustering cannot be performed using a single algorithm.. Some more algorithm can also be used to reduce the complexity and to increase the efficiency of the over all algorithm.

Acknowledgments

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] Bin Jiang; Jian Pei ; Yufei Tao ; Xuemin Lin (2013): Clustering Uncertain Data Based on Probability Distribution Similarity. IEEE Trans on knowledge and data engineering, Vol. 25, No. 4.

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

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

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

[5] Jiawei Han and Micheline Kamber (2006): *Data Mining: Concepts and Techniques Second Edition*. pp.418-424.

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

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

[8] Smith Tsang ; Ben Kao; Kevin Y. Yip; Wai-Shing Ho; Sau Dan Lee (2011): Decision Trees for Uncertain Data. IEEE Trans on knowledge and data engineering, vol.23,No.1..pp.64.

[9] Xiao-Feng Wang and De-Shuang Huang (2009): A Novel Density-Based Clustering Framework by Using Level Set Method. IEEE trans on knowledge and data engineering, Vol. 21, No.11.pp.1515.