ISSN:0975-9646

Clustering of Random Data: A New Concept

Bhumika Ingale, Antara Bhattacharya

CSE, RTMNU
Nagpur,Maharashtra,India

***Abstract*—Many algorithm exist for clustering of certain data. But less research is been done on algorithms which can be used for clustering of Random data. Data which lack predictability is known as Random data. Clustering of such data is difficult task. In this paper, we propose a new clustering algorithm called as Hybrid Hierarchical algorithm. This algorithm will be used for clustering of Random data. It will perform clustering of data which are similar to each other as well as it will also form a cluster of data which are dissimilar. It will be applicable for both text and graphical data. It will simplify data complexity.**

***Keywords*—Density Based Method, Knowledge discovery process, Cluster analysis**

I. INTRODUCTION

Clustering is the process of grouping data objects. Data objects are similar within a cluster where as data objects outside the cluster are dissimilar. Clustering is a form of unsupervised learning. It is important to understand the difference between supervised learning and unsupervised learning. Supervised learning, is a collection of labeled (preclassified) patterns; the problem is to label a newly encountered, yet unlabeled, pattern. Whereas unsupervised learning is learning from raw data where no classification is given whereas in supervised learning a classification of data is required. Clustering is a technique that has been widely studied and applied to many real-life applications. Many efficient algorithms exist , including k-means algorithm, have been devised to solve the clustering problem efficiently [4]. Association rule mining is an extremely popular data mining technique that can discover relationships between data[8].

Cluster Analysis is a branch of statistics that, in the last three decades, has been intensely studied and successfully applied to many applications[12]. Cluster analysis is the organization of a collection of patterns into cluster based on similarity. Cluster analysis is a process of partitioning a set of data objects into subsets. Each cluster is a subset, such that objects in a cluster are similar to one another, but it is dissimilar to objects in other clusters. Cluster analysis is used in many applications. For example, security, business intelligence, biology. A cluster is comprised of a number of similar objects collected or grouped together. Organizing data into sensible group is one of the most fundamental modes of understanding and learning.

Cluster analysis is the study of algorithm and methods for grouping objects. Clustering of Random data, is one of the essential tasks in data mining. Generally, an Random data object can be represented by a probability distribution . Data mining turns a large amount of data into knowledge. Data mining is an interdisciplinary subject. It can be defined in number of ways. Applying different algorithm on data may prove to be useful in the sense that it may come out with the previously unknown grouping with in the data. The problem of clustering random data objects according to their probability distribution happens in many scenarios.

For example, in marketing research, if a users are asked to evaluate mobile phone by scoring on number of aspects, such as, battery performance, image quality memory, screen resolution, shotting performance, quality of images and user friendliness. Each mobile phone may have different features and functions. As per to the need of user the user will score the mobile phone quality. Each mobile phone may be scored by many users on the bases of their functions. Thus, the user satisfaction to a mobile phone can be modelled as an uncertain object on the user score space. There are often a good number of mobile phone under a user study. Dealing with uncertain objects has raised several issues in data management and knowledge discovery. In particular, organizing uncertain objects is challenging due to the intrinsic difficulty underlying the various notions of uncertainty[10].

Most of the uncertain data mining and management algorithms use a variety of simplifying assumptions in order to allow effective design of the underlying algorithms. Examples of such simplifying assumptions could imply tuple or attribute independence. Dealing with Random data objects has raised several issues in knowledge discovery and data management process. Today in such an information age where huge amount of data is available but arrangement of such a large and complex data has become a tough job. In this paper, we will discuss about an algorithm which will overcome this difficulty.

II. LITERATURE SURVEY

[1] discusses Kullback-Leibler divergence method. KL divergence is very complicated method to implement. To tackle the problem, kernel density estimation and 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.

[6] This paper presents an up-to-date survey on evolutionary algorithms for clustering. It tries to reflect the

profile of this area by focusing more on those subjects that have been given more importance in the literature. Particularly, the paper has focused mainly on hard partitional algorithms, though overlapping (soft/fuzzy) approaches have also been covered. [5] 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 clustering to randomized ones. Also, it addressed versions of the problem where the output clustering is itself noisy.

III. ARCHITECTURE

The following Diagram is the Architecture of the system . Random data is unarranged data in the database to retrieve information from such data is difficult task. Clustering of random data is new concept less research is done on this topic. But today in such an information age to retrieve useful information from given raw data is important. Before forming cluster of information which are gathered from random data source. It is important to mine the data and to check whether the retrieve information is useful or not. Data mining deals with this concept. Data mining turns a large amount of data into knowledge. The architecture defines the actual working of the system.

The user fires a query to the query engine. The random data is collected from various different sources. The data is stored in the database in the form of raw data. The data collection and storage is done by the administrator. The data can be extended or reduced by the administrator. The administrator manages the data in the database. When user fires a query to the query engine the data gets collected in the database and then the algorithm is implemented.

Clustering process: The data so collected is first pre-processed. In pre-processing of data first the data cleaning is done. The unwanted files are removed or deleted. Then the files which are required is integrated after removing the outliers. We implement the Hybrid Hierarchical Algorithm. This algorithm is a combination of DBSCAN and Hierarchical algorithm. The groups of data are formed as a result which is stored in the database. And thus, the result is passed to the user.

The data collection is done by the administrator. The administrator manages the data in the database. When user fires a query to the query engine the data gets collected in the database and then the algorithm is implemented. The random data is collected from various different sources. The data is stored in the database in the form of raw data. Hybrid Hierarchical algorithm will be used for clustering of Random data. The cluster of data so formed as a result is sorted data.

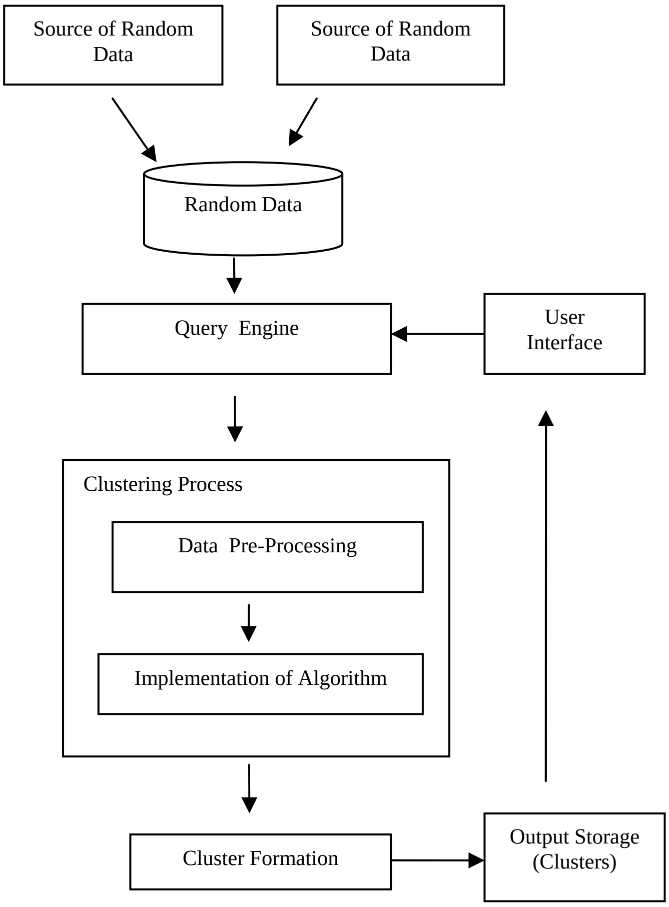

Fig. 1. Architecture of the system

IV. PROPOSED METHOD

Clustering of Random data requires a hybrid algorithm. There are many algorithms which can be used for clustering purpose but very few algorithms gives accurate results with minimum outliers and maximum accuracy. To achieve greater accuracy and minimum outliers following algorithms are used for developing hybrid algorithm.

A) Hybrid Hierarchical Algorithm

Hybrid Hierarchical algorithm will be used for clustering of Random data. This algorithm will give cluster with minimum error. This algorithm is the combination of two algorithm. It uses Density based method along with hierarchical method for clustering purpose.

1) Density based method: Clustering based on density, such as density-connected points. Each cluster has higher density of points than outside of the cluster. It discovers cluster of arbitrary shape, it can handle noise. 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[11]. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can be used for clustering purpose. Here we use DBSCAN algorithm for clustering of random data. It relies on a density-based notion of cluster. It is very sensitive to input parameters.

In density-based clustering methods, regions with high densities of data points are generally treated as cluster centers, while areas with sparse distributions of data points are boundaries to keep cluster centers divided from one another. So, they don't require any prior knowledge concerning the data and are able to identify clusters with arbitrary shape[9]. In DBSCAN algorithm the input parameter is hard to determine. In a spatial index, the computational complexity of DBSCAN algorithm is O(n log n), where n is the number of database object. Run time complexity of DBSCAN algorithm is O(n2) for each point it has to be determine. The performance of DBSCAN is affected by high dimensional datasets. It does not work well in high dimensional datasets. Density based method helps in detecting outliers.

2) Hierarchical algorithm: A hierarchical clustering method perform grouping of data objects into a tree of clusters. Hierarchical clustering methods can be classified as agglomerative or divisive. In agglomerative method each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained. Agglomerative method is also known as bottom up approach. In divisive method splitting of large data set is done that is all objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained. Divisive method is also known as top-down approach.

Related Issues: Implementation of Hierarchical algorithm does not require a apriori information about the number of clusters and it is easy to implement and gives best result in some cases but this algorithm can never undo what was done previously. The advantage of using hierarchical algorithm is that it embeds flexibility regarding the level of granularity and it is easy to handle any form of similarity or distances. It is also applicable to any type of attribute. 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[10]. Density based method and Hierarchical algorithm can be used together for clustering of random data. As DBSCAN algorithm can perform clustering in one scan without outliers it will consume less time for clustering of data and will perform clustering with higher accuracy. Hierarchical algorithm will give a tree structure to the data. In [7], it shows a systematic 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 hierarchical clustering and co-clustering. Hybrid Hierarchical algorithm arranges musical data along with text files.

There are many issues in clustering of random data. Methodology that is used for mining purpose. Number of algorithm exist for clustering of certain data. But clustering Random data and forming the output of it is our concern. So, the methodology which will be used should be able to perform clustering of Random data. (we can also term it as arrangement of data in certain format). There are many algorithms which do not support scalability. When data increases to certain extend the algorithm fails to works for clusterisation of random data our algorithm should be scalable the algorithm should be able to cluster data which are large in size. All this issues should be removed or handled well when random data is clustered in certain format. Hybrid hierarchical algorithm is such an algorithm which removes all those difficulties. This algorithm is used for clustering of Random data.

V. Conclusion

To improve the performance of clusterisation of random data with the help of Hybrid Hierarchical algorithm. The informal description of clustering as finding meaningful groups does not suggest a straightforward way for evaluating clusters or defining an objective function. Many definitions of good clustering exist. Prominently, clustering has been posed as an optimization problem for minimum error or maximum attribute predictability.

In this paper, a new algorithm is introduced called hybrid hierarchical algorithm. This algorithm is the combination of two algorithm i.e density based method and hierarchical algorithm. This algorithm is used for clustering random data. Clustering of random data is difficult task. Hybrid Hierarchical algorithm can be implemented to Random data for cluster formation. The clusters so obtained will be sorted data in certain data format. The Random data gets converted into a well arranged data. This algorithm works with great efficiency and with less complexity.

Acknowledgment

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

References

[1] 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.

[2] Jianbin Huang, Heli Sun, Qinbao Song, Hongbo Deng, and 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.

[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] Ben Kao, Sau Dan Lee, Foris K.F. Lee, David Wai-lok Cheung, and Wai-Shing Ho, "Clustering Uncertain Data Using Voronoi Diagrams and R-Tree Index" IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 9, pp.1219. September 2010.

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

[6] Eduardo Raul Hruschka,, Ricardo J. G. B. Campello , Alex A. Freitas, and Andr´e C. Ponce Leon F. de Carvalho, "A Survey of Evolutionary Algorithms for Clustering" IEEE Transactions on

Systems, man, and cybernetics—part c: applications and reviews, Vol. 39, no. 2. March 2009.

[7] 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.

[8] 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, pp.352. Nov.16-18, 2012

[9] Xiao-Feng Wang and De-Shuang Huang, “A Novel Density-Based Clustering Framework by Using Level Set Method” IEEE Transactions on Knowledge and Data Engineering, Vol. 21, no. 11. pp. 1518. November 2009

[10] 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. pp. 821,822. 2008

[11] Jiawei Han and Micheline Kamber : Data Mining: Concepts and Techniques Second Edition. pp.418. 2006

[12] Raymond T. Ng and Jiawei Han, “CLARANS: A Method for Clustering Objects for Spatial Data Mining” IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 5. pp.1003. September/October 2002.