Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

DENSITY BASED AND PARTITION BASED CLUSTERING OF UNCERTAIN DATA BASED ON KL-DIVERGENCE SIMILARITY MEASURE

Sinu T S[1], Mr.Joseph George
[1,2]*Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, India*

***Abstract*—** Data mining problems are significantly influenced by the uncertainty in data. Clustering certain data has been well studied in many field of data mining, but there is an only preliminary study in clustering uncertain data. Traditional clustering algorithms are mainly on geometric locations. So such methods will not able to find the similarity of uncertain objects that have different distribution and geometrically indistinguishable. In this paper we introduce a divergence method called KL-divergence for finding the similarity of uncertain objects. And this similarity is integrated into both density based and partition based clustering. And also we are comparing the accuracy level of both clustering methods using KL-divergence and using geometric distances as similarity measure and will find better and efficient method for clustering the uncertain objects.

Keywords- Clustering, Cluster validity, KL-Divergence, Uncertain Data.

I. INTRODUCTION

Data mining deals with the difficulty of extracting patterns from the information by paying suspicious attention to computing, communication and human-computer interface issues. In data mining mainly we have two types of data namely certain data and uncertain data. Mining certain data has been well studied in the various areas such as data mining, machine learning, Bioinformatics, and pattern recognition. However, there is only preliminary research on uncertain data. In many applications, data contain intrinsic uncertainty. Numerals of factors contribute the uncertainty such as the random nature of the physical data creation and collection procedure, measurement of error, and data staling. In many cases, the underlying uncertainty can be easily measured and collected. When this is the case, it is possible to use the uncertainty in order to improve the results of data mining algorithms. This is because the uncertainty provides a probabilistic measure of the relative importance of different attributes in data mining algorithms. Many data mining and management techniques need to be carefully re-designed in order to work effectively with uncertain data. This is because the uncertainty in the data can change the results in a subtle way, so that deterministic algorithms may often create misleading results. Data mining techniques used for certain data are also used namely Clustering, Classification, Frequent pattern mining etc.

Clustering is one of the major data mining tasks to group the similar information or data. All clustering algorithms aim of dividing the collection all data objects into subsets or similar clusters. A cluster is a collection of objects which are „similar" between them and are „dissimilar" to the objects belonging to

other clusters.

Clustering certain data has been well studied in the various areas such as data mining, machine learning, Bioinformatics, and pattern recognition. However, there is only preliminary research on clustering uncertain data. The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms considered for certain data. The main characteristics of uncertain data are, they change continuously, we cannot predict their behavior, the accurate position of uncertain objects is not known and they are geometrically indistinguishable. Because of these reason it is very difficult to cluster the uncertain data by using the traditional clustering methods. Clustering of uncertain data has recently attracted interests from researchers. This is driven by the need of applying clustering techniques to data that are uncertain in nature, and a lack of clustering algorithms that can cope with the uncertainty.

In this paper we are using probability distributions, which are essential characteristics of uncertain objects and are considered in measuring similarity between uncertain objects. The well-known Kullback-Leibler divergence is used to measure similarity between uncertain objects and integrate it into partitioning and density based clustering methods to cluster uncertain objects. And also we are comparing the accuracy level of both clustering methods using KL-divergence and using geometric distances as similarity and will find better and efficient method for clustering the uncertain objects using two cluster validity analyses namely Xie-Beni's index and the Fukuyama-Sugeno's Index. This paper also includes a comparison study of the algorithms based on the execution time and memory space usage.

II. STATE OF ART

The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms designed for certain data. Thus, the studies that extended traditional algorithms to cluster uncertain data are limited to using geometric distance-based similarity measures, and cannot capture the difference between uncertain objects with different distributions specifically, three principal categories exist in literature, namely partitioning-based clustering approaches, density-based clustering approaches and possible world approaches.

A) Partitioning-based clustering approaches

It constructs various partitions and evaluates them by using some criteria. In partition based clustering algorithm uses geometric distance to similarity between two uncertain objects. In this clustering algorithm only center for each object is taken. Extend the k-mean algorithm by using expected distance to measure a similarity between two uncertain data objects.UK-mean is an extension to the traditional K-mean algorithm to handle uncertain data object. UK-mean algorithm require to compute expected distance between each object and to obtaining expected distance is very costly because computation of ED function involves probability function. Probability density functions are different and arbitrary. The major computational cost of the UK-mean algorithm is the evaluation of Expected distance Improve the efficiency of the UK-mean algorithm by integrating some pruning techniques, to reduce many Expected Distance(ED) computations. But pruning effectiveness is not guaranteed, as it depend on the distribution of data.

B) Density-based clustering approaches

The presence of uncertainty changes the nature of the underlying clusters, since it affects the distance

function computations between different data points and proposed a technique based on the density functions inorder to find the clusters. And mainly there exists two clustering techniques based on the density namely DBSCAN and FDBSCAN. However, in our case, objects heavily overlap. There is no clear sparse regions to separate object into clusters. Therefore, the density-based approaches cannot work well.

C) Hierarchical clustering techniques

DBSCAN is extended to a hierarchical density based clustering method referred to as OPTIC. An effective (deterministic) density based hierarchical clustering algorithm is OPTICS. Here, the core idea in OPTICS is quite similar to DBSCAN and it is based on the concept of reachability distance between Data points. While the method in DBSCAN defines a large-scale density parameter which is used as a threshold in order to define reachability. It ensures the DBSCAN algorithm is used for different values with this ordering, then a consistent result is obtained. The output of OPTICS algorithm is not the cluster membership, but it is the orders of data points are processed. OPTICS algorithm shares so many characteristics with the DBSCAN algorithm, it is comparatively easy to extend the OPTICS algorithm to the uncertain case using the same approach as that was used for extending the DBSCAN algorithm. It is referred to as the FOPTICS algorithm. In the uncertain case, this value is defined probabilistically, and the consequent expected values are used to order the data points.

D) Possible world approaches

This follows the possible world semantics. A set of possible worlds are sampled from an uncertain dataset. Each possible world consists of an instance from each object. Clustering is conducted individually on each possible world and the final clustering is obtained by aggregating the clustering results on all possible worlds into a single global The goal is to minimize the sum of the difference between the global clustering and the clustering of every possible world. Clearly, a sampled possible world does not consider the distribution of a data object since a possible world only contains one instance from each object. The clustering results from different possible worlds can be drastically different. The most probable clusters calculated using possible worlds may still carry a very low probability.

III. METHODOLOGY

Here we have use KL-divergence as similarity measure for clustering instead of geometric distances in the traditional clustering algorithm. And it is integrated into both partition based and density based clustering. Also we had done a comparison of both algorithm using KL-divergence and traditional K-means algorithm using geometric distances for clustering the uncertain objects. And will find better and efficient method for clustering the uncertain objects using two cluster validity analysis namely Xie-Beni's index and the Fukuyama-Sugeno's Index. This paper also includes a comparison study of the algorithms based on the execution time and memory space usage.

A. Modeling Uncertain Objects and Probability distribution.

This section first models uncertain objects as random variables in probability distributions. Then we consider an uncertain object as a random variable following a probability distribution. We consider both the discrete and continuous cases. If the data is discrete with a finite or Countable infinite number of values, the object is a discrete random variable and its probability distribution is described by a

probability mass function (pmf). Otherwise, if the domain is continuous with a continuous range of values, the object is a continuous random variable and its probability distribution is described by a probability density function (pdf) .Here it takes pdf because we use weather data set for clustering process.

The probability density function of an uncertain data can be calculated by using the following equation.

P(X) =∫P(x) x

B. KL Divergence

In general, KL divergence between two probability distributions is defined as follows:

Kullback-Leibler Divergence:

In the discrete case, let f and g are two probability mass functions in a discrete domain ID with a finite or countable infinite number of values. The Kullback-Leibler divergence between f and g is given as:

$$D(f||g) = \sum f(x) \log f(x)/g(x) \quad (1)$$

In the continuous case, let f and g are two probability density functions in a continuous domain with a continuous range of values. The Kullback-Leibler divergence between f and g is defined as:

$$D(f||g) = \int f(x) \log f(x)/g(x) \quad (2)$$

a) Using KL Divergence as Similarty

It is natural to quantify the similarity between two uncertain objects by KL divergence. Given two uncertain objects P and Q and their corresponding probability distributions, D (P||Q) evaluates the relative uncertainty of Q given the distribution of P. D (P||Q) = E which is the expected log-likelihood ratio of the two distributions and tells how similar they are. The KL divergence is always nonnegative, and satisfies Gibbs‟ inequality. That is, D (P||Q) ≥0 with equality only if P = Q. Therefore, the smaller the KL divergence, the more similar the two uncertain objects.

C. CLUSTERING ALGORITHM

In this section, we present the clustering methods using KL divergence to cluster uncertain objects in these two categories. Here, we present the uncertain k-medoids method which belong to popular partitioning clustering method by using KL divergence .Then in the next Section we presents the uncertain DBSCAN method which integrates KL divergence into the framework of a typical Density-based clustering method DBSCAN. We describe the algorithms of the methods and how they use KL divergence as the similarity measure.

a) Partitioning Clustering Methods

A partitioning clustering method organizes a set of n uncertain object into k clusters, Using KL divergence as similarity, a partitioning clustering method tries to partition objects into k clusters and chooses the best k representatives, one for each cluster, to minimize the total KL divergence. The smaller the value of KL similarity means the better the clustering.

b) Density based Clustering Methods

To demonstrate density based clustering methods based on distribution similarity, we develop the uncertain DBSCAN method which integrates KL divergence into DBSCAN. Different to the FDBSCAN method which is based on geometric distances and finds dense regions in the original geometric space, the uncertain DBSCAN method transforms objects into a different space where the distribution differences are revealed. Initially, every core object forms a cluster. Two clusters are merged together if a core object of one cluster is density reachable from a core object of the other cluster. A noncore object is assigned to the closest core object if it is direct density reachable from this core object. The algorithm iteratively examines objects in the data set until no new object can be added to any cluster.

c) Partition clustering based on geometric distances.

Here we use traditional partition clustering algorithm called k-means for clustering the uncertain data objects. Here we use geometric distances to find the similarity of data objects.

REFERENCES

[1] H.-P. Kriegel, and M. Pfeifle. Hierarchical Density Based Clustering of Uncertain Data. In ICDM Conference, 2005.
[2] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In PAKDD Conference, pp. 199–204, 2006.
[3] Ben Kao Sau Dan Lee Foris K. F. Lee David W. Cheung and Wai-Shing Ho.” Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index” IEEE, 2010.
[4] S.D. Lee, B. Kao, and R. Cheng, “Reducing Uk-Means to k- Means,” Proc. IEEE Int‟l Conf. Data Mining Workshops (ICDM), 2007.
[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” Proc. Second Int‟l Conf. Knowledge Discovery and Data Mining (KDD), 1996.
[6] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander “Optics: Ordering Points to Identify the Clustering Structure,” Proc. ACM SIGMOD Int‟l Conf. Management of Data (SIGMOD), 1999.
[7] http://rda.ucar.edu/data sets/ds512.0/: weather data set from the National Center for Atmospheric Research data archive.
[8] Bin Jiang, Jian Pei, Senior Member, IEEE, Yufei Tao,Member, IEEE, and Xuemin Lin, Senior Member, IEEE“Clustering Uncertain Data Based on Probability Distribution Similarity” IEEE Transactions on knowledge and data engineering, vol. 25, no. 4, APRIL 2013
[9] C. C. Aggarwal and P. S. Yu. A Framework for Clustering Uncertain Data Streams. In *ICDE Conference*, 2008.
[10] Guha, S., Rastogi, R., and Shim K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of the ACM SIGMOD Conference.