A Review of Clustering Algorithms for Clustering Uncertain Data

Ajit B. Patil
ME Computer Student
Department of Computer Engineering
JSCOE PUNE
Pune, India
ajitbpatil99@gmail.com

Prof. M. D. Ingle.
Assistant Professor
Department of Computer Engineering
JSCOE PUNE
Pune, India
ingle.madhav@gmail.com

Abstract— Clustering is an important task in the Data Mining. Clustering on uncertain data is a challenging in both modeling similarity between objects of uncertain data and developing efficient computational method. The most of the previous method extends partitioning clustering methods and Density based clustering methods, which are based on geometrical distance between two objects. Such method cannot handle uncertain objects that are cannot distinguishable by using geometric properties and Distribution regarding to object itself is not considered. Probability distribution is an important characteristic is not considered during measuring similarity between two uncertain objects. The well known technique Kullbak-Leibler divergence used to measures the similarity between two uncertain objects. The goal of this paper is to provide detailed review about clustering uncertain data by using different methods & showing effectiveness of each algorithm.

Keywords-*Clustering, Uncertain data, probability distribution.*

*****

I. INTRODUCTION

Clustering is a process of partitioning a set of objects into a set of meaningful subclasses is called cluster. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. Clustering is also called as "unsupervised classification" i.e. there is no predefined classes. A good clustering method produces cluster with high quality in which intra-cluster similarity is high and inter-cluster similarity is low. Clustering used in wide range of applications such as pattern reorganization, clustering web log data to find groups of similar access pattern, create thematic map in GIS by clustering features of spaces. In many modern application ranges, e.g. the clustering of moving objects or sensor databases, only uncertain data is available.

Clustering of uncertain data is recognized as important issue. The problem of clustering uncertain data has been studied and find out solutions on this problem. The most of the previous clustering algorithm for clustering uncertain data are extension to the existing clustering algorithms which are designed for clustering uncertain data. But extended existing algorithms to clustering uncertain data are limited because they use geometric distance to measure similarity.

II. CLSTERING UNCERTAIN DATA

Uncertainty in data brings new in clustering of uncertain data. The most of the methods for clustering of uncertain data is based on a measurement of similarity between two objects of a uncertain data. Only few methods are using a divergence to measuring the similarity between two objects.

In this paper we provide a survey on different clustering algorithms for uncertain data. There are three main classes for clustering uncertain data. All this methods are based on a "Measurement of similarity".

1. Partition based Clustering methods:-
   Construct various partitions and evaluate them by using some criteria.
2. Density based clustering methods:-
   Clustering is based on density or radius (local cluster criterion), such as density-connected points
3. Possible world methods: - It is by taking a set of possible world are sampled from an uncertain data set.

1) Partition Based Clustering

Construct various partitions and evaluate them by using some criteria. In partition based clustering algorithm uses geometric distance to similarity between two uncertain objects.

a) UK-mean:-

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 [101] 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(ED).

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. Improve the performance of UK-mean not by pruning technique but by deriving a simple formula for ED computation. The new formula can reduce the computation time of UK-mean.

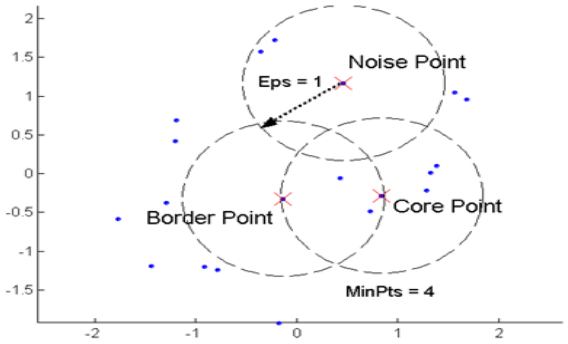
If every object has the same cluster representative, the partition based approaches cannot distinguish two sets of uncertain object having different distributions.

2) **Density Based Clustering**

These methods are based on the density and connectivity functions. Density based clustering algorithm are very popular because their ability to find clusters in arbitrary shape. Core idea is Density/Radius (range) is a constraint on objects to form a cluster. For each object of cluster contains at least minimum number neighborhood objects in given radius. We will discuss two types density based methods FDBCAN and FOPTICS.

a) FDBSCAN

Traditional DBSCAN [1] clustering algorithm clusters a data set by always adding objects to the current cluster which are directly density reachable from query object o..



The core idea of DBSCAN is for each object in cluster that there are minimum numbers of a point in neighborhood of that point. It has a problem to identify a cluster for varying density. Extend the traditional FDBSCAN [2] clustering algorithm to cluster uncertain data. Represent distance between two uncertain objects by "Distance density function" and "Distance Distribution function".

1) Distance density function:

Express distance between two objects probability density function. Let d be the d be the density function.

P (a<=d (O, O') <=b) denote the probability that d (O, O') is between a and b.

2) Distance distribution function:

It captures the probability that is the distance between two uncertain objects is smaller than or equal to a value of b.

The fuzzy version of DBSCAN [1] algorithm is known as FDBSCAN [2] algorithm. It works similar to the DBSCAN algorithm, except the density of a given point because uncertainty in data points. This corresponds to the core idea of a number of points within the density or neighborhood of a given point which are obtained only probabilistically and is an uncertain variable. Reachability of one point from another point is no longer deterministic. Other data points with certain probability are lies within neighborhood of a given point which may be less than 1. The additional constraint on reachability and probability both are must be greater than 0.5 is added. It is a generalized version DBSCAN algorithm in which reachability and probability are set 1.

b) FOPTICS

The OPTICS [6] algorithm cannot produce a clustering of a data set explicitly, but instead it creates the augmented ordering the database to represents its density based clustering structures. The proper cluster structure on many real data sets cannot be obtained by using global density parameter. There are different local densities are needed to show clusters in different region of data space.

In FOPTICS [7] algorithm show the similarity between two fuzzy objects by using probability function. It uses fuzzy distance function to measure the similarity between two uncertain data objects. Integrate the fuzzy distance function into the OPTICS algorithm, resulting clustering algorithm is also called as "FOPTICS".

The OPTICS [6] algorithm can share many characteristics with DBSCAN [1] algorithm; It is easy to extend the OPTICS algorithm in uncertain data using same approach that was used for extending DBSCAN.

The objects are heavily overlap, there is a no clear sparse to divide objects into clusters. So, the density based clustering approaches cannot work well.

c) Possible World Clustering

A Set of possible world are taken or sampled from an uncertain data set [13] to cluster uncertain data. Instance of each object consist in the each possible world. Clustering process is performed individually on each possible world. To obtaining the final clustering by combining results of all possible world [13] into a single global clustering. But sampled possible world does not consider distribution of data objects because it contains only one instance from each object. So the clustering is resulting from a different possible worlds can be different for a different possible world. The most of probable clusters are obtained using possible world approach show a very a low probability. The possible world does not provide meaningful and stable clustering result at the object level, is this method computationally infeasible because of there are exponential number of possible worlds.

 ISSN: 2321-8169
3643 - 3646

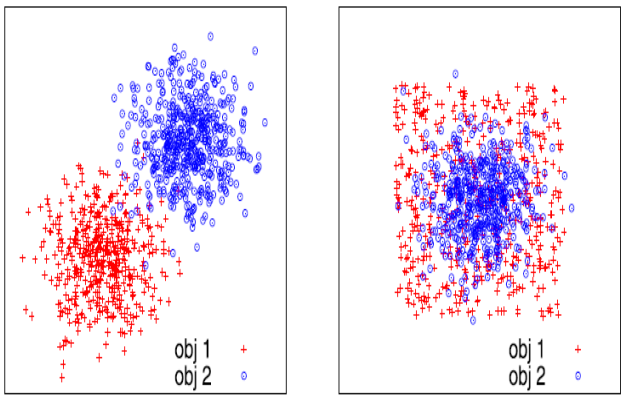
III. COMPARISION OF DIFFERENT CLUSTERING ALGORITHM FOR CLUSTERING UNCERTAIN DATA

Table1:Comparision of different algorithm used for clustering Uncertain Data.

Title	Algorithm	Limitation
Uncertain Data Mining: An Example in Clustering Location Data [1]	UK-mean	1) Computes expected distance each object-cluster pair in each iteration 2) Cannot distinguish the two sets of objects having different distribution.
Reducing UK-means to k-means [4]	K-mean	1) Effectiveness is not guaranteed, as it is depends on the distribution of data. 2) The restriction to use Mean squared distance for computation of expected distance.
Density based clustering algorithm for discovering cluster in large spatial database with Noise [1]	DBSCAN	1) Has a problem if identifying cluster with varying density. 2) Doesn't work in high dimensional database 3) Needs a large number of input parameter.
Density based clustering of uncertain data [2]	FDBSCAN	1) Proper cluster structure cannot be obtained by using global density parameter. 2) Selection of input parameter is complex task.
Efficient Clustering of Uncertain data [5]	UK-mean	1) Pruning min-max technique does not consider geometric structure and spatial relationship among cluster center.
Clustering with Bregman Divergence [11]		1) Did not provide method for efficiently evaluating Bregman Divergence nor calculating mean of set of objects.
Cluster based Language Model for Distributed Retrieval [10]	K-Mean	Multinomial distribution based methods not applied to general cases.

Each of the above algorithm is limited it considers only geometric characteristics of uncertain data objects. They cannot consider distribution similarity between uncertain data object for clustering. As an object in certain data set is a single point, the distribution regarding to the object itself is not considered in traditional algorithm. They focus on the geometrical properties of objects and they do not take into account probability distribution of object.

For example we have two data sets A and B of an uncertain object. As shown in figure 1. The objects are in A and B fallows the different distribution but all the objects in both dataset have same mean value or cluster center. Geometric location of objects are heavily Overlap. These two sets of object form two different clusters due they have different distribution



(a) (b)

Figure1:- different probability distribution and geometric locations

IV. CLUSTERING BASED IN DISTRIBUTION SIMILARITY

Cluster the uncertain object by measuring similarity between probability distributions of a two uncertain object. Clustering based on distribution similarity done by using Bregman Divergence. The Bregman divergence [11] summarizes clustering framework for measuring similarity. Efficiency of this algorithm has linear complexity with respect to the number of objects.

Measure the similarity between probability distribution of the object by using "Kullback-Leibler" divergence [12]. The KL divergence capture difference between two objects easily. Kullbak-Leibler [8] divergence has more effective and scalable than Bregman divergence.

V. CONCLUSION AND FEATURE WORK

The most of the previous algorithm for clustering uncertain data based are extension to Partition based & Density based algorithm. These algorithms are limited because measure similarity using geometric distance and they cannot capture the distribution similarity between uncertain object.

Our feature work is integrating effectiveness of Kullbak-Leibler Divergence into the both Density based and based clustering algorithm and reduce the computation of KL-Divergence for clustering uncertain data based distribution similarity.

REFERENCES

[1] 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.

[2] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.

[3] Michael Chau1, Reynold Cheng2, Ben Kao3, and Jackey Ng1 "Uncertain Data Mining: An Example in Clustering Location Data"

[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] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.

[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] H.-P. Kriegel and M. Pfeifle, "Hierarchical Density-Based Clustering of Uncertain Data," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2005.

[8] Bin Jiang, Jian Pei "Clustering Uncertain Data Based on Probability Distribution Similarity" IEEE 2005

[9] B. Kao, S.D.Lee, D.W.Cheung, W.-S. Ho, and K.F. Chan, "Clustering Uncertain Data Using Voronoi Diagrams," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2008.

[10] J. Xu and W.B. Croft, "Cluster-Based Language Models for Distributed Retrieval," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 1999.

[11] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.

[12] S. Kullback and R.A. Leibler, "On Inf2132ormation and Sufficiency," The Annals of Math. Statistics, vol. 22, pp. 79-86, 1951.

[13] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering Uncertain Data with Possible Worlds," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009.