ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 6, June 2013

An Approach for clustering uncertain data objects: A Survey

Samir N. Ajani, Prof. Mangesh Wanjari

***Abstract* - Recently, uncertain data objects is used in various applications such as VANET environment, sensors applications, image processing based system etc. Clustering of uncertain data is a major concept in data mining since more and more applications, such as sensor database, location database, biometric information systems, and produce vague and imprecise data. Clustering of uncertain data objects is a challenge in spatial data bases. Clustering is a process of organizing objects into groups whose members are similar in some way. There are lots of approaches used to classify the uncertain data by hard classifiers; few of them address the classification of the uncertain data by soft classifier. This paper describes the clustering of uncertain dataset .clustering of object is done by using indexing technique. Probability Density Functions (PDF) is used to represent uncertain data objects. K-Means algorithm is used to generate the clusters. Voronoi diagram is an important technique for answering nearest-neighbor queries for spatial databases. To improve the performance of k-Means, this algorithm is combined with Voronoi diagram.In this paper, we study how the Voronoi diagram can be used on uncertain data, which are inherent in scientific and business applications. Then we conclude our clustering approach**.

Index Terms: *Uncertain data objects, Voronoi Diagrams, Clustering*.

I. INTRODUCTION

Clustering of uncertain data is a great topic in data mining since more and more applications, such as sensor database, location database, biometric information systems, produces vague and imprecise data. Clustering is method to group the data in a similar element groups. In this group, the group members are having some similarity between them. Recently, there is a drastic need to handle a variety of types of data kept in large computer system which results the development of multimedia database systems having target of uniform management of voice, video, image, text, and numerical data. Among the many research challenges which the multimedia technology entails -including data placement, presentation, synchronization, etc. - content-based retrieval plays a dominant role. In order to satisfy the information needs of users, it is of vital importance to effectively and efficiently support the retrieval process devised to determine which portions of the database are relevant to users' requests .In particular, there is an urgent need of indexing techniques able to support execution of similarity queries. Since multimedia applications typically require complex distance functions to quantify similarities of multi-dimensional features, such as shape, texture, color , image patterns, sound ,text, fuzzy values, set values , sequence Data etc., multi-dimensional (spatial) access methods (SAMs), such as R-tree .Rest of the paper has section-ii which describes the related work , section –iii which describes the proposed approach and working, section iv has partial experimental result, section v has conclusion and result of proposed approach. The Voronoi diagram is an important technique for answering nearest-neighbor queries for spatial databases. In this paper, we study how the Voronoi diagram can be used on uncertain data, which are inherent in scientific and business applications. In particular, we propose the *Uncertain-Voronoi Diagram* (or *UV-diagram* in short). Conceptually, the data space is divided into distinct "UV-partitions", where each UV-partition P is associated with a set S of objects; any point q located in P has the set S as its nearest neighbor with non-zero probabilities. The UV-diagram facilitates queries that inquire objects for having non-zero chances of being the nearest neighbor of a given query point. It also allows analysis of nearest neighbor information, e.g., finding out how many objects are the nearest neighbors in a given area. However, a UV-diagram requires exponential construction and storage costs. To tackle these problems, we devise an alternative representation for UV-partitions, and develop an adaptive index for the UV-diagram. This index can be constructed in polynomial time. We examine how it can be extended to support other related queries.

II. LITERATURE SURVEY

Author Le Li Zhiwen discussed automatic soft classifier to classify the uncertain data. The automatic soft classifier first combines Fuzzy C means with a fuzzy distance function to assign the uncertain data into their corresponding clusters. [1]Carson Kai-Sang describes the Naive approach it is for finding constraint frequent pattern from uncertain data is to find all the frequent patterns first

, and the checks these frequent patterns against the user constraints as a post-processing step – to filter out the patterns that do not satisfy the constraints. [2]Author Haiquan Chen gives the techniques to Construct the generation tree based on possible world and make use of branch-and-bound technique to prune the tree [3] Author Paolo Ciaccia Marco Patella Pavel Zezula gives a new indexing for accessing data, called M-tree to organize and search large data sets from a generic "metric space", i.e. where object proximity is only defined by a distance function satisfying the positivity, symmetry, and triangle inequality postulates. The M-tree partitions objects on the basis of the irrelative distances, as measured by a specific distance function d, and stores these objects into fixed-size nodes,2 which correspond to constrained regions of the metric space. The M-tree is fully parametric on the distance function d, so the function implementation is a black-box for the M-tree. l it is a paged, balanced, and dynamic secondary memory structure able to index data sets from generic metric spaces [2][3].Author Prithviraj Sen describes the Framework based on probabilistic models, for explicitly modeling correlations among tuples based on probabilistic database.Charu C. Aggarwal Provide a survey of uncertain data mining and management applications. [1]Author Le Li Zhiwen Yul Zijian Fengl Xiaohang Zhang describes the automatic classification technique by soft classifier for the classification of uncertain data which appears in databases such as sensor ,location biometrics information databases with uncertainties .This data is generally imprecise in nature. This soft classifier technique is based on fuzzy c-means method with a fuzzy distance function to classify uncertain objects. The advantage of this method is that it works very well in uncertain data objects but not in certain data objects.[4] Author Charu C. Aggarwal, Philip S. Yu, discussed various methodologies for processing and mining uncertain data. In this a probabilistic database is used. A probabilistic-information database is a finite probability space whose outcomes are all possible database instances consistent with a given schema. This probabilistic database is also called as "possible world's model" .The specification of such databases is unrealistic, since an exponential number of instances would be needed to represent the table. Therefore, the natural solution is to use a variety of simplified models which can be easily used for data mining and data management purposes. We will discuss more on this issue slightly later. The information stored in probabilistic database can be easily represented by Probabilistic tables. After reviewing various indexing techniques and clustering techniques w found that if a clustering of uncertain data should be done with the combination of some indexing technique then the clustering of uncertain data can be done very easily.

III. PROPOSED PLAN FOR APPROACH

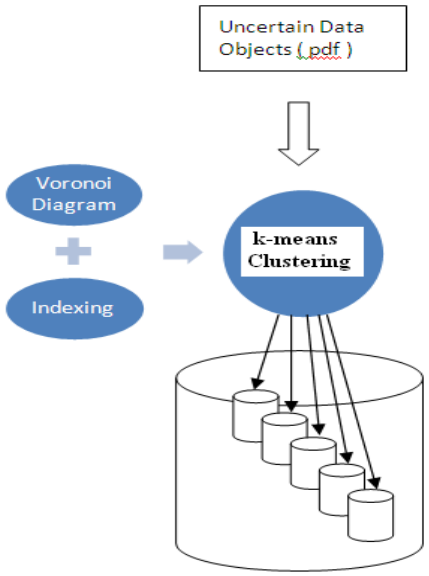

Fig.1. Proposed Approach plan.

Figure (1) shows the architecture of proposed plan. As shown in diagram a uncertain data set is used for clustering. Uncertain data set can be set of any existing objects in the world. For ex. various types of trees in a dense forest, mobile vehicles in the vanet environment, and various sensor devices in the sensor network etc. For clustering K-means algorithm is used , in which the goal is to minimize sum of square error (SSE).which is discussed in section –I. The basic idea behind the uncertain k-means algorithm is to minimize the expected sum of squared errors. UK-Means algorithm is a previous k-means algorithm to handle uncertain data objects. The UK-Means algorithm could be characterized as the least robust of all the methods. Its insensitivity to variance within a distribution can be viewed as a major flaw, especially given that the distributions of the features in the cells are extremely variable. It has been observed that simply applying K-means algorithm for clustering uncertain data set will more time .Hence for reducing the execution time and increasing the performance time we are proposing this plan in which K-means algorithm is integrated with Voronoi diagram concept. The Voronoi diagram concept is discussed in section –II. This proposed approach can be implemented for increasing the performance of the K-means algorithm in terms of execution time and better clustering results. This K-means algorithm is combined with indexing and voronoi to increase the performance of the K-means algorithm

.

IV. CONCLUSION AND FUTURE WORK

In this paper the various clustering techniques, indexing techniques is discussed. Also this paper presents models used for representing var. To handle these problems a new system uncertain data objects. It has been observed that if the clustering algorithm is combined with indexing method then the clustering of uncertain data objects can be done very easily. This paper proposes a plan in which a K-means algorithm is used with Voronoi Diagram and indexing method to increase the performance of K-Means algorithm. In future this proposed plan can be implemented to prove the increased performance of K-Means algorithm

REFERENCES

[1] "Clustering Uncertain Data With Possible Worlds"Peter Benjamin Volk, Frank Rosenthal, Martin Hahmann, Dirk Habich, Wolfgang Lehner,IEEE International Conference on Data Engineering.
[2] "An Efficient Distance Calculation Method for Uncertain Objects", Lurong Xiao, Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007).
[3] "A Survey of Uncertain Data Algorithms and Applications" ,Charu C. Aggarwal, IEEE Transactions on knowledge and data engineering, VOL. 21, NO. 5, MAY 2009.
[4] "UV-Diagram: A Voronoi Diagram for Uncertain Data" Reynold Cheng, Xike Xie , IEEE ICDE Conference 2010.
[5] "Automatic Classification of Uncertain Data by Soft Classifier" Le Li Zhiwen Yul, Zijian Fengl, Xiaohang Zhangl,Proceedings of the 2011 International Conference on machine Learning and Cybernetics, GuiJin, 10-13 July, 2011
[6] "Distance-Based Outlier Detection on Uncertain Data", Bin Wang, Gang Xiao, Hao Yu, Xiaochun Yang, IEEE Eleventh International Conference on Computer and Information Technology,2011.
[7]"Clustering Uncertain Data With PossibleWorlds", Peter Benjamin Volk, Frank Rosenthal, Martin Hahmann, Dirk Habich, Wolfgang Lehner, IEEE International Conference on Data Engineering 1084-4627/09 $25.00 © 2009 IEEE.
[8] "Clustering Uncertain Data using Voronoi Diagrams", Ben Kao Sau Dan Lee David W. Cheung Wai-Shing Ho K. F. Chan, IEEE 2010.
[9] "An Efficient Distance Calculation Method for Uncertain Objects", Lurong Xiao, Edward Hung Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)
[10]. "Automatic Classification of Uncertain Data by Soft Classifier", Le Li\ Zhiwen Yul2* Zijian Fengl Xiaohang Zhangl, Proceedings of the 2011 International Conference on machine Learning and Cybernetics, GuiJin, 10-13 July, 2011
[11]" UV-Diagram: A Voronoi Diagram for Uncertain Data", Reynold Cheng , Xike Xie , Man Lung Yiu , Jinchuan Chen , Liwen Sun, ICDE Conference2010IEEE .

First Author : Samir N. Ajani, Computer Science & Engineering Department, Autonomous, SRCOEM, Nagpur, Maharashtra -440013, India.

Second Author : Prof. Mangesh Wanjari, Computer Science & Engineering Department, Autonomous, SRCOEM, Nagpur, Maharashtra -440013, India.