Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Clustering of Uncertain Data Objects using Improved K-means Algorithm

**Samir N. Ajani**
*Computer Science & Engineering Department, Autonomous, SRCOEM, Nagpur, India.*

**Prof. Mangesh Wanjari**
*Computer Science & Engineering Department, Autonomous, SRCOEM, Nagpur, India.*

***Abstract - Recently data mining over the uncertain data attracts more attention of the data mining. The uncertainty occurs in a information because of the inaccurate measurement of the results, like scientific results, data gathered from sensor network, measuring temperature, humidity, pressure and so on. from such a sources there is possibility of getting the uncertainty in a data. Main task is to handle the uncertainty of the data in order to classify or cluster it. Many efficient algorithms, including the well known and widely applied k-means algorithm is widely used. Simply applying K-Means algorithm clusters the data takes large computation time. Hence if the indexing techniques are applied to the K-means algorithm then the cluster generation time is significantly reduced and the clustering will be done more clearly. This paper describes a brief idea of clustering, indexing, k-means algorithm, clustering of uncertain dataset. Also this paper proposes approach which will minimizes the computation time for clustering uncertain data. Then we conclude our approach.***

***Keywords - Uncertain data objects, K-means, Voronoi Diagrams, Clustering, Indexing.***

## I. INTRODUCTION

The topic of managing uncertain data has been explored in many ways. Different methodologies for data storage and query processing have been proposed. As the Availability of management systems grows, the research on analytics of uncertain data is gaining in importance. Unsupervised classification is a method where there are no predefined classes subsist. Grouping means creating a set of data into cluster is called cluster analysis. A good clustering technique produces high quality cluster with high intra-class similarity and lower inter-class similarity. Clustering applications are widely used in pattern reorganization; spatial data object analysis, Document classification, and Real world application like Marketing, City planning etc. Clustering when applied to a mobile node distributed sensor networks and wireless technology [5] forms a group network with a cluster representative and cluster members. The cluster representative exchanges data and centroid information with the server etc. in the form of batch mode for efficient telecommunication. Traditionally, clustering algorithms deal with a set of objects whose positions are accurately known. The goal is to find a way to divide objects into clusters so that the total distance of the objects to their assigned cluster centers is minimized. Similar to the challenges faced in the field of data management, algorithms for uncertain data mining also have a high performance degradation compared to their certain algorithms. K-means algorithm is widely for the clustering of uncertain data set objects. In K-means K indicates the no of clusters and means indicates the per cluster. K-means clustering algorithm (for point-valued data) is an iterative procedure. Each iteration consists of two steps. In step 1, each object $O_i$ is assigned to a cluster whose representative (a point) is the one closest to $O_i$ among all representatives. We call this step cluster assignment. In step 2, the representative of each cluster is updated by the means of all the objects that are assigned to the cluster. In cluster assignment, the closeness between an object and a cluster is measured by some simple distance such as Euclidean distance. K-Means simply used with clustering method takes long computation time for creating clusters. Hence the efficiency of K-means can be improved by combining some pruning techniques with K-Means algorithm. The rest of the paper is organized as follows. Section II gives a brief overview of the K-means Algorithm with clustering on uncertain data objects. Section III describes the proposed approach and description of techniques clustering used for implementing this approach .Section IV gives the conclusion & future work. The pruning techniques make use of bounding boxes over objects as well as the triangular inequality to establish lower- and upper-bounds of the EDs. Using these bounds, some candidate clusters are eliminated from consideration when UK-means determines the cluster assignment. of an object. The corresponding computation of expected distances from the object to the pruned clusters are thus not necessary and are avoided.

• Interval data
• Initialize means (e.g. by picking K samples at random)
• Iterate:
(1) Assign each point to nearest mean
(2) Move "mean" to center of its cluster

 

## II. CLUSTERING OF UNCERTAIN DATA SET WITH K-MEANS ALGORITHM:

As discussed in section -1 clustering is the main topic for grouping uncertain data objects. K-means is widely used for grouping the uncertain data element in various application. K-means method uses K prototypes, the centroids of clusters, to characterize the data. They are determined by minimizing the sum of squared errors,

$$M_k = \sum_{k=1}^{K} \sum_{i \in C_k} (Z_i - S_k)^2$$

where (z1,z2, ….. ,zn) = Z is the data matrix and $S_k = \sum i \in Ck\ xi/n_k$ is the centroid of cluster Ck and $n_k$ is the number of points in Ck. Standard iterative solution to K-means suffers from a well-known problem: as iteration proceeds, the solutions are trapped in the local minima due to the greedy nature of the update algorithm Fig. 1 shows the clustering of uncertain data set of forest given in fig. 1 .

**Example 1:**

Data set used is: Forest Cover type data
No. of objects: 5,81,012.
No. of Attributes: 54 columns of data:
10 quantitative variables,
4 binary wilderness areas
40 binary soil type variables

**Output:**

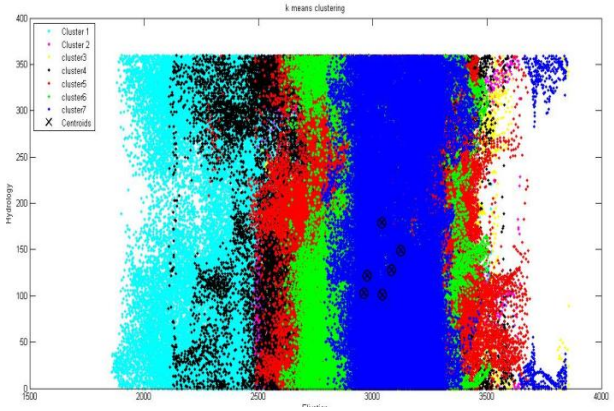


Fig. 1 Output of K-means algorithm.

In fig. 1, it clear that when simply K-means is applied to clustering technique it clustering of data is set is not done properly. here in fig. 1, 7 clusters are formed named red, blue, green, cyan ,black, pink,& yellow respectively. Here although a clusters are formed but the cluster element of clusters is overlapping with another cluster. As shown black color element is overlapping with cyan group of cluster, red is overlapping with green and pink and so on.As this algorithm is based on iterative methods it takes more time for creating clusters. Approximately it takes 42.817398 seconds for 7 clusters .i.e.

Cluster Number 1, Elements: 98941
Cluster Number 2, Elements: 94048
Cluster Number 3, Elements: 72108
Cluster Number 4, Elements: 104358
Cluster Number 5, Elements: 97324
Cluster Number 6, Elements: 49386
Cluster Number 7, Elements: 64847
Elapsed time is 42.817398 seconds.

The above stated problems of clustering can be overcome by using some indexing technique and Voronoi clustering with K-means algorithm for cluster uncertain data set.

## III. PROPOSED CLUSTERING APPROACH:

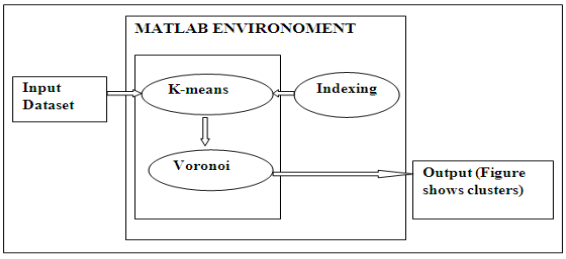


Fig. 2 Proposed Approach plan.

 

Fig. 2, shows the proposed architecture of experiment which is performed on Matlab. This proposed approach is a combination of K-means algorithm with Indexing and vornoi clustering. As shown in fig. 2 the input block will provide the uncertain data set object as a input. In that K-Means algorithm is applied before that an indexing in done on the input given. We create an index table based on the input dataset. K-means algorithm will generate clusters on which again an vornoi clustering is applied to further refine the clustering results. As indexing is done on input data set iteration time of k-means algorithm is reduced which will increase the efficiency of K-means algorithm. By following this proposed approach we will get the good results and also the computation time for creating clusters is reduced.

## IV. CONCLUSION AND FUTURE WORK

In this paper we have discussed clustering technique with K-means algorithm. Although K-means algorithm is widely used for clustering object but it takes large computation time when simply used and also we found that clusters are generated are not proper. Cluster elements of one group are overlapped with another group. Hence to overcome these problems of K-means algorithm some indexing technique may be added to k-means algorithm. We are proposing an approach which will overcome their problems. The approach given in this paper when implemented may improve the efficiency 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 Conference 2010 2010 IEEE 796*


 