An Approach to Cluster data using Voronoi and applying SVM for Outlier's

Maahi A. Talreja[1] Sheetal Rathi[2]
[1,2]Department of Computer Engineering
[1,2]TCET Mumbai, India

Abstract— Data mining an interdisciplinary research area spanning several disciplines such as expert system, database system, statistic, machine learning and intelligent information systems. In last few decades Data mining has becomes a very important and active area of research because of previously unknown and interesting knowledge from very huge data present in day to day life. Different approaches of the data mining are studied and used in several real time applications. When the data sets are analysis different aspects are consider. Clustering is important techniques to form a group of similar or items having similar properties. But the main problem with these clustering of data sets is outliers. Outliers are the unwanted data or similar data in the clusters. In this paper we proposed a system using K-Means clustering algorithm, Voronoi indexing techniques to form a structured cluster and then use the Support Vector Machine to find out the outliers from the clusters.

Key words: K-Means Clustering, Voronoi Indexing, Data Mining, Support Vector Machine

I. INTRODUCTION

In the present situation the technologies are used by humans to adequate in there society. Everyday a vast amount of data is used and these data are in various forms. It may be in the form of graphical formats, may be documents, may be the records (varying array) or video. In the real world different data is available in the different formats so that the proper actions have to be taken for better utilization of the datasets available. The data should be stored such that when the humans need that data it can be retrieved easily.

The technique of extracting knowledge from any type of data is called as data mining or knowledge hub or simply KDD. The important aspect that attracted a great deal of attention in information technology is the discovery of useful information from large amount of data produced in the industry. The perception "Data Mining" is due to the quote "we are data rich but information poor". A wide variety of rich data is available but we can hardly turn them into useful information and knowledge for decision making in the industry. To generate information it requires very huge datasets. Databases are available in variety of formats like audio/video, numbers, hypertext, figures, and text formats. To make complete advantage of this data, it requires a tool for extraction of the essence/ information/ knowledge of data stored in the database and discovering the pattern and actual data by summarizing automatically.

An outlier is an observation seen in the data, it is that data which deviates from other observed result so much that it arouses suspicions that it was generated by a different mechanism from the most part of data [1]. Whereas Inlier, on the other hand, is defined as an observation that is explained by using probability density function. This function represents probability distribution of main part of data observations [2].

In many real time applications clustering is mostly used for the grouping the elements with similar properties. Moat popular algorithm for the clustering is the K-Means algorithm. However, all these algorithms deal with objects whose distance is well known. The goal of this paper is to work with uncertain database where distance between objects is dynamic and not known. Grouping objects into clusters will make the very easy to use them in application. For example mobile devices can be grouped and one of the devices can be elected as leader for better coordination among them.

In this paper, the problem of clustering is considered for uncertain objects whose locations are specified by uncertainty regions over which arbitrary probability density functions are defined. So in this paper we used voronoi indexing to remove the uncertainty from the clusters

II. RELATED WORK

The uncertainty of the data is represented with the help of the algorithm in [3] the major challenge here was to handle the uncertainty of the data. To handle the uncertainty of data the PDF has important role in it.

When the traditional clustering methods were used their main aim is to find a unique factor which will form a cluster for the same object with totally new cluster or exiting by minimalizing the sum of the square error (SSE)[3] in case of the k-mean algorithm. In the data mining field the uncertainty is a major cause, whereas the tuple values are most important factors. The presence of the uncertainty anywhere is the major concern of the field. Here if the data is having less uncertainty than it is given more importance than the data having more uncertainty. The uncertainty factor has majorly two types, existential uncertainty and value uncertainty. If it is uncertain that whether an object or a data tuple exists there with uncertainty, such situation is called as existential uncertainty.

The tuples of the relational database could be associated with the probabilistic data; the probability here represents the existence of the object there in database [4]. When the data is present but its value is not known precisely then the value of the uncertainty is unity [4]. If there exists a data item who is having a value of uncertainty is usually represented by the PDF function over the constrained finite field of the possible values. UK-Mean algorithm [5], helps in getting the important data from the uncertain data, it takes multiple attempts to do this. There has been growing interest in uncertain data mining. In [6], a technique is provide which extends the k-means algorithm including the UK-Means algorithm so that the clusters of uncertain data and unknown data can be made. In different papers, it is empirically shown that clustering results are improved if

data uncertainty is taken into account during the clustering process. The data uncertainty is usually captured by the PDF function, and they are usually defined by the set of values. Since there is an explosion of information in uncertain data, the mining of uncertain data is difficult and computationally costly. (set's of samples vs. singular values).

To improve the performance of UK-means, CK-means [7] introduced a novel method for computing the EDs efficiently. But, this method proposed by author only works for a specific form of distance function. For general distance functions, [8] takes the pruning approach, and proposed different techniques such as min-max-dist pruning. In [9], there are various approximation algorithms have been used for clustering uncertain data using k-means, k-median as well as k-centre methods. Clustering of uncertain data is also related to fuzzy clustering, fuzzy clustering is been studied by different researcher for a long time in fuzzy logic. In fuzzy clustering, a cluster is represented by a fuzzy subset of objects. Each object has a "degree of belongingness" with respect to each cluster.

The fuzzy clustering methods widely used is the fuzzy c-means algorithm. To produce fuzzy clusters multiple different fuzzy clustering methods have been applied on normal or fuzzy data set. A major difference between the algorithm used in this paper and the fuzzy clustering is about focus, in general we focus on the hard clustering. Due to hard clustering it is important that each object belongs to exactly one cluster.

The formulation targets for applications such as mobile device clustering, in which each device should report its location to exactly one cluster leader from the structure. Voronoi indexing is a well-known geometric structure in computational geometry of the datasets. This technique is also been used in clustering. For example, Voronoi trees [10] have been proposed to answer Reverse Nearest Neighbors (RNN) queries. Given a set of data points and a query point q, the RNN problem is to find all the data points whose nearest neighbor is q.

To solve this problem efficiently the TPL algorithm has proposed more advanced pruning techniques. An R- tree resembles a B+- tree which is a self-balancing tree structure, except that it is devised for indexing multi-dimensional data points to facilitate proximity-based searching, such as k-nearest neighbour (kNN) queries. Rtrees are well studied and widely used in real applications. SQLite, MySQL and Oracle are different RDBMS products available. An R-tree conceptually groups the underlying points hiearchically and records the minimum bounding rectangle (MBR) of each group to facilitate answering of spatial queries. Since the major use of Rtree concentrate on optimizing the tree and answering the queries, here in this paper we will use Rtree in innovative way. We exploit the hierarchical grouping of the objects organized by an R-tree to help us check pruning criteria in batch, thereby avoiding redundant checking.

III. CLUSTERING ALGORITHM

Clustering is the most important and unsupervised learning problem occurred in data mining. That is all the problems deals with the finding a structure that is cluster of the data items in a collection of data. The definition of the clustering states that "the process of organizing objects into groups whose members are similar to each other". In a general way of speaking we can say that in a cluster similar kind of data items occur which are dissimilar to the other data objects in other clusters.

We can explain this with a very simple example as follows:

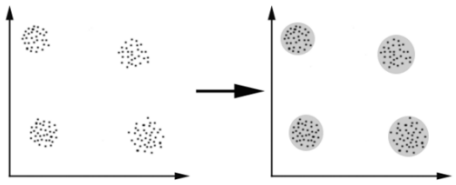

Fig. 2: Graphical Presentation of clustering

In the above case we can easily observe that there are 4 clusters into which the data can be divided accordingly. One of the criterion is the distance that is if two or more objects are close according to the geometric distance then this clustering is called distance-based clustering. Conceptual clustering is another type of clustering here there is a concept defined through which the nodes are compared and are clustered. Here the similarity factor is ignored.

The main aim or the goal of clustering is to determine the grouping constraint of a set of set of data. The aim leads to the important question that how to choose a better clustering algorithm/concept.? The solution is purely based on the user and the work. As according to the result the user will specify the data which he needs in a particular set so the user has to define the cluster concept and should work accordingly.

There are different clustering algorithms proposed by different authors in the last few decades. In this paper we are going to implement the K-Means Clustering algorithm on the different datasets to obtain the clusters from those datasets:

A. K-Means Clustering

The k-means method is simple and is globally accepted for employing a squared error criterion. The algorithm follows the steps. Firstly the k-means algorithm randomly partitions the data it then keeps reassigning the patterns that is clusters are made on the basis of the similarity between the data and the cluster concepts. The process is completed when there is do data to be divided and saturation occurs. (e.g., there is no reassignment of any pattern from one cluster to another, or the squared error ceases to decrease significantly after some number of iterations). The time complexity of the algorithm is showed as O(n), here n specifies the number of patterns. The problem occurs when the initial partition is not properly taken leading to the convergence of a local minima of the criteria.

In the partition clustering algorithm it obtains a single partition of the data instead of clustering them the example of this is the dendrogram produced by a hierarchical technique. This technique is advantageous as if there is a set of large data sets it is easy to compute such as in the dendrogram. Here problem occurs when we have to choose the number of the output clusters. To implementation of partition algorithm is as follows, Firstly the partition system produces the clusters not by general clustering but it optimizes the criterion function. The function can be defined locally of globally. Here we can use a set of data again and again to get the best output of the all. Therefore, here with

the data set the configuration and the starting states can be changed so that we get the best results and can be compared.

The partition clustering technique popularly uses the criterion function in the squared error criterion, it works best with the compact and isolated clusters. The squared error for a clustering y of a pattern set x (containing *K* clusters) is

$$e^2(x,y) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} ||x_i^{(j)} - c_j||^2 ,$$

Where $x_i^{(j)}$ is the j^{th} pattern belonging to the j^{th} cluster and c_j is the centroid of the j^{th} cluster.

IV. SUPPORT VECTOR MACHINE

If the machine is capable of learning then its main objective is to achieve good generalization performance when given adequate data for training and by striking a balance between the goodness of fit attained on a given training data set. It should also achieve error free recognition on all the dataset one it is trained. With this concept as its base the support vector machines have proved to achieve good generalized performance with the no knowledge of the data input.

The SVM usually place the input data into a higher dimensional feature space which is non linearly related to the input space. It also determines a separate hyper-plane with maximum margin between the two classes in the feature space [11]. This result is always displayed in a non linear boundary within the input data space. Without any computation the optimal separating hyper-plane can be determined, in the high dimensional feature space, it uses the kernel function in the input space. Commonly used kernels include:

1) *Linear Kernel:*

$$K(x, y) = x.y$$

2) *Radial Basis Function (Gaussian) kernel:*

$$K(x, y) = \exp(-||x - y||^2/2\sigma^2)$$

3) *Polynomial Kernel:*

$$K(x, y) = (x.y + 1)^d$$

An SVM in its elementary form can used for binary classification. It may, however, be extended to multiclass problems using the one-against-the-rest approach or by using the one-against-one approach.

V. VORONOI INDEXING DIAGRAM

To find the centroid of all the clusters of dataset in data mining we will be using the voronoi indexing. The base of the coverage problem is the voronoi property. As shown below in the varonoi diagram all the points inside a cell are closest to the generating cluster that lies within the cell. Thus, when we complete creating the voronoi diagram of the cluster, if in there is a point of any of the voronoi cell which is not covered by its generating cluster then this point is not covered by any cluster[12][13]. In the voronoi diagram the computation of area of voronoi cell is easier whereas the computation of the area of the uncovered region is complicated because the sensing regions may clash with each other.Figure 2 shows the sample voronoi diagram.

Fig. 2: Sample VORONOI Diagram

VI. PROPOSED WORK

In this paper we proposed the methods to identify the outliers from the different datasets. Outliers are the duplicate data or unwanted data available in the datasets, which need to identified and remove from the datasets to improve the performance. In this paper K-means clustering algorithm is used to form the clusters of the similar elements from the datasets. The biggest drawback of the system is that it generates random clusters. To remove this drawback in proposed method we implemented K-means algorithm with the Voronoi indexing. Voronoi indexing is used to calculate the centroid of the different parts of the datasets. The clustering is done according to the calculated centroid. Figure 3 shows the cluster formation of the datasets with fixed centroid according to the voronoi indexing.

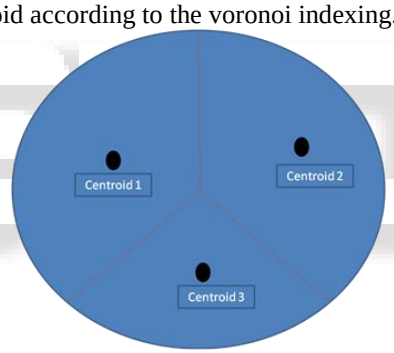

Fig. 3: Formation of cluster centroid using voronoi indexing.

Figure 4 shows the overall implementation block diagram of the proposed work.

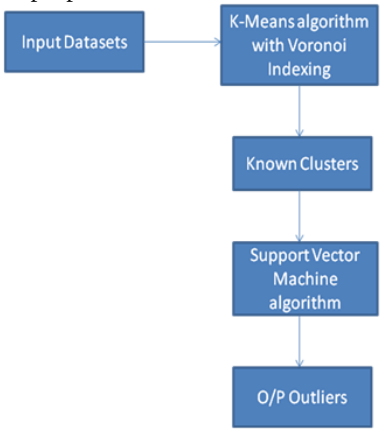

Fig. 4: Block Diagram of Overall Proposed Work

In the proposed method datasets are given as input to the system. Then on the datasets the K-means clustering algorithm is applied with the voronoi indexing. Voronoi indexing finds the centroid for the each cluster and then cluster is from across this centroid. This gives the known clusters instead of random clusters. Then SVM (Support

Vector Machine) on the clusters to find the outliers from the clusters. SVM selects the duplicates elements and an unwanted element from the clusters and gives them as the output of the system.

VII. EXPERIMENTAL RESULT

Dataset	Time	Memory
1000	24	8.24
2000	171	15.489
5000	133	10.48
10000	235	16.92

Table 1: Experimental results of proposed Top K Rules Method

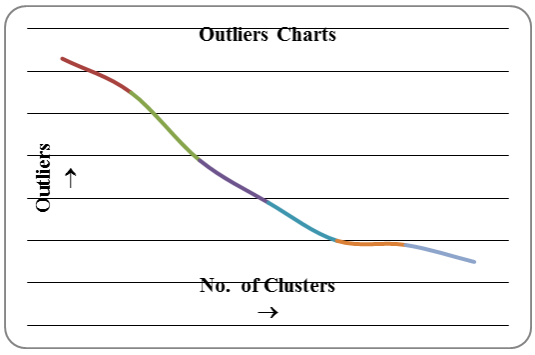

Fig. 5: Graph shows the actual result:

VIII. CONCLUSION

In this paper we proposed a system using K-Means clustering algorithm, Voronoi indexing techniques to form a structured cluster and then use the Support Vector Machine to find out the outliers from the clusters. In proposed method we used voronoi indexing with K-mean Clustering. Centroid of the clusters are calculated with the help of voronoi indexing, because of this delay time for calculating outliers from the datasets are reduces. To optimize the outlier from the structured clusters support vector machine is used which improved the results.

REFERENCES

[1] Williams, G., Baxter, R., He, H., Hawkins, S., and Gu, L.2002. A Comparative Study for RNN for Outlier Detection in Data Mining. In Proceedings of the 2nd IEEE International Conference on Data Mining, Maebashi City, Japan, pp.709.

[2] Ajani S., Wanjari, M., An Efficient Approach For Clustering Uncertain Data Mining Based On Hash Indexing And Voronoi Clustering " at 5th International Conference on Computational Intelligence and Communication Networks (CICN), 2013, page No. 486 – 490,CD. 27-29 Sept. 2013.

[3] Cheng Zhang, Ming Gao, Aoying Zhou "Tracking high quality clusters over uncertain data streams" IEEE international conference on data engineering 2009.

[4] Ben Kao, Sau Dan Lee,Foris K. F Lee, "Clustering uncertain data using Voronoi diagrams and R-tree indexing" IEEE transaction on knowledge and Data Engineering Vol.22 No.9 September 2010.

[5] Wang kay Ngai, Ben Kao, Chun Kit Chui, Michael Chau, Reynold Cheng, Kevin Y. Yip "Efficient clustering of uncertain data" Sixth international conference on data mining. (ICDM 2006).

[6] P.Misra and P. Enge, Global Positioning System: Signals, Measurements, and Performance, 2nd ed. Ganga-Jamuna Press, 2006, iSBN 0-9709544-1-7.[2] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless micro sensor networks," in 33rd Annual Hawaii International Conference on System Sciences (HICSS), IEEE, Maui, Hawaii, U.S.A., 4th-7th Jan. 2000.

[7] H.-P. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in KDD. Chicago, Illinois, USA: ACM, 21–24 Aug. 2005, pp.672–677.

[8] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," in The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), San Francisco, U.S.A., 30th Mar.–3rd Apr. 2003.

[9] J. Chen and R. Cheng, "Efficient evaluation of imprecise location dependent queries," in ICDE. Istanbul, Turkey: IEEE, 15-20 Apr. 2007, pp. 586–595.

[10] O. Wolfson and H. Yin, "Accuracy and resource consumption in tracking and location prediction," in SSTD, ser. Lecture Notes in Computer Science, vol. 2750. Santorini Island, Greece: Springer, 24-27 Jul. 2003, pp. 325–343.

[11] Gerardo Miramontes-de Le´on, Arturo Moreno-B´aez, "Assessment in subsets of MNIST Handwritten Digits and their Effect in the Recognition Rate", Journal of Pattern Recognition Research 2 (2011) 244-252

[12] Wang G, Cao G, Porta L. Movement-assisted sensor deployment. In: Proceedings of twenty-third annual joint conference of the IEEE computer and communications societies (INFOCOM); 2004

[13] Soreanu P, Volkovich Z, Barzily Z. Energy-efficient predictive jamming holes detection protocol for wireless sensor networks. In: Proceedings of second international conference on sensor technologies and applications (SENSOR- COMM); 2008. p. 306–11.

