ISSN:0975-9646

Uncertain Numerical Data Clustering with VORONOI Diagram and R-Tree with Ensemble SVM

D. M. PAdulkar, V. Z. Attar
Computer Science, COE Pune

Abstract: - Uncertainty in the data is the major issue of the research now a days, the data gathered from the different sources like satellite, sensor networks and many more application were get the uncertainty in the data. The uncertainty may because of the precision of the device with which the readings are taken. Clustering or classification of the normal data with any algorithm like k-means is simple, but if the data is uncertain then it is quite difficult to cluster or classify the data. So here we study the voronoi diagram based approach which will solve the problem uncertain data with minimum number of expected distances while adding the object to the specific cluster.
Keywords: VORONOI Diagram, R-Tree, Ensemble SVM, MBR.

I. INTRODUCTION

Clustering is the technique widely used for different applications in the different areas. It is very much essential to study the clustering of uncertain data. That uncertainty may occur because of the inaccurate measurements, precision of the device or a stream of the data that is coming from the sources like remote sensor networks that may present the uncertainty. Here we are studying how to handle this type of the uncertainties of the data objects. The traditional clustering methods which will deal with the point valued data which is certain one. But they don't handle the uncertain data efficiently [1].

In this paper we are concentrating on the uncertain numerical data only. The uncertainty of the data shown in the following table1

Certain Data	***Uncertain Data***
Age of Ram is 23 years	Age of Ram is in between [20-30]
Temperature of Pune at morning was 120F	Temperature of Pune at morning was 120±10F
Salary of Ram is 6lacs/pa	Salary of Ram is 5-7lacas/pa
Can't play cricket today, because the temperature is 100 degree Celsius.	Can/cant play cricket because the temp. is 50-90 degree Celsius.
Annual income is 110	Annual income is [90-120]
Season is rainy	Season is[rainy, summer, winter]

Table 1. Uncertain data examples

The uncertain data represented in the above table can't be handled with the algorithms like k-means. It's very difficult to handle such a data efficiently and effectively. If it desires to handle it, it has to take any one of the value in the range or a part of the value specified. It may average the values specified in the range but that much effective solution we may not get it. The first algorithm for handling the uncertain data was proposed by [2], which is the extension of the traditional k-mean algorithm. The main disadvantage of this algorithm is to require large amount of the Expected Distance calculations (EDs)[2]. As there are n number of the objects that is to be clustered in to different clusters and initial centroid as k number so it requires the n*k number of the steps for each iteration of the ED calculations. The ED of each object with all such a centroids present over there is calculated so the total time required to compute the desired clusters for each data object is very large one. So in this paper we try to minimize the number of ED calculations for the data objects so that up to some extend we can reduce the time complexity of the algorithm, which will cluster the uncertain data objects.

II. RELATED WORK

The uncertainty of the data firstly represented with the help of the algorithm in [2] the major challenging was to handle the uncertainty of the data. To handle uncertainty of the data the PDF has important role
every were. The most traditional clustering methods aims to find a unique factor which will form a cluster for the same object with existing or with totally new cluster by minimizing the sum of squared error(SSE)[2] in case of the k-mean algorithm. The uncertainty in the data mining field, the tuples value is not only the important factor but also the presence of the uncertainty any where is the major concern of the field. The data having less uncertainty is more important than having the more uncertainty. There are two different types of the uncertainties, existential uncertainty and value uncertainty. The existential uncertainty occurs when it is uncertain whether an object or a data tuple exists there with uncertainty. The tuples in the relational database could be associated with the probabilistic data; the probability represents the existence of the object there in database [1]. The value uncertainty is the one, where the data is in existence but its value is not known precisely [1]. A data item with the value uncertainty is usually represented by the PDF functions over the bounded finite field of the possible values. Getting the valuable results from the uncertain data, till there are different attempts made for the same in terms of the UK-Mean algorithms [3]. The UK-Mean is the first attempt to the data uncertainty. This is the extension of the existing K-Mean algorithm. For fuzzy data clustering there was another attempt called as CK-Mean. As we do the theoretical but algorithmic analysis of this algorithms we will find that, if there are ***n*** number of the objects that is to be clustered over the ***k*** number of the clusters, then for UK-Mean algorithms requires the ***n*k*** number of the computations per iterations so

that the computational time is very large for all such Expected Distance(ED) calculation. So that in terms of the time it is expensive one. If we have some sort of another pruning techniques which will reduce the number of computations for the UK-Mean algorithm. Clustering of uncertain data is also studied in the fuzzy data clustering. Because in the fuzzy data, the different instances are present for the same data tuples but with probabilistic values. The degree of the belongingness of the object is important in case of the fuzzy data clustering. The belongingness of the object depends upon the subset belongingness, to which subset it belongs. The fuzzy clustering concentrate on the hard clustering of the objects means it belong to only one of the cluster leader in it[1]. The voronoi diagram is the well known geometric structure which can be applied easily for the clustering problems as well. Here we are proposing the model with voronoi diagram which will give one of the centroid for the clusters. Once we know the centroid of a cluster then to minimize the distance of the data objects to its belonging cluster centroid. In this paper we concentrate on the uncertain numerical attributes only, which gives the uncertain data.

III. UNCERTAINTY MODEL

In this section we will concentrate on the uncertain model for numerical data only. Which is the most common uncertain attribute type occurring in the data mining field. When the value of the numerical attribute is uncertain then attribute is called as uncertain numerical attribute (UNA)[4]. These uncertain numerical values represent the range of the values specified in the above table, table1. By applying the Probability Distribution Function (PDF) over the range specified in the tuples. Continuous probability Distribution function is applied on the range of the data items. It generates the probability of the single attribute specified in the dataset. To handle the uncertainty of the range valued numerical data, we apply the continuous PDF on it so that it gives the probability of occurring the tuple in the data set. The goal of clustering is to group the same data items into the single cluster. So that the expected distance with cluster representative will be minimized. The expected distance for an object Oi with the cluster representative is represented as

$$E\left(\sum_{j=1}^{k}\sum_{i\in C_j}\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2\right)=\sum_{j=1}^{k}\sum_{i\in C_j}\int\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2 f(\mathbf{x}_i)d\mathbf{x}_i$$

This is computing the expected distance of the object to be clustered with respect to the pdf value of the uncertain numerical attribute. Once the expected distance calculation are over the next task is to have the clustering of the object. The next section discusses the clustering process with voronoi diagram for uncertain numerical attributes and pruning techniques with voronoi and pdfs are associated with the nodes of the R- tree.

The uncertainty varies explicitly for many more applications. Traditional clustering algorithms like K-means require the definition of the distance metrics. For example the object x is added to the cluster based on the distance between x and the cluster representative Ci. The object that is to be added to the Cluster is one dimensional where the single dimensional uncertainty presents. But if the uncertainty is multidimensional at that time the object is not a single point in the space but it is represented in terms of the PDF. So the PDF will fix the uncertainty region, within that the there is most possibility to get the point inside it. So that PDF is most powerful way to represent tackles with the uncertain data. The algorithms like K -mean for certain data and its next version firstly presented by[2] called as UK- Mean – uncertain K mean algorithm, it follows all the things as same as that of the K- mean algorithm except the use of the Expected Distance calculations. The UK-Mean calculates the ***ED(Oi Ci)*** where ***Oi*** is the object to be clustered and Ci is the cluster representative. In UK-mean it tries to minimizing is the ED of the object, were it is minimized there it add the object as a part of the clustering. In UK-Mean algorithm the number of ED calculations is done. Our task is to reduce the number of ED calculations so that it reduces the time for total computations [3].

IV. CLUSTERING MODEL

The Figure shows the clustering process to be carried out. The block diagram explains about the clustering process how that is carried out. Uncertain numerical data is supplied to the PDF generator algorithm, which generates the PDFS for all the records those who falls in valued uncertainty.

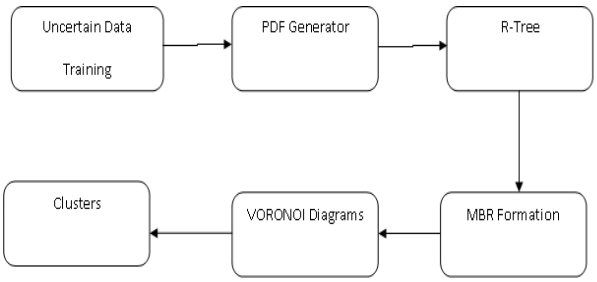


Figure1. Uncertain Data Clustering Process

As shown in the above figure the generated PDFs are bulk loaded into the R-Tree, which goes through the different iterations and fix the final root node for the tree. all the PDFs are stored into the leaf node of the tree and all other intermediate nodes including the root node represents some Bounding box called as Minimum Bounding Rectangle (MBR), which holds more than one records from the given dataset. As the process of MBR generation finishes our next starts called as pruning process with the help of the VORONOI diagram. This process fixes the centroid from the given dataset only and passes over reading the MBRs. If that MBR completely lies inside one the cell then there is no need to find out EDs for all the data items present into the same MBR that is the biggest advantage of using the VORONOI diagrams for clustering process. So that when we compare the UK- mean algorithm do the clustering of uncertain data. It passes every data records through the K number of iterations

as K clusters are assumed to be generated there. So the total time requirement of the UK-Means theoretically come to O(kXn), where the n- number of records and K- number of clusters want to be generated. As explained previous the MBR represents the more than one uncertain data record so that, number of EDs calculated will less as compared to the previous algorithm, this is the major reason of choosing VORONOI

Probably we are generating ten clusters because we want to use ten SVMs for classification purpose. So next step start called as classification process.

V. CLASSIFICATION MODEL

The Figure2 explains the process of classification of uncertain numerical data. The attempts made to classify the uncertain data, classify it but the approach we have used cluster the data items first and then classify. we proved that the supplying the data items as it is to the classification model, and do the clustering and supply the clusters as input to the classification system it increases the accuracy of the model. That is represented in the next section result and analysis.

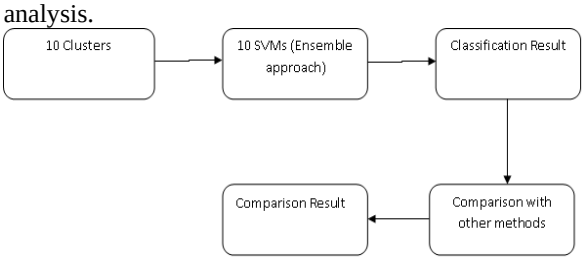


Figure 2. Uncertain Data Classification Process.

VI. VORONOI DIAGRAMS FOR UNCERTAIN NUMERICAL DATA.

suppose that in the D- dimensional region R we have set of P uncertain data points, we have set C of centers and distance function d(Pi,Cj) A Voronoi diagram V(P, C, d(*, *)) assigns each data point pi to the center ci with smallest distance d(pi , cj).This induces a partition of the data points into clusters. For the continuous case, it can be more natural to refer to regions rather than clusters. in both cases, we mean the subsets of the data points that are closest to a particular center. Assuming we are using the distance function, All regions are convex (no indentations). All regions are polygonal. Each region is the intersection of half planes; the infinite regions have centers on the convex hull. In some ways, a Voronoi diagram is just a picture; in other ways, it is a data structure which stores geometric information about boundaries, edges, nearest neighbors, and so on. How would you compute and store this information in a computer?

It's almost a bunch of polygons, but of different shapes and sizes and orders - and some polygons are “infinite". If you need to compute the exact locations of vertices’ and edges, then this is a hard computation, and you need someone else's software. The voronoi diagram seems to exhibit the ordering imposed by a set of center points that attract their nearest points. However, since the placement of the centers is arbitrary, the overall picture can vary a lot. The shape and size of the regions in particular seem somewhat random. Is there a way that we can impose more order? Yes! if we are willing to allow the centers to move. The voronoi diagram is really a “snapshot" of the situation at the beginning of one step of the continuous K-Means algorithm, when we have updated the clusters. As mentioned in the previously, we can take a given set of centers with a “disorderly" Voronoi diagram, and repeatedly replace the centers by centroid. This will gradually produce a more orderly pattern in which the clusters are roughly the same size and shape, and the centers are truly at the center. The key to this approach is to be able to compute or estimate the centroid of a set of points that are defined implicitly by the nearness relation. This algorithm needs the centroid (xi ; yi) of each cluster Ci . If Ci is a polygon, we use geometry to find the centroid. If Ci is more complicated, we must use calculus.

$$\bar{x}_i = \frac{\int_{C_i} x\,dx\,dy}{\int_{C_i} dx\,dy}, \quad \bar{y}_i = \frac{\int_{C_i} y\,dx\,dy}{\int_{C_i} dx\,dy},$$

But exact calculations are limited to simple, small, regular problems! We can estimate the centroid using sampling. Generate a large number of sample point’s p in the entire region. The centroid of cluster Ci is approximately the average of the sample point’s p that belong to Ci.

$$\bar{x}_i = \frac{\sum_{p \in C_i} x_p}{\sum_{p \in C_i} 1}, \quad \bar{y}_i = \frac{\sum_{p \in C_i} y_p}{\sum_{p \in C_i} 1}$$

By using density functions, we can vary the size of the regions. Two regions will not have the same area, but rather the same “weight", defined as the integral of the density. This is like dividing a country into provinces of equal population. Using this idea, for instance, you can generate a mesh for a computational region, and force the mesh to be fine near the boundaries, and very fine near corners or transition zones. The voronoi diagrams are the most powerful mathematical computational tool which will cluster the uncertain data. But in this paper we are concentrating on the voronoi diagram based pruning that is to applied on the UK-mean algorithm[1][3][6], where the uncertain data items are hold by the R- tree. All the leaf nodes of the R- tree represents the probability density function values for each records present into the clustering process of uncertain data. R- Tree leaf nodes represents the pair of keys are pdf's required for clustering process.

VII. R-TREE INDEXING

An R-tree [23] is a self-balancing tree structure that resembles a B+-tree, except that it is devised for indexing multi-dimensional data points to facilitate proximity-based

searching, such as k-nearest neighbour (kNN) queries. R-trees are well studied and widely used in real applications. They are also available in many RDBMS products such as SQLite, MySQL and Oracle. An R-tree conceptually groups the underlying points hierarchically and records the minimum bounding rectangle (MBR) of each group to facilitate answering of spatial queries. While most existing works on R-tree concentrate on optimising the tree for answering spatial queries, we use R-trees in the thesis in a quite innovative way: We exploit the hierarchical grouping of the objects organised by an R-tree to help us check pruning criteria in batch, thereby avoiding redundant checking.

VIII. Ensemble SVM for Classification

Once the ten clusters are generated from uncertain data supplied, are supplied to the SVM with ensemble approach. This SVM are with linear regression kernel are used.

IX. Experimental and Results.

The series of experiments are carried out with the synthetic and real dataset.

A. Dataset used: we have used a synthetic dataset generated by WEKA for experimentation and added the uncertainty in the data. This synthetic dataset with four attributes we have used and all the four attributes are uncertain. With the same setup used real dataset form KEEL miner low quality dataset. All the results represented here are based on this dataset with 10000 data records with four attributes and all are uncertain. The results are shown as bellow.

B. Results of classification uncertain data.

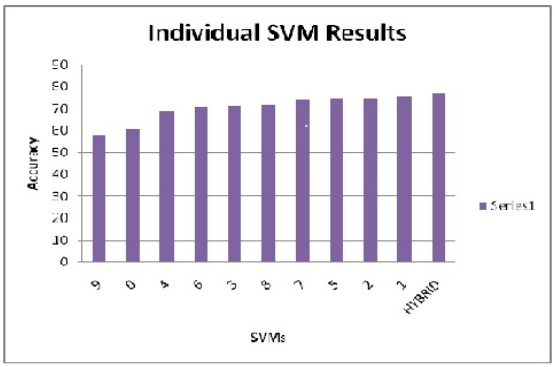


Figure 3. Individual SVM and hybrid result.

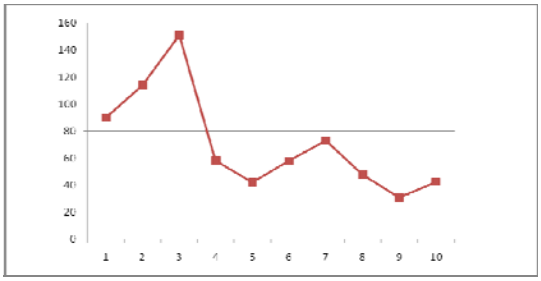


Figure4: avg. cluster distance vs. accuracy

As shown in the above graphs the accuracy of individual SVMs and Hybrid or ensemble SVM. The individual accuracy is less than the ensemble accuracy of classification. The second graph shows that as the average cluster distance is minimum then cluster accuracy is genrally greater. This is true for almost all the clusters.

Conclusion:-

The concept of voronoi diagram and R-tree used here helps to reduce the number of mathematical computations required for the item assignment to the particular cluster. The second step of the model is to classify the clustered items, here the it shows that the result of classification without clustering is lesser accurate than the result of classification after supplying clustered items to the classification model.

References

[1] 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.

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

[3] 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 minnig. (ICDM 2006)

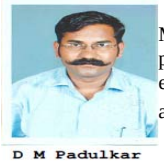
[4] Biao Qin, Yuni Xia, Sunil Prabhakar, Yicheng Tu "A rule based classification algorithms for uncertain data" IEEE international conference on data engineering 2009.

[5] Osamu takata, Sadaki miyamoto, Kazutaka Umayahara "Fuzzy clustering of data with uncertainties using minimum and maximum distance based on l1 metrics" IEEE international conference 2001.

[6] Graham Cormode, Andrew McGregor "Approximation algorithms for clustering uncertain data" PODS June 2008 ACM conference

Authors Profile.

Prof. Vahida Attar is working as Assistant Professor, Department of Computer Engineering and Information Technology, College of Engineering, Pune, Maharashtra, India. She has 18 years of experience in teaching and research. Her research areas include Stream Data Mining, Application of Data Mining for Healthcare and Agriculture. She has about 25 publications at international journals/conferences. Currently she is working on different funded projects from AICTE and BARC, Mumbai. She is member IEEE. life-member ISTE, CSI and IE.



Mr. D. M. Padulkar currently perusing his post graduate program in computer engineering for college of engineering pune. His research interest is data mining, algorithms. He is life member of ISTE,CSI.