Indian Journal of Science and Technology, Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/71972, December 2015 ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Approach to Perform Uncertainity Analysis on a Spatial Dataset using Clustering and Distance based Outlier Detection Technique

K. R. Manjula[1*], Amrita Kumari Keshari[2] and Atul Pahlazani[2]

[1]School of Computing, SAP, CSE, Sastra University, Thanjavur - 613401, Tamil Nadu, India; manjula@cse.sastra.edu
[2]B. Tech, CSE, Sastra University, Thanjavur - 613401, Tamil Nadu, India;
amrita1222.kumari@gmail.com, atul.pahlazani@gmail.com

Abstract

Background: In past years, many methods have been implemented for maintaining and supervising uncertain data that may occur due to collection of data in new ways which results in missing values, erroneous data. The main aim of this work is to help the end user to get correct information about spatial data. **Method**: The behaviour of data as an outlier is the result of uncertainty. The challenge in spatial data sets is to cluster uncertain objects. Hence, unsupervised clustering can be used to deal with this type of data. In this paper, the difficulty of outlier detection with uncertain data is examined. **Finding**: To improve the performance and quality, Voronoi Diagram is used which partition the objects into each cell and helps to see the exact location of an object. The integral part is the pre-processing step of removing uncertainty to avoid wrong interpretation. Furthermore, CLARA (Clustering LARge Applications) algorithm is applied to produce the high quality clusters. It has an in-built function of outlier detection too and it is suitable for large data set. This algorithm uses Mahalanobis Distance to calculate the distance between cluster and its members, to remove outliers and reduce uncertainty for feasible and supporting inputs. This procedure can be a valid provision to be use in geo-database creation. **Improvement**: The methodology can be enhanced by designing the procedure to develop a Decision Support System (DSS) for spatial database creation.

Keywords: CLARA Algorithm, Clustering, Mahalanobis Distance, Spatial Uncertainty, Varonooi Polygon

1. Introduction

The spatial applications like locating health centres, hotspot areas of natural calamities, finding roots, and travel impedes visualization etc., are need to maintain databases to store information. This information includes digital images, satellite images, statistical data and topology information. All these databases constitute some sort of spatial components that have a geographic impact. This impact can affect the quality of information if not mentioned correctly. And these spatial components are spatial objects made up of points, polygons, and lines and viewed in the form of healthcare centres, countries, roads, rivers, etc. Geo - spatial attributes can be used to model uncertainties of its geographic objects using data mining. It analyses large amount of spatial data to find patterns with respect to location[10]. It also identifies position or motion of objects in a coordinate system with respect to geographical information. Accordingly, the research based on uncertainty in spatial data is important. The uncertainty in spatial data is a major element for determining the quality of spatial data. The quality includes accuracy, consistency, correctness and completeness that can be handled using appropriate clustering algorithms.

Uncertain can be divided into two categories such as existential uncertainty and value uncertainty. Existential uncertainty occurs when it is uncertain that the object may exist or not. Whereas value uncertainty occurs when an object exists but its values are not exactly known. Our main aim is to detect outliers based on value uncertainty

*Author for correspondence

(for example latitude and longitude of locations). Data uncertainty is considered during clustering process so that clustering results are improved. We focus on CLARA clustering as it is considered for large data set. The resulted clusters are introduced to Mahalanobis distance which further enhance the clustering quality and detects more number of outliers. We have illustrated related work on four aspects:

- Voronoi diagram of uncertain data.
- Clustering uncertain data.
- Algorithm to extract more outliers.
- Removal of outliers to reduce uncertainty.

1.1 Voronoi Diagram of Uncertain Data

Based on[11], a plane X with 'n' distinct points is considered. The Voronoi diagram divides the plane X into 'n' cells, called Voronoi cells. These cells are mainly bounded by line segments and each cell contains exactly one data point. For example, if these points represent the locations of all the stationary shops in the city, the Voronoi diagram divides the city into cells (locations) based on each stationary shop. For each person residing in exactly one cell, the shop represents the closest place to buy stationary. We extend this method for clustering uncertain spatial data i.e., for a particular location with two or more latitude and longitude the Voronoi diagram will have collinear points as shown in Figure 1.

Voronoi diagram can be combined with clustering techniques. In[8], it is used as a method to enhance the spatial relationships among clusters to improve the coherence of pruning algorithm. In that study, Voronoi cell pruning technique has been defined based on Voronoi diagram which discusses whether the Minimum Bounding Box (MBR) for each point lies completely inside any Voronoi cell V (C_i). If so, then that point is assigned to that cluster C_i and all other clusters are pruned. This pruning algorithm is used to delete the redundant points. Figure 2 and Figure 3 shows the assignment of data points and the assignment of each point to a polygon and after pruning.

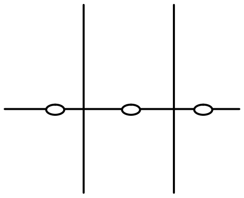

Figure 1. Voronoi diagram for non-collinear points.

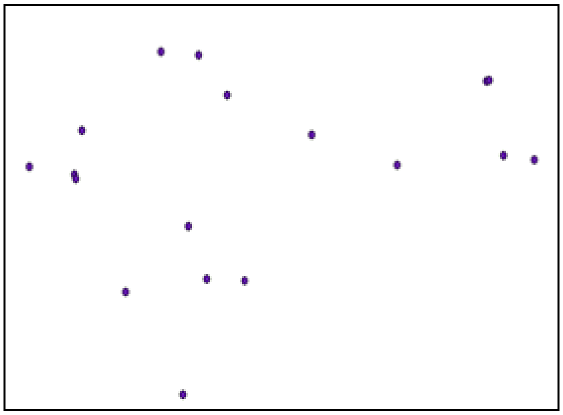

Figure 2. Input data points.

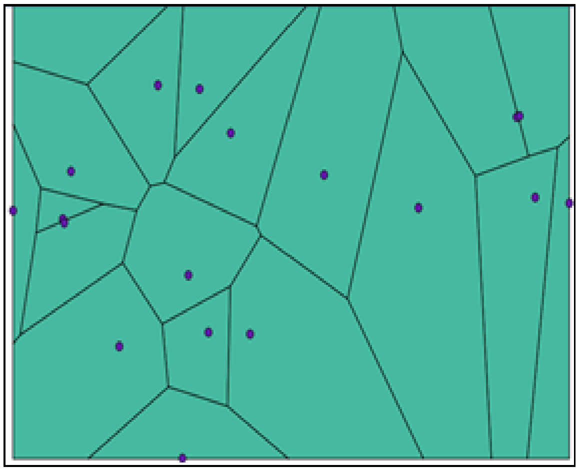

Figure 3. Output Voronoi cells.

In this paper, considering the latter approach as base, we introduced another clustering method using Voronoi diagram which is applicable for large spatial uncertain dataset[2]. Discuss a skeleton method based on Voronoi diagram for boundary points using Euclidean metrics. Skeletons represent the description of the structure of an object. These skeletons[13] provide a distinct and dense representation of the image.

1.2 Clustering Uncertain Data

Most of the clustering algorithms[14] optimize only clusters and improve their quality. CLARA, CLARANS, BIRCH and DBSCAN are exception handlers. Bin Jiang et al.[7] takes probability distributions of data into account and proposed the use of KL divergence to measure the similarities between uncertain data and combine it with clustering methods to cluster uncertain data. Both the continuous and discrete cases of uncertain data are considered[4] showed clustering technique for large numeric data sets using polygon approach by creating the polygon containing multiple triangles and computing density of each triangle and found the farthest points from the merged triangle. This procedure increased the quality of multi-density clusters but it is only limited to two-dimensional objects. In[2], the author proposed UCK-means algorithm[15] as an extension of K-means for clustering uncertain data and minimizes the sum of distance error. Hierarchical clustering does not support large datasets but K-means can extend some support to large datasets[12]. In[2], the author combined the traditional K-means algorithm with hash indexing and Voronoi diagram to improve the efficiency so that more refined and accurate results are obtained. Such methods are not able to handle uncertain data based on geometric information. All clustering algorithms aim at minimizing the distance between points in the same cluster and maximizing the distance between points in different clusters. Clustering uncertain multidimensional data set becomes costly.

1.2.1 Algorithm to Extract Outliers

The presence of multivariate outliers decreases the quality of clustering. As the dimensionality increases, identification of outliers becomes more difficult. Most of the existing techniques only deal with deterministic data, which cannot be applied on uncertain data. Density based technique considers dense region of objects and determine outliers which are far away from these regions and have less density regions proposed in[1]. The large data set is divided into micro-clusters to identify the outliers hence reduces the time consumption. Markus M. Breunig[6] introduced LOF which captures local outliers of uncertain data. Recently, only few studies explored outlier detection of spatial data, whereas, spatial data plays a major role in day-to-day life. The technique is used to extract valuable information. In[5] spatial - temporal data is discussed using ST-DBSCAN, an extension of DBSCAN and in[10] spatiotemporal association rules with spatial data is explored where scope of uncertainty is higher. He Mingke et al.[12] proposed mutually exclusive relation between uncertain data using distance based approach. Mutually exclusive relation means "if an object is present at a particular location at a particular time then that object cannot be present at any other location at that time". Ke Zhang et al.,[8] proposed a new distance based technique for scattered real world data which determines the deviation of objects from its neighbours based on location of objects.

1.3 Removal of Outliers to Reduce Uncertainty

Quality of spatial data plays a major role in data mining. Removal of outliers makes easier to evaluate data. The techniques of removing outliers are still unexplored. Atanassov R et al.,[3] discussed about outlier removal[16].

2. System Study

The existing system deals with finding uncertainties in data set considering mutually exclusive relations using distance based approach (Euclidean distance) to remove the outliers. The probability that instances of Y happen in (N(o)) is:

$$P(Y) = \sum_{(j=1)}^{n} \equiv \left[p(Y = y)i \right]$$

i.e., due to mutually exclusive relations if one event happens, the other event cannot take place.

In the existing system, an uncertain data set has been taken and instances of uncertain objects are considered independent of each other.

$$P_{j,S} = P_{j-1,S\backslash Y}\, P(Y) + P_{j\ S\backslash Y}(1\text{-}P(Y))$$

where $P_{j,S}$ denotes the probability that exactly j objects are present in uncertain dataset S which is visited while detecting t is an outlier and Y€S. The distance from an object to all its neighbours is calculated for each point in the data set. This distance decides the outliers and supports the process of uncertainty analysis.

For an observation to be valid, it must be within the range of threshold value taken according to the data set. The threshold value decides the accuracy of outlier detection. The observations outside the threshold value are marked as outliers given by: The probability that at least k objects appear in the neighbourhood of outlier o (N(o)) and it is computed by:

$$P_s(k) = 1 - P_{0,s} - P_{i-1,S,}$$

Where, $P_s(k) \geq \lambda$. If $P_S(k) < \lambda$, then o is considered to be an outlier.

Using Euclidean distance, this approach is limited to work on only limited datasets. These datasets take circular decision boundaries which do not suit for large spatial data. It is seen that if object (k) and threshold value (λ) is increased, the outliers and runtime are also increased. It is also observed that by increasing the distance (ud), the outliers are decreased but the runtime is increased. It is concentrated in removing only the outliers. Performance of the existing system can be improved further by considering large complex data sets and making corresponding modifications in the algorithm.

2.1 Proposed System

The method proposed can be used for wide collection of data. Instead of taking the mean value of the data points, it uses the most centrally located data point in a cluster as medoids. Unlike k-means, medoids move and adjust itself to improve the distance of the clusters. The process is repeated until best medoid is found. It efficiently detects noisy data and outliers.

2.1.1 Data Collection and Pre-processing

The spatial data set pertaining to districts in India is collected from the authorized source repository[9]. It contains the locations along with the latitude and longitude values given in degrees. To work upon this data with certain algorithms the values are converted from degrees to decimal values using the following formula:

$$\text{Decimal} = \text{Degrees} + (\text{Min}/60) + (\text{Sec}/3600)$$

2.1.2 Limitations of the Work

The distance based approach for uncertainty analysis can work with only numeric values. Use of huge data sets may result into possibility of getting overlapped clusters due to which accurate results cannot be generated. Moreover, due to distance computation over a non-uniform data may result in removal of some important objects too.

3. Methodology

3.1 Voronoi Diagram

The Voronoi diagram shows spatial relationship among the objects. It is used to know the exact location of objects. The properties and definition of Voronoi diagram is as stated. Given a set of points, Voronoi diagram divides the plane into cells. Each cell consists of an object closer to only one particular point. Formally, consider a dataset

$$P = \{o1, o2, o3, o4..., ok\}$$

as a set of k points in d-dimensional plane. The variable d(x, y) is denoted as the distance between two points x and y in the plane. Hence, Voronoi diagram of dataset P is the partition of plane into k cells. An object x falls into the cell with respect to another point y if and only if

$$d(x, y_m) < d(x, y_n), \forall m \neq n \text{ and } y_m, y_n \text{ € } P$$

Each cell consists of a point perpendicular bisector to other, given as $y_m|y_n$. The bisector is the perpendicular to the segment that joins y_m and y_n. The space is divided into two portions A and B. The portion containing y_m is denoted as $P_{m/n}$ and that containing y_n as $P_{n/m}$. This is how the Voronoi diagram is used. It can be explained by following pseudo algorithm:

- In a n-dimensional plane, a set of objects in data set $P = \{x_1, x_2, x_3, \ldots, x_m\}$ are considered with $d(a, b) \geq 0$ such that a, b € P. The steps in the algorithm as follows:

```
1. If (P==x_i) where i =1 \\ if the data set contains only one object
   1.1 Return
2. Else
   2.1 For all xi € P do
```

- Construct Voronoi diagram for A and B
- Prune the points lying in the opposite side of A and B respectively.

3.2 Clara

CLARA algorithm is used because of its efficiency in clustering large datasets and is reliable in detecting outliers. It takes a sample of objects from the data in account on which PAM algorithm is applied to get a suitable set of medoids for cluster formation. PAM is an iterative process that repeats until it gets the most suitable medoids (with least mean dissimilarity). Clusters obtained for the sample data are used to classify the rest of the data. Following algorithm explores the working of CLARA:

Input: Dataset P of d-dimension, number of clusters K, randomly selected number of samples S.

Output: Medoids M

Algorithm:

```
1. For i = 1 to S repeat
   • randomly draw a sample of size 40+2K from the dataset P;
2. Call PAM algorithm over this sample S to find medoids;
3. For each mi € M do
   • randomly assign each data point in dataset P to its nearest medoid
   • calculate the dissimilarity between data points and its medoid
```

$$Cost(M,P)=\sum\nolimits_{i=1}^{n} \equiv \left[dissimilarity\ (o_i)\ rep(m_i,o_j) \div \mathrm{n} \right]$$

- where -M is the set of medoids
- n is the total number of objects (data points) in dataset P
- rep (mi, Oi) gives the medoid closest to the object Oi

4. repeat the process until best medoid is not found.

3.3 Mahalanobis

It is the measure of distance between an object and a distribution of objects. The mean value or centre point of the distribution is taken to calculate the distance. The following steps illustrates the working of Mahalanobis distance:

Input: Generated clustered data C = {c_1,c_2,…,c_n}

Output: Computed distance

Algorithm:

1. Calculate the mean of the data

$$Mean(m)=\frac{1}{n}\sum_{i=1}^{n} p_i$$

2. Calculate covariance matrix of the variables

$$S^{-1}=\frac{1}{n}\sum_{i=1}^{n}(p_i-m)^2$$

3. Calculate the Mahalanobis distance using:

Where,

$$D^2=(x_i-m)^T\ S^{-1}(x_i-m)$$

x_i is each data point in data set P

m is the mean of data points

S^{-1} is the covariance matrix

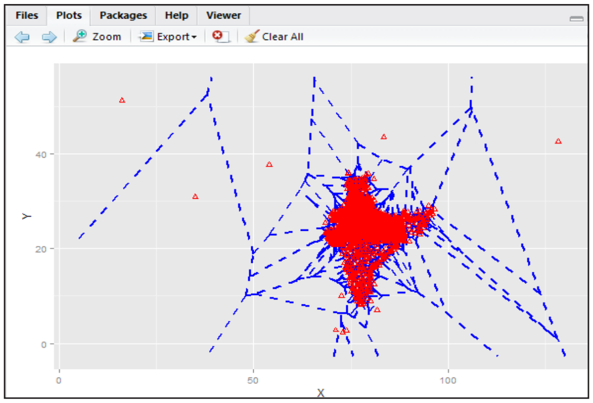

Figure 4. Voronoi polygon of each object.

3.4 Implementation

It deals with the spatial data set of districts in India with their latitude and longitude values in degrees. The data set is found to have uncertainties such as locations having multiple latitude and longitude values. This data set is subjected to produce Voronoi polygon. It divides each point from data set into each cell and shows the map of India as shown in Figure 4. The polygon constructed is subjected to CLARA algorithm and clusters are obtained as depicted in Figure 5. To further enhance the quality of clustering Mahalanobis distance is computed. For an object to be an outlier it should be less than the threshold value calculated from the distance. Figure 6 shows the objects plotted after the removal of outliers on the dynamic map.

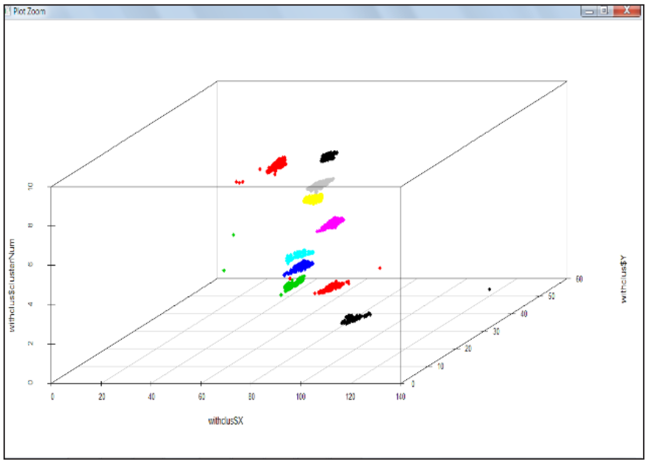

Figure 5. Clustered objects.

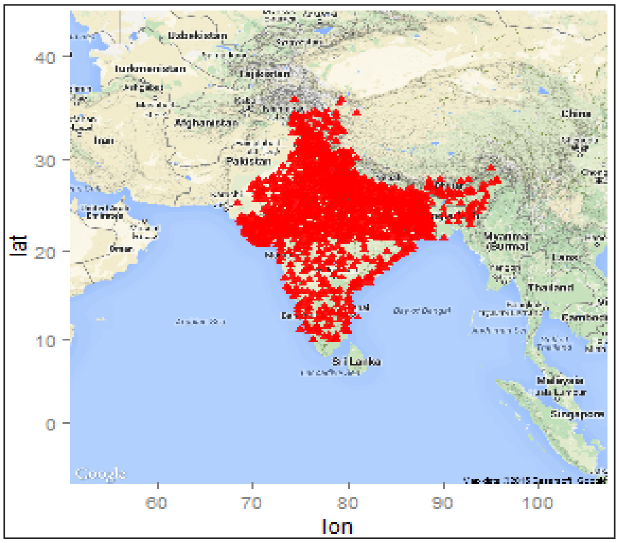

Figure 6. Data set without outliers.

4. Results and Discussion

Two methods can be used to measure the quality of the clusters: Extrinsic and intrinsic. Extrinsic method includes comparing the cluster with respect to the ground truth value. These ground truth value varies according to the type of data. If spatial data is considered, the ground truth value indicates the exact location of the point in coordinates on earth. Intrinsic method is an unsupervised method used when the ground truth value is not available. In that case, the quality of clusters is evaluated by taking into consideration how fine the clusters are separated. Here, the unsupervised method is used. The comparison is made between K-means and CLARA. Figure 7 and Figure 8 shows the number of objects in each cluster using two different clustering methods. The dataset contains some outliers due to which the clustering is not proper. It can be observed that using K-means, the number of outliers obtained is more than using CLARA in Figure 9. Hence, the proposed algorithm that uses CLARA algorithm with Mahalanobis distance is found to be more efficient in finding the outliers. Furthermore, if performance is taken into consideration, the running time of the proposed algorithm increases with the increase in number of clusters as compared to k-means depicted in Figure 10. In Figure 11, it can be observed that by increasing the distance, the number of outliers is decreased.

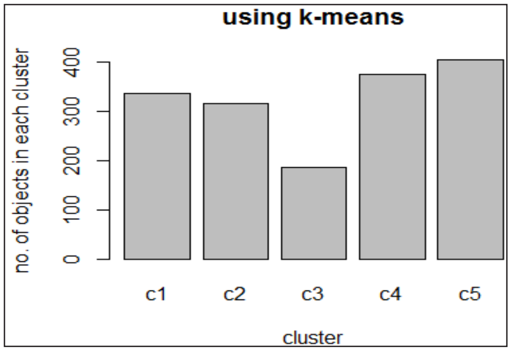

Figure 7. Objects in each cluster using K-means.

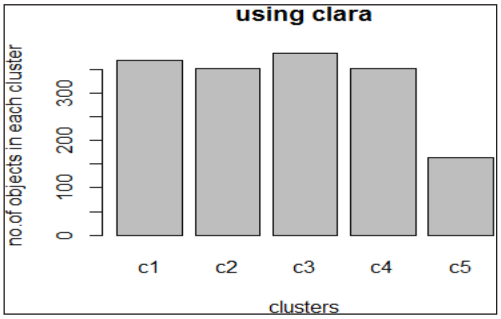

Figure 8. Objects in each cluster using CLARA.

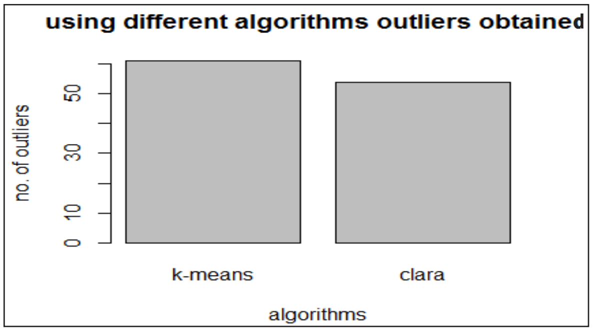

Figure 9. Number of outliers.

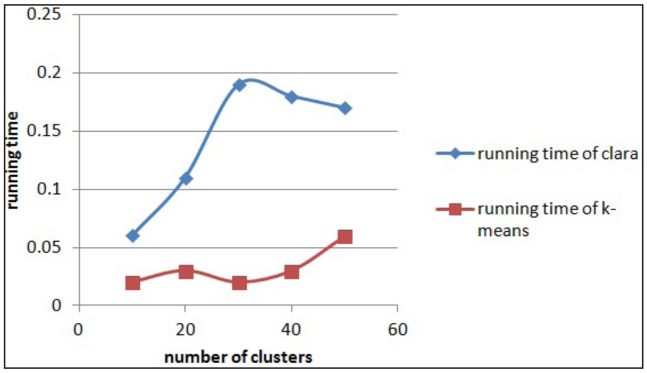

Figure 10. Comparison of running time.

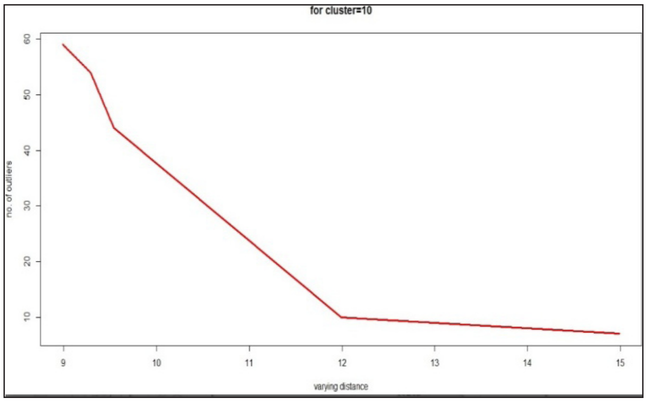

Figure 11. Variation in distance.

5. Conclusion

This paper has proposed a robust method aiding uncertainty analysis with a unique combination of Voronoi, CLARA and Mahalanobis Distance for outliers.

The first reason for combination is to find the clusters on spatial data. The second reason is being able to find the uncertainties and outliers when the clusters are obtained. Hence, this is the pre-processing step in the removal of uncertainty in spatial data set. This work used Voronoi polygon to detect the overlapping of data points.

Dividing each point into cell makes it easy to identify locations precisely. The experimental analysis shows that this approach can be applied to large spatial data sets. The work can be extended to cluster large multivariate spatial data sets to further enhance the quality of spatial data. The time complexity of the algorithm can be decreased when more number of clusters is considered.

6. References

1. Aggarwal CC, Yu PS. Outlier detection with uncertain Data. SDM, SIAM. 2008; p. 483–93.
2. Ajani S, Wanjari M. An Efficient approach for clustering uncertain data mining based on hash indexing and voronoi clustering. IEEE 5th International Conference on Computational Intelligence and Communication Networks; Mathura. 2013. p. 486–90.
3. Atanassov R, Bose P, Couture M, Maheshwari A, Morin P, Paquette M, Smid M, Wuhrer S. Algorithms for optimal outlier removal. Journal of Discrete Algorithms. 2009; 7(2):239–48.
4. Barun HB, Das DK, Sarmah SJ. A density based clustering technique for large spatial data using polygon approach. IOSR Journal of computer Engineering. 2012; 3(6):1–9.
5. Birant D, Kut A, ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data and Knowledge Engineering, Science Direct. 2007; 60(1):208–21.
6. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. ACM. 2000; 29(2):93–104.
7. Jiang B, Pei J, Tao Y, Lin X. Clustering uncertain data based on probability distribution similarity. IEEE Transactions on Knowledge and Data Engineering, 2013 Apr; 25(4):751–63.
8. Kao B, Lee SD, Cheung DW, Ho WS, Chan KF. Clustering uncertain data using voronoi diagrams. 8th IEEE International Conference on Data Mining; Pisa. 2008. p. 333–42.
9. Manjula KR, Jyothi S, Varma SAK, Varma SVK. Construction of spatial dataset from remote sensing using GIS for deforestation study. International Journal of Computer Applications. 2011 Oct; 31(10):26–32.
10. Manjula KR, Jyothi S, Varma AKS. Mining multilevel spatiotemporal association rules for analyzing the factors of deforestation. 2013; Available from: http://citeseerx.ist.psu.edu/viewdoc/download Doi: 10.1.1.259.2794. 2013.
11. Mayya N, Rajan VT. Voronoi diagrams of polygon: A framework for shape representation. IEEE IBM Research Report. 1996; 6(4):355–78.
12. Mingke H, Zheyuan D, Ni W. Distance based outlier detection on uncertain data of mutually exclusive relation. IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC); KunMing. 2013. p.1–5.
13. Ogniewicz R, IIg M. Voronoi skeletons: Theory and applications. Proc CVPR92; Champaign, IL. 1992. p. 63–9.
14. Radha M, Devi D, Thambidurai P. Similarity measurement in recent biased time series databases using different clustering methods. Indian Journal of Science and Technology. 2014; 7(2):189–98.
15. Yu P, Qinghua L, Xiyuan P. UCK-means :A customized K-means for clustering uncertain measurement data. 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD); Shanghai. 2011. p. 1196–200.
16. Zhang K, Hutter M, Jin H. A new local distance-based outlier detection approach for scattered real-world data. Berlin Heidelberg: Springer- Verlag. 2009. p. 813–22.