International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

AN APPROACH TO DETECT OUTLIERS IN OPENSTREETMAP DATA

Kriti Bhatia[1], Sukhjit Singh Sehra[2], Geetika[3], Sherry Chalotra[4]

[1] *M. Tech Scholar, Department of CSE, GNDEC, Ludhiana, Punjab (India)*

[2,3] *Assistant Professor, Department of CSE, GNDEC, Ludhiana, Punjab (India)*

[4] *Assistant Professor, Department of IT, PCBT, Phagwara, Punjab (India)*

ABSTRACT

The volume of spatial data is increasing day by day and is greatly challenging the ability to extract useful and implicit knowledge. The spatial outlier detection, an important branch of spatial data mining, aims to discover those objects whose non-spatial attribute values exhibits different behavior to a great extent. The neighboring data has a great influence on the spatial data objects. The identification of spatial outlier detection helps in discovering the informative knowledge which is quite useful for many researchers and decision making organizations. This paper proposed a new hybridized algorithm which combines both clustering and optimization technique. k-means clustering was used to cluster the whole dataset. After clustering, the ACO technique optimized and updated the cluster in order to get clusters more efficiently. The clusters formed were analyzed in a way that the data objects which were at the farthest position compared to the other data objects in each cluster were considered as outlier data objects. An experimental result showed that the proposed method was helpful in extracting outliers in a better way as compared to the existing k-means technique.

Keywords: ACO, Cluster, k-means clustering, Spatial data, Spatial outlier detection

I. INTRODUCTION

An outlier refers to that data object which exhibits different functionality or characteristics to a large extent from the other data objects present in a large data set [1]. An outlier either exists as a single isolated data object or forms a collective group of adjacent data objects together which forms a region. The correct identification and detection of outliers to have hidden and valuable knowledge about the data is one of the main concerns for many researchers and organizations [2]. The outlier detection problem is classified into two main categories: Traditional outlier detection approach and spatial outlier detection approach [3].

An OpenStreetMap data is a type of spatial geographic data [4]. Information is represented in the form of geometry and topology and the data can be mapped. Spatial data are often accessed, analyzed, or manipulated through Geographic Information Systems (GIS) [5]. A spatial outlier is that spatially referenced object whose non-spatial attribute values exhibits different behavior or characteristics of other spatial objects in its spatial neighborhood [6]. The identification of spatial outlier detection helps in discovering the informative knowledge which is quite useful for many researchers and decision making organizations. It can be used in many real life day to day applications like the detection of less developed region, in detecting the most literate district of any

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

state, issues related to public health, traffic related problem, any environmental issues, any criminal activities, location-based services, severe meteorological events [7]. Several clustering algorithms are used in the detection of outliers. The basic idea of clustering is to partition the dataset into different groups of data [8]. The data objects within each cluster shows similar properties and are different from other clusters. The most commonly used clustering algorithms are k-means [9] [10], PAM [11], CLARA, CLARANS [12] [13] [14] [15], DBSCAN [16] [17] [18]. K-means is a simple and easy to understand method which partitions the data objects into k clusters, where the value of k is specified by the user, such that it satisfies the minimum distance criterion. Several Swarm Intelligence (SI) algorithms have attracted researchers from the field of data mining. One of the nature inspired algorithm, ACO, has been used in combination to the clustering algorithm, which helps in the detection of outliers.

II. RELATED WORK

Bansal and Chugh [1] compared the outcomes from various clustering techniques and proposed a new method, taking the time complexity into consideration, which added fuzziness to already existing clustering methods.

Mankar and Ghuse [2] proposed a review of various outlier detection techniques. They showed how traditional methods and recent method works differently for outlier detection. They concluded that the most existing research focused on the algorithm based on special background. The efficiency of an outlier detection method depends greatly on the type of data and data distribution to be processed.

Dimble and Tidke [4] proposed an efficient clustering and density based outlier detection framework. The process has been categorized into two steps-Clustering of data based on any density based DBSCAN algorithm and Outlier detection is carried out using LOF. It had been shown that the proposed method works well with real world spatial data in terms of precision and recall values.

Sharma and Sejwal [7] have proposed a way of using spatial outlier detection technique to spot the less densely populated areas of the town or city. They have applied two advanced statistical approaches to the data set collected from different sites of Haryana.

Sumathi et al. [8] presented an overview of spatial data mining tasks, various techniques associated with it and also discussed spatial data mining trends and applications. Finally, it presented areas for further research needs in SDM.

Agrawal et al. [19] proposed a clustering based algorithm to detect outliers on high dimensional spatio-temporal data. The method first finds the k-nearest and k-shared nearest neighbor of each data object, and then builds the clusters around the identified center points. Once the clusters are formed, spatial outliers are identified and compared with temporal neighbors.

Hemalatha and Saranya [21] presented a thorough under- standing regarding SDM architecture, various methods, uses of various spatial operations, and algorithms needed for the discovery of knowledge in spatial databases. They also suggested that SDM is a promising field with many research areas and challenging issues.

Surya and Azhagusundari [22] presented a study of spatial outliers, its methods of detection and algorithms, including their complexity along with their pros and cons.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

III. PROPOSED METHODOLOGY

In the proposed method, the clustering algorithm is first performed which partition the dataset into the set of k clusters, based on the minimum euclidean distance measure. The value of k is then optimized with the help of ACO in order to represent k clusters in a more efficient manner. The outliers are then analyzed based on the final clusters thus formed which exhibits different characteristics. The overall structure of the proposed methodology is shown in Fig.1.

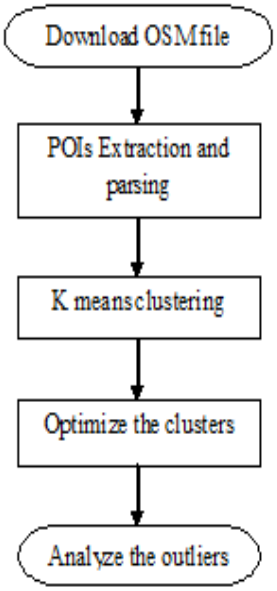

Fig.1 Schematic flow of proposed method

Step 1: Real world OSM data has been selected from OSM data repository (http://extract.bbbike.org/).

Step 2: From the given dataset, first the clusters are formed with the help of k-means clustering.

It is mainly divided into two parts: in the first phase the k cluster centers are to be chosen arbitrary where the k value is initialized before the process gets started. The next phase is the assigning of each data point to the k cluster centers according to some minimum distance calculation measure. The distance between each data point and cluster centers is calculated with the help of Euclidean distance. The reassignments of the cluster centers are done based on the averages of all the values of data points included in that particular cluster. This process continues repeatedly until there is no change in the values of clusters or minimum criterion function value reaches. The projected behavior to identify outliers is that they either do not be considered in any cluster, or belong to very small clusters, or are present forcefully in a cluster where they exhibits different behavior from the other data objects. Clustering based method has been frequently used in order to detect outliers more efficiently since they are few in number and exhibit different characteristics from other dataset features.

Step 3: After the cluster formation, Ant-Colony Optimization technique is applied to optimize the clusters. The ants are initialized to incrementally construct or to modify the clusters with the help of a probabilistic transition rule and on a local heuristic as:

$$P_{(i,D_n)} = \frac{\tau^{\alpha}_{(i,D_n)}\,\eta^{\beta}_{(i,D_n)}}{\sum_{j=0}^{k}\tau^{\alpha}_{(j,D_n)}\,\eta^{\beta}_{(j,D_n)}} \qquad (1)$$

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

where $P_{(i,D_n)}$ is the probability of data object D_n to reside in cluster i , $\tau(i, D_n)$ and $\eta(i, D_n)$ are the cluster and the distance related information, α and β are the constant parameters, and k is the value of clusters present.

$$\eta_{(i,D_n)} = \frac{K}{Dist(D_n C_i)} \qquad (2)$$

where D_n is n^{th} data object and C_i is i^{th} cluster center, $Dist(D_n, C_i)$ represents the distance calculations between D_n and C_i and Constant K is used to balance the values of η with τ accordingly.

Step 4: Once all the ants generate the clusters, a global pheromone updating rule is applied according to equation 3. It updates the k clusters.

$$\tau_{(i,D_n)} \leftarrow (1-\rho)\tau_{(i,D_n)} + \sum_i \Delta\, \tau_{(i,D_n)} \qquad (3)$$

where $\Delta\tau_{(i,D_n)}$ is the value of new added clusters, if any, to previous clusters by the next ant. The above mentioned equations from (1) to (3) are used with k-means algorithm to optimize the clusters to construct a clearer representation of clusters.

Step 5: The final clusters formed are then analyzed. The clusters having data objects at the farthest location are considered as outliers.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experiment is conducted on OpenStreetMap data which is selected from OSM data repository. The data set selected is of Ludhiana city with values between [minlng=”75.484”,minlat= ”30.618”] and [maxlng= ”76.309” maxlat=”31.141”]. The data set extracted is of XML data file format. The bank data nodes are extracted that contains 49 data objects with distinct node ids having information about the ATMs. There are two types of classes: class1 contains the information about the banks having attached ATMs with them; class2 represents the data about the banks which aren’t having the ATMs attached with them. As there are two types of datasets, so the number of clusters to be formed is taken as 2.The proposed method is implemented on the data set and the results obtained are represented below in Fig. 2.

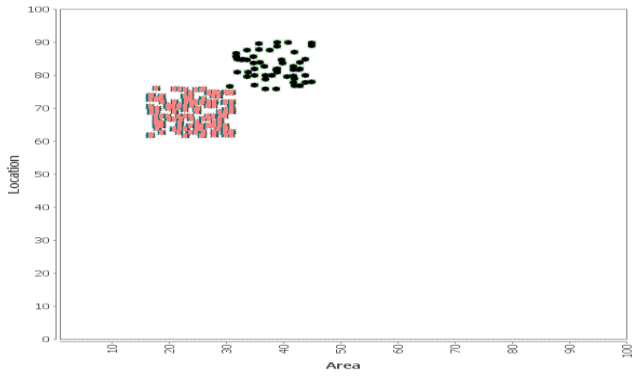

Fig.2 Dataset with 2 clusters

The Table I depict information of the number of data objects contained in each clusters. It also shows the number of outliers detected in each cluster.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Table I shows data objects and outlier details

Number of data objects in each cluster		Number of outliers detected
Cluster1	34	3
Cluster2	15	5

V. CONCLUSION

There are many techniques available for finding outliers in spatial data. Clustering is considered as the most frequently used technique. In this work, clustering algorithm i.e. k-means algorithm is combined with optimization technique in order to get effective clusters. It is because k-means algorithm is sensitive to outliers. The experimental results are conducted on bank dataset. The hybrid algorithm first cluster the non-spatial attribute values based on single common attribute value. After the cluster formation, the data objects which lies far away from the other data objects are considered as outlier objects. The hybrid technique identifies more outliers than the existing k-means algorithm.

The approach needs to be implemented to identify various other types of outliers present like spatio-temporal, temporal outliers whose detection holds an important research area. Future work needs to consider multiple non-spatial attributes to detect more efficient outliers.

REFERENCES

[1] N. Bansal, A. Chugh, Differentiate clustering approaches for outlier detection, International Journal of Innovative Research in Computer and Communication Engineering, 1(2), 2013, 193-196.

[2] A. B. Mankar and N. Ghuse, A review on detection of outliers over high dimensional streaming data using cluster based hybrid approach, International Journal of Science and Research, 3(11), 2014, 1850-1856.

[3] M. A. Khan, S. K. Pradhan and M. A. Khaleel, Outlier detection for business intelligence using Data Mining techniques, International Journal of Computer Applications, 106(2), 2014, 28-31.

[4] N. M. Dimble and B. Tidke, A framework for outlier detection in geographic spatial data, International Journal in Foundations of Computer Science & Technology, 5(2), 2015.

[5] N. Chaudhary, K. Sharma, M. Kumar and A. Tomar, Spatial outlier detection techniques, International Journal of Advance Research and Innovative Ideas In Education, 2(5), 2016.

[6] M. Aakunuri, G. Narasimha and S. Katherapaka, Spatial Data Mining: A recent survey and new discussions, International Journal of Computer Science and Information Technologies, 2(4), 2011, 1501-1504.

[7] A. Sharma and A. Sejwal, Multi-variant spatial outlier approach to detect less developed sites in given region, Global Journal of Business Management and Information Technology, 1(2), 2011, 77-83.

[8] N. Sumathi, R. Geetha and S. S. Bama, Spatial data mining-techniques trends and its applications, Journal of Computer Applications, 1(4), 2008, 28-30.

[9] S. Bhadoria and U. Datta, An analytic survey on current clustering technique of data categorizing and

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

retrieving, International Journal Of Engineering And Computer Science, 5(5), 2016, 16504-16508.

[10] A. Sharma and A. Sejwal, Multi-variant spatial outlier approach to detect less developed sites in given region, Global Journal of Business Management and Information Technology, 1(2), 2011, 77-83.

[11] G. Singh and V. Kumar, An efficient clustering and distance based approach for outlier detection, International Journal of Computer Trends and Technology, 4(7), 2013, 2067-2072.

[12] P. Dhivya and B. Rajdeepa, Spatial data mining using cluster analysis, International Journal of Innovative Research in Computer and Communication Engineering, 4(7), 2016, 13406-13410.

[13] M. Parimala, D. Lopez and N. C. Senthilkumar, A survey on density based clustering algorithms for mining large spatial databases, International Journal of Advanced Science and Technology, 31, 2011, 59-66.

[14] S. Vijayarani, P. Jothi, An efficient clustering algorithm for outlier detection in data streams, International Journal of Advanced Research in Computer and Communication Engineering, 2(9), 2013, 3657-3665.

[15] P. Sudha and K. Krithigadevi, Outlier detection using high dimensional dataset for comparision of clustering algorithms, International Journal of Advanced Research in Computer Science & Technology, 2(3), 2014, 283-288.

[16] S. Aggrwal and P. Kaur, Survey of partition based clustering algorithm used for outlier detection, International Journal for Advance Research in Engineering and Technology, 1(5), 2013, 57-62.

[17] M. H. Marghny and A. I. Taloba, Outlier detection using improved genetic K-means, International Journal of Computer Applications, 28, 2011, 33-36.

[18] J. Singh and S. Aggarwal, Survey on outlier detection in Data Mining, International Journal of Computer Applications, 67, 2013, 29-32.

[19] K. .P. Agrawal, S. Garg and P. Patel, Spatio-Temporal outlier detection technique, International Journal of Computer Science & Communication, 6(2), 2015, 330-337.

[20] P. Murugavel and M. Punithavalli, Improved hybrid clustering and distance-based technique for outlier removal, International Journal on Computer Science and Engineering, 3(1), 2011, 333-339.

[21] M. Hemalatha and N.N. Saranya, A recent survey on knowledge discovery in Spatial Data Mining, International Journal of Computer Science Issues, 8(3), 2011, 473-479.

[22] T. Surya and B. Azhagusundari, Spatial outlier detection approaches and methods: A survey, International Journal of Innovative Research & Development, 3(4), 2014, 24-29.