ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 7, July 2016

Extended DBSCAN Algorithm to Detect Cluster with Varied Density for Outlier Detection

Amey K.Redkar
PG Scholar
JSPM's Imperial College of Engineering & Research
Wagholi Pune, India

Prof.S.R.Todmal
Professor
JSPM's Imperial College of Engineering & Research
Wagholi Pune,India

***Abstract*— There are many methods on density based clustering. DBSCAN is one of the algorithm in which density based clustering method is used to detect outliers. DBSCAN is the simple of density based clustering method. Density based clustering helps in finding out and identifying dense clusters with any shape and size. In DBSCAN algorithm we have to give two parameters as input but it has one disadvantage that it fails to detect clusters with different densities. This paper propose a method EtDBSCAN based on DBSCAN that not only detect clusters with different densities but also in addition detect outliers correctly.**

Keywords - Component, density based clustering, outliers, rollback.

I. INTRODUCTION

In network security and data mining domains outlier's detection is one of the important area of research. Outliers are the collections in the patterns with dissimilar properties. Detecting outliers is one of the well known problems in data security. Clustering is the solution to detect outliers from the patterns and keeping aside the outliers in terms of noise. Clustering is used in many applications such as: Biology, Marketing, Libraries, Insurance, Planning, Earthquakes etc. Actually clustering is a unsupervised learning task.

Actually clustering can be used in different categories such as : Model based clustering, partitioned clustering, density based clustering, hierarchical clustering, cellular clustering. From this we are using density based clustering approach which is a nonparametric approach. In such methods cluster is granted as a high density area with density p(x). In such methods we don't have to give any input value for the number of clusters also it does need to make any assumption regarding the density required or regarding any variance in the given datasets or the nodes in the networks. Therefore density based clusters does not from any purposeful shape or any assume shape. Its shape can be a arbitrary shape and not compulsory it forms any assume shape. Whereas it forms a set of data objects which has points from the region covered by the high density area separated by the region or points of data objects from the region of low density[1].

There arise many number of problems in using density based methods in clustering one of the issue is the number of dimensions present and if large number of data items or objects needed to be scanned.[2] and [3] shows the which approaches can be used to handle large number of data items.

Density based clustering also has another problem as it considers clusters which has low number of nodes as noise although they are not the outliers.

DBSCAN(Density Based Spatial Clustering of Applications with Noise) is one of most recently used and simple approach in detection of outliers used in many fields of science. It is widely used in network security and data mining. DBSCAN has best ability to detect cluster with different shape and sizes.[4] It has one problem of detecting cluster with varied of different densities. In this paper we propose to detect outliers with varied or multiple densities.

II. LITERATURE SURVEY

Due to simplicity and best performance of DBSCAN various techniques have been developed to solve the problem of DBSCAN[5].

[6]OPTICS is one of the approach which is the famous one and is the version of DBSCAN. It doesn't need to find out the numbers of parameters needed. DBSCAN and OPTICS has one disadvantage of failure in determining the border points of the adjacent cluster. This approach uses hierarchical clustering technique to determine the clusters. OPTICS algorithm does not produce clustering of dataset but creates an arguments ordering of database using density based clustering, This method helps to detect cluster of various shapes and size and further helps in detecting outliers. This method derives a list of argument making use of broad range of parameters.

Further EnDBSCAN was developed to determine the border objects of the clusters. It was a an extension of DBSCAN. In this approach data distribution was used to detect the membership of the objects. In this approach also new features are added to increase the performance of DBSCAN. It can be further developed to polygon shaped datasets. Outlier detection and cleaning has been introduced in this approach which adds to enhancement of the DBSCAN algorithm.

DDSC[7] is another extension of DBSCAN.

It uses uniform density to find out the clusters in order to separate overlapped region. Therefore even if the adjacent cluster has different densities they would disjoint in two separate clusters. It calculates density variation of core object with respect to the cluster also it keeps track of density of the clusters. LDBSCAN[8],GDCIC[9] and GMDBSCAN[10] are the another extension of DBSCAN for solving the problems of DBSCAN.

UDBSCAN[14] density based clustering algorithm to uncertain objects.

Here a deviation function is proposed which approximates alll the model of the objects and it uses the standard deviation. Till now there is no cluster quality measurement.In this method a metric is used to calculate the density measurement of the cluster solution. Finally

experiments are carried to measure the density quality of the cluster solution. The rsults show that the UDBSCAN show good results to detect multi cluster than traditional approach.

[14]Grid and Density Based Clustering Algorithm with Relative Entropy Guoyan Huang, Ding Wang-

Mostly traditional clustering algorithms are based on grid and density approach which have a smaller time complexity. Here users need input parameters like density threshold and clustering precision is not high in the cluster edge. To improve the quality of clustering, in this method implement GDBRE (a Grid and Density Based clustering algorithm with Relative Entropy). Initially it defines grid relative entropy to calculate the density threshold then the density-connected dense grids are chosen to clustering according to the breadth-first search strategy. Finally adjacency of sparse grid is analyze. The data points of sparse grid are marked as boundary or noise points by calculating the Euclidean distance to adjacent dense grid centroid.

[16] An Efficient Algorithm for outlier detection S.Vijayarani,S.Nithya-

This method work with finding hidden patterns from the given large data sets.In this technique it works for finding hidden pattern for health related problem. PAM,CLARA,ECLARANS and CLARANS are used for outlier detection in this approach. It is used to extract data from lare datasets.

[17] Outlier detection in stream data by clustering method. Hossein Moradi Koupaie, Suhaimi Ibrahim-

This survey focus on the detection of outliers in stream data. It uses clustering method that detect outliers in given period. It work with the assumption of outliers that has been previously discovered and lastly find out the real outliers in stream data.

[18] The New Density Based Clustering Technique Rwand D. Ahmed-

A grid based technique is used in this approach which reduces the time complexity. Grid based technique divides the data space into cells and any number of points scattered within the grid are chosen. These scattered points extent the shape of the dataset taken. Than the whole data is considered as the cell and the whole clustering is done on the cell. Then all the cells are merged. MinPts are used to overcome the problem of clustering with multi density.

[19] Enhanced DBSCAN algorithm Meghana Sharma Priyamwada Palliwal

It is an extension of DBSCAN. In this approach detected cluster are not only separated by sparse region but also separated by varied density. Here the time complexity of DBSCAN algorithm is reduce and this algorithm takes both points and pixels. This paper has additional features for outlier detection. Further in this approach can be extended for cluster with varied density and also it can be further expanded for datasets of polygon shaped.

[20] Density based algorithm for discovering density varied cluster in large spatial database. Anant Ram, Sunita Jalal

DBSCAN is a base algorithm for density based clustering. It can detect the clusters of different shapes and sizes from the large amount of data which contains noise and outliers. However, it is fail to handle the local density variation that exists within the cluster. This paper, propose a density varied DBSCAN algorithm which is capable to handle local density variation within the cluster. It calculates the growing cluster density mean and then the cluster density variance for any core object, which is supposed to be expended further, by considering density of its neighborhood with respect to cluster density mean. If cluster density variance for a core object is less than or equal to a threshold value and also satisfying the cluster similarity index, then it will allow the core object for expansion.

III. DBSCAN

DBSCAN(Density Based Spatial Clustering of Applications with Noise) This algorithm keeps on growing forming regions of cluster according to the points covered in the density of the cluster and keeps on growing with high density cluster formation. It forms a cluster of high density which is of arbitrary shape of spatial data form with noise. Lastly it forms a cluster which is the maximal length cluster forms covering maximum points.

DBSCAN has to be given two user defined input parameters (neighborhood radius). It shows the neighbor which falls within the radius ε and also the minimum points required within the given radius to confirm if the selected point is core point or not. Therefore if the minimum points defined by the users fall within the radius ε of a point or object than it is called core object.

In DBSCAN when the object p is within neighborhood radius ε of q than q is core object and it is said that p is density reachable from q. Now if another core object o has p and q density reachable from o than p and q are said to be density connected.

Therefore in density based clustering any object present in any cluster is density connected or density reachable. And the objects not present in the cluster or not density reachable will consider outliers or noise[11].

IV. LIMITATION OF DBSCAN

DBSCAN is unable to detect clusters with different densities in dataset. Consider the figure given below.

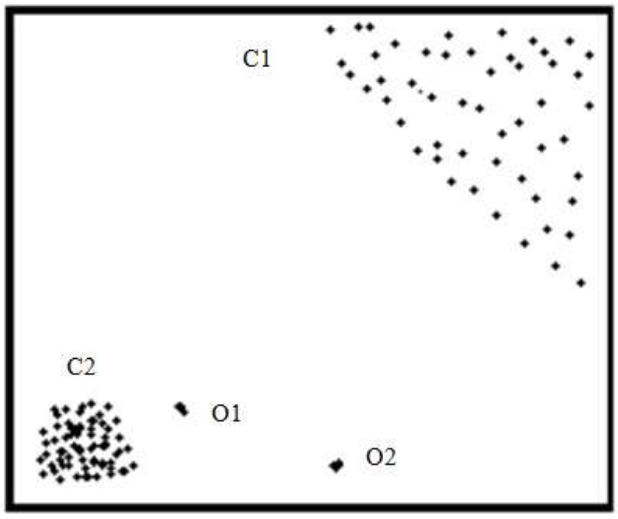


Fig 1: A dataset has multi-density clusters

From the figure if low value is selected for the distance between the objects of C2 than objects from the C1 cluster will be consider as the outliers on the other hand if high value is chosen for the distance between the objects between C2 than the O2 which is an outlier will be consider as the object of C2.

This problem takes place as the C1 and C2 has different densities for their clusters. While the DBSCAN uses constant user defined parameter while forming a cluster. OPTICS try to find the best density radius ε value through the calculation of the statistical approach but lastly OPTICS gives similar functioning as that of DBSCAN and cannot detect complex cluster with varied density for outlier detection.

V. PROPOSED SYSTEM

In the propose algorithm EtDBSCAN only one parameter is needed to be given to the algorithm. Minpts required in the given radius ε density has to be defined. Firstly keep radius ε small and further radius can be increase gradually so that most of the neighborhoods can be cover but one problem can occur that outliers can covered due to increase in radius ε and outlier can be considered as points in the cluster therefore statistical summary is carried on the data points to detect the outliers. The algorithm is described below.

During clustering firstly choose a point p randomly from the given points to form the first cluster and if p is core object than the with same radius ε cluster expansion is carried out and if chosen point is not core point than value Δε is added to the radius (ε + Δε). Now again p is examined to check if the it is a core object and if succeeded than with same radius (ε + Δε) the cluster is expanded and if not again Δε is added to the radius until p will be a core point or there is no possibility of increasing the cluster by neighborhood radius.

Condition may arise that the selected point may be a outlier so here after number of expansion of cluster the outlier becomes a core point. To solve this problem first detect the varied cluster and than find out the outliers. Now during expansion of the cluster after the k steps suppose the point is detected as core object after radius (ε + kΔε). From this point as that of DBSCAN cluster is expanded and the average distance is calculated between the points present in the cluster and without considering the distance between point p and its neighbor are calculated. If the calculated distance is much more lower than neighborhood (ε + kΔε) than it may concluded that the distance between the point p and other points in the cluster has no logic appearance and so p can be consider as outlier. Figure 2 shows the condition where p is detected as outlier.

After detection of outliers that generated cluster will be cancelled and here the cluster will be rollback. Again choose another points randomly and carry out the previous steps taking initial value for the radius to find out new core point.

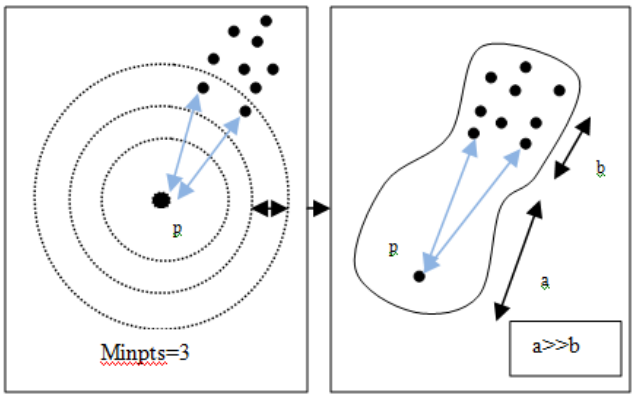


Fig 2:Detection of Outliers

The process is continued until all the points are assigned. The pseudo code for the propose method is shown below.

Algorithm for EtBDSCAN

Step 1.Function EtDBSCAN(int Minpts){

Step 2.eps= min_eps

Step 3.While there is a unlabel point{

Step 4. Select p as an unlabel random point

Step 5. eps = min_eps

Step 6. While(p isn't core) eps = eps + Δε

Step 7. Expand cluster C by start with p in DBSCAN way

Step 8. if $(\text{avg}_{i \in \text{neighbor}(p)}(N_{c(i)} - N_{c(\text{Neighbour}(i))}) << \text{eps})$

Step 9. Rollback(C)

Step 10. Label p as an outlier}

VI. MATHEMATICAL MODEL

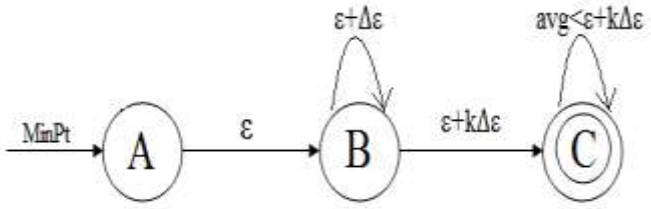


Figure : - Mathematical modelling DFA

A deterministic finite automaton M is a 5-tuple, (Q, ∑, δ, qo, F) consisting of

- A finite set of states (Q)={A, B, C}

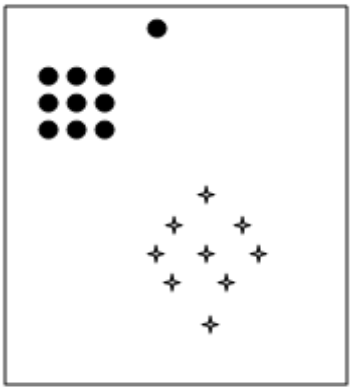
- A finite set of input symbols called the alphabet(∑)={MinPt, ε, ε+Δε, ε+kΔε, avg < ε+kΔε }
- A transition function (δ: Qx∑→Q)={ }
- A start state (qo € Q)={q0}
- A set of accept states (F ⊆ Q)={q2}

WHERE,
MinPt = Number of Minimum points
ε = Radius
ε+Δε = Increased value of radius
ε+kΔε = Radius of cluster at Kth point
avg < ε+kΔε = Average distance between points and cluster

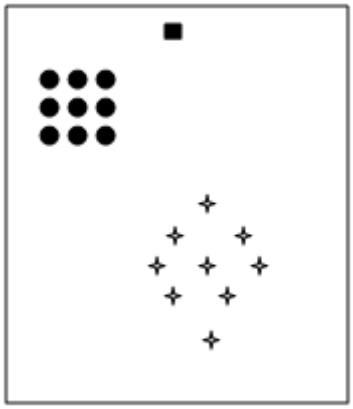
States / δ	MinPt	ε	ε+Δε	ε+kΔε	Avg < ε+kΔε
A	A	B	Ø	Ø	Ø
B	Ø	Ø	B	C	Ø
C	Ø	Ø	Ø	Ø	C

VII. Evaluation Strategy

The proposed method is implemented in Java language in netbeans environment. It has been tested on various data sets with our method.
First test was done on manually created datasets that is shown in the figure 3. Test was perform on data sets using DBSCAN and propose method. Figure shows that the DBSCAN method has difficulty in determining outliers but the propose method shows the outliers(Outliers are shown in black square shape).



a)DBSCAN



b)EtDBSCAN

Fig 3: Comparison of two methods on first data set to detect outliers

In fig 4 our method is tested on data sets and compare the time required to find the cluster with the K-means and K-medoid algorithm.

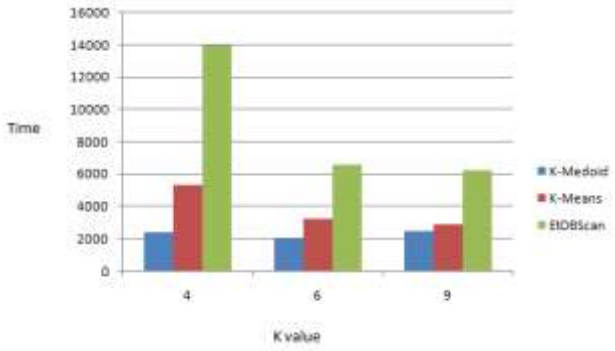


Fig 2

VIII. Conclusion

The approach propose through this paper is to overcome the formation of cluster for varied density in outlier detection. This method overcomes the problem of DBSCAN algorithm in case of multiple densities.
In this approach greedy technique is used to find out the density and further parameters of density are adjusted according to the cluster formation.
Detection of cluster and outliers using EtDBSCAN shows the advantage of this technique over DBSCAN.

References

[1] Pragati Shrivastava,Hitesh Gupta " A Review on Density Based Data Clustering in Spatial Data",International Journal of Advanced Computer,2012.

[2] Barnett V, Lewis T, "Outliers in Statistical Data"Wiley,3rd Edition,1995.

[3] C. Aggarwal, P. Yu, "Outlier Detection for High Dimensional Data",ACM SIGMOD Conference Proccedings,2001.

[4] M. M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, "LOF: Identifying Density-Based Local Outliers", ACM SIGMOD Conference Proccedings,2001.

[5] S. Ramaswamy, R. Rastogi, S. Kyuseok, "Efficient Algorithms for Mining Outliers from Large Data Sets", International Journal of Adnanced Research in Computer Science and Software Endineering, 2000.

[6] M. F. Jiang, S. S. Tseng, C. M. Su"Two-phase Clustering Process for Outliers Detection"Elsevier Science, 2001.

[7] Garfinkel, S. and H. Holtzman,"Understanding RFId Technology, in RFID: Applications, Security, and Privacy",International Journal of Technology and Design Education, 2005.

[8] Priyamvada Paliwal,Meghana Sharma,"Enhanced DBSCAN Algorithm",International Journal of Adnanced Research in Computer Science and Software Endineering,2013.

[9] Asieh Ghanbarpour,Behrooz Minaei,"ExDBSCAN an Extension of DBSCAN to Detect Cluster in Multi Density Datasets",IEEE Conference,2013.

[10] F. Angiulli, C. Pizzuti,"Fast Outlier Detection in High Dimensional Spaces"Springar-Verlag,2002.

[11] K. Yamanishi, J. Takeuchi, G. Williams, "On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms",IEEE Computer Society,2000.

[12] Henrik Bäcklund (henba892), Anders Hedblom (andh893), Niklas Neijman (nikne866), "A Density-Based Spatial Clustering of Application with Noise",International Journal of Computer Trends and Technology,2003.

[13] King Mongkuts,Thonburi, "U-DBSCAN- A Density Based Clustering Algorithm for Uncertain Objects", International Journal of Advanced Computer ,2010.

[14] Rajendra Pamula, Jatindra Kumar Deka "Grid and Density Based Clustering Algorithm with Relative Entropy",Advanced in Information Sciences and Service Sciences,2005.

[15] Carlos H. C. Teixeira, Gustavo H. Orair, "An Efficient Algorithm for Outlier Detection", International Journal of Advanced Computer,Researchgate:Information Systems,2011.

[16] Hossein Moradi Koupaie, Suhaimi Ibrahim ,"Outlier Detection in Stream Data by Clustering Method", International Journal of Advanced Computer Science and Information Technology ,2011.

[17] Rwand D. Ahmed, "The New Density Based Clustering Technique, International Journal of Advanced Research in Computer Science" ,2012.

[18] Meghana Sharma,Priyamwada Palliwal, "Enhanced DBSCAN Algorithm, International Journal of Computer Trends and Technology" ,2013.