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

A Survey on DBSCAN Algorithm To Detect Cluster With Varied Density.

Amey K. Redkar, Prof. S .R. Todmal

Abstract— **Density -based clustering methods are one of the important category of clustering methods that are able to identify areas with dense clusters of different shape and size.One of the basic and simple methods in this group is DBSCAN . This algorithm clusters dataset based on two received parameters from the user one is min points and second is the radius. One of the disadvantages of DBSCAN is its inability in identifying clusters with different densities in a dataset. In this paper, we are carrying out a survey to find out various algorithms related to DBSCAN and also other algorithm which can be used to detect clusters and outliers. Also their capability to detect cluster with varied density will be surved. Here multi density data set cluster detecting algorithm will be preffered.**

Index Terms—***component; density-based clustering; outlier;rollback.***

I. INTRODUCTION

In network security and data mining domains outliers 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, partitional 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 did not 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. 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. 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.

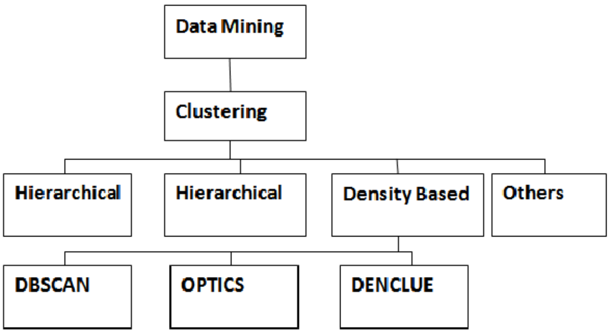

Fig1: DBSCAN in hierarchy of data mining.

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. Density based clustering also has another

Amey K. Redkar, *ME Research Scholar, Computer Engineering Department, University of Pune, Imperial College of Engineering and Research, Pune, Maharashtra, India, Mobile No +917507657841*
Dnyaneshwar A. Rokade, *Asst. Professor, Computer Engineering Department, University of Pune, Imperial College of Engineering and Research, Pune, Maharashtra, India.*

problem as it considers clusters which has low number of nodes as noise although they are not the outliers.

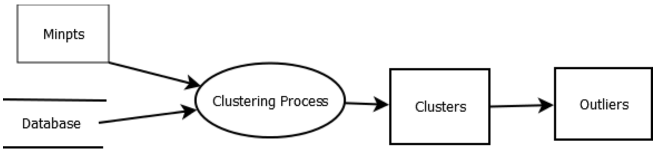

Fig 2: Working of Dbscan

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. 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.Data mining is the process of extracting hidden and interesting patterns or characteristics from very large datasets and using it in decision making and prediction of future behavior. This increases the need for efficient and effective analysis methods to make use of this information. One of the tasks is clustering where a set of objects is divided into several clusters where the intra-cluster similarity is maximized and the inter-cluster similarity is minimized]. The disadvantages in most of the traditional clustering algorithms are high computational complexity and that they do not scale well with the size of the very large datasets, so the development of enhanced clustering algorithms has received a lot of attention in the last few years. There are different clustering methods that can be used for handling very large datasets. These techniques can be categorized into partitioning,hierarchal,grid-based,density-based, model-based and constrain-based methods.
Techniques under the partitioning category are PAM, CLARA and CLARANS. These methods segment data into k groups where the value of k is supplied by the user. Density-based techniques were introduced to determine the arbitrary shaped cluster in spatial databases having noise. The DBSCAN , DENCLUE and OPTICS [11] are commonly used density-based clustering techniques.These algorithms optimize the best fit between the given data and a mathematical model. The constraint-based clustering techniques find the clusters that satisfy the userspecified preference or constraint Algorithms proposed under each of these categories try to challenge the clustering problems treating the large amount of data in large database. However,none of them are most effective. A clustering technique is considered to be good if it satisfies the following requirements 1)Minimal requirements of the domain knowledge to determine the values of its input parameters, which is very important problem especially for large data sets. 2)Discovery of arbitrary shaped clusters. 3)Good efficiency on large data sets. The density-based clustering algorithms are useful to discover clusters from the datasets with arbitrary shape and of large size. These algorithms typically cluster as dense regions of points in the data space that are separated by regions of low density. DBSCAN is the first density based clustering technique. It grows clusters according to a density based connectivity analysis.

Problem Statement

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is density-based clustering approaches used to detect outliers and clusters of different sizes and shapes but one of the problem of this approach is not realizing clusters with different densities.

II. RELATED WORK

Active Density-based Clustering By:Son T. Mai, Xiao He [2013][1]

This paper propose a novel active density-based clustering algorithm to deal with the sparseness (incompleteness) of the pairwise similarity matrix in many applications. Based on the availability of a LB similarity matrix, Act- DBSCAN iteratively selects the most informative pairwise LB similarities to update with their true similarities and refines the cluster structure.The general idea is to reach as close to the desired clustering result of DBSCAN as possible with each update. Act-DBSCAN contains an efficient probabilistic model and a scoring system called the Shared Core Object (SCO) score to evaluate the impact of the update of each pairwise LB similarity on the change of the intermediate cluster structure. Deriving from the monotonicity and reduction property of our clustering scheme and the SCO score, the two algorithms Splitting with SCO (SP-SCO) and Merging with SCO (MG-SCO) provide two different and efficient ways to actively select and update pairwise similarities and cluster results.

A Text Mining Model Based on Improved Density Clustering Algorithm By: Chen Qi,Lu Jianfeng, Zhang Hao[2013][2]

This paper has given a new clustering algorithm of text mining based on improved density clustering. The clustering algorithm based on density is widely used on text mining model for example the DBSCAN(densitybased spatial clustering of application with noise) algorithmDBSCAN algorithm is sensitive in choose of parameters, it is hard to find suitable parameters. In this paper a method based on k-means algorithm is introduced to estimate the E neighborhood and minpts. Finally an example is given to show the effectiveness of this algorithm.

The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters By: Tran Manh Thang, Juntae Kim[2011][3]

This paper introduce a new approach, DBSCAN-MP, provides new way to use DBSCAN with multiple parameters. It is suitable for anomaly detection in network traffic which contains multiple types The algorithm use also can update normal behavior by creating new clusters and updating clusters size. The results show that our approach creates new clusters and updates clusters size when we injected new attacks to remaining data parts. However the approach has a drawback in detecting DOS attack type. It also makes high false positive rate when network environment change overtime.

CGDBSCAN: DBSCAN Algorithm Based on Contribution and Grid By:Linmeng Zhang, Zhigao Xu[2013][4]

GbDBSCAN (an efficient grid-based DBSCAN algorithm) is an excellent improved DBSCAN algorithm, which makes up the defects that DBSCAN algorithm is sensitive to clustering parameters and unable to deal with large database, and retains the advantage of separating noise and finding arbitrary shape clusters. However, in GbDBSCAN, the grid technique treats the total number of points in one grid as the grid dense, and this simple treatment will depress the clustering accuracy. Therefore, CGDBSCAN is proposed in this paper, and within it migration-coefficient conception is presented firstly. With the optimization effect of contribution and migration-coefficient, the optimal selection of parameter Eps and the efficient SP-tree query index, the accuracy of clustering result is effectively improved while ensuring the operational efficiency of this algorithm.

An Improved DBSCAN, A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets By:Glory H.Shah[2012][5]

In this paper, an approach to improve dbscan clustering algorithm is introduced. Before forming the clusteres, the dataset is split based on threshold value. Also the proposed algorithm is tested on synthetic data set. The tests were performed considering all types of data sets i.e. ranging from low to high dimensional. It was found that euclidean distance works well on low dimensional data sets but does not work well on high dimensional data sets. Also it was found that,even though Euclidean distance forms more number of clusters, but the number of incorrectly clustered instances is more as compared to using distance as manhattan. The noise ratio remains high when distance measure as manhattan as compared to euclidean. But the time elapsed by the algorithm to form clusters is more when consider distance as euclidean as compared to manhattan.

DBSCAN: Past, Present and Future By:Kamran Khan,Simon Fong[2013][6]

This paper presents the summary information of the different enhancement of density-based clustering algorithm called the DBSCAN. The purpose of these variations is to enhance DBSCAN to get the efficient clustering results from the underlying datasets. In addition it also Highlights the research contributions and found out some limitations in different research works.Consequently, this work also depicts the critical evaluation in which comparison and contrast ve been taken out to show the similarities and differences among different authors works. The spatiality of this work is that it reveals the literature review of different DBSCAN modification and provides a vast amount of information under a single paper.

Evolutionary Clustering with DBSCAN By:Yuchao Zhang, Hongfu Liu[2013][7]

In this paper, initially propose an evolutionary clustering algorithm with DBSCAN to solve the density-based evolutionary clustering problem under the framework of temporal smoothness penalty. Through the experiments, proposed evolutionary approach show an advantage over the synthetic datasets. Compared with the other similar evolutionary clustering algorithms, such as the evolutionary K-means clustering, our method can not only resist to the noise, but also distinguish the clusters with arbitrary shapes during the evolution However, it is still necessary to take the parameters of the density-based algorithms into consideration, such as the EPS and MINPTS in DBSCAN, which are usually set prior to the clustering by users. Parameters selection is an important topic and sensitive problems for density-based clustering, especially for those contexts where the nodes number and density are both evolving with the time.

GCMDDBSCAN: DBSCAN Based on Grid and Contribution By:Linmeng Zhang,Zhigao Xu[2013][8]

DBSCAN (Density Based Spatial Clustering of Application with Noise) is an excellent densitybased clustering algorithm, which extends DBSCAN algorithm so as to be able to discover the different densities clusters, and retains the advantage of separating noise and finding arbitrary shape clusters. But, because of great memory demand and low calculation efficiency, Multi Density DBSCAN cannot deal with large database. Therefore, GCMDDBSCAN is proposed in this paper, and within it migration-coefficient conception is introduced firstly. In GCMDDBSCAN, with the grid technique, the optimization effect of contribution and migration coefficient,and the efficient SP-tree query index, the runtime is reduced a lot, and the capability of clustering large database is obviously enhanced, at the same time, the accuracy of clustering result is not degraded.

III. PROPOSED WORK

In the last decades clustering techniques has attracted the attention of the most researcher therefore new clustering algorithm has been proposed and DBSCAN is one of the proposed algorithm. Important methods of clustering are partitioning ,hierarchical and density based To challenge the clustering problem taking place in huge amount of data in large dataset, these algorithms are very useful. Clustering for large amount of data is an active research topic for last several years and still is.DBSCAN (Density Based Spatial Clustering of Applications with Noise) is one of most popular and classical density based clustering algorithm . DBSCAN algorithm used two important input parameters Epsilon (Eps) and minimum point (MinPts) and also used no. of cluster, unclustered instances, incorrectly instances well as time and noise ratio. Density-based clustering algorithms are proposed based on several concepts including: a)Core Point: A point is a core point if it has more than a specified number of points (MinPts) within Eps. These are points that are at the interior of a cluster.b)Border Point: A border point has fewer than MinPts within specified radius (Eps), but is in the neighborhood of a core point.c)Noise Point: A noise point is any point that is not a core point or a border point.d)Other is neighborhood of the point, directly density-reachable, density-reachable and density-connected, cluster. Compared with other clustering algorithms, density based clustering technique,such as DBSCAN, has several advantages as follows

1) The number of clusters in a data set is not required to be input before carrying out the clustering.

2) The detected clusters can be represented in an arbitrary.

3) Noise or outlier are detected or removed with help of filter.

4) DBSCAN requires only two parameters which is used find the Euclidean and Manhattan distance

.

IV. CONCLUSION

The approach propose through this project 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.This method uses the greedy technique for finding a cluster density and expand that cluster using found parameter. But determining a density for each cluster needs time, so our approach needs more time to run than DBSCAN, this is the disadvantages of our approach.So for future we suggest to use similar technique to greedy which can reduce the time required to run the algorithm..

REFERENCES

[1] Son T. Mai, Xiao He, "Active Density-based Clustering ," IEEE Transactions On Data Mining,vol.23,No. 3,July-December 2013.

[2] Chen Qi,Lu Jianfeng, Zhang Hao, "A Text Mining Model Based on Improved Density Clustering Tran Manh Thang, Juntae Kim, "The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters ," IEEE Transactions On Management of Data, Vol. 8, No. 12,December 2011.

[3] Linmeng Zhang, Zhigao Xu, "CGDBSCAN: DBSCAN Algorithm Based on Contribution and Grid," Sixth International Symposium on Computational Intelligence and Design, Vol.15, No. 2, Second Quarter 2013.

[4] Glory H.Shah, "An Improved DBSCAN, A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets ," INTERNATIONAL CONFERENCE ON ENGINEERING,, Vol. 1, No. 7, oct-Dec 2012.

[5] Kamran Khan,Simon Fong, "DBSCAN: Past, Present and Future ," IEEE Transactions On Data Mining, Vol. 11, No. 2, September 2014.

[6]Yuchao Zhanga, Hongfu Liu, "Evolutionary Clustering with DBSCAN ," Ninth International Conference on Natural Computation, Vol. 12, No. 5, 2013.

[7] Linmeng Zhang,Zhigao Xu, "GCMDDBSCAN: DBSCAN Based on Grid and Contribution ," International Conference on Dependable, Autonomic and Secure Computing, Vol. 1, No.9, 2013.

[8] Chetan Dharni,Meenakshi Bnasal, "An Improvement of DBSCAN Algorithm to Analyze Cluster for Large Datasets ," IEEE Transactions On Data Mining, Vol. 1, No. 6, September 2013.

[9] Apinya Tepwankul , Songrit Maneewongwattana , "U-DBSCAN : A Density-Based Clustering Algorithm for Uncertain Objects ," IEEE Conferece on ICDE workshop, Vol. 3, No. 9, September 2010.

AMEY K. REDKAR is M.E Research Scholar from the Imperial College of Engineering & Research, J.S.P.M Pune, Maharashtra, India. Main research areas: wireless networks, Cloud Computing, Computer network, Network Security.

S.R.TODMAL is an Professor (Computer Departmnt) from the Imperial College of Engineering & Research, J.S.P.M Pune, Maharashtra, India.Main research areas: Computer network, Data Mining, Network Security, cloud computing .