Contrivancing DBSCAN Algorithm on Spatial Data Using Matlab

Sarika Chaudhary,Pooja Batra Nagpal
Assistant professor, Amity school of engineering and technology
Amity University, Haryana

Abstract: With the headway of technology the huge volume of heterogeneous and diverse dimension data to be processed continues to show an exponential rise in all science and engineering domains resulting into spatial data. Data mining is the technique for discovering interesting patterns from large scale data. Clustering is an unsupervised learning of hidden data concept i.e. it divides data into clusters of similar objects; each cluster consists of objects that are similar between themselves and dissimilar to objects of other groups. In our paper we are implementing a Density based clustering technique i.e. DBSCAN on different datasets i.e. real and synthetic spatial data in order to manage data in efficient manner.

Keywords: Clustering, Data-mining, DBSCAN, Matlab.

I. INTRODUCTION

Data mining is the process of extracting hidden and interesting patterns or characteristics from large scale datasets and using it in decision making and prediction of future behaviour. Clustering plays a crucial role in the data mining. Clustering is the process of dividing a set of objects into different clusters in order to increase the intra-cluster similarity and to reduce the inter-cluster similarity [1]. But the problem with the traditional clustering techniques is that they are not suite well with spatial data because of their high complexity in term of size and execution time [1]. There are many clustering algorithms that can be used to handle spatial data such as grid based, density based, hierarchal and partitioning based methods.DBSCAN (Density Based Spatial Clustering of Applications with Noise) is one of most popular and classical density based clustering algorithm [2]. 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 [1]-[3] are proposed on several concepts including core point, border point and noise point. Compared with other clustering algorithms, density based clustering technique, such as DBSCAN [10], has several advantages and disadvantages as follows.

- As in K-means number of clusters should know in advance but in case of DBSCAN it is not required.
- Clusters of arbitrary shape can be detected.
- Noise or outlier are detected or removed with help of filter.
- DBSCAN requires only two parameters which can be found using Euclidean distance or Manhattan distance.
- Accurate initial value of Eps and MinPts are difficult to find.
- DBSCAN cannot cluster data sets well with large differences in densities [11].

This paper discuss about DBSCAN algorithm from field of data mining. The main idea of this paper is to evaluate and compare the runtime efficiency as well as noise ratio of different dataset. All the different dataset is run on MATLAB. The rest of this paper is organized as follows. Section 2 discusses literature review on clustering techniques. The proposed algorithm for DBSCAN using Matlab is presented insection 3 and experimental studies and performance evaluation is presented in section 4. Section 5 describes the conclusion and section 6 represented the references.

II. LITERATURE REVIEW

This section consist the literature review on DBSCAN [5], the first density based clustering technique. In DBSCAN clusters develops according to a density based connectivity analysis. In 2004, El-Sonbaty et al. [12] proposed an improved version of DBSCAN, which generate efficient clusters from spatial datasets using dataset partitioning as a pre-processing stage. The number of dataset scans and buffer size space is reduced thus improving the performance and become memory efficient [1].The most important advantage is that it is scalable and limitation is that the results are not evaluated on real data sets. In 2006, Liu et al. [6] proposed a fast density based clustering technique to minimize the time complexity and to maximize the quality of clusters. In FDBSCAN the dataset objects are sorted by certain dimensional co-ordinates. In improved DBSCAN algorithm, global Eps parameter is used. Few or single cluster consisting all object is formed when the range of Eps is small and if the range of Eps is high many small cluster are generated. It is time efficient algorithm; as it

decreases the time by ignoring the region objects already clustered. In 2012 Patwary et al. [7] proposed a new scalable parallel DBSCAN algorithm using graph algorithmic concepts. To construct clusters, a tree based bottom-up approach is used. The disjoint-set data structure is used to break the data access order and to perform the merging efficiently. In disjoint-set data structure, two main operations are used: FIND and UNION. This merging is performed using master-slave approach where master performs merging sequentially. The important advantage is that use of master salve method which is helps to speed up the process. The main limitation of this process is it will increase the I/O load and effects of cost is exits on it.In 2012 C.Havens et al. [8] has compared the three different technique based on efficiency to extend fuzzy c-means (FCM) clustering. In this comparison based on sampling, incremental technique and kernelized version of FCM that provide approximations based on sampling, including three proposed algorithms are done. syntactic dataset is used to conduct the numerical experiment that facilitate comparisons based time and space complexity, speed, quality of approximations to batch FCM, and assessment of matches between partitions and ground truth. In 2012 Glory H Shah [4] proposed clustering algorithm to detect cluster that exists within a cluster. He has detected the problem of clustering, in which clusters are of different size, density and shape. They evaluated the result by describing parameters such as number of clusters, unclustered instances as well as incorrectly clustered instances. For experimental work, they used five different datasets to evaluate the result.

III. PROPOSED ALGORITHM

In this section, DBSCAN (Density based Spatial Clustering of Applications with Noise) is designed to find out the spatial data cluster with noise. DBSCAN uses two user's specified parameters i.e. MinPts and Eps. The DBSCAN Algorithm is implemented in matlab which helps for formation of cluster using numerical data in form of matrix representation and handle the massive datasets along with multiple dimensions. The implemented algorithm works efficient with creating the high density cluster and discover cluster of any arbitrary shape in spatial database with noise.

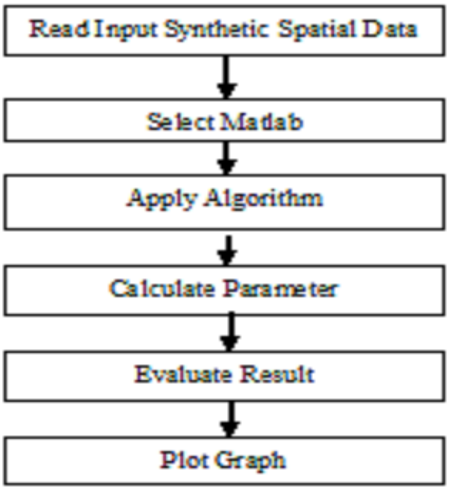

Fig.1. Flowchart for proposed algorithm

A.Input synthetic spatial dataset

We have used real and synthetic spatial dataset. There is total number of 5000 instances with 100 attribute in which 14 attributes are used to evaluate the result. All these instances are splitted into different dataset. The real dataset has been downloaded from UCI Machine Learning Repository Site.

B.Select matlab

After read the dataset the next step of our proposed algorithm is to run the dataset through the MATLAB and write the code for developing the script file for DBSCAN function used in Algorithm.

C.Apply DBSCAN algorithm

When all the dataset is loaded in MATLAB tool, then apply the DBSCAN algorithm on different dataset which configures the parameter value.

D. Calculate parameters

Before apply DBSCAN algorithm on the datasets user should configure the parameters value which will provide the different result as compare to default value of parameter.

E. Evaluate performance

In this step performance is evaluated and the result of dataset in the form of number of cluster formed, un-clustered instances, incorrectly clustered instances, time measure and analysis of noise with help of Eps and MinPt are compared with different datasets.

F.Plot graph

After result evaluation, performance is measure with graphical representation of different data sets.

IV. EXPERIMENTAL STUDIES

This section followed with various types of datasets with various numbers of points. All datasets load in DBSCAN algorithm and run on MatLab. The MatLab software is comparable to work for algorithm design and analysis. The detailed description of all the data set and detailed description of all parameters are given as shown in table. Mostly analysis done with numerical datasets and cluster pattern is evaluated. The detailed description of datasets is given as shown in table below.

Table 1: Description of data set

Data Set Type	Instances Value	Attribute Value	Attribute type	Run Time Efficiency
Multivariate	5000	14	Categorical, Integer, Numerical Data	12.093
Multivariate Matrix distribution	2000	2	Integer, Numerical Data	4.035
Normal Random distribution	1000	5	Integer, Numerical Data	2.451

In this paper three types of datasets i.e.Multivariate, Multivariate Matrix distribution, Normal Random distribution are compared and runtime efficiency is calculated using MatLab. According to experimental results run time efficiency of normal random distribution dataset is better than all.

V. PERFORMANCE EVALUATION

In this paper, different datasets are normalized and used with DBSCAN algorithm. After applying DBSCAN algorithm on different datasets performance of each dataset is evaluated. In fig.2, 5000 clusters are formulated using DBSCAN algorithm. It is found that proposed algorithm contains more clusters than normal DBSCAN. This concludes that the proposed algorithm can analyse the clusters for large datasets effectively.

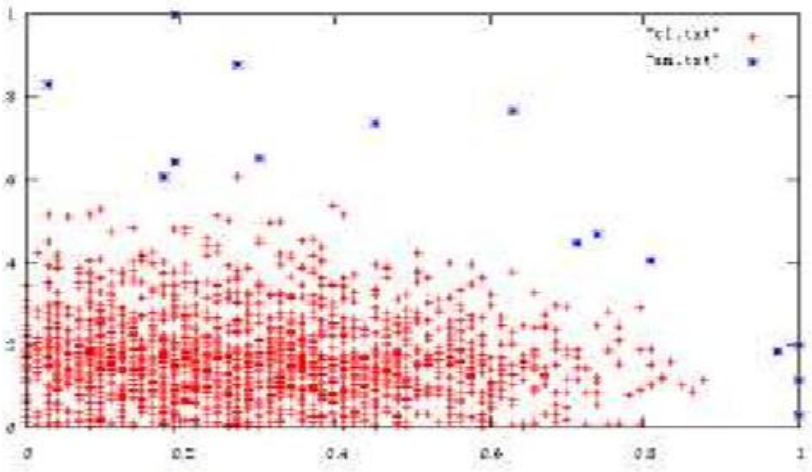

Fig.2.Formulation of 5000 cluster using DBSCAN algorithm

Then multivariate random clusters using DBSCAN algorithm are generated(figure 3).After multivariate random cluster generation, existance of noise is found out and eliminatedas shown in figure 4.In figure 5 formatted clusters are shown with the same value of input parameter which was taken in earlier implementation.

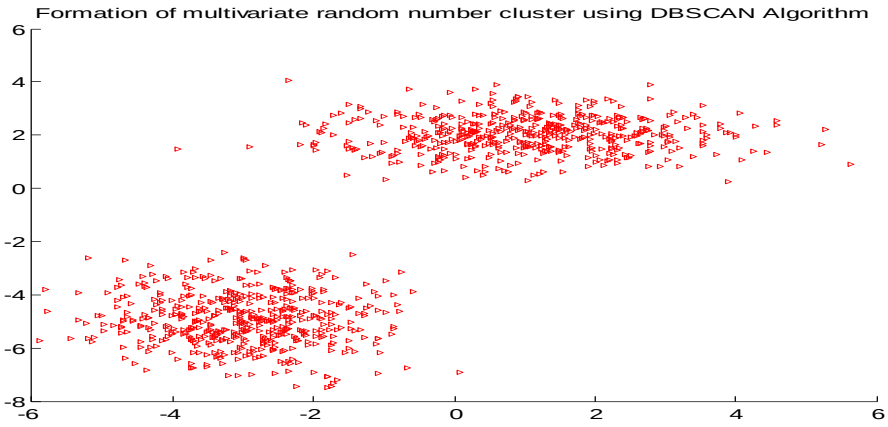

Fig.3. Multivariate random cluster using DBSCAN

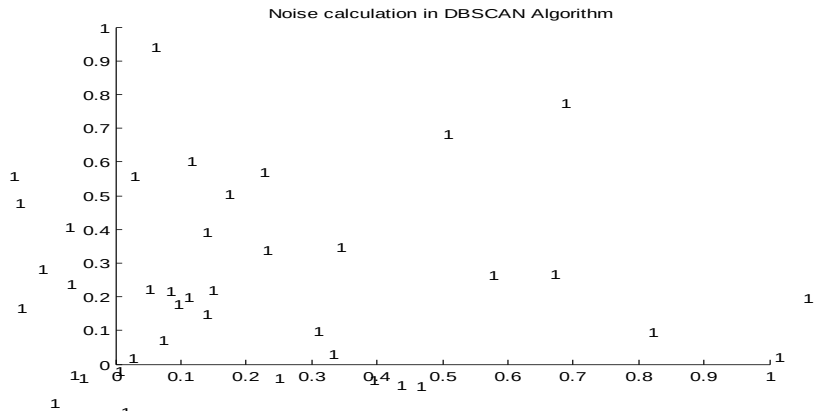

Fig.4. Elimination of noise from the cluster

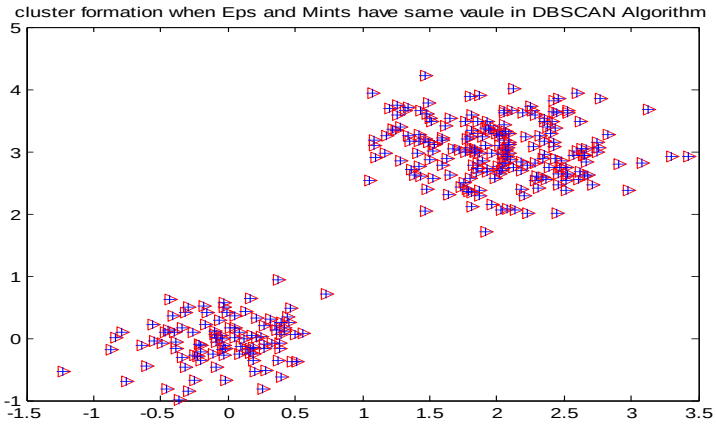

Fig.5. Cluster formation with same input parameter for EPs and Minpts in DBSCAN

VI. CONCLUSION

This paper concludes that total number of 7000 instances with 21 attributes is divided into different real and synthetic types of dataset before forming the clusters. Real dataset are downloaded from UCI site. Further synthetic dataset is divided into multivariate matrix distribution and normal random distribution .From the experimental results and analysis, it is concluded that the proposed algorithm is scalable as instead of working as a whole dataset it works on splitting the dataset. The proposed algorithm can analyze the cluster for large dataset effectively. The runtime efficiency of normal random distribution is better than multivariate matrix distribution which in turn is better than multivariate dataset.

VII. REFERENCES

[1] Chandra. E, Anuradha. V. P, A Survey on Clustering Algorithms for Data in Spatial Database Management System, International Journal of Computer Applications, Col. 24, June 2011.

[2] G. Karypis, E. H. Hanand, V. Kumar, "Chameleon: Hierarchical Clustering using Dynamic Modelling," Computer, Aug 1999.

[3] M. Parimala, D. Lopez, N. C. Senthilkumar, A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases, International Journal of Advanced Science and Technology, Vol. 31, June 2011.

[4] Glory H.Shah, "An Improved DBSCAN, "A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets"IEEE 2013.

[5] RashiChauhan,"A survey of density based clustering algorithm"IJCST vol.5 issue 2,April 2014.

[6] Bing Liu, "A Fast Density Based Clustering Algorithm For Large Databases", In: Proc. of IEEE Fifth International Conference on Machine Learning and Cybernetics, Dalian, August 2006.

[7] Md. Mostofa Ali Patwary, Diana Palsetia, Ankit Agrawal "A New Scalable Parallel DBSCAN Algorithm Using Disjoint-Set Data Structure", In: Proc. of IEEE International Conference, Salt Lake City,utah, USA November 2012.

[8] Timothy C. Havens, Senior Member, IEEE, James C. Bezdek,"Fuzzy c-Means Algorithms for Very Large Data", IEEE Transactions on Fuzzy systems,Vol.20,No.6N December 2012 Data", IEEE Transactions on Fuzzy systems, Vol. 20, No. 6, December 2012.

[9] Chetan Dharni, MeenakshiBnasal" An improvement of DBSCAN Algorithm to analyse Cluster for Large Datasets"IEEE 2013.

[10] Sonamdeep kaur,Sarika,"A survey: Clustering algorithm in data mining"IJCA. Cognition 2015.

[11] Zeng Donghai. The Study of Clustering Algorithm Based on Grid- Density and Spatial Partition Tree. XiaMen University, PRC, 2006.

[12] Yasser El-Sonbaty, M. A. Ismail, Mohamed Farouk "An Efficient Density Based Clustering Algorithm for Large Databases", IEEE, ICTAI 2004.