IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases

K. Nafees Ahmed[1], T. Abdul Razak[2]

Research Scholar, Department of Computer Science, Jamal Mohamed College, Tiruchirappalli, India[1]

Associate Professor, Department of Computer Science, Jamal Mohamed College, Tiruchirappalli, India[2]

Abstract: Clustering is an important data mining method for knowledge discovery in large databases. It is an exploratory data analysis tool which aims at categorizing different objects into groups (clusters) in such a way that the degree of association between two objects is high if they belong to the same group and low otherwise. Discovering clusters in spatial data is a challenging one because of its complexity nature. The clusters in spatial data are of different sizes, shapes and densities, and also contain noise and outliers. Different clustering techniques have been proposed for knowledge discovery from spatial databases. DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is a traditional and well known density-based clustering method. It defines a cluster as a maximal set of density-connected points. It can detect clusters of arbitrary shapes and filter out noise effectively. The clusters which are formed based on density are easy understandable and it does not limit the shapes. Besides its popularity, DBSCAN needs some improvements for better results. In this paper, we have analyzed and presented various significant enhancements of DBSCAN algorithm for our evaluation.

Keywords: Clustering, Cluster Analysis, Density based Clustering, DBSCAN, Spatial Clustering, Spatial Data.

I. INTRODUCTION

In recent years, due to the advancement of internet applications and widespread use of smart phones, huge amounts of spatial data have been generated from different sources. Earlier, the data mining and Geographic Information Systems (GIS) have presented as two different technologies, each with its own traditions and approaches for data analysis and visualization. Recently, the task of integrating these two technologies has become extremely importance. Spatial data mining refers to applying data mining techniques in spatial data, with an aim to find patterns corresponds to location. Finding patterns from spatial dataset is generally more complex than traditional datasets [1].

Spatial clustering is an important component of spatial data mining, a process of grouping set of objects in certain dimensional space into clusters such that the objects are highly similar in the same cluster and are dissimilar in other clusters. It plays a vital role in several applications such as crime analysis, population genetics, landscape ecology, spatial epidemiology, disease surveillance, geo-marketing, image exploration, social analysis and so on [2]. Finding clusters in data is difficult when the clusters are of different sizes, shapes, and densities. Although many algorithms exist, the density based clustering is more suitable for such clusters. The clusters which we got based on density are easy to understand and it does not limit the shapes of clusters. Density based methods are based on separating regions of high density (cluster) from that of low dense regions (noise).

The rest of this paper is organized as follows. Section 2 provides insight of ten different density based clustering algorithms. Section 3 concludes with a comparison of these algorithms and some direction for future work.

II. DENSITY BASED CLUSTERING

There are different types of clustering methods have been developed namely partitioning, hierarchical, density, grid, model, and constraint based. Among these, the density based method works based on the notion of density. Here, the clusters are formed as thick regions which are apart from thin regions. The fundamental idea is that increasing the identified cluster until the density (number of objects) in the neighborhood goes beyond some threshold.

A. DBSCAN (Density Based Spatial Clustering of Applications with Noise)

DBSCAN is a basic density based clustering algorithm proposed by Martin et al. [3] to find out arbitrary shaped clusters as well as to distinguish noise from large spatial databases. It accepts two parameters namely Eps (radius) and MinPts (minimum points-a threshold). It is based on center-based approach, where the density is estimated for a particular point in the dataset by counting the number of points within a specified radius, Eps. This allows us to classify a point as a core point, a border point, a noise point. The main idea is that for each point of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of points (MinPts).

Algorithm Description

1. Choose a random point p
2. Fetch all points that are density-reachable from p with respect to Eps and MinPts

DOI 10.17148/IJARCCE.2016.5277

3. A cluster is formed, if p is a core point
4. Visit the next point of the dataset, if p is a border point and non of the points are density reachable from p
5. Repeat the above process until all of the points have been examined

B. VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise)

A new improved density-based algorithm called VDBSCAN introduced by Liu et al. [4] for the intention of clustering datasets with varied densities. The basic design is that chooses suitable parameters for different density using k-dist plot and apply DBSCAN for every chosen parameter.

Algorithm Description
1. Find out and stores k-dist for each object and divide k-dist plots
2. The number of densities is given by k-dist plot
3. Chooses parameters Epsi automatically for each density
4. Scan the dataset and cluster varied densities using corresponding Epsi
5. Display the valid cluster

C. LDBSCAN(A Local Density Based Spatial Clustering Algorithm with Noise)

Daun et al. [5] proposed this new algorithm to discover different local-density clusters from a spatial database. It introduces the concept of local outlier factor (LOF) and local reachability density (LRD) to find clusters and noises.

Algorithm Description
1. Select an arbitrary point p
2. Retrieve all points local density-reachable from p with respect to LOFUB, pct, and MinPts
3. If p is a core point, a cluster is formed
4. If p is not a core point, check the next point of the database
5. Repeat the process until all the points have been processed

D. ST-DBSCAN(An Algorithm for Clustering Spatial-Temporal Data)

Birant et al. [6] proposed this algorithm to find out clusters according to spatial, temporal and non-spatial values of the objects. It improves DBSCAN in three ways such that it clusters spatial-temporal data corresponds to non-spatial, spatial and temporal attributes, detect clusters of different densities, and solve the conflicts in border objects. It requires four parameters namely Eps1 (distance parameter for spatial attributes like latitude and longitude), Eps2 (distance parameter for non-spatial attributes), MinPts (minimum points) and Δε (a threshold value)

Algorithm Description
1. Choose a random point p
2. Recover all points that are density-reachable from p with respect to Eps1 and Eps2
3. If p is a core point, a cluster is formed
4. If p is a border object, no points are density-reachable from p and then visits next point of the database
5. Repeat the process until all the points have been processed

E. DVBSCAN (Density Variation Based Spatial Clustering of Applications with Noise)

It is an extension of DBSCAN algorithm proposed by Ram et al. [7] to deal the local density variation within the cluster. The authors introduce the concept of cluster density mean (CDM) and cluster density variance (CDV) for cluster expansion. It takes minimum objects (m), radius (ε) and threshold values (α, λ) as input parameters.

Algorithm Description
1. A cluster is formed by selecting core object
2. Calculate cluster density mean (CDM) for growing cluster before expansion
3. Calculate cluster density variance (CDV) includes e-neighborhood of unprocessed core object with respect to CDM
4. If CDV of growing cluster is less than a specified threshold value α and if the difference between the minimum and maximum objects lying in e-neighborhood is less than a threshold λ
5. Then, only unprocessed core object is allowed for expansion; otherwise the object is simply included in the cluster

F. DBSCAN-DLP (Multi-density DBSCAN Algorithm Based on Density Levels Partitioning)

Zhongyang Xiong et al. [8] proposed a multi-density clustering method called DBSCAN-DLP which attempts to generalize the typical DBSCAN to automatically discover clusters of different densities through the concept of density level partitioning. The essential idea of this algorithm is that partitioning the dataset into different density level sets by examining statistical characteristics of its density variation, then estimates Eps for each density level set, and finally adopts DBSCAN on each density level set with corresponding Eps to get resulting clusters.

Algorithm Description
1. Initialize all objects in the database as unlabelled and unprocessed
2. Compute distance matrix, distMat
3. Compute k nearest neighbor distances, kdistList
4. Sort kdistList
5. Calculate density variation values list, DenVarList
6. Computer density variation threshold
7. Partition the database into a list of density level sets, DLSList
8. Refine DLSList through removal and merging process
9. Estimate Eps for each DLS, EpsList
10. Apply DBSCAN on each DLSList with respect to EpsList
11. Combine all clusters found in each iteration to get final clusters

G. PACA-DBSCAN (Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm)

A new hybrid algorithm based on partitioning based DBSCAN and Ant clustering is proposed by G. Chaudhari

DOI 10.17148/IJARCCE.2016.5277

Chaitali [9]. It applies one of the two partitioning methods namely PD-based partitioning and PACA partitioning depends on the dimension of datasets. That is, if the dataset is 2D, it uses PD-based partitioning method to partition the data; otherwise if the dataset is multi-dimensional, then it uses PACA method. And for each partition, this algorithm constructs R*-tree, plots k-dist graph and runs DBSCAN. Finally, the partial clusters will be merged based on predetermined rules.

H. DMDBSCAN (A Dynamic Method for Discovering Density Varied Clusters)
Mohameed T. H. Elbatta et al. [10] have given this new algorithm for the purpose of varied density datasets analysis. The main idea is that it uses dynamic method to find suitable value of Eps for each density level of the dataset.

1. Calculate suitable values of parameters Eps for different levels of densities according to k-dist plot by using dynamic method
2. For each value of Eps, apply DBSCAN to find all cluster with respect to corresponding density level
3. Then, all points which are clustered are ignored
4. Repeat the process until all points have been processed

I. MDBSCAN (Modified DBSCAN Using MST based value for ε in DBSCAN)
Nirmalya Chowdhury et al. [11] presented a modified version of DBSCAN algorithm for clustering datasets using minimum spanning tree (MST) based objective function and to discover natural grouping. A threshold based on MST of data points of each cluster thus obtained is used to remove noise from the final clustering.

1. Let D={X_1,$X_{2,}$ … … … , X_n} € R^m(m>2), where initially all data objects are marked as unvisited
2. Randomly choose an unvisited object p and mark as visited and verify whether the Eps-neighborhood of p contains at least MinPts objects
3. If not, p€NS, where NS is a noise set
4. Otherwise, all objects in all Eps-neighborhood of p are marked as visited
5. A new cluster C={y_1,y_2, … … … , y_k, p} € R^m, is created for p if the objects in all Eps-neighborhood of p do not overlap with any previous clusters
6. Else objects in Eps-neighborhood of p are merged with previously formed cluster
7. Objects in the noise set NS are excluded from the set if they belong in Eps-neighborhood of an unvisited core point
8. Repeat the steps from 2 to 7 until all objects are visited
9. Finally the objects that remains in the noise set are termed as noise points

J. MR-DBSCAN (A scalable MapReduce based DBSCAN algorithm for heavily skewed data)
An efficient scalable DBSCAN algorithm using MapReduce is presented by Yaobin HE et al. [12] with an aim of load balancing in large-scale datasets and efficient speed-up and scale-up for skewed big data. It works under three level namely data partitioning, local clustering, and global merging. In the first level, dataset is divided into smaller partitions based on spatial proximity. During second level, each partition is clustered independently. Then at the final level, the partial clustering results are combined to produce the global clusters.

III. CONCLUSION AND FUTURE WORK

In this study, we have presented a short review about recent improvements of DBSCAN algorithm like VDBSCAN, LDBSCAN, ST-DBSCAN, DVBSCAN, DBSCAN-DLP, PACA-DBSCAN, DMDBSCAN, MDBSCAN and MR-DBSCAN. Each algorithm has its own features. A comparative study in terms of merits of these algorithms is given in the Table 1.

TABLE I COMPARISON OF DIFFERENT DENSITY BASED ALGORITHMS

Algorithm Name	**Merits**
DBSCAN	Finding clusters of arbitrary shapes and handling noise effectively.
VDBSCAN	Identify clusters with varied density.
LDBSCAN	Finding similar local density clusters and detects noises effectively.
ST-DBSCAN	Discovers clusters on spatial-temporal data.
DVBSCAN	Finding clusters of varied density.
DBSCAN-DLP	Discovers clusters of different densities.
PACA-DBSCAN	Discovers clusters from multidimensional datasets.
DMDBSCAN	Discovers clusters of different density levels.
MDBSCAN	Discovers natural clusters based on MST objective function.
MR-DBSCAN	Finding clusters on heavily skewed data.

The aim of these variations is to enhance the DBSCAN algorithm for better clustering result. The future work can be focused on to improve the efficiency and reduce the time complexity of the algorithm by applying the concept of Metaheuristic technique like Particle Swarm Optimization (PSO).

REFERENCES

[1] Shashi Shekhar, Pusheng Zhang, Yan Huang and Ranga Raju Vatsavai, “Research accomplishmets and issues on spatial data mining,” 2003.

DOI 10.17148/IJARCCE.2016.5277

[2] S.E Spielman and J.C Thill, "Social area analysis, data mining and GIS," Computers Environment and Urban Systems, vol. 32, pp. 110-122, 2008.

[3] M. Ester, H-P. Kriegel, J. Sander, and X. Xu, "A Density-based algorithm for discovering clusters in large spatial databases with noise," in Proc of 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96), 1996.

[4] P. Liu, D. Zhou, and N. Wu, "Varied density based spatial clustering of applications with noise," in Proc of IEEE Conference (ICSSSM-07), pp. 528-531, 2007.

[5] L. Duan, L. Xu, F. Guo, J. Lee and B. Yan, "A local-density based spatial clustering algorithm with noise," Information Systems, vol. 32, pp. 978-986, 2007.

[6] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," Data and Knowledge Engineering, pp. 208-221, 2007.

[7] A. Ram, S. Jalal, Anand S. Jalal, and M. Kumar, "A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases," Int. Journal of Computer Applications (IJCA), vol. 3, no. 6, 2010.

[8] Zhongyang Xiong, Ruotian Chen, Yufang Zhang, and Xuan Zhang, "Multi-density DBSCAN Algorithm Based on Density Levels Partitioning," Journal of Information & Computational Science (JOICS), pp. 2739-2749, 2012.

[9] G. Chaudhari Chaitali, "Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm," Int. Journal of Engineering and Advanced Technology (IJEAT), vol. 2, no. 2, pp. 212-215, 2012.

[10] Mohammed T. H. Elbatta and Wesam M. Ashour, "A Dynamic Method for Discovering Density Varied Clusters," Int. Journal of Signal Processing, Image Processing, and Pattern Recognition, vol. 6, no. 1, pp. 123-134, 2013.

[11] Nirmalya Chowdhury and Preetha Bhattacharjee, "Using an MST-based Value for ε in DBSCAN Algorithm for Obtaining Better Result," Int. Journal of Information Technology and Computer Science, vol. 6, pp. 55-60, 2014.

[12] Yaobin HE, Haoyu TAN, Wuman LUO, Shengzhong FENG, and Jianping FAN, "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data," Research Article, Front. Computer Science, vol. 8, no. 1, pp. 83-99, 2014.

.

DOI 10.17148/IJARCCE.2016.5277