A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases

M.Parimala, Daphne Lopez, N.C. Senthilkumar
School of Information Technology& Engineering , VIT University,
Vellore 632 014, India
parimala.m@vit.ac.in, daphnelopez@vit.ac.in, ncsenthilkumar@vit.ac.in

Abstract

Density based clustering algorithm is one of the primary methods for clustering in data mining. The clusters which are formed based on the density are easy to understand and it does not limit itself to the shapes of clusters. This paper gives a detailed survey of the existing density based algorithms namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN and DBCLASD based on the essential parameters needed for a good clustering algorithm. We analyse the algorithms in terms of the parameters essential for creating meaningful clusters.

Keywords: *clustering; DBSCAN; VDBSCAN; DVBSCAN; ST-DBSCAN; DBCLASD*

1. Introduction

Data mining is a step in the Knowledge Discovery in Databases (KDD) process consisting of the application of data analysis and discovery of algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data [2].

Spatial Database Management (SDBS) [3] are database systems for the management of spatial data i.e., point objects or spatially extended in a 2D or 3D space or in some high dimensional vector space. Knowledge discovery becomes more and more important in spatial databases since increasingly large amounts of data obtained from satellite images, X-ray crystallography or other automatic equipment are stored in spatial databases.

Spatial Data Mining [4] is the process of discovering interesting and previously unknown but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding pattern from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships and spatial autocorrelation.

There is a rampant growth of spatial data and a number of needs arise as spatial data mining techniques, modeling semantic rich spatial properties such as topology, statistical interpretation models for spatial pattern, improving computational efficiency and model ,preprocessing spatial data and many others.

There are many techniques like classification, decision tree, fuzzy logic, neural networks applied for mining spatial data. Most of the recent work on spatial data has used various clustering techniques due to the nature of the data.

Clustering i.e., grouping the objects of a database into meaningful subclasses, is one of the major data mining methods [6]. Among many types of clustering algorithms density based algorithm is more efficient in detecting the clusters with varied density. There has been a lot of research on clustering algorithms for decades but the application to large spatial databases introduces the following requirements:

(i) Minimal number of input parameters. Because for large spatial databases it is very difficult to identify the initial parameters like number of clusters, shape and density in advance.
(ii) Discovery of clusters with arbitrary shape. Because the shape of clusters may be in any random shape.
(iii) Good efficiency should be achieved in very large databases.

2. Density-Based Algorithms for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN)

2.1. Introduction

DBSCAN [1] is a density based algorithm which discovers clusters with arbitrary shape and with minimal number of input parameters. The input parameters required for this algorithm is the radius of the cluster (Eps) and minimum points required inside the cluster (Minpts). The basic idea behind this DBSCAN [7] algorithm is as follows,

Definition 1: (Eps –neighborhood of a point) The Eps neighborhood of a point p, denoted by $N_{Eps}(p)$ is defined by

$$N_{Eps}(p) = \{p \in D \mid dist(p,q) \leq Eps\}$$

There are two kinds of points in the cluster, the points which is inside the cluster(core points), and points on the border of the cluster(border points).

Definition 2: (Directly density-reachable) A point p is directly density-reachable from a point q wrt. Eps, MinPts if

1) $p \in N_{Eps}(q)$ and

2) $|N_{Eps}(q)| \geq MinPts$ (core point condition).

Definition 3: (Density-reachable) A point p is density-reachable from a point q wrt. Eps and MinPts if there is a chain of points $p_1, ..., p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i.

Definition 4: (Density-connected) A point p is density-connected to a point q wrt. Eps and MinPts if there is a point o such that both, p and q are density-reachable from o wrt.Eps and MinPts.

Definition 5: (cluster) Let D be a database of points. A cluster C wrt. Eps and MinPts is a non-empty subset of D satisfying the following conditions:

1) $\forall$ p, q: if $p \in C$ and q is density-reachable from p wrt.Eps and MinPts, then $q \in C$. (Maximality)

2) $\forall p, q \in C$: p is density-connected to q wrt. Eps and MinPts (Connectivity)

Definition 6: (noise) Let $C_1, . . ., C_k$ be the clusters of the database D wrt. parameters Eps_i and $MinPts_i$, i = 1, . . ., k. The noise is defined as the set of points in the database D not belonging to any cluster C_i , i.e. noise = $\{p \in D \mid \forall i: p \notin C_i\}$.

2.2. Description of the Algorithm

In this section, the algorithm DBSCAN[7] Density Based Spatial Clustering of Applications with noise is designed to discover the spatial data clusters with noise . The steps involved in this algorithm are as follows,

(i) Select an arbitrary point p
(ii) Retrieve all points density-reachable from p w.r.t. Eps and Minpts.
(iii) If p is a core point, a cluster is formed.
(iv) If p is a border point, no points are density reachable from p and DBSCAN visits the next point of the database.
(v) Continue the process until all the points have been processed.

2.3 Impact of the Algorithm

DBSCAN requires two input parameters (Minimum points and radius) and supports the user in finding an approximate value for it using k-dist graph[7]. It discovers clusters of arbitrary shape. It holds good for large spatial databases.

2.4 Future Work

DBSCAN algorithm here considers [1] only point objects but it could be extended for other spatial objects like polygons. Applications of DBSCAN to high dimensional feature spaces should be investigated and radius generation for this high dimensional data also has to be explored. It also fails to detect clusters with varied density.

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

3.1. Introduction

The DBSCAN algorithm is not capable of finding out meaningful clusters with varied densities. VDBSCAN[9] algorithm detects cluster with varied density as well as automatically selects several values of input parameter Eps for different densities . Even the parameter k is automatically generated based on the characteristics of the datasets [8].

3.2. Description of the Algorithm

In general the algorithm has two steps, choosing parameters Eps_i and cluster with varied densities. The procedure for this algorithm [8] is as follows,

(i) It calculates and stores k-dist for each project and partition the k-dist plots.
(ii) The number of densities is given intuitively by k-dist plot.
(iii) The parameter Eps_i is selected automatically for each density.
(iv) Scan the dataset and cluster different densities using corresponding Eps_i
(v) Display the valid cluster with respect to varied density.

Algorithm:

```
1. Partition k-dist plot.
2. Give thresholds of parameters $Eps_i$(i=1,2,…..n)
3. For each $Eps_i$(i=1,2,…..n)
        a) Eps = $Eps_i$
```

b) Adopt DBSCAN algorithm for points that are not marked.
c) Mark points as c_i

4. Display all the marked points as corresponding clusters.

3.3. Impact of the Algorithm

The purpose of this algorithm is to find out meaningful clusters in databases with respect to widely varied densities. VDBSCAN has the same time complexity as DBSCAN and can identify clusters with different density which is not possible in DBSCAN algorithm. Even the input parameters (Eps) are automatically generated from the datasets.

3.4. Future Work

The behavior of parameter k in k-dist plot depends on the dataset. The consequence of the magnitude of parameter k for a particular dataset is one of the interesting challenges.

4. A Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Databases (DVBSCAN)

4.1. Introduction

DBSCAN [7] a pioneer density based clustering algorithm detects clusters with different shapes and sizes but fails to detect clusters with varied densities that exists within the cluster. DVBSCAN [10] algorithm handles local density variation within the cluster. The input parameters used in this algorithm are minimum objects(μ),radius, threshold values(α, λ).It calculates the growing cluster density mean and then the cluster density variance for any core object, which is supposed to be expanded further by considering density of its E-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 is also satisfying the cluster similarity index, then it will allow the core object for expansion.

4.2. Description of the Algorithm

(i) A cluster is formed by selecting core object.
(ii) Then it computes cluster density mean (CDM) is calculated for the growing cluster before allowing the expansion of an unprocessed core object.
(iii) Computation of the cluster Density variance (CDV) includes the E-neighborhood of the unprocessed core object with respect to CDM.
(iv) If CDV of growing cluster with respect to CDM is less than a specified threshold value α and the difference between the minimum and maximum object lying in the e-neighborhood of the object is less than a specified threshold value λ then only an unprocessed core object is allowed for expansion.
(v) Otherwise the object is simply added into the cluster.

4.3. Impact of the Algorithm

The DVBSCAN is able to handle the density variations that exist within the cluster. The clusters detected by this algorithm are having considerable density variation within the clusters. The detected clusters are not only separated by the sparse region but also separated by the regions having the density variation. It outperforms the DBSCAN, especially in case of local density[10]. as shown in Figure 1 and Figure 2.This algorithm finds the clusters that represent relatively uniform regions without being separated by sparse regions. The parameters α and λ are used to limit the amount of allowed local density variations within the cluster.

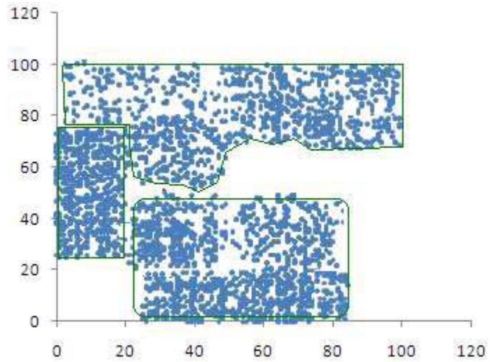


Figure 1. Clusters Generated by DBSCAN Algorithm

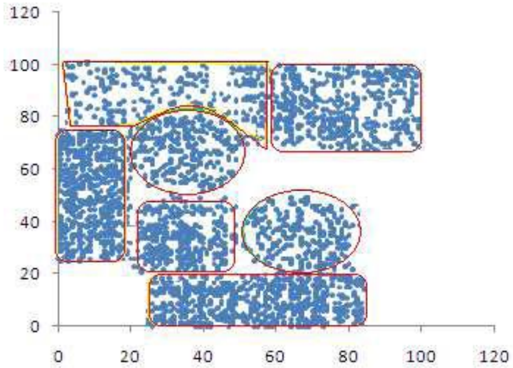


Figure 2. Clusters Generated by DVBSCAN Algorithm

4.4. Future Work:

As the time complexity is high it could be reduced. The input parameters can be determined automatically for better clustering.

5. A Distribution- Based clustering Algorithm for Mining Large Spatial Databases (DBCLASD)

5.1. Introduction

This new clustering Algorithm DBCLASD [11] detects clusters with arbitrary shape and it does not require any input parameters. The efficiency of DBCLASD on large spatial databases is also very attractive.

5.2. Description of the Algorithm

(i) DBCLASD is an incremental algorithm that is the assignment of a point to a cluster is based only on the points processed so far without considering the whole database.

(ii) It incrementally augments an initial cluster by its neighboring points as long as the nearest neighbor distance of the resulting cluster fits the expected distance distribution.

(iii) A set of candidates of a cluster is constructed using region queries which is supported by spatial Access Methods(SAM). The calculation of m is based on the model of uniformly distributed points inside the cluster C. Let A be the area of C and N be the number of its elements. A necessary condition for m is

$$N \times P(NNdist_C (P) > m) < 1$$

When inserting a new point p into cluster C, a circle query with center P and radius m is performed and the resulting points are considered as new candidates.

(iv) The incremental approach implies an inherent dependency of the discovering clusters from the order of generating and testing candidates. The order of testing the candidates is crucial. Candidates which are not accepted by the test for the first time are called unsuccessful candidates. To minimize the dependency on order of testing, the following two features are considered,

a) Unsuccessful candidates are not discarded but they are tried again later.
b) Points already assigned to some cluster may switch to another cluster later.

The testing of candidates are performed in two steps are as follows,

a) The current cluster is augmented by the candidate
b) Chi-square test is used to verify the hypothesis that the nearest neighbor distance set of the augmented cluster still fits the expected distance distribution.

5.3. Impact of the Algorithm

DBCLASD [6] Algorithm is based on the assumption that the points inside a cluster are uniformly distributed. The application of DBCLASD to earthquake catalogues shows that it also works effectively on real databases where the data is not exactly uniformly distributed. It is very efficient for large spatial databases. This algorithm fulfills all the requirements needed for designing a good clustering algorithm for spatial databases.

5.4. Future work

The existing algorithm is suitable for uniform distribution of points and can be extended to non-uniform points.

6. Spatial- Temporal Density Based Clustering (ST-DBSCAN)

6.1. Introduction

ST-DBSCAN algorithm is constructed by modifying DBSCAN [7] algorithm. In contrast to existing density-based clustering algorithm, ST-DBSCAN [12] algorithm has the ability of discovering clusters with respect to non-spatial, spatial and temporal values of the objects. The three modifications done in DBSCAN algorithm are as follows,

(i) ST-DBSCAN algorithm can cluster spatial-temporal data according to non-spatial, spatial and temporal attributes.
(ii) DBSCAN does not detect noise points when it is of varied density but this algorithm overcomes this problem by assigning density factor to each cluster.
(iii) In order to solve the conflicts in border objects it compare the average value of a cluster with new coming value.

6.2. Description of the Algorithm

The algorithm starts with the first point p in database D.

(i) This point p is processed according to DBSCAN algorithm and next point is taken.

(ii) Retrieve_Neighbors(object,Ep1,Ep2) function retrieves all objects density-reachable from the selected object with respect to Eps1,Eps2 and Minpts. If the returned points in Eps-neighborhood are smaller than Minpts input, the object is assigned as noise.
(iii) The points marked as noise can be changed later that is the points are not directly density-reachable but they will be density reachable.
(iv) If the selected point is a core object, then a new cluster is constructed. Then all the directly-density reachable neighbors of this core objects is also included.
(v) Then the algorithm iteratively collects density-reachable objects from the core object using stack.
(vi) If the object is not marked as noise or it is not in a cluster and the difference between the average value of the cluster and new value is smaller than Δ_{E} , it is placed into the current cluster.
(vii) If two clusters C1 and C2 are very close to each other, a point p may belong to both C1 and C2. Then point p is assigned to cluster which discovered first.

6.3. Impact of Algorithm

Spatial-temporal data refers to data which is stored as temporal slices of the spatial dataset. The knowledge discovery in spatial-temporal data is complex than non-spatial and temporal data. So this algorithm ST_DBSCAN [12] can be used in many applications such as geographic information systems, medical imaging and weather forecasting.

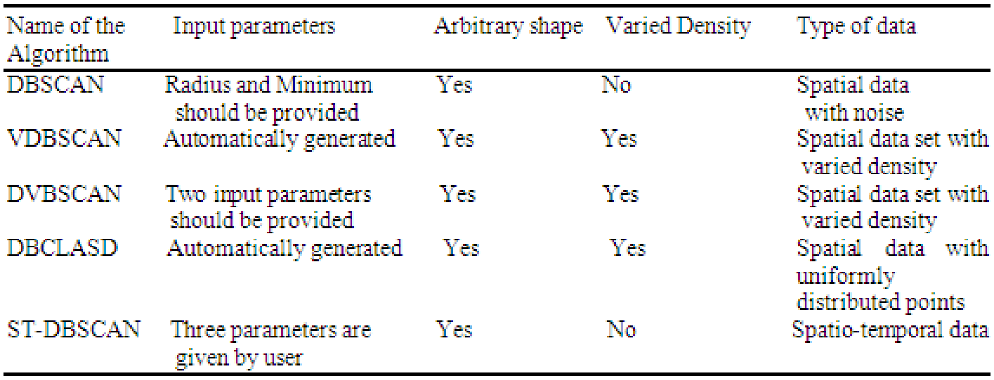
6.4. Future work

The input parameter has to be automatically generated. The performance of the algorithm also has to be improved.

7. Conclusion

This paper gives a detailed survey of five density based clustering algorithm like DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN and DBCLASD based on the essential requirements required for any clustering algorithm[11] in spatial data. Each algorithm is unique with its own features. A comparative study in terms of the input parameters, shapes of the cluster, density and the type of the data is given in Table 1.

Table 1. Comparison of Five Density Based Algorithms



Name of the Algorithm	Input parameters	Arbitrary shape	Varied Density	Type of data
DBSCAN	Radius and Minimum should be provided	Yes	No	Spatial data with noise
VDBSCAN	Automatically generated	Yes	Yes	Spatial data set with varied density
DVBSCAN	Two input parameters should be provided	Yes	Yes	Spatial data set with varied density
DBCLASD	Automatically generated	Yes	Yes	Spatial data with uniformly distributed points
ST-DBSCAN	Three parameters are given by user	Yes	No	Spatio-temporal data

References:

[1] Han J. Kamber,2001, “Data Mining : Concepts & Techniques”

[2] Fayyad U., Piatetsky-Shapiro G., and Smyth P. 1996. *“*Knowledge Discovery and Data Mining: Towards a Unifying Framework”. *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR,82-88.*

[3] Guting, ”An Introduction to Spatial Database Systems”, *VLDB 1994*

[4] Shashi Shekar, Pusheng Zhang, Ranga Raju Vatsavai , “Research Accomplishments and Issues on Spatial Data Mining”

[5] Shashi Shekar & Sanjay Chawla, “Spatial Databases a Tour”

[6] Matheus C.J., Chan P.K., and Piatetsky-Shapiro G. 1993. “Systems for Knowledge Discovery in Databases*”. IEEE Transactions on Knowledge and Data Engineering 5(6): 903-913.*

[7] Martin Ester,Han-peter Kriegel,Jorg Sander, Xiaowei Xu,”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, *2nd International conference on Knowledge Discovery and Data Mining (KDD-96)*

[8] A.K.M Rasheduzzaman Chowdhury, Md.Asikur Rahman, “An efficient Mehtod for subjectively choosing parameter k automatically in VDBSCAN”,*proceedings of ICCAE 2010 IEEE ,Vol 1,pg 38-41.*

[9] Peng Liu, Dong Zhou, Naijun Wu,” Varied Density Based Spatial Clustering of Application with Noise”, in *proceedings of IEEE Conference ICSSSM 2007 pg 528-531.*

[10] Anant Ram, Sunita Jalal, Anand S. Jalal, Manoj kumar, “A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases*”, International Journal of Computer Application Volume 3,No.6, June 2010.*

[11] Xiaowei Xu, Martin Ester, Hans-Peter Kriegal, Jorg Sabder, “ A Distribution Based Clustering Algorithm for Mining in Large Spatial Databases*” ICDE-98.*

[12] Derya Birant, Alp Kut, ”ST-DBSCAN: An Algorithm for Clustering Spatial-temporal data” *Data and Knowledge Engineering 2007 pg 208-221.*

Authors

Parimala is currently working as Assistant Professor at VIT University, India. She is an active researcher in the field of Spatial Databases. Her research interests include spatial data mining.

Daphne Lopez, Associate Professor and Division Leader at VIT University, India has vast experience in teaching, research and industry. To her credit there are a good number of publications in international conferences and journals.

N. C. Senthilkumar graduated from VIT University and is working as Assistant Professor(Senior) at VIT University, India. His research interests include dimensionality reduction of spatial data.