A Survey on Study of Enhanced DBSCAN Algorithm

Tanu Verma, Dr. Deepti Gaur
ITM University, Gurgaon, India

Abstract

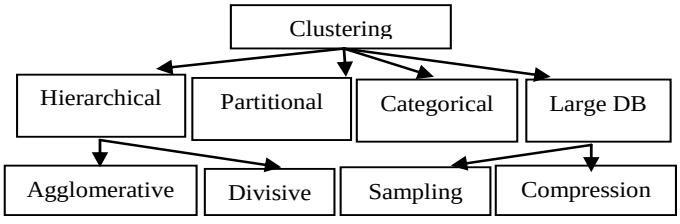
In Data minting, clustering plays a very important role. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Density Based Clustering is a well-known clustering algorithm which having advantages for finding out the clusters of different shapes and size from a large amount of data that contains noise and outliers. In this paper, I have presented a study of DBSCAN algorithm and its further enhancement based on the varied densities.

Index Terms—**Clustering, Density Based Clustering, DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN, VMDBSCAN, DMDBSCAN, VDDBSCAN**

1. Introduction

1.1. Cluster Definition and Types

Clustering is defined as the process of grouping of similar data or objects (either physical or abstract) into classes of similar objects according to characteristics found in the actual data [1]. Clustering has been useful in various application domains like biology, medicine, anthropology, marketing and economics. Clustering application include plant and animal classification, disease classification, pattern recognition, image processing, document retrieval and examining Web log data[2]. Basic requirement of good clustering algorithm are their ability to work with high dimensional patterns, scalability with large data sets, ability to find clusters with irregular shapes, ability to detect noisy outliers and should work in one scan or less data, time complexity, high dimensionality. Clustering methods can be categorized into two main types: fuzzy clustering and hard clustering and various different techniques are used to grouped data points is shown in figure 1:



" Figure 1. Clustering techniques"

1.2. Hierarchical Algorithm

Hierarchical clustering which creates sets of clusters, use the concept of "Dendrogram" **(**a tree data structure: tree of clusters) which is used to build cluster hierarchy. The root in a dendrogram tree contains one cluster where all elements are together. With the help of Agglomerative (merging) and Divisive (Dividing) approach dendrogram can be created. A merging or agglomerative clustering is a bottom up approach pairs of clusters are merged as one moves up the hierarchy [2]. A divisive clustering is a top down approach in which observations start in one cluster, and splits are recursively performed as one moves down the hierarchy. The process continues until a stopping criterion is achieved. Single link, complete link and average link techniques are perhaps the most well known agglomerative techniques based on well-known graph theory concepts. Also only spherical clusters can be obtained. The hierarchical algorithm's advantage includes versatile nature, it embedded flexibility with respect to a level of granularity and its integration with various validation indices, which can be defined on the clusters, and disadvantage is deriving appropriate parameters for the terminate conditions , object cannot be modified after its assignment to cluster. . The popular hierarchical clustering methods are BIRCH [3] and CURE [4], ROCK [5].

1.3. Partitional Algorithm

Nonhierarchical or partitional clustering creates the clusters in one step as opposed to several steps. Partitional clustering algorithms obtain a single partition of the data instead of a clustering structure(e.g dendrogram) produced by a hierarchical technique. The most popular partition-based clustering algorithms are the k-mean, Squared error, Nearest neighbour and PAM. The advantage of the partition-based algorithms is the use an iterative way to create the clusters, but its limitation is the number of clusters that has to be determined by user and only spherical shapes can be determined as clusters. Another limitation is that they suffer from a combinatorial explosion due to the number of possible solution. In K-means iterative clustering algorithm, items are moved among sets of clusters until the desired set is reached. Although the K-means algorithm often produces good results, it is not

time-efficient and does scale well. The squared error clustering algorithm minimizes the squared error. For each iteration in the squared error algorithm, each tuple is assigned to the cluster with closest centre. The PAM (partitioning around medoids) algorithm, also called the K-medoids algorithm, represents a cluster by a medoid.PAM does not scale well to large datasets because of its computational complexity. CLARA (Clustering Large Applications) improves on the time complexity of PAM by using samples of the dataset. CLARANS (Clustering large applications based upon randomized search) improves on CLARA by using multiple different samples[6]. CLARANS is shown to be more efficient than either PAM or CLARA for any size dataset. CLARANS assumes that all data are in main memory. This certainly is not a valid assumption for large database.

1.4. Density-based algorithms

DBSCAN [6] and OPTICS [7] find the core objects at first and grow the clusters based on these cores and search for objects that are in a neighborhood within a radius of a given object. In DBSCAN algorithm, first the number of objects presents within the neighbor region (Eps) is computed then the threshold value is compared with the neighbor objects counts, if counts appear less than threshold value then object is treated with noise. Else the new cluster that is formed from the core object by finding the group of density connected objects that are maximal w.r.t density reachability. OPTICS algorithm is an improvement of DBSCAN algorithm to deal with variance density clusters. OPTICS computes an ordering of the objects based on the reachability distance for representing the intrinsic hierarchical clustering structure. The advantage of these types of algorithms is that they can detect arbitrary forms of clusters and they can filter out the noise.

1.5. Grid based algorithm

Grid-based algorithms quantize the object space into a finite number of cells (hyper-rectangles) or grid and clustering is performed on grid instead of data object like in hierarchical and partitional. Since entire process accomplish at once for the calculation of statistical values of grid , it give fast processing time and good performance of clustering which depends only on grid not data objects.. The well known grid-based algorithms are STING [8], WaveCluster [9], and CLIQUE [10].

1.6. Model based algorithm

The assumption of model based algorithms is based on model which assumes that data is generated by model and original model is generated by data and model parameter can be categorized as partitional or hierarchical, which depend on the structure or model they hypothesize about the data set. Basically it provides a framework for incorporate knowledge about a domain and Expectation-Maximization (EM) algorithm [11] is most commonly used for clustering . The disadvantage of EM is that it get stuck in local optima if the seeds are not chosen well also EM algorithm lacks in computational efficiency. The common induction methods used in this algorithm is decision trees and neural networks. The common algorithm uses for this method are COBWEB and SOM (Self organizing Map). SOM is used for vector quantization and speech recognition.

1.7. Fuzzy algorithms

Fuzzy algorithms suggest soft clustering schema and it suppose that no hard clusters exist on the set of objects, but only one object can be assigned to more than one cluster. The known fuzzy clustering algorithm is FCM (Fuzzy C-MEANS) [12].FCM is considered better than harder K-means algorithm but it still converge to local minima of the squared error criterion.

The different clustering algorithms discussed are compared in Table 1. Comparison parameters include space, time complexity and general notes concerning applicability

Algorithm	Type	Space Complexity	Time Complexity	Notes
Single Link	Hierarchical	$O(n^2)$	$O(\mathbf{k}n^2)$	Not incremental
Average Link	Hierarchical	$O(n^2)$	$O(\mathbf{k}n^2)$	Not incremental
Complete Link	Hierarchical	$O(n^2)$	$O(\mathbf{k}n^2)$	Not incremental
Squared Error	Partitional	$O(n)$	$O(\mathbf{tk}n)$	Iterative
K-Means	Partitional	$O(n)$	$O(\mathbf{tk}n)$	Iterative
PAM	Partitional	$O(n^2)$	$O(tk(n-k)^2)$	Iterative; Adapted agglomerative ; Outliers
BIRCH	Partitional	O(n)	O(n)	CF-tree;Incremental
CURE	Hierarchical	O(n)	O(n)	Heap, k-D tree; Incremental ; Outliers
ROCK	Agglomerativ	$O(n^2)$	$O(n^2 \lg n)$	Sampling; Categorical;

	e			Links
DBSCAN	Mixed	$O(n^2)$	$O(n^2)$	Sampling; Outliers
CLARANS	Partitional	O(n)	O(n)	Iterative

"Table 1: Comparison between algorithms"

This paper is organized as follows: In the next section, a DBSCAN algorithm is explained in details. Section 3 covers the survey of enhanced DBSCAN algorithms and finally section 4 concludes the paper.

2. DBSCAN Algorithm

DBSCAN [6] is a density based algorithm which discovers clusters with arbitrary shape and with minimal number of input parameters. It considers regions with sufficiently high density as clusters of arbitrary shape in spatial databases (with noise). The input parameters required for this algorithm is the radius of the cluster (Eps) and minimum points required inside the cluster (Minpts).

2.1. Definition of DBSCAN Algorithm

It defines a cluster as a maximal set of density-connected points. Some basic definitions in DBSCAN are as follows:

Definition 1: The Eps (neighborhood of a point p), denoted by NEps(p) is defined by NEps(p) = {p ∈ D| dist(p,q)≤ 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: A point p is directly density-reachable from a point q wrt. Eps, MinPts if 1) p ∈ NEps(q) and 2) |NEps(q)| ≥MinPts (core point condition).

Definition 3: (Density-reachable) A point p is density-reachable from a point q wrt. Eps and MinPts if there exist a chain of points p1, ..., pn, p1 = q, pn = p such that pi+1 is directly density-reachable from pi.

Definition 4: (Density-connected object)

A object p is density-connected to a object q wrt. Eps and MinPts if there is a object o such that both, objects p and q are density-reachable from object o wrt.Eps and MinPts.

Definition 5: (cluster) Let D be a database of objects. A cluster C wrt. Eps and MinObjs is a non-empty subset of D satisfying the following conditions: 1) ∀ p, q: if p ∈ C and q is density-reachable from p wrt.Eps and MinObjs, then q ∈ C. (Maximality) 2) ∀ p, q ∈ C: p is density-connected to q wrt. Eps and MinObjs (Connectivity)

Definition 6: (noise) Let C1 ,. . ., Ck be the clusters of the database D wrt. parameters Epsi and MinPtsi, i = 1, . . ., k. The noise is defined as the set of points in the database D not belonging to any cluster Ci , i.e. noise = {p ∈ D|∀ i: p∉ Ci }

Definition 7: (border object) If an object is on the order of a cluster, then it is called a border object.

2.2. Detail in DBSCAN Algorithm

DBSCAN algorithm searches for clusters and outliers in the following steps:

(i) Select an arbitrary point p

(ii) Retrieve all points density-reachable from p w.r.t. Eps and Min

(iii) If p is a core point, a cluster is formed.

(iv) If there exist a border point p then 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. Disadvantage and improvement of DBSCAN Algorithm

As the first density-based clustering algorithm that discovers clusters with arbitrary shape and outliers, DBSCAN has certain limitations which are as follows:

(i) Within same database, when the number of samples is changed the two parameters Eps and MinPts have to be adjusted accordingly.
(ii) The Computational complexity of DBSCAN without any special structure is $O(n^2)$, where n is the number of database objects. If a spatial index is used, the complexity can be reduces to O (nlogn). However, the task of building a spatial index is time-consuming and less applicable to high dimensional data sets.

As the development of DBSCAN algorithm, there are several algorithm derived from DBSCAN algorithm aimed at reducing the computational complexity. In the next section we will look into those enhanced DBSCAN algorithm based on varied density.

3. Enhanced DBSCAN algorithm

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

3.1.1. Introduction

DBSCAN [6] 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 [13] algorithm handles local density variation within the cluster. The input parameters are taken as: minimum objects(μ), radius, threshold values(α, λ). Growing cluster density mean is calculated with the help of these parameters and after that the cluster density variance is calculated for any core object, which seems to be expanded further by considering density of its E-neighborhood with respect to cluster density mean. After this, comparison of cluster density variance of and is also satisfying the cluster similarity index, then expansion of core object is also allowed.

3.1.2. Description of the Algorithm

(i) A cluster is formed by selecting core object.
(ii) Then 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) Else the object is simply added into the cluster.
Database Approach

3.1.3. Impact of the Algorithm

The DVBSCAN is able to handle the density variations that exist within the cluster. The detection of clusters by this algorithm are having considerable density variation within the clusters. The detected clusters are separated by the sparse region as well as by the regions having the density variation. It increases the performance of the algorithm in comparison to DBSCAN, especially in case of local density. This algorithm finds the clusters that represent relatively uniform regions without being separated by sparse regions. α and λ parameters are used to limit the amount of allowed local density variations within the cluster.

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

3.2.1. Introduction

ST-DBSCAN algorithm is constructed by modifying DBSCAN [6] algorithm. In compare to existing density-based clustering algorithm, ST-DBSCAN [14] algorithm has the ability of discovering clusters with respect to spatial , non-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 spatial, non- spatial, and temporal attributes.
(ii) Noise input is not detected in DBSCAN when it is of varied density but this algorithm overcomes this problem by assigning density factor to each cluster.
(iii) To solve the conflicts in border objects, it compares the average value of a cluster with new coming value.

3.2.2. Description of the Algorithm

The algorithm starts with the first point p in database D.
(i) 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 Eps2, Eps1and Minpts. The objects is assigned as noise if the returned points in Eps-neighborhood are smaller than Minpts input.
(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 core object is selected, then a new cluster is constructed. Then all directly-density reachable neighbors of the 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 ΔE , it is placed into the current cluster.
(vii) If two clusters C1 and C2 are very close to each other, a point p may exist in C1 and C2. Then point p is assigned to cluster which discovered first.

3.2.3. Impact of Algorithm

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

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

3.3.1. Introduction

The DBSCAN [6] algorithm is not capable of finding out meaningful clusters with varied densities. VDBSCAN[15] algorithm detects cluster with varied density as well as automatically selects several values of input parameter Eps for different densities. Also parameter k is automatically generated based on the characteristics of the datasets [16] .

3.3.2. Description of the Algorithm

This algorithm is followed by choosing two steps i.e choosing parameters Epsi and choosing 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 Epsi is selected automatically for each density.
(iv) Dataset is scanned and cluster different densities using corresponding Epsi
(v) Result is displayed with valid cluster with respect to varied density.

3.3.3. Impact of Algorithm

The purpose of this algorithm is to find out meaningful clusters in databases with respect to widely varied densities. Time complexity of VDBSCAN and DBSCAN is same 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. Vibration Method DBSCAN (VMDBSCAN)

3.4.1 Introduction

VMDBSCAN[17] algorithm is constructed by modifying DBSCAN [6] algorithm. In contrast to existing density-based clustering algorithm, detects the clusters of different shapes, sizes that differ in local density. VMDBSCAN first finds out the “core” of each cluster – clusters generated after applying DBSCAN -. Then it “vibrates" points toward cluster that has the maximum influence on these points. The three modifications done in DBSCAN algorithm are as follows
(i) It first clusters the data points using DBSCAN.
(ii) Then, it finds the density functions for all data points within each cluster. The data point that has the minimum density function value will be the core for that cluster, since this point will be local maximum of the density function.
(iii) After that, it computes the density variation of the data point with respect to the density of core object of its cluster against all densities of other core's clusters. With reference to the density variance, we do the movement for data points toward the new core. New core is considered as one of other core's clusters, where maximum influence on the tested data point is exist.

3.4.2 Description of the Algorithm

1. Data sets input and Data standardization.
2. Calculate the Density Function for all the data points.
3. Do Clustering for the data points using traditional DBSCAN algorithm.
4. Find out the core of each cluster.
5. Calculate the Density Function for all the data points within each cluster generated by traditional DBSCAN.
6. For each data point, if its *E* with respect to its core's density function is greater than with respect to other core's density function, then vibrate the data points in that cluster toward the core which has the maximum influence on that point.

3.4.3 Impact of Algorithm

VMDBSCAN algorithm finds the correct number of clusters in varied densities. It gives far more stable estimates of the number of clusters than existing DBSCAN over many different types of data of different shapes and sizes. VMDBSCAN gives better efficiency results than DBSCAN, but it takes more time compared with DBSCAN. This is due that algorithm need to call DBSCAN algorithm to make initial clustering, then it needs to find cores of each returned clusters from DBSCAN.

3.5. Dynamic method DBSCAN (DMDBSCAN)

3.5.1. Introduction

It selects several values of the radius of a number of objects (Eps) for different densities according to a k-dist plot. For existing values of Eps, DBSCAN algorithm is adopted in order to make sure that all the clusters with respect to corresponding density are clustered. And the points for the next process that have been clustered are ignored, for avoiding marking of both denser areas and sparser ones as one cluster. DMDBSCAN[17] will use dynamic method to find suitable value of Eps for each density level of data set.

3.5.2. Description of the Algorithm

The basic idea of DMDBSCAN is that there is need of some methods to find the suitable values of parameters Eps for different densities according to k-dist plot, then traditional DBSCAN algorithm can be used to find clusters. For all value of Eps, DBSCAN algorithm is adopted to find all the clusters with respect to corresponding level of density. Then in the next step of algorithm, all points which clustered ignored. The final result will avoids marking both denser areas and sparser ones as one cluster. Formally, algorithm can describe our proposed to find suitable Epsi for each density level of data set as follow

1. Data sets input and Data standardization.
2. Calculates and stores k-dist for each point and partition k-dist plots.
3. The number of densities is given intuitively by k-dist plot.
4. Choose parameters Eps automatically for each density.

3.5.3. Impact of the Algorithm

DMDBSCAN gives better efficiency results than DBSCAN or DVBSCAN or VMDBSCAN clustering algorithms, but takes more time compared with DBSCAN and DVBSCAN. This is due that algorithm needs to call DBSCAN algorithm for each value of Eps.

4. Conclusion

This paper gives a detail study of clustering algorithms, their classification and comparison between them. Then a detail survey of density based clustering algorithm like DBSCAN, DVBSCAN, ST-DBSCAN, VMDBSCAN, DMDBSCAN is given based on essential requirements required for any clustering algorithm in spatial data. VMDBSCAN gives far more stable estimates of the number of clusters than existing DBSCAN over many different types of data of different shapes and sizes. DMDBSCAN overcomes on the problem of using one local value of Eps, by using local value of Eps for each level of density in a data set. Enhanced algorithm is effective and efficient and outperforms DBSCAN in detecting clusters of different densities and in eliminating noises.

5. References

[1] J., Data Mining Concepts and Techniques. Kaufman, 2006
[2] Margaret H. Dunham, Data Mining "Introduction and Advanced Topics".
[3] IEEE Trans Xu R. Survey of clustering algorithms. Neural Networks 2005;16.
[4] Rastogi R and Shim K, Guha S, CURE:an efficient clustering algorithm for large databases.
[5] R. Rastogi, K. Shim, and S. Guha, "ROCK: A robust clustering algorithm for categorical attributes".
[6] H.-P., Sander J., and Xu X. Ester M., Kriegel "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD"96), Portland: Oregon, pp. 226-231
[7] M. Breunig, H.P. kriegel, and J. Sander, M. Ankerst, "OPTICS: Ordering points to identify the clustering structure,"
[8] J. Yang, and R. Muntz, W. Wang,"STING: A statistical information grid approach to spatial data mining," Very Large Data Bases (VLDB'97), 1997.
[9] S. Chatterje, A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases,"
[10] "Automatic subspace clustering of high dimensional data for data mining applications," 1998, by Gehrke
[11] R.M. Neal and G. E. Hinton. "A new view of the EM algorithm that justifies incremental, sparse and other variants,"
[12] J. C. Bezdeck and W. Full, "Fcm: Fuzzy c-means algorithm," Computers and Geoscience, vol. 10, no. 2-3, pp. 191-203, 1984.
[13] 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.
[14] Derya Birant, Alp Kut, "ST-DBSCAN: An Algorithm for Clustering Spatial-temporal data" Data and Knowledge Engineering 2007 pg 208-221.
[15] 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.
[16] 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.
[17] Mohammad N. T. Elbatta, An improvement of DBSCAN algorithm for best results in varied densities.