A SURVEY ON STUDY OF ENHANCED PARTITION BASED DBSCAN ALGORITHM

Mr. Abhilash Kumar Pandey

Prof. Roshni Dubey

Abstract- Clustering is an important task in mining evolving data streams. 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. DBSCAN algorithm is very effective for analyzing large and complex spatial databases. DBSCAN need large volume of memory support and has difficulty with high dimensional data. Partitioning-based DBSCAN was proposed to overcome these problems. I have presented a study of DBSCAN & Enhanced Partition based DBSCAN algorithm.

Keyword- Clustering, Density Based Clustering, DBSCAN & PDBSCAN.

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 can 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:

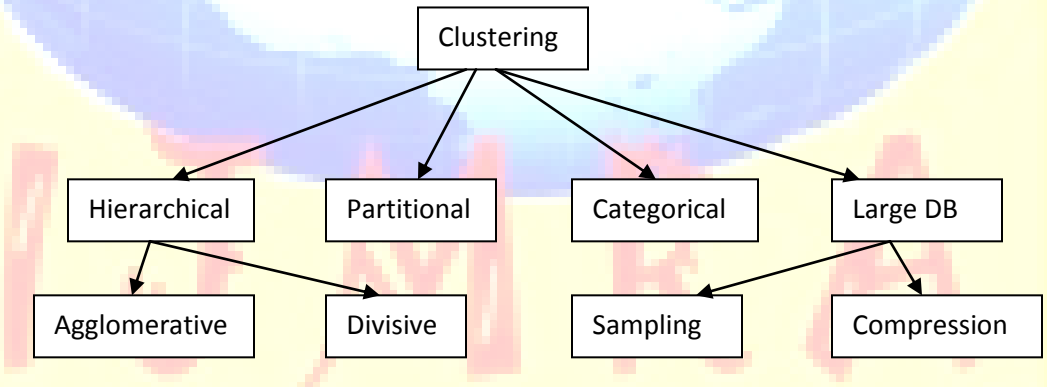

Figure 1: Clustering techniques

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 Kmeans 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 partitioned. 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], Wave Cluster [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 partition 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 gets 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.

This paper is organized as follows: In the next section, a DBSCAN algorithm is explained in details. Section 3 covers the survey of enhanced partition based 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) = $\{p \in D | dist(p,q) \leq Eps\}$ here 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 densityreachable 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

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. 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 (n2**)**, where n is the number of database objects. If a spatial index is used, the complexity can be educes 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 Partition Based DBSCAN

PDBSCAN [12] run DBSCAN algorithm on each partition which is partitioned by special rules. With PDBSCAN, the R⁄-tree should be built. DBSCAN requires specifying two global parameters Eps and MinPts. In order to reduce the computational complexity, MinPts is fixed to 4 usually. Then the k-dist graph must be plotted to decide the value of Eps. K-dist graph needs to calculate the distance of an object and its kth nearest neighbors for all the objects. Next, sort all the objects on the basis of the previous distances. Finally, plot the k-dist graph according to all the sorted objects and distances. Considering that building the R⁄-tree and plotting the k-dist graph have to cost much time especially for a large database, the initial database is partitioned into N partitions to reduce the time cost. Partitioning database can also alleviate the burden of memory and find more precise parameter Eps for every partition.

3.1. Partitioning Approach

Partitioning database is one of the most important steps for PDBSCAN. In this step, the algorithm needs to divide the database into N so that the parameter Eps of each partition can be specified more exactly.

3.2. **Enhanced Partition Based DBSCAN algorithm**

(i) Arbitrary select a point ***p***

(ii) Retrieve all points density-reachable from ***p*** wrt ***Eps*** and ***MinPts***.

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

(iv) If ***p*** does not match with cluster then add value to noise.

(v) If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.

(vi) Continue the process until all of the points have been processed

(vii) Create partitions based on relative density of points and distribute clusters accordingly.

4. Conclusion

This paper gives a detail study of clustering algorithms, their classification and algorithm. Then a detail survey of density based clustering algorithm DBSCAN and Enhanced Partition based DBSCAN algorithm. Among all clustering methods, density-based clustering algorithm is one of powerful tools for discovering arbitrary shaped clusters in large spatial databases. This algorithm can partition database in to N partitions according to the density of data. Enhanced Partition based DBSCAN 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] Chaudhari Chaitali G. "Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-2, December 2012