Study of Density based Algorithms

VIVEK S WARE
Computer Engineering Department, Mumbai University, K J Somaiya College of Engineering, Mumbai, Maharastra, India

BHARATHI H N
Computer Engineering Department, Mumbai University, K J Somaiya College of Engineering, Mumbai, Maharastra, India

ABSTRACT

Clusters formed on the basis of density are very useful and easy to understand and they do not limit to their shapes. Basically two types of density based algorithms are present. One is density based connectivity which focuses on Density and Connectivity and another is Density function which is total mathematical function. They work best in spatial database. In this paper, we are studying DBSCAN, VDBSCAN, DVBSCAN, UDBSCAN, OPTICS, DBNCLUE, GDBSCA and DBCLASD. We analyze some of the algorithms in terms of meaningful clusters.

Keywords

Dbscan, Vdbscan, Dvbscan, Udbscan, Optics, Dbnclue, Gdbsca And Dbclasd.

1. INTRODUCTION

New Term - 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. Clustering i.e., grouping the objects from a database into meaningful subclasses is one of the major data mining methods. Among many types of clustering algorithms density based algorithm is more efficient in detecting the clusters with varied density. Clustering analysis divides data into groups (Clusters) that are meaningful. If meaningful groups are goals, then cluster should capture the natural structure of the data. In some cases, however, cluster analysis is only useful as starting point for the other purpose such as data summarization. Cluster analysis long played an important role in wide variety of fields like Psychology and other social sciences, biology statistics, pattern recognition, information retrieval, machine learning and data mining.

Clustering is an unsupervised problem[1] and it deals with finding a structure in collection of unlabeled data. So simple definition of clustering can be as “the process of organizing objects into groups where members are similar”. Therefore a cluster can be a collection of objects which are similar between them and dissimilar with other clusters. Clustering algorithms need to satisfy following requirement: Scalability, different types of attribute, finding clusters with arbitrary shape, domain knowledge to determine input parameter, ability to handle noise and usability.

Cluster analysis work as group data objects based only on information found in the data that tells objects and their relationships. Goal is that greater homogeneity within a group and measures difference between groups. Cluster analysis is not a specific algorithm but involves general task to be solved and it is iterative process of knowledge discovery that involves trial and failure.

2. MOTIVATION AND RELATED WORK

There are many different categories in clustering. Some of them are 1.Partition, 2.Hierarcical, 3.Denstiy based, 4.Model based, 5.Grid based and 6 Ensemble.

A partition algorithm normally splits the data points into several subsets. Probabilistic, K-Medoids, K-Means are all sub class of partition methods. Some examples for Partition algorithm are PAM, CLARANS.

Hierarchical clustering algorithms cluster the data points in tree structure. Every cluster node will contain child cluster and sibling cluster. The Hierarchical clustering method can be classified as agglomerative and divisive. The Agglomerative clustering starts clustering one single point cluster and recursively merges two or more respective clusters. A divisive clustering starts with one cluster of all data points and repeatedly splits the most appropriate cluster. BRICH , CURE , CHAMELEON are Hierarchical clustering algorithms.

The mechanism behind the Density based clustering is that the open Euclidean space will be sub divided into set of data points. Further classification of this is algorithm is Density connectivity based and Density Function based. DBCLASD, OPTICS DBSCAN are density based clustering algorithms. [4]

Different algorithms in density based clustering like UDBSCAN [8], GDBSCAN, P-DBSCAN and current algorithms OPTICS are discussed and studied [1]. Three different algorithms of same categories i.e. DBSCAN, DBCLASD and mathematical based algorithm DENCLUE [7] are compared using different parameters [2]. Density algorithms for mining Large spatial database like VDBSCAN, DVBSCA are discussed [3] [4].

3. DENSITY BASED CLUSTERING ALGORITHMS

Generally clustering is classified into two categories i.e. non-exclusive (overlapping) and exclusive (non-overlapping). Exclusive clustering is further divided into two categories i.e. extrinsic (supervised) and intrinsic (unsupervised). Now intrinsic clustering is further divided into hierarchical and partition methods. Here is some introduction of both methods.

Hierarchical clustering, as its name suggests is a sequence of partitions in which each partition is nestled into the next partition in the sequence i.e. maintaining hierarchy. Hierarchical clustering is depicted by binary tree or dendrogram . Whole data set is represented by root node and rest of the leaf node is represented as data object. At any level cutting a dendrogram defines clustering and identifies new clusters. Partition clustering is of non hierarchical type. It generates a single partition of the data in order to achieve natural groups present in the data.

Density based algorithm belongs to partitional clustering. In Density based clustering there is partition of two regions i.e. low density region to high density region .A cluster is defined as a connected dense component that grows in any direction where a density leads. This is the reason that density based algorithms are capable of discovering clusters of arbitrary shapes and provides natural protection to outliers. Basically, density based clustering is divided into two categories i.e. density based connectivity and density function. In density based connectivity, density and connectivity are two main concepts comes under this and both measured in terms of local distribution of nearest neighbors. Density based connectivity algorithm examples are DBSCAN, GDBSCAN, OPTICS and DBCLASD algorithms and density function includes DENCLUE algorithm.

3.1 DENSITY BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE (DBSCAN) [1]

It is of Partitioned type clustering where more dense regions are considered as cluster and low dense regions are called noise.

Algorithm

Steps of algorithm of DBSCAN are as follows

- Arbitrary select a point r.
- Retrieve all points density-reachable from r w.r.t Eps and MinPts.
- If r is a core point, cluster is formed.
- If r is a border point, no points are density-reachable from r and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been Processed

In K-mean DBSCAN does not require prior knowledge of number of cluster. Disadvantage is that here we need to specify global parameter (Eps, MinPts) in advance. Single and two dimensional K-mean is present.

3.2 DENSITY BASED CLUSTERING FOR UNCRTAIN OBJECTS (UDBSCAN) [1]

Existing traditional clustering algorithm were designed to handle static objects. UDBSCAN extends the existing DBSCAN algorithm to make use of their derived vector deviation function which defines deviation in each direction from the expected representative.

3.3 GENERALIZED DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (GDBSCAN) [1]

It is a generalized version of DBSCAN. It can cluster point objects as well as polygon objects using spatial and non-spatial attributes. GDSCAN generalized DBSCAN in two ways; first if symmetric and reflexive are the two properties of neighborhood then we can use any notion of the neighborhood of an object. The two properties that are symmetric and reflexive are termed as binary predicate. Another (Second) by calculating the non spatial attributes with defining cardinality of the neighborhood. GDBSCAN has five important applications. In the first application we cluster a spectral space (5D points) created from satellite images in different spectral channels which are a common task in remote sensing image analysis. The second application comes from molecular biology. The points on a protein surface (3D points) are clustered to extract regions with special properties. To find such regions is a subtask for the problem of protein-protein docking. The third application uses astronomical image data (2D points) showing the intensity on the sky at different radio wavelengths.

The task of clustering is to detect celestial sources from these images. The last application is the detection of spatial trends in a geographic information system. GDBSCAN is used to cluster 2D polygons creating so-called influence regions which are used as input for trend detection. Spatial index structures such as R-trees may be used with GDBSCAN to improve upon its memory and runtime requirements and when not using such a structure the overall complexity is O (n log n).

3.4 ORDERING POINTS TO IDENTITY THE CLUSTERING STRUCTURE (OPTICS) [1]

The basic idea of OPTICS and DBSCAN algorithm is same but there is one drawback of DBSCAN i.e. while densities are varying it is difficult to detect meaningful clusters which were overcome in OPTICS algorithm. It not only stores the core distance but also a suitable reach ability distance for each object and creates a proper ordering of database. Advantage of algorithm is that it can be used to extract basic clustering information. As it requires wide range of parameter setting it become its disadvantage. Complexity is O (*kN2*) where k is no of dimensions Run time is *O (n log n)* [1]

3.5 DISTRIBUTION BASED CLUSTERING ALGORITHM FOR MINING LARGE SPATIAL DATABASES (DBCLASD) [3]

This Algorithm 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.

DBCLASD 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.[3] It is very efficient for large spatial databases. This algorithm fulfills all the requirements needed for designing a good clustering algorithm for spatial databases.

3.6 VARIED DENSITY

BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (VDBSCAN) [3]

The DBSCAN algorithm is not capable of finding out meaningful clusters with varied densities. Algorithm detects cluster with varied density as well as automatically selects users. Algorithm purpose 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. [3] Even the input parameters (Eps) are automatically generated from the datasets.

3.7 DENSITY BASED ALGORITHM FOR DISCOVERING DENSITY VARIED CLUSTERS IN LARGE SPATIAL DATABASES (DVBSCAN) [3]

Algorithm detects clusters with different shapes and sizes but fails to detect clusters with varied densities that exists within the cluster [9]. DVBSCAN 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.

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. The parameters α and λ are used to limit the amount of allowed local density variations within the cluster.

3.8 DENSITY BASED CLUSTERING (DENCLUE) [1]

In this algorithm concept of influence and density function is used. In this influence of each data point can be modeled formally using a mathematical function and that is called an influence function. Influence function describes the impact of data point within its neighborhood after that we calculate density function which is sum of influences of all data points. According to DENCLUE two types of clusters are defined i.e. centre defined and multi centre defined clusters .In centre defined cluster a density attractor x*(f^D_B(X*)> ξ) is the subset of the database which is density attracted by x* and in multicenter defined cluster it consist of a set of center defined clusters which are linked by a path with significance and ξ is noise threshold. The influence function of a data object y Є Fd is a function fY : Fd → R^+_0 which is defined in terms of a basic influence function Fb

$$f^y_B(x) =- fB(x, y).$$

The density function is defined as the sum of the influence functions of all data points.

$$\hat{f}^D_B(x) = \sum_{x_i \in near(x)} f^{x_i}_B(x) .$$

DENCLUE also generalizes other clustering methods such as density based clustering; partition based clustering, hierarchical clustering. In density based clustering DBSCAN is the example and square wave influence function is used and multicenter defined clusters are here which uses two parameter σ = Eps, ξ = MinPts. In partition based clustering example of k-means clustering is taken where Gaussian influence function is discussed. Here in center defined clusters ξ=0 is taken and σ is determined. In hierarchical clustering center defined clusters hierarchy is formed for different value of σ. Faster than DBSCAN by a factor of up to 45.

Algorithm

The DENCLUE algorithm works in two steps.

Step one is a pre-clustering step, in which a map of the relevant portion of the data space is constructed. The map is used to speed up the calculation of the density function which requires to efficiently accessing neighboring portions of the data space.

The second step is the actual clustering step, in which the algorithm identifies the density-attractors and the corresponding density attracted points.

4. CONCLUSION

In this paper an up to date survey on Density based clustering algorithm is done. It tries to focus not only the renewed algorithms such as DBSCAN, GDBSCAN, DBCLASD, OPTICS, DENCLUE but also algorithms like UBSCAN that having improvements in efficiency and runtime other than existing algorithms. In addition advantages and disadvantages of some algorithms are discussed and also impacts of some algorithms are also present.

In future, DBSCAN 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. IN DBCLASD, The existing algorithm is suitable for uniform distribution of points. Each algorithm is unique with its own features.

A comparative study in terms of input parameters, shapes of cluster, Density type and types of data used is shown below.

Table1: Comparison of Five density based algorithms

Name of Algorithm	Density Based Spatial Clustering of Applications with Noise	Distribution- Based clustering Algorithm for Mining Large Spatial Databases	Density based Clustering	Varied Density Based Spatial Clustering of Applications with Noise	Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Databases
Shape of Cluster	Arbitrary	Arbitrary	Arbitrary	Arbitrary	Arbitrary
Type of data	Spatial Data with Noise	Spatial Data with uniformly Distributed points	Large no. of data	Spatial Data with Varied Density	Spatial Data with Varied Density
Density type (varied)	No	YES	YES	YES	YES
Input parameter	Radius and minimum size	Automatically Generated	Two input Parameters	Automatically Generated	Two input Parameters
Complexity	O(n^2)	O($3n^2$)	O(log |D|)	Same as DBSCAN	Higher than DBSCAN
Handling of noise	NOT Very Well	Good	Very Well	Good	Good
Purpose of algorithm	To discover clusters with arbitrary shape	Design good cluster for spatial database	Can discover other clustering algorithms like hierarchical clustering, partition based clustering etc.	Find out meaningful cluster in database w.r.t widely varied density	Find out the density variations that exit within the cluster
Advantage	DBSCAN doesn't require no. of cluster in the data at prior stage	DBCLASD requires no user input	Good clustering properties in data sets with large amount of noise	Automatically select several input parameter and detect cluster with varied density	Handles local density variation within the cluster
Dis-advantage	Does not respond data with varied density	Slower than DBSCAN	Data points are assigned by hill climbing, it make unnecessary small steps	If parameter selection goes wrong then it has problem	High time complexity
Cluster Testing	No	Yes, with 2 features	No	Yes. Display cluster w.r.t varied density	Yes. Density are check with threshold value

5. REFERENCES

[1] Pooja Nagpal and Priyanka Mann, 2011, "Survey of Density Based Algorithms" in International Journal of Computer Science and Applications.

[2] Pooja Nagpal and Priyanka Mann, august 2011, "Comparative Study of Density Based Clustering Algorithms" in International Journal of Computer Science and Applications, volume-27.

[3] M. Parimala, Daphne Lophne and N. C. Senthilkumar, June 2011, "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases" in International Journal of Advanced Science and Technology, Volume-31.

[4] S. Vijayalakshmi, Dr. M. Punithavali, 2010, "Improved Varied Density Based Clustering Algorithms for Mining Large Spatial Databases" in Research and Development Center Bharathiar University Coimbatore, India IEEE.

[5] Peng Liu, Dong Zhou and Naijun Wu, "Varied Density Based Spatial Clustering of Application with Noise", in proceedings of IEEE Conference ICSSSM 2007 Pg 528-531.

[6] K.Jain, M. N. Murty and P. J. Flynn, 1999, "Data Clustering: A Review", ACM, 31(1999), PP.264-323.

[7] Hinneburg and D. Keim, 1998, "An efficient approach to clustering Large multimedia databases with noise" ,in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD'98), PP. 58-65.

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