ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Comparative Analysis of Density Based Clustering Techniques for Outlier Mining

R.Prabahari*, Dr.V.Thiagarasu
Assistant Professor*, Associate Professor, PG & Research Department of Computer Science, Gobi Arts & Science College, Gobi – 638 452, Tamil Nadu, India

Abstracts

Density based Clustering Algorithms such as Density Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points to Identify the Clustering Structure (OPTICS) and DENsity based CLUstering (DENCLUE) are designed to discover clusters of arbitrary shape. DBSCAN grows clusters according to a density based connectivity analysis. OPTICS, which is an extension of DBSCAN used to produce clusters ordering obtained by setting range of parameter. DENCLUE clusters object is based on a set of density distribution functions. The comparison of the algorithms in terms of essential parameters such as complexity, clusters shape, input parameters, noise handle, cluster quality and run time are considered. The analysis is useful in finding which density based clustering algorithm is suitable in different criteria.

Keywords: Clustering, Density based clustering, DENCLUE, OPTICS, DBSCAN.

Introduction

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups . It is main task of exploratory data mining and a common technique for statistical data analysis is used in many fields including pattern recognition, image analysis, information retrieval, machine learning and bioinformatics. Cluster analysis itself is not one specific algorithm [Chandra E & Anuradha. V. P, 2011] but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.

The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. A cluster is an ordered list of objects, which have some common characteristics.

There are many types of clustering techniques but there are two types of clustering which is relevant to this paper i.e. Hierarchical and Partitional Clustering.

Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of cluster. Strategies for hierarchical clustering generally fall into two types.

- **Agglomerative**: This is a "bottom up" approach- each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy
- **Divisive**: This is a "top down" approach- all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy

Hierarchical clustering is an agglomerative (top down) clustering method. As it name suggests, the idea of this method is to build a hierarchy of clusters, showing relations between the individual members and merging clusters of data based on similarity. In the first step of clustering, the algorithm l looks for the two most similar data points and merge them to create a new "pseudo-data point", which represents the average of the two merged data points. Each Iterative step takes the next two closest data points (or pseudo-data points) and merges them. This process is generally continued until there is one large cluster containing all the original data points. Hierarchical clustering results in a "tree", showing the relationship of all of the original points [Hinneburg A & Keim D,1998].

Partitional Clustering

Partitional clustering decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K ($N \geq$K) partitions of the data, with each partition representing a cluster. That is, it classifies the data into K groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group. Notice that for fuzzy partitioning [Smite et al.,2013], a point can belong to more than one group. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. To achieve global optimality in partitioned-based clustering, an exhaustive enumeration process of all possible partitions is required.

Many partition clustering algorithms try to minimize an objective function [Chaudhari Chaitali, 2012]. For example, in *K*-means and *K*-medoids the function (also referred to as the distortion function) is

$$\sum_{i=1}^{K}\sum_{j=1}^{|C_i|}(\mathrm{Dist}\,(x_{j,}\mathrm{center}(i))),$$

Where $|C_i|$ is the number of points in cluster *i*, Dist(x_j, center (*i*)) is the distance between point x_j and center *i*. Many distance function can be used, such as Euclidean distance and L_1 norm. Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. This algorithm typically determines all clusters at once. Most applications adopt one of two popular heuristic methods like k-mean algorithm k-medoids algorithm.

Density based methods which is the main concern of our paper belong to Partitional clustering. Density based clusters are defined as clusters which are differentiated from other clusters by varying densities [Ram A et al.,2010] that means a group which have dense region of objects may be surrounded by low density regions. Density based methods are of two types [Pragati Shrivastava & Hitesh Gupta, 2012] Density based Connectivity and Density based Functions.

Density based Connectivity is related to training data point. DBSCAN [Chaudhari Chaitali G, 2012] and OPTICS [Mihael Ankerst et al.,1999] comes under Density Connectivity while Density function is related to data points to computing density functions defined over the underlying attribute space. DENCLUE [Santhisree K & Damodaram,2011] comes under Density function.

This paper is organized as follows. Literature surveys are given in section 2. In section 3 discusses the comparison of density methods in detail. Experimental results are reported in section 4. Conclusions are presented in section 5.

Literature survey

DBSCAN

DBSCAN [Chakraborty S & Nagwani N. K, 2011] is a data clustering algorithm and it is a density based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN [Parimala M et al., 2011]is one of the most common clustering algorithms and also most cited in scientific literature. OPTICS can be seen as a generalization of DBSCAN to multiple ranges, effectively replacing the ε parameter with a maximum search radius.

DBSCAN's definition of a cluster is based on the notion of density reach ability. Basically, a point q is directly density-reachable from a point p if it is not farther away than a given distance ϵ (i.e., is part of its ϵ -neighborhood) and if p is surrounded by sufficiently many points such that one may consider p and q to be part of a cluster. q is called density-reachable (note the distinction from "directly density-reachable") from p if there is a sequence p_1, p_2, …. p_n of points with $p_1 = p_n$ and $p_n =$ q where each p_{i+1} is directly density-reachable from p_i.Note that the relation of density-reachable is not symmetric. q might lie on the edge of a cluster, having insufficiently many neighbors to count as dense itself. This would halt the process of finding a path that stops with the first non-dense point. By contrast, starting the process with q would lead to p (though the process would halt there, p being the first non-dense point). Due to this asymmetry, the notion of density-connected is introduced the two points p and q are density-connected if there is a point O such that both p and q are density-reachable from O. Density-connectedness *is* symmetric[Domeniconi C & Gunopulos D,2004].

A cluster, which is a subset of the points of the database, satisfies two properties.

- All points within the cluster are mutually density-connected.
- If a point is density-connected to any point of the cluster, it is part of the cluster as well.

Obviously clusters are define on some criteria which is as follows

Core: Core points lie in the interior of density based clusters and should lie within *Eps* (radius or threshold value), *MinPts* (minimum no of points) which are user specified parameters.

Border: Border point lies within the neighbourhood of core point and many core points may share same border point.

Noise: The point which is neither a core point nor a border point

Directly Density Reachable: A point r is directly density reachable from s w.r.t Eps and MinPts if belongs to NEps(s) and |NEps (s)| >= MinPts.

Density Reachable: A point r is density reachable from r point s wrt.Eps and MinPts if there is a sequence of points $r_1 \dots r_n$, $r_1 = s$, $r_n = s$ such that r_{i+1} is directly reachable from r_i.

Algorithm:

Step 1: Pre-Processing

Firstly, a pre-processing step must be applied to the removal of noise and diffuse emission. As stated before, this might be accomplished by using a threshold.

Step 2: DBSCAN Clustering

Secondly, the DBSCAN algorithm can be applied on individual pixels to link together a complete emission area at the images for each channel of the electromagnetic spectrum. This is done by setting the eps parameter to some value that will define the minimum area required for a source to be considered. The eps parameter will define the distance metric in terms of pixels. Each of the generated cluster will define a celestial entity.

Step 3: Multi-spectral Correlation

After identifying all clusters, one can apply a multi-spectral correlation process in order to consider the results (generated clusters) from every electromagnetic wavelength. It will not be detailed here, but a common approach would be only considering clusters which have one or more counterparts close enough with respect to some threshold on the other channels of the electromagnetic spectrum.

OPTICS

Ordering Points to Identify the Clustering Structure (OPTICS) is an algorithm for finding density-based clusters in spatial data. The basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses, the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are linearly ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster[Anant Ram et al., 2010]. OPTICS generalizes DB clustering by creating an ordering of the points that allows the extraction of clusters with arbitrary values for ε.

The **core-distance** is the smallest distance ε*’* between *p* and an object in its ε-neighborhood such that *p* would be a core object.

The **reachablity-distance** of *p* is the smallest distance such that *p* is density-reachable from a core object *o*.

The **generating-distance** ε is the largest distance considered for clusters. Clusters can be extracted for all ε *i* such that $0 \leq \varepsilon\, i \leq \varepsilon$

Algorithm:

Step 1: Find core distance of an object p is the smallest ε’ value that makes {p} a core object. If p is not core object , the core distance of p is undefined

Step 2 : The reachablity distance of an object q with respect to another object p is the greater value of the core–distance of p and the Euclidean distance between p and q. If p is not a core object, the reachablity distance between p and q is undefined.

DENCLUE

Closeness to a dense area is the only criterion for cluster membership DENCLUE has two variants[Heneburg & Keim D., 1998]. Arbitrary-shaped clusters & Centered defined cluters. Arbitrary shaped clusters similar to other density based methods. Center-defined clusters, similar to distance-based methods. The DENCLUE algorithm employs a cluster model based on Gaussian influence function. A cluster is defined by a local maximum of the estimated density function. Data points are assigned to clusters by hill climbing, i.e. points going to the same local maximum are put into the same cluster. A disadvantage of DENCLUE 1.0 is used hill climbing may make unnecessary small steps in the beginning and never converges exactly to the maximum, it just comes close. A new hill climbing procedure is introduced, which adjusts the step size automatically at no extra costs, prove that the procedure converges exactly towards a local maximum by reducing it to a special case of the expectation maximization algorithm. Experimentally that the new procedure needs much less iterations and can be accelerated by sampling based methods with sacrificing only a small amount of accuracy.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

In this algorithm concept of influence and density function is used. The 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 calculate density function which is sum of influences of all data points.

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 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.

Algorithm:

Step 1: Find the influence of each data point can be modeled using a Gaussian influence function

Step 2: The overall density of the data space can be modeled analytically as the sum of the influence function applied to all data points

Step 3: Clusters can then be determined by identifying density attractors where density attractors are local maximum of the overall density function.

Comparative study

All the experiments are done on Intel Core 2 Duo CPU having processor speed of 2.0 GHz with 0.99 GB of RAM. Implementation is done in MATLAB7.0. Iris Data set has been used for all experiments.

Data preprocessing

Data reduction technique using by Principal Component Analysis (PCA) [Linsay I Smith, 2007]. Reducing data results into completeness and simplicity of data help in getting accuracy in results.

Results and discussions

Experiment is done on six parameters which are defined. First one is complexity. Run time complexity of DBSCAN is O(n^2) when these are not using any accelerating index. But when noise increases run time complexity gets worse. OPTICS is an equivalent with DBSCAN in structure they have the same complexity and DENCLUE is O(log(|n|) and when noise increases performances gets better. Second parameter is shape of clusters. All the three algorithms support arbitrary shape of clusters. Third one is Input Parameter. To declare input parameter in advance is very typical job and that parameters effects efficiency as well as quality of clusters. So either there should be an efficient way to tell these parameters or there should no predefined parameters. All the three algorithms require two input parameters. Fourth parameter is Handling of Noise as noise increases DENCLUE performs very well and OPTICS also perform good but in case of DBSCAN, it does not perform so well. Fifth one is Cluster quality that is defined in terms of F score [Domeniconi C & Gunopulos D, 2004] value. DBSCAN have highest disagreement value of F score and after that OPTICS and then DENCLUE follows. So, DENCLUE is superior than DBSCAN and OPTICS. Sixth and last parameter is run time of algorithms. DENCLUE having least run time, after that OPTICS's run time. DBSCAN's run time is nearly equal to three times the run time of OPTICS.

Algo Rithms	Comp lexity	Cluters Shape	Input para meters	Noise Handle	Cluster Quality	Run Tim e (ms)
DB SCAN	O(n^2)	Arbit rary	Two	Not good	91.3%	120
OPTI CS	O(n^2)	Arbit rary	Two	Good	94.3%	40
DEN CLUE	O(logn)	Arbit rary	Two	Very Good	97.08%	30

The table shows that run time of DENCLUE algorithm is lowest while OPTICS and DBSCAN having highest run time. In terms of cluster quality DENCLUE leads while OPTICS and DBSCAN is lacking behind.

Conclusions

Three important density based algorithms are analyzed. From the comparative study, it is observed that the DENCLUE algorithm leads in terms of cluster quality. DENCLUE uses hill-climbing to calculate the density attractors of a density function. Clusters are formed by associating data objects with density attractors during the hill climbing procedure. The experimental evaluations in this study show that

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

outperforms the efficiency and effectives of DENCLUE when compared to other algorithms.

References

1. 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 Vol. 3, No.6, June 2010.
2. 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, Vol.2, Issue-2, pp. 212 – 215, December 2012.
3. Chakraborty S., Prof. Nagwani N. K., Analysis and Study of Incremental DBSCAN Clustering Algorithm, International Journal of Enterprise Computing And Business Systems, Vol. 1, July 2011.
4. Chandra. E, Anuradha. V. P, A Survey on Clustering Algorithms for Data in Spatial Database Management System, International Journal of Computer Applications, Col. 24, June 2011.
5. Domeniconi C, Gunopulos D, “An efficient density-based approach for data mining tasks.”, Knowl. Inform Syst 6(6):750–770, 2004.
6. Hinneburg A and D. Keim, “An efficient approach to clustering Large multimedia databases with noise” in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD 98), 1998, pp. 58–65.
7. Linsay I Smith, “A tutorial on Principal components Analysis”, June 10,2007
8. Mihael Ankerst , Markus M. Breunig , Hans-peter Kriegel , Jorg Sander,, “OPTICS: Ordering Points To Identify the Clustering Structure”, Proceedings of the 1999 ACM SIGMOD international conference on Management of data, 1999.
9. Mihael Ankerst , Markus M. Breunig , Hans-Peter Kriegel , Jörg Sander, “OPTICS: ordering points to identify the clustering structure”, Proceedings of the 1999, ACM SIGMOD international conference on Management of data, p.49-60, May 31-June 03, 1999, Philadelphia, Pennsylvania, United States.
10. Parimala M., Lopez D., Senthilkumar N. C., A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases, International Journal of Advanced Science and Technology, Vol. 31, June 2011.
11. Peter J. H., Antonysamy A., An optimized Density based Clustering Algorithm, International Journal of Computer Applications, Vol. 6, September 2010.
12. Pragati Shrivastava and Hitesh Gupta, “A Review of Density-Based clustering in Spatial Data”, International Journal of Advanced Computer Research, pp. 2249-7277, September 2012.
13. Ram A., Jalal S., Jalal A. S., Kumar M., A Density based Algorithm for Discovering Density varied clusters in Large Spatial Databases, International Journal of Computer Applications, Vol. 3, June 2010.
14. Santhisree K., Dr. Damodaram A., SSM-DBSCAN and SSM-OPTICS : Incorporating new similarity measure for Density based clustering of Web usage data, in International Journal on Computer Sciences and Engineering, August 2011.
15. Smiti, Abir, and Zied Eloudi, "Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory", In the IEEE 6th International Conference on Human System Interaction 2013, pp. 380-385, 2013.