Comparative Study of Density based Clustering Algorithms

Pooja Batra Nagpal
Department Of Computer Science
NIT Kurukshetra

Priyanka Ahlawat Mann
Department Of Computer Science
NIT Kurukshetra

ABSTRACT

This paper presents a comparative study of three Density based Clustering Algorithms that are DENCLUE, DBCLASD and DBSCAN. Six parameters are considered for their comparison. Result is supported by firm experimental evaluation. This analysis helps in finding the appropriate density based clustering algorithm in variant situations.

General Terms
Algorithms .

Keywords
Clustering Algorithms, Density based Algorithms, Clustering in Presence of Noise.

1. INTRODUCTION

The act of dividing meaningful groups of objects that share common properties is called clustering and groups having that objects which share common properties are called clusters. There are many types of clustering but we will discuss here only two type of clustering which is relevant to our topic i.e. Hierarchical and Partitional Clustering [1]. Both are of exclusive and of intrinsic type.

In Hierarchical clustering clusters which are in nested form are organized as a tree where root is the cluster which contains all the objects and internal nodes or clusters are the union of its sub clusters. Suppose we have N objects to be clustered and are denoted as $\sum$.

$\sum = \{ x_1 \ldots x_i \ldots x_n \}$ where xi is the ith object. A partition ξ breaks $\sum$ into subsets $\{C_1, C_2 \ldots .Cm\}$ satisfying the following

$C_i \cap C_j = \Phi$ where $i \neq j$ and

$C_1 \cup C_2 \cup \ldots\ldots C_m = \sum$

In Partitional clustering data is divided into subsets or clusters and subset should be of non- overlapping type in such a way that each no two subset will share a common data object. In it problem can be stated as n patterns in a d-dimensional space are given and to determine partitions of patterns into k clusters. K may be defined or not. Now according to Partitional clustering solution of this problem is to choose a criterion and evaluate it for all partitions and pick the partition that suits most the criteria.

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 that means a group which have dense region of objects may be surrounded by low density regions. Density based method are of two types: Density based Connectivity and Density based Functions [1].

Density based Connectivity is related to training data point and DBSCAN [2] and DBCLASD [3] comes under this while Density Functions is related to data points to computing density functions defined over the underlying attribute space and DENCLUE[4] comes under this.

2. RELEVENT RESEARCH ARTICLE

2.1 DBSCAN

DBSCAN[2] (Density Based Spatial Clustering of Applications with Noise) It is of Partitional type clustering where more dense regions are considered as cluster and low dense regions are called noise. 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 a 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 $r_1 \ldots r_n$, r_1 = s, r_n = s such that r_{i+1} is directly reachable from r_i.

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

2.2 DBCLASD

DBCLASD [3] (Application Based Clustering Algorithms for Mining in Large Spatial Databases) Basically DBCLASD is an incremental approach. A point is assigned to a cluster that processed incrementally without considering the cluster.

In DBCLASD cluster is defined by three properties which are as follows:

1) Expected Distribution condition NNDistSet(C) which is set of nearest neighbour of cluster C has the expected distribution with required confidence level.
2) Maximality Condition Every point that comes into neighbouring of C does not fulfil condition (1).
3) Connectivity Condition Each pair (a,b) are connected through grid cell structure.

Algorithm

- Make set of candidates using region query
- If distance set of C has expected distribution then point will remain in cluster.
- Otherwise insert point in list of unsuccessful candidates.
- In the same way expand cluster and check condition
- Now list of unsuccessful candidates is again checked via condition.
- If passes then put in cluster otherwise remain in that list

There are two main concepts in DBCLASD. First one is generating candidates and candidate generation is done on the basis of region query that specifies some radius for circle query to accept candidates. Second one is testing the candidates which is done through chi square testing. Points that lie under the threshold value are considered right candidates while those lies above threshold are remain in unsuccessful candidates' list. In last unsuccessful candidate list is again checked and every point go through test and points passes test are considered in cluster while left remains in unsuccessful candidates' list.

2.3. DENCLUE

DENCLUE [4] (Density based clustering) Main concepts are used here i.e. influence and density functions. Influence of each data point can be modelled as mathematical function and resulting function is called Influence Function. Influence function describes the impact of data point within its neighbourhood. Second factor is Density function which is sum of influence 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. The influence function of a data objects y Є F is a function. Which is defined in terms of a basic influence function F, $F(x) = F(x, y)$.

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

Algorithm

- Take Data set in Grid whose each side is of 2σ
- Find highly densed cells i.e.
- Find out the mean of highly populated cells
- If $d(\text{mean}(c_1), \text{mean}(c_2)) < 4a$ then two cubes are connected.
- Now highly populated or cubes that are connected to highly populated cells will be considered in determining clusters.
- Find Density Attractors using a Hill Climbing procedure.
- Randomly pick point r.
- Compute Local 4 σ density
- Pick another point (r+1) close to previous computed density.
- If den(r) < den(r+1) climb.
- Put points within (σ /2) of path into cluster.
- Connect the density attractor based cluster.

3. EXPERIMENATL SETUP

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 JAVA. Fisher's Iris Flower Data set is chosen for all experiments.

3.1 Data Pre-processing

Iris Flower data set is a four attribute data composed of the width of flower stalk, length stalk, the width of petal and the length of petal. In our experiment we are using data reduction technique by using Principal Component Analysis or PCA [5].Reducing data results into completeness and simplicity of data which helps in getting accuracy in results. Waveform representation of Iris Data set is shown in fig 1.

3.2 Framework

These are the some parameters that comes under framework:

- Complexity
- Shape of Clusters
- Input Parameters
- Handling of Noise
- Cluster Quality
- Run Time
-

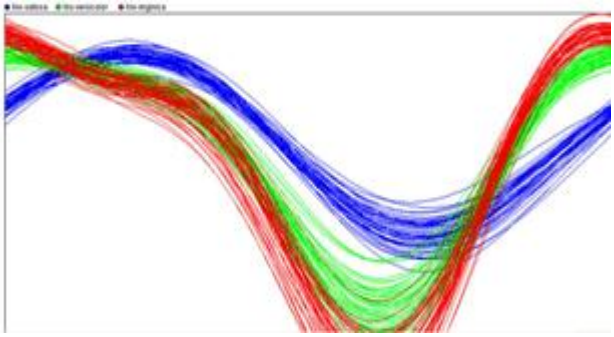


Fig. 1 Waveform representation of Iris Data set

4. EXPERIMENTAL EVOLUTION

Experiment is done on six parameters which are defined above. First one is complexity. Run time complexity of DBSCAN is O(n2) when we are not using any accelerating index. But when noise increases run time complexity gets worse. Run time complexity of DBCLASD is O(3n2) and of DENCLUE is O(log(|D|) and when noise increases performances gets better.

Second parameter is shape of clusters. All the three algorithms support arbitrary shape of clusters which is shown in figure2, figure3 and figure4.

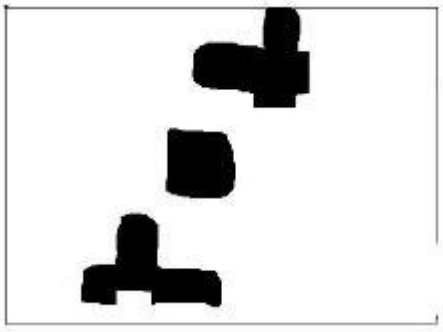
Third one is Input Parameter. To declare input parameter in advance is very typical job and that parameters effects efficiency as well as quality of clusters. So either there should be an efficient way to tell these parameters or there should no predefined parameters. DBSCAN requires two parameters while DBCLASD requires no parameter and DENCLUE requires two input parameters.

Fourth parameter is Handling of Noise as noise increases DENCLUE performs very well and DBCLASD 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 [6] [7] [8] [9] value. DENCLUE have highest disagreement value of F score and after that DBSACN and then DBCLASD follows.

Sixth and last parameter is run time of algorithms. DENCLUE having least run time, after that DBSCAN's

run time. DBCLASD's run time is nearly equal to three times the run time of DBSCAN.



σ = 0.2 , ξ = 0.01

Fig 2 Diagrammatic representation of cluster shapes in DENCLUE

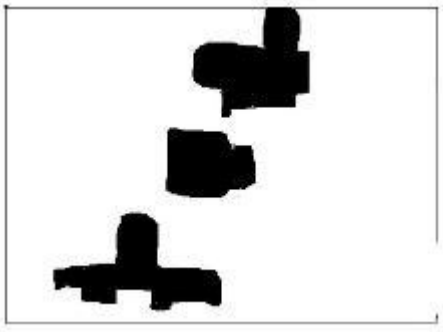


Fig 3 Diagrammatic representation of cluster shapes in DBCLASD

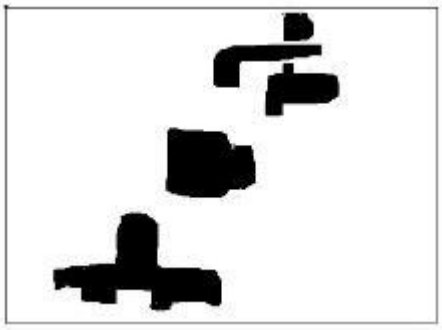


Fig. 4 Diagrammatic representation of cluster shapes in DBSCAN

5. CONCLUSION

Following table shows result of this comparative study for three algorithms using six parameters.

Table 1 Comparative analysis of three density based algorithms

Name Of the Algo-rithm	Comp Lexity	Shape Of Clust ers	Input Para-meters	Hand - ling Of noise	Clus-Ter Quali-Ty (F Disag)	Ru n Tim e(m s)
DBSC-AN	$O(n^2)$	Arbit-rary	Two Input Para-meters	Not Very well	8.7%	50
DBCL-ASD	$O(3n^2)$	Arbit-rary	No Input Para-meters	Good	5.8%	130
DENC -LUE	O(log|D|)	Arbit-rary	Two Input Para-meters	Very well	15.94 %	31

Above table shows that run time of DENCLUE algorithm is lowest while DENCLUE having highest run time. In terms of cluster quality DBCLASD leads while DENCLUE is lacking behind.

Our paper is useful in finding which density based clustering algorithm is suitable in different situations. For example where time does not matter and cluster quality is required there DBCLASD algorithm can be used. In other

Scenario noise and outliers are not of concern and still quality and time matters then DBSCAN is appreciable. Where run time is most important factor there DENCLUE will be the best option.

6. ACKNOWLEDGMENTS

My sincere thanks to Mrs. Priyanka Ahlawat Mann for her guidance in formulating the paper.

7. REFERENCES

[1] A.K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[2] Ester M. Kriegel H.-P., Xu X.: "Knowledge Discovery in Large Spatial Databases: Focusing Techniques for efficient Class Identification", Proc. 4th Int. Symp. on large Spatial Databases, Portland, ME, 1995, in: Lecture Notes In Computer Science, Vol. 951, Springer, 1995, pp. 67-82.

[3] XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J. 1998. A distribution-based clustering algorithm for mining in large spatial databases. In Proceedings of the 14th ICDE, 324-331, Orlando, FL.[10]A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, CM, 31 (1999), pp. 264–323.

[4] A. Hinneburg 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.

[5] Linsay I Smith, "A tutorial on Principal components Analysis",Retrived June 10,2007, http://csnet.otago.ac.nz/cosc453/stydent_tutorials/principal_components.pdf , pp.12-20

[6] McCallum, A, K. Nigam, and L.H. Ungar. Efficient Clustering of High-dimensional Data Sets with Application to Reference Matching. in Knowledge Discovery and Data Mining. 2000.

[7] Rijsbergen, C.J.v., Information Retrieval. 1975: Butter Worths

[8] Steinbach,M.,G. Karypis, and V. Kumar. A comparison of document clustering techniques in Text Mining workshop,KDD 2000

[9] Zhao, Y. and G. Karypis. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. in CIKM. 2002. McLean, Viginia.