International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-2, December 2012

Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm

Chaudhari Chaitali G.

Abstract- ***Clustering is the process of organizing similar objects into the same clusters and dissimilar objects in to different cluster. Similarities between objects are evaluated by using the attribute value of object, a distance metric is used for evaluating dissimilarity. DBSCAN algorithm is attractive because it can find arbitrary shaped clusters with noisy outlier and require only two input parameters. 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. But both DBSCAN and PDBSCAN algorithms are sensitive to the initial parameters.***

Keyword: Clustering, DBSCAN, PDBSCAN, Ant clustering algorithm

I. NTRODUCTION

Clustering is the process of dividing data in to groups such that objects in the same clusters are as similar as possible and different clusters are as dissimilar as possible. Clustering algori- thms can widely applied in many fields including: Information retrieval, pattern reco- gnition, machine learning, image processing and so on.

Clustering algorithms are classified in to four categories: a) Partitional clustering b) Density-based and grid based clustering c) Hierarchical clustering d) Other clustering

Traditional optimization algorithms are used with clustering algorithms to improve clustering effect. This algorithm includes: greedy algorithm, exhaustive search algorithm, local search heuristics, method of dynamic programming and so on. Traditional optimization algorithms are designed for specific issues and only effective on certain type of issues.

New optimization algorithm have been pro- posed, such as simulated annealing (SA), particle swarm optimization (PSO), ant clustering algorithm (ACA), genetic algorithm (GA).

II. II.RELATED WORK

Clustering is the unsupervised classification of patterns in to groups. There are four categories of clustering: partitional clustering, density based and grid based clustering, hierarchical clustering.

Manuscript published on 30 December 2012.
***** Correspondence Author (s)
Chaitali Chaitali, Department of Information Technology Parul Institute of Engg. And Tech Gujarat Technological university, Gujarat, India.

Other clustering algorithms.

Partitional clustering such as, k-means algorithm, can only discover spherical clusters. It is sensitive to the noise and the center points. Hierarchical clustering algorithms can be used to discover well-separated isotropic clusters. Hierarchical algorithms do not usually consider the attributes of spatial objects. Density based algorithm aims to identify dense regions that are separated by low density region. Two most advantages of density based algorithms are that they can discover clusters of arbitrary shape and they do not require a user to input the number of clusters. Grid based algorithm like WaveCluster and STING, usually employ a gridded data structure to enhance the effiency of clustering.

A. DBSCAN: Density based spatial clustering algorithm with noise

Density-based clustering defines cluster as region, the objects of the region are dense. The clusters are separated from one another by low-density regions. The reason we choose density-based clustering is that it has significant advantages over partitional and hierarchical clustering algorithms. It can discover clusters of arbitrary shapes. The computational complexity can be reduced to O(n*l*gn) by building some special data structures. In addition it is able to effectively identify noise points. But density-based clustering algorithms easily lead to memory problem when facing large databases. Some researches show that current density-based clustering algorithms often have difficulties with complex data sets in which the clusters are different densities.

The key idea in DBSCAN is that for each data object of a cluster, the neighborhood of a given radius (Eps) has to contain at least a minimum number (MinPts) of objects.

Eps-neighbourhood of point: The Eps-neighborhood of a point p, denoted by NEps(p), is defined by NEps(p) = {q 2 Djdist(p,q) 6 Eps}.

Directly density-reachable: An object p is directly density-reachable from an object q wrt. Eps and MinPts in the set of objects D if

(1) p 2 NEps(q) (NEps(q) is the Eps-neighborhood of q),
(2) jNEps(q)jPMinPts (Core point condition).

Core object & border object: An object is core object if it satisfies condition (2) of Definition 2, and a border object is such an object that is not a core object itself but is density-reachable from another core object.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

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

Density-connected: An object p is density-connected to an object q wrt. Eps and MinPts in the set of objects D if there is an object o 2 D such that both p and q are density-reachable from o wrt. Eps and MinPts in D.

Cluster: Let D be a database of points. A cluster C wrt. Eps and MinPts is a non-empty subset of D satisfying the following conditions:

(1) "p, q: if p 2 C and q is density-reachable from p wrt. Eps and MinPts, then q 2 C.
(2) "p, q 2 C: p is density-connected to q wrt. Eps and MinPts.(Connectivity).

Noise: Let C1,... ,Ck be the clusters of the database D wrt. parameters Epsi and MinPtsi,i = 1,...,k. Then the noise is the set of points in the database D not belonging to any cluster Ci, i.e. noise = {p 2 Dj"i : p R Ci}.

DBSCAN algorithm starts from an arbitrary point q, and retrieves all points density-reachable from q wrt. Eps and MinPts. If q is a core point, create a new cluster and assign the point q and its neighbors into this new cluster. Then the algorithm iteratively collects the neighbors within Eps distance from the core points. The process is repeated until all of the points have been processed.If q is a border point, no points are density-reachable from q and DBSCAN visits the next point of the database.

B. PDBSCAN: Partitioning based DBSCAN algorithm

PDBSCAN run DBSCAN algorithm on each partition which is partitioned by special rules. With PDBSCAN, the R⁄-tree should be built. DBSCAN requires to specify 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.

The steps of PDBSCAN are as follow:

(1) Partitioning the initial database into N partitions.
(2) For each partition, building local R⁄-tree, analyzing and selecting local Eps and MinPts, and then clustering it with DBSCAN.
(3) Merging the partial clusters.

In the first step of PDBCAN, some articles partition database over the data dimensions. This method will lead to many problems.

III. ANT CLUSTERING ALGORITHM

Ant clustering algorithm defines two dimensional grid. Size of grid is dependent on number of object. This algorithm scatters the data objects onto this grid and makes a group of agent ants work on this two-dimensional grid at the same time.

Every agent ant on the grid will occur in the following two situations:

(1) An agent ant holds an object i and evaluates the probability of dropping it on its current position;
(2) An agent ant is unloaded and evaluates the probability of picking up an object i.

ACA algorithm can cluster data without any initial knowledge. The agent ants just pick up and drop the objects influenced by the similarity and density of the objects within the agent ant's current neighborhood. The probability of picking up an object increases with low density neighborhoods, and decreases with high similarity among objects in the surrounding area.

IV. PROBLEM OF PARTITIONING APPROACH IN PDBSCAN:

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. If a database is partitioned at random, the results of second step that run DBSCAN algorithm with every data partition will be affected, and the third step that merge the partial clusters will be difficult.

Partitioning technique has three serious defects:

(1) The algorithm has to check the situation of every dimension and plot histogram for every dimension, this process not only costs much time but also reduces the accuracy of algorithm.
(2) It is quite difficult to merge the partial clusters in the third step because of the multi-dimensional data.
(3) When the shape of database is ring type or spirality entangled together, this approach that make data project to one dimension cannot be effective.

V. NEW ALGORITHM BASED ON PARTITIONING-BASED DBSCAN WITH ANT CLUSTERING

The new algorithm divides data according to point density use modified ant clustering algorithm to partition multi dimensional data to reduce time cost and improve accuracy.

A. Partitioning method based on point density

γ -neighborhood of a point: The c-neighborhood of a point i, denoted by Nc(i), is defined by Nc (i) = {q 2 Djdist(q, i) 6 c}.

Point density, or PD: The relative density with point i, denoted by Den(i), is defined by Den(i) = jNc(i)j/jDj. The steps of the PD-based partitioning method are as follow:

(1) Set the initial parameter N (N is the number of partitions);
(2) Calculate the Den(i) for each object i;
(3) Run the k-means algorithm to cluster data into N partitions based on the value of Den(i).

B. Partitioning method based on modified ant clustering algorithm

To deal with high-dimensional data, we give another partitioning method which combine modified ant clustering algorithm and PD-based partitioning method (PACA).

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The PACA algorithm first uses modified ant clustering algorithm to present multi-dimensional data on a two-dimensional grid, then PD-based partitioning method is employed to calculate and partition the objects. Because this partitioning method need not consider each dimension respectively as usually done, so it save much time and improve the accuracy.

C. The new hybrid method based on partitioning-based DBSCAN and ant clustering (PACA-DBSCAN)

Sections A and B have presented the PD-based partitioning method and PACA partitioning algorithm. These two methods can divide data points with similar density into the same area. Our new PACA-DBSCAN algorithm will employ one of these two partitioning methods according to the number of data dimension. If the data is two-dimensional, the algorithm uses PD-based partitioning method to partition data directly. Else if the data is multi-dimensional, the algorithm will partition data with PACA algorithm. Then for each partition, PACA-DBSCAN algorithm builds R*-tree, plots k-dist graph and runs DBSCAN algorithm. At last, the partial clusters will be merged based on predetermined rules. The PACADBSCAN algorithm makes a better use of the advantages of both PD-based partitioning and PACA partitioning.

VI. EXPERIMENTAL STUDIES

To test the effectiveness of PACA-DBSCAN algorithm we use five datasets. All datasets run with DBSCAN, PDBASCAN and PACADBSCAN algorithm.

A. datasets

Datasets are dataset1, dataset2 and dataset3. They are summarizing in table.

Data set	No.of data classes	No. of features	Size of data set
Data set1	3	2	500(257,32,211)
Data set2	4	2	1572(528,348,272,424)
Data set3	7	2	1043(343,30,38,241,72,157,155)

Table 1: datasets

All datasets are describe as follow:

1) Data set1(n=500,d=2,k=3): This is an artificial data set. It is a two-featured problem with three unique classes and 500 patterns. In order to prove the superiority of the new algorithm that it can be applied to non-uniform density of data and less sensitive to the input parameters, the shapes of three clusters is irregular.
2) Data set2 (n=1572,d=2,k=4): This is an artificial data set. It is a two-featured problem with four classes and 1572 patterns. These data is obtained using a data generator for multivariate Gaussian clusters. It can prove that the result of
PACA-DBSCAN is better than PDBSCAN and DBSCAN on standard Gaussian data set.
3) Data set3(n=1043,d=2,k=7) : this is an artificial data set. It is a two-featured problem with four classes and 1043 patterns. These data are obtained using a data generator. There are two kinds of shapes in Artset3. One is spiral, and another is annular. The data set also includes some noise points. We create it to show that PACA-DBSCAN can overcome the drawback of PDBSCAN which cannot recognize the special shape of data.

B. Experimental results:

In this experiment, three datasets are normalized according to the equation-1 and used with DBSCAN, PDBSCAN and PACA-DBSCAN algorithms.

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Compare the performance of DBSCAN, PDBSCAN and PACA-DBSCAN algorithm in terms of classical f-measure and a proposed criterion that is called Eps rate(ER):

1) F-measure: It is based the precision and the recall from the information retrieval The precision and the recall are defined as:

$$p(i,j) = \frac{nij}{nj}, \qquad \mathrm{r(i,j)} = \frac{\mathrm{nij}}{\mathrm{ni}}$$

Where,
Each class I = set of n_i items desired for query
Each cluster j= set of n_j items retrieved for a query
For class i and a cluster j, the F-measure is defined as:

$$F(i,j) = \frac{(b^2+1).p(i,j).r(i,j)}{b^2.p(i,j) + r(i,j)}$$

Take b=1 for equal weighting to p(i,j) and r(I,j). F-measure for dataset of size n is given by,

$$F = \sum_i \frac{ni}{n} \max\{F(i,j)\}$$

So, the bigger the F-measure is the better tha clustering algorithm is.

2) Eps rate (ER): it shows the sensitivity of DBSCAN algorithm to the parameter Eps.
For each subset of data ERi=

$$\frac{The\ scope\ of\ corresponding\ the\ optimal\ result}{the\ scope\ of\ Eps\ of\ partition\ i}$$

The bigger ER is the less sensitive to Eps the algorithm is.DBSCAN,PDBSCAN and PACA- DBSCAN need two parameter Eps and MinPts.Set minPts as 4. To set Eps we need to build R*-tree.

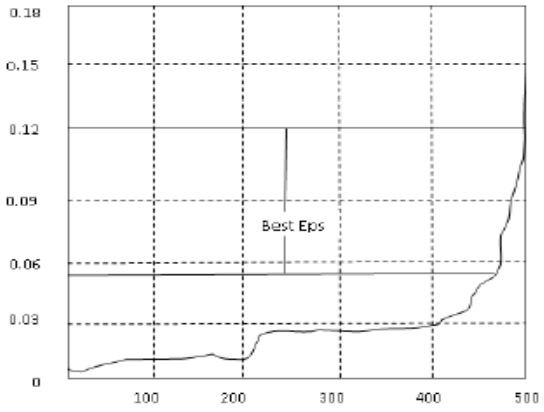

Figure 1: Dataset 1

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

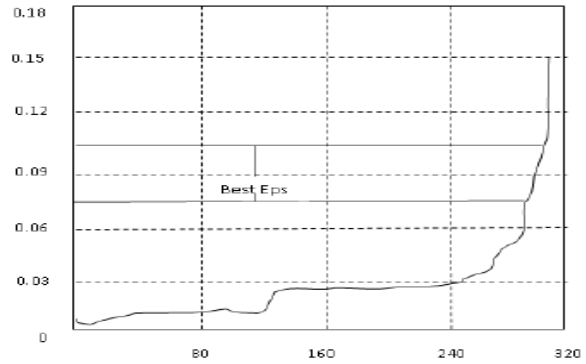

Figure 2: first partition of dataset 1

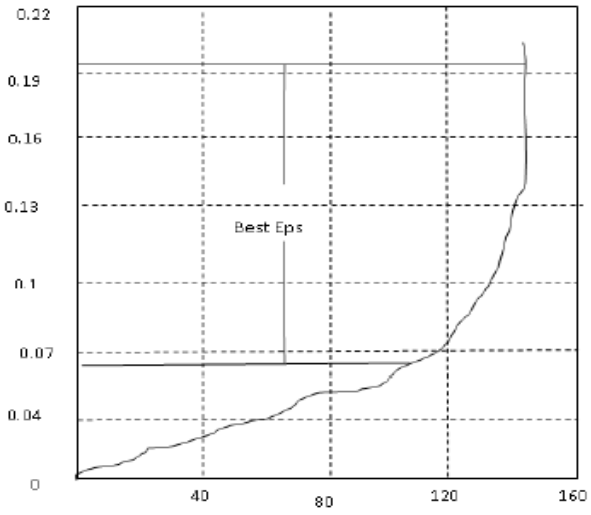

Figure 3: second partition of dataset 1

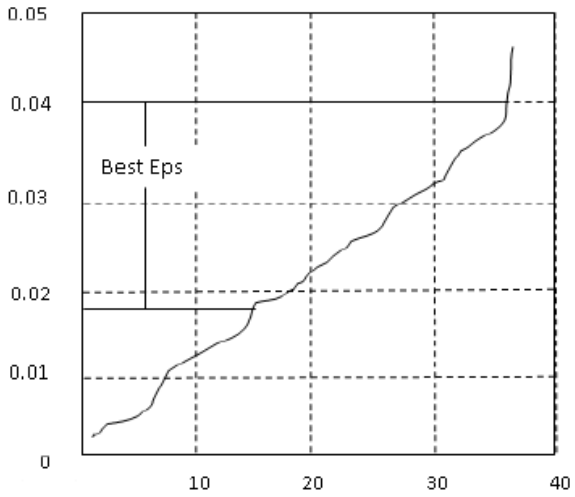

Figure 4: Third partition of dataset 1

Table: 2 ER of artificial data set 1

Algorithm	Data set	Scope of Eps	Total scop of Eps	ER of each area	final ER
DBSCAN	Dataset1	[0.05-0.12]	[0.02-0.163]	0.49	0.49
PACA- DBSCAN	Dataset1-1	[0.065-0.1]	[0.005-0.163]	0.22	
	Dataset1-2	[0.065-0.191]	[0.015-0.197]	0.692	
	Dataset1-3	[0.16-0.4]	[0.054-0.455]	0.599	0.503

Table 3: compairition of three algorithms

Data set	DBSCAN	PDBSCAN	PACA-DBSCAN
Dataset1 F-measure ER	0.0689 0.49	0.0689 0.497	0.996 0.503
Dataset2 F-measure ER	0.99 0.002	1.00 0.291	1.00 0.396
Dataset3 F-measure ER	0.551 0.114	0.608 0.216	0.962 0.425

VII. CONCLUSION

This paper present a new algorithm based on partitioning-based DBSCAN and Ant-clustering. This algorithm can partition database in to N partitions according to the density of data. New PACA-DBSCAN algorithm reduces the sensitivity to the initial parameters and also can deal with data of uneven density. This algorithm does not need to discuss the distribution of data on each dimension for multidimensional data. PACA-DBSCAN algorithm can cluster data of very special shape. To evaluate the performance of proposed algorithm we use 3 data set to compare with other algorithms. Comparison of three algorithm is shown in table-3. The result of PACA-DBSCAN algorithm is compared by F-measure and a proposed criterion. From experimental results that is proved that performance of PACA-DBSCAN algorithm is better than DBSCAN and PDBSCAN.

REFERENCES

[1] Birant, D., & Kut (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data and Knowledge Engineering, 60, 208–221.
[2] Dalli, A. (2003). Adaptation of the F-measure to cluster-based Lexicon quality evaluation. In EACL 2003. Budapest.
[3] Huang Darong, Wang Peng, Grid-based DBSCAN Algorithm with Referential Parameters
[4] Cao, F., Ester, M., Qian, W., & Zhou, A. (2006). Desity-based clustering over an evolving data stream with noise. In 2006 SIAM conference on data mining, Bethesda (pp. 328–339).
[5] Handl, J., & Meyer, B. (2007). Ant-based and swarm-based clustering. Swarm Intellegince, 1, 95–113.
[6] Viswanath, P., & Pinkesh, R. (2006). l-DBSCAN: A Fast Hybrid Density Based Clustering Method. Pattern Recognition, 1, 912–915.