Advances in Wireless and Mobile Communications.
ISSN 0973-6972 Volume 10, Number 5 (2017), pp. 933-942
© Research India Publications
http://www.ripublication.com

A Hybrid Framework using Fuzzy if-then rules for DBSCAN Algorithm

Dr. Anjali B. Raut
Head,
Department of Computer Science & Engineering
H.V.P.M'S, C.O.E.T., Amravati. (MS) India.

Abstract

DBSCAN is a density-based clustering algorithm. This algorithm clusters data of high density. For finding core objects traditional DBSCAN uses this core object as center core which extends outwards continuously. As core objects are growing, the unprocessed objects which are retained in memory, will occupy a lot of memory and I/O overhead which tends to low efficiency of algorithm. A data mining technique which is applied for large databases, a DBSCAN works on the basis of bivalent logic. Hence it can only identify objects which completely belonging to a particular cluster or not wholly belonging to it. In this paper, a framework of methodology of DBSCAN algorithm with the integration of fuzzy logic is presented. The improved version is hybridization of DBSCAN algorithm with fuzzy if-then rules.

Keywords: Data mining, DBSCAN, fuzzy logic, Clustering

I. INTRODUCTION

The term data mining is often used to refer to the entire knowledge discovery process perhaps because the term is shorter than knowledge discovery from data (KDD). Therefore, we adopt a broad view of data mining functionality: Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. Data mining includes examining data analysis, data discovery as well as deductive learning. Data mining can be categorized as predictive and descriptive. As

Partitioning and hierarchical methods are designed to find spherical-shaped clusters. They have difficulty finding clusters of arbitrary shape such as the “S” shape and oval clusters.

For such data it is necessary to identify regions in which noise or outliers are included in the clusters. Also we can have clusters as dense regions in the data space, separated by sparse regions for finding clusters of arbitrary shape. A density-based clustering method has this as main strategy for discovering clusters of non spherical shape.

In density based clustering algorithms, clustering process is done on the basis of categorization of points as border points, core points and noise points identified by the algorithm.

The directly density-reachable points using the Є threshold are provided by the user for each point of the dataset. Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a density based clustering algorithm that is very sensitive to input parameters. These parameters are difficult to decide. The run time complexity of DBSCAN algorithm is O(n2) and it has to be determined for each point. The computational complexity of DBSCAN algorithm is O(n log n) in a spatial index, where n is the total number of objects in the database. It requires two user defined parameters: Є, the radius that delimitate the neighborhood region of a particular point, and MinPts, the minimum number of points that are required in the Є-neighborhood.

II. LITERATURE REVIEW

Ester et al. [1] present DBSCAN algorithm in order to discover clusters of varying shape. Clustering algorithms generally make use of a similarity measure which is based on a distance metric. These algorithms divide the database in such a way that points which are in the same partition are more related to each other. Tsai et al. [10] propose a clustering method that executes fast than a Fast Self-Organizing Map (FSOM) merged with Genetic k-means algorithm (GKA) and k-means algorithm.

Different variants of DBSCAN have been studied. S.K.Popat et al. [3] described the concept of clustering as an automatic learning technique that groups similar objects into one cluster. X. P. Yu et al. [2] propose a new DBSCAN based on k-nearest neighbors (KNN). It combines KNN and DBSCAN so as to improve DBSCAN. In 2007, Liu et al. [4] proposed Varied Density Based Spatial Clustering of Applications with Noise (VDBSCAN) to examine the data set having different densities.

W.H. Wolberg et al. [8] analyze the analysis of breast cytology to exhibit the applicability to decision making and medical diagnosis. Some characteristics were evaluated. Thanh et al. [9] describe that the results of clustering obtained from DBSCAN is dependent on the order of processing of objects. A modified version of DBSCAN solves this issue for those data sets that contain dense composition with connected clusters.

A.Ram et al. [12] propose an algorithm that handles the local density variation inside a particular cluster by calculating density mean as well as variance of a particular cluster for any of the core objects. This algorithm gives optimized results.

K. Khan et al. [11] describe DBSCAN is one such density based algorithm that is used to extract meaningful pattern from different data sets. A. Ghanbarpour et al. [6] describe Extension of Density Based Spatial Clustering of Applications with Noise (ExDBSCAN) as an extension of DBSCAN-based method to cover multi -density data sets. It detects clusters with different densities as well as detects outliers correctly.

In order to ensure the high efficiency of DBSCAN clustering algorithm, and reduce its memory footprint. The original DBSCAN algorithm was improved, and the G-DBSCAN algorithm is proposed [13].

The DBSCALE algorithms are used for clustering of very large database. The clustering techniques are very proficient and also the rate of correctness is increases, but these algorithms suffered from noise and outlier problem. For the minimization of noise and outlier we modified DBSCALE algorithm using Naïve's Baye's theorem. According to this techniques, it compute maximum posterior hypothesis for the outlier data [5].

III. A HYBRID FRAMEWORK

3.1. The Basic Concept of DBSCAN

Each point in the density-connected set is density-reachable. If any point P which is not classified, choose that P, then check whether P is the core point. If the point is the core point, find all points, they are directly density-reachable from object P. With these points, create new cluster and assign ID to each of this cluster. If P is a boundary object, then continue to access the next data point. Continue this process until all points have been processed. Finally, no ID points as noise points.

The density of an object o can be measured by the number of objects close to o. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) which finds the core objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters.

A user-specified parameter Є > 0 is used to specify the radius of a neighborhood we consider for every object. The Є-neighborhood of an object o is the space within a radius Є centered at o. Due to the fixed neighborhood size parameterized by Є, the density of a neighborhood can be measured simply by the number of objects in the neighborhood. To determine whether a neighborhood is dense or not, DBSCAN uses another user-specified parameter, MinPts, which specifies the density threshold of dense regions. An object is a core object if the Є-neighborhood of the object contains at least MinPts objects. Core objects are the pillars of dense regions.

Given a set, D, of objects, we can identify all core objects with respect to the given parameters, Є and MinPts. The clustering task is therein reduced to using core objects and their neighborhoods to form dense regions, where the dense regions are clusters. To connect core objects as well as their neighbors in a dense region, DBSCAN uses the notion of density-connectedness.

“How does DBSCAN find clusters?” Initially, all objects in a given data set D are marked as “unvisited.” DBSCAN randomly selects an unvisited object p, marks p as “visited,” and checks whether the Є-neighborhood of p contains at least MinPts objects. If not, p is marked as a noise point else new cluster C is created for p, and all the objects in the Є-neighborhood of p are added to a candidate set, N. Algorithm iteratively adds those objects in N to C, that do not belong to any cluster. In this process, for an object p0 in N that carries the label “unvisited,” DBSCAN marks it as “visited” and checks its Є-neighborhood. If the Є-neighborhood of p0 has at least MinPts objects, those objects in the Є-neighborhood of p0 are added to N. DBSCAN continues adding objects to C until C can no longer be expanded, that is, N is empty. At this time, cluster C is completed, and thus is output.

To find the next cluster, DBSCAN randomly selects an unvisited object from the remaining ones. This process of clustering is continues till all objects are visited. The algorithm of the DBSCAN is given as below.

Algorithm: DBSCAN

Input:

D: a data set containing *n* objects,

Є: the radius parameter, and

MinPts: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
mark all objects as unvisited;
do
randomly select an unvisited object p;
mark p as visited;
if the Є-neighborhood of p has at least MinPts objects
create a new cluster C, and add p to C;
let N be the set of objects in the Є-neighborhood of p;
for each point p0 in N
if p0 is unvisited
```

```
mark $p0$ as visited;
if the Є-neighborhood of $p0$ has at least $MinPts$ points,
add those points to $N$;
if $p0$ is not yet a member of any cluster, add $p0$ to $C$;
end for
output $C$;
else mark $p$ as noise;
until no object is unvisited;
```

The main objective is to produce an improved DBSCAN algorithm by the hybridization of existing DBSCAN algorithm with fuzzy if-then rules. It will be based on multivalent logic by making use of the membership values

Step 1: Training data set will be loaded.
Step 2: DBSCAN algorithm works in the following way

Begin with an arbitrary starting point that is unvisited. Retrieve the ε-neighborhood. If it has enough points, a cluster is begins to form. If it does not have sufficient points, then the point is considered as noise. If any point is brought into being a part of a dense cluster, then Є-neighborhood of that point is also considered to be an element of it only. All the points found inside the Є-neighborhood are combined, just like their own Є-neighborhood when they are also dense. Repeat the above process till the density-connected cluster has been wholly created. Result is the detection of a cluster or noise.

3.2. Fuzzy If-Then Rules

Fuzzy set theory is a generalization of classical set theory. If-Then rule statements are used to formulate the conditional statements that comprise fuzzy logic.
A single fuzzy If-Then rule assumes the form

If p is a1 Then q is a2

where a1 and b2 are linguistic variables defined by fuzzy sets on the ranges (i.e. universe of discourse) X and Y respectively. The If part of the rule ‘p is a1’ is called the antecedent or premise and the Then-part of the rule ‘q is b2’ is called the consequent. The conditional statement can be expressed in a mathematical form

If a1 Then b2 or a1 → b2

Statements of if-then rules formulate the conditional statements. These conditional statements consist of fuzzy logic.

Step 1: Missing values of testing data set will be detected
Step 2: Apply fuzzy based DBSCAN algorithm on that data set

Step 3: Obtain the results of the missing values
Step 4: Evaluate the following parameters namely Accuracy, Geometric accuracy, Bit error rate, Specification, Sensitivity, Error rate, fmeasure and Execution Time

3.3. Fuzzy Based DBSCAN

The flowchart of Fuzzy Based DBSCAN is illustrated in Figure 1.

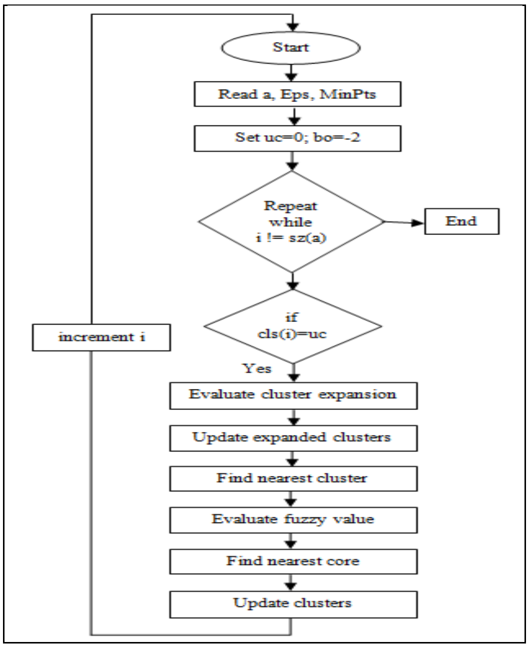

Figure 1: Flowchart of Fuzzy Based DBSCAN

The pseudocode of Fuzzy Based DBSCAN is as follows [9].

```
Begin fdbscan(a,€,MinPts)
uc= 0;
bo= -2;
n = sz(a,1);
cls = zr(n,1);
CId = 1;
for i=1:n
if cls(i) == uc
[ExpandClusterReturn cls]=
ExpandCluster(a,cls,i,CId,€, MinPts);
if ExpandClusterReturn
```

```
CId = CId +1;
end
end
end
cls_core = cls; core_idx = search(clscore > 0);
bo_pts = search(cls==bo);
for i=1:len(bo_pts)
currB = bo_pts(i);
d=dist(+a(currB,:),+a(core_idx,:),1);
[tmp nst_core]=fz_val(d);
nst_core_idx=core_idx(nst_core);
cls(currB)=cls(nst_core_idx);
end
end
Return cls cls_score
```

```
Begin [ExpandClusterReturn cls]=
ExpandCluster(a,cls,i,CId, Є, MinPts)
uc = 0;
ns = -1;
bo = -2;
d=dist(+a(i,:),+a(:,:),1);
sd = search(d < Є);
if sz(sd,2) < MinPts,
cls(i) = ns;
ExpandClusterReturn = 0;
else
while ~isemp(sd)
currP = sd(1);
d=dist(+a(currP,:),+a(:,:),1);
rslt = search(d <= Є);
if len(rslt) >= MinPts
cls(currP) = CId;
rslt_uc = rslt(search(cls(rslt)==uc));
rslt_ns = rslt(search(cls(rslt)==ns));
cls([rslt_uc rslt_ns]) = bo;
sd = union(sd,rslt_uc);
end
sd = sd(2:size(sd,2));
```

```
end
ExpandClusterReturn = 1;
end
end
Return
Begin Fz_val
If the dist> mean_val
Add to nst
Else if dist> median_val
Add to nst
End
Return tmp nst_core
```

IV. ADVANTAGES AND DISADVANTAGES OF DBSCAN

Advantages:

1. Algorithm can find clusters of arbitrary shapes and sizes, automatically determine the number of clusters, isolated noise points, high efficiency and one scan can complete the clustering.
2. Does not require a-priori specification of number of clusters.
3. It can even find clusters completely surrounded by (but not connected to) a different cluster.
4. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.
5. DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database.

Disadvantages:

1. In the process of clustering, DBSCAN once found the core object, then this core object as the center outward expansion, this process will continue to increase core objects, unprocessed objects are retained in memory.
2. If a large cluster exists in the database, it will require a lot of memory to store the core object information.
3. DBSCAN algorithm fails in case of varying density clusters.
4. Input parameters sensitive.
5. Parameter Є, MinPts difficult to determine.

V. CONCLUSION

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. It is attractive for the task of class identification in spatial databases. However, the other well-known algorithms suffer from severe drawbacks when applied to large spatial databases. The clustering algorithm DBSCAN relies on a density-based notion of clusters. For determining an appropriate value for it, DBSCAN takes only one input parameter. In this paper DBSCAN algorithm identifies objects based on bivalent logic. The existing techniques have not focused on the hybridization of DBSCAN with fuzzy if then rules. DBSCAN will be combined with fuzzy if then rules. The hybridization will allow DBSCAN to decide the cluster in more efficient manner.

REFERENCES

[1] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. 1996.

[2] X. P. Yu, D. Zhou, and Y. Zhou, “A New Clustering Algorithm Based on Distance and Density,” presented in proceedings of International Conference on Services Systems and Services Management (ICSSSM- 2005), Vol. 2.I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Popat, Shraddha K., and M. Emmanuel, "Review and Comparative Study of Clustering Techniques," presented in International Journal of Computer Science and Information Technologies. 2014.

[4] P. Liu, D. Zhou, and N. J. Wu,“VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise,” in proceedings of IEEE International Conference on Service Systems and Service Management, Chengdu, China, pp 1-4, 2007.

[5] Agrawal, Jitendra, Sanyogita Soni, Sanjeev Sharma, and Shikha Agrawal. "Modification of Density Based Spatial Clustering Algorithm for Large Database Using Naive Bayes' Theorem." In Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, pp. 419-423. IEEE, 2014.

[6] Ghanbarpour, Asieh, and Behrooz Minaei, "EXDBSCAN: An extension of DBSCAN to detect clusters in multi-density datasets." In Intelligent Systems (ICIS), 2014 Iranian Conference on, pp. 1-5. IEEE, 2014.

[7] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[8] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

[9] Thanh N. Tran*, Klaudia Drab, Michal Daszykowski, "Revised DBSCAN algorithm to cluster data with dense adjacent cluster", Chemometrics and Intelligent Laboratory Systems, 120:92:96., DOI: 10.1016 /j.chemolab.2012.11.006

[10] Tsai, Cheng-Fa, Han-Chang Wu, and Chun-Wei Tsai. "A new data clustering approach for data mining in large databases." Parallel Architectures, Algorithms and Networks, 2002. I-SPAN'02. Proceedings. International Symposium on. IEEE, 2002.

[11] Khan, Kamran, Saif Ur Rehman, Kamran Aziz, Simon Fong, S. Sarasvady, and Amrita Vishwa, "DBSCAN: Past, present and future." In Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the, pp. 232-238. IEEE, 2014.

[12] Ram, Anant, et al. "A density based algorithm for discovering density varied clusters in large spatial databases." Int. J. of Computer Applications 3.6 (2010): 1-4.

[13] International Journal of Database Theory and Application” MRG-DBSCAN: An Improved DBSCAN Clustering Method Based on Map Reduce and Grid” Vol.8, No.2 (2015), pp.119-128 http://dx.doi.org/10.14257/ijdta.2015.8.2.12