Volume 4, Issue 4, April 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

DBSCAN – Cluster Analysis of Spatial Data and Outlier Detection

Ankit Punia, M.Tech Student
DCSA, M. D. University,
Rohtak, India

Ms. Pooja Mittal,
Assistant Professor
DCSA, M. D. University, Rohtak, India

Abstract—Clustering is a major data mining technique for discovering trends in large databases. Outlier detection is a method that finds data objects that are inconsistent to the remaining data in the cluster. In this paper, we present DBSCAN algorithm which can deal with clusters of different densities and performs the clustering. It allows identifying clusters of arbitrary shapes and present outlier detection with stratification which ranks the different densities object. We show an implementation of this algorithm using Weka tool and present the data mining results.

Keywords: Data Mining, Clustering, DBSCAN, Algorithm

I. INTRODUCTION

Clustering is the method of partitioning of data into groups based on similar properties. A Cluster is a collection of data objects having similar properties. The data objects with in the same cluster have strong association between them and weak association between the data objects of different clusters. Clustering is an unsupervised learning so no training sample is available to partition the data. In this paper we focus on various cluster analysis techniques.

II. RELATED WORK

Major clustering methods has classified into partitioning, hierarchical, density-based and grid-based.

Partitioning methods divides the n elements into k partitions where each partitions represents a cluster and $k \leq n$ and each cluster contains at least one element. Partitioning methods works by first creating an early partitioning and then iteratively improves the clusters by moving objects from one partition to another. This method is suitable for finding only spherical shaped clusters in small to medium sized data sets. Examples of this method are k-means and k-mediods. K-means algorithm clusters data with the help of the mean value of the objects in the cluster. K-mediods algorithm assigns objects into clusters based on one of the objects positioned near the center of the cluster.

Hierarchical methods disassemble data objects into a set of nested clusters organized as a hierarchical tree. The two approaches, agglomerative (bottom up) and divisive (top down) decompose the given set of data objects. The agglomerative approach starts with each data object as a cluster. It recursively merges the objects or clusters of similar properties into one or until a stop condition hold. The divisive approach starts with all data objects as a single cluster. This cluster is then progressively splits into sub clusters. It stops when certain stopping conditions hold or till one data object is in each cluster.

Density based methods cluster data because their density (number of data objects or data points). These methods discover clusters of arbitrarily shaped, unlike partitioning and hierarchical methods. These methods help to filter out noise. Examples of these methods are DBSCAN, and OPTICS.

DBSCAN *(Density Based Spatial Clustering Applications with Noise)* is a density based clustering method aimed to discover clusters of arbitrary shape [6]. The key idea is that for each data object within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of data objects.

OPTICS *(Ordering Points to Identify Clustering Structure)* is a variation of DBSCAN which provides us cluster ordering of data to identify objects which are in a denser cluster.

Grid Based methods divide the object space into a finite number of cells.[6] These cells represent as a grid-like structure on which all clustering operations performed. This approach has faster processing time because it depends on the number of cells present in each dimension and not on the number of data objects. Examples are STING and CLIQUE.

STING *(Statistical Information Grid)* is a multi-resolution clustering technique [6] which divides the spatial area into rectangular cells. The cells form hierarchical like structure which stores statistical information of every attribute. The statistical parameters of the higher level cells calculated with the help of the parameters of the bottom level cells. Thus helps in processing the query effectively. The parameters include attribute independent parameters, such as count and the attribute dependent parameters such as mean, minimum, maximum. The layer which consists of small number of cells starts answering the queries. For each cell its calculated confidence interval reflects the cell's relevance to the given query. Only relevant cells examined to process the next lower level and this process repeated until the bottom layer reached [3].

CLIQUE *(CLustering In QUEst)* method is a dimension-growth subspace clustering. This technique starts at single dimensional subspaces and grows upward to higher-dimensional ones.[5] The key idea is that different subspaces may

contain different, significant clusters and, therefore, searches for groups of cluster within different subspaces of the same data set.

III. DBSCAN ALGORITHM

The basic key idea is that data objects in dense regions clustered together. The algorithm uses a fixed threshold value to decide dense regions. It discovers high density regions in space i.e. separated by low region density. The disadvantage of the algorithm is that it captures only certain types of noise when clusters of different densities exist. Unlike other clustering techniques, it does not require the predetermination of the number of clusters. It also discovers clusters of arbitrary shape in spatial databases with noise [1].

Variants of DBSCAN are:

Incremental DBSCAN acts as the core algorithm of query clustering tool. SDBDC (Scalable Density-Based Distributed Clustering method, first it works on each local site and then clusters distributed objects on global site.

Basic Definitions:

a) Ɛ-neighborhood: The neighborhood is distance between two points in a cluster. The neighborhood in a cluster is less than threshold input value, Ɛ. The neighborhood within a radius Ɛ of a given object is Ɛ-neighborhood.
b) MinPts: It presents the minimum number of data objects in any cluster.
c) Core Object: It refers Ɛ-neighborhood of an object contains at least MinPts of objects.
d) Directly density-reachable: A data object p is directly density-reachable from the data object q if p is within the Ɛ-neighborhood of q and q is a core object.
e) Density-reachable: An object p is density-reachable from the object q with respect to Ɛ and MinPts if there is a chain of objects $p_1, p_2, \ldots, p_n$, where $p_1 = q$ and $p_n = q$ such that p_{i+1} is directly density-reachable from p_i with respect to Ɛ and MinPts.[9]
f) Density Connected: An object p is density connected to object q with respect to Ɛ and MinPts if there exists an object $o \in D$.
g) Density Based Cluster: It is a set of density connected objects i.e. maximal with respect to density-reachability.
h) Border point: An object p is a border point if it is not a core object but density reachable from another core object.
i) Noise: The objects not assigned in any cluster act as noise.

The algorithm works as follows:

It first checks Ɛ-neighborhood of each point in the space. If the Ɛ-neighborhood of a point p contains more than MinPts, a new cluster created in which p acts as the core object. The algorithm iteratively gathers all the objects within Ɛ distance from the core objects. The process terminates when there is no new point to add to any cluster.[3,7]

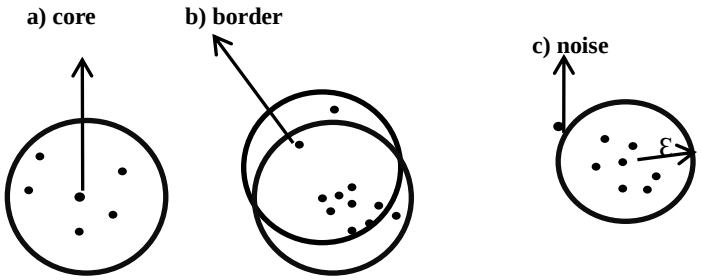

Fig. 1. a) Core, b) Border, c) Noise for MinPts 7 and radius Ɛ.

Algorithm:

Input:

D = {t_1, t_2, t_3 … t_n} // Set of elements
MinPts //Number of points in cluster
Ɛ // Maximum distance for density measure

Output:

K = {K_1, K_2, K_3 …, K_k } // Set of clusters

Method:

```
k=0;   // initially there are no cluster
for i = 1 to n do
  if t_i is not in a cluster, then
     X = {t_j | t_j is density-reachable from t_i};
     if X is a valid cluster, then
          k = k+1;
          K_k = X;
```

DBSCAN deals with outliers (data objects which are different with the remaining set of the data). The algorithm avoids the noise or outlier to insert into clusters. It's only capable to capture some types of outliers when different densities of clusters are present. This leads to the huge loss of important hidden information as sometimes the outlier are of particular interest. Examples are fraud detection, intrusion discovery.

Local Outlier Detection: This technique used to overcome the problems in analyzing different density distribution. An object is local outlier if it is outlying relative to its local neighborhood. [10]

Stratification Algorithm

The FindStrata method extracts strata from subset S. Then it checks if the average $AINFLO_k$ of the calculated strata is above δ; The objects in the last strata of S are candidate to be outlier.

```
STRATIFY(S,k)
Input
S: dataset of objects to be cluster, loaded into a data
structure having as fields data, layer, AINFLO_k , DF_k;
k: number of neighbors;
Output: Stratification of S;

n = |S|;
for i = 1 to n do
        S[i].DF_k = DF_k(S[i].data, k);
        S[i].AINFLO_k = AINFLO_k(S[i].data, k);
end for
SORT(S[1 ,,,n].data using DF_k as sorting key);
cut = 1;
layer = 1;
SORT(S[1 … n].AINFLO_k) +VAR(S[1 … n].AINFLO_k)
while cut< n do
  cut = FINDSTRATA(S,cut,layer, δ ,n);
  layer = layer + 1;
end while
return S;
```

```
FINDSTRATA(S, Start_Idx, newlayer, δ, n)
Input
S: dataset of objects to be cluster;
Start_Idx: cut point index;
newlayer: layer number;
δ : general threshold;
Output : A strata of the dataset;

newCut : = Start_Idx;
μ_DF = AVG(S[Start_Idx……n].DF_k);
while S[newCut]. DF_k< μ_DF do
   newCut = newCut + 1;
end while
μ_AINFLOk = AVG(S[Start_Idx……newCut - ].AINFLO_k);
for i = Start_Idx to newCut - 1 do
        S[i].layer = layer;
end for
μ_AINFLOk < δ then
return newCut;
end if
for i = newCut to n do
        S[i].layer = noise;
end for
```

Local Outlier Factor (LOF): The degree of outlierness with respect to the surrounding neighborhood. This factor also used to rank data objects with respect to their outlierness in the space.

The authors [11] introduce an algorithm (INFLO) to discover top-n outliers using clusters, for a given value of k. Density distribution estimated by considering neighbors. This results in outlier detection.

V. EXPERIMENTAL RESULTS

This analysis conducted on the sample of 150 data objects and 5 features used in the WEKA environment.

The results of clustering with DBSCAN were different, 4 clusters found and some objects identified as noise. DBSCAN seems unsuitable for these data objects, as two clusters are slightly covering one another and there is no clear separator between them.

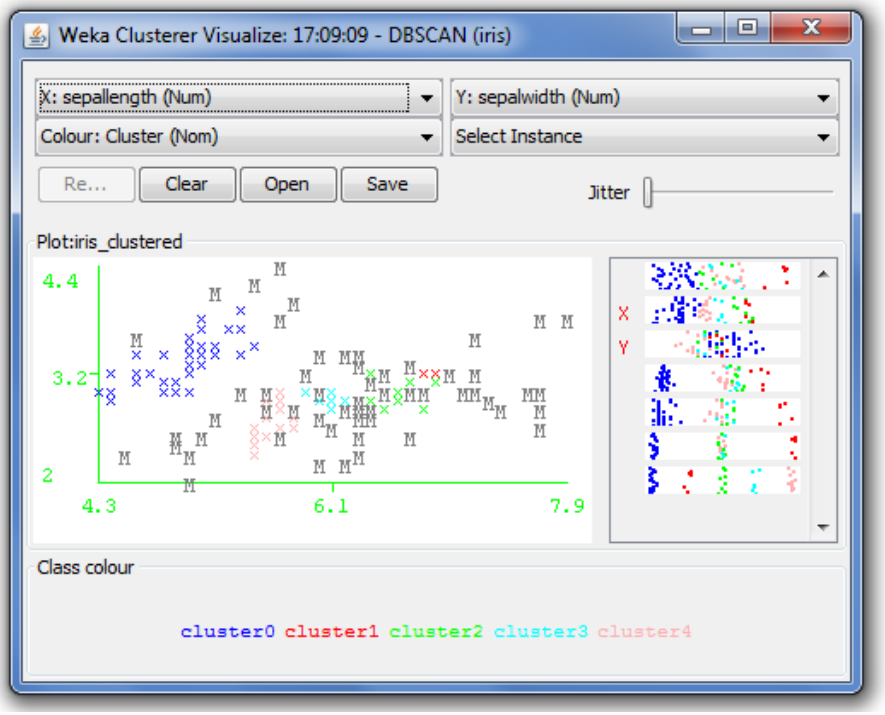

Fig. a) Epsilon = .1
minPts = 5
Clusters Generated = 4
M represents Noise in the figure

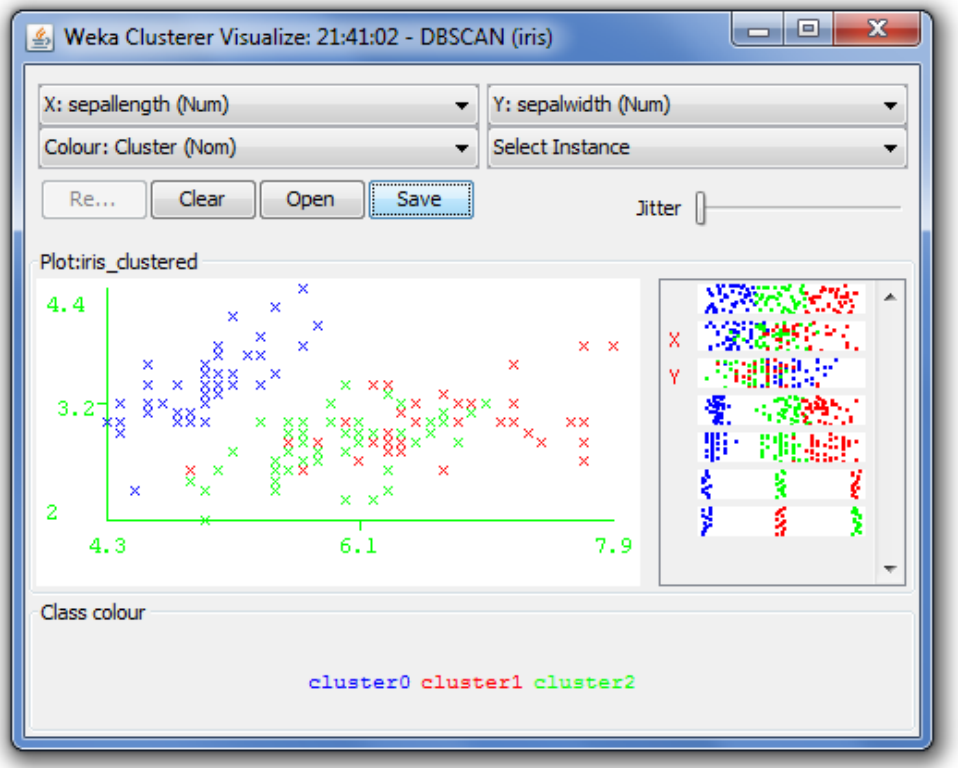

Fig. b) Epsilon = 0.4
minPts = 8
Clusters Generated = 3

In fig. c) only one large cluster of object present as the value of epsilon is much large and low density of different objects.

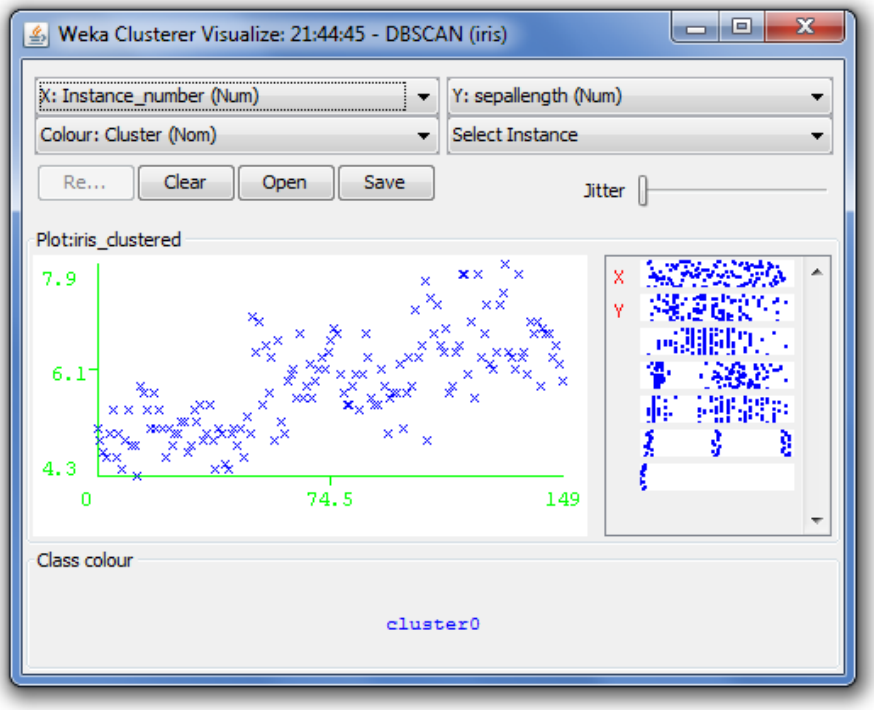

Fig. c) Epsilon = 2.0
minPts=6
Clusters Generated = 1

VI. CONCLUSION

Clustering is one of the most important techniques in data mining. It is the method of grouping data objects based on likeness within a cluster and dissimilarities between clusters. In this paper we delivered the wide classification of clustering methods such as partitioning, hierarchical, density based and grid based methods. This paper presents a density-based clustering and outlier detection algorithm. In future studies, it is intended to run the algorithm in parallel to improve the performance.

REFERENCES

[1] Data Mining: Concepts and Techniques, 3rd Edition, Jiawei Han and Micheline Kamber, Jian Pei, 2007.
[2] A Survey of Grid Based Clustering Algorithms, Ilango and V Mohan, International Journal of Engineering Science and Technology Vol. 2(8), 2010.
[3] A density-based algorithm for discovering clusters in large spatial databases with noise, Martin Ester, Hans-peter Kriegel , Jörg S, Xiaowei Xu, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining,1996.
[4] Survey of Clustering Data Mining Techniques, Pavel Berkhin, 2002.
[5] OPTICS: Ordering Points To Identify the Clustering Structure, Mihael Ankerst , Markus M. Breunig , Hans-peter Kriegel , Jörg Sander, Proceedings of the 1999 ACM SIGMOD international conference on Management of data, 1999.
[6] STING: A statistical information grid approach to spatial data mining, Wei Wang , Jiong Yang , Richard Muntz, Proceeding VLDB '97 Proceedings of the 23rd International Conference on Very Large Data Bases, 1987.
[7] Daszykowski, M., B. Walczak, and D. L. Massart. "Looking for natural patterns in data: Part 1. Density-based approach." *Chemometrics and Intelligent Laboratory Systems* 56.2 (2001): 83-92
[8] Kriegel, Hans-Peter, and Martin Pfeifle. "Density-based clustering of uncertain data." *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005.
[9] Ram, Anant, et al. "A density based algorithm for discovering density varied clusters in large spatial databases." *International Journal of Computer Applications* 3.6 (2010): 1-4.
[10] W. Jin, K. H. Tung and J. W. Han: Mining Top-n Local Outliers in Large Databases. *KDD* 2001
[11] W. Jin, Tung, J. Han, W. Wang, Ranking Outliers using symmetric neighborhood relationship, in: Advances in Knowledge Discovery and Data Mining,2006, p.p. 577 593.

