International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 10, October-2014, pp: (20-27), Impact Factor: 1.296, Available online at: www.erpublications.com

DBSCAN (Density Based Clustering Method with Connective Regions) - A Survey

Ms. Sneha Sharma
Assistant Professor, Computer Science, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan

Abstract: Data mining has suit an important in research area because of its ability to get valuable information from the data. The data mining uses various clustering algorithms for grouping related objects. One of the most important clustering algorithm is density based clustering algorithm, which groups the related objects in non linear shapes structure based on the density. But it has the problem of varied density, which does not find out meaningful clusters. To overcome this problem an improved NDCMD(A unified novel density based clustering using multidimensional spatial data) is used. In this paper, we also present P- DBSCAN, a new density-based clustering algorithm based on DBSCAN for analysis of places and events using a collection of geo-tagged photos. We thereby introduce two new concepts: (1) density threshold, which is defined according to the number of people in the neighborhood, and (2) adaptive density, which is used for fast convergence towards high density regions.

Keywords: P-DBSCAN, NDCMD, Data mining.

I. INTRODUCTION

Today data is received automatically from many different kinds of equipments. Satellites, x-rays and traffic cameras are just a few of them. To make this information/data understandable for us, it has to be processed. When working with large data sets it is in most scenarios useful to be able to separate information by dividing the data into smaller categories, and eventually, to do class identification. Not least is this important when treating large spatial databases. A satellite, for example, gathers images as it travels around our earth. It is desired to classify what parts of the images are houses, cars, roads, lakes, forests, etc. Since the image database is big, a good classification algorithm is needed. Classification can, for instance, be done with the help of clustering algorithms, which clumps similar data together into different clusters. However, using clustering algorithms involves some problems: It can often be difficult to know which input parameters that should be used for a specific database, if the user does not have enough knowledge of the domain. Furthermore, spatial data sets can contain huge amounts of data, and trying to find cluster patterns in several dimensions is very computationally costly. Short computing time is always favorable. Last, the shapes of the clusters can be arbitrary and in bad cases very complex. There are some well-used clustering algorithms out there; one of them is the famous DBSCAN. DBSCAN algorithms can handle all these mentioned problems in a good way. To overcome the varying density problem, NDCMD is introduced in this paper. Section 3 will discuss how the DBSCAN algorithm works in. Section 4 contains new approach that is P-DBSCAN for analyzing the path through which images are shared. Section 5 presents the concept of NDCMD. Section 6 present the comparison between DBSCAN and NDCMD, in terms of performance. Finally, section 7 contains the conclusion of this paper and sums up the positive and negative aspects of the DBSCAN algorithm, P-DBSCAN and NDCMD.

2. Related Work

Density-based clustering methods first established a little more than a decade ago. DBSCAN-based algorithms is that every point in a database should contain a minimum number of MinPts points in its neighborhood of radius E.

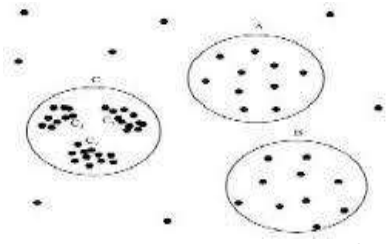


Fig. 1: Clusters having minimum no. of minpts

Improvements suggested in later research aimed at generalization of clustering approaches, efficient selection of input parameters, solving the problem of local densities or introducing a specialization for a particular task, such as moving clusters, trajectory clustering, spatio-temporal analysis of seawater characteristics, seismic activity, etc. For P-DBSCAN, Visualization of concentration of tourists using grid-based clustering was performed. Representative landmark images were found on the city and country scales in combining coordinates of geo-tagged photos with content based and textual analysis. After that weight was attached to images. To overcome the varying density problem, NDCMD was introduced.

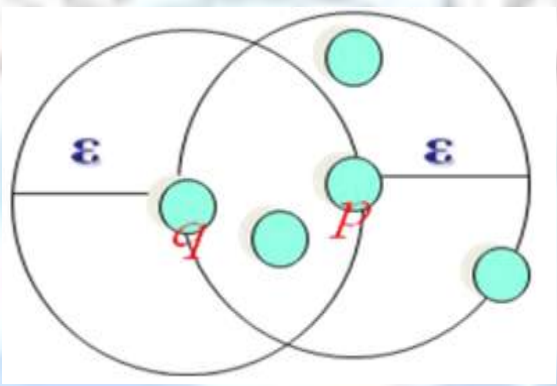
3. DBSCAN Algorithm

DBSCAN is a density based clustering method with connective regions. It is used to find out the cluster of arbitrary shape of data. DBSCAN method can find or handle the cluster of non spherical shape data.

3.1. Basic Terminology

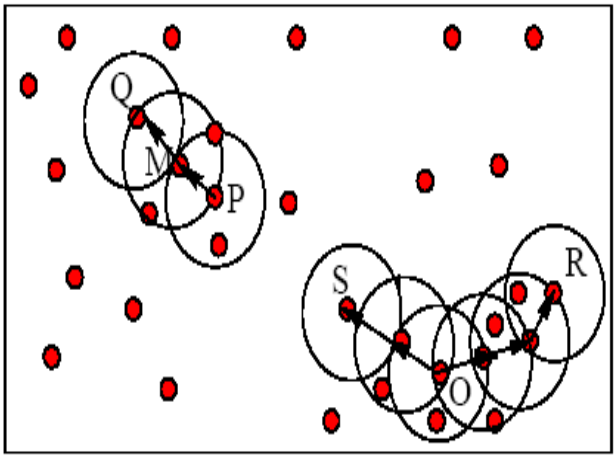
1. E- Neighborhood:

If any object lies within a radius E,Then It will be in E- neighborhood. In a given Example objects p,q are in E-neighborhood.



2. Core Object:

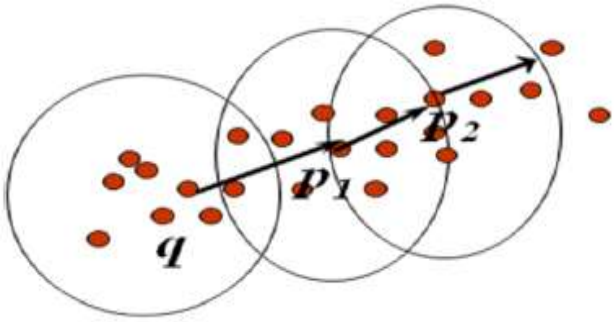
If object lie in E-neighborhood and having minimum minpts then it will be a core object. In given example, M,P,O,R are the core objects



3. Density-reachable:

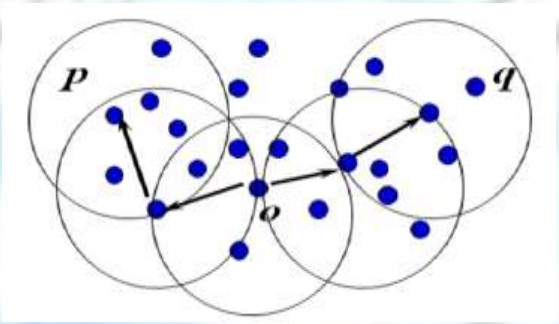
Point P is directly density reachable from point q.

- A point p is directly density-reachable from p2;
- p2 is directly density-reachable from p1;
- p1 is directly density-reachable from q;
- p ← p2 ← p1 ← q form a chain.



4. Density-connected:

P and q are density connected with each other by o.



3.2 DBSCAN Algorithm

DBSCAN(D, eps, MinPts)
C = 0
for each unvisited point P in dataset D
mark P as visited
NeighborPts = regionQuery(P, eps)
If sizeof(NeighborPts)< MinPts
mark P as NOISE
else
C=nextcluster expandCluster(P,NeighborPts,
C,eps, MinPts)
expandCluster(P,NeighborPts,C,eps, MinPts)
add P to cluster C
for each point P' in NeighborPts
if P' is not visited
mark P' as visited
NeighborPts' = regionQuery(P', eps)
if sizeof(NeighborPts') >= MinPts
NeighborPts = NeighborPts joined
with NeighborPts'
if P' is not yet member of any cluster
add P' to cluster C
regionQuery(P, eps)
return all points within P's eps- neighborhood (including P)

3.3 DBSCAN Advantage

- Cluster can have arbitrary shape and size.
- Number of cluster determined automatically
- Can separate cluster from surrounding noise
- Can be supported by spatial index structure.

3.4 DBSCAN Disadvantage

- Input parameter may be difficult to determine
- Cannot deal with varying density

3.5 Future Work

DBSCAN algorithm considers only point objects but it could be extended for other spatial objects like polygons.

4. P-DBSCAN

4.1 History

P-DBSCAN was introduced by Slava Kisilevich, Florian Mansmann, Daniel Keim in June 2010.

4.2 Approach

(1) density threshold:
which is denned according to the number of people in the neighborhood.
(2) adaptive density:
which is used for fast convergence towards high density regions.

4.3 P-DBSCAN Algorithm

Input: D - dataset of points with coordinates and ownership attributes, _ - neighborhood radius, Ad - adaptive density ag, Addt - adaptive density drop threshold

Output: Set of clusters

```
1 cluster-id = 0
2 while ((p = getUnprocessedPhoto(D)) =2 ;) do
3 CurrentDensity = Addt
4 if (jNeighborhood(p)j < MinOwners) then
5 MarkPhotoAsNoise(p)
6 else
7 cluster-id = cluster-id + 1
8 AssignPhotoToCluster(p,cluster-id)
9 UniqueQueue(Q,GetNeighborhoodPhotos(p))
10 while (Q is not empty) do
11 p = DeQueue(Q)
12 AssignPhotoToCluster(p,cluster-id)
13 if (jNeighborhood(p)j >= MinOwners) then
14 if (Ad == true) then
15 AdaptiveDensity(...)
16 else
17 UniqueQueue(Q,GetNeighborhoodPhotos(p))
```

18 end
19 end
20 end
21 end

4.4 Benefits

PDBSCAN not only reduces memory and I/O cost but also has the ability to deal with multi-density datasets.

4.5 Limitation

- This algorithm needs much human intervention in dataset partitioning
- It fails to deal with ring-shaped datasets.

5. NDCMD

It is a novel approach towards DBSCAN.NDCMD overcomes the varying density problems of dbscan.

5.1 Proposed System

In addition to DBSCAN the following definitions are required in NDCMD (A Unified Novel Density Based Clustering Using Multidimensional Spatial Data) to allow the considerable forming the same cluster and wide density variation.

Definition 1: Since there exists a variety of different types of data, a number of distance measures have been introduced. The most commonly used is Euclidean distance which is defined by the following equation:
Dist(p,q)=sqrt root((pk-qk)^2) for k=1 to d
Where p and q are data points and d is a number of dimensions.

Definition 2: (Cluster Density Mean): It is denoted by CDM(C).The Cluster Density Mean (CDM) of a growing cluster is defined as follows:
CDM (C)=(|C|
Where the N(o) is the density of the object o around in the ε- neighbourhood.

5.2 NDCMD Algorithm

Input : Data set D
Minimum points required to neighbourhood object x
Radius required to find nearest neighbourhood object ε

Output: No of clusters
Algorithm NDCMD (D,x,ε)

1. Initially all objects are unclassified
2. For each unclassified object x Є D
3. If Core(x) then
4. Generate new Cluster ID & Assign the clusterID to x
5. Insert x into the Queue
6. While Queue ≠ Empty
7. Extract front object p from the Queue
8. N = get Neighbors (p, ε)
9. If (size of(N) < ε)
10. mark p as NOISE
11. else

12. Increment x
13. mark p as visited
14. add p to cluster x
15. recourse (N)
16. Output as No. of clusters
17. for each detected x clusters
18. Find the cluster centers CDM
19. Find the total number of points in each cluster
20. If (no of clusters < define clusters)
21. unite clusters
22. else
23. subtract clusters from desire clusters and store into queue
24. split one or more as follows
25. Result as no of clusters

6. Performance Analysis

To judge against the performance of the proposed algorithm, we have also implemented the well known DBSCAN algorithm as well as novel algorithm. JAVA is used as a language to implement the algorithms. The performances of above two algorithms are evaluated by using the 2- Dimensional synthetic dataset in .arff file format. The 2-Dimensional synthetic dataset is containing varying objects from 14 to 5250 in 2-Dimensional plane. We comes to compare the time taken to built clusters using two different algorithm as well as compare the no of clusters form by the algorithms.

Table 1: Execution time of DBSCAN & NDCMD

Data Set	Instances	DBSCAN	NDCMD
		Time (seconds)	Time (seconds)
Weather	14	0.03	0.02
Cpu	209	0.03	0.02
Glass	214	0.02	0.01
Vote	435	0.02	0.02
Soybean	683	0.09	0.06
Diabetes	768	0.08	0.06

A3	3005	1.74	1.1
Super-Market	4627	1.67	0.94
A2	5249	1.94	1.2

Table 2: Number of clusters- DBSCAN & NDCMD

Data Set	Instances	DBSCAN	NDCMD
		No of Clusters	No of Clusters
Weather	14	2	2
CPU	209	5	6
Glass	214	5	6
Vote	435	3	3
Soybean	683	6	7
Diabetes	768	7	9
A3	3005	13	14
Super-Market	4627	2	3
A2	5249	13	14

7. Conclusions

In this paper we proposed the idea of DBSCAN,P-DBSCAN and NDCMD algorithms. DBSCAN form the clusters of objects. Using P-DBSCAN, a person can easily understand the concept behind sharing of images. DBSCAN cannot solve the problem of varying density. hence a novel approach toward dbscan is used i.e. NDCMD. In this paper, performance analysis between DBSCAN and NDCMD is also conducted. DBSCAN algorithm can be use in various fields like medical, market segmentation etc.

Acknowledgement

I would like to thank Ms. Sneha sharma and Mr. Krutibash Nayak to encourage me and provide useful information to complete this review.

References

[1]. S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In Proceedings of the 7th ACM/IEEE joint conference on Digital libraries, pages 1{10, 2007.
[2]. M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. SIGMOD Rec., 28(2):49{60, 1999.
[3]. D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial-temporal data. Data Knowl. Eng., 60(1):208{221, 2007.
[4]. D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In WWW, 2009.
[5]. L. Duan, L. Xu, F. Guo, J. Lee, and B. Yan. A local-density based spatial clustering algorithm with noise. Inf. Syst., 32(7):978{986, 2007.
[6]. Sanjay Chakrobarty, Prof. N.K.Nagwani," Analysis and Study of Incremental DBSCAN Clustering Algorithm", International Journal of Enterprise Computing and Business,Vol. 1, Issue 2 ,July 2011
[7]. K. Mumtaz et al, "A Novel Density based improved kmeans Clustering Algorithm – Dbkmeans", International Journal on Computer Science and Engineering, Vol. 02, pp. 213-218, 2010
[8]. Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra ,"An Empirical Evaluation of Density-Based Clustering Techniques",International Journal of Soft Computing and Engineering, Vol. 2, pp. 216-223, March 2012
[9]. B.G.Obula Reddy1, Dr. Maligela Ussenaiah2, " Literature Survey On Clustering Techniques", IOSR Journal of Computer Engineering, Vol. 3,pp. 01-12, July-August 2012
[10]. M.Parimala, Daphne Lopez, N.C. Senthilkumar "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology Vol. 31, pp. 59-66, June-2011
[11]. www.StudyMode.com
[12]. www.ijser.com
[13]. www.iosrjournals.com
[14]. www.ijoart.com