NDCMD: A Novel Approach Towards Density Based Clustering Using Multidimensional Spatial Data

Khushali Mistry, Swapnil Andhariya, Prof. Sahista Machchhar
[1]*Student, Marwadi Education Foundation's Group of Institution, Rajkot,*
[2]*Student, Marwadi Education Foundation's Group of Institution, Rajkot,*
[3]*Assistant prof., Marwadi Education Foundation's Group of Institution, Rajkot*

***Abstract*-- Density based clustering algorithm is one of the primary methods for clustering in data mining. The clusters which are formed based on the density are easy to understand and it does not limit itself to the shapes of clusters. One of them is DBSCAN which is a well known DENSITY-based clustering algorithm used for mining of unsupervised data. The DBSCAN algorithm suffers from several deficiencies whenever the database size is large. Also, DBSCAN does not respond well to data sets with varying densities. For this reason its complexity in worst case becomes $O(n^2)$. The PROPOSED novel algorithm NDCMD (A Unified Novel Density Based Clustering Using Multidimensional Spatial Data): this outperforms DBSCAN for varying density. This is motivated by the current state-of-the-art density clustering algorithm DBSCAN. Ultimately we show how to automatically and capably extract not only 'traditional' clustering information, such as representative points, but also the fundamental clustering structure. Extensive experiments on some synthetic datasets show the validity of the proposed algorithm.**

Keywords-Clustering, DBSCAN, NDCMD

1. Introduction

Data Mining is the method of identifying valid, novel, useful, and knowledgeable information as of data. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining functionalities like Data characterization, Data discrimination, Association analysis, Classification, prediction, Cluster analysis, Outlier analysis, Evolution analysis etc., discover patterns from the data. Clustering techniques widely used in numerous applications like including pattern recognition, market research, image processing and data analysis [14].

Clustering techniques classified into five major types –Hierarchical, Partitioned, Density-Based, Model-Based and Grid-Based. For Density-based methods developed to discover clusters with arbitrary shape. DBSCAN algorithm grows regions with high density forms and discovers clusters of arbitrary shape in spatial databases handle noise. OPTICS computes an enhanced cluster ordering for automatic and interactive cluster investigation, Contains information from OPTICS equivalent to density-based clustering obtained from a wide range of parameter settings. DENCLUE clustering method based on a set of density distribution functions [14].In this paper is covers DBSCAN algorithm and the novel approach towards with density based clustering.

In the proposed work, we present a new density based clustering algorithm which is successful to handle density variation in the dataset objects from low to high density from multidimensional data. First it finds the neighbors to form the clusters and then it calculates the growing cluster density mean. After it check if the number of clusters is more then so it will merge the two clusters otherwise it separates the clusters.

Second section of this paper focus on the related works for DBSCAN algorithm. Third section relates brief introduction to DBSCAN algorithm. Forth section gives novel approach to proposed algorithm. In fifth section give performance analysis of two algorithm and gives comparison of two algorithms. Finally last section six represents conclusion and future work.

2. Related work

J.Hencil Peter, A.Antonysamy combines a Fast DBSCAN Algorithm and Memory effect in DBSCAN algorithm to speed up the performance as well as improve the quality of the output. As seen that the Region Query operation takes long time to process the objects, only few objects are considered for the expansion and the remaining missed border objects are handled differently during the cluster expansion[1].

Yan Ren, Xiaodong Liu, wanquan Li relates Two novelties for the proposed algorithm, One is to adopt the Mahalanobis metric as distance measurement instead of the Euclidean distance in DBSCAN and the other is its effective merging approach for leaders and followers defined . This Mahalanobis metric is closely associated with dataset distribution. In order to overcome the unique density issue in DBSCAN, proposed an approach to merge the

sub-clusters by using the local sub-cluster density information. Eventually shown how to automatically and efficiently extract not only 'traditional' clustering information, such as representative points, but also the intrinsic clustering structure[2].

Qiliang Liu,, Min Deng , Yan Shi, Jiaqiu Wang addresses the problem of how to accommodate geometrical properties and attributes in spatial clustering. A new density-based spatial clustering algorithm (DBSC) is developed by considering both spatial proximity and attribute similarity. Delaunay triangulation with edge length constraints is first employed for modeling the spatial proximity relationships among spatial objects. A modified density-based clustering strategy is then designed and used to identify spatial clusters[3].

Bidyut Kr. Patra, Sukumar Nandi, P. Viswanath propose a distance based clustering method, l-SL to find arbitrary shaped clusters in a large dataset. In this method, first leaders clustering method is applied to a dataset to derive a set of leaders; subsequently single-link method (with distance stopping criteria) is applied to the leaders set to obtain final clustering. The l-SL method produces a flat clustering.Clustering result of the l-SL may deviate nominally from final clustering of the single-link method (distance stopping criteria) applied to dataset directly[4].

Sanjay chakra borty, Prof. N.K.Nagwani describes the incremental behaviors of Density based clustering. It specially focuses on the DBSCAN algorithm and its incremental approach. DBSCAN relies on a density based notion of clusters. In incremental approach, the DBSCAN algorithm is applied to a dynamic database where the data may be frequently updated [6].

3. DBSCAN algorithm

DBSCAN, A Density Based Spatial Clustering of Application with Noise which is a density based clustering technique for finding clusters of arbitrary shapes. DBSCAN starts with an arbitrary object in the dataset and checks neighbor objects within a given radius i.e. Eps. If the neighbors within that Eps are more than the minimum number of objects required for a cluster, it is marked as core object and if the objects in it surrounding within given *Eps* are less than the minimum number of objects required, then this object is marked as noise. The search continues for all the objects in the dataset. In DBSCAN, the distance of two points is determined by a distance metric, such as the Euclidean distance. For two points p and q in a dataset D, the distance between them is denoted by dist(p,q). Usually the distance is only dependent on these two points and independent on the dataset distribution[14].

Definition 1. (Eps-neighborhood). The Eps-neighborhood of a point p is defined by $\{q \in D | dist(p, q) \leq Eps\}$.

Definition 2. (Core point). A core point contains at least a minimum number (MinPts) of other points within its Eps-neighborhood.

Definition 3. (Directly density-reachable). A point p is directly density-reachable from a point q if p is within the Eps-neighborhood of q, and q is a core point.

Definition 4. (Density-reachable). A point p is density-reachable from the point q with respect to Eps and MinPts if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$ and $p_n = q$ such that p_{i+1} is directly density reachable from p_i with respect to Eps and MinPts, for $1 \leq i \leq n$, $p_i \in D$.

Definition 5. (Density-connected). A point p is density-connected to a point q with respect to Eps and MinPts if there is a point $o \in D$ such that both p and q are density-reachable from o with respect to Eps and MinPts.
With this concept state the following DBSCAN algorithm:-

```
DBSCAN (D, eps, MinPts)
C = 0
 For each unvisited point P in dataset D
    Mark P as visited
    N = regionQuery(P, eps)
    if sizeof(N) < MinPts
      Mark P as NOISE
    else
      C = next cluster
      expandCluster(P, N, C, eps, MinPts)
 expandCluster(P, N, C, eps, MinPts)
  add P to cluster C
  for each point P' in N
    if P' is not visited
      mark P' as visited
      N' = regionQuery(P', eps)
        if sizeof(N') >= MinPts
        N = N joined with N'
    if P' is not yet member of any cluster
      add P' to cluster C
```

The time complexity of the DBSCAN method is $O(n^2)$, where n is the number of objects in the dataset; With the support of spatial access methods such as R*-tree, its time complexity can reduce to O(nlogn).

4. 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.

Defination 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:

$$dist(p,q) = \sqrt{\sum_{k=1}^{d} (pk - qk)^2} \qquad (1)$$

Where p and q are data points and d is a number of dimensions.

Defination 2: (Cluster Density Mean): It is denoted by CDM(C).The Cluster Density Mean (CDM) of a growing cluster is defined as follows:

$$CDM(C) = \frac{\sum O \in C \left| N\varepsilon(o) \right|}{|C|} \qquad (2)$$

Where the N(o) is the density of the object o around in the ε-neighbourhood.

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

5. 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)	
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	
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

6. Conclusion and Future work

In this paper, we considered novel density based clustering algorithm that outperformance of the DBSCAN algorithm. Also we tested the algorithm on synthesis database. While testing the performance evaluation of this algorithm, the datasets that were taken ranges from low to high data set. The algorithm was tested on synthetic datasets. The results of these experiments demonstrate performance over the DBSCAN. Moreover, the biggest disadvantage of DBSCAN is that it cannot work on varied densities. We can solve this problem by using novel density based algorithm as well as time to form cluster is less compared to DBSCAN. The future work is lot of scope for the proposed NDCMD clustering algorithm in different application areas such as medical image segmentation and medical data mining. Future works may address the issues involved in applying the algorithm in a particular application area.

7. References

[1] J. Hencil Peter, A. Antonysamy, "An Optimised Density Based Clustering Algorithm", International Journal of Computer Applications ,Vol. 6– No.9, pp. *0-25,* September 2010

[2] Yan Rena, Xiaodong Liua, Wanquan Liuc," DBCAMM: A novel density based clustering algorithm via using the Mahalanobis metric", Applied soft computing,pp.1542-1554,2012

[3] Qiliang Liu,Min Deng ,Yan Shi,Jiaqiu Wang ,"A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity ",Computers & Geosciences ,vol. 46,pp.296–309,2012

[4] Bidyut Kr.Patra,Sukumar Nandi, P. Viswanath ,"Pattern Recognition ",vol.44 ,pp. 2862–2870 ,2011

[5] Anant Ram, Sunita Jalal , Anand S. Jalal, Manoj Kumar," A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases ",International Journal of Computer Applications, Vol. 3 – No.6, June 2010

[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 k-means 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] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of

[12] Various Clustering Algorithms in Data Mining ",International Journal of Engineering Research and Applications, Vol. 2,pp.1379-1384, May-Jun 2012

[13] Rahmat Widia Sembiring, Sajadin Sembiringand Jasni Mohamad Zain,"An efficient dimensional reduction method for data clustering", Bulletin of Mathematics, Vol. 04,pp. 43–58,2012

[14] K.Mumtaz et. al., "An Analysis on Density Based Clustering of Multi Dimensional Spatial Data", Indian Journal of Computer Science and Engineering,Vol 1, pp. 8-12,2011

[15] Jiawei Han and Micheline Kamber ,"Data Mining Concepts &Techniiques",Elsevier,2011

[16] Arun K Pujari , "Data Mining Techniques", University Press Private Limited 2001 ,pp.42-46,2009

[17] http://www.cs.waikato.ac.nz/ml/weka/arff.html

[18] http://cs.joensuu.fi/sipu/datasets/a2

[19] http://cs.joensuu.fi/sipu/datasets/a3