Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 92 (2016) 136 – 141

Procedia Computer Science

2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

An Efficient And Scalable Density-Based Clustering Algorithm For Normalize Data

Nidhi[a]*, Km Archana Patel[a]

[a]National Institute of Technology, Kurukshetra, 136119, Haryana, India

Abstract

Data clustering is a method of putting same data object into group. A clustering rule does partitions of a data set into many groups supported the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Finding clusters in object, particularly high dimensional object, is difficult when the clusters are different shapes, sizes, and densities, and when data contains noise and outliers. This paper provides a new clustering algorithm for normalized data set and proven that our new planned clustering approach work efficiently when dataset are normalized.

Peer-review under responsibility of the Organizing Committee of ICCC 2016

*Keyword:*Density based Clustering Algorithm, Computational complexity.

1. Introduction

Clustering is most typically used and a lot of powerful unsupervised learning technique in data processing [1]. It is helpful method that aims to arrange the input data set in to a collection of finite range of semantically consistent group with respect to some similarity . These algorithms will be roughly classified into seven classes, particularly Hierarchical algorithms, Density-based algorithms, Partitional algorithms, Graph-based algorithms, combinational algorithms, Grid-based algorithms , and Model-based algorithms [2]. Several issues related with use of these clustering mechanism are describe in [20]. Among these varieties of algorithms , Density-based

* Corresponding author. Tel.: +91-9050476789.
*E-mail address:*nidhik.kashyap@gmail.com.

Peer-review under responsibility of the Organizing Committee of ICCC 2016
doi:10.1016/j.procs.2016.07.336

algorithms are measure renowned for his or her easy explanation and therefore the relative straightforward implementation. Another two vital benefits of this algorithms are measure it is able of discovering clusters of various shapes and different size even in outlier data set and it does not need users to specify the amount of clusters. The purposed Density-based algorithms are distinguishing dense regions that are measure separated by low-density regions. DBSCAN provides high-quality performance but it depends on two specified parameters, r and MinPts [3]. It is time consuming for looking the closest neighbors of every object is unbearable within the cluster expansion and choosing different beginning points outcome in quite different consequences. Thus our objective is improved DBSCAN rule and projected a new DBSCAN rule for normalized dataset. Data normalization is method for linear transformation of data to a particular range.

The rest paper is organized as follows. Section 2 gives a summary of related work and described the traditional DBSCAN in detail. In Section 3, we present our improved clustering algorithm for normalized data based on Influence Space in detail. Finally, Section 4 concludes the paper and gives some future research.

2. Related Works

The documented DBSCAN algorithm is widely employed in applications such as social science[21, 22], anomaly detection[23, 24], biomedical image analysis[25] and spectroscopy. where it's needed to spot outliers and characterize clusters having impulsive shapes [4]. The main disadvantage of DBSCAN is that the high complexity within the neighborhood query for every object to construct the similarity matrix. DBSCAN algorithm cluster a low dimensional area whereas its performance degrades when managing with high-dimensional and large-scale information sets [5]. During this section, we have a tendency to first summary the most plan of DBSCAN then we have a tendency to discuss our planned DBSCAN algorithm for normalized information.

The DBSCAN Algorithm

DBSCAN is a good Density-based clustering rule initially for spatial catalog systems owing to its capability of checking out clusters with discretionary shapes. There are two major parameters in DBSCAN which are needed to be fixed, r and MinPts in which r represents the radius of a vicinity from the observing degree and MinPts suggests that the minimum variety of information degrees contained in such a vicinity. Suppose we tend to measure a given data set of n degrees Dataset ={ y1, y2,…,yn}. In DBSCAN, three completely dissimilar relationships between any two different degrees are measure outlined as follows:

2.1 Directly density reachable:A degree q is directly densible reachable from a degree p if q belong to Nr(p) and Nr(p) >=MinPts, where Nr(p)={q| distance (p,q)<=r}. Values of distance (p, q) are different with various distance functions

2.2 Density reachable: A degree q is density reachable to a degree p with regard to r and MinPts, if there is a series of degrees q1,…,qn, q1 = p, qn = q, such that qi+1 is directly density- reachable from qi with regard to r and MinPts, for1<=i<=n, qi belong to Dataset.

2.3 Density connected: A degree q is density connected to a degree p with regard to r and MinPts if there is a degree m belong to Dataset such that both q and p are density reachable from m with respect to r and MinPts.

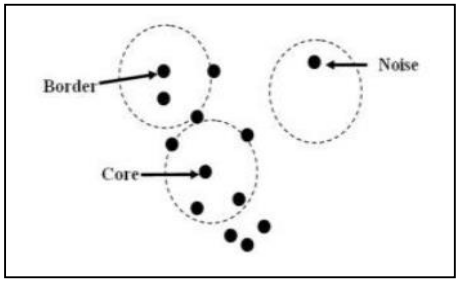

Fig. 1.Ouliers, Border and Core degree

According to these three relationships, all degrees in DBSCAN may be classified in to three categories: Core degree, Border degree and Outliers degree. Core degree- If the Quantity of degrees that are measure directly density reachable from q is bigger than MinPts within the r- neighborhood of a degree q, then q may be a Core degree. Border degree- If the Quantity of degrees within the r-neighborhood of a degree q is not over MinPts, and q is directly density reachable from a Core degree ,then q may be a Border degree. Noise degree- If a degree q is neither a Core degree nor a Border degree, then q may be a Outliers degree.

DBSCAN works as follows :

```
1.  Require: Dataset={y1,y2,…yn} the dataset. r: the maximu
radius of aneighborhood. MinPts: the minimum number of data degree
contained in ε- neighborhood.
2.  Ensure: CS = {CS1, CS2,…CSk}: set of clusters.
3.  ClusterID1=0
4.   Tick all degreesyi belong to Dataset as “UNVISITED”
5.   for all yi belong to Dataset do
6.  if yi is ticked as “UNVISITED” then
7.   mark yi as “VISITED”
8.   count the degrees in the r-neighborhood of yi as |Nr(yi)|
9.  if |Nr(yi)| <MinPts then
10.   tick all degrees in the r-neighborhood of yi as “outliers”
11.   else
12.   ClusterID1++
13.   ticked all degrees in the r-neighborhood of yi as
      ClusterID1
14.   insert degree ticked as “UNVISITED” in the r- neighborhc
of yi in to a queue
15.   while the queue is not blank do
16.   select the head degreeyp in the queue
17.    tick yp as “VISITED”
18.   calculate the degree in the r- neighborhood of yp as |Nr(yp
19.   remove yp from the q
20.   if |Nr(yp)|>=MinPts then
21.   tick all degrees in the r-neighborhood of yp as ClusterID1
22.   insert degrees ticked as “UNVISITED” in the r-
neighborhood of yp in to the queue
23.   endif
24.   endwhile
25.   endif
26.   endif
27.   end for
```

DBSCAN Algorithm scans whole dataset only one time and needs to calculate the distance of any pair of objects in the dataset. Hence, the computational complexity of the whole algorithms O(n2), where n is the number of degrees in the data set. If effective index structures are used and the dimension of degrees is low (d<=5), the computational complexity of DBSCAN can be reduced to O (n logn).

Limitations of DBSCAN:

- The performance of the algorithmic program depends on two parameters, r and MinPts
- The time consumption for looking the closest neighbors of every object is intolerable within the cluster enlargement.
- Selecting completely different beginning degrees leads to quite different consequences.
- DBSCAN is unable to spot adjacent cluster therefore numerous densities.

According to the shortcomings mentioned on top of, we have a tendency to projected a replacement algorithmic program on the idea of the standard DBSCAN algorithmic program and IS-DBSCAN [26] that success solves these above problem. In our projected algorithmic program we have a tendency to measure exploitation new neighborhood relationship supported the influence area, reduces variety of parameters to only one.The number of k-nearest neighbors. Meanwhile, the influence area is sensitive to native density changes, therefore it improves the matter of selecting adjacent clusters of various densities and native Outliers.

3. Proposed Algorithm:

Our projected algorithm not resolved the matter of selecting input parameters however conjointly obtained clusters with varied densities for normalized dataset. For the aim of handling with high-dimensional datasets, some new index structures are using. One of these new index structures is that local sensitive hashing methodology (LSH) [28]. In our paper, we tend to exploitation following ways to boost the standard of DBSCAN algorithm:

Search the closest neighbors of every object within the data set by using the improved p-stable locality sensitive hashing methodology to boost the quantifiability of DBSCAN.

1. Determine adjacent clusters with numerous density by taking advantage of the new neighborhood relationship : the IS_k-neighborhood relationship. ISk represents the influence area of every object. Moreover, this association is symmetric that perform the random choice of inputting degrees and with effectively reduces the quantity of parameters.
2. The difficulty of detecting contiguous clusters of various density is additionally a hard restriction of DBSCAN take away this restriction by local-density-based
3. Discover border objects by creating use of the new concept—Core density reachable
4. Improved p-stable locality sensitive hashing scheme—RLSH. It uses two steps

Step1: Generation of hash tables step2: Nearest neighbor look for the query p:

Overall, RLSH prevents the loss of the closest neighbors of the query degree the maximum amount as attainable to ensure the high accuracy. A further superiority of this strategy is that RLSH offer higher chance of finding the closest neighbor even with fewer hash tables for the aim of reducing the space needs due to the range of the candidate set. Thus we are using RLSH in our projected DBSCAN algorithm for normalized facts. In our paper, *only one parameter k* is outlined to represent each conditions (r radius and mindegree). *IS_k-neighborhood*- The neighborhood of q is defined by p belong to Dataset intersection, p belong to $IS_k(q)$.*Core degree*- If the quantity of degrees within the influence area of q ($IS_k(q)$) is bigger than 2k/3 (i.e. $IS_k(q)>2k/3$), then q is a Core degree.

In our projected algorithm we are using core density reachable, for the target of disconnecting the last border objects with density reachable chain. In this paper the definition of core density reachable relies on the r-neighborhood relationship, whereas we tend to outline core density reachable within the IS_k-neighborhood. Next, many completely different definitions in our planned methodology are represented.

Directly density reachable - A degree q is directly density reachable from a degree p if q belong ISk(p) and |ISk(p)|>=2k/3. Core density reachable - A degree q is core density reachable to a degree p with respect to k, if there are a sequence of degrees q1,…qn; q1 = p; qn = q, such qi+1 is directly density reachable from qi with relation to k and qi are measure core degrees for 1< i <n degree q belong to Dataset. Density connected - A degree q is density connected to a degree p with respect to k if there is a degree M belong to Dataset such that both p and q are core density reachable from M with respect to k.

3.1 Our proposed algorithm works as follows:

The procedure of *DBSCAN FOR NORMALIZED* data are often divided into three steps.

The first step: It is data initialization step and scanning the total dataset. The first step consists of two aspects, marking all degree as "UN_CLASSIFIED" and calculating the influence area of every degree. *The second step*: The most vital step is that the enlargement of core clusters. For the present observant degree y_i, if $|IS_k(y_i)| > 2k/3$, then y_i is the Core degree with the present ClusterID1. Afterwards, the algorithm searches degree y_j within the influence area of y_i that compose SeedList1. If $|IS_k(y_j)| > 2k/3$, then y_j is ticked a similar label as x_i. degree within the influence area of y_j that do not seem to be classified are additional into SeedList1. This method ensures that every one degree that are core density reachable from the observing degree y_i are allotted into a similar cluster.*The third step*: The last step of completely identifies border degrees. After the second step all core degrees detected. Degrees that are ticked as UN_CLASSIFIED after previous two steps are border degrees. In our algorithm, border degrees are labeled as cluster label which are the same as the label of then Core degree to border degrees.

Procedure of our proposed algorithm DBSCAN for normalized data

Algorithm: DBSCAN for normalized data

Require: Dataset ={y1, y2,…,yn}: the data set. Data set is normalized, k: the number of neighbors.

Ensure: CL={CL1, CL2,…,CLk}: set of clusters.

function DBSCAN for normalized data set D, k

```
1.  ClusterID1= 1
2.  Tick all degrees yi belong to Dataset as "UNCLASSIFIED"
3.  calculate the influence space ISk (yi) of each degreeyi belong
Dataset
4.  for all yi belong to Dataset do
5.  ifyi is ticked as "UNCLASSIFIED" then
6.   if ExpandCoreClusteryi,ClusterID1  then
7.  ClusterID1++
8.   end if
9.  end if
10. end for
11. for all yi belong to Dataset do
12. if yi is ticked as "UNCLASSIFIED" then
13. search each degree yj belong to ISk(yi) in the influence space of
14.  end if
15.  end for
16. end function
17. function EXPANDCORECLUSTER1  (yi, ClusterID1)
18.  SeedList1=ISk(yi)
19. if |SeedList1|>2k/3 then
20.yi is labeled as ClusterID1
21.else
22.false
23.end if
24.for all yj belong to SeedList1 do
25.if  |ISk(yj)|>2k/3 then
26.yj is labeled as ClusterID1
27.for all zm belong to |ISk(yj)|do
28.ifzm is labeledas "UNCLASSIFIED"
 29. If zm is not in SeedList1 then
30. addzm into SeedList1
31.end if
 33. end if
33. end for
 34.end if
35.end for
 36.true
 37. end function
```

Note that, the algorithm offered in our paper needs just one parameter k. This parameter represents the quantity of k-nearest neighbors. $IS_k(q)$ represents the influence space of a degree p. The planned algorithm can even be enforced for different kinds of normalization procedure, by merely modifying first two steps in the algorithm. Our planned DBSCAN for normalized data filtering all border degrees in every cluster expansion step. Border degrees are distinguished till all core degrees are known. Moreover, the new concept— Core density reachable guarantees that the Density-reachable chains of objects include Core degree that are measure Core density reachable from the observing Core degree are ticked as the same label as the observing core degree. Border degrees are allotted to the cluster labels that are similar to nearest core degree to border degrees in the influence areas of border degrees is in.

4. Conclusion and Future Works

Density based algorithm is not appropriate for data with high variance in density. DBSCAN needs two parametres r and MinPts. Generally this algorithm depends upon ordering of degrees within the dataset and it cannot cluster data sets well large distinction in densities. However our projected algorithm for normalized data, takes one parameter and deals with high dimensional information set. Additionally free to the starting degrees and able to notice adjacent clusters with varied densities.It also reduces the number of iteration. Our projected algorithm is easy but efficient and effective algorithm and it boosts the performance of DBSCAN in adjacent clusters with different densities. Future work will be dedicated to automatically identify only one parameter k when apply DBSCAN on normalized data , instead of specified by users.

References

[1] Manivara Kumar Parsha, SreenivasuluPacha: Recent advances in clustering algorithms,a review: International Journal of Conceptions on Computing and Information Technolog(November 2013)

[2] Suman andMrs.Pooja Mittal: Comparison and Analysis of Various Clustering Methods in Data mining On Education data set: International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)(March – April 2014)

[3] Bharat Chaudhari, Manan Parikh: A Comparative Study of clustering algorithms :International Journal of Application or Innovation in Engineering & Management (Oct 2012)

[4]Aastha Joshi and RajneetKaur: A Review Comparative Study of Various Clustering Techniques in Data Mining :International Journal of Advanced Research in Computer Science and Software Engineering(March 2013)

[5]YinghuaLv ,TinghuaiMa, MeiliTang , JieCao , YuanTian , Abdullah Al-Dhelaan, MznahAl-Rodhaan: An efficient and scalable density-basedclustering algorithm for datasts with complex structures:Neurocomputing(2015)

[6] Nida Rashid:An Algorithm Analysis on Data Mining: international Journal of Recent Research in Mathematics Computer Science andInformation Technology(sept 2015)

[7] R.Roseline, G.Jenitha, Henri Amirhta raj: Analysis and Application of Clustering Techniques in Data Mining: International Journal of Computing Algorithm(03, May 2014)

[8] Li ma,Lei-gu,Bolo,Souyi,qiao,Jin Wang: G-DBSCAN: An Improved DBSCAN Clustering Method Based On Grid: Advanced Science and Technology(2014)

[9]Anand V. Saurkar, VaibhavBhujade, PritiBhagat and AmitKhaparde: A Review Paper on Various Data Mining Techniques:International Journal of Advanced Research in Computer Science and Software Engineering (April 2014.)

[10]Wang Gui-Zh, Zhang Jian-Wei: Clustering-boundary-detection Algorithm Based onCenter-of-gravity of Neighborhoo: TELKOMNIKA(December 2013)

[11] P. IndiraPriyaand Dr. D.K.Ghose:A Survey on Different Clustering Algorithms in Data Mining Technique: International Journal of Modern Engineering Research (IJMER) (Feb.2013)

[12] R.C Mishra, Professor M.D.U Rohtak:Performance Evaluation of Clustering Algorithms: International Journal of Engineering Trends and Technology (7July 2013).

[13] NavneetKaur: Survey Paper on Clustering Techniques: Engineering and Technology Research (IJSETR)(April 2013)

[14] RajnetKaur,sri Guru: A Review Comparative Study of Various Clustering Techniques in Data Mining: International Journal of Advanced Research in Computer Science and Software Engineering(March 2013)

[15] binGu, VictorS. Sheng:Feasibility and finite convergence analysis for accurate on-line—support vectorlearning,IEEETran: NeuralNetw(2013).

[16] Shiv Pratap Singh Kushwah, KeshavRawat, Pradeep Gupta: Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining:International Journal of InnovativeTechnology and Exploring Engineering(August 2012)

[17] Shiv Pratap Singh Kushwah, KeshavRawat, Pradeep Gupta: Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining:International Journal of Innovative Technology and Exploring Engineering (IJITEE),(3 August 2012).

[18]S.Ganapathy, K.Kulothungan, P.Yogesh, A.Kannan: A Novel Weighted Fuzzy C-Means Clustering Based on Immune Genetic Algorithm for Intrusion Detection: Proceeding Engineering,Elsevier(2012).

[19] Manish Verma, Nidhi Gupta: A Comparative Study of Various Clustering Algorithms in Data Mining: International Journal of Engineering Research and Applications,(2012)

[20] A.K.Jain,M.N.Murty,P.J.Flynn,Dataclustering:areview,ACMComput.Surv. 31(1999)264–323.

[21]M. Mete, S. Kockara, K. Aydin, Fast density – based lesion detection in dermo scopy images, Comput. Med. Imaging Graph. 35 (2011) 128–136.

[22]J.Gong, C. H. Caldas, Data processing forreal – time construction site spatial modeling, Autom.Constr.17(2008)526–535.

[23] S. Handra, H. Ciocarlie, Anomaly detection in data mining. Hybrid approach between filtering – and - refinement and DBSCAN, in : 2011 6th IEEE International Symposiumon Applied Computational Intelligence and Informatics (SACI), Timisora, Romania, May19–21,2011,pp.75–83.

[24]M. Celik, F. Dadaser- Celik, A. Dokuz, Anomaly detection in temperature data using DBSCAN algorithm, in : 2011 International Symposiumon Innovations in Intelligent Systems and Applications (INISTA), Istanbul, Turkey, June15–18, 2011, pp. 91–95.

[25]T. N. Tran, T. T. Nguyen, T. A. Willemsz, G. vanKessel, H. W. Frijlink, K. V. D. V. Maarschalk, Adensity-based segmentation for 3D images, an application for X-raymicro-tomography, Anal. Chim. Acta725(2012)14–21.

[26]M. Ester, H. – P . Kriegel , J. Sander, X.Xu, Adensity – based algorithm for discovering clusters in large spatial data bases with noise, in KDD, Portland, Oregon, USA,August 2–4, 1996,pp.226–231.