IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ENHANCE VDBSCAN BY QUICK SELECT ALGORITHM FOR K VALUE SELECTION

Mrs. R.Bharathi M.Sc.,Mphil., ***Mrs.K.Karthika,M.Sc.,M.phil.,***
Mrs.C.Chandrapriya,M.sc.,Mphil.,
Assistant Professor, *Assistant Professor,* *Assitant Professor*
PG and Research Department of computer science,
DR.R.K.Shanmugam College of Arts and Science, Kallakurichi.-India.

Abstract *– Data mining is important topic in research area because Data mining is the process of analyzing a large batch of information t discern trends and patterns. The data mining uses various clustering algorithms for grouping related objects. One of the most importa algorithm is density based clustering algorithm which groups the related objects in non linear shapes structure based on the density. But it ha the problem of varied density, which does not find out meaningful clusters. To overcome this problem an improved VDBSCAN algorithm used. The main drawback of VDBSCAN algorithm is the value of parameter 'K' is user input dependent parameter, it largely degrades th efficiency of permanent Eps ('K' value is used for Eps selection). In our proposed method, the parameter 'K' is dataset dependent paramete which finds the characteristics of dataset. The characteristics in the proposed method are distance measure, average resolution and quic select method for 'K' selection. The distance measurement is done using jaccard index, which gives the better cluster formation, goo computation time and accuracy in large datasets.*

Index Terms - Data mining, clustering, DBSCAN, VDBSCAN, quick select based method.

I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data [1].There are five areas of clustering.

- Partitioning
- Hierarchical
- Density
- Grid
- Model

Density based algorithm it finds out the clusters in arbitrary shape. The cluster contained the minimum number of input parameters. The input parameter referred as Eps and minpts.eps is the radius of the cluster. Minpts referred as the inside of the cluster [9].

II. LITERATURE SURVEY

A.DBSCAN

DBSCAN is a pioneer density based clustering algorithm. It finds out the clusters of different shapes and sizes from the large amount of data which is containing noise and outliers [6][3].

Description of the algorithm:

- The algorithm first selects the center point of the cluster.
- Repair all points' density reachable from center point.

- If the center point is a core point, a cluster is formed. If the center point is border point, no points are density reachable from point and DBSCAN visits the next point of the database [4].

- Continue the process until all point has been processed.

Min pts=4
ϵ=1 unit

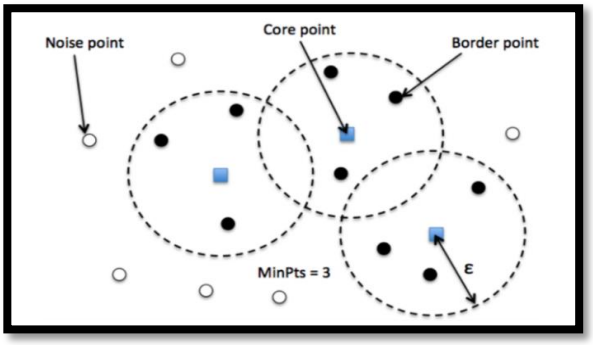


Fig 1: DBSCAN

Main concepts of DBSACN

DBSCAN uses density reachability and density connectivity.

Density reachability:

A point "p" is an density reachable from a point "q" if the point "p" is within ε distance from point "q" and "q" has enough number of points in its neighbors which are within the distance ε [5].

Density connectivity:

A point "p" and "q" are connected based on density, if there exist a point "r" which has enough number of points in its neighbors and both the points "p" and "q" are within the ε distance. The process works continuously. So, if "q" is neighbor of "r" and "r" is the neighbor of "s", and "s" is the neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p"[11].

Advantages:

- Does not require a-priori specification of number of clusters.
- Able to identify noise data while clustering.
- DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.

Disadvantages

- DBSCAN algorithm fails in case of varying density clusters.
- Fails in case of neck type of dataset.

B. VDBSCAN ALGORITHM

VDBSCAN algorithm for the purpose of varied density dataset analysis as well as selects automatically several values of parameters Eps for different densities [2].

Description of the algorithm:

- It calculates and stores k-dist for each project and partition the k-dist plots.

- The number of densities is given intuitively by k-dist graph.

- The parameters Eps_i is selected automatically for each density.

- Scan the dataset and cluster different densities using corresponding Eps_i.

- Display the valid cluster with respect to varied density.

Drawbacks of VDBSCAN:

The VDBSCAN algorithms have the problem of real life dataset [3]. Calculating the values of Eps_i the value of k is a user dependent input parameter in VDBSCAN algorithm. The performance and efficiency is largely hampered any examining dataset. In the k-dist plot some little changes show up for changing the density level of examining dataset it consider as the outliers. To overcome this problem, K is used as dataset dependent parameter.

III. OUR MOTIVATION

Our motivation is to propose 'K' as a dataset dependent parameter in VDBSACN. The dataset property based K in VDBSCAN also considers the same process. But in our proposed method select the better K value compares to dataset property based K selection.

The proposed method for the distance measurement using the jaccard index distance and using the quick select algorithm based K selection. Jaccard index distance to find more efficient distance and it is a statistical measure of the extent to which cases are multivariate outliers. It is a rule for calculating the distance between two points in data which is better adapted than the usual Euclidian distance and finds the probability based selection.

It measures similarity between data sets and distance between the means of two distributions. These have been used to find distance in dataset and more efficient.

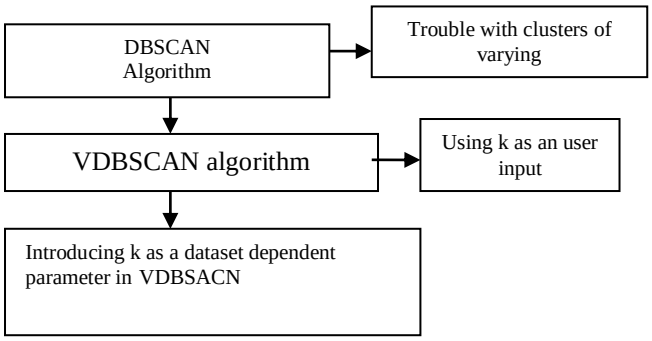


Fig 3: dataset property based 'k'

In this figure we show the overall process of the VDBSCAN clustering algorithm for datasets. we consider the input dataset to perform the VDBSCAN algorithm to find the distance of the each cluster . Using the K parameter we find the distance for dataset, jaccard index distance calculates the best distance values, also considered the probability of target points.

IV. PROPOSED ALGORITHM

Step 1: Create clusters from datasets of varying densities with different shapes and sizes, using VDBSCAN algorithm.

Step 2: To determining the value of K in varied density based spatial cluster analysis algorithm, we calculate the one point and find distance to all the other points from it and average it to find the average distance.

$$\mathbf{d\ (P_i) = \sum_{i=1}^{n} distance\ (Pi, Xi)\ /n\text{-}1}$$

Step 3: After finding the d (Pi) we have to calculate avg (d). Which is the average of all d (Pi), it is required to find out the Target Point (Ti).

$$\mathbf{Avg\ (d) = \sum_{i=1}^{n} d(Pi)/n}$$

Step 4: We have to determine the closest point which is nearest to the circumference of each circle by the following equation.

$$\mathbf{min|\ (distance\ (r - xi))\ |}$$

Step 5: We have to determine the Ti (Pos) of Ti for all Pi in the dataset, to determine the mode of Ti (Pos) for the whole dataset.

Step 6: it there is more than one mode values found, we compute the mean of maximum repeated Ti (Pos) or modes in the dataset of the different densities.

$$\mathbf{Mean\ value\ Ti\ (pos) = \sum_{i=1}^{m} T(mi)}$$

Maximum repeated modes /total number of modes

Step 7: The quick select based selection concept will determine the best k value in VDBSCAN algorithm.

```
function quickSelect(list, left, right, k)
 if left = right
 return list[left]
Select a pivotIndex between left and right
 pivotIndex := partition(list, left, right,
                         pivotIndex)
  if k = pivotIndex
    return list[k]
  else if k < pivotIndex
    right := pivotIndex - 1
  else
    left := pivotIndex + 1
```

Here ni is the total number of the values in the every position.

ALGORITHM DESCRIPTION

In the first step we found the different size, spaces of the dataset for performing the cluster using the VDBSCAN algorithm. To determine the value of the K at the different or varied dataset in the VDBSCAN, calculate the distance position

(d(pi)) for each of the position to other position ,find the average distance for each the Target Point (Ti).we determine the closest point which nearest to the circle at every position Xi. after found the Xi, calculate the mode Ti(pos) or the maximum number of the repeated modes in the whole dataset, if the dataset contains the more number of the repeated mode values ,calculate the mean value of the mode, for example 5,4,2 is the modes ,

Mean value Ti (pos) = (5+4+3)/3 =4

This type of the mode value is not optimal value of the K in the VDBSCAN so we propose our method to found the quick select density based selection calculation to best K value in the result.

Input: [7, 4, 6, 3, 9, 1]
k=2

Output: k'th smallest array element is 3

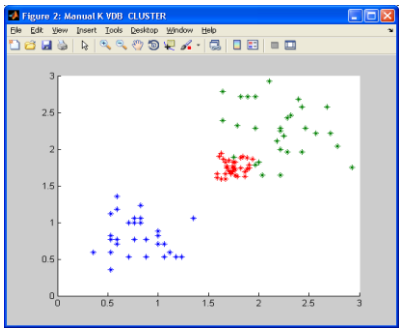
Input: [7, 4, 6, 3, 9, 1]
k=1

Output: k'th smallest array element is 1In this way we reduce the value of the k in the VDBSCAN algorithm and found the best value K.

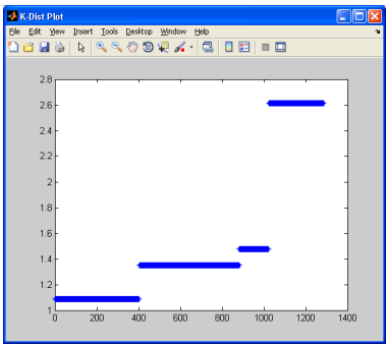
V. RESULTS AND DISCUSSION

DATASET

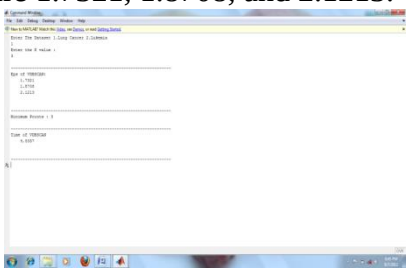
The proposed work has been implemented by using microarray gene expression dataset. Lung cancer dataset and leukemia dataset has been taken for this implementation. The algorithms were applied to gene expression data: the leukemia dataset, and the lung cancer dataset. Our proposed method using two types of datasets first we select the lung cancer dataset for clustering. The VDBSCAN algorithm uses the parameter K as user independent parameter. The k value is selected as 3 for Eps selection the three clusters are formed.



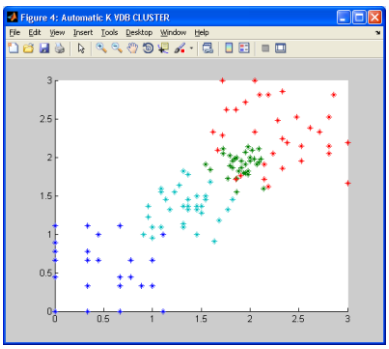
The VDBSCAN clustering k-dist plots are formed. In k-dist plot some little changes show up for changing the density level clustering it just discarded as outliers.



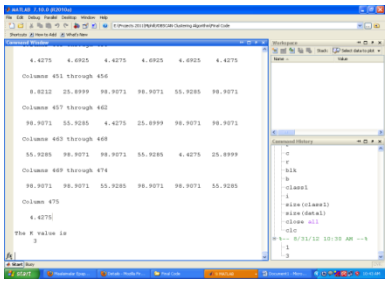
VDBSCAN algorithm three Eps values are 1.7321, 1.8708, and 2.1213.



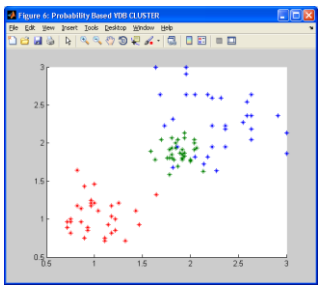
After calculating the distance and finding average determination of lung cancer dataset the dataset dependent k value is 4.the four clusters are formed.



Our proposed method using the same K as dataset dependent parameter is computed from average determination and distance measurement using the jaccard index distance with probability based K selection. The probability based K selection determining the k as 3.the command window is displayed in below,

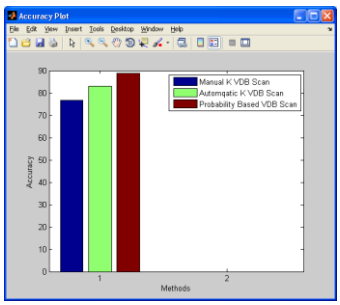


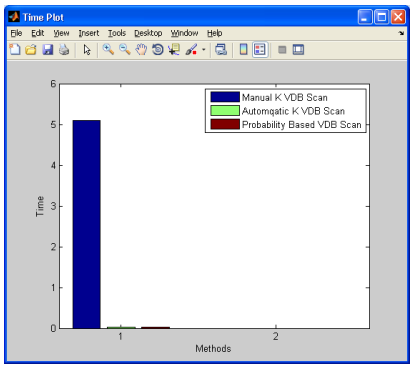
The clusters are formed.



VI. COMPARATIVE RESULTS

We compare the VDBSCAN, dataset property based K selection and quick select algorithm based K selection with ran index. Our proposed gives good computation time and accuracy and good cluster formation.





VII. CONCLUSION

VDBSCAN algorithm is one of the most efficient methods for creating clusters from dataset of varying density it create clusters of different shapes and sizes. But in VDBSCAN parameter K as a user input dependent parameter it largely degrades the efficient of permanent Eps so, the proposed method using k as a dataset dependent parameter with quick select algorithm for K value selection.

VIII. REFERENCES

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introducing to Data Mining”, Pearson Education Asia LTD, 2006.

[2] Peng Liu, Dong Zhou, Naijun Wu, “Varied density Based Spatial Clustering of Application with Noise”, 2007 IEEE.

[3] M.Parimala, Daphne Lopaz, N.C. Senthilkumar, "Survey on Density based Clustering Algorithm for mining large spatial databases", IJAST 2011.

[4] "An Optimized Density Based Clustering Algorithm", Volume 6– No.9, September 2010.

[5] Whelan, M.; Nhien-An Le-Khac; Kechadi, "Comparing two density-based clustering methods for reducing very large spatio-temporal dataset", Spatial Data Mining and Geographical Knowledge Services (ICSDM), IEEE International Conference, 2011, Page(s): 519 – 524.

[6] Xiaoping Yang; Lingmin He; Huijuan Lu, "A Clustering Algorithm for Datasets with Different Density", Computer Technology and Development, ICCTD '09, 2009, Page(s): 504 – 507.

[7] Ram, A.; Sharma, A.; Jalal, A.S.; Agrawal, A.; Singh, "An Enhanced Density Based Spatial Clustering of Applications with No", Advance Computing Conference, IACC 2009, Page(s): 1475 - 1478 .

[8] Jingke Xi, "Spatial Clustering Algorithms and Quality Assessment", Artificial Intelligence, JCAI '2009, Page(s): 105 - 108.

[9] (DBSCAN) M Ester, H-P. Kriegel. J. Sander, and X, Xu. "A density-based algorithm for discovering clusters in large spatial databases. KDD'96 ", 1996.

[10] Xiao-Feng Wang and De-Shuang Huang, "A Novel Density-Based Clustering Framework by Using Level Set Method", VOL. 21, NO. 11, NOVEMBER 2009

[11]Christopher C. Yang and Tobun Dorbin Ng, "Analyzing and Visualizing Web Opinion Development and Social Interactions with Density-Based Clustering ", VOL. 41, NO. 6, NOVEMBER 2011.

[12]Bert L. Hartnell," Improving some bounds for dominating Cartesian products **"**Discussion's Mathematicians Graph Theory" 23 (2003) 261.

[13] Jason. Peterson, "Clustering overview", http://www.cs.ndsu.nodak.edu/~jasonpet/CSCI779/Clustering.pdf.

[14] Abu wahid md.Masud Parvez "data set property based 'K' in VDBSCAN Clustering Algorithm", WCSIT, 2012.