World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 2, No. 3, 115-119, 2012

Data set property based 'K' in VDBSCAN Clustering Algorithm

Abu Wahid Md. Masud Parvez

Software Quality Architect of Software quality department
Tech Prolusion Labs
San Francisco, USA

Abstract— The term cluster analysis (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. Among different types of cluster the density cluster has advantages as its clusters are easy to understand and it does not limit itself to shapes of clusters. But existing density-based algorithms are lagging behind. The main drawback of traditional clustering algorithm which was largely recovered by VDBSCAN algorithm. But in VDBSCAN algorithm the value of parameter 'K' which was a user input dependent parameter. It largely degrades the efficiency of permanent Eps. In our proposed method the Eps is determined by the value of 'k' in varied density based spatial cluster analysis by declaring 'k' as variable one by using algorithmic average determination and distance measurement by Cartesian method and Cartesian product on multi dimensional spatial dataset where data are sparsely distributed. The basic idea of calculated 'k' which is computed from the characteristics of the examining dataset instead of a static user dependent parameter for increasing the efficiency of the VDBSCAN cluster analysis algorithm. By calculating value of 'k' with our newly developed arithmetic and algebraic method, user will obtain the most optimal value of Eps for determining cluster for the sparsely distributed dataset. This will add significant amount of efficiency of the VDBSCAN cluster analysis algorithm.

Keywords— Data mining; Cluster analysis; Clustering algorithm; DBSCAN algorithm & Ep.

I. INTRODUCTION

There are mainly five types of clustering methods. They are partition method (the most popular partition methods are K-means method and K-Medoids Method), Hierarchical Method (Agglomerative and Divisive Hierarchical Clustering, BIRCH: Balanced Iterative Reducing and Clustering, ROCK: A Hierarchical Clustering Algorithm for Categorical Attributes, Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modelling), Grid based method (STING: Statistical Information Grid Method, WaveCluster: Clustering Using Wavelet Transformation), Model-base method (EM (Expectation-Maximization) Method, COBWEB Method, SOM (Self-Organizing Feature Map) Method). Density-

Based Method (DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density, OPTICS: Ordering Points to Identify the Clustering Structure, DENCLUE: Clustering Based on Density Distribution Functions).

II. PREVIOUS WORK

DBSCAN's definition of a cluster is based on the notion of density reach ability. Basically, a point *p* is directly density-reachable [2,3,4,10] from a point *q* if it is not farther away than a given distance ε (that is *p* is the part of *q' s* ε-neighbourhood), and if ε-neighbourhood of *q* contains at least minimum number of points (>=minPts) such that one may consider *p* and *q* be part of a cluster and *q* as a core point then *p* is called directly density-reachable from *q*.
And again a point *p* is density-reachable [108,26] from point *q* if there is a chain of points P1....Pn, $p_1 = q$ and $p_n = p$ such that p_{i+1} is directly density-reachable from p_i.
Here the relation of density-reachable is not symmetric (since *q* might lie on the edge of a cluster, having insufficiently many neighbours to count as a genuine cluster element).
So another term comes which is known as density-connected. It is defined in this way:

Though there was huge improvement in approach was done in clustering aspect by DBSCAN for spherical date but if we go though then we can find the following main points regarding our concern. DBSCAN contains various disadvantages in creating or forming clusters. The disadvantages [2,5,3,11] of DBSCAN algorithm is given below—

- DBSCAN can only result in a good clustering as good as its distance measure that used in its function of getting neighbors. The most common distance metric used is the Euclidean Distance measure.

Especially for high-dimensional data, this distance metric can be rendered almost useless.

- DBSCAN does not respond well to data sets with varying densities (which is also called hierarchical data sets)

III. VDBSCAN ALGORITHM

To overcome the disadvantages of DBSCAN algorithm 3 Chinese scientists introduced a new algorithm named VDBSCAN. The DBSCAN algorithm [4,14,11] can form clusters of different shapes and sizes. But the DBSCAN algorithm had the problem of determining clusters from datasets of varying densities. Although DBSCAN algorithm can form clear clusters from datasets where density of the datasets is not much varied but it cannot form clusters from datasets of varying densities. Also it had the problem of determining clusters from high dimensional dataset. So to solve these problems an improved version of DBSCAN algorithm was created which can create clusters from datasets of varying densities. This is known as VDBSCAN algorithm. To work with VDBSCAN we have to work [2,3,4] in the following way—

- Firstly, VDBSCAN calculates and stores k-dist for each project and partition k-dist plots.
- Secondly, the number of densities is given intuitively by k-dist plot.
- Thirdly, choose parameters $\mathbf{Eps}_i$ automatically for each density.
- Fourthly, scan the dataset and cluster different densities using corresponding $\mathbf{Eps}_i$
- Finally, display the valid clusters corresponding to varied densities.

To work with VDBSCAN algorithm we have to follow two steps [10] regard. These two steps are—

- Choosing parameters $\mathbf{Eps}_i$
- Clustering according to varied density

IV. LIMITATIONS OF VDBSCAN

There were certain problems in DBSCAN algorithm. To overcome those problems VDBSCAN algorithm was introduced. But then also the VDBSCAN algorithm contained some problem with this real life dataset. The main problems of VDBSCAN algorithm is given below—

- For calculating the value of $\mathbf{Eps}_i$**,** the value of **K** is required. But again the value of **K** is a user dependant input parameter in VDBSCAN algorithm. So the performance and efficiency is largely hampered for any examining dataset because we are considering the value the value of **K** without considering the characteristics (density, dimension etc.) of the examining dataset.
- In the K-dist plot some little changes show up for the changing density level of the examining dataset. But finally after a certain time a sharp change shows up and according to the VDBSCAN algorithm the data corresponding to this sharp changed level are
- just discarded as outliers. But some of these data may also be important for us. They can be part of a cluster other than considering as outliers.

V. MOTIVATION

Our motivation for proposing `k` as a dataset dependent parameter in VDBSCAN algorithm. That is described with the aid of a figure (Fig: 1) below—
From this diagram it is shown that how we were motivated to introduce `K` as a dataset dependent parameter in VDBSCAN algorithm. Here in our proposed method we tried to offer some benefits regarding VDBSCAN algorithm and minimize the drawbacks that it had when we used `K` as a user input dependent parameter.

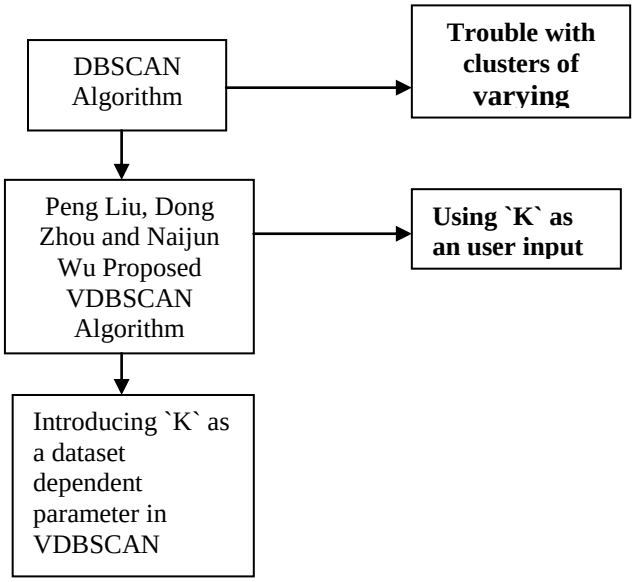

Figure 1: Motivation for Introducing `K` as a Dataset Dependent Parameter in VDBSCAN Algorithm

VI. OUR PROPOSED METHOD

Our proposed method introduces an efficient method for determining the value of K in varied density based spatial cluster analysis algorithm. In our proposed method, we are declaring K as variable which is determined by algorithmic average determination and distance measurement by Cartesian method and Cartesian product on multi dimensional spatial dataset which are sparsely distributed.

VII. OUR DEVELOPMENT

First let's take a multidimensional data plot. For your mathematical simplicity let's take a two dimensional date plot. Suppose it has n points. And we will find out all the points average one to all other points distance to other points. So first let's consider one point and find distance to all the other points from it and average it to find the average distance.

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

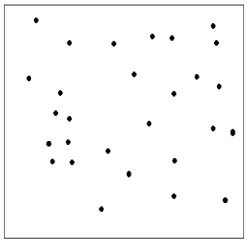

Fig 2: Sparsely Distributed Two-Dimensional Assumed Dataset

Here,

d(Pi)=Average distance from **Pi** to all other points in the data set.

We have to find out **d(Pi)** for all **Pi**.

Now we have to calculate **avg(d)**. Which is the average of all **d(Pi)**.And it is required to find out the **Target Point** (**Ti)** .

$$\mathbf{avg(d)} = \frac{\sum_{i=1}^{n} \mathbf{d(Pi)}}{\mathbf{n}}$$

For every **Pi** in the datasets we will draw a circle and the centre of the circle will be the points itself means **Pi**, and the radius of each circle will be the **avg(d)**.So area of each circle will be same. Here we conceive only the circumference of each circle.

Here,

Pi=Subjective Point or Centre of the Circle

r =**avg(d)** (Radius of each Circles.)

For every circle we have to determine the closest point which is nearest to the circumference of each circle by the following equation.

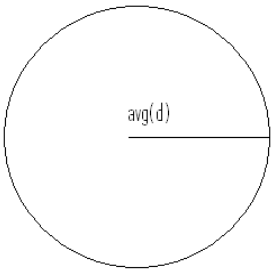

Fig 3: The Role of avg(d) in One Supposed Core

min| (distance (r – xi)) |

Xi is the point which has minimum distance from the circumference of a particular circle for the corresponding **Pi** which is the centre of that circle. And for that **Pi** we make **Xi** as a **Target Point** and tag as **Ti** .We have to find out **Ti** for every **Pi**.

Then we have to determine the position of **Ti** relative to the **Pi** for that particular circle.

Ti(Pos)=Position of the **Ti** relative to the **Pi** of a particular circle.

In this way we will determine the **Ti (Pos)** of **Ti** for all **Pi** in the dataset.

Now we have to determine the mode of **Ti(Pos)**.That means we have to find out maximum repeated **Ti(Pos)**.If there is more than one mode then we have to compute the mean of maximum repeated **Ti(Pos)**s or modes.

Mode of **Ti(Pos)** is basically our expected value of parameter **K** in the **K-dist** plot

VIII. PERFORMANCE ANALYSIS

To analyse our proposed method's performance practically we coded our proposed method mechanism in C++ language and added it with the VDBSCAN algorithm. There we generated K-dist graphs from the value of **`K`** determined by our developed method and also we will generated K-dist graphs by taking the value of **`K`** as a user input dependent parameter. After taking the value of **`K`** in these three ways, three different graphs was simulated. And then depending on the graphs we practically analysed the performance.

A. Taking random points for simulation

First, we took 20 random input points in the three-dimensional plane. Then we applied our developed method to find the most optimal value of **`K`** for this particular dataset containing with those 20 random points.

According to our developed system we found that the appropriate value of **`K`** is 4 for the data sets introduced in **Fig 4.**

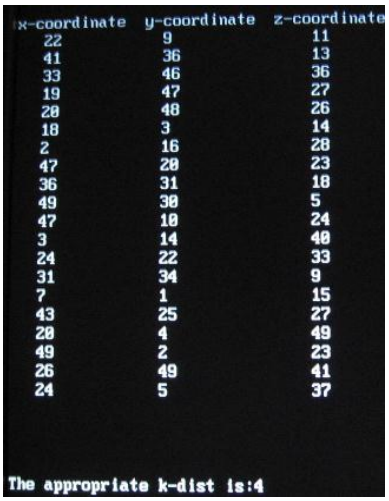

Fig 4: Initial Input Points (Three-Dimensional) for Determining the Value of `K`.

B. Simulation of K-dist Graph (According to the value of 'K' Determined by our proposed method)

The values to plot against the x-axis which are returned from the VDBSCAN scan algorithm integrated with our proposed method in Figure 5.

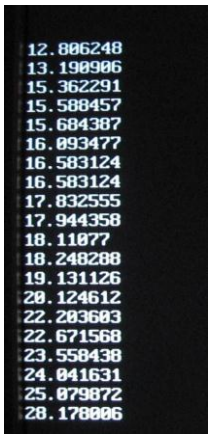

Fig 5: Values returned by our proposed method to build up the graph

When we put the retuned values against the x-axis then we get the following graph

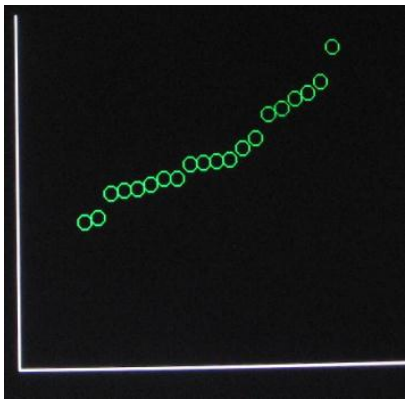

Fig 6**:** Simulated K-dist Graph (According to the Value of **`K`** (K=4) Determined by Our Method).

Graph Evaluation: From this graph (Fig 6) we can see that the graph level turning line is less varied and organized. And those small jumps took place only few times in our examining dataset. And also we have considered almost all points to define clusters that have reduced the probability for a particular point to becoming an outlier. The sharp change took place only once at the end (where we place the points on the X-axis in ascending order) of the k-disk graph and the points corresponding to the sharp change is going to be discarded as they are considered as outliers by this algorithm.

C. Stimulation of K-dist Graph (taking the value of "K" as a user input parameter)

For K=6, the simulated K-dist graph was followed for our examining dataset given in Fig 7

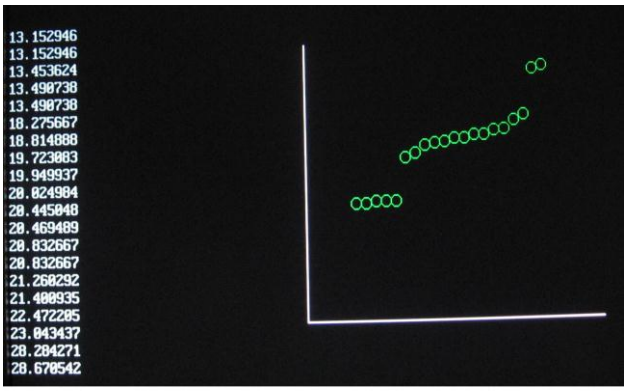

Fig 7: Simulated K-dist Graph According to the Value of `K` (K=6)

Graph Evaluation: From this K-dist graph **(Fig: 7)** we get the sharp change two times. This defines that in this dataset a lot of data are described as noise and outliers and these beginning points and the ending points will be discarded corresponding to the sharp change. That's why a lot of data can be discarded as outliers though they can be very important in several cases. Here the level turning lines are also not clearly shown.

D. Stimulation of K-dist Graph (taking the value of "K" as a user input parameter)

For K=10 the simulated K-dist graph was followed for our examining dataset given in Fig: 8.

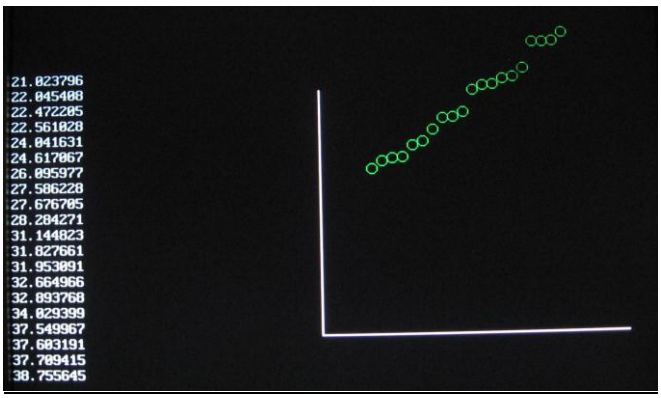

Fig 8: Simulated K-dist Graph (According to the Value of **`K`** (K=10) as User Input Parameter)

Graph Evaluation: From this K-dist graph **(Fig: 8)** we can see that there is mainly two problems. One is the lack of density in the cluster and other is there is no clear sharp change. In the graph we can see that the total graph is growing upwards. And for this it is really difficult to calculate the actual density level and the level turning lines. And as there is no particular sharp change in the graph. So it is also very difficult to calculate the value of **Eps**.

IX. DRAWBACK MINIMIZED BY OUR PROPOSED METHOD

Finally we are able to reach to a decision that our proposed method minimized several drawbacks from VDBSCAN algorithm. They are given below—

For calculating the value of $\mathbf{Eps_i}$, the value of **K** is required. But again the value of **K** is a user dependent input parameter in VDBSCAN algorithm. So the performance and efficiency is largely hampered for any examining dataset

because we are considering the value the value of **K** without considering the characteristics (density, dimension etc.) of the examining dataset. Our proposed method has minimized this problem by introducing `K` as a dataset dependent parameter rather than user input dependent parameter.

In the **K-dist** plot some little changes show up for the changing density level of the examining dataset. But finally after a certain time, a sharp change shows up and according to the VDBSCAN algorithm the data corresponding to this sharp changed level are discarded as outliers. But if that process is not enough efficient then we may lose some of those data which may also be important for us. They can be part of cluster other than considering as outliers. In our proposed method this problem has been minimized

X. CONCLUSION

VDBSCAN algorithm is one of the most efficient methods for creating clusters from dataset of varying density. Also it can create clusters of different shapes and sizes. But taking the parameter `K` as a user input dependent parameter and without considering the characteristics of the dataset into account, made the algorithm less efficient. But in our proposed method we introduced the value of `K` from the examining characteristics of the dataset. We welcome others to work with the two Sharpe changes attitude and duration in the k-dist graph regarding its probability and its position with multidimensional varied density date set.

ACKNOWLEDGMENT

Special thanks to Prof. Hawlader Abdullah Al-Mamun for his kind guide line on this research work.

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] Hai-Dong Meng; Yu-Chen Song; Fei-Yan Song; Hai-Tao Shen, "Application research of cluster analysis and association analysis", Software Engineering and Data Mining (SEDM), 2010 2nd International Conference, 2010 , Page(s): 597 – 602,IEEE conference publication.

[5] Yang Fan; Rao Yutai, "A Density-based Path Clustering Algorithm", Intelligent Computation and Bio-Medical Instrumentation (ICBMI), 2011, IEEE conference publication.

[6] 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

[7] Xiaobing Yang; Lingmin He; Huijuan Lu, "A Clustering Algorithm for Datasets with Different Densit", Computer Technology and Development, ICCTD '09, 2009, Page(s): 504 - 507

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

[9] Stephen Haag et al. (2006). Management Information Systems for the information age. Toronto: McGraw-Hill Ryerson. pp. 28. ISBN 0-07-095569-7. OCLC 63194770.

[10] http://en.wikipedia.org/wiki/DBSCAN

[11] 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

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

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

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

Abu Wahid Md. Masud Parvez received his Graduation degree at Computer Science and Information Technology from Islamic University of technology (IUT). He was bored at 23rd April 1986.

Masud parvez is currently working as Software Quality Architect in Tech propulsion labs (USA), currently he is posted at Asia brunch of the company. Previously he was working as Research Engineer in Electronics Research and development center, Walton.