Extended Varied DBSCAN on Multiple Data Sources

B.Yugandhar [1], P.Vinod Babu [2]

[1]*Final M Tech Student,*[2]*Asst. Professor*

Dept of Computer Science and Engineering1, Dept of Computer Science and Engineering2,
1 Swarnandhra college of Engineering and Technology, seetharampuram, Narspur-534280,w.g.dt,A.P.
[2]Swarnandhra college of Engineering and Technology, seetharampuram,Narspur-*534280,w.g.dt,A.P.*

Abstract:**Density based clustering algorithms are one of the primary method for data mining. VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise) is introduced to compensate problem because Existing density based algorithms are not capable of finding out all meaningful clusters whenever the density is so much varied. It is same as DBSCAN (Density Based Spatial Clustering of Applications with Noise) but only the difference is VDBSCAN selects several values of parameter *Eps* for different densities according to k-dist plot. The problem is the value of parameter k in k-dist plot is user defined. This paper introduces a new combined approach EVDBSCAN (Extended Varied Density Based Spatial Clustering of Applications with Noise) to find out the value of parameter k automatically based on the characteristics of the datasets from two different data sources for further Cluster analysis.**

In this method we consider spatial distance from a point to all others points in the datasets from two different data sources to get the integrated optimum solution. The proposed method has potential to find out optimal value for parameter k from different data sources .In this paper we are considering two different types data sources for the proposed approach for demonstration.

Keywords-- Density based clustering, DBSCAN, VDBSCAN, data mining, center based approach. Data integration.

1. Introduction:

Clustering analysis is a primary method for data mining. There are five areas of clustering, which are Partitioning, Hierarchical, Density, Grid, and Model methods. Density clustering methods are very useful to find clusters of any Shape, giving the correct parameters (yet hard to determine them). the goal of a clustering algorithm is to group the objects of a database into a set of meaningful subclasses). Traditional algorithms, such as DBSCAN can have trouble with density if the density of clusters varies widely. To resolve the problems of DBSCAN and VDSCAN, EVDBSCAN is introduced for the purpose of effective clustering analysis of datasets with varied densities from different data sources by integrating information from systems. It selects suitable parameters for different density, using k-dist plot, and adopts VDBSCAN algorithm for each chosen parameter. But the value of parameter k in k-dist plot is user defined. There is no logical derivation for choosing the value of parameter k.

We introduced a new combined method for determining the value of parameter k in k-dist plot of two different source datasets by integrating the data sources for further analysis i.e. forecasting, decision-making purpose. Our proposed method can select suitable value for parameter k based on the characteristics of the dataset by automatically scanning the data source datasets.

This paper is organized as follows: Section 2 describesrelated work, Section 3 describes the Proposed System architecture and system nature, Section 4 presents experimentand analysis. Section 5 presents conclusion and future scope

II. RELATED WORKS

Before going to the EVDBSCAN (Extended Varied Density Based Spatial Clustering of Applications with Noise) algorithm we need to understand how DBSCAN and VDBSCAN work.DBSCAN algorithm is based on center-based approach. In the center-based approach, density is estimated for a particular point in the dataset by counting the number of points within a specified radius, *Eps*, of that point. This includes the point itself. The center-based approach to density allows us to classify a point as a core point, a border point, a noise or background point. A point is core point if the number of points within *Eps*, a user-specified parameter, exceeds a certain threshold, *MinPts*, which is also a user specified parameter.

The basic approach of how to determine the parameters *Eps* and *MinPts* is to look at the behavior of the distance from a point to its **k**th nearest neighbor, which is called **k**-dist. The **k**-dists are computed for all the data points for some k, sorted in ascending order, and then plotted using the sorted values; as a result, a sharp change is expected to see. The sharp change at the value of **k**-dist corresponds to a suitable value of *Eps*. Line A in "Fig.2" shows a sample k-dist line. Note that the value of *Eps* that is determined in this way depends on **k**, but does not change dramatically as **k** changes. DBSCAN can find many clusters that could not be found using some other clustering algorithms, like K-means. However, the main weakness of DBSCAN is that it has trouble when the clusters have greatly varied densities. To sweep over the limitations of DBSCAN, VDBSCAN is acquainted.

The DBSCAN algorithm is not capable of finding out meaningful clusters with varied densities. VDBSCAN algorithm detects cluster with varied density as well as automatically selects several values of input parameter Eps for different densities. Even the parameter k is automatically generated based on the characteristics of the datasets.

VDBSCAN has two steps: choosing parameters *Epsi*and cluster in varied densities. Details are given in "Fig.3".

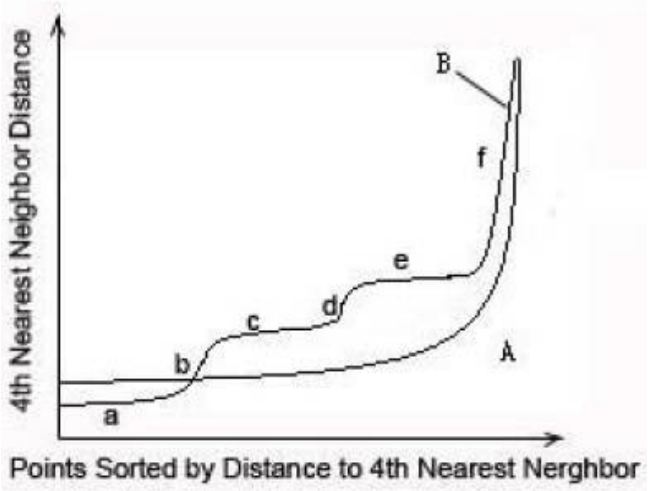


Fig 1. A Simple k-dist plot

EVDBSCAN is a three tier architectural process shown in fig 2, the proposed model first connects to the two different data sources with proper connections and accesses the databases at same time to process the combined approach. The proposed method adopts the VDBSCAN after the database selection process is completed and determining the value of parameter k in k-dist plot. Our proposed method can select suitable value for parameter k based on the characteristics of the dataset.

III. DESCRIPTION OF PROPOSED METHOD

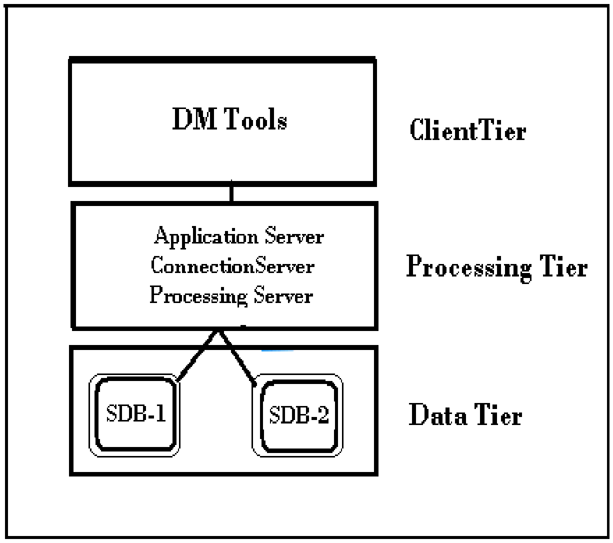


Fig 2EVDBSCAN System Architecture

3.1 Description of the Architecture

Tier-1: Client Tier: The client tier consists of the Cluster client tools of Data mining, The System connects to the Databases by creating valid connections to the databases view the optimal result.

Tier-2: processing Tier: The processing tier process to establish connection to the different data sources to merge required data sets extracting data from different databases and proceses the VDBSCAN method to detects cluster with varied density as well as automatically selects several values of input parameter Eps fordifferent densities. The tier sends cluster result information to the client Viewers.

Tier-3: Data Tier: The Data tier is made up of databases that contained data used in the EVDBSCAN.

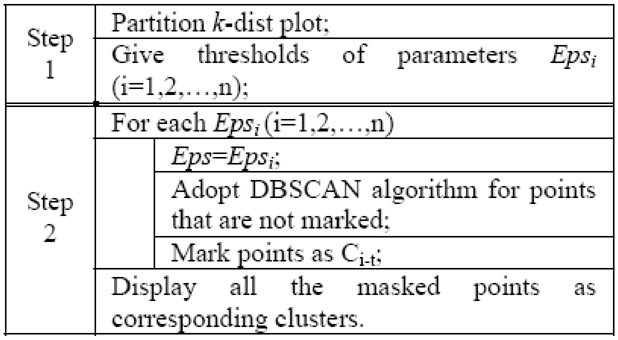
3.2 Work Nature of EVDBSCAN

The main goal of the proposed system is to detect effective clusters from different data source data sets by integrating the data from the different sources of information. The system incorporates the VDBSCAN can select suitable value for parameter k based on the characteristics of the dataset.

The user may select one or more databases from the client tool of EVDBSCAN and connects to databases through valid connections specified in the system and apply cluster Analysis method of proposed system to effectively detect clusters and can view valid cluster with respect to varied density.

The proposed system accepts the data specified by user and combines the required data for the datasets and applies theVDBSCAN algorithm detects cluster with varied density as well as automatically selects several values of input parameter Eps for different densities. Even the parameter k is automatically generated based on the characteristics of the datasets.

VDBSCAN has two steps: choosing parameters *Epsi*and cluster in varied densities. Details are given in “Fig.3”.



Step 1	Partition *k*-dist plot;	
	Give thresholds of parameters Eps_i (i=1,2,…,n);	
Step 2	For each Eps_i (i=1,2,…,n)	
		Eps=Eps_i;
		Adopt DBSCAN algorithm for points that are not marked;
		Mark points as $C_{i\text{-}t}$;
	Display all the masked points as corresponding clusters.	

3.3Description of the VDBSCAN Method

A. Step1

In our two dimensional data grid there are n points. We notate every points P_i as Subjective points.

P_i=Subjective point (i=1,2…,n)

$$d(P_i) = \frac{\sum_{i=1}^{n} \text{distance} \quad (P_i, x_i)}{n-1}$$

$d(P_i)$=Average distance from P_i to all other points in the data set.
We have to find out $d(P_i)$ for all P_i.

B. Step 2

Now we have to calculate avg(d). Which is the average of all $d(P_i)$.And it is required to find out the Target point T_i .

$$avg \quad (d) = \frac{\sum_{i=1}^{n} d(P_i)}{n}$$

C. Step 3

For every P_i in the datasets we will draw a circle and the center of the circle will be the points itself means P_i, and the radius of each circle will be the avg(d).So area of each circle will be same but it's not concern. Here we conceive only the circumference of each circle.

Here

P_i=subjective point or center of the circle
r =avg(d) (radius of each circles.)

D. Step 4

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

$$\min \left| \ (\text{distance}(r - x_i)) \ \right|$$

X_i is the point which has minimum distance from the circumference of a particular circle for the corresponding P_i which is the center of that circle. And for that P_i we make X_i as a Target point and tag as T_i .We have to find out T_i for every P_i.

E. Step 5

Then we have to determine the position of T_i relative to the P_i for that particular circle.

T_i(Pos)=Position of the T_i relative to the P_i of a particular circle.

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

F. Step 6

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

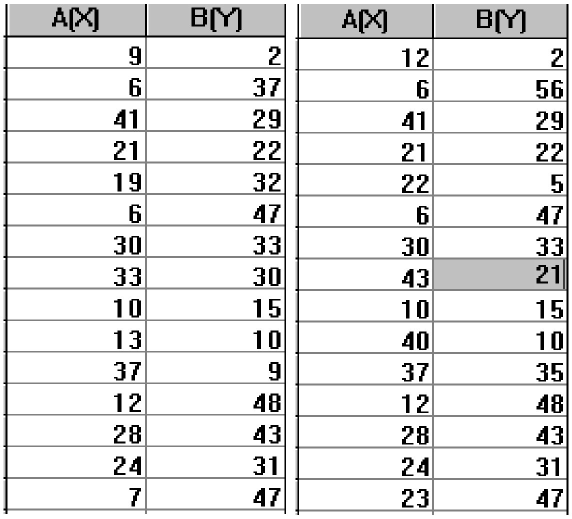
Mode of T_i(Pos) is basically our expected value of parameter k in the k-dist plot.

IV. EXPERIMENT AND ANALYSIS

A. Experimental Data: In order to observe and analyze experimental results directly, 2-dimension data of sample spatial information is chosen to prosecute our experiment. First, the system combines the data from different source to perform VDBSCAN on the required data set to perform clustering .the points are uniformly distributed in the data grid. We have implemented our proposed method in .Net And the coordinate of our observational points are given in Table 1.

B. Experiment Process and Result:According to our experimental data the value of parameter k for the k-dist plot is 10.Because we have pursued our experiment only with 30 points, 15 points from SDB1 and another 15 points from SDB2 for the simplicity. And they are sparsely distributed. That's why the value of the parameter k is slightly bigger. If the dataset is much denser then the value will be smaller. The value of the parameter k is 10, means for plotting k-dist we will consider the distance of the 10th nearest neighbor. Here we have sorted the distance in ascending order and plot them in y axis and plot the points in x axis according to their respective distance.

TABLE I TABLE 1. DATA FOR EXPERIMENT



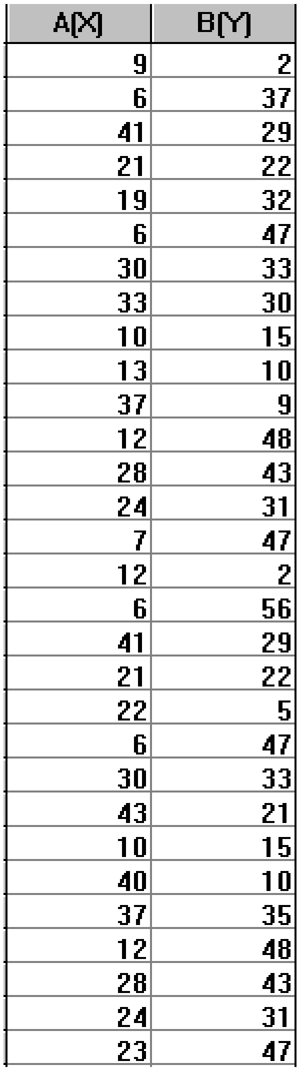
A[X]	B[Y]
9	2
6	37
41	29
21	22
19	32
6	47
30	33
33	30
10	15
13	10
37	9
12	48
28	43
24	31
7	47

A[X]	B[Y]
12	2
6	56
41	29
21	22
22	5
6	47
30	33
43	21
10	15
40	10
37	35
12	48
28	43
24	31
23	47

SDB-1 Sample data **SDB-2Sample data**

Integrated data from different sources shown in table2

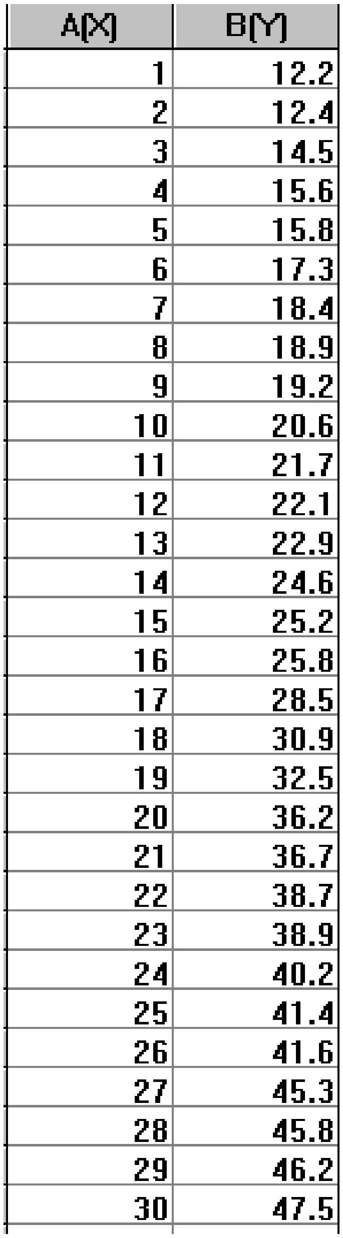
TABLE II



A[X]	B[Y]
9	2
6	37
41	29
21	22
19	32
6	47
30	33
33	30
10	15
13	10
37	9
12	48
28	43
24	31
7	47
12	2
6	56
41	29
21	22
22	5
6	47
30	33
43	21
10	15
40	10
37	35
12	48
28	43
24	31
23	47

The values are shown in Table 3.

TABLE III POINTS SORTED BY DISTANCE



A[X]	B[Y]
1	12.2
2	12.4
3	14.5
4	15.6
5	15.8
6	17.3
7	18.4
8	18.9
9	19.2
10	20.6
11	21.7
12	22.1
13	22.9
14	24.6
15	25.2
16	25.8
17	28.5
18	30.9
19	32.5
20	36.2
21	36.7
22	38.7
23	38.9
24	40.2
25	41.4
26	41.6
27	45.3
28	45.8
29	46.2
30	47.5

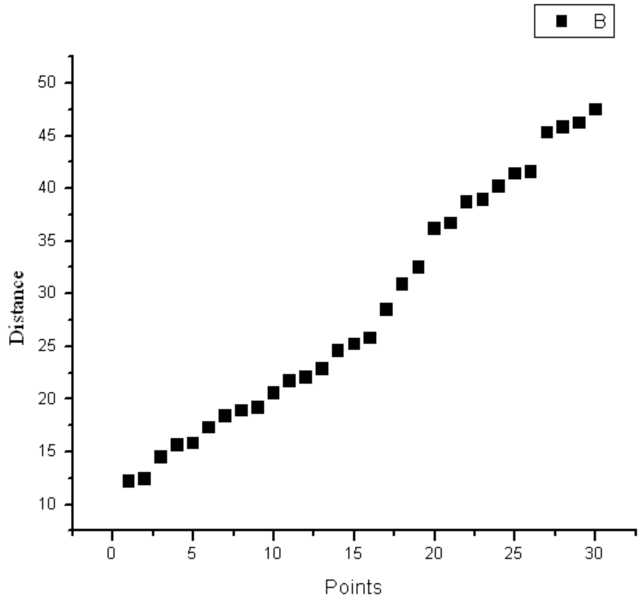


Figure 4. K-dist plot(k=10)

C. Experiment conclusion

Synthetic database with two 2-dimension data sample data is extracted from different data sources is used for demonstration. There are several incisive changes at the value of k-dist I .So we can get several incisive changes at the value of k-dist that's mean our proposed method can identify the k value for a particular datasets based on the demeanor of the datasets for measuring the value of *Epsi* which will be used for VDBSCAN to find outall meaningful clusters for that datasets which has varieddensities.

V. CONCLUSION

In this paper we proposed an extended method EVDBSCAN to perform VDBSCAN on merged datasets from different two data sources. The method to choose the value of parameter k for the k-dist plots to find out meaningful clusters for VDBSCAN algorithm. The experiment shows that the value of the parameter k for the k-dist plot should be chosen based on the characteristics of the datasets and our proposed method is capable of finding out the value which is based on the behavior of the datasets. However, there are several opportunities for future research. What are the consequences of the magnitude of parameter k for a particular datasets is one of the interesting challenges as well as the relationship with the value of parameter k and character of changes at the value of k-dist plot.

REFERENCES

[1] Peng Liu, Dong Zhou, Naijun Wu, “Varied Density Based SpatialClustering of Applications with Noise”, 2007, IEEE.

[2] Pang-Ning Tan, Michael Steinbach, VipinKumar, Introduction toData Mining, Pearson Education AsiaLTD, 2006.

[3] ain A. K., Dubes R. C., Algorithms for clusteringData, Prentice- Inc.,1988.

[4] MihaelAnkerst, Markus M. Breunig, Hans-Peter Kriegel, JörgSander, “OPTICS: Ordering points to identify the clusteringstructure”, proceeding of 1999 ACM-SIGMOD InternationalConference, ACM Press, pp.49-60, 1999.

[5] Sun Xue-gang, Chen Qun-xiu, and Ma Liang, “study on topic-basedweb clustering”, The Journal of Chinese Information Processing,Vol 17, No. 3, pp.21-26, 2003.

[6] Hinneburg A., Keim D., “An efficient approach to clustering in largemultimedia databases with noise”, Proceeding of 4th InternationalConference on Knowledge Discovery and Data Mining, New York City, NY, 1998.

[7] Sheikholeslami G., Chatterjee S., Zhang A., “WaveCluster: Amultiresolution clustering approach for very large spatial databases”,Proceeding 24th International Conference on Very Large Data Bases, pp. 428-439, New York City, NY,1998.