VDBSCAN:
Varied Density Based Spatial Clustering of Applications with Noise

Peng Liu, Dong Zhou, Naijun Wu

School of Information Management and Engineering,
Shanghai University of Finance and Economics, Shanghai, 200433, China
liupeng@mail.shufe.edu.cn

ABSTRACT

Clustering analysis is a primary method for data mining. Density clustering has such advantages as: its clusters are easy to understand and it does not limit itself to shapes of clusters. But existing density-based algorithms have trouble in finding out all the meaningful clusters for datasets with varied densities. This paper introduces a new algorithm called VDBSCAN for the purpose of varied-density datasets analysis. The basic idea of VDBSCAN is that, before adopting traditional DBSCAN algorithm, some methods are used to select several values of parameter *Eps* for different densities according to a k-dist plot. With different values of *Eps*, it is possible to find out clusters with varied densities simultaneity. For each value of *Eps*, DBSCAN algorithm is adopted in order to make sure that all the clusters with respect to corresponding density are clustered. And for the next process, the points that have been clustered are ignored, which avoids marking both denser areas and sparser ones as one cluster. Finally, a synthetic database with 2-dimension data is used for demonstration, and experiments show that VDBSCAN is efficient in successfully clustering uneven datasets.

Keywords: density-based clustering, DBSCAN, VDBSCAN, data mining

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) [1].

Roughly speaking, the goal of a clustering algorithm is to group the objects of a database into a set of meaningful subclasses. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a traditional and widely-accepted density-based clustering method. It can find clusters of arbitrary shapes and sizes, yet may have trouble with clusters of varying density. The density-based algorithms still suffer from several problems. Traditional algorithms, such as DBSCAN and DENCLUE, can have trouble with density if the density of clusters varies widely. There are also some improvements which can handle clusters of different densities, like OPTICS and Jarvis-Patrick clustering algorithm [4]. However, they lower the cluster validity simultaneously.

We introduce a new improved density-based algorithm in the following, for the purpose of effective clustering analysis of datasets with varied densities. The algorithm is named VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise). It selects suitable parameters for different density, using k-dist plot, and adopts DBSCAN algorithm for each chosen parameter.

2. RELATED WORKS

Density-based approaches apply a local cluster criterion and are very popular for the purpose of database mining. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise).

A common way to find regions of high-density in the data space is based on grid cell densities [7]. The basic idea for the algorithm is that the data space is partitioned into a number of non-overlapping regions or cells, and cells containing a relatively large number of objects are potential cluster centers. However, the success of the method depends on the size of the cells which must be specified by the user.

DBSCAN algorithm is based on center-based approach, one of definitions of density [5]. 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 details of DBSCAN algorithm are given in Figure 1 [4]. And it can be informally described as follows. Any two core points that are close enough within a distance *Eps* of one another are put in the same cluster. Likewise, any border point that is close enough to a core point is

1-4244-0885-7/07/$20.00 ©2007 IEEE.

put in the same cluster as the core point. Noise points are discarded.

DBSCAN algorithm.

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters with its associated core points.

Figure 1. DBSCAN algorithm

OPTICS (Ordering Points to Identify the Clustering Structure) is an improved method upon DBSCAN, which is practically the same in runtime and process, but represents the clusters of objects by the ordering of objects in the database [2]. The main advantage of OPTICS is that the algorithm does not limit itself to one holistic parameter setting, which is a limitation of traditional DBSCAN, but just displays the point-order information based on densities, instead of clusters. And OPTICS is weak at finding out information of clusters in sparse datasets though it is good at finding them in dense areas [3].

Another improvement of the DBSCAN algorithm is DENCLUE [6], which is older than OPTICS. DENCLUE introduces the idea of an influence function that describes the impact of a data point upon its neighborhood. The algorithm has a good mathematical foundation, and scales well because it is able to process highly sparse datasets with minimal work (use of grid cells) while it requires very careful selection of parameters *MinPts* and *Eps* [1].

WaveCluster [8] is another density-based approach, which applies wavelet transform to the feature space. The algorithm is grid-based and only applicable to low-dimensional data. Also the density and grid-based clustering technique CLIQUE [9] has been proposed for mining in high-dimensional data spaces.

3. DESCRIPTION OF VDBSCAN ALGORITHM

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 Figure 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.

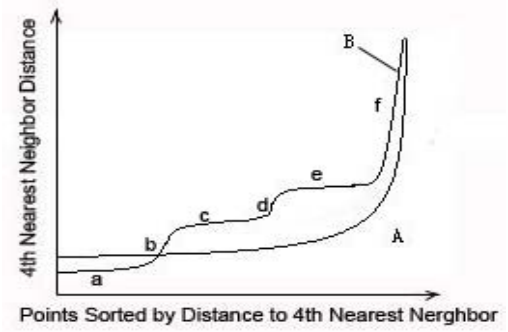

Figure 2. A sample k-dist plot

Because DBSCAN uses a density-based definition of a cluster, it is relatively resistant to noise and can handle clusters of different shapes and sizes. Thus, 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. Suppose that the noise around the denser cluster C_1 has the same density as the other cluster C_2. If the *Eps* threshold is low enough that DBSCAN finds C_2 as cluster, then C_1 and the points surrounding it will become a single cluster. If the *Eps* threshold is high enough that DBSCAN finds C_1 as a separate cluster, and the points surrounding are marked as noise, then C_2 and the points surrounding it will also be marked as noise. DBSCAN also has trouble with high-dimensional data because density is more difficult to define for such data. In this paper, we only focus on finding a solution for the main weakness when DBSCAN meets greatly varied densities.

In many real datasets, clusters with respect to different densities are all useful to analysis. It is necessary to find out both dense clusters and sparse ones. Very different partial densities may be needed to reveal clusters in different regions of the data space. So we introduce VDBSCAN, which is a new improved algorithm of traditional DBSCAN but goes beyond the limitation of single global parameter of DBSCAN, for the purpose of varied density-based clustering and analysis.

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 Eps_i automatically for each density. Fourthly, scan the dataset and cluster different densities using corresponding Eps_i. And finally, display the valid clusters corresponding with varied densities.
VDBSCAN has two steps: choosing parameters Eps_i and cluster in varied densities. Details are given in Figure 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-t};
	Display all the masked points as corresponding clusters.	

Figure 3. VDBSCAN algorithm

Step 1: Choosing Eps_i

This is a key step in the process. K-dist plot is drawn for not only selection of parameters Eps_i but also analysis of density levels of the dataset. For datasets with widely varied density, note that there will be some variation, depending on the density of the cluster and the random distribution of points, but for points of the same density level, the range of variation will not be huge while a sharp change is expected to see between two density levels. Thus there will be several smooth curves connected by greatly variational ones. If there are n (natural number n>1) different smooth curves in the k-dist plot, the dataset has n density levels. A dataset is of varied-density if it has several density levels and of n varied-density if it has n density levels. Specially, a dataset is of single-density if its density does not vary widely, or there is only one smooth curve in its k-dist plot. Figure 2 shows a sample k-dist plot. Line A shows a sample k-dist line of a single-density dataset while line B shows a sample line of a three varied-densities dataset.

For points that are not in a cluster, such as noise points, the corresponding k-dist line rockets, connecting two smooth curves which stand for two density levels. Line b and d in Figure 2 are such lines, which can be called level-turning lines. Line b connects line a and c, and line d connects c and e, while a, c and e stand for different density levels. Note that line f shows the k-dists of outliers and is not a level-turning line for it does not connect two smooth lines.

For different density levels Di, select suitable Eps_i. For example, in Figure 2, there are three density levels. Line a shows the densest density level and e shows the sparsest one. Combine line a and b as a sub-k-dist plot to select Eps_1, and then take line c and d as a sub-k-dist plot for Eps_2, e and f for Eps_3 finally.

Step 2: Varied-density clustering

Adopt algorithm DBSCAN for each Eps_i (natural number i=1,2,3,…,n. n is the number of density levels). Note that parameters Eps_i have been ordered as k-dist line curves, that is $Eps_i < Eps_{i+1}$ (i<n). Before adopt DBSCAN for Eps_{i+1}, mark points in clusters corresponding with Eps_i as C_{i-t} (t is a natural number), which indicates that the points belong to the cluster t in density level i. Marked points will not be processed by DBSCAN again. Non-marked points after all the Eps_i process are recognized as outliers. And all the C_{i-t} are displayed as the results.

4. EXPERIMENT AND ANALYSIS

4.1 Description of Data

In order to observe and analyze experimental results directly, 2-dimension data is chosen in our experiment. Figure 4 shows the data. Obviously, there are two regions with respect to different densities in the data. And data points of each region are uniformly distributed. The dataset provides a clustering standard to estimate the accuracy of the result, for it has strong regularity and obvious clusters. In addition, as it has been already acknowledged that density-based clustering algorithms can find out clusters with any shape, this paper will not discuss the problems of cluster shapes in clustering.

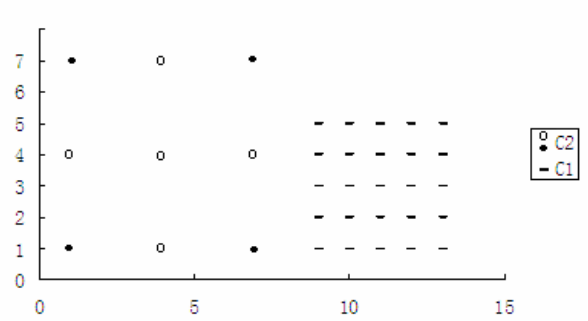

Figure 4. Data for experiment

4.2 Experiment Process and Result

Compute k-dist. Select k=3. Calculate the distance from each point to its 3rd nearest neighbor, which is called 3-dist, sort points by 3-dist and plot the sorted values. Figure 5 shows the k-dist plot (k=3).

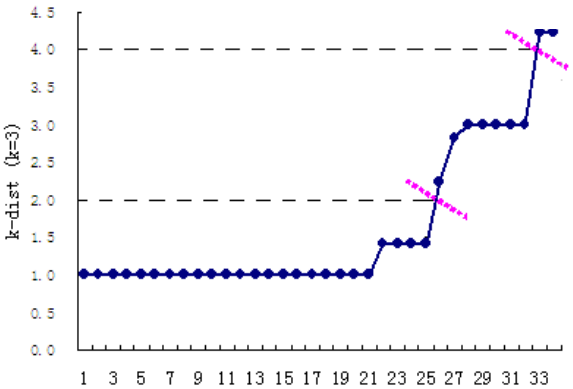

Figure 5. K-dist plot (k=3)

According to k-dist plot (k=3) shown in Figure 5, there are two sharp changes at the value of k-dist that correspond to suitable values of *Eps*. So select Eps_1=2, Eps_2 =4, and take the value of k as the *MinPts* parameter, that is, *MinPts*=3.

Adopt DBSCAN algorithm twice for two different values of *Eps*. In the first round, Eps_1 =2, points marked ' - ' in Figure 4 are identified as core points, and are clustered as C_1. At the same time, those marked ' o ' or ' . ' are identified as noise points. Note that, all the clustered points should be given cluster marks for the further DBSCAN operation. In the second round, Eps_2 =4, and only non-cluster-marked points are processed. Points marked ' o ' are identified as cores and those marked ' . ' as border points. Both types of points are clustered as C_2 and are given cluster marks. The density of cluster C_1 is 1 per unit area while that of C_2 0.25 per unit area. Non-cluster-marked points are recognized as noise points or anomaly.

4.3 Experiment Conclusion

A synthetic database with 2-dimension data is used for demonstration. The experiment shows that VDBSCAN is good at finding out clusters corresponding with varied densities, and it has the same time complexity as DBSCAN.

VDBSCAN algorithm successfully overcomes one of the main problems of traditional DBSCAN which limits itself to varied-density datasets. For the experimental data, traditional DBSCAN algorithm selects only a value of parameter *Eps*. Supposing using DBSCAN algorithm in this experiment, as the process of selection of the value of *Eps* using k-dist plot is the same as that in VDBSCAN, which has been described before, *Eps* may be either 2 or 4. If *Eps* is 2, cluster C_1 will be found out correctly while all points of C_2 will be identified as noise points or anomaly. If the value of *Eps* is 4, points of both C_1 and C_2 are clustered as a cluster. Thus, traditional DBSCAN algorithm can not distinguish clusters with respect to different densities, or can not find out them simultaneously. The experimental results show that varied-density datasets can be clustered successfully by VDBSCAN.

5. CONCLUSIONS

In this paper, we proposed a new density-based algorithm called VDBSCAN in purpose of finding out meaningful clusters in databases in respect with widely varied densities. VDBSCAN has the same time complexity as DBSCAN, and can identify clusters with different densities while DBSCAN can not. The experiment shows the efficiency of the new algorithm. However, there are several opportunities for future research. How to select all the parameters automatically is one of the interesting challenges as parameter k has to be chosen subjectively in VDBSCAN.

REFERENCES

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

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining,* Pearson Education Asia LTD, 2006.

[3] Jain A. K., Dubes R. C., *Algorithms for clustering Data*, Prentice-Hall, Inc., 1988.

[4] Ester M., Kriegel H.-P., Sander J., Xu X., "A density-based algorithm for discovering clusters in large spatial databases with noise", *proceeding of 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Pres*s, pp. 226-231, 1996.

[5] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, "OPTICS: Ordering points to identify the clustering structure", *proceeding of 1999 ACM-SIGMOD International Conference, ACM Press*, pp.49-60, 1999.

[6] Sun Xue-gang, Chen Qun-xiu, and Ma Liang, "study on topic-based web clustering", *The Journal of Chinese Information Processing*, Vol 17, No. 3, pp.21-26, 2003.

[7] Hinneburg A., Keim D., "An efficient approach to clustering in large multimedia databases with noise", *Proceeding of 4th International Conference on Knowledge Discovery and Data Mining*, New York City, NY, 1998.

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

[9] Agrawal R., Gehrke J., Gunopulos D., Raghavan P., "Automatic subspace clustering of high dimensional data for data mining applications", Proceeding ACM SIGMOD '98 International Conference on Management of Data, pp. 94-105, Seattle, WA, 1998.