Implementation of Adaptive DBSCAN for Cluster Analysis

Priya Dharshini D A
Assistant Professor
Department of Information Technology
KCG College of Technology, Chennai, India

Sudha T
Assistant Professor
Department of Information Technology
KCG College of Technology, Chennai, India

Usha G
Assistant Professor
Department of Information Technology
KCG College of Technology, Chennai, India

Adline Freeda R
Assistant Professor
Department of Information Technology
KCG College of Technology, Chennai, India

Abstract

The ability to monitor students’ academic performance as well as their participation in other events and managing a record for the same is a critical issue to the staff members. Encouraging the students to improve in the area were they are lacking behind is necessary. In this paper we propose a system where the staff members can update the student records in areas like academics, technical skills and participation in extracurricular activities and viewing the same whenever necessary. Further we implemented the adaptive DBSCAN for clustering the students based on their performance in various areas. Based on the resulting clusters the staffs identify the performance of each student and the students’ progress can also be monitored.

Keywords: Data mining, Clustering, Adaptive DBSCAN

I. INTRODUCTION

Data’s are generated very faster and larger every day. With rapid growth of data, the people are really interested only in the relation between data or the information data gives. Data mining is basically a concept of extracting the information needed from large blocks of data. Clustering is one of the primary tasks in the data mining field which groups data into several classes. It is the process of dividing objects into different groups such that the objects in the same group are similar to each other according to the same characteristics while objects in different groups are dissimilar. Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together. However, there are two key parameters for DBSCAN algorithm: eps and Minpts. Though users do not need to know the number of clusters, these two parameters are user-determined. Eps play an important role in DSCAN and the result may vary significantly according to different eps value. Sometimes it is hard to decide an appropriate eps because dataset are different from each other. Also DBSCAN has difficulties in finding clusters for dataset with different densities, because the eps value is a global parameter [4]. VDBSCAN [5] was proposed to deal with the different density situation. Different density should have different eps and corresponding Minpts value. Determine the eps value automatically according to the distribution of data under the multi-density situation is a desired property.

In this paper, we introduce a modified DBSCAN for determining an appropriate eps value automatically. The modified DBSCAN which is known as Adaptive DBSCAN select eps according to k-distance graph [6][7].If the dataset demonstrates different density, the Adaptive DBSCAN will cluster with different eps value and corresponding different Minpts values. And for dataset with single density level, we also use different eps and Minpts value to cluster the students based on their performance in various skill sets irrespective of the density of the dataset. It generates automatic cluster based on the density of the student details.

II. ADAPTIVE DBSCAN ALGORITHM

There are roughly five categories of clustering techniques: partition based (such as K-Mean, K-Medoids), hierarchical based (DBSCAN, OPTICS, DENCLUE), grid based (STING), and model based [2]. DBSCAN (Density Based Spatial Clustering of Application with Noise) is a widely used density based clustering methods. DBSCAN is based on the concept of density reachability [2] and it can find clusters of arbitrary shape and size and detect noise. Some clustering methods requires user to provide a priori information about the dataset. For example, K-Means is a widely used simple yet effective partition based clustering methods. But K-Mean require that user need to know the value of k which refer to the number of total cluster in advance, and it can only detect spherical cluster [3]. Compare to K-means, the advantage of DBSCAN is that DBSCAN does not require prior information. User do not need to know the number of clusters before applying DBSCAN and it can find cluster of any shape as well as distinguish noise.
However, there are two key parameters for DBSCAN algorithm: eps and Minpts. Though users do not need to know the number of clusters, these two parameters are user-determined. Eps play an important role in DBSCAN, and the result may vary significantly according to different eps value. Sometimes it is hard to decide an appropriate eps because dataset are different from each other. Also DBSCAN has difficulties in finding clusters for dataset with different densities, because the eps value is a global parameter [4]. VDBSCAN [5] was proposed to deal with the different density situation. Different density should have different eps and corresponding Minpts value. Determine the eps value automatically according to the distribution off data under the multi-density situation is a desired property.

In this Algorithm we introduce a modified DBSCAN for determining an appropriate eps value automatically. The modified DBSCAN which is known as Adaptive DBSCAN select eps according to k-distance graph [6][7]. If the dataset demonstrates different density, the Adaptive DBSCAN will cluster with different eps value and corresponding different Minpts values. And for dataset with single density level, we also use different eps and Minpts value to cluster the students based on their performance is various skill sets irrespective of the density of the dataset. It generate automatic cluster based on the density of the student details.

**A. Determining the eps value**

This section explains Adaptive DBSCAN. The first step is to draw a k-dist graph [6]. The distance function employed in DBSCAN is the Euclidian distance where data object x and y are two n-dimension data points:

$$D(x, y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + ... + (x_n-y_n)^2}$$ (1)

The k-dist is a function that calculates the distance between a point and its kth nearest point. After getting the k-dist value of every point in the dataset and sorting it in descending order, we get a sorted k-dist graph. Figure 1 show a sample 4-dist graph. A k-dist graph gives some clues about the density distribution of the dataset. A k-dist graph shows the k-dist value of every point in descending order. The k-dist returns the distance from the point to its kth-nearest points, if there is a suddenly increased distance in the k-dist graph, that mean the kth-nearest neighbor of a point become farther, and the density changes. If the points whose k-dist is higher than the threshold, they are regarded as noise and such threshold is the point in the first valley [8]. In the Figure 1, the arrow points out the valley. However, the valley sometimes is difficult to detect automatically. Though, the user can see the valley in a graphical representation, the value may not be accurate.

![Fig. 1: A sample 4-dist graph.](image)

if the density level start to vary, there will be some sharp variations in the k-dist graph; however, for the points belonging to the same density level, the variation may not be large [6]. Figure 2 is a k-dist graph sorted in ascending order as shown in figure 1. Because the original DBSCAN runs with single eps and Minpts, we only need to find the first valley in the k-dist graph. Hence, the k-dist graph in descending order for our ease to find several eps (i.e. the valley), so we sort the k-dist in ascending order. Similar density level does not cause larger variation in k-dist values, thus a sharp change between two smooth curves indicates density level varies [6]. We can find several smooth curves connected by rising curves connected by a sharply rising line b, indicating a and c are two density levels; line c and line e are similar case. Three smooth curves refers to the three density level and line f shows the k-dist of the noise since it does not connect the two smooth lines [6].
In k-dist graph, it only consider the distance of a point from its kth nearest point, instead we calculate the average of the distance from a point to its entire k-nearest neighbor (KNN). Considering the entire KNN neighbor of a point and the average leads to a smooth curve with the removal of noise and it is easier to detect the threshold of the density level [6][7]. To find the eps value for each density level, we need to calculate the slopes and to check if there exists a sharp difference between slopes. If we detect the slopes of two lines is the threshold for the certain density level. In [6], the slopes and the difference are calculated at a fixed interval. In some case, if the slope in one interval almost remain horizontal (i.e., the slope is too small and close to 0), even the slope in the next interval increases just slightly, it will detect a high changing rate. Hence, calculating the slope in small interval will find many small sudden changing slopes. Such slopes are detected because the previous interval is close to a horizontal line. Instead of using the fixed interval for calculating the slopes, we propose to calculate the slope between every point rather than an interval. Because there might be some points rising in the interval in case and using the fixed interval for calculating the slope may ignore such rising points. When a large change is detected between two slopes, it is necessary to look the slope afterwards to see whether the slope really start to rise.

When detecting a large change between two neighboring slopes and making sure the slope is really increasing, the point connected to the two lines is the threshold for the density level and we select it as the eps value. Finding all these kind of points in the k-dist graph, we find the value of eps for each density levels. The Adaptive DBSCAN starts by clustering from the lowest eps to the highest eps. The lower eps value we get in k-dist graph indicates the denser region because the average distance between a point and all of its k-neatest neighbors is small. First we cluster with lowest eps, the points belonging to that densest density level will be clustered together and the lists of noise are those points belonging to sparser region. Next, we apply DBSCAN on the noise list with the next eps and the points that have been clustered will be ignored in the following process. The eps value is the threshold to certain density level, so applying DBSCAN on the noise list obtained before with remaining eps can cluster points that belong to similar density level.

B. Evaluating Minpts

After determining the eps values, the next step is to calculate the minpts value. To determine the minpts values, [6] calculate the number of data points in the eps neighborhood of every point in the dataset.

\[ \text{Minpts} = \frac{1}{n} \sum_{1}^{n} P_i \]  

(2)

Where \( P_i \) is the number of data point in the eps-neighborhood of point \( i \) and \( n \) denotes the total number of data points in the dataset. For each eps determined from the k-dist graph, we can calculate the corresponding Minpts according to Equation (2). Because the k-dist is sorted in ascending order, we get the eps value from smaller value to larger value, which cause the Minpts value to increase. We cluster with smaller eps and then cluster remaining data points with larger eps value, so we will get Minpts with lower eps first. However, the larger eps indicated sparser density level, when we get the Minpts for larger eps values, the Equation 2 above will also consider the points in the density region since it calculate the Minpts with all data points. With larger eps, the points in dense region will collect more points, causing the Minpts to become larger. But the remaining points in the noise list after applying DBSCAN are the points with sparser density level when compared to the points that are just clustered in the last process. Thus the situation that sparser region have larger Minpts will happen. Since Minpts refer to the number of points in the eps-neighborhood of a point, it is not helpful to cluster sparser region with larger Minpts values. Hence, we proposed to modify the calculation of Mipts value as follows.

For each eps, the corresponding Minpts is:

\[ \text{Minpts}_j = \frac{1}{n} \sum_{1}^{n} P_i, \ \text{eps}_{j-1} \leq k\text{dist}(i) \leq \text{eps}_j \]  

(3)

Where \( P_i \) is the number in epsj neighborhood of point I and k-dist(i) is the average of distance between point I and all its k-nearest neighbors.
For Minptsj corresponding to epsj, we calculate the number of data points in epsj –neighborhood of every point such that point I satisfying

\[ \text{eps}_j,1 \leq k\text{dist}(i) \leq \text{eps}_j \]

Thus, the Minpts value is local to corresponding eps value and it considers the points with same density level, ignoring the points in denser region.

### III. Proposed Work

The Adaptive DBSCAN system consists of a mining server, Mining Students’ database and dividing them into groups based on their performance. Clustering Algorithm: Groups the students based on their overall performance.

- **Technology used:** Html5, CSS3, php, Ms excel
- **Attributes to be measured and analyzed:**
  - Educational Performance (CGPA)
  - Participation in technical events (Paper Presentations, coding)
  - Awards won in extracurricular activities (music, dance, sports)
  - Students grouped together based on their performance.
  - Test for Data in the Database (CGPA, No. of events participated (Technical and non Technical))
  - Clear understanding of every student’s performance.
  - Students can be guided accordingly.

### IV. Implementation and Result

The parameter of Adaptive DBSCAN eps of each dataset is obtained from the corresponding k-dist graph of the dataset. The eps is the threshold of each density level, which is the sharply changing place in the k-dist graph. We can discover the changing points from the k-dist graph. The eps values of other datasets are obtained in the same way. The Adaptive DBSCAN for cluster analysis is implemented to analyze individual students’ performance and overall class performance by the department staffs and to monitor their progress. The administrator is the one who stores all students’ details which includes student CGPA, number of awards won in technical events (coding, paper presentation ) and other extracurricular prizes won (music, dance, sports) in database provided a login id and password. The department staffs can view each student’s performance by logging into the staff account, update student’s record and monitor their progress. Here we use Adaptive DBSCAN to form clusters based on the student’s details (CGPA, No. of awards won in Technical events and other extracurricular awards won). We provide credits for each CGPA and other awards and we use this credits to form clusters. Thus the clustering is done based on the credits obtained and the final result is generated in a graph form. The graph is thus obtained for Individual student’s performance graph (Figure 3) and overall performance graph (Figure 4).

![Fig. 3: Individual performance analysis graph](image-url)
Implementation of Adaptive DBSCAN for Cluster Analysis

V. CONCLUSION AND DISCUSSIONS

DBSCAN is a widely used density-based clustering method. However, it is hard to determine the parameters eps and Minpts for DBSCAN. Also DBSCAN has difficulties in finding clusters with different densities. We use modified DBSCAN (Adaptive DBSCAN) so that it can detect clusters of different density levels. We decide the eps values according to data distributions without priori information. For each eps value, we modify and calculate the corresponding Minpts values. With different eps and Minpts, the modified DBSCAN can find clusters with different density levels. After applying DBSCAN, we cluster remaining noise with additional eps and Minpts value to try to find other clusters. The value of k for k-dist graph can affect the results significantly, and finding an optimal k is necessary. Hence, we propose an inference of finding an optimal k. The future enhancement is to make the system more accurate by introducing new concepts and techniques in Adaptive DBSCAN system.

REFERENCES