SUBSPACE CLUSTERING ANALYSIS USING DBSCAN AND SUBCLU FOR PROJECTING STMIK JAKARTA ALUMNI JOB

Sanjaya Alamsyah  
*Master of information Systems Management*  
gunadarma University  
Depok, Indonesia  
*E-mail:* salamsyah2@gmail.com

Novita Sulistyowati  
doctoral information technology  
gunadarma University  
Depok, Indonesia

Abstract  
Subspace clustering is projected as a search technique to classify the data or attributes in different clusters, grouping is done by determining the density of data and also identify outliers or data that is not relevant, so that each cluster exists in its own subset. This thesis proposes innovation subspace clustering algorithm based on density connection. In the early stages will be counted dimensional density, density results will be used as input data dimensions to determine the initial clusters based on density-dimensional, ie, by using the DBSCAN algorithm. Data on each cluster will then be tested whether having a relationship with the other data on the cluster, by using Algorithm SUBCLU.

Results of this study found bahawa SUBCLU not have an unclustered real dataset, so the perception of the results of the cluster will produce more accurate information for job satisfaction dataset whereas DBSCAN takes more time than the method SUBCLU. For larger and more complex the data, the performance SUBCLU looks more efficient than DBSCAN.

**Keywords:** Subspace clustering; DBSCAN; SUBCLU

I. Introduction

Technological developments, especially in information and communications technology, fuel the needs of the educated workers in the field of computer science and information technology. As a consequence, the growth of higher education institutions based on computer science and information technology (Diploma and Bachelor) also experienced a very rapid growth. This situation triggered intense competition among college alumni to seize the job market. The intense competition for college alumni to reach the job market is one of the issues that are often the topics of discussion among university leaders. So raises the question how the design of teaching and learning programs for students to stock the knowledge to students so as to produce alumni with competence in accordance with the available job market.

Subspace clustering is projected as a search technique for categorizing data or attributes on different clusters, grouping is done by determining the data density level and also identifying outliers or irrelevant data. so that each cluster is in its
own subset. This thesis proposes an innovation of subspace clustering algorithm based on density connection. In the early stages will be calculated dimensional density, dimension density results will be used as input data to determine the initial cluster based on dimensional density, i.e., by using DBSCAN Algorithm.

II. Data Mining and Clustering

A. Data Mining

Data mining is the process of extracting the data from large databases, used as technology to generate the required information. Data mining methods can be used to predict future data trends, estimate its scope, and can be used as a reliable basis in the decision-making process. Functions of data mining are association, correlation, prediction, clustering, classification, analysis, trends, outliers, and deviation analysis, and similarity and dissimilarity analysis.

B. Clustering

Clustering is a method for grouping similar objects into Class - Class. A cluster is a set of data that resembles one another and is not similar to other clustered data. Clustering is different from classification due to clustering no pre-set target classes. Clustering algorithms will attempt to divide the existing data into data groups where data in the same cluster is relatively more homogeneous when compared to the data in other groups. Clustering seeks to maximize the similarity of the data on the same cluster and minimize its similarity to the data in other clusters (Larose, 2005). The main inherent idea of this paper is to compare the percent of time needed to run the whole DBSCAN algorithm with respect to insertion of new coming data objects and determines which approach takes fewer amounts of time and effort. Due to the density-based nature of DBSCAN, the insertion or deletion of an object affects the current clustering only in the neighborhood of this object.

III. Results and Discussion

There is no model of job projection for college alumni. So it takes a model that can project work for alumni of college. How to obtain cluster subspace model from multidimensional data based on data density. What method is best used for multidimensional data subspace clustering process? In selecting the sample, this study used a convenient technique of convenient sampling and random sampling, the researchers had no other consideration for using convenient sampling except on the basis of convenience only. Someone was taken as a sample because by chance the person was in place or coincidentally the researcher knew the person. This type of sample is excellent if used for exploratory research, followed by a follow-up study whose samples are drawn at random. However, some cases of research using this type of sample, indicating the result turned out to be less objective (Mustafa, 2000). For that researchers also used a random sampling technique conducted in an online questionnaire through the website on the internet. For the purposes of data analysis, a minimum of 1000 alumni is required with the consideration of the amount of data required for the analysis process using DBSCAN and SUBCLU is at least 100 (Ferdinand, 2002) or other researchers suggest at least 5 times the number of indicators or research instruments (Sekaran, 2003). This research uses 6 research variables consisting of 3 exogenous variables (Glass, Liver-dis, Job satisfaction), and 3 moderator variable (age, age, year). Then from the variables have each have the indicators that become the
measurement instrument All indicators have been validated its compatibility with the constructs it represents.

A. Data Implementation

The data is taken from alumni data of STMIK Jakarta, STI & K Jakarta, where the data of alumni taken limit only 10 years from 2007-2016. And data obtained about 2400 alumni, where the purpose of the thesis is to project alumni who absorbed the work in the company according to the field. in managing data up to grouping based on density connections, how cluster job opportunities are based on subspace clusters, and to the extent of its analysis the author uses the help of WEKA applications.

B. Subspace Clustering Analysis

- The result of grouping in one strategy so that above will be processed by mining association rules, to find associative rules between certain combinations. A particular type of reasoning that uses the "if-then-else" rule statement will be implemented, just as a pattern and search engine inference for patterns in the corresponding pattern rules in the data. This research will apply subspace clusters, as learning operators of both nominal and numerical data

- An alternative approach to determining clustering that best fits the set of data x is to consider all possible clustering and select the one that makes the most sense according to the criteria and rationality.

<table>
<thead>
<tr>
<th>dimensions</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
<th>( d_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>7.4</td>
<td>4.6</td>
<td>5.7</td>
<td>5.5</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>6.4</td>
<td>3.7</td>
<td>8.2</td>
<td>8.0</td>
</tr>
<tr>
<td>( s_3 )</td>
<td>8.1</td>
<td>3.9</td>
<td>8.6</td>
<td>5.8</td>
</tr>
<tr>
<td>( s_4 )</td>
<td>5.2</td>
<td>6.0</td>
<td>5.5</td>
<td>5.8</td>
</tr>
<tr>
<td>( s_5 )</td>
<td>8.0</td>
<td>3.2</td>
<td>7.8</td>
<td>8.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dimension</th>
<th>clustered stream values</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>{5.2, 6.4}, {6.4, 7.4, 8.0, 6.1}</td>
</tr>
<tr>
<td>( d_2 )</td>
<td>{3.2, 3.7, 3.9, 4.6}, {4.6, 6.0}</td>
</tr>
<tr>
<td>( d_3 )</td>
<td>{5.5, 6.7}, {6.7, 7.8, 8.2, 6.6}</td>
</tr>
<tr>
<td>( d_4 )</td>
<td>{5.5, 5.8, 6.8}, {8.0, 8.3}</td>
</tr>
</tbody>
</table>

Figure 1 Cluster Initialization

- Some clustering algorithms have been developed, some of which are single clustering, and the other is clustering hierarchy. The following classification contains most of the well known clustering algorithms

- Single Clustering algorithm : Sequential algorithms, with a simple concept, work on a single data set or very little data; Cost function optimization algorithms, which adopt a J-cost function by quantizing a sensible term and generating clustering with optimization J. Included in this category are hard clustering algorithms such as k-means, fuzzy clustering algorithms such as
fuzzy c-means (FCM), probabilistic clustering algorithms such as EM and probabilistic algorithms.

C. DBSCAN

DBSCAN is one of the density-based clustering algorithms. The algorithm extends the region with high density into the cluster and places the irregular cluster on the spatial database with noise. This method defines the cluster as the maximal set of density-connected dots. DBSCAN has two parameters: Eps (maximum radius of the neighborhood) and MinPts (minimum number of points in the Eps-neighborhood of a point). The basic idea of density-based clustering relates to several new definitions:

Neighborhood with an Eps radius of an object is called Epsneighborhood of an object. If the Eps-neighborhood of an object contains a point at least the minimum number, MinPts, then an object is called a core object. Given a set of D objects, the p object is said to be directly density-reachable from an object q if p is included in the Eps-neighborhood of q and q is the core object.

![Figure 2 Density Reachable](image.jpg)

Sometimes DBSCAN may be applied on dynamic database which is frequently updated by insertion or deletion of data. After insertions and deletions to the database, the clustering discovered by DBSCAN has to be updated. Incremental clustering could improve the chances of finding the global optimum. In this approach, first we form clusters based on the initial objects and a given radius(eps) and minimum number of points(Min pts).

D. Subclu

- Subspace Clustering is an effective and efficient method in subspace clustering problems.
- Subclu is an effective and efficient method in subspace clustering problems with grid-based approach, this method can detect the shape and position of clusters. Monotonous nature of the relationship density, bottom up approach used to pruning subspace and produces clusters that are connected with density.
IV. PERFORMANCE EVALUATION

We have tested SUBCLU, FIRES and INSCY using data sets, with processor 1.66 GHz and 1 GB of RAM.

A. Esearch Result

To verify the quality of clustering obtained through our technique (Damira) and to speed up the first stage, we run DBSCAN, FIRES, INSCY, SUBCLU, and Damira. The Setup parameter is performed on subspace cluster bracketing, and the mean dimensions and number of clusters are defined. Table 1 shows the property of the dataset, we are implemented in 3 real datasets, and 6 higher education datasets. For each test study using MinPoints = 6 and Epsilon = 0.9, based on previous experiments these criteria make the best clusters.

Table 2 Experiment Result Performance
B. Data Sets

The data is taken from alumni data of STMIK Jakarta, STI & K Jakarta, where the data of alumni taken batasin only 10 years from 2007-2016. And data obtained about 2400 alumni, where the purpose of the thesis is to project alumni who absorbed the work in the company according to the field. in mengola data up to grouping based on density connections, how the cluster work opportunities based on subspace cluster, and until ketahap analysis writer use the help of WEKA application.

An important aspect of the proposed method is the number of clusters that are generated. Based on the test results, DBSCAN cluster generated only 4 clusters obtained on the dataset of alumni that work in accordance with the field as well, sum clusters identified very low, ranging from 5 clusters. Meanwhile Damira Damira managed to build a very large number of clusters for each dataset.
Using the Open subspace produces a visualization cluster, as shown in Figure 6, seen the number of clusters generated is always less than other subspace methods. Similarly, the number of dataset groups of higher education institutions forms another method of subspace clusters such as FIRES and INSCY, tendency to fail to form groups in each of the subspace spaces as shown in Figure 6, while Damira produces many clusters outside of other method.

By using Opensubspace, it produces a cluster of visualizations, like the ones shown in Figure 4.13, the number of clusters generated is always less from other subspace methods. Similarly, the number of groups of institutional datasets higher education, forming other subspace cluster methods, such as FIRES and INSCY, the tendency to fail to form groups in every space section, as shown in Figure 6, while Damira, produces many clusters, outside of other methods.
C. PERFORMANCE EVALUATION

Performance evaluation of data mining becomes very important, predicting the correct number Unattended cluster learning process is an obstacle, however it can be cleaned using efficiency, accuracy, cluster coverage and index F1- Entropy to assess the quality of the cluster.

From the experimental results, we can see that the time for grouping glass datasets and liver datasets, while for job satisfaction Damira dataset takes short time than the SUBCLU and INSCY methods as shown in Figure 9.

For larger and more complex data, Damira's performance looks more efficient than SUBCLU as shown in Figure 9. Still, still lower than FIRES and INSCY, especially compared to the method DBSCAN, which can be done very quickly, averaging less than 1 second, but tend to a lot of un-clustered data.

In grouping experiments for glass datasets, the FIRES method has accuracy which is better than INSCY, and SUBCLU. While Damira's method is more accurate than INSCY and SUBCLU methods, as shown in Fig 9. But for the experiment of dataset grouping Damira's accuracy is lower compared to FIRES, INSCY and SUBCLU methods.
D. CONCLUSION

The use of clustering algorithms is to measure the similarity of high-dimensional data or attributes often achieve the desired result, causing attributes to be unrelated or too close together. Closed data can form overlapping groups to form solid groups, data can be found in different groups and also in different parts of the space. Subspace clustering is projected as a technique for finding data or grouping attributes in different clusters. Grouping is done by determining the data density level and also done to identify outliers or irrelevant data that will make each cluster to exist in a separate subset.

This study proves innovative clustering subspace clustering algorithm (SUBCLU) based on density connections. The research presence estimates the density dimension and the results are used as input data.

To determine the initial cluster based on density connections using the DBSCAN algorithm. Each dimension is tested to investigate whether it has a relationship with data on another cluster by using the proposed clustering subspace algorithm. If the data has a relationship, it will be classified as a subspace. This study uses multidimensional data, such as benchmark datasets and real datasets. Real data sets from education, particularly on the perceptions of the student industry training and from industry.

To verify the clustering quality obtained through the author's technique and to speed up the first stage, we run DBSCAN, INSCY, SUBCLU. tendency to fail to form groups in every part of the space. SUBCLU and Damira do not have un-clustered real datasets, so the perception of cluster results will yield more accurate information, whereas for job satisfaction DBSCAN dataset takes longer than SUBCLU method. For larger and more complex data, the performance of SUBCLU looks more efficient than DBSCAN.

References

[4] Berka, Petr, Jan Rauch, Djamel Abdelkader Zighed, Data Mining And Medical Knowledge Management, Medical Information science reference, New York, 2009


