Application of K -Means Clustering for Students Graduate Group Analysis

SeftyWijayanti, Azahari
The Computer Engineering Department
STMIK Widya Cipta Dharma
Samarinda, Indonesia
seftywicid@gmail.com

***Abstract*—Grouping analysis (cluster analysis) is one of the multivariate data analysis which attracted many people and growing very rapidly . Grouping method that is being developed at this time is the fuzzy clustering analysis , which is capable of grouping by using a certain degree of membership . Fuzzy clustering algorithm that is commonly used is the fuzzy C -Means (FCM) and K -Means , which is capable of detecting a group with a different shape than the FCM . This study examines the comparative application of methods FCM and K -Means in a case study , namely graduate student grouping STMIK WidyaCipta Dharma based on the characteristics of the GPA , the Old Study , Study Programs and Predicate . Determination of the number of groups is done through a validity index . This study also will make FCM algorithm and K -Means with MATLAB software . The results showed that in some ways still superior to FCM K - Means . The number of most Optimal Group is a total of 4 groups.**

Keywords— Analysis of the group name; k –means; graduate student grouping , FCM

I. Introduction

Analysis grouping or Cluster analysis is one of the data analysis aimed at determining group or a group of a group of data based on common characteristics. The object of bias in the form of goods and services, objects, people or region. This analysis has been widely used to solve problems and research in several disciplines, such as academic, medical fields, including the field of territorial marketing field. Data clustering paradigm attracted many circles and written in various papers and journals Shihab, Maxwell, Pryor and Smith never applied analysis in the research field of social group to group social status in society based on equality and cross-cultural differences in the data.

The development of group analysis starts from the method hierarchy that outlines form a tree diagram is commonly called the dendrogram based on the distance to describe the group. Determination of the number of groups is to make the cut off of the dendrogram. While the method nonhierarkhi better known by means of a partition, for example, k-means. This method of determining in advance the number of groups that will be formed to fit the purpose of research.

With the pile of data in a college that is not used, then used a pile of data to search for new information. Piles of data used is the academic data in Engineering Program, Information Systems at STMIK WidyaCipta Dharma Samarinda, the IP (GPA) data students from one semester to the end by using one of the techniques of data mining is clustering method using K-Means.

Utilization of existing data in information systems to support decision-making activities, not enough to just rely on operational data alone, required the analysis of data to explore the potential of existing information.

II. Basis of Theories

A. Data Mining

Data mining is the process of discovering interesting patterns from massive amounts of data. As a knowledge discovery process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation [1].

Data mining is the science, art and technology of exploring large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective and accurate [2].

B. Data Warehouse

One emerging data repository architecture is the data warehouse. This is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)—that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis—for example, data mining tools that provide data classification, clustering, outlier/anomaly detection, and the characterization of changes in data over time [3].

C. Clustering

Clustering is the method by which a data set is divided into a number of smaller, more similar subgroups or clusters. The goal in cluster detection is to find previously unknown similarities in the data. Clustering data is a very good way to start analysis on the data because it can provide the starting point for discovering relationships among subgroups [4][5].

D. Cluster K-Means

The K-means method assigns cluster membership by distance. As shown previously in Figure 1, an object belongs to the cluster whose center it is closest to (which is measured using a simple Euclidean distance). When all objects have been assigned to clusters, the center of the cluster is moved to the mean of all assigned objects, thus the name K-means — *K* being the typical denomination for the number of clusters to look for [1].

K-means clustering algorithm is applied to the object that is represented in the form of points in d-dimensional vector space. K-means cluster perform all data within each dimension which point in the same segmentation given cluster ID. The value of k is the basic input of the algorithm that determines the number of segments to be formed. Partitions will be formed from a set of n objects into k clusters so formed in common objects in each segmentation k [6].

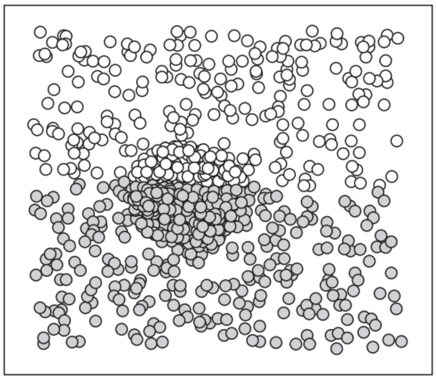

Fig. 1. Clustering of a dense region using K-means

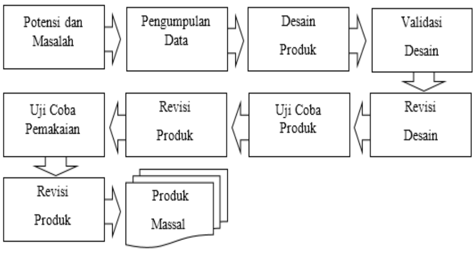

Fig. 2. Research and Development

III. Research Method

Methods of research and development (Research and Development) is a research method used to produce a particular product, and test the effectiveness of the product, measures the R & D can be seen in Figure 2.

In general the steps of research and development include:

A. Potential and Problems

Identify the problem at this stage the problem identified is something that when usefully be put would have added value. Issues that must be obtained in accordance with the data that can be based on empirical research reports of others, or documentation of other people's research reports, activity reports or documentation of new individuals or institutions.

B. Data collection

Collect information after the problem can be shown to be factual and up to date, then collected a variety of information that is used as the planning of specific products are expected to tackle the issue and requires its own research methods.

C. Product Design

The design of the products produced in research R & D vary. Product design can be realized in the form of a picture or a chart, so that it can be used as a handle to assess and make it. Or in the form of system accompanied with an explanation of the mechanism of the use of the system, how to work, as well as the advantages and disadvantages

D. Design Validation

Design validation process is an activity to assess whether the rational design of a product will be more effective than the old ones or not. Is said to be rational because it is still in the process of this assessment can be done by bringing in seasoned experts to assess the product so that further known weaknesses and strength in a discussion forum. For example, the progression of research models and expert team learning device in question is the instructional technology specialist, an expert in the field of studies on the same subject and expert evaluation of learning.

E. Revised Design

Design improvement after the known weaknesses of the product is carried out experiments to design improvements. Which served to improve the design are researchers who want to produce the products.

F. Test Product

Test products early stages of product trials conducted by simulating the use of such products. After simulated, it can be tested on a limited group. Testing was conducted to obtain information on whether the product is more effective than the old products.

G. *Revision Product*

Product revisions performed because of the tests conducted are limited, so it does not reflect the actual circumstances, the trial found the weaknesses and shortcomings of the products developed, and data for the product can be captured through the product.

H. *Trials*

Test after a revised use of the product will be applied to a wider group. In this trial the product remains to be assessed deficiencies and bottlenecks which appear for further improvement.

I. *Products Revised*

TahapThe final stage when the product usage in a wider group there is a shortage, then the product makers have to reevaluate how the performance of the product. From the results of the evaluation of the product can be used for the improvement and creation of new products again.

J. *MassProduct*

Eg product manufacturing this stage is the final stage of research R & D. If the product has been declared effective in some time testing the product can be applied to the group eg by making mass products.

Validity and rehabilitation in this study can be obtained melali expert judgment with techniques as done in the research survey, quasi-experimental, action research.

This research is experimentation, and in its implementation requires several instruments, namely data taken from multiple tables then formed one of the table so that meet the necessary attributes. Attribute data required in the study are as Table 1.

Data retrieved from the database in STMIK WidyaCipta Dharma with the year 2000-2014. In this study there weight method and attributes needed to determine who will be selected as the best graduate.

TABLE I. ATTRIBUTE DATA NOT PROCESSED

NIM	**IPK** **(XI)**	**PROGRAM STUDI** **(X2)**	**LAMA STUDI** **(X3)**	**PREDIKAT** **(X4)**
07.43.193	3,21	TeknikInformatika	4 Tahun	SangatMemuaskan
05.43.114	2,61	TeknikInformatika	6 Tahun	Memuaskan
04.43.105	2,80	TeknikInformatika	7 Tahun	SangatMemuaskan
07.43.906	3,39	TeknikInformatika	4 Tahun	SangatMemuaskan
06.43.086	2,81	TeknikInforamatika	5 Tahun	SangatMemuaskan
07.41.001	3,51	SistemInformasi	4 Tahun	DenganPujian
06.41.072	3,24	SistemInformasi	5 Tahun	SangatMemuaskan
07.41.911	3,68	SistemInformasi	4 Tahun	DenganPujian
05.41.009	3,11	SistemInformasi	6 Tahun	SangatMemuaskan
07.41.041	2,82	SistemInformasi	Tahun	SangatMemuaskan

TABLE II. STUDENT PASS THE SAMPLE DATA WICH WILL BE SEGMENTED BY K-MEANS

No	**X1**	**X2**	**X3**	**X4**
1	2.93	41	4	90
2	2.42	41	4	80
3	2.95	41	6	90
4	2.88	41	6	90
5	3.02	41	4	90
6	3.01	41	5	90
7	2.55	41	5	80
8	2.82	41	5	90
9	2.59	41	4	80
10	3.57	41	4	100
17	2.73	43	4	80
18	2.68	43	5	80
19	3.32	43	5	90
20	3.17	43	5	90
..	…			
455	2.55	43	4	80

IV. RESULT AND ANALYSIS

A. *Output Performace Based on the value of Confudion Matrix*

From the beginning of 1171 a number of data records after experiencing the initial (pre-processing) of data into a 1141 record. Test data is divided into 2-year graduate student STMIK WidyaCipta Dharma who graduated in 2005-2009 were 455 graduates Tier One Program Information Systems and Information Engineering that have the attributes X1: GPA, X2: Program, X3: Old Study, X4: Predicate Graduates.

From table 2 can be classified into several groups according to the attributes that have been specified in the form, GPA, Major, Old Studies, and Graduate Predicate.

Tests using the K-Means clustering is done with the help of tools from MATLAB applications. Data previously stored into csv format and inserted into the k-means clustering tools in MATLAB and inserted into the function testing of K-Means clustering by incorporating some initial parameters before the clustering is done. The number of clusters, according to the number of existing concentration in the study program Informatics and Information Systems. The maximum number of iterations is setup as much as 100 times the smallest error expected at 0.00001, with a value of powers is 2 initial objective function is 0 (zero) and the initial iteration is 1 (One).

B. *End Display Graphics*

Evalclusters charts used for referrals from some of the existing cluster. The first chart evalclusters highest peak which will be used to determine the best cluster of a few existing clusters, cluster evalcusters the best according to the above data is in cluster 4 can be seen in Figure 3.

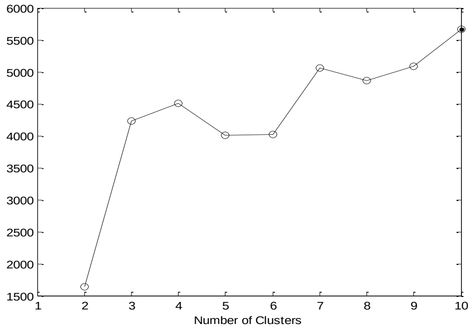

Fig. 3. Objective Function Value graph (Graph Evalclaster)

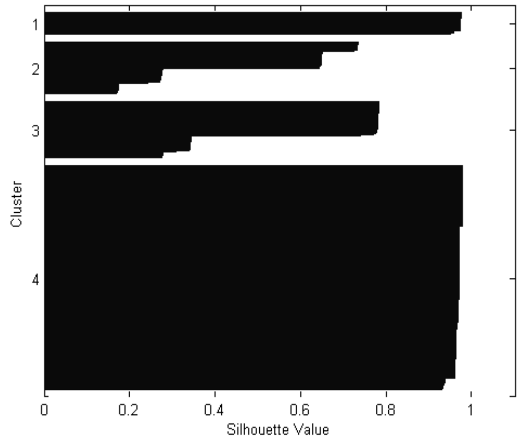

Fig. 4. Silhoutte With 4 Cluster

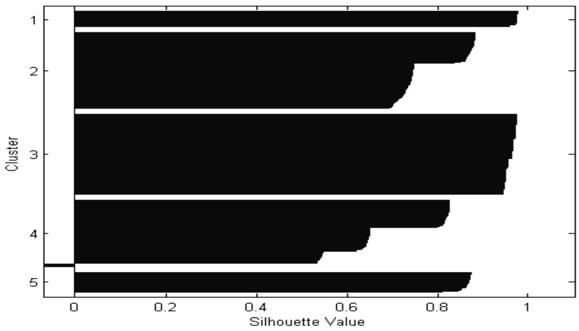

Fig. 5. Silhoutte With 5 Cluste

The graph in Figure 3 shows that the recommendations of the best clusters in cluster 4, where the highest peak in the first cluster position 4. The test results show the student data STMIK WidyaCipta Dharma for graduates of 2005-2009 recommended for admission in cluster 4.

Based on the cluster formed. Students can be grouped into four groups according to values that meet in each of the variables in each cluster and its silhouette can be seen in cluster 4 in Figure 4.

Silhouette cluster 4 shows the number of students who passed the cluster formed at 4 with the largest number means that students graduate on time 4-year study period according to the standard.

For Silhouette cluster 5 shows the number of students is in the negative area, namely in the area of -4 and -5 can be seen in Figure 5.

Looking for the right number of groups with a silhouette. Silhouette is used to see the uniformity of the cluster is formed. If the image silhouette has a tip in negative territory, the clusters formed is not so good because there is no data outliers are included in the cluster. Conversely, if all the cluster does not have a tail in the negative territory, then the results of the cluster can be considered quite good.

Two silhouette of comparison obtained the highest score of the first-Harabasz Calinski is at number four clusters, so that clustering will be made into 4 groups. From the silhouette image is seen that there is no element clusters that are in negative territory compared with 5 clusters. Thus the results of this cluster is quite good and represent the similar groups.

C. Output T-test

K-Means method For graduate students STMIK WidyaCipta Dharma In the 2-year period following graduation:

CLASSIFICATION METHOD OF K-MEANS OF PASS 2005-2014

1) Grouping using K-Means can be concluded for the study program Information Engineering and Information Systems graduate students have a GPA above 3.50

2) Grouping using K-Means can be concluded for the study program Information Engineering and Information Systems graduate students have a very satisfying predicate Graduates

3) Grouping Students For Informatics Engineering Program for the number of graduates have average study period - average 5 years.

REFERENCES

[1] J. Maclennan, Z. Tang, B. Crivat, "*Data Mining with Microsoft SQL Server 2008",* Indiana: Wiley Publishing Inc., 2009.

[2] L. Rokach, O. Maimon, "*Data Mining With Decission Trees Theory and Application",* USA: World Scientific Publishing Co.Ptc.Ltd., 2008.

[3] H. Jiawei, K. Micheline, P. Jian, "*Data Mining: Concepts and techniques",* M. Kaufmann, USA, 2012.

[4] O.P. Rud, "*Data Mining Cookbook:modeling data for marketing, risk, and costumer relationship management",* New York : John Wiley & Son Ltd, 2001.

[5] A. B. Downey, " *Physical Modeling in Matlab",* Needham MA: Gren Tea Press, 2011.

[6] D. L. Olson, D. Delen,*"Advanced Data Mining Techniques",* USA: Spinger, 2008.