

Analysis of New Student Selection using Clustering Algorithms

I M Suartana* and A I N Hidayat

Department of Informatics, Faculty of Engineering, Universitas Negeri Surabaya Jl. Ketintang, Surabaya 60231, Indonesia

*Madesuartana@unesa.ac.id

Abstract. This research describes the analysis and implementation of clustering method which will be used to process data Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) a new student selection, at Surabaya State University. Processing a large number of new student data becomes an annual issue in Surabaya State University. Based on data in 2016, the number of applicants reached 29,779 people. With a large amount of data takes a long time in processing the data to determine the participants who are selected. Our approach uses a clustering method to process participant data and determine the applicant who selected as a new student at Surabaya state university. For analysis and evaluation the accurate and appropriate clustering methods, we selected different clustering techniques that were previously used as benchmarks. The use of clustering may also reduce the cost spent on the application processing and the time the applicants have to wait for the outcome, and could further increase the chances of high-quality applicants getting admission to courses for which they chose. These result also expected can be applied to solve the problem with a similar case.

1. Introduction

Recruitment of new students is an annual-mandatory process for the universities. The recruitment process aims to select prospective students who will be accepted to attend as new students, by defined criteria. One of the new student selection process usually used by the university, based on student grade from high school or vocational high school. In this selection process, processing involves large amounts of data

Clustering is one of the methods used in data processing. Some field that uses Clustering algorithm such as: Finding similar documents using different clustering [1]. Determining the minimum stock and profit margin by building a model that can group items into categories ‘fast moving’ and slow moving’ [2]. Preliminary analysis of interaction and human observation data gathered from students using an Aplusix, an intelligent tutoring system for algebra [3].

In this study clustering algorithm is used to process a new student selection data, and classify data based on similarity of value owned by attribute each data reduction. Cluster analysis can contribute in compression of the information included in data. In several cases, the amount of available data is very large, and its processing becomes very demanding. Clustering can be used to partition the data set into some “interesting” clusters. Then, instead of processing the data set an entity, we adopt the representative soft the defined clusters in our process.

This paper organized as follows. Section 2 discusses clustering techniques. Section 3 presents the methodology in this study. In Section 4, the results and discussion of the experiments. Section 7 describes the research conclusions.



2. Clustering

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data. Analysis clustering is a process finding groups in a set of objects. Some fields use cluster analysis for various purposes. Figure. 1 shows stage in clustering.

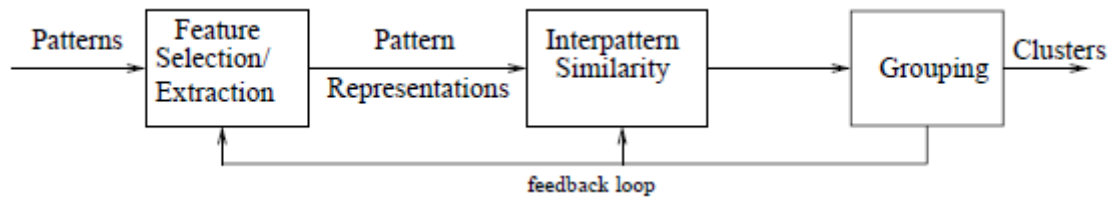


Figure 1. Clustering Steps [clustering]

Feature selection/Extraction is a process to identify the most influential features to use clustering data. Measuring similarity pattern using a distance function. The grouping process can be performed using some algorithms. Clustering output can be a partition of the data, which are grouped into groups or a group data where each pattern has a variable degree of membership in each of the output clusters [4].

2.1. K-Means clustering

K-means is a clustering algorithm, which is mostly applied to solve the clustering problems [5]. K-means is one of the common and simplest clustering algorithms. K-means procedure is grouping data set into some clusters defined by the user. The objective k-means function described by the equation

$$E = \sum_{i=1}^c \sum_{x \in C_i} d(x, m_i) \quad (1)$$

From the equation, m_i is the center of cluster C_i , $d(x, m_i)$ is a distance between a point x and m_i . E function attempts to minimize distance of each point from the each cluster centre which the point belongs. Fig. 2 shows the procedure of k -means algorithm.

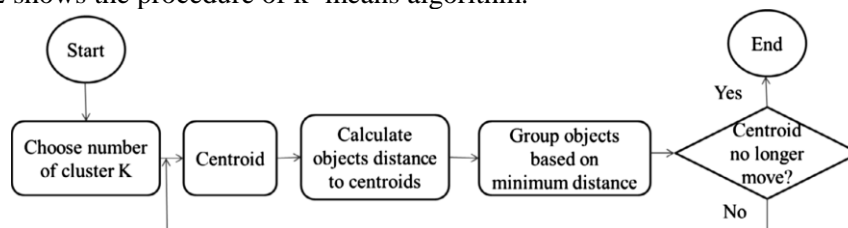


Figure 2. The procedure of k -means algorithm [4].

K-means intent to create clusters of roughly equal size and spherical shape, whereas methods looking for high-density areas may produce clusters of very variable size and shape [6].

2.2. Self-Organization map

Self-Organization Map (SOM) learn to classify data according to how they group in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution (as do competitive layers) and topology of the input data. This algorithm performs clustering process by forming coherent or SOM network used to group data by characteristics or data features.

2.3. Expectation-Maximization

Expectation-maximization (EM) algorithm is an unsupervised learning algorithm that can search knowledge from a set of data that does not have a specific label or target class. EM looking at the value

of each instance distributed into the Gaussian distribution, more precisely the Gaussian mixture, then an ascending iteration to find the value Highest likelihood for each instance.

3. Methods

Cluster analysis is used in many different areas with many different aims; this study aims at clustering new student selection data, at the Universitas Negeri Surabaya. The Problem in new student selection every year is the amount of data to be processed. With a large amount of data takes a long time in processing the data to determine the participants who are selected. The clustering method is selected for compression a large of data into smaller ones, based on similarity information that included in data. Clustering based on student registration data, which consist of course grades data from general high school or vocational high school. After forming student clusters, a cluster analysis was carried out to examine the result.

Cluster analysis can contribute in compression of the information included in data. In several cases, the amount of available data is very large, and its processing becomes very demanding. Clustering can be used to partition the data set into some grouped data called clusters then instead of processing the data set as an entity. We adopt the representative soft the defined clusters in our process, to achieve data compression. Figure 3 Shows Clustering process in this study.

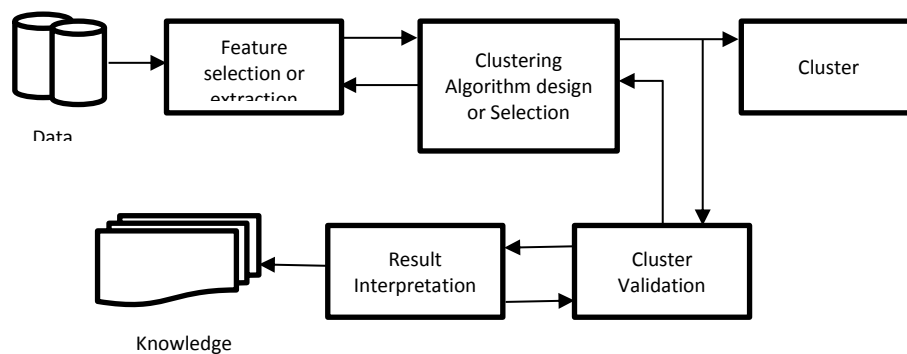


Figure 3. Clustering Process

3.1. Data collection and data processing

The data set is obtained from the student registration data from the University. The data set consist of new student registration data from three departments of the academic year 2016. Data set attribute contains: Ratio of applicants & received in *Seleksi Bersama Masuk perguruan tinggi Negeri* (SBMPTN) a new student selection, The average grade of SBMPTN, the average of high school final exam grade, the average of high school main subject grade, the average high school report grade, and academic excellence student. In this research, trials are conducted using 2000 data records.

3.2. Feature selection

From the initial data of student registration from the University, in this study using the six attributes of data as described in section 3.1 the data collection and data processing. The feature selection chooses differentiates feature, from multiple candidates, while feature extraction uses multiple transformations to generate useful and new features from the original. Both are essential for the effectiveness of data clustering. The appropriate feature selection can greatly reduce the workload and simplify the next design process. Ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret [7].

3.3. Clustering algorithm

In this study, k-means, self-organization map, and Expectation Maximization algorithm is occupied to perform a clustering process. Clustering algorithm design or selection usually select clustering algorithms which are combined with the selection of a corresponding proximity measure. Data in

clustering are grouped according to whether they resemble each other. Obviously, the proximity measure directly affects the formation of the resulting clusters. Almost all clustering algorithms are explicitly or implicitly connected to some definition of proximity measure.

3.4. Cluster validation

Test mode: Classes to clusters evaluation on training data is occupied in this study to evaluate cluster result. Class data obtained from the calculation of data selection by using weighting, the process currently used to determine the participants who are selected. Different approaches, even use the same algorithm usually lead to different clusters, identification, and selection parameter or the presentation order of input patterns may lead the final difference results. Therefore, an effective evaluation standard and criteria are required to provide a level of confidence to the users, against the clustering algorithm used that used.

3.5. Results interpretation.

The final goal of clustering is to provide users with purposeful insights from the original data, so users can effectively solve the problems being faced with. The expected result of data processing selection in the classification of students in several classes according to the characteristics of data carried on by students. The results of this Class will be used to determine the group of students who meet the criteria and are elected to the new students, and the data class declared rejected.

4. Results and discussion

Clustering result using k-means, self-organization map, and Expectation Maximization algorithm, discussed in this section. The data attributes used in clustering are shown in Table 1.

Table 1. Cluster data attributes

Attributes	Symbols
Ratio of applicants & received in (SBMPTN)	nrasiosbmptn
The average grade of SBMPTN	nratasbmptn
the average of high school final exam grade	nrata_un
the average of high school main subject grade	Nratabidstudi
the average high school report grade	Nratarpt
academic excellence student	BM

Cluster analysis can contribute in compression of the information included in data. The main result of our experiments is that the k-means, self-organization map, and Expectation Maximization algorithm could be used to group data based on six data attribute, into two clusters according to user input. Table 2 shows the results of the data cluster.

Table 2. Cluster result

	The academic year 2016	
	Cluster 0(%)	Cluster 1(%)
Department 1	42	58
Department 2	40	60
Department 3	17	83
Department 1	6	94
Department 2	7	93
Department 3	17	83
Department 1	41	69
Department 2	39	61
Department 3	17	83

In Table 3, some iteration represents the count of the objective function needs to satisfy the constraints; cluster centers and Time taken to build a model (full training data).

Table 3. Some iteration and the Time is taken to build a model (full training data).

	The academic year 2016	
	Iteration	Time (second)
Department 1	15	0.001
Department 2	9	0.001
Department 3	6	0.001
Department 1	11	0.02
Department 2	15	0.03
Department 3	1	0.02
Department 1	-	0.39
Department 2	-	0.33
Department 3	-	1.17

Analysis the results clustering algorithm using Classes to clusters evaluation on training data shown in table 4.

Table 4. Result of class to clusters evaluation on training data

	The academic year 2016	
	Incorrectly clustered instances (%)	
Department 1	28	
Department 2	31	
Department 3	27	
Department 1	24	
Department 2	25	
Department 3	27	
Department 1	28	
Department 2	31	
Department 3	27	

For analysis and evaluation the accurate and appropriate clustering methods, we selected Classes to clusters evaluation on training data as benchmarks. The maximum value of classification error based on result classes to clusters evaluation from six experiments with six different data is 31% on 2016 dataset.

5. Conclusion

This study aims to use a clustering algorithm to process data *Seleksi Nasional Masuk Perguruan Tinggi Negeri* (SNMPTN) a new student selection in Universitas Negeri Surabaya. The final goal of clustering utilization to reduce the cost spent on the application processing and the time the applicants have to wait for the outcome, and could further increase the chances of high-quality applicants getting admission to courses for which they chose. From the experiments, result cluster analysis can contribute in compression of the information included in data, and by the result of cluster evaluation using class to cluster evaluation, maximum value of classification error is 33.7%

References

- [1] Al-Anazi S, AlMahmoud H and Al-Turaiki I 2016 Finding Similar Documents Using Different Clustering Techniques *Procedia Computer Science* **82** 28-34
- [2] K Kusriani 2015 Grouping of Retail Items by Using K-Means Clustering *The Third Information Systems International Conference Indonesia*
- [3] Rodrigo M M T, Anglo E A, Sugay J O and Baker R 2008 Use of unsupervised clustering to characterize learner behaviors and affective states while using an intelligent tutoring system In *International Conference on Computers in Education* 57-64
- [4] Jain A K, Murty M N and Flynn P J 1999 Data clustering: a review *ACM computing surveys (CSUR)* **31** 3 264-323
- [5] Saxena A, Prasad M, Gupta A, Bharill N, Patel O P, Tiwari A and Lin C T 2017 A review of clustering techniques and developments *Neurocomputing* **267** 664-681
- [6] C Hennig 2015 What are the true clusters? *Pattern Recognition Letters* 53-62 2015
- [7] D W I Rui Xu 2005 Survey of Clustering Algorithms *IEEE TRANSACTIONS ON NEURAL NETWORKS* **16** 3 645-678