Scientific Journal of Informatics
Vol. 4, No. 1, May 2017

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

K-Medoid Algorithm in Clustering Student Scholarship Applicants

Sofi Defiyanti[1], Nurul Rohmawati W[2], Mohamad Jajuli[3]

[1, 2,3]Informatics Faculty of Computer Science, Universitas Singaperbangsa Karawang
Email: [1]sofi.defiyanti@staf.unsika.ac.id, [2]nurul.rohmawati@student.unsika.ac.id,
[3]mohamad.jajuli@staf.unsika.ac.id

Abstract

Data Grouping scholarship applicants Bantuan Belajar Mahasiswa (BBM) grouped into 3 categories entitled of students who are eligible to receive, be considered, and not eligible to receive scholarship. Grouping into 3 groups is useful to make it easier to determine the scholarship recipients fuel. K-Medoids algorithm is an algorithm of clustering techniques based partitions. This technique can group data is student scholarship applicants. The purpose of this study was to measure the performance of the algorithm, this measurement in view of the results of the cluster by calculating the value of purity (purity measure) of each cluster is generated. The data used in this research is data of students who apply for scholarships as many as 36 students. Data will be converted into three datasets with different formats, namely the partial codification attribute data, attributes and attribute the overall codification of the original data. Value purity on the whole dataset of data codification greatest value is 91.67%, it can be concluded that the K-Medoids algorithm is more suitable for use in a dataset with attributes encoded format overall.

Keywords: Scholarships, Clustering, Data Mining, K-Medoids, Purity Measure

1. INTRODUCTION

One of the reasons many students apply for academic leave even drop out the about the high tuition fees that affect the continuity of learning activities at a higher education institution. Scholarship assistance is given to students who are less able to meet its obligations during the period of study. The scholarship is of course also have to pay attention to certain criteria before it is given to the students concerned. The criteria depend on the conditions set by the scholarship. Another function of these scholarships as well as awards to outstanding students both in academic and non-academic. In this study scholarships that will be discussed is about BBM scholarship or Student Learning Assistance. Where this scholarship is a scholarship reserved for underprivileged students and have achievements in the field of academic and non-academic [1]. Algorithm k-means and K-Medoids of Technik clustering can help in classifying students are eligible to receive the scholarship, students in consider receiving and students who are not eligible to receive a scholarship.

The comparing PAM (Partition Around Medoids) and k-means clustering to tweets,it is known that an algorithm in clustering can be judged good or not based on the value of purity. value purity this is used to measure clustering results of each algorithm (k-means and partition around medoids [2]. Based on these studies will be conducted research using the value of purity to assess an algorithm K-Medoid but with different

data formats, so that can know better results clustering (from several different data formats) to determine the scholarship recipients.

The purpose of this study was to compare the value of purity to find out the results cluster from each attribute in determining the scholarship recipients. So, they will know the attributes of different formats, which has been generated.

This study used a methodology data mining CRISP-DM which consists of six stages, because this study aimed to compare the results of clustering,the CRISP-DM phases only until on stage 5. the stages as follows, business understanding, understanding of data, data processing, modeling and evaluation of [3].

2. METHODS

Methods are six CRISP-DM process data mining as illustrated in Figure 1 below:

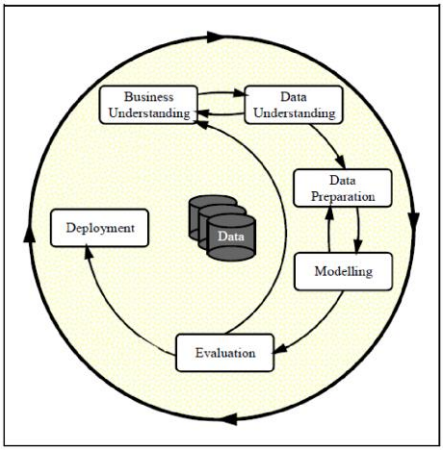

Figure 1. Model Crips- DM

a. Bussiness understanding

In this phase focuses on understanding and perspective of the business processes of a system. Namely the determination of project goals, translating the objectives, and prepare a strategy for the delivery destination.

b. Data Understanding

In this phase focusing on learning the existing data, collecting and sorting data.

c. Data Preparation

The phase of data preparation is the phase that consists of a selection of data, data cleansing, integrating data, and transformation of data to be continued into the modeling phase.

d. Modeling

In this phase of the process that occurs is the selection of an appropriate model. Modeling herein can be calibrated to optimize the results. Modeling with algorithm K-Medoids will be made to a group of recipients.

e. Evaluation
In this phase will be the evaluation process from the previous phase. the phase of this evaluation will be conducted comparative quantitative by considering the value of purity (Purity Measure).

f. Deployment
In this phase the process is happening is the preparation of a report or presentation of knowledge gained from the evaluation of the process data mining [3].

3. RESULTS AND DISCUSSION

3.1. Business Understanding

The purpose of business is based a description of the function of scholarships, among others, to help ease the burden of students in lectures, so bear the cost of reducing the number of students who dropped out of college because of financial problem. The purpose of this study was to compare the value of purity to find out the results cluster from each- each format attribute in determining the scholarship recipients. clustering to be used in cluster students who apply for scholarships fuel. Then the results of clustering are will be known the algorithm which has the result of cluster better so that it can be in the know students right receive scholarships fuel based cluster that right.

3.2. Data Understanding

From the results of data collection has been performed the data obtained as many as 36 students who apply for scholarships. Then from this data will have the criteria required for entry into the next stage. These criteria are, NPM, GPA, the number of credits that have been taken, the amount of parental income and number of dependents of parents.

3.3. Processing Data

From the data collected, there is some missing value on the criterion of the income of the parents, then missing value will be filled using techniques mean imputation or filled with value - the average of the criteria income parents with formula 1.

$$\bar{X} = \frac{\frac{Jumlah\ total\ atribut}{penghasilan\ orang\ tua}}{Jumlah\ data} \quad (1)$$

$$\bar{X} = \frac{62208900}{36}$$

$$\bar{X} = 1728025$$

So, value-average parental income criteria is Rp. 1,728,025, -. The categorization criteria parents income divided by the number of dependent parent (in this study abbreviated to JP) [4] and each of the criteria then categorized Based on Table 1:

Table 1. Categorization JP

Category	Qualifications	Codification
Category 4	JP≤ x - S	4
Category 3	x - S <JP <x	3
Category 2	JP ≤ x <x + S	2
Category 1	JP ≥ x + S	1

After calculating the unknown:
Mean JP(x): 1025394.4
JP Standard Deviation*(S)*: 705,913.89
and the results obtained to JP categorization presented in Table 2.

Table 2. table categorization JP

Category	Qualifications	Codification
Category 4	JP≤ Rp. 319,480.5	4
Category 3	Rp. 319,480.5 <JP <IDR. 1025394.4	3
Category 2	Rp. 1025394.4 ≤ JP <IDR. 1731308.3	2
Category 1	JP ≥ Rp. 1731308.3	1

and categorization criteria credits by finding value standard deviation and the mean of each criterion and then categorized Based on Table 3:

Table 3. Categorization SKS

Category	Qualifications	Codification
Category 5	SKS≤ x - 2S	5
Category 4	X - 2S ≤ SKS <x - S	4
Category 3	x - S ≤ SKS <x + S	3
Category 2	x + S ≤ SKS <x + 2S	2
Category 1	Credit ≥ x + 2S	1

After calculating the unknown:
Mean credits*(x)*: 75.78
SKS Standard Deviation (S):18.897
AAnd the results obtained for SKS categorization presented in Table 4.

Table 4. results categorization SKS

Category	Qualifications	Codification
Category 5	SKS≤ 38	5
Category 4	38 <SKS <56.89	4
Category 3	SKS ≤ 56.89 <94.67	3
Category 2	94.67≤ SKS 113.56	<2
Category 1	Credit113.56 ≥	1

After the categorization of the attributes of SKS and JP (earnings divided by the number of dependent elderly parents), then create a dataset with the name. dataset partial codification And to make the dataset whole codification to attribute GPA categorized based on the rule-making number of credits based on the CPI, with provisions such as in Table 5 below:

Table 5. Rule-making SKS based GPA

Credits	GPARange	Category
24	3:00 to 4:00	1
21	2:50 - 2.99	2
18	2:01 to 2:49	3
15	1.90 - 2:00	4
12	<1:49	5

This study was conducted for the three types of datasets, the dataset codified in part, the dataset codification overall and dataset original data (attributes that are not categorized).

3.4. Modeling

Modeling data mining in this study were made using the software RapidMiner Studio 5. in this application has been available algorithms clustering such as algorithms. k-medoid

Algorithms K-Medoids

a. Dataset partial codification
   Medoids the end produced as in table 6.

Table 6. *Medoids dataset* partial codification

	GPA	SKS	JP
Cluster 1	3,880	3	3
Cluster 2	3,450	3	1
Cluster 3	3,440	3	3

b. Dataset codification overall
   Medoids the end produced as in Table 7.

Table 7. *Medoidsdataset whole codification*

	GPA	SKS	JP
Cluster 1	1	3	2
Cluster 2	1	3	1
Cluster 3	1	3	3

c. Dataset original data
   Medoids generated end ie as in Table 8.

Table 8. *Medoids dataset* original data

	GPA	SKS	JP
Cluster 1	3,690	84	1,750,000
Cluster 2	3450	84	3,500,000
Cluster 3	3880	84	1,000,000

3.5. Evaluation

Using equation 2 for testing *purity measure* (r) for algorithm *K-Medoids* comparison value *purity* (r) the *dataset* with attribute data is codified in part, the overall codification of data and the original data. It can be concluded the higher the R value (closer to 1), the better the quality of their *cluster.*

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i \qquad (2)$$

Where:
r: accuracy level *clustering*
k: number of *the clusters*
a_i: objects that appear within the *cluster* C_i and the label *class* accordingly.

Result values *purity measure* algorithm *K-Medoids* shown in Table 9 and Figure 2 shows a comparison chart value *purity measure.*

Table 9. Purity Measure algorithm K-Medoids

Purity Measure (r)	
Dataset	*K-Medoids*
Codification most	0833
Codification Overall	0917
Original Data	0778

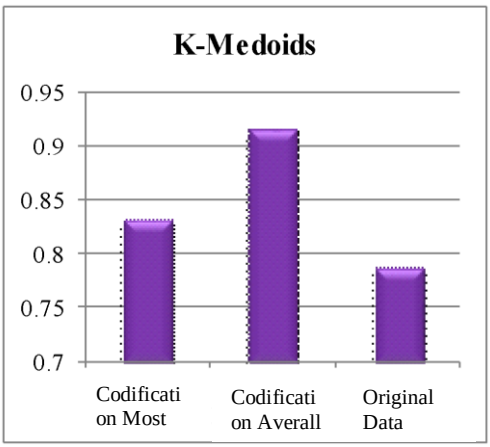

Figure 2. Graph comparison of *Purity Measure*

3.6. Discussion

Based on the counter value comparison purity measure the results of clustering algorithm K-Medoids the dataset attribute codification mostly known for 0833 or 83.33%. And for the dataset, the codification of the entire the results of cluster

algorithm K-Medoids known by 0917, or 91.76%. For dataset, original data in this study contain outliers, the known value of purity (r) for the results of cluster algorithm K-Medoids of 0778 or 77.78%.

So, we can conclude that for an algorithm K-Medoids, the dataset with attribute data that codified a whole have the results cluster better. This is because the algorithm K-Medoids using object selected randomly as the centers can clusters (medoid),as well the Euclidean as a function of distance to calculate the distance between the proximity of an object with medoid. Therefore members of a cluster are generated by an algorithm K-Medoids more likely similar to the object medoid her which was an object is selected randomly.

4. CONCLUSION

The comparing results of cluster algorithm K-Medoids based on the clustering of each format dataset Different (codified in part, the overall codification and the original data) to measure the accuracy rate clustering which calculates the value of purity measure of the results of the cluster. The greater the value of purity (closer to 1) the better the quality of the clusters produced by an algorithm.

Based on the counter value comparison purity measure the results clustering of the algorithm K-Medoids by formatting different attributes datasets (partly data attribute in codified, attributes codified data, and entirely original data attribute). Unknown value purity on a dataset of data codification part to the results of cluster algorithm k-medoids of 0833 or 83.33%. On the dataset overall value of the codification purity results of cluster algorithm K-Medoids of 0917 or 91.67%. For dataset, original data grades purity result from cluster algorithm K-Medoids of 0778 or 77.78%. It can be concluded that the level of accuracy of clustering the results clusters algorithm K-Medoids based on the purity measure, the dataset which codified the entire better than dataset that in the codification partial and the datasets original data.

5. REFERENCES

[1] DIKTI. 2015. *Pedoman umum Beasiswa dan Bantuan Biaya Pendidikan Peningkatan Prestasi Akademik (PPA)*. http://belmawa.ristekdikti.go.id/dev/wp-content/uploads/2015/11/PEDOMAN-BEASISWA-BBP-PPA-2015.pdf, diakses 15 Januari 2016.

[2] Wibisono, Y., 2011. Perbandingan Partition Around Medoids (PAM) dan K-means Clustering untuk Tweets. *Prosiding Konferensi Nasional Sistem Informasi*, pp.25-26.

[3] Budiman, I., Kom, M., Prahasto, I.T., ASc, M. and Yuli Christiyono, S.T., 2012. *Data Clustering Menggunakan Metodologi CRISP-DM untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma (Doctoral dissertation*, Universitas Diponegoro).

[4] Rohmawati, N. Defiyanti, S. Jajuli, M. 2015. Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jitter Jurnal Ilmiah Teknologi Informasi Terapan*. Vol. I (2). 62-68.