Utilization of Data Mining in the Processing of Average Values of High School Level National Examination at the Department of Natural Sciences in Indonesia

by Siti Choiriyah

Submission date: 11-Jan-2020 09:50AM (UTC+0700)
Submission ID: 2109315428
File name: Scopus_2020_Siti_Choiriyah_Utilization.doxc (109.46K)
Word count: 2702
Character count: 14210

Utilization of Data Mining in the Processing of Average Values of High School Level National Examination at the Department of Natural Sciences in Indonesia

Lukman Nasution[1], Siti Choiriyah[2], Abdul Rahmat[3], Heldy Vanni Alam[4], Salam[5]*

[1]Universitas Muslim Nusantara Al Washliyah Medan, Indonesia.
[2]Institut Agama Islam Negeri Surakarta, Indonesia.
[3]Gorontalo State University, Indonesia.
[4]Universitas Negeri Gorontalo, Indonesia,
[5]Universitas Negeri Gorontalo, Indonesia.

Abstract-The Computer Standards National Examination (abbreviated as UNBK) computer standard is a phenomenon of technological progress that has a big impact and impact on all aspects of life such as the world of education in Indonesia which is demanded to always develop every year so that Indonesian people get better quality education. The purpose of this research is to apply data mining techniques in analyzing the processing of the national standardized high school-level computerized national exam score in the Department of Natural Sciences in Indonesia. The data used is the data of the Ministry of Education and Culture processed by the Central Statistics Agency. The attribute used is the average UNBK grade at the high school level majoring in Natural Sciences by province 2018/2019 academic year consisting of 35 provinces. The method used is K-Medoids which is part of clustering. By using Davies-Bouldin Index (DBI) obtained 2 cluster labels namely C1: high cluster and C2: low cluster with the best DBI is 0,427. The results of the study stated that only 14,3% or 5 provinces were included in the high cluster (C1) namely: Bali, DKI Jakarta, Central Java, Riau islands and Yogyakarta. While 85,7% or 30 other provinces are included in the lower cluster (C2). For the final centroid value used for each cluster are C1 = 65,35 and C2 = 47,93. It is expected that the results of the study can provide information to the government that some provinces in western, central and eastern Indonesia are still in the lowest position that has an impact on the quality of education in Indonesia.The analysis of each clustering experiment is presented in this paper.

Keywords: Data mining, Clustering, K-Medoid Method, Average of National Standardized Computer Exam

Introduction

Based on data from the ministry of education and culture, the implementation of the Computer-Based National Examination (abbreviated UNBK) was first held in 2014 online in Indonesia. This UNBK system uses the computer as a test medium and in its implementation the UNBK is different from the paper based national test system (Paper Based Test (PBT) which has been running so far. With the progress of the current era (industrial revolution), the implementation of the Computer Based National Examination is a rapid technological advancement and has a profound impact and impact on various aspects of life such as advances in the world of education in Indonesia which are demanded to always develop each year so that Indonesian people get quality education the better. Based on data from the Ministry of Education and Culture, the average high school level UNBK examination results at the Department of Natural Sciences by region are still below the national average. The purpose of this study was to analyze the average pattern of UNBK grades at the high school level at the Department of Natural Sciences by region using mapping in the form of clusters. The results of the study are in the form of mapping the region to the average UNBK examination scores so that the region that has a UNBK score below the national average is obtained.

In the development of computer science, there are many methods that can be used in mapping clusters. One of them is data mining [1]. Data mining is a technique used to examine large databases as a way to find new and versatile patterns [2], [3]. There are several data mining techniques that are used such as clustering, classification, estimation and association [4]. Each technique has different methods of solving problems. The research that has the outcome in the form of mapping of the cluster that has been determined is clustering [5]–[7]. The use of clustering depends on the type of data available at the destination [8]. One of the most popular methods of applying clustering is K-Medoids and K-Means [6]. K-Medoids and K-Means have similarities in grouping data according to characteristics and measuring the similarity between data in one group [9]. In addition, both methods have advantages in fast computing time. In several studies mentioned that the K-Medoids

DOI: 10.5373/JARDCS/V12I6/S20201170
**Corresponding Author: Lukman Nasution , Email: lukmannasution@umnaw.ac.id*
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

method is better than K-Means where K-Means has a sensitive weakness to outlier data that can be overcome K-Medoids that can almost work on any type of matrix data and are able to overcome outliers [10]. In addition, several previous studies that used the advantages of K-Medoids were [11] titled Clustering of patient trajectories with an auto-stopped bisecting K-Medoids algorithm. The results mentioned that the method can be applied to classify patients into manageable groups. In addition, research [12]with the title Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm. The results of the study mentioned that the K-Medoids method can be applied and is better than K-Means in analyzing the spatial pattern of earthquake distribution in Indonesia. Based on this, it is expected that the research results can provide information on the results of optimal cluster mapping on the processing of the average value of the national high school-level computerized standardized national exam at the Department of Natural Sciences in Indonesia.

Methodology

1.1. Data

Sources of data in the study were obtained from data from the Ministry of Education and Culture processed by the Central Statistics Agency (https://www.bps.go.id/) in the average category of Computer-Standards National Examination (abbreviated UNBK) for high school level majoring in Natural Sciences according to the Povinsi 2018/2019 school year as many as 35 data records. The result of the K-Medoids method is to find out the provincial grouping in the form of cluster mapping of the average value of the Computer Standard National Examination (abbreviated UNBK) at the high school level in the natural science majors in Indonesia. The attributes used are in Table 1.

Table 1. Attribute used

Role	Name	Type	Range	Missings
-	UNBK Average Value	real	= [42.740...66.900]; mean =50.949	no missing values
id	Region Name	polynominal	= {Aceh, Bali, Bangka Belitung, Banten, Bengkulu, Central Java, Central Kalimantan, Central Sulawesi, DKI Jakarta, ...}	no missing values

The following research data are used after pre-processing data to maximize cluster results using Microsoft Excel software. The average data table of the National Standardized Computerized National Examination for high school level at the Department of Natural Sciences can be seen in the following table:

Table2. Research data

No	Region Name	UNBK Average Value
1	Aceh	43,03
2	Bali	57,59
3	Bangka Belitung	54,55
4	Banten	53,25
5	Bengkulu	50,38
6	DKI Jakarta	66,9
7	Gorontalo	46,81
8	Jambi	49,34
9	West Java	53,54
10	Central Java	59,32
11	East Java	56,28
12	West Kalimantan	51,36
13	South Kalimantan	53,65
14	North Kalimantan	50,22
15	Central Kalimantan	48,8
16	East Kalimantan	55,56
17	Riau islands	59,91
18	Lampung	50,32
19	Maluku	43,71
20	North Maluku	42,74
21	National	52,43
22	NTB	46,22
23	NTT	46,23
24	Papua	45,08

No	Region Name	UNBK Average Value
25	West Papua	47,7
26	Riau	51,67
27	West Sulawesi	44,37
28	South Sulawesi	46,21
29	Central Sulawesi	45,93
30	Southeast Sulawesi	47,03
31	North Sulawesi	47,03
32	West Sumatra	54,81
33	South Sumatra	47,98
34	North Sumatra	47,93
35	Yogyakarta	65,35

Source: Ministry of Education and Culture

Based on the introduction described, this research through several stages, the following stages of the framework in the preparation of research as shown in the following figure:

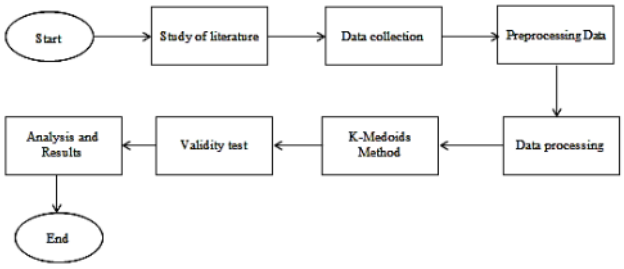

Figure 1. Research methodology

In Figure 1, an explanation of the stages in the process of applying the k-medoids method to the processing of the average national high school level computerized standardized exam in the science department in Indonesia is:

a) Carry out the process of analyzing a problem by using all references such as libraries, books or journals as a medium for reference material used in the research being carried out;
b) Preprocessing data on the data used (cleaning and data transformation) to maximize the results of data clustering;
c) Conduct a clustering process using the K-Medoids method;
d) Display the output results from clustering with cluster quality using validity test in determining the number of clusters (k) with Davies-Bouldin Index (DBI).

1.2. K-Medoids Method

The K-Medoids (Partitioning Around Medoids) method is a method similar to k-means because both of these methods break down the dataset into groups that have similarities [13]. Some advantages are can handle sensitive outliers due to an object with a large value [6] and can work on any type of data matrix [12]. The K-medoids method is more suitable for grouping data than the K-Means method [14]. Some steps to complete the K-Medoids method[10]:

a) Initialise the cluster center as many as the number of clusters (k).
b) Each data or object is allocated to the nearest cluster using the Euclidian Distance equation:

$$d(x,y) = |x-y| = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

Where:

d = the distance between x and y
x = cluster data center
y = data on the attribute
i = every data
n = amount of data
xi = data in the cluster center
yi = data on each data

c) Choose objects on each cluster randomly as new medoid candidates.
d) Calculate the distance of each object contained in each cluster with the new medoid candidate.

DOI: 10.5373/JARDCS/V12I6/S20201170
*Corresponding Author: Lukman Nasution , Email: lukmannasution@unnaw.ac.id
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

e) Calculate the total deviation (S) by calculating the total new distance value - the total old distance. If S <0 is obtained, swap objects with data
f) Cluster to create a new set of k objects as medoids.
g) Repeat steps c through e until there is no change in the medoid, so that clusters and cluster members are obtained.

Results and Discussion

After conducting the data pre-processing stage, the data is then processed with RapidMiner 5.3 software to show the Davies Bouldin Index (DBI) value as a reference for the best cluster grouping. Tests carried out on the number of clusters k = 2 to k = 4. The following results of data processing are shown in the following table:

Table2. Comparison of Davies Bouldin Index Results

Cluster	K-Medoids	K-Means
k=2	0,427	0,574
k=3	0,488	0,517
k=4	0,841	0,484

Based on the results of a comparison of the Davies Bouldin Index, shows that the number of k = 2 is the best cluster. Table 2 also explains the comparison of the Davies Bouldin Index with the k-means method. This is the reason why the k-medoids method is far better than k-means. After determining the best cluster number (k = 2), the calculation process is performed using the RapidMiner 5.3 software through the available Data Import Wizard tool. The import process can be done with various extensions such as xls, spss, csv, access, ARFF, xml, XRFF and others. In this study the import process uses excel data as shown in table 2. On the main process sheet, the k-medoids model in processing UNBK values is designed as shown in the following figure:

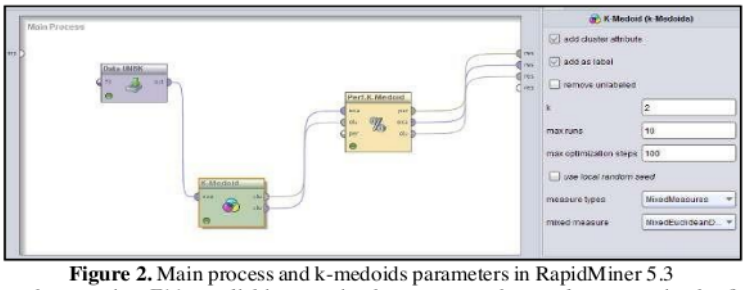

Figure 2. Main process and k-medoids parameters in RapidMiner 5.3

Run the process by pressing F11 or clicking run in the program, the results appear in the form of cluster initialization mapping for each object with the number of clusters (k = 2). The final centroid results on k-medoids can also be displayed as shown in the following image:

Attribute	cluster_0	cluster_1
UNBK Average Value	65.350	47.930

Figure 3.the final centroid results

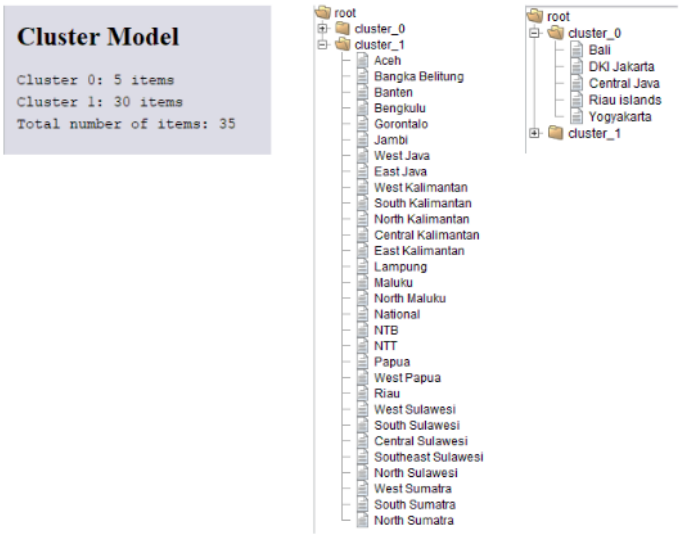

Figure 4. Final cluster results

Based on Figure 4, the final results of cluster mapping for processing the average national high school-level computerized national exam score in Indonesia are C1: high cluster (cluster_0) consisting of 5 provinces (Bali, DKI Jakarta, Central Java, Riau islands and Yogyakarta) and C2 : low cluster (cluster_1) consisting of 30 provinces (Aceh, Bangka Belitung, Banten, Bengkulu, Gorontalo, Jambi, West Java, East Java, West Kalimantan, South Kalimantan, North Kalimantan, Central Kalimantan, East Kalimantan, Lampung, Maluku, North Maluku, National, NTB, NTT, Papua, West Papua, Riau, West Sulawesi, South Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi, West Sumatra, South Sumatra, North Sumatra) with the centroid end result for C1 = 65,35 and C2 = 47,93. The following also results of the visualization of cluster mapping based on UNBK and provincial values as shown in the following figures 5 and 6:

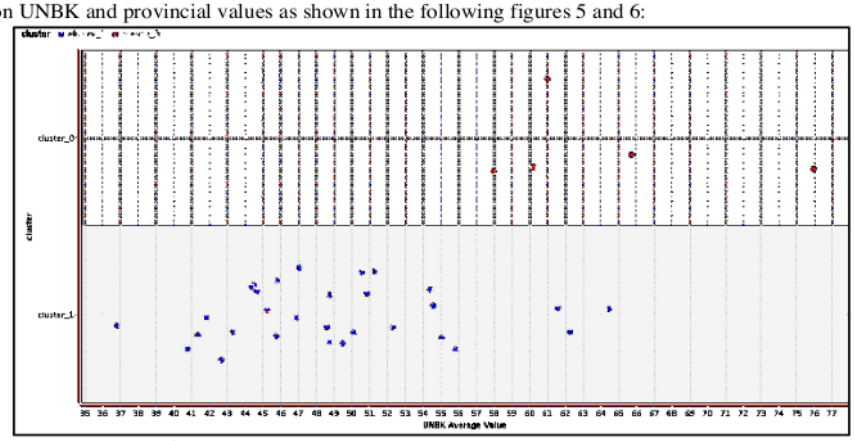

Figure 5. Cluster visualization based on UNBK average values

DOI: 10.5373/JARDCS/V12I6/S20201170
*Corresponding Author: Lukman Nasution , Email: lukmannasution@umnaw.ac.id
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

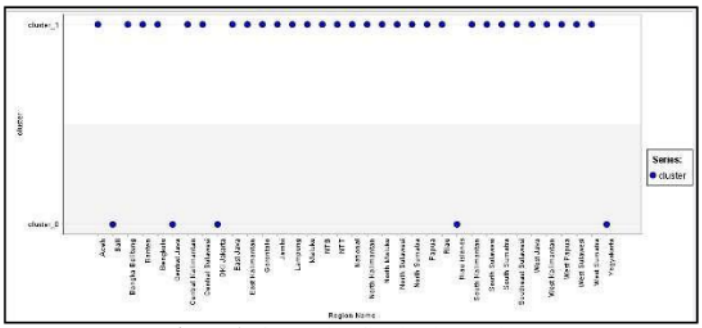

Figure 6. Cluster visualization by province

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -17.461
Avg. within centroid distance_cluster_0: -25.715
Avg. within centroid distance_cluster_1: -16.085
Davies Bouldin: -0.427
```

Figure 7. Davies Bouldin's vector performance at the number k= 2

In Figure 7, the results of the Davies-Bouldin Index (DBI) on the Cluster Distance Performance operator are used to evaluate the performance of the centroid-based clustering method by providing a list of performance criteria values based on the centroid cluster. In processing UNBK values by province, the Davies-Bouldin Index has an optimal clustering (k = 2): 0.427. This result is far better than the amount of clustering (k = 3).

Conclusion

The results of the study mentioned that the processing of the average value of the National High School Level Computerized National Examination at the Department of Natural Sciences in Indonesia can be applied using the K-Medoids method. By using 2 mapping clusters, the results obtained are 5 provinces in the C1 category (high cluster) and 30 provinces in the C2 category (low cluster). The value of low cluster (C2) is Aceh, Bangka Belitung, Banten, Bengkulu, Gorontalo, Jambi, West Java, East Java, West Kalimantan, South Kalimantan, North Kalimantan, Central Kalimantan, East Kalimantan, Lampung, Maluku, North Maluku, National , NTB, NTT, Papua, West Papua, Riau, West Sulawesi, South Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi, West Sumatra, South Sumatra, North Sumatra. From these results 85% of regions in Indonesia still have UNBK values below the national average which is mostly located in the eastern, western and central parts of Indonesia.

References

[1] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, "Classification of natural disaster prone areas in Indonesia using K-means," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.

[2] A. P. Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method," *Int. J. Artif. Intell. Res.*, vol. 1, no. 2, pp. 26–33, 2017.

[3] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.

[4] R. Suryawanshi, "A Novel Approach for Data Clustering using Improved K- means Algorithm," vol. 142, no. 12, pp. 13–18, 2016.

[5] S. Defiyanti, M. Jajuli, and N. Rohmawati, "K-Medoid Algorithm in Clustering Student Scholarship Applicants," *Sci. J. Informatics*, vol. 4, no. 1, pp. 27–33, 2017, doi: 10.15294/sji.v4i1.8212.

[6] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016, doi: 10.1016/j.procs.2016.02.095.

[7] E. M. Rangel, W. Hendrix, A. Agrawal, W. K. Liao, and A. Choudhary, "AGORAS: A fast algorithm for estimating medoids in large datasets," *Procedia Comput. Sci.*, vol. 80, pp. 1159–1169, 2016, doi: 10.1016/j.procs.2016.05.446.

[8] S. Harikumar and P. V. Surya, "K-Medoid Clustering for Heterogeneous DataSets," *Procedia Comput. Sci.*, vol. 70, pp. 226–237, 2015, doi: 10.1016/j.procs.2015.10.077.

[9] L. Peng, G. Y. Dong, F. F. Dai, and G. P. Liu, *A new clustering algorithm based on ACO and K-medoids*

optimization methods, vol. 19, no. 3. IFAC, 2014.

[10] A. Wanto *et al*., *Data Mining : Algoritma dan Implementasi*. Medan: Yayasan Kita Menulis, 2020.

[11] H. Fei, N. Meskens, and C. H. Moreau, "Clustering of patient trajectories with an auto-stopped bisecting K-medoids algorithm," *IFAC Proc. Vol.*, vol. 13, no. PART 1, pp. 355–360, 2009, doi: 10.3182/20090603-3-RU-2001.0281.

[12] F. R. Senduk and F. Nhita, "Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm," *Ind. J. Comput.*, vol. 4, no. 2016, pp. 65–76, 2019, doi: 10.21108/indojc.2019.4.3.359.

[13] B. Wira, A. E. Budianto, and A. S. Wiguna, "Implementasi Metode K-Medoids Clustering Untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 Di Universitas Kanjuruhan Malang," *Rainstek*, vol. 1, no. 3, pp. 54–69, 2019.

[14] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018, doi: 10.24014/coreit.v4i2.4498.

DOI: 10.5373/JARDCS/V12I6/S20201170

**Corresponding Author: Lukman Nasution , Email: lukmannasution@umnaw.ac.id*

Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

Utilization of Data Mining in the Processing of Average Values of High School Level National Examination at the Department of Natural Sciences in Indonesia

ORIGINALITY REPORT

11%	13%	6%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	cibg.org.au Internet Source	7%
2	digitalcommons.unl.edu Internet Source	5%

Exclude quotes	On	Exclude matches	< 5%
Exclude bibliography	On		