Chapter 21
Stepwise Iterative Maximum Likelihood Clustering Based on *Kompetisi Sains Madrasah'* Scores for Identifying Quality of Junior High School Grading Distribution

Kusaeri, Nanik Puji Hastuti, Ali Musthofa, Ahmad Fauzi, Ahmad Hanif Asyhar, Dian Candra Rini Novitasari, Dwi Rolliawati, Zakiyatul Ulya, Ahmad Yusuf, Nurissaidah Ulinnuha, and Noor Wahyudi

Abstract Kompetisi Sains Madrasah (KSM) is a science competition organized by the Ministry of Religious Affairs Indonesia. KSM scores can help to determine the quality distribution of students based on KSM data. Clustering is a method that can group data based on the similarity of data. One method of student clustering that has never been done is Stepwise Iterative Maximum Likelihood. SIML does clustering based on the average value and standard deviation of each cluster. SIML grouping is done several times by creating several different clusters. The best grouping experimental results are in the number of clusters 3. Cluster data is divided into 3 groups consisting of very good, good, and fair students. Silhouette values indicate the best grouping using 3 clusters. The silhouette value is 0.423745 for natural sciences subject data. Social sciences data has a silhouette value of 0.415654, and a silhouette value of 0.487071 for mathematics subjects. Three silhouette values are the best silhouette values when compared to other experiments.

21.1 Introduction

Education has a role in the progress of a nation. A country experiences underdevelopment in various aspects if a country does not pay attention to the development of education. Some developed countries make education as an investment to avoid reducing the quality of society. As a country in the process of becoming a developed

Kusaeri (✉) · A. Musthofa · A. Fauzi · A. H. Asyhar · D. C. R. Novitasari · D. Rolliawati ·
Z. Ulya · A. Yusuf · N. Ulinnuha · N. Wahyudi
UIN Sunan Ampel Surabaya, Ahmad Yani 117, Surabaya, Indonesia
e-mail: kusaeri@uinsby.ac.id

N. P. Hastuti
Kementerian Agama RI, Lapangan Banteng 3-4, Jakarta, Indonesia

Y.-D. Zhang et al. (eds.), *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020*, Smart Innovation, Systems and Technologies 182,
https://doi.org/10.1007/978-981-15-5224-3_21

country, Indonesia has carried out various educational developments. The progress of Indonesian education can affect the progress of the Indonesian state [1]. Increasing the quality of education is one of the efforts aimed at obtaining quality, advanced, and independent communities. One of Indonesia's efforts in developing education is to increase the quantity and quality of education. Enhancing the quality of education is realized by evaluating and developing educational strategies, while one of the quantitative efforts is to provide formal school facilities from urban to rural areas [2].

Formal schools established in Indonesia have various types including public schools and private schools. State schools are schools operated by the government. The government handles all school needs. While private schools are schools that are operated by private parties [3]. In its development, various schools were established to help to advance Indonesian education. The number of schools increases every year. The number of schools from elementary school to senior high school (SLTA), special school in Indonesia, reaches 307,655 schools. This amount is based on the basic education data of the Ministry of Education and Culture. When viewed from the type of school, there are 169,378 public schools and 138,277 private schools [4]. Several educational problems follow the growth of schools in Indonesia. One issue that arises is the school at the junior high school level.

One form of treatment in junior high school problems is holding a competition between students. *Kompetisi Sains Madrasah* (KSM) is a competition held by the Ministry of Religion of the Republic of Indonesia. KSM is followed starting from elementary school, junior high school, and senior high school. Madrasah Science Competition is a science competition that began in 2016. Students participate in KSM from elementary, junior high, and high schools under the auspices of the Ministry of Education and Culture [5]. KSM competition results are a sample of the quality scores of students from various provinces. The distribution of student grades in each region has a different score. A grouping system of score data is needed to see the diversity of student quality in Indonesia.

Clustering is a process of grouping data into a data class with high similarity and having data differences with other clusters. Previous research has conducted clustering of the quality of elementary school students using the k-means algorithm [6]. The clustering aims to identify the quality of elementary school students. In this research, clustering is done using the Stepwise Iterative Maximum Likelihood (SIML) algorithm. Maximum likelihood classifies data by looking at the average and standard deviation values. A random value is raised to be the center point of the cluster. The optimal value of the center cluster is obtained if it has an average and standard deviation [7]. Previous research stated that a maximum likelihood method can distinguish overlapping data. In addition, the maximum likelihood can group the number of samples that are lower than the data dimension [8]. Other studies use maximum likelihood clustering as a method for classifying people with leukemia. Maximum likelihood clustering groups data as 2, 3, 4, and 5 dimensions with excellent success. The success rate of clustering using the maximum likelihood can reach 90% more, as in previous research studies [8].

The previous clustering of education was carried out using the *k*-means method. Clustering is intended to help to reduce the number of students dropping out of school, and increase the quality of student learning [9]. The use of the *k*-means method is also provided to classify the quality of students using online learning data [10]. Another method used in the clustering process is the fuzzy c-means method. The clustering of students is aimed at providing evaluations of students needed by a teacher in improving the quality of learning [11]. Some student quality analysis that has been done has not used the SIML method. This study proposes the SIML method for classifying junior high school students based on the score of KSM.

The results of clustering using maximum likelihood require an evaluation system to see the level of success. One form of clustering evaluation is calculating data distance. The calculated distance is the distance of data in one cluster, as well as with other clusters. In addition, evaluation can also be seen from the distance of data to the cluster center point. The silhouette method can calculate all distances [12]. The results of the silhouette evaluation on the SIML clustering are used as a reference for analyzing the results of clustering. KSM data that has been clustered is analyzed for each cluster. Analysis of the results of clustering is intended to determine the quality of students. In addition, the results of the clustering of data can be used as a reference to provide action in improving the quality of education in Indonesia.

21.2 Stepwise Iterative Maximum Likelihood (SIML)

Maximum Likelihood Clustering (MLC) is a method that classify data by looking the average values and standard deviations from the center of the cluster [13–15]. Stepwise Iterative Maximum Likelihood (SIML) is a clustering method developed from MLC. SIML finds the best cluster center by finding the optimal partition repeatedly. Partition search is made by shifting the partition from the initial data to another partition. Partition optimization is given by looking at the log-likelihood value. If the partition shift results in a high log-likelihood value, the partition is directed toward the latest partition point [7]. Suppose the data is grouped into 2 groups, each cluster is marked with a red and orange circle. Each cluster has a midpoint as the center of the initial cluster. Cluster center points have coordinates in the form of a mean (μ) and standard deviation (σ). The log-likelihood value is obtained by adding up the possibilities in the first cluster (L_1), and the likelihood in the second cluster (L_2). If the new log-likelihood (L_{new}) value is greater than the old log-likelihood (L_{old}), the cluster point is changed to the new position. The mean and standard deviation values were obtained using Eqs. (21.1) and (21.2). While the likelihood value in cluster *i* is calculated using Eq. (21.3) [16].

$$\mu_i = \frac{1}{n_i} \sum_{x \in X_i} x \tag{21.1}$$

$$\sigma_i = \frac{1}{n_i} \sum_{x \in X_i} x(x - \mu_i)(x - \mu_i)^{\mathrm{T}} \tag{21.2}$$

$$L_i = -\frac{1}{2} n_i d - \frac{n_i d}{2} \log 2\pi - \frac{n_i}{2} \log|\sigma_i| + n_i \log \frac{n_i}{n} \tag{21.3}$$

Shifting the cluster center point requires calculations to find a better point. Calculation of new likelihood (L*) is obtained by adding up the value of the old likelihood (L) with changes in likelihood (ΔL), and constants (C), and mathematically can be written as in Eq. (21.4) [17]. The value of the change in likelihood itself is obtained using Eq. (21.5), and constants (C) are obtained using Eq. (21.6).

$$L_j^* = L_j + \left(\Delta L_j + C\right) \tag{21.4}$$

$$\begin{aligned} \Delta L_j = & -\frac{1}{2} \log|\sigma j| - \frac{n_j + 1}{2} \log\left(1 + \frac{1}{n_j + 1}\left(x - \mu_j\right)^{\mathrm{T}} \sigma_j - 1\left(x - \mu_j\right)\right) \\ & + \log \frac{n_j}{n} + \left(n_j + 1\right)\left(\frac{d}{2} + 1\right) \log \frac{\left(n_j + 1\right)}{n_j} \end{aligned} \tag{21.5}$$

$$C = -\frac{d}{2} - \frac{d}{2} \log 2\pi \tag{21.6}$$

21.3 Silhouette

Silhouette is one method of system evaluation. Silhouette is used to evaluate the results of data clustering. Silhouette evaluation is obtained by calculating the distance of each data to the center of the cluster. Each data has calculated the distance at each cluster center. The distance of the two points can be calculated using the Euclidean distance function [18].

The evaluation value of the silhouette aims to see how close the data is to the center of the cluster itself and the center of the other clusters. The results of the silhouette have intervals of -1 to 1. A value of -1 indicates poor clustering results and a value of 1 indicates clustering has the right results. To get the silhouette value, each cluster is calculated using Eqs. (21.7) and (21.8). The value of $a(i)$ indicates the distance of data $i(X_i)$ to other data (X_p)that are in one cluster (C_h). The value of $b(i)$ is the distance of data $i(X_i)$ to other data (X_p) in the other cluster [19].

$$\mathrm{sil}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{21.7}$$

$$\mathrm{Sil} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{sil}(i) \tag{21.8}$$

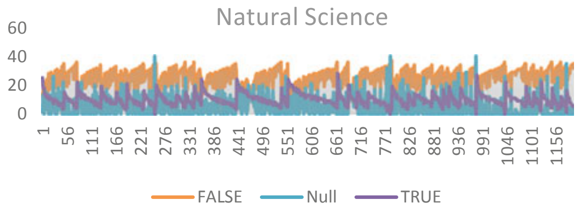
21.4 Results and Discussion

This research is quantitative research. This research is quantitative because it uses numerical data. The intended statistical data is KSM score data as material for analyzing research results using KSM participant data in 2019. The level of competition used is the junior high-level equivalent. The junior high school level has three types of competition. The first competition is the natural sciences subject. This competition has a total data of 1192 participants. Data on science competitions are shown in Fig. 21.1. The second competition is a competition in social sciences subjects. The SMP-level competition in social studies subjects has a total number of 1183 participants. In the third competition, there were 1254 participants with mathematics subjects. The participant data used is the provincial-level participants from the district/city-level selection results. Each data has three parameters. The parameters used include the number of questions answered correctly, answered incorrectly, and not answered. These parameters are used as input values to determine the quality of Indonesian students. The amount of data used comes from 34 provinces. Figure 21.1 presents natural science subjects data for number of questions answered correctly (TRUE), number of questions answered incorrectly (FALSE), and number of not answered (NULL) questions.

Looking at the data shown in Fig. 21.1, participants in science competitions have a higher tendency to answer false questions, followed by the number of questions answered correctly, and questions that were not answered. As with science participants, IPS participants have the same tendency. The difference between science participants is the number of questions answered incorrectly and the number of correct and blank questions. All three parameters have a high enough difference in value. In the mathematical participant data, the score parameters are not answered and the questions answered incorrectly have an almost equal number, while the number of questions answered correctly is relatively small. Data from the three competitions in clustering uses the SIML method which the number of questions answered correctly, the number of questions answered incorrectly, and the number of questions not answered as input data.

SIML clustering requires the determination of some initial parameters before clustering data. The parameters required are the determination of maximum iteration, determination of many classes, and determination of the initial cluster center. This study uses a maximum iteration of 1000 iterations, while many classes in this study were given several experiments. The trial amount is given starting from 3 classes, 4

Fig. 21.1 Natural science participant



classes, 5 classes, and 6 classes. Experiments for the number of classes are intended to see the optimal number of classes in KSM data clustering. The update value is also applied to the cluster center point. The clustering process stops if the L value does not change.

Clustering results are then evaluated using the silhouette method. The output value in Eq. (21.7) is the silhouette for each cluster. To see the overall silhouette value can be obtained using Eq. (21.8). As a reference to see the optimal number of classes, see the overall silhouette (S) silhouette value, and silhouette standard deviation. S value is used to examine the average success of the system in clustering data. While the silhouette standard deviation is to see how high the distribution of the silhouette is. The higher the S value the stronger the clustering results. If you have a large standard deviation silhouette value, it indicates that there are high-value or low-value data far from the average of other data.

The results of clustering experiments using several classes are shown in Table 21.1. Looking at the results of the experiments, it can be seen that clustering using 3 clusters has the highest silhouette value compared to silhouettes in the number of other clusters. In addition, clustering with 3 clusters has a relatively small silhouette value distribution. The distribution of silhouette values is indicated by the standard

Table 21.1 Silhouette values in several experiments

Number of clusters	Subjects	Silhouette	Silhouette standard deviation
3 cluster	Natural sciences	0.423745	0.231354
	Social sciences	0.415654	0.188862
	Mathematics	0.487071	0.202107
4 cluster	Natural sciences	0.450291	0.205557
	Social sciences	0.429888	0.195371
	Mathematics	0.401584	0.226638
5 cluster	Natural sciences	0.409574	0.193355
	Social sciences	0.419559	0.196411
	Mathematics	0.375758	0.200149
6 cluster	Natural sciences	0.427796	0.201309
	Social sciences	0.386807	0.239797
	Mathematics	0.365801	0.215890

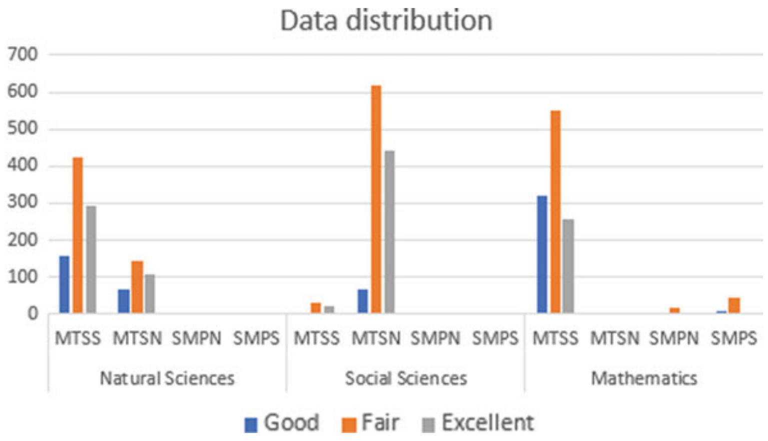


Fig. 21.2 Data distribution from clustering result

deviation of the silhouette values. The standard deviation value in the experiment 3 cluster has a small value.

The results of silhouette calculations in the 3 cluster experiment show that a lot of data has been classified well. Positive silhouette values show a good clustering. The number of data that has a value of less than 0 is 48 data from 1192 data. Overlapping data is 34 data from the 3rd cluster, 2 data from the 2nd cluster, and 12 data from the 1st cluster. In the social studies, subjects have 7 data with a value of less than 0 (overlapping) consisting of 3 data in cluster 1, 3 data from cluster 2, and 1 data from cluster 3. In the data competition, mathematics has overlapping data only 19 data. The overlapping data comes from 19 data in cluster 1, and 1 data from cluster 3. The results of clustering are obtained cluster groups based on the similarity of the data. Clustering results obtained as cluster 1 shows the cluster moderately, while cluster 2 is a cluster that shows data with minimal value. The last cluster is cluster 3 which states a group of data with high values. The distribution of student quality can be seen using the results of clustering. The distribution of clustering results is presented in Fig. 21.2. The distribution of students is divided into 3 categories, namely excellent, good, and fair. In the natural sciences comparison, the number of fair groups is the highest in private MTS, while in social science, the data is in public MTS. In mathematics competitions, the fair group on private MTS data is more dominant.

21.5 Conclusion

Clustering using the SIML method in KSM data has a maximum number of clusters of 3. The first cluster is good student data, followed by the second cluster of fair students, and the final cluster which shows excellent student data. The results of clustering using 3 clusters have the highest silhouette value compared to the silhouette values in the number of other clusters. Judging from the Kaufman table, the clustering value shows that the cluster structure formed is still weak. The silhouette value with cluster number 3 is 0.423745 for natural science, 0.415654 for social science, and 0.487071 for mathematics subjects.

References

1. Sudarsana, I.K.: Pemikiran tokoh pendidikan dalam buku lifelong learning: policies, practices, and programs (Perspektif Peningkatan Mutu Pendidikan di Indonesia). J. Penjaminan Mutu. **2**(2), 44–53 (2016)
2. Raharjo, S.B.: Evaluasi trend kualitas pendidikan di indonesia. J. Penelit. dan Eval. Pendidik. **16**(2), 511–532 (2012)
3. Dapodikbud, T.: Temukan Informasi Sekolah di seluruh Indonesia [Internet]. (2019). Available from: http://sekolah.data.kemdikbud.go.id/
4. Kemendikbud. Jumlah Sekolah Menurut Jenjang Tahun Ajaran 2017/2018 [Internet]. Databooks (2018). Available from: https://databoks.katadata.co.id/datapublish/2019/06/23/berapa-jumlah-sekolah-di-indonesia
5. Zakwandi, R.: Analisis Konsep Pesawat Sederhana Pada Pembelajaran Ilmu Pengetahuan Alam Berbasis Tradisi Sains Islam Di Madrasah Tsanawiyah. BELAJEA J. Pendidik. Islam. **2**(1), 21–34 (2017)
6. Luthfi, E.T.: Fuzzy C-Means untuk Clustering Data (studi kasus: data performance mengajar dosen). In: Seminar Nasional Teknologi, pp. 1–7 (2007)
7. Sharma, A., Shigemizu, D., Boroevich, K.A., et al.: Stepwise iterative maximum likelihood clustering approach. BMC Bioinform. **17**(1), 319 (2016)
8. Sharma, A., Lopez, Y., Tsunoda, T.: Divisive hierarchical maximum likelihood clustering. BMC Bioinform. **18**(16), 546 (2017)
9. Shovon, M.H.I., Haque, M.: Prediction of student academic performance by an application of k-means clustering algorithm. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(7) (2012)
10. Khalil, M., Ebner, M.: Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. J. Comput. High. Educ. **29**(1), 114–132 (2017)
11. Varela, N., Montero, E.S., Vasquez, C., et al.: Student performance assessment using clustering techniques. In: International Conference on Data Mining and Big Data, pp. 179–188. Springer, Berlin (2019)
12. Luna-Romera, J.M., del Mar Martinez-Ballesteros, M., Garcia-Gutierrez, J., Riquelme-Santos, J.C.: An approach to silhouette and dunn clustering indices applied to big data in spark. In: Conference of the Spanish Association for Artificial Intelligence, pp. 160–169. Springer, Berlin (2016)
13. Ding, M., Hu, K.: Susceptibility mapping of landslides in Beichuan County using cluster and MLC methods. Nat. Hazards **70**(1), 755–766 (2014)
14. Shivakumar, B.R., Rajashekararadhya, S.V.: Investigation on land cover mapping capability of maximum likelihood classifier: a case study on North Canara, India. Procedia Comput. Sci. **143**, 579–586 (2018)
15. Wang, Y., Jamshidi, M., Neville, P., Bales, C., Morain, S.: Multispectral Landsat image classification using a data clustering algorithm. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826), pp. 4380–4384. IEEE (2004)
16. Xu, J., Gangnon, R.E.: Stepwise and stagewise approaches for spatial cluster detection. Spat. Spatiotemporal. Epidemiol. **17**, 59–74 (2016)
17. Lozano, S., Calzada-Infante, L.: Computing gradient-based stepwise benchmarking paths. Omega **81**, 195–207 (2018)

18. Sharma, D., Thulasiraman, K., Wu, D., Jiang, J.N.: Power network equivalents: a network science based K-means clustering method integrated with silhouette analysis. In: International Conference on Complex Networks and their Applications, pp. 78–89. Springer, Berlin (2017)
19. Swindiarto, V.T.P., Sarno, R., Novitasari, D.C.R.: Integration of fuzzy C-means clustering and TOPSIS (FCM-TOPSIS) with silhouette analysis for multi criteria parameter data. In: 2018 International Seminar on Application for Technology of Information and Communication, pp. 463–468. IEEE (2018)