







# Chapter 20

## Graph Degree Linkage Clustering for Identify Student's Performance on *Kompetisi Sains Madrasah* in Indonesia



Ahmad Hanif Asyhar , A. Umar , Dian Candra Rini Novitasari ,  
Kusaeri , Ahmad Fauzi , Nurissaidah Ulinnuha , Dwi Rolliawati ,  
Noor Wahyudi , Ahmad Yusuf , Ali Mustofa , and Zakiyatul Ulya 

**Abstract** Graph degree linkage (GDL) algorithm is a development of agglomerative clustering and graph. The clustering algorithm can be applied to various problems, such as student performance. *Kompetisi Sains Madrasah* (KSM) is one of the competitions by integrating science and Islam held in Indonesia. The final score of the competition will be used in this study to identify the students' performance who participate in the competition. The main goal of the study is obtaining the results of the participant's performance clusters based on the final score of KSM and obtain the information about the quality of students in schools that are participating in the competition. Based on the research, we obtain the best  $K$ -neighborhood value of three clusters is equal to 25. The research obtains the silhouette coefficient value for clustering evaluation. They are 0.5104, 0.4838, 0.6853, 0.5943, 0.6605, 0.8037, 0.6455, 0.6723, 0.6996, 0.5767, and 0.6695 in mathematics, natural science, mathematics, natural science, social science, mathematics, biology, physic, chemistry, economics, and geography subjects. The identification of school student performance in each subject tested shows that State Islamic School student performance in KSM has the best performance in elementary and middle-high grade. In senior grade, the best students' performance in KSM is from Private Islamic School.

---

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

A. Umar  
Kementerian Agama RI, Lapangan Banteng 3-4, Jakarta, Indonesia

© The Editor(s) (if applicable) and The Author(s), under exclusive license  
to Springer Nature Singapore Pte Ltd. 2021

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\\_20](https://doi.org/10.1007/978-981-15-5224-3_20)

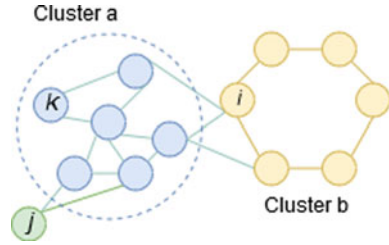
## 20.1 Introduction

Clustering is one of the analysis methods in data mining to grouping data sets based on similar characteristics of data. Two common techniques used in clustering algorithms are agglomerative hierarchical and partial clusters [1]. Agglomerative clustering also known as linkage method has several types, they are complete linkage, average linkage, etc. [2]. Agglomerative clustering algorithm has some disadvantages to solving high-dimensional space clustering problem because the neighborhood of the characteristics of the data is calculated based on pairwise distances between the data [3]. Therefore, Zhang et al. [4] proposes an agglomerative clustering algorithm based on the graph and it is called as graph degree linkage (GDL) algorithm. The algorithm proposed based on several studies on graph data representations that have been developed widely in various machine learning topics [5–9], but it rarely applied in agglomerative clustering. Furthermore, Zhang et al. [4] made improvements from [10–12] who did research before and tried to use agglomerative clustering on graph data representations. The algorithm has three main advantages. First, the algorithm has good performance even though it was applied to noisy and multiscale data. Second, the algorithm is a simple algorithm to implement. Then, last, the calculation time of the algorithm is very fast. In the algorithm, pairwise distances between sample data are used to construct  $K$ -nearest neighbors ( $K$ -NN) graphs, where indegree and outdegree become the similar characteristic of two clusters [4].

Clustering can be applied to solve various types of problems, such as data analysis, pattern recognition, image processing, market research, and student performance [13, 14]. The application of clustering algorithms to student performance problems is used to monitor the quality of education in a country [15]. The results of the monitoring can be used as evaluations to further improve the learning process [16], especially for schools and educational institutions. Student performance evaluations between schools are not the same as others, so the wide evaluation is needed. One of the methods to determine students' performance is school competition. One of the competitions held in Indonesia to appraise student performance is *Kompetisi Sains Madrasah* (KSM). The competition measures students' knowledge at various grades by integrating science and Islam. The final score of the competition will be used in this study to identify the students' performance who participate in the competition.

The main goal of the study is to obtain the results of the participant's performance clusters based on the final score of KSM and obtain the information about the quality of students in schools that are participating in the competition. In addition, the result of clustering the students' performance can be used as an evaluation of the committee and it can be used to plan for future activities for the better.

**Fig. 20.1** Illustration of indegree and outdegree of graph degree linkage [4]



## 20.2 Graph Degree Linkage

### 20.2.1 Algorithm Overview

Graph degree linkage (GDL) algorithm is part of the agglomerative clustering, and the algorithm starts by initializing several small clusters. Then, in every two clusters which have the similarity, they are chosen to be merged. To determine the neighborhood among data, it was calculated based on the indegree and outdegree of the vertex using  $K$ -NN graphs. The small clusters initialization are simply constructed as a weakly connected component of the  $K^0$ -NN graph, where the neighborhood size of  $K^0$  is small [4] (Fig. 20.1).

### 20.2.2 Neighborhood Graph

In graph theory, a vertex called adjacent to the vertex  $v$  in a graph if the vertex is connected to  $v$  with an edge. The neighborhood of vertex  $v$  in graph  $G$  is subgraphs of  $G$  consist of vertices adjacent to  $v$  with each edge connected to  $v$  [17].

Given a set of data  $X = \{x_1, x_2, x_3, \dots, x_n\}$  and directed graph  $G = (V, E)$ , where  $V$  is a set of vertices that correspond to the sample in  $X$ , and  $E$  is a set of vertices that connecting vertices. In the algorithm, the graph is connected with adjacency weight matrix  $W = [w_{ij}]$ , where  $w_{ij}$  is the weight of the edge at the vertex  $i$  to vertex  $j$ . In the problem of high-dimensional space,  $K$ -NN graphs are used to determine the neighborhood, where the weight of the edge is defined as follows [4].

$$w_{ij} = \begin{cases} \exp\left(\frac{-\text{distance}(i,j)^2}{\frac{a}{nk} \left[ \sum_i^n \sum_{j \in N_i^K} \text{distance}(i,j)^2 \right]}\right) & \text{if } x_j \in N_i^K \\ 0 & \text{otherwise} \end{cases} \quad (20.1)$$

where  $K$  and  $a$  are the parameters that free to be set, with  $\text{distance}(i, j)$  are distances between  $x_i$  and  $x_j$ , and  $N_i^K$  is a set of  $K$ -NN from  $x_i$ .

### 20.2.3 Adjacency Measure

The important element of the agglomerative clustering algorithm is the adjacency measure between two clusters. In [4], the adjacency measure is calculated based on indegree and outdegree of the graph. Indegree measures the density near the sample  $i$  and outdegree characterizes similarity of  $K$ -NN from vertex  $i$  to cluster  $C$ . Average indegree is defined in Eq. (20.2) and average outdegree is defined in Eq. (20.3), where  $|C|$  is the cardinality of the set  $C$ .

$$\text{deg}_i^-(C) = \frac{1}{|C|} \sum_{j \in C} w_{ij} \quad (20.2)$$

$$\text{deg}_i^+(C) = \frac{1}{|C|} \sum_{j \in C} w_{ij} \quad (20.3)$$

Furthermore, adjacency is defined as the product of the average indegree and average outdegree in Eq. (20.4). So, the adjacency of cluster  $C_b$  to cluster  $C_a$  by summing all vertices in  $C_b$ .

$$\mathcal{A}_{i \rightarrow C} = \text{deg}_i^-(C) \text{deg}_i^+(C) \quad (20.4)$$

$$\mathcal{A}_{C_b \rightarrow C_a} = \sum_{i \in C_b} \mathcal{A}_{i \rightarrow C_a} = \sum_{i \in C_b} \text{deg}_i^-(C_a) \text{deg}_i^+(C_a) \quad (20.5)$$

$$\mathcal{A}_{C_a, C_b} = \mathcal{A}_{C_b \rightarrow C_a} + \mathcal{A}_{C_a \rightarrow C_b} \quad (20.6)$$

The following is an algorithm of Graph Degree Linkage (GDL).

1. Initialize a set of  $n$  small cluster data  $X = \{x_1, x_2, x_3, \dots, x_n\}$  with  $n_T$  as the target number of clusters
2. Build the  $K$ -NN graph by initializing the number of clusters and  $K$ -neighborhood value, and get the weighted adjacency matrix  $W$ . A set of initial clusters define as  $V^c = \{C_1, \dots, C_n\}$ , where  $n_c$  is the number of clusters
3. Search two clusters  $C_a$  and  $C_b$  to be merged if has a similar character such  $\{C_a, C_b\} = \text{argmax}_{c_a, c_b \in V^c} \mathcal{A}_{c_a, c_b}$ . Then, update  $V^c$  and  $n_c$
4. Do step 3 repeatedly until  $n_c \leq n_T$ .

## 20.3 Clustering Evaluation: Silhouette

Silhouette is one method to evaluate the system of clustering. The value of the silhouette coefficient ( $sc$ ) can show the quality of a cluster system based on how the algorithm placed the objects on a cluster. The method is used to validate data, a single cluster, and the whole cluster [18]. In [19], the calculation of  $sc$  value starts

by calculating the average distance of an object  $i$  with each object in one cluster and calculates the average distance from object  $i$  with all objects in other clusters to obtain the smallest value. Then, calculate  $s(i)$  value of each data, and the average of the result  $s(i)$  value is the value of  $sc$ . The results of the calculation of  $sc$  value are in the range  $-1$  to  $1$ . Cluster results are called as good if the value of  $sc$  is positive. If  $s(i) = 1$  means that the object  $i$  is already in the right cluster. If  $s(i) = 0$  means that the object  $i$  is between two clusters. However, If the value of  $s(i) = -1$  means that the result of the cluster structure is overlapping. The criteria of  $sc$  value by Kaufman table [20].

## 20.4 Result and Discussion

The dataset processed in this study is the final score of KSM. The competition participated by students of public schools, Islamic schools, private schools, and state schools in various grade, they are Madrasah Ibtidaiyah (MI) or Islamic Elementary School, Madrasah Tsanawiyah (MTs) or Islamic Junior High School, and Madrasah Aliyah (MA) or Islamic Senior High School. Based on 540 data, it consists of scoring on multiple-choice questions and exploration of 11 subjects tested.

Based on the proposed algorithm, the first step in clustering is to determine the number of clusters and the value of  $K$ -neighborhood to build the  $K$ -NN Graph. In this study, the number of clusters used was three clusters, where each cluster represented the knowledge of the students who participated in the KSM. Clusters are identified as excellent, good, and fair clusters. Then, we used various  $K$ -neighborhood values of 10, 15, 20, and 25. It aims to find the best cluster that can represent the data. The results of each clustering experiment are evaluated using the silhouette shown in Table 20.1.

Based on the results of the evaluation of the clustering experiment, the best clustering system is using the value of  $K$ -neighborhood equal to 25. Illustration of the distribution of clustering results is shown in Fig. 20.2.

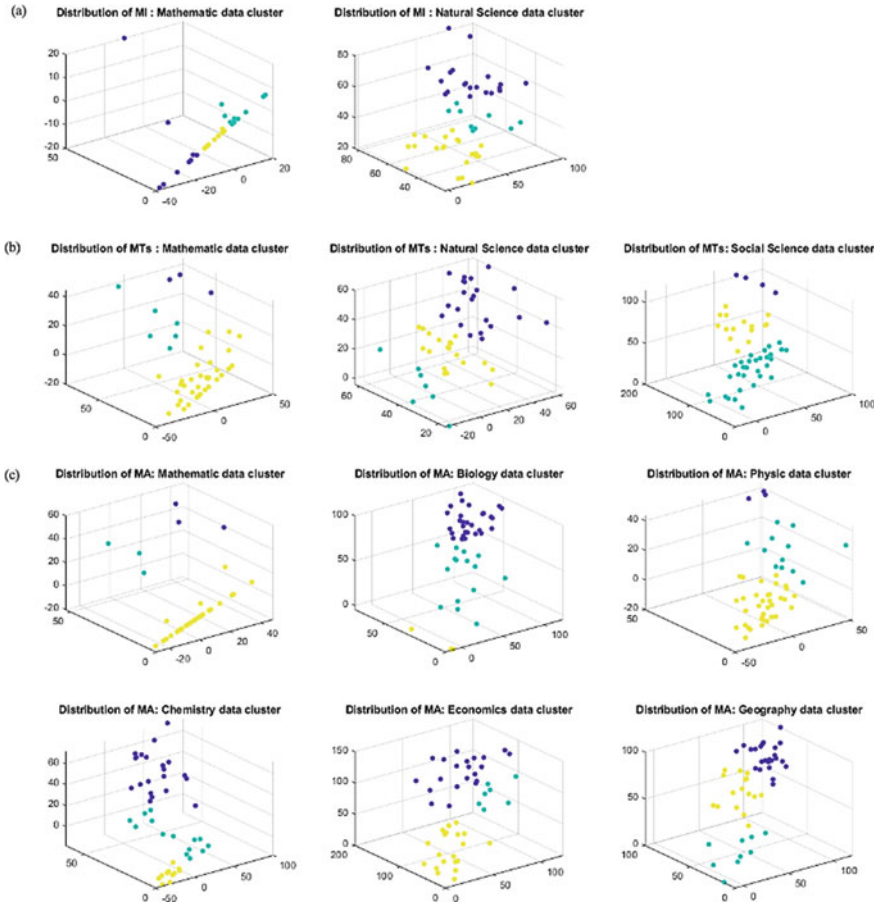
Besides obtaining the results of KSM participant's performance clustering, we identified the schools of each participant. Based on the information from the school and the results of the clustering, we identify the school students' performance is an excellent, good, and fair cluster in each subject tested in the competition.

Based on the graphs of elementary school student performance in Fig. 20.3, in natural science subjects, State Islamic Elementary School has the best student performance in the excellent cluster, then followed by private Islamic Elementary School and Private Elementary School. In mathematics subjects, private Islamic Elementary School is superior in student's performance. However, the number of students from private Islamic Elementary School is not few in the fair cluster.

Based on Fig. 20.4, it represents student performance at the middle grade, and the average number of State Islamic Junior High School student at Mathematics and Social Sciences is in the "Good" cluster, but in natural science subjects, they are in "Excellent" cluster. In participant's performance of State Islamic Junior High

**Table 20.1** Results of clustering evaluation using Silhouette

Education grade	Subjects	Number of clusters = 3											
		K = 10			K = 15			K = 20			K = 25		
		Mean	Varian	Mean	Varian	Mean	Varian	Mean	Varian	Mean	Varian		
MI	Mathematic	0.4257	0.2343	0.2396	0.2980	0.5104	0.2702	0.5104	0.2702	0.5104	0.2702		
	Natural Science	0.5957	0.0617	0.4838	0.0571	0.4838	0.0571	0.4838	0.0571	0.4838	0.0571		
	Mathematic	0.2972	0.2836	0.5926	0.1010	0.5926	0.1010	0.6853	0.1010	0.6853	0.0735		
MTs	Natural Science	0.4366	0.1514	0.4366	0.1514	0.6057	0.1034	0.5943	0.1071	0.5943	0.1071		
	Social Science	0.6402	0.0721	0.6402	0.0721	0.6402	0.0721	0.6605	0.0445	0.6605	0.0445		
	Mathematic	0.6502	0.0503	0.5308	0.1936	0.8037	0.0359	0.8037	0.0359	0.8037	0.0359		
MA	Biology	0.4386	0.2668	0.6400	0.0679	0.5414	0.1574	0.6455	0.1476	0.6455	0.1476		
	Physic	0.5667	0.1373	0.6595	0.0569	0.6595	0.0569	0.6723	0.1084	0.6723	0.1084		
	Chemistry	0.5899	0.1811	0.6144	0.1914	0.6313	0.0719	0.6996	0.0639	0.6996	0.0639		
	Economic	0.6203	0.1299	0.6419	0.0934	0.5767	0.1026	0.5767	0.1026	0.5767	0.1026		
	Geography	0.5769	0.1443	0.6146	0.0914	0.6146	0.0914	0.6695	0.0722	0.6695	0.0722		



**Fig. 20.2** a Illustration of the distribution of KSM participant performance in Madrasah Ibtidaiyah; b illustration of the distribution of KSM participant performance in Madrasah Tsanawiyah; and c illustration of the distribution of KSM participant performance in Madrasah Aliyah

School, the most participants are clustered in “Good” cluster at Mathematics and Social Science subjects, the Natural Science subjects are clustered in the “Fair” cluster (Fig. 20.5).

In the identification of Senior High School student’s performance, based on all subjects tested in KSM (except mathematics), State Islamic Senior High School student performance is superior to the other schools. At mathematics subject, the best student performance is Private Senior High School students. Private Islamic Senior High School student’s performance is clustered in “Good” cluster at all subjects tested. Meanwhile, State Senior High School student performance cannot be identified as well because there are not many students who netted in the national competition.

### Elementary Student's Performance

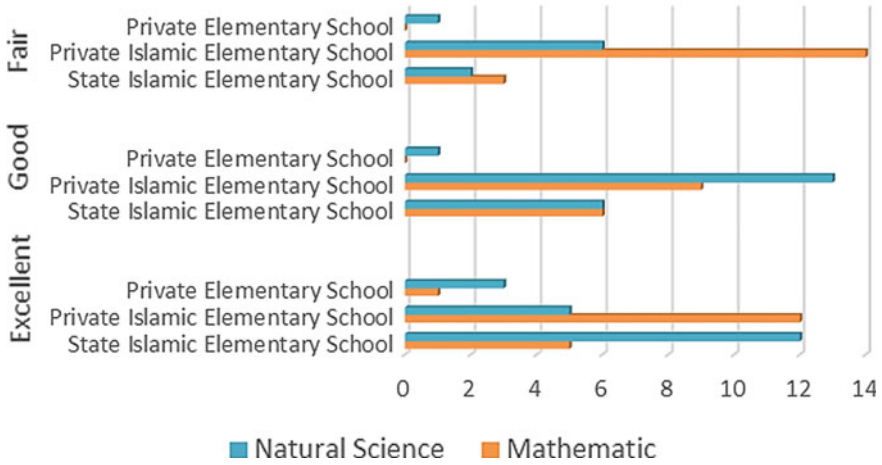


Fig. 20.3 Elementary School student's performance graph

### Junior High School Student's Performance

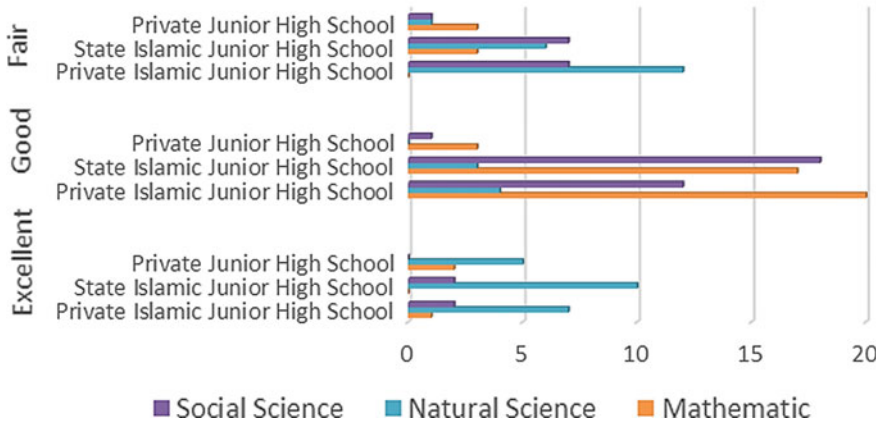


Fig. 20.4 Junior High School student's performance graph

## 20.5 Conclusion

The identification of students' performance who participates in KSM at the national stage can be used to monitor and evaluate the quality of education in Indonesia. The result of study obtained the best *K*-neighborhood value of three clusters which is equal to 25 with silhouette coefficient values that are 0.5104, 0.4838, 0.6853, 0.5943,



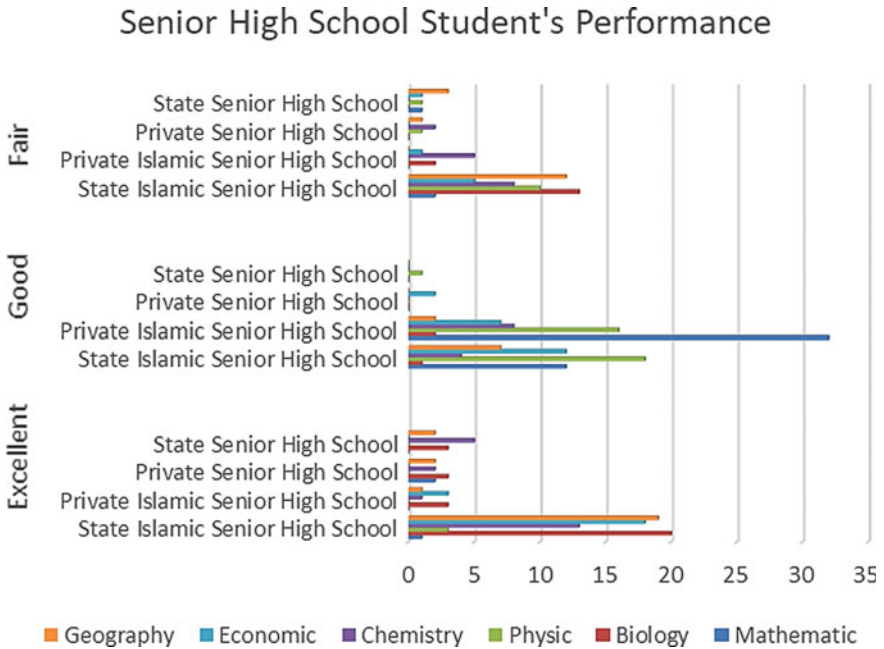


Fig. 20.5 Senior High School student’s performance graph

0.6605, 0.8037, 0.6455, 0.6723, 0.6996, 0.5767, and 0.6695 in all subjects. Based on the clustering result, the student performance generally clustered in excellent cluster in elementary and middle grade is from State Islamic School. In senior high school grades, the student’s performance that clustered in “excellent” cluster generally is from Private Islamic School.

## References

1. Omran, M.G.H., Engelbrecht, A.P., Salman, A.: An overview of clustering methods. *Intell. Data Anal.* **11**(6), 583–605 (2007)
2. Alfiani, A.P., Wulandari, F.A.: Mapping student’s performance based on data mining approach (a case study). *Agric. Agric. Sci. Procedia.* **3**, 173–177 (2015)
3. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* **27**(2), 83–85 (2005)
4. Zhang, W., Wang, X., Zhao, D., Tang, X.: Graph degree linkage: agglomerative clustering on a directed graph. In: *European Conference on Computer Vision*, pp. 428–441. Springer (2012)
5. Shi, J., Malik, J.: Normalized cuts and image segmentation. *Dep. Pap.* 107 (2000)
6. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: *Advances in Neural Information Processing Systems*, pp. 849–856 (2002)
7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003)

8. Grady, L., Schwartz, E.L.: Isoperimetric graph partitioning for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(3), 469–475 (2006)
9. Zhang, W., Lin, Z., Tang, X.: Learning semi-Riemannian metrics for semisupervised feature extraction. *IEEE Trans. Knowl. Data Eng.* **23**(4), 600–611 (2010)
10. Karypis, G., Han, E.-H.S., Kumar, V.: Chameleon: hierarchical clustering using dynamic modeling. *Computer (Long. Beach. Calif)* **8**, 68–75 (1999)
11. Zhao, D., Tang, X.: Cyclizing clusters via zeta function of a graph. In: *Advances in Neural Information Processing Systems*, pp. 1953–1960 (2009)
12. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **59**(2), 167–181 (2004)
13. Nagesh, A.S., Satyamurty, C.V.S.: Application of clustering algorithm for analysis of student academic performance. *Int. J. Comput. Sci. Eng.* **6**(January), 381–384 (2018)
14. Nagesh, A.S., Satyamurty, C.V.S., Akhila, K.: Predicting student performance using KNN classification in bigdata environment. *CVR J. Sci. Technol.* **13**, 83–87 (2017)
15. Shahiri, A.M., Husain, W.: A review on predicting student's performance using data mining techniques. *Procedia Comput. Sci.* **72**, 414–422 (2015)
16. Jayabal, Y., Ramanathan, C.: Clustering students based on student's performance-a partial least squares path modeling (PLS-PM) study. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 393–407. Springer (2014)
17. Fronček, D.: Locally linear graphs. *Math. Slovaca.* **39**(1), 3–6 (1989)
18. Prasetyo, E.: *Data Mining Mengolah Data Menjadi Informasi Menggunakan MATLAB*. Andi, Yogyakarta
19. Anggara, M., Sujaini, H., Nasution, H.: Pemilihan distance measure Pada K-Means clustering Untuk Pengelompokkan member di Alvaro fitness. *J. Sist. dan Teknol. Inf.* **4**(1), 186–191 (2016)
20. Swindiarso, V.T.P., Sarno, R., Novitasari, D.C.R.: Integration of Fuzzy C-means clustering and TOPSIS (FCM-TOPSIS) with Silhouette analysis for multicriteria parameter data. In: *2018 International Seminar on Application for Technology of Information and Communication*, pp. 463–468. IEEE (2018)