Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.6 No.1 January 2021, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Optimizing K-Means Algorithm by Using Particle Swarm Optimization in Clustering for Students Learning Process

Rudi Hariyanto[1], Mohammad Zoqi Sarwani[2]

[1,2]*Informatics Engineering Department, Universitas Merdeka Pasuruan, Indonesia*

[1]rudihariy4nt0@gmail.com*

[2]zoqi.sarwani@unmerpas.ac.id

Abstract— In the implementation of learning, several factors affect the student learning process, including internal factors, external factors, and learning approach factors. For example, the physical and spiritual condition of students. Physiological aspects (body, eyes and ears and talents of students, student interests). External factors, for example, environmental conditions around students, family, teachers, community, friends) Thus, learning achievement is significant because educational institutions' success can be seen from how many students learning achievement. This research's first focus is to do student clustering based on their learning process using 11 parameters. Second, using the PSO algorithm to get maximum clustering results. The research data were obtained from vocational secondary education institutions in the city of Pasuruan. The data is obtained from the results of school reports and questionnaires as much as 100 student data. Data attributes include environmental features, social features, and related school features to group student data for learning data processing. From the classification results using the PSO method, the silhouette value is 0.97140754, very close. These results indicate that the PSO method can improve the K-Means clustering method's performance in the classification process of student learning interest.

Keywords— Learning process, Optimization Algorithm, PSO, K-Means, Clustering

I. INTRODUCTION

Today, education is one of the essential aspects in changing one's mindset. The quality of education can be observed from the obtained internal support (family) and external support (environment). Education is an effort, influence, protection, and assistance for children to let them be independent, or instead of helping children to be able to carry out their life tasks [1]. Educational institutions are considered to be successful as the learning process quality is given to the students. Therefore, several factors affect the learning process, including internal factors, external factors, and learning approach factors [2]. Internal factors (factors within students), for example, students' physical and spiritual condition. Namely: physiological aspects (body, eyes, and ears) and psychological characteristics (student intelligence, student attitudes, student talents, student interests, and student motivation). External factors (factors from students' outside), for example, students' environment. Namely: social environment (family, teachers, community, friends) and non-social environment (home, school, equipment, and nature).

Meanwhile, students' learning approach factors include students' strategies and methods to learn subjects, including a high approach, medium approach, and low approach. Meanwhile, data mining is a term to describe the information contained in a set of data. Data mining is the process of using static techniques, mathematics, artificial intelligence, and machine learning to separate and identify useful information and related knowledge from large databases [3].

K-Means is one of the clustering methods in data mining. The determination of K's value must be determined at the beginning of the study by considering each group's differences. Also, the parameter being chosen is the cluster center, which was randomly assigned. The better the centroid determination, the more precise and faster the grouping process will be. Since the centroid is determined randomly, the accuracy level is sometimes not valid and often appears local optima (local solution) [4].

Particle Swarm Optimization Algorithm (PSO method) can optimize the centroid value on K-Means with promising results[5], [6]. PSO can also optimize the centroid value by referring to the local optima in real numbers[6]. Some related studies have been carried out by [7][8] regarding the implementation of K-Means to classify students based on students' learning process [8][9].

The results of research with 100 students and 11 attributes as the participants, it was known that as many as 120 students were in a right learning achievement cluster, 104 students were in medium learning achievement cluster, and 125 students were in low learning achievement cluster with a silhouette score of 0.669253108828133. Thus, the current study's first focus was to classify students based on their learning process using 11 parameters. The second focus was using the PSO algorithm to get maximum clustering results.

II. RESEARCH METHODOLOGY

This study used (1) literature study, (2) Survey and Questionnaire, (3) Application Design, (4) Algorithm PSO and K-Means (5) Test result (6) Analysis of Test Results. With the existence of a research concept framework, it is hoped that it can clarify the research contribution that will be carried out, as shown in Figure 1.

The initial stage for conducting research was to conduct literature studies based on the research topics taken. The studied literature is related to data clusters using the K-Means method. The research data were gathered from secondary vocational schools in Pasuruan.

DOI : 10.25139/inform.v6i1.3459

The data was obtained from the results of school reports and questionnaires as much as 100 student data. The data attributes included environmental features, social features, and related school features.

An analysis is a method after literature study and data collection on cases in the study. This study's analysis started from input data used and needed to produce a grouping of each data [10.8]. The data inputted were nominal and numeric. Each value that exists was the data category. The following description is an overview of the data presented in the current study.

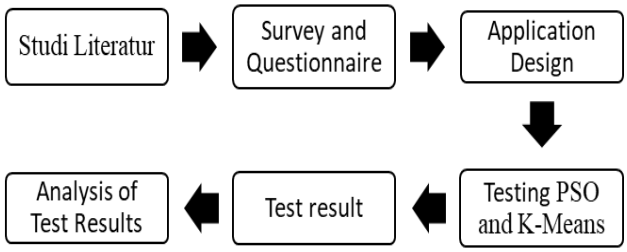

Figure 1 Flowchart of research work

A. *Particle Swarm Optimization (PSO)*

PSO is an optimization method that represents solutions to problems in the form of particles with stochastic properties. The workflow of the PSO can be explained as follows [14]:

- We are using particle best and global best available.
- Updating particles' position by adding the new velocity and the previous position with the following Equation (1).

$$v_i(t) = vw_i(t-1) + c_i r_i \left(x_{pi} - x_i\right) + c_2 r_2 \left(x_{gi} - x_i\right)$$
$$x_i(t) = x_i(t-1) + v_i(t) \quad (1)$$

where:
i = particle index
t = iteration
w = inertia
v_i = velocity of the *i-th* particle
x_i = position of the *i-th* particle
x_{gi} = best position of all particles (global best)
x_{pi} = best position of *i-th* particle (particle best)
$c_{1,2}$ = learning rate *r1,2* = random number [0.1]

- Evaluating the fitness of each particle.
- I was comparing and updating particle best and global best for each particle based on fitness.
- If the stop criteria were met, it would stop. Otherwise, go back to step 1.

B. *K-Means Algorithm*

The K-Means method was used to search for clustering data. It was starting by determining the number of clusters (*K*) and the initial centroid that was randomly selected. The centroid was the average of observations that were in a cluster. In cluster formation [10] suppose that a data matrix $\{X_{ij}\}$ was $n \times p$ where $i = 1, 2, \ldots, N,$ and $j = 1,2, \ldots, p$.

To know the data, the researcher did some steps, those are:

- Assume the initial cluster number *K*.
- Find the Ck centroid
- Calculate each object's distance to each centroid using the Euclidean distance, or it could be written as follows Equation (2).

$$d(x_i, c_i) = \sqrt{(x_i - c_i)^2} \quad (2)$$

- Arrange each object is arranged to the nearest centroid, and the collection of objects formed a cluster.
- Determine the new centroid of the newly formed cluster, where the new centroid was obtained from the average of each object located in the same cluster.
- Repeat step 3, if the initial and new centroids were not the same, to produce the best silhouette in its class.

III. Result and Discussion

In analyzing the students' interests and talents, it is necessary to consider several aspects: internal factors, external factors, social environment, and learning approach factors. From these factors, some information was obtained. Then a conclusion was drawn in determining students' interest in learning [8][11]. In this case, the researchers conducted student clustering based on the learning process using 11 parameters, namely: Family Condition (k1), Family Support (k2), Internet Access (k3), Mother's Education (k4), Father's Education (k5), Study Time (k6), Family Relations (k7), Silent Time (k8), Frequent Leave (k9), Health (k10), Absence (k11).

Table I
Clustering in Determining Student Interest in Learning

K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11
2	2	2	4	4	2	4	3	4	3	6
1	1	1	1	1	2	5	3	3	3	4
1	2	1	1	1	2	4	3	2	3	10
1	1	1	4	2	3	3	2	2	5	2
1	1	2	3	3	2	4	3	2	5	4
1	1	1	4	3	2	5	4	2	5	10
1	2	1	2	2	2	4	4	4	3	0
2	1	2	4	4	2	4	1	4	1	6
2	1	1	3	2	2	4	2	2	1	0
1	1	1	3	4	2	5	5	1	5	0
1	1	1	4	4	2	3	3	3	2	0
1	1	1	2	1	3	5	2	2	4	4
1	1	1	4	4	1	4	3	3	5	2
1	1	1	4	3	2	5	4	3	3	2
2	1	1	2	2	3	4	5	2	3	0

Description of the dataset in Table I:

- The condition of the family (1: Living together, 2: Divorced).
- Support from family (1: yes, 2: no).
- Internet access while studying (1: yes, 2: no).

DOI : 10.25139/inform.v6i1.3459

- Mother's last education (0: no school, 1: elementary school, 2: junior high school, 3: senior high school, 4: college).
- Father's last education (0: none, 1: elementary school, 2: junior high school, 3: senior high school, 4: college).
- Student duration (1: < 2 hours, 2: 2 hours - 5 hours, 3: 5 hours-10 hours, 4: > 10 hours).
- Students' relationships with their families (1: very bad, 2: bad, 3: fair, 4: good, 5: very good).
- The duration of silence after studying (1: very much, 2: a lot, 3: enough, 4: a little, 5: little).
- Chance to play with friends (1: very much, 2: a lot, 3: enough, 4: a little, 5: little).
- Students' health (1: very bad, 2: bad, 3: fair, 4: good, 5: very good).
- Student attendance at school (attendance 0-93).

The test was carried out after the input data analysis, and system design was carried out. The grouping process was carried out into 3 clusters (*K = 3*). The 3 clusters can produce the best Silhouette coefficient value. Based on the results of the cluster. The K variable random center points or centroids were determined. Measurement of each data distance to the center points was carried out using the Euclidean distance calculation. The smallest distance value for one centroid was obtained from the distance calculation results so that the data would be affiliated with the cluster data from the nearest cluster. After making sure that data 1 to the number of data was included in group one, then the new centroid's determination was based on each cluster's existing data. This process was repeated until the data was entered in 3 groupings, as in the K-means and PSO processes.

```
Cdistance jarak = new Cdistance(input,centroid);
    Ccluster kluster = new Ccluster(jarak);
    kluster.kluster();
    System.out.println("Hasil Kluster");
    // for(int i=0;i<kluster.getKluster().length;i++){
    //  for(int j=0;j<kluster.getKluster()[i].length;j++){
    //     if(kluster.getKluster()[i][j]>0)
    //             System.out.print(kluster.getKluster()[i][j]+" ");
    //  }
    //  System.out.println();
    // }
    ArrayList<ArrayList<Integer>> dataKluster;
    dataKluster = kluster.getKlusterDinamis();
    for(int i=0;i<dataKluster.size();i++){
       for(int j=0;j<dataKluster.get(i).size();j++){
          System.out.print((dataKluster.get(i).get(j)+1)+" ");
       }
       System.out.println();
    }
    for(int n=0;n<10;n++){
       System.out.println("iterasi "+(n+2));
       centroid.nextCentroid(kluster);
       System.out.println("Centroid");
       for(int i=0;i<centroid.getData().length;i++){
          for(int j=0;j<centroid.getData()[i].length;j++){
             System.out.print(centroid.getData()[i][j]+" ");
          }
          System.out.println();
       }
       kluster.kluster();
       dataKluster = kluster.getKlusterDinamis();
       System.out.println("Hasil Kluster");
       for(int i=0;i<dataKluster.size();i++){
          for(int j=0;j<dataKluster.get(i).size();j++){
             System.out.print((dataKluster.get(i).get(j)+1)+" ");
          }
          System.out.println();
       }
    }
```

Each process determined the centroid value based on this grouping; each result of the silhouette value determined the closest distance. In determining interest in learning, the K value would vary in determining each cluster's closest distance. Here are the results of the trial process in the application in Table II.

TABLE II
RESULTS OF PSO AND K-MEASN ALGORITHM IN STUDENTS 'LEARNING INTEREST PROCESS

The Number of K	Clustering	Silhouette Value
1	2	0.97140754
	6	0.267029056
	7	0.738606824
2	2	0.97140754
	6	0.639433666
	7	0.617561222
3	2	0.97140754
	6	0.594086774
	7	0.121294603
4	2	0.97140754
	6	0.582772673
	7	0.595558189
5	2	0.97140754
	6	0.594086774
	7	0.550932925
6	2	0.97140754
	6	0.267029056
	7	0.60066234
7	2	0.97140754
	6	0.267029056
	7	0.600113095
8	2	0.97140754
	6	0.594086774
	7	0.230460594
9	2	0.97140754
	6	0.609257434
	7	0.595558189
10	2	0.97140754
	6	0.594086774
	7	0.738606824
	2	0.97140754

DOI : 10.25139/inform.v6i1.3459

It was based on testing the system to find each variable's proximity value processed using K-Means and the Particle Swarm Optimization Algorithm. It optimized the centroid value on K-Means with good results, with some interactions, the Silhouette value was obtained in each of the interactions.

IV. CONCLUSION

From the results of the application trial, it can be concluded that for the classification of student interest in learning obtained from the results of school reports and questionnaires, as much as 100 student data. The external cluster is 0.97140754, and the Internal factor is 0.594086774, and the learning approach factor is 0.738606824. The silhouette value of that size is obtained because the distance between the data is very close. These results indicate that the PSO method can improve the k-means clustering method's performance in the classification process of student learning interest.

REFERENCES

[1] Langeveld, M.J. 1980. Exceptional Children: an Introductory. Survey of Special Education (Sixth Ed). Macmillan Publishing. New York.
[2] Syah Muhibbin,. 2006. Psikologi Belajar , Jakarta: PT. Raja Grapindo Persada.
[3] Efraim Turban, dkk. 2005. Decision Support Systems and. Intelligent Systems. Edisi 7, Jilid 1, New Jersey: Pearson Education.
[4] Y. A. Auliya, "Improve Hybrid Particle Swarm Optimization and K-Means by Random Injection for Land Clustering of Potato Plants," *Proc. - 2019 Int. Conf. Comput. Sci. Inf. Technol. Electr. Eng. ICOMITEE 2019*, vol. 4, no. 1, pp. 192–198, 2019, DOI: 10.1109/ICOMITEE.2019.8921207.
[5] M. Sarwani and D. Sani, "Implementasi Metode K-Means Sebagai Pengelompokan Siswa Berdasarkan Proses Belajar Siswa," pp. 1131–1135, 2018.
[6] J. Aranda and W. A. G. Natasya, "Penerapan Metode K-Means Cluster Analysis Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk Mahasiswa International Class Stmik Amikom Yogyakarta," *Semin. Nas. Teknol. Dan Multimed. 2016*, vol. 4, no. 1, pp. 4-2–1, 2016, [Online]. Available: https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/1293.
[7] F. Y. Bisilisin, Y. Herdiyeni, and B. I. B. P. Silalahi, "Optimasi K-Means Clustering Menggunakan Particle Swarm Optimization pada Sistem Identifikasi Tumbuhan Obat Berbasis Citra K-Means Clustering Optimization Using Particle Swarm Optimization on Image Based Medicinal Plant Identification System," *Ilmu Komput. Agri-Informatika*, vol. 3, no. 2002, pp. 38–47, 2014.
[8] Sujoto, T.S.Si., M.M.Kom. 2011. "Kecerdasan Buatan". Penerbit ANDI yogjakarta.
[9] A. Saidul and J. L. Buliali, "Implementasi Particle Swarm Optimization pada K-Means Untuk Clustering Data Automatic Dependent Surveillance-Broadcast," *Eksplora Inform.*, vol. 8, no. 1, p. 30, 2018, DOI: 10.30864/eksplora.v8i1.150.
[10] Rahmawati, dkk. 2019. Optimasi K-Means untuk Pengelompokan Data Kinerja Akademik Dosen menggunakan Particle Swarm Optimization, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 3, No. 4, April 2019, hlm. 4102-4110 http://j-ptiik.ub.ac.id
[11] Achyani, 2018. Penerapan Metode Particle Swarm Optimization Pada Optimasi Prediksi Pemasaran Langsung. JURNAL INFORMATIKA, Vol.5 No.1 2355-6579, E-ISSN: 2528-2247
[12] S. Kusumadewi, Membangun Jaringan Syaraf Tiruan Menggunakan MATLAB & EXCEL LINK. Yogyakarta: Graha Ilmu, 2004ISSN:
[13] Kusumadewi S, Hartati S, Harjoko A, Wardoyo R. 2006. Fuzzy Multi-Attribute Decision Making (FUZZY MADM). Graha Ilmu, Yogyakarta.
[14] Daihani, DU., 2001, Komputerisasi Pengambilan Keputusan, PT Elex Media Komputindo Gramedia, Jakarta.
[15] Kusrini, 2007, Konsep dan Aplikasi Sistem Pendukung Keputusan, Penerbit Andi, Yogyakarta.

DOI : 10.25139/inform.v6i1.3459