Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 12, No. 1, Juli 2023
ISSN 0216 – 0544
e-ISSN 2301– 6914

OPTIMIZATION OF K-MEANS CLUSTERING USING PARTICLE SWARM OPTIMIZATION ALGORITHM FOR GROUPING TRAVELER REVIEWS DATA ON TRIPADVISOR SITES

[a]I Made Satria Bimantara, [b]I Made Widiartha

[a,b] Informatics Department, Faculty of Math and Science, Udayana University, Bukit Jimbaran, Bali, Indonesia-803611, Telp: 081237787141
E-mail: satriabimantara.md@gmail.com, madewidiartha@unud.ac.id

Abstract

K-Means Algorithm can be used to group tourists based on reviews on tourist destination objects. This algorithm has a weakness that is sensitive to the determination of the initial centroid. The initial centroid that is determined at random will decreasing the level accuracy, often gets stuck at the local optimum, and gets a random solution. Optimization algorithms such as PSO can overcome this by determining the optimal initial centroid. The Elbow method is used to determined the optimal number of clusters (K) by calculating the sum squared error (SSE) value of the resulting clusters. The quality of the clusters produced by the K-Means Algorithm with and without PSO Algorithm is measured using the average Silhouette Coefficient (SC). Travel Reviews Data Set which consists of 980 records and 10 attributes is retrieved from online repository and used as secondary data in this study. The test results show that K=2 is the best number of clusters. The hybrid PSO and K-Means gives an average SC value of 0.300358 which is better than without the PSO Algorithm of 0.300076. The optimal PSO hyperparameter generated is the number of particles=30, φ_1*=2.2, and* φ_2*=3 at maximum iteration of 100.*

Key words: Clustering, K-Means, Optimization, Particle Swarm Optimization.

INTRODUCTION

The economic development of an area can be influenced by several factors, one of which is tourism. In several countries in the world, tourism is an important economic sector. Tourism is a sector that supports a source of foreign exchange for a country [1]. Tourism occupies the third position for foreign exchange earnings in Indonesia in 2009 [2].

Tourist visits to tourist destinations are important for the tourism sector. Increased tourist visits will increase income for the tourism sector and have implications for the development of a region's economy. Tourist visits to tourist destinations are influenced by several things, one of which is a review of the tourist attraction they will visit. Tourists will see the condition of the tourist destination object through its reviews before visiting it [1].

The rapid development of technology and communication makes it easier for tourists to see reviews of tourism destination objects. Tripadvisor application is one of them. Tourists can write their experiences at a particular tourist destination and see reviews of other tourist destinations in this application. Tripadvisor has an important role for tourists as their consideration before visiting tourism destinations [3].

Reviews about tourism destination objects that exist in applications such as Tripadvisor can be used to improve the quality of a tourist destination object and as a marketing medium for tourism service providers. One of them is by grouping tourists based on the reviews they give to tourist destinations. With this it can be seen the tendency of a tourist's preference for a particular tourist destination object, so that it can be used as a medium for promoting tourist destinations to tourists based on their favorite tourist destination objects.

Nurjanah and Arifin (2021) conducted a study to classify tourists based on reviews on tourist destinations using K-Means algorithm. Secondary data from online repository is retrieved and is used for this study. The data consist of 980 records and 10 attributes. The ten attributes are list of tourist destinations in East Asia, namely art galleries, juice bars, parks or picnic areas, restaurants, resorts, dance clubs, beaches, cinemas, museums, and religious institutions. Their research shows that K-Means algorithm can be used to group tourists based on reviews on tourist destinations into two clusters [1].

The K-Means algorithm has a weakness that is sensitive to the determination of the center of the cluster (centroid). Centroids that are determined randomly in K-Means Algorithm cause the level of accuracy to decrease and are often trapped at the local optimum [4]–[10]. If there is an error in the determination, the optimal solution or random results will not be obtained [11], [12]. Optimization algorithms can be used to solve the K-Means problem [7], [8]. Initial optimum centroids are found using an optimization algorithm and can be used in calculations of K-Means algorithm subsequent.

Particle Swarm Optimization algorithm (abbreviated PSO) is a method for solving optimization problem introduced by Kennedy and Eberhart in 1995 [13]. The search for optimal solutions in the PSO Algorithm is based on the social behavior of bird and fish populations in survival [14]. The PSO algorithm has been widely used as an optimization algorithm [15]–[17]. The PSO algorithm combined with the Naïve Bayes Algorithm provides higher accuracy than without the PSO Algorithm, such as research [18] on classification of hoax news on social media, research [19] on prediction of chronic kidney disease, and research [20] on text classification on e-government social media. PSO is also considered to solve scheduling problem with Ant Colony Optimization as a hybrid model [21].

The PSO algorithm has several advantages over other optimization algorithms. The PSO algorithm can solve minimum spanning tree problem better in terms of execution time complexity than Firefly algorithm (FA) [22]. The PSO algorithm is better in determining the position of the robot than the Genetic algorithm (GA) in terms of less computation and relatively fast time [14]. The Neural Network training process using the PSO Algorithm is better than GA in terms of faster computation time [23], [17]. Compared to other optimization algorithms, the PSO algorithm has better efficiency [24].

This study will classify tourists based on reviews on tourist destination objects using K-Means algorithm as done [1]. In this study, the PSO algorithm is proposed to optimize the initial centroid of K-Means algorithm before the next process is carried out. The data used in research [1] will be used in this study, but the data will go through a pre-processing stage before the data can be grouped.

MATERIAL AND METHODS

Research Flowchart

Research flowchart in this study is show in Figure 1.

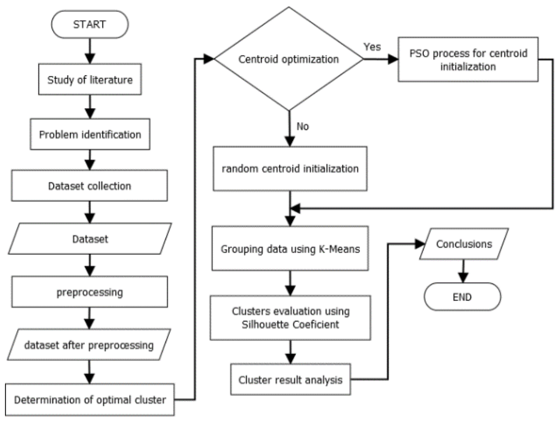

Fig 1. Research flowchart

Data Collection

Secondary data named Travel Reviews Data Set retrieved from online repository is used in this st. This data has been used in previous research by [1]. This data is sourced from the UCI Machine Learning Repository with a total of 980 records and 10 attributes with the name Travel Reviews Data Set. The ten attributes in the research data are of numeric type. The ten attributes are list of tourist destinations in East Asia, namely art galleries, juice bars, parks or picnic areas, restaurants, resorts, dance clubs, beaches, cinemas, museums, and religious institutions. The value for each attribute is the average value of user feedback on each attraction.

Preprocessing Data

Exploratory Data Analysis

The data exploration analysis stage is the first stage of the initial data processing. This stage is carried out to see whether or not there is a missing value in the data. There is no missing value from 980 records obtained after checking. At this stage, the calculation of the central tendency value for each attribute in the data is also carried out using a library from Python called Pandas. The calculated central tendency values include the mean value indicated by ‘mean’, the standard deviation indicated by ’std’, the quartile value 1 (Q1) indicated by ’25%’, the quartile value 2 (Q2) or median indicated by ‘50%’, the quartile value 3 (Q3) indicated by ‘75%’, ’min’ is indicating the minimum value indicated and ’max’ is indicating the maximum value. Figure 2 shows the results of calculating the central tendency of the data.

	count	mean	std	min	25%	50%	75%	max
Category 1	980.0	0.893194	0.326912	0.34	0.67	0.83	1.0200	3.22
Category 2	980.0	1.352612	0.478280	0.00	1.08	1.28	1.5600	3.64
Category 3	980.0	1.013306	0.788607	0.13	0.27	0.82	1.5725	3.62
Category 4	980.0	0.532500	0.279731	0.15	0.41	0.50	0.5800	3.44
Category 5	980.0	0.939735	0.437430	0.06	0.64	0.90	1.2000	3.30
Category 6	980.0	1.842898	0.539538	0.14	1.46	1.80	2.2000	3.76
Category 7	980.0	3.180939	0.007824	3.16	3.18	3.18	3.1800	3.21
Category 8	980.0	2.835061	0.137505	2.42	2.74	2.82	2.9100	3.39
Category 9	980.0	1.569439	0.364629	0.74	1.31	1.54	1.7600	3.17
Category 10	980.0	2.799224	0.321380	2.14	2.54	2.78	3.0400	3.66

Fig 2. The calculation results of the central tendency of each attribute in the data

Outliers Data Handling

Data Outlier data can be detected using the central tendency value. If a value is outside the lower and upper limit ranges, then the data is said to be outlier data. The lower and upper limits can be found using Equation (1) and Equation (2), respectively. The limit of three times the standard deviation of the average value of an attribute is the threshold that is usually used to identify outliers in data with normal distribution or resembling a normal distribution. Outlier data will be handled by discarding it and not included in the next data processing.

$$ba = \bar{A} + (3 \times \sigma_A) \quad (1)$$

$$bb = \bar{A} - (3 \times \sigma_A) \quad (2)$$

Determining the Optimal Number of Clusters

One of important parameter in K-Means is number of clusters to be choose. This parameter must be determined at the beginning before the grouping process using K-Means algorithm is carried out. The results of data grouping are influenced by the initial value of *K*, therefore it is important to determine the optimal *K* at the beginning.

There are several ways to determine the *K* optimal value. One of them uses Elbow method. With this method, the optimal *K* is determined based on the SSE value with a significant decrease and is angled. Determination of the *K* optimal value often uses the calculation of the SSE value [25]. SSE value calculated using Equation (3).

$$\mathrm{SSE} = \sum_{j=1}^{K} \sum_{x_i \epsilon C_j} \left\| x_i - P_j \right\|^2 \quad (3)$$

The steps of the Elbow method to determine the best value of *K* are as follows [26]:

1. Repeat steps 2 to 4 for a number of predefined range of *K* values;
2. Initialize the value of *K* at the beginning;
3. Calculate the resulting SSE value using *K* clusters;
4. Increase the value of *K*;
5. *Plotting* the resulting SSE values for each *K* cluster on the graph;
6. Analysis of the resulting SSE value which experienced a significant decrease; and

The *K* optimal value is determined from the point that forms the elbow on the graph.

K-Means Clustering Algorithm

The K-Means algorithm is one of the algorithm that is used for clustering purposes and is the most commonly used. This algorithm will partition a number of data into a number of *k* groups or clusters based on their proximity to the centroid. The centroid value in K-Means algorithm is determined by selecting a number of *k* data randomly. The pseudo code of K-Means clustering is presented in Table 1 [10], [26], [27].

The stages of the K-Means clustering algorithm to group data in more detail are as follows [6], [7], [28].

1. Specify the data you want to group.
2. Choose a number of k clusters.
3. Determine a random number of *k* data to serve as the centroid.
4. Perform steps 4.a to 4.b for each existing data.

4.a Calculate the data distance with all k centroids that exist using the Euclidean distance calculation using Equation (4).

$$D_{i,j} = \sqrt{\sum_{k=1}^{N}\left(x_{ik} - x_{jk}\right)^2} \quad (4)$$

Information:

$D_{i,j}$: distance between *i*-th data and *j*-th centroid

N : the number of data attributes

x_{ik} : value from *i*-th data in *k*-th attribute

x_{jk} : value from *j*-th centroid in *k*-th attribute

4.b Group data on a cluster that has the minimum distance with it.

5. Update the centroid based on the membership of each new cluster.
6. Repeat step 4 until centroid does not change.

Data from the last grouping was taken as a result when the centroid did not change anymore.

Particle Swarm Optimization (PSO)

PSO is one of the algorithms that can be used to solve different optimization problems [29]. This algorithm belongs to the metaheuristic method and was discovered by Kennedy and Eberhart in 1995. This algorithm is inspired by the social behavior of schools of fish swimming and flocks of birds flying in groups [9].

The centroid of K-Means clustering algorithm is optimized using PSO in this study. The PSO algorithm will determine a number of centroids randomly at the beginning of a predetermined number of particles. The centroids value in each particle will go to an optimal value as the velocity and position vectors of the particle are updated. The objective function used in this study to represent the optimality is minimizing the SSE value [9]. The smaller the SSE value produced by a particle, the more optimal the particle is. Therefore, fitness function used in this study is shown in Equation (5), where *f(p)* represents fitness value of *p*-th particle. Figure 3 shows the particle representation used to optimize the centroid in K-Means.

$$f(p) = \frac{1}{SSE} \quad (5)$$

The stages of the PSO algorithm to optimize the centroid on the K-Means are as follows:

1. Data preprocessing (D), number of clusters (optimal *K*), number of particles (*Np*), the maximum number of iterations (*max_iter*), parameters $\varphi_1$and φ_2 will be entered as inputs at this stage.
2. Initialize a number of *Np* particles representing the centroids to be optimized as in Figure 3.
3. Perform steps 3.a to 3.d as long as the *iter* value is less than *max_iter*.

3.a Performs the following steps for each particle.

(i) Calculate the fitness value of the *p*-th particle using Equation (5); and

(ii) Update the *Pbest* value from *p*-th particle using Equation (6). If the current *Pbest* value is better than before, then set the current fitness as *Pbest*. Otherwise, the *Pbest* value uses the previous.\

$$\mathrm{Pbest}_i^{t+1} = \begin{cases} \mathrm{Pbest}_i^t, & \mathrm{fitness}(x_i^{t+1}) \leq \mathrm{fitness}(\mathrm{Pbest}_i^t) \\ x_i^{t+1}, & \mathrm{fitness}(x_i^{t+1}) > \mathrm{fitness}(\mathrm{Pbest}_i^t) \end{cases} \quad (6)$$

3.b Update the *Gbest* value using Equation (7). If there is a particle that has a fitness value more than the previous *Gbest*, then set the current *Gbest* value from that *Pbest* particle's. Conversely, if no then the current *Gbest* value is the before *Gbest* value.

$$\mathrm{Gbest}^{t+1} = \begin{cases} \mathrm{Gbest}^t, \mathrm{argmax}(\mathrm{fitness}(\mathrm{Pbest}_i^{t+1})) \leq \mathrm{fitness}(\mathrm{Gbest}^t) \\ \mathrm{Pbest}_i^{t+1}, \mathrm{argmax}(\mathrm{fitness}(\mathrm{Pbest}_i^{t+1})) > \mathrm{fitness}(\mathrm{Gbest}^t) \end{cases} \quad (7)$$

3.c Perform the following steps for each particle.

(i) Update the particle velocity using Equation (8), (9), dan (10) [9], [30]; and

$$v_{i,j}^{t+1} = \chi\left[v_{i,j}^t + \varphi_1 \times r_1 \times \left(\mathrm{Pbest}_{i,j}^t - x_{i,j}^t\right) + \varphi_2 \times r_2 \times \left(\mathrm{Gbest}_j^t - x_{i,j}^t\right)\right] = \begin{cases} v_{max}, & v_{i,j}^{t+1} \geq v_{max} \\ v_{min}, & v_{i,j}^{t+1} \leq v_{min} \\ v_{i,j}^{t+1}, & v_{min} < v_{i,j}^{t+1} < v_{max} \end{cases} \quad (8)$$

$$\chi = \frac{2}{\left|2 - (\varphi_1 + \varphi_2) - \sqrt{(\varphi_1 + \varphi_2)^2 - 4(\varphi_1 + \varphi_2)}\right|} \quad (9)$$

$$\varphi = \varphi_1 + \varphi_2 > 4 \quad (10)$$

(ii) Update particle position using Equation (11).

$$x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1} = \begin{cases} T_{max(j)}, & x_{i,j}^{t+1} \geq T_{max(j)} \\ T_{min(j)}, & x_{i,j}^{t+1} \leq T_{min(j)} \\ x_{i,j}^{t+1}, & T_{min(j)} < x_{i,j}^{t+1} < T_{max(j)} \end{cases} \quad (11)$$

3.d Update *iter* value using Equation (12).

$$iter_{new} = iter_{old} + 1 \quad (12)$$

4. Optimal centroid is represented by the *Gbest* value that is generated in last iteration of the PSO mechanism. This centroid then used as the initial centroid in K-Means algorithm.

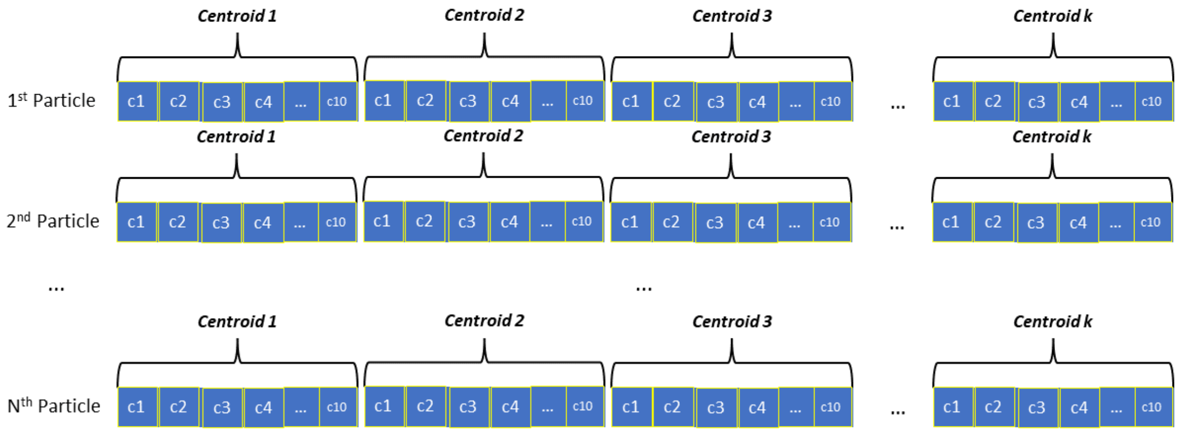

Fig 3. Particles representation for centroid optimization in K-Means using PSO Algorithm

Clustering Quality Measurement

The cluster quality that is generated using a clustering algorithm can be measured using extrinsic method and intrinsic method. Measures that can be used in the extrinsic method are BCubed precision and BCubed recall. The extrinsic method can only be used if there are references to the ideal clustering results produced by experts. The intrinsic method can be used if there is no reference to the clustering results produced by the expert. One measure that can be used in the intrinsic method is SC [10], [27].

This research uses the intrinsic method with the SC measurement. The quality of the clusters is measured using SC of all objects according to Equation (16) SC value is calculated using Equation (15). The values of a(o) and b(o) were obtained using Equation (13) and (14) espectively. The steps to determine silhouette coefficient value are as follows [6], [7], [10], [27]:

1. Calculate the *a(o)* value which is the average distance between object *o* and all other objects in the same cluster with it;
2. Calculate the *b(o)* value which is the minimum value of the average distance of the object *o* with all data in different clusters; and
3. Calculate the *s(o)* value which is the silhouette coefficient value of object *o*.

$$a(o) = \frac{\sum_{ó \in C_i, o \neq ó} dist(o,ó)}{|C_i - 1|} \quad (13)$$

$$b(o) = \min_{C_j: 1 \leq j \leq k,\, j \neq i} \left\{ \frac{\sum_{ó \in C_j} dist(o,ó)}{|C_j|} \right\} \quad (14)$$

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \quad (15)$$

$$\overline{SC} = \frac{\sum_{i=1}^{N} s(d_i)}{N} \quad (16)$$

RESULT AND DISCUSSION

Optimal Number of Clusters Test Results

Testing the optimal number of clusters or *K* is done using the Elbow method with 10 iterations. The highest of the average SSE differences from 10 iterations is used to choose the *K* optimal value. The range of *K* values used from 1 to 10 [26]. Table 2 and Table 3 are the results for the optimal number of clusters.

The highest results for the average value of the SSE difference is produced by K equal to 2, which is 556.41069 with a standard deviation of 0.01091. Based on Figure 4, *K* equal to 2 is the point that forms the elbow. Therefore, the optimal number of clusters that will be used in testing stage between K-Means with and without PSO is equal to 2.

Table 2. SSE value for each *K* for 10 iterations

K	SSE value for each *K* for 10 iterations										Mean	Std
	1	2	3	4	5	6	7	8	9	10		
1	1736,2	1736,2	1736,2	1736,2	1736,2	1736,2	1736,2	1736,2	1736,2	1736,2	1736,2	0,E+00
2	1179,8	1179,8	1179,8	1179,8	1179,8	1179,8	1179,8	1179,8	1179,8	1179,8	1179,8	1,E-02
3	1057,2	1053,3	1056,4	1063,8	1053,3	1063,8	1056,4	1056,8	1056,7	1056,4	1057,4	4,E+00
4	973,0	975,6	972,8	974,9	973,9	987,1	973,1	972,9	961,7	966,7	973,2	6,E+00
5	909,6	891,0	898,2	926,7	893,3	905,0	916,7	900,5	909,1	926,0	907,6	1,E+01
6	846,6	863,7	854,9	858,5	838,0	841,9	839,9	873,7	848,1	842,7	850,8	1,E+01
7	800,3	802,6	804,7	797,1	796,0	793,0	822,5	815,3	798,3	838,0	806,8	1,E+01
8	777,9	763,9	770,5	761,2	763,7	760,8	754,2	768,2	771,7	777,1	766,9	8,E+00
9	722,8	738,2	766,9	756,8	733,4	749,7	727,7	731,7	757,2	720,5	740,5	2,E+01
10	688,3	696,1	719,8	716,2	711,1	723,4	712,6	710,9	749,3	698,7	712,6	2,E+01

Table 3. The difference in SSE value for each *K* with (*K*-1) for 10 iterations

K	SSE difference value in each test *K* for 10 iterations										Mean	Std
	1	2	3	4	5	6	7	8	9	10		
1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,E+00
2	556,4	556,4	556,4	556,4	556,4	556,4	556,4	556,4	556,4	556,4	556,4	1,E-02
3	122,6	126,5	123,4	116,0	126,5	116,0	123,4	123,0	123,0	123,4	122,4	4,E+00
4	84,2	77,7	83,6	88,9	79,4	76,7	83,3	83,9	95,0	89,7	84,2	6,E+00
5	63,4	84,6	74,5	48,1	80,6	82,1	56,4	72,4	52,6	40,7	65,5	2,E+01
6	63,0	27,3	43,3	68,2	55,3	63,1	76,8	26,8	61,0	83,4	56,8	2,E+01
7	46,2	61,1	50,2	61,5	42,0	48,9	17,4	58,4	49,8	4,6	44,0	2,E+01
8	22,5	38,7	34,2	35,9	32,3	32,2	68,3	47,1	26,6	61,0	39,9	1,E+01
9	55,1	25,7	3,6	4,4	30,3	11,1	26,5	36,5	14,5	56,6	26,4	2,E+01
10	34,5	42,1	47,0	40,5	22,3	26,3	15,2	20,8	7,9	21,8	27,8	1,E+01

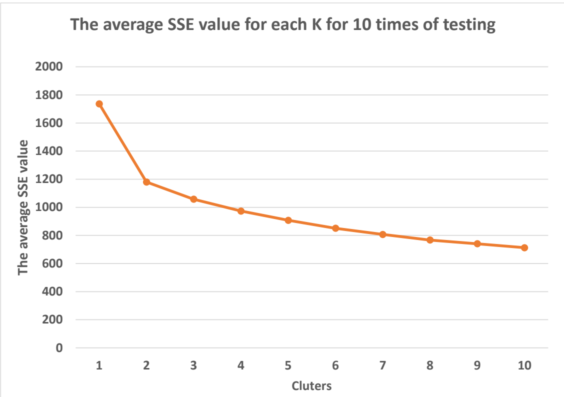

Fig 4. The average SSE value plot results for each K for 10 iterations of testing

PSO Algorithm Hyperparameter Test Results

Number of Particles Test Results

The number of particles (n_p), to be tested are 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 particles. In testing the n_p, the number of PSO iterations is set at 100 iterations, the value φ_1 is 2,3 and the value φ_2 is 2.5. Figure 7 shows the results of testing the number of particles for 10 trials. Based on Table 4, the number of particles of 30 gives the highest Silhouette Coefficient average value of 0.3004681. The resulting standard deviation is also very low and almost close to zero, which is 0.000545888.

Test Results for Combination of φ_1 and φ_2

The combination scheme of φ_1 and φ_2 values used in this test is a full combination. The values φ_1 and φ_2 range from 2.0 to 3.4 with an interval of 0.2. Each value in φ_1 will be paired with each value in φ_2, so that the full combination is obtained. In this test, the number of particles is set at 30 particles obtained from the result of the previous particle number test and the number of iterations is 100. Table 5 shows the results of testing the combination of φ_1 and φ_2 values for 10 trials.

Based on Table 5, the value of φ_1 equal to 2.2 and the value of φ_2 equal to 3 gives the highest SC average value of 0.30073138. The resulting standard deviation is also very low and almost close to zero, which is 0.000357509.

The trend graphs for the value of φ_1 and the value of φ_2 are shown in Figures 5 and 6, respectively. Based on Figure 5, the highest SC average value is obtained at the value of φ_1 equal to 2.2 and the value of φ_2 is kept constant equal to 3. Trend decreases if the value of φ_1 is less or more than 2.2. Based on Figure 6, the highest SC average value is obtained at the value of φ_2 equal to 3 and the value of φ_1 being kept constant equal to 2.2. The trend is decreasing if the value of φ_2 is less or more than 3.

Table 4. Number of PSO particles test results for 10 iterations of testing

n_p	Silhouette Coefficient value for each test										Mean	Std
	1	2	3	4	5	6	7	8	9	10		
10	0,2998	0,3010	0,2998	0,2998	0,2998	0,3009	0,3009	0,2998	0,3009	0,2998	0,3002	6,E-04
20	0,2998	0,2998	0,2998	0,2998	0,2998	0,3009	0,3000	0,2998	0,3010	0,2998	0,3000	5,E-04
30	0,3010	0,2998	0,3010	0,3009	0,3009	0,3009	0,3009	0,3000	0,2998	0,2998	**0,3005**	5,E-04
40	0,2998	0,2998	0,3009	0,2998	0,3009	0,2998	0,3009	0,3009	0,2998	0,3010	0,3003	6,E-04
50	0,2998	0,3000	0,3009	0,3009	0,2998	0,3009	0,2998	0,3009	0,3009	0,3000	0,3004	5,E-04
60	0,3009	0,3009	0,3009	0,3009	0,3009	0,2998	0,2998	0,2998	0,3000	0,2998	0,3003	6,E-04
70	0,3009	0,3009	0,3005	0,3009	0,3010	0,3000	0,3005	0,2998	0,2998	0,2998	0,3004	5,E-04
80	0,2998	0,3010	0,3009	0,3005	0,3009	0,2998	0,2998	0,2998	0,2998	0,2998	0,3002	5,E-04
90	0,3009	0,3009	0,2998	0,3000	0,2998	0,2998	0,3010	0,3000	0,3009	0,3009	0,3004	5,E-04
100	0,3000	0,3009	0,2998	0,3009	0,3005	0,3009	0,3007	0,3005	0,2998	0,2998	0,3004	5,E-04

Table 5. the test results for the combination of values φ_1 and φ_2

φ_1	φ_2	Silhouette Coefficient value for each test										Mean	Std
		1	2	3	4	5	6	7	8	9	10		
2	2	0,301	0,300	0,301	0,301	0,300	0,300	0,301	0,300	0,301	0,301	0,3005	5,E-04
	2,2	0,300	0,300	0,300	0,301	0,300	0,300	0,301	0,300	0,301	0,301	0,3004	5,E-04
	2,4	0,301	0,301	0,300	0,301	0,300	0,301	0,301	0,301	0,300	0,301	0,3005	5,E-04
	2,6	0,300	0,300	0,301	0,300	0,300	0,300	0,301	0,301	0,301	0,300	0,3002	6,E-04
	2,8	0,300	0,300	0,300	0,300	0,300	0,300	0,301	0,300	0,301	0,300	0,3001	5,E-04
	3	0,300	0,301	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,2999	3,E-04
	3,2	0,300	0,300	0,301	0,301	0,300	0,301	0,300	0,300	0,301	0,300	0,3003	5,E-04
	3,4	0,300	0,301	0,301	0,301	0,300	0,301	0,301	0,300	0,300	0,300	0,3003	5,E-04
2,2	2	0,301	0,300	0,301	0,301	0,301	0,300	0,300	0,300	0,301	0,301	0,3004	6,E-04
	2,2	0,301	0,301	0,301	0,301	0,301	0,300	0,301	0,301	0,300	0,300	0,3006	5,E-04
	2,4	0,300	0,301	0,300	0,300	0,301	0,300	0,300	0,301	0,300	0,301	0,3004	5,E-04
	2,6	0,301	0,301	0,300	0,300	0,301	0,301	0,301	0,300	0,301	0,301	0,3006	5,E-04
	2,8	0,300	0,300	0,300	0,300	0,301	0,301	0,301	0,301	0,300	0,300	0,3003	5,E-04
	3	0,300	0,301	0,301	0,301	0,300	0,301	0,301	0,301	0,301	0,301	**0,3007**	4,E-04
	3,2	0,300	0,301	0,301	0,301	0,300	0,300	0,301	0,300	0,300	0,300	0,3003	6,E-04
	3,4	0,300	0,300	0,301	0,300	0,301	0,300	0,300	0,300	0,301	0,300	0,3001	5,E-04
2,4	2	0,300	0,301	0,300	0,300	0,300	0,300	0,301	0,301	0,300	0,301	0,3004	5,E-04
	2,2	0,300	0,300	0,300	0,301	0,301	0,301	0,300	0,300	0,300	0,300	0,3002	5,E-04
	2,4	0,301	0,300	0,300	0,300	0,301	0,301	0,300	0,300	0,300	0,301	0,3002	5,E-04
	2,6	0,300	0,301	0,300	0,300	0,300	0,300	0,300	0,301	0,300	0,301	0,3003	5,E-04
	2,8	0,301	0,300	0,300	0,300	0,300	0,301	0,301	0,301	0,300	0,300	0,3003	5,E-04
	3	0,300	0,301	0,301	0,300	0,300	0,300	0,301	0,301	0,301	0,300	0,3004	5,E-04
	3,2	0,301	0,300	0,300	0,301	0,301	0,301	0,300	0,301	0,300	0,300	0,3004	5,E-04
	3,4	0,300	0,301	0,301	0,301	0,300	0,300	0,301	0,300	0,301	0,300	0,3004	5,E-04
2,6	2	0,300	0,300	0,301	0,300	0,300	0,301	0,300	0,300	0,300	0,301	0,3002	5,E-04
	2,2	0,301	0,300	0,301	0,300	0,300	0,300	0,301	0,300	0,300	0,301	0,3003	5,E-04
	2,4	0,301	0,300	0,300	0,300	0,301	0,301	0,300	0,301	0,300	0,300	0,3003	5,E-04
	2,6	0,300	0,300	0,301	0,301	0,300	0,301	0,300	0,300	0,300	0,300	0,3001	5,E-04
	2,8	0,301	0,301	0,301	0,301	0,301	0,300	0,300	0,300	0,301	0,300	0,3004	6,E-04
	3	0,300	0,301	0,300	0,301	0,300	0,301	0,301	0,301	0,300	0,300	0,3003	6,E-04
	3,2	0,300	0,301	0,300	0,301	0,300	0,301	0,300	0,300	0,300	0,300	0,3002	5,E-04
	3,4	0,300	0,301	0,300	0,300	0,300	0,301	0,301	0,300	0,300	0,301	0,3002	6,E-04
2,8	2	0,301	0,301	0,301	0,300	0,301	0,300	0,300	0,301	0,301	0,300	0,3004	6,E-04
	2,2	0,301	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,301	0,300	0,3001	4,E-04
	2,4	0,300	0,300	0,300	0,301	0,300	0,301	0,300	0,300	0,301	0,301	0,3003	5,E-04
	2,6	0,301	0,300	0,301	0,300	0,301	0,300	0,300	0,300	0,300	0,300	0,3002	5,E-04
	2,8	0,301	0,300	0,301	0,301	0,300	0,300	0,300	0,300	0,300	0,301	0,3003	5,E-04
	3	0,301	0,300	0,300	0,300	0,301	0,301	0,300	0,301	0,301	0,300	0,3003	5,E-04
	3,2	0,300	0,301	0,300	0,301	0,300	0,300	0,300	0,300	0,300	0,300	0,3000	5,E-04
	3,4	0,301	0,301	0,301	0,300	0,300	0,301	0,301	0,300	0,300	0,301	0,3005	5,E-04
3	2	0,300	0,301	0,300	0,301	0,300	0,301	0,301	0,301	0,301	0,300	0,3005	5,E-04
	2,2	0,301	0,300	0,300	0,301	0,300	0,301	0,300	0,301	0,301	0,300	0,3003	5,E-04
	2,4	0,300	0,301	0,300	0,300	0,300	0,301	0,301	0,300	0,300	0,301	0,3003	5,E-04
	2,6	0,300	0,300	0,300	0,300	0,300	0,300	0,301	0,301	0,300	0,301	0,3001	5,E-04
	2,8	0,301	0,301	0,301	0,300	0,301	0,301	0,300	0,301	0,300	0,301	0,3006	5,E-04
	3	0,300	0,301	0,300	0,301	0,301	0,300	0,301	0,301	0,300	0,300	0,3003	6,E-04
	3,2	0,300	0,301	0,301	0,300	0,300	0,301	0,301	0,301	0,300	0,300	0,3003	6,E-04
	3,4	0,301	0,300	0,301	0,300	0,300	0,300	0,301	0,300	0,300	0,301	0,3004	5,E-04
3,2	2	0,301	0,301	0,301	0,300	0,300	0,300	0,300	0,300	0,300	0,301	0,3002	6,E-04
	2,2	0,300	0,301	0,301	0,301	0,300	0,300	0,300	0,301	0,300	0,301	0,3004	5,E-04
	2,4	0,300	0,300	0,300	0,300	0,301	0,301	0,300	0,301	0,301	0,301	0,3004	5,E-04

	2,6	0,301	0,301	0,301	0,301	0,300	0,300	0,300	0,300	0,301	0,300	0,3004	5,E-04
	2,8	0,300	0,300	0,300	0,300	0,301	0,300	0,300	0,300	0,301	0,300	0,3000	4,E-04
	3	0,301	0,301	0,301	0,300	0,300	0,301	0,301	0,300	0,300	0,301	0,3005	5,E-04
	3,2	0,301	0,300	0,301	0,301	0,300	0,300	0,300	0,301	0,301	0,301	0,3005	5,E-04
	3,4	0,300	0,300	0,301	0,301	0,300	0,300	0,300	0,301	0,301	0,301	0,3003	6,E-04
	2	0,300	0,300	0,300	0,300	0,301	0,300	0,301	0,300	0,301	0,300	0,3001	5,E-04
	2,2	0,301	0,300	0,301	0,301	0,301	0,300	0,301	0,301	0,300	0,301	0,3005	5,E-04
	2,4	0,301	0,300	0,300	0,300	0,301	0,300	0,301	0,300	0,300	0,301	0,3003	5,E-04
3,4	2,6	0,300	0,300	0,300	0,301	0,301	0,300	0,301	0,301	0,300	0,300	0,3003	5,E-04
	2,8	0,300	0,301	0,300	0,300	0,300	0,301	0,301	0,300	0,300	0,300	0,3002	5,E-04
	3	0,301	0,300	0,301	0,301	0,301	0,301	0,301	0,300	0,300	0,300	0,3005	5,E-04
	3,2	0,301	0,300	0,301	0,300	0,300	0,300	0,300	0,300	0,301	0,300	0,3002	5,E-04
	3,4	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,300	0,2999	2,E-04

Table 6. Comparison test results of the K-Means clustering algorithm with (M1) and without (M2) PSO Algorithm for 10 times of testing

Methods	**Silhouette Coefficient value for each test**										**Mean**	**Std**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		
M1	0,2998	0,3005	0,2998	0,3009	0,2998	0,3010	0,2998	0,2998	0,2998	0,2998	0,3001	5,E-04
M2	0,3000	0,2998	0,2998	0,2998	0,3009	0,3009	0,3010	0,3010	0,3009	0,2998	0,3004	6,E-04

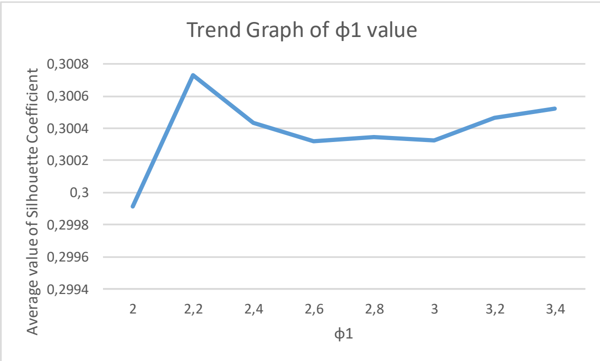

Fig 5. Trend graph of φ_1 when φ_2 is 3

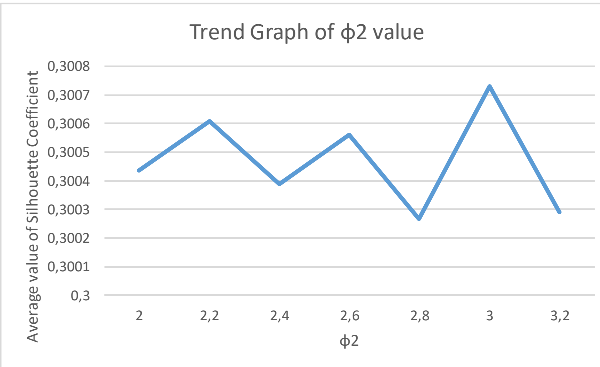

Fig 6. Trend graph of φ_2 when φ_1 is 2,2

Algorithm Comparison Test Results

The results of clustering generated using the K-Means Algorithm with and without optimization of the PSO Algorithm will be compared. The optimal PSO hyperparameters on the previously generated will be used in this test. The maximum number of iterations is set at 100; the number of particles is 30; and the value of φ_1 is 2.2 and the value of φ_2 is 3 will be set at the beginning before testing the two methods. This test will be carried out 10 times. Comparison test results of the K-Means with and without PSO is shown in Table 6

Based on Table 6, the K-Means Algorithm optimized using the PSO gives better results. This is indicated by the average SC value for 10 times higher testing than the K-Means Algorithm without optimization of the PSO Algorithm. The resulting standard deviation is also low and close to 0.

CONCLUSION

The conclusion of this study is that the results of the clustering of tourist review data on the Tripadvisor site generated by the K-Means Algorithm with optimization of the PSO Algorithm provide better results than without the optimization of the PSO Algorithm. Optimized K-Means Algorithm PSO Algorithm with hyperparameter number of particles of 30; and the value of φ_1 is 2.2 and the value of φ_2 is 3 giving the SC average value of 0.300358 and those without optimization of the PSO Algorithm giving the SC average value of 0.300076.

REFERENCES

[1] M. Nurjanah and T. Arifin, "PENERAPAN ALGORITMA K-MEANS UNTUK ANALISIS DATA ULASAN DI SITUS TRIPADVISOR," *JURNAL RESPONSIF*, vol. 3, no. 1, pp. 75–82, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti

[2] A. Kustanto, "PARIWISATA : SEBAGAI SALAH SATU UPAYA MEMPEROLEH DEVISA BAGI PEMERINTAH INDONESIA," 2019.

[3] S. A. Azzahra and A. Wibowo, "ANALISIS SENTIMEN MULTI-ASPEK BERBASIS KONVERSI IKON EMOSI DENGAN ALGORITME NAÏVE BAYES UNTUK ULASAN WISATA KULINER PADA WEB TRIPADVISOR," vol. 7, no. 4, 2020, doi: 10.25126/jtiik.202071907.

[4] F. K. Wardhani, E. Suryani, and A. Mukhlason, "Penerapan metode GA-Kmeans untuk pengelompokan pengguna pada Bapersip Provinsi Jawa Timur," *JURNAL TEKNIK ITS*, vol. 1, no. 1, pp. 545–550, 2012.

[5] S. Rustam, H. A. Santoso, and C. Supriyanto, "OPTIMASI K-MEANS CLUSTERING UNTUK IDENTIFIKASI DAERAH ENDEMIK PENYAKIT MENULAR DENGAN ALGORITMA PARTICLE SWARM OPTIMIZATION DI KOTA SEMARANG," 2018.

[6] B. Santoso, I. Cholissodin, and B. D. Setiawan, "Optimasi K-Means untuk Clustering Kinerja Akademik Dosen Menggunakan Algoritme Genetika," 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[7] D. A. Kuntjoro, B. Darma Setiawan, and R. S. Perdana, "Algoritme Genetika Untuk Optimasi K-Means Clustering Dalam Pengelompokan Data Tsunami," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[8] Y. P. Anggodo, W. Cahyaningrum, A. N. Fauziyah, I. L. Khoiriyah, O. Kartikasari, and I. Cholissodin, "HYBRID K-MEANS DAN PARTICLE SWARM OPTIMIZATION UNTUK CLUSTERING NASABAH KREDIT," vol. 4, no. 2, pp. 2355–7699, 2017.

[9] Suyanto, *Swarm Intelligence Komputasi Modern Untuk Optimasi dan Big Data Mining*. Bandung: Informatika Bandung, 2017.

[10] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*, Revisi. Bandung: Informatika Bandung, 2018.

[11] I. Hidayatin, S. Adinugroho, and C. Dewi, "Pengelompokan Wilayah berdasarkan Penyandang Masalah Kesejahteraan Sosial (PMKS) dengan Optimasi Algoritme K-Means menggunakan Self Organizing Map (SOM)," 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[12] A. Primandana, S. Adinugroho, and C. Dewi, "Optimasi Penentuan Centroid pada Algoritme K-Means Menggunakan Algoritme Pillar (Studi Kasus: Penyandang Masalah Kesejahteraan Sosial di Provinsi Jawa Timur)," 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[13] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "OPTIMASI NAÏVE BAYES CLASSIFIER DENGAN MENGGUNAKAN PARTICLE SWARM OPTIMIZATION PADA DATA IRIS," vol. 4, no. 3, pp. 180–184, 2017.

[14] N. F. Istighfarin, R. A. Rahmastati, and H. Nugroho, "PENERAPAN METODE PARTICLE SWARM OPTIMIZATION (PSO) DAN GENETIC ALGORITHM (GA) PADA SISTEM OPTIMASI VISIBLE LIGHT COMMUNICATION (VLC) UNTUK MENENTUKAN POSISI ROBOT," *Jurnal SIMETRIS*, vol. 11, no. 1, pp. 279–286, Apr. 2020.

[15] M. Sharif, J. Amin, M. Raza, M. Yasmin, and S. C. Satapathy, "An integrated design of particle swarm optimization (PSO) with fusion of features for detection of brain tumor," *Pattern Recognit Lett*, vol. 129, pp. 150–157, Jan. 2020, doi: 10.1016/j.patrec.2019.11.017.

[16] T. M. Phan, P. T. Ha, T. L. Duong, and T. T. Nguyen, "Improved particle swarm optimization algorithms for economic load dispatch considering electric market," *International Journal of Electrical and Computer Engineering*,

vol. 10, no. 4, pp. 3918–3926, 2020, doi: 10.11591/ijece.v10i4.pp3918-3926.

[17] A. Rahman and H. M. Asih, "Optimizing shipping routes to minimize cost using particle swarm optimization," *International Journal of Industrial Optimization*, vol. 1, no. 1, pp. 53–60, 2020.

[18] R. Wati, "PENERAPAN ALGORITMA NAIVE BAYES DAN PARTICLE SWARM OPTIMIZATION UNTUK KLASIFIKASI BERITA HOAX PADA MEDIA SOSIAL," *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, vol. 5, no. 2, pp. 9–14, 2020, [Online]. Available: www.bsi.ac.id

[19] T. Arifin and D. Ariesta, "PREDIKSI PENYAKIT GINJAL KRONIS MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER BERBASIS PARTICLE SWARM OPTIMIZATION," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 26–30, Apr. 2019, doi: 10.36787/jti.v13i1.97.

[20] K. S. Nugroho, I. Istiadi, and F. Marisa, "Naive Bayes classifier optimization for text classification on e-government using particle swarm optimization," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 1, pp. 21–26, Jan. 2020, doi: 10.14710/jtsiskom.8.1.2020.21-26.

[21] F. Yunita, Pranowo, and A. J. Santoso, "Hybrid model of particle swarm and ant colony optimization in lecture schedule preparation," in *AIP Conference Proceedings*, American Institute of Physics Inc., Jun. 2018. doi: 10.1063/1.5042895.

[22] D. Novianti, D. Anggraini, and P. Hapsari, "Analisis Perbandingan Algoritma Particle Swarm Optimization Dan Firefly Algorithm Dalam Menentukan Minimum Spanning Tree," *Rang Teknik Journal*, vol. I, no. 2, 2018, [Online]. Available: http://joernal.umsb.ac.id/index.php/RANGTEKNIKJOURNAL

[23] W. A. Setyowati and W. F. Mahmudy, "Optimasi Vektor Bobot Pada Learning Vector Quantization Menggunakan Particle Swarm Optimization Untuk Klasifikasi Jenis Attention Deficit Hyperactivity Disorder (ADHD) Pada Anak Usia Dini," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 11, pp. 4428–4437, Nov. 2018.

[24] I. Romadhona, I. Cholissodin, and Marji, "Penerapan Algoritme Particle Swarm Optimization-Learning Vector Quantization(PSO-LVQ) Pada Klasifikasi Data Iris," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 12, pp. 6418–6428, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[25] A. T. Rahman, "Coal Trade Data Clusterung Using K-Means (Case Study PT. Global Bangkit Utama)."

[26] B. Purnama, *Pengantar Machine Learning Konsep dan Praktikum Dengan Contoh Latihan Berbasis R dan Python*. Bandung: Informatika, 2019.

[27] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung, 2018.

[28] E. Susilowati, A. T. Hapsari, M. Efendi, and P. E. Kresnha, "DIAGNOSA PENYAKIT KANKER PAYUDARA MENGGUNAKAN METODE K-MEANS CLUSTERING," *JUST IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer*, vol. 10, no. 1, pp. 27–32, 2019, [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it

[29] T. Z. Khalaf, H. Çağlar, A. Çağlar, and A. N. Hanoon, "Particle swarm optimization based approach for estimation of costs and duration of construction projects," *Civil Engineering Journal (Iran)*, vol. 6, no. 2, pp. 384–401, Feb. 2020, doi: 10.28991/cej-2020-03091478.

[30] Suyanto, A. Arifianto, R. Rita, and S. Andi, *Evolutionary Machine Learning Pembelajaran Mesin Otonom Berbasis Komputasi Evolusioner*. Bandung: Informatika Bandung, 2020.