International Journal of
Mathematical Modelling & Computations
Vol. 12, No. 02, Spring 2022, 143- 152

Clustering with K-means Hybridization Ant Colony Optimization (K-ACO)

D. J. Ratnaningsih*

Universitas Terbuka, Jl. Cabe Raya, Pondok Cabe, Pamulang, Kota Tangerang Selatan, Indonesia.

Abstract. One of the well-known techniques in data mining is clustering. The very popular clustering method is K-means cluster because its algorithm is very easy and simple. However, the K-means cluster has some weaknesses, one of which is that the clustering result is sensitive toward centroid initialization so that the clustering result tends to be locally optimal. This paper explains the modification of the K-means cluster, that is, K-means hybridization with ant colony optimization (K-ACO). Ant Colony Optimization (ACO) is an optimization algorithm based on ant colony behavior. Through K-ACO, the weaknesses of cluster result which tends to be local optimal can be overcome well. The application of the hybrid method of K-ACO with the use of the R program gives better accuracy compared to the K-means cluster. K-means cluster accuracy yielded by Minitab, Matlab, and SAS at iris data is 89%. Meanwhile, K-ACO hybrid clustering with the R program simulated on 38 treatments with 3-time repetitions gives an accuracy result of 93.10%.

Received: 28 February 2022, Revised: 23 April 2022, Accepted: 26 May 2022.

Keywords: Clustering; Data mining; K-means; Ant colony optimization; Program R; Iris data.

Index to information contained in this paper

1. Introduction

Clustering is a method of grouping based on the size of the proximity (similarity). In data mining, clustering is the unsupervised grouping. The purpose of clustering is to group data with similar characteristics to a specific 'cluster' and data with different characteristics to another 'cluster'.

Many cluster methods are used in data mining. The most popular method is the K-means cluster because the algorithm is simple and easy. However, the K-means cluster has a weakness because, among the cluster results, they are sensitive to the initial determination (initialization) of the cluster center (centroid). In addition, another drawback is that the K-means cluster cannot resolve an optimal problem locally.

Determination of the clustering initialization is very important to produce a cluster corresponding to the actual data. Therefore, other methods are needed to address weaknesses in the K-means cluster. One method proposed in this paper is the K-means Ant Colony Optimization (abbreviated K-ACO). This method is used to determine the optimal cluster membership where optimal local problems can be solved better than K-means.

Ant (ant) is a metaheuristic optimization method. This algorithm is inspired by the behavior of ants living in colonies (colony) for determining the shortest route gained in bringing food to the nest of ants. This algorithm is known as Ant Colony Optimization

*Corresponding author. Email: djuli@ecampus.ut.ac.id

http://ijm2c.iauctb.ac.ir

(ACO). ACO algorithm was first discovered by Moyson and Mendeick and further developed by Dorigo et al. [1]. This algorithm is proved to be extremely flexible in overcoming problems of optimization permutations on a large-scale data.

Today the use of ACO is more widely in many fields. ACO Implementation has been proven both in overcoming problems of the traveling salesman, scheduling travel routes, and urban transportation systems. The application of ACO is also more widely in data mining, namely clustering, and classification. This paper will explain how the performance of K-ACO hybridization methods in determining the clustering of data compared to the clustering of data generated by the method of K-means.

2. Literature review

2.1 *K-means cluster*

K-means is one example of a non-hierarchical clustering data method that seeks to partition the data into the form of one or more clusters. K-means clustering method is simple and easy in the classification of data into several clusters. The main idea is to define the K-means centroid, one for each cluster. This method partitions the data into clusters so that the data, which have the same characteristics, are grouped into the same cluster and the data that have different characteristics are grouped into other clusters (Varma and Kumar, 2014).

The purpose of data clustering is to minimize the objective function that is set in the clustering process. In general, clustering is trying to minimize variations in a cluster and maximize inter-cluster variation. The basic algorithm K-means cluster is as follows:

1. Determine the number of clusters
2. Allocate data into the random cluster
3. Calculate the centroid/average of the data in each cluster
4. Allocate each data to the centroid/average nearby
5. Go back to Step 3, if there are still data that move into cluster or change value, if there are data above a threshold value or if the value changes in the objective function used above a specified threshold value.

2.2 *Ant colony optimization (ACO)*

Ant Colony Optimization (ACO) is an algorithm optimization based on the behavior of ant colonies developed by Dorigo and Caro in the early 90s. ACO is a branch of artificial intelligence called swarm intelligence (SI). Definition of swarm intelligence is a problem-solving method that utilizes the behavior of a set of agents working together. ACO algorithm is inspired by the social behavior of ant colonies in reaching the nearest route of the food source from the nest. The closest route selection is done by utilizing a chemical material called pheromone that is released from the foot while walking. The pheromone will attract the attention of other ants to follow a route. The more the amount of pheromone that exists on a route, the more potential the route is to be followed by other ants. Muñoz et al. [6] had shown that pheromone tracks left by ant colony make it possible to optimize the route where every agent (ant) is to contribute in finding the best solution. The behavior of ant colonies in bringing food to the nest is shown in Figure 1.

ACO algorithm was developed by Dorigo [2] to resolve the case of the Traveling Salesman Problem (TSP). ACO concept is applied in many cases, especially for classification and grouping. In general, the stages in the ACO algorithm include: (1) The status transition rules, (2) an update rule of pheromone, and (3) the repetition of transition rules to achieve the desired criteria (Liu [4]). The third stage in the ACO is described as follows.

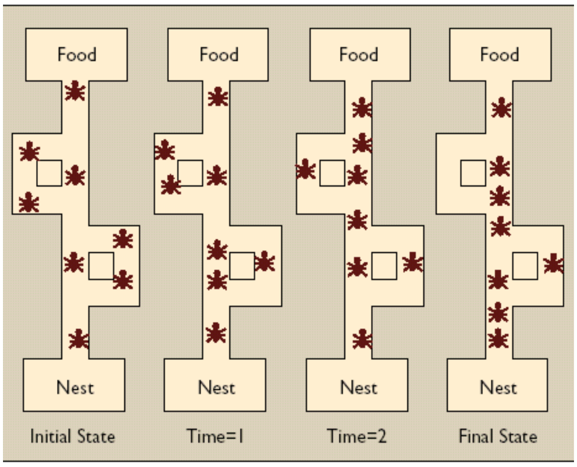

Figure 1. Colony behaviour of ants bringing food to their nest [4].

2.2.1 Status transition rules

In a status transition rule of an ant agent at the position *i* would choose to move to a point *j* with a probability as follows:

$$P_{ij}^{k}(t) = \begin{cases} \dfrac{\left[\tau_{ij}(t)\right]^{\alpha}[\eta_{tk}(t)]^{\beta}}{\sum_{s\subset J_k(i)}[\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}}. & \text{if } j \in J_k(i) \\ 0 & \text{if others} \end{cases} \tag{1}$$

$$\eta_{ij} = \frac{1}{d_{ij}} \tag{2}$$

$$d_{ij} = \sqrt{\left(x_{1i} - x_{1j}\right)^2 + \left(x_{2i} - x_{2j}\right)^2 + \cdots + \left(x_{pi} - x_{pj}\right)^2} \tag{3}$$

with:
$P_{ij}^{k}(t)$ is the probability of ant transition status from point i to point j on iteration of-*t*.
$\tau_{ij}(t)$ is the value of pheromone between point i and point j on interation of-*t*
d_{ij} is the distance between point i to point j
$J_k(i)$ is the various points that haven't been visited by ants when they reached point i.
α and β is the heuristic parameter.

2.2.2 Pheromone update rules

There are two rules in the ACO pheromone update, namely local pheromone update rules and global pheromone update. Local pheromone update aims to randomize the direction of the tracks being built, so that the points that have been covered previously by a tour of an ant may be passed later by a tour of the other ants. In other words, the influence of the local pheromone update is to make the level of interest in the existing segments change dynamically; that is, every time an ant uses a segment then this segment will soon be reduced in the level of interest (for this segment has lost a number of its pheromones). Thus, indirectly, the other ants will choose other segments that have not been visited. Consequently, the ant would have a tendency to converge on the same trajectory. This fact is explained through an experiment conducted by Dorigo and Gambardella [2].

Meanwhile, the global pheromone update is intended to provide more pheromone on shorter ant tours. In this system, the global pheromone update is only done by ants that make the shortest tour since the beginning of the trial. At the end of an iteration, after all, ants finish their tour, a number of pheromone are placed in segments that have been passed by an ant that has found the best tour (the other segments are not changed). Thus, at this stage, segments that are part of the global best tour will receive additional pheromone. The following is an explanation of the details of the local and global pheromone update on ACO.

a. Local pheromone update

When an ant moves from point i to point j, it will change the level of pheromone with a formula as follows:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) \tag{4}$$

with

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{N} \Delta\tau_{ij}^{k}(t) \tag{5}$$

$$\Delta\tau_{ij}^{k}(t) = \begin{cases} \frac{Q}{L_k}. & \text{jika } t \rightarrow t+1.\ \ point\ i \rightarrow \text{point } j \\ 0\,. & \text{the others} \end{cases} \tag{6}$$

N is the number of ants.
Q is the constant (level of pheromone).
L_k is the distance reached by the k ant.
ρ is the residual value of pheromone.

b. Global pheromone update

The interest in global pheromone update is to improve the level of ant pheromone that has been passed to produce the best performance. The update of global pheromone used is by the following formula.

$$\tau_{ij} = (1-\rho)\tau_{ij} + \sum_{k=1}^{N} \Delta\tau_{ij} \tag{7}$$

with $\rho\ \epsilon\ (0{,}1)$ as the level of pheromone evaporation on the global update and $\Delta\tau_{kij}$ is the amount of pheromone that derived from the best trajectory obtained. The amount of pheromone on the best path is calculated using the following formula.

$$\Delta\tau_{kij} = \begin{cases} \frac{Q}{L_k}. & \text{if } k\ is\ an\ ant\ used\ at\ a\ tour\ of\ \text{i. j} \\ 0\,. & \text{the others} \end{cases} \tag{8}$$

L_k is the length of the path taken by the ants. Updates of pheromones by ants are referred to as the global pheromone update.

2.3 *Traveling salesman problem using ant colony system*

Traveling Salesman Problem (TSP) is the most popular issue in the field of combinatorial optimization. A problem in the TSP is to determine the shortest path that can be taken by the salesman to the entire municipal customer. Then they returned to their hometowns if all town customers may only be visited once in each tour. The TSP result to be obtained is to find the shortest path interconnecting of n cities. Each city may only be visited once.

Several studies on the application of ACO in the TSP have a lot to do, among which is Dorigo et al. [1] which is the originator of the first application of the ACO in the TSP. In addition, the ACO application in the TSP has been done by Syamala and Prabha (2014) and Mohan and Remya [5]. In the TSP, if there are n cities that will be visited by then there will be as many as $(n \times (n-1))/2$ fruit segments and will have as many as $(n-1)!/2$ routes that are possible. The distance used in the standard TSP is asymmetrical distance, which means that the distance between city *i* to city *j* is equal to the distance between the cities *j* to city i, $d(i,j) = d(j,i)$.

Broadly speaking, the ACO process in the TSP is as follows. First, each ant at the beginning of the search for the shortest path is to put yourself in each city beginning at random. Then, each ant visited other towns that had never been seen until all the cities were visited. Each ant would have a list of visited cities that were never missed. The selection of the cities that have never been visited is based on a rule called transition rule status. Under this rule, it is the inverse distance of one city to another and the amount of pheromone within each segment. Each ant is only allowed to visit the city once.

The next stage is the process of updating pheromone both locally and globally in each segment traversed by ants. The shorter a tour generated by ant, the more the amount of pheromone left on sections traversed. In conclusion, ants will find the shortest path from the amount of pheromone produced.

3. Research method

Data used in this research is the iris data obtained from: https://archive.ics.uci.edu/-ml/machine-learning-databases/iris/iris.data. The reason is that the use of data slices that has international standard classification/clustering of data and clusterization is the absolute truth. In the application of the proposed hybridization method, the iris data are standardized because the size used is different for each characteristic.

There are three stages of K-ACO methods examined in this paper, namely (1) the K-means clustering, (2) optimization of the centroid clustering with the ACO, and (3) ends with K-means clustering again. The first stage, i.e. the use of K-means, is to obtain the initial centroid and the cluster along with its members. In the second stage, the resulting cluster and a member of the K-means (stage 1) serve as input to the ACO (stage 2). Then run the ACO process to obtain the best service.

The ACO process used in this paper is the ACO process that is applied to the Traveling Salesman Problem (TSP). The best of this new cluster centroid is obtained in the form of the start point and endpoint of the best route of each cluster. After that, it is resumed with K-means clustering (stage 3). The last cluster is formed by means such as at stage 1, the observed objects are placed back into the cluster using the new centroid produced from the ACO (phase 2) based on the proximity of each object. In detail, the stages of K-ACO are as follows.

Stage 1: Clustering K-means to determine cluster members

1. Determine the value of *k* which is the number of clusters generated from the *K*-means clustering algorithm.
2. Select the object *k* at random to be the centroid initial value.
3. Place the objects into a cluster that has the shortest distance to the centroid.
4. Redefine the centroid of each cluster by calculating the average value of the whole object members.
5. Repeat steps 4 and 5 to produce a centroid that is no longer changed (or until a certain iteration value).

Stage 2: The use of ant colony optimization (ACO) for new centroid

1. Determine the number of iterations of the ACO, ants, pheromone evaporation levels,

and heuristic and historical constants.
2. For each cluster formed from the K-means (stage 1), carried out by the ACO process parameters, it has been determined.
3. Save these routes of ACO best results for each cluster.
4. Specify the new cluster centroid of the starting point of these routes best obtained in step 3.

Having obtained a new centroid of Key Stage 2, the researcher then carried back the K-means clustering as Stage 1. This was done to obtain the final results of the cluster as the result of the process of K-ACO. Visually, the flow chart of algorithms of the K-ACO hybridization method is presented in Figure 2.

The method used in this study is a simulation. Simulations were conducted in 38 treatment combinations of ACO parameters and repeated 3 times. Studied parameters and tested combinations of treatment parameters are presented in Table 1. The software used is *the R* Program.

Table 1. Parameter Simulation of K-ACO hybridization method.

Parameter	Explanation	Parameter combination
t	The number of iteration	100, 500
N	The number of ant	30, 50, 100, 150
α	Heuristic parameter (pheromone weight), $\alpha \geq 0$	0,5; 2,0; 2,5; 3,0
β	Heuristic Parameter (visibility weight of each track), $\beta \geq 0$	3; 4; 4,5; 5; 6
ρ	Pheromone evaporation parameter	0,1; 0,4; 0,5; 0,6
Q	Constanta (the level of pheromone) to update global pheromone	0,2; 0,5; 0,9

The selection of some combination of parameters as shown in Table 1 is based on the results of research conducted by Dorigo and Stuzle [3], Tanjung [8], Wicaksana and Widiartha [10]. Meanwhile, the criteria used to determine the performance of clustering results of the K-ACO hybridization method is to calculate the percentage of the resulting clustering (clustering accuracy percentage). This clustering accuracy percentage is compared with the data clustering iris which has been internationally standardized. Results of the K-ACO clustering accuracy compared well with the results of K-means clustering of some of the software are SAS, Minitab, and Matlab.

4. Result and discussion

As described in research methods, the simulation conducted in this study was 38 treatments with replications in each treatment as many as 3 times. The average percentage of accuracy in three replications is presented in Table 2, Table 3, and Table 4. Table 2 presents the results of simulating the average accuracy of clustering iteration of 100 with the number of 30 ants. Table 3 presents the results of simulating the average accuracy of clustering iteration of 100 with the number of 50 ants. The treatment of combinations of parameters in Table 2 and Table 3 is a simulation conducted by Dorigo and Stutzle [3]. Meanwhile, a simulation of a combination treatment of the parameters in Table 4 was conducted by Tanjung [8], Wicaksana, and Widiartha [10]. However, in this study, iterations were performed 500 times.

Based on Table 3, shows that the average percentage of clustering in simulation accuracy with the number of ants is as many as 50 fish, larger than the number of ants with as many as 30 individuals. The average percentage of accuracy of clustering at this treatment amounted to 93.56%. The highest accuracy percentage was 94.67% and the lowest was 88.67%. In detail, each percentage clustering accuracy on some combination treatment is tested on a number of ants of as many as 50 animals which can be seen in Table 3.

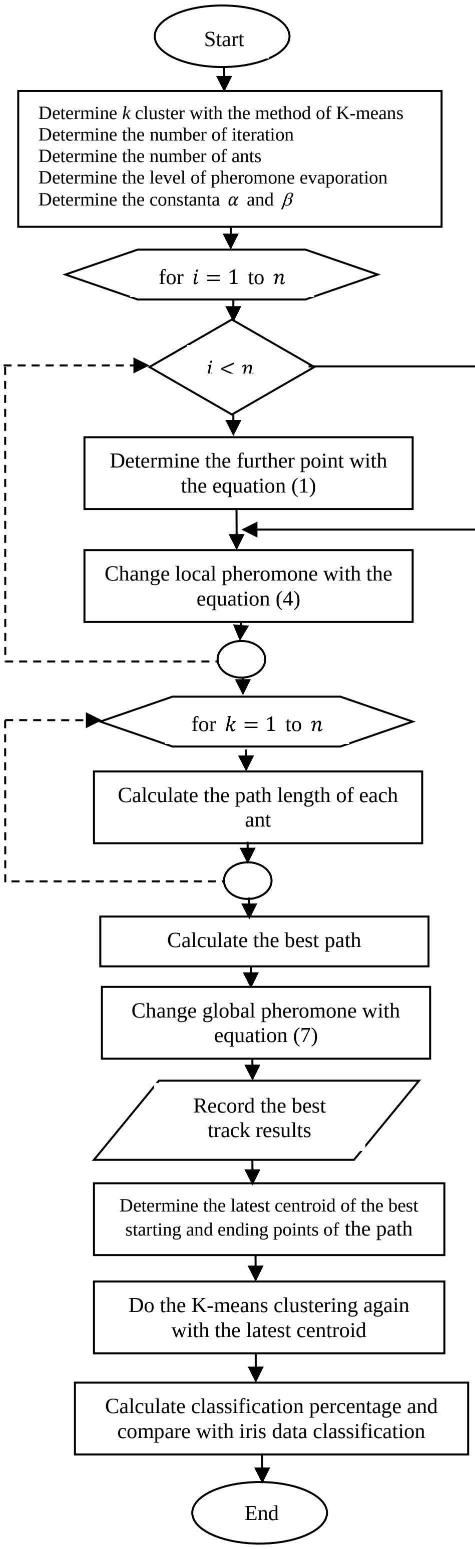
Start
Determine k cluster with the method of K-means
Determine the number of iteration
Determine the number of ants
Determine the level of pheromone evaporation
Determine the constanta α and β
for $i = 1$ to n
$i < n$
Determine the further point with the equation (1)
Back to the previous point
Change local pheromone with the equation (4)
for $k = 1$ to n
Calculate the path length of each ant
Calculate the best path
Change global pheromone with equation (7)
Record the best track results
Determine the latest centroid of the best starting and ending points of the path
Do the K-means clustering again with the latest centroid
Calculate classification percentage and compare with iris data classification
End

Figure 2. The Flow Diagram of K-ACO hybridization algorithm.

Table 2. The result of the average simulation of clustering accuracy at some treatments of parameter ($t = 100$ and $N = 30$).

No	Iteration (t)	Number of Ant (N)	α	β	ρ	Q	Accuracy percentage
1	100	30	2	3	0.1	0.9	92.67
2	100	30	2	3.5	0.1	0.9	91.33
3	100	30	2	4	0.1	0.9	93.33
4	100	30	2	4.5	0.1	0.9	89.33
5	100	30	2	5	0.1	0.9	89.33
6	100	30	2.5	3	0.1	0.9	91.33
7	100	30	2.5	3.5	0.1	0.9	90.67
8	100	30	2.5	4	0.1	0.9	89.33
9	100	30	2.5	4.5	0.1	0.9	94.00
10	100	30	2.5	5	0.1	0.9	94.67
11	100	30	3	3	0.1	0.9	90.67
12	100	30	3	3.5	0.1	0.9	93.33
13	100	30	3	4	0.1	0.9	89.33
14	100	30	3	4.5	0.1	0.9	97.33
15	100	30	3	5	0.1	0.9	89.33

Table 4 shows the percentage accuracy of clustering in 500 iterations with several combinations of parameters. Based on Table 4 it shows that the highest percentage of accuracy is in the combination treatment of $\alpha = 2$; $\beta = 5$; $\rho = Q = 0.5$. The percentage value clustering accuracy is 99.33%. Visibility from some combinations of treatment produced a greater percentage of the simulation of iteration of 100. This value is greater than that of the K-means clustering method.

Based on the simulation results shown in Table 2, Table 3, and Table 4, show that the percentage accuracy of clustering produced by the K-ACO hybridization method is better than K-means clustering. In general, the average percentage of clustering generated by the hybridization method K-ACO is 93.10%. Meanwhile, the percentage of accuracy of the results of the clustering K-means method amounted to 89.33%. Thus, it can be said that the performance of clustering generated by the hybridization method K-ACO performed on the data slices is better than clustering produced by the method of K-means.

5. Conclusion

Ant colony optimization (ACO) is a metaheuristic optimization method that has been widely applied to various fields. The ability of the ACO algorithm in solving the optimization problem makes it popular and wider in its application, including the clustering of data mining.

In clustering, merging the ACO algorithm with the K-means clustering method (K-ACO) can overcome the problem of local sensitivity and optimal centroid initialization. This is indicated by the percentage of the resulting clustering accuracy. The simulation results using data slices show that the percentage accuracy of clustering produced by the K-ACO hybridization method is better than the K-means clustering method. The percentage accuracy of the data clustering by K-means slices is 89.33%, while the K-ACO is 93.10%.

Table 3. The result of the average simulation of clustering accuracy percentage of some parameter treatments ($t = 100$ and $N = 50$).

No	Iteration (t)	Number of Ant (N)	A	β	ρ	Q	Accuracy Percentage
1	100	50	2	3	0.1	0.9	94.67
2	100	50	2	3.5	0.1	0.9	94.67
3	100	50	2	4	0.1	0.9	94.00
4	100	50	2	4.5	0.1	0.9	93.33
5	100	50	2	5	0.1	0.9	94.67
6	100	50	2.5	3	0.1	0.9	88.67
7	100	50	2.5	3.5	0.1	0.9	92.67
8	100	50	2.5	4	0.1	0.9	94.67
9	100	50	2.5	4.5	0.1	0.9	91.33
10	100	50	2.5	5	0.1	0.9	93.33
11	100	50	3	3	0.1	0.9	93.33
12	100	50	3	3.5	0.1	0.9	94.67
13	100	50	3	4	0.1	0.9	94.67
14	100	50	3	4.5	0.1	0.9	94.00
15	100	50	3	5	0.1	0.9	94.67

Table 4. The result of average simulation of clustering accuracy percentage at some parameter treatments ($t = 500$).

No	Iteration (t)	Number of Ant (N)	α	β	ρ	Q	Accuracy Percentage
1	500	30	0,5	4,5	0,4	0,2	90,67
2	500	50	2	5	0,5	0,5	99,33
3	500	50	2	6	0,6	0,5	96,67
4	500	100	2,5	4,5	0,5	0,2	94,67
5	500	100	2,5	6	0,6	0,5	94,67
6	500	100	2	6	0,5	0,5	90,67
7	500	150	2	6	0,5	0,5	94,67
8	500	150	2,5	6	0,6	0,2	90,67

References

[1] M. Dorigo, V. Maniezzo and A. Coloni, The ant system: optimization by a colony of cooperating agents, IEEE Trans Syst Man, Cybernetics- Part B, **26 (1)** (1996) 29-41, doi:10.1109/3477.484436.

[2] M. Dorigo and L. M. Gambardella, Ant colony system: A cooperative learning approach to the traveling salesman problem, IEEE Transaction on Evolutionary Computation**, 1 (1)** (1997) 53-66, doi:10.1016/j.eswa.2008.01.066.

[3] M. Dorigo and T. Stutzle, Ant Colony Optimization. A Bradford Book. London: The MIT Press Cambridge, **1 (4)** (2004) 28-39, doi:10.1109/MCI.2006.329691.

[4] X. Liu, Ant colony optimization based on dynamical pheromones for clustering analysis. *International Journal of Hybrid Information Technology*. **7 (2)** (2014) 29-38, doi:10.14257/ijhit.2014.7.2.04.

[5] A. Mohan, G. Remya, A parallel implementation of ant colony optimization for TSP based on MapReduce framework. *International Journal of Computer Applications*. **88 (8)** (2014) 9-12.

[6] MA. Muñoz, JA. López, EF. Caicedo, Swarm intelligence: problem-solving societies (a review), **28 (2)** (2008) 119-130.

[7] K. Shyamala, SS. Prabha, An ant colony optimization approach to solve traveling salesman problem. *International Journal on Recent and Innovation Trends in Computing and Communication*. **2 (12)** (2014) 3966–3971.

[8] W. N. Tanjung, Implementasi Algoritma K-Ant Colony Optimization untuk menyelesaikan masalah alokasi-alokasi. Tesis, Universitas Indonesia (2012).

[9] A. Verma, A. Kumar. Performance enhancement of K-means clustering algorithms for high dimensional data sets. *International Journal of Advanced Research in Computer Science and Software Engineering*. **4 (1)** (2014) 791-796.

[10] IMK. Wicaksana, IM. Widiartha, Metode ant colony optimization pada metode K-harmonic means untuk klasterisasi data. *Jurnal Ilmu Komputer*. **5 (1)** (2012) 55-62.