Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 6, No. 3, Januari 2012 **ISSN 0216 – 0544**

ADAPTIVE DATA CLUSTERING METHOD BASED ON ARTIFICIAL BEE COLONY AND K-HARMONIC MEANS

[a] I Made Widiartha, [b] Agus Zainal Arifin, [c] Anny Yuniarti
[a]Jurusan Ilmu Komputer, FMIPA, Universitas Udayana
Kampus Bukit, Gedung BJ Lt.I, Jimbaran Bali,
[b,c] Informatics Department, Faculty of Information Technology
Institute of Technology Sepuluh Nopember
E-Mail: [a]imdewidiartha@cs.unud.ac.id

Abstrak

Berbagai metode telah dibuat untuk dapat melakukan klasterisasi data. Salah satu metode tersebut adalah *K-Harmonic Means Clustering* (KHM). KHM merupakan metode klasterisasi data yang menyempurnakan *K-Means Clustering* (KM). Metode KHM telah mampu mengurangi permasalahan KM dalam hal sensitifitas pada inisialisasi titik pusat awal, meskipun demikian dalam KHM masih terdapat kemungkinan solusi yang dihasilkan merupakan suatu lokal optimal. Permasalahan lokal optimal ini dapat diatasi dengan memanfaatkan suatu metode yang memiliki karakteristik pencarian solusi global ke dalam metode KHM. *Artificial Bee Colony* (ABC) merupakan suatu metode *swarm* yang berbasis pada perilaku mencari makan dari koloni lebah madu yang memiliki karakteristik untuk menghindari kemungkinan konvergensi terhadap lokal optimal. Dalam penelitian ini diusulkan sebuah metode baru untuk klasterisasi data yang berbasis pada metode ABC dan KHM (ABC-KHM). Kinerja metode ABC-KHM ini telah dibandingkan dengan metode KHM dan ABC dengan memanfaatkan lima dataset. Dari hasil penelitian didapatkan hasil dimana metode ABC-KHM ini telah berhasil mengoptimalkan posisi titik pusat klaster KHM yang mengarahkan hasil klaster menuju suatu solusi global.

Kata kunci: *K-Means Clustering, K-Harmonic Means Clustering, Artificial Bee Colony,* ABC-KHM.

Abstract

Various methods have been made to cluster the data. One such method is K-Harmonic Means Clustering (KHM). KHM *is a clustering method that improves K-Means Clustering* (KM). KHM *method was able to reduce the problem of* KM *in terms of sensitivity to the initialization of the initial center point nevertheless there is still a possibility that the result of* KHM *is a local optimum. The local optimal problem can be solved by utilizing a method that has characteristic of a global search into* KHM *method. Artificial Bee Colony* (ABC) *is a swarm method based on foraging behavior of honey bee colony that has characteristics to avoid the possibility of local optimum convergence. In this research, a new method for data clustering based on* ABC *and* KHM (ABC-KHM) *is proposed. The performance* ABC-KHM *method has been compared with ABC and* KHM *by using five datasets. The results show that* ABC-KHM *method is able to optimize the position of the cluster center and directs the center to a global solution.*

Key words: K-Means Clustering, K-Harmonic Means Clustering, Artificial Bee Colony, ABC-KHM.

INTRODUCTION

Data clustering is a process to classify the data into several clusters/groups so that the data in a cluster has a maximum level of similarity, and the data between clusters has a resemblance to a minimum [1]. Clustering methods can generally be divided into two, namely hierarchical clustering and partitional clustering [1].

One partitional clustering method which is very popular is K-Means Clustering (KM). This method is widely used as a simple implementation, it can handle large amounts of data, and the relatively short time. However, if considered at the stage of KM to get the final cluster, there are weaknesses which the accuracy of the results of a cluster depends on the determination of the starting point of the cluster center. Besides that, a random starting point of centroid can cause the cluster results converge to local optimal [2].

To overcome the problem that occurs in initial cluster centers, Zhang, Hsu, and Dayal [3] propose a new method called K-Harmonic Means (KHM) is then modified by Hammerly and Elkan [4]. The purpose of this method is to minimize the harmonic mean of all points in the data set around the central cluster. Although the KHM can reduce the initial problem, but the KHM still have the possibility of optimal local problem [5]. To solve local optimal issues we need a method that has the ability to avoid the possibility of convergence to local optimum.

Artificial Bee Colony (ABC) is a method adopted foraging behavior of honey bee colonies. In this method there are three types of bees, they are employed bee, onlooker bee, and scout [6]. ABC method has been shown to have the ability to deal with local issues and the performance of ABC is better or equal to other similar methods such as Genetic Algorithm, Particel Swarm Optimization, Differential Evolution, and Evolution Strategies [7].

The characteristics of bee on the ABC method for finding solutions of Clustering can be used to help KHM out of local optimal problem. This research aims to produce a new method based on ABC method and the KHM (ABC-KHM). ABC-KHM method is expected to optimize the position of the cluster center that lead to global optimal solution.

K-HARMONIC MEANS CLUSTERING

K-Harmonic Means Clustering (KHM) is a method introduced by Zhang, Hsu, and Dayal which is made to overcome the existing problems in K-Means Clustering [3]. KHM is one example of center-based cluster, and is a method in which the clusters are formed by improving the location of center point of each cluster iteratively. In KHM, the objective function value generated by the search for total harmonic average of all data points to the distance between each data point to the whole point of the existing cluster centers [3]. This is different from the K-Means where the objective function is obtained from the total distance of all data to a cluster central point. Harmonic mean is defined as Equation (1).

$$HA(\{a_i \mid i = 1,...,K\}) = \frac{K}{\sum_{i=1}^{K} \frac{1}{a_i}} \quad (1)$$

In a harmonic function,: in harmonic function, if in $a_1, a_2 \ldots, a_n$ has one small value member then the average value of harmonic is small, otherwise if there is no small value member then the average of harmonic value is large [8]. Harmonic mean is very sensitive to the circumstances in which there are two or more adjacent central point. This method naturally places one or more central point to the area of data points away from the central points that existed before. This will make the objective function will be smaller [8]. The KHM method steps are as follows [5],

1. Initialize the initial cluster centers randomly.
2. Calculate the value of objective function by Equation (2), where p is the input parameter. p value is usually ≥ 2.

$$KHM(X,C) = \sum_{i=1}^{N} \frac{K}{\sum_{l=1}^{K} \frac{1}{\| x_i - c_l \|^p}} \quad (2)$$

3. For each data x_i, calculate the membership value $m(c_l \mid x_i)$ for each cluster center c_l by Equation (3).

$$m(c_l \mid x_i) = \frac{\| x_i - c_l \|^{-p-2}}{\sum_{l=1}^{k} \| x_i - c_l \|^{-p-2}} \quad (3)$$

4. For each data x_i, compute the weights $w(x_i)$ based on Equation (4)

$$w(x_i) = \frac{\sum_{l=1}^{K} \| x_i - c_l \|^{-p-2}}{\left(\sum_{l=1}^{K} \| x_i - c_l \|^{-p} \right)^2} \quad (4)$$

5. For each point of the center c_l, recalculate for the position of cluster center points of all data based on their membership and weight values as show in Equation (5).

$$c_l = \frac{\sum_{i=1}^{N} m(c_l \mid x_i).w(x_i).x_i}{\sum_{i=1}^{N} m(c_l \mid x_i).w(x_i)} \quad (5)$$

6. Repeat steps 2 through 5 until there is no significant change for the objective function value.
7. Set the membership of data x_i in a cluster with cluster center c_l according to the membership value x_i of c_l.

x_i is a member of a cluster with cluster center c_l if its membership value $m(c_l \mid x_i)$ is the largest compared to the value of membership to another cluster. Membership value $m(c_l \mid x_i)$ in the KHM method is very useful when the data are not well separated [9].

ARTIFICIAL BEE COLONY

Artificial Bee Colony (ABC) is a method that adopts foraging behavior of honey bee colonies to solve optimization problems of multidimensional and multimodal [7]. This method was introduced by Karaboga in 2005[6]. Artificial bee colony consists of three groups of employed bee, onlooker bee, and scout bee. Bees that wait in the dance area to make a decision in choosing the food source is called onlooker bee, and bees that go to the source of the food is called employed bee. While the bees that are doing a random search is called scout bee. Half of the first part of the bee colony consists of employed bee, and half of the second part includes the onlooker bee. This ABC Method can be described as in Figure 1.

The first step in the ABC method is sending employed bee (as a scout) in the search area to generate the initial population is randomly distributed. Each solution x_i where i = 1, 2, ..., *SN* (number of food source solution) is a *D*-dimensional vector. *D* is the number of optimized parameters. After the initialization phase is complete the determination of the population of the position of the next solution obtained through repeated cycles, C = 1, 2, ... , *MCN*. At the end of each cycle, the employed bee will do the calculation of fitness value (the value of nectar) from the resulting solution and the employed bee get nectar and share information about their position with a onlooker bee in the dancing area. Fitness values can be found by Equation (6).

$$fit_i = \frac{1}{1 + f_i} \quad (6)$$

F_i is the variable cost function value of solution *i*. Onlooker bee evaluates the information taken from all the employed bee and a food source with the probability of selecting the appropriate number of nektarnya. Such case the employed bee, onlooker bee also produces modifications in the position of food sources (solutions) in memory and check the amount of nectar from a new food source candidate.

If the value is higher than ever nectar, the bees will remember the new position and forget about the old position.

The bees choose a food source based on the probability p_i and roulette wheel selection method [10]. The p_i value is calculated by the Equation (7).

$$p_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \quad (7)$$

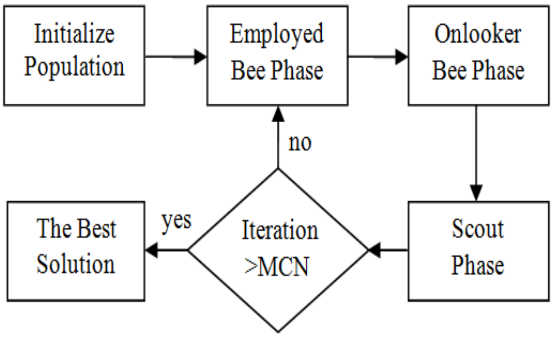


Figure 1. ABC Method.

In generating new candidate food position, the ABC uses Equation (8).

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \qquad (8)$$

The value of k {1, 2, ..., SN} and j {1, 2, .., D} is the index which is chosen randomly. Although k is determined randomly, but k must be different from i. ϕ_{ij} is a random number between [-1,1], which controls the production of neighbor's food sources positions around x_{ij}.

The food source which is left by the employed bee, is replaced with a new food source by the scout bee. In the ABC method, if a position could not be further enhanced through a number of cycles (cycles) that have been determined (limit), then the food source is assumed to be abandoned. This is simulated by generating a new food source positions randomly to replace the abandoned food source. Eg food sources left are x_i and j {1, 2, ..., D}, then the scout bee will look for new food sources to be replaced by x_i. This operation is performed using Equation (9).

$$x_i^j = x_{\min}^j + \text{rand}[0,1](x_{\max}^j - x_{\min}^j) \qquad (9)$$

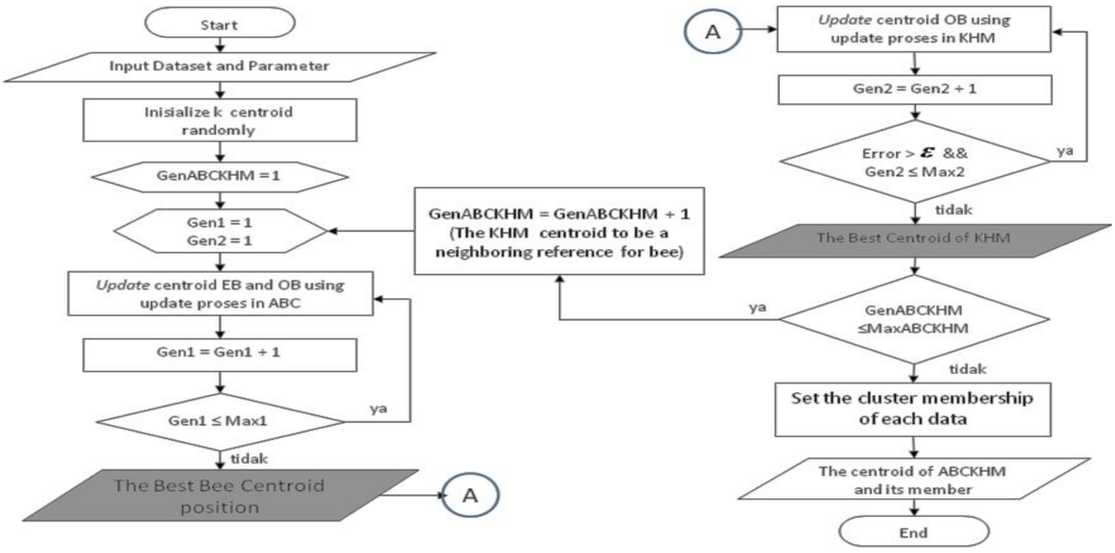
After each candidate position v_{ij} food sources are produced and evaluated by the employed bee, compared to the value of x_{ij} fitnesnya. If a new food source of nectar have the same or better than the old sources, then the source of the old will be replaced with new ones in memory, otherwise it is long maintained. In other words, rnekanisme greedy selection is used as a selection operation between the current source of food and the old food sources.

ARTIFICIAL BEE COLONY-K-HARMONIC MEANS CLUSTERING (ABC-KHM)

This study proposed a new method to process data clustering namely ABC-KHM. This method produced through hybridization between the ABC method and KHM method. The underlying done this hybridization is still focused on the presence of weakness in the KHM method for addressing issues of convergence to local optimum. These weaknesses arise from initial point of a random initial cluster centers. One way to overcome this drawback is to use an algorithm/method that has a global solution to the KHM method.

Various studies on the swarm method to obtain a global solution of optimization problems have been carried out. One such study conducted by Karaboga [6] which produced the ABC method. ABC method is a method of swarm that has ability to search for a global solution [7].

Numerous studies have been conducted to look at the performance of the ABC method, one of which is the study involving 25 benchmark function. The results of this study show the ABC method is able to surpass or equal to other methods [7]. This is the rationale for the use of bee behavior that existed at the ABC method to be applied in ABC-KHM method.



Start
Input Dataset and Parameter
Inisialize k centroid randomly
GenABCKHM = 1
Gen1 = 1
Gen2 = 1
Update centroid EB and OB using update proses in ABC
Gen1 = Gen1 + 1
Gen1 ≤ Max1
ya
tidak
The Best Bee Centroid position
A
GenABCKHM = GenABCKHM + 1
(The KHM centroid to be a neighboring reference for bee)
A
Update centroid OB using update proses in KHM
Gen2 = Gen2 + 1
Error > ε && Gen2 ≤ Max2
ya
tidak
The Best Centroid of KHM
GenABCKHM ≤MaxABCKHM
ya
tidak
Set the cluster membership of each data
The centroid of ABCKHM and its member
End

Figure 2. ABC-KHM Method.

In ABC-KHM method, the cluster is obtained by utilizing the reciprocal relationship between the two methods, namely ABC and KHM. The resulting position of the center point on the phase of the bees will be optimized with the phase of update process on the KHM method. The results obtained from the central point of KHM phase will be utilized in the next phase of the bee as a reference neighbor of the employed bee to do the exploration in the search space.

In implementing of ABC-KHM method, there are several variables that are used to limit the amount of each phase are available. Variable *Max1* is used to limit the iteration number of the search by the bees. If the value of the variable *Gen1* already exceeded the value of *Max1* then the process to find the center point of the bees are stopped. The results from the central point of this stage will be the focal point early in the next stage, namely stage KHM.

Iterations stage in KHM is limited by the two criterions. First, the iteration is stopped when the difference in the position of the center point between iterations is smaller than the threshold. The second, *Gen2* variable value has exceeded the limit value iteration KHM (*Max2*). Phase of the bees and the KHM will be carried out continuously until the value of the variable *GenABCKHM* already exceeded the maximum iteration is *MaxABCKHM*. ABC-KHM method can be described as in Figure 2.

If viewed from two sides hybridization of the ABC and KHM, the search process in ABC-KHM method is the incorporation of the advantages of both methods. The first thing is, the ABC method searches the optimal point by randomizing central points, this will allow the center in the end of iteration limit has not been at the optimum point, so ABC-KHM utilizes KHM update phase to further optimize the center position of the data. The second thing is, the weakness of the KHM that converges to local optimal clustering can be avoided by the use phase of the bees that has the characteristics of global search.

Data

Dataset used in this study consisted of the dataset Iris, Wisconsin Breast Cancer (Cancer), Contraceptive Method Choice (CMC), Glass, and Wine. The data in this study was taken from the UCI Machine Learning Repository (ftp://ftp.ics.uci.edu./pub/machine-learning-databases/).

Information on the number of features, classes, and data are listed in Table 1. In this study, 80% of the data will be used as training data and the rest are used as data testing. Training data is used to view the performance of the three methods of doing data clustering. Performance appraisal is viewed from three viewpoints, namely the objective function value of KHM (*X*,*C*), F-Measure, and the execution time (running time). The data used for testing only see the external correlation (class label) that is how the results of testing data classification using the cluster center by using training data.

RESULT AND DISCUSSION

In this study, the parameters value for ABC method refers to the parameter values used by Zhang. Limit parameters is 10 and the number of MCN is 2000 [11]. For the ABC-KHM method, the value of Limit, Max1, Max2, and Max3 are obtained by testing the values of this parameter. From the results that have been obtained, it was found that the best results obtained by using the *Limit* = 3, Max_1=20, Max_2 = 10, and Max_3 = 20.

Several scenarios are done to obtain the performance of these three methods by using the three benchmarks. This scenario is created by using different objective function. The difference lies in parameter *p* of objective function. The parameter *p* on the KHM method is a key parameter in the objective function [5]. This became the basis for scenarios conducted on the scores of different *p*.

Table 1. The division of DataSet.

Dataset	Fitur	Kelas	Jumlah Data	
			Training	Testing
Iris	4	3	120	30
Cancer	9	2	547	136
CMC	9	3	1179	294
Glass	9	6	172	42
Wine	13	3	143	35

Table 2. Average and Standard Deviation Test Results with *p* = 2.

Dataset	Measurement	Objective Function			F-measure			Time		
		KHM	ABC	ABC-KHM	KHM	ABC	ABC-KHM	KHM	ABC	ABC-KHM
Iris	Mean	144,456	156,329	144,450	0,9021	0,8862	0,9021	0,12	9,38	4,53
	Std. Dev.	0,001	8,43	0	0	0,03	0	0,02	1,21	0,08
Cancer	Mean	22.955,649	37.246,522	22.955,648	0,9510	0,9186	0,9510	0,14	17,05	5,63
	Std. Dev.	0	4900,95	0	0	0,027	0	0,037	0,917	0,153
Cmc	Mean	918,585	1044,602	918,570	0,3830	0,3811	0,3843	3,06	43,23	18,16
	Std. Dev.	1,589	83,887	1,583	0,006	0,017	0,010	0,255	2,448	0,230
Glass	Mean	32,0980	33,3767	31,9209	0,4269	0,4319	0,4395	0,29	45,52	16,76
	Std. Dev.	0,157	0,489	0,089	0,010	0,010	0,025	0,047	0,941	0,690
Wine	Mean	64,9765	73,3712	64,9743	0,9276	0,8275	0,9324	0,07	21,04	8,26
	Std. Dev.	0,001	2,155	0	0,004	0,093	0	0,009	0,380	0,183

Table 3. Average and Standard Deviation Test Results with *p* = 3.

Dataset	Measurement	Objective Function			F-measure			Time		
		KHM	ABC	ABC-KHM	KHM	ABC	ABC-KHM	KHM	ABC	ABC-KHM
Iris	Mean	97,7489	114,3611	97,7391	0,9107	0,8846	0,9107	0,07	13,99	5,66
	Std. Dev.	0,009	13,673	0	0	0,019	0	0,011	13,99	0,907
Cancer	Mean	91.294,7446	159.383,0011	91.294,7337	0,9569	0,9175	0,9569	0,30	35,52	17,67
	Std. Dev.	0,002	32671,729	0,014	0	0,033	0	0,019	0,843	0,936
Cmc	Mean	479,3195	525,6812	476,0932	0,3853	0,3693	0,3854	1,01	88,64	53,48
	Std. Dev.	0,043	63,289	2,916	0	0,031	0	0,173	1,373	1,119
Glass	Mean	10,1453	10,1165	9,4951	0,3673	0,3917	0,3825	0,25	93,18	40,04
	Std. Dev.	0,426	0,341	0,197	0,019	0,045	0,030	0,012	1,443	0,355
Wine	Mean	18,6671	22,0259	18,6543	0,9111	0,8683	0,9193	0,13	31,63	14,77
	Std. Dev.	0,004	1,014	0,003	0,003	0,048	0,005	0,017	1,339	0,400

In this study, there are three scenarios the value of p. They are *p* = 2, *p* = 3 and *p* = 4. This study used *F*-measure assessment in terms of external assessment (class labels). *F*-measure values obtained from Equation (10) [12].

$$F(i,j) = \frac{(b^2+1).(p(i,j).r(i,j))}{b^2.p(i,j)+r(i,j)} \qquad (10)$$

$p(i,j) = n_{ij}/n_j$ and $r(i,j) = n_{ij}/n_i$ where n_i is the amount of data from class *i* that is expected as a result of the query, n_j is the amount of data generated from cluster *j* by the query, and n_{ij} is the number of elements of class *i* in cluster *j*. To obtain a balanced weighting between precision and recall the value of *b* = 1 is used in calculating the value of *F*-measure [1].

To obtain the final conclusion about existing methods, the clustering experiment conducted 10 times for each scenario created. Conclusion The performance of the method will be obtained through the average (mean) and standard deviation of 10 experiments. Table 2 to Table 4 is the average and standard deviation of experiments have been conducted, results of studies involving five data sets and objective functions of these three scenarios showed that ABC-KHM method is generally superior than the other two methods. In terms of performance measurement through the objective function value, the ABC-KHM superior in every scenario.

The result of objective function value of ABC-KHM method are also relatively stable. This is evidenced by the acquisition value of the relatively small standard deviation on the ABC-KHM. Comparison of the cluster with the class label using the F-measure can also be seen that the ABC-KHM dominate in every scenario. In terms of measuring the time it takes we got the result that the KHM method requires the fastest time, followed by ABC-KHM method and the last is ABC method. The time difference between the process and the ABC-KHM KHM relatively large enough that it requires further research to improve the efficiency of the time required in the case of ABC-KHM data clustering.

The next section will be displayed on the table the results of research that has been done. As discussed in the previous section that the data used for testing only look externally (class label) how the results of testing data classification of each clustering method. Mechanical testing of data classification is to compare the data with the distance between the centers of existing clusters.

Testing data which has the shortest distance to a point in the center of the data is classified into the nearest classroom. To shorten the name of the method in each column of the table in the presentation of the results of the classification error method name is replaced with the sequence I = KHM, II = ABC, ABC-III = KHM. Table 5 to Table 7 is the result of misclassification of the three methods from ten trials conducted in each scenario.

Table 5 to Table 7 shows there is certainty that the value of the objective function and F-Measure which are more optimal will give a minimum misclassification. This of course can occur due to the different between data clustering and data classification process, where in obtaining the optimal value of a clustering process there is no use of the class label of the data but only by using the values of any existing features. This can make the cluster center position is worse to have a more minimal classification error.

Table 4. Average and Standard Deviation Test Results with $p = 4$.

Dataset	Measurement	Objective Function			F-measure			Time		
		KHM	ABC	ABC-KHM	KHM	ABC	ABC-KHM	KHM	ABC	ABC-KHM
Iris	Mean	100,7958	91,3177	76,4438	0,9014	0,8757	0,9014	0,29	13,08	6,82
	Std. Dev.	2,614	11,821	4,709	0	0,021	0	0,014	0,518	0,477
Cancer	Mean	432.161,612	725.778,9403	432.106,889	0,9550	0,8897	0,9550	0,78	38,92	18,21
	Std. Dev.	0,003	126.546,714	25,558	0	0,038	0	0,054	2,979	0,162
Cmc	Mean	266,6825	325,0230	258,2306	0,3919	0,3431	0,3989	1,47	88,92	57,01
	Std. Dev.	0,155	38,746	3,233	0,001	0,024	0,009	0,335	1,347	0,778
Glass	Mean	3,2938	3,2969	3,1089	0,3191	0,3508	0,2290	0,29	94,38	39,69
	Std. Dev.	0,121	0,129	0,048	0,060	0,056	0,034	0,041	3,288	0,633
Wine	Mean	6,3107	7,5468	6,2962	0,8946	0,8693	0,9003	0,11	31,02	15,18
	Std. Dev.	0,004	0,467	0,008	0,006	0,029	0,003	0,014	0,564	0,182

Table 5. Number of Classification Errors with the objective function *p* = 2.

Eksp.	Iris			Cancer			Cmc			Glass			Wine		
	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
1	4	2	4	2	3	2	233	225	233	29	27	27	3	2	3
2	4	4	4	2	5	2	233	228	233	30	26	26	3	3	3
3	4	6	4	2	2	2	233	224	233	30	29	25	3	2	3
4	4	4	4	2	6	2	233	239	233	30	23	26	3	14	3
5	4	2	4	2	8	2	211	234	233	29	26	26	3	3	3
6	4	10	4	2	2	2	233	224	233	30	26	26	3	3	3
7	4	4	4	2	12	2	233	223	233	24	23	25	3	7	3
8	4	6	4	2	11	2	233	236	211	30	26	26	3	3	3
9	4	5	4	2	5	2	233	228	233	29	26	25	3	2	3
10	4	4	4	2	5	2	233	239	233	25	29	26	3	5	3

Table 6. Number of Classification Errors with The Objective Function *p* = 3.

Eksp.	Iris			Cancer			Cmc			Glass			Wine		
	I	II	III	I	II	III	I	II	III	I	II	III	I	I	III
1	4	13	4	2	9	2	232	218	232	35	38	36	2	4	2
2	4	4	4	2	6	2	232	231	231	34	38	32	2	3	2
3	4	8	4	2	12	2	232	237	232	36	38	32	2	2	2
4	4	2	4	2	21	2	232	228	233	29	38	36	2	8	2
5	4	2	4	2	7	2	232	238	231	34	38	36	2	3	3
6	4	4	4	2	18	2	232	228	231	29	31	32	2	2	2
7	4	8	4	2	8	2	232	240	230	34	31	32	2	4	2
8	4	2	4	2	12	2	232	217	232	34	38	36	2	2	2
9	4	4	4	2	4	2	232	233	232	35	31	36	2	3	3
10	4	3	4	2	4	2	232	226	230	34	30	32	2	2	2

Table 7. Number of Classification Errors with The Objective Function p = 4.

Eksp.	Iris			Cancer			Cmc			Glass			Wine		
	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
1	4	4	4	3	7	3	234	218	228	38	32	38	2	2	2
2	4	3	4	3	13	3	234	222	233	38	32	38	2	4	2
3	4	8	4	3	20	3	234	237	233	40	31	38	2	2	2
4	4	4	4	3	6	3	234	220	228	38	30	40	2	2	2
5	4	2	4	3	4	3	234	240	235	38	39	38	2	3	2
6	4	5	4	3	32	3	234	218	234	40	31	38	2	2	2
7	4	4	4	3	36	3	234	209	233	38	31	38	2	2	2
8	4	2	4	3	20	3	234	218	219	36	30	38	2	2	2
9	4	5	4	3	4	3	234	235	233	38	32	38	2	7	2
10	4	0	4	3	6	3	234	218	234	38	40	40	2	2	2

CONCLUSION

This study proposed a new method for data clustering based on the method of ABC and KHM namely ABC-KHM. ABC-KHM method has been successful in optimizing the position of the cluster center and directs the center to a global solution.. This is evidenced by the results of studies that show the objective function values of the ABC-KHM method is the smallest than two other methods, namely ABC, and KHM. In terms of external assessment of the cluster using the F-measure, ABC-KHM method has also shown its dominance of the two other methods.

The time performance of ABC-KHM method is between the two methods. ABC-KHM method requires a relatively much longer when compared with the KHM method, so this is a weakness of the ABC-KHM. The running time optimization of ABC-KHM method will be the focus of future studies.

REFERENCES

[1] Tan PN, StainbachM, and Kumar V. *Introduction to Data Mining 4th edition*. Pearson Addison Wesley. 2006

[2] Pen JM, Lozano JA, and Larranaga P. An Empirical Comparison of Four Initialization Methods for The K-Means Algorithm. *Pattern Recognition Letters*. 20:1027-1040. 1999.

[3] Zhang B, Hsu M, and Dayal U. K-Harmonic Means – A Data Clustering Algorithm. Technical Report HPL-1999-124. Hewlett-Packard Laboratories. 1999.

[4] Hammerly G and Elkan C. Alternatives to The K-Means Algorithm that Find Better Clusterings. *Proceedings of the 11th international conference on information and knowledge management*. 600–607. 2002.

[5] Yang F, Sun T, and Zhang C. An Efficient Hybrid Data Clustering Method Based on K-Harmonic Means and Particle Swarm Optimization. *Expert Systems with Applications*. 36:9847–9852. 2009.

[6] Karaboga D. An Idea Based on Honey Bee Swarm for Numerical Optimization. Technical Report-TR06, Erciyes University. Engineering Faculty, Computer Engineering Department. 2005.

[7] Karaboga D and Akay B. A Comparative Study of Artificial Bee Colony Algorithm. *Applied Mathematics and Computation*. 214:108–132. 2009.

[8] Gungor Z. And Unler A. K-Harmonic Means Data Clustering with Simulated Annealing Heuristic. *Applied Mathematics and Computation*. 184:199–209. 2007.

[9] Jiang H, Yi S, Li J, Yang F., and Hu X. Ant Clustering Algorithm with K-Harmonic Means Clustering. *Expert Systems with Applications*. 37:8679–8684. 2010.

[10] Karaboga D and Basturk B. On The Performance of Artificial Bee Colony ABC Algorithm. *Applied Soft Computing*. 8:687–697. 2008.

[11] Zhang C, Ouyang D, and Ning J. An Artificial Bee colony Approach for Clustering. *Expert Systems with Applications*. 37:4761–4767. 2009.

[12] Dalli A. Adaptation of the F-Measure to cluster-based Lexicon quality evaluation. In EACL. Budapest. 2003.