A New Hybrid Approach for Data Clustering

Danial Yazdani
Islamic azad university, Shirvan Branch, Iran
d.yazdani@iau-shirvan.ac.ir

Sara Golyari
Islamic azad university, Shirvan Branch, Iran
S.golyari@iau-shirvan.ac.ir

Mohammad Reza Meybodi
Department of Computer Engineering and Information Technology
Amirkabir University of Technology
Tehran, Iran
mmeybodi@aut.ac.ir

Abstract— **Data clustering has been applied in multiple fields such as machine learning, data mining, wireless sensor networks and pattern recognition. One of the most famous clustering approaches is K-means which effectively has been used in many clustering problems, but this algorithm has some problems such as local optimal convergence and initial point sensitivity. Artificial fishes swarm algorithm (AFSA) is one of the swarm intelligent algorithms and its major application is in solving optimization problems. Of its characteristics, it can refer to high convergent rate and insensitivity to initial values. In this paper a hybrid clustering method based on artificial fishes swarm algorithm and K-means so called KAFSA is proposed. In the proposed algorithm, K-means algorithm is used as one of the behaviors of artificial fishes in AFSA. The proposed algorithm has been tested on five data sets and its efficiency was compared with particle swarm optimization (PSO), K-means and standard AFSA algorithms. Experimental results showed that proposed approach has suitable and acceptable efficacy in data clustering.**

Keywords; Artificial fishes swarm algorithm, data clustering, optimization, K-means, PSO.

I. INTRODUCTION

Data clustering has various applications in data mining [1], wireless sensor networks [2], pattern recognition [3], data compression [4], machine learning [5] and so on. Clustering importance in different sciences and also the type of used data, clustering rate, accuracy and many other parameters have caused in introduction of various methods and algorithms of data clustering [6][7][8]. Clustering is a sorting technique without supervision in which data set that usually is vectors in multidimensional spaces, based on a similarity or dissimilarity criterion, are divided into specified clusters. When the number of clusters is K and we have N data of M dimensions, data clustering algorithm would allocate each data to one of the clusters, based on the similarity among allocated data to one cluster rather than other data in other clusters. K-means algorithm is one of the most famous clustering data which is applied in many problems [9][10][11]. K-means algorithm is starts with K random cluster center and divides a collection of objects into K subsets. This method is on the most popular and most used clustering techniques, since it is easily understandable and can be performed and has linear time complexity. But K-means algorithm has various fundamental problems. Some of these problems are such as being trapped in local optimums and being sensitive to initial values of clusters center. Data clustering belong to NP problems. Finding solution of NP problems is difficult. Algorithms such as swarm intelligent algorithms have solved this problem to some extent. Solutions are found by these algorithms which are close to the answer. One of the most famous and applicable swarm intelligent algorithms is PSO that was represented by Kennedy and Eberhart in 1995 [12]. This algorithm is an efficient technique for solving optimization problems which works on probability rules and population. One of the other swarm intelligent algorithms which have been represented so far is artificial fish swarm algorithm (AFSA). AFSA is some type of inspired algorithms from nature which was represented by Dr. Li Xiao Lie in 2002[13]. This algorithm is a technique based on swarm behaviors that has been inspired from social behaviors of fish swarm in nature. This algorithm has high convergence rate, insensitive to initial values, flexibility, and high fault tolerance. This algorithm has been used in applications of optimization such as data clustering [14][15], PID control [16], data mining [17], DNA sequence encoding [18] and so on. This algorithm has major difference with PSO algorithm, in fact the structure and the type of performance of AFSA is completely different with PSO. The major difference between these two algorithms is that particles in PSO depend on their past for their next movement, indeed for next movement, each particle from its previous velocity, uses the best individual experience and the best group experience, but in AFSA artificial fishes perform completely independent from past for the next movement and the next movement only depends on the current situation of artificial fish and other artificial fishes of the swarm. Since AFSA has much more complexity than similar algorithms such as PSO and its efficiency was not better than them, was not much considered. In this paper it is tried to obtain comparable results with other algorithms, by presenting a hybrid approach in which AFSA has been used.

In this paper, a new hybrid clustering approach based on AFSA and K-means, called KAFSA, is proposed. In KAFSA, K-means is used as a behavior in AFSA, and after performing AFSA basic behaviors in each iteration, one iteration of K-means is performed on a percentage of randomly selected artificial fishes. Thus, in proposed algorithm, weaknesses of k-means were removed, in fact since in each iteration, K-means algorithm is performed on some artificial fishes that each of them is representative of K cluster center, so its sensitivity to initial point has been removed. Also by performing AFSA behaviors and changing the position of cluster center,

preceding convergence is avoided. experimental results PSO [19], K-means, standard AFSA and proposed KAFSA method algorithms on standard datasets of Wine, Irish, Pima [20] and two artificial data sets of 2-dimensions and 3-dimensions show that the proposed algorithm has a higher efficacy than the other tested algorithms. The following of the paper is as below: In section two, AFSA is described. In section 3, k-means is described. Section 4 is dedicated to the proposed algorithm description. In section 5, experimental results are analyzed and the last section argues the conclusions.

II. ARTIFICIAL FISH SWARM ALGORITHM

In water world, fishes can find areas that have more foods, which it is done with individual or swamp search of fishes. According to this characteristic, artificial fish model is represented by prey search, swarm movement and following behaviors which the problem space is searched by them. The environment which the artificial fish lines in, substantially is solution space and other artificial fishes domain. Food consistence degree in water area is AFSA objective function. Finally, artificial fishes reach to a point which its food consistence degree is maximum (global optimum).

As in figure 1 is observed, artificial fish perceive external concepts with sense of sight. Current situation of artificial fish is shown by vector $X=(X_1,X_2,\ldots,X_n)$. The visual is equal to sight field of artificial fish and X_v is a position in *visual* where the artificial fish wants to go there. Then if X_v has better food consistence than current situation, we go on step toward it which causes change in artificial fish situation from X to X_{next} , but if the current position is better than X_v, we continue browsing in visual area. The step is equal to maximum length step, the distance between two artificial fishes which are in X_i and X_j positions is shown by $d_{ij}=\| X_i-X_j\|$ (Euclidean distance).

Artificial fish model consists of two parts of variables and functions that variables include X (current artificial fish situation), step (maximum length step), visual (sight field), try-number (maximum test interactions and tried) and crowd factor δ ($0<\delta<1$). Also functions consist of prey behavior, fre move behavior, swarm behavior and follow behavior.

In each step of optimization process, artificial fish look for locations with better fitness values in problem search space by performing these four behaviors based on algorithm procedure[14][15][17][18].

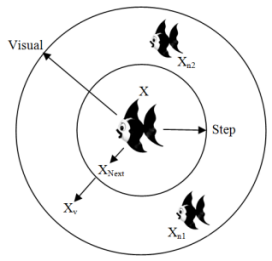

Figure 1. Artificial Fish and the environment around it.

III. CLUSTERING AND K-MEANS ALGORITHM

Clustering in D-dimensional Euclidean space is a process in which a set of N member based on a similarity criterion is divided into K groups or clusters. Various clustering methods are represented so far. The base of clustering algorithms is measuring the similarity between data and it is determined how much similar these two data vectors are by a function. K-means algorithm is one of the oldest and most famous clustering methods. This method sorts data vectors in D-dimensional space in clusters which their number was determined before, this clustering is based on Euclidean distance between data and cluster center which are considered as similarity criterion.

Euclidean distance between data vectors of a cluster with the center of that cluster is less than their Euclidean distance with other cluster centers. Standard k-means algorithm is as below:

- Initial positions of K cluster centers are determined randomly. Following phases are repeated:

a) For each data vector: the vector is allocated to a cluster which its Euclidean distance from its center is less than the other cluster centers. The distance to cluster center is calculated by:

$$Dis(X_p.Z_j)=\sqrt{\sum_{i=1}^{D}(X_{pi}-Z_{ji})^2} \quad (1)$$

- In relation (1), Xp specifies p-th data vectors, Zj specifies j-th cluster center and D is the dimension of data and cluster center.

b) Cluster center are updated by :

$$Z_j=\frac{1}{n_j}\left[\sum_{\forall Xp\in Cj}X_p\right] \quad (2)$$

- In relation (2), nj is the number of data vectors corresponding to j-th cluster and Cj is a subset of the total data vectors which constitute j-th cluster and are in it.
- Phases (a) and (b) are repeated until stop criterion is satisfied [18].

IV. PROPOSED ALGORITHM

In this method, in order to find optimal values of cluster center (which their number was specified before), relation (3) is used, in fact relation (3), is the fitness function which would be optimized (minimized) and calculates Intra cluster distance.

$$J(C_1,C_2,\ldots,C_K)=\sum_{i=1}^{K}\left(\sum_{X_j\in C_i}\| Z_i-X_j\|\right) \quad (3)$$

In relation (3), the sum of total Euclidean distance of all data vectors from cluster centers which they belong to, is calculated and is added to each other. In this relation we have K cluster (C) which of N data vectors (X) is classified based on

the distance from every cluster center (Z), and is placed in one of the clusters. Thus the purpose is to determine cluster centers that minimize relation (3), so optimal cluster centers are determined. Since data is D-dimensional and there are K clusters, so each artificial fish has K×D dimension. Figure 2, shows an artificial fish vector which consists of K cluster center of D dimensions.

In the proposed algorithm, first artificial fishes are initialized randomly, thus each of artificial fishes includes K random cluster centers. Then according to the allocation of data vectors to each of the clusters in artificial fishes based on Euclidean distance and fitness function of relation (3), behaviors of AFSA are performed for artificial fishes. At the end of each iteration of algorithm performance, after performing AFSA behaviors, K-means algorithm is performed on a specified percentage of artificial fishes. In fact, k-means algorithm is performed as a behavior on some random selected artificial fishes. Thus KAFSA has both AFSA and K-means algorithms capabilities. KAFSA is not sensitive to its initial points of cluster centers, has acceptable convergence speed based on number of iterations and local optimum with more ability passes.

But AFSA has a major weakness for clustering problems. In AFSA, visual and step parameters are numerical and have one value. Data clustering is one of the problems that AFSA needs for artificial fishes with high dimensions to solve. If interval values of the problem space is same for all variables, appropriate values could be determined for step and visual parameters. But in some clustering problems, different data vectors dimensions have different intervals. For instance, suppose that i-th dimension of all data vectors is in [1, 2] internal and j-th vector of these data vectors is in [40,100]. In this case if we consider small parameters values of visual and step to be appropriate for i-th dimension, search in j-th dimension gets into trouble since the parameters values of visual and step is small, then the convergence speed decreases so much and the probability of getting trapped in local optimums increases a lot in this dimension. If we consider large value of visual and step parameters to be appropriate for j-th dimension, search in i-th dimension faces problem, because in this case the value of visual and step parameters for searching i-th dimension is so large and this leads to searching in an space out of interval values of this dimension and the probability of finding values with better fitness decreases so much in this dimension. Thus it cannot determine the value of visual and step parameters such that search are performed well in all dimensions with different intervals.

To remove this problem we consider visual and step parameters as vectors. Dimensions number of visual and step parameters equals to artificial fish dimensions number. Thus the value of visual and step parameters in i-th dimension based on interval of changes of this dimension is determined in problem space.

$$\left[Z_{1,1}, Z_{1,2}, ..., Z_{1,D}, Z_{2,1}, Z_{2,2}, ..., Z_{2,D}, ..., Z_{K,1}, Z_{K,2}, ..., Z_{K,D}\right]$$

Figure 2. Structure of an artificial fish position in clustering problem space.

V. EXPERIMENTAL RESULTS

Experiments have been performed on five data sets which among them, three real data sets consist of Iris, Pima and Wine that were selected from standard data set UCI [20] and two data sets have been provided experimentally which the characteristics of each of them are described in the following:

Iris: This data set is according to the Iris flowers recognition that has three different classes and each class consists of 50 samples. Every sample has four attributes.

Pima: This data set is allocated to recognize diabetic patients that totally has 768 samples which is classified into two classes consisting of 500 and 268 samples, respectively. Every sample in this data set has 8 attributes.

Wine: This data set is regarding to drinks recognition that totally has 178 samples classified into three different classes including 59, 71 and 48 samples, respectively. In this data set, each sample has 13 attributes.

Art 2D: This experimental data set has four classes and every class consists of 100 samples that each of which has two attributes. Sample arrangement of this data set has been performed by uniform distribution in two dimensional space. This data set samples are shown in figure 3-a. As it is observed, in this data set, there are common boundary samples.

Art 3D: In this experimental data set, there are five classes and each class has 50 samples consisting of three attributes. Different properties of distribution of each class samples are as class1 ~ Uniform (25, 40), class 2 ~ Uniform (40, 55), class 3 ~ Uniform (55, 70), class 4 ~ Uniform (70, 85) and class 5 ~ (85, 100). This data set is shown in figure 3-b.

In the performed experiments, the population in standard AFSA, proposed algorithm of KAFSA, and PSO is five times of problem space dimensions. Problem space dimensions for each data set is the number of data set classes multiplied by the number of sample attribute of that data set. Maximum iteration for clustering each data set is 10 times of problem space dimensions.

In standard AFSA and KAFSA, crowd factor value is 0.5 and maximum number of attempts (try-number) is 10. In standard AFSA visual parameter value is 20% of the variation range of values of the sample dimensions and parameter value of step is considered half of visual value. In the proposed algorithm, KAFSA, the value of visual vector parameters in each dimension is 20 percent of the variation range of values of sample vectors on that dimension and step vector parameter's value on each dimension is considered equal to half of visual value on that dimension. In PSO c1 and c2 values are considered 2 and inertia weight on each attempt is obtained by W=rand/2 + 0.5 [21]. Fitness function is equal to intra-cluster distance and is calculated by equation (10). Experiments were repeated 50 times and the best, mean and standard deviation of optimization results of intra-cluster distance and clustering error for standard AFSA, PSO, K-means and the proposed algorithm of KAFSA are shown in tables 1-5 on presented data sets. In figure 4-8, graph of fitness function average value (intra-cluster distance) during the performance of algorithms in 50 executions is presented.

As it is observed in tables 1-5, the proposed algorithm of KAFSA has more appropriate efficiency than other tested algorithms on Iris, Pima, Wine, Art 2D and Art 3D data sets. In fact, in all cases, the proposed algorithm of KAFSA has obtained more appropriate value of intra-cluster distance and error rate. Obtained standard deviation in 50 attempts of performing k-means, PSO, standard AFSA algorithms and proposed algorithm of KAFSA on tested data sets shows that proposed algorithm of KAFSA has more stability in converging to the solution and the different obtained results difference in various attempts is less than the other algorithms. In fact, in proposed algorithm of KAFSA, by adjusting visual and step parameters as vectors, the proposed algorithm by performing basic behaviors of AFSA has more ability of crossing local optimums and convergence is toward global optimum and because of k-means algorithm performance on some artificial fishes in each iterations, convergence rate, ability to cross local optimums and accuracy of proposed algorithm is acceptable. Results of k-means algorithm, for low ability of it in crossing local optimum and its high sensitivity to initial values of cluster centers are of great difference in different attempts.

In figure 4-8, mean graphs of intra-cluster distance corresponding to k-means, standard AFSA, PSO algorithms and proposed algorithm of KAFSA through 50 times of their performance are presented on Iris, Pima, Wine, Art 2D and Art 3D data sets. To show better and separate different lines in graphs, from iteration 3 of performing algorithms to the last is shown in all of them. As it can be seen in these graphs, k-means algorithm and the proposed algorithm of KAFSA have faster convergence than standard AFSA and PSO, but k-means algorithm causes preceding convergence so fast and is trapped in local optimums. Since in the proposed algorithm of KAFSA, k-means algorithm is performed on several artificial fishes in every iteration that each of them are consisting of different cluster centers, thus, its convergence rate is higher than that of k-means. In standard AFSA, convergence rate is much low. In this algorithm also the ability of local search is not appropriate. In general, experimental results show that the proposed algorithm of KAFSA has high accuracy, stability, resistance and convergence rate.

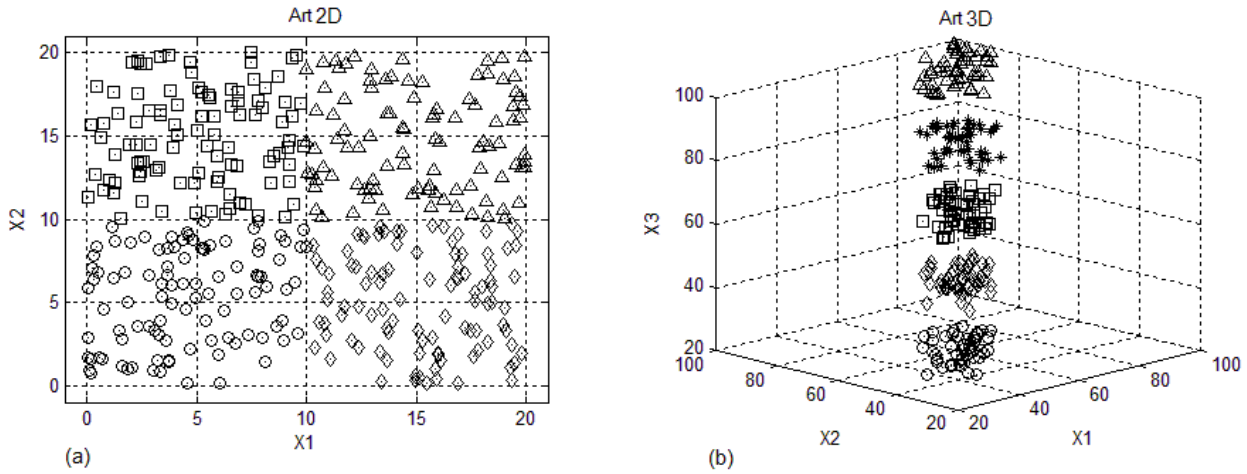

Figure 3. (a) Art 2D data set, (b) Art 3D data set

TABLE I. BEST, MEAN AND STANDARD DEVIATION OF INTRA-CLUSTER DISTANCE AND ERROR RATE FOR 50 PERFORMANCE OF K-MEANS, STANDARD AFSA, PSO ALGORITHMS AND THE PROPOSED METHOD OF KAFSA ON IRIS DATA SET.

Algorithm	Criteria	Intra Cluster Distance	Error Rate
K-means	***Best***	97.3259	10.6666
	Average	102.5686	16.0533
	Std.Dev	11.3378	10.9981
Std-AFSA	***Best***	96.9158	10.3333
	Average	112.3214	27.8888
	Std.Dev	5.4639	9.0092
PSO	***Best***	97.1044	10
	Average	102.2617	10.6444
	Std.Dev	5.8123	4.5028
KAFSA	***Best***	96.8707	10
	Average	97.0067	10.1777
	Std.Dev	0.0477	0.2998

TABLE II. BEST, MEAN AND STANDARD DEVIATION OF INTRA-CLUSTER DISTANCE AND ERROR RATE FOR 50 PERFORMANCE OF K-MEANS, STANDARD AFSA, PSO ALGORITHMS AND THE PROPOSED METHOD OF KAFSA ON PIMA DATA SET.

Algorithm	Criteria	Intra Cluster Distance	Error Rate
K-means	***Best***	52072.2439	43.9843
	Average	55076.5213	44.1059
	Std.Dev	7790.3743	0.3151
Std-AFSA	***Best***	47899.5163	38.7604
	Average	47974.2164	39.9322
	Std.Dev	47.8758	0.9272
PSO	***Best***	47627.7280	38.1510
	Average	48153.1790	39.3967
	Std.Dev	523.5348	0.8282
KAFSA	***Best***	47598.2707	37.8020
	Average	47632.1574	38.1927
	Std.Dev	18.0921	0.1981

TABLE III. BEST, MEAN AND STANDARD DEVIATION OF INTRA-CLUSTER DISTANCE AND ERROR RATE FOR 50 PERFORMANCE OF K-MEANS, STANDARD AFSA, PSO ALGORITHMS AND THE PROPOSED METHOD OF KAFSA ON WINE DATA SET.

Algorithm	Criteria	Intra Cluster Distance	Error Rate
K-means	***Best***	16555.6794	29.7752
	Average	17662.7283	34.3820
	Std.Dev	18780.6769	6.0837
Std-AFSA	***Best***	16771.0191	28.0898
	Average	16862.9259	29.1385
	Std.Dev	44.5073	0.3595
PSO	***Best***	16307.1622	28.0898
	Average	16320.6672	28.7453
	Std.Dev	9.5276	0.392
KAFSA	***Best***	16295.2924	28.0898
	Average	16298.9773	28.5112
	Std.Dev	2.2246	0.4417

TABLE IV. BEST, MEAN AND STANDARD DEVIATION OF INTRA-CLUSTER DISTANCE AND ERROR RATE FOR 50 PERFORMANCE OF K-MEANS, STANDARD AFSA, PSO ALGORITHMS AND THE PROPOSED METHOD OF KAFSA ON ART 2D DATA SET.

Algorithm	Criteria	Intra Cluster Distance	Error Rate
K-means	**Best**	1535.6569	2
	Average	1540.2182	3.4500
	Std.Dev	26.8721	6.5015
Std-AFSA	**Best**	1541.5976	2
	Average	1548.6530	8.6250
	Std.Dev	3.6503	3.4552
PSO	**Best**	1535.3019	2
	Average	1539.6964	4.5500
	Std.Dev	4.2290	2.3738
KAFSA	**Best**	1535.0403	2
	Average	1536.0283	2.9750
	Std.Dev	0.4752	0.8493

TABLE V. BEST, MEAN AND STANDARD DEVIATION OF INTRA-CLUSTER DISTANCE AND ERROR RATE FOR 50 PERFORMANCE OF K-MEANS, STANDARD AFSA, PSO ALGORITHMS AND THE PROPOSED METHOD OF KAFSA ON ART 3D DATA SET.

Algorithm	Criteria	Intra Cluster Distance	Error Rate
K-means	***Best***	1742.7046	0
	Average	2855.2478	30.0933
	Std.Dev	609.6909	13.0157
Std-AFSA	***Best***	1944.6312	0
	Average	2032.4155	2.5199
	Std.Dev	55.9909	6.0116
PSO	***Best***	1803.4374	0
	Average	2169.6334	9.0533
	Std.Dev	302.0498	11.4515
KAFSA	***Best***	1742.5095	0
	Average	1742.6940	0
	Std.Dev	0.1034	0

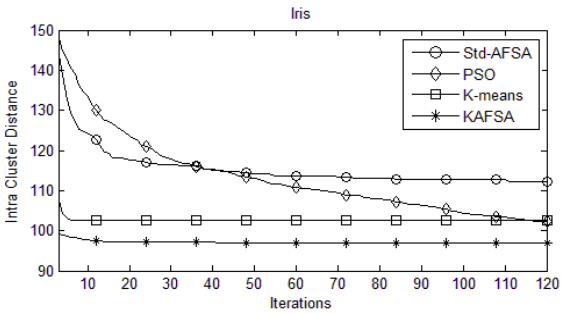

Figure 4. Average graph of intra-cluster distance during performance of k-means, standard AFSA, PSO algorithms and the proposed method for KAFSA on Iris data set for 50 performances.

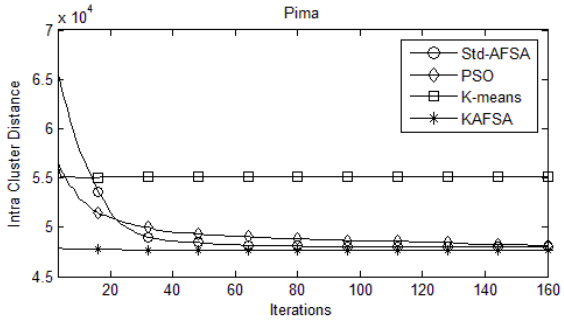

Figure 5. Average graph of intra-cluster distance during performance of k-means, standard AFSA, PSO algorithms and the proposed method for KAFSA on Pima data set for 50 performances.

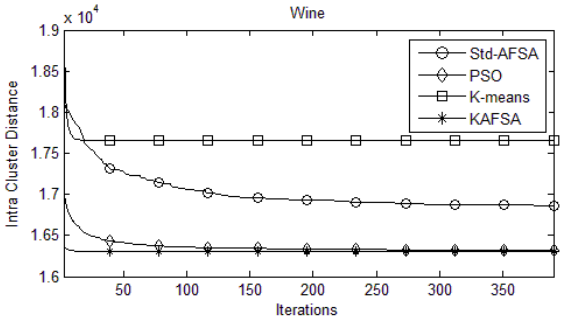

Figure 6. Average graph of intra-cluster distance during performance of k-means, standard AFSA, PSO algorithms and the proposed method for KAFSA on Wine data set for 50 performances.

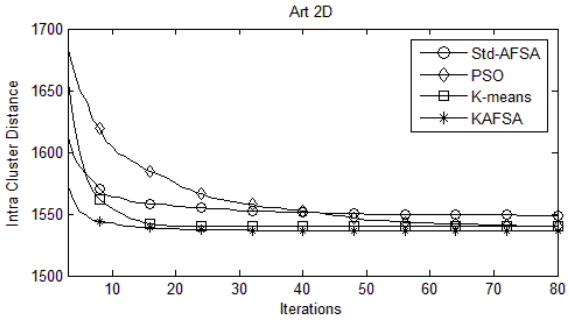

Figure 7. Average graph of intra-cluster distance during performance of k-means, standard AFSA, PSO algorithms and the proposed method for KAFSA on Art 2D data set for 50 performances.

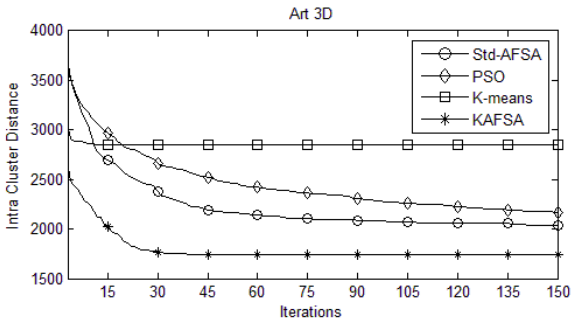

Figure 8. Average graph of intra-cluster distance during performance of k-means, standard AFSA, PSO algorithms and the proposed method for KAFSA on Art 3D data set for 50 performances.

In fact, in KAFSA, since in every iteration of algorithm implementation, one iteration of K-means is performed on some artificial fish, the convergence rate increases highly and because each time it works on a random artificial fish, early convergence and being trapped in local optimums is prevented and sensitivity to initial clusters centers is removed because various artificial fishes have different clusters centers.

performance of standard AFSA in 2 dataset Art2D and Art3D has acceptable beginning because of the similarity of search space bounds in different dimensions. But this procedure can't last to the end because of weakness of local search capability of standard AFSA.

Standard AFSA has two weak point in solving data clustering problem. First weak point is low capability in acceptable searching in problems with different search space bounds in their dimension. Second weak point is low capability in local searching.

Both of these weak points are solved in KAFSA. With changing parameters of visual and step to vectors the first weak point will be solved and with adding K-means as a behavior to artificial fish will improve the capability of local searching so the second weak point will be solved.

VI. CONCLUSION

In this paper, a new hybridized model was proposed for data clustering based on artificial fish swarm algorithm and k-means algorithm. In the proposed model, artificial fishes structure was configured for clustering in AFSA and k-means was used as behavior in AFSA that is performed on some artificial fishes on each iteration. Experimental results for optimizing fitness function related to intra-cluster distance showed that the proposed algorithm crossed well and with high rate from local optimums and converged toward global optimums and obtained results that are relatively stable in different performance. Generally, experimental results showed that the proposed algorithm had better efficiency than Standard AFSA, k-means and PSO.

REFERENCES

[1] C. Pizzuti and D. Talia, "P-AutoClass: scalable parallel clustering for mining large data sets", in IEEE transaction on Knowledge and data engineering, Vol. 15, pp. 629-641, May 2003.

[2] M. Kumar, S. Verma and P. P. Singh, "Data Clustering in Sensor Networks Using ART", in Wireless Communication and Sensor Networks (WCSN08), pp. 51-56, Allahabad, India, Februrary 2009.

[3] A. K. C. Wong and G. C. L. Li, "Simultaneous Pattern and Data Clustering for Pattern Cluster Analysis", in IEEE Transaction on Knowledge and Data Engineering, Vol. 20, pp. 911-923, Los Angeles, USA, June 2008.

[4] J. Marr, "Comparison Of Several Clustering Algorithms for Data Rate Compression of LPC Parameters", in IEEE International Conference on Acoustics Speech, and Signal Processing, Vol. 6, pp. 964-966, January 2003.

[5] X. L. Yang, Q. Song and W. B. Zhang, "Kernel-based Deterministic Annealing Algorithm For Data Clustering", in IEEE Proceedings on Vision, Image and Signal Processing, Vol. 153, pp. 557-568, March 2007.

[6] L. Wu, L. Peng and Y. Ye, "An Evolutionary Immune Network Based on Kernel Method for Data Clustering", in International Conference on Machine Learning and Cybernetics, Vol. 3, pp. 1759-1764, Hong Kong, October 2007.

[7] W. Z. Altun, G. Harrison, R. Tai and P. C. Y. Pan, "Improved K-means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property", in IEEE Transaction on NanoBioscience, Vol. 4, pp. 255- 265, September 2005.

[8] L. Zhu, F. Chung and S. Wang, "Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions", in IEEE Transactions on System, Man, and Cybernetics, Vol. 39, pp. 578-591, June 2009.

[9] S. Datta, C. R. Giannella and H. Kargupta, "Approximate Distributed K-Means Clustering Over a Peer-to-Peer Network", in IEEE Transaction on Knowledge and Data Engineering, Vol. 21, pp. 1372-1388, October 2009.

[10] A. T. Z. Nehorai and B. Porat, "K-means Clustering-based Data Detection and Symbol-timing Recovery for Burst-Mode Optical Receiver", in IEEE Transaction on Commiunications, Vol. 54, pp. 1492-1501, August 2006.

[11] S. S. Shahapurkar and M. K. Sundareshan, "Comparison of Self-Organizing Map with K-means Hierarchical Clustering for Bioinformatics Applications", in IEEE International Joint Conference on Neural Networks, Vol. 2, pp. 1221-1226, January 2005.

[12] J. Kennedy and R. Eberhart, "Particle Swarm Optimization", in IEEE International Conference on Neural Networks, Vol. 4, pp. 1942-1948, Perth, November 1995.

[13] L. X. Li, Z. J. Shao and J. X. Qian, "An Optimizing Method Based on Autonomous Animate: Fish Swarm Algorithm", In Proceeding of System Engineering Theory and Practice, Vol. 11, pp. 32-38 , 2002.

[14] S. Hi, N. Belacel, H. Hamam and Y. Bouslimani, "Fuzzy Clustering with Improved Artificial Fish Swarm Algorithm", In International Joint Conference on Computational Sciences and Optimization 09, Vol. 2, pp. 317-321, Hainan, China, 2009.

[15] L. Xiao, "A Clustering Algorithm Based on Artificial Fish School", in 2nd International Conference on Computer Engineering and Technology, Vol. 7, pp. 766-769, Chengdu, china, aprile 2010.

[16] Y. Luo, J. Zhang and X. Li, "The Optimization of PID Controller Parameters Based on Artificial Fish Swarm Algorithm", In IEEE International Conference on Automation and Logistics, pp. 1058-1062, Jinan, china, 2007.

[17] M.Zhang, C.Shao, M.Li and J.Sun, "Mining Classification Rule with Artificial Fish Swarm", In 6th World Congress on Intelligent Control and Automation, Vol. 2, pp. 5877-5881, Dalian, China, 2006.

[18] G.Cui, X.Cao, J. Zhou and Y.Wang, "The Optimization of DNA Encoding Sequences Based on Improved Artificial Fish Swarm Algorithm", In IEEE International Conference on Automation and Logistics, pp. 1141-1144, Jinan, China, 2007.

[19] D. W. van der Merwe and A. P. Engelbrecht, "Data Clustering Using Particle Swarm Optimization", in the 2003 Congress on Evolutionary Computation, Vol. 1, pp. 215-220, December 2003.

[20] http://archive.ics.uci.edu/ml/

Y. T. Kao, E. Zahara and I. W. Kao, "A Hibridized Approach to Data Clustering", in Elsevier Journal on Expert System with Applications, pp. 1754-1762, 2008.