A WK-Means Approach for Clustering

Fatemeh Boobord, Zalinda Othman, and Azuraliza Abu Bakar
Data Mining and Optimization Research Group, Center for Artificial Intelligence Technology, University Kebangsaan Malaysia, Malaysia

Abstract: *Clustering is an unsupervised learning method that is used to group similar objects. One of the most popular and efficient clustering methods is K-means, as it has linear time complexity and is simple to implement. However, it suffers from gets trapped in local optima. Therefore, many methods have been produced by hybridizing K-means and other methods. In this paper, we propose a hybrid method that hybridizes Invasive Weed Optimization (IWO) and K-means. The IWO algorithm is a recent population based method to iteratively improve the given population of a solution. In this study, the algorithm is used in the initial stage to generate a good quality solution for the second stage. The solutions generated by the IWO algorithm are used as initial solutions for the K-means algorithm. The proposed hybrid method is evaluated over several real world instances and the results are compared with well-known clustering methods in the literature. Results show that the proposed method is promising compared to other methods.*

Keywords: *Data clustering, K-means algorithm, IWO, hybrid evolutionary optimization algorithm, unsupervised learning.*

Received November 28, 2012; accepted August 12, 2013; published online September 4, 2014

1. Introduction

Clustering is a data analysis tool for grouping similar data. It has been used for storing and representing large amounts of information as data. Cluster analysis can be defined as discovering natural hidden groups of objects. It has been used for assigning the same objects to the same groups [22], where different objects are in different groups. Clustering has been applied in many fields, like engineering, computer science, economics, life and medical sciences, astronomy, earth science and social science [3]. Clustering algorithms are traditionally classified into two groups: Hierarchical clustering and partitional clustering. Hierarchical clustering divides objects into a tree of clusters. Since, this is not the subject of this study, we will not mention it in detail. The partitional method typically categorizes objects into K groups, which satisfy the following requirements: Each group has at least one object and each object must be a member of just one group [1, 22].

K-means is the most popular partitional method as it is easy to implement and very efficient with linear time complexity. However, the objective function of the K-means method is non-convex and it may contain many local minima. So, in the minimization process, the objective function of the K-means method may be to trap local optima. Therefore, the outputs of the K-means method greatly depend on the initial choice of cluster centres [12]. To overcome this drawback, many clustering algorithms have been introduced [4]. For example, a novel approach called Genetic K-means Algorithm (GKA) was proposed, which defines a basic mutation operator that is specific to clustering [13]. Nguyen and Cios [16] introduced Genetic Algorithm (GA) K-means logarithmic regression expectation maximization algorithm that mixed the best individuality of each method. A Tabu Search (TS) based algorithm was proposed for fuzzy K-means [15]. This algorithm is able to explore the solution space beyond local optimality in order to find a global solution to the fuzzy clustering problem. Niknam and Amiri [17] proposed a clustering algorithm based on the hybridization of Simulated Annealing (SA) and Ant Colony Optimization (ACO). They combined the advantages of ACO and SA to overcome the shortcomings of the K-means method [19]. Fathian *et al*. [6] proposed a Honey Bee Mating Optimization algorithm (HBMO) for the K-means method. The search algorithm is inspired by the process of HBMO in clustering. Kao *et al*. [12] proposed a method based on K-means, Nelder Mead simplex search and Particle Swarm Optimization. Niknam and Amiri [17] used a combination of fuzzy adaptive Particle Swarm Optimization (PSO), ACO and K-means algorithm. A two-step algorithm was presented by Zalik. The method extends the cost function of the K-means method and assigns at least one object to each cluster at the first step and then tries to minimize the cost function by adjusting seed points in the second step [22]. Pham *et al*. [20] developed a new algorithm called the Bee Algorithm (BA) that is capable of locating near optimal solutions efficiently. Zhang *et al.* [23] presented an Artificial Bee Colony (ABC) clustering algorithm. They used deb's rules to direct the search direction of each candidate solution. Niknam *et al*. [18] again proposed an algorithm based on combining Modify Imperialist Competitive Algorithm (MICA) and the K-means method. Hatamlou [9] developed a new algorithm based on black hole phenomenon. Hatamlou [10] again proposed a binary search algorithm to find optimal centroids of the K-means.

However, most ordinary evolutionary methods, like TS, GA, etc., are slow in converge [23]. In recent years, new methods, such as ACO, PSO, ABC and MICA, were introduced to obtain better solutions and converge more quickly [18]. Invasive Weed Optimization (IWO) is one of these new evolutionary algorithms, which was developed by Mehrabian and Lucas [14]. IWO was applied for clustering and the scores obtained by this method were either less or equal to the other clustering algorithm's scores [5].

As mentioned, the K-means algorithm is sensitive to initial cluster centres and may trap local optima. To overcome this drawback, we propose a hybrid method that combines IWO and K-means algorithm. We integrate the IWO output into the K-means algorithm. To increase the quality of the initial cluster centre for the K-means algorithm, the output of the IWO algorithm is used as an initial state of the K-means method. The performance of the algorithm was tested on several real world instances and the result was compared with other well-known clustering methods i.e., imperialist competitive algorithm, PSO, SA, TS, GA, ant colony and K-means. This paper is organized as follows: Section 2 provides steps for a proposed method with a quick review of the clustering problem, IWO and WK-means algorithm. In section 3, we present our results for optimization on several real data sets. Comparison and discussion with other evolutionary algorithms are also summarized in this section. Finally, section 4 concludes the paper.

2. Proposed Method

2.1. Cluster Analysis Problem

K-means is a simple, fast and very popular clustering method. The procedure for this algorithm first starts with placing *K* objects randomly, as initial cluster centres (*K* is a fixed number as a parameter). Next, the objects are assigned to their closest cluster centre. Then, the algorithm calculates the average of each cluster as a new cluster centre. The last two stages continue until a termination condition is reached. These steps are shown in Figure 1, The goal of the K-means algorithm is to minimize the sum of the distance between cluster centres and objects over all *K* clusters as follows[11]:

$$pref(X, C) = \sum_{i=1}^{N} \min\{\|X_i - C_i\|^2 \; i = 1, ..., K\} \tag{1}$$

Where *pref*(*X*, *C*) is a performance function (fitness function) of the K-means method that is defined on both data items and centre locations. X_i, i=1, ..., N is a data object and C_l, l=1, ..., K *is* a cluster centre [11]. However, during the minimization process of the K-means method, the fitness function of the K-means method is non-convex, which may lead to local optima. In other words, the output of the K-means method is strongly dependant on its initial cluster centres [18]. As such it is important to generate a reasonable initial solution to achieve a good quality cluster centre. We present the K-means algorithm as follows:

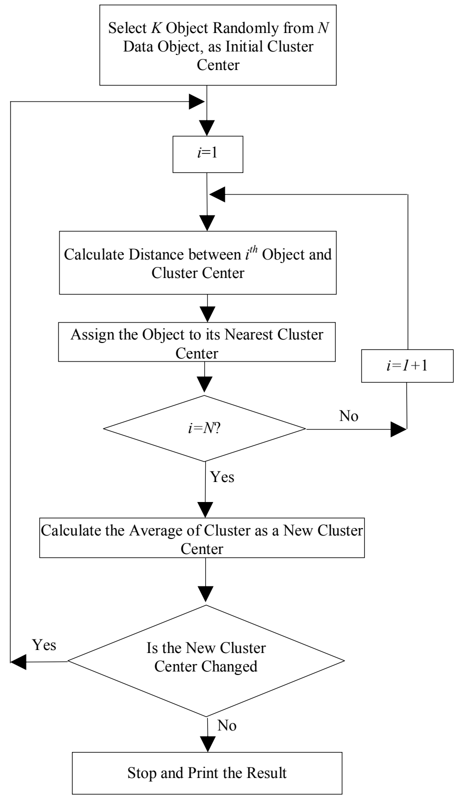

Figure 1. Flowchart of the K-means.

1. Put *K* points as *K* cluster centres.
2. Calculate the distance of each object to cluster centres and assign clusters according to minimum distance.
3. Recalculate the cluster centre according to the mean.
4. Repeat steps 2 and 3 until the maximum number of iterations is reached.

2.2. IWO

IWO is a recent numerical stochastic optimization algorithm. It was developed by Mehrabian and Lucas [14]. The algorithm has a simple process with good exploration and diversity [8]. IWO simulates natural behaviour of weeds in colonizing and finding a suitable place for growth and reproduction [14]. The optimization process is initialized by randomly generating solutions in the space. Then, each individual produces seed according to its fitness. The number of seeds grow linearly from S_{min} (for the worst individual) to S_{max} (for the best). In the next step, the produced seeds are scattered over the search area following the normal distribution, with mean equal to zero and adaptive standard deviation, according to the equation:

$$\delta_{iter} = \frac{(iter_{max} - iter)^n}{(iter_{max})^n}(\delta_{initial} - \delta_{final}) + \delta_{final} \tag{2}$$

Where $iter_{max}$ is the maximum number of iterations, δ_{iter} is the standard deviation at the current iteration and n is the non-linear modulation index. These newly produced seeds, with their parent, compose a potential solution for the next iteration. Producing seeds by this method continues until the maximum population is achieved. An elimination mechanism is employed, where the seeds and their parents are ranked together, and those with better values can survive and reproduce. Figure 2 shows the flowchart of the IWO Algorithm.

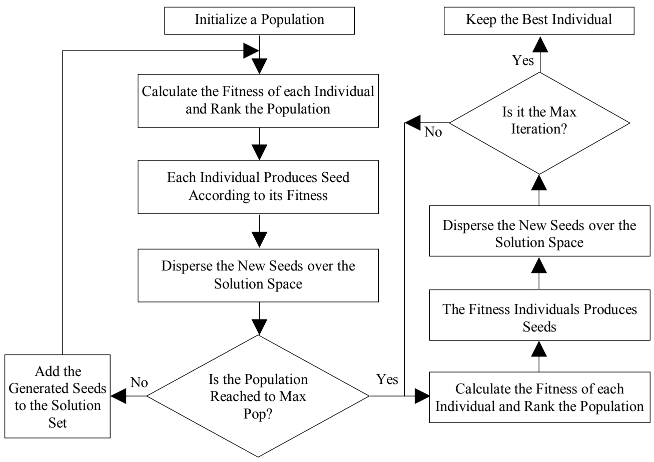

Figure 2. Flowchart of IWO.

2.3. WK-Means

The proposed method produces hybrids through the IWO algorithm and K-means. First, IWO is used to produce a good quality solution (the solution that is near to optimal). Then, the output of the IWO is used as an initial cluster centre for the K-means method. The method exploits the search capability of the IWO algorithm to overcome the local optimum problem of the K-means algorithm. More specifically, the task searches for a good approximate initial solution for the K-means algorithm. We present the WK-means algorithm in two stages as follows:

Algorithm 1: WK-Means

- *Stage 1. IWO Algorithm:*
  1. *Initialize the solution population.*
  2. *Evaluate the fitness of the population.*
  3. *Every member of the population produces seeds according to its fitness.*
  4. *The seeds are spread over the search area randomly by normal distribution and adaptive standard deviation.*
  5. *This process continues until the maximum number of plants is reached. Then, only the fittest plants can survive and reproduce seed, until the maximum iteration is reached.*
- *Stage 2. K-means Algorithm:*
  6. *The best solution obtained from the last stage is selected as an initial solution of the K-means.*
  7. *Assigning objects to clusters.*
  8. *Calculate the new position of K cluster centres.*

Repeat steps 7 and 8 until the maximum iteration is reached.

3. Experimental Results

In this section, the performance of the WK-means algorithm is compared with seven simple (not hybrid) clustering algorithms, including ICA, PSO, SA, TS, GA, ACO and K-means. The algorithm is implemented using Matlab 7.7 on a 2.27 GHz, 2.00 GB RAM laptop.

The Tables 1, 3, 5 and 7 show the results of the comparison among WK-means and ICA [2] ACO [21], PSO [12], SA [19], GA [13], TS and K-means [17, 19], for 100 runs on four real-instances datasets. The best centroid found by the proposed algorithm is also shown in Tables 2, 4, 6 and 8. For comparison of the results, in the Iris dataset, the best, average and worst results found by the algorithm are 96.6555, 96.6565 and 96.6704, respectively. Meanwhile, the results of the nearest algorithm, which is the ICA, are 96.6997, 96.8466 and 97.0059, for the same dataset, respectively. The most notable thing is that none of the other algorithms reach the worst solution found by the WK-means algorithm, even in their best solutions. At the same time, the standard deviation of solutions found by the WK-means algorithm is the smallest of all algorithms. This means that the WK-means algorithm is more reliable than the other methods and converges to the global optimal solutions in all of the runs. For the CMC dataset, the best solution found by the WK-means algorithm is 5694.6 and the nearest result for this value is 5700.9853, which belongs to the PSO. The best, average and the worst results found by the WK-means algorithm on the Wine dataset are 16,294, 16,297 and 16,304, respectively. Meanwhile, the results of the algorithm for ICA are 16,295.24, 16,297 and 16,304. For the vowel dataset, the best and average solutions found by the WK-means algorithm are 148,967.5 and 149,502.238, respectively. Meanwhile, the nearest results for these values are 150,991.6147 and 151,547.0511, which belong to the ICA dataset. As seen from results, the proposed method compares well with the other well-known algorithms. The WK-means algorithm outperforms other algorithms on several datasets.

Table 1. Result obtained by the algorithms for 100 different runs on Iris data set.

Method	Best Function Value	Average Function Value	Worst Function Value	Standard Deviation
WK-means	96.6555	96.6565	96.6704	0.00251
ICA	96.6997	96.8466	97.0059	0.11149
PSO	96.8942	97.2328	97.8973	0.34716
SA	97.4573	99.975	102.01	2.018
TS	97.3659	97.8680	98.5694	0.53
GA	113.9865	125.1970	139.7782	14.563
ACO	97.1007	97.1715	97.8084	0.367
K-means	97.333	106.05	120.45	14.6311

Table 2. Centers obtained by the algorithms for the best result on Iris data set.

Center 1	Center 2	Center 3
6.8231	5.0060	5.9033
3.0667	3.4180	2.7475
5.7256	1.4640	4.3820
2.0795	0.2440	1.4180

Table 3. Result obtained by the algorithms for 100 different runs on CMC data set.

Method	Best Function Value	Average Function Value	Worst Function Value	Standard Deviation
WK-means	5694.6	5751.04	5988.3	57.9428
ICA	5725.7	5736.36	5752.94	8.00056
PSO	5700.9	5820.96	5923.24	46.9596
SA	5849.0	5893.48	5966.94	50.8672
TS	5885.0	5993.59	5999.80	40.8456
GA	5705.6	5756.59	5812.64	50.3694
ACO	5701.9	5819.13	5912.43	45.6347
K-means	5842.2	5893.43	5934.43	47.16

Table 4. Centers obtained by the algorithms for the best result on CMC data set.

Center 1	Center 2	Center 3
43.8021	33.7219	24.4088
2.8369	3.0316	2.9730
3.3262	3.4576	3.4713
4.8235	3.7811	1.8294
0.8102	0.7968	0.9223
0.7701	0.6884	0.7889
1.8904	2.1321	2.2990
3.3369	3.2268	2.9257
0.1150	0.0750	0.0473
1.6123	2.0631	1.9916

Table 5. Result obtained by the algorithms for 100 different runs on Wine data set.

Method	Best Function Value	Average Function Value	Worst Function Value	Standard Deviation
WK-means	16,294	16,297	16,304	2.0219
ICA	16,295	16,298	16,304	2.9345
PSO	16,345	16,417	16,562	85.4974
SA	16,473	17,521	18,083	753.084
TS	16,666	16,785	16,837	52.073
GA	16,530	16,530	16,530	0
ACO	16,530	16,5305	16,530	0
K-means	16,555	18,061	18,563	793.213

Table 6. Centers obtained by the algorithms for the best result on Wine data set.

Center 1	Center 2	Center 3
13.8	12.5	12.9
1.9	2.4	2.6
2.4	2.3	2.4
17.1	20.8	20.0
106.6	92.5	102.1
2.9	2.1	2.1
3.0	1.8	1.5
0.3	0.4	0.4
1.9	1.5	1.4
5.6	4.1	5.6
1.1	0.9	0.9
3.1	2.5	2.3

Table 7. Result obtained by the algorithms for 100 different runs on Vowel data set.

Method	Best Function Value	Average Function Value	Worst Function Value	Standard Deviation
WK-means	148,967.5	149,502.238	153,053.1	1,139.966
ICA	150,991.6	151,547.051	152,735.16	704.0907
PSO	148,976.0	151,999.825	158,121.18	28,813.4692
SA	149,370.4	161,566.281	165,986.42	2847.08594
TS	149,468.2	162,108.538	165,996.42	2846.23516
GA	149,513.7	159,153.498	165,991.65	3105.5445
ACO	149,395.6	159.458.143	165,939.82	3485.3816
K-means	149,422.2	159,242.89	161,236.81	916

Table 8. Centers obtained by the algorithms for the best result on Vowel data set.

Center1	Center 2	Center 3	Center 4	Center 5	Center 6
388.8	368.9	618.6	445.1	516.1	404.1
2142.6	2298.1	1320.3	993.8	1833.6	1027.2
2674.1	2986.4	2345.2	2664.8	2556.4	2320.7

4. Conclusions

This paper presented a new clustering method based on a hybrid IWO and K-means algorithm. IWO a good global search algorithm, while K-means is a simple and fast local search clustering method. The proposed algorithm uses the advantages of IWO and K-means to prevent algorithms from getting to local optima. Experimental results using different datasets are shown in the tables and the results compare well with several other clustering algorithms, such as K-means, ACO, GA, TS, SA, PSO and ICA.

Acknowledgment

This research was supported by University Kebangsaan Malaysia, under grant number FRGS/1/2013/ICT02/UKM/01/1.

References

[1] Abu Abbas O., "Comparison Between Data Clustering Algorithms," *the International Arab Journal of Information Technology*, vol. 5, no. 3, pp. 320-325, 2008.

[2] Atashpaz H. and Lucas C., "Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired by Imperialistic Competition," *in Proceedings of IEEE Congresson Evolutionary Computation*, Singapore, pp. 4661-4667, 2007.

[3] Berkhin P., *Survey of Clustering Data Mining Techniques*, Grouping Multidimensional Data, 2006.

[4] Boobord F., Othman Z., and Abu Bakar A., "Metaheuristic Method for Clustering Problem," *in 2th National Doctoral Seminar on Artificial Intelligence Technology*, Malaysia, pp. 122 -125, 2012.

[5] Chowdhury A., Bose S., and Dos S., "Automatic Clustering based on Invasive Weed Optimization Algorithm," *in Proceedings of the 2nd International Conference*, *Swarm*, *Evolutionary and Memetic Computing*, Andhra Pradesh, India, pp. 105-112, 2011.

[6] Fathian M., Amiri B., and Maroosi A., "Application of Honey-Bee Mating Optimization Algorithm on Clustering," *Applied Mathematics and Computation*, vol. 190, no. 2, pp. 1502-1513, 2007.

[7] Han J. and Kamber M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, California, USA, 2006.

[8] Hajimirsadeghi H. and Lucas C., "A Hybrid IWO/PSO Algorithm for Fast and Global Optimization," *in Proceedings of IEEE Conference EUROCON*, St.-Petersburg, Russia, pp. 1964-1971, 2009.

[9] Hatamlou A., "Black Hole: A New Heuristic Optimization Approach for Data Clustering," *Information Sciences*, vol. 222, pp. 175-184, 2013.

[10] Hatamlou A., "In Search of Optimal Centroids on Data Clustering using a Binary Search Algorithm," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1756-1760, 2012.

[11] Jain A., "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters,* vol. 31, no. 8, pp. 651-666, 2010.

[12] Kao Y., Zahara E., and Kao I., "A Hybridized Approach to Data Clustering," *Expert Systems with Applications* vol. 34, no. 3, pp. 1754-1762, 2008.

[13] Krishna K. and Murty M., "Genetic K-Means Algorithm," *IEEE Transactions on System*, *Man*, *and Cybernetics-Part B*, vol. 29, no. 3, pp. 433-439, 1999.

[14] Mehrabian A. and Lucasc C., "A Novel Numerical Optimization Algorithm Inspired from Weed Colonization," *Ecological Informatics*, vol. 1, no. 4, pp. 355-366, 2006.

[15] Ng M. and Wong J., "Clustering Categorical Data Sets using Tabu Search Techniques," *Pattern Recognition*, vol. 35, no. 12, pp. 2783-2790, 2002.

[16] Nguyen C. and Cios K., "GAKREM: A Novel Hybrid Clustering Algorithm," *Information Sciences*, vol. 178, no. 22, pp. 4205-4227, 2008.

[17] Niknam T. and Amiri B., "An Efficient Hybrid Approach based on PSO, ACO and K-Means for Cluster Analysis," *Applied Soft Computing*, vol. 10, no. 1, pp. 183-197, 2010.

[18] Niknam T., Firouzi B., and Nayeripour M., "An Efficient Hybrid Algorithm based on Modified Imperialist Competitive Algorithm and K-Means for Data Clustering," *Engineering Application of Artificial Intelligence*, vol. 24, no. 2, pp. 306-317, 2011.

[19] Niknam T., Olamaie J., and Amiri B., "A Hybrid Evolutionary Algorithm Based on ACO and SA for Cluster Analysis," *the Journal of Applied Sciences*, vol. 8, no. 15, pp. 2695-2702, 2008.

[20] Pham D., Otri S., Afifi A., and Al-Jabbouli H., "Data Clustering using the Bees Algorithm," *in Proceedings of the 40th CIRP International Manufacturing Systems*, Liverpool, UK, pp. 1-8, 2007.

[21] Shelokar P., Jayaraman V., and Kulkarni B., "An Ant Colony Approach for Clustering," *Analytica Chimica Acta,* vol. 509, no. 2, pp. 187-195, 2004.

[22] Zalik K., "An Efficient K'-Means Clustering Algorithm," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1385-1391, 2008.

[23] Zhang C., Ouyang D., and Ning J., "An Artificial Bee Colony Approach for Clustering," *Expert Systems with Applications*, vol. 37, no. 7, pp. 4761-4767, 2010.

Fatemeh Boobord received the BS degree in applied mathematics from Islamic Azad University of Rasht Branch, Iran in 2005 and the MS degree in applied mathematics from Islamic Azad University of Lahijan Branch in 2010. She is PhD candidate in computer science at University Kebangsaan Malaysia (UKM) from 2010. Her research interests are artificial intelligence, data mining and optimization, operation research, data envelopment analysis (DEA).

Zalinda Othman received the BS degree in quality control and instrumentation from University Science Malaysia, Penang in 1994, and the MS degree in quality engineering, from University of Newcastle upon Tyne, United Kingdom, in 1996 and the PhD degree in artificial intelligence from University Science Malaysia, Penang, in 2002. She is Head of Industry and Community Partnership in Faculty of Information Science and Technology at University Kebangsaan Malaysia, where she is currently an associate professor. Her main research topics are the study of production optimization, artificial intelligence in manufacturing and data mining in production planning and control.

Azuraliza Abu Bakar is a Professor in data mining at University Kebangsaan Malaysia. She received her PhD degree (artificial intelligence) from University Putra Malaysia in 2002. Her research interests are in time series data mining, outbreak detection and deviation detection model employing nature inspired computing techniques.