5th SASTech
Iran | Mashhad
May 12 - 17 | 2011
5th Symposium on Advances in Science & Technology

An optimized clustering algorithm based on K-means using Honey Bee Mating algorithm

Ebrahim Teimoury
teimoury@iust.ac.ir

Mohammad Reza Gholamian
gholamian@iust.ac.ir

Bizhan Masoum
b.masoum@gmail.com

Mojgan Ghanavati
mojgan.ghanavati@gmail.com

Paper Reference Number: 17

Abstract

Clustering is a technique to categorize objects in k groups so that objects with most similar attribute values are placed in one group. Partitioning algorithms are a group of clustering algorithms and k-means algorithm is one of the most popular algorithms in this group that is very simple and fast but has some drawbacks too. In this paper we tried to propose an optimized hybrid clustering algorithm based on Honey Bee Mating algorithm and K-means in order to resolve these drawbacks. Finally, the performance of this optimized algorithm has been evaluated and compared with some other meta-heuristic clustering algorithms.

Key words: Clustering algorithm, Honey Bee Mating algorithm, K-means, Meta-heuristics

1. Introduction

Considering the large volume of information in today's world, the need for new techniques of data analysis and acquiring useful information and knowledge has become more obvious and thus newer optimized algorithms are continuously offered. Clustering is one of the widely used data analysis techniques and many clustering algorithms have been presented according to their different applications. Generally clustering algorithms are divided into two categories: partitioning-based and density-based. Partitioning-based algorithms try to divide the whole data into smaller groups based on a series of common criteria between all data to provide useful information and patterns about the data. K-means algorithm is one of the most popular partition-based algorithms which is very simple and fast but has 3 important drawbacks too. First, it is possible for this algorithm to be converged to local optimum solutions. Second, the results obtained from this algorithm are strongly dependent to its initial points. Third, the numbers of clusters should be predetermined. Many researchers tried to resolve these problems through combined algorithms and some of them used meta-heuristic algorithms. Zhang et al. (2009) proposed an ant colony optimization (ACO) algorithm to organize sensor nodes in a wireless sensor network into clusters based on their energy consumption. They evaluated this algorithm on several datasets and claimed that it could always be converged to the global optimum

solutions. Santosa et al. (2009) used a novel clustering algorithm which is based on major behavioral traits of cats such as seeking and tracing modes. They modified the formulas used in previous versions of this algorithm and used experimental data to evaluate it. Results show that this modified algorithm can perform better than PSO and K-means algorithms. Niknam et al. (2010) suggested a hybrid algorithm based on PSO and SA to resolve the drawbacks in K-means algorithm. Experimental results of this hybrid algorithm show that this combined algorithm can perform better and faster than SA, PSO and even K-means algorithms.

Zhang et al. (2010) proposed a simulated clustering algorithm based on Artificial Bee Colony (ABC) in order to group N items in k cluster. This algorithm was tested on some famous datasets and compared with some meta-heuristic algorithms such as Genetic Algorithm, Simulated Annealing, Taboo Search, Ant Colony and K-NM-PSO.

Another algorithm which has been simulated from honey bee behaviors is Honey Bee Mating Optimization algorithm (HBMO) first proposed by Abbass (2001a, b) and then completed by Bozorg Haddad et al. (2006) for water resources optimization problem. Bozorg Haddad and Marino (2007) again used this algorithm to another version of water resources optimization problem with dynamic penalty function. Bozorg Haddad et al. (2008a, b, c) used this algorithm to solve different versions of water reservoirs problem. Bozorg Haddad et al (2008d) used this algorithm to find the shortest path in project management problems. Also Niknam et al. (2008) proposed a combined algorithm based on HMBO and fuzzy multi-objective in order to find the configuration of power feeders.

Finally, Fathian and Amiri (2007) used HBMO algorithm to resolve one of the K-means drawbacks and called it HBMK. Then they compared its performance versus some other meta-heuristic algorithms like GA, SA, TS and ACO. Also Amiri and Fathian (2007) combined this novel clustering algorithm with SOM neural networks for market segmentation.

In this paper, we proposed an optimized hybrid clustering algorithm in order to resolve all the K-means drawbacks using a modified version of Honey Bee Mating algorithm and K-means combination. This paper is organized as follows: section 2 provides some information about the data used to evaluate the proposed algorithm. Section 3 describes the proposed hybrid algorithm. Section 4 shows the results achieved through running this algorithm on some datasets and compares it with some other meta-heuristic clustering algorithms. Finally, section 5 includes a brief conclusion.

2. Data and Material

Three famous datasets are used here to evaluate the proposed algorithm including Iris, Wine and Breast Cancer datasets. All data in Iris and Wine datasets are located in 3 different clusters and Breast Cancer dataset consists of 2 clusters.

3. Research Methodology

In this section we try to describe our improved HBMK algorithm. The basis of this algorithm is somehow similar to the one proposed by Fathian and Amiri (2007) but has been improved in two aspects in order to resolve all the K-means drawbacks. First, we empowered the mutation stage in the algorithm in order to improve the trial solutions (broods) created in each generation. Second, as mentioned by Fathian and Amiri (2007) the main drawback in their algorithm was that the number of clusters (k) must be predetermined. We used a measure called Silhouette Coefficient to resolve this problem.

3.1. Honey Bee Mating Optimization (HBMO) algorithm

A honey-bee colony typically consists of a single egg laying long-lived queen, anywhere from zero to several thousand drones (depending on the season) and usually 10,000 to 60,000 workers. The queen is the most important member of the hive because she is the one that keeps the hive going by producing new queen and worker bees. The mating process occurs during mating-flights far from the nest. A mating flight starts with a dance where the drones follow the queen and mate with her in the air. After every mating with about 7 to 20 drones, which all die after the mating process, the sperm from each drone is planted inside a pouch in queen's body (spermatheca) and she uses the stored sperms to fertilize the eggs. Each time a queen lays fertilized eggs, she retrieves at random a mixture of the sperms accumulated in the spermatheca to fertilize the egg.
In an optimization problem, this mating–flight may be considered as a set of transitions in a state-space (the environment) where the queen moves between the different states in some speed and mates with the drone encountered at each state probabilistically. At the start of the flight, the queen is initialized with some energy content and returns to her nest when her energy is within some threshold from zero or when her spermatheca is full.
Here, the functionality of workers is restricted to brood care and so each worker can be considered as a heuristic which tries to improve a set of broods. A drone mates with a queen probabilistically using an annealing function as Eq. 1 (Abbass, 2001a):

$$\Pr ob(Q,D) = \exp\left[-\Delta(f) \Big/ S(t) \right] \tag{1}$$

Where Prob (*Q*, *D*) is the probability of adding the sperm of drone *D* to the spermatheca of queen *Q* or the probability of a successful mating; $\Delta(f)$ is the absolute difference between the fitness of *D* and the fitness of *Q*; and *S*(*t*) is the speed of the queen at time t. It is apparent that the probability of mating is high when either is still in the start of her flight or when the fitness of the drone is as good as the queen's. After each transition in space, the speed and energy of queen will decay using Eqs. 2~3.

$$E(t+1) = \alpha \times E(t) \tag{2}$$
$$S(t+1) = \alpha \times S(t) \tag{3}$$

Where α is a factor $\in [0, 1]$ which is the amount of energy and speed reduction after each transition.
As shown in Fig. 1, the HBMO algorithm is a novel combined algorithm consists of GA, SA, local search and some innovations for its self-adaption.

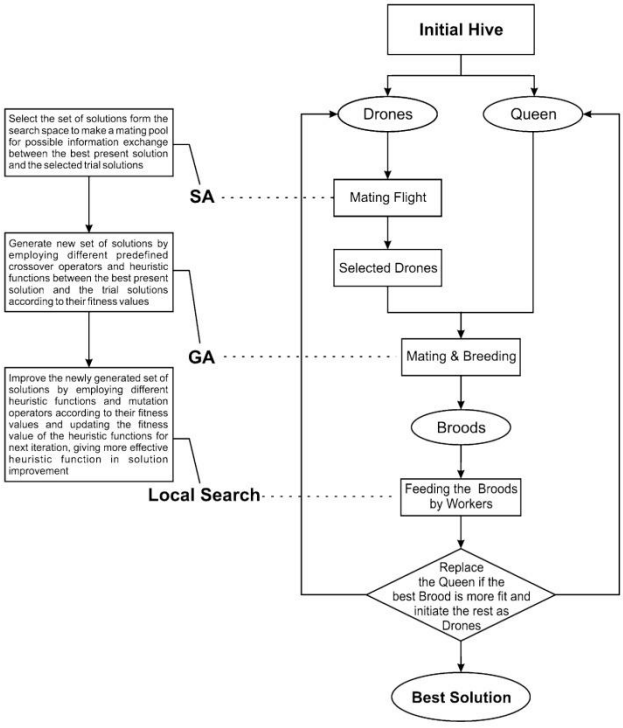

Fig.1 HBMO algorithm flowchart (Bozorg Haddad, 2008d)

The HBMO process consists of two repetitive features: reproduction and improvement. In the first process, the queen (the best up to now solution) and the drones (some selected trial solutions) mate (produce new solutions) and generate the broads (new trial solutions) by different genome combinations (crossover operators). Drones can be either the best broods of the previous generation or some selected solutions by the mating flight. The improvement process applies workers (different mutation operators) on queen and broods, generating some potential broods (test solutions) to be used as the queen or the drones for the next generation. The fitness of the resulted genotype is determined by evaluating the value of its fitness function.

The algorithm starts with three user-defined parameters (the number of queens, the queen's spermatheca size and the number of broods that will be born by all queens) and one predefined parameter (the number of workers or heuristics). The number of queens is usually equal to the number of problem's variable. The heuristics used here are: Random Flip, Random New, 1-Point Crossover, 2-Point Crossover, GSAT and Random Walk. The Random Flip randomly chooses a variable and changes its value to its complement. The Random New replaces the brood's genotype with a new randomly generated genotype. The 1-point and 2-point crossover heuristics, crossover the brood's genotype with a randomly generated genotype at random crossover points. GSAT is a kind of greedy local search and Random Walk that simulates GSAT operation while considering probability.

Fathian and Amiri (2007) utilized HBMO algorithm to find a set of optimal initial cluster centers to be used in K-means algorithm. They considered the number of queens equal to the number of cluster centers. After finding these initial centers, using the K-means algorithm, data can be easily clustered. The number of cluster centers for each dataset must be predetermined for their algorithm but here we proposed a method to resolve this problem too.

3.2. The Silhouette Coefficient

Silhouette Coefficient[1] is one of the efficient tools to evaluate the clusters in clustering process. SC uses intra-compactness and inter-separation to evaluate clusters. In general calculating the SC is a 3 steps procedure as follow:

a. Calculate a_i as the average distance between ith data point and all other data points of its cluster.

b. Calculate b_i as the average distance between ith data point and all data points of closest cluster.

c. Calculate SC of ith data point based on Eq. 4.

$$S_i = (b_i - a_i) \Big/ \max(a_i, b_i) \tag{4}$$

SC can have a value between -1 and 1. Larger SC values show better clustering performance. Calculating the SC value for large datasets is very complex so we used an

[1] SC

efficient approach proposed by Al-Zoubi and Al-Rawi (2008) in order to decrease this calculation complexity about 50%.
To calculate the SC of nth cluster, we calculate the average of SC for all of data points of nth cluster and to compute the overall SC of a clustering process, the average of SC for all clusters is calculated.
To find the optimal number of clusters, the SC value for different clustering solutions with various numbers of clusters is calculated. The largest calculated SC value shows the optimal number of clusters. We use this number of clusters as an input value in the HBMK algorithm. Fig. 2 shows the procedure of the modified algorithm.

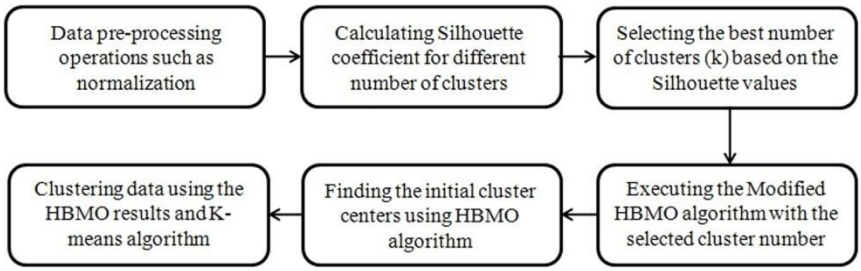

Fig.2 Total procedure of the improved HBMO clustering algorithm

4. Results and Analysis

To evaluate the performance of Silhouette Coefficient, we applied it on 3 standard datasets. We supposed that numbers of clusters are between 2 and 10 and then calculated the SC for each clustering solution. Result shows that SC can successfully find the optimum and correct number of clusters. Table 1 shows the values of SC for different clustering solutions with cluster numbers between 2 and 10 on 3 standard datasets.

	2	3	4	5	6	7	8	9	10
Iris	0.995592	0.996799	0.995294	0.981471	0.981401	0.982042	0.982631	0.983474	0.983681
Wine	0.999888	0.999998	0.974616	0.974627	0.979740	0.976481	0.978435	0.978408	0.980333
B.C.	0.999483	0.993162	0.993399	0.991256	0.991291	0.991254	0.992416	0.992979	0.992201

Table 1. Silhouette coefficient value of different clustering solutions

In this section also the performance of the proposed algorithm on different datasets has been compared with some other clustering algorithms. To do this, five different indexes have been used as indictors to evaluate the performance of algorithms:
1- Sum of square error for each data point based on Eq. 5.

$$SSE(x)=\sum_{i=1}^{N}\min\left\{\| Y_i - X_j \|^2\right\},\ j=1,2,...,k \tag{5}$$

Where $\|Y_i - X_j\|^2$ is the Euclidian distance between data point, Y_i, and cluster center, X_j. Our goal is to minimize the SSE using the proposed algorithm.
2- F-measure index that uses precision and recall idea of information retrieval based on Eq. 6.

$$F(i,j)=\frac{(b^2+1)(p(i,j)\times r(i,j))}{b^2(p(i,j)+r(i,j))} \tag{6}$$

Where

$$r(i,j) = n_{ij} / n_j \quad (7)$$

$$p(i,j) = n_{ij} / n_i \quad (8)$$

Where n_i is the item numbers of a default i class and n_j is the item numbers of j cluster and n_{ij} is item numbers of i class that exist in j cluster.
In order to have an equal weight for p and r indexes, we suppose that b = 1. Total value of F-measure for a dataset with n data points is calculated based on Eq. 9. Larger value of F-measure shows the better clustering solution.

$$F = \sum_i \frac{n_i}{n} MAX_j \{F(i,j)\} \quad (9)$$

In order to obtain more reliable results, we executed each clustering algorithm 100 times on all three datasets and then the average results considered as the final result for each algorithm.
3- Average run time.
4- Number of times that the fitness function has been executed in each algorithm in order to find the solution.
5- Average objective function value.
Final comparison results have been shown in Tables 2~4.

Algorithm	Avg. objective function value	SSE	Avg. run time (second)	No. of fitness function execution	F-measure
PSO	**97.2327**	**0.3471**	**~31**	**4953**	**0.780**
SA	**99.9571**	**2.018**	**~33**	**5314**	**0.776**
TS	**97.8671**	**0.5241**	**~137**	**20211**	**0.776**
GA	**125.197**	**14.567**	**~142**	**37301**	**0.777**
ACO	**97.1792**	**0.3541**	**~78**	**11001**	**0.779**
K-Means	**106.55**	**14.6331**	**~0.5**	**123**	**0.777**
HMBK	**98.3121**	**0.3130**	**~45**	**6401**	**0.783**

Table 2. Results obtained by different algorithms on Iris dataset

Algorithm	Avg. objective function value	SSE	Avg. run time (second)	No. of fitness function execution	F-measure
PSO	**16416.13**	**86.1012**	**~124**	**16533**	**0.518**
SA	**17521.14**	**754.013**	**~130**	**17266**	**0.514**
TS	**16785.12**	**52.173**	**~141**	**22701**	**0.515**
GA	**16531.11**	**0.0001**	**~174**	**33501**	**0.515**
ACO	**16530.97**	**0.0001**	**~120**	**15407**	**0.519**
K-Means	**18062.33**	**793.313**	**~0.7**	**390**	**0.520**
HMBK	**16333.79**	**0.0001**	**~140**	**21012**	**0.520**

Table 3. Results obtained by different algorithms on Wine dataset

Algorithm	Avg. objective function value	SSE	Avg. run time (second)	No. of fitness function execution	F-measure
PSO	**3050.11**	**110.88**	**~124**	**16290**	**0.818**
SA	**3238.88**	**230.13**	**~126**	**17332**	**0.818**
TS	**3251.13**	**232.45**	**~131**	**18801**	**0.818**
GA	**3248.89**	**229.01**	**~136**	**20119**	**0.819**

ACO	**3040.18**	**90.19**	**~123**	**15988**	**0.820**
K-Means	**3251.19**	**251.78**	**~0.5**	**180**	**0.829**
HMBK	**3268.88**	**47.13**	**~126**	**17918**	**0.829**

Table 4. Results obtained by different algorithms on Breast Cancer dataset

Results show that the modified clustering algorithm has an acceptable performance compared with other clustering algorithms and except for Average run time measure, in all other measures can outperform other known clustering algorithms.

5. Conclusions

Clustering is one of the important and applicable techniques in data mining. K-means is one of the most popular clustering algorithms because of its simplicity and speed. However K-means has some drawbacks too. Its result highly depends on its initial points and it may converge to the local optimums. Different combined algorithms have been presented to resolve these drawbacks. A group of these combined algorithms is based on meta-heuristic algorithms like a recently proposed algorithm called HBMK. However, HBMK still needs some improvement. In this paper, we combined an improved version of HBMK algorithm with Silhouette Coefficient to enhance its efficiency. Performance evaluation results show that this modified algorithm can outperform some other famous clustering algorithms.

References

Abbass, H. A. (2001a). Marriage in honey bees optimization (MBO): A haplometrosis polygynous swarming approach. *In The Congress on Evolutionary Computation, CEC2001*. Seoul, Korea, 207–214.

Abbass, H. A. (2001b). Amonogenous MBO approach to satisfiability. *In The International Conference on Computational Intelligence for Modelling, Control and Automation, CIMCA*. Las Vegas, NV, USA.

Al-Zoubi, M. B. & Al-Rawi, M. (2008). An Efficient Approach for Computing Silhouette Coefficients. *Computer Science*, 4(3), 252-255.

Amiri, B. & Fathian, M. (2007). Integration of self organizing feature maps and honey bee mating optimization algorithm for market segmentation. Journal of Theoretical and Applied Information Technology, 70-86.

Bozorg Haddad, O. & Afshar, A. & Marino, M. A. (2006). Honey-Bees Mating Optimization (HBMO) Algorithm: A New Heuristic Approach for Water Resources Optimization. *Water Resources Management*, 20, 661–680.

Bozorg Haddad, O. & Marino, M. A. (2007). Dynamic penalty function as a strategy in solving water combinatorial optimization problems with honey-bee optimization (HBMO) algorithm. *Journal of Hydroinformatics*. 9(3), 233-250.

Bozorg Haddad, O. & Afshar, A. & Marino, M.A. (2008a). Design-operation of multi-hydropower reservoirs: HBMO approach. *Water Resources Management*, 22 (12), 1709-1722.

Bozorg Haddad, O. & Afshar, A. & Marino, M.A. (2008b). Honey-bee mating optimization (HBMO) algorithm in deriving optimal operation rules for reservoirs. *Journal of Hydroinformatics*, 10(3), 257-264.

Bozorg Haddad, O. & Adams, B. J. & Marino, M.A. (2008c). Optimum rehabilitation strategy of water distribution systems using the HBMO algorithm. *Journal of Water Supply: Research and Technology - AQUA*, 57(5), 327-350.

Bozorg Hadad, O. & Mirmomeni, M. & Zarezadeh, M. & Marino, M. A. (2008). Finding the shortest path with honey-bee mating optimization algorithm in project management problems with constrained/unconstrained resources. *Computer Optimization Applications*, 47(1), 97-128.

Fathian, M. & Amiri, B. (2008). A Honeybee-mating Approach for Cluster Analysis. *Advance Manufacture Tech*. 1(38), 809–821.

Niknam, T. & Amiri, B. & Olamaei, J. & Arfei, A. (2009). An efficient hybrid evolutionary optimization algorithm based on PSO and SA for clustering. *Journal of Zhejiang University SCIENCE A*. 10(4), 512-519.

Niknam, T. & Olamaei, J. & Khorshidi, R. (2008). A Hybrid Algorithm Based on HBMO and Fuzzy Set for Multi-Objective Distribution Feeder Reconfiguration. *World Applied Sciences*. 4(2), 308-315.

Santosa, B. & Ningrum, M. K. (2009). Cat Swarm Optimization for Clustering. *International Conference of Soft Computing and Pattern Recognition*.

Tan, P. N. & Steinbach, M. & Kumar, V. (2006). *Introduction to Data Mining*. Michigan state university, chapter 8.

Zhang, C. & Ouyang, D. & Ning, J. (2010). An artificial bee colony approach for clustering. *Expert Systems with Applications*. 37, 4761–4767.

Zhang, C. & Xu, Q. (2009). Clustering Approach for Wireless Sensor Networks Using Spatial Data Correlation and Ant-Colony Optimization. *International Conference on Networks Security, Wireless Communications and Trusted Computing*.