Use of the Improved Frog-Leaping Algorithm in Data Clustering

Sahifeh Poor Ramezani Kalashami*, Seyyed Javad Seyyed Mahdavi Chabok

Faculty of Engineering, Department of Artificial Intelligence, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Received 5 February 2016; accepted 24 March 2016

Abstract

Clustering is one of the known techniques in the field of data mining where data with similar properties is within the set of categories. K-means algorithm is one the simplest clustering algorithms which have disadvantages sensitive to initial values of the clusters and converging to the local optimum. In recent years, several algorithms are provided based on evolutionary algorithms for clustering, but unfortunately they have shown disappointing behavior. In this study, a shuffled frog leaping algorithm (LSFLA) is proposed for clustering, where the concept of mixing and chaos is used to raise the accuracy of the algorithm. Because the use of concept of entropy in the fitness functions, we are able to raise the efficiency of the algorithm for clustering. To perform the test, the four sets of real data are used which have been compared with the algorithms K-menas, GA, PSO, CPSO. The results show better performance of this method in the clustering.

Keywords: Sales Forecast, ANFIS, Time Series Analysis, PSO & BPN methods.

1. Introduction

Clustering is one of the techniques known in the field of data mining. This process is actually a type of grouping of objects in different clusters that objects in a cluster are similar to each other. In clustering tries to data be divided into clusters that the similarities between the data in each cluster be maximized and similarity between data in different clusters be minimized [1]. When we apply clustering algorithm on a set of objects, clusters obtained can be used to emergence the inherent structure in data used in many fields, clustering has many applications such as pattern recognition, machine learning, data mining, information retrieval and bio-informatics. In recent years, various algorithms are provided for clustering, but unfortunately the classic model-based clustering techniques have shown disappointing behaviors in clustering. k-means algorithm is one of the most common algorithms for clustering, which is presented by Macgueen in 1967. k-means is simple and flexible, and easy to understand, but some disadvantages of this algorithm is being sensitive to initial values of the clusters and converging to a local optimum [2]. In order to overcome the shortcomings of k-means, many heuristic methods have been used in the past two decades. For example, in 2000, features of genetic algorithm are used to find cluster centers [3]. In 2007, a combination of simulated annealing algorithm (SA) and k-harmonic means has been used for clustering which problems related to local

* Corresponding author. Email: s_poorramezani@yahoo.com

optimum have been fixed in this method [4]. Iiu et al in 2008 using tabu search algorithm has provided a method to deal with the issue of the Minimum Sum-of-Squares Clustering [5]. In 2011 a method was presented for clustering data using ant colony algorithm (Aco) which was also able to achieve good results [6] In 2011, Bee Colony Optimization algorithm (Bco) was used for classification problems [7]. Gravitational Search Algorithm (GSA) has also been used in clustering problems [8], as well as the composition of this algorithm with k-means was able to achieve good results [9]. The particle swarm optimization algorithm (PSO) is also successfully applied in clustering problems [10]. Also, in 2011, and combination of chaos and particle swarm optimization algorithms was used to avoid falling into local optimum [11]. Metaheuristic algorithms that have been inspired by nature have become popular and powerful algorithms for solving optimization problems [12, 13]. After the research conducted, it was discovered that ideas used in SCE and PSO algorithm can be combined, and a stronger Metheuristic can be generated to solve discrete or hybrid problems. The new Metheuristic is known as shuffled frog leaping algorithm (SFLA). SFLA algorithm is a new meta-heuristic algorithm with efficient mathematical function and global search functionality. This algorithm has been developed by Eusuff and Lansey [14]. In this study, an improved algorithm is used to perform clustering. The improved algorithm is related to research conducted at the Islamic Azad University of Mashhad, which with chaos and combination operator has been able to act properly in optimization problems [15] Also, because the use of concept of entropy in the objective function, we have been able to raise the efficiency of the algorithm for clustering. For the efficiency of the proposed algorithm, the four real datasets are used. In the second section, the concept of entropy is explained, and in the third section, SFLA algorithm is discussed. The fourth section outlines the improved-type SFL algorithm (LSFLA) and its application in clustering and in the fifth section, the results of the proposed method on four real data sets are presented, and it was compared with previous methods. Finally, a summary of the work performed in this study was explained.

2. Entropy

Entropy is considered as one of the effective ways to analyze uncertainties. Because of the conceptual and computational complexity of this theory to the early twentieth century, there was little interest in using it, until, Shannon in 1948, developed extensive research on the use of the theory in various fields of engineering (16). Entropy is a measure for the degree of disorder in any system, so that, entropy is more for higher the degree of disorder. In general we can say that, entropy is a measure of the number of internal modes, which a system can have, and can be calculated using equation (1).

$$\text{Entropy} = \sum_{i=1}^{c} p_i \log_2 p_i \qquad (1)$$

Where i is number of categories and pi is likely to be a member in the desired category. In this study, threshold -type Entropy is used. ($0 \leq p$) where p is the allowable threshold.

3. Shuffled Frog Leaping Algorithm (SFLA)

The combination of EA with local search is called memetic algorithms. SFL algorithm is a memetic metaheuristic algorithm memetic which mimics Evolutionary behavior of a group of frogs when looking for a place with the greatest amount of food. In SFLA search begins with random selection of a population of frogs that have covered the whole lagoon. Population is divided to m Memeplex that any Memeplex includes n frog. The total population is obtained from the equation. Frogs are arranged according to their fitness value. In each Memeplex, frog with the best and worst fitness and, is characterized with Xb and Xw respectively. Frog also is determined with the best fitness in the population

with Xg. Local search runs in each memeplex and the worst frogs are updated according to equation (2) and (3).

$$D_i = \text{rand}().(X_b - X_w) \quad (2)$$

$$X'_w = X_w + D_i, \; D_{min} \le D_i \le D_{max} \quad (3)$$

Where, rand is a random number between zero and one. Dmin and D max respectively, are lower and upper limit of the allowable range which a frog only in this area can change its position. If this process produces a better frog, the frog is replaced with the worst; otherwise, the best frog is replaced with the worst frog. If improvement is not achieved, a practical solution is randomly generated. In the end, all the frogs are mixed for global information exchange, and the steps continue to achieve pre-defined convergence criteria.

4. Improved Shuffled Frog Leaping Algorithm

In LSFLA, to fix flaws of SFLA and improve its capabilities, it was tried that changes be made on an algorithm, so that the random behavior of the algorithm to be replaced with chaotic behavior also, the combination operator was used to produce a better position to replace with the worst situation. The logistics function is a reverse mapping, as two values are obtained per, which is shown in equation (4). With applying this mapping in each phase on the response of before stage, different patterns in response can be observed over the time. The models depend on the value of the parameter α and also on the starting point.

$$X_{n+1} = \alpha X_n (1 - X_n) \quad (4)$$

Given that, on the basis SFLA algorithm, the initial population is generated randomly, so the random production has been replaced by a chaotic production. In the basis SFLA algorithm, the steps to improve the situation the worst member of the sub-memeplex are done in three stages, in the third step, a new position in space is randomly generated, and be replaced with the worst alternative. The combination is an exploratory operator that allows great leaps in the area related to the parents. In the LSFLA algorithm, the best member of the memeplex and the best member of the entire population is vonsidered as parents. Then considering the possibility p, moving takes place between parental chromosomes which is shown in Figure 1. The resulting child if you have better fitness, will be replaced with the worst alternative in the sub-populations.

4.1. The Use of LSFLS in Data Clustering

In this study, the LSFLA algorithm search capabilities have been used to find a specified number clustering. In general, the goal of this study is clustering the data properly and based on correct metric clustering data Euclidean distance between the centers of the clusters. Flowchart algorithm is shown in Figure 2. Next, details of LSFLA algorithm used in clustering are described.

Step 1: initialization: We choose two variables m, n so that m is number of memeplex and n is the number of frogs in any memeplex.

Step 2: Generating a virtual population using the logistics function. The number of F virtual frogs in the space, where k is the number of clusters and d is the number of decision variables. The i-th frog is displaced by a vector from the amounts of decision variables $U(i) = (U_i^1, U_i^2, ..., U_i^{kd})$ that may contain a possible solution for k centers of the cluster.

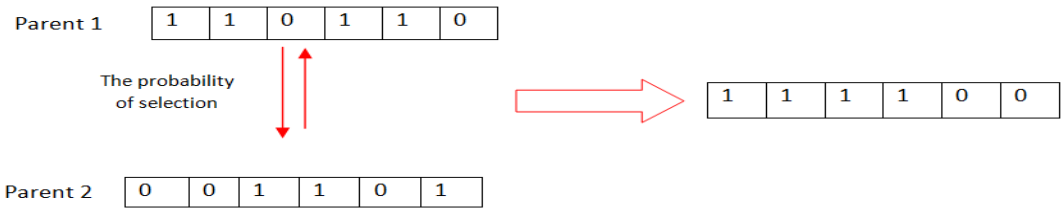

Fig.1. performance of combination operator in the SFLA algorithm

Step 3: Calculation of the Euclidean distance between generated centers (frogs) and data based on the equation (5).

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{n_d} (z_{pk} - m_{jk})\text{^}2} \qquad (5)$$

Where, Z_P represents the p-th data vector, m_j represents the j-th cluster center vector, nd input dimensions, ie the number of parameters of each vector and the index k represents the dimension. After classification of the data based on Euclidean distance, the fitness value for each frog F (i) is calculated based on the formula (6) as:

$$Fitness = Max(Entropy(k_i)),\ i = 1,2,\ldots,n \qquad (6)$$

Where, k_i is the number of clusters. In each category, the amount of entropy is calculated, the greater disorder will be associated with more entropy value and therefore, maximum entropy is selected as the fitness function.

Step 4: Sorting the frogs. We sort the F number frogs in descending order with respect to their performances. We keep them in an array $X = \{U(i), f(i), i = 1,\ldots F\}$. So i = 1 represents the frog with the highest level of performance. The place to keep the best frog in the general population, is P_X where $P_X = U(1)$.

Step 5: Categorizing Frogs into memeplex. X array is divided into m memeplex that each has n frog.

Step 6: Local search operations.

Step 7: mixing the memeplexes. After completion of certain stages of memetic evolution on each of the memeplexes, we insert $Y^1, Y^2, \ldots Y^m$ in the X. X array are arranged in a descending order based on performance and the place of the best member of the population be updated.

Step 8: Checking the convergence. If the convergence is satisfied, the program is over. Otherwise, return to Step 3. Decision on "when the algorithm must be finished?" is made through the numbers of the maximum number of time rings.

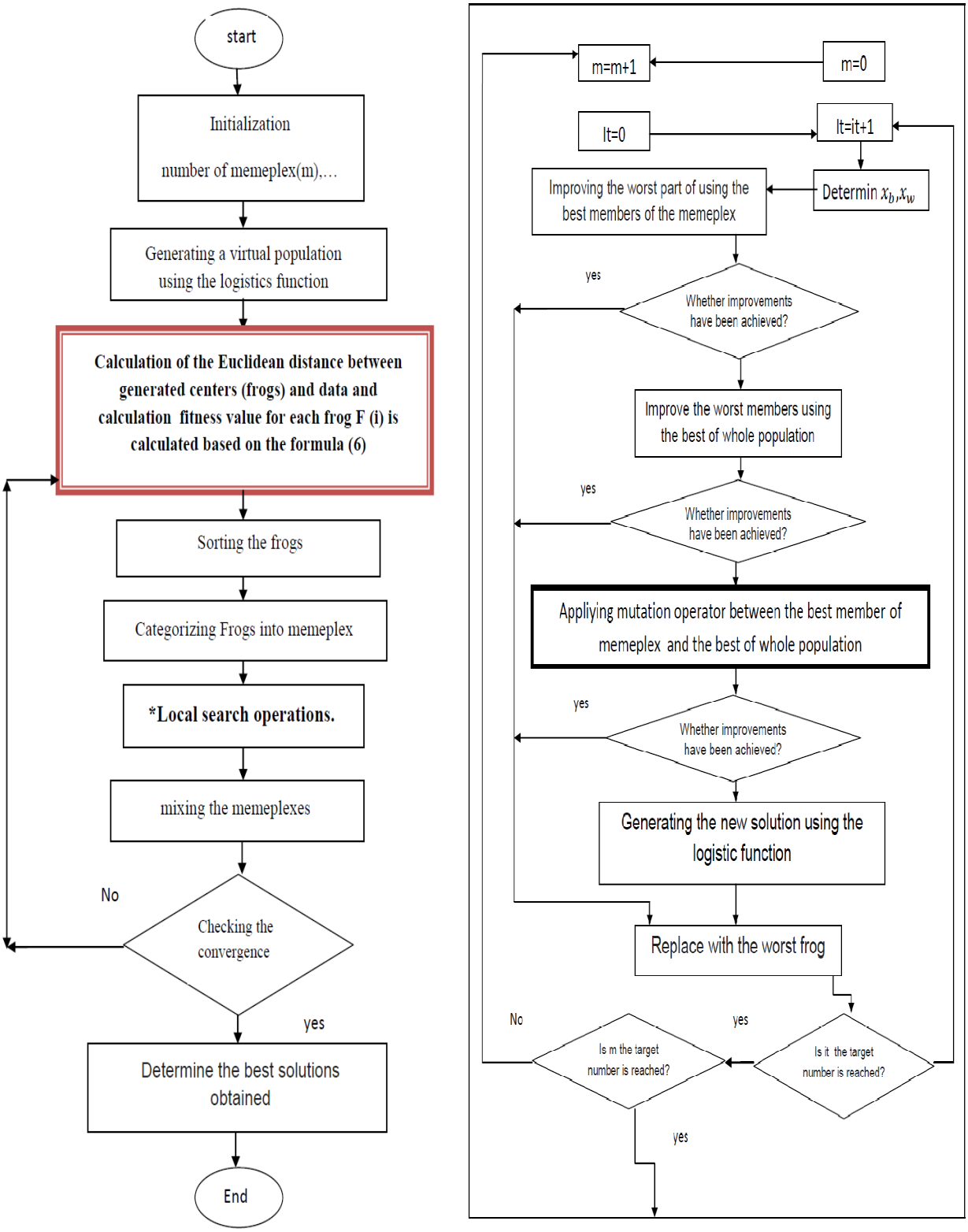

Fig. 2. Structure of LSFLA algorithm in the clustering

Left: Start, Setting the initial parameters (the numbers of memeplexes and the number of repetition in each memeplex), Generating a virtual population using the logistics function, Calculation of the Euclidean distance between generated centers (frogs) and data and calculation of the fitness value for each frog F (i) based on the formula (6), Sorting the frogs in descending order based on the fitness value, Categorizing Frogs into m memeplex, Mixing the memeplexes, Is there a convergence condition? Determinng the best solution, the end.

Right (local search): Improving the worst member using the best member of the memeplex, whether improvements have been achieved? Improving the worst member using the best member of the entire population, whether improvements have been achieved? Generating a new solution using the logistics function, Replacing with the worst frog? Has m reached to the desired size?

5. Simulations and Results

In this study, to evaluate the performance of the algorithm, the proposed method has been applied on four real datasets and compared with the algorithms k-means, GA, PSO and CPSO. Desired data is shown in Table 1.

Table 1
The data used

	The number of sample	The number of cluster	The number of feature
Iris	150	3	**4**
Wine	178	3	**13**
Glass	214	6	**9**
Cancer	683	2	**9**

5.1. Setting the Initial Parameters

MATLAB 2013 software was to implement algorithm and the algorithm runs on the similar hardware. Parameters required for the LSFLA algorithm are considered as follows. The number of memeplexes (m) is equal to 10 and the number of frogs (n) in a memeplex is equal to 20. Variable (q), the number of frogs selected for sub-memeplex is equal to 18 and local search in each sub-memeplex is repeated for 10 times. The α-parameter for logistic function is equal to 4 and the value of p for combiination operations is equal to 0.8. The scale parameter of research c at Leap step is equal to 2, and the threshold value at fitness function is considered equal to 0.15.

5.2. Evaluation Criteria

The proposed method is evaluated based on the error criterion. This criterion is calculated based on the equation (7).

$$error\ rate = \left(\frac{\sum_{i=1}^{n} I(A_i \neq B_i)}{n}\right) * 100 \quad (7)$$

Where I(0) denotes the indicator function.

In this equation, n is the number of total data. According to this equation, if i-th data is clustered properly, it is equal to 0, otherwise it will be 1. So whatever amount of error in clustering increases, error rate is higher.

5.3. Simulation Results

Table 2 shows average error rate and the best error rate for 20 reps. As can be seen, LSFLA algorithm has the best performance among other algorithms. For each data sets, the worst performance is related to the genetic algorithm, also k-meams algorithm tends to be in a local optimum. LSFLA method increases convergence rate and thus enables the LSFLA be better than other methods. This method classifies the data set with a mean value of 7.33, 28.08, 40.53 and 3.42 respectively for data Iris, Wine, Glass and Cancer. In other words, the method introduced in this study has the error rate less than other algorithms

which have led to a practical method in data clustering.

Table 2

Comparison of clustering algorithms error rate

dataset	Criteria	K-means	GA	PSO	CPSO	LSFLA
Iris	Averegae	17.80	18.2	12.53	10.00	**7.33**
	Best	10.67	10.02	10.00	10.00	**4.00**
Wine	Averegae	31.12	34.52	28.71	28.62	**28.08**
	Best	29.78	31.53	28.09	28.09	**28.08**
Glass	Averegae	39.20	46.89	43.35	40.38	**40.53**
	Best	37.41	41.67	39.48	34.12	**39.28**
Cancer	Averegae	4.39	8.59	5.11	3.51	**3.42**
	Best	4.35	5.27	3.66	3.51	**2.81**

6. Conclusion

In this study, to remove flaws such as falling into the local optimum, LSFLA algorithm is used for data clustering. The improved algorithm using the chaos and combination operators in the local search could improve the algorithm efficiency to achieve the optimal solution and acceptable results. As well as using entropy in the fitness function could increase the efficiency of the algorithm for clustering and the obtained results show that this method compared with k-means, GA, PSO, CPSO algorithms has less error rate and can be used as an efficient method for clustering problems.

References

[1] A. K. Jain, "Data clustering: 50 years beyond K-means " Pattern Recognition Letters, vol. 31, pp. 651-666, 2010. vol. 31, pp. 651-666, 2010.

[2] B. O. Pritesh Vora "A Survey on K-mean Clustering and Particle Swarm Optimization," International Journal of Science and Modern Engineering (IJISME), vol. 1, pp. 24-26, 2013.

[3] S. B. Ujjwal Maulik, "Genetic algorithm-based clustering technique," Pattern Recognition Letters, vol. 33, pp. 1455-1465, 2000.

[4] A. Ü. Zülal Güngör, "K-Harmonic means data clustering with tabu-search method," Applied Mathematical Modelling, vol. 32, pp. 1115-1125, 2008.

[5] Z. Y. Yongguo Liu, H. Wu, M. Ye, K. Chen, "A tabu search approach for the minimum sum-of-squares clustering problem," Information Sciences, vol. 178, pp. 2680-2704, 2008.

[6] Q. C. Lei Zhang, "A novel ant-based clustering algorithm using the kernel method," Information Sciences, vol. 181, pp. 4658-4672, 2011.

[7] C. O. Dervis Karaboga "A novel clustering approach: Artificial Bee Colony (ABC) algorithm," Applied Soft Computing, vol. 11, pp. 652-657, 2011.

[8] S. A. Abdolreza Hatamlou, H. Nezamabadi-pour, "Application of Gravitational Search Algorithm on Data Clustering," Rough Sets and Knowledge Technology, Springer, Berlin/Heidelberg, vol. 6954, pp. 337-346, 2011.

[9] S. A. Abdolreza Hatamloua, H. Nezamabadi-pourc, "A combined approach for clustering based on K-means and gravitational search algorithms," Swarm and Evolutionary Computation, vol. 6, pp. 47-52, 2012.

[10] Y. J. S. R.J. Kuoa, Zhen-Yao Chenc, F.C. Tien, "Integration of particle swarm optimization and genetic algorithm for dynamic clustering," Information Sciences, vol. 195, pp. 124-140, 2012.

[11] L.Y Chuanga, C.J. Hsiaob, "Chaotic particle swarm optimization for data clustering," Expert Systems with Applications, vol. 38, pp. 14555–14563, 2011.

[12] X.S. Yang, Metaheuristic Optimization: Nature-Inspired Algorithms and Applications: Springer Berlin Heidelberg, 2013.

[13] K. S. Debarati Kundua, S. Ghosha, S. Dasa, B.K. Panigrahib, Sanjoy Das, "Multi-objective optimization with artificial weed colonies," vol. 181, pp. 2441-2454, 2011.

[14] K. L. a. F. P. Muzaffar Eusuff, "Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization," Engineering Optimization, vol. 38, pp. 129-154, 2006.

[15] S. Poor Rajab, S. J. Sayyed Mahdi Chabok, G. Veysi, " Improvement of performance of Shuffled frog-leaping algorithm using the combination and chaos operator ", the second International Congress on Technology, Communication and Knowledge (ICTCK 2015), Islamic Azad University of Mashhad, 2015.

[16] C. Shannon, A mathematical theory of communication, bell System technical Journal vol. 27, pp. 379-423 and 623–656. Mathematical Reviews (MathSciNet): MR10, 133e, 1948.