Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 4 (2013) 243 – 249

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Intelligent Systems and Control

MKF-Cuckoo: Hybridization of Cuckoo Search and Multiple Kernel-Based Fuzzy C-Means Algorithm

Binu D[a*], Selvi M[a] and Aloysius George[a]

[a]*Aloy Labs, Bengaluru, India.*

Abstract

Discovering of optimal cluster through the help of optimization procedure is a recent trend in clustering process. Accordingly, several algorithms have been developed in the literature to mine optimal clusters. Most of the optimization-based clustering algorithms presented in the literature are only focused on the same objective given in the well-known clustering process, k-means clustering. Instead of k-means objective, some more effective objective functions are designed by the researchers for clustering. So, hybridization of those effective objectives with optimization algorithms can lead the effective clustering results. With the aim of this, we have presented a hybrid algorithm, called MKF-Cuckoo which is the hybridization of cuckoo search algorithm with the multiple kernel-based fuzzy c means algorithm. Here, MKFCM objective is taken and the same objective is solved through the cuckoo search algorithm which is one of the recent optimization algorithm proved effective in many optimization problems. For proving the effectiveness of the algorithm, the performance of the algorithm is comparatively analyzed with some other algorithm using clustering accuracy, rand coefficient, jaccard coefficient and computational time with iris and wine datasets. From the results, we can prove that the hybrid algorithm obtained 96% accuracy in iris data and 67% accuracy in wine data.

Keywords: Clustering, optimization, Cuckoo search, Multiple Kernel-Based Fuzzy C Means Algorithm (MKFCM), Clustering accuracy.

1. Introduction

Clustering is significantly needed for various applications due to the immense growth of data usage. With this significant application of clustering process, k-means clustering [1] was introduced initially with an

[*] Corresponding author
E-mail address: altimatebinu@gmail.com

doi:10.1016/j.aasri.2013.10.037

iterative procedure by considering the objective of minimizing the sum of squared error. After that, various authors were tried to find some problems in k-means clustering and introduced improved algorithms. One of the modified algorithms based on the fuzzy logic was fuzzy c-means clustering [2] that makes use of membership function and centroid computation procedure to iteratively find the best centroid. After introducing various algorithms by the researchers, the clustering algorithms are classified into two major categories, such as, partitional clustering and hierarchical clustering. On the other hand, some other different clustering algorithms were also presented in the literature which is grid-based clustering [3], projection-based clustering [4], subspace clustering [5] and density clustering [6].

Then, optimization algorithms were introduced for clustering process. In optimization-based clustering, researchers took clustering objective as minimum sum of squared error and they have used optimization procedure defined in their algorithm for solving clustering objective. Based on the similar kind of procedure, genetic algorithm [7], Particle Swarm optimization [8], bacterial foraging optimization [9], simulated annealing [10], artificial bee colony [11, 12] and firefly algorithm [13] were applied for clustering. Here, most of the algorithms have taken centroid sets as solutions to generate optimal cluster centroids. Meanwhile, some authors used hybrid optimization algorithms for clustering [14, 15], where two different optimization algorithms are combined to achieve effective results in clustering. At the same time, some simple modifications were done in standard optimization algorithm and then, it has been applied for clustering [16]. While obtaining more effective cluster results, k-means and FCM operator are additionally added into the optimization procedure by some authors [17, 18] with the objective of improving the convergence in clustering process.

After analyzing the literature completely about the clustering process, we have taken the objective defined in MKFCM algorithm [19] and this objective is solved using recent optimization algorithm, cuckoo search [20]. Here, solutions are encoded based on cluster centroid and optimal cluster centroids are achieved through the help of MKFCM objective function. The advantages of incorporating MKFCM objective into the clustering process is that kernel space can identify variance among the data point even better compared with the clustering done through only data space. Also, multiple kernels into one distance function can also lead to effective distance measurement so that effective clustering can be achieved since distance finding is the core process of any clustering process. According to the cuckoo search behavior in finding the best nest to lay egg, the searching of best cluster is done with the fitness of MKFCM objective. Here, new solutions are introduced based on levy flight equation that is effective in generating nearest neighbor solution with the inclusion of clustering objective. The paper is organized as follows: Section 2 presents the hybrid algorithm developed for clustering and section 3 provides experimentation performed for hybrid clustering and detailed results along with comparative analysis. Section 4 presents conclusion of the paper.

2. Proposed Methodology: MKF-cuckoo Algorithm

As per the literature review, the clustering process can be easily reformulated as optimization problem. The objective clustering optimization is that, to search of best cluster centroid with the help of iterative procedure. Accordingly, several techniques have been proposed in the literature for searching the optimal centroid in the problem space. But, most of the algorithm uses the same objective to find the optimal centroid with the help of minimizing the sum of squared distance in between the data objects. In our work, we have used an effective objective function that can lead to better results in searching the optimal centroid. With the motivation of this, we have utilized multiple kernel-based FCM algorithms and its objective has been taken for our research and it is combined with the recent cuckoo search algorithm to find the optimal centroid very quickly. The main contribution of the paper is hybridizing the cuckoo search which is one of the better algorithms for optimal search and with the MKFCM algorithm, which is one of the recent clustering algorithms proved effective in clustering. When hybridizing both the algorithms, the optimal search can be easy and at the same time, effective solution can be achieved because we have used MKFCM objective in optimal cluster centroid search. The whole process of the algorithm can be explained in three important phases such as 1) Initial solution encoding 2) Objective function for clustering 3) Algorithmic procedure for MKF-Cuckoo search.

2.1. Initial solution encoding

The searching of optimal solution through the help of any optimization algorithm needed an effective solution representation so that the searching efficiency to find quick solution can be improved. With the aim of this, the initial solution for our clustering process is random centroid taken from the input dataset. Suppose, initial nests given for the cuckoo search is P. Then, the algorithm takes P initial solutions from the input database. Every solution taken from the dataset contains $m*k$ matrix. Here, m represents the number of centroid needed and k represents the number of attributes given in the database. So, $m*P$ records are taken from the database and represented in P solution.

2. 2 Objective function for clustering

The objective function is taken from the clustering method given in [19] and it has been used as fitness function. Here, objective function considers distance minimization among the data points with its nearest neighbor cluster centroid. But, the distance finding makes use of kernel-based distance and fuzzy membership function. The overall objective of the hybrid clustering algorithm is denoted as below:

$$OB = \sum_{i=1}^{n}\sum_{j=1}^{m} u_{ij}{}^{b}(1 - k_{com}(x_i - m_j)) \qquad (1)$$

Here, $k_{com}(x_i - m_j)$ is represented as robust distance measurement in multiple kernel space. For the above objective function, the membership matrix $u_{ij}{}^{b}$ is computed as follows.

$$u_{ij}{}^{b} = \frac{(1 - k_{com}(x_i - m_j))^{\frac{-1}{b-1}}}{\sum_{j=1}^{m}(1 - k_{com}(x_i - m_j))^{\frac{-1}{b-1}}} \qquad (2)$$

The combined robust distance measurement in kernel space is the mathematical addition of two distances achieved from two kernels. Here, we have used two kernel such as, Gaussian kernel and tangential kernel.

$$k_{com}(x_i - m_j) = k_1(x_i - m_j) + k_2(x_i - m_j) \qquad (3)$$

$$k_1(x_i - m_j) = \exp\left(- \frac{\left\| x_i - m_j \right\|^2}{r^2} \right) \qquad (4)$$

$$k_2(x_i - m_j) = \tanh(\left\| x_i - m_j \right\| \qquad (5)$$

2.3. Algorithmic procedure for MKF-Cuckoo search

The detailed steps of the proposed MKF-Cuckoo search algorithm are described in this section. The pseudo code for the algorithm is given in figure 3.

Step 1: Initial set of P host nests: Initially, P solutions are given in an initial set of nests, every nest represents with $m*k$ matrix.

Step 2: Choose a random number (j) through levy flight: Here, through levy flight equation, a random number (j) is generated in between 1 to P and the corresponding solution (centroids sets) is chosen.

Step 3: Evaluate the fitness of nest located in the population corresponding to the random number. The fitness given in section 3.2 is used to find the fitness of the solution given in the j^{th} location of initial population.

Step 4: Choose a random nest (i) among P : Here, a random number generated in between 1 to P blindly and the solution given in i^{th} location of initial population is chosen to find the fitness of the solution.

Step 5: Replacing nest j by new solution if the fitness belongs to j is less than i : The evaluation of the fitness of both solutions taken from the previous steps is found out by comparing the fitness. The solution which is having less fitness is replaced with a new set of solution ($m*k$ matrix) using the following equation. The new solution $x^{(t+1)}$ for the worst nest is performed by,

$$x^{(t+1)} = x^{(t)} + \alpha \oplus Levy(\lambda) \tag{6}$$

Where, $\alpha > 0$ is the step size which should be related to the scales of the problem of interest. The product $\oplus$ means entry-wise multiplications. Levy flights essentially provide a random walk while their random steps are drawn from a Levy distribution for large steps,

$$Levy \sim u = t^{-\lambda}, (1 < \lambda \leq 3) \tag{7}$$

Once a new solution is generated, lower bound and upper bound conditions are checked. Every value placed in new solution of $m*k$ matrix is checked. If the value obtained in new solution based on levy flight is greater than upper bound, the values are replaced with upper bound value. If the value is less than lower bound, new value is replaced with lower bound value. The upper bound and lower bound value for every attributes are initially found out from the datasets. LB and UB is the maximum and minimum values of the every attributes presented in the input database.

Step 6: Remove worst nests based on the probability p_a : Based on the probability given in the algorithm, the worst set of nests from the population are identified and build new one in the corresponding location based on the step given in step 5.

Step 7: Keep the best one: The best set of nest is maintained in every iteration based on the fitness function and the process is continued from step 2 to 6 until the maximum iteration is reached.

3. Results and Discussion

The experimental results of the proposed MKF-Cuckoo are discussed in this section. The proposed algorithm is implemented in MATLAB and executed on a core i5 processor, 2.13 GHZ, 3 GB RAM computer.

3.1 Dataset Description

The two datasets such as iris and wine datasets are taken from Machine Learning Laboratory [21] for experimentation of the MKF-Cuckoo algorithm in clustering. *Iris dataset:* The data set contains three categories with equal number of data objects, where each category refers to a type of iris plant. There are 150 instances with four numeric features in iris data set. There is no missing attribute value. The attributes of the iris data set are sepal length in cm, sepal width in cm, petal length in cm and petal width in cm. *Wine dataset:* This dataset contains chemical analysis of 178 wines, derived from three different cultivars. Wine type is based on 13 continuous features derived from chemical analysis: alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyaninsm, color intensity, hue, OD280/OD315 of diluted wines and praline. The quantities of objects in the three categories of the data set are 59, 71 and 48, respectively.

3.2 Evaluation Metrics

For performance evaluation of the hybrid algorithm, we have used three different metrics, namely clustering accuracy, rand coefficient, jaccard coefficient.

Rand coefficient (R): It determines the degree of similarity between the known correct cluster structure and the results obtained by a clustering algorithm. It is defined as:

$$R = \frac{SS + DD}{SS + SD + DS + DD} \tag{8}$$

Where, SS represents both the data points belong to the same cluster and same group and SD represents both the data points belong to the same cluster but different groups. DS represents both the data points belong to different clusters but same group and DD represents both the data points belong to different clusters and different groups.

Jaccard coefficient (J): It is the same as rand coefficient except that it excludes DD and is defined as:

$$J = \frac{SS}{SS + SD + DS} \tag{9}$$

Clustering Accuracy (CA): The clustering accuracy computes the accuracy of grouping as compared with original labeled data. It is given as follows:

$$CA = \frac{1}{n}\sum_{j=1}^{m} \max_{i=1,2,\ldots,K} \left\{ \left| C_i \cap m_j \right| \right\} \tag{10}$$

Here, $C = \{C_1, C_2, \ldots, C_K\}$ is a labeled data set that offers the ground truth and $m = \{m_1, m_2, \ldots, m_m\}$ is a partition produced by a clustering algorithm for the data set.

3.3 Parameter Investigation

Cluster size and Number of iterations

The performance of the proposed hybrid clustering algorithm is analyzed with different cluster size such as 2, 3, 4 and 5. The proposed method is executed for different iterations and the performance is analyzed for iteration 10 and 100. The gradual results obtained are shown in the following table 1.

Table 1: Performance analysis in cluster size and number of iterations

Cluster size	Evaluation metrics	Iris dataset		Wine dataset	
		Iteration 10	Iteration 100	Iteration 10	Iteration 100
C=2	CA	0.680	0.667	0.669	0.517
	RC	0.555	0.518	0.554	0.498
	JC	0.306	0.274	0.328	0.249
	F	446.967	446.398	320.631	315.200
	Time	3.572	32.749	4.253	39.290
C=3	CA	0.893	0.787	0.517	0.511
	RC	0.423	0.516	0.571	0.572
	JC	0.327	0.330	0.250	0.247
	F	595.757	595.519	300.659	289.042
	Time	5.226	48.934	6.241	58.408
C=4	CA	0.953	0.653	0.410	0.399
	RC	0.364	0.561	0.592	0.599
	JC	0.321	0.282	0.200	0.212
	F	744.967	744.334	224.113	228.837
	Time	6.859	63.718	8.256	77.076
C=5	CA	0.959	0.887	0.455	0.354
	RC	0.583	0.425	0.628	0.612
	JC	0.360	0.322	0.233	0.177
	F	893.284	892.667	209.347	125.874
	Time	8.437	79.388	10.190	95.348

Here, the analysis can be done for both iris and wine datasets with different cluster size and different iterations. The accuracy for iris dataset is 68% in cluster size 2. So in order to increase the accuracy we increase the cluster size to 3, 4 and 5. So clustering accuracy increases to 89%, 95% and 96% simultaneously. When the cluster size increases the time automatically increased. But we obtained less accuracy values for 100 iterations when compared to 10 iterations. The accuracy is 89% in 100 iterations. For wine data set the accuracy is 67%. When comparing both data sets, iris has better results.

4. Conclusion

We have presented a hybrid algorithm, called MKF-Cuckoo which is the hybridization of cuckoo search algorithm with the multiple kernel-based fuzzy c means algorithm. Here, solutions are encoded based on the cluster centroid and the optimal cluster centroids are achieved through the cuckoo search algorithm which is one of the recent optimization algorithm proved effective in many optimization problems. The advantages of incorporating MKFCM objective into the clustering process is that kernel space can identify the variance of among the data point even better compared with the clustering done through only data space. Here, experimentation has been done using iris and wine dataset. For proving the effectiveness of the hybrid algorithm, comparative evaluation was done with the help of clustering accuracy, rand coefficient and jaccard coefficient. From the results, we identified that the hybrid algorithm obtained 95% accuracy in iris data and 67% accuracy in wine data. The future direction of this work is to incorporate the effective objective function and the modification in cuckoo search algorithm.

References

[1] J. McQueen. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of Fifth Berkeley Symposium on Math.Statistics and Probability. 1967;1 281-297.

[2] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981, ISBN: 0306406713.

[3] Wei-Keng Liao, Ying Liu. Alok Choudhary: A Grid-based Clustering Algorithm using Adaptive Mesh Refinement. Appears in the 7th Workshop on Mining Scientific and Engineering Data Sets. 2004.
[4] Mohamed Bouguessa and Shergrui Wang. Mining Projected Clusters in High-Dimensional Spaces. IEEE Transactions on Knowledge and Data Engineering. 2009;21:4.
[5] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the ACM-SIGMOD Int. Conf. Management of Data.1998; 94–105.
[6] M. Ester, H.-P.Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings Second International Conference on Knowledge Discovery and Data Mining. 1996; 226-231.
[7] Mualik, U., & Bandyopadhyay, S. Genetic algorithm based clustering technique. Pattern Recognition. 2002;33 1455–1465.
[8] K. Premalatha and A.M. Natarajan. A New Approach for Data Clustering Based on PSO with Local Search. Computer and Information Science. 2008;1:4.
[9] Miao Wan, Lixiang Li, Jinghua Xiao, Cong Wang,Yixian Yang. Data clustering using bacterial foraging optimization. J Intell Inf Syst. 2012;38 321–341.
[10] S. Z. Selim and K. Alsultan. A simulated annealing algorithm for the clustering problem. Pattern Recognit. 1991;10:24 1003–1008.
[11] Changsheng Zhang, Dantong Ouyang, Jiaxu Ning. An artificial bee colony approach for clustering. Expert Systems with Applications. 2010;37 4761–4767.
[12] Dervis Karaboga, Celal Ozturk. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied Soft Computing2011;11 652–657.
[13] J. Senthilnath, S.N. Omkar, V. Mani. Clustering using firefly algorithm: Performance study. Swarm and Evolutionary Computation. 2011; 164–171.
[14] S. Paterlini, T. Krink. Differential evolution and particle swarm optimisation in partitional clustering. Comput. Stat. Data Anal. 2006;50 1220–1247.
[15] T. Niknam, M. Nayeripour and B.Bahmani Firouzi. Application of a New Hybrid optimization Algorithm on Cluster Analysis. International Journal of Electrical and Computer Engineering. 2009;4:4.
[16] Swagatam Das, Ajith Abraham, Amit Konar. Automatic Clustering Using an Improved Differential Evolution Algorithm. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans. 2008;38:1.
[17] Krishna, K., & Murty. Genetic K-means Algorithm. IEEE Transactions on Systems Man and Cybernetics B Cybernetics, 1999;29 433–439.
[18] Hesam Izakian, Ajith Abraham, Vaclav Snasel Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization. World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, IEEE Press. 2009; 1690-1694.
[19] Long Chen, C. L. Philip Chen, and Mingzhu Lu. A Multiple-Kernel Fuzzy C-Means Algorithm for Image Segmentation.IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics. 2011;41:5 1263- 1274 .
[20] Yang, X. S. and Deb, S. Cuckoo search via L´evy flights. Proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009, India), IEEE Publications, USA/.2009; 210-214.
[21] UCI data Repository “http://archive.ics.uci.edu/ml/datasets.html”.