IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

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

Satish Chander[1], Vijaya P[2]

Waljat College of Applied Sciences, Rusayl, Muscat[1, 2]

Abstract: Clustering find various application in the variety fields like, telecommunication, medical image processing, bioinformatics and so on. This high demand poses a challenge to the researchers to develop an effective and efficient clustering algorithm for grouping the data objects. Accordingly, literature presents various algorithms for data clustering using partitional-based approaches. This paper presents a new clustering algorithm, namely, MKF-Firefly which is developed by combining the multiple kernel-based objective function and firefly algorithm. The firefly algorithm finds the optimal cluster centroids using the multiple kernel-based objective function. The centroids obtained from the firefly algorithm are then utilized for clustering process. The proposed clustering process is evaluated using Rand coefficient, Jaccard coefficient and Clustering Accuracy on the two different datasets like iris and wine. The proposed MKF-firefly achieved the clustering accuracy of 97% on the iris dataset.

Keywords: Clustering, Firefly, optimization, Multiple Kernel-Based clustering, Rand coefficient.

I. INTRODUCTION

The widespread use of the information technology leads to the amassing of the huge volume of data in many fields such as production, marketing, business etc. The bulk data must be grouped for the valid use. This leads to the development of the innovative methods to renovate the huge data into valuable information and knowledge. Data mining and machine learning communities are the approach to meet the requirement of the data clustering [10]. Clustering is useful for dividing large multidimensional data into distinguishable representative clusters. [9]. Clustering analysis has identified as a significant task in an extensive range of fields, whether for understanding or utility. In the framework of clustering, clusters are latent classes and cluster analysis is the process of datasets for automatically identifying classes. The following are some examples: Biology, Information retrieval, climate, Psychology and medicine. In context of clustering for utility, cluster analysis is the study of methods for discovering the most representative cluster types. They are summarization, Compression, Efficiently finding nearest neighbours etc [7, 8]. Cluster analysis groups data objects based only on the information that describes the object and their relationship. The aim is that the data records within a cluster be similar to one another and different from the data records in the other clusters [11, 12].

Generally, clustering algorithms are categorized into two major types, i) partitional clustering ii) hierarchical clustering. In partitional clustering, data groups are generated on every iterations by dividing the data into small subgroups. In hierarchical clustering, the data points are grouped based on the dendrogram which is a tree like structure for grouping the data. In partitional methods, K-means is one of the accepted clustering algorithms which is based on the iterative methods by dividing the data at every iteration. After the discovery of k-means algorithm [1], fuzzy c-means [2] was introduced by adding the fuzzy membership function within the clustering. After the development of FCM, various clustering algorithms have been developed in the literature by including the kernel function and various theories like, rough set and so on. Very recently, optimization algorithm has been included for clustering process. Here, the cluster centroids are found out through the heuristic search over the input data space using different search algorithms like, genetic algorithm, particle swarm optimization, cuckoo search, group search optimization, firefly algorithm and so on [4, 14, 15]. These heuristics algorithms are tried to find the centroids based on the different objective function which considers the kernel distance, the mean square distance and various clustering validity function.

Accordingly, in this paper, we have included multiple kernel-based clustering algorithm with firefly optimization. This paper considers the firefly algorithm [3] instead of cuckoo search algorithm which is utilized in [13]. At first, the data is pre-processed by removing the missing variables. Then, the input data is directly applied to firefly optimization for finding the cluster centroids. Here, firefly algorithm generates the random centroid and search towards the optima centroids using multiple kernel-based objective functions which utilizes two kernels such as, exponential and tangential kernel to find the distance among the data points. Every firefly is validated using this objective function and a firefly which is having the

DOI 10.17148/IJARCCE.2016.5742

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

minimum fitness is taken as the best centroid to perform clustering process. The selected cluster centroid is then used to group the data objects based on the minimum distance. The paper is organized as follows: section 2 discusses the proposed MKF-Firefly algorithm and section 3 discusses the results. The conclusion is presented in section 4.

II. PROPOSED METHODOLOGY: MKF-FIREFLY ALGORITHM FOR DATA CLUSTERING

This section presents the proposed MKF-firefly algorithm for data clustering. Here, input data is directly converted to object versus attribute format and the converted data is given to the firefly algorithm. The firefly algorithm search the given space to find the optimal centroid based on the MKF-based objective function. The obtained cluster centroid is then validated using three different metrics. The block diagram of the proposed data clustering is given in figure 1.

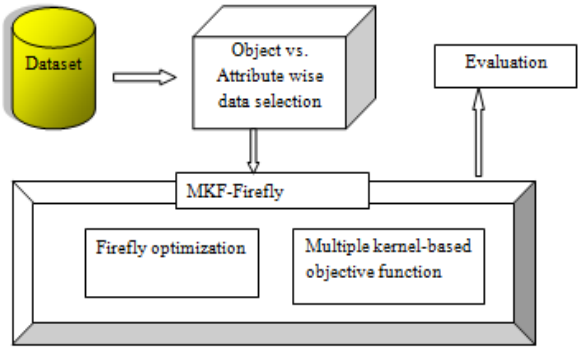

Figure 1. Block diagram of the proposed data clustering

A. Solution representation for firefly algorithm

The finding of optimal centroid through the help of any optimization processes requires an effective solution representation so that the searching efficiency to discover quick solution can be enhanced. With the aim of this, the initial solution for the clustering algorithm is random centroid obtained from the input dataset. Suppose, initial fireflies assigned for the firefly is P . Then, the algorithm reads P initial solutions from the input database. Every solution taken from the dataset contains $m * k$ matrix. Here, m indicates the required number of groups and k indicates the total features available in the database. So, $m * P$ data objects are considered from the database and given in P solution.

B. Intensity evaluation using the objective function

The intensity of the fireflies is validated by the objective function which is considered from the clustering algorithm given in [5] and it has been used as fitness function. Here, objective function utilizes distance minimization between the data points with its nearest neighbour cluster centroid. But, computing the distance utilizes kernel-based distance and fuzzy membership function. The main objective function of the hybrid clustering algorithm is given as below:

$$I = \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij}^{\ b} (1 - O_{com}(x_i - m_j)) \quad (1)$$

Here, $O_{com}(x_i - m_j)$ is the distance between the data points and centroid in multiple kernel space. For the above mathematical function, the membership matrix $u_{ij}^{\ b}$ is calculated as follows.

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

The integrated 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.

$$O_{com}(x_i - m_j) = O_1(x_i - m_j) + O_2(x_i - m_j) \quad (3)$$

$$O_1(x_i - m_j) = \exp\left(- \frac{\left\| x_i - m_j \right\|^2}{r^2} \right) \quad (4)$$

$$O_2(x_i - m_j) = \tanh(\left\| x_i - m_j \right\| \quad (5)$$

C. Algorithmic procedure for MKF-Firefly

The detailed steps of the proposed MKF-Firefly search algorithm are described in this section.

Step 1: At first, P number of fireflies is initialized randomly. These fireflies are known as population and every firefly within this population is represented as $m * k$ matrix.

Step 2: Find the light intensity of every firefly (I) using fitness function defined in section 2.2.

Step 3: Select two fireflies i and j to update it based on the following equation,

$$x(t+1) = x(t) + \beta * \exp\left[- \gamma * r_{ij}^{2} \right]\left(x_i(t) - x_j(t)\right) + \alpha_t \varepsilon_t \quad (6)$$

Where, α_t is a parameter for controlling the step size. ε_t is a vector drawn from Gaussian distribution. γ is absorption coefficient and r_{ij}^{2} is the distance between two fireflies.

Step 4: Once a new solution is generated, the algorithm ensure that each elements of new solution does not violate the lower bound and upper bound constraints. If the elements presented in new solution are beyond the upper limit, the elements are filled with upper bound value. If the element is smaller than lower bound, new element is filled

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

with lower bound value. The upper and lower bound elements of each feature are computed from the input data.

Step 5: The newly generated fireflies are given for fitness function which is used to find the intensity for every fireflies. If the newly generated fireflies have the greater intensity than the previous iteration, the new firefly is replaced in the same location. This process builds a new population of fireflies.

Step 6: Keep the best one: The best set of fireflies is kept up in each iteration based on the fitness function and the iteration is continued from step 2 to 5 until the maximum iteration is reached.

III. RESULTS AND DISCUSSION

The improvement of the proposed MKF-firefly algorithm is validated using two various datasets with three different evaluation metrics.

A. Experimental set up

Dataset Description: MKF-Firefly algorithm is experimented with two widely accepted datasets such as iris and wine which are obtained from Machine Learning Laboratory [6]. Iris dataset: The iris data consists of 150 data objects with four numeric attributes. Wine dataset: The wine data consists of 178 data objects which are obtained from three various cultivars. Here, 13 continuous attributes with one class attributes is presented.

Evaluation Metrics: Three different metrics, namely clustering accuracy, rand coefficient, jaccard coefficient are utilized here for validating the improvement of the proposed clustering algorithm. Rand coefficient (R): This measure intends to find the percentage of similarity level between the predefined cluster properties with the cluster results achieved by the clustering algorithm. It is defined as:

$$R = \frac{SS + DD}{SS + SD + DS + DD} \quad (7)$$

Where, SS represents the number of data records presented in both the cluster results and predefined classes. SD represents the number of data records presented in the same cluster but different classes. DS represents the number of data records available in different cluster results but same classes and DD represents the number of data records available in different clusters and different classes. Jaccard coefficient (J): This measure is similar to the rand coefficient but this one does not include DD and is defined as:

$$J = \frac{SS}{SS + SD + DS} \quad (8)$$

Clustering Accuracy (CA): The clustering accuracy is a metric used to measure the accuracy of grouping based on the original labelled data. It is computed based on the following equation:

$$CA = \frac{1}{n}\sum_{j=1}^{m} \max_{i=1,2,\dots,K} \left\{ \left| C_i \cap m_j \right| \right\} \quad (9)$$

Here, $C = \{C_1, C_2, \dots, C_K\}$ is a labelled data set that provides the ground truth and $m = \{m_1, m_2, \dots, m_m\}$ is groups generated by a clustering algorithm for the input data.

B. Experimental results

Table 1 shows the performance analysis of the proposed MKF firefly algorithm for various number of cluster size and number of iterations. When the number of iteration is fixed to 10, the proposed MKF firefly achieved the accuracy of 0.7. 0.9, 0.96 and 0.96 for the various number of cluster sizes from 2 to 5.

Similarly, for the iteration of 100, the proposed MKF-firefly obtained the accuracy of 0.71, 0.91, 0.965 and 0.97 for various number of cluster sizes in iris dataset. The jaccard coefficient of 0.4, 0.4, 0.4 and 0.5 is achieved by the proposed MKF-firefly in iris dataset for various number of cluster sizes.

Similarly, in wine dataset, the proposed MKF-firefly obtained the maximum accuracy of 0.68 when the number of cluster size is fixed to two and number of iteration is equal to 10. Also, for the iteration of 100, the proposed MKF-firefly obtained the maximum accuracy of 0.69 when the number of cluster is fixed to two. In terms of rand coefficient, the proposed MKF-firefly obtained the maximum value of 0.64 when the number of iteration is fixed to 100 and number of cluster is fixed to five.

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

Cluster size		**C=2**			**C=3**			**C=4**			**C=5**		
Evaluation metrics		CA	RC	JC	CA	RC	JC	CA	RC	JC	CA	RC	JC
Iris dataset	**Iteration 10**	0.7	0.56	0.4	0.9	0.43	0.4	0.96	0.5	0.4	0.96	0.6	0.5
	Iteration 100	0.71	0.56	0.3	0.91	0.52	0.5	0.96	0.6	0.45	0.97	0.62	0.6
Wine dataset	**Iteration 10**	0.68	0.6	0.4	0.55	0.6	0.4	0.5	0.6	0.5	0.56	0.63	0.5
	Iteration 100	0.69	0.61	0.5	0.56	0.62	0.5	0.55	0.61	0.51	0.59	0.64	0.55

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

IV. CONCLUSION

This paper presented a multiple kernel-firefly algorithm for the centroid estimation in the data clustering. Here, firefly algorithm was combined with the multiple kernel-based objective function which is based on exponential and tangential kernel function. The firefly algorithm was successfully adapted within the clustering procedure based on the solution encoding procedure. This algorithm found the optimal cluster centroid and it was used for grouping the data objects. The experimentation was performed using benchmarked iris and wine datasets and evaluation is done using the clustering accuracy, rand and jaccard coefficient. The results showed that the proposed algorithm is effective with improved clustering accuracy for various numbers of required clusters. In future, firefly algorithm can be modified to find the centroids vary rapidly.

REFERENCES

[1] J. McQueen, "Some Methods for Classification and Analysis of Multivariate Observations", In Proceedings of Fifth Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 281-297, 1967.
[2] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms". New York: Plenum Press, 1981, ISBN: 0306406713.
[3] J. Senthilnath, S.N. Omkar, V. Mani, "Clustering using firefly algorithm: Performance study", Swarm and Evolutionary Computation, pp. 164–171, 2011.
[4] Krishna, K., & Murty, "Genetic K-means Algorithm", IEEE Transactions on Systems Man and Cybernetics B Cybernetics, vol. 29, pp. 433–439, 1999.
[5] 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, vol. 41, no. 5, pp. 1263- 1274, 2011.
[6] UCI data Repository "http://archive.ics.uci.edu/ml/datasets.html".
[7] S. Kotsiantis, P. Pintelas, "Recent advances in clustering: a brief survey", WSEAS Trans. Information Science & Applications, vol. 1, no. 1, pp. 73–81, 2004.
[8] J.S. Lee, S. Olafsson, "Data clustering by minimizing disconnectivity", Information Sciences, vol. 181, pp. 732–746, 2011.
[9] RuiXu and Donald Wunsch, "Survey of Clustering Algorithms", IEEE transactions on neural networks, Vol. 16, No. 3, pp. pp. 645-677, May 2005.
[10] M. Yuwono, S. W. Su, B. D. Moulton, and H. T. Nguyen, "Method for increasing the computation speed of an unsupervised learning approach for data clustering," in Proc. IEEE CEC, Jun. 2012, pp. 2957–2964.
[11] PradiptaMaji, "Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data", IEEE transactions on systems, man, and cybernetics—part b: cybernetics, Vol. 41, No. 1, pp. 222-253, February 2011.
[12] Xuesong Yin, Songcan Chen, Enliang Hu, Daoqiang Zhang, "Semi-supervised clustering with metric learning: An adaptive kernel method", Pattern Recognition, vol. 43, pp. 1320–1333, 2010.
[13] D. Binu, M. Selvi, Aloysius George, "MKF-Cuckoo: Hybridization of Cuckoo Search and Multiple Kernel-based Fuzzy C-means Algorithm", AASRI Procedia, AASRI Conference on Intelligent Systems and Control, Vol. 4, pp. 243–249, 2013.
[14] 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, vol. 4, no. 4, 2009.
[15] 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, vol. 38, no. 1, 2008.