Decision Science Letters 3 (2014) 319–324

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

An efficient approach based on differential evolution algorithm for data clustering

Maryam Hosseini*, Mehdi Sadeghzade and Reza Nourmandi-Pour

Department of Computer Engineering, Science and Research Branch of Sirjan, Islamic Azad University, Sirjan, Iran

C H R O N I C L E

Article history:
Received October 15, 2013
Received in revised format
March 6 2014
Accepted March 24, 2014
Available online
March 26 2014

Keywords:
Data clustering
K-means algorithm
Differential evolution algorithm

A B S T R A C T

Clustering plays an essential role for data analysis and it has been widely used in different fields like data mining, machine learning and pattern recognition. Clustering problem divides some data, which is more similar to each other in terms of parameters under consideration. One of available methods about this area is k-means algorithm. Despite dependency of this algorithm on initial condition and convergence to local optimal points, it classifies *n* data to *k* clusters with high speed. Since we encounter a huge volume of data in clustering problems, one of suitable methods for optimal clustering is to use a meta-heuristic algorithm, which improves clustering operation. In this paper, differential evolution algorithm is utilized for solving available problems in k-means algorithm. In this paper, meta-heuristic algorithm has been used for solving data clustering problems. The applied algorithm has been compared with k-means algorithm on six known dataset from UCI database. Results show that this algorithm achieves better clustering than k-means algorithm.

1. Introduction

Data clustering is one of the most complicated engineering problems and it is considered as an NP-hard problem. Data clustering is challenging issue as the size of the data set problem increases, it becomes a time consuming issue and there are literally tremendous efforts to reduce the complexity of this class of problems (Leung et al., 2000; Qian, 2008). Clustering algorithms are generally classified as hierarchical clustering and partition clustering (Han et al., 2006; Frigui & Krishnapuram, 1999). Hierarchical clustering normally groups data objects with a sequence of partitions, using singleton clusters either to a cluster including all individuals or vice versa. Hierarchical procedures can be either agglomerative or divisive: agglomerative algorithms start with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and divides it into successively smaller clusters (Jain et al., 1999; Rokach & Maimon, 2005). Partition procedures that we concerned in this paper, try to divide the data set into a set of

* Corresponding author.
E-mail addresses: maryamhosseini37@yahoo.com (M. Hosseini)

doi: 10.5267/j.dsl.2014.3.006

disjoint clusters without the hierarchical structure. The most popular partition clustering algorithms are the prototype-based clustering algorithms. In this method, each cluster is normally represented by the center of the cluster and the used objective function is the sum of the distance from the pattern to the center (Mirkin, 1998; Eiben & Smith, 2003). K-means algorithm is one of well-known algorithms in solving data clustering problem because of its implementation simplicity and high speed to locate local optimum and the method has been successful in clustering many problems. However, high dependency of the results on initial state of this algorithm and convergence possibility of problem to local optimum (instead of global optimum) creates difficulty to solve many problems (Liu et al., 2011). Therefore, to handle data clustering problem, researchers use possible optimization methods such as genetic algorithm, firefly algorithm and imperialist competitive algorithm (Shelokar et al., 2004; Bin et al., 2002; Yang, 2009). In this article, differential evolution has been implemented for solving data clustering problem (Storn & Price, 1997; Jiawei & Kamber, 2001; Noman & Iba, 2008). This algorithm is a random algorithm based on population and it is one of the evolutionary algorithms. Differential evolution algorithm has been implemented to improve previous works. The results from this algorithm are compared with results from K-means algorithm. This paper is organized as follows. We first introduce the clustering problem in section 2. The implementation of the differential evolution algorithm is introduced in section 3 and 4. Finally, the experiments and results presented and are discussed in sections 5 and 6.

2. The clustering problem

Clustering is used by organizing models in humongous clusters for discovering intra-relation and inters-collection of samples and models. Clustering can be used in some areas such as analysis of similarity or dissimilarity and decreasing volume and data dimensions for data pre-processing. Many factors must be considered for solving clustering problem such as similarity standards or initial conditions. One of usual standards for measuring similarity among samples is Euclidean distance which is defined by Eq. (1).

$$d(x,c) = \sqrt{\sum_{i=1}^{m} \left(x_i - c_j\right)^2} \tag{1}$$

where x is a data vector, c is cluster centers, and m is the number of data. In fact, in an N-dimensional space, we consider each n sample as a point in the space and then we assign these points to k cluster based on pre-defined standards. One of important and famous algorithms in solving data clustering problem is k-means algorithm, which preserve a simple and relatively high speed to obtain local optimal solution. The primary objective of this algorithm is to find k center of the cluster. However, despite its success in solving most of the clustering problem, k-means algorithm cannot solve many problems due to dependency of the procedure on initial value of centers and early convergence and being in local optimum.

3. Differential evolution algorithm

Storn and Price (1997) are believed to be the first who proposed differential evolution algorithm. This algorithm is a random algorithm based on population and it is one of the evolutionary algorithms. It is similar to the other evolutionary algorithms and it is unique method for producing new solution in differential evolution algorithm. For solving optimization problems, it uses sampling target function in multiple selective points, randomly. Predetermined parameters constraints specify some regions in which initial population must be produced. Differential evolution algorithm produces a new solution in d-dimensional space and this new solution results from difference of available points. This method works by regulating three parameters. Parameter CR is probability of doing crossover, parameter NP is population size and parameter F is mutation weigh, which is multiplied by difference of tow vectors and added to the third vector.

3.1. Mutation

In this section, three vectors are selected randomly, two by two. Mutated vector is generated by Eq. (1) for each vector within population. Another operator has been proposed for mutation, too. This design increases greedy degree of algorithm movement towards optimization by using an optimum vector. Convergence is very suitable in order to increase speed especially for problems where their general optimum is found, very easily (Gong et al., 2008)

$$V_{i,G+1} = X_{r1,G} + F * (X_{r2.G} + X_{r3,G}) \quad (2)$$

3.2. Crossover

In the crossover step, each of mutated vector components is transferred to candidate vector (by probability CR). CR is the crossover constant $\in$ [0, 1], otherwise, equivalent component is substituted for main vector.

$$u_{ji,G+1} = \begin{cases} v_{ji,G+1} & If\ rand(j) \leq CR\ or\ j = randb() \\ x_{ji,G} & Else \end{cases} \quad (3)$$

In Eq. (3), rand (j) is j^{th} call of random function, which is a number within [0, 1]. However, for ensuring that, at least, one component is transferred to experimental vector, one component is transferred from mutated vector to experimental vector (without regard to CR), randomly. Therefore, one component is selected randomly for each candidate vector by using function randb() for transferring to the next generation (Bergey & Ragsdale, 2005; Storn & Price, 1997; Bäck & Schwefel, 1993).

3.3. Selection

Greedy method is applied for selecting survivors where each vector is compared with related candidate vector and each one, which is more competent, is transferred to the next generation. Fig.1 shows the DE algorithm.

$$x_{i,G+1} = Fitness\ Vector(u_{i,G+1}, x_{i,G}) \quad (4)$$

Differential Evolution Algorithm
Generate P=(x_1,x_2,…,x_n); N(point in D)
Repeat
  For *i=1* ***to*** *N* ***do***
    Compute a mutant ***u****;*
    Create ***y*** *by the crossover of* ***u*** *and* $\mathbf{x_i}$*;*
    If *f(**y**) < f(*$\mathbf{x_i}$*)* ***then*** *insert* ***y*** *into Q*
      Else *insert* $\mathbf{x_i}$ *into Q*
  End *if;*
 End *for;*
 P:= Q;
Until *stopping condition;*

Fig. 1. Pseudo-code of the DE algorithm

4. Proposed clustering method

Differential evolution algorithm uses a differential operator for producing new solutions and this operator exchanges information among population members. One of advantages of this algorithm is to have a memory, which keeps information of suitable solution in the recent population. The other advantage of this algorithm is associated with operator of its selection. In this algorithm, all members

of one population have equal chance to be chosen as one of the parents. In this algorithm, in lieu of each member, a gene donor is produced which performs exchanging by that member.

4.1. The first cost function

One of the parameters used to evaluate the clustering problem is the distance to within cluster and the proposed method of this paper uses Eq. (5) as follows,

$$Cost\ Function1 = \frac{1}{M}\sum_{i=1}^{M} ||Data_i - Cl_i|| \quad (5)$$

where Cl_i is a cluster center in $Data_i$, M is number of data and the maximum distances among centers of clusters are obtained.

4.2. The second cost function

In this study, the average distance between the clusters centers are also considered to be Eq. (6) is calculated.

$$Cost\ Function2 = \frac{1}{c*(c-1)}\sum_{i=1}^{c}\sum_{j=1}^{c} ||Cl_i - Cl_j|| \quad (6)$$

4.3. The third cost function

The third cost function to minimize the total cost function values is obtained from both the first and second as follows,

$$Final\ Cost = (Cost\ Function\ 1 + \frac{1}{Cost\ Function\ 2}) \quad (7)$$

5. Experimental Results

In this section, the results of the proposed algorithm are compared with algorithm k-means for solving clustering problem. Implementation of this algorithm has been accomplished by MATLAB software package. The parameters of DE algorithm are N_{pop}=50, number of cluster centers=3, CR= 0.9, mutation factor=0.01, iteration=100. Used data set includes Pima, Glass, Wine, Breast cancer (original), Breast cancer (diagnostic) and Iris, which have been extracted from known data base UCI. The characteristics of each dataset are described next.

Brest Cancer Wisconsin (Original): This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. This data set has 699 samples and 2 different classes. Every sample has 9 attributes.

Pima (Pima Indians Diabetes Database): this data set is about Pima Indians Diabetes that has totally 768 samples in 2 classes. These classes are about Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/ (height in m) ^2), Diabetes pedigree function, and Age (years) and each data has 8 attributes.

Iris (fisher's iris plants database): This data set is according to the Iris flowers recognition that has three different classes and each class consists of 50 samples. Every sample has four attributes.

Glass (glass identification database): this data set is about several types of glass that has totally 214 samples in 6 classes.
These classes are about building_windows_float_processed, vehicle_windows_float_processe, containers, ableware, building_windows_non_float_processed and headlamps and each data has 9 attributes.

Wine (wine recognition data): This data set is regarding to drinks recognition with 178 samples classified into three different classes including 59, 71 and 48 samples, respectively. In this data set, each sample has 13 attributes.

BCW (Wisconsin Diagnostic Breast Cancer): this data set is about Diagnostic breast cancer collected at the University of Wisconsin and it contains two different classes including 357 and 212 samples. In this data set, each sample has 30 features. By this change, clustering producer is done with more accuracy. In addition, clustering centers are determined with higher accuracy.

Table 1
The results of the implementation of the algorithms DE and k-means, based on inter-cluster distance

Dataset/Algorithm	K-means	DE
BCW (original)	0.5438	0.4494
Pima	0.7257	0.3434
Glass	0.5517	0.3045
Iris	0.4650	0.1954
Wine	0.6248	0.5177
BCW (Diagnostic)	0.6269	0.6050

Table 2
The results of the implementation of the DE algorithm based on inter-cluster distance and extra-cluster distance

Dataset/Algorithm	Inter- cluster	Extra-cluster	Inter- cluster& Extra-cluster
BCW (original)	0.4494	0.4122	0.4848
Pima	0.3434	0.4340	0.4384
Glass	0.3045	0.4109	0.4305
Iris	0.1954	0.6150	0.3651
Wine	0.5177	0.3503	0.5873
BCW (Diagnostic)	0.6050	0.2425	0.6235

Continue on differential evolution algorithm is executed for each of the data sets and the results are shown in a diagram convergence for 150 iterations in Fig. 2.

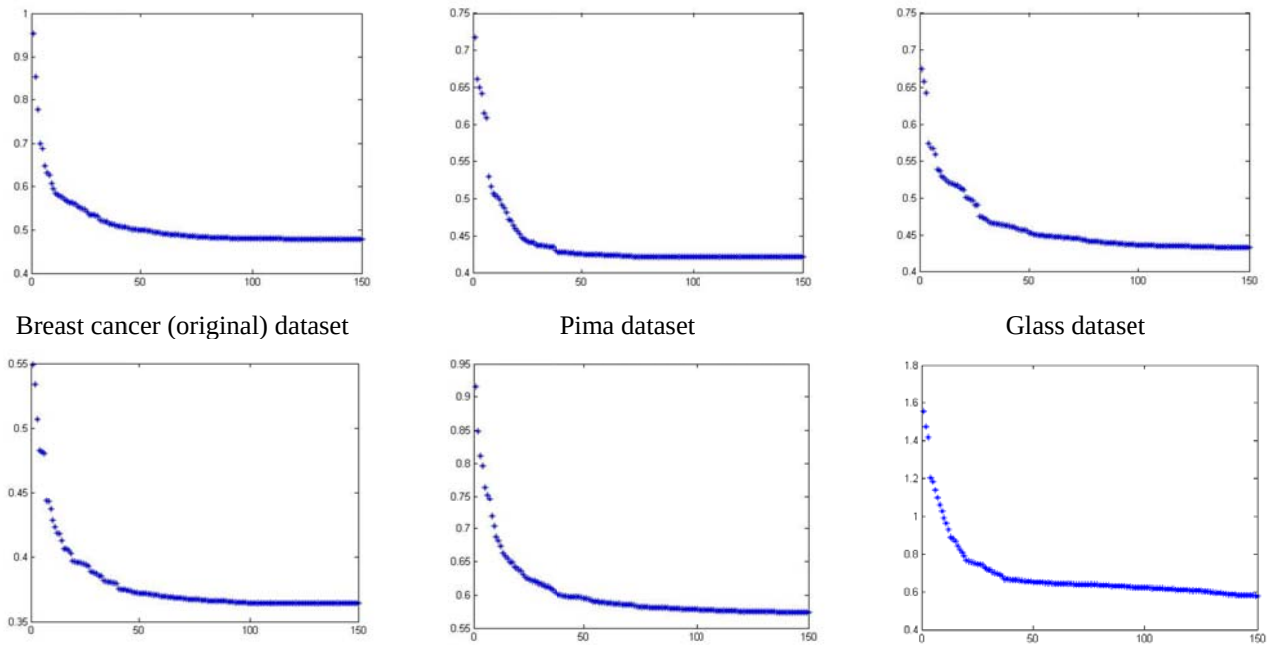

Fig. 2. Convergence diagram of differential evolution algorithm for different dataset

As we can observe from Fig. 2, differential evolution algorithm has provided more accurate clustering than algorithm k-means. According to results, this algorithm can be applied in clustering problems for finding optimal or near-optimal solutions for dividing *N* object in *K* clusters.

6. Conclusions

In this paper, we have presented a new differential evolution algorithm for solving problem of data clustering. The proposed algorithm measures the least distance between the available data and their centers. In addition, the farthest distance among centers of clusters was also obtained. Next, the proposed method considered the sum of these distances by some coefficients for each one as target function. In order to measure the performance of the proposed model, the results were compared with k-means algorithm based on cluster distance. The preliminary results indicate that the proposed algorithm provides more acceptable results than k-means algorithm.

References

Bergey, P. K., & Ragsdale, C. (2005). Modified differential evolution: a greedy random strategy for genetic recombination. *Omega*, *33*(3), 255-265.

Bin, W., Yi, Z., Shaohui, L., & Zhongzhi, S. (2002, May). CSIM: a document clustering algorithm based on swarm intelligence. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on* (Vol. 1, pp. 477-482). IEEE.

Bäck, T., & Schwefel, H. P. (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary computation*, *1*(1), 1-23.

Eiben, A. E., & Smith, J. E. (2003). *Introduction to evolutionary computing*. springer.

Frigui, H., & Krishnapuram, R. (1999). A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(5), 450-465.

Gong, W., Cai, Z., & Jiang, L. (2008). Enhancing the performance of differential evolution using orthogonal design method. *Applied Mathematics and Computation*, *206*(1), 56-69.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Jiawei, H., & Kamber, M. (2001). Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, *5*.

Leung, Y., Zhang, J. S., & Xu, Z. B. (2000). Clustering by scale-space filtering.*Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(12), 1396-1410.

Liu, Y., Wu, X., & Shen, Y. (2011). Automatic clustering using genetic algorithms. *Applied Mathematics and Computation*, *218*(4), 1267-1279.

Shelokar, P. S., Jayaraman, V. K., & Kulkarni, B. D. (2004). An ant colony approach for clustering. *Analytica Chimica Acta*, *509*(2), 187-195.

Mirkin, B. (1998). *Mathematical classification and clustering: From how to what and why* (pp. 172-181). Springer Berlin Heidelberg.

Noman, N., & Iba, H. (2008). Accelerating differential evolution using an adaptive local search. *Evolutionary Computation, IEEE Transactions on*, *12*(1), 107-125.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer US.

Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, *11*(4), 341-359.

Qian, W. (2008). Adaptive differential evolution algorithm for multiobjective optimization problems. *Applied Mathematics and Computation*, *201*(1), 431-440.

Yang, X. S. (2009). Firefly algorithms for multimodal optimization. In *Stochastic algorithms: foundations and applications* (pp. 169-178). Springer Berlin Heidelberg.