Performance Analysis of Standard PSO based K-means Algorithm for Clustering

Ashanta Ranjan Routray

Fakir Mohan University
Email: ashan2r@yahoo.co.in

Abstract : Since the last two decades, K-means clustering algorithm is playing a major role in the data mining research community. Unfortunately, some factors like improvement in the quality of cluster centers and minimal intra cluster distance has been a major limitations of K-means algorithm. By inspiring this in this paper, we have integrated the standard Particle swarm optimization algorithm with K-means algorithm to get the optimal cluster centers. The proposed hybrid algorithm able to produce global solution and the results of the SPSO-K-means is found to be superior than others.

Keywords: K-means, Particle swarm optimization, GA

I. INTRODUCTION

Clustering is an unsupervised technique whose goal of is to recognize the structure in an unlabeled data set by independently organizing data into homogeneous groups. Clustering is necessary when no labeled data are available regardless of whether the data are binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, multimedia, or mixtures of the above data types. Among the other clustering algorithms, K-means is one of the simple, efficient center based hard clustering technique used to solve different real world problems. In K-means, the no. of clusters are considered as ‘k’, which helps to group the similar objects in a closer fashion as well as make distance from the dissimilar type. Depending on the distance measure from the center, the k sets of clusters are divided into another k sets of subset clusters. Different optimization techniques can be used iteratively to form new cluster centers. It has been applied in various real life applications including all the fields of engineering. Some recently proposed k-means clustering algorithms and their applications relevant to the article have been studied.

A hybrid clustering method based on k-means and PSO for better convergence of the developed algorithm have been proposed by Ahmadyfard and Modares(2008). Santosa and Ningrum(2009) have shown that the performance of their proposed CSO based K-means performs better than PSO based clustering. Wang et. al. (2012) introduced a parallel map reduce K-PSO by combining the traditional k-means and PSO algorithm. Yao et. al. have combined the k-means method and mathematical morphology for fish image optimization. A modified k-means method for quasi unsupervised learning by controlling the size of the cluster partitions and adjusted by means of the Levenberg–Marquardt algorithm have been proposed by

This paper proposes an evolutionary standard PSO based K-means algorithm for data clustering. The parameters of both the K-means and PSO have been suitably chosen during the experimental analysis. We have calculated the inter and intra cluster distances between the clusters by using the proposed method. The rest of the paper is organized as follows: section 2 outlines the preliminaries; section 3 defines the proposed work. The experimental analysis and the result discussion have been indicated in section 4 and section 5 concludes our work.

II. PRELIMINARIES

2.1 K-means Algorithm

In a multidimensional space, the algorithm takes k number of input parameters and performs the partition on a set of n objects. First, a number of objects called clusters (K no.) are selected which are represented as cluster means. From these clusters, a new cluster mean is computed based on the distance metric between the object and the cluster mean. This process is continued until the convergence of criterion function is met for which k-means is able to find the best cluster center points in the space.

Steps of k-means Algorithm

1. Randomly select the predefined number of cluster centers from the dataset.

2. Calculate the Euclidian distances of each clusters from cluster centers.

3. Assign cluster number to each instance based on Euclidian distance. An instance i_j is assigned to cluster ck if Euclidian distance is minimum between i_j to c_k.

4. Find out new cluster center by computing the mean of all instances in a cluster.

5. If the previous sets of cluster centers are same as new clustering center, then go to step-7.

6. Else go to step-2

7. Exit

2.2 Standard Particle Swarm Optimization (SPSO)

Kennedy and Eberhart first introduced the Particle Swarm Optimization, which is inspired by the behavior of flying birds. Basically PSO is a population based algorithm whose complexity is less than other evolutionary algorithms, as it takes less parameter. The principle of PSO is basically lied on the following assumptions: (i) Birds flying at a location having no mass or dimension in a multidimensional space, by adjusting their positions and exchanging messages about the current position in search space according to their own earlier experience and their neighbors (ii) While travelling in a group for either food or shelter , they will not collide between themselves and also adjust both their position and velocity. In this mechanism, the swarm members modify their positions as well as the velocities after communicating their group information according to the best position appeared in the current movement of the swarm. The swarm particles would gradually get closer to the specified position and finally reach the optimal position with the help of interactive cooperation. Each particle has to maintain their local best positions lbest and the global best position gbest among all of them.

$$V_i^{(t+1)} = V_i^{(t)} + c_1 * rand(1) * \left(l_{best_i}{}^{(t)} - X_i{}^{(t)}\right) + c_2 * rand(1) * \left(g_{best}{}^{(t)} - X_i{}^{(t)}\right) \quad (1)$$

$$X_i{}^{(t+1)} = X_i{}^{(t)} + V_i{}^{(t+1)} \quad (2)$$

The cognition and social behavior of particles are being controlled by eq. (1) and next position of the particles are updated using eq. (2), V_i (t) and V_i(t+1) are the velocity of i^{th} particle at time t and t+1 in the population respectively, c_1 and c_2 are acceleration coefficient normally set between 0 to 2(may be same), X_i(t) is the position of i^{th} particle and lbest$_i$(t) and gbest(t) denotes the local best particle of i^{th} particle and global best particle among local bests at time t, rand(1) generates a random value between 0 to 1.

III. PROPOSED WORK

This section deals with the development of proposed SPSO-K-means for data clustering. The steps of standard PSO have been integrated with the K-means. The main purpose of the proposed algorithm is to compare the technique with other techniques along with to obtain the effective cluster centers. The intra cluster distance is computed based on the mean of the maximum distance between two data vectors within a instances of clusters. The aim is to minimize the intra cluster distance. Likewise, the intercluster distance is computed based on the minimum distance between centroids of clusters. The objective function for the fitness calculation of the clusters is given by eq. (3).

$$F(X_i) = \frac{k}{\left(\sum_{j=1}^{m} \sum_{i_k \in C_{i,j}} \left(i_k - C_{i,j}\right)^2\right) + d} \quad (3)$$

Proposed SPSO-K-means Algorithm

```
Initialize the position P and velocity V of particles randomly. Each particle is a potential solution for the clustering problem. A single particle represents the centroids of clusters. Hence the population of particles is initialized as follows (eq. 4):
$P = \{X_1, X_2 ... X_n\}$                (4)
Where $X_i$ represents the centroids of clusters which is a single possible solution (particle) in the search space and can be denoted in eq. (5).
$X_i = \{C_{i,1}, C_{i,2} ... C_{i,m}\}$                (5)
Where $C_{i,j}$ represents jth cluster center among m clusters in the datasets.
Iter=1;
While ($n_i$<=maxIter) (where $n_i$ is the no. of iteration)
Compute fitness of all particles $X_i$ in population P by using eq. (3).
If ($n_i$==1)
            Assign Local best particle lbest=P.
Else
            Evaluate fitness of P and P'.
            Compare the fitness of particles based on their fitnesses.
If fitness of $i^{th}$ particle $X_i$ in P is less that fitness of a particle in P'
Then assign $L_{best}$ (i) = P'(i).
Else assign $L_{best}$ (i) = P(i).
End of if
End of if
Select particles with best fitness value from Lbest as Gbest particle.
Compute new velocity $V_{new}$ of the particle by using eq. (1).
Generate next positions of particles by using eq. (2).
$n_i$ = $n_i$ +1;
End of while
```

IV. EXPERIMENTAL SET UP AND RESULT ANALYSIS

In this section, the detailed experimental set up and the obtained result analysis of the proposed method has been discussed. The performance of the proposed SPSO-K-Means has been compared with other existing methods like K-Means and GA-K-Means. For testing the proposed method, we have considered the data sets from UCI repository [20] and have been compared in terms of fitness value of the cluster centers from eq. (3). All the resultant values including the proposed SPSO-K-means algorithm has been listed in table-1. The proposed method has been implemented using MATLAB 9.0 on a system with an Intel Core Duo CPU T5800, 2GHz processor, 2 GB RAM and Microsoft Windows-2007 OS.

In the objective function equation (eq. (3)), are the used to calculate the fitness of clustering methods are calculated by using the parameters k and d. For the simulation purpose, the following values of k=50 and d=0.1 are considered. The acceleration coefficients c1 and c2 are set to 1.6 for early convergence during SPSO iteration. The inertia weight is taken between 1.8 to 2 for early convergence. The proposed Standard PSO method is able to produce a good cluster center of an abject as compared to other existing methods.

Table 1. Performance Comparison of SPSO-K-means with other techniques

Dataset	Algorithm	Fitness Value	Intra-cluster distance	Inter-cluster distance
Breast cancer data	K-means	5.211876209	18.58314	13.83092
	GA-K-means	5.229876211	18.29017	13.82344
	SPSO-K-means	5.240981201	18.27012	13.81029
Iris	K-means	0.012395396	2.65109	1.81029
	GA-K-means	0.013826351	2.66091	1.80287
	SPSO-K-means	0.014528017	2.63153	1.79028
Haberman	K-means	0.000317745	50.92872	17.03827
	GA-K-means	0.000328364	50.91827	17.04729
	SPSO-K-means	0.000328364	50.91639	17.02981
Glass	K-means	0.892658471	3.087683	3.65342
	GA-K-means	0.897209857	3.030937	3.71339
	SPSO-K-means	0.912389808	3.069736	3.63928
Contraceptive Method Choice	K-means	7.80139E-05	19.35789	8.46826
	GA-K-means	8.03819E-05	19.31086	8.52918
	SPSO-K-means	8.20198E-05	19.28053	8.46378
Wine	K-means	101.0578291	458.76281	327.57811
	GA-K-means	101.1652809	458.72105	327.59872
	SPSO-K-means	101.1789326	458.70309	327.51973
Spect heart	K-means	0.069341756	2.301928	1.08368
	GA-K-means	0.072648917	2.360837	1.09329
	SPSO-K-means	0.076041565	2.290367	1.02797

V. CONCLUSION

This paper investigated the performance of the proposed SPSO-K-means algorithm by considering different real world data sets. The SPSO-K-means method is compared with the other two standard techniques like K-means and GA-K-means. From the simulation results, it is observed that the proposed method gives superior performance as compared to the other techniques. The comparisons are made on the basis of fitness value, intra cluster distance and inter cluster distance among the clusters. In future, some more parameters can be tuned to obtain the better results with the PSO technique.

REFERENCES

[1] T.Warren Liao, Clustering of time series data—a survey, Pattern Recognition 38 (2005) 1857 – 1874.

[2] Hartigan, J.A. : Clustering algorithms. 1975, John Wiley & Sons, Inc.

[3] Hartigan, J. A., Wong, M. A. : Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society, Series C 28 (1): 100–108. JSTOR 2346830 (1979).

[4] A., Ahmadyfard, Modares, H. : Combining PSO and k-means to Enhance Data Clustering. 2008 International Symposium on Telecommunications, DOI: 10.1109/ISTEL.2008.4651388 , (2008) 688 - 691 .

[5] Santosa, B., Ningrum, M. K. : Cat Swarm Optimization for Clustering, 2009 International Conference of Soft Computing and Pattern Recognition. DOI: 10.1109/SoCPaR.2009.23.

[6] Wang, J., Yuan, D., Jiang, J. : Parallel K-PSO Based on Map Reduce. DOI: 10.1109/ICCT.2012.6511380 (2012) 1203 - 1208

[7] Yao, H., Duan, Q., Li, D., Wang, L. : An improved K-means clustering algorithm for fish image segmentation. Mathematical and Computer Modelling. **58** (2013) 790–798.

[8] Monedero, D.R., Solé, M., Nin, J., Forné, J. : A modification of the k-means method for quasi-unsupervised learning. Knowledge-Based Systems. **37** (2013) 176–185.

[9] Shahbaba, M., Beheshti, S.: MACE-means clustering, Signal Processing. **105**(2014)216–225.

[10] Tzortzis, G., Likas, A. : The Min Max k-Means clustering algorithm. Pattern Recognition. **47**(2014) 2505–2516.

[11] Naldi, M.C., Campello, R.J.G.B. : Evolutionary k-means for distributed datasets, Neurocomputing. **127** (2014)30–42.

[12] Hartigan, J.A. : Clustering algorithms. John Wiley & Sons. Inc (1975).

[13] Hartigan, J. A., Wong, M. A. (1979): Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C 28 (1): 100–108. JSTOR 2346830.

[14] Kennedy, J., Eberhart, R.: Particle swarm optimization. Proceedings of the 1995 IEEE International Conference on Neural Networks. vol. 4 (1995) 1942–1948.

[15] Wei, J., Guangbin, L., Dong, L.: Elite particle swarm optimizaion with mutation, IEEE Asia Simulation Conference – 7th Intl. Conf. on Sys. Simulation and Scientific Computing. (2008) 800–803.

[16] Khare, A., Rangnekar, S. : A review of particle swarm optimization and its applications in Solar Photovoltaic system. Applied Soft Computing. 13 (2013) 2997–3006.

[17] Babaei, M. : A general approach to approximate solutions of nonlinear differential equations using particle swarm optimization. Applied Soft Computing 13 (2013) 3354–3365.

[18] Neri, F., Mininno, E., Iacca, G.: Compact Particle Swarm Optimization. Information Sciences 239 (2013) 96–121.

◈◈◈