An Improved Swarm Based Hybrid K-Means Clustering for Optimal Cluster Centers

Janmenjoy Nayak, Bighnaraj Naik, D.P. Kanungo and H.S. Behera

Abstract Clustering is a frequently used unsupervised pattern recognition technique based on the grouping properties of data. K-means is one of the best known, simple and efficient method of data clustering. But this method is more sensitive to the initial cluster partitioning and suffers in local optimal cluster centers. In this paper, an attempt has been made to hybridize the K-means algorithm with the improved Particle Swarm Optimization (PSO) to improve fitness of cluster centers. The strategy of finding global best solution of IPSO is used to avoid the possibility of falling at local optimal cluster centers. The proposed method IPSO-K-means have been compared with K-means, GA-K-means and PSO-K-means and found better in terms of objective value than the others. Simulation result shows that the proposed method is effective, steady and stable and is more suitable for cluster analysis.

Keywords Clustering · K-means · Improved PSO

1 Introduction

Clustering is an unsupervised data mining technique and is based on the concept of similarity measures between the cluster groups. The aim of the clustering is to distinguish and reform the clusters of either similar or dissimilar type relying on their distance from the cluster center. K-means clustering is one of the competent clustering

J. Nayak (✉) · B. Naik · D.P. Kanungo · H.S. Behera
Department of Computer Science Engineering and Information Technology,
Veer Surendra Sai University of Technology Burla, Sambalpur 768018, Odisha, India
e-mail: mailforjnayak@gmail.com

B. Naik
e-mail: mailtobnaik@gmail.com

D.P. Kanungo
e-mail: dpk.vssut@gmail.com

H.S. Behera
e-mail: mailtohsbehera@gmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_54

techniques for solving large scale non convex optimization problems [1]. This method is useful to reduce the sum of intra cluster distances between the clusters. The algorithm follows a simple concept of classification of a data set into a number of clusters in a dimensional space. The features of the cluster are represented through a data point and relying on the homogeneity condition for which the clusters are separated. The numbers of clusters are considered as 'k' (called prior knowledge) helps to group the similar objects in a closer fashion as well as make distance from the dissimilar type. Based on the distance measure from the center, the k sets of clusters are divided into another k sets of subset clusters. Each time the newly formed cluster centers can be iteratively updated by using various optimization techniques. Many researchers have shown their key interest in developing k-means algorithm for diversified application areas. A number of recently proposed k-means clustering algorithms and their applications relevant to the article have been studied in this literature.

To achieve the appropriate cluster centers in the feature space for optimizing the similarity metrics, a no. of GA based clustering algorithm have been developed [2, 3]. The Ahmadyfard and Modares [4] have discussed a hybrid clustering method based on k-means and PSO for better convergence. A novel cat swarm optimized clustering algorithm have been proposed by Santosa and Ningrum [5] for better accuracy as compared to PSO. Kader [6] presented a hybrid two-phase GAI-PSO with k-means data clustering algorithm which performs fast data clustering and can avoid premature convergence to local optima. An improved PSO based k-means algorithm was developed by Zheng and Jia [7] to avoid the local optima problem in normal k-means clustering. Wang et al. [8] introduced a parallel map reduce K-PSO by combining the traditional k-means and PSO algorithm. Naik et al. [9] have proposed a hybrid PSO—K-means clustering algorithm to get optimal cluster centers for cluster analysis. An improved k-means with a hybridized PSO algorithm for web document clustering has been introduced by Jaganathan and Jaiganesh [10]. After the combination of k-means method and mathematical morphology, Yao et al. [11] have developed an improved k-means method for fish image optimization. Monedero et al. [12] presented a modification of the celebrated k-means method for quasi unsupervised learning by controlling the size of the cluster partitions and adjusted by means of the Levenberg–Marquardt algorithm. Shahbaba and Beheshti [13] introduced a novel minimum ACE k-means (MACE) clustering method which has the advantage for the use in synthetic and real data. Tzortzis and Likas [14] developed a minmax k-means algorithm where the cluster weights are set according their variance. To deal with distributed data and overcome the limitations of k-means, Naldi and Campello [15] proposed an evolutionary k-means algorithm for clustering.

Although k-means is a highly influential clustering algorithm used in various real life applications compared to other algorithms, still it has some major limitations like sensitivity to local optimal solutions in which area more works need to be done. By inspiring this, an improved swarm based k-means algorithm has been proposed for more effective and competent real world data clustering. The remaining part of the paper is organized as following manner. Section 2, describes the basic preliminary concepts like k-means, PSO and IPSO. In the Sect. 3, the proposed method (IPSO-k-means)

has been presented. Section 4 presents the experimental set up along with the results obtained. Section 5 gives the conclusion of our work.

2 Preliminaries

2.1 K-Means Algorithm

The k-means algorithm [16, 17] receives k number of input parameters and performs the partition on a set of n objects in the dimensional space. The method of k-means starts with the random selection of k no of objects and are represented as cluster means. Depending on the distance metric between the object and the cluster mean, for each of the residual objects, a similar object is being assigned which helps to compute a new cluster mean. This process will be continued till the convergence of criterion function. Hence, k-means is able to find the best cluster center points in the space.

Steps of k-means Algorithm

1. **Select** predefined number of cluster centers randomly from the dataset.
2. **Compute** Euclidian distances of each instance from cluster centers.
3. **Assign** cluster number to each instance based on Euclidian distance. An instance i_j is assigned to cluster c_k 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 PSO Algorithm

Particle Swarm Optimization is bird inspired metaheuristic with random selection of initial populations proposed by Kennedy and Eberhart [18]. Due to lesser parameter settings, the complexity of this population based algorithm is quite less than others. The epitome for the expansion of PSO was to consider a location having no mass or dimension, flying like a bird in multidimensional space, by adjusting its position and exchanging information about the current position in search space according to its own earlier experience and that of its neighbors [19]. While travelling in a group for either food or shelter [20], not only the behavior of various types of swarms indicates a unique indication towards the noncoliding nature between themselves, but also they 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 [21]. The swarm particles would gradually get closer to the specified

position and finally reach the optimal position with the help of interactive cooperation [22]. Each particle has to maintain their local best positions lbest and the global best position gbest among all of them.

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

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

Equation 1 controls both cognition and social behavior of particles and next position of the particles are updated using Eq. (2), Vi(t) and Vi(t + 1) are the velocity of ith particle at time t and t + 1 in the population respectively, c1 and c2 are acceleration coefficient normally set between 0 and 2 (may be same), Xi(t) is the position of ith particle and $\text{lbest}_i(t)$ and gbest(t) denotes the local best particle of ith particle and global best particle among local bests at time t, rand(1) generates a random value between 0 to 1.

2.3 Improved PSO Algorithm

In traditional PSO, the basic three steps like calculation of velocity, position and the fitness value will be iterated till the required criteria of convergence are met. The ending criteria may be the maximum change in the best fitness value. However, if the velocity of the swarm will be fixed to zero or nearer to that and the best position will have a fixed value, and then the PSO may lead to be trapped at some of local optima. This happens only due to the swarm's experience on the current and global positions. This experience is to be avoided and should be based on the mutual cooperation among all the swarms in a multidirectional manner [23].

So, in IPSO a new inertia weight factor λ is introduced to control both the local and global search behavior. The value of λ may be decreased quickly [24] during the initial iterations and slowly during the optimal iterations.

The new velocity and position updation can be realized through the Eqs. (3) and (4).

$$V_i^{(t+1)} = \lambda * V_i^{(t)} + c_1 * \text{rand}(1) * \left(l_{best_i}^{(t)} - X_i^{(t)}\right) + c_2 * \text{rand}(1) * \left(g_{best}^{(t)} - X_i^{(t)}\right) \quad (3)$$

$$X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \quad (4)$$

3 Proposed Algorithm (IPSO-K-Means)

The proposed IPSO-K-means algorithm is a hybrid algorithm based on the combination of improved PSO with K-means algorithm for real world data clustering. Due to the slow convergence speed of basic PSO and easier finding of a local

optimal solution in K-means algorithm, the hybridization of Improved PSO along with K-means algorithm will improve the convergence speed as well as helps to find the global optimal solution. So, the advantages of both the algorithms have been used in this paper, which may lead to achieve an efficient result than the use of any individual algorithms.

Pseudo Code of IPSO-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. 5):

$$P = \{X_1, X_2 \ldots X_n\} \tag{5}$$

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. (6).

$$X_i = \{C_{i,1}, C_{i,2} \ldots C_{i,m}\} \tag{6}$$

where $C_{i,j}$ represents jth cluster center among m clusters in the datasets.

Iter=1;
While (iter<=maxIter)
Compute fitness of all particles X_i in population P by using the following objective function in eq. 7:

$$F(X_i) = \frac{k}{\left(\sum_{j=1}^{m} \sum_{i_k \in C_{i,j}} \left(i_k - C_{i,j}\right)^2\right) + d} \tag{7}$$

If (iter==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 Lbest (i) = P'(i).
Else assign Lbest (i) = P(i).
End of if
End of if
Select particles with best fitness value from Lbest as Gbest particle.
Compute new velocity Vnew of the particle by using P, Lbest and gbest as follows:

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

Generate next positions of particles P' by using P and Vnew as follows:

$$X_i^{(t+1)} = X_i^{(t)} + Vnew_i^{(t+1)}$$

Iter = iter+1;
End of while

4 Experimental Setup and Result Analysis

In this section, the proposed IPSO-K-Means has been implemented in MATLAB and compared with other alternatives (K-Means, GA-K-Means, PSO-K-Means). All the clustering methods are tested with multidimensional real world datasets (Table 1) from UCI repository [25] and have been compared in terms of fitness value of the cluster centers from Eq. (7). The comparison of clustering methods is listed in Table 2. The proposed method has been implemented using MATLAB 9.0 on a system with an Intel Core Duo CPU T5800, 2 GHz processor, 2 GB RAM and Microsoft Windows-2007 OS.

Table 1 Dataset information

Datasets	No. of pattern	No. of clusters	No. of attributes
Iris	150	3	4
Lenses	24	3	4
Haberman	306	2	3
Balance scale	625	3	4
Wisconsin breast cancer	699	2	10
Contraceptive method choice	1473	3	9
Hayesroth	132	3	5
Robot navigation	5456	4	2
Spect heart	80	2	22

Table 2 Performance Comparison of IPSO-K-means with the other clustering methods

Datasets	Fitness values of clustering algorithms			
	K-means	GA-K-means	PSO-K-means	IPSO-K-means
Iris	0.012395396	0.013826351	0.014528017	0.014580183
Lenses	0.339904827	0.351735427	0.360239542	0.360282035
Haberman	0.000317745	0.000328364	0.000348162	0.000363902
Balance scale	0.002573387	0.002628475	0.002810827	0.002920182
Wisconsin breast cancer	7.25935E-14	7.26287E-14	7.28928E-14	7.32602E-14
Contraceptive method choice	7.80139E-05	8.03819E-05	8.20198E-05	8.21983E-05
Hayes roth	4.59807E-05	4.70825E-05	4.73918E-05	4.74029E-05
Robot navigation	0.001583094	0.001828362	0.001898018	0.001928362
Spect heart	0.069341756	0.072648917	0.076041565	0.078284661

In the Eq. (7), k and d are the parameters used to calculate the fitness of clustering methods along with the proposed method. The simulation has been carried out by setting the values k = 50, d = 0.1 and proposed clustering model found better from all existing methods. The acceleration coefficients c1 and c2 are set to 1.4 for early convergence during IPSO iteration. The inertia weight is set between 1.8 and 2 for early convergence. The proposed Improved PSO based technique is able to produce a good cluster center of an abject. But there is no certain time to meet the convergence criteria. With the increase in the number of iterations, the cluster center (initially chosen) will be attracted towards its corresponding similar clusters which will lead to obtain the final cluster center with best fitness value. The change in local and global best solution will result the updation in the new position and velocity of the cluster.

5 Conclusion

In this paper, a hybrid Improved swarm based K-means algorithm has been designed for the purpose of real world data clustering. The datasets have been considered from the UCI machine learning repository and are tested by various clustering methods like K-means, GA-K-means and PSO-K-means. The fitness value of the clusters obtained by the proposed method helped to get the more nearer and optimal cluster centers. The proposed method not only produces good fitness values but also it improves the cluster accuracy. The procedure to find the optimal cluster center in this paper is quite different and innovative as compared to other existing methods. The results shown in Table 2 from selected data sets indicate that the IPSO-K-means technique is able to find the global optimum solution with small standard deviations as compared to other methods. However, the future work may be planned for optimization of the initial cluster centers of k-means algorithm with the use of any other hybrid techniques.

Acknowledgments This work is supported by the Department of Science & Technology (DST), Ministry of Science & Technology, New Delhi, Govt. of India, under grants No. DST/INSPIRE Fellowship/2013/585.

References

1. Bai, L., Liang, J., Sui, C., Dang, C.: Fast global k-means clustering based on local geometrical information. Inf. Sci. **245**, 168–180 (2013)
2. Maulik, U., Bandyopadhyay, S.: Genetic algorithm-based clustering technique. Pattern Recogn. **33**, 1455–1465 (2000)

3. Bandyopadhyay, S., Maulik, U., Mukhopadhyay, A.: Multiobjective genetic clustering for pixel classification in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. **45**(5), 1506–1511(2007)
4. Ahmadyfard, A., Modares, H.: Combining PSO and k-means to Enhance Data Clustering. In: 2008 International Symposium on Telecommunications, pp. 688–691 (2008). doi:10.1109/ISTEL.2008.4651388
5. Santosa, B., Ningrum, M. K.: Cat swarm optimization for clustering. In: 2009 International Conference of Soft Computing and Pattern Recognition. doi:10.1109/SoCPaR.2009.23
6. Kader, A.R.F.: genetically improved pso algorithm for efficient data clustering. In: 2010 Second International Conference on Machine Learning and Computing. doi:10.1109/ICMLC.2010.19
7. Zheng, X., Jia, Y.: A study on educational data clustering approach based on improved particle swarm optimizer. In: 2011 International Symposium on IT in Medicine and Education (ITME), pp. 442–445 (2011). doi:10.1109/ITiME.2011.6132144
8. Wang, J., Yuan, D., Jiang, J.: Parallel K-PSO based on map reduce, pp. 1203–1208 (2012). doi: 10.1109/ICCT.2012.6511380
9. Naik, B., Swetanisha, S, Behera, D.K., Mahapatra, S., Padhi, B.K.: Cooperative swarm based clustering algorithm based on PSO and k-means to find optimal cluster centroids. In: National Conference on Computing and Communication Systems (NCCCS), pp. 1–5 (2012). doi: 10.1109/NCCCS.2012.6413027
10. Jaganathan, P., Jaiganesh, S.: An improved K-means algorithm combined with particle swarm optimization approach for efficient web document clustering, pp. 772–776 (2013). doi: 10.1109/ICGCE.2013.6823538
11. Yao, H., Duan, Q., Li, D., Wang, L.: An improved K-means clustering algorithm for fish image segmentation. Math. Comput. Model. **58**, 790–798 (2013)
12. Monedero, D.R., Solé, M., Nin, J., Forné, J.: A modification of the k-means method for quasi-unsupervised learning. Knowl.-Based Syst. **37**, 176–185 (2013)
13. Shahbaba, M., Beheshti, S.: MACE-means clustering. Sig. Process. **105**, 216–225 (2014)
14. Tzortzis, G., Likas, A.: The min max k-means clustering algorithm. Pattern Recogn. **47**, 2505–2516 (2014)
15. Naldi, M.C., Campello, R.J.G.B.: Evolutionary k-means for distributed datasets. Neurocomputing **127**, 30–42 (2014)
16. Hartigan, J.A.: Clustering algorithms. Wiley, New York (1975)
17. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: a K-means clustering algorithm. J. R. Statis. Soc. Ser. C **28**(1), 100–108 (1979). (JSTOR 2346830)
18. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the 1995 IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
19. Wei, J., Guangbin, L., Dong, L.: Elite particle swarm optimization with mutation. In: IEEE Asia Simulation Conference— 7th International Conference on System Simulation and Scientific Computing, pp. 800–803 (2008)
20. Khare, A., Rangnekar, S.: A review of particle swarm optimization and its applications in Solar Photovoltaic system. Appl. Soft Comput. **13**, 2997–3006 (2013)
21. Babaei, M.: A general approach to approximate solutions of nonlinear differential equations using particle swarm optimization. Appl. Soft Comput. **13**, 3354–3365 (2013)
22. Neri, F., Mininno, E., Iacca, G.: Compact particle swarm optimization. Inf. Sci. **239**, 96–121 (2013)
23. Yue-bo, M., Jian-hua, Z., Xu-sheng, G., Liang, Z.: Research on WNN aerodynamic modeling from flight data based on improved PSO algorithm. Neurocomputing **83**, 212–221 (2012)

24. Dehuri, S., Roy, R., Cho, S.B., Ghosh, A.: An improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification. J. Syst. Softw. **85**, 1333–1345 (2012)
25. Bache, K., Lichman, M.: UCI machine learning repository [http://archive.ics.uci.edu/ml], Irvine, CA, University of California, School of Information and Computer Science (2013)