Volume 4, Issue 12, December 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Hybrid Technique for Data Clustering Using Genetic Algorithm with Particle Swarm Optimization

Sundararajan S *
Associate Professor and Head/MCA
SNS College of Technology, Coimbatore
Tamil Nadu, India

Dr. Karthikeyan S
The Director/Computer Applications
Karpagam University, Coimbatore
Tamil Nadu, India

Abstract—Data clustering is useful in several areas such as machine learning, data mining, wireless sensor networks and pattern recognition. The most famous clustering approach is K-means which successfully has been utilized in numerous clustering problems, but this algorithm has some limitations such as local optimal convergence and initial point understanding. Clustering is the procedure of grouping objects into disjoint class is known as clusters. So, that objects within a class are extremely similar with one another and dissimilar with the objects in other classes. Firefly algorithm is mainly used for clustering problems, but it also has disadvantages. To overcome the problems in firefly this work used a proposed method of Hybrid K-Mean with GA/PSO. The hybrid method merges the standard velocity and modernizes rules of PSOs with the thoughts of selection from GAs. They compare the hybrid algorithm to the standard GA and PSO approaches. Experimental results show that the proposed method used to reduce the limitations and improve accuracy rate.

Keywords— Clustering, K-Mean, Firefly algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO).

I. INTRODUCTION

Clustering is one of the unsupervised learning branches where a set of sample, frequently vectors in a multi-dimensional dimensions, are joined into clusters in such a way that sample in the same cluster are similar in some sense and patterns in different clusters are dissimilar in the same sense. Cluster analysis is a difficult problem due to a variety of ways of measuring the similarity and dissimilarity concepts, which do not have a universal definition. Data clustering, also called cluster analysis, segmentation, taxonomy investigation, or unconfirmed classification, is a technique of constructing groups of objects, or clusters, in such a way that objects in different clusters are quite distinct and objects in one cluster are very similar. Data clustering is often difficult in classification, in which objects are allotted to predefined classes. In data clustering, the classes are also to be different.

Both GA and PSO, however, have their own set of strengths and weaknesses. The PSO algorithm is conceptually simple and implemented in a few lines of code. PSOs also have memory, whereas in a GA if an individual is not selected the information contained by that individual is lost. However, without a selection operator PSOs may waste resources on a poor individual that is stuck in a poor region of the search space. A PSO's group interactions enhance the look for an optimal solution, whereas GAs have problem identifying a current solution and are good at reaching a global region.

Here they propose a hybrid algorithm (GA/PSO) combining the strengths of GAs with PSO to evolve the performance. The hybrid algorithm joins the standard velocity and location update rules of PSOs with the ideas of selection and crossover from GAs. The algorithm is planned so that the GA done a global search and the PSO operates a local search. Other hybrid GA/PSO algorithms have been proposed, and have been tested on function minimization problems.

The remainder of this paper is organized as follows. Section II summarizes the concepts and related works. Section III details the proposed method, and Section IV discusses the experiments and the achieved results. Finally, Section V presents the conclusions of the work.

II. LITERATURE SURVEY

Zhang *et al* (2014) propose a novel fuzzy hybrid quantum artificial immune clustering algorithm based on cloud model (C-FHQAI) to solve the stochastic problem. Abdel-Kader and Rehab (2010) offers a hybrid two-phase GAI-PSO+ k-means data clustering algorithm that execute fast data clustering and can omitted early convergence to local optima. Mahmood and Amjad (2010) believe the trouble of placing copies of objects in a shared web server scheme to diminish the cost of allocation read and write requirements when the web servers have restricted storage capacities.

Particle Swarm Optimization is a population based globalized look for algorithm that mimics the ability (cognitive and social performance) of swarms. PSO manufacture improved results in complex and multi-peak troubles. An effort is completed to offer a direct for the researchers who are effective in the area of PSO and data grouping. PSO variants are also explain in this work are offered by Rana *et al* (2011). Frank Leung *et al* shos the performance of GA/PSO.

Akhshabi et al (2014), propose a particle swarm optimization (PSO) based on Memetic Algorithm (MA) that hybridizes with a local look for technique for work out a no-wait flow scheduling difficulty. The major objective is to diminish the whole flow time. Within the structure of the planned algorithm, a local edition of PSO with a ring-shape topology arrangement is utilized as global search. Li *et al* (2014), In the current study, particle swarm optimization (PSO) was invoked to meliorate PLS-DA via simultaneously selecting the optimal variable subset as well as the associated weights and the best number of latent variables in PLS-DA, forming a new algorithm named PSO-PLSDA. Chuang *et al* (2012), suggest fresh particle swarm optimization (CPSO) algorithms that discover the best SNP arrangement for cancer connection studies containing seven SNPs.

Muthukaruppan *et al* (2012), gives a particle swarm optimization (PSO)-based fuzzy expert scheme for the analysis of coronary artery disease (CAD). The planned scheme is based on the Cleveland and Hungarian Heart Disease datasets. Because the datasets contains numerous input attributes, decision tree (DT) was utilized to untangle the attributes that donate towards the diagnosis. Marinakis *et al* (2013), this introduce a fresh algorithmic environment inspired techniques that uses a hybridized Particle Swarm Optimization algorithm with a fresh neighborhood topology for effectively solving the Feature Selection Problem (FSP). Goldberg (1989) shows Genetic Algorithms-in Search, Optimization and Machine Learning. Improved GA and PSO Culled Hybrid Algorithm are given by Li *et al* (2008) and Shi *et al* (2005).

Mohamad *et al* (2013), they convey that from the gene expression data, choosing a small subset of instructive genes do classification. However, lots of the computational techniques faces difficulty in choosing small subset since the small amount of samples needs to be evaluated to the vast amount of genes (high-dimension), irrelevant genes and noisy genes.

III. PROPOSED DATA CLUSTERING FOR HYBRID GA/PSO ALGORITHM

The proposed method is more successful in research. The approaches are discussed in this research learning's are

3.1K-Mean

The K-means algorithm clusters D-dimensional data vectors into a predefined number of clusters on the basis of the Euclidean distance as the comparison criteria. Euclidean distances among data vectors are smallest amount for data vectors within a group as compared with distances to other vectors in various clusters. Vectors of the similar cluster are connected with one centroids vector, which shows the middle of that group and is the mean of the data vectors that belong jointly.

3.2 Modified firefly algorithm (MFA)

In the standard Firefly Algorithm (FA), in every iteration the brighter firefly applies it's manipulated over additional fireflies and attracts them towards itself in maximization problems. In fact, in the standard FA, fireflies move despite of the global optima and it reduce the ability of the firefly algorithm to discover global best. In this work, to remove weaknesses of FA and get better the collective movement of fireflies, propose a Modify Firefly Algorithm (MFA). In the proposed algorithm, use global optima in firefly's progress. Global optimum is associated to optimization difficulty and it is a firefly that has the greatest or smallest amount value. And the global optima will be modernized in any iteration of algorithm. In the proposed algorithm, when a firefly compare with another firefly as an alternative of the one firefly being permitted to authority and to attract its neighbors, global optima in every iteration is permitted to influence others and change in their association. In the MFA, when a firefly perform with correspond firefly, if the match firefly be brighter, the compared firefly will shift toward correspond firefly, measured by global optima.

3.3 Proposed Hybrid K-Mean with GA/PSO

A k-means clustering is the most popular and easy method for data clustering. In the k-means algorithm, initially k random cluster middle distinct and then every data vector will be assigning to every cluster based on Euclidean distance. Because of this the algorithm may trapped in local optima in the k-means clustering K center of cluster initial randomly. In this work to develop and enhance the accuracy of k-means algorithm, initialize the k-means algorithm with best centers, which evaluated by Hybrid GA/PSO algorithm.

$$\left[Z_{1,1}, Z_{1,2}, \dots, Z_{1,d}, Z_{2,1}, Z_{2,2}, \dots, Z_{2,d}, \dots, Z_{K,1}, Z_{K,2}, \dots, Z_{K,d}\right]$$

Figure1: Structure of a firefly position in clustering.

```
Initialize GA/PSO with random K*D centers
    While (t<max generation)
        For i=1: n (all n GA/PSO)
        For j=1: n (all n GA/PSO)
Calculate objective function of each GA/PSO by equation 4,
    If (ij>ii)
Move GA/PSO I toward j based on equation 7 to refine position of
GA/PSO (clusters center)
    End if
        End for j
        End for i,
Ranks the GA/PSO and find the current best to update current best to
next iteration
    End while
Rank the GA/PSO and find global best and extract the position of
global best
    Repeat
Initialize the k-means center with position of global best
Allocate each vector to a cluster by equation 4,
Refined the clusters by equation 5
    Do until predefined iteration.
```


In the planned technique, the clustering has two stages. First stage is initializing Hybrid K-Mean with GA/PSO with random values. Figure 1 show, data is D-dimensional and there are K clusters, so every GA/PSO has K×D dimension. The mechanism of GA/PSO algorithm must do till predefine iteration. The k-means will start with the location of best GA/PSO in the second stage. The k-means clustering process the centers. The proposed hybrid clustering algorithm is summarizing as the pseudo code show in following pseudo code.

IV. EXPRERIMENTAL RESULTS

In this section, the experimental results on several synthetic and real processes are discussed. There are a total of two algorithms used in this section, i.e., standard Firefly and Hybrid GA/PSO. The proposed method of Hybrid GA/PSO is used for data clustering. Experiments were behavior on two datasets from the UCI repository: *Iris* and *Wine*. This result shows the comparison of Firefly algorithm, GA/PSO and proposed Hybrid GA/PSO.

Table 1: Comparison of the Inter Cluster Distance values

Data points	KFA algorithm	Proposed Hybrid GA/PSO
Iris Dataset	0.05261	0.06473
Wine Dataset	0.03837	0.07087

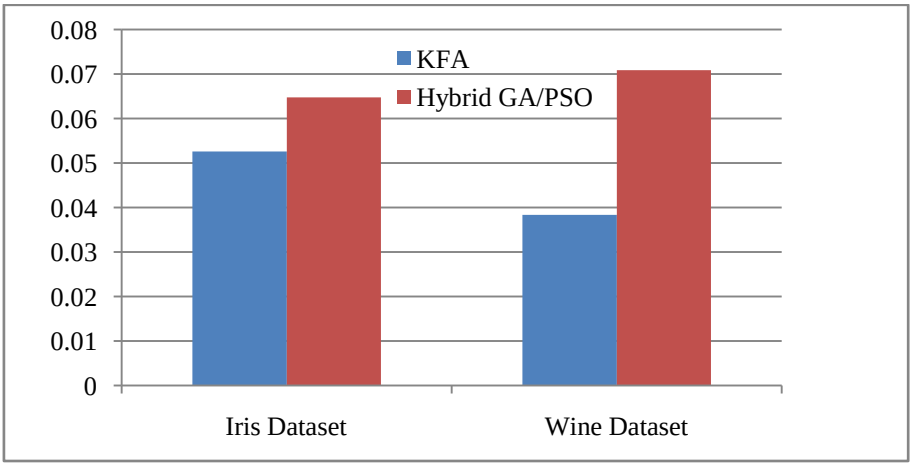

Figure 1: The Inter-Clustering distance

Table1 and Figure 1 reveal the comparison of the inter-cluster distance of the proposed Hybrid GA/PSO with KFA. The inter cluster distance values are larger in the proposed method.

Table 2: No on Clustering and Execution Time for proposed method

Methods	Iris Datasets		Wine Dataset	
	No. of Clustering	Execution Time (in secs.)	No. of Clustering	Execution Time(in secs.)
PSO	0.0293	58	0.0473	45
KFA	0.0448	42	0.0628	31
Hybrid K-Mean with GA/PSO	0.0766	18	0.0836	10

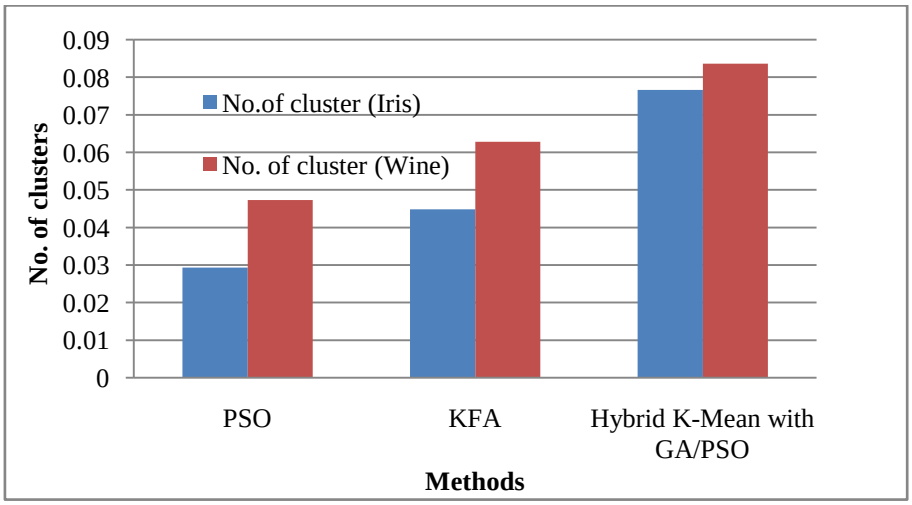

Figure 2: No. of Clustering for the approaches

Table 2 and figure 2 shows the Number of Clustering for PSO, FFA and Hybrid K-Mean with GA/PSO with Iris and Wine dataset. The proposed method of Hybrid GA/PSO have high clustering.

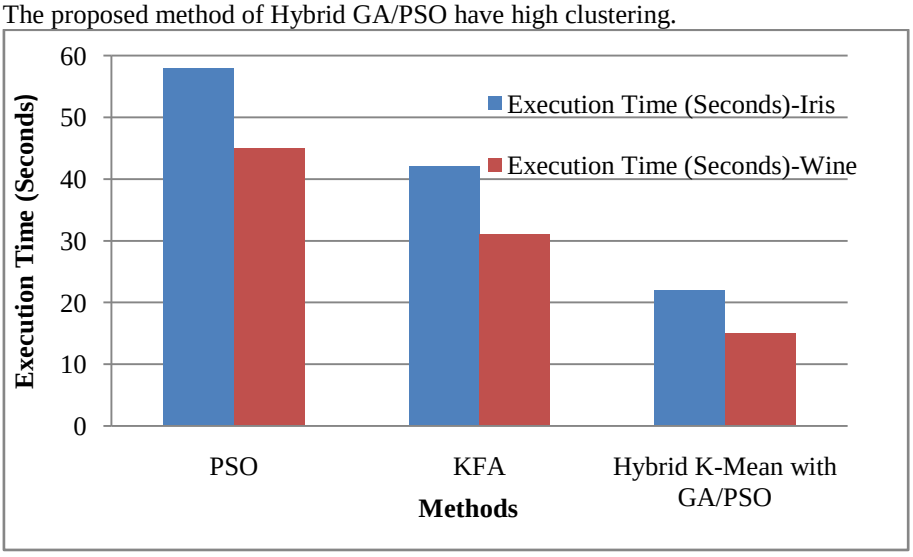

Figure 3: Execution Time

Table 2 and figure 3 shows the Execution Time for PSO, FFA and Hybrid GA/PSO with Iris and Wine dataset. The proposed methods of Hybrid K-Mean with GA/PSO have less execution time when compare with other approaches.

V. CONCLUSION

Genetic algorithm and particle swarm optimization are greatly related to their inherent parallel characteristics, both algorithms perform the function with a group of randomly created population, both have a fitness rate to calculate the population. PSO methodology is observed for document clustering limitation. It is found that the document clustering problem is successfully tackled with PSO methodology by optimizing for clustering process. A most useful advantage of the PSO is its capacity to cope with local optima by maintain, recombining and evaluation numerous candidate solutions concurrently. The Hybrid K-Mean with GA/PSO algorithm merges the capability of fast convergence of the PSO algorithm with the competence of ease to exploit preceding solution of GA for eliminating the early convergence. So PSO and GA are used in this research to expand efficient, robust and flexible algorithms to resolve a document clustering problems.

REFERENCES

[1] Abdel-Kader, Rehab F, 2010, "Genetically improved PSO algorithm for efficient data clustering", In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, pp. 71-75.

[2] Zhang, Ren-Long, Mi-Yuan Shan, Xiao-Hong Liu, and Li-Hong Zhang, 2014, "A novel fuzzy hybrid quantum artificial immune clustering algorithm based on cloud model", *Engineering Applications of Artificial Intelligence* 35: 1-13.

[3] Akhshabi, M., Tavakkoli-Moghaddam, R., & Rahnamay-Roodposhti, F. 2014, "A hybrid particle swarm optimization algorithm for a no-wait flow shop scheduling problem with the total flow time", *The International Journal of Advanced Manufacturing Technology*, *70*(5-8), 1181-1188.

[4] Li, Y. Q., Liu, Y. F., Song, D. D., Zhou, Y. P., Wang, L., Xu, S., & Cui, Y. F, 2014, "Particle swarm optimization-based protocol for partial least-squares discriminant analysis: Application to< sup> 1</sup> H nuclear magnetic resonance analysis of lung cancer metabonomics", *Chemometrics and Intelligent Laboratory Systems*, *135*, 192-200.

[5] Chuang, L. Y., Chang, H. W., Lin, M. C., & Yang, C. H, 2012, "Chaotic particle swarm optimization for detecting SNP–SNP interactions for CXCL12-related genes in breast cancer prevention", *European Journal of Cancer Prevention*,*21*(4), 336-342.

[6] Muthukaruppan, S., & Er, M. J, 2012, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease", *Expert Systems with Applications*, *39*(14), 11657-11665.

[7] Mohamad, M. S., Omatu, S., Deris, S., & Yoshioka, M, 2013, "A Constraint and Rule in an Enhancement of Binary Particle Swarm Optimization to Select Informative Genes for Cancer Classification", In *Trends and Applications in Knowledge Discovery and Data Mining* (pp. 168-178). Springer Berlin Heidelberg.

[8] Marinakis, Y., & Marinaki, M, 2013, "A hybridized particle swarm optimization with expanding neighborhood topology for the feature selection problem", In *Hybrid Metaheuristics* (pp. 37-51). Springer Berlin Heidelberg.

[9] Mahmood, Amjad, 2010, "Replicating web contents using a hybrid particle swarm optimization", *Information processing & management* 46, no. 2: 170-179.

[10] Rana, Sandeep, Sanjay Jasola, and Rajesh Kumar, 2011, "A review on particle swarm optimization algorithms and their applications to data clustering", *Artificial Intelligence Review* 35, no. 3: 211-222.

[11] Frank H.F.Leung, 2003, "Tuning of the Structure and Parameters of a Neural Network Using an Improved Genetic Algorithm", In IEEE Transaction on Neural Networks, Vol 14,No1, page 79-88

[12] Goldberg D.E, 1989, "Genetic Algorithms-in Search, Optimization and Machine Learning", Addison- Wesley Publishing Company Inc., London.

[13] Akhshabi, M., Tavakkoli-Moghaddam, R., & Rahnamay-Roodposhti, F, 2014, "A hybrid particle swarm optimization algorithm for a no-wait flow shop scheduling problem with the total flow time", *The International Journal of Advanced Manufacturing Technology*, *70*(5-8), 1181-1188

[14] Li W. T., Shi X. W., Xu L. and Hei Y.Q 2008, "Improved GA and PSO Culled Hybrid Algorithm for Antenna Array Pattern Synthesis", Progress In Electromagnetics Research, PIER 80, pp. 461–476.

[15] Shi, X.H., Liang Y.C., Lee H.P., Lu C. and Wang L.M., "An Improved GA and a Novel PSO-GA-Based Hybrid Algorithm", *Information Processing Letters*, Vol. 93, No. 5, (2005), pp. 255-261.

