Hybrid Particle Swarm Optimization (HPSO) for Data Clustering

Sandeep U. Mane
Assistant Professor, Dept. of CSE,
RIT, Rajaramnagar
Dist. Sangli, MS, India

Pankaj G. Gaikwad
M. Tech Scholar, Dept. of CSE
RIT, Rajaramnagar
Dist. Sangli, MS, India

ABSTRACT

Data mining is the collection of different techniques. Clustering information into various cluster is one of the data mining technique. It is a method, in which each cluster must contain more similar data and have much dissimilarity between inter cluster data. Most of traditional clustering algorithms have disadvantages like initial centroid selection, local optima, low convergence rate etc. Clustering with swarm based algorithms is emerging as an alternative to more conventional clustering techniques. In this paper, a new hybrid sequential clustering approach is proposed, which uses PSO - a swarm based technique in sequence with Fuzzy k - means algorithm in data clustering. Experimentation was performed on standard dataset available online. From the result, the proposed approach helps to overcome limitations of both algorithms, improves quality of formed cluster and avoids being trapped in local optima.

General Terms:

Computer Science, Data Mining, Algorithms, Swarm Intelligence

Keywords:

Data Clustering, Particle Swarm Optimization, Fuzzy k-means, Hybrid Particle Swarm Optimization

1. INTRODUCTION

Technically, data mining is the procedure of finding correlation and patterns among fields in relational databases by using traditional and non-traditional techniques. Some of the data mining strategies are summarization, association, clustering etc. Among these, data clustering is the most popular and widely used method. Data clustering algorithms have been used in data mining with many applications arising from a wide range of problems, including bioinformatics, image segmentation, security, medical image analysis, web data handling, etc [16].

The clustering algorithms belong to two groups: hierarchical clustering and partitioned clustering. In hierarchical clustering, an objective function is used locally as the merging or splitting criterion. Generally, hierarchical algorithms failed to provide optimal partitions along with selected criterion. Contradictory, partitioned methods assume the given number of clusters to be found and then look for the optimal partitions based on the objective function[13]. The use of classical optimization methods suffers from the problem of sticking to local minima; also the initialization of classical methods is another important issue. However, if good initial clustering centroids can be obtained using any of the other techniques, the partition based methods like fuzzy k-Means would work well in refining the cluster centroids to find the optimal clustering centers. To achieve better performance in data clustering applications, researchers are working on different traditional clustering techniques as well as on nature inspired or swarm based techniques.

Evolutionary and bio-inspired algorithms are introduced to overcome limitations of classical techniques and are quickly replacing the traditional techniques for practical solutions. Particle swarm optimization (PSO) is one of the nature-inspired stochastic optimization techniques. It is a Swarm Intelligence (SI) technique based on the observations of the collective behavior in decentralized and self-organized systems [7]. The PSO algorithm is used to generate good initial cluster centroid for the K-Means algorithm. In this paper, a hybrid sequential clustering approach is presented, it helps to avoid trapping algorithm in a local optimal solution.

The remaining part of the paper is organized as follows: Section II gives an overview of literature on Particle Swarm Optimization for data clustering. Section III briefly introduces basics of PSO and its limitations. Hybridization of PSO with Fuzzy k-Means is introduced in section IV. In section V implementation details and results have been represented. Finally, section VI concludes proposed work and gives future direction.

2. RELATED WORK

Different researchers have introduced and successfully applied so many traditional algorithms to solve clustering problems in different domains. Such divers clustering algorithms are due to the diversity in the induction principle and clustering models [3]. Swarm intelligence attracted many researchers in the field of engineering and management to have better performance over classical problem solving techniques used in pattern recognition and clustering. In this section, few authors work associated with Particle Swarm Optimization (PSO) for data clustering from the literature is presented.

Merwe and Engelbrecht [11] have proposed data clustering using PSO; the authors mainly focused on the objectives like use of arbitrary data as an input, minimizing intra-cluster distance and max-

imizing inter-cluster distance. The results of the algorithm were compared with k-means algorithm. The author's concluded that, proposed approaches have better conversion and low quantization error in comparison to k-means algorithm.

Esmin et al.[9] also proposed two new methods (gbest PSO cluster algorithm and the evaluation function) with PSO for data clustering. The gbest PSO clustering algorithm and evaluation function, these two approaches are used to do some modifications in fitness function of the algorithm proposed by Merwe and Engelbrecht for better results of clustering.

Kao et al. [5] stated that PSO gives better clustering results, when it is applied in one dimensional small dataset, but it fails to give good results for large dataset. The reason behind this is, when the search for a good solution reaches to the search space boundary; the particles tend to stay there and do not move in other direction for a good solution. To overcome these problem, two reflex schemes namely "PSO + parameter pulling reflex scheme" is to pull back particles which are out of the search space and "PSO + global pulling reflex scheme" which decides the reflecting range by comparing the particle location with the global best particle location which are used in PSO. This algorithm gives better results than other algorithms such as PSO, KGA, SAKM etc.

In [14],[15] researchers have proposed a hybrid sequential approach using k-means and particle swarm optimization algorithm. The author's objective was analyzing limitations of k-means clustering algorithm and to propose a new hybrid approach to address these limitations. The motivation for this idea of hybridization is from PSO algorithm, at the beginning stage of the algorithm, clustering process is started due to its fast convergence speed and then the result of PSO algorithm is tuned by the k-means to near optimal solution. Squared Euclidian distance measure was used to find distance between objects and centroids. The work carried out on Iris and Wine datasets.

Ahmady and Modares [1] proposed a combination of PSO and k-means algorithm, known as PSO-KM algorithm. It takes the advantages of both algorithms. Initially global solution is found by PSO, after that k-means algorithm is used for faster convergence. As long as the particles in the swarm being close to the global optimum, the algorithm switches to k-means.

The limitations of PSO discussed in [5] can be overcome using another approach presented in [17]. Here authors have proposed a hybrid technique based on combining the k-means algorithm, Nelder-Mead (NM) simplex search, and particle swarm optimization (PSO) called K-NM-PSO. The new K-NM-PSO algorithm is tested on nine datasets, and its performance is compared with those of PSO, NM-PSO, K-PSO and K-means clustering. Results show that K-NMPSO is both robust and suitable for handling large datasets.

Taher Niknam et al. [12] proposed a hybrid method based on fuzzy adaptive PSO and ACO called FAPSO-ACO. Authors used fuzzy rules to change adaptively the inertia weight and the learning factor, and attain the best position of the particles to calculate the transfer probability of ant colony. Then, the method takes the result of FAPSO-ACO as the initial value of k-means to find a better evaluation value. Experiments shows that the FAPSO-ACO-K algorithm attains better clustering evaluation values on UCI datasets.

Khoshdel et al. [8], proposed new hybrid learning-based algorithm for data clustering. In this paper author used learning automata technique. Khoshdel et al. described it as, "A learning automaton (LA) is a machine that can do finite actions. Every selected action is evaluated by a probability environment and evaluation result is given as a positive or negative signal to LA and LA utilizes this response in choosing next action and thereafter approaches toward selecting an action which gets the most reward from environment". In [8], the advantages of k-means for increasing convergence rate and PSO efficiency are used. Khoshdel et al. examined k-means, PSO, KAFSA, KPSO and their proposed work for iris, glass, wine, sonar, Pima and WDBC datasets.

Lifen and Changming [10] proposed integrated SOM/PSO clustering approach. During clustering process authors have firstly selected the important features using binary particle swarm optimization (BPSO) and mutual information (MI), due to which dimension of dataset get reduced. SOM is used to cluster the dataset and PSO is used to improve the clustering result. The integrated SOM/PSO also used Euclidean distance as a major of similarity.

3. BASICS OF PSO ALGORITHM

Particle Swarm Optimization (PSO) was firstly developed by Eberhart and Kennedy in 1995 which is a population based stochastic optimization method. It is motivated by social behavior of organisms such as bird flocking and fish schooling. The particle changes its position with time. All particles fly through multidimensional search space where each particle is adjusting its position according to its own experience and that of neighbors. During flight, particles velocity is stochastically accelerated toward its previous best position and towards a neighborhood best solution. Following Equation (1) and Equation (2) are velocity and position updating equations respectively, taken from [7]:

$$V_i(t+1) = W * V_i(t) + c_1 r_1 (pbest_i(t) - X_i(t)) + c_2 r_2 (gbest_i(t) - X_i(t)) \quad (1)$$

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

Where t is iteration count,$V_i(t)$ is velocity of particle i at time t, $X_i(t)$ is position of particle i at time t, W is Inertia weight which plays an important role in balancing local and global search to avoid stagnation of particles at local optima, $pbest_i$(t) is the best position found by particle itself so far, $gbest_i$(t) is the best position found by whole swarm so far, random values r_1, r_2 in the range of (0,1) are used to make sure that particles explore wide search space before converging around optimal solution and c_1, c_2 are positive acceleration constant and control the weight balance of $pbest_i$(t) and $gbest_i$(t).

Equation (1) is used for recording its current position X_i, and velocity V_i indicates speed along dimensions in a problem space. The best fitness values are updated at each generation, according to Equation (3),

$$P_i(t) \begin{cases} P_i(t) & f(X_i(t+1)) \leqslant f(X_i(t)) \\ X_i(t+1) & f(X_i(t+1)) > f(X_i(t)) \end{cases} \quad (3)$$

Where, f indicates the fitness function, P_i(t) indicates best fitness values and the position where value was calculated, and t indicates iteration count.

In literature, it seen as clustering problem is an optimization problem that locates optimal centroids of centers instead of finding optimal partitions. This gives us an opportunity to apply particle swarm

optimization (PSO) algorithm for clustering problems. PSO clustering algorithm performs a globalized search in the whole solution space. The Equation (4) - (7) represent the whole process, formulated from [2],[6].

For the data clustering process, single particle represents the Nc cluster centroid vectors. So, each particle xi looks as follows;

$$X_i = \{m_{i1}, m_{ij}, ...m_{iN_c}\} \tag{4}$$

where m_{ij} indicates j^{th} cluster centroid vector of i^{th} particle in cluster C_{ij}. Fitness value of particle is measured by using quantization error expression as follows:

$$J_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{\forall z_p \in C_{ij}} d\left(z_p, m_j\right) / \left|C_{ij}\right|\right]}{N_c} \tag{5}$$

Where, $|C_{ij}|$ indicates number of data vectors that belongs to cluster C_{ij}, d indicates Euclidean distance between each data vector to the centroid. This Euclidean distance can calculate by following expression:

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2} \tag{6}$$

Where, k indicates the dimension, N_d indicates number of parameter of each data vector, N_0 indicates number of cluster centroid to be formed, z_p indicates p^{th} data vector and m_j indicates centroid vector of cluster j. The cluster centroid vectors are recalculated by using following expression:

$$m_j = \frac{1}{n_j} \sum_{\forall z_p \in C_j} z_p \tag{7}$$

Where,n_j indicates number of data vectors in cluster j and C_j indicates subset of data vectors from cluster j.

The working of PSO algorithm for data clustering is as follows:

Algorithm

(1) Initialize a population of particles with random positions and velocities in the search space.
(2) While (termination conditions are not met)
{
for each particle i do
{
Calculate Euclidean distance d to all cluster centroids C_{ij}
Make assignment of data vector to cluster Cij such that minimum distance between data vectors within a cluster
Calculate fitness value according to equation (5)
}
Update pbesti(t) and gbesti(t) positions
Update cluster centroids according to equation (1) and (2)
}

Starting with random population, each particle moves in search space and keeps the best position it has seen. The PSO algorithm stops either when maximum number of function evaluations has been reached or when there is no significant improvement over a number of iterations. Among different swarm based techniques particle swarm optimization has advantage of faster convergence at initial stages of search process. But at near global optimum, the convergence speed becomes very slow. The formed numbers of cluster are forward to fuzzy k-means algorithm for further clustering process.

4. BASICS OF FUZZY K-MEANS ALGORITHM

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (soft clustering), data elements can belongs to more than one cluster and associated with each element is a set of membership levels. These indicate the strength of association between that data element and a particular cluster. Fuzzy clustering is the process of assigning these membership levels, and then using them to assign data elements to one or more clusters. Fuzzy cluster analysis therefore allows gradual memberships of data to clusters in (0, 1). Membership degrees can also express how ambiguously or definitely a data point should belong to a cluster. The concept of these membership degrees is clarified by interpretation of fuzzy sets [4]. For Fuzzy k-means phase, the number of cluster to be formed is output of PSO algorithm.

The fuzzy k-means clustering process is presented as follows:

Algorithm

(1) Initialize k clusters
(2) Until converged
  (a) Compute the probability of a point belong to a cluster for every $<$ point , cluster $>$ pair.
  (b) Re-compute the cluster centers using probability membership values of points (from a) to clusters.

It works into two steps as follows:

1) Form Clusters (assign membership degrees) and
2) Move Centroids (move each centroid towards its proper position).
To form clusters i.e. to assign membership function to each data vector, following mathematical expression is used.

$$\mu_k(x_i) = \frac{1}{\sum_{n=1}^{k} \left(\frac{d(m_k, x_i)}{d(m_h, x_i)}\right)^{\frac{2}{q-1}}} \forall i \in \{1...n\}, k \in \{1...k\} \tag{8}$$

And to move/update formed cluster centroids, following expression is used in the algorithm.

$$m_k = \frac{\sum_{i=1}^{n} \mu_k^q(x_i) x_i}{\sum_{i=1}^{n} \mu_k^q(x_i)} k = 1, 2, ..k \tag{9}$$

Where, $\mu_k(x_i) \in [0, 1]$ is a fuzzy membership function, 'q' is membership exponent that controls the amount of fuzziness, k-number of clusters, $d(m_k, x_k)$ is distance measure between the centre 'm_k' of clusters and the pattern $x_i \in x$.
Equation (8) and (9) are used in iterative fashion to update memberships and cluster centers. This continues till values of all patterns becomes negligible or required number of iterations is over. However, fuzzy k-means algorithms have several limitations like to provide number of clusters initially, trapped in local optima, etc. This can be minimized by using PSO at the initial stage and final results are tuned by applying fuzzy k-means in later stage.

5. PROPOSED TECHNIQUES

In this section the proposed work is discussed. The hybridization process is structured into two phases. In first phase, particle swarm optimization is implemented which makes use of whole search space for global solution. When the region of global optimum is found by PSO algorithm, we continue the clustering process using fuzzy k-means algorithm in second phase. When the value of fitness function of PSO for the number of successive iteration changes

negligibly the clustering algorithm switches to fuzzy k-means algorithm.

The PSO algorithm is hybridized with fuzzy k - means, which takes benefit of fast convergence speed of PSO and then it is tuned by fuzzy k - means to determine near optimal solution. Flow chart of proposed hybrid PSO algorithm for clustering is shown in Figure 1.

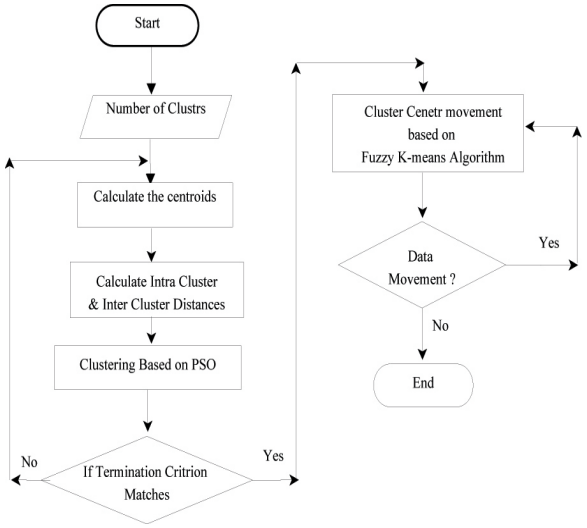

Fig. 1. Flow diagram of Hybrid PSO Clustering Algorithm

6. IMPLEMENTATION DETAILS AND RESULTS

This section presents the analysis of proposed work along with basic PSO and fuzzy k-means clustering algorithm. Also the experimental setup and dataset used for experimentation is discussed.

6.1 Problem Formulation

The data clustering is based on unsupervised learning, where data is divided into training set and it is used for generating cluster centers. The objective function is presented in equation (5) and cluster center is represented by equation (7). The objective function measures the quality of formed cluster with two criteria:

—*The objective function should minimize intra-cluster distance.*

—*The objective function should maximize inter-cluster distance.*

6.2 Experimental Setup

The parameter setting for particle swarm optimization (PSO) algorithm for data clustering problem is given as:

(1) Number of particles: 6
(2) Cognitive Parameter: 1.49
(3) Social Parameter: 1.49
(4) Max. number of iteration: 600
(5) w = 0.72

To test the proposed data clustering algorithm, standard datasets are available, among these following dataset were used:

Iris Plants: This is best known database found in pattern recognition literature. It's the database with 4 inputs, 3 classes and 150 data vectors. There are few missing values.

Glass: Its very well known database for machine learning. The study of classification types of glass was motivated by criminological investigation. It's the database with 10 attributes, 2 classes and 178 data vectors. There is also missing attributes.

Cancer: The Wisconsin breast cancer database contains 9 relevant inputs and 2 classes. The objective is to classify each data vector into benign or malignant tumors. The attributes are Class-id, Sample code number, Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses.

Wine: These datasets are the results of a chemical analysis of wines grown in the Italy but derived from three different cultivators. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

6.3 Experimental Results

The experimentation result of three algorithms on four different datasets are shown in Figure 2.

Data Set Name	Algorithms	Intra Cluster Distance	Inter Cluster Distance	Time Required in ms
Iris	Fuzzy k-means	0.3951	0.4872	133.5
	PSO	0.3884	0.5086	1092
	Proposed Hybrid PSO	0.3988	0.5454	1269.33
Glass	Fuzzy k-means	0.3100	0.4684	95
	PSO	0.3540	0.5184	1776.4
	Proposed Hybrid PSO	0.3576	0.5434	1914.6
Cancer	Fuzzy k-means	0.3191	0.4552	93.75
	PSO	0.3139	0.4497	1587.4
	Proposed Hybrid PSO	0.2999	0.4556	1546
Wine	Fuzzy k-means	0.4363	0.5683	167.5
	PSO	0.3417	0.3910	1164
	Proposed Hybrid PSO	0.3249	0.4421	1093.5

Fig. 2. Result of fuzzy k-means, PSO and Hybrid PSO Clustering Algorithms

For all mentioned results, average 30 simulations were performed. PSO and hybrid PSO algorithm run for 600 function evaluations. For PSO and hybrid PSO, values of c1, c2 = 1.49 and w = 0.72 are selected to ensure good convergence.

The intra-cluster distance ensures the compact cluster with small deviation within cluster while inter-cluster distance ensures larger

separation between other clusters. With reference to those criteria, the hybrid PSO approach succeeded most in finding clusters with larger separation than other two.

7. CONCLUSION

In this paper, hybridization of the particle swarm optimization (PSO) and the fuzzy k-means algorithm is proposed. The hybrid method has the advantage of both PSO and fuzzy k-means methods while it does not inherent their drawbacks. In hybridization process, PSO algorithm is used at initial stage because it successfully searches whole solution space during the initial stages of global search. As long as particles in swarm being near to global optimum, the algorithm switches to fuzzy k-means for further improving the quality of formed clusters. On the basis of result and experiments performed on four datasets namely iris, glass, cancer and wine data, author concludes that hybrid algorithm outperforms than fuzzy k-means and PSO clustering mainly in maximizing inter-cluster distances. In future, proposed hybrid PSO algorithm can be parallelized because from result it is observed that for data clustering the time required is more.

8. REFERENCES

[1] Ahmadyfard, Alireza, and Hamidreza Modares. Combining pso and k-means to enhance data clustering. In *IEEE International symposium on telecommunications*, pages 688–692, 2008.

[2] A. Carlisle and G. Dozier. An off-the shelf pso. In *Workshop on Particle Swarm Optimization, Indianapolis, IN*, pages 1–6, 2001.

[3] Vladimir Estivill Castro. Why so many clustering algorithms a position paper. *SIGKDD Exploration*, 4(1):66–75, January 2002.

[4] Doring Christian, Marie-Jeanne Lesot, and Rudolf Kruse. Data analysis with fuzzy clustering methods. *Computational Statistics Data Analysis*, 51(1):192–214, 2006.

[5] Kao IW Tsai CY and Wang YC. An effective particle swarm optimization method for data clustering. In *IEEE International Conference Industrial Engineering and Engineering Management*, pages 548–552, 2–4 December 2007.

[6] Shi Y. H. and Eberhart R. C. *Parameter Selection in Particle Swarm Optimization*, volume VII of *Evolutionary Programming*. Springer Berlin Heidelberg, 1998.

[7] Kennedy J. and Eberhart R. C. Particle swarm optimization. In *IEEE International Conference on Neural Networks, Perth Australia*, volume 4, pages 1942–1948, 1995.

[8] Hamed Khoshdel and Barat Saman. A new hybrid learning-based algorithm for data clustering. In *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, 2012.

[9] Esmin A Pereira D L and de Araujo F. Study of different approach to clustering data by using particle swarm optimization algorithm. In *IEEE World Congress on Computational Intelligence Evolutionary Computation, CEC 2008*, volume 1, pages 1817–1822, 1–6 June 2008.

[10] Li Lifen and Zhang Changming. Alert clustering using integrated som/pso. In *International Conference on Computer Design and Applications (ICCDA)*, volume 2, pages 517–574, 2010.

[11] Van Der Merwe and A.P Engelbrecht. Data clustering using particle swarm optimization. In *The 2003 Congress on Evolutionary Computation, 2003. CEC 03*, volume 1, pages 215–220, 8–12 December 2003.

[12] Taher Niknam and Babak Amiri. An efficient hybrid approach based on pso, aco and k-means for cluster analysis. *Applied Soft Computing*, 10.1:183–197, 2010.

[13] Jain A. R., Murthy M. N., and Flynn P. J. Data clustering: A review. *ACM Computing Surveys*, 31(3):265 –323, 1999.

[14] Sandeep Rana, Sanjay Jasola, and Rajesh Kumar. A hybrid sequential approach for data clustering using k-means and particle swarm optimization algorithm. *International Journal of Engineering, Science and Technology, MultiCraft Limited*, 2(6):167–176, 2010.

[15] Vinod Sharma, Nitish Salwan, Sandeep Singh, Navneet Singh Babra, and Prabhsimran Singh. A review of data clustering techniques and enhancement of data clustering using hybrid clustering model of k-means and pso clustering. *International Journal on Advanced Computer Theory and Engineering (IJACTE),*, 2(2):2319 – 2526, 2013.

[16] He Y. Pan W. and Lin J. Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computer Stat Data Anal 51*, pages 641–658, 2006.

[17] Kao Y-T, Zahara E, and Kao I-W. A hybridized approach to data clustering. *Expert Syst with Appl*, 34(3):1754–1762, 2008.