MultiCraft

International Journal of Engineering, Science and Technology
Vol. 2, No. 6, 2010, pp. 167-176

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest-ng.com

A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm

Sandeep Rana[1*], Sanjay Jasola[1], Rajesh Kumar[2]

[1]School of ICT, Gautam Buddha University, Greater Noida, INDIA*
[2] Department of Electrical and Computer Engineering, National University of Singapore, SINGAPORE
[*]Corresponding Author: e-mail: srana.it@gmail.com, Tel +91-120-2344246, Fax +91-120-2344300

Abstract

Clustering is a widely used technique of finding interesting patterns residing in the dataset that are not obviously known. The K-Means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient in terms of the execution time. However, due to its sensitiveness to initial partition it can only generate a local optimal solution. Particle Swarm Optimization (PSO) technique offers a globalized search methodology but suffers from slow convergence near optimal solution. In this paper, we present a new Hybrid Sequential clustering approach, which uses PSO in sequence with K-Means algorithm for data clustering. The proposed approach overcomes drawbacks of both algorithms, improves clustering and avoids being trapped in a local optimal solution. Experiments on four kinds of data sets have been conducted. The obtained results are compared with K-Means, PSO, Hybrid, K-Means+Genetic Algorithm and it has been found that the proposed algorithm generates more accurate, robust and better clustering results.

Keywords: Cluster Centroid, Global Optimization, K-Means clustering, Particle Swarm Optimization (PSO)

1. Introduction

In general, clustering involves partitioning of a given multidimensional data vector set into subsets based on the closeness or similarity among the data of same kind (Mitra and Acharya, 2004). Clustering algorithms have been used in data mining and machine learning with many applications arising from a wide range of problems, including exploratory data analysis, image segmentation, security, medical image analysis (Zhang and Chen, 2004), web handling and mathematical programming (Pyle, 1999), (Panov et al., 2008). Owing to the huge amount of data collected in databases, cluster analysis has recently become a highly active area of research. Clustering has been defined as the process of grouping a data set in a way that the similarity between data within a cluster is maximized while the similarity between data of different clusters is minimized (Rajan and Saravanan, 2008), (Xindong, 2004), so the clustering algorithms have to focus on the in-house grouping based on certain criteria. The research in this area has focused on finding an efficient, fast and effective cluster analysis algorithm to handle large databases.

Most clustering algorithms belong to two groups: hierarchical clustering and partitioned clustering. The hierarchical approach produces a nested series of partitions consisting of clusters either disjoint or included one into the other. In hierarchical clustering, an objective function is used locally as the merging or splitting criterion. In general, hierarchical algorithms cannot provide optimal partitions for their criterion. In contrast, partitioned methods assume the given number of clusters to be found and then look for the optimal partitions based on the object function (Jain et al., 1999). However, in many applications, hierarchical approaches are unpractical for clustering. In such circumstances, the partitioned clustering approach which directly minimizes the sum of squares distance is more applaudable. The traditional way to deal with such problems is to use some heuristics such as the well-known K-Means algorithm (Zalik, 2008).

The K-Means algorithm is one of the most popular methods for clustering multivariate quantitative data (Tsai and Chiu, 2008). It is a method commonly used to automatically partition a data set into k groups. K-Means algorithm generates a fast and efficient solution. The basic K-Means algorithm works with the objective to minimize the mean squared distance from each data point to its nearest centre. There are no efficient solutions known to any of these problems and some formulations are NP-hard. 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. These two drawbacks are also present in the K-Means algorithm and hence the cluster result is sensitive to the selection of the initial cluster centroids and converges to the local optima. Therefore, the initial selection of the cluster centroids decides the main processing of K-Means algorithm and the partition result of the dataset as well. The K-Means algorithm searches the local optimal solution in the vicinity of the initial solution to refine the partition result. An approximation algorithm for solving clustering problem with arbitrary dimensions was proposed (Kumar et al., 2010). A filtering algorithm based on kd-tree increased the speed of clustering process (Kanungo et al., 2002). Local approximation based heuristic was used for K-Means clustering and proved it through an empirical study (Kanungo et al., 2002). However, if good initial clustering centroids can be obtained using any of the other techniques, the K-Means would work well in refining the clustering centroids to find the optimal clustering centers. The same idea is proposed in this paper to determine initial points for K-Means algorithm by some other global optimization search algorithms.

Evolutionary and bio-inspired algorithms eradicate some of the above mentioned difficulties and are quickly replacing the classical methods in solving practical problems (Chen and Fun, 2004). The Particle Swarm Optimization (PSO) is one of the nature-inspired population based stochastic optimization algorithms. It is a Swarm Intelligence (SI) technique based on the observations of the collective behavior in decentralized and self-organized systems (Kennedy and Eberhart, 1995). Its examples can be found in nature, including bee colonies, ant colonies, bird flocking, animal herding, bacteria modeling and fish schooling (Kennedy et al., 2002). The particles search locally but the interaction with each other leads to the emergence of global behavior (El-abd and Kamal, 2005). The PSO algorithm can be used to generate good initial cluster centroids for the K-Means. In this paper, we present a hybrid sequential clustering approach that can avoid being trapped in a local optimal solution.

This paper is organized as follows. The K-Means algorithm is most commonly used algorithm because of its ease of implementation. Section 2 details the working of K-Means algorithm and also describes major drawbacks which are to be rectified. Section 3 details the standard PSO and the related issues about accuracy and convergence to optimal solutions. Section 4 describes the basic requirements of sequential clustering approach. The development and working of the approach is elaborated in the section 4. Section 5 discusses simulation and experimental results made on some standard test systems and draws inferences on the cluster formation from the results obtained. Finally, section 6 concludes the paper.

2. The K-Means Clustering Algorithm

Developed between 1975 and 1977 by J. A. Hartigan and M. A. Wong, K-Means clustering is one of the older predictive modeling methods (Mitra and Acharya, 2004). In K-Means clustering a set of n observations in d-dimensional space (an integer d) is given and the problem is to determine a set of c points to minimize the mean squared distance from each data point to its nearest center with which each observation belongs. No exact polynomial-time algorithms are known for this problem. The problem can be set up as an integer programming problem but because solving integer programs with a large number of variables is time consuming, clusters are often computed using a fast, heuristic method that generally produces good (but not necessarily optimal) solutions (Jain et al., 1999). The K-Means algorithm is one such method where clustering requires less efforts. In the beginning, number of cluster c is determined and the centre of these clusters is assumed. Any random objects as the initial centroids can be taken or the first k objects in sequence can also serve as the initial centroids.

Given a set of observations $(x^1, x^2, ..., x^n)$, where each observation is a *d*-dimensional real vector, then K-Means algorithm clustering aims to partition the *n* observations into c sets ($c < n$) as $Z = (z^1, z^2, ..., z^c)$ to minimize a measure of dispersion within the clusters. The standard K-Means algorithm minimizes the within-cluster sum of squares distance according to the equation (1) given below.

$$f_1 = \arg_z \min\left(\sum_{j=1}^{c} \sum_{X^i \in Z^j} \left\| X^i - \mu^j \right\|^2\right) \tag{1}$$

where μ^j is the mean of z^j.

There are two issues in creating a K-Means clustering algorithm: the optimal number of cluster and the centre of cluster. In many cases, number of cluster is given then the important part is where to put cluster centre so that scattered points can be grouped properly. Centre of cluster can be obtained by first assigning any random point and then optimizing the mean distance as given in equation (1). The process is repeated until all the centre positions are optimized.

The drawback of standard clustering algorithm is that they ignore measurement errors, or uncertainty, associated with the data. If these errors exist, then these can play a significant role in deciding clusters and cluster centers. In general, the algorithm does not achieve a global minimum of f_1 over the assignments. In fact, since the algorithm uses discrete assignment rather than a set of continuous parameters, the "minimum" it reaches cannot even be properly called a local minimum (Cui et al., 2005). Despite these drawbacks, the algorithm is used fairly frequently because of its ease of implementation (Tsai and Chiu, 2008). The result of K-Means algorithm is highly dependent upon its initial selection of cluster centers and before clustering it must be previously known and fixed.

3. The Particle Swarm Optimization Algorithm

PSO exploits a population of individuals to probe promising regions of the search space. In analogy with evolutionary computation methods, a swarm is similar to population and a particle is similar to an individual. PSO follows a stochastic optimization method based on Swarm Intelligence (VenderMerwe and Engelbrecht, 2003). The fundamental idea is that each particle represents a potential solution which it updates according to its own experience and that of neighbors. The PSO algorithm searches in parallel using a group of individuals. Individuals or particles in a swarm, approach to the optimum through their present velocity, previous experience and the experience of its neighbors (Shi and Eberhart 1998). PSO searches the problem domain by adjusting the trajectories of moving points in a multidimensional space. The motion of individual particles for the optimal solution is governed through the interactions of the position and velocity of each individual, their own previous best performance and the best performance of their neighbors.

In PSO, swarm is composed of a set of particles $P = \{ p^1, p^2 , p^3 , ... p^n \}$. The position of particle corresponds to a candidate solution of the optimization problem. At any time step k, the particle p^i has two vectors associated: position $\vec{X}_k^i$ and velocity $\vec{V}_k^i$. Both the information vectors have been recorded in every time step and help in further movement of particle. The best position that particle p^i has ever visited till time step k is known as personal best and represented by vector $\vec{pbest}_k^i$. The best position of all the particles is known as global best and represented by $\vec{gbest}_k^i$. The movement of particle in search space depends on the information it receives from neighborhood $N^i \subseteq P$. The neighborhood relations between particles are commonly represented as a graph $G = \{V, E\}$, where each vertex in V corresponds to a particle in swarm and each edge in E relates connections between them.

The basic PSO algorithm consists of three steps, namely, generation of particles and their information, movements and new information vector. This can be considered as generating particle's positions and velocities, velocity update, and finally, position update. First, the positions, $\vec{X}_k^i$, and velocities, $\vec{V}_k^i$, of the initial swarm of particles are randomly generated using upper and lower bounds on the search variables values, LB and UB, as expressed in equations (2) and (3). In equations (2) and (3), *rand* is a uniformly distributed random variable that can take any value between 0 and 1. This initialization process allows the swarm particles to be randomly distributed across the search space.

$$X_0^i = LB + rand\,(UB - LB) \tag{2}$$

$$V_0^i = \frac{LB + rand\,(UB - LB)}{\Delta t} \tag{3}$$

The movement of particle in the next time step is function of its current velocity and particle current position which is the objective function to be optimized. There are three parts in velocity update: the first part shows the current speed of particle i.e. shows its present state, the second part is known as the cognition term which shows the thought of the particle itself and the last part is social term that shows the ability of information sharing among the swarms. The initial velocity $\vec{V}_k^i$ is updated first using the information of $\vec{pbest}_k^i$ and $\vec{gbest}_k^i$ to $\vec{V}_{k+1}^i$ for next iteration. Good convergence of the search space and avoiding trapping in local minima can be ensured by using some random parameters, represented by the uniformly distributed variables, *rand*. The velocity update formula uses the current velocity, particle personal memory and swarm memory influence as given in equation (4).

$$\vec{V}_{k+1}^i = \underbrace{w\vec{V}_k^i}_{\text{Current Velocity}} + \underbrace{c_1\, rand\, \frac{(\vec{pbest}_k^i - \vec{X}_k^i)}{\Delta t}}_{\substack{\text{Particle Personal Memory} \\ \text{Consideration}}} + \underbrace{c_2 rand\, \frac{(\vec{gbest}_k^i - \vec{X}_k^i)}{\Delta t}}_{\substack{\text{Swarm Memory} \\ \text{Consideration}}} \tag{4}$$

Where c_1 and c_2 are two positive acceleration constants responsible for degree of information consideration of personal and swarm memory respectively and w is an inertia weight which is usually linearly decreasing during the iterations. The inertia weight w plays a role of balancing the local and global search. Tsai and Chiu (2008) proposed generalized models and techniques for tuning these parameters. Position update is the last step in each iteration. The Position of each particle is updated using its velocity vector given by equation (5).

$$\vec{X}^{i}_{k+1} = \vec{X}^{i}_{k} + \vec{V}^{i}_{k+1}\,\Delta t \tag{5}$$

The three steps of velocity update, position update, and fitness calculations are repeated until a desired convergence criterion is met. PSO algorithm is very fast, simple and easy to understand and implement. It also has a very few parameters to adjust (Kennedy et al., 2002) and requires little memory for computation. PSO also has major draw backs, such as when the search space is high its convergence speed becomes very slow near global optimum. Another PSO problem is its nature to a fast and premature convergence in mid optimum points.

4. Hybrid Sequential Clustering Algorithm

The issues related to global and local minimum play an important role when data sets and attributes associated are very large and the classification based on clustering is important and critical. In case of certain data sets like medical, security, finance etc. the error generated because of K- Means clustering algorithm is not acceptable. The objective function of the K-Means algorithm is not convex and hence it may contain many local minima. Bio-inspired algorithms have advantages of finding global optimal solution. The process of random searching and information sharing make these algorithms best tool for finding global solutions (Sadu et al., 2009). We have used one of such algorithm i.e. Particle Swarm Optimization (PSO) for data clustering. In this section we aim to propose a hybrid sequential clustering algorithm based on combining the K-Means algorithms and PSO algorithms. The motivation for this idea is the fact that PSO algorithm, at the beginning stage of algorithm starts the clustering process due to its fast convergence speed and then the result of PSO algorithm is tuned by the K-Means near optimal solutions. Flow chart of proposed algorithm is shown in Figure 1.

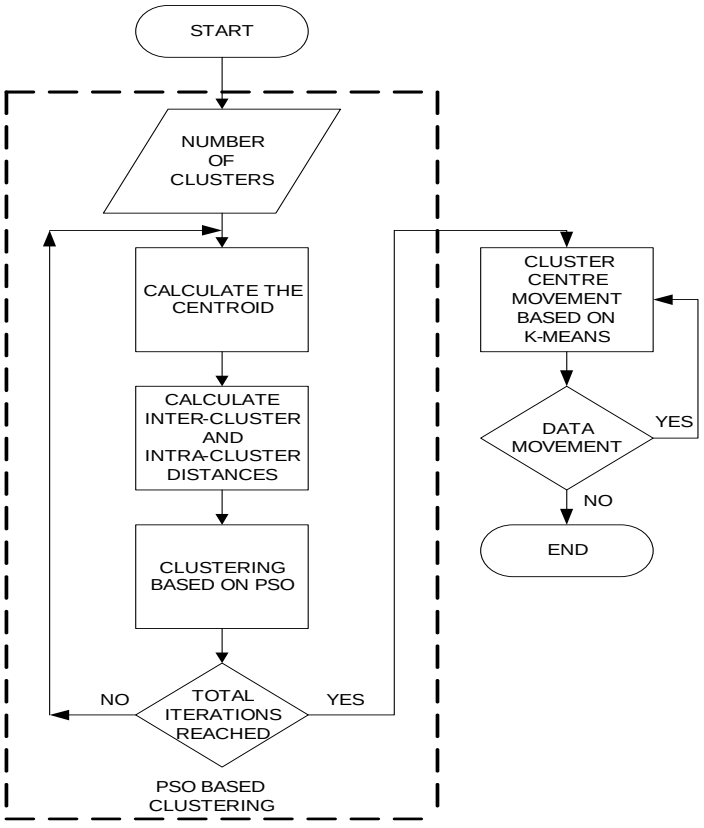

Figure.1. Flow chart of Hybrid Sequential Clustering Algorithm

The combination of K-Means algorithm and PSO will generate the better result compared to the result of individual algorithm. This algorithm will remove the drawbacks of both algorithm (K-Means Algorithm and PSO Algorithm) and uses the advantage of both algorithms for producing the best optimized result. The algorithm of the proposed scheme is given below.

Algorithm : Hybrid Sequential Clustering Algorithm

1. Initialization: Randomly generate particles where each particle represents a feasible solution i.e. cluster solution. The number of particles is taken as product of dataset dimension and number of clusters to be generated.
2. Initialization of particle position and velocity: Each candidate solution possesses a position, which represents the solution in search space and velocity for the movement of particles for finding global optimal solution. The position and velocity initialization is made by using equations (2) and (3).
3. Evaluation of fitness: The fitness value of each particle is computed by the following fitness function.

$$objective\ function(f_2) = \sum \left\| x^i - z^j \right\|,\ i = 1,....,n,\ j = 1,....,c \qquad (6)$$

   Where n and z are the number of datasets and clusters, respectively and x^i is data point and z^j is cluster centre. The value of objective function is stored as particle personal best and best of all personal best is recorded as global or swarm best.
4. Position and velocity update: The search for the global optimal solution is made through dynamically updating the particles in swarm. The velocity update will be made using equation (4) which is function of initial velocity, the particle own best performance and the swarm best performance. Position update will be made using equation (5) by adding incremental change in position in each step. Though particles have been initialized by equation (2) and (3) forcing them to search them within the boundary but in case they move out of boundary they are reset to the boundary value.
5. Steps 2-4 are repeated till the termination condition is reached.
6. Place n points into the space represented by the objects that are clustered with cluster centre obtained from PSO algorithm. These points represent initial group centroids.
7. Assign each object to the group that has the closest centroid.
8. When all objects have been assigned, recalculate the positions of the c centroids using equation (1).
9. Repeat Steps 6 and 7 until the centroids no longer move.

PSO algorithm is a probabilistic approach to find the optimal solution and hence in every run it generates a new optimal solution near around global optimal point. It is normally suggested to take 10 runs of the algorithm and find the mean value of it for further processing. Although PSO is a good clustering method, it does not perform well when the dataset is large or complex. K-Means is added in sequence to the PSO to obtain better result through further refinement in cluster formation. The PSO algorithm is used at the initial stage to help discovering the vicinity of the optimal solution by a global search. The result from PSO is used as the initial seed of the K-Means algorithm, which is applied for refining and generating the final result.

5. Result and Discussion

In this section, details of the overall results of the proposed algorithm are discussed. A complete program using MATLAB has been developed to find the optimal solution. This section has been divided into two subsections. Firstly the working of the proposed scheme and refinement in the cluster centers is illustrated. Secondly to evaluate the performance of the proposed clustering algorithm, few experiments have been conducted on two artificial generated data set problems and another two with standard data mining benchmark problems.

Subsection 1: Only six data with two attributes are selected to create a dataset, to give a graphical view of the working of proposed Hybrid Sequential clustering algorithm to frame two clusters. The data set is developed by random number generation in

a range [0, 1]. PSO is applied first to the data set and the obtained results are shown in Figure 2 for two cluster formulation. The obtained results from PSO are then processed through K-Means algorithm for further refining the cluster formation. Figure 3 shows the working of proposed Hybrid Sequential clustering algorithm. It can be seen that the cluster centers are further shifted. It can also be seen that the centers are moving more toward the centre of cluster and both centers are moving far away from each other i.e. maximization of distance between the cluster centers.

It is found that the further refinement in the cluster centers lead to more composite and condensed cluster formation also it is also observed that the cluster formation only using PSO is not sufficient. In PSO, the value of $w=0.5$ and $c_1=c_2=1.5$ has been taken for obtaining best results. The population size is chosen to be 10 and the entire algorithm is run for 10 iterations. The average results of 10 simulations runs are then passed to K-Means algorithm.

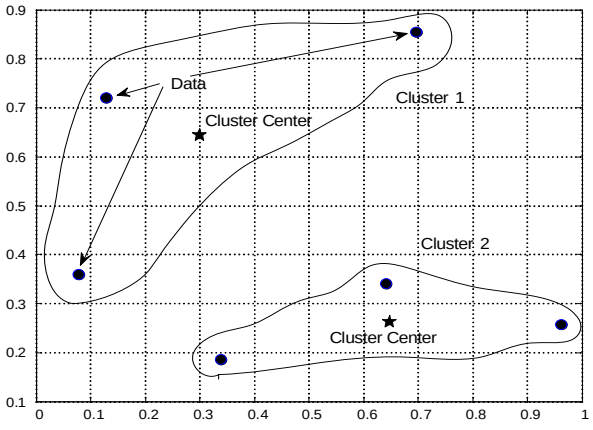

Figure 2. Cluster Centre and cluster formation by PSO for 6-data with 2-attributes

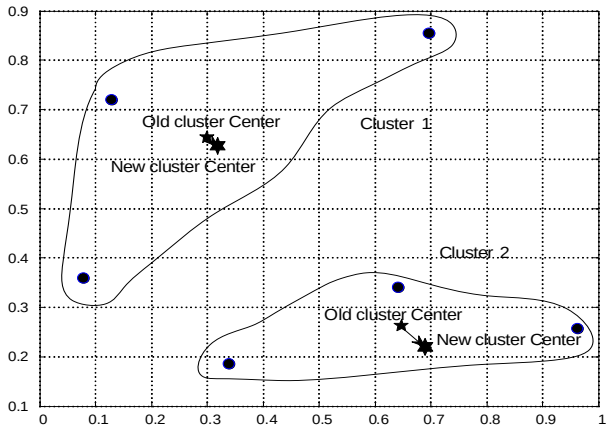

Figure 3. Cluster Centre refinement by K-Means Algorithm for 6-data with 2-attributes

For better understanding of the algorithm the complexity is further increased in the problem. Again 6 data are taken but with increase in attributes. The considered data has 3 attributes now. The results have been presented in Figure 4 with PSO algorithm. Figure 5 shows the result of our proposed hybrid sequential clustering algorithm in three dimensional spaces with increased attributes. The Figures 3 and 5 clearly show the improvement in the cluster centers. It is also observed that the intra cluster distance increased and inter cluster distance is minimized. The proposed algorithm results in the formation of more compact and more separable clusters and thus increases accuracy while new data have been added up.

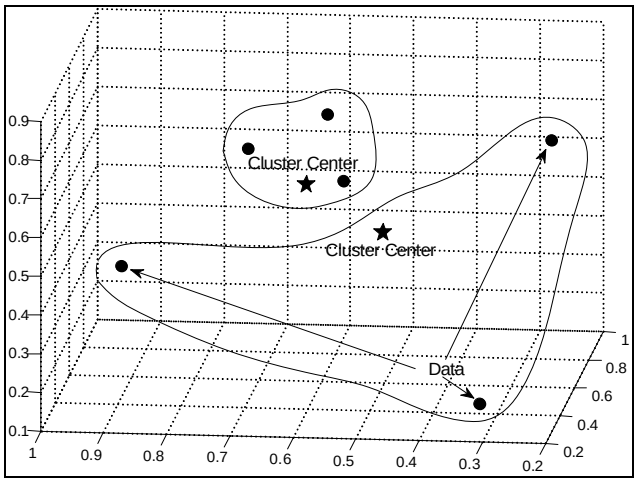

Figure 4. Cluster Centre and cluster formation by PSO Algorithm for 6-data with 3-attributes

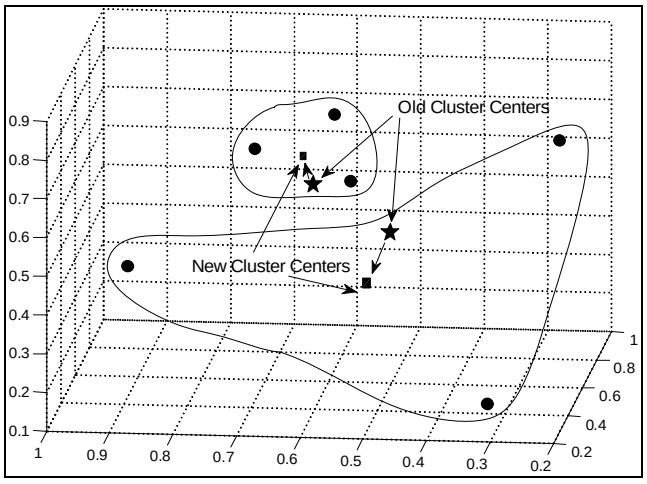

Figure 5. Cluster Centre refinement by K-Means Algorithm for 6-data with 3-attributes

Subsection 2*:* This Subsection presents comparison of the proposed scheme with K-Means, PSO, Hybrid, K-Means+Genetic algorithm. The accuracy and robustness of our proposed algorithm have been tested on four different problems. The classification problems are as follows:

(i) Artificial problem I: This problem formulation is made as per following classification rule (VenderMerwe and Engelbrecht, 2003).

$$y(z)=\begin{cases}1 & if\ (z_1 \geq 0.7) or((z_1 \leq 0.3) and(z_2 \geq -0.2 - z_1)) \\ 0 & otherwise\end{cases} \quad (7)$$

A total of 400 data vectors were randomly created, with z_1 and z_2 in a range [-1, 1].

(ii) Artificial problem II: This is a 2-dimensional problem with 4 unique classes (Merwe and Engelbrecht, 2003). The problem is interesting in that only one of the inputs is really relevant to the formation of the classes. A total of 600 patterns were drawn from four independent bivariate normal distributions, where classes were distributed according to equation (8) for *i* = 1, . . . ~ 4, where ***p*** is the mean vector and is the covariance matrix; *m1 = -3, m2 = 0, m3 = 3* and *m4* = 6.

$$N_2\left(\mu=\begin{pmatrix}m_i\\0\end{pmatrix}, \sum=\begin{bmatrix}0.50 & 0.05\\0.05 & 0.50\end{bmatrix}\right) \quad (8)$$

(iii) Wine Problem: These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents (inputs) found in each of the three types of wines (classes). These data are collected of 178 instances (data vectors) (http://archive.ics.uci.edu/ml/datasets.html). Hence, this is a classification problem with "well behaved class structures. There are 13 inputs, 3 classes and 178 data vectors.

(iv) Iris Data Set: This is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 150 instances each, where each class refers to a type of iris plant (http://archive.ics.uci.edu/ml/datasets.html). One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

The main purpose of our proposed Hybrid Sequential clustering algorithm is to compare the quality of the respective clustering, where quality is measured according to the following three criteria:

- The quantization error as defined in equation (9)

$$Q_e=\frac{\sum_{j=1}^{c}\left[\sum d(\overrightarrow{X^p}, z^j)/N_0\right]}{c} \quad (9)$$

  Where $d(\overrightarrow{X^p}, z^j)$ is distance to centroid, N_0 is number of data vectors to be clustered, c is the number of cluster to be formed.

- The intra-cluster distances, i.e. the distance between data vectors within a cluster, where the objective is to minimize the intra-cluster distances and is given by equation (10).

$$Intra=\frac{1}{n}\sum_{j=1}^{c}\left\|\overrightarrow{X^j}-z^j\right\|^2 \quad (10)$$

- The inter-cluster distances, i.e. the distance between the centroids of the clusters, where the objective is to maximize the distance between clusters is given by equation (11)

$$Inter=\min(\left\|z^i-z^j\right\|^2 \quad (11)$$

The results obtained from the five clustering algorithms (K-Means, PSO, Hybrid, K-Means+Genetic algorithm and Hybrid Sequential clustering algorithm) are summarized in Tables (1-3). In these algorithms, every run generates a new solution so the values reported are averaged over *30* simulations, with standard deviations to indicate the range of values to which the algorithms converge. Table 1 presents the comparison on the fitness of solutions, i.e. the quantization error. It can be noted that the results obtained through the Hybrid Sequential clustering algorithm has the smallest average quantization error.

The deviations in the results obtained are minimized in the proposed algorithm. All other algorithms have better solution in one or another case but there is no uniformity in the solution obtained. It is only the proposed Hybrid Sequential clustering algorithm which generates best among them.

Table 1. Comparison of K-Means, PSO, Hybrid, K-Means+Genetic Algorithm and Hybrid Sequential clustering algorithm with Quantization Error

Algorithm	Artificial problem I	Artificial problem II	Wine	Iris Data Set
K-Means	0.984 ±0.032	0.264 ±0.001	1.139 ±0.125	1.139 ±0.125
PSO	0.769 ±0.031	0.252 ±0.001	1.493 ±0.095	0.774 ±0.094
Hybrid	0.768 ±0.048	0.250 ±0.001	1.078 ±0.085	0.633 ±0.143
K-Means+Genetic algorithm	0.772 ±0.05	0.260 ±0.001	1.384 ±0.099	0.982 ±0.128
Hybrid Sequential clustering algorithm	0.764 ±0.031	0.250 ±0.001	1.072 ±0.084	0.628 ±0.092

Table 2. Comparison of K-Means, PSO, Hybrid, K-Means+Genetic Algorithm and Hybrid Sequential clustering algorithm with Intra-cluster Distance

Algorithm	Artificial problem I	Artificial problem II	Wine	Iris Data Set
K-Means	3.678 ±0.085	0.911 ±0.027	4.202 ±0.223	3.374 ±0.245
PSO	3.826 ±0.091	0.873 ±0.023	4.911±0.353	3.489 ±0.186
Hybrid	3.823 ±0.081	0.869 ±0.018	4.199 ±0.514	3.304 ±0.204
K-Means+Genetic algorithm	3.892 ±0.089	0.899 ±0.025	4.231 ±0.467	3.378 ±0.235
Hybrid Sequential clustering algorithm	3.647 ±0.080	0.864 ±0.016	4.199 ±0.223	3.300 ±0.204

Table 3. Comparison of K-Means, PSO, Hybrid, K-Means+Genetic Algorithm and Hybrid Sequential clustering algorithm with Inter-cluster Distance

Algorithm	Artificial problem I	Artificial problem II	Wine	Iris Data Set
K-Means	1.771 ±0.046	0.796 ±0.022	1.010 ±0.146	0.887 ±0.091
PSO	1.142 ±0.052	0.815 ±0.019	2.977 ±0.241	0.881 ±0.086
Hybrid	1.151 ±0.043	0.814 ±0.011	2.799 ±0.111	0.852 ±0.097
K-Means+Genetic algorithm	1.151 ±0.049	0.815 ±0.018	2.898 ±0.189	0.863 ±0.097
Hybrid Sequential clustering algorithm	1.779 ±0.043	0.815 ±0.022	2.983 ±0.113	0.894 ±0.089

Table 2 and Table 3 prersent the comparison of algorithms considering intra- and inter-cluster distances. These parameters are cosidered to ensure compact clusters with little deviation from the cluster centroids and larger separation between the different clusters. It can be seen from the results that Hybrid Sequential clustering algorithm successfully obtain better results than its counterparts.

It has been seen that for first two problems PSO generate better solution than K-Means, hybrid or K-Means+Genetic algorithm but for the other two the other algorithms are better while the proposed algorithm generates better solution among all of them. It is also seen that the deviation in results obatined by proposed Hybrid Sequential clustering algorithm is much less than its counter parts and hence proves its stability. It is because initial clustering made by PSO is further tuned with K-Means algorithm which has capability of obtaining better local optimal solution. Hence, the proposed solution always generates better solution than its counter algorithms.

6. Conclusion

This paper investigated the application of the PSO in sequence with K-Means to clustering problem. Five algorithms are tested, namely a standard K-Means, PSO, K-Means+ Genetic algorithm, Hybrid approach and the Hybrid Sequential clustering algorithm, where the swarms find the clusters centre and further refining is obtained through K-Means algorithm. The Hybrid Sequential

clustering algorithm is compared with two PSO approaches, K-Means clustering and with Genetic algorithm, which shows that the proposed clustering algorithm have better convergence to lower quantization errors, and in general, larger inter-cluster distances and smaller intra-cluster distances. The variation in the solutions obtained for different cases is also reported minimum in the proposed algorithm. It can be concluded that the drawback of finding optimal solution by K-Means can be minimized by using PSO over it. The variations in PSO algorithm and its hybridization with K-Means algorithm is proposed for future research.

References

Chen C. Y. and Fun Y., 2004. Particle swarm optimization algorithm and its application to clustering analysis. IEEE *International Conference on networking sensing and Control*, pp.789-79.

Cui X., Potok, T. E. and Palathingal. P., 2005. Document clustering using particle swarm optimization. *International Journal of Pattern Recognition and Artifcial Intelligence*, Vol. 19, No 3, pp. 185-191.

El-abd M. and Kamel M., 2005. Information exchange in multiple cooperating swarms. *IEEE swarm Intelligence Symposium*, pp. 138-142.

Jain A. R., Murthy M. N. and Flynn P. J., 1999. Data clustering: A Review. *ACM Computing Surveys*, Vol. 31, No 3, pp. 265-323.

Kanungo T., Mount D.M., Netanyahu N., Piatko C., Silverman R. and Wu A.Y., 2002. An efficient K-Means clustering algorithm: Analysis and implementation. *IEEE Trans. Patterns Analysis and Machine Intelligence*, Vol. 24, No 7, pp. 881-892.

Kanungo T., Mount D.M., Netanyahu N., Piatko C., Silverman R. and Wu A.Y., 2002. A local search approximation algorithm for K-Means clustering. *Computational Geometry: Theory and Applications*, SoCG'02, pp. 89-112.

Kumar A., Sabharwal Y. and Sen S., 2010. Linear time approximation schme for clustering problems in any dimensions. *Journal of ACM*, Vol.57, No 2 .pp 5:1-32.

Kennedy J. and Eberhart R. C., 1995. Particle swarm optimization. *IEEE International Conference on Neural Networks*, Perth Australia, Vol. 4, pp. 1942-1948.

Kennedy J., Eberhart R. C. and Shi Y., 2002. *Swarm intelligence*. Morgan Kaufmann.

Mitra S. and Acharya T., 2004. *Data Mining*. Wiley Publications.

Panov P., Dzeroski S. and Soldatova L., 2008. OntoDM: An ontology of data mining. *IEEE International Conference on Data Mining Workshops*, pp. 752-760.

Pyle D., 1999. Data preparation for data mining. *Morgan Kaufmann*.

Rajan J. and Saravanan V., 2008. A framework of an automated data mining system using autonomous intelligent agents. *International Conference on Computer Science and Information Technology*, pp. 700-704.

Sadu A., Kumar R. and Kavasseri R.G., 2009. Optimal placement of phasor measurement units using particle swarm Optimization. *World Congress on Nature & Biologically Inspired Computing*, pp.1708-1713.

Shi Y. and Eberhart R. C., 1998. Parameter selection in particle swarm optimization. *Evolutionary Programming*, Vol. 1441 of *Lecture Notes in Computers Science, Springer*. pp. 591-600.

Tsai C. Y. and Chiu C. C., 2008. Developing a feature weight self-adjustment mechanism for a K-Means clustering algorithm. *Computational Statistics and Data Analysis*, Vol. 52, pp. 4658-4672.

Vander Merwe D.W. and Engelbrecht A. P., 2003. Data clustering using particle swarm optimization. *Congress on Evolutionary Computation*, Vol. 1, pp. 215-220.

Xindong W., 2004. Data mining: Artifcial intelligence in data analysis. *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*.

Zalik K. R., 2008. An effcient k'means clustering algorithm. *Pattern Recognition Letters*, Vol. 29, pp. 1385-1391.

Zhang D. Q. and Chen S. C., 2004. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artifcial Intelligence in Medicine*, Vol. 32, No 1, pp. 37-50.

http://archive.ics.uci.edu/ml/datasets.html.

Biographical notes

Sandeep Rana received his master degree from U.P Technical University INDIA in Computer Application. He served as a lecturer in Sharda University. Currently he is working as Assistant System Manager in Gautam Buddha University (GBU) and also persuing Ph. D. from GBU. He is life time member of Indian Society of Technical Education and member of International Association of Computer Science and Information Technology. His area of interest is Artificial Intelligence and Data Mining.

Sanjay Jasola is Professor and Dean of School of Information and Communication Technology at Gautam Buddha University, Greater Noida, India. His research papers have been published in several International and national journals. He has also worked in Wawasan Open University, Malaysia and completed several international consultancy assignments. He is recipient of Gold Medal for innovation in open and distance learning from IGNOU, New Delhi. He is a fellow of

Institution of Engineers and Institution of Electronics and communication Engineers, India. His research interest includes wireless networking, mobile communication, Open educational resources.

Rajesh Kumar is Associate Professor in the Department of Electrical Engineering at the Malaviya National Institute of Technology (MNIT), INDIA. Presently he is Post Doctorate Research Fellow in the Department of Electrical and Computer Engineering at the National University of Singapore (NUS), SINGAPORE on leave from MNIT. He has been active in the research and development of Intelligent Systems and applications more than ten years, and is internationally known for his work in this area. Dr. Kumar has published over a hundred and twenty articles on the theory and practice of intelligent control, evolutionary algorithms, bio and nature inspired algorithms, fuzzy and neural methodologies, power electronics, electrical machines and drives. He has received the Career Award for Young Teachers in 2002 from Government of India. Dr. Kumar is a Senior Member IEEE, Member IE (INDIA), Fellow Member IETE, Senior Member IEANG and Life Member ISTE.

Received August 2010
Accepted October 2010
Final acceptance in revised form November 2010