Computer and Information Science
www.ccsenet.org/journal.html
Vol. 1, No. 4
November 2008

A New Approach for Data Clustering Based on PSO with Local Search

K. Premalatha (Corresponding Author)
Kongu Engineering College
Perundurai, Erode, TN, India
E-mail: kpl_barath@yahoo.co.in

A.M. Natarajan
Bannari Amman Institute of Technology
Erode, TN, India

Abstract

Data clustering is a popular approach for automatically finding classes, concepts, or groups of patterns. The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of the clustering process and the context in which clustering is used. This paper looks into the use of Particle Swarm Optimization (PSO) for cluster analysis. In standard PSO the non-oscillatory route can quickly cause a particle to stagnate and also it may prematurely converge on suboptimal solutions that are not even guaranteed to local optimal solution. In this paper a modification strategy is proposed for the particle swarm optimization (PSO) algorithm and applied in the data sets. This paper provides a method for particles to steer clear off from local stagnation and the local search is applied to improve the goodness of fitting. The effectiveness of this concept is demonstrated by cluster analysis. Results show that the model provides enhanced performance and maintains more diversity in the swarm and thereby allows the particles to be robust to trace the changing environment.

Keywords: PSO, Roulette-Wheel selection, K-Means, Local Search, Stagnation, Optimization

1. Introduction

Clustering algorithms can be categorized as either hierarchical or optimization. Hierarchical clustering techniques proceed by either a series of successive merges or a series of successive divisions. The result is the construction of a tree like structure or hierarchy of clustering's which can be displayed as a diagram known as a dendogram. Agglomerative hierarchical methods begin with the each observation in a separate cluster. These clusters are then merged, according to their similarity (the most similar clusters are merged at each stage), until only one cluster remains. Divisive hierarchical methods work in the opposite way. An initial cluster containing all the objects are divided into sub-groups (based on dissimilarity) until each object has its own group. Agglomerative methods are more popular than divisive methods.

Unlike hierarchical techniques, which produce a series of related clustering's, optimization techniques produce a single clustering which optimizes a predefined criterion or objective function. The number of clusters in this clustering is either specified a priori or is determined as part of the clustering method. Optimization methods start with an initial partition of objects into a specified number of groups. Objects are then reassigned to clusters according to the objective function until some terminating criterion is met. These methods differ with respect to the starting partitions, the objective functions, the reassignment processes, and the terminating criteria. Unlike hierarchical clustering techniques, optimization methods do not store similarity matrices. Thus the size of the data is not limited by storage space. However, there are a number of disadvantages affecting optimization methods:

(i) Some methods require the number of clusters a priori, and will divide the data into this number of clusters regardless of the data structure;

(ii) Certain clustering criterion are biased towards particular cluster shapes, and will impose these shapes on the data; and

(iii) The performance of optimization techniques is highly dependent on the initial partition.

In this study, a data clustering algorithm based on Simple PSO, Roulette Wheel Selection and K-Means algorithm. The remainder of this paper is organized as follows: Section 2 provides the related works in clustering. Section 3 gives a general overview of the PSO. The proposed PSO clustering algorithm is described in Section 4. Section 5 presents the detailed experimental setup and results for comparing the performance of the proposed PSO algorithm with the K-means approaches.

2. Related works

Even though there is an increasing interest in the use of clustering methods in pattern recognition [1], image processing [2] and information retrieval [4], clustering has a rich history in other disciplines [5] such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing. Other terms more or less synonymous with clustering include *unsupervised learning* [5], *numerical taxonomy* [6], *vector quantization* [7], and *learning by observation* [8]. The field of spatial analysis of point patterns [9] is also related to cluster analysis. The importance and interdisciplinary nature of clustering is evident through its vast literature. A survey of the state of the art in clustering *circa* 1978 was reported in Dubes and Jain [10]. A comparison of various clustering algorithms for constructing the minimal spanning tree and the short spanning path was given in Lee [11]. Cluster analysis was also surveyed in Jain et al. [12]. A review of image segmentation by clustering was reported in Jain and Flynn [2]. Comparisons of various combinatorial optimization schemes, based on experiments, have been reported in Mishra and Raghavan [13] and Al-Sultan and Khan [16].

3. Particle Swarm Optimization

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates. Particle Swarm Optimization (PSO) incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior, from which the idea is emerged [17][18]. PSO is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems. As an algorithm, the main strength of PSO is its fast convergence, which compares favorably with many global optimization algorithms like Genetic Algorithms (GA) [22], Simulated Annealing (SA) [20.] and other global optimization algorithms. For applying PSO successfully, one of the key issues is finding how to map the problem solution into the PSO particle, which directly affects its feasibility and performance.

Bird flocking optimizes a certain objective function. Each particle knows its best value so far (pbest) and its position. This information is analogy of personal experiences of each particle. Moreover, each particle knows the best value so far in the group (gbest) among pbests. This information is analogy of knowledge of how the other particles around them have performed. Namely, each particle tries to modify its position using the following information:

- current positions
- current velocities
- distance between the current position and pbest
- distance between the current position and gbest

This modification can be represented by the concept of velocity. Velocity of each particle can be modified by the following equation:

$$v_{id} = w * v_{id} + c_1 * rand() * (P_{id} - X_{id}) + c_2 * rand() * (P_{gd} - X_{id}) \quad (1)$$

Where, v_{id} : velocity of particle

x_{id} : current position of particle

w : weighting function,

c_1 & c_2 : determine the relative influence of the social and cognitive components

p_{id} : pbest of particle i,

p_{gd} : gbest of the group.

The following weighting function (2) is usually utilized in

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{iter_{\max}} x\ iter \quad (2)$$

Where, w_{max} : initial weight,

w_{min} : final weight,

$iter_{max}$: maximum iteration number,

iter : current iteration number.

Using the above equation, a certain velocity, which gradually gets close to pbest and gbest can be calculated. The current position (searching point in the solution space) can be modified by the following equation (3):

$$X_{id} = X_{id} + V_{id} \quad (3)$$

Fig. 1 shows the general flow chart of PSO.

The features of the searching procedure of PSO can be summarized as follows:

(a) As shown in equation (1)(2)(3), PSO can essentially handle continuous optimization problem.

(b) PSO utilizes several searching points like genetic algorithm (GA) and the searching points gradually get close to the optimal point using their pbests and the gbest.

(c) The first term of right-hand side (RHS) of (1) is corresponding to diversification in the search procedure. The second and third terms of that are corresponding to intensification in the search procedure. Namely, the method has a well-balanced mechanism to utilize diversification and intensification in the search procedure efficiently.

The above feature (c) can be explained as follows [18]. The RHS of (2) consists of three terms. The first term is the previous velocity of the particle. The second and third terms are utilized to change the velocity of the particle. Without the second and third terms, the particle will keep on "flying" in the same direction until it hits the boundary. Namely, it tries to explore new areas and, therefore, the first term is corresponding to diversification in the search procedure. On the other hand, without the first term, the velocity of the "flying" particle is only determined by using its current position and its best positions in history. Namely, the particles will try to converge to the pbests and/or gbest and, therefore, the terms are corresponding to intensification in the search procedure.

4. Proposed PSO for Data clustering

The original PSO described in section 3 is basically developed for continuous optimization problems. However, lots of practical engineering problems are formulated as combinatorial optimization problems. Kennedy and Eberhart developed a discrete binary version of PSO for the problems (Kennedy, 1997). The proposed system employs Discrete Binary PSO with gobalized and localized search.

4.1 Problem Formulation

The fitness of panicles is easily measured as the quantization error. The fitness function of the data clustering problem is given as follows:

$$f = \frac{\sum_{i=1}^{N_c} \{ \frac{\sum_{j=1}^{p_i} d(0_i, m_{ij})}{p_i} \}}{N_c} \quad (4)$$

The function f should be minimized.

where m_{ij} : jth data vector belongs to cluster i

O_i : Centroid vector of the i^{th} cluster

$d(O_i, m_{ij})$: the distance between data vector *mij* and the cluster centroid *Oi.*

Pi : stands for the number of data set, which belongs to cluster ***Ci;***

Nc : number of clusters.

4.2 Particle Representation

In the context of clustering, a single particle represents the cluster centroid vectors. That is, each particle $\mathbf{X}_{ij}$, is constructed as follows:

$X_{ij} = (mi_1, m_{i2}, ..., mi_m)$

where mij refers to the j-th cluster centroid vector of the i-th particle in cluster Cij. Therefore, a swarm represents a number of candidates clustering for the current data vectors.

4.3 Initial Population

One particle in the swarm represents one possible solution for clustering. Therefore, a swarm represents a number of candidate clustering solutions for the data set. At the initial stage, each particle randomly chooses k different data set from the collection as the initial cluster centroid vectors and the data sets are assigned to cluster based on one iteration of K-Means.

4.4 Local search

After finding the solutions of N particles, a local search is performed to further improve fitness of these solutions. Local search helps to generate better solutions, if the heuristic information can not be discovered easily. Local search is applied on all generated solutions or on a few percent N. In this work, local search is performed on 20% of the total solutions. So in the test data set of N data, local search is applied on the 20% of solutions based on roulette-wheel selection. The requirement is that the fittest individuals have a greater chance of selection than weaker ones. In the local search procedure, the objective function values selected particles are computed again. These solutions can be accepted only if there is an improvement on the fitness, namely, if the newly computed objective function value is lower than the first computed value, newly generated solution replaces the old one.

4.5 Personal best & Global best positions of particle

The personal best position of particle is calculated as follows

$$P_{id}(t+1) = \begin{cases} P_{id}(t) & \text{if } f(X_{id}(t+1)) \geq f(P_{id}(t)) \\ X_{id}(t+1) & \text{if } f(X_{id}(t+1) < f(P_{id}(t)) \end{cases}$$

The particle to be drawn toward the best particle in the swarm is the global best position of each particle. At the start, an initial position of the particle is considered as the *personal best* and the *global best* can be identified with minimum fitness function value.

4.6 Finding new solutions

According to its own experience and those of its neighbors, the particle adjusts the centroid vector position in the vector space at each generation. The new velocity is calculated based on equation (1) and changing the position based on equation(3)

Generally, in PSO algorithm, operations described above are iterated in main loop until a certain number of iterations are completed or all particles begin to generate the same result. This situation is named as stagnation behavior, because after a point, algorithm finishes to generate alternative solutions. The reason of this situation is, after a certain number of iterations, particles generate continuously the same solutions. Aiming minimizes the stagnation behavior of particles, the proposed technique follows the Quantization error of particles and if there is no change on the error after last 10 iterations, it moves the particles with the random velocities. In other words, to improve the solution, a feedback technique is applied on the algorithm. Fig 2 demonstrates the proposed Hybrid PSO for data clustering.

5. Experiment Results

In this section, results from the proposed PSO method and the K-Means on well-known test data sets are reported. The choice of the parameter values seems not to be critical for the success of the methods; it appears that faster convergence can be obtained by proper fine-tuning. The balance between the global and local exploration abilities of the proposed system is mainly controlled by the inertia weight, since the positions of the particles are updated according to the classical PSO strategy. A time decreasing inertia weight value, i.e., start from 0.9 and gradually decrease towards 0.4, proved to be superior to a constant value. The optimal solution (fitness) is determined with N=20, c1= 2.1 & c2 = 2.1. The test data sets are obtained from UCI's machine learning repository [23]. The Results obtained from test data sets by K-Means and the proposed system are shown in Table 1 & Table 2 respectively.

Iris plants database: This is a well-understood database with **4** inputs, **3** classes and 150 data vectors.

Wine: 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 found in each of the three types of wines.

Glass identification: From USA Forensic Science Service; 6 types of glass; defined in terms of their oxide content.

For each data set with two different distance measures 50 runs have been performed using the proposed PSO and the performance is exhibited in terms of the Fitness value , Inter and Intra Cluster similarity. Results for all of the aforementioned datasets are reported with the conventional cluster algorithm K-Means. Table 1 illustrates the analysis of the results for K-Means and Table 2 shows for Proposed PSO system

Conclusion

The advantages of the PSO are very few parameters to deal with and the large number of processing elements, so called dimensions, which enable to fly around the solution space effectively. On the other hand, it converges to a solution very quickly which should be carefully dealt with when using it for combinatorial optimization problems. In this study, the proposed PSO algorithm developed for data-clustering problem is verified on the datasets. It is shown that it increases the performance of the clustering and the best results are derived from the proposed technique. Consequently, the proposed technique markedly increased the success of the data-clustering problem.

References

Carlisle, and G. Dozier, (2001). "An off-the-shelf particle Swarm Optimization", *Proceedings of the Workshop on Particle Swarm Optimization*, Indianapolis, IN: Purdue School of Engineering and Technology, IUPUI.

Al-Sultan, K. S. And Khan, M. M. (1996). Computational experience on four algorithms for the hard clustering problem. *Pattern Recogn. Lett. 17*, 3, 295–308.

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY.

Dubes, R. C. And Jain, A. K. (1980). Clustering methodology in exploratory data analysis. In *Advances in Computers*, M. C. Yovits,, Ed.Academic Press, Inc., New York, NY, 113–125.

E. S. Gelsema And L. N. Kanal, Eds. 425–436.Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley and Sons, Inc., New York, NY.

Jain, A. K. And Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

Jain, A. K. And Flynn, P. J. (1996). Image segmentation using clustering. In *Advances in Image Understanding: A Festschrift for Azriel*

Jain, N. C., Indrayan, A., Goel, L. R. (1986). Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recogn. 19*, 1 (Jan./Feb. 1986), 95–99.

Kennedy J and Eberhart R (2001). Swarm intelligence. Morgan Kaufmann Publishers,Inc., San Francisco, CA.

Lee, R. C. T. (1981). Cluster analysis and its applications. In *Advances in Information Systems Science*, J. T. Tou, Ed. Plenum Press, New York, NY.

Michalski, R., Stepp, R. E., And Diday, E. (1983). Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5*, 5 (Sept.), 396–409.

Mishra, S. K. And Raghavan, V. V. (1994). An empirical study of the performance of heuristic methods for clustering. In *Pattern Recognitionin Practice*.

Oehler, K. L. And Gray, R. M. (1995). Combining image compression and classification using vector quantization. *IEEE Trans. Pattern Anal. Mach. Intell. 17*, 461–473.

R. Eberhart and J. Kennedy, (1995). "A new optimizer using particle swarm theory", *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, Nagoya, Japan, pp.39-43, Piscataway, NJ: IEEE Service Center.

R. Eberhart and Y. Shi, (2000). "Comparing inertia weights and constriction factors in particle swarm optimization", *Proc. of Congress on Evolutionary Computation (CEC2000)*, San Diego, CA, pp 84-88.

Rasmussen, E. (1992). Clustering algorithms. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419–442.

Ripley, B. D., Ed. (1989). *Statistical Inference for Spatial Processes*. Cambridge University Press, New York, NY.

Rosenfeld, N. Ahuja and K. Bowyer, Eds, IEEE Press, Piscataway, NJ, 65–83.

S. Naka, T. Genji, T. Yura, and Y. Fukuyama, (2001). "Practical Distribution State Estimation Using Hybrid Particle Swarm Optimization", *Proc. Of IEEE Power Engineering Society Winter Meeting*, Columbus, Ohio.

Sneath, P. H. A. And Sokal, R. R. (1973). *Numerical Taxonomy*. Freeman, London, UK.

Spath, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification*. Ellis Horwood, Upper Saddle River, NJ.

UCI Repository for Machine Learning Databases retrieved from the World Wide Web: http://www.ics.uci.edu/~mlearn/MLRepository.htm.

Y. Shi and R. Eberhart, (1998). "A Modified Particle Swarm Optimizer", *Proceedings of IEEE International Conference on Evolutionary Computation (ICEC'98)*, pp.69-73, Anchorage.

Table 1. Analysis with K-Means

Data sets	Distance Measure	K-Means Clustering		
		FV	Intra	Inter
Iris	Euclidean	0.8013	0.0616	5.2805
	Chebychev	0.6873	0.1902	4.7052
Wine	Euclidean	126.14	11.4103	759.170
	Chebychev	124.68	11.0918	759.008
Glass	Euclidean	1.5968	0.49094	6.2713
	Chebychev	1.1856	0.2544	5.0068

Table 2. Analysis with Proposed PSO System

Data sets	Distance Measure	Proposed PSO System		
		FV	Intra	Inter
Iris	Euclidean	0.5439	0.0616	9.8228
	Chebychev	0.4209	0.0537	9.2193
Wine	Euclidean	83.826	5.4399	831.25
	Chebychev	83.416	3.9643	822.12
Glass	Euclidean	0.5991	0.4909	10.2561
	Chebychev	0.4209	0.1569	9.8352

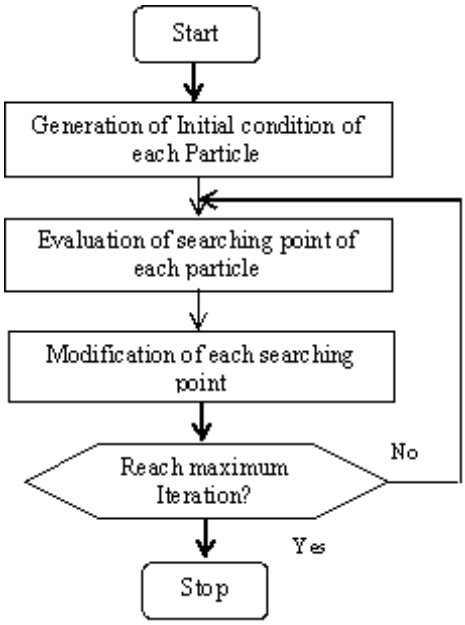

Figure 1. Simple PSO

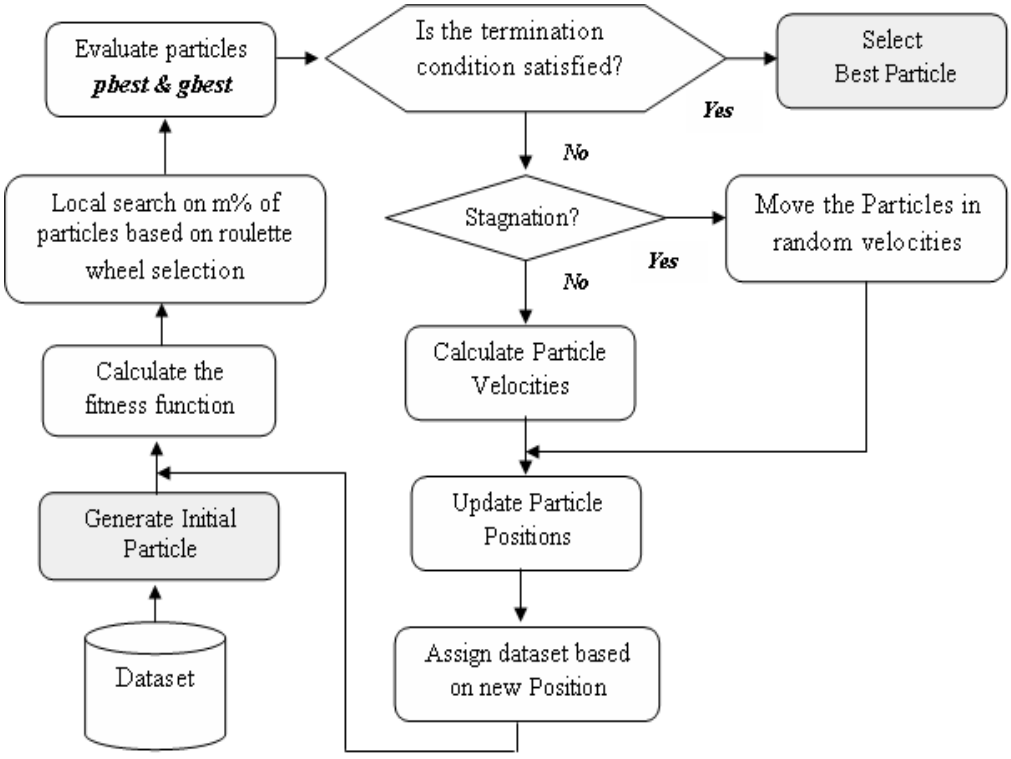

Figure 2. Hybrid PSO for Data Clustering