Data clustering based on particle swarm optimization with Lévy mechanism

Xiaoyong Liu*

Department of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, 510665, China

Received 6 September 2014, www.cmnt.lv

Abstract

Clustering analysis is a popular approach in data mining field. It is often used to automatically find classes or groups for unlabeled datasets. 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, Lévy Mechanism is proposed for the particle swarm optimization (PSO) algorithm and applied in the data sets. Results show that the new PSO model, named LPSO, provides enhanced performance for clustering data.

Keywords: Particle Swarm Optimization, Data clustering, Lévy Mechanism

1 Introduction

Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling [1, 2]. PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). The algorithm is initialized with a population of random solutions firstly, and then searches for optimal solutions by updating population. However, unlike GA, PSO has no evolution operators such as crossover and mutation. In PSO, the potential solutions, named particles, search through the problem space by following the current optimum particles. PSO has been successfully applied in many areas: function optimization, artificial neural network training, fuzzy system control, and other areas where GA can be applied, etc [3, 4].

Cluster analysis is a method for clustering a data set into groups of similar individuals [5, 6]. It is a branch in multivariate analysis and an unsupervised learning in pattern recognition. Cluster analysis identifies and classifies objects individuals or variables on the basis of the similarity of the characteristics they possess. It seeks to minimize within-group variance and maximize between-group variance. Its aim is to establish a set of clusters such that cases within a cluster are more similar to each other than they are to cases in other clusters.

Accurate diagnosis and effective treatment of disease is an important issue in life science research and has a positive meaning for human health. Recently, medical experts pay more attention to early diagnosis of disease and propose many new methods to deal with disease diagnosis problem. Using clustering analysis methods to diagnose disease is rapid development of a novel research branch of machine learning. Researchers have applied artificial intelligence and computer technology to develop some medical diagnostic systems, which improve the efficiency of diagnosis and become practical tools.

2 Clustering algorithm based on an adaptive pso with lévy mechanism

2.1 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 behaviours observed in flocks of birds, schools of fish, or swarms of bees, and even human social behaviour, from which the idea is emerged [9][10]. 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 favourably with many global optimization algorithms like Genetic Algorithms (GA), Simulated Annealing (SA). 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

**Corresponding author* e-mail: lxyong420@126.com

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:

$$v_{id} = w \cdot v_{id} + c_1 \cdot rand() \cdot (P_{id} - X_{id}) \\ + c_2 \cdot rand() \cdot (P_{gd} - X_{id}), \tag{1}$$

where v_{id} is the velocity of particle, X_{id} is the current position of particle, w is the weighting function, c_1 & c_2 determine the relative influence of the social and cognitive components, P_{id} is the pbest of particle *i*, P_{gd} is the gbest of the group.

The following weighting function (2) is usually utilized in

$$w_i = w_{\max} - \frac{w_{\max} - w_{\min}}{iter_{\max}} \cdot iter_i \quad , \tag{2}$$

where $w_{\max}$ is the initial weight, $w_{\min}$ the final weight, $iter_{\max}$ is the maximum iteration number, $iter_i$ is the 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):

$$X_{id} = X_{id} + v_{id}, \tag{3}$$

Figure 1 shows the general flow chart of PSO.

2.2 L'EVY MECHANISM

Various studies have shown that flight behaviour of many animals and insects has demonstrated the typical characteristics of Lévy flights. A recent study by Reynolds and Frye shows that fruit flies or *Drosophila melanogaster* explore their landscape using a series of straight flight paths punctuated by a sudden 90° turn, leading to a Lévy-flight-style intermittent scale free search pattern. The typical feature of Lévy flights is also showed on human behaviour. Even light can be related to Lévy flights.

Subsequently, such behaviour has been applied to optimization and optimal search, and preliminary results show its promising capability.

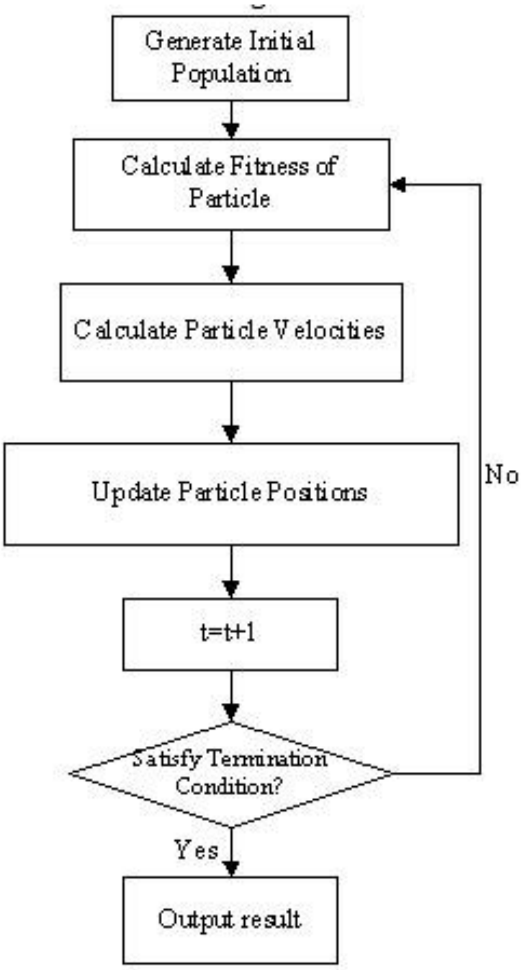
Generate Initial Population
Calculate Fitness of Particle
Calculate Particle Velocities
Update Particle Positions
t=t+1
Satisfy Termination Condition?
No
Yes
Output result

FIGURE 1 Simple PSO

A Lévy flight is performed as Formula (4) [7, 8]:

$$X_i^{t+1} = X_i^t + \alpha \oplus \text{Levy}(\lambda), \tag{4}$$

where, $\alpha > 0$ is the step size which should be related to the scales of the problem of interests. In most cases, we can use $\alpha = 1$. The above equation is essentially the stochastic equation for random walk. In general, a random walk is a Markov chain whose next location only depends on the current location (the first term in the above equation) and the transition probability (the second term). The symbol $\oplus$ means entrywise multiplications. This entrywise product is similar to those used in PSO, but here the random walk via Lévy flight is more efficient in exploring the search space as its step length is much longer in the long run. The Lévy flight essentially provides a random walk while the random step length is drawn from a Lévy distribution (see Figure 2).

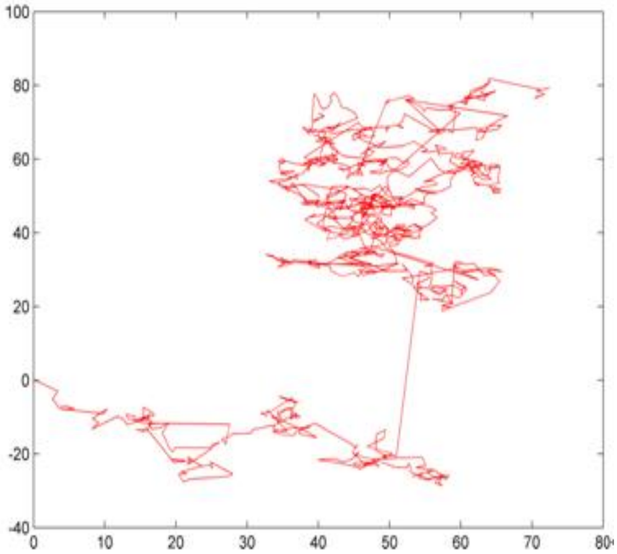

FIGURE 2 Random walk trajectories of Levy flight ($\alpha = 2$)

2.3 PSO WITH L'EVY MECHANISM FOR CLUSTERING

The original PSO described in section 2.1 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. The proposed system employs Discrete Binary PSO with globalized and localized search ［9， 10］.

2.3.1 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:

$$f = \frac{\sum_{i=1}^{N_c}\left(\frac{\sum_{j=1}^{p_i} d\left(C_i, m_{ij}\right)}{p_i}\right)}{N_c} . \tag{5}$$

The function *f* should be minimized, where m_{ij} is the *j*th data vector belongs to cluster *i*, C_i is the centroid vector of the *i*th cluster, $d\left(C_i, m_{ij}\right)$ is the distance between data vector m_{ij} and the cluster centroid C_i , p_i stands for the number of data set, which belongs to cluster C_i , N_c is the number of clusters.

2.3.2 Particle Representation

In the context of clustering, a single particle represents the cluster centroid vectors. That is, each particle X_{ij} is constructed as follows: X_{ij} = (m_{i1} , m_{i2} , ⋯, m_{ij}), where m_{ij} refers to the *j*-th cluster centroid vector of the *i*-th particle in cluster C_{ij} . Therefore, a swarm represents a number of candidates clustering for the current data vectors.

2.3.3 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.

2.3.4 Personal best and Global best positions of particle

The personal best position of particle is calculated as follows

$$P_{id}(t+1) = \begin{cases} P_{id}(t) & \left(f\left(X_{id}(t+1)\right) \geq f\left(P_{id}(t)\right)\right) \\ X_{id}(t+1) & \left(f\left(X_{id}(t+1)\right) < f\left(P_{id}(t)\right)\right) \end{cases} . \tag{6}$$

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.

2.3.5 Finding new solutions

According to its own experience and those of its neighbours, 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 (7) and equation (8):

$$w_i = w_{\min} + \frac{(w_{\max} - w_{\min})}{\exp\left(\tau \cdot \frac{iter_i}{iter_{\max}}\right)^2} , \tag{7}$$

$$X_{id}(t+1) = X_{id}(t) + Levy(\lambda) \cdot v_{id}(t+1) . \tag{8}$$

Figure 3 demonstrates the proposed PSO with Lévy Mechanism for data clustering.

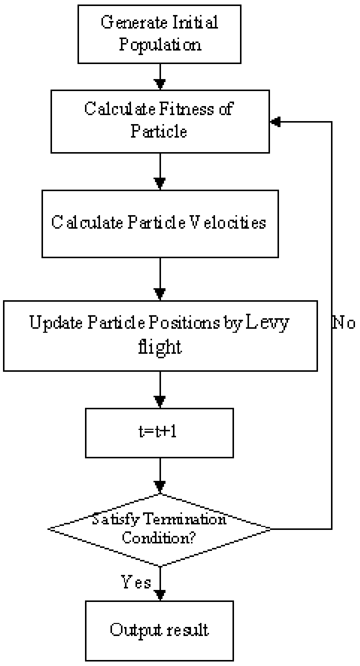
Generate Initial Population
Calculate Fitness of Particle
Calculate Particle Velocities
Update Particle Positions by Levy flight
t=t+1
Satisfy Termination Condition?
No
Yes
Output result

FIGURE 3 PSO with Lévy Mechanism (LPSO) for Data Clustering

3 Numerical examples

For the compare of performance between PSO and LPSO, two standard datasets are chosen to test the new algorithm. The program of the new algorithm is written by Matlab 2012b and run on a computer with 2.0 GHz CPU, 1GB DDR RAM.

TABLE 1 Parameters setting of PSO and LPSO Algorithm

Parameter	**PopSize**	**Iteration**	c_1	c_2
Value	**50**	**1000**	**1.2**	**1.2**

In this study, numerical experiments use two datasets, iris dataset and breast cancer dataset from UCI Machine Learning Repository (http://archive.ics.uci.edu/ ml/). The iris dataset consists of 150 data points with four attributes, and it is stored in a text file. It is one of the best known datasets to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each class, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other. Breast cancer dataset has 699 instances. Number of attributes of each instance is nine. There are thirteen numerical attributes including radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), smoothness (local variation in radius lengths), compactness, concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension. There are 458 as "benign" and 241 as "malignant". Table2 shows the detail of two datasets. Continuous attributes of two datasets are normalized firstly, and then used to train and test.

TABLE 2 Dataset

Dataset	**Number of Instances**	**Number of Attributes**	**Class**
Iris Data Set	150	4	3
Breast Cancer Data Set	699	9	2

TABLE 3 The compare of numeric results in Heart

	Best Value	**Best time (s)**
PSO	126.2	**76.5**
LPSO	**86.8**	**74.3**

TABLE 4 The compare of numeric results in Breast

	Best Value (10^2)	**Best time (s)**
PSO	197.93	**283.13**
LPSO	**197.29**	**281.34**

4 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 search all of 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 with Lévy mechanism developed for data-clustering problem is verified on the disease 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.

Acknowledgements

The author would like to thank anonymous reviewers for their constructive and enlightening comments, which improved the manuscript. This work has been supported by grants from Program for Excellent Youth Scholars in Universities of Guangdong Province (Project No. Yq2013108). The authors are partly supported by the Key grant Project from Guangdong provincial party committee propaganda department, China (Project No.LLYJ1311).

References

[1] Eberhart R, Kennedy J 1995 A new optimizer using particle swarm theory *Proceedings of the Sixth International Symposium on Micro Machine and Human Science* 39-43

[2] Eberhart R, Shi Y 2000 Comparing inertia weights and constriction factors in particle swarm optimization *Proceedings of Congress on Evolutionary Computation* 84-8

[3] *Deleted by CMNT Editor*

[4] Khan A, Bawane N G, Bodkhe S 2010 An Analysis of Particle Swarm Optimization with Data Clustering-Technique for Optimization in Data Mining *International Journal on Computer Science and Engineering* **2**(7) 2223-6

[5] Kaufman L, Rousseeuw P J 2009 *Finding Groups in Data: An Introduction to Cluster Analysis*. Press: Wiley, New York Ch 1,1-2

[6] Niknam T, Amiri B 2010 An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis *Applied Soft Computing* **10**(1) 183-97

[7] *Deleted by CMNT Editor*

[8] Saida I B, Nadjet K, Omar B 2014 A new algorithm for data clustering based on cuckoo search optimization *Genetic and Evolutionary Computing Springer-Verlag* 55-64

[9] Yuan K H, Shu Y X, Wei W 2014 Particle swarm optimization clustering for cement kilning system fault recognition *International Journal of Industrial and Systems Engineering* **17**(4) 477-94

[10] *Deleted by CMNT Editor*

Author

Xiaoyong Liu, born in 1979, Guangdong, China

Current position, grades: the associate Professor of Guangdong Polytechnic Normal University, China.
University studies: He received his master's degree from South China University of Technology in China. He received the Ph.D. degree from Graduate University of Chinese Academy of Sciences in China.
Scientific interests: His research interest fields include data mining, ant colony optimization and genetic algorithm.
Publications: more than 20 papers published in various journals.
Experience: He has teaching experience of more than 7 years and has completed 2 scientific research projects.