International Journal of Hybrid Information Technology
Vol.8, No.7 (2015), pp.215-224
http://dx.doi.org/10.14257/ijhit.2015.8.7.20

A New Clustering Algorithm of Hybrid Strategy Optimization

Li Yi-ran[1] and Zhang Chun-na[2]

1. College of Applied Technology, University of Science and Technology Liaoning, Anshan Liaoning 114011, China
2. School of Software, University of Science and Technology Liaoning, Anshan Liaoning 114051, China
E-mail：lyr7879@163.com

Abstract

Normally, improving the performance of clustering depends on improvement of the algorithm. On the basis, this paper presents a hybrid strategy optimization algorithm that K-means algorithm effectively combined with PSO algorithm, which not only has played their respective advantages, but also reflected a hybrid performance. First of all, combined with a semi-supervised clustering idea, to optimize the clustering center of particle by K - means in the iteration of algorithm, enhanced the searching capability of the particles. Secondly, improved the traditional K - means enhance the ability of the algorithm to deal with the concave and convex points. Finally, the algorithm is introduced into the particle state determination mechanism, on implementing mutation for unstable particles, so that the algorithm to obtain stable performance. Experimental results show that the hybrid algorithm optimization ability is outstanding, and the convergence and stability can be effectively improved.

Keywords: *K-means algorithm; PSO; convergence; hybrid strategy*

1. Introduction

Clustering analysis is an important research direction in the field of machine learning, and is widely used in data mining, pattern recognition, etc. For a given sample set of no classless to identify , by setting a good criterion function, to re-plan the sample set, to distinguish it into different subsets of the similarity degree in the same kind of the data is high, and in different classes is low. A semi-supervised clustering is a new kind of clustering method, which integrated the advantage of without supervision learning and supervised learning, can use a small amount of the tag information processing unlabeled data, in order to effectively improve the quality of clustering.

Currently, the semi-supervised clustering can be roughly divided into two types: based-constraint and based-distance [1]. The former is a clustering with constraint, the domain-specific knowledge in an agreed manner introduced into the clustering process, so that the algorithm to get more instructive domain knowledge, more targeted searches and clustering effect is guaranteed. Here, the pairwise constraints proposed by Wagstaff et al. widely used. The latter is based on the distance, through the analysis of the label data obtained metrics to implement clustering. In practice, the more is the combination of the two [2, 3].

There are some common problems of the two algorithms, such as, over-reliance on the Euclidean distance, making part of a priori knowledge of the tag data has not been put to good use. Furthermore, for high spatial complexity of large high-dimensional data processing ineffective, at the same time, the convergence is slow. Based on this, this paper presents a hybrid strategy optimization algorithm that K-means algorithm effectively combined with PSO algorithm, searching in the space of algorithm, the position of each particle from the cluster centers, iteration using the optimized K-means algorithm to optimize the new particles, in order to enhance the convergence of the

ISSN: 1738-9968 IJHIT

algorithm, to determine the state of the particle, and accordingly adjust the speed[4]. This method not only considers the over-reliance on the Euclidean distance problems, also take into account the convergence problem of the search space, to ensure the clustering results are very accurate.

2. Improved Semi-Supervised Clustering Algorithm

2.1 K-Means Algorithm

Clustering is divided the n wait for classified objects into different sections according to the rules, each section is a cluster. The most commonly used is the K-means clustering algorithm [5, 6]. The basic strategy of the algorithm is free to choose k objects as the initial cluster centers in the sample's space, the remaining objects compared distance with the nearest cluster center and divided into the nearest cluster, then accompanied by changes of the cluster center, the object's position repeatedly moving until meet the requirements of criterion function, clustering end[7].

Set Q -dimensional space S^Q limited set $X=\{x_1,x_2,\cdots,x_n\}$, initialized randomly divided into k class, is defined as $C_1,C_2,\cdots,C_k$, if the class has n objects, the cluster centers of the i class are defined as $Z_1,Z_2,\cdots,Z_k$, $Z_i=\frac{1}{n}\sum_{j=1}^{n}x_j, j\in[1,k]$ the objective function is defined as follows:

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n}D^2_{x_j,Z_i} \tag{1}$$

Among them, $D^2_{x_j,Z_i}$ represents a distance between the j -th text and the i -type cluster center, namely the Euclidean distance.

The core idea of algorithm is that through continuous iterative to find the k-best cluster center of sample data sets, other data moves to the cluster center, until the value of the objective function is minimized.

2.2 Improve Algorithm

The chaos phenomenon is widespread in nature, it belongs to the category of nonlinear and has the characteristics of ergodicity and regularity, sensitive to initial conditions, can search to all internal state in accordance with the laws of its own in a given area, and does not repeat[9,10]. In this way, can take advantage of the nature of the chaos to search optimization, search steps are as follows:

The traditional K - means algorithm is suitable for processing rules clusters, for the concave and convex points between two points is not too much to consider, thus, when calculating the distance easily lead to inaccurate results, affect the selected of cluster centers. Based on this, this paper correct the distance function in formula (1), set a new distance function $D'(x_j,Z_i)$, as the shortest distance that two points in space across the concave and convex points, therefore, the objective function amended as follows:

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n}D'^2_{x_j,Z_i} \tag{2}$$

For the interpretation of the amended distance function can be described as $D(a,b)$, among them, a,b are two points that waiting for calculation, related instructions are as follows:

1) If does not exist the concave and convex points between two points, then two points can be connected a straight line, formula (1) does not need to be modified;
2) Finding each concave and convex point that can connect a straight line with point a $\{B_1, B_2, \cdots, B_n\}$, calculate the distance to the point b, recorded as $d(B_i, a)$, in the formula, $1 \le i \le n$;
3) Calculating straight-line distance from each point in $\{B_1, B_2, \cdots, B_n\}$ can directly reach the point b, denote $d(B_i, b)$, using recursive method for the point of not directly reach, hierarchical iterative calculation distance, finally obtained the closest distance between a, b points, $d(B_i, a) + \cdots + d(B_k, b)$.

The improved algorithm can significantly improve the accuracy of clustering, but time efficiency will be reduced, at the same time, it is still sensitive to the initial value because based on the traditional K-means, and easy to produce clustering deviation for isolated point.

3. The Clustering Algorithm Combined Number of Particle

3.1 Particle Swarm Optimization

Particle Swarm Optimization, referred to as PSO, is a recently developed evolutionary algorithm, the idea originated foraging behavior from birds, fish and other groups [8-11]. Algorithm assumes that each particle in the space can be used as a solution, the optimal solution produced in the iterative process, the change of the particle's position determines by its own speed in each iteration, PSO is keeping up with current optimum particle, and conduct their own searches, finally found the extreme value of individual and global[12,13].

Setting a P -dimensional space, the size of the particle swarm is $\{x_1, x_2, \cdots x_n\}$, position and velocity of the particle are expressed as $x_i = \{x_{i1}, x_{i2}, \cdots, x_{iQ}\}$, $v_i = \{v_{i1}, v_{i2}, \cdots, v_{iQ}\}$, $i = 1, 2, \cdots, n$. The direction of particle is random in traveling, but it will follow the trajectory of the local extreme value Pb and global extreme value Gb, current position and velocity of particles as the iteration, according to the fitness value is updated every time, iteration also need to determine whether the flight termination condition is satisfied, and eventually get the optimal solution[14]. The current state formula of particle is as follows:

$$v_{iq}^{'} = \alpha v_{iq}^{"} + \beta^0 (x_{Pb} - x_{iq}) + \beta^1 (x_{Gb} - x_{iq}) \tag{3}$$

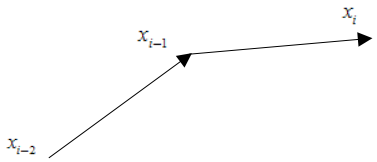
$$x_{iq}^{'} = x_{iq}^{"} + v_{iq}^{'} \tag{4}$$

In the formula, α is the inertia weight used to balance the velocity relationship between the current particle and the past in the process of moving. β^0、$\beta^1 \in [0,1]$, randomly distributed. Velocity of the particles is limited to a specified range $[v_{\min}, v_{\max}]$, if the speed is too small it is very easy to fall into local optimum, but the speed is too large, it is easy to cross the global optimal solution. The particle's flight direction depends on the three values of the formula (3), own speed $v_{iq}^{"}$, the distance between local optimum and current position $x_{Pb} - x_{iq}$, the distance between global optimum and current position $x_{Gb} - x_{iq}$, and with three weight coefficients.

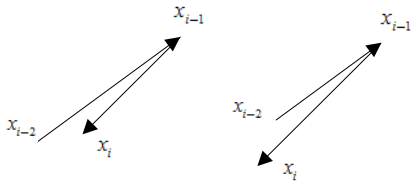
It is worth noting that the particles in the process of evolution, its performance directly affects the optimal solution is obtained.

3.2 Particle Swarm Optimization

In evolution, the state of the particle can be divided into two categories: normal forward and oscillation not moving, the particle's the obvious symbol of performance degradation appear the state of oscillation not moving at a certain dimension, as shown in the figure below:



（a）Normal State



（b）Abnormal State

Figure 1. Particle Path Diagram

Figure 1 is the abstract of the particle trajectory, the figure (a) is normal motion state, while figure (b) shows the particle reciprocating oscillation in a certain area. Determine a particle whether there is oscillation phenomenon by the formula (5):

$$\frac{\max(|x_i - x_{i-1}|, |x_{i-1} - x_{i-2}|)}{|x_i - x_{i-2}|} > h \tag{5}$$

If the position of the particle's before and after generation satisfies the formula (5), then the particle have the phenomenon of oscillation, led to the decrease of the search ability, even convergence in advance. x_i、x_{i-1}、x_{i-2} Indicates the position of the particle's before and after generation.

For determining the performance of the particles, the paper set up a basic principle: in the particle flight, every generation followed the optimal particle, it explain the performance of the particle is good; On the contrary, if the successive generations particle appear reciprocating oscillation or stagnation phenomenon, it indicates that the particle's performance fall. Here, defines two parameters for analysis: oscillation factor and stagnation factor. The following are described:

(1) Oscillating factor, if before and after two generations of particle meet formula (5) in iteration, it is determined that the particles oscillate, initialize oscillation factor $\theta_1 = 0$, if there is oscillation, θ_1 is incremented 1. For each generation particle in iteration are suitable formula (5), until the particles appears variation so far.

(2) Stagnation factor, this factor is used to detect the working condition of optimal particles, initialize stagnation factor $q_2 = 0$, detect whether coincidence the position of current optimal particle with the position of the previous iteration in iteration, if there is no change, q_2 is incremented 1. Otherwise q_2 would be set 0.

In order to better illustrate the performance of the particle, the following define the particle's performance indicators, combining oscillation factor and stagnation factor, and to set the both weight parameters, the formula is as follows:

$$P = d_1 q_1 + d_2 q_2 \tag{6}$$

In the formula, P is the particle's performance index, d_1、d_2 are the weighting parameters of oscillation factor and stagnation factor, this paper set d_1=0.5, d_2 = 2.5.

Above is applicable to determine for each particle, here, you can set a threshold value ε, according to the formula (6) to measure the state of the current particle, if it does not meet the requirements, then update the particle's velocity to change the unstable state of the particle.

3.3 Improved Particle

Based on the part particles exists instability in the flight, this paper proposes an adaptive mutation optimize operation. Determine the state of the current particle, and calculate the corresponding fitness value, to mutation, stimulate the unstable particles falling into local optimum to re-search, mutation need to consider the variance value of the fitness.

$$\rho_G = (\rho_M - \rho_m)\frac{\nu}{Q} + (\rho_m - \rho_M)\frac{2\nu}{Q} + \rho_M \tag{7}$$

In the formula, ρ_G is global extreme mutation probability; ρ_M is maximum mutation probability; ρ_m is minimal mutation probability, ν is fitness variance. In the formula (7), ν is smaller, the closer the distance of the particles, the greater the probability of global extremum *Gb* mutation. Conversely, ν is farther, namely the population diversity is strong, the smaller the probability of global extremum *Gb* mutation. Algorithms determine steady state of the particle, then the particle adaptive mutation operation get rid of the current state of the particle.

3.4 Clustering Encoding

Set the vector set of samples $\{X \mid x_1, x_2, \cdots, x_n\}$ in p-dimensional space, Clustering problem boils down to find a group division $\{C \mid c_1, c_2, \cdots, c_m\}$ so that the value of the criterion function in the formula (2) is the smallest. Particle swarm optimization coding around the cluster center, is a real-coded, the position of the particle not only consider cluster center, but also consider the speed and the corresponding fitness value, together, particle coded as: $\{z_{i1}, z_{i2}, \cdots, z_{in}, v_{i1}, v_{i2}, \cdots, v_{in}, f(X_i)\}$.

When the algorithm clustering, first to determine the clustering center, then one by one search the closer point away from the cluster center closer, get the distance from the point to be clustering and cluster centers, making it meet $\min(X_i - z_{ij}) \mid j \in [1, n]$. Algorithm in iterative process, the position and velocity of the particles are constantly change, fitness should also be updated, and the calculation steps are as follows:

(1) Calculate the cluster center, and get the cluster division of all points;

(2) calculate the criterion function value in the function (2);

(3) The individual fitness value is negatively related to the criterion function, set $f(X) = \frac{\lambda}{J}$, in the formula, λ is a fitness factor, criterion function value is smaller, the greater the individual fitness value.

3.5 Algorithm Step

Taking into account the problem that the traditional K-means clustering sensitive to the initial value and slow convergence speed, in this paper combines improved K-mean and PSO algorithm, proposed a new improved algorithm IPSO-KS, at the same time, improve

the algorithm, Solve the problem for the distance deviation of concave and convex point, algorithm steps are as follows:

(1) Sample set initialization, determine the population size, initialize speed and position of particle, and specify the cluster center of each particle, set the initial local and global optimization;

(2) Use of the fitness function to calculate the fitness of each particle, compare the experienced position of the particle during the flight, get local and global extremes;

(3) determine the state of the current particle according to the formula (5), (6) in iteration, and according to the formula (3), (4), (7) to adjust the state of particle;

(4) Using the modified K-means algorithm is optimized new generation particle in iteration;

(5) Judgment the termination condition of algorithm, if satisfied, then terminate; otherwise, return to step (2) to continue the calculation.

For the improved K-means algorithm, should be to implement clustering based on the encoding rules of the particle cluster center, constantly updated fitness value of the individual in the iteration, and replace the coding of the previous iteration. In addition, the clustering process will occasionally appear empty cluster, if appear, then traverse the furthest individual away from cluster center in the other cluster, will be divided into the cluster, the last iteration must meet no empty cluster.

4. Experimental Analysis

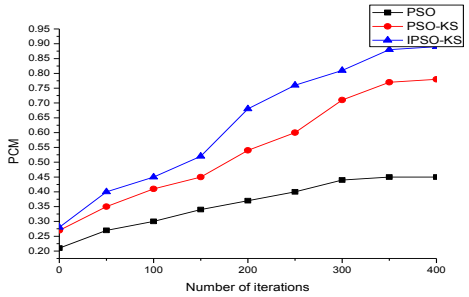
Experimental basis in a manner that is divided into two parts, the clustering effect, convergence and optimization capabilities. The former use data set of common database UCI, Iris, Wine and Similar, Iris is a low-dimensional data sets, dimension is 4, can be divided into three categories, the number of samples in each class is 50; Wine is a low-dimensional data sets, dimension is 13, can be divided into three categories, the number of samples in each class is 188; Similar high-dimensional data sets, dimension is 16 090, can be divided into three categories, the number of samples in each class is 288. Study on the clustering effect, adopt the pairwise comprehensive measure (PCM), it combines the accuracy and recall rate, the range set $[0,1]$, the larger the value, the better the clustering effect. Convergence will test the stability of the algorithm, the more gentle the better. For optimization ability to test using two benchmark functions, namely *Rosenbrock* and *Griewank* . The process of testing is random, each group is 100 times, the number of maximum iteration is 400, respectively test three data sets corresponding the values of benchmark function in the optimization tests, algorithm termination condition is the optimal solution smaller than -0.9999. Experiments compare the effects of three algorithms: traditional particle swarm algorithm, denoted PSO; the simple hybrid algorithm of K-means and PSO, denoted PSO-KS; this paper proposed the improved hybrid algorithm based on particle state determine denoted IPSO-KS. The following is the relevant formulas and the experimental results, figure 2 is a comparison of the respective data sets PCM values, figure 3 is a comparison of the algorithm's convergence, table 1 is the optimal value of the three benchmark functions.

(1) The range of the *Rosenbrock* function is $[-10,10]^D$, formula is as follows:

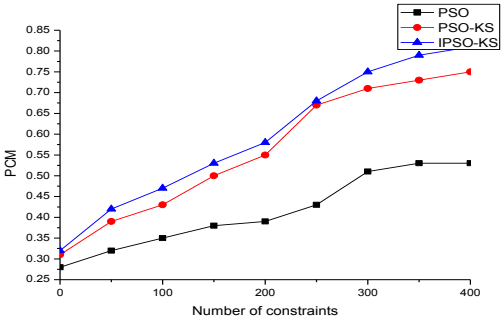
$$Rosenbrock(x) = \sum_{i=1}^{D} (100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2)$$

(2) The range of the *Griewank* function is $[-600,600]^D$, formula is as follows:

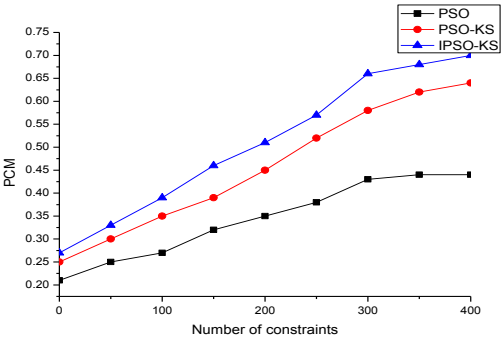
$$Griewank = \frac{1}{4000} \sum_{i=1}^{D} (x_i)^2 - \prod_{i=1}^{D} \cos(x_i / \sqrt{i}) + 1$$



(a) PCM Value Of The Data Set Iris



(B) PCM Value of the Data Set Wine



(C) PCM Value of the Data Set Similar

Figure 2. The Comparison about PCM Values of the Data Sets

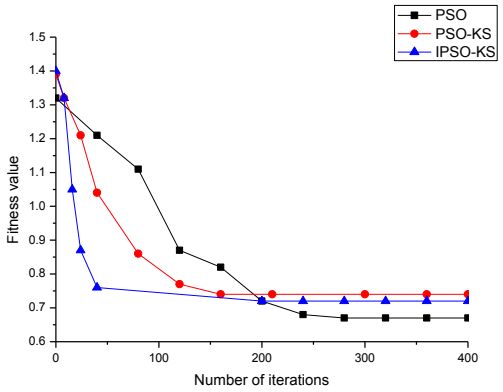


Figure 3. The Comparison about the Comprehensive Convergence of the Data Sets

Table 1. The Number of *Rosenbrock* Function about Successful Search

the data sets	Dimensions	PSO	PSO-KS	IPSO-KS
Iris	4	51	85	95
Wine	13	48	83	94
Similar	16090	47	81	92

Table 2. The Number of *Griewank* Function about Successful Search

the data sets	imensions	PSO	PSO-KS	IPSO-KS
Iris	4	54	87	96
Wine	13	50	84	95
Similar	16090	51	85	95

The experiment respectively determine the clustering effect, the convergence and optimization ability of algorithm. Among them, Figure 2 is the comparison of the PCM value, the results showed that the comparison of three algorithms in the three group's dataset, IPSO-KS has obvious advantages. When the number of iterations is zero, because no room for improvement, though PSO slightly different than the other two algorithms, but no big difference overall. With the increase of iteration, IPSO-KS and PSO-KS advantage gradually reflected, when the number of iterations is 150, the test values of two algorithms have increased significantly; and because IPSO-KS algorithm for particle performance is optimized, when the number of iterations to reach 300, the rise speed has accelerated signs, improved results will be reflected.

Figure 3 compares the convergence of three algorithm, the results show that the speed and stability of convergence about the improved algorithms IPSO-KS are better than the other two algorithms. When the number of iterations is 100, IPSO-KS basically complete convergence, while the worst number of iterations for PSO is 200, begin converge. Here, the particles after the K-means algorithm to optimize convergence is enhanced, meanwhile, the determination of the particle's state in the flight and corresponding mutation makes clustering effect of the algorithm that is more in line with expectations.

Table 1 and Table 2, using the benchmark function analysis the optimization ability of three algorithms, the comprehensive results of the two functions show that IPSO-KS optimization ability significantly stronger than the other two algorithms, in comparison, the gap of PSO-KS and IPSO-KS is smaller, the embodiment of the improved effect benefited from the K-means effectively combination with the PSO, cluster centers to be redefined and enhanced particle activity, making the algorithm performance is improved.

5. Conclusion

This paper proposed a hybrid optimization algorithm based on the semi-supervised clustering idea that K-means algorithm effectively combined with PSO algorithm. Among them, the K - means algorithm is used to calculate the particle clustering center, improves the performance of the new particle is stronger, at the same time to determine particle flight condition in the algorithm, and through the mutation operation to improve the performance of the particle, not only solve the problem of slow convergence speed, and ensure stability and the clustering effect of the algorithm. Experimental results show that the improved IPSO - KS algorithm is simple and efficient and in the cluster effect, the convergence and optimization performance are very outstanding ability.

The hybrid algorithm used in this paper, for the treatment of isolated points in the calculation of the clustering center leaves a lot to be desired, which requires further study follow-up work.

References

[1] E. Murat, C. Nazif and S. Sadullah, "A new algorithm for initial cluster centers in k-means algorithm [J]", Pattern Recognition Letters, vol. 32, no. 14, (**2011**), pp. 1701-1705.
[2] S. Faußer and F. Schwenker, "Semi-Supervised kernel clustering with sample-to-cluster weights[C]", PSL'11 Proceedings of the First IAPR TC3 conference on Partially Supervised Learning, (**2011**), pp. 72-81.
[3] M. S. Baghshah and S. B. Shouraki, "Metric learning for semi-supervised clustering using pairwise constraints and the geometrical structure of data [J]", Intelligent Data Analysis, vol. 13, no. 6, (**2009**), pp. 887-899.
[4] S. Kiranyaz, J. Pulkkinen and M. Gabbouj, "Multi-dimensional particle swarm optimization in dynamic environments [J]", Expert Systems with Applications, vol. 38, no. 3, (**2011**), pp. 2212-2223.
[5] A. M. Bagirov, "Modified global k-means algorithm for minimum sum-of-squares clustering problems [J]", Pattern Recognition, vol. 41, no. 10, (**2008**), pp. 3192-3199.
[6] M. Abdeyazdan, "Data clustering based on hybrid K-harmonic means and modifier imperialist competitive [J]", The Journal of Supercomputing, vol. 68, no. 2, (**2014**), pp. 574-598.
[7] M. E. Celebi, H. A. Kingravi and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm [J]", Expert Systems with Applications, vol. 40, no. 1, (**2013**), pp. 200-210.
[8] I. Montalvo, J. Izquierdo, R. Perez-Garcia and M. Herrera, "Improved performance of PSO with self-adaptive parameters for computing the optimal design of Water Supply Systems [J]", Engineering Applications of Artificial Intelligence, vol. 23, no. 5, (**2010**), pp. 727-735.
[9] H. Modares, A. Alfi and M.-B. Naghibi Sistani, "Parameter estimation of bilinear systems based on an adaptive particle swarm optimization [J]", Engineering Applications of Artificial Intelligence, vol. 23, no. 7, (**2010**), pp. 1105-1111.
[10] A. Elhossini, S. Areibi and R. Dony, "Strength pareto particle swarm optimization and hybrid ea-pso for multi-objective optimization [J]', Evolutionary Computation, vol. 18, no. 1, (**2010**), pp. 127-156.
[11] I. L. Schoeman and A. P. Engelbrecht, "A novel particle swarm niching technique based on extensive vector operations [J]", Natural Computing, vol. 9, no. 3, (**2010**), pp. 683-701.
[12] I. X. Tassopoulos and G. N. Beligiannis, "A hybrid particle swarm optimization based algorithm for high school timetabling problems [J]", Applied Soft Computing, vol. 12, no. 11, (**2012**), pp. 3472-3489.
[13] J. L. Fernandez-Martinez and E. Garcia-Gonzalo, "Stochastic Stability Analysis of the Linear Continuous and Discrete PSO Models [J]", IEEE Transactions on Evolutionary Computation, vol. 15, no. 3, (**2011**), pp. 405-423.
[14] S. N. Sivanandam and P. Visalakshi, "Dynamic task scheduling with load balancing using parallel orthogonal particle swarm optimization [J]", International Journal of Bio-Inspired Computation, vol. 1, no. 4, (**2009**), pp. 276-286.

Authors

Y. R. Li, he received the Master's degree in computer application technology from University of Science and Technology Liaoning, in 2008. Currently, he is a lecturer at School of applied technology college at University of Science and Technology Liaoning. His research interests include Distributed computing and data mining.

C. N. Zhang, she received the Master's degree in computer application technology from University of Science and Technology Liaoning, in 2007. Currently, she is a lecturer at School of Software Engineering at University of Science and Technology Liaoning. Her research interests include Distributed computing and data mining.