Clustering Analysis Based on Chaos Immune Algorithm

Wu Yanbin
School of Management Science and Engineering, Hebei University of Economic and Business
Shijiazhuang, China, telp: 86-0311-87655561
e-mail: wuyanbin080@163.com

Abstract

To improve the accuracy of clustering classification, the Chaos Immune Algorithm was proposed. In this algorithm, the ergodic property of chaos phenomenon is used to optimize the initial population, so it can accelerate the convergence of Immune Algorithms. Chaotic systems are sensitive to initial condition system parameters. Through the clone selection operator, antibody circulation and supplement, Clone operator and excellent individual chaotic disturbance, local optimums were avoided, so the global optimization was obtained. As for this issue, chaotic immune algorithm, ALECO-2, BPMA and BP algorithm were applied to test the algorithm performance in two group experiments. Theory and experiment showed that the Chaos Immune Algorithm can get global optimum clustering center, and greatly improve the amplitude of operation.

Keywords: *chaos, immune algorithm, cluster classification*

1. Introduction

Clustering Analysis is a non-supervisory pattern recognition method[1,2]. Data is grouped by clustering. And each group of data thus generated is termed as a cluster. Each data in a cluster is named as an object. The purpose of clustering is to make the objects from the same cluster to resemble each other in characteristic as much as possible, while objects from different clusters to differentiate each other in characteristic as much as possible. The task of clustering is to divide an unmarked mode into some subclasses according to certain rules. It is required that analogous samples be classified into the same group while non-analogous samples into different groups. Therefore, it is termed as non-supervisory classification. Currently, different methods of clustering have been applied into many fields like data mining, pattern recognition, image processing, Laser Radar targets detection, and remote sensing technique [3-6].

Overview of cluster analysis as follows:

The clustering can be simply described as follows: if classify n vectors $X_p\left(p=1,2,\cdots,n\right)$ into c categories $G_j\left(j=1,2,\cdots,c\right)$, and the clustering center of each categories should be c_j. This pattern consists of the following steps:

(1) Select appropriately the initial center $G_1^0,G_2^0,\cdots,G_c^0$ of c categories.

(2) In k iterations, if we adjust any sample vector X_p, into any category of c categories, as for all $i\neq j,i,j\in\left[1,c\right]$, if $\left\|X_p-G_j^k\right\|<\left\|X_p-G_i^k\right\|$, then $X_p\in S_j^k$, of which, S_j^k is the category with G_j^k as the center.

(3) Recalculate the new center G_j^{k+1} : $G_j^{k+1}=\frac{1}{N_j}\sum_{X_p\in S_j^k}X_p$ of category S_j^k that we got from step 2, N_j in the formula is the number of samples in category S_j^k, the condition for ending the iteration is $i\neq j,\ j=1,2,\cdots,c\ G_j^k=G_j^{k+1}$.

Received August 16, 2013; Revised October 6, 2013; Accepted October 14, 2013

As is shown, the clustering is an issue of optimization of grouping in nature. And the selection of the initial center influences the result of clustering greatly. Optimization of grouping can be solved by immune algorithm, and chaotic optimization arithmetic operators are introduced into immune algorithm. The ergodic property of chaos phenomenon is used to optimize the quality of the initial population and improve the efficiency of calculation.

2. Research Method

The core of Chaos Immune Algorithm is based on the biologic immune mechanism reflected in mathematics. Combining with random choices and determinacies, the Artificial Immune Algorithm is a heuristic random searching algorithm which has the ability to develop. The chaotic immune algorithm is indicated as Figure 1.

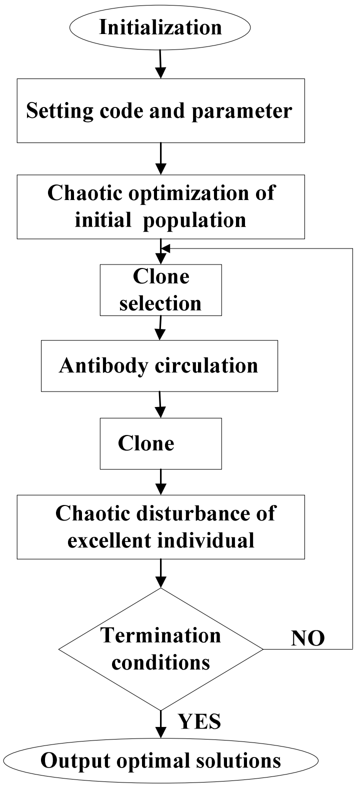
Initialization
Setting code and parameter
Chaotic optimization of initial population
Clone selection
Antibody circulation
Clone
Chaotic disturbance of excellent individual
Termination conditions
NO
YES
Output optimal solutions

Figure 1. The process of Chaos Immune Algorithm

Major operators of Chaotic Immune Algorithm:

(1) Chaotic optimization of the initial population

Ag was defined initial antigen and Ab were defined initial antibody populations. M was represented the scale of antibody populations.

$X_m = \{x_{m1}, x_{m2}, \cdots, x_{mi}, \cdots, x_{mn}\}$ was represented an initial antibody, $m = 1, 2, \cdots, M$ $i = 1, 2, \cdots, n$. n is the dimension of variable X_m . Similar with genetic algorithm, x_{mi} was called allele, $x_{mi} \in [a_{mi}, b_{mi}]$. Antibody bit string was divided into l segments and every segment length was n_i , so, $n = \sum_{i=1}^{l} n_i$ was represented total length of antibody gene segment. $Y(0) = (y_{0,1}, y_{0,2}, \cdots, y_{0,i})$ ($i = 1, 2, \cdots, n$) was defined initial antibody populations center.

After k iterations, the euclidean distance reciprocal between individual and antibody populations center was defined by affinity function:

$$\frac{1}{aff\left(X_m\right)}=\left(\sum_{i=1}^{n}\left|x_{mi}-y_{ki}\right|^2\right)^{1/2} \quad m=1,2,\cdots,M \quad i=1,2,\cdots,n \tag{1}$$

Main operators as follows:
1) Chaos optimization of initial population
The initial Ab were optimized through chaos operator.
2) The logistic map was used to generate chaotic variable γ_i^k.

$$\gamma_i^{k+1}=\mu\gamma_i^k\left(1-\gamma_i^k\right) \quad i=1,2,\cdots,n \quad k=0,1,2,\cdots \tag{2}$$

Where μ is the control parameter, after determined the value of μ, with the arbitrary initial value $\gamma_i^0\in\left(0,1\right)$ (except 0.25, 0.5, 0.75 the fixed point of equation(2)), an assured time series $\gamma_i^1,\gamma_i^2,\cdots\gamma_i^k$ can be iterated [7].

A chaotic-type initial value γ_{mi}^0 was produced by Equation (3).

$$\gamma_{mi}^0=\left(x_{mi}-a_{mi}\right)/\left(b_{mi}-a_{mi}\right) \quad m=1,2,\cdots,M \quad i=1,2,\cdots,n \tag{3}$$

Through the inherent properties of chaotic variables to fulfill the overall search,Chaos optimization algorithm can map the chaos space to the solution space. According to Equation (4), the chaos variables were maped from chaos space to the solution space.

$$x_{mi}=a_{mi}+\left(b_{mi}-a_{mi}\right)\gamma_{mi}^k \quad m=1,2,\cdots,M \quad i=1,2,\cdots,n \tag{4}$$

x_{mi}^* was set as the optimal solution at the current phase of coarse-grained search, aff^* was the optimal objective function value for current phase. After each iteration, the individual affinity function $aff\left(k\right)$ was calculated, iteration termination condition was $aff\left(k\right)\le aff^*$. The antibody population after optimization was denoted by Ab'.

(2) Clone selection operator

After the combination of antibody with antigen, antigen can be destroyed through a series of reactions which are based on antibody concentration. Antibody concentration $d\left(X_m\right)$ was defined by the following equation [8].

$$d\left(X_m\right)=\left[\frac{1}{M}\sum_{q=1}^{M}\frac{1}{1+\sqrt{\sum_{i=1}^{n}\left(x_{mi}-x_{qi}\right)}}\right]^{\alpha\cdot\left(1-\frac{g}{G}\right)} \tag{5}$$

Where α was system parameter to adjust algorithm convergence speed, g was evolution generation and G was the maximum evolution generation.

Affinity function $aff\left(X_m\right)$ was adjusted through the following equation. After the adjustment, $aff\left(X_m\right)$ was denoted by $\tilde{aff}\left(X_m\right)$.

$$\tilde{aff}\left(X_m\right)=aff\left(X_m\right)/d\left(X_m\right) \tag{6}$$

These antibodies which have bigger individual affinity value and the lower concentration can be promoted, on the contrary, those antibodies with smaller individual affinity value and the higher concentration can be inhibited, thus this process ensured the diversity of antibody group, so as to escape from local optimums.

(2) Antibody circulation and supplement

The antibody circulation and supplement mechanism of biological immune system was simulated in order to ensure the diversity of antibody group and realize global search. Each time before antibody group Ab' were cloned, M_s highest affinity antibodies selected in a random generation antibody group Ab_r with population size of M_r were used to replace M_s lowest affinity antibodies in Ab'.

(4)Clone operator

According to $\tilde{aff}$, antibodies in Ab' were ordered by sort descending. Take top m_c antibodies to be cloned, the antibody group cloned were denoted by Ab_c. As the following equation, the population size of Ab_c can be calculated.

$$M_c=\sum_{j=1}^{m_c} round\left(\frac{\beta M}{j}\right) \tag{7}$$

Where M_c was the population size of Ab_c. β was proliferation coefficient to control antibody group size owing to its influence of algorithm iteration and calculating time. j was the sequence number of antibody by sort descending. $round(\cdot)$ was rounding operation.

(5) Variation Operator

$\bar{X}=\left(\bar{x}_1,\cdots,\bar{x}_{i-1},\bar{x}_i,\cdots,\bar{x}_n\right)$ was set as a parent entity. According to Equation (8), after variation, the offspring individual was $\tilde{X}_m$, the antibody group mutated was denoted by Ab_m.

$$\tilde{X}_m=\bar{X}_m+\eta N\left(0,1\right)e^{-\hat{aff}} \tag{8}$$

After $(0,1)$ standardization, aff of $\bar{X}_m$ was denoted by $\hat{aff}$. $N(0,1)$ is normal random function with mean value $\mu=0$ and variance $\sigma=1$. The proportionality constant η can control attenuation of negative exponential function. According to Equation (8), the bigger the antibody affinity value, the smaller variation, that was beneficial to maintain the stability of the local optimal solution.

(6) Excellent individual chaotic disturbance

After the clone selection, clone and variation, the current optimal solution was $X^*=\left(x_1^*,\cdots,x_{i-1}^*,x_i^*,\cdots,x_n^*\right)$. According to Equation (9), chaotic disturbances to x_i^* was made.

$$x_i\left(k'\right)=x_i^*+\varphi u_{k,i} \quad i=1,2,\cdots,n \tag{9}$$

In the equation, $x_i(k')$ was the chaotic variable of a smaller range of ergodicity relative to Equation (9), φ was the adjustment coefficient related to the iteration number k'. in this paper φ was set by $\varphi=1-\left(\frac{k'-1}{k'}\right)^n$. Continue to carry out iterative search with $x_i(k')$, the termination condition was $aff\left(k'\right)\le aff^*$.

3. Results and Analysis

As for this issue, chaotic immune algorithm, ALECO-2, BPMA and BP algorithm were applied to test the algorithm performance in two group of experiments.

3.1. Tset 1

By inputting a 4 that is in normal distribution into 2 to carry out the simulated experiment. There are 20 samples, and the samples' vectors of average value is $\begin{bmatrix}-1.5 & -1.5 & -1.5 & -1.5\end{bmatrix}$ and $\begin{bmatrix}2.2 & 2.2 & 2.2 & 2.2\end{bmatrix}$, the sample's covariance matrix is $\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & 3 & 0\\ 0 & 0 & 0 & 3\end{bmatrix}$ and $\begin{bmatrix}5 & 0 & 0 & 0\\ 0 & 5 & 0 & 0\\ 0 & 0 & 5 & 0\\ 0 & 0 & 0 & 5\end{bmatrix}$, there are overlaps among these samples [9].

As for this issue, the chaotic immune algorithm, ALECO-2, BPMA and BP algorithm were applied. On the basis of large number of simulated operations, the optimum studying parameter for each algorithm were found. With the optimum parameter, the comparative curve of convergence of each algorithm were shown in Figure 2. the average value and standard balance of the times of iteration when converge after operating for 10 times at random were shown in Table 1. CIA was tested in another group with 20 samples for each patter, the rate of correct identification is 95.5%.

Table 1. The Statistics of Convergence of Different Algorthm

Algorithm	CIA	ALECO-2	BPMA	BP
Average Iteration Times	24.1	58.2	60.14	128.2
Standard Balance of Iteration Times	10.17	22.27	20.03	47.45

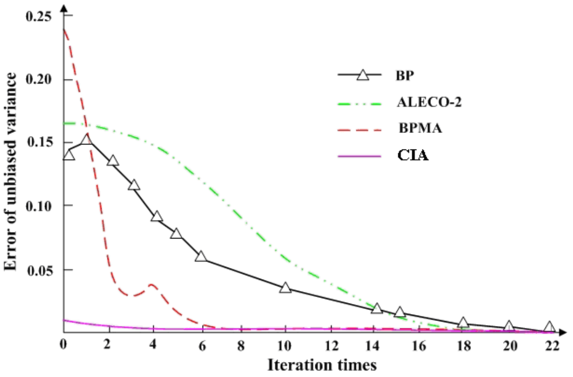

Figure 2. Comparative Curve of Convergence

3.2. Test 2

The market competitor was divided by marketing theory into 4 types. They are market Leader, market challenger, market follower and market filler. After the repeated research and comprehensive analysis of the 4 types of market competitors five characteristic factors can be extracted: (1) x_1 is the relative market share; 2) x_2 is the level of price changes for the enterprise; (3) x_3 is the capability for enterprise in new product development; 4) x_4 is the capability for enterprise distribution channels and physical distribution; 5) x_5 is the capability for a comprehensive marketing. Various characteristic factors and the corresponding grade of the actual meaning and values have been further divided, as shown in Table 2.

Table 2. Featured Factors of Enterprise Competitive Position

	Featured factor	Grade	Real meaning	Corresponding value
x_1	Relative market share		The comparison of market share in the largest market competitors	0.1-10
x_2	The level of price change	1	High	9-10
		2	Medium	8-8.9
		3	General	6-7.9
		4	Weak	0-5.9
x_3	capability of new product development	1	High	9-10
		2	Medium	8-8.9
		3	General	6-7.9
		4	Weak	0-5.9
x_4	Capability of saling channels and physical distribution	1	High	9-10
		2	Medium	8-8.9
		3	General	6-7.9
		4	Weak	0-5.9
x_5	Capability of comprehensive marketing	1	High	9-10
		2	Medium	8-8.9
		3	General	6-7.9
		4	Weak	0-5.9

Table 3. Accuracy of Different Algorthms

Algorithm		Market leader	Market challenger	Market follower	Market filler	Average correct rate
BP	Iterations	90	2000	2200	200	
	Correct rate	76.10%	72.38%	71.77%	73.34%	73.39%
BPMA	Iterations	82	1900	2100	194	
	Correct rate	80.30%	80.42%	81.75%	83.55%	82.39%
ALECO-2	Iterations	77	1800	2040	183	
	Correct rate	86.10%	82.38%	81.77%	83.34%	83.39%
CIA	Iterations	70	1788	2000	175	
	Correct rate	87.10%	89.08%	88.98%	88.75%	88.54%

280 of the garment industry manufacturers were randomly divided into the training sample set (140 samples) and test sample set (140 samples), four characteristic factor values were calculated in each sample. ALECO-2, BPMA and BP algorithm were applied to do a classified comparative test. The results were shown in Table 3.

4. Conclusion

Focusing on data clustering, this article tries to combine chaotic operator with immune algorithm. This chaotic immune algorithm integrates the advantages of both chaotic operator and clone operators. The outputs of theory and simulation both indicate that the chaotic immune algorithm can ensure the global optimum clustering and improve the amplitude of operation greatly.

Acknowledgment

This paper is part of the research results of Youth Fund of Science and Technology Research of Hebei Higher School (No: 2010236).

References

[1] Cheung YM. K Means. A New Generalized k-Means Clustering Algorithm. *Pattern Recognition Letters.* 2003; 24(15): 2883-2893.

[2] Kanade M, Hall O. Fuzzy ants and clustering. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.* 2007; 37(5): 758-769.

[3] Hung CC, Kulkarni S, Kuo BC. A New Weighted Fuzzy C-means Clustering Algorithm for Remotely Sensed Image Classification. *IEEE Journal of Selected Topics in Signal Processing*. 2011; 5(3): 543-553.

[4] Tseng VS, Kao CP. Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method. *IEEE/ACM Trans. On Computational Biology and Bioinformatics*. 2005; 2(4): 355-365.

[5] Farnaz F, Ebrahim P, Shahram T. Retinal Identification Based on Density Clustering and Fuzzy Logic and Connecting This information to Electronic Health Record. *International Journal of Electrical and Computer Engineering*. 2013; 3(3): 359-365.

[6] Cuncun W. Intelligent Search Technology Combining Semantic Grid and Clustering. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11(8): 4803-4809.

[7] Alatas B, Akin E, Ozer AB. Chaos Embedded Particle Swarm optimization Algorithms. *Chaos Solitons Fractals*. 2009; 40(4): 1715-1734.

[8] Riccardo C. Chaotic Sequences to Improve the Performance of Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation.* 2003; 7(3): 289-304.

[9] Leandro DC, Fernando VZ. Learning and Optimization Using the Clonal Selection Principle. *IEEE Transactions on Evolutionary Computation.* 2002; 6(3): 239-251.

[10] CCarlotta D, Dimitrios G, Sheng M. Locally Adaptive Metrics for Clustering High Dimensional Data. *Data Mining and Knowledge Discover*. 2007; 14(1): 63-97.