Union Coding Based Immune Clone Selection Unsupervised Clustering Algorithm

Lichao Mou
automation school
xi'an university of posts & telecommunications

Xiaoying Pan
School of Computer Science & Technology
Xi'an University of Posts & Telecommunications

Abstract-**A new union based coding based immune clone selection clustering algorithm is proposed in this paper. The algorithm designs a coding method which synthesizes number of clusters and cluster centers, without first specifying the number of clusters, effectively overcoming the dependency on the domain knowledge. At the same time, generate clustering results by immune clone operator, and get the final results by the lopping operation of minimum spanning tree. Experimental results show that the methods can cluster for different types of data sets properly and can determine the appropriate number of clusters.**

Keywords-Union coding, Immune, Unsupervised clustering, Minimum spanning tree

I. INTRODUCTION

Cluster analysis is one of the methods of multivariate statistical analysis, and also is an important branch of the statistical pattern recognition. Existing clustering algorithms can be divided into the following categories: partition clustering algorithms, hierarchical clustering algorithms, density and grid based clustering algorithms and other clustering algorithms. Partition clustering algorithms gradually reduce the error value of the objective function by iterative calculation, when the value of the objective function converges, the algorithms will get the final clustering results. Hierarchical clustering algorithms uses the data concatenate rule to split or polymerize data set repeated, and finally form a kind of hierarchical sequence clustering problem. Density-based clustering algorithms are devised to discover arbitrary shaped clusters, in such algorithms; a cluster is regarded as a region in which the density of data objects exceeds a threshold. In fact, cluster analysis is an unsupervised classification, it has no domain knowledge available, and most clustering algorithms rely on domain knowledge.

This paper put forward a Union Coding based Immune Clone Selection Unsupervised Clustering Algorithm. The algorithm individually designed a new encoding scheme, immune genetic manipulation based on single point variation, annexation operator and the optimization scheme of clustering result based on the principle of artificial immune network.

II. ARTIFICIAL IMMUNE SYSTEM

Castro, Kim and Du have successively put forward the clone selection algorithms inspired by the clone selection theory. Clone selection operator is the most important operator of clone selection algorithms, the operator consists of three basic operations: clone operation, immune genetic operation and clone selection operation. Under clone selection operator, the evolution of antibody population can be expressed as the following stochastic process.

$$CSA: A(k) \xrightarrow{T_c^C} Y(k) \xrightarrow{T_g^C} Z(k) UA(k) \xrightarrow{T_s^C} A(k+1)$$

In 1974, Jerne put forward the idiotypic network regulation theory. The theory considers any antibodies or antigen receptors on lymphocytes exist idiotypic determinants, they can be recognized by the other lymphocytes and produce anti-idiotypic antibodies. Through the idiotypic and anti-idiotypic recognize each other, stimulate each other, and condition each other, immune system forms a network.

III. THE NEW PROPOSED ALGORITHM

In real application, especially in the case of absence of sufficient domain knowledge, it is difficult to get a good number of clusters. Therefore, this paper will encode number of clusters with clustering result. Then the proposed algorithm searches for the best clustering result, meantime searches the optimal number of cluster.

A. Coding Scheme

For any finite data set $X=\{x_1, x_2, \cdots, x_n\}$, n is the number of data points. $A_i(k) \in \boldsymbol{A}(k)$ is one antibody of the antibody population $\boldsymbol{A}(k)$. Character string $a_1 a_2 \cdots a_n$ is the coding string of antibody $A_i(k)$, a_i is the cluster of ith data point, $1 \le a_i \le K$. For denot the number of cluster, add a gene locus a_0 in the coding string $a_1 a_2 \cdots a_n$ of antibody $A_i(k)$. Now the coding of $A_i(k)$ is $A_i(k) = a_0 a_1 \cdots a_n$, where a_0 is the number of clusters, $a_i (i=1,2,\cdots,n)$ is the cluster of ith data point, $1 \le a_i \le a_0 (i=1,2,\cdots,n)$.

B. Clone Selection Operator

If the antibody $A_i(k) \in \boldsymbol{A}(k)$, the population is $\boldsymbol{A}(k) = [A_1(k)\ A_2(k)\ \cdots\ A_m(k)]^T$. Then clone operation can be described as $Y_i(k) = T_c^C(A_i(k)) = \boldsymbol{I}_i \times A_i(k), i=1,2,\cdots,m$, where $\boldsymbol{I}_i$ is q_i dimension row vector. In this paper, $q_i(k)$ is given by:

$$q_i(k) = Int\left(m_c \bullet \frac{f(A_i(k))}{\sum_{j=1}^{m} f(A_j(k))} \right), i=1,2,\cdots,m$$

Published by Atlantis Press, Paris, France.

where $Int(\bullet)$ is a rounding up function, $m_c > m$ is a given integer related to the clone size, $f(\bullet)$ is a antibody-antigen function.

Based on the characteristics of clustering, we put forward a new immune genetic operation based on single point variation:

$$\boldsymbol{Z}(k) = T_g^C(\boldsymbol{Y}(k)) \quad , \quad p_g(Y_{ij}(k) \to Z_{ij}(k)) = 1 \quad ,$$
$$Z_{ij}(k) = T_g^C(Y_{ij}(k)) = T_g^C(a_{random})$$

where $a_{random} \in Y_{ij}(k)$, $random$ is a random number and $1 \le random \le n$. Immune genetic operation based on single point variation performs better than the one based on random variation for dealing clustering.

Figure 1 is the test results. Figure (a) is the initial data set, (b) is the clustering result based on random variation and (c) is the clustering result based on single point variation. We can see from the figures easily, algorithm based on single point variation can accelerate the rate of convergence significantly.

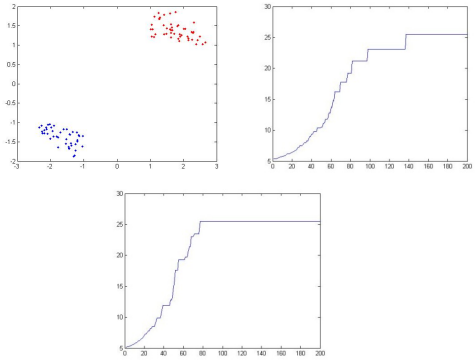


Figure 1. Test results

Suppose $\forall i = 1, 2, \cdots, m$, $B_i(k) = \max\{\boldsymbol{Z}_i(k)\} = \{Z_{ij}(k) | \max f(Z_{ij}(k)), j = 1, 2, \cdots, q_i\}$. Then the probability of $B_i(k)$ replaces the antibody $A_i(k)$ in the aboriginal population is:

$$p_s^k(A_i(k+1) \leftarrow B_i(k)) = \begin{cases} 1, & f(A_i(k)) < f(B_i(k)) \\ \exp\left(-\dfrac{f(A_i(k)) - f(B_i(k))}{a}\right), & f(A_i(k)) \ge f(B_i(k)), A_i(k) \text{ is the best in current population.} \\ 0, & f(A_i(k)) \le f(B_i(k)), A_i(k) \text{ is not the best in current population} \end{cases}$$

where a is a value related to the diversity of antibody population, better diversity, greater value a.

C. Annexation Operator

If $Best(k) = \max\{\boldsymbol{A}(k)\} = \{A_i(k) | \max f(A_i(k))\}$, $Worst(k) = \min\{\boldsymbol{A}(k)\} = \{A_i(k) | \min f(A_i(k))\}$, annexation operator can be described as: If $g\%T == 0$, then $Worst(k) \leftarrow Best(k)$, where g is iterations, T is interval generations. Annexation operator draws on Darwinian theory, can accelerate the rate of algorithm.

D. Construction of Fitness Function

In the case of given number of clusters, the fitness function can be expressed like this:

$$F = \sum_{j=1}^{n}\sum_{i=1}^{k} u_{ij} d_{ij} \Big/ \frac{1}{K}\sum_{i \ne j} \|c_i - c_j\|$$

where $d_{ij} = \|x_j - c_i\|$ is the distance from data point $x_j (j = 1, 2, \cdots, n)$ to cluster center $c_i (i = 1, 2, \cdots, k)$, $c_i \in R^p$; $u_{ij} \in \{0, 1\}$ represents data point x_j belong to cluster i or not. $\|c_i - c_j\|$ is the distance from cluster center c_i to c_j .

E. Artificial Immune Network Optimization

Suppose $c_1, c_2, \cdots, c_k$ are the best cluster centers through clone selection algorithm. We regard cluster centers as the nodes and regard the link between the two cluster centers as the edges, so we can get a totally connected graph G. Then Prime algorithm can generate the minimal spanning tree G1 of graph G.

All nodes of G1, namely that is the cluster centers equal to antibodies of artificial immune network. According to the requirements, conduct several lopping operations for the minimal spanning tree. After lopping operations, the antibodies which still connect each other belong to the same cluster and the disconnected antibodies belong to different clusters. Figure 2 is the example of making sure the clustering result by using the lopping operations.

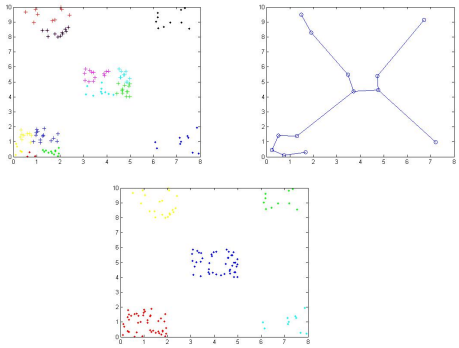


Figure 2. The example of the lopping operations.

IV. Experiment Results

In order to test the performance of the proposed algorithm, we test it on artificial data sets and remote sensing image. And, in the experiment of cluster analysis for remote sensing image, we compare the proposed algorithm with K-means algorithm.

A. Artificial Data Sets

First we test the proposed algorithm on three different artificial data sets, experiment results are as follow. Figure (a) is original data set, (b) is the clustering result generated by clone selection algorithm in the case of not given number of clusters, (c) is the minimum spanning tree, (d) is the most suitable clustering result generated by lopping operator.

These experiments show that the clustering results generated by using the proposed algorithm can be satisfactory.

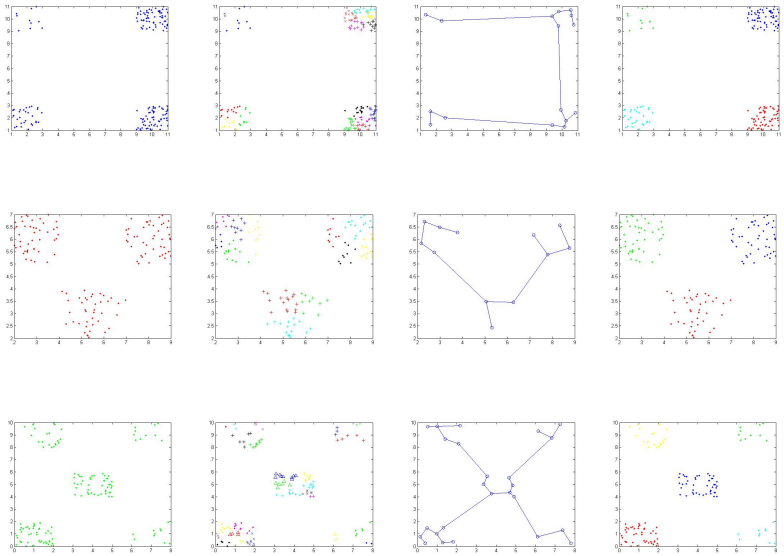
B. *Remote Sensing Image*

In order to further test the performance of the proposed algorithm, we use the remote sensing image to test it and compare with K-means algorithm. Experiment result is as follow. Figure (a) is original remote sensing image, (b) is the result of K-means algorithm and (c) is the result of the proposed algortithm.

Refer to the original remote sensing image (Figure a), it can be seen, the accuracy of clustering result based on the proposed algorithm is higher than the result based on the K-means algorithm. The proposed algorithm can distinguish different land cover well and the boundaries of the various ground objects are visible. Cluster analysis of remote sensing image based on K-means algorithm, its accuracy is not high and the boundaries of the various ground objects are blurred. For example, the water segment shown in the circle in figure a, get a good reflection in figure c, but the corresponding part of the water segment have been a drying phenomenon in figure b. Then comparing the rectangular area, part of the bridge was represented well in figure c, while the corresponding region in figure b do not reflect the bridge structure.

V. CONCLUSIONS

For most clustering algorithms need the number of clusters, this paper designs a coding scheme which synthesizes the number of clusters and cluster centers, meantime introduces the immune clone selection mechanism and artificial immune network idea. We can get the clustering result through immune clone selection algorithm and finally get the most suitable clustering result by using lopping operator. The algorithm is simple and practical and can get suitable result on different types of data sets, so is an effective clustering algorithm.



(a) (b) (c) (d)

Figure 5. The testing results.

ACHNOLGMENT

This work is partially supported by the Natural Science Project by Shaanxi Province Office of Education under Grant 2010JK837, the Natural Science Foundation of xi'an University of posts and telecommunications under Grant 103-0402, 000-1273.

REFERENCE

[1] SUN Ji-Gui, LIU Jie, ZHAO Lian-Yu. New Progress in Clustering Algorithm Research, Journal of Software, Vol.19, No.1, January 2008, pp. 48-61.

[2] Castro L.N.D, Zuben F.J.V. An Evolutionary Network for Data Clustering Proceedings, Sixth Brazilian Symposium on Neural Networks, 2000 Page(s): 84-89.

[3] Kim J, Bentley P.J. Towards an artificial immune system for network intrusion detection: an investigation of clone selection with a negative selection operator. Proceedings of the 2001 Congress on Evolutionary Computation, 2001, 2: 1244-1252.

[4] DU Hai-Feng, JIAO Li-Cheng, WANG S. Clone Operator and Antibody Clone Algorithms. Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002: 506-510.

[5] Maulink U, Bandyopadhyay S. Genetic algorithm-based clustering technique. Pattern Recognition, 2000, 33(9): 1455-1465.

[6] PAN Xiao-Ying, LIUFang, JIAO Li-Cheng. Density Sensitive Based Multi-Agent Evolutionary Clustering Algorithm. Journal of Software, Vol.21, No. 10, October 2010, pp. 2420-2431.

[7] SHI Yun-Song, SHI Yu-Feng. Classification of Remote Sensing Image based On Kernel Fuzzy C-means, Journal of Nan Jing Forestry University (Natural Science Edition), Vol.34, No.6, Nov.2010, pp. 164-166.

[8] HUANG Zhen-Hua, XIANG Yang, ZHANG Bo, WANG Dong, LIU Xiao-Ling. An Efficient Method for K-means Clustering, RP&AI, Vol.23, NO.4, Aug. 2000, pp. 516-521.

[9] MA Wen-Ping, SHANG Rong-Hua, JIAO Li-Cheng. Immune Clonal Optimization Clustering Technique, JOURNAL OF XIDIAN UNIVERSITY, Vol.34, NO.6, Dec. 2007, pp. 911-915.

(a) (b) (c)

Figure 6. The testing results of remote sensing image.

Published by Atlantis Press, Paris, France.