A NOVEL CLUSTERING METHOD WITH NETWORK STRUCTURE BASED ON CLONAL ALGORITHM*

LI Jie, GAO Xinbo, JIAO Licheng

(School of Electronic Engineering, Xidian Univ., Xian 710071, China)

Email: leejie@mail.xidian.edu.cn

Abstract

In the field of cluster analysis, objective function based clustering algorithm is one of widely applied methods. However, this type of algorithms need the priori knowledge about the cluster number and the form of clustering prototypes, which can only process data sets with the same type of prototypes. Moreover, these algorithms are very sensitive to the initialization and easy to get trap into local optima. For the purpose, this paper presents a novel clustering method with network structure based on clonal algorithm to realize the automatization of cluster analysis. By analyzing the neurons of the obtained network with minimal spanning tree, one can easily get the cluster number and the related classification information. The test results with various data sets illustrate that the novel algorithm achieves more effective performance on cluster analyzing the data set with mixed numeric values and categorical values.

1. Introduction

Cluster analysis is one of multivariate statistical analysis method. In traditional cluster analysis methods, the objective function based clustering algorithms convert the cluster analysis into an optimization problem. Due to having profound functional foundation, the study of this type of algorithms becomes the main topic of cluster analysis, in which *c*-means or *k*-means algorithm is one of representative algorithms [1]. However, *c*-means algorithm cannot detect clusters in nonlinear subspaces. To this end, the clustering prototypes are generalized from points to lines, planes, shells and conics, and some *c*-means type algorithms are proposed with various prototypes, such as *c*-lines, *c*-planes, *c*-shells and *c*-conics algorithms [2,3,4]. Therefore, these *c*-means type algorithms can perform cluster analysis on data sets with different prototypes.

Although the above *c*-means type algorithms extend the application range of objective function based clustering algorithms, it costs at the increasing of requirements of clustering priori information, otherwise these algorithms will be misled and result in a wrong partition of data set [1]. Meanwhile, these algorithms require that all the prototypes should be with the same type. Such requirement limits their practical application of the objective function based clustering algorithms. In many fields, the data set to be analyzed often contains unknown number of subsets with different prototypes. Unfortunately, the number of clusters is difficult to automatically determine, especially in the high-dimensional feature space.

Recently, a fuzzy clustering algorithm with multi-type prototypes is proposed [5], which integrates the available prototype-based clustering algorithms together. Since the type of clustering prototypes in data set is not specified, the algorithm has to automatically analyze and obtain the parameters of candidate prototypes with switch regression fashion. Therefore, the performance of the proposed clustering algorithm almost completely depends on the effect of the prototype initialization. It is evident that the objective function of prototype-based clustering algorithm possesses many local optima, so it is easy to get trap into the local optima from improper initialization and results in dissatisfied partition [6,7]. For this purpose, with the advent of the genetic algorithm (GA), some GA-based clustering algorithms were proposed, which can converge into the global optima with a high probability. However, GA-based methods converge very slowly and easily occur premature phenomenon [8].

For automatic determining the cluster number *c*, some researchers present self-organized feature mapping (SOFM) to realize the cluster analysis [9]. Although the SOFM can resolve the problem of determining the cluster number, it does not work well for analyzing the data set with multi-type prototypes.

Cluster analysis is well known as an unsupervised learning method. However, the above clustering algorithms need priori information, which of course affects their unsupervised nature. With the rise of the artificial immune system, an artificial immune network is presented to realize the real unsupervised cluster analysis [10]. This new method can obtain good performance for the data set with distinct boundaries among the subsets. For the data set with indistinct boundaries among subsets, it is difficult to achieve the effective network structures.

As an effective tool of data mining, cluster analysis often needs to process large high-dimensional data set. Moreover, such data set maybe consists of the numeric and

* This work was supportedby a grant from the National Natural Science Foundation of China

the categorical attributes. Few algorithms can handle both data types well, such as *k*-prototypes and etc [11]. However, such method also asks the priori information about cluster number and clustering prototypes.

To overcome the drawbacks of the existing clustering algorithms, a novel algorithm with network structure is proposed based on the clonal algorithm. It can perform cluster analysis with multi-type prototypes. Moreover, it can automatically determine the proper cluster number and cluster structures. By defining a new clonal operator, the forbidden clone, and combining it with the clonal selection algorithm, the proposed clustering algorithm can converge to the global optima at high convergence rate and with a higher probability. The forbidden clone operator makes the clustering network creating the characteristics of patience, which reduces the effect of indistinct boundaries on classification and result in a good clustering performance. In addition, by defining a new distance measurement function for samples with different attributes, the proposed algorithm can also analyze the data set with the mixed attributes.

The rest of this paper is organized as follows. The clustering algorithm based on the evolutionary immune networks is introduced in next section. Section 3 presents the novel clustering algorithm with network based on the clone algorithm. In section 4, the experimental results are given with comparison of performance. The final section is the concluding remarks and the future works.

2. Clustering Algorithm based on Evolutionary Immune Networks

It is Jerne who proposed the immune network theory in 1974 [12]. Based on this immune network theory, in 2000, Leandro presented an evolutionary artificial immune network (EAIN) [10]。The main idea of EAIN is as follows. Let $X=\{x_1,x_2,\cdots,x_n\}$ denote a set of *n* objects, where each object $x_i=[x_{i1},x_{i2},\cdots,x_{im}]^T$ is described by m attributes, to characterize a molecular configuration as a point $s\in S^m$. The possible interactions within the system will be represented in the form of a connectivity graph. The network model can be defined as follows.

Definition 1: The evolutionary artificial immune network is an *edge-weighted graph*, not necessarily fully connected, composed of a set of nodes, called *cells*, and sets of node pairs called *edges* with a number assigned called *weight*, or *connection strength*, specified to each connected edge.

The clusters in the network will serve as *internal images* (*mirrors*) for mapping existing clusters in the data set into existing clusters in the network of cells. Such a network structure based clustering method can automatically determine the cluster number and the priori information of clustering prototypes. Unfortunately, when there exists noise in data set, or the boundaries among clusters are indistinct, if noise samples or samples on the cluster boundaries are selected as antigen, the immune system will be activated greatly and make cell proliferation and antibody secretion. One thus cannot achieve the clear network structures and the correct classification. In addition, this method cannot process data set with categorical attributes either. In order to solve this problem, a novel clustering method with network structure is presented based on the clonal algorithm。

3. A Novel Clustering Algorithm with Network Structure based on the Clonal Algorithm

At first, a new distance measure is defined for the novel algorithm. Let $X=\{x_1,x_2,\cdots,x_n\}$ denote a set of *n* objects and $x_i=[x_{i1},x_{i2},\cdots,x_{im}]^T$ be an object represented by *m* attribute values. Let c be a positive integer.

3.1 The definition of distance measure

3.1.1 Distance measure for numeric data clustering

The widely used distance measure is the Euclidean distance [13]. For the data set with real attributes, the Euclidean distance can be written as

$$d^2(x_j,p_i)=(x_j-p_i)^T\cdot(x_j-p_i) \tag{1}$$

In the clustering algorithm based on network structure, let $P_i=[p_{i1},p_{i2},\cdots,p_{ii_n}]$, $i=1,2,\cdots c$, i_n is the number of cell, which is contained in the *i*-th network structure P_i; $p_{ig}=[p_{ig,1},p_{ig,2},\cdots,p_{ig,m}]^T$, $g=1,2,\cdots i_n$ is the *g*-th neuron of the *i*-th network. The dissimilarity measurement between x_j and the *i*-th cluster is modified as:

$$d^2(x_j,P_i)=\min\{(x_j-p_{ig})^T\cdot(x_j-p_{ig}),\quad g=1,2,\cdots i_n\} \tag{2}$$

3.1.2 Distance measure for mixed data clustering

When *X* has mixed attributes, assuming that each object is denoted by $x_i=[x_{i1}^r,\cdots,x_{it}^r,x_{i,t+1}^c,\cdots,x_{im}^c]^T$, the dissimilarity between mixed-type objects x_j and the network cell p_{ig} can be measured by the Eq.(3).

$$d^2(x_j,p_{ig})=\sum_{l=1}^{t}|x_{jl}^r-p_{ig,l}^r|^2+\lambda\cdot\sum_{l=t+1}^{m}\delta(x_{il}^c,p_{ig,l}^c) \tag{3}$$

$\delta(\cdot)$ is defined as

$$\delta(a,b)=\begin{cases}0 & a=b\\ 1 & a\neq b\end{cases} \tag{4}$$

The λ is used to avoid favoring either type of attribute. Then we rewrite Eq.(2) as

$$d^2(x_j,P_i)=\min\left\{\sum_{l=1}^{t}|x_{jl}^r-p_{ig,l}^r|^2+\lambda\cdot\sum_{l=t+1}^{m}\delta(x_{il}^c,p_{ig,l}^c),\quad g=1,2,\cdots i_n\right\} \tag{5}$$

In this way, by modifying the distance measurement function, the categorical attributes can be treated as well as numeric attributes. With the new distance function, the available clustering algorithm can process the data set with mixed attributes.

3.2 A Novel Clustering Algorithm with Network Structure based on the Clonal Algorithm

Based on the phenomenon of the clonal selection and the forbidden clone, this paper presents a novel clustering

algorithm with network structures based on the clonal algorithm to analyze the data set with indistinct boundaries among clusters.

Let each sample x_j in the data set $X=\{x_1, x_2, \cdots, x_n\}$ is a different antigen. According to the principle of the immune network, once a new antigen x_j raise in body, the existing network neurons, *i.e.*, antibody will recognize it. The successfully recognized antibody will activate the network, which leads proliferation of antibody. If the antigen corresponds to noise data or samples in the indistinct boundaries among clusters, the forbidden clone will be performed and the corresponding neuron will be excluded. In addition, by reducing the antibodies with Ab-Ab infinity less than a threshold σ_s, the network structure will be predigested.

The distance measure of our clustering algorithm indicates that the smaller the distance measure is, the better the clustering partition. For this case, the clonal algorithm asks for a bigger affinity value. Hence, we define the Ab-Ag affinity function by using the dissimilarity measure.

$$f(x_j, p_{ig}) = \frac{1}{1+\sum_{l=1}^{t} |x_{jl}^r - p_{ig,l}^r|^2 + \lambda \cdot \sum_{l=t+1}^{m} \delta(x_{il}^c, p_{ig,l}^c)} \quad i=1,2,\cdots c,\ g=1,2,\cdots i_n \quad (6)$$

The Ab-Ab affinity is defined as:

$$D_{ij} = \left\| p_{ig} - p_{jl} \right\| \quad i,j=1,2,\cdots c \quad g=1,2\cdots i_n \quad l=1,2,\cdots j_n \quad (7)$$

Where, $\|\cdot\|$ is any a norm; $D=(D_{ij})_{N\times N}$ is the antibody-antibody affinity matrix, and $N=\sum_{i=1}^{c} i_n$ is the number of neurons of networks.

After obtaining the final neurons of networks, we have to solve the following two problems. (1) How many clusters contained in data set on earth? (2) How to partition the obtained neurons into categories? Here, we apply the minimal spanning tree (MST) technique to explore the relationship among neurons. The MST is often used to detect and describe the network structure of clusters [14].

Definition 2: A tree is a *spanning tree* of a graph if it is a sub-graph containing all the vertices of the graph. A *minimal spanning tree* of a graph is a spanning tree with minimum weight. The weight of a tree is defined as the sum of the weights of its constituent edges.

After obtaining the MST of the networks, by detecting the number of valleys in the bar plot of the MST, one can determine the cluster number of the given data set. In addition, we detect each distance D_{ij} between any neuron pair (i,j). If the D_{ij} is large enough, the neuron pair (i,j) will be disconnected. Finally, the neurons connected in the MST form a category. The number of connected part in the MST corresponds to the cluster number. Thus, if x_j belongs to the *i*-th cluster, we have

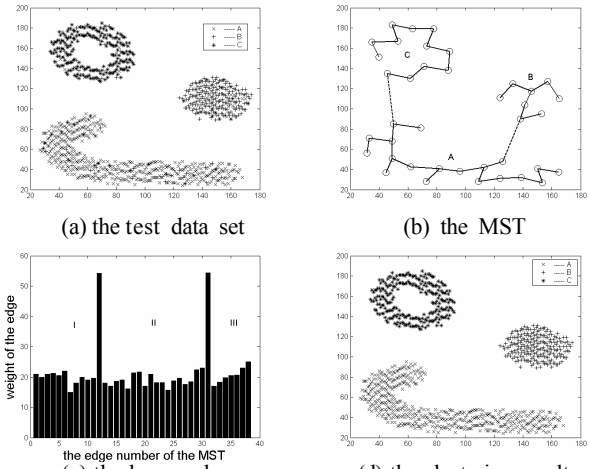
$$d^2(x_j, P_i) = \min\left\{ d^2(x_j, P_l),\quad l=1,2,\cdots c \right\} \quad (8)$$

4. Experimental Results and Discussion

To test the effectiveness of the proposed novel clustering algorithm with network structure based on the clonal algorithm, we present some preliminary experimental results in this section. The experimental results with various prototype data sets demonstrate the good performance of the novel algorithm.

4.1 Data set with distinct boundaries

To simplify illustration, we use data records having only three attributes, two numeric values and one categorical value. First, we generate an erose-shaped data subset, a ring-shaped subset and a spheral subset respectively. Then these points are expanded to 3D by adding a categorical value to each point (see Figures 1(a)).



(a) the test data set (b) the MST

(c) the bar graph (d) the clustering result

Figure 1 The test result with distinct boundaries

The analysis results of the data set with the proposed algorithm are presented in Figure 1 (b)-(d). In this experiment, the termination condition is the evolutionary generation equal to 5. From Figure 1(c), it can be obtained that the valley number of the bar plot is just equal to the cluster number. The final classified result is shown in Figure 1 (d), in which all the samples are classified correctly. It is obvious that the proposed clustering algorithm can effectively analyze data set with mixed attributes and erose-shaped prototypes.

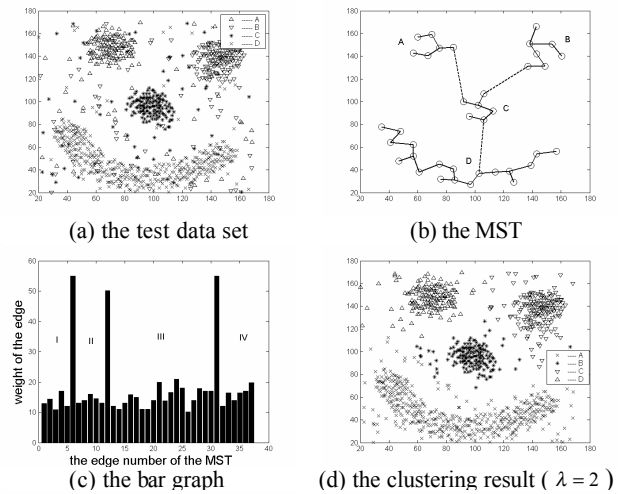
4.2 Data set with indistinct boundaries

For the data set with multi-type prototypes, if the clusters are well separated, *i.e.*, each cluster has distinct boundary, the traditional fuzzy *c*-means algorithm can also achieve good classification result. However, in most cases the data set encountered are with indistinct boundaries among clusters. In this case, the available algorithms will not be able to obtain satisfied result.

Based on the same reason as the above experiments, we construct a test data set with three attributes as shown in Figure 2 (a). Figure 2 (b),(c) show the produced MST of network structures and its bar plot respectively by the clustering algorithm based on the clonal algorithm. From Figure 2(b), it can be found that the obtained synthetic network structure well agrees with the structure of the

original data set, and the proposed algorithm can explore the interior structure of data set very well.

Figure 2 (d) shows the clustering result of the proposed novel algorithm. Although the given data set contains different prototypes with indistinct boundaries, for introducing the forbidden clone operator, the proposed clustering algorithm can still obtain good result. When the samples on the boundaries are selected as antigen, the forbidden clone operator makes the neuron in suppression state, which guarantees the final network neurons can stand for the representative samples of data set.



(a) the test data set (b) the MST

(c) the bar graph (d) the clustering result ($\lambda = 2$)

(e) the clustering result of the k-prototype algorithm (f) the MST obtained by the EAIN with $\sigma_s = 10$

Figure2 The test result with indistinct boundaries

Figure 2 (e) is the result of the traditional k-prototype algorithm. Since the data set contains different prototypes and ambiguous boundaries among clusters, the k-prototype algorithm analyzes the data set with the same prototypes and results in many samples misclassified. Figure 2 (f) shows the MST of network neurons generated by the standard EAIN σ_s . Since the standard EAIN only emphasize the effect of the clonal selection, if the samples on the boundaries are selected as antigens, for the higher Ab-Ag affinity, the corresponding network neurons will be activated and make the proliferation of antibodies with specificity. Moreover, the corresponding network neurons will not be able to remove by either clone compression or network compression. Therefore, the obtained networks cannot clearly reflect the structure of data set. Also the related cluster numbers and classification information will not able to obtained. It thus is impossible to analyze and classify the data set correctly.

5、Conclusion

This paper presents a novel clustering algorithm with network structure based on the clonal algorithm. Since the new algorithm combines the clonal selection algorithm and the forbidden clone operator, the obtained network has not only the specificity but also the tolerance of immunity. The experimental results illustrate that the novel algorithm can effectively explore the cluster structures of the data set. Moreover, it does not depend on the prototype initialization and the priori information of cluster number, which makes it as a real unsupervised learning.

REFERENCE

[1] Gao Xinbo, "Studies of optimization and applications of fuzzy clustering algorithm," Doctoral Dissertation, Xidian University, Xi'an, China, 1999. (In Chinese)

[2] Bezdek J.C, "Patten Recognition with Fuzzy Objective Function Algorithms," Plenum Press, New York, 1981.

[3] Dave R.N. and Bhaswan K, "Adaptive fuzzy c-shells clustering and detection of ellipses," IEEE Trans. NN, 1992, 3(5): 643-662.

[4] Krishnapuram R., Frigui H. and Nasraoni O, "Fuzzy and possiblistic shell clustering algorithms and their application to boundary detection and surface approximation-Part I," IEEE Trans. FS, 1995, 3(1): 29-43.

[5] Gao Xinbo, Xue Zhong, Li Jie, "An initialization method for fuzzy clustering with multi-type prototypes," Journal of Chinese Electronics, 27(12): 72-75, 1999. (In Chinese)

[6] Hathaway R.J. and Bezdek J.C,"Switching regression models and fuzzy clustering," IEEE Trans. FS, 1993, 3(1): 195-204.

[7] Gath I. and Geva A. B, "Unsupervised optimal fuzzy clustering," IEEE Trans. SMC, 1989, 11(7): 773-781.

[8] Li Jie, "A GA-based clustering algorithm for large data set with mixed attributes," Technical Report, School of Electronic Engineering, Xidian University, 2002.

[9] William H. H, Loretta S. A, William M. P, David T, and Michael W, "Self-Organizing Systems for Knowledge Discovery in Large Databases," http://www.kddresearch.org/Publications/Conference

[10] Leandro N.C. and Fernando J.Z, "An Evolutionary Immune Network for Data Clustering," Proceedings of the IEEE Computer Society Press, SBRN'00, 2000, vol.1: 84-89.

[11] Zhexue Huang, "Clustering large data sets with mixed numeric and categorical values," Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore: World Scientific, pp. 21-34.

[12] Jerne N.K, "Towards a Network Theory of the Immune System," Ann. Immunol. 1974. (Inst. Pasteur) 125C, pp.373-389.

[13] B. Everitt, "Cluster Analysis," Heinemann Educational Books Ltd., 1974.

[14] Zahn C.T, "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters," IEEE Trans. on Computers, 1971, 20(1): 68-86.