基于克隆算法的网络结构聚类新算法

李 洁, 高新波, 焦李成
(西安电子科技大学电子工程学院, 陕西西安 710071)

摘 要: 基于目标函数的聚类算法是目前应用最为广泛的聚类分析方法之一. 然而这类算法都需要类别数和聚类原型的先验知识, 且只能分析具有相同原型的数值型数据. 此外这类算法还存在对初始化敏感, 易陷入局部极值点等弱点. 为此, 本文提出一种基于克隆算法的网络结构聚类新算法以实现聚类分析的自动化. 由于新算法将克隆选择与禁忌克隆相结合, 使网络既具有免疫的特异性又具有免疫的耐受性, 通过分析网络神经元的最小生成树, 能够快速准确地获得类别数以及相关的分类信息. 对各种类型的数据集的测试结果均表明, 本文提出的新算法对于处理具有混和特征的数据集聚类分析问题是相当便捷有效的.

关键词: 聚类分析; 数值特征; 类属特征; 克隆选择; 禁忌克隆
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2004) 07-1195-05

A Novel Clustering Method with Network Structure Based on Clonal Algorithm

LI Jie, GAO Xin-bo, JIAO Li-cheng
(School of Electronic Engineering, Xidian Univ., Xi. an, Shaanxi 710071, China)

Abstract: In the field of cluster analysis, objective function based clustering algorithm is one of widely applied methods so far. However, this type algorithms need the prior knowledge about the cluster number and the type of clustering prototypes, and can only process data sets with the same prototypes. Moreover, these algorithms are very sensitive to the initialization and easy to get trapped into local optima. To this end, this paper presents a novel clustering method with network structure based on clonal algorithm to realize the automation of cluster analysis. Since the new algorithm combines the clonal selection algorithm and forbidden clonal operation, the obtained network has not only the specificity but also the tolerance of immunity. By analyzing the neurons of obtained network with minimal spanning tree, one can easily get the cluster number and related classification information. The test results with various data sets illustrate that the novel algorithm achieves more effective performances on cluster analyzing the data set with mixed numeric values and categorical values.

Key words: cluster analysis; numeric feature; categorical feature; clonal selection; forbidden clone

1 引言

聚类分析是多元统计分析的方法之一, 也是统计模式识别中非监督模式分类的一个重要分支[1]. 在传统的聚类方法中, 基于目标函数的聚类算法由于把聚类问题归结为一个优化问题, 具有深厚的泛函基础, 从而成为聚类算法研究的主流. c-均值算法就是其中最具代表性的一种[2], 但由于它是以聚类中心为原型, 因而不能检测特征空间中非线性子空间中存在的聚类, 为此人们对聚类原型模式进行了相应的扩展, 形成了从特征空间中的点到线、面、壳以及二次曲线等诸多类型的原型, 提出了象 c-线、c-面、c-壳、c-二次曲线[3~5]等一系列的针对各种原型的聚类算法. 实现了对各种不同原型的聚类分析.

但是这种聚类算法存在的最大问题就是对聚类先验知识要求的增加, 因为在聚类分析之前, 必须给定聚类原型的类型以及聚类类别数 c, 否则将会对算法产生误导[2], 从而得到一个错误的划分; 同时这类算法要求在数据集中不同类别的样本只能呈同一种类型的分布模式, 只是参数有所差别, 从而限制了其实际应用的范围. 因为在实际的应用中, 很多情况下同一数据集中往往含有多种不同原型的样本子集, 而且当样本点处于高维空间时, 很难获得聚类类别数 c.

近期, 人们提出了一种多类原型模糊聚类算法[6], 将现有的原型聚类算法集成在一起. 但这种方法的成败几乎全部依赖于初始化. 而且研究表明, 基于原型的聚类算法求解过程极不稳定, 从而对原型初始化要求很高, 否则稍有不慎就会陷入局部极值点. 为了解决局部极值问题, 随着遗传算法的出现,

收稿日期: 2003-05-19; 修回日期: 2004-01-20
基金项目: 国家自然科学基金(No. 60073053; 60202004)、863 计划资助课题(No. 2002AA135080)

人们提出了基于遗传算法的聚类方法, 尽管该方法能以较高的概率收敛到全局最优点, 但收敛速度较慢, 而且还容易出现早熟[7]现象.

针对类别数 c 未知的问题, 有人提出采用自组织映射神经网络进行聚类[8], 虽然能解决预先设定类数问题, 但对多类原型的数据分析却是无能为力.

聚类分析以非监督学习著称, 而上述的各种算法显然都需要先验知识. 随着人工免疫系统方法的兴起, 人们又提出利用人工免疫网络进行聚类分析的方法[9], 真正实现了无监督分类, 但这种方法仅适合于各样本子集边界清晰的情况, 如果各子集间交集非空, 就得不到有效的网络结构.

同时作为数据挖掘的一种有利工具, 聚类分析常常需要处理大量高维数据集, 而且这些数据通常既包含数值也包含类属值. 传统的将类属值转化为数值的方法不是总能得到有效的结果, 这是因为类属域是无序的.

只有很少几种算法能较好的处理混合属性数据聚类问题, 例如 k2原型算法等, 但这些方法同样要求聚类类别数 c 和聚类原型先验知识.

鉴于此, 本文提出一种基于克隆算法的网络结构聚类新算法, 该算法能同时处理多类原型的数据聚类分析问题, 而且在聚类的过程中自动获得类数信息; 同时我们定义了一种新的克隆算子- 禁忌克隆(Forbidden Clone), 将该算子与克隆选择算子相结合, 由于克隆选择算法是群体搜索策略, 本质上固有并行性和搜索变化的随机性, 在搜索中不易陷入局部极小值, 最终能以较大的概率获得问题的全局最优解, 且具有较快的收敛速度; 而通过禁忌克隆运算使网络产生免疫耐受性, 从而克服了边界模糊对分类效果的影响; 并定义了一种新的距离测度函数, 将不同属性特征相结合, 从而达到对具有混合属性特征的数据进行聚类分析的目的.

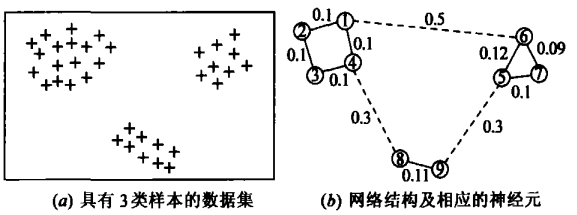
本文的安排如下: 下一节介绍基于进化免疫网络的聚类算法, 第三节讨论基于克隆算法的网络结构聚类新算法, 第四节为实验结果, 将本文提出的新方法与标准网络结构算法和 k2原型算法进行了性能比较, 并测试了禁忌克隆运算对基于网络结构聚类算法的收敛性能的影响. 最后总结全文, 并指出进一步的研究方向.

2 基于进化免疫网络的聚类算法

免疫网络理论是由 Jerne 在 1974 年最先提出的[10], Leandro 根据 Jerne 的免疫网络理论, 在 2000 年提出一种人工免疫进化网络(Evolutionary Artificial Immune Network)[9]. 其主要思想是: 假设 $X=\{x_1, x_2, \cdots, x_n\}$ 是待分析的数据集, 其中 $x_i=[x_{i1}, x_{i2}, \cdots, x_{im}]^T$ 表示第 i 个样本的 m 个特征值, x_i 可以用状态空间 S^m 中的一个点 s 来表示. 将这个点 s 作为抗原, 来决定抗体- 抗体以及抗原- 抗体之间的相互作用. 并且系统内部的相互作用可以用一个连通图来表示, 这时网络模型可以定义为:

定义 1 人工免疫进化网络是一个加权的图 G, 该图由一组不完全连接的称作神经元的节点构成, 每对节点产生一条边, 边的长度称为权值或者连接强度.

数据集中不同的样本就可以分别映射到网络内部相应的神经元上, 彼此亲合度高的神经元归为一类, 这样样本在数据集中也相应地分为不同的类. 假设数据集中包含 3 类样本(如图 1(a)所示), 经过学习后得到的网络结构显示于图 1(b)中, 圆圈里的数字是该神经元的标号, 连线上的数字表示其连接强度.



(a) 具有 3 类样本的数据集 (b) 网络结构及相应的神经元

图 1 人工免疫网络示意图

从图 1(b)中可以看出, 该网络结构中明显包含 3 类神经元, 且每一类中神经元的数目、连接方式及其强度都不相同. 这些神经元分别是原始数据集中不同样本的映射. 我们注意到网络结构中神经元的个数远远小于数据集中的样本数, 因此, 这种人工免疫网络具有数据压缩的功能, 也适合于大数据集聚类分析.

这种基于网络结构的聚类方法虽然解决了聚类分析前需要先验知识的问题, 但该方法存在一个较大的弱点就是当数据集中各类间边界不明显, 或者存在噪声时, 这些样本作为抗原, 能够极大的激活免疫系统, 引起细胞增殖以及抗体分泌, 使我们的网络结构不再清晰, 从而不能进行正确的分类. 另外, 该方法不能对具有类属特征的数据进行分析. 为了解决上述问题, 本文提出一种基于克隆算法的网络结构聚类新算法.

3 基于克隆算法的网络结构聚类新算法

首先, 我们定义了一种新的距离测度函数, 来实现将不同属性特征相结合的目的. 令 $X=\{x_1, x_2, \cdots, x_n\}$ 表示一组具有 n 个样本的数据集, 其中 $x_i=[x_{i1}, x_{i2}, \cdots, x_{im}]^T$ 表示第 i 个样本的 m 个特征值. 令 c 是一个正整数, 那么对 X 进行聚类的目的就是要找到一个划分, 将 X 中的目标分为 c 类.

3.1 距离测度函数的定义

3.1.1.1 数值数据聚类的距离测度函数 目前大家广泛使用的距离测度函数是欧几里德距离[11]. 对于具有实特征的数据集, 即 $X<R^m$, 则样本 x_j 与第 i 类的原型 p_i 间的距离测度函数如下式(1)所示.

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

在基于网络结构的聚类算法中, 设 $P_i=[p_{i1}, p_{i2}, \cdots, p_{ii_n}]$, $i=1, 2, \cdots, c$, i_n 表示第 i 类的网络结构中包含的神经元的个数, $p_{ig}=[p_{ig,1}, p_{ig,2}, \cdots, p_{ig,m}]^T$, $g=1, 2, \cdots, i_n$ 是第 i 类的网络结构中第 g 个神经元, 则可将 x_j 与第 i 类的相异性测度修正为:

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

3.1.1.2 混合数据聚类中的距离测度函数 当样本具有混和特征时, 设每个样本用 $x_i=[x_{i1}^r, \cdots, x_{it}^r, x_{i,t+1}^c, \cdots, x_{im}^c]^T$ 表示, 其中前 t 个分量是数值特征, 后 m- t 个分量是类属特征, 则

x_j 和第 i 类网络结构中第 g 个神经元的 p_{ig} 间的相异性测度可由公式(3)计算:

$$d^2(x_j, p_{ig}) = \sum_{l=1}^{t} |x_{jl}^r - p_{ig,l}^r|^2 + K\# \sum_{l=t+1}^{m} D(x_{jl}^c, p_{ig,l}^c) \quad (3)$$

式中第一项是数值特征上的欧几里德距离平方, 第二项是类属特征上的简单的相异匹配测度. D(#)定义为:

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

权值 K 用来调节两种特征的比例, 以避免偏向任何一种特征. 关于 K 取值对分类结果的影响我们将另文讨论.

对于混合类型的目标, 我们可以通过修正式(2)得到 x_j 和第 i 类网络结构 P_i 之间的相异性测度为:

$$d^2(x_j, P_i) = \min_{g=1,2,\cdots,i_n} \left\{ \sum_{l=1}^{t} |x_{jl}^r - p_{ig,l}^r|^2 + K\# \sum_{l=t+1}^{m} D(x_{jl}^c, p_{ig,l}^c) \right\}, \quad (5)$$

这样通过修改距离测度函数, 就把数值特征和类属特征相结合, 从而达到对混和属性数据聚类的目的.

3.2　基于克隆算法的网络结构聚类新算法

1958 年 Burnet 提出体内有针对各种抗原的相应细胞系, 抗原进入机体后, 选择相应细胞系与之结合, 使之增殖并产生特异性抗体, 这就是著名的克隆选择学说. 若在胚胎期间由某抗原选择相应细胞系接触后, 这些细胞系处于抑制状态, 而机体失去针对这种抗原的反应性, 形成耐受, 则称为禁忌克隆[12]. 本文即根据机体的克隆选择与禁忌克隆现象, 提出基于克隆算法的网络结构聚类新算法来分析类间边界不明显的数据集聚类问题.

令原始数据集 $X = \{x_1, x_2, \cdots, x_n\}$ 中的每个样本 x_j 都是不同的抗原, 根据免疫网络原理, 当机体内出现一个新的抗原 x_j 时, 现有的网络神经元(即抗体)要对其进行识别, 成功识别的抗体将激活网络, 并导致相应抗体的增殖. 若该抗原对应为噪声点或者是处在不明显边界上的样本点, 则进行禁忌克隆运算, 将相应的神经元排除. 此外, 通过将抗体-抗体亲合度小于门限 R_s 的抗体清除, 可以达到简化网络结构的目的.

我们借助距离测度函数来构造抗体-抗原亲合度函数如下式.

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

抗体2抗体亲合力定义为:

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

+#+ 为任意范数, $D = (D_{ij})_{N \times N}$ 为抗体-抗体亲合力矩阵, $N = \sum_{i=1}^{c} i_n$ 为网络神经元的个数.

基于禁忌克隆的网络结构聚类新算法可描述为:

Step 1　l= 1, 设定算法参数, 初始化网络神经元 A(1);
Step 2　对每一个抗原 x_i, 根据式(6)分别计算其与所有网络神经元的亲合度;
　Step 211　选择亲合度最高的 k 个神经元为抗体, 按照亲合度越高, 抗体克隆规模越大的原则进行克隆, 克隆总数为 N_c;
　Step 212　利用式(8), 通过减小 N_c 个克隆抗体与抗原 x_i 之间的距离, 来增加其抗体2抗原亲合度;
　Step 213　计算改良的抗体与抗原 x_i 间的亲合度, 并将亲合度最高的 N% 个抗体作为网络记忆细胞, 存储于 M_p 中;
　Step 214　在 M_p 中, 将抗体2抗原亲合度小于门限 R_d 的记忆细及抗体2抗体亲合度小于门限 R_s 的记忆细胞死亡;
　Step 215　将原始网络神经元 A(1)与记忆细胞 M_p 结合, 形成新的网络神经元 A(1) z [A(1); M_p];
Step 3　计算每个神经元与其它神经元间的亲合度, 若亲合度小于门限 R_s 的神经元个数小于 R_f, 则将该神经元排除;
Step 4　计算网络神经元间的亲合度, 将亲合度小于门限 R_s 的神经元死亡;
Step 5　用随机产生的神经元替代 r% 个亲合度最差的神经元;
Step 6　l= l+ 1, 若满足终止条件, 停止计算; 否则, 返回 step2.

上述算法中, R_s 是网络压缩门限, 一般先取较小值, 然后不断增大其值, 并分析获得结果, 以便最终确定合适的网络参数; R_d 是克隆死亡率, 可以适当简化网络结构; R_f 是禁忌克隆门限, 可使网络产生免疫的耐受性.

对于给定抗原 x_i, 其与网络中各个抗体的亲合度通过下式进行改良:

$$A(1) = A(1) - A(A(1) - x) \quad (8)$$

式中 $x = x_i \# I_N$, $N = \sum_{i=1}^{c} i_n$, 为网络神经元的个数, I_N 为 N 维全 1 行向量, A 是 N 维向量, 称作变异率.

当数据集与相应网络神经元之间的距离测度和达到最小时终止循环. 在得到最终的网络神经元后, 我们还需要解决下面两个问题(1)数据集中究竟包含几类样本? (2)如何确定各个网络神经元分别属于哪一类? 本文采用最小生成树(Minimal Spanning Tree, MST)来分析网络神经元间的关系[13, 14].

定义 2　连通图 G 的一个子图如果是一棵包含 G 的所有顶点的树, 则该子图称为 G 的生成树. 生成树各边的权值总和称为生成树的权. 权最小的生成树称为最小生成树.

在得到网络结构的最小生成树后, 通过寻找 MST 的 bar 图中山谷的数目就可以确定数据集中样本的类别数; 另外我们检测相邻神经元(i, j)之间的距离 D_{ij}, 若 D_{ij} 足够大, 则将 MST 中(i, j)间的连接断开, 这样 MST 中连通的神经元就属于

同一类, 不连通的神经元分属于不同的类. 这样, 当样本 x_j 属于第 i 类时, 应有:

$$d^2(x_j, P_i) = \min\{d^2(x_j, P_l), \quad l = 1, 2, , , c\} \quad (9)$$

4 实验结果与分析

4.1.1 具有相同原型的数据集分类性能检验

为便于直观显示, 我们构造的数据样本仅具两个数值型和一个类属型的特征. 首先产生六组正态分布的二维数据点, 共包含 600 个样本, 如图 2(a) 所示. 然后给每一个点迭加一个类属特征而扩展到三维(如图 2(b)).

我们对图 2(b) 所示的数据集用本文 SC= 5 提出的新算法进行分析, 该实验终止条件是循环次数 , 图 2(c) 是得到的网络神经元的最小生成树, 其中虚线表示要断开连接, 图 2(d) 是相应的 bar 图, 从该图中可以看出, 山谷的个数恰好等于数据集中包含的类别数. 图 2(e) 和(f) 分别是最终的合成网络结构以及相应的分类结果.

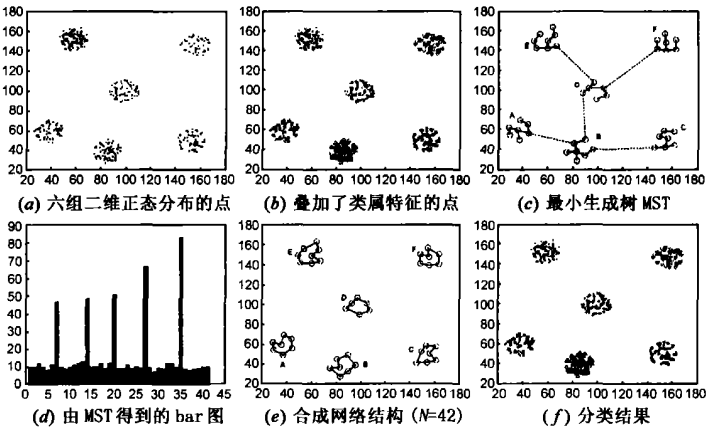


图 2 具有相同原型的数据集测试结果

4.1.2 具有不同原型的数据集分类性能检验

在需要聚类分析的数据集中, 常常同时包含各种不同的原型. 在本实验中, 我们将分两种情况进行讨论, 一种是原型种类不同, 但各类间边界清晰的数据集, 另一种是原型种类不同, 而且各类间边界模糊的数据集.

4.1.2.1 数据集边界清晰

同样为了便于直观显示, 我们构造的数据样本仅具有两个数值特征和一个类属特征. 首先, 我们构造分别包含一个任意形状、环形以及团状分布 3 种不同原型的二维数据点(共包含 1500 个样本), 在此基础上, 按照 4.1.1 节的方法产生一个类属特征, 作为测试数据集, 分别如图 3(a) 和(b) 所示.

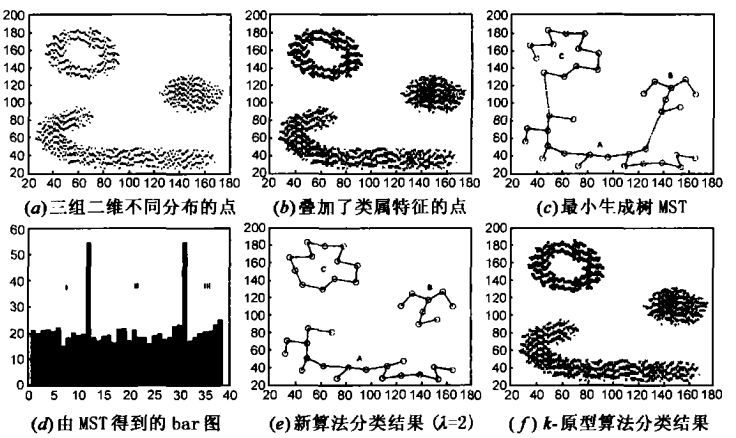


图 3 具有不同原型的数据集测试结果(一)

对图 3(b) 所示数据集分析得到的结果分别显示于图 3(c)~(f), 该实验终止条件是 , 从图 3(d) 中也可以看出, 山谷的个数恰好等于数据集中包含 SC= 5 的类别数. 图 3(f) 是分类结果, 我们看到所有样本均被正确分类.

4.1.2.2 数据集边界模糊

基于同样的原因, 我们首先构造了具有 3 个特征的检测数据集, 如图 4(a) 所示. 图 4(b)~(d) 分别是基于克隆算法的网络结构最小生成树、合成网络结构以及相应的 bar 图.

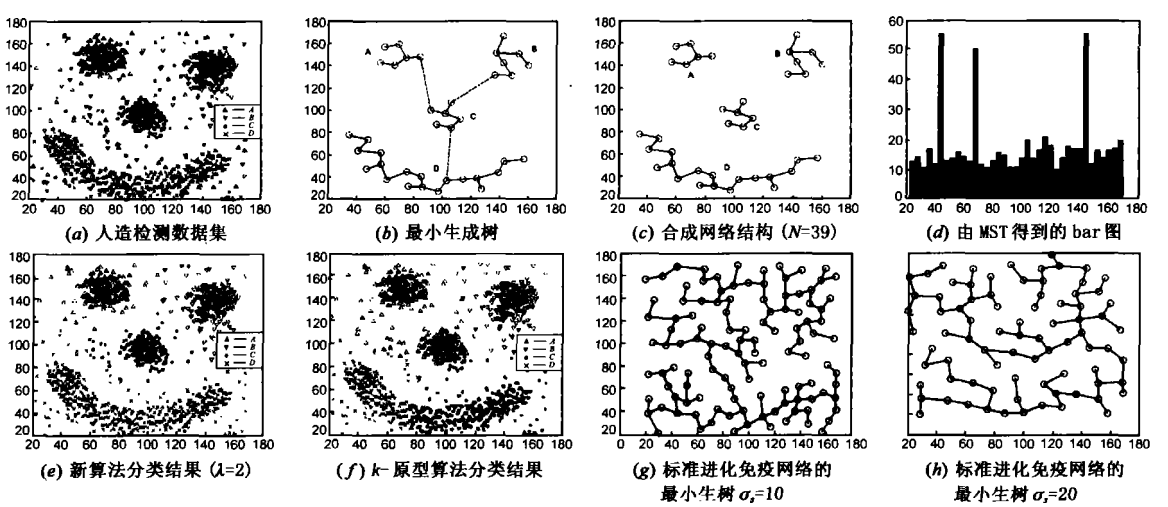


图 4 具有不同原型的数据集测试结果(二)

图 4(e) 显示了本文新算法的聚类结果, 由于我们的新算法中增加了禁忌克隆运算, 当样本处于不明显的边界上时, 该样本点作为抗原, 使相应的神经元处于抑制状态, 从而保证了最终的网络神经元都能代表各类的典型样本, 因此达到较好的分类效果. 图 4(f) 是传统的 k2 原型算法的分类结果, 由于数据集中包含了不同原型, k2 原型算法仅采用同一种原型进行分析显然是不合适的. 图 4(g) 和(h) 都是根据标准进化免疫网络原理得到的网络神经元最小生成树, 只是门限 R_s 的取值不同, 从中我们看到, 由于标准进化免疫网络算法中, 只强调克隆选择的作用, 处在模糊边界上的样本作为抗原, 具有很高的抗原2抗体亲合度, 能够强烈地活化相应的网络神经元, 这时不论是克隆压缩还是网络压缩都不能除去相应的网络神经元, 从而使最终得到的网络不能清晰表现数据的结构, 也不能从中得到相关的类别数以及各种分类信息, 因而达不到正确分类的目的.

5　结论

本文提出一种基于克隆算法的网络结构聚类新算法. 该算法既包括克隆选择又包括禁忌克隆, 从而使得到的网络既具有免疫特异性又具有免疫耐受性. 实验结果表明该方法能够有效地发现数据中的聚类结构. 且不依赖初始原型的选择, 也无需类数的先验知识, 可以真正做到无监督自学习, 这在实际应用中是非常方便的.

本文的重点放在如何将克隆算法应用于网络结构聚类算法中, 来解决具有混合属性特征数据的聚类问题. 但在实验过程中, 我们发现网络对压缩门限 R_s 和禁忌克隆门限 R_f 比较敏感, 如何自适应的设定相关参数以及 K 的最优取值问题, 这将是我们进一步研究的重要方向.

参考文献:

[1] 何清. 模糊聚类分析理论与应用研究进展[J]. 模糊系统与数学, 1998, 12(2): 89- 94.

[2] 高新波. 模糊聚类算法的优化及应用研究[D]. 西安: 西安电子科技大学, 1999.

[3] Bezdek J C. Patten Recognition with Fuzzy Objective Function Algo2rithms[M]. New York: Plenum Press, 1981.

[4] Dave R N, Bhaswan K. Adaptive fuzzy c2shells clustering and detection of ellipses[J]. IEEE Trans NN, 1992, 3(5): 643- 662.

[5] Krishnapuram R, Frigui H, Nasraoni O. Fuzzy and possiblistic shell clustering algorithms and their application to boundary detection and surface approximation2Part I[J]. IEEE Trans FS, 1995, 3(1): 29- 43.

[6] 高新波, 薛忠, 李洁. 一种多类原型模糊聚类的初始化方法[J]. 电子学报, 1999, 27(12): 72- 75.

[7] 李洁. 一种基于GA的混和属性特征大数据集聚类算法[R]. 西安: 西安电子科技大学, 2002.

[8] William H H, Loretta S A, William M P, David T, Michael W. Self2Or2ganizing Systems for Knowledge Discovery in Large Databases[OL]. http://www.kddresearch.org/Publications/Conference/HAPTW1.pdf

[9] Leandro N C, Fernando J Z. An evolutionary immune network for data clustering[A], Proceedings of the IEEE Computer Society Press[C]. USA: IEEE Press, 2000. 84- 89.

[10] Jerne N K. Towards a Network Theory of the Immune System[M]. ANN. Immunol, Paris(Inst Pasteur), 1974, 125C: 373- 389.

[11] B Everitt. Cluster Analysis[M]. Heinemann Educational Books Ltd., 1974.

[12] 吴敏毓, 刘恭植, 等. 医学免疫学[M]. 合肥: 中国科学技术大学出版社, 1999.

[13] Zahn C T. Graph2theoretical methods for detecting and describing gestalt clusters[J]. IEEE Trans on Computers, 1971, 20(1): 68- 86.

[14] Leclerc B. Minimum spanning trees for tree metrics: Abridgements and adjustments[J]. Journal of Classification, 1995, 12: 207- 241.

作者简介:

李　洁　女, 1972 年生于陕西西安, 工学硕士, 西安电子科技大学讲师, 现为西安电子科技大学电子工程学院博士研究生, 主要从事人工智能、模式识别, 数据挖掘等方面的研究.

高新波　男, 1972 年生于山东莱芜, 工学博士, 西安电子科技大学教授, 博士生导师, IEEE 会员, 中国电子学会高级会员, 主要从事智能信息处理、计算机视觉、基于内容的图像与视频信息检索等领域的研究.

焦李成　男, 1959 年 10 月出生于陕西白水, 1984 年和 1990 年在西安交通大学分别获得硕士和博士学位, IEEE 高级会员, 现为西安电子科技大学教授, 博士生导师, 主要从事非线性科学和智能信号处理以及神经网络与大规模并行处理等领域的研究.