基于空间聚类的动物疫点分布划分算法研究

郭茂耘[1] 潘丽娟[1] 江红旗[2] 柴 毅[1]
(重庆大学自动化学院 重庆 400030)[1] (重庆出入境检验检疫局 重庆 400030)[2]

摘 要 防疫管理资源(人员和设备等)的合理有效配置是动物疫情防疫管理关注的问题之一。根据疫点的空间分布情况,基于空间聚类和最大夹角边界确定方法,提出了对疫点分布进行分类划分的方法。首先研究了将K-Means聚类分析方法应用于疫点的空间聚类分析,实现了疫点按空间亲疏关系的分类。在此基础上,根据最大夹角原理,在聚类结果中确定了每一个分类的边界线,实现了对疫点按空间关系进行分类划分区域的方法,从而为疫情管理人员有效地监测与分析疫情空间分布模式、控制管理和预防动物疫情的扩散提供了支持。

关键词 疫情,空间聚类,边界点搜索,分类区域
中图法分类号 TP301.6 **文献标识码** A

Partition Algorithm of Animal Foci Distribution Based on Spatial Clustering Analysis

GUO Mao-yun[1] PAN Li-juan[1] JIANG Hong-qi[2] CHAI Yi[1]
(College of Automation, Chongqing University, Chongqing 400030, China)[1]
(Chongqing Entry Exit Inspection and Quarantine Bureau, Chongqing 400030, China)[2]

Abstract The reasonable effective configuration of epidemic prevention and management resources (personnel and equipment, etc.) is one of the main concerns about animal epidemic prevention system. This paper proposed a method to classify and divide foci based on a spatial clustering method and a principle of maximum included angles. Firstly, the K-Means clustering analysis method was applied to the foci distribution and realized the classification of foci by intimate or distant relationship. Secondly, based on the principle of maximum angle, determined the boundary line of each classification. By above methods, realized the foci classification division according to the spatial relationship and got divided area, so that epidemic management personnel could get decision support to effectively monitoring and analyzing epidemic spatial distribution model, controlling and preventing the spread of the animal epidemic information.

Keywords Epidemics, Spatial clustering, Boundary points search, Classification zone

1 引言

近年来,世界各地频繁出现各种重大动物疫病,如口蹄疫、禽流感、高致病性蓝耳病等。动物疫病的爆发不仅将影响全球的食物供应,给社会经济造成最大损失,甚至会严重危害人类健康[1-3]。WHO指出,只要在环境中存在着禽流感病毒,就存在着新的人禽流感病毒流行的危险。动物疫情的爆发不仅影响农业生产,也会给人类的健康和安全造成直接的威胁。因此,对动物疫情及防控方法的研究是农业检验和公共卫生管理人员关注的热点之一。

动物疫情的发生往往受生态环境、人口分布、气候、河流等地理因素的影响,疫情的发生和流行具有地域性特点。因此,相关学者开展了动物疫情的空间分布情况和分布模式研究。在国内,王丽萍等人将空间聚类分析方法应用到疫情分析中,研究了1990—2006年安徽疟疾疫情时空分布特点[4];叶敏等人采用系统聚类方法对2004年我国各地区麻疹发病率进行聚类分析[5];武继磊等人根据北京市SARS病例数据,试图分析出SARS的空间过程[6]等。在国外,美国国防部建立了以社区为基础的流感电子监测系统以及流感监测系统,开展了流感疫情检测和应急响应[7];新泽西理工大学研究人员通过社交网络系统提取数据源,开发了疫情爆发和蔓延检测系统(EOSDS)[8]。

对疫点分区进行防疫管理,在动物疫情防控管理中具有重要的现实意义。通过确立疫点的空间分类管控区域,使防疫工作人员掌握疫情在空间上的分布情况和模式,针对不同地区采取不同的防制措施,提前做好疫苗的接种,合理有效地配置卫生人力和物力资源等,为动物疫情管控决策提供科学依据,以及时有效地预防和控制疫情的扩散。

本文针对动物疫情防控管理中对疫点分布进行空间分类划分区域的问题,利用地理信息系统(GIS)对疫情爆发情况进行可视化描述,应用空间聚类分析方法对疫点的空间分布特征按空间亲疏关系进行分类,实现在空间上对疫点的划分。

本文受国家质检总局科技计划项目(2011IK024)资助。

郭茂耘(1973—),男,副教授,主要研究方向为空间信息处理及应用、智能信息处理、复杂系统建模与仿真等;**潘丽娟**(1987—),女,硕士生,主要研究方向为决策支持与数据挖掘;**柴 毅**(1962—),男,教授,主要研究方向为信息处理、融合与控制、工业过程控制理论与技术、智能系统理论及其应用等。

并以此为基础，基于最大夹角原理，确定各疫点划分子集的边界点，得到各划分区域边界。

2 疫点的 K-Means 空间聚类分析方法

空间聚类分析是将数据挖掘分析中的聚类分析方法用于地理空间信息的处理中，已广泛应用于地震模式分析、气象、环境、社会经济及公共卫生等诸多领域[9-12]。本文以常用的 K-Means 聚类分析方法为基础，通过改进其中的数据距离计算公式，将空间实体的空间数据（疫点空间坐标）带入改进的聚类分析算法，来实现疫点的空间聚类分析。

2.1 K-Means 算法

K-Means 算法是一种常用的聚类分析方法，具有算法简单且收敛速度快的特点。基本思想是：首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心；而对于所剩下的其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止[13,14]。k 个聚类具有以下特点：各聚类本身尽可能紧凑，而各聚类之间尽可能分开。

K-Means 算法的流程如下：

(1)从待聚类数据 $G=\{G_1,G_2,\cdots,G_N\}$ 中，N 为数据数目，任意选择 K 个对象作为初始聚类中心，记 $Z_1(m),Z_2(m),\cdots,Z_i(m),\cdots,Z_K(m),i=1,2,\cdots,K,K$ 为聚类中心数目。$Z_i(m)$ 表示第 m 次迭代后获得的第 i 个聚类中心。

(2)在待聚类数据 G 中，按最小距离原则将样本分配给以上 K 个聚类中心，若：

$$\| G_i-z_j(m) \|_d=\min \| G_i-z_j(m) \|_d \tag{1}$$

式中，$j=1,2,\cdots,K;i=1,2,\cdots,N$，则 $G_i\in C_j(m)$。m 为迭代次数，$C_j(m)$ 为经过 m 次迭代得到的第 j 个聚类，其聚类中心为 $z_j(m)$。

(3)计算每个聚类中的样本均值作为新的聚类中心，即：

$$z_j(m+1)=1/N*\sum_{G_h\in C_j(m)}G_h \tag{2}$$

式中，$j=1,2,\cdots,K;h=1,2,\cdots,N_j;N_j<N,N_j$ 为第 i 个聚类 $C_j(m)$ 所包含的样本数，且 $G_h\in G$。

(4)若 $z_j(m+1)\neq z_j(m),j=1,2,\cdots,K$，令 $m=m+1$，重复步骤(2)、(3)，直到聚类中心不再变化，即 $z_j(m+1)=z_j(m)$。

(5)结束，得到 k 个聚类。

可以看出，K-Means 聚类过程是一个不断迭代处理的过程，其终止条件是聚类中心在两次迭代过程中，其变化值小于一个指定的阈值。

2.2 K-Means 空间聚类分析方法

K-Means 聚类分析应用于空间实体所属分类的基本思想就是将空间实体中与空间位置相关的属性特征进行聚类分析，找出属性特征上的共同点。针对所涉及的疫点区域划分问题，本文将疫点的空间坐标作为待聚类的数据，代入 K-Means 聚类分析算法进行聚类。因此，根据本文 2.1 节中的 K 均值聚类算法，可做如下改进：

(1)为完成疫点的空间聚类分析，可令 2.1 节式(1)中的 $G_1,G_2,\cdots,G_N$ 为：$G_i=(x_i,y_i)$。其中，x_i 为第 i 个疫点的经度坐标，y_i 为第 i 个疫点的纬度坐标，$z_j(m)=(cx_i,cy_i)$ 为聚类中心。其中，cx_i 为第 i 个聚类中心的经度坐标，cy_i 为第 i 个聚类中心的纬度坐标。

(2)对于 2.1 节式(1)的 $\| \cdot \|_d$，在不需要严格要求，即不必考虑地球曲率和坐标投影系影响的情况下，空间点与点间距离可以改为如下算式：

$$\| G_i-z_j(m) \|_d=\sqrt{(x_i-cx_j)^2+(y_i-cy_j)^2} \tag{3}$$

于是，将各个疫点的坐标代入式(3)，按照本文 2.1 节给出的 K-Means 聚类算法，可得到对各个疫点的聚类结果。

本文利用 Google Map 平台，将得到空间聚类分析结果可视化，图 1 中为疫点的分布情况，图 2 为 K-Means 空间聚类分析的结果，这里分类数 $K=4$，各分类用不同颜色标号的图标标注。

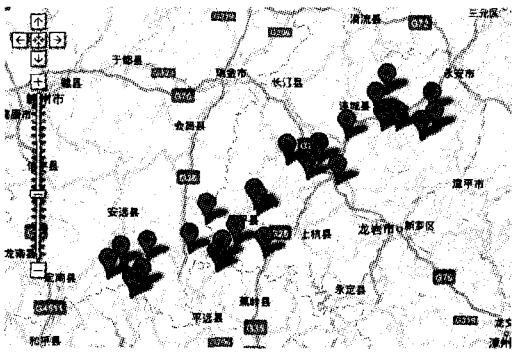

图 1 疫点空间分布

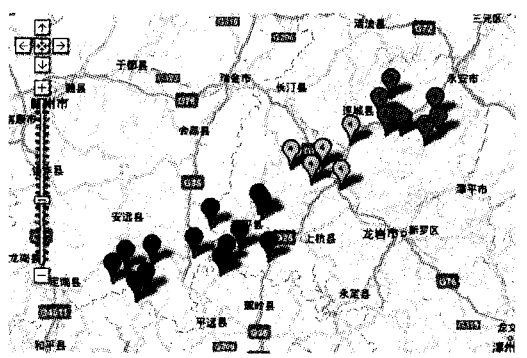

图 2 疫点空间聚类结果

3 基于最大夹角的疫点分类区域边界确定方法

3.1 基于最大夹角的聚类结果边界点的确定方法

如图 2 所示，K-Means 空间聚类分析只是解决了疫点的分类问题，下一步要确定每一个分类的区域边界。这里，首先需要确定的是各个分类离散点的边界点。黄先锋等人提出了一种基于三角形边长比约束的离散点边界追踪算法[15]；袁满等人提出一种基于行列法离散点边界搜索算法[16]。这些算法适用于海量的离散点，而且离散点分布比较均匀的情况。本文针对所涉及疫点非均匀分布的情况，提出一种基于最大夹角原理获取各疫点分类边界点的方法，其流程如下：

(1)取任意一个分类 F，即：$F=\{F(i),i=0,1,\cdots,p-1\}$，$p$ 表示该分类有 p 个疫点，F 表示二维经纬度坐标 (x,y)。取该分类中的任意点，即：F_0，找出与 F_0 最远距离的一个点记为 $F(0)$，满足 $F(0)-F_0=\max\{F(i)-F_0,i=0,1,\cdots,p-1\}$，则 $F(0)$ 为该分类中的一个边界点。

(2)F_0 与 $F(0)$ 组成向量 $v_1=F_0-F(0)$，找出与 v_1 构成

最大夹角的向量 v_2，其中 $v_2=F(i)-F(0)$，且 $i=0,1,\cdots,p-1$，满足 $\theta=\max\{\cos^{-1}(\|v_2*v_1\|/(v_2*v_1))\}$，则找到了分类的第二个边界点 $H_m(1)$。

(3)找出与向量 v_2 最大夹角的向量 v_3，重复(2)。

(4)直到最后找到的向量边界点为 $F(0)$时结束，顺序得到 j 个边界点：$F(0),F(1),\cdots,F(j-1)$，且 $F(j)\in F$。顺序连接各边界点，得到每个分类最外围点构成的区域。算法如图 3 所示，实验结果如图 4 所示。

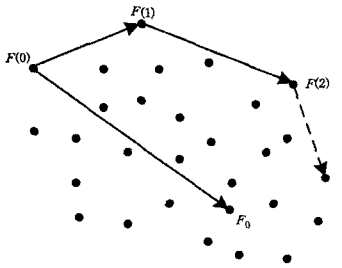

图 3　基于最大夹角原理的边界点搜索方法

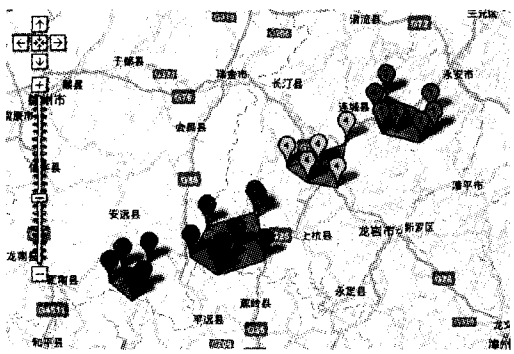

图 4　聚类边界

3.2　分类区域边界的确定

由 3.1 节可以知道，基于最大夹角原理，可以确定一个分类点集合的边界点。由于疫点的影响一般是以该点为中心、具有一定半径的圆(即疫区)。这个半径大小跟疫病有关，一般为 3～5 公里。因此，分类区域的边界应该是以分类结果中各个分类的边界点连线、向外平行扩散得到的一个与边界点连线的平行线，即分类区域的边界线。算法如下：

(1)获取分类边界点构成的多边形重心，且重心必定在边界点构成的凸多边形内。取上述 3.1 节中任意一个分类的边界点，即：$F(i)=\{(x(i),y(i)),i=0,1,\cdots,j-1\}$，其中 j 表示该分类有 j 个边界点。将以 $F(0)$为顶点的 j 边形划分解成 $j-2$ 个三角形。每个三角形面积为 $S(i)$，多边形的重心记为：

$$O=(x_c,y_c)$$

其中：

$$\begin{cases} x_c=\dfrac{1}{3}*\left(\sum_{i=1}^{i=j-2}(x(0)+x(i)+x(i+1))*S(i)\right)\Big/\sum_{i=1}^{i=j-2}S(i) \\ y_c=\dfrac{1}{3}*\left(\sum_{i=1}^{i=j-2}(y(0)+y(i)+y(i+1))*S(i)\right)\Big/\sum_{i=1}^{i=j-2}S(i) \end{cases}$$

$S(i)$为 $F(0)$、$F(i)$、$F(i+1)$3 点构成的面积。

(2)计算每条边的外点。满足条件：每条边的外点与重心构成的直线与对应边垂直，该点到对应边的距离为 r(根据疫点影响区域半径确定)，且外点在多边形外侧。由重心及每条边的端点得到每条边的外点，即：$F_o(i)=(x_o(i),y_o(i))$，其中 $i=0,1,\cdots,j-1$，计算公式略。

(3)计算过外点且与外点所对应边平行的直线，并计算相邻直线的交点，即：$F_p(i)=(x_p(i),y_p(i))$。$F_o(i)$为上述(2)得到的多边形的外点，$k(i)$为外点对应边的斜率。则：

当 $k(i)\in\emptyset$ 时，

$$\begin{cases} x_p(i)=x_o(i) \\ y_p(i)=y_o(i+1)+k(i+1)(x_o(i)-x_o(i+1)) \end{cases}$$

当 $k(i+1)\in\emptyset$ 时，同理。

当 $k(i)\notin\emptyset$ 且 $k(i+1)\notin\emptyset$ 时，得到：

$$\begin{cases} x_p(i)=(k(i)x_o(i)-k(i+1)x_o(i+1)-y_o(i)+ \\ \qquad y_o(i+1))/(k(i)-k(i+1)) \\ y_p(i)=k(i)x_p(i)-k(i)x_o(i)-y_o(i) \end{cases}$$

顺序连接 $F_p(i)$，即为分类区域边界线，如图 5 所示。

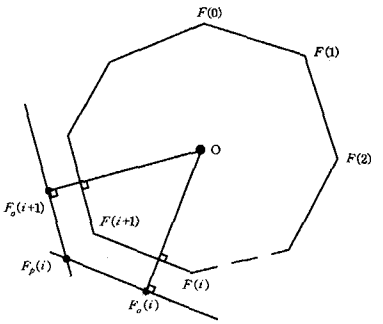

图 5　分类区域边界线

实验结果如图 6 所示，经聚类划分的 4 个分类区域中，区域内边界线为分类边界点连线，外边界线为分类区域边界线。

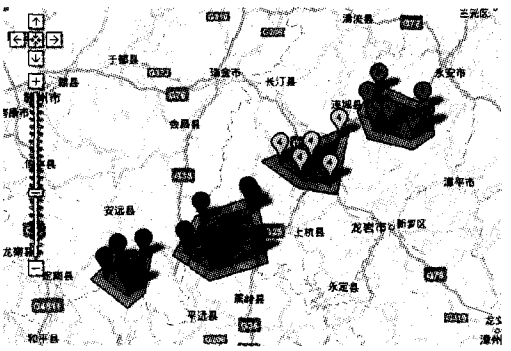

图 6　分类区域边界确定

结束语　针对如何合理有效地配置动物防疫资源的问题，本文从空间关系的角度，提出了基于 K-Means 空间聚类和最大夹角方法的疫情分类区域划分方法。该方法将疫点按照其空间分布情况进行聚类，再基于最大夹角原理，确定每一分类的边界节点，进而得到分类区域边界，从而为动物疫情防控部门时时查看疫点的空间分布模式、合理有效地配置防控资源、有效地预防与控制疫情的扩散、提高疫病防治管理水平，提供了技术手段。

作为进一步研究的方向，可以在疫病的传播、山地、河流以及行政区划等影响下的疫点的空间聚类分析方法及分类区域边界的模糊化处理方法等方面开展研究。

参 考 文 献

[1]　中国动物卫生与流行病学中心国际兽医事务处，2012 年 3－4 月全球重大动物疫情综述[J]. 中国动物检疫，2012，29(5)

[2]　姚海潮，曾江勇，张成福. 浅议禽流感的危害及防控[J]. 西藏科技，2008(6)

[3] 疫情动态[J]. 中国动物保健,2012(5)
[4] 王丽萍,徐友富,王建军,等. 1990－2006 年安徽疟疾疫情时空分布特点研究[J]. 疾病控制杂志,2008,12(2)
[5] 叶敏,李晓松,殷菲. 2004 年我国麻疹发病情况分析[J]. 现代预防医学,2008,35(8)
[6] 武继磊,王劲峰,孟斌,等. 2003 年北京市 SARS 疫情空间相关性分析[J]. 浙江大学学报,2005,31(1)
[7] MacIntosh V H. Enhancing Influenza Surveillance Using Electronic Surveillance System for the Early Notification of Community-Based Epidemics[C]//NTIS. 2004:1-17
[8] Xiang Ji,Chun S A,Geller J. Epidemic outbreak and spread detection system based on twitter data[C]//ICHIS. 2012:152-163
[9] 赵红蕊,唐中实. 基于图像空间聚类的滤波技术[J]. 计算机应用,2006,26(11)
[10] 王海军,张德礼. 基于空间聚类的城镇土地定级方法研究[J]. 武汉大学学报,2006,31(7)
[11] 梅新,崔伟宏,高飞,等. 基于空间聚类的物流配送决策研究[J]. 武汉大学学报,2008,33(4)
[12] Figuera C,Lillo J M,Mora-Jiménez I,et al. Multivariate spatial clustering of traffic accidents for local profiling of risk factors [C]//14th International IEEE Conference on Intelligent Transportation Systems(ITSC). 2011
[13] Cornélis B. Framing Spatial Decision-Making and Disaster Management in Time, Geo-information for Disaster Management [M]. Springer,2005,Part 3:281-293
[14] 郭茂耘. 航天发射安全控制决策的空间信息分析与处理研究[D]. 重庆:重庆大学,2011
[15] 黄先锋,程晓光,张帆,等. 基于边长比约束的离散点准确边界追踪算法[J]. 武汉大学学报,2009,34(6)
[16] 袁满,袁志华. 一种基于行列法离散点边界搜索算法[J]. 计算机应用研究,2010,27(1)

(上接第 28 页)

Step4 将获得的初值赋给 LVQ 算法。

利用免疫克隆算法对初值进行选择,将选择结果赋值给 LVQ 作为其初始权值。论文结合免疫克隆聚类算法对 LVQ 聚类算法进行改进,改进后的算法简称为 ICALVQ 算法。ICALVQ 算法的流程图如图 3 所示。

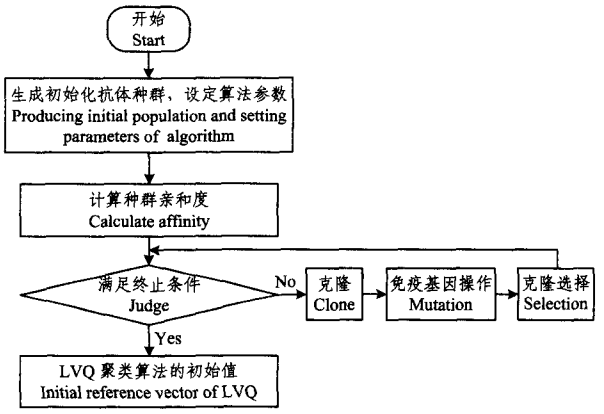

图 3 ICALVQ 算法流程图

4 实验结果

为了验证 ICALVQ 算法的有效性,本文选择 UCI 数据集中的 IRIS 数据进行测试。IRIS 数据集包含 150 个数据组,分为 3 类,每类 50 个数据,每个数据包含 4 个属性,是在数据挖掘、数据分类中非常常用的测试集、训练集。

设定免疫克隆算法的克隆规模 q 为定值 5,LVQ 算法的学习次数为 500。分别采用文献[5]中的初始权值(3 个聚类中心的值均为 0 或 9)以及接近聚类中心的初始权值进行试验,测试结果分别见表 1 和表 2。

由表 1 表 2 看出,当选择的初值远离聚类中心时,传统 LVQ 算法学习过程中只有 v_1 得到了调整,v_2 和 v_3 由于未成为获胜神经元而始终没有得到调整,只有当选择的初值接近聚类中心时,传统的 LVQ 最终学习后得到的聚类中心才较为合理。而 ICALVQ 算法则始终保持稳定性,很好地克服了 LVQ 算法所存在的对初值敏感的缺点。

表 1 初始权值远离聚类中心时两种算法最终结果的比较

初值	LVQ 算法的聚类中心	ICALVQ 算法的聚类中心
v_1=(0 0 0 0) v_2=(0 0 0 0) v_3=(0 0 0 0)	v_1=(5.836 3.057 3.752 1.201) v_2=(0.0 0.0 0.0 0.0) v_3=(0.0 0.0 0.0 0.0)	v_1=(6.599 2.985 5.564 2.037) v_2=(5.926 2.671 4.351 1.406) v_3=(5.006 3.415 1.472 0.252)
v_1=(9 9 9 9) v_2=(9 9 9 9) v_3=(9 9 9 9)	v_1=(5.836 3.057 3.752 1.201) v_2=(9.0 9.0 9.0 9.0) v_3=(9.0 9.0 9.0 9.0)	v_1=(6.599 2.985 5.564 2.037) v_2=(5.926 2.671 4.351 1.406) v_3=(5.006 3.415 1.472 0.252)

表 2 初始权值接近聚类中心时两种算法最终结果的比较

初值	LVQ 算法的聚类中心	ICALVQ 算法的聚类中心
v_1=(6 3 5 2) v_2=(6 3 4 1) v_3=(4 3 1 0.1)	v_1=(6.846 3.076 5.745 2.076) v_2=(5.896 2.732 4.431 1.447) v_3=(5.004 3.426 1.461 0.246)	v_1=(6.599 2.985 5.564 2.037) v_2=(5.926 2.671 4.351 1.406) v_3=(5.006 3.415 1.472 0.252)

结束语 论文首先对 LVQ 算法进行了深入的分析,基于免疫克隆算法具有较强群体搜索能力的特点,将免疫克隆算法运用到 LVQ 算法的优化问题中,以克服传统 LVQ 算法对初始权值敏感的缺点。将基于免疫克隆算法的 LVQ 新算法应用到 IRIS 数据集的分类实验上,实验结果表明,ICALVQ 算法较传统 LVQ 算法有着很强的稳定性,从而有效地克服了 LVQ 算法对初始权值敏感的问题,显示出其良好的应用前景。

参 考 文 献

[1] Jiao L C,Wang L. A novel genetic algorithm based on Immunity [J]. IEEE Transaction on Systems, Man, and Cybernetics-part A:Systems and Humans,2000,30(5):552-551
[2] 焦李成,杜海峰,刘芳,等. 免疫优化计算、学习与识别[M]. 北京:科学出版社,2006:92-99
[3] 马文萍,尚荣华,焦李成. 免疫克隆优化聚类技术[J]. 西安电子科技大学学报:自然科学版,2007,34(6):911-918
[4] 张敏灵,陈兆乾,周志华. SOM 算法、LVQ 算法及其变体综述[J]. 计算机科学,2002,29(7):97-100
[5] Pal N R,Bezdek J C,Tsao E C K. Generalized clustering networks and Kohonen's Self-organizing scheme[J]. IEEE Transactions on Neural Networks,1993,4(4):549-557