

文章编号:1672-3961(2011)02-0012-06

一种带克隆选择的粒子群动态聚类算法

范会联¹, 罗跃国², 李献礼²

(长江师范学院 1. 数学与计算机学院; 2. 网络信息中心, 重庆 408100)

摘要:针对聚类数不确定的高维、大规模数据聚类问题,提出以粒子群优化算法为基础、引入克隆选择算子的聚类分析算法。该算法利用粒子群的优化搜索机制搜索聚类中心向量,并根据适应度高低控制粒子的克隆数量和变异幅度,达到有效避免陷入局部最优的目的,并能克服传统聚类算法对初始值敏感的缺点,提高了算法的稳定性。仿真实验结果表明,该算法不仅能正确得出聚类簇数,而且聚类正确率较对比算法提高了至少7.0%。

关键词:克隆选择;粒子群;聚类有效性;亲合力

中图分类号:TP391.4 **文献标志码:**A

A dynamic clustering algorithm based on a particle swarm optimization with clonal selection

FAN Hui-lian¹, LUO Yue-guo², LI Xian-li²

(1. School of Mathematics and Computer; 2. Computer Network Information Center, Yangtze Normal University, Chongqing 408100, China)

Abstract: In order to achieve cluster analysis with high-dimensional and unknown number of clusters, a new clustering algorithm based on a particle swarm optimization algorithm(PSO) with clonal selection operator was proposed. Directed by the nature of PSO, this new algorithm could randomly search the clusters centers, and control the clone numbers and variation range by affinity. This algorithm could also avoid being trapped in local optima and could overcome being sensitive to initialization. Experimental results on benchmark clustering problems showed that this new algorithm could adaptively determine the amount and the center's positions of clustering. The results also showed that the average correct rate of the new algorithm was higher than the compared algorithm by at least 7.0%.

Key words: clonal selection; particle swarm; clustering validity; affinity

0 引言

作为非监督模式识别的一个重要分支,聚类分析是依据某种相似性测度规则将一个没有类别标记的数据集划分成若干个具有类别标识的聚类簇,使同一聚类簇中的对象具有较高的相似度,不同聚类簇中的对象具有较高的相异性^[1]。因此,在确定相似性测度规则后,聚类分析需解决两个关键问题:数据集中存在多少聚类簇及寻找这些簇的中心

位置。

在聚类分析的研究过程中,传统的经典聚类算法,如 k -means 聚类算法、FCM(Fuzzy C-means)聚类算法等,在获得成功应用的同时,也存在一些问题,如对初始值敏感、容易陷入到局部最优点、算法伸缩性较差、无法处理大规模数据等。为解决这些问题,人们将智能计算引入聚类研究领域,提出了一系列基于进化计算的聚类算法,如基于遗传算法的聚类方法^[2-3]、基于蚁群系统的聚类算法^[4]、基于免疫网络的聚类算法^[5-6]等。这些聚类方法在处理大

收稿日期:2010-02-14

基金项目:重庆市教委科学技术研究资助项目(KJ091309)

作者简介:范会联(1971-),男,重庆石柱人,副教授,硕士,主要研究方向为软件工程、智能信息处理。E-mail:fhlmx@163.com

规模数据集上取得很好的成果,但是在一些复杂数据集特别是高维、聚类数不确定的情况下,对聚类算法尤其是基于智能计算的聚类算法所采用的聚类中心搜索策略提出了更高的要求^[7]。因此,在高维数据集且类别数不确定的聚类分析中,基于进化计算的聚类算法在提高搜索效率、防止陷入局部最优方面还亟待改进。

本研究将克隆选择与粒子群优化算法结合用于聚类分析,每个粒子看作是待聚类数据集的一种聚类簇划分方式,同时粒子也被视为抗体,对粒子同时进行粒子群优化和免疫进化。进化过程中,粒子按其亲和度高低成正比进行克隆、按抗体浓度成反比进行克隆抑制、按亲和力和高低成反比进行局部变异,从而达到加快搜索过程、有效防止陷入局部最优的目的。

1 粒子群聚类算法

1.1 聚类问题的描述及目标函数的改进

设待聚类的数据集 $X = \{X_1, X_2, \dots, X_n\}$, 其中 $X_i = \{x_{i1}, x_{i2}, \dots, x_{id}\}$ 为 d 维特征空间 \mathbf{R}^d 中的一个点, 聚类的目的就是要找到一个划分 $C = \{C_1, C_2, \dots, C_k\}$, 在满足

$$X = \bigcup_{i=1}^k C_i \text{ 并且 } C_i \neq \emptyset, \forall i \neq j, C_i \cap C_j = \emptyset \quad (1)$$

情况下,使得总的类内离散度和

$$J_v = \sum_{j=1}^k \sum_{X_i \in C_j} d(X_i, Y_j) \quad (2)$$

达到最小,其中 Y_j 是第 j 个聚类中心, $d(X_i, Y_j)$ 为 X_i 到对应聚类簇中心点的距离^[8], 采用不同的距离定义,可用于不同分布类型数据的聚类分析。本研究用欧氏距离对类簇分布具有球状特征的数据集进行聚类分析,以验证算法的有效性。为了改变式(2)中将每一个属性对聚类的贡献看作是相同的现状,并降低因为聚类数据属性的量纲选取不同或者取值范围差距较大给聚类分析带来的影响,本研究做了以下改进:

首先对数据集 X 中的所有数据 x_{ij} 按式(3)作归一化处理转变为 X' , 式中 $\max_{i=1}^n x_{ij}$ 表示所有 n 个数据中第 j 维的最大值。

$$x'_{ij} = \frac{x_{ij}}{\max_{i=1}^n x_{ij}}, 1 \leq i \leq n, 1 \leq j \leq d. \quad (3)$$

其次,如式(4)所示,引入加权因子以区分不同属性在聚类过程中的贡献^[9],

$$J_{cv} = \sum_{j=1}^k \sum_{X'_i \in C_j} e^{w_i} \times d(X'_i, Z_j), 1 \leq t \leq d. \quad (4)$$

其中, e^{w_i} 表示第 t 个属性的加权因子,并且满足

$$\sum_{t=1}^d w_t = 0, \text{ 即 } \prod_{t=1}^d e^{w_t} = 1. \quad (5)$$

1.2 粒子群聚类算法

粒子群优化算法^[10-12] 求解聚类问题的基本思想是把每个粒子所处位置看作是待聚类数据集 X' 的一种聚类簇划分方式,对于由 k 个类簇中心构成的 d 维数据聚类的第 i 种聚类划分方式,本研究用 $Z_i = \{z_{i1}^i, z_{i2}^i, \dots, z_{id}^i, z_{21}^i, z_{22}^i, \dots, z_{2d}^i, \dots, z_{k1}^i, z_{k2}^i, \dots, z_{kd}^i\}$ 表示第 i 个粒子,其中 $[z_{j1}^i, z_{j2}^i, \dots, z_{jd}^i]$ 表示第 i 种聚类划分方式的第 j 个类簇中心的位置。第 i 个粒子的“飞行”速度记为 $V^i = \{v_{i1}^i, v_{i2}^i, \dots, v_{id}^i, v_{21}^i, v_{22}^i, \dots, v_{2d}^i, \dots, v_{k1}^i, v_{k2}^i, \dots, v_{kd}^i\}$, 其中 $[v_{j1}^i, v_{j2}^i, \dots, v_{jd}^i]$ 表示第 i 种聚类划分方式的第 j 个类簇中心的位置在各自 d 维方向的速度。第 i 个粒子的历史最优位置记为 $p_i = \{p_{i1}^i, p_{i2}^i, \dots, p_{id}^i, p_{21}^i, p_{22}^i, \dots, p_{2d}^i, \dots, p_{k1}^i, p_{k2}^i, \dots, p_{kd}^i\}$, 整个粒子群迄今为止搜索到的最好位置记为 $g = \{g_{11}, g_{12}, \dots, g_{1d}, g_{21}, g_{22}, \dots, g_{2d}, \dots, g_{k1}, g_{k2}, \dots, g_{kd}\}$, 则粒子群聚类算法就是按式(6)和(7)不断调整“飞行”速度和更新位置以寻找类簇中心,以实现在满足式(1)的约束条件下,使目标函数(4)取得最小值的优化问题。

$$V_i(t+1) = w \cdot V_i(t) + c_1 \cdot r_1 \cdot (p_i(t) - Z_i(t)) + c_2 \cdot r_2 \cdot (g(t) - Z_i(t)), \quad (6)$$

$$Z_i(t+1) = Z_i(t) + V_i(t+1), \quad (7)$$

其中,式(6)中 w 是惯性权值,取值通常在 $0.40 \sim 0.95$, $Z_i(t)$ 和 $V_i(t)$ 分别表示第 i 个粒子在 t 时刻的位置和速度, c_1 和 c_2 为加速因子, r_1 和 r_2 是取值在 $[0, 1]$ 范围内的两个随机数, $p_i(t)$ 是第 i 个粒子到 t 时刻所找到的最佳适应值对应的位置, $g(t)$ 是粒子群到 t 时刻所找到的最佳适应值对应的位置。

1.3 初始类簇中心及聚类簇数的选取

初始类簇中心从数据集 X' 中随机选取,而聚类簇数则采用以下方法动态确定:从聚类簇数 $k=2$ 开始,依次增加给定的类簇数目,直到 $k \leq \sqrt{n}$ (n 为数据集中样本数)^[13], 依次用算法求出各种 k 值对应的类簇中心,再根据文献[14]提出的聚类效果判断指标(式(8))计算聚类有效性,确定最满足指标的聚类簇数。算法在增加类簇数的过程中,采用将前次聚类结果中类内离散度最大的类簇先拆分成2个类,再连同其它类簇中心组成新的初始类簇中心。

$$I(k) = \left(\frac{1}{k} \times \frac{E_1}{J_{cw}} \times D_k \right)^r, \quad (8)$$

其中, k 为聚类簇数,指数 $r \geq 1$, 算法中取 2 ; E_1 对于给定的数据集是一个常数; J_{cw} 按式(4)计算,表示各

类内的点与聚类中心的距离的加权和; $D_k = \max_{i,j=1}^k \|C_i - C_j\|$ 表示各聚类中心两两之间的最大距离。好的分类应该是类内高内聚(类内距离尽可能小),类间低耦合(类间距离尽可能大),因此,式(8)的 $I(k)$ 值越大说明聚类效果越好^[15]。

2 带克隆选择的粒子群动态聚类算法

在粒子群聚类算法的基础上,把待分类的数据对象视为抗原,每个粒子视为抗体,把表示粒子位置的每一个 d 维向量看成抗体的 1 个基因片段,则每个抗体由 k 个这样的基因片段构成(k 为聚类簇数)。设第 m 次迭代得到种群规模为 n 的抗体种群 $A(m) = \{A_1(m), A_2(m), \dots, A_n(m)\}$, 那么该种群在克隆选择作用下,其演化过程如图 1 所示^[16]。

$$A(m) \xrightarrow{\text{克隆}} n \times A(m) \xrightarrow{\text{变异}} Z(m) \cup A(m) \xrightarrow{\text{选择}} A(m+1)$$

图 1 克隆选择的演化过程

Fig. 1 The process of clonal selection

2.1 带克隆选择的粒子群聚类算法基本思想

首先产生 N 个粒子(分别对应一种聚类划分方式),在粒子群算法的作用下,调整这 N 个粒子的位置,计算相应的适应度,然后对这 N 个粒子根据其抗体-抗原亲合力及抗体-抗体相似度进行不同数量的克隆,实现搜索空间的扩张;克隆得到的抗体在经过针对基因的变异操作后,与原抗体(克隆前的抗体)比较各自的抗体-抗原亲合力,通过克隆选择保留亲合力最高的粒子(达到种群空间压缩的目的),进入下一次迭代,直到最后产生出可以捕获抗原的最佳抗体或者达到指定迭代次数为止。

算法一方面利用粒子群算法引导搜索方向,另一方面对粒子群算法每次迭代产生的结果进行克隆,将粒子群算法的搜索结果扩张到更大的种群空间,通过对部分基因进行不同程度的变异实现更为精细的局部搜索,然后经过选择,将搜索结果再压缩到原种群空间大小,从而保证算法具有良好的全局搜索和局部搜索性能。

2.2 编码方案

算法中的抗体就是一个粒子,采用基于聚类中心的实数编码方案,根据各自取值范围,将其量化值编码成串,第 i 个粒子代表第 i 种聚类中心的划分方式为 $Z^i = \{Z_1^i, Z_2^i, \dots, Z_k^i\}$, 其中 k 为聚类个数, $Z_m^i = \{z_{m1}^i, z_{m2}^i, \dots, z_{md}^i\}$ 表示第 i 种聚类方式的第 m 个聚类

中心的位置, d 为向量维数。

2.3 抗体-抗原亲合力函数

由式(4)可知,聚类目标函数值越小,则聚类效果越好,而此时抗体-抗原亲合力应该越大。因此,借助目标函数构建抗体-抗原亲合力,第 i 种聚类划分(对应第 i 个粒子的位置)的抗体-抗原亲合力函数为

$$f(Z^i) = M/J_{cw}, \quad (9)$$

其中, J_{cw} 按式(4)计算, M 是抗体-抗原亲合力系数,主要用于使 $f(Z)$ 的值介于区间 $[0, 1]$ 之间,并且不致于太大和太小。对于给定的数据集 M 是一个常数。

2.4 抗体-抗体相似度函数及抗体浓度

抗体-抗体相似度是抗体与抗体之间的相似程度,在本算法中,每个粒子代表一个抗体,故抗体-抗体相似度指粒子与粒子间各自所代表的聚类划分方式的相似程度。粒子 i 与粒子 j 的相似度定义为

$$S(i, j) = e^{-d(Z^i, Z^j)}, \quad (10)$$

其中 $d(Z^i, Z^j)$ 表示粒子 i 与 j 的欧氏距离,距离越近,相似度越大。

在抗体-抗体相似度定义的基础上,用抗体浓度表示抗体在整个抗体群中的比例,抗体 i 的浓度 C_i 定义如下^[17]:

$$C_i = \frac{\sum_{j=1}^N S_{ij}}{N}, \quad (11)$$

其中 N 为抗体种群的大小, S_{ij} 定义如式(12),它表示当抗体 i 与 j 的相似度除以抗体 i 与当前抗体群中抗体的最大相似度的商如果大于抗体相似性系数的预设阈值 η 时取值为 1, 否则取值为 0。

$$S_{ij} = \begin{cases} 1, & \frac{S(i, j)}{\max_{v \in N} \{S(i, v)\}} > \eta; \\ 0, & \text{else.} \end{cases} \quad (12)$$

C_i 的值越大,表示该类抗体在抗体群中的比例越大,在克隆时应加以抑制。

3 克隆选择的关键问题

3.1 克隆操作中的克隆规模

本算法通过克隆操作实现抗体种群空间的扩张,为产生新的抗体提供基础^[18]。针对不同抗体,依据其抗体-抗原亲合力、抗体-抗体相似度以及抗体浓度计算该抗体的克隆数量,基本思想是:抗体-抗原亲合力高的抗体克隆数量多,同时兼顾抗体的多样性,引入抗体浓度以抑制过度克隆,避免早熟收

敛。抗体 i 的克隆数量计算式为

$$q_i = \text{Int}[M_c \times f(\mathbf{Z}^i) \times (1 - C_i)], \quad (13)$$

其中 M_c 为设定的克隆上限,即最大克隆数量, $f(\mathbf{Z}^i)$ 按式(9)计算, C_i 按式(11)计算。 $\text{Int}[*]$ 表示向上取整数。

3.2 基因变异操作

虽然粒子群算法本身具有引导搜索方向的作用,但粒子群算法在对粒子位置更新的时候是针对该粒子位置的所有矢量方向都更新,容易跳过更优或者最优位置,故本算法中基因变异操作是对粒子群算法每次迭代后新的粒子(即抗体)的部分基因片段(每个基因片段对应一个聚类中心)按式(14)进行变异操作,从而有效加大了对该粒子当前位置的局部搜索。第 i 个粒子的第 m 个聚类中心 \mathbf{Z}_m^i 的变异操作定义如下:

$$\mathbf{Z}_m^{i'} = \begin{cases} \mathbf{Z}_m^i + \lambda \times \left(1 - \frac{t}{T}\right) \times e^{-f(\mathbf{Z}_m^i)} \times \text{Rand}, & r_{im} \geq p_i; \\ \mathbf{Z}_m^i, & \text{else。} \end{cases} \quad (14)$$

$$p_i = \frac{f(\mathbf{Z}^i)}{\max_{k \in N}(f(\mathbf{Z}^k))}。 \quad (15)$$

其中, λ 为控制变异幅度的系数, Rand 是取值在 $[-1, 1]$ 的 d 维随机向量, t 表示当前迭代次数, T 为最大迭代次数, r_{im} 是针对第 i 个粒子的第 m 个聚类中心产生的一个取值在 $[0, 1]$ 的随机数,式(15)定义了第 i 个粒子的每个基因片段发生变异的概率 p_i , $f(\mathbf{Z}^i)$ 按式(9)计算。从式(14)、(15)可以看出,一方面抗体发生变异的基因片断的多少受抗体-抗原亲合力大小的影响,亲合力越大,变异的机会越少,故发生变异的基因片段越少;另一方面,变异幅度的大小受抗体-抗原亲合力和迭代次数的控制,抗体-抗原亲合力 $f(\mathbf{Z}^i)$ 越大,则变异幅度越小;随着迭代次数的不断增加,变异幅度也逐渐减小。

3.3 选择操作

对于任意粒子 i ,若该粒子的克隆群体中,具最大抗体-抗原亲合力的粒子 i' 的亲合力 $f(\mathbf{Z}^{i'}) > f(\mathbf{Z}^i)$,则用粒子 i' 代替 i 参与下一次粒子群进化。

3.4 动态聚类算法的实现

算法采取针对待聚类数据设置聚类簇数 $k = 2$ 递增到 $k \leq \sqrt{n}$,分别寻找各自的最佳类簇中心,最后通过每种 k 值下的聚类效果指标 $I(k)$ 确定聚类数和相应的聚类中心。算法实现步骤如下:

初始化粒子群大小 N ,最大迭代次数 T ,变异幅度系数 λ ,抗体相似性系数 η 等参数;
读入待聚类数集 \mathbf{X} 并按式(3)归一化为 $\mathbf{X}' = \{\mathbf{X}'_1,$

$\mathbf{X}'_2, \dots, \mathbf{X}'_n\}$;

初始聚类簇数 $k = 2$;

While($k \leq \sqrt{n}$)

初始化粒子群的中每个粒子的位置 $\mathbf{Z}^i = \{\mathbf{Z}_1^i, \mathbf{Z}_2^i, \dots, \mathbf{Z}_k^i\}$ 与速度 $\mathbf{V}^i = \{\mathbf{V}_1^i, \mathbf{V}_2^i, \dots, \mathbf{V}_k^i\}$;

While(当前迭代次数 $t < T$)

For $i = 1$ to 粒子数 N

按最小距离原则将 \mathbf{X}' 中的所有矢量 \mathbf{X}'_j 分到某一个聚类中心 \mathbf{Z}_j^i 代表的类簇中;

按式(9)计算每个粒子(即抗体)的适应值(即抗体-抗原亲合力);

按式(13)计算克隆数量,对粒子进行克隆;

按式(14)进行基因变异操作;

克隆选择;

更新每个粒子的当前最优解 \mathbf{p}_i ;

更新群体当前最优解 \mathbf{g} ;

按式(6)更新粒子的速度;

按式(7)更新粒子的位置;

endfor

endwhile

按式(8)计算聚类效果的指标 $I(k)$;

增加聚类数, $k++$;

endWhile

聚类效果指标 $I(k)$ 取得最大值的 k 作为聚类数,对应的分类结果就是数据集 \mathbf{X} 的最佳聚类结果。

4 仿真实验与分析

为了测试本研究提出算法的有效性,从 UCI (University of California) 机器学习库中选取 Iris、Wine 两个数据集,测试算法的动态聚类能力,并将本研究的算法和改进的模糊聚类算法(FCM)^[19]、基于粒子群的模糊聚类算法(particle swarm optimization based fuzzy C-means, PSOFM)^[8]以及基于克隆选择的模糊聚类算法(clonal selection based fuzzy C-means, CSFCM)^[15]在聚类正确率方面进行比较。

4.1 参数设置

算法在保持最大迭代次数不变的情况下,粒子数量随测试数据集的样本个数及属性个数增加而增加。测试中不考虑属性的重要性,故式(4) w_i 的值设为 0。抗体-抗原亲合力系数 M 的值设为大于 J_{cw} ,目的是使式(9)中 $f(\mathbf{Z})$ 的值介于区间 $[0, 1]$ 之间,并且不致于太大和太小。最大克隆数量 M_c 及变异幅度 λ 越大,则搜索的空间越大,但算法的时

间开销越大。抗体相似性系数 η 影响的是种群中粒子的相似程度,值越大,则放宽对相似程度的约

束。算法具体参数设置见表1。

表1 算法参数设置
Table 1 Parameters setting for different data sets

数据集	种群大小 N	最大迭代次数 T	惯性权值 w	加速因子 c_1, c_2	式(4)中的 w_i	式(8)中的 E_1	式(8)中的 r	式(9)中的 M	克隆数量 M_c	变异幅度 λ	相似性系数 η
Iris	10	50	0.65	1	0	100	2	15	10	1	0.7
Wine	30	50	0.65	1	0	1 000	2	50	10	1	0.7

4.2 实验结果及分析

本研究提出的算法在针对 Iris 和 Wine 数据集进行动态聚类的测试结果如表2所示。

表2 数据集在聚类簇 k 取不同值时 $I(k)$ 的部分对应值
Table 2 Comparison of partial $I(k)$ with different k values

数据集	聚类簇数为 k 时分别对应的 $I(k)$ 值				
	2	3	4	5	6
Iris	1.185	1.786	1.451	1.029	0.925
Wine	1.292	2.338	1.684	1.326	0.698

Iris 是包含 150 个植物的真实数据集,分成 3 类、每个数据项 4 个属性,其中 Iris-setosa 和 Iris-versicolor、Iris-virginical 是线性分离的,Iris-versicolor、Iris-virginical 之间存在一定的重叠。Wine 数据集有 178 个数据,分成 3 类,每个数据项有 14 个属性,各个属性的取值差距较大,尤其是 Proline 属性的聚值的聚值范围是 [278 ~ 1 680],如果不按本研究的方法做预处理,则该属性对聚类结果有很大影响。

本研究的算法与其他算法在运行 10 次后得到的聚类正确率的平均值如表3所示,其中聚类正确率定义为

$$\text{正确率} = \frac{\text{被分配到正确聚类簇的数据个数}}{\text{总的的数据个数}} \times 100\%$$

表3 四种算法的聚类结果比较

Table 3 Comparison of the clustering results of different algorithms

数据集	聚类算法	正确率/%		
		最低	最高	平均
Iris	本算法	96.0	96.0	96.0
	FCM	89.0	89.0	89.0
	PSOFCM	89.0	89.0	89.0
	CSFCM	89.0	89.0	89.0
Wine	本算法	92.1	93.3	92.7
	FCM	78.6	84.5	82.1
	PSOFCM	81.2	88.3	84.8
	CSFCM	82.5	88.9	85.2

从表2和表3可以看出,根据式(8)的聚类有效性判断规则,本研究提出的算法可以正确得出 Iirs 和 Wine 数据集的最佳聚类数 3,并且该算法针对 Iris 数据集的聚类正确率保持在 96.0%,较其他

算法得出的 89.0% 的正确率提高了 7.0%;针对 Wine 数据集的聚类正确率最低达到 90.1%,较其它算法均有显著提高。另一方面,本研究的算法在对这两个数据集的 10 次测试中,最低正确率与最高正确率的非常接近,说明了本研究的算法具有很好的稳定性。

5 结论

聚类问题在一定条件下可以归结为一个带约束的优化问题。本研究将进化理论运用于聚类问题求解,在改进目标函数的基础上,结合克隆选择机理,构造了针对数据聚类分析的粒子群动态聚类算法。该算法一方面利用粒子群算法的优化搜索机制在数据集中有指导地随机搜索聚类中心向量;另一方面,通过克隆算子在同一粒子周围的多个方向进行全局或局部搜索,促使种群中的粒子快速进化,从而使该算法在解决高维数据的聚类问题时,不仅克服了传统聚类算法对初始值敏感的缺点,而且能有效地避免陷入局部最优,提高聚类算法的稳定性及可靠性。今后我们将进一步研究改进目标函数,并将该算法运用到文本聚类和类内模式为非球状散布的聚类分析问题中。

参考文献:

[1] 武佳薇,李雄飞,孙涛,等. 邻域平衡密度聚类算法[J]. 计算机研究与发展,2010,47(6):1045-1052.
WU Jiawei, LI Xiongfei, SUN Tao, et al. A density-based clustering algorithm concerning neighborhood balance [J]. Journal of Computer Research and Development, 2010, 47(6):1045-1052.

[2] TSENG L, YANG S. A genetic approach to the automatic clustering problem [J]. Pattern Recognition, 2001, 34(2):415-424.

[3] UJJWAL M, SANGHAMITRA B. Genetic algorithm-based clustering technique [J]. Pattern Recognition, 2000, 33(9):1455-1465.

[4] 沙露,鲍培明,李尼格. 基于蚁群系统的聚类算法研究[J]. 山东大学学报:工学版,2010,40(3):13-18.

- SHA Lu, BAO Peiming, LI Nige. The research of a clustering algorithm based on the ant colony system[J]. Journal of Shandong University: Engineering Science, 2010, 40(3):13-18.
- [5] 公茂果, 焦李成, 马文萍, 等. 基于流形距离的人工免疫无监督分类与识别算法[J]. 自动化学报, 2008, 34(3):367-375.
- GONG Maoguo, JIAO Licheng, MA Wenping, et al. Unsupervised classification and recognition using an artificial immune system based on manifold distance[J]. Acta Automatic Sinica, 2008, 34(3):367-375.
- [6] FERREIRA C. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex Systems, 2001, 13(2):87-129.
- [7] 刘若辰, 沈正春, 贾建, 等. 基于免疫优势的克隆选择聚类算法[J]. 电子学报, 2010, 38(4):960-964.
- LIU Ruochen, SHEN Zhengchun, JIA Jian, et al. Immuno domain based clonal selection clustering algorithm [J]. Acta Electronica Sinica, 2010, 38(4):960-964.
- [8] 刘琼, 罗可. 一种带变异操作的粒子群聚类算法[J]. 计算机工程与应用, 2010, 46(16):130-132.
- LIU Qiong, LUO Ke. Clustering algorithm based on particle swarm optimization with mutation[J]. Computer Engineering and Applications, 2010, 46(16):130-132.
- [9] 唐成龙, 王石刚, 徐威. 基于数据加权策略的模糊聚类改进算法[J]. 电子与信息学报, 2010, 32(6):1277-1283.
- TANG Chenglong, WANG Shigang, XU Wei. Improved fuzzy clustering algorithm based on data weighted approach[J]. Journal of Electronics & Information Technology, 2010, 32(6):1277-1283.
- [10] 吕艳萍, 李绍滋, 陈水利, 等. 自适应扩散混合变异机制微粒群算法[J]. 软件学报, 2007, 18(11):2740-2751.
- LÜ Yanping, LI Shaozi, CHEN Shuili, et al. Particle swarm optimization based on adaptive diffusion and hybrid mutation[J]. Journal of Software, 2007, 18(11):2740-2751.
- [11] LIANG J J, QIN A K. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions[J]. IEEE Trans on Evolutionary Computation, 2006, 10(3):281-295.
- [12] 刘建华, 黄添强, 严晓明. 融合 PSO 算法思想的进化算法[J]. 山东大学学报:工学版, 2010, 40(5):34-40.
- LIU Jianhua, HUANG Tianqiang, YAN Xiaoming. Evolutionary algorithm based on the idea of particle swarm optimization[J]. Journal of Shandong University: Engineering Science, 2010, 40(5):34-40.
- [13] 于剑, 程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学: E 辑, 2002, 32(2):274-280.
- YU Jian, CHEN Qiansheng. Optimal number of clusters in fuzzy clustering [J]. Science in China: Series E, 2002, 32(2):274-280.
- [14] MAULIK U, BANDYOPADHYAY S. Performance evaluation of some clustering algorithms and validity indices [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(12):1650-1654.
- [15] 张旭, 郭晨. 基于克隆选择的快速动态聚类算法[J]. 计算机工程, 2007, 33(23):16-18.
- ZHANG Xu, GUO Chen. Fast dynamic clustering algorithm based on clone selection[J]. Computer Engineering, 2007, 33(23):16-18.
- [16] 刘琼, 吴小俊. 一种改进的免疫克隆选择算法[J]. 山东大学学报:工学版, 2009, 39(6):8-13.
- LIU Qiong, WU Xiaojun. An improved immune clonal selection algorithm[J]. Journal of Shandong University: Engineering Science, 2009, 39(6):8-13.
- [17] 于宗艳, 王丽. 免疫优化算法及应用[J]. 计算机仿真, 2009, 26(12):154-156.
- YU Zongyan, WANG Li. An immune optimization algorithm and its application[J]. Computer Simulation, 2009, 26(12):154-156.
- [18] De CASTRO L N, Von ZUBEN F J. Learning and optimization using the clonal selection principle[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(3):239-250.
- [19] PAKHIRA M K, BANDYOPADHYAY A, MAULIK U. Validity index for crisp and fuzzy clustering[J]. Pattern Recognition, 2004, 37(3):487-501.

(编辑:胡春霞)