基于混合生物地理学优化的聚类算法①

温肖谦, 黄发良, 李超雄, 汪 焱

(福建师范大学 软件学院, 福州 350100)

摘 要: 聚类分析是数据挖掘的重要任务之一, 而具有易早熟与收敛速度慢等缺陷的传统生物地理优化算法(Biogeography-Based Optimization, BBO)很难满足具有 NP(Non-deterministic Polynomial)性质的复杂聚类问题需求, 于是提出了一种基于混合生物地理学优化的聚类算法(Mixed Biogeography-Based Optimization, MBBO), 该算法构造了一个基于梯度下降局部最优贪心搜索的新迁移算子, 以聚类目标函数值作为个体适应度进行数据集内隐簇结构寻优. 通过在 4 个标准数据集(Iris、Wine、Glass 与 Diabetes)的实验, 结果表明 MBBO 算法相对于传统的优化算法具有更好的优化能力和收敛度, 能发现更高质量的簇结构模式.

关键词: 生物地理学优化算法; 局部优化; 数据挖掘; 聚类

Mixed Biogeography-Based Optimization Algorithm for Data Clustering

WEN Xiao-Qian, HUANG Fa-Liang, LI Chao-Xiong, WANG Yan

(School of Software, Fujian Normal University, Fuzhou 350100, China)

Abstract: Cluster analysis is an important task of data mining, however, traditional biogeography-based optimization algorithm with limitations such as prematurity and poor convergence can not satisfy the demands of solving the NP (Non-deterministic Polynomial) clustering problem. A novel algorithm Mixed Biogeography-Based Optimization (MBBO) is proposed. The algorithm integrates a new migration operator, which is constructed on gradient descent local search, and uses clustering validity index as the individual fitness to optimize implicit cluster structures in datasets. Experimental results on the four benchmark datasets (Iris, Wine, Glass and Diabetes) show that MBBO algorithm outperforms the traditional optimization algorithms such as PSO, BBO, and K-means in terms of clustering validity and convergence, and can acquire the higher quality cluster structures of the datasets.

Key words: biogeography-based optimization; local optimization; data mining; clustering

聚类是将数据对象分成类或者簇的过程, 使得簇内对象间的相似度尽可能高, 而簇间对象间的相似度尽可能低, 通过比较数据的相似性和差异性, 能发现数据的内在特征及分布规律, 从而获得对数据更深刻的理解与认识. 作为数据挖掘中重要的研究内容之一, 聚类分析技术受到了广泛的关注, 在机器学习、图像处理、商业决策等领域得到了广泛的应用.

聚类方法类别大致可分为划分方法、层次方法、基于模型、基于密度以及基于网格的方法. 在现有的聚类算法中, K-means 算法[1]以其简单和高效占有重要地位, 但因其在寻找聚类中心的过程中采用了启发式方法, 使得该算法对初始聚类中心的选择较为敏感, 容易陷入局部最优解. 聚类问题的计算 NP 性质启发着大量研究人员试图从超启发式优化的角度寻找数据内隐藏簇结构, 从而涌现出一系列诸如基于遗传算法、蚁群优化算法、微粒群算法等的各种进化聚类算法. Masoud 等[2]提出一种基于变长编码的 GA 算法并运用该算法进行聚类, Cui 等[3]提出利用 PSO 算法进行聚类, 通过实验证明聚类效果比 K-means 等的聚类方法更理想. 这类搜索算法利用群体智能的全局优化特

① 基金项目:教育部人文社会科学研究青年基金项目(12YJCZH074);福建省教育厅科技项目(JA13077)
收稿时间:2014-11-13;收到修改稿时间:2015-01-12

性能寻找到较好的全局最优解, 但相比于基于固定模板的搜索算法计算复杂度高.

近年来, 生物地理优化算法[4]作为一种新的群体智能优化方法备受关注, 该算法的提出是在对生物物种迁移数学模型的研究基础上, 借鉴其他仿生智能优化算法的框架而形成的. 与其他智能算法(PSO, GA等)相比, 尽管BBO算法具有更好的全局搜索能力, 但是 BBO 算法在搜索过程中也容易因早熟而陷入局部最优. Gong 等[5]将用进化变异机制替换 BBO 算法的突变操作, 提出一种具有改进突变操作的生物地理学优化算法. Boussaid 等[6]结合差分进化算法与生物地理学优化算法, 提出一种两阶段差分生物地理学优化算法. Ma 等[7]引入多种迁移率模型, 提出一种改进迁移模型的生物地理学优化算法. Vijay 等[8]利用 K-Means 算法为 BBO 算法初始化聚类中心来提高聚类的有效性(BBOKMI). 尽管这些改进方法不同程度地提高了 BBO 的寻优能力, 但在面对复杂的数据聚类任务时仍然很难取得令人满意的效果.

为了克服算法的缺陷不足, 本文提出一种基于局部优化算法和生物地理学优化算法的混合 BBO 算法, 即: MBBO 算法, 该算法利用局部优化思想改进传统算法的迁移模型, 克服 BBO 算法因早熟而陷入局部最优解和收敛慢等问题, 提高 BBO 算法的寻优能力. 实验结果表明, MBBO 具有较强的寻优能力, 能有效提高聚类的准确性.

1　生物地理学优化算法

生物地理学数学模型是建立在由 Alfred Wallace 和 Charles Darwid 于 19 世纪提出的生物地理学基础之上的数学模型, 其核心思想是通过模拟物种的产生、灭绝与迁移等过程来实现数学问题的优化求解. 在生物生存区域内, 一个栖息地(habitat)适合生物生存的优劣程度被简称为生存适宜度(HSI). 具有较高适宜度的栖息地可以容纳较多的物种, 而具有较低适宜度的栖息地仅可以供养较少的物种. 影响栖息地 HSI 的因素有很多, 比如其宿主区域的降雨量、植被多样性、地貌特征、土地面积和温度等. 当栖息地的适宜度较高的时候, 将最终表现为同种强势物种饱和; 一直保持较低适宜度的栖息地, 可能由于某种自然灾害的发生导致这个栖息地的物种的灭绝, 此时也会导致其他物种的大量迁入.

下面以图 1 来说明栖息地物种迁移原理, 当物种的种类为 0 时, 物种的迁出率 μ 为 0, 此时迁入率 λ 达到最大; 当物种种类最大时, 则相反. 当物种的种类为 S0 时, 迁出率 μ 和迁入率 λ 相等, 此时达到动态平衡; 如果由于某个外在的因素打破了这种平衡, 则重新进行物种迁移操作, 然后经过一段时间又会达到新的平衡. 在实际应用中, 往往假设λ和μ是线性的且具有相同的最大值.

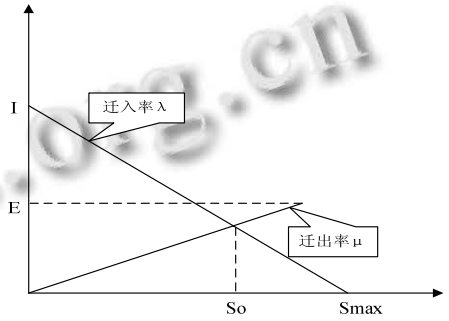


图 1　生物地理学种群迁移模型

图 1 中 I 和 E 分别表示最大迁入率和最大迁出率, S_{max} 为栖息地所能容纳的最大物种数量. 令 k 为栖息地 i 居住的种群个体数, 物种的最大数 $S_{max}=n$, 则迁出率和迁入率的演化过程可形式化为公式(1)与(2).

$$\mu_i = \frac{E \cdot k}{n} \tag{1}$$

$$\lambda_i = I(1-\frac{k}{n}) \tag{2}$$

由迁出率 μ 和迁入率 λ 可进一步推算栖息地 i 种群数量概率 P_i, 则有

$$P_i = \begin{cases} -(\lambda_i+\mu_i)P_i + \mu_{i+1}P_{i+1} & S=0 \\ -(\lambda_i+\mu_i)P_i + \lambda_{i-1}P_{i-1} + \mu_{i+1}P_{i+1} & 1 \le S \le S_{\max}-1 \\ -(\lambda_i+\mu_i)P_i + \lambda_{i-1}P_{i-1} & S=S_{\max} \end{cases} \tag{3}$$

BBO 突变操作的关键问题是如何根据柄息地的种群数量概率给出相应的突变率. 与处于稳态的栖息地相比较, 物种种类较少或者较多的栖息地更易受到栖息地降雨量、植被多样性、地质多样性和气候等外界因素的干扰而发生突变. 适宜度较高的栖息地和适宜度较低的栖息地对应的种群数量概率都较低, 平衡点(即图 1 中的 S_0)对应的种群数量概率则较高. 每个栖息地的物种数量概率表示给定特征存在的可能性, 若一个栖息地种群数量概率较低, 则该特征存在的概率较小, 如果发生突变, 它很有可能突变成更好的特征. 相反地, 具有较高种群数量概率的特征个体则具有很

小的可能性突变到其它特征. 因此, 突变概率函数与该栖息地的种群数量概率成反比. 基于此, 构造如下变异算子(公式(4)), 使得处于低适宜度栖息地的物种获得更多生存发展的机会.

$$m_i = m_{\max}(\frac{1-p_i}{p_{\max}}) \tag{4}$$

式(4)中, m_{max} 为已知定义突变率的最大值, 引入突变可以增加物种的多样性, 这个突变公式使得较低适宜度的栖息地以较大的概率发生突变, 为这个栖息地增加寻找更优解的概率. 但是, 突变公式可能会破坏较优的栖息地的特征, 所以可以在算法的迭代过程中, 保留精英个体, 使得这些适宜度较高的栖息地特征得到有效保护.

2 MBBO算法

2.1 编码方法

假设有 N 个数据样本, 每个样本数据 X_i 含有 D 个属性, 则 $X_i=(X_{i1},X_{i2},\ldots,X_{iD})$(i=1,2,…,n), 则数据样本集合为 $P=(X_1,X_2,\ldots,X_n)$. 算法将 n 个向量 X_i(i=1,2,…,n)分为 k 个组 C_j(j=1,2,…,k), n_j表示第 j 类的样本个数, 并求每组的聚类中心 $C_j=\frac{1}{n_j}\sum_{i=1}^{n_j}X_i$ (i=1,2,…,n_j). 考虑到编码长度和数据两方面的因素, 本文采用聚类中心的方式编码, C_j 代表一组聚类中心, 聚类中心的个数为 k, 每个中心有 D 维实数编码的属性, 那么编码长度就为 k·D, 编码方式如图 2 所示.

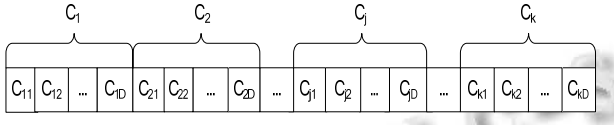


图 2 编码方式

其中, $C_j=(C_{j1}, C_{j2}, \ldots, C_{jD})$表示的是第 j 个簇的中心向量.

2.2 适应度函数

适应度函数是用来评价个体的适应度和判别群体中个体优劣的标准, 具有较高适宜度的栖息地可以容纳较多的物种, 而具有较低适宜度的栖息地仅可以供养较少的物种. 显然, 个体的适应度值本质就是对该个体所对应的数据划分方案的质量测度. 在此, 我们选择 l2 范数作为数据对象 X_i 与其相应聚类中心 C_j 的距离测度, 进而构造如下个体适应度函数:

$$f_c=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\beta_{ji}\left\|X_i-C_j\right\|^2 \tag{5}$$

$$\beta_{ji}\begin{cases}1 & \text{对于每个}m\neq j\text{时,}\left\|X_i-C_j\right\|^2\le\left\|X_i-C_m\right\|^2\\ 0 & \text{其它}\end{cases} \tag{6}$$

其中, β_{ji}为权重, 表示的是向量 X_i 与组 C_j 的相关程度. 对于聚类问题, 与中心的距离越小说明相似性越大, 即目标函数 f_c 的值越小越好.

2.3 改进的迁移操作

传统 BBO 算法采用随机轮盘选择策略来确定迁入迁出栖息地, 若将一个不合适的特征迁入到迁入率高的栖息地, 可能导致迁入率高的栖息地适宜度指数更低. 同样, 若一直都从某个高迁出率的栖息地迁出特征, 随着迁移次数增加, 可能导致种群多样性降低, 收敛速度变慢, 从而增加了陷入局部最优解的几率. 因此, 迁移特征的选择对算法寻优性能起着举足轻重的作用. 基于此, 我们引入局部优化思想对传统迁移算子进行改进, 即: 将局部优化算法搜索产生的新个体来替代较劣个体, 使其跳出局部最优, 提高优化效率.

作为局部优化算法的典型代表, 梯度下降法由于其简单高效而被广泛使用, 其核心思想是: 在求目标函数最优解时始终沿负梯度方向修正解向量. 据此可以将梯度下降搜索适应度函数中的解向量 C 的过程形式表示为 GDSearch 算法, 进而有改进的 GDS_Migrator 迁移算子.

GDSearch 算法

Step 1: 初始化迭代次数m =0, 阈值θ, α

Step 2: 求出特征X_i 所对应的适应度函数f_c (X_i)

Step 3: 迭代次数增加: m = m+1

Step 4: 沿负梯度的方向寻找下一个接近极值的特征 $X_{jk}=X_{ij}-\alpha\nabla f_c(X_i)$

Step 5: 若有$|\alpha\nabla f_c(X_i)|<\theta$, 则有$X_{ij}=X_{jk}$, $f_c(X_i)=f_c(X_i')$, 否则转到Step 3.

GDS_Migrator算子

Step 1: For (i =1 to 种群大小)

Step 2: For (j =1 to 特征维数)

Step 3: If rand < λ

Step 4: 选择需要改变X_i的特征X_{ik}

Step 5: If rand < μ

Step 6: X_{ik}=GDSearch(X_{ik})

Step 7: End

Step 8:　　End

Step 9:　　End

Step 10:End

2.4 算法描述

综上所述, 以 BBO 算法为基本计算框架, 基于混合生物地理优化的聚类算法(MBBO)流程可归纳为以下步骤:

Step1. 初始化基本参数, 设置栖息地数量 n、每个栖息地可容纳的最大物种数量 S_{max}, 设置迁入率最大值 I 和迁出率最大值 E、最大变异率 m_{max} 等.

Step2. 个体编码操作.

Step3. 利用公式(5)计算栖息地的适应值, 并计算栖息地对应的迁入率 λ 以及迁出率 μ 等.

Step4. 执行迁移操作: 根据我们改进的迁移算子, 计算迁入率和迁出率决定栖息地的个体是否进行迁移, 若是则执行迁移操作, 用新个体代替旧个体, 若否则不执行迁移操作.

Step5. 执行突变操作: 根据式(3)更新每个栖息地的种群数量概率 P. 如果满足突变条件, 则根据变异公式(4)对当前个体进行变异.

Step6. 是否满足停止条件, 如果不满足, 则跳转到 Step3, 如果满足, 则输出聚类结果.

3 实验结果与分析

本文中实验的环境为: 操作系统为 Windows 7, 处理器为 Intel Core2 Duo E3700 @ 2.66Ghz, 内存为 2GB, 实验软件为 Matlab7.0.

3.1 实验数据

本文采用 UCI 数据库中的 Iris、Wine、Glass 与 Diabetes 作为实验数据集, 从有效性、收敛性与聚类群体数目三个方面对 MBBO 算法进行性能评价. 各数据集所含数据样本个数、数据样本的属性个数、类别个数如表 1 所示.

表 1　实验数据集特征描述

数据集	样本数	属性数	类别数
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
Diabetes	768	8	2

3.2 参数设置

为了比较的有效性和公平性, 文中各个算法的初始实验参数设置为: 步长 α=0.5, 阈值 $\theta=10^{-4}$, 群体大小 NP=50, D=20, 变异率 m=0.01, 最大迁入率 I=1, 最大迁出率 E=1, 取最大迭代次数 G=100 为终止条件.

3.3 算法有效性分析

聚类评价算法包括外部评价法、内部评价法以及相对评价法三种. 本实验采用的是外部评价法中的 F-Measure 评价方法. 它组合了信息检索中的查准率与查全率的思想来进行聚类评价. F-Measure 法适用于对测试集进行聚类的聚类结果进行评价, F-Measure 值越高, 说明聚类效果越好. 为了评价不同模型的聚类性能, 我们选择了传统的 K-means 算法、PSO 算法、BBO 算法以及 BBOKMI 算法与本文提出的 MBBO 聚类算法在 4 个常用 UCI 数据集(Iris、Wine、Glass 与 Diabetes)进行实验比较, 结果如表 2 所示.

表 2　聚类有效性比较

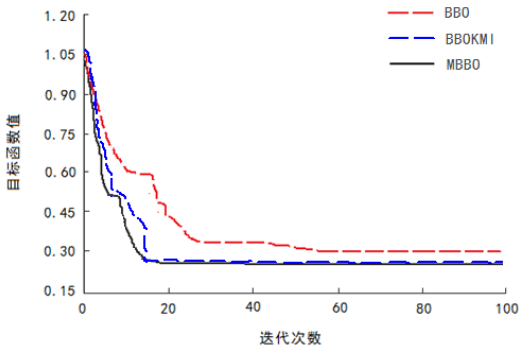
数据集	K-means	PSO	BBO	**BBOKMI**	MBBO
Iris	0.8087	0.8308	0.8436	**0.8910**	0.8997
Wine	0.6681	0.7037	0.7155	**0.7190**	0.7634
Glass	0.5244	0.5350	0.5521	**0.5586**	0.5742
Diabetes	0.5768	0.6219	0.6487	**0.6610**	0.6656

从表 2 的实验数据可以看出: 五种算法在对 UCI 数据集进行实验的时候, MBBO 算法、BBOKMI 算法和 BBO 算法在大部分情况下聚类结果 F-Measure 值要比 K-means、PSO 算法好, 提高幅度大, 效果更明显, PSO 算法的结果比 K-means 算法好. 在四个数据集上, MBBO 算法相比于 BBOKMI 算法聚类有效性均有提高, 在 Wine 数据上提高了 4.44%. 对于数据集 Iris 和 Wine, MBBO 算法相比于传统的 BBO 算法, 聚类结果分别提高了有 5.61%和 4.79%. 而对于数据集 Glass 和 Diabetes, MBBO 算法和传统的 BBO 算法相比, 有 2.21%和 1.69%的提高. 实验结果表明本算法聚类结果 F-Measure 值更高, 得到更加理想的聚类结果.

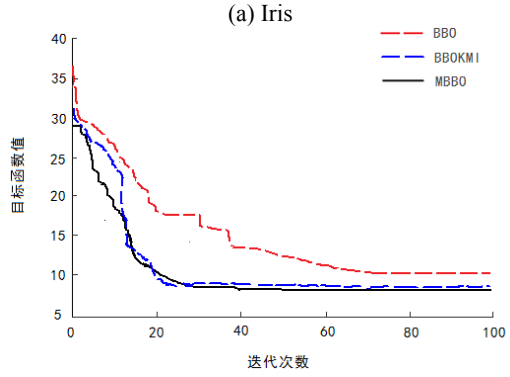
3.4 算法收敛性分析

考虑到 BBO 算法、BBOKMI 算法与 MBBO 算法的聚类有效性比较接近, 本小节进一步对这三者之间的收敛性进行实验比较, 实验结果见图 3. 由图 3 可知, 三种算法在不同数据集上表现出来的收敛性不同, 从总体上看, MBBO 算法都比 BBOKMI 算法、传统 BBO 算法具有更快的收敛速度. 结合上一小节的聚类有效性分析, 不难得出, 替换合适的的特征能够使得在保证聚类质量不降低或者更高的情况下, 有效的提高收

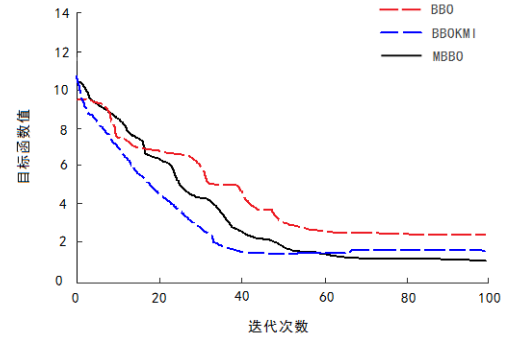
敛的速度, 并得到质量更好的聚类结果.



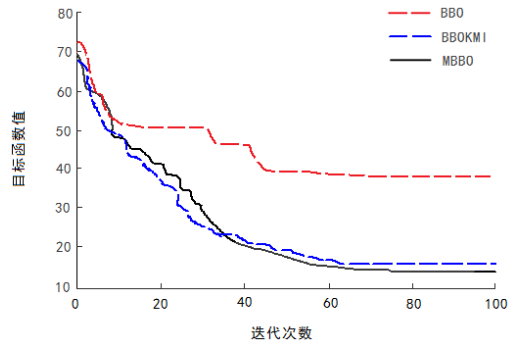
(a) Iris



(b) Wine



(c) Glass



(d) Diabetes

图 3 不同数据集的平均最佳适应度值优化图

3.5 参数对聚类有效性的影响

3.5.1 群体数目对聚类有效性的影响

为了研究聚类群体数目对聚类有效性的影响, 在本实验中, 我们将算法 MBBO 的聚类群体的数目分别设为 5、10、15、20, 依次加 5 递增, 其它参数不变. 实验在 Iris 数据集、Wine 数据集、Glass 数据集以及 Diabetes 数据集上进行, 比较不同群体数目下的差异, 其结果如图 4 所示.

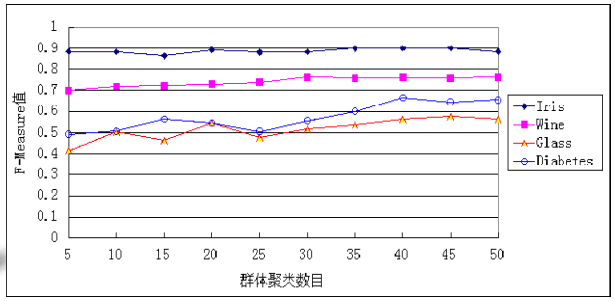


图 4 聚类群体数目对 MBBO 算法的影响

从图 4 可以看出: MBBO 算法在数据集 Iris 和 Wine 上对聚类群体的大小不敏感, 结果没有很大变化. 而数据集 Glass 和 Diabetes 在聚类群体规模小于 25 的时候, 算法不是太稳定, 结果有波动, 群体规模在 25-40 呈上升趋势, 群体规模大于 40 时呈平稳状态. 所以本算法应该应用于大小大于 40 以上的聚类群体, 以保证算法的稳定性.

3.5.2 突变率取值对聚类有效性的影响

考虑到算法中参数突变率的大小设置可能会影响算法的准确率, 我们将实验中四种不同数据集在取不同的突变率 m 情况下进行对比分析, 结果如图 5 所示.

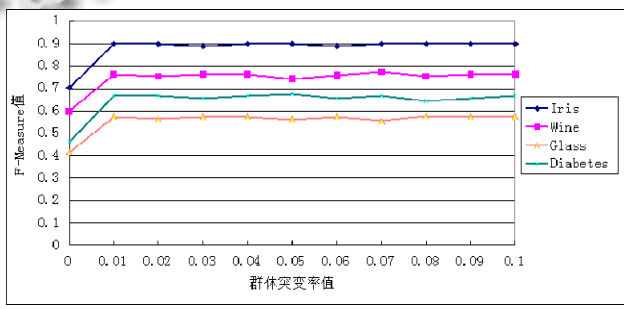


图 5 不同突变率值对算法有效性的影响

由图 5 的实验结果可以发现, 突变率的取值对于算法的 F-Measure 值的影响并不大, 但是进行突变操作, 会增加适宜度较低的栖息地的物种的多样性, 提高算法的寻优能力.

4 总结

为了提高聚类的质量, 提出了一种融入生物地理优化的聚类算法. 针对传统 BBO 算法的缺陷不足, 本文提出了一种基于 BBO 算法结合局部优化思想的聚类算法(MBBO). 该算法改善了 BBO 算法的局部最优解问题, 提高了聚类质量. 在 UCI 数据集上的实验结果证明了 MBBO 算法的有效性和收敛性. 但是该算法的稳定性等方面还有待于提高, 这些是我们下一步要研究的内容.

参考文献

1 MacQueen JB. Some methods for classification and analysis of multivariate observations. Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 281–297.

2 Masoud A, Setayeshi S, Hossaini Z. A web page classification and clustering by means of genetic algorithm a variable size page representation approach. Computational Intelligence for Modeling, 2008: 436–440.

3 Cui X, Potok T, Palathingal P. Document clustering using particle swarm optimization. Proc. of IEEE Swarm Intelligence Symposium(SIS-2005). 2005. 186–191.

4 Simon D. Biogeography-based optimization. IEEE Trans. on Evolutionary Computation , 2008, 12(6): 702–713.

5 Gong WY, Cai ZH, Ling CX, et al. A real-coded biogeography-based optimization with mutation. Applied Mathematics and Computation, 2010, 216(9): 2749–2758.

6 Boussad I, Chatterjee A, Siarry P, et al. Two-stage update biogeography-based optimization using differential evolution algorithm(DBBO). Computers and Operations Research, 2011, 38(8): 1188–1198.

7 Ma HP. An analysis of the equilibrium of migration models for biogeography-based optimization. Information Sciences, 2010, 180(18): 3444–3464.

8 Kumar1 V, Chhabra JK, Kumar D. Advances in Computing, Communication and Control. Berlin: Springer Berlin Heidelberg, 2011: 448–456.

9 Snyman JA. Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms. Springer Publishing.

10 马海平,陈子栋,潘张鑫.一类基于物种迁移优化的进化算法.控制与决策,2009,24(11):1620–1624.

11 周涛,陆惠玲.数据挖掘中聚类算法研究进展.计算机工程与应用,2012,48(12):100–111.

12 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48–61.

13 林剑,徐力.基于混合生物地理优化的混沌系统参数估计.物理学报,2013,62(3):305–305.