Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2020, 9(2), 92-99
Published Online May 2020 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2020.92011

Hans汉斯

Improved Hybrid Clustering Algorithm Based on Artificial Bee Colony Algorithm and K-Means Algorithm

Wanying Bao, Xiaoling Luo*, Xin Pan

College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hhhot Inner Mongolia
Email: *baowy1995@163.com

Received: Mar. 23rd, 2020; accepted: Apr. 30th, 2020; published: May. 7th, 2020

Abstract

In order to overcome the disadvantages of K-Means clustering algorithm, such as over dependence on the initial clustering center, easily falling into local optimum, and the premature and slow convergence of the artificial bee colony algorithm due to the limitations of search strategies, a hybrid clustering method combining the improved global artificial bee colony algorithm and K-Means++ algorithm is proposed, which makes full use of the characteristics of the improved global artificial bee colony algorithm and K-Means++ algorithm. It can optimize the location of the initial clustering center and the convergence speed is fast. By combining the two, K-Means can search globally and jump out of the local optimal solution. The experiments are carried out with the Wine data set and balance-Scale data set in the UCI database. The results show that the improved global artificial bee colony algorithm has faster convergence speed and better optimization effect than the standard artificial bee colony algorithm. Compared with the original K-Means algorithm, the hybrid clustering algorithm proposed in this paper has better stability, fewer iterations, faster convergence speed and better clustering effect.

Keywords

Global Artificial Bee Colony Algorithm, K-Means, Fitness Function, Cluster Analysis

基于人工蜂群与K-Means的改进混合聚类算法

包婉莹，罗小玲*，潘　新

内蒙古农业大学，计算机与信息工程学院，内蒙古　呼和浩特
Email: *baowy1995@163.com

*通讯作者。

文章引用: 包婉莹, 罗小玲, 潘新. 基于人工蜂群与 K-Means 的改进混合聚类算法[J]. 人工智能与机器人研究, 2020, 9(2): 92-99. DOI: 10.12677/airr.2020.92011

收稿日期：2020年3月23日；录用日期：2020年4月30日；发布日期：2020年5月7日

摘　要

为了克服K-Means聚类算法过度依赖初始聚类中心、容易陷入局部最优的缺点以及人工蜂群算法因为搜索策略的局限而导致的易早熟，收敛速度慢的问题，提出了改进的全局人工蜂群算法与K-Means++算法相结合的混合聚类方法，充分利用改进的全局人工蜂群算法可以全局寻优的特点与K-Means++算法能够优化初始聚类中心位置并且收敛速度快的特点，将二者融合，使得K-Means可以进行全局搜索，跳出局部最优解，并用UCI数据库中的Wine数据集和Balance-Scale数据集进行实验。结果表明，改进的全局人工蜂群算法较标准人工蜂群算法收敛速度更快，寻优效果更好；本文提出的混合聚类算法与原始K-Means算法相比，稳定性更好，迭代次数减少，收敛速度更快，而且聚类效果也有了明显改善。

关键词

全局人工蜂群算法，K-Means，适应度函数，聚类分析

1. 引言

数据挖掘是从大量数据中挖掘有趣模式和知识的过程，聚类分析是数据挖掘的一种重要技术手段，它将一个数据集对象或观测划分成若干子集[1]。K-Means 算法是聚类分析中一种基于划分的无监督学习算法。具有思想简单、效果好、容易实现的优点，广泛应用于机器学习等领域[2]。但其对初始中心敏感、易陷入局部最优等问题也是近年来研究者对此算法进行改进的重点。文献[3] [4] [5] [6] [7]采用改进群体智能等方法与 K-Means 混合聚类以及从优化初始中心的角度考虑改进 K-Means。人工蜂群算法(ABC)因其搜索速度较快、鲁棒性好、且易于实现的优点被广泛应用。但由于搜索策略的局限性，人工蜂群算法虽然在前期性能较好，但在后期易陷入局部解。文献[8] [9] [10]对于 ABC 算法的问题进行改进，优化蜂群的搜索策略，但算法的收敛速度偏慢。文献[11]将人工蜂群算法改进后与 K-Means 算法相结合，一定程度提高了聚类质量，但是该算法随机产生食物源，降低了种群的多样性。文献[12]为避免聚类中心可能是同类样本而采用最大最小距离积法产生初始聚类中心。本文提出了一种结合人工蜂群与 K-Means 的改进混合聚类算法，IGABC-K-Means++算法，降低了初始聚类中心的依赖性和陷入局部最优解的可能性，有效提高了算法的聚类效果和稳定性，缩短了聚类时间。

2. 相关算法简介

2.1. 原始 K-Means 算法

设样本集为 $X=\left(X_1,X_2,\cdots,X_N\right)$，其中 $X_i=X_{i1},X_{i2},\cdots,X_{iD}$ $(i=1,2,\cdots,N)$，现将这 N 个 D 维的数据聚类：

Step 1：随机选取 k 个簇中心点；

Step 2：遍历所有数据，将每个数据划分到欧式距离最近的中心点中；

Step 3：计算每个聚类的平均值，并作为新的簇中心点；

Step 4：重复 2~3，直到某个中止条件。

2.2. 人工蜂群算法

人工蜂群算法具体实现步骤：

Step1：初始化蜂群数 NP、食物源个数 SN，最大迭代次数 MCN、食物源停滞的最大次数 Limit 和确定搜索空间 D，通常有 SN = NP/2；在设置的区间内按照公式(1)随机生成 SN 个食物源并根据公式(2)计算其适应度值 fitness。

$$X_{ij} = b_j + rand(0,1)(a_i - b_j);\ (i = 1,2,3,\cdots,\mathrm{SN}) \tag{1}$$

$$fitness = \begin{cases} \dfrac{1}{1+fitness} \\ 1+abs(fitness) \end{cases} \tag{2}$$

公式(1)中，$rand(0,1)$ 表示 $(0,1)$ 之间分布均匀的随机数。a_j 和 b_j 分别表示第 j 维数据的上限和下限 ($j = 1,2,3,\cdots,D$)。

Step2：引领蜂由式(3)进行邻域搜索，产生一个候选食物源并计算 fitness，

$$V_{ij} = x_{ij} + R_{ij}(x_{mj} - x_{kj}) \tag{3}$$

公式(3)中，$k \neq i$；R_{ij} 为 $[-1,1]$ 之间的随机数。对比新解与旧解的 fitness，如果新解较好，就用新解取代旧解，否则，仍用旧解，同时将食物源的停滞次数加 1。

Step3：跟随蜂按照轮盘赌选择引领蜂，根据公式(3)更新当前位置，fitness 较大，被更新的可能越大，选择概率为公式(4)：

$$P_i = fit_i \Big/ \sum_{i=1}^{SN} fit_i;\ i = 1,2,3,\cdots,\mathrm{SN} \tag{4}$$

Step4：经过 Limit 次循环后某个解没有被更新，则放弃当前食物源，此引领蜂转成侦查蜂。

Step5：完成 MCN 次迭代后，输出 fit_{max} 的最优解。

3. 一种基于人工蜂群与 K-Means 的改进混合聚类算法(IGABC-K-Means++)

3.1. K-Means++算法

为了解决因为初始化带来的 K-Means 算法的问题，K-Means++算法主要是让随机选取的中心点不再趋于局部最优解，而是让其尽可能的趋于全局最优解，解决 K-Means 算法的初始化问题。算法中，并不是直接选择距离最远的点作为新的簇中心，只是让这样的点被选做簇中心的概率更大。

具体步骤如下：

Step 1：随机寻找一个点作为中心点；

Step 2：计算其他点到目前的全部簇中心点的距离(最开始只有一个中心点)；

Step 3：利用公式(5)计算出映射到对应点的概率。

$$P_i = \frac{D(k)^2}{\sum_{i=0}^{m} D(i)^2} \tag{5}$$

其中 $D(k)$就是第 k 个点到其他中心点的最短距离。

Step 4：根据 Step 3 中计算出的概率利用轮盘法随机选择出一个中心点，然后重复步骤 2，3，4，直至找到全部中心点。

3.2. 适应度函数设计

适应度函数的选取是影响算法稳定性和收敛性的关键因素[1]，本文结合人工蜂群迭代搜索过程以及K-Means 算法思想提出一种新的适应度函数，如式(6)所示。

$$fitness_i = \sum_{j=1}^{k} \sum_{x_i \in c_j} d\left(x_i, Center_j\right) \Big/ CN_j \tag{6}$$

$\sum_{j=1}^{k} \sum_{x_i \in c_j} d\left(x_i, Center_j\right)$ 表示第 j 类的类内距。CN_j 表示属于第 j 类的样本数。

如果仅以类内距作为适应度函数，就会忽略类内样本数对于类内离散度的影响。当各类样本数目差距较大时，仅仅用类内距作为适应度函数是不合理的，本文将平均类内距离作为适应度函数。Fitness 值越小，表示类内离散度越小，类内点密度越大，说明粒子的位置越好，聚类结果也就越精确。

3.3. GABC 算法

朱国普等学者提出了全局最优解引导的 ABC 算法(GABC 算法) [13] [14]，该算法在位置搜索公式中添加了全局最优项来指导算法的搜索过程，通过添加全局最优项来加强算法在全局最优解附近的搜索能力而且收敛速度也有所增加[15]。使用公式(7)取代原始 ABC 算法中的公式(3)。

$$V_{ij} = x_{ij} + R_{ij}\left(x_{mj} - x_{kj}\right) + \varphi\left(x_{best} - x_{ij}\right) \tag{7}$$

R_{ij} 是 $[-1,1]$ 中的一个随机数，x_{best} 表示全局最优食物源，φ 是 $[0,C]$ (C 是一个正数)中的一个随机数，经多次实验，当 $C=1.5$ 时效果最好。

3.4. 贪婪选择

在原始人工蜂群算法中，跟随蜂按照轮盘赌选择引领蜂时，适应度值越大，被搜索到的概率越大。在本文中，将跟随蜂的选择概率定为公式(8)：

$$P_i = 1 - \frac{fit_i}{\sum_{i=1}^{SN} fit_i},\quad i = 1,2,3,\cdots,\mathrm{SN} \tag{8}$$

适应度值越小表示引领蜂的位置越好，被选择的概率越大。

3.5. IGABC-K-Means++算法的实现

本文将改进的全局人工蜂群算法与 K-Means++算法相结合，提出一种新的混合聚类方法，算法描述如下：

Step 1：模型开始，设置引领蜂、跟随蜂的数量(一般情况下，二者相等)；最大迭代次数 MCN 及控制参数 Limit；聚类簇数 K，Cycle = 1；利用 K-Means++初始化 SN 个蜂群。

Step 2：对初始蜂群进行聚类划分，根据公式(6)计算每只蜜蜂的 fitness，将值较小的 50%为引领蜂，值较大的 50%为跟随蜂。

Step 3：引领蜂利用式(7)进行邻域搜索，得到新位置，对比两个位置的 fitness，按照贪婪选择原则，如果 $\mathrm{fit}_{新} < \mathrm{fit}_{旧}$，则新位置取代原位置；否则，保持原位置。并且将食物源的停止次数加 1；当所有引领蜂完成邻域搜索后，根据式(8)计算概率 Pi。

Step 4：跟随蜂利用 Pi 并基于轮盘赌原则选择引领蜂。当跟随蜂完成选择后，利用式(7)对邻域进行搜索，同样按照贪婪选择原则选择 fit_i 小的位置。

Step 5：跟随蜂搜索结束，用 K-Means 对得到的位置进行聚类划分，更新蜂群。

Step 6：如果某引领蜂在 Limit 次迭代后，结果都没有改变，则变为侦察蜂，并随机产生一个新食物源。

Step 7：如果当前迭代次数达到最大次数 MCN 转向步骤 8，否则转向步骤 3，Cycle = Cycle + 1。

Step 8：输出聚类中心和对应的 fitness，算法结束。算法流程图如图 1 所示：

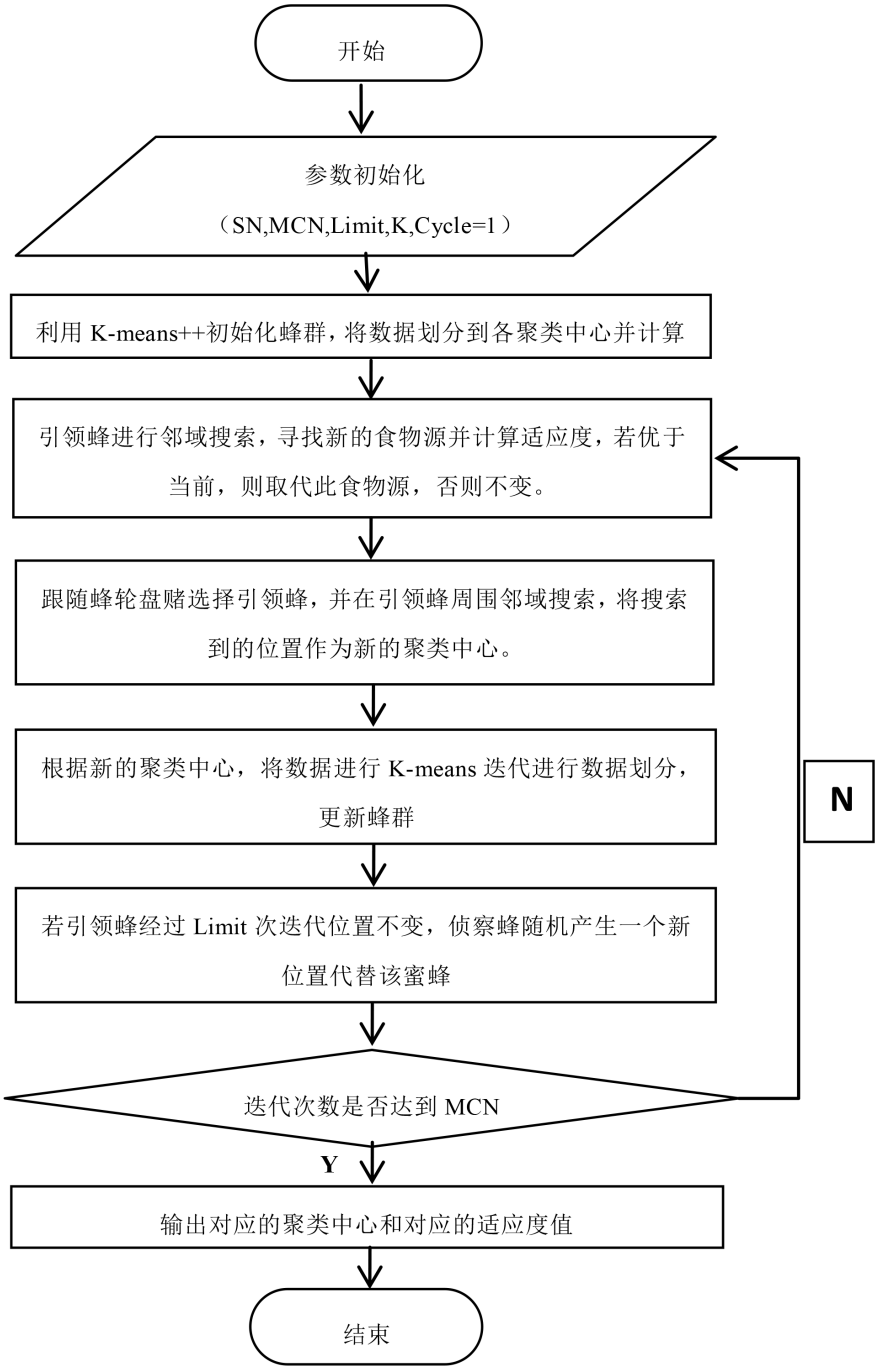

Figure 1. IGABC-K-Means++ algorithm flowchart
图 1. IGABC-K-Means++算法流程图

4. 实验结果及分析

4.1. IGABC 算法性能测试

在函数优化时，设 IGABC 算法和原始 ABC 算法的种群个数 NP = 20，SN = 10，Limit = 100，MCN = 2000。两种算法分别在 Sphere、Griewank 两个标准测试函数上进行性能测试，测试函数属性如下：

Griewank 函数：

$$f(x)=\frac{1}{4000}\sum_{i=1}^{D}X_i^2-\prod_{i=1}^{D}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1 \quad [-600,600]^D$$

Sphere 函数：

$$f(x)=\sum_{i}^{D}x_i^2 \quad [-100,100]^D$$

其中 Sphere 是单峰函数，Griewank 是多峰函数，用适应度来评价改进算法的性能，得出结果如图 2 和图 3 所示。

结果分析：

从图 2 和图 3 可以看出原始 ABC 算法在两种测试函数上表现出了容易陷入局部最优并且有不同程度的收敛速度变慢的问题；IGABC 在适应度函数和全局引导因子的作用下，迭代次数更少，可以找到的位置更优。

4.2. IGABC-K-Means++算法性能测试

为了验证 IGABC-K-Means++聚类算法的有效性，实验选取 UCI 机器学习数据库中著名的 Wine 数据集和 Balance-scale 数据集，数据集属性如表 1 所示。算法参数设置：最大迭代次数 MCN = 100；食物源个数 SN = 10；食物源停滞的最大次数 Limit = 100。

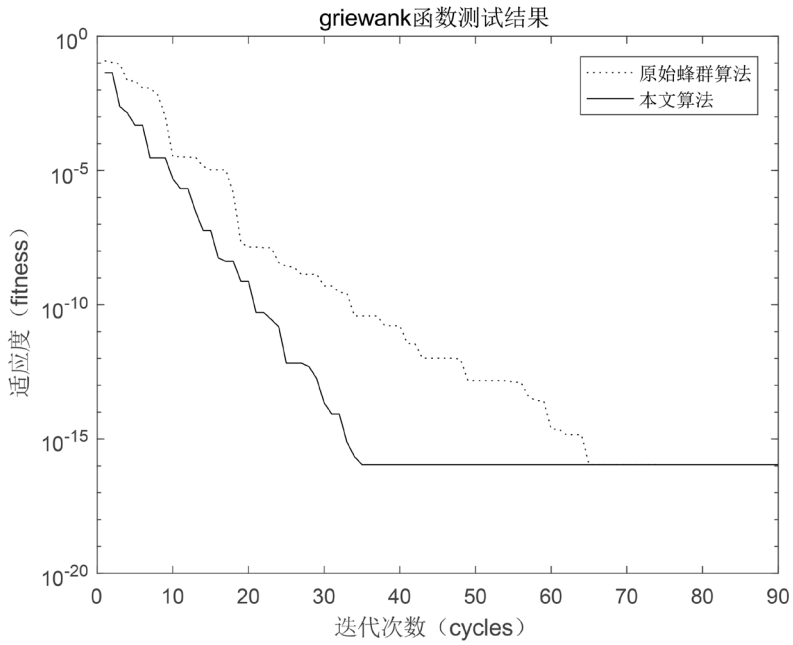

Figure 2. Fitness change of different algorithms in the Griewank function
图 2. 不同算法在 Griewank 函数的适应度变化图

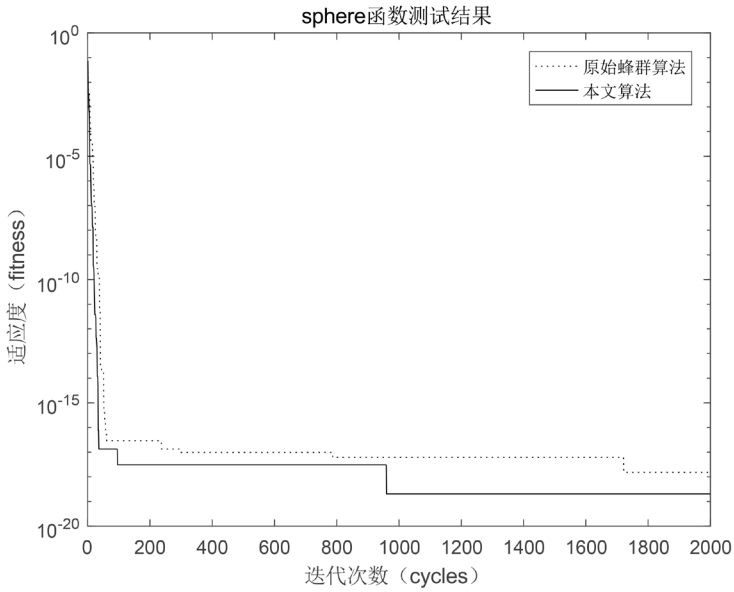

Figure 3. Fitness change of different algorithms in sphere function
图 3. 不同算法在 Sphere 函数的适应度变化图

Table 1. Data sets used in the experiment
表 1. 实验中使用的数据集

数据集名称	数据个数	属性维数	类别个数
Wine data	**178**	**13**	**3**
Balance-scale	**625**	**4**	**3**

实验的软件环境 Matlab2016b，Windows10 操作系统。在相同的数据集下，都进行了 20 次的单独实验，取平均值，在收敛时间、最大值、最小值、平均值、标准差 5 个方面进行比较，结果如表 2、表 3 所示：

Table 2. Clustering comparison results of wine data
表 2. Wine 数据聚类对比结果

算法	收敛时间	最大值	最小值	平均值	标准差
K-Means	1.3622	1.8558	1.6007	1.7275	53.6903
ABC+K-Means	10.9185	1.6714	1.6488	1.6514	0.0924
IABC-K-Means	5.3274	1.6568	1.6555	1.6562	0.0515
IGABC-K-Meanss++	3.1284	1.6485	1.6460	1.6468	0.0176

Table 3. Cluster comparison results of Balance-Scale data
表 3. Balance-Scale 数据聚类对比结果

算法	迭代时间	最大值	最小值	平均值	标准差
K-Means	1.0835	1.1874	0.4262	0.9761	1.7460
ABC+K-Means	11.0268	1.2835	0.9075	1.2442	0.0608
IABC-K-Means	7.68478	1.3337	1.0383	1.3271	0.0287
IGABC-K-Meanss++	5.1476	1.2633	1.0321	1.2376	0.0096

结果分析：

在表 2 和表 3 中可以看出，K-Means 算法聚类耗时短但结果的标准差较大，每次实验结果差距较大，主要是由于初始聚类中心的影响，所以效果并不理想；

ABC+K-Means 算法是把原始的 ABC 算法与传统的算法融合在一起，想比较而言，此算法的寻优能力比传统 K-Means 算法好一些，标准差减小，但 ABC 算法易早熟的问题仍然存在，所以算法后期收敛速度缓慢，需要消耗的时间更长；

IABC-K-Means 算法是文献[12]提出的算法，采用最大最小距离积法产生初始聚类中心，克服了传统 K-Means 鲁棒性较差的缺点，聚类效果得到改善；

IGABC-K-Means++算法在选择的初始点时可以更好的反映数据实际分布，既保证了聚类准确度和算法效率，而且算法表现出很好的稳定性。

5. 结论

本文将改进的人工蜂群算法与 K-Means++算法结合，从优化聚类中心位置入手，从蜂群的初始化、适应度函数、位置更新，贪婪选择四个方面进行改进以较大概率跳出局部极值，寻找更优的聚类中心，解决了 K-Means 算法全局搜索能力差的问题。实验对比结果表明 IGABC-K-Means++算法效率改善的较为明显、性能也得到较大提高，优化了 K-Means 聚类效果和决策分析的准确度。下一步的研究目标是，利用改进的全局人工蜂群算法和 K-Means++算法结合的混合聚类算法优势，将其应用于植物叶片高光谱图像数据中，研究融合算法的实用性。

基金项目

国家自然科学基金(No. 61962048, No. 61562067)。

参考文献

[1] 廖伍代, 朱范炳, 王海泉, 孙雪凯. 基于人工蜂群优化的 K 均值聚类算法[J]. 计算机测量与控制, 2018, 26(4): 136-138.

[2] 杨俊闯, 赵超. K-Means 聚类算法研究综述[EB/OL]. 计算机工程与应用(网络首发论文). http://kns.cnki.net/kcms/detail/11.2127.TP.20191015.1136.006.html, 2019-10-15.

[3] 刘薇, 刘伯嵩, 王洋洋. 基于改进鱼群和 K-Means 的混合聚类算法[J]. 计算机工程与应用, 2013, 49(22): 119-122.

[4] 喻金平, 张勇, 廖列法, 等. 一种改进的混合蛙跳和 k 均值结合的聚类算法[J]. 计算机工程与科学, 2016, 38(2): 356-362.

[5] 熊众望, 罗可. 基于改进的简化粒子群聚类算法[J]. 计算机应用与研究, 2014, 31(12): 3550-3552.

[6] 邓海, 覃华, 孙欣. 一种优化初始中心的 K-Means 聚类算法[J]. 计算机技术与发展, 2013, 23(11): 42-45.

[7] Lu, B. and Ju, F. (2012) An Optimize Genetic K-Means Clustering Algorithm. *CSIP* 2012: *Proceedings of the* 2012 *International Conference on Computer Science and Information Processing*, 2012, 1296-1299.

[8] 刘三阳, 张平, 朱明敏. 基于局部搜索的人工蜂群算法[J]. 控制与决策, 2014, 29(1): 123-128.

[9] 葛宇, 梁静, 王学平. 基于极值优化策略的改进的人工蜂群算法[J]. 计算机科学, 2013, 40(6): 247-251.

[10] 周长喜, 毛力, 吴滨, 等. 基于当前最优解的人工蜂群算法[J]. 计算机工程, 2015, 41(6): 147-151.

[11] 毕晓君, 宫汝江. 一种结合人工蜂群和 k-均值的混合聚类算法[J]. 计算机应用研究, 2012, 29(6): 2040-2042.

[12] 喻金平, 郑杰, 梅宏标. 基于改进人工蜂群算法的 k 均值聚类算法[J]. 计算机应用, 2014, 34(4): 1065-1069.

[13] Zhu, G.P. and Kwong, S. (2010) Gbest-Guided Artificial Bee Colony Algorithm for Numerical Function Optimization. *Applied Mathematics and Computation*, **27**, 1341-1348.

[14] 罗可, 易斌. 一种基于改进蜂群的 K-Means 聚类算法[J]. 长沙理大学学报(自然科学版), 2016, 13(4): 85-89.

[15] 常扣扣. 基于改进人工蜂群算法的 K-Means 聚类算法[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2017.