Journal of Computer Applications ISSN 1001-9081 2014-04-10
计算机应用,2014,34(4):1065-1069,1088 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2014)04-1065-05 doi:10.11772/j.issn.1001-9081.2014.04.1065

# 基于改进人工蜂群算法的 K 均值聚类算法

喻金平,郑 杰*,梅宏标

(江西理工大学 信息工程学院,江西 赣州 341000)

(* 通信作者电子邮箱 zjhxp_1990@163.com)

**摘 要:** 针对 K 均值聚类(KMC)算法全局搜索能力差、初始聚类中心选择敏感,以及原始人工蜂群(ABC)算法的初始化随机性、易早熟、后期收敛速度慢等问题,提出了一种改进人工蜂群算法(IABC)。该算法利用最大最小距离积方法初始化蜂群,构造出适应 KMC 算法的适应度函数以及一种基于全局引导的位置更新公式以提高迭代寻优过程的效率。将改进的人工蜂群算法与 KMC 算法结合提出 IABC-Kmeans 算法以改善聚类性能。通过 Sphere、Rastrigin、Rosenbrock 和 Griewank 四个标准测试函数和 UCI 标准数据集上进行测试的仿真实验表明,IABC 算法收敛速度快,克服了原始算法易陷入局部最优解的缺点;IABC-Kmeans 算法则具有更好的聚类质量和综合性能。

**关键词:** 人工蜂群算法;K 均值聚类算法;适应度函数;位置更新公式;聚类

**中图分类号:** TP18;TP301.6 **文献标志码:** A

## K-means clustering algorithm based on improved artificial bee colony algorithm

YU Jinping, ZHENG Jie*, MEI Hongbiao

(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi* 341000, *China*)

**Abstract:** In order to overcome the disadvantages of the K-Means Clustering (KMC) algorithm, such as the poor global search ability, being sensitive to initial cluster centroid, as well as the initial random, being vulnerable to trap in local optima and the slow convergence velocity in later period of the original Artificial Bee Colony (ABC) algorithm, an Improved ABC (IABC) algorithm was proposed. IABC algorithm adopted the max-min distance product algorithm for initial bee colony to form a fitness function, which is adapted to the KMC algorithm, and a position updating method based on the global leading to enhance the efficiency of the iterative optimization process. The combination of the IABC and KMC (IABC-Kmeans) would improve the efficiency of clustering. The simulation experiments were conducted on the four standard test functions including Sphere, Rastrigin, Rosenbrock and Griewank and the UCI standard data sets. The experimental results indicate that IABC algorithm has a fast convergence speed, and overcomes the defect of the original algorithm being easily falling into local optimal solution; IABC-Kmeans has better clustering quality and general performance.

**Key words:** Artificial Bee Colony (ABC) algorithm; K-Means Clustering (KMC) algorithm; fitness function; position update rule; clustering

## 0 引言

聚类作为一种无监督学习,是数据挖掘领域的一个重要研究方向。聚类就是将数据对象分组成多个簇(类),同一簇内的对象相似度尽可能大,不同簇间的对象相似度尽可能小。K 均值聚类(K-Means Clustering,KMC)算法是一种基于划分思想的聚类算法,它具有思路简单、聚类快速、局部搜索能力强的优点;但也存在对初始聚类中心选择敏感、全局搜索能力较差、聚类效率和精度低的局限性问题。对于 KMC 算法对初始点敏感和全局搜索能力较差问题,吸引了很多学者对该问题的研究与改进。文献[1]基于谱图理论思想,采用密度敏感的相似性度量来计算对象的密度,启发式地生成样本初始中心,可以得到较高质量的初始中心;文献[2]在每个类内都有一个数据稠密区的假设基础上,提出了一种基于最小支撑树的聚类中心初始化方法,该方法提高了 KMC 算法的模式识别率,但增加了时间复杂度;文献[3]提出一种改进的粒子群优化(Particle Swarm Optimization, PSO)和 KMC 混合聚类算法,算法在运行过程中通过引入小概率随机变异操作增强种群的多样性,提高了 KMC 算法的全局搜索能力;文献[4]通过将遗传算法(Genetic Algorithm, GA)的编码、交叉和变异思想融入 KMC,充分结合 KMC 的局部寻优能力和遗传算法的全局寻优能力,提出了一种基于遗传算法的优化 KMC 算法,有效地解决了 KMC 易陷入局部收敛的问题。

群体智能与仿生算法以其进化过程与初始值无关、搜索速度快、对函数要求低的优点,成为进化算法的一个重要分支,并吸引了各个领域学者对其研究。目前,比较常见的群体智能与仿生算法有粒子群算法(Particle Swarm Optimization, PSO)、细菌觅食算法(Bacterial Foraging Algorithm, BFA)、人工鱼群算法(Artificial Fish Swarm Algorithm, AFSA)、遗传算法(GA)和蚁群算法(Ant Colony Optimization, ACO)等[5]。近年来,在优化领域中出现了一种新的随机型搜索方法——蜂

收稿日期:2013-10-24;修回日期:2013-12-17。

基金项目:江西省教育厅自然科学基金资助项目(DJJ12346);江西省研究生创新专项基金资助项目(YC2013-S198)。

作者简介:喻金平(1964-),男,江西南昌人,教授,主要研究方向:数据挖掘;郑杰(1990-),男,安徽六安人,硕士,主要研究方向:数据挖掘、群体智能;梅宏标(1976-),男,江西南昌人,副教授,博士,主要研究方向:大规模仿真系统工程。

群算法。Seeley 于 1995 年最先提出了蜂群的自组织模拟模型，在该模型中，虽然各社会阶层的蜜蜂只完成了一种任务，但是蜜蜂以“摆尾舞”、气味等多种方式在群中进行信息的交流，使得整个群体可以完成诸如喂养、采蜜、筑巢等多种工作。Karaboga 于 2005 年将蜂群算法成功应用于函数的极值优化问题，系统地提出了人工蜂群算法(Arificial Bee Colony，ABC)，该算法简单，全局搜索能力好，鲁棒性强；但是，人工蜂群算法也存在着后期收敛速度较慢、容易陷入局部最优的问题。文献[6]通过引入反学习的初始化方法，有效提高了求解效率和解的质量；文献[7]通过引入人工蜂群的粒子群算法，利用粒子群的局部搜索能力和人工蜂群的全局搜索能力，使得算法具有较快的收敛速度和很强的跳出局部最优的能力。

鉴于 KMC 和 ABC 算法各自的特性，本文首先提出了一种改进的 ABC(Improved ABC，IABC)算法，利用提出的最大最小距离积法初始化蜂群，保证初始点的选择能够尽可能代表数据集的分布特征，并在迭代过程中使用新的适应度函数和位置更新公式完成寻优进化；然后将 IABC 算法应用到 KMC 中提出了 IABC-Kmeans 算法，以改善聚类性能。

## 1 相关算法简介

### 1.1 原始的 KMC 算法

设聚类样本 $x=\{\boldsymbol{x}_i\in\mathbf{R}^d, i=1,2,\cdots,n\}$，划分为多个不同类别 $C=\{C_1,C_2,\cdots,C_k\}$，其中 $\boldsymbol{x}_i(i=1,2,\cdots,n)$ 为 $d$ 维向量，$k$ 为聚类个数。

聚类中心计算公式如下：

$$C_j=\frac{1}{C_j}\sum_{i=1}^{n}\boldsymbol{x}_i;\ j=1,2,\cdots,k \tag{1}$$

KMC 的准则函数如下：

$$J=\sum_{j=1}^{k}\sum_{\boldsymbol{x}_i\in C_j}d(\boldsymbol{x}_i,C_j) \tag{2}$$

其中：$d(\boldsymbol{x}_i,C_j)$ 表示数据 $\boldsymbol{x}_i$ 与所属类中心 $C_j$ 的距离；$J$ 表示类内距之和。

### 1.2 人工蜂群算法

人工蜂群(ABC)算法是模拟蜜蜂的觅食(采蜜)过程，其组成要素有食物源、引领蜂、跟随蜂和侦察蜂。其中，引领蜂和跟随蜂各占一半，并且每一个蜜源仅有一只引领蜂开采。

算法的思想：设有 $N$ 个食物源 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，每个食物源 $\boldsymbol{x}_i(i=1,2,\cdots,N)$ 都是一个 $d$ 维向量，设定蜜蜂总的循环搜索次数为 $MCN$，每个蜜源的可重复开采次数为 $Limit$。

蜂群的初始化：随机产生 $N$ 个解并计算其适应度，将适应度由大到小排列，前一半作为引领蜂，后一半作为跟随蜂和侦察蜂。

随后，引领蜂在食物源邻域进行搜索，并根据式(3)产生一个新的食物源。比较两个食物源的优劣，保留质量较好的食物源。

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

其中：$v_{ij}$ 表示在 $x_{ij}$ 附近产生的一个新的位置；$k\in\{1,2,\cdots,N\}$，$k$ 和 $j$ 都是通过随机公式产生的随机数，并且 $k\neq i$；$r_{ij}\in[-1,1]$，它使新产生的位置约束在 $x_{ij}$ 附近。

在跟随阶段，跟随蜂根据引领蜂传达的食物源丰富度信息，基于轮盘赌原则，依据式(4)中概率选择引领蜂。选中引领蜂后，跟随蜂也在食物源邻域依据式(3)生成一个新的食物源，比较两食物源的优劣，保留质量较好的食物源。

$$P_i=\frac{fitness_i}{\sum_{i=1}^{N}fitness_i};\ i=1,2,\cdots,N \tag{4}$$

其中：$fitness_i$ 是第 $i$ 个解的适应度值，$P_i$ 是跟随蜂选择引领蜂的概率。

当引领蜂连续经过 $Limit$ 次循环后食物源没有更新时，则放弃该食物源成为侦察蜂。

## 2 改进的人工蜂群算法

针对原始人工蜂群的初始化、适应度函数和位置更新公式，本文提出的 IABC 算法使用最大最小距离积法初始化蜂群，克服原始人工蜂群算法初始化的随机性；并利用新的适应度函数及引入全局引导因子的位置更新公式进行迭代寻优。

### 2.1 使用最大最小距离积法初始化

种群初始化在进化算法中显得尤为重要，因为它影响算法的全局收敛速度和解的质量。所以本文在文献[8-9]的基础上提出最大最小距离积法并用其初始化蜂群，这里的初始化处理不仅克服了蜂群初始化的随机性，也为后面的 K 均值聚类降低了对初始点的敏感性。文献[8]采用了最大最小距离法搜索出最佳初始聚类中心，降低了对初始聚类中心的敏感性，在收敛速度和准确率上都有较大提高；但是由于其遵从最小距离的思想，可能使得初始聚类中心的选取过于稠密而出现聚类冲突现象。文献[9]针对最大最小距离法的缺陷提出了最大距离积法搜索初始聚类中心，使得初始点的选择更加符合数据分布特征，并有效减少了迭代次数；但是最大距离积法也存在缺陷，比如会出现两个距离积相等而它们所在区域的点密度又相差很大的情况，一些参数需要用户自己输入，选择的初始点有偏向点集外围的倾向而无法准确反映数据实际分布。

针对最大最小距离法和最大距离积法的不足，提出一种最大最小距离积法，其中：$D$ 是包含所有数据的集合；$N$ 是初始蜂群的个数；$k$ 是要选取的初始点个数；$Z$ 是存储待加入的 $k$ 个初始点的集合，算法开始前为空集；$Temp$ 是存储 $Z$ 中各个元素到 $D$ 中各个元素乘积结果的数组。算法流程如图 1 所示。

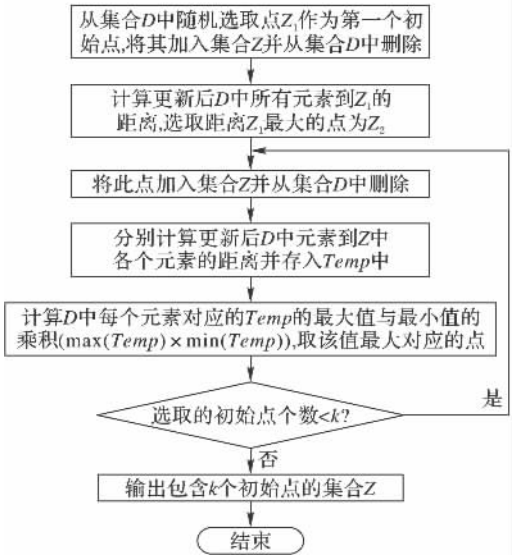



图 1 最大最小距离积算法流程

通过该方法的思想和步骤可以看出,该算法所需参数少,使用 ( max( *Temp*) × min( *Temp*) ) 的乘积能选取到点密度较大的点,稀疏了初始点的分布; 而且通过这样的处理,不仅可以大概率地避免文献 [9] 中出现两个距离积相等而它们所在区域的点密度又相差很大的情况,同时也能用乘积放大点与点之间的差异,使得选取过程更具区分度。

### 2.2　适应度函数

适应度函数将引导群体进化的方向,直接决定了群体的进化行为、迭代的次数和解的质量,不同的适应度函数会得出不同优劣程度的解。所以,结合人工蜂群迭代搜索过程以及 KMC 算法思想提出一种新的适应度函数,如式( 5) 所示。

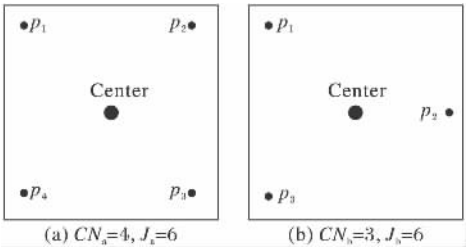
$$fitness_i = CN_i / J_i;\ i = 1,2,\cdots,N \tag{5}$$

其中: $CN_i$ 表示属于第 $i$ 类的点的个数; 由 $J = \sum_{j=1}^{k}\sum_{x_i \in C_j} d(x_i, C_j)$ 得出 $J_i = \sum_{x_j \in C_i} d(x_j, C_i)$ 表示第 $i$ 类的类内对象到中心点 $C_i$ 的距离之和。

如果仅仅以点数或是类内距作为适应度函数都会有其不足之处,举例如下:

1) $J_i$ 相同, $CN_i$ 不同的情况。

图 2( a) 和图 2( b) 中虽然 $J$ 相等,但是 $CN_a = 4 > CN_b = 3$, 如果仅以类内距为适应度函数,则在选优的过程中会失去精确度; 由式( 5) ,显然有 $fitness_a > fitness_b$, 所以迭代会向图 2( a) 所示的趋势搜索进化,减少了迭代次数并提高精确度。

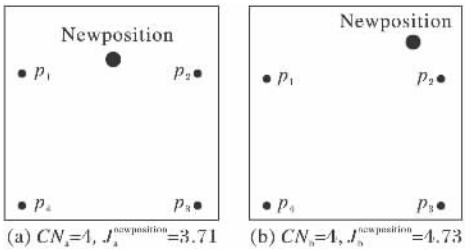



(a) $CN_a$=4, $J_a$=6　　(b) $CN_b$=3, $J_b$=6

图 2　*CN* 不等,*J* 相等

2) $CN_i$ 相同, $J_i$ 不同的情况

图 3( a) 和图 3( b) 中虽然 $CN$ 都相等,但是 $J_a^{\text{newposition}} = 3.7 < J_b^{\text{newposition}} = 4.73$, 如果仅以点个数作为适应度函数, 则函数在辨识( a) 、( b) 情况时适应力会下降; 同理,由式( 5) ,有 $fitness_a > fitness_b$, 使迭代过程更精确。




(a) $CN_a$=4, $J_a^{\text{newposition}}$=3.71　　(b) $CN_b$=4, $J_b^{\text{newposition}}$=4.73

图 3　*CN* 相等,*J* 不等

### 2.3　位置更新公式

位置更新公式决定着蜜蜂能否快速准确地找到新的蜜源。式( 3) 位置更新公式具有很强的搜索能力,但是探索能力欠缺,在搜索邻域时具有迭代随机性、易陷入局部最优解、更新速度缓慢的缺点 [10-11] 。所以,针对这一问题,提出一种引入全局因子的位置更新式( 6) :

$$V_{ij} = x_{ij} + r_{ij}(x_{mj} - x_{kj}) + \varphi(x_{\text{best},j} - x_{ij}) \tag{6}$$

其中: $v_{ij}$ 表示在 $x_{ij}$ 附近产生的一个新的位置; $k, m \in \{1,2,\cdots,N\}$, $k$、$m$ 和 $j$ 都是通过随机公式产生的随机数, $k$、$m$ 互斥且都不等于 $i$; $r_{ij} \in [-1,1]$; $\varphi \in [0,1]$ 是一个随机数; $x_{\text{best},j}$ 代表食物丰富度最高的食物源。

式( 3) 在邻域搜索时仅仅向着 $r_{ij}(x_{mj} - x_{kj})$ 的矢量方向迭代,没有考虑迭代前后位置的优劣比较,在整个搜索过程中,每只引领蜂只能获得自己的历史最优位置和当前的位置信息,缺乏对于整个蜂群全局最优的考虑。从群体智能的进化角度来看,群体中的每个个体都可以从群体中所有其他个体经验中收益。所以,式( 6) 在( 3) 的基础上加上了全局引导因子 $(x_{\text{best},j} - x_{ij})$, 使蜜蜂的搜索具有很强的方向性和目的性,在全局因子前面加入了影响因子 $\varphi$, 用于约束寻优的幅度。从因子组成可以看出,如果当前位置与最优位置差距大,则更新的步长会动态增加; 反之,则缓慢逼近。

## 3　基于改进人工蜂群算法的 KMC 算法

基于以上三点改进提出了 IABC-Kmeans 算法。该算法的基本思想是: 通过 IABC 算法进行一次迭代,将迭代得到的新位置作为 KMC 的初始点并进行一次 K 均值聚类,再用聚类获得的新的中心点更新蜂群; 如此交替执行 IABC 算法和 K 均值聚类,直到算法结束。

算法基本步骤描述如下:

步骤 1　设置引领蜂、跟随蜂和侦察蜂的数量( 一般有引领蜂数量 = 跟随蜂数量) ; 最大迭代次数 *MCN* 以及控制参数 *Limit*; 当前迭代次数 *Cycle*,初始值为 1; 聚类类别数 $k$; 利用最大最小距离积法初始化蜂群,产生 $\{Z_1, Z_2, \cdots, Z_N\}$ 个初始蜂群。

步骤 2　对初始蜂群进行一次聚类划分,根据式( 5) 计算每只蜜蜂的适应度,按照适应度大小排序,将前一半作为引领蜂,后一半作为跟随蜂。

步骤 3　引领蜂利用式( 6) 对其邻域进行搜索,得到新的位置,按照贪婪选择原则,如果新的位置的适应度大于原先位置的适应度,则用新的位置更新原位置; 否则,保持原位置不变。当所有引领蜂完成邻域搜索后,根据式( 4) 计算概率 $P_i$。

步骤 4　跟随蜂利用算得的概率 $P_i$ 并基于轮盘赌原则选择引领蜂,原则上, $P_i$ 越大,表明引领蜂 $i$ 的适应度值越大,被跟随蜂选中的概率也越大。当跟随蜂完成引领蜂选择后,利用式( 7) 对邻域搜索,同样按照贪婪选择原则选择适应度高的位置。

步骤 5　在所有跟随蜂完成搜索后,将得到的位置作为聚类中心,对数据集进行一次 K 均值迭代聚类,根据聚类划分,用每一类的新的聚类中心更新蜂群。

步骤 6　如果某引领蜂在 *Limit* 次迭代后,结果都没有改变,则由引领蜂变为侦察蜂,并随机产生一个新的位置取代原位置。

步骤 7　如果当前迭代次数大于最大次数 *MCN*, 则迭代结束,算法结束; 否则转向步骤 2, $Cycle = Cycle + 1$。

## 4　实验仿真与分析

实验采用 CPU 为 Intel Core 2 Duo 2.00 GHz、内存为 2 GB 的

计算机，操作系统为 Windows XP，编译软件为 Matlab 7.10.0。

### 4.1 IABC 算法性能测试

在函数优化时，设改 IABC 算法和原始 ABC 算法的蜂群群体规模为 20，即引领蜂和跟随蜂的数量均为 10；*Limit* 值为 100，即在同一食物源搜索迭代超过 100 次则个体由引领蜂变为侦察蜂；迭代次数为 2000。对本文的 IABC 算法、文献[12]算法和原始 ABC 算法分别在 Sphere、Rastrigin、Rosenbrock、Griewank 四个标准测试函数上进行测试，对比效果并作分析，各测试函数特性如表 1 所示。其中 Sphere 和 Rosenbrock 是单峰函数，Rastrigin 和 Griewank 是多峰函数，采用适应度评价改进算法的性能，得出四种标准测试函数的适应度变化趋势如图 4～7 所示。

表 1 四个测试函数的表达式、搜索范围、最小值

| 函数名称 | 函数表达式 | 搜索空间 | 最小值 |
|---|---|---|---|
| Sphere | $f_1(x)=\sum_{i=1}^{n}x_i^2$ | [-100,100] | 0 |
| Rastrigin | $f_2(x)=\sum_{i=1}^{n}(x_i^2-10(\cos(2\pi x_i))+10)$ | [-5.12,5.12] | 0 |
| Rosenbrock | $f_3(x)=\sum_{i=1}^{n}100(x_{i+1}-x_i^2)^2+(1-x_i)^2$ | [-100,100] | 0 |
| Griewank | $f_4(x)=\frac{1}{4000}\sum_{i=0}^{n}x_i^2-\prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1$ | [-600,600] | 0 |

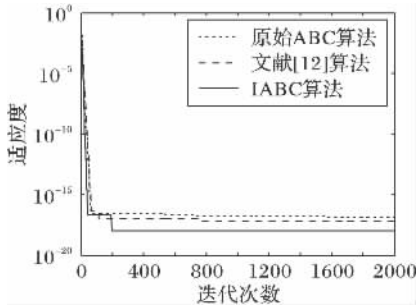



图 4 不同算法在 Sphere 函数的适应度变化图

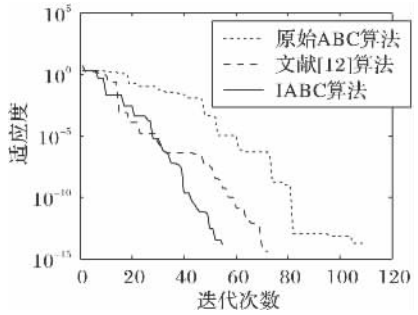



图 5 不同算法在 Rastrigin 函数的适应度变化

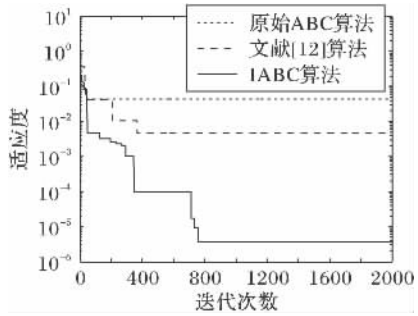



图 6 不同算法在 Rosenbrock 函数的适应度变化

从图 4～7 可以看出原始 ABC 算法在单峰函数以及多峰函数上都会出现不同程度的收敛速度较为缓慢、易陷入局部最优解的情况；文献[12]算法虽然相对原始算法在收敛速度有所提高，迭代次数更少，但是在全局寻优能力上稍显薄弱；本文的 IABC 算法采用新的适应度函数和位置更新公式，避免了食物源位置更新邻域的随机性，通过全局引导因子可以使蜜蜂能够快速向最优食物源所在区域移动。所以从图 4、图 6 可以看出利用 IABC 算法相对原始 ABC 算法和文献[12]算法能够找到更优的位置，图 5、图 7 中 IABC 算法在寻得最优位置的过程中迭代次数更少，节省了时间开销。

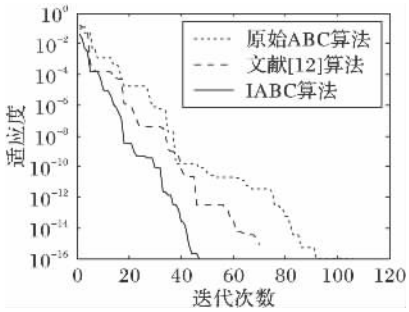



图 7 不同算法在 Griewank 函数的适应度变化

### 4.2 IABC-Kmeans 算法性能测试

IABC-Kmeans 算法参数设置如下：最大迭代次数 $MCN=100$，蜂群规模 $N=20$；Iris 中聚类数目 $k=3$，Balance-scale 中聚类数目 $k=3$，Glass 中聚类数目 $k=6$；控制参数 *Limit* = 10。

UCI 标准数据集[13]中的 Iris、Balance-scale 和 Glass 的特征如表 2 所示，本文的 IABC-Kmeans 算法在 Iris、Balance-scale、Glass 三个标准数据集上迭代 100 次的适应度收敛趋势如图 8 所示。

表 2 实验中涉及的数据集

| 数据集名称 | 样本数目 | 属性维数 | 类别个数 |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Balance-scale | 625 | 4 | 3 |
| Glass | 214 | 10 | 6 |

从图 8 可以看出，无论是具有 625 个样本数的 Balance-scale 数据集，还是有 10 个属性数的 Glass 数据集，本文的 IABC-Kmeans 算法在各个数据集上的适应度值变化幅度都很小，采用新的位置更新公式使得算法能够以全局最优解为目标快速调整搜索步长并迭代到最优解的位置；其中也可以看出在 Balance-scale 数据集寻优的过程中也有跳出局部最优

解,由引领蜂变为侦察蜂搜索到全局最优解。由此可见,本文的 IABC-Kmeans 算法能够快速准确地向全局最优解位置进化,具有跳出局部最优解的能力,迭代进化次数少,算法的稳定性强。

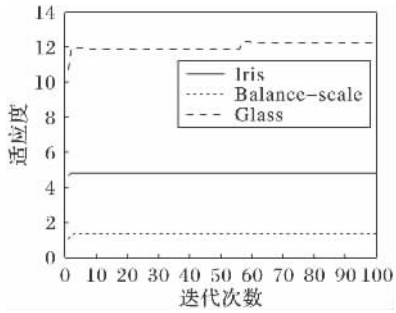



图 8　Iris、Balance-scale、Glass 的适应度收敛趋势

对 KMC 算法、ABC + KMC 算法、文献 [12] 算法和本文的 IABC-Kmeans 算法在经典数据集 Iris、Balance-scale、Glass 上进行测试并作对比分析。各种算法在上述三种标准数据集上的实验相关指标如表 3 ~ 5 所示。

表 3　Iris 数据聚类对比结果

| 算法 | 最小值 | 最大值 | 平均值 | 标准差 |
|---|---|---|---|---|
| KMC | 2.954 5 | 4.434 7 | 4.309 6 | 1.441 0 |
| ABC + KMC | 3.951 7 | 4.556 3 | 4.455 4 | 0.097 3 |
| 文献 [12] 算法 | 4.069 4 | 4.692 5 | 4.643 2 | 0.010 5 |
| IABC-Kmeans | 4.735 5 | 4.809 5 | 4.805 8 | 0.001 1 |

表 4　Balance-scale 数据聚类对比结果

| 算法 | 最小值 | 最大值 | 平均值 | 标准差 |
|---|---|---|---|---|
| KMC | 0.426 2 | 1.187 4 | 0.976 1 | 1.746 0 |
| ABC + KMC | 0.907 5 | 1.283 5 | 1.244 2 | 0.060 8 |
| 文献 [12] 算法 | 0.948 8 | 1.325 4 | 1.305 9 | 0.018 3 |
| IABC-Kmeans | 1.120 3 | 1.333 7 | 1.327 1 | 0.003 4 |

表 5　Glass 数据聚类对比结果

| 算法 | 最小值 | 最大值 | 平均值 | 标准差 |
|---|---|---|---|---|
| KMC | 5.638 1 | 10.154 3 | 8.648 7 | 2.029 3 |
| ABC + KMC | 7.842 9 | 10.654 4 | 9.950 1 | 0.374 1 |
| 文献 [12] 算法 | 7.762 4 | 10.721 5 | 10.685 5 | 0.162 6 |
| IABC-Kmeans | 10.852 6 | 11.891 9 | 11.889 7 | 0.058 2 |

由表 3 ~ 5 可以看出: KMC 算法的聚类效果依赖于初始点的选择,所以聚类的标准差相对较大,全局寻优能力较差,而且需要达到稳定值的迭代次数多、耗时长; ABC + KMC 算法与 KMC 算法相比,算法的寻优能力增强,标准误差减小,但是由于 ABC 易早熟,所以算法后期收敛速度缓慢,耗费时间较长,很难达到全局最优解; 文献 [12] 算法引入了线性调整策略可以较快地定位到最优解,但是全局搜索能力不明显,仍然存在易早熟的问题; 本文的 IABC-Kmeans 算法由于加入初始化过程、适应度公式和全局引导因子,使得全局搜索能力增强,能够跳出局部最优解,迭代次数少,收敛速度和聚类精准度都有提高,标准差最小。

综上所述可以看出,改进的人工蜂群算法( IABC) 在各个测试函数上相比原始人工蜂群算法( ABC) 和文献 [12] 算法在效率和精确度上都有提高,克服了算法易陷入局部最优解和后期收敛速度缓慢的缺陷,提高了算法的鲁棒性和综合性能; 而本文的 IABC-Kmeans 算法不仅具有改进后的人工蜂群的全局搜索能力,而且也使 K 均值聚类的迭代次数减少,节省了时间开销。

## 5　结语

本文提出了一种改进的人工蜂群算法,分别从蜂群的初始化、适应度函数、位置更新公式三个方面对其改进,从而克服了原始算法初始化随机性和易陷入局部最优解等问题。将改进的人工蜂群算法与 KMC 算法结合解决了 KMC 算法全局搜索能力差的问题。实验对比结果表明了本文算法的有效性,而且在优化效率、优化性能上都有较大改善。但是,该算法也有其自身的局限性,由于引入基于全局引导的位置更新公式,使得算法更侧重探索能力而忽视了其开发能力; 同时将改进人工蜂群算法与 KMC 算法结合来完成聚类过程的耗时相对较长。如何在保证利用人工蜂群算法和 KMC 算法各自优势的同时降低时间复杂度,将是下一步的研究内容。

参考文献:

[1] WANG Z, LIU G, CHEN E. A K-means algorithm based on optimized initial center points [J]. Pattern Recognition and Artificial Intelligence, 2009, 22(2): 299 – 304. (汪中,刘贵全,陈恩红. 一种优化初始中心点的 K-means 算法[J]. 模式识别与人工智能, 2009,22(2): 299 – 304.)

[2] LI C, WANG Y. New initialization method for cluster center [J]. Control Theory & Applications, 2010, 27(10): 1435 – 1440. (李春生,王耀南. 聚类中心初始化的新方法[J]. 控制理论与应用, 2010,27(10): 1435 – 1440.)

[3] TAO X, XU J, YANG L, *et al.* Improved cluster algorithm based on K-means and particle swarm optimization [J]. Journal of Electronics & Information Technology, 2010, 32(1): 92 – 97. (陶新民,徐晶,杨立标,等. 一种改进的粒子群和 K 均值混合聚类算法[J]. 电子与信息学报,2010,32(1): 92 – 97.)

[4] LU B, JU F. An optimized genetic K-means clustering algorithm [C]// CSIP 2012: Proceedings of the 2012 International Conference on Computer Science and Information Processing. Piscataway: IEEE, 2012: 1296 – 1299.

[5] YANG S, ZHANG H. Bionic swarm intelligence and computing — Matlab technology [M]. Beijing: Publishing House of Electronic Industry, 2012: 236 – 243. (杨淑莹,张桦. 群体智能与仿生计算——Matlab 技术实现[M]. 北京: 电子工业出版社,2012: 236 – 243)

[6] HUANG L, LIU S, GAO W. Differential evolution with the search strategy of artificial bee colony algorithm [J]. Control and Decision, 2012, 27(11): 1644 – 1648. (黄玲玲,刘三阳,高卫峰. 具有人工蜂群搜索策略的差分进化算法[J]. 控制与决策,2012,27(11): 1644 – 1648.)

[7] GAO W, LIU S, JIAO H, *et al.* Particle swarm optimization with search operator of artificial bee colony algorithm [J]. Control and Decision, 2012, 27(6): 833 – 838. (高卫峰,刘三阳,焦合华,等. 引入人工蜂群搜索算子的粒子群算法[J]. 控制与决策, 2012,27(6): 833 – 838.)

(下转第 1088 页)

间效率的进一步提升,以及可并行执行的加速版本。

**参考文献:**


[1] CHICKERING D M, GEIGER D, HECKERMAN D. Learning Bayesian network is NP-hard, MSR-TR-94-17 [R]. [S.l.]: Microsoft Research, 1994.

[2] CHOW C K, LIU C N. Approximating discrete probability distributions with dependence trees [J]. IEEE Transactions on Information Theory, 1968, 14(3): 462 - 467.

[3] de WAAL P R, van der GAAG L C. Inference and learning in multi-dimensional Bayesian network classifiers [C]// Proceedings of the 9th European Conferences on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, LNCS 4724. Berlin: Springer-Verlag, 2007: 501 - 511.

[4] REBANE G, PEARL J. The recovery of causal poly-trees from statistical data [C]// UAI '87: Proceedings of the 3rd Conference on Uncertainty in Artificial Intelligence. New York: Elsevier Science Inc., 1987: 175 - 182.

[5] ZARAGOZA J C, SUCAR E, MORALES E. A two-step method to learn multidimensional Bayesian network classifiers based on mutual information measures [C]// Proceedings of the 24th International FLAIRS Conference. Menlo Park, California: AAAI Press, 2011: 644 - 649.

[6] RODRIGUEZ J D, LOZANO J A. Multi-objective learning of multi-dimensional Bayesian classifiers [C]// HIS '08: Proceedings of the 8th International Conference on Hybrid Intelligent Systems. Washington, DC: IEEE Computer Society, 2008: 501 - 506.

[7] BIELZ C, LI G, LARRANGA P. Multi-dimensional classification with Bayesian networks [J]. International Journal of Approximate Reasoning, 2011, 52(6): 705 - 727.

[8] BORCHANI H, BIELZA C, LARRANGA P. Learning CB-decomposable multi-dimensional Bayesian network classifiers [C]// PGM'10: Proceedings of the 5th European Workshop on Probabilistic Graphical Models. Helsinki: HIIT Publications, 2010: 23 - 32.

[9] ZARAGOZA J H, SUCAR L E, MORALES E F, *et al.* Bayesian chain classifiers for multidimensional classification [C]// IJCAI '11: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Menlo Park, California: AAAI Press, 2011, 3: 2192 - 2197.

[10] BORCHANI H, BIELZA C, MARTÍNEZ-ARTÍN P, *et al.* Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: an application to predict the European Quality of life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39) [J]. Journal of Biomedical Informatics, 2012, 45(6): 1175 - 1184.

[11] BORCHANI H, BIELZA C, TORO C, *et al.* Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers [J]. Artificial Intelligence in Medicine, 2013, 57(3): 219 - 229.

[12] ALIFERIS C F, TSAMARDINOS I, STATNIKOV A. HITON: a novel Markov blanket algorithm for optimal variable selection [C]// Proceedings of the 2003 Annual Symposium on American Medical Informatics Association (AMIA). Washington, DC: AMIA Publications, 2003: 21 - 25.

[13] PENA J M, NILSSON R, BJORKEGREN J, *et al.* Towards scalable and data efficient learning of Markov boundaries [J]. International Journal of Approximate Reasoning, 2007, 45(2): 211 - 232.

[14] VERMA T, PEAL J. Equivalence and synthesis of causal models [C]// UAI '90: Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence. New York: Elsevier Science Inc., 1990: 255 - 268.

[15] van der GAAG L C, de WAAL P R. Multi-dimensional Bayesian network classifiers, UU-CS-2006-056 [R]. Utrecht: Utrecht University, Department of Information and Computing Sciences, 2006.

[16] FU S, DESMARAIS M C. Tradeoff analysis of different Markov blanket local learning approaches [C]// PAKDD '08: Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer-Verlag, 2008: 562 - 571.

[17] FU S, DESMARAIS M C. Fast Markov blanket discovery algorithm via local learning within single pass [C]// Canadian AI '08: Proceedings of the 21st Conference of the Canadian Society for Computational Studies of Intelligence. Berlin: Springer-Verlag, 2008: 96 - 107.

[18] FU S, DESMARAIS M C. Markov blanket based feature selection: a review of past decade [C]// Proceedings of the 2010 World Congress on Engineering, LNCS 2183. Berlin: Springer-Verlag, 2010: 321 - 328.

[19] PEARL J. Causality: models, reasoning, and inference [M]. Cambridge: Cambridge University Press, 2000.

[20] SPIRTES P, GLYMOUR C, SCHEINES R. Causation, prediction and search [M]. 2nd ed. Cambridge: MIT Press, 2001.


(上接第 1069 页)


[8] ZHOU J, XIONG Z, ZHANG Y, *et al.* Multiseed clustering algorithm based on max-min distance means [J]. Journal of Computer Applications, 2006, 26(6): 1425 - 1427. (周涓,熊忠阳,张玉芳,等. 基于最大最小距离法的多中心聚类算法[J]. 计算机应用,2006,26(6):1425 - 1427)

[9] XIONG Z, CHEN R, ZHANG Y. Effective method for cluster centers' initialization in K-means clustering [J]. Application Research of Computers, 2011, 28(11): 4188 - 4190. (熊忠阳,陈若田,张玉芳. 一种有效的 K-means 聚类中心初始化方法[J]. 计算机应用研究,2011,28(11):4188 - 4190.)

[10] BABAYIGIT B, OZDEMIR R. A modified artificial bee colony algorithm for numerical function optimization [C]// ISCC 2012: Proceedings of the 2012 IEEE Symposium on Computers and Communications. Piscataway: IEEE, 2012: 000245 - 000249.

[11] HE D, JIA R, SHI S. An artificial bee colony optimization algorithm guided by complex method [C]// ISCID 2012: Proceedings of the 2012 Fifth International Symposium on Computational Intelligence and Design. Piscataway: IEEE, 2012, 1: 348 - 351.

[12] BI X, GONG R. Hybrid clustering algorithm based on artificial bee colony and K-means algorithm [J]. Application Research of Computers, 2012, 29(6): 2040 - 2042. (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.)

[13] Center for Machine Learning and Intelligent Systems. Machine learning repository [EB/OL]. [2013 - 09 - 12]. http://archive.ics.uci.edu/ml/datasets/.