

引入改进飞蛾扑火的 K 均值交叉迭代聚类算法

黄鹤^{1,2}, 李昕芮¹, 吴琨¹, 郭璐³, 王会峰¹, 茹锋¹

(1. 长安大学电子与控制工程学院, 710064, 西安; 2. 陕西省道路交通智能检测与装备工程技术研究中心, 710064, 西安; 3. 西北工业大学无人机系统国家工程研究中心, 710072, 西安)

摘要: 针对现有 K 均值聚类(KMC)算法在选取初始聚类中心时随机性较大、全局搜索能力差、聚类精度低等问题,提出了一种引入改进飞蛾扑火的 K 均值交叉迭代聚类(IMFO-KMC)算法。利用最大最小距离积法初始化聚类中心,避免了 KMC 算法对随机初始聚类中心较为敏感的问题;利用样条插值预测的思想改进飞蛾扑火算法,提高了算法的收敛速度及寻优精度;以类内平均距离为适应度函数,引导插值扑火算法优化 KMC 迭代过程中的聚类中心,提高了聚类精度。将 IMFO-KMC 与 KMC、 K -means++ 算法、模糊 c 均值聚类算法在国际标准数据集 Iris、Wine 和 Seeds 上进行了实验对比,结果表明:IMFO-KMC 算法在 Iris 数据集上的性能提升最为明显,相比其他算法准确率提高了 0.67%~4.18%,标准化互信息提高了 1.5%~4.01%。

关键词: 飞蛾扑火算法;聚类中心; K 均值聚类;类内平均距离;最大最小距离积法
中图分类号: TP301.6 **文献标志码:** A
DOI: 10.7652/xjtuxb202009003 **文章编号:** 0253-987X(2020)09-0032-08



Hybrid Iterative K -Means Clustering with Improved Moth-Flame Optimization

HUANG He^{1,2}, LI Xinrui¹, WU Kun¹, GUO Lu³, WANG Huifeng¹, RU Feng¹

(1. School of Electronic Control, Chang'an University, Xi'an 710064, China; 2. Shaanxi Road Traffic Intelligent Detection and Equipment Engineering Technology Research Center, Xi'an 710064, China;
3. UAV National Engineering Research Center, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: A new K -means clustering algorithm with improved moth-flame optimization is proposed to solve the problem that the current K -means clustering (KMC) algorithm has great randomness in selecting the initial clustering center, poor global search ability and low clustering accuracy. The maximum and minimum distance function is adopted to initialize the clustering center to avoid the problem that KMC algorithm is sensitive to the random initial clustering center. Then the moth-flame optimization is improved by spline interpolation to heighten the convergence rate and optimization accuracy. The average distance category is taken as the fitness function to guide interpolation moth-flame optimization to optimize the clustering center in the process of KMC iteration so as to improve the clustering accuracy. Compared with KMC algorithm, K -means++ algorithm and fuzzy c -means clustering algorithm on the international standard data sets Iris, Wine and Seeds, the experimental results show that the IMFO-KMC algorithm achieves the most significant improvement on the Iris data set. Compared with the

other algorithms, the accuracy of IMFO-KMC algorithm is improved by 0.67%-4.18%, and the normalized mutual information is improved by 1.5%-4.01%.

Keywords: moth-flame optimization; cluster center; K-means clustering; average distance
category; maximum and minimum distance product

作为统计学习领域的重要组成部分,聚类分析有着广泛的应用前景,其可以针对不同对象分析差异,并可根据某种特定规则进行模式分类。随着科技的进步,对数据精度的要求逐步提高,提高聚类算法精度成为了许多学者不断研究的热点问题^[1]。

经典K均值聚类(KMC)算法具备收敛快和易于实现的优点,但该算法存在易受初始聚类中心影响^[2]和精度低的问题。为此,许多学者进行了深入的研究,并提出了相应的改进算法,例如:陈兴蜀等提出了利用特征关联度优化传统KMC算法聚类中心初始化的方法^[3];肖洋等给出了一种基于最小方差的自适应K均值初始化方法,利用样本方差得到密度信息,并根据样本密度得到聚类初始中心^[4];郑丹等根据数据点的K-DIST图选取初始中心^[5];张强等根据类内数据对象的重要性得到自适应权值,提高了的聚类精度^[6];于金霞等通过遗传算法和模糊算法改进了聚类算法^[7]。

为了改善K均值初始中心的随机性并提高算法的聚类的性能,本文提出了一种引入改进飞蛾扑火算法的K均值交叉迭代聚类算法。该算法通过最大最小距离积法初始化聚类中心,降低KMC算法对初始聚类中心的敏感性。同时,利用改进的飞蛾扑火算法优化聚类过程,有效提高聚类精度。

1 基于最大最小距离积法的K均值聚类初始化方法

传统KMC算法的初始聚类中心选取随机性较大,影响聚类精度。为改善聚类性能,用最大最小距离积法替代KMC任意选取初始聚类中心的步骤。

1.1 传统K均值聚类算法

KMC的目的是将所给的数据集按照数据类型分为多个子集,具体步骤是:从数据集 $X=\{x_1, x_2, \dots, x_n\}$ 中将任意 k 个数据点划为初始聚类中心 C_j ($j=1, 2, \dots, k$);计算给定数据集 X 中聚类中心 C_j 与其他各点 x_i ($i=1, 2, \dots, n$)之间的欧氏距离,将各点归类到与其最近的聚类中心所属的类别中;计算各子集数据到其对应聚类中心的距离均值作为新的聚类中心。重复上述步骤,直到聚类中心在下次迭代中不变为止。从上述过程可知,传统KMC算法

的初始聚类中心选取随机性较大,鲁棒性较差。

聚类中心 C_j 和 C_j 与 x_i 的欧氏距离分别为

$$C_j = \frac{1}{C_j} \sum_{i=1}^n x_i, \quad j = 1, 2, \dots, k \quad (1)$$

$$d(x_i, C_j) = \sqrt{\sum_{k=1}^d (x_{ik} - C_{jk})^2} \quad (2)$$

1.2 利用最大最小距离积法初始化K均值聚类

针对人工蜂群算法的初始化具有较大随机性的问题,文献[8]提出了最大最小距离积法,并用其初始化算法种群,很好地克服了蜂群初始化的随机性,提高了算法的聚类精度。因此,本文将最大最小距离积法用于KMC算法的聚类中心初始化过程,降低初始聚类中心对算法的影响,具体步骤如下。

步骤1 从集合 M 中任取一个元素作为初始聚类中心 z_1 ,将此元素放入集合 Z 并从 M 中删除。

步骤2 计算更新后 M 中所有元素到 z_1 的距离,选取距离 z_1 最远的元素为 z_2 。

步骤3 将 z_2 放入集合 Z 并从 M 中删除。

步骤4 分别求得更新后 M 中的元素到 Z 中各元素的距离,并放入集合 T 中。

步骤5 找出 M 中各元素对应的集合 T 中的最大和最小距离,求其乘积,将最大乘积对应的元素从 M 中删除,放入 Z 中。若 Z 中元素数小于 k ,则转到步骤3;若 Z 中元素数大于 k ,则初始点选取结束,输出包含 k 个初始点的 Z ,即为求得的初始聚类中心。

可以看出,采用最大最小距离积法能够很好地避免初始聚类中心选取的随机性,提高聚类精度。

2 改进的飞蛾扑火算法

群体智能算法是指通过模拟自然界中的蚂蚁、蜜蜂等生物群体行为机制的一类算法,利用群体智能算法搜索最优解的方法受到了学者们的热切关注。2015年,Mirjalili根据飞蛾围绕火焰做对数螺旋曲线轨迹的行为提出了飞蛾扑火(MFO)算法^[9]。目前,该算法已经在网络、电子、经济模型、工业制造、人脸识别等方面有了具体应用^[10-14],但在聚类算法的应用较少。因此,本文将MFO算法进行了改进,并结合KMC算法的特点设计了相关的聚类算法。

2.1 飞蛾扑火算法

MFO 算法是根据生活中飞蛾围绕火焰做对数螺旋线形式运动的行为构建的数学模型,通过迭代更新飞蛾位置求解得到最优火焰位置^[15]。

2.1.1 飞蛾种群初始化 MFO 算法中, $\mathbf{M}$ 为飞蛾种群, $m_{i,j}$ 为第 i 只飞蛾在第 j 维变量空间的位置, $\mathbf{M}$ 的第 i 行为 $\mathbf{M}_i$ 。 $\mathbf{M}$ 对应的适应度矩阵为 $\mathbf{O}_M$, O_{Mi} 为第 i 只飞蛾的适应度,由适应度函数得到。 $\mathbf{F}$ 用来存储火焰, $\mathbf{F}$ 对应的适应度矩阵为 $\mathbf{O}_F$ 。 $\mathbf{M}$ 、 $\mathbf{O}_M$ 、 $\mathbf{F}$ 、 $\mathbf{O}_F$ 的公式为

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,d} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,d} \\ \vdots & \vdots & & \vdots \\ m_{n,1} & m_{n,2} & \cdots & m_{n,d} \end{bmatrix}; \mathbf{O}_M = \begin{bmatrix} O_{M1} \\ O_{M2} \\ \vdots \\ O_{Mn} \end{bmatrix} \quad (3)$$

$$\mathbf{F} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,d} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,d} \\ \vdots & \vdots & & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,d} \end{bmatrix}; \mathbf{O}_F = \begin{bmatrix} O_{F1} \\ O_{F2} \\ \vdots \\ O_{Fn} \end{bmatrix} \quad (4)$$

式中: n 为飞蛾数及最初的火焰数; d 为样本维度。火焰矩阵 $\mathbf{F}$ 由飞蛾矩阵 $\mathbf{M}$ 根据适应度排序得到,因此火焰矩阵存储的是飞蛾当前搜索得到的最优解。

2.1.2 位置更新机制 飞蛾围绕火焰更新位置的机制可以分为飞蛾扑火和飞蛾弃焰两个过程。

(1) 飞蛾扑火。受趋光生物特性的影响,飞蛾寻找与它距离最近的火焰位置 $\mathbf{F}_j$, 并绕其做对数螺旋曲线运动。对数螺旋曲线的定义为

$$\mathbf{M}_i = \mathbf{D}_i e^{bt_0} \cos(2\pi t_0) + \mathbf{F}_j \quad (5)$$

式中: b 为调整对数螺旋曲线的常数参数; 参数 t_0 为 $[-1, 1]$ 内的一个任意值, 表示飞蛾的下一位置与火焰的远近, $t_0 = -1$ 时两者相距最近, $t_0 = 1$ 时两者相距最远; $\mathbf{D}_i$ 为第 j 个火焰与第 i 只飞蛾之间的距离, 公式为

$$\mathbf{D}_i = |\mathbf{M}_i - \mathbf{F}_j| \quad (6)$$

利用式(5)更新 $\mathbf{M}_i$ 的各元素得到第 i 只飞蛾的新位置, 根据适应度对飞蛾位置进行排序, 更新火焰矩阵, 通过多次迭代搜索得到最优解。飞蛾运动轨迹如图 1 所示。

(2) 飞蛾弃焰。火焰数在迭代过程中自适应减少, 飞蛾逐渐舍弃劣解, 仅围绕保留下来的火焰做旋转运动, 提高算法寻优的收敛速度。当前火焰数为

$$N_F = f_{\text{round}} \left(N - l \frac{N-1}{T'} \right) \quad (7)$$

式中: N 为飞蛾种群的总数; f_{round} 为四舍五入取整函数; T' 为最大迭代数; l 为当前迭代数。

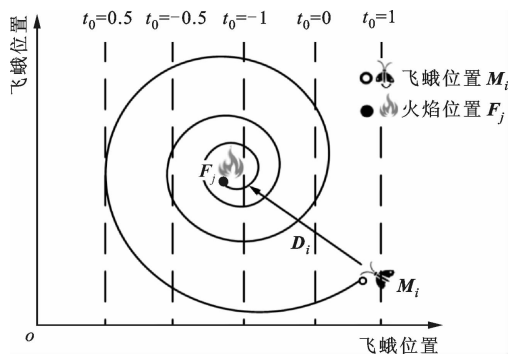


图1 飞蛾运动轨迹

2.2 插值飞蛾扑火算法

在基本飞蛾扑火算法的位置更新机制中,飞蛾寻找距离自己最近的火焰,并对其做式(5)的对数螺旋线运动。由于式(5)中 t_0 为随机数,导致飞蛾及火焰位置的更新具有较大随机性。考虑到火焰历史位置的影响,本文提出了一种插值飞蛾扑火(IMFO)算法,通过火焰历史位置来插值预测下一代火焰位置,加快了飞蛾扑火算法的收敛速度,并提高了精度。IMFO 中利用的样条插值是由一系列不连续的数据点拟合出一条光滑曲线,并将其用来预测估计那些无法从自变量直接获取因变量的方法。

设 f_j^t 为第 t ($t \geq 0$) 代中第 j 个火焰的位置。保存未来 g 代火焰 j 的位置 $[f_j^{t+0}, f_j^{t+1}, \dots, f_j^{t+g}]$, 对第 $g+1$ 代火焰 j 的位置进行插值预测,然后将第 $g+1$ 代预测得到的火焰 j 的位置与基本飞蛾扑火算法迭代更新得到的火焰 j 的位置进行比较,取较好的位置作为最终第 $g+1$ 代火焰 j 的位置 f_j^{t+g+1} 。

本文进行了大量实验,考虑到算法的复杂程度,最终设定只存储迭代过程中近 3 代的最优火焰的历史位置,即 $g=3$, 对第 4 代最优火焰的位置进行插值预测。最优火焰位置通过对火焰适应度排序比较得到。每 3 次迭代后进行一次插值预测,每 4 次为一轮,依次循环,直到算法达到预设的迭代数。

本文提出的 IMFO 算法仅对原始飞蛾扑火算法迭代过程中的最优火焰位置进行插值预测,其他火焰位置仍由飞蛾通过随机搜索得到,保留了 MFO 算法的随机性。此外,IMFO 算法迭代过程中,飞蛾围绕各自对应的火焰进行旋转搜索,并将飞蛾随机搜索得到的新位置和火焰位置重新比较排序,更新火焰列表,以免所有飞蛾围绕非真实全局最优的火焰旋转而出错,导致寻优陷入局部极值。由此可以看出,插值飞蛾扑火算法不仅增强了迭代搜索过程中最优值的局部搜索能力,而且具有较好的广域搜索能力,保留了飞蛾扑火算法的全局搜索能力,加快

了算法的收敛速度,提高了算法寻优的精确度。IMFO 算法的流程如图 2 所示。

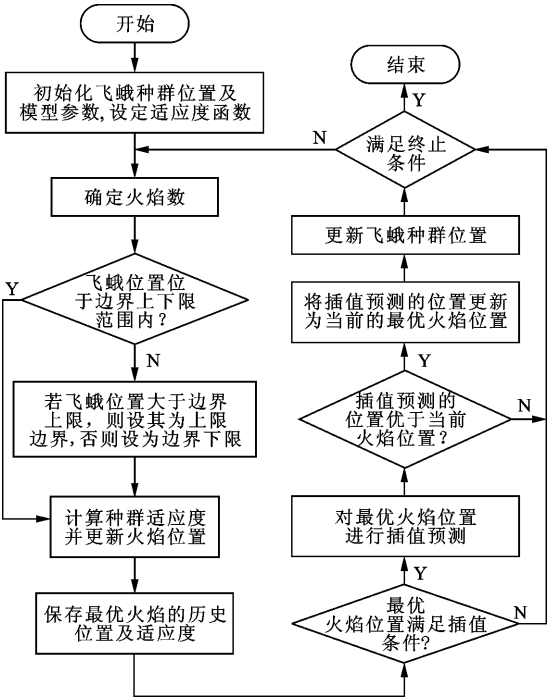


图 2 IMFO 算法流程

3 插值飞蛾扑火算法与 K 均值聚类算法的交叉迭代

本文结合 IMFO 算法与 KMC 算法的特点,提出了一种引入改进飞蛾扑火的 K 均值交叉迭代聚类(IMFO-KMC)算法。IMFO-KMC 算法利用插值飞蛾扑火算法优化聚类中心,需要将数据的分类信息引入到 IMFO 算法中。因此,利用类内平均距离定义 IMFO 算法的适应度函数,具体形式为

$$G_{F_i} = J_i / C_{N_i}, i = 1, 2, \dots, k \tag{8}$$

$$J_i = \sum_{x_j \in C_j} d(x_j, C_j) \tag{9}$$

式中: J_i 为第 i 类的点到该类聚类中心的距离之和; C_{N_i} 为第 i 类点的数量。

从式(8)可以看出:在类内点数据量相同的情况下,聚类中心与类内其他点的距离之和越小,则适应度越小;在聚类中心到类内其他点的距离之和不变的情况下,类内点数据量越多,则适应度越小,即得出的聚类中心越精确。因此,该适应度函数能够很好地评价数据的聚类效果,并且作为 IMFO 算法的适应度函数能够很好地引导飞蛾种群进化方向,得到良好的优化效果。IMFO-KMC 算法的基本步骤为:

- (1)输入数据集 M ,根据其确定类别数 k ;
- (2)用最大最小距离积法初始化聚类中心;

(3)计算 M 中的某个点到每个聚类中心的欧氏距离,当该点与某一初始聚类中心的距离最小时,则把该点与这个初始聚类中心归为一类,依照此方法将所有的点归并到对应的类中;

(4)使用 IMFO 算法重新计算归并之后的每个类的新聚类中心;

(5)判断迭代数是否达到规定数 I_2 ,若满足,则循环结束,程序结束,否则,跳到步骤(3),继续执行。

IMFO-KMC 算法的流程如图 3 所示。

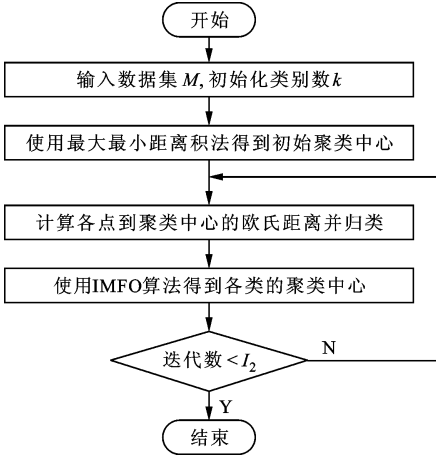


图 3 IMFO-KMC 算法的流程

4 实验结果分析与应用

实验平台采用 CPU 为 Intel Core i7 2.60 GHz、内存为 16 GB 的计算机,操作系统为 Windows 10,编译软件为 Matlab R2018a。

4.1 插值飞蛾扑火算法性能测试

为了验证 IMFO 算法的优化性能,将 IMFO 算法与文献[9]提出的 MFO 算法、文献[16]提出的粒子群(PSO)算法、文献[17]提出的遗传粒子群(GAPSO)算法和文献[18]的模拟退火粒子群(SAPSO)算法进行了实验对比,实验采用单峰函数 Sphere、Schwefel's2.22 及多峰函数 Ackley、Penalized1 这 4 个基准测试函数^[9]。设各种群大小均为 30,迭代 500 次,PSO、GAPSO 及 SAPSO 算法中惯性权值 $\omega = 0.5$,学习因子 $c_1 = c_2 = 2$,GAPSO 算法的交叉概率 $p_c = 0.6$,变异概率 $p_m = 0.01$,SAPSO 算法中模拟退火降温常数为 0.5。

利用 5 种算法针对每个基准测试函数进行了独立的 20 次实验。不同算法在不同测试函数上的种群最优适应度变化曲线如图 4 所示,20 次实验得到的最优值以及每次实验得到的最优适应度的均值和标准差如表 1 所示。

由图 4 可以看出:在优化精度方面,本文提出的 IMFO 与 MFO 算法的实验结果明显优于 PSO、

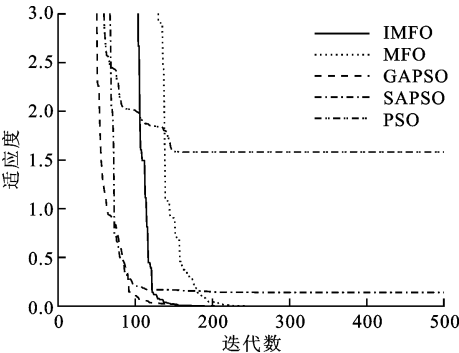
GAPSO 与 SAPSO 算法的;在收敛速度方面,PSO 与 SAPSO 算法虽收敛较快,但优化精度较低, GAPSO 算法由于在迭代过程容易滞留在局部极值,收敛速度及优化精度不稳定,而对于优化精度较高的 MFO 与 IMFO 算法,IMFO 算法能够以更快的速度搜索到全局的最优解。

表 1 不同算法对 4 种函数的实验结果

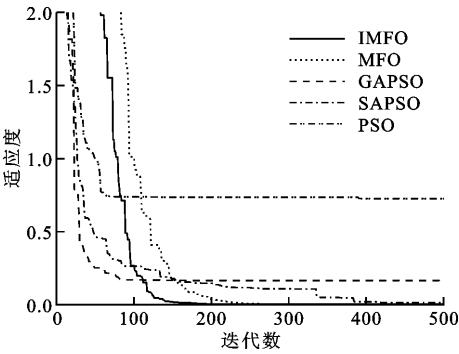
测试函数	算法	最优 适应度	适应度 均值	适应度 标准差
Sphere	PSO	0.116 7	4.623 2	5.679 7
	SAPSO	0.001 3	0.362 8	1.684 7
	GAPSO	5.12×10^{-11}	1.78×10^{-8}	3.79×10^{-8}
	MFO	3.97×10^{-16}	5.07×10^{-13}	1.36×10^{-12}
	IMFO	2.70×10^{-19}	2.82×10^{-15}	1.05×10^{-14}
Schwefel's- 2.22	PSO	0.478 8	2.053 1	1.407 8
	SAPSO	0.155 7	0.705 5	0.578 6
	GAPSO	9.89×10^{-4}	0.048 2	0.064 0
	MFO	4.66×10^{-10}	7.66×10^{-9}	8.74×10^{-9}
	IMFO	1.50×10^{-13}	1.10×10^{-10}	2.30×10^{-10}
Ackley	PSO	2.582 3	5.151 8	1.519 0
	SAPSO	1.783 6	3.978 6	1.384 4
	GAPSO	1.155 1	2.674 9	0.741 1
	MFO	1.39×10^{-8}	2.92×10^{-7}	3.06×10^{-7}
	IMFO	6.14×10^{-11}	2.70×10^{-8}	4.91×10^{-8}
Penalized1	PSO	0.047 9	3.248 2	3.089 1
	SAPSO	0.057 9	1.252 1	1.280 8
	GAPSO	7.41×10^{-10}	0.317 6	0.662 6
	MFO	1.10×10^{-16}	3.06×10^{-14}	7.65×10^{-14}
	IMFO	3.70×10^{-19}	1.30×10^{-14}	4.74×10^{-14}

由表 1 可以看出:对于单峰函数 Sphere 和 Schwefel's2.22,IMFO 算法得到的最优适应度达到了 10^{-19} 和 10^{-13} ,明显优于其他 4 种算法的,在适应度均值方面,IMFO 算法的精度达到了 10^{-15} 和 10^{-10} ,远高于其他 4 种算法的,在适应度标准差方面,IMFO 算法的标准差均最小,优化效果最好;对于多峰函数 Ackley 和 Penalized1,IMFO 算法的最优精度达到了 10^{-11} 和 10^{-19} ,均高于其他 4 种算法 3 个数量级以上,同时在均值上也表现最优,在适应度标准差方面,IMFO 算法对 2 个多峰函数测试的标准差值均最小,稳定性最好。

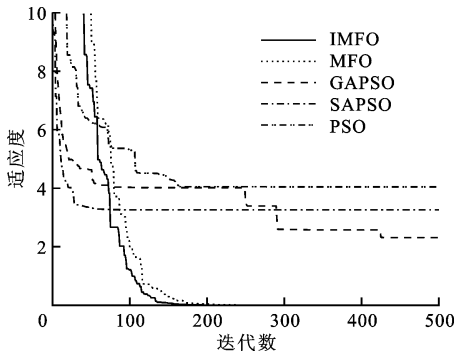
由 4.1 小节可知:与原 MFO 算法相比,IMFO 算法能够在加强局部搜索能力的同时提高广域搜索能力;与其他 3 种算法相比,IMFO 算法的优化精度



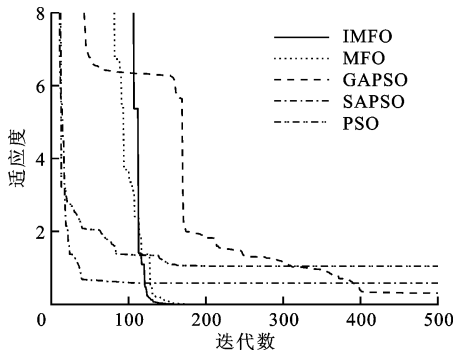
(a) Sphere 函数



(b) Schwefel's2.22 函数



(c) Ackley 函数



(d) Penalized1 函数

图 4 不同算法在不同测试函数上的种群最优适应度变化曲线

更高且稳定性更好。

4.2 引入改进飞蛾扑火的 K 均值交叉迭代聚类算法性能评价

为了验证本文 IMFO-KMC 算法的聚类效果,将 IMFO-KMC 算法与 KMC 算法、K-means++ 算法^[19]和模糊 c 均值(FCM)聚类算法^[20]进行实验对比,实验数据采用 Iris、Wine 和 Seeds 国际标准数据集。设定各算法实验次数为 20,最大迭代数均为 100,IMFO-KMC 算法中的飞蛾扑火算法的迭代数为 50,FCM 算法中的权重指数 m 取 2.126。标准数据集特征如表 2 所示。

表 2 标准数据集特征

数据集	样本数	属性维数	类别数	各类别样本数
Iris	150	4	3	50,50,50
Wine	178	13	3	59,71,48
Seeds	210	7	3	70,70,70

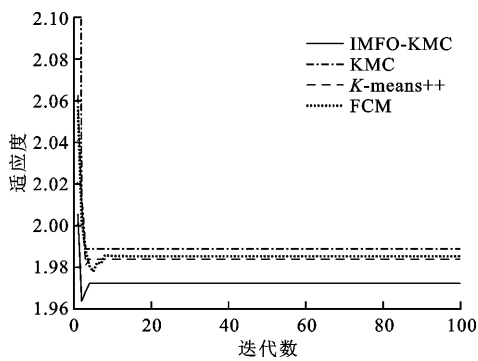
采用准确率 P_{ACC} 、调整兰德系数 S_{ARI} 、标准化互信息 R_{NMI} 对不同聚类算法进行测评。 P_{ACC} 表示聚类结果和真实数据类别的一致性; S_{ARI} 表示计算结果与真实数据的相似度; R_{NMI} 表示两组数据之间的关联度。各指标越接近于 1 则表示聚类效果越好。

由各算法根据式(7)所示的适应度函数得到收敛曲线,结果如图 5 所示。采用聚类评价指标对各算法的聚类性能进行评价,结果如表 3 所示,表中各评价指标由 20 次独立实验数据求均值得到。

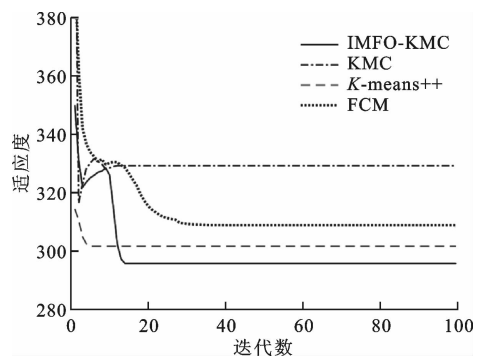
表 3 不同算法对 3 个数据集的评价指标 %

数据集	算法	P_{ACC}	S_{ARI}	R_{NMI}
Iris	IMFO-KMC	90.00	74.37	76.36
	KMC	85.82	67.90	72.25
	K-means++	88.93	72.28	74.86
	FCM	89.33	72.94	74.82
Wine	IMFO-KMC	70.79	37.15	41.93
	KMC	69.63	36.43	42.71
	K-means++	70.05	36.92	42.83
	FCM	69.66	36.02	42.13
Seeds	IMFO-KMC	89.52	71.66	69.49
	KMC	89.38	71.47	69.57
	K-means++	89.48	71.60	70.02
	FCM	89.52	71.66	69.49

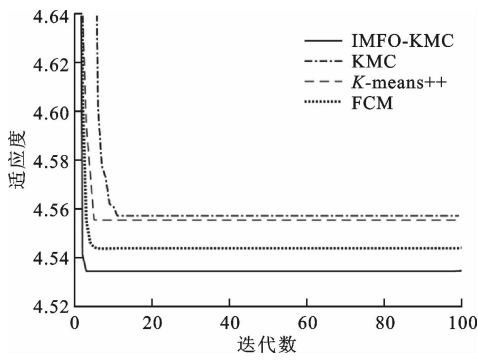
由图 5 可以看出,相比于 KMC、K-means++ 和 FCM 算法,本文 IMFO-KMC 算法在低维数据集 Iris、Seeds 和高维数据集 Wine 下,聚类结果均优于



(a) Iris 数据集



(b) Wine 数据集



(c) Seeds 数据集

图 5 不同算法在不同数据集上的适应度变化曲线

其他 3 种算法的。由表 3 可以看出:在 Iris 数据集上,本文 IMFO-KMC 算法测得的 P_{ACC} 、 S_{ARI} 、 R_{NMI} 均为最优,相比其他算法的提升了 0.8%~9.5%;在 Wine 和 Seeds 数据集上,本文算法的 P_{ACC} 、 S_{ARI} 最优,其中在 Wine 数据集上,本文算法的 P_{ACC} 比其他 3 种算法的提升了 1.0%~1.7%, S_{ARI} 提升了 0.6%~3.1%。从图 5 和表 3 可知,本文算法最优。

由 4.2 小节可知:KMC 算法的稳定性最差、精度最低;K-means++ 算法要求初始聚类中心的距离尽可能的远,所以相比 KMC 算法较稳定、精度略有提升;FCM 算法引入模糊控制的概念,使得适应度变化平稳,但精度较低;本文 IMFO-KMC 算法可

以快速跳出局部最优,稳定性较好,聚类准确度较高。

4.3 引入改进飞蛾扑火的 K 均值交叉迭代聚类算法的应用

聚类算法作为常见的数据分析方法,在生物、工程、商业等众多领域具有重要的作用,如:商业中依据客户消费特征将客户分为不同消费类群;根据学生各科成绩对其进行分类从而制定有针对性的培养计划;依据不同的经济发展指标聚类划分区域等。

选取文献[21]中的 data11-03 数据进行实验,该数据为在 $12\times 28.350\text{ g}$ (1 盎司 = 28.350 g)啤酒

中,包含热量(1 卡 = 4.186 J)、钠质量、酒精度(乙醇的体积分数)、价格 4 个维度的 20 种啤酒的数据。利用本文提出的 IMFO-KMC 算法对该数据集进行分析,结果如表 4 所示。可以看出:第 1 类啤酒的序号为 1、2、3、5、6、8、11、13、14、15、17、18,这类啤酒的热量和钠质量偏高、酒精度居中;第 2 类啤酒的序号为 16、19,这类啤酒的热量和酒精度偏低;第 3 类啤酒的序号为 4、7,这类啤酒的热量和酒精度较高;第 4 类啤酒的序号为 9、10、12、20,这类啤酒的酒精度低于第 1、3 类的,高于第 4 类的,价格居中。

表 4 啤酒数据特征及分类结果

序号	啤酒	热量/J	钠质量/mg	酒精度/%	价格/美元	类别
1	Budweiser	144×4.186	15	4.7	0.43	1
2	Schlitz	151×4.186	19	4.9	0.43	1
3	Ionenbrau	157×4.186	15	4.9	0.48	1
4	Kronensourc	170×4.186	7	5.2	0.73	3
5	Heineken	152×4.186	11	5.0	0.77	1
6	Old-milnaukee	145×4.186	23	4.6	0.26	1
7	Aucsberger	175×4.186	24	5.5	0.4	3
8	Strchs-bohemi	149×4.186	27	4.7	0.42	1
9	Miller-lite	99×4.186	10	4.3	0.43	4
10	Sudeiser-lich	113×4.186	8	3.7	0.44	4
11	Coors	140×4.186	18	4.6	0.44	1
12	Coorrs-light	102×4.186	15	4.1	0.46	4
13	Michelob-light	135×4.186	11	4.2	0.5	1
14	Becks	150×4.186	19	4.7	0.76	1
15	Kirin	149×4.186	6	5.0	0.79	1
16	Pabst-extra-light	68×4.186	15	2.3	0.38	2
17	Hamms	136×4.186	19	4.4	0.43	1
18	Heilemans-old-sky	144×4.186	24	4.9	0.43	1
19	Olympia-gold-light	72×4.186	6	2.9	0.46	2
20	Schlitz-light	97×4.186	7	4.2	0.47	4

5 结 论

本文以最大最小距离积法对聚类中心进行初始化解,解决了 KMC 算法对初始聚类中心敏感的缺陷。引入样条插值预测算法及类内平均距离对 MFO 算法进行改进,解决了飞蛾种群个别解较差及容易陷入局部最优解等问题。将 IMFO 与 KMC 算法交叉迭代聚类,提高了鲁棒性和聚类精度。但是,本文还存在 IMFO 算法在求解某些复杂函数时精度不足的遗留问题,有待在未来的工作中进一步改善。

参考文献:

[1] YU Q Y, LUO Y L, CHEN C M, et al. Outlier-eliminated k -means clustering algorithm based on differential privacy preservation [J]. Applied Intelligence, 2016, 45(4): 1179-1191.

[2] PATEL G K, DABHI V K, PRAJAPATI H B. Clustering using a combination of particle swarm optimization and K -means [J]. Journal of Intelligent Systems, 2017, 26(3): 457-469.

[3] 陈兴蜀, 吴小松, 王文贤, 等. 基于特征关联度的 K -means 初始聚类中心优化算法 [J]. 四川大学学报(工

- 程科学版), 2015, 47(1): 13-19.
- CHEN Xingshu, WU Xiaosong, WANG Wenxian, et al. An improved initial cluster centers selection algorithm for K -means based on features correlative degree [J]. Journal of Sichuan University (Engineering Science Edition), 2015, 47(1): 13-19.
- [4] 肖洋, 李平, 王鹏, 等. 基于最小方差的自适应 K -均值初始化方法 [J]. 长春理工大学学报(自然科学版), 2015, 38(5): 140-144, 149.
- XIAO Yang, LI Ping, WANG Peng, et al. An adaptive K -means initialization method based on minimum deviation [J]. Journal of Changchun University of Science and Technology (Natural Science Edition), 2015, 38(5): 140-144, 149.
- [5] 郑丹, 王潜平. K -means 初始聚类中心的选择算法 [J]. 计算机应用, 2012, 32(8): 2186-2188, 2192.
- ZHENG Dan, WANG Qianping. Selection algorithm for K -means initial clustering center [J]. Journal of Computer Applications, 2012, 32(8): 2186-2188, 2192.
- [6] 张强, 王红卫, 陈游, 等. 基于自适应权重的 RFCM 聚类算法 [J]. 微电子学与计算机, 2016, 33(12): 80-84.
- ZHANG Qiang, WANG Hongwei, CHEN You, et al. Rough fuzzy c -means clustering algorithm based on self-adaptive weights [J]. Microelectronics & Computer, 2016, 33(12): 80-84.
- [7] 于金霞, 蔡自兴, 段琢华. 一种用于激光雷达特征提取的改进遗传聚类算法 [J]. 模式识别与人工智能, 2008, 21(4): 535-540.
- YU Jinxia, CAI Zixing, DUAN Zhuohua. An improved genetic clustering algorithm for feature extraction of laser scanner [J]. Pattern Recognition and Artificial Intelligence, 2008, 21(4): 535-540.
- [8] 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, 1088.
- YU Jinping, ZHENG Jie, MEI Hongbiao. K -means clustering algorithm based on improved artificial bee colony algorithm [J]. Journal of Computer Applications, 2014, 34(4): 1065-1069, 1088.
- [9] MIRJALILI S. Moth-flame optimization algorithm: a novel nature-inspired heuristic paradigm [J]. Knowledge-Based Systems, 2015, 89: 228-249.
- [10] KHALILPOURAZARI S, PASANDIDEH S H R. Multi-item EOQ model with nonlinear unit holding cost and partial backordering: moth-flame optimization algorithm [J]. Journal of Industrial and Production Engineering, 2017, 34(1): 42-51.
- [11] SINGH P, PRAKASH S. Optical network unit placement in Fiber-Wireless (FiWi) access network by Moth-Flame optimization algorithm [J]. Optical Fiber Technology, 2017, 36: 403-411.
- [12] ALLAM D, YOUSRI D A, ETEIBA M B. Parameters extraction of the three diode model for the multi-crystalline solar cell/module using moth-flame optimization algorithm [J]. Energy Conversion and Management, 2016, 123: 535-548.
- [13] YILDIZ B S, YILDIZ A R. Moth-flame optimization algorithm to determine optimal machining parameters in manufacturing processes [J]. Materials Testing, 2017, 59(5): 425-429.
- [14] ZHANG L, MISTRY K, NEOH S C, et al. Intelligent facial emotion recognition using moth-firefly optimization [J]. Knowledge-Based Systems, 2016, 111: 248-267.
- [15] GUO L, LI X R, HUANG H, et al. Intelligent penetration of UAV based on moth suppression algorithm [C] // 2018 Chinese Automation Congress (CAC). Piscataway, NJ, USA: IEEE, 2018: 4044-4049.
- [16] SHINODA M, MIYATA Y. PSO-based stability analysis of unreinforced and reinforced soil slopes using non-circular slip surface [J]. Acta Geotechnica, 2019, 14(3): 907-919.
- [17] GHAMISI P. A novel method for segmentation of remote sensing images based on hybrid GAPSO [J]. International Journal of Computer Applications, 2011, 29(2): 7-14.
- [18] 靳一, 王继武, 吴乐南. 基于 IM-SAPSO 和 SVM 的 EBPSK 检测器设计 [J]. 应用科学学报, 2012, 30(2): 141-145.
- JIN Yi, WANG Jiwu, WU Lenan. EBPSK signal detector based on IM-SAPSO and SVM [J]. Journal of Applied Sciences, 2012, 30(2): 141-145.
- [19] BAHMANI B, MOSELEY B, VATTANI A, et al. Scalable k -means++ [J]. Proceedings of the VLDB Endowment, 2012, 5(7): 622-633.
- [20] 肖春景, 张敏. 基于减法聚类与模糊 c -均值的模糊聚类研究 [J]. 计算机工程, 2005, 31(S07): 135-137.
- XIAO Chunjing, ZHANG Min. Research on fuzzy clustering based on subtractive clustering and fuzzy c -means [J]. Computer Engineering, 2005, 31(S7): 135-137.
- [21] 卢纹岱. SPSS for Windows 统计分析 [M]. 北京: 电子工业出版社, 2006: 429-438.

(编辑 陶晴 刘杨)