



基于混合身份搜索黏菌优化的模糊C-均值聚类算法

贾鹤鸣, 张棕淇, 姜子超, 冯榆淇

引用本文:

贾鹤鸣,张棕淇,姜子超,冯榆淇. 基于混合身份搜索黏菌优化的模糊C-均值聚类算法[J]. 智能系统学报, 2022, 17(5): 999–1011.

JIA Heming,ZHANG Zongqi,JIANG Zichao,FENG Yuqi. An optimization fuzzy C-means clustering algorithm based on the hybrid identity search and slime mold algorithms[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(5): 999–1011.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202107011>

您可能感兴趣的其他文章

基于改进PSO和FCM的模糊辨识

Fuzzy identification based on improved PSO and FCM

智能系统学报. 2019, 14(2): 378–384 <https://dx.doi.org/10.11992/tis.201707025>

一种融合邻域信息的模糊C-均值图像分割算法

Fuzzy C-means image segmentation algorithm incorporating neighborhood information

智能系统学报. 2019, 14(2): 273–280 <https://dx.doi.org/10.11992/tis.201712035>

基于混合距离学习的鲁棒的模糊C均值聚类算法

Robust FCM clustering algorithm based on hybrid-distance learning

智能系统学报. 2017, 12(4): 450–458 <https://dx.doi.org/10.11992/tis.201607019>

基于改进的遗传算法的模糊聚类算法

Fuzzy clustering algorithm based on the improved genetic algorithm

智能系统学报. 2015(4): 627–635 <https://dx.doi.org/10.3969/j.issn.1673-4785.201503033>

融合并行混沌萤火虫算法的K-调和均值聚类

K-harmonic means clustering merged with parallel chaotic firefly algorithm

智能系统学报. 2015, 10(6): 872–880 <https://dx.doi.org/10.11992/tis.201505043>



微信公众平台



期刊网址

DOI: 10.11992/tis.202107011

网络出版地址: <https://kns.cnki.net/kcms/detail/23.1538.TP.20220623.1004.002.html>

基于混合身份搜索黏菌优化的模糊 C-均值聚类算法

贾鹤鸣¹, 张棕淇², 姜子超², 冯榆淇²

(1. 三明学院 信息工程学院, 福建 三明 365004; 2. 东北林业大学 机电工程学院, 黑龙江 哈尔滨 150040)

摘要: 针对模糊 C-均值聚类算法 (fuzzy C-means clustering, FCM) 对于初始化聚类中心敏感、收敛速度慢, 聚类效果不稳定且容易陷入局部最优等问题, 提出了一种将黏菌 (SMA) 与青少年身份搜索 (AISA) 相融合的自适应优化模糊 C-均值算法 (AISA-SMA-FCM)。该算法首先通过引入 AISA 算法中的青少年社会机制, 改善 SMA 算法中的全局搜索和局部开发性能。克服了 SMA 对于高维数据及部分混峰数据不敏感的缺陷, 通过标准测试函数验证改进后的混合 AISA-SMA 算法寻优求解性能更为优秀; 其次此算法用于 FCM 聚类算法的迭代机制中, 通过将 AISA-SMA 聚类环节加入 FCM 算法聚类中心迭代过程中, 使 FCM 算法获得自适应优化算法相同的特性, 即算法在每次迭代中都将具有探索和开发两个过程, 并依据循环迭代次数调节比重, 求解聚类结果; 最后通过 UCI 标准数据集仿真测试, 利用适应度平均值与聚类正确率评价所提算法的稳定性与有效性, 结果表明, AISA-SMA 算法用于 FCM 聚类问题效果较好, AISA-SMA-FCM 算法较其他聚类方式和相应的优化技术具有收敛速度快、求解精度高的优点。

关键词: 模糊 C-均值聚类; 启发式优化; 黏菌算法; 青少年身份算法; 社会机制; 模糊策略; UCI 数据库; 融合算法
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673-4785(2022)05-0999-13

中文引用格式: 贾鹤鸣, 张棕淇, 姜子超, 等. 基于混合身份搜索黏菌优化的模糊 C-均值聚类算法 [J]. 智能系统学报, 2022, 17(5): 999-1011.

英文引用格式: JIA Heming, ZHANG Zongqi, JIANG Zichao, et al. An optimization fuzzy C-means clustering algorithm based on the hybrid identity search and slime mold algorithms[J]. CAAI transactions on intelligent systems, 2022, 17(5): 999-1011.

An optimization fuzzy C-means clustering algorithm based on the hybrid identity search and slime mold algorithms

JIA Heming¹, ZHANG Zongqi², JIANG Zichao², FENG Yuqi²

(1. School of Information Engineering, Sanming University, Sanming 365004, China; 2. College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: The fuzzy C-means clustering algorithm (FCM) has many shortcomings, such as high sensitivity to initial clustering centers, slow convergence, unstable clustering results, and ease of falling into local optimums. To address these problems, an adaptive optimization fuzzy C-means algorithm based on the fusion of the slime mold algorithm and the adolescent identity search algorithm (AISA-SMA-FCM) is proposed in this paper. First, the algorithm improves the global search and local development performance of SMA by introducing the youth social mechanism in AISA, thus overcoming the limitation of SMA of not being sensitive to high-dimensional data and some mixed peak data, along with an excellent performance of the improved AISA-SMA algorithm in optimizing and solving problems verified by the standard test function. Second, the novel proposed algorithm adds the AISA-SMA clustering link to the iteration process of FCM, enabling it to have the same characteristics as the adaptive optimization algorithm—undergoing two processes of exploration and development in each iteration, adjust the proportion as per the number of iterations, and solve the clustering results. Finally, through the simulation test on the UCI standard data sets, the stability and effectiveness of the algorithm are evaluated based on the fitness average and the clustering accuracy rate. The results show that the AISA-SMA algorithm demonstrates a good effect when used in the iteration mechanism of the FCM algorithm, with a faster convergence speed and a higher solution accuracy when compared with other clustering methods and corresponding optimization technologies.

Keywords: fuzzy C-mean clustering; heuristic optimization; slime mold algorithm (SMA); adolescent identity search algorithm (AISA); social mechanism; fuzzy strategy; UCI database; fusion algorithm

收稿日期: 2021-07-06. 网络出版日期: 2022-06-24.

基金项目: 福建省自然科学基金面上项目 (2021J011128); 福建省本科高校教育教学改革研究项目 (FBJG20210338); 三明市科技计划引导性项目 (2021-S-8); 三明学院教育教学改革重点项目 (J2010305); 三明学院高教研究课题 (SHE2013); 福建省农业物联网应用重点实验室开放研究基金项目 (ZD2101).

通信作者: 贾鹤鸣. E-mail: jiaheminglucky99@126.com.

在计算科学中, 机器学习是计算方法的重要研究领域。随着人工智能技术的飞速发展, 海量数据涌现于人类社会, 数据较易存储, 但数据处理的计算准确率与效率受数据维度与数目等原因的影响较大, 而机器学习可以有效提升数据处

理的效果^[1-2],但是要想通过机器学习获得准确的模型,就需要大量具有标签和规律特点的数据。这种需求正是数据预处理过程所能提供的。所以将优化技术用于数据预处理过程以获得高质量数据,具有广阔的应用场景及需求。如何进一步显著提高数据预处理的性能已成为目前机器学习领域中的热点内容。聚类分析因其独特的优势:能够在无标签、海量的数据中快速、准确挖掘数据自身的规律及特点,完成数据统计与分类任务,进而大幅度提升机器学习的准确性与效率^[3],成为数据预处理的重要技术之一。

作为聚类技术的其中一种方法,无监督机器学习聚类的实质是将物理或抽象对象的集合分成由类似的对象组成的多个类别的过程,这些由聚类所生成的簇是一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中的对象相异,能够揭示数据内部特有的信息价值^[4],完成数据分类任务。经过诸多学者深入研究,目前的主流聚类分析算法包含:基于密度的方法、基于网格的方法、基于层次的方法、基于模型的方法和基于划分式方法等。其中,C-均值聚类^[5]是一种广泛应用于探索性数据分析、数据挖掘、机器学习、图像检索和数学编程中的聚类技术,但此方法计算和分类都局限于硬子空间集合,所以其存在时间复杂度高、对噪声不敏感,处理数据维度较低等缺点与局限性。为改善其聚类效果,Dunn等^[6]基于模糊数学的概念提出了模糊C-均值聚类(fuzzy C-mean, FCM)技术,使得分类问题不再局限于硬子空间集,每个样本可同时隶属于不同的类别,有效克服了C-均值聚类算法的缺点与局限,具有良好的聚类效果,但FCM依旧没有解决经典聚类技术中所存在的初始化敏感、迭代的过程中易陷入局部最优、稳定性不足等问题。Kennedy等^[7]通过模仿鸟群的觅食运动方式提出了一种粒子群优化(particle swarm optimization, PSO)算法,并将其应用于聚类问题中,效果良好;Niknam等^[8]将蚁群优化和模拟退火相结合,提出了一种有效的混合进化算法,同样应用于聚类问题,提高了数据分类的准确率;Ozturk等^[9]利用改进后的二进制人工蜂群算法对聚类问题进行优化,显著改善了数据处理的效率;蒙祖强等^[10]提出一种混合蛙跳与阴影集优化的粗糙模糊聚类算法,效果显著;张新明等^[11]将灰狼优化(grey wolf optimizer, GWO)算法用于FCM的聚类问题,提升了实际聚类性能。以上所述研究表明,使用启发式群智能优化算法对聚类分析问题进行优化

可获得良好的分类准确率与效率,因此本文基于以上研究的有效性以及模糊C均值自身不足引入新颖的启发式群智能优化对FCM聚类技术加以改进研究。

启发式群智能优化发展已近20余年,黏菌优化算法(slime mould algorithm, SMA)是Li等^[12]最近基于黏菌觅食的策略提出的一种新颖、有效的优化算法,因其优化求解效果好,已有许多学者将其算法用于实际的工程优化问题中,如Zhou等^[13]将SMA与WOA^[14]算法相结合用于X射线图像分割;Mostafa等^[15]将SMA模型用于太阳能电池板最优模型参数提取;Mohamed Abdel-Basset等将SMA算法进行优化得到了一种高效的二进制黏菌算法^[16]。Bogara等^[17]基于青少年的成长轨迹进行建模分析,提出了一种青少年身份优化算法(adolescent identity search algorithm, AISA),其中特有的社会行为、榜样机制与身份特征更新策略提高了群智能类算法的优化求解精度。近年来启发式群智能优化算法多数应用于实际工程问题的参数优化,如贾鹤鸣等^[1,3]利用此类算法同步优化支持向量机的参数选取与特征选择,极大改善了数据分类的准确率,但此类算法难以改变自身的逻辑收敛效果,而聚类算法的核心是将一组数据集划分成各个子集,两者均为闭环迭代计算模式,因此本文将SMA优化作为一种聚类的迭代策略融入FCM算法中,对其进行计算优化。但SMA算法的收敛是以各黏菌领域作为主要搜索区域,若各粒子之间信息交流不够紧密,容易导致收敛速率慢、求解精度低,搜索域重复等问题。因此本文通过融合AISA算法中的社会群体性迭代方式,提出了混合青少年搜索黏菌优化(adolescent identity search algorithm-slime mould algorithm, AISA-SMA)算法,以提高SMA算法寻优求解性能,减少算法的不稳定因素。在启发式群智能优化FCM聚类中,利用AISA-SMA算法优秀的收敛效果,将其引入FCM聚类迭代过程中,提出新的融合聚类算法AISA-SMA-FCM来改善算法精度,以获得更优秀的聚类效果。

本文首先将AISA算法与SMA算法进行混合改进研究,改善SMA算法的高维及混峰求解能力,并通过测试函数仿真实验证明所提混合算法求解精度高,收敛能力强;随后将本文混合AISA-SMA算法应用于FCM聚类技术当中,使得FCM算法每次迭代都具有开发和搜索两种行为,并对经典UCI数据集进行聚类仿真测试,分析所得适应度值等指标,对比其他聚类方法证明AISA-

SMA-FCM 算法具有更优秀的聚类能力。

1 相关算法

1.1 模糊C-均值算法(FCM)

模糊C-均值聚类算法是用隶属度来确定每个数据点属于某个聚类程度的一种聚类算法,首先给出样本观测数据矩阵:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

式中矩阵每行为一个样本,每列为一个变量的 n 个观测值,即式(1)是由 n 个样本的 p 个观测值构成的矩阵。

聚类目的是将这 n 个样本划分为 c 类($2 < c < n$),记 $V = (v_1, v_2, \dots, v_c)$ 为 c 个类的聚类中心,其中 $v_i = (v_{i1}, v_{i2}, \dots, v_{ip})$ ($i = 1, 2, \dots, c$)。在模糊C-均值聚类中,每个样本是以一定的隶属度划分为某一类,其定义目标函数:

$$J(\mathbf{U}, \mathbf{V}) = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^m d_{ik}^2 \quad (2)$$

其中, $\mathbf{U} = (u_{ik})_{c \times n}$ 为隶属度矩阵, $d_{ik} = \|x_k - v_i\|$ 。 $J(\mathbf{U}, \mathbf{V})$ 表示各类中的样本到该聚类中心的加权平方距离之和,权重是样本 x_k 属于第 i 类隶属度的 m 次方。模糊C-均值聚类算法的聚类准则是求取 \mathbf{U} 、 \mathbf{V} ,使得 $J(\mathbf{U}, \mathbf{V})$ 取得最小值。模糊C-均值聚类算法的具体步骤如下:

1) 确定类的个数 c ,幂指数 $m(m > 1)$ 和初始隶属度矩阵 $\mathbf{U}^{(0)} = (u_{ik}^{(0)})$,通常可以取 $[0, 1]$ 上均匀分布随机数来确定隶属度矩阵 $\mathbf{U}^{(0)}$,令 $l = 1$ 来表示第1步迭代(l 迭代次数)。

2) 计算第1步的聚类中心 $V^{(l)}$:

$$v_i^{(l)} = \frac{\sum_{k=1}^n (u_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^n (u_{ik}^{(l-1)})^m}, (i = 1, 2, \dots, n) \quad (3)$$

3) 更新隶属度矩阵 $\mathbf{U}^{(l)}$:

$$u_{ik}^{(l)} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}^{(l)}}{d_{jk}^{(l)}} \right)^{\frac{2}{m-1}}}, (i = 1, 2, \dots, c; k = 1, 2, \dots, n) \quad (4)$$

4) 计算目标函数 $J^{(l)}$,其中 $d_{ik}^{(l)} = \|x_k - v_i^{(l)}\|$:

$$J^{(l)}(\mathbf{U}^{(l)}, \mathbf{V}^{(l)}) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik}^{(l)})^m (d_{ik}^{(l)})^2 \quad (5)$$

5) 本文不设置停止精度,达到迭代上限,停止迭代,否则重复步骤2)。

6) 经过不断的迭代之后,可以求得最终的隶属度矩阵 \mathbf{U} 和聚类中心 \mathbf{V} ,使得目标函数 $J(\mathbf{U}, \mathbf{V})$ 达

到最小。根据最终的隶属度矩阵的取值可以确定所有样本的分类,当 $u_{jk} = \max_{1 \leq i \leq c} \{u_{ik}\}$ 时,可以将 x_k 归为第 k 类,以完成分类过程。

1.2 黏菌算法(SMA)

黏菌算法是一种新颖的元启发式群智能优化算法。该算法受启发于黏菌扩散和觅食行为进而获得连接食物的最佳路径。其捕食过程主要分为两个阶段。

在黏菌觅食活动的第一阶段,当黏菌搜索食物时,会根据空气中的气味来接近食物。在搜索第二阶段中,黏菌开始包围食物,模拟黏菌静脉结构中的收缩模式,并根据食物质量调整位置。依据黏菌的搜索行为,可以利用式(6)来模拟其觅食行为

$$\mathbf{S}(t+1) = \begin{cases} \mathbf{S}_b(t) + \mathbf{v}_b \times (\mathbf{W} \times \mathbf{S}_A(t) - \mathbf{S}_B(t)), & r < p \\ \mathbf{v}_c \times \mathbf{S}(t), & r \geq p \end{cases} \quad (6)$$

$$\mathbf{v}_b = [-a, a] \quad (7)$$

$$a = \arctan h \left(-\left(\frac{t}{t_{\max}} \right) + 1 \right) \quad (8)$$

式中: \mathbf{v}_c 从1线性减少到0; t 表示当前迭代; t_{\max} 为最大迭代数; \mathbf{X}_b 为当前发现气味浓度最高的位置; $\mathbf{S}(t)$ 为当前黏菌的位置; $\mathbf{S}(t+1)$ 为黏菌将要占据的下一个位置; \mathbf{S}_A 、 \mathbf{S}_B 代表从黏菌群体中随机选择的两个个体; r 为0和1间随机数; $\mathbf{v}_b \in (-a, a)$, \mathbf{W} 为权重:

$$W(\text{smellindex}(l)) = \begin{cases} 1 + r \times \log \left(\frac{bF - S(i)}{bF - \omega F} + 1 \right), & \text{condition} \\ 1 - r \times \log \left(\frac{bF - S(i)}{bF - \omega F} + 1 \right), & \text{其他} \end{cases} \quad (9)$$

$$\text{smellindex} = \text{sort}(\mathbf{S}) \quad (10)$$

当食物浓度较高时,区域权重较大;反之权重转移到其他区域。 bF 、 ωF 分别为当前迭代中获得的最佳、最差适应度值,condition为 $\mathbf{S}(i)$ 在整个种群中排名靠前的半部分。参数 p 的表达式如下:

$$p = \tanh(|f(i) - \text{DF}|) \quad (11)$$

式中: $i \in 1, 2, \dots, n$, $f(i)$ 是当前的位置适应度,DF是所有迭代中获得的最佳适应度值。

最后增补黏菌的全局搜索行为。此时对黏菌位置重新建模,得到完整的黏菌算法如下式所示。

$$\mathbf{S}^*(t+1) = \begin{cases} \text{rand}^*(\text{UB} - \text{LB}) + \text{LB}, & \text{rand} < z \\ \mathbf{S}_b(t) + \mathbf{v}_b^*(\mathbf{W}^* \mathbf{S}_A(t) - \mathbf{S}_B(t)), & r < p \\ \mathbf{v}_c^* \mathbf{S}(t), & r \geq p \end{cases} \quad (12)$$

式中:rand、 r 属于 $[0, 1]$,UB和LB是搜索空间的上限和下限, z 是黏菌将寻找其他食物来源或在当前最佳食物来源附近进行搜索的概率, \mathbf{W} 、 \mathbf{v}_b 、 \mathbf{v}_c 用于模拟黏菌静脉宽度变化。

1.3 青少年身份搜索算法 (AISA)

AISA 算法通过身份特征 (行为、喜好、思想、能力、信念等) 来描述青少年的身份, 本算法中青少年通过推理观察社会行为、榜样机制和不良特征选择这 3 种策略来形成自己的身份。

1) 推理观察社会行为: 青少年可以通过观察并推理同伴群体的行为, 来形成自己的身份。假定青少年可以通过识别组中的最佳特征并模仿他们, 其中青少年的最佳身份为局部最佳适应度。通过 Chebyshev 函数连接功能网络, 并通过最小二乘法估计实现局部优化, 所涉及切比雪夫多项式 $\{T_k(x)\}_{k=0,1,2,\dots}$ 递归方程如下:

$$k(x) := \begin{cases} 1, & k=0 \\ x, & k=1 \\ 2xT_{k-1}(x) - T_{k-2}(x), & k \geq 2 \end{cases} \quad (13)$$

其中 k 是 Chebyshev 多项式系数。归一化后的初始种群矩阵定义为 \hat{X} , 算法每个输入元素定义如式 (14):

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^N & x_2^N & \cdots & x_n^N \end{bmatrix}_{N \times n} \quad (14)$$

$$\phi = \begin{bmatrix} \phi_1^1 & \phi_2^1 & \cdots & \phi_n^1 \\ \phi_1^2 & \phi_2^2 & \cdots & \phi_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1^N & \phi_2^N & \cdots & \phi_n^N \end{bmatrix} \quad (15)$$

本文采用 $w^j \in R^{1 \times k}$ 为权重输入, 通过式 (16) 获得局部适应度值, ϕ 为次回归向量, 如式 (15) 中矩阵所示。

$$f_j^i = \phi_j^i w^j \quad (16)$$

当前群体的最佳身份向量由种群矩阵 \hat{X} 中 \hat{F} 对应矩阵中每一行最小索引量按照式 (17) 加和所得。此时局部优化后的群体最佳身份为 x^* 。

$$m^j = \arg \min \{f_j^i | i = 1, 2, \dots, N\} \quad (17)$$

$$x_j^* = x_j^{m^j} \quad (18)$$

所得出青少年的最佳身份 x_{new}^i 为

$$x_{\text{new}}^i = x^i - r_1 (x^i - x^*) \quad (19)$$

2) 榜样机制: 青少年通过模仿具有较高地位、权力和威望的榜样来形成自己身份, 榜样为具有最佳适应度值的个体。

$$x_{\text{new}}^i = x^i - r_2 (x^p - x^{\text{rm}}) \quad (20)$$

式中: r_2 是 $[0,1]$ 中的随机数; x^{rm} 表示榜样 (最佳个人)。 $p \neq \text{rm}$, x^p 表示 p 个青少年。

3) 不良特征选择: 青少年可能会具有吸烟、吸毒和霸凌等不良特征, 假定不良特征 x^q 是从总体矩阵中随机选择的元素, 青春期身份可以通过式 (21) 表示:

$$x_{\text{new}}^i = x^i - r_3 (x^i - x^q) \quad (21)$$

式中: r_3 为 $[0,1]$ 内均匀分布的 $1 \times n$ 维数字行向量; x^q 表示如下所示的负同一性向量。

$$x^q = [x^{q1} \ x^{q2} \ \cdots \ x^{qn}]_{1 \times n}^T \quad (22)$$

将 3 种结果组合, 得到新的青少年身份, 如式 (23) 所示, 其中 r_4 是 $[0,1]$ 内的随机数, 用于 3 种不同策略的选择。

$$x_{\text{new}}^i = \begin{cases} x^i - r_1 (x^i - x^*), & r_4 \leq 1/3 \\ x^i - r_2 (x^p - x^{\text{rm}}), & 1/3 < r_4 < 2/3 \\ x^i - r_3 (x^i - x^q), & 2/3 < r_4 \end{cases} \quad (23)$$

2 混合青少年搜索群体-黏菌算法

2.1 青少年-黏菌算法混合策略

黏菌算法是受到黏菌捕食中行为和形态的变化启发。在捕食的过程中, 根据食物量的多少, 黏菌的捕食行为将会产生正反馈和负反馈, 从而在黏菌算法中总结出了 3 种捕食形态。对比于其他群智能类算法, 如粒子群、灰狼、人工鱼群或蝗虫算法等依托于个体行为与群体行为相结合作为搜索方式的群智能优化算法, 黏菌算法更类似于鲸鱼算法、樽海鞘算法等在自身算法中凭借个体本身的多种探索、开发机制作为优化方式的算法。但这种以个体搜索作为优化策略核心的算法往往出现精度高的搜索特点, 但是同时每个个体搜索机制中的广域搜索和局部开发无法针对当前的迭代情况做到有效的平衡, 个体与群体的联系较弱, 无法有效地学习其他个体的经验, 导致每次迭代效果较差, 稳定性不足, 针对混峰和多峰问题性能较弱等情况。因此本文将侧重于将 SMA 算法与社会性质较强的 AISA 算法相结合, 替换黏菌算法的部分搜索机制, 使得 SMA 算法的每个个体具有社会属性。从社会群体的角度调整探索与开发的比重, 避免陷入局部最优, 改善优化求解精度, 增强稳定性。

在黏菌算法中, 黏菌个体通过式 (6)、(12) 来进行迭代运算, 其中式 (12) 是负责 SMA 算法的全局探索, 黏菌个体概率 z 进行全局内的搜索。但 SMA 算法本身缺乏有效的探索机制, 全局搜索是一种效率较低的搜索机制, 单纯的式 (12) 无法保证黏菌算法拥有足够大搜索比重。所以本文将 AISA 算法中青少年不良特征选择与 SMA 算法的全局域搜索相结合, 拓展搜索范围。更新后的公式为

$$\begin{cases} \text{rand}^* (\text{UB} - \text{LB}) + \text{LB}, & \text{rand} < R_1 \\ x^i - r_3 (x^i - x^q), & \text{rand} \geq R_1 \end{cases} \quad (24)$$

式中: x^q 为不良个体集合, x^q 表示负同一性向量, r_3 是一个 $1 \times n$ 区间中均匀分布的数字行向量 $[0,1]$ 。 R_1 为社会选择量, 具体参数因数据而定, 本文中

为 0.3。更新后的搜索公式将有较大的几率根据目前黏菌群体的位置确定与开发区域相反的探索度较低区域,使得其他个体执行进行广域搜索时可以提升搜索效率。

SMA 算法中式 (6) 主要模仿了黏菌个体通过静脉液宽进行包围食物。在此过程中黏菌个体将通过向食物中心 $S_b(t)$ 靠拢的方式进行包围。但 $S_A(t) - S_B(t)$ 使得黏菌个体在向最佳点靠拢时依旧具有局部探索的能力。这个策略本身十分优秀,但是在 SMA-AISA 算法的全局搜索部分已经加强了其搜索能力,则此部分算法中,将引入社会属性,以获得更为激进的开发策略以保证开发与搜索的平衡。将 AISA 算法中的青少年推理观察社会行为与黏菌的食物包围机制相结合,通过使用式 (19) 替换包围机制中的食物最佳点 $S_b(t)$,使得包围机制可以向整个群体中局部最佳的位置进行包围,提升探索效率。

$$\begin{cases} S_b(t) + v_b^*(W^*S_A(t) - S_B(t)), & \text{rand} < R_2 \\ x^* + v_b^*(W^*S_A(t) - S_B(t)), & \text{rand} \geq R_2 \end{cases} \quad (25)$$

其中 x^* 的计算式 (19) 与上文中 AISA 的青少年群体观察机制相同,黏菌通过观察群体行为,确定群体特征局部最优情况进行包围。为防止过拟合的发生,设置 R_2 为群体学习概率,群体学习概率决定策略烈度,具体参数因数据而定,本文设定为 0.5。

SMA 算法中式 (6) 主要模仿了黏菌个体对于食物的捕食机制,此公式是 SMA 算法主要的开发策略。但是在针对数据维度高且复杂的情况时,黏菌的开发能力不是十分理想。为了保证本文的收缩曲线能够保持平滑和快速,将 AISA 算法中的榜样学习机制作为增补机制,防止算法在前期过度收缩。具体计算公式如下:

$$\begin{cases} v_c^*S(t), R_3 > 0.5 \\ x^i - \partial(x^p - x^m), R_3 \leq 0.5 \end{cases} \quad (26)$$

此更新公式中, ∂ 是青少年学习能力因子,公式为式 (11)。学习因子保证了增补的公式 (11) 具有前期学习能力强,后期能力弱的特点,保证收敛速度。 $\partial(x^p - x^m)$ 调整收敛的步长,防止个体粒子在前期过度收敛陷入局部最佳,故设置 R_3 为榜样学习意愿,调整开发策略,具体参数因数据决定,本文设定为从 1 线性减少到 0。综合 3 种更新公式,得到新的启发式混合群智能优化算法即青少年身份黏菌算法,实际运行过程如下所示。首先随机初始化种群数值,产生隶属度矩阵,确认参数,随后将进入 While 循环,计算每个个体的适应度;并对个体进行操作,通过个体适应度计算其本身的黏菌权重如式 (9) 所示;随后进入两个判断过程判断几率是否大于 z ;判断随机数是否

大于 R_1 , 通过式 (24) 进行全球域搜索。最后若式 (9) 中几率小于 z ;判断随机数是否大于 R_2 , 若大于则通过式 (25) 中第一策略对目前最优位置进行包围和开发,反之则通过式 (25) 中第二策略通过青少年身份形成机制,优化收敛位置,进行包围和开发。完成前两个步骤后,个体通过式 (26) 进行捕食,通过剧烈收缩得到最佳适应值。完成循环,若迭代次数不满足最大迭代次数。则跳转回到 While 循环。

2.2 青少年-黏菌混合策略可行性分析

对 AISA 算法的搜索策略进行分析,可以发现其推理观察社会行为机制如式 (19),是具有一定的创新性的。在本策略中通过 Chebyshev 函数连接功能网络和最小二乘法估计实现局部优化。式 (17) 和式 (18) 使得算法可以根据当前群体适应度及特征,对最优点进行预估计,使得搜索个体在迭代的前期可以快速的向最优方向包围和开发。这个策略相较于 SMA 算法所使用的策略,即利用目前个体最优解作为当前循环的收敛中心,在收敛速度方面具有很大优势。所以本文在式 (25) 基础上引入式 (19) 的局部预测机制,以改善其收敛速度较慢的弱点。

从 SMA 算法搜索策略进行分析,式 (6) 作为 SMA 算法核心公式,式 (9) 独特的收缩包围策略,使得 SMA 算法在局部的开发与探索通过权重的调整达到了很好的平衡,这使得该算法在处理数据结构相对简单的局部寻优问题时有极其优异的性能。但也是因为这个机制,所以算法在迭代的过程中,局部的开发与探索占用了大量的计算资源,所以在面对数据结构复杂,维度较高的全局寻优问题时,无论精度还是收敛速度都显示出了疲态。而 AISA 算法恰巧与其相反,得益于群体信息共享机制、相对高效的全局搜索机制,可使得 AISA 算法在面对数据结构复杂,维度较高的全局寻优问题时可以获得良好的效果,这正是 SMA 算法所缺少的部分。所以本文在式 (24) 中引入了不良特征选择策略,并且在式 (26) 中引入榜样策略。这样的改进目的是提升原算法的稳定性,及高纬度、复杂数据的处理能力。

3 AISA-SMA 优化 FCM 聚类问题

FCM 算法受模糊理论的影响,相较于 C-均值 (K-means) 聚类为代表的硬子空间集聚类算法提供了更为灵活的聚类思路。但传统的 FCM 算法在聚类的迭代时通常是一个中心连续移动的过程,这种聚类的方式无法兼顾全局的搜索,算法对初始化依赖程度较大并且容易陷入局部最佳;同时在高维度的聚类过程中,原算法通常无法有效地进行聚类分析。针对以上缺点,本文利用

AISA-SMA 算法精度高、收敛快和稳定性强的特点来优化原始 FCM 算法, 通过调整迭代过程, 即将 AISA-SMA 算法的搜索和开发特性引入 FCM 的每一次迭代过程, 改善算法性能, 这样既引入了全新的搜索模式也保留模糊 C 均值算法的模糊数学思路。提出融合算法 AISA-SMA-FCM, 意在通过 AISA-SMA 算法较好的开发、探索能力和较强的收敛策略, 提升 FCM 算法的单次循环聚类效果, 以及尽量减少参数优化所增加的时间复杂度带来的影响。使得原 FCM 算法聚类精度、稳定性及分类准确率得到提升。

3.1 混合 AISA-SMA 算法实现聚类功能

首先, AISA-SMA 算法在初始化处理聚类问题时与其他聚类算法类似, 首先在算法的初始阶段读入数据矩阵 \hat{D} :

$$D = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^N \\ x_2^1 & x_2^2 & \cdots & x_2^N \\ \vdots & \vdots & \ddots & \vdots \\ x_a^1 & x_a^2 & \cdots & x_a^N \end{bmatrix} \quad (27)$$

所得矩阵为 $N \times a$, 其中 N 为数据特征数, a 为数据总量。此时需要定义 AISA-SMA 算法迭代过程中的个体, 赋予每个个体初值 \hat{G} 。这一步与其他启发式群智能优化算法中的初始化相同。

$$G = \left(\left[\begin{array}{cccc} d_1^1 & d_1^2 & \cdots & d_1^N \\ \vdots & \vdots & \ddots & \vdots \\ d_k^1 & d_k^2 & \cdots & d_k^N \end{array} \right]^1 \cdots \left[\begin{array}{cccc} d_1^1 & d_1^2 & \cdots & d_1^N \\ \vdots & \vdots & \ddots & \vdots \\ d_k^1 & d_k^2 & \cdots & d_k^N \end{array} \right]^n \right)^T \quad (28)$$

\hat{G} 为一个 $N \times K \times n$ 的矩阵, 其中 K 为需要聚类

的中心数, n 为 AISA-SMA 算法中总共设定的个体数。每一个个体都为 K 个聚类中心, 每个中心在初始化阶段由式 (29) 随机产生。

$$d = \text{rand}(K, N) \times (UB - LB) + LB \quad (29)$$

其次, 所得到的初始化个体通过 ASIA-SMA 算法中数据更新式 (24)~(26) 迭代得到全新的聚类中心点位置, 此过程可以视为 N 维空间中 $K \times n$ 个搜索粒子围绕 K 个聚类中心进行的探索和开发。此时聚类中心点的适应度值应为 K 个聚类中心一组, 由式 (30) 计算:

$$F = \sum_{k=1}^K \sum_{i=1}^N \|x_i - u_k\|^2 \quad (30)$$

适应度越高则代表此时聚类结果的组内距最小, 组间间距最大, 即聚类结果的种群内部相似度更高, 聚类效果越好。

具体 AISA-SMA 算法处理 FCM 聚类问题流程图如图 1 所示, 其中 AISA-SMA 算法中的个体可以看做搜索粒子, 通过式 (24)~(26) 对聚类中心的分布情况进行迭代优化, 通过式 (5) 计算每个点的适应度值, 找到适应度值最好的聚类中心分布, 将聚类中心作为迭代结果输出, 完成数据分类任务。AISA-SMA 算法具体的思路如下所示, 首先读取所聚类的数据, 设置聚类数目, 通过聚类数目, 以聚类中心为个体进行 AISA-SMA 的初始化及参数设定。随后通过式 (30) 计算每个个体的适应度。通过式 (9) 对每个个体计算权重, 并判断随机数是否大于 z , 判断是否进行全球化搜。最后通过 R_1, R_2, R_3 决定式 (24)~(26) 的策略选择。

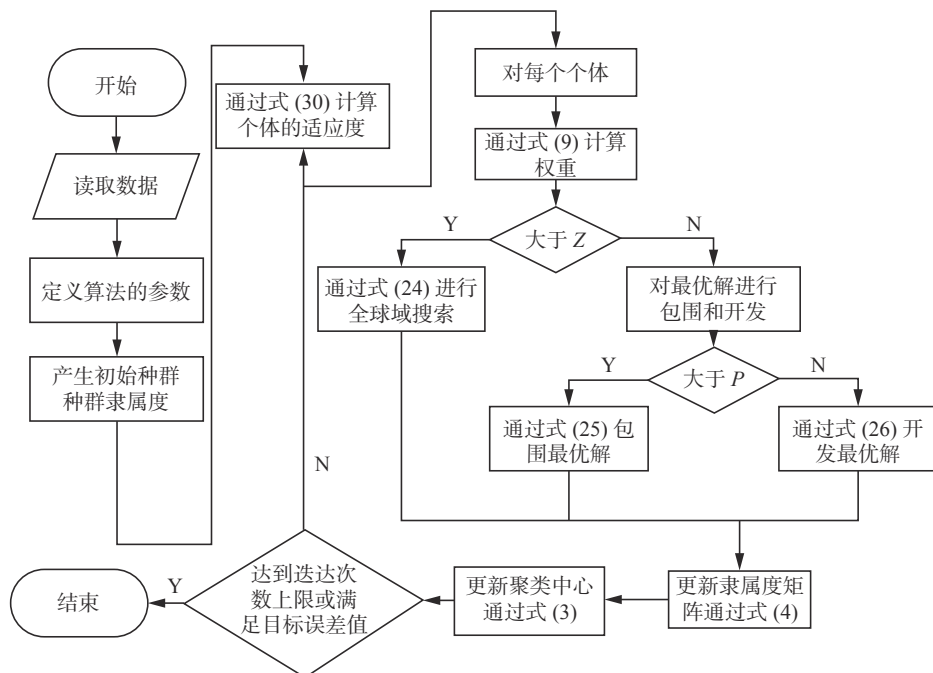


图 1 启发式智能优化用于 FCM 聚类

Fig. 1 Heuristic intelligent optimization for FCM clustering

3.2 混合 AISA-SMA 算法实现聚类功能

所提 AISA-SMA-FCM 聚类算法通过增补 AISA-SMA 聚类策略, 替换部分 FCM 算法中随机属性高的策略。个体模仿 AISA-SMA 算法中的包围捕食猎物的过程, 探索开发聚类中心分布, 对聚类中心的适应度进行迭代优化, 具体实现过程如下:

算法初始阶段, 随机生成隶属度矩阵 B , 每个隶属度矩阵应为 $a \times K \times n$, 隶属的矩阵中, n 为 n 个社会黏菌搜索粒子, 每个搜索粒子所属的隶属度矩阵代表数据 D 中每个数据对于聚类中心的隶属度。

$$B = \left(\left[\begin{array}{cccc} b_1^1 & b_1^2 & \cdots & b_1^k \\ \vdots & \vdots & \ddots & \vdots \\ b_a^1 & b_a^2 & \cdots & b_a^k \end{array} \right]^1 \cdots \left[\begin{array}{cccc} b_1^1 & b_1^2 & \cdots & b_1^k \\ \vdots & \vdots & \ddots & \vdots \\ b_a^1 & b_a^2 & \cdots & b_a^k \end{array} \right]^n \right)^T \quad (31)$$

获得隶属的矩阵 B 后, 通过式 (29) 生成聚类中心 G , 此时算法获得了与搜索粒子数相同的聚类中心数。聚类中心将作为 AISA-SMA 聚类算法中的初值进行迭代运算, 通过式 (24)、(25)、(26) 将获得新的聚类中心 G , 聚类中心 G 通过 FCM 算法中的式 (4) 更新隶属度矩阵 B , 通过新的隶属计算每个粒子的归属情况, 某个中心的隶属度最高, 则该粒子隶属于某个聚类中心, 随后通过式 (3) 再次更新聚类中心 G 完成一个完整周期。随着算法迭代, AISA-SMA 算法将快速收敛, 搜索粒子所代表的聚类中心将通过优化策略, 寻找适应度最

高区域, 并逐渐稳定并影响后续过程发散程度。此过程中的式 (25)、(26) 主要为整个算法提供局部的收敛精度, 并且迭代公式 (24)、(25) 的搜索能力, 将避免止算法陷入局部最优。

4 仿真实验与结果分析

为全面验证改进算法的有效性, 本节共设计两部分对比实验, 一是选择目前流行的启发式智能优化算法进行标准测试函数实验, 测试算法理论寻优性能; 二是利用经典 UCI 数据集测试对比其他算法用于聚类分析的优化效果。仿真实验运行环境为 Windows 7 系统下使用 MATLAB2016a, CPU: i5-6500@3.3GHz, RAM: 4GB。

4.1 标准测试函数寻优对比实验

本节选择 AISA-SMA、SMA^[12]、AISA^[17]、MRFO^[18]、GWO^[11]、PSO^[19] 算法在 15 个标准测试函数^[3] 上进行对比测试分析, 其中测试集 F1~F7 为单峰函数, F8~F11 为多峰函数, F12~F15 为动态多峰函数。实验测试评价基于运行 10 次搜索最优解的平均值。种群搜索个体设定为 30, 最大迭代次数为 500, 函数维度为 30。

AISA-SMA 与其他 4 种优化算法在 15 个标准测试函数的寻优实验统计结果如表 1 所示。收敛曲线结果如图 2 所示。

表 1 AISA-SMA 与其他优化的测试函数对比结果

Table 1 Comparison results between AISA-SMA and other optimized test functions

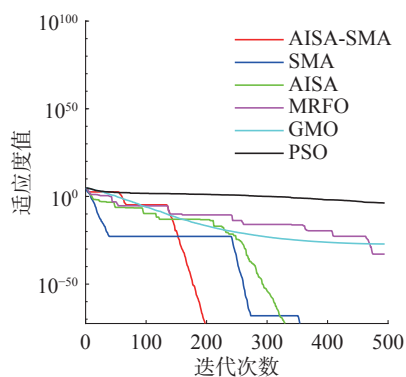
| 函数 | MRFO | GWO | PSO | SMA | AISA | AISA-SMA |
|----|------------------------|-----------------------|-----------------------|-------------------------|-------------------------|-------------------------|
| 1 | 2.78×10^{-18} | 1.43×10^{-5} | 1.15×10^3 | 1.42×10^{-143} | 4.56×10^{-106} | 4.21×10^{-296} |
| 2 | 1.20×10^{-9} | 5.34×10^{-4} | 2.15×10^{54} | 3.45×10^{-80} | 1.59×10^{-56} | 4.06×10^{-197} |
| 3 | 3.96×10^{-17} | 8.77×10^4 | 2.18×10^5 | 1.39×10^{-32} | 2.68×10^{-78} | 5.45×10^{-293} |
| 4 | 1.14×10^{-10} | 4.48×10^1 | 2.33×10^1 | 4.28×10^{-55} | 2.39×10^{-51} | 6.32×10^{-161} |
| 5 | 2.99×10^2 | 2.98×10^2 | 3.61×10^6 | 2.52×10^{-2} | 4.54×10^{-2} | 2.72×10^{-3} |
| 6 | 6.45×10^1 | 4.86×10^1 | 1.17×10^3 | 4.66×10^{-1} | 2.11×10^{-4} | 2.32×10^{-3} |
| 7 | 6.11×10^{-1} | 2.55×10^{-2} | 1.98×10^4 | 1.78×10^{-4} | 1.26×10^{-4} | 6.96×10^{-4} |
| 8 | -9.54×10^3 | -4.08×10^4 | -1.77×10^4 | -1.25×10^5 | -1.25×10^5 | -1.26×10^5 |
| 9 | 0.00×10^0 | 3.02×10^1 | 3.31×10^3 | 0.00×10^0 | 0.00×10^0 | 0.00×10^0 |
| 10 | 4.58×10^{-10} | 2.04×10^{-4} | 8.59×10^0 | 8.88×10^{-16} | 8.88×10^{-16} | 8.88×10^{-16} |
| 11 | 8.05×10^{-2} | 1.16×10^{-1} | 1.51×10^{-6} | 2.99×10^{-3} | 9.59×10^{-6} | 3.54×10^{-6} |
| 12 | 9.98×10^{-1} | 3.64×10^0 | 2.09×10^0 | 9.98×10^{-1} | 9.86×10^{-1} | 9.98×10^{-1} |
| 13 | 1.09×10^{-3} | 4.83×10^{-4} | 3.60×10^{-4} | 6.05×10^{-4} | 3.31×10^{-4} | 8.46×10^{-4} |
| 14 | -6.66×10^{-1} | -3.86×10^0 | -3.86×10^0 | -3.86×10^0 | -3.86×10^0 | -3.86×10^0 |
| 15 | -5.62×10^{-1} | -1.05×10^1 | -9.33×10^0 | -1.05×10^1 | -6.12×10^0 | -1.05×10^1 |

数据表明, 从均值上看, 原 SMA 算法测试函数搜索最优解的表现良好, 与本文算法具有一定的竞争力, 但本文所提出的混合 AISA-SMA 在多

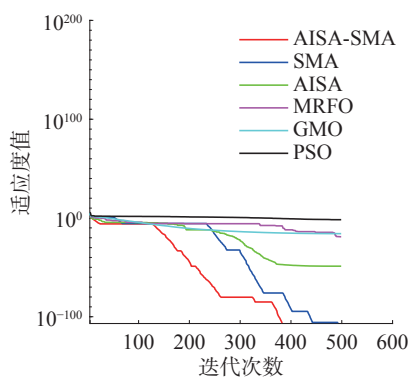
数的单峰值与多峰值函数测试中可以搜寻到理论最优值, 在动态多峰实验中也基本可以搜寻到理论最佳值, 性能相比 SMA 算法有了一定提升。

AISA-SMA 算法相比原 AISA 算法, 在单峰的数据开发中具有很大优势, 在多峰和混峰问题中也良好地继承了其良好的开发特性, 很好地弥补了原 SMA 算法在面对复杂多峰问题开发能力不足的缺点。结合收敛曲线分析, 结合后的 AISA-SMA

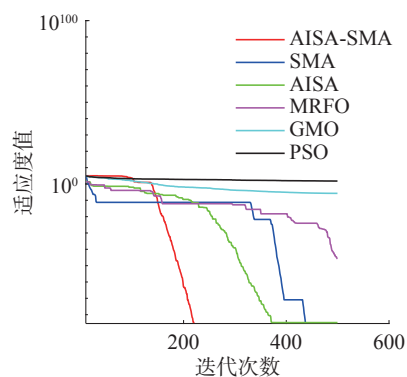
算法也继承了 AISA 算法在前期激进的收敛策略, 使得 AISA-SMA 算法收敛速度进一步提升, 通过节约迭代次数的策略, 相较于 SMA 算法提升计算效率。这进一步说明本文所提出的混合算法在空间搜索与计算求解中较其他算法具有明显的优势。



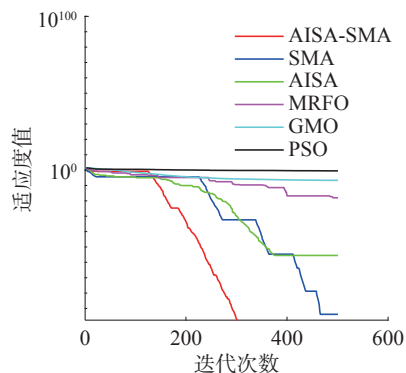
(a) F1



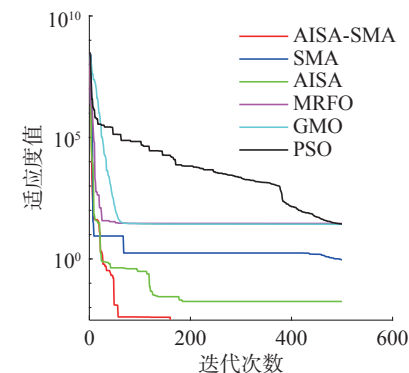
(b) F2



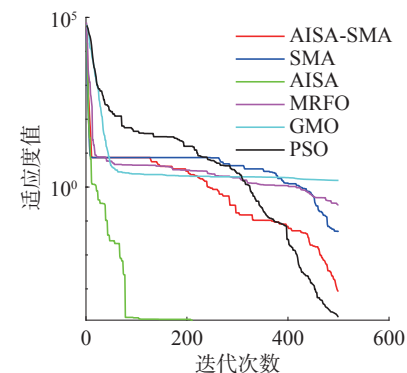
(c) F3



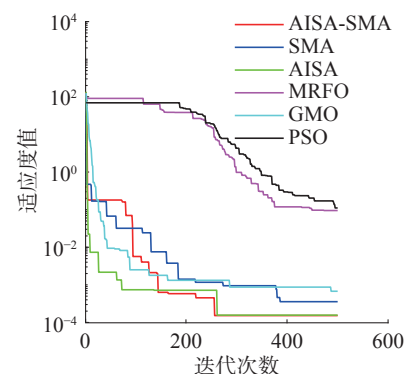
(d) F4



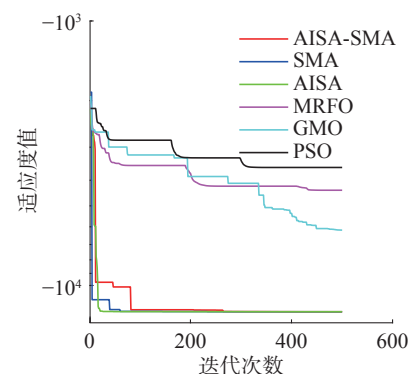
(e) F5



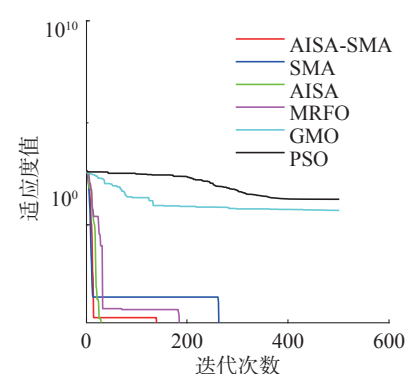
(f) F6



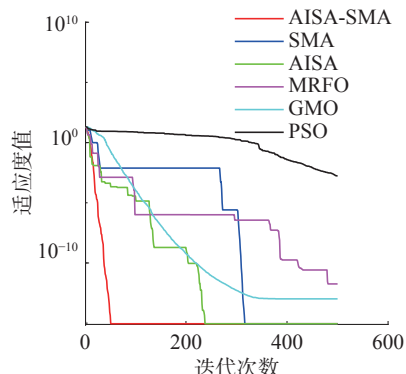
(g) F7



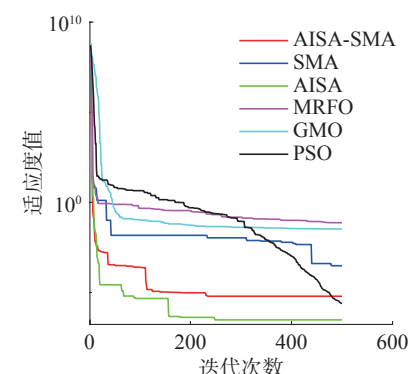
(h) F8



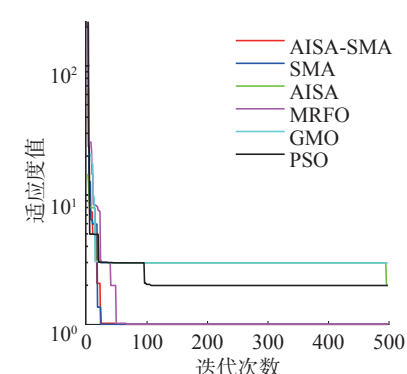
(i) F9



(j) F10



(k) F12



(l) F14

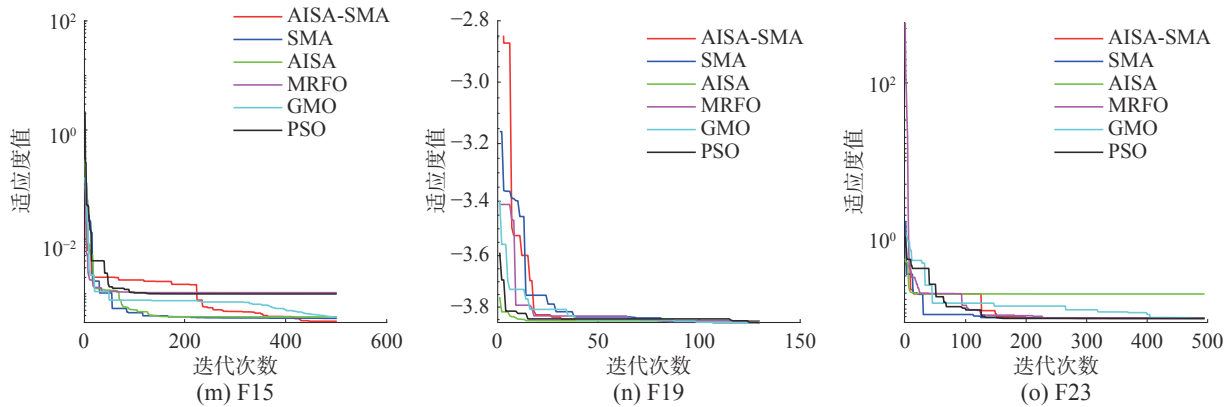


图 2 本文算法与其他算法适应度函数曲线

Fig. 2 The fitness function curves of the proposed and other algorithms

4.2 聚类分析对比实验

为进一步验证本文所提出混合优化用于 FCM 聚类, 即 AISA-SMA-FCM 算法的有效性与优越性, 选择 UCI 数据库^[20]中的 6 个标准数据集进行聚类仿真实验, 具体数据集说明列于表 2 中。种群设置为 50, 最大迭代次数为 200, 运行环境与上节实验相同。对比验证的聚类方法采取 SMA 优化 K-means 算法与 SMA 优化 FCM 算法、FCM 算法、以及 MRFO 优化 FCM 算法。实验中每个算法独立运行 15 次, 统计计算适应度数据的平均值作为评价指标 1, 平均值越小表明算法聚类效果好, 即聚类误差值小; 将 15 次最优平均适应度值的方差结果采取 ANOVA 箱型图形式进行分析, 作为评价指标 2。每个图中, 箱图的高度代表算法 15 次运算的数据分散程度, 越短的箱图表明 10 次结果越集中, 同时也证明箱图的噪声越小, 离群值较少, 即算法稳定性越高; 同时将适应度函数收敛曲线作为评价算法聚类效果的指标 3, 平滑、下降迅速, 无较长延迟的曲线证明算法拥有更好的优化聚类性能; 为了进一步体现算法的性能差异, 将通过算法 15 次独立运行的结果, 进行非参数统计显著性测试 (Wilcoxon 秩和检验), 以及进行参数双向方差分析的非参数模拟弗里德曼检验得出统计 p 值作为评价指标 4。将

算法 15 次独立运行聚类结果的准确率, 进行统计作为评价指标 5。最后对比 AISA-SMA-FCM 及 FCM 算法过程散点图对算法收敛速度进行评价为指标 6, 综合指标 1~6 对算法进行综合评价。

表 2 UCI 数据集的详细信息
Table 2 Details of UCI data sets

| 数据 | 特征 | 样本数 | 分类 | 编号 |
|-------------------------|-----|-----|----|----|
| Iris | 3 | 150 | 3 | 1 |
| Breast cancer wisconsin | 9 | 683 | 2 | 2 |
| Urban Land Cover | 147 | 507 | 9 | 3 |
| Haberman's survival | 3 | 306 | 2 | 4 |
| ART1 | 2 | 600 | 4 | 5 |
| ART2 | 3 | 250 | 5 | 6 |

基于 AISA-SMA-FCM 算法与其他对比算法的聚类所得适应度值如表 3 所示。表中数据结果表明, AISA-SMA-FCM 在平均适应度值中表现最优, 相较于其他经典算法更为突出, 得益于 FCM 算法主体思路的优势与 AISA-SMA 算法优越性, 表中用于优化 FCM 的算法都较优化 K-means 算法表现出更好的性能。因此, 本文所提出的混合 AISA-SMA 算法更适合 FCM 聚类分析的优化处理。

表 3 AISA-SMA 其他优化算法用于聚类的适应度值

Table 3 The fitness value of AISA-SMA and other optimization algorithms for clustering

| 算法 | Iris | Breast CW | UrbanLC | HS | ART1 | ART2 |
|------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| SMA-Kmeans | 2.2×10^2 | 2.46×10^4 | 2.31×10^4 | 2.73×10^4 | 2.84×10^3 | 1.22×10^5 |
| MRFO | 8.58×10^1 | 1.60×10^4 | 2.31×10^4 | 2.36×10^4 | 1.30×10^3 | 6.12×10^4 |
| FCM | 7.99×10^1 | 1.61×10^4 | 2.31×10^4 | 2.31×10^4 | 1.28×10^3 | 5.38×10^4 |
| SMA | 7.91×10^1 | 1.56×10^4 | 2.31×10^4 | 2.20×10^4 | 1.25×10^3 | 5.35×10^4 |
| AISA-SMA | 7.84×10^1 | 1.55×10^4 | 2.31×10^4 | 1.52×10^4 | 1.23×10^3 | 5.34×10^4 |

基于 AISA-SMA-FCM 算法与其他聚类方法的 ANOVA 箱型图的测试对比结果如图 3 所示。在 6 个数据集中, 本文所提出的混合算法用于聚类时的计算稳定性最好, 虽在 Haberman's Survival

数据集中 SMA 的稳定性最佳, 但其适应度值不如本文方法优秀, 因此, 综合其他聚类来看, 本文提出的方法用于 FCM 聚类分析问题能够获得良好的聚类稳定性。

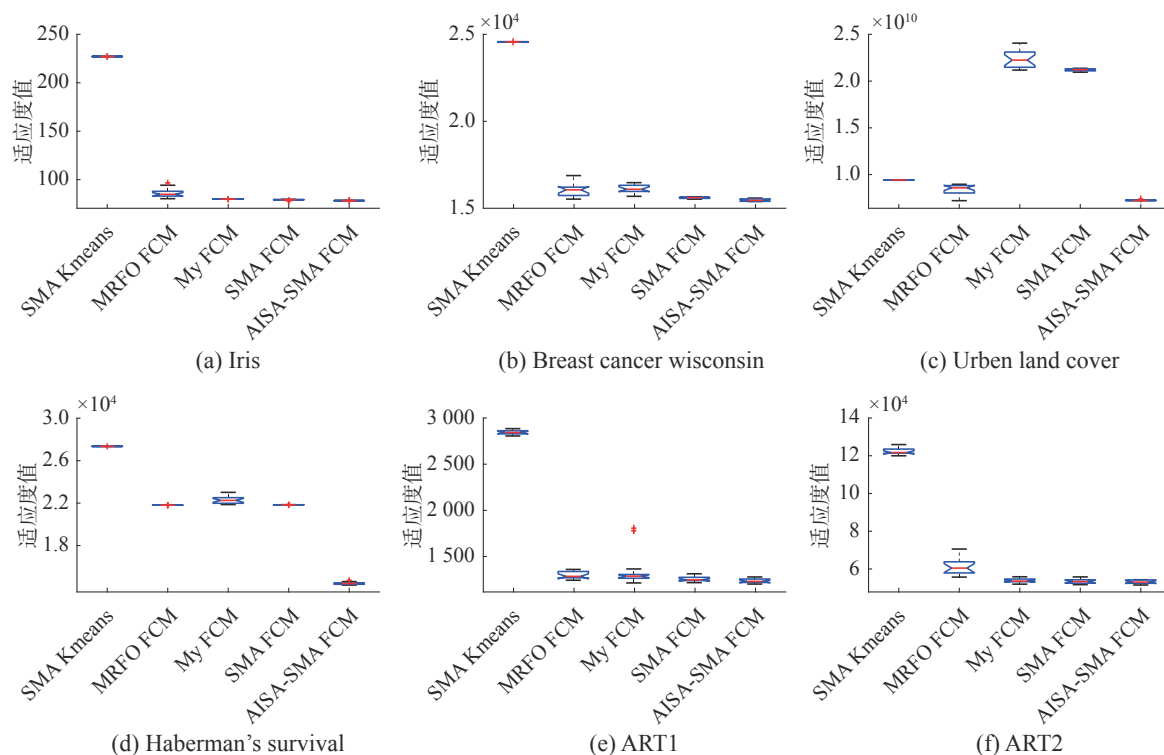


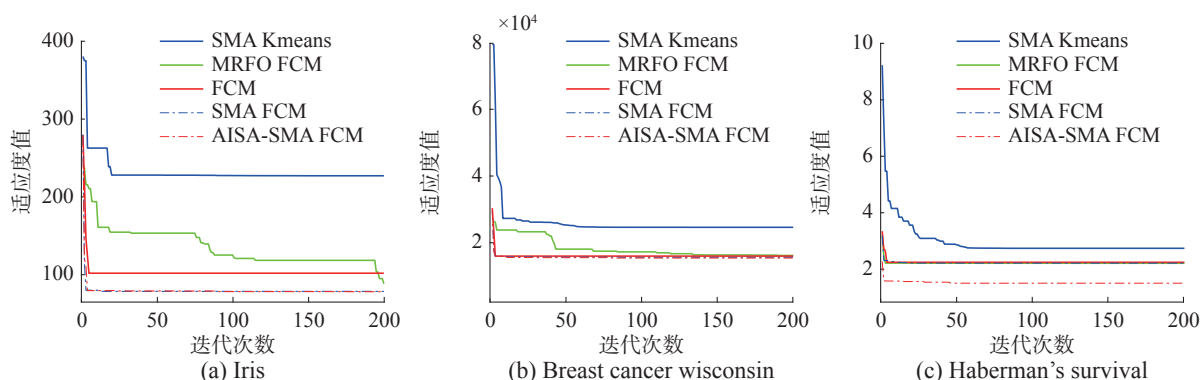
图 3 本文算法与其他算法适应度函数的 ANOVA 图

Fig. 3 The ANOVA graphs of the fitness function of the proposed and other algorithms

基于 AISA-SMA-FCM 算法与其他聚类方法的适应度函数值收敛曲线如图 4 所示。在数据集 2 与 3 中, 本文所提算法收敛曲线效果最佳, 用于解决 FCM 的聚类问题时能够达到快速、准确收敛, 同时搜寻到最佳的适应度值, 数据集 1 与 3 中的结果表明, 与经典的 K-means 聚类效果相比, FCM 聚类技术更为优秀, 本文算法用于 FCM 聚类优化效果显著, FCM 的搜索计算能力强。进一步验证本文混合算法优化。

表 4 中给出了 Wilcoxon 秩和检验产生的 p 值, 该值用于比较的两个值为连续分布的样本且具有相

等的中间值。表中的本组实验数据几乎所有的值都低于 0.05, 因 5% 为显著性水平值, 零假设不成立, 所以数据表明 AISA-SMA-FCM 算法在统计上是显著的, 实验结果具有真实可信性。为进一步证明本文算法的优越性, 验证两种算法的差异性, 进行参数双向方差分析的非参数模拟弗里德曼检验, 所得结果如表 5 所示。AISA-SMA-FCM 算法秩结果为 1.9, 表明其具有最良好的性能。综上所述, 实验结果证明 AISA-SMA 算法在针对 FCM 聚类问题时, 能够提升 FCM 算法的性能, 并且 AISA-SMA-FCM 算法相较于其他算法在聚类优化过程中效果显著。



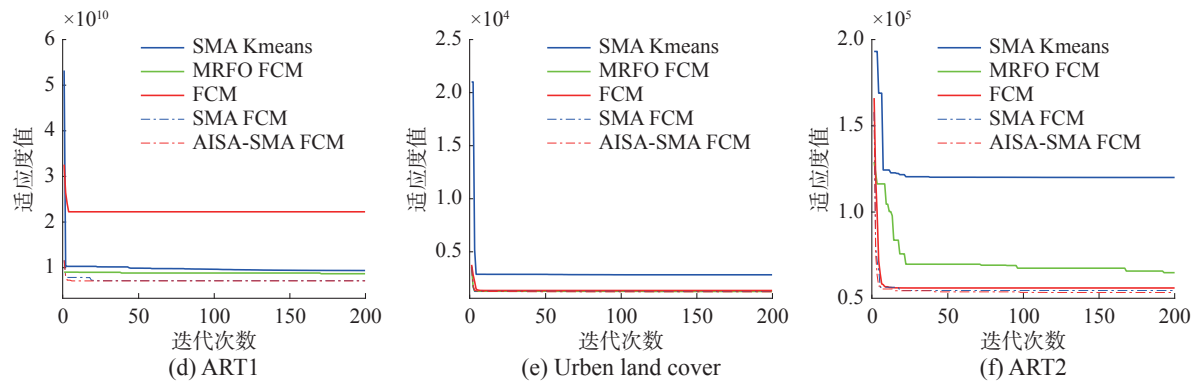


图4 本文算法与其他算法适应度函数曲线

Fig. 4 The fitness function curves of the proposed and other algorithms

表4 本文算法与其他优化 Wilcoxon 检验 p 值Table 4 The Wilcoxon rank sum test p -value of the proposed and the other methods

| 数据 | SMA-Kmeans | MRFO | FCM | SMA | AISA-SMA |
|----|-----------------------|-----------------------|-----------------------|-----------------------|-------------------|
| 1 | 6.80×10^{-8} | 6.81×10^{-8} | 6.33×10^{-8} | 9.73×10^{-6} | 1.0×10^0 |
| 2 | 6.77×10^{-8} | 6.77×10^{-8} | 6.18×10^{-8} | 1.04×10^{-6} | 1.0×10^0 |
| 3 | 8.01×10^{-9} | 5.74×10^{-1} | 8.01×10^{-9} | 4.57×10^{-9} | 1.0×10^0 |
| 4 | 6.80×10^{-8} | 8.60×10^{-6} | 6.80×10^{-8} | 6.80×10^{-8} | 1.0×10^0 |
| 5 | 3.37×10^{-6} | 5.72×10^{-5} | 4.21×10^{-4} | 2.25×10^{-2} | 1.0×10^0 |
| 6 | 3.39×10^{-6} | 3.39×10^{-6} | 7.58×10^{-4} | 6.19×10^{-2} | 1.0×10^0 |

表5 本文算法与其他优化的弗里德曼检验

Table 5 The Friedman tests of the proposed and others

| 算法 | SMA-Kmeans | MRFO | FCM | SMA | AISA-SMA |
|----|------------|------|-----|-----|----------|
| 评级 | 6.8 | 4.6 | 4.4 | 2.3 | 1.9 |

4.3 聚类结果对比实验

4.3.1 聚类准确率对比试验

本节将进一步对比 AISA-SMA-FCM 算法真实聚类效果。为了直观对比 5 种算法的聚类效果,统计 6 个数据集的正确率,每个算法运行 15

次,取平均正确率,详细结果如表 6 所示。由该表统计结果可以看出,在本文的 6 个数据集中,相较于其他算法, AISA-SMA 算法分类准确率较高。AISA-SMA 算法对于 UrbanLC、ART1、ART2 数据集,相较于基础 FCM 及 SMA 算法提升明显。但对于聚类中心数较高的情况例如 UrbanLC,本算法依旧无法进行有效的分类,同时对于聚类趋势不明显的数据集如 HS,本文算法也无法做到有效分类。综上所述, AISA-SMA 针对多中心、混杂数据集,依然还具有一定的提升空间。

表6 本文算法与其他算法的平均准确率

Table 6 Average correct rate of the proposed and others

| 算法 | Iris | BreastCW | UrbanLC | HS | ART1 | ART2 |
|------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SMA-Kmeans | 64.8 | 67.9 | 36.8 | 63.9 | 70.7 | 66.5 |
| MRFO | 85.5 | 95.4 | 43.0 | 52.5 | 93.1 | 92.3 |
| FCM | 84.4 | 94.7 | 51.8 | 52.6 | 64.8 | 85.3 |
| SMA | 87.0 | 95.2 | 48.8 | 53.2 | 96.8 | 94.3 |
| AISA-SMA | <u>88.9</u> | <u>95.7</u> | <u>55.4</u> | <u>57.3</u> | <u>97.2</u> | <u>95.3</u> |

4.3.2 聚类进程对比试验

本节将对聚类过程进行探讨,选择 Iris、HS、ART1、ART2 4 个数据集,统计 FCM 及 AISA-SMA 算法在 5 次及 20 次迭代时的聚类效果,具体结果如图 5 所示。从图中 ART1、ART2 及 HS 数据集

可知,在 5 次迭代时 AISA-SMA 算法已经具有基本分类模型,相较于 FCM 收敛较快;分析 Iris 数据集可知,两聚类算法在 20 次时都已经收敛,但分类效果 AISA-SMA 算法明显优于 FCM 算法,聚类群落清晰。因此通过测试结果可知, AISA-SMA

算法相较于 FCM 算法具有响应速度快, 分类效果好的优点, 在相同的迭代次数下, 聚类群体形态明显好于原生 FCM 算法, 其中的 ART2 及 Iris 数

据集效果更为突出。综上所述, AISA-SMA 算法在聚类准确度、收敛速度及稳定性方面具有一定提升, 效果较好。

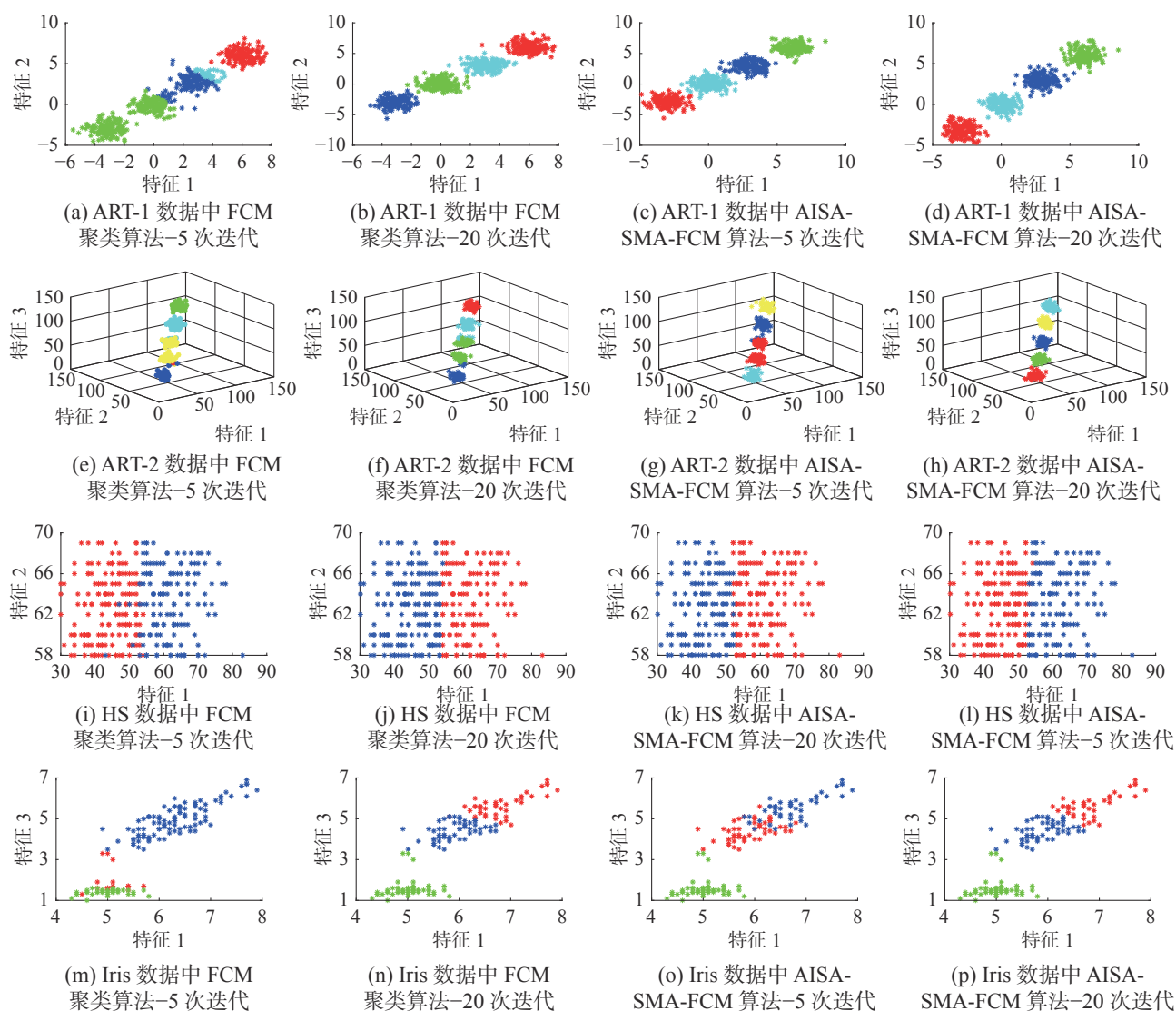


图 5 本文算法与其他算法的聚类过程

Fig. 5 The Clustering process of the proposed and others

5 结论

本文提出了一种混合 SMA 与 AISA 算法来优化模糊 C-均值的聚类方法, 算法首先将 SMA 与 AISA 进行混合优化, 利用 AISA 算法中的青少年社会机制, 改善 SMA 算法中的全局搜索和局部开发性能, 测试函数实验表明混合算法计算求解性能优越; 对于原 FCM 算法采取 AISA-SMA 算法对其迭代优化, 提出 AISA-SMA-FCM 聚类算法以改善原聚类方法的聚类效率及精度, 仿真结果证明 AISA-SMA-FCM 聚类算法在计算稳定性和收敛精度方面均具有较大的提升。如何提升算法在更高维度数据聚类分析中仍具有稳定、准确的效果是未来的研究方向, 同时可探索将混合算法

AISA-SMA 运用到更多工程优化计算中。

参考文献:

- [1] 贾鹤鸣, 李瑶, 孙康健. 基于遗传乌燕鸥算法的同步优化特征选择.[J]. 自动化学报, 2022, 48(6): 1601-1615.
JIA Heming, LI Yao, SUN Kangjian. Simultaneous feature selection optimization based on hybrid sooty tern optimization algorithm and genetic algorithm[J]. Acta automatica sinica, 2022, 48(6): 1601-1615.
- [2] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述 [J]. 模式识别与人工智能, 2014, 27(4): 327-336.
HE Qing, LI Ning, LUO Wenjuan, et al. A survey of machine learning algorithms for big data[J]. Pattern recognition and artificial intelligence, 2014, 27(4): 327-336.
- [3] 贾鹤鸣, 姜子超, 李瑶. 基于改进秃鹰搜索算法的同步

- 优化特征选择[J]. 控制与决策, 2022, 37(2): 445–454.
- JIA Heming, JIANG Zichao, LI Yao. Simultaneous feature selection optimization based on improved bald eagle search algorithm[J]. Control and decision, 2022, 37(2): 445–454.
- [4] 汤正华. 基于改进蝙蝠优化自确定的模糊 C-均值聚类算法[J]. 计量学报, 2020, 41(4): 505–512.
- TANG Zhenghua. Self-defined fuzzy clustering C-means algorithm based on improved bat optimization[J]. Acta metrologica sinica, 2020, 41(4): 505–512.
- [5] 董发志, 丁洪伟, 杨志军, 等. 基于遗传算法和模糊 C 均值聚类的 WSN 分簇路由算法[J]. 计算机应用, 2019, 39(8): 2359–2365.
- DONG Fazhi, DING Hongwei, YANG Zhijun, et al. WSN clustering routing algorithm based on genetic algorithm and fuzzy C-means clustering[J]. Journal of computer applications, 2019, 39(8): 2359–2365.
- [6] DUNN J C. A graph theoretic analysis of pattern classification via tamura's fuzzy relation[J]. IEEE transactions on systems, man, and cybernetics, 1974, SMC-4(3): 310–313.
- [7] POLI R, KENNEDY J, BLACKWELL T. Particle swarm optimization[J]. Swarm intelligence, 2007, 1(1): 33–57.
- [8] NIKNAM T, OLAMAEI J, AMIRI B. A hybrid evolutionary algorithm based on ACO and SA for cluster analysis[J]. Journal of applied sciences, 2008, 8(15): 2695–2702.
- [9] OZTURK C, HANCER E, KARABOGA D. Dynamic clustering with improved binary artificial bee colony algorithm[J]. Applied soft computing, 2015, 28: 69–80.
- [10] 蒙祖强, 胡玉兰, 蒋亮, 等. 基于混合蛙跳与阴影集优化的粗糙模糊聚类算法[J]. 控制与决策, 2015, 30(10): 1766–1772.
- MENG Zuqiang, HU Yulan, JIANG Liang, et al. Shuffled frog leaping algorithm and shadowed sets-based rough fuzzy clustering algorithm[J]. Control and decision, 2015, 30(10): 1766–1772.
- [11] 张新明, 王霞, 康强. 改进的灰狼优化算法及其高维函数和 FCM 优化[J]. 控制与决策, 2019, 34(10): 2073–2084.
- ZHANG Xinming, WANG Xia, KANG Qiang. Improved grey wolf optimizer and its application to high-dimensional function and FCM optimization[J]. Control and decision, 2019, 34(10): 2073–2084.
- [12] LI Shimin, CHEN Huiling, WANG Mingjing, et al. Slime mould algorithm: a new method for stochastic optimization[J]. Future generation computer systems, 2020, 111: 300–323.
- [13] ZHOU Yongquan, WU Haizhou, LUO Qifang, et al. Automatic data clustering using nature-inspired symbiotic organism search algorithm[J]. Knowledge-based systems, 2019, 163: 546–557.
- [14] ABDEL-BASSET M, CHANG V, MOHAMED R. HSMA_WOA: a hybrid novel Slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images[J]. Applied soft computing, 2020, 95: 106642.
- [15] MOSTAFA M, REZK H, ALY M, et al. A new strategy based on slime mould algorithm to extract the optimal model parameters of solar PV panel[J]. Sustainable energy technologies and assessments, 2020, 42: 100849.
- [16] ABDEL-BASSET M, MOHAMED R, CHAKRABORTTY R K, et al. An efficient binary slime mould algorithm integrated with a novel attacking-feeding strategy for feature selection[J]. Computers & industrial engineering, 2021, 153: 107078.
- [17] BOGAR E, BEYHAN S. Adolescent Identity Search Algorithm (AISA): a novel metaheuristic approach for solving optimization problems[J]. Applied soft computing, 2020, 95: 106503.
- [18] ZHAO Weiguo, ZHANG Zhenxing, WANG Liying. Manta ray foraging optimization: an effective bio-inspired optimizer for engineering applications[J]. Engineering applications of artificial intelligence, 2020, 87: 103300.
- [19] 徐艳. Kmeans 算法与群体智能算法 (PSO) 融合的研究与应用[D]. 呼和浩特: 内蒙古农业大学, 2019.
- XU Yan. Research and application of kmeans algorithm and swarm intelligence algorithm(PSO)fusion[D]. Hohhot: Inner Mongolia Agricultural University, 2019.
- [20] DUA D, GRAFF C. UCI Machine Learning Repository [EB/OL]. [2020–10–21]. <http://archive.ics.uci.edu/ml>.

作者简介:



贾鹤鸣, 教授, 主要研究方向为群体智能优化算法与工程应用。主持福建省自然科学基金等项目 10 余项。发表学术论文 60 余篇。



张棕淇, 硕士研究生, 主要研究方向为群体智能优化、机器学习和聚类技术。



姜子超, 硕士研究生, 主要研究方向为群体智能优化、机器学习、特征选择和聚类技术。