Journal of System Simulation

Volume 35 | Issue 3 Article 6

3-22-2023

# Multi-strategy Hybrid ABC for Microarray High-Dimensional Feature Selection

Chuandong Qin
*1.School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China;2.Ningxia Key Laboratory of Intelligent Information and Big Data Processing, Yinchuan 750021, China;*, qinchuandong123@163.com

Baosheng Li
*1.School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China;*, daishuli163@163.com

Baole Han
*1.School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China;*

# Multi-strategy Hybrid ABC for Microarray High-Dimensional Feature Selection


## Abstract

Abstract: Traditional feature selection approaches have major limitations for high-dimensional microarrays, and it is difficult to accurately and efficiently propose the best feature subset. To address this problem, *a multi-strategy hybrid artificial bee colony (ABC) algorithm based on wrapper is proposed*, *which mixes chaotic opposition-based learning strategy, elite guidance strategy, and Mantegna Lévy distribution strategy*, and proposes two new search strategies in the employed and onlooker bee phases respectively. *A new objective function is proposed for the microarray high-dimensional feature selection problem*, which balances the optimal performance of the model with the minimization of the feature subset size. Experimental results show that the algorithm is able to achieve high classification accuracy while still satisfying the feature subset size minimization objective to some extent. Moreover, it outperforms improved algorithms such as GABC and six new intelligent algorithms such as the salp swarm algorithm.



## Keywords

artificial bee colony(ABC) algorithm, high-dimensional feature selection, chaotic opposition-based learning, elite guidance strategy, Mantegna Lévy distribution

## Recommended Citation

Chuandong Qin, Baosheng Li, Baole Han. Multi-strategy Hybrid ABC for Microarray High-Dimensional Feature Selection[J]. Journal of System Simulation, 2023, 35(3): 515-524.

第 35 卷第 3 期 系统仿真学报© Vol. 35 No. 3
2023 年 3 月 Journal of System Simulation Mar. 2023

# 微阵列高维特征选择的多策略混合人工蜂群算法

秦传东[1,2]，李宝胜[1*]，韩宝乐[1]
(1. 北方民族大学 数学与信息科学学院，宁夏 银川 750021；2. 宁夏智能信息与大数据处理重点实验室，宁夏 银川 750021)

**摘要：**传统的特征选择方法对于高维微阵列具有较大的局限性，难以准确高效地提出最佳特征子集。针对该问题，提出了基于*wrapper*的多策略混合人工蜂群算法，该算法混合了混沌反向学习策略、精英引导策略、*Mantegna Lévy*分布策略，分别在雇佣蜂与观察蜂阶段提出了两种新的搜索策略。针对于微阵列高维特征选择问题，提出新的平衡模型性能最优与特征子集规模最小化目标函数。实验结果表明：该算法能够达到较高的分类准确率，可在一定程度上取得特征子集规模最小化的目标，且优于GABC等改进算法与樽海鞘群等六种新型智能算法。

**关键词：**人工蜂群算法；高维特征选择；混沌反向学习策略；精英引导策略；Mantegna Lévy分布
中图分类号：TP181; TP391.9 文献标志码：A 文章编号：1004-731X(2023)03-0515-10
DOI: 10.16182/j.issn1004731x.joss.21-1188
**引用格式:** 秦传东, 李宝胜, 韩宝乐. 微阵列高维特征选择的多策略混合人工蜂群算法[J]. 系统仿真学报, 2023, 35(3): 515-524.
**Reference format:** Qin Chuandong, Li Baosheng, Han Baole. Multi-strategy Hybrid ABC for Microarray High-Dimensional Feature Selection[J]. Journal of System Simulation, 2023, 35(3): 515-524.

## Multi-strategy Hybrid ABC for Microarray High-Dimensional Feature Selection

*Qin Chuandong*[1,2], *Li Baosheng*[1*], *Han Baole*[1]
(1. School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China;
2. Ningxia Key Laboratory of Intelligent Information and Big Data Processing, Yinchuan 750021, China)

**Abstract:** Traditional feature selection approaches have major limitations for high-dimensional microarrays, and it is difficult to accurately and efficiently propose the best feature subset. To address this problem, *a multi-strategy hybrid artificial bee colony (ABC) algorithm based on wrapper is proposed*, *which mixes chaotic opposition-based learning strategy, elite guidance strategy, and Mantegna Lévy distribution strategy*, and proposes two new search strategies in the employed and onlooker bee phases respectively. *A new objective function is proposed for the microarray high-dimensional feature selection problem*, which balances the optimal performance of the model with the minimization of the feature subset size. Experimental results show that the algorithm is able to achieve high classification accuracy while still satisfying the feature subset size minimization objective to some extent. Moreover, it outperforms improved algorithms such as GABC and six new intelligent algorithms such as the salp swarm algorithm.

**Keywords:** artificial bee colony(ABC) algorithm; high-dimensional feature selection; chaotic opposition-based learning; elite guidance strategy; Mantegna Lévy distribution

收稿日期：2021-11-18 修回日期：2022-01-06
基金项目：宁夏自然科学基金一般项目(2021AAC03230)
第一作者：秦传东(1976-)，男，副教授，博士，研究方向为智能计算与大数据分析。E-mail：qinchuandong123@163.com
通讯作者：李宝胜(1996-)，男，硕士生，研究方向为智能计算与大数据分析。E-mail：daishuli163@163.com

# 0 引言

在基因表达数据中，数据的规模主要是其庞大的特征数量造成的[1]。由于“维数灾难”的存在，许多机器学习算法在该数据集上的表现并不理想。不相关特征会提供误导信息，它会直接降低K近邻分类器(K-nearest neighbor, KNN)的分类效果，使得高斯核或者多项式核支持向量机(support vector machine, SVM)过拟合，降低其鲁棒性。冗余特征会覆盖有效特征，尽管决策树、随机森林等算法会在训练过程中剔除冗余特征，却延长了计算时间[2]。因此，特征选择不仅能够提高模型的性能，而且能够提高处理大规模数据集的效率。

传统的特征选择方法主要是根据距离度量、信息度量、相关性度量等标准提出的。如Relief算法采用欧氏距离度量特征之间及特征与类别之间的线性关系；决策树与随机森林等集成方法采用信息增益、Gini指数、互信息[3]等度量特征之间及特征与类别之间的非线性关系；Pearson相关系数法、方差分析、Fisher得分等采用统计原理度量特征与类别之间的相关性。与上述方法不同，以算法整体性能作为标准的方法直接与分类器交互，将特征选择与机器学习任务融合在一起，如文献[4]使用SVM进行癌症分类的基因选择。文献[5]使用Bagging C4.5算法做特征选择。然而，这种方法对于分类器算法依赖性强，只能获取局部最优解，而且在高维数据集上效率较低。

特征选择是一个NP-hard问题[2]，其搜索空间会随着特征数的增加呈指数级扩张，若特征数量为，则总的特征子集数量为。当维数较小时，采用穷举法或者分支定界法能获得全局最优的特征子集。面临高维问题，学者们提出利用序列搜索，如前向搜索、后向搜索于双向搜索等，但是该方法只能获得局部最优子集。为了解决这些问题，群智能算法如粒子群优化(particle swarm optimization, PSO)算法[6]、人工蜂群(artificial bee colony, ABC)算法[6]等被用于特征选择。其中，ABC作为一种较新的元启发式算法，具有易于实现，对初始化具有鲁棒性，收敛速度快的优点。文献[7]提出多种群的ABC算法来解决高维符号回归问题，提高了模型的泛化能力。文献[8]提出一种快速多目标ABC特征选择算法，保证了算法获得高质量的特征子集。文献[9]将ABC与梯度提升树(gradient boosting decision tree, GBDT)结合，利用ABC识别特征信息，实现决策树输入的全局优化，有效地较低了数据集的维度并取得了较好的分类精度。文献[10]为提取生物医学数据的最佳特征子集，在化学反映优化算法过程中加入人工蜂群的精英策略，使得特征子集的识别与分类精度上有一定的提高。

为了同时平衡2个关系：最大化算法性能与最小化特征子集的关系以及ABC探索与开发的关系，本文提出了一种基于Wrapper方法的多策略竞争的人工蜂群算法。首先，设计了使分类精度与特征子集规模达成共赢局面的加权目标函数。其次，提出多种不同的具有竞争关系的搜索策略。

# 1 人工蜂群算法

人工蜂群(ABC)算法已经被用于解决各个领域的复杂问题。基本的蜂群由雇佣蜂、观察蜂和侦察蜂组成，每个蜜源是目标问题的一个候选解。

(1) 初始化阶段

在该阶段，$SN$个种群的蜜源通过式(1)随机产生：

$$X_i^d = X_{\min}^d + rand(0,\ 1)\cdot(X_{\max}^d - X_{\min}^d) \tag{1}$$

式中：$i \in \{1, 2, \cdots, SN\}$；$d \in \{1, 2, \cdots, D\}$；$X_{\max}^d$, $X_{\min}^d$分别为第$d$维的上界与下界。

(2) 雇佣蜂阶段

在初始化阶段之后，雇佣蜂按照式(2)在给定的空间中搜索新的蜜源：

$$V_i^d = X_i^d + \phi_i^d (X_i^d - X_k^d) \tag{2}$$

式中：$k \in \{1, 2, \cdots, SN\}$；$d \in \{1, 2, \cdots, D\}$；且$i \neq k$，$\phi$为[−1, 1]的随机数。如果$V_i$比$X_i$更好，那么$V_i$将替换$X_i$，与$X_i$相关的计数器将重置为0，否则




DOI: 10.16182/j.issn1004731x.joss.21-1188

增加1。

(3) 观察蜂阶段

雇佣蜂蜜源搜索完成之后，观察蜂会在一些选中的蜜源邻域空间内搜索新的蜜源，而选择蜜源的概率依赖于适应值，一般选择锦标赛选择策略：

$$p_i = fit_i \Big/ \sum_{i=1}^{SN} fit_i \tag{3}$$

式中：$p_i$为选择蜜源$X_i$的概率；$fit_i$为该蜜源的适应度值。

$$fit_i = \begin{cases} \dfrac{1}{1+f(X_i)}, & f(X_i) \geqslant 0 \\ 1+abs(f(X_i)), & f(X_i) < 0 \end{cases} \tag{4}$$

式中：$f(X_i)$为$X_i$的目标函数值。同样，如果新的蜜源比之前的蜜源更好，那么将会保留新蜜源，并将计数器设置为0，否则增加1。

(4) 侦察蜂阶段

当计数器的次数大于预先设定的数，蜜源将会被舍弃，雇佣蜂成为侦察蜂并通过式(1)产生新的蜜源。

算法1　基本的ABC算法

step 1：根据式(1)初始化种群，设置初始化参数：种群数*SN*，最大函数评估次数*MNFE*，最大停滞开采次数*limit*，初始迭代*t*=1。

step 2：雇佣蜂根据式(2)搜索新的蜜源，并根据式(4)计算适应度值。

step 3：如果$f(V_i)>f(X_i)$，$X_i$替换$V_i$，*trial*=0，否则保持原有蜜源，*trial* = *trial* +1。

step 4：观察蜂根据式(3)计算的概率选择蜜源，并由式(2)进行邻域搜索产生新的蜜源，计算其适应度，执行step 3。

step 5：若*trial* >*limit*，则雇佣蜂放弃当前蜜源，成为侦察蜂，并根据式(1)产生新的蜜源。

step 6：保留最优解，*t*=*t*+1，若满足迭代终止条件，则输出最优解，否则转step 2。

# 2　多策略混合的ABC算法

原始的ABC特征选择算法没有做到同时平衡模型性能最大化与特征子集规模的最小化，与平衡提高全局探索与提高局部开发能力这两者之间的关系。本文提出的多策略混合的人工蜂群特征选择(multi-strategy hybrid artificial bee colony feature selection，MSHABC-FS)算法做了2个改进：①提出新的加权目标函数缓解模型性能最大化与特征子集规模的最小化的冲突关系；②提出多策略混合的ABC，同时兼顾了全局探索与局部开发。

## 2.1 目标函数

当模型性能的权重过大，特征子集规模最小化会在实际的最佳蜜源搜索过程中被轻视，甚至忽略掉，具体表现为特征子集规模在多次实验中频繁波动，不相关特征或冗余特征仍然有较大概率被保留，这些特征反过来会降低模型的性能。反之，若特征子集规模最小化目标权重过大，则会剔除掉关键特征信息，从而使分类性能更大概率会维持在较低水平。黄金分割法的最大特性在于和谐，它能够有效的平衡2个冲突目标间的关系。因此，本文采用黄金分割法的思想，设计出特征选择的目标函数：

$$F = \alpha^2 \cdot Error + \alpha \cdot \frac{n}{N} \tag{5}$$

式中：$\alpha^2+\alpha=1$；*Error*为分类错误率；*n*为特征子集个数；*N*为特征总数。

## 2.2 混沌反向学习策略

混沌映射具有随机性、遍历性的特点，本文根据文献[11]提出的利用Tent混沌算法来初始化ABC种群：

$$Z_{k+1} = \begin{cases} z_k/\beta, & 0 < z_k \leqslant \beta \\ (1-z_k)/(1-\beta), & \beta < z_k \leqslant 1 \end{cases} \tag{6}$$

将式(6)融入到式(1)中可得新的初始化方程为

$$X_i^d = X_{\min}^d + Z_i^d \cdot (X_{\max}^d - X_{\min}^d) \tag{7}$$

种群的初始化在群智能算法中会影响其全局收敛性与解的质量[12]，本文将混沌初始化与反向学习策略相结合，提出混沌反向学习策略，反向



 

学习的引入可以提高算法的开发能力。反向解的数学式为

$$X_{\text{op}}^d = X_{\min}^d + X_{\max}^d - X_i^d \tag{8}$$

### 2.3 精英引导策略

为进一步提高ABC的开发能力，分别考虑引用差分进化算法的变异算子与遗传算法中的杂交算子[13]，2种算子既能够保留精英，又能够基于精英引导种群搜索，提出的搜索方程为

$$V_i^d = X_{\text{best}}^d + \phi_i^d (X_{\text{best}}^d - X_k^d) \tag{9}$$

式中：$X_{\text{best}}^d$为当前迭代中种群最优个体。式(9)在保留精英解的同时，能够将种群快速地引导到最优个体邻域位置，因此该策略能够大大提高ABC的开发能力。然而，该搜索策略明显会失去种群的多样性，在处理复杂优化问题时，易陷入局部最优。为解决该问题，本文进一步引入Mantegna Lévy分布策略。

### 2.4 Mantegna Lévy分布

Mantegna Lévy分布是一种随机游走模型，具有马尔可夫性质，本文将其引入产生新的蜜源，增加解的多样性。其分布为

$$L(\alpha) \sim |\alpha|^{-1-\beta}$$

式中：$0<\beta\leqslant 2$；$\alpha$为随机步长，$\alpha \sim \dfrac{\mu}{\beta^{-1}|\upsilon|}$，$\begin{cases}\mu \sim N(0,\ \phi_\mu^2)\\ \upsilon \sim N(0,\ \phi_\upsilon^2)\end{cases}$，$\phi = \left(\dfrac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma((1+\beta)/2)\cdot\beta\cdot 2^{(\beta-1)/2}}\right)^{\beta^{-1}}$，其中：$\phi_\mu = \phi$；$\phi_\upsilon = 1$；$\Gamma$为标准的gamma函数。

首先，将该分布引入到混沌反向学习策略中，得到

$$X_i^d = X_{\min}^d + \lambda_i^d \cdot Z_i^d \cdot (X_{\max}^d - X_{\min}^d) \tag{10}$$

式中：$\lambda$为由Mantegna Lévy分布产生的随机数。然后，为平衡探索能力与开发能力，本文分别在雇佣蜂阶段和观察蜂阶段结合Mantegna Lévy分布对式(9)进行改进：

$$X_i^d = \lambda_i^d \cdot (X_{\text{best}}^d + \phi_i^d \cdot (X_{\text{best}}^d - X_k^d)) \tag{11}$$

$$V_i^d = X_{\text{best}}^d + \lambda_i^d \cdot \phi_i^d \cdot (X_{\text{best}}^d - X_k^d) \tag{12}$$

通过分析式(11)，(12)可以发现，本文提出的更新公式不仅保留了精英引导的性质，加快收敛速度，又借助马尔可夫的“无记忆”性维持了蜂群的多样性，使得2个冲突体达成了统一。

## 3 多策略混合的ABC特征选择算法

### 3.1 特征选择的编码策略

利用人工蜂群算法进行特征选择时，每个蜜源对应的是特征选择问题的一个解决方案，因此，合适的编码策略会影响特征子集的选择效果，本文使用连续编码策略如图1所示。

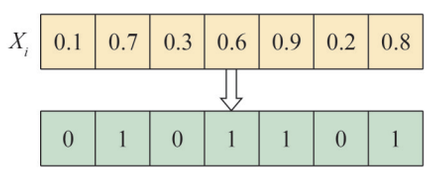


图1 特征选择的编码过程

Fig. 1 Coding sequence for feature selection

$$X_i = \begin{cases} 0, & x_{id} \leqslant 0.5 \\ 1, & x_{id} > 0.5 \end{cases} \tag{13}$$

式中：$x_{id} \in [0,\ 1]$为选择第$d$个特征作为特征子集的概率，$d \in \{1,\ 2, \cdots,\ D\}$。

### 3.2 MSHABC-FS算法

多策略混合的人工特征选择算法(MSHABC-FS)的步骤与流程如图2所示。

step 1：设置初始化ABC参数，种群规模*SN*，蜜源数量*SN*/2，最大函数评估次数*MNFE*，初始迭代*iter*=1，开采次数*limit*，初始标记*trail*=0，并分别利用式(1)和式(7)初始化。

step 2：输入数据集并划分为训练集与测试集，KNN对训练集进行训练，并将训练好的模型对测试集进行预测，根据式(5)计算每个蜜源的适应度值，确定最优个体蜜源。

step 3：雇佣蜂$i$根据式(11)搜索新的蜜源，若蜜源超出搜索空间，则利用边界阻尼法处理，计算其适应度值$f(V_i)$；

step 4：如果$f(V_i)>f(X_i)$，$X_i$替换$V_i$，$trial$=0，否则保持原有蜜源，$trial$= $trial$ +1；

step 5：观察蜂根据式(3)计算的概率选择蜜源，并由式(12)进行邻域搜索产生新的蜜源，若蜜源超出搜索空间，则利用边界阻尼法处理，计算其适应度$f(V_i)$，执行 step 4；

step 6：如果 $trial$ >$limit$，则雇佣蜂放弃当前蜜源成为侦察蜂，分别根据式(1)、式(7)和式(10)初始化蜜源，利用贪婪策略选择最佳蜜源；

step 7：保持当前迭代最优解，$iter$=$iter$+1，若满足迭代终止条件，则输出最优解及特征子集，否则转 step 2。

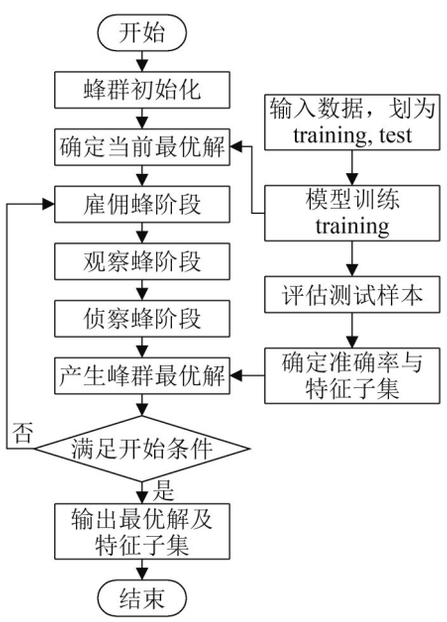


图2 MSHABC-FS 算法流程图
Fig. 2 MSHABC-FS flow chart

### 3.3 算法复杂性分析

本文提出的算法包含 4 个阶段：初始化；雇佣蜂阶段；观察蜂阶段与侦察蜂阶段。在初始化阶段，KNN 训练时间复杂度为$O(N)$，使用式(1)和(7)产生$N$个蜜源，又每个解的维度为$D$，则该阶段时间复杂度为$O(2ND)$；在雇佣蜂阶段，利用式(11)产生新蜜源，并使用贪婪选择，因此该阶段的时间复杂度为$O(N)$；在观察蜂阶段，需要计算选择每个解的概率，并根据式(12)产生新的蜜源，故时间复杂度为$O(N+ND)$；在侦察蜂阶段，需要分别利用式(1)、(7)和(10)初始化新的蜜源，故时间复杂度为$O(3ND)$。又后 3 个阶段需要重复迭代$M$次，所以本文所提出多策略混合 ABC 算法的总时间复杂度为$O(2ND+M(N+N+3ND))=O(MND)$。

## 4 实验结果与分析

首先，本文选择模型最为简单的$K$近邻作为分类器，其次分别选择了二分类和多分类的数据集，参照文献[10]的方法使用分类误差评估分类器性能：

$$Error=\frac{FP+FN}{TP+FP+TN+FN}$$

### 4.1 实验数据

本文既针对二分类问题进行分析、又对多分类问题进行说明。该实验数据选自 NCBI 数据库(https://www.ncbi.nlm.nih.gov/)公开数据，分别为 2 个二分类与 3 个多分类生物医学上基因组的微阵列数据集，具体数据描述见表 1。

表 1 数据概况
Table 1 Data profile

| 数据集 | 实例数 | 特征数 | 类别数 |
|---|---|---|---|
| GLI | 85 | 22 283 | 2 |
| GLA-BRA | 180 | 49 151 | 4 |
| CLL-SUB | 111 | 11 340 | 3 |
| TOX | 171 | 5 748 | 4 |
| SMK-CAN | 187 | 19 993 | 2 |

### 4.2 实验参数与设计

实验平台为 Windows 10 处理器 Intel(R) Core (TM) i5-8500 CPU @ 3.00GHz，内存 8GB。所有的程序是在 MATLAB 9.10(R2021a)上运行。本文将所提出的 MSHABC-FS 算法与传统的 PSO 算法对比，同时又与最新提出的算法进行对比，包括樽海鞘群算法(salp swarm algorithm, SSA)[14]、原子搜索优化算法(atom search optimization, ASO)[15]、探路者算法(pathfinder algorithm, PFA)[16]、蝴蝶优化算法(butterfly optimization algorithm, BOA)[17]与



 

灰狼优化算法(grey wolf optimizer, GWO)[18]。具体的实验参数设置见表2。

### 4.3 结果分析

为避免模型过度拟合，本文采用留出法将数据集按8:2划分为2个互斥的训练集与测试集进行验证。每种算法在每个数据集上独立运行30次。将MSHABC-FS算法与基本的ABC算法及其改进算法，包括GABC(gbest-guided ABC) [19]、CABC(candidate-based ABC) [20]、EABC(enhancing ABC)[13]、COABC (converge-onlookers ABC)[21]进行对比。具体对比结果见表3。可以看出，在独立运行30次后，与ABC及其4个改进算法相比，MSHABC-FS算法在5个数据集中搜索到的平均特征子集规模最小，平均适应度最少，直观地说明了本文算法所搜寻到的特征子集较其他5种算法更优；而且，在GLI数据集的平均准确率为100%，方差为0，说明该算法在该数据集上的性能较好，比较稳定。在GLA-BRA数据集的平均准确率为82.22%，仅次于ABC，在其余的3个数据上的平均准确率分别是80.91%，82.65%，85.14%，都达到了最高，表明了MSHABC-FS既达到了较高的分类准确率，又在一定程度上取得了特征子集规模最小化的目标。同时经过计算，在相同实验参数下，本文算法的平均运行时间在5个数据集上的平均排名约为1.8，也取得了较好效果。所以，MSHABC-FS的在4个指标上的综合表现都是比较可观的。

表2 多种算法参数设置
Table 2 Various algorithm parameter settings

| 算法 | 参数 |
|---|---|
| ABC及改进 | 初始种群$SN$=20；最大函数评估次数$MNFE$=2 000；开采次数$limit$=50 |
| SSA | 初始种群$SN$=20；最大函数评估次数$MNFE$=2 000；控制参数$c_1=2\exp[-(4L/iter_{\max})]$ |
| ASO | 初始种群$SN$=20；最大函数评估次数$MNFE$=2 000；深度权重$\alpha=50$；乘数权重$\beta=0.2$ |
| PFA | 初始种群$SN$=20；最大函数评估次数$MNFE$=2 000；$\alpha,\beta\in[1, 2]$；$u\in[-1, 1]$ |
| BOA | 初始种群$SN$=20；最大函数评估次数$MNFE$=2 000；感官模态$c=0.01$；幂指数$a=0.1+0.2t/iter_{\max}$；开关概率$P=0.8$ |
| GWO | 初始种群$SN$=20；最大函数评估次数$MNFE$=2 000；收敛因子$a=2(1-iter/iter_{\max})$ |

表3 与ABC及其改进算法对比
Table 3 Comparison with ABC and its improved algorithms

| 数据集 | 指标 | ABC | | GABC | | CABC | | EABC | | COABC | | MSHABC-FS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* |
| GLI | 1 | 10 959 | 106.1 | 11 100 | 142.1 | 10 580 | 70.8 | 10 969 | **57.23** | 10 660 | 71.8 | **734.4** | 222.2 |
| | 2 | 0.198 | 0.016 | 0.206 | 0.019 | 0.214 | 0.011 | 0.228 | 0.012 | 0.023 | 0.212 | **0.013** | **0.003** |
| | 3 | 98.24 | 0.028 | 87.65 | 0.065 | 94.71 | 0.019 | 93.53 | 0.018 | 95.29 | 0.037 | **100** | **0** |
| | 4 | 25.77 | **0.075** | 43.29 | 0.336 | 23.55 | 0.095 | 296.22 | 1.299 | **23.50** | 0.115 | 31.10 | 0.27 |
| GLA-BRA | 1 | 24 367.6 | 159.7 | 24 483.8 | 125.1 | 23 687.8 | 186.2 | 24 362.2 | 153.8 | 23 884.8 | **85.45** | **1 628** | 705.8 |
| | 2 | 0.292 | **0.001** | 0.302 | 0.009 | 0.304 | 0.010 | 0.312 9 | 0.008 | 0.302 | 0.008 | **0.123** | 0.014 |
| | 3 | **83.33** | **0** | 78.33 | 0.030 | 80.56 | 0.019 | 80.00 | 0.012 | 81.11 | 0.012 | 82.22 | 0.025 |
| | 4 | 102.32 | **0.31** | 142.49 | 0.85 | 93.81 | 0.57 | 688.39 | 1.907 | 94.58 | 0.458 | **79.72** | 3.11 |
| CLL-SUB | 1 | 5 545.6 | **37.5** | 5 620.6 | 57.46 | 5 301.2 | 50.2 | 5 569.2 | 49.33 | 5 353.9 | 44.03 | **666.9** | 424.7 |
| | 2 | 0.327 | **0.001** | 0.328 | 0.002 | 0.319 | 0.013 | 0.348 | 0.019 | 0.329 | 0.019 | **0.140** | 0.032 |
| | 3 | 77.27 | **0** | 65.91 | 0.096 | 77.27 | 0.021 | 74.09 | 0.031 | 75.91 | 0.307 | **80.91** | 0.036 |
| | 4 | 19.03 | 0.099 | 28.04 | 0.105 | **17.23** | 0.109 | 156.57 | 1.093 | 17.45 | 0.128 | 20.02 | 0.433 |

续表

| 数据集 | 指标 | ABC | | GABC | | CABC | | EABC | | COABC | | MSHABC-FS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* | *Mean* | *Std* |
| TOX | 1 | 2 854.7 | 36.7 | 2 866.7 | 32.43 | 2 658.6 | 56.68 | 2 833 | 46.41 | 2 691 | **26.85** | **284.4** | 136.45 |
| | 2 | 0.328 | **0.009** | 0.341 | 0.018 | 0.333 | 0.022 | 0.363 | 0.023 | 0.310 | 0.032 | **0.126** | 0.011 |
| | 3 | 77.65 | **0.015** | 59.12 | 0.058 | 74.71 | 0.034 | 71.76 | 0.037 | 78.82 | 0.051 | **82.65** | 0.017 |
| | 4 | 17.66 | 0.227 | 22.62 | 0.152 | 16.101 | 0.240 | 88.17 | 0.934 | 15.81 | **0.068** | **14.75** | 0.314 |
| SMK-CAN | 1 | 9 889.5 | 102.4 | 9 401.4 | 58.4 | 9 569.7 | **58.2** | 9 871.8 | 62.1 | 9 550.4 | 87.0 | **857.1** | 317.33 |
| | 2 | 0.327 | **0.007** | 0.323 | 0.016 | 0.323 | 0.014 | 0.332 | 0.008 | 0.329 | 0.006 | **0.108** | 0.012 |
| | 3 | 77.57 | 0.013 | 76.76 | 0.024 | 77.30 | 0.023 | 76.76 | 0.015 | 76.22 | **0.012** | **85.14** | 0.019 |
| | 4 | 48.27 | **0.147** | 44.32 | 0.196 | 43.82 | 0.165 | 287.09 | 1.029 | 43.93 | 0.249 | **37.19** | 0.862 |

注：1：特征数；2：适应度值；3：准确率/%；4：运行时间/s

图 3(a)~(e)为 MSHABC-FS 与 ABC 及其 4 种改进算法在 5 个微阵列高维数据集上的收敛曲线对比。可以直观地看出，5 种算法 ABC、CABC、GABC、EABC、COABC 在经过 10~20 次的迭代后便开始陷入局部最优，而本文所提 MSHABC-FS 算法可以跳出局部最优，更接近于全局最优值。同时，由于该算法加入了混沌初始化及反向学习策略，在初始阶段更容易找到全局最优点。与最近提出的算法对比，可以发现本文对传统 ABC 算法的改进使其在收敛性能上更具有优势。

将 MSHABC-FS 与最近几年提出的算法，包括 SSA、ASO、PFA、BOA、GWO 进行对比，具体对比结果见表 4。

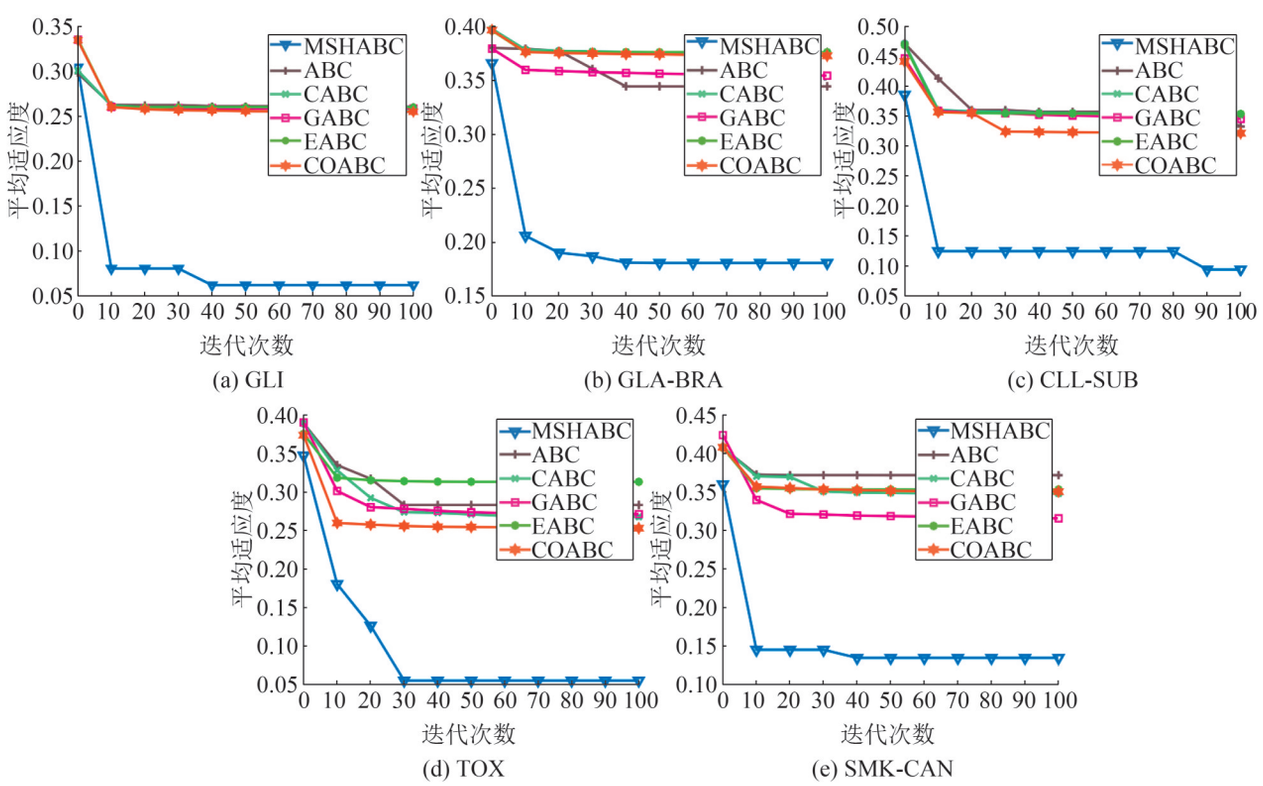


图 3　本文算法与 ABC 及其 4 种改进算法的收敛曲线

Fig. 3　Convergence curves of MSHABC with ABC and its 4 improved algorithms

由表 4 可以看出，在经过 30 次独立运行后，与 SSA 等 5 个算法进行对比，MSHABC-FS 算法在 5 个数据中提取的平均特征子集规模最小，平均适应度也最小；GWO 与 MSHABC-FS 在 GLI 数据集的平均分类准确率并列第一，在 GLA-BRA 等 4 个数据集上，MSHABC-FS 仅次于 GWO 算法，关于准确率的综合排名第二；同时，MSHABC-FS 算法在 5 个数据集关于运行时间的平均排名为 3.6，处于中间位置。说明在微阵列高维特征选择中，本文所提算法在特征数、适应度值、准确率方面具有较强的优势。

图 4(a)~(e)为 MSHABC-FS 与 SSA 等 5 种算法在 5 个微阵列高维数据集上的收敛曲线对比。可以看出，在 GLI 数据集上，本文所提 MSHABC-FS 算法与 GWO 算法的收敛点基本持平，但前者收敛到最优点的速度更快，在仅仅迭代 10 次之后，便靠拢到最优点附近。在 CLL-SUB 数据集上，前 40 次迭代中,该算法比 GWO 的收敛性能更好，在 40~90 次迭代中，GWO 的平均适应度优于前者，在后面 10 次迭代中，该算法取得微弱优势。在 GLA-BRA、TOX、SMK-CAN 数据集上 MSHABC-FS 算法的收敛性能都明显优于 SSA 等 5 种算法。说明本文所提算法在这 5 个微阵列高维数据集上的收敛性能更好，更容易确定最优特征子集。

表 4　与其他算法对比
Table 4　Comparison with other algorithms

| 数据集 | 指标 | SSA | | ASO | | PFA | | BOA | | GWO | | MSHABC-FS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Mean* | *std* | *Mean* | *std* | *Mean* | *std* | *Mean* | *std* | *Mean* | *std* | *Mean* | *std* |
| GLI | 1 | 10 992 | **50.1** | 10 677 | 120.4 | 11 013 | 115.2 | 3 810.9 | 754.7 | 2 658.1 | 336.1 | **734.4** | 222.2 |
| | 2 | 0.212 | 0.017 | 0.205 | 0.017 | 0.203 | 0.017 | 0.076 | 0.013 | 0.046 | 0.006 | **0.013** | **0.003** |
| | 3 | 95.88 | 0.03 | 96.47 | 0.03 | 97.65 | 0.03 | 98.24 | 0.028 | **100** | **0** | **100** | **0** |
| | 4 | 23.43 | **0.07** | 30.73 | 0.39 | 26.26 | 0.08 | **18.04** | 0.13 | 18.05 | 0.63 | 31.10 | 0.27 |
| GLA-BRA | 1 | 24 347.6 | **121.9** | 23 875.2 | 227.9 | 24 396.2 | 132.1 | 6 577.2 | 1 382.8 | 6 184 | 524.5 | **1 628** | 705.8 |
| | 2 | 0.306 | 0.007 | 0.299 | 0.008 | 0.293 | **0.001** | 0.164 | 0.006 | 0.141 | 0.009 | **0.123** | 0.014 |
| | 3 | 81.11 | 0.012 | 81.67 | 0.015 | 83.33 | **0** | 81.67 | 0.015 | **85.00** | 0.015 | 82.22 | 0.025 |
| | 4 | 94.04 | **0.26** | 98.53 | 3.96 | 100.37 | 0.37 | 65.47 | 0.393 | **52.61** | 1.61 | 79.72 | 3.11 |
| CLL-SUB | 1 | 5 557.8 | 73.2 | 5 436.5 | 73.4 | 5 549.1 | **33.32** | 1 508.2 | 396.5 | 1 474.1 | 174.5 | **666.9** | 424.7 |
| | 2 | 0.339 | 0.013 | 0.324 | 0.003 | 0.327 | **0.001** | 0.211 | 0.019 | 0.159 | 0.015 | **0.140** | 0.032 |
| | 3 | 75.45 | 0.023 | 77.27 | **0** | 77.27 | **0** | 74.09 | 0.037 | **82.27** | 0.026 | 80.91 | 0.036 |
| | 4 | 17.19 | 0.120 | 21.18 | 0.597 | 18.58 | 0.047 | **13.04** | **0.082** | 13.34 | 0.345 | 20.02 | 0.433 |
| TOX | 1 | 2 826.3 | 33.4 | 2 769.8 | 54.99 | 2 838.3 | **20.29** | 705.4 | 113.6 | 965.6 | 117.8 | **284.4** | 136.45 |
| | 2 | 0.344 | 0.016 | 0.302 | 0.015 | 0.337 | 0.012 | 0.201 | 0.026 | 0.135 | 0.025 | **0.126** | **0.011** |
| | 3 | 74.71 | 0.030 | 80.88 | 0.025 | 75.88 | 0.019 | 75.00 | 0.044 | **88.53** | 0.039 | 82.65 | **0.017** |
| | 4 | 16.24 | 0.302 | 20.16 | 0.432 | 17.11 | 0.148 | 12.71 | **0.120** | **12.26** | 0.457 | 14.75 | 0.314 |
| SMK-CAN | 1 | 9 867.2 | **34.24** | 9 624 | 66.2 | 9 898 | 71.9 | 2 367.8 | 208.9 | 2 923.9 | 342.7 | **857.1** | 317.33 |
| | 2 | 0.347 | 0.014 | 0.322 | 0.015 | 0.326 | 0.014 | 0.192 | 0.007 | 0.123 | 0.020 | **0.108** | 0.012 |
| | 3 | 74.32 | 0.014 | 77.57 | 0.026 | 77.84 | 0.025 | 76.22 | **0.011** | **89.19** | 0.031 | 85.14 | 0.019 |
| | 4 | 44.33 | **0.010** | 42.14 | 1.601 | 47.44 | 0.054 | 31.24 | 0.217 | **27.20** | 1.395 | 37.19 | 0.862 |

注：1：特征数；2：适应度值；3：准确率/%；4：运行时间/s

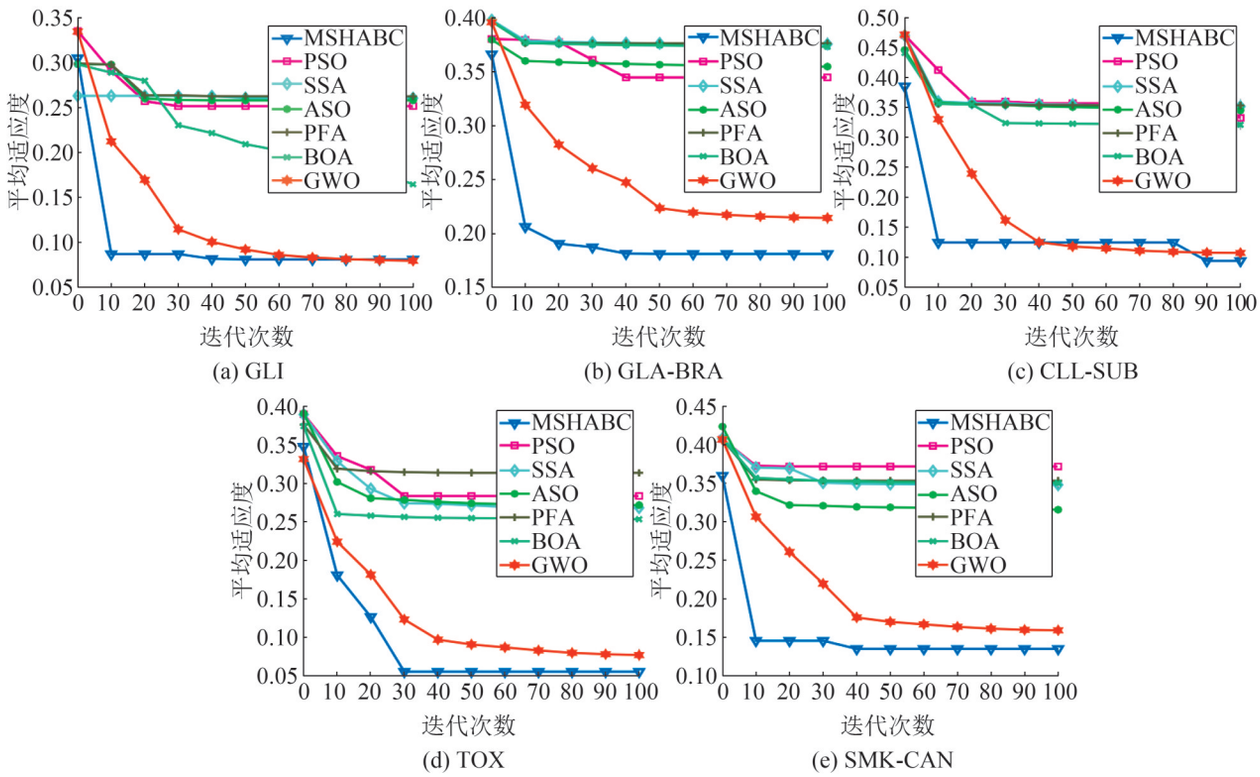


图4　本文算法与SSA等5种算法的收敛曲线

Fig. 4　Convergence curves of MSHABC and 5 algorithms such as SSA

## 5　结论

特征选择是模式识别与数据挖掘的一个难题，针对微阵列的高维特征选择问题，本文提出了多策略混合的人工蜂群结合KNN分类器进行特征选择(MSHABC-FS)算法。该算法在基本ABC算法的基础之上，将混沌初始化、反向学习策略、精英引导策略、Mantegna Lévy分布策略进行混合，使算法既保留了精英引导的性质，加快收敛速度，又维持了种群的多样性，较好地平衡了两者之间的关系。同时，又根据黄金分割法的思想，设计了能够平衡模型性能与特征子集规模关系的目标函数。实验结果综合表明了本文算法收敛性能与跳出局部最优的能力较高，不仅达到了较高的分类准确率，又在一定程度上取得了特征子集规模最小化的目标。

与其他最新算法相比较，本文所提算法在高维特征选择上具有一定准确率与时间优势，较好地解决了微阵列的高维特征选择的实际问题。但高维特征处理问题是一个复杂的研究领域，在未来工作中的重点将会进一步优化算法性能，降低高维特征选择的计算复杂度。

## 参考文献：

[1] Sun L, Zhang X, Qian Y, et al. Feature selection Using Neighborhood Entropy-Based Uncertainty Measures for Gene Expression Data Classification[J]. Information Sciences (S0020-0255), 2019, 502: 18-41.

[2] Nguyen B H, Xue B, Zhang M. A Survey on Swarm Intelligence Approaches to Feature Selection in Data Mining[J]. Swarm and Evolutionary Computation (S2210-6502), 2020, 54: 100663.

[3] Meyer P E, Schretter C, Bontempi G. Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity[J]. IEEE Journal of Selected Topics in Signal Processing (S1932-4553), 2008, 2(3): 261-274.

[4] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification Using Support Vector Machines[J]. Machine Learning (S1573-0565), 2002, 46(1): 389-422.

[5] Lee S J, Xu Z, Li T, et al. A Novel Bagging C4. 5 Algorithm Based on Wrapper Feature Selection for Supporting Wise Clinical Decision Making[J]. Journal of Biomedical Informatics (S1532-0464), 2018, 78: 144-155.

[6] 李炜, 巢秀琴. 改进的粒子群算法优化的特征选择方法[J]. 计算机科学与探索, 2019, 13(6): 990-1004.
Li Wei, Chao Xiuqin. Improved Particle Swarm Optimization Method for Feature Selection[J]. Frontiers of Computer Science and Technology, 2019, 13(6): 990-1004.

[7] Arslan S, Ozturk C. Multi Hive Artificial Bee Colony Programming for High Dimensional Symbolic Regression with Feature Selection[J]. Applied Soft Computing (S1568-4946), 2019, 78: 515-527.

[8] Wang X, Zhang Y, Sun X, et al. Multi-objective Feature Selection Based on Artificial Bee Colony: An Acceleration Approach with Variable Sample Size[J]. Applied Soft Computing (S1568-4946), 2020, 88: 106041.

[9] Rao H, Shi X, Rodrigue A K, et al. Feature Selection Based on Artificial Bee Colony and Gradient Boosting Decision Tree[J]. Applied Soft Computing (S1568-4946), 2019, 74: 634-642.

[10] 张戈, 王建林. 基于混合ABC和CRO的高维特征选择方法[J]. 计算机工程与应用, 2019, 55(11): 93-101.
Zhang Ge, Wang Jianlin. Feature Selection of High-dimensional Data Based on ABC and CRO Algorithm[J]. Computer Engineering and Applications, 2019, 55(11): 93-101.

[11] 匡芳君, 金忠, 徐蔚鸿, 等. Tent混沌人工蜂群与粒子群混合算法[J]. 控制与决策, 2015, 30(5): 839-847.
Kuang Fangjun, Jin Zhong, Xu Weihong, et al. Hybridization Algorithm of Tent Chaos Artificial Bee Colony and Particle Swarm Optimization[J]. Control and Decision, 2015, 30(5): 839-847.

[12] 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.
Gao Weifeng, Liu Sanyang, Jiao Hehua, et al. Particle Swarm Optimization with Search Operator of Artificial Bee Colony Algorithm[J]. Control and Decision, 2012, 27 (6): 833-838.

[13] Gao W, Liu S, Huang L. Enhancing Artificial Bee Colony Algorithm Using More Information-Based Search Equations[J]. Information Sciences (S0020-0255), 2014, 270: 112-133.

[14] Abualigah L, Shehab M, Alshinwan M, et al. Salp Swarm Algorithm: A Comprehensive Survey[J]. Neural Computing and Applications (S1433-3058), 2020, 32 (15): 11195-11215.

[15] Zhao W, Wang L, Zhang Z. Atom Search Optimization and Iits Application to Solve a Hydrogeologic Parameter Estimation Problem[J]. Knowledge-Based Systems (S0950-7051), 2019, 163: 283-304.

[16] Yapici H, Cetinkaya N. A New Meta-heuristic Optimizer: Pathfinder Algorithm[J]. Applied Soft Computing (S1568-4946), 2019, 78: 545-568.

[17] Arora S, Singh S. Butterfly Optimization Algorithm: A Novel Approach for Global Optimization[J]. Soft Computing (S1433-7479), 2019, 23(3): 715-734.

[18] Mirjalili S, Mirjalili S M, Lewis A. Grey Wolf Optimizer [J]. Advances in Engineering Software (S0965-9978), 2014, 69: 46-61.

[19] Zhu G, Kwong S. Gbest-Guided Artificial Bee Colony Algorithm for Numerical Function Optimization[J]. Applied Mathematics and Computation (S0096-3003), 2010, 217(7): 3166-3173.

[20] Gao W, Liu S, Huang L. A Novel Artificial Bee Colony Algorithm Based on Modified Search Equation and Orthogonal Learning[J]. IEEE Transactions on Cybernetics (S2168-2275), 2013, 43(3): 1011-1024.

[21] Luo J, Wang Q, Xiao X. A Modified Artificial Bee Colony Algorithm Based on Converge-Onlookers Approach for Global Optimization[J]. Applied Mathematics and Computation (S0096-3003), 2013, 219 (20): 10253-10262.




DOI: 10.16182/j.issn1004731x.joss.21-1188