Computer Science and Application 计算机科学与应用, 2023, 13(10), 1855-1869
Published Online October 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1310184

Hans汉斯

改进二进制沙丘猫群优化特征选择算法

周子航*，王丽娜

河北地质大学信息工程学院，河北 石家庄

收稿日期：2023年9月11日；录用日期：2023年10月10日；发布日期：2023年10月18日

摘　要

特征选择在机器学习的分类任务中被广泛应用，选择出的特征子集会直接影响后续学习算法的性能。针对沙丘猫群优化算法(SCSO)全局搜索能力弱、收敛速度慢问题，本文提出一种改进的二进制沙丘猫群优化特征选择算法。首先改进控制沙丘猫在搜索阶段和攻击阶段转换参数的调整方法，使用两阶段的改进收敛因子策略代替线性递减策略，以提升算法的全局搜索能力。其次受PSO算法位置更新公式的启发，引入社会学习因子和认知学习因子策略，提高算法的收敛速度。为了验证新提出算法在求解特征选择问题上的性能，本文选择了4种经典算法在8个UCI数据集上进行了对比测试，实验结果表明新提出算法的性能优于对比算法。

关键词

沙丘猫群优化算法，收敛因子，学习因子，特征选择

Improved Binary Sand Cat Swarm Optimization Feature Selection Algorithm

Zihang Zhou*, Lina Wang

School of Information and Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: Sep. 11th, 2023; accepted: Oct. 10th, 2023; published: Oct. 18th, 2023

Abstract

Feature selection is widely used in classification tasks of machine learning, and the selected feature sets directly affect the performance of subsequent learning algorithms. Aiming at the issues of weak global search ability and slow convergence speed of Sand Cat Swarm Optimization (SCSO), an improved binary sand cat swarm optimization feature selection algorithm is proposed in this

*第一作者。

文章引用: 周子航, 王丽娜. 改进二进制沙丘猫群优化特征选择算法[J]. 计算机科学与应用, 2023, 13(10): 1855-1869.
DOI: 10.12677/csa.2023.1310184

paper. Firstly, the adjustment method of controlling the transition parameters of sand cat in the search phase and attack phase is improved.This method employs a two-stage improved convergence factor strategy, replacingthe linear decrement strategy, aiming to enhance the algorithm's global search capability. Secondly, inspired by the position update formula of the PSO algorithm, social learning factor and cognitive learning factor strategies are introduced to improve the convergence speed of the algorithm. In order to verify the performance of the newly proposed algorithm in solving the feature selection problem, this study conducted comparative tests on eight UCI datasets using four classical algorithms. The experimental results demonstrate that the performance of the newly proposed algorithm outperforms the compared algorithms.

Keywords

Sand Cat Swarm Optimization Algorithm, Convergence Factor, Learning Factor, Feature Selection

1. 引言

特征选择是一种数据的预处理方式，用于从初始的特征空间中选择出与研究问题相关、有代表性的特征，剔除那些和研究问题无关、冗余的特征[1]。选择出来的特征子集具有更低的维度，能够提高分类算法的性能。

依据特征子集的搜索策略是否和后续的学习器相结合，可以将特征选择的方法分为过滤式特征选择、包装式特征选择和嵌入式特征选择[2]。过滤式特征选择首先对已知的数据集进行特征选择，然后将选好之后的特征子集用于模型训练，这两个过程是相互独立的[3]。过滤式特征选择的核心是选用某种准则对特征子集进行度量，如 Zheng K 等人[4]提出了一种结合最大信息熵(MIE)和最大信息系数(MIC)的过滤式特征子集选择方法。包装式特征选择中特征子集的选择标准会依赖于后续的学习器，所以首先需要确定后续的学习器，如 K 近邻(K-Nearest Neighbor, KNN)、支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest)等[5]。Wei J 等人[6]提出了一种新的变异增强 BPSO-SVM 算法，增加粒子的变异概率，从而使算法跳出局部最优，获得高质量的特征。徐明等人[7]对正余弦算法进行改进，首先设计出一种新的非线性递减函数替代原有线性递减函数，其次引入个体最优位置引领个体位置的更新，最后引入翻筋斗觅食机制以增加群体多样性，在解决高维特征选择问题上取得较好的效果。嵌入式特征选择就是将前面的特征选择算法嵌入后续的学习器中，即在学习器训练的过程中，同时完成特征选择。由于包装式特征选择算法具有可直接对算法本身进行优化、易于实现、精度与其它两种算法相比较高等优点，所以本文采用的是包装式的特征选择方法。

特征子集的搜索策略包括穷举法、分支定界法、前向搜索法、后向搜索法、随机搜索法等[8]。穷举法随着特征数量的增加，特征子集的数量会呈现指数级的增长趋势，算法的复杂度很高，只适用于维数低的情况。分支定界法较穷举法复杂度相对较低，但是随着特征数量的增大，复杂度也会呈现出指数级的增长趋势。前向搜索法：每次选择一个表现最好的特征加入到已选好的特征子集中。后向搜索法：每次从已选好的特征子集中剔除一个表现最差的特征。随机搜索法：使用一定的随机优化算法如遗传算法[9] (Genetic Algorithm, GA)、蚁狮算法[10] (Ant Lion Optimizer, ALO)、粒子群优化算法[11] (Particle Swarm Optimization, PSO)、灰狼算法(Grey Wolf Optimizer, GWO)等生成特征子集，再利用特定的评价函数去评

定所选出的特征子集的优劣，通过不断地迭代使得选出来的特征子集的变化趋于稳定，最终得到最优特征子集。徐明等人[12]对灰狼算法进行改进并将其用于求解特征选择问题中，设计一种基于正弦函数的非线性过渡参数策略代替原来的线性递减策略，且在最优灰狼个体的选取上，引入小孔成像学习策略产生新的候选个体。改进算法能有效地提高分类精度，选择最优特征。随机搜索可以防止算法陷入局部最优，找到近似最优解。随机搜索算法被广泛的用于求解优化问题，路雪刚等人[13]对鲸鱼优化算法进行改进，并将其用于求解畜禽废弃物运输路径优化问题。MPanda 等人[14]将灰狼算法用于求解路径规划问题。和其它的搜索方法相比，随机搜索的搜索效率远高于其它搜索方法。

受沙丘猫搜索和捕食猎物行为的启发，Amir Seyyedabbasi 等人[15]于 2022 年提出了沙丘猫群优化算法(Sand Cat Swarm Optimization, SCSO)。该算法通过一种自适应机制，控制算法在搜索阶段和攻击阶段之间的过渡，具有较好的全局寻优能力，在求解高维和多目标问题中表现良好，可以将其用于求解特征选择问题。YIMING LI 等人[16]提出了一种基于随机变异和精英协作的沙丘猫群优化算法，该算法首先引入了一种非线性周期调整机制，以平衡算法的全局探索能力和局部开发能力，加快算法的收敛速度。其次引入随机变异的精英协作策略，使算法能够跳出局部极值，进一步提高了算法的寻优精度和收敛速度。并与文献中其它群智能优化方法进行了对比实验，验证了改进策略的有效性。Dijana Jovanovic 等人[17]提出了一种基于改进的沙丘猫群优化算法的入侵检测特征选择，在 SCSO 算法的基础上嵌入了著名的人工蜂群算法(Artificial Bee Colony, ABC)的搜索机制。通过在两个著名数据集(UNSW-NB15 和 CICIDS-2017)上验证所提出的方法，并将结果与处理相同问题并在类似配置下工作的其他前沿算法的报告结果进行比较，证明了性能改进。综合来说，SCSO 算法具有较强的优化问题求解能力，但是其解存在精度低、容易陷入局部最优、迭代后期收敛速度慢等缺点，算法性能具有较大的提升空间。

2. 基本的沙丘猫群优化算法

沙丘猫群优化算法是受沙丘猫的觅食行为启发而提出的一种新的随机优化算法。沙丘猫利用它们奇妙的听觉特性，可以探测到地下活动的猎物。沙丘猫的觅食行为分为搜索猎物和攻击猎物两个阶段，并通过一种机制去控制两种行为之间的平衡。算法的数学模型如式(1)所示：

$$\overrightarrow{\boldsymbol{Pos}}(t+1)=\begin{cases}\overrightarrow{\boldsymbol{Pos}_b}(t)-\vec{\boldsymbol{r}}\cdot\overrightarrow{\boldsymbol{Pos}_{rnd}}\cdot\cos(\theta), & \text{if } |R|\le 1;\\ \vec{\boldsymbol{r}}\cdot\left(\overrightarrow{\boldsymbol{Pos}_{bc}}(t)-rand(0,1)\cdot\overrightarrow{\boldsymbol{Pos}_c}(t)\right), & \text{otherwise.}\end{cases} \tag{1}$$

其中 $\overrightarrow{\boldsymbol{Pos}}(t+1)$ 表示第 $t+1$ 次迭代后沙丘猫的位置，其中 $\overrightarrow{\boldsymbol{Pos}_b}(t)$ 为第 t 代时种群的最优解的位置，$\overrightarrow{\boldsymbol{Pos}_c}(t)$ 表示第 t 代时个体当前位置，$\overrightarrow{\boldsymbol{Pos}_{bc}}(t)$ 表示第 t 代时种群一个候选解的位置，$\overrightarrow{\boldsymbol{Pos}_{rnd}}$ 为当前位置和最优位置之间的一个随机位置，计算公式如式(6)所示，$\vec{\boldsymbol{r}}$ 表示每只猫的灵敏度范围，计算公式如式(4)所示，θ 为通过轮盘赌算法选择出的随机角度。参数 R 用来控制沙丘猫在搜索阶段和攻击阶段之间的过渡。R 的计算公式如式(2)所示：

$$\vec{\boldsymbol{R}}=2\times\overrightarrow{\boldsymbol{r}_G}\times rand(0,1)-\overrightarrow{\boldsymbol{r}_G} \tag{2}$$

其中 $\overrightarrow{\boldsymbol{r}_G}$ 为沙丘猫的常规的灵敏度范围，$\overrightarrow{\boldsymbol{r}_G}$ 的计算公式如式(3)所示：

$$\overrightarrow{\boldsymbol{r}_G}=S_M-\frac{S_M\times iter_c}{iter_{max}} \tag{3}$$

$\overrightarrow{\boldsymbol{r}_G}$ 随着迭代次数的增加线性下降。其中 $iter_c$ 为当前迭代次数，$iter_{max}$ 为最大迭代次数，S_M 是由沙丘猫的听觉特征激发的，初始时设置其值为 2。$\vec{\boldsymbol{r}}$ 表示每只猫的灵敏度范围，计算公式如式(4)所示：

$$\vec{\boldsymbol{r}}=\overrightarrow{\boldsymbol{r}_G}\times rand(0,1) \tag{4}$$

当参数 R 的绝对值小于等于 1 的时候，处于攻击阶段，使用式(5)进行位置更新。

$$\overrightarrow{\boldsymbol{Pos}}(t+1)=\overrightarrow{\boldsymbol{Pos}_b}(t)-\vec{\boldsymbol{r}}\cdot\overrightarrow{\boldsymbol{Pos}_{rnd}}\cdot\cos(\theta) \tag{5}$$

其中 $\overrightarrow{\boldsymbol{Pos}_b}(t)$ 为第 t 代时种群的最优解的位置，θ 为利用轮盘选择算法为每只沙丘猫选择出的一个随机角度，$\overrightarrow{\boldsymbol{Pos}_{rnd}}$ 为当前位置和最优位置之间的一个随机位置，以确保沙丘猫可以靠近猎物，计算公式如式(6)所示：

$$\overrightarrow{\boldsymbol{Pos}_{rnd}}=\left|rand(0,1)\cdot\overrightarrow{\boldsymbol{Pos}_b}(t)-\overrightarrow{\boldsymbol{Pos}_c}(t)\right| \tag{6}$$

当参数 R 的绝对值大于 1 的时候，处于搜索阶段，使用公式(7)进行位置更新。

$$\overrightarrow{\boldsymbol{Pos}}(t+1)=\vec{\boldsymbol{r}}\cdot\left(\overrightarrow{\boldsymbol{Pos}_{bc}}(t)-rand(0,1)\cdot\overrightarrow{\boldsymbol{Pos}_c}(t)\right) \tag{7}$$

其中 $\overrightarrow{\boldsymbol{Pos}_{bc}}(t)$ 表示第 t 代时种群一个候选解的位置，$\overrightarrow{\boldsymbol{Pos}_c}(t)$ 表示第 t 代时个体当前位置。

综上所述，沙丘猫的位置更新分为搜索和攻击两个阶段。在攻击阶段时，使用轮盘赌算法可以避免算法陷入局部最优陷阱，引入随机位置可以保证沙丘猫在不断地向猎物位置靠近。在搜索阶段时，选择一个随机候选解来引导沙丘猫的位置更新，沙丘猫能够找到其它的可能的猎物位置，防止算法陷入局部最优。

3. 改进的二进制沙丘猫群优化算法

本文通过引入两阶段的改进收敛因子策略和改进学习因子策略，提高了算法的全局搜索能力，加快了算法的收敛速度，并在特征子集的评价函数中加入了特征和类别之间的关联性作为评价函数的一部分。从标准 UCI 数据集值选取 8 个样本数和特征数量均不同的数据集来测试算法的性能，实验结果表明，改进后的算法具有更好的分类效果。

3.1. 基本二进制沙丘猫群优化特征选择算法

3.1.1. 基本二进制沙丘猫群优化算法

基本的沙丘猫群优化算法只能用于处理连续的问题，为了将沙丘猫群优化算法用于求解离散型问题，需要将沙丘猫的位置离散化。在初始化及位置更新时，将每只沙丘猫的位置离散化，经过离散化处理后可得到二进制沙丘猫群优化算法(Binary Sand Cat Swarm Optimization, BSCSO)，算法的具体实现如下所示。

在种群初始化时，随机生成每只沙丘猫的初始位置，每只沙丘猫每一维的位置都为 0 或 1。初始时的位置生成公式如式(8)所示：

$$P_{ij}=randint(0,1) \tag{8}$$

其中 P_{ij} 表示第 i 个个体在第 j 维中的取值。$randint(0,1)$ 表示 0 或 1。

在位置更新时，根据公式(1)计算出每只沙丘猫的位置，然后通过文献 18 中 8 种不同的 Sigmoid 函数(如式(9)~(16)所示)将每只沙丘猫的位置离散化[18]。具体的计算公式如式(17)所示。

$$S_1:S(x)=\frac{1}{1+\mathrm{e}^{-2x}} \tag{9}$$

$$S_2:S(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{10}$$

$$S_3:S(x)=\frac{1}{1+\mathrm{e}^{-x/2}} \tag{11}$$

$$S_4: S(x)=\frac{1}{1+e^{-x/3}} \tag{12}$$

$$V_1: V(x)=\left|\frac{\sqrt{\pi}}{2}\int_0^{(\sqrt{\pi}/2)x} e^{-t^2}dt\right| \tag{13}$$

$$V_2: V(x)=\left|\tanh(x)\right| \tag{14}$$

$$V_3: V(x)=\left|x/\sqrt{1+x^2}\right| \tag{15}$$

$$V_4: V(x)=\left|\frac{2}{\pi}\arctan\left(\frac{\pi}{2}x\right)\right| \tag{16}$$

$$PN_{ij}^t=\begin{cases}0, & \text{if } Sig\left(P_{ij}^t\right)\le rand(0,1);\\ 1, & \text{otherwise.}\end{cases} \tag{17}$$

其中 P_{ij}^t 代表离散化前沙丘猫群第 i 个个体在第 j 维的取值，Sig 表示不同的激活函数，PN_{ij}^t 代表离散化后沙丘猫群第 i 个个体在第 j 维的取值，$rand(0,1)$ 表示 0 到 1 之间的随机数。

3.1.2. 基本二进制沙丘猫群优化算法求解特征选择问题

基本二进制沙丘猫群优化算法求解特征选择问题的伪代码如下所示：

算法 1：BSCSO 伪代码

1：初始化参数。最大迭代次数 $iter_{max}$、每只猫的灵敏度范围 $\vec{\boldsymbol{r}}$ 、沙丘猫的常规的灵敏度范围 $\overrightarrow{\boldsymbol{r}_G}$ 。
2：利用式(9)~(17)对种群进行初始化，使用 2.3 节中的特征子集的评价函数计算初始时每个个体的适应度值。
3：while t <$iter_{max}$ do
4：计算每个个体的适应度值，并将最佳位置 $\overrightarrow{\boldsymbol{Pos}_b}(t)$ 及其适应度值保存。
5：　for 每个代理 do
6：根据轮盘赌算法选择一个随机角度
7：根据公式(3)更新收敛因子 $\overrightarrow{\boldsymbol{r}_G}$
8：　　if (abs |R| ≤1)
9：根据公式(5)更新每个个体的位置。
10：将更新后的位置根据 2.1.1 节中离散化方法将个体的位置离散化。
11：计算每个个体的适应度值。如果更新后的个体更优，则将其替换为最佳位置 $\overrightarrow{\boldsymbol{Pos}_b}(t)$ 。
12：　　else
13：根据公式(6)更新每个个体的位置。
14：将更新后的位置根据 2.1.1 节中离散化方法将个体的位置离散化。
15：计算每个个体的适应度值。如果更新后的个体更优，则将其替换为最佳位置 $\overrightarrow{\boldsymbol{Pos}_b}(t)$ 。
16：　　end if
17：　end for
18：t++
19：end while

首先对种群进行初始化，根据特征子集的评价函数计算出每个个体的适应度值，挑选出适应度值最

好的个体并记录其所在的位置，根据位置更新公式更新每个个体的位置及全局最优位置，不断地重复上述过程直至达到最大迭代次数。

3.2. 改进的二进制沙猫群优化算法特征选择算法

3.2.1. 改进策略

沙丘猫独特的觅食方式使得沙丘猫群优化算法具有良好的局部寻优能力，但同时算法也存在容易陷入局部最优的问题。为了提高算法的全局搜索能力，本文提出了一种两阶段的改进收敛因子的策略，增加了处于搜索阶段的迭代次数，提高了算法的全局搜索能力。

在搜索阶段和攻击阶段时的位置更新公式中，SCSO 算法通过随机解引领个体位置的更新，这样虽然可以防止算法陷入局部最优，但同时也会导致算法出现收敛速度慢、精度低等问题。受 PSO 算法中粒子的位置更新公式的启发，我们引入社会学习因子和认知学习因子策略，通过全局最优位置和局部最优位置共同引领个体位置的更新，加快了算法的收敛速度。

(1) 两阶段的改进收敛因子策略

沙丘猫的觅食行为分为搜索和攻击两个阶段，搜索阶段类似全局搜索，即对整个可行域进行搜索。攻击阶段类似局部搜索，即对某一部分区域进行密切搜索。如何协调算法在搜索和攻击阶段之间的过渡是至关重要的。

由公式(1)可知，R 控制算法在两个阶段之间的过渡，在$|R| > 1$ 时，算法的搜索范围广，算法具有较好的全局搜索能力，当$|R| < 1$ 时，算法具有较好的局部开发能力。由公式(2)可知，R 的取值由 $\overrightarrow{r_G}$ 决定，由公式(3)可知，$\overrightarrow{r_G}$ 在算法的迭代过程中由 2 线性递减到 0。而在实际的沙丘猫捕食猎物的过程中，收敛因子 $\overrightarrow{r_G}$ 的线性变化并不能满足实际的要求。本文提出了一种两阶段的改进收敛因子的策略，在第一阶段，采用线性递减的策略，在第二阶段，采用对数函数非线性调整收敛因子，与原有的线性递减的策略相比，增加了处于搜索阶段的迭代次数，提高了算法的全局搜索能力，加快了算法的收敛速度。收敛因子 $\overrightarrow{r_G}$ 的更新公式如式(18)所示：

$$r_G = \begin{cases} S_M - S_M * t/iter_{Max}, & \text{if } r_G > 1.2; \\ 1.2/\lg 1.2 * \lg\left(1.2 - \left(t/iter_{max}\right)^k\right), & \text{otherwise.} \end{cases} \tag{18}$$

其中 S_M 值设为 2，k 的值设为 3，t 表示当前迭代次数，$iter_{max}$ 表示算法的最大迭代次数。

(2) 引入社会学习因子和认知学习因子策略

由公式(7)可知，在搜索阶段，通过一个随机候选解引领沙丘猫个体位置的更新，这样虽然可以保证算法的随机性较强，不易陷入局部最优，但同时会使得算法的收敛速度较慢。为了解决这一问题，我们引入了全局最优位置，使用随机候选解和全局最优解共同引领位置的更新。改进后的位置更新公式如式(19)所示：

$$\overrightarrow{\boldsymbol{Pos}}(t+1) = \vec{\boldsymbol{r}} \cdot \left(a * \overrightarrow{\boldsymbol{Pos}_{bc}}(t) + (1-a) * \overrightarrow{\boldsymbol{Pos}_{b}}(t) - rand(0,1) \cdot \overrightarrow{\boldsymbol{Pos}_{c}}(t)\right) \tag{19}$$

权重系数 a 值设置为 0.9。由式(1)可知，沙丘猫的位置更新方式具有过大的随机性，在维度较高的条件下，算法的性能会很差。为了加快算法的收敛速度，受粒子群算法影响，让搜索方向保持一定的惯性，不易轻易改变，我们引入了惯性权重 w。为了加快算法的寻优速度，能够更快的找到最优解，我们引入了学习因子 c_1、c_2。通过随机位置、局部最优位置和全局最优位置共同引领沙丘猫位置的更新，使沙丘猫能够更快的向最优位置靠近。改进后的位置更新公式如式(20)~(21)所示。公式(20)用于搜索阶段的位置更新，公式(21)用于攻击阶段的位置更新。

$$\overrightarrow{\boldsymbol{Pos}}(t+1)=w\cdot\vec{\boldsymbol{r}}\cdot\left(a*\overrightarrow{\boldsymbol{Pos}_{bc}}(t)+(1-a)*\overrightarrow{\boldsymbol{Pos}_{b}}(t)-rand(0,1)\cdot\overrightarrow{\boldsymbol{Pos}_{c}}(t)\right)+c_1\cdot rand(0,1)\cdot\left(\overrightarrow{\boldsymbol{Pos}_{b}}(t)-\overrightarrow{\boldsymbol{Pos}_{c}}(t)\right)+c_2\cdot rand(0,1)\cdot\left(\overrightarrow{\boldsymbol{Pos}_{p}}(t)-\overrightarrow{\boldsymbol{Pos}_{c}}(t)\right) \tag{20}$$

$$\overrightarrow{\boldsymbol{Pos}}(t+1)=w\cdot\left(\overrightarrow{\boldsymbol{Pos}_{b}}(t)-\vec{\boldsymbol{r}}\cdot\overrightarrow{\boldsymbol{Pos}_{rnd}}\cdot\cos(\theta)\right)+c_1\cdot rand(0,1)\cdot\left(\overrightarrow{\boldsymbol{Pos}_{b}}(t)-\overrightarrow{\boldsymbol{Pos}_{c}}(t)\right)+c_2\cdot rand(0,1)\cdot\left(\overrightarrow{\boldsymbol{Pos}_{p}}(t)-\overrightarrow{\boldsymbol{Pos}_{c}}(t)\right) \tag{21}$$

其中惯性权重 w 的值设置为 0.9，学习因子 c_1、c_2 的值设置为 0.5，$\overrightarrow{\boldsymbol{Pos}_{b}}(t)$ 表示第 t 代时的局部最优位置。

将上述两种策略融入到算法中，在增强算法的全局搜索能力的同时加快算法的收敛速度。

3.2.2. 改进二进制沙丘猫群优化算法求解特征选择问题

改进的二进制沙丘猫群优化算法(Improved Binary Sand Cat Swarm Optimization, PBSCSO)求解特征选择问题的伪代码如下所示：

算法 2：PBSCSO 伪代码

1：初始化参数。最大迭代次数 $iter_{max}$、每只猫的灵敏度范围 $\vec{\boldsymbol{r}}$ 、沙丘猫的常规的灵敏度范围 $\overrightarrow{\boldsymbol{r}_G}$ 。
2：利用式(9)~(17)对种群进行初始化，使用 2.3 节中的特征子集的评价函数计算初始时每个个体的适应度值。
3：while t <$iter_{max}$ do
4：计算每个个体的适应度值，并将全局最优位置 $\overrightarrow{\boldsymbol{Pos}_{b}}(t)$ 和局部最优位置 $\overrightarrow{\boldsymbol{Pos}_{p}}(t)$ 及其适应度值保存。
5：　for 每个代理 do
6：根据轮盘赌算法选择一个随机角度
7：根据公式(3)更新收敛因子 $\overrightarrow{\boldsymbol{r}_G}$
8：if (abs |R| ≤1)
9：根据 2.2.1 节中公式(21)更新每个个体的位置。
10：将更新后的位置根据 2.1.1 节中离散化方法将个体的位置离散化。
11：根据 2.3 节中特征子集评价函数计算每个个体的值。如果更新后的个体更优，则将其替换为最佳位置 $\overrightarrow{\boldsymbol{Pos}_{b}}(t)$ 。
12：　else
13：根据 2.2.1 节中公式(20)更新每个个体的位置。
14：将更新后的位置根据 2.1.1 节中离散化方法将个体的位置离散化。
15：根据 2.3 节中特征子集评价函数计算每个个体的值。如果更新后的个体更优，则将其替换为最佳位置 $\overrightarrow{\boldsymbol{Pos}_{b}}(t)$ 。
16：　end if
17：　end for
18：t++
19：end while

3.2.3. 改进策略的验证

从标准 UCI 数据集中选取 8 个数据集测试算法的性能，数据集详情见表 3 所示。使用 S_3 型函数作为沙丘猫位置离散化的转换函数，用 2.3 节中的特征子集评价函数计算个体的适应度值。设置算法的执行次数为 30 次，种群规模为 10，最大迭代次数为 100，使用 K-NN 分类器以及 holdout 验证的方式对筛选出来的特征子集的性能进行评估，K 的值设置为 5。

表 1 是在 BSCSO 算法的基础上引入两阶段的改进收敛因子策略(Convergence Binary Sand Cat Swarm

Optimization, CBSCSO)与 BSCSO 算法的比较。使用 3.3 节中的特征子集的评价指标作为评估标准，其中 AVG 表示运行 30 次所得的最优适应度值的平均值，AVGALL 表示在 8 个数据集上最优适应度值的平均值。从结果可以看出 CBSCSO 在 7 个数据集上表现较优，且在所有数据集上的平均性能较好。说明本文提出的两阶段的改进收敛因子策略是有效的，能够加强算法的全局搜索能力，找到更优的特征子集。

Table 1. The average of the optimal fitness values of the two algorithms
表 1. 两种算法的最优适应度值的平均值

数据集名称	指标	BSCSO	CBSCSO
Breastcancer	AVG	0.037	0.036
BreastEW	AVG	0.062	0.060
HeartEW	AVG	0.169	0.168
Lymphopgraphy	AVG	0.177	0.176
SpectEW	AVG	0.164	0.165
IonosphereEW	AVG	0.149	0.148
WineEW	AVG	0.068	0.058
Zoo	AVG	0.089	0.080
	AVGALL	0.114	0.111

表 2 是在 BSCSO 算法的基础上引入社会学习因子和认知学习因子策略(Study Binary Sand Cat Swarm Optimization, SBSCSO)与 BSCSO 算法的比较。其中 AVG 表示运行 30 次所得的适应度值收敛时的迭代次数的平均值。从结果可以看出 SBSCSO 在 8 个数据集上表现都较优，说明本文提出的引入改进学习因子的策略能够提高算法的收敛速度，更快找到最优特征子集。

Table 2. The number of iterations when the fitness values of the two algorithms converge
表 2. 两种算法适应度值收敛时的迭代次数

数据集名称	指标	BSCSO	SBSCSO
Breastcancer	AVG	39.2	26.4
BreastEW	AVG	28.4	19.2
HeartEW	AVG	45.7	40.8
Lymphopgraphy	AVG	39.3	32.6
SpectEW	AVG	51.1	40.5
IonosphereEW	AVG	41.0	33.7
WineEW	AVG	33.0	26.4
Zoo	AVG	29.0	23.4

3.3. 特征子集评价函数

分类问题的特征选择中，为了提高算法的性能，在特征子集的评价函数设计时，不仅需要考虑分类的

准确率，还需将特征子集的维度、特征和类别的相关性考虑在内。其中特征和类别的相关性通过特征和类别之间的互信息来计算。本文利用改进的二进制沙丘猫群优化算法搜索特征子集，用特征值为 1 或 0 代表特征是否被选中，引入 *K*-NN 分类器验证分类的准确率，构造了一种新的特征子集的评价函数如式(22)所示。

$$feature\_val = 0.98 * ERate + 0.01 * \frac{SelNum}{Sum} + 0.01 * \frac{\sum\left(1 - I\left(X;Y\right)\right)}{SelNum} \tag{22}$$

其中 *ERate* 表示分类的错误率，*SelNum* 表示被选中的特征子集(即离散化后特征值为 1)中特征的数目，*Sum* 表示数据集中特征的数目，*I*(*X*;*Y*)表示每个被选中的特征和类别之间互信息。在分类问题中，分类准确率是最重要的评价标准，因此将分类错误率的权重设置为 0.98，特征子集的维度和特征和类别之间的相关性的权重设置为 0.01。

4. 实验及结果分析

4.1. 测试文本

使用 python 编写程序，从标准 UCI 数据集值选取 8 个样本数和特征数量均不同的数据集来测试算法的性能。表 3 给出了数据集的名称、特征值的数量、样本的数量、类别数量。

Table 3. Data set introduction
表 3. 数据集介绍

数据集名称	特征值数量	样本数量	类别数量
Breastcancer	9	699	2
BreastEW	30	569	2
HeartEW	13	270	2
Lymphopgraphy	18	148	4
SpectEW	22	267	2
IonosphereEW	34	351	2
WineEW	13	178	3
Zoo	16	101	7

4.2. 测试算法及相关参数

将 PBSCSO 与 GA、BPSO [19]、BGWO [20]、BSCSO 对比，使用 KNN 分类器以及 holdout 验证的方式对筛选出来的特征子集的性能进行评估，*K* 的值设置为 5。五种群智能算法的初始种群数为 10，最大迭代次数为 100，算法运行次数为 30。GA 算法中交叉概率为 0.9，变异概率为 0.1。PSO 算法中设置惯性权重 *w* 为 1.0，学习因子 $c_1 = c_2 = 1.8$。BSCSO 中角度范围设置为[0, 360]。

除此之外，在 BSCSO 算法中沙丘猫的位置离散化时，我们采用式(9)~(16)中 8 中不同的转换函数，并将结果最好的转换函数用于 PBSCSO 中。

4.3. 评价指标

使用分类准确率、适应度值、最优特征子集的维度作为评价指标与其他的 4 种群智能优化算法进行比较，证明 PBSCSO 算法的有效性。

(1) 分类准确率

分类准确率是分类正确的样本数占样本总数的比例，是度量分类问题最主要的评价指标。分类准确率计算公式如式(23)所示：

$$avgAcc = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N}\sum_{j=1}^{N} compare\left(O_j, R_j\right) \tag{23}$$

其中 K 表示算法的执行次数，N 表示用于验证的样本的总数量，O_j 表示输出得到的类标签，R_j 表示真实的类标签。$compare$ 将算法输出得到的类标签与数据的真实标签进行对比，如果相同为 1，不同为 0。

(2) 适应度值

适应度值由三部分组成，分别是分类的准确率、特征子集的维度、特征和类别之间的相关性。算法执行 K 次，求每次执行完后所求得的最佳位置的适应度值的平均值，计算公式如式(24)所示：

$$avgFit = \frac{1}{K}\sum_{i=1}^{K} w_1 * ERate + w_2 * \frac{SelNum}{Sum} + w_3 * \frac{\sum\left(1-I\left(X;Y\right)\right)}{SelNum} \tag{24}$$

其中 K 表示算法的执行次数，w_1 表示分类准确率所占的权重，取值为 0.98，w_2 表示特征子集的维度所占的权重，取值为 0.01，w_3 表示被选中的特征和类别之间的相关性的权重，取值为 0.01。$ERate$ 表示分类的错误率，$SelNum$ 表示被选中的特征子集的维度，Sum 为特征的总数目，$I(X;Y)$表示每个被选中特征和类别之间的相关性。

(3) 最优特征子集的维度

最优特征子集的维度就是每次迭代结束后所挑选出来的特征子集中值为 1 的特征的个数。算法执行 K 次，求特征子集维度的平均值，计算公式如式(25)所示：

$$avgSize = \frac{1}{K}\sum_{i=1}^{K} size\left(f_i\right) \tag{25}$$

其中 K 表示算法的执行次数，f_i 表示第 i 次执行所选出的特征子集，$size(f_i)$表示所选出的特征子集的维度。

4.4. 实验分析

本文采用五种不同的算法在表 3 中 8 种不同的 UCI 数据集上进行了 30 次独立重复实验。表 4 是 BSCSO 算法在 8 种测试数据集上利用八种不同的转换函数最终得到的分类准确率的情况。其中 AVG 表示运行 30 次所得的分类准确率的平均值，STD 表示其对应的标准差，AVGALL 表示在 8 个数据集上分类准确率的平均值，STDALL 表示在 8 个数据集上的分类准确率标准差的平均值。从结果可以看出，虽然使用 S_3 型转换函数进行离散化在个别数据集上表现并不突出，但其所得到的准确率的平均值最高，对于不同数据集上进行特征子集选择的平均性能较好。因此，选取 S_3 型转换函数进行离散化，并将该转换函数用于以下 PBSCSO 中沙丘猫位置的离散化。

表 5 是 PBSCSO 与其它的四种算法分类准确率的比较，其中 AVG 表示运行 30 次所得的分类准确率的平均值，STD 表示其对应的标准差，AVGALL 表示在 8 个数据集上的分类准确率的平均值，STDALL 表示在 8 个数据集上的分类准确率标准差的平均值。从结果可以看出，BGWO 在 Breastcancer 数据集上表现最好，BSCSO 在 BreastEW 和 IonosphereEW 数据集上表现最好，PBSCSO 在 Breastcancer、HeartEW、Lymphopgraphy、SpectEW、WineEW 和 Zoo 数据集上表现都是最好的，在 BreastEW 和 IonosphereEW 数据集上的表现仅次于最优值。相比较于 GA、BPSO 和 BGWO，本文提出的 PBSCSO 算法的特征选择机制在大多数数据集上都可以进一步的提升分类的效果，说明在进行特征选择时，PBSCSO 能够有效的提取出和类别相关信息，找到最优特征子集。

Table 4. Classification accuracy of BCSCSO algorithm under eight conversion functions
表 4. BSCSO 算法在 8 种转换函数下的分类准确率及其标准差

数据集名称	指标	S1	S2	S3	S4	V1	V2	V3	V4
Breastcancer	AVG	0.961	0.958	0.960	0.963	0.963	0.964	0.958	0.951
	STD	0.008	0.010	0.012	0.014	0.012	0.015	0.009	0.010
BreastEW	AVG	0.938	0.942	0.946	0.932	0.937	0.933	0.938	0.936
	STD	0.010	0.013	0.019	0.011	0.011	0.019	0.014	0.016
HeartEW	AVG	0.804	0.801	0.793	0.781	0.766	0.741	0.766	0.765
	STD	0.045	0.045	0.037	0.044	0.041	0.040	0.039	0.041
Lymphopgraphy	AVG	0.788	0.799	0.791	0.824	0.753	0.802	0.786	0.753
	STD	0.070	0.062	0.042	0.048	0.043	0.044	0.066	0.045
SpectEW	AVG	0.793	0.808	0.801	0.807	0.791	0.783	0.780	0.787
	STD	0.024	0.040	0.043	0.050	0.038	0.044	0.041	0.043
IonosphereEW	AVG	0.850	0.845	0.845	0.840	0.890	0.884	0.879	0.869
	STD	0.023	0.040	0.042	0.042	0.038	0.033	0.037	0.031
WineEW	AVG	0.907	0.907	0.914	0.896	0.907	0.920	0.907	0.912
	STD	0.022	0.038	0.040	0.026	0.028	0.021	0.033	0.024
Zoo	AVG	0.841	0.867	0.896	0.883	0.835	0.845	0.861	0.880
	STD	0.062	0.057	0.062	0.074	0.076	0.031	0.050	0.072
	AVGALL	0.860	0.865	0.868	0.865	0.855	0.859	0.859	0.856
	STDALL	0.033	0.038	0.037	0.038	0.035	0.030	0.039	0.035

Table 5. The mean value and standard deviation of classification accuracy of the five algorithms
表 5. 5 种算法的分类准确率的平均值及其标准差

数据集名称	指标	GA	BPSO	BGWO	BSCSO	PBSCSO
Breastcancer	AVG	0.961	0.959	0.965	0.960	0.965
	STD	0.015	0.015	0.010	0.012	0.010
BreastEW	AVG	0.936	0.934	0.929	0.946	0.942
	STD	0.013	0.016	0.020	0.019	0.015
HeartEW	AVG	0.781	0.789	0.718	0.793	0.819
	STD	0.054	0.060	0.085	0.037	0.049
Lymphopgraphy	AVG	0.777	0.754	0.748	0.791	0.811
	STD	0.055	0.058	0.050	0.042	0.043
SpectEW	AVG	0.796	0.789	0.806	0.801	0.811
	STD	0.041	0.045	0.038	0.043	0.034

Continued

IonosphereEW	AVG	0.835	0.839	0.836	0.845	0.844
	STD	0.034	0.032	0.031	0.042	0.033
WineEW	AVG	0.916	0.915	0.924	0.914	0.926
	STD	0.026	0.031	0.045	0.040	0.034
Zoo	AVG	0.876	0.876	0.893	0.896	0.903
	STD	0.056	0.057	0.064	0.062	0.048
	AVGALL	0.859	0.856	0.852	0.868	0.877
	STDALL	0.036	0.039	0.042	0.037	0.033

分类准确率是衡量分类问题的性能的最重要的指标，从表 5 可以看出，PBSCSO 在不同数据集上进行特征子集选择的平均准确率最高、平均标准差最小，这表明使用 PBSCSO 算法的特征选择机制的分类效果最好且比较稳定。

表 6 是 PBSCSO 与其他的四种算法适应度值的比较，其中 AVG 表示运行 30 次所得的最优适应度值的平均值，STD 表示其对应的标准差，AVGALL 表示在 8 个数据集上的适应度值的平均值，STDALL 表示在 8 个数据集上的适应度值标准差的平均值。从结果可以看出，BPSO 在 WineEW 数据集上表现最好，BSCSO 在 HeartEW 数据集上表现最好，PBSCSO 在 Breastcancer、BreastEW、Lymphopgraphy、SpectEW、IonosphereEW 和 Zoo 数据集上表现都是最好的，在 WineEW 和 HeartEW 数据集上表现仅次于最优值。相比较于 GA、BPSO 和 BGWO，本文提出的 PBSCSO 算法的特征选择机制在大多数数据集上都可以进一步的提高算法的寻优能力，说明在进行特征选择时，PBSCSO 能够选出最优特征子集。相比于 BSCSO，本文提出的 PBSCSO 算法在大部分数据集上适应度值都较小，说明本文提出的两阶段的改进收敛因子策略是有效的，能够提升算法的寻优能力。

Table 6. The mean value and standard deviation of the optimal fitness values of the five algorithms
表 6. 5 种算法的最优适应度值的平均值及其标准差

数据集名称	指标	GA	BPSO	BGWO	BSCSO	PBSCSO
Breastcancer	AVG	0.037	0.036	0.041	0.037	0.036
	STD	0.003	0.004	0.004	0.003	0.003
BreastEW	AVG	0.062	0.063	0.074	0.062	0.061
	STD	0.007	0.007	0.010	0.009	0.006
HeartEW	AVG	0.185	0.181	0.255	0.169	0.174
	STD	0.021	0.016	0.053	0.015	0.018
Lymphopgraphy	AVG	0.186	0.179	0.217	0.177	0.171
	STD	0.019	0.023	0.027	0.024	0.020
SpectEW	AVG	0.164	0.163	0.176	0.164	0.156
	STD	0.019	0.018	0.025	0.014	0.012

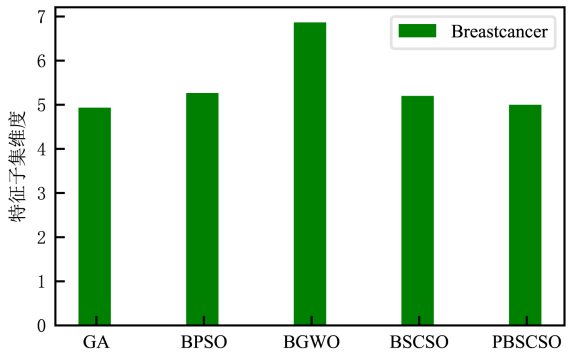
Continued

IonosphereEW	AVG	0.144	0.150	0.160	0.149	0.141
	STD	0.013	0.013	0.020	0.013	0.020
WineEW	AVG	0.066	0.059	0.080	0.068	0.064
	STD	0.011	0.010	0.019	0.011	0.010
Zoo	AVG	0.101	0.093	0.123	0.089	0.088
	STD	0.016	0.015	0.034	0.009	0.014
	AVGALL	0.118	0.115	0.112	0.114	0.111
	STDALL	0.018	0.013	0.024	0.012	0.012

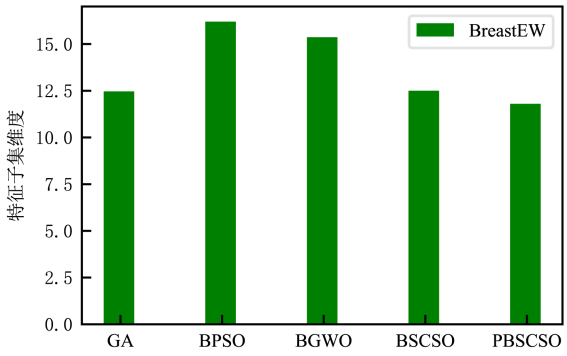
适应度值综合考虑分类的准确率和所选出来的最优特征子集的维度，从表 6 可以看出，PBSCSO 在不同数据集上进行特征子集选择的平均适应度值最小、平均标准差最小，这表明使用 PBSCSO 算法的寻优能力最好且比较稳定。

图 1 是在 8 种不同的数据集上利用 5 种不同的算法得到的最优特征子集的数量平均值情况。和 BPSO、BGWO 和 BSCSO 相比，PBSCSO 在大多数数据集上的特征子集的维度都是最小的，和 GA 相比，两种算法在 8 个数据集上的特征子集的维度基本一致。从结果可以看出，本文提出的 PBSCSO 算法能够在保证分类准确率的前提下，有效的降低所选出的最优特征子集的维度。

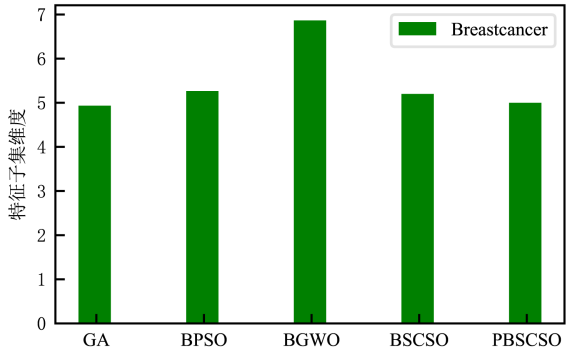
综上所述，PBSCSO 算法能够在保证特征子集维度较小的情况下提高分类的准确率，寻到更优的特征子集。将 PBSCSO 算法用于求解特征选择问题能取得较好的效果。



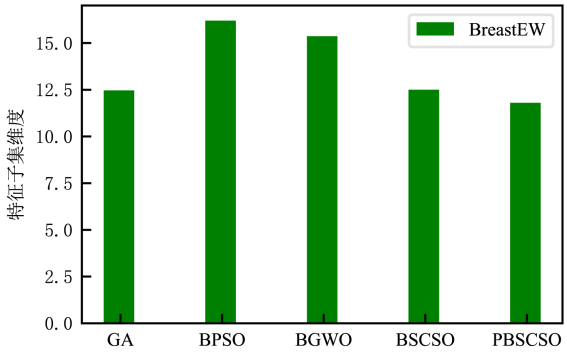
(a) Breastcancer



(b) BreastEW



(c) HeartEW



(d) Lymphopgraphy

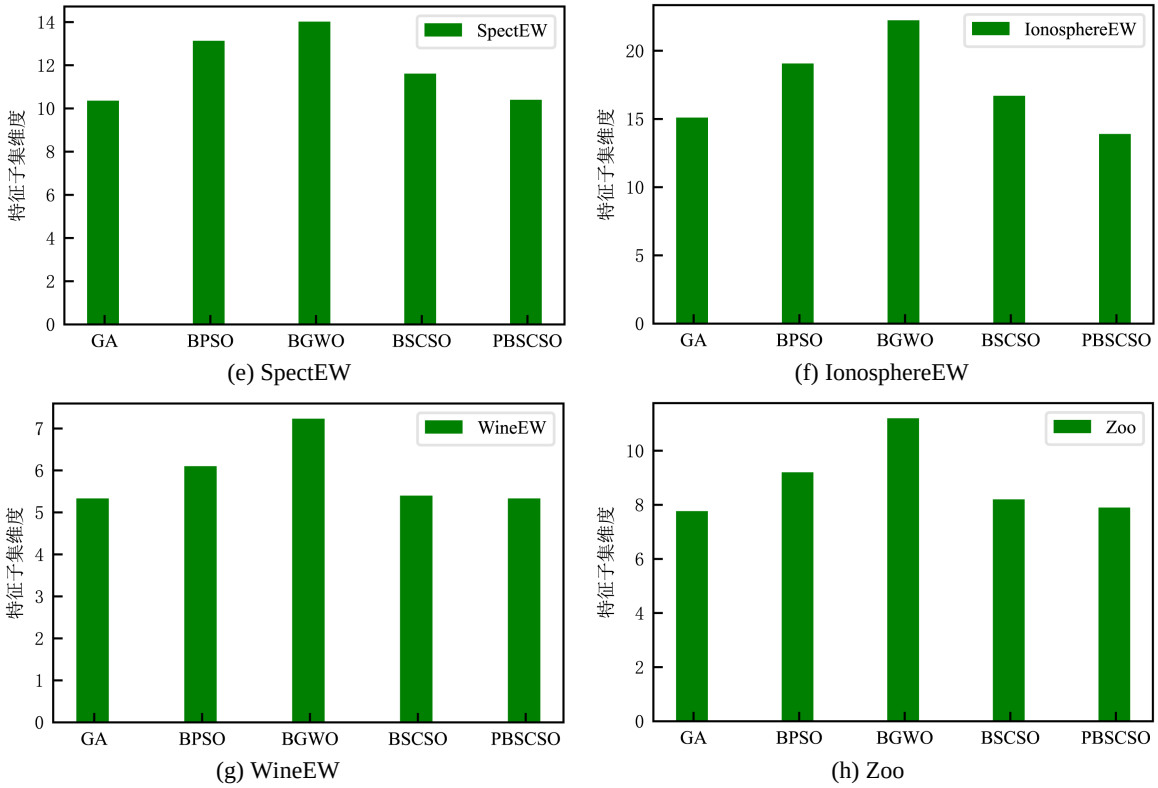


Figure 1. Number of feature subsets
图 1. 特征子集数量

5. 结论

PBSCSO 特征选择算法通过引入两阶段的改进收敛因子策略，增加了处于搜索阶段的迭代次数，提高了算法的全局搜索能力。通过引入社会学习因子和认知学习因子策略，有效地降低了个体位置更新时的随机性，加快了算法的收敛速度。实验结果表明，引入两阶段的改进收敛因子策略能够增强算法的寻优能力，引入社会学习因子和认知学习因子策略能够加快算法的收敛速度。与 GA、BPSO、BGWO 和 BSCSO 相比，PBSCSO 特征选择算法在提高分类准确率、降低最优特征子集的维度、鲁棒性和收敛性等方面均表现较好。

参考文献

[1] Wang, L., Qian, Y., *et al*. (2019) Feature Selection Using Lebesgue and Entropy Measures for Incomplete Neighborhood Decision Systems. *Knowledge-Based Systems*, **186**, Article ID: 104942. https://doi.org/10.1016/j.knosys.2019.104942

[2] 刘艳, 程璐, 孙林. 基于KS检验和邻域粗糙集的特征选择方法[J]. 河南师范大学学报: 自然科学版, 2019, 47(2): 21-28.

[3] Jin, Z., Teng, S., Zhang, J., *et al*. (2022) Structural Damage Recognition Based on Filtered Feature Selection and Convolutional Neural Network. *International Journal of Structural Stability and Dynamics*, **22**, Article ID: 2250134. https://doi.org/10.1142/S0219455422501346

[4] Zheng, K., Wang, X., Wu, B., *et al*. (2020) Feature Subset Selection Combining Maximal Information Entropy and Maximal Information Coefficient. *Applied Intelligence*, **50**, 487-501. https://doi.org/10.1007/s10489-019-01537-x

[5] Wang, J. and Zhang, L. (2021) Discriminant Mutual Information for Text Feature Selection. In: *International Conference on Database Systems for Advanced Applications*, Springer, Cham, 136-151. https://doi.org/10.1007/978-3-030-73197-7_9

[6] Wei, J., Zhang, R., Yu, Z., *et al*. (2017) A BPSO-SVM Algorithm Based on Memory Renewal and Enhanced Mutation Mechanisms for Feature Selection. *Applied Soft Computing*, **58**, 176-192. https://doi.org/10.1016/j.asoc.2017.04.061

[7] 徐明，羊洋，龙文．解决高维优化和特征选择的多策略改进正弦余弦算法[J]．科学技术与工程，2023，23(13): 5632-5640.

[8] Mao, Y. and Yang, Y. (2019) A Wrapper Feature Subset Selection Method Based on Randomized Search and Multi-layer Structure. *BioMed Research International*, **2019**, Article ID: 9864213. https://doi.org/10.1155/2019/9864213

[9] Katoch, S., Chauhan, S.S. and Kumar, V. (2021) A Review on Genetic Algorithm: Past, Present, and Future. *Multimedia Tools and Applications*, **80**, 8091-8126. https://doi.org/10.1007/s11042-020-10139-6

[10] Abualigah, L., Shehab, M., Alshinwan, M., *et al*. (2021) Ant Lion Optimizer: A Comprehensive Survey of Its Variants and Applications. *Archives of Computational Methods in Engineering*, **28**, 1397-1416. https://doi.org/10.1007/s11831-020-09420-6

[11] Jain, M., Saihjpal, V., Singh, N., *et al*. (2022) An Overview of Variants and Advancements of PSO Algorithm. *Applied Sciences*, **12**, Article No. 8392. https://doi.org/10.3390/app12178392

[12] 徐明，龙文．基于多策略融合灰狼优化算法的特征选择方法[J]．科学技术与工程, 2021, 21(20): 8544-8551.

[13] 路雪刚，张雪花，张梦桃．基于改进鲸鱼优化算法的畜禽废弃物运输路径优化问题[J]．科学技术与工程，2022, 22(25): 11120-11129.

[14] Panda, M., Das, B. and Pati, B. (2020) Global Path Planning for Multiple AUVs Using GWO. *Archives of Control Sciences*, **30**, 77-100.

[15] Seyyedabbasi, A. and Kiani, F. (2022) Sand Cat Swarm Optimization: A Nature-Inspired Algorithm to Solve Global Optimization Problems. *Engineering with Computers*, **39**, 2627-2651. https://doi.org/10.1007/s00366-022-01604-x

[16] Arai, K. (2022) Improved ISODATA Clustering Method with Parameter Estimation Based on Genetic Algorithm. *International Journal of Advanced Computer Science and Applications*, **13**, 187-193. https://doi.org/10.14569/IJACSA.2022.0130523

[17] Jovanovic, D., Marjanovic, M., Antonijevic, M., *et al*. (2022) Feature Selection by Improved Sand Cat Swarm Optimizer for Intrusion Detection. 2022 *International Conference on Artificial Intelligence in Everything* (*AIE*), Nicosia, 2-4 August 2022, 685-690. https://doi.org/10.1109/AIE57029.2022.00134

[18] 王泽昆，贺毅朝，李焕哲，等．基于新颖S型转换函数的二进制粒子群优化算法求解具有单连续变量的背包问题[J]．计算机应用, 2021, 41(2): 461-469.

[19] Bharti, K.K. and Singh, P.K. (2016) Opposition Chaotic Fitness Mutation Based Adaptive Inertia Weight BPSO for Feature Selection in Text Clustering. *Applied Soft Computing*, **43**, 20-34. https://doi.org/10.1016/j.asoc.2016.01.019

[20] Emary, E., Zawbaa, H.M. and Hassanien, A.E. (2016) Binary Grey Wolf Optimization Approaches for Feature Selection. *Neurocomputing*, **172**, 371-381. https://doi.org/10.1016/j.neucom.2015.06.083