控制与决策
Control and Decision

基于改进秃鹰搜索算法的同步优化特征选择

贾鹤鸣，姜子超，李瑶

引用本文:
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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1025

基于改进秃鹰搜索算法的同步优化特征选择

贾鹤鸣[1,2†], 姜子超[2], 李 瑶[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，哈尔滨 150040)

摘 要: 针对传统支持向量机在封装式特征选择中分类效果差、子集选取冗余、计算性能易受核函数参数影响的不足,利用元启发式优化算法对其进行同步优化. 首先利用莱维飞行策略和模拟退火机制对秃鹰搜索算法的局部搜索能力与勘探利用解空间能力进行改进,通过标准函数的测试结果验证其改进的有效性;其次将支持向量机核函数参数作为待优化目标,利用改进后的算法在封装式特征选择模型中搜寻最优核函数参数,同时获得相对应的最优特征子集;最后对UCI存储库的12个标准数据集进行特征选择仿真实验,在平均分类准确率、所选特征个数及适应度值上进行综合评估分析. 实验结果表明,所提算法可有效降低特征维度,能够更准确地实现数据分类,在空间搜索与求解精度方面较原算法及其他非线性最优化算法表现优秀,具有一定的工程应用价值.

关键词: 秃鹰搜索优化；莱维飞行；模拟退火；支持向量机；封装式特征选择

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1025 **开放科学(资源服务)标识码(OSID):**

引用格式: 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.

Simultaneous feature selection optimization based on improved bald eagle search algorithm

JIA He-ming[1,2†], *JIANG Zi-chao*[2], *LI Yao*[2]

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

Abstract: Aiming at the shortcomings of support vector machines in wrapper feature selection, such as poor classification effect, redundant subset selection, and computational performance that are easily affected by kernel function parameters, the meta-heuristic optimization algorithm is used to optimize it simultaneously. Firstly, the local search ability and the exploration and utilization solution space ability of the bald eagle search algorithm are improved by using the Levy flight strategy and simulated annealing mechanism, the test results of the standard function prove that the improvement is effective. Then, the kernel function parameters of the support vector machine are taken as the optimization objective, and the improved algorithm is used to search for the optimal kernel function parameters in the wrapper feature selection model and it obtains the corresponding feature subset simultaneously. Finally, a feature selection simulation is performed on the 12 standard data sets of the UCI repository, and the average classification accuracy, the number of selected features and the fitness value are comprehensively evaluated and analyzed. The experimental results show that the proposed algorithm can effectively reduce the feature dimension and achieve data classification more accurately. Compared with the original algorithm and other nonlinear optimization algorithms, it has certain engineering application value.

Keywords: bald eagle search optimization；Levy flight；simulated annealing；support vector machine；wrapper feature selection

0 引 言

在商业、社交媒体和科学研究等领域中,随着数据计算需求增加,数据特征维度数以千计,样本数量也随之不断增加,机器学习正面临极具挑战的“维数灾难”等问题. 作为数据挖掘的一部分,数据预处理可显著改善数据挖掘性能[1]. 特征选择基于某种评价准则从原始特征空间中选择特征子集,通过消除数据中的不相关特征,减少计算时间,提升学习准

收稿日期: 2020-07-24；**录用日期:** 2020-12-03.

基金项目: 福建省自然科学基金面上项目(2021J011128)；福建省教育厅中青年教师教育科研项目(JAT200618)；三明市科技计划引导性项目(2021-S-8, 2020-G-61)；三明学院引进高层次人才科研启动经费项目(20YG14)；三明学院科学研究发展项目(B202009)；福建省农业物联网应用重点实验室开放研究基金项目(ZD2101).

责任编委: 林崇.

†通讯作者. E-mail: jiahemingluckу99@126.com.

确率,进而大幅度、连续改善数据挖掘能力,是重要的数据预处理工作. 文献[2]提出了一种基于最大信息系数和Gram-Schmidt正交化的过滤式特征选择方法,以消除不相关冗余,并将其应用于高维生物医学数据挖掘中,提高了计算精度和效率;文献[3]提出了一种基于鲸鱼优化的封装式特征选择方法,利用二进制鲸鱼优化算法搜索最优特征子集完成分类任务,效果良好. 针对封装式特征选择分类问题,支持向量机[4](support vector machine, SVM)在计算非线性高维的数据分类时具有出色表现,但效果受其内部参数影响较大,故学者通过优化方法对参数进行有效选取,以获取最佳分类性能. 在SVM优化中,文献[5]验证了随机梯度下降(stochastic gradient descent, SGD)算法在解决机器学习中大规模SVM非线性优化问题时的有效性和可行性;文献[6]基于次梯度和改进的拟牛顿法求解SVM基本核的权系数,所提拟牛顿多核学习(quasi newton-multiple kernel learning, QN-MKL)算法具有良好的收敛速度和分类精度. 上述为数学应用类优化方法在求解SVM分类问题中的研究工作,虽在求解复杂非线性问题时有一定优势,但从工程应用角度思考,此类优化方法仍较为复杂,易陷入局部最优解. 文献[7]在粒子群算法中引入自适应速度收敛因子改善其收敛效果,用于优化SVM参数选取,提升了分类性能;文献[4]通过蝗虫优化算法同步优化支持向量机的参数选取与封装式特征选择,显著改善了特征选择能力,但未对原算法做出改进研究,因此优化能力仍有待改善. 以上优化应用表明,元启发式智能优化在解决某一类非线性优化问题时,能够不以固定、系统的方式求解某个邻域的最优解,而是利用智能优化机制自身的探索与开采能力去搜索最佳的解决方案,增加求解的多样性和可能性,故元启发式智能优化方法更适合解决SVM参数调优的分类问题.

基于上述启发式智能优化的特点,特征选择亦可用其进行优化. 文献[8]设计了一种混合海鸥优化算法(hybrid seagull optimization algorithm, HSOA),通过引入热交换原理增强海鸥局部搜索能力,并将其应用于封装式特征选择,该方法能够有效地寻找最优特征子集,减少计算时间,提高数据分类准确率;文献[9]针对高维度数据设计了一种结合两相变异的灰狼优化算法,并用于封装式分类的特征选择中,显著改善了计算求解精度. 以上研究表明,启发式优化在封装式特征选择中的分类计算效果具有一定优势,但多数仍有优化机制繁琐复杂、内部预设参数偏多等不足,在实际应用中,这会影响算法在局部和全局搜索过程中的求解精度和计算效率. 因此,探究模型简洁、运算高效、求解精准的优化算法是封装式特征选择的重要研究方向之一.

在过去20余年中,受生物行为和自然现象启示的元启发式优化算法有效地解决了现实工程应用的复杂问题,主要包含进化类和群智能类. 进化类算法模仿自然界的进化操作,以遗传算法[10](genetic algorithm, GA)较为流行,通过模拟达尔文的生物进化过程,在解空间内搜寻最优解,具有较强的鲁棒性;群智能优化源于生物群体寻找食物或捕杀猎物的行为,以粒子群优化[7](particle swarm optimization, PSO)为经典算法,模仿鸟类群体的搜索食物机制,个体极值通过迭代计算最优解,具有良好的通用性,此类算法还包括灰狼优化算法[9](grey wolf optimizer, GWO)、鲸鱼优化算法[3](whale optimization algorithm, WOA)、斑点鬣狗优化算法[11](spotted hyena optimizer, SHO)等. 尽管两类优化算法机制不同,但搜索终值均为某域内最优解,以上各元启发式优化算法在优化中有一定优越性,但No-Free-Lunch定理[12]表明,没有一种算法可以解决所有优化问题,这促使研究人员不断探索更优秀的算法以解决不同领域的问题. 秃鹰搜索(bald eagle search, BES)优化[13]是马来西亚学者Alsattar提出的一种新型元启发式算法,该算法具有较强的全局搜索能力,能够有效地解决各类复杂数值优化问题. 对于上述支持向量机参数选取及封装式特征选择而言,均可视为最优化问题. 而元启发式智能优化算法在其中均表现出一定的优势,故本文对BES算法进行SVM参数调优及封装式特征选择方面的探索研究与应用.

本文主要研究内容如下. 首先,将莱维飞行策略与模拟退火机制引入秃鹰搜索优化算法,增强局部搜索和全局收敛能力,提出改进秃鹰搜索优化算法,通过测试函数验证改进效果;其次,利用改进算法优化支持向量机学习器,并将这种融合模型应用于封装式特征选择中,同步处理特征子集的选取和支持向量机参数调整;最后,通过标准数据集进行测试实验,对比其他最优化算法,验证改进算法在克服局部最优缺陷、提高全局搜索精度,以及解决封装式特征选择问题时的有效性.

1 背景知识

1.1 秃鹰搜索优化算法

秃鹰遍布于北美洲地区,飞行中视力敏锐、观察能力优秀. 以捕食鲑鱼为例,秃鹰首先会基于自身位

置或跟踪其他鸟群到鲑鱼附近来选择搜索空间,朝一个特定区域飞行;其次在选定搜索空间内搜索水面,直到发现合适的猎物;最后秃鹰会逐渐改变飞行高度,快速向下俯冲,从水中成功捕获鲑鱼等猎物.

BES算法以秃鹰捕食猎物的行为进行模拟,将其分为选择搜索空间、搜索空间猎物和俯冲捕获猎物3个阶段,数学模型如下所示.

1) 选择搜索空间:秃鹰随机选择搜索区域,通过判断猎物数目确定最佳搜寻位置,便于搜索猎物,该阶段秃鹰位置$P_{i,\mathrm{new}}$更新由随机搜索的先验信息乘以α来确定. 该行为数学模型描述为

$$P_{i,\mathrm{new}} = P_{\mathrm{best}} + \alpha r(P_{\mathrm{mean}} - P_i). \tag{1}$$

其中:α为控制位置变化参数,变化范围为(1.5, 2);r为(0, 1)间随机数;P_{best}为当前秃鹰搜索确定的最佳搜索位置;P_{mean}为先前搜索结束后秃鹰的平均分布位置;P_i为第i只秃鹰位置.

2) 搜索空间猎物(探索):秃鹰在选定搜索空间内以螺旋形状飞行搜索猎物,加速搜索进程,寻找最佳俯冲捕获位置. 螺旋飞行数学模型采用极坐标方程进行位置更新,如下所示:

$$\theta(i) = a \cdot \pi \cdot \mathrm{rand}; \tag{2}$$

$$r(i) = \theta(i) + R \cdot \mathrm{rand}; \tag{3}$$

$$xr(i) = r(i) \cdot \sin(\theta(i)),\ yr(i) = r(i) \cdot \cos(\theta(i)); \tag{4}$$

$$x(i) = xr(i)/\max(|xr|),\ y(i) = yr(i)/\max(|yr|). \tag{5}$$

其中:$\theta(i)$和$r(i)$分别为螺旋方程的极角和极径;a和R为控制螺旋轨迹的参数,变化范围分别为(0, 5)、(0.5, 2);rand为(0, 1)内随机数;$x(i)$和$y(i)$为极坐标中秃鹰位置,取值均为(−1,1). 秃鹰位置更新如下:

$$P_{i,\mathrm{new}} = P_i + x(i) \cdot (P_i - P_{\mathrm{mean}}) + y(i) \cdot (P_i - P_{i+1}), \tag{6}$$

其中P_{i+1}为第i只秃鹰下一次更新位置.

3) 俯冲捕获猎物(利用):秃鹰从搜索空间的最佳位置快速俯冲飞向目标猎物,种群其他个体也同时向最佳位置移动并攻击猎物,运动状态仍用极坐标方程描述,如下:

$$\theta(i) = a \cdot \pi \cdot \mathrm{rand},\ r(i) = \theta(i); \tag{7}$$

$$xr(i) = r(i) \cdot \sin(h(\theta(i))),$$
$$yr(i) = r(i) \cdot \cos(h(\theta(i))); \tag{8}$$

$$x_1(i) = xr(i)/\max(|xr|),\ y_1(i) = yr(i)/\max(|yr|). \tag{9}$$

俯冲中秃鹰位置更新公式为

$$\begin{cases} \delta_x = x_1(i) \cdot (P_i - c_1 P_{\mathrm{mean}}), \\ \delta_y = y_1(i) \cdot (P_i - c_2 P_{\mathrm{best}}). \end{cases} \tag{10}$$

$$P_{i,\mathrm{new}} = \mathrm{rand} \cdot P_{\mathrm{best}} + \delta_x + \delta_y. \tag{11}$$

其中:c_1和c_2为秃鹰向最佳和中心位置的运动强度,取值均为(1, 2).

1.2　莱维飞行策略

莱维飞行源于Levy的对称莱维稳定分布积分,是一种生成特殊的随机步长方法,飞行步长服从重尾的指数概率分布(Levy分布),其服从参数(步长)为s的分布公式,即

$$\mathrm{Levy}(s) \sim u = t^{-1-\beta},\ \beta \in (0, 2]. \tag{12}$$

文献[14]提出了一种用正态分布求解随机步长的方法,应用效果良好,公式如下:

$$s = u/|v|^{1/\beta}. \tag{13}$$

其中:$\beta = 1.5$;u、v均服从$N(0, \delta_u^2)$和$N(0, \delta_v^2)$的正态分布. 参数方差如下:

$$\delta_u = \Big[\frac{\Gamma(1+\beta) \cdot (\sin(\pi\beta/2))}{\Gamma((1+\beta)/2) \cdot \beta \cdot (2^{(\beta-1)/2})}\Big]^{1/\beta},\ \delta_v = 1, \tag{14}$$

其中Γ为标准的Gamma函数积分运算.

1.3　模拟退火机制

模拟退火[11](simulated annealing, SA)以一定的概率接受比当前解更差的解,因此可以跳出局部最优解并且获得全局最优解. 当邻域解比当前解更好时,将其认为是最新的解,否则由Boltzmann概率确定最新解概率,Boltzmann概率[11]如下:

$$P = \exp(-\theta/T). \tag{15}$$

其中:θ为最佳解与产生的邻域解的适应度之间的差异,T(温度)是在搜索过程中按一定规律周期性减小的参数.

1.4　支持向量机

文献[4]以统计学习理论为基础提出了一种支持向量机学习方法,对于给定训练样本集,找到最优划分超平面,需将此类问题转化为求解有约束的最优化,故通过以下数学理论求取最优间隔超平面.

对于线性可分样本,存在一个超平面将其完全分开,平面数学方程为

$$f(x) = \omega \cdot (x) + b = 0. \tag{16}$$

其中:x为特征向量;ω为m维法向量,决定超平面的方向;b为截距,表示超平面与原点的距离.

利用式(16)计算临界超平面,其中两个异类支持向量间隔$\gamma = 2/|\omega|$最大即为约束条件,故构造求解有约束的最优化方程

$$\begin{cases}\min\limits_{\omega,b}(1/2|\omega|^2);\\ \text{s.t. } y_i(\omega\cdot(x)+b)\geqslant 1,\ i=1,2,\ldots,m.\end{cases}\tag{17}$$

所以,求解使式(17)成立的ω、b就是最优超平面.

对于线性不可分样本,SVM采取非线性函数$\varphi(x)$将数据从原始空间映射到高维的特征空间中使其线性可分. 假设$\varphi(x)$是将x映射后的特征向量,则样本特征空间中的超平面数学方程为

$$f(x)=\omega\cdot\varphi(x)+b=0.\tag{18}$$

引入拉格朗日乘子求对偶问题的最优目标函数为

$$\max_{\alpha}\ \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\varphi(x_i)^{\mathrm{T}}\varphi(x_j),\tag{19}$$

使得$0\leqslant\alpha_i\leqslant C,\sum\limits_{i=1}^{m}\alpha_i y_i=0$. 其中:$\alpha$为约束中添加的拉格朗日乘子,$C$为惩罚因子. 因高维空间中$\varphi(x_i)^{\mathrm{T}}\varphi(x_j)$内积求取困难,引入核函数$K(x_i,x_j)$,构造有约束最优化方程

$$\begin{cases}\max\limits_{\alpha}\ \sum\limits_{i=1}^{m}\alpha_i-\frac{1}{2}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j);\\ \text{s.t. } 0\leqslant\alpha_i\leqslant C,\ \sum\limits_{i=1}^{m}\alpha_i y_i=0.\end{cases}\tag{20}$$

求得最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots,\alpha_m^*)^{\mathrm{T}}$,从中选择$\alpha_i^*$计算$b^*$和$f(x)$分类决策函数即可完成分类.

SVM中常用核函数有5种:线性核函数、多项式核函数、径向基核函数、高斯核函数以及Sigmoid核函数[4]. 本文对高斯核函数进行优化,表达式如下:

$$K(x_i,x_j)=\exp(-|x_i-x_j|^2/g^2),\tag{21}$$

其中g为核参数.

2 改进秃鹰搜索算法优化SVM

在原BES优化算法中,秃鹰的狩猎过程可划分为选择搜索空间、在搜索空间中搜查待捕获猎物个体、快速俯冲捕获猎物3个过程. 而在元启发式智能优化算法中,算法的局部和全局搜索能力可作为优化效果是否良好的评价依据. 本文针对前两阶段优化机制做出相关改进研究.

2.1 改进秃鹰搜索算法

文献[13]在原BES算法中指出,秃鹰选择搜索空间利用上个阶段的可用信息确定下个搜索区域,在随机选择另一个搜索区域时,会依据之前的搜索域确定相应的邻域,进而完成秃鹰第1阶段的搜索过程,因此针对秃鹰选择搜索空间而言,如果秃鹰群体陷入局部最优状态,则在全局的搜索寻优过程中便无法准确捕获猎物,即无法在某类特定优化问题中求得最优解,极大降低了启发式智能优化算法的最优化效果. 故基于以上分析,可以增强搜索邻域的局部搜索能力来提高秃鹰全局的搜索优化能力,进而综合改善BES的最优化能力. 具体改进如下.

改进1:秃鹰搜索优化在选择搜索空间处理复杂计算时,控制位置变化参数α和随机参数r易使种群选择搜索空间解时过早收敛,陷入局部最优状态. 文献[14]在磷虾群算法中引入莱维飞行策略,扩大了群体的搜索范围,使算法能够及时跳出局部最优点,以改进原算法易陷入局部极值和搜索效率低的不足,故莱维飞行策略用于启发式群智能优化算法中效果显著. 因此,为增强秃鹰个体间的信息交流,改善全局最优化,本文将莱维飞行策略引入秃鹰搜索优化中,并作用于参数α、r部分,使原算法改善收敛效果,跳出局部最优. 引入后式(1)改为

$$P_{i,\mathrm{new}}=P_{\mathrm{best}}+\alpha\cdot r(P_{\mathrm{mean}}-P_i)\times\mathrm{Levy}.\tag{22}$$

元启发式群智能优化算法中,求解空间的探索开发与开采利用是两个重要过程[3],如何能够有效地在两者间平衡转换直接决定了优化求解的能力. 在BES算法中,一旦秃鹰选定好搜索空间,完成最佳求解空间探索开发,便会进入搜索猎物阶段的开采利用,秃鹰通过独特的螺旋飞行搜索方式寻找待捕获猎物,这种方式增加了搜索遍历的多样化,以进一步跳出局部最优,持续不断地在解空间的利用中获取有效解. 因此,对解空间的探索开发与开采利用过渡而言,若能增强秃鹰的解空间局域搜索能力,则在选定解空间内能够跳出局部极值点,增加获得最优解的概率,找到最佳猎物位置,从而解决最优化问题.

对于搜索猎物阶段的开采利用不足所做改进如下.

改进2:在极坐标螺旋式搜索最优空间猎物时,秃鹰可探索并利用新的搜索空间解. 模拟退火具有强大的局部搜索能力,广泛应用于工程优化. 文献[11]利用模拟退火对斑点鬣狗算法进行混合优化,改善其局部寻优能力,平衡算法的解空间探索与利用过程,并将混合算法用于特征选择中,表现良好. 这表明此算法机制可以提高解空间局部邻域的利用开采过程,适合最优化问题的求解. 故针对BES算法的改进研

究而言,为提高秃鹰在既定空间内搜索猎物能力,全面探索并利用解空间,将秃鹰在该阶段所得种群位置解作为模拟退火初始解,累次迭代寻优,增强原算法在局部邻域内求解精度,找到搜索空间最优解. 其中模拟退火可视为秃鹰搜索内优化算子,使秃鹰能够更精准俯冲捕获猎物,改善全局求解质量.

基于以上关于初始搜索空间选择避免局部最优以及搜索解空间最值改善算法开采利用寻优能力的两处改进,将新算法命名为IBES(improved bald eagle search).

2.2　IBES算法优化SVM的特征选择

支持向量机在工程应用领域具有强大的学习能力. 对于分类而言,其性能主要取决于核的类型和约束参数. 惩罚因子C与核参数g是影响高斯核函数映射分类的重要参数,其中C与分类容错性成反比,该值过高可导致过拟合现象,降低算法泛化能力,过低则会产生欠拟合效果,错分样本数据;参数g控制核函数宽度,选择不当仍会造成不良分类效果. 因此,若用IBES算法同步优化SVM参数g和C的选择,则可提高SVM的性能和效率,更准确地对数据进行分类计算.

基于封装式的特征选择方法结合优化算法可提高数据分类的精度和效率,这种特征选择方法主要包括3个部分:搜索方法、归纳算法、评价措施[4]. 本文中,将改进的秃鹰搜索算法作为搜索方法,用来选择最佳特征子集;SVM可作为归纳算法,提高准确率和学习效率;采用分类准确率作为评价措施,用来评价特征子集选择的效果. 因此,将IBES算法同步优化支持向量机参数和特征选择,能够更好地解决特征选择问题,故将这种新模型命名为IBES_SVM.

2.3　种群个体与适应度函数设计

在元启发式群智能算法优化计算时,对于不同的优化问题,种群个体代表不同含义. 对于特征选择问题而言,其实质是二元优化问题,优化后的选择特征结果表示仅限于“0”和“1”,值“0”表示未选择该特征,值“1”表示选择该特征. 利用IBES算法同时优化SVM参数和特征选择时,输入的种群搜索个体设计分为两部分[4],前部分为优化SVM时的核参数g和惩罚因子C,后部分为原始数据集中的特征. 优化选择特征时,种群的个体解可视为一维向量,每个维度的原始数据值与0.5比较,大于等于0.5则选择该特征,否则剔除该特征.

特征选择可视为多个目标优化问题,当分类结果中分类准确率较高,选择特征子集个数较少时,说明所得分类效果优秀. 在算法迭代过程中,一般采用适应度函数评估每个解的质量. 改进秃鹰搜索优化算法和支持向量机可平衡分类准确率和特征子集个数这两个指标. 因此,根据SVM分类器所得的解的分类准确率和特征选择的所选特征子集个数,设计适应度函数[11]如下:

$$\mathrm{Fitness} = \alpha \cdot \mathrm{acc} + \beta(1 - R/N). \tag{23}$$

其中:acc为正确分类率;R和N分别为特征选择所选特征子集个数和原始数据集特征总数;参数$\alpha \in (0,1)$为分类精确性,本文α取0.99;参数$\beta = 1 - \alpha$为所选特征重要性.

通过上述设计,本文提出的基于IBES_SVM模型特征选择流程如图1所示.

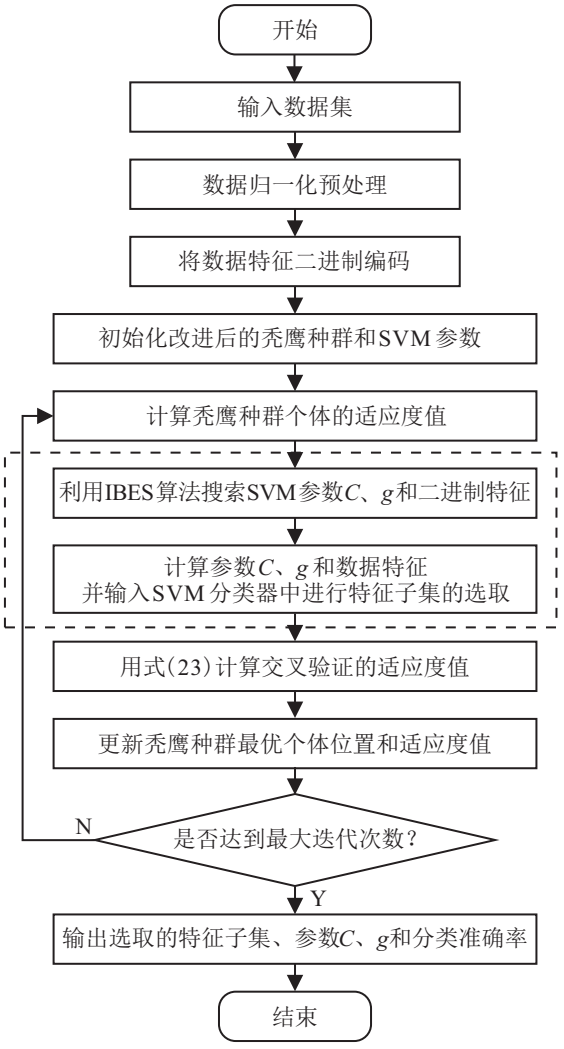

图1　基于IBES_SVM的特征选择流程

3　实验与分析

3.1　标准测试函数实验

为验证本文所提算法的优化性能,对原始算法与改进算法进行标准函数[8]仿真测试实验,同时对

比其他4种优化算法性能,包括PSO、GWO、SHO、HSOA. 其中:函数1～函数5为单峰值,可测试算法的收敛效果;函数6～函数10为多峰值,可测试算法的局部寻优和搜索能力. 测试算法的优化性能基于函数所寻找的全局最优解,粒子种群大小设为30,迭代次数设为500,每种算法运行30次,运行环境为Matlab 2016a,采用适应度平均值mean和标准差值std评价算法的优化效果.

基于10个标准测试函数实验测试结果如表1所示. 由仿真实验数据可知:在平均值上,所提出的IBES算法在4个单峰值函数中均可搜索到理论最优解,表明该算法具有良好的收敛效果和较高的求解精度. 对于多峰值函数而言:所提IBES算法在计算搜索过程中收敛效果良好且寻优能力较强,可找到全局最优解;在适应度标准差中,无论是单峰值函数还是多峰值函数,IBES算法求解效果较其他算法更稳定. 综上所述,本文IBES算法比原始算法及其他4种算法具有更好的优化效果.

表1 IBES算法与其他优化算法的测试函数结果

测试函数		PSO	GWO	SHO	HSOA	BES	IBES
1	mean	5.68e-22	1.12e-59	3.59e-38	2.92e-03	**0.00e+00**	**0.00e+00**
	std	1.04e-22	2.04e-60	6.56e-39	5.39e-03	**0.00e+00**	**0.00e+00**
2	mean	8.59e-12	3.35e-33	1.80e-14	1.01e-02	2.48e-197	**0.00e+00**
	std	1.11e-11	6.48e-33	9.63e-14	3.06e-02	**0.00e+00**	**0.00e+00**
3	mean	2.30e-06	1.24e-24	5.23e+01	5.00e+02	5.06e-10	**0.00e+00**
	std	3.00e-06	3.98e-24	4.29e+01	6.02e+02	2.77e-09	**0.00e+00**
4	mean	1.40e-05	6.05e-18	3.37e-01	1.14e+00	9.75e-197	**0.00e+00**
	std	1.46e-05	1.74e-17	6.58e-01	1.16e+00	**0.00e+00**	**0.00e+00**
5	mean	1.01e+02	2.70e+01	2.53e+01	4.94e+00	6.16e+00	**1.53e+00**
	std	9.76e+01	**6.50e-01**	9.09e+00	4.96e+00	1.09e+01	2.15e+00
6	mean	−2.25e+03	−2.65e+03	−3.94e+03	**−4.18e+03**	−1.88e+03	−3.59e+03
	std	4.62e+02	**3.06e+02**	4.39e+02	1.70e+03	8.22e+02	5.62e+02
7	mean	4.61e+00	9.52e-01	5.33e-02	5.27e-03	**0.00e+00**	**0.00e+00**
	std	2.15e+00	2.07e+00	2.92e-01	3.10e-03	**0.00e+00**	**0.00e+00**
8	mean	6.17e-11	7.64e-15	3.41e-09	4.18e-03	1.13e-15	**8.88e-16**
	std	9.79e-11	2.16e-15	1.87e-08	7.08e-03	9.01e-16	**0.00e+00**
9	mean	2.01e-01	2.74e-02	1.10e-09	9.88e-01	**0.00e+00**	**0.00e+00**
	std	1.54e-01	2.91e-02	6.02e-09	1.80e-01	**0.00e+00**	**0.00e+00**
10	mean	3.23e+00	3.25e+00	2.18e+00	3.46e+00	2.45e+00	**1.48e+00**
	std	2.45e+00	3.69e+00	2.24e+00	3.65e+00	1.76e+00	**9.30e-01**

3.2 特征选择实验

为验证IBES_SVM的特征选择效果,本文采取UCI[15]数据存储库中的12个数据集进行测试,每个数据集的特征个数、样本个数和类别数目如表2所示.

实验前数据集需作预处理以改善特征选择的结果,将所有特征值归一化在0～1之间,最后依据种群个体的二进制设计实验. 归一化后的特征值如下:

$$F_{\text{norm}} = (F - F_{\min})/(F_{\max} - F_{\min}). \quad (24)$$

其中:F_{norm}为归一化后特征;F为原始特征;$F_{\max}$和$F_{\min}$为特征取最大值、最小值.

表2 实验数据集列表

序号	数据集	样本数	特征数	类别数
1	Iris	150	4	3
2	ILPD	583	10	2
3	Zoo	101	16	7
4	Flags	194	30	8
5	Soybean	307	35	19
6	Connectionist	208	60	2
7	Urban	168	148	9
8	MUSK	476	168	2
9	SCADI	70	206	7
10	LVST-voice	126	309	2
11	DMEAV	373	513	2
12	Micromass	360	1 300	10

在特征选择实验中,本文将种群大小和最大迭代值设为30和100. 通过交叉实验能够验证分类器的性能,将数据集合理划分可以评估算法的有效性. K 折交叉验证能有效避免过学习以及欠学习状态的发生,因此实验采取基于SVM分类器的 K 折交叉验证来评价算法的性能,其中用于训练和验证的子集数为 $K-1$,测试子集数为1,重复 M 次,则每种算法对每个数据集进行 $K\cdot M$ 次评估,采取LIBSVM工具箱和Matlab 2016a仿真.

特征选择共设计3部分实验: 第1部分为原始算法和改进算法的特征选择实验; 第2部分为对比两种非线性最优化方法(SGD、QN-MKL)的特征选择实验; 第3部分为对比其他分类算法朴素贝叶斯(naive bayes, NB)、K 近邻(k-nearest neighbor, KNN)、决策树(decision tree,DT)、传统SVM数据分类实验. 为验证本文提出的方法在特征选择过程中的优化效果,采用以下指标[11]评估算法的性能(M 均为算法运行次数).

平均分类准确率: 描述所选特征集合分类准确率的平均值,计算如下:

$$\text{mean}=\frac{1}{M}\sum_{i=1}^{M}\text{acc}(i), \tag{25}$$

其中 $\text{acc}(i)$ 为第 i 次分类准确率.

平均选择特征个数: 表示特征选择所选特征子集个数的平均值,如下所示:

$$\text{mean}=\frac{1}{M}\sum_{i=1}^{M}\text{Size}(i), \tag{26}$$

其中 $\text{Size}(i)$ 为第 i 次运行选择的特征个数.

适应度平均值mean: 表示运行算法时,计算所得解的适应度平均值,计算公式如下:

$$\text{mean}=\frac{1}{M}\sum_{i=1}^{M}\text{Fitness}(i), \tag{27}$$

其中 $\text{Fitness}(i)$ 为算法第 i 次运行适应度.

适应度标准差std: 表示在运行算法后得到的最优解的变化情况,计算公式为

$$\text{std}=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(\text{Fitness}(i)-\text{mean})^2}. \tag{28}$$

3.2.1　原始算法和改进算法的特征选择实验

基于原始和改进算法在平均选择特征个数和平均选择特征个数上的实验结果如图2和图3所示.

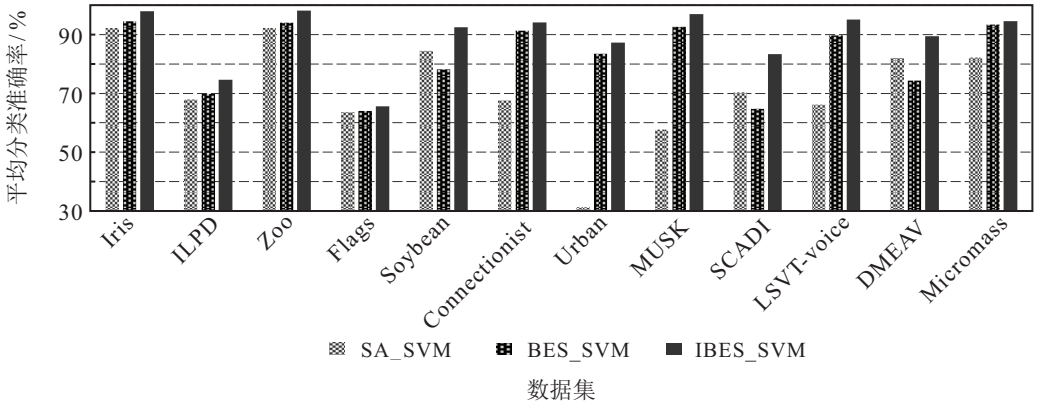

图 2　原始和改进算法在平均分类准确率上的测试结果

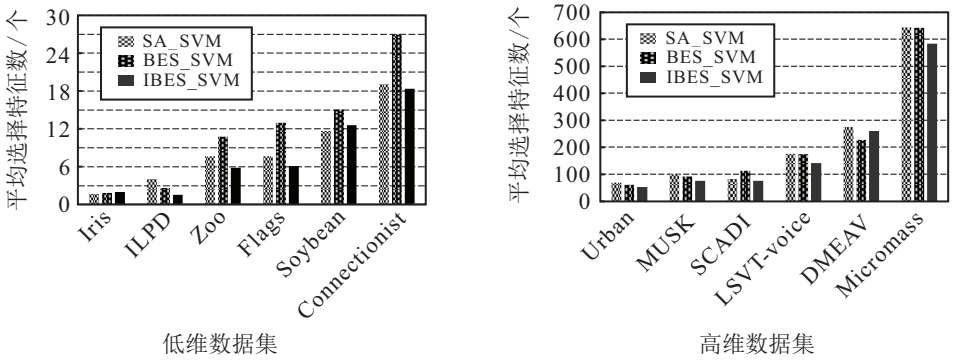

图 3　原始和改进算法在平均选择特征个数上的测试结果

图2和图3数据表明: 在分类准确率上,改进算法可对数据集进行更准确地分类; 在选择特征数上,改进算法可对不同维度的数据子集选取更少特征. 因此,IBES算法在特征选择中表现良好,验证了所提算

法有效. 基于原始和改进算法在特征选择中的适应度(平均值、标准差)实验结果如表3所示.

表3数据显示: 除个别数据集外, 所提出的改进算法在用于特征选择时, 适应度(mean、std)值上表现良好, 因此改进算法比原始算法具有更好的运行稳定性.

综上所述, 仿真实验测试结果表明: 在特征选择中, 本文改进算法有效且较原始算法表现优越.

表3 原始和改进算法的适应度(平均值、标准差)值测试结果

数据集	适应度值					
	SA_SVM (mean ± std)		BES_SVM (mean ± std)		IBES_SVM (mean ± std)	
Iris	0.914 2	0.019 5	0.950 6	0.013 4	**0.972 5**	**0.004 4**
ILPD	0.676 8	0.033 5	0.695 4	0.0211	**0.743 3**	**0.009 0**
Zoo	0.918 3	0.016 2	0.935 0	0.018 3	**0.977 4**	**0.012 1**
Flags	0.637 5	0.029 3	0.644 8	0.029 1	**0.651 3**	**0.016 4**
Soybean	0.843 1	0.030 3	0.778 5	0.186 0	**0.920 7**	**0.003 0**
Connectionist	0.676 4	0.113 0	0.910 8	0.023 2	**0.936 3**	**0.014 5**
Urban	0.315 1	0.287 0	0.832 5	**0.010 9**	**0.868 2**	0.011 8
MUSK	0.572 2	0.024 5	0.919 8	0.0172	**0.968 6**	**0.009 9**
SCADI	0.703 1	0.054 9	0.646 5	0.208 4	**0.828 9**	**0.046 4**
LSVT-voice	0.659 0	0.026 7	0.890 4	0.056 3	**0.948 3**	**0.006 6**
DMEAV	0.813 1	0.029 2	0.737 2	0.093 6	**0.888 4**	**0.014 6**
MicroMass	0.818 8	0.095 4	0.929 7	**0.019 7**	**0.941 1**	0.042 0

3.2.2 改进算法和最优化算法的特征选择实验

在前文中, 所提改进算法在特征选择中表现良好, 为能够更好地验证IBES_SVM方法用于特征选择的有效性, 本节选择其他两种非线性最优化(SGD[5]、QN-MKL[6])算法结合支持向量机进行特征选择对比, 所用评价标准与前文实验相同. 基于IBES_SVM算法和SGD_SVM、QN-MKL_SVM最优化算法在封装式特征选择中的适应度值结果如表4所示.

表4 IBES与SGD、QN-MKL结合SVM的适应度(平均值、标准差)值测试结果

数据集	适应度值					
	SGD_SVM(mean ± std)		QN-MKL_SVM(mean ± std)		IBES_SVM(mean ± std)	
Iris	0.966 3	0.006 9	0.912 6	0.018 9	**0.972 5**	**0.004 4**
ILPD	0.735 2	0.027 3	0.709 2	0.060 5	**0.743 3**	**0.009 0**
Zoo	0.952 5	0.016 9	0.922 3	0.013 9	**0.977 4**	**0.012 1**
Flags	**0.731 2**	0.030 7	0.698 5	0.048 8	0.651 3	**0.016 4**
Soybean	0.809 4	0.017 6	0.838 6	0.071 6	**0.9207**	**0.003 0**
Connectionist	0.812 9	0.145 3	0.883 4	0.040 3	**0.936 3**	**0.014 5**
Urban	0.757 6	0.053 0	0.830 5	**0.009 9**	**0.868 2**	0.011 8
MUSK	0.852 1	0.050 6	0.887 7	0.048 5	**0.968 6**	**0.009 9**
SCADI	0.762 4	0.173 8	**0.896 6**	0.102 5	0.828 9	**0.046 4**
LSVT-voice	0.863 2	0.090 9	0.891 3	0.006 7	**0.948 3**	**0.006 6**
DMEAV	0.751 1	**0.010 2**	0.671 6	0.013 9	**0.888 4**	0.014 6
MicroMass	0.795 7	0.112 1	0.742 9	0.136 7	**0.941 1**	**0.042 0**

从适应度平均值可以看出: 本文所提的算法能够在绝大多数低、高维数据集中表现良好, 平均提高了8.2%的适应度值; 在该值的标准差上, 虽有个别数值表现不佳, 但就整体而言, IBES_SVM封装式特征选择方法仍具有良好的稳定性.

图4和图5分别为本文所提算法与SGD、QN-MKL结合SVM封装式特征选择后的平均分类准确率和选择特征个数结果.

在平均分类准确率上, IBES_SVM平均提升了8个百分点, 这表明该方法能够针对不同维度的数据进行有效分类; 同时在高、低维数据选择特征子集的结果中, IBES_SVM较另外两种最优化方法平均减少了13.75个特征, 有效地降低了数据维度. 因此, 基于以上分析可知, 所提的改进方法能够平衡分类准确率和所选特征子集, 较数学类最优化方法更适合解决封装式特征选择问题.

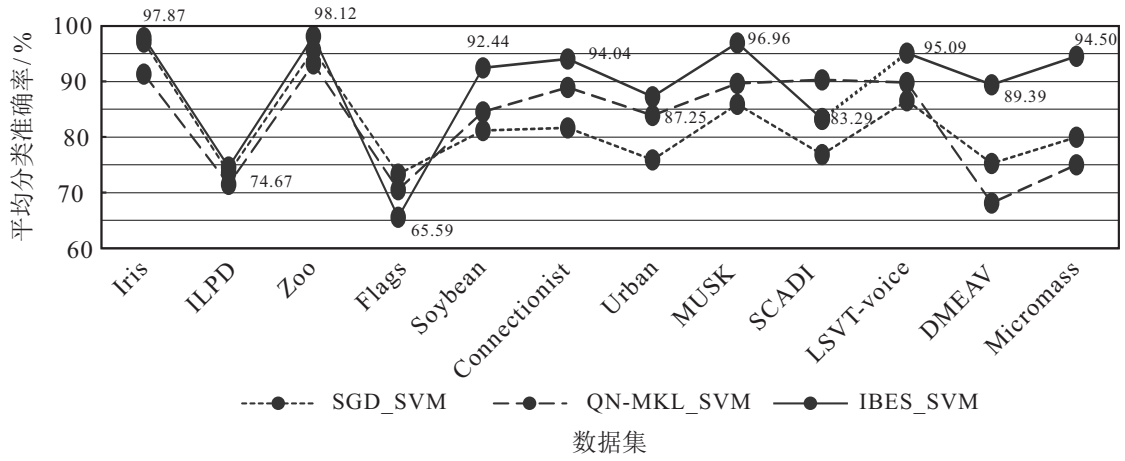

图 4　IBES与SGD、QN-MKL结合SVM在平均分类准确率上的测试结果

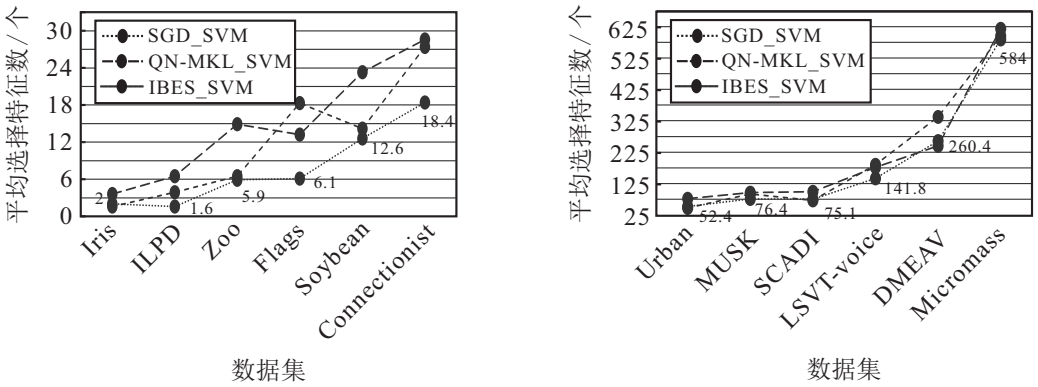

图 5　IBES与SGD、QN-MKL结合SVM在平均选择特征个数上的测试结果

3.2.3　改进算法与其他分类算法对比实验

为进一步验证本文提出的IBES算法优化支持向量机在特征选择中的分类效果,将目前机器学习领域中常用的分类算法NB、KNN、DT、传统SVM算法与本文方法在前文的12个数据集上进行数据分类对比. 基于平均分类准确率评价的实验测试数据如表5所示.

表 5　改进算法与其他分类算法平均分类准确率的比较

数据集	平均分类准确率/%				
	NB	KNN	DT	SVM	IBES_SVM
Iris	88.33	33.33	96.59	96.57	**97.87**
ILPD	53.91	71.61	**78.79**	74.83	74.67
Zoo	48.60	39.25	68.10	77.90	**98.12**
Flags	60.23	29.60	57.72	57.57	**65.59**
Soybean	56.67	3.29	34.63	76.72	**92.44**
Connectionist	65.00	83.04	89.37	56.05	**94.04**
Urban	80.15	11.11	**88.80**	76.13	87.25
MUSK	75.11	88.86	88.37	59.71	**96.96**
SCADI	67.14	2.85	41.42	74.69	**83.29**
LSVT-voice	51.67	33.59	89.77	82.75	**95.09**
DMEAV	55.06	32.44	69.24	34.55	**89.39**
MicroMass	37.75	9.86	10.06	83.38	**94.50**

表5结果表明:在绝大多数的数据集中,本文所提的IBES_SVM方法准确率高,尤其在高维度、多类别数据上,明显优于其他经典算法,因此本文提出的方法能够有效完成数据分类任务.

4　结　论

针对机器学习中的数据预处理技术,本文提出了一种基于改进秃鹰搜索算法同步优化的特征选择方法,通过多个仿真实验验证了所提方法的性能. 首先,与原始SA、BES算法,其他非线性最优化的SGD、QN-MKL算法验证进行对比,实验结果表明,在不同维度和容量的样本数据中,所提改进算法能够有效提升分类准确率,降低选择数据特征维度;其次,与经典的NB、KNN、DT、SVM算法进行分类比较,所提方法分类更准确. 如何进一步改善特征选择能力,提升算法在更高容量和维度的数据中仍具有准确分类效果,将是未来研究的主要内容.

参考文献(References)

[1]　姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-166.
(Yao X, Wang X D, Zhang Y X, et al. Summary of

feature selection algorithms[J]. Control and Decision, 2012, 27(2): 161-166.)

[2] Lyu H Q, Wan M X, Han J Q, et al. A filter feature selection method based on the maximal information coefficient and gram-schmidt orthogonalization for biomedical data mining[J]. Computers in Biology and Medicine, 2017, 89: 264-274.

[3] Mafarja M M, Mirjalili S. Whale optimization approaches for wrapper feature selection[J]. Applied Soft Computing, 2017, 62: 441-453.

[4] Aljarah I, Al-Zoubi A M, Faris H, et al. Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm[J]. Cognitive Computation, 2018, 10(3): 478-495.

[5] Bottou L. Large-scale machine learning with stochastic gradient descent[C]. Proceeding of the 19th International Conference on Computational Statistics. Paris: Physica-Verlag HD, 2010: 177-186.

[6] 胡庆辉, 丁立新, 刘晓刚, 等. 基于原问题求解的非稀疏多核学习方法[J]. 华南理工大学学报: 自然科学版, 2015, 43(5): 78-85.
(Hu Q H, Ding L X, Liu X G, et al. A non-sparse multi-kernel learning method based on primal problem[J]. Journal of South China University of Technology: Natural Science Edition, 2015, 43(5): 78-85.)

[7] Liu J, Liu Z, Xiong Y. Method of parameters optimization in SVM based on PSO[J]. Transactions on Computer Science & Technology, 2013, 2(1): 9-16.

[8] Jia H M, Xing Z K, Song W L. A new hybrid seagull optimization algorithm for feature selection[J]. IEEE Access, 2019, 7: 49614-49631.

[9] Abdel-Basset M, El-Shahat D, El-Henawy I, et al. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection[J]. Expert Systems with Applications, 2020, 139: 112824.1-112824.14.

[10] Li S J, Wu H, Wan D S, et al. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine[J]. Knowledge-Based Systems, 2011, 24(1): 40-48.

[11] Jia H M, Li J D, Song W L, et al. Spotted hyena optimization algorithm with simulated annealing for feature selection[J]. IEEE Access, 2019, 7: 71943-71962.

[12] Alabert A, Berti A, Caballero R, et al. No-free-lunch theorems in the continuum[J]. Theoretical Computer Science, 2015, 600: 98-106.

[13] Alsattar H A, Zaidan A A, Zaidan B B. Novel meta-heuristic bald eagle search optimisation algorithm[J]. Artificial Intelligence Review: An International Science and Engineering Journal, 2020, 53(3): 2237-2264.

[14] 王秋萍, 丁成, 王晓峰. 一种基于改进KH与KHM聚类的混合数据聚类算法[J]. 控制与决策, 2020, 35(10): 2449-2458.
(Wang Q P, Ding C, Wang X F. A hybrid data clustering algorithm based on improved krill herd algorithm and KHM clustering[J]. Control and Decision, 2020, 35(10): 2449-2458.)

[15] Blake C L. UCI machine learning repository[EB/OL]. (2013-12-23) [2019-12-17]. http://archive.ics.uci.edu/ml.

作者简介

贾鹤鸣 (1983−), 男, 教授, 博士, 从事群智能优化、特征选择、多阈值图像分割等研究, E-mail: jiaheminglucky99@126.com;

姜子超 (1995−), 男, 硕士生, 从事机器学习、群智能优化、特征选择的研究, E-mail: jiangzichao@nefu.edu.cn;

李瑶 (1997−), 女, 硕士生, 从事群智能优化、特征选择的研究, E-mail: liyao@nefu.edu.cn.

(责任编辑：闫 妍)