计算机科学
COMPUTER SCIENCE

改进的森林优化特征选择算法在信用评估中的应用

黄宇航，宋友，王宝会

引用本文

黄宇航，宋友，王宝会.改进的森林优化特征选择算法在信用评估中的应用[J]．计算机科学, 2023, 50(6A): 220600241-6.

HUANG Yuhang, SONG You, WANG Baohui. Improved Forest Optimization Feature Selection Algorithm for Credit Evaluation [J]. Computer Science, 2023, 50(6A): 220600241-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600241

改进的森林优化特征选择算法在信用评估中的应用

黄宇航　宋　友　王宝会
北京航空航天大学软件学院　北京 100191
(ITcathyh@buaa.edu.cn)

摘　要　信用评估是金融领域的一个关键问题,它可以预测出一个用户是否存在拖欠风险,从而减少坏账损失。信用评估的关键挑战之一就是数据集存在着大量无效或冗余特征。为了解决该问题,提出了一种改进的森林优化特征选择算法(Improved Feature Selection using Forest Optimization Algorithm,IFSFOA)。该算法针对原始算法 FSFOA 的不足,在初始化阶段使用基于卡方校验的初始化策略代替随机化初始,提升算法寻优的能力;在局部播种阶段利用多层级变异策略,优化局部搜索能力,解决 FSFOA 的搜索空间受限和局部性问题;在更新候选森林时,使用贪婪选取策略挑选优质树,淘汰劣质树,收敛搜索发散过程。最后在涵盖了低维、中维和高维的公开信用评估数据集上设置对比实验,结果表明 IFSFOA 在分类和维度缩减方面的能力的综合表现均优于 FSFOA 和近年提出的较为高效的特征选择算法,验证了 IFSFOA 的有效性。

关键词:森林优化算法;特征选择;信用评估;演化计算;包裹式方法
中图法分类号　TP3-05

Improved Forest Optimization Feature Selection Algorithm for Credit Evaluation

HUANG Yuhang,SONG You and WANG Baohui
College of Software,Beihang University,Beijing 100191,China

Abstract　Credit evaluation is a key problem in finance,which predicts whether a user is at risk of defaulting and thus reduces bad debt losses. One of the key challenges in credit evaluation is the presence of a large number of invalid or redundant features in the dataset. To solve this problem,an improved feature selection using forest optimization algorithm(IFSFOA) is proposed. It addresses the shortcomings of the original algorithm FSFOA by using a cardinality check-based initialization strategy instead of randomized initialization in the initialization phase to improve the algorithm's search capability;using a multi-level variation strategy in the local seeding phase to optimize the local search capability and solve the problems of restricted search space and localization of FSFOA;using a greedy selection strategy to select high-quality trees and eliminate low-quality trees when updating the candidate forest. In updating the candidate forest,we use the greedy selection strategy to select high-quality trees and eliminate low-quality trees,and converge the search dispersion process. Finally,the results show that IFSFOA outperforms FSFOA and more efficient feature selection algorithms proposed in recent years in terms of classification ability and dimension reduction ability,and validates the effectiveness of IFSFOA by setting up comparison experiments on public credit evaluation datasets covering low,medium and high dimensions.

Keywords　Forest optimization algorithm,Feature selection,Credit evaluation,Evolutionary computation,Wrapper methods

1　引言

近年来,随着科技的进步与经济的发展,人们的消费水平迅速提高,消费理念也逐渐升级,信用卡逐渐成为人们日常生活中重要的信贷产品之一。随之而来的是愈发严重的信用风险问题:2021 中国信用卡逾期未偿还金额高达 1 000 亿,逾期比例仍持续高涨。信用评估作为风险控制的主要工具,能够有效地应对信用风险问题。在实际业务中,客户数据通常具有无效与冗余特征多和特征分布不均两个特点,未经处理的数据不仅会带来额外的计算成本,还会劣化学习任务的效果。特征选择的主要目标是选择能够更好地了解分类模型的潜在特征的最佳组合,剔除冗余和无效特征,降低学习难度,提高预测精度,避免维度诅咒的发生[1-4]。

特征选择方法主要分为 3 类:过滤式、嵌入式和包裹式。过滤式方法不依赖于学习算法,它根据每个特征的特点进行评分与排序,这类方法执行简单、快速,但最终的选择结果通常难以获得令人满意的分类精度[5-7]。嵌入式方法将特征选择过程与学习算法耦合,在训练过程中完成特征的选择,分类精度较过滤式方法高,但适用场景较为受限[8-9]。包裹式方法依赖于预先定义好的学习算法来评估特征子集,能够针对不同的选择问题挑选不同的评价方案,在获得较高分类精度的同时,提升算法的灵活性[10-11]。因此,包裹式方法在特征选择工作中得到了广泛应用[12-13]。

其中演化计算由于具有高效和泛化能力强的特点,也多被应用在特征选择问题上[14-19]。例如:Babatunde 等提出利用改进的遗传算法开展特征选择[20];Hafez 等参考正余弦

通信作者:王宝会(wangbh@buaa.edu.cn)

优化的原理，将其应用在特征选择问题[21]；Tubishat 等将鲸鱼优化算法应用在如特征选择的单目标问题上[22]；Ghaemi 等则提出了森林优化特征选择算法(Feature Selection using Forest Optimization Algorithm，FSFOA)[23]；Too 等模拟哈里斯鹰的捕猎过程，提出了基于哈里斯鹰算法的特征选择方案 NHHOFS[24]；Hegazy 等给出了基于樽海鞘群优化算法的特征选择算法 SCAFS[25]。

在实践中，FSFOA 在分类精确度和维度缩减率上表现出色，仅需较小的计算代价就能完成比较充分的特征空间搜索，并能保证一定的泛化能力[23]。但是 FSFOA 也存在一定的不足：1)在初始化森林阶段，FSFOA 采用随机选取策略，存在一定的盲目性，影响了后续整体搜索过程；2)在局部播种阶段，FSFOA 每次只对单特征进行状态取反，因而限制了搜索范围的扩散，降低了搜索效率；3)在规模限制阶段，FSFOA 将所有被淘汰的树都加入候选森林后用以全局播种，这样不仅会提高计算成本，还会重复使用劣质树，使得算法容易陷入局部解。

基于卡方校验的特征筛选对不平衡数据选取特征子集具有良好的指导作用[26-27]；同时，改进的遗传算法中提出了通过多特征全变异的更新规则，能够有效拓展搜索空间，提高特征子集的选择能力[20]。基于以上思想，本文提出了一种改进的森林优化特征选择算法(IFSFOA)，在算法训练初始化、局部播种和规模限制 3 个方面进行改进，在提高算法分类性能的同时，选取维度更小的特征子集。在公开数据集上进行对比实验，实验表明 IFSFOA 在分类精度、维度缩减率和搜索效率方面均有更优异的表现。

2 森林优化特征选择算法

FSFOA 使用树来表征每个特征子集，每棵树使用 0/1 序列记录特征的选择状态：0 代表该特征不参与后续的学习任务，1 代表该特征参与后续的学习任务。FSFOA 主要由 5 个阶段组成：森林初始化、局部播种、规模限制、全局播种和更新最优树。

初始化森林：生成 N 颗树构成一个森林，每颗树由特征选择状态、年龄(Age)和适应度(Fitness)组成，其中特征选择状态是随机生成的 0/1 序列，年龄置为 0，对选中特征计算适应度。

局部播种：仅森林中年龄为 0 的树会参与播种。对于每颗参与播种的树，首先复制出 LSC 棵树，并将复制出的树年龄置为 0，随机选取一个特征状态进行取反。然后将新树放入森林中，将森林中的树年龄加 1。

规模限制：FSFOA 通过年龄上限(life time)和区域上限(area limit)两种方式来限制森林的规模。首先将年龄大于年龄上限的树放入候选森林，如果森林中树的数量仍超过区域上限，则会按适应度从大到小排序，选出 area limit 棵树留在森林，剩余的树移入候选森林。

全局播种：根据参数转移率(transfer rate)从森林中随机挑选一定比例的树并复制，将复制出的每棵树随机选取 GSC 个特征状态进行取反生成新树，将新树加入森林。

更新最优树：将森林中适应度最大的树作为最优树，将其年龄置为 0，并将其重新放回森林，用于下一轮的局部播种。

3 改进的森林优化特征选择算法

本文对 FSFOA 存在的不足做了分析，提出了对应的改进策略，给出了一个更有效的特征选择算法 IFSFOA，在提高算法分类能力的同时，优化维度缩减能力。IFSFOA 在森林初始化阶段采用基于卡方校验的初始化策略代替随机选取策略，辅助演化过程有效求解；在局部播种阶段使用多层级变异策略，提高搜索能力；在规模限制阶段引入贪心思想，收敛搜索范围，提升最优求解能力。

3.1 基于卡方校验的初始化

卡方检验用于描述两个事件的独立性或描述实际观察到的值与期望值的偏差程度，较高的值表明实际观察到的值与预期值更偏离，且表明这两个事件彼此之间并不完全独立。卡方校验的主要步骤为计算观察值与理论值的偏差，该计算方式如式(1)所示：

$$x^2=\sum_{i=1}^{k}(A_i-E_i)^2/E_i \tag{1}$$

其中，k 为观察值的数量，A 为观察值，E 为理论值，通过式(1)能够计算出统计量。x^2 越接近 0，代表计算偏差越小，假设成立的概率越高；反之 x^2 越大，代表偏差值越大，假设成立的概率越低，即代表该特征与预测目标的关联性越强。

IFSFOA 提出了一种基于卡方校验的初始化策略。初始化策略分为两步：第一步，计算数据所有特征与标签之间的卡方值，并根据卡方值对特征进行逆序排序；第二步，为每颗初始化的树从第一步排序后的特征依据卡方值从大到小选取 n 个特征，并设置为当前树所使用的特征，其中 n 为随机生成的不大于特征总数的非零整数。

通过基于卡方校验的初始化，能够保证初始化选用的特征与标签有着较强的关联性；同时引入随机因子，使得初始化种子能更好地在搜索空间扩散，避免集中在局部空间。

3.2 多层级变异的局部播种

FSFOA 在局部播种时每次只对单一特征进行取反，即每一颗新树只有一个特征状态发生了变化，这样不仅影响搜索效率，还会在 LSC 参数较小时导致搜索空间极度受限，容易陷入局部解。IFSFOA 提出了多层级变异策略来克服局部播种扩散能力不足的局限。

对于森林内每颗年龄为 0 的树首先复制 LSC 棵新树，新树使用参数局部播种分段(Local Seeding Segment，LSS)，并根据式(2)等比递增地选取 n 个特征进行选用状态的改变，n_i 代表复制出的第 i 棵新树需要处理的特征数。

$$n_i=(LSC/LSS)\times i \tag{2}$$

在选取出所需变换选取状态的特征后，不再使用取反算式，而是模拟自然界的基因变异过程，采用全随机变化的策略。即对于任意被挑选特征的选用状态取值与旧值无关，而是 0/1 随机取值。当 LSC 取值为 2，LSS 取值为 2，其具体过程如图 1 所示。

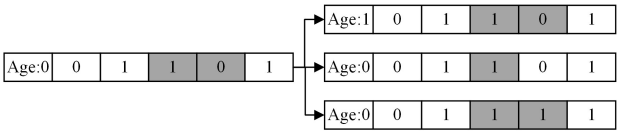

图 1　多层级变异策略($LSC=2$,$LSS=2$)

Fig. 2　Multi-level variation strategy with $LSC=2$ and $LSS=2$

首先根据原树拷贝出 LSC 棵树 $T1$ 和 $T2$，并将原树的年龄加 1。根据式(2)，从 $T1$ 任选一个特征并随机设置特征使用状态，放入森林中；接着，从 $T2$ 任选两个特征并随机设置特征使用状态，放入森林中。此时完成了该树的局部播种，继续其他树的播种过程。

通过多层级变异策略不仅能够充分提高局部搜索能力，还能够辅助提升全局播种最优树的寻优特性，使得解空间更容易收敛到最优解。

3.3　贪心取优

FSFOA 在完成局部播种后，会根据年龄上限和区域上限将所有被淘汰的树都加入候选森林。因为 FSFOA 未对候选森林做进一步的筛选，所以有大量劣质树存在于候选森林中，影响了全局搜索的有效性。

为了剔除劣质树，充分发挥优质树的作用，提高搜索能力和效率，IFSFOA 提出了贪心取优策略：增加候选森林规模上限参数 max candidates，若候选森林规模大于 *max candidates* 棵，则将候选森林内的树按照适应度逆序取出 *max candidates* 棵树放回候选森林，丢弃剩余树。

通过贪心取优限制候选森林规模，能够保证后续全局播种时都会使用适应度较高的优质树做进一步的搜索范围拓展，从而保证能在解集空间内进行有效搜索，避免劣质树的干扰，提高寻优能力。

3.4　算法流程

在完成初始化后，IFSFOA 就会进入主要的迭代流程：局部播种、规模限制、贪心取优、全局播种和更新最优树，持续迭代至满足停止条件。在完成迭代后，输出最优特征子集，完成特征搜索过程。

算法 1 给出了 IFSFOA 算法的伪代码，并用粗体标出了改进的地方。第一段粗体是基于卡方校验的初始化，第二段粗体是多层级变异的局部播种策略，第三段粗体是贪心取优。

算法 1　IFSFOA

输入：LSC，lifetime，arealimit，transferrate，GSC，LSS，maxcandidates

输出：适应度最高的特征子集

1. 初始化森林
2. **计算各特征与标签之间的卡方值**
3. **根据卡方值对特征进行逆序排序**
4. **For** 每颗森林中的树 **do**
5. 　**k=rand(1,特征数)**
6. 　**根据卡方值逆序选择 k 个特征**
7. 　**将选定特征的状态置为 1**
8. 　将树的 Age 置为 0
9. **End for**
10. **While** 满足迭代条件 **do**
11. 　将森林中 Age 为 0 的树进行局部播种
12. 　**For** i=1 to LSC **do**
13. 　　**根据式(2)获得特征变异数 n_i**
14. 　　**在选定树中随机选取 n_i 个特征**
15. 　　**将选定特征的状态随机 0/1 取值**
16. 　**End for**
17. 　将森林中所有树 Age 加 1
18. 　通过 life time 和 area limit 进行种群规模限制
19. 　**根据适应度逆序选择 max candidates 颗候选森林中的树**
20. 　**丢弃候选森林中未被选中的树**
21. 　对候选森林进行全局播种
22. 　更新最优树
23. 　将适应度值最高的树 Age 置为 0
24. **End while**
25. 返回适应度值最高的特征子集

4　实验结果与分析

本文使用 python3.7 实现代码，所有实验均在一台配置为 Intel i7、16 GB 内存、250 GB 硬盘的计算机上完成。

4.1　数据集与对比算法

本文实验均使用公开数据集，包括神州信息提供的汽车贷款数据集、UCI 提供的德国信用和澳大利亚信用数据集，以及 Kaggle 提供的 AER 信用数据集，并根据特征数量将数据集划分为低维数据、中维数据和高维数据，分别对应的范围是[0,19]、[20,49]、[50,∞][23]，使用数据集的特征数、实例数和分类数如表 1 所列。

表 1　实验中使用的数据集

Table 1　Dataset used in experiment

Dataset	Feature	Instance	Class
Car Loan	50	3968	2
German	20	1000	2
Australian	14	690	2
AER	11	1319	2

为了验证本文算法在信用评估问题中能在降低数据维度的同时提高分类能力，本文将其与较为高效的特征选择算法进行对比，对比算法的详细信息如表 2 所列。

表 2　对比算法信息

Table 2　Information of comparison methods

Name	Year
SCAFS[21]	2016
WOAFS[22]	2018
FOAFS[23]	2018
QBHHOFS[24]	2019
SSAFS[25]	2019

4.2　评价指标

考虑到信用评估通常为不平衡数据集，本文主要采用 3 个评价指标进行综合考量：分类精确度(CA)、AUC 值和维度缩减率(DR)。分类精确度和 AUC 值用于评估算法的分类能力，维度缩减率用于评估算法的降维能力。

分类精确度代表所有正确分类个数在数据集中的占比，其定义如式(3)所示。其中，NCC 代表正确的分类数，NAS 代表数据集的实例数。

$$CA=NCC/NAS \tag{3}$$

通过混淆矩阵，我们计算得到假阳性率 FPR 和真阳性率 TPR：

$$\begin{cases} TPR=TP/(TP+FN) \\ FPR=FP/(FP+FN) \end{cases} \tag{4}$$

以假阳性率为横坐标、真阳性率为纵坐标，可以绘制出ROC曲线，曲线下的面值即为AUC值。AUC值通常反映了模型的分类效果，AUC值越大代表模型分类效果越佳。

维度缩减率的定义如式(5)所示。其中，NSF 代表参与学习的特征数，NAF 代表总特征数。

$$DR=1-NSF/NAF \tag{5}$$

在实际业务使用中，人们会综合权衡分类能力和维度缩减能力所带来的收益。所以为了统一评价标准，并进一步验证所提算法的有效性，本文采用综合权重评价来评价算法在实际应用中的有效性，综合权重评价定义如式(6)所示：

$$Fitness=CA+0.01\times DR \tag{6}$$

4.3 参数设置

FSFOA认为 *life time*，*area limit* 和 *transfer rate* 这3个参数用于控制搜索范围和迭代周期，受数据集的影响较小，所以将 *life time* 固定为15、*area limit* 固定为50、*transfer rate* 固定为5%[23]。同理，将 *max candidates* 设置为150，*LSS* 设置为10。IFSFOA其他参数设置如表3所列。

表3 IFSFOA的参数信息

Table 3 Specific information of parameters of IFSFOA

Dataset	LSC	GSC
Auto Loan	12	24
German	5	10
Australian	3	6
AER	2	4

本文实验使用信用评估中最常用的逻辑回归作为分类器，正则化设置为L2，正则化强度设置为1.0，求解方式使用LBFGS。同时，为了避免因过拟合问题导致的实验数据不可靠问题，保证对比实验的有效性，本文使用5折交叉验证方法评估结果。

4.4 实验结果分析

具体实验对比结果如表4－表7所列，IFOAFS-init代表仅初始化策略的优化，IFOAFS-local代表仅局部播种的优化，IFOAFS-limit代表仅规模限制的优化。表中的粗体代表该组对比实验最佳的分类精度、AUC值或维度缩减率，评价指标均为实验5次后所取的平均值。

表4 汽车贷款数据集的对比结果

Table 4 Comparison results for auto loan dataset

Algorithm	CA	AUC	DR
Origin	0.710	0.716	—
SCAFS	0.709	0.710	0.400
WOAFS	0.709	0.708	0.120
QBHHOFS	0.708	0.706	0.300
SSAFS	0.710	0.712	0.400
FOAFS	0.713	0.725	0.440
IFOAFS-init	0.714	0.729	0.420
IFOAFS-local	0.715	0.733	0.440
IFOAFS-limit	0.708	0.709	0.420
IFOAFS	**0.716**	**0.737**	**0.500**

表5 德国信用数据集的对比结果

Table 5 Comparison results of German dataset

Algorithm	CA	AUC	DR
Origin	0.749	0.656	—
SCAFS	0.753	0.660	0.350
WOAFS	0.756	0.670	0.250
QBHHOFS	0.756	0.669	0.150
SSAFS	0.754	0.659	0.550
FOAFS	0.752	0.663	0.300
IFOAFS-init	0.756	0.667	0.400
IFOAFS-local	0.762	0.675	0.450
IFOAFS-limit	0.757	0.667	0.450
IFOAFS	**0.764**	**0.677**	**0.650**

表6 澳大利亚信用数据集的对比结果

Table 6 Comparison results of Australian dataset

Algorithm	CA	AUC	DR
Origin	0.869	0.869	—
SCAFS	0.869	0.871	0.428
WOAFS	0.868	0.869	**0.500**
QBHHOFS	0.860	0.862	**0.500**
SSAFS	0.868	0.869	0.286
FOAFS	0.862	0.864	0.428
IFOAFS-init	0.862	0.864	0.428
IFOAFS-local	0.869	0.869	0.357
IFOAFS-limit	0.868	0.870	**0.500**
IFOAFS	**0.874**	**0.874**	0.286

表7 AER信用数据集的对比结果

Table 7 Comparison results of AER dataset

Algorithm	CA	AUC	DR
Origin	0.981	0.985	—
SCAFS	0.982	0.988	0.363
WOAFS	0.982	0.988	0.455
QBHHOFS	0.982	0.988	0.363
SSAFS	0.982	0.988	0.090
FOAFS	0.982	0.988	0.272
IFOAFS-init	0.982	0.988	0.272
IFOAFS-local	0.982	0.988	0.455
IFOAFS-limit	0.982	0.988	0.455
IFOAFS	0.982	0.988	**0.545**

通过表4－表7可以看出，IFOAFS在各数据集的分类精确度和AUC值都优于其他特征选择算法，在德国信用和澳大利亚信用数据集上更是有着明显优势，较FSFOA有着较大提升，分类精确度的提升范围为0.4%～1.6%，AUC值的提升范围为1.7%～2.1%。

接着对比维度缩减率。IFSFOA在汽车贷款、德国信用和AER信用数据集上均为最优解，在澳大利亚信用数据集上表现不佳。澳大利亚信用数据集本身维度较低，维度的降低对计算复杂度影响有限，所以为了保证分类能力，IFSFOA选择了一个分类能力更佳，但维度稍高的特征子集。

观察表8的综合权重评价指标，可以看到IFSFOA在各数据集上均为最优值，这也进一步验证了IFOAFS具备较佳的泛化能力，在信用评估上有着良好的应用成果，能够有效降低数据维度，并提高分类能力。

表 8　对比算法在不同数据集上的综合权重评价

Table 8　Evaluation of combined weights of comparison algorithms on different datasets

Algorithm	Auto Loan	German	Australian	AER
SCAFS	0.713	0.756	0.873	0.985
WOAFS	0.710	0.758	0.873	0.986
QBHHOFS	0.711	0.757	0.865	0.985
SSAFS	0.714	0.759	0.870	0.982
FOAFS	0.717	0.755	0.866	0.984
IFOAFS-init	0.718	0.76	0.866	0.984
IFOAFS-local	0.719	0.766	0.872	0.986
IFOAFS-limit	0.712	0.761	0.873	0.986
IFOAFS	**0.721**	**0.771**	**0.877**	**0.987**

通过消融实验证明了本文提出的 3 个改进方案的有效性，将 3 个改进方案组合得到的 IFSFOA 也有着更优异的表现，其寻优能力明显高于 FSFOA。

在实际应用中，算法的执行效率也是主要关注点之一。为了进一步验证 IFSFOA 的有效性，本文对比了在不同数据集上特征搜索的收敛轮次，结果如表 9 所列。通过表 9 可以看到，在没有改变主体流程和搜索空间的前提下，IFSFOA 在各个数据集的搜索收敛速度都快于 FSFOA，说明 IFSFOA 的寻优能力和搜索效率明显优于 FSFOA。

表 9　迭代轮数对比

Table 9　Comparison results of AER dataset

Dataset	FSFOA	IFSFOA
Auto Loan	40	14
German	120	12
Australian	120	4
AER	15	7

综上，从实验结果可以得出，IFSFOA 较 FSFOA 有着明显优势。IFSFOA 不仅在分类能力和维度缩减能力方面获得了明显提升，而且有着更高的搜索效率，能有效提高信用评估的性能。

结束语　本文针对 FSFOA 在特征选择过程中进行了分析，并针对其存在的不足给出了 3 个优化方案，提出了基于改进森林优化算法的特征选择算法 IFSFOA，并将其应用到信用评估领域中。IFSFOA 通过改进的初始化策略和局部播种策略，有效提高了特征选择能力；通过在规模限制阶段引入贪心策略，在提升算法寻优能力的同时，降低了算法复杂度。

在不同维度的公开信用评估数据集上进行的对比实验结果表明，所提方法在分类精确度、AUC 值和维度缩减率上均有明显优势，能够有效提高信用评估效果。未来将进一步研究算法在超高维度的搜索空间过大、提高维度缩减能力等问题，并考虑将所提方法应用到其他领域中。

参 考 文 献

[1] LI Z Q, DU J Q, NIE B, et al. Summary of Feature Selection Methods[J]. Computer Engineering and Applications, 2019, 55(24): 10-19.

[2] KHAIRE U M, DHANALAKSHMI R. Stability of feature selection algorithm: A review[J]. Journal of King Saud University-Computer and Information Sciences, 2019, 34(4): 1060-1073.

[3] ABUKWAIK H, BURGER A, ANDAM B K, et al. Semi-automated feature traceability with embedded annotations[C]// Proceedings of the 2018 IEEE International Conference on Software Maintenance and Evolution(ICSME). IEEE, 2018: 529-533.

[4] AGRAWAL P, ABUTARBOUSH H F, GANESH T, et al. Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019)[J]. IEEE Access, 2021, 9(1): 26766-26791.

[5] GHOSH M, GUHA R, SARKAR R, et al. A wrapper-filter feature selection technique based on ant colony optimization[J]. Neural Computing and Applications, 2020, 32(12): 7839-7857.

[6] MALDONADO J, RIFF M C, NEVEU B. A review of recent approaches on wrapper feature selection for intrusion detection [J]. Expert Systems with Applications, 2022, 198(1): 116822.

[7] HANCER E, XUE B, ZHANG M. Differential evolution for filter feature selection based on information theory and feature ranking[J]. Knowledge-Based Systems, 2018, 140(1): 103-119.

[8] MALDONADO S, LÓPEZ J. Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification[J]. Applied Soft Computing, 2018, 67(1): 94-105.

[9] LU M. Embedded feature selection accounting for unknown data heterogeneity[J]. Expert Systems with Applications, 2019, 119(1): 350-361.

[10] EL ABOUDI N, BENHLIMA L. Review on wrapper feature selection approaches[C]// Proceedings of the 2016 International Conference on Engineering & MIS(ICEMIS). IEEE, 2016: 1-5.

[11] BOUZOUBAA K, TAHER Y, NSIRI B. Dos attack forecasting: A comparative study on wrapper feature selection[C]// Proceedings of the 2020 International Conference on Intelligent Systems and Computer Vision(ISCV). IEEE, 2020: 1-7.

[12] KARUNAKARAN V, RAJASEKAR V, JOSEPH S. Exploring a filter and wrapper feature selection techniques in machine learning [M]// Computational Vision and Bio-Inspired Computing. Springer, 2021: 497-506.

[13] BALOGUN A O, BASRI S, JADID S A, et al. Search-based wrapper feature selection methods in software defect prediction: an empirical analysis[C]// Proceedings of the Computer Science On-line Conference. Springer, 2020: 492-503.

[14] LIU J H, LIN M L, ZHANG J, et al. A kind of heuristic local random feature selection algorithm[J]. Computer Engineering and Applications, 2016, 52(2): 170-174.

[15] KOWAL M, SKOBEL M, NOWICKI N. The feature selection problem in computer-assisted cytology[J]. International Journal of Applied Mathematics and Computer Science, 2018, 28(4): 759-770.

[16] SHARMA M, KAUR P. A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem [J]. Archives of Computational Methods in Engineering, 2021, 28(3): 1103-1127.

[17] ROSTAMI M, BERAHMAND K, NASIRI E, et al. Review of swarm intelligence-based feature selection methods[J]. Engineering Applications of Artificial Intelligence, 2021, 100(1): 104210-104210.

[18] TELIKANI A, TAHMASSEBI A, BANZHAF W, et al. Evolu-

tionary Machine Learning: A Survey[J]. ACM Computing Surveys(CSUR), 2021, 54(8): 1-35.

[19] EIBEN A E, SMITH J E. What is an evolutionary algorithm? [M]// Introduction to Evolutionary Computing. Springer, 2015: 25-48.

[20] BABATUNDE O H, ARMSTRONG L, LENG J, et al. A genetic algorithm-based feature selection[J]. International Journal of Electronics Communication and Computer Engineering, 2014, 5(4): 899-905.

[21] HAFEZ A I, ZAWBAA H M, EMARY E, et al. Sine cosine optimization algorithm for feature selection[C]// Proceedings of the 2016 International Symposium on Innovations in Intelligent Systems and Applications(INISTA). IEEE, 2016: 1-5.

[22] TUBISHAT M, ABUSHARIAH M A, IDRIS N, et al. Improved whale optimization algorithm for feature selection in Arabic sentiment analysis[J]. Applied Intelligence, 2019, 49(5): 1688-1707.

[23] GHAEMI M, FEIZI-DERAKHSHI M R. Feature selection using forest optimization algorithm[J]. Pattern Recognition, 2016, 60(1): 121-129.

[24] TOO J, ABDULLAH A R, MOHD SAAD N. A new quadratic binary harris hawk optimization for feature selection[J]. Electronics, 2019, 8(10): 1130.

[25] HEGAZY A E, MAKHLOUF M, EL-TAWEL G S. Feature selection using chaotic salp swarm algorithm for data classification [J]. Arabian Journal for Science and Engineering, 2019, 44(4): 3801-3816.

[26] LIU H, ZHOU M, LIU Q. An embedded feature selection method for imbalanced data classification[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(3): 703-715.

[27] ADO A, DERIS M M, SAMSUDIN N A, et al. Adaptive and Global Approaches Based Feature Selection for Large-Scale Hierarchical Text Classification[C]// Proceedings of the International Conference of Reliable Information and Communication Technology. Springer, 2022: 105-116.

HUANG Yuhang, born in 1998, postgraduate. His main research interests include data mining and software engineering.

WANG Baohui, born in 1973, senior engineer, master supervisor. His main research interests include software architecture, big data, artificial intelligence, etc.