基于混合式特征选择的辐射源个体识别

顾楚梅，曹建军，王保卫，徐雨芯

引用本文

顾楚梅，曹建军，王保卫，徐雨芯. 基于混合式特征选择的辐射源个体识别[J]. 计算机科学，2024，51(5): 267-276.

GU Chumei, CAO Jianjun, WANG Baowei, XU Yuxin. Specific Emitter Identification Based on Hybrid Feature Selection [J]. Computer Science, 2024, 51(5): 267-276.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300216

基于混合式特征选择的辐射源个体识别

顾楚梅[1,2,3]　曹建军[1,2]　王保卫[3]　徐雨芯[1,2,3]
1 国防科技大学第六十三研究所　南京 210007
2 国防科技大学大数据与决策实验室　长沙 410073
3 南京信息工程大学计算机学院网络空间安全学院　南京 210044
(m15261820030@163.com)

摘　要　为提高辐射源个体识别的准确率和运算效率,提出了一种基于混合式特征选择的辐射源个体识别。封装式特征选择方法分类正确率高,但计算复杂度高,处理高维数据时效率低。嵌入式特征选择方法计算复杂度低,但依赖于特定分类器。针对上述问题,综合封装式和嵌入式特征选择方法的特点,首先对信号数据使用3种嵌入式方法(随机森林、XGBoost和LightGBM)初选特征,分别得到随机森林子集、XGBoost子集和LightGBM子集。然后使用封装式方法对初选后得到的子集进行第二次降维,其中搜索策略分别使用序列后向搜索策略和蚁群优化算法,分类算法使用LightGBM。混合式方法共得到6种特征选择模型,通过对比各个模型得到的分类正确率和最优子集中的特征个数,确定最佳混合式特征选择模型。

关键词:辐射源个体识别;特征选择;随机森林;XGBoost;LightGBM;序列后向搜索策略;蚁群优化
中图分类号　TP391

Specific Emitter Identification Based on Hybrid Feature Selection

GU Chumei[1,2,3],CAO Jianjun[1,2],WANG Baowei[3] and XU Yuxin[1,2,3]
1 The Sixty-third Research Institute,National University of Defense Technology,Nanjing 210007,China
2 Laboratory for Big Data and Decision,National University of Defense Technology,Changsha 410073,China
3 School of Computer Science,Nanjing University of Information Science and Technology,Nanjing 210044,China

Abstract　To improve the accuracy and computational efficiency of specific emitter identification,a specific emitter identification based on hybrid feature selection is proposed. Wrapped feature selection methods have high classification accuracy,but it has high computational complexity and low efficiency in processing high-dimensional data. Embedded feature selection methods have low computational complexity, but rely on specific classifiers. To address the above problems, combining the characteristics of wrapped and embedded feature selection methods,firstly,three embedded methods(Random Forest,XGBoost,and LightGBM) are used to initially select features for signal data,and a random forest subset,an XGBoost subset and a LightGBM subset are obtained respectively. Secondly,the wrapped methods are used to perform a second dimensionality reduction on the subset obtained after the primary selection. Sequential backward selection and an ant colony optimization algorithm are used as research strategies respectively,while LightGBM is used as the classification algorithm. A total of six feature selection models are obtained from the proposed hybrid feature selection method. The optimal hybrid feature selection model is determined by comparing the classification accuracy and the number of features in the optimal subset obtained by each model.

Keywords　Specific emitter identification,Feature selection,Random forest,XGBoost,LightGBM,Sequential backward selection,Ant colony optimization

1　引言

随着通信工程的快速发展和移动通信设备的普及,无线通信在现代社会中越来越不可或缺。近年来,无线通信领域中的辐射源个体识别(Specific Emitter Identification,SEI)技术受到国内外的广泛关注[1]。在该领域中,辐射源发出的信号不仅包含了所需的信号信息,还承载了辐射源内部器件的固有硬件信息,通过提取这部分信息特征来识别不同辐射源个体的过程称为辐射源个体识别[2]。SEI一般由3个步骤组成:对接收到的原始辐射源信号进行预处理;从经过预处理后

到稿日期:2023-03-29　返修日期:2023-09-18
基金项目:国家自然科学基金(71901215,61371196);中国博士后科学基金(20090461425,201003797)
This work was supported by the National Natural Science Foundation of China(71901215,61371196) and China Postdoctoral Science Foundation Foundation(20090461425,201003797).
通信作者:曹建军(caojj@nudt.edu.cn)

的信号中提取辐射源物理层本质的细微特征，称为射频指纹(Radio Frequency Fingerprint，RFF)特征；使用分类器识别信号以确定发射此信号的辐射源个体[3]。

而辐射源信号数据量大、提取到的射频指纹特征维数高等问题给后续辐射源个体分类识别增加了一定难度，利用高维特征进行学习会花费大量计算时间和较大存储空间，不利于科学研究和实际应用。在高维特征中有一部分特征与学习任务高度相关，但也存在不相关特征或冗余特征，这些特征的输入会大大降低学习性能，引发维数灾难问题[4]。通常使用特征选择(Feature Selection，FS)方法来剔除上述特征，以达到降低特征维数并提高辐射源个体识别正确率的目的。特征选择旨在根据某种评价标准从原始特征空间中消除不相关和冗余特征、选出高相关性特征组成特征子集来获得比使用所有特征更好的性能[5]。

根据特征选择与学习算法的结合方式，特征选择算法主要分为过滤式、封装式和嵌入式。过滤式方法通过评估每一特征的鉴别能力来过滤鉴别能力差的特征，该方法运算时间短、独立于学习算法且具有高泛化性，但依赖具体的度量标准，典型的过滤式特征选择方法有 Relief、Pearson 相关系数、方差分析、卡方检验和互信息等[6]。封装式方法由搜索策略和学习算法组成，将特征选择封装到学习算法中，通过学习算法的预测结果进行评估，并使用搜索策略调整特征子集。该方法所得特征子集的性能高且考虑了特征间的相互关系，但计算复杂度高。典型的封装式方法有递归特征消除等[7]。嵌入式方法将特征选择嵌入到学习算法中，学习算法结束的同时得到了特征重要性值，该方法效率较高、特征分辨力好，但依赖于指定学习算法。典型的嵌入式方法有基于惩罚项的方法，例如 Lasso 等及基于树模型的方法包括决策树(Decision Tree，DT)、随机森林(Random Forest，RF)、梯度提升决策树(Gradient Boosting Decision Tree，GBDT)、极端梯度提升(eXtreme Gradient Boosting，XGBoost)和轻量级梯度提升(Light Gradient Boosting Machine，LightGBM)等[8]。

为提高辐射源个体识别的效果和效率，仅使用一种特征选择方法进行研究是不够的，可以结合两种或多种特征选择方法形成混合式特征选择方法。常见的为过滤式方法结合封装式方法和嵌入式方法结合封装式方法。在研究混合式特征选择方法时，间接地研究了两种典型的特征选择方法，本文对由嵌入式和封装式方法组成的混合式特征选择方法进行深入综合研究，促进了两种方法的进一步发展，为数据量大和特征高维带来的运算时间长、预测性能差的问题提供了解决方法，在辐射源个体识别的过程中具有重大意义。从研究背景出发，特征选择是提高辐射源信号处理效率，进一步减少数据通信量，提高辐射源个体识别准确率的关键步骤。研究辐射源信号个体的特征选择方法，有助于改造现有的通信电子侦察设备与研制新一代电子对抗情报侦察设备，在军事应用和民用方面具有重大意义。

混合式特征选择方法在多个领域和领域数据集上都能得到较优的特征子集。为了有效地消除帕金森病数据集中无用或有噪声的数据，文献[9]提出了一种结合过滤式和封装式方法优点的混合特征选择算法。采用 3 种不同的过滤方法对原始特征进行预排序，然后在封装式方法中使用人工蜂群算法作为学习算法，以获得最佳特征子集。混合方法不仅减小了特征子集的大小，还提升了特征子集分类的性能。为了应对高维特征空间对特征选择方法带来的巨大挑战，文献[10]提出了一种由过滤式方法和元启发式优化方法组成的混合特征选择方法。首先使用方差分析过滤式方法来识别方差为零或方差很低的特征，然后使用多目标灰狼优化算法以及互信息将初选后的子集优化为最佳相关特征子集。为了选择适当的特征来开发准确的金融危机预测模型，文献[11]提出了一种混合过滤式和封装式的方法。在过滤步骤中，信息增益和斯皮尔曼相关性被用作评价标准；在封装步骤中，使用水波优化算法作为搜索策略。为了减少高维特征集中的不相关或冗余特征并提高分类正确率，文献[12]提出了一种用于文本分类的混合过滤式和封装式文本特征选择方法。首先使用信息增益来选择一些得分高的特征，其次使用灰狼优化算法来继续优化初选特征子集。为解决医学数据中高维特征空间和高特征冗余的问题，文献[13]提出了一种结合信息增益比和遗传算法的混合特征选择算法。首先，使用基于信息增益比的过滤式方法对原始特征集进行排序；然后，根据等分割的密度原则对排序特征进行分组；最后，利用群体进化遗传算法对排序后的特征组分别进行特征选择。

在辐射源个体识别领域，为进一步提升辐射源个体识别的效果和效率，提出了一种混合式特征选择方法。该方法结合了嵌入式特征选择方法和封装式特征选择方法，主要过程如下。

1)选取 12 个统计特征参数和标准化相对能量，结合提升小波包分解与重构方法提取特征并构建特征参数体系。使用 Z-score 标准化方法对特征数据集进行预处理。

2)使用 3 种嵌入式方法，即 RF，XGBoost 和 LightGBM，分别对训练集进行第一次降维，选取前若干个重要性值最大的特征作为初选特征子集。

3)对初选特征子集使用封装式特征选择方法进行第二次降维，搜索策略分别采用序列后向搜索策略(Sequential Backward Selection，SBS)和改进的蚁群算法，结合一次降维的过程，生成 6 种混合式特征选择模型，根据最大分类正确率和最小特征子集规模选出最适合当前数据集的混合式方法。

4)采用不同信噪比下的电台数据集，相比特征全集、单一特征选择方法，综合考虑分类正确率和特征个数等评估指标。实验结果证明了所提方法的性能更优。

2 数学模型

特征选择的过程即从集合 set 中选择基数为 q 的一个特征子集 $subset^q$，使输入该子集时可以满足某种目标函数。根据所研究的问题即辐射源个体识别问题的本质是分类问题，特征选择和分类器的性能相关联。因此，使用分类器的分类正确率和所选特征子集中特征的个数作为特征选择的目标函数。实验中的数据集为从两个电台发出的辐射源信号数据，旨在通过辐射源个体识别技术识别两个电台，进而分析电台的性质、属性和危险等级等，本质上是一个二分类问题。

对使用分类器的分类正确率和所选子集中特征个数作为

目标函数的特征选择方法，特征选择问题可以描述为：从集合 set 中根据目标函数得到一个基数为 q 的特征子集 $subset^q$，输入 $subset^q$ 时得到的分类正确率 A 最大且特征个数 q 最小。具体数学模型如下：

$$\max A \tag{1}$$

$$\max q \tag{2}$$

$$\text{s.t. } |subset^q| = q, 1 \leqslant q \leqslant n \tag{3}$$

$$A = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

数据输入测试集前，已知正类(Positive)数据和负类(Negative)数据，模型预测的数据也分为正负两类。可以得出 4 个指标：样本真实类别为正类，模型识别结果也为正类(True Positive，TP)的个数；样本真实类别为正类，但模型识别结果为负类(False Negative，FN)的个数；样本真实类别为负类，但模型识别结果为正类(False Positive，FP)的个数；样本真实类别为负类，模型识别结果也为负类(True Negative，TN)的个数。

3 混合式特征选择

图 1 给出了本章的研究框架图，首先对原始辐射源信号进行特征提取得到特征全集，并使用 Z-score 标准化对特征数据集进行预处理。然后分别使用 RF，XGBoost 和 LightGBM 这 3 种嵌入式特征选择方法计算每一特征重要性并选取前 100 个特征重要性值大的特征组成初选特征数据集。接着使用封装式特征选择方法进行第二次降维，其中分类算法使用 LightGBM，搜索算法分别使用序列后向搜索策略和 GBAS，共得到 6 种特征选择模型，通过比较辐射源个体识别正确率和最优特征子集中特征的个数来选出最佳特征选择模型。

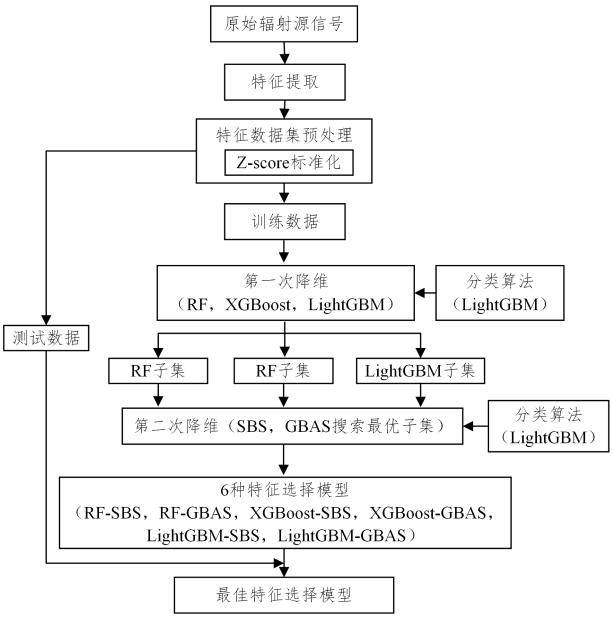


图 1 混合式特征选择框架图

Fig. 1 Framework of hybrid feature selection

3.1 特征提取

特征提取是辐射源个体识别的关键步骤，提取到的特征直接影响到特征选择和分类器的性能。原始辐射源信号数据量往往很大，辐射源个体识别的关键不在于使用所有数据对辐射源进行描述，而在于使用信号数据中的有效特征来识别辐射源个体。特征提取旨在通过变换来提取有效的识别特征，使原始信号从高维数据空间转移到低维特征空间[14]。

辐射源个体识别中常见的特征提取方法有双谱法、提升小波包分析法、Wigner-Ville 和经验模态分解法等。基于提升小波包分析法较好的时频分辨能力和较高的运算效率等优点，本文采用此方法，其分解与重构方法可以获得更多特征信息，扩大有效的识别特征集合[15]。

本文选取 12 个统计特征参数，即均值、平均幅值、方根幅值、标准差、有效值、峰-峰值，波形指标、脉冲指标、峰值指标、偏斜度、峭度和峪度指标，并使用标准化相对能量。依据统计特征参数和标准化相对能量，结合提升小波包分解与重构，给出特征参数体系。对于辐射源发出的信号，首先使用先序分解后序搜索算法[16]对数据进行最优基分解得到最佳小波包树；然后通过分解与重构将最佳小波包树调整为一棵两层的满二叉树，满二叉树的叶子节点分别记作(2,0)，(2,1)，(2,2)和(2,3)，计算 4 个频带内系数各统计特征值和标准化相对能量；接着分别对每一频带内的系数进行单支重构并分别提取相应的统计特征，最后对原信号重构并提取重构原信号的统计特征。

将重构原信号的 12 个统计特征参数(标号为 1－12)，小波包分解的第二层 4 个节点系数的各 12 个统计特征参数(标号为 13－60)，4 个单支重构信号的 12 个统计特征参数(标号为 61－108)及小波包分解的第二层 4 个节点系数的标准化相对能量(标号为 109－112)，共 112 个特征依次编号。为全面描述辐射源信号信息，在幅值信号、I 路信号和 Q 路信号上分别提取 112 个特征并构建特征集 $set=\{v_i \mid v_i=v_1, v_2, \cdots, v_n\}$，$i=1,2,\cdots,n$，此时 $n=336$。

3.2 Z-score 标准化

为统一数据样本的数量级、增加可比性及加快算法收敛速度，采用 Z-score 标准化对特征值进行处理[17]，其表达式如下：

$$v'_{ir} = \frac{v_{ir} - \bar{X}_{v_i}}{\delta_{v_i}} \tag{5}$$

其中，v_{ir} 为特征 i 的第 r 个特征值，$\bar{X}_{v_i}$ 和 σ_{v_i} 分别为特征 i 的均值和标准差。

Z-score 标准化将数据转换到某个范围，且不会改变原始数据的排列顺序。标准化后，不同数量级的特征在数值上进行了统一，寻优过程更为平缓，更容易正确地收敛到最优解。

3.3 嵌入式特征选择

3.3.1 RF

在决策树中，样本按其特征值进行分类。树的节点表示数据集的特征，分支表示分区的决策规则。基于多决策树构造随机森林，可以有效地防止过度拟合问题。随机树中各决策树之间没有任何关联，树分裂过程中具有随机性，通过预测每棵树的结果并根据投票原则获得最优解，主要用于分类问题[18]。具体流程如下[19]。

1)随机抽样(放回抽样)。对训练集样本进行随机抽样，得到多个子样本集，然后对每一个子样本集构造决策树模型。

2)随机选择特征。在每个子样本集中依据设置比例随机

选择一部分特征，并根据基尼指数等准则选择特征并将其作为分割节点，基尼指数的计算式如下。：

$$G=\sum_{i=1}^{C}p(i)\times(1-p(i)) \tag{6}$$

其中，C 为类别总数，并且数据属于 i 类的概率为 $p(i)$。当叶子节点所含样本均属于同一类或者达到设置的约束（如最大树的深度）时，决策树停止分裂。

3）构建决策树。对每个节点重复步骤 2），直到无法拆分，然后生成决策树。

4）投票。重复步骤 1）－步骤 3），建立大量决策树，生成随机森林，并根据“少数服从多数”的原则做出决策。

随机森林作为分类器时基于集成的思想，集合了多个决策树的分类结果，具有不受特征维数影响和精度较高等优点。

在每轮随机抽样中，训练集中有一部分数据没有被采样集采集中，没有参与决策树的建立，对于这部分数据，通常称为袋外数据（Out Of Bag，OOB）[20]。这些数据没有参与训练集模型的拟合，因此可以用来检测模型的泛化能力。在 RF 中，通过计算每棵决策树的袋外数据可以得到一个误差率，以衡量预测特征的重要性程度。具体步骤可以概括为：首先，对于每一棵决策树，选择相应的袋外数据并计算袋外数据误差，记为 err_{OOB1}。然后，随机对袋外数据所有样本的特征 X 加入噪声干扰（可以随机改变样本在特征 X 处的值），再次计算袋外数据误差，记为 err_{OOB2}。最后，假设随机森林中共有 N 棵树，则特征 X 的重要性如下：

$$imp=\sum_{i=1}^{N}(err_{\mathrm{OOB2}}-err_{\mathrm{OOB1}})/N \tag{7}$$

加入随机噪声后，如果袋外数据的准确率大幅度下降（即 $\mathrm{err}_{\mathrm{OOB2}}$ 上升），说明这个特征对于样本的预测结果有很大影响，进而说明该特征的重要程度高，故可用式（7）来衡量每一特征的重要性。

3.3.2 XGBoost

XGBoost[21]是一种 boosting 算法，其在 GBDT 算法的基础上，对损失函数进行二阶泰勒展开，并加入正则项，避免了过拟合，有效地加快了收敛速度。通过不断添加新的决策树来拟合先前预测的残差，XGBoost 算法减少了预测值和实际值之间的残差，提高了预测精度并获得了特征重要度分数。XGBoost 的预测描述如下：

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i),f_k\in F \tag{8}$$

其中，K 是决策树的数量，f_k 是第 k 个子模型，x_i 是第 i 个输入样本，F 是所有决策树的集合。

XGBoost 的目标函数由损失函数和正则项组成：

$$L^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t) \tag{9}$$

其中，t 是迭代次数，l 是测量预测值 $\hat{y}_i$ 和目标值 y_i 之间差异的可区分损失函数，$\hat{y}_i^{(t-1)}$ 是前一次迭代 $t-1$ 的预测，$\Omega(f_t)$ 是第 k 次迭代的正则项，其表达式如下：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2 \tag{10}$$

其中，γ 和 λ 是可以防止决策树过于复杂的正则项系数，T 是叶节点的数量，w 是叶权重。

XGBoost 对损失函数进行二阶泰勒展开[22]，然后找到目标函数的最小值，并通过以下公式计算相应的最优值。

$$\tilde{L}^{(t)}(q)=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_j+\lambda}+\gamma T \tag{11}$$

其中，q 表示将示例映射到对应叶索引的每棵树的结构，$I_j=\{i|q(x_i)=j\}$ 是叶子 j 的实例集，g_i 是样本 x_i 的一阶导数，h_i 是样本 x_i 的二阶导数。式（11）可用于评估树结构，其值越小，模型越好。

损失函数（也称为拆分后的增益）的计算式如下：

$$L_{\mathrm{split}}=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_j+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_j+\lambda}-\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_j+\lambda}\right]-\gamma \tag{12}$$

其中，右侧的 4 项分别表示分裂后的左子树得分和右子树得分、分裂前树的得分和复杂调整惩罚系数。式（12）中，$I=I_L\cup I_R$，当达到深度限制或 $L_{\mathrm{split}}<0$ 时，树将停止分裂。

XGBoost 特征的重要性得分如下：

$$IS_i=\{x|x=w_iv_i\} \tag{13}$$

其中，v_i 是特征集，w_i 是对应特征的权重。

3.3.3 LightGBM

LightGBM 是 GBDT 算法的一种优化，主要用于解决 GBDT 方法在处理大量数据或高维特征时效率低和扩展性差的问题。LightGBM 中直方图（Histogram）、基于梯度的单侧采样（Gradient-Based One-Side Sampling，GOSS）等方法的提出解决了上述问题[23]，该方法提供了度量标准来衡量模型中特征的重要性。

LightGBM 提供了两个度量标准来衡量模型中特征的重要性：1）split，每个特征在所有决策树中被分割的总次数；2）gain，特征在所有决策树中作为分裂点所得到的信息增益。一个特征在所有决策树中被分裂的次数越多或得到的信息增益值越大，此特征的重要性程度就越高，对预测结果的影响也就越大[24]。

对于特征 j，决策树选择最优分裂点 $d_{j^*}=\arg\max_d V_j(d)$ 并计算得到最大信息增益 $V_j(d_{j^*})$，然后在点 d_{j^*} 处将数据分成左右孩子节点。特征 j 在单棵决策树中节点 d 的重要性为：

$$IMP_{\mathrm{im}}=w_d\cdot\Delta V \tag{14}$$

其中，w_d 表示节点 d 的数据量与总数据量的比值，ΔV 表示节点 d 分裂后左右叶子节点与分裂前原节点的信息增益值。然后将每棵决策树中特征 j 的重要性相加得到特征 j 基于模型 LightGBM 的特征重要性评分，评分越高，该特征对预测结果越有效[25]。

3.3.4 特征选择

在搜索特征子集阶段，首先将评估特征阶段得到的重要性值进行排序，然后采用基于 3 种嵌入式特征选择中特征重要性值的序列后向搜索策略和改进蚁群搜索法进行特征选择，综合考虑输入该特征子集时得到的分类正确率和子集中特征的个数来确定最优特征子集。

3.4 封装式特征选择

3.4.1 序列后向搜索策略

序列后向搜索策略属于封装式特征选择方法。该方法首先需要与分类算法结合，然后从特征全集（特征总数 n）开始，

从特征集中剔除一个特征 x，计算剔除特征 x 后的评价函数值，遍历共 $n-1$ 个特征组合并选择评价函数值最大的特征组合进行下一步遍历，直至特征个数剔除到 1。最后比较所有遍历得到的结果，选出评价函数值最高的特征组合，即最优特征子集。序列后向搜索策略只能去除特征而不能加入特征，属于贪心算法，容易陷入局部最优解。当数据集是高维时，序列后向搜索策略运算量大，可以通过过滤式特征选择方法或嵌入式特征选择方法对特征集进行初选，得到一个较低维度的特征集合，再通过序列后向搜索策略继续从初选后的特征集中筛选特征。

3.4.2 蚁群算法搜索策略

引用文献[26]中的基于图的蚂蚁系统(Graph-Based Ant System，GBAS)算法求解。引入 GBAS，根据辐射源信号特征选择问题构造有向图，如图 2 所示。

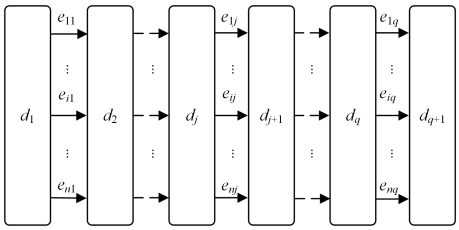


图 2 子集问题构造图的有向图

Fig. 2 Constructing directed graphs of graphs for subset problems

图 2 中，n 是备选集的基数，q 是所获得子集的最大可能基数，节点为 $d_k(k=1,2,\cdots,q+1)$，有向图的边代表备选特征，$e_{ij}=\langle d_j,d_{j+1}\rangle$，$j=1,2,\cdots,q$ 表示备选集的第 i 个元素，路径映射为一个求得的辐射源信号特征子集。

在时间 t，节点 d_1 处的蚂蚁数量被设置为 M，并且在约束条件下，每个蚂蚁根据边上的信息素和启发式因子独立地选择一个边到下一个节点。在时间 t，第 a 只蚂蚁通过线段 e_{ij} 从 d_j 转移到 d_{j+1} 的概率如下：

$$P_{ij}^{a}(t)=\begin{cases}\dfrac{[\tau_{ij}(t-1)]^{\alpha}\eta_i^{\beta}}{\sum\limits_{e_{bj}\notin tabu_a}[\tau_{bj}(t-1)]^{\alpha}\eta_b^{\beta}}, & e_{ij}\notin tabu_a\\ 0, & \text{其他}\end{cases}\tag{15}$$

其中，$\tau_{ij}(t)$ 是在时间 $t(t=0,1,2,\cdots)$ 时边 e_{ij} 上的信息素的量，初始化信息素量 $\tau_{ij}(0)=D(D$ 是一个常数)。启发式因子 η_i 表示选择第 i 个元素的期望程度，α 和 β 分别是信息素量和启发式因子的重要程度。第 a 只蚂蚁经过的边用禁忌表 $tabu_a$ 记录。

针对本文研究的信号数据的特点，启发式因子 η_i 的计算式如下：

$$\eta_i=\min\left\{\frac{|\mu_{1i}-\mu_{2i}|}{\sqrt{\sigma_{1i}^2+\sigma_{2i}^2}}\right\}\tag{16}$$

其中，采用最小 Fisher 标准判别率作为特征 i 的启发式因子，μ_{1i} 和 μ_{2i} 为两状态类特征 i 的均值，σ_{1i}^2 和 σ_{2i}^2 为方差。

根据文献[26]中的定义 2 和定理 1，在信息素更新阶段，当一条路径的信息素增强时，等效路径的信息素也需要增强，即基于等效路径增强的信息素更新策略的表达式如下：

$$\tau_{ij}(t)=\begin{cases}(1-\rho)\tau_{ij}(t-1)+\dfrac{\Phi'(tabu^t)}{Q}, & e_{ij}\in\Gamma(tabu^t)\\ (1-\rho)\tau_{ij}(t-1), & \text{其他}\end{cases}\tag{17}$$

其中，ρ 是信息素挥发系数，$\Phi'(tabu^t)/Q$ 是信息素增量公式，$\Phi'(tabu^t)$ 是信息素增强路径的目标函数值，Q 是常数(Q 根据 ρ 值确定，用于调节信息素增量)。

3.4.3 特征选择

混合式特征选择方法在第二阶段使用封装式特征选择方法，序列后向搜索和蚁群算法作为搜索策略，LightGBM 作为分类器，对初选后的特征子集进行进一步筛选得到最优特征子集。

4 实验结果与分析

4.1 数据准备

实验在 1 台 i7-4770 3.40 GHz 4 核处理器、24 GB 内存的电脑上运行，开发环境为 Python 3.8。

实验数据来源于两个相同型号的电台辐射源，采集环境为无噪声的环境，两个电台发出的信号在 10 种不同的采集状态下获得。10 种采集状态下信号的具体参数如表 1 所列。

表 1 信号参数

Table 1 Signal parameters

载波/MHz	调制方式	信号带宽	采样频率/M	间隔时间/ms
55	QPSK	25 kHz	1	20
75	QPSK	25 kHz	1	20
420	QPSK	2 MHz	20	2
420	QPSK	5 MHz	50	1
420	QPSK	10 MHz	80	1
420	QPSK	20 MHz	100	2
2000	QPSK	2 MHz	20	2
2000	QPSK	5 MHz	50	1
2000	QPSK	10 MHz	80	1
2000	QPSK	20 MHz	100	1

对于每个电台，在每种采集状态下选择 200 组(每组 4096 个数据)数据(共 4000×4096 个)，并选择 75%(3000)进行训练，选择 25%(1000)进行测试。原始数据在无噪声的环境下采集得到，为验证所提方法在噪声数据上的效果，添加高斯白噪声将信噪比分别调整为 10 dB 和 5 dB，并分别进行特征提取得到特征集合。信噪比为 10 dB 的特征集合 $set_{10\text{dB}}=\{v_i\mid v_i=v_1,v_2,\cdots,v_n\}$，$i=1,2,\cdots,n=336$ 和信噪比为 5 dB 的特征集合 $set_{5\text{dB}}=\{v_i\mid v_i=v_1,v_2,\cdots,v_n\}$，$i=1,2,\cdots,n=336$。

4.2 特征重要性值

对于无噪声特征集合 set_{original}，嵌入式方法 RF，XGBoost 和 LightGBM 都可以根据相关公式计算每一特征的重要性值。图 3 给出了 LightGBM 方法前 20 个最重要特征的特征标号-特征重要性值柱状图[27]。

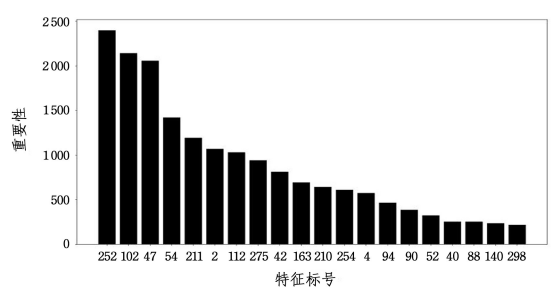


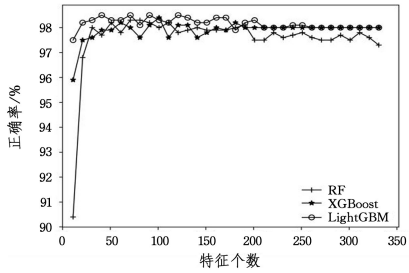
图 3 特征重要性排序

Fig. 3 Feature importance ranking

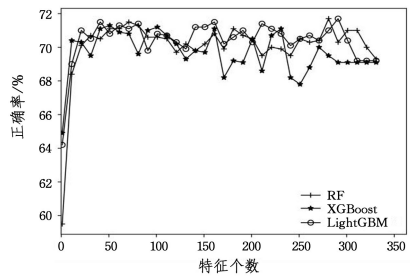
为获得最优特征子集，将计算得到的特征重要性值从小到大进行排序，并按照一定步长(10 维)依次剔除特征。

4.3 第一次降维

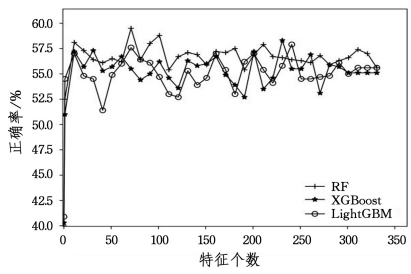
第一次降维使用嵌入式特征选择方法，分别对数据集 $set_{original}$，set_{10dB} 和 set_{5dB} 基于 RF，XGBoost 和 LightGBM 方法计算每一特征的重要性值，并对得到的重要性值进行排序。对于特征全集，根据特征重要性值按照一定步长(10 维)依次剔除特征，并绘制特征个数-分类正确率变化折线图(见图 4)，综合考虑分类正确率和特征个数，得到第一次降维的结果。



(a)数据集 $set_{original}$ 分类正确率的变化趋势



(b) 数据集 set_{10dB} 分类正确率的变化趋势



(c)数据集 set_{5dB} 分类正确率的变化趋势

图 4 不同数据集特征个数-分类正确率的变化趋势图

Fig. 4 Trend chart of the number of features-classification accuracy in different datasets

图 4(a)给出了在特征集 $set_{original}$ 中，3 种嵌入式方法(RF，XGBoost 和 LightGBM)分别计算特征重要性值并进行排序，然后采用维度逐步递减方法得到分类正确率的变化趋势。在特征个数相同时，LightGBM 方法得到的正确率最高，XGBoost 方法得到的正确率居中，RF 方法得到的正确率最低，因此对于 $set_{original}$ 特征集，LightGBM 方法的一次降维效果较好。从图 4(a)的整体趋势可以看出，随着特征个数的增加，3 种方法的分类正确率均基本呈现先增后减并趋于平缓的趋势，特征个数在 1～100 区间内，3 种嵌入式方法可以获得最高的分类正确率，特征个数在 101～336 区间内，分类正确率基本呈现振动下降的趋势，且变化较小。因此，对于 $set_{original}$ 数据集，选取前 100 个特征重要性值大的特征作为初选特征子集。

图 4(b)给出了在特征集 set_{10dB} 中，特征个数与分类正确率的变化趋势。在特征个数相同时，RF 方法和 LightGBM 方法得到的正确率较高，XGBoost 方法得到的正确率较低，因此对于 set_{10dB} 特征集，RF 方法和 LightGBM 方法的一次降维效果较好。从图 4(b)的整体趋势可以看出，随着特征个数的增加，3 种方法的分类正确率均基本呈现先增后振动平缓的趋势，特征个数在 1～100 区间内，3 种嵌入式方法基本可以获得较高的分类正确率，特征个数在 101～336 区间内，分类正确率呈现变化较小的上下振动趋势，且此区间的最高分类正确率比 1～100 区间内的最高分类正确率提高不大。因此，对于 set_{10dB} 数据集，选取前 100 个特征重要性值大的特征作为初选特征子集。

图 4(c)给出了在特征集 set_{5dB} 中，特征个数与分类正确率的变化趋势。在特征个数相同时，RF 方法得到的正确率最高，XGBoost 方法和 LightGBM 方法得到的正确率较低，因此对于 set_{5dB} 特征集，RF 方法的一次降维效果较好。从图 4(c)的整体趋势可以看出，随着特征个数的增加，3 种方法的分类正确率均基本呈现先增后平缓振动的趋势，特征个数在 1～100 区间内，3 种方法基本可以获得最高的分类正确率，特征个数在 101～336 区间内，分类正确率在 55.0%周围呈现变化较小的上下振动趋势，且此区间的最高分类正确率低于区间 1～100 的最高分类正确率。因此，对于 set_{5dB} 数据集，可以选取前 100 个特征重要性值大的特征作为初选特征子集。

根据上述分析，每种数据集均选取前 100 个特征重要性值大的特征作为第二次降维的特征输入。表 2 列出了每种数据集下每种方法选取的 100 个特征的编号。

表 2 第一次降维得到的初选特征子集

Table 2 Primary feature subset obtained from the first dimensionality reduction

数据集	RF	XGBoost	LightGBM
$set_{original}$	{112,54,100,102,101,52,4,50,53,42,5,334,3,51,90,2,275,323,252,151,210,163,162,211,199,88,140,250,322,187,253,198,262,222,186,141,300,64,65,254,98,263,62,109,111,138,41,99,335,15,17,188,18,86,30,298,311,89,150,301,87,40,190,94,6,16,251,29,63,201,14,227,310,110,76,139,223,152,302,274,78,74,58,38,165,92,104,189,312,117,95,200,41,28,105,264,75,47,313,55}	{47,252,102,312,62,305,53,211,240,193,274,54,254,52,38,188,76,88,275,98,104,136,28,60,90,51,4,207,210,163,2,42,100,94,245,138,199,15,44,3,7,12,40,150,162,46,194,316,139,151,64,6,198,87,39,322,169,43,311,103,250,78,58,202,114,299,115,217,119,55,33,281,178,107,154,191,226,105,323,50,71,263,91,140,306,142,22,30,16,82,310,127,95,302,266,18,146,321,57,66}	{252,102,47,54,211,2,112,275,42,163,210,254,4,94,90,52,40,88,140,298,227,53,322,100,58,30,312,116,199,51,142,38,167,311,151,198,87,250,230,228,3,274,14,107,238,302,86,300,59,152,108,43,118,66,18,114,98,115,263,251,190,262,200,15,91,109,323,16,5,242,64,50,11,6,46,179,226,264,266,65,150,139,145,110,314,99,103,157,95,257,169,92,206,22,219,78,186,293,26,281}

（续表）

数据集	RF	XGBoost	LightGBM
set_{10dB}	{28,26,4,74,5,2,27,29,40,3,76,110,39,89,77,75,38,86,88,87,53,41,52,111,15,50,34,112,17,100,62,98,99,14,101,16,109,64,51,65,94,63,218,226,64,321,229,282,42,225,264,271,154,25,312,115,213,22,294,262,170,258,211,165,228,32,167,35,215,326,116,103,331,246,330,323,146,276,90,241,314,165,212,126,46,254,293,182,54,114,234,102,261,163,317,269,162,307,31,169}	{4,3,26,38,86,52,2,200,39,310,14,40,114,99,75,260,28,16,66,300,226,311,100,228,74,62,227,98,298,251,238,92,174,210,131,32,34,154,220,276,312,322,280,156,118,225,178,230,239,42,128,299,44,130,246,19,293,91,262,88,48,22,324,244,198,331,172,51,180,284,233,47,137,139,133,215,94,212,305,167,23,102,15,257,148,196,144,175,211,72,202,264,214,182,104,10,57,76,82,164}	{4,26,38,34,3,28,40,52,74,86,15,88,76,110,51,99,321,50,14,16,294,100,87,39,2,215,75,273,178,42,27,106,249,82,203,211,206,111,155,258,239,114,115,307,209,137,119,218,116,225,250,326,170,200,330,191,46,109,259,228,279,105,98,175,11,242,79,31,161,10,297,293,164,154,327,6,234,97,73,230,25,113,134,256,22,150,306,226,270,195,246,318,176,47,146,202,251,163,312,95}
set_{5dB}	{3,4,2,5,89,38,53,40,50,109,39,17,52,51,112,76,88,74,86,123,273,16,22,209,29,58,15,87,124,321,77,110,91,14,197,28,99,42,306,309,100,27,41,43,261,330,118,119,10,94,64,315,101,54,148,139,75,70,258,98,26,65,82,198,106,146,111,1,73,184,97,181,147,85,267,113,47,150,281,130,282,158,218,90,314,191,232,172,143,274,194,227,30,34,63,325,182,104,80,125}	{53,2,86,127,133,3,153,138,198,27,276,50,63,38,87,81,103,80,4,52,71,286,220,240,175,150,84,329,217,157,272,163,179,242,76,42,33,292,228,243,73,205,92,176,288,28,57,9,141,159,246,232,294,332,307,328,184,259,142,70,270,287,283,114,261,139,31,274,154,60,72,130,186,49,250,215,78,23,11,164,39,51,56,278,236,162,321,200,165,161,54,132,68,16,234,113,147,17,64,247}	{2,3,58,52,243,258,38,91,321,86,51,50,261,273,242,118,97,209,194,42,73,95,267,6,330,142,25,106,206,119,270,297,278,22,225,4,171,78,7,294,43,218,202,123,234,306,26,16,113,271,312,140,146,207,150,303,226,251,185,203,46,84,315,178,190,307,27,158,249,1,31,15,81,94,299,211,39,256,133,137,219,275,183,2821,291,215,230,159,136,155,30,122,49,127,70,126,76,279,54,166}

4.4 第二次降维

将3个数据集分别使用3种嵌入式方法进行第一次降维并得到初选特征子集（见表2），将初选特征子集输入到第二次降维的6个特征选择模型（RF-SBS，RF-GBAS，XGBoost-SBS，XGBoost-GBAS，LightGBM-SBS，LightGBM-GBAS）中。6个模型的命名方式为“第一次降维使用的嵌入式方法-第二次降维使用的封装式方法”。其他对比方法为特征全集（336个特征）、使用嵌入式方法进行一次降维并使用序列后向搜索策略进行二次降维（DT-SBS，GBDT-SBS）的方法和改进的使用嵌入式方法进行一次降维并使用序列后向搜索策略进行二次降维（ACO_XGBoost-SBS[28]，ACO_LightGBM-SBS[29]）的方法、仅使用封装式方法进行一次降维（GBAS）的方法。

根据表2，分别对3个数据集使用3种嵌入式方法得到的初选特征子集均为100维。为进一步提高辐射源个体识别的准确率，减小特征子集规模并提高分类效率，使用封装式特征选择对初选特征集进行二次降维。搜索策略分别使用SBS和GBAS，其中，初始化GBAS中的参数，即 $\tau_{ij}(0)=1$，$\alpha=1$，$\beta=1$，$\rho=0.2$，$Q=0.02$，蚂蚁数 $M=45$，最大迭代次数 $Nc=100$。分类器使用LightGBM。在每类数据集下的6种混合式特征选择方法和GBAS方法中，分别令特征子集规模 q 为1～100，得到的最大分类正确率 A 和最小特征子集规模 q 如表3所列，并通过表3计算得到最优特征选择方法与其他方法性能的差值，如图5所示（图中对比方法中的“all”表示特征全集，其余对比方法的命名方式为取首字母）。

表3 第二次降维各模型的分类正确率和子集规模

Table 3 Classification accuracy and subset size of each model for the second dimensionality reduction

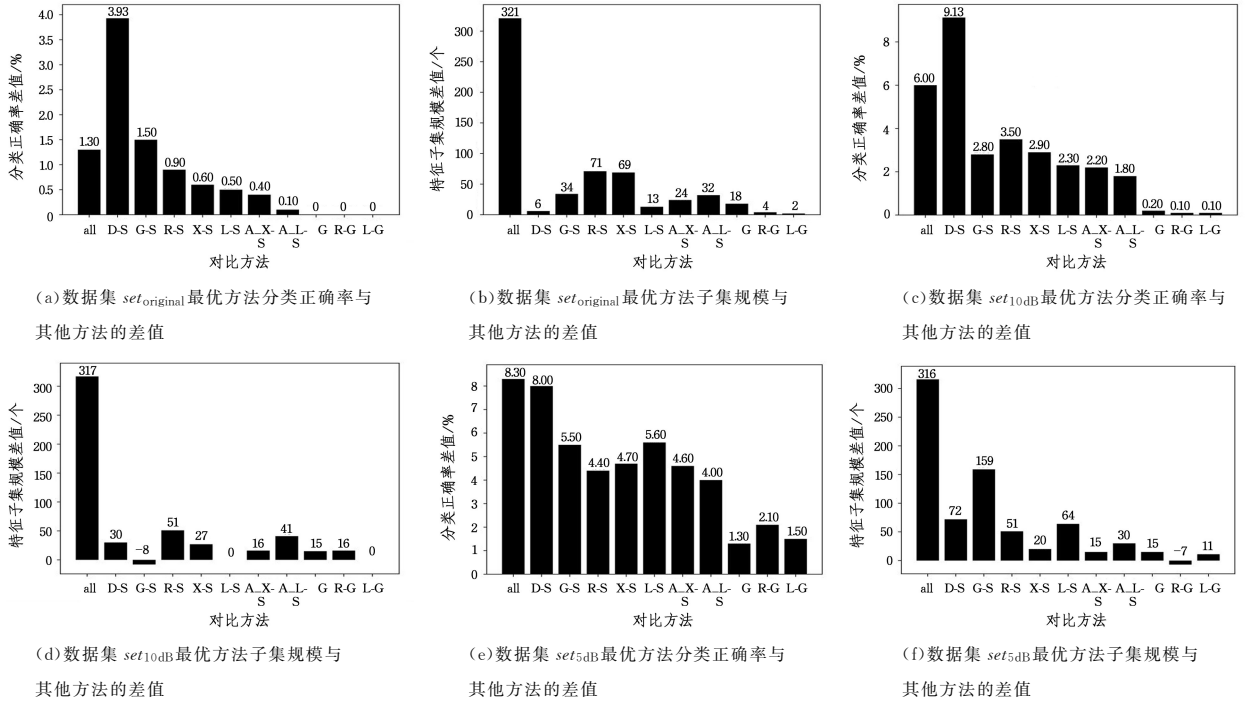
数据集	评估准则	全集	DT-SBS	GBDT-SBS	RF-SBS	XGBoost-SBS	LightGBM-SBS	ACO_XGBoost-SBS	ACO_LightGBM-SBS	GBAS	RF-GBAS	XGBoost-GBAS	LightGBM-GBAS
$set_{original}$	A/%	98.00	95.37	97.80	98.40	98.70	98.80	98.90	99.20	**99.30**	**99.30**	**99.30**	**99.30**
	q	336	21	49	86	84	28	39	47	33	19	**15**	17
set_{10dB}	A/%	69.20	66.07	72.40	71.70	72.30	72.90	73.00	73.40	75.00	75.10	**75.20**	75.10
	q	336	49	**11**	70	46	19	35	60	34	35	19	19
set_{5dB}	A/%	55.60	55.90	58.40	59.50	59.20	58.30	59.30	59.90	62.60	61.80	**63.90**	62.40
	q	336	92	179	71	40	84	35	50	35	**13**	20	31

表3和图5中，对比其他11种特征选择方法，使用特征全集进行分类得到的正确率最低，因此使用特征选择是必要的。合适的特征选择方法不仅可以提高辐射源个体的识别正确率，还可以缩小特征子集规模以达到提高模型效率的目的。

为解决封装式特征选择方法效率低的问题，首先使用嵌入式特征选择对特征集进行特征初选，使用初选后的特征子集进行分类，不仅提高了辐射源个体识别的分类正确率，还降低了特征维度，为二次降维缩小了搜索空间，进而提高了辐射源个体识别的效率。第二次降维使用封装式特征选择方法，搜索策略选用序列搜索策略中的序列后向搜索策略和随机搜索策略中基于蚁群算法的GBAS方法。

对于数据集 $set_{original}$，依据表3、图5(a)和图5(b)，最佳特征选择模型为XGBoost-GBAS，此时最大分类正确率 A 为99.30%，特征子集规模 q 为15。对比特征全集，XGBoost-GBAS方法分类正确率提高了1.30%，特征子集个数减少了321。对比二次降维使用序列后向搜索策略（DT-SBS，GBDT-SBS，RF-SBS，XGBoost-SBS，LightGBM-SBS，ACO_XGBoost-SBS，ACO_LightGBM-SBS），XGBoost-GBAS方法分类正确率分别提高了3.93%，1.50%，0.90%，0.60%，0.50%，0.40%和0.10%，特征子集个数分别减少了6，34，71，69，13，24和32。结合RF-GBAS和LightGBM-GBAS方法，可见使用SBS进行二次降维得到的最优辐射源个体识别正确率均低于使用GBAS进行二次降维得到的最优辐射源个体识别正确率，且使用SBS进行二次降维得到的最优特征子集规模

均大于使用 GBAS 进行二次降维得到的最优子集规模。对比仅使用封装式方法 GBAS 和使用两次降维的混合式特征选择方法，可以得到相同的最大分类正确率，而 XGBoost-GBAS 方法得到的最小特征子集维度比 GBAS 方法得到的最小特征子集维度减少了 18。由此可见，一次降维不仅缩小了搜索空间的大小，进而提高了分类效率，还为后续特征选择过程提供了特征重要性值较大的特征，进一步在降低维度的同时提高了分类正确率。对比 RF-GBAS 和 LightGBM-GBAS，XGBoost-GBAS 方法与这两种方法得到的最高分类正确率相同为 99.30%，特征子集个数分别减少了 4 和 2。



(a)数据集 $set_{original}$ 最优方法分类正确率与其他方法的差值

(b)数据集 $set_{original}$ 最优方法子集规模与其他方法的差值

(c)数据集 set_{10dB} 最优方法分类正确率与其他方法的差值

(d)数据集 set_{10dB} 最优方法子集规模与其他方法的差值

(e)数据集 set_{5dB} 最优方法分类正确率与其他方法的差值

(f)数据集 set_{5dB} 最优方法子集规模与其他方法的差值

图 5 最优特征选择方法与其他方法性能的差值

Fig. 5 Performance difference between the optimal feature selection method and other methods

对于数据集 set_{10dB}，依据表 3、图 5(c)和图 5(d)，最佳特征选择模型为 XGBoost-GBAS，此时最大分类正确率 A 为 75.20%，特征子集规模 q 为 19。对比特征全集，XGBoost-GBAS 方法分类正确率提高了 6.00%，特征子集个数减少了 317。对比二次降维使用序列后向搜索策略(DT-SBS，GBDT-SBS，RF-SBS，XGBoost-SBS，LightGBM-SBS，ACO_XGBoost-SBS，ACO_LightGBM-SBS)，XGBoost-GBAS 方法的分类正确率分别提高了 9.13%，2.80%，3.50%，2.90%，2.30%，2.20%和 1.80%，特征子集个数分别减少了 30、增加了 8、减少了 51，27，0，16 和 41。其中虽然 GBDT-SBS 方法得到的最优特征子集规模较 XGBoost-GBAS 方法的更小，但分类正确率远不如后者，故 XGBoost-GBAS 仍为最佳方法。结合 RF-GBAS 和 LightGBM-GBAS 方法，可见使用 SBS 进行二次降维得到的最优辐射源个体识别正确率均低于使用 GBAS 进行二次降维得到的最优辐射源个体识别正确率，且使用 SBS 进行二次降维得到的最优特征子集规模基本大于使用 GBAS 进行二次降维得到的最优子集规模。对比仅使用封装式方法 GBAS 和使用两次降维的混合式特征选择方法，XGBoost-GBAS 方法的分类正确率提高了 0.20%，特征子集个数减少了 15。对比 RF-GBAS 和 LightGBM-GBAS，XGBoost-GBAS 方法与这两种方法得到的最高分类正确率均提高了 0.10%，特征子集个数分别减少了 16 和 0。

对于数据集 set_{5dB}，依据表 3、图 5(e)和图 5(f)，最佳特征选择模型为 XGBoost-GBAS，此时最大分类正确率 A 为 63.90%，特征子集规模 q 为 20。对比特征全集，XGBoost-GBAS 方法分类正确率提高了 8.30%，特征子集个数减少了 316。对比二次降维使用序列后向搜索策略(DT-SBS，GBDT-SBS，RF-SBS，XGBoost-SBS，LightGBM-SBS，ACO_XGBoost-SBS，ACO_LightGBM-SBS)，XGBoost-GBAS 方法的分类正确率分别提高了 8.00%，5.50%，4.40%，4.70%，5.60%，4.60%和 4.00%，特征子集个数分别减少了 72，159，51，20，64，15 和 30。结合 RF-GBAS 和 LightGBM-GBAS 方法，可见使用 SBS 进行二次降维得到的最优辐射源个体识别正确率均低于使用 GBAS 进行二次降维得到的最优辐射源个体识别正确率，且使用 SBS 进行二次降维得到的最优特征子集规模基本大于使用 GBAS 进行二次降维得到的最优子集规模。对比仅使用封装式方法 GBAS 和使用两次降维的混合式特征选择方法，XGBoost-GBAS 方法的分类正确率提高了 1.30%，特征子集个数减少了 15。对比 RF-GBAS 和 LightGBM-GBAS，XGBoost-GBAS 方法与这两种方法得到的最高分类正确率分别提高了 2.10%和 1.50%，特征子集个数分别增加了 7 和减少了 11，其中虽然 RF-GBAS 方法得到的最优特征子集规模较 XGBoost-GBAS 方法更小，但分类正确率远不如后者，故 XGBoost-GBAS 仍为最优模型。

根据上述分析，将经由特征选择方法选出的特征输入分类器中比使用特征全集进行分类得到的分类正确率更高。

鉴于封装式特征选择方法效率低和嵌入式特征选择方法分类效果差等问题，本文提出的混合式方法有效解决了上述问题。使用嵌入式方法计算每一特征的重要性值并进行特征初选不仅提高了辐射源个体识别的分类正确率，还缩小了二次降维的搜索空间，进而提高了效率，二次筛选使用封装式方法。由实验可以证得，使用 GBAS 搜索策略比使用序列后向搜索策略得到的二次降维结果更好。混合式特征选择方法不仅比单一特征选择方法得到的分类正确率高，特征子集规模也进一步减少。

结束语 为解决嵌入式特征选择方法所得辐射源个体识别正确率低、封装式特征选择方法在识别辐射源个体时运算效率低的问题，提出了一种混合式特征选择方法。使用提升小波包变换提取特征并对特征值进行标准化，以最大分类正确率和最小特征子集规模为目标函数，建立了混合式特征选择的数学模型。首先使用 RF，XGBoost 和 LightGBM 这 3 种嵌入式方法对特征集中的所有特征计算重要性值，并按照重要性值进行排序，每种方法下选取前 100 个最重要的特征。然后将进行初选后的特征子集使用封装式特征选择方法进行二次降维，封装式方法中，搜索策略分别采用序列后向搜索策略和 GBAS，结合一次降维和二次降维，得到 6 种特征选择模型（RF-SBS，RF-GBAS，XGBoostSBS，XGBoost-GBAS，LightGBM-SBS，LightGBM-GBAS），对 3 个数据集分别使用 6 种模型进行特征选择。实验结果表明，相比特征全集，特征选择是必要的；相比单一嵌入式特征选择方法，混合式特征选择的分类正确率有较大的提升；相比单一封装式特征选择方法，混合式特征选择方法减小了搜索空间，提高了辐射源个体识别的效率。

本文针对基于混合式特征选择的辐射源个体识别问题做了一定的工作，但仍有许多问题需要进一步研究。下一步工作可以从以下几个方面展开：1）分类算法是辐射源个体识别的关键步骤，不同分类器的选择会影响最终识别的结果，下一步可以对比信号数据集在多个不同分类器上得到的分类结果，以确定最适合当前信号数据集的分类器；2）采集数据来源于两个电台，故本文属于二分类问题，而现实辐射源信号数据种类复杂，多分类问题将是之后研究的重点。

参 考 文 献

[1] ZHANG M，LUO Z H，HUANG J G，et al. A fingerprint extraction method based on I/Q imbalance[J]. Acta Electronica Sinica，2020，48(4)：717-722.

[2] TANG Z，LEI Y K. Method of individual communication transmitter identification based on maximum correntropy[J]. Journal on Communication，2016，37(12)：1-5.

[3] LIU J F，YU H Y，DU J P，et al. Specific emitter identification under dynamic noise based on domain adaptation[J]. Journal of Signal Processing，2021，37(6)：1000-1007.

[4] DASGUPTA A，DRINEAS P，HARB B，et al. Feature selection methods for text classification[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2007：230-239.

[5] HE T，HU J，XIA P，et al. Feature selection of Emg signal based on ReliefF algorithm and genetic algorithm[J]. Journal of Shanghai Jiao Tong University，2016，50(2)：204-208.

[6] WOZNICA A，NGUYEN P，KALOUSIS A. Model mining for robust feature selection[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2012：913-921.

[7] TANG C，ZHENG X，LIU X W，et al. Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection[J]. IEEE Transactions on Knowledge and Data Engineering，2022，34(10)：4705-4716.

[8] JOHANNES H，MARTIN P，KLAUS B，et al. Leveraging model inherent variable importance for stable online feature selection[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2020：1478-1488.

[9] LI H，PUN C M，XU P，et al. A hybrid feature selection based on a discrete artificial beecolony for Parkinson's diagnosis[J]. ACM Transactions on Internet Technology，2021，21(3)：1-22.

[10] SHARMA D K，VARSHNEY R P，GARY A，et al. Hybrid feature selection method with multi-objective grey wolf optimizer for high dimension data[C]// Proceedings of the 9th Conference on Computing for Sustainable Global Development (INDIACom). IEEE，2022：854-859.

[11] LU X Q，HE W D，LU Q Q，et al. Hybrid filter-wrapper feature selection using water wave optimization for financial crisis prediction in enterprises[C]// Proceedings of the 16th Conference on Intelligent System and Knowledge(ISKE). IEEE，2021：193-199.

[12] ALYASIRI O M，CHEAH Y，ABASI A K. Hybrid filter-wrapper text feature selection technique for text classification[C]// Proceedings of the International Conference on Communication and Information Technology(ICICT). IEEE，2021：80-86.

[13] XU Z Z，SHEN D R，NIE T Z，et al. Hybrid features election algorithm combining information gain ratio and genetic algorithm [J]. Journal of Software，2022，33(3)：1128-1140.

[14] MCNAMARA Q，VEGA A D L，YARKONI T. Developing a comprehensive framework for multimodal feature extraction [C]// International Conference on Knowledge Discovery and Data Mining(SIGKDD). ACM，2017：1567-1574.

[15] CAO J J，ZHANG P L，ZHANG Y T，et al. Feature extraction of an engine cylinder head vibration signal based on lifting wavelet package transformation[J]. Journal of Vibration and Shook，2008，27(2)：34-37.

[16] XU Z L，JIN R，YE J P，et al. Non-monotonic feature selection [C]// Proceedings of the 26th Annual International Conference on Machine Learning. ACM，2009：1145-1152.

[17] JUN Y J，LEAU Y，ALIAS S，et al. A multi-filter feature selection in detecting distributed denial-of-service attack[C]// Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering. ACM，2019：57-62.

[18] FAN X，FENG Z Q，YANG X H，et al. Hazw weather recognition based on multiple features and random forest[C]// Procee-

dings of International Conference on Security, Pattern Analysis, and Cybernetics. IEEE, 2018: 485-488.

[19] FENG D D, DENG Z F, WANG T X, et al. Identification of disturbance sources based on random forest model[C]// Proceedings of International Conference on Power System Technology. IEEE, 2018: 3370-3375.

[20] ZHAI Y B, ZHENG X H. Random forest traffic classification method in SDN[C]// Proceedings of International Conference on Cloud Computing, Big Data and Blockchain. IEEE, 2018: 1-5.

[21] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[22] LI L, LI C, WU Y, et al. Spectroscopy-based food internal quality evaluation with XGBoost algorithm[C]// APWeb-WAIM 2018. Springer, 2018: 56-64.

[23] LI Z S, YAO X, LIU Z G, et al. Feature selection algorithm based on LightGBM[J]. Journal of Northeastern University (Natural Science), 2021, 42(12): 1688-1695.

[24] YUAN J, RAO Z, LIN H, et al. Classification of Chinese dialect regions from L2 English speech[C]// International Conference on Acoustics, Speech and SP(ICASSP). IEEE, 2019: 8117-8121.

[25] SHEN S, QIAN Y, ZHENG J, et al. Accurately predicting circRNA-disease associations using variational graph auto-encoders and LightGBM[C]// International Conference on Bioinformatics and Biomedicine(BIBM). Piscataway, IEEE, 2021: 522-527.

[26] CAO J J, ZHANG P L, WANG Y X, et al. Graph-based ant system for subset problems[J]. Journal of System Simulation, 2008, 20(22): 6146-6150.

[27] CHEN Z, LV N. Network intrusion detection model based on Random Forest andXGBoost[J]. Journal of Signal Processing, 2020, 36(7): 1055-1064.

[28] CAO J J, GU C M, WANG B W, et al. Specific emitter identification based on ACO-XGBoost [C]// APWeb-WAIM 2022. Springer, 2023: 76-90.

[29] GU C M, CAO J J, WANG B W, et al. Specific emitter identification of LightGBM based on colony parameters optimization[J]. Journal of Computer Engineering and Science, 2023, 45(1): 9-18.

GU Chumei, born in 1997, postgraduate, is a member of CCF(No. I7490G). Her main research interests include intelligent data analysis and application.

CAO Jianjun, born in 1975, Ph.D, associate researcher, is a member of CCF (No. 13414S). His main research interests include data quality control and data intelligent analysis.

(责任编辑:喻藜)