◆交通场景◆

基于特征选择和聚类的动态选择性集成模型*

徐雨芯[1,2],曹建军[1**],王保卫[2],翁年凤[1],顾楚梅[1,2]
(1. 国防科技大学第六十三研究所,江苏南京 210007; 2. 南京信息工程大学计算机与软件学院,江苏南京 210044)

摘要:为提高辐射源个体识别的准确率,降低动态选择性集成的计算复杂度,本文提出基于特征选择和聚类的动态选择性集成模型(FSC-DES)。利用归一化皮尔森相关系数法度量不同基分类器间混淆矩阵的差异性,以各基分类器准确率最高及基分类器间差异性最大为目标,得到基分类器集合和对应特征子集集合。利用聚类方法将验证集划分为若干类,以验证集分类准确率最高为目标,为每簇验证集选择最优的基分类器子集和对应的特征子集。在测试阶段,对测试集进行聚类,仅比较每簇测试样本和每簇验证样本数据分布的最大均值差异值,减少运算时间。每簇测试样本在相似度最高的验证集所对应的特征子集集合和基分类器子集下进行预测,并根据不同权重基分类器预测结果的加权和进行最终决策。为验证方法的必要性和优越性,将本文方法与传统集成学习方法进行对比,结果表明,本文方法在信噪比分别为 10、5 dB 的条件下,分类准确率均提升约 5%,具有更好的分类效果和泛化性能。

关键词:特征选择;动态选择性集成;支持向量机;蚁群优化算法;辐射源个体识别;二分类问题

中图分类号:TP391 **文献标识码**:A **文章编号**:1005-9164(2024)05-1002-09
DOI:10.13656/j.cnki.gxkx.20241122.001

电子技术在现代战争中具有显著优势。在目前的实战场景中,电子设备的应用更加广泛,电子设备体制复杂,且相互之间的干扰使得所处的电磁环境更加复杂。在复杂的电磁环境中,不仅需要对电子设备进行侦查,还需要对其进行精准打击和信号干扰,即电子对抗和电子反对抗。随着国际关系的演变和各国军事的发展,各种电台、雷达是战场环境中指挥、控制、通信、情报探测、电子监控等军事活动的基础前端和数据来源,具有举足轻重的地位[1]。利用信号处理技术对所截获的辐射源信号进行分析,提取其中包含个体信息的特征,并对辐射源个体进行区分识别显得尤为重要。

近年来,深度学习在特定辐射源识别(Specific Emitter Identification,SEI)领域有广泛的发展和应用[2],深度神经网络能够通过具有非线性激活函数的多个隐含层来检索抽象特征,有利于提取辐射源信号

收稿日期:2022-10-06 修回日期:2022-10-20
*国家自然科学基金项目(61371196),中国博士后科学基金特别资助项目(2015M582832)和国家重大科技专项(2015ZX01040201-003)资助。
【第一作者简介】
徐雨芯(1998—),女,在读硕士研究生,主要从事智能数据分析与应用研究,E-mail:2801343036@qq.com。
【**通信作者简介】
曹建军(1975—),男,副研究员,硕士研究生导师,主要从事智能数据分析与应用研究,E-mail:caojj@nudt.edu.cn。
【引用本文】
徐雨芯,曹建军,王保卫,等. 基于特征选择和聚类的动态选择性集成模型[J]. 广西科学,2024,31(5):1002-1010.
XU Y X,CAO J J,WANG B W,et al. A Dynamic Ensemble Selection Model Based on Feature Selection and Clustering [J]. Guangxi Sciences,2024,31(5):1002-1010.

深层次的细微特征[3]。因此,Tu 等[4]和 Merchant 等[5]使用基于深度神经网络的端到端的 SEI 方法将复数信号的同相和正交分量(In-phase and Quadrature components,IQ)组成二维矩阵,在二维矩阵上训练实值神经网络(Real-Valued Neural Network,RVNN)。然而,RVNN 不能直接处理复杂的基带信号,通常会忽略 IQ 之间的相关性,导致识别性能下降。Wong 等[6]和 Gong 等[7]将原始信号的同相/正交(I/Q)数据直接作为卷积神经网络的输入,完成 SEI 任务。何遵文等[8]提出一种通信辐射源个体识别的自编码器构造方法,提高了通信辐射源个体识别任务中自编码器的性能。这类方法利用神经网络端到端的特性,具有一定的整体性,但过于依赖神经网络的设计,并且对不同的原始数据类型要设计不同的神经网络,可扩展性不强。另外,常用的卷积神经网络更擅长识别二维图像数据,直接处理 I/Q 数据效果会有所降低。

因此,更有效的方案是将 SEI 任务分为两步,一是数据的预处理和特征提取,二是分类器的设计和训练。由于从不同角度提取的辐射源个体特征具有差异,单个分类算法无法对所有特征做出较好的分类效果,在分类过程中使用单一分类器得到的分类准确率效果欠佳。通过将多个具有差异性的分类器组合,使组合分类器学习多样化,可以获得比使用单个分类器更好的预测结果[9,10]。

集成学习主要通过助推(Boosting)法、袋装(Bagging)法或堆叠(Stacking)法将多个弱学器融合,通过平均、加权、最值等方法对各基分类器(Support Vector Machine,SVM)的输出结果进行整合得到一个强学习器。选择性集成方法在前者基础上增加了基学习器的筛选阶段,通过静态或动态选择方法获得比使用全部基分类器集成效果更好的基分类器集成子集。静态选择方法主要分为基于排序的方法和基于优化的方法。基于排序的方法主要是通过验证集错误率、边距等准则对基分类器进行排序,仅选择值最优的基分类器输入最终集成子集中,如使用多样性和边距的评估方法对分类器进行排序。基于优化的方法将集合选择视为优化问题,通过启发式优化或数学规划来解决,如 Zhou 等[11]提出基于遗传算法的选择性集成学习算法(Genetic Algorithm Based Selective Neural Network Ensemble,GASEN),为每个基分类器指派一个权重,通过遗传算法对权重进行迭代计算,权重高于阈值的基分类器予以保留,仅使用保留下来的基分类器进行集成。

动态选择方法考虑到每个样本间的差异性,根据样本的不同特点,寻找待测样本的局部相似样本,动态挑选最适合待测样本的基学习器进行预测,以提高性能,减少基学习器间的冗余。相比静态选择方法,动态选择方法能够获得更好的分类性能,因此动态选择方法受到越来越多的学者关注。如何为待测样本确定相似邻域并选择集成效果最优的基分类器子集是动态选择方法的关键。Ko 等[12]提出 k-Nearest-ORAcles(kNORA)方法,将分类结果的精确性作为基分类器选择的标准。对每一个测试样本,先寻找到其 k 个近邻作为验证样本,然后选出能够将该 k 个样本正确分类的基分类器子集,将其作为该测试样本的基分类器集合。这类方法针对每个测试样本的预测,都需在整个训练样本集或验证样本集中遍历,寻找其 k 个近邻,时间开销较大[13]。

针对以上问题,本文提出基于特征选择和聚类的动态选择性集成模型(Dynamic Ensemble Selection model based on Feature Selection and Clustering,FSC-DES)。以各基分类器准确率最高及基分类器间差异性最大为目标进行特征选择,得到特征子集集合和对应基分类器集合。利用聚类方法将验证集划分为若干类,以验证集分类准确率最高为目标,为每簇验证集选择最优的特征子集及对应的基分类器子集。在测试阶段,对测试集进行聚类,仅比较每簇测试样本和每簇验证集间的最大均值差异值,每簇测试样本在相似度最高的验证集所对应的特征子集集合和基分类器子集下进行预测,降低相似度计算的对比次数,减少运算时间,根据不同权重基分类器预测结果的加权和进行最终决策,提高分类准确率。

1 基于特征选择的动态选择性集成模型

1.1 基分类器池模型设计

集成学习思想要求组合分类器中,各基分类器有较高的准确率且基分类器之间具有高的差异性,即组合分类器中的各基分类器分类准确率越高,同时它们之间输出的差异性越大,则此组合分类器的分类性能越好[9]。常采用 Bagging 法对样本或特征进行扰动,从而生成各基分类器。然而基于 Bagging 法生成的基分类器池,缺少对在不同特征子集下样本分布差异性和分类准确率的考虑,导致集成模型的分类准确率受到一定限制。

本文从特征选择的角度给出一种新的基分类器

设计方法。在不同训练样本集下,以基分类器分类准确率最高、基分类器输出的混淆矩阵差异性最大为目标选择不同的特征子集,生成 $N(N \in Z^+)$个高差异性、高准确率的基分类器及对应的最优特征子集。

记特征集大小为 $Q(Q \in Z^+)$,第 $n(1 \leqslant n \leqslant N, n \in Z^+)$个基分类器所选的特征子集为 s_n,其大小为 $q(1 \leqslant q \leqslant Q)$,对应样本集为 x_n,分类准确率为 $\mathrm{Acc}(\mathrm{SVM}_n(x_n, s_n))$,输出的混淆矩阵为 $\boldsymbol{M}$,记第 n 个混淆矩阵为 $\boldsymbol{M}_n$,目标函数设计如下:

$$\max \mathrm{Acc}(\mathrm{SVM}_n(x_n, s_n)), \tag{1}$$

$$\max[1 - \max_{j=1}^{n-1} S_c(\boldsymbol{M}_n, \boldsymbol{M}_j)], \tag{2}$$

$$\min q, \tag{3}$$

$$\text{s.t. } 1 \leqslant q \leqslant Q, 1 \leqslant n \leqslant N, q, n \in Z^+。 \tag{4}$$

式(1)目标为在当前特征子集 s_n 下,基分类器 SVM_n 对样本集 x_n 有最高分类准确率;式(2)采用归一化皮尔森相关系数法度量不同基分类器间混淆矩阵的相似性,比较当前分类器与前 $n-1$ 个分类器的差异性,选择使得第 n 个分类器与其他 $n-1$ 个分类器之间具有最大差异性的特征子集,即互补特征子集,从而最大化分类器的多样性,其中皮尔森相关系数公式见式(5)至式(7);式(3)目标为当不同特征子集下目标函数值均相同,选择基数较小的特征子集。

$$S_c(V_1, V_2) = \frac{\sum_{i=1}^{2}\sum_{i'=1}^{2}(V_{1,ii'} - \bar{V}_1)(V_{2,ii'} - \bar{V}_2)}{\sqrt{\sum_{i=1}^{2}\sum_{i'=1}^{2}(V_{1,ii'} - \bar{V}_1)^2 \sum_{i=1}^{2}\sum_{i'=1}^{2}(V_{2,ii'} - \bar{V}_2)^2}} + 1}{2}, \tag{5}$$

$$\bar{V}_1 = \frac{1}{4}\sum_{i=1}^{2}\sum_{i'=1}^{2}V_{1,ii'}, \tag{6}$$

$$\bar{V}_2 = \frac{1}{4}\sum_{i=1}^{2}\sum_{i'=1}^{2}V_{2,ii'}。 \tag{7}$$

式(5)中,V_1、V_2 分别为两个不同基分类器的混淆矩阵。由式(5)可知 $S_c \in [0,1]$,当 $S_c = 0$ 时,表示 V_1、V_2 全负相关,即两个基分类器的输出分布矩阵完全相反,分类结果完全不同,两个基分类器的差异性最大;当 $S_c = 1$ 时,表示 V_1、V_2 完全正相关,即两个基分类器的输出分布矩阵完全相同,分类结果完全一致,两个基分类器的差异性最小。

1.2 动态选择性集成模型设计

通过 1.1 节基分类器池模型设计得到一组高差异性、高准确率的基分类器集合和对应特征子集集合。基于特征选择的动态选择性集成模型就是从基分类器集合中,为每个测试样本选择部分基分类器和对应特征子集,构成基分类器子集和特征子集集合,以此获得更好的预测结果。

考虑到聚类就是将样本分为多簇,同簇样本的数据分布具有较高的相似度,因此若先将样本聚为 $k(3 \leqslant k \leqslant 10)$簇,对同簇样本选用相同的基分类器组合及其对应的特征子集,可以大幅减少对比次数,提高分类效率。

为避免过拟合,在给定验证集下,以分类准确率最高为目标,设计动态选择性集成模型。具体过程如下:记第 $m(1 \leqslant m \leqslant k, m \in Z^+)$簇验证样本 Z_m 在指定 $T(1 \leqslant T \leqslant N, T \in Z^+)$个基分类器下,组成的组合分类器及对应特征子集分别为 C_m^T、S_m^T,以分类准确率最高为目标,得到每簇验证集下的组合基分类器。目标函数如下:

$$\max \mathrm{Acc}(C_m^T(Z_m, S_m^T)), \tag{8}$$

$$\text{s.t. } 1 \leqslant m \leqslant k, 1 \leqslant T \leqslant N, m, T \in Z^+。 \tag{9}$$

当组合基分类器个数 T 大于 1 时,需对各基分类器输出结果进行集成。例如,在现实生活中,某些议题需要专家委员会做出决定,但不同的专家(不同的基分类器)对相同问题(相同样本)有不同的知识背景和专业水平。当我们知道一位专家在某一特定领域非常博学时(分类准确率高),会相信这位专家的建议,即便他对当前的建议不完全有信心(置信度相对低)。另一方面,当某个专家的知识相对不够丰富(分类准确率低)时,我们会在他非常确定(置信度高)的情况下考虑其目前的建议[14]。因此,本文以基分类器分类准确率为权重,基于各基分类器分类置信度的加权和做最终预测。具体过程如下:记第 $t(1 \leqslant t \leqslant T)$个基分类器分类准确率为 Acc_t,权重为 w_t,对第 i 个样本预测结果为 $y_t^i \in \{1, -1\}$,预测为 1 类的概率为 $p_t^i \in [0,1]$。组合基分类器将第 i 个样本预测为 1 类的概率为 P_i,预测结果为 Y_i,采用式(10)至式(12)对不同权重的各基分类器输出结果进行集成:

$$w_t = \mathrm{Acc}_t / \sum_{j=1}^{T}\mathrm{Acc}_j, \tag{10}$$

$$P_i = \sum_{t=1}^{T} w_t \times p_t^i, \tag{11}$$

$$Y_i = \begin{cases} 1, P_i \geqslant 0.5 \\ -1, P_i < 0.5 \end{cases}。 \tag{12}$$

式(10)为当前基分类器的准确率与所有基分类器准确率之和的比值,表示第 t 个基分类器的权重。式(11)表示 T 个不同权重的基分类器预测第 i 个测

试样本为第1类的置信度。式(12)为组合基分类器对第 i 个样本的最终预测值 Y_i，当预测为1类的置信度 P_i 大于0.5时，第 i 个样本被分为第1类；反之，分为第2类。

在该模型下可以得到每簇验证集最好的基分类器子集和对应的特征子集集合。当对测试样本进行预测时，同样先将测试样本聚为 k 类，在每簇测试样本下，比较其与各簇验证集的数据分布相似度，选择分布相似度高的验证集所对应的基分类器子集和特征子集集合对其进行预测，并采用式(10)至式(12)对各不同权重的基分类器输出结果进行集成，可以有效减少对比次数和运算时间。

衡量数据的分布差异一般通过数据的一阶、二阶统计信息即均值、方差来衡量，而低阶的统计信息不足以完全表达出数据分布的信息[15]。最大均值差异(Maximum Mean Discrepancy，MMD)是一种非参数度量，它通过将数据分布投影到再生核希尔伯特空间来计算它们之间均值(高阶统计量)的差距[16]，本文选用MMD方法来计算测试集和验证集间的相似度。记 $\phi X\rightarrow H$ 是从特征空间到再生核希尔伯特空间的投影，X、Y 表示两个数据集，两个数据集之间的MMD定义为

$$D(X,Y)=\left\| \frac{1}{l_1}\sum_{x\in X}\varphi(x)-\frac{1}{l_2}\sum_{y\in Y}\varphi(y)\right\|_H, \tag{13}$$

式中，l_1、l_2 分别表示两个数据集 X、Y 的大小，x、y 分别表示数据集中的某一样本。当两个数据集 X、Y 属于同一数据分布时，$D=0$。

2 模型设计求解

蚁群优化(Ant Colony Optimization，ACO)算法是受自然界蚂蚁觅食行为启发而广泛应用的一种启发式算法，该算法的优点主要是信息正反馈、较强鲁棒性及并行分布式计算等。最早被用于解决旅行商问题，随后其他组合优化问题如背包问题和特征选择问题也能运用蚁群优化算法得到解决。Gretton等[17]提出一种基于图的蚂蚁系统(Graph-Based Ant System，GBAS)，该算法基于构造图提出等效路径的概念，将问题的无序信息和有向图的路径相结合，实现将无序信息转化为有序信息，提高蚁群优化算法的性能。利用蚁群优化算法对模型求解，分析如下。

①曹建军[18]基于提升小波包分解与重构算法，提取原时域信号的12个统计特征参数，小波包分解第2层4个节点系数的各12个特征参数，4个单支重构信号的各12个统计特征参数，及4个标准化相对能量，共112个特征参数。为获得更全面的信号特征，在复数信号的幅值、I路、Q路3种形式下，分别提取以上112个特征，共336个特征，即 $Q=336$。

②在生成基分类器池前，需要设定基分类器的个数 N。Roy等[19]基于从分类问题中提取复杂性度量来预测基分类器池的最佳大小，结果表明，使用平均大小为20的基分类器池，选择性集成分类结果表现更好。因此本文基分类器个数 N 设定为20。

③对二分类器而言，特征子集大小 q 在5－10具有较好的运算效率和分类精度，为了不丢失边缘解，将 q 值的搜索范围限定在1－15。

④在特征子集基数 q 确定的情况下，当求解第一个基分类器时，式(2)并不存在，因此可以直接使用式(1)作为目标函数。当基分类器个数大于1时，将式(1)、式(2)加权求和转化为单目标优化函数，如式(14)所示。

$$\begin{aligned}&\max\{r_1\mathrm{Acc}(\mathrm{SVM}_n(x_n,s_n))+r_2[1-\\&\max_{j=1}^{n-1}S_c(\boldsymbol{M}_n,\boldsymbol{M}_j)]\}\\&r_1>0,r_2>0,r_1+r_2=1,\end{aligned} \tag{14}$$

式中，r_1 与 r_2 是聚合参数，在为第 n 个基分类器选择大小为 q 的特征子集过程中，通过聚合参数控制当前基分类器的分类准确率和基分类器间差异性之间的平衡，本文设定 $r_1=0.8$，$r_2=0.2$。

⑤为分析模型中超参数 k 的值对实验结果的影响，在电台原始采集信号下，进行不同 k 值实验结果的对比分析(图1)。当 $k=6$ 时，分类准确率较高，因此本文设定 k 值为6。

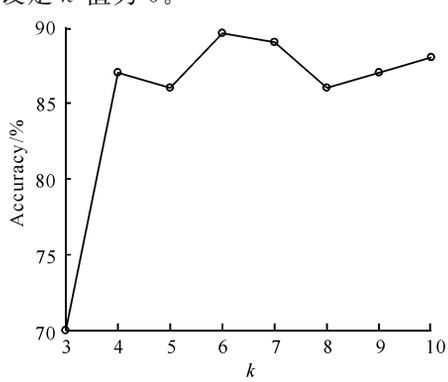

图1 不同 k 值下的准确率

Fig. 1 Accuracy under different values of k

根据以上分析，设计的模型求解算法描述如下：

算法1 动态选择性集成模型求解

输入：信息素重要程度值 $\alpha=1$，启发式信息重要

程度值 $\beta=1$,当前蚂蚁编号为 ant,迭代蚂蚁个数 Ant=150,蚁群优化算法当前迭代次数为 ite,最大迭代次数 Ite=100;训练集、验证集、基分类器个数 $N=20$,聚类个数 $k=6$;

输出:最优基分类器子集和对应特征子集集合。

① FOR $1\leqslant n\leqslant N$ DO

② 初始化蚁群优化算法信息素矩阵、启发式信息;

③ FOR $1\leqslant q\leqslant 15$ DO

④ WHILE ite < Ite DO

⑤ FOR $1\leqslant$ ant $\leqslant$ Ant DO

⑥ 从特征集合中搜索大小为 q 的特征子集;

⑦ END FOR

⑧ 按分析④确定指定子集大小下的第 n 个基分类器的特征子集;

⑨ 更新信息素矩阵;

⑩ END WHILE

⑪ END FOR

⑫ 比较不同子集大小下特征子集的分类准确率,选择分类结果最优的特征子集并将其输入特征子集集合中;

⑬ 将最优特征子集对应的基分类器输入基分类器池中;

⑭ END FOR

⑮ 得到基分类器池和对应的特征子集集合;

⑯ 采用均值聚类方法将验证集聚为 k 类;

⑰ FOR $1\leqslant m\leqslant k$ DO

⑱ 初始化蚁群优化算法信息素矩阵、启发式信息;

⑲ FOR $1\leqslant T\leqslant N$ DO

⑳ WHILE ite < Ite DO

㉑ FOR $1\leqslant$ ant $\leqslant$ Ant DO

㉒ 从基分类器池中搜索大小为 T 的基分类器子集;

㉓ 得到基分类器子集及其对应的特征子集集合;

㉔ 按式(10)至式(12)计算每簇验证样本分类准确率;

㉕ END FOR

㉖ 以式(8)为目标确定大小 T 下的基分类器子集和对应的特征子集集合;

㉗ 更新信息素;

㉘ END WHILE

㉙ END FOR

㉚ 比较不同基分类器子集大小下的分类准确率,选择最优基分类器子集和对应的特征子集集合;

㉛ END FOR

㉜ 得到每簇验证集最优基分类器子集及其特征子集集合。

3 验证实验

3.1 实验数据准备及评价指标

为说明方法的有效性,本文从两个电台中真实采集信号数据,两个电台工作性能、采样参数及各参数采样数均相同,如表 1 所示。

表 1 电台参数设置

Table 1 Radio parameter setting

载波/MHz Carrier/MHz	调制方式 Modulation	信号带宽 Signal bandwidth	采样频率/MHz Sampling frequency/MHz	间隔时间/ms Interval time/ms	采样数 Number of samples
55	QPSK	25 kHz	1	20	200
75	QPSK	25 kHz	1	20	200
420	QPSK	2 MHz	20	2	200
420	QPSK	5 MHz	50	1	200
420	QPSK	10 MHz	80	1	200
420	QPSK	20 MHz	100	2	200
2 000	QPSK	2 MHz	20	2	200
2 000	QPSK	5 MHz	50	1	200
2 000	QPSK	10 MHz	80	1	200
2 000	QPSK	20 MHz	100	1	200

两个电台采样数共2×2 000组,将其中3 000组数据作训练集,500组数据作测试集,500组数据作验证集。实验配置如下:CPU Intel E5-2609 V2×2、内存64 GB、硬盘960 GB、系统CentOS 7,运用Python 3.8和Pycharm进行编程。

实验采用分类准确率(Accuracy)、样本平均测试时间(Average test time)、接受者操作特征(Receiver Operating Characteristic,ROC)曲线及曲线下面积(Area Under Curve,AUC)进行分析。

$$\text{Accuracy}=\frac{\text{正确分类测试样本数}}{\text{测试样本总数}},\tag{15}$$

$$\text{Average test time}=\frac{\text{测试总时间}}{\text{测试样本数}}。\tag{16}$$

ROC曲线是针对二分类模型的一种坐标图示的分析工具,将二分类结果定义为阳性(Positive)和阴性(Negative)。X轴定义为伪阳性率(False positive rate),即在所有实际为阴性的样本中,被错误地判断为阳性的比率;Y轴定义为正阳性率(True positive rate),即在所有实际为阳性的样本中,被正确地判断为阳性的比率。AUC表示ROC曲线下的面积,AUC值越大,模型表现越好。

3.2 消融实验

在电台原始采集信号下,分别将本文方法(FSC-DES)与表2中方法1、方法2、方法3的分类准确率和分类性能进行对比,多次实验取均值,结果如表3所示。

表2 消融实验设置

Table 2 Ablation experiments setting

方法 Method	特征选择 Feature selection	静态集成 Static integration	动态选择 Dynamic selection	聚类 Clustering
Method 1	√			
Method 2	√	√		
Method 3	√		√	
FSC-DES	√		√	√

从表3可以看出,方法1仅采用特征选择在单一支持向量机下分类,该方法的分类准确率为86.0%;方法2在方法1的基础上结合静态选择性集成思想,即对所有测试样本均选择相同的基分类器集合,该方法的分类准确率为88.6%,可以看出加入集成学习思想后,分类准确率得到了提升;方法3相较方法2,采用了动态选择方法,但未对测试样本采用聚类方法,即方法3对每个测试样本都需寻找相似度最高的k个验证集,并利用其对应的基分类器子集和特征子集进行分类,该方法的分类准确率为89.2%,可以看出采用动态选择方法的分类准确率较静态选择有一定的提升,但是样本平均测试时间相差过大,运行速度较慢;而FSC-DES模型由于先对验证集和测试集分别进行了聚类,极大减少了测试样本的相似度计算对比次数,样本平均测试时间由方法3的0.083 s降低至0.006 s,分类准确率较方法2、方法3分别提高了1.0%、0.4%。可见,FSC-DES模型在提高分类准确率的同时,提升了运行速度。

表3 消融实验结果

Table 3 Results of ablation experiments

方法 Method	分类准确率/% Classification accuracy/%	平均测试时间/秒 Average test time/s
Method 1	86.0	—
Method 2	88.6	0.004
Method 3	89.2	0.083
FSC-DES	89.6	0.006

3.3 多个集成算法对比

为验证本文提出的FSC-DES模型的有效性,人为添加信噪比分别为10 dB和5 dB的高斯白噪声,在多个集成算法,包括梯度决策提升树(Gradient Boosting Decision Tree,GBDT)、随机森林(Random Forest,RF)、eXtreme Gradient Boosting(XGboost)、Light Gradient Boosting(LightGBM)进行对比实验。

分类准确率对比结果如图2所示。在10 dB信噪比数据下,FSC-DES模型的分类准确率较RF、GBDT、XGboost、LightGBM分别提高3.8%、5.4%、5.8%、4.4%。在5 dB信噪比数据下,FSC-DES的分类准确率较RF、GBDT、XGboost、LightGBM分别提高5.2%、7.6%、9.4%、8.6%。

最后,为了更直观地比较不同算法在相同数据集下的分类能力,使用ROC曲线来评估预测模型的好坏,结果如图3、图4所示。

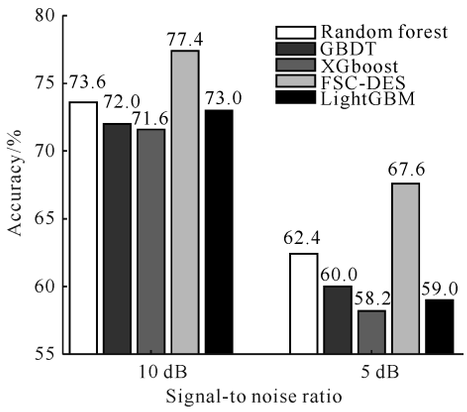

图 2　分类准确率对比结果

Fig. 2　Comparison results of classification accuracy

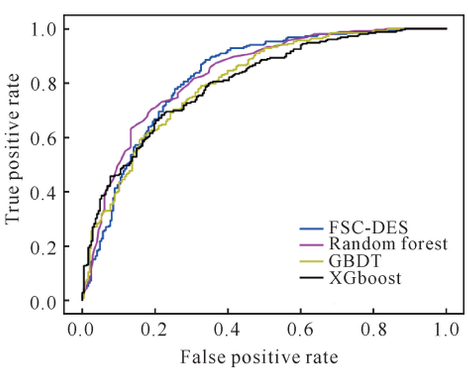

图 3　10 dB 下 ROC 曲线对比图

Fig. 3　Comparison results of ROC curve under 10 dB

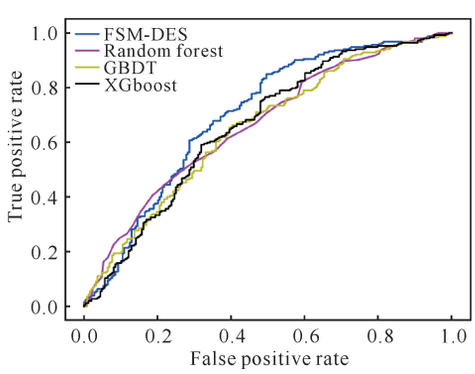

图 4　5 dB 下 ROC 曲线对比图

Fig. 4　Comparison results of ROC curve under 5 dB

由图 3、图 4 可知,在不同信噪比下,FSC-DES 模型的 ROC 曲线均比其余方法更靠近左上角。AUC 值对比结果如表 4 所示,FSC-DES 模型的 AUC 值,即 ROC 曲线下的面积值,也均大于其余方法。由此可知,本文提出的 FSC-DES 模型在不同信噪比数据下均具有更好的分类效果。

表 4　AUC 值对比结果

Table 4　Comparison results of AUC value

方法 Method	信噪比 Signal-to-noise ratio	
	10 dB	5 dB
Random forest	0.80	0.64
GBDT	0.78	0.65
XGboost	0.79	0.62
FSC-DES	0.82	0.70

4　结论

近年来,机器学习技术在辐射源个体识别领域取得了一定的成果。其中,集成学习由于利用多个分类器解决同一个问题,显著提高了泛化能力,成为机器学习领域的一个热点。

本文提出的 FSC-DES 模型,结合特征选择方法设计基分类器生成模型,提高了特征利用率和基分类器间的差异性。同时结合聚类思想,提高了动态选择性集成过程中的时间性能。在真实采集数据下的对比实验可以看出本文方法的有效性和必要性。

目前,FSC-DES 模型主要针对二分类识别问题,后续研究可以由二分类扩展到多分类,提高模型适用性。同时,模型在基分类器训练过程中,由于特征空间较大,采用蚁群优化算法进行求解时训练时间较长,后续工作可以着眼于缩短训练时间。

参考文献

[1]　孙丽婷,黄知涛,王翔,等. 辐射源指纹特征提取方法述评[J]. 雷达学报,2020,9(6):1014-1031.

[2]　周鑫,何晓新,郑昌文. 基于图像深度学习的无线电信号识别[J]. 通信学报,2019,40(7):114-125.

[3]　黄健航,雷迎科. 通信辐射源个体识别的自编码器构造方法[J]. 火力与指挥控制,2018,43(11):108-112.

[4]　TU Y,LIN Y,HOU C B,et al. Complex-valued networks for automatic modulation classification [J]. IEEE Transactions on Vehicular Technology,2020,69(9):10085-10089.

[5]　MERCHANT K,REVAY S,STANTCHEV G,et al. Deep learning for RF device fingerprinting in cognitive communication networks [J]. IEEE Journal of Selected Topics in Signal Processing,2018,12(1):160-167.

[6]　WONG L J,HEADLEY W C,MICHAELS A J. Specific emitter identification using convolutional neural network-based IQ imbalance estimators [J]. IEEE Ac-

cess,2019(7):33544-33555.

[7] GONG J L,XU X D,QIN Y F,et al. A generative adversarial network based on framework for specific emitter characterization and identification [C]//2019 11th International Conference on Wireless Communications and Signal Processing. Piscataway,NJ:IEEE,2019:1-6.

[8] 何遵文,侯帅,张万成,等.通信特定辐射源识别的多特征融合分类方法[J].通信学报,2021,42(2):103-112.

[9] 毕凯,王晓丹,姚旭,等.一种基于 Bagging 和混淆矩阵的自适应选择性集成[J].电子学报,2014,42(4):711-716.

[10] 陶晓玲,亢蕊楠,刘丽燕.基于选择性集成的并行多分类器融合方法[J].计算机工程与科学,2018,40(5):787-792.

[11] ZHOU Z H,WU J X,JIANG Y,et al. Genetic algorithm based selective neural network ensemble [C]// Proceedings of the 17th international joint conference on artificial intelligence. San Francisco, CA: Morgan Kaufmann,2001:797-802.

[12] KO A H R,SABOURIN R,BRITTO A S. From dynamic classifier selection to dynamic ensemble selection [J]. Pattern Recognition,2008,41(5):1718-1731.

[13] 曹建军,张培林,王艳霞,等.一种求解子集问题的基于图的蚂蚁系统[J].系统仿真学报,2008,20(22):6146-6150.

[14] 王宁燕,韩晓霞.聚类的动态分类器集成选择[J].计算机系统应用,2015,24(4):205-208.

[15] NGUYEN T T,LUONG A V,DANG M T,et al. Ensemble selection based on classifier prediction confidence [J]. Pattern Recognition,2020,100:107104.

[16] 丁毅,王明亮,张道强.差异性随机子空间集成[J].计算机科学与探索,2018,12(9):1434-1443.

[17] GRETTON A,BORGWARDT K M,RASCH M J,et al. A kernel method for the two-sample-problem [C]// Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge, Mass:MIT Press,2006:513-520.

[18] 曹建军.基于提升小波包变换和改进蚁群算法的自行火炮在线诊断研究[D].石家庄:军械工程学院,2008.

[19] ROY A,CRUZ R M O,SABOURIN R,et al. Meta-regression based pool size prediction scheme for dynamic selection of classifiers [C]//2016 23rd International Conference on Pattern Recognition (ICPR 2016). Cancun,Mexico:IEEE,2016:216-221.

A Dynamic Ensemble Selection Model Based on Feature Selection and Clustering

XU Yuxin[1,2],CAO Jianjun[1**],WANG Baowei[2],WENG Nianfeng[1],GU Chumei[1,2]

(1. The 63rd Research Institute,National University of Defense Technology,Nanjing,Jiangsu,210007,China;2. College of Computer Science and Technology,Nanjing University of Information Science and Technology,Nanjing,Jiangsu,210044,China)

Abstract:To improve the accuracy of emitter individual recognition and reduce the computational complexity of dynamic selective ensemble,a Dynamic Ensemble Selection model based on Feature Selection and Clustering (FSC-DES) is proposed. The normalized Pearson correlation coefficient method is used to measure the difference of the confusion matrix between base classifiers,and the base classifier set and the corresponding feature subset set are obtained with the goal of maximizing the accuracy of each base classifier and the difference between base classifiers. The validation set is divided into several classes by the clustering method,and the optimal base classifier subset and corresponding feature subset are selected for each cluster validation set with the goal of maximizing classification accuracy of the validation set. In the testing phase,clustering is performed for the test set,and only the maximum mean difference of the data distribution is compared between each cluster of test samples and each cluster of validation samples to reduce the operation time. Each cluster of test samples is predicted under the feature subset set and base classifier subset corresponding to the valida-

tion set with the highest similarity,and the final decision is made according to the weighted sum of the prediction results of different weight base classifiers. Furthermore,the method proposed in this paper is compared with the conventional integrated learning method to assess the necessity and superiority of the method. The results show that the method proposed in this paper has the improvement of about 5% in the classification accuracy when the signal-to-noise ratio is 10 dB and 5 dB,respectively,demonstrating better classification effect and generalization performance.

Key words: feature selection;dynamic ensemble selection;support vector machine;ant colony Optimization;specific emitter identification;binary classification

责任编辑:陆媛峰