基于 SMOTE 和 gcForest 的医疗小样本数据分类研究

刘文昌[1]，魏赟[1]，袁浩轩[2]，高跃[2]
（1. 上海理工大学光电信息与计算机工程学院，上海 200093；
2. 复旦大学计算机科学技术学院，上海 200438）

摘 要：针对传统机器学习模型在医疗小样本数据上由浅层模型结构和复杂数据特征导致的分类表现不佳的问题，提出了一种联合多粒度改进级联森林（cgicForest，combine multi-grained improved cascade forest）模型。通过在多粒度扫描中加入随机抽样环节以及对变换特征进行优化来提高模型表征学习能力，并改进级联森林部分的层级结构来提升模型分类能力。针对存在类别不平衡问题的数据集，提出安全边界过采样（SBS, safe-borderline-SMOTE）算法在属于安全边界的少数样本周围进行动态插值，提高训练数据质量，再通过 cgicForest 模型进行训练学习，最终得到支持不平衡医疗小样本数据的 SBS-cgicForest 分类模型。在 3 种医疗数据集上应用 SBS-cgicForest 分类模型进行测试，结果表明，cgicForest 模型在具有复杂特征的医疗小样本数据上分类的性能指标较多粒度级联森林（gcForest, multi-grained cascade forest）模型提升了 4.1～5.4 个百分点，与 SBS 算法结合后各性能指标提升 6.6～11.2 个百分点，比与传统采样方法结合后的 F_1 评分高出 2～2.5 个百分点，为解决医疗小样本数据的分类问题提供了参考，并为智慧医疗场景下的物联网应用提供了支持。

关键词：医疗数据；小样本；SMOTE；gcForest
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.2096−3750.2023.00337

Research on medical small sample data classification based on SMOTE and gcForest

LIU Wenchang[1], WEI Yun[1],YUAN Haoxuan[2], GAO Yue[2]
1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
2. School of Computer Science and Technology, Fudan University, Shanghai 200438, China

Abstract: Aiming at the problem of poor classification performance in traditional machine learning models caused by shallow model structure and complex data characteristics in small medical sample data, an combine multi- grained improved cascade forest (cgicForest) model was proposed. It enhances the representation learning ability of the model by adding random sampling into the multi-grained scanning and optimizing the transformation features. It also enhances the model's classification ability by updating the cascade forest's hierarchical structure. Considering category imbalance problems in datasets, the safe-borderline-SMOTE (SBS) algorithm was proposed to dynamic interpolate around the few class samples belonging to the safety boundary, which can improve the quality of training data. The cgicForest was applied for training and learning, thus the SBS-cgicForest classification model was obtained which can support imbalanced medical small samples data. The model is used on three medical datasets for classification experiments. The results show that the performance indexes of the cgicForest model in the classification of medical small sample data with complex

收稿日期：2022−12−09；**修回日期：**2023−03−07
通信作者：刘文昌，wenchangit@163.com
基金项目：国家重点研发计划（No.2018YFB1700902）；国家发展和改革委员会资助的“基于 5G 网络特大型城市区域智慧医疗应急救援体系建设”项目
Foundation Items: The National Key Research and Development Program of China (No.2018YFB1700902)；This work was supported by the National Development and Reform Commission of China (NDRC) under grant “5G Network Enabled Intelligent Medicine and Emergency Rescue System for Giant Cities”

characteristics have increased by 4.1～5.4 percentage points, compared with the multi-grained cascade forest (gcForest) model. The performance indexes have increase by 6.6～11.2 percentage points after the combination with SBS algorithm, the F_1 score was 2～2.5 percentage points higher than that obtained by traditional sampling methods. It provides a reference for solving the classification problem of small medical sample data, and includes support for internet of things applications in smart medical scenarios.

Key words: medical data, small sample, SMOTE, gcForest

0　引言

随着信息技术的快速发展，数字智能化已然成为当下研究热点，例如，在智慧医疗领域利用机器学习[1]和深度学习方法实现辅助诊断、病情预测等，从而达到提高医疗效率、扩充医疗资源的效果。科研人员通常使用深度学习模型对大规模医疗数据进行分析，然而在很多情况下，构建一个大规模的医疗数据集难度很大，甚至是难以实现的，由于部分类型疾病信息的保密性和罕见性，往往只能获得小规模的医疗数据集。但是深度学习模型依靠大量数据的训练才能达到理想的分类效果，在训练过程中需要学习很多训练参数，当训练数据的规模较小时，容易导致模型出现过拟合的现象。相比之下，传统机器学习模型的参数要少得多，并且对数据更加敏感，训练起来也更加高效，更适用于小样本数据的分类，但是与其他领域的数据不同，医疗数据具有多样性且数据特征较为复杂，而传统机器学习模型都是浅层结构，在医疗小样本数据上的分类表现欠佳。

近年来，小样本数据的分类问题愈发引起人们的重视，文献[2]提出一种卷积自编码神经网络，通过使用大量未标注数据训练特征学习网络，而只使用少量标注数据微调网络解决由医疗图像标注稀缺引起的识别问题；文献[3]通过将两两图像中的特征组合生成新的样本，从而将小数据集扩展为适合训练神经网络模型的大规模数据集，这些深度神经网络方法都是从扩充数据样本的角度出发，根本上还是依赖于大规模数据；文献[4]提出将多个传统机器学习模型融合应用于小规模数据集，尽管其分类效果有所提升，但仍然无法避免传统机器学习的缺陷。当遇到特征维度过高、关系较为复杂的数据集时，浅层模型的表现不如深度模型[5]。文献[6]提出的多粒度级联森林（gcForest, multi-grained cascade forest）是一种集成决策树的深度模型，具有很少的超参数，而且对参数的设置包容度高[7]，适用于解决小数据集分类问题。对于数据特征较为复杂的医疗小样本数据，文献[8]提出使用 gcForest 构建相关分类模型，并得到优于深度神经网络的结果；文献[9]通过使用特定位置评分矩阵与小波变换技术结合来提取重要特征信息，进而利用 gcForest 模型进行预测，从而提高疾病诊断的准确率；文献[10]提出一种自适应加权森林（AWDF, adaptive weighted deep forest），通过为训练实例添加权重进行置信度筛选，从而增强了模型在部分医疗小样本数据上的分类表现，文献[11]提出提升级联深度森林（BCDForest, boosting cascade deep forest）模型，通过多类别粒度扫描和实例权重增强策略提升了模型在医疗小样本数据上的分类能力；受 DenseNet 思想的启发[12]，文献[13]提出密度自适应级联森林（daForest, dense adaptive cascade forest）模型，通过堆叠级联森林每层的输出特征，增强层级间的信息传递，从而提高模型在高维医疗小数据集上的分类准确性；文献[14]提出孪生深度森林（SDF, siamese deep forest）模型，通过度量向量对之间的距离并对每个森林配置权重，使相同类别实例数据分布更加集中，而不同类别实例数据更加松散，从而提高模型在医疗小样本数据上的训练效果；文献[15]提出两次提升的深度森林（TBDForest, two boosting deep forest）模型，通过计算每个森林中最重要特征的标准差构成一个新的特征传入下一层，并将每层结构一分为二进行训练，以提升模型在医疗小样本数据上的分类效果。对于类别不平衡的医疗小样本数据，文献[16]通过使用重采样技术处理训练数据，从而提高 gcForest 对不平衡医疗小样本数据的评估质量；文献[17]通过结合合成少数类过采样技术（SMOTE, synthetic minority over-sampling technique）[18]与 Tomek link[19]技术对医疗数据进行平衡处理，进而利用级联森林进行分类，从而提高集成模型对少数类样本的敏感度；文献[20]通过结合 Easy Ensemble 技术[21]将多粒度扫描生成的不平衡训练数据划分成若干子集，分别使用 SMOTE 进行平衡处理，进而对

级联森林中的基学习器进行集成训练来提升在不平衡医疗数据上的分类效果。尽管 gcForest 模型已经有了不少改进，但是仍存在一些不足之处，限制了它在医疗小样本数据上的分类表现。

- 医疗数据普遍存在数据不平衡问题，这会导致模型更偏向于多数类样本的表征学习，对少数类样本的拟合欠佳，而传统的 SMOTE 和 Borderline-SMOTE 采样方法容易产生噪声样本[22]，并不能很好地处理不平衡数据。
- 医疗数据的部分特征之间存在关联关系，多粒度扫描部分只能通过滑动窗口获取连续的特征片段，因而会导致重要特征组合丢失的现象。
- 经多粒度扫描生成的变换特征维度较高，其中包含一些非重要特征，这在一定程度上会影响级联森林的训练性能。
- 由于级联森林中基学习器训练的随机性会影响层级的拟合质量，级联森林部分的分类表现欠佳。

针对上述问题，本文提出了安全边界过采样（SBS, safe-borderline-SMOTE）算法，通过对少数类样本进行安全边界的划分，并对插值范围进行动态调整，从而生成具有更多有效信息的新样本，提高模型在不平衡数据上的分类表现。此外，本文提出了联合多粒度改进级联森林（cgicForest, combine multi-grained improved cascade forest）模型，通过在多粒度扫描中加入随机抽样环节提升模型的特征提取能力；结合随机森林的特征重要度分析对多粒度扫描生成的变换特征进行筛选优化，提高模型的训练性能；通过向级联森林每一层级添加子层并重复利用每层输出的类概率向量，增加数据训练机会的同时也对数据进行了增强，从而提高模型的分类能力，使其在医疗小样本数据上的分类表现更加出色，使智慧医疗领域中的物联网应用在资源受限场景下也能得到广泛使用。

1　相关模型理论

1.1　gcForest 模型

gcForest 模型由多粒度扫描和级联森林两部分组成，gcForest 模型结构如图 1 所示。多粒度扫描部分主要进行数据特征的增强，生成与分类结果关系更密切的特征。模型首先利用不同大小的滑动窗口对原始数据进行扫描，将提取的低维特征向量组成训练集；然后基于这些不同维度大小的训练集，分别训练一个随机森林[23]和完全随机森林[24]，这两种随机森林在构建子树时选取特征的方式不同，前者通过随机选择 $\sqrt{d}$ 个特征作为候选（d 为输入特征的数量），进而基于基尼系数建立分裂节点，而后者通过不断从输入特征中随机选择一个特征进行分割，直到满足停止条件，从而使得模型中的随机森林具有多样性，避免模型出现过拟合的现象；最后将各个随机森林生成的所有类概率向量连接作为级联森林模块的输入数据。

级联森林部分基于多粒度扫描模块生成的数据进行训练学习，其融合了 Stacking[25]集成学习方法，构成逐层递进的非神经网络式深度学习模型，每一层都由相同数量的随机森林和完全随机森林组成，除了第一层直接采用模块输入数据作为输入数据外，往后每一层则将上一层输出的类概率向量和模块输入数据相连接作为输入数据，以此通过向原始特征中加入上一层分类结果的方

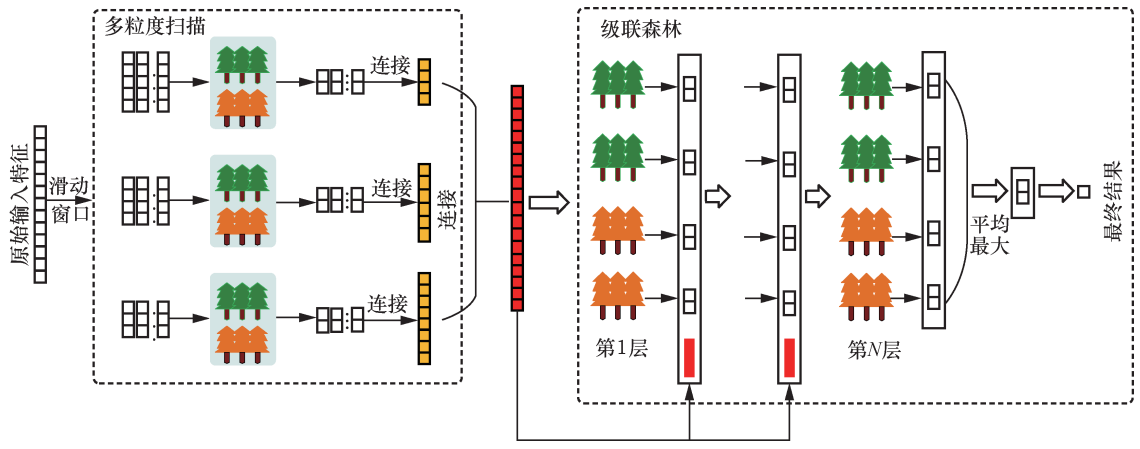

图 1　gcForest 模型结构

式实现特征的增强，层层迭代训练学习，而且每当新增了一个层级，整个级联森林都会通过验证集进行评估，如果评估指标较新增层级前模型的评估指标没有得到提升，则会自动删除新增层级并终止训练。最终将级联森林最后一层输出的类概率向量加和求平均，选择最大的一维作为最终预测结果。

gcForest 与深度神经网络的模型结构有异曲同工之妙，其将随机森林作为基学习器进行集成学习，随机森林充当了“神经元”这个角色。而随机森林亦融合改进了 Bagging 集成思想[26]，在构建过程中通过随机抽取不同训练子集和特征来训练子树，使随机森林中每个分类和回归树（CART, classification and regression tree）具备多样性[27]，CART 每个分支节点对应一个属性判断规则，每个叶子节点表示一种决策结果。而 CART 是基于基尼系数建立分裂节点的[28]，相较于 ID3、C4.5 算法[29]基于信息增益和信息增益比建立分裂节点的方式更加高效[30]。

$$\mathrm{Gini}(D)=1-\sum_{k=1}^{K}P_k^2=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{1}$$

$$\mathrm{Gini}(D,A)=\frac{|D_1|}{|D|}\mathrm{Gini}(D_1)+\frac{|D_2|}{|D|}\mathrm{Gini}(D_2) \tag{2}$$

其中，P_k 表示第 k 类的概率，C_k 表示第 k 类的集合，A 为样本集 D 的某一特征，根据特征 A 把样本集 D 分割成 D_1 和 D_2，基尼系数值越小样本集纯度越高，特征分割数据越彻底。在构建 CART 节点时，计算当前训练样本集现有特征在所有可能取值下的基尼系数，并选择最小基尼系数对应的特征和对应分割点作为当前节点的分裂条件，然后按照节点分裂条件将训练样本集划分。依次通过递归遍历特征子集，按照上述步骤不断构建子节点，直到满足停止条件。每个叶子节点都代表一个预测标签，其标签值是由划分到叶子节点的训练子集的类概率表示。

1.2　SMOTE 算法原理

SMOTE 算法原理如图 2 所示。假设 x 是少数类样本，先找到 x 的 k 个最近邻样本，根据过采样率 N 在其 k 个最近邻样本中随机挑选 n 个样本，分别记为：$x_1,x_2,\cdots,x_n$；然后在少数类样本 x 和 $x_i(i=1,2,\cdots,n)$之间进行随机线性插值，生成一个新的少数类样本 x_{new_i}

$$x_{\mathrm{new}\_i}=x+\mathrm{rand}(0,1)\times(x_i-x) \tag{3}$$

其中，rand(0,1)代表 0～1 之间任意一个小数，而构建的新样本则是选取的 x_i 与原样本 x 两个样本点之间的任意一点。

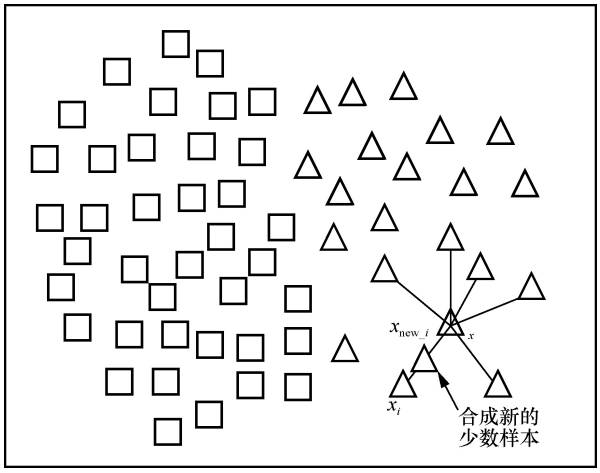

图 2　SMOTE 算法原理

2　SBS-cgicForest 分类模型

2.1　整体模型分类流程

本文将提出的 SBS 算法与 cgicForest 模型结合建立适用于医疗小样本数据的分类模型。整体模型分类流程如图 3 所示。

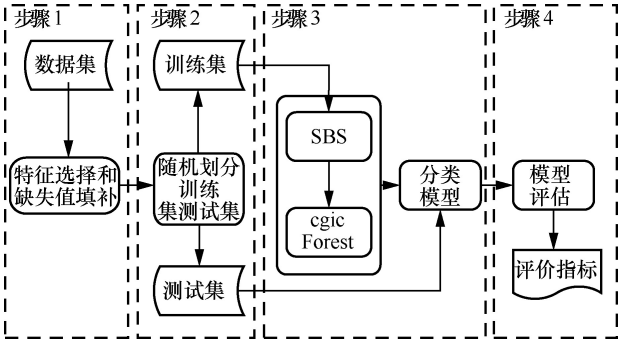

图 3　整体模型分类流程

首先，对数据集进行特征选择、缺失值填补等预处理操作；其次，将处理后的数据集按照指定比例随机划分为训练集与测试集，通过 SBS 算法对训练集中的少数类样本进行过采样，使训练集样本类别达到平衡状态；然后，使用数据平衡的训练集对 cgicForest 模型进行训练，生成支持不平衡医疗小样本数据的分类模型，并通过测试集进行验证；最后，通过各种评价指标对分类模型进行评估。

2.2 SBS 算法

由于传统的 SMOTE 算法[18]对少数类样本缺乏独特选择，容易在少数类样本较聚集的区域生成少数类的新样本，导致少数类样本数据分布没有明显改善，而 Borderline-SMOTE 在少数类样本的危险边界生成新样本，容易生成噪声样本，导致对分类模型的训练产生负面影响。针对上述问题，本文提出了 SBS 算法，通过对少数类样本进行安全边界的划分，并对插值范围进行动态调整，避免产生噪声数据，从而提高新样本的质量，SBS 算法见算法 1。

算法 1　SBS 算法

输入　多数类样本集 D_{maj}、少数类样本集 D_{min}、过采样率 N

输出　新的样本集 D

for X_i in D_{min} do
　$KNN_i \leftarrow$ Find Minoritys $KNN(X_i)$
　if $0.5 <= \text{CalMajAndMinRate}(KNN_i) <= 1$ then
　　$x \leftarrow X_i$
　　$Z(x) \leftarrow \text{FindNearestNeighbor}(x, N)$
　　for x_i in $Z(x)$ do
　　　if x_i is minority then
　　　　$x_{\text{new\_}i} = x + \text{rand}(0,1) \times (x_i - x)$
　　　else do
　　　　$x_{\text{new\_}i} = x + \text{rand}(0,0.5) \times (x_i - x)$
　　　end if
　　　$x_{\text{new\_}i}$ join D_{new}
　　end for
　end if
end for
return $D = D_{\text{maj}} \bigcup D_{\text{min}} \bigcup D_{\text{new}}$

2.3 cgicForest 模型

本文所提 cgicForest 模型旨在修复传统 gcForest 模型在医疗小样本数据分类任务上存在的缺陷，主要针对多粒度扫描、变换特征、级联森林 3 个部分进行了改进。cgicForest 模型结构如图 4 所示。

2.3.1 联合多粒度扫描

由于传统 gcForest 模型多粒度扫描结构化数据存在重要特征组合丢失的问题，本文提出了联合多粒度扫描，其结构如图 4 中联合多粒度扫描部分所示。在多粒度扫描的基础上加入随机抽样环节进一步对原始输入数据进行特征提取，随机抽取特征的规模设置与多粒度扫描中滑动窗口的大小设定类似，均不能超过原始输入数据的维度，在此范围内设置固定的随机抽取特征规模大小，并设置随机抽取次数，以此重复从原始输入数据中随机抽取不同的特征片段。将滑动窗口和随机抽样获取的特征片段组成不同训练集，进而分别训练一个随机森林和一个完全随机森林。通过训练好的森林模型对滑动窗口和随机抽样获取的特征片段进行预测，并将生成的类概率向量聚合形成相应的变换特征，最终将所有的变换特征连接生成与分类结果关系更密切的特征，从而达到增强多粒度扫描特征表达能力的效果。其中，添加的随机抽取特征环节通过对特征进行随机抽取组合的方式，使得每一个特征都能被充分学习利用，从而有效降低了重要特征组合丢失的风险，提高了多粒度扫描部分的表征学习能力。

2.3.2 特征优化

经多粒度扫描生成的变换特征维度较高，其中包含一些非重要特征，这在一定程度上会影响级联森林的训练性能。为了改进级联森林的拟合性能，提高训练质量和效率，本文通过结合随机森林所带的特征重要度分析功能，对级联森林的输入数据进

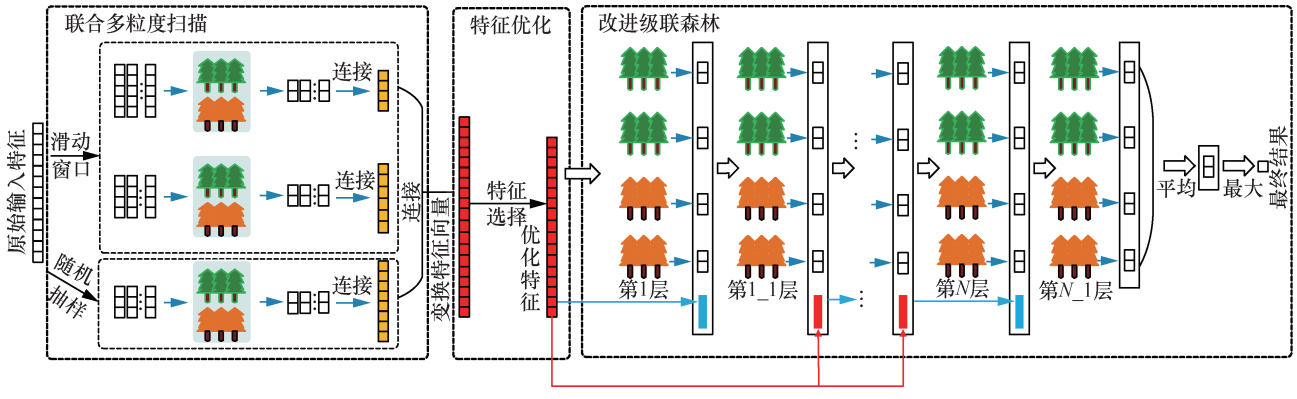

图 4　cgicForest 模型结构

行特征的筛选优化操作，具体流程如图 4 特征优化部分所示，其原理是通过比较原始验证集和加入特征噪声后的验证集的预测误差来评估特征的重要性，预测误差越大说明加入噪声的特征重要度越高，反之则越低。最终将重要值大于零的特征筛选出来，并去掉重要值为零的特征，将得到的特征子集作为级联森林的输入数据进行训练，从而降低了计算成本，使得级联森林可以更加高效准确地训练学习。

2.3.3　改进级联森林

针对因级联森林中基学习器训练的随机性使层级的拟合质量下降，导致级联森林部分的分类表现欠佳的问题，本文所提改进级联森林的模型结构如图 4 改进级联森林部分所示。分别在每一级联层中再新增一个子层，并且将父层的输入数据与输出数据连接作为子层的输入数据进行训练，这样，不仅通过添加新的子层提高了数据的训练机会，增强了模型的表征学习能力。通过将父层的输入数据与输出数据连接作为子层的输入数据进行训练，充分利用了每层输出的类概率向量，并且将特征再次增强进而改进了下一层的输入特征，通过向子层的输入特征中加入更多高质量的判别特征，使得新模型在医疗小样本数据上可以获得更强大的分类能力，并且提高了模型的稳定性和泛化能力，保证了模型在其他类型数据集上也可以稳定输出。

将上述联合多粒度扫描与特征优化以及改进级联森林部分整合，构成本文提出的 cgicForest 模型。以二分类为例，具体步骤如下。

1) 假设原始输入特征的大小为 100 维，联合多粒度扫描过程中分别采用 40、50 维大小的滑动窗口，以 1 维大小的步长进行特征提取。针对每一条原始输入数据，分别生成 61 个 40 维的特征片段和 51 个 50 维的特征片段。同时采用 30 维大小的随机抽样特征规模对原始输入特征进行 71 次的重复抽取，生成 71 个 30 维的特征片段。利用以上不同规模的特征片段分别训练出一个随机森林和一个完全随机森林，通过训练好的森林模型将每一个特征片段转化为相应二维大小的类概率向量，将其连接从而生成对应的变换特征。因此，以上不同规模大小的特征片段将分别转化为 244、204、284 维的变换特征，将所有向量连接形成 732 维的变换特征向量。

2) 通过随机森林对联合多粒度扫描生成的 732 维变换特征进行筛选，假设获得 600 维优化的特征，将其作为改进级联森林的输入数据进行训练。

3) 改进级联森林每一层设有两个随机森林和两个完全随机森林，第 1 层的输入数据是优化后的包含 600 维特征的输入数据，而第 1 层子层的输入数据则是父层输出数据与输入数据相连接组成的 608 维数据，往后第 i 层的输入数据都将是 i–1 层的子层输出数据与模块优化的 600 维特征输入数据的连接组成的 608 维数据，而子层的输入数据则是父层的输出数据与输入数据相连接组成的 616 维数据。不断按照上述方式依次新增级联层级，直到在验证集上的测试指标不再提高，最终将改进级联森林最后一层输出的类概率向量加和求平均，选取占比最大的类概率作为模型最终的分类结果。

3　实验与结果分析

3.1　实验数据

本文从美国加州大学欧文分校机器学习存储库（UCI, UC Irvine Machine Learning Repository）上下载了 Diabetes、ECG、Cirrhosis 3 种医疗数据集，其中，Diabetes 是糖尿病患者数据集、ECG 是心律失常患者数据集、Cirrhosis 是肝硬化患者数据集，3 种医疗数据集的样本数量都比较少，适用于验证本文提出的分类模型。3 种 UCI 医疗数据集的详细信息见表 1，本文对每个数据集都进行了预处理操作，如缺失值填补、数据类型转换、特征选择等，以确保实验模型都能够达到最好的训练效果。由于 ECG 和 Cirrhosis 这两个数据集样本的异常类别有多种，本文统一将其设置为一种异常类别，从而转化为二分类问题方便后续实验结果的对比分析。

表 1　3 种 UCI 医疗数据集的详细信息

数据集	实例数	特征数	少数类数量	数据不平衡比例
Diabetes	768	8	268	1.87
ECG	452	176	207	1.18
Cirrhosis	418	17	144	1.90

3.2　实验设置

本文的所有实验都基于 Python 语言实现，运行环境为 Windows 10 操作系统，处理器型号为 Intel Core i5-6300HQ CPU 2.30 GHz，内存大小为 8 GB。在实验过程中，对于每个数据集，都将其划分为两个互斥的集合，其中，80%作为训练集，20%作为

测试集。然后，将使用逻辑回归（LR, logistic regression）、支持向量机（SVM, support vector machine）、gcForest、TBDForest 与本文所提 cgicForest 模型在不同数据集上建立相关分类模型。其中，传统 gcForest 模型和 cgicForest 模型多粒度扫描部分滑动窗口以及随机抽样规模大小分别设为 d/3、d/2、2d/3（在联合多粒度扫描中的随机抽样规模大小），所有的随机森林都设置为 30 个决策树，级联森林部分所有的随机森林都设置为 100 个决策树。为了公平比较每个模型，对其他模型的参数都进行了调优，使其分类表现都达到最优。本文分别采用正确率（ACC, accuracy）、精确率（PRE, precision）、召回率（REC, recall）、F_1 评分、曲线下面积（AUC, area under the curve）这 5 个评价指标对模型的分类性能进行评估。这些指标都建立在混淆矩阵的基础之上，混淆矩阵见表 2。

正确率、精确率、召回率、F_1 评分分别为

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

$$\mathrm{PRE}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\mathrm{REC}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

$$\mathrm{F_1}=\frac{2\times\mathrm{PRE}\times\mathrm{REC}}{\mathrm{PRE}+\mathrm{REC}} \tag{7}$$

表 2 混淆矩阵

对比项	实际为正例	实际为负例
预测为正例	TP	FP
预测为负例	FN	TN

AUC 表示受试者操作特征（ROC, receiver operator characteristic）曲线与坐标轴围成的面积，值越大，模型的分类效果越好，ROC 空间以假阳性率为 x 轴，真阳性率为 y 轴。其中假阳性率等于 FP/(FP+TN)，真阳性率等于召回率。

3.3 实验结果与分析

为了检验 cgicForest 模型的性能和分类效果，本文先不结合 SBS 算法进行数据平衡处理，直接基于原始训练集对各个模型进行训练，并通过测试集对模型进行评估，各个模型在数据集上的实验结果见表 3。

通过对比表 3 中的数据可以看出，本文所提 cgicForest 模型在前两个数据集中的表现较其他模型在各个指标上均有明显提升，尤其是在 ECG 数据集上提升效果最大，与标准 gcForest 模型相比正确率提高了约 4.4 个百分点，精确率提高了约 5.4 个百分点，召回率提高了约 4.1 个百分点，F_1

表 3 各个模型在数据集上的实验结果

数据集	指标	LR	SVM	gcForest	TBDForest	cgicForest
Diabetes	ACC	0.776 0	0.760 4	0.781 2	0.786 5	**0.802 1**
	PRE	0.759 9	0.729 2	0.756 9	0.756 9	**0.779 7**
	REC	0.699 6	0.696 5	0.720 3	0.745 3	**0.752 6**
	F_1	0.714 7	0.707 0	0.732 4	0.750 4	**0.763 0**
	AUC	0.849 9	0.805 1	0.853 4	0.838 5	**0.857 1**
ECG	ACC	0.708 0	0.699 1	0.761 1	0.769 9	**0.805 3**
	PRE	0.708 5	0.698 6	0.761 8	0.771 9	**0.815 9**
	REC	0.701 1	0.692 9	0.756 0	0.764 2	**0.797 0**
	F_1	0.702 0	0.693 7	0.757 3	0.765 8	**0.799 6**
	AUC	0.767 0	0.798 2	0.810 8	0.814 6	**0.822 2**
Cirrhosis	ACC	**0.781 0**	0.742 9	0.752 4	0.752 4	0.763 3
	PRE	**0.781 5**	0.715 7	0.736 9	0.732 2	0.753 9
	REC	**0.713 8**	0.691 4	0.685 4	0.692 0	0.692 0
	F_1	**0.728 9**	0.699 2	0.696 7	0.702 5	0.702 5
	AUC	**0.807 2**	0.755 2	0.759 3	0.756 8	0.788 4

评分提高了约 4.2 个百分点，与 TBDForest 模型相比正确率提高了约 3.5 个百分点，精确率提高了约 4.4 个百分点，召回率提高了约 3.3 个百分点，F_1 评分提高了约 3.4 个百分点，这表明 cgicForest 模型在处理具有高维复杂特征的医疗小样本数据分类问题方面具有显著优势，ROC 曲线对比如图 5 所示，可以看出，cgicForest 模型的 ROC 曲线下面积比传统 gcForest 模型都要大，而且 AUC 值较标准 gcForest 高出约 1.1 个百分点，较 TBDForest 高出约 0.8 个百分点，体现了 cgicForest 模型的稳定性能。

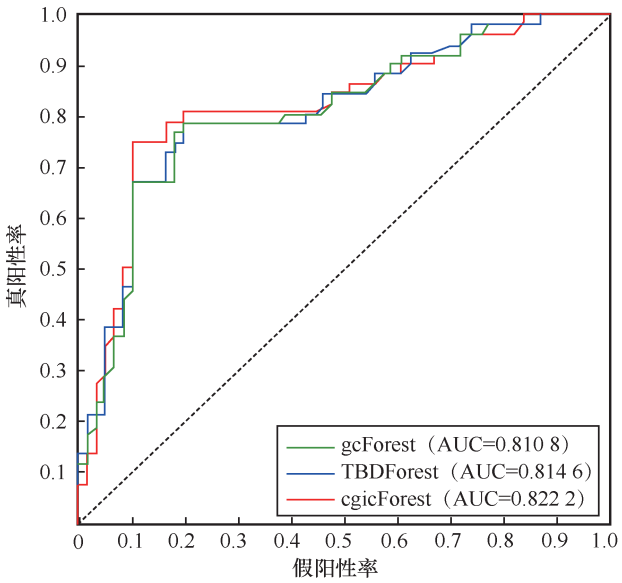

图 5　ROC 曲线对比

传统 gcForest 模型在多粒度扫描部分通过滑动窗口提取不同维度的特征片段，进而利用基学习器生成与分类结果关系更加密切的特征向量，并将其连接作为级联森林的输入数据，但是由于滑动窗口只能获取连续的特征片段，容易导致一些重要特征组合的丢失。本文所提 cgicForest 模型通过加入随机抽样特征环节保障提取特征片段的多样性，有效降低了重要特征组合丢失的风险，增强多粒度扫描生成变换特征向量的质量，并通过特征的重要度分析对生成的变换特征向量再次进行优化，从而确保级联森林部分的输入数据质量。级联森林部分每一层级通过对相同数量不同类型的随机森林进行集成训练学习，并且对原始特征进行扩展，层层递进提高模型分类性能。由于随机森林训练的随机性，即从数据集中随机地选择子样本并从完整的属性中随机选取部分属性构建子树，使得随机森林具有多样性，并与随机性更强的完全随机森林组合构成级联层级，保障了级联森林中基学习器的多样性，避免模型出现过拟合，但也由于这个原因，在一定程度上会降低随机森林拟合的质量，影响了级联森林层层递进训练学习的稳定性，导致级联森林部分的分类表现欠佳。本文所提 cgicForest 模型通过修改级联森林的层级结构，增加了数据的训练机会，并且使数据的特征再次得到增强，提高了其在小样本、高维度和类别不平衡的医疗数据上的分类能力。

但是，在第 3 个数据集上，各个集成森林学习模型表现得都不如逻辑回归模型，这是因为集成森林模型在数据不平衡比例较高且特征数量较少的数据集上会出现拟合不佳的情况。具体来说，随机森林作为 gcForest 模型的基学习器，而 CART 同样作为随机森林的基学习器，当遇到数据不平衡比例较高且特征数量较少的医疗小样本数据时，由于数据类别与特征值的分布不均匀，会导致求得的基尼系数值不可靠。在此基础上构建的子树质量较差，这必然会使随机森林整体拟合质量下降，直接影响了多粒度扫描部分生成的数据质量，也间接影响后续级联森林部分进行表征学习，导致整体模型的分类效果欠佳。总而言之，由于数据样本类别的不平衡比例较高，致使模型更偏向于多数类样本的表征学习，而对少数类样本的拟合欠佳。为弥补这一缺陷，本文先通过 SBS 算法对少数类样本进行扩充，使训练集的样本类别达到平衡状态，然后再对实验模型进行训练，最后通过测试集进行模型评估，各个模型结合 SBS 算法后在数据集上的实验结果见表 4。

通过对比表 3 与表 4 的实验结果可以看出，各个模型在结合 SBS 算法之后，在 3 个数据集上的实验结果均有显著提高，只有逻辑回归模型在第 3 个数据集上的表现变差，这是因为逻辑回归模型易受极端值的影响。可以发现本文提出的 SBS-cgicForest 分类模型在各个数据集上的表现均是最佳的，并且与 cgicForest 模型相比正确率平均提升了 6.6 个百分点，精确率平均提升了 7.6 个百分点，召回率平均提升了 11.2 个百分点，F_1 评分平均提升了 10.1 个百分点，AUC 平均提升了 9.9 个百分点。实验结果表明，SBS 算法有效弥补了多粒度扫描部分训练不平衡数据的缺陷，充分保障了级联森林模块输入数据的质量。

表 4　各个模型结合 SBS 算法后在数据集上的实验结果

数据集	指标	LR	SVM	gcForest	TBDForest	cgicForest
Diabetes	ACC	0.804 0	0.868 0	0.868 0	0.873 6	**0.884 0**
	PRE	0.804 1	0.869 0	0.870 7	0.873 7	**0.885 8**
	REC	0.804 1	0.868 2	0.868 3	0.870 3	**0.884 3**
	F_1	0.804 0	0.867 9	0.867 8	0.872 9	**0.883 9**
	AUC	0.876 2	0.927 0	0.933 9	0.934 2	**0.938 7**
ECG	ACC	0.780 5	0.772 4	0.821 1	0.826 7	**0.837 4**
	PRE	0.793 4	0.779 3	0.832 5	0.835 6	**0.844 9**
	REC	0.791 0	0.780 2	0.831 3	0.835 8	**0.846 0**
	F_1	0.780 4	0.772 3	0.821 1	0.824 9	**0.837 4**
	AUC	0.840 6	0.869 8	0.902 9	0.904 1	**0.9 080**
Cirrhosis	ACC	0.715 3	0.766 4	0.832 1	0.835 3	**0.846 7**
	PRE	0.712 9	0.772 7	0.831 8	0.835 8	**0.846 4**
	REC	0.718 8	0.772 0	0.834 0	0.835 9	**0.848 7**
	F_1	0.715 3	0.766 4	0.831 8	0.833 9	**0.846 4**
	AUC	0.796 2	0.834 4	0.913 0	0.914 7	**0.918 1**

本文所提 cgicForest 模型主要通过融合联合多粒度扫描、特征优化以及改进级联森林 3 种方法增强模型的分类能力，为了验证各个方法的有效性，本文在 ECG 数据集上对 cgicForest 模型进行消融实验。将 cgicForest 模型分别与融合联合多粒度扫描和特征优化两种方法的联合多粒度挑选级联森林（cgscForest, combine multi-grained selected cascade forest）模型、只融合联合多粒度扫描的联合多粒度级联森林（cgcForest, multi-grained cascade forest）模型以及标准 gcForest 模型进行对比，cgicForest 模型消融实验结果如图 6 所示。cgcForest 模型较标准 gcForest 模型表现更加出色，精确率提升了约 2 个百分点，由于召回率没有明显变化，F_1 评分只提高约 0.5 个百分点，cgscForest 模型精确率进一步提升了约 1.6 个百分点，且召回率与 F_1 评分分别提高了约 3.6 与 3.1 个百分点，最终加入改进级联森林方法的 cgicForest 模型使分类效果进一步提升，精确率提升了约 1.8 个百分点，召回率与 F_1 评分分别提升了约 0.5 与 1.2 个百分点。结果表明，本文提出的 cgicForest 模型通过融合联合多粒度扫描随机提取不同的特征，有效降低了重要特征组合丢失的风险，增强了变换特征向量的质量，从而提升了模型的表征学习能力，并进一步通过特征优化对生成的变换特征进行筛选，有效确保级联森林部分的输入数据质量，提升了模型的分类性能，最终通过改进级联森林的层级结构增加了数据的训练机会，并且使数据的特征再次得到增强，进一步提高了模型在医疗小样本数据上的分类能力。

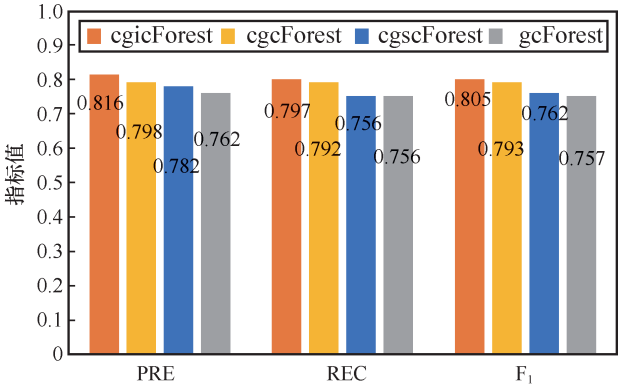

图 6　cgicForest 模型消融实验结果

为了验证 SBS 算法处理不平衡数据的优越性，本文对 SBS 算法进行了消融实验，首先将 cgicForest 模型分别与 SBS、传统 Borderline-SMOTE（为便于

展示在图 7 中以 BSMOTE 表示）以及 SMOTE 算法结合，SBS 算法消融实验结果如图 7 所示，cgicForest 模型与 SBS 算法结合的表现比与传统的 BSMOTE 或 SMOTE 算法结合的表现更加出色，在 3 个数据集上的 F_1 评分均有 2～2.5 个百分点的提升。结果表明 SBS 算法通过在定义的安全边界范围内动态调整插值范围进行少数类样本过采样，充分保障了生成样本的质量，避免产生噪声数据，从而有效提高了 cgicForest 模型在不平衡数据上的分类表现。实验证明，SBS-cgicForest 分类模型更加适用于医疗小样本数据。

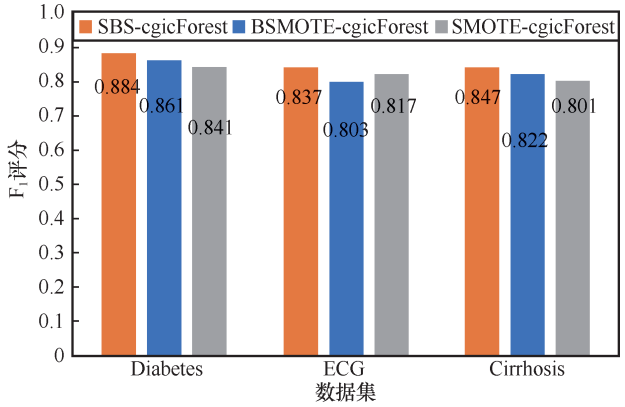

图 7　SBS 算法消融实验结果

3.4　模型复杂度分析

本文所提出的分类模型主要包含 SBS 算法和 cgicForest 模型两部分，其中 SBS 算法的时间复杂度为 $O(n^2)$，cgicForest 模型的复杂度为 $O(n^3)$，n 表示训练样本的数量，整体模型的复杂度为 $O(n^3)$。SBS 算法和 cgicForest 模型的改进都未涉及距离的计算，但是由于增加了级联森林部分的层数，这会在训练过程中耗费更多的时间，不过本文通过对变换特征的优化在一定程度上也提升了级联森林的训练效率。在解决医疗小样本数据的分类问题时，由于数据量较少，对模型的复杂度要求较低，SBS-cgicForest 分类模型的训练效率与传统机器学习算法相比差别不大，在资源受限场景下也可以稳定运行。

4　结束语

本文将 SBS 算法与 cgicForest 模型结合进行医疗小样本数据分类。当遇到训练样本类别不平衡的状况时，通过 SBS 算法对属于安全边界的少数类样本进行动态过采样使训练集样本类别达到平衡状态，从而有效保障了输入数据的质量，提高了 cgicForest 模型在不平衡数据上的分类表现。本文所提 cgicForest 模型通过联合多粒度扫描提升特征提取能力，并对变换特征进行优化提高模型训练性能，改进级联森林增加数据训练机会的同时也对数据进行了增强，充分利用了每层输出的类概率向量，使得模型具有更强的分类能力，并且提高了模型的稳定性和泛化能力，保证了模型在医疗小样本数据上可以稳定输出。实验结果表明，cgicForest 模型的分类表现较标准 gcForest 模型明显提升，将 cgicForest 模型与 SBS 算法结合后，分类效果提升显著，并且比与其他过采样方法结合更有优势，为解决医疗小样本数据的分类问题提供了参考，并为智慧医疗场景下的物联网应用提供了支持。未来研究考虑将模型的应用范围扩大，尝试与神经网络模型结合构建端到端的小样本数据分类模型，并尝试优化模型复杂度。

参考文献：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Publishing House, 2016.

[2] CHEN M, SHI X B, ZHANG Y, et al. Deep feature learning for medical image analysis with convolutional autoencoder neural network[J]. IEEE Transactions on Big Data, 2021, 7(4): 750-758.

[3] HU G S, PENG X J, YANG Y X, et al. Frankenstein: learning deep face representations using small data[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2018, 27(1): 293-303.

[4] 李春生, 曹琦, 于澍. 针对小规模数据集的多模型融合算法研究[J]. 计算机技术与发展, 2020, 30(2): 63-66.
LI C S, CAO Q, YU S. Research on multi-model fusion algorithm for small scale data sets[J]. Computer Technology and Development, 2020, 30(2): 63-66.

[5] 薛参观, 燕雪峰. 基于改进深度森林算法的软件缺陷预测[J]. 计算机科学, 2018, 45(8): 160-165.
XUE C G, YAN X F. Software defect prediction based on improved deep forest algorithm[J]. Computer Science, 2018, 45(8): 160-165.

[6] ZHOU Z H, FENG J. Deep forest: towards an alternative to deep neural networks[EB]. 2017.

[7] 何宏, 陈叔达. 面部表情的深度卷积级联森林识别[J]. 小型微型计算机系统, 2021, 42(4): 805-809.

HE H, CHEN S D. Deep convolutional cascade forest for facial expression recognition[J]. Journal of Chinese Computer Systems, 2021, 42(4): 805-809.

[8] 颜建军, 刘章鹏, 刘国萍, 等. 基于深度森林算法的慢性胃炎中医证候分类[J]. 华东理工大学学报(自然科学版), 2019, 45(4): 593-599.
YAN J J, LIU Z P, LIU G P, et al. Syndrome classification of chronic gastritis based on multi-grained cascade forest[J]. Journal of East China University of Science and Technology, 2019, 45(4): 593-599.

[9] CHEN Z H, LI L P, HE Z, et al. An improved deep forest model for predicting self-interacting proteins from protein sequence using wavelet transformation[J]. Frontiers in Genetics, 2019(10): 90.

[10] UTKIN L, KONSTANTINOV A, MELDO A, et al. A deep forest improvement by using weighted schemes[C]//Proceedings of 2019 24th Conference of Open Innovations Association (FRUCT). Piscataway: IEEE Press, 2019: 451-456.

[11] GUO Y, LIU S H, LI Z H, et al. BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data[J]. BMC Bioinformatics, 2018, 19(Suppl 5): 118.

[12] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2261-2269.

[13] WANG H Y, TANG Y, JIA Z Y, et al. Dense adaptive cascade forest: a self-adaptive deep ensemble for classification problems[J]. Soft Computing, 2020, 24(4): 2955-2968.

[14] UTKIN L V, RYABININ M A. A Siamese deep forest[J]. Knowledge-Based Systems, 2018, 139: 13-22.

[15] FAN Y M, QI L, TIE Y. The cascade improved model based deep forest for small-scale datasets classification[C]//Proceedings of 2019 8th International Symposium on Next Generation Electronics (ISNE). Piscataway: IEEE Press, 2019: 1-3.

[16] LIU H, ZHANG N, JIN S G, et al. Small sample color fundus image quality assessment based on gcforest[J]. Multimedia Tools and Applications, 2021, 80(11): 17441-17459.

[17] 刘超, 吴申, 郑一超, 等. 基于深度森林和DNA甲基化的癌症分类研究[J]. 计算机工程与应用, 2020, 56(13): 189-193.
LIU C, WU S, ZHENG Y C, et al. Classification of cancer based on deep forest and DNA methylation[J]. Computer Engineering and Applications, 2020, 56(13): 189-193.

[18] XU Z Z. A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data[J]. Information Sciences, 2021(572): 574-589.

[19] PEREIRA R M, COSTA Y M G, SILLA C N Jr. MLTL: a multi-label approach for the Tomek Link under sampling algorithm[J]. Neurocomputing, 2020 (383): 95-105.

[20] YUAN Z W, ZHAO P. An improved ensemble learning for imbalanced data classification[C]//Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Piscataway: IEEE Press, 2019: 408-411.

[21] REN X Y, YUAN Z Y, HUANG J M. Research on fake reviews detection based on feature construction and Easy Ensemble-RF[C]//Proceedings of 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE). Piscataway: IEEE Press, 2022: 478-482.

[22] XU X L, CHEN W, SUN Y F. Over-sampling algorithm for imbalanced data classification[J]. Journal of Systems Engineering and Electronics, 2019, 30(6): 1182-1191.

[23] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[24] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[25] BREIMAN L. Stacked regressions[J]. Machine Learning, 1996, 24(1): 49-64.

[26] 吴辰文, 梁靖涵, 王伟, 等. 一种顺序响应的随机森林: 变量预测和选择[J]. 小型微型计算机系统, 2017, 38(8): 1762-1766.
WU C W, LIANG J H, WANG W, et al. Random forest algorithm for sequential response: prediction and selection of variables[J]. Journal of Chinese Computer Systems, 2017, 38(8): 1762-1766.

[27] 乔健, 诸佳慧, 严康桓. 基于随机森林 CART 特征选择改进算法的电信客户流失预测模型[J]. 电信工程技术与标准化, 2022, 35(3): 78-82.
QIAO J, ZHU J H, YAN K H. Telecom customer churn prediction model based on improved random forest cart feature selection algorithm[J]. Telecom Engineering Technics and Standardization, 2022, 35(3): 78-82.

[28] TAHMASSEBI A, GANDOMI A H, SCHULTE M H J, et al. Optimized naive-Bayes and decision tree approaches for fMRI smoking cessation classification[J]. Complexity, 2018: 1-24.

[29] HSSINA B, MERBOUHA A, EZZIKOURI H, et al. A comparative study of decision tree ID3 and C4.5[J]. International Journal of Advanced Computer Science and Applications, 2014, 4(2): 126-133.

[30] 李孝伟, 陈福才, 李邵梅. 基于分类规则的C4.5决策树改进算法[J]. 计算机工程与设计, 2013, 34(12): 4321-4325, 4330.
LI X W, CHEN F C, LI S M. Improved C4.5decision tree algorithm based on classification rules[J]. Computer Engineering and Design, 2013, 34(12): 4321-4325, 4330.

[作者简介]

刘文昌（1998– ），男，上海理工大学光电信息与计算机工程学院硕士生，主要研究方向为机器学习、数据分类等。

袁浩轩（1996– ），男，复旦大学计算机科学技术学院博士生，主要研究方向为深度学习、智能频谱感知等。

魏赟（1976– ），女，博士，上海理工大学副教授，主要研究方向为分布式系统、网络信息控制等。

高跃（1978– ），男，博士，复旦大学教授，主要研究方向为卫星互联网、天空地一体化网络、压缩感知与机器学习、智能天线。