ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 5,Sep. 2025,pp. 1222－1238
DOI:10. 16337/j. 1004-9037. 2025. 05. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多重随机性与隐私保护的栈式随机森林算法

宋奕霖，王士同

(江南大学人工智能与计算机学院,无锡 214122)

摘　要：作为一种针对分类和回归任务行之有效的集成学习算法,随机森林(Random forest，RF)还面临着泛化能力提升和隐私保护的挑战。本文提出了一种改进的基于多重随机性与隐私保护的栈式随机森林(Bernoulli-multinomial stacked random forest,BMS-RF)算法。基本思想是在构造决策树分裂特征和分裂点选择阶段引入伯努利分布Dropout部分特征向量选择候选特征向量,通过两个多项分布随机选择分裂特征与分裂点,每棵决策树采用非数值查询的指数机制添加噪声维持其隐私保护机制,在集成分类器时引入多层栈式结构将前一层的输出随机投影和源训练集拼接作为新的输入,使得每一森林可以共享源样本空间信息,逐层提高基学习器分类性能。通过对此算法的一致性以及隐私能力的理论分析表明BMS-RF可以通过栈式结构显著提高分类性能。14个中小规模数据集合上的实验结果验证了该算法不但能降低运行时间且具有更好的泛化性能,隐私保护水平较强时可以在简化结构和提高运行速度的基础上达到与RF变体基本一致的分类性能。

关键词：随机森林;集成分类;栈式结构;隐私保护;决策树

中图分类号：TP181　　**文献标志码**：A

Trestle Random Forest Based on Multiple Randomness and Privacy Protection

SONG Yilin，WANG Shitong

(School of Artificial Intelligence and Computer Science，Jiangnan University，Wuxi 214122，China)

Abstract：As an effective ensemble learning algorithm for classification and regression tasks，the random forest (RF) also faces challenges in improving generalization ability and privacy protection. In response to this challenge，this paper proposes an improved Bernoulli-multinomial stacked random forest (BMS-RF) algorithm based on multiple randomness and privacy protection. The basic idea is to introduce Bernoulli distribution Dropout partial feature vectors to select candidate feature vectors in the stage of constructing decision tree splitting features and splitting point selection. By randomly selecting splitting features and splitting points through two polynomial distributions，each decision tree adopts a non numerical query index mechanism to add noise for maintaining its privacy protection mechanism. When integrating classifiers，a multi-layer stack structure is introduced to randomly project the output of the previous layer and concatenate the source training set as new inputs，so that each forest can share the spatial information of the source samples and improve the classification performance of the base learner layer by layer. Theoretical

基金项目:国家重点研发专项计划(2022YFE0112400)。
收稿日期:2023-12-05;**修订日期**:2024-04-09

analysis of the consistency and privacy ability of this algorithm shows that BMS-RF can significantly improve classification performance through a stack structure. Experimental results on 14 small and medium-sized datasets verify that the algorithm not only reduces running time but also has better generalization performance. When the privacy protection is strong, it can achieve classification performance similar to RF variants on the basis of simplifying the structure and improving running speed.

Key words: random forest (RF); integrated classification; stack structure; privacy protection; decision tree

引　言

单个分类器的传统机器学习分类方法包含决策树算法和支持向量机算法,由于其过拟合及性能不佳衍生出多个分类器集成算法,如装袋法(Bagging)和提升法(Boosting)。Bagging基分类器是从原始数据中均匀采样,互相独立并行集成,而Boosting中基分类器的数据集通过带权重的采样得到,下一个基分类器依赖上一个分类器的预测结果串行集成。随机森林(Random forest, RF)是由Breiman[1]提出的一种基于决策树的集成机器学习方法。实践证明,随机森林因其运行速度快、泛化错误率低、可处理多类问题,以及易于并行硬件架构分布而被普遍认为是最准确的通用分类器之一。现有的随机森林在处理许多实际问题上表现异常出色,例如监测航空进场调度[2]、预测交通数据、读取医疗数据预测患者再次入院概率以及预测金融市场波动等,而且在许多跨领域的学科中也得到广泛的应用[3-5]。

Bremain[1]提出一致性确保了随着样本量的增加,RF结果趋于最佳,然而传统RF的树节点分割过程易产生数据依赖树。为解决一致性问题,Biau等[6]分析了各种简化版本的随机森林和其他随机预测模型,并给出了一致性定理。Denil等[7]发现了与Breiman RF类似的随机森林新变种,通过泊松分布选择每个节点候选特征的子空间以增加随机性并简化确定性树的构建过程。现有随机森林方法在决策树的构造阶段渗透更多的随机性来简化确定性树构建过程,导致分类性能与RF也会有较大差距,因此一致性随机森林泛化性能提升仍会是随机森林领域的研究热点。随机森林在低维数据集上具有卓越的泛化性能和不确定性处理能力,然而处理高维数据是导致大多数分类法性能退化的原因之一,随机森林是为数不多的可以扩展为高维数据建模的方法之一,现已有许多改进的高维数据建模随机森林算法。旋转随机森林(Rotating random forest, RoF)算法[8]对K个子集应用主成分分析(Principal component analysis, PCA),K轴旋转获得基分类器的新特征,在集成同时保证个体准确性和多样性。分层抽样随机森林算法[9]采用分层抽样的方法选择具有高维数据的随机森林特征子空间,特征子空间选择时按比例从每个组中随机选择特征,确保每个子空间包含足够的信息特征,以便在高维数据中进行分类。子空间选择随机森林算法[10]采用p值评估特征在寻找信息性特征之间界限时的重要性,提出新的特征加权子空间选择方法。然而,选择分裂特征与分裂点时传统随机森林会遍历所有特征,这将消耗太多的时间,如何提高算法在高维数据集上的运行速度,通过树结构引入随机性,在不牺牲其学习性能的情况下减少数据依赖性仍是亟待解决的问题。

此外保护数据隐私已经成为机器学习发展的一个不可或缺的方向,传统隐私保护算法包括k-匿名[11]、l-多样性[12]、t-接近性[13]和m-不变性[14]等。传统隐私保护基于攻击者未被严格定义且未完全量化攻击者的背景知识,面对新的攻击类型,模型需要不断改进,而差分隐私不仅提供了隐私保护水平的严格定义与定量评估办法,也证明了需要添加的噪声量与数据集本身的大小无关,其中隐私预算规模、分类器所需查询数量和数据中微小变化的敏感性决定了差分隐私对分类器的影响。Bai等[15]提出基于不纯度的噪声本质上是满足差分隐私的指数机制,如何利用差分隐私添加扰动来提供隐私保护,又不

会导致随机森林算法的分类精度下降仍然是数据挖掘领域的研究热点。

一致性与隐私保护随机森林仍然面临以下具有挑战性的问题:(1)选择分裂特征与分裂点时需遍历所有特征导致训练速度较慢。因此本文在构建树阶段增加额外的随机性,添加伯努利分布选择部分候选分裂特征与分裂点,基于众多BMS-Tree树来构建森林。(2)一致性导致的泛化性能下降问题。为解决此问题,基于栈式泛化原理[16]引入栈式结构改进单个分类器逐层提高分类性能,根据每层构建的森林生成栈式结构的BMS-RF。(3)隐私保护和数据可用性难以平衡问题。本文提出BMS-Tree以享有一致性和隐私保护机制,在实验部分,将BMS-RF与其他随机森林在公开数据集上进行了比较,实验结果证明了BMS-RF的有效性。

1 相关理论

1.1 随机森林

在数据挖掘领域,决策树是常用的分类回归算法,主要有基于香农熵的ID3、C4.5算法和基于基尼系数分裂的CART算法。随机森林算法是受到特征随机选择、随机子空间和随机分裂选择技术启发,使用Bagging将几种决策树组合在一起的方法。Breiman RF通过结合许多多样化的独立训练的决策树来形成森林,总的数据空间是$\mathbf{R}^d$,将每棵决策树的构建过程作为数据空间的一个划分,那么每一个叶子节点是$\mathbf{R}^d$的一个分区,也是数据空间的超矩形单元。

随机森林算法主要由3个部分组成:原始样本有放回采样、随机特征子空间选择和多数投票。首先从给定数据集D中有放回的等权重随机抽样n个数据点,得到$\delta_1,\delta_2,\cdots,\delta_M$,采样得到的数据集构成$M$棵CART决策树;其次从原始$d$个特征中随机抽样出$m$个特征,其中$m<d$,根据最大不纯度或最大均方差(Mean squared error, MSE)从候选特征的特征子空间中选择分裂特征和分裂点,不断循环直至达到停止条件,这样就构成了M个基分类器;最后将各基分类器的预测结果多数投票输出最终预测结果。

1.2 栈式泛化集成学习算法

传统随机森林框架组合单棵决策树,随着深度学习的兴起,深度架构及神经网络与RF逐渐融合,提出深度森林[16-18],神经网络是端到端的模式,而栈式泛化结构是有监督形成特征指导分类。Stacking堆叠算法[19]是由Woplert提出的一种串行集成学习框架,当单个泛化器使用堆叠算法时,堆叠泛化[20-21]是一种纠正泛化器错误的方案。本文中泛化器在特定训练集上训练,然后提出特定问题,使用堆栈泛化来改进单个泛化器的行为,而非将其用作组合泛化器的手段。

栈式结构就是以深度学习的方式堆叠多层随机森林,根据栈式泛化原理[16,22]将每层的预测结构作为下一层的输入,每个随机森林作为神经网络的基础神经元。该框架第1层采用多个基分类器对数据集进行训练,每个基泛化器输出相应的预测结果;将第1层的预测输出作为第2层的输入,然后对第2层重新构造的预测模型泛化器进行训练;再用第2层的模型输出预测结果,通过多层堆叠模型结果进行泛化,以此来提高模型预测精度。

1.3 隐私保护

差分隐私的实现机制是在输入或输出阶段加入随机化的噪音,屏蔽两个相邻数据集D和D'之间的差异,常见的隐私保护机制包括拉普拉斯机制[23]、指数机制[24]、随机响应机制[25]以及高斯机制[26],这些机制都可以在一定程度上保护数据的隐私,同时保持数据的可用性。拉普拉斯噪声[23]用于实现数值即连续型数据查询,指数机制[24]适用于非数值查询,处理的都是离散型的数据,这些机制适用于各种数据类型和隐私保护级别,因此在实际应用中必须根据具体的场景和需求选择最适合的机制来确保隐私得

到有效保护。BMS-Tree选择分裂特征和分裂值时需要对训练数据集的输出结果赋予不同的概率来实现隐私保护,因此本文需要对非数字查询的指数机制添加噪声,从而保护训练数据集中的敏感信息,避免在模型训练过程中泄露个人隐私。

定义1(ε-差分隐私) 对于任意相邻数据集D、D'的所有可能输出的集合为P,对于P的任何子集S,随机机制f满足式(1)则称随机机制f提供ε-差分隐私保护。

$$\Pr[f(D)\in S]\leqslant \exp(\varepsilon)\cdot\Pr[f(D')\in S] \tag{1}$$

式中:ε为隐私保护预算参数;$\Pr[\cdot]$为事件发生的概率。

定义2(全局敏感度) 全局敏感度作为量化噪声大小的参数,能反映个人信息保障的强度,表示删除数据集中某一记录对查询结果造成的影响。D是$\mathbf{R}^d$上具有d维特征的数据,$Q:(D,r)\to\mathbf{R}_d$是用于衡量输出r质量的数据集D的评分函数,对于相邻数据集D和D',满足

$$\Delta Q=\max_{D,D'}\|Q(D,r)-Q(D',r)\|_1 \tag{2}$$

式中$\|Q(D,r)-Q(D',r)\|_1$表示$Q(D,r)$和$Q(D',r)$之间的一阶范式距离。

定义3(指数机制) 当接收到一个查询之后,通过一定的概率值返回结果,从而实现差分隐私。该机制目标返回一个$r\in S$,在满足差分隐私的前提下,评分函数$Q(D,r)$得分高的输出概率高,得分低的输出概率低。指数机制选择r的概率与$\exp\left(\frac{\varepsilon Q(D,r)}{2\Delta Q}\right)$成正比,即

$$\Pr[f(D)=r]=\frac{\exp\left(\frac{\varepsilon Q(D,r)}{2\Delta Q}\right)}{\sum\limits_{r'\in S}\exp\left(\frac{\varepsilon Q(D,r')}{2\Delta Q}\right)} \tag{3}$$

2 本文算法

2.1 BMS-RF模型

BMS-RF模型旨在将若干随机森林进行栈式集成以提高分类的泛化性能,同时通过实现对每个随机森林内的单棵决策树的隐私保护来达到对整个栈式随机森林进行隐私保护的目的。BMS-RF模型构建的基本思想是在每一层随机森林的每一棵决策树构造过程中,使用了三重随机性:通过使用伯努利分布来选取候选特征以提升训练速度并降低存储开销;按照差分隐私的指数机制,对候选分裂特征和分裂点添加噪声以提供隐私保护;运用多项分布来随机选择分裂特征以及分裂点的方式来构造相应的决策树,进而形成随机森林。因为添加的随机性及噪声会对随机森林的性能产生一定影响,为此依据栈式泛化原理,通过栈式结构来构造后继的随机森林并进而形成栈式随机森林,即BMS-RF,进一步提高了模型的泛化性能。由于每棵决策树的构建过程能够提供一致性与隐私保护,因而整个栈式随机森林依然具有一致性和隐私保护机制,能够有效地解决分类问题中隐私保护与数据可用性难以平衡的问题。

图1展示了BMS-RF模型。模型从左向右由N层构成,第1层随机森林并行生成M个BMS-Tree,构造树分裂特征与分裂点的选择如上述决策树构造过程,根据M个BMS-Tree输出预测向量并进行随机投影,输出第1层的随机森林的预测并与原始数据集拼接形成第2层的输入,第2层同样并行生成M个不同的BMS-Tree,构造树过程同第1层一样,如此往复到第Layer_N层随机森林,最后顺序集成N层栈式随机森林的预测结果得到最终预测结果。

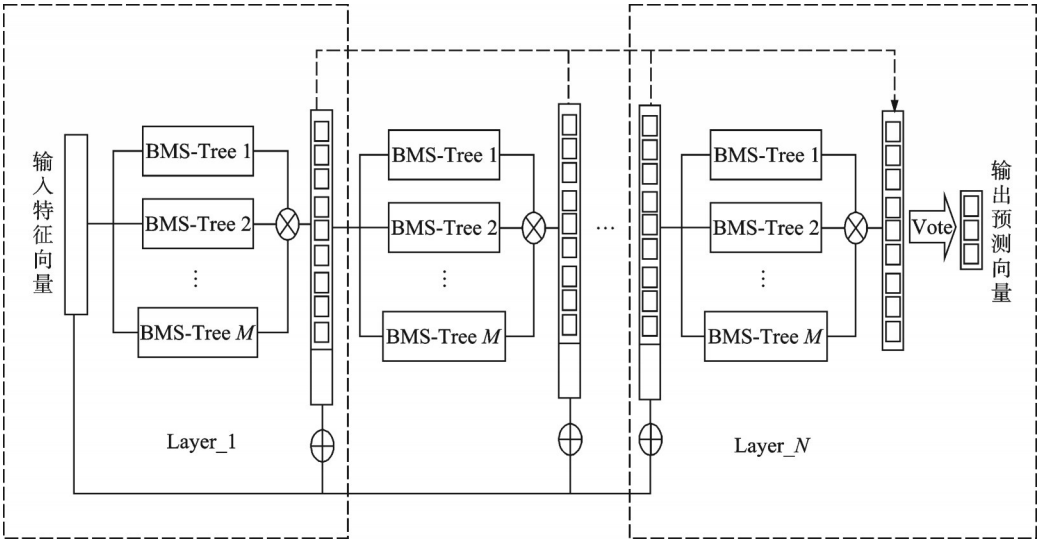

图1 BMS-RF模型

Fig.1 BMS-RF model

2.2 栈式随机森林集成学习

基于上述模型,下面具体讨论栈式随机森林集成学习过程。随着样本量及数据维度的增加,传统单层RF很难满足高精度分类要求,用栈式泛化原理将多个随机森林堆叠在一起,本文讨论的是多分类问题。以三分类为例,通过输入特征向量,最终输出为输入$\boldsymbol{X}$属于各个类别的概率,假设栈式堆叠层数为N层,每层包含M个BMS-Tree,每一层的输入是前一层的输出和原始特征,输出新特征向量包括M个基分类器和3个类别,通过原有预测结果的随机投影获得特征向量的随机位移,每个森林输出分类概率分布,也就是新特征,新特征向量与原始的输入特征向量$\boldsymbol{X}$拼接组成新的输入,每层训练数据的输入都与原始数据在数据空间上保持一致,进行下一层的生成。每生成并训练完一层的森林,在该层森林的测试集上验证表现是否继续提升,迭代到性能不能提升停止。由所有森林投票产生最终预测值,栈式结构随机森林串行集成各基分类器获得更好的泛化性能,由于数据集维度及样本规模的限制,本文栈式堆叠层数为2~4层。算法1以伪代码格式总结了BMS-RF集成学习过程,是一个完整的栈式结构随机森林集成学习过程。

算法1 BMS-RF集成学习

输入:训练集矩阵$\boldsymbol{X}$,随机投影系数α,森林的数目M,栈式堆叠层数N

输出:BMS-RF预测值

(1)$\boldsymbol{X}_{\text{Layer\_}N}=\boldsymbol{X}$

(2)for $i=1$ to N do

(3)根据式(4)更新输入向量,把上一层输入与本轮输出拼接作为下一层的输入

(4) for $j=1$ to M do

(5) 调用算法2构建BMS-Tree

(6) 根据式(5)输出第Layer_i层各类别概率

(7) end for

(8)根据式(5)计算第Layer_N层输出类别概率,根据式(6)对BMS-RF结果投票获得最终预测值

(9)end for

(10)返回:BMS-RF预测值

通常情况定义训练数据集$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$是$n$对数据构成的训练集,$\boldsymbol{x}_n$是特征向量,$y_n \in \{1, 2, \cdots, C\}$为相应特征的标签,其中标签共有$C$类。方便起见,定义矩阵$\boldsymbol{X}$为训练集输入特征矩阵,$Y$为训练样本的类标签,$\boldsymbol{X}_{\text{Layer_}i}$为第Layer_$i$层的新特征向量,$\boldsymbol{Z}_{\text{Layer_}i}$为第Layer_$i$层的输出,$\boldsymbol{W}_{\text{Layer_}i}$为第Layer_$i$层的随机投影矩阵,是元素为0至1的随机实数,$\alpha$为给定的较小常数,实验中设为0.35,依据栈式泛化原理得到新的特征向量为

$$\boldsymbol{X}_{\text{Layer\_}i+1} = \boldsymbol{X} + \alpha * \boldsymbol{Z}_{\text{Layer\_}i} * \boldsymbol{W}_{\text{Layer\_}i} \tag{4}$$

预测过程中,BMS-RF构建的分类器中未标记测试样本$\boldsymbol{x}$,可以通过后续算法2知道它落在决策树的哪个叶子节点上,使用投票策略获得最终输出,属于c类的概率见式(5),式(6)通过最大化$w^{(c)}(\boldsymbol{x})$给出预测。

$$w^{(c)}(\boldsymbol{x}) = \frac{1}{|N^E(\boldsymbol{x}, \Theta)|} \sum_{(X,Y) \in N^E(\boldsymbol{x}, \Theta)} \Pi\{Y = c\} \tag{5}$$

$$\hat{y} = \arg \max_{c \in \{1, 2, \cdots, C\}} \sum_{j=1}^{M} \Pi\{f^{(j)}(\boldsymbol{x}) = c\} \tag{6}$$

式中:$|N^E(\boldsymbol{x}, \Theta)|$表示样本$\boldsymbol{x}$的叶子节点中估计点的数目,其中$\Theta$表示构建树中的随机化变量;$\Pi(\cdot)$是判别函数,$\Pi(\cdot)$为真时取值1,否则为0;$f^{(j)}(\boldsymbol{x})$是第$j$棵树$f^{(j)}$对$\boldsymbol{x}$的预测,第Layer_$i$层森林输出一个向量$\boldsymbol{Z}_{\text{Layer_}i}$,最后将$N$层输出向量投票,输出类别$\hat{Y}$,单棵决策树停止条件与Breiman RF相同,与最小叶大小有关,条件限制仅限于估计点而不是整个训练数据集上。对于每个叶子,估计点的实际大小大于k,其中k是训练实例数n的低阶无穷大,$k \to \infty$且在$n \to \infty$时$k/n \to 0$。

基学习器的输出数据作为新的数据样本,构成了栈式结构的的下一层输入数据,能够有效保持源域空间特征,而且在栈式结构中,前一层预测结果随机偏移到原始训练集中,有助于更好地实现线性可分。下一层的算法能够发现并且纠正第一层学习中的预测误差,而上一层的训练结果也能充分运用于下一层的归纳过程,以提升模型的泛化能力。

2.3 决策树构造

现有的随机森林随机性主要体现在自举抽样、决策树构造和预测类型阶段。本文用训练集分区来代替原始随机森林的样本有放回抽样,训练集D随机分为两个不重叠的子集:D^{S}和D^{E},用于树的构造,选择最佳分裂特征与分裂点,D^{E}用于构造决策树之后对其进行规则分割并做出预测,对树的构造没有影响,定义两部分的比值为Partition rate=结构点/估计点,构建另一棵决策树时,训练集被随机独立地重新划分。随机森林在高维数据中的性能并没有在低维数据中显著,针对在高维数据中随机森林性能的提升,并且随着特征维度的增加,计算所有信息增益的时间消耗也会显著增加,在决策树构造阶段引入非均匀分布的Dropout[27]可以解决上述问题。算法2中以伪代码格式总结BMS-RF中单棵决策树的详细构造过程。分裂特征选择阶段在所有特征上进行伯努利试验获取候选特征,引入两个抖动变量p_1、$p_2 \in (0, 1)$,以p_1的概率从由结构点构成的向量$\hat{\boldsymbol{I}}$中选择$p_1 \times d$个候选特征向量,选择一部分特征归一化转换成概率生成候选特征,最后通过多项分布再随机选择一个最佳分裂特征,m个分裂点选择阶段以p_2的概率选择候选分裂点,与分裂特征同理。

算法2 BMS-Tree($D^{\text{S}}, D^{\text{E}}, k, B_1, B_2, B_3$)

输入：结构点D^{S}，估计点D^{E}，超参数k、B_1、B_2、B_3

输出：BMS-Tree及决策树预测值

(1)while不满足建树停止条件do

(2) 计算所有分裂点的不纯度减少量V_{ij}，用向量$\hat{\boldsymbol{I}}$表示d个特征向量的不纯度最大降幅

(3) 针对d个候选分裂特征进行伯努利试验，p_1概率选择$p_1 \times d$个特征向量，计算归一化向量$\hat{\boldsymbol{I}}$，计算概率$\phi = \mathrm{soft\,max}\left(\frac{B_1}{2}\hat{\boldsymbol{I}}\right)$，$B_1 \geqslant 0$是隐私预算相关的超参数

(4) 根据多项分布随机选择一个分裂特征

(5) 针对m个候选分裂点进行伯努利试验，以p_2的概率选择候选分裂点，计算所选分裂特征A_j的归一化向量$\hat{\boldsymbol{I}}^{(j)}$并计算概率$\varphi = \mathrm{soft\,max}\left(\frac{B_2}{2}\hat{\boldsymbol{I}}^{(j)}\right)$，$B_2 \geqslant 0$是与隐私预算相关的超参数

(6) 根据多项分布随机选取分裂值，D^{S}和D^{E}相应地分别被分成两个不相交的子集D^{Sl}、D^{Sr}和D^{El}、D^{Er}

(7) T.leftchild ← BMS-Tree($D^{\mathrm{Sl}}, D^{\mathrm{El}}, k, B_1, B_2, B_3$)

(8) T.rightchild ← BMS-Tree($D^{\mathrm{Sr}}, D^{\mathrm{Er}}, k, B_1, B_2, B_3$)

(9) if D^{El}和D^{Er}中估计点的数量都大于k then

(10) 执行第1步递归建树

(11) else

(12) 计算概率$\psi = (\psi_1, \psi_2, \cdots, \psi_m) = \mathrm{soft\,max}\left(\frac{B_3}{2}\hat{P}\right)$，$B_3 \geqslant 0$是与隐私预算相关的超参数

(13) 根据多项分布为每个叶子随机选择1个标签作为其代表

(14) End if

(15)End while

(16)返回：BMS-Tree及决策树预测值

决策树采用由基尼系数衡量的不纯度作为分裂准则，而且采用二分分裂的方式递归建树，同时把该节点分裂为左、右两个子节点，D^{Sl}和D^{Sr}分别是节点的左子集和右子集，$T(\cdot)$是基尼指数函数，以此类推，直到达到建树的停止条件。单棵决策树采用递归划分，$\{v_{ij}\}$表示所有候选分裂点集合，v_{ij}表示第j个特征的第i个值，I_{ij}表示相应分裂点的不纯度减少量，节点的不纯度减少量定义为

$$I(D^{\mathrm{S}}, v) = T(D^{\mathrm{S}}) - \frac{|D^{\mathrm{Sl}}|}{|D^{\mathrm{S}}|}T(D^{\mathrm{Sl}}) - \frac{|D^{\mathrm{Sr}}|}{|D^{\mathrm{S}}|}T(D^{\mathrm{Sr}}) \tag{7}$$

则$\hat{\boldsymbol{I}} = \{I_1, I_2, \cdots, I_d\} = \{\max\{I_{i,1}\}, \max\{I_{i,2}\}, \cdots, \max\{I_{i,d}\}\}$，其中$\max\{I_{i,j}\}$是特征$A_j$的最大不纯度减少量，估计点构成每片叶子预测向量$\hat{\boldsymbol{P}} = (p_1, p_2, \cdots, p_C)$，$i = 1, 2, \cdots, C$。

2.4 时间复杂度分析

对于n个样本、特征维度是d的数据集，平衡二叉树在最佳情况下深度为$O(\log n)$，构造树理论计算复杂度为$O(dn\log n)$。由于BMS-RF的复杂度是构造单个树的复杂度总和，因此分析基分类器也就是算法2的复杂度。算法2的时间消耗主要集中在步骤5和步骤7，候选分裂特征选择需要$O(p_1 d)$计算

量,候选分裂点选择需要$O(p_2 n\log n)$计算量,拥有M棵决策树的森林复杂度为$O(M\cdot(p_1 d)\cdot(p_2 n\log n))$。这种分析忽略了为每个节点选择分裂特征和值所涉及的计算时间。由于$p_1 d\approx\sqrt{d}$,因此Dropout过后的BMS-RF比多项随机森林(Multinomial RF,MRF)运行快。算法1由外循环和内循环组成,内循环通过调用算法2来构建每层随机森林,外层循环N次,外循环通过步骤5生成随机投影矩阵需要$O(nd)$计算量,拼接生成新特征向量需要$O(nCd)$计算量,因此栈式堆叠N层需要时间复杂度是$O\left(\sum_{\text{Layer}_i=1}^{N} M*(p_1 d)\cdot(p_2 n\log n)+nd+nCd\right)$。与大多数模型复杂度固定的不同,栈式结构可以通过在适当时终止训练来自适应地决定其模型复杂度。

2.5 BMS-RF的一致性和隐私分析

2.5.1 单层一致性

一致性是一种学习算法的基本理论性质,它保证了当数据的大小趋近于无穷大时,算法的输出趋向收敛至最优值。

定义4 D表示n组训练集合$(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)$,$\Theta$表示构建树中的随机化变量,$L^*_{\text{Bayes}}$表示Bayes误差,$(\boldsymbol{X},Y)$表示第一层的训练样本及类别,则分类器$f$的误差概率$L$为

$$\mathbb{E}(L)=\mathbb{P}\{f(\boldsymbol{X},\Theta,D)\neq Y\mid D\}\to L^*_{\text{Bayes}}\ \text{as}\ n\to\infty \tag{8}$$

引理1 投票分类器$\overline{f(M)}$对M个单独训练的分类器$\{f(i)\}_{i=1}^{M}$进行多数投票,其中f表示不同随机性Θ下的分类器,如果分类器f一致,则$\overline{f(M)}$具有一致性。

引理2 如果投票数据的标签对分类规则的结构没有影响,则$n\to\infty$时,$\mathbb{E}(L)\to L^*_{\text{Bayes}}$当且仅当满足两个条件,$\text{diam}(N(\boldsymbol{X},\Theta))\to 0$且$|N^{\text{E}}(\boldsymbol{X},\Theta)|\to\infty$,其中$N(\boldsymbol{X},\Theta)$代表包含$\boldsymbol{X}$的叶子节点,$|N^{\text{E}}(\boldsymbol{X},\Theta)|$表示其中估计点的数目。

证明 设$\text{Size}'(e)$是$N(\boldsymbol{X},\Theta)$第e个属性的大小,对于任意给定$e\in\{1,2,\cdots,d\}$,由伯努利随机森林可得[28]

$$\mathbb{E}[\text{Size}'(e)]\leqslant\left(1-0.25\frac{p_1*p_2}{d}\right)^{\omega} \tag{9}$$

式中:p_1、p_2是引入的两个抖动变量,ω表示从根节点到叶子节点的层数,$\omega\to\infty$,$\lim_{n\to\infty}\mathbb{E}[\text{Size}'(e)]=0$,也就是叶子节点的直径在概率上趋向于0。又因$\omega\to\infty$时,$n\to\infty$,所以分裂过程可以无限期地进行下去。

引理3 如果在分裂特征选择阶段引入的超参数B_1是上有界的,则存在下界J_1对于任意给定特征A被选择在每个节点分割的概率都有$J_1>0$,因此每个特征被选中的概率都是非零的。

引理4 如果所有未标记d维样本都满足$[0,1]^d$,在BMS-RF中一旦有1个特征A被分割,那么这个特征会从小到大分裂为m'个同等的部分,即$A^{(1)},A^{(2)},\cdots,A^{(m')},A^{(i)}=\left[\frac{i-1}{m'},\frac{i}{m'}\right]$。如果在分裂值选择阶段引入的超参数$B_2$是上有界的,则存在下界$J_2$对于任意给定特征$A$被选择在每个节点分割的概率都有$J_2>0$。

引理2表明分裂过程可以无限期进行下去,引理3[15]表明单层BMS-RF每个特征具有非零被选择概

率,引理4[15]表明叶子上的每个超立方体足够小且包含无限数量的估计点。由引理2~4可证明决策树的一致性,由引理1决策树的一致性可推知随机森林的一致性,即可证BMS-RF单层在概率上的一致性。

2.5.2 栈式多层一致性

引理5 对于足够大的样本,栈式结构中分裂点的累积分布函数(Cumulative distribution function, CDF)在0处右连续,在1处左连续,决策树的每个节点在概率上被分裂无限次。

证明 在样本大小为n的数据集中,落在叶子节点A中的样本遵循二项分布$B(n,p')$,一般性的情况下假设分区率为1,则A中估计点的数量是$np'/2$。设置每个叶子节点中样本数量的最小值为k,由于栈式堆叠过后训练样本与原始特征拼接增加且仍能保持源域空间特征,可以增强分类器的泛化性能,也就是训练样本下$(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1, Y)$包含的叶子节点中估计点的个数增加,$|N^{\mathrm{E}}(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1,\Theta)|-\sigma<k,\sigma>0$,其中$\sigma$表示栈式结构后估计点的增加个数。根据切比雪夫的不等式,可以求出$N^{\mathrm{E}}(\boldsymbol{X},\Theta)$的界,有

$$\mathbb{P}\left(|N^{\mathrm{E}}(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1,\Theta)|<k\right)=\mathbb{P}\left(|N^{\mathrm{E}}(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1,\Theta)|-\sigma-np'/2<k-\sigma-np'/2\right)\leqslant$$
$$\frac{1}{2}\mathbb{P}\left(\left||N^{\mathrm{E}}(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1,\Theta)|-\sigma-np'/2\right|>|k-\sigma-np'/2|\right)\leqslant\frac{1}{|k-\sigma-np'/2|^2}\tag{10}$$

当$n\to\infty$时,$k-\sigma-np'/2$仍然为负,当$n\to\infty$时,式(10)上限为0,根据BMS-RF构建树的停止条件,如果节点估计点的采样数量大于k,树将会无限频繁地生长分裂。

引理6 当基分类器f是一致的,栈式结构的BMS-RF仍然是一致的。

证明 为了证明BMS-RF的一致性,只需要确保栈式堆叠后各个决策树的一致性。对于第2层的训练数据D_1及训练样本及类别$(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1, Y)$,由多项随机森林引理得结构点估计点数据点切分中的随机性不会影响基决策树的一致性,那么假设在条件U下,分类器f具有一致性,其中U代表训练集栈式结构随机偏移输入数据的随机性。如果随机偏移训练数据时具有概率为1的一致性,那么分类器就具有一致性,栈式结构随机偏移的输入数据也不会影响基决策树的一致性,即

$$\mathbb{E}(L)=\mathbb{P}\{f(\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1,\Theta,D_1)\neq Y|U\}\to L^*_{\mathrm{Bayes}}\ \mathrm{as}\ n\to\infty\tag{11}$$

根据栈式泛化原理,$\boldsymbol{X}_{\mathrm{Layer}_i+1}=\boldsymbol{X}+\alpha*\boldsymbol{Z}_{\mathrm{Layer}_i}*\boldsymbol{W}_{\mathrm{Layer}_i},\alpha\to0$时,输入数据不进行偏移,由于$\boldsymbol{W}_{\mathrm{Layer}_i}$随机投影矩阵为0至1的实数随机值,$\alpha\to1$时,源域空间特征则会大概率改变,因此$\alpha$是个偏小的常数是合理的,第Layer_$i$层经过随机偏移的输入数据$\boldsymbol{X}+\alpha*\boldsymbol{Z}_1*\boldsymbol{W}_1$仍然分布在$[0,1]^d$上,并且处处都有非零密度,分裂点的CDF在0处右连续,在1处左连续,此部分证明见引理5。当$n\to\infty$时,$k\to\infty$且$k/n\to0$,BMS-RF每层的各个决策树就是一致的。因此栈式堆叠过后增强随机森林泛化能力的基础上,仍能保持BMS-RF的一致性。

2.5.3 BMS-RF隐私分析

面对较为复杂的隐私问题,对一个数据集进行了T次独立的查询,总的隐私预算控制在设定的阈值ε以内,这些单个查询的隐私预算就是$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T$,每个隐私机制$f_T$提供$\varepsilon_t$隐私保护,需要运用差分隐私的两个基本性质:

(1)顺序组合:假设随机机制$f=\{f_1,f_2,\cdots,f_T\}$依次作用于给定数据集D的相交子集,那么随机机制f满足$\sum_{t=1}^{T}\varepsilon_t$差分隐私;

(2)并行组合:假设随机机制$f=\{f_1,f_2,\cdots,f_T\}$作用于给定数据集D的不相交子集,那么随机机制f满足$\max(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T)$差分隐私。

引理7[15] 根据MRF及差分隐私组合性质可得,决策树中基于不纯度的分裂特征选择本质上是差分隐私的指数机制,满足B_1-差分隐私,基于不纯度的分裂点选择满足B_2-差分隐私,决策树中每个叶子的标签选择满足B_3-差分隐私。

引理8 单层随机森林满足$\sum_{t=1}^{M}\varepsilon_t$-差分隐私,栈式集成后仍满足$\varepsilon$-差分隐私。

证明 引理7证明单棵决策树的构造满足差分隐私的指数机制,而单层随机森林有M棵决策树,M个随机机制对每棵决策树提供隐私保护,由顺序组合性质可得单层随机森林满足$\sum_{t=1}^{M}\varepsilon_t$-差分隐私,栈式结构将单层分类器学习的知识添加到原始特征中形成$X_{\text{Layer_}t}$,源数据的特征信息进行保留并增强,新组合形成的训练数据集D_1是对原有训练集产生的随机较小偏移,因此仍然满足在数据集上顺序执行的一系列函数,栈式结构为N层,其中任意一层的BMS-RF满足$\sum_{t=1}^{N}\varepsilon_t$-差分隐私,由差分隐私顺序组合性质可得栈式集成后仍然满足ε-差分隐私。

3 实验与分析

3.1 UCI数据集

实验从UCI分类数据集中选取14个数据集进行测试,实例按照样本量从小到大进行排序,具体数据如表1所示。数据集来自不同的实际应用领域,涵盖低中高维样本、二分类和多分类任务,具有足够的代表性,以此评估BMS-RF的分类表现。数据集基本信息包括样本量、特征维度和类别数,为了减弱每个数据集的随机性,实验进行10倍交叉验证,每个算法运行10次取平均值。

表1 数据集描述

Table 1 Dataset description

数据集	样本量	特征维度	类别数
Zoo	101	16	7
Wpbc	198	34	2
Wdbc	569	30	2
Sonar	208	60	2
Movement_libras	360	90	15
Clean1	476	166	2
Australian	690	14	2
Vehicle	846	18	4
Banknote	1 372	4	2
WineQuality-red	1 599	11	6
Image	2 310	19	7
Wilt	4 839	5	2
WineQuality-White	4 898	11	7
Texture	5 500	40	11

3.2 实验结果

3.2.1 实验参数及环境设置

本文实验均在同一环境下完成,在Windows 11环境下搭建系统,Python版本为3.8,CPU为i7-12700,2.10 GHz,内存为16 GB,泛化能力分析中,M表示决策树的数量,d表示树的深度,树的深度$\leqslant 10$,隐私保护超参数$B_1=B_2=10$,$B_3\to\infty$,参数敏感性分析中,设置差分隐私预算ε,将其平均分配到森林中每棵决策树上,每层每棵树的隐私预算则为ε/M,根据差分隐私顺序组合性质,每个节点分配到的隐私预算二等分得到$B_1=B_2=\varepsilon/2(d\cdot M)$,$B_3=\varepsilon/M$,设置隐私预算搜索范围为$\{0.1,0.5,1\}$,详细实验参数设置见表2。

表2 实验参数设置

Table 2 Parameter settings

参数名称	参数设置	描述
M	$M\in\{15,20,30\}$	每层决策树数量
k	5	决策树最小叶子大小
Partition rate	1	分区率=结构点/估计点
p_1,p_2	$p_1,p_2\in\{0.05,0.1,0.2,0.4,0.6,0.8,1\}$	伯努利分布引入随机性
B_1,B_2,B_3	$B_1=B_2=10,B_3\to\infty$	隐私保护超参数
d	10	树最大深度
Layer_i	Layer_$i\in\{2,3,4\}$	栈式结构层数
ε	$\varepsilon\in\{0.1,0.5,1\}$	隐私预算

3.2.2 泛化能力分析

在本次实验中,不同的数据集模型分类精度最优时的参数也不尽相同,评价指标采用测试精度(Accuracy, Acc)和运行时间(Time),为了评估BMS-RF在低维和高维数据集上的泛化性能,在14个数据集上比较BMS-RF和其他对比算法的测试精度。

表3列出了Breiman RF和两个一致性RF在不同数据集上的指标对比,其他对比算法设置树的数量均为100,所有一致性RF变体的分区率设置为1。对比算法采用随机均匀地选择单个特征和分裂点的Breiman RF[1]算法,利用两个伯努利分布选择最佳分裂点和分裂值的Breiman随机森林(Breiman random forest, BRF)[28],通过基于杂质的多项式分布、分别随机选择1个分裂特征和1个分裂值的MRF[15],现有一致的RF变体采用不同级别的Breiman RF简化来保证一致性。对于伯努利分布留下的分裂特征与分裂点,栈式堆叠层数一般设为2~4层可以具有更好的泛化能力。同时本文还进行了Wilcoxon signed-rank检验[29],以展示MRF和标准RF的结果在显著性水平0.05上的差异,带"*"的结果是检验出有效的结果。表3结果显示,BMS-RF在隐私保护水平较低时在10个数据集上泛化能力都有较显著提升,特别是90维的Movement_libras数据集从73.43%上升至78.49%,166维的Clean1从84.09%上升至86.68%,而精度小范围提升的Wdbc、Australian数据集平均运行时间也得到大幅降低。BMS-RF能够在证明一致性与隐私保护的基础上维持甚至超越RF性能,平均运行时间比其他一致性RF都少。

除了与一致性随机森林和隐私保护随机森林进行比较外,还将BMS-RF在隐私预算为1时的结果与一些先进的随机森林的测试精度进行比较,结果如表4所示,此时BMS-RF的构成为30~60棵树,其他对比算法设置树的数量均为100,包括RoF[30]和倾斜随机森林[31-32]。RoF包含主成分分析随机森林(Rotating random forest-principal component analysis, RoF-PCA)与线性判别随机森林(Rotating random forest-linear discriminant analysis, RoF-LDA);基于多表面近似支持向量机(Multi surface approximate support vector machine, MPSVM)的正则化随机森林包括吉洪诺夫正则化随机森林(Tikhonov regularized random forest, MPSVM-T)、轴并行正则化随机森林(Axis parallel regularization random forest, MPSVM-P)以及零空间正则化随机森林(Zero space regularized random forest, MPSVM-N)。由表4可以看出,BMS-RF通过较少的基分类器数量可以达到先进随机森林相持平甚至更优越的性能,MPSVM-T算法在Clean1数据集上测试精度为85.92%,而BMS-RF仅通过60棵树(栈式结构(30+30))能达到

表3 不同RF在UCI数据集上的评价指标对比

Table 3 Comparison of evaluation indices of different RFs on the UCI dataset

数据集	RF		BRF		MRF		BMS-RF	
	Acc/%	Time/s	Acc/%	Time/s	Acc/%	Time/s	Acc/%	Time/s
Zoo	91.76	6.90	80.21	7.58	90.93	7.35	94.06*	6.90
Wpbc	79.82	126.87	76.24	56.09	76.26	296.90	79.11*	24.60
Wdbc	95.69	809.53	94.84	336.88	95.51	1 633.30	95.93	183.69
Sonar	82.05	141.61	77.10	71.09	77.07	488.65	79.47*	23.75
Movement_libras	77.22	544.04	53.86	343.25	73.43	3 057.00	78.49*	227.15
Clean1	83.66	6 229.80	79.86	465.48	84.09	6 229.85	86.68*	202.81
Australian	86.94	170.80	86.31	101.49	86.06	354.64	86.82	26.70
Vehicle	74.33	369.59	70.31	179.59	73.04	679.93	74.98*	172.00
Banknote	99.21	1 755.73	98.19	986.03	99.44	1 657.41	99.42	94.44
WineQuality-red	68.48	914.46	60.38	518.65	63.08	1 923.90	65.34*	178.37
Image	97.54	5 161.86	95.52	3 329.76	97.53	12 935.15	97.64*	3 406.50
Wilt	98.30	19 676.44	97.12	10 531.78	98.58	28 076.55	98.30	3 177.48
WineQuality-White	67.24	4 451.37	56.80	2 682.82	60.69	11 908.56	63.55*	2 543.55
Texture	97.36	—	97.40	—	97.15	—	98.37*	—

表4 在UCI数据集上先进的RF和BMS-RF的测试精度对比

Table 4 Accuracy comparison of the more advanced RFs and BMS-RF on the UCI dataset %

数据集	RoF-PCA	RoF-LDA	MPSVM-T	MPSVM-P	MPSVM-N	BMS-RF
Zoo	86.13	86.55	82.09	85.73	80.01	94.06
Wpbc	76.65	79.34	77.22	76.60	76.97	79.11
Wdbc	95.68	97.27	96.76	96.84	97.04	95.93
Sonar	79.83	81.30	81.91	80.39	79.21	79.47
Movement_libras	77.42	82.83	74.17	76.03	71.11	78.49
Clean1	83.98	85.95	85.92	84.50	85.13	86.68
Australian	86.84	87.32	86.99	87.04	86.91	86.82
Vehicle	75.95	74.30	74.36	74.27	72.04	74.98
Banknote	99.88	99.81	99.91	99.89	99.88	99.42
WineQuality-red	65.52	66.14	61.51	64.96	58.10	65.34
Image	96.08	96.24	96.08	96.74	93.60	97.64
Wilt	98.54	98.65	98.08	98.08	98.11	98.30
WineQuality-White	63.56	63.86	58.45	62.52	44.88	63.55
Texture	98.10	98.16	99.12	98.14	98.90	98.37

86.68%,从而达到简化结构和降低存储开销的目的。Vehicle数据集能够在满足一致性与隐私保护的基础上达到74.98%的测试精度,与先进随机森林达到基本一致的性能。

3.2.3 参数敏感性分析

表5列出了在14个数据集上BMS-RF参数p_1、p_2及栈式结构每层决策树个数M的取值。实验表明p_1、p_2取值在0.2～0.4之间,能够更好地平衡BMS-RF的测试精度和运行时间,每层决策树个数确定在15～30棵树之间能够取得更好的泛化性能,同时栈式结构在2～4层能够获得更好的效果,太多层次的栈式结构可能会导致过拟合并延长算法所需时间,因此总的决策树个数控制在30～60棵。

表5 不同数据集下BMS-RF的p_1,p_2及每层决策树个数M值

Table 5 p_1,p_2 and the number of decision trees M in each layer of BMS-RF under different datasets

数据集	p_1、p_2	M
Zoo	0.2,0.8	60: (20+20+20)
Wpbc	0.4,0.2	60: (20+20+20)
Wdbc	0.2,0.8	60: (30+30)
Sonar	0.2,0.2	60: (30+30)
Movement_libras	0.2,0.4	60: (15+15+15+15)
Clean1	0.2 ,0.2	60: (30+30)
Australian	0.2,0.2	30: (15+15)
Vehicle	0.6,0.2	60: (20+20+20)
Banknote	0.2,0.2	60: (30+30)
WineQuality-red	0.4,0.2	60: (30+30)
Image	0.4,0.4	40: (20+20)
Wilt	0.4,0.2	30:(15+15)
WineQuality-white	0.4,0.2	60: (20+20+20)
Texture	0.4,0.2	60: (20+20+20)

选取3个代表性数据集(Movement_libras、Sonar、WineQuality-red)作为本次实验的数据集,观察未进行栈式结构的BMS-RF在数据集上的测试精度和运行时间的变化,评估这两个参数p_1、p_2对泛化性能和运行速度的影响,结果如图2,3所示。p_1、p_2是为了解决泛化性能与运行时间问题所引入的两个抖动变量,实验证明p_1、p_2值为0.2～0.6时,伯努利分布对BMS-RF的精度影响不大,Movement_libras、Sonar、WineQuality-red数据集的最佳取值分别为0.2,0.4、0.2,0.2、0.4和0.2,Dropout过后在大部分数据集上可以与MRF性能持平并显著提高运行速度,一定比例的Dropout对于高维数据集Movement_libras、Sonar、Australian数据集都取得了更好实际性能。图2(a)看出维度不是太高的WineQuality-red数据集p_1、p_2相对较小,测试精度也相对下降;从图2(b)看出,90维的Sonar数据集随着p_1、p_2值的减小测试精度保持平稳提升,并未出现数据集的严重失真。由图3(a～c)可看出,当p_1不变,随着p_2减小,BMS-RF的运行时间始终保持下降的趋势,这证明了伯努利Dropout在提高BMS-RF的运行速度方面具有较好的可行性;d、m取值较大时,伯努利分布更倾向于高斯分布,往往能为分类器提供更强的泛化性能。对于维度较小的数据集,p_1、p_2为0.05时扰动较大,特征选择阶段渗透的抖动都可能导致数据的失真,测试

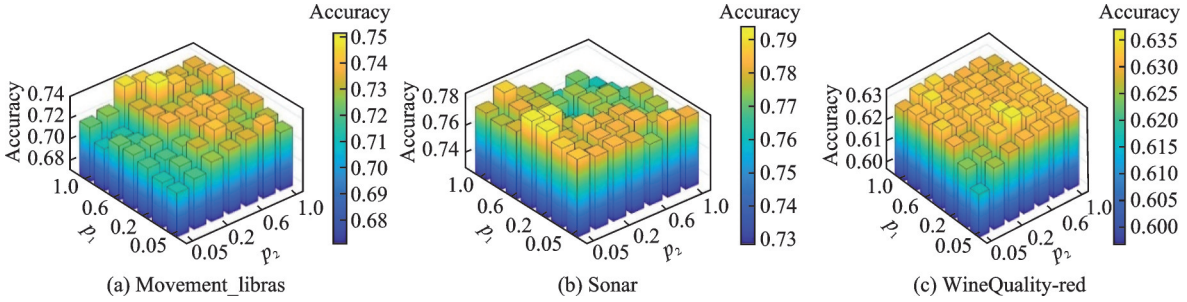

图2 BMS-RF在参数p_1、p_2下的测试精度

Fig.2 Test accuracy of BMS-RF under p_1 and p_2 parameters

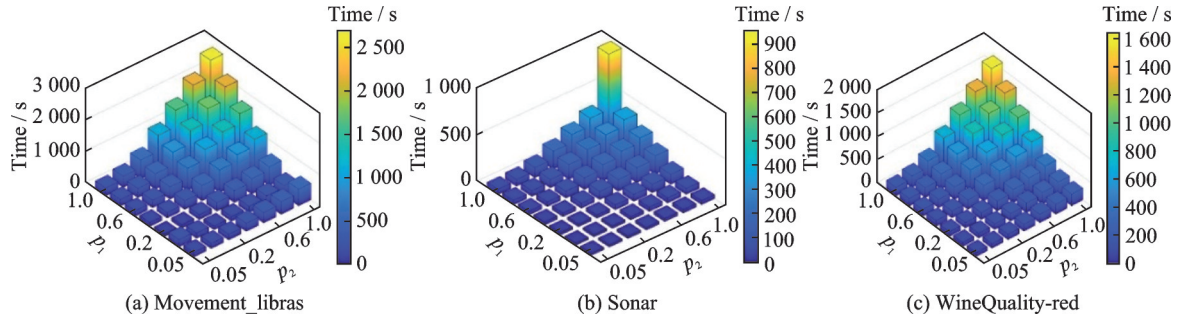

图3 BMS-RF在参数p_1、p_2下的运行时间

Fig.3 Running time of BMS-RF under p_1 and p_2 parameters

精度会有反向变化。当p_1、p_2概率逐渐扩大,也就是保留的特征向量逐渐增多,泛化性能会逐渐增大。由图2(a~c)可看出,p_1、p_2取值为0.2~0.4是合理的;p_1、p_2取值趋向于1时,BMS-RF泛化能力接近于Breiman;p_1、p_2取值逐渐减小至0时相当于随机均匀选择单个分裂特征与分裂点生长树的Biau08[33],Dropout后划分节点会计算更少的候选分裂特征与分裂点,所有数据集在测试精度基本一致的情况下运行速度得到显著加快,因此Bernoulli dropout是一个比较好的选择。

图4展示了BMS-RF在$B_1=B_2=10$时总的决策树数量为{3, 10, 20, 30, 40, 50, 60}的测试精度。BMS-RF栈式结构默认为3层,$B_1=B_2=10$时隐私预算较大,也就是隐私保护程度较小。由上述8个数据集可得,在基分类器数量小于20时,随机扰动可能会导致BMS-RF的分类性能降低;在决策树数量大于30时(栈式结构为(10,10,10))后,BMS-RF相较MRF泛化性能显著提升,由此可见栈式结构能显著提高基分类器的泛化性能;当决策树数量达到60棵、栈式结构2~4层时BMS-RF算法能够取得最佳的泛化性能。其中Sonar数据集经过栈式结构(20,20,20)精度从75.29%提升至79.22%。

为了平衡数据可用性与安全性的问题,图5展示了8个数据集在不同隐私预算下的分类效果。根据图5可以看到,对于多数数据集,BMS-RF能够在较高的隐私保护水平下具有比多项随机森林MRF更好的泛化性能。当决策树的数量较小时,对于隐私预算ε越小的情况,即添加噪声越多,隐私保护程度越高,其对应的BMS-RF的泛化性能呈现下降的趋势。反之,当隐私保护预算ε越大时,测试精度越高。当决策树数量超过30棵时,各数据集在不同隐私预算下,其测试精度的提升趋于平稳,可见在隐私预算较小、隐私保护程度较高的情况下,随着决策树数量M的不断增大,BMS-RF的分类性能仍能与MRF持平甚至超越,并且在不同隐私预算下测试分类精度之间的差距逐渐缩小。随着栈

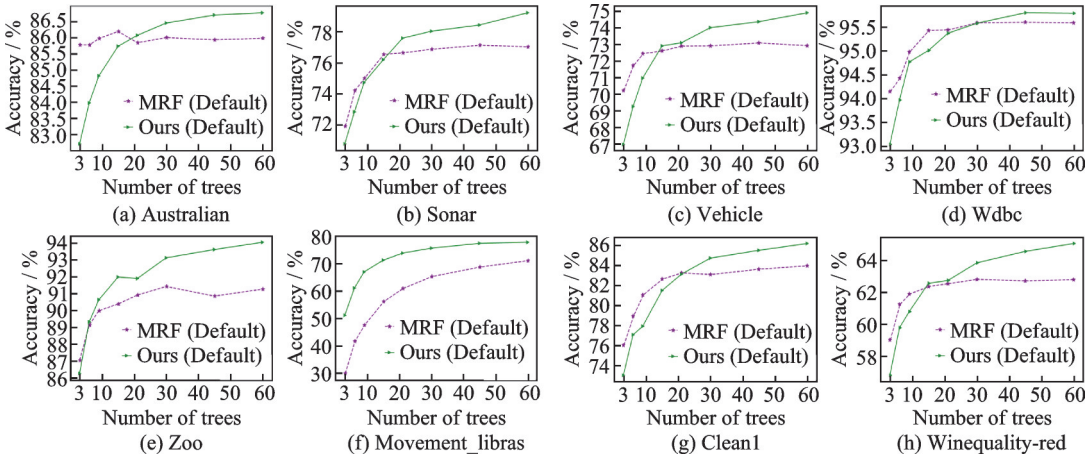

图4 BMS-RF在$B_1=B_2=10$时8个数据集不同决策树数量的测试精度

Fig.4 Testing accuracy of BMS-RF with different number of decision trees on eight datasets under $B_1=B_2=10$

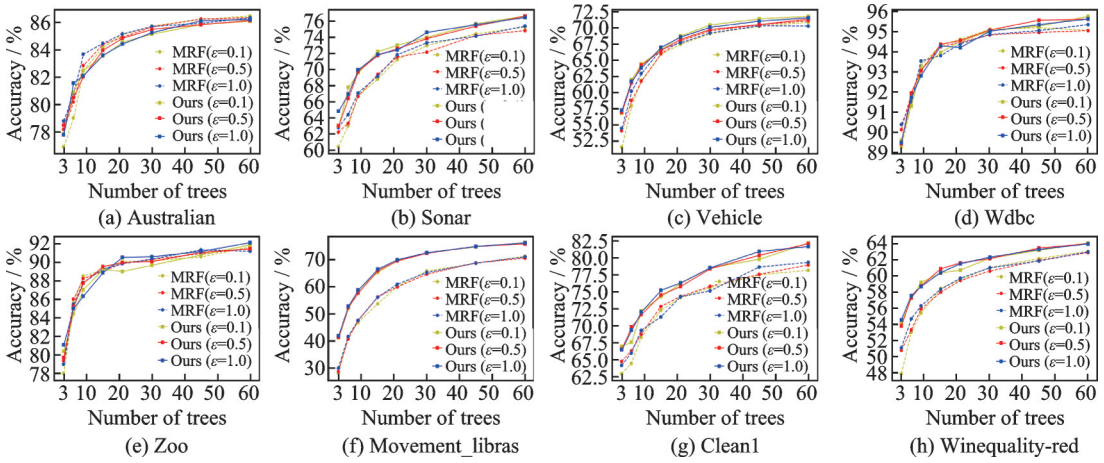

图5 不同隐私预算与决策树数量下算法在8个数据集上的分类准确性

Fig.5 Classification accuracy of the algorithm on eight datasets under different privacy budgets and numbers of decision trees

式结构中的每层决策树数量减少，在总的隐私预算固定的情况下每棵决策树所能分得的隐私预算将随之增多。按照图5所示，栈式结构的随机森林BMS-RF确实在大多数情况能够提高隐私机制下的泛化性能。特别地，对于Australian、Zoo数据集，BMS-RF的测试精度与MRF相比，并不总是提升甚至有所下降。对于Sonar、Vehicle数据集，BMS-RF在$\varepsilon=0.1$的测试精度甚至比$\varepsilon=1$时还高。这些特殊情况的出现可能与数据本身及当时的随机扰动有关，更多的分析和验证将在未来研究中进一步进行探索。

4 结束语

本文针对随机森林所面临的隐私保护下的泛化能力提升问题，以Dropout、栈式结构原理为基础，通过多随机森林中每棵决策树引入多重随机性以及栈式结构来集成多个随机森林，提出了一种基于多重随机性与隐私保护的栈式随机森林BMS-RF。在UCI数据集上的实验结果表明：在满足隐私机制的

前提下,由于栈式结构的使用,BMS-RF在大多数情况下能够提升分类的泛化能力并降低训练时间。此外,与Breiman RF以及一些先进随机森林的相比,BMS-RF的泛化效果尚有改进的空间。因此,如何在理论和实践中获得更好的泛化性能,以及针对数据集如何确定合适的堆叠层数和决策树个数,将是未来研究的方向。

参考文献:

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

[2] 杜卓铭,张军峰,杨春苇.基于分类与优化的进场航空器调度方法[J].南京航空航天大学学报,2023,55(6): 1065-1071.
DU Zhuoming, ZHANG Junfeng, YANG Chunwei. Approach aircraft scheduling method based on classification and optimization[J]. Journal of Nanjing University of Aeronautics &. Astronautics, 2023, 55(6): 1065-1071.

[3] PRASAD A M, IVERSON L R, LIAW A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction[J]. Ecosystems, 2006, 9: 181-199.

[4] ROTA BULO S, KONTSCHIEDER P. Neural decision forests for semantic image labelling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 81-88.

[5] XIONG C, JOHNSON D, XU R, et al. Random forests for metric learning with implicit pairwise position dependence[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2012: 958-966.

[6] BIAU G, SCORNET E, WELBL J. Neural random forests[J]. Sankhya A, 2019, 81: 347-386.

[7] DENIL M, MATHESON D, DE FREITAS N. Narrowing the gap: Random forests in theory and in practice[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2014: 665-673.

[8] RODRIGUEZ J J, KUNCHEVA L I, ALONSO C J. Rotation forest: A new classifier ensemble method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630.

[9] YE Y, WU Q, HUANG J Z, et al. Stratified sampling for feature subspace selection in random forests for high dimensional data[J]. Pattern Recognition, 2013, 46(3): 769-787.

[10] NGUYEN T T, ZHAO H, HUANG J Z, et al. A new feature sampling method in random forests for predicting high-dimensional data[C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Berlin: Springer, 2015: 459-470.

[11] SWEENEY L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571-588.

[12] LEBANON G, SCANNAPIECO M, FOUAD M R, et al. Beyond k-anonymity: A decision theoretic framework for assessing privacy risk[C]//Proceedings of Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006. Rome, Italy: Springer Berlin Heidelberg, 2006: 217-232.

[13] LI N, LI T, VENKATASUBRAMANIAN S. t-closeness: Privacy beyond k-anonymity and l-diversity[C]//Proceedings of 2007 IEEE 23rd International Conference on Data Engineering. [S.l.]: IEEE, 2006: 106-115.

[14] XIAO X K, TAO Y F. m-invariance: towards privacy preserving re-publication of dynamic datasets[C]//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2007: 689-700.

[15] BAI J, LI Y, LI J, et al. Multinomial random forest[J]. Pattern Recognition, 2022, 122: 108331.

[16] VINCENT P, BENGIO Y. Kernel matching pursuit[J]. Machine Learning, 2002, 48: 165-187.

[17] SU R, LIU X, WEI L, et al. Deep-resp-forest: A deep forest model to predict anti-cancer drug response[J]. Methods, 2019, 166: 91-102.

[18] ZHOU Z H, FENG J. Deep forest: Towards an alternative to deep neural networks[C]//Proceedings of IJCAI. Melbourne: International Joint Conferences on Artificial Intelligence Organization, 2017: 3553-3559.

[19] LEJEUNE D, JAVADI H, BARANIUK R. The implicit regularization of ordinary least squares ensembles[C]//Proceedings of International Conference on Artificial Intelligence and Statistics. [S.l.]: PMLR, 2020: 3525-3535.

[20] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.

[21] ZHANG Guoling, WANG Xiaodan, Li Rui, et al. Extreme learning machine based on stacked noise reduction sparse autoencoder[J]. Computer Engineering, 2020,46 (9): 61-67.

[22] 段友祥,赵云山,马存飞,等.基于多层集成学习的岩性识别方法[J].数据采集与处理,2020,35(3): 572-581.
DUAN Youxiang, ZHAO Yunshan, MA Cunfei, et al. Lithology identification method based on multi-layer ensemble learning [J]. Journal of Data Acquisition and Processing, 2020, 35(3): 572-581.

[23] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Proceedings of Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006. New York: Springer Berlin Heidelberg, 2006: 265-284.

[24] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). [S.l.]: IEEE, 2007: 94-103.

[25] WANG Y, WU X, HU D. Using randomized response for differential privacy preserving data collection[C]//Proceedings of EDBT/ICDT Workshops. Aachen: CEUR-WS.org, 2016, 1558: 0090-6778.

[26] LIU F. Generalized Gaussian mechanism for differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(4): 747-756.

[27] SHEN X, TIAN X, LIU T, et al. Continuous dropout[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(9): 3926-3937.

[28] WANG Y, XIA S T, TANG Q, et al. A novel consistent random forest framework: Bernoulli random forests[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(8): 3510-3523.

[29] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets[J]. The Journal of Machine Learning Research, 2006, 7: 1-30.

[30] ZHANG L, SUGANTHAN P N. Random forests with ensemble of feature spaces[J]. Pattern Recognition, 2014, 47(10): 3429-3437.

[31] KATUWAL R, SUGANTHAN P N, ZHANG L. Heterogeneous oblique random forest[J]. Pattern Recognition, 2020, 99: 107078.

[32] ZHANG L, SUGANTHAN P N. Oblique decision tree ensemble via multisurface proximal support vector machine[J]. IEEE Transactions on Cybernetics, 2014, 45(10): 2165-2176.

[33] BIAU G, DEVROYE L, LUGOSI G. Consistency of random forests and other averaging classifiers[J]. Journal of Machine Learning Research, 2008, 9(9): 2015-2033.

作者简介:

宋奕霖(1999-),女,硕士研究生,研究方向:人工智能、模式识别,E-mail: 2979308103@qq.com。

王士同(1964-),**通信作者**,男,教授,博士生导师,研究方向:人工智能、模式识别、生物信息,E-mail: wxwangst@ali-yun.com。

(编辑:刘彦东)