

DOI:10.11918/201912140

基于差分隐私下包外估计的随机森林算法

李玉强¹, 陈鋆昊¹, 李琦¹, 刘爱华²

(1. 武汉理工大学 计算机科学与技术学院, 武汉 430063; 2. 武汉理工大学 能源与动力工程学院, 武汉 430063)

摘要: 针对差分隐私随机森林算法在对高维数据进行分类时准确率不理想的问题, 本文通过引入差分隐私下的包外估计来计算决策树权重以及特征权重, 从而提出一种基于差分隐私下包外估计的随机森林算法 (random forest under differential privacy based on the out-of-bag estimate, RFDP_OOB). 本算法首先在差分隐私保护下生成一部分的随机森林, 利用差分隐私下包外估计的特性对决策树和特征的重要性进行评估, 从而计算出决策树权重以及特征权重, 然后通过特征权重对特征进行划分, 得到非重要特征集. 接着在生成剩下的一部分随机森林的过程中, 对最佳特征为非重要特征的结点进行预剪枝操作, 使其成为叶子结点, 从而减小噪声、提高决策树分类准确率, 并具有较好的执行效率. 最后在预测分类结果时, 取所对应的决策树权重最大的分类结果作为随机森林算法的分类结果, 从而提高随机森林算法的分类准确率. 本文还对算法的有效性和隐私性进行了理论分析, 并通过实验结果验证了本算法的有效性, 本算法可以在保护数据隐私性的同时提高算法的分类准确率.

关键词: 差分隐私; 随机森林; 包外估计; 高维数据; 数据挖掘

中图分类号: TP391

文献标志码: A

文章编号: 0367-6234(2021)02-0146-09

Random forest algorithm under differential privacy based on out-of-bag estimate

LI Yuqiang¹, CHEN Junhao¹, LI Qi¹, LIU Aihua²

(1. School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China;

2. School of Energy and Power Engineering, Wuhan University of Technology, Wuhan 430063, China)

Abstract: Since the accuracy of random forest algorithm under differential privacy is undesirable when classifying high-dimensional data, the out-of-bag estimate was introduced to calculate the weights of decision trees and features, and the random forest algorithm under differential privacy based on the out-of-bag estimate (RFDP_OOB) was proposed. First, the algorithm generates a part of random forest under differential privacy, and the importance of decision trees and features is evaluated by utilizing the out-of-bag estimate under differential privacy, so as to calculate the weights of the decision trees and features. Then, the features are re-divided into non-essential features through feature weights. Next, in the process of generating the remaining part of the random forest, the pre-pruning operation is performed on the nodes whose best features are non-important features to make them leaf nodes, so as to reduce noise and improve the classification accuracy of the decision tree with better efficiency. Finally, in predicting the classification results, the classification result with the maximum weight of the corresponding decision tree is taken as the classification result of the random forest algorithm, thereby improving the classification accuracy of the random forest algorithm. The privacy and effectiveness of the algorithm were analyzed theoretically, and the experimental results verified the effectiveness of the algorithm. The proposed algorithm can improve the classification accuracy and protect the privacy of data.

Keywords: differential privacy; random forest; out-of-bag estimate; high-dimensional data; data mining

随着大数据时代的来临, 在利用各种新技术把生活中丰富的数据搜集存储起来, 以便于进行研究的同时, 个人隐私数据的泄露也成为了当今社会的一大问题^[1]. 近年来, 隐私泄露事故的不断发生在国内外都造成了很大的社会恐慌, 如 2018 年 3 月 15 日, Facebook 被曝涉嫌泄露用户隐私数据, 带来了严重的经济问题和不良的社会影响. 于是, 隐私保护问题引起了人们的高度重视, 尤其在数据挖掘领

域, 当数据挖掘者对数据进行处理并获取有价值的信息时, 必然会给数据集中的隐私信息带来泄露的风险^[2-3]. 因此差分隐私^[4]以其严格性在数据挖掘领域得到了广泛的应用^[5]. 其中随机森林作为数据挖掘领域中一种重要的分类方法, 在数据预测分析中起着关键作用^[6], 将差分隐私技术应用到随机森林算法中可以保护数据的隐私性, 但是这样必然会造成算法分类准确率的大幅度下降, 尤其是对高维数据进行分类的时候, 对此研究者们做了许多工作.

Jagannathan 等^[7]最早将差分隐私保护应用在

收稿日期: 2019-12-26

作者简介: 李玉强 (1977—), 男, 副教授, 硕士生导师

通信作者: 陈鋆昊, 472769019@qq.com

随机森林中, Patil 和 Singh^[8] 进一步提出了一种保证差分隐私条件下的多类别分类算法 (DiffPRF), 但是没有提出进一步的改进方法. 穆海蓉等^[9] 在此基础上对算法进行了改进, 提出了一种面向随机森林的差分隐私保护算法 (DiffPRFs), 该算法利用指数机制在结点划分时选择分裂特征和分裂值, 从而降低了因离散化预处理而产生的系统性能消耗, 但是两次调用指数机制使得隐私预算过小, 导致噪声过大, 从而降低了随机森林的分类准确率. 为了提高算法的分类准确率, Xin 等^[10] 通过划分不相交的子集, 再利用差分隐私的并行组合性提高决策树的隐私预算, 提出了差分隐私贪婪决策森林算法 (DPGDF). 但是子集数量受到数据数量和隐私预算限制, 在子集数量很小的情况下此算法会退化为决策树算法, 这也意味着该算法本质上为决策树算法, 不具备随机森林算法的随机性. 在前人研究的基础上, Xiang 等^[11] 提出了差分隐私下的协作随机森林算法 (CRFsDP), 该算法用验证集来计算决策树权重, 并对决策树进行后剪枝处理, 将剪掉的分支的隐私预算分配给父结点以达到减少噪声、提升分类性能的效果. 但是需要先生成满二叉树, 这使得算法的时间复杂度较高, 另外通过验证集得到的决策树权重不能精确地估计决策树的分类效果.

因此, 在决策树权重计算和剪枝方式上还有待改进, 这也导致这些算法在具有高维度的数据集上的分类效果仍然不太理想. 而如何在差分隐私下对具有高维度的数据集进行更精确地分类, 一直是研究者们研究的重点^[12].

针对未添加差分隐私的随机森林算法中特征筛选和决策树筛选的方面, Paul 等^[13] 提出了一种改进的随机森林算法 (IRF), 利用包外估计^[14] 计算决策树权重和特征权重, 然后通过不断迭代更新, 选择表现良好的特征以及决策树. 但是该算法没有应用差分隐私保护, 不受隐私预算限制, 可以不断迭代. 而在差分隐私保护机制下需要预先向决策树分配隐私预算, 所以需要在有限的迭代次数中完成; 同时为了保护数据隐私, 包外估计也更需要加入差分隐私保护.

鉴于此, 本文通过引入差分隐私下的包外估计, 计算决策树权重以及特征权重, 从而提出一种基于差分隐私下包外估计的随机森林算法 (RFDP_OOB). 一方面利用特征权重减少决策结点上非重要特征的使用, 并以此来指导先剪枝操作, 进而在较低时间复杂度的前提下提高决策树的分类准确率; 另一方面利用决策树权重进行集成提升, 从而提高整个随机森林的分类准确率.

1 算法相关定义

1.1 差分隐私及其特性

定义 1 (差分隐私) 对于一个算法 M , 若其满足 ε -差分隐私, 则

$$P_r[M(D) \in S_m] \leq e^\varepsilon P_r[M(D') \in S_m]. \quad (1)$$

式中: S_m 为算法 M 可以输出的所有值集合的任意子集, D 和 D' 为差别至多为一条记录的两个数据集, ε 为隐私保护预算.

差分隐私保护技术本身蕴含着序列组合性与并行组合性两种重要的组合性质^[15]. 这两种性质在证明算法的隐私性以及隐私预算分配过程中起着重要作用.

性质 1 (序列组合性)^[15] 给定数据库 D 与 n 个随机算法 $A_1, \dots, A_n$, 且 $A_i (1 \leq i \leq n)$ 满足 ε_i -差分隐私, 则 $\{A_1, \dots, A_n\}$ 在 D 上的序列组合满足 ε -差分隐私, 其中 $\varepsilon = \sum \varepsilon_i$.

性质 2 (并行组合性)^[15] 设 D 为一个隐私数据库, 被划分成 n 个不相交的子集, $D = \{D_1, \dots, D_n\}$, 设 A 为任一个随机算法满足 ε -差分隐私. 则算法 A 在 $\{D_1, \dots, D_n\}$ 上的系列操作满足 ε -差分隐私.

1.2 差分隐私下的包外估计及其应用

1.2.1 差分隐私下的包外估计

在生成每棵决策树时需要随机且有放回地抽取样本, 因此会有大概 1/3 的数据未抽取到, 这些数据就是该决策树的包外数据. 在随机森林算法下, 数据集中特征的重要性、决策树分类能力和相关性计算都依赖于包外数据^[16]. 而用包外数据在该决策树上进行决策, 错误分类数据占包外数据总数的比率就是包外估计 (out-of-bag estimate). 经验证, 包外估计是对集成分类器泛化误差的无偏估计^[14], 可以准确地估计决策树的分类能力. 但是为了保护包外数据的隐私, 本文引入差分隐私下的包外估计, 在计算包外估计时添加噪声以满足差分隐私条件, 定义如下:

定义 2 (差分隐私下包外估计) 对随机森林中的一棵决策树, 差分隐私下的包外估计 B 为

$$B = \frac{M + N(\varepsilon)}{O}. \quad (2)$$

式中: O 为包外数据大小, M 为错误分类数据大小, ε 为该决策树隐私预算, $N(\varepsilon)$ 为差分隐私噪声函数.

可以看出, 加入的差分隐私噪声 $N(\varepsilon)$ 扰乱了真实的错误分类数据量, 但是根据差分隐私定义, 只需要添加少量的噪声就可以达到保护隐私的目的, 而不会使数据丧失可用性、失去规律, 在本文中选用 Laplace 噪声函数. 所以 B 越小, 则代表该决策树的

分类准确率更高.

1.2.2 决策树筛选

由于包外估计可以简单有效地估计决策树分类能力,本文参照 Paul 等^[13]的方法计算决策树权重,并用权重来代表该决策树的分类能力,进而进行决策树筛选.因此,将决策树 t 的权重 W_t 定义为

$$W_t = \frac{1}{B_t}. \quad (3)$$

式中 B_t 为该决策树差分隐私下的包外估计.不难看出,决策树权重越高,则代表着该决策树分类时错误分类的包外数据量越少,也就代表着该决策树的分类能力越好.

1.2.3 特征筛选

除了决策树筛选,本文还利用包外估计可以有效地评估数据集中特征的重要性的能力,在决策树构成的过程中参照 Paul 等^[13]提出的方法对数据集中的特征进行筛选.

1) 特征权重. 要进行特征筛选首先需要给出特征的评价标准,本文将其定义为特征权重. 参照文献 [13] 中的思想,首先定义特征在结点上的权重,其具体定义如下.

定义 3(特征在结点上的权重) 在决策树中某个非叶子结点上,特征 j 在该结点上的权重为

$$W_j = 1 - G(j). \quad (4)$$

式中 $G(j)$ 为基尼指数. 从基尼指数的定义可以知道,其值越小则表明按照此特征划分的效果越好. 所以当特征权重越大时,代表着基尼指数越小,分类效果越好. 而一棵决策树上有多个结点,为了更好地衡量特征权重,取结点上特征权重的均值作为该特征在决策树中的权重,定义如下.

定义 4(特征在决策树中的权重) 特征 j 在决策树中的权重为

$$\overline{W_j} = \frac{\sum W_j}{N}. \quad (5)$$

式中: N 为决策树中非叶子结点数量, W_j 为非叶子结点上特征 j 的特征权重.

但是每棵决策树的分类性能各不相同,因此取决策树中的特征权重的加权平均值,来衡量该特征在随机森林算法中的重要性,定义如下.

定义 5(特征在随机森林算法中的权重) 对 t' 棵决策树,某特征在随机森林算法中的权重为

$$\frac{\sum_t \overline{W_j} * W_t}{t'}. \quad (6)$$

式中, $\overline{W_j}$ 为特征 j 在决策树中的权重, W_t 为决策树

权重.

2) 特征划分. 在本文中用特征在随机森林算法中的权重(以下简称特征权重)来衡量特征的重要性,并据此对特征进行划分,找出非重要特征,进而对决策树进行改进. 而在划分的过程中需要找出离群特征,定义如下.

定义 6(离群特征) 对已有权重的特征子集 R , 离群特征满足条件

$$W_j < (\mu - 2\sigma). \quad (7)$$

式中: W_j 为特征 j 的特征权重, μ 为 R 中权重的平均值, σ 为标准差.

从定义中可以看出,离群特征的权重较小,而且与特征权重的平均值都相差甚远,所以在本文中离群特征认定为相比于其他特征重要性更低的特征. 进而可以根据特征的权重值对特征进行划分

$$\Gamma = \Gamma - A. \quad (8)$$

$$R = R + A. \quad (9)$$

式中: Γ 为权重前 f 大的特征集,其中 f 为决策树特征数量, R 为 Γ 的补集, A 为 Γ 中的离群特征集. 然后找出 R 中的离群特征,记为非重要特征 Γ' . 这些特征一旦被标记为重要特征或非重要特征,标记将不会被改变.

在随机森林生成的过程中,需要更新重要特征 Γ 和非重要特征 Γ' ,即:保持 Γ 和 Γ' 中已有特征不变,对其余特征 R ,将特征放入重要特征集 Γ 中,若其满足

$$W_j > W_{\min}. \quad (10)$$

式中 W_j 为特征 j 的特征权重, $W_{\min}$ 为重要特征 Γ 中的最小权重. 然后取出 R 中的离群特征,放入非重要特征集 Γ' 中.

3) 特征筛选下决策树的生成. 在对特征进行了划分后,就可以利用非重要特征在决策树构成的过程中对数据集中的特征进行筛选. 在决策树生成过程中,如果某个结点最优决策特征是非重要特征,则将该结点变为叶子结点,所有隐私预算用于叶子结点上计数值的加噪,使得噪声总量减小,从而提高分类准确率.

如图 1 所示,左侧为不进行特征筛选时生成的决策树,此时在根结点的右孩子结点上,最优决策特征是一个非重要特征. 如果进行特征筛选,则会在该结点上停止子结点的生成,即将该结点变为叶子结点,并将该结点所有的隐私预算对计数值添加噪声,最后得到如图 1 中右侧所示的决策树.

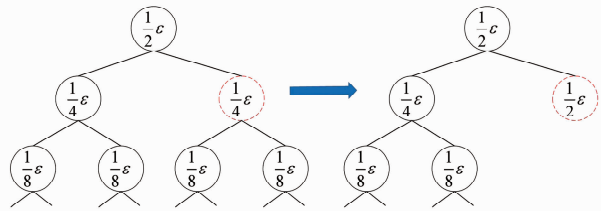


图 1 特征筛选下决策树的生成

Fig. 1 Generation of decision tree under feature selection

从图 1 中可以看出,若不进行特征筛选,根结点的右孩子结点消耗的隐私预算为 $\varepsilon/4$,剩下的隐私预算将会被分配到它的子结点中;而通过特征筛选可以知道该结点使用的特征为非重要特征,此时如果继续根据此特征划分子结点,分类效果并不理想,而且隐私预算会因为分配到子结点中而变小,根据差分隐私定义可知,当隐私预算越小时,加入的噪声量越大,分类的准确率就会降低,因此将该结点变为叶子结点可以提高分类准确率。

另外,不难看出,本文提出的特征筛选属于预剪枝策略,即不需要生成完整的决策树。而如 CRFsDP 算法^[11]中使用的后剪枝策略则需要先生成一棵完整的决策树,然后自底向上的对决策树进行剪枝。因此在执行效率方面,本文提出的特征筛选具有相对较好的性能。

2 算法实现及分析

2.1 RFDP_OOB 算法描述

RFDP_OOB 算法的主要思想是先差分隐私

保护下生成一部分的随机森林,同时根据加入噪声的包外估计计算决策树权重以及特征权重,并根据特征权重划分出重要特征与非重要特征,接着根据特征筛选生成剩下的随机森林,最后根据决策树权重进行分类预测,具体步骤如下。

1) 将隐私预算平均分配到每棵决策树上,生成一部分随机森林。对决策树中的结点判断:若该结点不是叶子结点,则取一半当前隐私预算对决策结果加噪;若该结点是叶子结点,用当前隐私预算对计数值添加噪声,确定分类类别。在构建完成后计算决策树中的特征权重、差分隐私下的包外估计以及决策树权重。

2) 对于已生成的一部分随机森林,获得随机森林中的特征权重之后,对特征进行划分,找出非重要特征。

3) 根据非重要特征构建决策树:对非叶子结点,若该结点的最优划分特征在非重要特征集中,则该结点变为叶子结点,并用当前隐私预算对计数值添加噪声,确定分类类别。同时对非重要特征集进行更新。

4) 将所有决策树以及对应的权重组合成随机森林。对要预测的数据,用随机森林中每棵决策树对其进行分类预测:从根结点开始,根据当前结点的分类属性决定该条数据应该进入到哪一个子结点,直到达到叶子结点,然后获得叶子结点的标签。将其相应决策树的权重线性相加,取权重最大的那个分类结果作为整个随机森林算法的分类结果。

算法伪代码如下:

```
输入:训练数据集  $X$ , 决策特征集  $F$ , 隐私预算  $\varepsilon$ , 决策树的数量  $t$ , 每棵树的深度  $d$ , 决策树特征数量  $f$ 
输出:满足  $\varepsilon$  - 差分隐私的随机森林
①  $\varepsilon' = \varepsilon/t$ ;
② For  $i = 1$  to  $t'$ : // 先生成一部分随机森林,  $t' < t$ 
③ 从  $X$  中有放回地取出大小为  $|X|$  的数据集, 并切分为决策属性集  $X(i)$  和对应的分类属性集  $y(i)$ 
④ 从决策特征集  $F$  中选出大小为  $f$  的特征子集  $F(i)$ 
⑤  $tree = Build\_tree(X(i), y(i), F(i), d, \varepsilon')$ ; // 构建 CART 树, 记录局部特征权重
⑥  $B = tree.predict\_oob(X - X(i), y - y(i), F(i), \varepsilon')$ 
⑦  $W_i = 1/B$  // 计算决策树权重
⑧ End For
⑨ 计算全局特征权重
⑩ 划分重要特征  $\Gamma$  及非重要特征  $\Gamma'$ 
⑪ For  $i = t' + 1$  to  $t$ :
⑫ 同上取出决策属性集  $X(i)$ 、对应的分类属性集  $y(i)$ 、特征子集  $F(i)$ 
⑬  $tree = Build\_tree(X(i), y(i), F(i), d, \varepsilon', \Gamma')$ ; // 根据非重要特征构造 CART 树
⑭  $B = tree.predict\_oob(X - X(i), y - y(i), F(i), \varepsilon')$ 
⑮  $W_i = 1/B$ 
⑯ 更新重要特征  $\Gamma$ 、非重要特征  $\Gamma'$ 
⑰ End For
⑱ Return { tree }
```

2.2 算法分析

2.2.1 隐私性分析

RFDP_OOB 算法中有 t 棵决策树, 每棵决策树分配到的隐私预算 $\varepsilon' = \varepsilon/t$, 用来保护两部分的数据隐私.

1) 用来生成决策树的训练数据. 对决策树中同一层的不同结点, 消耗当前一半隐私预算 $\varepsilon_h = \varepsilon'/2^{h+1}$, 其中 h 为结点的深度, 所以这些结点是符合 ε_h -差分隐私要求的. 又因为这些结点的数据都是不相交的, 根据差分隐私的并行组合性, 将数据不相交的结点组合后仍然满足 ε_h -差分隐私.

而决策树中不同层结点的数据存在交叉, 因此根据差分隐私的序列组合性, 将其组合满足的差分隐私所要求的隐私预算为各层隐私预算之和, 即 $\sum \varepsilon_h = \varepsilon'$. 所以用来生成决策树的训练数据是符合 ε' -差分隐私的.

2) 用来计算决策树包外估计的包外数据. 用包外数据计算单棵决策树的包外估计时, 用分配的隐私预算 ε' 对分类结果正确的计数进行加噪, 所以也是满足 ε' -差分隐私的.

由此这两部分数据都满足 ε' -差分隐私, 又由包外估计定义可知, 包外数据为生成决策树的训练数据的补集, 两者之间不存在交叉, 所以根据差分隐私的并行组合性可知, 将这两部分数据组合的决策树符合 ε' -差分隐私.

最后, 由于每棵树所用的训练数据是随机选择的, 因此所用数据会有交叉, 根据序列组合性, 整个随机森林消耗的隐私预算为每棵决策树消耗隐私预算的叠加, 即为 $t * \varepsilon' = \varepsilon$, 所以 RFDP_OOB 算法满足 ε -差分隐私.

2.2.2 执行效率分析

算法先生成一部分随机森林, 即 t' 棵决策树. 对每棵决策树, 深度为 d , 数据量大小为 X , f 是特征的数量. 决策树生成时, 把所有特征值都作为分裂候选, 为其添加噪声并计算基尼指数. 令特征 j 值的种类为 $V(j)$, 则每层时间复杂度为 $O(X * \sum_j V(j))$, 其中 $V(j)$ 最小为常数, 最大为数据量大小 X . 所以 d 层的决策树生成的时间复杂度是 $O(d * X * \sum_j V(j))$.

决策树生成后计算包外估计, 因为包外数据期望值为原始数据的 $1/3$, 所以计算包外估计的时间

复杂度为 $O(X)$, 远小于决策树生成时间, 而添加噪声只需要 $O(1)$ 的时间复杂度, 所以不影响时间复杂度的规模.

综上可知, 这一部分的时间复杂度为 $O(t' * d * X * \sum_j V(j))$.

接着需要计算全局特征权重并划分特征, 而全局特征权重只与特征数量和决策树权重有关, 所以时间复杂度为 $O(t' * f)$; 得到全局特征权重之后进行特征划分, 时间复杂度为 $O(\min(t' * f, F))$, 其中 F 为原始数据中的特征数, 这里因为用到的特征数量 $t' * f$ 不会超过原始数据中的特征数量 F , 所以取较小的值. 所以这一部分的时间复杂度为两者累加, 故而取较大的值, 即为 $O(\max(t' * f, F))$.

最后根据特征筛选生成剩下随机森林, 即 $\Delta t = t - t'$ 棵决策树. 与之前决策树生成的不同点在于每个结点处分裂特征进行筛选, 如果是非重要特征则不继续分裂, 所以此时要对特征进行查找, 时间复杂度为 $O(\min(t' * f, F))$, 仍然远小于决策树生成时间, 并且进行了筛选的决策树达不到深度 d , 执行时间会减少. 至于包外估计和权重更新不再赘述, 消耗时间仍会远小于决策树生成时间. 所以特征筛选下的随机森林的时间复杂度为 $O(\Delta t * d * X * \sum_j V(j))$.

将以上 3 部分的时间复杂度累加可以得出, 此算法的时间复杂度为

$$O(t * d * X * \sum_j V(j)).$$

式中: t 为随机森林中决策树数量, d 为决策树深度, X 为数据量大小, f 是特征的数量, $V(j)$ 是特征 j 中值的种类.

可以看出, 算法的主要模块, 无论是包外估计的计算还是特征的筛选都不影响算法时间复杂度的规模, 与 DiffPRFs^[9] 算法具有相同的时间复杂度, 与 CRFsDP 算法^[11] 最好情况下即不进行剪枝的情况下的时间复杂度相同, 所以此算法具有较高的性能.

3 实验结果

3.1 实验环境

本文实验操作系统为 Windows 10, CPU 为 i5-3230M @ 2.6 GHz, 内存为 8 G, 使用 PyCharm 2018, Python 3.7. 实验使用 3 个不同的高维度数据集, 都是来自 UCI 数据库的真实数据集, 而且数据集中都有一些隐私信息, 因此选用这些数据集进行实验在

一定程度上体现了本研究的现实意义.第1个数据集是罗彻斯特理工学院在2017年发布的关于癫痫病发作识别的数据集,共有11 500条数据,每条数据有179个特征,判定类别为癫痫病是否发作,以下记为ES.第2个数据集是伊斯坦布尔大学医学院神经病学系2018年发布的,记录了帕金森患者的基本信息及语音记录,共有756条数据,经过预处理删除id属性后共有754个特征,可以根据特征判定该人员是否患有帕金森,以下记为PDC.第3个数据集是互联网广告数据集,删除不完整数据后共有2 359条,每条数据有1 558个特征,可以根据数据判断它是否为广告,以下记为ADS.

本实验采用 F_1 作为评价随机森林性能的指标,它能综合衡量准确率 P 和召回率 R ,其取值范围为 $[0,1]$,定义如下:

$$F_1 = \frac{2 * P * R}{P + R}. \tag{9}$$

实验中使用该指标衡量融合分类器在测试集上的准确度. F_1 值越高,表明准确率和召回率越高,分类器正确分类能力越强.

RFDP_OOB算法的目的在于高维数据下保证数据隐私性的同时提高随机森林算法的分类准确率,因此实验中主要对随机森林算法的分类准确率进行对比验证.所以首先考虑RFDP_OOB算法中独有的参数预生成决策树数量 t' 对算法分类准确率的影响,找出最优值.然后根据文献[11]中的设定,分别在不同决策树深度和隐私预算下,将RFDP_OOB算法与其他差分隐私随机森林算法进行对比,观察算法的分类准确率,其中包括穆海蓉等^[9]提出的DiffPRFs算法、Xiang等^[11]提出的CRFsDP算法.为减少随机性带来的影响,对每组实验进行100次,取 F_1 值的平均值作为最终结果.另外通过比较RFDP_OOB算法与DiffPRFs算法^[9]、CRFsDP算法^[11]的执行时间,来验证RFDP_OOB算法的执行效率.同样进行100次实验,比较平均执行时间.

3.2 实验结果

3.2.1 参数与算法分类性能关系

首先观察RFDP_OOB算法中预生成决策树数量 t' 对分类准确率的影响.参照文献[11]的设定,令隐私预算 $\epsilon = 1$,决策树深度 $d = 5$,决策树数量 $t = 100$,决策树的特征数量 $f = \sqrt{F}$, F 为数据集的特征数.然后预生成的决策树数量 t' 与总数量 t 的比

值为 $1/6、1/3、1/2、2/3、1$,在3个数据集上的结果见图2.

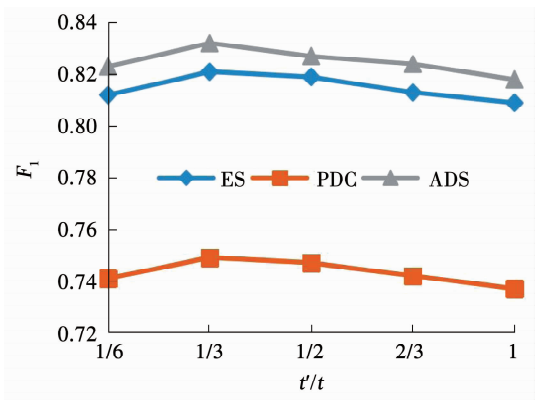


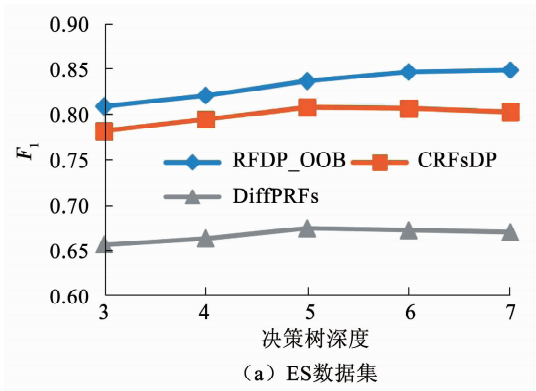
图2 3个数据集下的 t' 对分类准确率的影响

Fig. 2 Influence of t' on classification accuracy under three datasets

从图2中可看出,随着预生成决策树数量 t' 的增加, F_1 值先上升后下降,在 $t'/t = 1/3$ 的时候 F_1 值最大,而当 $t'/t = 1$ 时 F_1 值最小.这是因为,当预生成决策树数量过小时,使用的特征数量较小,不利于划分特征,从而不能充分利用特征筛选对决策树进行改进,进而降低了RFDP_OOB算法的分类准确率.而当预生成决策树数量过大时,则可以利用特征筛选进行改进的决策树数量过小,尤其是 $t'/t = 1$ 时,没有可以根据特征筛选构建的决策树,因此无法利用特征筛选达到改进的效果.综上所述,为了使RFDP_OOB算法达到最优,令预生成决策树 t' 为随机森林中决策树数量的 $1/3$.

3.2.2 算法分类性能对比

根据文献[11]中的设定,令决策树数量 $t = 100$,隐私预算 $\epsilon = 1$,决策树深度 d 为 $3 \sim 7$ 之间时,将提出的算法RFDP_OOB与CRFsDP算法^[11]、DiffPRFs算法^[9]在3个数据集上进行对比,实验结果见图3.



(a) ES数据集

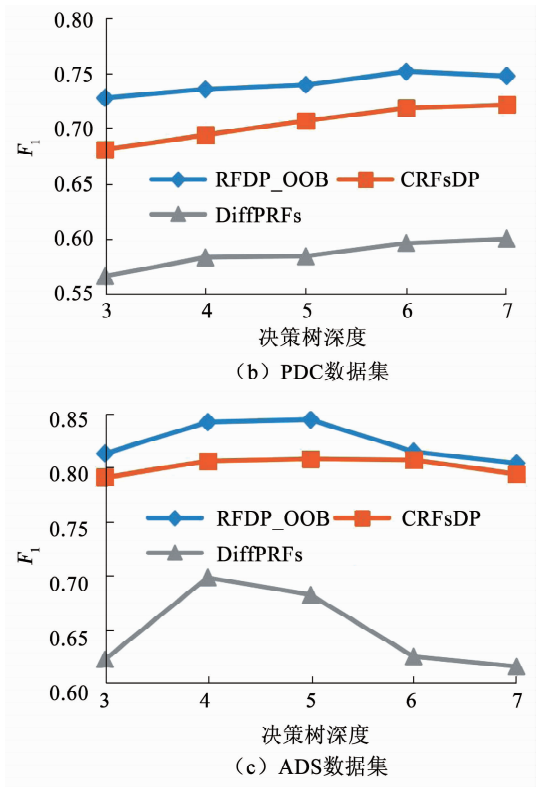


图 3 不同深度下的分类性能

Fig. 3 Classification performance under different depths

从图 3 中可以看出, RFDP_OOB 算法在 ES 和 ADS 数据集上的 F_1 值最高达到 0.85, 在 PDC 数据集上最高为 0.75, 并且在相同条件下比 CRFsDP 算法^[11]提高了 5%, 比 DiffPRFs 算法^[9]提高了 20%。这是因为 RFDP_OOB 算法和 CRFsDP 算法^[11]都进行了剪枝操作以及集成提升, 所以在高维数据下仍具有良好的性能。而且普遍来看, RFDP_OOB 算法的分类效果要高于 CRFsDP 算法^[11], 这也验证了 RFDP_OOB 算法的有效性。

另外, 3 种算法具有相似的趋势, 随着决策树深度的增加, 这 3 个算法的分类准确率先增加后不变甚至下降, 这也是隐私预算导致的。由前文分析可知, 决策树的高度越高时虽然有利于对数据进一步分类, 但是也导致最底层叶子结点的隐私预算越小, 从而使得加入噪声总量更大, 分类结果发生改变。

接下来仍然参照文献[11]中的设定, 比较不同隐私预算下算法的分类准确率, 令决策树数量 $t = 100$, 决策树深度 $d = 5$, 隐私预算 $\varepsilon = 0.1, 0.25, 0.5, 1$, 分别在 3 个数据集上进行实验。实验结果如图 4 所示。

从图 4 中可看出, RFDP_OOB 算法的 F_1 值在 ES 数据集上最高为 0.836, 在 PDC 数据集上最高为

0.74, 在 ADS 数据集上最高为 0.844。并且在相同隐私预算下 RFDP_OOB 算法比 CRFsDP 算法^[11]的 F_1 值要高 3% ~ 6%, 比 DiffPRFs 算法^[9]的 F_1 值高 20%。而且 3 种算法的 F_1 值都随着隐私预算的增加而增大, 这是因为隐私预算越大则添加的噪声量越小, 从而提高决策树的分类准确率, 进而提高随机森林算法的分类准确率。

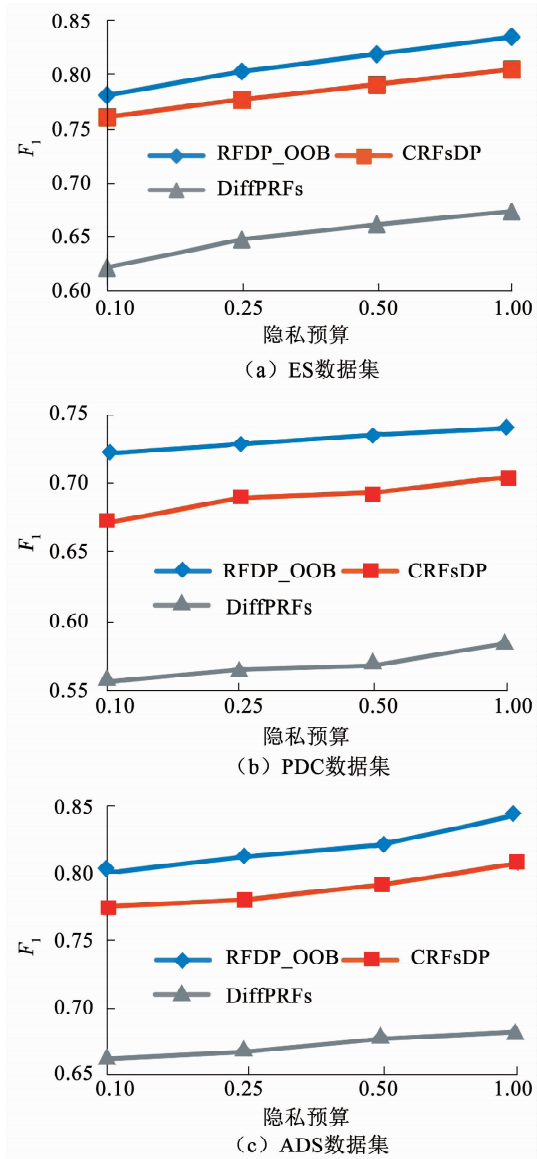


图 4 不同隐私预算下的分类性能

Fig. 4 Classification performance under different privacy budgets

综上所述, 无论是不同决策树深度还是不同隐私预算, 在其他参数相同的情况下, RFDP_OOB 算法的分类准确率都是要高于 CRFsDP 算法^[11]和 DiffPRFs 算法^[9]的, 由此可见, RFDP_OOB 算法在相同的隐私保护程度下, 能够有效提高分类的准确率, 也再次验证了本文改进思路的有效性。

3.2.3 算法执行效率对比

最后验证算法的执行效率,仍然参照文献[11]中的设定,令决策树数量 $t=100$ 、隐私预算 $\epsilon=1$,决策树深度 d 为3~7之间,测试3种算法的执行效率,实验结果见表2~4.

表2 ES数据集下的执行效率

Tab. 2 Execution efficiency under dataset ES min			
深度	RFDP_OOB	CRFsDP ^[11]	DiffPRFs ^[9]
3	7.3	24.8	6.9
4	8.6	26.6	8.3
5	10.3	28.9	9.9
6	12.9	31.8	12.4
7	15.6	35.9	14.5

表3 PDC数据集下的执行效率

Tab. 3 Execution efficiency under dataset PDC s			
深度	RFDP_OOB	CRFsDP ^[11]	DiffPRFs ^[9]
3	62	157	58
4	81	206	82
5	103	255	101
6	125	308	129
7	147	361	144

表4 ADS数据集下的执行效率

Tab. 4 Execution efficiency under dataset ADS s			
深度	RFDP_OOB	CRFsDP ^[11]	DiffPRFs ^[9]
3	5.04	7.67	4.61
4	5.98	7.96	5.42
5	6.84	8.80	6.35
6	7.85	9.40	7.49
7	9.06	9.69	8.85

综合表2~4可以看出,RFDP_OOB算法的运行时间差别较大,在ADS数据集上只需要7 s,而在ES数据集上却需要10 min.这是因为,根据前面执行效率的分析可知,执行效率与数据量大小、特征数量以及特征值种类有关,而ES数据集虽然特征数量较少,但是数据量以及特征值种类都远远大于另外两个数据集,因此时间复杂度较高.而ADS数据集中很多特征只有0和1两种值,所以执行时间反而是最短的.

另外容易看出,RFDP_OOB算法和DiffPRFs算法^[9]的执行时间要低于CRFsDP算法^[11],这是因为CRFsDP算法^[11]中需要先生成满二叉树再使用后剪枝方法来改进决策树,所以时间复杂度较高.RFDP_OOB算法与DiffPRFs算法^[9]具有相近的执行效率,这之前所分析的它们具有相同规模的时间复杂度相符合,进一步验证了RFDP_OOB算法的可行性.

4 结 论

本文引入了差分隐私下的包外估计来计算特征权重以及决策树权重,从而提出了一种基于差分隐私下包外估计的随机森林算法RFDP_OOB,并给出了算法的具体思想与描述.然后针对RFDP_OOB算法的隐私性进行了理论分析,再通过对比实验验证了算法的可行性及有效性.这种算法能够进行有效的特征筛选、集成提升,从而提高差分隐私随机森林算法的分类准确率,但在特定数据集上执行时间过长,而特征筛选模块又无法应用到并行方法下.因此,如何在并行模式下进行特征筛选,从而提高算法的执行效率,是日后需要研究的重点.

参考文献

[1] CAI Bo, WANG Xuanxuan, LI Peng, et al. A summary of data analysis based on differential privacy[C]// IEEE 4th International Conference on Big Data Security on Cloud. Omaha: IEEE, 2018: 79. DOI: 10.1109/BDS/HPSC/IDS18.2018.00029

[2] ZAINAB S S E, KECHADI T. Sensitive and private data analysis: a systematic review [C]// 3rd International Conference on Future Networks and Distributed Systems. Paris: ACM, 2019: 1. DOI:10.1145/3341325.3342002

[3] KAIROUZ P, OH S, VISWANATH P. The composition theorem for differential privacy[J]. IEEE Transactions on Information Theory, 2017, 63(6): 4037. DOI: 10.1109/TIT.2017.2685505

[4] DWORK C. Differential privacy [C]// Automata, Languages and Programming. Berlin: Springer, 2006: 1. DOI: 10.1007/11787006_1

[5] KRISHNAN S, WANG Jiannan, FRANKLIN M J, et al. PrivateClean: data cleaning and differential privacy [C]// 2016 International Conference on Management of Data. San Francisco: ACM, 2016: 937. DOI: 10.1145/2882903.2915248

[6] RANA S, GUPTA S K, VENKATESH S. Differentially private random forest with high utility [C]// 2015 IEEE International Conference on Data Mining. Atlantic City: IEEE, 2015: 955. DOI: 10.1109/ICDM.2015.76

[7] JAGANNATHAN G, PILLAI PAKKAMNATT K, WRIGHT R N. A practical differentially private random decision tree classifier [C]// 2009 IEEE International Conference on Data Mining Workshops. Miami: IEEE, 2009: 114. DOI: 10.1109/ICDMW.2009.93

[8]PATIL A, SINGH S. Differential private random forest[C]// 2014 International Conference on Advances in Computing, Communications and Informatics. New Delhi: IEEE, 2014: 2623. DOI: 10.1109/ICACCI.2014.6968348

[9]穆海蓉, 丁丽萍, 宋宇宁, 等. DiffPRFs: 一种面向随机森林的差分隐私保护算法[J]. 通信学报, 2016, 37(9): 175
MU Hairong, DING Liping, SONG Yuning, et al. DiffPRFs: random forest under differential privacy [J]. Journal on Communications, 2016, 37(9): 175. DOI: 10.11959/j. issn. 1000-436x. 2016169

[10]XIN Bangzhou, YANG Wei, WANG Shaowei, et al. Differentially private greedy decision forest [C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 2672. DOI: 10.1109/ICASSP.2019.8682219

[11]XIANG Tao, LI Yang, LI Xiaoguo, et al. Collaborative ensemble learning under differential privacy[J]. Web Intelligence, 2018, 16(1): 73. DOI: 10.3233/WEB-180374

[12]张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014, 37(4): 927
ZHANG Xiaojian, MENG Xiaofeng. Differential privacy in data publication and analysis[J]. Chinese Journal of Computers, 2014, 37(4): 927. DOI: 10.3724/SP.J.1016.2014.00927

[13]PAUL A, MUKHERJEE D P, DAS P, et al. Improved random forest for classification [J]. IEEE Transactions on Image Processing, 2018, 27(8): 4012. DOI: 10.1109/TIP.2018.2834830

[14]GWETUM V, VIRIRI S, TAPAMO J R. Purity and out of bag confidence metrics for random forest weighting[C]// International Conference on Computational Collective Intelligence. Cham: Springer, 2018: 491. DOI: 10.1007/978-3-319-98443-8_45

[15]DWORK C, ROTHBLUM G N. Concentrated differential privacy [Z]. arXiv:1605.02065

[16]TRUEX S, LIU Ling, GURSOY M E, et al. Privacy-preserving inductive learning with decision trees [C]// 2017 IEEE International Congress on Big Data. Honolulu: IEEE, 2017: 57. DOI: 10.1109/BigDataCongress.2017.17

(编辑 苗秀芝)

封面图片说明

封面图片来自本期论文“高离化率电-磁场协同增强 HiPIMS 高速沉积特性”,是采用电-磁场协同增强 HiPIMS 新技术制备的钼膜 AFM 三维立体表面形貌图. 高功率脉冲磁控溅射技术(HiPIMS)是一门新兴的高度电离化物理气相沉积技术,具有峰值功率密度高、脉冲频率和占空比低的放电特征,可获得较高溅射金属离化率和放电系统等离子体密度,有利于制备高硬度、高耐摩擦磨损性、高耐腐蚀性及抗高温氧化性优异的高质量薄膜/涂层. 然而,HiPIMS 存在的沉积速率较低的问题限制了其发展. 基于此所提出的电-磁场协同增强 HiPIMS 技术采用了电场和磁场双外场协同增强 HiPIMS 放电与沉积过程,该新技术在保持高离化率的条件下实现了 HiPIMS 高速沉积过程.

(图文提供:李春伟^{1,2},田修波²,姜雪松¹,徐淑艳¹(1. 东北林业大学工程技术学院;2. 先进焊接与连接国家重点实验室(哈尔滨工业大学))