面向隐私安全的联邦决策树算法

郭艳卿[1)] 王鑫磊[1)] 付海燕[1)] 刘航[1)] 姚明[2)]

[1)](大连理工大学信息与通信工程学院 辽宁 大连 116024)

[2)](深圳市洞见智慧科技有限公司数据智能部 北京 100007)

摘 要 根据用户信息进行资质审查是金融领域的一项重要业务，银行等机构由于用户数据不足和隐私安全等原因，无法训练高性能的违约风险评估模型，从而无法对用户进行精准预测. 因此，为了解决数据不共享情况下的联合建模问题，本文提出一种基于联邦学习的决策树算法 FL-DT（Federated learning-Decision Tree）. 首先，构造基于直方图的数据存储结构用于通信传输，通过减少通信次数，有效提升训练效率；其次，提出基于不经意传输的混淆布隆过滤器进行隐私集合求交，得到包含各参与方数据信息的联邦直方图，并建立联邦决策树模型. 最后，提出多方协作预测算法，提升了 FL-DT 的预测效率. 在四个常用的金融数据集上，评估了 FL-DT 算法的精确性和有效性. 实验结果表明，FL-DT 算法的准确率比仅利用本地数据建立模型的准确率高，逼近于数据集中情况下模型的准确率，而且优于其他联邦学习方法. 另外，FL-DT 的训练效率也优于已有算法.

关键词 联邦学习；决策树；混淆布隆过滤器；隐私安全；数据不共享

中图法分类号 TP181 DOI 号

Federated Decision Tree Algorithm for Privacy Security

GUO Yan-Qing[1)] WANG Xin-Lei[1)] FU Hai-Yan[1)] LIU Hang[1)] YAO Ming[2)]

[1)](School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning 116024)

[2)](Data Intelligence Department of InsightOne Tech Co, Ltd, Beijing 100007)

Abstract In recent years, with the vigorous development of technology and its related industries, Internet finance has increasingly highlighted its advantages. For a long time, qualification review based on the user information has been a fairly important business in the financial field. In most cases, when an individual applies for a loan from a bank, the bank will evaluate him or her through the actual situation based on the established predictive model to determine whether to grant the loan. In this process, a high-quality default risk assessment can avoid unnecessary losses for the banks. However, there are still many deficiencies in the current research on the assessment of default risks of borrowers by banks and other lending institutions. On the one hand, it is difficult to build a high-quality prediction models due to the lack of user data; on the other hand, people are paying more and more attention on the privacy protection of personal data, it is also a tough work for banks to obtain a large amount of relative data, and because of that, they cannot carry out the prediction models to accurately predict users' situation. In order to solve the problem of joint modeling in the case of data is not

收稿日期：2020-12-01；最终修改稿收到日期：年-月-日. 本课题得到国家自然科学基金(No.62076052, No.U1736119)、中央高校基本科研业务费(No.DUT20TD110, No.DUT20RC(3)088)资助. 郭艳卿，男，博士，教授，CCF会员（NO.18163M),主要研究领域为机器学习，网络空间安全. E-mail: guoyq@dlut.edu.cn. 王鑫磊，男，硕士研究生，主要研究领域为机器学习，计算机视觉. E-mail: wangxinlei@mail.dlut.edu.cn. 付海燕，女，博士，高级工程师，CCF会员（NO.18162M），主要研究领域为计算机视觉. E-mail: fuhy@dlut.edu.cn. 刘航，男，博士，副教授，主要研究领域为医学信号处理. E-mail: liuhang@dlut.edu.cn. 姚明，男，硕士，深圳市洞见智慧科技有限公司创始人、董事长，主要研究领域为大数据、多方安全计算、联邦学习. E-mail: yaoming@insightone.cn.

shared, this paper introduces the idea of the federated learning to effectively utilize the value of other participants data without the leaving of local data to establish a shared predictive model. Because decision tree algorithms are widely used in financial risk controlling and fraud identification, this paper proposes a decision tree algorithm FL-DT (Federated Learning-Decision Tree) based on federated learning. Federated learning is the concept put forward by Google in 2016, which can complete joint modeling without data sharing. Specifically, the data of each owner will not leave the local place, and the global sharing model will be jointly established through the parameter exchange method under the encryption mechanism in the federal system (in the case of not violating data privacy protection regulations). Moreover, each participant only serves for the local targets. Firstly, a data storage structure based on histogram is presented for communication transmission, which can effectively improve training efficiency by reducing the number of communications. Secondly, the garbled Bloom filter based on an oblivious transfer is proposed to performed the privacy set intersection, and then we can obtain the federated histogram containing the data information of each participant, and establishes the federated decision tree model. Finally, a multi-party collaboration prediction algorithm is put forward to improve the prediction efficiency of FL-DT. Based on four commonly used data sets in the financial field, this article assesses the accuracy and effectiveness of the FL-DT algorithm. The experimental results show that the prediction accuracy of the FL-DT model is higher than that of the model established using only local data, which is close to the model built in the case of data concentration. In addition, the prediction accuracy of the FL-DT methods is better than other federated learning methods, and the training efficiency and prediction efficiency are also better than other algorithms.

Key words Federated learning; Decision tree; Garbled bloom filter; Privacy security; Data not sharing

1 引言

近年来，互联网金融发展迅速，大数据背景下对借款人进行准确的贷前风险评估是各大金融机构的关注重点，但是银行等放贷机构对借款人的违约风险评估[1,2,3,4]方面的研究仍存在很大不足. 一方面由于缺少用户数据很难构建高质量的预测模型；另一方面由于对个人数据的隐私保护，银行等很难获得大量的用户数据. 例如中国在 2017 年提出的《中华人民共和国网络安全法》中，对数据的收集和处理提出了严格的约束和控制要求；美国加利福尼亚州也于 2020 年 1 月正式生效了《加利福尼亚州消费者隐私法》（California Consumer Privacy Act，CCPA）[5].

个人向银行申请贷款时，银行根据建好的预测模型对贷款人进行评估判断是否给予贷款. 数据共享的情况下，银行根据自己的数据库构建预测模型，通过该模型对新用户进行预测. 然而由于数据孤岛的存在，单个银行往往没有如此详细的用户属性信息. 如图 1 所示，银行只有用户的某些信息（例如：账单和房产信息），而其它金融公司持有这个用户的其他信息（例如：年龄和收入）. 银行想利用金融公司的用户数据扩展自己的数据库，而金融公司也可以根据自己的贡献从模型中收益. 由于要遵守数据安全法规或者保持专有数据的竞争优势，双方都不想共享自己的数据给对方.

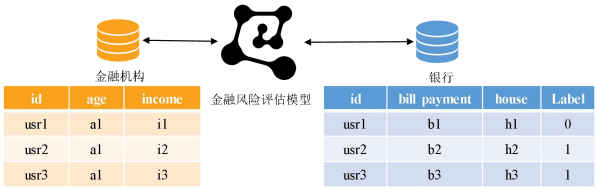


图 1 联邦场景下的联合建模

针对上述多方联合建模问题，文献[6]假设在训练过程中所有参与方可以以明文形式共享标签信息，但是现实情况下，标签只存在于某一个参与方中，并且由于隐私问题等不能透漏给其他参与方. 文献[7,8]中提出了一种在垂直分布的数据上隐私保护的决策树算法. 但是此方法存在两个问题：一是该算法只能处理属性值是离散的数据集，这对于现实情况不太实用；二是他们提出的方法必须揭示属性特征的类别分布，这将导致潜在的数据泄露风险.

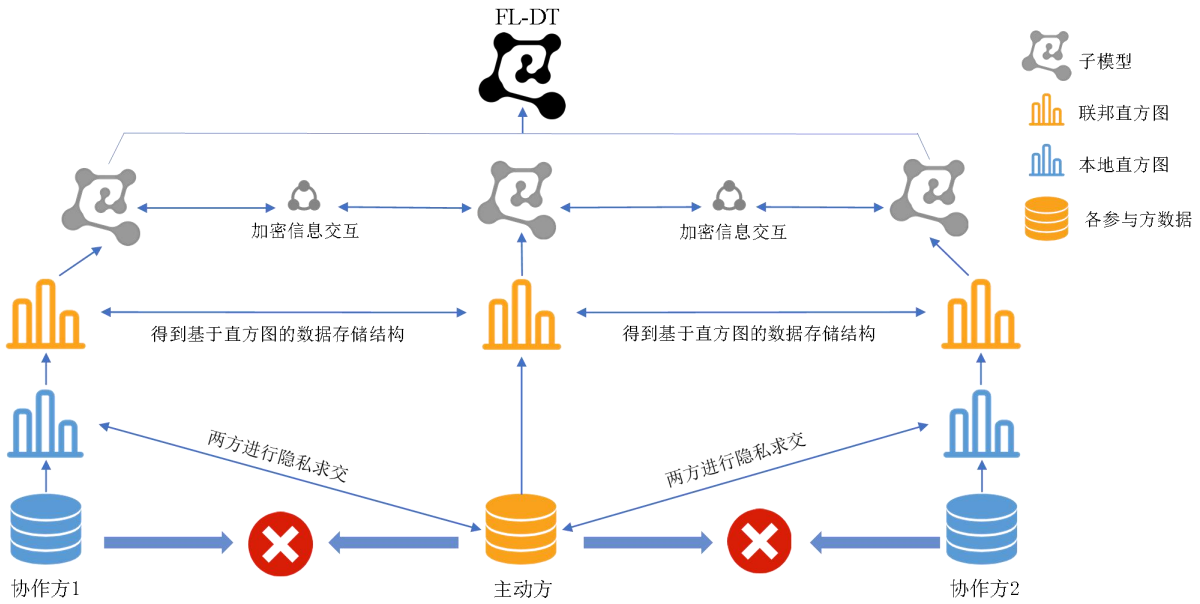


图 2 基于联邦学习的决策树方法整体框架

苹果公司提出使用差分隐私[9]来解决保护隐私问题. 差分隐私的基本思想是当第三方交换和分析数据时，通过向数据中添加经过适当校准的噪声来消除可能暴露用户身份的信息. 但是，差分隐私仅在一定程度上防止用户数据泄露，而不能完全消除个人信息. 此外，差分隐私仍要求不同组织之间进行数据交换，这也存在隐私泄露的风险.

仅仅依靠传统方法难以解决数据不共享情况下违约风险评估的联合建模问题，本文引入联邦学习思想，在数据不离开本地的情况下有效利用其它参与方的数据价值，建立共享的预测模型. 由于决策树算法被广泛地应用在金融风控和诈骗识别[5]等领域，因此本文提出一种基于联邦学习的决策树算法 FL-DT，整体流程框架如图 2 所示.

论文的主要贡献如下：1）提出一种基于联邦学习框架的决策树算法 FL-DT，解决数据不共享情况下的违约风险评估联合建模问题；2）提出一种基于直方图的数据存储结构进行通信传输，通过减少通信次数，有效提升训练时间. 并且以此直方图结构建立联邦决策树模型；3）采用秘密共享和基于不经意传输的混淆布隆过滤器两种多方安全计算技术加密训练联邦树模型，保证各参与方的信息安全.

本文章节安排如下：第 2 节介绍相关工作；第 3 节对如何建立联邦决策树模型进行详细介绍；第 4 节在四个常用金融数据集上进行实验，验证本文方法的有效性；第 5 节为结论及未来工作展望.

2 相关工作

2.1 联邦学习

联邦学习是由谷歌[10,11,12]在 2016 年提出的概念，该技术在数据不共享的情况下完成联合建模. 具体来讲，各个数据拥有者的自有数据不会离开本地，通过联邦系统中加密机制下的参数交换方式（不违反数据隐私保护法规的情况下）联合建立全局共享模型，建好的模型在各参与方只为本地的目标服务[13].

联邦学习框架包括两个模块：首先是数据对齐[14,15,16]，然后利用联邦协同的机器学习算法[17,18, 19,20]进行训练. 根据数据拥有者所持有数据的分布特点，联邦学习可分为横向联邦学习[21]、纵向联邦学习[22]和联邦迁移学习[23]. 1）横向联邦学习：当两个参与方的用户特征重叠较多而用户重叠较少时，取出双方用户特征相同而用户不完全相同的那部分数据进行训练； 2）纵向联邦学习：当两个参与方的用户重叠较多而用户特征重叠较少的情况下，取出双方用户相同而用户特征不完全相同的数据进行训练； 3）联邦迁移学习：当两个参与方的用户和用户特征都重叠较少时，不对数据进行划分，而是利用迁移学习来克服数据或标签不足的情况.

已有研究提出了一些基于联邦学习的分类方法. Pivot-DT[19]是一种联邦决策树分类算法，采用多方安全计算（MPC）和门限同态加密（TPHE）两种加密技术的混合联邦学习框架. 文献[24]中提出

的方法因为采用复杂的加密方法加密多个参与方的数据，导致训练过程非常耗时，并且带来了过高的计算开销. 文献[25]通过引进第三方平台进行密钥发放，但是并不能确定第三方是完全可靠的. RDTs[26]和 SecureBoost[17]中假设训练过程中的一些中间结果可以以明文形式发送，但是对于存在恶意攻击目的的参与方可以利用中间结果反推出其它参与方敏感信息，导致隐私数据泄露. Djatmiko 等人通过泰勒展开近似非线性逻辑损失，对加密后垂直划分的数据建立逻辑回归，除此之外，开源联邦学习框架 FATE[27]也实现了联邦逻辑回归算法 FL-LR（Federated Learning Logistic Regression），但是这种方式将导致精度损失严重.

通过以上分析，现有的联邦分类方法大都基于逻辑回归，Xgboost 等经典机器学习算法，并结合不同的加密技术保证联邦学习的安全性. 本文主要研究各参与方通过联邦协同的方式建立精准违约风险评估模型，其应用场景是纵向联邦学习，是在数据对齐的基础上，提出满足隐私保护和提升通信效率的联邦决策树算法. 与其他联邦学习相比，本文所选用机器学习算法不同和所使用的加密技术不同.

2.2 隐私集合交集

隐私集合交集（Private set intersection，PSI）[28,29,30]是指两方P_1和P_2各自拥有数据X和Y，在不暴露交集以外数据的情况下求得两方交集$X \cap Y$. 隐私求交计算是安全多方计算领域的一个重要方面，有着很广泛的应用场景. 例如隐私保护相似文档检测、安全的人类基因检测、隐私保护的社交网络关系发现等. 在数据由不同管理者持有的条件下，PSI 计算让用户既可以享受大数据时代的各种网络服务，也可以保护隐私数据的安全[31].

Freedman 等人[32]首先提出了隐私求交算法的概念和第一个基于不经意多项式估值协议的 PSI，随后又有研究人员提出了基于电路的 PSI[33,34]和基于公钥加密的 PSI[35,36]. 目前速度最快最流行的是基于不经意传输的 PSI.

3 基于联邦学习的决策树算法

3.1 问题描述

假设m个数据拥有方$(p_1, p_2, \cdots, p_m)$想要于n个样本$\{X_i, y_i\}_{i=1}^{n}$共同训练模型，所有属性特征$X_i \in \mathbb{R}^{1\times d}$分布在$m$个独立的参与方之间$\{X^k \in \mathbb{R}^{n\times d_k}\}_{k=1}^{m}$. 每个参与方拥有相同的用户集合，并且已经对齐，即每个用户在各个参与方具有相同的索引 id. 但是不同参与方持有同一个用户的不同属性信息. 任意两个参与方之间拥有的属性信息都不相同. 不失一般性，在构建机器学习模型时，约定只有一方提供类别标签，标签属性为$y \in \mathbb{R}^{n\times 1}$，由$p_m$方持有. 表 1 总结了本文所使用符号及其含义.

表 1 本文所使用符号及其代表含义

符号	代表含义
m	参与方的数量
n	样本总数
$\{p_i\}_{i=1}^{m}$	各个参与方
X^i	参与方p_i拥有的数据
X_j^i	参与方p_i的第j个特征
d_i, d	p_i的特征数量，总特征数量
y	类别标签
H^i	参与方p_i建立的直方图
b	直方图分桶数

约定 1. 主动方 由于标签属性对于监督学习必不可少，因此必须有一个主动方可以访问标签属性. 将同时拥有属性特征和类别标签的参与方称为主动方. 在联邦学习中，主动方担任主导服务器训练的责任.

约定 2. 协作方 将仅具有属性特征的参与方称为协作方，协作方在联邦学习中扮演客户的角色，比如其他金融机构. 协作方也需要通过模型来预测用户类别，因此，他们必须与主动方合作建立共同的模型，用来对未来的用户进行预测.

对于分布在$(p_1, p_2, \cdots, p_m)$等$m-1$个彼此独立的协作方数据$\{X^k\}_{k=1}^{m-1}$和分布在p_m上的带有标签的主动方数据$\{X^m, y\}$，本文通过联邦学习建立联邦决策树模型. 联邦学习框架要求满足模型无损[37,38]：

$$|P_r(DT|FL) - P_r(DT)| \le \varepsilon \quad (1)$$

其中$P_r(DT|FL)$为联邦决策树模型的精度，$P_r(DT)$为数据集中情况下决策树模型的精度，ε为任意正数. 在多方协作建立联邦决策树的过程中，为解决通信效率和隐私保护的问题，本文提出基于直方图形式的数据存储方式解决大规模稀疏属性特征的有效结构化问题，提出基于不经意传输的混淆布隆过滤器解决数据不共享情况下的隐私集合求交问题.

3.2 基于直方图的数据存储结构

联邦学习要求各方数据必须保存在本地，通过

交互模型参数进行模型聚合，当通信次数非常频繁或者传输参数过多时就会增加传输成本，降低训练速度，因此，解决通信成本问题是非常重要的优化环节[39,40]. 受 LightGBM[41]启发，本文提出联邦框架下的直方图算法，将各参与方数据转换成直方图形式的数据结构，并以此建立决策树.

主动方p_m由于拥有数据和标签$\{X^m, y\}$可以直接建立联邦直方图. 例如将p_m方的属性A划分到t个桶内，假设标签属性y具有c个类别$\{C_k \in \mathbb{R}^{|c_k|\times 1}\}_{k=1}^{c}$，其中$C_k$为第$k$类标签的索引集合,$|c_k|$为第$k$类标签的数量, 则建立的直方图为

$$H_A^m = \{|A_l| \in \mathbb{R}^{1\times c}\}_{l=1}^{t} \quad (2)$$

其中$|A_l|$为第l个桶内所包含c个类别的数量，即联邦直方图内存储的是属性A的所有桶，桶内包含各个标签类别的数量.

对于其他协作方$\{X^i\}_{i=1}^{m-1}$，由于没有标签信息，先建立本地直方图. 例如将X^i参与方的属性B划分到T个桶内，则建立的本地直方图为

$$H_B^i = \{B_l \in \mathbb{R}^{1\times c}\}_{l=1}^{T} \quad (3)$$

B_l表示第l个桶内所包含c个类别的索引集合. 即协作方直方图桶内存储的是该桶内样本的索引集合，这些索引值再通过隐私求交算法得到类别标签的数量.

直方图结构除了用于存储数据，还有减少通信次数和节省内存消耗的优势.

减少通信次数 在进行隐私求交过程中，如果不使用直方图数据存储结构，需要将协作方的每个样本索引逐一进行隐私求交，确定其所属的类别，此时通信的次数为样本数n.使用直方图存储结构可以将通信的次数由样本数n降低为直方图的桶数b，由于$\#b \ll \#n$，所以直方图算法极大地降低了通信的次数.

节省内存消耗 在建立决策树过程中，最耗时的步骤就是寻找最佳分裂点，目前最流行的是预排序算法[42]. 预排序算法需要 32 位浮点数（4Bytes）保存每个特征值，并且对每一列特征，都需要一个额外排好序的索引列，所以预排序消耗的总内存为：$2\times\#n\times\#d\times\#4Bytes$. 而直方图算法不需要存储原始的特征值，也不需要排序，仅需要存储离散后的特征分桶值，而分桶值只要 8 位整型（1Bytes）存储即可，直方图算法消耗的总内存为：$\#n\times\#d\times\#1Bytes$. 所以直方图算法消耗的内存仅为原始的 $1/8$.

3.3 基于不经意传输的混淆布隆过滤器

3.3.1 秘密共享

秘密共享(Secret Sharing)[43,44]是基本的密码原语. 秘密拥有者将秘密s分成n个子份额，使用t个或者更多份额的任何子集即可有效的恢复秘密. 对于少于t个份额的任何子集，该秘密是不可恢复的，并且每个子份额不提供有关该秘密的任何信息. 这种系统称为(t,n)-秘密共享.

当$t=n$时，通过简单的 XOR（异或）运算就可以得到一种有效且广泛使用的秘密共享方案[45]. 该方案进行秘密共享时：首先生成和秘密s相同长度的$n-1$个随机比特串$(r_1, r_2, \cdots, r_{n-1})$，由$(r_1, r_2, \cdots, r_{n-1})$和$s$计算得到

$$r_n = r_1 \oplus r_2 \oplus \cdots \oplus r_{n-1} \oplus s \quad (4)$$

每个r_i都是秘密的一部分. 秘密s通过计算$r_1 \oplus r_2 \oplus \cdots \oplus r_{n-1} \oplus r_n$恢复，且任何少于$n$个份额的子集不会显示秘密的信息.

3.3.2 混淆布隆过滤器

混淆布隆过滤器的作用包含两方面，一是把协作方直方图中每个桶内存储的索引集合（集合中的每个索引值称为元素），编码映射到m位的过滤器中；二是主动方对协作方进行查询，完成隐私求交，生成联邦直方图.

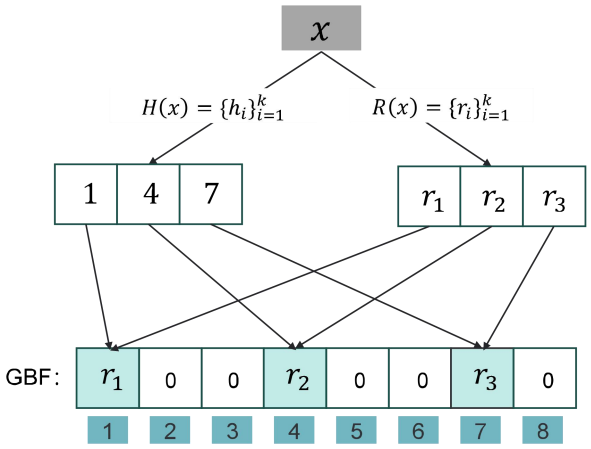


图 3 混淆布隆过滤器对元素的映射过程

对于一个(m,n,k,H,λ)参数化的混淆布隆过滤器GBF：m为混淆布隆过滤器的位数，n为索引集合S的元素个数，k表示哈希函数的个数，H为哈希函数集合，λ表示混淆布隆过滤器每个位置的比特数. 初始时，混淆布隆过滤器内存放的都是随机数. 图

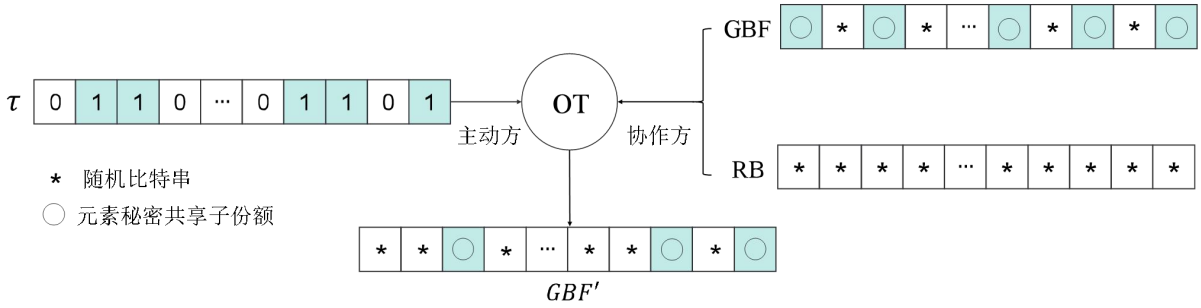


图 4 m位二选一不经意传输协议

3 所示是混淆布隆过滤器对协作方直方图桶内元素的映射过程，图中$k=3$，首先将直方图桶内任意元素x根据上述的秘密共享方案拆分成k个子份额$\{r_i\}_{i=1}^{k}$，并利用k个函数$H=\{h_1,\cdots,h_k\}$进行映射，得到其在混淆布隆过滤器中的k个位置

$$h_k(x)=mod(F_k(x),m) \tag{5}$$

其中$F_k(\cdot)$可以是任何哈希加密算法，如 HD5[46]，SHA256[47]等. 然后再将秘密共享子份额插入混淆布隆过滤器对应的索引位上

$$GBF[h_i(x)]=r_i \quad 1\le i\le k \tag{6}$$

当主动方查询元素y是否在集合s中，首先对y进行与x相同的哈希操作$\{h_i(y)\}_{i=1}^{k}$，得到y在混淆布隆过滤器中的k个位置. 然后将k个位置上的r_i取出来进行 XOR 运算，如果运算结果等于y值，则y在集合S中，否则不在集合S中，这一过程即为隐私求交.

联邦学习框架下，由主动方P_m主导，其他协作方$(p_1,p_2,\cdots,p_{m-1})$协同完成模型训练. 各个参与方在本地利用自己的数据建立直方图，主动方直接建立联邦直方图$\{H_i^m\in\mathbb{R}^{t\times c}\}_{i=1}^{d_m}$，而协作方建立桶内存储索引的直方图$\{\{H_i^j\in\mathbb{R}^{t\times c}\}_{i=1}^{d_j}\}_{j=1}^{m-1}$.主动方接受到各方建立的直方图后，与本地的标签属性$\{C_k\in\mathbb{R}^{|c_k|\times 1}\}_{k=1}^{c}$进行隐私求交，得到协作方每个桶对应主动方标签类别的数量，建立协作方的联邦直方图，最后整合所有直方图为$\{\{H_i^j\in\mathbb{R}^{t\times c}\}_{i=1}^{d_j}\}_{j=1}^{m}$.

从查询算法可知，如果协作方将生成的GBF直接发送给主动方，主动方可以通过暴力求解的方式求得GBF内存储的元素，所以存在数据泄露的风险. 为了防止隐私泄露，本文采用基于不经意传输的通信协议获得两方交集.

不经意传输(Oblivious transfer，OT)[48]协议是一种基于公钥密码体制的密码学协议，是安全多方计算的一种基础协议. 基础的 N 选 1 不经意传输协议可以描述为：两个参与方，其中发送方持有n条数据$s_1,\cdots,s_n$，而接受方持有选择向量τ，经过不经意传输，接受方能够获得s_τ，但是无法获得除s_τ以外的信息，发送方未输出其他信息且无法获得接受方的选择向量τ.

本文采用m位 2 选 1 的不经意传输解决上述可能暴露信息的问题. 如图 4 所示，其中GBF为协作方生成的混淆布隆过滤器，τ为主动方的选择向量，RB为随机比特串. 协议过程如下：主动方利用$H(y)$得到要查询的元素y在混淆布隆过滤器中的k个位置，将其设置为 1，其余位置设置为 0，即得到选择向量τ. 运行不经意传输：

$$GBF'[i]=\begin{cases}GBF[i], & \text{if } \tau[i]=1\\ RB[i], & \text{if } \tau[i]=0\end{cases} \tag{7}$$

从而主动方获得一个仅包含两方交集的混淆布隆过滤器GBF'，而协作方没有得到任何信息.

除了安全性，正确性也是显而易见的. 如果$y\notin S$，则经过相同的哈希函数，从GBF中取出k个λ比特进行异或运算的结果却等于y的概率为$2^{-\lambda}$（λ通常设置为 32），是可以忽略不计的. 当$y\in S$时，因为哈希函数是相同的，则在查询阶段每个$GBF[h_i(y)]$必定为y的秘密共享子份额，所以一定可以恢复y. 但是存在一种情况，由于GBF中存储的不是秘密共享子份额就是随机数，将GBF的特定位置存储为秘密共享子份额的概率为

$$p'=1-(1-\frac{1}{m})^{kn} \tag{8}$$

当集合S内某个元素哈希后的索引位全被其他元素的秘密共享子份额占据后，查询时就会出现本来在集合内，但是查询结果却不在集合内，这种假阴性的概率为

$$p=p'^{k}\times(1+O(\frac{k}{p'}\sqrt{\frac{\ln m-k\ln p'}{m}})) \tag{9}$$

3.4 联邦决策树的构造

各参与方通过隐私求交得到联邦直方图，以此直方图结构建立联邦决策树模型 FL-DT. 决策树算法是一种对高维数据进行有效分类的数据挖掘方法. 通过对输入的特征信息进行分析训练，构造决策树模型作为分类规则. 传统的 CART 树[49]是一种应用广泛的决策树学习算法，主要通过计算基尼指

数来选择最优特征，得到二叉树判断准则进行分类. 本文以 CART 树为基础，采用基尼指数选择最优特征建立联邦树模型.

在分类问题中，假设有K个类，样本点属于第k类的概率为p_k，则概率分布的基尼指数定义为

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2 \quad (10)$$

对于二分类问题，若样本点属于第一类的概率为p，则概率分布的基尼指数为

$$Gini(p)=2p(1-p) \quad (11)$$

对于给定的直方图集合H，其基尼指数为

$$Gini(p)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|H|}\right)^2 \quad (12)$$

其中$|C_k|$表示H中属于第k个类别的数量，K为类别的数量，$|H|$表示直方图内总样本数量. 根据直方图集合H计算出基尼指数最小的最优特征H_i，最佳分裂桶为$bins_j$. 此时直方图集合H根据最优特征H_i是否取值最佳分裂桶$bins_j$被分割为H_l和H_r两个子集，重新建立左边子集直方图

$$H_l=\{\{H_i(bins_j)\}_{j=1}^{t}\}_{i=1}^{d}\in H|H_i(bins)bins_j\} \quad (13)$$

右边子集直方图H_r可以根据直方图加速算法得到

$$H_r=H-H_l \quad (14)$$

那么在最优特征直方图H_i的条件下，直方图集合 H 分裂后的基尼指数为

$$Gini(H,H_i)=\frac{|H_l|}{|H|}\mathrm{Gini}(H_l)+\frac{|H_r|}{|H|}\mathrm{Gini}(H_r) \quad (15)$$

其中$Gini(H)$表示集合的不确定性，$Gini(H,H_i)$表示经$H_i(bins)=bins_j$分割后集合的不确定性，基尼指数越大，样本的不确定性越大，即越不适合作为最优分裂特征. 我们规定只有当分裂前的基尼指数$Gini(H)$与分裂后的基尼指数$Gini(H,H_i)$大于给定的阈值时，才继续分裂，否则判断当前节点为叶子节点，即

$$Gini(H)-Gini(H,H_i)>\varepsilon \quad (16)$$

ε为最小分裂阈值. 重复上述步骤，直到建立联邦决策树模型完成，整体算法流程如算法 1 所示.

算法 1. FL-DT 算法.

输入：分布在$p_1,\cdots,p_m$各参与方的数据$X^1,\cdots,X^m$，$prune\ conditions$ 剪枝条件

输出：FL-DT{联邦决策树模型}

$Hist_i=BuildHistogram()$ //各方参与建立本地直方图

FOR each $Hist_1,\cdots,Hist_{m-1}$ DO

  $FLHist_i=PSI(Hist_i)$ //使用混淆布隆过滤器隐私求交，建立联邦直方图

END FOR

$FLHist=|FLHist_1\cap\cdots\cap FLHist_m|$ //整合所有联邦直方图

IF $prune\ condition$ satisfied THEN

  return leaf node with majority class //满足剪枝条件，则返回叶子结点

ELSE

  $BestAtt,BestBins=FindBestSplit(FLHist)$ //找到最优特征和最佳分桶值

  IF $Gini(before)-Gini(after)>\varepsilon$ THEN

    Create Interior Node $p_i.A=BestAtt$ //创建中间节点

    IF $BestAtt$ at party p_i THEN

      At p_i: Split X into 2 partitions X_l,X_r

      $FLHist\_left=BuildHistogram(X_l)$

      $FLHist\_right=FLHist-FLHist\_left$ //直方图加速算法

    END IF

    FL-DT($FLHist\_left$) //递归

    FL-DT($FLHist\_right$) //递归

  END IF

END IF

3.5 多方协作预测算法

在模型预测过程中，各参与方共享同一个联邦树模型. 当前树节点的最优特征属于哪个参与方，就需要到哪个参与方执行树模型，所以各参与方之间需要频繁交互. 如果预测样本量过多，时间成本急剧上升，实际应用中根本无法投入使用. 在联邦学习框架中，通信所消耗的时间成本比计算消耗的时间成本大的多，因此本文提出多方协作预测算法，通过增加本地计算量，减少通信次数，从而提升预测效率.

如图 5 所示，左图为建立好的联邦树模型，右图为预测单个样本的流程. 从根节点所在方 Party 1 开始，对所有测试集样本逐一预测，如果当前节点还属于 Party 1，则继续下一节点，直到当前节点不属于 Party 1 为止；如果当前节点属于其他参与方，则记录当前节点的 Tree ID 和 Party ID，然后继续预测下一个样本. 在当前参与方遍历完所有样本后，Party 1 根据 Party ID 分别将预测样本转到对应的参与方继续预测，其他参与方直接从 Tree ID 节点继续预测即可. 重复此步骤，直到所有样本都得到预测结果. 利用这种方法可以将预测的时间复杂度

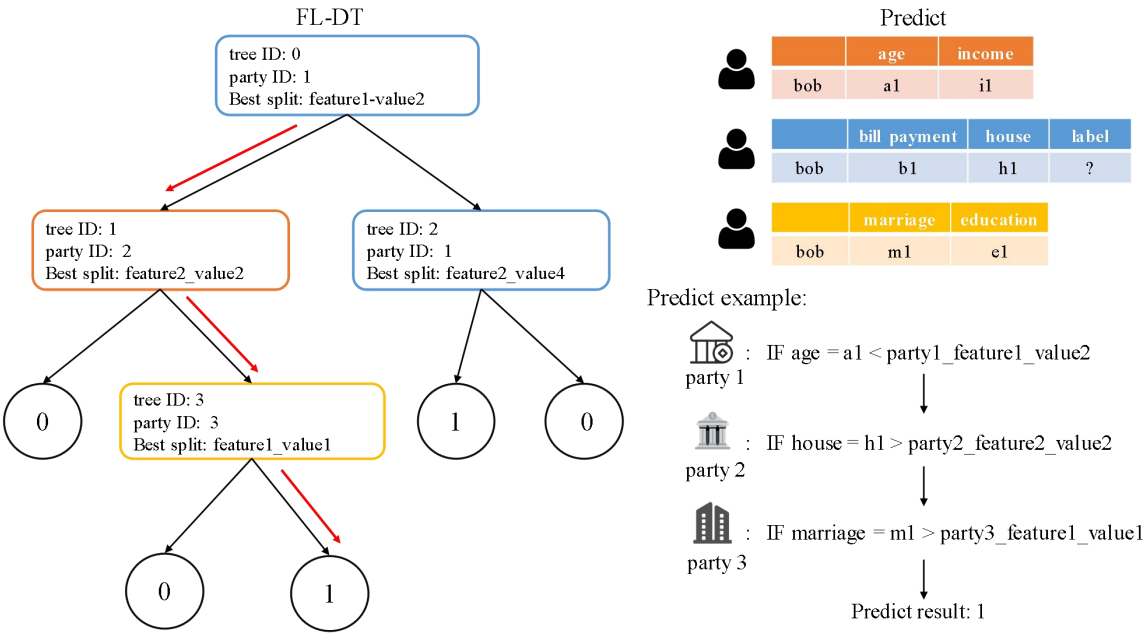


图 5 模型预测

从$O(\#n) \times O(\#h)$降低到$O(\#h)$，在大规模样本量预测上有很大优势.

4 实验及结果分析

4.1 数据集及评价指标

4.1.1 数据集介绍

本节通过四个常用的金融领域数据集对基于联邦学习的决策树模型进行性能评估：

Creditcard：是关于客户信用分数的数据集，与预测用户是否按时还款的分类任务相关. 包含30000 个用户样本、23 个属性信息和 1 个标签信息.

Bankmarket：与银行机构的电话营销活动有关，目的是预测客户是否会订阅定期存款. 包含4521 个用户样本，16 个属性信息和 1 个标签信息.

GMSC：是银行用来评估客户是否会遭受严重财务问题的分类数据集. 包含 150000 个用户样本、10 个属性和 1 个标签信息. 由于原始数据存在数据缺失，特征值异常等问题，所以我们对此数据集进行了特征工程处理，处理后包含 149165 个用户样本、10 个属性信息和 1 个标签信息.

CarEvaluation：该数据集是对汽车进行评估的分类数据集，包含 1728 个用户样本，6 个属性信息和 1 个标签信息.

实验过程中，将整体数据集的 70%用于训练，其余 30%用于测试. 并将数据集按照θ的比例将特征拆分成两部分，分别分配给参与方 Party A 和 Party B，其中 Party A 为主动方. 本文所有实验都是在两台华为云服务器上进行，配置如下：系统为 Centos 7.6.1810； CPU 为 Intel(R) Gold 6278C，2.6GHz，80GB 内存；GPU 为 Cirrus Logic GD 5446.

4.1.2 对比方法

本文采用以下对比方法验证 FL-DT 的性能：1）**PartA-DT**，**PartB-DT**：每个参与方仅通过本地数据训练模型，其中 PartA-DT、PartB-DT 分别为 Part A 参与方的本地数据和 Part B 参与方的本地数据训练的决策树模型，此时获得的决策树模型是非联邦协同情况下的模型，目的是验证联邦框架的有效性；2）**NoFL-DT**：所有参与方数据集中情况下训练决策树模型，目的是验证联邦后的模型精度无损；3）**其他联邦学习方法**：将 Pivot-DT[19]、FL-LR[27]、SecureBoost[17]作为其他联邦学习对比方法，用于评估 FL-DT 模型的精确性.

4.1.3 评价指标

本文主要从精确性和有效性两个方面评价联邦树模型 FL-DT：

精确性：采用模型在测试集上预测正确数量所占总数的百分比作为评价指标. 为了公平比较，SecureBoost 仅采用一棵树，所有对比方法都采用相同的实验参数，如树的最大深度，剪枝条件等.

有效性：采用训练模型所需时间和模型预测单

表 2 模型精度对比

数据集	NoFL-DT	PartA-DT	PartB-DT	Pivot-DT	FL-LR	SecureBoost	FL-DT
Credit card	82.7444	82.50	79.8556	82.1526	78.7778	82.7333	82.7356
Bank market	89.0824	88.7168	87.1681	88.6077	87.5368	88.5693	89.0118
GMSC	93.4635	93.4054	93.1931	-	93.3785	93.4301	93.4321
Car Evaluation	79.9228	72.5868	66.0231	-	67.9536	79.1505	79.1627

个样本所需时间作为指标，用于评估 FL-DT 的有效性.

4.2 模型的精确性实验

为了验证模型的精确性，我们将提出的联邦决策树 FL-DT，部分决策树 Part-DT 和数据共享决策树 NoFL-DT 在上述四个数据集上与 Pivot-DT、FL-LR 和 SecureBoost 进行比较，实验结果如表 2 所示. 主要参数设置如下：特征分割比例 $\theta = 70\%$，最大树深为 5，桶数为 256，最小信息增益为 1e-3，混淆布隆过滤器的比特数 λ 设置为 32. 我们对每个实验分别进行 10 次独立重复实验，并取平均结果.

从表 2 中分析可以得到以下结论：1）FL-DT 的精确度总是比 PartA-DT 和 PartB-DT 的精确度高，而且拥有总数据集 70%的主动方 PartA-DT 比 PartB-DT 精度也高，验证了数据特征维度越多，模型训练的效果越好，即数据不共享情况的联邦框架是有效的；2）联邦决策树 FL-DT 的精确度非常接近数据集中决策树 NoFL-DT，验证了联邦后的模型精度是无损的；3）与三种其他联邦学习对比方法相比，FL-DT 在四个数据集上的精度都得到了提升，验证了本文提出的基于直方图和混淆布隆过滤器的联邦决策树模型优于其他联邦学习方法. 本实验在 GMSC 上效果不明显的原因主要有两点：一是 GMSC 特征维度太少，而样本量却较大. 二是标签类别数量分布极度不均衡，两类标签数量比为 9:1.

为了进一步探索分桶数 b，最大树深h和数据分割比例θ对模型精确度的影响，我们在 Credit card 数据集上进行以下消融实验，实验结果如图 6 所示.

图 6（a）体现了分桶数对模型精度的影响，随着桶数的增加，模型精度也在不断上升并趋于数据集中情况下的最优精度. 原因在于，桶数较少时，量化误差较大，正则化效果较强. 当桶数增加到一定程度，量化间隔趋于特征值数量，每个桶内的样本数量更少，划分更精确，所以精度也达到最优. 但是桶数太大也会增加训练的时间成本.

图 6（b）是最大树深对模型精度的影响，随着树深的增加，模型对数据拟合效果越好，但是随着树深 h 的继续增加，就会出现过拟合，导致在测试集上测试精度下降.

图 6（c）表明了不同的数据分割比例对模型精度的影响. θ 代表的是主动方所持有的特征比例，随着 θ 的提升，精度变化不大. 主要原因是无论主动方的数据还是协作方的数据都建立直方图形式的数据存储结构，并且以直方图结构建立联邦树模型，所以只要特征总数一定，无论属予哪个参与方，建立的直方图结构都一样，最终模型也一样.

图 6（d-f）所示为选取最大树深为 5，最佳分桶数为 256，分割比例 $\theta = 70\%$ 等参数，FL-DT、FL-LR 和 SecureBoost 分别在三个数据集上得到的 ROC 曲线图. 实验结果标明，FL-DT 在三个数据集上的预测精度均高于其他联邦学习方法.

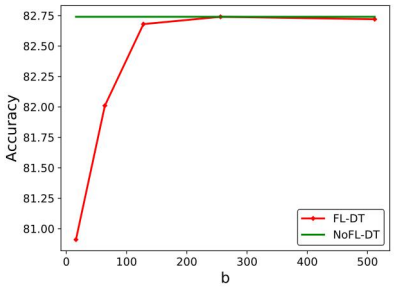
4.3 模型的有效性实验

联邦决策树模型的有效性主要受三个因素影响：1）属性分桶的数量 b；2）决策树的深度 h；3）数据集的样本量 n. 在本小节中，我们分别研究以上三个变量对训练时间的影响.

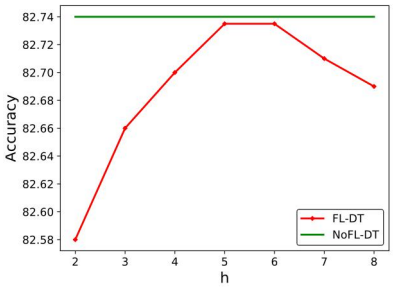
图 7（a）表明了分桶数 b 对训练时间的影响，从三个数据集整体来看，随着桶数的增加训练时间也相应的增加. 主要有两个原因：1）对于相同的样本量，分桶数越多，通信的次数也随之增加；2）分桶数越多，在计算基尼指数时，需要遍历的样本量也就越多，计算成本也就越高. 从图中还可以得到，对于相同的桶数，GMSC 建模所需时间总是比 Credit card 时间长，并且随着桶数的增加，差距也在逐渐增大.

图 7（b）体现了树的最大深度 h 对训练时间的影响，随着最大树深的增加，中间节点的数量也逐渐增多，训练消耗时间也自然增大. 其中在相同的树深下，GMSC 训练时间总是比 Credit card 多 20min 左右.

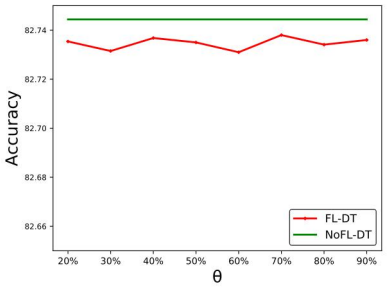
图 7（c）为样本数量 n 与训练时间的关系，训练样本越多，训练所需时间也就越长. 其中 Bank



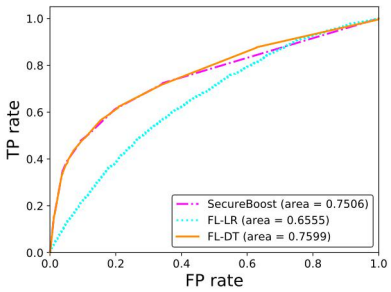
(a) 模型精度 vs. b



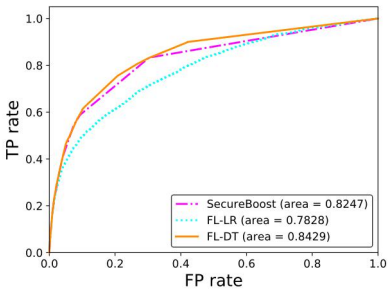
(b) 模型精度 vs. h



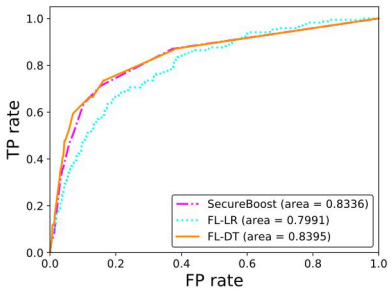
(c) 模型精度 vs. θ



(d) Credit card ROC 曲线



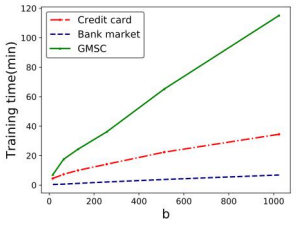
(e) Bank market ROC 曲线



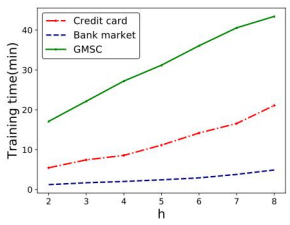
(f) GMSC ROC 曲线

图 6 （a）b对测试集精度的影响；（b）h对测试集精度的影响；（c）θ对测试集精度的影响
（d）Credit card ROC 曲线；（e）Bank market ROC 曲线；（f）GMSC ROC 曲线

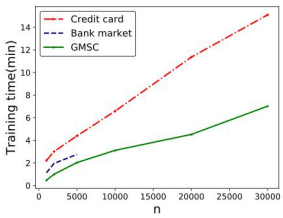
market 只有 4521 个样本量. 由于 Credit card 和 Bank market 分别比 GMSC 多了 13 个和 7 特征信息，所以在相同的样本量下，GMSC 所消耗时间最少.



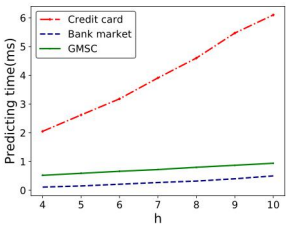
(a) 训练时间 vs. b



(b) 训练时间 vs. h



(c) 训练时间 vs. n



(d) 预测单个样本时间 vs. h

图 7 b, h, n对 FL-DT 效率的影响：（a）b对训练时间的影响；（b）h对训练时间的影响；（c）n对训练时间的影响；（d）h对预测时间的影响

图 7（d）为最大树深h对模型预测效率的影响. 在 Credit card 上，随着树深的增加，预测单个样本的时间上升趋势明显. 原因在于随着树深的增加，树模型的中间节点的数量也在增多，所以预测时间也随之增加. 而对于 GMSC 来说，树深最大的情况下单个样本预测时间也小于 1ms，可能的原因主要有两个：1）GMSC 本身特征属性较少，虽然样本量很大，但是分桶数是固定的，随着树深的增加，树模型中间节点的数量并没有增加很多；2）本文采用增加本地计算量，减少通信次数的预测方法，这种方法更适合预测大规模样本量的情况.

5 结论与展望

本文提出一种基于联邦学习框架的决策树算法 FL-DT 解决数据不共享情况下的联合建模问题. FL-DT 采用基于直方图形式的数据存储结构进行通信传输，通过减少通信次数，有效的提升训练效率. 并且以此直方图结构建立联邦决策树模型. 还提出基于不经意传输的混淆布隆过滤器进行隐私集合求交，不依赖任何可信的第三方，保证各参与方的信息安全. FL-DT 适用于半诚实模型，中间结果处于加密状态，能够保护参与方的信息不被其他存在恶意的参与方攻击. 本文在两个数据集上对 FL-DT 模型的精确性和有效性进行了评估. 实验

结果标明，FL-DT 模型比各方单独训练的模型精度都高，证明联邦学习的方法是有效的，并且与数据共享情况下训练的模型相比几乎是没有损失的. 除此之外，FL-DT 的训练效率和预测效率也优于其他算法.

本文提出的基于联邦学习的决策树算法还有不足之处. 由于一棵决策树的分类效果有限，所以在未来的工作中，考虑引入集成学习的思想，在建立的单棵树基础上实现随机森林（RF）、梯度提升树（GBDT）等方法.

致 谢 感谢国家自然科学基金(No.62076052,No.U1736119)、中央高校基本科研业务费(No.DUT20TD110,No.DUT20RC(3)088)的资助.

参 考 文 献

[1] Cheng Da-Wei, Niu Zhi-Bin, Zhang Li-Qing. Research on the risk of large-scale unbalanced secured online loans. Chinese Journal of Computers, 2020, 43(04): 668-682(in Chinese)
(程大伟,牛志彬,张丽清. 大规模不均衡担保网络贷款的风险研究. 计算机学报,2020,43(04):668-682)

[2] Xie Chen-Xin. Study on adaptability of borrower credit risk assessment model of P2P online lending platform. Wuhan Finance, 2019(3):23-29(in Chinese)
(谢陈昕.P2P网贷平台借款人信用风险评估模型适应性研究. 武汉金融，2019(3):23-29)

[3] Srinivasan V, Yong H K. Credit Granting: A Comparative Analysis of Classification Procedures. Journal of Finance, 2012, 42(3)

[4] Nalić J, Martinovic G. Building a Credit Scoring Model Based on Data Mining Approaches.2020,30(02):23

[5] California consumer privacy act. bill no. 375 privacy: personal information: businesses. https://leginfo.legislature.ca.gov. 2018:06-28

[6] Hu Y, Niu D, Yang J, and Zhou S. FDML: A collaborative machine learning framework for distributed features. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA, 2019,2232-2240

[7] Vaidya J, and Clifton C. Privacy-preserving decision trees over vertically partitioned data. Proceedings of the Data and Applications Security and Privacy. Berlin, Germany, 2005,139–152

[8] Vaidya J, Clifton, C, Kantarcioglu M, and Patterson A S. Privacy-preserving decision trees over vertically partitioned data. Proceedings of the ACM Transactions on Knowledge Discovery from Data. New York, USA, 2008,2(3):14

[9] Dwork C, Roth A, et al. The algorithmic foundations of differential privacy. Foundations and Trends R in Theoretical Computer Science. 2014,9(3–4):211–407

[10] Mcmahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging. arXiv preprint arXiv:1602.05629,2016

[11] Jakub K, H.Brendan M, Daniel R, et al. Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527,2016

[12] Jakub K, H Brendan M, et al. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016

[13] Mcmahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. Proceedings of the Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017,1273-1282

[14] Pinkas, Benny and Schneider, et al. Scalable Private Set Intersection Based on OT Extension. ACM Transactions on Privacy and Security. 2018,1243-1255

[15] Freedman M J, Nissim K, and Pinkas B. Efficient private matching and set intersection. Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques. Berlin, Germany,2004,1–19

[16] Pinkas B, Schneider T, Zohner M. Faster private set intersection based on OT extension. Proceedings of the 23rd USENIX Security Symposium. Berkeley, USA:2014:797-812

[17] Cheng K, Fan T, Jin Y, Liu Y, Chen T, and Yang Q. Secureboost: A lossless federated learning framework. arXiv preprint arXiv:1901.08755,2019

[18] Hardy S, Henecka W, Ivey-Law H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv: 1711.10677,2017

[19] Wu Y, Cai S, Xiao X, et al. Privacy preserving vertical federated learning for tree-based models. arXiv preprint arXiv:2008.06170.2020

[20] Hartmann V, Modi K, Pujol J, et al. Privacy-preserving classification with secret vector machines. arXiv preprint arXiv:2019.1907.03373

[21] Gao D, Ju C, Wei X, et al. HHHFL: hierarchical heterogeneous horizontal federated learning for electroence phalography. arXiv preprint arXiv:1909.05784,2019

[22] Liu Y, Kang Y, Zhang X, et al. A communication efficient vertical federated learning framework. arXiv preprint arXiv:1912.11187,2019

[23] Shreya S, Xing C, Yang L, et al. Secure and efficient federated transfer learning. arXiv preprint arXiv:1910.13271,2019

[24] Liu Y, Ma Z, Liu X, Ma S, Nepal S and Deng R. Boosting privately: Privacy-preserving federated extreme boosting for mobile crowdsensing. arXiv preprint arXiv:1907.10218,2019

[25] Ohrimenko O, Schuster F, Fournet C, Mehta A, Nowozin S, Vaswani K and Costa M. Oblivious multi-party machine learning on trusted processors. Proceedings of the USENIX Security Symposium. Austin, USA, 2016,619-636

[26] Vaidya J, Shafiq B, Fan W, Mehmood D, and Lorenzi D. A random decision tree framework for privacy-preserving data mining. IEEE Transactions on Dependable and Secure Computing,2014,11(5):399-411

[27] https://github.com/FederatedAI/FATE

[28] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao Xiao-Kui. A review of research on privacy protection for database applications. Chinese Journal of Computers, 2009, 32(05): 847-861(in Chinese)
(周水庚,李丰,陶宇飞,肖小奎. 面向数据库应用的隐私保护研究综述.计算机学报,2009,32(05):847-861)

[29] Gong Lin-Ming, Wang Dao-Shun et al. PSI calculation based on no matching errors. Chinese Journal of Computers, 2020, 43(09): 1769-1790(in Chinese)
(巩林明,王道顺等. 基于无匹配差错的PSI计算. 计算机学报,2020,43(09):1769-1790)

[30] Pinkas, Benny and Schneider, et al. Scalable Private Set Intersection Based on OT Extension. ACM Transactions on Privacy and Security. 2018,21(2),1243-1255

[31] Zhou Su-Fang, Li Shun-Dong et al. Efficient calculation of the intersection of confidential sets. Chinese Journal of Computers, 2018,41(02):464-480(in Chinese)
(周素芳,李顺东等. 保密集合相交问题的高效计算. 计算机学报,2018,41(02):464-480)

[32] Freedman M J, Nissim K, and Pinkas B. Efficient private matching and set intersection. Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques. Berlin, Germany,2004,1–19

[33] Huang Y, Evans D, Katz J. Private set intersection: Are garbled circuits better than custom protocols? Proceedings of the Network and Distributed System Security Symposium. Reston,USA,2012:1-15

[34] Dong C, Chen L, Wen Z. When private set intersection meets big data: An efficient and scalable protocol. Proceedings of the ACM SIGSAC Conference on Computer & Communications Security. New York，USA, 2013: 789-800

[35] Dana D, Tal M, Mariana R, et al. Efficient robust private set intersection. Proceedings of the International Conference on Applied Cryptography and Network Security. Berlin, Germany, 2009:125-142

[36] Hazay C, Nissim K. Efficient set operations in the presence of malicious adversaries. Proceedings of the International Conference Workshop on Public Key Cryptography. Berlin,Germany, 2010:312-331

[37] Li T, Sahu A K, Talwalkar A, et al. Federated learning: challenges, methods, and future directions. IEEE Signal Processing Magazine, 2020, 37(3): 50-60

[38] Kairouz P, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977.2019

[39] Wang Jian-Zong, Kong Ling-Wei, Huang Zhang-Cheng, et al. Summary of Federated Learning Algorithms. Big Data. 2020:1-22(in Chinese)(王健宗，孔令炜，黄章成等人．联邦学习算法综述．大数据. 2020:1-22)

[40] Kairouz P, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977.2019

[41] Ke G, Men Q and Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the Conference and Workshop on Neural Information Processing Systems. Long Beach, USA, 2017:3146-3154

[42] Li P, Burges C, Wu Q, et al. Learning to rank using multiple classification and gradient boosting. Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada 2007:845–852.

[43] Stadler, Markus. Publicly Verifiable Secret Sharing. Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques. Berlin, Germany, 1996:190-199

[44] Shamir Adi. How to Share a Secret. Communications of the ACM. 1979,22(11):612-613

[45] Bruce Schneier. Applied cryptography: protocols, algorithms, and source code in C. New York, USA, 2007

[46] Ronald R and Dusses. The MD5 message-digest algorithm. MIT Laboratory for Computer Scienc. Cambridge, USA. 1992: 330-344

[47] Donald Eastlake and Paul Jones. US secure hash algorithm 1 (SHA1). RFC 3174, Milford USA. 2001:0-22

[48] Naor M and Pinkas B. Efficient oblivious transfer protocols. Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, Pennsylvania, USA 2001:448–457

[49] Timofeev, Roman. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin. 2004,1-40

Guo Yan-Qing, Ph.D professor. His research interests include Machine learning, computer vision and Cyberspace security.

Wang Xin-Lei, M.S. His research interests include Machine learning and Computer vision.

Fu Hai-Yan, Senior engineer. Her research interests include Computer vision.

Liu Hang, associate professor. His research interests include Medical signal processing.

Yao Ming, Masters Degree. His research interests include Big data, multi-party secure computing and federated learning.

Background

Federated learning (FL) is an emerging paradigm for machine learning that enables multiple data owners to jointly train a model without revealing their private data to each other. The basic idea of FL is to iteratively let each client (i) perform some local computations on her data to derive certain intermediate results, and then (ii) exchange these results with other clients in a secure manner to advance the training process, until a model is obtained.

The challenges in federated learning at first glance resemble classical problems in areas such as privacy, large-scale machine learning, and distributed optimization. For instance, numerous methods have been proposed to tackle expensive communication in the machine learning, optimization, and signal processing communities. However, these methods are typically unable to fully handle the scale of federated networks, much less the challenges of systems and statistical heterogeneity. Similarly, while privacy is an important aspect for many machine learning applications, privacy-preserving methods for federated learning can be challenging to rigorously assert due to the statistical variation in the data, and may be even more difficult to implement due to systems constraints on each device and across the potentially massive network.

Existing work on FL has mainly focused on the horizontal setting, which assumes that each client's data have the same schema, but no tuple is shared by multiple clients. In practice, however, there is often a need for vertical federated learning, where all clients hold the same set of records, while each client only has a disjoint subset of features. in order to solve the problem of joint modeling in the case of data not sharing, this paper proposes a decision tree algorithm FL-DT based on Vertical federated learning. The method proposed in this paper improves training efficiency and protects the information security of all participants.

This paper is partially supported by National Natural Science Foundation of China (No.62076052, No. U1736119) and Fundamental scientific research business expenses of central universities (No. DUT20TD110, No. DUT20RC (3)088)