# 基于 BCP 的联合委托学习模型及协议

高胜[1,2]，向康[1,3]，田有亮[1,3]，谭伟杰[1]，冯涛[4]，吴晓雪[5]

（1. 贵州大学计算机科学与技术学院公共大数据国家重点实验室，贵州 贵阳 550025；
2. 中央财经大学信息学院，北京 100081；3. 贵州大学密码学与数据安全研究所，贵州 贵阳 550025；
4. 兰州理工大学计算机与通信学院，甘肃 兰州 730050；5. 贵州省计量测试院，贵州 贵阳 550000）

**摘 要**：为了实现数据安全共享的同时减少客户端在数据挖掘过程中的计算成本，基于 BCP 同态加密算法提出了联合委托学习模型及协议。首先，针对决策树模型的安全构造提出了基于虚假记录的隐私保护方法。其次，根据数据垂直分布与水平分布的情况，基于隐私保护委托点积算法和隐私保护委托求熵算法提出了相应的委托学习协议。最后，给出了委托学习协议及决策树模型结构的安全性证明和性能分析。结果表明，基于虚假记录的隐私保护方法不会影响最终模型的构建，并且各客户端最终获得的模型与真实数据构建的模型具有一致性。

**关键词**：委托学习；决策树；BCP 同态加密；数据挖掘

**中图分类号**：TP309

**文献标识码**：A

**DOI:** 10.11959/j.issn.1000−436x.2021089

# BCP-based joint delegation learning model and protocol

GAO Sheng[1,2], XIANG Kang[1,3], TIAN Youliang[1,3], TAN Weijie[1], FENG Tao[4], WU Xiaoxue[5]

1. State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2. School of Information, Central University of Finance and Economics, Beijing 100081, China
3. Institute of Cryptography and Data Security, Guizhou University, Guiyang 550025, China
4. School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
5. Institute for Metrology and Calibration of Guizhou, Guiyang 550000, China

**Abstract:** In order to realize data security sharing and reduce the computing costs of clients in data mining process, a joint delegation learning model and protocol based on BCP homomorphic encryption algorithm was proposed. Firstly, a privacy preserving method based on false records was proposed for the security of decision tree model. Secondly, in view of the vertical and horizontal distribution of data, the corresponding delegation learning protocols based on privacy preserving delegation dot product algorithm and privacy preserving delegation entropy algorithm was proposed. Finally, the security proof and the performance analysis of delegation learning protocols and decision tree model structure were given. The results show that the privacy protection method based on false records does not affect the final model construction, and the final model obtained by each client is the same as that constructed by real data.

**Keywords:** delegation learning, decision tree, BCP homomorphic encryption, data mining

---

**收稿日期**：2020−12−15；**修回日期**：2021−03−26

**通信作者**：田有亮，youliangtian@163.com

**基金项目**：国家自然科学基金资助项目（No.61662009, No.61772008, No.U1836205, No.62072487）；贵州省科技重大专项计划基金资助项目（No.20183001）；贵州省科技计划基金资助项目（黔科合基础[2019]1098, 黔科合平台人才[2019]5703）；贵州省高层次创新型人才基金资助项目（黔科合平台人才[2020]6008）；贵州大学引进人才科研基金资助项目（贵大人基合字[2020]61）；贵州大学培育基金资助项目（贵大培育[2019]56）

**Foundation Items:** The National Natural Science Foundation of China (No.61662009, No.61772008, No.U1836205, No.62072487), Science and Technology Major Support Program of Guizhou Province (No.20183001), Science and Technology Program of Guizhou Province (No.[2019]1098, No.[2019]5703), Project of High-level Innovative Talents of Guizhou Province (No.[2020]6008), Research Project of Guizhou University for Talent Introduction (No.[2020]61), Cultivation Project of Guizhou University (No.[2019]56)

## 1 引言

机器学习等相关技术的发展使大数据中的有利信息得以被挖掘和利用。在实际生活中，数据的分布并不是只存在于一个数据站点，而是多样化地分布于多个数据站点。因此，数据共享[1-2]已成为数据挖掘等相关领域的研究热点，而数据隐私泄露等安全问题是数据共享技术中的发展瓶颈。传统的基于安全多方计算（SMC, secure multi-party computation）[3-4]的解决方案效率低下且可行性较差，无法真正实现对大数据[5]的处理。

在实际的数据特征学习过程中，对数据进行较复杂的分析、模型构造以及优化通常是比较困难的，导致客户端背负着沉重的计算成本。甚至部分企业或用户由于受限于自身对数据处理的能力而无法挖掘出有用的信息，只能依托于云服务[6-7]提供商进行特征提取和模型训练。因此，基于传统的委托计算思想引出数据多点分布时的数据外包挖掘方法具有重要的实际应用价值。本文将这种数据外包挖掘的方式命名为联合委托学习，如图 1 所示。

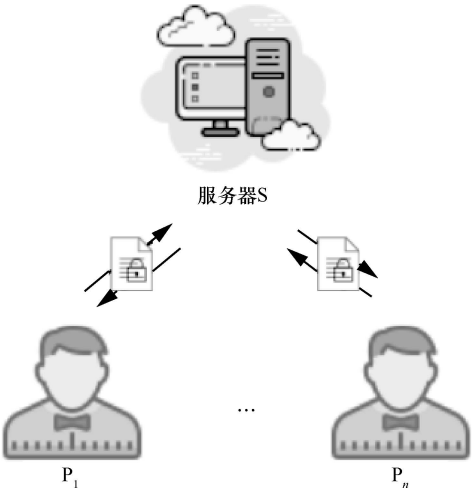



图 1 联合委托学习

在数据多点分布时进行联合委托学习需要考虑以下需求。

1) 在数据共享中，不仅要避免用户隐私数据的泄露，而且必须保证加密后的数据满足数据挖掘的条件。

2) 避免服务器从计算的中间结果推测出最终构建的模型信息。

3) 尽量将计算任务委托给云服务提供商，以降低客户端的计算成本。

### 1.1 本文的贡献

针对联合委托学习中的安全性需求，本文的主要贡献如下。

1) 基于传统的委托计算思想提出了一种联合委托学习模型，并针对决策树的构造设计了一种基于虚假记录的隐私保护方法（FRPPM, false-based records privacy protection method），该方法利用少量的虚假记录扰乱最终构建的模型结构，增强了数据和模型结构的安全性。

2) 基于 BCP（Bresson, Catalano, Pointcheval）同态加密算法分别设计了隐私保护委托点积算法（PPDDPA, privacy preserving delegation dot product algorithm）和隐私保护委托求熵算法（PPDEA, privacy preserving delegation entropy algorithm），降低了客户端的隐私数据在数据共享中的泄露风险。

3) 针对数据垂直和水平分布的情况，根据上述 2 种算法分别提出了对应的委托学习协议，降低了客户端在数据挖掘中的计算成本。

### 1.2 相关工作

隐私保护数据挖掘技术[8-9]可分为基于数据扰动的方法和基于安全多方计算的方法。基于数据扰动方面，Agrawal 等[10]提出了用加入随机噪声的方法来进行隐私保护决策树挖掘的方案。但此种加入随机噪声的方法过于简单，Kargupta 等[11-12]对加入随机噪声方法的安全性提出了质疑，并基于随机矩阵理论提出了从扰动后的数据估计真实数据的方法。Bu 等[13]给出了一种基于函数扰动的方法，该方法采用反函数变换方式来将扰动数据上的虚假决策树还原为真实数据上的决策树。

基于安全多方计算方面，Hamada 等[14]及 Bost 等[15]分别研究了可以应用于 SMC 中的决策树分类算法，在 SMC 中向各参与方隐藏了输入向量，但有关决策树的信息被假定对各方公开。其中，Bost 等[15]研究了使用同态加密通过决策树对信息进行安全分类的方法。随后，Wu 等[16]及 Backes 等[17]各自将其扩展为一个随机森林。此外，在与本文的工作相似的研究中，Ichikawa 等[18]提出了一种新颖的安全多方协议，同样隐藏了输入向量和输出类以及树的结构。Zheng 等[19]及 Li 等[20]分别设计了基于云服务器的分类模型，可以保护树模型和客户数据隐私。另外，Liu 等[21]在此基础上设计了支持离线服务的分类模型提高了系统的可伸缩性。特别地，Li 等[22]

针对数据的水平和垂直分布分别设计了外包隐私保护加权平均协议（OPPWAP, outsourced privacy preserving weighted average protocol）和外包安全集交叉协议（OSSIP, outsourced secure set intersection protocol），但不能保护训练模型结构的安全性。

## 2　预备知识

### 2.1　BCP 同态加密

BCP 同态加密算法是由 Bresson、Catalano 和 Pointcheval 于 2003 年提出的，属于同态密码体制，具有以下性质。

$$\mathrm{Enc}_{\mathrm{pk}}\left(m_1+m_2\right)=\mathrm{Enc}_{\mathrm{pk}}\left(m_1\right)\odot\mathrm{Enc}_{\mathrm{pk}}\left(m_2\right) \tag{1}$$

其中，$m_1$ 和 $m_2$ 表示明文信息，$\odot$ 表示在同一公钥下加密域中的算术乘法运算。BCP 同态加密算法的形式化描述包括以下 4 个部分。

1) Setup($\lambda$)。首先给定安全参数 $\lambda$ 表示模数 $N$ 的位长，再选定 2 个不同的素数 $p'$ 和 $q'$ 分别计算 $p=2p'+1$，$q=2q'+1$ 以及 $N=pq$；随后选择 $g\in Z_{N^2}^{*}$（$Z_{N^2}^{*}$ 表示在 $[1,N^2]$ 中所有与 $N^2$ 互质的数），并使 $g^{p'q'} \bmod N^2=1+kN$，$k\in[1,N-1]$；最后得到公共参数 $\mathrm{pp}=(N,k,g)$，主密钥 $\mathrm{MK}=(p',q')$。

2) KeyGen(pp)。随机选择 $t\in Z_{N^2}$，根据公共参数得到公钥 $\mathrm{pk}=g^t \bmod N^2$，私钥 $\mathrm{sk}=t$。

3) $\mathrm{Enc}_{\mathrm{pk}}(m)$。明文 $m\in Z_N$，选择随机数 $r\in Z_{N^2}$，利用公钥加密得到密文 $(A,B)$，其中 $A=g^r \bmod N^2$，$B=g^{t\,r}(1+mN) \bmod N^2$。

4) $\mathrm{Dec}_{\mathrm{sk}}(A,B)$。利用私钥 $\mathrm{sk}=t$ 解密，获得明文 $m=\dfrac{B/(A^t)-1 \bmod N^2}{N}$。

### 2.2　数据的分布形式

通常数据的分布类型包括 2 种情况：水平分布类型，即每个站点仅包含一部分元组，但每个元组都是完整的；垂直分布类型，即各个站点包含所有元组，但每个元组都不是完整的，仅包含一部分属性。如表 1 中前 14 条记录所示，为方便叙述，此处假设数据分布在站点 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 处。数据水平分布时，站点 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 分别包含部分完整的记录，同时各站点都知道数据所对应的属性名称，即表 1 中的第二行信息，但各个站点对其他站点所包含的具体数据一无所知。数据垂直分布时，站点 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 都包含所有记录，但每条记录都是不完整的，对于所有特征属性而言站点 $\mathrm{P}_1$ 只包含前 2 个属性的数据，站点 $\mathrm{P}_2$ 只包含后 2 个属性的数据，但它们都包含标签项数据，即表 1 中的最后一列信息，同样各个站点不愿意对其他站点透露自己所包含的具体数据。

**表 1**　数据集的分布形式

| 站点 | id | $\mathrm{P}_1$ | | $\mathrm{P}_2$ | | |
|---|---|---|---|---|---|---|
| | | outlook | temperature | humidity | windy | play |
| $\mathrm{P}_1$ | 1 | sunny | hot | high | weak | no |
| | 2 | sunny | hot | high | strong | no |
| | 3 | overcast | hot | high | weak | yes |
| | 4 | rainy | mild | high | weak | yes |
| | 5 | rainy | cool | normal | weak | yes |
| | 6 | rainy | cool | normal | strong | no |
| | 7 | overcast | cool | normal | strong | yes |
| $\mathrm{P}_2$ | 8 | sunny | mild | high | weak | no |
| | 9 | sunny | cool | normal | weak | yes |
| | 10 | rainy | mild | normal | weak | yes |
| | 11 | sunny | mild | normal | strong | yes |
| | 12 | overcast | mild | high | strong | yes |
| | 13 | overcast | hot | normal | weak | yes |
| | 14 | rainy | mild | high | strong | no |
| $\mathrm{P}_1$ | 15 | foggy | cold | low | calm | yes |
| $\mathrm{P}_2$ | 16 | foggy | cold | low | breezy | no |

## 3　联合委托学习模型

### 3.1　系统模型

联合委托学习的系统模型框架如图 2 所示，系统模型包含 2 个服务器和 $n$ 个客户端（用户），并且相互之间采用安全信道连接。其中 $\mathrm{S}_1$ 是主服务器，$\mathrm{S}_2$ 是副服务器，分别由不同的服务提供商提供。首先，由 $\mathrm{S}_2$ 生成公共参数并发送给各客户端，客户端根据公共参数计算出各自的公私钥对。其次，客户端根据 FRPPM 对各自的私有数据集进行扰动处理，并利用各自的公钥对扰动后的数据统计信息进行加密。然后，将 $A$、$B$ 这 2 类密文分别发送给 $\mathrm{S}_2$ 和 $\mathrm{S}_1$ 进行计算，并由 $\mathrm{S}_1$ 综合计算后返回结果。最后，各客户端根据返回的结果构建同样的决策树模型。需要注意的是，无论数据垂直分布还是水平分布，本文在模型的构建过程中都以 ID3[23]算法为例。为方便后续描述，假设总体数据集 $D$ 分布在 $n$ 个客户端处，即 $D=D_1\cup D_2\cup\cdots\cup D_n$，并且共有 $t$ 条记录、$d$ 个属性和一个分类标签项 $C$，其中属性集

$a=\{a_1,a_2,\cdots,a_d\}$，且各属性有 $U$ 个可能的取值 $\{a_i^1,a_i^2,\cdots,a_i^U\}$，$i=\{1,2,\cdots,d\}$。

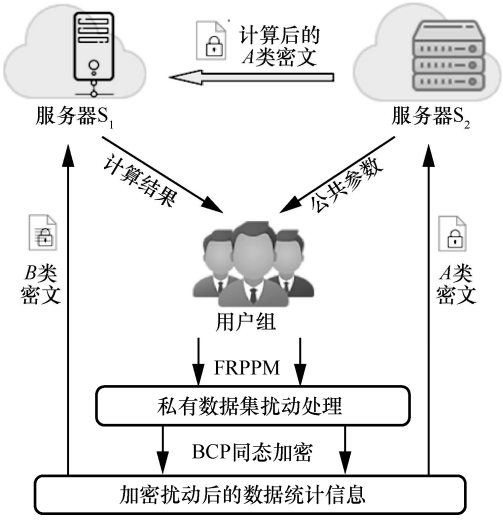



图 2 联合委托学习的系统模型框架

### 3.2 安全模型

假设所有的参与者和服务器都是半诚实的并且不存在共谋行为，即服务器会认真完成客户端的计算任务但对计算结果好奇，客户端会诚实地提供自己的密文数据但同样好奇其他客户端的数据，并且服务器不具备数据集属性名及其取值类别信息等先验知识。

模型假设每个客户端都有私有数据对 $(x_i,y_i)$，其中 $i=(1,2,\cdots,n)$，客户端之间不愿透露数据真实值但又希望利用其他人的数据计算最终结果 $R$。

$$R=\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}y_i} \tag{2}$$

初始化阶段。由服务器 $S_2$ 生成公共参数 $\mathrm{pp}=(N,k,g)$ 并分发给每一个客户端。各客户端选择随机数 $t_i\in Z_{N^2}$，生成各自的公钥 $\mathrm{pk}_i$ 和私钥 $\mathrm{sk}_i$，即 $\mathrm{pk}_i=g^{t_i}\bmod N^2,\mathrm{sk}_i=t_i$。

加密阶段。各客户端选择随机数 $r_i,s_i\in Z_{N^2}$，随后对自己的敏感数据加密获得

$$\begin{cases}A_{x_i}=g^{t_i r_i}\bmod N^2\\ B_{x_i}=g^{t_i r_i}(1+x_iN)\bmod N^2\end{cases} \tag{3}$$

$$\begin{cases}A_{y_i}=g^{t_i s_i}\bmod N^2\\ B_{y_i}=g^{t_i s_i}(1+y_iN)\bmod N^2\end{cases} \tag{4}$$

最后将 $B_{x_i}$ 和 $B_{y_i}$ 发送给服务器 $S_1$，同时将 $A_{x_i}$ 和 $A_{y_i}$ 发送给服务器 $S_2$。

计算阶段。服务器 $S_2$ 接收到各客户端发送的数据后计算 $A_x=\prod_{i=1}^{n}A_{x_i}$，$A_y=\prod_{i=1}^{n}A_{y_i}$，并将 $A_x$ 和 $A_y$ 发送给服务器 $S_1$。服务器 $S_1$ 接收到各客户端及服务器 $S_2$ 发送的数据后计算 $B_x=\prod_{i=1}^{n}B_{x_i}$，$B_y=\prod_{i=1}^{n}B_{y_i}$，$X=\frac{B_x}{A_x}-1\bmod N^2$，$Y=\frac{B_y}{A_y}-1\bmod N^2$ 以及 $R=\frac{X}{Y}$。

输出阶段。服务器 $S_1$ 将最终计算结果 $R$ 返回给每一个客户端。

## 4 基于虚假记录的隐私保护方法

本节针对决策树的构造提出了一种新的隐私保护方法。该方法类似于数据扰动的方法，但与之不同的是客户端不在原始数据中做扰动操作，而是添加完整的虚假记录来达到扰动效果。

以表 1 中的数据为例，前 14 条记录是真实的数据，最后 2 条是添加的虚假记录。换句话说，在原始数据属性 outlook 的取值情况中并没有 foggy 类型，同样其他属性也都没有对应的 cold、low、calm 以及 breezy 的取值类型。简而言之，最后 2 条记录是虚构的，其目的在于通过添加虚假记录的方式来生成干扰树枝，使服务器无法辨认决策树分支的真实性。

以 ID3 算法为例，表 1 中的数据可以生成如图 3 所示的决策树 T'。

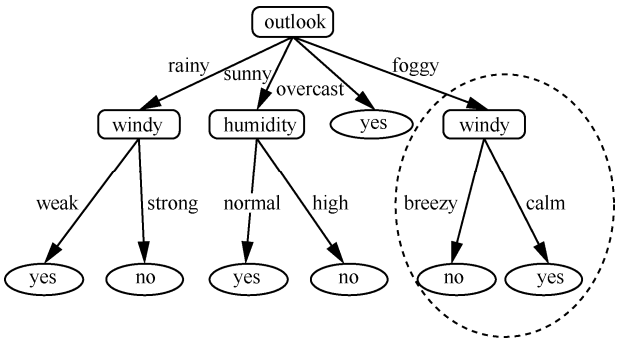



图 3 具有干扰分支的决策树 T'

图 3 中，虚线框中的分支就是生成的干扰分支，当所有节点以及分支信息处于加密状态时，服务器

是很难猜测或分辨分支的真实性的，而客户端解密后通过后剪枝的方式剪掉虚假的分支可以轻松获得真实决策树 T。

在添加虚假记录进行扰动时，有以下两点是需要注意的。

1) 由于 ID3 算法的核心是以属性的信息增益大小来确定决策树各节点的划分属性的，因此，在添加虚假记录后要保证真实数据中原本信息增益最高的属性仍保持最高。例如，在表 1 的真实数据中，属性 outlook 的信息增益最大，当加入虚假记录后仍要保证属性 outlook 的信息增益最大，否则就破坏了真实决策树的结构，根节点不再是属性 outlook，也就是说，降低了决策树模型的分类精度。

2) 添加虚假记录的方法在提高安全的同时，也由于增加虚假数据集 $D'$ 导致挖掘计算量增大。因此本文限定添加的虚假数据集为原真实数据集的 2%～15%，当增加的计算量在可接受的范围内时，客户端应尽量增加更多根节点可能取值的虚假类别以提高决策树的安全性，这一点将在后续的安全性分析章节进行具体介绍。另外，$n$ 个客户端在联合委托学习之前可以通过协商确定添加的虚假数据集 $D'$ 或由某一个客户端设定虚假数据集分发给其他客户端。当数据垂直分布时，各客户端也应将数据集 $D'$ 垂直分割，因此各客户端添加的记录数为 $|D'|$；当数据水平分布时，则每个客户端添加的数据量为 $\frac{|D'|}{n}$。

## 5　联合委托学习协议

根据第 3 节和第 4 节提出的联合委托学习模型及基于虚假记录的隐私保护方法，可以设计以下数据在不同分布形式时的委托学习协议。

### 5.1　数据垂直分布的委托学习协议

当数据垂直分布时，各客户端虽然包含的记录信息不完整，但可以使用布尔化向量表示数据在各属性上的取值情况。例如，在表 1 中，$P_1$ 使用向量表示所有记录在属性 outlook 上取值为 sunny 的情况

$$\boldsymbol{V}_{\mathrm{s}}=[1,1,0,0,0,0,0,1,1,0,1,0,0,0,0,0]^{\mathrm{T}} \tag{5}$$

其中，“1”表示该记录在属性 outlook 上取值为 sunny，“0”表示取其他值。同理，$P_2$ 可以表示所有记录在属性 humidity 上取值为 high 的情况

$$\boldsymbol{V}_{\mathrm{h}}=[1,1,1,1,0,0,0,1,0,0,0,1,0,1,0,0]^{\mathrm{T}} \tag{6}$$

尽管 $P_1$ 和 $P_2$ 之间都不愿透露各自的数据信息，但可以通过向量点积 $\boldsymbol{V}_{\mathrm{s}}\cdot\boldsymbol{V}_{\mathrm{h}}$ 来获得同时满足在属性 outlook 和 humidity 上分别取值为 sunny 和 high 的记录数。其中，求解点积 $\boldsymbol{V}_{\mathrm{s}}\cdot\boldsymbol{V}_{\mathrm{h}}$ 的过程可以看作第 3.2 节安全模型中假设的一部分，如 $\sum_{i=1}^{n}x_i$。

根据以上描述，各客户端可以将各自的私有向量加密后发送给服务器，由服务器代理进行点积运算，具体如算法 1 所示。

**算法 1**　隐私保护委托点积算法

**输入**　每个客户端 $\{P_i\,|\,1\leqslant i\leqslant n\}$ 分别输入各自 $(t+|D'|)$ 维的隐私向量 $\boldsymbol{V}_i$

**输出**　结果 $R$

1) 服务器 $S_2$ 采用 BCP 同态加密算法生成公共参数 pp 并分发给各客户端。

2) 各客户端基于公共参数生成各自的密钥对 $(\mathrm{pk}_i,\mathrm{sk}_i)$。

3) 各客户端利用各自的公钥对向量 $\boldsymbol{V}_i$ 中每一个元素进行加密，得到密文向量对 $(\boldsymbol{V}_i^A,\boldsymbol{V}_i^B)$，并将 $\boldsymbol{V}_i^A$ 和 $\boldsymbol{V}_i^B$ 分别发送给服务器 $S_2$ 和 $S_1$。

4) 服务器 $S_2$ 接收到各方发送的向量 $\boldsymbol{V}_i^A$ 后做如下计算。

令 $j=1$，while($j\leqslant t+|D'|$) do {取 $\boldsymbol{V}_i^A$ 中的第 $j$ 个元素 $v_{i,j}$ 计算 $v_j=\prod_{i=1}^{n}v_{i,j}$，将 $v_j$ 加入向量 $\boldsymbol{V}^A$ 中，$j=j+1$}，最后将向量 $\boldsymbol{V}^A$ 发送给服务器 $S_1$。

5) 服务器 $S_1$ 接收到各客户端及服务器 $S_2$ 发送的向量 $\boldsymbol{V}_i^B$ 及 $\boldsymbol{V}^A$ 后做如下计算。

令 $j=1$，$R=0$，while($j\leqslant t+|D'|$) do {取 $\boldsymbol{V}_i^B$ 及 $\boldsymbol{V}^A$ 中第 $j$ 个元素 $v_{i,j}$ 及 $v_j$ 计算 $v=\prod_{i=1}^{n}v_{i,j}$，

$$m=\frac{v/v_j-1\bmod N^2}{N}$$

if($m\leqslant 1$) 令 $m=0$,else 令 $m=1$, $R=R+m$, $j=j+1$}

6) 服务器 $S_1$ 将结果 $R$ 返回给各客户端。

注：向量 $\boldsymbol{V}_i^A$ 和 $\boldsymbol{V}_i^B$ 表示隐私向量 $\boldsymbol{V}_i$ 中每一个元素加密之后生成的密文对 $(A,B)$ 分别组成的向量，即 $A$ 类密文组成向量 $\boldsymbol{V}_i^A$，$B$ 类密文组成向量 $\boldsymbol{V}_i^B$。

根据上述算法，可以设计数据垂直分布时的联合委托学习协议，具体介绍如下。

1) 各客户端利用自身的数据计算出数据集 $D$ 的信息熵

$$\mathrm{Ent}(D)=-\sum_{k=1}^{K}p_k\mathrm{lb}p_k \tag{7}$$

其中，$K$ 表示数据集 $D$ 中分类标签项可能取值的类别数，$p_k$ 表示第 $k$ 个类别的样本所占的比例，$k=\{1,2,\cdots,K\}$。

2) 各客户端计算出各自所具有的属性信息增益大小，以此来共同确定信息增益最大的属性并作为决策树的根节点。

$$\mathrm{Gain}(D,a_i)=\mathrm{Ent}(D)-\sum_{u=1}^{U}\frac{|D^u|}{|D|}\mathrm{Ent}(D^u) \tag{8}$$

其中，$D^u$ 表示在第 $u$ 个分支中包含的 $D$ 中所有在属性 $a_i$ 上取值为 $a_i^u$ 的样本集，$u=\{1,2,\cdots,U\}$。

3) 各客户端共同协商决定添加虚假记录的数量以及虚假数据的具体值，并将数据集 $D$ 和 $D'$ 布尔化，用布尔型向量表示所有记录在某一个属性上的取值情况。

4) 由当前划分（只有一个节点时指根节点）属性所属的客户端计算出各分支的信息熵并发送给其他客户端。若某分支的信息熵为零，则直接标记该分支为叶子节点，否则共同委托服务器计算该分支的其他信息，以便选取该分支的划分属性。

5) 各客户端将各自的私有向量加密后发送给服务器，并由服务器返回计算结果。

6) 各客户端根据结果 $R$ 计算出各属性的信息增益并确定信息增益最大的属性为该分支节点。再返回到步骤 4)，以类似的方式递归地构造树的其他节点。

下面，以表 1 中的数据为例，说明上述协议的具体执行过程。

1) 由于站点 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 都拥有标签项数据和不完整的记录数据，因此各客户端可以利用自身的数据计算出数据集 $D$ 的信息熵

$$\mathrm{Ent}(D)=-\frac{9}{14}\mathrm{lb}\frac{9}{14}-\frac{5}{14}\mathrm{lb}\frac{5}{14}=0.94 \tag{9}$$

2) 站点 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 也可以计算出各自所具有的属性信息增益大小，并发布给其他客户端。

$\mathrm{P}_1$ 计算：
$$\begin{cases}\mathrm{Gain}(D,\mathrm{outlook})=0.246\\ \mathrm{Gain}(D,\mathrm{temperature})=0.029\end{cases} \tag{10}$$

$\mathrm{P}_2$ 计算：
$$\begin{cases}\mathrm{Gain}(D,\mathrm{humidity})=0.151\\ \mathrm{Gain}(D,\mathrm{windy})=0.048\end{cases} \tag{11}$$

因此选择属性 outlook 为根节点。

3) 虽然 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 可以在不透露具体数据的情况下共同确定根节点，但余下的所有分支节点必须利用整个数据集信息计算才能确定。因此在确定根节点后，$\mathrm{P}_1$ 和 $\mathrm{P}_2$ 需要将各自数据的统计信息委托给服务器进行整合计算，为了保证数据以及模型的安全，首先 $\mathrm{P}_1$ 和 $\mathrm{P}_2$ 需要共同协商决定添加虚假记录并将数据集布尔化。

4) 由 $\mathrm{P}_1$ 计算 outlook 属性划分的 4 个分支的信息熵，其中包括 foggy 分支。例如，式(12)计算的 sunny 分支不为零，因此该分支下的节点为非叶子节点。

$$\mathrm{Ent}(D_{\mathrm{s}})=-\left(\frac{2}{5}\mathrm{lb}\frac{2}{5}+\frac{3}{5}\mathrm{lb}\frac{3}{5}\right)=0.971 \tag{12}$$

$\mathrm{P}_1$ 将 $\mathrm{Ent}(D_{\mathrm{s}})$ 发送给 $\mathrm{P}_2$，并联合 $\mathrm{P}_2$ 将各自数据的统计信息发送给服务器计算其余 3 个属性在该分支下的信息增益情况。例如计算 humidity 属性的信息增益值。用 $D_{\mathrm{s\_h}}$ 和 $D_{\mathrm{s\_n}}$ 分别表示当属性 outlook 取 sunny、属性 humidity 取 high 和 normal 时的样本集，则

$$\mathrm{Gain}(D_{\mathrm{s}},\mathrm{humidity})=\mathrm{Ent}(D_{\mathrm{s}})-\frac{|D_{\mathrm{s\_h}}|}{|D_{\mathrm{s}}|}\mathrm{Ent}(D_{\mathrm{s\_h}})-\frac{|D_{\mathrm{s\_n}}|}{|D_{\mathrm{s}}|}\mathrm{Ent}(D_{\mathrm{s\_n}}) \tag{13}$$

其中，$\mathrm{Ent}(D_{\mathrm{s\_h}})$ 和 $\mathrm{Ent}(D_{\mathrm{s\_n}})$ 的计算过程为

$$\mathrm{Ent}(D_{\mathrm{s\_h}})=-\frac{|D_{\mathrm{s\_h\_y}}|}{|D_{\mathrm{s\_h}}|}\mathrm{lb}\frac{|D_{\mathrm{s\_h\_y}}|}{|D_{\mathrm{s\_h}}|}-\frac{|D_{\mathrm{s\_h\_n}}|}{|D_{\mathrm{s\_h}}|}\mathrm{lb}\frac{|D_{\mathrm{s\_h\_n}}|}{|D_{\mathrm{s\_h}}|} \tag{14}$$

$$\mathrm{Ent}(D_{\mathrm{s\_n}})=-\frac{|D_{\mathrm{s\_n\_y}}|}{|D_{\mathrm{s\_n}}|}\mathrm{lb}\frac{|D_{\mathrm{s\_n\_y}}|}{|D_{\mathrm{s\_n}}|}-\frac{|D_{\mathrm{s\_n\_n}}|}{|D_{\mathrm{s\_n}}|}\mathrm{lb}\frac{|D_{\mathrm{s\_n\_n}}|}{|D_{\mathrm{s\_n}}|} \tag{15}$$

其中，$D_{\mathrm{s\_h\_y}}$ 和 $D_{\mathrm{s\_h\_n}}$ 分别表示当属性 outlook 及 humidity 取值为 sunny 和 high 时，标签项 play 取值为 yes 和 no 的样本集。同理，$D_{\mathrm{s\_n\_y}}$ 和 $D_{\mathrm{s\_n\_n}}$ 也分别表示对应的样本集。

5) 在计算各属性的信息增益过程中，各站点需要知道当前分支的样本数等数据信息，但由于各站点之间不愿透露自己的数据样本信息，因此通过第三方服务器进行代理计算。例如，若站点 $\mathrm{P}_2$ 想要知道在 sunny 分支中的样本数，则可以联合站点 $\mathrm{P}_1$ 分别将向量 $\boldsymbol{V}$ 和 $\boldsymbol{V}_{\mathrm{s}}$ 加密后发送给服务器，让其代理

计算出在 sunny 分支中的样本数$|D_{\mathrm{s}}|=\boldsymbol{V}_{\mathrm{s}}\cdot\boldsymbol{V}$，并返回给站点 $P_2$。其中，向量$\boldsymbol{V}$是$(t+|D'|)$维的通用向量，其所有元素都为 1。

类似地，站点 $P_1$ 使用向量

$$\boldsymbol{V}_{\mathrm{s\_y}}=[0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0]^{\mathrm{T}} \quad (16)$$

表示所有记录在属性 outlook 及标签项 play 上的取值情况，其中“1”表示该条记录同时满足在属性 outlook 及标签项 play 上分别取值为 sunny 和 yes，“0”表示不满足。同理，向量$\boldsymbol{V}_{\mathrm{s_n}}$表示所有记录在属性 outlook 及标签项上分别取值 sunny 和 no 的情况。在站点 $P_2$ 的数据中，向量$\boldsymbol{V}_{\mathrm{h}}$表示所有记录在 humidity 属性上取值为 high 的情况，向量$\boldsymbol{V}_{\mathrm{n}}$表示所有记录在 humidity 属性上取值为 normal 的情况。站点 $P_1$ 和 $P_2$ 分别将

$$|D_{\mathrm{s\_h}}|=\boldsymbol{V}_{\mathrm{s}}\cdot\boldsymbol{V}_{\mathrm{h}},\ |D_{\mathrm{s\_n}}|=\boldsymbol{V}_{\mathrm{s}}\cdot\boldsymbol{V}_{\mathrm{n}}$$

$$|D_{\mathrm{s\_h\_y}}|=\boldsymbol{V}_{\mathrm{s\_y}}\cdot\boldsymbol{V}_{\mathrm{h}},\ |D_{\mathrm{s\_h\_n}}|=\boldsymbol{V}_{\mathrm{s\_n}}\cdot\boldsymbol{V}_{\mathrm{h}}$$

$$|D_{\mathrm{s\_n\_y}}|=\boldsymbol{V}_{\mathrm{s\_y}}\cdot\boldsymbol{V}_{\mathrm{n}},\ |D_{\mathrm{s\_n\_n}}|=\boldsymbol{V}_{\mathrm{s\_n}}\cdot\boldsymbol{V}_{\mathrm{n}}$$

加密后发送给服务器，可以得到对应的计算结果。

6) 站点 $P_1$ 和 $P_2$ 收到服务器返回的结果后，各自可以计算出相同的 humidity 属性信息增益数据。再返回到步骤 4)，以相同的方式，站点 $P_1$ 和 $P_2$ 可以计算出余下 2 个属性的信息增益大小，并选择信息增益最大的属性作为 sunny 分支下的划分节点。以类似的方式递归地构造树的其他节点，最后站点 $P_1$ 和 $P_2$ 都可以构造出图 3 中完整的决策树 T'，经过剪掉虚假的分支后获得真正的决策树 T。

### 5.2　数据水平分布的委托学习协议

与数据垂直分布的情况不同，数据水平分布时各客户端无法根据自身的数据计算出总体数据的信息熵和各属性的信息增益，只能委托服务器作为中间节点进行代理计算。

根据式(7)可以看出，计算总体数据的信息熵需要各客户端提供各自数据的统计信息，即

$$\mathrm{Ent}(D)=-\sum_{k=1}^{K}\frac{\sum_{i=1}^{n}|D_{i,k}|}{\sum_{i=1}^{n}|D_i|}\mathrm{lb}\frac{\sum_{i=1}^{n}|D_{i,k}|}{\sum_{i=1}^{n}|D_i|} \quad (17)$$

其中，$D_{i,k}$表示在第$i$个客户端的数据中属于第$k$类样本的数据集。式(17)同样可以看作第 3.2 节中安全模型的假设形式

$$\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}y_i}\mathrm{lb}\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}y_i} \quad (18)$$

假设各客户端使用向量

$$\boldsymbol{V}_{i,D}=[|D_{i,1}|,|D_{i,2}|,\cdots,|D_{i,K}|,|D_i|] \quad (19)$$

表示该客户端的数据所属类别的统计信息，其中，$|D_i|$表示该客户端具有的数据量；使用向量

$$\boldsymbol{V}_{i,j}=[|D_{i,j}^{1}|,|D_{i,j}^{2}|,\cdots,|D_{i,j}^{U}|] \quad (20)$$

表示该客户端的数据在第$j$个属性上取各个属性值的数据量，其中，$|D_{i,j}^{u}|$表示第$i$个客户端的数据在属性$a_j$上取值为$a_j^u$的样本数，$u=\{1,2,\cdots,U\}$，$j=\{1,2,\cdots,d\}$；使用向量组

$$\boldsymbol{V}_{i,D_j}\begin{cases}\boldsymbol{V}_{i,a_j^1}=[|D_{i,a_j^1,1}|,\cdots,|D_{i,a_j^1,K}|,|D_{i,j}^{1}|]\\ \boldsymbol{V}_{i,a_j^2}=[|D_{i,a_j^2,1}|,\cdots,|D_{i,a_j^2,K}|,|D_{i,j}^{2}|]\\ \quad\vdots\\ \boldsymbol{V}_{i,a_j^U}=[|D_{i,a_j^U,1}|,\cdots,|D_{i,a_j^U,K}|,|D_{i,j}^{U}|]\end{cases} \quad (21)$$

表示该客户端的数据在第$j$个属性上取不同属性值时所属类别信息，其中，$\boldsymbol{V}_{i,a_j^1}$表示该客户端的数据在第$j$个属性上取$a_j^1$时的数据所属类别信息，$|D_{i,a_j^1,k}|$表示在第$j$个属性上取值为$a_j^1$并且属于第$k$类的样本数。

下面，给出多个客户端委托服务器代理计算的具体算法，如算法 2 所示。

**算法 2**　隐私保护委托求熵算法

**输入**　每个客户端$\{P_i\,|\,1\leqslant i\leqslant n\}$分别输入各自的隐私向量$\boldsymbol{V}_{i,D}$、$\boldsymbol{V}_{i,j}$以及向量组$\boldsymbol{V}_{i,D_j}$

**输出**　以第$j$个属性划分的信息增益$\mathrm{Gain}(D,a_j)$

1) 服务器$S_2$采用 BCP 同态加密算法生成公共参数 pp 并分发给各客户端。

2) 各客户端基于公共参数生成各自的密钥对$(\mathrm{pk}_i,\mathrm{sk}_i)$。

3) 各客户端利用各自的公钥将向量$\boldsymbol{V}_{i,D}$中每一个元素进行加密，得到密文向量对$(\boldsymbol{V}_{i,D}^{A},\boldsymbol{V}_{i,D}^{B})$，并将$\boldsymbol{V}_{i,D}^{A}$和$\boldsymbol{V}_{i,D}^{B}$分别发送给服务器$S_2$和$S_1$。

4) 两服务器联合计算数据集$D$的信息熵。当服务器$S_2$和$S_1$分别接收到各方发送的向量$\boldsymbol{V}_{i,D}^{A}$和

$\boldsymbol{V}_{i,D}^{B}$ 后做如下计算。

令 $k=1$， while($k \leqslant K+1$) do {$S_2$ 和 $S_1$ 分别从 $\boldsymbol{V}_{i,D}^{A}$ 和 $\boldsymbol{V}_{i,D}^{B}$ 中取出第 $k$ 个元素 $v_{i,k}^{A}$ 和 $v_{i,k}^{B}$ 计算 $v_k^A=\prod_{i=1}^{n} v_{i,k}^{A}$， $v_k^B=\prod_{i=1}^{n} v_{i,k}^{B}$，并将 $v_k^A$ 和 $v_k^B$ 分别加入向量 $\boldsymbol{V}_D^A$ 和 $\boldsymbol{V}_D^B$ 中， $k=k+1$}。随后， $S_2$ 将向量 $\boldsymbol{V}_{\boldsymbol{D}}^{\boldsymbol{A}}$ 发送给服务器 $S_1$。由服务器 $S_1$ 计算整体数据集中各类别所占的比例。$S_1$ 分别从向量 $\boldsymbol{V}_D^A$ 和 $\boldsymbol{V}_D^B$ 中取出最后一个元素，即第 $K+1$ 个元素，计算可得整体数据集的数据量为 $|D|=\dfrac{v_{k+1}^{B}/v_{k+1}^{A}-1 \bmod N^2}{N}$。

令 $k=1$，while($k \leqslant K$) do {分别从 $\boldsymbol{V}_D^B$ 及 $\boldsymbol{V}_D^A$ 中取出第 $k$ 个元素 $v_k^B$ 及 $v_k^A$，计算第 $k$ 个类别所占的比例 $p_k=\dfrac{v_k^B/v_k^A-1 \bmod N^2}{|D|N}$， $k=k+1$}。

最后服务器 $S_1$ 可计算出整体数据集的信息熵 $\mathrm{Ent}(D)=-\sum_{k=1}^{K} p_k \mathrm{lb} p_k$。

5) 返回步骤 3)，客户端采用同样的方式对向量组中的每一个向量加密，并发送给服务器计算出采用第 $j$ 个属性划分时每一个分支所包含的数据的信息熵 $\mathrm{Ent}(D_j^u)$。

6) 各客户端以同样的方式对向量 $\boldsymbol{V}_{i,j}$ 加密，并发送给 2 个服务器计算数据在第 $j$ 个属性上取不同属性值时所占的比例 $p_u$。令 $u=1$， while($u \leqslant U$) do {$S_2$ 和 $S_1$ 分别从 $\boldsymbol{V}_{i,j}^{A}$ 和 $\boldsymbol{V}_{i,j}^{B}$ 中取出第 $u$ 个元素 $v_{i,u}^{A}$ 和 $v_{i,u}^{B}$，计算 $v_u^A=\prod_{i=1}^{n} v_{i,u}^{A}$ 和 $v_u^B=\prod_{i=1}^{n} v_{i,u}^{B}$， $S_2$ 将 $v_u^A$ 发送给服务器 $S_1$，由 $S_1$ 计算比例 $p_u=\dfrac{v_u^B/v_u^A-1 \bmod N^2}{|D|N}$， $u=u+1$}。

7) $S_1$ 计算出以第 $j$ 个属性划分时的信息增益 $\mathrm{Gain}(D,a_j)=\mathrm{Ent}(D)-\sum_{u=1}^{U} p_u \mathrm{Ent}(D_j^u)$。

根据算法 2，可以设计如下数据水平分布的联合委托学习协议。

1) 各客户端以向量的形式表示自身数据所属类别信息。经过各自公钥加密后将 $A$、$B$ 两类密文分别发送给服务器 $S_2$ 和 $S_1$，计算得到整体数据集（或分支包含的子数据集）的信息熵。

2) 类似地，各客户端以向量 $\boldsymbol{V}_{i,j}$ 和 $\boldsymbol{V}_{i,a_j^u}$ 的形式表示数据在第 $j$ 个属性上的取值情况。通过委托服务器进行代理计算，可以获得所有属性的信息增益数据，将信息增益最大的属性作为节点（当前没有节点时作为根节点）属性。

3) 确定根节点后，各客户端协商确定添加虚假记录的数量 $|D'|$ 及具体数据值，并且每一个客户端添加的虚假记录数都为 $\dfrac{|D'|}{n}$。

4) 返回执行步骤 1)，客户端采用同样的方式发送虚假的统计信息委托服务器计算根节点下各分支的信息熵。若某分支的信息熵为零，即表示该分支所包含的样本属于同一类别，则客户端直接标记该分支为此类别的叶子节点。否则执行步骤 2)，委托服务器计算出该分支下的所有属性信息增益，选择信息增益最大的属性作为该分支的节点属性。

5) 反复执行步骤 1)和步骤 2)，以类似的方式递归地构造树的其他节点。

## 6 安全性及性能分析

### 6.1 安全性分析

本节从客户端的数据与最终构建的决策树模型 2 个方面分析本文所提出的隐私保护委托算法和学习协议的安全性。

1) 当服务器 $S_2$ 不能同时获得数据的密文 $A$ 和 $B$ 时，客户端的数据是安全的。

**证明** 服务器 $S_2$ 利用主密钥 $\mathrm{MK}=(p',q')$ 解密的过程如下。

①利用客户端的公钥计算出对应的私钥。

$$(\mathrm{sk}) \bmod N=\frac{(\mathrm{pk}^{p'q'}-1) \bmod N^2}{N} k^{-1} \bmod N \tag{22}$$

其中， $k^{-1}$ 表示 $k$ 模 $N$ 的逆。

②利用密文 $A$ 计算出客户端在加密过程中选择的随机数 $r$。

$$(r) \bmod N=\frac{(A^{p'q'}-1) \bmod N^2}{N} k^{-1} \bmod N \tag{23}$$

③令 $\delta$ 表示 $p'q'$ 模 $N$ 的逆，并且 $\gamma=(\mathrm{sk} \cdot r) \bmod N$，则明文为

$$m=\frac{((B/(g^{\gamma}))^{p'q'}-1) \bmod N^2}{N} \delta \bmod N \tag{24}$$

从上述解密过程可以看出，当服务器 $S_2$ 利用主密钥解密时，必须同时具有密文 $(A,B)$ 才能获得明文 $m$，但在本文所设计的安全模型中，各客户端是将其密文 $A$ 和 $B$ 分别发送给不同的服务器做求和

运算，并最终由服务器$S_1$计算出构建决策树模型的中间结果。因此当服务器$S_1$和$S_2$之间不存在共谋行为时，任何一个服务器都不会具备解密数据的基本条件。综上所述，客户端的数据是安全的。

2) 当数据集的属性个数$d$等基本参数足够大时，最终构建的决策树模型是安全的，即服务器不能从中间结果推测出真正的模型。

**证明**　在数据垂直分布的情况中，隐私保护委托点积算法只要求服务器对布尔化后的向量做内积运算，因此服务器并不了解数据的真实意义和计算目的，所以，很难猜测出有关决策树模型的任何信息。而在数据水平分布的情况中，服务器$S_1$根据计算信息熵和信息增益的结果，可以构造出如图 4 所示的空模型框架。

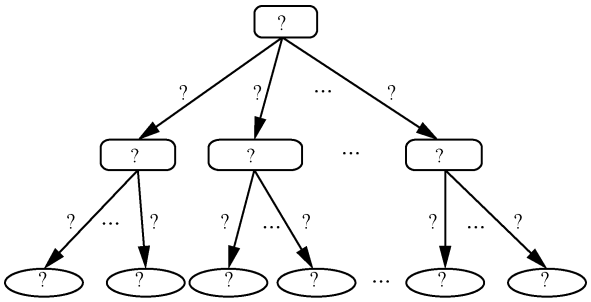


图 4　决策树空模型框架

为方便描述，假设客户端在添加虚假记录的过程中，将根节点属性的取值类别增加了$l$个可能的虚假取值，并且最终构建的决策树模型有$e$个非叶子节点、$f$个中间叶子节点和$h$个底层叶子节点。首先，服务器能够正确匹配所有非叶子节点对应的节点属性的概率可以表示为

$$p_1 = \frac{1}{d(d-1)^{e-1}} \tag{25}$$

其次，服务器能正确匹配所有叶子节点对应的类别信息的概率可表示为

$$p_2 = \left(\frac{1}{K}\right)^f \left(\frac{1}{K^{U-1}(K-1)}\right)^{\frac{h}{U}} \tag{26}$$

其中，$\frac{h}{U}$表示连接底层叶子节点的非叶子节点个数。类似地，服务器能正确匹配所有分支信息的概率可表示为

$$p_3 = \left(\frac{1}{U!}\right)^e \tag{27}$$

最后，服务器能正确剪掉虚假分支的概率可以表示为

$$p_4 = \frac{(U-l)!}{U!} \tag{28}$$

因此，服务器能正确获得完整的决策树模型的概率为

$$p = p_1 p_2 p_3 p_4 = \frac{1}{d(d-1)^{e-1}} \left(\frac{1}{K}\right)^f \left(\frac{1}{K^{U-1}(K-1)}\right)^{\frac{h}{U}} \left(\frac{1}{U!}\right)^e \frac{(U-l)!}{U!} \tag{29}$$

根据式(29)可知，当数据集的属性个数$d$等基本参数以及根节点属性可能的虚假取值数$l$足够大时，服务器能猜测模型的概率$p$是可以忽略的，同时也说明了当增加虚假记录的数量在可接受的范围时，$l$的值越大越能提高模型的安全性。

另外，值得注意的是，上述服务器能够正确获得完整决策树的概率$p$是基于服务器了解数据集基本信息的情况下才成立的。即只有当服务器知道该数据集有哪些具体的属性名及各属性可能的取值时，才能了解该数据集的用途并对模型框架进行匹配和猜测。然而在本文提出的算法中，客户端并未透露任何关于数据集的基本信息，因此进一步降低了服务器根据中间计算结果对模型进行推测的概率。

综上所述，本文提出的联合委托学习协议构建的决策树模型是安全的。

### 6.2　性能分析

本节通过对比客户端与服务器的时间开销来评估本文所提出的算法和协议的性能。实验测试中使用 Python 实现了 PPDDPA 和 PPDEA，建立了安全参数$\lambda$为 1 024 的 BCP 密码系统，并在 Ubuntu 18.04（CPU 主频为 2.6 GHz，型号为 Core i5-3230M，内存为 4 GB）的设备上进行了测试。为了避免网络时延的影响，本文在同一设备上模拟所有客户端和服务器。

首先，对 PPDDPA 和 PPDEA 的性能进行了测试，设定每个客户端的隐私向量是 1 000 维，相当于数据集的记录数为 1 000，特征属性个数$d > 500$。如图 5 所示，在 PPDDPA 的性能测试中可以看出，两服务器的时间花销总和及各参与的客户端的时间花销几乎不随着客户端的数量增加而增加，这表明 PPDDPA 的性能几乎不受客户端数量的影响。其实在该算法的执行过程中也可以看出这一点，每当该

算法执行一次时，不管客户端的数量是多少，实际上只有 2 个客户端参与其中并只对各自的向量进行加密操作，同时服务器也只对 2 个向量做点乘运算。

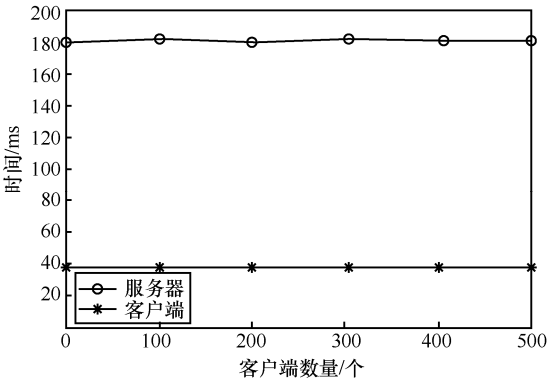



图 5 PPDDPA 性能测试

在 PPDEA 的测试中，每个客户端都有 1 000 条数据记录，因此每一个客户端都会参与其中，从图 6 可以看出，服务器的时间花销随着客户端数量的增加而显著增加，而各客户端的时间花销几乎不受影响。同样也可以从该算法的执行过程看出，各客户端仅执行向量加密操作，而服务器的计算量随着客户端数量的增加而增大。综上所述，本文提出的算法不仅适用于少量客户端联合委托学习的情况，而且在大量客户端参与时也能保证各客户端的数据加密成本不随客户端数量的增加而增大。

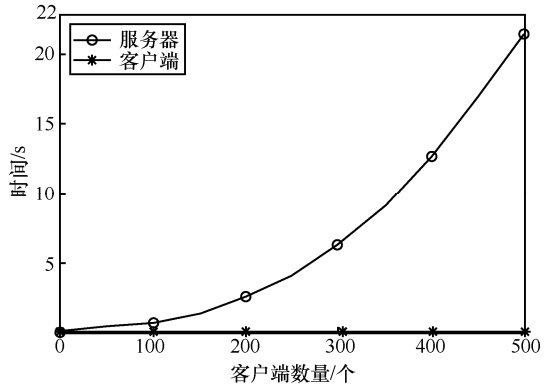



图 6 PPDEA 性能测试

其次，利用急性肝功能衰竭疾病预测数据集对本文提出的协议与 OPPC4.5[22]协议进行了对比测试。由于该数据集只包含 29 个特征属性，因此设定参与的客户端数量最大为 29，并且插入的虚假记录数为 200 条。如图 7 所示，当数据垂直分布时，在本文所提出的协议中由于客户端需要提前计算出整体数据集的信息熵并布尔化数据集，因此计算成本略高于 OPPC4.5 协议，但从整体来看客户端与服务器对模型训练的总成本略低于 OPPC4.5 协议。从图 8 中可以看出，当数据水平分布时，在本文所提出的协议中客户端的计算成本显著低于 OPPC4.5 协议，因为模型训练的计算过程几乎完全由服务器处理，客户端仅需对数据进行统计和加密操作。因此也说明当数据水平分布时，客户端的计算负担得到了显著改善。

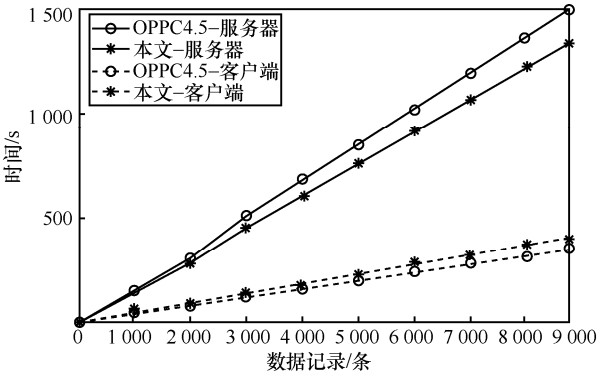



图 7 数据垂直分布的协议性能测试

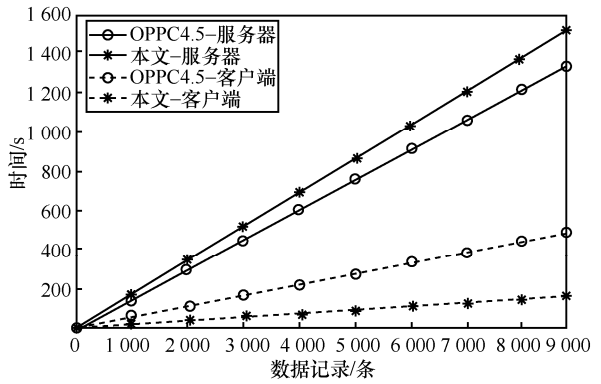



图 8 数据水平分布的协议性能测试

最后，以表 1 中的数据对本文提出的基于虚假记录的隐私保护方法进行了测试。如图 9 所示，无论数据分布情况如何，客户端最终都可以构建正确的决策树模型 T。而在服务器侧，当数据垂直分布时，两服务器都得不到任何关于模型的信息。当数据水平分布时，只有服务器 $S_1$ 可以推测出如图 10 所示的模型框架，且仅能使用不确定的信息（字母）代替节点和分支的信息。对于机器学习模型训练而言，学习的过程就是对模型参数的调参过程，而没有参数的模型是毫无价值的。通常隐私保护的决策树挖掘方法均使用隐藏节点名称达到保密的目的，本文在此基础上增加了虚假分支的方法（图 10 中 A 节点下的分支中(a, c, d) 必有一个分支是虚假的）以此来扰动决策树模型结构，进一步提升了决策树模型的安全性。

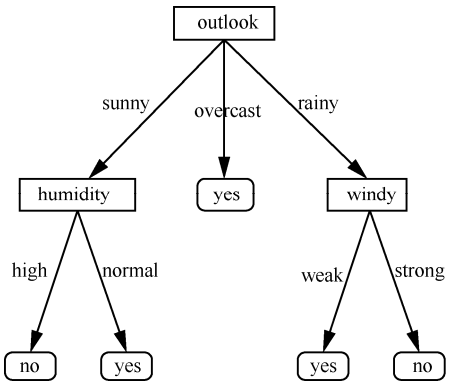



图 9　客户端最终获得的模型 T

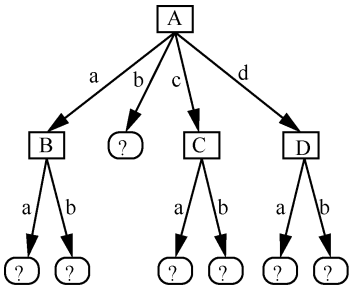



图 10　$S_1$ 可推测的虚假模型框架

## 7　结束语

为降低用户隐私数据在数据共享过程中的泄露风险，同时减少客户端在数据挖掘过程中的计算成本，本文基于传统的委托计算思想和 BCP 同态加密算法提出了一种联合委托学习模型，该模型采用双服务器分别计算客户端的部分密文信息的方式来降低数据共享中的隐私泄露风险。进一步地，针对决策树的安全构造，提出了一种基于虚假记录的隐私保护方法，该方法利用少量的虚假记录改变了数据统计的真实结果，并对决策树的模型结构进行扰动，避免了服务器获得真实的中间计算结果和最终训练的模型结构。另外，分别对数据垂直分布和水平分布的情况设计了隐私保护委托算法及联合委托学习协议，在保证数据安全共享的同时降低了客户端的计算成本。最后，通过实验测试结果表明，在联合委托学习过程中，各客户端的数据加密成本不随客户端数量的增加而增大，并且最终获得的模型与真实数据构建的模型具有一致性，即最终挖掘得到的模型准确度没有任何损失，而服务器很难推测和匹配出真实的模型结构。

## 参考文献：

[1] JIN H, LUO Y, LI P L, et al. A review of secure and privacy-preserving medical data sharing[J]. IEEE Access, 2019, 7: 61656-61669.

[2] SUN Y, YIN L H, SUN Z, et al. An IoT data sharing privacy preserving scheme[C]//IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops. Piscataway: IEEE Press, 2020: 984-990.

[3] ZHAO C, ZHAO S N, ZHAO M H, et al. Secure multi-party computation: theory, practice and applications[J]. Information Sciences, 2019, 476: 357-372.

[4] VU D H, LUONG T D, HO T B. An efficient approach for secure multi-party computation without authenticated channel[J]. Information Sciences, 2020, 527: 356-368.

[5] REDDY G T, REDDY M P K, LAKSHMANNA K, et al. Analysis of dimensionality reduction techniques on big data[J]. IEEE Access, 2020, 8: 54776-54788.

[6] WANG C, WANG A D, XU J, et al. Outsourced privacy-preserving decision tree classification service over encrypted data[J]. Journal of Information Security and Applications, 2020, 53: 102517.

[7] OLAKANMI O O, DADA A. An efficient privacy-preserving approach for secure verifiable outsourced computing on untrusted platforms[J]. International Journal of Cloud Applications and Computing, 2019, 9(2): 79-98.

[8] MANIKANDAN V, PORKODI V, MOHAMMED A S, et al. Privacy preserving data mining using threshold based fuzzy c-means clustering[J]. ICTACT Journal on Soft Computing, 2018, 9(1): 1813-1816.

[9] TEO S G, CAO J, LEE V C S. DAG: a general model for privacy-preserving data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(1): 40-53.

[10] AGRAWAL R, SRIKANT R. Privacy-preserving data mining[C]//2000 ACM SIGMOD International Conference on Management of Data. Texas. New York: ACM Press, 2000: 439-450.

[11] KARGUPTA H, DATTA S, WANG Q, et al. On the privacy preserving properties of random data perturbation techniques[C]//Third IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2003: 99-106.

[12] KARGUPTA H, DATTA S, WANG Q, et al. Random-data perturbation techniques and privacy-preserving data mining[J]. Knowledge and Information Systems, 2005, 7(4): 387-414.

[13] BU S, LAKSHMANAN L V S, NG R T, et al. Preservation of patterns and input-output privacy[C]//2007 IEEE 23rd International Conference on Data Engineering. Piscataway: IEEE Press, 2007: 696-705.

[14] HAMADA K, HASEGAWA S, MISAWA K, et al. Privacy-preserving fisher's exact test for genome-wide association study[C]//International Workshop on Genome Privacy and Security. Piscataway: IEEE Press, 2017: 99-102.

[15] BOST R, POPA R A, TU S, et al. Machine learning classification over encrypted data[C]//Network and Distributed System Security Symposium. Piscataway: IEEE Press, 2015: 4324-4325.

[16] WU D J, FENG T, NAEHRIG M, et al. Privately evaluating decision trees and random forests[J]. Proceedings on Privacy Enhancing Technologies, 2016, 2016(4): 335-355.

[17] BACKES M, BERRANG P, BIEG M, et al. Identifying personal DNA methylation profiles by genotype inference[C]//2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2017: 957-976.

[18] ICHIKAWA A, OGATA W, HAMADA K, et al. Efficient secure mul-

ti-party protocols for decision tree classification[C]//Australasian Conference on Information Security and Privacy. Piscataway: IEEE Press, 2019: 362-380.

[19] ZHENG Y F, DUAN H Y, WANG C. Towards secure and efficient outsourcing of machine learning classification[C]//European Symposium on Research in Computer Security. Berlin: Springer, 2019: 22-40.

[20] LI P, LI J, HUANG Z G, et al. Privacy-preserving outsourced classification in cloud computing[J]. Cluster Computing, 2018, 21(1): 277-286.

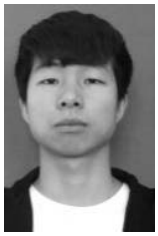
[21] LIU L, SU J, CHEN R, et al. Secure and fast decision tree evaluation on outsourced cloud data[C]//International Conference on Machine Learning for Cyber Security. Piscataway: IEEE Press, 2019: 361-377.

[22] LI Y, JIANG Z L, YAO L, et al. Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties[J]. Cluster Computing, 2019, 22(1): 1581-1593.

[23] PHU V N, TRAN V T N, CHAU V T N, et al. A decision tree using ID3 algorithm for English semantic analysis[J]. International Journal of Speech Technology, 2017, 20(3): 593-613.

[作者简介]

**高胜**（1987− ），男，湖北黄冈人，博士，中央财经大学副教授、硕士生导师，主要研究方向为数据安全与隐私保护、区块链技术及应用等。



**向康**（1993− ），男，湖北仙桃人，贵州大学硕士生，主要研究方向为委托机器学习、委托计算与博弈论。

**田有亮**（1982− ），男，贵州六盘水人，博士，贵州大学教授、博士生导师，主要研究方向为算法博弈论、密码学与安全协议、大数据安全与隐私保护等。

**谭伟杰**（1981− ），男，陕西合阳人，博士，贵州大学讲师，主要研究方向为通信信号处理、通信网络安全、阵列信号处理。

**冯涛**（1970− ），男，甘肃临洮人，博士，兰州理工大学研究员、博士生导师，主要研究方向为网络与信息安全、密码学。

**吴晓雪**（1977− ），男，江苏南通人，贵州省计量测试院工程师，主要研究方向为能源计量。