面向纵向联邦学习的隐私保护数据对齐框架

高 莹*①② 谢雨欣① 邓煌昊① 朱祖坤① 张一余①
①(北京航空航天大学网络空间安全学院 北京 100191)
②(中关村实验室 北京 100194)

摘 要：纵向联邦学习中，各个客户端持有的数据集中包含有重叠的样本ID和不同维度的样本特征，需要进行数据对齐以适应模型训练。现有数据对齐技术一般将各方样本ID交集作为公开信息，如何在不泄露样本ID交集的前提下实现数据对齐成为亟需解决的问题。基于可交换加密和同态加密技术，该文构造了隐私保护的数据对齐框架ALIGN，包括数据加密、密文盲化、密文求交和特征拼接等步骤，使得相同的原始样本ID经过双重可交换加密可变换为相同的密文，并且对样本特征经同态加密后又进行了盲化处理。ALIGN框架能够对参与方样本ID的密文求交，将交集内样本ID对应的全部特征数据进行拼接并以秘密分享形式分配给参与方。相比现有数据对齐技术，该框架不仅能够保护样本ID交集的隐私性，同时能安全地删除样本ID交集外的样本信息。对ALIGN框架的安全性证明表明，除数据规模外，各客户端不能通过数据对齐获得关于对方数据的任何信息，保证了隐私保护策略的有效性。与现有工作相比，每增加10%的冗余数据，ALIGN框架利用所得数据对齐结果可将模型训练时间缩短约1.3秒，将模型训练准确度稳定在85%以上。仿真实验结果表明，通过ALIGN框架进行纵向联邦学习数据对齐，有利于提升后续模型训练的效率和模型准确度。

关键词：纵向联邦学习；数据对齐框架；隐私保护；可交换加密；同态加密
中图分类号：TN918;TP309 **文献标识码**：A **文章编号**：1009-5896(2024)00-0001-09
DOI: 10.11999/JEIT231234

A Privacy-preserving Data Alignment Framework for Vertical Federated Learning

GAO Ying①② XIE Yuxin① DENG Huanghao① ZHU Zukun① ZHANG Yiyu①
①(*School of Cyber Science and Technology*, *Beihang University*, *Beijing* 100191, *China*)
②(*Zhongguancun Laboratory*, *Beijing* 100094, *China*)

Abstract: In vertical federated learning, the datasets of the clients have overlapping sample IDs and features of different dimensions, thus the data alignment is necessary for model training. As the intersection of the sample IDs is public in current data alignment technologies, how to align the data without any leakage of the intersection becomes a key issue. The proposing private-preserving data alignment framework is based on interchangeable encryption and homomorphic encryption technologies, mainly including data encryption, ciphertext blinding, private intersecting, and feature splicing. The sample IDs are encrypted twice based on an interchangeable encryption algorithm, where the same ciphertexts correspond to the same plaintexts, and the sample features are encrypted and then randomly blinded based on a homomorphic encryption algorithm. The intersection of the encrypted sample IDs is obtained, and the corresponding features are then spliced and secretly shared with the participants. Compared to the existing technologies, the privacy of the ID intersection is protected, and the samples corresponding to the IDs outside intersection can be removed safely in our framework. The security proof shows that each participant cannot obtain any knowledge of each other except for the data size, which guarantees the effectiveness of the private-preserving strategies. The simulation experiments demonstrate that the runtime is shortened about 1.3 seconds and the model accuracy keeps higher than 85% with every 10% reduction of the redundant data. The simulation experimental results show that

收稿日期：2023-11-07；改回日期：2024-04-02
*通信作者： 高莹 gaoying@buaa.edu.cn
基金项目：北京市自然科学基金 (No. M21033)，腾讯微信犀牛鸟基金
Foundation Item: Natural Science Foundation of Beijing Municipality (M21033),Tencent Rhino-Bird Joint Research Program.

using the ALIGN framework for vertical federated learning data alignment is beneficial for improving the efficiency and accuracy of subsequent model training.

Key words: Vertical federated learning; Data alignment; Privacy protection; Commutative encryption; Homomorphic encryption

1 引言

随着大数据时代的到来和计算能力的空前发展，人工智能已经成为引领新一代产业变革的新型技术，其将对企业竞争、社会发展，甚至国家战略产生深远的影响。然而，由于训练数据的隐私问题和参与方之间的信任问题，使得隐私数据难以流通，无法充分发挥数据的潜在应用价值，制约着人工智能技术的进一步发展。联邦学习[1–5]作为人工智能技术的新分支，能够在满足隐私保护和数据安全的前提下，在多参与方之间进行数据共享和模型训练，为同时解决数据孤岛问题和隐私保护问题提供了新的解决方案。通常地，根据具体应用场景下数据的组织和使用形式不同，联邦学习应用服务技术可分为“纵向”和“横向”两种。其中，纵向联邦学习[6–10]，也称作按特征划分的联邦学习，适用于各参与方的训练数据有重叠的样本ID，而数据特征有所不同的场景，广泛应用于企业间合作进行人工智能模型训练。

然而，在纵向联邦学习的数据对齐阶段，各个参与方需要找到共有样本ID以完成数据(特征、标签)对齐，而样本ID是样本中能够唯一标识每个用户身份的信息字段，为重要的隐私信息，如用户名、手机号、身份证号等，数据对齐阶段对各参与方公开共有样本ID造成了严重的隐私安全隐患。在参与方拥有样本规模相差悬殊时，共有样本ID的泄漏使得样本规模大的参与方可能得到样本规模小的参与方的大部分样本ID，将导致更加严重的隐私安全问题。另外，对于联邦学习的模型训练来说，样本ID为无效信息，不会对模型训练过程和结果产生影响。因此，避免数据对齐阶段的样本ID隐私泄漏能够保护隐私数据隐私性，使纵向联邦学习同时适用于对称及非对称场景，为纵向联邦学习中亟需解决的研究问题。纵向联邦学习数据对齐及训练阶段如图1所示。

为了保证数据对齐过程中样本ID的隐私安全，可采用隐私集合求交(Private Set Intersection, PSI)技术使各参与方得到重叠的样本ID, PSI允许两个或多个参与者秘密地计算他们的交集而不泄露除交集之外的其他任何信息。目前，两个参与方场景下的PSI已有多种实现方案，如基于公钥加密体制的PSI[11]，基于混淆电路的PSI[12]，基于OT的PSI[13,14]等。PSI 技术可以保护交集以外的信息，但当直接用于数据对齐时并不能保护交集中样本ID的信息，即各参与方可获得哪些样本ID存在于交集之中的信息，仍然造成隐私泄漏。Lu等人[15]基于轻量级密码原语设计了一种新的PSI协议，可有效提升纵向联邦学习数据对齐效率，但未对样本ID交集的隐私性保护做出提升。为了进一步保护交集隐私，Hardy 等人[16]在得到PSI求交后对交集结果进行了加密，Liu 等人[17]在PSI求交后对交集结果进行了混淆。然而，额外增加的加密和盲化操作又降低了数据对齐的整体效率。

随着PSI技术的发展，Circuit-PSI[12,18–20]技术由于可以在不泄露交集信息的前提下对交集进行任意函数计算而逐渐引起关注。基于Circuit-PSI技术的纵向联邦学习数据对齐过程中，用1或0作为标签表示每个样本ID处于交集之内或之外。对于交集内样本，将标签和特征拼接作为秘密值进行秘密分享；对于交集外样本，将标签和随机数拼接作为秘密值进行秘密分享。因此，基于Circuit-PSI技术的数据对齐得到的结果中，不仅包含交集内样本的特征，还包含了冗余的随机数据，这部分数据无法直接移除，将继续参与后续的模型训练。为了改进数据对齐结果中存在冗余数据的问题，Liu等人[21]设计了iPrivJoin框架。该框架首先基于布谷鸟哈希和简单哈希对样本进行预处理，然后类似于Circuit-PSI，对样本特征进行秘密分享。为了去除冗余数据，在基于不经意置换技术对秘密分享结果进行混淆后，通过标签重构去除交集之外的样本。这虽然能够解决冗余数据的问题，但数据清洗过程增加了额外的计算和通信开销，同时增加了整个流程的复杂性和工程实现上的困难性；另外，由于使用了布谷鸟哈希和简单哈希进行预处理，iPrivJoin还在安全性方面存在一定的隐患，即当布谷鸟哈希或者简单哈希失败时，敌手可以通过这一信息推断诚实方的输入信息。同时由于布谷鸟哈希存在一定的假阳性误判概率导致过程中生成了1.4倍的无效数据。

针对上述问题，本文基于可交换加密[22,23]、同态加密[24]、秘密分享[25,26]等基本的密码学原语，设计了一个新的数据对齐框架ALIGN，能够保护样本ID交集的隐私，拼接交集内样本ID对应的特征并以秘密分享形式分配给参与方，得到不含冗余信

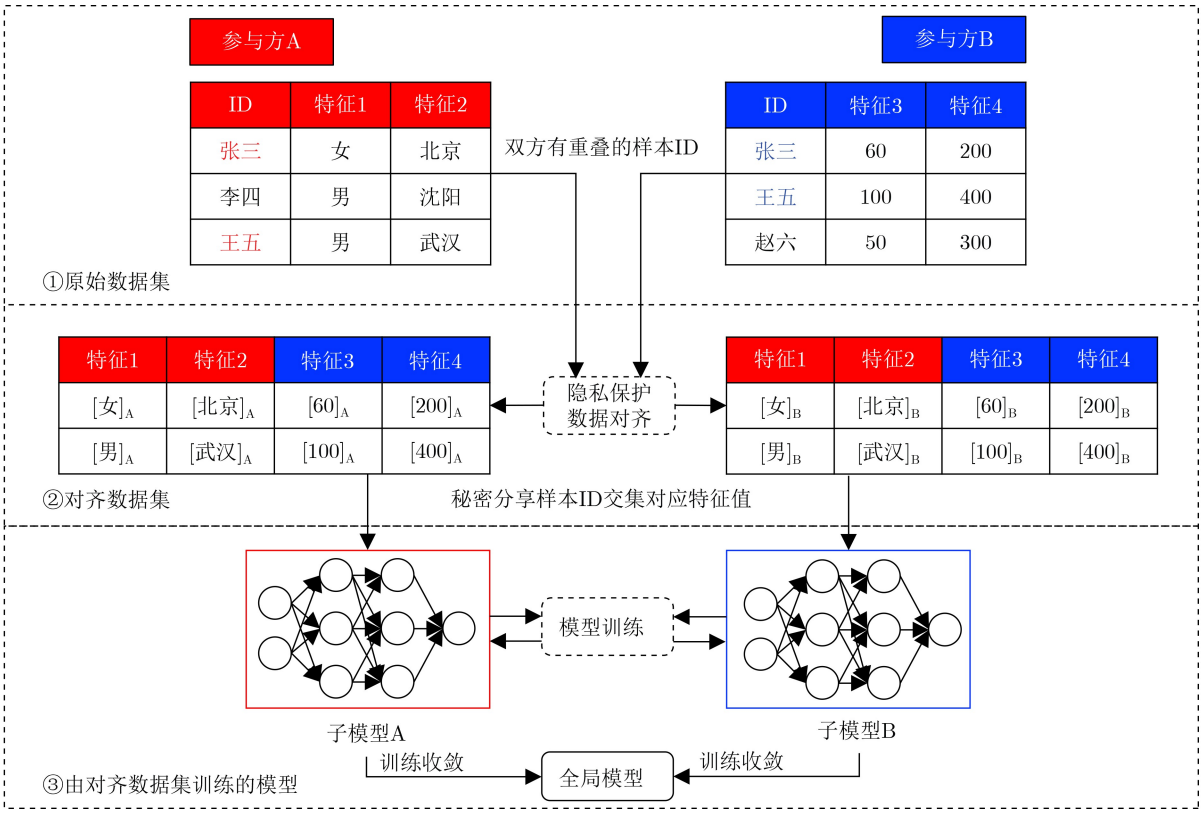


图 1 纵向联邦学习数据对齐及训练阶段示意图

息的对齐结果。相比于现有数据对齐技术，ALIGN框架的主要优势有：具有较强的简洁性，不存在由布谷鸟哈希导致的数据扩张问题，也不需要额外的步骤删除冗余数据；安全性由同态加密模块和可交换加密模块保证，不存在哈希碰撞的安全隐患；底层的同态加密和可交换加密算法均具有可插拔性，框架能够在具体场景下灵活实现。本文贡献主要包含3个方面：

(1)设计了一种隐私保护的数据对齐框架ALIGN，基于可交换加密技术使得两个参与方获取加密的样本ID交集，基于同态加密技术使得特征数据在数据对齐过程中安全传输，解决了数据对齐阶段的样本ID泄漏问题；

(2)在理论上对ALIGN框架的安全性和复杂性进行分析，通过安全性证明保证了框架的安全性，并对框架计算复杂度和通信复杂度提供理论计算结果；

(3)提供框架的实例化实现，通过与Circuit-PSI[20]的仿真实验对比发现，每增加10%的冗余数据，ALIGN框架利用其数据对齐结果可将模型训练时间缩短约1.3秒，将模型训练准确度稳定在85%以上。表明ALIGN框架在提升后续模型训练的效率和模型准确度方面更有优势。

2　预备知识

本节首先简要介绍可交换加密[22,23]、同态加密[24]和加法秘密分享[25,26]，最后描述本文的研究问题。

2.1　可交换加密

设可交换加密算法为可多项式时间内可计算的一组加密函数$f:K\times F\to F$，其中，密钥空间为K，明密文空间为F。记$f_e(x)\triangleq f(e,x)$，f满足如下性质：

(1)可交换性：对任意密钥$e,e'\in K$，明文$x\in F$，有$f_e(f_{e'}(x))=f_{e'}(f_e(x))$；

(2)双射：对任意密钥$e\in K$，函数$f_e:F\to F$是双射的；

(3)可逆：对任意密钥$e\in K$，函数f_e的逆运算$f_e{}^{-1}$在多项式时间内可计算；

(4)不可区分性：随机选择密钥$e\in K$明文$x,y,z\in F$，$(x,f_e(x),y,f_e(y))$与$(x,f_e(x),y,z)$的分布不可区分。

可交换加密算法一般有3类构造，分别是基于RSA的SRA算法[27]、基于DDH假设的Pohlig-Hellman指数加密算法[28]和基于DLP假设的DH密钥交换协议[29]，其中后者可以扩展到基于椭圆曲线的ECDH密钥交换协议，也是一种较为常见的可交换加密算法[23]。

2.2　同态加密

假设加密算法Enc(·)Enc(·)的明文空间为代数

结构M，密文空间为代数结构N，$\cdot$为M上的二元运算，$\odot$为N上的二元运算。对任意$m_1, m_2 \in M$ $\mathrm{m}_1, \mathrm{m}_2 \in \mathrm{M}$，有$\mathrm{Enc}(m_1 \cdot m_2) = \mathrm{Enc}(m_1) \odot \mathrm{Enc}(m_2)$，则称加密算法$\mathrm{Enc}(\cdot)$具有同态性质。

常见的同态加密算法中，有分别支持加法同态和乘法同态的半同态加密算法(Somewhat Homomorphic Encryption, SFE)，也有同时支持加法同态和乘法同态并且允许任意多次乘法和加法运算的全同态加密算法(Fully Homomorphic Encryption, FHE)。Paillier同态加密算法[24]是一个支持加法同态的半同态加密算法，可在明密文空间上保持加法运算和数乘运算，在本文框架的实例实现中可以发挥作用。

2.3 加法秘密分享

设$(G, +)$为加法交换群。针对两个参与方的秘密分享算法将秘密值$x (x \in G)$分为两个份额x_1和x_2，使得仅当同时已知份额x_1和x_2时，才能够恢复秘密值x，记为$x = x_1 + x_2$；而当已知x_1或x_2其中之一时，不能恢复秘密值x。

2.4 问题描述

本文设计数据对齐框架ALIGN。假设存在两个参与方P_A和P_B，双方原始数据集分别为$D_A = \{(\mathrm{ID}_{A,i}, X_{A,i})\}_{i \in [n_A]}$和$D_B = \{(\mathrm{ID}_{B,j}, X_{B,j})\}_{j \in [n_B]}$，其中，$\mathrm{ID}_A, \mathrm{ID}_B$为双方各自样本ID集合，$X_A, X_B$为双方各自样本特征集合，$n_A, n_B$为双方各自样本数量。设双方样本ID交集为$\mathrm{Int} = \mathrm{ID}_A \cap \mathrm{ID}_B$，设交集内ID对应的特征拼接后为$X' = \{x'_m = (X_{A,i}, X_{B,j}) | \mathrm{ID}_{A,i} = \mathrm{ID}_{B,j} \in \mathrm{Int}\}$。数据对齐目标为，使$P_A$ $\mathrm{P_A}$、P_B $\mathrm{P_B}$各自得到X'的秘密分享，过程中不破坏样本ID和样本特征隐私，双方对于Int不能获得除交集大小外的其余相关信息。本文符号汇总见表1。

3 框架设计和分析

本节首先描述ALIGN数据对齐框架的总体设计和详细步骤，其次对框架的计算复杂度和通信进行分析，最后给出框架在半诚实场景下的安全性证明。

表 1 本文符号汇总

符号	含义
P_A, P_B	纵向联邦学习参与方
n_A	参与方P_A的原始数据集包含的样本数量
n_B	参与方P_B的原始数据集包含的样本数量
$[n] (n \in \mathbb{Z})$	集合$\{1, 2, \cdots, n\}$
m	参与方P_A和P_B的样本ID交集大小
$\langle \cdot \rangle_A$	参与方P_A的秘密分享份额
$\langle \cdot \rangle_B$	参与方P_B的秘密分享份额
$[x]_k$	以密钥k加密明文x得到的密文

3.1 框架设计

ALIGN框架总体分为4个步骤：(1)ID加密和特征加密；(2)ID重加密和特征密文盲化；(3)ID密文求交和特征拼接；(4)解密生成秘密份额。双方样本ID经过步骤(1)和(2)的两次可交换加密，相同的原始样本ID具有相同的密文值，则可通过比对求出ID交集。样本特征经过步骤(1)的同态加密后得到密文，又经步骤(2)对密文数据进行随机盲化，构造了双方的秘密分享份额。步骤(2)中还对密文的顺序进行了混淆，避免双方获得明文与密文的映射关系。经过步骤(1)和(2)，样本ID和特征均经过了加密和混淆，则可通过步骤(3)实现隐私保护的ID求交和特征拼接。最终经过步骤(4)的解密运算，双方可得到最终的秘密分享份额。框架流程图见图2。

设参与方P_A和P_B的可交换加密密钥分别为k_A和k_B，同态加密的公私钥密钥对分别为$(\mathrm{pk}_A, \mathrm{sk}_A)$和$(\mathrm{pk}_B, \mathrm{sk}_B)$。ALIGN框架详细描述见算法1。

(1)ID加密和特征加密

参与方P_A和P_B各自利用可交换加密算法处理样本ID，则D_A中第$i (i \in [n_A])$个ID加密为$[\mathrm{ID}_{A,i}]_{k_A}$，$D_B$中第$j (j \in [n_B])$个ID加密为$[\mathrm{ID}_{B,j}]_{k_B}$。双方各自利用己方公钥通过同态加密算法处理样本特征，

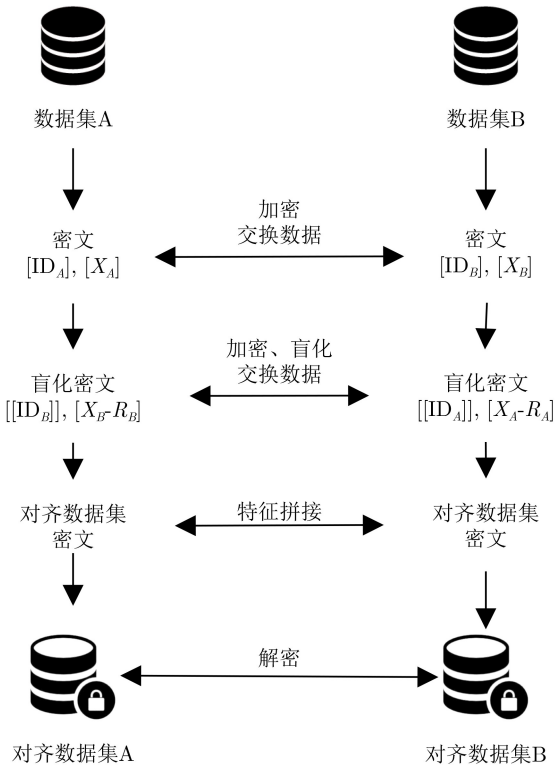


图 2 ALIGN框架流程图

算法 1　ALIGN框架任意参与方数据对齐算法

输入：数据集(ID, X)，可交换加密算法$E(\cdot)$及密钥k，同态加密算法$H(\cdot)$及公私钥对(pk,sk)

输出：数据对齐结果res

```
Begin
  # ID加密和特征加密
  P sends pk to P' and receives pk' from P';
  (ID, X) ← (E_k(ID), H_pk(X));
  P sends (ID, X) to P' and receives (ID_1, X_1) from P';
  # ID重加密和特征密文盲化
  ID_1 ← E_k(ID_1);
  R = [];
  for i in range(|X_1|):
    r_i ← RandomGenerator();# RandomGenerator()表示随机数生成器
    R.append(r_i);
    X_1[i] ← X_1[i] − H_pk(r_i);
    end for
  dict ← {ID_1[i] : R[i] for i in range(|ID_1|)} ;
  π ← PermutationGenerator(); #PermutationGenerator()表示随机置换函数生成器，用于随机置换列表中元素的顺序
  (ID_1, X_1) ← π(ID_1), π(X_1);
  P sends (ID_1, X_1) to P' and receives (ID_2, X_2) from P';
  # ID密文求交和特征拼接
  res_1, res_2 = [], [];
  for i in range(|ID_2|):
    for id in ID_1:
      if ID_2[i] == id:
        res_1.append(X_2[i]);
        res_2.append(dict[id]);
      end if
    end for
  end for
  # 解密生成秘密份额
  result ← (H^-1_sk(res_1),res_2);
End
```

则D_A $\mathrm{D_A}$中的第$i(i \in [n_A])$ $\mathrm{i(i \in [n_A])}$个特征加密为$[X_{A,i}]_{\mathrm{pk}_A}$ $\mathrm{X_{A,i}^{'} = [X_{A,i}]_{pk_A}}$，$D_B$ 中的第$j(j \in [n_B])$ $\mathrm{j\,(j \in [n_B])}$个特征加密为$[X_{B,j}]_{\mathrm{pk}_B}$。双方交换密文，记双方发送的加密数据分别为$X_A{}^1$和$X_B{}^1$，其中

$$X_A^1 = \{([\mathrm{ID}_{A,i}]_{k_A}, [X_{A,i}]_{\mathrm{pk}_A})\}_{i\in[n_A]},$$
$$X_B^1 = \{([\mathrm{ID}_{B,j}]_{k_B}, [X_{B,j}]_{\mathrm{pk}_B})\}_{j\in[n_B]}$$

(2)ID重加密和特征密文盲化

参与方P_A接收到的样本ID密文为$[\mathrm{ID}_{B,j}]_{k_B}$，利用己方密钥对其进行第2次可交换加密，得到$[[\mathrm{ID}_{B,j}]_{k_B}]_{k_A}$，由可交换加密的性质，$[[\mathrm{ID}_{B,j}]_{k_B}]_{k_A} = [\mathrm{ID}_{B,j}]_{k_Ak_B}$。$P_A$生成一组随机数$\{R_{B,j}\}_{j\in[n_B]}$，为接收到的原本特征密文$[X_{B,j}]_{\mathrm{pk}_B}$添加掩码，由同态性质得$[X_{B,j}]_{\mathrm{pk}_B} - [R_{B,j}]_{\mathrm{pk}_B} = [X_{B,j} - R_{B,j}]_{\mathrm{pk}_B}$。$P_A$得到处理后的密文$\{([\mathrm{ID}_{B,j}]_{k_Ak_B}, [X_{B,j} - R_{B,j}]_{\mathrm{pk}_B})\}_{j\in[n_B]}$，保留加密的样本ID密文与随机数的映射关系$\{([\mathrm{ID}_{B,j}]_{k_Ak_B}, R_{B,j})\}_{j\in[n_B]}$。

P_A随机选取置换函数π_A对集合中数据的顺序进行混淆，将计算结果记作X_B^2，有

$$X_B^2 = \{([\mathrm{ID}_{B,\pi_A(j)}]_{k_Ak_B}, [X_{B,\pi_A(j)} - R_{B,\pi_A(j)}]_{\mathrm{pk}_B})\}_{j\in[n_B]}$$

同样地，将参与方P_B计算结果记作X_A^2，有

$$X_A^2 = \{([\mathrm{ID}_{A,\pi_B(i)}]_{k_Ak_B}, [X_{A,\pi_B(i)} - R_{A,\pi_B(i)}]_{\mathrm{pk}_A})\}_{i\in[n_A]}$$

其中，π_B为参与方P_B随机选取的置换函数，用于混淆集合X_A^2中数据的顺序。P_A保留映射关系$\{([\mathrm{ID}_{A,i}]_{k_Ak_B}, R_{A,i})\}_{i\in[n_A]}$。双方交换$X_A^2$和$X_B^2$。

(3)ID密文求交和特征拼接

参与方P_A经上一步计算了两次可交换加密后的样本ID集合$\{[\mathrm{ID}_{B,\pi_A(j)}]_{k_Ak_B}\}_{j\in[n_B]}$，并接收了$P_B$计算的加密样本ID集合$\{[\mathrm{ID}_{A,\pi_B(i)}]_{k_Ak_B}\}_{i\in[n_A]}$。若$[\mathrm{ID}_{A,\pi_B(i)}]_{k_Ak_B}=[\mathrm{ID}_{B,\pi_A(j)}]_{k_Ak_B}$，则参与方$P_A$和$P_B$各自根据ID与随机数的映射关系进行特征拼接得到$([X_{A,\pi_B(i)}-R_{A,\pi_B(i)}]_{\mathrm{pk}_A}, R_{B,\pi_A(j)})$和$(R_{A,\pi_B(i)}, [X_{B,\pi_A(j)}-R_{B,\pi_A(j)}]_{\mathrm{pk}_B})$。由此得到了ID密文的交集和拼接后的特征。设ID密文交集大小为m，将特征拼接后集合重新编号，则双方得到结果可分别记作${X_A}^3$和${X_B}^3$，其中

$$X_A^3=\{([X_{A,l}-R_{A,l}]_{\mathrm{pk}_A}, R_{B,l})\}_{l\in[m]},$$
$$X_B^3=\{(R_{A,l}, [X_{B,l}-R_{B,l}]_{\mathrm{pk}_B})\}_{l\in[m]}$$

(4)解密生成秘密份额

双方根据自己的私钥对X_A^3或X_B^3进行解密，则可得到最终的秘密份额，分别记作$\langle X'\rangle_A$和$\langle X'\rangle_B$，其中

$$\langle X'\rangle_A=\{(X_{A,l}-R_{A,l}, R_{B,l})\}_{l\in[m]},$$
$$\langle X'\rangle_B=\{(R_{A,l}, X_{B,l}-R_{B,l})\}_{l\in[m]}$$

3.2 复杂度分析

设任意参与方的样本数量为n，特征数量为k，双方ID交集的大小为m，单个ID密文长度为l比特，单个特征密文长度为l'比特。本节对计算复杂度和通信复杂度进行分析。

计算复杂度：ALIGN框架的操作类型分为同态加密操作、可交换加密操作和明文操作3种。3.1节的步骤(1)中，双方各自执行nk次同态加密操作以加密样本特征，n次可交换加密操作以加密样本ID；步骤(2)中，双方各自执行nk次同态加密操作以盲化样本特征，n次可交换加密操作以重加密样本ID；步骤(3)中，双方各自求交集需要n次明文操作，密文拼接需要m次明文操作；步骤(4)中，双方各自解密样本特征需要mk次同态加密操作。因此总计算复杂度为$mk+2nk$次同态加密，$2n$次可交换加密和$n+m$次明文操作。由于同态加密计算复杂度较高，总复杂度可近似为$O(nk)$次同态加密。

通信复杂度：3.1节的步骤(1)和(2)中，双方均交换了长度为$nkl'+nl$比特数据，总通信数据量为$4nk(l+l')$比特。由于同态加密导致密文长度扩张，即$l'\gg l$，通信复杂度约$O(nkl')$比特。

3.3 安全性分析

ALIGN框架安全性由安全的同态加密算法和可交换加密算法以及秘密分享份额的随机性保证。本节在半诚实场景下，根据基于模拟的标准安全多方计算证明方法[22]证明ALIGN框架的安全性。

在半诚实敌手模型中，每一个参与方都诚实地执行协议，但对其他参与方的数据是好奇的。在协议执行过程中，可能对协议的中间结果以及接收到的消息进行充分的分析和推理，以获取其他参与方的隐私信息。

假设参与方P_A $\mathrm{P_A}$和P_B $\mathrm{P_B}$的输入数据分别为D_A和D_B，有函数$f=(f_A,f_B)$，通过ALIGN框架计算f。设ALIGN的输入为D_A和D_B，参与方P_A执行ALIGN的视图为

$$\mathrm{View}_{\mathrm{ALIGN},A}(D_A,D_B)=(D_A,r_A,m_1,\cdots,m_t)$$

其中r_A是协议执行过程中生成的随机数，m_i表示P_A收到的第i个消息。P_A在ALIGN执行过程中的输出可以通过$\mathrm{View}_{\mathrm{ALIGN},A}(D_A,D_B)$计算得到。相应地，$P_B$在ALIGN执行过程中的输出也可以通过$\mathrm{View}_{\mathrm{ALIGN},B}(D_A,D_B)$计算得到。基于模拟的证明方法的关键是参与方是否能从输入和输出中模拟出协议可以计算的任何信息，使模拟出来的理想视图与真实视图不可区分。具体地，如果存在概率多项式时间算法Sim_A和Sim_B分别满足

$$\begin{aligned}&(\mathrm{Sim}_A(D_A,f_A(k,D_A,D_B)),\\&f_B(k,D_A,D_B))_{k\in\mathbb{N},X,Y\in\{0,1\}^*}\\&=\{(\mathrm{View}_{\mathrm{ALIGN},A}(k,D_A,D_B),\\&\quad\mathrm{Output}_{\mathrm{ALIGN},B}(k,D_A,D_B))\}_{k\in\mathbb{N},X,Y\in\{0,1\}^*}\end{aligned}$$

和

$$\begin{aligned}&(f_A(k,D_A,D_B),\mathrm{Sim}_B(D_B,\\&f_B(k,D_A,D_B)))_{k\in\mathbb{N},X,Y\in\{0,1\}^*}\\&=\{\mathrm{Output}_{\mathrm{ALIGN},A}(k,D_A,D_B),\\&\quad\mathrm{View}_{\mathrm{ALIGN},B}(k,D_A,D_B)\}_{k\in\mathbb{N},X,Y\in\{0,1\}^*}\end{aligned}$$

其中，$k\in\mathbb{N}$为安全参数。那么当存在半诚实敌手时，协议可以安全地计算f。模拟器可以在不需要任何额外信息的前提下构造出无法区分的理想视图，那么根据ALIGN执行过程中所获得的中间结果无法推测出任何额外信息，则ALIGN是安全的。

定理1 假设参与方P_A $\mathrm{P_A}$和P_B $\mathrm{P_B}$的输入数据分别为D_A和D_B，经ALIGN框架得到的目标数据集为X'，定义数据对齐函数f为

$$f(\mathrm{ALIGN},D_A,D_B)=(\langle X'\rangle,\perp)$$

存在一个模拟器，拥有(或P_B $\mathrm{P_B}$)在方案执行过程中获取的信息，使得P_A $\mathrm{P_A}$(或P_B)的真实视图分布与通过D_A和$|D_B|$(或$|D_A|$和D_B)模拟出来的理想视图是不可区分的，即除数据规模外，参与方

P_A P$_{\rm A}$和P_B不能通过数据对齐获得关于对方数据的任何信息。

证明：不妨假设P_A为敌手(假设P_B为敌手的证明过程相同)，P_A的真实视图包含的信息有：P_A的输入D_A，P_B的样本数量n_B，特征数量k_B，P_A接收的密文X_B^1和X_A^2，以及基于这些信息的运算结果。

构造模拟器S，X_B^1的模拟过程为：(1)$\mathcal{S}$生成n_B个随机数模拟样本ID密文，记为$S([\mathrm{ID}_j]_{k_B}), j\in[n_B]$；(2)$\mathcal{S}$生成$n_B\times k_B$个随机数模拟被加密的样本特征，记为$S([X_{B,j}]_{\mathrm{pk}_B}), j\in[n_B]$；(3)将两份数据拼接起来$S([\mathrm{ID}_j]_{k_B}||[X_{B,j}]_{\mathrm{pk}_B}), j\in[n_B]$，得到结果。由于可交换加密和同态加密在选择明文攻击下安全，则模拟视图中的消息与X_B^1的计算不可区分。

$X_A{}^2$的模拟过程为：(1)$\mathcal{S}$生成n_A个随机数模拟经过双重可交换加密的样本ID密文，记为$S([\mathrm{ID}_i]_{k_A}), i\in[n_A]$；(2)$\mathcal{S}$生成$n_A\times k_A$个随机数，模拟$P_B$生成的随机数，记为$S(R_{A,i}), i\in[n_A]$；(3)$\mathcal{S}$利用$pk_A$加密秘密份额，并与经过双重加密的样本ID密文拼接，得到$S([\mathrm{ID}_i]_{k_A}||R_{A,i}), i\in[n_A]$。由于可交换加密在选择明文攻击下安全，并且秘密份额随机分割，则模拟视图中的消息与X_A^2 X$_{\rm A}^2$的计算不可区分。

综上，P_A的真实视图分布与通过D_A和D_B模拟的视图分布不可区分。　　证毕

4　实例化与实验对比

本节首先选用具体的加密算法给出ALIGN框架的实例化实现，然后将ALIGN框架与Circuit-PSI框架的对齐结果应用于机器学习模型训练并进行仿真实验对比，最后对实验结果进行了分析与讨论。

4.1　实例实现

选用基于ECDH的可交换加密算法[23]和密钥长度为1024bit的Paillier同态加密算法[24]对ALIGN框架进行实例化。采用Java语言，基于BigInteger类库实现大整数运算，基于Security类库实现安全随机数生成。实验基于公开数据集MNIST[30]，包含像素为28 × 28的图像。设参与双方各自持有样本数为256，ID交集大小为m，随机选取数据集内m张图像并进行平均纵向分割，作为双方持有样本；对任意参与方随机选取数据集内其他图像256 − m张并进行纵向平均分割，取其中1份作为ID交集外元素对应样本。表2总结了不同ID交集大小下，ALIGN框架的运行时间。

实验结果表明，运行时间随ID交集大小增长呈线性增长趋势，ID交集大小每增长50，运行时间增长约2.2 s。实验结果验证了3.2节的复杂度分析。

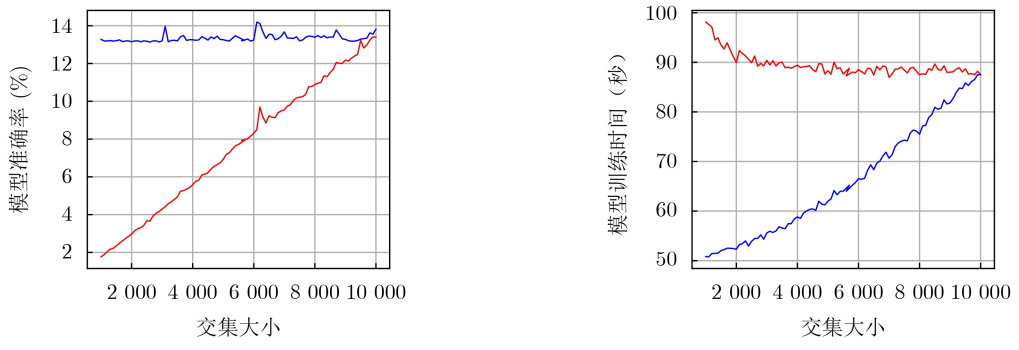
4.2　仿真实验对比

本文选取Circuit-PSI[20]作为对比，分别将4.1节数据对齐结果和Circuit-PSI的数据对齐结果应用于机器学习模型训练。选用Crypten框架[31]下标准参数的支持向量机模型，模拟纵向联邦学习在数据对齐后的模型训练过程。设定样本数为10^4，通过仿真实验评估了不同ID交集大小下模型的训练情况。图3对比了Circuit-PSI(蓝色实线)和ALIGN(红色实线)数据对齐下，模型训练的时间和模型准确率。

由仿真结果可见，随着ID交集大小增大，相同训练轮数下，训练数据为ALIGN数据对齐结果时训练时间不断增长，模型准确率均在85%以上；而训练数据为Circuit-PSI结果时，训练时间稳定在13～14 s，模型准确率不断增长。当ID交集不等于

表 2　ALIGN框架运行时间

ID交集大小	运行时间(s)
50	2.30
100	4.49
150	6.69
200	8.82
250	11.07



(a) 模型准确率　　(b) 模型训练时间

图 3 ALIGN和Ciruit-PSI对齐数据的模型训练对比

全集时，相比于Circuit-PSI, ALIGN框架数据对齐结果将训练时间缩短了1.1～11.5 s，将模型准确度提升了6.7%～47.3%。当ID交集大小越小时，ALIGN框架的优势越明显，ID交集大小每减小10^3，即每增加10%的冗余数据，ALIGN将训练时间缩短1.3 s，而模型准确度几乎不变。

4.3 实验结果分析

由表2中ALIGN框架的实例化实现结果可见，ALIGN框架的运行时间与ID交集大小有线性关系。而在基于PSI的数据对齐方案中，运行时间与ID交集的大小无关，仅与样本总量有关。这是由于ALIGN框架仅对交集内ID对应特征进行秘密分享，而在基于PSI的数据对齐方案中，则对全部ID对应样本均进行了秘密分享操作。其中仅ID交集内的特征秘密分享准确有效，对于交集外样本则将随机数进行秘密分享，这些随机数为对齐过程中生成的冗余的数据，将对后续的模型训练造成不利影响。

图3的实验结果对比了ALIGN和Circuit-PSI的数据对齐结果对于模型训练不同影响，表明了ALIGN框架下的对齐结果相较于Circuit-PSI更加有利于后续模型训练。当设定交集内ID大小不同而样本数不变时，ALIGN框架的数据对齐结果中始终不包含冗余数据，有效的数据量等于ID交集大小。因此，ID交集越大，训练时间越长但模型准确度不受影响。而Circuit-PSI的数据对齐结果中包含交集外ID对应的冗余数据，应用于训练的数据量等于样本数。因此，当ID交集递减时，Circuit-PSI的数据对齐结果数据总量不变而包含更多的冗余数据，导致模型训练时间不变但准确率不断下降。仿真实验结果说明，通过ALIGN框架进行纵向联邦学习数据对齐，有利于提升后续模型训练的效率和模型准确度。

5 结束语

本文为纵向联邦学习设计了一种新的隐私保护的数据对齐框架ALIGN，能够在保护样本ID和特征隐私性的同时实现数据对齐和特征拼接。ALIGN框架基于可交换加密和同态加密技术，实现对样本ID和特征隐私的加密和盲化，参与方在数据对齐过程后，能够得到交集内样本ID对应特征的秘密分享份额。利用基于模拟的标准安全多方计算证明方法给出了安全性证明。通过仿真实验评估了应用不同数据对齐结果进行模型训练的情况，结果表明每增加10%冗余数据，ALIGN框架可将模型训练时间缩短约1.3秒，模型准确率稳定在85%以上。本文可为纵向联邦学习数据对齐框架的设计提供一种新思路。通过计算和通信复杂度分析，计算复杂度主要来源于同态加密，通信复杂度依赖于同态密文大小。后续的工作可以结合半同态加密算法的高效实现(如Paillier加密方案的高效实现方法)以及可交换加密算法的加速实现来进一步提升ALIGN框架应用落地的可能性。

参 考 文 献

[1] YANG Qiang, LIU Yang, CHEN Tianjian, *et al*. Federated machine learning: Concept and applications[J]. *ACM Transactions on Intelligent Systems and Technology*, 2019, 10(2): 12. doi: 10.1145/3298981.

[2] 刘艺璇, 陈红, 刘宇涵, 等. 联邦学习中的隐私保护技术[J]. 软件学报, 2022, 33(3): 1057–1092. doi: 10.13328/j.cnki.jos.006446.
LIU Yixuan, CHEN Hong, LIU Yuhan, *et al*. Privacy-preserving techniques in federated learning[J]. *Journal of Software*, 2022, 33(3): 1057–1092. doi: 10.13328/j.cnki.jos.006446.

[3] LI Tian, SAHU A K, TALWALKAR A, *et al*. Federated learning: Challenges, methods, and future directions[J]. *IEEE Signal Processing Magazine*, 2020, 37(3): 50–60. doi: 10.1109/MSP.2020.2975749.

[4] KAIROUZ P, MCMAHAN H B, AVENT B, *et al*. Advances and open problems in federated learning[J]. *Foundations and Trends® in Machine Learning*, 2021, 14(1/2): 1–210. doi: 10.1561/2200000083.

[5] BELTRÁN E T M, PÉREZ M Q, SÁNCHEZ P M S, *et al*. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges[J]. *IEEE Communications Surveys & Tutorials*, 2023, 25(4): 2983–3013. doi: 10.1109/COMST.2023.3315746.

[6] 陈晋音, 李荣昌, 黄国瀚, 等. 纵向联邦学习方法及其隐私和安全综述[J]. 网络与信息安全学报, 2023, 9(2): 1–20. doi: 10.11959/j.issn.2096-109x.2023017.
CHEN Jinyin, LI Rongchang, HUANG Guohan, *et al*. Survey on vertical federated learning: Algorithm, privacy and security[J]. *Chinese Journal of Network and Information Security*, 2023, 9(2): 1–20. doi: 10.11959/j.issn.2096-109x.2023017.

[7] LIU Yang, KANG Yan, ZOU Tianyuan, *et al*. Vertical federated learning: Concepts, advances and challenges[J]. arXiv: 2211.12814, 2023. **(查阅网上资料, 未能确认文献类型, 请确认)**.

[8] ROMANINI D, HALL A J, PAPADOPOULOS P, *et al*. Pyvertical: A vertical federated learning framework for multi-headed splitNN[J]. arXiv: 2104.00489, 2021. **(查阅网上资料, 未能确认文献类型, 请确认)**.

[9] LI Qun, THAPA C, ONG L, *et al*. Vertical federated learning: Taxonomies, threats, and prospects[J]. arXiv: 2302.01550, 2023. **(查阅网上资料, 未能确认文献类型, 请确认)**.

[10] WEI Kang, LI Jun, MA Chuan, *et al*. Vertical federated learning: Challenges, methodologies and experiments[J].

arXiv: 2202.04309, 2022. **(查阅网上资料, 未能确认文献类型, 请确认)**.

[11] FREEDMAN M J, NISSIM K, and PINKAS B. Efficient private matching and set intersection[C]. Proceedings of International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, 2004: 1–19. doi: 10.1007/978-3-540-24676-3_1.

[12] PINKAS B, SCHNEIDER T, TKACHENKO O, *et al*. Efficient circuit-based PSI with linear communication[C]. Proceedings of the 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, 2019: 122–153. doi: 10.1007/978-3-030-17659-4_5.

[13] PINKAS B, SCHNEIDER T, and ZOHNER M. Scalable private set intersection based on OT extension[J]. *ACM Transactions on Privacy and Security*, 2018, 21(2): 7. doi: 10.1145/3154794.

[14] PINKAS B, ROSULEK M, TRIEU N, *et al*. PSI from PaXoS: Fast, malicious private set intersection[C]. Proceedings of the 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, 2020: 739–767. doi: 10.1007/978-3-030-45724-2_25.

[15] LU Linpeng and DING Ning. Multi-party private set intersection in vertical federated learning[C]. Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications, Guangzhou, China, 2020: 707–714. doi: 10.1109/TrustCom50675.2020.00098.

[16] HARDY S, HENECKA W, IVEY-LAW H, *et al*. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption[J]. arXiv: 1711.10677, 2017. **(查阅网上资料, 未能确认文献类型, 请确认)**.

[17] LIU Yang, ZHANG Xiong, and WAN Libin. Asymmetrical vertical federated learning[J]. arXiv: 2004.07427, 2020. **(查阅网上资料, 未能确认文献类型, 请确认)**.

[18] RINDAL P and SCHOPPMANN P. VOLE-PSI: Fast OPRF and circuit-psi from vector-ole[C]. Proceedings of the 40th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, 2021: 901–930. doi: 10.1007/978-3-030-77886-6_31.

[19] CHANDRAN N, GUPTA D, and SHAH A. Circuit-PSI with linear complexity via relaxed batch OPPRF[J]. *Proceedings on Privacy Enhancing Technologies*, 2022, 2022(1): 353–372. doi: 10.2478/POPETS-2022-0018.

[20] RAGHURAMAN S and RINDAL P. Blazing fast PSI from improved OKVS and subfield vole[C]. Proceedings of 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, USA, 2022: 2505–2517. doi: 10.1145/3548606.3560658.

[21] LIU Yang, ZHANG Bingsheng, MA Yuxiang, *et al*. iPrivJoin: An ID-private data join framework for privacy-preserving machine learning[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 4300–4312. doi: 10.1109/TIFS.2023.3288455.

[22] AGRAWAL R, EVFIMIEVSKI A, and SRIKANT R. Information sharing across private databases[C]. Proceedings of 2003 ACM SIGMOD International Conference on Management of Data, San Diego, USA, 2003: 86–97. doi: 10.1145/872757.872771.

[23] AHIRWAL R R and AHKE M. Elliptic curve diffie-hellman key exchange algorithm for securing hypertext information on wide area network[J]. *International Journal of Computer Science and Information Technologies*, 2013, 4(2): 363–368.

[24] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classes[C]. Proceedings of International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic, 1999: 223–238. doi: 10.1007/3-540-48910-X_16.

[25] SHAMIR A. How to share a secret[J]. *Communications of the ACM*, 1979, 22(11): 612–613. doi: 10.1145/359168.359176.

[26] BONAWITZ K A, EICHNER H, GRIESKAMP W, *et al*. Towards federated learning at scale: System design[C]. Proceedings of Machine Learning and Systems 2019, Stanford, USA, 2019.

[27] SHAMIR A, RIVEST R L, and ADLEMAN L M. Mental poker[M]. KLARNER D A. The Mathematical Gardner. Boston: Springer, 1981: 37–43. doi: 10.1007/978-1-4684-6686-7_5.

[28] POHLIG S and HELLMAN M. An improved algorithm for computing logarithms overGF(p)and its cryptographic significance (Corresp.)[J]. *IEEE Transactions on information Theory*, 1978, 24(1): 106–110. doi: 10.1109/TIT.1978.1055817.

[29] DIFFIE W and HELLMAN M. New directions in cryptography[J]. *IEEE Transactions on Information Theory*, 1976, 22(6): 644–654. doi: 10.1109/TIT.1976.1055638.

[30] LECUN Y. The MNIST database of handwritten digits[EB/OL]. http://yann.lecun.com/exdb/mnist/, 1998.

[31] BARNES R. CrypTen[EB/OL]. https://github.com/facebookresearch/CrypTen, 2020. **(查阅网上资料,未找到作者信息,请确认)**.

高　莹：女，副教授，博士生导师，研究方向为密码学、隐私计算和区块链.

谢雨欣：女，博士生，研究方向为联邦学习.

邓煌昊：男，博士生，研究方向为联邦学习.

朱祖坤：男，硕士生，研究方向为联邦学习.

张一余：男，硕士，研究方向为纵向联邦学习.

责任编辑：马秀强