基于安全两方计算的隐私保护线性回归算法

魏立斐,李梦思,张 蕾,陈聪聪,陈玉娇,王 勤
上海海洋大学 信息学院,上海 201306

摘 要:随着数据安全与隐私泄露事件频发,泄露规模连年加剧,如何保证机器学习中数据和模型参数的隐私引发科学界和工业界的广泛关注。针对本地存储计算资源的有限性及云平台的不可信性所带来的数据隐私问题,基于秘密共享技术提出了一种安全两方计算的隐私保护线性回归算法。利用加法同态加密和加法掩码实现了秘密共享值的乘法计算协议,结合小批量梯度下降算法,最终实现了在两个非共谋的云服务器上的安全线性回归算法。实验结果表明,该方案同时保护了线性回归算法训练及预测阶段中的数据及模型参数,且模型预测性能与在明文域中进行训练的模型相近。

关键词:线性回归;安全两方计算;秘密共享;加法同态加密;隐私保护
文献标志码:A **中图分类号:**TP309 doi:10.3778/j.issn.1002-8331.2007-0337

Privacy-Preserving Linear Regression Algorithm Based on Secure Two-Party Computation

WEI Lifei, LI Mengsi, ZHANG Lei, CHEN Congcong, CHEN Yujiao, WANG Qin
College of Information, Shanghai Ocean University, Shanghai 201306, China

Abstract: With the frequent occurrence of data security and privacy leaks and the increasing scale of leaks year after year, how to ensure the privacy of data and model parameters in machine learning has aroused widespread concern in the scientific and industrial community. Aiming at the data privacy problem caused by the limitation of local storage and computing resources and the untrustworthiness of cloud platforms, this paper proposes a privacy protection linear regression algorithm for secure two-party computation based on secret sharing technology. This paper uses additive homomorphic encryption and additive masks to realize a multiplication calculation protocol of the secret shared values, and combines the mini-batch gradient descent algorithm to realize the secure linear regression algorithm on two non-collusive cloud servers. The experimental results show that the scheme protects the data and model parameters in the training and prediction phases of the linear regression algorithm at the same time, and the model prediction performance is similar to the model trained in the plaintext domain.

Key words: linear regression; secure two-party computation; secret sharing; addition homomorphic encryption; privacy protection

线性回归(Linear Regression)作为一种广泛使用的基础型机器学习算法,是通过对多个影响因素和结果进行拟合,从而以线性模型来建模一个或多个自变量与因变量之间相关关系的一种方法。为了提高和优化回归模型性能,通常需要训练大量的原始数据,但本地计算资源及存储资源有限,使得一些企业或机构无法满足独自训练模型的需求。而云计算[1]的迅速发展,使得这一问题得到很好地解决,目前有许多云计算服务商业平台(如亚马逊和谷歌等)允许客户端上传数据到云服务器进行各种机器学习任务。但因为云计算的不可信性[2],它可以查看并记录用户的数据信息,甚至可能遭受到敌手的攻击而泄露用户数据,所以研究能够保护隐私的线性回归方案尤为重要。目前已经有许多学者提出了具有隐私保护的线性回归方案。基于差分隐私的方法[3-5]

基金项目:国家自然科学基金(61972241,61802248,61672339);上海市自然科学基金面上项目(18ZR1417300)。

作者简介:魏立斐(1982—),男,博士,副教授,硕士生导师,CCF高级会员,主要研究领域为信息安全和密码学;李梦思(1994—),女,硕士研究生,CCF学生会员,主要研究领域为机器学习和信息安全;张蕾(1983—),通信作者,女,博士,讲师,主要研究领域为应用密码学、大数据安全性和访问控制,E-mail:lzhang@shou.edu.cn;陈聪聪(1996—),男,硕士研究生,CCF学生会员,主要研究领域为机器学习、信息安全和安全计算;陈玉娇(1996—),女,硕士研究生,CCF学生会员,主要研究领域为机器学习和信息安全;王勤(1996—),男,硕士研究生,CCF学生会员,主要研究领域为信息安全和安全计算。

收稿日期:2020-07-20 **修回日期:**2020-11-03 **文章编号:**1002-8331(2021)22-0139-08

是通过对回归模型添加适当噪声的方式来实现隐私保护,但引入噪声的同时会导致模型性能有所下降。基于同态加密(Homomorphic Encryption,HE)的方案[6-8]则通常需要客户端使用同态加密算法对训练数据加密之后,由云服务器利用同态性质在密文上进行训练,但使用同态加密算法对大量的数据实现加密对于客户端来说计算开销太大,并且由于同态加密计算本身的限制,无法实现任意次的加法和乘法,在现实环境中并不实用。

安全多方计算(Secure Multi-party Computation,SMC)起源于1982年Yao[9]提出的百万富翁问题,核心思想是在不泄露任何参与方的私有输入情况下正确地计算目标函数。作为安全计算领域里的核心内容之一,SMC是构造多方计算协议的基础,可用于解决现实世界中的实际问题,但它需要借助现代密码学中的其他技术来实现[10-12],比如基于秘密共享(Secret Sharing)、混淆电路(Garbled Circuits,GC)、同态加密等技术的安全多方计算线性回归方案[13-15]。2011年,Hall等人[16]基于同态加密首次提出了一种可以达到安全性定义的安全两方计算线性回归协议,但该方案过于依赖计算开销巨大的同态加密,无法应用到数据条目庞大的数据集中。Martine等人[17]基于文献[16]在数据集分布于多个参与方的情境下,提出了一种能够保护数据隐私的线性回归方案,各计算方可以在不共享自己私有数据集的情况下协同训练线性回归模型。Dankar[18]通过引入一个半可信第三方,在理论上提出了一种支持多个数据提供者参与的隐私保护线性回归方案。Adrià等人[19]提出一种用于任意分布于多个参与方的训练集的隐私保护线性回归方案,该方案结合了Yao的混淆电路和全同态加密方案。之后,Mohassel等人[20]提出的SecureML方案基于混淆电路和不经意传输(Oblivious Transfer,OT)[21]协议,设计了支持安全两方计算的随机梯度下降算法,实现了线性回归、逻辑回归以及神经网络的模型训练任务。该方案由数据拥有者将私有数据通过秘密共享的方式分发给两个服务器,由两个服务器用安全多方计算的方式训练模型,实现了加法和乘法的分布式计算。在SecureML的基础上,唐春明等人[22]借助基于OT协议生成的乘法三元组[23],提出了具有隐私性的回归模型训练算法,同时实现了对训练数据及模型参数的隐私保护。Akavia等人[24]提出一种能够从多个数据所有者提供的数据集中学习线性回归模型的数据隐私保护方案,该方案使用两个非共谋服务器和线性同态加密(Linearly Homomorphic Encryption)来学习正则化线性回归模型。Dong等人[25]提出了一个可以适应半诚实和恶意环境下的分布式机器学习框架,每个参与者将自己的梯度分成共享份额,并分配给多个参数服务器,由参数服务器聚合梯度后发还给参与者,参与者在本地更新参数。

受安全多方计算核心思想的启发,本文拟采用安全两方计算技术来解决线性回归方案中的隐私保护问题。在此之前,Mohassel[20]和唐春明[22]提出的基于安全两方计算的隐私保护方案,由两个非共谋云服务器协作完成线性回归任务,但他们均使用通信复杂度较高的OT协议,因此在规模比较大的数据集上使用会有一定的局限性。另外,文献[20]中的线性回归协议虽然解决了数据隐私保护问题,但需要两个云服务器直接重构模型参数,因此该方案无法保证模型参数的隐私性。不同于文献[20]和文献[22],本文避免使用通信复杂度较高的OT协议,而是通过使用加法同态加密和加法掩码相结合的方法实现秘密共享值的乘法计算,避免两方服务器私有信息的泄露。相比之下,本文方案在保证数据和模型参数隐私不被泄露的同时,所需要的通信开销更低。本文主要贡献包括以下两方面:

(1)使用安全两方计算的方式执行小批量梯度下降算法更新模型参数,通过将加法同态加密与加法掩码相结合的方法,实现了秘密共享值之间的乘法计算。

(2)提出并实现了隐私保护的预测方案,确保云服务器在预测过程中无法获得预测数据的具体信息,同时在预测结束后无法获得真正的预测结果,实现了隐私保护线性回归预测。

1 预备知识

1.1 线性回归

给定一个包含 n 条数据的训练集 $(\boldsymbol{X},\boldsymbol{y})$,其中 $\boldsymbol{X}\in\mathrm{R}^{n\times d}$ 表示具有 d 个特征的样本集特征矩阵,$\boldsymbol{y}\in\mathrm{R}^{n}$ 表示 n 条数据样本对应的标签向量,线性回归任务的目标是从训练集 $(\boldsymbol{X},\boldsymbol{y})$ 学习模型 M 的一组回归系数 $\theta\in\mathrm{R}^{d}$,使得目标值 $\boldsymbol{y}\approx\boldsymbol{X}\theta$。

为了衡量模型的好坏,需要对训练出的模型进行性能评价,一个常用的评价标准是平方误差和,即目标值和预测结果之间差距的平方和,因此可以定量化损失函数:

$$L(\theta)=\frac{1}{2n}\sum_{i=1}^{n}(x_i\theta-y_i)^2 \tag{1}$$

线性回归的任务就是寻求使得 L 最小化时的 θ 值。

梯度下降是一个用来求目标函数最小值的优化算法,本文使用小批量梯度下降算法(Mini-Batch Gradient Descent,MBGD)来求出损失函数 $L(\theta)$ 的最小值。它在更新每一参数时都使用一部分样本进行更新,相比较随机梯度下降算法(Stochastic Gradient Descent,SGD)和批量梯度下降算法(Batch Gradient Descent,BGD),该算法可以缩减模型收敛所需要的迭代次数,同时使收敛的结果更接近梯度下降的效果。对于小批量数据集 $(\boldsymbol{X}^B,\boldsymbol{y}^B)$ 的梯度下降算法的参数更新方式为:

$$\theta^{e+1}=\theta^{e}-\frac{\alpha}{|B|}(\boldsymbol{X}^B)^{\mathrm{T}}\times(\boldsymbol{X}^B\times\theta^{e}-\boldsymbol{y}^B) \tag{2}$$

其中,$\boldsymbol{X}^B$ 和 $\boldsymbol{y}^B$ 分别表示小批量样本集的特征值和目标值,e 表示当前迭代次数,α 表示学习率,$|B|$ 表示小批量样本数量。

1.2　加法同态加密

同态加密最早是由Rivest等人[26]提出,这种加密方法允许直接在密文上进行某些特殊类型的计算而获得密文结果,并且将密文结果解密后,其值与在明文上执行的函数结果一致。本文使用支持加法同态加密的Paillier加密系统[27],它是由Paillier于1999年基于复合剩余类困难问题建立的概率公钥加密系统。该加密系统工作原理如下:

(1)密钥生成 $KeyGen(\cdot)\rightarrow(pk,sk)$:随机生成两个大素数 p、q 满足 $\gcd(pq,(p-1)(q-1))=1$,计算 $N=pq$,$\lambda=\mathrm{lcm}(p-1,q-1)$,随机选择整数 $g\in Z_{N^2}^*$,则公钥 $pk=(g,N)$,私钥 $sk=\lambda$。

(2)加密过程 $Enc(pk,m)\rightarrow c$:选择一个随机数 $r\in Z_N^*$,则明文 m 对应的密文 $c=g^m r^N \bmod N^2$。

(3)解密过程 $Dec(sk,c)\rightarrow m$:对于密文 c 解密后的明文 $m=(L(c^\lambda \bmod N^2)/L(g^\lambda \bmod N^2))\bmod N$,其中 $L(u)=(u-1)/N$。

对于明文 m 有:

$$(Enc(m))^k=\underbrace{Enc(m)\times Enc(m)\times\cdots\times Enc(m)}_{k个}=$$
$$(g^m r_1^N \bmod N^2)\times(g^m r_2^N \bmod N^2)\times\cdots\times$$
$$(g^m r_k^N \bmod N^2)=$$
$$g^{km}(r_1 r_2\cdots r_k)^N \bmod N^2=Enc(km)$$

本文Paillier同态加密利用python-paillier库(https://python-paillier.readthedocs.io/)实现,该加密库支持浮点数的计算,因此对于实数 k 和明文 m 有 $(Enc(m))^k=Enc(km)$。

1.3　秘密共享

秘密共享就是指共享的秘密在多个计算方之间进行合理分配,以达到由所有参与方共同掌管秘密的目的。Sharmir在1979年最早提出 t-out-of-n 秘密共享方案[28],允许将秘密 s 进行分割并在 n 个参与者中共享,使得至少任意 t 个参与者合作才能够还原秘密,而任何少于 t 个参与者均不可以得到秘密的任何信息。具体地,该方案由两种算法组成:共享算法 $Share(\cdot)$ 和重构算法 $Recon(\cdot)$,算法描述如下:

(1) $Share(s,t,n)\rightarrow(s_1,s_2,\cdots,s_n)$:给定秘密 s、阈值 t 以及共享份额数 n,可以产生一组秘密共享值 $\{s_1,s_2,\cdots,s_n\}$。

(2) $Recon(\Theta,t)\rightarrow s$:给定秘密共享值的子集 Θ,其中 $\Theta\in\{s_1,s_2,\cdots,s_n\}$ 且 $|\Theta|\geqslant t$,则可以重构出原始秘密 s。

本文方案涉及两方计算任务,由两个云服务器进行交互式协作计算,因此本文采用2-out-of-2秘密共享方案。即对于秘密 a,通过共享算法 $Share(a,2,2)\rightarrow(a_0,a_1)$ 得到其对应的两个共享值 $a_i(i=0,1)$;反之,由秘密共享值 $a_i(i=0,1)$ 恢复出原始秘密 a 的过程就叫作 $Recon(\{a_0,a_1\},2)\rightarrow a$。其中 $a=a_0+a_1$。

2　系统模型

本文采用诚实且好奇的非共谋双云服务器模型,即云服务器诚实地执行预置的计算任务,同时出于好奇会查看并记录数据信息,但不会向另一方透露任何自己的输入、中间计算参数以及输出信息。如图1所示,本文系统模型包含一个数据提供者(Data Provider)、一个用户(User)和两个云服务器(Cloud Server,CS)。数据拥有者发布线性回归模型的训练任务并提供必要的训练数据,在与云服务器 $CS_i(i=0,1)$ 建立基于TSL/SSL协议的安全信道并进行模型训练任务协商之后,利用秘密共享原理将训练数据分发给它们,由两个云服务器协作完成训练任务;在预测阶段,具有预测请求的用户在与服务器建立安全信道之后,将待预测数据通过秘密共享的方式分发给它们,两个云服务器进行协作预测,并将各自的预测值返还给用户,由用户重构出最终的预测结果。

3　本文方案

本章主要描述基于安全两方计算的隐私保护线性回归方案,包括秘密共享值的乘法计算以及基于两方计算的线性回归安全训练和预测阶段。其中秘密共享值的乘法计算(Multiplication of Secret Shared Values,MoSSV)协议作为训练及预测阶段的基础协议,主要用于双云服务器的安全两方计算。如图2所示,训练阶段和预测阶段分别由三个主要模块构成。在训练阶段,鉴于数据需要脱离数据提供者本地,但又不能向云服务器泄露任何

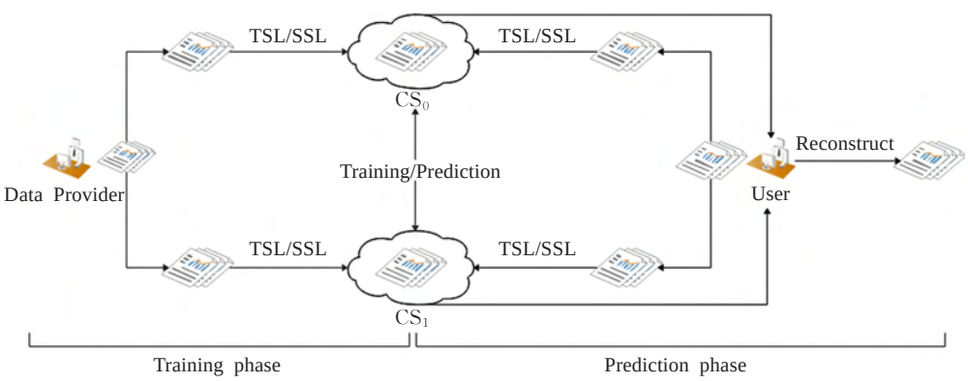


图1　系统模型

Fig.1　System model

数据信息,因此首先通过 *Share*(·) 算法将数据以秘密共享的方式进行划分;收到秘密共享数据之后,两个云服务器共同协商并预置相同的训练参数,即学习率、小批量样本数目、最大迭代次数及损失阈值;最后由两个云服务器执行安全的小批量梯度下降算法进行模型参数更新,直至模型收敛。在预测阶段,同样出于保护预测数据的隐私,首先将预测数据通过秘密共享 *Share*(·) 算法分发给双云服务器;之后由双云服务器执行 *CalPred*(·) 模块,得到预测结果的秘密共享值;最后由用户执行 *Recon*(·) 算法重构最终的预测结果。

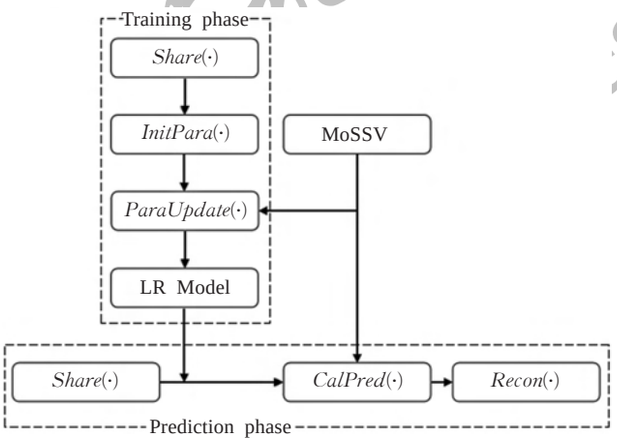


图2　隐私保护线性回归方案框架图

Fig.2　Framework of privacy protection linear regression

3.1　秘密共享值的乘法计算协议

假设两个计算方分别持有给定矩阵和向量的秘密共享值,那么如何在保证计算方各自的秘密共享值不被泄露的情况下,安全地完成秘密共享值之间的乘法运算呢?Du 等人[29]在仅有两个计算方参与的情形下,基于OT协议提出了一系列解决分布式线性代数问题的方案。由于OT协议通信复杂度较高,陈莉等人[30]基于同态加密的性质设计了用于求解分布式线性方程组问题的安全两方计算协议。基于文献[30],本文利用加法同态加密和加法掩码实现了适用于线性回归任务中秘密共享值的乘法计算协议(MoSSV),其核心思想是在仅有两个计算方参与的情况下,利用加法同态加密保护其中一个计算方的私有信息,利用加法掩码掩盖另一计算方的私有信息,最终计算双方获得矩阵-向量乘积的秘密共享值。下面给出了协议执行过程的详细描述,协议执行过程如图3所示。

3.1.1　问题描述

秘密共享值的乘法计算问题可以描述为:对于给定的矩阵 M 和向量 v(其中 M 的第二维度与 v 的第一维度一致),M_i 和 $v_i\,(i=0,1)$ 是它们的秘密共享份额且分别为计算方 $P_i(i=0,1)$ 所拥有,其中 $M=M_0+M_1$,$v=v_0+v_1$,即计算方 P_0 拥有私有矩阵 M_0 和私有向量 v_0,另一计算方 P_1 拥有私有矩阵 M_1 和私有向量 v_1。执行该协议之后,$P_i(i=0,1)$ 可以获得乘积 Mv 的秘密共享份额 $p_i=Multi(M_0,M_1,v_0,v_1)$。

3.1.2　秘密共享值的乘法计算协议

输入:P_0 的私有矩阵 M_0 和私有向量 v_0,P_1 的私有矩阵 M_1 和私有向量 v_1。

输出:$P_i(i=0,1)$ 可得到 $p_i=Multi(M_0,M_1,v_0,v_1)$。

协议过程描述:

步骤1　P_i 各自生成同态加密密钥对 (pk_i,sk_i),其中 pk_i 和 sk_i 分别表示 P_i 的公钥和私钥,并将公钥 pk_i 发给对方。

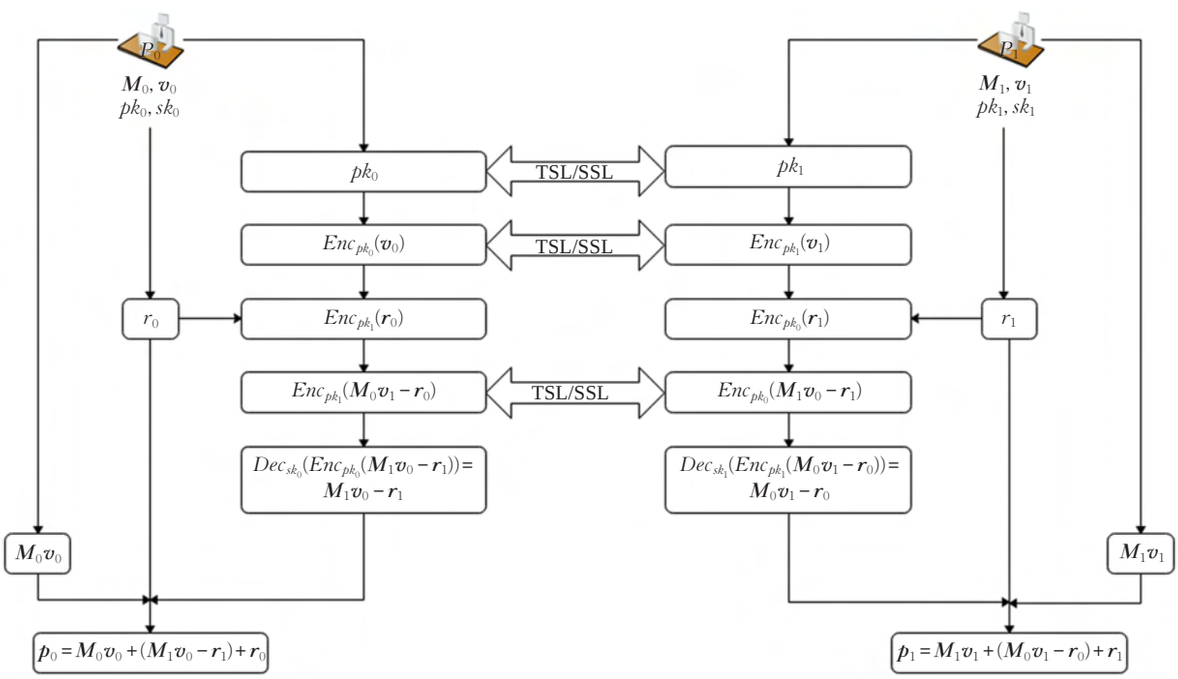


图3　秘密共享值的乘法计算协议流程图

Fig.3　Flowchart of multiplication calculation protocol for secret shared values

步骤2　P_i 使用自己的公钥 pk_i 加密私有向量 v_i 并将 $Enc_{pk_i}(v_i)$ 发给对方。

步骤3　P_{1-i} 收到对方公钥 pk_i 和加密向量 $Enc_{pk_i}(v_i)$ 后,随机生成向量 r_{1-i},并使用对方的公钥 pk_i 加密得到 $Enc_{pk_i}(r_{1-i})$。

步骤4　P_{1-i} 计算 $Enc_{pk_i}(M_{1-i}v_i-r_{1-i})$ 并将结果发给对方。

步骤5　P_i 收到 $Enc_{pk_i}(M_{1-i}v_i-r_{1-i})$ 后,使用自己的私钥 sk_i 解密得到 $M_{1-i}v_i-r_{1-i}$。

步骤6　P_i 计算得到 $p_i=M_iv_i+(M_{1-i}v_i-r_{1-i})+r_i$,协议结束。

3.1.3　正确性

对于任意两个秘密 a、b,要求在避免使用可信第三方且不泄露 a、b 值的情况下求 $c=a+b$。通过随机数 r,可以进行以下构造:$a'=a-r$,$b'=b+r$。因为随机数可以相抵消,所以有 $c=a'+b'=(a-r)+(b+r)=a+b$。

根据以上原理,对于MoSSV协议,有:

$$\begin{aligned}p_0+p_1=&[M_0v_0+(M_1v_0-r_1)+r_0]+\\&[M_1v_1+(M_0v_1-r_0)+r_1]=\\&M_0v_0+[r_0+(M_0v_1-r_0)]+\\&[(M_1v_0-r_1)+r_1]+M_1v_1=\\&M_0v_0+M_0v_1+M_1v_0+M_1v_1=\\&(M_0+M_1)(v_0+v_1)=Mv\end{aligned}$$

因此该协议是正确的。

3.1.4　安全性

在MoSSV协议执行之前,计算方 $P_i(i=0,1)$ 之间通过协商建立基于TSL/SSL协议的安全通道,以确保他们之间发送的任何敏感数据的安全性及完整性。该协议利用加法同态加密和加法掩码的性质保护计算双方的私有信息,其安全性主要体现在协议计算过程步骤2~步骤4中。在步骤2中计算方 P_i 利用加法同态加密技术,使用己方公钥加密私有向量并发送给对方,因为对方不知道密文对应的私钥,所以无法解密,从而起到保护私有向量隐私性的作用;在步骤4中,对方不能直接将加密的矩阵向量乘积发送回来,而是在步骤3利用加法掩码的原理使用随机向量将乘积进行盲化,从而起到保护对方私有矩阵信息的作用。

综上,在协议执行过程中,$P_i(i=0,1)$ 可以获得的信息如表1所示。

表1　在MoSSV协议步骤中参与方所获得的信息

Table 1　Information obtained by each participant in MoSSV protocol

阶段	P_0	P_1
步骤1	pk_1	pk_0
步骤2	$Enc_{pk_1}(v_1)$	$Enc_{pk_0}(v_0)$
步骤3	$Enc_{pk_1}(r_0)$	$Enc_{pk_0}(r_1)$
步骤4	$Enc_{pk_0}(M_1v_0-r_1)$	$Enc_{pk_1}(M_0v_1-r_0)$
步骤5	$M_1v_0-r_1$	$M_0v_1-r_0$
步骤6	p_0	p_1

3.2　训练阶段

(1) $Share((X,y),2,2)\rightarrow((X_0,y_0),(X_1,y_1))$

数据提供者将私有训练数据 (X,y),利用2-out-of-2秘密共享的原理随机拆分为与原始数据维度大小相同的两部分子数据 (X_0,y_0) 和 (X_1,y_1),并通过基于TSL/SSL协议建立的安全信道分发给云服务器 CS_0 和 CS_1。

(2) $InitPara(CS_0,CS_1)\rightarrow(\theta_0,\theta_1,\alpha,|B|,E)$

由于在线性回归模型训练之前,参与训练模型的计算方需要共同预置一些必要的参数,以高效准确地完成回归任务。因此 $CS_i(i=0,1)$ 首先共同协商学习率 α、小批量样本数目 $|B|$、最大迭代次数 E,并分别初始化模型参数 $\theta_i\in\mathrm{R}^d$(全0/1向量或者任意随机数)。

(3) $ParaUpdate(\cdot)\rightarrow(\theta_0,\theta_1)$

为了优化模型收敛速度,云服务器 $CS_i(i=0,1)$ 之间使用安全的小批量梯度下降(MBGD)算法,根据式(2)更新模型参数。具体子步骤如下所示:

① CS_i 分别选取索引 B 相匹配的小批量样本数据 (X_i^B,y_i^B);

② CS_i 利用双方的批量样本数据 (X_i^B,y_i^B) 和模型参数 θ_i,使用MoSSV协议分别得到秘密共享值 $\hat{y}_i^B=Multi(X_0^B,X_1^B,\theta_0,\theta_1)$;

③ CS_i 计算 $\hat{y}_i^B$ 与真实秘密共享值 y_i^B 之间的误差 $err_i^B=\hat{y}_i^B-y_i^B$,并将误差向量 err_i^B 发给对方 CS_{1-i};

④ CS_i 重构误差向量 $err^B=err_0^B+err_1^B$;

⑤ CS_i 计算梯度变化共享值 $G_i^B=(X_i^B)^{\mathrm{T}}\cdot err^B$;

⑥ CS_i 根据式(2) $\theta_i^{e+1}=\theta_i^e-\dfrac{\alpha}{|B|}G_i^B$ 更新本地模型参数 θ_i。

之后,CS_i 根据当前更新的模型参数 θ_i,利用MoSSV协议计算损失函数值 L_i 并发给对方,两个云服务器根据 $Recon(\cdot)$ 重构出整个训练集的损失值,并判断模型是否收敛,若收敛,训练结束,当前 $CS_i(i=0,1)$ 所拥有的参数 θ_i 为线性回归模型参数的秘密共享值;否则,以上子步骤会循环执行。当模型或者训练达到最大迭代次数 E 时,强行停止训练。

3.3　预测阶段

(1) $Share(X^U,2,2)\rightarrow(X_0^U,X_1^U)$

已知云服务器 CS_0 和 CS_1 分别拥有模型参数的秘密共享值 θ_0 和 θ_1,具有预测任务的用户User可以利用云服务器强大的计算力进行线性预测。首先将预测数据集 X^U 进行拆分预处理,得到两个子数据集 X_0^U 和 X_1^U 并分别发送给云服务器 CS_0 和 CS_1。

(2) $CalPred(X_0^U,X_1^U,\theta_0,\theta_1)\rightarrow(\hat{y}_0^U,\hat{y}_1^U)$

$CS_i(i=0,1)$ 利用MoSSV协议得出预测结果的秘密共享值 $\hat{y}_i^U=Multi(X_0^U,X_1^U,\theta_0,\theta_1)$。

(3) $Recon(\hat{y}_0^U,\hat{y}_1^U)\rightarrow\hat{y}^U$

$CS_i(i=0,1)$ 分别将秘密共享结果 $\hat{y}_i^U$ 发送给用户,由用户根据秘密共享值重构出真实的预测结果 $\hat{y}^U$。

4 性能评估

4.1 安全分析

本文方案实现了数据及模型参数的隐私保护,如表2所示。本文的线性回归任务涉及两个计算方安全地执行小批量梯度下降算法,即由两个云服务器进行交互式协作计算,但因为云服务器是诚实且好奇的,对训练数据的安全存在一定的威胁,所以方案首先利用加法秘密共享将训练用的数据以适当形式拆分后分发给不同计算方。本方案中两个云服务器 CS_0 和 CS_1 非共谋,因此有效地避免了云服务器恢复原始数据信息的问题,实现了对训练数据的隐私保护。

表2 基于两方计算的线性回归方案对比

Table 2 Comparison of linear regression schemes based on two-party computation

方案	技术	数据隐私	模型参数隐私	通信成本
文献[20]	乘法三元组,OT,GC	√	×	高
文献[22]	乘法三元组,OT	√	√	高
本文	HE,加法掩码	√	√	低

在参数更新过程中,涉及到两方秘密共享值需要同时使用的计算操作,比如 $\hat{y}_i^B$,需要在保护各自秘密共享数据以及模型参数的情况下进行安全计算。本文利用安全的MoSSV协议,使用同态加密对各自模型参数 θ_i 进行加密处理,根据加法掩码的原理使用随机向量掩盖秘密共享数据 X_i^B 的信息,防止对方使用私钥解密获得私有数据信息,根据表1可知,云服务器无法获取到任何关于对方的隐私信息。判断模型收敛时,云服务器无法根据损失函数值 L_i 恢复出对方的预测结果 $\hat{y}_i$ 以及真实标签 y_i 。预测阶段仅涉及到秘密共享数据和模型参数,其安全性分析同理。

值得注意的是,在训练阶段结束之后,若数据提供者有模型使用需求,则可以直接向云服务器请求发还模型参数的秘密共享值,并在本地使用 $Recon(\cdot)$ 重构出模型参数。因为数据提供者本身是具有模型训练任务的,所以该操作并不会涉及到模型参数私有信息的泄露问题。

4.2 性能分析

与本文最接近的方案是文献[20]和文献[22],均属于基于安全多方计算的隐私保护方案,但它们均使用通信复杂度较高的OT协议,方案通信成本过高,因此本文避免采用OT协议,而是使用加法同态加密和加法掩码技术,对比如表2所示。传统使用同态加密的隐私保护方案的一般做法是使用同态加密技术将原始训练数据加密处理后以密文的形式进行训练,最后将模型同态解密。这种方法虽然可以实现数据及模型的隐私保护,但是在密文数据上的训练会使得通信和计算开销呈指数级增长,因此为了平衡计算通信开销与隐私保护之间的矛盾,本文不论是在隐私保护线性回归算法的训练阶段还是预测阶段,均首先引入加法秘密共享技术,使用 $Share(\cdot)$ 算法将原始数据转换为非敏感型数据,并将拆分后的秘密共享数据直接以明文的形式分发给云服务器 CS_0 和 CS_1 ,即通过明文传输的方法保护原始数据,而不是直接将数据进行加密处理,有效避免了用户与服务器之间的密文传输,从而既保护了原始数据的隐私,又极大降低了数据提供者(或用户)与服务器之间的通信开销。之后在模型训练 $ParaUpdate(\cdot)$ 和预测结果计算 $CalPred(\cdot)$ 模块中,借助同态加密可以对密文直接进行处理的特性,结合加法掩码技术,利用加法秘密共享值之间的加法和乘法计算特性,将模型参数更新公式进行分解,使用MoSSV协议保护计算双方的私有信息。相较于传统的同态加密方案,这种方法不需要在双云服务器之间进行多维度密文数据的传送,只需要在每一轮迭代过程发送单一维度的密文向量即可,从而不仅保护了数据和模型参数的隐私,而且大幅度降低了双云服务器之间的通信开销。

由于在不同的数据集上模型收敛的速度不同,本文仅针对一轮迭代训练过程中的时间及通信开销进行方案性能分析。在小批量梯度下降算法中,小批量样本数目 $|B|$ 以及学习率 α 的选择是很重要的, $|B|$ 太大或者太小都会导致训练时间过长。经过大量的实验验证,最终本文将小批量样本数目 $|B|$ 、最大迭代次数 E 及学习率 α 分别设置为10、100和0.1,并设置模型收敛条件(即损失阈值)为 10^{-5} 。在本文的方案中,通信开销主要来自于 $ParaUpdate(\cdot)$ 阶段中的MoSSV协议。 ct 表示一条密文大小, pt 表示一条明文大小。那么在计算 $\hat{y}_i^B$ 时, $CS_i(i=0,1)$ 需要分别将向量 θ_i 和盲化的乘积向量 $X_i^B\theta_{1-i}-r_i$ 的密文形式发给对方,因此通信开销为 $(d+|B|)\times ct$;对于 err_i^B 的发送,通信开销为 $|B|\times pt$ 。当判断模型是否收敛时,需要计算训练集上的损失函数,计算 $\hat{y}_i$ 的过程需要 $(n+d)\times ct$ 的通信成本,另外 $CS_i(i=0,1)$ 将损失函数值 L_i 发送给对方需要 $n\times pt$ 。综合以上,一轮迭代训练过程的总通信开销为 $2(\lceil n/|B|\rceil d+n)\times ct+4n\times pt$ 。文献[20]和文献[22]的通信成本主要出现在小批量梯度下降算法的执行以及OT协议中。其中文献[20]执行SGD_Linear协议时, $CS_i(i=0,1)$ 在每轮迭代过程中发送盲化之后的权重及误差向量所需要的通信成本为 $(|B|+d)\times pt$,而使用OT协议计算乘法三元组时通信量为 $|B|dl\times pt$,则每轮迭代过程所需要的总通信成本为 $2\lceil n/|B|\rceil(|B|+d+|B|dl)\times pt$ 。在文献[22]的线性回归算法迭代训练过程中, $CS_i(i=0,1)$ 计算预测值和梯度变化量时的通信成本均为 $(|B|d+|B|+d)\times pt$,而双云服务器使用OT协议计算乘法三元组需要通信成本为 $|B|dl\times pt$,故每次迭代总通信成本为 $2\lceil n/|B|\rceil(|B|d+|B|+d+|B|dl)\times pt$,其中 $\lceil a\rceil$ 表示比 a 大的最小整数, l 表示数据长度。

表3 明文和密文下的实验结果对比

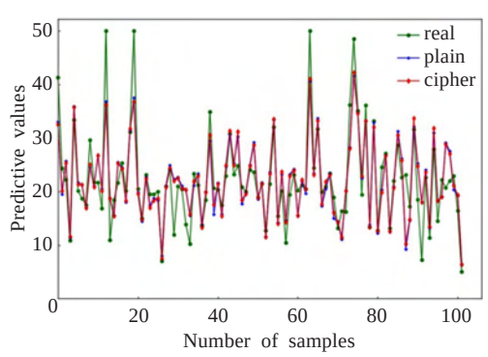
Table 3 Comparison of implementation results between plaintext and ciphertext

Dataset	Number of samples n	Number of features d	Data state	Number of iterations	Average training time/ms	Results			
						MAE	MSE	RMSE	R-Square
Boston	506	13	Plain	12	1.50	3.050	18.255	4.273	0.754
			Cipher	11	157 912.00	2.969	17.652	4.201	0.763
Diabetes	442	10	Plain	19	1.37	40.293	2 518.251	50.182	0.402
			Cipher	27	128 340.00	40.949	2 568.132	50.677	0.390

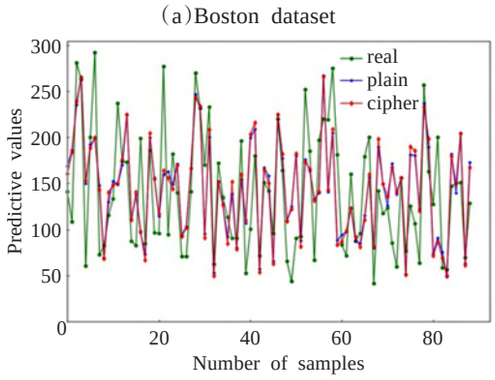
4.3 实验结果

为了证明本文方案的有效性,对基于安全两方计算的数据隐私保护线性回归算法进行了实验验证。实验平台配置为Intel® Core™ i5-4200M、2.50 GHz、8 GB内存的计算机,使用Python语言进行编程,通过两个类分别模拟数据提供者和云服务器的行为。实验数据选用Python的Scikit-learn库提供的Boston数据集和Diabetes数据集。其中Boston数据集涉及美国人口普查局收集的美国马萨诸塞州波士顿住房价格的有关信息,包含506条样本数据,每条数据包含有13个输入变量和1个输出变量。Diabetes数据集包含404条医疗记录,每条记录有10个输入变量和1个输出变量。本文随机选取数据集的80%用于训练模型,剩余的20%用于测试模型性能,并对其进行归一化处理。

图4展示了预测数据集真实标签值与明文域及密文域的预测值的对比曲线图。其中明文状态实验结果是指由客户端本地独自训练模型的情形,曲线图证明密文下的预测结果几乎与明文下的预测结果一致。同时,本文以均方误差(Mean-Square Error,MSE)、均方根误差(Root-Mean-Square Error,RMSE)、平均绝对误差(Mean Absolute Error,MAE)和R平方(R-Squared)作为线性回归模型的评估指标。与明文下的结果相比,本文方案几乎实现了相同的预测性能,这表明本文基于两方计算的保护数据隐私线性回归方案是可行的。如表3所示,Boston数据集完成一轮训练平均需要157.912 s,Diabetes数据集完成一轮训练需要128.340 s。虽然密文域的训练速度与明文相比慢,但是训练一次得到的模型可以用于多次预测,因此针对注重隐私性的医疗、基因、财务等数据而言,密文域的训练是可以接受的。



(a)Boston dataset



(b)Diabetes dataset

图4 明文域及密文域预测结果与真实标签值的对比图

Fig.4 Comparison of real and predicted label values in plaintext and ciphertext

5 结束语

本文提出了一种基于安全两方计算的数据隐私保护线性回归方案。为了在两方计算过程中不泄露数据、模型参数及中间参数的信息,本文利用加法同态加密和加法掩码保护两个云服务器的秘密共享值,实现了秘密共享值的乘法计算协议MoSSV。实验结果表明,本文方案在保证模型准确度的情况下,实现了数据及模型参数的隐私保护。本文提出的方案保证了训练和预测过程的高效性,并达到了较高的准确度。对于下一步的工作,计划针对逻辑回归、岭回归等回归算法的数据隐私保护问题展开研究,并在隐私回归问题的时间及通信成本方面进一步优化。

参考文献:

[1] MELL P,GRANCE T.The NIST definition of cloud computing:SP 800-145[S].National Institute of Standards and Technology,Information Technology Laboratory,2011.

[2] 吴吉义,沈千里,章剑林,等.云计算:从云安全到可信云[J].计算机研究与发展,2011,48(S1):229-233.
WU J Y,SHEN Q L,ZHANG J L,et al.Cloud computing:cloud security to trusted cloud[J].Journal of Computer Research and Development,2011,48(S1):229-233.

[3] 郑剑,邹鸿珍.差异化隐私预算分配的线性回归分析算法[J].计算机应用与软件,2016,33(3):275-278.
ZHENG J,ZOU H Z.Linear regression anylysis algorithm of differential privacy budget allocation[J].Computer Applications and Software,2016,33(3):275-278.

[4] DANDEKAR A,BASU D,BRESSAN S.Differential privacy for regularised linear regression[C]//LNCS 11030:29th

International Conference on Database and Expert Systems Applications.Cham:Springer,2018:483-491.

[5] 葛宇航.基于差分隐私的线性回归分析[J].科技经济导刊,2019,27(14):163-164.
GE Y H.Linear regression analysis based on differential privacy[J].Technology and Economic Guide,2019,27(14):163-164.

[6] 李娟,马飞.基于同态加密的分布式隐私保护线性回归分析模型[J].微电子学与计算机,2016,33(1):110-113.
LI J,MA F.A model on distributed privacy preserving linear regression anylysis based on homomorphic encryption[J].Microelectronics & Computer,2016,33(1):110-113.

[7] KIKUCHI H,HAMANAGA C,YASUNAGA H,et al.Privacy-preserving multiple linear regression of vertically partitioned real medical datasets[J].Journal of Information Processing,2018,26:638-647.

[8] YANG H M,HE W C,ZHOU Q X,et al.Efficient and secure outsourced linear regression[C]//LNCS 11336:18th International Conference on Algorithms and Architectures for Parallel Processing.Cham:Springer,2018:89-102.

[9] YAO A C.Protocols for secure computation[C]//Annual Symposium on Foundations of Computer Science.Piscataway:IEEE,1982:160-164.

[10] 蒋瀚,徐秋亮.基于云计算服务的安全多方计算[J].计算机研究与发展,2016,53(10):2152-2162.
JIANG H,XU Q L.Secure multiparty computation in cloud computing[J].Journal of Computer Research and Development,2016,53(10):2152-2162.

[11] 蒋瀚,刘怡然,宋祥福,等.隐私保护机器学习的密码学方法[J].电子与信息学报,2020,42(5):1068-1078.
JIANG H,LIU Y R,SONG X F,et al.Cryptographic approaches for privacy-preserving machine learning[J].Journal of Electronics and Information Technology,2020,42(5):1068-1078.

[12] 谭作文,张连福.机器学习隐私保护研究综述[J].软件学报,2020,31(7):2127-2156.
TAN Z W,ZHANG L F.Survey on privacy preserving techniques for machine learning[J].Journal of Software,2020,31(7):2127-2156.

[13] DANKAR F,BRIEN R,ADAMS C,et al.Secure multiparty linear regression[C]//7th International Workshop on Privacy and Anonymity in the Information Society,Athens,Mar 28,2014:406-414.

[14] FU Z F.Linear regression protocol for privacy protect[C]//2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics,Hangzhou,Aug 26-27,2017.Piscataway:IEEE,2017:215-218.

[15] BLOOM J M.Secure multi-party linear regression at plaintext speed[J/OL].arXiv:1901.09531,2019.

[16] HALL R,FIENBERG S E,NARDI Y.Secure multiple linear regression based on homomorphic encryption[J].Journal of Official Statistics,2011,27(4):669-691.

[17] MARTINE D C,RAFAEL D,ANDERSON N,et al.Fast,privacy preserving linear regression over distributed datasets based on pre-distributed data[C]//8th ACM Workshop on Artificial Intelligence and Security.New York:ACM,2015:3-14.

[18] DANKAR F K.Privacy preserving linear regression on distributed databases[J].Transactions on Data Privacy,2015,8(1):3-28.

[19] ADRIÀ G,SCHOPPMANN P,BALLE B,et al.Privacy-preserving distributed linear regression on high-dimensional data[J].Nephron Clinical Practice,2017,4:345-364.

[20] MOHASSEL P,ZHANG Y.SecureML:a system for scalable privacy-preserving machine learning[C]//2017 IEEE Symposium on Security and Privacy.Piscataway:IEEE,2017:19-38.

[21] RABIN M.How to exchange secrets by oblivious transfer:TR-81[R].Aiken Computation Lab,Harvard University,1981.

[22] 唐春明,魏伟明.基于安全两方计算的具有隐私性的回归算法[J].信息网络安全,2018,18(10):10-16.
TANG C M,WEI W M.Regression algorithm with privacy based on secure two-party computation[J].Netinfo Security,2018,18(10):10-16.

[23] BEAVER D.Efficient multiparty protocols using circuit randomization[C]//LNVS 576:11th Annual International Cryptology Conference.Berlin:Springer,1991:420-432.

[24] AKAVIA A,SHAUL H,WEIS M,et al.Linear-regression on packed encrypted data in the two-server model[C]//7th ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography.New York:ACM,2019:21-32.

[25] DONG Y,CHEN X J,SHEN L Y,et al.Privacy-preserving distributed machine learning based on secret sharing[M]//Information and communications security.Berlin:Springer,2020:684-702.

[26] RIVEST R L,ADLEMAN L M,DERTOUZOS M L.On data banks and privacy homomorphisms[C]//Foundations of Secure Computation,New York,1978:169-179.

[27] PAILLIER P.Public-key cryptosystems based on composite degree residuosity classes[C]//LNCS 1592:18th International Conference on the Theory and Applications of Cryptographic Techniques.Berlin:Springer,1999:223-238.

[28] SHAMIR A.How to share a secret[J].Communications of the ACM,1979,22(11):612-613.

[29] DU W L,ATALLAH M J.Privacy-preserving cooperative scientific computations[C]//14th IEEE Computer Security Foundations Workshop,Cape Breton,2001:273-282.

[30] 陈莉,林柏钢.基于分布式线性方程组求解的安全多方计算协议[J].信息网络安全,2013,13(9):2-5.
CHEN L,LIN B G.Secure protocols for resolving distributed system of linear equations[J].Netinfo Security,2013,13(9):2-5.