面向算力网络的多方安全协同线性回归研究

潘洁[1]，侯慧芳[1]，陈曦[1]，薛嬰[1]，徐连坤[2]
（1. 中国移动通信集团设计院有限公司，北京 100080；
2. 中国移动通信集团有限公司，北京 100032）

摘　要：随着科技的飞速发展，机器学习已经成为推动企业进步的关键因素。然而，对于中小企业而言，数据量和算力的限制常常成为其应用机器学习的障碍。算力网络的兴起则为企业带来了新的机遇，但也伴随着数据安全等新的挑战。提出一种面向算力网络的多用户安全协同计算线性回归方案，该方案允许多个用户使用敏感数据在算力网络中实现安全的联合训练，从而构建线性回归模型。该方案采用低成本盲化手段与同态加密技术对用户敏感数据进行加密处理，从而保护了敏感数据的安全性。

关键词：线性回归；协同计算；数据安全
中图分类号：TN915.08；TP393.0
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024204

Research on multi-party security collaborative linear regression for computing power networks

PAN Jie[1], HOU Huifang[1], CHEN Xi[1], XUE Zhao[1], XU Liankun[2]
1. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China
2. China Mobile Communications Group Co., Ltd., Beijing 100032, China

Abstract: With the rapid development of science and technology, machine learning has become a key factor driving the progress of enterprises. However, for small and medium-sized enterprises, the amount of data and computing power often become obstacles to apply machine learning. The rise of computing power networks has brought new opportunities for enterprises, accompanied by new challenges such as data security. A linear regression scheme for multi-user security collaborative computing based on computing power networks was proposed. The scheme allowed multiple users to achieve secure joint training in a computing power network using sensitive data to build a linear regression model. The scheme adopts low-cost blinding method and homomorphic encryption technology to encrypt user sensitive data to protect the security of sensitive data.

Key words: linear regression, collaborative computing, data security

收稿日期：2024−06−11；修回日期：2024−08−15

0　引言

随着大数据驱动时代的快速发展，机器学习技术已成为企业获取竞争优势的重要手段。机器学习能够从大量数据中自动提取有用信息，并应用于预测和决策中。机器学习技术可以帮助企业用户更好地理解客户需求、优化业务流程，以及提高运营效率[1]。例如，通过分析客户数据，借助机器学习，企业可以制订更精准的营销策略，从而提升客户满意度和忠诚度[2-3]。然而，对于众多中小企业来说，由于缺少大规模数据集和强大的计算能力，往往难以充分发挥机器学习技术的优势。中小企业在应用过程中常面临数据量和算力的双重限制。数据量小可能导致模型训练不足，进而影响预测的准确性；算力不足则可能延长模型训练时间，甚至导致某些复杂模型无法有效运行。

算力网络的出现为这些企业带来了新的解决思路。算力网络通过网络将分散的计算资源连接起来，为中小企业提供了弹性且强大的算力支持[4-5]。这种网络化的计算方式不仅降低了中小企业的硬件投入成本，还能根据需求灵活调整算力规模。同时，算力网络还促进了企业间的数据共享和协作，有效缓解了数据量不足的问题[6]。然而，算力网络是不可信实体，将明文数据直接交给算力网络进行计算会带来数据泄露风险。此外，在算力网络中有多个用户参与计算过程，如何保护敏感数据在训练过程中不被其他用户获取或推导出明文信息，成为亟待解决的关键问题[7-8]。因此，设计一种安全的数据保护协同计算方法，确保中小企业能安全、高效地完成机器学习算法或应用的训练任务，具有重要的现实意义。

本文针对线性回归模型，在算力网络场景下，设计了一种多用户安全协同计算方案。本文提出的方案采用了特殊的矩阵盲化技术对用户数据盲化处理，使得算力网络在无法获取原始数据信息的同时可以基于盲化数据进行模型训练。同时，方案还利用同态加密技术对训练后的子模型进行加密，使模型可以在密文状态下进行聚合，从而保护模型的安全。本文所提方案充分考虑了安全性与高效性，能够使多个算力不足的用户在不泄露各自原始数据的前提下，高效地进行线性回归模型训练，降低了数据安全风险，充分发挥了算力网络的潜力，从而助力中小企业在大数据驱动时代取得更大的成功。

1　关键技术

1.1　线性回归

线性回归在统计学中是用来估计一个标量响应和一个或多个解释变量之间线性关系的统计模型[9]。在线性回归中，使用线性预测函数对关系进行建模，线性预测函数的未知模型参数是根据数据估计的[10]。对于一个样本集$(\boldsymbol{X},\boldsymbol{y})$，其中$\boldsymbol{X}$为由样本特征值构成的矩阵，$\boldsymbol{y}$为对应的标签向量。线性回归模型可以表示如下。

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\omega} \tag{1}$$

其中，$\boldsymbol{\omega}$表示决定预测准确度的权重向量。线性回归模型可以使用最小二乘法来求解，求解方式如下。

$$\boldsymbol{\omega}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{2}$$

1.2　Paillier同态密码系统

Paillier同态密码系统，由法国密码学家Pascal Paillier于1999年首次提出并以其名字命名，是一种用于公钥密码的概率非对称密码系统[11]。该系统中算法的安全性基于复合剩余类困难问题，具有加法与点乘的同态性质。Paillier同态密码系统由以下3个算法构成。

（1）$\mathrm{KeyGen}(k)\rightarrow(\mathrm{pk},\mathrm{sk},\mathrm{parameters})$：给定一个安全参数$k$，生成一个元组$(\mathrm{pk},\mathrm{sk},\mathrm{parameters})$，其中pk表示公钥，sk表示私钥，parameters表示

参数。

（2）Encrypt$(m, \mathrm{pk}, \mathrm{parameters}) \rightarrow c$：给定一个公开密钥pk，一个明文空间内的明文$m$，参数parameters，加密成一个密文$c$。

（3）Decrypt$(c, \mathrm{sk}, \mathrm{parameters}) \rightarrow m$：给定一个私钥sk，一个密文空间内的密文$c$，参数parameters，解密得到明文$m$。

给定任意由KeyGen算法生成的元组(n, λ, n)，Paillier同态密码系统满足以下两个性质。

- 正确性：对于任意的明文m，都有：

$$\mathrm{Decrypt}\left(\mathrm{Encrypt}\left(m, n, n\right), \lambda, n\right) = m \tag{3}$$

- 半同态性质：对于任意两个明文空间内的明文m_1和m_2，令：

$$c_1 = \mathrm{Encrypt}\left(m_1, n, n\right)$$

$$c_2 = \mathrm{Encrypt}\left(m_2, n, n\right) \tag{4}$$

其中，c_1、c_2满足$\mathrm{Decrypt}\left(c_1 \odot c_2, \lambda, n\right) = m_1 + m_2$，“+”为定义在明文空间$Z_n$下的加法，“⊙”为定义在密文空间$Z_{n^2}$下的乘法。

2 系统框架

面向算力网络的多方安全协同线性回归系统框架如图1所示。基于算力网络的多用户安全协同计算线性回归方案包含3种实体：算力网络$U=\{U_1, U_2, \cdots, U_n, U_r\}$，数据提供方$D=\{D_1, D_2, \cdots, D_n\}$，结果需求方$R$。算力网络负责提供稳定可靠的算力资源，数据提供方负责提供计算线性回归模型所需的数据，结果需求方负责接收计算好的线性回归模型。由于自身算力不足、数据规模复杂等，数据提供方无法在本地进行大规模数据运算。因此，多个数据提供方共同在算力网络中进行线性回归模型训练，然而，出于对隐私数据安全的考虑，数据提供方不能直接将明文数据上传到算力网络中进行运算。基于此，每个数据提供方首先在本地对数据盲化处理，再上传至算力网络中进行模型训练。计算出初步的子模型后，对子模型进行同态加密，然后发送至结果需求方的算力节点中进行聚合。根据同态加密的性质，结

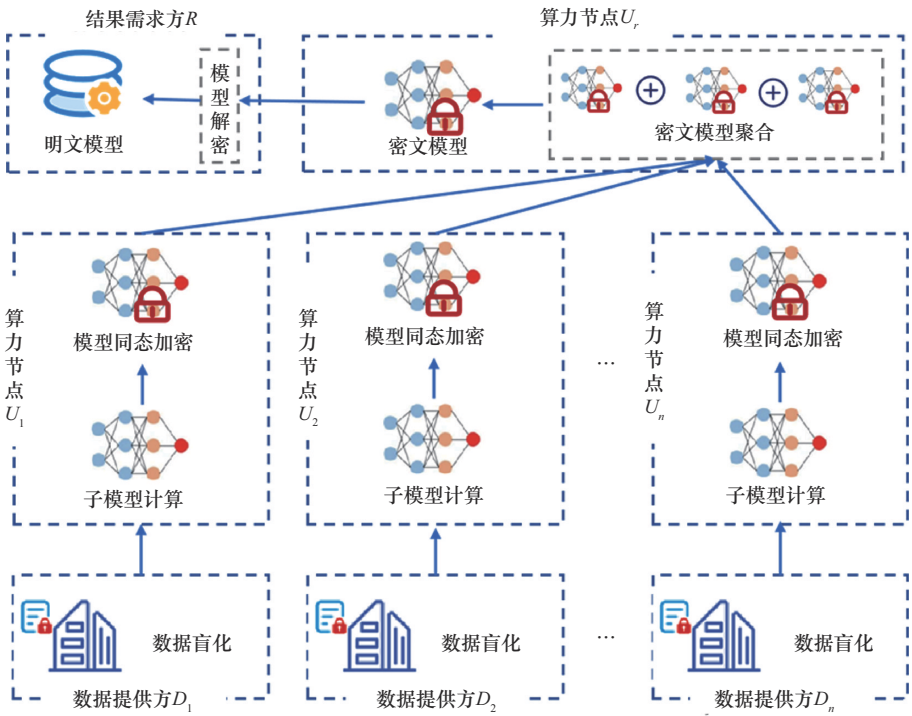

图1 面向算力网络的多方安全协同线性回归系统框架

果需求方可以在密文状态下，将子模型进行聚合，得到总的密文模型后，可以对其进行解密，最终得到线性回归模型。

3　方案设计

3.1　问题描述

线性回归，作为一种经典的机器学习算法，正逐渐改变着人们的生活方式、交流习惯和工作模式。然而，由于其涉及复杂矩阵运算，训练过程在计算资源上相对较重，尤其是在处理大规模数据集时，对于那些计算资源有限的用户群体来说，可能难以承担如此巨大的计算负荷。在此背景下，借助算力网络进行线性回归的训练成为一种理想的解决方案。一些数据资源有限的实体在训练线性回归模型时，由于数据量较少可能导致模型不够精确。因此，可以多个用户进行合作，使用各自的数据共同训练出一个更加精准的线性回归模型，但也引发了一系列新的安全问题。多个用户的训练数据集可能包含各自的敏感信息，这些信息不能泄露给算力网络以及其他用户。因此，如何使多个资源有限的用户能够安全、高效地将线性回归算法部署在算力网络中进行联合训练，成为一个亟待解决的重要问题。基于此，本文提出的方案满足以下目标。

（1）输入隐私性：方案需要确保算力网络不能根据数据提供方上传的加密数据恢复出用户的数据，结果需求方无法根据模型结果恢复出数据提供方的明文数据。

（2）输出隐私性：方案需要确保算力网络不能根据数据提供方的加密数据与密文模型推导出明文模型信息，数据提供方无法根据数据计算结果推出明文模型信息。

（3）高效性：数据提供方执行方案的效率应该远远高于数据提供方在本地执行线性回归算法的效率。

3.2　方案描述

假设有 n 个数据提供方 D_i（$1 \leqslant i \leqslant n$），一个结果需求方 R，他们在算力网络 U 中进行联合线性回归训练。

输入：数据提供方 D_i 的样本数据为 $(\boldsymbol{X}_i, \boldsymbol{y}_i)$（$1 \leqslant i \leqslant n$）。

输出：线性回归模型参数，权重向量 $\boldsymbol{\omega}$。

具体过程如下。

步骤 1　注册算力节点

数据提供方与结果需求方分别在算力网络中注册算力节点。设数据提供方 D_i 在算力网络中注册算力节点 U_i，结果需求方在算力网络中注册算力节点 U_r。

步骤 2　生成密钥

数据提供方 D_i（$1 \leqslant i \leqslant n$）生成一个随机矩阵 $\boldsymbol{Q}_i$ 作为自己的秘密矩阵，$\boldsymbol{Q}_i$ 是一个稀疏正交矩阵（$\boldsymbol{Q}_i$ 满足 $\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{Q}_i = \mathbf{I}$，$\boldsymbol{I}$ 是单位矩阵）。结果需求方 R 生成 Paillier 算法的公私钥对（sk,pk），生成一个公共密钥 $\boldsymbol{P}$，$\boldsymbol{P}$ 是一个稀疏的随机置换矩阵，将（pk,$\boldsymbol{P}$）公开给所有数据提供方。为了保护密钥的安全生成与传输，各方共同选择一个公钥密码算法（如 RSA、SM2 等）。基于该公钥密码算法，各方首先交换自己的公钥信息，然后利用公钥信息进行密钥协商协议，共同计算出会话密钥。这里的会话密钥有两个，数据提供方与结果需求方共享会话密钥 sek1，数据提供方、结果需求方与算力网络共享会话密钥 sek2。数据提供方 D_i 保留 $\boldsymbol{Q}_i$ 作为密钥不能泄露给结果需求方与算力节点。结果需求方使用会话密钥 sek1 加密密钥矩阵 $\boldsymbol{P}$，并发送给数据提供方，但不能泄露给算力节点。结果需求方使用会话密钥 sek2 加密 Paillier 公钥 pk，然后发送给数据提供方与算力网络。

步骤 3　盲化数据

数据提供方 D_i 使用密钥 $\boldsymbol{P}$、$\boldsymbol{Q}$ 对自己的样本

数据$(\boldsymbol{X}_i,\boldsymbol{y}_i)$盲化，得到$\boldsymbol{X}'$、$\boldsymbol{y}'$。

$$\begin{aligned}\boldsymbol{X}_i'&=\boldsymbol{Q}_i\boldsymbol{X}_i\boldsymbol{P}\\ \boldsymbol{y}_i'&=\boldsymbol{Q}_i\boldsymbol{y}_i\end{aligned}\tag{5}$$

数据提供方D_i将（$\boldsymbol{X}_i'$,$\boldsymbol{y}_i'$,pk）上传至U_i中。

步骤4 模型计算&加密

数据提供方D_i在U_i中计算：

$$\begin{aligned}A_i&=\boldsymbol{X}_i'^{\mathrm{T}}\boldsymbol{X}_i'\\ b_i&=\boldsymbol{X}_i'^{\mathrm{T}}\boldsymbol{y}_i'\end{aligned}\tag{6}$$

数据提供方D_i使用pk对（A_i,b_i）进行加密，得到密文矩阵$\boldsymbol{E}_{Ai}$、$\boldsymbol{E}_{bi}$。

数据提供方D_i发送（$\boldsymbol{E}_{Ai}$,$\boldsymbol{E}_{bi}$）给结果提供方R。

步骤5 模型密文聚合

结果需求方R接收所有参与方发送的（$\boldsymbol{E}_{Ai}$,$\boldsymbol{E}_{bi}$）后，在U_r中进行聚合，得到$\boldsymbol{E}_A$、$\boldsymbol{E}_b$。

$$\begin{aligned}\boldsymbol{E}_A&=\boldsymbol{E}_{A1}+\boldsymbol{E}_{A1}+\cdots+\boldsymbol{E}_{An}\\ \boldsymbol{E}_b&=\boldsymbol{E}_{b1}+\boldsymbol{E}_{b2}+\cdots+\boldsymbol{E}_{bn}\end{aligned}\tag{7}$$

步骤6 模型解密

结果需求方R在本地使用私钥sk对$\boldsymbol{E}_A$、$\boldsymbol{E}_b$进行解密，得到A^*、b^*。

步骤7 计算模型

结果需求方R在U_r中计算：

$$\boldsymbol{\omega}^*=A^{*-1}b^*\tag{8}$$

$$\begin{aligned}\boldsymbol{\omega}&=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\\ &=\left(\begin{bmatrix}X_1^{\mathrm{T}} & X_2^{\mathrm{T}} & \cdots & X_n^{\mathrm{T}}\end{bmatrix}\cdot\begin{bmatrix}X_1\\X_2\\\cdots\\X_n\end{bmatrix}\right)^{-1}\begin{bmatrix}X_1^{\mathrm{T}} & X_2^{\mathrm{T}} & \cdots & X_n^{\mathrm{T}}\end{bmatrix}\cdot\begin{bmatrix}y_1\\y_2\\\cdots\\y_n\end{bmatrix}\\ &=\left[X_1^{\mathrm{T}}X_1+X_2^{\mathrm{T}}X_2+\cdots+X_n^{\mathrm{T}}X_n\right]^{-1}\left[X_1^{\mathrm{T}}y_1+X_2^{\mathrm{T}}y_2+\cdots+X_n^{\mathrm{T}}y_n\right]\\ &=\left[A_1+A_2+\cdots+A_n\right]^{-1}\left[b_1+b_2+\cdots+b_n\right]\\ &=A^{-1}b\end{aligned}\tag{14}$$

其次，证明样本数据盲化的正确性。数据提供方D_i使用密钥$\boldsymbol{P}$、$\boldsymbol{Q}_i$将（$\boldsymbol{X}_i$,$\boldsymbol{y}_i$）盲化成（$\boldsymbol{X}_i'$,$\boldsymbol{y}_i'$），然后计算（A_i,b_i）并加密成（$\boldsymbol{E}_{Ai}$，$\boldsymbol{E}_{bi}$）：

$$\begin{aligned}A_i&=\boldsymbol{X}_i'^{\mathrm{T}}\boldsymbol{X}_i' & b_i&=\boldsymbol{X}_i'^{\mathrm{T}}\boldsymbol{y}_i^{\mathrm{T}}\\ &=(\boldsymbol{Q}\boldsymbol{X}_i\boldsymbol{P})^{\mathrm{T}}(\boldsymbol{Q}\boldsymbol{X}_i\boldsymbol{P}) & &=(\boldsymbol{Q}\boldsymbol{X}_i\boldsymbol{P})^{\mathrm{T}}(\boldsymbol{Q}\boldsymbol{y}_i)\\ &=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}_i\boldsymbol{P}, & &=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{y}_i\\ &=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{P} & &=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{y}_i\end{aligned}\tag{15}$$

步骤8 解盲化

结果需求方R在本地ω^*进行解盲化，得到最终模型$\boldsymbol{\omega}$。

$$\boldsymbol{\omega}=\boldsymbol{P}\boldsymbol{\omega}^*\tag{9}$$

3.3 安全性分析

定理1 在相同的数据下，所提出的方案计算出的线性回归模型结果与原始线性回归模型的结果是相同的。

证明： 首先证明数据聚合的正确性。方案中，数据提供方D_i的样本数据为$(\boldsymbol{X}_i,\boldsymbol{y}_i)$（$1\leqslant i\leqslant n$）。在明文的视角下，$D_i$需要计算：

$$\begin{aligned}A_i&=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\\ b_i&=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{y}_i\end{aligned}\tag{10}$$

然后R_i聚合所有参数，得到（A,b）：

$$\begin{aligned}A&=A_1+A_2+\cdots+A_n\\ b&=b_1+b_2+\cdots+b_n\end{aligned}\tag{11}$$

最后计算权重模型$\boldsymbol{\omega}$：

$$\boldsymbol{\omega}=A^{-1}b\tag{12}$$

将所有数据提供方的数据样本聚合成总的样本集$(\boldsymbol{X},\boldsymbol{y})$：

$$\boldsymbol{X}=\begin{bmatrix}X_1\\X_2\\\cdots\\X_n\end{bmatrix},\boldsymbol{y}=\begin{bmatrix}y_1\\y_2\\\cdots\\y_n\end{bmatrix}\tag{13}$$

那么，权重模型$\boldsymbol{\omega}$为：

结果需求方进行密文聚合计算（E_A,E_b），解密后得到（A^*,b^*）。

$$\begin{aligned}\boldsymbol{E}_A&=E_{A1}+E_{A2}+\cdots+E_{An}=E(A_1+A_2+\cdots+A_n)=E(A^*)\\ \boldsymbol{E}_b&=E_{b1}+E_{b2}+\cdots+E_{bn}=E(b_1+b_2+\cdots+b_n)=E(b^*)\end{aligned} \tag{16}$$

$$\begin{aligned}A^*&=D(\boldsymbol{E}_A)=A_1+A_2+\cdots+A_n=\boldsymbol{P}^{\mathrm{T}}A\boldsymbol{P}\\ b^*&=D(\boldsymbol{E}_b)=b_1+b_2+\cdots+b_n=\boldsymbol{P}^{\mathrm{T}}b\end{aligned} \tag{17}$$

其中，$E(\cdot)$表示Paillier密码的Encrypt算法，$D(\cdot)$表示Paillier密码的Decrypt算法。计算权重模型$\boldsymbol{\omega}$：

$$\begin{aligned}\boldsymbol{\omega}&=\boldsymbol{P}\boldsymbol{\omega}^*\\ &=\boldsymbol{P}A^{*^{-1}}b^*\\ &=\boldsymbol{P}(\boldsymbol{P}^{\mathrm{T}}A\boldsymbol{P})^{-1}\boldsymbol{P}^{\mathrm{T}}b\\ &=\boldsymbol{P}\boldsymbol{P}^{-1}A^{-1}\boldsymbol{P}^{\mathrm{T}^{-1}}\boldsymbol{P}^{\mathrm{T}}b\\ &=A^{-1}b\end{aligned} \tag{18}$$

定理2　数据提供方的样本数据具有隐私性，即在整个方案执行过程中，算力网络和结果需求方无法获得样本数据的明文信息。

证明：对于算力网络，可以获得数据提供方上传的（$\boldsymbol{X}_i'$,$\boldsymbol{y}_i'$）（$1\leqslant i\leqslant n$）值，该值由密钥矩阵$\boldsymbol{P}$、$\boldsymbol{Q}_i$对原始数据盲化。$\boldsymbol{P}$、$\boldsymbol{Q}_i$由数据提供方持有，算力网络无法获得。因此，算力网络随机化矩阵$\boldsymbol{P}'$、$\boldsymbol{Q}_i'$。那么，算力网络可以对（$\boldsymbol{X}_i'$,$\boldsymbol{y}_i'$）解密，得到（$\boldsymbol{X}_i^*$,$\boldsymbol{y}_i^*$）：

$$\begin{aligned}\boldsymbol{X}_i^*&=\boldsymbol{Q}_i'^{-1}\boldsymbol{X}_i'\boldsymbol{P}'^{-1}\\ \boldsymbol{y}_i^*&=\boldsymbol{Q}_i'^{-1}\boldsymbol{y}_i'\end{aligned} \tag{19}$$

由此可得，算力网络可以破解数据提供方的样本数据的概率为：

$$\begin{aligned}&\Pr[(\boldsymbol{X}_i^*=\boldsymbol{X}_i)\cap(\boldsymbol{y}_i^*=\boldsymbol{y}_i)]\\ &=\Pr[(\boldsymbol{P}'=\boldsymbol{P})\cap(\boldsymbol{Q}_i'=\boldsymbol{Q}_i)]\\ &=\prod_{j=1}^{2k}\frac{1}{\sigma}\end{aligned} \tag{20}$$

其中，$\frac{1}{\sigma}$为算力网络能猜到$\boldsymbol{P}$、$\boldsymbol{Q}_i$矩阵中一个元素的概率，k为构造P、$\boldsymbol{Q}_i$矩阵的单位向量的元素数量。由于$\boldsymbol{P}$、$\boldsymbol{Q}_i$矩阵的元素由实数构成，此概率在多项式时间内是可忽略不计的，因此，算力网络无法获得数据提供方的样本数据。

对于结果需求方，可以获得A^*，且已知$\boldsymbol{P}$，那么可以解密得到$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$值。然而结果需求方不知道$\boldsymbol{X}$的值，因此尝试随机化$\boldsymbol{X}$的值得到$\boldsymbol{X}^*$，那么结果需求方可以破解数据提供方的样本数据的概率为：

$$\Pr[((\boldsymbol{X}^*)^{\mathrm{T}}\boldsymbol{X}^*=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})]=\Pr[(\boldsymbol{X}^*=\boldsymbol{X})]=\prod_{j=1}^{h}\frac{1}{\sigma} \tag{21}$$

其中，h为矩阵$\boldsymbol{X}$的元素数量。由于$\boldsymbol{X}$的元素由实数构成，此概率在多项式时间内是可忽略不计的。因此，结果需求方无法获得数据提供方的样本数据。

定理3　结果需求方的权重模型具有隐私性，即在整个方案执行过程中，算力网络和数据提供方无法获得权重模型的明文信息。

证明：对于算力网络，可以获得经过计算后的加密权重模型$\boldsymbol{\omega}^*$。该加密模型是由密钥矩阵$\boldsymbol{P}$进行盲化的，且$\boldsymbol{P}$由结果需求方持有。算力网络无法获得$\boldsymbol{P}$的值。因此，算力网络尝试随机化矩阵来模拟$\boldsymbol{P}$的值，得到$\boldsymbol{P}'$。那么，算力网络可以使用$\boldsymbol{P}'$对$\boldsymbol{\omega}^*$进行解密：

$$\boldsymbol{\omega}'=\boldsymbol{P}'^{-1}\boldsymbol{\omega}^* \tag{22}$$

由此可得，算力网络可以破解结果需求方的权重模型的概率为：

$$\Pr[(\boldsymbol{\omega}'=\boldsymbol{\omega})]=\Pr[(\boldsymbol{P}'=\boldsymbol{P})]=\prod_{j=1}^{k}\frac{1}{\sigma} \tag{23}$$

由于矩阵$\boldsymbol{P}$的元素由实数构成，此概率在多项式时间内是可忽略不计的，因此，算力网络无法获得结果需求方的权重模型。

对于算力网络，仅仅将自己的盲化数据上传至算力网络，并使用自己的数据进行计算，无法获取别的数据提供方的数据，无法进行模型的聚合，无法得到最终的模型结果。因此，数据提供方想要获取结果需求方的权重模型，只能随机化一个权重$\boldsymbol{\omega}'$。那么数据提供方可以破解结果需求方的权重模型的概率为：

$$\Pr[(\boldsymbol{\omega}'=\boldsymbol{\omega})]=\prod_{j=1}^{t}\frac{1}{\sigma} \tag{24}$$

其中，t为权重模型的元素数量。由于$\boldsymbol{\omega}$的元素由实数构成，此概率在多项式时间内是可忽略不计的，因此，数据提供方无法获得结果需求方的权重模型。

3.4 实验评估

本次实验使用Python 3.8实现了所提出的方案，并从时间效率、模型精度等方面对所提出的方案进行评估。实验采用Intel(R) Xeon(R) Gold 5 218 CPU、2.30 GHz主频、32 GB内存、Linux操作系统作为实验环境进行评估。为了评估方案训练出的线性回归模型精度情况，实验采用经典的Iris鸢尾花数据集进行评估。该数据集共有150个样本，3个分类。实验抽取30个样本作为测试样本集，剩余120个样本作为训练样本集。由于方案涉及多个用户，因此，实验设定数据提供方的数量为2，结果提供方数量为1，算力网络数量为1，包含3个计算节点。针对训练样本集，将其中50个样本点作为第1个数据提供方的样本数据集，剩余70个作为第2个数据提供方的样本数据集。

不同模式下求解的权重模型值见表1。表1中的原始数据集是指使用全部的原始数据集训练出来的权重模型，即前述的120个样本数据集训练所得权重模型；明文方案是指不执行方案中涉及的加密解密技术，只执行模型训练与聚合；所提出的方案是指完整地执行本文提出的方案。由表1可知，3种模式下所得的权重模型与测试样本正确率完全相同，明文方案与原始数据集的对比验证了方案拆分聚合方法的正确性，所提出的方案与原始数据集的对比说明了所提出的方案在模型精度上是无损的。

为了评估方案的效率情况，本文统计了方案执行过程中每个实体、每个步骤的耗时情况。由于Iris鸢尾花数据集样本数量较少，为了更好地评估效率指标，本文通过随机采样生成了5个样本集，数量分别为5 000、6 000、7 000、8 000、10 000。不同实体执行方案时的耗时情况如图2所示。算力网络耗时最多、数据提供方的耗时最少。算力网络需要进行模型计算、Paillier加密操作以及模型聚合，承担了方案中的所有耗时操作。结果需求方需要进行Paillier的解密操作后对最后的模型解盲化操作。Paillier解密操作是仅次于加密操作的复杂操作，尽管低于算力网络的耗时，但是也远远高于数据提供方。数据提供方仅需要使用密钥矩阵对自己的数据盲化操作。由于密钥矩阵均是稀疏的，因此盲化操作的耗时是最少的。

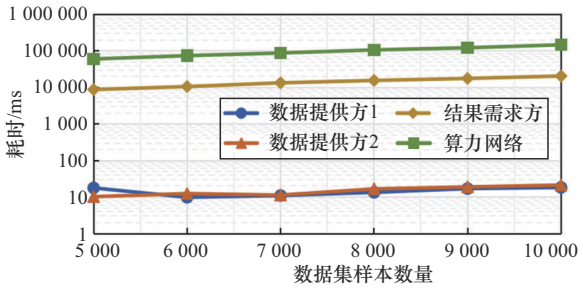

图2 不同实体执行方案的耗时情况

方案中各步骤的耗时情况如图3所示。模型计算和加密步骤耗时最高，这是因为涉及了矩阵运算以及Paillier加密运算，包含了方案最复杂的操作。模型解密步骤耗时次之，是因为其涉及了Paillier解密操作。盲化数据、模型密文聚合、计算模型以及解盲化耗时较少，是因为它们只涉及一些系数矩阵的乘法以及向量乘法。这些耗时较

表1 不同模式下求解的权重模型值

模式	权重模型值ω	测试样本正确率
原始数据集	[0.098 888 53, 0.019 636 44, 0.184 313 49, 0.555 801 63]	96.67%
明文方案	[0.098 888 53, 0.019 636 44, 0.184 313 49, 0.555 801 63]	96.67%
所提出的方案	[0.098 888 53, 0.019 636 44, 0.184 313 49, 0.555 801 63]	96.67%

少的步骤在数据提供方或结果需求方本地进行，而耗时较多的步骤则转移到算力网络进行，因此，对于数据提供方与结果需求方而言，大大提高了计算线性回归模型的效率。

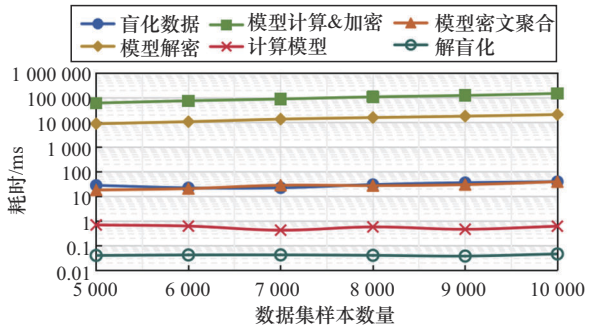

图3　方案中各步骤的耗时情况

为了进一步展示方案的高效性，本文对 pplr 方案[12]与本文提出的方案进行对比。由于两个方案提出的实体不完全相同，为了能够公正地展示两个方案的效率情况，本文对部分实体进行等效对齐。具体地，将 pplr 方案中的评估者（evaluator）与密码服务提供者（cryptographic service provider，CSP）合并为一个实体，将本文方案的算力网络与结果需求方合并为一个实体，合并的实体称为“求解者”。相同数据集不同规模下两个方案对比情况如图 4 所示。本文方案不论是在总耗时还是各实体的耗时方面，均比 pplr 方案耗时低，这是因为 pplr 方案的数据提供方需要计算 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 与 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$，且需要进行 Paillier 加密操作；求解者在计算线性回归模型中把一个线性方程组求解拆分成两个同样大小的线性方程组求解，这样虽然可以保证安全，但是计算量大大增加。本文提出的方案中，数据提供方只需要采取轻量级数据盲化操作，计算量非常小；求解者在求解线性回归模型时本质上只进行一个线性方程组的求解，与 pplr 方案相比计算量较小。综上所述，本文提出的方案相较于 pplr 方案是高效的。

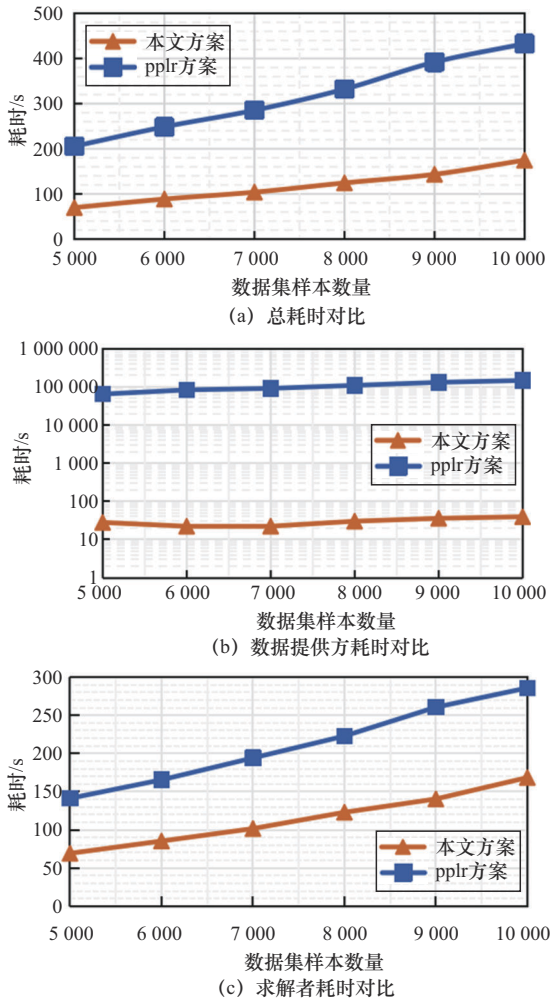

图4　相同数据集不同规模下两个方案对比情况

4　相关工作

不同学者在算力网络或者云服务器空间内解决机器学习效率与安全方面开展了大量研究。Zhou 等[2]提出了一种基于非线性同态认证加密的逻辑回归 Linear Regression Decision Diffie-Hellman（LR-DDH）分类模型，该模型利用椭圆曲线加密和非线性同态认证加密机制，实现了数据加密传输和计算，同时支持在不可信云服务器中进行逻辑回归模型的密文训练和更新。然而，模型在数据外包场景下，无法确保与多个云之间的安全交互以及防止中间人攻击。Arouj 等[3]针对计算密集型和延迟敏感型任务，提出了一种融合

边缘计算和分片学习的新型联邦学习框架。然而，该框架虽然优先选择剩余电量多的设备参与训练以减少能源消耗，但这一策略并未直接降低这些设备的功耗，设备仍有可能因电池耗尽而中途退出训练，导致dropout问题。Qiao等[13]提出了一种结合Hyperledger Fabric区块链和基于数据转换的共享方法，以保护多参与方信用数据的隐私和安全性，该方法通过数据转换、n-out-of-N OT协议及1-out-of-N盲目传输等技术，实现敏感数据的安全聚合和选择性共享。然而，该方法使用Rabin's Oblivious Transfer协议使得通信效率较低。Xu等[14]提出了一种应用于边缘计算的基于环形拓扑的协作群学习框架。该框架利用深度神经网络（deep neural network，DNN）进行热舒适性预测，通过损失函数和随机梯度下降优化模型，同时在Raspberry Pi上实现分布式边缘计算。然而，该框架在交换本地模型参数时没有进行保护，存在被反推原始数据的安全风险。Sim等[15]提出了一种通过估值和激励机制来确保各参与方有动力分享数据的同时保护隐私的方法，协调方通过充分统计量引发的关于模型参数的贝叶斯惊奇来评估其价值。然而，过于强烈的差分隐私保证可能导致模型准确性损失，影响协作的公平性和整体利益。

5 结束语

本文针对线性回归模型提出了一个基于不可信算力网络的多用户安全协同计算方案。该方案利用特殊的盲化手段，使每个资源受限的用户能够在本地高效地完成数据盲化，保护了数据隐私性。为了保护子模型安全，方案进一步采用同态加密技术对子模型进行加密，不仅保护了子模型的安全，而且实现了在密文上进行子模型的聚合。实验结果表明，本文提出的方案在模型精度上与原始线性回归模型相比是无损的，在效率方面，将线性回归中的复杂、耗时的操作转移到算力网络中进行计算，大大提高了数据提供方与结果需求方的计算效率。

参考文献：

[1] 郑士芹. 机器学习算法在数据挖掘中的应用[J]. 互联网周刊, 2024(9): 30-32.
ZHENG S Q. The application of machine learning algorithms in data mining[J]. Internet Weekly, 2024(9):30-32.

[2] ZHOU Y S, SONG L Y, LIU Y N, et al. A privacy-preserving logistic regression-based diagnosis scheme for digital healthcare[J]. Future Generation Computer Systems, 2023(144): 63-73.

[3] AROUJ A, ABDELMONIEM A M. Towards energy-aware federated learning via collaborative computing approach[J]. Computer Communications, 2024(221): 131-141.

[4] 朱文阅, 董江帆, 李玉华, 等. 算力网络安全与数据安全治理技术研究[J]. 电信工程技术与标准化, 2024, 37(1): 12-17.
ZHU W Y, DONG J F, LI Y H, et al. Research on computing power network security and data security governance technologies[J]. Telecom Engineering Technics and Standardization, 2024, 37(1): 12-17.

[5] 张焘, 许长桥, 连一博, 等. 基于深度强化学习的算力网络主动防御方法[J]. 中国科学(信息科学), 2023, 53(12): 2372-2385.
ZHANG T, XU C Q, LIAN Y B, et al. Deep reinforcement learning-based moving target defense method in computing power network[J]. Scientia Sinica (Informationis), 2023, 53 (12): 2372-2385.

[6] 沈昕炎, 林亚捷, 许方敏, 等. 面向工业大模型的算力网络架构与关键技术[J]. 自动化博览, 2024, 41(2): 43-47.
SHEN X Y, LIN Y J, XU F M, et al. Computing power network architecture and key technologies for large-scale industrial models[J]. Automation Panorama, 2024, 41(2): 43-47.

[7] 周旭, 李琢. 面向算力网络的云边端协同调度技术[J]. 中兴通讯技术, 2023, 29(4): 32-37.
ZHOU X, LI Z. Cloud-edge-end collaborative scheduling technology for computing power network[J]. ZTE Technology Journal, 2023, 29(4): 32-37.

[8] ZHENG L, CUI Y X, JIN S S, et al. High-performance computing-based open-source power transmission and distribution grid co-simulation[J]. IEEE Transactions on Power Systems, 2024, 39(5): 6144-6153.

[9] FREEDMAN D. Statistical models: theory and practice[M].

2nd ed. Cambridge: Cambridge University Press, 2009.

[10] IZENMAN A J, IZENMAN A J. Multivariate regression[J]. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, 2008: 159-194.

[11] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classes[C]//Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 223-238.

[12] QIU G W, GUI X L, ZHAO Y L. Privacy-preserving linear regression on distributed data by homomorphic encryption and data masking[J]. IEEE Access, 2020, 8: 107601-107613.

[13] QIAO Y C, LAN Q J, JIA S Y, et al. Blockchain-based multi-institution collaborative credit evaluation modeling method[J]. Computer Networks, 2024, 240: 110187.

[14] XU R X, JIN W Q, KHAN A N, et al. Cooperative swarm learning for distributed cyclic edge intelligent computing[J]. Internet of Things, 2023, 22: 100783.

[15] SIM R, ZHANG Y, HOANG N, et al. Incentives in private collaborative machine learning[J]. Advances in Neural Information Processing Systems, 2024, 2404: 01676.

[作者简介]

潘洁（1978– ），女，中国移动通信集团设计院有限公司高级工程师，主要研究方向为算力网络安全和信息安全。

侯慧芳（1986– ），女，现就职于中国移动通信集团设计院有限公司，主要研究方向为基于算力网络的新一代网络安全关键技术。

陈曦（1989– ），男，现就职于中国移动通信集团设计院有限公司，主要研究方向为网络与数据安全。

薛曌（1992– ），女，中国移动通信集团设计院有限公司工程师，主要研究方向为网络安全监测、算力网络安全、商用密码。

徐连坤（1976– ），男，中国移动通信集团公司高级工程师，主要研究方向为网络安全。