# 隐私计算场景下数据质量治理探索与实践

张燕，杨一帆，伊人，罗圣美，唐剑飞，夏正勋
星环信息科技（上海）股份有限公司，上海 200233

**摘要**
隐私计算是一种新型数据处理技术，可以在保护数据隐私及安全的前提下，实现数据价值转化和流通。然而隐私计算场景中“数据可用不可见”的特性给传统的数据质量治理工作带来了很大的挑战，业界尚缺乏完善的解决方案。针对上述问题，提出一种适用于隐私计算场景的数据质量治理方法与流程，构建了本地与多方两个层级的数据质量评估体系，能够兼顾本地域及联邦域的数据质量治理工作，同时提出了一种数据贡献度衡量方法，对隐私计算的长效激励机制进行探索，从而提升隐私计算的数据质量，并提高计算结果的精度。

**关键词**
隐私计算；联邦学习；数据质量治理；数据贡献度

中图分类号：D922.16，TP18　　文献标志码：A　　doi: 10.11959/j.issn.2096–0271.2022073

# *Exploration and practice of data quality governance in privacy computing scenarios*

**ZHANG Yan, YANG Yifan, YI Ren, LUO Shengmei, TANG Jianfei, XIA Zhengxun**
Transwarp Technology (Shanghai) Co., Ltd., Shanghai 200233, China

***Abstract***
Privacy computing is a new data processing technology, which can realize the transformation and circulation of a data value on the premise of protecting data privacy and security. However, the invisible feature of data in private computing scenarios poses a great challenge to traditional data quality management. There is still a lack of perfect solutions. To solve the above problems in the industry, a data quality governance method and process suitable for privacy computing scenarios were proposed. A local and multi-party data quality evaluation system was constructed, which could take into account the data quality governance of the local domain and the federal domain. At the same time, a data contribution measurement method was proposed to explore the long-term incentive mechanism of privacy computing, improve the data quality of privacy computing, and improve the accuracy of computing results.

***Key words***
privacy computing, federated learning, data quality governance, data contribution

## 0 引言

随着全球数字经济的蓬勃发展，数据作为生产要素的重要性日益凸显，其已渗透到人类生活的方方面面。近年来，政府及企业不断加强对数据安全、数据资产、数据隐私的保护[1]，使得数据主体之间、主体内部的“数据孤岛”现象日益突出，影响了数据价值的变现。隐私计算作为一种新型数据处理技术，能够在保护数据隐私的前提下，为跨域数据处理提供安全可靠的计算环境，实现多方协同数据处理，改变数据流通及使用的模式。隐私计算增强了数据流通过程中对个人隐私和数据安全的保护，其技术实现不仅涉及数据处理算法、处理流程的改变，还涉及数据预处理、特征工程、数据贡献度等细分领域的改造，当前业界对隐私计算算法、流程的讨论较多[2-5]，对数据质量治理、数据贡献度等方面的研究较少。

隐私计算对参与计算的数据质量有更高的要求。首先，隐私计算是一种多方协同计算，任何一方的数据质量出现问题，都很容易成为隐私计算的“短板”，“木桶效应”显著；其次，隐私计算通过加密中间数据实现数据流通，加密以及中间数据的信息传递方式在一定程度上减少了有效信息量，因此对数据质量提出了更高的要求。此外，隐私计算通常是跨部门、跨组织的协作计算，且相互之间不能见到对方的数据，这提高了隐私计算前期工作沟通及协调的复杂性，特别是数据预处理工作。因此，有必要对隐私计算场景下数据质量治理的相关工作展开研究，在“数据可用不可见”的情况下，实现多方数据的数据质量评估和优化。针对上述问题，本文研究了隐私计算场景下的数据质量评估及优化方法，并提出从数据质量评估、数据质量优化、数据贡献度评估3个方面构建隐私计算场景下的数据质量治理框架。该框架兼顾本地域及联邦域的数据质量治理工作，从而提升隐私计算的数据质量。在此基础上，本文还提出一种数据贡献度衡量方法，对隐私计算的长效激励机制进行探索。

## 1 隐私计算场景下的数据质量治理背景

随着数据规模、计算模式的变化，不同时期的数据质量治理工作有不同的内涵[6-10]。在数据仓库时代，数据大多为结构化数据，规模小且存储在单机系统中，此时数据质量治理主要是指数据质量评估和优化[11]，通常采用定量[12]或不定量[13]的方法评估数据质量，从数据源、数据预处理和元数据管理等方面优化数据质量[14]。数据仓库时代下的数据质量治理主要围绕数据的一致性、完整性、准确性和及时性开展，很少从数据相关性、数据价值等维度评估数据质量[15]。随着大数据技术的出现，数据规模成倍增加，数据质量治理面临多源、异构、海量、高时效的挑战[16]，数据质量治理的内容也因此扩展到数据标准定义、数据整合与清洗、数据质量评估、数据质量监控等数据质量管理全过程[17]，通过制订数据质量标准，定义数据质量规则库，构建数据质量评价指标体系，制订数据质量管理策略，实现全流程的数据质量治理[18]。但是，这种大数据质量治理的处理方式需要将多个组织的数据进行集中存储、集中处理，不可避免地存在数据安全及隐私泄露的风险，也给数据管理引入了合规风险[19]。随着国家、个人对数据安全和隐私保护的重视，隐私计算的应用越来越广泛。隐私计算是一种跨密码学、数据

科学、人工智能等多学科的技术[1]，多方协作进行联合计算和联合建模。隐私计算从机制上实现了原始数据不出库，从根源上降低了隐私泄露的风险[1]，但也提高了数据质量治理的技术复杂性和实施难度。在隐私计算场景中，联邦特征工程是传统特征工程算法在隐私计算环境下的重构，常用于对参与方的数据进行优化[20]。

数据贡献度常用来衡量数据参与方提供的数据价值，是数据质量治理中必不可少的一部分。传统的数据贡献度评估方法通常只使用数据量维度作为数据贡献度指标，忽略了数据质量的影响。在隐私计算场景中，为了让数据所有者持续提供数据，公平有效地评估每个参与方的数据贡献度至关重要。合理的贡献评价指标可以使激励机制公平分配联邦收益，激励数据所有者提供更有价值的数据[21]。当前，有专家研究本地数据质量与多方计算结果之间的影响关系，通过层次化影响分析，检测出本地数据中的负影响数据[22]或评估各参与方数据对多方计算结果的正向贡献[23]。也有专家将数据信息熵用于衡量数据集中包含的信息量[24]，以此作为数据参与方的数据贡献度，或从模型训练效果和训练成本角度确定数据参与方的数据贡献度[25]。

目前，针对隐私计算场景下数据质量治理的研究比较零散，不同于传统的数据质量治理方法，本文充分考虑了隐私计算场景下数据治理面临的诸多问题和挑战，例如如何在数据不可见的情况下实现联邦数据质量评估？如何在保护隐私的前提下，根据数据质量评估完成数据质量优化？完成数据质量治理之后，如何评估隐私计算过程中各参与方的数据贡献度，进而建立一种有效的激励机制？在传统方法的基础上，结合隐私计算“本地计算、联邦协同”的计算特点，本文提出从本地域和联邦域两个维度研究隐私计算场景下的数据质量治理问题，涵盖数据质量评估、数据质量优化、贡献度激励全流程。本文构建了本地与多方两个层级的数据质量评估体系，使用多个维度的综合评分度量数据质量，并依据本地数据质量评估结果和联邦数据质量评估结果，分别对数据质量进行本地优化和联邦优化，在数据不出本地、保障数据安全的前提下，实现隐私计算场景下的数据质量评估和优化。同时，本文从建模的视角出发，通过数据集贡献度、样本贡献度、特征贡献度等多个层次来量化参与方的总体数据贡献度。

## 2 隐私计算场景下的数据质量治理技术实现

隐私计算的主流技术[26-27]包括联邦学习（federated learning，FL）、多方安全计算（secure multi-party computation，MPC）[28]、可信执行环境（trusted execution environment，TEE）[29]3种，其中联邦学习被视为下一代人工智能协同算法和协作网络的基础[30]，是当下研究和应用的热点。因此，本文选择联邦学习作为重点场景来描述隐私计算场景下数据质量治理技术的具体实现，从数据质量评估、数据质量优化、数据贡献度评估3个方面构建隐私计算场景下的数据质量治理框架，如**图1**所示。

隐私计算场景下，数据质量治理需要综合考虑本地计算及联邦计算两种计算过程对数据质量的要求。本文分别从本地域和联邦域两个维度对各参与方数据进行质量评估，前者为本地数据质量评估，后者为联邦数据质量评估。基于本地数据质量评估结果可对参与方数据进行初步筛选，基于联邦数据质量评估结果可预判多方数据对联邦计算结果的增益。依据数据质

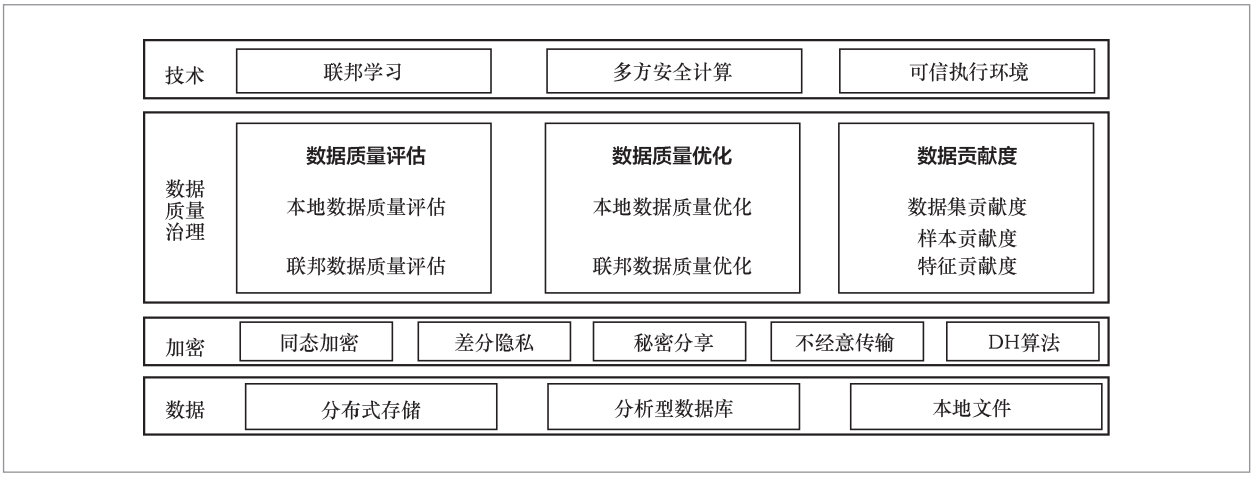


图1 隐私计算场景下的数据质量治理框架

量评估结果，指导各参与方进行本地和联邦数据质量优化工作，进一步提升数据质量。此外，为了鼓励更多的数据方积极参与到隐私计算中，非常有必要设计一套科学合理的贡献度衡量标准，衡量各参与方数据的贡献度，从而进行公平公正的联邦收益分配。

上述方法经过少量调整可适用于多方安全计算和可信执行环境场景下的数据质量治理。与联邦学习相比，它们的区别在于采用的密码学算法不同。多方安全计算场景下的联邦数据质量评估和优化一般采用不经意传输和秘密共享这两种经典的多方安全计算技术和方案，可信执行场景下的联邦数据质量评估和优化、贡献度评估主要依赖硬件算法实现。

## 2.1 隐私计算场景下的数据质量评估技术实现

联邦学习数据质量评估体系包括本地数据质量评估和联邦数据质量评估两个层级，质量评估的具体流程如**图2**所示。

如**图2**所示，联邦学习的参与方A和B先分别进行本地数据质量评估，再进行联邦数据质量评估。在本地数据质量评估层级，参与方A和B综合重复值评分、缺失值评分、异常值评分和单一值评分后，得到各自的本地数据质量评分。系统可以根据上报的本地数据质量评分，判断各参与方是否达到参与联邦学习的标准。在联邦数据质量评估层级，满足参加条件的参与方先进行样本对齐，再从数据重合度、信息量和线性相关性等维度考虑多方数据之间的相互影响，评估联邦数据质量。最终将参与方的本地数据质量评分和联邦数据质量评分进行加权计算，得到参与方的综合数据质量评分。

### 2.1.1 本地数据质量评估

本地数据质量评估包括计算重复值评分$S_r$、缺失值评分$S_m$、异常值评分$S_a$和单一值评分$S_s$ 4种，最终以4种评分的总分作为本地数据质量评分。4种评分的具体实现方法如下。

- 重复值评分$S_r$。每个参与方统计本地样本数据中重复的样本数量，计算重复的样本数量与总样本数量的比值，计算式如下：

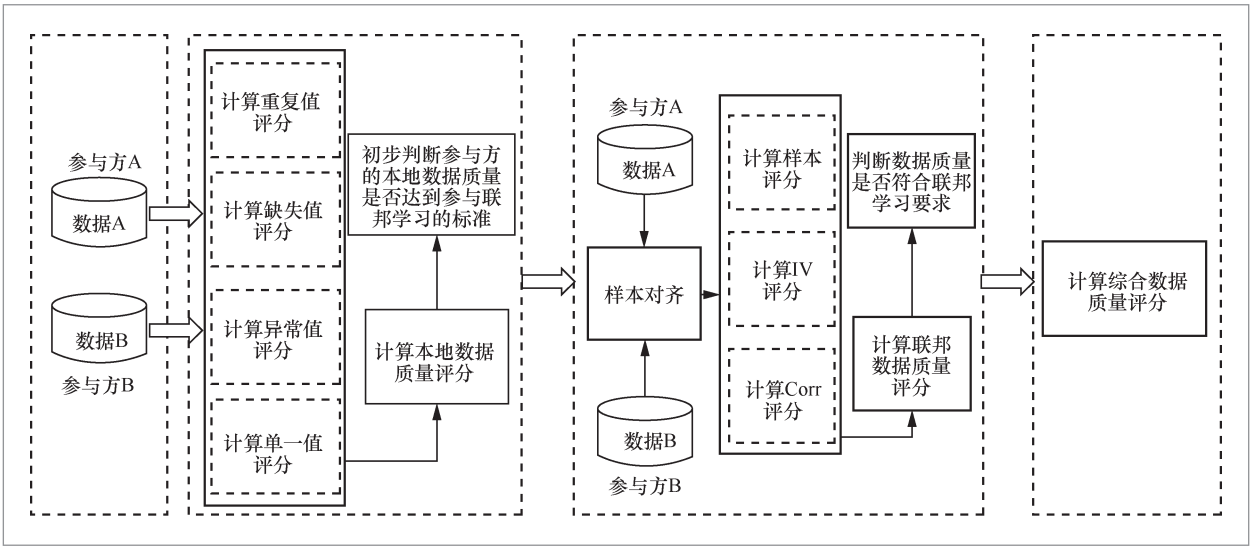


图 2 联邦学习数据质量评估流程

$$S_r = \text{round}\left(1-\frac{D_R}{D_T},2\right) \quad (1)$$

其中，$D_T$是参与方的本地样本数，$D_R$是重复样本数（出现重复则计数加1，不是“不同的重复样本数”），round函数将数字四舍五入到指定的位数。假设参与方A共有2 000个本地数据样本，其中有87个重复样本，那么参与方A的重复值评分为 $\text{round}\left(1-\frac{87}{2\,000},2\right)=0.96$；参与方B共有3 000个本地数据样本，其中有645个重复样本，那么参与方B的重复值评分为 $\text{round}\left(1-\frac{645}{3\,000},2\right)=0.78$。重复值评分越高，本地数据中重复出现的样本越少。

- 缺失值评分$S_m$。每个参与方对本地数据的每一维度特征的缺失值进行统计处理，即统计每一维度特征中特征值缺失或数值类型为“NULL”的样本数量占总样本数据的比例，计算式如下：

$$S_m = \text{round}\left(\frac{1}{p}\times\sum_{i=1}^{p}\left(1-\frac{D_{M_i}}{D_T}\right),2\right) \quad (2)$$

其中，$D_{M_i}$是第$i$维特征中特征值缺失或数值类型为“NULL”的样本数，$p$表示样本维数。假设参与方A的2 000个本地数据样本有3维特征，其中第1维特征中有39个数值类型为“NULL”的样本，第2维特征中有12个数值类型为“NULL”的样本，第3维特征中有10个数值类型为“NULL”的样本，则参与方A的缺失值评分为$\text{round}\left(\frac{1}{3}\times\left[\left(1-\frac{39}{2\,000}\right)+\left(1-\frac{12}{2\,000}\right)+\left(1-\frac{10}{2\,000}\right)\right],2\right)=0.99$；参与方B的3 000个本地数据样本有2维特征，其中第1维特征中有72个数值类型为“NULL”的样本，第2维特征中有75个数值类型为“NULL”的样本，则参与方B的缺失值评分为$\text{round}\left(\frac{1}{2}\times\left[\left(1-\frac{72}{3\,000}\right)+\left(1-\frac{75}{3\,000}\right)\right],2\right)=0.98$。缺失值评分越高，本地数据中特征值缺失或数值类型为“NULL”的样本越少。

- 异常值评分$S_a$。每个参与方对本地数据的每一维度特征的异常值进行统计。对于连续型数据，可以使用绝对中位差（median absolute deviation，MAD）方法（一种非参数方法）、枢轴量法（即常

见的3-$\sigma$法则）、四分位距（interquartile range，IQR）方法（一种非参数方法）等进行评分。这里以联邦学习IQR方法[31]为例，定义IQR为上75%分位数$\xi_{75\%}$与下25%分位数$\xi_{25\%}$的差值，$t$为阈值，将超过上限$\xi_{75\%}+t\times\mathrm{IQR}$或下限$\xi_{25\%}-t\times\mathrm{IQR}$的值定义为异常值，其中$\xi$为维度特征的特征值排序集合。对于离散型数据，若数据是编码类型的，将超出编码取值范围（超过上下限或者出现未定义编码）的值定义为异常值。然后，计算特征属于异常值的样本数量占总样本数量的比例，根据该比值计算异常值评分，计算式如下：

$$S_{\mathrm{a}}=\mathrm{round}\left(\frac{1}{p}\times\sum_{i=1}^{p}\left(1-\frac{D_{\mathrm{A}_i}}{D_{\mathrm{T}}}\right),2\right) \quad (3)$$

其中，$D_{\mathrm{A}_i}$是第$i$维特征为异常值的样本数。假设参与方A的2 000个本地数据样本有3维特征，假设阈值$t$取1.5，则上限为$\xi_{75\%}+1.5\times\mathrm{IQR}$，下限为$\xi_{25\%}-1.5\times\mathrm{IQR}$，其中第1维特征有658个异常值，第2维特征有426个异常值，第3维特征有200个异常值，那么参与方A的异常值评分为round$\left(\frac{1}{3}\times\left[\left(1-\frac{658}{2\,000}\right)+\left(1-\frac{426}{2\,000}\right)+\left(1-\frac{200}{2\,000}\right)\right],2\right)=0.79$；参与方B的3 000个本地数据样本有2维特征，其中第1维特征有665个异常值，第2维特征有649个异常值，那么参与方B的异常值评分为$\mathrm{round}\left(\frac{1}{2}\times\left[\left(1-\frac{665}{3\,000}\right)+\left(1-\frac{649}{3\,000}\right)\right],2\right)=0.78$。异常值评分越高，本地数据中有异常值的样本越少。

- 单一值评分$S_{\mathrm{s}}$。每个参与方对本地数据的每一维度在规定量纲条件下的标准差进行统计。若某一维度特征的标准差小于阈值，则该维特征的单一值评分为0，反之为1。将所有维度特征的单一值评分的平均值作为本地数据的单一值评分，计算式如下：

$$S_{\mathrm{s}}=\mathrm{round}\left(\frac{1}{p}\times\sum_{i=1}^{p}I\left[v_i\geqslant t_i\right],2\right) \quad (4)$$

其中，$v_i$是参与方本地样本第$i$维特征的标准差，$t_i$是第$i$维特征的阈值。假设参与方A的本地数据有3维特征，阈值$t$取$10^{-8}$，其中第1维特征的标准差为186，第2维特征的标准差为37，第3维特征的标准差为$9\times10^{-9}$，那么参与方A的单一值评分为$\mathrm{round}\left(\frac{1}{3}\times(1+1+0),2\right)=0.67$；参与方B的本地数据有2维特征，其中第1维特征的标准差为$3\times10^{-10}$，第2维特征的标准差为$5\times10^{-6}$，那么参与方B的异常值评分为$\mathrm{round}\left(\frac{1}{2}\times(0+1),2\right)=0.50$。单一值评分越高，本地数据的规范性越高。

综合上述指标的评分，计算本地数据质量评分，本地数据质量评分=重复值评分+缺失值评分+异常值评分+单一值评分，即：

$$S_{\mathrm{local}}=S_{\mathrm{r}}+S_{\mathrm{m}}+S_{\mathrm{a}}+S_{\mathrm{s}} \quad (5)$$

各参与方可事先约定本地数据质量评分阈值（既可设定单一评分阈值，也可以是总分阈值），若参与方的本地数据质量评分低于该阈值，说明其数据质量不高，其他参与方可拒绝与之一起进行联邦学习。

### 2.1.2 联邦数据质量评估

联邦数据质量评估旨在判断参与方对总体数据质量是否有增益作用，具体做法为利用隐私集合求交[32-34]、联邦IV（information value）、联邦线性相关系数等算法，分别计算数据样本评分、IV评分和Corr评分，综合上述3种评分，最终得到联邦数据质量评估结果。

进行联邦数据质量评估时，首先利用隐私集合求交技术将所有参与方数据进行样本对齐处理，再进行多维度评分，从而评估联邦环境下的数据质量。其中，隐私集合求交是在不共享原始数据的前提下，实现对所有参与方数据的交集运算，达到样本对齐的目的。样本对齐后，就可以计算样本评分、IV评分、Corr评分，具体如下。

（1）样本评分 $S_{\text{sample}}$

样本对齐后，计算样本重合比例。假设参与方A无标签，参与方B有标签，将A与B的数据进行样本对齐处理，然后使用样本重合比例计算样本评分，计算式如下：

$$S_{\text{sample}}=\begin{cases}2\times\dfrac{\left|C_{\text{A}}\cap C_{\text{B}}\right|}{t\left|C_{\text{B}}\right|}, \dfrac{\left|C_{\text{A}}\cap C_{\text{B}}\right|}{\left|C_{\text{B}}\right|}<t\\ 2, \text{其他}\end{cases} \tag{6}$$

其中，$C_{\text{A}}$表示参与方A的样本数量，$C_{\text{B}}$表示参与方B的样本数量，$\dfrac{\left|C_{\text{A}}\cap C_{\text{B}}\right|}{\left|C_{\text{B}}\right|}$表示样本重合比例，$t$为给定阈值。样本评分越高，联邦数据中的对齐样本比例越大。

（2）IV评分 $S_{\text{IV}}$

IV用于衡量特征变量的目标预测能力的大小。一般来说，IV越大，该特征的预测能力越强，信息贡献度越高。通过计算参与方数据每一列特征的IV，对联邦特征的信息量进行评估，同时，可以根据IV对特征变量进行筛选。在二分类场景下，IV的计算式[35]如下：

$$\text{IV}=\sum_i\left[\left(\frac{b_i}{b_T}-\frac{g_i}{g_T}\right)\log\left(\frac{b_i/b_T}{g_i/g_T}\right)\right] \tag{7}$$

其中，$b_i$和$g_i$分别表示第$i$个分组中属于类别1和属于类别2的样本数量，$b_T$和$g_T$分别表示属于类别1和属于类别2的样本总数。

与传统的IV计算方式不同，在联邦学习场景下，需要通过加密条件下的数据交互来实现IV计算。纵向联邦学习场景下的联邦IV[20]计算流程如**图3**所示，假设参与方A只有特征$X$没有标签，参与方B同时拥

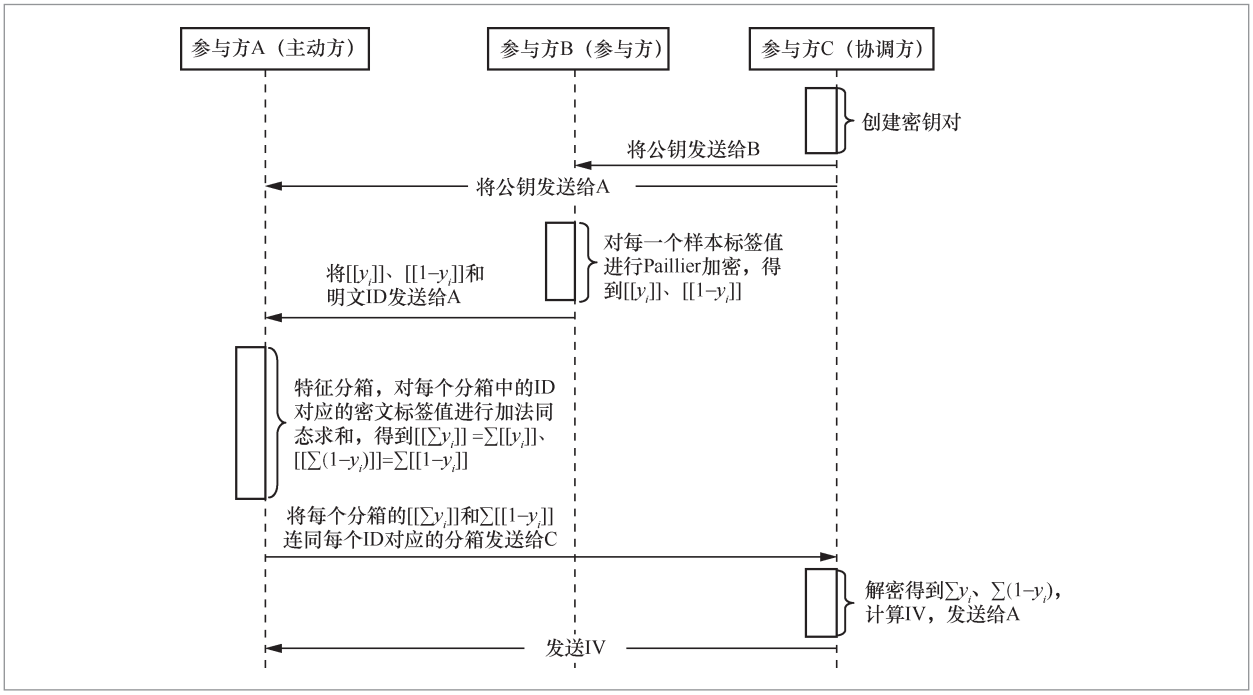


图 3 联邦 IV 计算流程

有特征$X$和标签$Y$，C是协调方。

- C先创建密钥对，并将公钥发送给A和B。
- B采用同态加密方法（如Paillier算法等）加密每一个样本$i$的标签值：$y_i$和$1-y_i$，并得到$[[y_i]]$和$[[1-y_i]]$，将其与明文ID一起发送给A。这是因为A没有标签，需要B提供密文标签值。
- A在本地对所有特征进行特征分箱，在接收到B的密文标签值和ID后，对每个分箱中的ID对应的密文标签值进行加法同态求和，得到每个分箱中的$[[\sum_i y_i]]=\sum_i[[y_i]]$和$[[\sum_i 1-y_i]]=\sum_i[[1-y_i]]$，再将其连同每个ID对应的分箱发送给C。
- C在接收到A的分箱求和结果后，解密得到$\sum_i y_i$和$\sum_i(1-y_i)$。此时$\sum_i y_i$ $\sum_i(1-y_i)$分别表示第$i$个分箱的正样本数和负样本数。

本文针对单个特征的评分标准为：

$$S_i=\begin{cases}0.5, & \mathrm{IV}<0.02\\ 1, & \mathrm{IV}\in[0.02,0.1)\\ 1.5, & \mathrm{IV}\in[0.1,0.3]\\ 2, & \mathrm{IV}>0.3\end{cases} \tag{8}$$

在应用实践中，IV小于0.02的特征变量对预测几乎没有效果，IV位于[0.02,0.1)区间的特征变量预测效果较弱，IV位于[0.1,0.3]区间的特征变量预测效果中等，如果IV大于0.3，那么这个特征变量的预测能力很强[20]。

本文使用的IV评分就是用联邦IV评估数据的信息量，具体计算式如下：

$$S_{\mathrm{IV}}=\frac{1}{p}\sum_{i=1}^{p}S_i \tag{9}$$

其中，$p$是特征数，$S_i$是第$i$个特征的IV评分值。

（3）Corr评分$S_{\mathrm{Corr}}$

线性相关系数表示特征变量之间的线性相关程度，计算式[36]如下：

$$\mathrm{Corr}(X,Y)=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2}\sqrt{\sum_i(y_i-\bar{y})^2}}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{10}$$

其中，$x_i$表示变量$X$中第$i$个样本的值，$\bar{x}$表示变量$X$的均值，$y_i$代表变量$Y$中第$i$个样本的值，$\bar{y}$表示变量$Y$的均值，$\mathrm{Cov}(X,Y)$表示$X$与$Y$的协方差，$\mathrm{Var}(X)$表示$X$的方差，$\mathrm{Var}(Y)$表示$Y$的方差。Corr为线性相关系数（简称Corr值），其绝对值的取值范围为0~1。通常来说，Corr的绝对值越接近1，变量$X$和$Y$之间的线性相关程度越高；Corr绝对值越接近0，$X$和$Y$之间的线性相关程度越低。也可以将多项式回归系数[28]作为$\mathrm{Corr}(X,Y)$。

针对联邦学习场景下的线性相关系数计算，同样需要通过加密条件下的数据交互来实现。纵向联邦学习场景的联邦Corr值计算流程如**图4**所示，假设参与方A只有特征$X$没有标签，参与方B同时拥有特征$X$和标签$Y$，C是协调方。

- C先创建密钥对，并将公钥发送给A和B。
- A计算本地特征$X$的方差$\mathrm{Var}(X)$，使用同态加密方法（如Paillier算法等）加密$\mathrm{Var}(X)$，得到$X$的密文方差$[[\mathrm{Var}(X)]]$，并将其发送给B。
- B先计算本地特征$Y$的方差$\mathrm{Var}(Y)$，接收到A的特征$X$的密文方差$[[\mathrm{Var}(X)]]$后，计算$[[\mathrm{Var}(X)]]\times\mathrm{Var}(Y)$，并将结果发送给C。
- C在接收到B的$[[\mathrm{Var}(X)]]\times\mathrm{Var}(Y)$后，进行乘法同态解密，得到$\mathrm{Var}(X)\times\mathrm{Var}(Y)$。

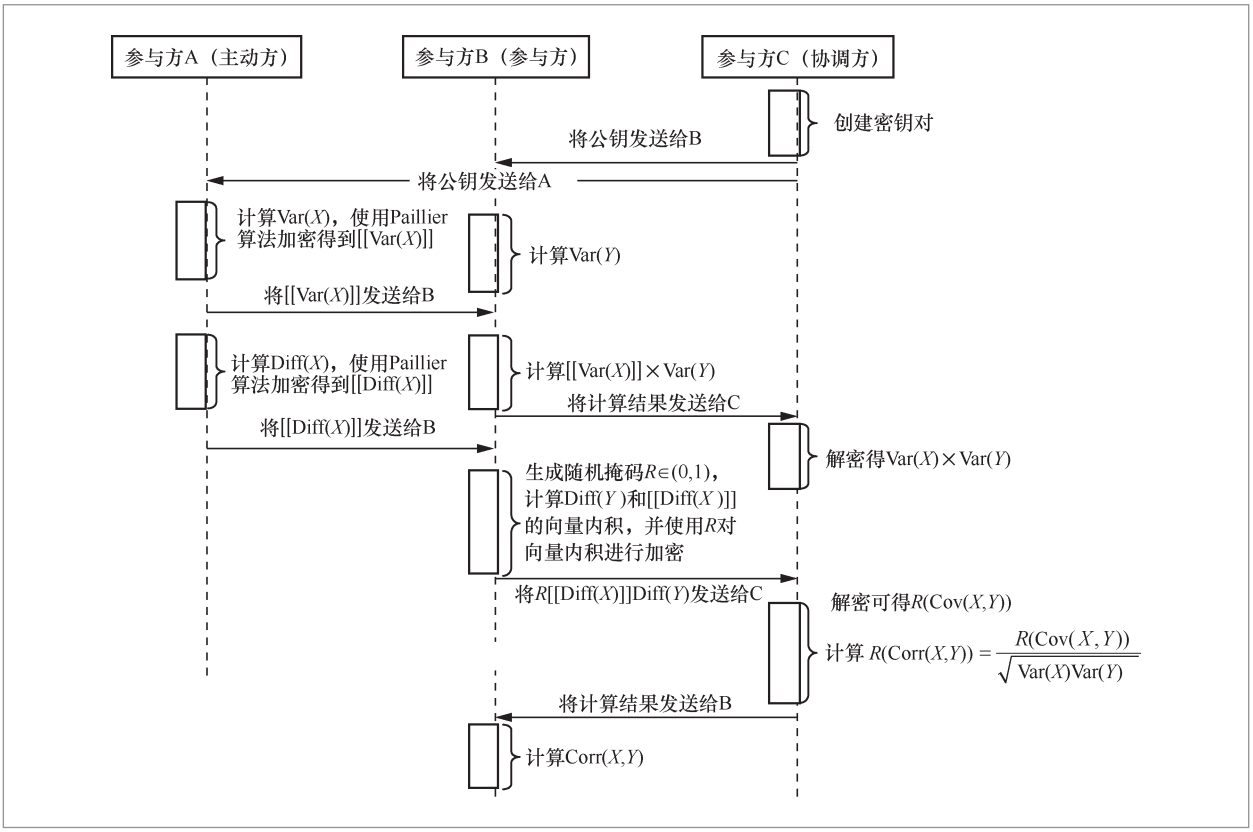


图 4 联邦 Corr 值计算流程

- A计算本地特征$X$与其均值的差值$\mathrm{Diff}(X)$，使用同态加密方法（如Paillier算法等）加密$\mathrm{Diff}(X)$，得到密文差值$[[\mathrm{Diff}(X)]]$，并将其发送给B。
- B在本地生成随机掩码$R$，$R$的取值范围为$(0,1)$，并计算特征$Y$与其均值的差值$\mathrm{Diff}(Y)$，在接收到A的密文差值$[[\mathrm{Diff}(X)]]$后，计算$[[\mathrm{Diff}(X)]]$与$\mathrm{Diff}(Y)$的向量内积$\sum_i (x_i-\bar{x})(y_i-\bar{y})$，即$\mathrm{Cov}(X,Y)$，利用生成的随机掩码$R$对$\mathrm{Cov}(X,Y)$进行加密，即$R(\mathrm{Cov}(X,Y))$，并将加密后的$[[R(\mathrm{Cov}(X,Y))]]$发送给C。
- C接收到B的密文$[[R(\mathrm{Cov}(X,Y))]]$后，进行乘法同态解密，得到$R(\mathrm{Cov}(X,Y))$，计算$R(\mathrm{Corr}(X,Y))=\frac{R(\mathrm{Cov}(X,Y))}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$，并将结果发送给B。
- B收到$R(\mathrm{Corr}(X,Y))$后，使用随机掩码$R$解密得到$\mathrm{Corr}(X,Y)$。

本文利用联邦Corr值计算Corr评分$S_{\mathrm{Corr}}$，计算式如下：

$$S_{\mathrm{Corr}}=2\times\frac{1}{p}\sum_{i=1}^{p}\mathrm{Corr}_i \qquad (11)$$

其中，$p$是$X$的特征数，$\mathrm{Corr}_i$表示第$i$个特征与$Y$的Corr值。

基于上述指标评分，计算联邦数据质量评分：联邦数据评分=样本评分+IV评分+Corr评分，即：

$$S_{\mathrm{Federated}}=S_{\mathrm{sample}}+S_{\mathrm{IV}}+S_{\mathrm{Corr}} \qquad (12)$$

根据联邦数据质量评分，判断参与方数据对于总体数据质量是否有增益作用。各参与方可事先约定联邦数据质量评分阈值（既可设定单一评分阈值，也可以是总分

阈值），若参与方的联邦数据质量评分超过该阈值，则说明参与方数据能提升总体数据质量；反之，参与方数据可能降低总体数据质量，需进一步排查原因。

## 2.2 隐私计算场景下的数据质量优化技术实现

### 2.2.1 本地数据质量优化

本地数据质量优化主要基于本地数据质量评估结果，从完整性、规范性、一致性、准确性、唯一性等维度，对各参与方的数据进行本地优化[37]。关键技术包括重复样本去重[38]、缺失值填充[39]、异常值清除[40]、数据标准化和归一化[41]等。

### 2.2.2 联邦数据质量优化

针对本地数据质量评分较低的情况，除本地数据质量优化外，还可以进行联邦数据质量优化。具体如下。

- 联邦缺失值填充：针对本地数据质量评估结果中缺失值评分较低的情况，除本地缺失值填充外，还可以进行联邦缺失值填充，具体做法是对所有参与方的数据进行联调统计分析，计算全局均值，然后采用全局均值对缺失值进行填充。
- 联邦异常值处理：针对本地数据质量评估结果中异常值评分较低的情况，除本地异常值清除外，还可以进行联邦异常值处理，具体做法是对所有参与方的数据进行联调统计分析，计算每个特征的全局IQR值，将全局IQR值的上下限作为异常值的判断标准，并使用全局均值对异常值进行填充。
- 联邦标准化：针对数据质量评估结果中单一值评分较低的情况，除本地数据标准化处理外，还可以进行联邦标准化处理。标准化是指计算目标列的均值$\mu$和标准差$\sigma$，并对该列每个元素$x$进行$(x-\mu)/\sigma$变换。标准化的作用是使处理后的数据服从标准正态分布。与本地数据标准化相比，联邦标准化的不同之处在于利用所有参与方的全局数据计算均值$\mu$和标准差$\sigma$，而不仅仅是各参与方的本地数据。

针对联邦数据质量评分较低的情况，可以采取联邦去重、联邦特征筛选、联邦字符串索引进行优化。具体如下。

- 联邦去重：在联邦数据之间去除重复样本或无关特征。在横向联邦学习中，各参与方的数据特征要保持一致，同时要求数据样本要保持唯一性。在纵向联邦学习中，所有参与方需要找到具有共同ID的样本，样本ID不重合的数据不会参与到联邦建模中。因此，各参与方除了要在本地去除重复样本，还需要对联邦数据进行去重处理。隐私集合求交技术在保护数据隐私安全的前提下，完成多方数据的交集运算，实现横向联邦数据特征对齐和纵向联邦样本对齐，在实现特征或样本对齐的基础上，去除多余数据，直到联邦数据质量评估中的样本评分达到要求。
- 联邦特征筛选：特征筛选是为了从原始特征中找出最有效的特征，帮助减少特征的维度、降低数据冗余度，从而提升模型的性能。联邦数据质量评估中的IV评分和Corr评分可分别用于衡量特征变量预测能力以及特征变量与预测变量之间的相关程度。因此，当联邦数据质量评估结果中的IV评分或Corr评分较低时，可以基于联邦IV和联邦Corr值进行特征筛选，这有助于联邦任务发起方确保参与联合建模的特征维度能够有效提升模型效果。具体做法是计算每一列特征的联邦IV和联邦Corr值，筛选出IV或Corr值较高的特征作为联邦特征，继续参与联邦建模。

● 联邦字符串索引：字符串索引的作用是将$k$个不同的字符串映射到区间[0,$k$−1]的$k$个整数上，从而完成从字符串到数字的转变。联邦字符串索引在联邦学习场景下找到目标列出现的所有取值，并进行从字符串到数字的映射。

完成本地和联邦数据质量优化后，再重新评估参与方的数据质量评分，只有参与方的数据质量评分达到或超过规定阈值，才允许该参与方的数据参与到联邦建模中。例如，若某参与方本地数据质量评估中的重复值评分低于规定阈值，则可以要求该参与方进行样本去重，直到重复值评分超过规定阈值。

## 2.3 隐私计算场景下的数据贡献度评估技术实现

本文从建模的视角出发，通过计算参与方提供的数据对模型性能的贡献来决定收益分配。因此，本文从数据集贡献度、样本贡献度、特征贡献度等维度来量化参与方总体的数据贡献度。

● 数据集贡献度$C_{\text{Data}}$。数据集贡献度是指从数据量、数据质量两个维度评估参与方在训练样本集方面的贡献。数据集贡献度有助于更好地激励参与方贡献更多高质量数据。具体做法是使用加权法计算数据集贡献度，计算式如下：

$$\psi_j = \beta_1 \frac{T_j}{T_m} + \beta_2 \frac{\varphi_j}{\varphi_m} \tag{13}$$

其中，$\psi_j$表示第$j$个参与方的数据集贡献度，$m$表示参与方数量，$T_j$表示第$j$个参与方贡献的数据量，$T_m$表示所有参与方贡献的数据总量，$\varphi_j$表示第$j$个参与方的数据质量评分，$\varphi_m$表示所有参与方的数据质量总分之和，$\beta_1$和$\beta_2$分别为数据量和数据质量评分的权重。

● 样本贡献度$C_{\text{Sample}}$。样本贡献度将各参与方训练数据对模型效果的提升程度作为联邦建模贡献的评价标准，基本做法是将参与方训练数据中的实例样本删除后重新训练模型，并计算新模型的预测效果，可使用缺失法[23]计算各参与方数据样本对模型效果的提升程度。具体实现如下。

假设第$i$个实例对模型预测结果的影响表示[23]为：

$$\phi^{-i} = \frac{1}{n}\sum_{j=1}^{n}\left\|\hat{y}_j - \hat{y}_j^{(-i)}\right\| \tag{14}$$

其中，$n$表示样本量大小，$\hat{y}_j$表示第$j$个实例的预测结果，$\hat{y}_j^{(-i)}$表示第$i$个实例被删除时新模型的预测结果。假设该参与方贡献的数据集合为$D$，则其对建模的贡献可定义为：

$$\phi^{(-D)} = \sum_{i \in D}\phi^{(-i)} \tag{15}$$

也可以使用近似法估计每个参与方对建模效果提升的影响，具体做法是先从所有参与方中去除任意一个参与方，然后评估重新训练的模型预测效果，最后将其与之前所有参与方数据参与训练的模型预测效果进行对比。

● 特征贡献度$C_{\text{Feature}}$。特征贡献度通过分析样本中每个数据特征与模型预测结果之间的关系来量化数据特征对模型预测结果的贡献度，可用Shapley值方法等[42-43]量化各参与方数据对模型预测结果的贡献度。对于具体实例的特征变量$x_j$，其Shapley值是该特征在所有可能的特征组合上对模型预测结果贡献度的加权和，计算式[44]如下：

$$\phi_j = \sum_{S \subseteq \{x_1,\cdots,x_N\}/\{x_j\}} \frac{|S|!\left(N-|S|-1\right)!}{N!} \cdot \left(f\left(S \cup \{x_j\}\right) - f(S)\right) \tag{16}$$

其中，$N$表示特征维度，$x_j$表示第$j$个特征变量，$S\subseteq\{x_1,\cdots,x_N\}/\{x_j\}$表示不包含$x_j$的特征集合的子集，$f(S)$表示对特征子集$S$的预测结果，$f(S\cup\{x_j\})$表示对包含$x_j$的特征组合的预测结果，$f(S\cup\{x_j\})-f(S)$表示$x_j$的边际贡献，$\frac{|S|!(N-|S|-1)!}{N!}$表示权重。因此，特征贡献度的计算式为：

$$\varphi=\sum_{j=0}^{m}\left(\frac{1}{n}\sum_{i=0}^{n}\phi_j^i\right) \tag{17}$$

其中，$\varphi$表示参与方总特征贡献度，$m$表示参与方的特征维度，$n$表示参与方的数据样本总数，$\phi_j^i$表示参与方第$i$个样本中第$j$个特征的Shapley值。参与方的特征越多，其特征贡献度越大。

基于上述3个贡献度可以得到参与方的数据贡献度$C$，计算式为：

$$C=\alpha_1C_{\text{Data}}+\alpha_2C_{\text{Sample}}+\alpha_3C_{\text{Feature}} \tag{18}$$

其中，$\alpha_1$、$\alpha_2$、$\alpha_3$为权重系数。

对于联邦而言，参与方持续地参与联邦学习进程是其成功的关键所在。参与方加入联邦，构建一个机器学习模型，训练出的模型可以产生收益，参与方可以共享收益，以此为激励。根据本文提供的贡献度评估标准，可有效计算出各参与方数据对联邦模型的贡献度，可按照数据贡献度比例进行收益分配。

### 2.4 小结

第2节围绕数据质量评估、数据质量优化、数据贡献度评估3个方面描述了隐私计算场景下的数据质量治理技术实现。其中，数据质量评估从本地域和联邦域两个层面考虑，建立了本地与联邦两个层级的数据质量评估体系，使用多个维度的综合评分度量数据质量。同时，依据数据质量评估结果，分别对数据质量进行本地优化和联邦优化，在数据不出本地、保障数据安全的前提下，联合各方数据进行数据清洗及特征工程，全面提升参与方的数据质量。为了鼓励更多的数据方积极参与到联邦学习中，又从建模的视角出发，通过量化数据集贡献度、样本贡献度、特征贡献度，评估各参与方数据对整个联邦模型的贡献度，从而制订一种公平公正的联邦收益分配机制。

## 3 应用案例

某电力公司系统经过多年的信息化建设和完善，积累了大量数据资产，为了提质增效，公司决定挖掘电力数据的潜在商业价值。该公司联合水务部门采用联邦学习的方式，基于用电数据和用水数据进行群租房识别，但实际效果并不理想。通过对电力公司数据和水务部门数据的深度调研分析发现，参与联邦学习的参与方中，每个参与方存在数据粒度不同、样本标准不统一以及异常值、缺失值数据较多等问题，导致各参与方的数据质量参差不齐，严重影响联邦建模的性能。因此，如何对各参与方进行数据质量评估，提升参与联邦学习建模的数据质量，避免因数据质量问题降低模型性能，成为亟待解决的问题。

本应用案例基于星环科技联邦学习平台Transwarp Sophon FL对群租房识别模型进行联合训练，Transwarp Sophon FL框架如**图5**所示。

Transwarp Sophon FL采用分布式的数据计算与存储管理，集成同态加密、差分隐私、秘密分享、不经意传输、DH（Diffie-Hellman）算法等多种加密算法，

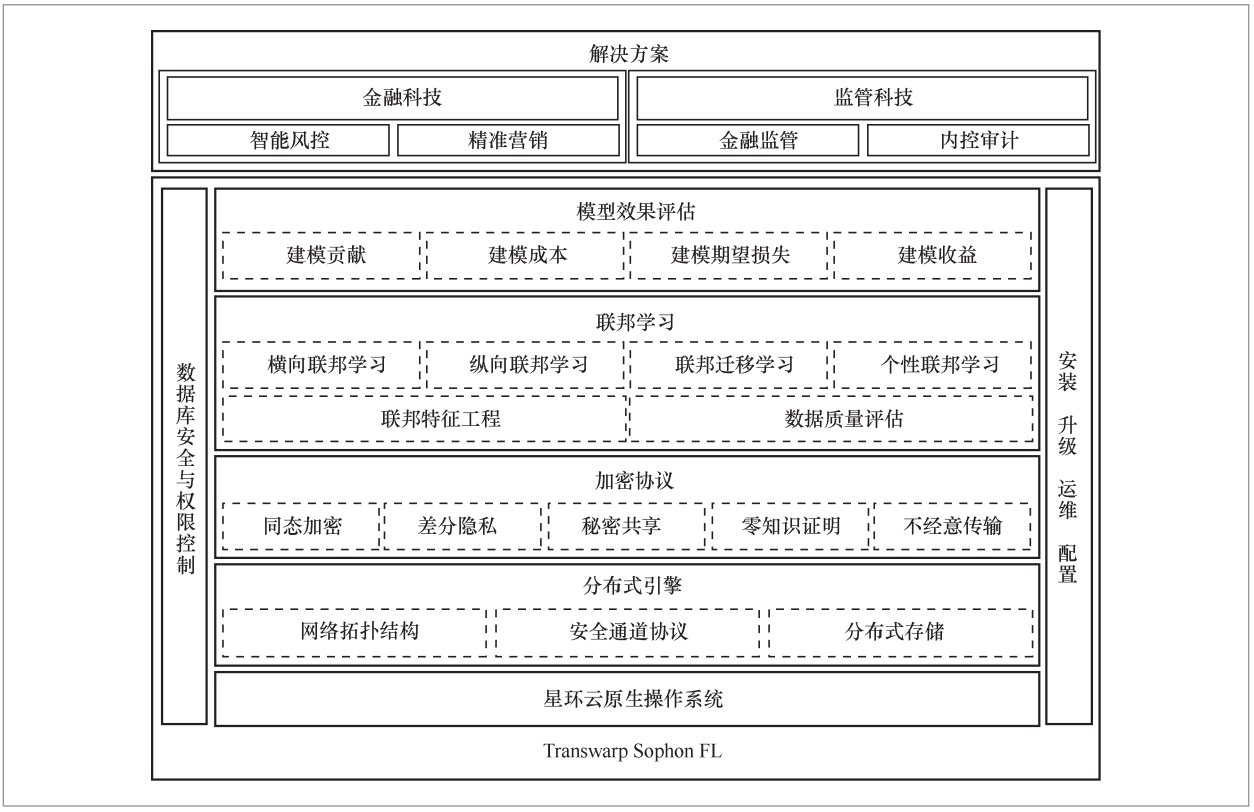


图 5 Transwarp Sophon FL 框架

保护数据隐私安全，使用联邦学习、多方安全计算、隐私计算、加密网络通信等多种功能，为多方安全建模提供完整的解决方案。同时，该平台还提供了一整套数据质量治理方法，方便用户在联邦框架下进行数据质量评估、数据质量优化、贡献度评估等工作，为AI模型的训练提供大量优质数据，大大提升联邦模型的性能。

在联邦建模过程中，电网公司为主动方，水务部门为参与方，采用纵向联邦学习模式，融合用电数据和用水数据，联合构建群租房识别模型，部署方式如**图6**所示。

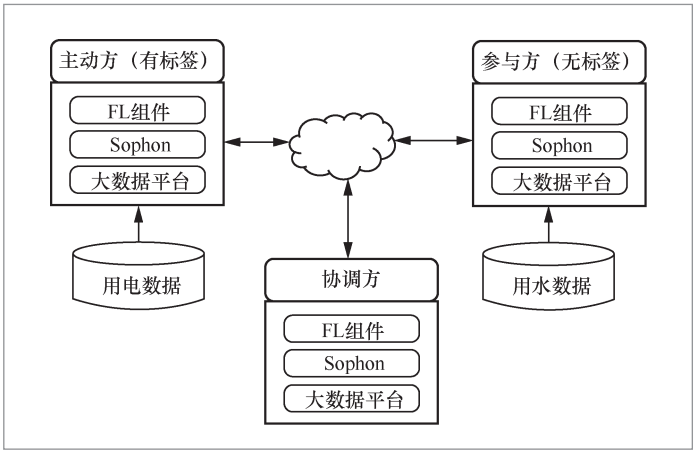


图 6 群租房识别应用部署方式

为了提高联邦学习模型的性能，本应用案例从数据质量评估、数据质量优化、数据贡献度评估3个方面对用电数据和用水数据进行数据质量治理。其中，在数据质量的综合评分中，本地数据质量评分的权重系数为0.4，联邦数据质量评分的权重系数为0.6。在数据贡献度评分中，数据集贡献度、样本贡献度、特征贡献度的权重系数均设置为1/3。本应用案例先分别计算电力公司和水务部门的本地数

据质量评分和联邦数据质量评分，然后依据各参与方的本地和联邦数据质量评估结果，分别对用电数据和用水数据进行数据清洗以及联邦特征工程等数据质量优化工作，并使用优化后的数据进行联合建模，最后评估训练数据的贡献度，并分配收益。其中，群租房识别模型的数据质量治理流程如图7所示。

在模型训练完毕后，双方协同使用用电数据和用水数据进行联合测试，生成群租房预测名单，测试流程如图8所示。

通过对比数据质量治理前后的群租房识别模型效果，验证了Transwarp Sophon FL数据质量治理框架在隐私计算场景下的优势。进行数据质量治理前，群租房识别模型的模型评估指标AUC[45]是0.7349，如图9所示；进行数据质量治理后，群租房识别模型的AUC是0.8188，如图10所示。进行数据质量治理后，群租房识别模型的AUC较之前提升了11.4%，为政府有效排查群租房，消除群租房造成的消防安全隐患，打造和谐、安全、美丽的生活环境做出了突出贡献。同时，在联

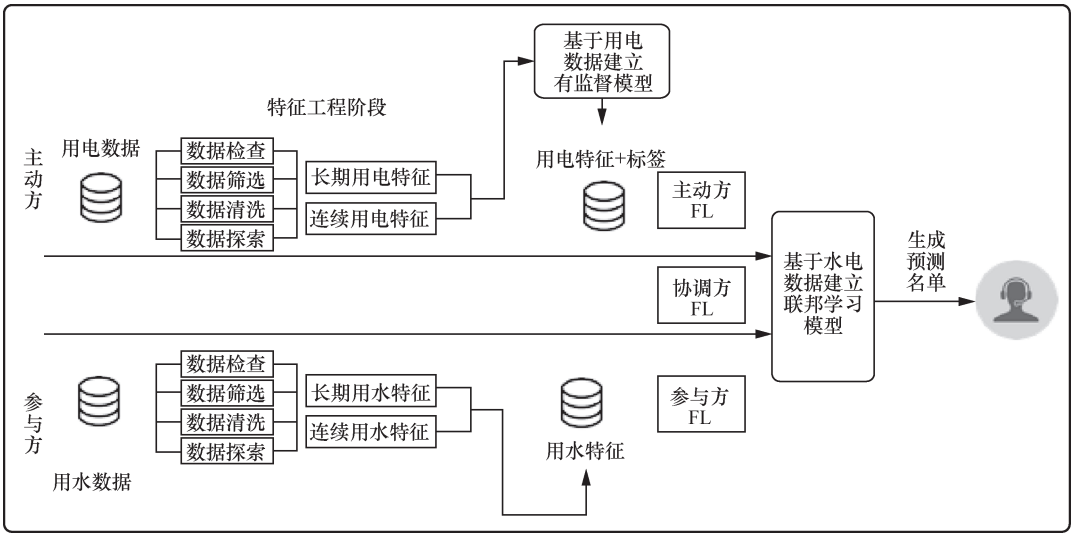


图 7 群租房识别模型数据质量治理流程

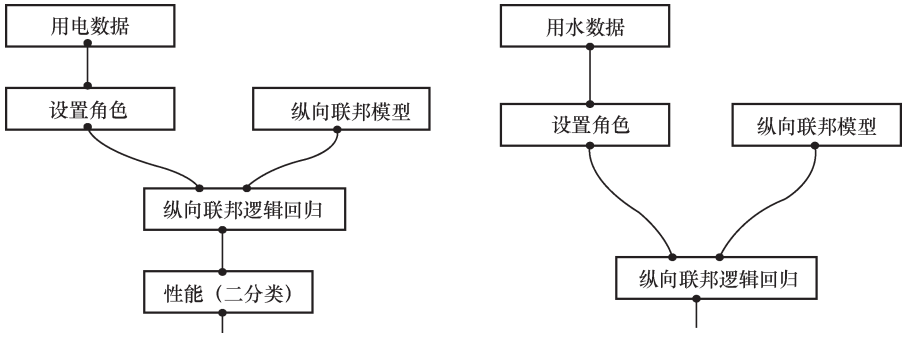


图 8 群租房识别模型测试流程

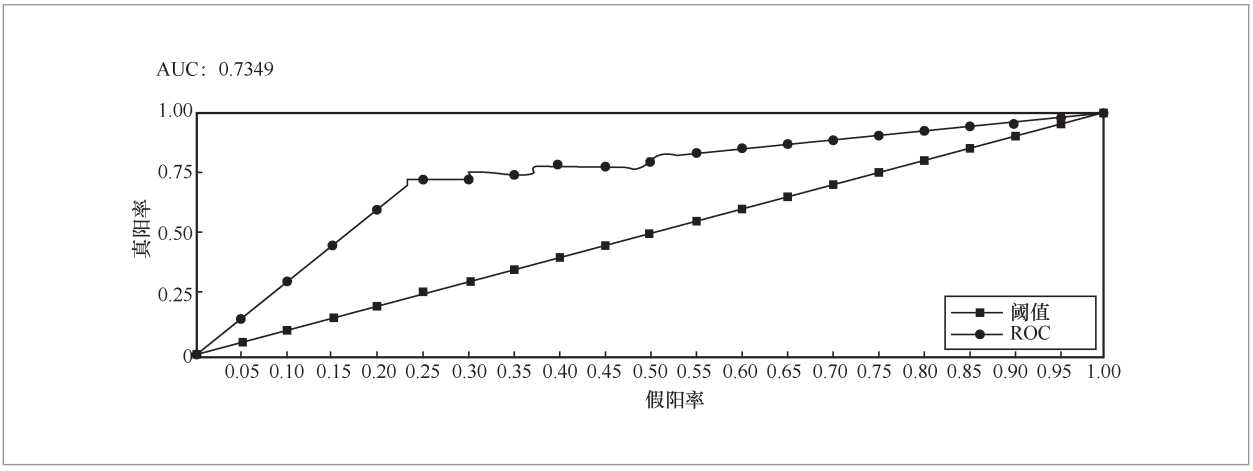


图 9 数据质量治理前群租房识别模型 AUC

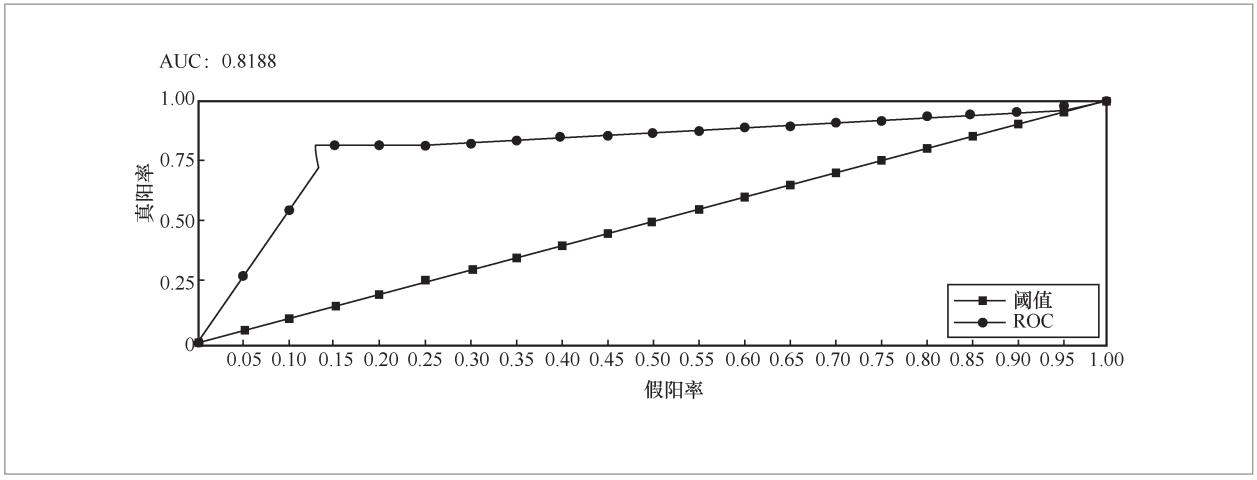


图 10 数据质量治理后群租房识别模型 AUC

合建模过程中，全程明文数据不出本地数据库，有效保护了居民用水用电的数据隐私。

## 4 结束语

本文对隐私计算场景下的数据质量治理工作进行了研究和探索，围绕数据质量评估、数据质量优化、数据贡献度3个维度构建了一种隐私计算场景下的数据质量治理框架，通过实践证明其在保护数据隐私的前提下，可实现隐私计算场景下的数据质量评估和优化，全方位提升了参与方的数据质量，提高了计算结果的精度。本文提出的隐私计算场景下的数据质量治理框架可被广泛应用到金融风控、联合医疗、保险智能定价、工业联合运维、供应链管理等场景中，具有广阔的应用前景。当然本文的研究尚有不足之处，比如本文考虑的隐私计算场景下的数据质量治理涉及大量的密文计算，计算效率还有待进一步提升；如何从数据治理视角防御多方隐私计算模式中的数据毒化[46]，尚缺乏完善的解决方案。这些问题也是下一阶段的重点工作。

## 参考文献:


[1] 中国信息通信研究院, 隐私计算联盟. 隐私计算白皮书(2021年)[R]. 2021.
China Academy of Information and Communication Technology, Privacy Computing Alliance. Privacy computing white paper(2021)[R]. 2021.

[2] 符芳诚, 侯忱, 程勇, 等. 隐私计算关键技术与创新[J]. 信息通信技术与政策, 2021, 47(6): 27-37.
FU F C, HOU C, CHENG Y, et al. Key technology and innovation of privacy preserving computing[J]. Information and Communications Technology and Policy, 2021, 47(6): 27-37.

[3] 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 preprint, 2017, arXiv:1711.10677.

[4] 李凤华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4): 1-11.
LI F H, LI H, JIA Y, et al. Privacy computing: concept, connotation and its research trend[J]. Journal on Communications, 2016, 37(4): 1-11.

[5] YANG S W, REN B, ZHOU X H, et al. Parallel distributed logistic regression for vertical federated learning without third-party coordinator[J]. arXiv preprint, 2019, arXiv:1911.09824.

[6] WAND Y, WANG R Y. Anchoring data quality dimensions in ontological foundations[J]. Communications of the ACM, 1996, 39(11): 86-95.

[7] PIPINO L L, LEE Y W, WANG R Y. Data quality assessment[J]. Communications of the ACM, 2002, 45(4): 211-218.

[8] 刘金晶, 王梅. 大数据下的数据质量评价指标构建实践[J]. 计算机技术与发展, 2019, 29(10): 46-50.
LIU J J, WANG M. Practice of data quality evaluating index construction under big data[J]. Computer Technology and Development, 2019, 29(10): 46-50.

[9] 中国信息通信研究院,大数据技术标准推进委员会.数据资产管理实践白皮书(4.0)[R]. 2019.
China Academy of Information and Communication Technology, Big Data Technology and Standerd Committee. Data asset management practices white paper(4.0)[R]. 2019.

[10] Firstlogic. Data quality assessment: a methodology for success[Z]. 2003.

[11] HEER J, HELLERSTEIN J M, KANDEL S. Data wrangling[M]//Encyclopedia of big data technologies. Cham: Springer, 2019: 584-591.

[12] 杨青云, 赵培英, 杨冬青, 等. 数据质量评估方法研究[J]. 计算机工程与应用, 2004, 40(9): 3-4, 15.
YANG Q Y, ZHAO P Y, YANG D Q, et al. Research on data quality assessment methodology[J]. Computer Engineering and Applications, 2004, 40(9): 3-4, 15.

[13] WANG R Y, STOREY V C, FIRTH C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640.

[14] 方幼林, 杨冬青, 唐世渭, 等. 数据仓库中数据质量控制研究[J]. 计算机工程与应用, 2003, 39(13): 1-4.
FANG Y L, YANG D Q, TANG S W, et al. Data quality managements in data warehouse[J]. Computer Engineering and Applications, 2003, 39(13): 1-4.

[15] 包阳, 齐璇, 李海龙. 大型软件系统数据质量问题研究[J]. 计算机工程与设计, 2011, 32(3): 963-967, 987.
BAO Y, QI X, LI H L. Research on data quality of large-scale software system[J]. Computer Engineering and Design, 2011, 32(3): 963-967, 987.

[16] 宗威, 吴锋. 大数据时代下数据质量的挑战[J]. 西安交通大学学报(社会科学版), 2013, 33(5): 38-43.
ZONG W, WU F. The challenge of data quality in the big data age[J]. Journal

of Xi'an Jiaotong University (Social Sciences), 2013, 33(5): 38–43.

[17] 吴信东, 董丙冰, 堵新政, 等. 数据治理技术[J]. 软件学报, 2019, 30(9): 2830–2856.
WU X D, DONG B B, DU X Z, et al. Data governance technology[J]. Journal of Software, 2019, 30(9): 2830–2856.

[18] 中国信息通信研究院. 数据安全治理实践指南(1.0)[R]. 2001.
China Academy of Information and Communication Technology. Data security governance practice guide (1.0)[R]. 2001.

[19] 黄刘生, 田苗苗, 黄河. 大数据隐私保护密码技术研究综述[J]. 软件学报, 2015, 26(4): 945–959.
HUANG L S, TIAN M M, HUANG H. Preserving privacy in big data: a survey from the cryptographic perspective[J]. Journal of Software, 2015, 26(4): 945–959.

[20] 彭南博, 王虎, 等. 联邦学习技术及实战[M]. 北京: 电子工业出版社, 2021.
PENG N B, WANG H, et al. Federated learning techniques and practices[M]. Beijing: Publishing House of Electronics Industry, 2021.

[21] 杨强, 刘洋, 程勇, 等. 联邦学习[M]. 北京: 电子工业出版社, 2020.
YANG Q, LIU Y, CHENG Y, et al. Federated learning[M]. Beijing: Publishing House of Electronics Industry, 2020.

[22] 李安然. 面向特定任务的大规模数据集质量高效评估[D]. 合肥: 中国科学技术大学, 2021.
LI A R. Efficient task-oriented quality assessment for large-scale datasets[D]. Hefei: University of Science and Technology of China, 2021.

[23] WANG G, DANG C X, ZHOU Z Y. Measure contribution of participants in federated learning[C]//Proceedings of 2019 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2019: 2597–2604.

[24] 朱建明, 张沁楠, 高胜, 等. 基于区块链的隐私保护可信联邦学习模型[J]. 计算机学报, 2021, 44(12): 2464–2484.
ZHU J M, ZHANG Q N, GAO S, et al. Privacy preserving and trustworthy federated learning model based on blockchain[J]. Chinese Journal of Computers, 2021, 44(12): 2464–2484.

[25] 王鑫, 周泽宝, 余芸, 等. 一种面向电能量数据的联邦学习可靠性激励机制[J]. 计算机科学, 2022, 49(3): 31–38.
WANG X, ZHOU Z B, YU Y, et al. Reliable incentive mechanism for federated learning of electric metering data[J]. Computer Science, 2022, 49(3): 31–38.

[26] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[J]. arXiv preprint, 2016, arXiv:1610.05492.

[27] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. arXiv preprint, 2019, arXiv:1908. 07873.

[28] YAO A C. Protocols for secure computations[C]//Proceedings of 23rd Annual Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 1982: 160–164.

[29] Open Mobile Terminal Platform Consortium. Advanced trusted environment: OMTP TR1[Z]. 2009.

[30] 杨强. 联邦学习：人工智能的最后一公里[J]. 智能系统学报, 2020, 15(1): 183–186.
YANG Q. Federated learning: the last on kilometer of artificial intelligence[J]. CAAI Transactions on Intelligent Systems, 2020, 15(1): 183–186.

[31] 杨一帆, 邵一淼, 施宇. 一种分位数的获取方法,设备及存储介质: CN202111153418[P]. 2021–09–29.
YANG Y F, SHAO Y M, SHI Y. A method, device and storage medium for obtaining quantiles: CN202111153418[P]. 2021–09–29.

[32] CRISTOFARO E, TSUDIK G. Practical private set intersection protocols with linear computational and bandwidth complexity[C]//Proceedings of the 14th International Conference on Financial Cryptography and Data Security. Heidelberg: Springer, 2010: 143–159.

[33] CRISTOFARO E, TSUDIK G. On the performance of certain private set intersection protocols[C]//Proceedings of the 5th International Conference on Trust & Trustworthy Computing. [S.l.:s.n.], 2012.
[34] FREEDMAN M J, NISSIM K, PINKAS B. Efficient private matching and set intersection[C]//Proceedings of the 2014 International Conference on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer, 2004: 1–19.
[35] GOOD I J. Weight of evidence: a brief survey[J]. Bayesian Statistics 1985(2): 249–270
[36] RODRIGUEZ-LUJAN I, HUERTA R, ELKAN C, et al. Quadratic programming feature selection[J]. The Journal of Machine Learning Research, 2010, 11(2): 1491–1516.
[37] JOHNSON T, DASU T. Data quality and data cleaning[C]//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2003: 681.
[38] 叶焕倬, 吴迪. 相似重复记录清理方法研究综述[J]. 现代图书情报技术, 2010(9): 56–66.
YE H Z, WU D. A survey of approximately duplicate data cleaning method[J]. New Technology of Library and Information Service, 2010(9): 56–66.
[39] 朱晓峰. 缺失值填充的若干问题研究[D]. 桂林: 广西师范大学, 2007.
ZHU X F. Studies on missing data imputation[D]. Guilin: Guangxi Normal University, 2007.
[40] 程开明. 统计数据预处理的理论与方法述评[J]. 统计与信息论坛, 2007, 22(6): 98–103.
CHENG K M. The theory and methods of data preparation: an overview[J]. Statistics & Information Forum, 2007, 22(6): 98–103.
[41] 贾俊平, 何晓群, 金勇进. 统计学(第六版)[M]. 北京: 中国人民大学出版社, 2015.
JIA J P, HE X Q, JIN Y J. Statistics[M]. Beijing: China Renmin University Press, 2015.
[42] LIPOVETSKY S, CONKLIN M. Analysis of regression in game theory approach[J]. Applied Stochastic Models in Business and Industry, 2001, 17(4): 319–330.
[43] ŠTRUMBELJ E, KONONENKO I. Explaining prediction models and individual predictions with feature contributions[J]. Knowledge and Information Systems, 2014, 41(3): 647–665.
[44] LUNDBERG S, LEE S I. A unified approach to interpreting model predictions[J]. arXiv preprint, 2017, arXiv:1705.07874.
[45] 汪云云, 陈松灿. 基于AUC的分类器评价和设计综述[J]. 模式识别与人工智能, 2011, 24(1): 64–71.
WANG Y Y, CHEN S C. A survey of evaluation and design for AUC based classifier[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(1): 64–71.
[46] 张义莲, 颜晟, 朱旻捷, 等. 机器学习系统毒化攻击综述[J]. 通信技术, 2020, 53(3): 535–542.
ZHANG Y L, YAN S, ZHU M J, et al. Overview on poisoning attacks against machine learning system[J]. Communications Technology, 2020, 53(3): 535–542.

作者简介

**张燕**(1985– ), 女, 星环信息科技(上海)股份有限公司人工智能研究员, 主要研究方向为隐私计算、可解释AI、因果分析等。

**杨一帆**(1985- ),男,博士,星环信息科技(上海)股份有限公司产品总监、首席科学家,主要研究方向为统计、图计算、强化学习等。

**伊人**(1989- ),女,博士,星环信息科技(上海)股份有限公司隐私计算首席科学家,主要研究方向为隐私计算、联邦学习、知识图谱等。

**罗圣美**(1971- ),男,博士,星环信息科技(上海)股份有限公司大数据研究院院长,主要研究方向为大数据、并行计算、云存储、人工智能等。

**唐剑飞**(1986- ),男,星环信息科技(上海)股份有限公司大数据技术标准研究员,主要研究方向为大数据、数据库、图计算等。

**夏正勋**(1979- ),男,星环信息科技(上海)股份有限公司高级研究员,主要研究方向为人工智能、大数据、数据库、流媒体处理技术等。

**收稿日期:** 2022-04-02
**通信作者:** 张燕, yan.a.zhang@transwarp.io