

基于人在回路的纵向联邦学习模型可解释性研究

李晓欢^{1,2}, 郑钧柏^{1,2}, 康嘉文³, 叶进², 陈倩^{1,4}

(1. 广西高校智能网联与场景化系统重点实验室(桂林电子科技大学信息与通信学院), 广西 桂林 541004;

2. 广西综合交通大数据研究院, 广西 南宁 530025;

3. 广东工业大学自动化学院, 广东 广州 510006;

4. 桂林电子科技大学建筑与交通工程学院, 广西 桂林 541004)

摘要: 纵向联邦学习(vertical federated learning, VFL)常用于高风险场景中的跨领域数据共享, 用户需要理解并信任模型决策以推动模型应用。现有研究主要关注VFL中可解释性与隐私之间的权衡, 未充分满足用户对模型建立信任及调优的需求。为此, 提出了一种基于人在回路(human-in-the-loop, HITL)的纵向联邦学习解释方法(explainable vertical federated learning based on human-in-the-loop, XVFL-HITL), 通过构建分布式HITL结构将用户反馈纳入VFL的基于Shapley值的解释方法中, 利用各参与方的知识校正训练数据来提高模型性能。进一步, 考虑到隐私问题, 基于Shapley值的可加性原理, 将非当前参与方的特征贡献值整合为一个整体展示, 从而有效保护了各参与方的特征隐私。实验结果表明, 在基准数据上, XVFL-HITL的解释结果具有有效性, 并保护了用户的特征隐私; 同时, XVFL-HITL对比VFL-Random和直接使用SHAP的VFL-Shapley进行特征选择的方法, 模型准确率分别提高了约14%和11%。

关键词: 纵向联邦学习; 可解释性; 人在回路; Shapley值

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202345

Research on the explainability of vertical federated learning models based on human-in-the-loop

LI Xiaohuan^{1,2}, ZHENG Junbai^{1,2}, KANG Jiawen³, YE Jin², CHEN Qian^{1,4}

1. Guangxi University Key Laboratory of Intelligent Networking and Scenario System (Guilin University of Electronic Technology), Guilin 54004, China

2. Guangxi Research Institute of Integrated Transportation Big Data, Nanning 530025, China

3. School of Automation, Guangdong University of Technology, Guangzhou 510006, China

4. School of Architecture and Transportation Engineering, Guilin University of Electronic Technology, Guilin 54004, China

Abstract: Vertical federated learning (VFL) is commonly used for cross-domain data sharing in high-risk scenarios. Users need to understand and trust model decisions to promote the application of models. Existing research primarily focuses on the trade-off between explainability and privacy within VFL, and fails to fully meet the needs of users for establishing trust and fine-tuning models. To address these issues, we proposed an explainable vertical federated learning method based on human-in-the-loop (XVFL-HITL), which incorporated user feedback into the VFL's Shapley value-based explainability approach through a distributed HITL structure, using the knowledge of all VFL participants to correct training data and enhance model performance. Furthermore, considering privacy concerns, this paper employed the additive principle of Shapley values to integrate the feature contribution values of all entities other than the target participant

收稿日期: 2023-09-12; 修回日期: 2023-12-07

通信作者: 陈倩, chenqian@mails.guet.edu.cn

基金项目: 国家自然科学基金项目(No. U22A2054); 广西科技重大专项(No. AA22068101)

Foundation Items: The National Natural Science Foundation of China (No. U22A2054), The Key Science and Technology Project of Guangxi (No. AA22068101)

into an aggregated measure, which effectively protected the feature privacy of each participant. Experimental results indicated that on benchmark data, the explainability results of XVFL-HITL were effective and could well protect the feature privacy of user. Additionally, compared to VFL-Random and VFL-Shapley, the model accuracy of XVFL-HITL improved by approximately 14% and 11%, respectively.

Key words: vertical federated learning, explainability, human-in-the-loop, Shapley value

0 引言

近年来,人工智能的迅猛发展推动了超五代移动通信(beyond 5G, B5G)^[1]、元宇宙(meta-universe)^[2]等多个技术的发展,这些新兴技术有望在医疗^[3-5]、交通^[6-7]、智慧城市^[8]及工业生产^[9-10]等多个领域应用。但受隐私保护需求和法律法规限制^[11-13],分布式系统中各参与方不愿共享本地数据,从而导致数据壁垒问题。联邦学习技术(federated learning, FL)的出现有效解决了数据壁垒问题,成为在保护隐私前提下实现数据共享的有效解决方案^[14]。纵向联邦学习(vertical federated learning, VFL)作为FL的一种类型,适用于样本空间相同但特征空间不同的场景^[15],广泛应用于医疗^[16]、金融^[17]和广告^[18]等领域。VFL在多个领域中的应用日益增多,但其决策依据对用户而言仍然是一个黑盒。了解VFL的决策依据可以为用户提供更丰富的信息支持并增强用户对模型的信任。可解释性方法的引入有助于人类理解黑盒模型^[19],这恰好是复杂的VFL模型所需要的,因此在VFL背景下研究模型可解释性具有重要意义。

机器学习的可解释性可以分为内在可解释性和事后可解释性^[20]。内在可解释性指模型本身具备较好的解释能力,通过参数或结构可以直观得到解释结果,线性模型、决策树等均属于此类模型^[21]。对于基于内在可解释性的设计,文献[22]提出了基于有界约束的逻辑回归VFL框架实现风险评估,但其受限于场景和具体模型种类。事后可解释性主要通过分析模型的输入与输出之间的关系理解模型决策,不直接依赖于模型内部参数或结构,主要解释对象为神经网络^[23],代表方法有LIME^[24]和SHAP^[25]。文献[26]提出了一种使用反事实分析的可解释VFL框架,为理解VFL提供了非常有价值的思路。实际上,由于博弈论中的Shapley值具有公平属性及完备的数学过程^[27],事后可解释性方法大多基于Shapley值思想衡量数据样本中每个特征在不同特征子集内的边际贡献,该方法以数据样

本为单位进行解释。文献[28]提出了一种特征分组方法,该方法将具有低解释性的原始特征转换为可解释的特征组,最后使用Shapley值的方式展示了转换后特征的重要性。文献[29]提出利用Shapley值在保障数据隐私的前提下解释联邦学习模型,为实际应用提供了重要指导。文献[30]在VFL中使用Shapley值来计算分组特征的重要性。文献[31]为解释LR-BA的有效性,使用SHAP方法解释恶意潜在表示如何影响顶层模型的预测。尽管上述研究在探讨VFL模型可解释性方面提供了宝贵的思路,但大多数工作都侧重于应用或者设计新的可解释性算法,以实现VFL中可解释性和隐私保护之间的权衡,忽视了可解释性算法以人为本的初衷,解释结果没有被用户更好地利用。因此将可解释性研究与人类行为相结合可以有效改善VFL的应用效果。

为了使可解释性算法在VFL模型和用户之间更好地建立链接,需要纳入用户反馈,即引入人在回路(human-in-the-loop, HITL)的思想。HITL被证明可以在学习和推断中改善数据分析,通过人类知识进一步提高模型性能^[32]。因此,HITL将支持用户依据解释结果进行更深入的操作,可以进一步加深用户对模型的理解和信任^[33]。文献[34]提出了一种将HITL应用于漏洞检测等任务的概念框架,通过HITL整合专家的决策以提高系统的性能。在可解释性方法中纳入HITL的大多数方法往往是集中式地采集用户反馈,不适用于VFL这种分布式架构。

为了解决以上问题,本文以算法可解释性的评价应以服务人的理解出发点,深入探讨了HITL在VFL可解释性框架中的指导作用,提出了一种基于人在回路的纵向联邦学习解释方法(explainable vertical federated learning based on human-in-the-loop, XVFL-HITL)来评估VFL模型的特征贡献,HITL的引入促进了用户对数据分析的理解和模型性能的提升。XVFL-HITL方法强调了用户反馈在训练数据调整和模型训练指导中的核心作用,体现了在提升模型可解释性方面的创新性和有效性。

本文的主要贡献为：首先，将可解释性引入VFL中，提出了具有可解释性的纵向联邦学习框架（explainable vertical federated learning framework, ExVFL），如图1所示；其次，通过构建分布式HITL结构将用户反馈纳入基于Shapley值的解释方法中，VFL的各参与方利用用户知识校正本地数据以提高模型性能；最后，为应对传统SHAP方法中汇聚特征贡献值后再进行解释评价导致的特征特性泄露的风险，利用Shapley值的可加性原理改进解释方法XVFL-HITL，将非当前参与方的特征贡献值整合为一个整体，有效地保护了各参与方的特征隐私。

1 可解释纵向联邦学习框架 ExVFL

本文考虑 M 个参与方 $P=\{P_1, P_2, \dots, P_M\}$ 协同训练VFL模型 Q 的场景。首先，各参与方数据均配备唯一样本标识符（ID）以确保数据对齐。各方对相同样本拥有不同的特征，通过ID确保特征与其对应的样本一致。因此各方本地样本数量均为 n 。接着，定义参与方 P_i 本地数据集为 $X^i=\{x_1^i, x_2^i, \dots, x_n^i\} \in \mathbb{R}^{n \times d_i}$ ，其中 x_j^i 表示 P_i 本地数据集 X^i 中第 j 个实例， d_i 表示 P_i 中样本的特征数量。定义参与方 P_i 本地数据的特征集合为 $F^i=\{f_1^i, f_2^i, \dots, f_{d_i}^i\} \in \mathbb{R}^{1 \times d_i}$ ，其中 f_j^i 为 P_i 本地数据集 X^i 中第 j 列的特征索引。因此，各参与方本地数据集为 $X=\{X^1, X^2, \dots, X^M\} \in \mathbb{R}^{n \times k}$ ，

其中 $k=\sum_{i=1}^M d_i$ 为总特征数，具体系统参数说明如表1所示。

表1 参数说明

参数	说明
M	参与方数量
P	参与方集合
P_i	集合 P 中第 i 个参与方
X	所有参与方本地数据集
n	每个参与方本地数据数量
k	所有参与方本地数据特征数量
X^i	P_i 的本地数据集
d_i	X^i 的特征数量
x_j^i	X^i 中第 j 个样本实例
F^i	X^i 的特征集合
f_j^i	X^i 中第 j 列的特征索引
Q	本地模型

为了更符合实际场景，进行如下约定。

(1) VFL中仅云端含标签集合 $Y=\{y_1, y_2, \dots, y_n\}$ ，各参与方仅有数据的特征项，且云端是诚实可信的。

(2) VFL中各方数据特征不能存在交集，即对于任意的 $P_i, P_j \in P$ ，如果 $i \neq j$ ，则 $F^i \cap F^j = \emptyset$ 。

(3) 为了适应复杂非线性数据的处理，本文选择了神经网络作为本地模型。

如图1所示，本文研究的可解释纵向联邦学习

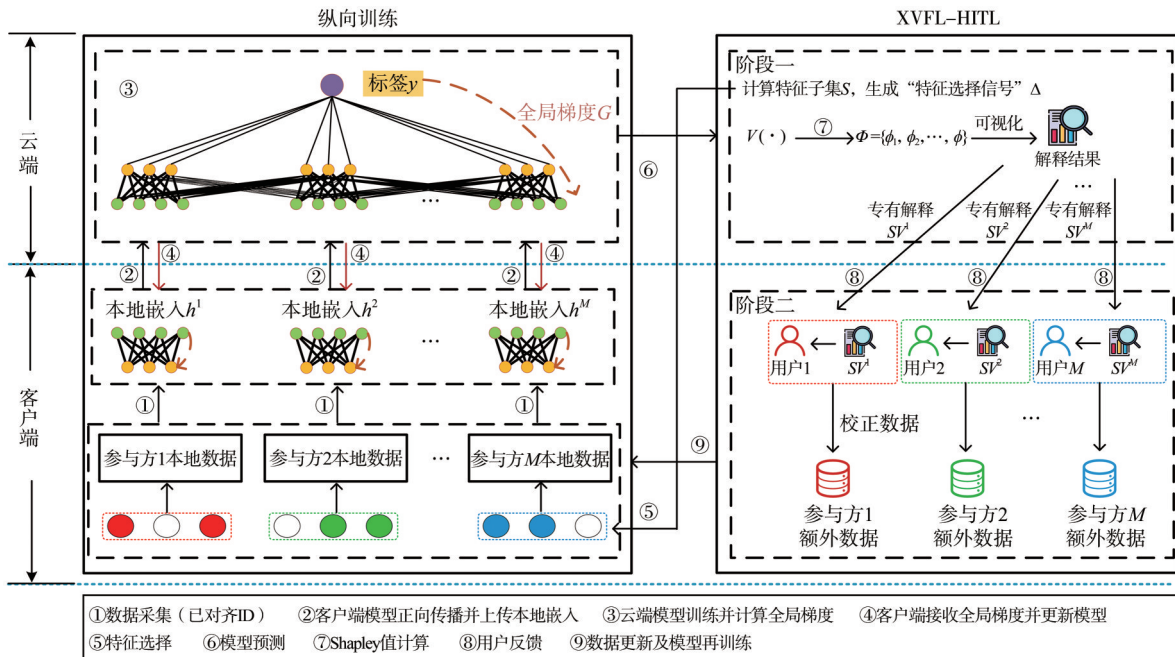


图1 可解释纵向联邦学习框架（ExVFL）

框架 ExVFL 中数据垂直分布在各参与方，各方协作训练模型并对其解释。框架主要包括纵向训练和模型解释两部分。以下是框架具体描述。

(1) 纵向训练：本部分主要聚焦于在 VFL 模型中进行数据训练。为了更好地适应 VFL 的训练过程和隐私保护需求且便于说明，本文以基于分裂神经网络^[35]的结构为例，进行 VFL 架构的设计，即模型 Q 为分裂神经网络。分裂神经网络的特点为将神经网络分割成底层和顶层模型，底层模型分布在每个参与方的客户端，各参与方利用本地数据在底层模型前向传播中计算出本地嵌入。顶层模型则部署在云端，云端收集并联合各本地嵌入形成全局嵌入，将全局嵌入作为顶层模型的输入进行前向传播，然后进行反向求导计算出顶层模型的模型参数及全局嵌入层的全局梯度。云端将全局梯度分割并广播给各参与方客户端对应的部分，各客户端根据各自的部分梯度使用反向求导更新各方的本地模型。重复上述过程，不断完成前向传播、反向求导，直到模型收敛。在该过程中，各参与方不需要共享彼此的原始数据，有效地保护了各方隐私。

进一步细化模型 Q 的设计，定义顶层模型参数为 θ_S ，各参与方 P_i 的客户端中底层模型参数为 $\Theta = \{\theta_1, \theta_2, \dots, \theta_M\}$ 。以单条样本 z 为例， P_i 的本地嵌入为 $h_z^i = Q(\theta_i; x_z^i)$ ，全局嵌入为 $H_z = [h_z^1, h_z^2, \dots, h_z^M]$ 。综上，对于基于分裂神经网络的 VFL 模型，其问题定义为最小化模型的损失函数，具体表示如下：

$$\min_{\Theta, \theta_S} \ell(\Theta, \theta_S; X, Y) = \frac{1}{n} \sum_{i=1}^n \ell(Q_S(\theta_S; H_i), y_i) + \lambda \sum_{j=1}^M \gamma(\theta_j, \theta_S) \quad (1)$$

其中， Q_S 为顶层模型， ℓ 为损失函数， γ 为正则化，防止模型过拟合， λ 是控制 γ 强度的超参数。

式 (1) 主要从顶层模型的角度展示 VFL 的训练过程。而对于各客户端的底层模型参数 θ_i ，则由顶层模型反向传播计算出的全局梯度 $G = \partial \ell / \partial H$ 更新。由于 H 由 M 个参与方的本地嵌入 h^i 联合而成， G 可被分解为与各本地嵌入相对应的部分梯度，即：

$$G = \left[\frac{\partial \ell}{\partial h^1}, \frac{\partial \ell}{\partial h^2}, \dots, \frac{\partial \ell}{\partial h^M} \right] \quad (2)$$

因此各方接收的本地梯度 g^i 为：

$$g^i = \frac{\partial \ell}{\partial h^i} \quad (3)$$

各客户端根据各自的本地梯度 g^i 继续反向求导

更新各方的底层模型。

(2) 模型解释：VFL 可解释性的目标是寻找一种能在 VFL 中最大化模型理解的解释方法 S ，旨在为 VFL 部分由式 (1) 训练好的模型 Q 提供解释。 S 表示各特征的贡献或者参与方在协同训练过程中的作用。具体来说，对于学习任务 T ，在给定数据集 X 的情况下，用户 User 能对该模型 Q 实现最大程度的理解，即效用函数 U 最大。具体表示如下：

$$\operatorname{argmax}_S U(S|Q, \text{User}, X, T) \quad (4)$$

$$T = \min_{\Theta, \theta_S} \ell(\Theta, \theta_S; X, Y) \quad (5)$$

其中， U 度量了 S 对 VFL 模型的解释性，学习任务 T 为式 (1) 的问题定义，解释方法 S 为各种解释方案。本文提出的解释方法 S 由两阶段实现。

第一阶段为 Shapley 值计算，主要进行 VFL 模型解释。云端首先获取特征数量，并统计出全部特征子集。对于各特征子集中缺失的特征采用 0 值作为基准值代替，这一过程中云端向客户端发送多个二进制向量作为特征选择信号来实现对本地样本特征的选择。然后云端获取各特征子集下模型的预测值并计算各特征的贡献值。

特征贡献的计算采用 SHAP 的基本思路，但考虑到直接汇聚特征计算 Shapley 值，容易导致特征特性泄露。具体来说，即所有参与方均从云端接收相同的解释结果，那么当前参与方能够识别到其他参与方的特征 ID 及其对应的特征贡献值。因此，存在被其他参与方推断出特征特性的可能，从而导致隐私泄露。为了更好地保护隐私，本文基于 Shapley 值的可加性^[25]，进一步将同一参与方特征的贡献值累加为一个整体值，并发送给其他参与方。累加后的特征整体称为“封装特征”。最终云端将为每个参与方生成专有可解释性结果 SV^i ，当前参与方的本地特征会具体显示特征贡献，而其他参与方的特征将被封装为一个整体并返回一个总体特征贡献值。进一步，云端聚合相同参与方特征的贡献值并将其封装为一个整体值，即“封装特征”的贡献值。最后，云端广播给每个参与方包含其他参与方封装特征的专有解释结果，避免了特征名称及对应贡献值的泄露。

第二阶段为用户反馈，各参与方的用户将并行对解释结果进行反馈，进一步调优模型。云端将解释样本的预测结果 y 和对应各参与方专有解释结果

SVⁱ发送给各参与方。参与方用户将基于专业知识和解释结果进行反馈，反馈涉及本地样本的特征调整，例如特征项的剔除或特征值的调整，或者提供反馈评价，告知云端预测结果的准确性。用户反馈将用作训练数据校正以形成新的数据样本。新的数据样本将替换旧的样本，并整合到本地数据集中形成新一轮的训练集。然后各方在规定时间内完成数据更新后，再次进行VFL的训练以更新模型。

算法具体流程如下。

①数据对齐：使用私有集合交集（private set intersection, PSI）对齐各方数据，使各参与方拥有相同ID和不同特征。

②客户端模型正向传播：对齐各参与方数据后，各参与方对其客户端模型使用本地数据完成前向传播过程，得到本地嵌入。各参与者将其本地嵌入传输给云端。

③云端模型训练：云端联接各本地嵌入，形成全局嵌入。将全局嵌入输入云端模型，完成在云端模型中的正向传播。然后云端模型反向求导计算梯度至云端模型的输入层，得到全局梯度。

④全局梯度广播：云端模型根据各参与方嵌入的维度对全局梯度进行分割，得到各参与方的本地梯度。然后将本地梯度发送给各参与方客户端，各参与方基于客户端模型使用全局梯度反向传播更新模型参数。

重复步骤②~④，直到模型收敛。

⑤特征选择：云端确定特征集合并计算所有特征子集。云端根据特征子集为各参与方发送特征选择信号，指导各参与方在本地数据中进行特征选择。子集中存在的特征使用真实值，缺失的特征使用基线值。

⑥样本预测值：云端获取在不同特征子集下模型对样本的预测值。

⑦Shapley值计算：云端对不同特征子集下样本输出值进行统计，计算各特征的贡献值，并为各参与方计算出“封装特征”的贡献值。然后云端为每个参与方生成一个专有的解释结果，该结果只包括其本地特征和其他参与方的“封装特征”的贡献值。

⑧用户反馈：各参与方的用户根据云端广播的解释结果结合专业经验校正解释样本。然后将新样本作为额外数据替换旧数据，并与其余原始样本合并，形成新的数据集。

⑨数据更新与模型再训练：各参与方利用更新后的数据集在已收敛模型的基础上再次对模型训练，即重复步骤②~④，直到模型再次收敛，更新后模型准确率将获得提升。

2 XVFL-HITL解释方法设计

定义各参与方总特征集合为 $F=\{F^1, F^2, \dots, F^M\}$ 。为了衡量 F 中各特征的影响，使用Shapley值计算各特征的贡献值， F 中全部特征将被视为参与博弈的玩家。因此，待解释样本 z 中特征 j 的Shapley值 $\phi_j(z)$ 如下：

$$\phi_j(z) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [V_z(S \cup \{j\}) - V_z(S)] \quad (6)$$

其中， $|F|$ 为样本特征的数量， $S \subseteq F \setminus \{j\}$ 表示从总特征集合 F 中移除特征 j ，而 S 是不含特征 j 的 F 的特征子集，即 S 是 F 中除了特征 j 之外特征的任意组合。 $V(\cdot)$ 为特征集合内特征参与博弈所获取的奖励值，在此为模型 Q 的预测值。 $V_z(S)$ 表示 S 中特征取真实值， S 外特征取基线值时模型的预测值。本文将基线值设置为与解释样本同维度的全0向量。 $V_z(S \cup \{j\})$ 表示当特征 j 加入 S 后模型的预测值。因此解释样本 z 中各特征贡献值为 $\Phi_z = \{\phi_1(z), \phi_2(z), \dots, \phi_i(z)\}, i \in F$ 。该特征贡献值向量将作为模型对 z 的解释结果。

为遵循VFL的数据隐私原则，确保数据不出本地，本文设计了“特征选择信号”机制，允许在不传输本地数据的情况下对样本进行解释。本文定义特征选择信号为 $\Delta = \{\delta_1, \delta_2, \dots, \delta_M\}$ ，其中每个 δ_i 是与对应参与方 P_i 本地特征维度相匹配的二进制向量。首先，云端根据总特征集合 F 确定所有可能的特征子集。当针对特定的特征子集进行计算时，云端会分析该子集中哪些特征是缺失的，并基于这一分析来设置 δ_i 。随后，云端将根据当前的子集需求向各参与方发送对应的 δ_i ，引导其选择特征：

$$X^{i'} = X^i \odot \delta_i \quad (7)$$

其中， $\odot$ 表示哈达马积，向量 δ_i 中元素与参与方 P_i 中样本特征值对应相乘，以选择性地“激活”或“禁用”某些特征，当 δ_i 中的某个元素为1时，对应的 X^i 中的特征值保持不变，当 δ_i 中的某个元素为0时，对应的 X^i 中的特征值被置为0。这样的机制可以有效地按需选择特征，而不需要在各参与方之间传输原始数据，从而保护了数据隐私。各参与方使用 $X^{i'}$ 在本地执行前向传播，然后将中间嵌入

结果传输到云端，继续前向传播，从而得到模型的预测值。云端保存每个特征子集 S 下的模型预测值，并根据式 (6) 计算解释样本 z 中所有特征的贡献值。值得注意的是，特征贡献值的计算可以通过仅一轮通信实现。云端预先为所有可能的特征子集情况计算 δ_i ，并一次性将其发送给各个参与方。然后，各参与方在本地批量处理数据，并将中间结果统一返回给云端。云端计算出所有特征的特征贡献值后，根据参与方的数量及其本地包含的特征情况，将其包含特征的特征贡献值相加并保存为一个整体，即对应每个参与方的“封装特征”。每个参与方只能访问到本地特征的具体贡献值，对于其他参与方的特征贡献值，只能显示对应参与方的“封装特征”贡献值。这样设计旨在保护每个参与方的数据隐私，同时允许他们获取足够的信息理解模型，并参考解释结果对模型调试。

上述过程实现了为多参与方提供包含不同“封装特征”的专有解释结果 SV^i 。为增强人类的表现和感知，需引入用户反馈替代自动化且不透明的人工智能系统，用户反馈可以用于交互过程中的控制或作为评价反馈，以增强用户对系统的信任，并提升系统的性能^[36-37]。为更好地利用解释结果以提升模型性能，进一步引入分布式的 HITL，各参与方用户根据解释结果处理本地训练数据，实现对模型的进一步调优。用户参考解释结果后使用 HITL 重新编码分类标签，生成新的人工修订标签并替换到原始数据集中。因此，定义参与方 P_i 被解释的样本集合为 EX^i ，解释结果经过用户反馈后 EX^i 被校正得到新样本集合 CX^i ，具体表示如下：

$$CX^i = \text{HITL}(EX^i | SV^i) \quad (8)$$

其中，用户反馈 $\text{HITL}(\cdot)$ 具体设计如下：云端首先获取高置信度且预测结果和实际标签不一致的样本，将这些样本的专有解释结果分别发送给各参与方，参与方用户根据解释结果生成反馈信号 μ^i ， μ^i 为一维二进制向量。如果解释结果符合某方用户的认知，则发送值为1的反馈信号，云端收集所有用户的评价，如果超过50%用户的反馈信号值为1，则认为模型的预测是正确的，将对原始数据的标签进行修正，反之则不需要修正原始数据的标签。然后各方将从本地原始数据集中剔除样本集合 EX^i ，并将新样本集合 CX^i 整合到原始数据集中，形成新的扩展数据集 D^i ，具体表示如下：

$$D^i = \{X^i \setminus EX^i, CX^i\} \quad (9)$$

其中， $X^i \setminus EX^i$ 表示从 X^i 中移除 EX^i 后剩下的样本集合。各参与方同步进行数据更新，更新后的总数据集表示为 $D = \{D^1, D^2, \dots, D^M\}$ 。在保护隐私的前提下，本文将 HITL 纳入可解释性方法，实现了模型性能的提升。因此，VFL 可解释性的最终目标将变为提升模型性能，本文以模型准确率为例进行说明。即式 (4) 中的效用函数 U 被设置为模型准确率 Acc ，式 (4) 重写如下：

$$\begin{aligned} & \underset{Q, D, \Theta, \theta_s}{\operatorname{argmax}} \quad \text{Acc}(\text{XVFL} - \text{HITL}) \\ & \min_{\Theta, \theta_s} \ell(\Theta, \theta_s; D, Y) \end{aligned} \quad (10)$$

综上所述，XVFL-HITL 通过重新设置“封装特征”为各参与方生成专有的解释结果，在保护隐私的情况下实现了对 VFL 模型的解释。进一步，通过引入 HITL 对解释结果反馈，在对训练数据校正后能有效提升模型的准确率。XVFL-HITL 算法伪码如下。

输入：待解释样本索引 z ，参与方数量 M ，参与方总特征集合 $F = \{F^1, F^2, \dots, F^M\} = \{f_1^1, \dots, f_1^2, \dots, f_1^M, \dots, f_{dM}^M\}$ ，解释样本总维度 k ，各参与方本地样本维度 d_i ；

输出：待解释样本 z 中各特征 Shapley 值 $\phi_i(z)$ ，校正后的数据集 D^i ；

- 1: 初始化：特征子集 $S = \emptyset$ ，特征选择信号 Δ 为零向量集合，反馈信号 $\mu^i = 0$ ；
- 2: 计算所有特征子集并存入 S 中；
- 3: 根据 S 中特征缺失情况，生成 Δ ；
- 4: **for** $j=1, \dots, f_{dM}^M$ **do**
- 5: **for** $i=1, \dots, M$ **do**
- 6: 云端发送 δ_i 给对应参与方 P_i ；
- 7: P_i 对 z 特征选择得到 $X^{i'}$ (式 (8))；
- 8: $X^{i'}$ 前向传播得到本地嵌入 h_z^i ；
- 9: 云端接收各参与方本地嵌入，并聚合为全局嵌入 $H_z = [h_z^1, h_z^2, \dots, h_z^M]$ ；
- 10: **end**
- 11: 全局嵌入 H_z 作为云端本地模型输入得到预测结果 $V_z(\cdot)$ ；
- 12: 计算特征 j 的特征贡献值 $\phi_j(z)$ (式 (6))；
- 13: **end**
- 14: **for** $i=1, \dots, M$ **do**
- 15: 云端将 P_i 拥有特征的特征贡献值累加并保存为“封装特征 i ”；

```

16: end
17: for  $i=1, \dots, M$  do
18:   云端整合  $P_i$  的特征及其他方“封装特征”的特征贡献值, 作为专有解释  $SV^i$  发送给  $P_i$ ;
19:    $P_i$  根据  $SV^i$  评估特征贡献值对标签的影响, 并生成反馈信号  $\mu^i$ ;
20: end
21: 云端接收各参与方反馈信号  $\{\mu^1, \dots, \mu^M\}$ ;
22: if 值为1的反馈信号数量/值为0的反馈信号数量  $\geq 0.5$ :
23:   原始标签翻转;
24: else
25:   原始标签不变;
26: 将校正后的数据样本替换原始样本(式(9))。

```

3 实验与分析

3.1 实验设置

(1) 数据集: VFL 研究中常用数据集大多是来自金融和医疗领域的表格数据集, 以便于在 VFL 中形成多方场景。本文采用公开数据集 Adult^[38], 该数据集是一个经典的二分类数据集, 用于预测一个人的收入是否超过 5 万美元。该数据集包含了一些与人口统计学相关的特征, 例如年龄、受教育程度、婚姻状况和职业等。数据集共有 48 842 个样本, 其中包含 14 个特征和一个二元类别标签, 其中特征“Education”与“Education-Num”是映射关系, 因此在数据预处理时剔除特征“Education”。在特征划分方面, 本文依据特征内容的性质进行了合理的分配。将基本人口统计信息相关特征, 如“Age”“Sex”“Education-Num”和“Marital Status”分配给参与方 1, 这些特征通常作为个体基本属性; 将家庭及社会背景等信息相关特征, 如“Relationship”“fmlwgt”“Country”和“Race”分配给参与方 2, 这类特征反应了个体的社会属性; 将个人工作及经济状况等信息相关特征, 如“Occupation”“Capital Gain”“Capital Loss”“Workclass”和“Hours per week”分配给参与方 3, 这些特征通常与个体经济状况直接相关。本文使用 80% 的数据进行训练, 其余数据用于测试。

(2) 模型设置: 本文在 3 个客户端及单个中心服务器的 VFL 环境模拟多个参与方, 以 3 个参与方为例。假设样本 ID 已经在各方对齐, 在 3 个客户端

上分离特征, 使用 PyTorch 实现多客户端的 VFL 环境, 各客户端共同训练一个四层全连接神经网络模型。模型分为两部分, 各参与建模方客户端处理前半部分网络, 即底层网络。模型发起方(服务器)负责后半部分网络, 即顶层网络。底层网络和顶层网络均为两个全连接层, 分别是 32、16 和 16、8 的神经元设置。另外, 设置激活函数为 ReLU, 优化器使用 ADAM, 学习率为 0.001, VFL 训练周期为 100。

(3) 环境配置: 实验在一台搭载 Intel(R) Core (TM) i7-10875H CPU @ 2.30GHz 处理器、16 GB 内存和 NVIDIA GeForce RTX 2070 显卡的计算机上进行, 实验环境为 Python 3.10。

3.2 实验分析

(1) 可解释性分析

本文提出的 XVFL-HITL 作为计算样本实例中特征重要性的方法适配了 VFL 过程, 有效保护了各方隐私。XVFL-HITL 可以为多个参与者揭示具有隐私保护的贡献值结果。图 2 通过代表性样例展示了 XVFL-HITL 计算的贡献值结果, 其中 Shapley 值为正时, 表示该特征对模型预测产生促进作用, 即增加该特征值会导致预测值的提升; 相应地, Shapley 值为负时, 表示抑制作用, 即特征值的增加会降低模型的预测值。但不论正负, Shapley 值的绝对值大小均可视为该特征对预测结果的贡献程度。图 2 展示了各参与方视角下, 接收到 Adult 数据集中第 300 条样本的解释结果热力图。可以看出, 对于参与方 1, 只有本地特征能显示具体的特征贡献值, 而其余参与方则显示为“参与方 2 封装特征”“参与方 3 封装特征”这样的整体特征贡献值, 而整体贡献是该参与方本地特征贡献值的累加。对于参与方 2 和参与方 3 也是同样的效果。这种方法防止了具体特征项信息的泄露, 避免了特征攻击等安全隐患。

图 3 则展示了 1 000 条样本下的解释结果。为了更清晰地比较, 仅展示云端尚未进行特征封装时的总解释结果。通过对每条样本中对应特征的特征贡献值取绝对值并平均, 可以看出特征“fmlwgt”和“Capital Loss”的贡献是最小的。“fmlwgt”在原始数据集中表示人口普查员的编号, 是一个与任务无关的标识符, “Capital Loss”在数据集的所有样本中几乎都为 0, 基本没有信息量。这样的观察结果证实了 XVFL-HITL 的有效性, 同时也为特征选择提供了有价值的见解。

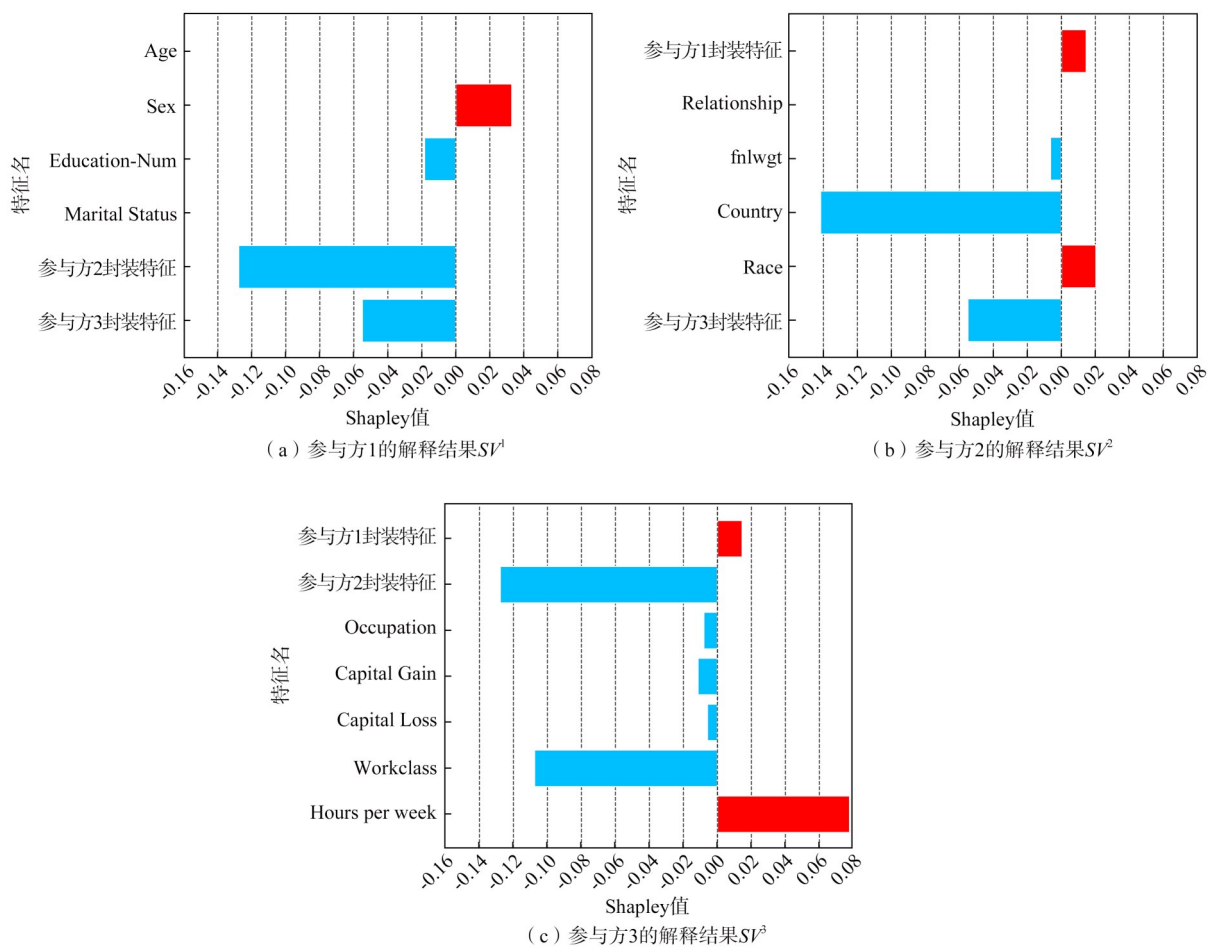


图2 XvFL-HITL为各参与方提供专有解释结果

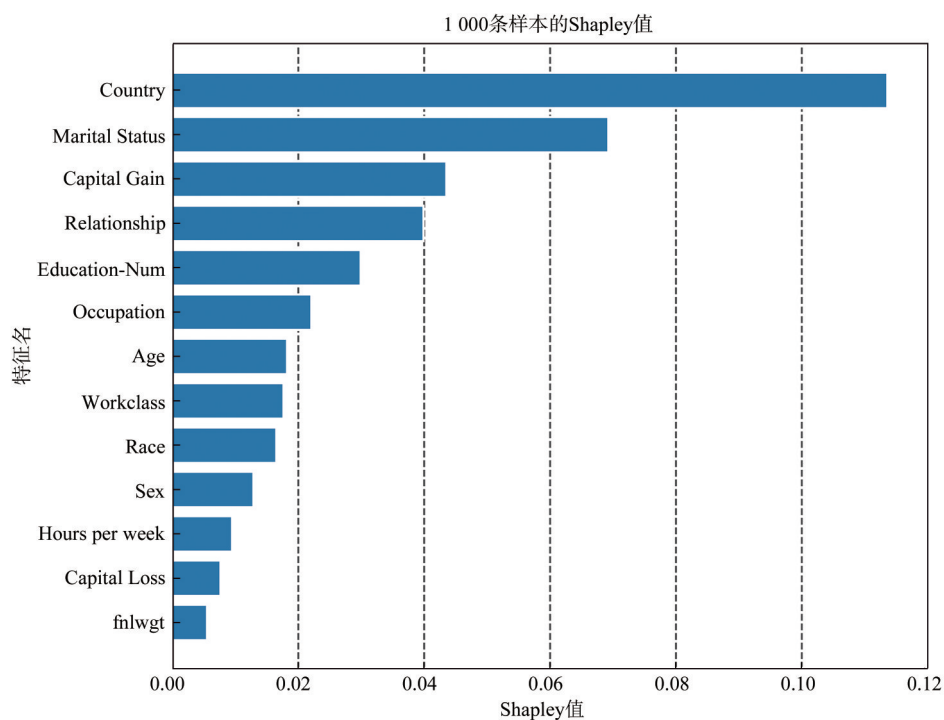


图3 1 000条样本下的各特征贡献值

(2) 模型准确率分析

本部分主要分析了XVFL-HITL对VFL模型性能的影响。对比了随机剔除特征项的模型训练(VFL-Random)、VFL、基于Shapley值进行特征项剔除模型训练(VFL-Shapley)和XVFL-HITL这4种方法在Adult数据集下的准确率和损失,以证明XVFL-HITL对VFL模型准确率提升的作用。

为了更清晰地说明HITL如何具体实施并影响模型性能,提供以下例子:考虑一个样本,其特征明显指向高收入群体,例如,该样本表示一个具有高教育程度和高级职位的个体,其原始标签是“ ≤ 5 万美元”。然而,模型对该样本的预测结果是“ > 5 万美元”,并且置信度非常高。这种情况下,用户有理由怀疑原始标签可能是错误的。因此,用户在观察到解释结果中“Education-Num”和“Occupation”的高影响后,可以判定原始标签是错误的,便对其进行标签翻转。另外,在这一过程中如果判别某些特征值可能是不准确的,还可以利用用户知识修正样本的特征值。通过这种方式,HITL允许用户纠正可能的标签错误或不准确特征值,从而提高模型的性能。

图4展示了4种方法下VFL模型在Adult数据集下的准确率对比结果。VFL模型准确率为83.67%,VFL-Random处理后模型准确率为80.65%,VFL-Shapley下的模型准确率为83.69%,XVFL-HITL处

理后模型的准确率为94.99%。VFL-Random方法是随机进行特征选择的,可能会剔除与标签关联性较强的特征,因此结果显然会变差。VFL-Shapley则依据解释结果进行特征选择,云端自动依据特征贡献值剔除无关特征“fnlwgt”“Capital Loss”,这样的操作虽然损失了一部分信息,但这些特征的低贡献值意味着它们对模型性能的影响有限,而更重要的是,减少数据的稀疏性还有助于加强其他特征与标签之间的关联性。因此,模型的性能反而会稍有提升。在XVFL-HITL中,用户将参考解释结果对训练数据进行优化。首先用户依据应用场景规定模型设置置信度区间,然后根据收到的解释结果分析高置信度中误分类样本的原始标签是否错误,最后利用更新的训练集在已训练好的模型基础上重新进行模型训练。这样的处理模式纳入了人的反馈对原始数据校正,更正了错误的训练数据,剔除了异常值。实验结果表明,对比传统随机特征选择方案VFL-Random和依据Shapley值自动进行特征选择方案VFL-Shapley,XVFL-HITL的模型准确率分别提高了约14%和11%,说明了将HITL纳入可解释性方法中能显著提升模型的性能。

图5展示了4种方法下VFL模型在Adult数据集下的模型损失对比结果。损失曲线反映模型收敛性,其中XVFL-HITL方法得到的模型损失最小。这意味着XVFL-HITL使模型的预测结果更接近真实标

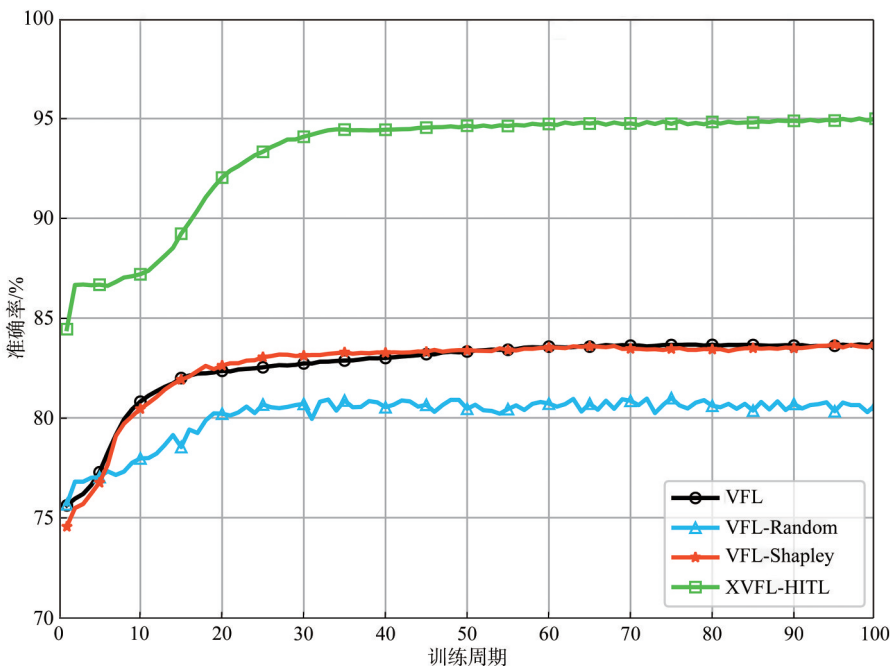


图4 不同方法下模型准确率对比

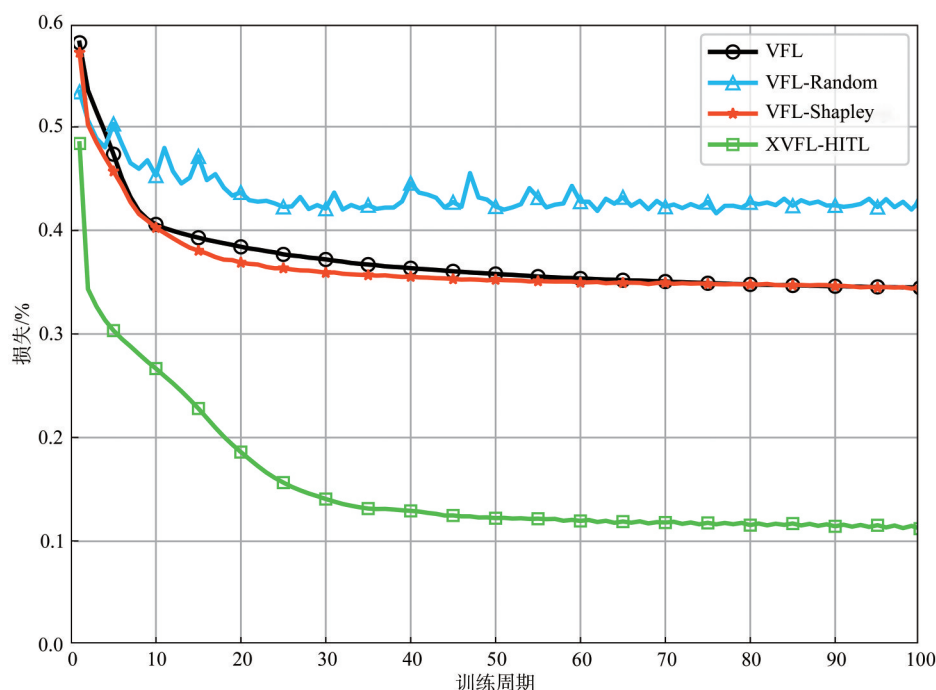


图5 不同方法下模型损失对比

签。这是因为XVFL-HITL通过用户反馈来重新调整训练数据，模型吸收额外的先验知识，从而更好地引导训练过程。

综上所述，XVFL-HITL能在保护用户隐私的前提下为各参与方提供模型解释。相比传统的VFL-Random和直接使用SHAP的VFL-Shapley，XVFL-HITL的模型准确率分别提高了约14%和11%。

4 结束语

本文提出了基于人在回路的纵向联邦学习解释方法XVFL-HITL，旨在提高VFL的模型性能及用户对模型的理解，推进VFL模型在实际高风险场景中的应用。实验结果表明，XVFL-HITL在保护数据隐私的同时利用用户反馈对训练数据进行校正，相较于使用VFL-Random和VFL-Shapley进行特征选择，模型准确率分别提高了约14%和11%。综上，XVFL-HITL的解释结果具有有效性且更好地保护了用户的特征隐私，同时还能VFL的参与者提供更高的模型准确性。

参考文献：

[1] XU W, YANG Z H, NG D W K, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing[J]. IEEE Journal of Selected Topics in Signal Processing, 2023, 17(1): 9-39.

[2] DU H Y, MA B H, NIYATO D, et al. Rethinking quality of experience for metaverse services: a consumer-based economics perspective[J]. IEEE Network, 2023(99): 1-8.

[3] BHARDWAJ R, NAMBIAR A R, DUTTA D. A study of machine learning in healthcare[C]//Proceedings of 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). Piscataway: IEEE Press, 2017: 236-241.

[4] MALEKLOO A, OZER E, ALHAMAYDEH M, et al. Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights[J]. Structural Health Monitoring, 2022, 21(4): 1906-1955.

[5] ABDULLAH, ALANAZI. Using machine learning for healthcare challenges and opportunities[J]. Informatics in Medicine Unlocked, 2022, 30: 100924.

[6] YANG Y, He K, WANG Y P, et al. Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods[J]. Physica A: Statistical Mechanics and Its Applications, 2022, 595: 127083.

[7] ABU AL-HAIJA Q, KRICHEN M, ABU ELHAIJA W. Machine-learning-based darknet traffic detection system for IoT applications[J]. Electronics, 2022, 11(4): 556.

[8] SALEEM M, ABBAS S, GHAZAL T M, et al. Smart cities: fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques[J]. Egyptian Informatics Journal, 2022, 23(3): 417-426.

[9] FERNANDES M, CORCHADO J M, MARREIROS G. Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review[J]. Applied Intelligence, 2022, 52(12): 14246-14280.

[10] GAROUANI M, AHMAD A, BOUNEFFA M, et al. Towards big industrial data mining through explainable automated machine learning[J]. The International Journal of Advanced Manufacturing Technology, 2022,

- 120(1/2): 1169-1188.
- [11] YANG F, MUELLER M L. Internet governance in China: a content analysis[J]. Chinese Journal of Communication, 2014, 7(4): 446-465.
- [12] MULDER T, TUDORICA M. Privacy policies, cross-border health data and the GDPR[J]. Information & Communications Technology Law, 2019, 28(3): 261-274.
- [13] 陆康, 刘慧, 任贝贝, 等. 智慧图书馆用户数据隐私保护研究: 基于《中华人民共和国网络安全法》和《一般数据保护条例》的文本启示[J]. 图书馆理论与实践, 2020(3): 17-21.
- LU K, LIU H, REN B B, et al. Research on user data privacy protection of smart libraries: based on the 《enlightenment of cybersecurity law of the people's republic of China》 and 《general data protection regulation》[J]. Library Theory and Practice, 2020(3): 17-21.
- [14] 杨强, 童咏昕, 王晏晟, 等. 群体智能中的联邦学习算法综述[J]. 智能科学与技术学报, 2022, 4(1): 29-44.
- YANG Q, TONG Y X, WANG Y S, et al. A survey on federated learning in crowd intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(1): 29-44.
- [15] 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.
- CHEN J Y, LI R C, HUANG G H, et al. Survey on vertical federated learning: algorithm, privacy and security[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 1-20.
- [16] VEPAKOMMA P, GUPTA O, SWEDISH T, et al. Split learning for health: distributed deep learning without sharing raw patient data[EB]. arXiv preprint, 2018, arXiv:1812.00564.
- [17] LONG G D, TAN Y, JIANG J, et al. Federated Learning for Open Banking[M]//YANG Q, FAN L, YU H. Federated Learning. Cham: Springer, 2020: 240-254.
- [18] ZHOU J, LI X L, ZHAO P L, et al. KunPeng: parameter server based distributed learning systems and its applications in alibaba and ant financial[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1693-1702.
- [19] MOLNAR C. Interpretable Machine Learning[M]. North Carolina: Lulu Press, Inc., 2020.
- [20] DAS A, RAD P. Opportunities and challenges in explainable artificial intelligence (xai): a survey[EB]. arXiv preprint, 2020, arXiv:2006.11371.
- [21] FREITAS A A. Comprehensible classification models[J]. ACM SIGKDD Explorations Newsletter, 2014, 15(1): 1-10.
- [22] ZHENG F, LI K, TIAN J, et al. A vertical federated learning method for interpretable scorecard and its application in credit scoring[EB]. arXiv preprint, 2020, arXiv:2009.06218.
- [23] MONTAVON G, SAMEK W, MÜLLER K R. Methods for interpreting and understanding deep neural networks[J]. Digital Signal Processing, 2018, 73: 1-15.
- [24] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust You?": explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144.
- [25] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 4768-4777.
- [26] CHEN P, DU X, LU Z, et al. EVFL: an explainable vertical federated learning for data-oriented artificial intelligence systems[J]. Journal of Systems Architecture, 2022, 126: 102474.
- [27] DUBEY P. On the uniqueness of the shapley value[J]. International Journal of Game Theory, 1975, 4(3): 131-139.
- [28] KANG Y, HE Y, LUO J, et al. Privacy-preserving federated adversarial domain adaptation over feature groups for interpretability[J]. IEEE Transactions on Big Data, 2022(99): 1-2.
- [29] WANG G. Interpret federated learning with shapley values[EB]. arXiv preprint, 2019, arXiv:1905.04519.
- [30] 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 (Big Data). Piscataway: IEEE Press, 2019: 2597-2604.
- [31] GU Y, BAI Y. LR-BA: backdoor attack against vertical federated learning using local latent representations[J]. Computers & Security, 2023: 103193.
- [32] WU X, XIAO L, SUN Y, et al. A survey of human-in-the-loop for machine learning[J]. Future Generation Computer Systems, 2022, 135: 364-381.
- [33] 张俊, 许沛东, 陈思远, 等. 物理-数据-知识混合驱动的人机混合增强智能系统管控方法[J]. 智能科学与技术学报, 2022(4): 571-583.
- ZHANG J, XU P D, CHEN S Y, et al. A hybrid physics-data-knowledge driven approach for human-machine hybrid-augmented intelligence-based system management and control[J]. Chinese Journal of Intelligent Science and Technology, 2022(4): 571-583.
- [34] NGUYEN T N, CHOO R. Human-in-the-loop XAI-enabled vulnerability detection, investigation, and mitigation[C]//Proceedings of 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE). Piscataway: IEEE Press, 2021: 1210-1212.
- [35] ROMANINI D, HALL A J, PAPADOPOULOS P, et al. PyVertical: a vertical federated learning framework for multi-headed splitnn[EB]. arXiv preprint, 2021, arXiv:2104.00489.
- [36] TSIKAKIS K, MURRAY-RUST D. Using human-in-the-loop and explainable AI to envisage new future work practices[C]//Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments. New York: ACM, 2022: 588-594.
- [37] VALTONEN L, MÄKINEN S J. Human-in-the-loop: explainable or accurate artificial intelligence by exploiting human bias? [C]//Proceedings of 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference. Piscataway: IEEE Press, 2022: 1-8.
- [38] KOHAVI R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid[C]//Proceedings of KDD. [S.l.:s.n.], 1996: 202-207.

[作者简介]



李晓欢 (1982-), 男, 博士, 桂林电子科技大学信息与通信学院教授、博士生导师, 主要研究方向为智能计算、工业物联网和空天地网络等。



郑钧柏 (1997-), 男, 桂林电子科技大学信息与通信学院硕士生, 主要研究方向为可解释性机器学习和联邦学习等。



康嘉文（1989-），男，博士，广东工业大学自动化学院青年百人教授，主要研究方向为隐私保护、区块链和工业物联网等。



陈倩（1984-），女，硕士，桂林电子科技大学信息与通信学院副教授，主要研究方向为物联网、交通大数据和智能计算等。



叶进（1970-），女，博士，广西综合交通大数据研究院教授、博士生导师，主要研究方向为网络协议设计和智能计算等。