联邦忘却学习研究综述

王鹏飞[1),2)] 魏宗正[1),2)] 周东生[3)] 宋 威[1),2)] 肖蕴明[4)]
孙 庚[5),6)] 于 硕[1),2)] 张 强[1),2)]

[1)](大连理工大学计算机科学与技术学院 辽宁 大连 116024)
[2)](大连理工大学社会计算与认知智能教育部重点实验室 辽宁 大连 116024)
[3)](大连大学先进设计与智能计算教育部重点实验室 辽宁 大连 116622)
[4)](美国西北大学计算机科学系 埃文斯顿 60208 美国)
[5)](吉林大学计算机科学与技术学院 长春 130012)
[6)](吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 数据已经成为与土地、劳动力、资本、技术等并列的重要生产要素之一.利用数据分析挖掘数据的潜在价值,有助于推动产业创新、技术升级和区域经济发展.然而,在数据使用过程中,隐私泄露等风险限制了数据的流通和共享.因此,如何在数据流通和共享过程中保护数据隐私已成为研究热点.联邦忘却学习(Federated Unlearning)撤销用户数据对联邦学习模型的训练更新,可以进一步保护联邦学习用户的数据安全.本文综述了联邦忘却学习的研究工作,首先简要阐述了联邦学习架构,并引出忘却学习和联邦忘却学习的概念和定义;其次,根据修正对象的不同将联邦忘却学习算法分为面向全局模型和面向局部模型两类,并详细分析各类算法的实现细节以及优缺点;然后,本文还详述联邦忘却学习中常用评价指标,将评价指标划分为模型表现指标、遗忘效果指标和隐私保护指标三类,并分析不同类型评价指标的优缺点;最后,本文对联邦忘却学习未来的研究方向进行展望.

关键词 联邦学习;联邦忘却学习;数字经济;隐私保护;边缘智能
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.00396

A Survey on Federated Unlearning

WANG Peng-Fei[1),2)] WEI Zong-Zheng[1),2)] ZHOU Dong-Sheng[3)] SONG Wei[1),2)]
XIAO Yun-Ming[4)] SUN Geng[5),6)] YU Shuo[1),2)] ZHANG Qiang[1),2)]

[1)](*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)
[2)](*Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, Dalian, Liaoning* 116024)
[3)](*Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education), Dalian University, Dalian, Liaoning* 116024)
[4)](*Department of Computer Science, Northwestern University, Evanston* 60208 *USA*)
[5)](*School of Computer Science and Technology, Jilin University, Changchun* 130012)
[6)](*Key Laboratory of Symbolic Computing and Knowledge Engineering (Ministry of Education), Jilin University, Changchun* 130012)

收稿日期:2023-04-20;在线发布日期:2023-11-28.本课题得到国家重点研发计划(2021ZD0112400)、国家自然科学基金联合基金项目(U1908214)、国家自然科学基金青年项目(62202080)、中国博士后科学基金面上项目(2023M733354)、中央高校基本科研业务费(DUT23YG122)资助.**王鹏飞**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为分布式人工智能、大数据智能处理,E-mail:wangpf@dlut.edu.cn.**魏宗正**,硕士研究生,主要研究方向为联邦学习.**周东生**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为智能计算、计算机图形学及视觉.**宋 威**,硕士研究生,主要研究领域为联邦学习.**肖蕴明**,博士研究生,主要研究领域为边缘计算、网络测量.**孙 庚**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为群体智能、协作通信.**于 硕**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为数据科学、图学习,数据安全.**张 强**(通信作者),博士,长江学者特聘教授,国家杰出青年科学基金获得者,中国计算机学会(CCF)高级会员,主要研究领域为新一代人工智能、生物智能计算.E-mail:zhangq@dlut.edu.cn.

Abstract Data has become an important factor of production alongside land, labor, capital, technology, etc. By leveraging data analysis to mine potential value, we can uncover profound insights into consumer behavior, market trends, and production efficiency, thereby promoting industrial innovation, technology upgrades, and regional economic development. However, it may cause privacy leakage problems when we use and share data. This oversight has also led to more serious issues, such as the leakage of sensitive data and illegal cross-border data transfers. For instance, some financial companies, due to the absence of comprehensive privacy protection mechanisms in the processes of collecting, circulating, and utilizing user data, have experienced incidents where data is used and traded without user consent. As a result, it severely stops the data circulation and sharing. To further protect user data privacy, federated unlearning can rollback the data-generated training updates to the machine learning model, which can further protect the data privacy and security of users. In this paper, we review the research work of federated unlearning. Firstly, we conduct an in-depth analysis of the federated learning training architecture, highlighting the specific types of privacy leakage threats. To reduce the risk of privacy leaks, we introduce the concept and definition of unlearning, and list different unlearning scenarios, thereby seamlessly transitioning to the concept of federated unlearning. On this basis, we outline the processes involved in federated unlearning and introduce unlearning granularity and challenges. Secondly, the federated unlearning algorithms are divided into two categories, including global model-oriented and local model-oriented algorithms according to the modified object. We further subdivide into several subcategories based on two major categories and analyze the implementation details of each algorithm in depth. To further compare the strengths and weaknesses, we conduct detailed comparative analyses across different categories of algorithms, focusing on aspects such as algorithm performance, types of requesters, and forgetting requests. Additionally, we also conducted an experiment to show the performance of different categories of federated unlearning algorithms in terms of model accuracy. Thirdly, the commonly used performance metrics are divided into three categories, including model performance metrics, forgetting effect metrics, and privacy protection metrics. We conduct a detailed comparison and analysis of these metrics in terms of the unlearning stage, as well as their advantages and drawbacks. Fourthly, we summarize the research and applications of federated unlearning in privacy protection and attack resistance, including the protection of commercial information privacy, federated recommendation systems and federated clustering, etc. Finally, this paper looks forward to the future research directions of unlearning algorithms and applications from the personalized perspective, including promoting the market circulation of data elements, deletion of low-quality data, forgetting applications in cross-domain machine learning, customized services, and federated unlearning in special scenarios.

Keywords federated learning; federated unlearning; digital economy; privacy preserving; edge intelligence

1 引 言

数据的分析挖掘和开放共享对于推动各社会领域的发展具有重要意义. 长期以来,谷歌公司研发的数据搜索和分析工具Google Trends[1]被广泛用于通信、医学、健康和经济等各领域的分析和决策,不仅提高了分析效率,并且提供了有价值的见解. 长期从事智能决策理论与技术研究的中国工程院院士杨善林指出[2],大数据的价值体现在通过分析和挖掘数据所获得的深层知识,这些知识可为各种实际应用提供有效的决策支持.

然而,数据在流通和共享使用的过程中也存在

泄露的风险.Facebook数据泄露[3]、普渡大学案[4]等一系列数据隐私泄露案例使人们不断意识到自身数据安全的重要性,这也导致数据所有者不再愿意共享自己的数据,数据隐私泄露成为机器学习中最严重的问题之一[5].我国在《"十四五"大数据产业发展规划》指出[6],我国大数据产业发展迅速,但存在敏感数据泄露、违法跨境数据流动等问题,例如部分金融公司在用户数据采集、流转和使用中因未建立健全隐私保护机制,而导致数据在未经用户授权下被使用、交易等现象.

如何在实现数据流通共享的同时保护用户隐私已经成为当前数字经济建设的瓶颈.目前的相关研究尽力保证用户的数据隐私不受侵害,但仍然难以有效解决隐私泄露问题.例如,数据加密技术(包括同态加密[7]、多方安全计算[8]等)通过加密计算将明文数据转化为密文,进而实现安全的数据传输、存储和使用.但是数据加密算法的设计通常在安全性和计算速度上进行取舍[9],安全的加密算法往往需要复杂的计算,因此难以在大规模数据的机器学习任务[10]中广泛应用.近年来联邦学习[11]作为一种分布式的隐私保护技术,能在不共享原始数据的情况下实现数据知识的共享.

作为数据"可用不可见"的分布式人工智能学习新范式,联邦学习通过聚合用户本地训练的局部模型来完成机器学习训练.在此过程中,联邦学习不需要获取用户的实际数据信息,因此可以有效减少用户隐私泄露的可能性,为解决隐私泄露问题提供了有效的解决方案.然而,机器学习模型通过训练的方式保留了用户数据中隐含的信息或者知识[12],攻击者可以通过模型推断数据等方法"偷窥"用户隐私.例如,Ren等人[13]提出了一种回归神经网络模型,该模型能够通过用户向服务器传递的模型参数信息来恢复联邦学习训练过程的敏感数据.为了确保用户参与联邦学习训练过程中的个人数据安全,研究人员仍需要探索可靠且有效的方法来保护用户数据隐私.

联邦忘却学习[14](Federated Unlearning)提供了一种从联邦学习全局模型中删除特定数据贡献的方案,可以进一步保护用户数据隐私.联邦忘却学习在联邦学习框架的基础上,通过迭代训练[15]、直接删除[16]等方式,撤销用户本地局部模型对全局模型的训练更新.除此之外,国内外相关学者也针对联邦忘却学习在后门攻击抵抗[17]、数据撤销验证[18]等多个方向展开研究,目前联邦忘却学习已成为国际前沿研究领域.考虑到联邦忘却学习研究的发展迅速,以及国内仍缺少对这一领域进行系统梳理这一现状,本文综述联邦忘却学习的研究成果,以便于研究人员能够更系统、更快速地了解联邦忘却学习领域以及未来趋势.

本文是联邦忘却学习研究的第一篇中文综述性文章,主要从联邦忘却学习的实现方法、性能指标、应用领域以及未来展望等方面展开.本文具体章节安排如图1所示,第2节主要阐述机器学习训练模式从联邦学习到联邦忘却学习的发展,并给出联邦忘却学习的基本定义;第3节将联邦忘却学习算法根据修正对象分为面向全局模型和面向局部模型的联邦忘却学习算法,并对每种类型算法的发展方向、优缺点进行详细分析;第4节将评价指标分为模型表现指标、遗忘效果指标和保护隐私指标,并对不同类型评价指标的优缺点进行分析;第5节是对联邦忘却学习应用进行详细阐述;第6节对联邦忘却学习的未来研究进行展望并总结研究面临的挑战;最后,本文第7节进行全文总结.

2 联邦忘却学习概述

本节首先阐述联邦学习的一般训练流程,然后介绍忘却学习的相关研究,最后详述联邦忘却学习的流程、忘却粒度以及面临挑战.

2.1 联邦学习

传统的机器学习方法通常要求用户将原始数据上传至高性能云服务器进行集中式训练,但这种做法往往会导致数据流动不受控制和敏感信息泄露等问题.为了应对这些问题,谷歌于2017年提出联邦学习[11],以在保护用户隐私数据的前提下实现模型训练,进而促进数据的流通和共享.

联邦学习架构如图2所示,中心服务器协同由N个持有训练数据的用户组成的集合$U=\{u_1,u_2,\cdots,u_N\}$共同训练机器学习模型$f(w)$,得到模型最优参数w^*,其中每个用户u_i持有训练数据$D_i=\left\{\left(x_k,y_k\right)\right\}_{k=1}^{n_i}$,其中$n_i$为$D_i$中样本的数量.

联邦学习的具体训练流程如下:

(1)服务器初始化任务.中心服务器初始化模型参数w_0,选择$m(m\leqslant N)$个用户参与联邦学习任务,同时向各用户发送全局模型参数w_0.

(2)用户训练局部模型.在第t轮训练中,用户u_i获取全局模型参数w_t,并在本地数据集D_i上训练

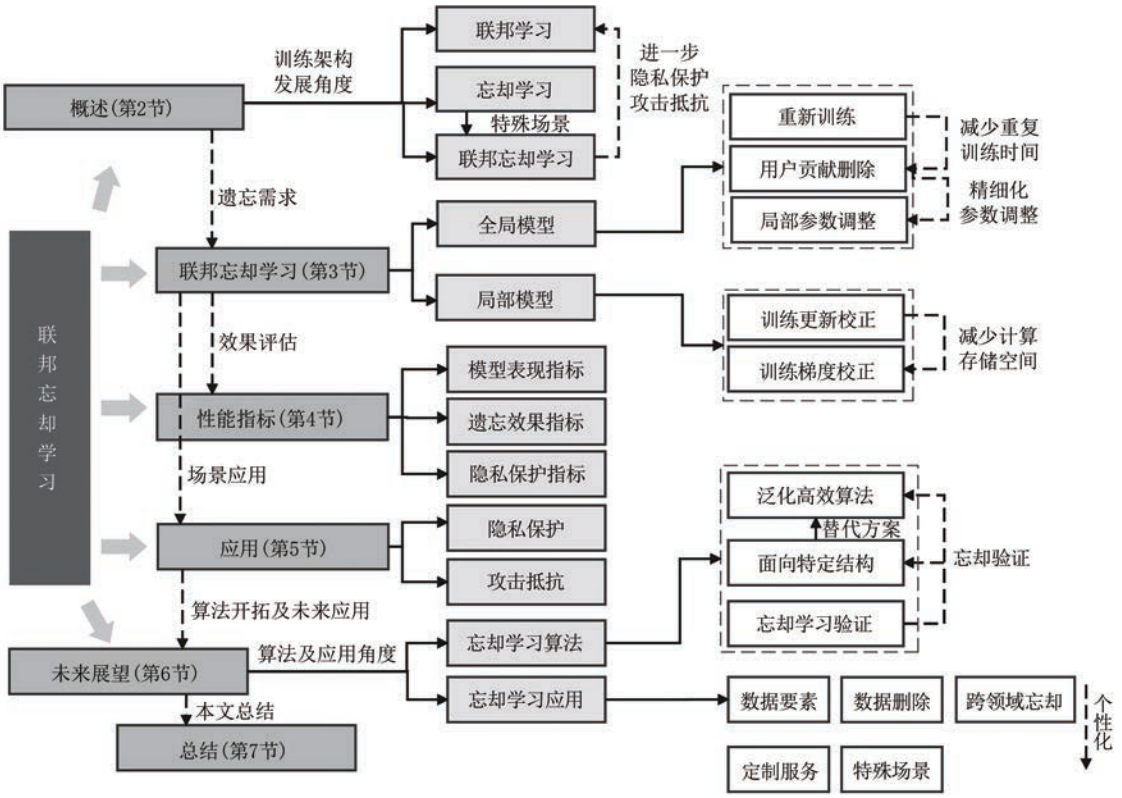

图1　文章结构安排

局部模型$f(w_t)$，以最小化目标函数：

$$\min_{w} F_i(w)=\frac{1}{n_i}\sum_{k\in D_i} l_k(w) \tag{1}$$

其中$l_k(w)$是局部模型对训练样本(x_k,y_k)的损失函数，表示模型预测值与实际值之间的误差．为优化公式(1)，SGD、Adam等优化器[19]被广泛使用．例如，在采用SGD优化器更新模型参数时，迭代公式为

$$w_{t+1}^i=w_t-\eta\nabla F_i(w_t^i,x_i) \tag{2}$$

其中w_{t+1}^i为用户u_i训练后的局部模型参数，η为学习率，x_i为随机抽取的数据．最后，用户计算局部模型的参数更新$\Delta w_t^i=w_t-w_{t+1}^i$，并发送至中心服务器．

(3)服务器聚合局部模型参数．中心服务器使用聚合规则(例如FedAvg等[20-22])对参与训练用户局部模型的参数更新进行聚合：

$$w_{t+1}=w_t+\frac{1}{m}\sum_{i=1}^{m}\Delta w_t^i \tag{3}$$

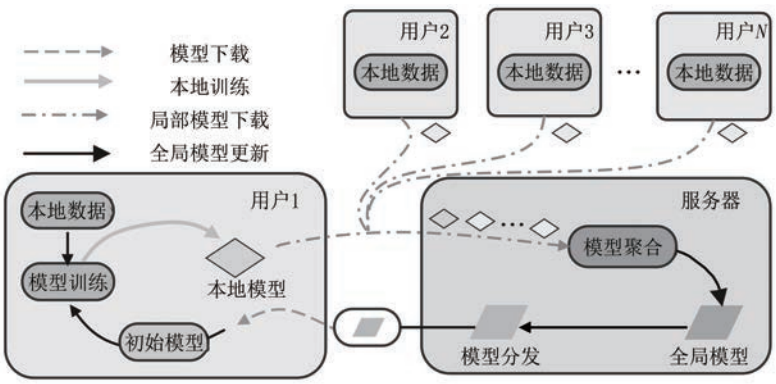

图2　联邦学习架构

然而，联邦学习所有用户共享全局模型的训练方式带来了新的数据隐私泄露安全隐患[23]．一方面，共享的全局模型可能泄露用户隐私数据．联邦学习的机器学习模型对所有参与训练的用户开放，

攻击者可以在本地训练过程中访问全局模型，根据模型反推参与者的数据信息，进而导致数据隐私泄露. Gong等人[24]指出，训练后的机器学习模型能够记忆训练数据的秘密信息，并且不法分子可以通过成员推断攻击[25]、模型翻转攻击[26]等方法窃取用户的隐私数据. 另一方面，机器学习模型也容易遭受恶意模型的攻击. 在聚合过程中，如果不能识别恶意用户上传的虚假模型，联邦学习训练的机器学习模型可能受到数据投毒[27]、模型投毒[28]等攻击，进而影响模型预测的准确率. 此外，联邦学习中训练数据可能受到数据中毒攻击而被污染或操纵，或存在训练数据过时，甚至在训练后被确定为错误等问题，这些数据被训练会严重降低联邦学习模型的准确率[16].

因此，如何有效地撤销训练数据对联邦学习全局模型的贡献，对于保护用户隐私数据、优化全局模型意义重大. 近年来，忘却学习[29](Unlearning)的概念被提出，它通过撤销训练数据对机器学习模型更新，为保护用户数据隐私和全局模型的安全提供了新的解决方案.

2.2 忘却学习

机器学习致力于从数据中提取有价值知识，而忘却学习则赋予模型“忘记”特定数据的能力. 忘却学习[29]的核心在于通过调整模型参数，实现与特定数据未参与训练相同的效果，同时避免完全的重复性训练. 基于忘却学习方法，服务提供者可以依据用户请求，使机器学习模型遗忘由特定用户数据训练产生的数据贡献，确保无法从模型中追踪到该用户数据的踪迹，达到保护用户隐私的目的. 同时，服务提供者可以通过让机器学习遗忘因错误数据或低质量数据训练产生的模型更新，进一步增强模型的安全性.

忘却学习的难度取决于不同的机器学习场景. 在所有数据都可获取的情况下，忘却学习可以利用数据、模型等有效信息进行忘却. 机器忘却学习[30](Machine Unlearning)不受隐私保护或数据加密的限制，将所有训练数据集中处理，对特定数据进行忘却学习. 如果待遗忘的数据样本量极小，例如仅有3到5张图片，这种场景称为小样本忘却学习[31](Few-Shot Unlearning). 除了忘却数据量的需求外，针对不同模型、快速忘却等需求的方法也被学者广泛研究，例如回归模型忘却[32]、推荐模型忘却[33]、快速忘却学习[34]等.

在数据受到隐私或使用限制的情况下，实施忘却学习的有效信息减少，忘却难度大大增加. 零次观察忘却[35](Zero-Glance Unlearning)不允许使用请求遗忘的数据，只能利用剩余训练数据，这在人脸识别系统等场景中是常见的. 如果所有参与训练的数据都无法被访问，这种情况被称为零样本忘却[36](Zero-Shot Unlearning)，它的条件更为严格，只能在数据不可见的情况下进行忘却学习.

在不同场景下，忘却学习算法设计的难度与获取数据有效信息的难度相关. 在数据均可获取的情况下，我们能够直接针对需要遗忘的数据进行集中处理，并深入挖掘更有效的信息，例如记忆矩阵学习[37]可以实现任何类别的忘却行为. 在此场景下，主要的挑战在于如何提高忘却速度与遗忘效果，并解决灾难性遗忘等负面影响. 然而，当数据使用权限受到限制时，忘却算法只能借助数据在满足隐私和控制条件下获得的信息. 相比于所有数据都可获取的情况，限制数据的信息提取和利用都需要针对模型和实际场景进行精细化的设计，尤其在零样本忘却的场景下，这一挑战更为显著[38].

联邦忘却学习[14]是零样本忘却的一个特例，它考虑数据完全不可见的实际场景，研究在联邦学习场景下的忘却学习. 联邦学习场景不仅无法直接获取数据信息，而且数据还分布在不同用户的手中. 对这种分散的数据集中实施忘却，尤其是当需要被遗忘的数据分布在不同的用户中时，有效信息更难以进行整合和利用. 但是，联邦忘却学习符合现实世界的数据管理模式. 现实场景下用户对自身数据拥有绝对的控制权，因此联邦忘却学习的研究具有更高的应用价值.

2.3 联邦忘却学习

本节阐述联邦忘却学习的流程、目标、所能实现的忘却粒度以及面临的挑战.

2.3.1 联邦忘却学习流程

联邦忘却学习通过撤销用户数据对全局模型产生的训练更新，来解决联邦学习中存在的隐私泄露问题. 如图3所示，联邦忘却学习的流程如下：

(1)联邦学习. 中心服务器协同用户集合$U=\{u_1,u_2,\cdots,u_N\}$进行联邦学习，训练得到最优模型参数w^*.

(2)数据更新撤销. 用户u_i可以在任意时刻向中心服务器发送撤销数据D_i^f请求. 中心服务器根据用户的请求完成以下目标[16]：

定义1. 联邦忘却学习. 设w^r表示联邦学习

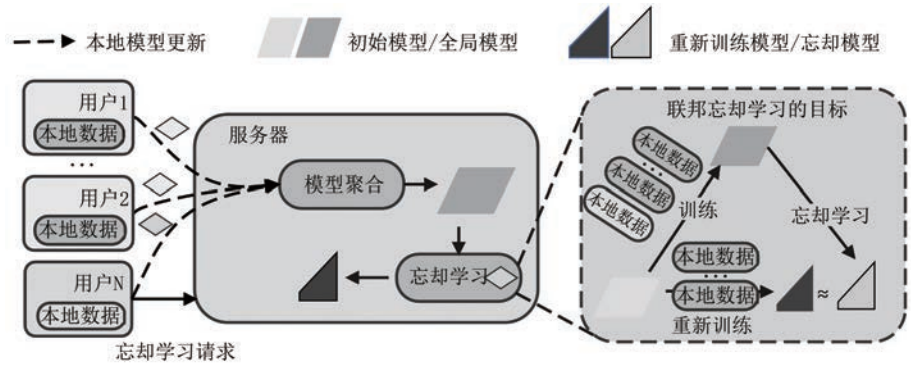

图3　联邦忘却学习架构

模型在数据集 $U_{i=1}^{N}(D_i-D_i^r)$ 上重新训练得到的模型参数，w^u 表示联邦忘却学习得到的全局模型参数. 给定阈值 $\varepsilon\geqslant 0$，忘却学习算法 $\mathcal{A}$，存在满足以下不等式成立：

$$\left|T(w^r)-T_{\mathcal{A}}(w^u)\right|\leqslant\varepsilon \tag{4}$$

其中 $T(w^r)$ 表示模型参数 w 的数据分布，ε 表示很小的正数. 当 $T(w^r)$ 和 $T_{\mathcal{A}}(w^u)$ 两分布 ε 相同时，算法 $\mathcal{A}$ 实现忘却学习.

(3)用户 u_i 将数据 D_i^r 从联邦学习系统中删除，防止数据 D_i^r 被重新训练.

(4)服务器继续发起联邦学习任务，恢复全局模型的预测准确率.

2.3.2　联邦忘却学习忘却粒度

用户在不同场景下提出的联邦忘却学习请求有所不同，主要体现在联邦忘却学习对全局模型的忘却粒度.

(1)样本忘却. 样本忘却[29]是从联邦学习模型中撤销特定数据样本对模型的训练更新，是细粒度的忘却学习请求，也是联邦忘却学习中常见的请求之一[36].

(2)类别忘却. 类别忘却[17]用于分类任务中撤销单个或多个类别的数据对全局模型的训练更新. 例如，人脸识别模型[39]中记录数量庞大的人脸数据，每一个用户的人脸就是一个类别信息. 当用户请求撤销自身的人脸数据时，联邦忘却学习需要删除模型对该类人脸信息的记忆.

(3)任务忘却. 机器学习模型不仅可以针对单个任务进行训练，还可以对多任务训练. 在多任务训练模式下，模型学习多个任务时可以因任务之间的相关性获得性能上的收益[40]. 任务忘却是粗粒度的忘却学习请求，用于撤销某个任务所有数据对模型的训练更新，但会因大量的数据撤销而产生灾难性的忘却[41-42].

2.3.3　联邦忘却学习面临挑战

联邦忘却学习在不掌握用户数据的情况下，撤销特定数据对全局模型的贡献，其研究中主要存在以下挑战：

(1)分布式的训练方式[43]导致联邦忘却学习难以完全撤销目标用户(请求忘却学习的用户)对模型的数据贡献. 联邦学习是在保护用户隐私的前提下进行分布式训练，每个参与的用户都会保留全局模型的参数并在本地进行训练. 因此，仅仅在全局模型中撤销某个用户的训练更新是远远不够的，因为其他用户的掌握模型参数仍然保留着待撤销的用户更新，在之后的训练中仍会被聚合到全局模型中.

(2)模型训练的增长过程使联邦忘却学习难以利用已训练的信息. 联邦学习模型的训练是一个增长的过程[44]，先前所有用户数据的训练更新都决定了当前用户在本地的模型参数更新，并且其他用户对模型的训练也会因增长的训练过程包含目标用户数据的踪迹. 因此，仅撤销目标用户的本地更新会极大地降低全局模型的准确率，而且撤销更新后的模型仍然会保留对目标数据的记忆.

(3)大量的数据撤销导致灾难性的忘却[41]. 一般而言，联邦忘却学习后的模型在准确率的表现要比重新训练的模型差. 当忘却学习的数据量增加时，全局模型的准确率会急剧下降. 尽管一些研究[45-46]通过设计特殊的损失函数来缓解灾难性的忘却，但灾难性的忘却问题仍然需要被解决.

3　联邦忘却学习算法

本节对当前联邦忘却学习的算法开展综述. 首

先,我们根据修正对象(在忘却过程中被直接修改的特定元素)将各类联邦忘却学习算法划分为面向全局模型和面向局部模型两种.面向全局模型的忘却算法直接对全局模型参数进行修改并利用用户数据调整全局模型,而面向局部模型的算法则利用用户训练的局部模型参数对全局模型参数进行间接修改.其次,本节对每一类算法在适用模型、发起者、忘却学习请求类型、优缺点等方面进行详细的对比和分析.最后,本节基于算法分析和实验验证进一步对比不同类型算法之间的差异.

3.1 面向全局模型的联邦忘却学习算法

面向全局模型的联邦忘却学习算法通过直接修改全局模型参数来实现对目标数据的忘却.最初的研究方法是将模型完全初始化,然后在剩余用户上进行重新训练.然而,这种方法在计算和通信方面具有较高的复杂性,因为重新进行联邦学习需要大量的计算资源.为了降低不必要的训练开销,当前的研究主要集中在如何消除目标用户对模型的影响,然后利用再训练方法恢复模型性能.

3.1.1 重新训练

如图4所示,重新训练通过模型回退和再训练方法能够完全撤销特定用户数据对模型的训练更新.首先,服务器通过联邦学习在服务器持续更新全局模型参数w_t以完成模型训练.其次,用户u_i根据自身隐私需求向服务器发出忘却学习请求.接着,服务器响应用户u_i的忘却请求,将全局模型参数w_t回退到某个训练时刻训练所对应的模型参数$w_j(j<t)$.最后,为防止因参数回退产生的灾难性遗忘,服务器依赖回退后的模型参数在其他用户的数据上进行再训练,从而实现联邦忘却学习.然而,当联邦忘却学习处理海量数据样本时,再训练过程中传输大规模模型参数时会产生巨大的通信压力,并且也会增加训练的时间成本.

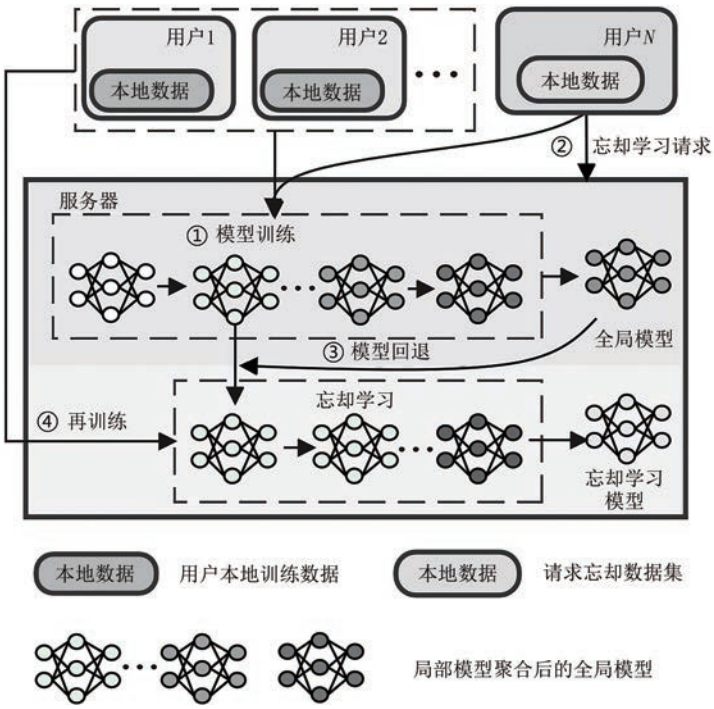

图4 重新训练基本架构

早期的研究采用初始化模型参数和再训练方法,实现联邦忘却学习.Liu等人[16]提出快速再训练算法Rapid Retraining.该算法采用近似估计Hessian矩阵的方法来快速构建联邦学习模型.定义联邦数据删除操作,允许用户将本地的部分数据移除联邦学习系统:

定义2. 联邦数据删除操作.一个删除$r=(\tau,o)$,其中$\tau\in D$为数据集D的一个数据样本,o为一个删除请求.对于用户u_i,给定局部数据集D_i和一个删除序列$R=\{r_1,r_2,\cdots,r_N\}$,删除操作被定义为

$$D_i^r=D_i\circ R=D_i\setminus\left\{\Omega\right\}_{j=1}^{|r_i|} \tag{5}$$

其中$|r_i|$为用户u_i删除请求r_i所删除的数据样本数量.

用户u_i发送忘却学习请求后,服务器将机器模型参

数完全初始化为w_0^u，并协同所有用户在剩余数据集上进行快速地重新训练. 在t时刻，用户u_i利用小批量梯度下降[47]和Hessian矩阵优化参数w_0^u，如公式(6)所示.

$$w_{t+1}^{u,i}=w_t^{u,i}+\frac{1}{B-\Delta B_t}H_i^{-1}\mu_i \tag{6}$$

其中$\mu_i=\nabla F_k(w_t^{u,i})$为梯度，$B$为最小批量的大小，$\Delta B_t$表示在$t$轮训练中删除的子集大小.

然而，Hessian矩阵的计算成本仍然昂贵. 为解决这一问题，利用有限记忆BFGS(L-BFGS)算法的思想估计Hessian矩阵：

$$H_t^i\approx\frac{1}{B-\Delta B}\sum_{(x_k,y_k)\in D_i'}\nabla F_k(w_t^{u,i})\nabla F_k(w_t^{u,i})^T \tag{7}$$

其中，H_t^i是通过参数$w_t^{u,i}$对Hessian矩阵的近似估计.

虽然Rapid Retraining能够消除用户数据对模型的更新，但仍然需要大量的时间和频繁的通信用于重新训练. 鉴于此，研究人员对重新训练的方法进行改进，以减少重新训练产生的时间和通信消耗. 为了减少重新训练时参与的用户数量，Su等人[48]提出一种异步联邦学习的聚合机制KNOT. KNOT建立并解决线性规划问题，以优化用户与集群的分配，该方法将用户划分为集群，仅在集群内执行联邦学习的模型聚合. 集群划分的训练方式使一个用户训练后产生的数据贡献仅能够影响同一集群的用户，其他集群的用户不受影响. 当用户请求联邦忘却学习时，数据忘却而引起的重新训练可以被限制在每个集群内. 因此，仅在集群中执行忘却学习可以极大地减少再训练过程中的通信消耗和时间. 为了减少不必要的重复训练次数，Fraboni等人[49]提出了Informed Federated Unlearning(IFU). IFU算法建立在模型的有界敏感度概念之上，有界敏感度由用户u_k是否参加联邦学习而产生的模型差计算，如公式(8)所示.

$$\varphi(t,i)=\sum_{j=0}^{t-1}\left\|\Delta w_j-\Delta w_j^r\right\|_2 \tag{8}$$

其中Δw_j^r是在删除用户u_i重新训练后j时刻的模型更新.

根据给定的阈值φ^*，IFU寻找最大的时间t满足$\varphi(t,i)\leqslant\varphi^*$，并将机器学习模型回退到该时间训练的模型. 在此基础上，IFU在剩余数据集上进行重新训练，从而实现联邦忘却学习.

3.1.2　用户贡献删除

针对重复性的再训练工作，用户贡献删除算法可以减少用户重复训练的时间和通信开销. 如图5所示，服务器仍然通过用户的数据训练全局模型，聚合用户数据训练产生的局部模型参数$\{w_t^1,w_t^2,\cdots,w_t^N\}$，并修改全局模型参数$w_t$. 在此基

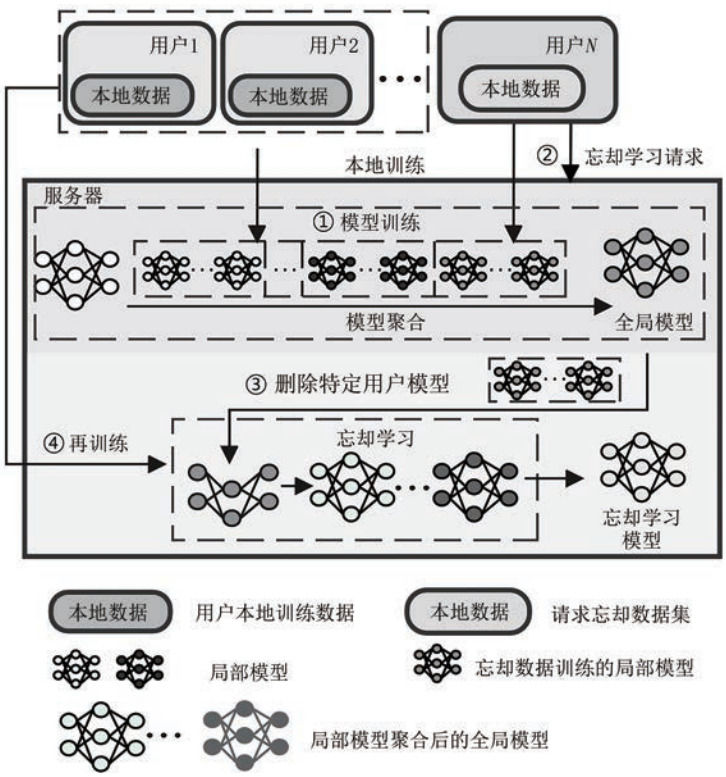

图5　用户贡献删除基本架构

础上，特定用户u_i向服务器发起忘却学习请求，服务器从全局模型w_t中直接删除用户u_i产生的局部模型参数$\{w_0^i, w_1^i, \cdots, w_t^i\}$，实现对用户$u_i$数据的忘却．最后，服务器利用再训练提升模型准确率．用户贡献删除算法不再进行模型参数的回退，而是撤销局部模型对全局模型的训练更新，在此基础上进行再训练．因此，用户贡献删除算法是在已经具有训练基础的模型上进行再训练，可以减少重新训练算法中重复的训练工作．

在早期的研究中，用户贡献删除方法通过直接删除用户训练的局部模型和再训练来实现忘却学习．Wu等人[50]提出了Federated Unlearning with Knowledge Distillation(FUKD)，该算法需要在中心服务器内存储所有用户模型参数的更新，通过直接删除的方式实现联邦忘却学习．FUKD是完全在服务器上执行，所以运行时间和成本消耗优于需要用户和服务器之间额外通信的方法，并且该方法不受神经网络结构的限制．具体而言，根据公式(9)考虑忘却前后模型参数之间的关系：

$$w^u = w^* - \frac{1}{N}\sum_{t=1}^{T}\Delta w_t^i + \sum_{t=1}^{T}\epsilon_t \tag{9}$$

其中w^*和w^u分别为忘却前后的模型参数，$\Delta w_t^i = w_t^i - w_{t-1}$为用户$u_i$在第$t$轮的参数更新，$\epsilon_t$表示删除参数更新$\Delta w_t^i$对模型造成偏差的修正，$T$表示训练时间．

因此，服务器利用目标用户u_i历史的平均更新$\frac{1}{N}\sum_{t=1}^{T}\Delta w_t^i$和修正$\sum_{t=1}^{T}\epsilon_t$可以删除用户$u_i$数据对模型的贡献．然而，修正$\sum_{t=1}^{T}\epsilon_t$难以计算．为此，FUKD在删除用户的更新后，通过知识蒸馏方法[51]恢复产生的修正，其中知识蒸馏将原模型看作教师模型，将忘却后模型看作学生模型．

模型训练是一个逐步增长的过程，聚合后的用户模型将影响后续所有参与训练的用户更新方向．因此，直接删除某一用户的局部模型更新可能导致整体模型产生偏差，而短时间内通过训练来弥补这一性能损失非常困难．针对这一问题，Zhu等人[52]提出FedLU．FedLU基于认知神经科学理论通过回溯干扰和被动衰减删除特定的知识．回溯干扰是指记忆中的信息是有限的，当新信息不断涌入时，就会与旧信息相互干扰，使得旧信息逐渐难以检索．回溯干扰通过定义硬损失反向优化请求忘却的样本．被动衰减认为遗忘是由于时间的推移和记忆轨迹的消失而导致的．被动衰减步骤通过知识蒸馏方法恢复机器学习模型的泛化性能，衰减记忆中的旧知识．当用户u_i请求忘却学习时，服务器要求用户u_i通过回溯干扰和被动衰减来撤销数据对全局模型的贡献，并利用知识蒸馏的方法代替再训练过程以恢复全局模型性能，进而降低忘却学习对全局模型准确率的影响．

3.1.3 局部参数调整

参数局部调整方法通过局部性地修改全局模型参数，解决特定模型结构的忘却学习问题．该类算法通过结构信息计算用户数据贡献的参数位置，准确地删除用户数据对全局模型的贡献，从而实现数据的有效遗忘，详细过程如图6所示．首先，服务器联合所有用户进行联邦学习训练，得到全局模型参数w_t．其次，用户u_i向服务器发起忘却学习请求．再次，服务器接受忘却请求后，会根据模型结构对参数w_t进行评估，识别出包含用户u_i参数贡献的部分．然后，服务器将全局模型中用户u_i的参数贡献部分消除，形成新的参数w_t'．最后，服务器联合用户进行再训练，调整参数w_t'．一般而言，此类算法在对机器学习贡献的删除能力方面优于面向全局模型和局部模型的联邦忘却学习算法．目前该类算法的研究既涉及决策树等传统机器学习模型，也涉及深度神经网络模型．

决策树[53]、支持向量机[54]等传统机器学习模型在工业场景得到了广泛的应用，但是也存在着因模型记忆而导致的利益冲突．针对这一问题，研究者将联邦忘却学习应用于传统机器学习模型．例如，Li等人[55]提出了一个针对随机森林的联邦忘却学习框架RevFRF，在撤销用户数据贡献时同时满足以下两个要求：(1)其余用户无法继续使用包含目标用户数据的模型．(2)目标用户无法再使用过去训练的模型．具体而言，RevFRF利用决策树的结构通过递归遍历的方式将目标用户提供的所有节点销毁，然后在剩余的用户中重新构建决策树，以恢复删除后的模型准确率损失．该算法仅适用于决策树这一类模型，存在一定的局限性．

深度神经网络[56]中各层上的参数通过加权、乘积等方式构建了输入与输出之间的映射．目前的研究针对卷积层等特定结构，探索参数对模型输入输出的贡献，实现了高效的联邦忘却学习．

例如，利用卷积层的结构特点进行联邦忘却学习，Wang等人[17]提出了针对图像分类任务的联邦忘却学习算法Federated Unlearning via Class-discriminative Pruning(FUCP)，用于完成类别忘却

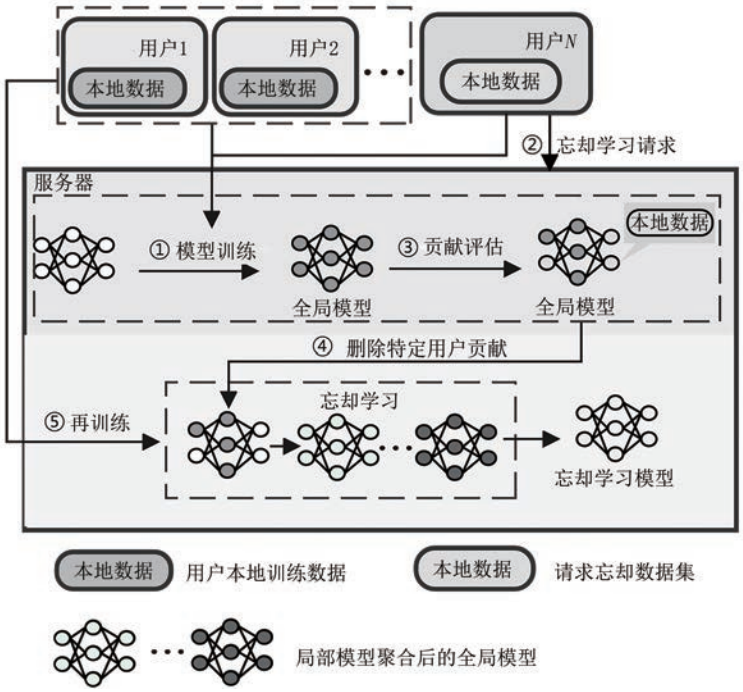

图6　局部参数调整基本架构

任务.针对图像分类任务,FUCP依据卷积神经网络结构中通道对类别的评分剪枝通道参数,使联邦学习模型有选择地"遗忘"特定类别的贡献.然而,FUCP只能删除某个类别数据对全局模型的贡献,而无法剔除类别中的特定样本子集,并且该算法只适合卷积神经网络分类模型.

FUCP通过术语频率-逆文档频率所给出的得分来判断卷积神经网络通道对某一类别的贡献,从而进行剪枝操作."术语频率-逆文档频率"[57](Term Frequency Inverse Document Frequency, TF-IDF)的概念被应用于量化信道对类别的贡献.考虑第l层卷积$A^{*}\in R^{|O|\times C_{out}^{l}}$,其中$|O|$表示分类类别的数目,$C_{out}^{l}$表示输出通道数,则术语频率TF的计算公式为

$$TF_{l}^{u'}=\frac{A_{l}^{*o'}}{\sum_{j=0}^{C_{out}^{l}}A_{l}^{*o',j}} \tag{10}$$

其中$TF_{l}^{u'}\in R^{C_{out}^{l}}$表示每个通道对特定类别$o'$的贡献.

但是,一些在TF中得分较高的频道也可能对目标类别以外的其他类别做出贡献.为了获得目标类别的最具辨别力的通道,可通过公式(11)计算第l层的IDF:

$$IDF_{l}^{j}=\log\frac{1+|O|}{1+\left|o_i\in U:A_{l}^{*o_i,j}\geqslant Avg\left(A_{l}^{*o_i}\right)\right|} \tag{11}$$

其中$IDF_{l}^{j}\in R^{1}$表示通道j在整个类别中罕见的程度.

IDF_{l}^{j}越接近0说明该通道在其他类中也有很高的贡献.将$TF_{l}^{o'}$与IDF相乘计算得到通道对类别的评分$TF-IDF_{l}^{o'}$:

$$TF-IDF_{l}^{o'}=TF_{l}^{o'}*\left\{IDF_{l}^{j},\cdots,IDF_{l}^{C_{out}^{l}}\right\} \tag{12}$$

其中$TF-IDF_{l}^{o'}\in R^{C_{out}^{l}}$表示通道和类别之间的相关分数.

高TF-IDF得分的通道对目标类别有更好的识别能力,因此FUCP将高得分通道的参数初始化为0并确保其不再进行反向传播[58].剪枝过程不需要任何迭代训练或搜索,因此具有较高的计算效率.在完成通道修剪后,FUCP通过再训练对模型进行修正,恢复模型的性能.

参数局部调整方法依赖于模型结构,采用可解释的数学方法判断特定用户数据贡献高的参数位置,并进行剪枝.此类方法能够借助模型结构和训练的数据信息,因此能够实现更好的忘却效果.然而,参数局部调整方法的研究较少,缺少针对不同类型模型的有效忘却方案.

表1对比具有代表性的面向全局模型算法在算法性能、适用模型、发起者和忘却学习请求类型的差异.其中算法性能包括遗忘效果、忘却速度、占用空间和稳定性.遗忘效果是指机器学习模型执行忘却学习算法后对数据的记忆程度;忘却速度是指忘却

表1 面向全局模型的联邦忘却学习算法对比

类型	算法名称	算法性能				适用模型	发起者		忘却请求		
		遗忘效果	忘却速度	占用空间	稳定性		用户	服务器	样本	类别	任务
重新训练	Rapid retraining[16]	●	○	○	○	机器学习模型	√	×	√	√	√
	KNOT[48]	●	◐	◐	○	机器学习模型	√	×	√	√	√
	IFU[49]	●	◐	●	◐	机器学习模型	√	×	√	√	√
用户贡献删除	FUKD[50]	●	●	●	◐	机器学习模型	×	√	√	×	√
	FedLU[52]	◐	◐	◐	◐	机器学习模型	×	√	√	×	√
局部参数调整	RevFRF [55]	●	○	●	○	决策树	×	√	√	×	√
	FUKD[17]	●	◐	◐	◐	卷积神经网络	×	√	×	√	×

学习算法的快慢；占用空间指忘却算法过程中服务器占用的空间大小；稳定性是指随联邦学习用户数量对遗忘效果、忘却速度和占用空间的敏感性. 此外，实心圆圈表示“程度很重”，半实心圆圈表示“程度一般”，空心圆表示“程度很轻”.

通过对比发现，重新训练算法遗忘效果好，但忘却速度较慢且算法稳定性差，其发起者均为用户，主要用于保护用户隐私，并且支持更多类型的忘却学习请求. 用户贡献删除算法占用空间大、忘却速度较快. 用户贡献删除算法的发起者均为服务器，支持较少类型的忘却学习请求，因为其主要用于解决异构数据、中毒攻击等问题. 局部参数调整算法需要根据模型结构计算用户数据对模型参数贡献，算法遗忘效果好，忘却时间较长，占用空间较大. 此外，局部参数调整算法的设计根据需求制定，任务发起者和忘却请求类型根据不同的需求而改变.

3.2 面向局部模型的联邦忘却学习算法

面向局部模型的联邦忘却学习算法通过增加联邦学习训练，在局部模型中获取新的知识，根据新知识整体性地修改全局模型，不需要进行再训练. 在最初面向局部模型的联邦忘却学习研究中，所有用户上传的局部模型都会被修改，然后聚合到全局模型中. 为了减少服务器中复杂的计算，研究人员利用用户训练的局部模型实现忘却学习，仅调整一部分用户局部模型，并将其聚合到全局模型.

3.2.1 训练更新校正

训练更新校正算法在现有模型的基础上增加额外的联邦学习训练，对训练过程中产生的模型参数进行修正，并通过聚合修正后的模型修改全局模型的参数，具体过程如图7所示. 首先，服务器协同所有用户进行联邦学习，完成全局模型的训练. 然后，用户 u_i 向服务器发起忘却学习请求. 接着，服务器协同用户继续进行一定轮次的联邦学习训练，接受用户上传的模型参数 $\{w_{t+1}^i, w_{t+2}^i, \cdots, w_{t_{end}}^i\}$. 最后，服务器对训练的局部模型参数 $\{w_{t+1}^i, w_{t+2}^i, \cdots, w_{t_{end}}^i\}$ 进行修正，通过聚合修正后的模型参数 $\{\tilde{w}_{t+1}^i, \tilde{w}_{t+2}^i, \cdots, \tilde{w}_{t_{end}}^i\}$ 更新全局模型，从而实现联邦忘却学习.

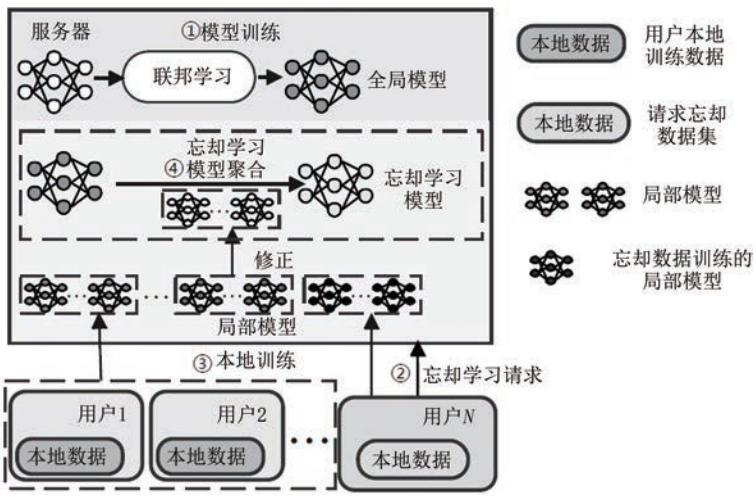

图7 训练更新校正算法基本架构

早期的研究致力于对所有用户的模型参数进行校正，并将校正后的参数进行聚合，以达成联邦忘却学习的目标. 例如，Liu等人[14,59]提出了FedEraser算法，基本思想是对用户的模型参数更新进行校准. 在用户u_i发起联邦忘却请求时，中心服务器协同用户集合$U_{train}=U\backslash\{u_i\}$使用初始的全局模型参数$w_0$进行$n$轮本地训练. 在第$t(1\leqslant t\leqslant T)$轮训练时，每一个用户$u_j\in U_{train}$计算模型更新$\Delta w_t^j$并发送至中心服务器. 中心服务器利用已训练的更新$\Delta w_t^i$对其他用户的训练参数$\Delta w_t^j$进行校正：

$$\Delta\tilde{w}_t^j=\left|\Delta w_t^i\right|\frac{\Delta w_t^j}{\left|\Delta w_t^j\right|} \tag{13}$$

其中Δw_t^i的范数$\left|\Delta w_t^i\right|$表示参数需要更改的程度，$\Delta w_t^j$的标准化$\frac{\Delta w_t^j}{\left|\Delta w_t^j\right|}$表示模型参数应该更新的方向.

在所有用户模型的更新校正后，中心服务器对所有校正更新$\Delta\tilde{w}_t^j$进行聚合，公式如(14)所示：

$$\Delta\tilde{w}_t=\frac{1}{(N-1)\sum_i\Omega_i}\sum_j\Omega_j\Delta\tilde{w}_k^j \tag{14}$$

其中Ω_i为联邦学习框架中用户u_i的权重，N为联邦学习系统中的用户数量.

聚合的校正$\Delta\tilde{w}_t$对全局模型参数进行更新：

$$w_{t+1}=w_t+\Delta\tilde{w}_t \tag{15}$$

然而，FedEraser难以完全撤销用户数据对模型的贡献. FedEraser利用最近局部模型更新来近似删除目标用户对模型的贡献. 如果要撤销的数据样本更新在一开始就用于模型训练，那么联邦忘却学习得到的模型可能仍然包含这些数据样本的贡献.

为了删除用户数据对模型的贡献，Gao等人[18]提出联邦忘却学习与验证的统一框架VERIFI，并提供名为Scale-to-Unlearn (S2U)的联邦忘却学习算法. 具体而言，在t_m时刻，用户u_i发起联邦忘却学习请求撤销自身数据D_i^r对模型参数w_{t_m}的更新. 中心服务器接受请求并发起联邦学习训练，在模型聚合时缩小目标用户模型的贡献比例，放大其他用户模型的贡献比例：

$$w_{t+1}\leftarrow Agg\left(\Psi\left(w_{t+1}^j-w_t\right)_{j=1}^n\right) \tag{16}$$

其中$\Psi(w_{t+1}^j-w_t)$为修正参数的更新项，$Agg(\cdot)$表示参数聚合，计算公式为

$$\Psi\left(w_{t+1}^j-w_t\right)=\begin{cases}\alpha(w_{t+1}^j-w_t), if\, j\, is\, u_i\\ \beta(w_{t_m}-w_t), if\, j\in U\backslash u_i\end{cases} \tag{17}$$

其中$\alpha\in(0,1)$为缩小模型贡献的比例，$\beta\in[1,+\infty]$为放大模型贡献的比例. 此外，VERIFI判断联邦忘却学习任务完成的条件是通过设计的判断函数$\Phi(\cdot)$，当$\Phi(D_i^r,w_{t_m})-\Phi(D_i^r,w_{t+1})\geqslant\delta$时，即为服务器实现联邦忘却学习任务. 在上面公式中，S2U在分类任务中将$\Phi(\cdot)$设计为$\Phi(D_i^r,w)=D_{KL}(f(D_i^r,w),P)$，$f(\cdot)$为机器学习模型，$P=(\frac{1}{p},\frac{1}{p},\cdots,\frac{1}{p})$表示等概率分布，$D_{KL}(\cdot\|\cdot)$表示相对熵[60].

为了减少服务器存储模型参数更新历史记录的负载，近期的研究仅修改请求忘却用户的参数. 例如，Li等人[15]基于子空间梯度上升实现了一种新型的联邦忘却学习算法. 如图8所示，服务器收集目标用户u_i在执行梯度上升之后生成的梯度μ，以及其他用户的表示矩阵R. 服务器通过奇异值分解[61]获取矩阵R的梯度子空间S，并根据公式(18)修改全局模型参数.

$$w_{t+1}=w_t-\mu_2 \tag{18}$$

其中μ_2为梯度μ在子空间S上的投影. 该方法只修改目标用户u_i的梯度，有效地减少服务器的计算量.

3.2.2　训练梯度校正

训练梯度校正算法的思想是在现有模型的基础上增加联邦学习训练，通过调整部分用户的训练方法，以直接聚合的方式来更新全局模型参数，详细过程如图9所示. 首先，服务器协同所有用户进行联邦学习，完成全局模型参数w_t的更新. 然后，用户u_i请

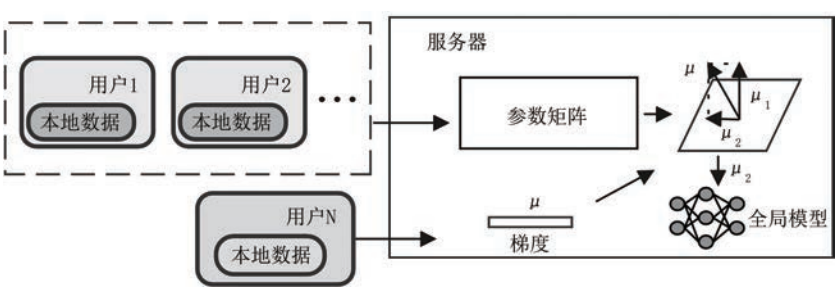

图8　子空间梯度上升

求服务器撤销自身数据对模型的更新. 接着,服务器要求所有用户继续进行联邦学习训练,在此过程中调整用户的训练策略(例如损失函数、训练梯度等). 最后,服务器接收并聚合用户的局部模型参数$\{w_{t+1}^1, w_{t+1}^2, \cdots, w_{t+1}^N\}$,利用聚合后的模型参数更新全局模型,以此实现联邦忘却学习.

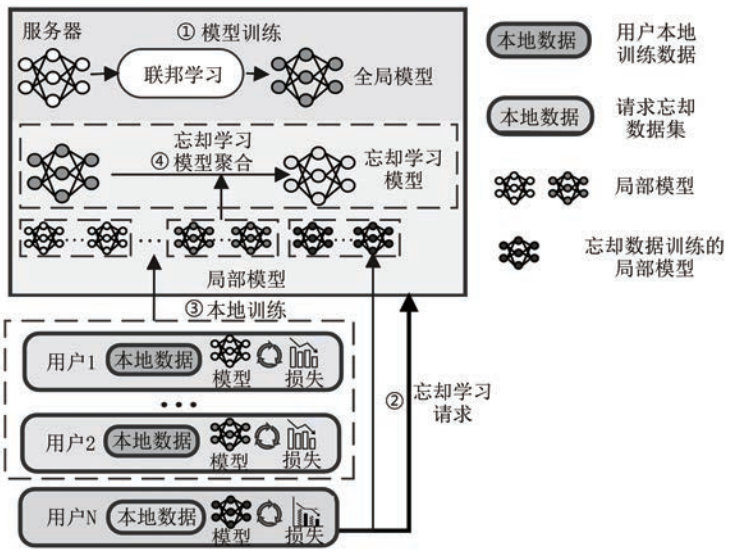

图9 训练梯度校正算法基本架构

考虑到梯度信息在机器学习模型训练中的关键作用,研究人员在本地训练过程中通过改变梯度信息来达成忘却学习的目标. 例如,Liu等人[62]提出Forsaken,该算法在参与训练的用户中配置可训练的虚拟梯度产生器,以根据目标函数的改变产生对应的虚拟梯度:首先,用户u_i请求联邦忘却学习删除本地部分数据D_i^r,下载中心服务器的全局模型参数w_t,初始化虚拟梯度μ_0.

其次,用户使用数据集D_i^r对机器学习模型进行T轮训练更新. 在t轮训练时,用户u_i在本地更新模型参数:

$$w_t^i = w_{t-1}^i - \eta \frac{1}{\left|D_i^r\right|}\mu_t \tag{19}$$

其中η为学习率,$\left|D_i^r\right|$表示训练数据集D_i^r的样本数量. 用户u_i通过模型对数据进行预测:

$$Y = \left\{ y_k \middle| y_k = f_t(x_k, w_t^i), x_k \in D_i^r \right\} \tag{20}$$

其中y_k为对数据x_k的预测标签. 根据预测标签,优化目标函数:

$$L_0 = argmin_{\mu_d}(\left\| \mu_d - w_t \right\|^2 + \left\| Y - P \right\|^2 + \delta) \tag{21}$$

产生虚拟梯度μ_t. 在公式(21)中,$\|Y-P\|^2$表示训练优化的目标,其中$P=(\frac{1}{p}, \frac{1}{p}, \cdots, \frac{1}{p})$表示等概率分布;$\|\mu_d - w_t\|^2$约束优化时的最大步长和优化方向;$\delta = \|w_{t+1} - w_t\|^2$为正则化项,用于避免模型精度损失. 最后,用户向中心服务器上传累计的虚拟梯度$\sum_{t=0}^{T}\mu_t$,中心服务器利用梯度信息修改全局模型的参数.

为了减少用户训练过程中的复杂性,早期的研究通过梯度上升策略更新模型参数,并在此过程中限制参数的修改幅度. Halimi等人[63]翻转机器学习模型训练过程,利用梯度上升方法实现联邦忘却学习,提出Unlearning with Projected Gradient Ascent (UPGA). 请求联邦忘却学习的用户u_i解决优化问题:

$$\max_{w_t^i \in \left\{ v \in R^d : \left\| v - w_i^{ref} \right\|_2 \leqslant \delta \right\}} F\left(w_t^i\right) \tag{22}$$

其中$w_i^{ref} = \frac{1}{N-1}\sum_{i \neq j}^{N} w_{t-1}^j$表示$(t-1)$轮训练其他用户对参数的贡献;$\|v - w_i^{ref}\|_2 \leqslant \delta$表示二范数[64]约束机器学习模型忘却学习的参数,避免模型参数因梯度上升而急剧改变;δ为超参数.

在解决问题(22)时,UPGA采用梯度上升训练的方法. 具体而言,定义w_i^{ref}为圆心,半径为δ的二范数球$\Omega = \{v \in R^d : \|v - w_i^{ref}\|_2 \leqslant \delta\}$,以及投影操作$\varphi: R^d \to R^d$. 给定学习率$\eta$,用户$u_i$迭代地更新局部模型参数:

$$w_{t+1}^i = \varphi\left(w_t^i + \eta \nabla F_i\left(w_t^i\right)\right) \tag{23}$$

迭代终止条件为模型在目标数据集上的准确率小于给定的阈值δ. 由于梯度上升会导致模型的准

确率下降，因此忘却学习后仍然需要采用联邦学习继续训练的方式恢复全局模型的准确率.

此外，Wu等人[65]提出EWC-SGA忘却学习框架，同样使用梯度上升的方法实现联邦忘却学习. 具体而言，忘却学习的过程是优化新的损失函数：

$$F_i^u(w)=F_i(w)+\frac{\lambda}{2}Fisher\left(w_t^i-w_{t-1}\right) \quad (24)$$

其中λ表示约束强度；$Fisher(\cdot)$表示使用Fisher信息矩阵计算模型中每个参数的重要性因子之和，用于限制参数改变的幅度.

除了常用的机器学习模型，训练梯度校正方法还应用于变分机器学习模型的忘却学习. 变分贝叶斯联邦学习[66]是针对概率模型的联邦学习，其求解的目标为最小化整体的自由能，利用概率分布理论求解模型参数分布$q(\theta)$的特殊联邦学习算法：

$$min_{q(\theta)}\{F(q(\theta))=\sum_{k=1}^{K}E_{\theta\sim q(\theta)}\left[L_k(\theta)\right] +\alpha\cdot D_{KL}(q(\theta)\|p_0(\theta))\} \quad (25)$$

其中$\alpha>0$为超参数；$D_{KL}(\cdot\|\cdot)$表示KL距离[61]；$p_0(\theta)$为先验分布.

针对变分贝叶斯联邦学习进行忘却学习，已有学者对其进行了广泛的研究. 例如，Gong等人[67]在针对分类任务中提出Forget-SVGD算法. 该算法通过改变局部训练的目标函数来实现联邦忘却学习. 当用户u_i请求中心服务器撤销本地数据集D_i^r对模型的贡献时，中心服务器在后续的训练过程中最小化忘却学习的自由能[68]：

$$min_{q(\theta)}\{F(q(\theta))=\sum_{k=1}^{K}E_{\theta\sim q(\theta)}\left[-L_k(\theta)\right] +\alpha\cdot D_{KL}(q(\theta)\|p(\theta))\} \quad (26)$$

其中$\sum_{k=1}^{K}E_{\theta\sim q(\theta)}\left[-L_k(\theta)\right]$是参数分布在忘却数据集上的梯度上升，$D_{KL}(q(\theta)\|p(\theta))$是为了限制参数改变的幅度.

面向局部模型的联邦忘却学习算法通过对用户上传的局部模型参数或用户本地训练的梯度进行修改，利用模型聚合修改全局模型参数，从而实现联邦忘却学习. 相比于面向全局模型算法，面向局部模型算法能够借助已训练的数据信息，因此能够保证模型的精度不会快速下降.

表2对比了面向局部模型的代表算法在算法性能、适用模型、发起者和忘却学习请求类型的不同. 面向局部模型算法的发起者均是用户，这是因为此类算法是对用户的局部模型或产生的梯度进行修改以保护用户隐私. 训练更新校正算法需要占用较大的空间，并且遗忘效果较差，但适用更多类型的模型. 训练梯度校正算法占用空间较少，稳定性较高，并且支持所有的忘却学习请求，这是因为大部分的计算工作是在用户本地进行的，服务器只需要进行聚合操作就可以实现忘却学习.

表2　面向局部模型的联邦忘却学习算法对比

类型	算法名称	算法性能				适用模型	发起者		忘却请求		
		遗忘效果	忘却速度	占用空间	稳定性		用户	服务器	样本	类别	任务
训练更新校正	FedEraser[59]	○	◐	●	○	机器学习模型	√	×	√	×	√
	VERIFI[18]	◐	◐	◐	◐	机器学习模型	√	×	√	×	√
训练梯度校正	Forsaken[62]	◐	◐	○	◐	分类模型	√	×	√	√	√
	UPGA[63]	◐	◐	○	●	机器学习模型	√	×	√	√	√
	Forget-SVGD[67]	◐	◐	○	◐	贝叶斯机器学习	√	×	√	√	√

3.3　联邦忘却学习算法对比总结

本节根据修正对象将联邦忘却学习算法分为面向全局模型和面向局部模型算法两类，阐述忘却学习的原理和发展方向，对不同联邦忘却学习算法进行对比，并分析它们的优点和不足. 在联邦学习过程中，全局模型和局部模型是联邦学习中仅可以共享部分，分别对应面向全局模型和局部模型的忘却学习算法，并且这两种算法是相互互斥的，可以全面地对现有的忘却算法进行分类.

如图10所示，我们复现各类算法在MINIST数据集[69]上对比面向全局模型和面向局部模型算法执行忘却学习操作时模型准确率的变化. 我们分别选择Rapid Retraining[16]和UPGA[63]作为面向全局模型和面向局部模型联邦忘却学习算法代表去删除某一类数据对模型的影响，在剩余类别上进行准确率评估. 此外，面向全局模型算法中局部参数调整是对全局模型的参数进行局部性的调整，而其他算法均是对全局模型进行整体性的调整，鉴于其特殊性，

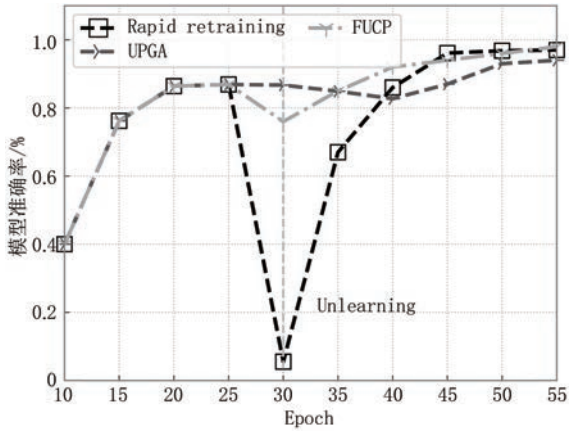

图10 联邦忘却学习后准确率变化曲线

我们增加FUCP[17]作为局部参数调整算法的代表进行对比. 我们进行一定的联邦学习训练,并在Epoch为30时开始联邦忘却学习进而删除某一类数据对模型的参数更新. 根据模型准确率的变化,我们发现:

(1)面向全局模型的算法会存在模型准确率急剧下降的问题,而面向局部模型避免了这一问题. 此外,局部参数调整方法能够精确地删除目标数据对模型的贡献,在此基础上进行再训练. 该方法借助模型结构和训练的数据信息,能够取得最佳的效果.

(2)面向全局模型的算法大部分时间用于机器学习模型的准确率恢复,而面向局部模型大部分时间用于撤销目标用户对模型的训练更新.

表3总结不同种类联邦忘却学习算法的优点、缺点、解决问题以及适应场景,具体如下:

表3 不同种类联邦忘却学习算法对比

忘却方式	类型	优点	缺点	解决问题	适用场景
面向全局模型	重新训练	能够有效地消除用户数据对模型的贡献	模型准确率会在短时间内大幅下降;随着客户端的增加,时间和通信量大幅增加	保护用户隐私,消除用户数据对模型的影响	适用于用户数量少、不考虑网络延迟、高度关注隐私等场景
	用户贡献删除	能够快速地实现联邦忘却学习,有效地删除特定用户对模型的贡献	占用服务器的大量空间;模型准确率会在短时间内大幅下降	忘却学习时间长	适用于服务器具有高计算能力和存储空间等场景
	局部参数调整	能够有效、快速地删除目标数据对模型的贡献	仅适用于特定的模型结构;占用服务器的计算资源	解决模型仍然保留部分用户数据贡献	仅适用于特定任务、特定模型的忘却学习的场景
面向局部模型	训练更新校正	能够快速地删除目标数据对模型的贡献	占用服务器的大量空间;模型仍然保留部分用户数据的贡献	忘却学习时间长	适用于低延迟、开销相对较小、服务器具有高计算能力和存储空间的场景
	训练梯度校正	仅通过一定的联邦学习训练删除用户隐私,能够快速删除用户对模型的数据贡献	模型仍然保留部分用户数据的贡献;部分算法仅适用于限定的任务,在其他任务不一定有效	忘却学习大幅改变模型参数,降低模型准确率	对模型准确率和泛化性能要求较高的场景

(1)面向全局模型的联邦忘却学习算法虽然能够有效地删除目标用户对全局模型的训练更新,但模型准确率会在短时间内大幅降低,而且恢复后的模型难以在短时间内达到撤销模型更新之前的映射效果,主要适用于不考虑用户延迟、高度关注用户隐私等场景. 此外,局部参数调整方法可以利用模型的结构信息和训练产生数据信息,因此在特定的模型上往往表现出良好的忘却学习效果,同时能够尽可能地减少准确率损失,但其仅适合特定的模型结构,存在着一定限制.

(2)面向局部模型的联邦忘却学习算法利用模型训练产生的数据信息,在此基础上通过训练实现联邦忘却. 因此,面向局部模型的联邦忘却学习算法可以防止模型准确率的急剧下降. 面向局部模型算法适用于服务器具有较高计算能力和空间等场景.

4 性能指标

本节详细阐述评价联邦忘却学习算法的性能指标,将性能指标分为模型表现指标、遗忘效果指标以及隐私保护指标. 模型表现指标的研究对象是机器学习模型,用于评估联邦忘却学习算法对机器学习模型的影响程度,包括准确率、损失函数等;遗忘效果指标的研究对象为忘却学习算法,遗忘效果是对联邦忘却学习算法执行效率和数据贡献删除能力的评价,包括相对熵、遗忘率等;隐私保护指标是联邦忘却学习在隐私保护方面抵抗攻击入侵能力的评

价，其研究对象是模型的鲁棒性.

4.1 模型表现指标

4.1.1 准确率

准确率(Accuracy)作为机器学习的重要指标，在联邦忘却学习中同样重要. 准确率用于衡量训练所产生的分类器在测试集中的预测能力，即多少的样本被分类器正确地预测. Liu等人[50]提到联邦忘却学习在撤销用户数据对全局模型的更新时会导致全局模型准确率大幅下降，从而产生错误的预测，影响整个参与联邦学习系统用户的利益. 因此，准确率评估全局模型的变化，可以有效评价联邦忘却学习算法对模型的影响. 针对于分类任务而言，准确率的计算如公式(27)：

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{27}$$

其中 TP, TN, TP, FN 分别表示真阳性、真阴性、伪阳性以及伪阴性的属性数量.

准确率差[62](Accuracy Difference)是联邦忘却学习前后全局模型在测试集中的预测准确率差值，能够直观地判断联邦忘却学习前后对全局模型准确率变化，其计算公式如公式(28)所示：

$$Accd=Acc-Acc_{unlearning} \tag{28}$$

与准确率相关，联邦忘却学习还存在两个使用较少的度量指标：召回率[70](Recall Rate，RR)和F1-Score[71](F1)，分别用于评估模型对正样本的预测能力，以及评估模型对正样本和准确率之间综合的预测能力，计算公式如(29)和(30)所示：

$$RR=\frac{TP}{TP+FN} \tag{29}$$

$$F1=\frac{2TP}{2TP+FP+FN} \tag{30}$$

4.1.2 损失函数

损失函数(Loss Function)通常是非负实值函数，表示训练样本的真实值与预测值之间的误差. 目前损失函数可分为基于距离度量的损失函数和基于概率分布的损失函数.

(1)基于距离度量的损失函数通常将数据映射为基于距离度量的特征空间上的样本点，并在特征空间上计算样本真实值和预测值之间的距离. 常用的损失函数包括均方误差(MSE)[72]、曼哈顿距离(L1损失)[73]等. 例如在回归任务中，数据 $\{(x_k,y_k)\}_{k=1}^{N}$ 与预测样本 $\left\{f(x_k)\right\}_k^N$ 的 MSE 误差为

$$MSE=\frac{1}{N}\sum_{i=k}^{N}(y_k-f(x_k))^2 \tag{31}$$

(2)基于概率分布度量的损失函数利用样本间的相似性，度量样本的真实分布与预测分布之间的距离. 该类损失函数在分类模型中应用广泛，常用的损失函数包括交叉熵损失[74]、焦点损失[75]等. 例如在分类任务中，数据 x_k 在类别 O 上的真实分布 p_{x_k} 和预测分布 q_{x_k} 的交叉熵损失为

$$H\left(p_{x_k},q_{x_k}\right)=-\sum_{j=1}^{|O|}p_{x_k}\left(o_j\right)\log\left(q_{x_k}\left(o_j\right)\right) \tag{32}$$

其中 $p_{x_k}(o_j)$ 为数据 x_k 预测为类别 o_j 的概率.

4.2 遗忘效果指标

4.2.1 相对熵

相对熵[60](Relative Entropy)，也称为KL距离(Kullback-Leibler Divergence，KL)，用于衡量两个概率分布相似性的评价指标. 一般情况下，KL距离越小说明忘却学习算法的效果越好. 例如，计算重新训练模型和忘却学习后模型的KL距离，如公式(33)所示：

$$D_{KL}(p\|q)=\int p(w)(\log p(w)-\log q(w))\mathrm{d}w \tag{33}$$

其中，$p(w)$ 和 $q(w)$ 分别表示重新训练后的参数分布与联邦忘却学习算法后的参数分布.

4.2.2 遗忘率

遗忘率[62](Forgetting Rate，FR)是联邦忘却学习前后“已知”和“未知”之间的转换率，衡量在联邦忘却学习前后有多少样本从全局模型的记忆集变为未知集，其计算公式如(34)所示：

$$FR=\frac{\mathrm{BT}}{\mathrm{BT}+\mathrm{BF}}\times\frac{\mathrm{AF}}{\mathrm{AT}+\mathrm{AF}}\times 100\% \tag{34}$$

其中BT、BF分别表示执行联邦忘却学习前全局模型在数据集中记忆和未记忆的样本数量；AT、AF分别表示执行联邦忘却学习后全局模型在数据集中记忆和未记忆的样本数量.

4.2.3 曝光误差

在语言模型中，曝光误差[76](Exposure)是测量模型对给定序列 $s[r]$ 记忆程度的指标，其计算公式如(35)所示：

$$exposure_\theta\left(s[r]\right)=-\log_2\int_0^{P_\theta(s[r])}\rho(x)\mathrm{d}x \tag{35}$$

其中 $\rho(x)$ 表示一个具有均值 μ、标准差 σ 和斜度 α 的斜正态函数，$P_\theta\left(s[r]\right)$ 表示在机器学习模型 f 下序列 $s[r]$ 的对数困惑度，计算公式为

$$P_\theta\left(s[r]\right)=\sum_{i=1}^{n}(-\log_2 Pr(x_i|f(x_1,\cdots,x_{i-1}))) \tag{36}$$

其中 $Pr(x_i|f(x_1,\cdots,x_{i-1}))$ 表示在模型输入前

$(i-1)$序列，预测出现单词x_i的概率.

4.2.4 时间

时间(Time)就是联邦忘却学习的执行时间，用于衡量联邦忘却学习算法的执行效率. 联邦忘却学习的执行时间主要包括撤销数据对模型更新的时间T_u和联邦学习提升模型精度的时间T_p. 其中，T_u是指在满足定义1执行所需要的忘却时间，T_p是指在剩余的数据集上继续训练使模型在测试集上的准确率达到给定阈值δ的时间.

不同的算法在T_u和T_p上的表现不同. 面向全局模型的联邦忘却学习算法执行时间主要花费在T_p，T_u的执行时间几乎可以忽略，例如Rapid Retraining[16]直接初始化参数就能够完成撤销用户数据对模型的训练更新，但需要大量时间进行重新训练；面向局部模型的联邦忘却学习在T_u和T_p上的时间花费相对均匀，与忘却学习的阈值ε和准确率阈值δ有关.

4.3 隐私保护指标

4.3.1 信息损失

信息损失[77]是攻击者可以获得的信息量，它是对个体隐私性的预先评估. 隐私保护程度越高，攻击者可以获得的信息越少. 信息损失主要通过考虑攻击者的推断推理能力、数据相似性和分辨能力等方式进行计算.

基于攻击者推断能力的信息损失通过估计攻击者可以获得的信息量，来评估自身的隐私泄露风险. 例如，泄露信息数量[78]仅估计自身数据可能泄露的数量，计算公式如(37)所示.

$$pri_{ali}=|D_{disclose}| \tag{37}$$

其中$|D_{disclose}|$表示可能泄露的个体数据数量. 如果可以估计攻击者的推测数据Y，那么用户可以计算真实分布X与推测分布Y之间的相对熵来判断攻击者的估计与真实样本之间的距离. 除此之外，互信息[79]也可以根据真实分布X与推测分布Y测量从隐私机制中泄露的信息量，计算公式如(38)所示.

$$pri_{ml}=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{38}$$

除此之外，条件互信息[80]和系统匿名度[81]等指标作为互信息的进一步发展，可以为用户提供更多维度的数据敏感性度量标准.

基于数据相似性的信息损失是以观察到的或公开发布的数据属性进行度量，独立于攻击者，仅从公开数据的特征得出隐私级别. 例如，k-匿名性[82]是一种用于评估数据集隐私保护级别的量化指标. 如果一个数据集满足k-匿名性，那么至少有k条记录在所有标识符属性上与任何特定记录相同. 因此，k的大小可以视为数据集隐私保护级别的量化指标，k值越大，提供的隐私保护级别就越高. 为了改善基于属性攻击或背景知识攻击的抵抗能力，研究人员提出l-多样性[83]，这一概念要求数据集中的每一个等价类(在所有标识符属性上相同的记录集合)，其敏感属性包含至少l种不同的值. 这意味着在任何标识符属性相同的记录集合中，敏感属性的值有足够的多样性，从而防止攻击者利用背景知识准确地确定个体的敏感信息. 在此基础上，研究者也提出t-贴近度[84]、(k,e)-匿名[85]等指标，以解决在不同需求下的隐私评估问题.

基于分辨能力的信息损失是对攻击者区分两个感兴趣数据的能力评价，与正式的隐私机制有直接关联. 例如，差分隐私[86]的定义依赖于用两个最多只差一行的数据集D_1和D_2，即两个数据集之间的汉明距离[87]最多为1. 如果对于所有的查询响应S，一个随机函数K运行在D_1和D_2数据集，两个数据集的查询响应的最大差值为$exp(\epsilon)$，则称函数K满足ϵ-差分隐私. 在ϵ-差分隐私下，隐私强度的计算公式可以表示为

$$priv_{DP}\equiv\forall S\subseteq Range(K):p\big(K(D1)\in S\big)\leqslant exp(\epsilon)\cdot p\big(K(D2)\in S\big) \tag{39}$$

基于差分隐私原则，研究学者还进一步扩展出近似差分隐私[88]、分布式差分隐私[89]等指标衡量数据集泄露的可能性.

在现代的机器学习中，信息损失指标在联邦学习中被广泛应用，以量化数据的隐私程度，并确保用户在数据处理和机器学习过程中的敏感信息不被泄露. 然而，信息损失指标仅专注于特定数据和攻击者，与忘却学习在修改全局模型参数上的方式有本质区别，因此在评估联邦忘却学习算法时使用较少.

4.3.2 攻击抵抗

联邦学习在学习过程中容易受到攻击者的攻击，攻击者发起攻击主要用于模型攻击和数据窃取. 模型攻击者可以通过改变他们的局部模型参数、数据等方式来操纵全局模型的更新，影响着模型安全，导致模型准确率降低，例如后门攻击[90]、拜占庭攻击[91]等. 此外，数据窃取攻击者通过观察全局模型的更新，推断出其他参与者的数据信息，威胁用户的隐私安全，例如成员推理攻击[92]、模型窃取攻

击[93]等.

联邦忘却学习通过撤销数据对模型的训练更新，保护模型安全和用户隐私. 联邦忘却学习可以撤销攻击者的恶意数据或恶意模型，使全局模型脱离攻击者的威胁. 除此之外，联邦忘却学习可以在模型中消除全局模型中用户隐私数据的踪迹，进而保护数据隐私安全. 目前，联邦忘却学习撤销数据抵御攻击的研究主要集中在后门攻击和成员推理攻击两种攻击方式.

后门攻击成功率[90](Backdoor Attack)即后门攻击成功次数与后门攻击发起次数之比. 模型后门是通过参与训练得到的深度神经网络中的隐藏模式. 后门攻击是在训练期间设置模型后门，使模型产生错误的预测：在后门未激活的情况下，恶意节点与正常节点具有相同的表现；而一旦后门被激活，训练模型的输出将变为攻击者预先设置的标签. 例如Wu等人[50]在MINIST数据集中修改一类数据用于训练，向模型添加隐藏模式，使模型在标签为“1”的样本预测为其他标签. 为抵御这一攻击，Wu等人通过FUKD算法撤销攻击者训练数据对全局模型的更新，在MINIST数据集上测试，成功将后门攻击成功率从97.7%降为0%.

成员推断攻击[21,92](Membership Inference Attack)本质是判断输入数据是否用于目标模型训练的二分类问题. 具体来说，攻击者构建判断模型$\mathcal{D}(w)$，根据不同输入数据$\{x_1, x_2, \cdots, x_n\}$在目标模型$f(w)$上的预测行为来训练模型$\mathcal{D}(w)$，其中$w$表示生成模型或判断模型的参数.

攻击者通过执行推理攻击来推断数据集D中的数据是否被用于训练目标模型. 具体来说，一次成员推理攻击是使用已训练的判断模型$\mathcal{D}(w)$对数据集D中的单个测试数据点x'进行预测，即计算$D(w, x')$. 如果$D(w, x')=1$，那么可以判断数据点x'为目标模型$f(w)$的训练数据；如果$D(w, x')=0$，那么数据点x'则被判断为目标模型$f(w)$的非训练数据. 在这种情况下，如果预测准确，那么成员推理攻击被视为成功，否则被视为失败. 值得注意的是，即使攻击者对目标模型$f(w)$的参数、结构不了解，该攻击仍然有效. 成员推断攻击成功率即成员推断攻击成功次数与成员推断攻击发起次数之比，计算方法如(40)所示.

$$Acc_{inference}=\frac{TP_{inference}+TN_{inference}}{|D|} \tag{40}$$

其中$TP_{inference}$表示$x'\in D$作为真实训练样本时$\mathcal{D}(w, x')$预测为1的次数，$TN_{inference}$表示$x'\in D$作为非训练样本时$\mathcal{D}(w, x')$预测为0的次数.

针对成员推理攻击的威胁，联邦忘却学习可以通过减少目标模型包含的训练数据信息抵御攻击. Wang等人[17]根据模型对样本类别的记忆程度测量成员推断攻击的成功率，并通过FUCP算法删除特定类别对神经网络贡献度较高的参数. 与重新训练相比，FUCP在CIFAR10数据集[94]上成员推断攻击成功率降低0.63%.

4.4 联邦忘却学习性能指标对比总结

本小节根据研究对象的不同将联邦忘却学习的评价指标分为模型表现指标、遗忘效果指标和隐私保护指标. 表4列出了常用性能指标的类型、优点和缺点. 模型表现指标只能判断联邦忘却学习算法对模型预测能力的影响程度，难以计算对目标用户数据预测能力的影响，因此无法判断对目标用户数据的遗忘程度，需要结合其他类别的指标综合使用. 现有遗忘效果指标能够反应模型对数据的记忆程度，但部分指标局限于特定模型，例如遗忘率仅适用于分类模型.

此外，由于不同数据对联邦学习模型的贡献不同，模型表现指标、遗忘效果指标以及隐私保护指标三者在处理不同数据忘却请求时的变化关系有所不同. 当被请求遗忘的数据对全局模型有积极贡献时，随着遗忘效果指标的提高，由于有价值数据的移除，模型表现指标会逐渐下降，同时，模型抵御攻击的能力会提升. 另一方面，当被请求遗忘的数据对模型有消极影响(如低质量数据或随机生成的数据)时，随着遗忘效果指标的提升，模型表现指标会逐步提高，但由于模型减少了对消极数据的记忆，增强对其他数据的记忆，其抵御攻击的能力会降低.

三种指标分别在联邦忘却学习的不同阶段衡量联邦忘却学习算法的能力. 在忘却学习准备阶段和执行阶段中，模型表现指标通过监视模型预测能力在忘却算法实施过程中的动态变化，从而衡量忘却学习模型对模型性能的影响. 遗忘效果指标是在忘却执行阶段和完成阶段考察，旨在评估机器学习算法对数据记忆程度的变化，以此来衡量忘却学习算法的效果. 最后，隐私保护指标在忘却完成阶段进行评估，以了解忘却学习后机器学习模型在保护数据隐私方面的性能.

表4 联邦忘却学习评价指标对比

类型	名称	评估忘却学习阶段			优点	缺点
		准备阶段	执行阶段	完成阶段		
模型表现指标	准确率/准确率差	√	√	×	能够衡量忘却学习对模型预测准确的影响	无法衡量联邦忘却学习算法性能，只能评估对模型预测能力的影响
	召回率	√	√	×	能够衡量模型预测正样本的能力	
	F1-score	√	√	×	能够在准确率和正样本预测中进行综合地衡量忘却学习对模型的影响	
	损失函数	√	√	×	能够更细致地计算忘却学习在模型预测值与真实值的偏离	无法直接衡量忘却学习对模型预测能力的影响
遗忘效果指标	相对熵	×	√	√	通过与模型参数真实分布的比较，精准地反映忘却学习的效果	一般无法预知模型参数的真实分布，只能通过重新训练得到
	遗忘率	×	√	√	能够判断联邦忘却学习前后模型对训练数据的记忆程度	无法衡量真实值偏离的大小
	曝光误差	×	√	√	能够直接衡量模型的忘却程度	只能局限于语言模型
	时间	×	√	√	能够直接地衡量忘却学习的执行速度	不能直接衡量忘却学习算法的执行效率，需要与重新训练进行对比
隐私保护指标	后门攻击成功率/成员推断攻击成功率	×	×	√	能够衡量模型对入侵攻击的抵抗能力	无法衡量忘却学习算法对模型的影响

5 联邦忘却学习应用

本节从遗忘价值的角度，探讨联邦忘却学习的应用. 在联邦忘却学习中，用户和服务器是两个关键参与者. 从用户的角度看，移除用户数据在模型中的影响可以减弱模型对该数据的记忆，从而有助于保护用户隐私. 从服务器的角度来看，消除劣质数据对全局模型的影响，可以显著提升模型的泛化能力. 因此，我们将从用户隐私保护和模型攻击抵抗两个方面来详细阐述联邦忘却学习的应用场景.

5.1 隐私保护

联邦学习模型通过训练更新产生对用户数据的记忆，即使用户退出联邦学习系统，机器学习模型仍然保留因数据训练模型而产生的用户数据记忆，进而存在用户数据隐私泄露的风险. 针对这一问题，联邦忘却学习可以为退出联邦学习系统的用户撤销其数据对模型的训练更新，因此用户可以放心地进行数据流通，不必担心自身退出系统后的隐私泄露问题.

例如，在不同企业间进行机器学习模型的协同训练时，可能会出现利益冲突. 在训练过程中，恶意企业有可能通过对模型进行逆向推导，获取其他企业的隐私数据，从而导致数据泄露和企业损失. 此外，一旦企业退出合作，机器学习模型仍会保留其数据贡献，这可能引发进一步的利益纠纷. 为解决此问题，Liu等人[55]针对随机森林模型提出了一种联邦学习算法，RevFRF. 在撤销目标用户数据贡献后，全局模型将不再包含该用户的数据信息，同时目标用户也无法再次访问联邦学习模型. 因此，RevFRF实现了对目标用户数据的隔离，能够保护用户隐私.

Yuan等人[95]提出一种高效的联邦忘却学习方法FRU，去解决联邦推荐系统的数据安全问题. 联邦推荐系统[96]旨在保护用户数据隐私的前提下为用户提供个性化的推荐服务. 这类系统通过联邦学习技术将用户数据分布式存储在各个设备上，降低了数据泄露的风险. 然而，尽管现有的联邦推荐系统在保护用户隐私方面取得一定的成果[97-99]，但它们还没有充分考虑如何有效地消除用户对机器学习模型的贡献. FRU基于FedEraser[59]的思想，通过回滚和校准历史更新来移除用户数据对模型的贡献. 考虑到个人设备存储资源的有限性，FRU仅需存储重要的模型更新就可以实现忘却学习. 在实践中，在联邦推荐系统中提供一种有效的忘却学习策略，既可以保护用户隐私，又可以维持推荐系统的性能. 未来的研究可以继续关注联邦推荐系统中的忘却学习问题，进一步优化算法性能和隐私保护能力，以满足不断增长的用户隐私需求.

5.2 攻击抵抗

数据中毒[28]会引起联邦学习模型准确率下降. 在联邦学习的迭代训练过程中，所涉及的数据可能会失效、被污染、过时或被数据中毒攻击操控. 联邦学习的全局模型将会因学习错误数据的贡献而造成准确率下降. 因此，采用联邦忘却学习撤销这些错误数据所产生的模型更新，可以提高模型的泛化性能.

例如，联邦知识图嵌入[100]可以在保护各客户端本地数据隐私的同时，充分挖掘和整合来自不同客户端的知识，从而为知识图提供更为丰富和全面的信息. 然而，联邦知识图嵌入在实际应用中面临诸多挑战，如数据异构[101]、知识遗忘[102]等问题. 其中，数据异构是指不同客户端中的数据可能具有不同的结构、格式和语义，这可能导致在整合过程中出现信息丢失或模型性能下降. 知识遗忘是指当用户希望删除与某些实体或关系相关的知识时，应当确保这些知识能够从知识图嵌入模型中完全移除，以保护数据隐私. 为了解决这一问题，Zhu等人[52]基于回溯干扰和被动衰减的方法在客户端删除特定的知识，并扩散到全局模型中. 忘却学习完成后的模型在链接预测和知识遗忘方面都取得了更好的效果.

联邦聚类[103]对分布式数据进行聚类分析，能够有效地挖掘数据中隐藏的模式和规律. 针对模型中毒攻击[26]的威胁，为确保能够提供更高泛化性能的模型，服务器端实现安全的局部模型聚合在联邦学习中显得尤为关键. 在这一背景下，解决联邦聚类中的忘却学习问题显得尤为重要. 忘却学习的目标是在不影响模型整体性能的前提下，从模型中移除特定数据或用户信息. 为了实现这一目标，Pan等人[104]提出通过采用特殊的初始化操作来完全删除用户的模型更新，并在服务器端进行安全的自监督训练. 然而，这种方法在实现联邦忘却学习时需要较长的时间，因此未来的研究仍需进一步探讨优化方案，以提高算法效率并更好地保护用户隐私.

6 未来展望

为进一步推动数据要素在市场的流通，保护用户隐私和数据安全已经成为一个重要的科学研究问题. 联邦忘却学习通过遗忘用户本地训练的模型更新，提供保护用户隐私的新思路. 本节结合现有的工作，从联邦忘却学习算法和其应用角度进行未来展望.

6.1 联邦忘却学习算法

6.1.1 泛化高效的联邦忘却学习算法

泛化高效的联邦忘却学习算法需要被广泛地研究. 面向全局模型的算法虽然可以很好地撤销特定用户数据对模型的训练更新，但全局模型的准确率会在短时间内大幅下降；面向局部模型的算法虽然避免准确率的迅速下降，但全局模型仍然有可能包含目标用户对模型的训练更新. 然而，在实际应用中用户可能需要将联邦忘却学习应用于不同的机器学习模型. 由于不同模型的结构存在差异，忘却学习算法在不同模型上的表现也会具有优劣之分，这导致忘却学习算法无法为所有用户提供高质量服务. 综合所有类型算法的优缺点，设计友好的、泛化高效的联邦忘却学习应用方案尤为重要. 在设计泛化高效的联邦忘却学习算法时，应该考虑不同机器学习的结构、输入输出等特征，例如贝叶斯机器学习是推测样本的分布，与机器学习的值预测存在较大差异. 此外，忘却学习算法也应该考虑不同的机器学习任务[105]，如分类、回归、聚类等. 不同机器学习任务的目标不同，所要求的忘却性能、忘却请求也不同，这是泛化高效联邦忘却学习算法设计所面临的难点. 因此，设计泛化高效的联邦忘却学习算法存在很大的挑战.

6.1.2 局部参数调整忘却学习算法

局部参数调整方法可作为当前的备用方案. 鉴于泛化高效算法设计的困难性，局部参数调整方法可利用模型结构，寻找与特定用户数据高度相关的模型参数位置再删除或剪枝. 例如，FUCP[17]针对卷积神经网络进行联邦忘却学习，利用TF-IDF衡量用户数据对各通道的贡献，通过初始化具有高贡献的通道参数进而快速地完成忘却学习.

局部参数调整方法最重要的一步是去定位与用户数据高度相关的模型参数. 受到机器学习模型可解释性的限制，我们很难仅利用训练过程中的模型参数去发现这些与特定用户数据高相关性的参数位置[106]. 针对这一问题，联邦忘却学习算法可以考虑与高可解释性的数学方法相结合，例如奇异值分解[107]、主成分分析[108]等. 我们通过将可解释性算法参与学习、训练等方式去捕获机器学习模型的结构信息. 例如，可以通过奇异值分解将机器学习模型的参数矩阵进行分解，得到关键的秩. 根据秩去寻找与用户高相关度的模型参数.

针对现有算法进一步探索，我们发现局部参数调整的思想可在面向局部模型算法中得到扩展应

用. 现阶段,局部参数调整的思想仅在面向全局模型的方法中使用,这样直接操作全局模型参数可能会损失模型参数之间的关键信息,进而影响全局模型的准确率. 研究人员可以将局部参数调整的思想引入到面向局部模型算法中,用户通过本地训练过程来调整局部模型的参数. 在此基础上,服务器可以聚合用户局部模型去修改全局模型的特定参数,进而实现联邦忘却学习. 与直接修改全局模型参数的方式相比,其可以在精准定位用户数据贡献的同时,保留模型内部参数间的信息联动性,从而维持模型的整体性能.

6.1.3 联邦忘却学习验证

为了更有效地保护用户隐私,建立联邦忘却学习和忘却学习验证的体系至关重要. 忘却学习验证旨在确认特定数据的训练更新是否已成功地从模型中移除,这有助于确保数据安全和用户隐私得到充分保护. 尽管如此,在现有研究中,联邦忘却学习验证的工作相对较少[18],而大部分忘却学习研究方案仅通过不同的评价指标来衡量算法性能,而忽略了忘却学习验证的研究. 为了让用户更加信任忘却学习算法能够消除机器学习模型对他们隐私数据的记忆,深入研究联邦忘却学习的验证工作显得尤为重要.

为了实现联邦忘却学习的验证,未来研究可以制定统一的评价指标,以便在不同的忘却学习研究方案中进行比较和评估,例如可以使用相对熵判断机器学习对某一用户数据的记忆程度,根据记忆程度来验证是否完成忘却学习.

6.2 联邦忘却学习应用

6.2.1 数据要素的市场流通

随着人工智能的不断发展和应用,越来越多的企业和机构需要利用海量的数据进行训练和优化,以提高人工智能算法的性能和应用效果. 然而,数据的共享和流通过程中的隐私泄露问题[23-25]成为制约数据要素市场流通的重要因素.

未来的研究可以利用联邦忘却学习实现数据要素的市场流通. 联邦忘却学习算法将作为联邦学习系统的重要组成部分,为数据隐私保护和数据安全共享提供重要支持. 联邦忘却学习允许用户随时退出联邦学习系统,同时无需担心全局模型中仍然存在自身隐私数据的痕迹. 因此,用户的数据隐私得到充分保护,用户可以放心地将自身数据流通并用于使用.

6.2.2 低质量数据删除

由于人的不可靠性,大量的低质量数据可能会被引入到机器学习训练中. 由于人的主观偏见、知识局限等因素[109]的影响,训练数据在收集、处理和标注过程中,可能存在标注错误、数据异常等问题. 在这些低质量数据上进行训练会影响模型的准确率,进而使模型进行错误的预测和识别,影响用户体验. 为了避免重新训练,通过联邦忘却学习去解决低质量数据的问题并提升用户体验将成为未来研究趋势.

联邦学习中识别低质量数据仍需进一步的研究. 当前的识别低质量数据的工作,例如贡献值[110],需要频繁的数据流通才能够评估用户的数据质量. 此外,如果所有用户均具有一定数量的低质量数据,根据贡献值的评估方法是不奏效的[111]. 因此,如何设计有效的低质量数据鉴别方法是阻碍联邦忘却学习去除低质量数据对模型影响的关键挑战. 未来的研究可以探讨在联邦学习环境下,使用模型损失值来识别低质量数据. 例如,研究者可以尝试使用模型在特定数据训练前后损失值的正负变化来衡量数据的质量.

6.2.3 跨域机器学习的忘却应用

联邦忘却学习在数据可用不可见下实现跨设备的数据遗忘,基于此跨设备的隐私保护方法,未来研究可以针对相似的跨域机器学习场景采用忘却学习实现进一步的数据隐私保护. 例如,域自适应任务中黑盒域自适应方法[112-113]实现源域(源数据分布)转移到目标域(目标数据分布)的知识迁移,在保护源域数据隐私的基础上完成目标域无标签数据的机器学习. 黑盒域自适应方法和联邦忘却学习都具备数据可用不可见性、跨域机器学习的特性,但二者的区别在于,前者采用迁移学习的模式,需要利用带有标签的数据的知识去学习其他领域无标签的数据,而联邦忘却学习则是在多个带有标签的数据下共同训练一个机器学习模型. 根据联邦忘却学习的思想,可以利用源域的模型和目标域的数据,对模型参数进行修改,实现在目标域学习完成后的模型中删除源域的知识,进而保护源域的数据隐私.

6.2.4 定制化服务

利用联邦忘却学习可以为用户提供定制化的服务. 当前,联邦忘却学习的研究重点主要集中在用户数据隐私[16]和模型安全[50]方面. 事实上,联邦忘却学习可以根据用户的偏好为其提供定制化的机器学习模型. 因为用户偏好[114]在很大程度上反映了他

们的个性化需求和特定兴趣，这些需求和兴趣在不同的用户之间可能存在显著差异．为了提供定制化的服务，我们可以在训练过程中捕获用户的偏好信息．在训练完成后，我们可以通过联邦忘却学习来忘却那些与用户偏好差距较大的数据．据此，提供给用户的模型将更加贴合其个性化需求和特定兴趣，从而为用户实现定制化的模型．

用户偏好信息在联邦学习框架下难以获取[115]．直接获取用户数据判断训练数据相关性违背了联邦学习的范式，并且为多用户提供服务也会为中心服务器带来压力，因此根据数据直接判断是不可取的．针对这一问题，研究人员可以将训练数据的相似性转化为训练过程中模型参数的相似性，进而聚合相似的模型参数为用户提供更高质量的模型．

6.2.5 特殊场景下的联邦忘却学习

针对不同场景下的联邦忘却学习方法需要进一步的研究．目前，大部分联邦忘却学习算法都是针对普遍的联邦学习场景进行研究的，没有考虑其他方面的限制，例如有限的带宽[116]、限制的设备资源[117]等．然而，当面临特殊的联邦学习场景，例如有限带宽的元宇宙场景[118]，现有的联邦忘却学习算法难以应用．在这种场景下，张量分解方法，例如CP分解的快速算法和分布式随机Tucker分解方法[119]，可用于压缩和简化联邦学习中的模型，同时可在此基础上设计联邦忘却学习算法，减少元宇宙的数据传输，从而设计出更加有效和实用的联邦忘却学习算法．

7 总 结

数据在使用过程因未建立健全隐私保护机制而受到隐私泄露的威胁．联邦学习作为一种隐私保护技术，可以在保护数据隐私前提下实现数据共享．然而，联邦学习模型在训练过程中保留了用户数据的记忆，攻击者可以利用模型推断数据等方法窃取用户隐私，并攻击联邦学习模型．在此背景下，联邦忘却学习通过撤销用户数据对联邦学习模型训练更新的方式，进一步保护联邦学习的用户隐私和模型安全．本文对联邦忘却学习进行综述，将联邦忘却学习算法分为面向全局模型和面向局部模型两种算法，并对不同类型算法进行详尽的阐述．本文同时也对评价指标进行综述，将联邦忘却学习评价指标分为模型表现指标、遗忘效果指标和隐私保护指标，并进行对比分析．此外，本文阐述联邦忘却学习在实现用户隐私保护和全局模型攻击抵抗的应用．最后，我们对联邦忘却学习算法和应用进行未来展望．

致 谢 感谢《计算机学报》编辑和审稿专家，他们付出了辛勤工作．

参 考 文 献

[1] Pan Z, Nguyen H L, Abu-Gellban H, et al. Google trends analysis of covid-19 pandemic//Proceedings of the IEEE International Conference on Big Data (Big Data). Atlanta, USA, 2020: 3438-3446

[2] Yu Hong, He De-Niu, Wang Guo-Yin, et al. Big data intelligent decision-making. Acta Automatica Sinica, 2020, 46(5): 878-896 (in Chinese)
(于洪，何德牛，王国胤等．大数据智能决策．自动化学报．2020，46(5)：878-896)

[3] Ayaburi E W, Treku D N. Effect of penitence on social media trust and privacy concerns: The case of facebook. International Journal of Information Management, 2020, 50: 171-181

[4] Tripathi M, Mukhopadhyay A. Financial loss due to a data privacy breach: An empirical analysis. Journal of Organizational Computing and Electronic Commerce, 2020, 30(4): 381-400

[5] Hu R, Gong Y. Trading data for learning: Incentive mechanism for on-device federated learning//Proceedings of the IEEE Global Communications Conference. Taipei, China, 2020: 1-6

[6] Zhang Tao, Ma Hai-Qun. Analysis of the themes and development trends of big data policies in China. Information Studies: Theory & Application, 2022, 45(3): 72-80 (in Chinese)
(张涛，马海群．我国大数据政策主题分析及发展动向研判．情报理论与实践．2022，45(3)：72-80)

[7] Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Computing Surveys (CSUR), 2020, 53(4): 1-35

[8] Maurer U. Secure multi-party computation made simple. Discrete Applied Mathematics, 2006, 154(2): 370-381

[9] Elminaam D S A, Kader H M A, Hadhoud M M. Performance evaluation of symmetric encryption algorithms. International Journal of Computer Science and Network Security, 2008, 8(12): 280-286

[10] Peralta G, Cid-Fuentes R G, Bilbao J, et al. Homomorphic encryption and network coding in iot architectures: Advantages and future challenges. Electronics, 2019, 8(8): 827

[11] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017: 1273-1282

[12] Lu Y, Huang X, Dai Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT.

IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186

[13] Ren H, Deng J, Xie X. GRNN: Generative regression neural network—A data leakage attack for federated learning. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(4): 1-24

[14] Liu G, Ma X, Yang Y, et al. Federated unlearning. arXiv preprint arXiv:2012.13891, 2020

[15] Li G, Shen L, Sun Y, et al. Subspace based federated unlearning. arXiv preprint arXiv:2302.12448, 2023

[16] Liu Y, Xu L, Yuan X, et al. The right to be forgotten in federated learning: An efficient realization with rapid retraining//Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications. London, UK, 2022: 1749-1758

[17] Wang J, Guo S, Xie X, et al. Federated unlearning via class-discriminative pruning//Proceedings of the 2022 World Wide Web Conference. Singapore, 2022: 622-632

[18] Gao X, Ma X, Wang J, et al. Verifi: Towards verifiable federated unlearning. arXiv preprint arXiv:2205.12709, 2022

[19] Barnes L P, Inan H A, Isik B, et al. rTop-k: A statistical estimation approach to distributed SGD. IEEE Journal on Selected Areas in Information Theory, 2020, 1(3): 897-907

[20] Chen Y, Ning Y, Slawski M, et al. Asynchronous online federated learning for edge devices with non-iid data//Proceedings of the IEEE International Conference on Big Data (Big Data). Atlanta, USA, 2020: 15-24

[21] Li J, Meng Y, Ma L, et al. A federated learning based privacy-preserving smart healthcare system. IEEE Transactions on Industrial Informatics, 2021, 18(3): 2021-2031

[22] Wang P, Zhao Y, Obaidat M S, et al. Blockchain-enhanced federated learning market with social internet of things. IEEE Journal on Selected Areas in Communications, 2022, 40(12): 3405-3421

[23] Li Q, Wen Z, Wu Z, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(4): 3347-3366

[24] Gong M, Xie Y, Pan K, et al. A survey on differentially private machine learning. IEEE Computational Intelligence Magazine, 2020, 15(2): 49-64

[25] Hu H, Salcic Z, Sun L, et al. Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR), 2022, 54(11s): 1-37

[26] Zhang Y, Jia R, Pei H, et al. The secret revealer: Generative model-inversion attacks against deep neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 253-261

[27] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems//Proceedings of the 25th European Symposium on Research in Computer Security. Guildford, UK, 2020: 480-501

[28] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to {Byzantine-Robust} federated learning//Proceedings of the 29th USENIX Security Symposium (USENIX Security 20. Wilmington, USA, 2020: 1605-1622

[29] Ma Z, Liu Y, Liu X, et al. Learn to forget: Machine unlearning via neuron masking. IEEE Transactions on Dependable and Secure Computing, 2022(1): 1-14

[30] Bourtoule L, Chandrasekaran V, Choquette-Choo C A, et al. Machine unlearning//Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2021: 141-159

[31] Yoon Y, Nam J, Yun H, et al. Few-shot unlearning by model inversion. arXiv preprint arXiv:2205.15567, 2022

[32] Tarun A K, Chundawat V S, Mandal M, et al. Deep regression unlearning//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA, 2023: 33921-33939

[33] Chen C, Sun F, Zhang M, et al. Recommendation unlearning//Proceedings of the 2022 World Wide Web Conference. Singapore, 2022: 2768-2777

[34] Chen M, Gao W, Liu G, et al. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7766-7775

[35] Nguyen T T, Huynh T T, Nguyen P L, et al. A survey of machine unlearning. arXiv preprint arXiv:2209.02299, 2022

[36] Shah V, Träuble F, Malik A, et al. Unlearning via sparse representations. arXiv preprint arXiv:2311.15268, 2023

[37] Poppi S, Sarto S, Cornia M, et al. Multi-class explainable unlearning for image classification via weight filtering. arXiv preprint arXiv:2304.02049, 2023

[38] Wang W, Zheng V W, Yu H, et al. A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-37

[39] Cavazos J G, Phillips P J, Castillo C D, et al. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science, 2020, 3(1): 101-111

[40] Kudithipudi D, Aguilar-Simon M, Babb J, et al. Biological underpinnings for lifelong learning machines. Nature Machine Intelligence, 2022, 4(3): 196-210

[41] Liu Y, Fan M, Chen C, et al. Backdoor defense with machine unlearning//Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications. London, UK, 2022: 280-289

[42] Nguyen Q P, Oikawa R, Divakaran D M, et al. Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten//Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. New York, USA, 2022: 351-363

[43] Nguyen D C, Ding M, Pathirana P N, et al. Federated learning for internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1622-1658

[44] Dong J, Wang L, Fang Z, et al. Federated class-incremental learning//Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 10164-10173

[45] Du M, Chen Z, Liu C, et al. Lifelong anomaly detection through unlearning//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 1283-1297

[46] Golatkar A, Achille A, Soatto S. Eternal sunshine of the spotless net: Selective forgetting in deep networks// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Wilmington, USA, 2020: 9304-9312

[47] Jin R, Xing Y, He X. On the convergence of mSGD and AdaGrad for stochastic optimization. arXiv preprint arXiv: 2201.11204, 2022

[48] Su, Ningxin, and LiBaochun. Asynchronous federated unlearning//Proceedings of the IEEE INFOCOM 2023-IEEE Conference on Computer Communications. New York, USA, 2023

[49] Fraboni Y, Vidal R, Kameni L, et al. Sequential informed federated unlearning: Efficient and provable client unlearning in federated optimization. arXiv preprint arXiv:2211.11656, 2022

[50] Wu C, Zhu S, Mitra P. Federated unlearning with knowledge distillation. arXiv preprint arXiv:2201.09441, 2022

[51] Gou J, Sun L, Yu B, et al. Multilevel attention-based sample correlations for knowledge distillation. IEEE Transactions on Industrial Informatics, 2022, 19(5): 7099-7109.

[52] Zhu X, Li G, Hu W. Heterogeneous federated knowledge graph embedding learning and unlearning//Proceedings of the 2023 World Wide Web Conference. Singapore, 2023: 2444-2454

[53] Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2021, 2(1): 20-28

[54] Kurani A, Doshi P, Vakharia A, et al. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 2023, 10(1): 183-208

[55] Liu Y, Ma Z, Yang Y, et al. Revfrf: Enabling cross-domain random forest training with revocable federated learning. IEEE Transactions on Dependable and Secure Computing, 2021, 19 (6): 3671-3685

[56] Hacohen G, Weinshall D. On the power of curriculum learning in training deep networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 2535-2544

[57] Jabri S, Dahbi A, Gadi T, et al. Ranking of text documents using TF-IDF weighting and association rules mining// Proceedings of the 4th International Conference on Optimization and Applications (ICOA). Mohammedia, Morocco, 2018: 1-6

[58] Singh A, Kushwaha S, Alarfaj M, et al. Comprehensive overview of backpropagation algorithm for digital image denoising. Electronics, 2022, 11(10): 1590

[59] Liu G, Ma X, Yang Y, et al. Federaser: Enabling efficient client-level data removal from federated learning models// Proceedings of the 29th International Symposium on Quality of Service (IWQOS). Tokyo, Japan, 2021: 1-10

[60] Claici S, Yurochkin M, Ghosh S, et al. Model fusion with Kullback-Leibler divergence//Proceedings of the 37th International Conference on Machine Learning. 2020: 2038-2047

[61] Kornilova M, Kovalnogov V, Fedorov R, et al. Zeroing neural network for pseudoinversion of an arbitrary time-varying matrix based on singular value decomposition. Mathematics, 2022, 10 (8): 1208

[62] Liu Y, Ma Z, Liu X, et al. Learn to forget: User-level memorization elimination in federated learning. arXiv preprint arXiv:2003.10933, 2020

[63] Halimi A, Kadhe S, Rawat A, et al. Federated unlearning: How to efficiently erase a client in FL? arXiv preprint arXiv: 2207.05521, 2022

[64] Egwu N, Mrziglod T, Schuppert A. Neural network input feature selection using structured l 2-norm penalization. Applied Intelligence, 2023, 53(5): 5732-5749

[65] Wu L, Guo S, Wang J, et al. Federated unlearning: Guarantee the right of clients to forget. IEEE Network, 2022, 36(5): 129-135

[66] Lee S, Park C, Hong S N, et al. Bayesian federated learning over wireless networks. arXiv preprint arXiv:2012.15486, 2020

[67] Gong J, Kang J, Simeone O, et al. Forget-svgd: Particle-based bayesian federated unlearning//Proceedings of the 2022 IEEE Data Science and Learning Workshop (DSLW). Singapore, 2022: 1-6

[68] Kiefer A B. Psychophysical identity and free energy. Journal of the Royal Society Interface, 2020, 17(169): 20200370

[69] Mu N, Gilmer J. Mnist-c: A robustness benchmark for computer vision. arXiv preprint arXiv:1906.02337, 2019

[70] Cui Z, Zhao Y, Cao Y, et al. Malicious code detection under 5G HetNets based on a multi-objective RBM model. IEEE Network, 2021, 35(2): 82-87

[71] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 2020, 21(1): 1-13

[72] Ren J, Zhang M, Yu C, et al. Balanced mse for imbalanced visual regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 7926-7935

[73] Cheng W, Zhu X, Chen X, et al. Manhattan distance-based adaptive 3D transform-domain collaborative filtering for laser speckle imaging of blood flow. IEEE Transactions on Medical Imaging, 2019, 38(7): 1726-1735

[74] Li Y, Zou Y, Glorioso P, et al. Cross entropy benchmark for measurement-induced phase transitions. Physical Review Letters, 2023, 130(22): 220404

[75] Li X, Lv C, Wang W, et al. Generalized focal loss: Towards efficient representation learning for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3139-3153

[76] Carlini N, Liu C, Erlingsson Ú, et al. The secret sharer: Evaluating and testing unintended memorization in neural

networks//Proceedings of the 28th USENIX Security Symposium (USENIX Security 19. ClaraSanta, USA, 2019: 267-284

[77] Wagner I, Eckhoff D. Technical privacy metrics: A systematic survey. ACM Computing Surveys (CSUR), 2018, 51(3): 1-38

[78] Yussof N A M, Ho H W. Review of water leak detection methods in smart building applications. Buildings, 2022, 12 (10): 1535

[79] Zhou M, Yan K, Huang J, et al. Mutual information-driven pan-sharpening//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 1798-1808

[80] Kuwahara T, Kato K, Brandão F G S L. Clustering of conditional mutual information for quantum gibbs states above a threshold temperature. Physical Review Letters, 2020, 124 (22): 220601

[81] Ravi N, Krishna C M, Koren I. Enhancing vehicular anonymity in ITS: A new scheme for mix zones and their placement. IEEE Transactions on Vehicular Technology, 2019, 68(11): 10372-10381

[82] Wang J, Cai Z, Yu J. Achieving personalized k-anonymity-based content privacy for autonomous vehicles in CPS. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4242-4251

[83] Mehta B B, Rao U P. Improved l-diversity: Scalable anonymization approach for privacy preserving big data publishing. Journal of King Saud University-Computer and Information Sciences, 2022, 34(4): 1423-1430

[84] Ren W, Ghazinour K, Lian X. kt-Safety: Graph release via k-anonymity and t-closeness (Technical Report). arXiv preprint arXiv:2210.17479, 2022

[85] Sweeney L. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems, 2002, 10(05): 557-570

[86] Ponomareva N, Hazimeh H, Kurakin A, et al. How to dp-fy ml: A practical guide to machine learning with differential privacy. arXiv preprint arXiv:2303.00654, 2023

[87] Li J, Lin S, Yu K, et al. Quantum k-nearest neighbor classification algorithm based on Hamming distance. Quantum Information Processing, 2022, 21(1): 18

[88] Feldman V, McMillan A, Talwar K. Stronger privacy amplification by shuffling for rényi and approximate differential privacy//Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Florence, Italy, 2023: 4966-4981

[89] Ge Y F, Orlowska M, Cao J, et al. MDDE: Multitasking distributed differential evolution for privacy-preserving database fragmentation. The VLDB Journal, 2022, 31(5): 957-975

[90] Gong X, Chen Y, Wang Q, et al. Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions. IEEE Wireless Communications, 2023, 30 (2): 114-121

[91] Ma X, Jiang Q, Shojafar M, et al. DisBezant: secure and robust federated learning against byzantine attack in IoT-enabled MTS. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(2): 2492-2502

[92] Yang W, Wang N, Guan Z, et al. A practical cross-device federated learning framework over 5G networks. IEEE Wireless Communications, 2022, 29(6): 128-134

[93] Goldstein B F, Patil V C, Ferreira V C, et al. Preventing DNN model IP theft via hardware obfuscation. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2021, 11(2): 267-277

[94] Xue Y, Wang Y, Liang J, et al. A self-adaptive mutation neural architecture search algorithm based on blocks. IEEE Computational Intelligence Magazine, 2021, 16(3): 67-78

[95] Yuan W, Yin H, Wu F, et al. Federated unlearning for on-device recommendation//Proceedings of the 16th ACM International Conference on Web Search and Data Mining. New York, USA: 393-401

[96] Imran M, Yin H, Chen T, et al. ReFRS: Resource-efficient federated recommender system for dynamic and diversified user preferences. ACM Transactions on Information Systems, 2023, 41(3): 1-30

[97] Ammad-Ud-Din M, Ivannikova E, Khan S A, et al. Federated collaborative filtering for privacy-preserving personalized recommendation system. arXiv preprint arXiv: 1901.09888, 2019

[98] Muhammad K, Wang Q, O'Reilly-Morgan D, et al. Fedfast: Going beyond average for faster training of federated recommender systems//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA, 2020: 1234-1242

[99] Wu C, Wu F, Cao Y, et al. Fedgnn: Federated graph neural network for privacy-preserving recommendation. arXiv preprint arXiv:2102.04925, 2021

[100] Chen H, Li G, Jiang W, et al. Dynamic dual quaternion knowledge graph embedding. Applied intelligence, 2022, 52 (12): 14153-14163

[101] Chen M, Zhang W, Yuan Z, et al. Fede: Embedding knowledge graphs in federated setting//Proceedings of the 10th International Joint Conference on Knowledge Graphs. New York, USA, 2021: 80-88

[102] Peng H, Li H, Song Y, et al. Differentially private federated knowledge graphs embedding//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York, USA, 2021: 1416-1425

[103] Li C, Li G, Varshney P K. Federated learning with soft clustering. IEEE Internet of Things Journal, 2021, 9(10): 7773-7782

[104] Pan C, Sima J, Prakash S, et al. Machine unlearning of federated clusters. arXiv preprint arXiv:2210.16424, 2022

[105] Al-Sahaf H, Bi Y, Chen Q, et al. A survey on evolutionary machine learning. Journal of the Royal Society of New Zealand, 2019, 49(2): 205-228

[106] Kaur H, Nori H, Jenkins S, et al. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning//Proceedings of the 2020 CHI Conference on

Human Factors in Computing Systems. New York, USA, 2020: 1-14

[107] Zhang S, Gu Q, Wu X, et al. Non-uniform decomposition method used for obtaining the frequency-constrained matrix of broadband laguerre beamforming. IEEE Wireless Communications Letters, 2022, 11(7): 1359-1363

[108] Lu C, Feng J, Chen Y, et al. Tensor robust principal component analysis with a new tensor nuclear norm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(4): 925-938

[109] Jiang Y, Cong R, Shu C, et al. Federated learning based mobile crowd sensing with unreliable user data//Proceedings of the 18th International Conference on Smart City. Yanuca Island, Fiji, 2020: 320-327

[110] Liu Z, Chen Y, Yu H, et al. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(4): 1-21

[111] Liu Z, Chen Y, Yu H, et al. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(4): 1-21

[112] Wu K, Shi Y, Han Y, et al. Domain Adaptation without Model Transferring. arXiv preprint arXiv:2107.10174, 2021

[113] Zhang H, Zhang Y, Jia K, et al. Unsupervised domain adaptation of black-box source models. arXiv preprint arXiv: 2101.02839, 2021

[114] Jannach D, Manzoor A, Cai W, et al. A survey on conversational recommender systems. ACM Computing Surveys (CSUR), 2021, 54(5): 1-36

[115] Nguyen D C, Pham Q V, Pathirana P N, et al. Federated learning for smart healthcare: A survey. ACM Computing Surveys (CSUR), 2022, 55(3): 1-37

[116] Imteaj A, Thakker U, Wang S, et al. A survey on federated learning for resource-constrained IoT devices. IEEE Internet of Things Journal, 2021, 9(1): 1-24

[117] Saha R, Misra S, Deb P K. FogFL: Fog-assisted federated learning for resource-constrained IoT devices. IEEE Internet of Things Journal, 2020, 8(10): 8456-8463

[118] Xu M, Ng W C, Lim W Y B, et al. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. IEEE Communications Surveys & Tutorials, 2023, 25(1): 656-700

[119] Qiu Y, Zhou G, Zhang Y, et al. Canonical polyadic decomposition (CPD) of big tensors with low multilinear rank. Multimedia Tools and Applications, 2021, 80(15): 22987-23007

WANG Peng-Fei, Ph. D., associate professor. His main research interests include distributed artificial intelligence and big data intelligent processing.

WEI Zong-Zheng, M. S. candidate. His main research interest is federated learning.

ZHOU Dong-Sheng, Ph. D., professor. His main research interests include intelligent computing, computer graphics and vision.

SONG Wei, M.S. candidate. Her main research interest is federated learning.

XIAO Yun-Ming, Ph. D., candidate. His main research interests include edge computing and network measurement.

SUN Geng, Ph. D., associate professor. His main research interests include swarm intelligence and cooperative communication.

YU Shuo, Ph. D., associate professor. Her main research interests include data science, graph learning and data security.

ZHANG Qiang, Ph. D., professor. His main research interests include the new generation of artificial intelligence and biological intelligence computing.

Background

Federated unlearning plays an important role in the field of data computing and information security, which aims to remove the contribution of specific user data in the global model that has been trained by federated learning. Federated unlearning is specifically designed to align with the "right to be forgotten", a principle that allows users participating in federated learning to request the removal of their data's impact on the trained global machine learning model.

However, removing the contribution of data from a specific user in federated learning is extremely challenging due to the distributed nature of not grasping the user's local data. Firstly, federated learning involves distributed training while preserving user privacy, and simply revoking a user's training update in the global model is far from sufficient. This is because the model parameters held by other users still retain the updates of the unlearning user, and these will continue to be aggregated into the

global model during subsequent training. Secondly, the training of the federated learning model is a cumulative process. The training updates of all users up to a certain point determine the current local model parameter updates for the unlearning user. As a result, other users' contributions to the model training also include traces of the unlearning user's data due to this growing training process.

Federated unlearning was first proposed in 2020, and some research works have been done recently. However, it still does not have a literature review and survey to conclude the research works of federated unlearning and give future research directions.

This work is partly supported by the National Key R&D Program of China (2021ZD0112400), the Joint Funds of the National Natural Science Foundation of China (U1908214), the Young Scientists Fund of the National Natural Science Foundation of China (62202080), the China Postdoctoral Science Foundation (2023M733354), and the Fundamental Research Funds for the Central Universities of Ministry of Education of China (DUT23YG122). In this paper, we aim to do a comprehensive survey for federated unlearning. To achieve the goal, this paper analyzes the changes in the machine learning training architecture in recent years and gives the concept and definition of federated unlearning. Then, we elaborate on the federated unlearning algorithms by dividing it into three categories. In addition, we also introduce the performance metrics which are also divided into three categories, and the performance of these metrics is given exhaustively. The future research directions of federated unlearning are summarized at the end of this paper.