傅圣泽. 面向工业实体抽取的联邦学习优化算法[J]. 智能计算机与应用,2024,14(7):246-251. DOI:10.20169/j.issn.2095-2163.240740

面向工业实体抽取的联邦学习优化算法

傅圣泽
(上海工程技术大学 电子电气工程学院, 上海 201620)

摘 要: 人工智能技术在工业、医疗和金融等领域得到了广泛应用,并取得了巨大的成功。工业实体抽取任务是实现工业领域数字化转型的关键一环,然而其实现往往需要大量的数据支持,而这些数据往往分布在各个机构或组织之间。各行各业都产生了海量的有价值的数据,但是在实际的应用场景中,安全隐私、法律法规和行业竞争等多种因素往往导致各方的数据不能共享,从而形成所谓的"数据孤岛"。针对这一问题,联邦学习提供了一种解决方案,可以有效解决数据孤岛问题,但联邦学习目前仍然面临一些问题和挑战,其中最典型的问题就是数据异构问题。针对各行各业存在的数据孤岛问题以及联邦学习本身的数据异构问题,本文以工业领域实体抽取任务为对象研究联邦学习的异构问题,从本地优化的角度提出了一种基于本地修正的联邦学习算法 FedAmend,改善该联邦学习框架在面对工业领域数据非独立同分布时的表现,并在某汽车集团的工业设备故障数据上验证了 FedAmend 的可行性。

关键词: 实体抽取; 联邦学习; 数据异构; 本地优化
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)07-0246-06

Federated learning optimization algorithm for industrial entity extraction

FU Shengze
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Artificial Intelligence (AI) technologies have been widely used with great success in the industrial, healthcare and financial sectors. The industrial entity extraction task is a key part of achieving digital transformation in the industrial sector, yet its realization often requires the support of a large amount of data, which is often distributed among various institutions or organizations. All industries generate huge amounts of valuable data, but in actual application scenarios, security, privacy, laws and regulations, and industry competition often lead to the formation of so-called "data silos" where data cannot be shared. To address this problem, Federated Learning provides a solution that can effectively solve the problem of data silos, but Federated Learning is still facing a number of problems and challenges, the most typical of which is the problem of data heterogeneity. Aiming at the data silo problem existing in various industries and the data heterogeneity problem of federated learning itself, this paper studies the heterogeneity problem of federated learning with the entity extraction task in the industrial domain as the object, proposes a federated learning algorithm FedAmend based on local correction from the perspective of local optimization, improves the performance of this federated learning framework in the face of the non-independent and homogeneous distribution of the data in the industrial domain and verifies the performance of this federated learning algorithm on industrial equipment failure data of an automobile group. The feasibility of FedAmend is verified on the industrial equipment failure data of an automobile group.

Key words: entity extraction; federated learning; data heterogeneity; local optimization

0 引 言

工业生产过程中涉及到的数据规模日渐庞大。以往工业数据的处理主要依靠技术工人和领域专家,根据经验和知识进行数据诊断,效率较低。此外,工业领域涉及到工业设备的模式感知、故障监控等大量数据处理相关的工作,从这些分散在各处的工业数据中抽取到有用信息,为工业领域提供信息支撑,帮助工业领域相关人员快速有效地解决问题,是实现工业领域数字化转型的迫切需求。工业设备数据的复杂性与实时性是传统的数据库技术无法胜任的。知识图谱是一种展示信息中知识架构、知识要点的可视化技术,实现了从"数据"到"知识"的转变,其基本单位是"实体(Entity)-关系

作者简介: 傅圣泽(1998-),男,硕士研究生,主要研究方向:联邦学习。Email:1012439688@qq.com

收稿日期: 2023-04-21

(Relationship)-实体(Entity)”构成的三元组,这种展示数据结构关系的技术在工业领域逐渐受到青睐,可以有效解决上述问题[1-2]。实体抽取是构建知识图谱的第一步也是关键一步。实体是知识图谱中的最基本元素,其提取的完整性、准确率等都会对知识库的质量产生直接的影响。在现实落地的场景中,构建工业实体抽取模型需要来自各方的大量数据,而传统的集中式训练模式对各方的数据安全隐私造成了严重威胁。

虽然各种人工智能相关的技术得到了快速发展并获得了巨大成功,但是将这种以数据为驱动的人工智能技术落地并应用到各行各业十分困难。现有的人工智能技术需要大量优质数据作为支撑,但是在很多落地的场景中,数据量不足且数据质量差是非常普遍的现象。传统的方法是将各方的优质数据进行集中式训练,如果要收集并整合各行各业的数据进行训练,除了要面对商业竞争问题还要面对用户数据的隐私安全问题,因此彼此分散的数据难以整合利用,从而形成了“数据孤岛”[3]。

为了解决这些问题,克服传统方法的不足,联邦学习的框架应运而生,该技术可在保护各方数据隐私的前提下完成联合建模[4]。联邦学习是人工智能领域的一项基础技术,本质上是一种机器学习框架,通常由中心服务器和客户端组成,在保障各方数据、保护隐私安全、保证合法的前提下,协助多个参与方或多个计算结点开展机器学习,从而达到安全建模的目的。联邦学习秉持“只传递模型参数或者梯度”的思想,各方的数据只需要保留在本地,从而避免了数据泄露。作为一种基于客户端的机器学习方法,联邦学习具有很多优点,能够在保护数据隐私的前提下充分利用各方的私有数据。

尽管联邦学习可以有效解决数据孤岛问题,但联邦学习目前仍然面临一些问题和挑战[5-6]。在联邦学习框架中,最典型的问题就是数据异构问题,也叫数据非独立同分布(称为 Non-IID)问题,具体表现为联邦学习的各个参与方具有独立性,各客户端的数据分布情况通常是不相同的,各参与方数据量也是不均衡的,统称为数据异构。数据异构会使联邦学习的全局模型偏离全局最优解,导致联邦学习的效果变差。

经典的联邦平均算法在数据异构场景下往往不能发挥出良好的性能,本文主要研究面向工业领域实体抽取的联邦学习异构问题解决方法,聚焦于数据异构场景下的联邦学习优化算法。

1　相关内容

1.1　实体抽取

实体抽取又称命名实体识别,其主要任务是识别文本中的命名实体并分类到预先定义的实体类别中,例如人名、地名和组织机构名等[7]。常用的标注方式为 BIO 标注:B-begin,I-inside,O-outside;B 表示实体的开始部分,I 表示实体的内容,O 表示非实体部分。如图 1 所示,实体抽取模型识别到“拆卸”和“液压管”两个实体,按照 BIO 形式进行标注,并将其分别分类到操作(Operation,OPE)和设备(Device,DEV)这两个预先定义的标签类别,其中的关键问题是如何建立优质的实体抽取模型,以便从海量的数据源中抽取到自己想要的实体信息。然而构建工业领域实体抽取模型需要来自工业领域各方的大量数据,各方的数据往往由于行业本身的商业竞争问题和安全隐私问题不能共享从而形成了数据孤岛。

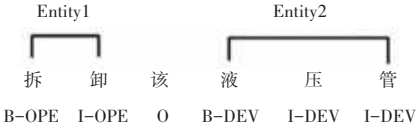

图 1　实体抽取示例

Fig. 1　Example of entity extraction

1.2　联邦学习

随着人工智能相关技术的飞速发展和数字化社会的到来,各行各业产生了大量有价值的数据。收集分散数据进行统一的机器学习训练时,除了需要面对各个行业本身的商业竞争问题,还要解决用户数据的安全隐私问题。受限于商业竞争、个人隐私、法律法规等方面的约束,各方的数据无法直接交换或者集中,从而形成“数据孤岛”现象,制约着人工智能技术在各行各业的落地。联邦学习很好地解决这一问题,该技术可在保护各方数据隐私的前提下完成联合建模[8-9]。在联邦学习框架中,各方的私有数据在训练过程的始终都保存在本地,避免了泄露数据的风险;通过调动多方联合训练,可以将分布在各方的高质量数据合理地利用起来,因此使用联邦学习框架的训练效果会优于各方单独训练的效果。Mc Mahan[10]提出了经典的联邦学习训练过程以及对应的联邦学习优化算法即联邦平均算法(FedAvg),该算法是联邦学习中最经典的算法。对联邦平均算法进行理论研究以及算法改进是联邦学习的一个主流方向。

联邦学习作为一种基于用户端的分布式机器学

习方法,能够直接基于用户的本地数据训练出效果较好的机器学习模型,并且可以充分利用来自各方的优质数据。联邦学习框架主要包括参与方和中心服务器,联邦学习的过程如图 2 所示。

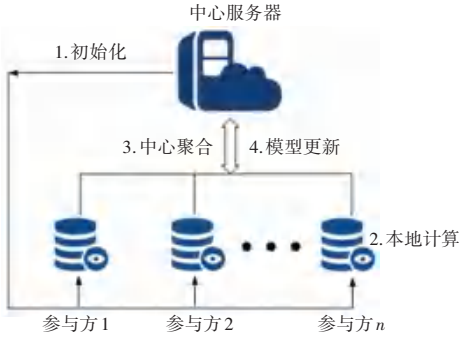

图 2 联邦学习框架

Fig. 2 Federated learning framework

(1)初始化:所有的参与方从中心服务器得到一个初始化模型,加入联邦学习并确定相同的任务及模型训练目标;

(2)本地计算:在每一轮联邦学习的通信过程中,参与方首先从中央服务器得到全局模型参数,然后又使用其私有数据对模型进行训练,更新模型,并将这些更新发送至中心服务器;

(3)中心聚合:聚合不同参与方本地训练得到的模型并更新得到下一轮的全局模型;

(4)模型更新:中心服务器根据聚合后的结果对全局模型进行更新,并将更新后的模型返回给参与联邦学习的参与方。

在联邦学习场景中,由于设备归属于某个企业、工厂、部门,因此其数据分布差异极大,这种现象被称为数据异构,也称为数据非独立同分布(Non-IID)。Zhao Y [11] 与 Sattler F[12] 通过各种实验表明,非独立同分布的数据会严重影响联邦学习的性能。

2 基于本地修正的联邦学习优化算法

联邦学习中各参与方数据往往是非独立同分布的,导致联邦学习框架的性能变弱。针对联邦学习数据异构问题导致客户端偏移现象,本文从本地优化的角度出发,提出了一种基于本地修正的联邦学习优化算法 FedAmend。

2.1 客户端偏移

联邦学习框架最核心的操作就是利用各客户端本地计算的结果来更新全局模型,因此各客户端本地计算的结果将直接影响到了全局模型的效果。在联邦学习框架中,各参与方在进行本地训练时,局部模型会根据自身的数据特性朝着局部模型的最优解方向靠拢。由于数据异构场景下各参与方数据分布差异较大,使得局部模型的更新朝着不同的方向,导致聚合之后得到的全局模型偏离全局最优解,从而影响到联邦学习的性能。如图 3 所示,初始的全局模型 Global 发送给两个客户端从而得到两个本地模型 X_1 与 X_2;在 Non-IID 情况下,模型 X_1 朝着局部最优解 x_1 方向移动,模型 X_2 朝着局部最优解 x_2 方向移动,经过 3 轮更新取两个客户端模型的均值得到全局模型 Y,此时发现联邦学习训练得到的全局模型 Y 与全局最优模型 y 存在偏差。说明客户端数据分布不一致,本地模型在更新时会朝着不同方向优化,使得联邦学习很难得到一个令人满意的全局模型。

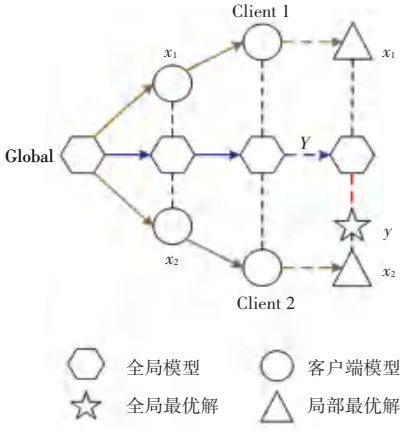

图 3 客户端偏移

Fig. 3 Client drift

2.2 联邦平均算法的局限性

联邦平均算法(FedAvg)与传统的联邦随机梯度下降法(FedSGD)相比,最大的区别在于 FedAvg 算法在客户端本地更新时进行多次迭代更新,而 FedSGD 算法只是在本地进行一次更新。虽然联邦平均算法减小了通信成本,提高了联邦学习的效率,但同时也带来了新的问题。

FedSGD 由于只在本地训练一次,不易受到本地数据异构问题的影响,更新路径与全局模型的理想更新路径相近,但通信成本过高;FedAvg 通过增加本地训练轮次来减少所需的通信轮次,极大的缓解联邦学习的通信压力,但也导致了新的问题。由于各参与方的数据是非独立同分布的,过多的本地训练轮次导致本地更新路径朝着不同方向移动,联邦学习的效果变差即客户端偏移现象。FedAvg 在面对数据异构场景时往往不能发挥出好的效果。

2.3 FedAmend 算法

联邦平均算法提出在联邦学习的客户端本地执

行多轮局部更新，增大本地计算从而减小联邦学习的通信压力，大大提高了联邦学习框架的效率。现实的各种场景中，不同的客户端在数据分布上存在的差异较大，这种单纯增加本地计算的方法会导致各客户端朝其局部最优解方向移动，使联邦学习的全局模型性能变差。本文从本地优化的角度出发，提出一种基于本地修正的联邦学习优化算法 FedAmend。

机器学习中过度训练也可能会导致过拟合现象，而联邦学习中本地过多的训练轮次可能会导致客户端偏移现象。机器学习中通过添加 L2 正则项来防止过拟合；同样，联邦学习中可以在本地训练阶段添加修正项来遏制客户端偏移现象。

正则化约束一定程度上等价于在训练时加入了先验知识，这种思想同样也可以类比运用到联邦学习中。在联邦学习中，希望各参与方的本地更新方向趋于一致，全局模型可以对各参与方的更新方向起到引导作用，让各参与方在本地多轮迭代之后也不会过分偏离全局模型，从而一定程度上聚拢了各参与方的更新方向。

本文所提出的 FedAmend 算法在 FedAvg 的基础上进行优化，在客户端本地训练的损失函数中加入修正项，遏制因客户端本地多轮迭代而导致的客户端偏移现象，从而缓解联邦学习的数据异构问题。FedAmend 算法对于联邦学习客户端本地训练的优化如公式(1)所示：

$$F(\theta) = L(\theta) + \lambda \parallel W_{\text{global}} \parallel_2 \qquad (1)$$

其中，$L(\theta)$ 表示客户端本地训练时的损失函数；W_{global} 表示这一轮联邦学习的全局模型参数；λ 为修正项的系数。

W_{global} 一定程度上使各参与方的聚合更新方向聚拢，让各参与方本地训练出来的模型不会过分的偏离当前的全局模型，遏制了客户端偏移的现象，从而一定程度上缓解了联邦学习数据异构问题带来的负面影响。

2.4　模型介绍

本文使用 BERT+BiLSTM+CRF 作为实体抽取模型，模型结构如图 4 所示。

模型的处理流程：首先，输入该汽车集团的工业设备故障工单语料数据作为训练数据，使用 BERT 预训练模型获取字向量，提取文本重要特征；其次，通过 BiILSTM 深度学习上下文特征信息，进行命名实体识别；最后，CRF 层对 BiLSTM 的输出序列处理，得到一个预测标注序列，对序列中的各个实体进行提取分类。

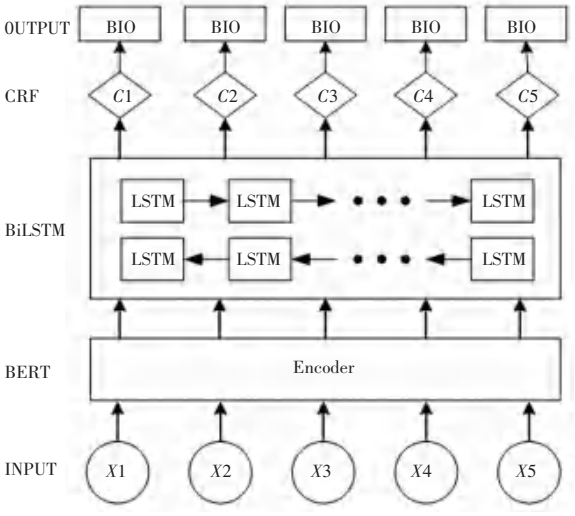

图 4　BERT+BiLSTM+CRF 模型结构

Fig. 4　Structure of BERT+BiLSTM+CRF

3　实验

3.1　实验数据介绍

实验数据使用某汽车集团的工业设备故障数据，包含各种语料信息以及实体类别。首先对原数据进行处理，将原数据转换为 BIO 格式的数据，数据示例如图 5 所示。原数据经过处理之后共有 ATTRIBUTE，NORMAL，UNNORMAL，FAULT，DEVICE，OPERATION 6 种实体标签以及非实体标签 0。实验数据分为训练集和测试集，其中训练集共有 54 526 行，测试集有 20 880 行。

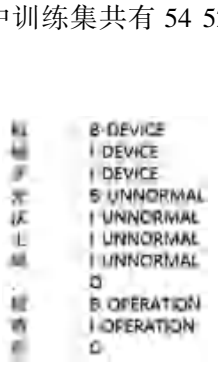

图 5　数据示例

Fig. 5　Sample data

3.2　联邦学习的数据异构场景

本实验使用某汽车集团工业设备故障工单数据，从标签分布偏移和数量分布偏移两个角度涵盖 Non-IID 的场景，示意图如图 6 所示。图中横坐标表示不同的客户端，横向的宽度表示数据量，越宽表示数据量越大；纵坐标表示各客户端拥有的标签类别数。图 6 中共有 Client1，Client2，Client3 3 个参与方，Client1 无论是标签数还是数据量都少于其他两

方,Clent2 和 Clent3 拥有相同的标签数但是两者的数据量不同。

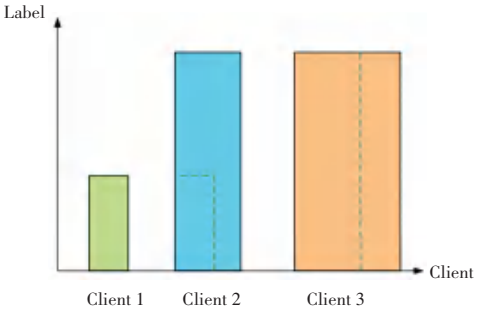

图 6 Non-IID 数据

Fig. 6 Non-IID data

本实验的数据分为两种情况,表示数据的异构程度不同,分别命名为 Low Non-IID 和 High Non-IID,其中 Low Non-IID 只存在数量偏移的情况,High Non-IID 既存在数量偏移也存在标签偏移。

对于 Low Non-IID 场景,有 Low1,Low2,Low3 3 个参与方,其中 Low1 共有1 000行数据,Low2 共有 875 行数据,Low3 共有 525 行数据;对于 High Non-IID 场景,有 High1,High2,High3 3 个参与方,其中 High1 共有 100 行数据,只有 O,DEVICE,UNNORMAL 3 种标签的数据;High2 共有 125 行数据,只有 O,DEVICE,NORMAL,OPERATION 4 种标签的数据;High3 共有 150 行数据,只有 O,DEVICE,FAULT,OPERATION,ATTRIBUTE 5 种标签的数据。

3.3 实验对比

在 Low Non-IID 和 High Non-IID 两种数据场景下进行实验,修正项系数 λ 都设置为 0.001。用 $D(1)$、$D(2)$、$D(3)$ 分别表示各参与方的聚合权重,FedAmend 各参与方的权重见表 1。

表 1 FedAmend 的权重

Table 1 Weight of FedAmend

场景	$D(1)$	$D(2)$	$D(3)$
FedAmend-Low Non-IID	$\frac{1\,000}{2\,400}$	$\frac{875}{2\,400}$	$\frac{525}{2\,400}$
FedAmend-High Non-IID	$\frac{100}{375}$	$\frac{125}{375}$	$\frac{150}{375}$

Low1、Low2、Low3 表示各参与方用上述数据单独训练的情况;FedAvg 表示对这 3 个参与方使用联邦平均算法的情况,FedAmend 表示使用本文提出方法的情况。对于 Low Non-IID 和 High Non-IID 这两种数据场景,通过实验得到 Low Non-IID 场景下的各方法的准确率见表 2 和表 3。

表 2 Low Non-IID 场景的实验结果

Table 2 Experimental results of the Low Non IID scenario

实验对比	准确率 / %
Low1	64.9
Low2	69.6
Low3	66.7
FedAvg	72.7
FedAmend	73.4

同理,High1、High 2、High 3 表示各参与方用上述数据单独训练的情况;FedAvg 表示对这 3 个参与方使用联邦平均算法的情况,FedAmend 表示使用本文提出方法的情况。本文的 FedAmend 算法在 Low Non-IID 场景下比传统的联邦平均算法高 0.7% 的准确率。

表 3 High Non-IID 场景的实验结果

Table 3 Experimental results of the High Non IID scenario

实验对比	准确率 / %
High1	37.9
High2	53.4
High3	67.1
FedAvg	69.4
FedAmend	70.3

从表 3 可以看出,本文的 FedAmend 算法在 High Non-IID 场景下比传统的联邦平均算法高 0.9%的准确率。

综上,使用了 FedAmend 优化算法之后,在 Low Non-IID 和 High Non-IID 两种场景的实验中,FedAmend 比传统的联邦学习优化算法有更好的表现,表明数据异构下的客户端偏移现象是联邦学习框架效果变差的原因之一。

4 结束语

联邦学习具有广泛的应用价值,数据异构问题作为联邦学习中的一大挑战一直都被广泛关注与研究。本文针对联邦学习的数据异构问题,提出 FedAmend 算法,相比传统的联邦平均算法有更好的表现,缓解了联邦学习数据异构问题带来的负面影响。

参考文献

[1] CAI H, ZHENG V W, CHANG K C. A comprehensive survey of graph embedding: Problems, techniques, and applications [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616-1637.

[2] LIU W, ZHOU P, ZHAO Z, et al. K-bert: Enabling language representation with knowledge graph [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 2901-2908.