



(12) 发明专利申请

(10) 申请公布号 CN 117910539 A

(43) 申请公布日 2024. 04. 19

(21) 申请号 202410310044.8

G06F 18/2413 (2023.01)

(22) 申请日 2024.03.19

G06F 18/2431 (2023.01)

G06F 18/213 (2023.01)

(71) 申请人 电子科技大学

地址 611731 四川省成都市高新区(西区)
西源大道2006号(72) 发明人 陈维龙 段雯静 袁伟民 张欣然
张彦如(74) 专利代理机构 电子科技大学专利中心
51203

专利代理师 曾磊

(51) Int. Cl.

G06N 3/0895 (2023.01)

G06N 3/098 (2023.01)

G06N 3/045 (2023.01)

G06N 3/082 (2023.01)

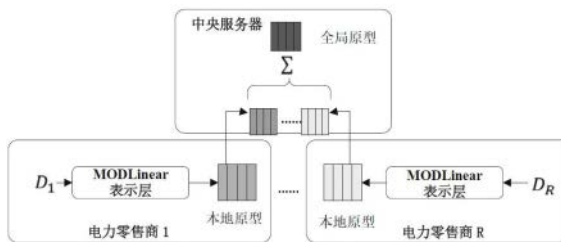
权利要求书3页 说明书8页 附图2页

(54) 发明名称

一种基于异构半监督联邦学习的家庭特征
识别方法

(57) 摘要

一种基于异构半监督联邦学习的家庭特征识别方法,属于联邦学习领域。本发明结合基于原型的联邦学习和半监督学习方法构建SF-Heter框架,用于基于异构电表数据的家庭特征识别。用基于原型的聚合取代了基于梯度的聚合,各电力零售商利用私有数据和全局原型优化本地的家庭特征识别模型;采用半监督学习方法,通过对未标记数据的标签预测实现数据增强,利用未标记数据信息。本发明基于原型联邦学习框架,在保护电力零售商数据隐私的同时,解决了各零售商之间数据与模型异构问题。本发明采用半监督学习方法,有效使用大量未标记数据信息,提高了识别电力消费者特征的性能。



1. 一种基于异构半监督联邦学习的家庭特征识别方法,其特征在于,包括如下步骤:

步骤S1:构建本地MODLinear模型,将智能电表数据输入MODLinear模型,输出预测的用户家庭特征,并计算初始本地原型;

步骤S2:聚合全局原型,将所有电力零售商的本地原型发送至中央服务器进行聚合,生成统一的、具有代表性的全局原型;

步骤S3:更新本地模型,构建损失函数,使电力零售商的本地原型和中央服务器的全局原型对齐,根据损失函数值更新本地家庭特征识别模型,迭代计算新的本地原型和全局原型。

2. 根据权利要求1所述的基于异构半监督联邦学习的家庭特征识别方法,其特征不在于,

MODLinear模型由两个组件组成:第一个组件为表示层,也称为特征提取层,通过神经网络将高维电表数据转换为低维向量;第二个组件是识别层,利用表示层得到的低维向量进一步预测用户家庭特征;

MODLinear模型被表示为 $\mathcal{E}(\phi_r)\mathcal{G}(v_r)$, $\mathcal{E}(\phi_r)$ 表示第 r 个零售商的表示层,其参数化形式为 ϕ_r , $\mathcal{G}(v_r)$ 表示第 r 个零售商的识别层,其参数化形式为 v_r ;定义第 r 个零售商拥有的智能电表数据集为 D_r , D_r 由数据集 D_r^u 和 D_r^l 组成,其中 D_r^u 是第 r 个零售商的未标记数据集, D_r^l 是第 r 个零售商的标记数据集;

在MODLinear模型中,使用移动平均核 k ,将智能电表数据 X 分解为趋势分量 X_k^{Tr} 和余数分量 X_k^{Re} ;其中, $X_k^{Re} = X - X_k^{Tr}$,电表数据的整体表示由分解出的趋势分量 X_k^{Tr} 和余数分量 X_k^{Re} 与 W_k^{Tr} 和 W_k^{Re} 分别相乘得到,具体表示为 $H_k = W_k^{Tr} X_k^{Tr} + W_k^{Re} X_k^{Re}$,其中 W_k^{Tr} 和 W_k^{Re} 为两个不同的线性层; $H_o = W^o X$ 表示通过直接将智能电表数据 X 输入另一个线性层 W^o 而不进行分解得到;定义 $H = \mathcal{E}(\phi_r)$ 为智能电表数据的最终表示,其由 H_o 和 H_k 连接起来表示为: $H = H_o \cup H_k, k \in \{1, 2, 3\}$ 。

3. 根据权利要求2所述的基于异构半监督联邦学习的家庭特征识别方法,其特征不在于,

本地原型采用如下方式计算,对于第 r 个零售商,定义 $Y_{r,m}$ 为第 r 个零售商的第 m 个特征,定义 $C_{r,m,n}$ 表示 $Y_{r,m}$ 的第 n 类的本地原型, $C_{r,m,n}$ 的值为表示层产生的属于第 n 类的智能电表数据的最终表示 H 的均值,则 $C_{r,m,n}$ 计算公式为:

$$C_{r,m,n} = \frac{1}{|D_{r,m,n}|} \sum_{X \in D_{r,m,n}} \mathcal{E}(\phi_r, X);$$

其中 $D_{r,m,n}$ 是智能电表数据集 D_r 的子集,由属于第 m 个特征的第 n 类的训练实例组成; $\mathcal{E}(\phi_r, X)$ 为 X 的向量表示。

4. 根据权利要求3所述的基于异构半监督联邦学习的家庭特征识别方法,其特征不在于,

全局原型的计算公式为: $\bar{C}_{m,n} = \frac{1}{|R|} \sum_{r \in R} C_{r,m,n}$;

其中 $\bar{C}_{m,n}$ 表示 $Y_{r,m}$ 的第 n 类的全局原型, 由所有零售商本地原型的平均值聚合而成; $|R|$ 表示参与聚合的零售商的总数。

5. 根据权利要求4所述的基于异构半监督联邦学习的家庭特征识别方法, 其特征在于, 构建损失函数表示为:

$$\mathcal{L} = \mathcal{L}_I + \lambda_1 \mathcal{L}_S + \lambda_2 \mathcal{L}_{semi-I};$$

其中 $\mathcal{L}_I$ 表示分类损失, $\mathcal{L}_S$ 表示对齐损失, $\mathcal{L}_{semi-I}$ 表示半监督学习损失, λ_1 和 λ_2 系数是用来平衡 $\mathcal{L}_I$ 、 $\mathcal{L}_S$ 和 $\mathcal{L}_{semi-I}$ 这三个损失函数的超参数。

6. 根据权利要求5所述的基于异构半监督联邦学习的家庭特征识别方法, 其特征在于,

$$\text{分类损失 } \mathcal{L}_I \text{ 的计算公式为: } \mathcal{L}_I = -\frac{1}{|D_r^l|} \sum_{x_i \in D_r^l} \sum_{m=1}^M \sum_{n=1}^{N_m} y_{i,m,n} \log p_{i,m,n};$$

其中, x_i 是第 r 个零售商的标记数据集 D_r^l 中的标记数据, M 为家庭特征的个数, N_m 为第 m 个特征类别数, $y_{i,m,n}$ 为真实标签, $p_{i,m,n}$ 为 x_i 属于第 m 个特征的第 n 个类的概率。

7. 根据权利要求6所述的基于异构半监督联邦学习的家庭特征识别方法, 其特征在于,

$$\text{对齐损失 } \mathcal{L}_S \text{ 的计算公式为: } \mathcal{L}_S = \frac{1}{M} \sum_{m=1}^M \frac{1}{N_m} \sum_{n=1}^{N_m} l_{r,m,n};$$

$l_{r,m,n}$ 定义为:

$$l_{r,m,n} = -\log \frac{\exp(\text{sim}(C_{r,m,n}, \bar{C}_{m,n}) / \tau)}{\sum_{t=1}^{N_m} \exp(\text{sim}(C_{r,m,n}, \bar{C}_{m,t}) / \tau)};$$

其中, $\bar{C}_{m,t}$ 表示 $Y_{r,m}$ 的第 t 类的全局原型, $t = 1, \dots, N_m$; 其中

$\text{sim}(C, \bar{C}) = C\bar{C} / \|C\| \|\bar{C}\|$ 表示经过 ℓ_2 范数归一化的 C 和 $\bar{C}$ 之间的点积, 即余弦相似度, τ 表示温度参数。

8. 根据权利要求7所述的基于异构半监督联邦学习的家庭特征识别方法, 其特征在于, 半监督学习损失 $\mathcal{L}_{semi-I}$ 的计算公式为:

$$\mathcal{L}_{semi-I} = -\frac{1}{|D_r^u|} \sum_{x_j \in D_r^u} \sum_{m=1}^M \sum_{n=1}^{N_m} \hat{y}_{j,m,n} \log p_{j,m,n};$$

其中, x_j 是第 r 个零售商的未标记数据集 D_r^u 中的未标记数据, $\hat{y}_{j,m,n}$ 为 x_j 的伪标签,

$p_{j,m,n}$ 为 x_j 属于第 m 个特征的第 n 类的预测概率。

一种基于异构半监督联邦学习的家庭特征识别方法

技术领域

[0001] 本发明属于联邦学习领域,尤其涉及一种基于异构半监督联邦学习的家庭特征识别方法。

背景技术

[0002] 随着智能电表的日益普及,电力零售商收集客户信息的方式发生了改变,智能电表收集关于用电模式的详细信息,提供了账单记录之外的大量信息。为了加强能源管理,并为客户提供个性化和多样化的服务,零售商需要利用细粒度的智能电表数据来准确识别客户的消费特征。

[0003] 为了分析智能电表数据,传统机器学习技术和深度学习算法被广泛应用,深度学习模型的训练需要大量的数据,零售商之间共享数据可以缓解数据不足的问题,并扩大深度训练的数据量。然而,电力零售商并非同质的,在他们的客户群、服务提供和他们经营的能源市场方面可能存在显著差异,统计数据的异质性对在所有零售商中实施通用能源管理解决方案提出重大挑战。此外,规模较小或资源不足的零售商可能会难以有效地访问和分析智能电表数据,这可能导致数据的访问和利用不平等。除了统计数据的异质性外,电力零售商通过竞争获取利润,因此可能不愿意彼此分享原始隐私敏感的客户信息。

[0004] 为了在不直接共享用电数据和家庭信息的情况下实现不同电力零售商之间的协作学习,联邦学习(Federated Learning, FL)被认为是一种很有前途的解决方案,即共享本地训练的模型,而不是原始的训练数据。

[0005] 虽然联邦学习通过共享由所有数据训练的全局模型,缓解了各零售商训练数据不足的问题。但由于标注成本高和数据保护问题,电力零售商在使用自己的数据训练本地模型时仍然面临数据量不足的挑战。现有技术采用联邦学习框架保护数据隐私,但忽略了数据和模型的异质性,未考虑电力零售和家庭特征识别场景下的异质性问题。较高的标注成本和用户隐私担忧,既造成未标注数据的浪费,又由于可用标注数据的不足导致分析不准确。基于转导支持向量机(Transductive Support Vector Machine, TSVM)的半监督学习方法,在标注数据量有限的情况下进行样本分类和信息提取,这是迄今为止唯一一项将半监督学习框架应用于智能电表数据的工作。然而,传统的转导支持向量机不能共享模型参数,不能应用于联邦学习框架。

发明内容

[0006] 针对现有技术中联邦学习未考虑数据和模型的异质性、标注成本高及传统的转导支持向量机的半监督学习方法不能共享模型参数的技术问题,本发明的目的在于提供一种基于异构半监督联邦学习的家庭特征识别方法。

[0007] 为解决上述技术问题,本发明的基于异构半监督联邦学习的家庭特征识别方法的具体技术方案如下:

本发明提出的半监督联邦学习的框架SF-Heter (Semi-Supervised Federated

Learning for Smart Meter Heterogeneous Data),用于异构智能电表数据的家庭特征识别任务。每个零售商的本地模型:MODLinear (Multi-Channel Origin Dlinear,MODLinear)由表示层和识别层两部分组成,表示层为特征提取层,将高维电表数据转换为低维向量,提取智能电表中的时间序列信息。表示层后为识别层,利用低维向量表示输出家庭特征的分类预测。数据存储和模型训练过程基于联邦学习框架以分散的方式实现,保护了零售商的数据隐私。采用基于原型的联邦学习方法,中央服务器和零售商通过原型而不是梯度信息进行交流,零售商可以根据自己的数据格式和容量设计个性化参数的维度,以解决数据和模型的异质性问题。此外,采用半监督学习,通过预测未标记的数据的伪标签进行数据增强,并引入统一的损失项,以充分利用未标记的智能电表数据,提高识别家庭特征的准确性。因此,本方案能够比较好地解决上述问题。

[0008] 本发明的基于异构半监督联邦学习的家庭特征识别方法具有以下有益的技术效果:

1) 构建SF-Heter框架处理电力零售商数据,其采用基于原型的联邦学习框架,中央服务器和电力零售商之间能够通过原型进行交流,在保护电力零售商数据隐私的同时,解决数据与模型异质性问题。

[0009] 2) 构建深度MODlinear学习模型提取智能电表中的时间序列信息,输出家庭特征的预测。采用半监督学习方法,通过预测未标记数据的伪标签进行数据增强,引入损失函数,充分利用未标记数据补充有限可用的标记数据,提高模型家庭特征识别的准确性。

附图说明

[0010] 为了更清楚地说明本发明实施例的技术方案,下面将对本发明实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0011] 图1为本发明的MODlinear本地模型结构示意图。

[0012] 图2为本发明的全局模型聚合结构示意图。

[0013] 图3 为本发明的本地模型更新框架示意图。

具体实施方式

[0014] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0015] 本发明提出的半监督联邦学习的SF-Heter框架,旨在应对异构数据和模型的挑战。该框架结合了半监督学习技术和原型框架来提高家庭特征分类精度。整体方法为:

1) 初始轮:每个零售商只使用本地电表数据来训练本地模型,并计算初始本地原型。将所有本地原型集 $\mathbf{C}$ 转发至中央服务器,对所有本地原型集 $\mathbf{C}$ 进行平均,聚合成全局原型 $\bar{\mathbf{C}}$ 。

[0016] 2) 随后训练轮:使用全局原型 $\bar{C}$ 来更新零售商的本地模型参数,并计算新的本地原型。每个零售商将本地原型集转发至中央服务器,聚合新的全局原型,再分发给所有零售商。从而开始新一轮的本地模型更新和全局原型聚合。不断重复该过程,直到训练完成。利用私有数据和全局原型,每个本地零售商都优化了各自的模型。

[0017] 3) 更新本地模型时,采用半监督学习方法,通过预测未标记电表数据的伪标签实现数据增强。将分类损失、对齐损失和半监督学习损失集成到一个的损失函数中,根据损失函数值更新本地家庭特征识别模型 $F(r, m)$, $F(r, m)$ 代表第 r 个零售商的第 m 个特征的神经网络模型。本地模型能够更有效地从标记和未标记的数据中学习,也鼓励每个本地模型生成更符合全局共识的原型,提高了家庭特征识别的准确性。

[0018] 为了达到上述目的,本发明提出的基于异构半监督联邦学习的家庭特征识别方法,即半监督联邦学习的SF-Heter框架,包括如下步骤:

步骤S1:构建本地MODLinear模型。将智能电表数据输入MODLinear模型,输出预测的用户家庭特征,并计算初始本地原型。

[0019] 具体的,如图1所示,所示步骤S1包括以下步骤:

MODLinear模型由两个组件组成:第一个组件为表示层,也称为特征提取层,通过神经网络将高维电表数据转换为低维向量;第二个组件是识别层,利用表示层得到的低维向量进一步预测用户家庭特征。MODLinear模型被表示为 $\mathcal{E}(\phi_r)\mathcal{G}(v_r)$, $\mathcal{E}(\phi_r)$ 表示第 r 个零售商的表示层,其参数化形式为 ϕ_r , $\mathcal{G}(v_r)$ 表示第 r 个零售商的识别层,其参数化形式为 v_r 。定义第 r 个零售商拥有的智能电表数据集为 D_r , D_r 由数据集 D_r^u 和 D_r^l 组成,其中 D_r^u 是第 r 个零售商的未标记数据集, D_r^l 是第 r 个零售商的标记数据集。

[0020] 表示层中,为提高电表数据的可预测性,采用时间序列预测的标准方法:季节趋势分解。具体分解方法为在输入序列上使用一个移动平均核提取时间序列的趋势分量,将原始序列与趋势分量的差值作为余数分量。在MODLinear模型中,使用移动平均核 k ,将智能电表数据 X 分解为趋势分量 X_k^{Tr} 和余数分量 X_k^{Re} ;其中, $X_k^{Re} = X - X_k^{Tr}$,电表数据的整体表示由分解出的趋势分量 X_k^{Tr} 和余数分量 X_k^{Re} 与 W_k^{Tr} 和 W_k^{Re} 分别相乘得到,具体表示为 $H_k = W_k^{Tr} X_k^{Tr} + W_k^{Re} X_k^{Re}$,其中 W_k^{Tr} 和 W_k^{Re} 为两个不同的线性层。为了提取不同时间维度信息,采用不同内核大小对电表数据进行分解,从而得到不同的表示。此外, $H_o = W^o X$ 表示通过直接将智能电表数据 X 输入另一个线性层 W^o 而不进行分解得到。定义 $H = \mathcal{E}(\phi_r)$ 为智能电表数据的最终表示,其由 H_o 和 H_k 连接起来表示为:

$$H = H_o \cup H_k, k \in \{1, 2, 3\}.$$

[0021] 识别层中,将智能电表数据的最终表示 H 输入到由三个全连接层模块(Fully Connected, FC)和一个输出层组成的网络中,全连接层模块使网络能够捕获数据中的复杂关系和模式,从而实现有效的识别和学习;输出层直接识别用户特征。

[0022] 原型(Prototypes)能有效地表示更广义的用户家庭特征信息;同时,在与服务器

的传输过程中,高度聚合的原型避免了用户信息在服务器上的分离,保证用户隐私的保护。本地原型通过如下方式计算,对于第 r 个零售商,定义 $Y_{r,m}$ 为第 r 个零售商的第 m 个特征,定义 $C_{r,m,n}$ 表示 $Y_{r,m}$ 的第 n 类的本地原型, $C_{r,m,n}$ 的值为表示层产生的属于第 n 类的智能电表数据的最终表示 H 的均值,则 $C_{r,m,n}$ 可计算为:

$$C_{r,m,n} = \frac{1}{|D_{r,m,n}|} \sum_{X \in D_{r,m,n}} \mathcal{E}(\phi_r, X)$$

[0023] 其中 $D_{r,m,n}$ 是智能电表数据集 D_r 的子集,由属于第 m 个特征的第 n 类的训练实例组成。 $\mathcal{E}(\phi_r, X)$ 为 X 的向量表示。在服务器聚合时,有必要使用一致的数据格式进行聚合。在传统的联邦学习中,不同的零售商需要相同的聚合模型。这个限制阻止了在服务器上共享不同的模型和不同格式的数据。但通过使用原型,只需要确保表示层生成的原型的格式是一致的。这使得模型设计具有更大的灵活性,更容易实现并适应不同场景中的不同模型和数据格式。

[0024] 步骤S2:聚合全局原型,将所有电力零售商的本地原型发送至中央服务器进行聚合,生成统一的、具有代表性的全局原型。

[0025] 具体的,如图2所示,所示步骤S2包括以下步骤:

为在原型上达成全局共识,中央服务器整理来自所有本地零售商的原型。全局原型的计算公式为:

$$\bar{C}_{m,n} = \frac{1}{|R|} \sum_{r \in R} C_{r,m,n}$$

[0026] 其中 $\bar{C}_{m,n}$ 表示 $Y_{r,m}$ 的第 n 类的全局原型,由所有零售商本地原型的平均值聚合而成。 $|R|$ 表示参与聚合的零售商的总数。这种协作方式确保每个零售商为统一的、具有代表性的全局原型做出贡献。

[0027] 步骤S3:更新本地模型,构建损失函数,使电力零售商的本地原型和中央服务器的全局原型对齐,根据损失函数值更新本地家庭特征识别模型,迭代计算新的本地原型和全局原型。

[0028] 具体的,如图3所示,所示步骤S3包括以下步骤:

每个零售商的本地模型的目标为:将其来自表示层的本地原型与通过聚合中央服务器上的所有本地原型生成的全局原型进行对齐。这通过应用一种新的损失函数来实现,该损失函数最小化本地原型和全局原型之间的距离,同时最大化分类准确性。该损失函数表示为:

$$\mathcal{L} = \mathcal{L}_I + \lambda_1 \mathcal{L}_S + \lambda_2 \mathcal{L}_{semi-I}$$

[0029] 其中 $\mathcal{L}_I$ 表示分类损失, $\mathcal{L}_S$ 表示对齐损失, $\mathcal{L}_{semi-I}$ 表示半监督学习损失。 λ_1 和 λ_2 系数是用来平衡 $\mathcal{L}_I$ 、 $\mathcal{L}_S$ 和 $\mathcal{L}_{semi-I}$ 这三个损失函数的超参数。

[0030] 构建分类损失 $\mathcal{L}_I$,目标是最小化标记数据的分类误差。使用交叉熵作为分类损失,

其计算公式为:

$$\mathcal{L}_I = -\frac{1}{|D_r^l|} \sum_{x_i \in D_r^l} \sum_{m=1}^M \sum_{n=1}^{N_m} y_{i,m,n} \log p_{i,m,n}$$

[0031] 其中, x_i 是第 r 个零售商的标记数据集 D_r^l 中的标记数据, M 为家庭特征的个数, N_m 为第 m 个特征的类别数, $y_{i,m,n}$ 为真实标签, $p_{i,m,n}$ 为 x_i 属于第 m 个特征的第 n 个类的概率。

[0032] 构建对齐损失 $\mathcal{L}_S$, 目标是使本地原型和全局原型之间的差异最小化。它鼓励每个零售商通过考虑全局共识原型来生成一个更准确、更有代表性的原型, 计算如下:

$$\mathcal{L}_S = \frac{1}{M} \sum_{m=1}^M \frac{1}{N_m} \sum_{n=1}^{N_m} l_{r,m,n}$$

[0033] $l_{r,m,n}$ 定义为:

$$l_{r,m,n} = -\log \frac{\exp(\text{sim}(C_{r,m,n}, \bar{C}_{m,n}) / \tau)}{\sum_{t=1}^{N_m} \exp(\text{sim}(C_{r,m,n}, \bar{C}_{m,t}) / \tau)}$$

[0034] 其中, $C_{r,m,n}$ 是 $Y_{r,m}$ 的第 n 类的本地原型, $\bar{C}_{m,t}$ 表示 $Y_{r,m}$ 的第 t 类的全局原型, $t = 1, \dots, N_m$, $\bar{C}_{m,n}$ 表示 $Y_{r,m}$ 的第 n 类的全局原型。其中 $\text{sim}(C, \bar{C}) = C\bar{C} / \|C\| \|\bar{C}\|$ 表示经过 ℓ_2 范数归一化的 C 和 $\bar{C}$ 之间的点积, 即余弦相似度, τ 表示温度参数。通过最小化本地原型和全局原型之间的距离, 对齐损失可以促进学习过程的收敛。对齐损失在原型上运行, 而不是原始数据, 提供了额外的隐私保护层。原型的使用也降低了计算和通信成本, 因为它们通常比原始数据或原始模型的维度要低得多。

[0035] 构建半监督学习损失 $\mathcal{L}_{\text{semi-I}}$, 利用未标记数据弥补标记数据的有限可用性, 并提高整体模型性能。通过预测未标记数据的伪标签, 增强了模型的鲁棒性。智能电表数据 X 包括标记数据 X^l 和未标记数据 X^u , 为了生成伪标签, 未标记数据 X^u 经历了数据增强, 从而得到增强后的数据 $W(X^u)$ 。模型预测类分布 $q_w = \mathcal{F}_i(W(X^u), w_i)$, $\mathcal{F}_i(W(X^u), w_i)$ 函数为预测类分布的神经网络模型, 其中 w_i 为模型的参数。 $\hat{q}_w = \text{argmax}(q_w)$ 表示 X^u 的伪标签。半监督学习损失 $\mathcal{L}_{\text{semi-I}}$ 表示为:

$$\mathcal{L}_{\text{semi-I}} = -\frac{1}{|D_r^u|} \sum_{x_j \in D_r^u} \sum_{m=1}^M \sum_{n=1}^{N_m} \hat{y}_{j,m,n} \log p_{j,m,n}$$

[0036] 其中, x_j 是第 r 个零售商的未标记数据集 D_r^u 中的未标记数据, $\hat{y}_{j,m,n}$ 为 x_j 的伪标签, $p_{j,m,n}$ 为 x_j 属于第 m 个特征的第 n 类的预测概率。这些伪标签 $\hat{y}_{j,m,n}$ 是模型在自训练过

程中生成的:最初对有标签数据进行训练,然后模型预测未标记数据的标签。

[0037] 通过优化过程使损失 $\mathcal{L}$ 最小化,从而更新每个零售商的本地模型,在本地模型更新之后,每个零售商计算其更新后的本地原型,并将其发送到中央服务器以进行全局原型更新。本地模型更新和全局原型更新的迭代过程产生了更好的整体模型,提供了一种稳健的机制来处理联邦学习框架中不同零售商之间的数据和模型异质性。

[0038] 本发明通过使用爱尔兰能源管理委员会 (CER) 数据集验证发明的SF-Heter框架,该数据集由2009年7月14日至2010年12月31日的智能计量电力用户行为试验期间获得。该数据集包括18个月来每隔30分钟的用电量记录,收集自4232个爱尔兰居民客户。

[0039] 从问卷调查中收集的信息揭示了爱尔兰居民客户的能源消费行为和习惯。对收集的数据进行分析,以揭示各种因素之间的模式和相关性,并分配标签将客户分为不同类别。提取关键标签,包括就业状况、房屋所有权和电器数量,以总结用户的用电特征。如表1所示,这些标签分为两类或三类。

[0040] 表1 家庭特征标签

编号	家庭特征	类标签	数量
1	收入者的年龄	1:青年(<35)	436
		2:中年(35 ~ 65)	2819
		3:老年(> 65)	953
2	卧室数量	1:少(< 3)	404
		2:中等(< 3)	3354
		3:多(> 4)	474
3	烹饪耗能	1:用电的	2960
		2:非用电的	1272
4	有孩子	1:有	1229
		2:没有	3003
5	居民人数	1:标准(≤ 2)	2199
		2:多数(> 2)	2033
6	社会阶层	1:AB	642
		2:C1C2	1840
		3:DEF	1760
7	娱乐设备数量	1:少(≤ 3)	2050
		2:中等(4 ~ 5)	1290
		3:多(> 4)	892
8	工作状态	1:在职的	1285
		2:失业的	2947

[0041] 将数据分成两部分:80%的数据用于训练模型,剩余20%数据用于测试模型的性能。设置10个零售商,将训练数据分为 $|R| = 10$ 的10个部分,每个部分约占消费者的8%。在半监督测试时,消费者的部分标签被保留,使用已标记和未标记的数据度量模型的性能。

[0042] 对于家庭特征分类问题,根据与分类器相关的混淆矩阵定义分类器性能,对于 K 分类问题,可以得到混淆矩阵 CM 。其元素 $CM(i, j)$ 表示在类 i 中被分类为 j 的样本数。根据混淆矩阵,分类精度ACC度量在评价非平衡数据集分类器的性能方面有局限性,马修斯相关系

数MCC考虑了各种混淆矩阵结果的比例,更适合不平衡的数据集。因此,采用MCC作为评估指标,其计算公式为:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

[0043] 其中TP(True Positive)和TN(True negative)表示正确分类的正例和负例的数量。FN(False negative)和FP(False positive)分别表示错误分类的正例和负例的数量。MCC为1表示完全分类,0表示不优于随机分类,-1表示分类与观察不一致。

[0044] 1) 进行智能电表数据异质性评估。研究异质性对本发明SF-Heter框架的影响,并将本发明与五种算法对比:

①SF-Heter:本发明提出的框架;

②Decentralized:消费者单独训练他们的模型,而不交换任何个人数据或本地模型参数;

③Centralized:消费者用所有的个人数据训练他们的模型;

④FedAvg:领先的联邦学习算法,采用基于梯度的方法,将所有局部模型简单地求均值为全局模型;

⑤FedProto:基于带有近端项的FedAvg框架的算法。

[0045] 这五个框架的比较如下表2所示:

表2 联邦学习框架对比

框架	隐私	异构性	未标记数据	全局信息	通信成本
Decentralized	✓	×	×	×	-
Centralized	×	×	×	✓	58.34s
FedAvg	✓	×	×	✓	5.94s
FedProto	✓	×	×	✓	0.01s
SF-Heter	✓	✓	✓	✓	0.01s

[0046] 表2中框架的比较表明,SF-Heter框架在隐私、异构性和未标记数据处理方面表现出色,且其通信成本非常低,仅为0.01s。同时,SF-Heter还可以通过其他零售商的样品获得全局信息。相反,Decentralized方法支持隐私,但不支持异构、未标记的数据或收集全局信息。Centralized方法不提供隐私保护,并且不能有效地处理异构或未标记的数据。FedAvg能够维护隐私、通过参数上传收集全局信息,但仍然缺乏处理异构和未标记数据的能力,因为FedAvg只与服务器共享模型。相反,FedProto是专门为处理异构数据而设计的,但FedProto没有对未标记数据进行处理。

[0047] 混合异质性是指三种异质性(不等长度、不等采样、数据不平衡)同时存在的情况。在这种情况下,每个零售商可能通过不同的抽样策略收集独特的数据,这些数据的长度不同,服务于不同的地区。在混合异质性情景下,SF-Heter框架的ACC、MCC指标如下表3所示:

表3 实验结果对比

框架	SF-Heter		Decentralized		centralized		FedAvg		FedProto	
指标	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	ACC
#1	0.9272	0.1811	0.9011	0.1162	0.6662	0.1076	0.6470	0.0	0.9002	0.1169
#2	0.8950	0.2369	0.8675	0.2099	0.7142	0.0597	0.6084	0.0	0.8688	0.2063
#3	0.8022	0.1457	0.8199	0.0568	0.6549	0.0868	0.6617	0.0027	0.8221	0.0528
#4	0.8408	0.3092	0.8416	0.1675	0.6696	0.1409	0.7107	0.0046	0.8448	0.1645
#5	0.8502	0.5005	0.8231	0.1955	0.6590	0.3161	0.6926	0.2821	0.8230	0.1934
#6	0.6600	0.1535	0.6420	0.1412	0.4648	0.1005	0.4682	0.0531	0.6381	0.1344
#7	0.7189	0.3857	0.6729	0.2805	0.4542	0.1409	0.4624	0.2174	0.6702	0.2719
#8	0.7297	0.3093	0.8326	0.1724	0.6772	0.1924	0.6592	0.0029	0.8320	0.1407
平均	0.8030	0.2770	0.7999	0.1675	0.6200	0.1431	0.6138	0.0704	0.7999	0.1601

[0048] 该实验结果验证了SF-Heter框架在解决所有三种类型的异质性方面的有效性，SF-Heter算法以最高的ACC和MCC得分优于其他五种算法。相比之下，FedAvg算法在案例#1和#2中未能准确预测不平衡标签，揭示了其在处理此类数据分布方面的局限性。案例#3#4#8的结果进一步突出了传统FL方法在处理强异质性时所面临的挑战。另一方面，FedProto和SF-Heter的表现证明了将异质性纳入学习过程的好处。这两种算法都优于传统方法，强调了在FL设置中考虑不同异质性因素的重要性。FedProto不仅改善了个人客户的个性化，还从全局的角度，通过类似集群的策略，聚合了标签相似的原型。该算法的辅助损失由 λ 值控制，可以在个性化和全局信息之间取得平衡。除FedProto外，SF-Heter充分利用了未标记的异构数据，实验结果表明，本发明SF-Heter框架是处理标记数据量有限的复杂异构场景的有效解决方案，同时能保护用户隐私。

[0049] 2) 进行未标记数据评估。采用四种比较的模型，分别为：

- ①supervised PNN:基于提出的FL框架的典型有监督多层感知器 (MLP) 模型；
- ②supervised CNN-LSTM:典型的监督随机森林模型；
- ③F-Heter:在FL框架和监督部分中提出的模型；
- ④SF-Heter:在FL框架和半监督部分中提出的模型。

[0050] 实验比较了SF-Heter的MCC与其他三种算法的不同水平的未标记数据和异质性。由实验结果知，随着未标记比率的增加，四种算法的性能都有所下降。

[0051] 由于更多的未标记数据为学习模型引入了更大的不确定性，这种性能下降是预期内的。在这四种算法中，SF-Heter甚至始终优于其他算法，实验结果证明了SF-Heter框架在联邦设置中处理未标记数据的有效性。

[0052] 可以理解，本发明是通过一些实施例进行描述的，本领域技术人员知悉的，在不脱离本发明的精神和范围的情况下，可以对这些特征和实施例进行各种改变或等效替换。另外，在本发明的教导下，可以对这些特征和实施例进行修改以适应具体的情况及材料而不会脱离本发明的精神和范围。因此，本发明不受此处所公开的具体实施例的限制，所有落入本申请的权利要求范围内的实施例都属于本发明所保护的范围内。

MODLinear本地模型

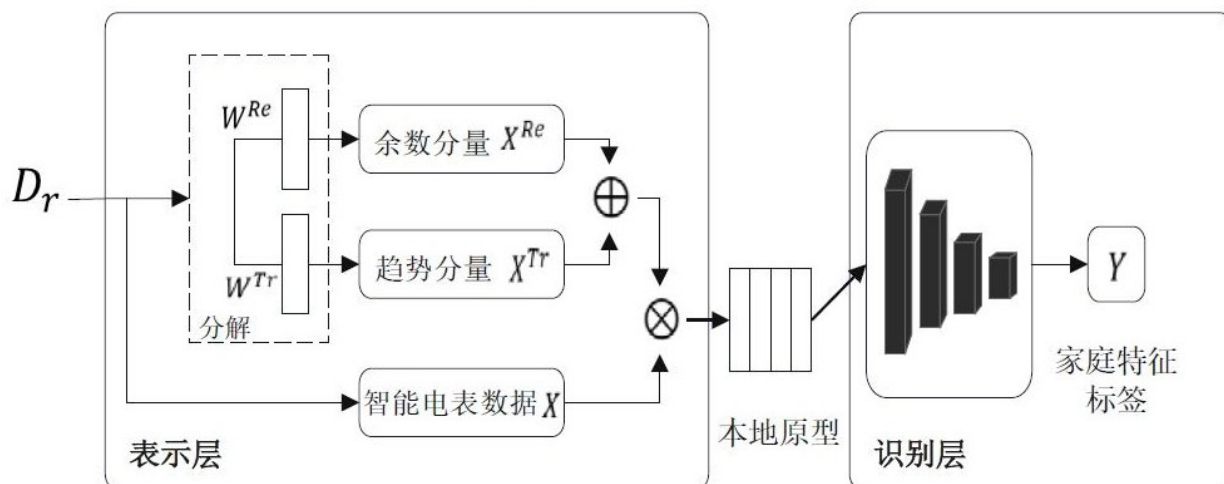


图 1

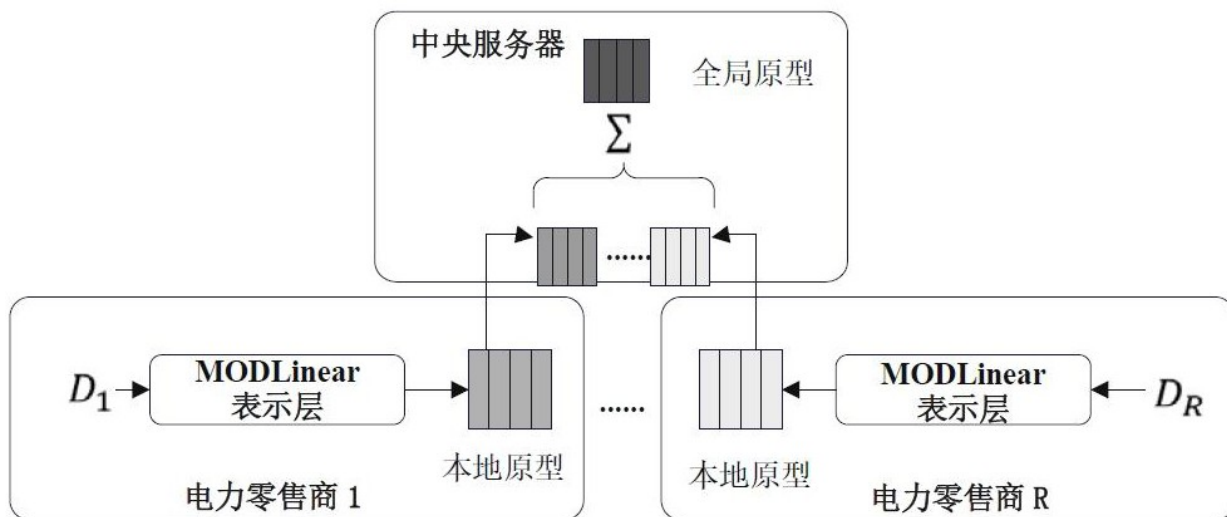


图 2

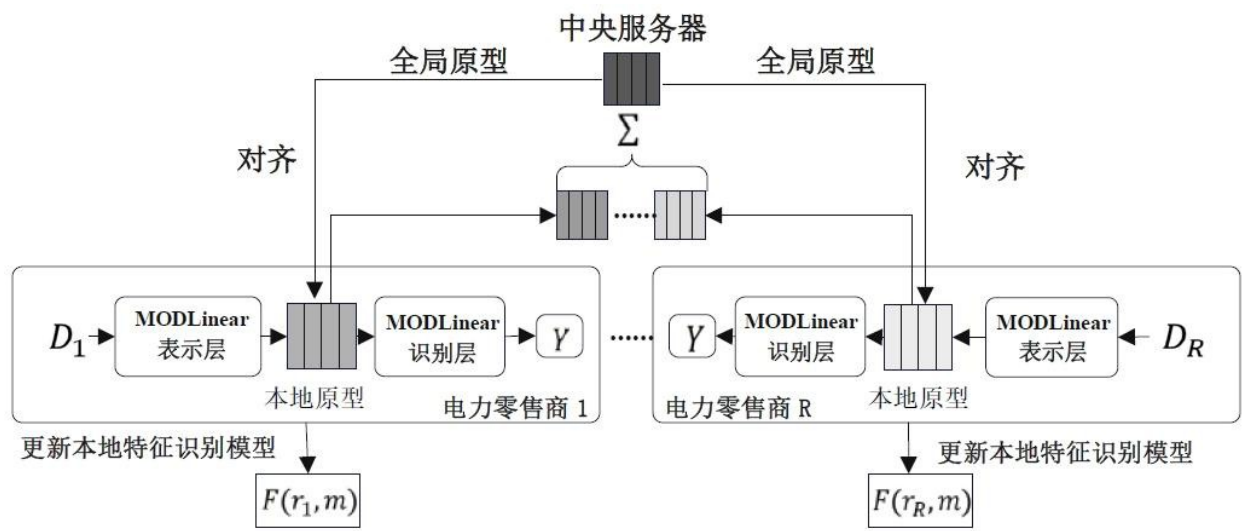


图 3