重庆邮电大学研究生综述报告

|  |  |
| --- | --- |
| **报告题目** | 联邦半监督学习方法研究综述 |
| **学 号** | S221201028 |
| **姓 名** | 吕九峦 |
| **学位层次** | 硕士研究生 |
| **学位类别** | 工学硕士 |
| **学科专业** | 软件工程 |
| **研究方向** | 视觉智能软件 |
| **研究生导师** | 韦庆杰 正高级工程师 |
| **报告日期** | 2023年 9月13日 |
| **报告分数** |  |
| **导师签名** |  |
| **学院审核** |  |

# 联邦半监督学习方法研究综述

吕九峦

重庆邮电大学软件工程学院 重庆 400065

E-mail: 1342409485@qq.com

联系电话: 17384775461

**摘要：**半监督学习，旨在缓解训练数据集中标签数据有限的问题。当数据集分布在多个不同数据拥有方而形成“数据孤岛”时，半监督学习模型训练和应用成为了一项挑战。在数据安全隐私保护需求下，针对联合多方数据进行半监督学习的研究方向，展开了联邦半监督学习方法综述调研。首先，详细介绍了半监督学习方法和联邦学习技术。然后，归纳和总结了联邦半监督学习方法，涉及两种场景：（1）客户端同时拥有带标签和不带标签数据（标签在客户端）；（2）带标签数据仅在服务器上可用（标签在服务器）。最后，探讨了联邦半监督在多种数据类型上推广和在纵向联邦架构等方向存在的问题及进一步工作。

**关键字：**半监督学习，联邦学习，机器学习，数据隐私

# A Review of Federated Semi-supervised Learning

**Abstract:** Semi-supervised learning, an intermediate between supervised and unsupervised learning, mitigates the dearth of labeled samples by incorporating unlabeled ones, optimizing model performance. However, practical applications frequently encounter data scarcity and isolation. Federated learning offers a solution, enabling multiple parties to collaboratively construct machine learning models, maintaining privacy. Hence, integrating semi-supervised learning with federated learning presents a novel strategy. This paper firstly delivers a comprehensive review of semi-supervised and federated learning techniques. Subsequently, it categorizes and summarizes semi-supervised learning methods based on federated learning, considering two scenarios: (1) clients possessing both labeled and unlabeled data, and (2) labeled data residing solely on the server. Lastly, it delves into issues and future directions in federated semi-supervision concerning credibility enhancement and vertical federation architecture.

**Keywords:** semi-supervised learning, federated learning, machine learning, data privacy1

## 引言

### 1.1 研究背景及意义

大多数应用领域的问题是没有足够多的标记数据，而未标记数据的获得相对容易。要获得有标记的数据非常困难，因为需要经验丰富的领域专家来标注未标记的数据。半监督学习（Semi-Supervised Learning）可以解决这个问题，它是介于监督学习和无监督学习之间的一种机器学习方法。与监督学习和无监督学习相比，半监督学习是最近才出现的。监督学习是从给定的示例集合中学习，每个示例包括问题实例及其标签（通常由该领域的专家给出）。例如，在分类问题中，待分类的数据元素表示为特征向量，而类别则表示为分类标签。示例集（也称为训练集或标记集）用于构建分类器，该分类器可用于对任何新的给定数据实例进行分类。在无监督学习中，我们得不到任何标签集，得到的只是无标签数据。无监督学习的主要任务是找到数据集中存在的结构，比如聚类结构。与有监督学习相比，无监督学习更加困难，因为没有标签来评估结果[1]。

目前，半监督学习方法是数据集集中在同一个数据拥有方，进行模型训练和应用。然而，在很多实际应用中，所需要的数据集分散在多个不同的数据拥有方，且各方都有数据安全隐私保护的需求。由此形成的“数据孤岛”为半监督学习模型训练和应用带来了挑战。我们需要有一种机制，可以在保护数据安全隐私的情况下，联合多方数据共同训练一个优化的半监督学习模型。联邦学习[2]（Federated Learning，FL）可以解决多方联合建模的问题，在保护用户隐私和数据安全的前提下，合理共享数据，使得多元数据得以融合和碰撞，最大程度的地发挥数据应有的价值。具体来讲，联邦学习可以在参与方数据不共享的情况下，让各个企业进行联合建立全局共享的机器学习模型，从而解决“数据孤岛”问题，提高预测机器学习模型的准确度。因此，将半监督学习与联邦学习结合，可以将传统的半监督学习扩展到更广泛的应用环境中，从而引发了一个新的研究课题，即联邦半监督学习。

### 1.2 文献调研

在联邦半监督学习领域，常见的中文关键词有：联邦学习、半监督学习，联邦半监督学习；常见的英文关键词有：Federated Learning, Semi-supervised Learning, Federated Semi-supervised Learning。为了系统地了解联邦半监督学习领域的相关研究，我们使用表1中的中文关键词在知网、万方和维普数据库上进行了相关中文文献的调研，调研结果如图1.1-1.3所示；我们使用表1中的英文关键词在SCI、IEEE Xplore、SpringerLink和ACM数据库上进行了相关英文文献的调研，调研结果如图1.4-1.7所示。

表1 文献调研使用的关键词

|  |  |  |  |
| --- | --- | --- | --- |
| 关键词 |  |  |  |
| 中文 | 联邦学习 | 半监督学习 | 联邦半监督学习 |
| 英文 | Federated Learning | Semi-supervised Learning | Federated Semi-supervised Learning |

图1.1 知网数据库中近五年每个中文关键词的文献数量

图1.2 万方数据库中近五年每个中文关键词的文献数量

图1.3 维普数据库中每年每个中文关键词的文献数量

图1.4 不同中文数据库近五年文献数占比

使用中文关键词在知网、万方和维普三个中文数据库上共检索出中文文献59、121、和30篇，检索的文献均属于计算机技术学科。以“联邦学习”、“半监督学习”以及“联邦半监督学习”为关键词的中文文献数分别为9363、7238和70，因此可以了解到，在3个中文关键词中，以“联邦学习”为关键词的文献最多。从图1.1，图1.2和图1.3可以看出，近五年，联邦半监督学习领域的中文文献数量整体呈上升趋势。

图1.5 不同中文关键词文献数占比

图1.6 SCI数据库中近五年每个英文关键词的文献数量

图1.7 IEEE数据库中近五年每个英文关键词的文献数量

图1.8 SpringerLink数据库中近五年每个英文关键词的文献数量

图1.9 ACM数据库中近五年每个英文关键词的文献数量

图1.10 英文数据库近五年文献数占比

图1.11 不同英文关键词文献数占比

使用英文关键词在SCI、IEEE、SpringerLink和ACM上分别检索出英文文献15340、25369和1252篇，。从图1.11可知，以Semi-Supervised Learning为关键词的论文数最多，其次是Federated Learning。从图1.6-1.9可以看出，近五年联邦半监督学习领域的文献数呈上升趋势。

综上所述，近五年联邦半监督学习领域的文献数呈上升趋势，越来越受到研究人员的关注。其中以Federated Learning和Semi-Supervised Learning作为关键词的英文文献数最多，因此我们主要使用Semi-Supervised Learning和Federated Learning检索英文文献。我们的参考文献主要有两种类型的文章构成:第一种是综述性质的，提到了现有问题、方法或提供了对现有技术的一般概述;第二种则是提出了新的研究思路。我们的方法分类参考多篇近三年的综述论文整理得出。

### 1.3 文章结构

文章的其余部分安排如下：第二章介绍半监督学习的概念、算法；第三章概述联邦学习的概念、类型、隐私保护方案和具体过程；第四章探讨当前主要的联邦半监督学习方法；第五章为当前存在的问题及进一步工作。

## 半监督学习概述

本章概述半监督学习，分别介绍半监督学习概念（2.1节）、常见的半监督学习方法（2.2节）。

### 2.1 半监督学习概念

半监督学习是一种机器学习方法，它同时使用标记和未标记的样本数据，这是机器学习领域中的一种较新的方法。在机器学习领域，通常将任务区分为两种主要类型：有监督学习和无监督学习[3]。有监督学习涉及对输入和对应输出的数据点进行训练，目标是建立一个分类器或回归器，预测未曾见过的输入的输出。相对地，无监督学习并未提供特定的输出值，而是试图从输入中推导出潜在的结构。作为机器学习的子领域，半监督学习旨在结合有监督和无监督学习的特性[4, 5]。通常，半监督学习算法通过利用与其他任务相关的信息，以提升某一任务的性能。例如，在处理分类问题时，会利用未标记的额外数据点来增强分类效果。

在传统的有监督学习问题中，有一个包含个标记数据点的集合，每个数据点包括一个来自给定输入空间的对象和一个关联的标签。在回归问题中，为实数值；在分类问题中，为类别。基于数据集，有监督学习方法试图推导出一个函数，能够成功预测未曾见过的输入的标签。在实际分类问题中，通常还有一个包含个未标记数据点的集合。这些数据点，预测的测试数据，其标签是未知的。半监督分类方法试图利用这些未标记的数据点来构建性能超越仅使用标记数据的学习器。

半监督学习的一个前提是，输入空间的底层边际数据分布包含关于后验分布的信息。如果满足这一条件，可以借助无标签数据获取的信息，从而获取的信息。反之，如果并未包含的任何信息，那么无法通过额外的无标签数据提高预测准确性[5]。幸运的是，如实践中半监督学习方法的成功应用所示，大多数实际学习问题似乎满足前述条件。然而，和的交互方式并不总是一致的。这导致了半监督学习假设的产生，它明确定义了预期的交互类型[6]。最知名的假设包括平滑性假设（即若两个样本和在输入空间中接近，则其标签和应相同）、低密度假设（决策边界不应穿过输入空间的高密度区域）以及流形假设（同一低维流形上的数据点应具有相同的标签）。这些假设构成了大多数（如果不是全部）半监督学习算法的基础，这些算法通常依赖于其中一个或多个假设的明确或隐含满足，下面将一一介绍半监督学习中的基本假设。

①平滑性假设：平滑性假设表明，在输入空间中，相邻的两个输入点和应具有相同的标签和。虽然这一假设在监督学习中也常用，但在半监督学习中的优势更大：平滑性假设可以直接应用于无标签数据。例如，设有一个标记数据点和两个未标记数据点和，其中接近，接近，但和不接近。由于平滑性假设，仍可预期具有与相同的标签，因为接近性——从而标签——是通过传递的。

②低密度假设：低密度假设意味着分类器的决策边界最好穿过输入空间的低密度区域。换句话说，决策边界不应通过高密度区域。该假设是在输入数据的真实分布上定义的。当考虑来自该分布的有限样本集时，这基本上意味着决策边界应位于观察到的数据点很少的区域。因此，低密度假设与平滑性假设密切相关；事实上，低密度假设可视为基础数据分布的平滑性假设的对应假设。设有一个低密度区域，即较低的区域。由于中的观测数据预期较少，观察到任何一对相似数据点的可能性不大。如果将决策边界设置在这个低密度区域，不会违反平滑性假设，因为它只涉及相似的数据点对。然而，对于高密度区域，预计会有大量的数据点。因此，将决策边界设置在高密度区域将违反平滑性假设，因为对于相似的数据点，预测的标签将不同。

③流形假设：在机器学习问题中，如果数据可以用欧几里得空间表示，那么高维输入空间中的观察数据点通常集中在低维子结构上，这些子结构被称为流形，它们是局部欧氏的拓扑空间。例如，在一个所有数据点都位于球面上的三维输入空间中，我们可以认为数据位于一个二维流形上。在半监督学习中，流形假设包括两个主要观点：(a) 输入空间由多个低维流形组成，所有的数据点都位于这些流形上；(b) 位于同一流形上的数据点共享相同的标签。因此，若能确定哪些流形存在以及哪些数据点位于哪个流形上，就可以从同一流形上的已标记数据点推断出未标记数据点的类别。

### 2.2 常见的半监督学习方法

在过去二十年中，人们已经提出了许多半监督学习算法。这些方法在半监督学习假设、未标记数据的使用方式以及与有监督算法的关系等方面存在差异。如图2.1所示，现有的半监督学习方法主要分类两大类：归纳方法（2.2.1节）和转导方法（2.2.2节）[7]。

#### 2.2.1 归纳方法

归纳方法的目标是构建一个可以对输入空间中的任何对象进行预测的分类器。虽然在训练这个分类器时可以使用未标记的数据，但一旦训练完成，对新的、先前未见过的例子的预测就是独立的。这与有监督学习方法的目标相对应，即在训练阶段构建一个模型，然后用该模型预测新数据点的标签。归纳方法分为两类：包装方法和本质上的半监督学习方法。

（1）包装方法

包装方法是半监督学习中最早且最广为人知的算法之一[5]。它们依赖于一个或多个有监督的基学习器，并通过迭代训练这些学习器，使用原始标签数据和先前未标注的数据，后者通常被称为伪标签数据。整个过程通常包括训练和伪标记两个交替步骤。在训练步骤中，一个或多个监督分类器在标记数据，以及可能的先前迭代的伪标签数据上进行训练。在伪标记步骤中，生成的分类器用于推断先前未标注的对象的标签；对预测最有信心的数据点进行伪标记，以供下一次迭代使用。

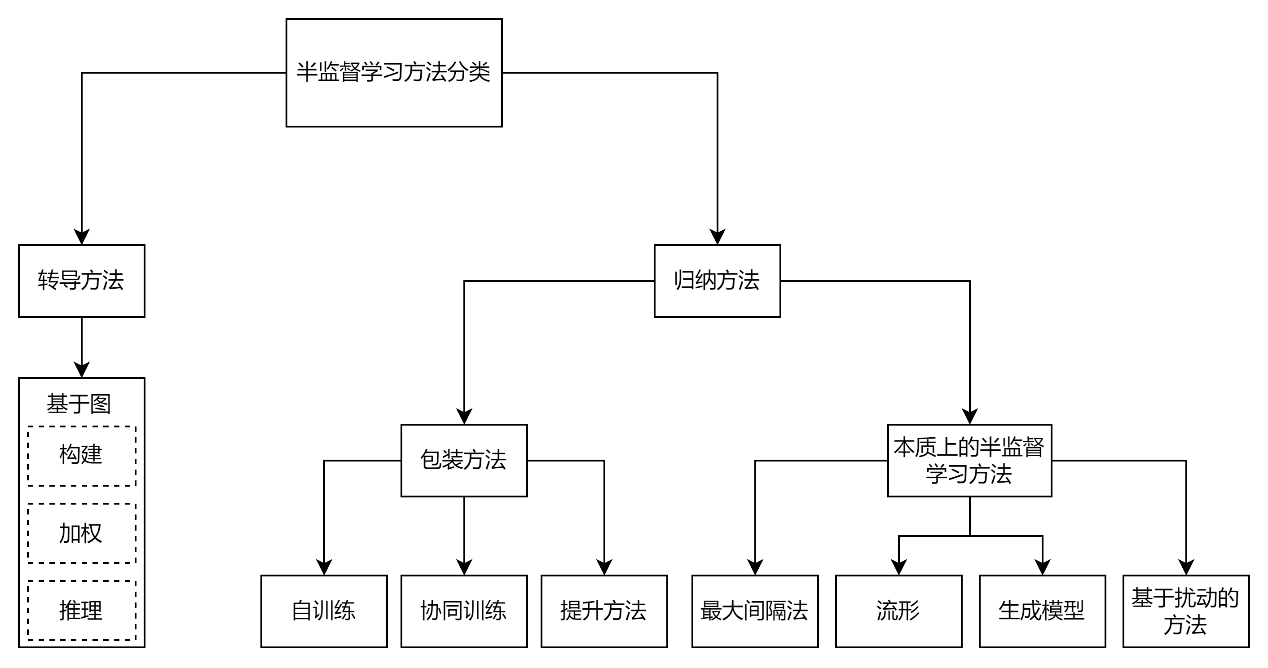


图2.1 半监督学习方法分类

包装方法的一个显著优势是，它们几乎可以与任何有监督的基学习器一起使用。有监督的基学习器可以完全不知道包装方法，包装方法只是将伪标签样本传递给基学习器，就像传递普通标签样本一样。虽然有些包装方法要求基学习器提供概率预测，但许多依赖于多个基学习器的包装方法并不要求。对于任何特定的包装方法，其所依据的半监督学习假设都取决于所使用的基学习器。从这个意义上说，包装方法本身不能被视为一种学习方法：只有当它与一组特定的基学习器相结合时，才能成为一种完整的学习方法。Triguero等人发表了一篇详尽的关于包装方法的调查。除了提供这些方法的概述，还提出了一个分类法，基于以下三个因素：（a）使用的分类器数量，（b）是否使用了不同类型的分类器，以及（c）是否使用单视图或多视图数据（即数据是否被划分为多个特征子集）。这个分类法为包装方法的研究提供了有价值的洞察。常见的包装方法有：自训练[8]，协同训练[9, 10]，提升方法[11]。

自训练方法（有时也称“自学习”方法）是最基本的伪标签方法[8]。它们由一个单一的监督分类器组成，该分类器在标注数据和早期迭代中伪标注的数据上进行迭代训练。在自训练过程的开始，仅对标注数据训练一个监督分类器。然后，用这个分类器预测未标记的数据点，并将最有把握的预测添加到标注数据集中。接着在原始标注数据和新获得的伪标注数据上重新训练分类器。这一过程通常持续进行，直到没有未标记的数据为止。自训练首先由Yarowsky提出[12]，用于文本中的词义消歧。后来，研究者为其提出了多种应用与变体。例如，Rosenberg等人在物体检测中应用自训练[13]，超越了当时的最先进模型。Dópido等人为高光谱图像分类设计了自训练策略，选取了一组候选无标签样本，并为信息量大的样本添加伪标签[14]。自训练范式包括多种设计决策，如选择伪标签数据、在后期迭代中重复使用伪标签数据以及停止标准[8, 13]。伪标签数据的选择尤为重要，因为它决定了哪些数据会被纳入分类器的训练集。在典型的自训练设置中，这种选择基于预测置信度，而置信度估计的质量对算法性能有重大影响。特别是，未标记样本的预测概率排序应反映实际的置信度排序。

协同训练作为自训练的扩展，适用于多个有监督分类器。它涉及两个或多个分类器在标记数据上的迭代训练，每次迭代中，分类器将其最有信心的预测加入到其他分类器的标记数据集。为确保成功，基学习器之间的预测不应有强相关性，避免限制交换有用信息的能力，这被称为多样性标准[9]。Zhou 和 Li[10]研究了依赖多个基学习者的半监督学习，称这些方法为基于分歧的，意指共同训练利用了学习者间的分歧，通过无标签数据交换不同的预测标签[10]。为了促进分类器的多样性，早期的协同训练方法主要依赖于数据的多个不同视图的存在，这些视图通常对应于特征集的不同子集。例如，在处理视频数据时，可以将数据自然地分解为视频数据和音频数据。这种协同训练方法属于更广泛的多视图学习方法，其中还包括广泛的监督学习算法。Xu等人[15]对多视角学习进行了全面调查。

提升方法由多个基础分类器组成，经过训练后合成预测结果[11]。最基本的形式是独立训练k个分类器并汇总其预测。技术上，提升方法是贪婪地构建分类器的加权集合。令 表示在时间时的分类器集合，其权重为。定义为对真实标签为的数据点的预测标签的损失函数。在算法的每次迭代中，都会向分类器集合中加入一个权重为的分类器，以优化成本函数。

（2）本质上的半监督方法

本质上的半监督方法直接优化包含标记和未标记样本的目标函数，不依赖任何中间步骤或监督学习基础。通常，这些方法是对现有监督方法的扩展，将未标记样本纳入目标函数。这些方法通常依赖于某种半监督学习假设。例如，最大边际方法依赖于低密度假设，而大部分半监督神经网络依赖于平滑性假设。

最大边际分类器的目标是最大化数据点与决策边界的距离，这与半监督学习的低密度假设相吻合：如果所有数据点（除个别异常值外）与决策边界的边距都较大，那么决策边界必然位于低密度区域[16]。因此，理论上，最大边际方法可以很好地应用于半监督学习：我们可以利用未标记的数据来识别低密度区域，从而确定可以获得大边距的区域。有监督最大边际分类器最突出的例子是支持向量机（SVM）：这种分类方法试图最大化决策边界与最接近决策边界的点之间的距离，同时鼓励数据点被正确分类。它是最早在半监督环境中提出的最大边际方法之一，自此以后，人们对它进行了广泛的研究[3]。

基于扰动的方法通常基于神经网络实现。神经网络能直接将额外的无监督损失项纳入目标函数，因此容易扩展至半监督学习。近年来，因其在多种应用领域的成功，神经网络再次受到重视[17-19]。因此，对半监督神经网络的研究兴趣也随之增加。特别是多层神经网络，即深度神经网络，为半监督学习提供扩展。

流形假设是许多本质半监督学习算法的基础。基于扰动的方法利用平滑性假设，对输入或分类器微小变化时的行为差异进行惩罚。然而，我们可以设想，不是所有的微小输入变化都会导致相似的输出。特别是，如果数据位于低维度流形上，我们预期分类器对流形内的微小变化应不敏感。这一观察与流形假设相符。

#### 2.2.2 转导方法

前面小节介绍的半监督学习方法都属于归纳方法：主要目标是利用标记和未标记数据构建一个能预测整个输入空间数据点标签的模型。因此，归纳学习器有明确的训练和测试阶段：在训练阶段，标记数据和未标记数据用于构建分类器；在测试阶段，此分类器用于对未标记数据或其他未见过的数据点进行分类。

本节将探讨半监督学习方法的另一大类——转导方法。与归纳算法不同，转导方法不会创建一个可以在整个输入空间运行的预测器。相反，它针对给定的一组未标记数据点生成预测结果。因此，与归纳算法不同，我们无法区分训练和测试阶段：转导方法接收标记数据和未标记数据，并专门为未标记数据输出预测结果。

转导方法通常在所有数据点（包括标记和未标记的数据点）上定义一个图，通过可能的加权边来编码数据点之间的相似性[20]。然后，定义并优化目标函数以实现两个目标：

①对于已标记的数据点，预测标签应与真实标签一致。

②根据相似性图的定义，相似的数据点应具有相同的标签预测。

基于图的半监督学习方法通常包含三个步骤：图创建、图加权和推理[21, 22]。首先，根据某种相似性度量，将图中的节点（代表数据点）相互连接。然后，对生成的边进行加权，得到权重矩阵。这两步通常被称为图构建阶段。完成后得到一个由节点集和权重矩阵W组成的图，其中节点对应数据点，权重矩阵包含所有节点对的边权重，权重为零表示无边。

一旦图构建完成，我们就可以利用它来获取未标记数据点的预测值。基于图的转导方法的目标函数通常包括：一部分用于惩罚预测标签与真实标签的不匹配，另一部分用于惩罚连接数据点的标签预测差异。具体来说，给定标记数据的监督损失函数和标记或未标记数据点对的非监督损失函数，基于图的转导方法试图找到一个标签，以最小化这些损失。

其中，λ代表监督项的重要性。另外，一些基于图的方法还对未标记预测增加了额外的单一正则项。这个基于图的方法的通用框架允许每个步骤有多种变体。这种表述在基于图的方法中很常见，且大多数基于图的推理算法都被证明符合这一框架[5, 23]。

基于图的半监督学习方法的范围可以根据两个主要阶段（即图构建和推理）的不同方法有效地进行划分。基于图的方法的早期研究主要集中在第二阶段，而对图构建的研究则很少。Zhu[5]指出，这种不平衡可能是不公正的，图构建会对分类器性能产生重大影响。后来的研究解决了这一不平衡问题，图构建从此成为一个备受关注的研究领域[24]。基于图的转导方法是在 2000 年代初提出的，而基于图的推理方法在随后的十年中得到了特别深入的研究。Zhu [24]的半监督学习调查报告和博士论文涵盖了这一领域的大部分研究成果。下面按照基于图的方法领域研究的一般时间顺序，首先介绍解决推理问题的方法，然后介绍图构建方面的研究。

（1）图的推理

推理过程包括对未标记数据点形成预测。如果标记数据的预测标签在推理过程中不固定为真实标签，则对整个预测标签集进行优化。一般来说，它们在损失函数和以及权衡参数的具体选择上有所不同。此外，一些方法只推断最可能的标签分配，而其他方法估计边际概率分布。总之，这些变化产生了大量不同的基于图的推理方法。尽管一般目标函数也适用于多类设置，但许多基于图的方法并不能自然地扩展到二进制分类之外。下面的推理方法主要考虑的是二元分类情况。

第一个基于图的半监督分类方法是由 Blum 和 Chawla[25]提出的，他们尝试使用近邻算法和邻域（连接距离小于）的数据点对来构建图。他们将边的权重保持固定和统一，但尝试改变未标记数据点之间的边相对于其他边的权重。一旦构建了图，就可以从最小切割的角度来解决优化问题。具体来说，添加一个源节点，以无限权重连接到正数据点；添加一个汇节点，以无限权重连接到负数据点。因此，确定最小切点相当于找到一组组合权重最小的边，去除这些边后，图中就没有从源节点到汇节点的路径了。在生成的图形中，所有位于包含的分量中的未标记节点都标记为正节点，而所有位于包含的分量中的未标记节点都标记为负节点。

（2）图的构建

图构建是基于图的方法中最重要的方面：为了使推理有效，构建的图必须准确捕捉局部相似性。基于图的方法最初的研究主要集中在推理阶段，对图的构建研究不多[20]。不过，近年来这种情况有所改变。人们对不同的图构建算法进行了广泛的实验，并引入了新的方法[21, 23, 24]。由于图的节点与数据点（包括已标记和未标记的数据点）相对应，因此图构建阶段相当于在节点之间形成边（产生邻接矩阵），并为它们附加权重（产生权重矩阵）。在很多情况下，节点间连通性的相似性度量也用于构建权重矩阵。

构建图形需要创建邻接矩阵，其元素表示节点对之间是否存在边。目前有三种常用的确定边的方法：邻接法，近邻法和-匹配。前两种都是局部性的，即每个节点都可以独立确定一组邻接。换句话说，一个节点的邻域构建不会影响另一个节点 的邻域构建（除非是的邻居）。另一方面，第三种方法（-匹配）优化的是全局目标，相距甚远的节点会对彼此的连接性产生重大影响。-邻接法 邻接法是最早用于图构建的方法之一，它简单地将每个节点连接到距离最大的所有节点[25]。换句话说，当且仅当 ，是某种距离度量（通常是欧氏距离）时，和之间才会产生一条边。结果图的结构在很大程度上取决于距离度量的选择。此外，由于是固定的，如果给定输入数据中的模式尺度发生变化，它就不能很好地工作[21, 24]。

## 3 联邦学习概述

联邦学习（Federated Learning，简称FL）旨在保护数据隐私的同时，通过整合不同参与者的分散数据来训练机器学习模型。本章首先介绍联邦学习的概念（3.1节），然后探讨其各种类型（3.2节），接着讨论如何通过联邦学习保护用户隐私（3.3节），最后详述联邦学习的整个过程（3.4节）。

### 3.1 联邦学习概念

在2016年，Google公司面临了一个挑战，该挑战涉及到如何在用户终端设备上进行模型训练[2]。在许多情况下，智能手机应用程序依赖于在后台运行的模型来提供服务。这些模型的训练完全依赖于用户数据的收集和分析。例如，考虑手机上的输入法软件，它需要根据每个用户的打字和拼音习惯进行持续更新，使得软件能够更好地匹配用户的打字习惯，从而让用户感觉输入法变得越来越智能。

随着智能手机和平板电脑的技术进步和广泛使用，终端设备现在已经具备了处理大量数据和进行复杂计算的能力。这意味着，这些设备可以在本地处理和存储用户的隐私数据，同时进行模型训练。这种方法不仅可以推动大数据的应用，提高服务质量，而且可以避免将数据发送到中心节点进行处理，保护用户的隐私。

然而，这种方法也带来了一个问题，即如何在保护用户隐私的同时，有效地利用分散在各个终端设备上的数据进行模型训练。为了解决这个问题，Google公司提出了一种名为联邦学习的方法。联邦学习的核心思想是，让模型在本地设备上进行学习，并将学习到的模型更新分享给其他设备，而不是共享原始数据。这种方法旨在保护用户隐私的同时，实现大规模的分布式学习。

在联邦学习的环境中，通常面对的是多个独立的参与方，每个参与方都拥有自己独特的数据集。相比传统的分布式机器学习方法，该方法通常需要将所有参与方的数据汇集到一起，以共同训练出一个统一的机器学习模型，可以将其称为模型。然而，联邦学习提供了一种新的解决方案：所有的参与方可以共同训练一个模型，第个参与方的数据集保留在本地，没有与其他参与方进行直接的数据交互。这种方法有效地保护了各参与方数据的隐私性，因为它们的数据不需要离开本地设备或被直接共享。通过这种方式，得到一个由所有参与方共同训练的机器学习模型，将其称为模型。这种模型的生成方式既保护了各参与方的数据隐私，又允许从各个独立的数据集中学习，从而提高模型的性能和泛化能力。

联邦学习允许训练得到的模型与将数据聚合在一起训练得到的模型有一定程度的性能差距，但各个参与方保护了自身数据隐私。若存在一个非负实数，使模型与的模型准确度和满足以下公式：

则称作该联邦学习算法达到-精度损失。

在研究联邦学习的应用模式时，通常会遇到两种主要的架构方式。首先，有客户端-服务器架构，这是一种中心化的模式，其中一台服务器负责协调多个客户端的学习过程[2]。在这种架构中，服务器通常负责模型的初始化和全局更新，而客户端则负责在本地数据上进行模型的训练和更新。其次，对等网络架构，这是一种更为分散的模式，其中没有中心服务器，各个节点（即客户端）直接进行通信和协作，共同完成模型的训练和更新[26]。这种架构的主要优点是它可以在没有中心服务器的情况下进行学习，从而增加了系统的鲁棒性和可扩展性。然而，这种架构也存在一些挑战，例如怎样有效地协调节点之间的通信和学习，以及如何处理可能出现的不同步问题。

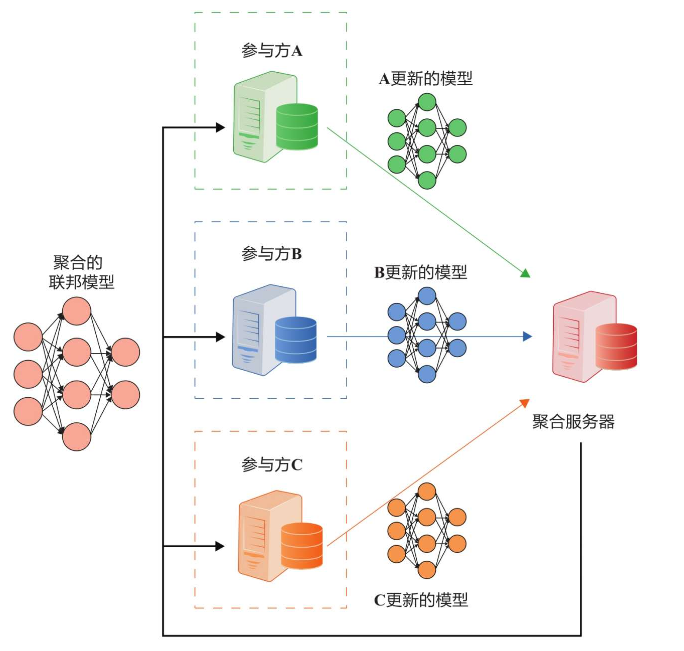


图 3.1 联邦学习客户-服务器架构

在客户端-服务器架构中，联邦学习的训练过程如图3.1所示。这个过程主要包括几个关键步骤。首先，每个数据持有者（也就是客户端）会根据其特定的数据、条件和规则在本地进行模型的训练。然后，这些客户端将它们的模型的脱敏参数发送到一个可信赖的第三方，这个第三方充当中央服务器，负责将这些参数进行整合和计算。这个中央服务器的角色主要是进行全局模型的更新，以及协调各个客户端的学习过程。完成这次全局更新后，中央服务器将更新后的模型参数发回各个客户端，以便它们更新自己本地的模型。这个过程会反复进行，直到全局模型达到收敛，即模型的性能不再有显著提高，或者达到预设的训练轮数为止。

对等网络架构中进行联邦学习训练的过程如图3.2所示。在这种架构中，各个参与方（即客户端）之间可以直接进行通信，而无需依赖第三方进行算。这种方式进一步提高了数据的安全性，因为所有的数据都保留在各自的节点上，没有任何数据需要在网络中传输。然而，这种模式也需要更多的计算资源，因为每个节点需要执行加密和解密的操作，以保证数据的安全性和隐私性[2, 27, 28]。此外，对等网络架构也需要一个有效的协调机制，以确保所有的节点能够同步地进行学习和更新，这也是这种架构的一个重要挑战。总的来说，对等网络架构为联邦学习提供了一种更为分散和安全的训练模式，但也带来了一些新的挑战和需求。

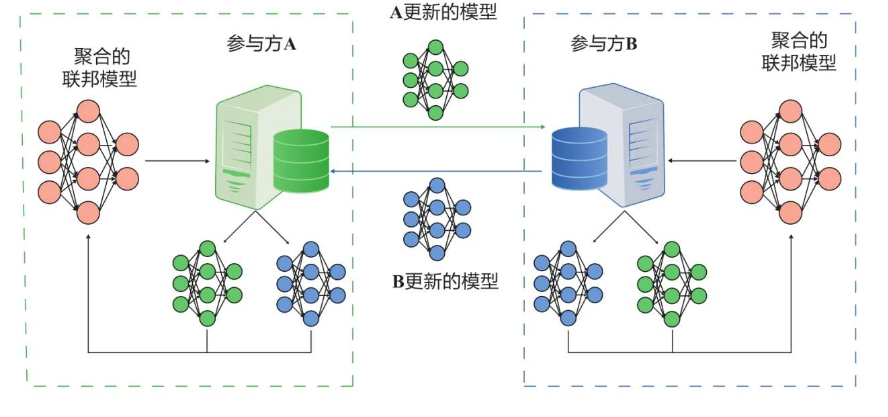


图3.2 联邦学习对等网络架构

### 3.2 联邦学习类别

在联邦学习的研究和应用中，根据每个参与方的数据分布的差异性（即不同参与方之间的数据重叠程度）将联邦学习划分为三个主要的类型[28]。首先，横向联邦学习（Horizontal Federated Learning，简称HFL）。这种类型的联邦学习主要应用于参与方之间有共享特征，但样本ID不同的情况。其次，纵向联邦学习（Vertical Federated Learning，简称VFL）。在这种类型的联邦学习中，参与方之间拥有相同的样本ID，但特征不完全相同。最后，联邦迁移学习（Federated Transfer Learning，简称FTL）。这种类型的联邦学习是一种特殊的情况，它主要适用于那些特征和样本ID都不相同，但又存在一定关联性的场景[29]。这三种类型的联邦学习提供了不同的方式来处理数据分布的异质性，从而使得联邦学习可以适应更多的应用场景。

如图3.3所示，横向联邦学习的核心概念是样本的联合，其适用于那些联邦学习的参与方拥有重叠的数据特征，即数据特征在各个参与方之间是对齐的，但每个参与方拥有的数据样本是不相同的[30]。这种情况类似于在表格数据中进行水平划分，因此称之为横向联邦学习。在横向联邦学习的应用场景中，参与方的业务通常是相同的，但客户或用户群体是不同的。也就是说，这种类型的联邦学习主要应用于那些特征重叠多，而用户重叠少的场景。

例如，考虑多个不同地区的银行，业务功能和服务可能非常相似，这意味着数据特征是相似的。然而，由于客户群体不同，因此拥有的数据样本也是不同的。在这种情况下，这些银行可以通过横向联邦学习的方式来协同建立一个机器学习或深度学习模型。利用这个模型，他们可以更准确地进行数据的回归预测或分类预测，从而提高他们的业务效率和客户满意度[28]。

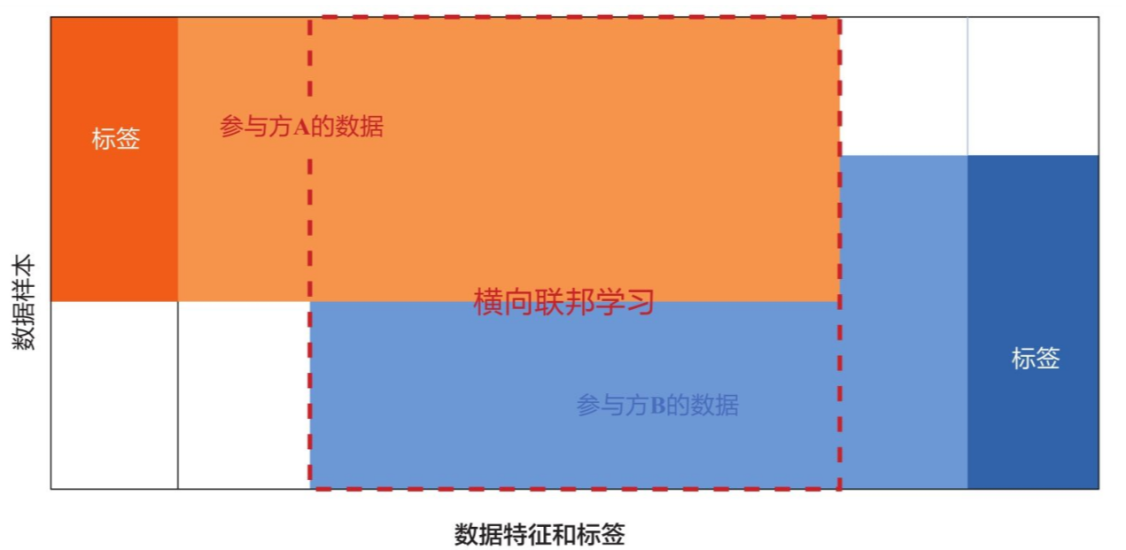


图3.3 横向联邦学习

联邦学习通常由两个以上的用户作为参与方，以及一个可信的第三方作为中心服务器进行组成。各个参与方的本地数据样本数以及本地数据特征数可能不足以训练一个优秀的机器学习模型，因此需要其他参与方来共同参与协同训练模型。可信第三方作为中心服务器的作用类似于分布式机器学习的服务器，收集各个参与方的梯度，并在中心服务器上进行聚合，完成聚合操作后返回给各个参与方相应的应更新的梯度。

在横向联邦学习中，具体步骤如下：

①各参与方在本地计算模型梯度，并使用同态加密、差分隐私或秘密共享等加密技术，对梯度信息进行掩饰，并将掩饰后的梯度信息（简称为加密梯度）发送给聚合服务器。

②服务器进行安全聚合操作对加密梯度进行聚合，使用基于同态加密的加权平均算法（如FedAvg算法）。

③服务器将安全聚合后得到的梯度信息发送给各个参与方。

④各个参与方对收到的加密梯度信息进行解密，并根据解密后的梯度结果信息更新各自的本地模型参数。

如图3.4所示，纵向联邦学习的核心思想是特征的联合，其主要适用于联邦学习的参与方拥有重叠的数据样本，即各参与方之间的数据样本是对齐的，但在数据特征上存在差异[30]。这种情况类似于在表格数据中进行垂直划分，因此将其被称为纵向联邦学习。在纵向联邦学习的应用场景中，参与方的用户或样本通常是重叠的，但特征可能存在差异。

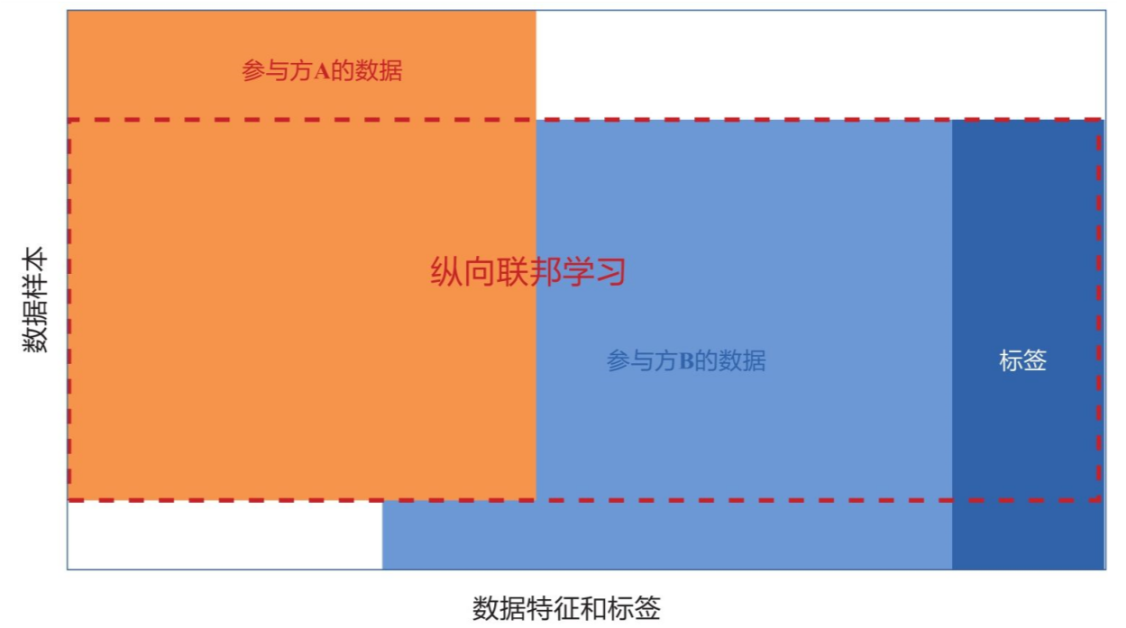


图3.4 纵向联邦学习

例如，考虑同一地区的超市和银行，用户群体可能都是该地区的居民，因此他们的数据样本是相同的。然而，由于业务类型和服务不同，因此数据特征可能会有所不同。在这种情况下，这些超市和银行可以通过纵向联邦学习的方式来协同建立一个机器学习或深度学习模型。利用这个模型，他们可以更准确地进行数据的回归预测或分类预测，从而提高业务效率和客户满意度[28]。总的来说，纵向联邦学习提供了一种有效的方式，使得那些有共享样本但特征不完全相同的参与方可以共同进行模型的训练和学习，而无需共享原始数据。

在纵向联邦学习中，首先要在不暴露各个参与方自身数据的情况下，进行样本对齐操作，得到多个参与方的共同用户。之后纵向联邦学习具体步骤如下：

①中心服务器创建加密密钥对，并将公共密钥发送给各个参与方。

②各个参与方计算中间信息，将中间信息进行加密和交换。中间信息用来帮助计算梯度和损失值。

③各个参与方计算己方模型的加密梯度并分别加入附加掩码。持有标签的参与方还会计算加密损失。参与方将加密的结果发送给中心服务器。附加掩码被用来使自身的梯度信息不暴露给中心服务器。

④中心服务器对梯度和损失信息进行解密，并将结果发送回各个参与方。各个参与方解除梯度信息上的掩码，并根据这些梯度信息来更新模型参数。

如图3.5所示，联邦迁移学习是联邦学习和迁移学习两种思想的结合。它主要适用于那些参与者之间的特征和样本重叠都非常少的情况。在这种情况下，需要寻找一种方法来利用这些参与者的数据进行模型的训练和学习，即使数据在特征和样本上都有所不同。

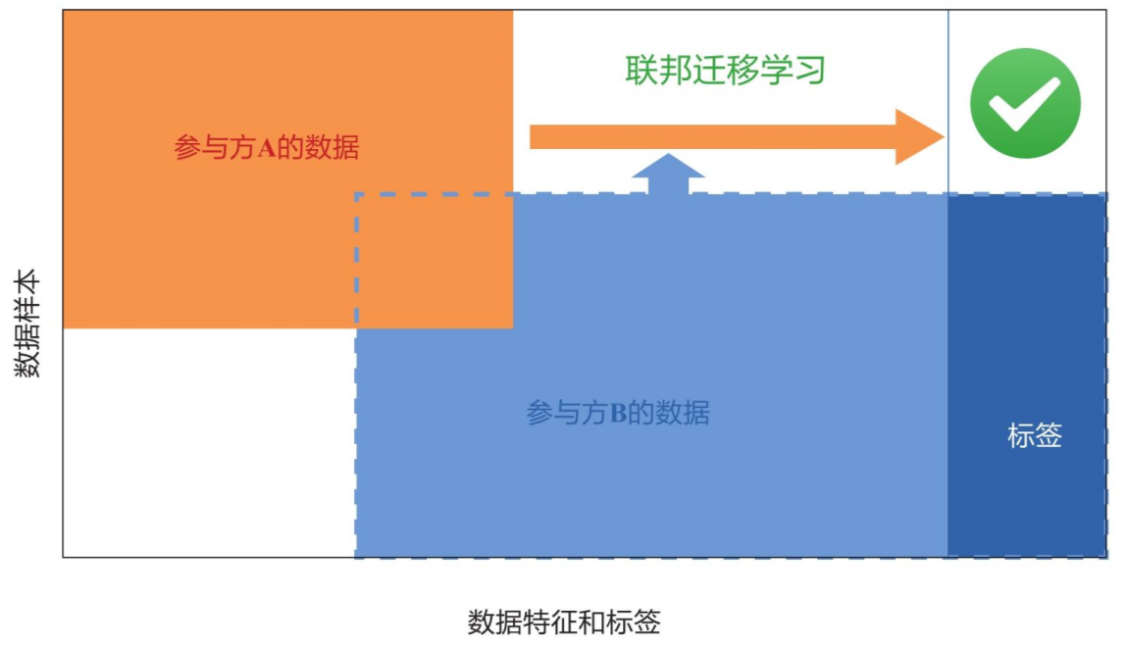


图3.5 联邦迁移学习

在目前，大部分联邦学习的研究都是基于横向联邦学习与纵向联邦学习的，在联邦迁移学习领域的研究暂时还不多。因此，主要讨论的联邦学习类型为横向联邦学习与纵向联邦学习。

### 3.3 联邦学习隐私保护方案

联邦学习对于数据的隐私保护方案主要分为三种，分别是安全多方计算、同态加密、差分隐私。

（1）多方安全计算

安全多方计算（Secure Multi-party Computation，简称SMC）是一种隐私保护方案，它最初是为了解决所谓的“百万富翁问题”[31]。这种计算框架的主要目标是使多个参与方能够协同计算一个函数，该函数基于每个参与方的私有输入，而不需要将这些输入直接传递给其他参与方。这样，每个参与方只能得到自己的函数输出值，而无法获取其他参与方的输入和输出信息。

在安全多方计算的过程中，允许计算私有输入值的函数，从而使得每个参与方只能获取到与其输入值对应的函数的输出值。这种方式保证了数据的隐私性，因为每个参与方无法获取到其他参与方的输入和输出信息。因此，安全多方计算为在保护隐私的同时进行协同计算提供了一种有效的解决方案。这在那些需要高度隐私保护，例如金融、医疗和政府等领域，具有重要的应用价值。

例如，假设一个私有的数值被分给个参与方作为共享方，则每一个参与方只能获知的内容，所有参与方能够协同地计算

参与方只能根据自己的输入来获取输出值，而不能知道其他参与方的输入。

（2）同态加密

同态加密是一种基于计算复杂性理论和数学难题的密码学技术。这种技术的核心特性在于，它允许在加密的数据上进行特定的运算，并得到一个加密的结果。当对这个加密的结果进行解密时，会得到一个值，这个值与在未加密的原始数据上进行同样的运算得到的结果是完全相同的。这种特性使得同态加密成为一种非常有用的工具，因为它允许在不知道原始数据的情况下处理和操作数据。

同态加密的概念最初是由Rivest等人在1978年提出的[31]。他们的目标是创建一种新的加密方法，这种方法可以在密文上进行操作，从而允许在加密内容上进行特定的代数运算。这种方法的优点在于，它可以保护数据的隐私，同时还可以进行必要的数据处理和运算。因此，同态加密技术在那些需要在保护数据隐私的同时进行数据处理和分析的场景中，例如云计算、数据挖掘和机器学习等领域，具有广泛的应用前景。

Paillier于1999年提出的同态加密方案中[32]，若用, 表示两个经过同态加密后的密文信息，对于加法同态运算，有

对于标量乘法同态，有

使用同态加密方法能够在不对密文解密的情况下就可以进行对密文进行计算，这样计算就不需要获得明文内容而只需要获得加密过的密文，这样可以很好的保护敏感数据和信息，同时又可以进行相应的计算操作，能够处理相应的加密数据。

（3）差分隐私

差分隐私（Differential Privacy）是一种由Dwork等人提出的隐私保护方案[33]。这种方案的核心思想是引入一种随机性机制，使得当输入数据中的单个样本发生变化时，输出的结果分布不会有显著的变化。这样，即使攻击者可以观察到输出结果，他们也无法确定输入数据中的任何特定样本的存在或状态，从而保护了数据的隐私。

差分隐私通过引入一种称为随机噪声的概念来实现这一目标。具体来说，当计算一个函数的结果时，会在结果中添加一些随机噪声，使得结果的精确值被掩盖。这种方法的优点在于，它可以在保护数据隐私的同时，仍然可以进行有效的数据分析和处理。因此，差分隐私技术在那些需要在保护数据隐私的同时进行数据处理和分析的场景中，例如数据挖掘、机器学习和统计分析等领域，具有广泛的应用前景。

对于只有一个记录不同的两个数据集和，一个随机算法，以及对于任意的输出，若称随机算法提供-差分隐私保护，则有：

其中表示隐私预算，表示失败概率，当时，就得到了性能更好的-差分隐私。

差分隐私提供了一种信息安全性保障，即函数的输出结果对数据集的任何特定记录都不太敏感。因此，差分隐私能够被用于抵抗成员推理攻击。

## 4 联邦半监督学习研究现状

当前的联邦半监督学习方法主要基于横向联邦架构（即参与方共享特征，但样本ID不同）。根据标注数据的位置分为标签在客户端和标签在服务器端两种情况[34]。

（1）标签在客户端

标注数据存于客户端，而服务器只能获取未标注数据。例如，某家公司欲利用智能手机拍摄的图片训练一个物体检测的联邦学习模型，但无法直接访问用户的本地数据，只能依赖用户的标注(见图4.1)。首先，对于不同的参与方（用户），样本ID是不同的，但参与方（用户）手机里的图片特征是一样的，所以构成了横向联邦的设置。其次，用户通常不愿为每张照片标注，给基于横向联邦的半监督学习创造了一个"客户端标注"（label-at-client）的环境。

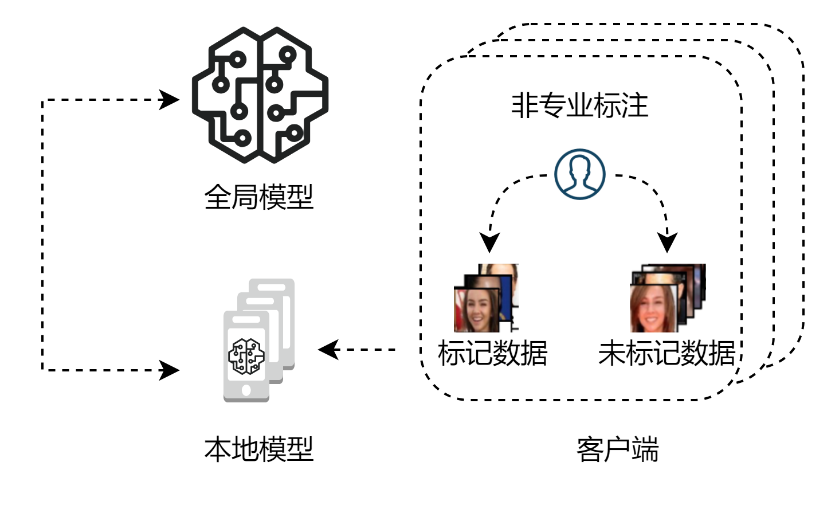


图4.1 标记数据在客户端的情况

RSCFed[35] 主要关注联邦半监督学习中的标签隔离问题和数据异质性问题。在局部训练中，采用师生模型[36]对无标签数据进行训练。为进一步解决数据异质性问题，RSCFed 提出了子共识抽样法和距离加权聚合法。在每一轮中，通过对所有参与者的多个子集进行独立抽样，从而聚合出多个子共识模型，这样每个子共识模型都有望包含拥有标记数据的参与者。此外，本地模型会根据它们与子共识模型的距离进行加权，这样偏差模型就会得到较低的权重，其影响也会降到最低。

FedSSL[37]解决了标签隔离问题、数据隐私问题和数据异构问题。为了便于对未标记的客户端进行本地训练，FedSSL 利用了伪标记技术。此外，为解决数据异构问题，FedSSL 学习全局生成模型，从统一的特征空间生成数据，从而通过生成的数据缓解数据异构问题。最后，为了防止生成模型造成的隐私泄露，FedSSL 利用差分隐私（DP）来限制生成模型中训练数据的信息泄露。

FedMatch[38]提出了一种客户端间一致性损失来解决数据异质性问题。具体来说，对每个客户端的前k个最近客户端进行采样，在每个数据样本上，本地模型的输出与前k个客户端模型的输出进行正则化，以确保一致性。此外，FedMatch 还提出了分离式学习方法，将标注数据和未标注数据的参数分开，未标注数据的参数是稀疏的。更新时，只有未标记数据的客户端会上传稀疏张量，从而降低通信成本。

FedPU[39]研究了半监督学习中更具挑战性的环境--正向和无标签学习，在这种环境中，每个客户端只有类的子集标签。在这种情况下，客户端只掌握所有类别中的一部分信息，从而导致严重的标签隔离问题。为了解决这个问题，FedPU 提出了一个新颖的目标函数，即把学习客户端负类的任务交给拥有负类标签数据的其他客户端。这样，每个客户端只负责学习正类，并可自行进行局部训练。根据经验，在正类和无标签学习设置中，所提出的 FedPU 优于 FedMatch[38]。

AdaFedSemi[40]提出了一种系统，可在联合半监督学习中利用服务器端无标记数据实现效率与模型准确性之间的权衡。在每一轮学习中，模型都是通过客户端的标注数据进行训练，并在服务器端进行汇总。服务器端未标注数据通过伪标注纳入训练过程。AdaFedSemi[40]确定了平衡效率和性能的两个关键参数，即客户端参与率P和伪标签的置信度阈值。较低的 P可以降低通信成本和模型准确性，而较高的可以降低服务器端计算成本，同时也限制了未标记数据的使用。。实验表明，AdaFedSemi 通过动态调整 P 和在不同的训练阶段实现了效率和准确性之间的良好平衡。

DS-FL[41]解决了与 AdaFedSemi 类似的问题，即客户端拥有标签数据，而服务器拥有非标签数据。它提出了一种集合伪标签解决方案来利用服务器端的非标签数据。具体来说，它不是对数据样本使用单一的伪标签，而是对所有客户端生成的伪标签进行平均，。这将创建一个客户端模型集合，并提供更好的性能。此外，由于只传输伪标签而不是模型参数，通信成本可以大大降低。此外，DS-FL 发现在伪标签上进行训练会导致预测熵增大。因此，DS-FL 提出了一种减少熵的聚合方法，即在聚合之前使局部输出。

（2）标签在服务器端

标注数据存于服务器，而客户端只有未标注数据。例如，一家可穿戴设备公司希望利用联邦学习训练健康监测模型（见图4.2）。在这种情况下，由于用户通常缺乏专业知识，无法标注健康相关数据，因此客户端的数据是未标注的。

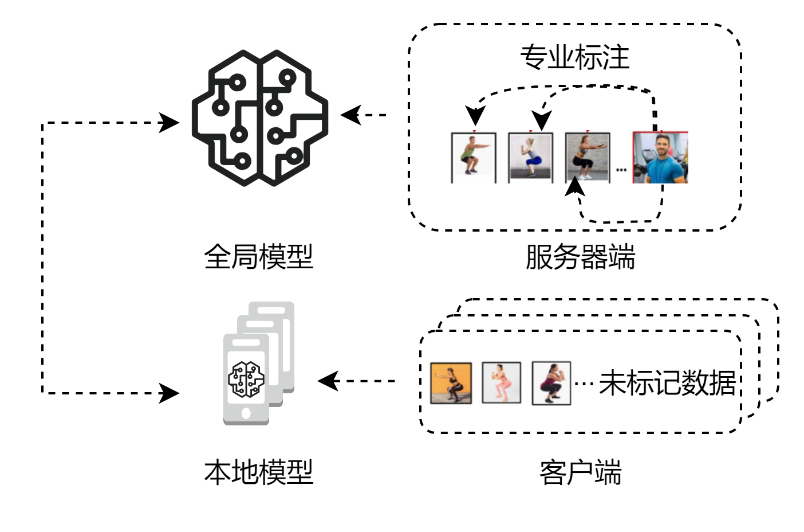


图4.2 标记数据在服务器端的情况

标签数据在客户端的情况比标签在服务器端设置更复杂，原因是所有客户端都只拥有未标记数据，无法为联邦模型提供额外的监督信号。如文献[38, 42]所示，仅使用无标签数据进行训练可能会导致遗忘从有标签数据中学到的知识，从而影响模型性能。

为了解决标签数据和非标签数据之间的隔离问题，FedMatch[38]提出了一种不相干的学习方案，即分别为标签数据和非标签数据设置两套参数。在非标签数据上进行训练时，标签数据的参数是固定的，反之亦然，以防止知识被覆盖。非标记数据的参数会在参与者和服务器之间传输，而非标记数据的参数设置为稀疏的，这为通信效率带来了额外的好处。此外，为了解决不同客户端持有的异构数据问题，FedMatch 提出了客户端间一致性损失，这样不同参与者的本地模型就能在相同数据上产生相似的输出。

SemiFL[42]采用另一种方法来解决这些挑战。它建议使用标注数据对全局模型进行微调，以提高其质量，并减轻客户端无监督训练所造成的遗忘。此外，SemiFL 建议最大限度地提高客户端模型与全局模型之间的一致性，而不是在客户端之间对模型输出进行正则化。具体来说，全局模型为客户端的未标记数据生成伪标签，而客户端的局部模型则根据伪标签进行训练。实证结果表明，与 FedMatch 相比，SemiFL能产生更有竞争力的结果。

（3）联邦半监督学习方法总结

表4.1总结了前文介绍的基于横向联邦的半监督学习方法，按照标签在客户端/服务器端划分。

表 4.1 基于横向联邦的半监督学习方法总结，×表示没有关注这个问题

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类别 | 方法 | 采用的半监督学习算法 | 隐私保护方案 | 数据异质问题解决办法 | 如何提高性能 |
| 标签在客户端 | RSCFed[35] | 教师学生模型[36] | × | 加权距离聚合 | × |
| FedSSL[37] | 伪标记[43] | 差分隐私 | 全局生成模型 | × |
| FedMatch [38] | 伪标记[43] | × | 客户端间一致性 | 分散学习和稀疏学习 |
| FedPU[39] | PU Learning[44] | × | 客户端间一致性 | × |
| AdaFedSemi[40] | 伪标记[43] | × | × | 调整置信度阈值和参与率 |
| DS-FL[41] | 集成未标记[45] | × | × | 传输日志、无参数 |
| 标签在服务器端 | SemiFL[46] | 伪标记[43] | × | 减少熵的平均值 |  |
| FedMatch[38] | 伪标记[43] | × | 客户端间一致性 | 分散学习和稀疏学习 |

## 5 当前存在的问题及进一步工作

与有监督的一般的联邦学习相比，联邦半监督学习问题仍未得到充分探索。在联邦半监督学习领域，存在以下几个问题：

①如何在特征空间不同且部分数据未标记的情况下分类：在某些场景中，多方机构各自持有部分重叠但特征空间不同的私有数据集，且在这些机构中，有的只拥有正样本数据，而其他机构持有的数据是未标记的。这些机构的共同目标是通过协作，利用他们各自的数据，训练出一个能够准确地从未标记数据中识别出可靠阳性样本的推荐模型，这带来了一个新的挑战，即利用纵向联邦半监督学习进行用户推荐。

②如何将现有的半监督学习方法推广到表格类数据：现有许多半监督学习方法主要基于图像数据。然而，在大数据领域，数据类型多样，可能是图像，也可能是表格数据。如何在半监督学习中应用这些方法并进行学习是一个挑战。尤其是当数据为表格类时，基于图像的方法可能无法适用。

③联邦学习本身架构的改进：在研究联邦半监督学习的过程中，联邦学习本身的过程和一些算法细节成为挑战。例如，如何更好地构建数据交换模式，以及如何优化联邦学习架构模式等，都是需要解决的问题。

针对联邦半监督学习领域存在的问题，进一步的工作可以包括研究基于纵向联邦半监督学习算法以处理特征空间不同且部分数据未标记的情况，探索适用于表格数据的半监督学习方法或调整现有的基于图像的方法，以及研究优化联邦学习的数据交换模式和架构模式以提高学习效率和模型性能。

## 参考文献

[1] Jain, A., M. Murty, and P.J.J. Flynn, *Data clustering: a review, ACM Computing Survey.* 1999. **31**(3).

[2] Konečný, J., et al., *Federated learning: Strategies for improving communication efficiency.* 2016.

[3] Bishop, C.M. and N.M. Nasrabadi, *Pattern recognition and machine learning*. Vol. 4. 2006: Springer.

[4] Chapelle, O., M. Chi, and A. Zien. *A continuation method for semi-supervised SVMs*. in *Proceedings of the 23rd international conference on Machine learning*. 2006.

[5] Zhu, X.J., *Semi-supervised learning literature survey.* 2005.

[6] Hady, M.F.A. and F.J.H.o.N.I.P. Schwenker, *Semi-supervised learning.* 2013: p. 215-239.

[7] Van Engelen, J.E. and H.H.J.M.l. Hoos, *A survey on semi-supervised learning.* 2020. **109**(2): p. 373-440.

[8] Triguero, I., et al., *Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study.* 2015. **42**: p. 245-284.

[9] Wang, W. and Z.-H. Zhou. *A New Analysis of Co-Training*. in *ICML*. 2010.

[10] Zhou, Z.-H., M.J.K. Li, and I. Systems, *Semi-supervised learning by disagreement.* 2010. **24**: p. 415-439.

[11] Zhou, Z.-H., *Ensemble methods: foundations and algorithms*. 2012: CRC press.

[12] Yarowsky, D. *Unsupervised word sense disambiguation rivaling supervised methods*. in *33rd annual meeting of the association for computational linguistics*. 1995.

[13] Rosenberg, C., M. Hebert, and H. Schneiderman, *Semi-supervised self-training of object detection models.* 2005.

[14] Dópido, I., et al., *Semisupervised self-learning for hyperspectral image classification.* 2013. **51**(7): p. 4032-4044.

[15] Xu, X., et al., *Co-labeling for multi-view weakly labeled learning.* 2015. **38**(6): p. 1113-1125.

[16] Ben-David, S., et al. *Learning low density separators*. in *Artificial Intelligence and Statistics*. 2009. PMLR.

[17] Collobert, R., et al., *Natural language processing (almost) from scratch.* 2011. **12**(ARTICLE): p. 2493− 2537.

[18] Krizhevsky, A., I. Sutskever, and G.E.J.A.i.n.i.p.s. Hinton, *Imagenet classification with deep convolutional neural networks.* 2012. **25**.

[19] LeCun, Y., Y. Bengio, and G.J.n. Hinton, *Deep learning.* 2015. **521**(7553): p. 436-444.

[20] Zhu, X. and J. Lafferty. *Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning*. in *Proceedings of the 22nd international conference on Machine learning*. 2005.

[21] Jebara, T., J. Wang, and S.-F. Chang. *Graph construction and b-matching for semi-supervised learning*. in *Proceedings of the 26th annual international conference on machine learning*. 2009.

[22] Liu, W., J. Wang, and S.-F.J.P.o.t.I. Chang, *Robust and scalable graph-based semisupervised learning.* 2012. **100**(9): p. 2624-2638.

[23] Chong, Y., et al., *Graph-based semi-supervised learning: A review.* 2020. **408**: p. 216-230.

[24] de Sousa, C.A.R., S.O. Rezende, and G.E. Batista. *Influence of graph construction on semi-supervised learning*. in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. 2013. Springer.

[25] Blum, A. and S. Chawla, *Learning from labeled and unlabeled data using graph mincuts.* 2001.

[26] Feng, S. and H.J.a.p.a. Yu, *Multi-participant multi-class vertical federated learning.* 2020.

[27] McMahan, B., et al. *Communication-efficient learning of deep networks from decentralized data*. in *Artificial intelligence and statistics*. 2017. PMLR.

[28] Yang, Q., et al., *Federated machine learning: Concept and applications.* 2019. **10**(2): p. 1-19.

[29] Kairouz, P., et al., *Advances and open problems in federated learning.* 2021. **14**(1–2): p. 1-210.

[30] Mammen, P.M.J.a.p.a., *Federated learning: Opportunities and challenges.* 2021.

[31] Ryffel, T., et al., *A generic framework for privacy preserving deep learning.* 2018.

[32] Paillier, P. *Public-key cryptosystems based on composite degree residuosity classes*. in *International conference on the theory and applications of cryptographic techniques*. 1999. Springer.

[33] Dwork, C., et al. *Calibrating noise to sensitivity in private data analysis*. in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. 2006. Springer.

[34] Jin, Y., et al., *Federated Learning without Full Labels: A Survey.* 2023.

[35] Liang, X., et al. *Rscfed: Random sampling consensus federated semi-supervised learning*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[36] Tarvainen, A. and H.J.A.i.n.i.p.s. Valpola, *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.* 2017. **30**.

[37] Fan, C., J. Hu, and J. Huang. *Private semi-supervised federated learning*. in *International Joint Conference on Artificial Intelligence*. 2022.

[38] Jeong, W., et al., *Federated semi-supervised learning with inter-client consistency & disjoint learning.* 2020.

[39] Lin, X., et al. *Federated learning with positive and unlabeled data*. in *International Conference on Machine Learning*. 2022. PMLR.

[40] Wang, L., et al. *Enhancing Federated Learning with In-Cloud Unlabeled Data*. in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 2022. IEEE.

[41] Itahara, S., et al., *Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data.* 2021. **22**(1): p. 191-205.

[42] Diao, E., J. Ding, and V.J.A.i.N.I.P.S. Tarokh, *SemiFL: Semi-supervised federated learning for unlabeled clients with alternate training.* 2022. **35**: p. 17871-17884.

[43] Lee, D.-H. *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*. in *Workshop on challenges in representation learning, ICML*. 2013. Atlanta.

[44] Bekker, J. and J.J.M.L. Davis, *Learning from positive and unlabeled data: A survey.* 2020. **109**: p. 719-760.

[45] Bachman, P., O. Alsharif, and D.J.A.i.n.i.p.s. Precup, *Learning with pseudo-ensembles.* 2014. **27**.

[46] Ganin, Y., et al., *Domain-adversarial training of neural networks.* 2016. **17**(1): p. 2096-2030.