去中心化学习中的可扩展数据点估值

Konstantin D. Pandl[a, b]、Chun-Yin Huang[b]、Ivan Beschastnikh[b]、Xiaoxiao Li[b]、Scott Thiebes[a]、Ali Sunyaev[a]

[a]*Karlsruhe Institute of Technology*，卡尔斯鲁厄，德国
[b]英属哥伦比亚大学，温哥华，加拿大

arXiv:2305.01657v1 [cs.LG] 1 May 2023

摘要

现有关于数据估值的研究 聚焦于对联邦学习与群体学习中的客户端 贡献进行估值，并且在客户端之间的数据是独立 且同分布的（IID）。实际上，数据很少按照IID分布。

我们提出了一种名为 DDVal 的去中心化数据估值方法，能够在联邦学习与群体学习中对个人数据点进行估值。DDVal 基于共享深度特征，并通过 k近邻近似方法对夏普利值进行近似。这为新的应用提供了可能，例如，同时奖励为去中心化机器学习任务提供数据的机构和个人。

通过 DDVal 对数据点进行估值，还可以得出关于机构贡献的层级性结论。我们通过实验证明，在估计机构贡献方面，DDVal 的准确性高于现有的联邦学习夏普利值近似方法。具体而言，在机构之间的 IID 和非独立同分布数据分布下，其对夏普利值的近似余弦相似度均达到 99.969%，而当前最佳方法分别为 99.301% 和 97.250%。

DDVal 的扩展能力取决于数据点的数量，而非客户端的数量，且具有对数线性复杂度。与具有指数复杂度的现有方法相比，这种扩展方式更为有利。我们表明，在客户端众多、每个客户端拥有的数据点较少的数据分布场景中，DDVal 尤其高效——例如，超过 16 个客户端各自拥有 8,000 个数据点。通过将 DDVal 集成到去中心化系统中，我们表明，它不仅适用于集中式联邦学习，也适用于去中心化群体学习；这与新兴互联网技术（如 Web3）的研究方向十分契合，即奖励向算法提供数据的用户。

1. 引言

机器学习（ML）训练需要大量多样且标注准确的数据。在某些行业中，这可能颇具挑战性，例如医疗保健行业，数据流动受隐私法规约束 [1]，而数据标注则需要专家，他们可能难以寻觅且费用高昂 [2]。通常，在某些司法管辖区，患者可能拥有自己的数据 [3],，使用这些数据训练模型可能需要经历漫长的同意流程。此外，患者分享数据以供人工智能系统使用的激励有限，这可能导致数据匮乏 [4]。

这些挑战引发了从业者和研究人员对相关系统的兴趣，这类系统让个人掌控自己的数据，并在其向人工智能算法提供数据时给予奖励 [5]。一个研究方向是采用基于区块链的代币经济和 Web3 技术 [6, 7]。早期方案将区块链网络与可信执行环境结合起来，在受保护的硬件隔离区中处理数据 [5]。处理内容包括训练机器学习模型、根据数据点对机器学习任务的影响为其估值，以及使用机器学习模型进行推理。因此，请求机器学习模型推理的模型使用者会向模型提供者和数据提供方付费，以使用机器学习模型。这一系列研究属于近期实践中的一个更大趋势：实体要求机器学习公司为其提供数据支付报酬 [8]。

另一个研究方向通过联邦学习（FL）解决训练数据可用性问题 [9]。联邦学习使机构能够在不共享数据的情况下协作训练机器学习模型。联邦学习过程既可以由中央服务器协调，也越来越多地通过客户端之间的点对点通信来实现。这也称为群体学习（SL），通常通过私有区块链网络上的智能合约实现。群体学习提高了联邦学习过程的稳健性和可审计性。从业者发布的首批白皮书表明，有必要为基于区块链的群体学习环境中的机构开发奖励系统 [10]。

目前，关于在保护隐私的同时激励个人贡献数据的研究十分有限。有一篇论文提出了一种基于可信执行环境（TEEs）的概念，在该环境中汇集数据、评估数据价值并训练机器学习模型 [5]。然而，该论文并未详细描述技术系统，而可信执行环境的可扩展性有限，并且存在通过侧信道攻击泄露私人信息的风险。总体而言，这引出了一个问题：是否存在其他方式，能够在保护个人隐私的同时激励其贡献数据。为保护隐私，我们以去中心化学习技术为基础，同时致力于评估数据价值。因此，我们与现有研究 [11, 12] 有所不同，重点在于评估数据点的价值，而非个人贡献。

近期研究提出了面向联邦学习（FL）中各机构的奖励系统，该系统依据其贡献质量进行奖励 [4, 13]，而贡献质量通常需要耗费大量计算资源才能量化。

然而，这些奖励系统依赖的数据估值技术最适用于独立同分布（IID）数据分布 [11, 12], ，而这在实际中并不现实。此外，联邦学习客户端的数据可能来自个人——例如，在医院参与联邦学习的情况下，数据可能由患者提供。目前尚不清楚，如何设计这类奖励系统，才能奖励联邦学习（FL）或分割学习（SL）场景中受资源限制（计算资源、存储资源、网络带宽）的客户端所对应的个人；也不清楚如何最好地利用区块链和分布式账本技术（DLT）的潜在优势（例如，机构通过质押代币作出承诺） [14]。

这类知识对于设计医疗保健等行业的数据供给激励机制，以及释放人工智能的全部潜力至关重要。初步研究表明，此类系统具有实用性和必要性，但并未提供详细的技术架构，且依赖可信执行环境等特殊技术 [5]。我们旨在开发一种名为 DDVal 的分层数据估值方法，用于评估去中心化学习中的数据点，而非机构的贡献。这样，我们便能够计算群组的分层奖励和数据价值，例如拥有多个数据点的患者所作的贡献，或整个机构所作的贡献。我们将 DDVal 集成到一个系统中，在该系统中，我们把估值后的数据转换为奖励，并开发一种融入 DDVal 的分割学习和代币经济架构。我们的核心贡献有四点：

- 我们提出了去中心化学习中数据点估值的概念，并介绍了一种名为 DDVal 的方法，对其进行了评估。
- 我们提出了联邦学习中的分层数据估值概念，并展示了我们还可以对机构的贡献进行估值，并将其与现有方法进行比较。
- 我们开发了一种智能合约系统，将分割学习与代币经济相结合，以透明的方式自动激励机构参与去中心化学习。
- 我们概述了该系统如何通过激励患者贡献数据，并促进医疗保健领域可信人工智能的发展。

本文其余部分的组织如下。第 2 节介绍 FL 和 SL、FL 中的贡献估值、集中式机器学习中的夏普利值估计以及代币经济的背景。随后，我们在第 3 节介绍系统设计。接着，在第 4 节介绍用于评估该系统的方法。第 5 节呈现结果。第 6 节进行讨论。最后，在第 7 节总结全文。

2. 背景

2.1. 联邦学习与群体学习

FL 支持在由不同客户端持有的相互分离的数据孤岛上训练机器学习模型。在其最初的应用场景中，FL 被提出用于训练智能手机键盘的语言模型 [9, 15] 同时保护用户的隐私。联邦学习（FL）过程由多个迭代的通信轮次组成，每个轮次包含四个步骤。首先，将全局机器学习模型分发给各个客户端。其次，各客户端在本地进行优化。第三，将客户端的梯度发送至服务器。第四，对模型和梯度进行聚合，通常通过取平均值的方式生成更新后的模型。通信轮次会持续重复，直至达到结束条件（例如，基于测试集的预测性能不再提升）。由于实际训练数据不会离开客户端的本地环境，联邦学习有助于保护数据所有者的隐私。因此，如今联邦学习在医疗保健等行业中的应用受到了广泛讨论 [16]。

分割学习（SL）是联邦学习（FL）的一种进一步去中心化的变体。与每个客户端连接到一个中心化服务器不同，各客户端以点对点的方式相互连接 [17]。分割学习通常通过私有区块链网络和独立的文件共享网络来实现。区块链是一种由复制状态机构成的网络，能够以防篡改的方式在去中心化网络中存储数据 [18]。在分割学习中，机器学习模型和梯度，或其哈希值，存储在私有区块链网络上。因此，文献中也将其称为基于区块链的联邦学习。这有助于记录模型和梯度更新，并对分割学习过程进行控制。与联邦学习相比，分割学习的优势包括更高的可审计性和可解释性、更强的网络鲁棒性，以及通过区块链网络实现访问控制。由于分割学习主要影响学习过程中的通信层，因此，分割学习所获得的模型在参数或属性方面通常与联邦学习并无差异。去中心化可能会增加网络的复杂性，但对于医疗保健等高度重视可信机器学习的特定行业而言，这一特性具有重要价值 [17, 19, 14]。由于复杂的机器学习模型通常过于庞大，无法存储在区块链上，因此，一种常见的解决方案是将区块链网络与星际文件系统（IPFS）网络相结合 [20]。IPFS 是一种点对点文件共享网络，其中文件通过哈希值进行索引。因此，如果客户端知道文件的哈希值，就可以从网络中检索该文件 [21]。对于基于区块链的联邦学习，模型和梯度的哈希值存储在链上。因此，网络中的所有客户端都能接收这些哈希值，并通过 IPFS 检索相应的文件 [14]。

2.2. 联邦学习中的贡献估值

联邦学习（FL）中的数据估值通常侧重于评估各个客户端的贡献。数据估值中的一个主流概念是夏普利值，其定义如下列公式所示，其中 D 表示由各个客户端组成的集合，z 表示其大小为N，而 U 表示效用函数（例如，机器学习模型的受试者工作特征曲线下面积，即 AUROC）：

$$\varphi_{Shap}(z_i) = \frac{1}{N} \sum_{S \subseteq D \setminus z_i} \frac{1}{\binom{N-1}{|S|}} (U(S \cup z_i) - U(S))$$

根据这一定义，夏普利值（SV）唯一地满足三个理想属性，这些属性与数据估值 [22, 23] 的背景高度相关。首先，群体理性意味着联邦学习联盟的全部价值增益都将完整分配给所有参与机构：$U(D) = \sum_{z_i \in D} \varphi(z_i)$。

其次，公平性意味着，如果机构 z_i 对联盟的所有子集都没有边际贡献，那么该机构的夏普利值应为零：即，对于所有 $S \subseteq D \setminus z_i$，若 $U(S \cup z_i) = 0$，则 $\varphi(z_i) = 0$。同时，如果两个机构 z_i 和 z_j 为联盟的某个子集带来的性能提升相同，那么它们的贡献量化值也应相同，即对于所有 $S \subseteq D \setminus z_i \setminus z_j$，若 $U(S \cup z_i) = U(S \cup z_j)$，则 $\varphi(z_i) = \varphi(z_j)$。

第三，可加性意味着，在多个效用下的夏普利值之和，等于这些效用之和所对应的效用函数下的夏普利值：$\varphi(v_1, z_i) + \varphi(v_2, z_i) = \varphi(v_1 + v_2, z_i)$。

在其规范形式中，计算夏普利值需要评估（训练和测试）实体的每个可能子集；对于 N 个联邦学习参与者，需要进行 $2^N - 1$ 次评估。因此，即使客户端数量很少，计算夏普利值的计算开销也可能非常大。

因此，研究人员开发了近似方法，以更少的计算资源估计夏普利值。大多数此类近似算法旨在高效近似联盟的训练模型 [11], ，某些方法还旨在减少需要测试的联盟数量 [24] ，或加快测试过程 [12]。Song 等人提出的第一种 SV近似方法 [11] 通过训练最大的联盟，并汇总梯度来近似较小的联盟。随后，对所有近似得到的联盟进行测试。这将训练复杂度从训练 $2^N - 1$ 个联盟大幅降低至仅训练 1 个联盟。因此，该方法被称为单轮（OR）近似。尽管仍需测试 $2^N - 1$ 个模型，但测试通常比训练所需的计算资源和通信开销少得多。

Kumar 等人提出的另一种名为 SaFE 的方法也采用类似的第一步：正常训练最大的联邦学习联盟 [12]。之后，每个客户端提取每个数据点的特征向量，并基于这些特征向量训练一个逻辑回归（LR）模型分类器。然后，将由逻辑回归（LR）分类器组合而成的 $2^N - 1$ 个联盟进行组合并测试。在测试数据集的特征向量上测试此类逻辑回归（LR）分类器的过程非常快速，因此，作者表明其方法可以扩展到 20 个及以上的客户端 [12]。

这些 SV 近似方法的准确性通常通过标准 Shapley 值向量与近似得到的夏普利值向量之间的余弦相似度来衡量 [12, 11]。余弦相似度能有效表示两个向量夹角的余弦值：精确近似时该值为 1，而该值为 -1 则表示近似值完全相反。以下公式给出了计算两个向量**a**和**b**之间余弦相似度的方法。

$$S_C(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

2.3. 集中式机器学习中的 *Shapley* 值估计

在集中式机器学习中，夏普利值估计通常用于在训练过程中评估各个数据点的价值。与联邦学习不同，联盟的规模要大得多（即数万、数十万甚至更多数据点）。因此，无法计算标准形式的夏普利值，需要采用高效的近似方法。集中式机器学习中一种常用的近似方法是基于 k 近邻（K 近邻）分类器的夏普利值近似方法 [22]。为此，首先使用全部训练数据点训练神经网络（NN）。随后，提取训练点和测试点的深度特征。深度特征是通过提取 NN 最后一层激活值，对训练数据点和测试数据点进行表征；这一方法与 Kumar 等人针对联邦学习采用的方法类似。之后，获得一个 K 近邻分类器来替代 NN。因此，与整个训练数据集相比，只有一小部分训练数据点会影响局部推理点上的决策。由此无需考虑所有子集，从而加快了计算速度 [22]。随后，可利用一个测试数据点和一个 K 近邻参数 K，按照以下公式递归计算N 个数据点的夏普利值 φ：

$$\varphi(z_N) = \frac{\mathbb{1}[y_{z_N} = y_{test}]}{N}$$

$$\varphi(z_i) = \varphi(z_{i+1}) + \frac{\mathbb{1}[y_{z_i} = y_{test}] - \mathbb{1}[y_{z_{i+1}} = y_{test}]}{K} \frac{min(K, i)}{i}$$

我们还在算法 1 中说明了这一过程。在实践中，测试数据点有多个，而训练数据点的夏普利值会根据夏普利值的可加性，在各个测试数据点上取平均。一般而言，对于 N 个待估值的数据点，计算 K 近邻夏普利值的复杂度为 $O(N \log N)$。成功的应用包括医疗保健领域的医学影像，例如高效识别误标图像，或移除不必要的数据，更好地保护患者隐私 [25, 26]。

算法 1 ComputeSVs 函数（KNN-SV [22]）

输入：训练数据：$D = \{(x_i, y_i)\}_{i=1}^{N}$，测试数据：$D_{test} = \{(x_{test,i}, y_{test,i})\}_{i=1}^{N_{测试}}$

输出：SV $\{\varphi_i\}_{i=1}^{N}$

对于 $j \leftarrow 1$到 N_{test} **执行**

  $(\alpha_1, \cdots, \alpha_N) \leftarrow$训练数据的索引，按升序排列使用 $d(\cdot, x_{test})$;

  $\varphi_{j,\alpha_N} \leftarrow \frac{\mathbb{1}[y_{\alpha_N} = y_{test}]}{N}$;

  对于 $i \leftarrow N - 1$ 到 1 **执行**

    $\varphi_{j,\alpha_i} \leftarrow \varphi_{j,\alpha_{i+1}} + \frac{\mathbb{1}[y_{\alpha_i} = y_{test,\ j}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{test,\ j}]}{min\{K,\ i\}} \frac{\{\ \}}{K}$;

<escape>6: </escape>**对于** $i \leftarrow 1$至 N **执行**

  $\varphi_i \leftarrow \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} \varphi_{j,i}$;

2.4. 代币经济

代币经济使主体能够基于分布式账本技术（DLT，即

区块链） [6]在主体（如个人或组织）之间转移资产（如加密货币）的所有权。因此，代币经济减少了对可信第三方的需求，并催生了新型商业模式和透明的业务流程，其中包括个人数据变现。这有助于主体之间的协作与合作。在许多情况下，代币经济系统由多个分布式账本组成，每个账本都针对特定用例进行了优化，并采用跨链技术实现分布式账本之间的通信 [6]。代币是代币经济的重要组成部分，用于表示资产，例如货币价值。比特币或以太坊等传统加密货币价格波动剧烈，可能不适用于支付活动。因此，以太坊等区块链会承载稳定币，这类稳定币与一种稳定货币（如美元）挂钩，旨在使其价格几乎不发生波动 [27]。以太坊上的这类稳定币通常采用 ERC-20 标准实现。该标准提供了标准化接口和库，以便开发可与 ERC-20 代币交互的智能合约。

早期研究提出了代币经济，以激励客户端 s 在联邦学习中作出贡献 [28, 29]。例如，在 FedCoin [29], 中，作者将夏普利值计算纳入联邦学习，并将其应用于区块链的区块生成；同时，根据联邦学习客户端所作贡献的质量向其支付报酬。其他早期研究则提出，使用基于可信执行环境的区块链，为数据估值提供安全的计算空间 [5]。模型使用方随后可以根据数据价值，通过区块链系统向模型和数据提供方付费。

3. 系统设计

3.1. 系统概述

我们的核心贡献是开发了一种用于联邦学习和分割学习的新型数据估值方法 DDVal，可对单个数据点进行估值。我们还通过一个新颖的系统，展示了 DDVal 在分割学习和代币经济中的集成。该系统包含机器学习部分和网络（区块链与 IPFS）部分。机器学习部分的核心是训练联邦联盟，并采用一种基于 KNN-SV 方法对数据进行分层估值的方法。我们将在第 3.2 节详细介绍该方法。

对于分割学习（SL），我们使用第 3.3 节所述的带有智能合约的私有区块链网络和私有 IPFS 网络。对于代币经济，我们为公有区块链设计并部署了一个智能合约，具体见第 3.4 节。分割学习（SL）和代币经济的整个过程也在图 5.4（见第 5.4 节）的时序图中进行了说明。

3.2. 分层数据估值方法

首先，我们旨在为联邦学习和分割学习评估数据点的价值。一种简单的方法是先评估客户端贡献，然后让每个客户端像在集中式机器学习中那样评估其数据点的价值，并将两者相乘。然而，这无法体现联邦学习中多个客户端协作所带来的效用。因此，我们采用了一种更为精细的方法，其灵感来自集中式机器学习中的 KNN-SV 方法。基于单个夏普利值对于我们还旨在推导数据点群组（例如客户端）的贡献。

第一步是让联邦学习（FL）或分割学习（SL）过程在所有客户端上进行多轮重复通信，直到达到结束条件为止。结束条件可以定义为：验证数据集上的评估指标（即 AUROC）达到平台期。这样，所有客户端都会获得相同的全局模型。随后，开始执行分层数据估值的特殊过程，其分别从联邦学习服务器的角度和客户端的角度在算法 2 和算法3 中进行了描述。为了首先获得深度特征，每个联邦学习客户端都会使用全局模型，对其各自的每个训练数据点和测试数据点执行推理。在此推理过程中，它们存储神经网络最后一个输出层的激活值，这些激活值被称为数据点的深度特征。然后，在集中式联邦学习场景中，客户端将深度特征提交给服务器；在去中心化分割学习场景中，则将其提交给其他客户端。接着，我们根据第2.3节中所述的 KNN-SV 计算过程，为每个训练数据点计算夏普利值（SV）。此外，我们对夏普利值进行归一化，使其总和与最大联盟的效用相匹配，如以下公式所示：

$$\varphi_{AUC}(z_i) = \varphi(z_i)\frac{AUC_{coalition} - 0.5}{\sum_i \varphi(z_i)}$$

因此，我们将联盟的有效效用定义为其 AUROC 得分减去 0.5，因为随机分类器的 AUROC 得分为 0.5，且有效效用的取值范围为 0 到 0.5。

在集中式、基于服务器的联邦学习中，服务器会将特定数据点的夏普利值发送给相应的客户端。在去中心化群体学习场景中，客户端可以自行计算夏普利值。为了估计任意层级结构（即数据点群组）的贡献，我们对各个数据点的夏普利值进行汇总：$\varphi_{AUC}(D) = \sum_{z \in D} \varphi_{AUC}(z)$，以表示由数据点构成的群组 D 的贡献。其背后的直觉是，奖励通常按照夏普利值的比例进行分配，而奖励总和应与群组的贡献 [22] 成正比。例如，在实践中，这意味着可以评估患者群组或整个医院所做出的贡献。有关详细信息，请参阅从服务器视角出发的算法 2 和从客户端视角出发的算法3。此外，有关计算夏普利值的具体过程，请参阅背景章节中的算法 1。

DDVal 还确保不同客户端中的相同数据点获得相同的夏普利值（SV）。这是因为无论深度特征来自哪个客户端，数据估值都是针对所有深度特征集中计算的。对于相同的深度特征，该算法将得出相同的夏普利值，从而满足夏普利值的公平性条件。

3.3. 群体学习方法

对于分割学习（SL），我们建立了一个基于权威证明共识机制的私有以太坊网络。该区块链实现了客户端之间点对点通信的目标，客户端之间不一定需要完全互连。此外，私有区块链还能够

算法 2 从以下视角评估联邦学习数据的数据估值过程服务器。

记号：
$\Psi = \{\psi_0, \cdots, \psi_N\}$: 每轮次的全局模型
AUC：验证结果集合
K：K近邻方法中的 k 参数

当endingCondition**时**
  对于每个客户端 c**执行**
    $\nabla_c \leftarrow$本地训练(ψ)
  $\psi_{i+1} \leftarrow$ 聚合(∇, ψ_i)
  对于每个客户端**执行**
    $AUC \leftarrow$验证(ψ_{i+1})
  如果$min(AUC^{i-endingConditionPeriod:i}) < AUC_i$ **则**
    结束条件 $\leftarrow 1$
对于每个客户端 c**执行**
  F_{tr}, y_{tr}, F_{te}, $y_{te} \leftarrow$ 客户端。提取深度特征(ψ_{i+1})
$SVs \leftarrow$ 计算$SVs(K, F_{tr}, y_{tr}, F_{te}, y_{te})$

算法3 联邦学习的数据估值过程，视角为客户端 .

符号说明：
(X_{tr}, y_{tr}): 训练数据对　　(X_{te}, y_{tr}): 测试数据对
F_{tr}: 深度训练特征　　F_{te}：深度测试特征
ψ：全局模型　　∇：梯度

函数本地训练(ψ)
  $\nabla = \text{train}(X_{tr}, y_{tr}, \psi)$
  return ∇

函数验证(ψ)
  曲线下面积 $= \text{validate}(X_{te}, y_{te}, \psi)$
  return 曲线下面积

函数extractDeepFeatures(ψ)
  $F_{tr} = \psi(X_{tr})$
  $F_{te} = \psi(X_{te})$
  return $F_{tr}, y_{tr}, F_{te}, y_{te}$

控制谁可以访问网络，并充当存储交易的不可变账本。我们的实验使用 Go-Ethereum (Geth) 1.10.25 版。

对于这个私有区块链，我们使用 Solidity 编程语言开发了一个智能合约，用于协调分割学习过程及后续的数据估值。代码包含在我们的 GitHub 代码库中[1]。此外，我们搭建了一个私有 IPFS 网络，以便在客户端之间传输较大的文件（即数兆字节大小的模型梯度以及深层特征向量），因为以太坊等区块链网络无法在交易中存储如此大的文件。通过 IPFS，客户端可以在网络中发布文件，其他客户端则可以通过哈希值请求该客户端传输文件。事实证明，这种将私有以太坊网络与私有 IPFS 网络相结合的架构已成功用于分割学习任务 [21, 14]。对于分割学习智能合约，合约所有者（即向区块链提交合约的机构）会指定获准参与分割学习过程的白名单以太坊地址，并为初始机器学习模型指定一个哈希值。合约创建后，白名单中的机构可以在合约中表明已准备好开始分割学习过程。随后，机构通过智能合约在多个通信轮次中迭代交换模型梯度哈希值和模型测试结果。实际的模型和模型梯度则通过 IPFS 网络交换。达到结束条件后，智能合约便会终止分割学习过程；在我们的实现中，结束条件是最近 10 个通信轮次内测试结果没有改善。

3.4. 代币经济方法

我们的目标是通过代币经济扩展分割学习方法，激励机构以及与机构对应的个人（例如患者）为分割学习过程提供数据。转移具有货币价值的资产时，我们通常希望使用成熟、大型且值得信赖的网络。大型公有无许可区块链（例如以太坊公有网络）具有较高的交易完整性，并拥有活跃的代币生态系统，这些代币代表货币价值。出于研究目的，我们使用以太坊 Sepolia 测试网。以太坊公有网络托管着庞大的去中心化金融应用生态系统，这些应用运行在智能合约上，其中包括与美元等非加密货币挂钩的稳定币，因此其价值通常波动较小，尤其是与传统加密货币相比。在医疗保健等行业的智能合约应用中，这一点非常理想。

除了部署在私有区块链上的智能合约外，我们还会向公有区块链提交一份智能合约。首先，某个机构将编译后的智能合约提交到区块链，并为该智能合约注入 ERC-20 稳定币。之后，在私有网络上开展分割学习。我们通过 DDVal 扩展私有区块链上的流程，DDVal 在分割学习之后进行。随后，每个机构

[1] https://github.com/kpandl/Scalable-Data-Point-Valuation-in-Decentralized-Learning

都会进行数据估值，各机构也会获知其他机构的夏普利值——每个机构都应获得相同的计算结果。随后，各机构将这些夏普利值提交到公有区块链上的智能合约，由该智能合约向各机构支付奖励。任何剩余代币（即分割学习的结果未能达到 AUROC 为 1 的理想效果时）都将退还给存款人，因此存款人也很可能会收到退款。

随后，机构可以将奖励全部或部分转交给个人数据提供方，例如医疗保健场景中的患者；奖励金额与其数据点的夏普利值总和成正比。这些款项既可以通过在交易所出售代币并使用传统银行系统来支付，也可以借助可扩展性相对较强、交易成本可能低于以太坊的区块链（例如莱特币）支付。

4. 实验方法

为了评估 DDVal，我们围绕四个研究问题展开，从不同角度进行考察，具体如下。在评估中，我们将 NIH [30], CheXpert [31], MIMIC-CXR [32]这三个胸部X光数据集组合起来，用于机器学习任务。这些数据集广泛应用于医疗保健领域的机器学习研究，使我们能够模拟真实的非独立同分布数据分布场景。机器学习模型方面，我们采用在 ImageNet 数据集上预训练的 121 层 DenseNet 架构 [33], 。该架构在胸部X光数据集 [34], 以及集中式机器学习中的基于KNN的数据估值任务中均表现良好 [26]。由此，我们提取最后一个（即第四个）密集块中的深度特征。

通过我们的实验方法，我们旨在解决以下研究问题：

研究问题1。DDVal 对数据点进行估值的效果如何？有效的数据估值机制可以通过为数据点赋值，成功区分对机器学习任务有益的数据点和有害的数据点。如上所述，这些值可以采用夏普利值的形式，因为夏普利值具有理想的数据估值属性。

对于数据估值任务而言，计算标准 Shapley 值的计算复杂度过高，因为这类任务涉及数万个甚至更多数据点 [22, 23]。因此，我们无法将近似的夏普利值与标准 Shapley 值进行比较。不过，我们可以专注于开展另外两项实验，以便对夏普利值估计的质量得出结论。在第一项实验中，我们随机翻转数据点的标签。具体而言，我们根据客户端的不同，随机翻转训练数据点中 0 % 到 25 % 的标签。每个数据点有 8 个不同的标签，因此一些客户端存在多个标签被翻转的情况。根据集中式机器学习数据估值方面的研究，我们预计，标签被翻转的数据点的平均夏普利值会低于标签未被翻转的数据点 [23]。其次，我们旨在分析各个数据点的夏普利值与其标签之间的关系。胸部X光数据集等真实世界数据集通常具有多维标签。例如，患者的一张 CXR扫描可能显示多种医学状况。根据集中式机器学习实验所得的认识，我们预计存在医学状况的扫描图像，其夏普利值会高于不存在医学状况的扫描图像 [26]。

研究问题2。DDVal 对机构贡献的估值有多准确？

与数据点的数量相比，联邦学习中的机构数量通常要少得多（即跨机构联邦学习中有 2–100 家机构 [35]）。因此，我们可以计算标准 Shapley 值，并将我们的分层贡献指数与标准 Shapley 值进行比较。此外，现有研究提出了针对联邦学习中客户端的夏普利值估计方法，我们也可以将其纳入比较。

在准确度比较实验中，我们使用 3 个客户端，这意味着需要评估 7 个不同的联盟才能计算标准 Shapley 值，这在计算上是可行的。我们重复实验 12 次，每次实验中客户端之间的数据分布都取决于不同的随机种子。这样，我们便能获得具有统计显著性的结果，包括均值和置信区间。因此，我们使用余弦相似度这一常用指标来衡量夏普利值估计的准确度 [11, 12]。

由于非独立同分布的数据分布给成功实施联邦学习（FL）带来了挑战，尤其是在联邦学习中的数据估值方面，因此我们分析了独立同分布和非独立同分布两种情形。联邦学习（FL）和分割学习（SL）在医疗行业中尤其受到广泛讨论，因此，我们从三个不同的胸部X光数据集中抽取训练数据：NIH [30], CheXpert [31], MIMIC-CXR [32]。这些数据集具有不同的输入分布和标签分布。在独立同分布情形下，我们有3个客户端，共包含27,999张扫描图像，每个数据集抽取9,333张扫描图像。我们确保具有多张潜在扫描图像的患者只关联到一个客户端。在非独立同分布情形下，每个客户端同样拥有27,999张扫描图像，但每个客户端仅从其中一个数据集中抽取扫描图像。

此外，我们还分析了 RQ1 中同一标签翻转实验里的机构贡献。因此，我们预期，对于翻转标签占比更高的客户端，机构贡献会更小。

研究问题3：DDVal 的可扩展性如何？

为了分析不同夏普利值近似的可扩展性，我们针对不同规模的数据集进行了运行时间实验，以评估数据估值。通常，标准夏普利值计算开销随实体（即数据点或联邦学习参与方）数量呈指数增长。虽然近似方法效率更高，但实体数量增加时，通常也需要更多计算资源。因此，联邦学习客户端的夏普利值估计通常会随着客户端数量增加而变慢。正如前文所述，DDVal 中的数据估值方法的扩展依据是数据点数量，而非客户端数量。

在我们的场景中，理论上的复杂度为 $N \log N$，其中有 N 个数据点 [22]。因此，从复杂度的角度来看，这些数据点分布在少数还是众多客户端上并无影响。因此，我们预计 DDVal 特别适用于客户端数量众多、且每个客户端的数据点相对较少的去中心化学习场景。我们旨在通过实验验证理论复杂度，并将其与现有的、随客户端数量扩展的基于夏普利值的客户端估值方法进行比较 [11, 12]。为了将这些方法与 DDVal 进行比较，我们假设实验中每个客户端固定拥有 8,000 个数据点。

在实验中，我们考察数据点数量方面的可扩展性。例如，16,000 个数据点对应一项包含 2 个联邦学习客户端的实验，而 32,000 个数据点则对应一项包含 4 个联邦学习客户端的实验。可扩展性实验中的测试数据集包含 3,600 张扫描图像。

在实验中，我们测量联邦学习服务器执行数据估值所需的计算时间，但不包括客户端侧所需的时间（例如，SaFE 和 DDVal 中的深度特征提取）。这是因为客户端级计算通常很快，而且每个客户端只需对自己的数据进行计算，这些计算会由各个客户端并行完成。然而，服务器计算的复杂度可能会迅速增加，因为服务器需要评估所有客户端的贡献。在去中心化协作学习场景中，服务器的任务会分摊到所有客户端上，因此联邦学习服务器所做的工作也由客户端完成。

RQ4. 在医疗保健去中心化学习场景中，如何应用 DDVal，奖励提供数据的机构和患者?

这个问题聚焦于分布式系统方面。因此，我们使用一组服务器而非高性能计算集群来运行实验。在此过程中，我们依次执行分割学习、数据估值以及奖励的分配与发放。

分割学习描述了去中心化学习中的点对点通信概念，而非基于服务器的通信。分割学习的一种常见实现方式涉及私有区块链网络 [17]。服务器运行私有区块链节点和 IPFS 节点。它们还连接到一个公共区块链。 n 节点。在实验中，我们旨在部署智能合约，并评估端到端流程的运行时间和货币成本。

因此，系统中有两种用户角色。第一种是协调者，负责在公有区块链和私有区块链上发布用于代币经济和分割学习的智能合约，并为公有区块链上的智能合约提供资金。第二种是机构。机构也可以承担协调者的角色。

我们开发并实现了一种系统架构，将 DDVal 集成到分割学习和代币经济中。在整个过程中，我们测量协调者和参与机构产生的交易成本。此外，我们还测量所耗费的时间。这两项测量的目的是确定相关数值是否适用于实际应用。对于时间而言，如果协作式去中心化学习所产生的额外开销远短于实际模型训练过程所需的时间，我们便可以认为其具有实际可行性。对于交易成本，我们还可以将公有区块链的成本与大型医疗机构通常承担的成本进行比较。

5. 实验结果

5.1. 评估数据点价值的有效性

首先，我们聚焦于标签翻转实验，其中有 6 个客户端，每个客户端有 28,000 个数据点，每个数据点因随机因素而具有不同数量的翻转标签。查看表 1 中各个数据点及其夏普利值时，数据没有翻转标签的数据点的夏普利值最高，平均值为 $2.43 \cdot 10^{-6}$。数据点的标签翻转得越多，其夏普利值就越小。对于翻转标签达到 5 个或以上的数据点，平均SV为负值。这初步表明，夏普利值估计得出的结果是合理的。

表 1：一次包含翻转标签实验中所有客户端单个数据点的平均SV。

翻转标签的数量	夏普利值 $\cdot 10^6$
0	2.43 ± 1.22
1	1.88 ± 0.53
2	1.34 ± 0.95
3	0.74 ± 0.80
4	0.14 ± 0.31
5	-1.17 ± 0.64
6	-2.17 ± 1.59

接下来，我们关注没有发生任何标签翻转时，数据点的平均SV与其标签之间的关系。在实验设置中，有 3 个客户端，每个客户端各有 28k 个数据点。因此，这些夏普利值无法与前一次实验中的数值直接比较；由于训练所得全局模型的相当效用仅分配给一半的数据点，其数值通常更高。结果如表 2 所示。由此可见，当表示存在某项发现的标签设为 true，但 7 个医疗状况标签均未设为 true 时（即扫描图像中显示的是未定义的医疗状况），数据点的平均SV最低，为 $1.1 \cdot 10^{-6}$。平均SV次低的是不属于任何医疗状况的数据点。除这两种特殊情况外，扫描图像中存在的医疗状况越多，平均SV通常越高。因此，平均SV最高的是包含 4 种医疗状况的扫描图像，为 $1.729 \cdot 10^{-5}$。

表 2：多维数据实例的平均SV。

查找现有内容标签	医疗数据数量条件标签	SV
错误	0	$2.892 \cdot 10^{-6}$
True	0	$1.100 \cdot 10^{-6}$
True	1	$3.932 \cdot 10^{-6}$
True	2	$8.117 \cdot 10^{-6}$
True	3	$1.424 \cdot 10^{-5}$
True	4	$1.729 \cdot 10^{-5}$

5.2. 评估机构贡献的价值

独立同分布数据分布情形下的机构贡献估值结果如图 5.2所示。三个机构的标准 Shapley 值均值较为接近，介于 11.047 % 至 11.125 % 之间。DDVal 的夏普利值介于 10.984 % 至 11.105 % 之间，而 SaFE 的范围为 10.649 % 至 11.423 %，OR近似方法的范围为 10.371 % 至 11.213 %。标准 Shapley 值和 DDVal 的 95%置信区间范围相对较小，前者为 0.192 %，DDVal 为 0.240 %；SaFE 的范围为 1.64 %，OR近似方法为 3.503 %。

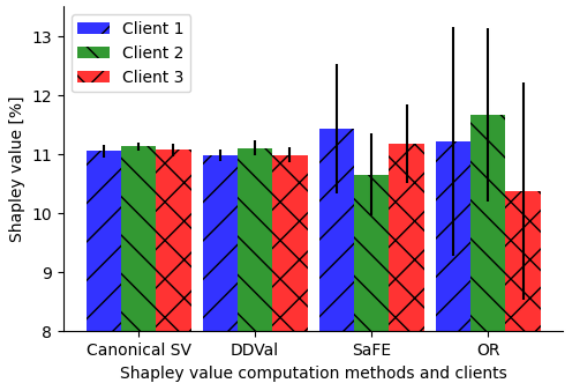


图1：独立同分布情形下客户端的平均SV、规范值及三种近似方法的95%置信区间。

采用非独立同分布数据分布进行的类似实验结果如图 5.2所示。对于采用同一近似方法的不同客户端，标准 Shapley 值的波动比独立同分布情形下更大。NIH 数据集的值最低，均值为 8.918 %；CheXpert 数据集的值最高，均值为 10.969 %。在 DDVal 中，NIH 的夏普利值最低，均值为 9.101 %；MIMIC 略高于 CheXpert，分别为 10.683 % 和 10.634 %。在另外两种近似方法中，CheXpert 客户端的夏普利值相对较低，SaFE 为 7.831 %，OR 为 6.246 %；而 NIH 客户端（7.831 % 和 6.246 %）和 MIMIC 客户端（11.682 % 和 11.890 %）的近似夏普利值均高于标准 Shapley 值和 DDVal 的估值。

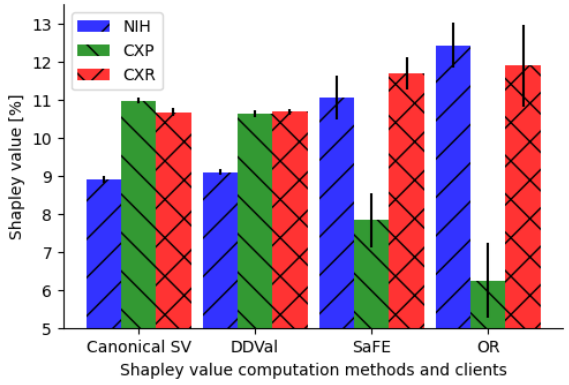


图2：非独立同分布情形下客户端的平均夏普利值（SV）及其95%置信区间，包括规范方法和3种近似方法。

为了更好地比较这 3 种近似方法与标准 Shapley 值之间的准确性，我们根据这些值计算了余弦相似度。表 3 展示了在 IID 和非独立同分布两种情形下，各近似方法与标准 Shapley 值之间的余弦相似度。值得注意的是，DDVal 的平均余弦相似度最高，达到 99.969%，并且在 IID 数据分布和非独立同分布数据分布下均表现出同样出色的性能。SaFE 在 IID 情形下的平均余弦相似度也相对较高，达到 99.301%，但受到非独立同分布数据分布的负面影响，其余弦相似度降至 97.250%。类似地，对于 OR 采用近似方法后，余弦相似度从 97.189 % 降至 93.753 %。

表 3：余弦相似度、均值和 95%置信区间。

SV近似方法	独立同分布 [%]	非独立同分布 [%]
DDVal	99.969 ± 0.006	99.969 ± 0.010
SaFE	99.301 ± 0.403	97.250 ± 0.906
OR近似	97.189 ± 0.756	93.753 ± 1.371

DDVal 在 IID 和非独立同分布数据分布下表现同样出色，原因在于所有站点的深度特征都会被收集并在一个站点进行估值。因此，数据划分的影响会减弱。相反，OR方法通过累积不同轮次的梯度来创建近似模型；当客户端的梯度差异很大时，这种方法通常会导致不准确的结果，因为在非独立同分布数据分布设置下，余弦相似度会降低。

作为 DDVal 机构估值部分的最后一部分，图 5.2 展示了第5.1节标签翻转实验的结果，并在机构层级进行分析。没有任何翻转标签的客户端平均 SV 最高，为 6.884 %。对于接下来 5 个翻转标签比例逐渐增加（每次增加 5%）的客户端，平均 SV 持续下降；翻转标签比例为 25 % 的客户端平均 SV 降至 3.548%。每增加 5% 的比例，SV 的绝对降幅都相对稳定，介于 0.58 至 0.713 个百分点之间。置信区间的绝对范围也相对稳定，介于 0.176% 至 0.232% 之间。因此，相对于平均 SV，SV 较低的客户端置信区间更大（例如，翻转标签比例为 25% 的客户端为 6.37%，而没有翻转标签的客户端为 3.14%）。

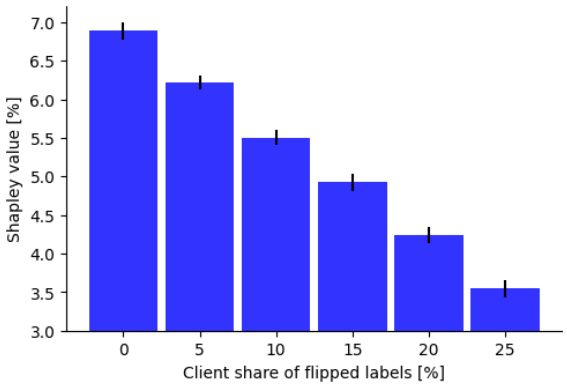


图3：标签翻转实验，按客户端划分的平均SV和95%置信区间。

5.3. 可扩展性

为了测量客户端级方法近似计算夏普利值所需的计算时间，我们假设每个客户端有 8,000 个数据点。根据定义，协作学习至少需要 2 个客户端。因此，我们测量了客户端

级方法 SaFE 和 OR 处理 16,000 个数据点所需的计算时间，随后每次增加 16,000 个数据点，相当于再增加两个客户端。

可扩展性实验的结果如图 5.3 所示。由于纵轴上的数值范围较大，我们将其设为对数刻度。

在图中，OR方法和 SaFE 的指数复杂度以及 DDVal 的对数线性复杂度显而易见。对于最多约 5 个客户端的范围，且每个客户端拥有 8,000 个数据点时，OR方法通常比 DDVal 为 40,000 个数据点估值更快。客户端数量更多时，DDVal 则更快。我们计算了 OR方法最多 8 个客户端的情况，计算耗时约 21 小时。相比之下，SaFe 方法具有相似的指数复杂度，但起始基线低得多。因此，在相同场景下，它仅耗费约 23 秒的计算时间，而 DDVal 则耗时约 2.5 小时。

DDVal 和 SaFE 在数据估值所需计算时间上的交叉点位于客户端数量超过 16 个或数据点超过 128,000 个时。在 SaFE 最全面的实验中，我们对 18 个联邦学习客户端的贡献进行了估值，这些客户端共拥有 144,000 个数据点，耗时约 18 小时。在 DDVal 最全面的实验中，我们使用了 20 个客户端，共计 160,000 个数据点，耗时不到 8 小时。

对于 10 个客户端，估值耗时 1.85 分钟。相比之下，DDVal 耗时 272 分钟。OR 估值在我们的集群上计算耗时过长，但在 8 个客户端的设置下，最终于 1255 分钟后完成。不过，SaFE 的复杂度呈指数增长，而 DDVal 的复杂度为对数线性。在数据点和客户端更多，或每个客户端的数据点少于 8,000 个的场景中，就计算开销而言，DDVal 可能比 SaFE 更有优势。客户端和数据点较少时，OR 近似比 DDVal 更快。对于总计 60,000 个数据点的 6 个客户端，基于 KNN-SV 的估值速度更快，耗时约 3.9 小时，而 OR 近似耗时 4.5 小时。从理论上看，OR 近似与 SaFE 一样具有指数复杂度，随着客户端数量增加，需要更多计算时间。

5.4. 集成到群体学习代币经济中

为将数据点估值整合到群体学习中，我们聚焦于群体学习的一种基于区块链的实现。为此，我们开发并部署了两个智能合约：一个用于在私有区块链上进行群体学习及后续通过 DDVal 开展的数据估值，另一个用于公有区块链上的代币经济。客户端与这两个合约交互，从而使两者的状态保持同步。我们在图5.4中展示了这一流程，并将在下文对其进行说明。

协调者在两个区块链上部署智能合约后

hain

私有和公有区块链中，协调者直接将合约地址提交给所有参与的机构服务器。随后，这些服务器加入智能合约。接着，协调者为公有智能合约注资，之后，机构开展分割学习和数据估值，并通过私有区块链和 IPFS 进行通信。在此过程中，我们测量通过的

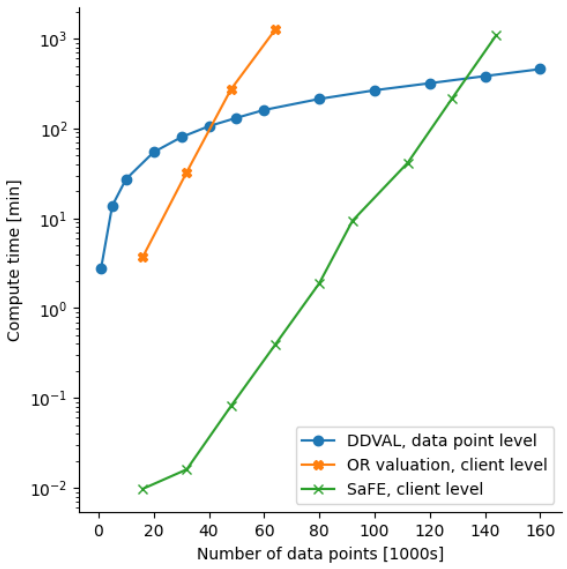


图4：可扩展性实验。对于客户端级近似，我们假设每个客户端的训练数据集固定为 8,000 张扫描图像。纵轴为对数刻度。

时间，这主要由同步等待时间（即等待所有机构上传新的梯度哈希值）和文件传输时间决定。完成分割学习和数据估值后，机构会在公有区块链上发布各机构贡献的量化结果。全部发布后，资金将进行分配。

通过使用私有区块链和 IPFS 网络，只有经授权的机构能够查看学习活动相关信息，包括元数据（即梯度哈希、学习轮次）和机器学习模型数据。在公有区块链上，公众可以查看测试结果和资金流向。这有助于增强个人的信任，例如，当机构承诺将收益转交给数据提供方时，如医院承诺将收益转交给患者。同时，这也能让机器学习模型本身等更敏感的数据免于公开。公有智能合约本身包含 83 行代码，因此可以接受审计，从而增强各实体对它的信任。

在实际实施时，以太坊公有区块链上的交易成本以 Gas 成本衡量。因此，我们按用户角色对交易进行划分。协调者首先创建智能合约，然后为其注资，之后发起付款。机构只负责报告贡献指数。到目前为止，Gas 消耗最高的是合约创建，数值相对较大，达到 1,473,028。撰写本文时，即2023年4月7日，以太坊主网上的 Gas 价格为 33 Gwei，其中，1 Gwei 等于 10^{-9} 以太币，且以太币价格为 1859 美元。以太币是以太坊区块链的主要加密货币，也是其核心组成部分。在以太坊公有区块链上，为 3 家机构部署代币经济智能合约的总成本约为 90.37 美元。增加机构数量时，成本线性增长，每增加一家机构，消耗 34,399 gas，成本约为 1.50 美元。其他交易成本为

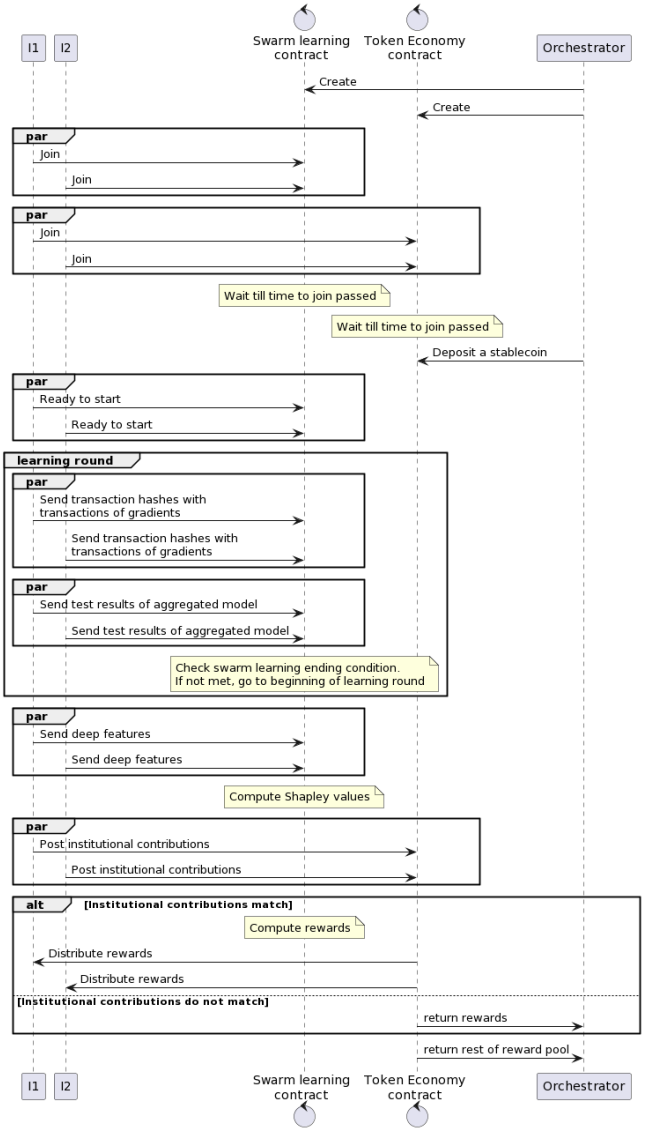


图5：SL智能合约与代币经济智能合约的时序图，其中协调者与两个示例机构（I1 和 I2）进行交互。

低得多，在此情况下，协调者为合约提供资金的成本为 3.20 美元，合约支付金额为 7.24 美元。对于每个机构，将夏普利值提交至智能合约的成本为 7.17 美元。

因此，通信模型、梯度和深度特征的流程相对较快，大约在 5 分钟内即可完成。这比机器学习模型本身的训练流程快得多，后者可能需要数小时。因此，网络开销可以认为很小。

6. 讨论

6.1. 隐私方面

使用联邦学习（FL）和分割学习（SL）的核心理念，是在保护数据所有者的信息隐私的同时，实现跨数据孤岛的机器学习。我们通过数据估值方法扩展分割学习的概念，并进一步深入探讨其隐私影响。在客户端之间共享深度特征，可能会增加成员推理攻击的风险，也就是判断某个数据点（即一张医学图像）是否参与了训练过程的攻击。不过，在医疗保健场景中，成员推理攻击或许并非重大隐患，因为两家机构需要拥有同一位患者的相同或极其相似的数据点，而攻击者最多只能得知该患者曾同时到访这两家机构。

更值得担忧的是重建攻击，在这种攻击中，SL客户端可能会从数据中重建训练数据点（例如医学图像）。由于在使用训练数据点进行推理时，仅共享神经网络的最后一层输出层，因此这里很难实施此类重建攻击。当前攻击需要访问更多层 [36]。

不过，仍可能存在针对这些假设性攻击场景的额外保护和保障措施。例如，可以采用与训练机器学习模型时使用差分隐私类似的方案来实现差分隐私（DP） [37]。具体而言，客户端向服务器提交每个训练数据点最后一层输出层的向量之前，可以向这些向量添加高斯噪声。噪声的标准差可以取决于每个向量维度的标准差，以及用于控制隐私保护级别的缩放因子。噪声的均值为 0。此外，还可以实施裁剪。我们将这一实现留待未来研究。未来研究的一个有趣问题是差分隐私对夏普利值估计准确性的影响，以及它如何影响夏普利值（例如，噪声越大，夏普利值越趋于均匀）。

6.2. 主要发现

通过本研究，我们提出了去中心化学习中的数据点价值评估概念，并为此开发了一种名为 DDVal 的技术。因此，该技术适用于联邦学习（FL）和分割学习（SL）。我们通过实验证明，该技术能够准确地评估数据价值：在三个胸部X光数据集上识别错误标记数据，并评估多维度标记数据点的价值。由此，我们发现，与包含较少医学状况或不包含任何医学状况的扫描图像相比，同时呈现多种医疗状况的扫描图像通常具有更高的夏普利值（SV）。

我们还发现，DDVal 有助于对贡献进行层级估值，例如在个人或机构层面。因此，实证结果表明，DDVal 在非独立同分布数据分布场景下也尤其有效，并且在机构层面的数据估值方面优于现有方法。

此外，我们发现，DDVal 具有有趣的可扩展性特性，因为它在去中心化学习中的扩展取决于数据点的数量，而非客户端的数量。因此，它尤其适用于拥有众多协作客户端、且每个客户端仅拥有有限数据量的去中心化学习场景。与现有在大规模联邦学习联盟中具有指数级复杂度的方法相比，DDVal 的对数线性复杂度具有显著优势。

此外，我们发现 DDVal 能够很好地集成到基于区块链的分割学习架构中。为此，一个 IPFS 网络和一个智能

私有区块链上的合约不仅可以控制分割学习过程，还可以控制后续的数据估值。此外，另一个智能合约可以管理代币经济，并根据各机构的贡献按比例向其支付资金。我们发现，以去中心化的方式进行数据估值会带来合理的时间开销，而在公有区块链上运行智能合约的成本也处于合理范围内，这表明了该方法的实用性。

6.3. 对研究与实践的启示

我们的研究为学术研究和实践提供了多方面的启示。在学术研究方面，我们提出了联邦学习（FL）和分割学习（SL）中的数据点估值概念，并展示了其在奖励为机器学习（ML）任务提供数据的个人以及识别错误标记数据点方面的实用应用。因此，去中心化机器学习中的数据点估值有望成为一个更加活跃的研究领域。为此，我们证明了一种基于 K 近邻（KNN）的夏普利值（SV）近似方法能够成功地对数据点进行估值，同时也有助于得出有关机构贡献值的结论。由此，我们为协作式机器学习中的数据估值开辟了新途径，而传统上，这类估值主要依赖于可信执行环境中的数据汇集等技术 [5]。此外，我们展示了数据点估值在医疗保健领域联邦学习中的应用前景，例如，奖励提供数据的机构和个人，以及检测可能存在的错误标注的数据实例。此外，我们强调了数据估值技术能够应对去中心化学习中非独立同分布数据分布的重要性。此外，我们还表明，去中心化系统架构研究不仅可以纳入协作学习流程，还可以进一步涵盖后续的数据估值和奖励分配。因此，要实现最优的系统设计，就需要对机器学习方面和分布式系统方面进行全面、整体的考量。

在实践方面，我们的结果表明，新颖的数据估值技术已可应用于各种场景，例如奖励医疗保健领域提供数据的患者和机构、检测优质数据源，以及识别错误标记数据。此外，我们展示了，通过引入数据估值和奖励系统对SL进行扩展，有助于在实践中实现Web3愿景，奖励信息系统用户提供数据。

6.4. 局限性与未来研究

我们的研究存在一些局限性，也为未来研究提供了可能性。首先，尽管我们使用了三个不同的数据集进行评估，但这三个数据集都是胸部X光数据集。因此，未来的研究可以重点评估更多医学和非医学数据集。其次，与经典的联邦学习和分割学习方法相比，DDVal 依赖于共享更多信息，这可能会带来隐私影响。我们已在第 6.1 节中对此进行了讨论。虽然我们认为，对于某些使用场景而言，这些隐私影响是可以接受的，但未来的研究可以通过差分隐私或安全多方计算等技术，致力于改善隐私保护，或提供隐私保证。对于差分隐私而言，观察其是否会降低夏普利值估计的准确性，以及夏普利值是否趋于一致。第三，组织和用户如何看待 DDVal，以及其实际应用价值究竟是什么，目前仍有待研究。因此，未来的研究可以致力于将我们的 DDVal 应用于实践，并研究数据估值及相关概念在现实世界中的实际采用情况。

7. 结论

在本研究中，我们开发了一种新颖的方法DDVal，通过共享深度特征近似夏普利值，从而能够对联邦学习和分割学习中的数据点进行估值。除数据点估值外，DDVal还可以帮助估计机构的夏普利值贡献。在我们的实证评估中，对于独立同分布和非独立同分布数据分布，该方法的表现同样出色，余弦相似度约为99.969%。此外，它具有理想的可扩展性，其复杂度相对于数据点数量呈对数线性增长；对于每个客户端拥有少量数据点的众多客户端场景，这一特性尤其具有吸引力，例如在我们的可扩展性实验中，16个客户端各拥有8,000个数据点。我们将DDVal集成到一个去中心化系统中，通过引入数据点估值和代币经济扩展了分割学习的概念，并展示了具有应用前景的场景。例如，在医疗保健环境中，DDVal可以奖励机构，并使机构能够奖励患者为机器学习任务共享数据。此外，它还可以帮助检测可能的错误标记数据。最终，DDVal有助于将Web3概念的愿景付诸实践，使用户因向机器学习系统共享数据而获得奖励。

致谢

我们感谢 Mathias L´ecuyer 就联邦学习（FL）中的分层数据估值方法及隐私方面的问题所开展的宝贵讨论。在计算资源方面，作者感谢巴登-符腾堡州通过 bwHPC 提供的支持，以及德国研究基金会（DFG）通过资助项目 INST 35/1597-1 FUGG 提供的支持。

参考文献

[1] James Zou、Mikael Huss、Abubakar Abid、Pejman Mohammadi、Ali Torkamani 和 Amalio Telenti。基因组学深度学习入门。《自然》遗传学，51(1):12–18，2019年。[2] Tim R¨adsch、Sven Eckhardt、Florian Leiser、Konstantin D Pandl、Scott Thiebes 和 Ali Sunyaev。放射科医生可能会遗漏什么：利用机器学习识别X射线图像中的错误标记实例。2021年。[3] Angela Ballantyne。我们应如何看待临床数据所有权？*Journal of Medical Ethics*，46(5):289–294，2020年。[4] Konstantin D Pandl、Florian Leiser、Scott Thiebes 和 Ali Sunyaev。可信医疗联邦学习的奖励机制。*arXiv*预印本 *arXiv:2205.00470*，2022年。[5] Nick Hynes、David Dao、David Yan、Raymond Cheng 和 Dawn Song。Sterling隐私保护数据市场演示。*VLDB*基金会学报，11(12):2086–2089，2018年。[6] Ali Sunyaev、Niclas Kannengießer、Roman Beck、Horst Treiblmaier、Mary Lacity、Johann Kranz、Gilbert Fridgen、Ulli Spankowski 和 Andr´e Luckow。代币经济。商业 & 信息系统工程，63(4):457–478，2021年。

[7] Se Young Jung、Taehyun Kim、Hyung Ju Hwang 和 Kyungpyo Hong。医疗保健区块链系统代币经济的机制设计：基于模拟真实场景的开发研究。《医学互联网研究杂志》，23(9):e26802，2021年。[8] Maria Diaz。Stack Overflow效仿Reddit和Twitter，向人工智能公司收取训练数据费用，2023年4月。[9] Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。基于去中心化数据的深度网络通信高效学习。载于人工智能与统计学，第1273–1282页。PMLR，2017年。[10] 慧与科技。群体学习：将分布式数据转化为竞争优势。[11] Tianshu Song、Yongxin Tong 和 Shuyue Wei。联邦学习中的利润分配。载于*2019*年*IEEE*国际大数据会议（*Big Data*），第2577–2586页。IEEE，2019年。[12] Sourav Kumar、A Lakshminarayanan、Ken Chang、Feri Guretno、Ivan Ho Mien、Jayashree Kalpathy-Cramer、Pavitra Krishnaswamy 和 Praveer Singh。通过集成学习提升医疗保健联邦学习中的数据估值效率。载于分布式、协作与联邦学习国际研讨会，面向ff资源多样的全球医疗保健的可负担医疗保健与人工智能研讨会，第119–129页。Springer，2022年。[13] Leon Witt、Mathis Heyer、Kentaroh Toyoda、Wojciech Samek 和 Dan Li。去中心化且具有激励机制的联邦学习框架：系统性文献综述。*arXiv*预印本 *arXiv:2205.07855*，2022年。[14] Konstantin D Pandl、Scott Thiebes、Manuel Schmidt-Kraepelin 和 Ali Sunyaev。人工智能与分布式账本技术的融合：范围综述与未来研究议程。《*IEEE Access*》，8:57075–57095，2020年。[15] Andrew Hard、Kanishka Rao、Rajiv Mathews、Swaroop Ramaswamy、Franc¸oise Beaufays、Sean Augenstein、Hubert Eichner、Chlo´e Kiddon 和 Daniel Ramage。用于移动键盘预测的联邦学习。*arXiv*预印本 *arXiv:1811.03604*，2018年。[16] Jie Xu、Benjamin S Glicksberg、Chang Su、Peter Walker、Jiang Bian 和 Fei Wang。医疗保健信息学中的联邦学习。《医疗保健信息学研究杂志》，5:1–19，2021年。[17] Stefanie Warnat-Herresthal、Hartmut Schultze、Krishnaprasad Lingada-halli Shastry、Sathyanarayanan Manamohan、Saikat Mukherjee、Vishesh Garg、Ravi Sarveswara、Kristian H¨andler、Peter Pickkers、N Ahmad Aziz等。用于去中心化且保密的临床机器学习的群体学习。《自然》，594(7862):265–270，2021年。[18] Niclas Kannengießer、Sebastian Lins、Tobias Dehling 和 Ali Sunyaev。分布式账本技术特征之间的权衡。《*ACM*计算调查》（*CSUR*），53(2):1–37，2020年。[19] Di Fan、Yifan Wu 和 Xiaoxiao Li。群体学习在皮肤病变分类中的公平性。载于临床图像处理方法、分布式与协作学习、用于抗击*COVID-19*的人工智能及安全与隐私保护机器学习：*MICCAI 2021*同期举办，法国斯特拉斯堡，*2021*年*9*月*27*日至*10*月*1*日，论文集第*2*卷，第120–129页。Springer，2021年。[20] Yang Zhao、Jun Zhao、Linshan Jiang、Rui Tan、Dusit Niyato、Zengxi-ang Li、Lingjuan Lyu 和 Yingbo Liu。面向物联网设备的隐私保护区块链联邦学习。《*IEEE*物联网杂志》，8(3):1817–1829，2020年。[21] Juan Benet。IPFS——内容寻址、版本化的点对点文件系统。*arXiv*预印本 *arXiv:1407.3561*，2014年。[22] Ruoxi Jia、David Dao、Boxin Wang、Frances Ann Hubis、Nezihe Merve Gurel、Bo Li、Ce Zhang、Costas J Spanos 和 Dawn Song。面向最近邻算法的高效任务特定数据估值。*arXiv*预印本 *arXiv:1908.08619*，2019年。[23] Amirata Ghorbani 和 James Zou。数据Shapley值：机器学习数据的公平估值。载于国际机器学习大会，第2242–2251页。PMLR，2019年。[24] Zelei Liu、Yuanyuan Chen、Han Yu、Yang Liu 和 Lizhen Cui。GTG-Shapley：联邦学习中高效且准确的参与方贡献评估。《*ACM*智能系统与技术汇刊》（*TIST*），13(4):1–21，2022年。[25] Siyi Tang、Amirata Ghorbani、Rikiya Yamashita、Sameer Rehman、Jared A Dunnmon、James Zou 和 Daniel L Rubin。使用夏普利值进行医学影像数据估值及其在大规模胸部X光数据集中的应用。《科学报告》，11(1):1–9，2021年。[26] Konstantin D Pandl、Fabian Feiland、Scott Thiebes 和 Ali Sunyaev。《面向医疗保健的可信机器学习：基于夏普利值的可扩展数据估值》。载于健康、推理与学习会议论文集，第47–57页，2021年。[27] Amani Moin、Kevin Sekniqi 和 Emin Gun Sirer。《稳定币设计的分类框架：现状综述》。载于金融密码学与数据安全：第*24*届国际会议（*FC 2020*），亚庇，马来西亚，*2020*年*2*月*10*日至*14*日修订精选论文集*24*，第174–197页。Springer，2020年。[28] Shashi Raj Pandey、Lam Duc Nguyen 和 Petar Popovski。《Fedtoken：联邦学习中的数据贡献代币化激励》。*arXiv*预印本 *arXiv:2209.09775*，2022年。[29] Yuan Liu、Zhengpeng Ai、Shuai Sun、Shuangfeng Zhang、Zelei Liu 和 Han Yu。《FedCoin：用于联邦学习的点对点支付系统》。载于《联邦学习：隐私与激励》，第125–138页。Springer，2020年。[30] Xiaosong Wang、Yifan Peng、Le Lu、Zhiyong Lu、Mohammadhadi Bagheri 和 Ronald M Summers。《Chestx-ray8：医院规模的胸部X光数据库，以及常见胸部疾病弱监督分类与定位的基准测试》。载于*IEEE*计算机视觉与模式识别会议论文集，第2097–2106页，2017年。[31] Jeremy Irvin、Pranav Rajpurkar、Michael Ko、Yifan Yu、Silviana Ciurea-Ilcus、Chris Chute、Henrik Marklund、Behzad Haghgoo、Robyn Ball、Katie Shpanskaya 等。《CheXpert：一个包含不确定性标签并用于专家比较的大型胸部X线影像数据集》。载于*AAAI*人工智能会议论文集，第33卷，第590–597页，2019年。[32] Alistair EW Johnson、Tom J Pollard、Seth J Berkowitz、Nathaniel R Greenbaum、Matthew P Lungren、Chih-ying Deng、Roger G Mark 和 Steven Horng。《MIMIC-CXR：一个经过去标识化、公开可用且包含自由文本报告的胸部X线影像数据库》。《*Scientific Data*》，6(1):317，2019年。[33] Gao Huang、Zhuang Liu、Laurens Van Der Maaten 和 Kilian Q Weinberger。《密集连接卷积网络》。载于*IEEE*计算机视觉与模式识别会议论文集，第4700–4708页，2017年。[34] Alexander Ke、William Ellsworth、Oishi Banerjee、Andrew Y Ng 和 Pranav Rajpurkar。《CheXtransfer：ImageNet模型在胸部X光解读中的性能与参数效率》。载于健康、推理与学习会议论文集，第116–124页，2021年。[35] Peter Kairouz、H Brendan McMahan、Brendan Avent、Aur´elien Bellet、Mehdi Bennis、Arjun Nitin Bhagoji、Kallista Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。《联邦学习的发展与开放问题》。《基础与趋势》® 机器学习，14(1–2):1–210，2021年。[36] Borja Balle、Giovanni Cherubin 和 Jamie Hayes。《利用知情攻击者重建训练数据》。载于*2022*年*IEEE*安全与隐私研讨会（*SP*），第1138–1156页。IEEE，2022年。[37] Martin Abadi、Andy Chu、Ian Goodfellow、H Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。《差分隐私深度学习》。载于*2016*年*ACM SIGSAC*计算机与通信安全会议论文集，第308–318页，2016年。