说 明 书

基于联邦学习的数据价值衡量机制

技术领域

[0001] 本发明涉及的是一种公平高效的数据价值衡量的方法，具体是一种基于联邦学习（Federated Learning）和夏普利值（Shapley Value）的卖方数据衡量机制，在卖方数据不暴露的前提下，买方能够衡量卖方数据对于其模型的价值，同时提出了高效准确的估计方法。

背景技术

[0002] 随着数据的总量不断扩大，价值不断提升，越来越多的公司依靠数据驱动型模型来做各种各样的商业决策。数据是这些模型的原料，大规模的训练数据能够让模型精度更高。为了支持数据资产的交换和交易，许多互联网平台涌现出来，例如 AWS，Dawex，WorldQuant。

[0003] 许多工作证明了不同训练数据对于模型的重要程度和价值也是不同的，我们能够通过挑选好的训练数据样本、去除不好的训练数据样本来提升模型的效果。在数据交换中，一个基本问题就是如何衡量数据对于模型的价值，从而在训练过程中提升模型训练效果。

[0004] 有许多工作提出了衡量数据价值的方法，例如 LOO（Leave-one-out）方法，基于影响函数（Influence Function）的方法，数据夏普利值（Data Shapley）方法。但是这些方法都需要接触到数据的详细信息，例如数据样本、数据分布等。但是这在真实的数据市场中是不可能的，因为信息的不对称性（卖方不希望泄露数据的详细信息直到买方付钱购买数据，但买方想在购买之前直到卖方数据对模型的贡献来最大化自己的礼仪）。

[0005] 随着人们对个人信息的重视程度和保护意识越来越强，联邦学习（Federated Learning）作为一种分布式、去中心化的机器学习方法吸引了许多研究者的注意。它允许许多客户联合地训练一个全球模型，同时保证每个客户的数据隐私不会泄露。

发明内容

[0006] 本发明针对现有技术存在的上述不足，提出一种基于联邦学习和夏普利值的数据价值定义及其高效准确的估计方案。

[0007] 本发明定义了每个卖方数据的联合夏普利值（Fed-Shapley）作为其数据价值：$\bar{\phi}_t(c_k) = \sum_{S\subseteq C\setminus c_k} \frac{\bar{w}_t(S\cup c_k)-\bar{w}_t(S)}{\binom{n-1}{|S|}} = E_{S\subseteq C\setminus c_k}[\bar{w}_t(S\cup c_k)-\bar{w}_t(S)]$，$\bar{\phi}_t(c_k)$为卖家$c_k$在第t轮的联合夏普利值；C为所有卖家的集合；$\bar{w}_t(S)$为只有卖家子集S参与到联邦学习训练过程时，全局模型在第t轮的参数。本发明证明了当联邦学习过程满足两个条件时，联合夏普利值和集中式学习得到的数据夏普利值没有差异，条件为：所有客户参与到训练过程；每一轮训练中每个客户只在本地训练

一次。

[0008] 本发明所述的估计方案为：联合夏普利值可以表示为$\bar{\phi}_t(c_k) = E_{S\subseteq C\setminus c_k}[\bar{w}_t(S\cup c_k) - \bar{w}_t(S)] = E_{Q\subseteq C\setminus c_k}[(\bar{w}_t(C\setminus Q) - \bar{w}_t(C)) - (\bar{w}_t(C\setminus\{Q, c_k\}) - \bar{w}_t(C))] = E_{Q\subseteq C\setminus c_k}(\epsilon_t^{-Q,*} - \epsilon_t^{-Q-c_k,*})$，$\epsilon_t^{-Q,*}$表示在训练过程中从总卖家集合C移除卖家子集Q后，模型在第t轮的参数变化。其值可以通过本发明的估计方法得到：$\epsilon_t^{-Q,*} \approx \epsilon_t^{-Q} = \sum_{k\subseteq C\setminus Q}\frac{n_k}{N(C\setminus Q)}\left[I - \eta\nabla_w^2 L\left(w_{t-1}^k(C_t), D_k\right)\right]\epsilon_{t-1}^{-Q} + \bar{w}_t(\mathrm{C_t}\rightarrow \mathrm{C_t}\setminus \mathrm{Q}) - \bar{w}_t(C)$，$\mathrm{n_k}$为第k个卖家的数据集大小；$N(C\setminus Q)$为卖家子集$\mathrm{C}\setminus\mathrm{Q}$的总数据集大小；η为学习率；$\mathrm{L}\left(\mathrm{w_{t-1}^k(C_t)}, \mathrm{D_k}\right)$表示当模型参数为$\mathrm{w_{t-1}^k(C_t)}$时，模型在数据集$\mathrm{D_k}$上的损失函数；$\bar{w}_t(\mathrm{C_t}\rightarrow \mathrm{C_t}\setminus \mathrm{Q})$表示只在第t轮将卖家数据集Q移除后全局模型的参数。同时，将传统的 Monte-Carlo 采样方法和本发明的估计方法结合，可以得到时间复杂度更低的估计方法。

[0009] 所述的蒙特卡洛采样是指：随机采样包含所有卖家的一个排列，按照顺序计算每一个排列当中每个卖家数据集对于之前所有数据集的边际贡献，采样多次求取平均值即为每个卖家的估计价值。

[0010] 所述边际贡献是指：加入此卖家数据集后全局模型参数的变化。

[0011] 通过联邦学习，本发明可以在不泄露卖方数据隐私的前提下，衡量卖方数据对于买方模型的价值，解决了数据市场中信息不对称的问题。

[0012] 本发明的数据衡量方法是基于博弈论的经典概念夏普利值（Shapley Value），具有与之类似的三条公平性定理：如果卖家$\mathrm{c_k}$的数据集对于模型性能没有影响，则其价值为 0；如果对于两个卖家$\mathrm{c_i}, \mathrm{c_j}$，将其数据集分别添加到任意子集$\mathrm{S}\subseteq \mathrm{C}\setminus \mathrm{c_i}, \mathrm{c_j}$后模型性能相同，则c_i和$\mathrm{c_j}$具有相同的价值；任意多种评估方法得到的数据集价值等于这些评估方法结合在一起得到的数据集价值。同时，为了解决夏普利值高时间复杂度的问题，本发明提出了高效准确的数据价值估计方法，在理论上具有有限误差上界，并在实验中得到验证。

[0013] 当模型损失函数为凸函数时，理论上本发明对联合夏普利值的估计误差上界与训练轮数t有线性关系；当模型损失函数为非凸函数时理论上本发明对联合夏普利值的估计误差上界与训练轮数有指数关系。

[0014] 本发明的估计误差还来源于卖方数据集的分布差异和方差。对于后者，本发明提出了减小误差的方法：在估计联合夏普利值时，不考虑当移除的客户子集规模过大的情况。

[0015] 本发明是通过以下技术方案实现的，本方法包括以下步骤：

[0016] 步骤 1、每个卖家k产生一个小型的无偏的数据集$\mathrm{d_k}$，产生方法为根据数据分布$\mathrm{p_k}$直接产生或者从原数据集$\mathrm{D_k}$中随机选取一小部分数据样本。

[0017] 步骤 2、对于每个卖家及其产生的小型数据集，买家执行一个专门设计的联邦学习过程：

在训练过程中的每一轮，每个卖家只执行一次本地参数更新并且所有卖家都参与到模型训练当中；在每一轮中，买家不仅与传统的联邦学习一样分发和收集模型，而且通过本发明的估计方法衡量出每个卖家数据的联合夏普利值$\bar{\phi}_t(c_k)$。

[0018] 步骤 3、当全局模型收敛后，买家得到了每个卖方的联合夏普利值，选择性价比最高的数据集购买。

技术效果

[0019] 与现有技术相比，本发明优点包括：解决了真实数据市场中信息不对称的问题，使得买家可以在不接触到卖家数据的前提下衡量各个数据集对自己模型的贡献；只需要训练一次即可估计出各个数据集的价值，解决了已有数据衡量方法需要重复训练的问题；证明了当联合学习的过程满足一定条件时，其全局模型参数和各个客户数据集的联合夏普利值等于集中式学习下的模型参数和各个数据集的数据夏普利值。

[0020] 本发明提出的高效的联合夏普利值估计方法经试验证明有着较小的估计误差，同时改进方法能够有效减小有较大数据方差的卖家的估计误差。

附图说明

[0021] 图 1 为当移除不同数目的卖方数据集后，本模型对全局模型参数变化的估计误差随训练轮数的变化。

[0022] 图 2 为当模型损失函数为凸函数时，且当卖方数据集分布相同且方差都较小、分布不同但方差都较小、分布不同且方差较大时，联合夏普利值的估计误差随训练轮数变化的关系。

[0023] 图 3 为应用本发明针对方差较大的改进方法后，联合夏普利值的估计误差随训练轮数变化的关系。

[0024] 图 4 为应用蒙特卡洛采样方法后，本发明对联合夏普利值估计的误差随训练轮数的变化关系以及应用改进方法后误差的变化。

[0025] 图 5 为当模型损失函数为非凸函数时联合夏普利值的估计误差随训练轮数的变化关系，其中：a 为卖家数据集独立同分布时的情况，b 为卖家数据集不独立同分布时的情况。

[0026] 图 6 为联合夏普利值的应用展示，当删除值较大、较小的数据集时，模型的性能变化

[0027] 图 7 为所用数据集的说明。

具体实施方式

[0028] 下面对本发明的实施例作详细说明，本实施例在以本发明技术方案为前提下进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

实施例 1

[0029] 本实施例包括 8 个卖方，其相关信息如图 7 所示，实施步骤如下所示：

[0030] 步骤 1、每个卖家k产生一个小型的无偏的数据集d_k，产生方法为根据数据分布p_k直接产生或者从原数据集D_k中随机选取一小部分数据样本。

[0031] 步骤 2、对于每个卖家及其产生的小型数据集，买家执行一个专门设计的联邦学习过程：在训练过程中的每一轮，每个卖家只执行一次本地参数更新并且所有卖家都参与到模型训练当中；在每一轮中，买家不仅与传统的联邦学习一样分发和收集模型，而且通过本发明的估计方法衡量出每个卖家数据的联合夏普利值$\bar{\phi}_t(c_k)$。

[0032] 所述联邦学习训练过程的优化目标为$\min_{\bar{w}}\{L(\bar{w},D)=\sum_{k\in C}\frac{n_k}{N(C)}L(\bar{w},d_k)\}$，其中$L(\bar{w},D)$为模型$\bar{w}$在数据集D上的损失函数，$n_k$为卖方k的数据集大小，$N(C)$为总卖方数据集C的大小，$d_k$为卖方k产生的小数据集。每一轮训练都包括两个过程：本地训练和模型聚合。

[0033] 所述本地训练过程为：在第t轮中，每个卖家$k\in C$从中心服务器下载第$t-1$的全局模型，表示未来$\bar{w}_{t-1}$。然后，每个卖家k执行m 次本地参数更新，第i次的参数更新为：$\bar{w}_{t,i}^k\leftarrow\bar{w}_{t,i-1}^k-\eta\nabla_{\bar{w}}L(\bar{w}_{t,i-1}^k,d_k)$，其中η表示学习率。

[0034] 所述模型聚合过程为：在第t轮的训练中，中心服务器随机选取一个卖方子集$C_t\subseteq C$作为参与者。经过m次本地参数更新后，每个卖家上传本地的模型$\bar{w}_{t,m}^k$。中心服务器将收集上来的模型聚合为新的全局模型$\bar{w}_t$，许多聚合算法被提出，如FedAvg，SCAFFOLD，FedBoost，FedNova，Fetchsgd 等。

[0035] 步骤 3、当全局模型收敛后，买家得到了每个卖方的联合夏普利值，选择性价比最高的数据集购买。

[0036] 所述的估计方案为：联合夏普利值可以表示为$\bar{\phi}_t(c_k)=E_{S\subseteq C\setminus c_k}[\bar{w}_t(S\cup c_k)-\bar{w}_t(S)]=E_{Q\subseteq C\setminus c_k}[(\bar{w}_t(C\setminus Q)-\bar{w}_t(C))-(\bar{w}_t(C\setminus\{Q,c_k\})-\bar{w}_t(C))]=E_{Q\subseteq C\setminus c_k}(\epsilon_t^{-Q,*}-\epsilon_t^{-Q-c_k,*})$，$\epsilon_t^{-Q,*}$表示在训练过程中从总卖家集合C移除卖家子集Q后，模型在第t轮的参数变化。其值可以通过本发明的估计方法得到：$\epsilon_t^{-Q,*}\approx\epsilon_t^{-Q}=\sum_{k\subseteq C\setminus Q}\frac{n_k}{N(C\setminus Q)}\left[I-\eta\nabla_w^2L\left(w_{t-1}^k(C_t),d_k\right)\right]\epsilon_{t-1}^{-Q}+\bar{w}_t(C_t\to C_t\setminus Q)-\bar{w}_t(C)$，$n_k$为第k个卖家的数据集大小；$N(C\setminus Q)$为卖家子集$C\setminus Q$的总数据集大小；η为学习率；$L\left(w_{t-1}^k(C_t),D_k\right)$表示当模型参数为$w_{t-1}^k(C_t)$时，模型在数据集$D_k$上的损失函数；$\bar{w}_t(C_t\to C_t\setminus Q)$表示只在第t轮将卖家数据集Q移除后全局模型的参数。同时，将传统的 Monte-Carlo 采样方法和本发明的估计方法结合，可以得到时间复杂度更低的估计方法。

[0037] 所述的蒙特卡洛采样是指：随机采样包含所有卖家的一个排列，按照顺序计算每一个排列当中每个卖家数据集对于之前所有数据集的边际贡献，采样多次求取平均值即为每个卖家的估计价值。所述边际贡献是指：加入此卖家数据集后全局模型参数的变化。

模拟实验结果

[0038] 图 7 展示了实验部分所涉及数据集和训练模型的相关信息。

[0039] 图 1 展示了当损失函数为凸函数时，本发明对移除不同卖家子集Q后模型参数变化ϵ_t^{-Q}的估计误差随训练轮数的变化关系。它证明了我们的理论分析：当损失函数为凸函数时，本发明对模型参数变化的估计误差上界与训练轮数t有线性关系。

[0040] 图 2 展示了当模型损失函数为凸函数且当卖方数据集分布相同且方差都较小、分布不同但方差都较小、分布不同且方差较大时，联合夏普利值的估计误差随训练轮数变化的关系。它与图 1 证明了卖家数据集的分布差异性越大，模型的参数变化越大，进一步使得联合夏普利值的平均估计误差从 0.004 上升到 0.15。当我们将小部分卖家数据集替换为方差更大的数据集时，联合夏普利值的平均估计误差上升到 4.0。这个异常大的误差来源于当移除的卖家数量过多时，对于模型变化ϵ_t^{-Q}的估计很不准确。

[0041] 为了解决上述问题，我们在通过式子$\bar{\phi}_t(c_k) = E_{Q\subseteq C\setminus c_k}[\epsilon_t^{-Q,*} - \epsilon_t^{-Q-c_{k'}*}]$计算每个卖方价值时，忽略当移除的卖方子集Q的数量即|Q|很大的情况。改进估计方法后，本发明对联合夏普利值的估计误差如图 3 所示。

[0042] 为了找到仅仅由本发明估计方法导致的误差，我们首先在计算每个卖方的联合夏普利值时考虑所有可能的边际贡献。由图 4 可以看到有着较大数据方差的卖家也有着较大的估计误差。然后，我们将估计方法与蒙特卡洛采样相结合来降低时间复杂度。本发明尝试了不同的采样数量，例如n^2, n^3，其中$n = |C|$为卖方数量。从图 4 中可以发现与估计方法导致的误差相比，采样带来的误差可以忽略不计。为了解决数据方差大带来估计误差大的问题，我们采用之前所述的改进方法并尝试了不同|Q|作为阈值，由图 4 可以看到，平均联合夏普利值的估计误差从 0.6 降到了 0.2，有较大方差卖家的估计误差从 2.5 降到 0.3

[0043] 图 5 中的 a 和 b 展示了当模型损失函数是非凸的且卖方数据集时独立同分布或者非独立同分布时，本发明对卖方联合夏普利值的估计误差随训练轮数的变化。它验证了我们的理论分析：损失函数为非凸时，估计误差与训练轮数t有指数关系。

[0044] 图 6 展示了联合夏普利值的现实意义和应用效果，我们分别移除有较大、较小联合夏普利值的卖家，并重复联邦学习的训练过程，比较模型性能和表现的变化。实验结果证明移除有较小联合夏普利值的卖方数据集能够提升模型性能，移除有较大联合夏普利值的卖方数据集会降低模型性能。

说 明 书 附 图

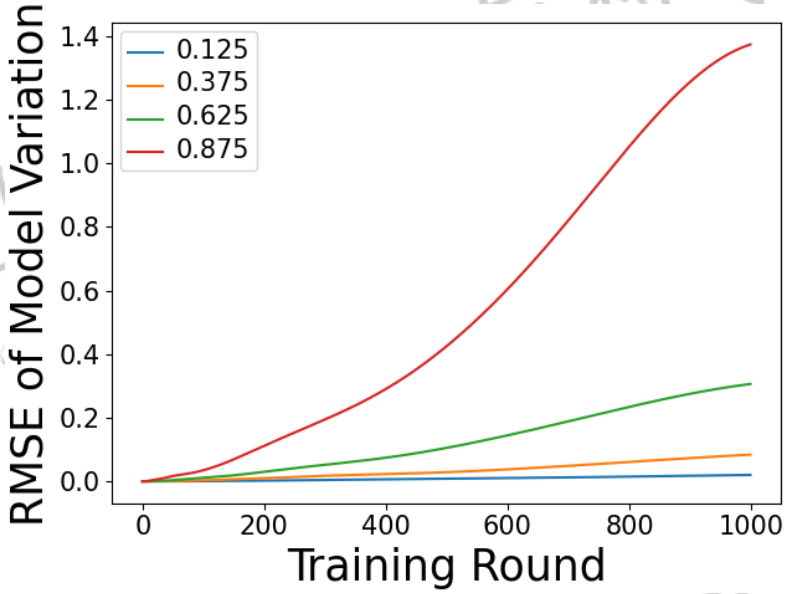
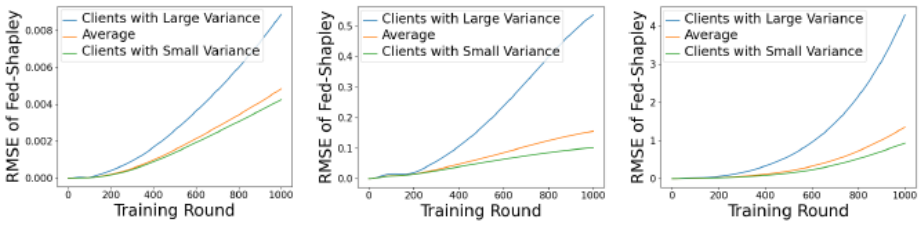


图 1



5

图 2

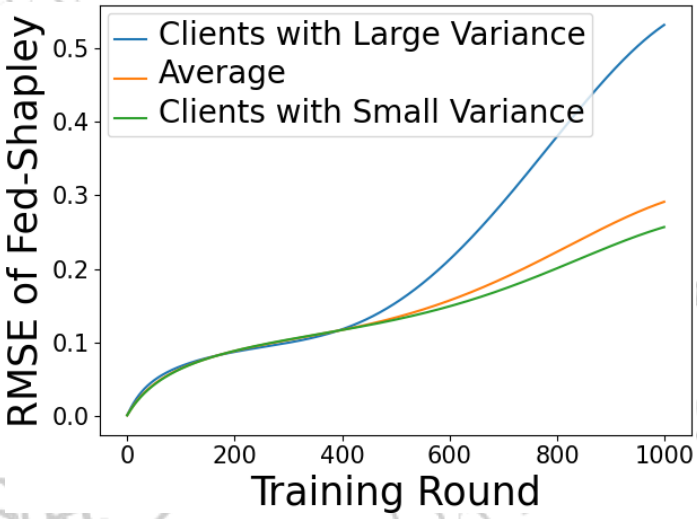


图 3

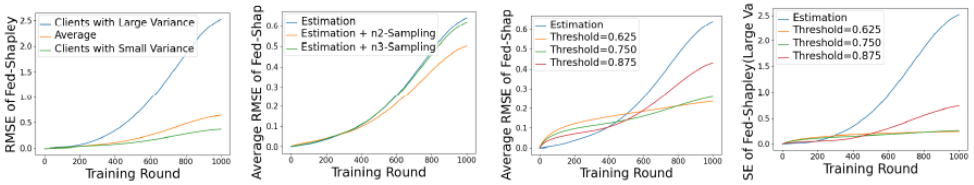
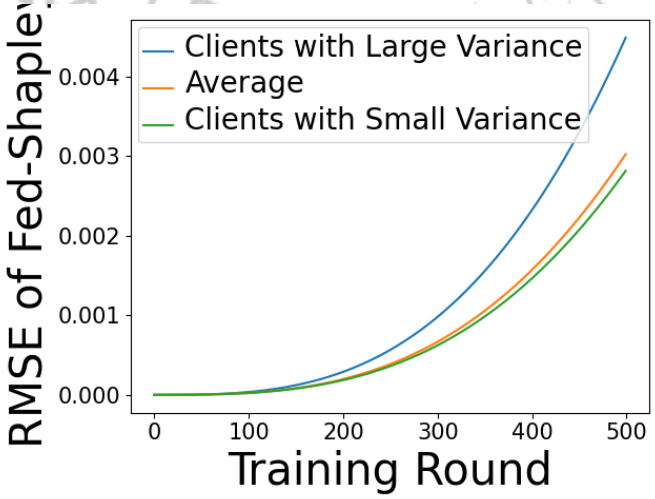

RMSE of Fed-Shapley
Clients with Large Variance
Average
Clients with Small Variance
Training Round
Average RMSE of Fed-Shap
Estimation
Estimation + n2-Sampling
Estimation + n3-Sampling
Training Round
Average RMSE of Fed-Shap
Estimation
Threshold=0.625
Threshold=0.750
Threshold=0.875
Training Round
SE of Fed-Shapley(Large Va
Estimation
Threshold=0.625
Threshold=0.750
Threshold=0.875
Training Round

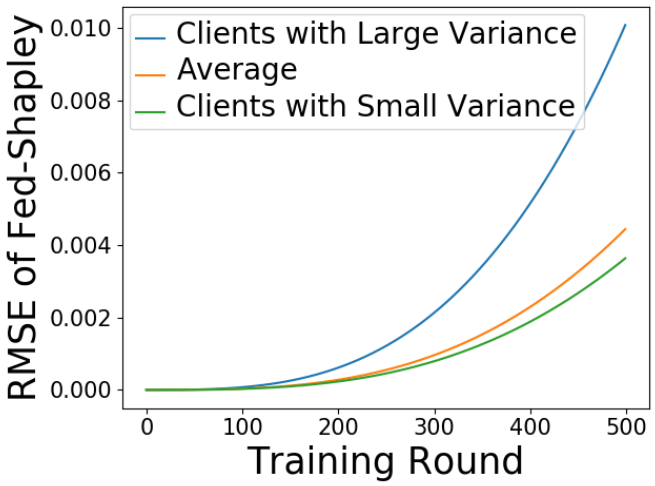
图 4


RMSE of Fed-Shapley
Clients with Large Variance
Average
Clients with Small Variance
0.000
0.001
0.002
0.003
0.004
0
100
200
300
400
500
Training Round

a

5

图 5


RMSE of Fed-Shapley
Clients with Large Variance
Average
Clients with Small Variance
0.000
0.002
0.004
0.006
0.008
0.010
0
100
200
300
400
500
Training Round

b

图 5

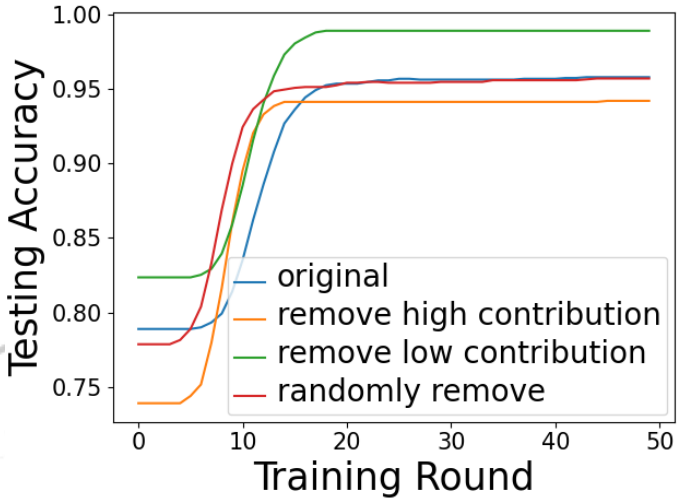


图 6

	Model	Dataset	Distribution	η	$\lvert C \rvert$
Setting 1	LogReg	Synthetic	Non-IID, UnBalance	0.003	8
Setting 2	CNN	FEMNIST	IID, Balance	0.02	8
Setting 3	CNN	FEMNIST	Non-IID, UnBalance	0.02	8

图 7

权 利 要 求 书

1、一种基于联合学习和夏普利值的适用于真实数据市场的数据价值衡量方法，其特征在于，包括以下步骤：

步骤 1、每个卖家k产生一个小型的无偏的数据集d_k，产生方法为根据数据分布p_k直接产生或者从原数据集D_k中随机选取一小部分数据样本。

步骤 2、对于每个卖家及其产生的小型数据集，买家执行一个专门设计的联邦学习过程：在训练过程中的每一轮，每个卖家只执行一次本地参数更新并且所有卖家都参与到模型训练当中；在每一轮中，买家不仅与传统的联邦学习一样分发和收集模型，而且通过本发明的估计方法衡量出每个卖家数据的联合夏普利值$\bar{\phi}_t(c_k)$。在上述条件下，理论证明联合学习得到模型参数和每个数据集的联合夏普利值等于集中式学习下得到的模型参数和每个数据集的数据夏普利值。

步骤 3、当全局模型收敛后，买家得到了每个卖方的联合夏普利值，选择性价比最高的数据集购买。

[0045] 2、根据权利要求 1 所述的方法，其特征是，所述联邦学习满足条件时，得到的模型参数和每个卖家的联合夏普利值等于集中式学习下计算得到的模型参数和每个买家数据集的价值（数据夏普利值）。所述条件为：在训练过程中的每一轮，每个卖家只执行一次本地参数更新并且所有卖家都参与到模型训练当中，

[0046] 3、根据权利要求 1 所述的方法，其特征是，所述的估计方法是指：联合夏普利值可以表示为 $\bar{\phi}_t(c_k) = E_{S\subseteq C\setminus c_k}[\bar{w}_t(S\cup c_k) - \bar{w}_t(S)] = E_{Q\subseteq C\setminus c_k}[(\bar{w}_t(C\setminus Q) - \bar{w}_t(C)) - (\bar{w}_t(C\setminus\{Q,c_k\}) - \bar{w}_t(C))] = E_{Q\subseteq C\setminus c_k}(\epsilon_t^{-Q,*} - \epsilon_t^{-Q-c_k,*})$，$\epsilon_t^{-Q,*}$表示在训练过程中从总卖家集合C移除卖家子集Q后，模型在第t轮的参数变化。其值可以通过本发明的估计方法得到：$\epsilon_t^{-Q,*} \approx \epsilon_t^{-Q} = \sum_{k\subseteq C\setminus Q}\frac{n_k}{N(C\setminus Q)}\left[I - \eta\nabla_w^2 L\left(w_{t-1}^k(C_t), D_k\right)\right]\epsilon_{t-1}^{-Q} + \bar{w}_t(\mathrm{C_t}\to\mathrm{C_t}\setminus\mathrm{Q}) - \bar{w}_t(C)$，$n_k$为第k个卖家的数据集大小；$N(C\setminus Q)$为卖家子集$\mathrm{C}\setminus\mathrm{Q}$的总数据集大小；η为学习率；$\mathrm{L}\left(\mathrm{w}_{t-1}^{k}(\mathrm{C_t}), \mathrm{D_k}\right)$表示当模型参数为$\mathrm{w}_{t-1}^{k}(\mathrm{C_t})$时，模型在数据集$D_k$上的损失函数；$\bar{w}_t(\mathrm{C_t}\to\mathrm{C_t}\setminus\mathrm{Q})$表示只在第t轮将卖家数据集Q移除后全局模型的参数。同时，将传统的 Monte-Carlo 采样方法和本发明的估计方法结合，可以得到时间复杂度更低的估计方法。所述的蒙特卡洛采样是指：随机采样包含所有卖家的一个排列，按照顺序计算每一个排列当中每个卖家数据集对于之前所有数据集的边际贡献，采样多次求取平均值即为每个卖家的估计价值。所述边际贡献是指：加入此卖家数据集后全局模型参数的变化。

说 明 书 摘 要

一种基于联邦学习和夏普利值的适用于真实数据市场的数据价值衡量方法。为了解决数据市场中的信息不对等性，本发明基于联合学习保护数据隐私的特点和夏普利值的公平性，定义了联合学习下的数据价值，联合夏普利值。并用理论证明，当联邦学习满足条件时，得到的模型参数和每个卖家的联合夏普利值等于集中式学习下计算得到的模型参数和每个买家数据集的价值（数据夏普利值）相等。同时，本发明通过数学推导，得到无需重复训练模型的联合夏普利值的估计方法。通过让每个卖家产生小型无偏数据集，对这些数据集进行联合学习的训练过程，在训练过程中用本发明的估计方法维护每一个卖家数据集的联合夏普利值，基于模型收敛后各个数据集的联合夏普利值选择性价比最高的数据集购买。经实验证明，本发明的估计方法有着较小、可容忍的误差，且所定义的联合夏普利值有着实际应用价值和意义。

摘 要 附 图