白泽问卷
让数据产生价值

2021FinTechathon
微众银行第三届金融科技高校技术大赛

作品介绍文档

作品名称：白泽问卷

参加赛道：人工智能赛道

团队名称：白泽

负 责 人：张瑞元

团队成员：刘家祥、阳家勋、魏秋亚、王宵

提交时间：2021 年 11 月 7 号

目录

1. 项目概述

1.1 项目名称

本参赛作品名称为白泽问卷[1]，是一款基于 FATE 平台的模型聚合与数据定价应用。“白泽”也是我们团队的队名，他是黄帝在泰山收的神兽，他上知天文，下知地理，能够预知未来，还能说话，是一种祥瑞的神兽，它就是祥瑞的象征，因此队名和作品名取“白泽”，寓意我们团队希望能够开启一扇联邦学习应用时代的大门，我们的作品能够解决联邦学习现存痛点，并且成为联邦学习界的工具类应用。

图 1-1 白泽问卷 Logo

上图是我们项目 Logo，其中白色鹿形表示我们这次项目的吉祥物白泽；主体上是一个蓝色的盾牌，表示白泽问卷是一款尊重数据隐私的、充分运用联邦学习、区块链、安全多方计算等数据隐私计算技术的软件；主体盾牌也可以看做一张纸质的问卷，其右上角被掀开，露出一个金黄色的小圆点，表示白泽问卷的特色功能，即白泽问卷通过问卷的形式来实现用户数据的收益和公平分配。

1.2 产品说明

白泽问卷是集数据[2]收集、模型训练与聚合、数据定价[3-6]于一体的联邦学习应用，包括小程序和 Web 端两部分。小程序端主要是以问卷形式向海量 C 端用户收集模型训练过程中所需的数据，通过在用户本地生成子模型的方式来保护数据隐私（通过加密传输的方式来保护数据隐私），小程序的特点是轻量级，可以方便地进行问卷的发布与填写。Web 端主要是通过 FATE 平台来聚合模型，并根据数据贡献来进行数据定价，在此过程中生成有价值的模型和可以交易的数据。因此，本产品一方面解决了模型训练数据短缺的痛点，另一方面使数据贡献者得

到应有的收益，促进数据的充分利用与流通，最重要的是让人工智能与公众不再那么遥不可及，让用户在白泽问卷中触碰最新深度学习、联邦学习等技术，解决现实难题。

2.产品核心竞争力

2. 1 产品创新点

2. 1. 1 实际性

1. 这是一个可以收集和分析本地数据的真实的、可推广的、可自定义的开源的联邦学习[7-8]项目。联邦学习的研究者们可以通过更改项目源代码、或直接使用在线平台，来完成联邦学习任务的实际实验。

2. 这是一个让用户可以快速地、无忧分享任何敏感的隐私数据的软件。利用联邦学习[9]、多方安全计算、区块链等隐私计算技术，通过小程序接口，实现用户可以快速地、无忧地分享其隐私数据而不用担心被泄露。

3. 白泽问卷是一个对社会发展有意义的产品。通过收集用户的数据，我们进一步提高了人工智能对医疗、心理健康、民生等领域的增速，比如利用抑郁症患者检测模型来实现用户自测、利用水果新鲜度检测模型来协助人们挑选水果，上述内容均不侵犯用户的隐私，可以放心使用。

2. 1. 2 技术性

1. 通过采用 Shapley 等数据定价方法。本产品实现了公平分配问卷收益，使得数据的提供者能根据数据质量的不同、数据量的不同来收获不同的收益。

2. 通过区块链保障了系统的安全性。本产品通过引入区块链[10]，进一步地保障了联邦学习过程的安全性，①避免用户的数据遭受攻击，②保证了数据定价过程的公开透明、可追溯、防篡改。

3. 使用了 computer vision 以及 nature language process 中最新的 sota[11-12]算法，为联邦学习助力。在算法领域准备使用 Swin transformer、Bert 等网络结构助力分类、分割网络。

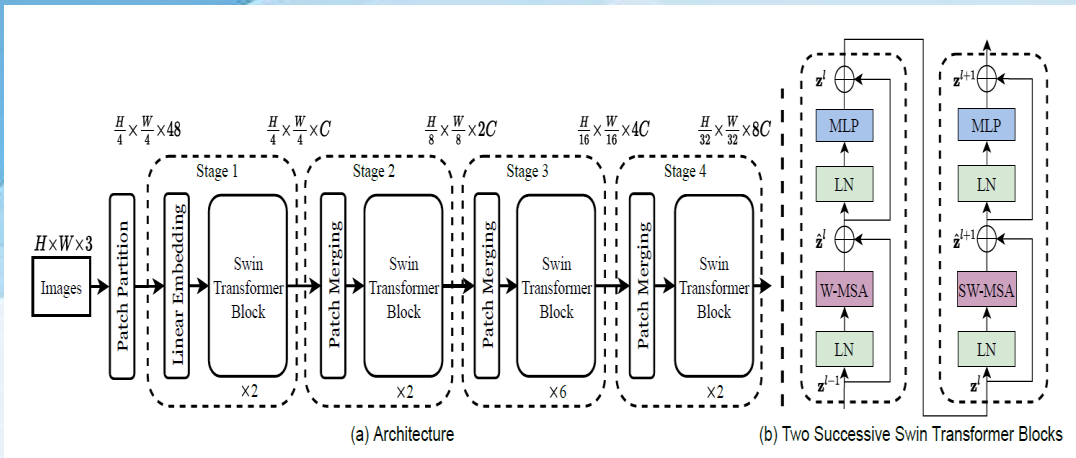

图 2-1 Swin transformer

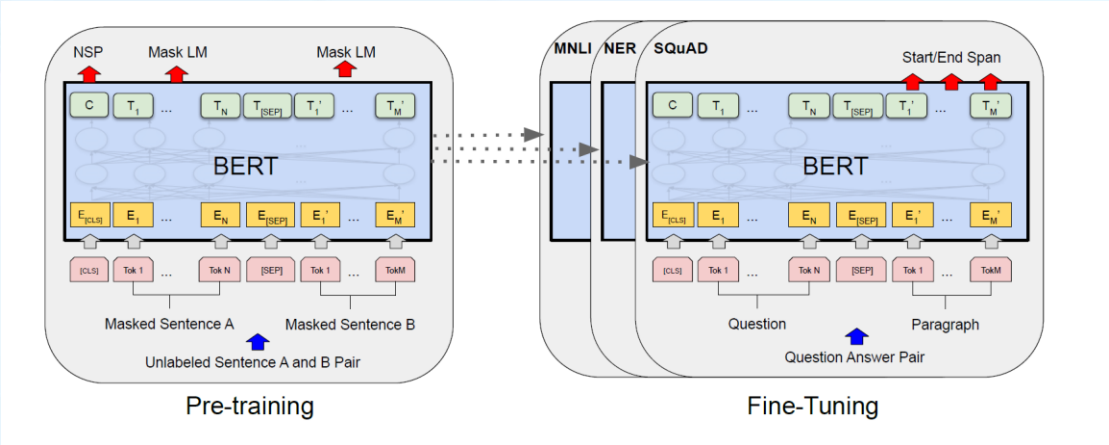

图 2-2 Bert

2. 2 商业应用场景与价值

2.2.1 市场背景

当前是人工智能产业的发展浪潮，主要是源于深度学习[13-14]算法的提出，在数据量和计算能力的基础上实现大规模计算，属于技术性突破。属于超级人工智能的，关于意识起源、人脑机理等方面的基础理论研究仍有继续突破的余地。目前，苹果、谷歌、微软、亚马逊和 Facebook 这五大巨头无一例外都投入了越来越多的资源，来抢占人工智能市场，甚至将自己整体转型为人工智能驱动型的公司。国内互联网领军者也将人工智能作为重点战略，凭借自身优势，积极布局人工智能领域。现今中国人工智能行业的创业公司发展领域各色各异，计算机视觉领域与自然语言处理领域拥有最多创业公司。计算机视觉技术是人工智能的重要核心技术之一，可应用到安防、金融、硬件、营销、驾驶、医疗等领域，而目前我

国计算机视觉技术水平已达到全球领先水平，广泛的商业化渠道和技术基础是其成为最热门领域的主要原因。而目前来看，自动驾驶、医疗、安防、金融、营销等领域是业内人士普遍比较看好方向。而联邦学习对深度学习加持与结合，能够较为完美得解决深度学习中涉及的隐私保护与数据安全问题，让基于数据驱动的深度学习离公众不再遥远。

联邦学习为过往 AI 技术逻辑带来的最大改变在于，它的数据结构可以在参与各方不披露底层数据的前提下，完成共建模型的搭建，之后利用整个数据联邦内的数据资源进行训练，使每个参与方都将获得能力提升。

而联邦学习最大的价值，就是改变了 AI 时代每个数据拥有方单打独斗的“常识”，将数据资源以可行的方式联合在了一起。将联邦学习投向产业应用，最直接的目标是可以改变重度数据安全领域，尤其是金融产业的智能化效率；向长远看，联邦学习可能改变每一家企业获取 AI、打造自身 AI 体系的方式与门槛，对智能社会有着举足轻重的价值。肩负着打破数据孤岛的重任，联邦学习很快成为了 AI 世界中的未来之星。

2.2.2 商业价值

白泽问卷利用联邦学习打开了一道公众与人工智能计算机视觉等最新技术桥接的门，联邦学习的火爆，原因在于它并不致力于改变机器学习和数据存储的基本实现方式，而是改变了不同 AI 模型之间的协作模式。

首个面向 C 端联邦学习工具类应用

为数据需求方与数据拥有者建立桥梁，使联邦学习走进广大 C 端用户，激励广大用户贡献数据，充分释放数据价值；在此过程中，建模方可以获取高质量数据集，并实时查看训练进程，极大提高了建模效率。

小型数据交易市场

在某次任务中积累的数据可能会用于其它模型的训练，因此可以根据系统中的数据定价规则来使数据集在各方之间规范流转。

模型体验社区

任务发起方可以将训练好的模型发布在小程序中，为广大 C 端用户提供各种各样的服务。用户可以在小程序中寻找成熟的模型，也可以发布模型需求。例如

情绪低落的人群可以利用抑郁症检测来判断自己是否已经患有抑郁症，从而及时进行治疗，在此过程中，模型运行在用户手机本地，因此检测结果只有本人知道，不会被泄漏。

2.2.3 目标用户

产品定位为联邦学习数据收集与训练可视化工具。目标用户分为两类：数据需求方和数据拥有者。数据需求方指在训练模型过程中缺少高质量数据集的研究机构或科技公司，数据拥有者指广大 C 端用户。产品为数据需求方提供任务发布、查看训练进展等主要功能，另一方面为数据拥有者提供填写问卷、参与训练、体验模型等功能。让白泽问卷有效地根据用户需求解决各种问题。

因此，本产品的创新点在于，根据任务发布者设置的数据格式生成问卷，C 端用户填写问卷来贡献数据，对 AI 不了解的 C 端用户可以不关心这背后的任务运作，但可以获得贡献数据的奖励；另一方面，想要参与任务的机构或个人也可以通过参与任务来与任务发布方联合建模，模型在各方本地生成，不泄露数据隐私的同时实现数据的融合与应用。最后，任务发布方、数据贡献者和任务参与方都可以获得模型的使用权，形成良性互动。

2.2.4 SWOT 分析

内部分析 外部分析	优势：S 白泽问卷通过联邦学习[3]技术打通了公众与人工智能的壁垒，让公众都能根据不同需求，以白泽问卷的方式定制不同的人工智能服务产品；团队协作能力强，富有凝聚力。	劣势：W 团队经验不足； 初步规模较小。
机会：O 人工智能为中国十四五规划中重点部署之一，随着人工智能的快速发展，其落地需要也愈发强烈。数据驱动的深度学习成为主流的现在，白泽问卷提供一种数据隐私保护的新途径，加速人工智能落地发展；服务对象广；项目具有很强的创新意识。	SO 战略： 提高算法的精确性与鲁棒性； 加大产品的宣传和推广； 加强服务质量； 多与用户交流不断改进加强。	WO 战略： 加强团队成员之间的配合与分工；加强算法效率与精确度的提升。
威胁：T 在技术上没有很强的技术壁垒。	ST 战略： 养成良好的服务态度，提供优质的服务；逐步形成自己的特色，多渠道的降低经营成本。	WT 战略： 多了解客户需求，发布需求量大的联邦学习问卷；在加强自身的同时注意市场变化。

3.产品与研发

3. 1 结构图

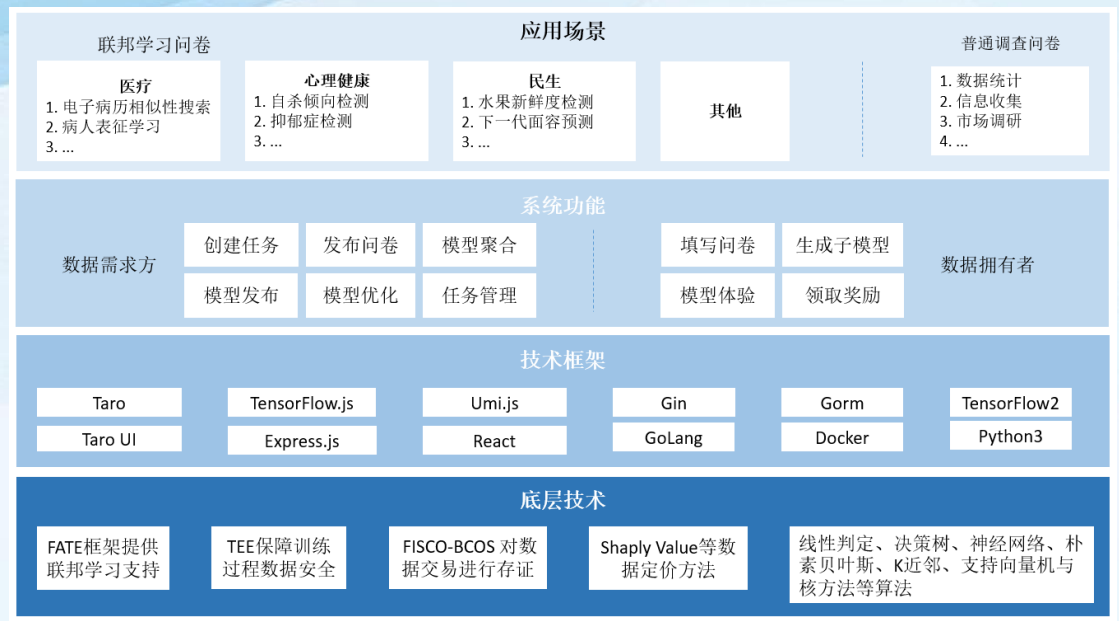

图 3-1 白泽问卷结构图

图 3-1 为白泽问卷的系统结构图，从下往上依次为底层技术、开发技术框架、白泽问卷系统功能、白泽问卷应用领域。底层技术层是白泽问卷中采用的底层技术，包含 FATE 框架提供联邦学习支持，TEE 技术来保障训练过程数据安全，FISCO-BCOS 来对数据交易进行存证，利用 Shapley 等数据定价方法来实现收益的公平，利用线性判定、决策树、神经网络、朴素贝叶斯、K 近邻、支持向量机与核方法等算法实现联邦学习问卷的构建；开发技术框架主要分为前端、后端，前端所用的技术有 Taro、Taro UI、TensorFlow. js、Express. js、Umi. js、React，后端所用的技术有 Gin、Golang、Gorm、Docker、TensorFlow2、Python3 等；白泽问卷提供的系统功能有：问卷分类管理、问卷管理、收益管理、模型预览展示、联邦学习任务调度管理、数据统计管理、系统管理、用户管理等；最后，应用层将服务各领域，提供人工智能模型支持，从联邦学习问卷的角度，可以为医疗、心理健康、民生等领域提供相关服务，从普通调查问卷的角度，亦可以提供数据统计、信息收集、市场调研等功能。

3. 2 联邦学习实现方案

FATE（Federated AI Technology Enabler）是微众银行 AI 部门发起的开源项目，为联邦学习生态系统提供了可靠的安全计算框架[15-19]。它使用多方安全计算（MPC）以及同态加密（HE）技术构建底层安全计算协议，以此支持不同种类的机器学习的安全计算，包括逻辑回归、基于树的算法、深度学习和迁移学习等。

FATE 目前支持三种类型联邦学习算法：横向联邦学习、纵向联邦学习以及迁移学习。FederatedML 是一个实用和可扩展的联邦机器学习库；FATE Serving 是一个可扩展的、高性能的联邦学习模型服务系统；FATEFlow 是一种用于联邦倾斜的端到端管道平台；FATEBoard 是一个面向最终用户的联邦学习建模的可视化工具；Federated Network 是一个联邦学习多方通信网络；KubeFATE 是一个使用云本地技术管理联邦学习工作负载。

Federatedml 模块包括许多常见机器学习算法联邦化实现。所有模块均采用去耦的模块化方法开发，以增强模块的可扩展性。具体来说，其提供了：

1）联邦统计：包括隐私交集计算，并集计算，皮尔逊系数，PSI 等

2）联邦特征工程：包括联邦采样，联邦特征分箱，联邦特征选择等。

3）联邦机器学习算法：包括横向和纵向的联邦 LR，GBDT，DNN，迁移学习等

4）模型评估：提供对二分类，多分类，回归评估，聚类评估，联邦和单边对比评估

5）安全协议：提供了多种安全协议，以进行更安全的多方交互计算。

图 3-2 联邦机器学习框架

FATE 支持 Linux 或 Mac 操作系统，当前 FATE 支持 Native 部署：单机部署和

集群部署;KubeFATE 部署。

在白泽问卷系统中，需要利用小程序来采集用户的数据，并将该数据与客户端优化代码一起打包，通过安全通信通道将打包的任务传送到可信的执行环境下，接下来使用 FATE 框架来对用户的数据进行联合建模。具体内容如下：

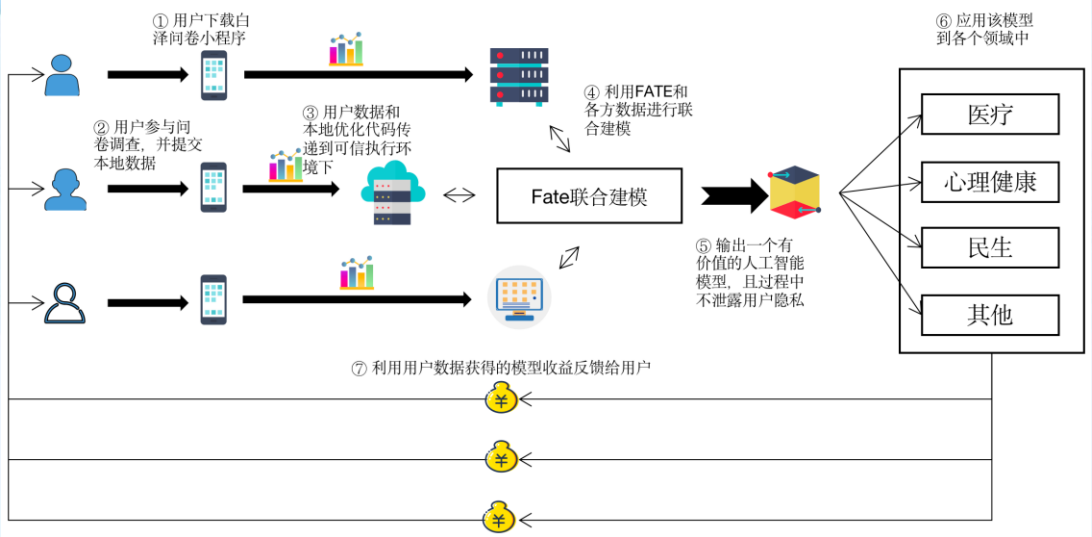

图 3-3 联合建模

如上图 3-3 所示，①用户首先通过微信或支付宝打开白泽问卷小程序，点击一项可以加入的问卷任务，即联邦学习任务；②用户提交本地数据到小程序中，并对数据进行加密打包；③将本地打包好的加密数据和优化代码传递到可信执行环境中，这个执行环境可以是用户信任的第三方服务器、或者是用户自己的电脑中；④接下来，就可以利用 Fate 来对各方的数据进行联合建模；⑤最终，FATE 会输出一个具有价值的人工智能模型，且上述过程中并不会泄露用户的隐私数据；⑥ 接下来，将该模型应用到相应的领域中，实现其价值；⑦最后，利用用户数据和计算资源训练的模型的收益将反馈给用户。

3.3 数据定价方案

Shapley 值法[20-21]是 Shapley L. S 于 1953 年提出，为解决多个局中人在合作过程中因利益分配而产生矛盾的问题，属于合作博弈领域。应用 Shapley 值的一大优势是按照成员对联盟的边际贡献率将利益进行分配，即成员 i 所分得的利益等于该成员为他所参与联盟创造的边际利益的平均值。图 3-4 代表 shapley 算法所实施的数据定价方案图。

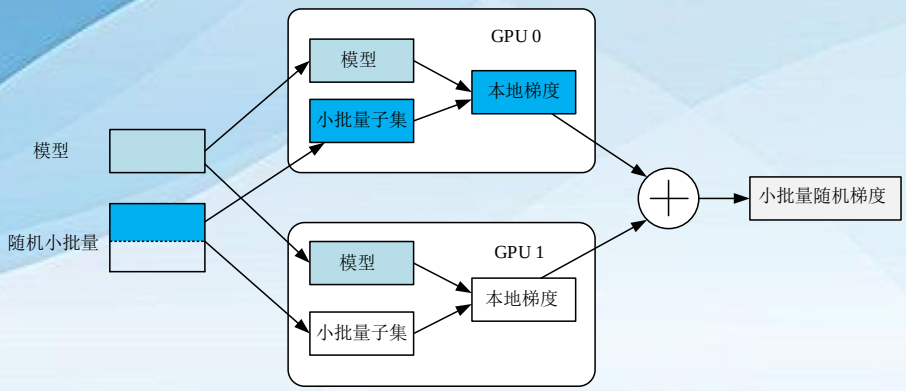

图 3-4 数据定价方案图

在白泽问卷中，最终形成的模型是不同客户上传数据训练后的各模型的聚合。但由于客户上传的数据量、数据质量的不同，对整体模型产生的影响也大有不同。因此，在联邦学习的过程中就涉及到利益分配的问题，这也是联邦学习所必须面对的。我们通过 shapley 算法，对不同客户的数据质量和数据量等进行综合考虑，通过其模型参数对最终模型的边际贡献值来达到较为合理地一种利益分配。相较于传统的只根据不同模型的准确率（未考虑模型合作效果）进行收益分配，shapley 算法显得更为公平。

另外，因为考虑到有些用户在上传数据时，虽然可能对最终模型不能产生很好的效果，传统的 shapley 算法也许就会认为该用户产生的边际贡献值为 0 而对该客户不分配收益。针对此情况，我们对 shapley 算法进行了改进，为这些（没有功劳也有苦劳）特殊用户自动分配一定比例的收益，从而调动不同用户对上传数据的积极性。

Shapley 值法是合作博弈模型中常用的成本分摊方法，在数据定价中应用广泛。假设集合 N= {1,2,⋯,n}表示组成联盟 S 的 n 个成员，X 表示集合 N 中第 i 个成员从联盟 S 总成本 C(S)中应分摊到的成本，则称$X = (X_1, X_2, \cdots, X_n)$为合作博弈对策的分配策略。$X_i$满足以下公式：

$$C(S) = \sum_{i=1}^{n} X_i \tag{1}$$

设各成员在在随机合作时正好形成联盟 S 的概率为W(|S|)，计算方法如下所示：

(2)

可得基于 shapley 值法成本分摊方式中，各成员分摊到的成本为

$$X_i = \sum_{S_i \in S} W(|S|)[C(S) - C(S-i)] \tag{3}$$

其中$[C(S) - C(S - i)]$为成员 i 的边际成本。利用 Shapley 值法分摊成本时，每个模型分摊到的成本取决于由于该成员加入联盟后，给联盟带来的成本增值大小，也就是该聚合模型的边际成本。

各模型边际成本与参数设置情况和各模型之间出力的相关性有关。当模型 i 与其他模型出力序列之间的负相关性越大时，模型 i 的加入会使得整体模型集群出力中呈现的平滑效应更加明显，可用皮尔逊值 p 表示不同模型之间的相关性，设模型 1 实际出力序列P^*_{W1}的平均值为 α 、模型 2 实际出力序列P^*_{W2}的平均值为 β，则 P^*_{W1}与P^*_{W2}之间的相关性系数$p_{1,2}$为：

$$p_{1,2} = \frac{\sum_{i=1}^{n} (P^*_{W1}(i) - \alpha)(P^*_{W2}(i) - \beta)}{\sqrt{\sum_{i=1}^{n} (P^*_{W1}(i) - \alpha)^2}\sqrt{\sum_{i=1}^{n} (P^*_{w2}(i) - \beta)^2}} \tag{4}$$

p 取值范围为(−1，1)。该公式是对两模型相对各出力均值的偏差矢量计算余弦相似度，去中心化的处理使得参数设置相差大的模型间也能有效衡量正负波动相似程度。当 p 值小于 0 表示两个模型出力偏差成负相关，2 者偏差矢量方向呈相反趋势，合作时正负波动易相互抵消，平滑效应明显，模型集群出力平缓，总波动成本下降;当 p 值大于 0 表示两个模型出力成正相关，2 者偏差矢量的方向大体相同，平滑效应不明显，易出现波动性的叠加使得集群出力曲线更加陡峭，但各时刻总出力变化为各成员出力变化的和因此，将模型集群因出力变化率过大引发的波动成本分摊到各成员后，各成员需要支付的成本仍不会高于不合作时的原始成本。

不同模型的的波动成本还与模型的聚合度 S 有关。综合聚合度值低的模型可通过和跟随聚合程度高的模型合作来改善出力曲线，使集群模型出力波动趋势向聚合规律改进;然而对综合聚合度值高的模型来说，加入集群后对出力曲线的改善有限，且成员间的关联程度导致集群模型聚合度小于成员原来的模型聚合度，

此时需要提出一种激励各成员适当控制参数聚合、改善综合聚合度的分摊机制来促进集群模型出力曲线的持续优化。而在 Shapley 值法下，边际成本与相关性配适情况有关，并不能充分表现各模型对整体模型集群的综合聚合度的影响。此时可以设计一个考虑综合聚合度的波动成本分摊系数 K，通过结合分摊系数与 Shapley 值法得到改进 Shapley 值。另外从监管的角度看，对模型跟随参数波动情况衡量系数的考核属于合格性考核，Shapley 值法下各模型指标合格后，持续改善跟随参数情况也不能进一步降低边际成本;而根据各模型影响集群模型的聚合情况进行再分摊则会激励各模型在达到合格性指标之后继续改善 S 值，提升最终模型质量。波动成本分摊系数 K 体现各模型成员影响集群出力跟踪参数波动的情况，首先与各成员的综合聚合度 S 有关。S 值大的成员使得模型集群出力波动特性向参数波动特性靠近，减少集群的波动成本。设考虑模型 i 与其等效模型聚合度的分摊系数为$K_{S,i}$。当整体模型聚合度越大时，模型可靠程度与参数波动规律差异越小，模型给最终模型带来的附加成本越低，不同模型分摊到的波动成本越少，$K_{S,i}$的值也应该更小。设模型 i 的聚合度为S_i，可得$K_{S,i}$，的计算公式为：

$$K_{s,i} = \frac{\frac{1}{S_i}}{\sum_{j=1}^{n} \frac{1}{S_j}} \tag{5}$$

波动成本分摊系数 K 还与模型集群成员的数据量有关。模型集群出力为各成员出力之和，数据量大的成员的出力在集群出力中占比高，其出力规律给模型集群出力规律带来的影响也更大。而 S 值经归一化和标准化处理后无法体现各成员数据量的大小情况，因此 K 的计算还需单独考虑数据量的影响。设衡量模型 i 数据量的分摊系数为$K_{Q,i}$。数据量越小的模型，在集群模型中出力占比小，对应的分摊系数$K_{Q,i}$值也更小，设模型 i 数据量为Q_i，可得$K_{Q,i}$计算公式为：

$$K_{Q,i} = \frac{Q_i}{\sum_{j=1}^{n} Q_j} \tag{6}$$

设$K_{Q,i}$和$K_{S,i}$的对应权重为w_1和w_2，可得模型 i 的聚合度分摊系数K_i为：

$$K_i = w_1 K_{Q,i} + w_2 K_{S,i} \tag{7}$$

w_1与w_2之和为 1。为表示$K_{Q,i}$和$K_{S,i}$之间的动态关系，本方案采用熵值法计算权重，权重取值与各模型之间的指标差异有关。当某一分摊系数差值越大，对应的权重更大，分摊时就会重点考虑此分摊系数。当各模型数据量相近时，各模型给集群模型出力带来的影响主要体现在成员的综合聚合度上，此时$K_{Q,i}$应分得的权重小;当各模型数据量相差较大而综合聚合度相近时，数据量越大的模型给集群模型带来的影响越大，此时$K_{Q,i}$应分得的权重大。由于模型的波动性，各模型数据量和综合聚合度并不确定，因此$K_{Q,i}$和$K_{S,i}$之间的权重应根据系数的变化进行相应改变，而熵值法正符合这实际需求。设有 n 个模型， m 个影响指标 x，可得指标矩阵X_{ij}为：

$$\boldsymbol{X}_{ij} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \tag{8}$$

先使用 mapminmax 对指标进行归一化处理，得指标矩阵$X_{ij}^{'}$。设第 j 项指标的熵值为e_j：

$$\boldsymbol{e}_j = -\boldsymbol{k}\sum_{i=1}^{n}\left(\frac{\boldsymbol{X}'_{ij}}{\sum_{i=1}^{n}\boldsymbol{X}'_{ij}}\right)\ln\left(\frac{\boldsymbol{X}'_{ij}}{\sum_{i=1}^{n}\boldsymbol{X}'_{ij}}\right) \tag{9}$$

其中 k 一般表示为$(\ln(n))^{-1}$，可得到第 j 项指标的权重w_j为：

$$w_j = \frac{g_j}{\sum_{j=1}^{m} g_j} \tag{10}$$

其中：$g_j = 1 - e_j$；w_j的取值在 0 到 1 之间；且 m 个指标权重之和为 1。

按照上述逻辑即可实现模型的数据定价。

3.4 产品实现方案

3.4.1 数据库设计

表 3-1：角色表 role

属性名	数据类型	是否为空	主键/外键	其他
id	int	否	主键	
name	varchar(50)	否		unique

表 3-2：用户表 user

属性名	数据类型	是否为空	主键/外键	其他
id	int	否	主键	
username	varchar(50)	否		unique
password	varchar(255)	否		
is_enabled	boolean	否		默认为 Ture

表 3-3：数据集表 dataset

属性名	数据类型	是否为空	主键/外键	其他
id	int	否		unique
name	Varchar(50)	否		
file_type	Varchar(20)	否		
path	Varchar(256)	否		
header	boolean	否		
split	boolean	否		
user_id	int	否	外键	

注：其他数据库表类似，不在赘述。

3.4.2 接口设计

表 3-4：分类管理-添加分类

标题	内容
简要描述	用户创建一个分类
请求 URL	http://xx. com/api/category/add
请求方式	POST

表 3-5：参数

参数名	必选	类型	说明
name	是	string	分类名称

参数名	必选	类型	说明
description	是	string	描述
file	否	string	图标路径

返回示例

```
{
  "error_code": 0,
  "data": {
  "id":123,
    "name":"军事",
    "description":"军事描述描述描述",
    "file":"图标文件地址"
  }
}
```

表 3-6：返回参数说明

参数名	类型	说明
id	int	任务分类的 id，由雪花算法生成
name	string	任务分类的名称
description	string	任务分类的描述
file	file	图标文件

备注：更多返回错误代码请看首页的错误代码描述

表 3-7：分类管理-分类列表

标题	内容
简要描述	任务分类列表
请求 URL	http://xx. com/api/category/list
请求方式	POST

表 3-8：参数

参数名	必选	类型	说明
page	是	int	第几页

参数名	必选	类型	说明
limit	是	int	每页大小

返回示例

```
{
    "error_code": 0,
    "data":{
            page:1,
            limit:10,
            list:[{
                    "id":123,
                    "name":"军事",
                    "description":"军事描述描述描述",
                    "file":"图标文件地址"
                    },{
                    "id":1234,
                    "name":"军事",
                    "description":"军事描述描述描述",
                    "file":"图标文件地址"
            }]
    }
}
```

表 3-9：返回参数说明

参数名	类型	说明	备注
page	int	第几页	
limit	int	每页大小	
id	int	任务分类的 id，由雪花算法生成	
name	string	任务分类的名称	
description	string	任务分类的描述	
file	file	图标文件地址	

备注：更多返回错误代码请看首页的错误代码描述

表 3-10：任务列表（简略信息）

标题	内容

简要描述	任务列表
请求 URL	http://xx. com/api/task/list
请求方式	POST

表 3-11：参数

参数名	必选	类型	说明
page	是	int	第几页
limit	是	int	每页大小

表 3-12：返回示例

```
{
  ″error_code″: 0,
  ″data″:{
  page:1,
  limit:10,
  list:[{
        id: 123,
        catagoryId:456,
        name:″mnist″,
        description:’测试使用’,
        file: ″public/task/1. png″
  }]
  }
}
```

表 3-13：返回参数说明

参数名	类型	说明	备注
page	int	第几页	
limit	int	每页大小	
id	number	联邦学习任务 id，由雪花算法生成	
catagoryId	是	number	任务分类编号
name	string	联邦学习任务名称	
description	string	任务详细描述，是富文本框生成的 html	
file	string	联邦学习任务图像	地址

备注：更多返回错误代码请看首页的错误代码描述

表 3-14：模型管理-提交梯度信息

标题	内容
简要描述	提交梯度信息
请求 URL	http://xx. com/api/gradient/add
请求方式	POST

表 3-15：参数

参数名	必选	类型	说明
taskId	是	number	
globalModelId	是	number	上一个全局模型 id
modelFile	是	string	模型地址，格式为任务 id/用户名/模型名称

返回示例

```
{
  ″error_code″: 0,
  ″data″: {
       clientModelId:123,
       use: false,
       modelFile: ’public/task1/user1/mnist01’,
  }
}
```

表 3-16：返回参数说明

参数名	类型	说明
clienModelId	number	客户端上传模型编号
use	boolean	模型更新的过程中是否使用
modelFile	string	模型所在地址

备注：更多返回错误代码请看首页的错误代码描述

表 3-17：接口 5：获取历史训练记录（模型记录）

标题	内容
简要描述	获取历史训练记录
请求 URL	http://xx. com/api/gradient/list
请求方式	POST

返回示例

```
{
  ″error_code″: 0,
   data: [{
        clientModelId: 123,
        clientModelFile: ’’,
        globalModelId: 345,
        globalModelFile: ’’
  }, {
        clientModelId: 123,
        clientModelFile: ’’,
        globalModelId: 345,
        globalModelFile: ’’
  }, {
        clientModelId: 123,
        clientModelFile: ’’,
        globalModelId: 345,
        globalModelFile: ’’
  }]
        ″groupid″: 2 ,
        ″reg_time″: ″1436864169″,
        ″last_login_time″: ″0″,
      }
    }
```

表 3-18：返回参数说明

参数名	类型	说明
clientModelId	number	客户端模型编号
clientModelFile	string	客户端模型文件地址
globalModelId	number	全局模型编号

参数名	类型	说明
globalModelFile	string	全局模型文件地址

备注：更多返回错误代码请看首页的错误代码描述

注：其他接口内容类似，不在赘述。

3.4.3 界面设计

界面设计内容不再展示，具体内容见下一章「项目展示」。

4.项目展示

4.1 操作步骤

用户从首页点击「发布问卷」，选择「联邦学习问卷」，此时可以选择「从模板创建」和「自定义创建」。

图 4-1 问卷发布

「从模板创建」：进入模板分类列表，选择要使用的模板，模板上会展示任务简介、参数、数据格式等，点击「使用此模板」，进入模板编辑界面，可以设置最大参与次数、每份数据奖励等，点击「预览」，可以看到发布后的问卷样式，点击「发布」，即可讲问卷分享给问卷填写者。

图 4-2 问卷样式

4.2 模型训练可视化

用户可以在小程序中查看任务的训练进展，分为「我发布的任务」和「我参与的任务」。包括任务信息等。

「我发布的任务」：点击「任务进展」可以查看当前聚合进展和历史聚合情况以及模型准确率变化。

图 4-3 任务进展

「我参与的任务」：点击「任务进展」可以查看我在每轮聚合中的参与情况，点击「长期运行」可以将任务打包后在电脑端执行；点击「数据」可以查看或添加本地数据，还可以点击「手动参与训练」来生成并上传本地模型。

图 4-4 参与的任务

4. 3 有价值的模型展示

首页会为用户推荐最热门的模型，用户点击模型可以进入模型体验界面，当前可以体验的模型包括抑郁症检测，水果新鲜度检测，未来宝宝长相预测等。

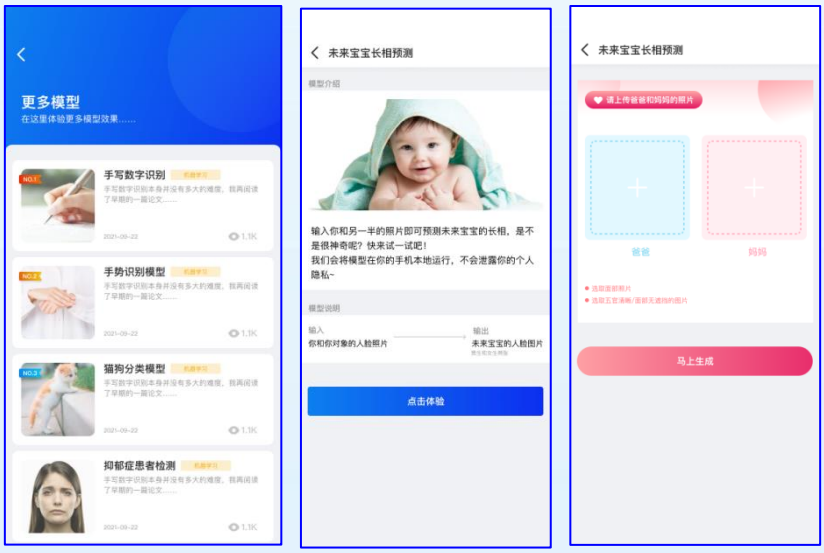

图 4-5 模型展示

5.产品营销

5. 1 定价策略

5.1.1 定价策略分析

价格通常是影响交易成败的重要因素，同时又是市场营销中最难以确定的因素。本项目的定价目标主要考虑服从于总体战略发展的目标，也即是通过与战略配套的定价策略，达到尽快进入市场，提高市场占有率这一目标。这要求项目策划既要考虑成本的补偿，又要考虑消费者对价格的接受能力，从而使定价策略具有买卖双方双向决策的特征，本次定价策略采用 3C 模型分析。

A）客户需求弹性分析——具备一定弹性

本问卷是传统问卷的改进品，无论从问卷设计、隐私保护等方面都优于传统问卷，但是传统问卷调查已经被普遍接受的现实对我们来说是一个最大的挑战。我们从客户需求弹性上分析可得分析本公司的产品属于具有一定需求弹性的产品，原因有以下 2 点：

产品重要性	问卷的改进和发布在总支出虽然占据一定比例，但其类似于 app 的更新换代，是真正面向客户的服务，对客户有一定的影响。
产品定制性	本次联邦学习的问卷处理包含传统问卷外，还面向特定的企业、医务等技术人员，相比较于传统问卷，本问卷具有定制性。

由于本产品的有一定弹性，且具备普通问卷所不具备的功能，因此，产品的定价应稍微高于市场水平价。

B）产品成本分析——规模效应

产品[22]采用自行设计结构、方案、自行发布的生产模式，选址在杭州下沙高教园区，交通便利、教育实验资源发达地区，有利于产品推广及应用。

C）竞争者分析——竞争较小

目前国内市场上在面向 C 端的联邦学习问卷发布和设计上较少细分市场甚至没有竞争者，而国外也很少有同类产品。即使白泽问卷以较高的成本利润率定价，相信在目标客户内也是可以接受的范围。

5.1.2 定价策略选择

综合考虑行业成本利润率和白泽问卷产品的技术价值等相关因素，公司将产品毛利率定为 50%~80%。，面对竞争，即使采取低价策略时，我们仍有较大的降价空间以保持自身优势。另外，项目组会根据不同的销售渠道对产品的定价做适当的调整。对于客户超出价格浮动范围的出价需进行申请。项目组将从战略性合

同规模大小等角度对该合同进行考量后决定是否给予相应的价格让步。

5.2 营销策略

A）导入期——直接销售

通过销售人员直接上门推销的方式进行销售，同时初期随着广告的投放公关的宣传工作，会有一定的企业主动联系进行详细了解，此时的销售工作都全权由销售人员进行，省去了中间商的环节大大降低了成本，使白泽问卷产品在价格上更有优势。

B）成长期——代理销售+网络销售

此阶段顾客对问卷已熟悉，销量大幅增加，利润大幅增加 同时竞争更加激烈。因而此时应该增加销售机构和网点拓展新的市场，同时应花重金在建设项目的网站上，尤其在电子商务发展壮大的今天，网购已逐渐进入了人们生活中，正是在这种趋势下在组织市场上网上购买问卷使用也将成为主力军，产品发布与推广需要响应时代的号召把握这个机会，不光在技术上领先在销售方式上也要走在时代的前沿。

C）成熟期——代理销售+产业联盟

在国内建立了一定的知名度与口碑后，我们将开拓国外市场，可通过在国外找个代理的形式进行销售，或直接在网站上与国外的问卷制作与发布的企业进行联系进行市场开发销售，为了将白泽问卷推向更广的市场，除了采用代理销售的方式，另一方面可以和其他企业联盟。由于产品已经到了成熟期，可以推行产业链联盟，比如和软件制作商结盟，以便后期新问卷的推发。同时，还可以和已有合作的企业继续合作，推行研发和市场合作联盟。

5.3 推广策略

5.3.1 初创期

1）试用期

与有购买意向的客户签订 1~3 月的试用期协议。在试用过程中，我们将定期到各个客户单位进行了解试用情况，采集产品使用数据进行分析报告。试用期结

束后，若客户无使用意向，可以不收取费用。客户如果后期继续使用，需支付一定的平台使用费用并签署相应的合同。

2）优惠活动

对于相关客户，如果吸引其他客户使用，会获得一定的奖励。在使用白泽问卷超过一年以上的客户可以提供 VIP 折扣优惠活动并获得相关的礼品。VIP 客户可以获得问卷最新提供的一些高性能服务，以及在使用问卷过程中遇到什么问题会第一时间得到解决。

3）专业推广

发布相关联邦学习整合进不同平台的论文和专利，在技术人才和高端职业领域产生较大的影响，从而吸引更多的人来关注产品。

5.3.2 成熟期

以人员营销为主，可以与银行、美团、高校工作者等各种企业和联盟进行更为密切的合作。通过广告推广，线下访谈等多种形式来进一步提高产品使用率。

6.产品规划

6.1 应用规划

1）图像分类和图像识别

机器学习、深度学习和神经网络[23-25]可以帮助人们理解图像。这种技术有着广泛的应用，从社交网站想要给其网站上的照片贴上标签，到安全团队想要实时识别犯罪行为，再到自动化汽车需要通畅的道路。零售商在图像分类和图像识别方面也有很多应用。配备具有计算机视觉和机器学习的机器人可以扫描货架以确定哪些物品是缺货或放错地方；使用图像识别可以确保从购物车中取出的所有物品被成功扫描，从而限制无意中的销售损失；通过分析图像还可以识别可疑活动，如入店行窃以及检测违反工作场所安全的行为（如未经授权使用危险设备）等。

2）水蔬新鲜度打分

用户通过手机拍摄水果蔬菜的照片，经过模型的计算和匹配，得出水蔬的一个新鲜度分数，通过分数，白泽问卷有以下五种状态的新鲜度打分推荐：’此水

果新鲜度过低，请勿食用’’此水果新鲜度低，谨慎食用’’此水果请尽快食用’’此水果新鲜度正常’’此水果非常新鲜’，能够帮助用户快速识别水果种类以及其新鲜度程度。后续将加入数据定价体系，通过联邦学习，学习当季当地水蔬的价格，计算得出一个推荐的价格。还可以挖掘历史定价数据和一系列其他变量的数据集，以了解特定的动态因素（从每天的时间、天气到季节）如何影响商品和服务的需求。机器学习算法可以从这些信息中学习，并将这些洞察力与其他市场和消费者数据结合起来，帮助企业根据这些庞大且众多的变量动态定价商品，这一策略最终将帮助企业实现收入最大化。动态定价(有时称为需求定价)最常发生在运输行业，例如网络约车会随着叫车人数增加而飙升定价或要求增加同乘人数，另外还有在学校假期期间飙升的机票价格等。

3）文本情感分析

文本情感分析(Sentiment Analysis)是指利用自然语言处理和文本挖掘技术，对带有情感色彩的主观性文本进行分析、处理和抽取的过程。目前，文本情感分析研究涵盖了包括自然语言处理、文本挖掘、信息检索、信息抽取、机器学习和本体学等多个领域，得到了许多学者以及研究机构的关注，近几年持续成为自然语言处理和文本挖掘领域研究的热点问题之一。情感分析任务按其分析的粒度可以分为篇章级，句子级，词或短语级；按其处理文本的类别可分为基于产品评论的情感分析和基于新闻评论的情感分析;按其研究的任务类型，可分为情感分类，情感检索和情感抽取等子问题。一般情况下，加权计算结果为正是正面倾向，结果为负是负面倾向 ，得分为零无倾向。如果其负值过大，可以判定其具有极其消极的情感倾向，这时可以考虑为有抑郁甚至自杀倾向。基于情感词典的方法和基于机器学习的分类算法相比，虽属于粗粒度的倾向性分类方法，但由于不依赖标注好的训练集，实现相对简单，对于普遍通用领域的网络文本可有效快速地进行情感分类。

6.2 展望

进一步，产品在功能上还需进行如下的优化提升：一是精准投放。根据用户的特征，为用户推荐相关任务；二是把控数据质量。使用 AI+人工的方式，多方

位监控问卷质量，使数据定价有依据，更客观；三是加速数据回收。提供更更富的模板，使任务快速创建，问卷一键投放，多维触达，快速回收目标样本数据。

在技术上，可以利用区块链来实现数据协作的全流程保护，将数据的使用与操作记录上链，以旁路的形式作为数据合规使用的证明。另外，使用可信计算为各方任务运行提供可靠容器，保护数据的隐私性与机密性。

参考文献：

[1] 顾纯存,余东江.《白泽》[J].美苑,2015(S2):178.

[2] 曾高雄,胡水海,张骏雪,陈凯.数据中心网络传输协议综述[J].计算机研究与发展,2020,57(01):74-84.

[3] Geoffrey MUKWADA,Desmond MANATSA.Acacia mearnsii management in a South African National Parks:SWOT analysis using hot topics in biological invasion as a guide[J].Journal of Mountain Science,2017,14(01):205-218.

[4] 尹鑫,田有亮,王海龙.面向大数据定价的委托拍卖方案[J].电子学报,2018,46(05):1113-1120.

[5] Yuanhua WANG, Daizhan CHENG, Xiyu LIU. Matrix expression of Shapley values and its application to distributed resource allocation[J]. Science China(Information Sciences),2019,62(02):46-56.

[6] 彭慧波,周亚建.数据定价机制现状及发展趋势[J].北京邮电大学学报,2019,42(01):120-125.

[7] 吴文峻,黄铁军,龚克.中国人工智能的伦理原则及其治理技术发展[J].Engineering,2020,6(03):212-229.

[8] QIU XiPeng, SUN TianXiang, XU YiGe, SHAO YunFan, DAI Ning, HUANG XuanJing. Pre-trained models for natural language processing: A survey[J]. Science China(Technological Sciences),2020,63(10):1872-1897.

[9] 芦效峰,廖钰盈,Pietro Lio,Pan Hui.一种面向边缘计算的高效异步联邦学习机制[J].计算机研究与发展,2020,57(12):2571-2582.

[10] 王晨旭,程加成,桑新欣,李国栋,管晓宏.区块链数据隐私保护：研究现状与展望

[J].计算机研究与发展,2021,58(10):2099-2119.

[11]Sen XU, Xiangjun LU, Kaiyu ZHANG, Yang LI, Lei WANG, Weijia WANG, Haihua GU, Zheng GUO, Junrong LIU, Dawu GU. Similar operation template attack on RSA-CRT as a case study[J]. Science China(Information Sciences),2018,61(03):131-147.

[12]Xin Li, Chongyin Li. Occurrence of two types of El Nino events and the subsurface ocean temperature anomalies in the equatorial Pacific[J].Chinese Science Bulletin,2014,59(27):3471-3483.

[13]刘彩霞,魏明强,郭延文.基于深度学习的三维点云修复技术综述[J/OL].计算机辅助设计与图形学学报:1-17[2021-11-07].http://kns.cnki.net/kcms/detail/11.2925.TP.20211101.1210.006.html.

[14]杨岳毅,王立德,王冲,王慧珍,李烨.基于深度主动学习的MVB网络故障诊断方法[J/OL].西南交通大学学报:1-8[2021-11-07].http://kns.cnki.net/kcms/detail/51.1277.U.20211103.1048.002.html.

[15]陈大卫,付安民,周纯毅,陈珍珠.基于生成式对抗网络的联邦学习后门攻击方案[J].计算机研究与发展,2021,58(11):2364-2373.

[16]刘庆祥,许小龙,张旭云,窦万春.基于联邦学习的边缘智能协同计算与隐私保护方法[J].计算机集成制造系统,2021,27(09):2604-2610.

[17]田家会,吕锡香,邹仁朋,赵斌,李一戈.一种联邦学习中的公平资源分配方案[J/OL].计算机研究与发展:1-14[2021-11-07].http://kns.cnki.net/kcms/detail/11.1777.tp.20210825.1542.007.html.

[18]史鼎元,王晏晟,郑鹏飞,童咏昕.面向企业数据孤岛的联邦排序学习[J].软件学报,2021,32(03):669-688.

[19]卢绍帅,陈龙,卢光跃,管子玉,谢飞.基于弱监督对比学习的小样本情感分类[J/OL].计算机研究与发展:1-13[2021-11-07].http://kns.cnki.net/kcms/detail/11.1777.tp.20211104.1703.017.html.

[20]庞传军,刘金波,张波,杨笑宇,余建明,刘艳.基于 Shapley 值的电力负荷预测结果溯源分析方法[J/OL].电力自动化设备:1-6[2021-11-07].https://doi.org/10.160

81/j.epae.202110001

[21]郑晨昕,江岳文.基于改进 Shapley 值的风电波动成本分摊策略[J/OL].电网技术:1-8[2021-11-07].https://doi.org/10.13335/j.1000-3673.pst.2021.0275.

[22]吴璠,王中卿,周夏冰,周国栋.基于用户和产品表示的情感分析和评论质量检测联合模型[J].软件学报,2020,31(08):2492-2507.

[23]李少波,杨磊,李传江,张安思,罗瑞士.联邦学习概述：技术、应用及未来[J/OL].计算机集成制造系统:1-29[2021-11-07].http://kns.cnki.net/kcms/detail/11.5946.TP.20210831.1414.006.html.

[24]汪航,田晟兆,唐青,陈端兵.基于多尺度标签传播的小样本图像分类[J/OL].计算机研究与发展:1-10[2021-11-07].http://kns.cnki.net/kcms/detail/11.1777.TP.20210825.1442.003.html.

[25]郭松,范存群.基于重要度理论的图像识别方法[J].计算机集成制造系统,2021,27(09):2736-2740.

白泽团队感谢您的阅读~

预祝本届金融科技高校技术大赛圆满结束

预祝白泽问卷团队取得优异成绩