

编号:

中国地质大学（武汉）计算机学院  
2019-2020 年度本科生科研立项

申  
报  
书

项目名称：基于联邦学习的数据隐私保护

申 报 人：常文瀚

联系方式：17695870530

指导老师：朱天清

中国地质大学（武汉）计算机学院

二〇一九年十二月

## 一、申报者情况

负 责 人	姓名	常文瀚	性别	男	学号	20181001095
	民族	汉族	政治面貌	团员	班级	191181
	邮箱	1378704731@qq.com				
	联系方式		电话：17695870530    QQ 号：1378704731			
团 队 成 员	姓名	性别	班级	学号	电话	
	周子荟	女	192181	20181001311	15531195363	
	李嘉诚	男	191183	20171000983	13296579812	
	田冰	男	191181	20181001910	13608410519	
	刘宇鹏	男	191181	20181003174	15623280275	
	杨彤	男	191181	20181000651	15927557508	
指 导 老 师	姓名	性别	职称	邮箱	电话	
	朱天清	女	教授	<a href="mailto:tianqing.e.zhu@foxmail.com">tianqing.e.zhu@foxmail.com</a>	17720484903	

## 二、申报项目情况

项目名称：
项目简介（包括项目独特性及创新性论述）
<p>大数据环境下面临着大量信息泄露的挑战。特别在机器学习的环境下，由于需要共享大量的训练数据，信息的隐私容易被轻易泄露。联邦学习采用了分布式的联合学习方式，只需要和中心共享参数，而不用共享数据，很好的保护了用户的隐私。但联邦学习依旧面临算法健壮性差，精确度不够，以及容易被攻击的弱点。目前的课题研究如何在保证隐私的情况下，增强联邦学习的抗攻击性。</p> <p>联邦学习的出处是金融机构的痛点，尤其是像“微众银行”这样的互联网银行。一个实用的例子是检测多方借贷。这在银行业，尤其是互联网金融一直是很头疼的一个问题。多方借贷是指某不良用户在一个金融机构借贷后还钱给另一个借贷机构，这种非法行为会让整个金融系统崩溃。要发现这样的用户，传统的做法是金融机构去某中心数据库查询用户信息，而各个机构必须上传他们所有用户，但这样做等于暴露金融机构的所有重要用户隐私和数据安全，这在GDPR下就不被允许。在联邦学习的条件下，没有必要建立一个中心数据库，而任何参与联邦学习的金融机构可以利用联邦机制向联邦内的其他机构发出新用户的查询，其他机构在不知道这个用户具体信息的前提下，回答在本地借贷的提问。这样做既能保护已有用户在各个金融机构的隐私和数据完整性，同时也能完成查询多头借贷的这个重要问题。</p> <p>针对不同数据集，联邦学习分为横向联邦学习(horizontal federated learning)、纵向联邦学习(vertical federated learning)与联邦迁移学习(Federated Transfer Learning, FmL)</p>
项目的实际应用价值、现实意义及国内外研究现状

联邦学习具有保护隐私和多方本地数据安全的极大优势。避免集中式存储数据，安全合规地从多源不互通的数据中创造新的价值，充分利用各方数据资源，优化机器学习训练结果，学习参与方可以在联合形成协同合作的联邦大数据环境，形成联邦学习生态。对金融、互联网、通信、零售、交通运输、工业生产等行业提供计算服务支持。我们可以从以下四方面窥探联邦学习的价值和意义。

1. 丰富的数据资源是联邦学习最大的金矿。原本分散在各规模企业的数据，通过联邦学习生态达成，可以发挥其自身作用，有了更好的用武之地。利用“联邦学习+人工智能”真正的赋能大数据并反哺个人和企业业务，用数据和科学提升业务效益。

2. 打破传统企业机构的数据边界，利用联邦学习提升智能化效果。改变过去商务智能和政府仅仅依靠机构内部数据的局面。

3. 达成各行业联手，共建全行业的联邦学习生态。联邦学习的出现已经开始改变大数据在各行各业的应用方式，联邦大数据生态的构建也离不开学界和工业界的共同探索和推动，使用联邦学习技术的各方应当携手，联合制定数据联邦行业规范，促成多方联邦数据协议，达成标准化、协同化、规范化的联邦学习环境。

在信息流通日益渗透到企业和个人的今天，联邦学习将逐渐成为金融、保险、投资、医疗等众多行业领域实现商业价值和隐私安全保护的最佳途径。

联邦学习的技术框架建设方面。谷歌首先提出开源的离散数据联邦学习应用框架 TensorFlow Federated (TFF)。TensorFlow Federated 主要支持利用如今数量众多的移动智能终端设备和边缘端计算设备的计算能力，保证数据不离开本地的同时训练本地机器学习模型，通过 Google 开发的 Federated Averaging 算法，即使在较差的通信环境下，也能实现保密、高效、高质量的模型汇总和迭代流程，且移动端和边缘端用户体验上不做任何牺牲和妥协。目前 Google 已经将联邦学习应用在移动设备键盘输入预测上。


在学术研究与行业应用上，腾讯发起的中国首家互联网银行——微众银行正在积极探索。在国际人工智能专家、微众银行首席人工智能官杨强教授带领下的 AI 团队开源了首个联邦学习“FATE (FederatedAI Technology Enabler)”工业框架，作为安全计算框架支持联合 AI 生态系统，该框架可以实现基于同态加密和多方计算的安全计算协议，在信贷风控、客户权益定价、监管科技等领域推出了相应的商用方案。

目前，联邦学习的国际标准化工作正在进行，随着 6 月 15 日 IEEE 联邦学习基础架构与应用标准工作组的第二次会议在美国洛杉矶的召开。海内外 13 家来自科技、金融、教育、医疗等不同行业的知名研究机构及企业从多角度探讨联邦学习技术的应用案例，对联邦学习标准草案的制定提出建设性意见，该标准草案预计在一年内出台，意味着将为立法和监管提供更多技术依据。

项目采用的研究和实验方法
<p><b>Federated learning</b> 旨在打破数据孤岛，建立安全的数据生态。我们参考已有的研究方向和理论，进行进一步的优化和探索。</p> <p>隐私保护的参考理论基于“同态加密”和“差分隐私”。“同态加密”可以同时解决隐私和小数据两方面的问题，支持加法同态和乘法同态。不断增长的噪音（error）会在增长到一定程度之后，使得密文无法破解，也就无法实现全同态。我们需要实现在加密过程中降低噪音来使得加密的继续进行同时保证密文可破解。“差分隐私”旨在提供一种当从统计数据库查询时，最大化数据查询的准确性，同时最大限度减少识别其记录的机会。联系算法中的歧视和损失函数，将歧视来源分解为泛化误差中的模型拟合能力、模型的方法以及噪声。通过研究歧视的三种来源，来实现对算法公平的控制。</p> <p>基于理论进一步寻求降低同态加密中降低噪音的实现方法。针对差分隐私设计公平算法隐私保护算法，在模拟真实的数据集中进行测试，以期有所突破。</p>
研究内容、研究计划及预期目标（目标要具体，包括阶段目标和最终目标）
<p><b>研究内容：</b> 本项目研究了在各联邦与中心共享参数不需共享数据的情况下，即在保证用户隐私的情况下，增强联邦学习的抗攻击性，以此来合法地解决数据碎片和隔离问题。</p> <p><b>研究计划：</b> 1）（2019.12-2020.6）了解联邦学习的基本背景并对现如今对联邦学习的研究成果进行思考；学习机器学习和深度学习的基本知识，初步了解联邦学习的系统架构。 2）（2020.6-2020.12）研究横向、纵向联邦学习以及联邦迁移学习基本模型，深入研究联邦学习系统架构的各部分模型。</p> <p><b>预期目标</b> 1）设计一种健壮性更强，精确性更好且较不容易被攻击的联邦学习算法。 2）设计一个多个参与者联合数据建立虚拟的共有模型，使得它们能共同获益。 3）在某些特定的领域内，如金融机构（互联网银行等），设计出可以避免非授权的数据扩散和解决数据孤岛问题的算法并将其实现。</p> <p><b>阶段目标：</b> 1）完成机器学习，深度学习基础知识的学习，深入理解各种常见模型，如支持向量机、决策树等 1）完成对隐私保护机器学习即保护隐私的分布式协作机器学习的学习，包括用于纵向分区数据的安全多方决策树的算法、安全支持向量机算法、安全的多方梯度下降方法等。 3）完成横向与纵向联邦学习以及联邦迁移学习的基本模型的学习</p> <p><b>最终目标：</b> 本项目研究的最终目标是为研究人员和行业从业人员在设计更好的机器学习产品方面提供帮助。例如，本项目可以为基于保证用户隐私且不违反相关法规的联邦学习的产品提供完整的框架。</p>

现有研究基础及条件		
<p><b>现有研究基础：</b>团队成员熟练掌握一种以及上高级程序语言，例如 C++ , python 等；目前已经学习了数据结构、计算机组成原理等专业基础课；初步了解机器学习相关内容。且初步了解联邦学习的基础知识，了解联邦机器学习又名联邦学习，联合学习，联盟学习。联邦机器学习是一个机器学习框架，能有效帮助多个机构在满足用户隐私保护、数据安全和政府法规的要求下，进行数据使用和机器学习建模</p>		
经费预算（所需总费用：4400 元）		
费用项目	金额（元）	用途
1. 设备费	600 元	购置服务器，移动硬盘，u 盘，相关图书和专业资料
2. 会议费	300 元	主要用于会议的组织，笔，纸的购买。
3. 专家咨询费	500 元	我们将针对相关问题咨询领域专家，目前专家咨询费跟据专家级别在几百至几千一次。
4. 出版 / 文献 / 信息传播 / 知识产权事务费	3000 元	主要用于论文出版的版面费（1000-2000 元一篇），专利申请和保护（2500-3000）以及相关文献资料的购置。

### 三、指导老师及学院意见

第一 指导 老师	姓名	朱天清	性别	女	年龄	40	职称	教授
	院系	计算机学院		联系电话		17720484903		
其他 指导 老师	姓名	性别	年龄	职称	所在院系		联系电话	
请对申报者科研能力及项目可行性作评价		<p>申报成员积极、热情，态度认真，准备得当，具备初步课题研究的专业能力，联邦学习可使用的机器学习算法不局限于神经网络，还包括随机森林等重要算法。联邦学习有望成为下一代人工智能协同算法和协作网络的基础。</p>						
请对项目的意义、技术水平、适用范围及推广前景作评价		<p>联邦学习具有保护隐私和多方本地数据安全的极大优势。避免集中式存储数据，安全地从多源不互通的数据中创造新的价值，充分利用各方数据资源，优化机器学习训练结果，学习参与方可以在联合形成协同合作的联邦大数据环境，形成联邦学习生态。对金融、互联网、通信、零售、交通运输、工业生产等行业提供计算服务支持。</p>						
指导老师签名		<div style="text-align: center;">  </div> <div style="text-align: right;">           年月日         </div>						
学院学生科技工作领导小组意见		<div style="text-align: center;">             负责人签名：              （学院盖章）              年月日           </div>						

