基于联邦机器学习的睡眠质量预测

第 5 组　组员

姓名	学号
宣朋羽（组长）	202021080520
蒋育韬	202022080503
沙泽鑫	202022080521
王亦明	202022080502
张奔	202022080531

正文

随着智能设备的广泛普及应用，各大厂商陆续发布了智能手表、智能手环等穿戴设备，这些贴身设备几乎无不与健康相关。目前，智能穿戴设备能够精准且持续地采集用户的生理数据信息，如睡眠时长、睡眠心率[1]、运动时长[2]等等，运用这些数据预测生理状态能帮助我们维护身心健康。同时，由于这些数据的敏感性，如何避免信息泄露也成为了当前智能穿戴设备使用这些数据面临的问题。本项目拟使用神经网络对睡眠质量进行预测，并通过联邦机器学习在保护数据隐私的情况下提升模型质量，同时运用差分隐私进一步提高数据安全性。

随着生活节奏的日益加快，人们的工作压力也越来越大，很多人长期处于亚健康状态，还可能因睡眠得不到保障而埋下健康隐患。睡眠是人体自我更新过程中一个很重要的环节，是恢复体力、维持心理和生理健康的必要条件。良好的睡眠可使疲惫的人体更好地进行自我修复，从而保证精力的充沛；睡眠不佳则可能加剧人体疲劳，久而久之还容易引起其他的健康问题，如肥胖、抑郁症等。最近人们逐渐认识到睡眠健康的重要性，睡眠质量是人们预防疾病、调节生活方式的重要指标[3]。不过，普通用户在没有专业人士的引导下常常没法正确判断睡眠质量，因此通过智能设备预测睡眠质量具有重要意义。

然而，睡眠数据属于健康数据，与用户的隐私息息相关。从个人隐私来说，卡内基梅隆大学 Latanya Sweeney 的研究显示，基于邮编、性别、出生日期有 87%几率识别出唯一个人身份，有 18%几率还原出更多个人信息，比如住院信息，购物信息等等[4]。根据 R B Altuman 教授研究统计，只需 75 个统计上独立的但核算多态性位点即可唯一确定一个人[5]。通过比较易得的个人信息（性别，年龄，生日等）与健康信息（睡眠状况，病史信息

等）结合可以轻易的推断一个人的个人隐私甚至身体状况、，这无疑是对个人隐私的极大侵犯。从政策来说，国务院办公厅把生物学资源和医疗大数据作为国家的基础战略资源，并将要求对数据使用进行严格控制。虽然睡眠数据不完全属于医疗数据，但是在某些条例中，如欧盟的 GDPR（一般数据保护条例）要求对睡眠数据等健康数据进行保护，因此直接使用睡眠数据进行数据挖掘具有政策风险。但常规的脱敏方法很难做到对数据隐私性百分百的保护，而更加深度的脱敏方法会加重数据的扭曲和失真，导致数据可用性降低甚至出现负面效果。所以需要一种更有效的更具有隐私保护性质的方法，可以高效率高效用的利用数据。

现实生活中，除了少数巨头公司拥有海量数据，绝大多数企业都存在数据量少、数据质量差的问题，只使用本地数据的话，就会导致训练出的模型效果不理想。另外由于隐私保护、利益以及政策法规的限制，持有用户数据的各方也不愿意与他方分享数据，数据也往往以孤岛形式出现。想要打破数据孤岛这种局面就需要将各方数据综合起来，但如果数据持有方都将数据上传至一个平台，会使得海量数据面临泄露的风险。为了解决上述问题，有人提出联邦的学习方法，联邦机器学习使多方数据在保证数据安全和合法合规的基础上，共同建模、间接集合大量的数据，从而提升机器学习模型的效果[6]。

结合问题特点以及前人工作的铺垫，本项目拟采用联邦机器学习进行睡眠质量预测模型的训练，并辅以差分隐私提高数据安全性。联邦机器学习技术及数据隐私保护大会上明确提出了“联邦机器学习”这个概念[7]。联邦机器学习是一种可以保护数据隐私的分布式机器学习，用以解决数据孤岛问题[8]，由于其通过模型融合可以学到更多信息 ，从而能够获得更好的模型[9]。训练流程大致如下，客户端在本地使用本地数据进行训练，并上传每轮次模型更新的梯度，服务器融合梯度并下发新的全局模型。训练过程中，原始数据始终保存在本地，没有数据直接泄露的风险 ，同时在客户端上传的梯度中加入噪声以避免差分攻击，可以进一步提高数据的安全性[10]。

数据集取自 Kaggle，数据来自 ios 端的 sleep cycle app[11]。数据包含睡前备注、睡后心情、睡眠时长、心率、当日步数以及睡眠质量评分。模型的评估标准采用 MSE 指标，并拟通过比较联邦机器学习模型与数据分散式机器学习以及数据集中式机器学习的模型指标做进一步分析。

参考文献

[1] 梁超, 王鹏, 曹贝贝, 等. 可穿戴智能睡眠质量检测系统[J]. 电子测量与仪器学报, 2018, 32(5): 159-167.

[2] 韩富强. 常规智能设备用于走跑运动能耗监控的效果研究[D]. 西安体育学院, 2019.

[3] 杨天骁,李金宝,胡悦.基于 Jawbone 手环的睡眠质量预测研究[J].黑龙江大学工程学报,2015,6(01): 74-79.

[4] L Sweeney, JS Yoo.De-anonymizing south korean resident registration numbers shared in prescription data[J/OL].https://techscience.org,2015-09-29.

[5] Olga Troyanskaya, Michael Cantor, Gavin Sherlock,etc. [J]. Bioinformatics, 2001,17(6): 520-525.

[6] 知乎 沐清予. 详解联邦学习 Federated Learning [EB/OL] (2019-08-22)[2021-03-14].
https://zhuanlan.zhihu.com/p/79284686?utm_source=wechat_session

[7] CCF. 回顾：CCFTF14 联邦学习技术及数据隐私保护研讨会 [EB/OL] .(2019-03-26)[2021-03-16]. https://www.ccf.org.cn/c/2019-03-26/661203.shtml

[8] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-effificient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artifificial Intelligence and Statistics, pages 1273–1282, 2017 (original version on arxiv Feb. 2016).
https://www.jddglobal.com/newsDetail/725b38a1999e4864a2fadfd980f9dee1

[9] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. arXiv preprint arXiv:1912.04977, 2019.

[10]李效光,李晖,李凤华,朱辉. 差分隐私综述. 信息安全学报, 2018, 3(05): 92-104.

[11]Kaggle Dana Diotte. Sleep Data[DB/OL] .(2018-03-13)[2021-03-16].
https://www.kaggle.com/danagerous/sleep-data