|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **分类号** |  | **密级** |  | |
|  |  |  |  |  | |
|  | | | | | |
| **重庆邮电大学研究生学位论文**  **开题报告** | | | | | |
|  | | | | | |
|  | **中文题目** | **针对PU问题的多方联邦半监督推荐方法研究** | | |  |
|  |  | | |
| **英文题目** | **Research on Multi-Party Federated Semi-Supervised Recommendation Methods for PU Problems** | | |
|  |  | | |
| **学 号** | **S221201028** | | |
| **姓 名** | **吕九峦** | | |
| **学位层次** | **硕士研究生** | | |
| **学位类别** | **工学硕士** | | |
| **学科专业** | **软件工程** | | |
| **研究方向** | **视觉智能软件** | | |
| **指导教师** | **韦庆杰 正高级工程师** | | |
| **完成日期** | **2023年12月20日** | | |
|  | | | |

|  |
| --- |
| **一、选题依据** |
| **1.1 研究背景、目的和意义**  推荐系统[1]是一种信息过滤系统，旨在预测用户对物品的“评分”或“偏好”。在电子商务[2, 3]、新闻[4, 5]和医疗保健[6, 7]等多个现实世界的商业领域中，推荐系统都发挥着至关重要的作用。为了实现这些推荐，机器学习模型经常被采用来生成预测。然而，训练这些模型需要大量的数据，包括用户的敏感信息，如个人信息、行为、社会关系，以及上下文数据，如当前位置/时间/活动。为了达到高准确率，多数的推荐系统将数据存储在中央服务器上。这种集中式的存储方法可能导致严重的隐私风险，如数据被盗、丢失和未经授权的使用。此外，由于各种监管限制，以及隐私数据由多个不同参与方拥有，开发跨多方数据集的推荐系统变得越来越具有挑战性。  联邦学习(Federated Learning, FL)[8]是谷歌公司提出的一种分布式学习方法，它为推荐系统提供了一种在保护隐私的前提下进行模型训练的解决方案。 该方法的核心思想在于，不同的数据拥有方(通常被称为客户端)可以通过共享模型参数和梯度来协作训练机器学习模型，而无需共享原始数据。目前，联邦学习的研究主要集中在有监督学习领域，以解决具有完全标注数据的问题。然而，在现实世界中，由于受到领域专业知识限制、资源的有限性或缺乏适当的标注工具等因素的影响，很难获取到完全标注的数据。  联邦半监督学习(Federated Semi-Supervised Learning, FSSL)[9]旨在解决标注数据有限的联邦学习问题，特别是在只涉及正样本(Positive)和未标记样本(Unlabeled)的情况下，也称为PU问题[10]。目前，关于联邦半监督学习的研究主要集中在横向联邦学习(Horizontal Federated Learning, HFL)[11]，在这种情况下，各方共享相同的特征空间，但样本ID空间不同。然而，在纵向联邦学习(Vertical Federated Learning, VFL)[11]，即各方共享相同的样本ID空间，但特征空间不同的情况下，应用半监督技术的研究较少。  我们探讨以下VFL场景： 1)存在多个参与方，各方拥有的数据集样本 ID 部分重叠，但特征空间不同；2)需要服务推荐的一方仅拥有少量正样本(Positive)，而其他各方拥有未标记样本(Unlabeled)；3)各方旨在共同训练一个推荐模型，该模型能够从未标记样本中识别可靠的正样本。  值得注意的是，上述第二个场景对应于PU问题，即从正样本(Positive)和未标记样本(Unlabeled)中学习的问题。传统的PU学习方法[10, 12-14]旨在有效利用未标记样本来解决PU问题。然而，当一方只拥有正样本，而无法获得其他方持有的未标记样本时，这又是一个新的挑战。我们将其称为未标记样本不足(Unlabeled-Data-Deficient)的PU学习问题(UDD-PU)，传统的 PU 学习方法无法解决这个问题。  UDD-PU问题是纵向联邦学习背景下的一个新挑战，通过解决UDD-PU问题，为电子商务、新闻推荐、医疗保健等领域的推荐系统提供一种新的解决方案，有助于提高推荐的准确性和用户满意度，同时保护用户的隐私和数据安全。  **1.2 国内外研究现状**  1)联邦学习：联邦学习(FL)最初由 McMahan 等人[8]提出，旨在促进协作式模型学习，而无需从个体收集数据。在具体实践上，此方法主要是在具有一致模型架构的本地数据上进行本地模型的训练，并通过平均所有本地模型来实现全局模型的更新。此外，联邦学习也强调在单个移动终端进行模型训练，以尽量避免数据的传输[15-17]。自从这一方法的提出，研究人员已经针对不同的联邦设置调整了多种机器学习模型，其中包括决策树[18, 19]、线性/逻辑回归[20-22]和神经网络[23, 24]。  Bonawitz 等人[16]引入了一种基于安全多方计算的聚合方法，以提高联邦学习的安全性。此方法可以在多个互不信任的参与方之间聚合隐私值，同时确保不会泄露各自隐私值的任何信息。另外，Geyer 等人[25]还结合了客户级差分隐私的概念，以进一步加强隐私保护，防止有人尝试利用全局模型来重建其他客户的隐私数据。除此之外，同态加密(Homomorphic Encryption, HE)[26] 也是一种保护用户数据隐私的有效方法，它允许在加密状态下进行参数交换[47]。  纵向联邦学习(VFL)的概念首次被Yang等人[11]提出，并针对线性模型[27, 28]和神经网络[29]设计了相关协议，早期的研究还探索了如何在垂直分区数据上实现隐私保护的决策树[30]。此外，Djatmiko等人[31]建议了一种在加密的垂直分区数据上联合运行逻辑回归的方法，该方法使用泰勒展开来近似非线性逻辑损失，但可能会对模型性能产生一定影响。梯度提升决策树(GBDT)[32]是机器学习和数据科学领域广泛使用的集成学习方法。目前，已经产生了基于GBDT的多种VFL算法，例如SecureBoost [33]、SecureBoost+ [34]和SecureGBM [35]。这些算法利用加法同态加密(HE)技术来保护主动方和被动方之间传输的残差和特征直方图。同时，隐私集合求交(Private Set Intersection , PSI)也被广泛用于VFL中的隐私保护实体对齐，以确保各方可以在不泄露任何额外信息的情况下确定共同的ID交集。PSI协议可以通过加密和签名策略、遗忘传输等多种技术来实现[36, 37,48]。  2)半监督学习：半监督学习(Semi-supervised learning, SSL)是一种机器学习方法，它同时利用标记数据和未标记数据进行模型训练。PU学习方法是一种特殊的半监督二值分类方法，适用于仅有正样本和未标记样本的场景。为了克服这种学习场景带来的挑战，研究人员提出了各种算法和技术。其中，Liu等人[10]提出的两步法技术是一种常用的方法，它基于正样本与负样本差异显著的假设，通过识别可靠的负样本和选择性地生成额外的正样本来提高学习效果。另外，有偏PU学习方法是另一类用于PU学习的算法，它将未标记的样本视为带有类标签噪声的负样本。具体来说，Liu和Tao[13]提出了一种方法，将未标记样本视为带有标签噪声的负样本，而Liu等人[10]则将未标记数据视为权重较小的负样本，并在对样本进行加权处理后应用逻辑回归算法。  负样本数据中存在的噪声会增加学习过程的复杂性，因为它可能会导致实际上是正样本的数据被赋予过多权重[46,49,50]。为了应对这个问题，研究人员提出了多种技术，其中包括装袋法(Bagging)和最小平方SVM(LS-SVM)[38]。装袋SVM是一种训练多个有偏差的SVM分类器的方法，每个分类器都针对正样本和样本的子集进行训练。Mordelet和Vert进一步扩展了这种方法，他们在装袋SVM的基础上对正负样本进行了重采样，并通过引入引导策略来提高算法的性能。  3)联邦半监督学习：当前的联邦半监督学习方法多是基于横向联邦学习架构，在联邦半监督研究领域，研究者们已经提出了多种方法来解决标签隔离、数据异质性等问题。根据标签的位置，我们可以将这些方法分为两种情况：标签在客户端和标签在服务器端[39]。  当标签位于客户端时，RSCFed[40,51,52]、FedSSL[41]、FedPU[42]、AdaFedSemi[43,53]和DS-FL[44,55]等方法都各自提出了相应的解决方案。具体而言，RSCFed采用师生模型进行局部训练，并提出了子共识抽样法和距离加权聚合法来处理数据异质性。FedSSL则通过伪标记技术和全局生成模型来解决标签隔离和数据异质性问题，并引入了差分隐私技术以确保隐私数据的安全。FedPU专注于解决PU学习中的标签隔离问题，通过提出新的目标函数，使客户端能够专注于学习正类。AdaFedSemi和DS-FL则通过利用服务器端的无标记数据，采用伪标注和集合伪标签技术，以在效率和模型准确性之间找到平衡点。  另一方面，当标签位于服务器端时，SemiFL[45]和FedMatch[9]提出了解决方案。SemiFL利用标注数据对全局模型进行微调，以提高模型的质量，并减少客户端在无监督训练过程中的遗忘问题。FedMatch则通过引入客户端间一致性损失和分离式学习方法，解决了数据异质性问题，并降低了通信成本。这些方法展示了在联邦半监督学习中处理标签隔离和数据异质性问题的不同策略，为实际应用提供了有益的参考。  **1.3 当前存在的主要问题**  ①如何在特征空间不同且部分数据未标记的情况下进行分类：目前，联邦半监督学习方法多是基于横向联邦学习架构，即特征空间相同，但样本ID空间不同。然而，在真实的应用场景中，不同的参与方可能拥有样本ID空间相同，但特征空间不同的数据。特别地，只有一方拥有带标签的正样本数据，而其它方的数据未标记。这些参与方期望通过协同合作，利用各自的数据集训练出一个高效的推荐模型，从而从未标记的数据中识别出有价值的正样本，现有的方法无法解决。这给我们提出了一个难题，即如何利用联邦半监督学习，在特征空间不同且部分数据未标记的情况下进行分类，来优化用户推荐并克服这一挑战。  ②如何将现有的联邦半监督学习方法扩展至表格类数据：目前，许多半监督学习方法主要集中于图像数据的应用。但在大数据领域，我们面对的数据类型丰富多样，如图像、表格等。因此，如何在半监督学习的框架下应用并优化这些方法成为了一个紧迫的问题。尤其是针对表格类数据，基于图像的方法可能并不适用，这要求我们开发专门针对表格数据的半监督学习方法。  ③联邦学习架构的完善和优化：在研究联邦半监督学习的过程中，联邦学习的具体执行过程和算法细节成为了我们面临的关键难题。例如，如何构建更为高效的数据交换模式以加强各方之间的协同合作，以及如何优化联邦学习的整体架构来提升其性能和效率，都是我们必须解决的重要议题。对这些问题的深入研究将有助于推动联邦半监督学习领域的持续发展，并在实际应用中实现更好的性能表现。 |
| **二、研究目标和主要研究内容** |
| **2.1 研究目标**  本文的研究目标是在纵向联邦学习(VFL)设置下，利用多方数据解决UDD-PU问题，使各方能够在保护数据隐私和安全的同时，协同训练机器学习模型，实现多方联邦半监督推荐。  **2.2 主要研究内容**  **(1)针对PU问题的半监督学习方法研究**  为了有效解决UDD-PU问题，我们首先在未标记样本充足、无需多方协同建模的情况下进行研究。具体来说，我们深入调查了针对PU问题的各种半监督学习方法，并提出了我们自己的解决方案。为确保这些方法在解决PU问题上的效果，我们会对其进行详尽的性能比较和评估。这一阶段的研究将后续处理更为复杂的UDD-PU问题奠定坚实的基础。  **(2)针对UDD-PU问题的多方半监督推荐方法研究**  在研究内容(1)的基础上，我们进一步筛选出了几种在解决PU问题上表现卓越的方法，并对这些方法进行深度分析和理解。接下来，我们将这些方法进行改造，使它们能够适用于纵向联邦学习环境，以便用于解决更为复杂的UDD-PU问题，其中我们提出的方法被称为VFPU(Vertical Federated Learning with Positive and Unlabeled data)。  这一过程包括了对所选方法的理论性质的深入理解，对算法性能的细致评估，以及对算法在不同数据集上的实验验证等多个步骤。我们的目标是，通过对这些方法的改造，使它们能够在多方数据环境中，有效地识别出可靠的正样本，从而提高推荐的准确性。  同时，我们还考虑到数据隐私和安全的问题。因此，我们的改造也包括了数据隐私保护和安全防护的设计。我们希望通过这些设计，使得各方能够在保护自己数据隐私和安全的前提下，共同训练出一个高效的推荐模型。  最后，我们将通过一系列实验，验证VFPU和其它基线方法在解决UDD-PU问题上的效果，包括它们的分类准确率，计算效率。我们期望通过这些实验，能够为解决UDD-PU问题提供有效的方法和实践指导。  **(3)VFPU的改进与优化**  近年来，基于深度学习的半监督学习方法在图像和文本数据领域取得了令人瞩目的成果。然而，在处理表格类数据时，尤其是在多方半监督推荐场景中，这些方法的应用尚不广泛。究其原因，表格类数据与图像和文本数据在结构上存在显著差异，导致许多基于深度学习的半监督学习方法无法直接适用于表格类数据。因此，我们打算对这些基于深度学习的半监督学习模型进行改进，以契合表格类数据的特性，并将这种优秀的特性融合到VFPU中，对VFPU方法进行进一步的完善和优化。  目前，VFPU方法仅适用于二元类分类PU问题，然而，在某些情况下，会产生多正类多负类PU(Multiple-Positive-Multiple-Negative, MPMN-PU)学习问题[42]。因此，扩展VFPU方法以处理MPMN-PU问题将是我们未来另一个重要的改进方向。 |

|  |
| --- |
| **三、拟解决关键问题及其研究方法** |
| **3.1 拟解决关键问题**  1）如何设计和应用合适的半监督学习算法，以利用未标记的样本进行模型的训练？  2）在多方环境下，如何设计或者引入一种学习机制，并结合半监督方法，使得多个参与方能够协同训练，共同构建一个机器学习模型？  **3.2 采取的研究方法**  针对研究内容(1)，我们提出了PU Bagging+两步法的方案来解决PU问题。  首先，我们得到一个初始的数据集，里面包含少量正样本以及大量的正样本标签。然后，我们PU Bagging方法对数据集进行采用训练预测，从未标记样本中的随机采样，采样集的大小与正样本数量相同，将采样集中的未标记样本看作负样本，采样集与所有正样本构成一个平衡的二分类训练集，得到每个未标记样本的得分，这个得分是指该未标记样本预测为正样本的概率。根据这个得分，可以将未标记样本中选择部分可靠正样本加入到正样本数据集中，并不断重复这个过程。  针对研究内容(2)，我们提出了VFPU解决方案，该方法包括数据预处理和样本加密对齐、执行VFPU算法核心流程两个阶段。具体地，VFPU从多方未标记数据中反复进行随机抽样，将抽样到的数据视为负样本。因此，它形成了多个正负样本均衡的训练数据集，以及多个包含这些未标记样本的测试数据集。对于每个训练数据集，VFPU都会反复训练一个基于纵向联邦架构的基学习器。我们使用训练好的基学习器为测试数据集中的每个样本生成预测分数。根据分数总和及其在测试数据集中的出现频率，我们计算出每个未标记样本的为正样本的概率。概率最高的样本被视为可靠的正样本。这些样本会被添加到正样本集合中，然后从未标记数据中删除。采样、训练和选择可靠正样本的过程反复进行。  针对研究内容(3)，对于VFPU的改进与优化，我们准备通过文献调研法研究半监督学习领域最新的模型，将这些模型改进使其适用于表格类数据以及解决PU问题，然后通过实验评估这些最新的模型在解决PU问题上的性能，进一步对VFPU进行改进与优化。  **3.3 可能的创新**  1)识别UDD-PU学习问题：未标记样本不足的PU学习问题(UDD-PU)，这是纵向联邦学习背景下的一个新挑战。通过认识这一问题，填补了现有文献的空白，并提出了解决这一问题的方案。  2)提出VFPU算法解决UDD-PU问题：该算法能让多方以保护隐私的方式协作训练机器学习模型。该算法包括随机抽样、平衡正负样本训练和迭代选择可靠的正样本。  **参考文献：**   1. Das D, Sahoo L, Datta S. A survey on recommendation system[J]. International Journal of Computer Applications, 2017, 160(7). 2. SARWAR B, KARYPIS G, KONSTAN J, et al. Analysis of recommendation algorithms for e-commerce; proceedings of the Proceedings of the 2nd ACM Conference on Electronic Commerce, F, 2000 [C]. 3. Schafer J B, Konstan J A, Riedl J. E-commerce recommendation applications[J]. Data mining and knowledge discovery, 2001, 5: 115-153. 4. ZHENG G, ZHANG F, ZHENG Z, et al. DRN: A deep reinforcement learning framework for news recommendation; proceedings of the Proceedings of the 2018 world wide web conference, F, 2018 [C]. 5. Liu J, Dolan P, Pedersen E R. Personalized news recommendation based on click behavior[C]//Proceedings of the 15th international conference on Intelligent user interfaces. 2010: 31-40. 6. Kim J, Lee D, Chung K Y. Item recommendation based on context-aware model for personalized u-healthcare service[J]. Multimedia Tools and Applications, 2014, 71: 855-872. 7. Yue W, Wang Z, Zhang J, et al. An overview of recommendation techniques and their applications in healthcare[J]. IEEE/CAA Journal of Automatica Sinica, 2021, 8(4): 701-717. 8. MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data; proceedings of the Artificial intelligence and statistics, F, 2017 [C]. PMLR. 9. Jeong W, Yoon J, Yang E, et al. Federated semi-supervised learning with inter-client consistency & disjoint learning[J]. arXiv preprint arXiv:2006.12097, 2020. 10. LIU B, DAI Y, LI X, et al. Building text classifiers using positive and unlabeled examples; proceedings of the Third IEEE international conference on data mining, F, 2003 [C]. IEEE. 11. Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.. 12. Mordelet F, Vert J P. A bagging SVM to learn from positive and unlabeled examples[J]. Pattern Recognition Letters, 2014, 37: 201-209. 13. Liu T, Tao D. Classification with noisy labels by importance reweighting[J]. IEEE Transactions on pattern analysis and machine intelligence, 2015, 38(3): 447-461. 14. XU Y, XU C, XU C, et al. Multi-Positive and Unlabeled Learning; proceedings of the IJCAI, F, 2017 [C]. 15. Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint arXiv:1610.05492, 2016. 16. BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning; proceedings of the proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, F, 2017 [C]. 17. Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1–2): 1-210. 18. ZHAO L, NI L, HU S, et al. Inprivate digging: Enabling tree-based distributed data mining with differential privacy; proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications, F, 2018 [C]. IEEE. 19. LI Q, WEN Z, HE B. Practical federated gradient boosting decision trees; proceedings of the Proceedings of the AAAI conference on artificial intelligence, F, 2020 [C]. 20. Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine learning and systems, 2020, 2: 429-450. 21. Hanzely F, Hanzely S, Horváth S, et al. Lower bounds and optimal algorithms for personalized federated learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 2304-2315. 22. MOHRI M, SIVEK G, SURESH A T. Agnostic federated learning; proceedings of the International Conference on Machine Learning, F, 2019 [C]. PMLR. 23. YUROCHKIN M, AGARWAL M, GHOSH S, et al. Bayesian nonparametric federated learning of neural networks; proceedings of the International conference on machine learning, F, 2019 [C]. PMLR. 24. Wang H, Yurochkin M, Sun Y, et al. Federated learning with matched averaging[J]. arXiv preprint arXiv:2002.06440, 2020. 25. Geyer R C, Klein T, Nabi M. Differentially private federated learning: A client level perspective[J]. arXiv preprint arXiv:1712.07557, 2017. 26. Rivest R L, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms[J]. Foundations of secure computation, 1978, 4(11): 169-180. 27. Hardy S, Henecka W, Ivey-Law H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption[J]. arXiv preprint arXiv:1711.10677, 2017. 28. Liu Y, Kang Y, Zhang X, et al. A communication efficient collaborative learning framework for distributed features[J]. arXiv preprint arXiv:1912.11187, 2019. 29. Liu Y, Kang Y, Xing C, et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems, 2020, 35(4): 70-82. 30. Vaidya J, Clifton C, Kantarcioglu M, et al. Privacy-preserving decision trees over vertically partitioned data[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2008, 2(3): 1-27. 31. Djatmiko M, Hardy S, Henecka W, et al. Privacy-preserving entity resolution and logistic regression on encrypted data[J]. Private and Secure Machine Learning (PSML), 2017. 32. Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232. 33. Cheng K, Fan T, Jin Y, et al. Secureboost: A lossless federated learning framework[J]. IEEE Intelligent Systems, 2021, 36(6): 87-98. 34. Chen W, Ma G, Fan T, et al. Secureboost+: A high performance gradient boosting tree framework for large scale vertical federated learning[J]. arXiv preprint arXiv:2110.10927, 2021. 35. FENG Z, XIONG H, SONG C, et al. Securegbm: Secure multi-party gradient boosting; proceedings of the 2019 IEEE international conference on big data (big data), F, 2019 [C]. IEEE. 36. PINKAS B, SCHNEIDER T, ZOHNER M. Faster private set intersection based on {OT} extension; proceedings of the 23rd USENIX Security Symposium (USENIX Security 14), F, 2014 [C]. 37. Pinkas B, Schneider T, Zohner M. Scalable private set intersection based on OT extension[J]. ACM Transactions on Privacy and Security (TOPS), 2018, 21(2): 1-35. 38. Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9: 293-300. 39. Jin Y, Liu Y, Chen K, et al. Federated Learning without Full Labels: A Survey[J]. arXiv preprint arXiv:2303.14453, 2023. 40. LIANG X, LIN Y, FU H, et al. Rscfed: Random sampling consensus federated semi-supervised learning; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2022 [C]. 41. FAN C, HU J, HUANG J. Private semi-supervised federated learning; proceedings of the International Joint Conference on Artificial Intelligence, F, 2022 [C]. 42. LIN X, CHEN H, XU Y, et al. Federated learning with positive and unlabeled data; proceedings of the International Conference on Machine Learning, F, 2022 [C]. PMLR. 43. WANG L, XU Y, XU H, et al. Enhancing Federated Learning with In-Cloud Unlabeled Data; proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), F, 2022 [C]. IEEE. 44. 王鑫,黄伟口,孙凌云.跨机构联邦学习的激励机制综述[J/OL].计算机科学,1-14 45. 穆旭彤,程珂,宋安霄等.抗拜占庭攻击的隐私保护联邦学习[J/OL].计算机学报,1-20[2024-01-18]. 46. 王国伟.基于半监督深度学习的信用预测优化方法[D].武汉理工大学,2021.DOI:10.27381/d.cnki.gwlgu.2021.001570. 47. 张哲.非独立同分布数据下半监督联邦学习的研究及应用[D].黑龙江大学,2023.DOI:10.27123/d.cnki.ghlju.2023.001577. 48. 钱栋炜.基于信息熵的半监督联邦学习性能优化[D].华东师范大学,2023.DOI:10.27149/d.cnki.ghdsu.2023.00104. 49. 于宏正.基于半监督联邦学习的行为识别研究[D].山东大学,2022.DOI:10.27272/d.cnki.gshdu.2022.002696. 50. LI J, XIONG C, HOI S C. Comatch: Semi-supervised learning with contrastive graph regularization; proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2021 [C]. 51. 邓丽钰.针对纵向联邦学习的半监督标签推断攻击及其防御方法研究[D].西安电子科技大学,2022.DOI:10.27389/d.cnki.gxadu.2022.000556. 52. 史鼎元,王晏晟,郑鹏飞等.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.DOI:10.13328/j.cnki.jos.006174 53. 高迢康,靳晓宁,赖英旭.模型异构的联邦学习入侵检测[J/OL].北京工业大学学报,1-15[2024-01-18] |

|  |
| --- |
| **四、进度安排和预期成果** |
| **4.1 预期成果**  目前已有一篇论文(Multi-party Federated Recommendation Based on Semi-supervised Learning)，被IEEE Transactions on Big Data期刊（SCI 2区）接收  后续针对UDD-PU这个点继续深挖，对VFPU进行改进与优化，争取发表一篇一区的论文。  **4.2 进度安排**   |  |  |  | | --- | --- | --- | | 时间段 | 工作内容 | 预期成果 | | 2023年11月30日——2023年12月30日 | 工作进展阶段一检查 | 对于前期工作开展情况进行汇总，并进行研究现状分析，对前期方案进行修订和扩展，完成成果提交检查。 | | 2024年1月1日——2024年1月15日 | 学位论文初稿 | 完成学位论文理论部分初稿，完成附加实验结果检查。 | | 2024年1月16日——2024年3月10日 | 工作进展阶段二检查 | 对于前期工作开展情况进行汇总，并进行差距分析，对前期方案进行修订和扩展，研究室内部开展研究成果检查，学位论文提纲定稿，指导教师签署通过 | | 2024年3月11日——2024年6月30日 | 学位论文及成果检查 | 完成学位论文初稿，并提交指导教师修改未提交，正式中期检查；成果检查确认正式成果物获得情况；向导师汇报职业方向。 | | 2024年7月1日——2024年11月10日 | 学位论文完稿 | 完成学位论文第一稿，达到送出查重、盲审标准，实验室内部开展查重，指导教师签署通过 |   **4.3 研究条件**  4.3.1软硬件条件  (1) 硬件条件  本研究实验环境为操作系统为Windows 10家庭中文版，处理器AMD Ryzen 7 5800H CPU @ 3.20GHz ，已安装16G内存，显卡为NVIDIA GeForce RTX 3060，系统类型为64位操作系统，基于x64的处理器。这些实验环境为搭建机器学习模型提供了强大的计算能力，能够很好满足其计算能力。同时支撑本研究的实验室有英伟达(NVIDIA)公司出品的GTX1660Ti,GTX1070和RTX2070图形加速卡各一个，以及集合两块英伟达旗舰级图形加速卡GTX1080ti的服务器一台，可以完成多个参与方的搭建，在联邦学习环境下运行实验。  (2) 软件条件  本研究需要采用Python语言进行代码的编写，选择的集成开发环境为Pycharm。搭建机器学习模型与神经网络的开源框架可选择TensorFlow框架与Pytorch框架。对联邦学习环境的搭建可选择FATE框架或FederatedScope框架。  4.3.2文献资料及数据来源  (1) 文献资料来源  学校图书馆和网上数字图书馆收集了顶会SIGMOD、SIGKDD、IJCAI等数据挖掘类与联邦学习类的权威期刊和大量的期刊论文、会议论文、专利、学位论文、电子图书、IEEE/IEL数字资源库、SCI-科学引文索引、ElseierSD资源数据库。可以网上浏览，也可以从远程数据库下载，学校丰富的数字资源完全能够满足我对期刊论文、会议论文、电子图书文献资料的查阅下载的需求，以使得有充足的文献资料支持我们的研究。  (2) 数据集来源  本实验的数据集来自一些数据挖掘与推荐系统相关论文中常用的数据集，常常将数据集按样本横向拆分或是按特征纵向拆分将其使用于联邦学习的研究中。具体数据集如下表所示。   | **数据集** | **样本数(行)** | **数值型特征(列)** | **类别型特征(列)** | | --- | --- | --- | --- | | Bank | 4.1M | 7 | 13 | | Credit | 4M | 10 | 14 | | AutoML-B | 0.82M | 17 | 7 | | Amazon | 32769 | 0 | 10 | | Adult | 48842 | 6 | 9 |   **Bank：**这是一个与葡萄牙银行直接营销活动相关的数据集。营销活动通过电话呼叫进行。它包括四个数据集，我们使用的是"bank-additional-full"数据集，包含41,188个样本和20个输入/特征。数据按日期排序，时间跨度从2008年5月到2010年11月。分类的目标是预测客户是否会订阅(是/否)定期存款，由变量"y"表示。  **Credit：**这个数据集包含40,000个信用卡客户的数据和24个变量，包括人口统计学和支付历史信息。该数据集用于构建预测模型，评估客户在下个月是否会违约贷款的可能性，由一个二进制目标变量表示违约状态。征，54个类别型特征。  **Malware：**原始数据位于https://www.kaggle.com/c/malware-classification。这是由微软提供的恶意软件数据集，根据数据集中软件相应的特征信息，来判断他们是属于什么类型的恶意软件。Malware数据集包含892万个样本，12个数值型特征，69个类别型特征。  **Amazon：**原始数据位于https://www.kaggle.com/competitions/amazon-employee-access-challenge/overview。由亚马逊公司提供的二分类数据集，通过数据集中员工的特征信息，判断员工是否应当被授予操作权限。Amazon数据集包括32769个样本，10个类别型特征。  **Adult：**原始数据位于https://archive.ics.uci.edu/ml/machine-learning-databases/adult/，是一个二分类数据集，目的为判断个人的收入是否大于50000美元。此数据集由Barry Becker从1994年美国人口普查数据库中提取的。Adult数据集包含48842个样本，包含6个数值型特征，8个类别型特征。 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **论文开题管理** | | | | | |
| 版本说明 | 提交版本：首次提交 | | | | |
| 学生承诺 | 学生签字： 年 月 日 | | | | |
| 导师意见 | 导师签字： 年 月 日 | | | | |
| 开  题  组  成  员 |  | 姓 名 | 职 称 | | 所 在 单 位 |
| 组 长 |  |  | |  |
| 成 员 |  |  | |  |
| 成 员 |  |  | |  |
| 成 员 |  |  | |  |
| 成 员 |  |  | |  |
| 开  题  报  告  组  意  见 | 组长签字： 年 月 日 | | | | |
| 导师团队意见 | 负责人： 年 月 日 | | | 学院意见 | 院长： 年 月 日 |