《PDMFRec: A Decentralised Matrix Factorisation with Tunable User-centric Privacy》复现报告

论文作者：Duriakova E, Tragos E Z, Smyth B, et al.

复现：李志涛

摘要

随着人工智能的发展，推荐系统在人们的生活中扮演着越来越重要的角色。一般情况下，推荐系统是数据驱动的，它的性能依赖于大规模数据的训练。然而，随着人们对个人隐私和数据安全的关注，个性化推荐与用户隐私保护之间的平衡变得越来越重要。联邦机器学习提出来应对上述挑战。为了保护推荐系统的用户隐私，许多研究人员将联邦学习引入推荐系统，并提出保护隐私的推荐系统模型（联邦推荐）。联邦推荐的研究分为两个方向：有服务器的联邦推荐和无服务器的联邦推荐。在无服务器的联邦推荐上，隐私增强型的去中心化分布式矩阵分解模型（PDMFRec）是一个具有以用户为中心和隐私增强的矩阵分解方法。它基于邻域梯度权重共享方法使模型收敛。在无服务器的联邦推荐研究中取得优秀的效果。本文研究并复现 PDMFRec，复现结果与原论文不一致，最后我对算法收敛性和复现结果为什么不一致进行分析。

关键词：联邦学习；推荐系统；协同过滤

1 引言

随着人工智能的发展，推荐系统在现代社会中扮演着越来越重要的角色，被广泛应用于商品、新闻、广告等领域。推荐系统可以根据用户的信息，即用户信息 (如身高、年龄) 和用户与物品之间的相互信息 (如评分记录)，可以帮助用户挖掘出最有用的信息，为用户提供个性化的推荐。其中，协同过滤是一种非常流行的推荐技术，并已广泛应用于各种商业应用中。

一般情况下，推荐系统是数据驱动的，它的性能依赖于大规模数据的训练。传统上，为了训练一个集中的推荐模型，组织采用数据共享策略，在大多数情况下从多个来源聚合用户数据，而不需要用户的确认。然而，随着人们对个人隐私和数据安全的关注，这样的数据聚合被许多法规和法律禁止。例如，近年来，随着数据安全和用户隐私法的颁布，以及隐私监管的加强，个性化推荐与用户隐私保护之间的平衡变得越来越重要。例如，欧盟的《通用数据保护条例》（GDPR）定制了一套有关用户数据隐私的规定，用户必须遵守这些规定。

联邦机器学习提出来应对上述挑战。联邦学习的主要思想是共享对隐私不敏感的知识 (如加密梯度和中间结果)，而不是原始数据。它在保护用户隐私的前提下，它可以有效地帮助多个组织利用不同的数据，并构建一些机器学习模型。它已被应用于一些高度关注隐私的领域，例如医疗保健，以及大量分布式数据的问题，例如车辆网络和边缘计算。为了保护推荐系统的用户隐私，许多研究人员将联邦学习引入推荐系统，并提出保护隐私的推荐系统模型，这也导致了对联邦推荐的一系列研究，联邦推荐是联邦学习和推荐系统的跨学科领域。对解决用户的隐私问题有重要意义。

2 相关工作

2.1 集中式的联邦推荐

FCF[1] 将所有未评分的项目视为负面反馈，这将导致模型训练中的偏差和高昂的通讯成本。FedRec[2] 会对一些未评分的项目进行采样，以保护用户的评分行为，即每个用户已评分的项目，但这会在模型训练中带来噪音，并导致推荐性能下降。FedRec++[3] 扩展了 FedRec[2]，并通过分配一些去噪客户端来提出无损联邦推荐框架，以保护隐私的方式消除 FedRec[2] 混合填充方法引入的噪音。FedMF[4]提出了一个矩阵分解的联合推荐模型，它通过同态加密来保护用户的原始评分记录。FedGNN[5]首先将 GNN 应用于联合推荐。它通过同态加密计算邻居，并通过随机抽取一些伪互动项目来保护用户的评分行为。还有一些其他的集中式联合推荐方法[6-14]，而我们则专注于去中心化的联邦推荐。

2.2 去中心化的联邦推荐

Hegedűs 等人[15] 提出了一种基于八卦的去中心化推荐算法，这是一种随机选择多个节点并定期发送信息的通信协议，但会导致额外的通信成本。DMF[16] 根据地理位置计算邻域用户的权重，并使用随机游走选择一些相邻节点来发送全局项目特定的梯度。但是，这种方法会泄露用户的地理位置和评分行为的隐私。PDMFRec[17] 计算用户之间的余弦相似性，然后根据相似性选择相邻项以发送梯度。PDMFRec[17] 是解决 DMF 隐私保护问题的好办法，但计算用户的相似性也会泄露用户的隐私。需要注意的是，去中心化的联邦推荐仍然需要服务器的帮助，例如计算相邻用户的权重[16] 和用户之间的余弦相似性[17] 。但是，去中心化联邦推荐中的服务器仅在训练过程之前工作，并且仅执行一些轻量级辅助工作。

3 本文方法

3.1 本文方法概述

PDMFRec[17]是一个分散的推荐框架，具有以用户为中心和隐私增强的矩阵分解方法。与 DMF[16]类似，它通过共享邻域梯度的方法来训练模型。在 PDMFRec 中，用户 u 对物品 i 的预测评分是通过本地物品特定的潜在特征向量 $V_{i\cdot}^{u} \in \mathbb{R}^{1\times d}$ 和用户特定的潜在特征向量 $U_{u\cdot} \in \mathbb{R}^{1\times d}$ 的内积计算的，即。$\hat{r}_{ui} = U_{u\cdot}V_{i\cdot}^{uT}$ 。具体来说，PDMFRec 根据两个相应用户的两套评分项目计算余弦相似度。在模型训练过程中，每个用户 u 选择相似度大于某一阈值的邻居，然后发送和接收特定项目的梯度。训练过程中，PDMFRec 的单向通信是匿名的，可以很好地保护用户评分行为的隐私。

3.2 符号解释

我们将一些用到的符号解释放在表 1 中。

3.3 目标函数

要最小化的目标函数如下,

$$\min_{\Theta} \sum_{u=1}^{n}\sum_{i=1}^{m} y_{ui} f_{ui} \tag{1}$$

其中 $f_{ui} = \frac{1}{2}(r_{ui} - \hat{r}_{ui})^2 + \frac{\alpha}{2}\|U_{u\cdot}\|^2 + \frac{\alpha}{2}\|V_{i\cdot}^{u}\|^2$。

表 1: 一些记号及其解释。

符号	解释
n	客户端数量 (i.e., 用户)
m	物品数量
$\mathfrak{R} = \{1, \dots, 5\}$	评分范围
$r_{ui} \in \mathfrak{R}$	用户 u 对物品 i 的评分
$\mathcal{R} = \{(u, i, r_{ui})\}$	训练数据的评分记录
$\mathcal{R}_u$	用户 u 在训练数据 $\mathcal{R}$ 的评分记录
$\mathcal{R}^{te} = \{(u, i, r_{ui})\}$	测试数据的评分记录
$\mathcal{I}$	全部物品集合
$\mathcal{I}_u$	用户 u 的物品集合
$\mathcal{N}_u^j$	最接近用户 u 并且在第 t 次迭代中发送梯度 ∇V_j 给用户 u 的领居集合
$\mathcal{N}_u$	最接近用户 u 的领居集合
$\mathcal{I}_{\mathcal{N}_u}^t$	在第 t 次迭代中用户 u 从它的领居中接受到的物品集合
$\mathbf{W}$	(用户, 用户) 相似度矩阵
$d \in \mathbb{R}$	特征向量的维度
$U_{u\cdot} \in \mathbb{R}^{1\times d}$	用户特征向量
$V_{i\cdot} \in \mathbb{R}^{1\times d}$	物品特征向量
$\hat{r}_{ui}$	用户 u to 物品 i 的预测评分
γ	学习率
λ	衰减系数
T	迭代次数

3.4 梯度计算

我们的模型参数梯度计算公式如下，

$$\nabla U_{u\cdot} = -e_{ui} V_{i\cdot}^u + \alpha U_{u\cdot} \tag{2}$$

$$\nabla V_{i\cdot}^u = -e_{ui} U_{u\cdot} + \alpha V_{i\cdot}^u \tag{3}$$

其中 $e_{ui} = r_{ui} - \hat{r}_{ui}$ 是误差。

3.5 更新规则

我们的模型参数梯度更新公式如下，

$$\theta = \theta - \gamma \nabla \theta, \tag{4}$$

其中 γ 是学习率, 并且 θ 可以是 $U_{u\cdot}$ 或者 $V_{i\cdot}^u$。

3.6 领居构建

(用户, 用户) 相似度矩阵 **W** 可以通过余弦相似度计算：

$$W_{uu'} = \frac{|\mathcal{I}_u \cap \mathcal{I}_{u'}|}{\sqrt{|\mathcal{I}_u| \cdot |\mathcal{I}_{u'}|}} \tag{5}$$

其中 $\mathcal{I}_u$ 表示用户 u 的已评分集合。

我们可以控制（用户，用户）相似性矩阵 **W** 的稀疏性，如下所示。

$$|\mathcal{I}_u \cap \mathcal{I}_{u'}| \geqslant threshold \tag{6}$$

其中 $threshold$ 越大，**W** 就越稀疏。

3.7 算法

Algorithm 1 PDMFRec 的伪代码描述。

for $t = 1, 2, \ldots, T$ **do**
  for 每一个用户 u 并行计算 **do**
    for $r_{ui} \in \mathcal{R}_u$ **do**
      分别通过 Eq.(2) and Eq.(3) 计算用户和物品的梯度.
      通过 Eq.(4) 更新用户和物品的特征向量.
      for $u' \in \mathcal{N}_u$ **do**
        发送 $W_{uu'}\nabla V_i$. 给用户 u'.
      end for
      Synchronize(). /* 等待所有用户完成物品梯度的交换.*/
      更新物品的特征向量
      via $V_{j\cdot}^u = V_{j\cdot}^u - \gamma \frac{\sum_{w \in \mathcal{N}_u^j} W_{wu} \nabla V_{j\cdot}^w}{|\mathcal{N}_u^j|} \quad j \in \mathcal{I}_{\mathcal{N}_u}^t$.
    end for
  end for
end for

4 复现细节

4.1 与已有开源代码对比

本复现没有参考任何相关源代码。

4.2 实验环境搭建

本复现基于 java jdk 1.8 以及操作系统 windows 10 完成。

4.3 数据集和评估指标

本复现在实验中使用了公共数据集 MovieLens 100K（ML100K）[1] 本复现通过以下步骤处理每个数据集。(i) 我们将数据集随机分成五个相等的部分。(ii) 我们将三部分作为训练数据，一部分作为验证数据，剩下的一部分作为测试数据；(iii) 我们重复第二步五次，得到五份不同的训练数据、验证数据和测试数据。

对于每种算法的性能评估，我们使用两个常用的指标，即平均绝对误差（MAE）和均方根误差（RMSE），并报告每个数据集的五个副本上的平均性能。

4.4 基线方法和参数设置

本复现的基线方法是中心化推荐方法 PMF[18]。对于参数设置，我们固定了特征向量的维度 $d = 20$，并通过验证数据的 MAE 性能搜索学习率的最佳值 $\gamma \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$。

5 实验结果分析

本复现的实验结果在表 2 中。我们可以看到关于 PDMFRec 有三个部分。分别是 PDMFRec，PDMFRec(-sim) 以及 PDMFRec(-sim,all)。PDMFRec 是按照原始的论文伪代码编写的结果。PDMFRec(-sim) 是原始的 PDMFRec 去掉了相似度的结果。PDMFRec(-sim,all) 在 PDMFRec(-sim) 的基础上将物品梯度发送给了所有用户。我们可以发现，PDMFRec(-sim) 比 PDMFRec 效果要好，我们分析是因为用户间的相似度太低导致的。梯度被乘了一个很小数值的相似度，模型无法继续收敛，导致了效果不好。PDMFRec(-

[1] https://grouplens.org/datasets/movielens/100k/

sim,all) 比 PDMFRec 和 PDMFRec(-sim,all) 都要好，我们分析是因为发送给所有用户在一定程度上保证了所有用户的物品特征向量的一致性，使得模型能更好的收敛。我们可以看到，PMF 的效果最好，是因为它是将所有数据放在一个地方训练的，是分散式训练的一个上限值。

Algorithm	MAE	RMSE
PDMFRec	$1.0194_{\pm 0.0040}$	$1.2722_{\pm 0.0039}$
PDMFRec(-sim)	$0.7832_{\pm 0.0036}$	$0.9975_{\pm 0.0040}$
PDMFRec(-sim,all)	$0.7332_{\pm 0.0038}$	$0.9360_{\pm 0.0056}$
PMF	$0.7286_{\pm 0.0031}$	$0.9314_{\pm 0.0047}$

表 2: PDMFRec 和 PMF 在 ML100K 的推荐性能。

6 总结与展望

本文研究去中心化联邦推荐问题，并对其中比较经典的去中心化联邦推荐算法 PDMFRec 进行复现。我们发现 PDMFRec 中的基于相似度的领居构建发送梯度的方式不能取得很好的效果，模型的收敛性有待考究。

参考文献

[1] AMMAD-UD-DIN M, IVANNIKOVA E, KHAN S A, et al. Federated Collaborative Filtering for Privacy-Preserving Personalized Recommendation System[J/OL]. CoRR, 2019, abs/1901.09888. https://arxiv.org/abs/2004.04256.

[2] LIN G, LIANG F, PAN W, et al. FedRec: Federated Recommendation with Explicit Feedback[J]. IEEE Intelligent Systems, 2021: 36(5):21-30.

[3] LIANG F, PAN W, MING Z. FedRec++: Lossless Federated Recommendation with Explicit Feedback [C]//AAAI'21: Proceedings of the 35th AAAI Conference on Artificial Intelligence. 2021: 4224-4231.

[4] CHAI D, WANG L, CHEN K, et al. Secure Federated Matrix Factorization[J]. IEEE Intelligent Systems, 2020, 36(5): 11-20.

[5] WU C, WU F, CAO Y, et al. FedGNN: Federated Graph Neural Network for Privacy-Preserving Recommendation[J/OL]. arXiv preprint arXiv:2102.04925, 2021. https://arxiv.org/abs/2102.04925.

[6] MINTO L, HALLER M, LIVSHITS B, et al. Stronger Privacy for Federated Collaborative Filtering With Implicit Feedback[C]//RecSys'21: Proceedings of the 15th ACM Conference on Recommender Systems. 2021: 342-350.

[7] KHAN F K, FLANAGAN A, TAN K E, et al. A Payload Optimization Method for Federated Recommender Systems[C]//RecSys'21: Proceedings of the 15th ACM Conference on Recommender Systems. 2021: 432-442.

[8] ZHANG S, YIN H, CHEN T, et al. PipAttack: Poisoning Federated Recommender Systems for Manipulating Item Promotion[C]//Proceedings of the 15th ACM International Conference on Web Search and

Data Mining. 2022: 1415-1423.

[9] MUHAMMAD K, WANG Q, O'REILLY-MORGAN D, et al. FedFast: Going Beyond Average for Faster Training of Federated Recommender Systems[C]//KDD'20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2020: 1234-1242.

[10] LIU S, XU S, YU W, et al. FedCT: Federated Collaborative Transfer for Recommendation[C]// SIGIR'21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 716-725.

[11] ANELLI V W, DELDJOO Y, NOIA T D, et al. Federank: User controlled feedback with federated recommender systems[C]//European Conference on Information Retrieval. 2021: 32-47.

[12] KHARITONOV E. Federated online learning to rank with evolution strategies[C]//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019: 249-257.

[13] JALALIRAD A, SCAVUZZO M, CAPOTA C, et al. A simple and efficient federated recommender system[C]//Proceedings of the 6th IEEE/ACM international conference on big data computing, applications and technologies. 2019: 53-58.

[14] GUO Y, LIU F, CAI Z, et al. PREFER: Point-of-interest REcommendation with efficiency and privacy-preservation via Federated Edge leaRning[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(1): 1-25.

[15] HEGEDÜS I, DANNER G, JELASITY M. Decentralized Recommendation Based on Matrix Factorization: A Comparison of Gossip and Federated Learning[C]//Proceedings of the Workshop on Decentralized Machine Learning at the Edge with ECML PKDD 2019. 2019: 317-332.

[16] CHEN C, LIU Z, ZHAO P, et al. Privacy Preserving Point-of-Interest Recommendation Using Decentralized Matrix Factorization[C]//AAAI'18: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018: 257-264.

[17] DURIAKOVA E, TRAGOS E Z, SMYTH B, et al. PDMFRec: A Decentralised Matrix Factorisation with Tunable User-Centric Privacy[C]//RecSys'19: Proceedings of the 13th ACM Conference on Recommender Systems. 2019: 457-461.

[18] MNIH A, SALAKHUTDINOV R R. Probabilistic Matrix Factorization[C]//NeurIPS'07: Proceedings of the 21st Conference on Neural Information Processing Systems. 2007: 1257-1264.