软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(12):3852−3868 [doi: 10.13328/j.cnki.jos.006128]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于注意力联邦蒸馏的推荐方法*

谌 明, 张 蕾, 马天翼

(浙江省同花顺人工智能研究院,浙江 杭州 310012)

通讯作者: 谌明, E-mail: chm@zju.edu.cn

摘 要: 数据隐私保护问题已成为推荐系统面临的主要挑战之一.随着《中华人民共和国网络安全法》的颁布和欧盟《通用数据保护条例》的实施,数据隐私和安全成为了世界性的趋势.联邦学习可通过不交换数据训练全局模型,不会泄露用户隐私.但是联邦学习存在每台设备数据量少、模型容易过拟合、数据稀疏导致训练好的模型很难达到较高的预测精度等问题.同时,随着 5G(the 5th generation mobile communication technology)时代的到来,个人设备数据量和传输速率预计比当前提高 10~100 倍,因此要求模型执行效率更高.针对此问题,知识蒸馏可以将教师模型中的知识迁移到更为紧凑的学生模型中去,让学生模型能尽可能逼近或是超过教师网络,从而有效解决模型参数多和通信开销大的问题.但往往蒸馏后的学生模型在精度上会低于教师模型.提出一种面向推荐系统的联邦蒸馏方法,该方法首先在联邦蒸馏的目标函数中加入 Kullback-Leibler 散度和正则项,减少教师网络和学生网络间的差异性影响;引入多头注意力机制丰富编码信息,提升模型精度;并提出一个改进的自适应学习率训练策略来自动切换优化算法,选择合适的学习率,提升模型的收敛速度.实验验证了该方法的有效性:相比基准算法,模型的训练时间缩短 52%,模型的准确率提升了 13%,平均误差减少 17%,NDCG 值提升了 10%.

关键词: 联邦学习;分布式学习;联邦蒸馏;推荐系统;注意力机制

中图法分类号: TP18

中文引用格式: 谌明,张蕾,马天翼.一种基于注意力联邦蒸馏的推荐方法.软件学报,2021,32(12):3852−3868. http://www.jos.org.cn/1000-9825/6128.htm

英文引用格式: Chen M, Zhang L, Ma TY. Recommendation approach based on attentive federated distillation. Ruan Jian Xue Bao/Journal of Software, 2021,32(12):3852−3868 (in Chinese). http://www.jos.org.cn/1000-9825/6128.htm

Recommendation Approach Based on Attentive Federated Distillation

CHEN Ming, ZHANG Lei, MA Tian-Yi

(Zhejiang HiThink RoyalFlush AI Research Institute, Hangzhou 310012, China)

Abstract: Data privacy protection has become one of the major challenges of recommendation systems. With the release of the Cybersecurity Law of the People's Republic of China and the general data protection regulation in the European Union, data privacy and security have become a worldwide concern. Federated learning can train the global model without exchanging user data, thus protecting users' privacy. Nevertheless, federated learning is still facing many issues, such as the small size of local data in each device, over-fitting of local model, and the data sparsity, which makes it difficult to reach higher accuracy. Meanwhile, with the advent of 5G (the 5th generation mobile communication technology) era, the data volume and transmission rate of personal devices are expected to be 10 to 100 times higher than the current ones, which requires higher model efficiency. Knowledge distillation can transfer the knowledge from the teacher model to a more compact student model so that the student model can approach or surpass the performance of teacher model, thus effectively solve the problems of large model parameter and high communication cost. However, the accuracy of student model is lower than teacher model after knowledge distillation. Therefore, a federated distillation approach is proposed with attentional mechanisms for recommendation systems. First, the method introduces Kullback-Leibler divergence and regularization term to the objective function of federated distillation to reduce the impact of heterogeneity between teacher network and student network; then it introduces multi-head

* 收稿时间: 2020-01-18; 修改时间: 2020-04-18; 采用时间: 2020-08-07

attention mechanism to improve model accuracy by adding information to the embeddings. Finally, an improved adaptive training mechanism is introduced for learning rate to automatically switch optimizers and choose appropriate learning rates, thus increasing convergence speed of model. Experiment results validate efficiency of the proposed methods: compared to the baselines, the training time of the proposed model is reduced by 52%, the accuracy is increased by 13%, the average error is reduced by 17%, and the NDCG is increased by 10%.

Key words: federated learning; distributed learning; federated distillation; recommendation systems; attentive mechanism

近年来,随着电商平台和移动互联网的迅猛发展,人们已经步入信息过载的时代.推荐系统作为连接用户和信息的桥梁,正变得越来越重要.目前,主流的推荐系统主要基于大数据下的离线和在线推荐[1,2],但该类推荐系统往往需要收集大量用户个人信息以及浏览、购买等用户行为记录,存在数据隐私泄露的风险.随着《中华人民共和国网络安全法》、欧盟《通用数据保护条例》等一系列严格的数据隐私保护法律法规出台,对此类数据的收集提出更多限制措施.另外,出于政策法规、商业竞争等因素,不同机构间的数据很难互通[3].针对以上问题,联邦学习范式被提出[4,5].该范式可使模型在不上传用户隐私数据的前提下进行联合建模,同时与领域和算法无关,可实现在不同数据结构、不同机构间协同建模,有效保护用户隐私和数据安全[6].

随着 5G(the 5th generation mobile communication technology)技术的普及,用户设备端数据的上传速度和下载速度将高达 10Gbps 级别,同时,移动设备的响应时间将降至仅 1 毫秒级别,相比 4G(the 4th generation mobile communication technology)下载速度快 6.5 万倍[7];用户数据的爆炸式增长对机器学习模型的训练速度提出更高要求,与此同时,推荐系统随着模型的复杂度越高,联邦学习需要交换的权重系数也越多,给联邦学习下的模型移动端通信开销带来了严峻的挑战[8].知识蒸馏可用于将参数大的复杂网络(教师模型)中的知识迁移到参数量小的简单网络(学生模型)中去,用更少的复杂度来获得更高的预测效果[9].针对联邦学习设备间模型参数多和通信开销大,Jeong 等人[10]将知识蒸馏引入联邦学习场景,用于压缩每台设备模型参数的体量并减少通信次数.但除上述挑战和问题外,推荐系统在数据上仍存在着如下问题.

(1) 用户间行为数据差异较大,通常行为数据体现为长尾分布,使得设备间数据存在高度异质性;

(2) 真实推荐场景下数据大都为非独立同分布(non-IID),但大部分推荐算法往往仍基于独立同分布(IID)假设[11],该假设忽略了非独立同分布可能造成的数据、模型上的异质性.

在联邦蒸馏的场景下,以上问题会造成不同设备数据之间的差异,进而造成设备模型之间的差异.而知识蒸馏的引入,会进一步地扩大教师模型与学生模型之间的分布差异,使全局模型收敛速度慢,准确率低.针对以上问题,还没有针对推荐场景的联邦蒸馏算法及框架被提出.

本文提出基于注意力联邦蒸馏的推荐方法,该方法相比 Jeong 等人[10]提出的联邦蒸馏算法做了如下改进.在联邦蒸馏的联合目标函数中加入 KL 散度(Kullback-Leibler divergence)和正则项,减少因教师网络和学生网络间的差异对全局模型造成的影响,提升模型稳定性和泛化性能;在联邦蒸馏设备端流程中引入改进的多头注意力(multi-head attention)机制,使特征编码信息更加丰富,提升整体模型精度;提出一种自适应学习率的训练策略,利用混合优化的方法优化联邦蒸馏的联合目标函数,提高模型收敛速度,抵消注意力编码增加的计算量.该方法是目前第一个面向推荐系统场景的联邦蒸馏方法.

1 相关研究

1.1 联邦学习

数据的隐私保护一直是推荐系统的重要研究方向,联邦学习可在不共享隐私数据的情况下进行协同训练,能够有效地解决数据隐私问题[12].国内外一些学者对其进行了研究.Google AI 团队提出了联邦学习方法,该方法在不收集用户数据的情况下,在每台设备上独立完成模型训练,再将梯度数据进行隐私保护加密传输到中心节点服务器(联邦中心),最后,中心节点根据汇总结果将更新后的梯度(全局模型)再回传到每台设备上,从而完成每台设备的梯度和模型更新,解决了用户数据孤岛问题[13−14].目前,机器学习的很多领域都已引入联邦学习,

如联邦迁移学习[15]、联邦强化学习[16]、联邦安全树[17]等.Yurochkin 等人[18]提出了贝叶斯无参联邦框架,通过实验证明了效率上的有效性,模型压缩比更低.Liu 等人[19]提出一种迁移交叉验证机制的联邦学习,能够为联邦内的设备模型带来性能提升;他们还提出灵活可拓展的方法,为神经网络模型提供额外的同态加密功能.Zhuo 等人[20]提出一种新的联邦强化学习方法,为每台设备构建新的 Q 网络,解决了构建高质量的策略难度大的问题;更新本地模型时对信息使用高斯差分保护,提升了用户的隐私保护能力.Kewei 等人[21]提出一个联邦提升树系统,可以让多个机构共同参与学习,可以有效地提升分类准确率,同时让用户对自己的数据有更多的控制权.也有学者在联邦学习中引入其他算法,并对联邦学习效率问题进行研究.Sharma 等人[22]提出一种隐私保护树的 Boosting 系统,能够在精度上与非隐私保护的算法保持一致.Ghosh 等人[23]提出了一种离群对抗方法,将所有节点和异常的设备一起考虑,解决了鲁棒异质优化问题,并给出了分析误差的下界.虽然联邦学习能够解决数据隐私问题,但随着用户数据量和模型复杂度的增加,存在着模型参数多和移动端通信开销大等问题.学者们希望使得通信负载与模型大小无关,只与输出大小有关.将教师模型中的知识迁移到学生模型中,降低复杂度的同时仍能保持较好的预测精度,知识蒸馏便是这样一种知识迁移的方法.

1.2 知识蒸馏

Hinton 等人[24]提出了知识蒸馏,将教师网络相关的软目标作为损失函数的一部分,以诱导学生网络的训练,实现知识迁移.Yim 等人[25]使用矩阵来刻画层与层之间的特征关系,然后用 L2 损失函数去减少教师模型和学生模型之间的差异,并让学生模型学到这种手段,而不仅仅是利用目标损失函数进行知识的迁移.Heo 等人[26]利用对抗攻击策略将基准类样本转为目标类样本,对抗生成的样本诱导学生网络的训练,从而有效提升学生网络对决策边界的鉴别能力.但硬标签会导致模型产生过拟合现象,对此,Yang 等人[27]提出了一个更合理的方法,并没有去计算所有类的额外损失,而是挑选了几个具有最高置信度分数的类来软化标签,提高模型的泛化性能.

近些年,有学者提出将联邦学习和知识蒸馏结合起来.Jeong 等人[10]提出了一种分布式模型联邦蒸馏训练算法,能够有效解决用户通信开销大的问题.采用生成对抗网络生成数据,解决用户生成的数据样本非独立同分布的问题.Han 等人[28]提出一种保护隐私的联邦强化蒸馏框架,由事先设置的状态和策略组成,通过交换每台设备的策略值,从而共同训练本地模型,解决了代理隐私泄露问题.然而这些方法在解决非独立同分布问题上主要采用生成对抗网络或强化学习的方法将非独立同分布数据转为独立同分布数据,在实际应用中复杂性较大.

针对前述文献和方法的不足,尤其是因联邦学习回传梯度参数的方法参数多、计算量大、模型训练过程无法自适应调节学习率、蒸馏算法训练速度慢等问题,本文在第 2 节提出并详细描述一种基于注意力机制的联邦蒸馏推荐方法.

2 一种基于注意力联邦蒸馏的推荐方法(AFD)

2.1 符号定义

推荐系统中通常包括召回和排序两个阶段:召回阶段对历史数据用协同过滤或其他召回算法召回一批候选 Item 列表,排序阶段对每个用户的候选 Item 列表进行 CTR(click-through rate)预测,最后选取排序靠前 Top-n 的 Item 作为推荐结果.

假设整个系统包含设备集 K(共$|K|$台设备),每台设备包含 Item 特征和用户特征,则设备 k($k\in K$)中的用户特征为 U^k,Item 特征为 I^k.r 为第 k 台设备上的特征总数.X^k 为设备 k 的本地数据,y^k 为设备 k 的本地数据对应的标签,p^k 为设备 k 上的本地数据对应的学生模型预测结果.E 为全局训练轮数,t 为所有数据的标签值($t\in T$,T 为标签集).S^k 为联邦中心收集到的设备 k 的 Logits 向量集合(本文的 Logits 向量皆为经过 *softmax* 操作后的归一化的向量值),即为学生模型;$S^{/k}$ 为除去设备 k 后其他设备的 Logits,$\overline{S}^{/k}$ 为教师模型,为除去设备 k 后其他设备的 Logits 平均值.本文提出方法所使用的主要符号定义见表 1.

Table 1 Definitions of main symbols

表 1 主要符号定义

符号	说明	符号	说明
$k,t,K,T,$ $\|K\|,\|T\|,E$	设备编号;标签编号;设备集;标签集;设备数量;标签数量;设备本地模型训练轮数	$S^k,S^{/k},\bar{S}^k,\bar{S}^{/k}$	设备 k 中学生网络输出的标签 Logits 集合;去除设备 k 的其他设备所有标签 Logits 加和的集合;设备 k 中学生网络输出的所有标签的平均 Logits 集合;去除设备 k 的其他设备所有标签的平均 Logits 集合
n_t^k,g^k	设备 k 中学生网络训练数据为标签 t 的数据数量和该设备数据总量	$y^k,p^k,p_{teacher}^k$	设备 k 的数据实际类别标签,学生模型预测标签和教师网络预测标签
$l(\cdot),L(\cdot)$	学生网络和教师网络的损失函数	GL	联合损失函数
$P(\cdot)$	查询项与搜索矩阵的相似度	Q,S_i,V_i,b,a_i	特征 i 的搜索矩阵,查询键值,结果矩阵,minibatch 数据量,*attention* 值
R_z,m_z,V_z	时刻 z 目标函数关于模型参数的梯度,一阶动量和二阶动量	$\varphi(R_{z-q}^2)$	z 时刻之前 q 时刻内的最优梯度
ω,δ_z	模型参数,时刻 z 的模型梯度	$\eta_z^{SGD},\eta_z,\eta_z^{Adam}$	z 时刻 SGD 阶段初始学习率,修正后的学习率及 Adam 的阶段学习率
X^k,Y^k,E	设备 k 上的本地数据集,实际标签和预测标签,全局训练轮数(epoch)	I^k,U^k,r	第 k 台设备上用户特征,商品特征和特征总个数

2.2 方法整体流程

本文提出的基于注意力和联邦蒸馏的推荐方法(AFD)运行在多个分布式设备中,包括在设备端运行的学生网络和运行在服务器端负责收集、整合、分发教师模型参数的联邦中心.协作流程如图 1 所示.

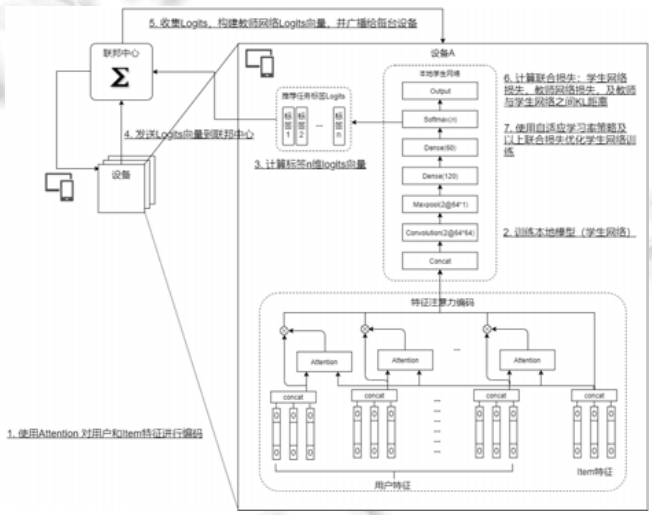


Fig.1 Collaboration flow of attentive federated distillation

图 1 注意力联邦蒸馏协作流程

具体描述如下:

(1) 每台设备初始化一个基于深度神经网络的推荐(或点击率预测)模型(如卷积神经网络、DeepFM[29]等)作为学生模型,使用设备本地数据进行模型训练.其中,本地设备使用 Attention 机制(见第 2.4 节)对本地用户特征和商品特征进行编码,融合特征交叉信息得到特征 Embedding 表达,并将这些表达作为本地模型的输入进行训练.使用 Attention 机制可捕捉更多兴趣特征,同时,编码本身可减少本地用户数据泄露的风险;

(2) 本地模型训练收敛后,设备获取模型参数,并将模型参数上传至联邦中心.这里,上传的模型参数与常规联邦学习中的不同:联邦蒸馏方法上传的参数为本地学生模型最后 *Softmax* 层计算出的 Logits 向量(每个推荐目标标签对应的 Logits 向量,取多轮训练的平均值),而联邦学习方法上传的则是模型权重

矩阵.对推荐标签数量较少或点击率预测任务(二分类),使用联邦蒸馏方法可大大减少上传参数的体量,缓解大规模设备下可能造成的通信拥堵;

(3) 联邦中心使用联邦学习算法将接收到的每台设备上传的标签平均Logits向量整合为新的全局Logits向量.具体地,针对每台设备,联邦中心将其他设备发送的Logits向量使用联邦学习算法构建出该台设备的教师模型,并将教师模型分发到每台设备中(该步骤具体流程详见表3);

(4) 设备接收教师模型,通过结合自适应学习率策略(见第2.5节)优化联合损失函数(见第2.3节),并以此指导学生网络的训练.联合损失函数包含教师网络、学生网络的损失,同时还包含教师网络与学生网络之间的差异度.该步骤算法流程详见表2.

以上描述中,步骤(1)和步骤(3)中的推荐算法和联邦学习算法不限,可根据实际需求自由组合.在下面的章节,我们将详细描述图1流程及表2、表3算法中使用的策略.

Table 2 Attentional federated distillation—Processes on devices

表2 注意力联邦蒸馏算法——设备流程

输入:设备集K,标签集T,设备模型迭代训练轮数E;

```
随机初始化每台设备每个标签的Logits S_t^k
for k in K do
    接收联邦中心分发的教师模型Logits S'^k = {S'_1^k,...,S'_|T|^k}
    for e←1 to E do
        l_e^k ← Eq.(1)   //计算学生模型损失
        GL(e,k)←Eq.(3,4)   //计算联合损失
        η←Eq.(20)   //自适应学习率策略选择合适学习率
        ω_e^k ← ω_e^k − η∇GL(e,k)
        for t in T do   //计算设备k每个标签的Logits
            S_t^k ← Eq.(5)
        end
    end
    for t in T do   //计算设备标签的平均Logits
        统计本设备标签为t的训练数据量 n_t^k
        S̄_t^k ← S_t^k / n_t^k
    end
    发送设备k的平均Logits S̄^k = {S̄_1^k,...,S̄_|T|^k} 到联邦中心
end
```

Table 3 Attentional federated distillation—Processes on the federated center

表3 联邦注意力蒸馏算法——联邦中心流程

输入:设备集K,标签集T,全局模型最大迭代轮数Max$Epoch$;

```
随机初始化每台设备除该设备外的Logits S'^k
for epoch←1 to MaxEpoch do
    for k in K do
        接收设备k的平均Logits S̄^k = {S̄_1^k,...,S̄_|T|^k}
        for t in T do
            S̄_t^k ← S̄_t^k + S̄'_t^k   //累加其他设备标签t的Logits
        end
    end
    //计算教师网络Logits并分发
    for k in K do
        for t in T do
            S'_t^k ← Eq.(6)
        end
        分发教师模型Logits S'^k = {S'_1^k,...,S'_|T|^k} 到设备k
    end
end
```

2.3 联邦蒸馏

现有的联邦学习算法是对模型权重进行平均,由于推荐系统中模型复杂,权重参数众多,分配到每台设备上,模型参数回传到联邦中心,会占用大量的资源,并且联邦中心计算权重平均值也是一笔巨大的时间开销.当采用现有联邦蒸馏算法的损失函数进行优化时,仅仅分别计算了教师网络和学生网络与真实标签的误差值,却忽略了教师网络和学生网络本身的差异性给模型带来了影响,容易造成模型过拟合.通过实验发现,教师网络和学生网络本身的差异性对模型的推荐效果具有较大的影响.为了减少学生网络和教师网络之间差异大造成的影响,本文提出了一种新的目标函数.相比于传统目标函数只计算本地设备预测值与真实值之间的误差,本文提出的目标函数除了利用其他设备作为教师模型来指导本地学生模型训练,还将学生模型与教师模型之间的差别作为优化目标的一部分加入损失函数,降低设备间数据差异造成的影响.

首先,设备 $k(k\in K)$的本地学生模型及联邦中心分发的教师模型在该设备上的损失函数可分别定义为

$$l^k=f(p^k,y^k) \tag{1}$$

$$L^k=f(p^k_{teacher},y^k) \tag{2}$$

其中,$f(\cdot)$为损失函数,p^k 和 y^k 分别为设备 k 中学生模型对本地测试数据的预测值及其真实值, $p^k_{teacher}$ 为教师模型对本设备测试数据的预测值.假设全局模型共需训练 E 轮,则训练 e 轮($e\in[1,E]$)后的联合损失函数由学生模型损失、教师模型损失以及学生模型与教师模型差异组成(表 2 第 6 行),具体定义如下:

$$GL(e,k)=\begin{cases}\alpha l_e^k+\dfrac{\lambda}{2}\|\omega^k\|^2, & \text{if } e=1\\ \underbrace{\alpha l_e^k}_{\text{学生模型损失}}+\underbrace{\beta L_e^k}_{\text{教师模型损失}}+\underbrace{(1-\alpha-\beta)KL(l_e\|L_e)}_{\text{学生模型与教师模型差异}}+\underbrace{\dfrac{\lambda}{2}\|\omega^k\|^2}_{\text{正则项}}, & \text{if } e>1\end{cases} \tag{3}$$

其中,α,β分别为学生模型和教师模型损失的权重参数,λ为正则项权重参数,ω^k 为设备 k 的模型参数(如神经网络中的 Weights 和 Bias),$\|\cdot\|^2$ 为 $L2$ 范数.为节省参数通信量(传统联邦学习算法如 FedAvg 需传输模型参数)并增强模型的泛化性能,本文方法在联合损失函数中增加了 $L2$ 正则项(见公式(3)).由于高度偏斜的非独立同分布(non-IID)数据会让学生模型之间的分布差异增大,降低整个模型的收敛效率,本文通过使用 KL 散度(Kullback–Leibler divergence)来衡量学生模型和教师模型之间的差异,并将该差异作为全局损失函数的一部分进行优化.差异计算方式如下:

$$KL(l_e^k\|L_e^k)=\sum_{k=1}^{|K|}l_e^k\log\frac{l_e^k}{L_e^k} \tag{4}$$

公式(3)中,当 e=1 时(即第 1 轮全局模型训练),此时联邦中心尚未收集首轮本地设备的模型 Logits,本地设备无需从联邦中心接受教师模型的 Logits,此时,联合损失仅包含本地学生模型的损失;当 e>1 时,联邦中心已完成首轮模型收集并分发教师模型,则本地学生模型的优化可同时使用学生模型、教师模型及学生-教师模型差异进行联合优化.同时,为加速模型收敛速度,本文提出一个可自动切换优化算法及选择合适学习率的优化策略,用于优化联合损失函数(见第 2.5 节).优化后的本地学生模型对本地设备数据进行预测,得出新本地模型对应每个数据标签的 Logits $\hat{S}_t^k$,并通过下式更新设备 k 对应标签 t 的 Logits(表 2 第 10 行):

$$S_t^k=\hat{S}_t^k+\overline{S}_t^{/k} \tag{5}$$

其中, $\overline{S}_t^{/k}$ 为从联邦中心接受到的去除第 t 台设备后的 Logits 平均值(即教师模型).最后,设备 k 将平均后的本地 Logits 作为设备 k 的学生模型发送到联邦中心进行整合.联邦中心通过计算除去每台设备本身的其他设备学生模型的 Logits 的平均值来得到该设备的教师模型,并分发给对应设备(表 3 第 16 行).具体计算方式如下:

$$S_t^{/k}=\frac{\overline{S}_t^k-\overline{S}_t^{/k}}{|K|} \tag{6}$$

联邦蒸馏的过程减少了传统联邦学习过程中的模型权重回收和分发造成的时间和通信开销,能够有效提升整体效率.同时,通过加入 KL 散度,将教师模型和学生模型之间的差异性加入到损失函数中进行优化,从而缓

解了数据差异带来的影响,提升模型的推荐性能.然而,联邦蒸馏虽然可以缓解 Non-IID 的影响,但若设备之间数据差异较大或数据量较少,仍然需要其他优化手段来提高模型的精度.在下面的章节中,本文方法利用特征注意力编码得到特征间更多的交互信息来丰富本地特征.

2.4 特征Attention编码

在推荐场景中,用户兴趣和产品的种类具有多样性,一个用户可能对多个种类产品感兴趣,一个种类可能有多个产品,但最终影响模型结果可能只有其中一部分.以付费服务推荐场景为例:当一个用户同时购买了两个付费产品,很难区分他对哪个产品更感兴趣;但如果其中一个产品连续购买多次,另一个产品只购买过一次,那么说明连续购买年数这一特征,对模型的分类具有更高的权重影响.同时,对于不同的用户,可能是因为不同的特征而决定最后是否会购买.这类场景下,对用户交互过的商品和候选商品做特征 Attention 编码尤为重要,可以有效地捕捉用户对不同商品及不同特征之间的差异性.由于不同的用户关注的兴趣点不同,用户兴趣呈现多样性变化,主流的深度神经网络(DNN)模型对用户的历史行为是同等对待,且忽略了时间因素对推荐结果的影响[30],离当前越近的特征越能反映用户的兴趣.然而,现有的基于联邦学习的推荐方法未考虑特征之间的交互关系.为了充分利用历史特征及特征交互信息,本文通过加入一个改进的 Attention 机制,在特征向量进入模型训练之前通过 Attention 机制计算用户行为权重,得出每个用户不同的兴趣表征.目前,基于 Attention 机制的方法[31,32]通常在输出层前加入 Attention 层,以捕捉用户和 Item 的二阶交叉信息.与这些方法不同,我们并未在模型输出层前加入 Attention 层,而是在模型输入前使用.这样做有如下目的:1) 保证框架灵活性,避免侵入现有本地模型的结构;2) 尽可能丰富输入特征的信息,提高模型精度.

对于每一个用户,有一个等长于特征总数 r 的 Attention 编码,其中,Attention 编码的每一个维度表示该特征的权重(即重要程度).由于用户线上的交互特征通常非常稀疏,当一个用户的特征值只有一个非零特征时,这个特征会得到很高的 Attention 得分;而当一个用户有多个非零特征时,受限于 *Softmax* 计算的 Logits 值,各个特征的 Attention 得分反而不高,重点信息难以全部保留.本文使用的 Attention 方法主要包含两点改进:1) 由于不同设备中数据特征维度空间不同,提出一种映射方法将不同设备数据映射到相同维度,进而允许其进行 Attention 操作;2) 增加 Attention 编码的维度,增强特征交互的表征能力.

(1) Attention 编码映射

由于每台设备的数据特征空间不相同,首先需要将所有特征统一映射到一个 *dim* 维的 Embedding 矩阵.具体地,通过创建特征 embedding 向量[*feature_count*,*dim*],将单阶或多阶特征映射到[*b*,*r*,*dim*].其中,*feature_count* 为所有特征的类别总数,*b* 为一个 minibatch 的数据量(如 16,32),每批次训练数据维度为[*b*,*max_feature*],*max_feature* 为所有特征的维度总和,*r* 为特征总数量.映射过程中,若特征为单阶,如连续数值型特征,则特征 Embedding 为该数字在 Embedding 向量中对应的特征;若特征为多阶,如 One-Hot 特征,则使用多阶特征所有特征值在 Embedding 向量中对应特征的和作为该多阶特征的 Embedding.具体搜索矩阵对第 i 个特征的 Attention 权重计算方式如下:

$$P(Q,S_i)=QS_i^T \tag{7}$$

其中,Q 为搜索矩阵,S_i 为特征 i 的查询键值,$(\cdot)^T$ 为矩阵转置.映射后的 Q 维度为[*b*,1,*dim*],S_i 维度为[*b*,*r*,*dim*],V_i 维度为映射到[*b*,*r*,*dim*].$P(\cdot)$为查询项与搜索矩阵的相似度,同时也为搜索矩阵 Q 对特征的权重系数,维度为[*b*,1,*r*].最后,再通过 *Softmax* 操作归一化到[0,1].具体如下:

$$a_i=softmax(P(Q,S_i))=\frac{P(Q,S_i)}{\sum_{i=1}^{r}P(Q,S_i)} \tag{8}$$

(2) 增加 Attention 的维度

传统的 self-attention 是在序列内部做 attention 操作,每次使用一个用户的特征去查询其和所有其他特征的匹配程度,共进行 r 轮相同操作得到 attention 值.对于每个用户,只有一个等长于 r 的 Attention 矩阵,Attention 矩阵的大小为[*b*,*r*].但推荐场景的数据集通常很稀疏,当一个用户只有一个非零特征时,这个特征会得到很高的分

值;而当一个用户有多个非零特征时,重点特征的权重值反而难以取得较高的得分.本文方法将得到的搜索矩阵 Q 做矩阵变换,首先将权重系数矩阵由[b,1,r]转为[b×r,1],再利用矩阵乘法将结果与[1,m]相乘得到[b×r,m],再将权重系数矩阵转为[b,m,r].其中,m 为新增加的 Attention 的维度.对于每个特征,有 m 个等长于 r 的 Attention 值,变换后矩阵的大小由[b,1,r]变为[b,m,r],从而增加 Attention 的维度 m,促使不同的 Attention 关注不同的部分,减少了因召回商品数量不同造成的影响.通过求均值,将[b,m,r]变为[b,1,r],得到 Attention 值 a_i,再根据权重系数对 V_i 进行加权求和,得到搜索矩阵 Q 的 Attention 值.具体如下:

$$Attention(Q,K_i,V_i)=\sum_{i=1}^{r}a_iV_i \tag{9}$$

虽然特征 Attention 编码能够丰富编码信息,提升模型精度,但由于增加了特征维度,可能会降低模型的训练速度.最后,本文提出一种分段自适应学习率训练策略,通过切换不同的优化器来加快模型收敛速度.

2.5 分段自适应学习率策略

目前,已有文献实证发现:在联邦学习及分布式训练中,Adam 等基于动量的优化方法会直接影响到联邦学习的效果.尤其在非独立同分布(non-IID)数据下,本地设备模型的更新方向可能与全局模型差别较大,从而造成全局模型效果下降[33,34].同时,推荐系统是一个复杂的非线性结构,属于非凸问题,存在很多局部最优点[35].

Bottou 等人指出:SGD 虽然可以加快训练速度,但因为 SGD 更新比较频繁,会造成严重的震荡陷入局部最优解[36,37].联邦学习需要在典型的异构数据的情况下,通过全局数据优化每台设备上的模型,因此需要一种快速、能适应稀疏和异构分布数据的优化策略.Gao 等人提出了多种自适应方法来缩放梯度,解决了在数据稀疏的情况下存在性能差的问题,但仅仅通过平均梯度平方值的方法无法提升收敛速度[38,39].Shazeer 等人提出了一种分段调整学习率方法,采用分段训练的方式,在不损失精度的情况下提升了训练速度,但需要根据经验来选择切换的时机和切换后的学习率[40,41].

针对以上问题,本文基于 Wang 等人的工作[38],提出了一种分段自适应学习率优化方法,该方法的主要创新点为:1) 优化梯度下降过程,改进动量的计算方法,解决正相关性带来的收敛困难问题;2) 让算法在训练过程中自动由 Adam 无缝转换到 SGD 的混合优化策略,从而保留两种优化算法的各自优势,大幅缩短联合损失函数的收敛时间,并且保证了模型的准确性.

本地设备学生模型的目标函数为最小化联合损失(见公式(3)),即 $\min GL,\omega$为学生模型参数(如神经网络模型中的 Weights,Bias 等),则在时刻 z 目标函数关于模型参数的梯度 R_z 为

$$R_z=\nabla GL \tag{10}$$

在基于动量的优化算法中,动量表示参数在参数空间移动的方向和速率.目标函数关于参数的梯度二阶动量等价于当前所有梯度值的平方和.目标函数关于模型参数的一阶动量 m_z 和二阶动量 V_z 分别为 R_z 和 R_z^2 的指数移动平均.二阶动量 V_z 通过除以 $\sqrt{V_z}$ 实现对 R_z 尺度的缩放控制,反映了梯度下降的速率.但在 Adam 算法中,动量的计算本质上为动量 V_z 与梯度 R_z 的正相关性计算,会导致大梯度的影响减弱,小梯度的影响增强,最终会让收敛变得困难.本文假设过去时刻的参数梯度相互独立,因此可以利用过去 q 时刻的参数梯度 R_{z-q} 计算 V_z,而无需引入相关性计算.具体地,该策略从最近的 q 时刻的参数梯度中选择一个最优值,即:

$$\phi(R_{z-q}^2)=\max\{R_{z-q}^2,R_{z-q+1}^2,...,R_z^2\} \tag{11}$$

为解决上面讨论的正相关性计算带来的收敛困难问题,本文提出了优化后的二阶动量计算方法:

$$V_z=\mu_1V_{z-q}+(1-\mu_1)\phi(R_{z-q}^2) \tag{12}$$

其中,μ_1 为权重参数.公式(12)使用最近 q 时刻的最优梯度代替当前梯度,避免了计算二阶动量所需的相关性计算.同样地,一阶动量的计算也可去相关性,即:在计算一阶动量时,也利用最近 q 时刻的参数梯度来更新 m_z.具体如下:

$$m_z=\frac{\sum_{i=z-q}^{z}\mu_2R_i}{\sum_{i=z-q}^{z}\mu_2} \tag{13}$$

其中,μ_2 为权重系数.由公式(12)和公式(13)可得到时刻 z 的下降梯度:

$$\delta_z = \frac{\mu_3 m_z}{\sqrt{V_z}} \tag{14}$$

其中,μ_3 为梯度下降的权重系数.最后,根据下降梯度更新 z+1 时刻的学生模型参数 ω_{z+1}:

$$\omega_{z+1}=\omega_z-\delta_z \tag{15}$$

由于基于动量的 Adam 算法会直接影响联邦学习的收敛效果,本文在学生模型训练过程前半段采用 Adam 优化,后半段采用 SGD 优化,同时解决训练过程中相关性导致的模型收敛困难和收敛速度慢的问题.其中,优化算法的切换条件及切换后 SGD 的学习率为该分段策略的两个关键点.

(1) 算法切换条件.

联邦学习中,利用自适应学习率的方法(如 Adam)存在切换时间选择困难的问题:如切换过快,则无法提升收敛速度;切换过慢,则可能陷入局部最优解,影响收敛效果.受 Wang 等人提出的从 Adam 切换到 SGD 的条件[38]的启发,当满足迭代轮数大于 1 且修正后的学习率与原始的学习率的绝对值小于指定阈值ξ时进行切换,即:

$$|\eta_z-\eta_z^{SGD}|<\xi \tag{16}$$

其中,η_z 为每个迭代都计算的修正后的 SGD 学习率,与原始的学习率 η_z^{SGD} 之差的绝对值小于阈值,则认为已经满足切换条件,则切换为 SGD 并以调整后的学习率继续训练.接下来介绍如何确定 SGD 切换后的学习率.

(2) 切换算法后,SGD 的学习率.

SGD 阶段需确定的学习率包括初始学习率及修正后的学习率.Wang 等人提出将 SGD 下降的方向分解为 Adam 下降的方向和其正交方向上的两个方向之和[38],本文方法与前者的区别在于对正交分解后的方向进行修正.由于 Adam 计算学习率使用的是二阶动量的累积,要想计算出 SGD 阶段学习率大小,需要对 SGD 的下降方向进行分解.本文将 SGD 下降的方向分解为 Adam 下降的方向和其正交方向上的两个方向分别乘以 0.5(cos60°)再求和,其余部分与 Wang 等人的方法一致[38].假设模型优化已由 Adam 切换为 SGD 阶段,首先要沿着模型预测方向(p^k)走一步,而后沿着其正交方向走完相应步数.在当前时刻 z,正交分解后的 SGD 在 Adam 下降方向上的正交投影为 $proj_{\eta_z}^{SGD}$,等价于 Adam 的下降方向 η_z^{Adam},即:

$$proj_{\eta_z}^{SGD}=\eta_z^{Adam} \tag{17}$$

求解该方程,得到 SGD 阶段的初始学习率 η_z^{SGD}:

$$\eta_z^{SGD}=\frac{(\eta_z^{Adam})^T\eta_z^{Adam}}{(\eta_z^{Adam})^T R_z} \tag{18}$$

为了减少扰动,使用移动平均值来修正对学习率的估计,修正后的学习率如下:

$$\eta_z=\frac{\eta_z^{SGD}}{1-\sigma} \tag{19}$$

其中,σ为 SGD 权重系数.

3 实验及分析

3.1 数据集及实验设置

我们在 Movielens[42]数据集和同花顺 Level2 数据集上验证 AFD 及策略的有效性.Movielens 数据集包含 2 000 个用户及用户特征、3 300 部电影以及电影的标签属性信息.实验中,选取电影评价数大于 15 的电影和评价电影数量大于等于 10 的用户作为训练样本[43].本文还在同花顺真实场景金融数据集中进行验证,数据集主要包含用户对 Level2 产品的购买情况统计,特征包括了用户 ID、用户历史购买信息、设备信息、用户对该产品的评价、用户自身属性特征、产品特征等.其中,离散特征 18 项,连续特征 22 项.实验中对特征进行预处理,包括缺失特征补全、去掉用户编码和标签字段缺失的用户、去掉用户非空特征数量小于 3 的数据等.预处理完成后,训练集共有 32 万用户及 40 项特征,共 78 万条样本数据;测试集共有 12 万用户,40 项特征共 25 万条样本数据.

实验过程中,对原始数据进行去噪和脱敏处理,采用交叉验证的方式,将训练集和测试集分成 4 份,并分发到 4 台模拟设备,模拟联邦实际应用场景,每台设备上的数据相互独立.

为了对比不同联邦推荐算法的推荐准确率,我们将本文提出的基于注意力联邦蒸馏的推荐算法 AFD 结合卷积神经网络(CNN)与结合联邦学习的其他 3 种推荐算法进行对比实验,这 3 种推荐算法包括:

(1) FWD:联邦学习(FedAvg)结合 Wide&Deep 算法[44];

(2) FDIN:联邦学习(FedAvg)结合深度兴趣网络(DIN)算法[45];

(3) FD+CNN:联邦蒸馏算法[10]结合卷积神经网络.AFD+CNN 方法在不使用本文提出的 3 个策略的情况下等价于 FD+CNN.

AFD 与以上 3 个模型的对比见表 4.

Table 4 Comparisons between AFD and baselines

表 4 AFD 算法和基准模型对比

算法	注意力机制	联邦学习算法	蒸馏机制	自适应学习率
AFD+CNN	√	√	√	√
FD+CNN	×	√	×	×
FWD	×	√	×	×
FDIN	√	√	×	×

本文模型及实验使用 Tensorflow 实现,并且在 Nvidia GeForce GTX 1080Ti GPU 上进行实验.AFD 及 3 种方法的实验设置如下.

(1) AFD+CNN:attention 的维度 m 设为 32.网络层参数设置,CNN 层数为 5,隐藏层的大小 hidden_units 设为 128,两个卷积核为[64,64],最大池化层为[64,1],3 个全连接层为 120,60 和 2;

(2) FWD:Deep 部分全连接层为 128,64 和 2;

(3) FDIN:隐藏层单元数为 32,全连接层为 80,40 和 2;

(4) FD+CNN:CNN 层数为 5,2 个卷积层,1 个最大池化层,2 个全连接层.

3.2 评价指标

本文采用如下指标作为实验结果的评价指标.

- Time:模型迭代指定轮数运行的时间;
- Loss:模型损失函数(为与其他模型统一,AFD 评估学生模型原始损失,而非联合损失);
- AUC:ROC 曲线下面积,用来反映分类器的分类能力;
- ACC:准确率,表示分类正确的样本数占样本总数的比例;
- NDCG(normalized discounted cumulative gain):归一化折损累积增益;
- MAE(mean average error):评估算法推荐质量的指标,通过计算实际分值与预测分值的差异,来衡量推荐是否准确.

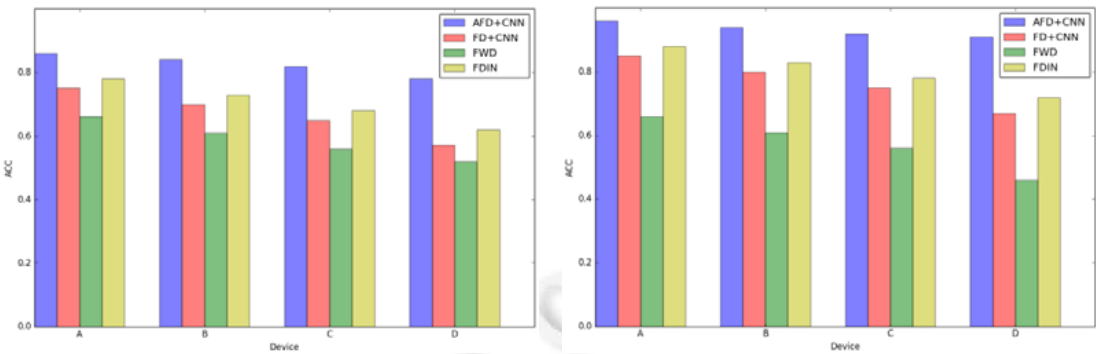
3.3 实验结果及分析

- 实验 1:不同联邦推荐算法下的精度实验.

在两个数据集上的准备率对比结果如图 2 所示,结果表明,本文提出的 AFD 算法准确率高于其他 3 种基准方法.在 Movielens 数据集上,AFD 算法的平均准确率最高达到了 0.84,FDIN 的准确率高于 FD 和 FWD 算法.在 Level2 数据集上,AFD 算法的准确率达到 0.92 左右,FD+CNN 的准确率为 0.81 左右,FWD 准确率仅为 0.67 左右,FDIN 约为 0.83 左右,AFD 相比不使用本文提出的 3 个策略的 FD+CNN 算法在准确率上提升了 13%.可以看出:FD+CNN 在使用联邦蒸馏机制后,模型精度与 FDIN 相当.FDIN 由于使用了 Attention 机制,总体精度优于除 AFD 外的其他方法.

表 5 为 4 台设备中的 MAE 及全局模型的 MAE.由表 5 可以看出:由于数据分布情况不同,4 台设备中模型精度有较大差别.同一设备,FWD 误差值最大,FDIN 和 FD 算法 MAE 均小于 FWD 算法.在 Movielens 数据集

上,FWD 算法的 MAE 值最大,推荐效果最差,而 AFD 算法 MAE 值比 FD 算法平均误差减少了约 20%.在 Level2 数据集上,FD 和 FDIN 算法 MAE 结果近似,而 AFD 算法比以上两种算法平均误差减少了约 17%.同时,AFD 在 4 台设备中均取得了最好的结果,表明 AFD 相对于其他 3 种基准算法推荐性能表现最佳.



(a) Movielens (b) Level2

Fig.2 ACC on different datasets

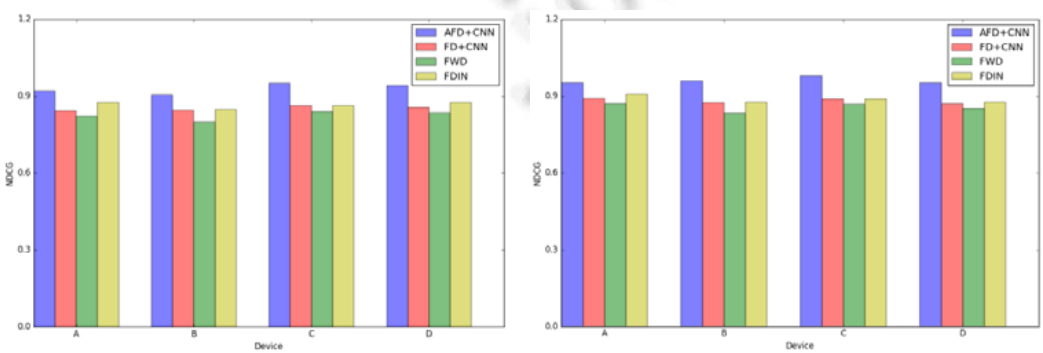
图 2 不同数据集下的 ACC

Table 5 MAE on Movielens and Level2 datasets

表 5 Movielens 和 Level2 数据集下的 MAE

数据集	算法	设备A	设备B	设备C	设备D	Global-*MAE*
Movielens	AFD+CNN	**0.15**	**0.23**	**0.20**	**0.17**	**0.19**
	FD+CNN	0.18	0.27	0.26	0.23	0.24
	FWD	0.24	0.33	0.31	0.28	0.29
	FDIN	0.16	0.25	0.22	0.21	0.21
Level2	AFD+CNN	**0.09**	**0.17**	**0.15**	**0.14**	**0.14**
	FD+CNN	0.12	0.23	0.17	0.15	0.17
	FWD	0.15	0.27	0.24	0.17	0.21
	FDIN	0.10	0.21	0.18	0.16	0.16

由图 3 可以看出:使用 NDCG@5 作为评价指标,AFD 算法在 4 台设备上的 NDCG 值均高于其他 3 种基准模型.其中,在 Movielens 数据集上,AFD 的 NDCG 平均值达到 0.92,FWD 的 NDCG 平均值为 0.82,FD 和 FDIN 的 NDCG 平均值接近(约为 0.85).AFD 比以上两种算法 NDCG 值提升了约 8%;在 Level2 数据集上,AFD 的 NDCG 平均值在 0.96,FWD 的 NDCG 平均值在 0.85,FD 和 FDIN 的 NDCG 平均值在 0.87.AFD 比以上两种算法 NDCG 值提升了 10%.



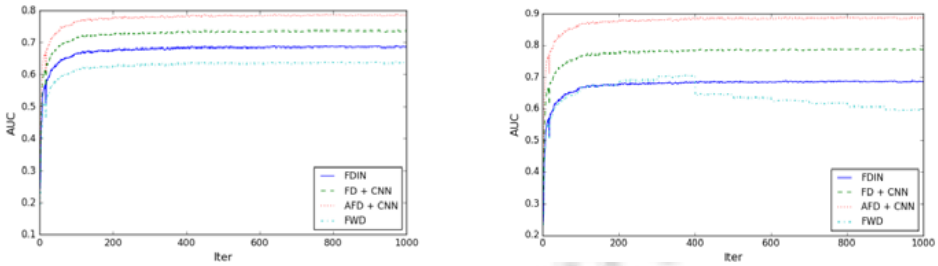
(a) Movielens (b) Level2

Fig.3 NDCG on different datasets

图 3 不同数据集下的 NDCG

由图 4 可以看出,AFD 算法 AUC 值均高于基准算法.其中,在 Movielens 数据集上,AFD 算法的 AUC 为 0.78;

在 Level2 数据集上,AFD 算法的 AUC 为 0.86,FDIN 和 FWD 算法的 AUC 仅为 0.66,FD+CNN 算法的 AUC 为 0.76.

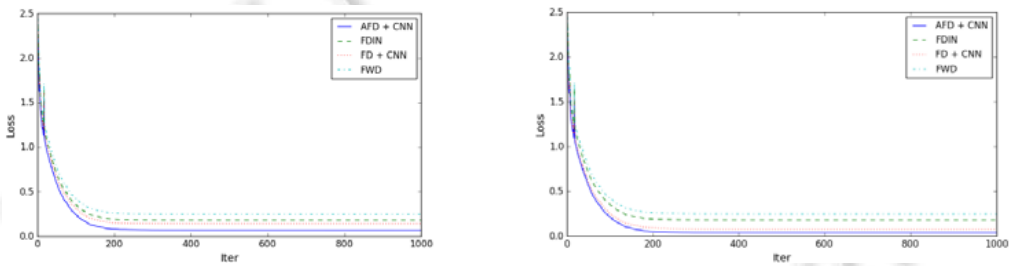


(a) Movielens (b) Level2

Fig.4 AUC on different datasets

图 4 不同数据集下的 AUC

由图 5 可以看出:随着迭代轮数的增加,AFD 可在迭代轮数小于 200 轮时收敛,收敛速度略优于其他 3 种算法.同时,AFD 在两个数据集上均取得了更低的损失:在 Movielens 数据集上,AFD 的 Loss 约为 0.2;在 Level2 数据集上,AFD 的 Loss 可达到 0.1 左右,均低于其他 3 种基准算法.以上实验结果表明:本文提出的 AFD 算法收敛速度更快,总体推荐性能更好.



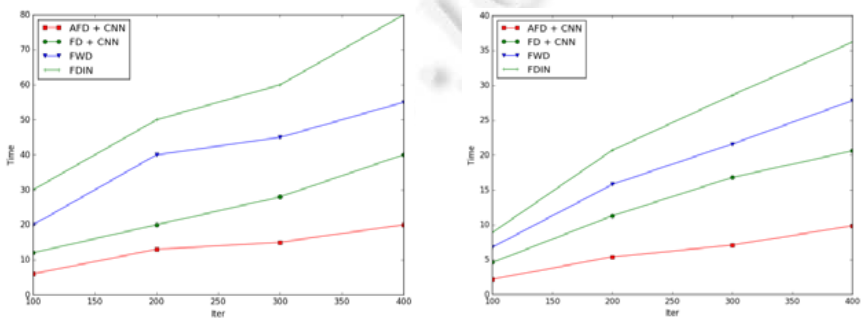
(a) Movielens (b) Level2

Fig.5 Loss on different datasets

图 5 不同数据集下的 Loss

- 实验 2:自适应学习率在联邦蒸馏中的有效性验证.

为了验证改进后的自适应学习率方法的有效性,将算法的运行时间作为评价指标,对比 AFD 与 FWD,FDIN 和 FD 不同迭代轮数下的运行时间.实验结果如图 6 所示.



(a) Movielens (b) Level2

Fig.6 Running time of algorithmson different datasets

图 6 不同数据集上算法运行时间

从结果中可看出:在 Movielens 数据集上,AFD 算法的耗时明显低于其他 3 种基准算法,耗时曲线较平缓;在

Level2 数据集上,FWD 和 FDIN 算法的运行时间较长,随着迭代轮数的增加,运行时长呈线性增长,FD 算法运行时长小于以上两种算法.而采用自适应学习率策略的 AFD 算法在相同轮数下耗时最短,同时,在 200 轮以后,运行时长曲线增长更缓慢.在迭代 400 轮左右,AFD 累计运行时长为 9.8 分钟,FD+CNN 运行时长为 20.6 分钟,AFD 算法较 FD+CNN 算法训练时间缩短 52%左右,说明自适应学习率的方法能够有效的提升训练速度.

- 实验 3:Attention 机制的有效性验证.

本实验将 AFD 中的 Attention 编码策略结合在其他 3 个基准模型中,分别为联邦蒸馏算法结合 CNN 及注意力机制(FD+CNN+ATN)、联邦学习结合 Wide&Deep 算法和注意力机制(FWD+ATN)和联邦学习结合深度兴趣网络和注意力机制(FDIN+ATN).将 NDCG、AUC、相同条件下训练时长(迭代次数 400,minibatch 大小 128,学习率 0.001)和设备端 MAE 作为对比指标,在 Movielens 数据集和 Level2 数据集上对比实验结果见表 6.

Table 6 Comparisons between baselines using attentional mechanism

表 6 各基准模型使用 Attention 机制后的效果对比

数据集	算法	NDCG	AUC	Time (分钟)	Global-MAE
Movielens	FD+CNN+ATN	0.88 (5%)	0.67 (13%)	41 (28%)	0.22 (−8%)
	FWD+ATN	0.86 (4%)	0.63 (15%)	64 (25%)	0.26 (−10%)
	FDIN+ATN	0.88 (2%)	0.73 (−1%)	92 (18%)	0.20 (−5%)
	AFD+CNN	0.92 (11%)	0.78 (8%)	13.4 (19%)	0.19 (−20%)
Level2	FD+CNN+ATN	0.92 (6%)	0.75 (20%)	23 (12%)	0.18 (6%)
	FWD+ATN	0.90 (6%)	0.63 (9%)	35 (29%)	0.19 (−9%)
	FDIN+ATN	0.92 (1%)	0.81 (1%)	46 (24%)	0.15 (−6%)
	AFD+CNN	0.96 (6%)	0.87 (7%)	9.8 (56%)	0.14 (−22%)

表 6 中,括号内的数字为加入 Attention 机制后的方法相比未加入之前方法的提升/减少幅度.NDCG 和 AUC 该数字越大越好,运行时间和 MAE 则越小越好.在 Movielens 数据集中,FD+CNN 加入注意力机制后,NDCG@5 值提升约 5%,AUC 值提升约 13%,Global-MAE 误差减少约 8%,相同条件下训练时长却增加了约 28%,说明加入注意力机制虽然对 FD+CNN 算法精度有明显提升,但增加了计算量.FDIN 加入注意力机制后,Global-MAE 有明显降低,但 NDCG 指标和 AUC 几乎不变,训练时长增加了约 18%,说明加入注意力机制对 FDIN 算法精度提升有限.这是由于 FDIN 已经在内部对集成了 Attention 操作.对比实验中除 FDIN 外,其他模型精度均有明显提升,但会增加算法的计算量,增加训练时间.从同花顺 Level2 数据分析,可以进一步得出相同的结论.

- 实验 4:Attention 编码后特征之间的关联性分析.

本实验对 Attention 编码后特征之间的关联性进行分析,结果见图 7.其中,图 7 的横纵坐标均为 Level2 用户和产品标签字段,颜色由浅到深表示两个特征的关联度逐级提高,关联度较高的特征能够获得较高的权重得分.

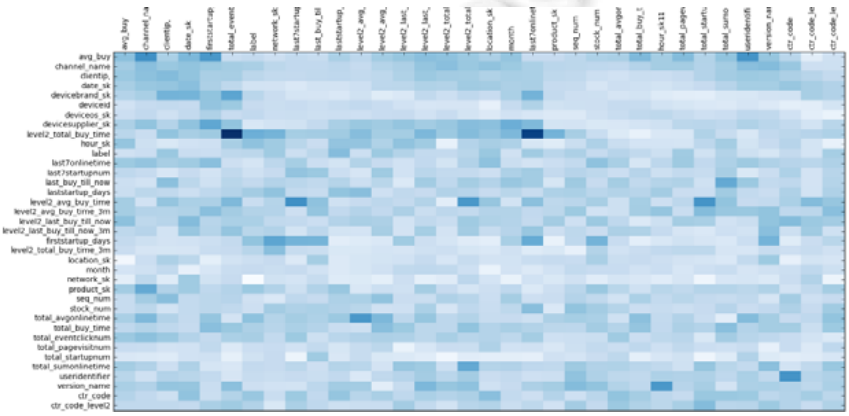


Fig.7 Visualization of feature interactions on Level2 dataset after attentional encoding

图 7 Level2 数据集下进行注意力编码后的特征交互可视化

可以看出,一些特征如 level2_total_buy_time(Level2 产品历史购买次数),total_eventclicknum(Level2 产品历

史点击次数),last7onlinetime(过去 7 天的在线时长)等之间存在较强的特征交互,表明用户活跃度如点击次数和在线时长等特征对产品购买影响较大,符合现实业务中的观察结论.该结果表明:本文提出的 Attention 策略可以提取出更丰富的特征表征信息(无需通过 Attention 网络进行训练),增强设备数据,提升模型精度.

- 实验 5:3 个改进策略对联邦蒸馏的有效性验证.

在最后一个实验中,验证本文 3 个策略对联邦蒸馏方法框架的贡献程度,分别为联邦蒸馏加入 KL 散度和正则项(FD+KLR)、联邦蒸馏加入改进后的注意力机制(FD+ATN)和联邦蒸馏中加入自适应学习率优化策略(FD+ADA).为了验证 3 个改进策略对联邦蒸馏的有效性,将 NDCG、AUC、相同条件下训练时长(迭代次数 400, minibatch 大小 128,学习率 0.001)和 MAE 作为对比指标.对比实验结果见表 7.

Table 7 Comparisons between three strategies

表 7 3 个策略的效果对比

数据集	算法	NDCG	AUC	Time (分钟)	Global-MAE
Movielens	FD	0.81	0.69	32	0.24
	FD+KLR	0.83 (2%)	0.72 (4%)	34.6 (8%)	0.23 (–4%)
	FD+ATN	0.87 (7%)	0.75 (8%)	37.2 (16%)	0.21 (–12%)
	FD+ADAM	0.81 (0%)	0.69 (0%)	12.7 (–60%)	0.24 (–)
	AFD	0.92 (13%)	0.78 (13%)	13.4 (–58%)	0.19 (–20%)
Level2	FD	0.88	0.76	20.6	0.17
	FD+KLR	0.90 (2%)	0.81 (6%)	22.6 (13%)	0.16 (–6%)
	FD+ATN	0.93 (5%)	0.84 (10%)	24.7 (20%)	0.15 (–12%)
	FD+ADAM	0.88 (–)	0.76 (–)	9.3 (–55%)	0.17 (–)
	AFD	0.96 (10%)	0.87 (14%)	9.8 (–52%)	0.14 (–17%)

从表 7 结果可以看出:在 Movielens 数据集中:比较加入改进算法前后的 NDCG@5 指标,AFD 最高为 0.92,FD+ATN 为 0.87,分值最低的是 FD+ADA 为 0.81;加入注意力机制比原始联邦蒸馏算法有约 7%的提升,其次是 FD+ KLR,相比原始联邦蒸馏算法有约 2%的提升;比较加入改进算法前后的 AUC 值,加入 FD+ATN 相比原始联邦学习算法有约 8%的提升;对比加入改进算法前后的 MAE,FD+KLR 和 FD+ATN 相比 FD+ADAM 误差减少了约 4%和 12%;从精度来看,提升最明显的是加入注意力机制(ATN),其次是引入 KL 和正则项的联合损失优化策略(KLR),而自适应学习率策略(ADA)对精度的提升有限;但从训练收敛速度角度,ADA 策略取得了最大的收益,较只加入 KLR 训练时间减少了约 60%,说明该策略能大大提升学生模型的训练速度;FD+ATN 耗时最多,说明 ATN 策略大幅提高了计算量;KLR 策略因只对目标函数做优化,对性能影响较少.从同花顺 Level2 数据结果分析可以进一步得出相同的结论.

综上所述,在联邦蒸馏框架中加入注意力机制可以大幅提升模型的性能;加入 KL 散度和正则项的联合优化策略可以减少特征之间的差异性带来的影响,从而提升模型的精度;最后,加入自适应学习率的训练策略在不损失或较小损失模型精度的情况下,可以大幅缩短模型的训练时间.加入 3 个改进策略后,本文提出的 AFD 在实验数据集上获得了最优的性能.

4 结 论

本文提出了一种改进的联邦蒸馏推荐方法,包括一个标准的模型优化联邦蒸馏算法.该算法引入了 3 种策略:(1) 为增强设备中的数据特征,引入了一个改进的注意力编码机制;(2) 针对设备间数据差异可能带来的影响,引入了一个评估学生模型与教师模型差异指标及正则项的联合优化方法;(3) 为抵消注意力编码机制带来的计算量提升,提出一个改进的自适应学习率方法来切换不同优化方法,选择合适的学习率来加快模型收敛速度,使得训练时间缩短了 52%左右.最后,通过实验在 Movielens 数据集和同花顺 Level2 线上数据集验证策略的有效性.实验结果表明:相比于 3 种基准算法,本文提出的算法相比于原始联邦蒸馏算法训练时间缩短 52%,模型的准确率提升了 13%,平均绝对误差减少了约 17%,NDCG 值提升了约 10%,展示了良好的收敛效率和推荐精度.在未来的研究中可尝试的方向是:将联邦蒸馏与强化学习结合起来,为不同的设备制定不同的策略,无需回传或

仅少量回传模型参数即可达到与回收模型相同的收敛效果,以大幅降低通信量.

References:

[1] Subramaniyaswamy V, Logesh R, Indragandhi V. Intelligent sports commentary recommendation system for individual cricket players. Int'l Journal of Advanced Intelligence Paradigms, 2018,10(1-2):103−117.

[2] Manogaran G, Varatharajan R, Priyan MK. Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. Multimedia Tools and Applications, 2018,77(4):4379−4399.

[3] Shin H, Kim S, Shin J, *et al*. Privacy enhanced matrix factorization for recommendation with local differential privacy. IEEE Trans. on Knowledge and Data Engineering, 2018,30(9):1770−1782.

[4] McMahan B, Moore E, Ramage D, *et al*. Communication-efficient learning of deep networks from decentralized data. In: Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: PMLR, 2017. 1273−1282.

[5] Bonawitz K, Ivanov V, Kreuter B, *et al*. Practical secure aggregation for privacy-preserving machine learning. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: ACM, 2017. 1175−1191.

[6] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. In: Proc. of the IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2019. 739−753.

[7] Radio Spectrum Policy Group. RSPG report on the results of the public consultation on the Reviewof the EU Telecommunications Framework. Technical Report, 2016. http://spectrum.welter.fr/international/rspg/reports/rspg-report-2016-framework-review.pdf

[8] Huang K, Zhu G, You C, *et al*. Communication, computing, and learning on the edge. In: Proc. of the IEEE Int'l Conf. on Communication Systems (ICCS). Chengdu: IEEE, 2018. 268−273.

[9] Song X, Feng F, Han X, *et al*. Neural compatibility modeling with attentive knowledge distillation. In: Proc. of the 41st Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2018. 5−14.

[10] Jeong E, Oh S, Kim H. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data. In: Proc. of the 2nd Workshop on Machine Learning on the Phone and other Consumer Devices (MLPCD 2). Montréal: JMLR, 2018.

[11] Luo L, Huang W, Zeng Q. Learning personalized end-to-end goal-oriented dialog. In: Proc.of the AAAI Conf. on Artificial Intelligence, Vol.33. Honolulu: AAAI, 2019. 6794−6801.

[12] Yang Q, Liu Y, Chen T, *et al*. Federated machine learning: Concept and applications. ACM Trans. on Intelligent Systems and Technology (TIST), 2019,10(2):1−19.

[13] Li T, Sahu AK, Talwalkar A, *et al*. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 2020,37(3):50−60.

[14] Smith V, Chiang CK, Sanjabi M. Federated multi-task learning. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Long Beach: Curran Associates, Inc., 2017. 4424−4434.

[15] Gao D, Liu Y, Huang A, *et al*. Privacy-preserving heterogeneous federated transfer learning. In: Proc. of the 2019 IEEE Int'l Conf. on Big Data. Los Angeles: IEEE, 2019. 2552−2559.

[16] Nadiger C, Kumar A, Abdelhak S. Federated reinforcement learning for fast personalization. In: Proc. of the 2019 IEEE 2nd Int'l Conf. on Artificial Intelligence and Knowledge Engineering (AIKE). Sardinia: IEEE, 2019. 123−127.

[17] Li Q, Wen Z, He B. Practical federated gradient boosting decision trees. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence (AAAI 2020). New York: AAAI, 2020. 4642−4649.

[18] Yurochkin M, Agarwal M, Ghosh S, *et al*. Bayesian nonparametric federated learning of neural networks. In: Proc. of the Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7252−7261.

[19] Liu Y, Kang Y, Xing C, *et al*. A secure federated transfer learning framework. IEEE Intelligent Systems. 2020,35(4):70−82. [doi: 10.1109/MIS.2020.2988525]

[20] Nadiger C, Kumar A, Abdelhak S. Federated reinforcement learning for fast personalization. In: Proc. of the IEEE 2nd Int'l Conf. on Artificial Intelligence and Knowledge Engineering (AIKE). Sardinia: IEEE, 2019. 123−127.

[21] Ke G, Meng Q, Finley T, *et al*. Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. Hangzhou: IEEE, 2017. 3146−3154.

[22] Sharma S, Chen K. Privacy-preserving boosting with random linear classifiers. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. Toronto: ACM, 2018. 2294−2296.

[23] Cho JH, Hariharan B. On the efficacy of knowledge distillation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4794−4802.

[24] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. In: Proc. of the NIPS Deep Learning and Representation Learning Workshop. 2015.

[25] Yim J, Joo D, Bae J, *et al*. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017. 4133−4141.

[26] Heo B, Lee M, Yun S. Knowledge distillation with adversarial samples supporting decision boundary. In: Proc. of the AAAI Conf. on Artificial Intelligence. Hawaii: AAAI, 2019. 33:3771−3778.

[27] Yang C, Xie L, Su C, *et al*. Snapshot distillation: Teacher-student optimization in one generation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2859−2868.

[28] Cha H, Park J, Kim H, *et al*. Federated reinforcement distillation with proxy experience memory. In: Proc. of the 1st Int'l Workshop on Federated Machine Learning for User Privacy and Data Confidentiality (FML 2019). 2019.

[29] Guo H, Tang R, Ye Y, *et al*. DeepFM: A factorization-machine based neural network for CTR prediction. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: IJCAI.org, 2017. 1725−1731.

[30] Hu C, Meng XW, Zhang YJ, *et al*. Enhanced group recommendation method based on preference aggregation. Ruan Jian Xue Bao/Journal of Software, 2018,29(10):3164−3183 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5288.htm [doi: 10.13328/j.cnki.jos.005288]

[31] Xiao J, Ye H, He X, *et al*. Attentional factorization machines: Learning the weight of feature interactions via attention networks. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: IJCAI.org, 2017. 3119−3125.

[32] Wang Q, Liu FA, Xing S, *et al*. A new approach for advertising CTR prediction based on deep neural network via attention mechanism. In: Proc. of the Computational and Mathematical Methods in Medicine. 2018.

[33] Li T, Sahu AK, Talwalkar A, *et al*. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 2020,37(3):50−60.

[34] Wang X, Han Y, Wang C, *et al*. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. IEEE Network, 2019,33:156−165.

[35] Wu B, Lou ZZ, Ye YD. Co-regularized matrix factorization recommendation algorithm. Ruan Jian Xue Bao/Journal of Software, 2018,29(9):2681−2696 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5274.htm [doi: 10.13328/j.cnki.jos.005274]

[36] Cui X, Zhang W, Tüske Z, *et al*. Evolutionary stochastic gradient descent for optimization of deep neural networks. In: Proc. of the Neural Information Processing Systems. Montréal: JMLR, 2018. 6048−6058.

[37] Cutkosky A, Orabona F. Momentum-based variance reduction in non-convex SGD. In: Proc. of the Neural Information Processing Systems. Jaipur: JMLR, 2019. 15236−15245.

[38] Wang Y, Zhou P, Zhong W. An optimization strategy based on hybrid algorithm of Adam and SGD. In: Proc. of the MATEC Web of Conf. 2018. 232.03007.

[39] Huang G, Liu Z, Van Der Maaten L. Densely connected convolutional networks. In: Proc.of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4700−4708.

[40] Li Q, Tai C, Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2017. 2101−2110.

[41] Bottou L. Stochastic gradient descent tricks. In: Proc. of the Neural Networks: Tricks of the Trade. Heidelberg: Springer-Verlag, 2012. 421−436.

[42] Harper FM, Konstan JA. The movielens datasets: History and context. ACM Trans. on Interactive Intelligent Systems (TIIS), 2015, 5(4):1−19.

[43] Yu H, Li JH. Algorithm to solve the cold-start problem in new item recommendations. Ruan Jian Xue Bao/Journal of Software, 2015,26(6):1395−1408 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4872.htm [doi: 10.13328/j.cnki.jos.004872]

[44] Cheng HT, Koc L, Harmsen J, *et al*. Wide & deep learning for recommender systems. In: Proc.of the 1st Workshop on Deep Learning for Recommender Systems. Boston: DLRS, 2016. 7−10.

[45] Zhou G, Zhu X, Song C, *et al*. Deep interest network for click-through rate prediction. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 1059−1068.

附中文参考文献:

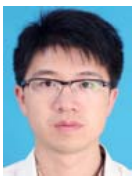
[30] 胡川,孟祥武,张玉洁,等.一种改进的偏好融合组推荐方法.软件学报,2018,29(10):3164−3183. http://www.jos.org.cn/1000-9825/5288.htm [doi: 10.13328/j.cnki.jos.005288]

[35] 吴宾,娄铮铮,叶阳东.联合正则化的矩阵分解推荐算法.软件学报,2018,29(9):2681−2696. http://www.jos.org.cn/1000-9825/5274.htm [doi: 10.13328/j.cnki.jos.005274]

[43] 于洪,李俊华.一种解决新项目冷启动问题的推荐算法.软件学报,2015,26(6):1395−1408. http://www.jos.org.cn/1000-9825/4872.htm [doi: 10.13328/j.cnki.jos.004872]

谌明(1983−),男,博士,主要研究领域为人工智能,机器学习,大数据.

马天翼(1986−),男,博士,主要研究领域为机器学习,联邦学习,推荐系统.



张蕾(1992−),男,硕士,主要研究领域为个性化推荐系统.