广告点击率预估的逐层残差交互网络

尹云飞[1),2)] 龙连杰[1)] 黄发良[2)] 吴开贵[1)]

[1)](重庆大学计算机学院 重庆 400044)

[2)](广西人机交互与智能决策重点实验室 南宁 530100)

摘 要 网络广告费的收取通常是以用户的点击次数来计算的,因此如何准确地预估点击率(CTR)是广告公司十分关心的问题.当前先进水平的方法集中在构建各种高阶特征交互模型来预估 CTR,但是高阶特征交互会丢失低阶信息,尤其是丢失原始特征的信息.为此,本文提出一个新的逐层残差交互网络,它在每次交互时都考虑原始特征的引导作用,被命名为逐层残差交互网(LRIN).LRIN 强调高阶特征交互应该建立在原始特征逐层交互的基础上.n 阶特征交互由原始特征与 $n-1$ 阶特征通过元素积运算得到.进而,本文引入了多尺度方法来设计注意力网络.受逐层交互的影响,注意力网络也被设计成多层,称之为逐层注意力网络.为了将二者结合起来,本文提出将逐层残差交互网络的输出作为逐层注意力网络的权重,由此形成了一种新的双网络训练模型.在多个 benchmark 数据集上的实验结果表明,LRIN 的性能比当前先进的方法在 Criteo 数据集上平均提高 1.24%,在 Avazu 数据集上平均提高 2.16%,在 MovieLens-1M 数据集上平均提高了 1.3%,在 Book-Crossing 数据集上平均提高了 1.27%.

关键词 残差网络;逐层;特征交互;CTR 预估;注意力

中图法分类号 TP301 **DOI 号** 10.11897/SP.J.1016.2024.00575

Layer-by-Layer Residual Interactive Network Approach for Advertisement Click-Through Rate Prediction

YIN Yun-Fei[1),2)] LONG Lian-Jie[1)] HUANG Fa-Liang[2)] WU Kai-Gui[1)]

[1)](*College of Computer Science, Chongqing University, Chongqing* 400044)

[2)](*Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision, Nanning* 530100)

Abstract Online advertising fees are charged based on the number of times that users click on ads, and therefore how to accurately predict Click-Through Rate (CTR) is a very concerned issue for advertising companies. Current state-of-the-art methods focus on constructing various high-order feature interaction models to predict CTR; however, high-order feature interactions will lose low-order information, especially the information of original features. To this end, a novel layer-by-layer residual interaction network framework is proposed in this paper, which leverages the guiding role of the original features at each interaction, and is named as the Layer-by-layer Residual Interaction Network (LRIN). LRIN emphasizes that higher-order feature interactions should be based on the interactions of original features layer by layer. The interaction of n-order features is obtained by the element-wise product between the original features and the $n-1$-order features. Moreover, a multi-scale approach is introduced to design attention network. Affected by layer-by-layer interaction, the attention network is also designed into multiple layers, which is called layer-by-layer attention networks. In order to combine the two, this paper proposes to take the outputs of the layer-by-layer residual interaction network as the weights of the layer-by-layer

收稿日期:2023-04-02;在线发布日期:2023-12-27.本课题得到国家自然科学基金(61962038)、中央高校基本科研业务费项目(2022-CDJKYJH023)、广西人机交互与智能决策重点实验室开放基金项目(GXHIID2208)资助.**尹云飞**(通信作者),博士,副教授,主要研究方向为数据分析、人工智能.E-mail:yinyunfei@cqu.edu.cn.**龙连杰**,硕士研究生,主要研究方向为推荐系统.**黄发良**(通信作者),博士,教授,主要研究领域为社交网络、数据挖掘.E-mail:flhuang@cqu.edu.cn.**吴开贵**,博士,教授,主要研究领域为区块链与数据分析.

attention network, and thus forms a novel dual-network training model. The experimental results on multiple benchmark datasets indicate that the performance of LRIN is on average 1.24% better than current advanced methods on the Criteo dataset, 2.16% better on the Avazu dataset, 1.3% better on the MovieLens-1M dataset, and 1.27% better on the Book-crossing dataset.

Keywords residual network; layer-by-layer; feature interaction; CTR prediction; attention

1 引 言

网络广告是以互联网为载体的广告,它的收费与用户的点击率(CTR)有关[1].广告公司对于准确地预估网络广告的 CTR 十分感兴趣.网络广告点击率预测的特点是针对不同的用户能够提供不同的广告.与其他预测方法相比,网络广告点击率预测的准确度较低、处理的数据比较复杂.现有方法通常采用高阶特征交互来提高 CTR 预估的准确度[2],但是特征经过多次高阶交互后,其语义信息丢失严重.例如,"姓名"与"爱好"的交互结果是一种不存在的抽象特征,进而这种抽象的特征再相互交互,其结果会是更抽象的特征;此时,交互结果与原始的"姓名"与"爱好"关联很少.

鉴于这些实际的问题,近几年,高阶特征交互方法的改进得到了越来越多的研究,其中交互算子设计是一个研究的热点.早期的工作试图利用内积、哈达玛积、外积、Kronecker product 等来提高高阶特征交互的性能[3-4].但是,高阶交互不但复杂度高而且也使原始信息丢失严重.Luo 等人[5]改进了高阶特征交互的信息丢失问题,提出了基于特征感知的交互方法.Wang 等人[6]将传统的特征交互运算扩展为可加可乘的运算,并使用 MaskBlock 结构来关注重要的特征.集成学习也是一种缓解交互信息丢失问题的方法,Xue 等人[7]提出了基于自适应哈希算法来选择重要特征.近年来,还出现了高阶特征交互和低阶特征交互相结合的神经网络方法.无论是改进交互算子还是改进高阶交互神经网络结构,其本质都是让相似的特征得到增强而不相似的特征得到削弱.高阶特征交互的优点是能快速地区分特征,但是,这样容易丢失原有信息.低阶特征交互的优点是能较好地保留原始特征的信息,但是它的分类能力较弱.文献[8]较好地将高阶特征交互和低阶特征交互结合起来,并在此基础上提出了分层因子分解机的特征交互方法.

在图 1 中,左边是通常的特征交互过程,右边是逐层特征残差交互过程.图中涉及了 f_1, f_2, $f_{1,2}$, $f_{1,1,2}$等记号,其中 f_1和 f_2是最初的特征,f_1和 f_1交互的结果记为 $f_{1,1}$,f_1和 f_2交互的结果记为 $f_{1,2}$,f_1和 $f_{1,2}$交互的结果记为 $f_{1,1,2}$,$f_{1,1}$和 $f_{1,2}$交互的结果记为 $f_{1,1,1,2}$,其他以此类推.显然,普通交互是在 $\{f_1, f_2\}$的基础上进行两两交互生成$\{f_{1,1}, f_{2,2}, f_{1,2}\}$,进而,又在$\{f_{1,1}, f_{2,2}, f_{1,2}\}$的基础上进行两两交互.与此不同的是,逐层特征残差交互是 f_1(或者 f_2)与 $\{f_{1,1}, f_{2,2}, f_{1,2}\}$ 交互生成 $\{f_{1,1,1}, f_{1,2,2}, f_{1,1,2}\}$,然后,$f_1$再与$\{f_{1,1,1}, f_{1,2,2}, f_{1,1,2}\}$交互生成$\{f_{1,1,1,1}, f_{1,1,2,2}, f_{1,1,1,2}\}$……,显然,逐层特征残差交互能在一定程度上缓解低阶特征语义丢失问题.

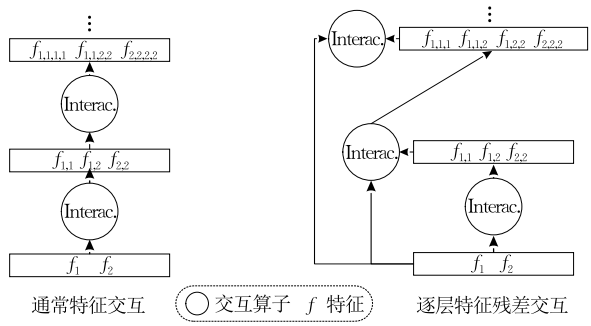

图 1 特征交互的过程

尽管取得了这些成功,但尚未考虑原始特征逐层地去引导的高阶特征交互,并将这种逐层引导机制与逐层注意力机制结合起来,这能够潜在地提供另一种高阶特征交互的手段.这促使本文探索逐层残差交互机制和注意力机制,它们能够相互配合而构成一种新的高阶特征交互方法——LRIN.如前所述,本文的设计动机是缓解高阶交互时低阶语义的丢失问题,因此,本文将每一次高阶交互的结果再次与原始特征交互,让原始特征引导高阶交互,由此形成逐层残差交互网络,如图 1 所示.显然,这种方法在已有工作[5-7]中并没有被提及.本文的另一个设计动机是有效地利用特征交互的结果来提高 CTR 预测的准确度,因此,本文引入了注意力机制,其中注意力权重取自逐层残差交互网络.据了解,这种方法目前尚未有文献提及.此外,本文还通过 MLP(Multi-Layer Perceptron)进行复杂的高阶非线性特征交互,并将

神经网络每一层的输出拼接起来作为最终的输出,这是为了充分利用高阶特征交互的优势.实验结果表明,本文的这些改进取得了很好的性能.本文提出的逐层交互残差网络不仅可以用于网络广告点击率预测,而且还能用于特征选择、语义增强、基于注意力机制的自然语言处理等领域.

综上所述,由于特征之间的相互影响,特征两两交互后生成的新特征会远离原始特征的含义.随着任意特征之间迭代式的两两交互,这种偏离的速度急剧提升.为了降低这种偏离速度,本文改变了原有的基于任意特征的两两交互机制,提出了逐层的特征交互,即 n 阶特征交互由原始特征与 $n-1$ 阶特征通过元素积运算得到.为了尽早利用逐层交互的结果,本文把同层的特征交互结果作为注意力去引导 CTR 预测网络的训练.因此,本文的主要工作在于引入了逐层残差,并设计了逐层残差引导的逐层注意力网络.本文的主要贡献可以归纳如下:

(1)本文提出逐层残差交互网络,它在每次交互时都考虑原始特征的引导作用,避免信息过早丢失.

(2)本文首次将逐层残差交互网络与逐层注意力网络结合起来,使逐层残差网络的每一层输出作为逐层注意力网络的权重,从而有利于进行多尺度的 CTR 预估,提高了预估性能.

(3)4 个公开数据集上的实验表明,本文的模型优于基线模型,取得了先进的性能.

本文第 2 节是相关工作的文献综述;在第 3 节中,介绍了本文提出的 LRIN 模型,包括问题描述、模型架构、组件设计、复杂度分析等;在第 4 节中,进行了实验研究,比较 LRIN 与其它先进的方法,并进行了消融研究,以证明其有效性;第 5 节是本文的研究结论和未来的研究方向.

2　相关工作

2.1　基于逻辑回归的模型

早期的方法是逻辑回归(LR)模型[9].LR 通过拟合历史数据得到线性方程(或者线性方程组),然后 LR 把实时数据代入线性方程以预估结果.Yin 等人[10]提出一种改进的逻辑回归模型,称为 Coupled Logistic Regression model.Wang 等人[11]将逻辑回归与神经网络结合起来,提出了自适应逻辑回归模型.Nanda 等人[12]将分类与逻辑回归结合起来提出了 Robust CTR 预估.LR 是 CTR 预估的基线方法,但是这是一种过拟合的方法.LR 在实际应用中很少使用,这是因为不同领域、不同时刻的数据有很大的差别,这导致 LR 方法在公开数据集上的预估准确度很低[3].

2.2　基于因子分解机的模型

因子分解机(FM)模型是一种二阶特征交互模型[13],它在特征两两交互的基础上使用内积来模拟特征交互.FM 综合了矩阵分解和支持向量机的优点,能够有效解决特征交互的稀疏性问题.He 等人[14]设计了一种双线性特征交互算子去改造了 FM 模型,取消了 FM 模型最后的特征求和操作.Juan 等人[15]提出了将特征分成不同的域(Fields),通过不同域中的特征交互来构建 FM 模型,并为每一个特征在域中学习了一个嵌入表征.Xiao 等人[16]提出了基于注意力机制的 FM 模型,其中注意力机制被用来学习特征交互的重要性.Guo 等人[17]提出并行学习一阶特征交互和二阶特征交互的思想.Lian 等人[18]在 DeepFM 的基础上,使用 CIN 网络替代原来的显式交互模块.Yu 等人[19]提出了样本感知 FM 的思想.Lu 等人[20]在样本感知 FM 的基础上,引入了 Transformer 机制.Long 等人[8]提出了层次结构的 FM 模型.Zhou 等人[21]提出了基于位置投影的 FM 思想.然而,受限于 FM 的多项式拟合时间,基于 FM 的模型仅适合于低阶特征交互.尽管已经有人提出 FM 之间可以相互交互形成高阶交互,但是模型的海量参数和低效的处理时间使这些模型无法成为有效的高阶交互模型.

2.3　基于高阶交互的模型

基于高阶交互的模型是通过深度神经网络来对原始特征进行多次交互(即高阶交互),它将深度学习方法引入 CTR 预估框架.Zhang 等人[3]使用 DNN 对 CTR 数据集的特征进行多次交互.Qu 等人[4]研究了基于内积操作的显式特征交互和基于外积操作的显式特征交互.Cheng 等人[22]提出了阶数自适应调整的特征交互方法.Cheng 等人[23]结合了 LR 方法和深度神经网络方法.Wang 等人[24]提出了一种显式特征交互和隐式特征交互相结合的方法.Zhao 等人[25]设计了一个注意力结构,通过它来计算嵌入向量对模型的影响.Wang 等人[6]设计了 MaskBlock 结构,它可以构建 MaskNet 以进行复杂的特征交互.Chen 等人[26]提出了多样性增强的高阶特征交互网络模型.尽管基于深度神经网络的模型可以大幅度提高 CTR 预估的性能,但是高阶交互造成了语义的严重丢失.原始特征经过多次交互后,原来的语义信息基本上丢失殆尽.

3 LRIN 模型

3.1 问题描述

CTR 是网络广告领域常用的广告费计算模式. 当网络用户点击了某个网络广告时,广告主就要向广告公司支付一笔广告费用[27-28]. CTR 预估的概念模型可以近似地公式化为 $\hat{y}=\mathcal{F}(\boldsymbol{x}|\boldsymbol{\theta})$,其中 $\boldsymbol{x}$ 是输入信息、$\boldsymbol{\theta}$ 是模型参数、$\hat{y}$ 是预估值. CTR 预估作为一个研究问题,它的输入信息有三类:物品信息、用户信息和场景信息. 在实际的 CTR 预估情形下,物品信息是指用户点击的具体广告,例如用户点击了一本图书,物品信息就是该图书的所有信息;而用户信息就是该用户的所有信息;场景信息就是点击的时间、点击后是否购买及购买记录等信息. 在互联网范围内,这种信息是异常巨大的. 因此,$\hat{y}$ 的实际意义是根据相关的用户信息、物品信息和场景信息,预测某用户未来最有可能点击的广告. CTR 预估的任务是建立预估模型. 由于输入信息包括离散特征和连续特征,其中连续特征也可以离散化,因此输入信息可以通过 one-hot 或者 multi-hot 进行编码,如图 2 所示[29].

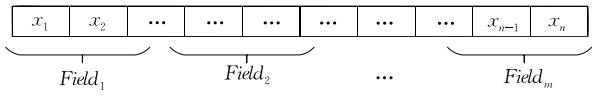

图 2 CTR 的信息表征

在图 2 中,$x_1,\cdots,x_n$ 是 n 个特征. 将这 n 个特征划分为 m 个特征域,即 $Field_1,Field_2,\cdots,Field_m$. 令 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m$ 分别是 $Field_1,Field_2,\cdots,Field_m$ 的向量代表,显然它们是划分出的子向量. 记 $\boldsymbol{x}=[x_1,\cdots,x_n]$,$\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m]$,$\boldsymbol{X}_i\in Field_i$,则 $\boldsymbol{X}$ 与 $\boldsymbol{x}$ 是等价的. 但是,$m\ll n$ 带来了 $\boldsymbol{X}$ 的更方便处理,并且每个特征能够在自己的特征域中进行局部的表征和优化. CTR 预估的本质是一个二分类问题,其标签值 $y\in\{0,1\}$ 表示用户的点击行为. 当 $y=1$ 时,表示用户点击了目标物品;当 $y=0$ 时,表示用户没有点击目标物品.

3.2 提出的 LRIN 模型

在本节中,对提出的 LRIN 模型进行详细的介绍,如图 3 所示. LRIN 模型包括输入层、嵌入层、逐层残差交互层、输出层等. 紧接着,还将探讨模型训练并分析模型的复杂度.

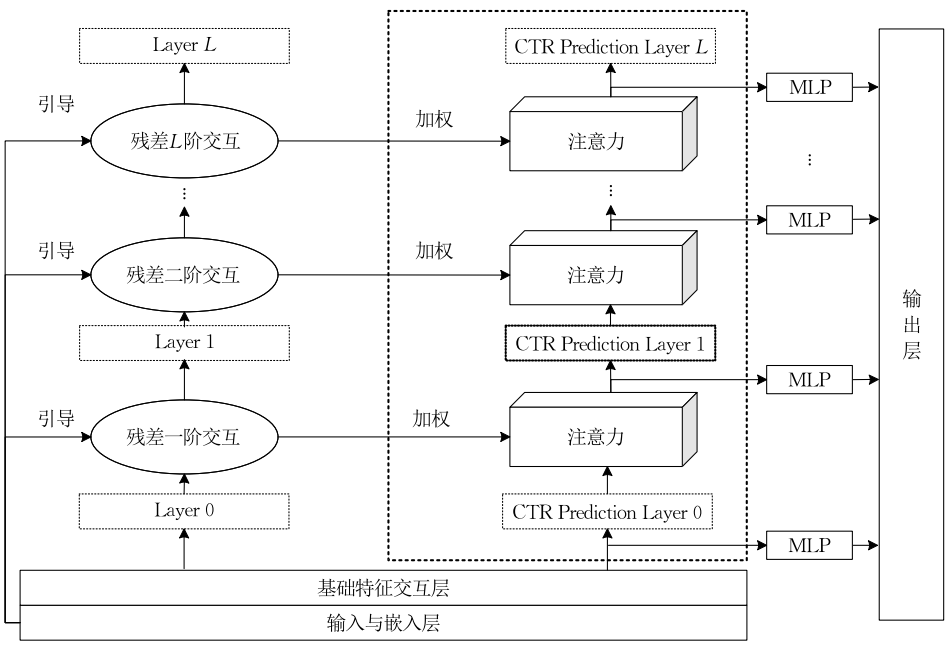

图 3 LRIN 模型

在图 3 中,左边是逐层残差网络,右边是逐层注意力网络. MLP 是 Multi-Layer Perceptron. 从模型的下面到上面,依次为特征嵌入层、基本特征交互层和双网络协同训练层. Output 为输出层. 主要部分是双网络协同训练层,它包括逐层残差交互网络和注意力机制. 原始嵌入特征经过两两交互后,其残差再由原始特征引导,产生引导权重. 该权重作用于 CTR 预估神经网络,形成注意力机制,这种注意力机制在神经网络的每一层都会产生. 在左边的逐层残差交互部分,特征交互的结果在各层之间传递;在每层,原始嵌入特征均发挥引导作用,形成残差交互效应. 在第 0 层,原始特征引导特征交互,称为 1 阶

残差交互;在第 1 层,原始特征引导 1 阶残差特征交互,形成 2 阶残差交互;依此类推.右边是基于注意力机制的 CTR 预估模型,它使用 MLP 进行更复杂的高阶非线性特征交互.最后将每一层的输出被拼接起来得到基于注意力机制的 CTR 预估输出.

3.2.1　输入拼接与嵌入层

CTR 预估的输入是用户点击事件的记录,称为 Instance.例如{用户标识:用户 id;性别:女;年龄:25 岁;学历:本科;物品标识:物品 id;物品类型:生活用品;物品价格:54;购买时间:2022-3-1 晚上 11 点}.从这个实例可以发现,一条记录可能包含很多信息,例如用户的特征(性别、年龄等),物品的属性(类型、价格等)以及上下文环境信息(时间、地点等).因此,需要对不同信息进行拼接.

通常,拼接后的信息是稀疏向量.假设拼接后的输入信息为 $\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m]\in\mathbb{R}^n$,如图 2 所示.由于 $\boldsymbol{X}$ 非常稀疏和高维,本文设计了嵌入层,嵌入层将稀疏特征映射成低维、密集的向量.嵌入层引入特征域机制,其输出表示为 $[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m]\in\mathbb{R}^{k\times m}$,其中 $\boldsymbol{v}_i\in\mathbb{R}^k$,如图 2 所示.$\boldsymbol{v}_i$ 表示当前输入在 $Field_i$ 中的嵌入表示,m 是特征域的数量,k 是维度.

3.2.2　基础特征交互层

在基础特征交互层,各个特征域的嵌入特征被显式地进行两两交互运算,如式(1)所示.

$$fI[i,j]=\sum_{f=1}^{k}\langle v_{i,f},v_{j,f}\rangle \tag{1}$$

在式(1)中,fI 是基础交互运算,$\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 分别表示两个嵌入向量,$\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle$ 是大小为 k 的两个嵌入向量的内积,用来刻画第 i 个和第 j 个特征之间的相互作用.根据文献[13],式(1)可以进一步的简化为式(2).

$$fI(\boldsymbol{x})=\frac{1}{2}\sum_{f=1}^{k}\left(\left(\sum_{i=1}^{n}v_{i,f}x_i\right)^2-\sum_{i=1}^{n}v_{i,f}^2x_i^2\right) \tag{2}$$

其中,$fI(\boldsymbol{x})$ 表示基础特征交互的结果.

3.2.3　逐层残差交互层

在现实应用中,不同的特征对用户的选择有不同的影响.例如,一个用户可能出于喜欢某个导演而选择该导演的某部电影,也可能出于喜欢某个演员而选择该演员的某部电影.当然,也可能选择该导演或演员的某类电影,例如爱情电影.特征交互也类似,经过多次特征交互就可以选择出更重要的特征.但是,多次特征交互会误导特征的选择方向.本文使用逐层残差交互层来防止这种误导,如图 4 所示.

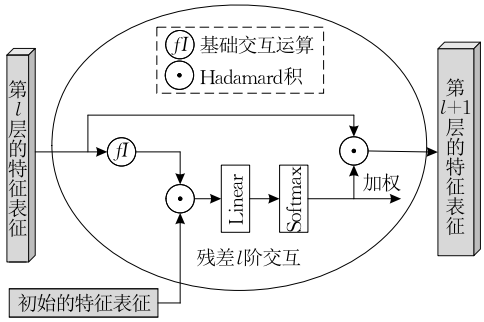

图 4　原始特征引导的逐层残差交互

图 4 展示了第 l 层的残差交互模型,$l=1,2,\cdots,L$.fI 是基础交互运算,它对第 l 层的特征进行两两交互,其计算方法参见式(1)和式(2).$\odot$ 是 Hadamard 积,通过它,原始特征能引导 fI 的后续交互.Linear 表示线性层用于学习参数,Softmax 表示激活函数.Weighting 表示对 CTR 预估进行加权.

根据逐层残差交互的思想,本文使用评估函数 $g:\mathbb{R}^k\times\mathbb{R}^k\to\mathbb{R}$ 计算初始嵌入特征 $\boldsymbol{v}_i$ 与 fI 之间的引导得分,如式(3).

$$\boldsymbol{\tau}_{\boldsymbol{v}_i}^{(l)}=\boldsymbol{W}^{(l)}(\boldsymbol{v}_i\odot\boldsymbol{fI}^{(l)})+\boldsymbol{b}^{(l)} \tag{3}$$

其中,其中 $\boldsymbol{v}_i\in\mathbb{R}^k$ 是初始嵌入特征的第 i 个特征向量,$\boldsymbol{fI}^{(l)}\in\mathbb{R}^k$ 表示神经网络的第 l 层的基础交互,l 是当前层,k 是维度.$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是权重和偏置.为了更清晰地描述特征之间的差异,可以将式(3)规范化为式(4).

$$\tilde{\boldsymbol{\tau}}_{\boldsymbol{v}_i}^{(l)}=\frac{\exp(\boldsymbol{\tau}_{\boldsymbol{v}_i}^{(l)})}{\sum\limits_{\boldsymbol{v}_i\in\boldsymbol{x}}\exp(\boldsymbol{\tau}_{\boldsymbol{v}_i}^{(l)})} \tag{4}$$

然后,通过线性层 Linear 和激活函数 Softmax 对 $\tilde{\boldsymbol{\tau}}_{\boldsymbol{v}_i}^{(l)}$ 进行处理,得到 $\tilde{\boldsymbol{\tau}}^l$,如图 3 所示.接着,根据式(5),$\tilde{\boldsymbol{\tau}}^l$ 与 l 层的特征 $\boldsymbol{x}^{(l)}$ 进行迭代运算就得到了 $l+1$ 层的特征表征.

$$\boldsymbol{x}^{(l+1)}=\tilde{\boldsymbol{\tau}}^{(l)}\boldsymbol{x}^{(l)} \tag{5}$$

3.2.4　基于注意力的 CTR 预估

由于 LRIN 是通过逐层残差交互网络完成从低阶到高阶传递信息,因此在 LRIN 模型的 CTR 预估部分,需将 $\tilde{\boldsymbol{\tau}}^l$ 视为权重,形成加权的 CTR 预估神经网络.这种设计体现了显式交互和隐式交互相融合的思想,是一种“显中有隐,隐中有显”CTR 预估模型,如式(6)所示.

$$\boldsymbol{x}^{(l+1)}=\boldsymbol{W}_{\mathrm{CTR}}^{(l)}\tilde{\boldsymbol{\tau}}^{(l)}\boldsymbol{x}^{(l)}+\boldsymbol{b}_{\mathrm{CTR}}^{(l)} \tag{6}$$

其中,$\boldsymbol{x}^{(l)}$ 和 $\boldsymbol{x}^{(l+1)}$ 是 CTR 预估网络的第 l 层和第 $l+1$ 层的结果.$\tilde{\boldsymbol{\tau}}^{(l)}$ 是注意力权重.$\boldsymbol{W}_{\mathrm{CTR}}^{(l)}$ 和 $\boldsymbol{b}_{\mathrm{CTR}}^{(l)}$ 分别是 CTR 预估网络的第 l 层权重和偏置.

3.2.5 输出层

如前所述,在 LRIN 的每一层都产生了一个残差交互权重,该权重作用于 CTR 预估神经网络的相关层.为了获得准确度的进一步提高,本文在 CTR 预估网络的每一层引入了 MLP,它用来完成复杂的高阶非线性交互,如图 3 和式(7)所示.

$$\begin{aligned}&\boldsymbol{o}_1^{(l)}=\delta(\boldsymbol{W}_1\boldsymbol{fI}^{(l)}+\boldsymbol{b}_1)\\&\boldsymbol{o}_2^{(l)}=\delta(\boldsymbol{W}_2\boldsymbol{o}_1^{(l)}+\boldsymbol{b}_2)\\&\cdots\\&\boldsymbol{o}_m^{(l)}=\delta(\boldsymbol{W}_m\boldsymbol{o}_{m-1}^{(l)}+\boldsymbol{b}_m)\\&\boldsymbol{o}^{(l)}=\boldsymbol{h}^{\mathrm{T}}\boldsymbol{o}_m^{(l)}+\boldsymbol{b}\end{aligned}\tag{7}$$

其中,$\boldsymbol{o}^{(l)}$是第 l 层的输出结果,m 是 MLP 的层数.$\boldsymbol{W}_i$和 $\boldsymbol{b}_i$分别是第 i 层的权重和偏置,$\boldsymbol{h}$ 和 δ 分别是投影权重和激活函数.

为了获得 LRIN 的最终输出,需要将 LRIN 每一层的输出拼接起来,如式(8)所示.

$$\boldsymbol{O}=[o^{(0)},o^{(1)},o^{(2)},\cdots,o^{(l)},\cdots,o^{(L)}]\tag{8}$$

其中,L 是 LRIN 的层数.需要注意的是,LRIN 的层数从 0 开始.当层数为 0 时表示没有采用逐层残差交互机制,而是直接将原始特征二阶交互进行输出,这与 NFM 模型[14]等效,如图 3 所示.

与通常的 CTR 预估一样,通过投影权重与偏置机制,并使用激活函数将它投影到预测分数 $\hat{y}$.最后的点击率预估值可以表示为

$$\hat{y}(\boldsymbol{x})=\delta(\boldsymbol{WO}+\boldsymbol{b})\tag{9}$$

其中,δ 是 sigmoid 激活函数,$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是投影权重和偏置.采用交叉熵损失函数,可以得到模型最终的优化目标函数

$$\min_{\Theta}\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}y_i\lg\hat{y}_i+(1-y_i)\lg(1-\hat{y}_i)+\lambda\|\Theta\|_2^2\tag{10}$$

其中,Θ 为总参数空间,包括所有的嵌入层、交互层、逐层注意力网络层的参数.N 为训练实例总数,$\lambda\|\Theta\|_2^2$ 为 $L2$ 正则化项.在模型训练时采用了 Dropout 和 BN 技术.

3.3 模型复杂度分析

模型复杂度分析是指对神经网络所含参数数量的分析[30].为了比较各种相关 CTR 预估模型的复杂度,本文列出了 LR、FM、NFM、FFM、AFM、DeepFM、xDeepFM、IFM、DIFM、FNN、AFN、WD、DCN、PNN、DRM、MaskNet、HAFM、LRIN 等模型的复杂度,如表 1 所示.

表 1 模型的复杂度比较(模型的复杂度由模型所含参数的数量决定)

Models	Number of parameters
LR[10]	n
FM[13]	$n+nk$
NFM[14]	$n+nk+\rho(k)$
FFM[15]	$n+n(m-1)k$
AFM[16]	$n+nk+\frac{mk(m-1)}{2}$
DeepFM[17]	$n+nk+\rho(mk)$
xDeepFM[18]	$n+nk+\rho(mk)+\sum_{t=1}^{L_{\mathrm{CIN}}}D_t(1+mD_{t-1})$
IFM[19]	$n+nk+\rho(mk)$
DIFM[20]	$n+nk+\rho(mk)+mkD_{\mathrm{ATT}}+D_{\mathrm{ATT}}HL_{\mathrm{STACK}}$
FNN[3]	$n+nk+\rho(mk)$
AFN[22]	$n+nk+mL_{\mathrm{LNN}}+kL_{\mathrm{LNN}}+\rho(ksL_{\mathrm{LNN}})$
WD[23]	$n+nk+\rho(mk)$
DCN[24]	$nk+\rho(mk)+2mkL_{\mathrm{DCN}}+mk$
PNN[4]	$nk+\rho\left(\frac{m(m-1)}{2}+mk\right)$
DRM[29]	$nk+\rho(mk)+\rho(2mk+D_{\mathrm{last}})$
MaskNet[6]	$nk+mk+(R+N+1)(mk)^2+N\rho(mk)$
HAFM[8]	$nk+(L_{\mathrm{HAN}}+1)\rho(k)+m^2kL_{\mathrm{HAN}}$
TAOA[1]	$nk+mk+E_{\mathrm{ctr}}[SumTopK(\{b_i\times Ctr_i\})]$
DICN[2]	$o(mDH+TH^2)$
LRIN	$nk+(L_{\mathrm{HAN}}+1)[\rho(k)+m^2k^2L_{\mathrm{HIN}}+mk^2]+m^2kL_{\mathrm{HAN}}$

在表 1 中,n 代表特征数目,m 代表特征域的数目.LR 模型的复杂度最低,时间复杂度为 $\mathcal{O}(n)$,参数数量仅为 n.FM 模型中的参数个数为 $n+nk$,其中 k 表示线性部分中每个特征的参数数目,nk 表示嵌入层的参数数目.FM 模型的复杂度为 $\mathcal{O}(kn)$.同理,IFM模型、FNN模型、WD模型的复杂度为 $\mathcal{O}(kn)$;NFM 模型、DeepFM 模型、xDeepFM 模型、DIFM 模型、AFN模型、DCN模型、PNN模型、DRM模型、MaskNet模型、HAFM模型的时间复杂度大于 $\mathcal{O}(kn)$ 而小于 $\mathcal{O}(kn^2)$;FFM 模型、AFM 模型的时间复杂度为 $\mathcal{O}(kn^2)$.

在 LRIN 模型中,模型参数是 $nk+(L_{\mathrm{HAN}}+1)[\rho(k)+m^2k^2L_{\mathrm{HIN}}+mk^2]+m^2kL_{\mathrm{HAN}}$,其中,$L_{\mathrm{HAN}}$ 表示残差交互网的层数,L_{HIN} 表示逐层注意力网络的层数.LRIN 模型的时间复杂度大于 $\mathcal{O}(kn)$ 而小于 $\mathcal{O}(kn^2)$,它与 NFM、DeepFM、xDeepFM、DIFM、AFN、DCN、PNN、DRM、MaskNet、TAOA、DICN、HAFM 属于同一个数量级,但是 LRIN 的性能远超这些模型.

4 实 验

通过上面的理论分析和模型设计,已经展示了 LRIN 模型的先进性.在本节中,进行实验研究.通过

实验,本文试图回答下面 3 个问题:(1)在性能方面,LRIN 模型是否比现有先进的模型好;(2)在模型组成部分的有效性方面,消融实验是否能证明模型的相关组成部分是有效的;(3)逐层残差交互网络和逐层注意力网络的层数对模型会产生什么影响.

4.1　实验设置

4.1.1　数据集

Criteo① 数据集是一个广告 CTR 预估数据集,由 Kaggle 平台公开发布[31]. 该数据集有 45 840 617 条用户点击广告的记录,其中每一条点击记录包含 26 个分类特征字段和 13 个数值特征字段.

Avazu② 数据集是由 Avazu 公司发布在 Kaggle 平台,它是一个关于广告 CTR 预估的数据集[32]. 该数据集有 40 428 967 条记录,每条记录包括 23 个特征字段.

MovieLens-1M③ 数据集是经典的电影评分数据集[33]. 该数据集有 1 000 209 条记录,涉及 6040 名匿名用户、3900 部电影等信息. 在研究中,本文对电影的评分进行了二值化,具体做法是:对于电影评分小于等于 3 的评分记录,将其标签值设为 0;对于电影评分大于 3 的评分记录,将其标签值设为 1.

Book-Crossing④ 数据集是一个公共数据集,它来源于图书交流网络社区[34]. 该数据集有 1 149 780 条评分数据,涉及 278 858 个匿名用户、271 379 本书等信息. 在研究中,本文对评分数据进行了二值化. 具体方法设计如下:评分小于等于 5 的评分记录,其标签值设为 0,评分大于 5 的评分记录,其标签设为 1.

在上述 4 个数据集中,前两个数据集是与广告相关,后两个是与推荐相关的. 由于本文进行的广告点击率预估研究属于一种特定领域的推荐方法研究,因此后两个数据集可以验证本文方法的推荐性能. 这种选择数据集的动机是为了让本文的方法能在更多的领域中得到推广. 此外,为了方便实验对比,采用与现有先进的方法[6,8,35]一致的数据集划分方法,即按照 8∶1∶1 比例随机划分训练集、验证集和测试集.

4.1.2　基线模型

根据模型的原理和组成,基线模型可以分为三大类:逻辑回归模型、基于 FM 的模型和基于深度神经网络的模型.

(1)逻辑回归模型包括 LR[10] 等,它们是点击率预估中经典的基线模型.

(2)基于FM的模型包括:FM[13],它在 LR 的基础上,引入了特征二阶交互操作;NFM[14],它在 FM 二阶交互后,取消求和操作;FFM[15],引入了特征域的概念,为每一个特征在每一个特征域都学习了一个嵌入表征;AFM[16],用逐层注意力网络为每一个交互特征学习一个权重;DeepFM[17],结合了 FM 和神经网络,既学习一阶特征交互又学习二阶特征交互;xDeepFM[18],在 DeepFM 的基础上,设计了一个 CIN 网络替代了显式交互模块;IFM[19],提出了样本感知思想;DIFM[20],在 IFM 的基础上,引入了 Transformer. HAFM[8],提出了层次结构的 FM.

(3)基于深度神经网络的模型包括:FNN[3],它利用 DNN 进行特征交互;AFN[22],它引入了自适应特征交互阶数调整方法;WD[23],它将深度学习网络和 LR 结合起来;DCN[24],它将特征的显式交互和隐式交互结合起来;IPNN[4],它采用内积(inner product)进行显式的特征交互;OPNN[4],它采用外积(outer product)进行显式的特征交互;DRM[29],它通过一个注意力结构来计算出每一个嵌入维度对于模型的重要性;MaskNet[6],它利用 MaskBlock 来进行复杂的特征交互. 最近,Bian 和 Wang[36] 提出了压缩交互网络组件和感知交互机组件,前者用于特征交互、后者用于特征编码. 实验表明这种方法的性能比 xDeepFM[18] 有一定提高. 基于图神经网络的 CTR 预测也是一类深度神经网络预测模型. 最近,Zhai 等人[37] 提出了一种图结构化网络模型来捕获高阶特征交互之间的因果关系.

4.1.3　评价指标

AUC 和 *Logloss* 是 CTR 预估任务中最常用的两个评价指标. *AUC* 是指 ROC 曲线下的面积,它是一种二分类模型性能的评价指标. *AUC* 可以表征预估的正例排在负例前面的概率,*AUC* 值越高表示性能越好. *Logloss* 是指交叉熵损失,可以表征真实值和预估值之间的差值. 在点击率预估中,*Logloss* 越小,表示模型预估越准确.

4.2　性能比较

性能比较实验基于 PyTorch 框架来训练所有模型. 遵照现有方法[3-4,6,8,14-20,24,29]的参数设置方法,设置实验参数. 为了比较的公平性,所有模型都采用 Adam 优化器,损失函数采用交叉熵损失函数,学习率在{0.0001,0.001,0.01}之间自动取值. 在嵌入层中,向量的维度 k 取 16,其他超参数与原始实验保持

① https://www.kaggle.com/c/criteo-display-ad-challenge/data
② https://www.kaggle.com/c/avazu-ctr-prediction/data
③ https://grouplens.org/datasets/movielens/1m
④ https://grouplens.org/datasets/book-crossing

一致. 对于含有 DNN 的模型,将 DNN 隐藏层的深度设为 2,隐藏层神经元的个数设为 16. 对于 Criteo 数据集,批处理大小(Batch Size)设置为 4096;对于 Avazu 数据集,Batch Size 设置为 2048;对于 Movielens-1M 和 Book-Crossing 数据集,Batch Size 均设置为 128. Dropout 在{0.1 ~ 0.9}之间取值,正则化参数 λ 在{0.00001,0.0001,0.001}之间取值.

在模型对比时,进行多组实验,再对结果取平均值. AFM 模型注意力因子大小设置为 16. DCN 模型的交互层层数设置为 3. xDeepFM 模型的 CIN 网络层参数与默认的 DNN 参数保持一致,设置为[16,16]. AFN 模型的对数神经网络维度设置为 1500. DIFM 模型中的多头注意力头数设置为 4,注意力层数设置为 3,注意力的维度大小设置为 16. MaskNet 模型的 MaskBlock 数目设置为 3. HAFM 模型采用 3 层注意力网络,LRIN 模型分别采用 3 层残差网络和 3 层逐层注意力网络. 模型的最大迭代周期(epoch)设置为 50,并采用早期停止法(step=2)避免模型过度拟合训练数据,缩短训练时间. 表 2 是 4 个数据集上 CTR 预估的 *AUC* 和 *LogLoss* 实验结果.

表 2 4 个数据集上 CTR 预估的性能比较

Models	Criteo		Avazu		MovieLens-1M		Book-Crossing	
	AUC	*LogLoss*	*AUC*	*LogLoss*	*AUC*	*LogLoss*	*AUC*	*LogLoss*
LR[10]	0.78963	0.45993	0.74984	0.39643	0.77994	0.56308	0.74279	0.54440
FM[13]	0.80165	0.44933	0.76711	0.38699	0.81292	0.51880	0.77532	0.52184
NFM[14]	0.80325	0.44771	0.76814	0.38630	0.79918	0.53497	0.77066	0.52170
FFM[15]	0.80603	0.44537	0.76966	0.38568	0.80975	0.52168	0.78588	0.50458
AFM[16]	0.80273	0.44730	0.77469	0.38416	0.80868	0.52340	0.77902	0.50795
DeepFM[17]	0.80681	0.44456	0.77510	0.38247	0.80454	0.52828	0.77193	0.52098
xDeepFM[18]	0.80879	0.44303	0.77195	0.38420	0.80376	0.52916	0.77511	0.52062
IFM[19]	0.79941	0.45289	0.78156	0.38010	0.80730	0.52292	0.78401	0.50410
DIFM[20]	0.80691	0.44557	0.78826	0.37501	0.81003	0.52007	0.78782	0.50042
FNN[3]	0.80086	0.45320	0.77109	0.38601	0.80154	0.53014	0.78315	0.50495
AFN[22]	0.80766	0.44375	0.77372	0.38312	0.79772	0.53884	0.76591	0.53030
WD[23]	0.80116	0.44986	0.77106	0.38481	0.80175	0.53037	0.76847	0.52234
DCN[24]	0.80324	0.44816	0.77299	0.38365	0.80526	0.52649	0.78365	0.50449
IPNN[4]	0.80754	0.44396	0.77579	0.38203	0.80629	0.52659	0.78667	0.50150
OPNN[4]	0.80863	0.44296	0.77676	0.38160	0.80811	0.52269	0.78600	0.50228
DRM[29]	0.80443	0.44698	0.78374	0.37764	0.80850	0.52132	0.78688	0.50117
MaskNetx[6]	0.80860	0.44330	0.77852	0.38059	0.80104	0.53492	0.78118	0.50719
MaskNety[6]	0.80847	0.44315	0.77905	0.38026	0.80106	0.53321	0.78131	0.50664
CausalGNN[37]	0.78440	0.45990	0.77280	0.37500	0.80553	0.52654	0.78501	0.52085
FRDFFM[36]	0.80880	0.44300	0.77193	0.38440	0.80380	0.52920	0.77510	0.52060
HAFM[8]	0.81118	0.44082	0.78933	0.37426	0.81293	0.51753	**0.78962**	**0.49834**
LRIN	**0.81382**	**0.43840**	**0.79117**	**0.37286**	**0.81473**	**0.51472**	0.78826	0.50008

注:*AUC* 和 *LogLoss* 是评估指标,黑色数字是最优结果.

在表 2 中,20 个模型在 4 个不同数据集上的性能被展示出来,其中,实验结果为多次实验后的平均值,最大方差小于 7×10^{-5}. 引入“下划线”标记 *top*3 的实验结果,引入“粗体”标记最优的实验结果. 对于每个数据集,使用两个评价指标 *AUC* 和 *Logloss* 来分别评估. 此外,与文献[2](2023)报道的性能比较,LRIN 模型在 Criteo 数据集上比它提高了 4.6%(*AUC*);在 Avazu 数据集上,LRIN 模型比它提高了 0.8%(*AUC*).

实验结果分析:

(1)从整体来看,本文提出的 LRIN 模型在 4 个数据集上都取得了较好的性能. 与最先进的模型相比,LRIN 在 *AUC* 指标上的改进为 2.23‰~6.22‰,平均改进了 3.61‰;在 *Logloss* 指标上的改进为 4.16‰~10.29‰,平均改进了 6.63‰.

(2)与 LR 模型相比,LRIN 在评估指标 *AUC* 上提升了 2.91%~6.13%,平均提升了 5.05%;在评估指标 *Logloss* 上提升了 3.91%~8.29%,平均提升了 6.55%. 由于缺少特征交互,LR 模型的性能较差. 因此,实验结果表明了特征交互操作对于 CTR 预估非常重要.

(3)与基于 FM 的高阶特征交互模型相比,LRIN 模型的表现显著地优于现有先进的模型. 这表明 LRIN 模型中逐层残差交互和注意力机制是有效的.

(4)与 HAFM 模型相比,在数据集 Criteo、Avazu、MovieLens-1M 上,LRIN 模型的表现远超 HAFM 模型. 在特征数目较少的 Book-Crossing 数据集上,HAFM 模型的表现稍微优于 LRIN 模型. 这很可能是因为

LRIN 模型是一种特征交互、注意力、MLP 的组合模型.由于在特征比较少时,特征交互对 CTR 预估的作用相对较小,所以 LRIN 模型的第一个环节即特征交互发挥得不够积极,导致后续环节的组合效果不佳.但是,随着特征数目的增加,LRIN 的优势迅速被发挥出来.这说明了,LRIN 模型更擅长于特征数目较多的数据集.

4.3　时间效率比较

由于 LRIN 模型增加了神经网络的层数、引入了逐层残差交互网络和逐层注意力网络,需要评估 LRIN 模型的时间效率.图 5 展示了现有先进模型的时间效率比较.

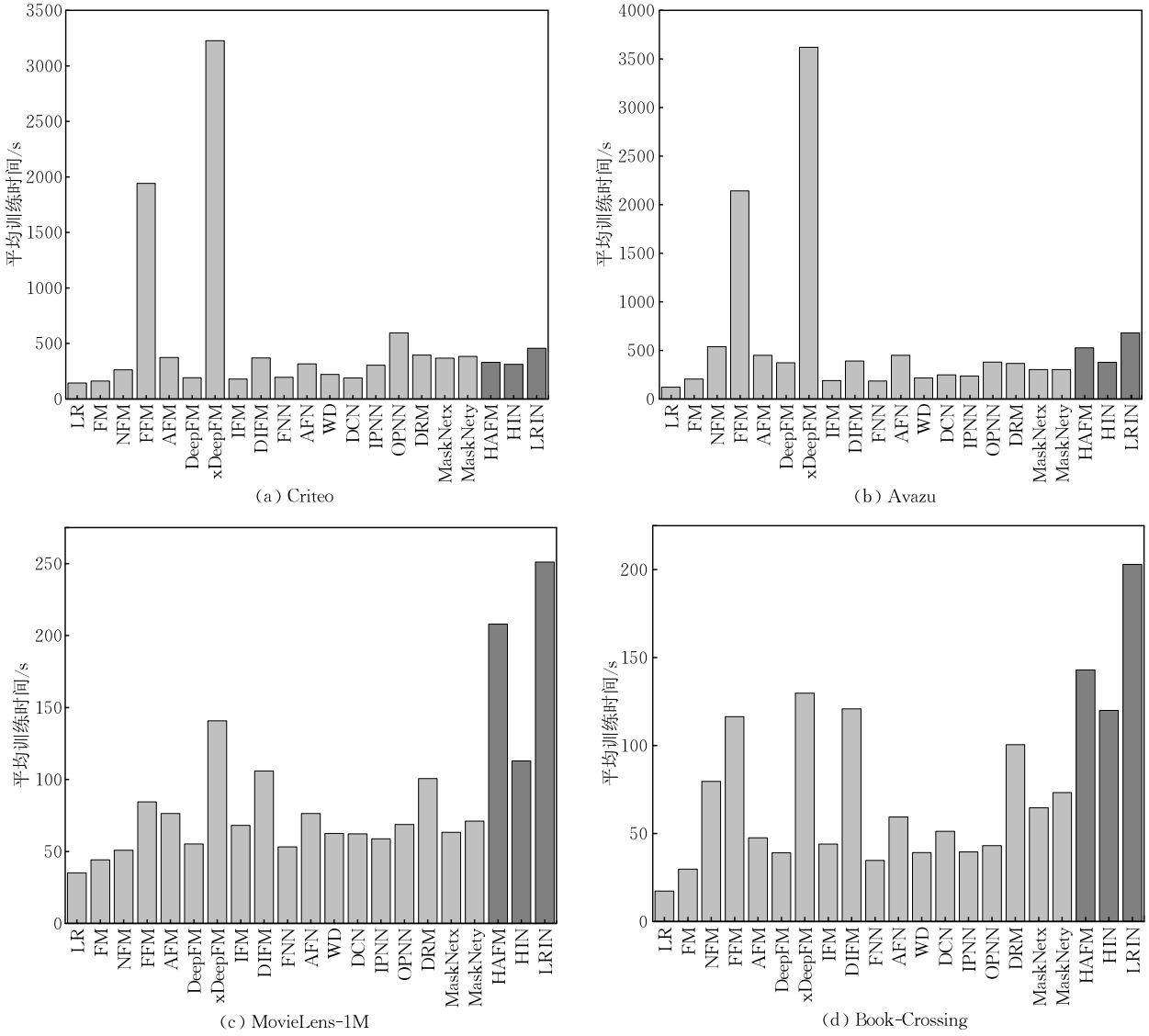

图 5　4 个数据集上的时间效率比较

在图 5 中,展示了在 4 个数据集上不同方法的运行时间(以迭代周期为单位).每一个柱形的横坐标表示模型名称,纵坐标表示模型一个迭代周期的平均训练时间,颜色加深的柱体代表最先进的模型.

实验结果分析:

(1)从整体看,所有模型在 Criteo 和 Avazu 数据集上的运行时间明显高于 MovieLens-1M 和 Book-Crossing 数据集.这是由于前两个数据集的特征数目和样本数量都较多,而后两个相对较少.因此,样本特征数目和样本数量是影响时间效率的重要因素.

(2)模型越简单,效率越高.例如,LR 的时间效率是最好,它是最简单的,但是它的性能是最差的.在 Criteo 和 Avazu 数据集上,xDeepFM 的时间效率最差,FFM 次之.因为这两个模型有较多的模型参数和计算操作:在 xDeepFM 模型中,CIN 网络的每一层计算都需要和输入层的所有向量做哈达玛(Hadamard)积;FFM 模型为每一个特征在不同的特征域提供了独立的表征向量.在 MovieLens-1M

和 Book-Crossing 数据集上，LRIN 模型的时间效率最低，HAFM 模型次之. 这是因为对于特征数目和样本较少的数据集，特征嵌入的影响显得非常大. 这导致了模型嵌入的计算开销占据大部分的运行时间. 因此，模型的时间效率被降低. HIN 是 LRIN 模型的简化，正是这种原因，其时间效率高于 HAFM 模型.

(3) 与先进的模型比较，由于实验所采用的神经网络规模基本一致，所以每一个 epoch 的运行时间差别不大. 特别是对于工业大型数据集，LRIN 模型的运行效率都远远好于 xDeepFM 和 FFM. LRIN 模型与 HAFM 模型的时间效率相当，但是它的性能好于 HAFM 模型.

4.4 消融研究

4.4.1 组成部分的消融实验

为了探索 LRIN 模型的各个组成部分对性能的影响，对 LRIN 模型进行消融研究. 即去除关键部分中的一个，观察其性能的变化. 表 3 展示了 LRIN 模型的消融实验. 表 3 中的"-"表示移除，例如"LRIN-A"表示 LRIN 模型移除了逐层注意力网络，"LRIN-RI"表示 LRIN 模型移除了逐层残差交互网络，"LRIN-MLP"表示 LRIN 模型移除了 MLP. 需要注意的是，移除了逐层注意力网络的 LRIN 模型，可以像 MMoE (Multi-gate Mixture of Experts)[35] 一样，被视为多专家混合模型，专家的数量与原初的 LRIN 模型中逐层注意力网络层的数量相同.

表 3 LRIN 模型的消融实验

Datasets	LRIN-A	LRIN-RI	LRIN-MLP	LRIN
Criteo	0.81118	0.81294	0.81274	0.81382
Avazu	0.78933	0.79075	0.78935	**0.79117**
MovieLens-1M	0.81293	0.81275	0.80586	**0.81473**
Book-Crossing	0.78539	0.78718	0.78326	**0.78826**

注：评估指标：*AUC* (LRIN-A 是没有逐层注意力网络的 LRIN 模型，LRIN-RI 是没有逐层残差交互网络的 LRIN 模型，LRIN-MLP 是没有 MLP 的 LRIN 模型). 黑色数字是最优结果.

实验结果分析：

(1) 在移除逐层注意力网络后，LRIN 模型的性能大幅度下降. 这表明了逐层注意力网络是 LRIN 模型的必要组成部分和有效的改进. 逐层注意力网络捕获高阶非线性特征交互的能力上优于 MLP. 为了获得更好的性能，LRIN 模型不但包括了逐层注意力网络，而且也包括了 MLP.

(2) 在移除逐层残差交互网络后，LRIN 模型的性能有一定下降，这表明逐层残差交互网络是 LRIN 模型的有改进效果的组成部分. 对于不同的数据集，对应的性能下降是不一样的. 例如，在 MovieLens-1M 和 Book-Crossing 数据集上，由于特征数目较少，这导致了模型性能的大幅度下降. 相反，在 Criteo 和 Avazu 数据集上，性能的下降小一些.

(3) 在移除 MLP 后，LRIN 模型的性能有显著下降，这表明 MLP 是 LRIN 模型的必要组成部分. 很多模型都将 MLP 作为组成部分，因为 MLP 的高阶、非线性交互能力较强. 实验结果表明，在 LRIN 模型中，移除 MLP 会对模型的性能产生不同的影响：对于小型的数据集，如 MovieLens-1M 和 Book-Crossing 数据集，删除 MLP 将导致模型的性能显著下降. 这是因为这些数据集的特征字段和特征数目较少，它们需要通过 MLP 产生更多的高阶、非线性交互特征. 相反，对于具有大量特征和大量特征字段的 Criteo 和 Avazu 数据集，删除 MLP 仅对模型的性能影响不大.

4.4.2 参数影响实验

由于 LRIN 模型主要包括逐层残差交互网络和逐层注意力网络，本文研究逐层残差交互网络和逐层注意力网络的层数(L_res, L_att)对 LRIN 性能的影响. 实验分别在数据集 Criteo、Avazu、MovieLens-1M 和 Book-Crossing 上进行，如图 6 所示.

在图 6 中，图 6(a)是在数据集 Criteo 上的实验，图 6(b)是在数据集 Avazu 上的实验，图 6(c)是数据集 MovieLens-1M 上的实验，图 6(d)是在数据集 Book-Crossing 上的实验. 其中，*AUC* 是性能指标，L_res 是逐层残差交互网络的层数，L_att 是逐层注意力网络的层数.

实验结果分析：

(1) 在 Criteo 数据集中(图 6(a))，当保持逐层残差交互网络层数不变的情况下，LRIN 模型的性能会随着逐层注意力网络的层数增加先提升而后再缓慢降低. 类似地，当保持逐层注意力网络层数不变的情况下，LRIN 模型的性能会随着逐层残差交互网络的层数增加先提升而后再缓慢降低. 当采用 4 层逐层注意力网络和 4 层逐层残差交互网络，LRIN 模型可以达到最优性能.

(2) 在 Avazu 数据集中(图 6(b))，当保持逐层注意力网络层数不变的情况下，随着逐层残差交互网络层数的增加，LRIN 模型保持先提升后降低的趋势. 但是，当固定逐层残差交互网络层数时，情况却不尽相同. 具体而言，当逐层残差交互网络的层数为 1 时，随着逐层注意力网络层数的增加，LRIN 模型的性能呈现缓慢下降而后上升的趋势. 当逐层

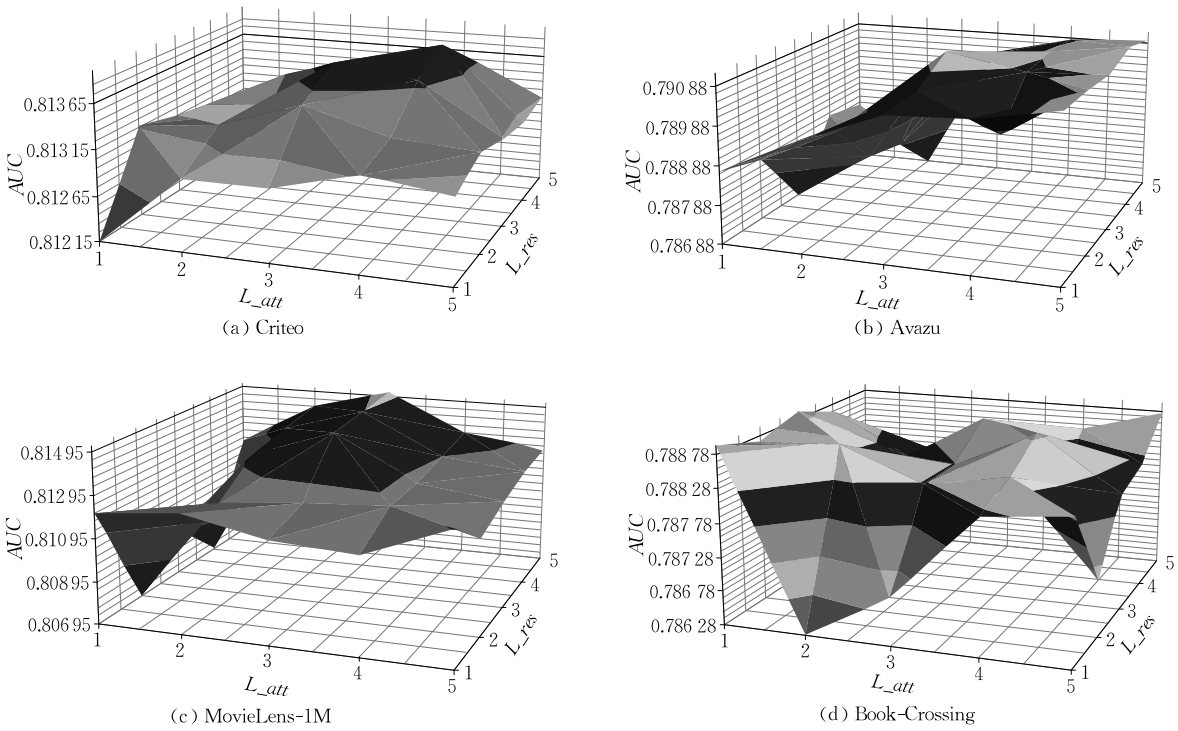

图 6　参数影响实验

残差交互网络的层数为 2 时,随着逐层注意力网络层数的增加,模型的性能一直呈下降趋势. 当逐层残差交互网络的层数为 3 时,随着逐层注意力网络层数的增加,LRIN 模型的性能呈现先上升而后缓慢下降的趋势. 当逐层残差交互网络的层数大于 3 时,随着逐层注意力网络层数的增加,LRIN 模型的性能是一个先缓慢下降后显著上升的趋势. 当采用 4 层逐层注意力网络和 5 层逐层残差交互网络,LRIN 模型可以达到最优性能.

(3) 在 MovieLens-1M 数据集中(图 6(c)),当保持逐层注意力网络层数不变的情况下,随着逐层残差交互网络层数的增加,模型大致保持先提升后降低的趋势. 当固定逐层残差交互网络层数时,随着逐层注意力网络的层数增加,LRIN 模型的表现为:当逐层残差交互网络的层数为 1 时,随着逐层注意力网络层数的增加,模型的性能先下降而后逐渐缓慢上升,并且模型整体性能较差;当逐层残差交互网络的层数为 2 和 5 时,LRIN 模型的性能随着逐层注意力网络层数的增加先下降后上升;当逐层残差交互网络的层数为 3 和 4 时,LRIN 模型的性能随着逐层注意力网络层数的增加而上升. 当采用 5 层逐层注意力网络和 3 层逐层残差交互网络,LRIN 模型的性能达到了最优.

(4) 在 Book-Crossing 数据集中(图 6(d)),在改变逐层注意力网络和逐层残差交互网络的层数时,性能变化不明显. 当逐层残差交互网络层数固定不变时,随着逐层注意力网络的层数增加,LRIN 模型的性能大致呈现先降低后上升的趋势. 当固定逐层注意力网络层数时,随着逐层残差交互网络层数的增加,LRIN 模型的性能变化规律不同. 具体而言,当逐层注意力网络的层数为 1 和 3 时,LRIN 模型随着逐层残差交互网络的层数增加,其性能变化呈现先降低而后上升最后再降低的双峰趋势. 当逐层注意力网络层数为 2 时,LRIN 模型随着残存交互网络层数的增加,其性能变化呈现上升和下降的波动趋势. 当逐层注意力网络层数为 4 时,LRIN 模型随着逐层残差交互网络层数的增加,其性能呈现下降和上升的波动趋势. 当逐层注意力网络层数为 5 时,LRIN 模型随着逐层残差交互网络层数的增加,其性能呈现先上升后下降再上升的趋势. 当采用 3 层逐层注意力网络和 1 层逐层残差交互网络,LRIN 模型可以达到最优性能.

5　结　论

本文首次提出了逐层残差交互网络,它的每一层都由原始特征指导,其中低阶交互信息逐层向高阶交互过程传递,并且每一层交互的结果都被作为

逐层注意力网络的权重.它利用了双网络协同训练机制,把逐层残差交互网络和逐层注意力网络有机地结合起来了.在四个主流的 CTR 预估数据集上进行了大量实验,实验结果表明本文提出的 LRIN 比当前先进的方法具有更高的性能.本文的局限性主要在于模型的复杂度相对于传统方法较大.未来继续探索以下工作:

(1)探索模型复杂度与性能的关系,以提供灵活的组合框架,让使用者根据自身需求定制不同性能和不同复杂度的模型.

(2)将新兴的时间序列神经网络模型纳入到结构中.随着深度学习技术的发展,新兴的神经网络结构不断被提出.因此,需要探索新的网络模型对 CTR 预估的作用,这必将激发出新的 CTR 预估方法.

参 考 文 献

[1] Wang Y Q, Liu X Y, Zheng Z Z, Zhang Z L, et al. On designing a two-stage auction for online advertising//Proceedings of the ACM Web Conference 2022 (WWW 2022). Lyon, France, 2022: 90-99

[2] Guan F, Qian C, He F Y. A knowledge distillation-based deep interaction compressed network for CTR prediction. Knowledge-Based Systems, 2023, 275: 1-9

[3] Zhang W, Du T, Wang J. Deep learning over multi-field categorical data—A case study on user response prediction//Proceedings of the 38th European Conference on Information Retrieval. Padua, Italy, 2016: 45-57

[4] Qu Y, Cai H, Ren K, et al. Product-based neural networks for user response prediction//Proceedings of the IEEE 16th International Conference on Data Mining. Catalonia, Spain, 2016: 1149-1154

[5] Luo L, Chen Y F, Liu X H, et al. Feature aware and bilinear feature equal interaction network for click-through rate prediction//Proceedings of the 27th International Conference on Neural Information Processing. Bangkok, Thailand, 2020: 432-443

[6] Wang Z, She Q, Zhang J. MaskNet: Introducing feature-wise multiplication to CTR ranking models by instance-guided mask. arXiv preprint arXiv:2102.07619, 2021

[7] Xue N M, Liu B, Guo H F, et al. AutoHash: Learning higher-order feature interactions for deep CTR prediction. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(6): 2653-2666

[8] Long L J, Yin Y F, Huang F L. Hierarchical attention factorization machine for CTR prediction//Proceedings of the 27th International Conference on Database Systems for Advanced Applications. 2022: 343-358

[9] Zhao X Y, Xia L, Tang J L, et al. Deep reinforcement learning for search, recommendation, and online advertising: A survey. arXiv preprint arXiv:1812.07127, 2018

[10] Yin N, Li H Y, Su H C. CLR: Coupled logistic regression model for CTR prediction//Proceedings of the ACM Turing 50th Celebration Conference—China. Shanghai, China, 2017: 1-9

[11] Wang X, Yang P, Chen S P, et al. Efficient learning to learn a robust CTR model for web-scale online sponsored search advertising//Proceedings of the 30th ACM International Conference on Information and Knowledge Management. 2021: 4203-4213

[12] Nanda M B, Mishra B S P, Anand V. Effects of binning on logistic regression-based predicted CTR models//Proceedings of the Biologically Inspired Techniques in Many Criteria Decision Making. Balasore, India, 2022: 483-493

[13] Rendle S. Factorization machines//Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia, 2010: 995-1000

[14] He X, Chua T S. Neural factorization machines for sparse predictive analytics//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan, 2017: 355-364

[15] Juan Y, Zhuang Y, Chin W S, et al. Field-aware factorization machines for CTR prediction//Proceedings of the 10th ACM Conference on Recommender Systems. Boston, USA, 2016: 43-50

[16] Xiao J, Ye H, He X, et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 3119-3125

[17] Guo H, Tang R, Ye Y, et al. DeepFM: A factorization-machine based neural network for CTR prediction//Proceedings of the International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 1725-1731

[18] Lian J, Zhou X, Zhang F, et al. xDeepFM: Combining explicit and implicit feature interactions for recommender systems//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 1754-1763

[19] Yu Y, Wang Z, Yuan B. An input-aware factorization machine for sparse prediction//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 1466-1472

[20] Lu W, Yu Y, Chang Y, et al. A dual input-aware factorization machine for CTR prediction//Proceedings of the International Joint Conference on Artificial Intelligence. Yokohama, Japan, 2020: 3139-3145

[21] Zhou J D, Li X, Wang X, et al. Locally weighted factorization machine with fuzzy partition for elderly readmission prediction. Knowledge-Based Systems, 2022, 242: 1-11

[22] Cheng W, Shen Y, Huang L. Adaptive factorization network: Learning adaptive-order feature interactions//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3609-3616

[23] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston, USA, 2016: 7-10

[24] Wang R, Fu B, Fu G, et al. Deep & cross network for ad click predictions//Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 1-7

[25] Zhao Z, Fang Z, Li Y, et al. Dimension relation modeling for click-through rate prediction//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2333-2336

[26] Chen L, Shi H Y. DexDeepFM: Ensemble diversity enhanced extreme deep factorization machine model. ACM Transactions on Knowledge Discovery from Data, 2022, 16(5): 1-17

[27] Jiang J H, Zheng Y F, Shi Z K, et al. A practical system for privacy-aware targeted mobile advertising services. IEEE Transactions on Services Computing, 2020, 13(3): 410-424

[28] Liu D X, Ni J B, Lin X D, et al. Towards private and efficient Ad impression aggregation in mobile advertising//Proceedings of the IEEE International Conference on Communications. Shanghai, China, 2019: 1-6

[29] Wang F Y, Gu H S, Li D S, et al. MCRF: Enhancing CTR prediction models via multi-channel feature refinement framework//Proceedings of the 27th International Conference on Database Systems for Advanced Applications. 2022: 359-374

[30] Tofigh S, Ahmad M Q, Swamy M N S. A low-complexity modified ThiNet algorithm for pruning convolutional neural networks. IEEE Signal Processing Letters, 2022, 29: 1012-1016

[31] Long L J, Huang F L, Yin Y F, et al. Multi-task learning for collaborative filtering. International Journal of Machine Learning and Cybernetics, 2022, 13(5): 1355-1368

[32] Long L J, Yin Y F, Huang F L. Graph-aware collaborative filtering for top-*n* recommendation//Proceedings of the International Joint Conference on Neural Networks. Shenzhen, China, 2021: 1-8

[33] Sun J Q, Yin Y F, Huang F L, et al. Self-residual embedding for click-through rate prediction//Proceedings of the 5th International Joint Conference on Web and Big Data. Guangzhou, China, 2021: 323-337

[34] Yin Y F, Cheng H, Huang F L. ECG pattern discovery algorithm based on local repeatability//Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine. Seoul, Republic of Korea, 2020: 1206-1209

[35] Ma J, Zhao Z, Yi X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 1930-1939

[36] Bian N, Wang L. Based on feature reconstruction and feature intersection, the Deep CTR Prediction Model//Proceedings of the 2nd International Conference on Big Data, Information and Computer Network. Xishuangbanna, China, 2023: 93-96

[37] Zhai P, Yang Y, Zhang C. Causality-based CTR prediction using graph neural networks. Information Processing and Management, 2023, 60(1): 1-19

YIN Yun-Fei, Ph. D., associate professor. His main research interests are data analysis and artificial intelligence.

LONG Lian-Jie, M. S. candidate. His main research interest is recommendation system.

HUANG Fa-Liang, Ph. D., professor. His main research interests include data mining and social media processing.

WU Kai-Gui, Ph. D., professor. His main research interests are blockchain and data analysis.

Background

Advertising Click-Through Rate (CTR) estimation is of great significance to the accurate deployment of online advertising. CTR estimation belongs to the research field of recommendation system, which is currently one of the most popular artificial intelligence technologies. While current state-of-the-art methods employ high-order feature interaction to discover similar features, and then, make recommendations based on similar features. The traditional feature interaction method utilizes the multiplication of feature vectors, so that the similar features are relatively strengthened in the process of vector multiplication, and the dissimilar features are relatively weakened in the process of multiplication. Existing methods

leverage deep neural networks to iteratively complete the multiplication process of feature vectors. Therefore, the essence of deep neural networks is feature interaction, and it is a high-order feature interaction. In recent years, high-order feature interaction methods have been widely studied, and the design of interaction operators is a research hotspot. Early work attempted to improve the performance of higher-order feature interactions using the inner product, the Hadama product, the outer product, the Kronecker product, and so force. Recently, neural network methods combining high-order feature interaction and low-order feature interaction have emerged. Whether it is an interaction operator or a neural network based on high-order interaction, its essence is to make similar features enhanced and not similar features weakened. The advantage of high-order feature interaction is that it can quickly distinguish features, but it is easy to lose the original information. The advantage of low-order feature interaction is that it can better retain the information of the original feature, while its classification ability is relatively weak. MaskNetx (Wang et al., 2021) and AutoHash (Xue et al., 2022) are state-of-the-art methods that combine high-order feature interactions with low-order feature interactions. On this basis, HAFM (Long et al., 2022) introduces the feature interaction method of hierarchical factorization machine, which greatly improves the accuracy of CTR prediction.

Despite these successes, researchers have not considered primitive feature-guided higher-order feature interactions, which could potentially provide another means of higher-order feature interactions. This motivates us to explore the layer-by-layer residual interaction mechanism and attention mechanism, which could cooperate with each other to form a new higher-order feature interaction method, viz., LRIN. In LRIN, in order to alleviate the semantic loss problem of high-order interaction, we interact with the original features again with the results of each high-order interaction, and let the original features guide the higher-order interaction, thereby forming a residual network of layer-by-layer interaction. Experimental results show that our improvement achieves very good performance.

This research was supported by the National Natural Science Foundation of China (No. 61962038), the Central University Basic Research Funding (No. 2022CDJKYJH023), and the Open Research Fund of Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision (No. GXHIID2208).