基于改进 FM 算法和注意力机制的深度点击率预估模型

李兴兵[1],谢 珺[1],续欣莹[2],李小飞[1],赵旭栋[1]

(1.太原理工大学 信息与计算机学院,山西 晋中 030600;2.太原理工大学 电气与动力工程学院,山西 太原 030024)

摘 要:针对目前的广告点击率预估模型未能充分学习低阶特征且忽略了不同高阶特征对模型准确率的影响不同的问题,提出了一种基于注意力机制和深度学习的点击率预估模型。该模型采用改进因子分解机(Factorization machine,FM)算法,将全息简化表示(Holographic reduced representation,HRR)的压缩外积用于 FM 中,从而更好地学习低阶特征,帮助模型获得更好地表示。采用深度神经网络(Deep neural network,DNN)对高阶特征建模学习。引入注意力神经网络区分不同高阶特征交互的重要性来更好地学习高阶特征,从而得到一种能够同时有效学习到低阶特征和高阶特的点击率(Click-through rate,CTR)模型——基于改进 FM 算法和注意力机制的深度点击率预估模型(Deep click rate prediction model based on attention mechanism and improved FM algorithm,DAHFM)以提升模型的预估性能。在 Criteo 和 MovieLens-1M 数据集上大量的实验表明,DAHFM 模型相比逻辑回归(Logistic regression,LR)、FM 和 DeepFM 等模型不仅有效学习了特征信息,而且一定程度上提升了模型的性能和点击率的预估效果。

关键词:点击率预估;因子分解机;注意力机制;深度神经网络;组合特征

中图分类号:TP391 **文章编号**:1005-9830(2021)04-0429-10

DOI:10.14177/j.cnki.32-1397n.2021.45.04.006

Deep click rate prediction model based on improved FM algorithm and attention mechanism

Li Xingbing[1],Xie Jun[1],Xu Xinying[2],Li Xiaofei[1],Zhao Xudong[1]

(1.School of Information and Computer,Taiyuan University of Technology,Jinzhong 030600,China;
2.School of Electrical and Power Engineering,Taiyuan University of Technology,Taiyuan 030024,China)

Abstract:Aiming at the current advertising click-through rate prediction model that fails to fully learn low-level features and ignores the different effects of different high-level features on the accuracy of the model,a click-through rate prediction model based on attention mechanism and deep

收稿日期:2021-05-08 **修回日期**:2021-07-07

基金项目:山西省应用基础研究计划项目(201801D221190;201801D121144)

作者简介:李兴兵(1996-),男,硕士生,主要研究方向:推荐系统与点击率预估,E-mail:970442510@qq.com;通信作者:谢珺(1979-),女,博士,副教授,主要研究方向:智能信息处理、机器学习、数据挖掘、文本情感分析与推荐、医学图像处理,E-mail:xiejun@tyut.edu.cn。

引文格式:李兴兵,谢珺,续欣莹,等. 基于改进 FM 算法和注意力机制的深度点击率预估模型[J]. 南京理工大学学报,2021,45(4):429-438.

投稿网址:http://zrxuebao.njust.edu.cn

learning is proposed. First, the model adopts an improved factorization machine(FM) algorithm, and uses the compressed outer product of the holographic simplified representation(HRR) in FM, so as to better learn low-level features and help the model obtain a better representation. Secondly, deep neural network(DNN) is used to model and learn high-level features. Finally, the attention neural network is introduced to distinguish the importance of different high-level feature interactions to better learn high-level features, so as to obtain a click-through rate(CTR) model based on attention mechanism and improved FM algorithm(DAHFM) that can effectively learn low-level features and high-level features at the same time to improve the estimated performance of the model. A large number of experiments on the Criteo and MovieLens-1M data sets show that the DAHFM model not only effectively learns the feature information, but also improves the performance of the model and the prediction effect of click-through rate to a certain extent compared with such models as logistic regression(LR) , FM and DeepFM.

Key words: click-through rate estimation; factorization machine; attention mechanism; deep neural network; combined features

自2020年以来,在线广告的收入不断增长,已经发展成为一项千亿美元的业务。点击率预估(Click-through rate, CTR)在广告行业至关重要,主要目标是在适当的环境下向适当的用户提供适当的广告。广告的精准投放,依赖于预估目标受众对相应广告的点击率预估[1]。因此,广告点击率预估的效果和准确率成为人们关注的焦点。点击率预测是预测用户点击推荐项目的概率。它在个性化广告和推荐系统中起着重要的作用。目前已经有很多模型被提出来解决这个问题,如逻辑回归(Logistic regression, LR)[2],Poly2[3]模型,因子分解机模型(Factorization machine, FM)[4],梯度提升树(Gradient boosting decision tree, GBDT) + LR[5]模型。近年来,利用神经网络进行点击率估计也是这个领域的一个研究趋势,并引入了一些基于深度学习的模型,如Wide&Deep[6]模型,基于因子分解机的神经网络(Factorization machine supported neural network, FNN)[7]模型和DeepFM[8]等模型。特征学习对于CTR任务至关重要,对于排序模型来说,有效地捕捉这些复杂的特征非常重要。但是这些浅层模型,例如LR和FM等模型只对低阶特征相互作用建模有效,对捕捉高阶特征相互作用没有什么效果。基于深度神经网络的深度模型则利用多层非线性神经网络捕获高阶特征相互作用,而无法有效学习到低阶特征交互。因此,本文提出一种既能利用改进FM算法来有效学习到低阶特征又能利用深度神经网络(Deep neural network, DNN)[9]和注意力机制[10]学习到高阶特征交互的点击率预估模型——基于改进FM算法和注意力机制的深度点击率预估模型(Deep click rate prediction model based on attention mechanism and improved FM algorithm, DAHFM)。不仅增强了模型的可解释性,而且提高了点击率预估的准确率。本文的主要工作包括以下4个方面。

(1) 提出了一种基于改进FM算法和注意力机制的深度点击率预估模型,通过大量实验判断其是否有效提升了点击率预估效果;

(2) 全息简化表示(Holographic reduced representation, HRR)[11]的压缩外积被应用到FM中来改进FM算法从而更好地学习低阶特征,利用快速傅里叶变换降低算法的事件复杂度,使模型训练效果更快更好;

(3) 利用DNN学习高阶特征,并构建合适的注意力网络来区分不同高阶特征的重要性,从而更好地利用高阶特征提高模型的准确率;

(4) 应用不同数据集的实验结果表明本文提出的新模型DAHFM可以有效学习低阶和高阶组合特征,在一定程度上提高了点击率预估的效果。

1 相关工作

早期人们解决点击率预估问题的方法是构建LR模型[2],没有考虑特征间的相互关系,而且需要人工特征,但算法简单也易于调参。Poly2[3]是考虑了二阶特征的模型,它可以在一定程度上解

决特征交叉组合的问题，但是此模型需要非常稀疏的特征向量作为输入，导致模型训练难度大，而且不容易收敛。2010 年 Rendle[4] 提出 FM，考虑了二阶特征的组合，模型的性能要优于线性模型，而且它的复杂性是线性的。其后，Facebook 提出 GBDT 模型，由其中的集成决策树自动学习有效特征组合信息，然后联合 LR 做出预测[5]，取得了不错的效果。近几年，随着深度神经网络的不断发展，其在文字、语音、图像等众多领域取得成功，人们提出了一些基于深度学习的模型来进行 CTR 预测。2016 年谷歌提出的 Wide 和 Deep[6] 结合了 LR 和多层感知器(Multi-layer perceptron，MLP)，其中 Wide 这部分采用了有比较强的记忆模型 LR，而 Deep 端则使用了有一定泛化能力的 MLP。FNN[7] 同时结合了预训练的 FM 和 MLP。2017 年 Guo 等[8] 提出的 DeepFM 可以同时学习到低阶和高阶特征信息，且 FM 部分和 Deep 部分共享输入和嵌入层，也加快了训练，提高了模型的效率。Wang 等[12] 提出的深度交叉网络(Deep & cross network，DCN) 使用交叉网络而不需要人工就能自动提取显示组合特征，且网络结构简单，节省内存。Xiao 等[13] 提出的注意因子分解机(Attentional factorization machine，AFM) 利用注意力机制对 FM 的二阶交叉特征进行加权有效地学习了组合特征。2018 年 Lian 等[14] 提出的 xDeepFM 通过一种新颖的压缩模型来模块化功能交互互动网络(Compressed interaction network，CIN) 部分。2019 年 Song 等[15] 进一步提出 AUTOINT模型，采用多头自注意力机制将高维稀疏特征映射到低维空间，然后自动构建特征交互。Liu 等[16] 提出的基于卷积神经网络的特征生成(Feature generation by convolutional neural network，FGCNN) 使用卷积神经网络(Convolutional neural network，CNN) 生成特征并将最新的深度分类器应用于扩展特征空间。Huang 等[17] 提出的 FiBiNet 通过双线性有效地学习特征相互作用功能。2020 年 Lu 等[18] 提出了一种双输入感知因子分解机(Dual input-aware factorization machine，DIFMs)，它可以按位和向量两个层次上同时自适应地重新加权原始特征表示。Tao 等[19] 提出基于高阶稀疏特征交互的模型(High-order attentive factorization machine，HoAFM)，将高阶特征交互注入到特征表示学习中，以建立具有表达性和信息性的交叉特征。杨妍婷等[20] 提出一种基于增强型因子分解向量输入神经网络的广告点击率预测模型，在基于因子分解向量输入神经网络的基础上增加了新特征生成层，采用一种针对 CTR 数据的卷积操作，对数据进行通道变换后引入 Inception 结构进行卷积，将生成的新特征和原始特征结合，提升了深度网络的学习能力。2021 年 Deng 等[21] 提出 DeepLight 模型通过在浅层组件中显式搜索信息特征交互，修剪 DNN 冗余参数和密集的嵌入向量，加速模型训练。Zhu 等[22] 提出了一种新的解耦自注意力神经网络(Disentangled self-attentive neural network，DSAN) 模型，通过解耦技术来促进学习特征的相互作用。本文采用改进的 FM 算法全息因子分解机(Holographic factorization machine，HFM)[23]，将 HRR 用于 FM 中，因为 HRR[11] 可以表示压缩的外积，可以帮助模型获得更好地表示，从而更好地学习到低阶特征得到更好的实验效果。同时结合 DNN 学习高阶特征，利用注意力机制[10] 处理来自不同层次的分层特征，然后自动选择出占优势的特征从而提升模型预测的准确率。

2　基于改进 FM 算法和注意力机制的深度点击率预估模型

本文所提出的 DAHFM 模型结构如图 1 所示。

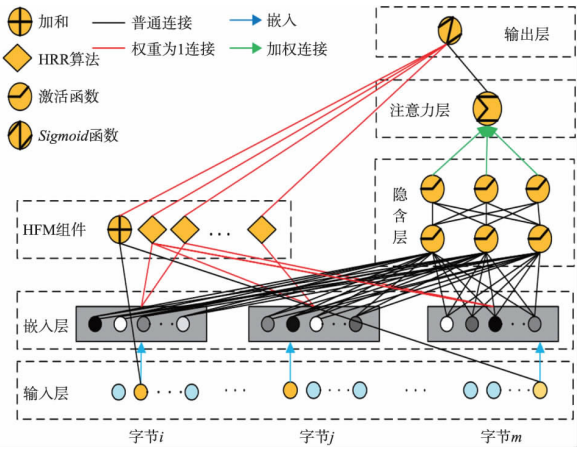

图 1　DAHFM 模型结构

图 1 中包括 6 大部分，分别为：(1) 输入层，稀疏特征由输入层输入；(2) 嵌入层，将高维稀疏特征转换成低维稠密向量输出；(3) HFM 组件，利用 HRR[11] 算法改进 FM，不仅可以建模一阶特征，还可以高效地获取二阶特征表示；(4) 注意力层，加权平均来判断不同特征之间交互的重要性，突出优势特征；(5) 隐含层，对输入特征多层次的

抽象,使不同类型的数据得到更好的线性划分,从而更好地学习高阶特征;(6) 输出层,将模型通过一系列操作得到的低阶和高阶特征信息通过 *Sigmoid* 函数做归一化输出。

2.1 输入层

点击率预估输入的稀疏特征首先需要经过独热编码[24]使其转化为独热向量,首先将用户的个人资料和商品的属性表示为一个稀疏向量。输入数据有两种值,第一种是数值,例如,[年龄=54]。第二种是分组分类值,例如,[职业=学生]或[专业=计算机]。分类值可以直接转换成二进制向量表示。通常情况下模型的输入是高维稀疏的,将 $\boldsymbol{x}\in\mathbf{R}^D$ 表示为稀疏特征向量,其中 D 表示特征空间的维数。对于 CTR 预测,这里假设 $y\in\{0,1\}$ 表示用户是否点击了给定的项目。如果 $y=1$ 则表明用户点击了该广告,$y=0$ 表明用户没有点击。

2.2 嵌入层

嵌入是指将高维稀疏特征转换成低维密集向量的常用技术[25]。嵌入层的输出可以看作 $K\times F$ 矩阵,其中每一列都是一个嵌入向量,K 是嵌入层的维数,F 是特征字节的个数。通过 $\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_F]$ 表示嵌入。其中:$\boldsymbol{e}_i\in\mathbf{R}^K$ 表示嵌入向量,K 是一个超参数,表示嵌入的维数。

2.3 HFM 组件

FM 是因子分解机,它的数学表达式为

$$F(x)=w_0+\sum_{i=1}^{n}w_i\cdot x_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i,v_j\rangle x_i x_j \quad (1)$$

式中:$x\in\mathbf{R}^N$;$\langle\cdot,\cdot\rangle$ 表示两个向量的点积操作;$\{v_1,v_2,\cdots,v_n\}$ 表示分解参数;w_0 是全局的偏差;$\sum_{i=1}^{n}w_i\cdot x_i$ 是模型的线性部分;$\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i,v_j\rangle x_i x_j$ 是因子分解机算法的核心部分。FM 将所有特征向量的内积进行简单的相加,这种对位相乘的特征交叉方式,会丧失比较多的信息,最终使模型的信息浪费,造成点击率预估效果不佳。1995 年 Plate 等提出了全息简化表示,简称为 HRR。其最主要的想法是一系列用于模拟全息存储和检索的编码和解码操作,其中主要的两个算子表达式如下

$$[a\otimes b]_k=\sum_{i=1}^{d-1}a_i b_{(k-i)\bmod d} \quad (2)$$

$$[a\bigstar b]_k=\sum_{i=1}^{d-1}a_i b_{(k+i)\bmod d} \quad (3)$$

式中:$\bigstar:R^d*\ R^d\to R^d$ 表示循环相关算子 CCOR;$\otimes:R^d*\ R^d\to R^d$ 表示循环卷子算子 CCOV。有了式(2)和(3)两个算子,假设有一个 m,如果

$$m=a\otimes b \quad (4)$$

$$a\bigstar m\approx b+n \quad (5)$$

式中:n 为噪声,式(4)和式(5)也被称作为相关性提取,CCOV 和 CCOR 可以作为编码解码对,再引入联想记忆运算符,那么

$$m=a_1\otimes b_1+a_2\otimes b_2+a_3\otimes b_3 \quad (6)$$

这样计算任一 b 都会相对容易,比如希望计算 b_1,只需要计算 $a_1\bigstar m$ 就可以得到一个带有噪声的 b_1。最简单计算外积和压缩外积需要的时间复杂度都为 $O(n^2)$,但是在这里,可以利用快速傅里叶变换(Fast Fourier transform, FFT)在 $O(n\log n)$ 的运行时间内计算得到结果

$$a\otimes b=F^{-1}(F(a)\odot F(b)) \quad (7)$$

$$a\bigstar b=F^{-1}(\overline{F(a)}\odot F(b)) \quad (8)$$

式中:F 和 F^{-1} 分别为傅里叶变换和傅里叶逆变换,$\odot$ 为哈达玛乘积,也就是矩阵对位相乘。最终选取快速傅里叶变换的实数值作为输出。通过 FFT 可以将 HRR 的计算时间复杂度压缩为 $O(n)$,提高了模型的效率。再看 CCOV 和 CCOR 的计算,发现它们就是在把外积压缩为向量表示,一方面可以节省大量内存,另一方面,相较于 FM 直接内积求和,压缩外积可以获得更多的信息,得到更强的特征表达,因而获得更佳的效果。此处将 HRR 替换 FM 中的内积的形式,得到 HFM 表达式

$$HFM(x)=L(x)+h^{\mathrm{T}}\left(\sum_{i=1}^{n}\sum_{j=i+1}^{n}(v_i\otimes v_j)x_i x_j\right) \quad (9)$$

式中:$L(x)=w_0+\sum_{i=1}^{n}w_i\cdot x_i$ 表示 FM 公式中的线性部分。在 HFM 中成对交叉特征是可以提取的,从 HFM 可以得到

$$m=\alpha_{12}(v_1\otimes v_2)+\alpha_{13}(v_1\otimes v_3)+\cdots \quad (10)$$

式中:α_{ij} 表示特征对(x_i,x_j)的交互,因此只需要计算 $v_1\bigstar m$,就可以得到 $v_2,v_3,\cdots,v_n$ 的组合信息。利用 HFM 组件对 $\mathbf{R}^K$ 的嵌入向量进行操作,得到 $2K$ 个 v,最后使用式(9)计算得到结果并输入到下一层。

2.4 隐含层

这部分的输入是上文定义的嵌入 E。实际上,这里连接嵌入向量

$$\boldsymbol{H}_0=contact(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_F) \quad (11)$$

这里采用了一个前馈神经网络，所有隐藏层的大小相同。设 $\boldsymbol{H}_k$ 表示隐藏层，其中 $k=1,2,\cdots,L$。$\boldsymbol{H}_k \in \mathbf{R}^d$，其中 d 是超参数。隐含层神经网络的结构如下

$$\boldsymbol{H}_1 = Relu(W^{(0)}\boldsymbol{H}_0 + b^{(0)}) \tag{12}$$

$$\boldsymbol{H}_{k+1} = Relu(W^{(k)}\boldsymbol{H}_k + b^{(k)}) \tag{13}$$

隐含层最终的输出是把不同层的输出连接起来作为一个总的输出。隐藏单元是高阶特征交叉，它比低阶相互作用包含更全面的信息。随着层越来越深，隐藏的单元会呈现更高阶的特征。那么总的输出表达式为

$$\boldsymbol{Out}_L = [\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_L] \tag{14}$$

式中：L 是网络的深度。然后将隐含层的输出送入注意力层。

2.5　注意力层

注意力机制的核心思想在于：让不同输入特征对结果的贡献程度不同，主要突出更加重要的特征。本文采用一种使用注意力机制的解决方案处理隐含层输出的高阶组合特征，实现高阶特征整合。不同层的分层权重可以定义为

$$\alpha_k' = \langle h, Relu(W_a\boldsymbol{H}_k + b_a) \rangle \tag{15}$$

$$\alpha_k = \frac{\exp(\alpha_k')}{\sum_{k=1}^{L}\exp(\alpha_k')} \tag{16}$$

式中：$W_a \in \mathbf{R}^{d\times e}$，$d$ 表示每一层隐藏单元的数目，$b_a \in \mathbf{R}^e$，$h \in \mathbf{R}^d$，e 表示注意力网络中隐藏单元的数目。利用 *Softmax* 函数的特性来对注意力机制的得分进行归一化处理，使用 *Relu*[26] 函数作为激活函数。α 代表层与层之间的层次关系。于是注意力层的输出就可以定义为

$$\boldsymbol{Out}_A = \sum_{k=1}^{L}\alpha_k \cdot \boldsymbol{H}_k \tag{17}$$

Sigmoid 函数表达式为

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{18}$$

2.6　输出层

经过 HFM 得到的低阶特征信息和经过注意力层输出的高阶组合特征信息经过 *Sigmoid* 函数输出，最终得到组合预测模型表达式

$$p = Sigmoid(HFM(x) + \boldsymbol{Out}_A) \tag{19}$$

式中：$HFM(x)$ 代表 HFM 组件的输出，$\boldsymbol{Out}_A$ 表示注意力层的输出。

2.7　模型分析

本文所提出的模型 DAHFM 在输入层和嵌入层上与 DeepFM 类似，学习低阶特征和高阶特征两部分共享同样的输入，不仅节省了内存而且提高了模型的训练效率。由式(1)可知，FM 的算法时间复杂度为 $O(n^2)$，但本文采用的 HFM 算法时间复杂度由式(7)、式(8)和式(9)可知，通过快速傅里叶变换可以将 HRR 的计算时间复杂度压缩为 $O(n)$，相较于传统的 FM 效率更高，算法的复杂度也更加小。这样不仅节约了整个模型的计算资源，也降低了模型训练的计算代价。DNN 与注意力机制的结合不仅可以有效学习到高阶组合特征，而且对不同高阶特征重要性进行区分，更加利于高阶特征的表达，使得模型的准确率更好。对模型整体来说，DAHFM 模型不仅分别有效学习了低阶特征和高阶特征，而且也降低了时间复杂度，提升了训练效率和模型点击率预估的效果。

2.8　损失函数和模型过拟合解决方法

为了对模型的权重和参数更好地学习，本文使用对数损失函数 *LogLoss* 作为模型的目标函数，其公式表示如下

$$Logloss = -\frac{1}{N}\sum_{i=1}^{n} y\log(p) + (1-y)\log(1-p) \tag{20}$$

式中：y 表示用户真实点击(1 代表点击，0 代表非点击)，p 表示模型的预测点击率，N 表示训练样本的总数。LR、FM 和 AFM 模型使用 L2 正则化[27,28]来解决过拟合问题，它的公式比较简单易于计算，直接在原来的对数损失函数基础上加上权重参数的平方和：$Logloss + \lambda\sum_j w_j^2$，其中 λ 是正则化参数，可以调节正则化强度。其他的深度神经网络模型通过 Dropout[29] 来训练解决过拟合问题。

3　实验

本次实验的实验环境是 Windows 10 操作系统，I5 处理器，基于 Python3，Tensorflow2.0 框架。为了验证本文所提模型 DAHFM 的性能，本节在两个公共数据集上进行了大量的实验来证明所提模型的总体性能，并与现有的一些模型进行了比较。

3.1　数据集和评价指标

本文使用两个数据集，第一个是 Criteo 数据集，它包含一个在线广告服务的 7 天点击日志，有

超过 4 500 万个样本，每个样本包含 13 个整数特征和 26 个分类特征。第二个是 MovieLens-1M 数据集，MovieLens-1M 数据集含有来自 6 000 名用户对 4 000 部电影的 100 万条评分数据。其中评分得分数值从 0 到 5，在分类过程中，将评分小于 3 的样本视为负样本，因为分数低表示用户不喜欢这部电影。同时将评分大于 3 的样本视为阳性样本，并移除中性样本，即评分等于 3 的样本。对于这两个数据集，将数据随机分为训练集(70%)、验证集(20%)和测试集(10%)。本次实验采用 *LogLoss* 和 *AUC*[30] 作为评价指标，其中 *LogLoss* 的计算公式见式(20)，*AUC* 是指 ROC 曲线下面积，它是随机选择的正样本比随机选择的负样本更高分数的概率。*AUC* 的大小与模型性能的优劣呈正相关。

3.2 实验对比方法简介

在实验中，将本文所提出的模型 DAHFM 与其他 7 个不同方法的模型进行了比较，以下是这些模型的简介。

LR: LR[2] 是工业应用中最广泛使用的线性模型。它易于实现，训练速度快，但不能捕捉非线性信息。

FM: FM[4] 使用因子分解技术来模拟二阶特征相互作用。对于稀疏数据具有很好的学习能力。

AFM: AFM[13] 是捕捉二阶特征相互作用的最先进的模型之一。它通过使用注意机制来区分二阶组合特征的不同重要性，从而扩展了 FM。

Wide&Deep: Wide&Deep[6] 包括 Wide 和 Deep 部分，其中 Wide 部分模拟线性低阶特征相互作用，Deep 部分模拟非线性高阶特征相互作用。然而，大部分仍然需要特征工程。

DeepFM: DeepFM[8] 结合了 FM 模型和 DNN 模型在特征学习方面的优势。可以同时学习低阶和高阶特征交互，而且显著减少了特征工程工作量。

DCN: DCN[12] 将 Wide&Deep 中的 Wide 部分替换为由特殊网络结构实现的 Cross，自动构造有限高阶的交叉特征，并学习对应权重，不需要手动特征工程。

AutoInt: AutoInt[15] 提出了一种基于自注意力神经网络的新方法，它可以自动学习高阶特征交互，并有效地处理大规模高维稀疏数据。

3.3 实验参数设置

Dropout 大小设置为 0.5，优化器为 Adam[31]，激活函数是 *Relu*[26]，学习率为 0.001。

3.4 对比实验及结果

表 1 中 AUC_1、AUC_2 和 $LogLoss_1$、$LogLoss_2$ 分别表示模型在 Criteo 和 MovieLens-1M 数据集上的 *AUC* 和 *LogLoss*。

表 1 不同模型在两个数据集上的表现

模型	AUC_1	$LogLoss_1$	AUC_2	$LogLoss_2$
LR	0.7652	0.4737	0.7632	0.4529
FM	0.7863	0.4625	0.8123	0.4146
AFM	0.7906	0.4587	0.8113	0.4127
Wide&Deep	0.7958	0.4553	0.8221	0.3983
DeepFM	0.7987	0.4529	0.8255	0.3965
DCN	0.8001	0.4513	0.8273	0.3923
AutoInt	0.8026	0.4498	0.8286	0.3874
DAHFM	**0.8037**	**0.4479**	**0.8314**	**0.3842**

由表 1 可知，作为唯一一个不考虑特征交互的模型，LR 与其他模型相比表现最差。说明同时并适当地学习高阶和低阶特征交互作用，可以提高点击率预测模型的性能。由图 2 和图 3 可以直观地看出，与其他模型相比，DAHFM 在 Criteo 数据集和 MovieLens-1M 数据集上的表现也更好。由此可见，同时有效学习高阶和低阶特征信息，提高了点击率预测模型的性能。

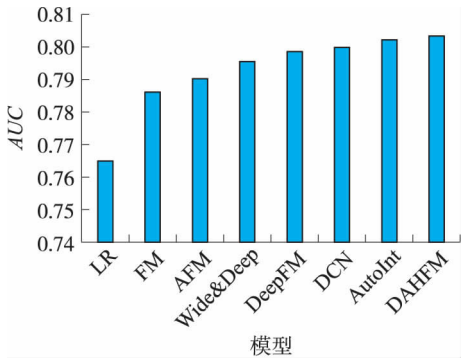

图 2 不同模型在 Criteo 上的表现

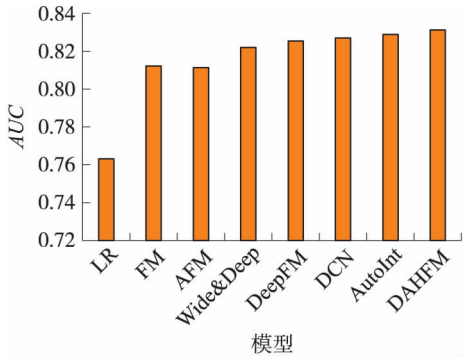

图 3 不同模型在 MovieLens-1M 上的表现

3.4　超参数

本节研究在 Criteo 数据集上 DAHFM 模型的不同超参数对点击率评估结果的影响。这里选取 *AUC* 作为评价指标。主要对以下超参数依次实验:(1) *Dropout*;(2) 隐藏层数量;(3) 注意力网络的层数。

3.4.1　*Dropout*

Dropout 指的是一个神经元在网络中保留的概率。它是一种折衷神经网络的精度和复杂性的正则化技术。这里将 *Dropout* 分别设置为 1.0、0.9、0.8、0.7、0.6、0.5。如图 4 和图 5 所示,当 *Dropout* 在 0.6 到 0.9 时,所有深度模型都能达到各自的最佳性能。实验结果表明,在模型中设置合适的 *Dropout* 可以增强模型的鲁棒性,从而提升了模型的性能,使最终的点击率预估效果更好。

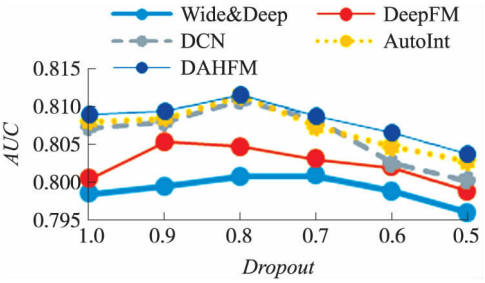

图 4　不同深度模型在不同 *Dropout* 下的 *AUC*

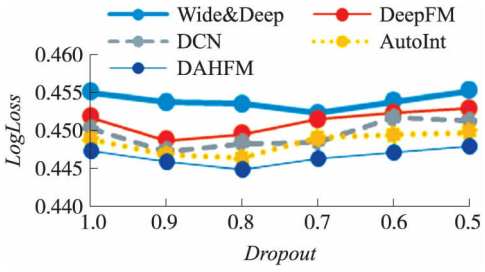

图 5　不同深度模型在不同 *Dropout* 下的 *LogLoss*

3.4.2　隐藏层数量

如图 6 和图 7 所示,增加隐藏层的数量可以在一开始提高模型的性能,但是如果隐藏层的数量不断增加,会出现过拟合现象,从而使模型的性能下降。这也说明了设置合适的隐藏层数量对模型的性能提升是有益的。

3.4.3　注意力网络的层数

表 2 中 AUC_1 和 AUC_2 分别表示模型在 Criteo 和 MovieLens-1M 数据集上的 *AUC*。

由表 2 可以看出,随着注意力网络层数的增大,模型性能可以继续提高。但注意力网络层数过大会造成过拟合使得预测的准确率降低,这也证明了适当的注意力网络层数会提高模型的性能,提高点击率预估的准确率。

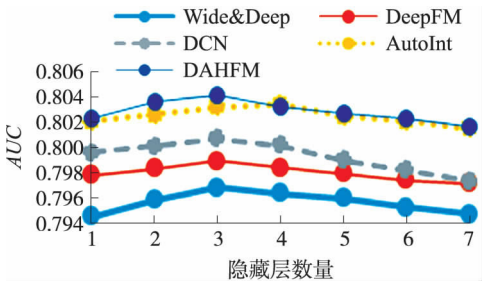

图 6　不同深度模型在不同隐含层数量下的 *AUC*

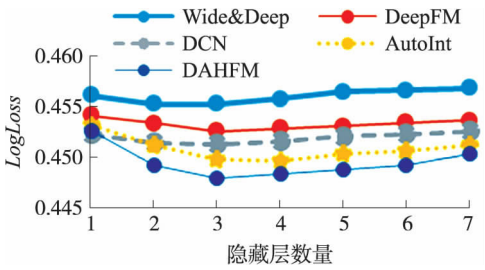

图 7　不同深度模型在不同隐含层数量下的 *Logloss*

表 2　不同注意力网络层数下 DAHFM 在两个数据集上的 *AUC*

注意力网络层数	AUC_1	AUC_2
32	0.7986	0.8286
64	0.8037	0.8314
128	0.8040	**0.8318**
256	**0.8042**	0.8312
512	0.7836	0.8087

4　模型关键组件效果分析

4.1　实验设置

本文提出的 DAHFM 模型可以划分为 3 大部分,第一部分是 HFM[23] 组件,第二部分可以理解为 DNN[9],第三部分为注意力网络(Attention net),故模型关键组件可以分为以下几类:

(1) HFM[23] 组件单独作为一个模型;

(2) DNN[9] 模块单独作为一个模型;

(3) HFM 组件+注意力网络作为一个模型,简写为 AHFM;

(4) DNN+注意力网络组成一个模型,简写为 ADNN;

(5) HFM 组件+DNN 构成一个模型,简写为 DHFM。

表 3 中 AUC_1、AUC_2 和 $LogLoss_1$、$LogLoss_2$ 分别表示模型在 Criteo 和 MovieLens-1M 数据集上的 *AUC* 和 *LogLoss*。

表 3 关键组件在两个数据集下的表现

模型	AUC_1	$LogLoss_1$	AUC_2	$LogLoss_2$
HFM[23]	0.7871	0.4618	0.8128	0.4152
AHFM	0.7914	0.4593	0.8119	0.4165
DNN[9]	0.7972	0.4587	0.8237	0.3981
ADNN	0.7984	0.4575	0.8246	0.3974
DHFM	0.7995	0.4523	0.8267	0.3957
DAHFM	**0.8037**	**0.4479**	**0.8314**	**0.3842**

4.2 实验结果分析

由表 1 和表 3 可知 HFM 的性能比 FM 的性能好,其中在 Criteo 数据集和 MovieLens-1M 数据集上 HFM 的 *AUC* 分别比 FM 提高了 0.8% 和 0.5%,而且 *LogLoss* 分别降低了 0.7% 和0.6%; 这也验证了 HFM 算法是优于 FM 的。从图 8和图 9 可以看出 AHFM 的性能优于 HFM,ADNN 的性能优于 DNN,DAHFM 的性能也比 DHFM 好。这也说明加入了注意力网络的模型是优于没有加入注意力机制的。证明了合适的注意力网络能够有效提升模型的性能,增强模型的点击率预估能力。从图 8 和图 9 也可以直观地看出后 4 个深度模型 DNN、ADNN、DHFM、DAHFM 的性能优于前两个浅层模型 AHFM 和 HFM,这说明了深度神经网络的优越性,它能够学习到浅层模型学习不到的高阶特征,虽然 DNN 和 ADNN 只学习到了高阶而没有学习到低阶特征,但他们的性能也比 HFM 和 AHFM 的性能好,这充分证明了高阶组合特征的重要性。但是低阶特征也不能忽视,由表 3 可知,DAHFM 模型的性能优于其他的消融模型,因为它不仅学习到了低阶特征而且也充分重视了高阶特征组合对模型性能的影响,使用深度神经网络对高阶特征建模,又加入注意力网络区分不同高阶组合特征的重要性,突出更加有用的高阶交叉特征,不仅增强了模型的性能也提高了模型的可解释性。通过消融实验也证明了 HFM 组件,DNN 和注意力网络对模型性能的重要性,融合了这 3 部分的 DAHFM 模型也明显优于其他的消融模型。

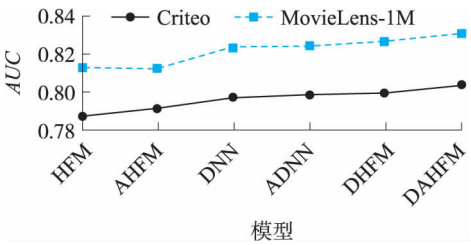

图 8 关键组件在两个数据集下的 *AUC*

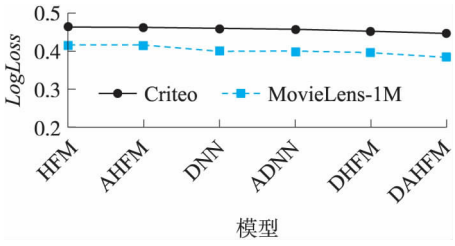

图 9 关键组件在两个数据集下的 *LogLoss*

5 结束语

针对目前点击率预估模型难以充分学习低阶特征也难以有效模拟高阶交叉特征的非线性关系,本文提出了一种基于注意力机制和深度学习的点击率预估模型 DAHFM。首先通过改进的 FM 算法 HFM 对低阶特征信息进行学习,通过实验证明 HFM 算法相对 FM 更好地学习了低阶特征。其次,利用深度神经网络来学习高阶特征信息,并加入注意力网络区分不同高阶组合特征的重要性,提高了高阶特征学习的效率。最后,构建同时学习低阶特征和高阶组合特征的模型来对广告点击率做预估。在 Criteo 和 MovieLens-1M 数据集上的大量实验结果表明了相对其他浅层和深层模型,DAHFM 的性能更加优异,点击率预估的效果也更好。

模型关键组件效果分析实验也表明,无论是低阶还是高阶特征的学习都对模型的性能以及点击率预估的准确率至关重要。因此下一阶段的研究将更加注重对特征的学习,不断研究新的特征学习方法来优化和改进现有模型,从而提高点击率预估效果。

参考文献:

[1] 陈杰浩,张钦,王树良,等. 基于深度置信网络的广告点击率预估的优化[J]. 软件学报,2019,30(12):3665-3682.
Chen Jiehao, Zhang Qin, Wang Shuliang, et al. Click-

through rate prediction based on deep belief nets and its optimization [J]. Journal of Software, 2019, 30(12): 3665-3682.

[2] Chapelle O, Manavoglu E, Rosales R. Simple and scalable response prediction for display advertising [J]. ACM Transactions on Intelligent Systems and Technology, 2015, 5(4): 1-34.

[3] 刘梦娟,曾贵川,岳威,等. 基于融合结构的在线广告点击率预测模型 [J]. 计算机学报,2019,42(7): 1570-1587.
Liu Mengjuan, Zeng Guichuan, Yue Wei, et al. A hybrid network based CTR prediction model for online advertising [J]. Chinese Journal of Computers, 2019, 42(7): 1570-1587.

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

[5] Xie Jianjun, Coggeshall S. Prediction of transfers to tertiary care and hospital mortality: A gradient boosting decision tree approach [J]. Statistical Analysis and Data Mining: the ASA Data Science Journal, 2010, 3(4): 253-258.

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

[7] Zhang Weinan, Du Tianming, Wang Jun. Deep learning over multi-field categorical data [C]//Advances in Information Retrieval. Cham, Germany: Springer International Publishing, 2016: 45-57.

[8] Guo Huifeng, Tang Ruiming, Ye Yunming, et al. DeepFM: A factorization-machine based neural network for CTR prediction [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2017: 1725-1731.

[9] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[10] Zhou Guorui, Zhu Xiaoqiang, Song Chenru, et al. Deep interest network for click-through rate prediction [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: Association for Computing Machinery, 2018: 1059-1068.

[11] Plate T A. Holographic reduced representations [J]. IEEE Transactions on Neural Networks. IEEE, 1995, 6(3): 623-641.

[12] Wang Ruoxi, Fu Bin, Fu Gang, et al. Deep & cross network for ad click predictions [C]//Proceedings of the ADKDD'17. New York, USA: Association for Computing Machinery, 2017: 1-7.

[13] Xiao Jun, Ye Hao, He Xiangnan, et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2017: 3119-3125.

[14] Lian Jianxun, Zhou Xiaohuan, Zhang Fuzheng, et al. xDeepFM: Combining explicit and implicit feature interactions for recommender systems [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: Association for Computing Machinery, 2018: 1754-1763.

[15] Song Weiping, Shi Chence, Xiao Zhiping, et al. AutoInt: Automatic feature interaction learning via self-attentive neural networks [C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, USA: Association for Computing Machinery, 2019: 1161-1170.

[16] Liu Bin, Tang Ruiming, Chen Yingzhi, et al. Feature generation by convolutional neural network for click-through rate prediction [C]//The World Wide Web Conference on-WWW'19. New York, USA: Association for Computing Machinery, 2019: 1119-1129.

[17] Huang Tongwen, Zhang Zhiqi, Zhang Junlin. FiBiNET: Combining feature importance and bilinear feature interaction for click-through rate prediction [C]// Proceedings of the 13th ACM Conference on Recommender Systems. New York, USA: Association for Computing Machinery, 2019: 169-177.

[18] Lu Wantong, Yu Yantao, Chang Yongzhe, et al. A dual input-aware factorization machine for CTR prediction [C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, 2020: 3139-3145.

[19] Tao Zhulin, Wang Xiang, He Xiangnan, et al. HoAFM: A high-order attentive factorization machine for CTR prediction [J]. Information Processing & Management, 2020, 57(6): 102076.

[20] 杨妍婷,韩斌. 基于增强型 FNN 的广告点击率预测模型 [J]. 南京理工大学学报,2020,44(1): 33-39.
Yang Yanting, Han Bin. Advertising click-through rate prediction model based on enhanced FNN [J]. Journal

of Nanjing University of Science and Technology, 2020,44(1):33-39.

[21] Deng Wei, Pan Junwei, Zhou Tian, et al. DeepLight: Deep lightweight feature interactions for accelerating CTR predictions in ad serving[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. New York, USA: Association for Computing Machinery, 2021: 922-930.

[22] Zhu Yanqiao, Xu Yichen, Yu Feng, et al. Disentangled self-attentive neural networks for click-through rate prediction [J]. arXiv preprint arXiv: 2101.03654, 2021.

[23] Tay Y, Zhang Shuai, Luu A T, et al. Holographic factorization machines for recommendation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 5143-5150.

[24] Zhang Weinan, Du Tianming, Wang Jun. Deep learning over multi-field categorical data [C]//Advances in Information Retrieval. Cham, Germany: Springer International Publishing, 2016: 45-57.

[25] Mikolov T, Chen Kai, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv: 1301.3781, 2013.

[26] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks [C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, USA: PMLR 15, 2011: 315-323.

[27] Han Song, Mao Huizi, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding [J]. arXiv: 1510.00149, 2015.

[28] Han Song, Pool J, Tran J, et al. Learning both weights and connections for efficient neural networks [J]. arXiv preprint arXiv: 1506.02626, 2015.

[29] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15: 1929-1958.

[30] Lobo J M, Jiménez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models [J]. Global Ecology and Biogeography, 2008, 17(2): 145-151.

[31] Kingma D, Ba J. Adam: A method for stochastic optimization [J]. Computer Science. arXiv preprint arXiv: 1412.6980, 2014.