面向展示广告的点击率预测模型综述

刘梦娟[1] 曾贵川[1] 岳 威[1] 仇笠舟[1] 王加昌[2]

(电子科技大学信息与软件工程学院 成都 610054)[1] (中国核动力研究设计院 成都 610213)[2]

摘 要 点击率预测模型的研究近年来备受学术界和工业界的关注。针对展示广告定向投放的点击率预测模型,研究了样本特征的预处理技术、基于传统机器学习模型的CTR预测方案、基于最新的深度学习模型的CTR预测方案、CTR预测模型的主要性能评价指标等,并基于一个开放数据集对其中的典型方案给出性能对比和量化分析,最后讨论了目前面向展示广告的点击率预测模型研究存在的问题和未来发展趋势。

关键词 点击率预测,定向广告,逻辑回归,因子分解机,深度学习

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.006

Review on Click-through Rate Prediction Models for Display Advertising

LIU Meng-juan[1] ZENG Gui-chuan[1] YUE Wei[1] QIU Li-zhou[1] WANG Jia-chang[2]

(School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)[1]

(Nuclear Power Institute of China, Chengdu 610213, China)[2]

Abstract In recent years, the study of the click-through rate prediction model has attracted much attention from academia and industry. As for the existing CTR prediction models for displaying targeted advertising, this paper studied the preprocessing techniques for features of samples, the CTR prediction schemes based on traditional machine learning models and the latest deep learning models, and the main performance evaluation indexes of CTR prediction models. Specially, these typical CTR prediction schemes were evaluated based on a public dataset, further some quantitative analysis and performance comparison were given. Finally, the problems and research trends in CTR prediction were discussed.

Keywords Click-through rate prediction, Targeted advertising, Logistic regression, Factorization machine, Deep learning

1 引言

近年来,在线广告(Online Advertising)在工业界取得了巨大成功,已经发展成为一个数十亿美元规模的产业[1]。目前,在线广告的投放主要分为两类:搜索广告(Sponsored Advertisement)和展示广告(Display Advertisement)。其中,搜索广告通常是根据用户的搜索关键词,将相应的广告与搜索结果页面同时显示,其主要面临的挑战是依照关键词与广告之间的匹配关系以及竞标价格对广告显示进行排位,这主要由搜索引擎提供商来实现并优化[2],例如百度和谷歌的搜索广告竞价机制;展示广告通常是以图像或者视频的形式展示在媒体提供的广告位上,通常由广告位的提供媒体或竞价平台将广告定向投放到广告产品的潜在客户正在浏览的页面的广告位上。对于广告商来说,展示广告除了追求品牌形象的打造,更重要的是追求每次广告投放的实际收益,即在广告展示后用户是否购买广告展示的产品[3]。因此,展示广告定向投放的一个重点就是将广告尽可能投放到那些会产生收益的曝光机会(Ad Impression)上,这就要求广告投放系统在投放之前对广告投放到曝光机会的转化率(Conversion Rate, CVR)进行预测,将广告定向投放到转化率高的曝光机会上。

图1展示了一次广告展示到转化的过程。用户在媒体页面的广告位上看到广告以后,如果产生兴趣,首先产生的是点击行为,广告点击与广告展现次数的比值称为点击率(Click Through Rate, CTR);点击行为发生以后,将会打开广告商的落地页(Landing Page),落地页打开次数与点击次数的比值称为到达率,这是在广告商网站上发生的;如果用户从落地页开始,进一步完成加购物车或下单等操作,则称为转化,转化次数与到达次数的比值称为转化率。

到稿日期:2018-07-05 返修日期:2018-10-23 本文受国家自然科学基金(61202445,61472064),四川省科技厅高新技术发展与产业化重点研发项目(2017FZ0004),桂林电子科技大学云计算与复杂系统重点实验室开放课题(170676)资助。

刘梦娟(1979-),女,博士,副教授,主要研究方向为机器学习、计算广告、推荐系统,E-mail:mjliu@uestc.edu.cn(通信作者);**曾贵川**(1993-),男,硕士生,主要研究方向为机器学习、计算广告;**岳 威**(1995-),男,硕士生,主要研究方向为机器学习、计算广告;**仇笠舟**(1995-),男,硕士生,主要研究方向为机器学习、计算广告;**王加昌**(1978-),男,硕士,高级工程师,主要研究方向为系统仿真、机器学习、大数据分析。

图 1　展示广告从投放到落地到转化的过程示意图

Fig. 1　Process of display advertisement from delivery to landing to conversion

遗憾的是，CVR 很难准确预测，这是因为转化行为比点击率行为发生的频率更低，且转化行为相对于点击行为可能会有较长的延迟（有时候会达到一周），这使得离线建模非常困难。因此，无论在学术界还是工业界，更多地是将广告投放到曝光机会的预测点击率作为广告定向投放或者实时竞价系统中出价的依据。工业界已经举办了多次点击率或转化率预测大赛，为学术界的研究提供了真实有效的广告投放/点击数据集，如表 1 所列。

表 1　CTR/CVR 预测大赛及公开数据集

Table 1　CTR/CVR predicting competition and public data set

数据集	说明
Avazu 2015	点击率预测大赛 https://www.kaggle.com/c/avazu-ctr-prediction/data
Criteo 2013	点击率预测大赛 http://labs.criteo.com/2013/12/download-terabyte-click-logs/
Criteo 2013	转换率预测大赛 http://labs.criteo.com/2013/12/conversion-logs-dataset/
iPinYou 2014	KDD 2014，点击日志可用于 CTR 预测 http://contest.ipinyou.com/

CTR 预测或 CVR 预测可以归纳为典型的回归问题。目前的研究方法主要分为基于传统机器学习模型的方案和基于最新的深度学习模型的方案。本文首先对 CTR 预测问题展开形式化描述；然后给出 CTR 预测模型建模的基本流程；在此基础上分别对传统的建模方案和最新的建模方案进行介绍，并对比各种 CTR 预测模型的特点；最后在真实数据集上对各种典型方案的预测性能进行量化分析，并对各种关键技术点进行了详细讨论。本文的特色之处在于：对目前主流的和最新的 CTR 预测模型进行了全面介绍，并对各种典型模型的实现代码进行了归纳整理；对目前已有的实验数据集进行了整理，为相关学者进行后续研究提供了数据基础；对各种典型算法的原理及性能进行了量化分析，在实验基础上对其中的关键技术点展开了讨论，并针对其中存在的问题及可行的解决思路进行了阐述。

2　点击率预测问题的形式化描述

点击率预测问题作为一个典型的回归问题，其建模方法如图 2 所示。学习系统基于给定的训练数据集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}$ 构建预测模型 $Y=f(\boldsymbol{x}_i)$，然后预测系统根据预测模型预测每个符合投放规则的广告对于新到达的曝光机会 $\boldsymbol{x}_{N+1}$ 的点击率 y_{N+1}。其中，$(\boldsymbol{x}_i,y_i)$ 表示训练样本 i，$\boldsymbol{x}_i$ 是样本 i 的特征向量，y_i 表示广告是否被点击的标签，$y_i=1$ 表示发生过点击，$y_i=-1$ 表示没有发生点击。

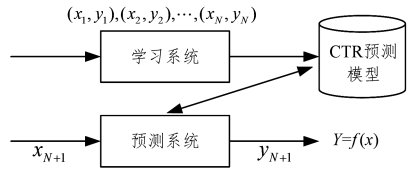

图 2　点击率预测问题

Fig. 2　Click-through rate predicting problem

在 CTR 预测模型的学习过程中，需要确定 3 个要素：1）模型，即 $f(\cdot)$ 的形式，CTR 预测可以采用逻辑回归模型（Logistic Regression，LR）、因子分解机模型（Factorization Machine，FM）、深度神经网络（Deep Neural Network，DNN）等；2）策略，即选择什么准则来学习最优模型，在有监督学习中通常引入损失函数 $L(\cdot)$ 来度量预测错误的程度，学习使得损失函数最小的模型参数，因此 CTR 预测问题可以描述为式（1）所述的形式，对于不同的模型，可以设计不同的损失函数，例如 LR 中通常使用对数似然损失函数[4]；3）求解最优模型的参数向量 θ 的计算方法，在 CTR 预测中常用的优化算法包括梯度下降（Gradient Descent）和随机梯度下降（Stochastic Gradient Descent）[5]。

$$\min_{\theta} L(f(x_i),y_i) \tag{1}$$

3　样本特征的预处理技术

在展示广告的点击率预测中，样本使用的特征主要是分类特征（Categorical Features），例如用户的性别（Gender）、所在的城市（City）等。分类特征不能直接用于预测计算，因此通常使用独热（One-hot）编码[6]对分类特征进行预处理。独热编码的方法如式（2）所示，假设分类特征 c 在数据集中有 M 种取值可能，将 c 编码为一个由二值元素组成的 M 维向量，每个元素 $b^i\in\{0,1\}$。如果数据集中还包括连续值的特征（例如 Criteo 数据集），则首先利用分箱技术将其转化为分类特征，再按照分类特征的预处理方法完成该特征的独热编码。

$$c=(b^1,b^2,\cdots,b^M) \tag{2}$$

其中，$\sum_{i=1}^{M}b^i=1$。

假设数据集中 Gender 特征有 2 种可能取值（Female/Male），因此 Gender 特征编码为 2 比特，[0,1]表示 Female，[1,0]表示 Male；City 特征有 3 种可能取值（Beijing/Shanghai/Chengdu），因此 City 特征可编码为 3 比特，对应为[0,0,1]，[0,1,0]，[1,0,0]。例如，一个位于北京的男性用户，其编码后的特征向量为：

$$\underbrace{[1,0]}_{Gender=Male}\quad\underbrace{[0,0,1]}_{City=Beijing}$$

在 CTR 预测中，通常将独热编码后的每个比特称为一个特征，例如[1,0]中第一个比特表示男性特征，第二个比特表示女性特征；而将相同物理属性的不同取值对应的若干比特称为一个特征域(field)，例如这里的男性特征和女性特征都属于性别特征域，记为$[b^1,b^2]$。对于一个样本，每个特征域只有样本中出现的特征才取值为 1，其余取值为 0，因此独热编码后的样本特征向量是一个超高维度的稀疏向量。在简单的逻辑回归模型和因子分解机模型中，通常使用独热编码后的原始特征向量作为输入。

最新的深度学习模型也被用于预测 CTR，然而在 DNN 中，输入层节点通常与隐层节点全连接，如果直接将独热编码后的原始特征向量作为输入，将会导致巨大的计算开销[7]。因此，研究者通常会在输入层和第一个隐藏层之间增加一个嵌入层(Embedding Layer)，用于减少 DNN 的输入单元数。这里引入嵌入向量(Embedding Vector)的概念，将独热编码后的每个特征映射为一个固定维度的嵌入向量，再将一个样本包含特征对应的所有嵌入向量拼接起来作为输入。假设独热编码后的特征数为 n，每个嵌入向量的维度为 D，则特征的嵌入向量 $\boldsymbol{v}$ 可以写为一个矩阵的形式，如式(3)所示，其中每个特征 $\boldsymbol{v}_i$ 可以表示为一个实数组成的稠密向量。

$$\boldsymbol{v}=\begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{v}_2\\ \vdots\\ \boldsymbol{v}_n\end{bmatrix}=\begin{bmatrix}\boldsymbol{v}_1^1 & \boldsymbol{v}_1^2 & \cdots & \boldsymbol{v}_1^D\\ \boldsymbol{v}_2^1 & \boldsymbol{v}_2^2 & \cdots & \boldsymbol{v}_2^D\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{v}_n^1 & \boldsymbol{v}_n^2 & \cdots & \boldsymbol{v}_n^D\end{bmatrix} \tag{3}$$

对于一个样本，由于属于相同特征域的特征只有 1 个有取值，因此 DNN 模型的输入单元数为 $N\times D$，其中 N 表示特征向量中特征域的个数。式(4)展示了将特征向量映射为嵌入向量后作为输入的一个示例。假设独热编码后的样本特征为[1,0,0,0,1]，其中前 2 个比特为 Gender 特征域，后 3 个比特为 City 特征域，嵌入向量的维度为 2，这里只有男性特征和 Beijing 特征的值为 1，因此只需要将男性特征和 Beijing 特征的嵌入向量的拼接 [0.2,0.8,0.6,0.4] 作为输入，映射后的嵌入单元的数量为 4。从稀疏高维二值向量到密集实数嵌入向量的映射一般有两种方法：1)基于因子分解机(FM)模型预训练完成，将 FM 训练得到的每个特征的隐含向量作为嵌入向量，早期的 FNN 和 PNN 等均使用 FM 模型进行嵌入向量的预训练；2)将原始特征到嵌入向量的映射作为预测模型的一部分，进行联合训练，嵌入向量的初始值随机设置，如图 3 所示，权重通过学习得到，最新的 Wide&Deep[8]，DeepFM[9]，Deep&Cross[10] 等方案均采用这种联合学习的方法来得到嵌入向量。

$$\begin{aligned}&[1,0]\Rightarrow[0.4,0.6][0,0,1]\Rightarrow[0.3,0.7]\\&[1,0,0,0,1]\Rightarrow[0.4,0.6,0.3,0.7]\end{aligned} \tag{4}$$

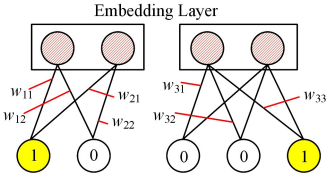

图 3 嵌入向量联合学习的示例

Fig. 3 Example of embedding vector in joint learning

4 基于传统机器学习模型的预测方案

针对点击率预测问题，最早的解决方案是利用逻辑回归(LR)[4,39-42]来学习点击率预测模型。LR 中，定义点击率预测公式如式(5)所示，其中 $\boldsymbol{x}_i$ 表示广告展示机会的特征向量，y_i 表示广告是否被点击的真值，$y_i=1$ 表示发生过点击，$y_i=-1$ 表示没有发生点击，$\boldsymbol{w}\in\boldsymbol{R}^{n+1}$ 表示模型中参数的向量，n 表示特征向量的维度，因此该次广告展示未发生点击的概率公式如式(6)所示；可将式(5)和式(6)合并为式(7)。

$$P(y_i=1|\boldsymbol{x}_i)=\frac{1}{1+\mathrm{e}^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}} \tag{5}$$

$$P(y_i=-1|\boldsymbol{x}_i)=1-\frac{1}{1+\mathrm{e}^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}}=\frac{1}{1+\mathrm{e}^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}} \tag{6}$$

$$P(y_i=\pm1|\boldsymbol{x}_i)=\frac{1}{1+\mathrm{e}^{-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)}} \tag{7}$$

在逻辑回归模型中，损失函数通常使用对数损失函数(即负的对数似然函数)，如式(8)所示，其中 m 表示训练集中的样本数。为了防止模型过拟合，通常在损失函数中加入 L2 正则化项，因此可以通过最小化损失函数来学习 LR 模型的参数 w，如式(9)所示，其中 λ 表示正则化参数。逻辑回归模型的参数求解算法非常多，除了牛顿法、拟牛顿法，还包括随机梯度下降法[11]、坐标下降法等[12]，也可以使用 FOBOS[13]，RAD[14]，FTRL[15]，FTRL-Proximal[15] 等在线算法求解。

$$\begin{aligned}L(\boldsymbol{w})&=-\sum_{i=1}^{m}\log\left(\frac{1}{1+\mathrm{e}^{-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)}}\right)\\&=\sum_{i=1}^{m}\log(1+\exp(-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)))\end{aligned} \tag{8}$$

$$\min_{w}\left\{\sum_{i=1}^{m}\log(1+\exp(-y_i(f(\boldsymbol{w},\boldsymbol{x}_i)))+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right\} \tag{9}$$

其中，$f(\boldsymbol{w},\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i$。

逻辑回归模型是一种广义线性模型，非常容易实现大规模的实时并行处理，因此在工业界获得了广泛应用，但是线性模型的学习能力有限，不能捕获高阶特征(非线性信息)。为此，文献[16]提出了 Poly2 模型，该模型不仅考虑了单个特征携带的信息，而且考虑了二阶组合特征 (Feature Conjunction)携带的信息，因此 Poly2 将式(9)中的 $f(\boldsymbol{w},\boldsymbol{x}_i)$ 改写为式(10)，不仅考虑了一阶特征对应的权重，而且考虑了二阶组合特征对应的权重，其中 $w_h(k,l)$ 表示样本的第 k 个特征和第 l 个特征组合对应的权重，x_i^k 表示样本 $\boldsymbol{x}_i$ 的第 k 个特征的值，因此 Poly2 模型需要为每个组合特征学习一个权重。Poly2 的问题在于，当样本的特征维度非常大时，二阶组合特征的权重计算的复杂度将变得非常大，为 $O(\bar{n}^2)$，其中 $\bar{n}$ 表示样本中非 0 元素的平均值；此外，如果某个特征组合在训练集中没有出现，那么对应项的权重将不能得到充分学习，从而降低了预测的准确性。

$$f(\boldsymbol{w},\boldsymbol{w}_h,\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\sum_{k=1}^{n}\sum_{l=k+1}^{n}(w_h(k,l)\cdot x_i^k\cdot x_i^l) \tag{10}$$

针对 Poly2 的问题，文献[17]提出了基于因子分解机(FM)[18]的 CTR 预测模型，其基本思想是将式(9)中的 $f(\boldsymbol{w},\boldsymbol{x}_i)$ 改写为式(11)的形式，其中 $\boldsymbol{v}_k$ 和 $\boldsymbol{v}_l$ 分别表示特征 k 和特征 l 的维度为 D 的隐含向量。因此，在 FM 中每个特征用一

个 D 维的隐含向量表示，从而使二阶的组合特征的权重分解为两个隐含向量的点积，将计算复杂度降低为 $O(D\cdot\bar{n})$；同时，即使训练集中没有出现某个特征组合，也由于两个特征的隐含向量是分别学习的而不会影响预测的准确性。FM 的参数求解通常采用随机梯度下降算法。

$$f(\boldsymbol{w},\boldsymbol{v},\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\sum_{k=1}^{n}\sum_{l=k+1}^{n}(\langle\boldsymbol{v}_k\cdot\boldsymbol{v}_l\rangle\cdot x_i^k\cdot x_i^l) \tag{11}$$

其中，$\langle\boldsymbol{v}_k\cdot\boldsymbol{v}_l\rangle=\sum_{d=1}^{D}\boldsymbol{v}_k^d\boldsymbol{v}_l^d$。

FM 的缺陷在于每个特征都只学习一个唯一的隐含向量，在与其他不同特征进行组合时，同一个特征产生的影响力都是相同的；而事实上，当与不同特征域的特征组合时，特征可能表现出不同的隐含特征分布。例如，对于样本“一个女性(Female)用户在发布媒体 Vogue 上浏览页面时，对投放的 Gucci 广告发生了点击行为”，在 FM 中，学习二阶的组合特征时，只需要学习 3 个隐含特征向量，无论与特征 Gucci 还是与特征 Female 进行组合时，特征 Vogue 的隐含向量都是相同的；而事实上，我们更希望在与不同领域的特征进行组合时考虑差异化的隐含特征向量。为此，文献[19]在 FM 模型的基础上引入了特征域(Field)的概念，提出了面向特征域的因子分解机(Field-aware Factorization Machines，FFM)模型。其基本思想是将特征分割为若干领域，例如将特征 Gucci 划分为 Advertiser 域，将特征 Female 划分为 Gender 域，将特征 Vogue 划分为 Publisher 域，每个特征将针对不同的特征域学习不同的隐含向量，因此式(11)中的 $f(\boldsymbol{w},\boldsymbol{v},\boldsymbol{x}_i)$ 改写为式(12)的形式，其中 $\boldsymbol{v}_{k,f_l}$ 和 $\boldsymbol{v}_{l,f_k}$ 分别表示特征 k 在特征 l 所属的 f_l 域以及特征 l 在特征 k 所属的 f_k 域的隐含向量。FFM 的参数求解通常采用随机梯度下降以及改进的 AdaGrad[19] 算法。

$$f(\boldsymbol{w},\boldsymbol{v},\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\sum_{k=1}^{n}\sum_{l=k+1}^{n}(\langle\boldsymbol{v}_{k,f_l}\cdot\boldsymbol{v}_{l,f_k}\rangle\cdot x_i^k\cdot x_i^l) \tag{12}$$

Poly2，FM，FFM 都是在 LR 基础上增加对二阶特征组合的权重自动学习的模型。除此之外，Facebook 的研究人员还提出了另一种筛选特征和特征组合的方式，称为 GBDT＋LR 方案[20]。该方案利用 GBDT(Gradient Boost Decision Tree)来帮助筛选有区分度的特征和特征组合，并将其作为 LR 模型的输入，从而增强 LR 的非线性学习能力，基本模型如图 4 所示。

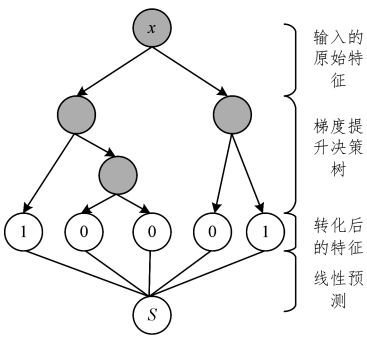

图 4　GBDT＋LR 方案

Fig. 4　Solution of GBDT＋LR

GBDT 是一种非线性模型，它基于集成学习中 boosting 的思想[21]，每次迭代都在减少残差的梯度方向新建立一棵回归树，每个叶子节点作为一个取值为 0/1 的输入特征，因此新特征向量的长度等于 GBDT 模型里所有回归树包含的叶子节点的总数。在图 4 的例子中，有两棵回归树，包含 5 个叶子节点，因此输入到 LR 的新特征向量的维度为 5。假设当样本 x 输入 GBDT 时，它在左边的树中落入第一个节点，在右边的树中落入第二个节点，则新的特征向量编码为[1，0，0，0，1]，再将其输入到 LR 中进行模型学习。除了 GBDT＋LR 外，2014 年 Kaggle CTR 的冠军队使用 GBDT＋FM 的融合方案[22]取得了非常好的预测效果。

基于传统机器学习模型的 CTR 预测方案又被称为基于浅层模型的方案，其优点是模型简单、预测性能较好、可解释性强；缺点主要在于很难自动提取高阶组合特征携带的信息，目前一般通过特征工程来手动提取高阶组合特征。

5　基于深度学习模型的预测方案

随着深度学习在计算机视觉[23]、语音识别[24]、自然语言处理[25]等领域取得巨大成功，其探索特征间高阶隐含信息的能力也被应用到了 CTR 预测中。较早的具有影响力的基于深度学习模型的 CTR 预测方案是 Zhang 等[7]在 2016 年提出的基于因子分解机的神经网络(Factorization Machine supported Neural Network，FNN)模型，如图 5 所示。该模型利用一个带嵌入层(Embedding Layer)的深度神经网络(DNN)来完成点击率预测，其特点是通过 FM 模型预先训练得到每个特征的稠密嵌入向量(Dense Vector)，将样本的所有嵌入向量拼接起来作为 DNN 的输入进行训练。FNN 的特点是每个特征的嵌入向量是预先采用 FM 模型训练的，因此在学习 DNN 模型时，训练开销降低，模型能够更快地达到收敛。这里的 DNN 是一个包含多个隐层的前馈神经网络，用于学习高阶特征之间的交互信息，其中嵌入层的每个节点与第 1 个隐层的每个节点全连接，第 1 个隐层中每个节点的输出值采用式(13)计算，其中 $\boldsymbol{h}_1\in\mathbb{R}^{n_1}$ 是第 1 个隐层节点的输出向量，n_1 是第 1 个隐层的节点数，$\boldsymbol{W}_0$ 表示嵌入层节点到第 1 个隐层节点的连接权重，$\boldsymbol{W}_0\in\mathbb{R}^{n_1\times n_0}$，$n_0$ 是嵌入层的节点数，$\boldsymbol{x}_0\in\mathbb{R}^{n_0}$ 是嵌入层的输出向量，$\boldsymbol{b}_0$ 表示第 1 个隐层的偏置向量，$\boldsymbol{b}_0\in\mathbb{R}^{n_1}$。

$$\boldsymbol{h}_1=f(\boldsymbol{W}_0\boldsymbol{x}_0+\boldsymbol{b}_0) \tag{13}$$

每个隐层的节点数和隐层的层数可调整，隐层之间每个节点均采用全连接，第 $l+1$ 个隐层节点的输出值按式(14)进行计算，其中 $\boldsymbol{W}_l$ 表示第 l 个隐层节点到第 $l+1$ 个隐层节点的连接权重，$\boldsymbol{W}_l\in\mathbb{R}^{n_{l+1}\times n_l}$，$n_l$ 和 n_{l+1} 分别是第 l 个隐层和第 $l+1$ 个隐层的节点数，$\boldsymbol{h}_l\in\mathbb{R}^{n_l}$ 是第 l 个隐层节点的输出值，$\boldsymbol{b}_l$ 表示第 $l+1$ 个隐层的偏置向量，$\boldsymbol{b}_l\in\mathbb{R}^{n_{l+1}}$，FNN 的隐层中所有节点的激活函数 $f(\cdot)$ 都采用 tanh 函数[26]。输出节点用于计算预测点击率，输出节点的激活函数采用 sigmoid 函数[26]，预测点击率 p 的计算公式如式(15)所示，其中 $\boldsymbol{W}_L^S$ 表示最后一个隐层到输出节点的权重向量，$\boldsymbol{W}_L^S\in\mathbb{R}^{n_L}$，$\boldsymbol{h}_L^S$ 表示最后一个隐层的输出向量，$\boldsymbol{h}_L^S\in\mathbb{R}^{n_L}$，$b_L^S$ 表示输出节点的偏置。

$$\boldsymbol{h}_{l+1}=f(\boldsymbol{W}_l\boldsymbol{h}_l+\boldsymbol{b}_l) \tag{14}$$

$$p=sigmoid(\boldsymbol{W}_L^S\boldsymbol{h}_L^S+b_L^S) \tag{15}$$

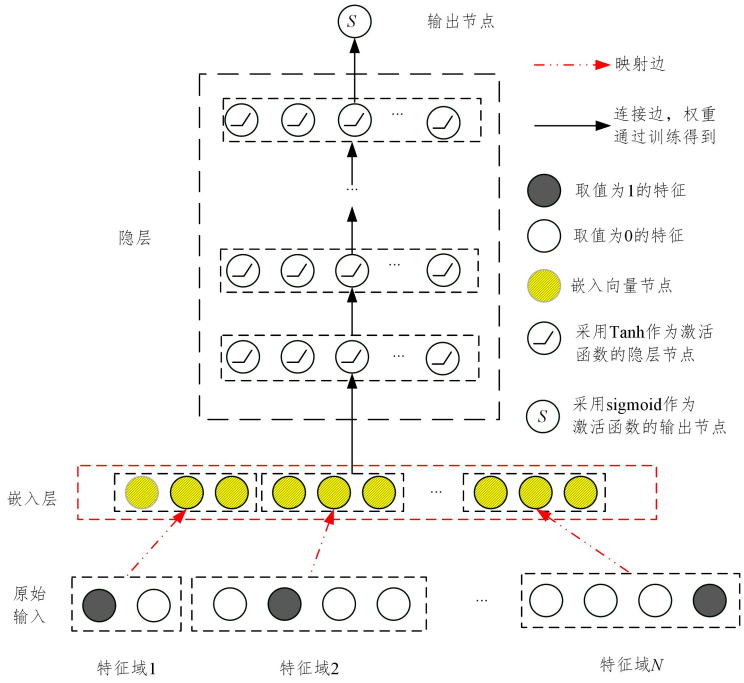

图 5 FNN 模型

Fig. 5 FNN model

文献[27]也提出了一个基于 DNN 的 CTR 预测模型，其在 DNN 的结构中引入了一个 product 层，DNN 的输入单元不仅包括每个特征的嵌入向量，还包括任意两个特征嵌入向量的积运算，这种方案称为 PNN(Product-based Neural Network)。根据积运算的不同类型，该方案有 3 种变化：IPNN，OPNN，PNN*。其中，IPNN 表示任意两个特征域的嵌入向量做内积，OPNN 表示任意两个特征域的嵌入向量做外积，PNN* 表示将内积和外积的输出结果拼接起来。不同于 FNN，PNN 在第 1 个隐层的输入中不仅考虑了一阶特征的嵌入向量，还考虑了任意两个特征嵌入向量之间的组合操作。PNN 的网络结构如图 6 所示，这里 PNN 的隐层节点采用了不同的激活函数 ReLU，任意两个特征的嵌入向量做内积的公式如式(16)所示，其中 $p_{i,k}$ 表示嵌入向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_k$ 的内积。PNN 的嵌入向量也是预先采用 FM 模型进行预训练的。

$$p_{i,k}=\langle \boldsymbol{v}_i \cdot \boldsymbol{v}_k \rangle=\begin{bmatrix} v_i^1 & v_i^2 & \cdots & v_i^D \end{bmatrix}\begin{bmatrix} v_k^1 \\ v_k^2 \\ \vdots \\ v_k^D \end{bmatrix}=\sum_{t=1}^{D} v_i^t v_k^t \quad (16)$$

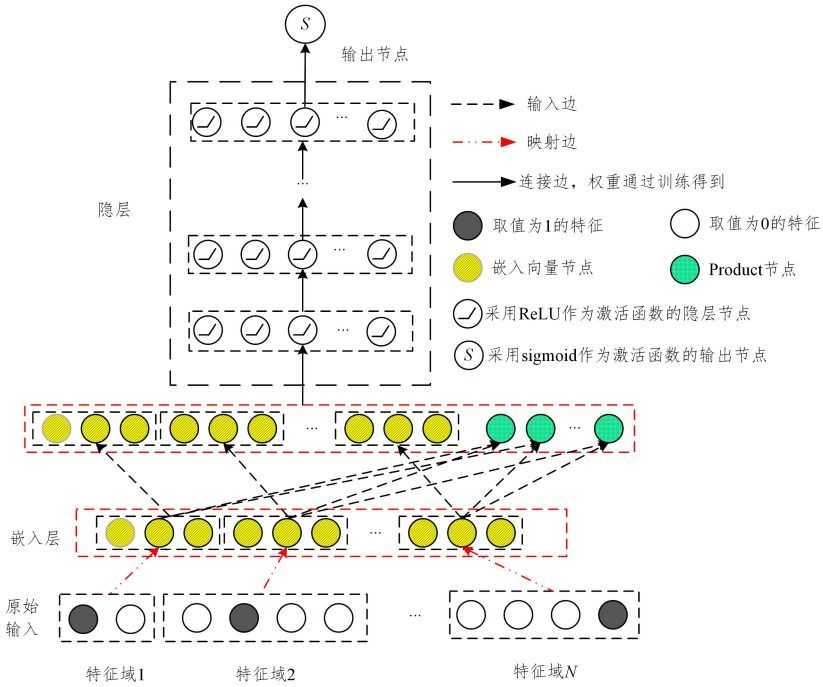

图 6 PNN 模型

Fig. 6 PNN model

FNN 和 PNN 充分利用了 DNN 对特征高阶隐含信息的表示能力，但忽略了一阶特征携带的信息，而文献[8]的实验证明一阶特征对于 CTR 的预测也是非常重要的。为此，Google 的研究人员[8]提出了一种深度学习模型与线性模型的融合结构 Wide & Deep。该结构将线性模型和深度学习模型进行巧妙的融合，如图 7 所示，通过将线性模型和 DNN 结合起来联合训练，不仅考虑了低阶特征携带的信息，也考虑了高阶特征之间的交互信息，因此能够获得超过 FNN 和 PNN 的预测性能，但是其 Wide 部分仍然依赖于手动的特征工程。Wide 部分的特征包括原始的输入特征和转换特征，其中转化特征是采用一种 cross-product 操作，需要由人工来确定哪些特征进行该操作。式(17)定义了 cross-product 操作，即“与”操作，只有选择组合的一阶特征都取值为 1 时，组合特征才为 1。Deep 部分，分类特征需要首先转换为嵌入向量，数值特征直接输入，嵌入向量随机初始化，然后参与训练学习，以最小化最终的损失函数。注意，Wide & Deep 的嵌入向量不是预训练的，而是通过模型学习权重参数得到的。

$$\phi_k(x)=\prod_{i=1}^{d}x_i^{c_{ki}},c_{ki}\in\{0,1\} \tag{17}$$

文献[9]在 Wide & Deep 的基础上，将线性模型(Wide)替换为 FM 模型，从而提出 DeepFM，如图 8 所示。该结构将 FM 模型和 DNN 模型结合起来联合训练，优点是不需要特征工程的支持，也可以同时学习低阶和高阶特征的相互作用。FM、Deep、嵌入向量都是作为模型组成部分进行端到端的联合训练的，因此这里的嵌入向量也不是预先训练的，而是通过训练学习的。

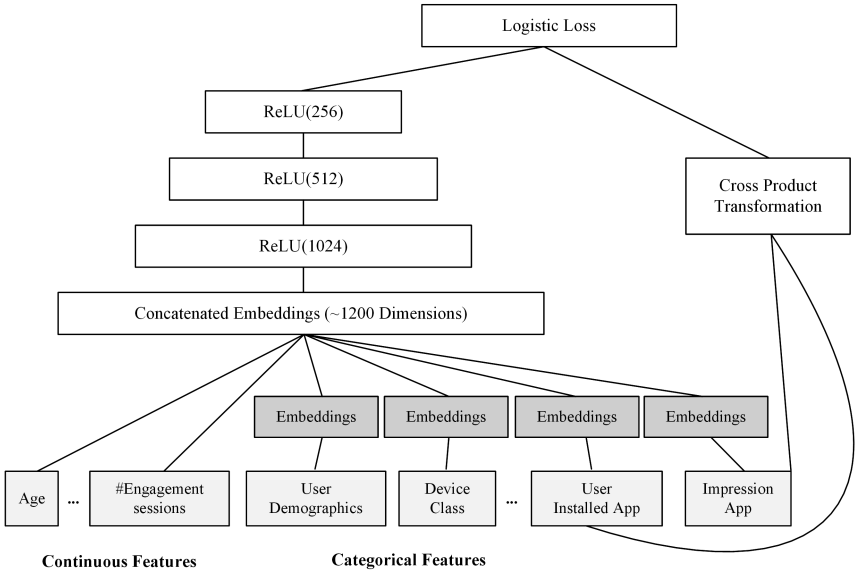

图 7　Wide & Deep 模型

Fig. 7　Wide & Deep model

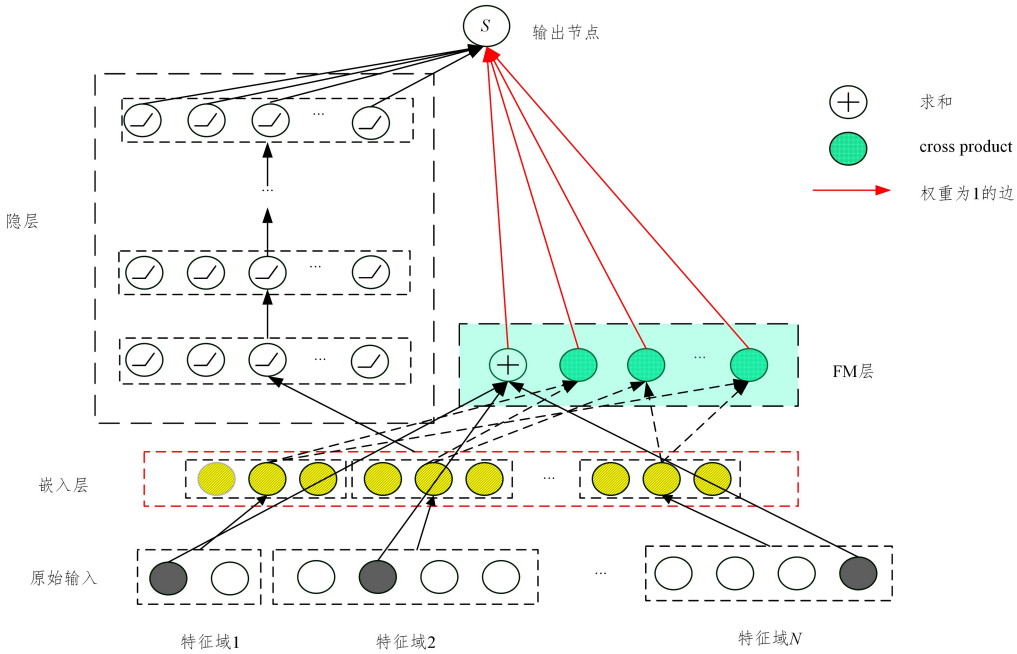

图 8　DeepFM 模型

Fig. 8　DeepFM model

目前基于 DNN 融合结构的 CTR 预测模型逐渐成为学术界和工业界研究的热点。Google 的研究人员进一步在 Wide & Deep 的基础上提出了 Deep & Cross 模型[10]。考虑到 DNN 是隐含地学习特征间的相互作用，并不是所有特征交叉都是有效的，该模型引入了 Cross 网络，可以自动、显式、有限度地进行特征交叉，并将 Cross 网络与普通 DNN 网络进行并行训练。该模型被称为深度交叉网络(Deep & Cross Network，DCN)，其结构如图 9 所示。

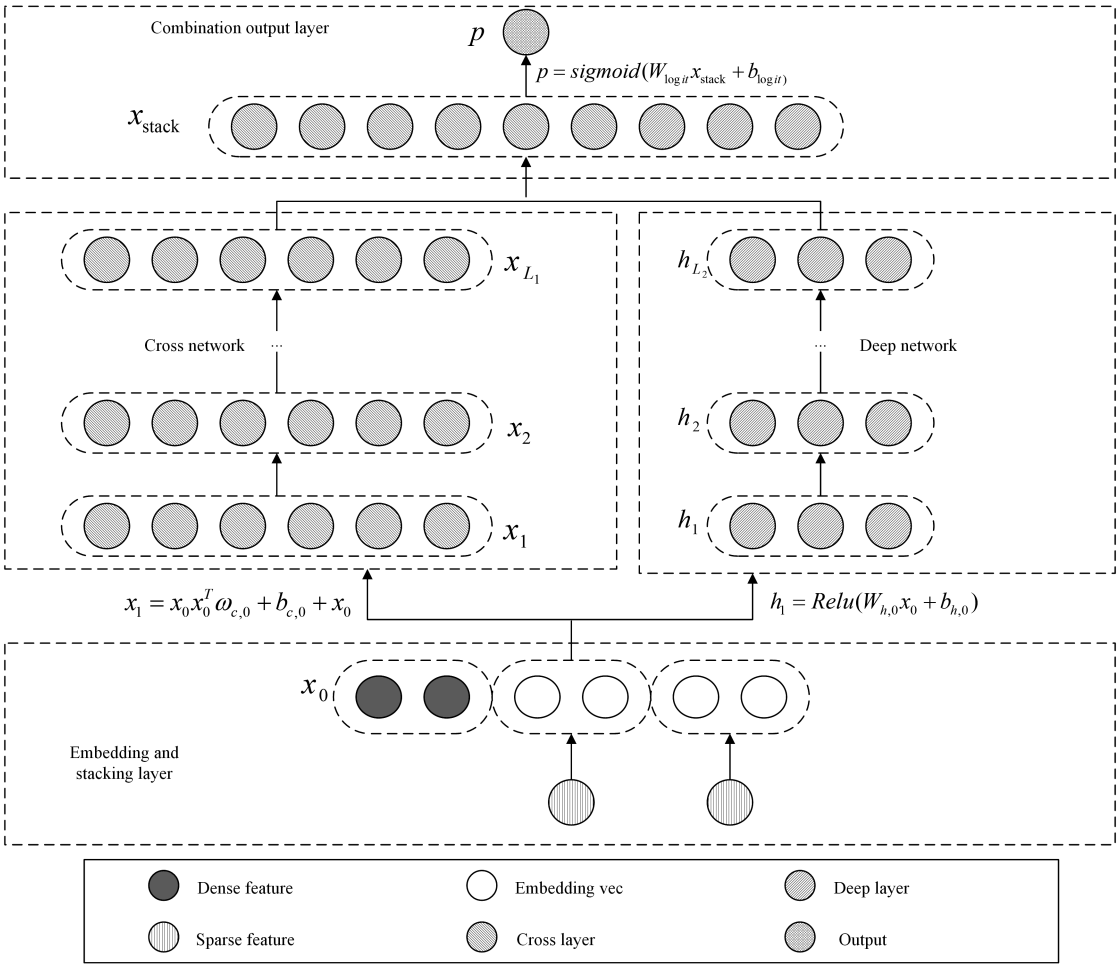

图 9 Deep & Cross 模型

Fig. 9 Deep & Cross model

连续值特征被直接输入到模型，稀疏分类特征被映射为嵌入向量后再输入到模型。其中，右边的 Deep Network 就是普通的 DNN 网络；左边的 Cross Network 是文献[10]新提出的结构(如图 10 所示)，用于自动完成特征的交叉组合，每层的输出如式(18)所示，其中 $\boldsymbol{x}_{l+1}, \boldsymbol{x}_l \in \mathbb{R}^D$ 是列向量，对应第 $l+1$ 和第 l 个交叉层的输出，$\boldsymbol{w}_l$ 和 $\boldsymbol{b}_l$ 分别表示第 l 层的权重向量和偏置向量。

$$\boldsymbol{x}_{l+1} = \boldsymbol{x}_0 \boldsymbol{x}_l^{\mathrm{T}} \boldsymbol{w}_l + \boldsymbol{b}_l + \boldsymbol{x}_l = f(\boldsymbol{x}_l, \boldsymbol{w}_l, \boldsymbol{b}_l) + \boldsymbol{x}_l \tag{18}$$

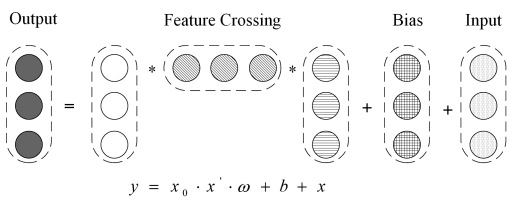

图 10 交叉层的示意图

Fig. 10 Schematic diagram of cross layers

实验表明，与单个 DNN 相比，DCN 可以以更低的计算开销更加有效地捕获到有效的特征组合，从而获得更优的预测性能；同时参数数量减少了一个数量级。需要注意的是，DCN 的嵌入向量与 Wide & Deep 学习方法的嵌入向量相同，是通过联合学习模型参数得到的。

6 典型方案的特点对比及量化分析

第 4 节和第 5 节对目前已经提出的 9 种典型 CTR 预测方案进行了介绍，本节将首先分析这些典型方案的特点；然后利用 iPinYou 公开数据集[28]对这 9 种典型方案的预测性能进行定量评价，并在 GitHub 上分享了这 9 种方案的实现代码[1)]。

6.1 典型方案的特点对比

首先分析这几种典型方案的共同点，可以发现无论是基于传统机器学习模型的方案，还是基于深度学习模型的方案，其关键技术都可以归纳为两个方面：1)模型结构的选择；2)输

1) https://github.com/mjliu-advertising

入模型的特征表示。在模型结构的选择上，CTR 预测模型可以设计为：单一浅层模型的结构，如 LR，FM，FFM；两个浅层模型的融合结构，如 GBDT＋LR；单一深度学习模型（DNN）的结构，如 FNN 和 DNN；浅层模型和深度学习模型的融合结构，如 Wide & Deep 和 DeepFM；深度学习模型和深度学习模型的融合结构，如 Deep & Cross。

在输入模型的特征表示上，点击率预测模型的训练样本通常以分类特征为主，有的也包含数值特征。对于分类特征，通常采用第 3 节所述的独热编码技术；对于连续值特征，大多数方案首先采用分箱技术将其转化为分类特征，再利用独热编码将其转化为二值向量，所有的特征向量拼接起来就是模型的输入特征向量。LR，FM，FFM 都是采用这种原始的稀疏高维二值特征向量作为输入，但不同于 LR 只考虑一阶特征和手动选择的组合特征，FM 和 FFM 在模型中还考虑了二阶组合特征，并能对二阶组合特征的权重进行自动学习。另一种思路是首先利用一种模型对有效的一阶特征及组合特征进行筛选，将筛选出的特征用于预测，GBDT＋LR 和 GBDT＋FM 均是这种思路，其中 GBDT 就是作为特征筛选的一种预处理技术。浅层模型的优点是物理含义可解释，能够得出每个特征和特征组合对预测的重要程度。

基于深度学习模型的 CTR 预测方案，其模型的输入不能直接采用独热编码后的稀疏高维特征向量，这是因为通常输入层节点到隐层节点采用的是全连接方式，如果使用原始特征，将导致需要学习的模型参数非常庞大，会降低学习效率。一种通用的方法是将原始特征转化为嵌入向量后再输入到 DNN 中，这就涉及到对嵌入向量的学习。目前，嵌入向量的学习有两种方式：1）在 FNN 和 PNN 模型中，嵌入向量的学习是通过 FM 模型预训练得到的，在训练 DNN 的模型参数时，嵌入向量直接代入；2）在 Wide&Deep，DeepFM 和 Deep&Cross 模型中，嵌入向量是作为训练模型的一部分，联合学习得到的。

基于深度学习模型的 CTR 预测方案，其本质是利用了深度学习模型对高阶特征的表达能力，因此 DNN 中最后一个隐层的输出可以理解为是以隐含方式进行组合的高阶特征，而这些高阶组合特征将全连接到最后的输出节点。不同于 FNN 和 PNN 只考虑高阶特征，Google 的研究人员通过实验发现低阶特征对于 CTR 的预测也非常重要，因此提出了 Wide & Deep 融合结构，既考虑高阶特征表示，也考虑低阶特征（一阶和手动的组合特征）。DeepFM 通过将融合结构的线性部分替换为 FM 模型，有效避免了对特征工程的依赖；Deep & Cross 则是将线性部分替换为一个深度的交叉网络，使得可以自动得到显式的、有限度的组合特征。

表 2 和表 3 分别给出了基于传统机器学习模型和基于深度学习模型的预测方案的特点对比。最后，可以发现典型方案都是采用 sigmoid 函数作为最终的输出映射函数，且最优化求解算法大多数都采用随机梯度下降算法。

表 2　基于传统机器学习模型的方案的特点

Table 2　Features of scheme based on traditional machine learning model

方案	模型结构	特征向量	特点
LR	单/浅层	原始稀疏超高维度	只考虑一阶特征，高阶特征信息需要通过特征工程手动增加
FM	单/浅层	原始稀疏超高维度	考虑一阶特征和二阶特征，自动完成
FFM	单/浅层	原始稀疏超高维度	考虑一阶特征和二阶特征，考虑不同特征域之间特征两两组合的权重倾向性
GBDT＋LR	融合结构，GBDT＋LR	原始稀疏超高维度	利用 GBDT 完成一阶特征和高阶特征组合的筛选

表 3　基于深度学习模型的方案的特点

Table 3　Features of scheme based on deep learning model

方案	模型结构	是否预训练	特点
FNN	单/DNN	是	只考虑高阶特征表示
PNN	单/DNN	是	只考虑高阶特征表示，DNN 输入不仅包含每个特征的嵌入向量，还包含任意两个嵌入向量的“积操作”
Wide & Deep	融合结构，LR＋DNN	否	同时考虑高阶特征和低阶特征，线性部分的“叉积”需要手动干预
DeepFM	融合结构，FM＋DNN	否	同时考虑高阶特征和低阶特征，低阶特征部分自动学习完成
Deep & Cross	融合结构，Cross＋DNN	否	同时考虑高阶特征和低阶特征，低阶特征通过自动学习完成

6.2　典型方案的性能分析

本节将进一步利用 iPinYou 公开数据集对典型方案的 CTR 预测性能进行量化分析，并就以下几个问题进行探讨：1）负样本采样对于提升预测性能是否有效；2）深度学习模型是否对预测性能有实质提升；3）融合模型是否对预测性能有实质提升。目前，点击率预测的评价指标主要包括两个：AUC 和 LogLoss[29]。其中，AUC 是 ROC 曲线下的面积，其值越大，说明 CTR 预测模型的性能越好；LogLoss 是交叉熵损失，其越小，说明预测模型的性能越好。

6.2.1　iPinYou 数据集

iPinYou 数据集是 iPinYou 公司在 2013 年发布的一个真实广告投放的数据集，包括曝光机会、竞价、点击、转化 4 类日志，其中的曝光机会和点击日志可用于点击率预测。具体来说，在本文实验中，每个样本对应了一次广告曝光，特征信息包括用户的相关信息（如用户类别标签、使用的浏览器、IP 地址、所在区域、城市等），广告位的相关信息（如广告位的宽度、高度、可见性、所在网站的域名以及 URL 等），投放的广告 ID，以及最终的点击情况（用户点击为 1，无点击为 0）。考虑到 CTR 预测模型是针对每个广告商的，本文采用了其中 *Advertiser ID*＝1458 和 *Advertiser ID*＝3386 的两个广告商的投放和点击日志分别建立了两个数据集。所有实验均采用前 7 天的样本作为训练集，后 3 天的样本作为测试集。表 4 展示了两个数据集中样本的统计情况。观察表 4 发现：真实互联网上广告投放的点击率是非常低的，即数据集中正负样本的比例严重不平衡，会使得模型对正样本的学习不充分，从而大幅降低 CTR 预测模型的精度[30]。因此，本文首先验证负采样对预测性能的提升能力。另一方面，可以发现独热编码后原始特征的数目是非常巨大的，达到了 55 万～56 万量级，将其直接输入 DNN 必然导致巨大的计算开销，经过嵌入向量映射后，输入 DNN 的单元数目将大幅减少，这里假设每个

嵌入向量的维度为 $D=10$。

表 4 两个数据集的统计情况

Table 4 Statistics of two data sets

	Advertiser ID=1458		*Advertiser ID*=3386	
	训练集	测试集	训练集	测试集
样本数	3083056	614638	2847802	545421
点击数	2454	515	2076	445
点击率($*10^{-3}$)	0.7960	0.8379	0.7290	0.8159
特征域数	16	16	16	16
特征数	560802	560802	556884	556884
嵌入层节点数	176	176	176	176

6.2.2 负采样对预测性能的影响

本组实验验证负采样是否能解决点击率预测数据集中普遍存在的正负样本不平衡的问题。该实验基于 *Advertiser ID* =1458 数据集，首先在不进行负采样的情况下测试不同方案的预测性能；然后对训练集按照正负样本比例为 1∶1000 进行负采样。实验结果如表 5 和图 11 所示，统计发现：无论是 AUC 还是 LogLoss 指标，各 CTR 预测方案在负采样后的预测性能都有所提升，除了 FFM 在 LogLoss 指标上略微增加了 1.75%。分析预测性能提升的原因，主要是负采样后，训练集中正负样本的比例更加平衡，使得预测模型对正样本的学习更充分。但是另一方面，负采样会导致训练样本数减少，特别会使训练样本的特征分布与测试集的样本特征分布出现差异，从而使得有些预测模型欠拟合。例如，FFM 由于需要学习特征的两两组合，每个特征对应不同特征域的嵌入向量，负采样后会导致部分组合特征的样本缺失，从而降低了模型的拟合度。

表 5 正负样本比例为 1∶1000 时典型方案的预测性能

Table 5 Predictive performance of typical schemes when the proportion of positive and negative samples is 1∶1000

方案	*Advertiser ID* =1458（无负采样）		*Advertiser ID* =1458（负采样）	
	AUC	*LogLoss*($*10^{-3}$)	*AUC*	*LogLoss*($*10^{-3}$)
LR	0.7017	6.558	0.8283	4.535
FM	0.7038	6.565	0.8308	4.539
FFM	0.8190	5.519	0.8209	5.616
GBDT+LR	0.6914	6.585	0.8138	4.747
FNN	0.7062	6.561	0.8321	4.553
PNN	0.7062	6.547	0.8282	4.509
Wide & Deep	0.6913	6.587	0.8277	4.474
DeepFM	0.6981	6.475	0.8199	4.570
Deep & Cross	0.6999	6.569	0.8261	4.491

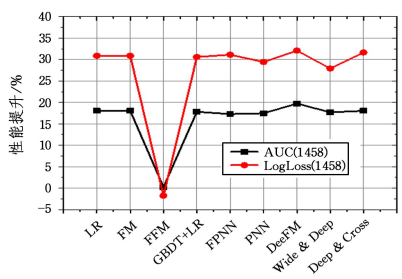

图 11 典型方案在进行负采样后的性能提升

Fig. 11 Performance improvement of typical schemes after negative sampling

进一步，图 12 展示了 LR 在不同负采样比例下预测性能的变化情况。观察发现，在 *Advertiser ID* =1458 训练集中由于正样本非常少，当采样比例达到1∶50时，预测性能大幅下降。因此，对于不同的数据集，应该根据数据集的实际情况，合理设计负采样的比例。

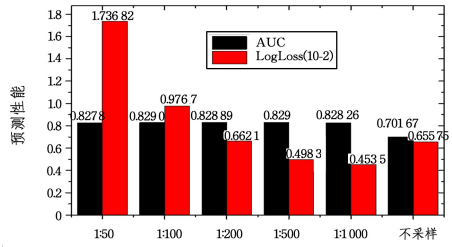

图 12 LR 方案在不同比例负采样后的性能提升

Fig. 12 Performance improvement of LR scheme after negative sampling in different proportions

6.2.3 模型结构对预测性能的影响

本组实验主要测试不同模型结构对预测性能的影响。本实验的训练集是按照正负样本1∶1000的比例进行负采样，测试集不变。首先，基于两个不同的数据集完成浅层模型结构方案的测试，结果如表 6 所列。观察发现：LR 和 FM 的性能优于 GBDT+LR 和 FFM；FFM 表现出的预测性能并不理想，这是因为负采样后的训练集可能导致样本的特征分布发生变化，从而与测试集的样本特征分布存在较大差异，导致预测性能下降。与预想不一致的是，考虑一阶特征和二阶组合特征的 FM 也没有表现出比只考虑一阶特征的 LR 更为优异的性能。分析原因，可能是不同数据集有其自身的特点，特别是对于 *Advertiser ID*=3386 的数据集，增加二阶组合特征后，可能不仅没有提取到更有效的信息，反而增加了一些噪声特征到模型中。

表 6 典型方案的性能对比

Table 6 Performance comparison of typical schemes

评价指标	LR	FM	FFM	GBDT+LR
Advertiser ID =1458 数据集				
AUC	0.8282	0.8308	0.8209	0.8138
LogLoss($*10^{-3}$)	4.535	4.539	5.616	4.747
AUC 提升/%	1.777	2.084	0.8713	—
LogLoss 提升/%	19.249	19.174	—	15.470
Advertiser ID =3386 数据集				
AUC	0.8048	0.8038	0.7839	0.7798
LogLoss($*10^{-3}$)	5.895	5.970	6.957	5.993
AUC 提升/%	3.200	3.071	0.521	—
LogLoss 提升/%	15.256	14.190	—	13.850

其次，实验比较了基于深度学习模型的方案的预测性能，结果如表 7 所列，表中的性能提升都是相对于该指标最差的性能值的，两个数据集中 AUC 指标值最小的是 GBDT+LR，LogLoss 指标值最大的是 FFM。

图 13 和图 14 给出了所有典型方案的性能折线图。观察发现：基于单一 DNN 的方案(FNN 和 PNN)的预测性能并没有比基于浅层结构的 LR 和 FM 模型的方案有显著提升；基于融合结构的 CTR 预测方案在性能上也没有比单一结构的

深度学习预测方案表现出明显的优越性。分析原因，可能是因为深度神经网络需要手动设定的超参数更多，例如每个隐层的神经元数目、每个神经元的激活函数、隐层的数目等；而本文的实验并没有特别地针对性能优化进行调参，而是所有深度神经网络采用统一的结构和激活函数。本文已将所有实验代码及数据集分享到 GitHub 上。

表 7　基于深度学习模型的方案的性能对比

Table 7　Performance comparison of schemes based on deep learning model

评价指标	FNN	PNN	Wide & Deep	DeepFM	Deep & Cross
Advertiser ID=1458 数据集					
AUC	0.8321	0.8282	0.8277	0.8199	0.8261
LogLoss($*10^{-3}$)	4.553	4.508	4.474	4.570	4.491
AUC 提升/%	2.252	1.771	1.707	0.748	1.516
LogLoss 提升/%	18.931	19.712	20.340	18.622	20.033
Advertiser ID=3386 数据集					
AUC	0.8067	0.8078	0.8099	0.803	0.8008
LogLoss($*10^{-3}$)	5.888	5.911	5.931	6.011	5.899
AUC 提升/%	3.446	3.583	3.855	2.963	2.687
LogLoss 提升/%	15.365	15.037	14.748	13.59	15.208

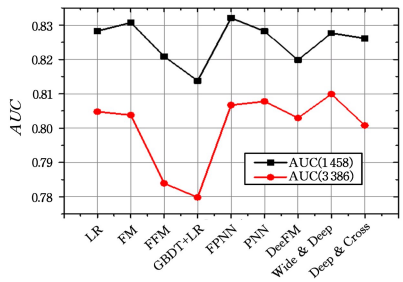

图 13　典型方案的 AUC 指标的对比

Fig. 13　Comparison of AUC metric in typical schemes

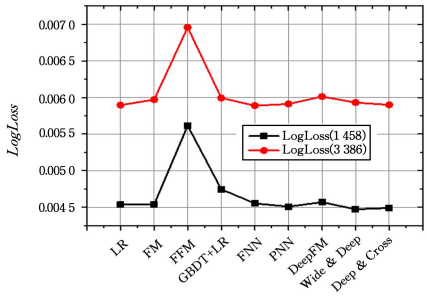

图 14　典型方案的 LogLoss 指标的对比

Fig. 14　Comparison of LogLoss metric in typical schemes

6.2.4　讨论

本节将针对以下几个问题展开讨论：首先，针对训练集中正负样本严重不平衡的问题，合理的负样本采样能够切实提升方案的预测性能，但是负采样会导致训练集的样本特征分布与真实样本特征的分布不一致，从而使某些模型的预测性能下降；其次，实验结果虽然不能说明单一的深度神经网络方案具有比浅层结构方案更优的预测性能，但是在所有典型方案中，考虑了深度学习模型的方案确实表现出了优于只考虑浅层结构的方案的性能，这说明深度学习模型确实能够挖掘高阶组合特征间的有效信息；最后，在本文的实验中，更为复杂的融合结构并没有表现出更为优异的预测性能，这可能与实验没有对超参数进行调优有关。

7　存在的问题及未来的研究方向

尽管 CTR 预测模型的研究已经引起了工业界和学术界的大量关注，但目前 CTR 预测模型的性能在实际广告投放系统中的表现仍然不太理想，还有较大的提升空间。本节对 CTR 预测中存在的难点问题及未来的研究方向进行了归纳。

1）真实场景下，广告定向投放的点击率是非常低的（通常小于 1%），这导致数据集中正负样本的比例极度不平衡。目前通常采用负采样来解决这个问题，但是如本文实验所示，不适当的负采样很可能导致训练集的样本特征分布与真实样本特征分布存在较大差异，从而降低训练的预测模型的性能。一种最新的研究思路是利用生成对抗网络（Generative Adversarial Networks，GAN）来生成正样本，从而达到正负样本比例平衡的目标[32]。

2）冷启动问题：对于新加入的广告，由于其历史信息太少，很难对其点击率做出有效预测。通常使用迁移学习来帮助解决这个问题[33]，即将从类型相似、数据充分的商品上学习到的知识迁移到新商品上，这样有助于预测新商品的点击率。此外，迁移学习的领域适用性也可用于解决实时竞价机制导致的训练数据集的样本“删失”问题[34]；同时，将迁移学习应用到深度学习模型中，也可以减少模型的训练时间与计算开销[35]。将迁移学习应用到点击率预测中，将是未来研究的一个热点方向。

3）模型结构：尽管本文对已有的典型方案的模型结构进行了总结和分析，并给出了一个简单的量化分析，但这还是一个初步的研究，未来还可以进一步针对不同数据集的特点，分析其对应的最优模型结构，以及哪些设计对提升预测性能有重要作用。设计出新的面向点击率预测的模型结构，仍将是未来研究的热点。另一种研究思路是在已有模型结构的基础上进一步完善细节设计。例如，文献[36]提出将 attention 机制引入到深度学习模型中，提出基于 attention 的因子分解机模型；文献[37]提出神经因子分解机模型；文献[38]除了将样本的数值特征与分类特征作为模型的输入以外，还将广告展示的图片信息作为特征，将其加以处理后作为特征输入到模型中，以提升预测精度。文献[31]借鉴“生成对抗”的思想，提出了一种由自编码器、判别器和预测器三部分构成的模型结构，自编码器在与判别器“伪装鉴别”的对抗中，从原始特征中抽象出有效的特征表示，该特征表示作为预测器的输入，进一步提高了模型的预测性能。

结束语　本文针对展示广告定向投放的点击率预测模型展开研究。首先，将点击率预测问题描述为一个最优化问题，并给出了建立点击率预测模型的关键要素；其次，介绍了建立预测模型的数据样本的特征预处理技术，包括独热编码技术以及嵌入向量技术；然后，对目前主要的点击率预测方案进行了详细介绍，包括基于传统机器学习模型的方案和基于深度

学习模型的方案;在此基础上,对目前主要的点击率预测方案的特点和性能进行了分析和验证;最后,对该方向存在的问题及未来的研究方向展开了讨论。本文的主要贡献在于对目前主流的和最新的 CTR 预测模型进行了全面介绍,并对各种典型模型的实现代码进行了归纳整理;对目前已有的实验数据集进行了整理,为相关学者进行后续研究提供了数据基础;对各种典型算法的原理及性能进行了量化分析,在实验基础上对其中的关键技术点展开了讨论,并针对其中存在的问题及可行的解决思路进行了阐述。

参 考 文 献

[1] OLIVIER C. Offline evaluation of response prediction in online advertising auctions[C]// The International Conference of World Wide Web. Florence, Italy, 2015: 18-22.

[2] LIU P, WANG C. Computational advertising: market and technology of Internet business realization [M]. Beijing: The People's Posts and Telecommunications Press, 2015. (in Chinese)
刘鹏,王超. 计算广告:互联网商业变现的市场与技术[M]. 北京:人民邮电出版社,2015.

[3] WANG J, ZHANG W, YUAN S. Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting [J]. Foundations & Trends® in Information Retrieval, 2017, 11(4-5): 297-435.

[4] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012.

[5] 伊恩·古德费洛,约书亚·本吉奥,亚伦·库维尔. 深度学习[M]. 赵申剑,黎彧君,符天凡,等译. 北京:人民邮电出版社,2017.

[6] BECK J E, WOOLF B P. High-level Student Modeling with Machine Learning[M]// Intelligent Tutoring Systems. Berlin, Germany, 2000: 584-593.

[7] ZHANG W, DU T, WANG J. Deep Learning over Multi-field Categorical Data: A Case Study on User Response Prediction[C]// Proceedings of European Conference on Information Retrieval. Switzerland Cham: Springer, 2016: 45-57.

[8] CHENG H T, KOC L, HARMSEN J, et al. Wide & Deep Learning for Recommender Systems[C]// The Workshop on Deep Learning for Recommender Systems. Boston, USA, 2016: 7-10.

[9] GUO H, TANG R, YE Y, et al. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 1725-1731.

[10] WANG R, FU B, FU G, et al. Deep & Cross Network for Ad Click Predictions[C]// Proceedings of AdKDD and TargetAd. Halifax, 2017: 1-7 .

[11] BOTTOU L. Online Learning and Neural Networks[M]. Cambridge, UK: Cambridge University Press, 1998.

[12] ZINKEVICH M. Online Convex Programming and Generalized Infinitesimal Gradient Ascent: Technical Report CMU-CS-03-110[R]. Carnegie Mellon University, 2003.

[13] DUCHI J, SINGER Y. Efficient Online and Batch Learning Using Forward Backward Splitting[J]. Journal of Machine Learning Research, 2009, 10(18): 2899-2934.

[14] XIAO L. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization[J]. Journal of Machine Learning Research, 2010, 11(1): 2543-2596.

[15] MCMAHAN H B, HOLT G, SCULLEY D, et al. Ad click prediction: a view from the trenches[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, 2013: 1222-1230.

[16] CHANG Y W, HSIEH C J, CHANG K W, et al. Training and Testing Low-degree Polynomial Data Mappings via Linear SVM [J]. Journal of Machine Learning Research, 2014, 11(11): 1471-1490.

[17] OENTARYO R J, LIM E P, LOW J W, et al. Predicting response in mobile advertising with hierarchical importance-aware factorization machine[C]// ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2014: 123-132.

[18] RENDLE S. Factorization Machines with libFM [J]. Acm Transactions on Intelligent Systems & Technology, 2012, 3(3): 1-22.

[19] JUAN Y, ZHUANG Y, CHIN W S, et al. Field-aware Factorization Machines for CTR Prediction[C]// ACM Conference on Recommender Systems. Boston MA, USA: ACM, 2016: 43-50.

[20] HE X R, PAN J F, JIN O, et al. Practical lessons from predicting clicks on ads at facebook [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 1-9.

[21] ZHOU Z H. Ensemble Methods: Foundations and Algorithms [M]. New York: CRC press, 2012.

[22] JUAN Y C, CHIN W S, ZHUANG Y. kaggle-2014-criteo[DB/OL]. [2018-07-12]. ttps://github. com/guestwalk/kaggle-2014-criteo.

[23] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2): 1097-1105.

[24] ALEX G, ABDEL-RAHMAN M, GEOFFREY H. Speech recognition with deep recurrent neural networks [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649.

[25] SHEN Y L, HE X D, GAO J F, et al. A latent semantic model with convolutional-pooling structure for information retrieval [C]// ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 2014: 101-110.

[26] 周志华. 机器学习[M]. 北京:清华大学出版社,2016:114.

[27] QU Y R, CAI H, REN K, et al. Product-based neural networks for user response prediction[C]// IEEE International Conference on Data Mining. Barcelona, Spain, 2016: 1-6.

[28] LIAO H R, PENG L X, LIU Z C, et al. Ipinyou global rtb bid-

ding algorithm competition dataset[C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA,2014:1-6.

[29] MURPHY K P. Machine Learning:A Probabilistic Perspective [M]. Boston:MIT,2012.

[30] HE H,GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering,2009,21(9): 1263-1284.

[31] XIE Q Z,DAI Z H,DU Y L,et al. Controllable Invariance through Adversarial Feature Learning[C]// 31st Conference on Neural Information Processing Systems. Long Beach,CA,USA, 2017.

[32] DENG Y,SHEN Y,JIN H,et al. Disguise Adversarial Networks for Click-through Rate Prediction[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne. Australia,2017:1589-1595.

[33] SU Y H,JIN Z M,CHEN Y,et al. Improving Click-Through Rate Prediction Accuracy in Online Advertising by Transfer Learning [C]// Proceedings of WI 17. Leipzig,Germany,2017.

[34] ZHANG W,ZHOU T,WANG J,et al. Bid-aware Gradient Descent for Unbiased Learning with Censored Data in Display Advertising[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA,2016:665-674.

[35] JASON Y,JEFF C,YOSHUA B,et al. How transferable are features in deep neural networks[C]// Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3320-3328.

[36] XIAO J,YE H,HE X N. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017:3119-3125.

[37] HE X G,CHUA T S. Neural Factorization Machines for Sparse Predictive Analytics[C]// The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku,Tokyo,Japan,2017:355-364.

[38] CHEN J,SUN B,LI H,et al. Deep CTR Prediction in Display Advertising[C]// The 2016 ACM Multimedia Conference. Amsterdam,Netherlands,2016:811-820.

[39] CHAPELLE O,MANAVOGLU E,ROSALES R. Simple and Scalable Response Prediction for Display Advertising [J]. ACM Transactions on Intelligent Systems and Technology, 2014, 5(4):1-34.

[40] LEE K C,ORTEN B,DASDAN A,et al. Estimating Conversion Rate in Display Advertising From Past Performance Data [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM press,2012:768-776.

[41] GRAEPEL T,CANDELA Q,BORCHERT T,et al. Web-scale Bayesian Click-throuth Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine [C]// Proceedings of the 27th International Conference on Machine Learning. Israel:Omnipress,2010:13-20.

[42] RICHARDSON M,DOMINOWSKA E,RAGNO R. Predicting Clicks:Estimating the Click-through Rate for New Ads[C]// International Conference on World Wide Web. Canada: ACM, 2007:521-530.