doi: 10.12068/j.issn.1005-3026.2019.07.006

基于混合模型的广告转化率问题研究

李雄飞, 周晋男, 张小利
(吉林大学 计算机科学与技术学院, 吉林 长春 130012)

摘 要: 现有广告转化率预估模型缺乏对深层特征间相互作用的研究,针对这一问题提出了一种新的混合模型. 通过高效的梯度提升机(light gradient boosting machine,LightGBM)模型提取高阶组合特征,并结合基于区域的因子分解机(field-aware factorization machines,FFM)模型有效处理稀疏数据的优点进行转化率的预估. 为了验证模型的有效性和泛化能力,在两个数据集上讨论了参数对预估结果的影响,并将模型与其他模型进行对比实验. 实验结果表明提出的混合模型的预估结果更准确.

关 键 词: 转化率预估;高效的梯度提升树;基于区域的因子分解机;混合模型;高阶组合特征
中图分类号: TP 391 **文献标志码**: A **文章编号**: 1005-3026(2019)07-0942-06

Research on Advertising Conversion Rate Based on Hybrid Model

LI Xiong-fei, *ZHOU Jin-nan*, *ZHANG Xiao-li*
(College of Computer Science and Technology, Jilin University, Changchun 130012, China. Corresponding author: ZHANG Xiao-li, E-mail: zhangxiaoli@jlu.edu.cn)

Abstract: Many existing models of predicting advertising conversion rate lack research on the interaction among deeper features. Hence, a new hybrid model was proposed for this problem. High-level combination features were extracted using a light gradient boosting machine (LightGBM) model, and combining with the advantages of field-aware factorization machines (FFM) model. It can effectively process sparse data and predict the conversion rate. In order to verify the effectiveness and generalization ability of the hybrid model, the model was tested on two data sets for discussing the influence of parameters on model prediction results and was compared with other models. The experimental results show that the hybrid model is more accurate.

Key words: conversion rate prediction; light gradient boosting decision tree; field-aware factorization machines; hybrid model; high-level combination features

随着平板电脑和智能手机的广泛使用,在APP中内置广告逐渐成为了移动数字营销的重要形式. 现有的许多广告平台在推荐广告时其目标是最大化点击次数,而移动APP广告期望用户进行的动作不仅仅是点击而是下载APP,这时广告的转化率就成为广告投放平台制定投放策略时需考虑的重要指标.

实际上,由于被转化的广告数据很少,所以相比于对广告点击率的预估,转化率的预估变得更加困难. 现有的对广告点击率(click through rate, CTR)和转化率(conversion rate, CVR)的预估模型可以分为两类:一类是线性模型,另一类是非线性模型. 在线性模型中,例如逻辑回归[1](logistic regression, LR),具有简单、可解释性强和伸缩性好等优点,已被Google,Facebook和Yahoo广泛使用. 近年来,因子分解机(factorization machine, FM)[2-3]、梯度增强决策树(gradient boosting decision tree, GBDT)、基于GBDT的扩展结构

收稿日期: 2018-05-04
基金项目: 国家自然科学基金资助项目(61801190); 吉林省自然科学基金资助项目(20180101055JC); 吉林省优秀青年人才基金资助项目(20180520029JH); 中国博士后科学基金资助项目(2017M611323).
作者简介: 李雄飞(1963-),男,吉林省吉林市人,吉林大学教授,博士生导师.

(XGBoost)和深度神经网络(DNN)[4]也被应用于工业上进行广告点击率和转化率预测. 由 Criteo 举办的 Kaggle 广告点击率预估比赛,获得优胜的模型 FFM[5]是在 FM 基础上增加了域的概念. 由于单一模型存在不同的优缺点,不能达到最佳的预估效果,所以通常将不同模型进行组合来提高预估的准确性. Google 的推荐系统采用 Wide & Deep 模型[6],这种混合模型结合了 LR 宽和 DNN 深的特性,提高了预估准确性;Facebook 上广告的点击率预估使用 GBDT 和 LR 混合模型,通过 GBDT 进行非线性特征转换,然后提供给 LR 进行最终预测,混合之后的模型比单一模型的预估准确度提升了 3% 左右. 模型 FNN[7]则将通过 FM 学习的低阶组合特征嵌入到 DNN 中. 这些混合模型偏向于分析高维组合特征,有些虽然同时考虑了低阶特征间的相互关系,但难以达到平衡. 本文通过分析数据构造特征,将 LightGBM[8]模型提取的高阶组合特征与低阶离散型特征作为 FFM 的输入进行广告转化率预估,使 FFM 在预估时不局限于二阶组合特征,改善了 FFM 的预估性能.

1　解决方案

本文针对现有广告转化率预测中的困难和模型的不足,提出了新的解决方案,其总体流程如图 1 所示,包括:数据筛选、数据划分、特征工程、建立模型.

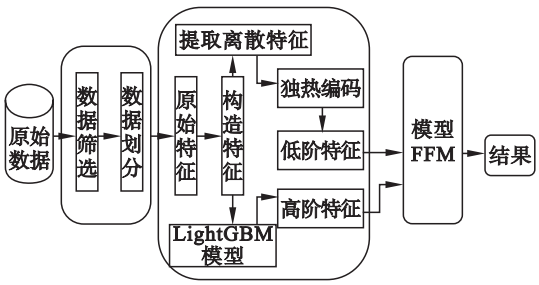

图 1　解决方案的总流程

Fig. 1　Overall flow chart of the solution scheme

由于在广告系统中存在转化回流时间,所以最后几天数据的转化标记可能不够准确,但是直接删除这些数据可能会破坏数据的完整性. 通过分析转化时间和应用程序(application,APP)之间的关系,对数据进行筛选来保留数据的真实性和完整性. 在划分训练和测试数据时,本文通过实验确定了合适的数据窗口大小,并且分析了数据新鲜度对预估结果的影响.

本文利用 LightGBM 模型天然的树型结构对特征进行组合来提取更高阶的组合特征,并且同时将低阶和高阶特征作为 FFM 模型的输入,得到了比较好的预估结果.

2　特征提取

2.1　低维特征提取

本文使用的数据集来自腾讯(Tencent)在 2017 年 4 月举办的“社交广告算法大赛”中的广告日志信息,出于对数据安全的考虑,数据中一些原始字段被进行了加密处理. 其中包括广告信息、用户信息和上下文信息. 用户信息描述了用户的属性和行为,例如用户 ID、年龄、性别、学历、兴趣爱好、历史 APP 安装列表等;广告信息描述了一条广告素材的信息,包括广告 ID、素材 ID、APP 分类、APP 平台等;上下文信息描述了广告所在环境的信息,包括广告位 ID、站点集合 ID、广告位类型、联网方式、运营商等. 本文将从原始数据信息中提取的低维特征分为以下几类,其中 ID 类特征和交叉特征是离散型特征,时间相关特征和统计特征是连续型特征.

1) ID 类特征:用户 ID、广告 ID、广告位 ID 和 APP 的 ID 等.

2) 交叉特征:考虑到用户和广告、广告和 APP、用户和 APP 之间存在密切的绑定关系,所以增加了用户 – 广告、广告 – APP 和用户 – APP 交叉特征.

3) 时间相关的特征:分别统计用户、用户 – 广告特征和用户 – APP 特征在一段时间内的点击次数和安装次数.

4) 统计特征:分别统计 ID 类特征、交叉类特征的历史转化率和点击率.

2.2　高维特征提取

本文利用 LightGBM 模型来提取高维组合特征. LightGBM 是一种集成学习模型,通过不断迭代训练多棵决策树以最小化损失函数. 已知广告数据都是高度稀疏的,也会出现数据缺失的情况,而 LightGBM 模型不但可以处理稀疏和缺失数据,同时还支持离散型数据的直接输入,提高了数据处理的效率. 提取高维特征的具体步骤如下:

步骤 1　将低维特征作为 LightGBM 模型的输入,并设置参数(迭代次数、损失函数、采样率 a,b 等),初始化模型.

步骤 2　将每个特征进行分桶并归一化.

步骤 3　计算特征的直方图.

步骤 4　从直方图中获得分裂增益,在树的叶子节点中选择最佳的分裂特征和特征值.

步骤 5　根据选择的最佳分裂特征和特征值将样本分割到左右子树中.

步骤 6　重复步骤 3 ~ 步骤 5 直到达到限制的叶子数或者限制深度.

步骤 7　将新生成的树放入模型.

步骤 8　根据生成模型预测每条样本,并且计算梯度值.

步骤 9　将数据按照梯度绝对值递增顺序排序.

步骤 10　从数据中取前 $a \times 100\%$ 有较大增益值的样本,并且在剩余的数据中随机采样 $b \times 100\%$ 个样本.

步骤 11　在采样的具有较小增益的样本中,将其增益值乘以常数 $(1-a)/b$,可以保证不改变数据分布.

步骤 12　将步骤 9,步骤 10 得到的数据作为新的输入数据,重复步骤 2 ~ 步骤 12,直到迭代结束.

在步骤 2 ~ 步骤 4 中采用直方图方法进行分裂特征和特征值的选择,在步骤 3 中计算子节点的直方图时可以通过将父节点和兄弟节点的直方图做差快速得到. 步骤 8 ~ 步骤 11 具体实现可参照文献[8].

3　算法框架

FFM 能够处理高度稀疏的数据,并且能够分析二阶特征的组合关系,通过将特征分域,使每个特征对应不同域中不同的隐向量,在广告点击率和转化率的预估中有很高的准确率. 本文通过将提取高阶特征的 LightGBM 模型与 FFM 集成,弥补 FFM 在高阶组合特征分析上的不足,从而提高模型对广告转化率预估的准确性. 模型集成过程如图 2 所示.

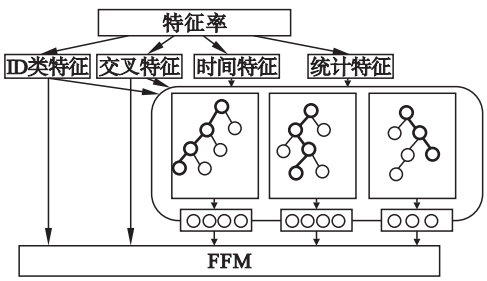

图 2　基于 FFM 的混合模型的框架

Fig. 2　Framework of FFM-based hybrid model

混合模型的算法框架如下:

1) 利用 LightGBM 提取高维组合特征(即图 2 中的树型结构模型,若某样本遍历 LightGBM 后到达叶节点为加粗的路径,那么路径中的特征即为从该样本提取的组合特征,组合特征用叶节点的索引值表示).

2) 模型初始化:将 LightGBM 提取高阶特征(每棵树叶节点的索引值)和低阶离散型特征同时作为输入数据,并设置参数.

3) 将输入数据归一化,防止溢出.

4) 对训练数据中每个样本做以下操作,样本被形式化表示为 $(\boldsymbol{x}_i, y_i)$. 其中:$\boldsymbol{x}_i = (x_{j_1}, x_{j_2}, \cdots, x_{j_m})$ 表示样本,$x_{j_1}, x_{j_2}, \cdots, x_{j_m}$ 表示样本的特征,一个样本由 m 维特征组成;y_i 是二值变量,取值为 0 或 1,表示真实的转化结果. 计算每个样本的 φ 值,公式为

$$\varphi(\boldsymbol{w}, \boldsymbol{x}_i) = \sum_{j_1=1}^{m} \sum_{j_1=j_2+1}^{m} (\boldsymbol{w}_{j_1,f_2} \cdot \boldsymbol{w}_{j_2,f_1}) x_{j_1} x_{j_2}. \quad (1)$$

其中,$\boldsymbol{w}_{j,f}$ 是某个特征对某个域的一个隐向量. 根据损失函数计算该样本的损失值,损失函数的公式为

$$L = \lg\{1 + \exp[-y_i \varphi(\boldsymbol{w}, \boldsymbol{x}_i)]\} + \frac{\lambda}{2} \| \boldsymbol{w} \|^2. \quad (2)$$

计算该样本的梯度值,梯度公式为

$$g_\varphi = -y_i \cdot \frac{\exp\{-y_i \varphi(\boldsymbol{w}, \boldsymbol{x}_i)\}}{1 + \exp\{-y_i \varphi(\boldsymbol{w}, \boldsymbol{x}_i)\}}. \quad (3)$$

用梯度值和初始的学习率更新 FFM 模型中的隐向量,公式为

$$\boldsymbol{w}'_{j,f} = \boldsymbol{w}_{j,f} - \frac{\eta}{\sqrt{1 + \sum_t (g^t_{w_{j,f}})^2}} \cdot g_\varphi. \quad (4)$$

其中:g_φ 是本次迭代所求的梯度值;$g^t_{w_{j,f}}$ 是之前第 t 次迭代的梯度值.

5) 同理,计算测试样本的损失值.

6) 重复步骤 3) 和 4) 直到迭代结束或者测试样本的损失值不再减少.

4　实　　验

4.1　实验准备

腾讯提供的数据集是 2015 年 9 月 17 日至 30 日的日志信息,为了模拟在线的训练和测试数据流,实验选择 17 日 ~ 28 日的数据作为训练数据,30 日的数据作为测试数据. 数据信息可被形式化表示为 $D = \{(\boldsymbol{x}_i, y_i)\}$ ($|D| = n, \boldsymbol{x}_i = \mathbf{R}^m, y_i = 0$ 或者 1),其中:$\boldsymbol{x}_i$ 表示一个样本,i 是样本 ID;n 是样本数;m 是特征维数;y_i 表示广告是否被转化,假

如用户对一条广告进行了转化操作，则 $y_i=1$；否则广告未被转化，则 $y_i=0$. 实验选取在广告点击率和转化率预估中，将逻辑损失（logistic loss，LL）作为评估指标，其公式如下：

$$\mathrm{LL}=-\frac{1}{N}\sum_{i=1}^{N}(y_i\lg(p_i)+(1-y_i)\lg(1-p_i)). \tag{5}$$

其中：N 是样本总数；y_i 是二值变量，取值 0 或 1，表示第 i 个样本的真实结果；p_i 是第 i 个样本的预估结果. 模型的 LL 值越低，模型预估的准确性越高，实验结果如表 1 所示. 从表中可以观察到，FFM 作为单一模型使用时比 LightGBM 的 LL 值低一些，而经过 LightGBM 提取高阶组合特征后的 LL 是最低的，也就是预估准确率最高.

表 1　混合模型和单一模型的逻辑损失值比较

Table 1　Comparison of logistic loss values of the hybrid model and the single model

模型	逻辑损失(LL)/%
FFM	10.197
LightGBM	10.203
LightGBM + FFM	9.87

4.2　数据新鲜度

在广告转化率预估系统中，数据会随着时间的变化而变化，那么数据的新鲜度就会影响模型对转化率预估的准确性. 为了探索训练和测试天数的延迟对模型预估性能的影响，实验将 17 日 ~ 28 日的数据每相邻 3 d 作为 1 组，一共分为 5 组训练数据，均以 30 日数据作为测试数据，分别在 LightGBM + FFM 和 FFM 上进行了实验，实验结果如表 2 所示.

表 2　延迟天数对模型逻辑损失值的影响

Table 2　Impact of the delay days on logistic loss values of the models　　%

延迟天数	FFM	LightGBM + FFM
5	10.48	10.35
4	10.38	10.20
3	10.27	10.19
2	10.20	10.12
1	10.19	9.97

从表中可以观察到，两个模型的逻辑损失值都会随着延迟天数的增加而变大，而一天的延迟会使逻辑损失增加 0.1% 左右，所以在广告预估系统中，需要新鲜的训练数据来保证模型的预估性能.

4.3　数据窗口

在选择数据训练模型时还需要考虑数据的大小，本文称之为数据窗口大小（data sizes）. 考虑到用户在下载 APP 之后可能过很长时间才会重新启动 APP，或者用户启动 APP 的行为需要广告商上报回广告系统的过程发生了延时. 基于这两种情况，如果将数据窗口设置得过小，则本来已经被转化的广告很可能因为广告商没有及时将 APP 激活数据上报给广告系统，也就是说这条广告的 label 本应被标注为 1 而实际却是 0. 所以数据窗口的负样本比真实要多，造成了预估的不准确.

相反地，数据窗口过大不仅会在等待数据过程时占用存储空间而造成浪费，还会由于数据新鲜度的下降使模型的预测性能变低. 因此，需要在数据的新鲜度和窗口大小之间达到一个平衡. 本文通过实验比较了不同数据窗口大小对预估性能的影响，实验结果如图 3 所示.

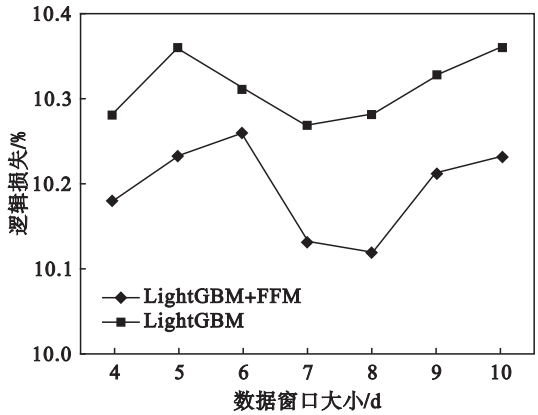

图 3　数据窗口大小对模型预估性能的影响

Fig. 3　Impact of the data sizes on prediction accuracy

从图 3 中可以看出，大体上数据窗口大小为 78 时，两个模型的逻辑损失值最小. 当数据窗口大小从 8 开始继续增大时，两个模型的逻辑损失值开始变大. 这是由于在数据窗口较大时，继续增大窗口反而会使窗口中产生许多不必要的数据信息，因此会影响预估模型的性能，则逻辑损失值会变大. 而在数据窗口较小时，通过增大数据窗口使原本因延迟而导致窗口中负样本比实际多的情况得以解决，所以逻辑损失值会减小.

5　优化参数

当不做特殊说明时，本节中所有实验的训练数据均选用 7 d 的数据量.

5.1　LightGBM 模型中参数的讨论

5.1.1　迭代次数的选择

通常模型训练时迭代次数越多预估的精度越

高,将模型迭代训练 200 次,限制每棵树的叶子总数不超过 64 个,观察逻辑损失值随迭代次数的变化. 实验结果如图 4 所示,从图中可以看出,当 LightGBM 模型的学习率为 0.05 和 0.1 时,逻辑损失值随着迭代次数的增加而减小. 并且迭代次数在 60 之前逻辑损失值下降比较明显,在 60 之后下降缓慢. 当学习率为 0.2 时,逻辑损失值先随着迭代次数的增加而降低,在迭代次数超过 120 后逻辑损失值逐渐上升,说明出现了过拟合.

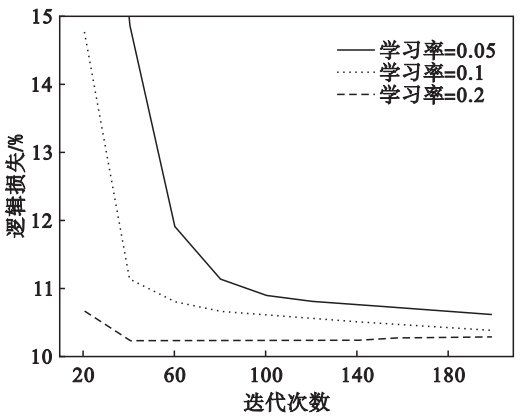

图 4 LightGBM 模型取不同学习率时逻辑损失值随迭代次数的变化

Fig. 4 Change of logistic loss values with iterations when LightGBM model takes different learning-rates

5.1.2 XGBoost 和 LightGBM 时间的比较

LightGBM 和 XGBoost[9] 的生长方式不同,LightGBM 只要在限制的深度内达到设置的叶子总数就可以停止生长,从而能够快速收敛. 如表 3 所示,记录了 XGBoost 和 LightGBM 取不同迭代次数时训练模型所用的时间.

表 3 不同迭代次数时模型所用时间的对比

Table 3 Comparison of the time at different iteration s

迭代次数	XGBoost	LightGBM
20	38	1
40	75	2
80	149	3
160	294	5
200	366	6

从表 3 中可以观察到,相同的迭代次数,XGBoost 所用的时间是 LightGBM 的几倍,并且随着迭代次数的增加倍数也在增长.

5.2 FFM 中参数的讨论

5.2.1 提前停止(early stopping)

在很多机器学习问题中,为了避免过拟合的出现,会采用提前停止的方法(early stopping)[10]. 在 FFM 中用到类似 early stopping 的策略:将数据分成训练集和验证集,在模型的每次迭代训练后用验证集计算损失函数值. 如果损失函数值上升则记录迭代次数,也可以按照记录的迭代次数用全部数据重新训练模型. 如果损失函数值下降则继续迭代训练直到损失函数值上升.

从图 5 可以观察到,FFM 在经过 5 次迭代训练后逻辑损失达到最低值,从第 5 次迭代之后随着迭代次数的增加逻辑损失值逐渐变大. 随着训练次数的增加模型变得越来越复杂,模型在训练数据集和验证集上的损失值也逐渐减小. 但当模型达到一定的复杂度时,模型在验证集上的误差反而会随之增大,也就是出现了过拟合. 所以 FFM 采用提早停止的方法来避免过拟合,当然有些时候,对于验证集最佳的迭代次数却不一定适合测试集,这里不做过多讨论.

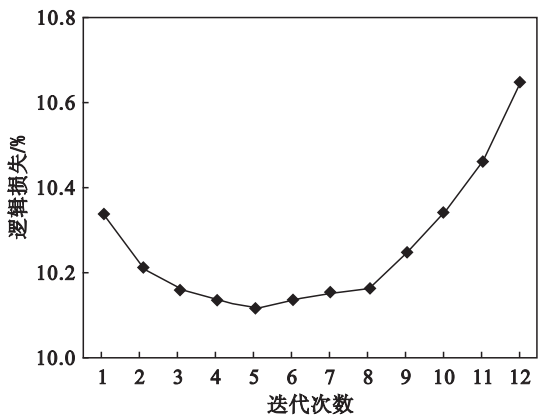

图 5 不同迭代次数对 FFM 模型预估性能的影响

Fig. 5 Impact of different iterations on prediction performance of FFM model

5.2.2 线程的选择

考虑到 FFM 模型在训练时选择多线程可能会出现不同的收敛情况. 本文通过实验讨论了多线程时得到的收敛结果,为了方便比较速度的提升情况,将速度提升比定义为单线程所用时间和多线程所用时间的比.

从图 6 中可以观察到,最开始增加线程数能够使模型的收敛速度有明显的提升,但当线程数增加到 12 之后收敛速度趋于平稳. 在使用了 22 个线程后收敛速度稍有降低,这是因为当多个线程同时访问内存空间时,一个线程必须等待另一个线程访问结束后才可以继续执行,当越来越多的线程同时执行时会加剧冲突的发生.

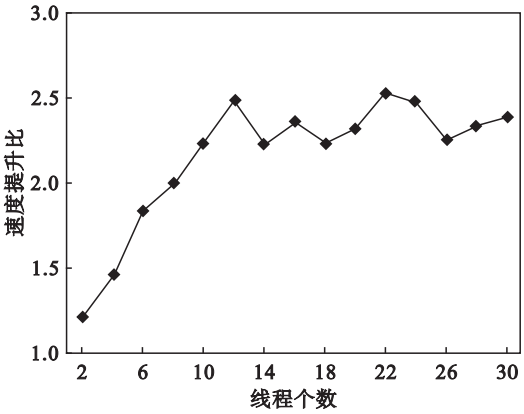

图 6　多个线程时的速度提升比

Fig. 6　Speedup ratio of using multi-threading

5.3　更多模型的比较

为了进一步验证本文提出的混合模型的有效性和泛化性，将 LightGBM + FFM 在 Criteo 数据集和 Tencent 数据集上与更多模型进行了比较. 实验结果如表 4 所示，可以观察到经过 LightGBM 提取高阶组合特征的 FFM 在两个数据集上都有最低的逻辑损失值.

表 4　更多模型预估结果(逻辑损失)的比较

Table 4　Comparison of more model predictions (logistic loss)　%

模型	Tencent 数据集	Criteo 数据集
GBDT + LR	10. 563	45. 281
XGBoost	10. 199	44. 351
XGBoost + FFM	10. 183	44. 258
LightGBM	10. 203	44. 562
FFM	10. 197	44. 107
LightGBM + FFM	9. 87	43. 715

6　结　　语

本文提出了一种新的混合模型用于广告转化率的预估，突破了 FMM 作为单一模型预估时只考虑二阶组合特征的局限性. 在保留低阶特征的前提下，通过 LightGBM 进行了高阶组合特征的研究，改善了 FFM 的预估性能. 并从数据和模型参数两个角度进行实验分析，优化了模型. 为了验证混合模型的有效性和泛化性，在不同数据集上与其他模型进行了比较，实验结果表明本文提出的混合模型的逻辑损失值最小，预估结果更准确.

参考文献：

[1] 董书超. 基于逻辑回归模型的广告点击率预估系统的设计与实现[D]. 哈尔滨：哈尔滨工业大学，2016.
(Dong Shu-chao. Design and implementation of click through rating system based on logistic regression model [D]. Harbin：Harbin Institute of Technology，2016.)

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

[3] Juan Y，Lefortier D，Chapelle O. Field-aware factorization machines in a real-world online advertising system[C]//Proceedings of the 26th International Conference on World Wide Web Companion. Perth，2017：680 – 688.

[4] Ling X L，Deng W W，Gu C，et al. Model ensemble for click prediction in bing search Ads[C]//Proceedings of the 26th International Conference on World Wide Web Companion. Perth，2017：689 – 698.

[5] Pan J，Xu J，Ruiz A L，et al. Field-weighted factorization machines for click-through rate prediction in display advertising[C]//Proceedings of the 2018 World Wide Web Conference on World Wide Web. Lyon，2018：1349 – 1357.

[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. Boston，2016：7 – 10.

[7] Zhang W，Du T，Wang J. Deep learning over multi-field categorical data[C]//European Conference on Information Retrieval. Padua，2016：45 – 57.

[8] Ke G，Meng Q，Finley T，et al. LightGBM：a highly efficient gradient boosting decision tree[C]//Advances in Neural Information Processing Systems. Barcelona，2017：3149 – 3157.

[9] Chen T，Guestrin C. XGBoost：a scalable tree boosting system[C]//Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco，2016：785 – 794.

[10] Raskutti G，Wainwright M J，Yu B. Early stopping and non-parametric regression：an optimal data-dependent stopping rule[J]. *Journal of Machine Learning Research*，2014，15(1)：335 – 366.