文章编号：2095-2163(2020)05-0067-05 中图分类号：TP391 文献标志码：A

基于轻量梯度提升机的广告转化率预估方法

刘恩伯，赵玲玲，苏小红
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：互联网广告转化率对于搜索引擎服务供应商和广告商都是一个重要的量化指标，大数据平台下互联网广告转化率预估的实现具有很强的理论研究价值和实际应用价值。由于互联网广告的转化是一个大量数据下的小概率事件，因此为了提高广告转化率预估，提出了轻量梯度提升机的预估方法。通过对大规模广告转化日志的分析，提取数据特征并构造数据集，成功应用轻量梯度提升机算法实现广告转化率的预估。实验结果表明，与传统工业界常用的机器学习方法相比，在相同的特征处理和数据集下，轻量梯度提升机的预估结果优于其他方法。

关键词：轻量梯度提升机；大数据；广告转化率；机器学习

Light GBM-based method for Internet advertising conversion rate prediction

LIU Enbo, ZHAO Lingling, SU Xiaohong
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Internet advertising conversion rate is an important quantitative indicator for search engine service providers and advertisers, the realization of Internet advertising conversion rate prediction under the big data platform has strong theoretical research value and practical application value. Since the conversion of internet advertising is a small probability event under a large amount of data, therefore, in order to increase the advertising conversion rate prediction, a LightGBM-based method is proposed. Through the analysis of large-scale advertising conversion logs, we extracted data features and constructed data sets, then applied LightGBM algorithm to achieve advertising conversion rate prediction successfully. The experimental results show that compared with the traditional machine learning methods in industry, LightGBM has better prediction results than other methods under the same features extraction and data sets.

【Key words】 LightGBM; big data; advertising conversion rate; machine learning

0 引 言

计算广告[1-4]是在给定具体的网页内容和用户下，通过大量的计算，将最佳广告内容匹配给用户的一种精准化广告投放机制。随着网络技术的快速发展，广告投放平台不断迭代，广告投放的形式加速改变，如今已经拥有庞大的市场体系，成为互联网重要的商业模式。广告投放效果的好坏通过转化、曝光和点击等指标来权衡，因此广告转化率预估是计算广告领域的关键问题。通过对用户的行为动作、兴趣爱好等的分析获得对用户的特征抽象，通过对特征的分析建模对特定的用户推送不同的个性化广告内容。不但可以提高广告的投放效果 ，使广告主的收益更大，将广告从无用的骚扰内容变身成有价值的用户感兴趣的内容，从而为用户的工作和生活等方面带来了极大的便利。

广告转换率估算方法依赖于用户、发布商和广告客户数据层次结构中的过去性能观察值[5]。更具体地说，是利用单独的二项分布在不同的选择层次上对转换事件建模并估计分布参数，使用逻辑回归将这些单独的估计量结合起来以准确识别转换事件。Amr Ahmed, Abhimanyu Das 等人提出了一个分层模型和可伸缩算法来执行多任务学习的推理。在联合稀疏设置中推断任务关联和子任务结构，通过一个分布式次梯度预言器以及与变量组和子组相关的 prox-operators 的连续应用来实现[6]。并将此算法应用于展示广告中的转化问题上，精度和准度得到了很大的提升。随着移动广告的增长，使得预测广告响应的任务对于最大化业务收入至关重要。由于广告的响应数据受限于历史记录的冷启动，阻碍预测的可靠性。为此，Richard J. Oentaryo, Ee-Peng Lim 等人开发了一个分层重要性意识因子分解机器（HIFM），它提供了一个有效的通用潜在因素框架，其中包含重要性权重和分级学习[7]。实证研究表明，HIFM 优于当前时间潜在因素模型，冷启动情景下的整体预测效果得到改善。Weinan Zhang, Tianming Du 等人提出了两种使用深度神经

作者简介：刘恩伯（1993-），男，硕士，主要研究方向：信息融合、机器学习；赵玲玲（1980-），女，博士，讲师，主要研究方向：智能信息处理与信息融合、图像处理等；苏小红（1966-），女，博士后，教授，博士生导师，主要研究方向：图像处理、任务规划、人工智能等。

收稿日期：2020-03-06

网络(DNN)的新模型,以自动学习来自类别特征的有效模式,并预测用户的广告转化率[8]。解决了用户响应预测模型必须将自身限制为线性模型或者需要手动构建高阶组合功能。Hongxia Yang,Quan Lu 等人提出一种新的概率生成模型,通过将自然语言处理,动态转移学习和可伸缩预测的组件紧密集成来预测转化率[9]。过度预测和过度出价是实时出价平台中的基本挑战。为了解决这个问题,Quan Lu,Shengjun Pan 等人[10]提出了一个安全的预测框架,其中包含转换分配调整以处理过度预测,并进一步缓解不同级别的过度出价[10]。

本课题采用 Tencent 公开的移动社交应用广告数据,预估广告点击后被激活的概率,即在给定广告信息、用户信息和上下文情况等外需信息和广告日志的情况下,预估广告被点击并发生转化的概率。

目前工业界常用的方法有很多,比如广点通精排使用的 LR 模型,Yahoo 和 Bing 使用传统的 GBDT 模型,Facebook 使用 GBDT+LR 的组合模型,百度凤巢采用 FM 模型。这些模型针对不同的应用场景和不同的广告数据效果不尽相同,各有自己的有缺点,将传统模型进行试验,并用实验结果与我们的方法进行效果对比。

轻量梯度提升机(Light Gradient Boosting Machine, LightGBM[11]),是一种基于 Gradient Boosting[12-13]的集成学习算法。传统的 Boosting 算法包括:AdaBoost, RankBoost, GBDT 等。由于广告转化日志特征维度很高且数据量庞大,数据稀疏性高,传统的 GBDT 不能满足搞得效率和可扩展,LightGBM 算法是传统梯度提升模型的改进,算法性能有了极大的提升,非常适合广告转化日志的属性特征。因此,本文将 LightGBM 算法应用于互联网广告转化率预估中,挖掘用户的行为和广告等有用信息,建立回归预估模型,调整优化参数从而得到理想的预估概率。实验结果表明,LightGBM 与传统的机器学习方法相比,具有精确度高、运行速度快、内存消耗低和可扩展等特点。

1 数据与模型

1.1 数据描述

原始数据集包括训练集和测试集,以及广告特征,用户特征,上下文特征等 8 个数据文件。训练数据和测试数据每行代表一个样本,各字段之间由逗号分隔,顺序依次为:"instanceID,label,clickTime,creativeID, userID, positionID, connectionType, telecomsOperator",其中,instanceID 唯一标识一个样本。数据集为广告系统中随机抽取某半个月的转化日志,并遵照运营中的应用 App 和用户特征维度进行随机采样。每一条训练样本代表一条广告转化日志记录,样本标签 label 取值 1 代表该条广告被点击并发生了转化,0 表示没有发生转化。

广告特征包括账户——推广计划——广告——素材四级结构。不同的账户对应不同的特定广告主;推广计划包含多个不同的广告,是广告的一个集合。广告主可以把预算额度情况、是否匀速投放、计划推广平台等条目一致的广告整理到同一个推广计划中,便于管理;广告是指广告商设计的广告素材或创意以及展示等相关的设置;素材是直接展现给用户的广告内容,同一条广告可以包含多个广告素材。广告特征还包含有各类 App 的相关特征。用户特征包括用户的基本特征:年龄、性别、学历、婚恋状态、育儿状态、家乡和籍贯、常住地等,还包括用户安装 App 流水等。上下文特征包括广告曝光的具体位置;移动工具等的上网模式如 4G、Wifi 等;移动工具的运营商如联通、电信等;多个广告位的聚合以及对于某些站点人工定义的一套广告位规格分类等。

基于对转化日志回流时间的分析,发现几乎 100%的回流时间发生在三天内,其次是两天之内,占到了 90%。因此,考虑到硬件因素,选择 28,29 两天为训练集,30 天为测试集。

1.2 特征提取

数据特征的提取,主要包括以下四部分特征。

(1)基础特征。基础特征即原数据集合包含的已知特征,如 age、creativeID、adID、positionType、appInstallList 等。用这些基础特征训练出来的模型,已经具备指导转化率预估的能力。即让分类器学习这些基本属性对于是否转化的分布,完成最基本浅层的预估。

(2)用户的统计特征。一条广告是否发生转化,主要取决于用户,因此提取与用户相关的特征属性是保证预估准度和精度的关键。用户的统计特征主要包括两部分:基本统计和时序统计。基本统计包括:用户的转化类别、转化次数、安装数量、点击数量、安装同类别 App 的数量等;时序统计包括:统计点击时间之前的 App 安装数量、种类、用户点击量等。

(3)Trick。由于网速延时、带宽等外部因素影响,用户在短期内可能不断重复的点击同一条广告,挖掘这些连续不断的点击日志的信息是很有价值的。例如,对短期内连续的多条重复记录进行编号,

记录当前点击分别与前一次点击和后一次点击的时间差，统计相同时段内的点击量等。

(4)贝叶斯平滑后的转化率。在某些特殊条件下，如统计同一广告位下某 App 的历史转化率，由于广告位上线时间有延时，往往上线慢的广告统计不充分，特征对其基本无影响，因此用户历史转化率并不能相对准确的表示该条件下的真实转化率。又如大多数用户只点击过某个 App 一次，历史转化率就可能会达到 1，使用这些记录训练模型即使用标签来训练模型，极大的影响训练结果。所以，对某些特定属性计算该属性下的贝叶斯平滑后的转化率。

按照上述提取方案一步步生成最终的数据集合，整体生成流程如图 1 所示。component 1 - n 是生成的中间数据集。

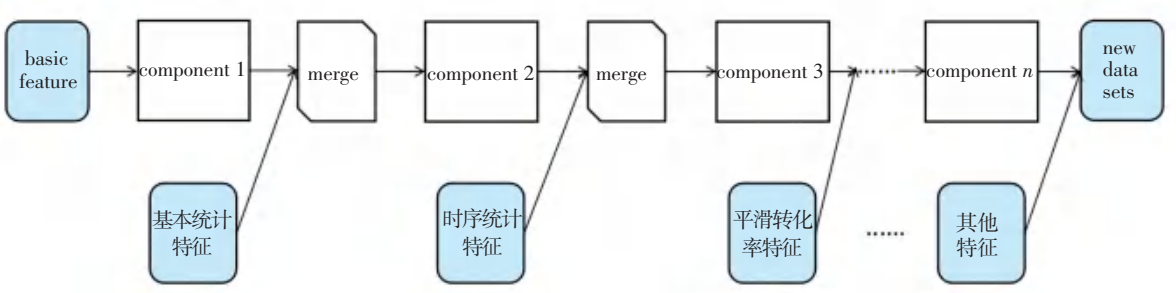

图 1　数据集生成流程图

Fig. 1　Data set generation flow chart

1.3　特征重要性分析

本文对生成的数据集，使用 XGBoost 工具对所选取特征进行重要性排序，从而可以判断出哪些特征与广告转化与否关系较大，这是实验的特征选取的关键一步。分析结果如图 2 所示。

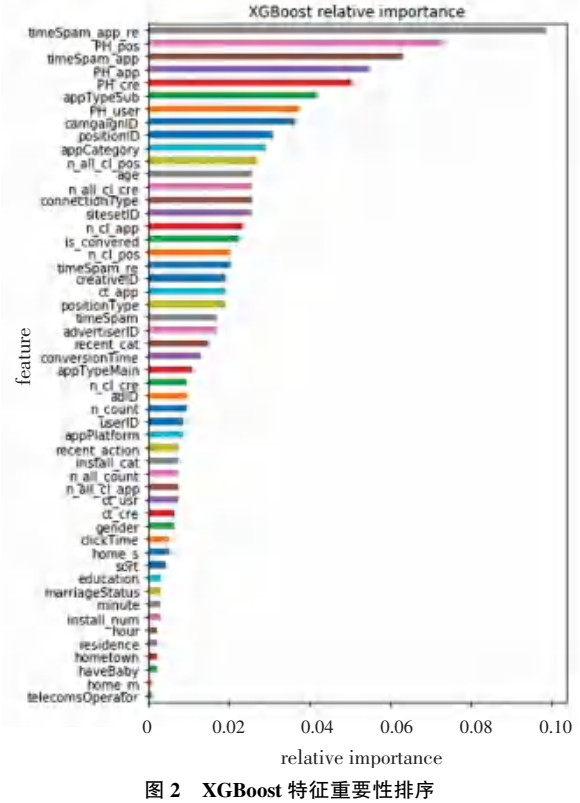

图 2　XGBoost 特征重要性排序

Fig. 2　Importance of feature variables by XGBoost

由图 2 可知，短期内重复点击的广告最可能会被转化，贝叶斯平滑对转化率的影响也极为关键，一些基础特征对转化率的指导也很重要。利用相关性特征，通过皮尔森相关系数对相关性排序低的特征进行筛选。

1.4　轻量梯度提升机

LightGBM 是基于 GBDT 的梯度提升算法。在此基础上 LightGBM 提出两种新方法：Gradient - based One - Side Sampling (GOSS) 和 Exclusive Feature Bundling (EFB)。

针对数量大，GOSS 保留所有梯度较大的实例，在梯度小的实例上使用随机采样。为了抵消对数据分布的影响，计算信息增益的时候，GOSS 对小梯度的数据引入常量乘数。GOSS 首先根据数据的梯度绝对值排序，选取 Top A 个实例，然后在余下的数据里通过随机采样 B 个，接着计算信息增益时为采样出的小梯度数据乘以(1-A)/B，这样算法就会更关注训练不足的实例，而不会过多改变原数据集的分布。所以 LightGBM 采用了基于 Leaf-wise 的决策树算法，这是一种按叶子生长并带有深度限制的生长策略。而大多数梯度提升模型使用 Level-wise 的决策树算法，这是一种按层生长的生长策略，如图 3 所示。Leaf-wise 是一种作用更好的生长策略，它每次从当前所有叶子节结点中，找到分裂增益最大的一个叶子结点进行分裂，如此循环。因此与 Level - wise 生长策略相比，在分裂次数等条件相同的情况下，Leaf-wise 可以得到更好的结果，训练速度更快。

针对特征维度高，高维的数据通常是稀疏的。特别的，稀疏特征空间中，许多特征是互斥的，例如

他们从不同时为非零值。EFB 算法能够将许多互斥的特征变为低维稠密的特征,能够有效的避免不必要零值特征的计算,能够极大地加速 GBDT 的训练过程而且损失精度。实际上,使用直方图算法,用表格来标记非零元素来忽略零值特征。通过对表格中的数据的扫描,建立直方图的时间复杂度将从 O(#data)降到 O(#non_zero_data)。从内存消耗上看,直方图算法只需(#data * #features * 1Bytes)的内存,在寻找分割点时,直方图算法的时间复杂度代价是 O(#feature * #data),而在数据分割时,直方图算法时间复杂度的代价只有 O(#data)。在计算上,分割结点次数得到很大的降低;在数据并行时,通信代价得到极大的降低。

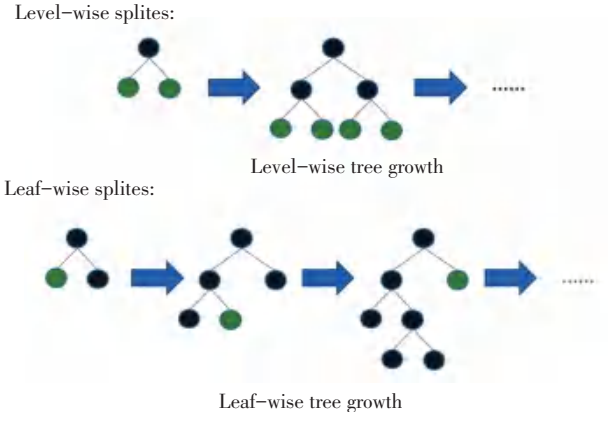

图 3 Level-wise 和 Leaf-wise 生长策略

Fig. 3 Level-wise and Leaf-wise growth strategies

LightGBM 还直接支持类别特征,不需要进行独热编码操作,从而极大地降低了数据维度。此外,Cache 命中率、网络通信和并行计算上都有一定程度的优化,且支持 GPU 加速。

1.5 参数选取

LightGBM 参数组成主要分为调节训练速度的参数,调节精度的参数,防止过拟合的参数三部分。在给定其他参数默认值情况下,分别使用网格搜索进行最佳参数选择,其中重要的参数设置如:bagging_fraction+ bagging _ freq 同时设置来提高 bagging 的速度,控制树决策树复杂度的参数 num_leaves 设置为 355,此时并未选择 max_depth 来防止过拟合,提高训练精度的参数学习率 learning_rate 设置为 0.02,提高速度的参数 feature_fraction 设置为 0.5 等。

2 实验结果与结论分析

2.1 评价标准

由于广告转化率数据具有数据量大且稀疏的特性,为了体现预测结果与真实值的吻合程度,评价指标采用对数损失(Logloss),公式(1)如下:

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i\log(p_i)) + (1-y_i)\log(1-p_i)). \quad (1)$$

其中,N 为测试样本总数,y_i 是二值变量,取值 0 或 1,表示第 i 个样本的 label,p_i 为模型预测第 i 个样本 label 为 1 的概率。

2.2 结果比较

使用 LR、GBDT、GBDT+LR、FM、FFM 分别对相同特征工程处理后的数据集进行建模,其中 FFM 模型是在 FM 模型的基础上进行改进,主要区别在于 FM 模型中,每一个特征会对应一个隐变量,而在 FFM 模型中,将不同类特征分为多个域,每个特征对应每个域分别对应一个隐变量。通过参数优化和交叉验证,对第 30 天的转化率进行预估。并计算得到各自的 Logloss 值与本文的模型 LightGBM 进行对比,得到的实验结果如图 4 所示。

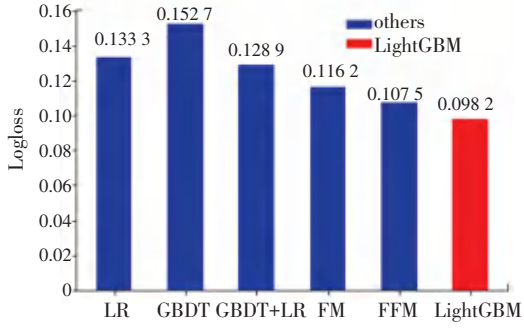

图 4 实验结果对比图

Fig. 4 Comparison of experimental results

实验结果表明,在所有实验模型中,通过对 Logloss 值的对比,发现 LightGBM 预估结果在准度和精度上都要好于其他五种模型。在实验过程中,LightGBM 的内存占用率与其他模型相比最低,CPU 利用率仅为 0.47,而运算速度仅次于 LR,但明显高于其他几种模型。

3 结束语

本文采用轻量梯度提升机算法,基于腾讯社交广告日志,对其数据进行特征选择构造和回归算法建模,并与 LR、GBDT、GBDT+LR、FM、FFM 算法进行对比,得到了更加精准的转化率预估结果。对于日后互联网广告转化率的提高具有重要的现实和指导意义。

参考文献

[1] BRODER A Z. Computational advertising and recommender systems [C]//Proceedings of the 2008 ACM conference on Recommender systems. 2008: 1-2.

(下转第 75 页)