Hans Journal of Data Mining 数据挖掘, 2019, 9(2), 60-67
Published Online April 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.92008

Hans汉斯

Research on Click-Through Rate Prediction in Display Advertising Based on Machine Learning

Zhiyue Zhang, Hao Huang

College of Information, University of International Business and Economics, Beijing
Email: zzyzzaii@163.com

Received: Apr. 5th, 2019; accepted: Apr. 18th, 2019; published: Apr. 25th, 2019

Abstract

Display ads are an important part of online advertising. Predicting clicks before a display ad can not only reduce the cost of ad serving but also increase the efficiency of Internet companies' resource utilization and increase revenue. As big data and machine learning technologies continue to mature, more and more companies are using technology to predict ad click-through rates. This paper studies the display ad click rate prediction problem from two aspects: feature importance and model suitability. Firstly, the article finds that the advertising characteristics are most important for the advertisement click rate prediction problem by comparing the characteristics of advertisement, user and context, and media characteristics. At the same time, adding media features and user context features can also improve the model effect. Secondly, this paper compares the advantages and disadvantages of the machine model commonly used in advertising click rate estimation, mainly from the two dimensions of model performance and model time consumption. This paper finds that the logistic regression model, the random forest model, and the gradient lifting decision tree model are the most suitable machine learning models for solving the problem of advertising click rate prediction.

Keywords

Display Advertising, Computing Advertising, Click-Through Rate Estimation, Machine Learning

基于机器学习的展示广告点击率预测研究

张芝悦，黄　浩

对外经济贸易大学信息学院，北京
Email: zzyzzaii@163.com

收稿日期：2019年4月5日；录用日期：2019年4月18日；发布日期：2019年4月25日

文章引用：张芝悦，黄浩．基于机器学习的展示广告点击率预测研究[J]．数据挖掘, 2019, 9(2): 60-67.
DOI: 10.12677/hjdm.2019.92008

摘　要

展示广告是网络广告的重要组成部分。在展示广告投放前对其点击情况进行预测不仅能够减少广告投放的成本也能够提高互联网公司的资源利用效率从而增加收入。随着大数据以及机器学习技术不断成熟，越来越多的公司采用相关技术预测广告点击率。本文从特征重要性以及模型适合性两个方面研究展示广告点击率预测问题。首先，文章通过对比广告特征、用户及上下文特征、媒体特征三大类特征发现广告特征对于广告点击率预测问题最为重要，同时加入媒体特征以及用户上下文特征也能够提升模型效果。其次，本文对比研究了常用于广告点击率预估的机器模型优劣，主要从模型性能以及模型耗时两个维度进行比较。本文发现逻辑回归模型、随机森林模型、梯度提升决策树模型是最适合解决广告点击率预测问题的机器学习模型。

关键词

展示广告，计算广告，点击率预估，机器学习

1. 引言

据统计[1]，2018 年中国网络广告市场规模达到 3750.1 亿元，在互联网核心企业中，网络广告收入占总体的 60%以上，网络广告仍是互联网产业的核心商业模式。以图片、文字等形式定向推送给用户的网络广告即为展示广告，它是网络广告中一个非常重要的组成部分。展示广告投放的主要目标是精准营销，即在一定预算成本下，将广告投放者的收益提升至最大化以提高收益节约成本，其中一个方法就是最大化展示广告的点击次数。因此，这就需要不断提升广告被点击的次数。在大数据时代，借助机器学习模型对广告日志数据进行分析以预测广告的点击情况不仅能够达到精准营销的目的，也能够优化广告投放分布，最终提升广告收入。

本文主要从广告日志特征选择、机器学习模型分析两个方面研究广告点击率问题。通过多个实验分析，本文得到预测广告点击率最佳的机器学习模型以及广告特征、上下文及用户特征、媒体特征对模型的不同影响。

2. 相关工作介绍

传统的机器学习模型在广告点击率预测任务上应用广泛。传统的机器学习方法主要分为单一模型预测以及模型组合预测两大类。在单一模型中，逻辑回归、决策树等是较为常见的模型。Richardson 等[2]使用广告特征、关键字以及用户特征与逻辑回归模型结合预测广告点击率，并分析了不同关键字对广告点击情况的影响。Dupret [3]等基于充足的历史广告点击数据应用决策树等模型进行广告点击率预测。此外，还有其他的机器学习模型如：支持向量机模型[4]等应用于广告点击率预测并都取得了不错的效果。在模型组合方面，Facebook [5]公司研究人员将梯度提升决策树与逻辑回归结合，将经过梯度提升树模型转化后的特征组合作为逻辑回归模型的输入也得到了不错的效果。Yahoo [6]公司研究人员将集成学习应用于广告点击率预测，通过改变各个分类器的权重以关注分类错误样本以提升预测精度。这些方法都取

得了不错的效果，但传统的机器学习模型更依赖人工对特征的处理，在应用模型前期需要大量的人工特征工程。

近年来，随着深度学习的兴起，越来越多的研究者将深度学习应用于广告点击率预测中。百度公司[7]研究人员将因子分解机于深度神经网络相结合用于预估广告点击率的 FNN 模型。谷歌公司[8]研究人员推出 Wide&Deep 模型以处理点击率预估问题。在 Wide&Deep 的 Embedding 层后加入特征交叉功能形成的 PNN 网络[9]也取得了不错的效果。此外，还有一些将深度学习与注意力机制结合的模型。AFM [10]模型将引入注意力机制削减无效特征，提升特征的有效性从而提升模型效果。近期，阿里巴巴公司研究人员提出的深度兴趣网络[11]也使用注意力机制提升有效信息权重并结合卷积神经网络的感受野特性进行广告点击率模型的设计。

3. 实验过程

3.1. 实验数据及描述

本实验所采用的数据源于实际广告点击日志数据以保证实验结果的准确性与严谨性。实验数据共105,491 条，其中共有原始特征 24 个，click 字段则代表广告是否被用户点击，是实验中要预测的字段。实验中广告点击日志部分原始数据如表 1 所示：

Table 1. Advertisement click log data
表 1. 广告点击日志数据

字段名称	字段
样本 id	instance_id
广告是否点击	click
广告 id	adid
广告主 id	advert_id
广告主名称	advert_name
广告主行业	advert_industry_inner
……	
点击时间戳	time
联网类型	nnt
操作系统	os
设备类型	devtype
手机品牌	make

3.2. 数据预处理

数据预处理包含正负样本均衡、数据清洗、特征处理三个部分。

正负样本均衡：在原始数据集中，负样本与正样本之比约为 10:1，若直接按照原始正负样本比例进行训练会影响模型准确率评估以及模型效果。基于此本文采取保留原数据集中的所有正样本(即 click 字段值为 1 的样本)，对负样本进行欠采样，以保证正负样本比例均衡。

数据清洗：此部分包含缺失数据的填充、重复数据的删除。本文的缺失数据处理主要有两种类型：删除 click 字段缺失的样本，以字段属性为依据选择不同类型的值填充，如使用众数填充 make 特征字段的缺失值。

特征处理：本文对广告日志中各特征进行相应提取与转化以适应模型训练要求。主要的处理方法为分桶以及独热编码、标签编码。本文对时间戳类型的字段 time 进行转化并提取时间段 hour_seg 特征，即将 0~7、8~12、13~17、17~24 分为四个不同的时间段。针对本数据集，本文分别使用独热编码以及标签编码对类别型特征(如手机品牌、省份、城市等)进行处理并使用逻辑回归模型分别验证效果，如表 2 所示。

Table 2. Label coding and one-hot coding effect comparison
表 2. 标签编码与独热编码效果对比

编码方式	机器学习模型	AUC 值	准确率
独热编码	逻辑回归	0.6920	0.6931
标签编码	逻辑回归	0.6957	0.6970

从表 2 中通过对比可看出，针对本数据集两种处理方法效果相当。为节约计算资源，本文使用标签编码进行特征处理。

此外，本文将广告主行业进行切分为广告一级行业、广告二级行业两个特征以更详细地描述广告信息。最后，本文将数据集中取值只有一个值的特征删除，处理后结果如表 3 所示。

Table 3. Post-processing feature list
表 3. 处理后特征列表

字段名称	字段
广告是否点击	click
广告 id	adid
广告主 id	advert_id
广告主名称	advert_name
广告一级行业	advert_industry_inner_1
广告二级行业	advert_industry_inner_2
活动 id	campaign_id
创意 id	creative_id
创意类型	creative_type
是否落地页跳转	creative_is_jump
是否为 js 素材	creative_is_js
创意宽	creative_width
创意高	creative_height
app 分类	app_cate_id
媒体 id	app_id
媒体广告位	inner_slot_id
请求来源地	city_new
点击时间段	hour_seg
运行商类型	carrier
点击时间戳	time
联网类型	nnt
操作系统	os
设备类型	devtype
手机品牌	make

3.3. 广告点击率预估模型

3.3.1. 单模型广告点击率预估算法

1) 逻辑回归模型

逻辑回归模型通过在传统的线性模型中加入 sigmoid 函数从而实现非线性变换，将模型结果值压缩至从 0 到 1 的区间，并通过不断优化损失函数训练模型，其模型公式如(1)、(2)所示。其中 $X_1, X_2, \cdots, X_n$ 是每个样本的 n 个特征，X 为权重 $W_0, W_1, \cdots, W_n$ 与 $X_1, X_2, \cdots, X_n$ 线性组合得到的结果[12]，之后将 X 用 sigmoid 函数处理得到最终的 $f(X)$值：

$$X = W_0 + W_1 X_1 + \cdots + W_n X_n \tag{1}$$

$$f(x) = \frac{1}{1+e^{-X}} \tag{2}$$

当结果越接近 1 时代表结果为正例的可能性越大，反之，当结果越接近 0 时则代表结果为反例的可能性越大。逻辑回归模型是工业界最常用的预测展示广告点击率的模型，在多模型对比实验中，通常选择逻辑回归模型作为基线模型进行比较。广告点击率预估问题中，输入 X 代表广告日志样本，其包含日志中各字段，输出 Y 代表该样本被模型预测为正例的概率即广告点击率。$f(X)$代表模型习得的函数。当 Y 的值大于 0.5 则将此广告分类为正类即会被点击，当 Y 的值小于 0.5 则将此广告分类为负类即不会被用户点击。应用逻辑回归进行广告点击率预估可表示为如式(3)所示[9]：

$$p(C|X) = \frac{1}{1+e^{-f(X)}} \tag{3}$$

其中 $p(C|X)$ 为在给定展示广告日志数据情况下，用户点击广告的概率即为输出 Y 的值。$p(C|X)$ 的值与分界线 0.5 进行比较即可判断并预测此条广告的点击情况获得分类结果。

2) 决策树模型

作为一种基本且高效的二分类模型，决策树以样本的特征为根据对样本进行划分识别，它在广告点击率预测中也有良好的效果。决策树的学习过程通常是根据一定的标准递归地选择最优广告日志特征，然后依据该特征对广告日志数据集进行分割直至所有子数据集都有能够被明确的归类为点击与未点击两大类别[13]。

在本文的广告点击率预测二分类问题中采用的是 CART 决策树，CART 决策树又称为分类回归树，它是一种典型的二叉树结构。在分类过程中，CART 决策树选择具有最小基尼指数的属性及其属性值作为最优分裂属性以及最优分裂属性值。基尼指数越小，则代表分类后的样本纯净度越高。对于给定的样本集合，基尼指数的计算如式(4)所示：

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{k} \left(\frac{|C_K|}{|D|} \right)^2 \tag{4}$$

其中 C_K 是 D 中属于第 k 类的样本，K 是类别的个数。如果样本集合 D 根据特征 A 是否取某一个可能的值 a 被分割成 D_1 和 D_2 两部分，则根据特征 A，集合 D 的基尼指数计算如式(5)所示：

$$\mathrm{Gini}(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{5}$$

CART 决策树通过上述方式不断选择最优划分特征从而得到一颗分类决策树，从而对样本进行分类与预测。

3.3.2. 集成学习模型广告点击率预估算法

集成学习通过建立和组合多个弱学习器完成学习任务，根据集成融合方法不同，它主要可分为两大类：自助聚集法和提升法。由于随机森林是自助聚集法的扩展且在工业界及学术界应用广泛，故本文主要应用随机森林模型进行训练。同时，本文选取提升法中的两大代表性模型自适应增强模型、梯度提升决策树模型进行训练。

1) 自助聚集法(Bagging)

随机森林模型：以决策树作为弱学习器并在并在自助聚集类集成方法的基础上通过随机选择特征子集进行划分得到随机森林模型(Random Forest)。在对多个弱学习结果进行投票后，随机森林模型才产生最终的模型结果。在广告点击率预测问题中，通过不同决策树对样本进行分类得到样本的点击与未点击情况，再将所有决策树的分类结果进行投票得出样本最终分类情况。

2) 提升法(Boosting)

自适应增强模型：自适应增强模型(Adaboost)针对给定广告日志数据训练不同的弱学习器，并通过弱学习器融合产生强学习器。在更新过程中，前一次分类不正确的样本权重将被不断提高从而使得模型逐渐修正预测结果，并继续依此训练其他的弱学习器。

梯度提升决策树模型：梯度提升决策树模型(GBDT)以决策树为弱学习器，每轮迭代从给定数据集中进行有放回抽样形成本轮训练数据，通过多次迭代产生多个弱学习器。模型最终结果由弱学习器通过投票决定。GBDT 模型被认为是机器学习中最有效的方法之一，其在各种比赛以及工业应用中都取得了良好的效果。

3.4. 模型评价标准

本文主要采用精度(accuracy)以及受试者工作特征曲线下面积(AUC)作为主要的模型评价指标。

精度(accuracy)：精度是分类任务中最常用的性能度量，它指的是分类正确的样本数与总样本数的比重。

当正负样本不均衡时，精度并不是一个很好的度量指标。例如，若样本中点击与未点击用户的比例为 1:10，模型将所有用户都分类为未点击，模型的精度仍能达到 90%，然而这并不符合实际。本文已经在数据预处理阶段进行正负样本均衡的处理消除了正负样本不均衡对精度的影响，故本文将采用精度作为模型评价标准之一。

受试者工作特征曲线下面积(AUC)：AUC 是另外一个应用广泛的二分类器评价标准。“真正例率”指在全部正样本中，分类无误的正样本所占的比重，“假正例率”指在全部负样本中，分类错误的负样本所占的比重。以“真正例率”为竖直方向，“假正例率”为水平方向构建坐标系形成的图像为受试者工作特征曲线(ROC)，对 ROC 曲线求取其线下面积即为 AUC。

AUC 值大于等于 0.5 且小于等于 1，当 AUC 为 0.5 则代表与随机猜想效果相同。因此，AUC 值通常应大于 0.5。在这种情况下，AUC 的值越大，则说明模型效果越好。

4. 实验结果

4.1. 不同特征组合对广告点击率预测的影响

选择合适的特征对于模型预测结果有重要影响。本文将广告点击日志中的特征分为用户及上下文特征、广告特征、媒体平台特征并依次分析不同特征下模型的效果。

本文首先将广告特征应用于训练模型，再依次加入用户特征与媒体特征。实验模型选取工业界最常用的逻辑回归模型。实验结果如表 4 所示，结果保留四位有效数字。

Table 4. AUC and accuracy with different feature combinations
表 4. 不同特征组合下模型的 AUC 值与准确率

特征组合	LR 模型 AUC 值	LR 模型准确率
广告特征	0.6931	0.6920
广告特征 + 用户特征	0.6932	0.6936
广告特征 + 用户特征 + 媒体特征	0.6957	0.6970

由表 4 结果可知，单独使用广告特征时模型能够取得不错的效果，AUC 值以及准确率能够接近 0.7。当加入用户特征与媒体特征后模型性能有所提升。加入用户特征后模型性能提升较小，可能由于本文所拥有的用户信息较少且都为边缘用户信息从而不能对用户进行更进一步刻画。若能够收集更加丰富的用户信息，则模型的预测效果会进一步提升。

4.2. 不同机器学习模型对比分析

本文对比了不同机器学习模型应用于广告点击率预测时的性能，如表 5 所示：

Table 5. Comparative analysis of different machine learning models
表 5. 不同机器学习模型对比分析

模型类别	机器学习模型	AUC 值	准确率
单模型	逻辑回归	0.6957	0.6970
	决策树	0.6509	0.6512
集成学习模型	随机森林	0.6971	0.6958
	自适应增强	0.6942	0.6930
	梯度提升决策树	0.7501	0.7001

由表 5 可知，五个模型性能由低到高依次为：决策树、逻辑回归、自适应增强模型、随机森林、梯度提升决策树。单模型中逻辑回归效果最好，集成学习模型中梯度提升决策树模型效果最好。

作为单模型代表之一的逻辑回归模型预测效果较好，准确率以及 AUC 值在所有模型中排第三位。集成学习中自助聚集法的代表随机森林模型在所有模型中排第二位，且其性能相比单决策树模型有较大提升。与单决策树相比，随机森林模型具有更好的性能，但与梯度提升决策树相比，随机森林模型性能仍有差距。提升法的两个代表自适应增强模型以及梯度提升决策树模型均有较好的性能，但自适应增强模型相比于梯度提升决策树模型以及随机森林模型仍有劣势。梯度提升决策树模型是所有模型中性能最好的，其准确率可达到 0.7001、AUC 值能达到 0.7501，远远超过其他所有模型。

五种机器学习模型在运行耗时上也有差别，将各个模型在训练集的总耗时以相对于逻辑回归模型耗时的比值展示，结果如表 6 所示。

Table 6. Different machine learning models consume time in test set
表 6. 不同机器学习模型在测试集耗时

机器学习模型	训练耗时(分钟)	相对耗时
逻辑回归	0.0225	1
决策树	0.0272	1.01
随机森林	0.0236	1.05
Adaboost	0.1458	6.48
梯度提升决策树	0.0302	1.34

由表 6 可看出，自适应增强模型耗时超出所有模型，若应用此模型进行广告点击率预测可能出现线上模型延时的问题。其他四个模型随着模型性能提升其耗时也有所增加。

综上所述，在这五个模型中，自适应增强模型训练耗时过长可能会产生延时现象对预测造成误差。提升法的代表梯度提升树模型具有最好的性能，相比于传统的自适应增强模型训练时间大大缩短且具有较高的精度。自助聚集法的代表随机森林模型相比于单颗决策树有较大性能提升。虽然逻辑回归模型较为简单，但在广告点击率预测问题中仍有较好的效果，且其可解释性及适应性较强。

5. 结语

本文从两个方面研究了展示广告点击率预测的问题。第一点，选择合适的特征对于广告点击率预测有重要影响。广告点击日志数据中有许多特征，主要是广告特征、用户及上下文特征、媒体特征三大类。这三类特征利用的越充分、组合的越好则模型的效果越好。因此，在实际业务中，应充分挖掘这三类信息以提高模型效果。

第二点，针对于广告点击率预测任务，可首先考虑逻辑回归、随机森林、梯度提升决策树三个模型，并依据实际环境条件，再从中选出合适的模型。

单模型广告点击率预测算法中逻辑回归模型最好，集成学习模型广告点击率预测算法中随机森林模型、梯度提升决策树模型较好。集成学习中的两个模型效果也都比逻辑回归效果好。但随着模型性能的提升其耗时也有所增加。逻辑回归虽然性能稍逊于随机森林模型以及梯度提升树模型，但其耗时最短且可解释性较强，也是一个不错的模型。

基金项目

国家重点研发计划资助(National Key R&D Program of China)，项目编号：2017YFB1400700。

参考文献

[1] 艾瑞咨询. 2018 年中国网络广告市场年度监测报告－简版. http://report.iresearch.cn/report/201808/3264.shtml

[2] Richardson, M., Dominowska, E. and Ragno, R. (2007) Predicting Clicks: Estimating the Click-Through Rate for New Ads. *International Conference on Word Wide Web*, *ACM*, 521-530. https://doi.org/10.1145/1242572.1242643

[3] 肖垚, 毕军芳, 韩易, 董启文. 在线广告中点击率预测研究[J]. 华东师范大学学报(自然科学版), 2017(5): 80-86+100.

[4] Dave, K. and Varma, V. (2010) Predicting the Click-Through Rate for Rare/New Ads. Center for Search and Information Extraction Lab International Institute of Information Technology, Hyderabad.

[5] He, X., Pan, J., Jin, O., *et al*. (2014) Practical Lessons from Predicting Clicks on Ads at Facebook. *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, *ACM*, 1-9.

[6] Bagherjeiran, A., Hatch, A., Ratnaparkhi, A., *et al*. (2010) Large-Scale Customized Models for Advertisers. *IEEE International Conference on. Data Mining Workshops* (*ICDMW*), 1029-1036. https://doi.org/10.1109/icdmw.2010.157

[7] Zhang, W., Du, T. and Wang, J. (2016) Deep Learning over Multi-Field Categorical Data. *European Conference on Information Retrieval*, Springer, Cham, 45-57.

[8] Cheng, H.T., Koc, L., Harmsen, J., *et al*. (2016) Wide & Deep Learning for Recommender Systems. *Proceedings of the* 1*st Workshop on Deep Learning for Recommender Systems*, *ACM*, 7-10. https://doi.org/10.1145/2988450.2988454

[9] Qu, Y., Cai, H., Ren, K., *et al*. (2016) Product-Based Neural Networks for User Response Prediction. *IEEE* 16*th International Conference on Data Mining* (*ICDM*), 1149-1154. https://doi.org/10.1109/icdm.2016.0151

[10] Xiao, J., Ye, H., He, X., *et al*. (2017) Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. arXiv:1708.04617. https://doi.org/10.24963/ijcai.2017/435

[11] Zhou, G., Song, C., Zhu, X., *et al*. (2017) Deep Interest Network for Click-Through Rate Prediction. arXiv:1706.06978.

[12] 施梦圜, 顾津吉. 基于平衡采样的轻量级广告点击率预估方法[J]. 计算机应用研究, 2014, 31(1): 33-36+39.

[13] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 55-60.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2163-145X，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjdm@hanspub.org