第 45 卷 第 1 期 计 算 机 工 程 2019 年 1 月
Vol. 45 No. 1 Computer Engineering January 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)01-0178-08 文献标志码：A 中图分类号：TP181

基于多源融合特征提取的在线广告预测模型

刘 冶[1,2,3]，刘 荻[1,2]，王砚文[3,4]，傅自豪[2,3]，印 鉴[1,2]

（1. 中山大学 数据科学与计算机学院，广州 510006；2. 广东省大数据分析与处理重点实验室，广州 510006；
3. 火烈鸟网络（广州）股份有限公司 数据中心，广州 510630；4. 香港理工大学 电子计算学系，中国 香港 999077）

摘 要： 针对智能移动终端应用平台上的广告点击率（CTR）预测问题，在传统 PC 端 Web 平台在线广告 CTR 预测方法的基础上，提出一个新的智能移动终端在线广告投放业务架构。基于此架构，构建基于机器学习的在线广告预测模型，对用户基本信息、广告内容、用户使用环境等多源特征进行融合提取，实现在线广告 CTR 的精确预测。结合移动 APP 应用环境的特点，将用户历史行为数据加入预测模型进一步提高 CTR 预测性能。实验结果表明，该模型具有较高的 CTR 预测准确率。

关键词： 计算广告；广告点击率；特征选择；机器学习；预测模型

中文引用格式： 刘冶，刘荻，王砚文，等. 基于多源融合特征提取的在线广告预测模型[J]. 计算机工程，2019，45(1)：178-185，191.

英文引用格式： LIU Ye，LIU Di，WANG Yanwen，et al. Online advertising prediction model based on multiple source fusion feature extraction[J]. Computer Engineering，2019，45(1)：178-185，191.

Online Advertising Prediction Model Based on Multiple Source Fusion Feature Extraction

LIU Ye[1,2,3]，LIU Di[1,2]，WANG Yanwen[3,4]，FU Zihao[2,3]，YIN Jian[1,2]

（1. School of Data and Computer Science，Sun Yat-sen University，Guangzhou 510006，China；
2. Guangdong Provincial Key Laboratory of Big Data Analysis and Processing，Guangzhou 510006，China；
3. Data Center，Flamingo Network Co.，Ltd.，Guangzhou 510630，China；
4. Department of Computing，The Hong Kong Polytechnic University，Hong Kong 999077，China）

【Abstract】 Aiming at the problem of advertising Click Through Rate（CTR）prediction on intelligent mobile devices application platform，this paper proposes a novel online advertising business architecture for intelligent mobile devices based on the traditional CTR prediction method on PC Web platform. With this architecture，an online advertising prediction model based on machine learning is designed to integrate and extract the multiple source features such as user information，advertising content and user usage environment，so as to achieve accurate prediction of online advertising CTR. Combined with the characteristics of the mobile APP application environment，the CTR prediction performance is improved by adding the user's historical behavior data into the prediction model. Experimental results show that this model has a high accuracy rate of CTR prediction.

【Key words】 computational advertising；advertising Click Through Rate（CTR）；feature selection；machine learning；prediction model

DOI： 10.19678/j.issn.1000-3428.0049207

0 概述

近年来，互联网已成为人们生活中的重要部分，在线广告也日益成为互联网经济的一个主要组成部分。随着技术的进步，在线广告的投放逐渐向精准化的方向演进[1-2]。在线广告的精准投放就是对投放的环境和给定的用户进行分析，通过不同算法来选择与给定用户最匹配的广告，并进行定向投放[3-4]。

计算广告的核心任务是在特定环境下为特定用户选择最合适的广告展示。点击率（Click Through

基金项目： 广东省科技计划项目（2012A010701013）；广州市科技计划项目（2013J4500059）；广州市天河区科技计划项目（201601YG152，201701YG127）；广东省大数据分析与处理重点实验室开放基金（2017017，201805）。

作者简介： 刘 冶（1989—），男，博士研究生，主研方向为机器学习、神经网络、网络挖掘；刘 荻，硕士；王砚文，博士研究生；傅自豪，硕士；印 鉴，教授、博士、博士生导师。

收稿日期： 2017-11-07 **修回日期：** 2018-01-08 **E-mail：** jourkliu@163.com

Rate,CTR)即广告展示后被点击的概率,一方面广告被点击与否可以有效地说明当前展示的广告是否符合用户的兴趣;另一方面,点击是后续广告其他操作的前提,而对后续操作的预测一般比较困难,因此准确预测广告的点击率成为计算广告的一个核心任务。

在线广告的投放因投放平台的不同分为 PC 端的在线投放和移动端的在线投放。PC 端的互联网计算广告可以分为付费搜索和内容匹配[5]。由于移动设备方便携带以及越来越多功能的应用(Application,APP)的出现,移动设备已成为人们生活和工作的重要工具。当前,与 PC 端较成熟的计算广告研究相比,移动端广告个性化精准推荐领域的研究还有较多的问题尚未解决。

移动端广告的个性化精准推荐主要通过广告平台实现,广告平台为智能移动终端开发者提供了广告服务,通过自身的优化策略和推送算法对广告业务进行聚合和精准推送,提升广告投放的收益。在移动应用分发平台上,火烈鸟网络的果盘游戏平台是一个典型的移动应用分发聚合平台,本文通过在果盘游戏平台的应用对移动 APP 广告的 CTR 预测问题进行研究。在传统计算广告学方法的基础上,结合移动 APP 应用环境的特点,提出一个新的智能移动终端在线广告投放业务架构,构建多源融合的特征处理方法,并选用合适的预测模型实现 CTR 的精确预测。

1　相关工作

计算广告的核心问题是为一系列用户与环境的组合找到最合适的广告投放策略,以提高整体广告活动的利润[2],形式化描述为:

$$\text{Maximize}\sum_{i=1}^{T}(r(a_i,u_i,c_i)-q(a_i,u_i,c_i)) \quad (1)$$

其中,r 代表收益,q 代表成本,a、u、c 分别代表广告、用户和环境变量,即对应广告活动的 3 个参与主体,T 为广告所展示的次数。对于大多数广告而言,成本相对稳定,可设为常数。因此,成本部分可以从优化表达式中省略,优化目标函数为:

$$\text{Maximize}\sum_{i=1}^{T}(r(a_i,u_i,c_i)) \quad (2)$$

通常,选取合适的广告进行投放,往往考虑的是该广告在一段时间内的整体收益表现,但是由于每次广告展示的效果无法事先获取,因此通常将整体的收入分解为每次展示的收入之和,在每次展示中对广告展示结果进行预测,选择预测收益最大的广告进行展示。

当前普遍采用基于机器学习的方法对广告 CTR 进行预测[6-7]。利用用户与广告的交互数据,提取用户、展示环境以及广告特征进行训练,构建分类或回归模型,当系统收到广告展示请求时,利用训练后的模型预测当前候选集中的广告是否会被用户点击以及被点击的概率。

对于 CTR 预测模型,文献[8]对搜索引擎广告的 CTR 预测问题进行了深入研究,通过收集日志数据得到的上下文以及广告的信息用于构建特征,最后建立逻辑回归模型对新广告的 CTR 进行预测。深度神经网络由于强大的拟合能力,越来越多的研究人员将其用于广告特征提取。文献[9]通过设计一种深度卷积神经网络结构提取图像广告特征,有效地捕捉了图像广告的特点,利用提取出的特征对广告进行推荐。在线广告数据的另一个特点是特征之间存在高度非线性关联关系。文献[10]提出一种融合决策树和逻辑回归的推荐模型。作者在应用中发现,利用决策树对连续型特征进行离散化,并将转化后的特征作为逻辑回归模型的输入,可以有效提高系统的整体性能。文献[11]基于学习排序方法对上下文广告点击率进行预测。在上下文广告投放场景中,广告通常以列表的方式展现,通过列表中的广告点击情况可以得到用户偏好信息,从而获得学习排序模型,利用逻辑回归模型转换广告的排序值得到 CTR。

随着移动互联网的迅速发展,移动端广告的研究也日渐受到关注。文献[7]对移动端广告的研究现状以及研究方向做了详细的介绍。目前,移动端广告的研究大部分集中在如何从广告设计方面提升一条广告的质量以增加被用户认可的概率,而专门对广告 CTR 预测的研究则相对较少。首先,作为广告展示的载体,移动端的 APP 通常有特定的功能类型,APP 的功能和类型对用户兴趣有较强的指示作用,因此,特定 APP 只展示特定内容类型的广告;其次,移动设备可以获取比 PC 更多的用户和展示环境的信息,这些都对广告的点击有较大影响。移动端的广告点击会有一些不同于 PC 端的特征,移动端的 CTR 预测是值得探索的领域。

文献[12]使用用户在一段时间内的行为数据作为特征,将用户的行为以及每 2 个行为间的时间间隔构造 1 个序列,以此对用户进行聚类,并对聚类结果进行分析,发现可以对不同行为的用户进行划分,找出不同使用习惯的用户。文献[13]使用用户近期每个时间间隔内是否使用特定应用的序列建模,构建特征训练模型,实时预测用户当前可能会进行的操作,得到很好的效果。文献[10]通过实验发现,历史数据会对 CTR 预测的结果产生重要的影响。对于用户历史数据的利用,目前主要使用了一些与广告相关的基于统计的信息,如用户的历史广告点击数等。事实上,用户的历史数据是一系列的连续用户行为,将用户的行为看作一个连续的序列更能反映出用户的特征。文献[14]根据用户历史 APP 的使用记录来预测用户的下一个将要打开的 APP,对

多个用户手机已安装的 APP 应用列表,计算概率分数,进行 Top-K 预测,将用户最可能打开的 APP 提前载入内存,以提高 APP 的反应速度,优化用户体验。

基于已有研究,本文提出适合于移动 APP 广告的通用预测模型,充分利用了用户历史使用记录信息,并在火烈鸟果盘平台的数据上分析各种特征的使用情况,实验结果表明,使用该模型可以得到很好的预测效果。

2 预测模型和系统设计

2.1 广告投放业务架构

鉴于移动设备的特殊性,原生广告已成为移动在线广告的一种重要形式[15]。比起传统的搜索广告和上下文广告,原生广告与平台融合度更高,但同时因其尺寸原因,移动设备可展示的广告数量更少。因为广告点击率是广告质量最重要的衡量指标,所以更准确地预测广告点击率成为广告投放技术的关键。

本文设计了移动广告聚合平台业务流程,如图 1 所示,广告聚合平台主要有 5 个模块:移动客户端,服务器,广告池,算法策略和数据仓库。其中:移动客户端负责收集和上报用户行为日志,并存储在数据仓库;广告池负责定时抽取广告,以保证广告数据有效;算法策略主要利用用户和广告数据进行建模,得到个性化推荐模型,并提供接口给服务器调用;数据仓库主要存储用户日志数据、算法策略数据以及特征数据等;服务器负责广告请求分发以及模型生成更新。

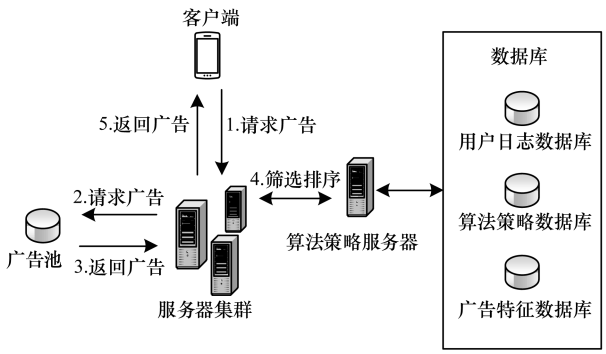

图 1 移动广告聚合平台业务流程

广告聚合平台业务流程具体如下:

步骤 1 用户打开移动客户端,并向服务器请求广告推荐列表。

步骤 2 服务器收到请求后,向广告池请求广告。

步骤 3 广告池向服务器返回广告。

步骤 4 服务器调用算法策略接口对广告进行点击率预测,然后进行筛选和排序。

步骤 5 服务器将排序后的广告发回给客户端,客户端接收后对其进行展示。

移动设备的 APP 客户端向服务器请求内容的同时,也会向服务器请求广告。服务器收到请求后,会使用当前客户端的一些信息,从广告源获取一批符合客户端投放条件的广告,再根据自身的投放策略对广告进行筛选和排序,最终选择一定数量的广告下发至客户端。在实际业务中,考虑到移动网络不稳定的情况,为了避免由于网络原因造成网络较差时无法请求广告而出现空窗现象,一般服务器会一次性向客户端发送多条广告,客户端将广告存放在本地并缓存一段时间,在 APP 有广告展示机会时将这些广告轮流进行展示。

从平台的收益考虑,通常的做法是对候选集的广告根据预期收益排序,将预期收益最高的一部分广告进行投放[8],但在大部分广告实行按点击次数计费(Cost per Click,CPC)、按行为计费(Cost per Action,CPA)的计费模式情况下,并不能准确得到广告的预期收益,因此需要对其进行预测。

对于 CPC 广告来说,收益为:

$$\Gamma = N \times CTR \times \tau \tag{3}$$

其中,τ 为每一次点击的收益,N 为广告被点击的次数。

对 CPA 广告来说,收益为:

$$\Gamma = N \times CTR \times \alpha \times \gamma \tag{4}$$

其中,α 为转化率,γ 为每次点击的转化价格。

由于在实际广告业务中,同类广告位的转化率相近,因此 CTR 预测成为关键步骤。得到预测点击率后,CPC 广告的预期收益便可确定,CPA 广告的预期收益也近似可以确定。如何得到 CTR 的预测值,机器学习领域的通用解决方法是构造数学模型,通过计算预测实际的 CTR 值。一般预测模型的构建包括 2 个主要的工作:特征数据集的选取和机器学习模型的构建。

2.2 预测模型

2.2.1 模型选择

CTR 的预测通常被看作是一个分类问题,一次特定的展示,如果用户点击了广告,则结果为 1;否则为 0,分类器通常可以给出一个 0~1 之间的数字表示结果为 1 的可能性,可以将该数字看作要预测的 CTR 值。具体做法为:从当前需要预估的广告(ad)、用户(user)和上下文环境(context)中提取各部分的特征或各部分之间的特征,构建对应的特征向量 $\boldsymbol{X}_{\text{ad}}$、$\boldsymbol{X}_{\text{user}}$、$\boldsymbol{X}_{\text{context}}$ 等,具体特征的使用则与应用场景有关,在下一节中将会详细介绍本文使用特征。在构建出各种特征向量后,将所有的特征融合得到最终作为模型输入的特征向量 $\boldsymbol{X}$:

$$\boldsymbol{X} = (\boldsymbol{X}_{\text{ad}}, \boldsymbol{X}_{\text{user}}, \boldsymbol{X}_{\text{content}}, \cdots) \tag{5}$$

由于广告点击受到多方面因素的影响,因此准确的预测结果要求使用准确的预测模型。但由于在

线广告通常有非常大的数据规模,同时又有很大的流量[10],因此就求模型要足够高效,能够快速完成大规模数据拟合,同时在实际使用时可以尽快完成预测。逻辑回归(Logistic Regression,LR)模型[16-17]因其简单高效的特点得到了广泛应用[8,10]。逻辑回归模型是一个线性分类模型,预测函数为:

$$p(y=1)=f(\boldsymbol{w};\boldsymbol{X})=\frac{1}{1+e^{-\boldsymbol{w}^{T}\boldsymbol{X}}} \tag{6}$$

其中,$\boldsymbol{w}$ 为模型参数,$\boldsymbol{X}$ 为从展示信息中提取的特征向量。通过在训练集 D 上最小化损失函数来找到最合适的参数 $\boldsymbol{w}$。损失函数的描述如下:

$$L(D,\boldsymbol{w})=-\sum_{<x,y>\in D}[y\times\lg(f(\boldsymbol{w};\boldsymbol{X}))+(1-y)\times\lg(1-f(\boldsymbol{w};\boldsymbol{X}))] \tag{7}$$

逻辑回归的优点在于简单,可以快速实现上亿规模数据的训练,但它是一个线性模型,不能获取特征间的非线性关系。

因子分解机(Factorization Machine,FM)[18-19]是一种基于矩阵分解的机器学习算法。其优势在于对稀疏数据有很好的学习能力,适合于在线广告等可能会遇到特征缺失的情景。FM 的特点在于对特征进行了组合,其预测函数为:

$$p(y=1)=\frac{1}{1+e^{-y'}} \tag{8}$$

$$y'=w_0+\sum_{i=1}^{n}\boldsymbol{w}_iX_i+\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle V_i,V_j\rangle X_iX_j \tag{9}$$

其中,参数为向量 $\boldsymbol{w}_i(i=1,2,\cdots,n)$ 和矩阵 $\boldsymbol{V}$,$\boldsymbol{V}$ 的大小为 $n\times k$,n 为特征维数,k 为超参数,$<V_i,V_j>$ 为矩阵 $\boldsymbol{V}$ 的第 i 行与第 j 行的点积。由于加入了矩阵,模型可以学习到不同特征分量之间的相互关系,因此使得模型的表达能力更强,但不足之处在于参数较多,训练时间较长。

梯度上升决策树(Gradient Boosting Decision Tree,GBDT)[20]是一种迭代的决策树模型,通过多棵回归树的组合,从而实现具有较好的学习能力和泛化能力的模型。

文献[10]指出提升决策树(Boosting Decision Tree,BDT)具有很好的特征转化和组合的功能,由此提出混合模型的思想,首先将原始特征通过 GBDT 树进行特征映射,使用样本将 GBDT 的每一棵树的输出作为分类特征,再放入其他分类器如 LR、FM 进行训练。

2.2.2　特征选择

特征选择是机器学习的重要基础,在基于模型的 CTR 预测中需要找到具有代表性的特征来构建模型。一个预测模型的预测效果很大程度上取决于使用的特征是否足够有效,是否具有预测能力[5,10]。

在不同的应用环境下,可以获得的特征也不同,因此针对特定的应用场景需要构建不同的特征集。特征按照一般类型可以分为广告特征、用户特征、用户-广告交互特征、上下文特征 4 种。广告特征和用户特征是 2 种分别描述广告和用户信息的基本特征。用户-广告交互特征一般指的是该用户对该广告的历史展示和点击情况。上下文特征指展示该条广告时的一些情况,如当前的时间、展示广告的页面内容等。上下文特征在传统的内容广告中是重要的部分[5],因为页面的内容是描述用户当前兴趣的重要指示,但在特定的移动 APP 中,广告展示的页面通常比较固定,展示环境也十分相似,而且由于客户端会预先拉取广告后展示,广告的展示时机也不确定,因此较难获取每次展示的上下文特征。本文研究主要使用广告特征($\boldsymbol{X}_{\text{ad}}$)、用户特征($\boldsymbol{X}_{\text{user}}$)和用户历史行为特征($\boldsymbol{X}_{\text{history}}$)。

广告特征和用户特征按时效特点可分为静态和动态特征,静态特征指一些比较固定的特征,如广告标题、用户使用的手机系统等;动态特征指一些会随时间变化的特征,如广告近期的点击率、用户登录与下载的数量等。

1)广告特征 $\boldsymbol{X}_{\text{ad}}$

(1)静态特征:广告唯一的 ID,提供广告的广告商,广告应用的国家代码,广告推广的应用标题、描述、大小等。

(2)动态特征:广告最近一段时间的展示数量、被点击数量等。

2)用户特征 $\boldsymbol{X}_{\text{user}}$

(1)静态特征:采集到的用户使用的设备号、软件版本、用户渠道、设备系统及版本、手机型号、网络运营商、IP 地址、年龄、性别、语言等信息。

(2)动态特征:用户近一段时间的广告展示数量、点击数量等。

3)用户历史行为特征 $\boldsymbol{X}_{\text{history}}$

用户历史行为属于用户的动态特征,显而易见用户的历史行为是预测用户行为的有力依据[12-13,21],但在传统环境下,历史行为特征不容易获取。在移动互联网中,可以根据设备的全局统一识别码筛选出属于不同用户的记录。平台软件可以记录下用户使用软件的一些行为记录,如对于一个游戏平台,可以记录用户的下载、安装、打开某个应用中的某个功能点等。经过整理,一个用户的历史记录为一个列表,列表的每一行代表用户的一次事件,如表 1 所示。

表 1 用户历史行为特征

事件	特征
1	时间 1,类型 1,参数 1
2	时间 2,类型 2,参数 2
⋮	⋮
n	时间 n,类型 n,参数 n

综上所述,本文将采用用户特征、广告特征和用户历史行为特征,并分别提取这 3 种特征中会影响广告点击率的特征数据进行建模。其中用户特征和广告特征又分为动态特征和静态特征进行提取,提取的特征如表 2 所示。

表 2 用户特征

特征类型	动态特征	静态特征
用户特征	近期广告的点击量、点击率等	设备号、软件版本、用户渠道、设备系统及版本、IP 地址、年龄、性别、用户使用的语言等
广告特征	近期的全局点击率、点击量等	广告的标题,广告的国家,推广的应用类型、描述等
历史行为特征	近期的行为事件记录	无

2.2.3 特征处理

通常在实际广告业务中,特征处理的数据主要包含广告特征数据、用户特征数据、上下文特征数据。这 3 种特征数据根据形式的不同又可以分为以下 4 类:

1)数值特征:取值为有限范围的数值(整数或小数)的特征,如用户历史的点击数量、广告推广应用的大小等。

2)枚举特征:取值为有限数量的不同可能值的特征,如广告推广商品的类别、用户使用语言等。

3)文本特征:文本内容的特征,如应用的标题、文本介绍等。

4)时序特征:取值为有序的若干个独立元素组成的序列特征,如用户历史使用记录,为由客户端软件上报的一个个事件组成的序列。

因为特征的处理方式对模型的预测效果有很大影响,所以对不同特征需要不同的处理方式。对以上 4 种数据类型,本文提出了通用的处理方法:

1)应用于数值特征的分块映射方法。数值特征一般可以直接作为模型输入,但本文对数值进行分段处理,将数值的取值范围划分为一个个区间[10,22],即将连续的数值特征转化为离散的分类特征,这样可以获取非线性的结构。实验证明,将数据分段处理可以提升 CTR 的预测效果。

对于数据分段,直观的方法是采用将取值的范围等分为若干段的方法,但数值的特征通常服从一类长尾分布,即绝大部分值都集中在较小的一段,而后面的一大段范围只有很少的示例。等分的做法会使绝大部分示例集中在少数几段。因此本文选择另一种划分方式,按照区间内实例的数量进行等分,每个区间中实例的数量近似。转化操作如下:给定排序后特征值 $features=[x_1,x_2,\cdots,x_n]$,其中 $x_i\leqslant x_j$,$1\leqslant i\leqslant j\leqslant n$。转化后的特征值为:

$$features\_transfom=\left[floor\left(\frac{index(x_1)}{features\_in\_bin}\right),floor\left(\frac{index(x_2)}{features\_in\_bin}\right),\cdots,floor\left(\frac{index(x_n)}{features\_in\_bin}\right)\right] \tag{10}$$

其中,$index$ 为计算特征排序后的索引值函数,$floor$ 为向下取整函数,$features\_in\_bin$ 为区间中实例的数量。

2)应用于枚举特征的 One-hot 映射或者哈希方法。比较通用的分类特征数值化方法是 One-hot 处理方式[8,23-24]:对于有 N 种不同值的特征,将每个值唯一地映射到一个 0 ~ $N-1$ 之间的数,对每一个实例构建一个 N 维全零向量,并将该实例取值对应的位置的值置为 1。转化操作如下:给定特征值 $features=[x_1,x_2,\cdots,x_n]$ 和散列函数 $hash(x)\in N-1$。转化后的特征为:

$$Onehot(x)=[u_1,u_2,\cdots,u_i,\cdots,u_{N-1}] \tag{11}$$

其中,$u_i=\begin{cases}1, i=hash(x)\\0, i\neq hash(x)\end{cases}$。

One-hot 方法的优点在于简单易行,但一些特征包含许多不同的离散值,而其中大部分值只出现很少的次数,若简单地对所有的值做映射,则会产生一个很长的向量。文献[23]中提出了随机映射的思想,证明该映射只会以较小的概率影响最后的结果,并且可以降低向量的长度,减轻数据稀疏度。本文中的部分数据也采用了这种随机映射方法,对于出现频率较高的值,仍然按照 One-hot 的方式映射,而对于频率较低的值则随机映射,实验结果证明该方法对于 CTR 预测效果较好。

3)应用于文本特征的 Word2vec 方法。每一个广告会包含一个标题和一段产品描述。文献[8]提出了热点词汇的方法,选取点击率较高的若干条广告,统计其中出现频率最高的 K 个词作为特征,对每一条广告的标题和描述在这些词上做 One-hot 映射和随机映射。在移动应用中,由于显示空间有限,描述的内容通常很短。在本文实验的火烈鸟果盘平台上,所有 APP 描述信息的平均长度仅为 13 个单词。考虑到由于广告展示的空间有限,标题中词的词义是很重要的部分,而高频词的做法并没有词义,为了获取标题的词义信息,本文使用 Word2vec 方法[25]。由于广告的标题和描述内容十分有限,本文首先使用 Wikipedia 的语料库进行词向量的训练,采用标题中词的词向量累加作为标题的词向量,然后对所有的广告标题进行聚类[26]。最后对标题根据聚类结果进行 One-hot 映射。

4)应用于时序特征的 Cluster 方法。对于时序型数据,目前并没有较好的处理方法。一方面,用户的行为随时间形成的有序序列反映了用户的默认特征和习惯;另一方面,不同于有语法限制的文字序列,用户的各种行为之间并不一定有严格的先后顺序。因此本文综合考虑了两方面的因素,通过处理得到 2 类特征:(1)使用文献[12]提出方案对用户行为序列进行聚类得到的特征;(2)统计用户各种行为的频率,直接作为特征。

获取用户的所有操作行为构成行为序列,所述用户的行为序列如下:

$$S=(s_1,s_2,\cdots,s_{2n-2},s_{2n-1}) \tag{12}$$

其中,设用户共有 n 个操作事件,序列 S 中奇数位置的元素为用户的各种操作行为,偶数位置的元素为 2 个操作事件发生的时间差。

使用序列 S 的所有子序列作为用户的特征进行聚类,提取 S 所有的 k-gram 集合,即所有在序列 S 中出现过的长度为 k 的子序列:

$$T_k(S)=\{(s_1,s_{i+1},\cdots,s_{i+k-1})\mid 1\leqslant i\leqslant n-k+1\} \tag{13}$$

对于一个给定的 k 和 2 个序列 S_1 和 S_2,首先计算得到 2 个序列的 k-gram 集合的并集:

$$T=T_k(S_1)\cup T_k(S_2) \tag{14}$$

分别统计 2 个序列 T 中每个元素出现的频率:

$$C_i=(c_{i,1},c_{i,2},\cdots,c_{i,m}) \tag{15}$$

其中,$c_{i,j}$ 为 T 中第 j 个元素在序列 S_i 中出现的频率。

由以上内容定义序列 S_1 和 S_2 的距离为:

$$D(S_1,S_2)=\frac{1}{\pi}\cos^{-1}\frac{\sum_{j=1}^{m}c_{1,j}\times c_{2,j}}{\sqrt{\sum_{j=1}^{m}(c_{1,j})^2}\times\sqrt{\sum_{j=1}^{m}(c_{2,j})^2}} \tag{16}$$

在得到每 2 个用户行为序列的距离后,可以依此使用基于距离的聚类方法对用户进行聚类,如谱聚类算法。在聚类后,将用户的行为序列的类别进行 One-hot 映射,映射后的向量作为该类特征。

2.3　模型构建

在特征提取和处理完成后,需要将数据导入合适的模型进行训练。由于在线业务的动态性,因此模型也需要实时更新。本文使用离线更新的模式,定时从业务日志中提取训练的样本,构建训练集对现有模型进行更新。

在实际场景中,训练数据量巨大,而且分布不均衡,正样本(产生点击的数量)的数量远少于负样本(未产生点击的数量)。二次采样[8,24]在一定程度上可以减缓这种不均衡所带来的影响。该方法以小于 1 的比例从原负样本集中选取部分负样本,以提高正样本在数据集所占的比例。经过二次采样后,样本集中正负样本的比例会发生变化,使预测的 CTR 发生偏差,因此需要对预测结果进行修正。如果用 p 表示二次采样后的分布,p_t 表示在数据集中实际的分布,那么采样前和采样后样本属于正样本的概率与属于负样本的概率之比需满足:

$$\frac{p(y=1)}{p(y=-1)}=\frac{p^t(y=1)}{p^t(y=-1)}\times\frac{1}{r} \tag{17}$$

其中,$p(y=1)$ 和 $p(y=-1)$ 分别为正、负样本的分布,r 为从原始负样本集中选取部分负样本时的比例,一般 $r\ll 1$,因此需要根据式(17)对模型预测得出的概率进行修正得到实际数据集上样本属于正样本的概率,即广告的 CTR。

$$p^t(y=1)=\frac{r\times p(y=1)}{1+r\times p(y=1)-p(y=1)} \tag{18}$$

2.4　系统设计

文献[6]提出一个基于配置语言的易扩展的数据处理系统。该系统可完成从原始数据源到最终作为预测模型输入的转化。本文系统的实现也采用该模式,将从原始数据处理到特征生成的过程分成几个连续的模块进行处理,可以方便地实现添加、删除特征,改变特征的处理方式等。该方法可以将特征处理部分与线上模型的训练和使用分开,最大限度地降低模块间的耦合,使得系统更易扩展和修改。

系统由 3 个主要部分组成:

1)提取器(Source),实现从任何外部数据源数据到数值向量的转化。数据源包括关系型数据库、结构化或非结构化的日志、网络等任何其中内容可以转化为数值向量的数据源。

2)转化器(Transformer),对 Source 产生的数值向量做进一步处理,如连续数据的离散化、特征组合等。Transformer 组件可以没有,也可以连续有几个,每一个对前一个的结果进行继续处理。

3)聚合器(Assembler),将 Transformer 或 Source 输出的各个部分的向量组合融合,生成最终作为模型输入的向量。

系统使用 JSON 文件进行配置,每个特征的配置文件中保存了生成该特征使用的数据来源,以及依次经过的若干转化器组件。在训练或线上服务时,根据配置内容依次生成该特征的特征向量,并由聚合器组件组合所有的特征向量,最后生成特征向量,传入预测模型进行训练或判断。

3　实验结果与分析

3.1　实验设置

为验证本文提出的移动互联网广告推荐系统的效果,使用火烈鸟网络的果盘游戏平台数据进行实验,实验中选取 2015 年—2017 年间连续 12 周的广告展示和点击数据进行实验。其中前 10 周的数据作为训练数据集,剩余 2 周的数据作为测试集。根据业务特点,一条广告可能会在一段时间内多次展示。实验以天为时间单位,预测一个用户在一天内对一个广告是否进行点击。从业务日志中提取一个

(用户、广告对)数据以及一天内是否点击作为一条样例。为了提高针对性,实验首先将对 12 周内没有任何点击记录的用户过滤,得到的活跃用户约 200 万,广告约 10 万条,用户广告对约 3 000 万对。同时对样本集中的负样本进行二次采样。

为研究不同模型的预测效果,实验分别使用5 种模型,其中 3 种经典模型和 2 种组合模型,并分别使用各种模型进行训练和预测。经典模型为 GBDT、LR、FM。组合模型为 GBDT + LR、GBDT + FM,先使用 GBDT 对特征进行映射,用 GBDT 的输出作为 LR 或 FM 的输入进行训练。实验目的为:1)分析比较不同模型的预测效果;2)比较特征重要性及其对结果的影响。曲线下面积(Area Under Curve,AUC)对于二分类问题是一个重要的指标,并且对于在线广告场景的不平衡样本集具有较好的适应能力,是在线广告 CTR 预测的常用标准[4-6],因此本文选择 AUC 作为结果评价指标。

3.2 结果分析

3.2.1 模型预测效果对比

本文对不同模型的预测效果进行了对比,如图 2 所示。可以看出,组合模型的效果要优于经典模型,证明了使用 GBDT 转化后的特征更具有预测能力。

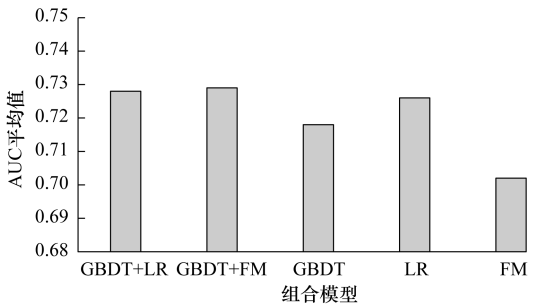

图 2 不同模型的预测效果对比

3.2.2 特征作用分析

为分析用户、广告、用户历史行为特征对预测效果的影响,对 3 类特征及其组合分别进行实验,观察使用不同的特征组合得到的预测效果,如图 3 所示。

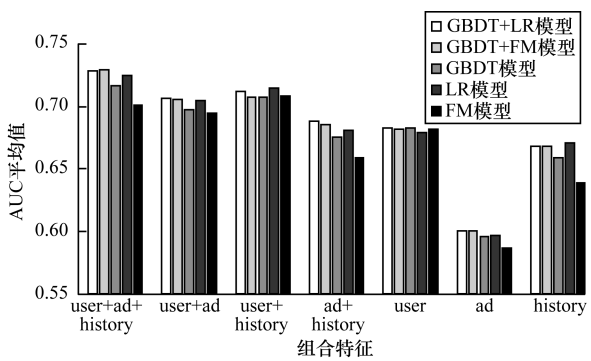

图 3 使用不同特征的预测效果对比

从图 3 可以看出,广告特征在单独使用时效果较差,但广告特征与其他特征组合仍有较好的预测性能。其主要原因为移动设备的屏幕尺寸有限,可以展示给用户的广告信息也有限,因此仅有广告数据很难决定用户是否有兴趣,也说明了深度理解应用场景以及用户需求是提升产品效果的重要途径。从图 3 还可以看到,加入用户历史行为特征后,预测结果有一定的提升。使用不同特征的预测效果对比实验的详细数值见表 3。

表 3 使用不同特征的 AUC 预测结果对比

模型	特征		
	user	ad	history
GBDT + LR	0.683 0	0.600 5	0.668 6
GBDT + FM	0.681 8	0.600 7	0.668 0
GBDT	0.683 4	0.596 0	0.658 8
LR	0.679 4	0.597 1	0.670 9
FM	0.681 7	0.586 8	0.639 1

3.2.3 模型参数选择

从表 4 的实验结果可以看出,混合模型的预测效果要优于基本模型,说明经过 GBDT 转换后的特征具有更强的预测能力。GBDT 对原始特征的转化是预测效果提升的主要原因,为了最大限度地提升预测效果,需要 GBDT 具有更好的转换功能。GBDT 的核心参数为其决策树的数量 n 以及决策树的最大深度 d。较大的参数 n 可以使模型在训练集上的误差更小,但过大可能会造成过拟合,而且会延长训练和预测时间。参数 d 决定了原始特征的组合维度,当 d 较大时,生成一棵树需要更多的条件,可以获取更多的原始特征组合方式。当 d 过大时,一棵树的可能输出很多,但每个样例在每棵树上只有一个输出,这样生成的特征向量会比较稀疏,一般控制一棵树的叶节点为 12 个左右[10]。

表 4 使用不同组合特征的 AUC 预测结果对比

模型	组合特征			
	user + ad + history	user + ad	user + history	ad + history
GBDT + LR	0.729 0	0.706 5	0.712 7	0.688 2
GBDT + FM	0.729 6	0.706 1	0.708 0	0.685 8
GBDT	0.717 3	0.697 4	0.708 0	0.676 0
LR	0.725 2	0.704 9	0.715 1	0.681 6
FM	0.701 6	0.695 3	0.709 1	0.659 2

本文对不同的 n、d 组合对结果的影响进行对比,如表 5 所示。可以看出,需要选取合适的 d 和 n 才能得到较好的 AUC 预测结果。当 d 和 n 的值偏离合适值时,预测结果呈现出减小的趋势。在本文实验中,经过多轮交叉校验,当 $d=4$、$n=10$ 时,AUC 值最高,预测效果最好。

表 5　GBDT 参数对 AUC 预测结果的影响

d	$n=10$	$n=20$	$n=30$	$n=40$
2	0.723 681	0.727 695	0.727 961	0.728 059
3	0.718 170	0.719 632	0.719 993	0.720 213
4	0.728 703	0.729 055	0.728 957	0.728 007
5	0.727 055	0.727 527	0.726 795	0.725 251
6	0.724 704	0.724 246	0.722 551	0.721 602
7	0.724 492	0.722 679	0.721 495	0.719 491

4　结束语

本文针对智能移动终端的互联网广告平台在线投放,提出一种新的在线广告投放的业务服务架构。基于该架构,提取用户基本信息、广告内容、用户使用环境等特征并进行挖掘。本文给出广告融合预测模型,将不同类型的特征进行多源融合实现在线广告的 CTR 预测,并设计实验对预测模型进行效果验证。实验结果验证了本文预测模型的有效性。下一步将引入用户时间序列行为数据特征对用户特点进行充分挖掘,以提高在线广告预测效果。

参考文献

[1] 周傲英,周敏奇,宫学庆. 计算广告:以数据为核心的 Web 综合应用[J]. 计算机学报, 2011, 34(10): 1805-1819.

[2] 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] ABBASSI Z, BHASKARA A, MISRA V. Optimizing display advertising in online social networks[C]// Proceedings of International Conference on World Wide Web. New York,USA:ACM Press,2015:1-11.

[4] MCMAHAN H B, HOLT G, SCULLEY D, et al. Ad click prediction: a view from the trenches[C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA:ACM Press,2013:76-77.

[5] LI C,LU Y, MEI Q, ET A L. Click-through prediction for advertising in twitter timeline[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA:ACM Press,2015:1959-1968.

[6] AGARWAL D,LONG B,TRAUPMAN J,et al. LASER:a scalable response prediction platform for online advertising[C]//Proceedings of ACM International Conference on Web Search and Data Mining. New York, USA:ACM Press,2014:173-182.

[7] GREWAL D, BART Y, SPANN M, ET A L. Mobile advertising: a framework and research agenda[J]. Journal of Interactive Marketing,2016,34:3-14.

[8] RICHARDSON M, DOMINOWSKA E, RAGNO R. Predicting clicks: estimating the click-through rate for new ads[C]//Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM Press,2007:521-530.

[9] MO K,LIU B,XIAO L,et al. Image feature learning for cold start problem in display advertising[C]// Proceedings of International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2015:1-9.

[10] HE X, PAN J, JIN O, et al. Practical lessons from predicting clicks on ads at facebook[C]//Proceedings of the 8th International Workshop on Data Mining for Online Advertising. New York,USA:ACM Press,2014: 1-9.

[11] TAGAMI Y, ONO S, YAMAMOTO K, et al. CTR prediction for contextual advertising: learning-to-rank approach[C]//Proceedings of International Workshop on Data Mining for Online Advertising. New York, USA:ACM Press,2013:1-8.

[12] WANG G, ZHANG X, TANG S, et al. Unsupervised clickstream clustering for user behavior analysis[C]// Proceedings of CHI Conference on Human Factors in Computing Systems. New York, USA: ACM Press, 2016:225-236.

[13] SUN Y, YUAN N J, XIE X, et al. Collaborative nowcasting for contextual recommendation[C]// Proceedings of International Conference on World Wide Web. New York,USA:ACM Press,2016:1407-1418.

[14] BAEZA-YATES R, JIANG D, SILVESTRI F, et al. Predicting the next app that you are going to use[C]// Proceedings of the 8th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press,2015:285-294.

[15] ZHOU K, REDI M, HAINES A, et al. Predicting pre-click quality for native advertisements[C]//Proceedings of International Conference on World Wide Web. New York,USA:ACM Press,2016:299-310.

[16] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[17] FAN R E,CHANG K W,HSIEH C J,et al. LIBLINEAR:a library for large linear classification[J]. Journal of Machine Learning Research,2008,9(9):1871-1874.

[18] RENDLE S. Factorization machines[C]//Proceedings of IEEE International Conference on Data Mining. Washington D. C. ,USA:IEEE Computer Society,2010: 995-1000.

[19] RENDLE S. Factorization machines with libFM[J]. ACM Transactions on Intelligent Systems and Technology,2012,3(3):219-224.

[20] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2016: 785-794.

(下转第 191 页)

[2] 张灵,丁伍洋.一种基于面部运动单元识别的驾驶员疲劳检测方法:CN 103479367 B[P].2016.
[3] WALECKI R,RUDOVIC O,PAVLOVIC V,et al. Copula ordinal regression for joint estimation of facial action unit intensity[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:4902-4910.
[4] WALECKI R,OGNJE N,RUDOVI C,et al. Deep structured learning for facial action unit intensity estimation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2017:5709-5718.
[5] MING Z,BUGEAN A,ROUAS J L,et al. Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine[C]//Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition. Washington D. C.,USA:IEEE Press,2015:1-6.
[6] GUDI A,TASLI H E,UYL T M D,et al. Deep learning based FACS action unit occurrence and intensity estimation[C]//Proceedings of IEEE International Conferences on Automatic Face and Gesture Recognition. Washington D. C.,USA:IEEE Press,2015:1-5.
[7] ZHAO K,CHU W S,ZHANG H. Deep region and multi-label learning for facial action unit detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:3391-3399.
[8] MOHAMMADI M R,FATEMIZADEH E,MAHOOR M H. Intensity estimation of spontaneous facial action units based on their sparsity properties[J]. IEEE Transactions on Cybernetics,2016,46(3):817.
[9] JENI L A,GIRARD J M,COHN J F,et al. Continuous AU intensity estimation using localized,sparse facial feature space[C]//Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition. Washington D. C.,USA:IEEE Press,2013:1-7.
[10] BENITEZQUIROZ C F,SRINIVASAN R,MARTINEZ A M. EmotioNet:an accurate,real-time algorithm for the automatic annotation of a million facial expressions in the wild[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:5562-5570.
[11] 丁伍洋.基于光流与 HMM 的疲态人脸中运动单元识别研究[D].广州:广东工业大学,2014.
[12] 仝少敏.基于受限玻尔兹曼机的面部运动识别方法研究[D].哈尔滨:哈尔滨工业大学,2014.
[13] MAVADATI S M,MAHOOR M H,BARTLETT K,et al. DISFA:a spontaneous facial action intensity database[J]. IEEE Transactions on Affective Computing,2013,4(2):151-160.
[14] ZHAO G,PIETIKAINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(6):915-928.
[15] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Computer Society,2005:886-893.
[16] CHANG C C,LIN C J. LIBSVM:a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.
[17] VALSTAR M F,ALMAEV T,GIRARD J M,et al. FERA 2015-second facial expression recognition and analysis challenge[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:1-8.
[18] SHROUT P E,FLEISS J L. Intraclass correlations:uses in assessing rater reliability[J]. Psychological Bulletin,1979,86(2):420-431.
[19] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-sale image recognition[EB/OL].[2017-10-21]. https://www.jianshu.com/.
[20] NIU Z,ZHOU M,WANG L,et al. Ordinal regression with multiple output CNN for age estimation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:4920-4928.

编辑　索书志

(上接第 185 页)

[21] YANG J,QIAO Y,ZHANG X,et al. Characterizing user behavior in mobile Internet[J]. IEEE Transactions on Emerging Topics in Computing,2015,3(1):95-106.
[22] SAPATINAS T. The elements of statistical learning[J]. Journal of the Royal Statistical Society:Series A (Statistics in Society),2004,167(1):192-192.
[23] WEINBERGER K,DASGUPTA A,LANGFORD J,et al. Feature hashing for large scale multitask learning[C]//Proceedings of International Conference on Machine Learning. New York,USA:ACM Press,2009:1113-1120.
[24] 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.
[25] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL].[2017-08-11]. http://cn.arxiv.org/abs/1301.3781.
[26] PEDREGOSA F,VAROQUAUX G,GRAMFORT A,et al. Scikit-learn:machine learning in Python[J]. Journal of Machine Learning Research,2013,12(10):2825-2830.

编辑　陆燕菲