基于用户相似度和特征分化的广告点击率预测研究

潘书敏 颜 娜 谢瑾奎
(华东师范大学计算机科学技术系 上海 200241)

摘 要 大数据环境下如何对互联网广告进行精准投放一直是计算广告学领域高度关注的问题。作为在线广告投放效果的一个重要指标,点击率的精确预测关系到媒体、用户和广告主三方的利益。目前的主流方法是通过抽取特征建立单一点击率预测模型,其不足之处在于使用单个权重来度量特征对点击率的影响过于片面。该研究基于分而治之的思想,提出了基于用户相似度和特征分化的混成模型。该模型首先根据混合高斯分布来评估用户相似度,将其划分为多个群体。针对不同群体,分别构建子模型并进行有效组合,从而挖掘同一特征对不同群体的差异化影响,进而准确地预测广告点击行为。通过使用真实互联网公司的广告数据集进行实验,并与主流方法做了详细的对比分析,检验了该方法的有效性。

关键词 计算广告学,点击率预测,用户相似度,特征分化,混成模型
中图法分类号 TP274 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.048

Study on Advertising Click-through Rate Prediction Based on User Similarity and Feature Differentiation

PAN Shu-min YAN Na XIE Jin-kui
(Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China)

Abstract Targeting the Internet advertising accurately is an eye-catching problem in the field of computational advertising. As an important evaluation criteria for online advertising effect, the precision of prediction for click through rate (CTR) benefits publishers, advertisers and users. Without considering feature differentiation, mainstream approaches are extracting features and establishing click prediction model, which use a single weight to measure the effect of a feature for CTR. According to the idea divide and conquer, a hybrid model based on user similarity and feature differentiation was proposed. The model divides users into several groups depending on user similarity evaluated by mixture gaussian distribution. For each group, model was built respectively and they were combined to excavate the different effects of a feature to different groups and improve predict CTR prediction accuracy. Several experiments on advertising data sets of an Internet companies were made and the effectiveness of the approach through detailed comparative analysis was verified with the mainstream approaches.

Keywords Computational advertising, CTR prediction, User similarity, Feature differentiation, Hybrid model

1 引言

互联网在线广告(Online Advertising)[1-3]是指广告主基于互联网中的图形、视频、文字等形式直接投放广告,主要分为搜索广告和展示广告两大类。搜索广告(Sponsored Advertising)是用户搜索关键词时展示的广告,只在搜索引擎中出现;展示广告(Display Advertising)是各中小型网站为了盈利,将自己网站的某些位置当作广告位来出售,也是长尾理论[4]在互联网广告中的一个成功运用。随着互联网行业的多年积累和不断发展,越来越多的广告主从传统媒体广告的投放系统转移到数字化营销上,在线广告也成为了如Google和Yahoo等互联网公司最具利润的商业模式之一。中国产业信息网发布的《2015—2020年中国互联网广告市场格局及投资咨询报告》中指出,中国互联网广告市场增速远高于电视、报纸、广播、杂志等传统媒体平台,目前约占整体广告市场的20%,预计2015年将进一步增长至26.6%,并在2017年达到2800亿元。

实时竞价(Real Time Bidding, RTB)[5]是在线广告中一种广泛应用的广告投放机制。它通过第三方技术对用户展示行为的评估来帮助广告主进行出价和广告投放,其主要流程如图1所示。

到稿日期:2015-12-07 返修日期:2016-04-24
潘书敏(1992—),女,硕士生,主要研究方向为社会计算、计算广告学,E-mail:panshumin829@163.com;颜 娜(1992—),女,硕士生,主要研究方向为社会计算、算法博弈论;谢瑾奎(1975—),男,博士,副教授,CCF会员,主要研究方向为社会计算、算法博弈论,E-mail:jkxie@cs.ecnu.edu.cn(通信作者)。

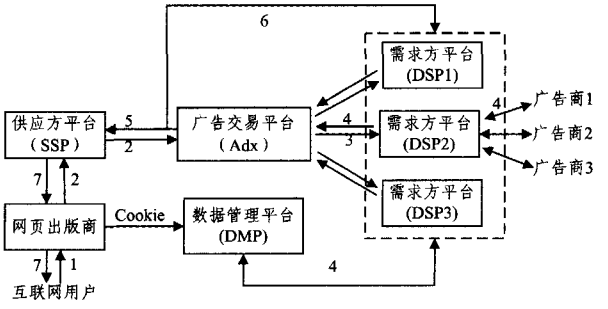

图 1 实时竞价广告投放流程简图

RTB 主要有三大组成部分：供应方平台(SellSide Platform，SSP)，主要服务于媒体；需求方平台(Demand Side Platform，DSP)，服务于广告主，为广告主提供广告投放的策略；广告交易平台(Ad Exchange，Adx)，将广告展示信息传递给需求方平台。DSP 在为广告主指定广告投放策略时需要考虑参与广告位竞拍的价格以及广告投放的效果。广告投放效果通常是通过点击率(Click-Through Rate，CTR)即广告点击次数与展现次数的比例来衡量的。由于目前 RTB 广泛运用的计费模型为按照每次点击计费(Cost Per Click，CPC)的模型，即

$$ecpc = cpc * ctr$$

因此，点击率也是 DSP 在帮助广告主制定竞价价格时的一个重要的参考依据。

与搜索广告可以根据当前用户的搜索关键词来实时获取用户的兴趣不同，展示广告中，广告主只能根据用户的历史行为来推测用户的兴趣，因此点击率更是广告精准投放的一个重要指标。最大化投放广告的点击率可以保证媒体、广告主和用户三方的利益：对于媒体而言，高点击率的广告可以提高网站收益；对于广告主而言，精确的点击率预估可以辅助其合理使用预算，获得广告曝光机会，将广告传递给目标人群；对用户而言，浏览自己感兴趣的广告也相应地提高了用户体验。但是，由于广告数据呈现天然的稀疏性(用户未发生点击的行为日志和发生点击的行为日志的比值达到几百，甚至达到千的级别)，使得展示广告中用户的点击率预测难以达到很高的准确率。

本文的主要工作是从用户和特征的角度，针对展示广告提出一种基于用户相似度和特征分化的点击率预估方法(Advertising Click-through Rate Estimation Based on User Similarity and Feature Differentiation，USFD)。通过对具有相似特征的用户进行聚类分析，探索同一特征对不同对象点击行为的差异化影响，以此来提高点击率预测的精度。同时本文在真实数据集上进行了大量的实验，并与其他方法进行了详细的比较分析，结果证明了该方法的有效性。

2 相关工作

目前的点击率预测主要采用的是机器学习中的分类技术，根据用户历史行为数据，抽取有效的特征对用户行为进行建模。Richardson 等人[6]提出了使用逻辑回归模型来学习基于搜索广告的 CTR。然而，展示类广告无法像搜索广告一样可以明确地获取用户的搜索需求，不存在查询关键词等重要的信息特征，因此该模型的准确度不高。

Lee 等人[7]为缓解数据稀疏性，引入数据分层(Data Hierarchy，DH)的思想来评估展示广告转化率。作者考虑广告展示数据存在上下层级关系的特性(例：一个网站下包括多个网页，每个网页下又存在多个广告位)，从媒体、用户和广告主三方对数据进行分层处理。同时，考虑到数据在同一层次保持着更多的相似性，集成数据在各层上的弱点击率来提高预测的准确性。Wang 等人[8]采用贝叶斯模型对点击率进行评估。同文献[7]的思想类似，Agarwal 等人[9]基于广告数据事先已有的不同粒度层次的概念，采用树状马尔可夫的方法对广告点击率进行预测。总结来看，以上几种模型更倾向于从用户历史行为来考虑点击率的预估，没有深入挖掘具体特征和用户行为之间的关系。Agarwal 等人[10]提出了一个时空模型来计算文档的点击率。该方法根据同一文章的重复展示特性，调节用户的疲惫指标，建立动态线性回归模型，对广告数据的点击率预测同样有重要的借鉴意义。

此外，随着神经网络和深度学习[11]的不断发展，很多学者开始使用深度学习的方法来学习特征间的高维关系从而提高点击率预估的准确率[12]。深度学习方法的一个主要劣势在于训练算法较为复杂，训练耗时较长。

在前人方法的基础上，本文以特征对用户行为的影响为主要着眼点，利用混合高斯刻画用户的特征分布，对用户行为进行聚类分析。因此，本文最大的特色之处在于突出了特征对于不同类型用户的差异化影响，同时针对单一分类器在分类准确率上的缺陷，采用了组合模型的思想，提出了基于用户相似度和特征分化的广告点击率预测混成方法。

3 基于用户相似度和特征分化的 CTR 预测

3.1 USFD 预测方法的基本框架

定义 1 考虑二分类问题，对于给定数据集 $\{x_i, y_i\}_{i=1}^n$，其中，$\forall x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\}$ 表示每一个数据对象由 m 个特征构成；y_i 是标签，取值 $+1$ 或 -1，分别代表正例或负例。

CTR 评估的任务就是根据用户的历史广告点击数据预测未来用户对广告的点击情况。此情境下，正例数据即为产生了点击的数据，反之为负例。传统的机器学习(如逻辑回归、支持向量机等)通常根据大量训练样本来训练特征的权重，这类模型的不足在于每个特征在所有样本对象中占有同样的权重，然而在实际生活中，特征对不同对象的重要程度显然不同。以购买电子产品为例，价格对所有群体的影响程度可能相同，但与此同时，男性群体更关注电子产品的性能，女性群体对外观的关注程度则远大于其本身的性能。显然，单一分类器无法区别该特征对不同性别群体影响的差异，因而降低了分类器预测结果的准确性。

基于上述现象，本文的设计思路如图 2 所示。对于历史用户的广告展现日志，提取用户特征，采用聚类的方法对用户进行相似性划分，每个子集被看作是具有相似行为的用户。对于各聚类子集，再运用传统机器学习的方法构造子分类器。对于新来的用户和广告信息，计算用户与各子集的相似度，再将其放入各子分类器进行训练得到相应的概率，最终对其进行合理组合，得到一个具有差异性特征效果的点击概率。

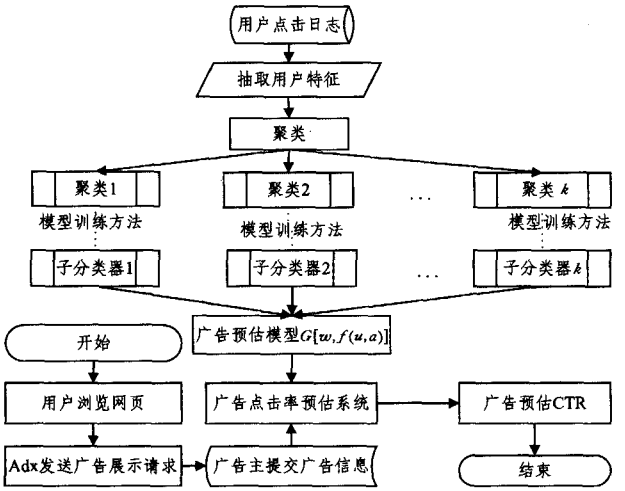

图 2　基于用户相似度和特征分化的 CTR 预测框架

综上所述，该方法的两个技术重点是如何准确地刻画用户相似性以及对子分类器预测概率进行合理组合。本文将在后面章节对这些问题的具体解决方案进行阐述。

3.2　USFD 预测方法的具体实现

3.2.1　基于混合高斯分布的用户聚类

高斯分布是统计学和机器学习领域的一个重要的概率分布模型。本文采用多高斯混合模型来刻画历史数据集中的用户分布。一条多维数据的概率密度通常可以使用多维单高斯模型表示，如式(1)所示：

$$\mathcal{M}(x;\mu,\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}\exp(-(x-\mu)^{T}\Sigma^{-1}(x-\mu)) \quad (1)$$

其中，x 为 d 维的样本向量，μ 是样本集的期望，Σ 是方差。然而，针对复杂数据集而言，一个单高斯模型往往不能完整地描述其数据的分布情况，因此需要采用多个高斯模型来逼近数据的分布，故称作高斯混合模型[13](Gaussian Mixture Model，GMM)。在 GMM 中，对于给定的训练样本 $\{x_1,\cdots,x_m\}$，其隐含类别标签用 $Z=\{z_1,\cdots,z_k\}$ 表示。假定 z_i 满足多项式分布 $z_i\sim Multinomial(\pi)$，其中 $p(z_i=j)=\pi_j$ 且 $\pi_j\geqslant 0$，$\sum_{j=1}^{k}\pi_j=1$，而且在给定 z_i 后，假定 x_i 满足多值高斯分布，可以得到 x_i,z_i 的联合概率分布，如式(2)所示：

$$p(x_i,z_i)=p(x_i|z_i)p(z_i) \quad (2)$$

整个 GMM 模型可以简单描述为：对于每个样例 x_i，先从 k 个类别中按多项式分布抽取一个 z，然后根据 z 所对应的 k 个高斯分布中的某个分布生成样例 x_i 的概率，这里的 z_i 是隐含随机变量。由此也可以看出，模型中有 3 个需要求解的变量，分别为 π,μ 和 Σ。参数的求解采用的是极大似然估计的思想，根据以上描述，可以得到似然函数，如式(3)所示：

$$\ell(\pi,\mu,\Sigma)=\sum_{i=1}^{m}\log p(x_i;\pi,\mu,\Sigma) \quad (3)$$

将式(2)代入似然函数得式(4)：

$$\ell(\pi,\mu,\Sigma)=\sum_{i=1}^{m}\log\sum_{z_i=1}^{k}p(x_i|z_i)p(z_i) \quad (4)$$

由于 z 是隐含变量，因此最后使用 EM 算法来估计各参数[14]。用 $\Theta=\{(\pi_1,\mu_1,\Sigma_1),\cdots,(\pi_k,\mu_k,\Sigma_k)\}$ 表示求解的参数。对于历史日志中的每条记录 x_i，其总体概率分布可以看作由 k 个加权 π_i 的高斯函数 $\mathcal{M}(x_i,\mu_i\Sigma_i)$ 的线性组合构成，如式(5)所示：

$$\Phi(x;\Theta)=\sum_{i=1}^{k}\pi_i\mathcal{M}(x_i,\mu_i\Sigma_i) \quad (5)$$

其中，$\mathcal{M}(x_i,\mu_i\Sigma_i)$ 是对象在第 i 个子类的高斯分布概率，本文使用用户相似度的概念来表示各样本在各子类中的概率分布。

定义 2(用户相似度)　对于每个用户，用 s_i 表示用户在第 i 个子高斯分布的概率密度，称作用户与用户子集 i 的相似度，向量 $\vec{s}=\{s_1,\cdots,s_k\}$ 称作用户的相似度向量。

根据 GMM 的聚类方法，可以按照用户特征相似度将样本划分为 k 个子集，且每个子集服从一个子高斯分布。在3.1节中提到，点击率的预测可以看作是机器学习中的经典的二分类问题，对于每个历史用户子集都可以训练出一个子点击率预估模型。这里使用最常用的逻辑回归模型作为基本的子分类器模型。逻辑回归模型[15]可以看作是线性回归模型通过 Sigmoid 函数将其值域映射到了[0,1]范围内。对于任何一条样本，该用户点击广告的概率如式(6)所示：

$$p(y=1|X)=\frac{1}{1+e^{-w^{T}X}} \quad (6)$$

其中，X 为特征向量，w 是要求解的特征权重向量，逻辑回归中求解 w 常采用梯度下降、L-BFGS 方法或者拟牛顿方法[16]等方法。本文的实验中采用 L-BFGS 方法对参数进行求解。

对于新来用户和广告，提取其特征放入 k 个点击率预估子模型中得到弱估计概率 $p_1,\cdots,p_k$，组成向量 $\vec{p}$。为了求解最终的点击率，需要对这 k 个弱估计概率进行合理组合。传统的投票机制中，各子分类器在最终的决策中占有相同的权重，这意味着各子集对预测对象的影响是相同的；然而本文中，各子用户集内用户特征具有相似性，而用户子集之间的差异性较大。本文假设特征相似度高的用户在行为上也具有较高的相似性，因此各用户子集对该用户行为的判断应取决于该用户与用户子集的相似程度。基于该假设，采用用户与该用户子集的相似度作为该分类器在评判用户行为时的权威值。

算法 1　USFD

Input：训练集 X，测试集 T，子类数目 k
Output：测试集中每条数据的点击率 $\tilde{p}$
Step1：Clusting
1. 读取训练集 x，提取用户特征组成用户特征样本集合 v
2. 计算似然函数 $\ell(\pi,\mu,\Sigma)=\sum_{i=1}^{m}\log\sum_{z_i=1}^{k}p(u_i|z_i)p(z_i)$
3. EM 算法估计高斯模型的参数 Θ
4. For each u_i in U：
5. $\vec{s}=\mathcal{M}(u_1;\Theta)$
6. 选取相似度最大对应的子集 γ_m，$\gamma_m=\gamma_m\cup\{x_i\}$
7. Return $\gamma=\{\gamma_1,\cdots,\gamma_k\}$ 和 Θ
Step2：Training
8. For Each γ_i in γ：
9. 训练逻辑回归模型 lr_i
10. Return $LR=\{lr_1,\cdots,lr_k\}$
Step3：Predicting
11. For each t_i in T：
12. $\vec{s}=\mathcal{M}(t_i;\Theta)$
13.　　For each lr_i in LR：
14. $p_i=lr_i.predict(t_i)$

15. $\vec{p}=\{p_1,\cdots,p_k\}$

16. pctr=$\vec{s}*\vec{p}^{T}$

根据定义 2,可以由高斯分布的各参数求解出用户相似度$\vec{s}$,由此可以求解出该用户最终的广告点击率,如式(7)所示:

$$pctr=\vec{s}*\vec{p}^{T} \tag{7}$$

总结以上阐述,得到 USFD 流程如算法 1 所示。

4 实验结果与分析

4.1 数据集预处理

4.1.1 实验数据集介绍

本文实验采用的数据集来自某广告公司的真实广告历史点击日志[17],共包括 7 天约 1200 多万条广告展示和点击数据。每一条数据中包括有用户、广告主、域名、广告创意等 24 个字段信息。每天的展示及点击情况如表 1 所列。

表 1 各天广告日志展示及点击信息

#Day	#Impression	#Click	CTR
1	1821350	1289	0.0007
2	1805953	1158	0.0006
3	1634830	1302	0.0008
4	1651524	1250	0.0008
5	1920370	1779	0.0009
6	1745722	1593	0.0009
7	1657338	1607	0.0009

表 1 中,Impression 是每天的广告展示数量,Click 为展示的广告被点击的数量,CTR 是平均点击率。由表 1 也可以看出,广告数据呈现了严重的稀疏性特征,各天的展示数据中,点击数和非点击的比例都约为 1000∶7。

定义 3(稀疏率,Sparse Ratio,SR) 对实验数据进行随机抽样得到样本集,该样本集包含一定量的点击数据和非点击数据,将样本集中的非点击数据和点击数据的比值称为该样本集的稀疏率。

4.1.2 特征预处理

(1)离散型特征处理

由于模型的训练过程中都需要数值型的数据输入,因此需要将非数值型数据转化为数值型的数据。本文对离散型特征采用了独热编码(One-Hot Encoder)[1]方法进行转换,即对每个离散的特征,如果它有 N 种取值,则可以将它转化为一个 N 维的二进制向量。以数据集中的用户浏览器属性为例,通过对浏览器字段的解析和统计,得到用户使用的浏览器的比例分布图,如图 3 所示。

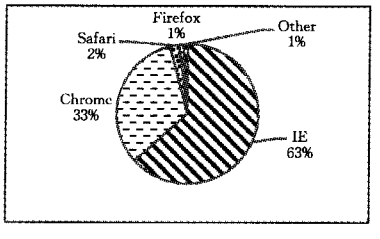

图 3 用户浏览器使用分布图

统计发现,IE 用户和 Chrome 用户占整个数据集的近 96%。根据经验,Safari 用户和 Firefox 用户虽然所占比例比较少,但是可能显著地代表某一类群体,故将这两者也与其他分开。据此,将浏览器这一个属性根据独热编码的方式扩展成五维属性,例如一个使用 IE 浏览器的用户可以用向量 $f=(1,0,0,0,0)$ 来表示,Chrome 的用户则可以用向量 $f=(0,1,0,0,0)$ 来表示,以此类推。同理,对数据集中的广告交易平台、广告主等其他离散特征,也做了同样的编码处理。

(2)连续型特征处理

连续型特征的处理主要采用归一化[18]的方法:

$$value_{new}=\frac{value-value_{\min}}{value_{\max}-value_{\min}}$$

其中,$value_{\max}$ 和 $value_{\min}$ 分别为该特征取值范围内的最大值和最小值。将特征进行归一化可以减小特征与特征之间数据数量级上的差异对实验结果造成的影响。

4.2 实验对比方法和评估指标的选择

4.2.1 实验对比方法的选择

为了评估本文方法的有效性,本文选取其他 3 种方法进行了对照实验。如表 2 所列,Basic_LR 为使用传统的机器学习方法——逻辑回归进行分类;DH 为文献[6]所提到的方法;Bagging[18]是机器学习中经典的分类器组合方法,与本文提出的 USFD 方法不同,Bagging 采用随机抽样,子样本集合内的数据不具备相似性,并且产生的各子分类器在最后的决策过程中占有相同的权重。

表 2 各实验方法的简单对比

点击率预估方法	采样方式	分类器权重
Basic_LR	基本的逻辑回归分类器	
Bagging	随机采样	各子分类权重相同
DH	对媒体、广告主、用户三方分别进行分层处理	
USFD	根据用户相似度聚成子类	样本用户与各子类用户的相似度为该分类器的权重

4.2.2 分类器评估指标[18]

(1)P-R 曲线

准确率(Precision)和召回率(Recall)是评估分类器性能的两个常用指标。准确率反映的是分类器对整个样本的判定能力;召回率反映了被正确判定的正例占总的正例的比重。基于准确率和召回率,可以作出 P-R(Precision-Recall)曲线,P-R 曲线越往坐标(1,1)的位置靠近则分类器性能越好。

(2)ROC 曲线和 AUC 值

针对严重稀疏的数据集而言,准确率作为评估指标存在一个严重的缺陷,即不能恰当地反映分类器的性能。例如:测试样本中有 A 类样本 90 个,B 类样本 10 个。分类器 C1 把所有的测试样本都分成了 A 类,分类器 C2 把 A 类的 90 个样本分对了 70 个,B 类的 10 个样本分对了 5 个。虽然 C1 的分类精度为 90%,C2 的分类精度为 75%,但是显然 C2 的分类器更具有实用性。

此外,本文还选取受试者工作特征曲线(Receiver Operating Characteristic Curve,ROC 曲线)和 AUC 值(Area Under roc Curve)作为分类器的另一项评估指标。ROC 曲线基于混淆矩阵,其横坐标为假正率,纵坐标为真正率。AUC 值是 ROC 曲线的一个直观反映,AUC 值越大则分类器的性能越好。

[1] http://scikit-learn.org

4.3 实验结果与分析

为平衡实验的效率，同时也为了突出数据集稀疏的特点，选取了稀疏率分别为 1，3 和 10 的数据组成不同大小的数据集分别进行实验。实验采用最简单的 Holdout 验证方式[18]，抽取数据集的 80%作为训练集，余下的 20%作为测试集，并统计各实验设置下不同方法的 PR 曲线、ROC 曲线和 AUC 的值进行比较，以下将对实验结果的各指标进行一一分析。

4.3.1 各子用户集特征权重分析

与单一分类器方法的区别在于，本文提出的 USFD 方法对用户进行相似性聚合时，易于发现同一特征对不同用户的影响力不同。实验中，分别选取了子类数目 $K=2$ 和 $K=3$ 时各子分类器具有代表性的特征权重与单一分类器的相应特征权重作比较，如图 4 和图 5 所示。

从图 4 中可知，特征 1，2，3 在各子集中都占有很重要的权重。值得注意的是，特征 4 在聚类集合 1 中的影响是积极的，在聚类集合 2 中的影响是消极的，然而在单分类器中该特征对所有对象的影响都是积极的；特征 5，7，8 也具有相同的规律。同理，分析图 5 中的统计结果，可以得到相同的结论。由此可以看出，本文提出的方法可以挖掘特征在不同群体中的不同影响力，以此来提高模型的预测性能。

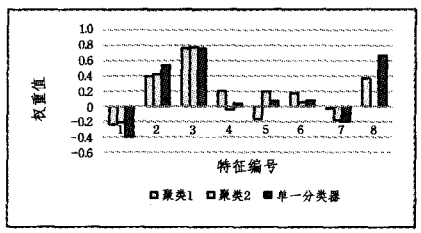

图 4　子类特征权重 VS 单一分类器特征权重($K=2$)

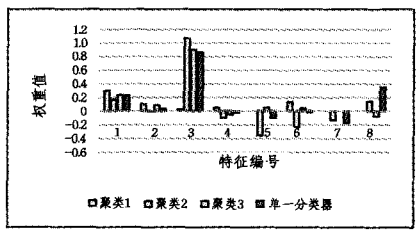

图 5　子类特征权重 VS 单一分类器特征权重($K=3$)

4.3.2 不同方法的 P-R 曲线分析

图 6 示出了 $SR=10$ 的数据集下的 P-R 曲线。

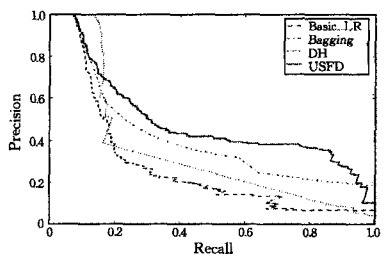

图 6　4 种方法的 P-R 曲线对比($SR=10$)

可以看出，分类器的组合方法如 Bagging 和 USFD 的性能要优于其他方法；同时由于本文提出的方法是以用户相似度进行聚合，更易于发现特征对不同用户对象的影响力不同，因此性能要优于 Bagging。Basic 和 DH 法的性能都较差，并且 DH 的曲线不够平滑，可能存在过拟合的现象。

4.3.3 不同方法的 ROC 曲线和 AUC 值分析

表 3 列出了 USFD 在不同数据集中选取不同聚合数目 K 的情况下 AUC 值的比较。为了能清晰地看出本文方法的预估质量在不同规模数据下及聚类数目不同时的差异，本文根据表 3 得到相应的预估 AUC 值直方图，如图 7 所示。

表 3　不同 SR 和 K 值下 USFD 方法的 AUC 值

	1	3	10
$K=2$	**0.763**	0.744	0.769
$K=3$	0.736	**0.782**	**0.792**
$K=4$	0.735	0.760	0.756
$K=5$	0.747	0.737	0.758

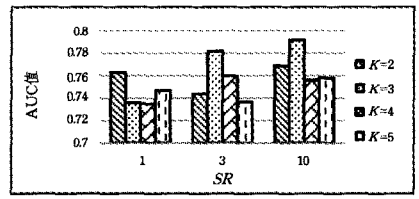

图 7　不同 SR 和 K 值下 USFD 方法的 AUC 值直方图

由实验结果可以看出，当 $SR=1$ 时，由于数据集本身数据量比较少，划分为多个集合导致各集合的数据量太少，使得各子分类器不够精确，从而影响最终的分类效果。

在数据集相对充足的情况下，选取聚类数目 $K=3$ 时分类器取得的性能较好，随着子类数目增加，AUC 值会降低。分析其原因，可归结为两点：1）由于子类数目增加，各子类数据量太少，导致分类器不够精确；2）可能由于子类划分过多导致训练存在过拟合现象，影响分类器的性能。

表 4 列出了不同数据集下，不同方法的 AUC 值比较，此处本文方法的 AUC 值选取的是表 3 中各设置下性能最好时的取值。各数据集下的 ROC 曲线如图 8—图 10 所示。

表 4　不同 SR 下 4 种方法 AUC 值的对比

SR	1	3	10
Basic_LR	0.737	0.726	0.723
DH	**0.775**	0.733	0.755
Bagging	0.744	0.728	0.746
USFD	0.763	**0.782**	**0.792**

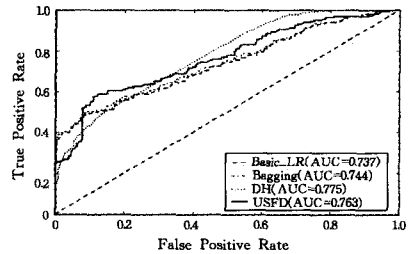

图 8　4 种方法 ROC 曲线的对比($SR=1$)

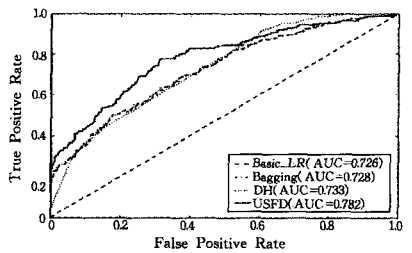

图 9　4 种方法 ROC 曲线的对比($SR=3$)

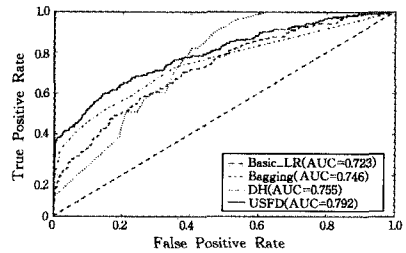

图 10 4 种方法 ROC 曲线的对比($SR=10$)

总体而言,基本分类器的性能 Basic_LR 比其他方法的效果都要差。可以看出在数据量较小的情况下,划分数据集不利于分类器的训练。数据分层是对数据进行的一种分层聚合,在较小的数据集下表现的效果优于其他方法。

随着数据集的增大,先对数据进行聚类划分,将相似的对象聚集在一起,有利于合理地进行用户定向,深度挖掘特征在不同用户群体中的不同影响力,使得预测更加精准,因此分类器的效果更佳。在 $SR=10$ 的情况下,USFD 比 Basic_LR 提高了约 9.5%,比其他两种方法也提高了近 5%。Bagging 的方法在数据抽样上采用的是随机抽样,每个样本集不具备相似性,且分类器在最后的组合中所占的比例相同,因此在性能上也不如本文提出的 USFD 方法。另外,DH 方法的 ROC 曲线不够平滑,存在一定的过拟合现象,这与之前分析的 P-R 曲线得到的结论相一致。

4.3.4 USFD 方法处理效率的分析说明

RTB 系统对实时性要求非常高,要求从用户浏览广告产生广告请求到最终的广告呈现这一系列流程在毫秒级的时间内完成。受实验环境的限制,本实验无法获得真实的在线广告投放场景。但是同其他方法,本文提出的 USFD 的参数估计和训练都可以在线下完成,因此满足了线上实时性的要求。值得一提的是,在 USFD 方法中,各子分类器间的训练互不干扰,并行处理较单一模型的训练而言也大大缩短了训练时间。

结束语 计算广告学蓬勃发展,精准的广告点击率对于媒体、广告主和用户都有着重要意义。本文的主要工作是从用户和特征的角度,针对展示广告提出一种基于用户相似度和特征分化的点击率预估方法。通过对具有相似特征的用户进行建模分析,挖掘特征的差异化对用户点击行为的影响,以此来提高展示广告的点击率预测精度。实验证明,本文提出的 USFD 方法比传统的逻辑回归方法在 AUC 值上提高了 9.5%,比文中提到的 Bagging 和 DH 两种方法也提高了近 5%。另一方面,本文提出的方法中的参数估计都可以在线下完成,满足线上计算实时性的要求;并且本文的方法具有一定的适应性,对于子分类器的选择并不仅仅局限于逻辑回归,而是可以根据数据特性选择适合的分类方法。

本文后续的重点研究工作有:将用户的反馈放入点击率模型中,构造实时动态的点击率反馈模型;同时,将准确的点击率预估运用到后续的广告投放竞拍中,为广告主指定合理的竞价策略。

参 考 文 献

[1] BRODER A Z. Computational advertising[C]// Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA. San Francisco, California, USA, 2008.

[2] ZHOU A Y, ZHOU M Q, GONG X Q. Computational Advertising: A Data-Centric Comprehensive Web Application [J]. Chinese Journal of Computers, 2011, 34(10): 1805-1819. (in Chinese)
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.

[3] JI W D, WANG X L, ZHOU A Y. Techniques for estimating click-through rates of Web advertisements: A survey[J]. Journal of East China Normal University(Natural Sciences), 2013(3): 2-14. (in Chinses)
纪文迪,王晓玲,周傲英. 广告点击率估算技术综述[J]. 华东师范大学学报(自然科学版),2013(3):2-14.

[4] ANDERSON C. The Long Tail: Why the Future of Business Is Selling Less of More[M]. Hyperion, 2006.

[5] YUAN Y, WANG F, LI J, et al. A survey on real time bidding advertising[C]// 2014 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). IEEE, 2014: 418-423.

[6] 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. ACM, 2007: 521-530.

[7] LEE K C, 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. ACM, 2012.

[8] WANG X, LI W, CUI Y, et al. Click-through rate estimation for rare events in online advertising [J]. Online Multimedia Advertising: Techniques and Technologies, 2011, 10: 1-12.

[9] AGARWAL D, BRODER A Z, CHAKRABARTI D, et al. Estimating rates of rare events at multiple resolutions[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA, 2007: 16-25.

[10] AGARWAL D, CHEN B C, ELANGO P. Spatio-temporal models forestimating click-through rate[C]// Proceedings of the 18th International Conference on World Wide Web. Madrid Spain, 2009: 21-30.

[11] YU K, JIA L, CHEN Y, et al. Deep Learning: Yesterday, Today, and Tomorrow[J]. Journal of Computer Research & Development, 2013, 50(9): 1799-1804.

[12] ZHANG Z Q, ZHOU Y, XIE X Q, et al. Research on Advertising Click-Through Rate Estimation Based on Feature Learning [J]. Chinese Journal of Computers, 2016, 39(4): 780-794. (in Chinese)
张志强,周永,谢晓芹,等. 基于特征学习的广告点击率预估技术研究[J]. 计算机学报,2016,39(4):780-794.

[13] STAUFFER C, GRIMSON W E L. Adaptive Background Mixture Models for Real-Time Tracking[C]// Cvpr. IEEE Computer Society, 1999.

[14] REDNER R A. Maximum likelihood estimation for mixture models[J]. Annals of Mathematical Statistics, 1980, 22(3): 583-590.

[15] BISHOP C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]// Springer-Verlag New York. 2006.

[16] HASTIE T, TIBSHIRANI R, FRIEDMAN J, et al. The elements of statistical learning [M]. New York: Springer, 2009.

[17] ZHANG W, YUAN S, WANG J, et al. Real-Time Bidding Benchmarking with iPinYou Dataset [J/OL]. https://arxiv.org/abs/1407/7073.

[18] HAN J, KAMBER M. Data Mining: Concepts and Techniques (Second Edition)[M]. San Francisco, 2006: 1-25.

(上接第 278 页)

通过对 MSLIR 模型的仿真和以往模型的对比以及各个转发概率的比较,可以看出 MSLIR 模型能更准确地描述信息传播规律,反映出信息在微博网络中的传播特性。同时在微博网络中提出不完整阅读行为,通过微博平台调节阅读率,可以有效筛选出危害公共安全的恶意信息,使得积极向上的信息可以得到广泛传播。而新增加的潜伏个体说明信息的广泛传播需要一个潜伏期,对恶意信息可以通过扼杀在潜伏期而有效地控制缩小恶意信息的传播范围。传播的直接免疫情况通过调节免疫概率可以有效遏制恶意信息的传播。以上情况都是传统传播模型未全面考虑的,而 MSLIR 模型通过增加潜伏节点和完善信息的传播途径,可以根据信息随着时间的传播而采取不同的控制措施。

结束语　为了更真实地反映微博网络中信息的传播规律,本文提出了一种更真实地反映微博信息传播过程的 MSLIR 模型。该模型根据现实中微博信息传播的情况,在原有 SIR 模型的基础上增加了粉丝的不完整阅读行为、潜伏节点和直接免疫过程,并把信息传播的途径进行了新的阐述,该模型更适合微博信息的实际传播过程。以新浪微博为例,通过相应的数据模拟和抓取,构建模型的微博信息传播过程,对提出的模型进行了实验仿真,在将其与以往模型对比的同时,分析了模型中各参数的变化对传播过程的影响。仿真结果表明,所提模型可以很好地描述微博信息的传播过程。转发概率 P_1 取值的增大,会增加网络中免疫节点的最终密度。转发概率 P_3 和直接免疫概率 P_2 的变化,都会影响网络达到稳定状态的时间,但不会影响各类节点的最终密度。

本文提出的微博传播模型未考虑看到的粉丝对一个博文事件的多次转发和微博信息传播的空间效果以及可信度等因素,以后的研究应融入这些因素,以进一步完善微博信息传播模型。

参 考 文 献

[1] ZHAO Y R, WANG Y T, WU M Z. Overlapping Community Detection Based on Node-influence Propagation in Heterogeneous Social Networks [J]. Journal of Chinese Computer Systems, 2015, 36(10): 2190-2196. (in Chinese)
赵玉蓉,王轶彤,吴铭泽. 异构社会网络中基于节点影响力传播的可重叠社团发现[J]. 小型微型计算机系统, 2015, 36(10): 2190-2196.

[2] LI L Y, SUN L J, YANG J H. Research on Online Social Network[J]. Computer Science, 2015, 42(11): 8-21. (in Chinese)
李立耀,孙鲁敬,杨家海. 社交网络研究综述[J]. 计算机科学, 2015, 42(11): 8-21.

[3] WANG Y Q, JIANG G P. Virus spreading on complex networks with imperfect immunization [J]. Acta Physica Sinica, 2010, 59(10): 6734-6743. (in Chinese)
王亚奇,蒋国平. 复杂网络中考虑不完全免疫的病毒传播研究[J]. 物理学报. 2010, 59(10): 6734-6743.

[4] SUDBURY A J. The proportion of the population never hearing a rumor[J]. Journal of Applied probability, 1985, 22(2): 443-446.

[5] ZHOU J, LIU Z H, LI B W. Influence of network structure on rumor propagation [J]. Physics Letters A, 2007, 368(6): 458-463.

[6] ZHANG Y C, LIU Y, ZHANG H F, et al. The research of information dissemination model on online social network[J]. Acta Physica Sinica, 2011, 60(5): 60-66. (in Chinese)
张彦超,刘云,张海峰,等. 基于在线社交网络的信息传播模型[J]. 物理学报, 2011, 60(5): 60-66.

[7] ZHOU X F, XU K, ZHANG L S, et al. Propagation Measurement and Cluster Analysis of Time Series in Social Networks [J]. Journal of Chinese Computer Systems, 2015, 36(7): 1545-1552. (in Chinese)
周雪峰,徐恪,张蓝珊,等. 社交网络的传播测量与时间序列聚类分析[J]. 小型微型计算机统, 2015, 36(7): 1545-1552.

[8] SU Q, HUANG J, ZHAO X. An information propagation model considering incomplete reading behavior in microblog [J]. Physica A, 2015, 419(2): 55-63.

[9] DING X, LIU Q C, ZHANG W. An improved model for information dissemination and prediction on micblog nrtworks [J]. Journal of University of Science and Technology of China, 2014, 42(7): 582-598. (in Chinese)
丁鑫,刘其成,张伟. 一种改进的微博网络信息传播与预测模型[J]. 中国科学技术大学学报, 2014, 42(7): 582-598.

[10] ZHAO L, WANG Q, CHENG J, et al. Rumor spreading model with consideration of forgetting mechanism: a case of online blogginglive journal[J]. Physica A, 2011, 390(13): 2619-2625.

[11] ZHAO L, CUI H, QIU X, et al. SIR rumor spreading model in the new media age[J]. Physica A, 2013, 392 (4): 995-1003.

[12] ZHAO L, WANG J, CHEN Y, et al. SIHR rumor spreading model in social networks[J]. Physica A, 2012, 391(7): 2444-2453.

[13] XIONG F, LIU Y, ZHANG Z, et al. An information diffusion model based on retweeting mechanism for online social media [J]. Physics Letters A, 2012, 376(6): 2103-2108.

[14] GU Y R, XIA L L. The propagation and inhibition of rumors in online social network[J]. Acta Physica Sinica, 2012, 61(23): 514-518. (in Chinese)
顾亦然,夏玲玲. 在线社交网络中谣言的传播与抑制[J]. 物理学报, 2012, 61(23): 514-518.

[15] SUN Y C, LIU H, ZHANG G J. Cross platform System for Real-time Crowd Simulation [J]. Journal of Chinese Computer Systems, 2015, 36(4): 862-867. (in Chinese)
孙云晨,刘弘,张桂娟. 支持跨平台的实时人群运动仿真系统[J]. 小型微型计算机系统, 2015, 36(4): 862-867.