文章编号: 1000-5641(2018)03-0077-11

基于特征工程的视频点击率预测算法

匡 俊[1], 唐卫红[2], 陈雷慧[1], 陈 辉[3], 曾 炜[3], 董启民[4], 高 明[1]

(1. 华东师范大学 数据科学与工程学院, 上海 200062;
2. 上海市农业技术推广服务中心, 上海 201103;
3. 深圳腾讯计算机系统有限公司, 北京 100080;
4. 林西县职业技术教育中心, 内蒙古 林西 025250)

摘要: 点击率预测技术在视频推荐系统中具有重要的作用. 视频推荐系统可以根据点击率预测的结果调整投放顺序, 从而提高用户的真实点击率. 在点击率预测问题中, 由于数据存在海量性以及不平衡性等问题, 点击率预测的精确度一般都较低. 针对以上问题, 使用特征工程和机器学习相结合的方法, 有效地改进了现有的视频点击率预测算法的性能. 首先, 使用特征工程方法, 从原始数据中提取特征, 并使用矩阵分解等方法生成交叉特征; 然后, 分别基于逻辑回归、因子分解机和梯度提升决策树-逻辑回归实现点击率预测模型. 实验结果表明, 基于因子分解机模型和基于梯度提升决策树-逻辑回归模型的预测精度要优于基于逻辑回归的模型, 并且将用户特征和视频特征进行交叉组合能够改进点击率预测的精度.

关键词: 点击率预测; 特征工程; 因子分解机; 梯度提升决策树
中图分类号: TP391 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2018.03.009

Algorithm for video click-through rate prediction

KUANG Jun[1], TANG Wei-hong[2], CHEN Lei-hui[1], CHEN Hui[3], ZENG Wei[3], DONG Qi-min[4], GAO Ming[1]

(1. *School of Data Science and Engineering, East China Normal University, Shanghai* 200062, *China*;
2. *Shanghai Agricultural Technology Extension and Service Center, Shanghai* 201103, *China*;
3. *Shenzhen Tencent Computer System Co. Ltd., Beijing* 100080, *China*;
4. *Vocational and Technical Education Center of Linxi County, Linxi Inner Mongolia* 025250, *China*)

Abstract: Click-through rate prediction has played an important role in video recommendation systems. A video recommendation system can suggest media to users based on the results of click-through rate prediction. In this way, users may be more likely to click

收稿日期: 2017-05-19
基金项目: 国家重点研发计划(2016YFB1000905); 国家自然科学基金广东省联合重点项目(U1401256); 国家自然科学基金(61672234, 61502236, 61472321)
第一作者: 匡 俊, 男, 硕士研究生, 研究方向为用户行为分析、点击率预测. E-mail: 15001830063@163.com.
通信作者: 董启民, 男, 中学一级教师, 研究方向为信息处理技术. E-mail: 418976195@qq.com.

the videos recommended by platforms. However, given the volume and imbalance of data in some applications, the accuracy of click-through rate prediction may be very low. To improve the performance, this paper proposes an integrated approach by combining feature engineering with techniques from machine learning. In the first stage, the algorithm uses feature engineering to extract user, video, and combinational features from the original dataset. In the second stage, the algorithm predicts the click-through rate by employing supervised models of logistic regression, factorization machine, and gradient boosting decision tree combined with logistic regression. The experimental results illustrate that the prediction accuracy of the factorization machine model and the gradient boosting decision tree combined with logistic regression model are better than the logistic regression model. Moreover, the cross combination of user and video features can improve the accuracy of the click-through rate prediction.

Keywords: click-through rate prediction; feature engineering; factorization machine; gradient boosting decision tree

0 引 言

随着互联网、物联网和云计算技术的发展, 每时每刻人们都被海量的信息所包围, 而且信息还在呈爆炸式地增长. 因此, 当前我们正处于一个信息严重过载的时代. 人们要从数量庞大并且纷繁复杂的信息中获取到自己所需要的信息, 是一件非常困难的事情. 就观看视频而言, 用户可能会花费大量的时间在寻找自己感兴趣的视频上; 而对于视频服务提供商来说, 如果能将用户感兴趣的视频精确地推送给用户, 可以大大改善用户观影体验, 从而留住更多的用户, 在增加网站流量的同时也增加网站收益. 视频的点击率预测技术正是为了应对上述问题而提出的, 该问题利用用户观影历史行为, 挖掘用户的兴趣和偏好等信息, 从而实现用户观影的预测.

然而, 视频点击率预测是一个极具挑战性的难题. 首先, 大部分用户点击的视频数量普遍较少, 用户历史行为数据中 Positive(点击的)数据远少于 Negative(未点击的)数据, 数据的稀疏性特征造成 Positive 和 Negative 的数据极度不平衡, 这给点击率预测带来很大的挑战. 其次, 用户和视频数据的类型和内容丰富多样, 呈现出异构性的特点, 点击率预测算法需要从纷繁复杂的数据中发掘出对点击率预测精度有积极作用的特征, 这也是一项艰难的问题. 最后, 由于新用户和新视频缺乏历史观看数据, 因此在预测点击率时, 准确度很难达到预期目标, 这种问题称为冷启动问题, 如何应对冷启动也是点击率预测问题中的一个难题.

为了解决上述问题, 本文提出了一种将特征工程和机器学习技术相结合的点击率预测算法. 在机器学习模型方面, 此算法分别基于逻辑回归、因子分解机[1]以及梯度提升决策树-逻辑回归[2-3]实现点击率预测模型. 在特征工程方面, 此算法基于特征选择方法从原始数据中选取出用户和视频的相关特征, 并采用人工组合以及基于矩阵分解两种方式形成用户和视频的交互特征. 另外, 此算法基于聚类方法选取训练数据, 从而优化点击率预测的效果. 在实验部分, 本文使用腾讯用户观看视频的真实行为数据测试本文提出算法的性能, 实验结果表明, 基于因子分解机模型和基于梯度提升决策树-逻辑回归模型的预测精度要优于基于逻辑回归的模型, 并且将用户特征和视频特征进行交叉组合能够改进点击率预测的精度.

1　相关工作

点击率预测等相关问题的研究已取得了大量的进展[4]. Richardson 等人提出了基于逻辑回归模型实现点击率预测算法[5], 这种方法较为简单, 易于实现, 且时间复杂度低, 但在稀疏的数据下效果不够好, 预测的精确度较低; Chapelle 等人提出了基于动态贝叶斯网络的点击率预测模型[6]; Graepel 等人提出了在线贝叶斯概率回归模型[7]. 以上这些方法主要是针对用户的行为进行建模, 有利于个性化的推荐. 但是基于贝叶斯方法的模型存在无法处理新用户和新视频的问题, 并且在处理稀疏数据时表现也不够好. 基于 SVM 模型[8]的好处是可以处理多维非线性的数据, 但在处理稀疏数据时表现不够好, 点击率预测精度较低. Shan 等人提出了一种基于张量分解的点击率预测模型[9], 该模型在处理稀疏数据时的表现相对较好, 但更新张量分解模型所需的时间较长, 且无法并行化; Yan 等人提出了一种基于 Group Lasso 的点击率预测模型[10], 该模型可以自动建立特征之间的关联, 在应对稀疏数据时表现较好, 但是计算复杂度较高; Angarwal 等人将普通的逻辑回归算法进行改进, 实现了名为“LASER”的点击率预测系统[11], 该系统在应对冷启动问题时表现良好, 但在处理稀疏数据时表现较差; 文献 [12] 提出了一种利用视频点击流数据预测用户行为的方法, 该方法从点击流数据中提取特征并进行特征选择, 利用经过筛选的特征预测用户行为; 文献 [13] 基于卷积神经网络构建模型, 预测搜索广告的点击率.

2010 年, Rendle 提出了因子分解机[1](Factorization Machine, FM) 模型, 该模型能够缓解数据稀疏性对预测模型的影响, 在数据稀疏的情况下表现良好. 2014 年, 文献 [3] 提出了一种将梯度提升决策树和逻辑回归融合的模型, 梯度提升决策树模型利用提升 (Boosting) 的方法[14]集成多棵决策树. 决策树[15]算法的特点是能高效地处理非线性的问题, 但是容易出现过拟合. 梯度提升决策树并不要求在单棵决策树上达到最优解, 而是集成多棵决策树的结果, 使集成的结果达到最优, 这种方法解决了过拟合的问题. 本文所提出的点击率预测算法, 就是基于以上两种模型实现的.

2　算法主要框架

本文提出的视频点击率预测算法主要包括特征工程和学习模型构建两部分, 其中特征工程是从原始数据中提取特征, 而学习模型构建旨在基于特征工程提取的特征, 运用监督学习框架训练点击率预测模型. 本文提出的算法的主要框架如图 1 所示.

原始数据中数据的格式复杂, 冗余数据多, 无法将原始数据直接应用到点击率预测算法中. 因此, 视频点击率预测算法首要解决的是从原始数据中获取特征. 本文算法使用特征工程方法从原始数据提取特征: 从用户行为数据中提取出用户特征, 包括用户自身属性、用户的兴趣偏好以及用户所使用的设备属性; 同时从视频数据中提取出视频特征, 包括视频的基本信息以及视频的类型特征. 交叉特征对于视频点击率预测的精度具有重要意义, 在提取出用户特征和视频特征之后, 将两者进行交叉组合形成交叉特征.

对于一些非数值型或者无数值意义的特征, 需要进行特征编码操作. 特征编码会将一个特征离散到多个维度, 使编码后的每一维特征都具有数值上的意义. 通过特征工程处理得到的特征数量如表 1 所示.

本文基于监督学习模型实现点击率预测算法, 并且由于点击率预测问题实质上是一个二分类问题, 即预测用户点击或者不点击视频. 因此需要构建 Label 为 1 的正例样本(用户点击视频的数据), 以及 Label 为 0 的负例样本(用户不点击视频的数据), 来训练监督学习模型.

正例数据可以直接从用户观看视频的历史记录中获取. 然而, 由于单个用户只会点击少量的视频, 对于单个用户来说, 绝大多数视频都是未点击的, 但并不表示用户一定不会点击这些视频. 因此, 需要从大量未点击的视频中挑选出具有代表性的视频用来产生负例训练样本. 若采用随机抽样的方法, 挑选出的视频可能不具有代表性. 针对这个问题, 本算法使用基于聚类[16]的方法选择用户未点击视频中的一部分, 用于产生负例数据. 具体方法将在第 3 节中详细说明.

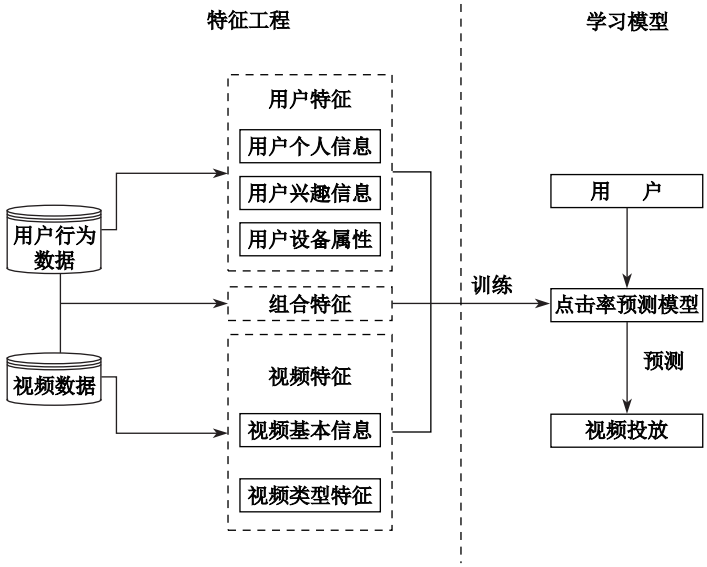

图 1 视频点击率预测算法框架

Fig. 1 Framework of algorithm for video click-through rate prediction

表 1 特征数量

Tab. 1 Number of features

特征		特征数量
用户特征	用户自身属性	4
	用户兴趣偏好	460
	用户设备属性	28
视频特征	视频类型特征	99
	视频基本信息	278
组合特征	交叉特征	8

离线训练好模型后, 点击率预测模型可以在线预测每个到访用户对每一个视频的点击概率, 并依据预测结果, 为用户推荐其最有可能点击的视频.

3 视频点击率预测算法

3.1 特征工程

特征工程是本文算法的首要任务, 也是算法的重点. 本文算法以各个特征的重要性为依据选择特征. 特征重要性使用袋外数据 (Out Of Bag, OOB) 误差作为度量指标[17]. 首先, 基于 Bagging 方法构建 N 棵决策树[18], 对每一棵决策树, 计算其袋外数据误差 errOOB1. 然后对每一个特征加入噪声干扰后, 再次计算每一棵决策树的袋外数据误差 errOOB2. 特征的重要性为

所有决策树 errOOB2 与 errOOB1 差值的平均值, 特征 x 重要性的表达式为

$$F(x)=\frac{\sum_{i=1}^{N}(\text{errOOB2}_i(x)-\text{errOOB1}_i(x))}{N}. \tag{1}$$

使用上述特征选择方法, 从原始数据中提取出对点击率预测重要的相关特征, 如表 2 所示.

表 2　用户和视频特征

Tab. 2　User and video features

特征类型		特征内容
用户特征	用户自身属性	性别、年龄、学历
	用户兴趣偏好	兴趣标签、视频类型、导演、演员、地区
	用户设备属性	CPU核数、内存、默认浏览器、共存安全软件
视频特征	视频类型特征	大类类型、子类型
	视频基本信息	演员、导演、上映年份、地区

对于一些具有数值意义的特征, 例如用户活跃天数和用户年龄, 可以直接从原始数据中切分出来; 但是对于一些无数值意义的特征, 例如视频大类的编号(编号在数值上的大小没有意义), 或者字符串类型的特征, 例如演员、导演等, 需要对特征进行编码操作. 特征编码采用一种类似于 One-Hot Encoding[19]的操作. 以本文所使用的数据集中用户感兴趣的演员标签为例: 某用户感兴趣的演员在原始数据中如表 3 所示, 其中, 用户感兴趣的演员用“#”隔开, “%”前面是演员名字, 后面是该用户喜欢该演员的权重值.

表 3　用户感兴趣的演员标签

Tab. 3　Tags for actors of user-interest

演员标签
海清%27#陈数%24#邬君梅%38#王源%54#易烊千玺%18#胡先煦%41#汪俊%34

对于这种类型的特征, 首先需要预处理 popular 的值, 例如最受欢迎的 100 位演员. 然后将这 100 个演员中的每一个当作特征向量中的一维特征, 特征的值为每个用户对该演员的喜爱程度. 按照这种编码方式, 表3中的演员标签将被编码成表 4 所示的形式.

表 4　用户感兴趣的演员编码

Tab. 4　Encoding of tags for actors of user-interest

唐嫣	吴建豪	···	王源	···	易烊千玺	···
0	0	0	54	0	18	0

除了用户特征和视频特征, 用户和视频的组合特征对视频点击率的预测也是至关重要的. 本文产生组合特征的方式为人工组合特征和基于矩阵分解产生组合特征.

(1) 人工组合交叉特征. 可以将用户的兴趣偏好和视频的特征交叉生成新的特征. 例如, 用户的演员偏好特征和视频的演员都经过特征编码, 形成了两个维数相同, 并且相应维度含义也相同的向量. 通过计算这两个向量的相似度, 可以得到用户对于视频中演员喜欢程度的“综合分数”, 该“综合分数”可以作为一个新的组合特征.

(2) 基于矩阵分解的组合特征. 根据用户观看视频的记录生成用户观看视频时长的 $n\times m$ 的矩阵, 其中 n 表示用户数量, m 表示视频数量. 将用户-视频矩阵分解成用户矩阵($n\times k$) 和视频矩阵 ($k\times m$). 通过这种方式将用户和视频映射到一个 k 维的隐空间中. 如图 2 所示, 每个用户对应用户矩阵中的一行, 每个视频对应视频矩阵中的一列, 用户和视频均可以表示为 k 维向量. 通过计算用户隐向量和视频隐向量之间的相似度, 可以发掘出用户和视频之间的潜在关联.

因此可以将隐向量之间的相似度作为组合特征.

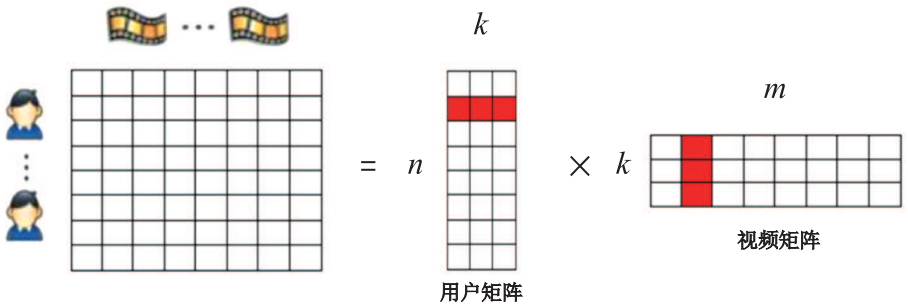

图 2 基于矩阵分解生成组合特征

Fig. 2 Schematic diagram for generating new features based on matrix factorization

3.2 模型的训练和测试

分别基于逻辑回归、梯度提升决策树-逻辑回归和因子分解机建立预测模型, 其中, 逻辑回归的优势是模型简单、易于实现且效率高. 给定特征 x, 用户点击某个视频的概率, 即

$$P(y=1|x)=\frac{1}{1+\mathrm{e}^{-g(x)}}, \tag{2}$$

其中 $g(x)$ 为特征的线性函数, 表示为

$$g(x)=\omega_0+\omega_1x_1+\omega_2x_2+\omega_3x_3+\cdots+\omega_px_p=\theta^{\mathrm{T}}x. \tag{3}$$

梯度提升决策树是一种利用梯度提升方法集成多棵决策树的算法. 其主要特点是能够处理高维非线性的数据, 并且解决了普通决策树算法容易出现过拟合的问题. 其表达式为

$$f_M(x)=\sum_{m=1}^{M}T(x;\Theta_m), \tag{4}$$

其中, M 为决策树的棵数, T 表示单棵决策树. 梯度提升决策树算法每一轮学习一棵决策树 T_m, 树 T_m 是通过拟合上一轮 T_{m+1} 的残差(即损失函数的负梯度值)来学习. 每一轮学习过后, 将新学习到的决策树加入表达式 (3) 所示的加法模型中.

梯度提升决策树-逻辑回归融合模型是一种将梯度提升决策树和逻辑回归相结合的一种模型, 通过梯度提升决策树, 可以发掘某些特征之间的潜在关系, 因此可以自动对某些特征进行交叉组合. 利用梯度提升决策树的这一特点, 使用梯度提升决策树生成逻辑回归模型的输入特征. 梯度提升决策树模型中, 每一棵决策树的每一个叶子节点维护逻辑回归的一维输入特征. 对于每个样本, 只需找到其在梯度提升决策树中的路径, 就能将该样本转化为逻辑回归的输入特征.

因子分解机模型的特点是能够缓解数据的稀疏性带来的影响, 能自动发掘特征之间的关联. 与逻辑回归类似, 对输入特征建立模型

$$\Phi(\omega,x)=\omega_0+\sum_{i=1}^{n}\omega_ix_i+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\omega_{ij}x_ix_j. \tag{5}$$

式 (5) 中等号右边前两项与逻辑回归类似, 为普通的线性模型部分, 最后一项为交叉项部分, x_i 和 x_j 分别代表第 i 维特征和第 j 维特征. 所有的交叉项系数 ω_{ij} 可以组成一个对称矩阵 $W(n\times n)$, 将矩阵 $W(n\times n)$ 分解成矩阵 $V(n\times k)$ 和矩阵 $V^{\mathrm{T}}(k\times n)$ 的乘积, 矩阵 V 的第 j 行

向量为第 j 维特征的隐向量. ω_{ij} 等于隐向量 v_i 和隐向量 v_j 的内积. 于是对于 ω_{ij} 的学习变成了对隐向量 v_i 和 v_j 的学习. 因子分解机模型可以表示为

$$\Phi(\omega, x) = \omega_0 + \sum_{i=1}^{n} \omega_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < v_i, \quad v_j > x_i x_j. \tag{6}$$

可以注意到, 对于隐向量 v_i 来说, 只要存在任意的 j 使 $x_i x_j$ 不为 0, 均能够学习 v_i. 因此, 这种方法大大地缓解了数据稀疏性带来的影响.

3.3　训练数据的生成

模型建立好之后, 需要进行训练. 如何产生训练样本也是点击率预测算法需要考虑的问题. 原始数据集中含有用户点击视频的历史记录, 因此可以直接用来产生正例数据. 但是负例数据的产生则要复杂很多. 由于单个用户点击的视频数量极少, 因此负例数据的数量庞大, 远多于正例数据, 而且用户未点击的视频并不意味着用户不会点击这些视频. 因此, 不能将所有未点击的视频作为负例样本用于训练. 为了解决这个问题, 需要从未点击的视频中挑选出一部分最可能成为负例样本用于模型训练.

本文使用一种基于聚类的方法挑选负例样本. 首先, 提取视频特征向量; 然后根据视频特征向量的欧氏距离将视频聚类, 分成 k 组, 根据聚类算法的原理可知, 每个聚类中的视频都有较强的相似性; 最后, 对于每个用户, 统计其点击每个类中视频的数量, 点击的视频数量少, 说明该用户对该种类型的视频不太感兴趣, 点击该聚类中的视频可能性较小. 因此, 在生成负例样本时, 要从每个用户点击视频数量最少的几个类中, 选择该用户未点击的视频. 通过这种方式产生的负例样本具有代表性, 能在一定程度上提升视频点击率预测的精度.

3.4　视频推荐

为了测试点击率预测算法的效果, 我们利用算法的预测结果为用户推荐视频: 对于每个用户, 使用训练好的模型预测该用户对每个视频的点击概率, 然后将该用户最有可能点击的视频推荐给该用户, 并通过用户点击视频的记录计算推荐的成功率. 推荐的流程如图 3 所示, 当用户访问视频网站时, 获取用户特征、视频特征和它们的组合特征, 计算用户点击视频的概率, 根据预测的点击概率, 向用户推荐视频.

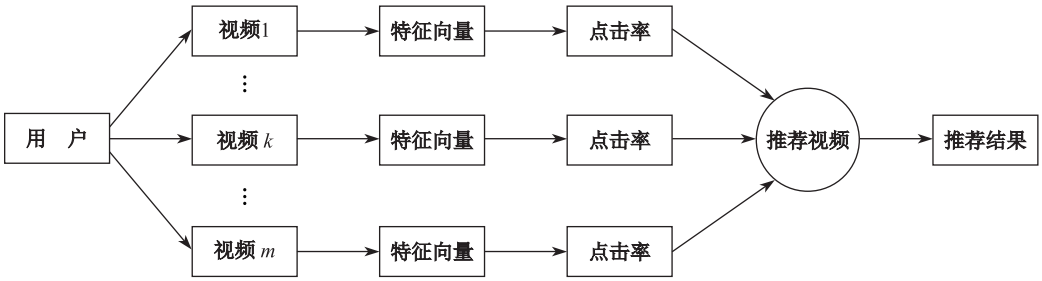

图 3　为用户推荐视频

Fig. 3　Workflow for recommending videos to users

4　实　　验

4.1　实验数据集

本实验使用的数据集为腾讯用户观看视频的真实行为数据, 其中包含 147 213 个用户和 67 027 个视频的数据, 用户数据中包括新注册的用户. 此外, 数据集中还包括 30 849 964 条用

户观看记录. 将其中观看时长不足 5 min 的无效观看记录去除后, 还剩下 19 389 876 条观看记录.

4.2 评价指标

分别使用精度 (Precision)、召回率 (Recall)、F1-Score、对数损失 (Log loss)、AUC (Area UnderCurve), 以及时间效率作为算法的评价指标. 精度的定义为, 点击率预测算法预测用户会点击的视频中, 被用户真正点击的视频比例, 即

$$P = \frac{TP}{TP + FP}, \tag{7}$$

其中, TP 为点击率预测算法预测用户会点击, 且用户真正点击的视频数量, FP 为点击率预测算法预测用户会点击, 但用户没有点击的视频数量. 另外, 令 FN 为点击率预测算法预测用户不点击, 但用户真实点击的视频数量.

召回率为用户实际点击的视频中被预测正确的视频比例, 即

$$R = \frac{TP}{TP + FN}. \tag{8}$$

F1-Score 为综合考虑精度和召回率的一个指标, 其表达式为

$$F1 = \frac{2 \times P \times R}{P + R}. \tag{9}$$

对数损失的计算公式为

$$\text{Log loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)), \tag{10}$$

其中, p_i 为预测的点击概率, y_i 为用户点击与否的真实值. 从式 (10) 可以看出, 对数损失越小, 预测结果越精确.

4.3 结果与分析

通过将本文算法的预测结果与朴素贝叶斯 (Naive Bayes, NB) 算法和决策树 (Decision Tree, DT) 算法的预测结果对比来验证本文算法的效果. 为了验证加入组合特征的作用, 我们分别在加入组合特征和不加入组合特征两种情况下测试算法的性能.

在实际应用中, 单个用户点击视频的数量普遍较少, 数据的正负样本比例会极度不平衡, 因此我们分别使用正负样本比例为 1∶300 和 1∶2 000 的测试数据集测试算法效果.

训练集正负样本比例为 1∶10, 测试集正负样本比例为 1∶300, 仅使用用户特征和视频特征而不加入组合特征. 基于因子分解机 (FM) 的点击率预测模型、基于梯度提升决策树-逻辑回归 (GBDT+LR) 的点击率预测模型和基于逻辑回归 (LR) 的点击率预测模型的预测效果如表 5 所示. 表 5 中还包括相同训练数据和测试数据下, 朴素贝叶斯 (NB) 算法和决策树 (DT) 算法的预测效果.

表 5 实验结果1

Tab. 5 Result 1

模型及算法	Precision	Recall	F1-Score	Log loss	AUC	Time/ms
FM	0.13	0.64	0.22	0.130 4	0.862	57 968
GBDT+LR	0.09	0.74	0.16	0.134 7	0.883	4 231
LR	0.08	0.14	0.10	0.204 2	0.682	409
NB	0.01	0.21	0.01	2.441 7	0.626	1 624
DT	0.01	0.77	0.01	15.121	0.666	3 471

图 4 为本文 3 种模型以及朴素贝叶斯 (NB) 算法和决策树 (DT) 算法预测结果的 ROC(Receive Operation Characteristic) 曲线.

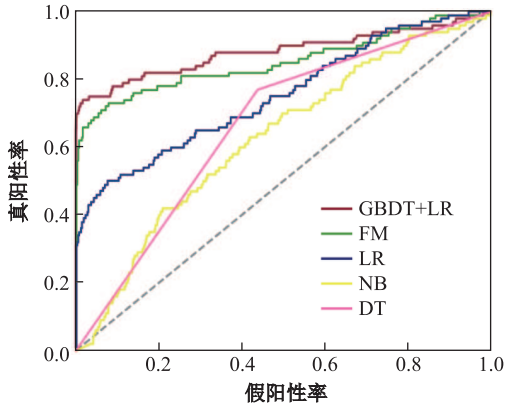

图 4　ROC曲线一

Fig. 4　ROC curve 1

训练集正负样本比例为 1∶10, 测试集正负样本比例为 1∶2 000, 仅使用用户特征和视频特征而不加入组合特征. 基于因子分解机 (FM) 的点击率预测模型、基于梯度提升决策树-逻辑回归 (GBDT+LR) 的点击率预测模型和基于逻辑回归 (LR) 的点击率预测模型的预测效果如表 6 所示. 表 6 中还包括相同训练数据和测试数据下, 朴素贝叶斯 (NB) 算法和决策树 (DT) 算法的预测效果.

表 6　实验结果2

Tab. 6　Result 2

模型及算法	Precision	Recall	F1-Score	Log loss	AUC	Time(ms)
FM	0.10	0.81	0.18	0.102 0	0.941	393 332
GBDT+LR	0.09	0.68	0.16	0.166 6	0.861	275 293
LR	0.02	0.35	0.04	0.215 8	0.803	2 628
NB	0.001	0.19	0	1.445 5	0.654	5 009
DT	0.001	0.77	0	15.13	0.668	33 295

本文 3 种模型以及朴素贝叶斯 (NB) 算法和决策树 (DT) 算法预测结果的 ROC 曲线如图 5 所示.

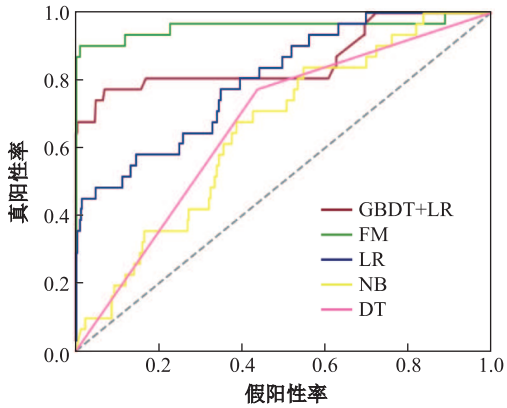

图 5　ROC曲线二

Fig. 5　ROC curve 2

从以上两组实验可以看出, 本文所提出的点击率预测算法表现良好. 基于梯度提升决策树-逻辑回归的模型和基于因子分解机的模型, 在点击率预测效果方面要优于基于逻辑回归的模型, 但基于逻辑回归的算法时间复杂度较低. 通过对比两组实验的结果发现, 面对正负样本比例不平衡的数据时, 很多负例样本会被预测成正例, 这会导致如下情况: 尽管召回率较高, 但是精度和 F1-Score 会很低. 这说明在正负样本极度不平衡的情况下, 点击率预测是一个非常困难的问题. 本文算法在正负样本不平衡的情况下依然能够有较好的表现.

在用户特征和视频特征的基础上, 加入用户和视频的组合特征. 在训练集正负样本比例为 1∶10, 测试集正负样本比例为 1∶300 的情况下, 基于因子分解机 (FM) 的点击率预测模型、基于梯度提升决策树-逻辑回归 (GBDT+LR) 的点击率预测模型和基于逻辑回归 (LR) 的点击率预测模型的预测效果如表 7 所示.

表 7 实验结果3

Tab. 7 Result 3

模型	Precision	Recall	F1-Score	Log loss	AUC	Time/ms
FM	0.16	0.90	0.28	0.162 3	0.979	53 065
GBDT+LR	0.14	0.94	0.24	0.139 7	0.987	4 154
LR	0.10	0.93	0.19	0.238 4	0.971	365

通过对比实验结果 3 和实验结果 1 发现, 加入组合特征后, 基于 3 种不同模型的算法效果都有了显著的提升. 由此可知, 在点击率预测问题中, 特征工程非常关键, 尤其是特征之间的交叉组合, 对点击率预测算法的效果有十分重要的影响.

基于本文所提出算法的点击率预测结果, 为每个用户推荐其最有可能点击的视频. 分别为点击视频的总次数大于 20、30、40 和 60 次的用户推荐视频, 推荐的成功率如表 8 所示.

表 8 推荐成功率

Tab. 8 Accuracy of recommendations

点击视频次数	推荐成功率
>20	0.03
>30	0.022
>40	0.07
>60	0.08

我们对特征进行重要性分析, 发现组合特征、用户兴趣偏好以及视频类型对视频点击率预测较为重要, 并且预测不同类型视频的点击率, 各个特征的重要性也不相同. 预测电影的点击率时, 导演特征和演员特征的重要性基本相同; 而预测电视剧的点击率时, 导演特征的重要性远低于演员特征的重要性. 基于以上分析, 我们将电影和电视剧从视频数据中分离出来, 分别做推荐, 推荐的成功率如表 9 所示.

表 9 推荐电影和推荐电视剧的成功率

Tab. 9 Accuracy of movie and TV recommendations

点击视频次数	推荐成功率	
	电影	电视剧
>20	0.02	0.05
>30	0.013	0.018
>40	0.07	0.08
>60	0.086	0.085

从表 9 可以看出, 将电影和电视剧分别做推荐, 成功率略有提高.

5 总　　结

本文基于特征工程和 3 种机器学习的模型实现了一种视频点击率预测算法. 算法利用特征工程的方法提取用户特征和视频特征, 并将用户特征和视频特征交叉组合, 且分别基于逻辑回归、梯度提升决策树-逻辑回归和因子分解机实现点击率预测模型. 本文算法在正负例样本不平衡且数据较为稀疏的情况下依然表现良好. 未来可以在特征工程方面进行更多的工作, 从而提高视频点击率预测算法的性能.

[参 考 文 献]

[1] RENDLE S. Factorization machines[C]// IEEE International Conference on Data Mining. IEEE Computer Society, 2010: 995-1000.

[2] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[3] 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. ACM, 2014: 1-9.

[4] 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): 1-14.

[5] RICHARDSON M, DOMINOWSKA E, RAGNO R. Predicting clicks: Estimating the click-through rate for new ads[C]// International Conference on World Wide Web. ACM, 2007: 521-530.

[6] CHAPELLE O, ZHANG Y. A dynamic bayesian network click model for web search ranking[C]// International Conference on World Wide Web. ACM, 2009: 1-10.

[7] GRAEPEL T, CANDELA J Q, BORCHERT T, et al. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing Search engine[C]// International Conference on Machine Learning. DBLP, 2010: 13-20.

[8] JOACHIMS T. Optimizing search engines using click-through data[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002: 133-142.

[9] SHAN L, LIN L, SUN C, et al. Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization[J]. Electronic Commerce Research & Applications, 2016, 16(C): 30-42.

[10] YAN L, LI W J, XUE G R, et al. Coupled group lasso for web-scale CTR prediction in display advertising[C]// International Conference on Machine Learning. 2014: 802-810.

[11] AGARWAL D, LONG B, TRAUPMAN J, et al. LASER: A scalable response prediction platform for online advertising[C]// ACM International Conference on Web Search and Data Mining. ACM, 2014: 173-182.

[12] AQUIAR E, NAGRECHA S, CHAWLA N V. Predicting online video engagement using clickstreams[C]//IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2015. DOI: 10.1109/DSAA.2015.7344873.

[13] 李思琴, 林磊, 孙承杰. 基于卷积神经网络的搜索广告点击率预测[J]. 智能计算机与应用, 2015(5): 22-25.

[14] SCHAPIRE R E. A brief introduction to boosting[C]// 16th International Joint Conference on Artificial Intelligence. [S.l.]: Morgan Kaufmann Publishers Inc, 1999: 1401-1406.

[15] QUINLAN J R. Induction on decision tree[J]. Machine Learning, 1986(1): 81-106.

[16] HARTIGAN J A, WONG M A. Algorithm AS 136: A k-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100-108.

[17] BREIMAN L. Out-of-bag estimation[R]. Berkeley: University of California, 1996.

[18] BREIMAN L. Bagging Predictors[M]. [S.l.]: Kluwer Academic Publishers, 1996.

[19] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]// ACM SIGKDD International Conference. ACM, 2016:785-794.

(责任编辑: 李　艺)