

CAT-RFE: 点击欺诈的集成检测框架

卢翼翔¹, 耿光刚¹, 延志伟², 朱效民³, 张新常⁴

(1. 暨南大学网络空间安全学院, 广东 广州 510632; 2. 中国互联网络信息中心, 北京 100190;
3. 山东齐鲁大数据研究院, 山东 济南 250001; 4. 山东省科学院, 山东 济南 250001)

摘要: 点击欺诈是近年来最常见的网络犯罪手段之一, 互联网广告行业每年都会因点击欺诈而遭受巨大损失。为了能够在海量点击中有效地检测欺诈点击, 构建了多种充分结合广告点击与时间属性关系的特征, 并提出了一种点击欺诈检测的集成学习框架——CAT-RFE 集成学习框架。CAT-RFE 集成学习框架包含3个部分: 基分类器、递归特征消除(RFE, recursive feature elimination)和 voting 集成学习。其中, 将适用于类别特征的梯度提升模型——CatBoost (categorical boosting) 作为基分类器; RFE 是基于贪心策略的特征选择方法, 可在多组特征中选出较好的特征组合; Voting 集成学习是采用投票的方式将多个基分类器的结果进行组合的学习方法。该框架通过 CatBoost 和 RFE 在特征空间中获取多组较优的特征组合, 再在这些特征组合下的训练结果通过 voting 进行集成, 获得集成的点击欺诈检测结果。该框架采用了相同的基分类器和集成学习方法, 不仅克服了差异较大的分类器相互制约而导致集成结果不理想的问题, 也克服了 RFE 在选择特征时容易陷入局部最优解的问题, 具备更好的检测能力。在实际互联网点击欺诈数据集上的性能评估和对比实验结果显示, CAT-RFE 集成学习框架的点击欺诈检测能力超过了 CatBoost 模型、CatBoost 和 RFE 组合的模型以及其他机器学习模型, 证明该框架具备良好的竞争力。该框架为互联网广告点击欺诈检测提供一种可行的解决方案。

关键词: 点击欺诈检测; 类别梯度提升; 递归特征消除; 集成学习

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022065

CAT-RFE: ensemble detection framework for click fraud

LU Yixiang¹, GENG Guanggang¹, YAN Zhiwei², ZHU Xiaomin³, ZHANG Xinchang⁴

1. College of Cyber Security, Jinan University, Guangzhou 510632, China

2. China Internet Network Information Center, Beijing 100190, China

3. Shandong Institute of Big Data, Jinan 250001, China

4. Shandong Academy of Sciences, Jinan 250001, China

收稿日期: 2021-09-23; 修回日期: 2022-01-05

通信作者: 耿光刚, gggeng@jnu.edu.cn

基金项目: 国家自然科学基金(92067108); 广东省自然科学基金(2021A1515011314)

Foundation Items: The National Natural Science Foundation of China (92067108), The Natural Science Foundation of Guangdong Province (2021A1515011314)

引用格式: 卢翼翔, 耿光刚, 延志伟, 等. CAT-RFE: 点击欺诈的集成检测框架[J]. 网络与信息安全学报, 2022, 8(5): 158-166.

Citation Format: LU Y X, GENG G G, YAN Z W, et al. CAT-RFE: ensemble detection framework for click fraud[J]. Chinese Journal of Network and Information Security, 2022, 8(5): 158-166.

Abstract: Click fraud is one of the most common methods of cybercrime in recent years, and the Internet advertising industry suffers huge losses every year because of click fraud. In order to effectively detect fraudulent clicks within massive clicks, a variety of features that fully combine the relationship between advertising clicks and time attributes were constructed. Besides, an ensemble learning framework for click fraud detection was proposed, namely CAT-RFE ensemble learning framework. The CAT-RFE ensemble learning framework consisted of three parts: base classifier, recursive feature elimination (RFE) and voting ensemble learning. Among them, the gradient boosting model suitable for category features-CatBoost was used as the base classifier. RFE was a feature selection method based on greedy strategy, which can select a better feature combination from multiple sets of features. Voting ensemble learning was a learning method that combined the results of multiple base classifiers by voting. The framework obtained multiple sets of optimal feature combinations in the feature space through CatBoost and RFE, and then integrated the training results under these feature combinations through voting to obtain integrated click fraud detection results. The framework adopted the same base classifier and ensemble learning method, which not only overcame the problem of unsatisfactory integrated results due to the mutual constraints of different classifiers, but also overcame the tendency of RFE to fall into a local optimal solution when selecting features, so that it had better detection ability. The performance evaluation and comparative experimental results on the actual Internet click fraud dataset show that the click fraud detection ability of the CAT-RFE ensemble learning framework exceeds that of the CatBoost method, the combined method of CatBoost and RFE, and other machine learning methods, proving that the framework has good competitiveness. The proposed framework provides a feasible solution for Internet advertising click fraud detection.

Keywords: click fraud detection, CatBoost, recursive feature elimination, ensemble learning

0 引言

广告是互联网最主要的商业模式,已经逐步形成互联网广告产业。近年来,整个产业的规模在持续快速增长。互联网用户打开网页或者移动手机应用都能看到各式各样的广告。广告中最常见的一种付费模式是点击付费(pay per click),即由广告商(advertiser)提供广告链接,发布者(publisher)可以在自己的网站或应用中发布该广告链接,以此来赚取广告商的点击费用^[1]。点击付费商业模式简单,只通过点击就能产生收入,且广告点击欺诈的惩罚风险相对较低,这让许多不法发布者有了可乘之机^[2]。

点击欺诈是近年来最常见的网络犯罪手段之一,互联网广告行业每年都会因为点击诈骗而遭受巨大的损失。为了减少广告商在点击付费模式中点击欺诈而造成的巨大损失,同时减少点击欺诈对网络和商业环境的不良影响,设计一种能够在海量点击中有效检测出欺诈点击的方法尤为重要。

本文针对广告点击欺诈检测问题,提出多种构建特征的方法,并探索一种适用于该问题的

CAT-RFE 集成学习框架。本框架使用 CatBoost (categorical boosting)^[3]模型与递归特征消除(RFE, recursive feature elimination)方法,在构建好的特征中选取多组较优的特征组合,将 CatBoost 作为基分类器,对这些特征组合的数据训练后进行 voting 集成。

本文提出的集成框架,在特征空间中探索多组较优的特征组合,同时将每组特征组合通过基分类器进行集成。基分类器拟采用梯度提升模型,使集成框架在点击欺诈检测中尽可能发挥出最大的优势。与现有工作不同,该框架综合考虑特征和模型,将特征的选取融入模型中,成为框架中的一部分,结合多组较优的特征组合,避免仅选取局部最优特征组合而导致降低检测精度。同时,该框架结合当前点击欺诈检测方面的最优的机器学习模型,即对梯度提升模型进行集成。为了避免多类基分类器性能的参差不齐而导致无法提升集成模型的检测能力,该框架对多个相同的梯度提升模型进行集成,以确保集成模型的稳定性。

在实际大规模互联网广告点击数据集上的对比实验显示,本文框架优于基线模型、CatBoost 模

型、CatBoost 与 RFE 组合的模型以及其他经典机器学习模型,证明了所提框架具备良好的竞争力。

1 相关工作

点击欺诈作为互联网广告面临的最严重的威胁之一,如何有效和精准地在海量的广告点击中检测出欺诈点击成为学术界和工业界广泛关注的问题。结合 Gohil 等^[2]对点击欺诈检测研究的观点,将相关检测方法分为如下4类。

1) 流量分析:该方法对批量的广告点击流量进行分析,即根据每个用户的点击流量或每个时间段的点击流量,找出不同于正常点击流量的点击数据。例如, Nagaraja 等^[4]针对时间序列点击流量对点击欺诈行为进行检测。

2) 数据挖掘:该方法针对广告点击数据挖掘出某种欺诈规则,并根据该规则识别未知点击中的欺诈数据。例如, Gabryel^[5]使用其改进的 TF-IDF 词频统计方法,对点击产生的数据构造出 TF-IDF 矩阵,并进一步使用 K 近邻分类器检测欺诈的点击。

3) 机器学习:根据广告点击产生的数据构建出相应的特征,再使用机器学习或深度学习分类器通过这些特征识别欺诈点击。例如, Mouawi 等^[6]利用有些广告发布者会诱导用户点击感兴趣的广告链接的特点,构造出相应的特征训练机器学习分类算法;董亚楠等^[7]采用“Fisher 分”算法选取重要的用户行为特征,实现了在模型精度基本不变的情况下,加快模型的训练和检测速度;Taneja 等^[8]将 RFE 方法与海灵格距离决策树(HDDT)结合,用于不平衡数据集的检测;张欣等^[9]使用了 Boosting-SVM 集成模型,解决了 SVM 在点击欺诈数据集中产生的过拟合现象;有研究者使用随机森林模型,说明其检测精度比 SVM 和逻辑回归等模型高^[10-11];Perera 等^[12]使用不同的机器学习模型,构造出6种集成学习方案,实验说明在点击欺诈中集成学习模型比单一模型更有效;另外,有不少研究者使用 xgboost 和 LightGBM 等梯度提升模型^[13-16],取得了比一般机器学习模型更为优异的效果;Thejas 等结合级联森林和 xgboost^[17],取得了比单一梯度提升模型更好的效果。

4) 蜜罐技术:广告商给发布者批量投放广告

链接时,可在其中增加一些“虚假”的广告链接,利用这种链接来判断欺诈的点击。通常真实用户不会点击“虚假”的广告链接,点击了这种广告链接的用户极有可能是软件或程序模拟的用户,进一步可认为该用户的所有点击均为欺诈的点击。例如, Haddadi 等^[18]将展示给用户的广告以一定的概率替换为“虚假”广告,以此来检测“虚假”用户。

综上所述,流量分析方法利用广告点击流量对批量的点击进行分析,虽然流量数据获取方便且检测方法简便,但该方法使用的数据较为单一,只能涵盖多数欺诈点击的规律,容易被不法发布者找到点击流量的规律;数据挖掘方法提取出的欺诈规则基本上是欺诈点击的通用规则,既无法涵盖特殊欺诈点击的特性,也容易被欺诈者规避;蜜罐技术利用“虚假”广告链接来分辨欺诈的点击,该方法易于实现,然而不法发布者可针对真实用户的点击进行分析,辨识真实的广告链接,从而控制程序只点击真实广告链接以逃避检测。

相较于流量分析、数据挖掘和蜜罐技术3类检测方法,机器学习方法在点击欺诈方面具有更好的检测能力、可适应性和鲁棒性^[2]。机器学习方法不单单根据一种或少量几种点击数据来辨认欺诈的点击,而是由多种数据构建而成的特征数据,这些特征数据由于种类较多,易于涵盖更多的欺诈点击特性。由于机器学习模型运用的点击数据种类较多,不法发布者很难找到某些规律来躲避点击欺诈检测。机器学习对于复杂数据的处理能力比其他方法相对较优,因此检测能力比其他方法更好。

根据以上综述的机器学习方法的文献可知,有的研究者专注于点击特征的构建和选取,表明了所提特征的有效性,同时证明了特征的提取或选取是至关重要的一步;有的研究者则专注于机器学习模型的选取或构造,研究结果证明了集成和梯度提升模型的检测能力优于普通机器学习模型。虽然上述提到的研究工作在提高点击欺诈检测能力的方面取得一定的效果,但每种方法都缺少将特征与模型结合的考虑,使模型很难进一步提升检测精度。此外,虽然在该问题下已存在梯度提升模型的集成模型^[17],但现有的模型是集成

了多种不同的梯度提升模型,这容易导致各种梯度提升模型之间相互制约,从而达不到更好的检测效果。因此,探索一种特征选取与检测模型结合的方法,尽可能提高点击欺诈检测精度,显得必要且有意义。

2 提出的集成架构

高精度点击欺诈检测,需要综合考虑特征或者模型两个层面。本文提出了多类点击欺诈特征的提取思路,并在此基础上提出了一个集成的点击欺诈检测框架。本节描述特征构建细节和集成框架构建步骤。

2.1 特征构建

在常规的点击欺诈数据集中,存在用户设备或网络等属性、发布者的网站或应用属性以及点击时间属性等,这些属性大多数是类别属性,而连续属性较少。

为了尽可能让检测模型提取到比较丰富的信息特征,提高检测精度,本文对点击欺诈数据集中类别属性和点击时间属性进行扩展,构建出多类通用且合理的特征。本文把构建的特征分为5类,以下详细描述这5类特征的构建思路。

1) 第1类特征:将数据集中除点击时间属性外的属性直接作为特征。

2) 第2类特征:考虑到有部分类别属性的不同取值数量很多,这可能会导致取到某些值的样

本数较少,这些样本在训练中可能会导致过拟合,因此预设一个阈值 c_1 ,将样本数量不大于 c_1 的属性值映射为同一个值,其余属性值保持不变。

3) 第3类特征:与第2类特征的构建方法类似,考虑到某些类别属性取到某些值的样本数特别少,这些样本很有可能是“欺诈”样本,因此预设一个阈值 c_2 ,将样本数不大于 c_2 的属性值映射为1,样本数大于 c_2 的属性值映射为2,同时将缺失值取为0。

4) 第4类特征:考虑到点击数据是否“欺诈”可能与时间点无关,而与时间段有关,如“欺诈”的点击可能会较频繁地出现在某个时间段,因此将点击时间属性离散化,将点击时间分为多个等长的时间段,设每段时间为 $c_3 \text{ min}$,不同时间段内的时间映射为不同值。

5) 第5类特征:同样利用点击时间属性值,考虑到“欺诈”的点击可能与每个时间段内的样本数量相关,因此将每个时间段内的时间映射为时间段内的样本数量。

2.2 提出的框架

为了更好地将特征与模型结合,本文提出了一种新的检测广告点击欺诈的集成框架,称为CAT-RFE集成学习框架。如图1所示,CAT-RFE集成学习框架主要由基分类器CatBoost、递归特征消除和Voting集成学习等组成。接下来介绍这3个组成部分。

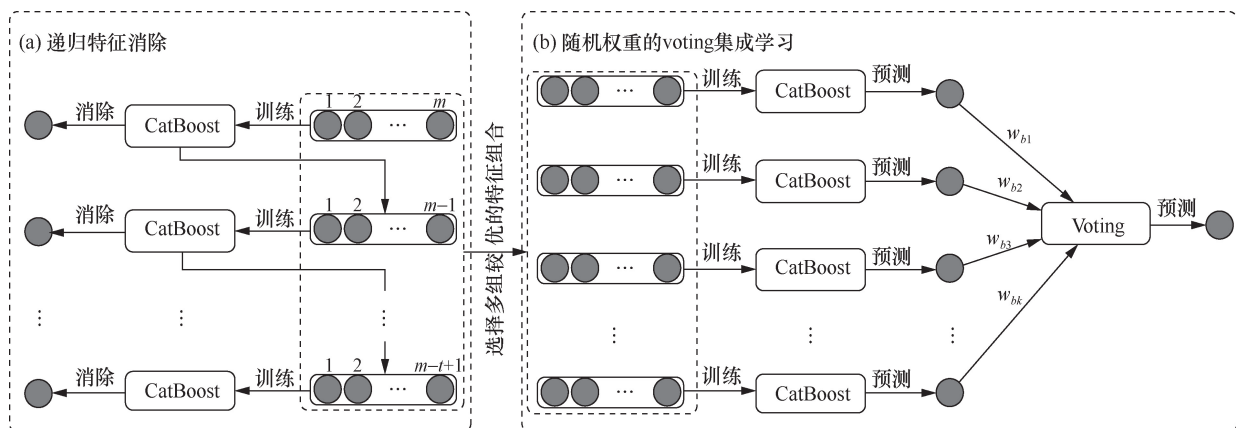


图1 CAT-RFE集成学习框架
Figure 1 CAT-RFE integrated learning framework

基分类器 CatBoost: CatBoost 是由 Dorogush 等在文献[3]中提出的一种梯度提升模型,针对现

有对类别特征处理的不足,引入了一个更有效的策略,使该模型能够很好地处理类别特征,同时

在训练中利用类别特征的优势,减少过拟合,使该模型在类别特征上的精度优于现有的梯度提升模型。

递归特征消除: RFE 是一种基于贪心策略的特征选择方法,其目的是希望找到一种最佳的特征组合用于模型训练,虽然基于贪心策略的 RFE 方法可能找不到最优特征组合,但随着每一次将贡献度最低的特征消除,RFE 方法能够找到贡献度尽可能高的特征组合,因此,RFE 是一种有效的特征选择方法。RFE 方法的步骤简单,每次用相同的机器学习模型进行训练,训练后将贡献度最低的特征剔除,再用剩余的特征继续训练,最终将交叉验证结果最好的特征组合作为目标特征组合。

Voting 集成学习: Voting 是最常见且最简单的集成学习之一,其思想是少数服从多数,将多个基分类器独立训练,对每个基分类器的预测结果进行投票,将票数多的结果作为集成学习的预测结果。Voting 集成学习存在两种投票方式,一种是 hard voting,另一种是 soft voting,前者是统计每个基分类器的预测结果,把每个基分类器的结果加权平均后作为集成学习的预测结果,而后者是统计每个基分类器的预测概率,将这些概率的加权平均值作为集成的预测概率。

本文框架采用 CatBoost 作为基分类器是考虑到 CatBoost 对类别特征的处理具有很好的效果;为了选择出贡献度较高的特征组合,CAT-RFE 进一步采用了 RFE 方法,将每一轮贡献度最低的特征剔除;为了弥补 RFE 方法可能无法找到最优特征组合的不足,CAT-RFE 使用基分类器 CatBoost,对从 RFE 方法得到的多个特征组合进行训练,并使用 voting 进行集成,从而得到集成预测结果。本文框架的算法步骤如下。

根据图 1 中递归特征消除部分,假设导入本框架的数据集的特征共有 m 个,RFE 方法执行 $t(1 \leq t < m)$ 轮。第一轮将具有 m 个特征的数据集代入 CatBoost 模型,在训练集上进行十折交叉验证得到验证分数 s_1 和每个特征的贡献度,同时将贡献度最低的特征剔除;第二轮重复第一轮的操作,将剩下 $m-1$ 个特征的数据集代入 CatBoost 基分类器,得到验证分数 s_2 ;重复执行直至第 t 轮,得到

验证分数 s_t 。RFE 执行结束后得到 t 组特征组合和对应的 t 个验证分数,在这 t 组特征组合中挑选出验证分数最高的前 k 个特征组合,由图 1 中随机权重 voting 集成学习部分,将包含这 k 个特征组合的数据集分别用 k 个 CatBoost 基分类器进行训练,然后使用 voting 对这 k 个基分类器的预测概率进行集成。为 k 个基分类器设置 c 组随机权重 $w_i = (w_{i1}, w_{i2}, \dots, w_{ik})$,其中 $i=1, 2, \dots, c$,每个基分类器对应预测概率为 $y_j (j=1, 2, \dots, k)$,用 voting 集成学习为每组随机权重按照如下公式计算预测概率。

$$z_i = \sum_{j=1}^k w_{ij} y_j \quad (1)$$

其中 $\sum_{j=1}^k w_{*j} = 1$,集成后得到 c 组预测概率,最后

在训练集上将 voting 集成的“交叉验证分数最高”的一组权重 $w_b = (w_{b1}, w_{b2}, \dots, w_{bk})$ 作为本框架的集成权重。

3 实验评估

3.1 数据集

本文所使用的数据集来源于飞桨“MarTech Challenge 点击反欺诈预测”任务,该数据集由模拟生成,对某些属性含义进行隐藏和脱敏处理。该数据集包含两个文件,训练集和测试集,训练集中包含 50 万次点击数据,每个点击数据包含 19 个属性和 1 个标签,测试集中包含 15 万次点击数据,每个点击数据仅包含 19 个属性(无标签)。在数据集包含的 19 个属性中,其中 18 个属性可作为点击数据的特征(样本 id 除外),数据集属性说明如表 1 所示。训练集的标签有两种取值(“0”和“1”),其中“0”表示正常的点击数据,约占训练集数据总数的 52%，“1”表示欺诈的点击数据,约占训练集总数的 48%。

该数据集的标签类别较为平衡,无须对数据进行采样处理。但是数据集中包含大量缺失值,同时存在较多类别属性的不同取值,这可能是该数据集检测精度不高的原因。到 2021 年 8 月,没有研究结果能在该数据集上达到 89.5% 以上的准确性。

3.2 数据预处理

在构建检测模型之前,首先需要进行数据预

处理,数据预处理能够帮助模型更完整地提取数据的特征,是提高模型精度的关键步骤之一。在数据预处理时,本文首先需要分析各个属性的数据类型,区分每个属性是连续属性还是类别属性,这有利于构建新特征以及后续选择和设计检测模型。

由表1中属性说明可知,除了“dev_height”“dev_width”“dev_ppi”和“timestamp”这4个属性是连续属性之外,其余属性均为类别属性。对于“dev_height”“dev_width”和“dev_ppi”这3个属性,由于其不同取值数量远小于数据总量,并且这些属性值的大小关系和点击是否为“欺诈”的相关性不大,因此可将这3个属性视为类别属性。而对于“timestamp”属性,虽然时间是连续的概念,但考虑到点击欺诈可能集中在一个时间段,而与时间点的关系不大,因此在构建特征时,可将该属性离散化。

表1 数据集属性说明
Table 1 Dataset attribute description

属性名称	说明	属性名称	说明
android_id	对外广告位 ID	dev_height	设备高
media_id	对外媒体 ID	dev_width	设备宽
apptype	app 所属分类	dev_ppi	设备屏幕分辨率
package	包名	lan	设备采用的语言
version	app 版本号	location	用户地理位置编码
ntt	网络类型	fea_hash	用户特征编码
carrier	设备使用的运营商	fea1_hash	用户特征编码
os	操作系统	cus_type	用户特征编码
osv	操作系统版本	timestamp	请求到达服务时间

有些类别属性的取值不能直接作为检测模型的输入,需要将属性值转化为模型能够识别的数值,因此可对类别属性采用数值编码,同时将属性缺失值视为该属性的另外一个类别,将属性类别编码为连续的整数。

按照上文提出的特征构建,将该数据集的属性扩展成5类特征。如表1所示,除去“timestamp”属性,构建的第一类特征有17个。如表2所示,构建的第2类至第5类特征共有24个,命名规则分别为“[属性名]_count_filter_[c_1]”“[属性名]_count_bi_[c_2]”“time_minute_[c_3]”“time_minute_count_[c_3]”。因此,特征总数为41个。

表2 新特征构建
Table 2 New feature construction

属性名称	新特征名称
android_id	(1) android_id_count_filter_20 (2) android_id_count_bi_1
media_id	(3) media_id_count_bi_1000
package	(4) package_count_filter_20 (5) package_count_bi_150
version	(6) version_count_bi_13000
osv	(7) osv_count_bi_2000
dev_height	(8) dev_height_count_bi_360
dev_width	(9) dev_width_count_bi_850
fea_hash	(10) fea_hash_count_filter_20 (11) fea_hash_count_bi_1
fea1_hash	(12) fea1_hash_count_filter_20 (13) fea1_hash_count_bi_60
timestamp	(14)~(19) time_minute_X(X=5,10,30,60,120,360) (20)~(24) time_minute_count_X(X=1,2,5,10,30)

3.3 评价指标

互联网广告点击反欺诈检测性能评估的常用指标是准确率(accuracy),该指标用于评价检测模型识别标签准确性的能力。

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (2)$$

其中,TP是实际标签为正例,预测为正例的数量;FP是实际标签为反例,而预测为正例的数量;TN是实际标签为反例,预测为反例的数量;FN是实际标签为正例,而预测为反例的数量。

3.4 基线模型

“MarTech Challenge 点击反欺诈预测”数据集给出了一个基于深度神经网络的点击欺诈模型,其在该大型数据集上表现出良好性能,本文将该模型作为基线模型,对比新提出的框架的有效性。

基线模型包含输入层、嵌入层、隐藏层和输出层。其中,采用嵌入层是由于数据集的属性基本上是类别属性,而每两个不同的类别属性值之间的距离无法用单个数值衡量,因此将类别属性值嵌入高维空间中作为单位向量,使每两个不同的类别属性值之间的欧几里得距离相等。

如图2所示,基线模型首先对每个输入数据分别做嵌入操作,每个嵌入操作分别提取出100维特征数据;然后由隐藏层分别提取出16维特征数据;最后将多个16维特征数据拼成一系列数据,再由输出层输出分类结果。

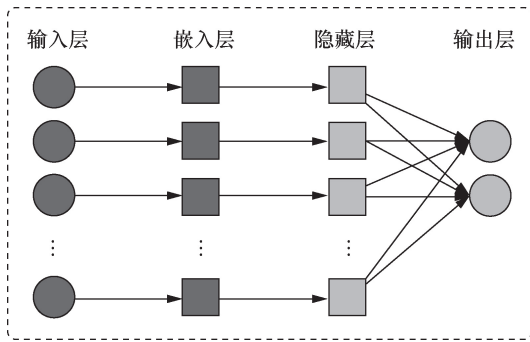


图 2 深度神经网络模型
Figure 2 Deep neural network model

3.5 提出框架的应用

在应用提出的框架之前，先比较 CatBoost 与其他机器学习模型的检测性能，以此来证明将 CatBoost 作为基分类器是最优的选择。为充分比较机器学习分类器的性能，根据分类模型的类型，选用其他 6 种分类器： K 近邻、逻辑回归、决策树、随机森林、LightGBM^[19]和 xgboost^[20]。其中， K 近邻属于统计模型，逻辑回归属于线性模型，决策树属于树型模型，随机森林属于树集成模型，LightGBM、xgboost 和 CatBoost 同属于树集成模型中的梯度提升模型。将总共 7 种分类模型对训练集做十折交叉验证，其中分类器的超参数除了将 CatBoost 的“iterations”设置为 200 以外，其余全设为默认值。实验结果如表 3 所示，可见 CatBoost 对该数据集的检测性能最优，因此选用 CatBoost 作为基分类器是比较合适的。

表 3 不同分类器的实验结果
Table 3 Experimental results of different classifiers

分类器名称	十折交叉验证准确率
K 近邻	62.244 2%
逻辑回归	69.255 0%
决策树	82.652 4%
随机森林	88.480 4%
LightGBM	88.274 0%
xgboost	88.811 6%
CatBoost	89.322 0%

在上述工作基础上，本文将数据集应用到第 2 节提出的框架中。首先，将基分类器 CatBoost 的超参数“iterations”设为 200，“learning_rate”设为 0.1636，“max_depth”设为 10；在 RFE 步骤中，设递归轮数 $t = 22$ ，即在 RFE 完成后会获得

22 组特征组合，表 4 列出了 22 个特征组合中训练集做十折交叉验证准确率最高的前 5 个；在 voting 集成步骤中，分别设特征组合数量 $k = 2, 3, 4, 5$ ，同时设随机权重组数 $c = 200$ 。为了方便读者复现所提框架，笔者已将本框架代码上传至 GitHub。

表 4 不同特征组合的实验结果
Table 4 Experimental results of different feature combinations

序号	特征个数	十折交叉验证准确率
1	26	89.396 4%
2	37	89.391 6%
3	35	89.386 0%
4	30	89.385 2%
5	39	89.382 2%

3.6 结果分析

在本文提出的 CAT-RFE 集成框架中，选取 voting 集成训练集“交叉验证准确率最高”的权重作为集成框架的权重。这里对“交叉验证准确率最高”有两种评判标准，一种是准确率“均值最高”，另一种是准确率的“均值减标准差最高”。表 5 中实验 a 和实验 b 表示这两种评判标准的 CAT-RFE 集成测试结果，其中，实验 a 表示选取十折交叉验证“均值最高”的权重，实验 b 表示选取十折交叉验证“均值减标准差最高”的权重，每一行表示本框架在 voting 集成步骤中分别选取前 2、3、4、5 个最好的特征组合进行集成的结果。设置实验 b 是考虑到集成结果的好坏可能容易受精度最差的一折影响，然而实验结果表明虽然实验 b 在本框架中取到了最高测试集准确率 89.381 3%，但该结果存在偶然性，在 k 值递增时的表现不如实验 a 稳定，因此在选取权重时还是按照“交叉验证均值最高”的方法。

表 5 不同集成权重的 CAT-RFE 结果
Table 5 Results of CAT-RFE with different ensemble weights

k	实验 a 测试集准确率	实验 b 测试集准确率
2	89.362 0%	89.381 3%
3	89.368 0%	89.347 3%
4	89.368 7%	89.356 0%
5	89.374 0%	89.374 0%

表 6 分别比较了基线模型、CatBoost、使用递归特征消除方法以及本文框架的实验结果。

其中,第1行是仅使用第一类特征在基线模型下的实验结果;第2行是仅使用第一类特征在CatBoost模型下的实验结果;第3行是构建新特征后,使用所有特征在CatBoost模型下的实验结果;第4行是运用RFE方法在所有特征中选择交叉验证均值最高的特征组合,然后将该特征组合应用于CatBoost模型的实验结果;第5行是使用所有特征在CAT-RFE集成学习框架下的实验结果。实验结果表明,新构建的特征在CatBoost模型下精度有所提高,证明新构建的特征是有效的。同时,单纯使用RFE方法选取的特征组合未必是最优的特征组合,本文框架的表现优于基线模型以及不使用和仅使用RFE的CatBoost模型,与基线深度神经网络模型相比在测试集上提升了1.35%的准确率。本文框架使用基分类器将多个特征组合进行训练再集成,基分类器选择了多个相同的梯度提升模型,避免了多种及模型因性能差异大而导致的不稳定问题,在互联网广告点击欺诈检测方面具备更优的辨识能力。

表6 模型结果对比
Table 6 Comparison of model results

序号	模型	特征	交叉验证准确率	测试集准确率
1	基线模型	第一类特征	88.040 0%	88.025 3%
2	CatBoost	第一类特征	89.322 0%	89.331 3%
3	CatBoost	所有特征	89.364 6%	89.358 7%
4	CatBoost+RFE	所有特征	89.396 4%	89.328 7%
5	提出的框架	所有特征	89.425 2%	89.374 0%

4 结束语

针对互联网广告点击欺诈检测,现有的集成模型是将不同类型的多个基分类器进行融合,但由于各个模型的检测能力差异较大,往往相互制约,很难组合成一个检测能力更好的模型。本文提出了CAT-RFE集成学习框架和一些关于点击欺诈的特征构建方法,框架不仅将特征与模型紧密结合,而且集成了多个相同的梯度提升模型,使框架更加鲁棒。所提出的框架在实际大规模点击欺诈检测中表现出良好的效果,在“MarTech Challenge 点击反欺诈预测”大型数据集上取得了比基线模型、常见统计检测模型和仅使用RFE方

法选择特征的模型更高的准确率。在未来的工作中,将尝试改进RFE方法或考虑其他特征选取方法,以降低特征选择的时间复杂度,并进一步研究对基线模型进行直接优化,构造更适合于点击欺诈的深度学习模型。

参考文献:

- [1] BORGHI M, DESSAI P, MALIK V, et al. Advertisement click fraud detection system: a survey[J]. International Journal of Engineering Research & Technology (IJERT), 2021, 10(5): 553-560.
- [2] GOHIL N, MENIYA A D. A survey on online advertising and click fraud detection[J]. Nayanaba Gohil Department of Information Technology Shantilal Shah Engineering, 2020.
- [3] DOROGUSH A V, ERSHOV V, GULIN A. CatBoost: gradient boosting with categorical features support[J]. arXiv preprint arXiv:1810.11363, 2018.
- [4] NAGARAJA S, SHAH R. Clicktok: click fraud detection using traffic analysis[C]//Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks. 2019: 105-116.
- [5] GABRYEL M. Data analysis algorithm for click fraud recognition[C]//International Conference on Information and Software Technologies. 2018: 437-446.
- [6] MOUAWI R, AWAD M, CHEHAB A, et al. Towards a machine learning approach for detecting click fraud in mobile advertising[C]//2018 International Conference on Innovations in Information Technology (IIT). 2018: 88-92.
- [7] 董亚楠, 刘学军, 李斌. 一种基于用户行为特征选择的点击欺诈检测方法[J]. 计算机科学, 2016, 43(10): 145-149.
DONG Y, LIU X, LI B. Click fraud detection method based on user behavior feature selection[J]. Computer Science, 2016, 43(10): 145-149.
- [8] TANEJA M, GARG K, PURWAR A, et al. Prediction of click frauds in mobile advertising[C]//2015 Eighth International Conference on Contemporary Computing (IC3). 2015: 162-166.
- [9] 张欣, 刘学军, 李斌, 等. 一种网络广告点击欺诈检测的SVM集成方法[J]. 小型微型计算机系统, 2018, 39(5): 951-956.
ZHANG X, LIU X J, LI B, et al. Application of SVM ensemble method to click fraud detection[J]. Journal of Chinese Computer Systems, 2018, 39(5): 951-956.
- [10] BERRAR D. Random forests for the detection of click fraud in online mobile advertising[C]//Proceedings of the 1st International Workshop on Fraud Detection in Mobile Advertising. 2012: 1-10.
- [11] SHAOHUI D, QIU G W, MAI H, et al. Customer transaction fraud detection using random forest[C]//2021 IEEE International Conference on Consumer Electronics and Computer Engineering (IC-CECE). 2021: 144-147.
- [12] PERERA K S, NEUPANE B, FAISAL M A, et al. A novel ensemble learning-based approach for click fraud detection in mobile advertising[M]//Mining Intelligence and Knowledge Exploration. 2013: 370-382.

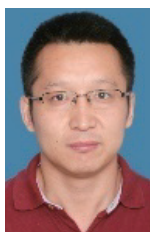
- [13] GOHIL N P, MENIYA A D. Click Ad fraud detection using xgboost gradient boosting algorithm[C]//International Conference on Computing Science, Communication and Security. 2021: 67-81.
- [14] VIRUTHIKA B, DAS S S, KUMAR E M, et al. Detection of advertisement click fraud using machine learning[J]. International Journal of Advanced Science and Technology, 2020, 29(5): 3238-3245.
- [15] MINASTIREANU E A, MESNITA G. Light GBM machine learning algorithm to online click fraud detection[J]. J Inform Assur Cybersecur, 2019, (2019): 263928.
- [16] ZHANG Y, TONG J, WANG Z, et al. Customer transaction fraud detection using Xgboost model[C]//2020 International Conference on Computer Engineering and Application (ICCEA). 2020: 554-558.
- [17] THEJAS G S, DHEESHJITH S, IYENGAR S S, et al. A hybrid and effective learning approach for Click Fraud detection[J]. Machine Learning with Applications, 2021, 3: 100016.
- [18] HADDADI H. Fighting online click-fraud using bluff ads[J]. ACM SIGCOMM Computer Communication Review, 2010, 40(2): 21-25.
- [19] KE G, MENG Q, FINLEY T, et al. Lightgbm: a highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30: 3146-3154.
- [20] CHEN T, HE T, BENESTY M, et al. Xgboost: extreme gradient boosting[J]. R Package Version 0.4-2, 2015, 1(4): 1-4.



耿光刚（1980—），男，山东泰安人，博士，暨南大学教授，主要研究方向为机器学习、大数据分析和互联网基础资源安全。



延志伟（1985—），男，山西兴县人，博士，中国互联网络信息中心研究员，主要研究方向为 IPv6 移动性管理、BGP 安全机制、信息中心网络架构。



朱效民（1982—），男，山东莱芜人，博士，山东齐鲁大数据研究院副研究员，主要研究方向为高性能计算、大数据分析。

[作者简介]



卢翼翔（1995—），男，广东潮州人，暨南大学硕士生，主要研究方向为统计机器学习、网络空间安全。



张新常（1975—），男，山东新泰人，博士，山东省科学院教授，主要研究方向为智能网络、网络架构与协议，工业互联网。