计算机科学
COMPUTER SCIENCE

基于攻击经济学的移动虚拟运营商诈骗检测

李洋, 李振华, 辛显龙

引用本文

李洋, 李振华, 辛显龙.基于攻击经济学的移动虚拟运营商诈骗检测[J]. 计算机科学, 2023, 50(8): 260-270.

LI Yang, LI Zhenhua, XIN Xianlong. Attack Economics Based Fraud Detection for MVNO[J]. Computer Science, 2023, 50(8): 260-270.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000103

基于攻击经济学的移动虚拟运营商诈骗检测

李　洋[1]　李振华[1]　辛显龙[2]
1 清华大学软件学院　北京 100084
2 小米科技有限公司　北京 100085
(liyang874235642@163.com)

摘　要　受电信资源充分利用和激发良性市场竞争的双重驱动,移动虚拟运营商(虚商)近年来迅速流行,其依靠基础运营商的基础设施为用户提供更灵活优惠的服务。考虑到线下实体店维护成本较高,虚商基本上采取完全线上的服务方式,这给用户监管带来很大困难;很多不法分子利用在线身份认证漏洞,大量购买虚商电话卡拨打诈骗电话,严重损害了虚商及其用户声誉,成为目前虚商存续发展的瓶颈。为解决该难题,与拥有超两百万用户的主流虚商"小米移动"合作研究,发现相关工作普遍假设诈骗电话是随意的、零散的或隐蔽的,导致其检测方法对于虚商场景低效甚至无效。然而,通过人工分析发现,不同于传统假设,虚商场景中几乎所有的诈骗电话都是有组织、按计划、成规模的,从而提出基于攻击经济学、合理分析诈骗电话时空特征的新型检测方法,成功提取出有效甄别的关键特征,再结合机器学习分类,将诈骗用户的比例降低至0.023‰,远低于基础运营商在信息充分的前提下所达到的0.1‰。在避免所提方案被破解的前提下,已将部分代码和数据开源,以帮助净化整个产业生态。

关键词:移动虚拟运营商;诈骗检测;攻击经济学;时空特征分析;机器学习
中图法分类号　TP393

Attack Economics Based Fraud Detection for MVNO

LI Yang[1], LI Zhenhua[1] and XIN Xianlong[2]
1 School of Software, Tsinghua University, Beijing 100084, China
2 Xiaomi Technology Co. LTD., Beijing 100085, China

Abstract　Driven by the full utilization of telecommunication resources and stimulating healthy market competition, mobile virtual network opera-tors(MVNOs) become popular rapidly in recent years. MVNOs rely on the infrastructures of mobile network operators(MNOs) to provide users with cheaper and more flexible services. Due to the high maintenance costs of physical stores, MVNOs mostly provide fully online service. However, scammers use vulnerabilities in online authentication to purchase SIM cards and make scam calls, which seriously affects the reputation of MVNOs and their users. This has become a bottleneck problem for the survival and development of MVNOs. To address this issue, we collaborate with a large commercial MVNO with over 2 million users named Xiaomi Mobile. Related work generally assumes that scam calls are random, scattered or hidden, making the detection methods inefficient or even invalid for the scenario of MVNOs. However, by analyzing the crowdsourced dataset, almost all scam calls are found to be organized, planned, and scaled. Thus, a method based on attack economics and reasonable analysis of the spatio-temporal characteristics of scam calls is proposed. This method successfully extracts the key features, and by combining with machine learning-based classification, it greatly reduces the proportion of scammers in Xiaomi Mobile to 0.023‰, which is far lower than the 0.1‰ achieved by the MNOs that have sufficient information. Under the premise of excluding the risk of being cracked, part of the code and data has been open sourced to help purify the ecology of entire telecom industry.

Keywords　MVNO, Fraud detection, Attack economics, Spatio-Temporal analysis, Machine learning

1　引言

受益于通信技术进步和经济全球化的积极影响,移动电信产业近30年来高速发展,5G将服务性能提升至新高度[1]。高速发展的另一面,则是基础运营商的寡头垄断,导致电信产业市场竞争不活跃、服务质量不尽如人意。为增加

到稿日期:2022-10-13　返修日期:2023-03-03
基金项目:国家重点研发计划(2022YFB4500703);国家自然科学基金(61902211,62202266)
This work was supported by the National Key R & D Program of China(2022YFB4500703) and National Natural Science Foundation of China (61902211,62202266).
通信作者:李振华(lizhenhua1983@gmail.com)

有效市场竞争，提升电信资源使用效率，提高行业服务质量，移动虚拟运营商（虚商）近年来应运而生，其依赖基础运营商的蜂窝基础设施，提供更便宜灵活的资费套餐，已获得一定的市场份额与商业成功。

考虑到线下实体店维护成本较高，虚商基本采取完全线上的认证购卡方式，方便快捷，却给了不法分子可乘之机。虚商在线认证分为个人信息认证和活体认证两个阶段，对于第一阶段，不法分子可以用较小的代价获取农村老人、流浪者等的身份信息来通过个人信息认证；对于第二阶段，如果采用金融支付软件广泛应用的先进炫彩活体技术[2]，其成本与维护线下实体店相当，因此虚商普遍使用成本较低的动作活体认证[3]，但此认证方式对基于视频合成换脸的活体攻击技术[4]防御能力较差，不法分子能够攻击成功，从而大量购买虚商电话卡拨打诈骗电话。这导致虚商用户中诈骗号码的比例（0.37‰）远大于基础运营商（0.1‰左右），对虚商及其用户的声誉产生严重不良影响。因此，虚商诈骗用户的检测和治理对其存续发展乃至整个电信行业的生态净化至关重要[5-6]。

为研究上述问题，本文作者和拥有超 200 万用户的商业虚商“小米移动”合作，对虚商诈骗用户的检测进行了广泛调研和深入探索。起初小米移动试图直接使用基础运营商已经十分成熟的监管服务为其提供诈骗号码列表，却因为缺少两类关键数据而无能为力：实体店线下现场认证数据和用户通话建立的信令数据。实际上，作为虚商能获得的相关数据基本上只位于通信会话层。因此，我们重点调研基于通信话单数据的诈骗检测方法，发现相关工作普遍假设诈骗电话是随意的、零散的或隐蔽的，进而从受害者角度选择比较直观的特征进行分析[7-8]，应用到虚商场景后精度（Precision）不高，召回率（Recall）极低，效果很不理想。

为了找到行之有效的检测方法，我们人工分析小米移动提供的话单数据，并借助与基础运营商合作获取的大量众包标注（是否为诈骗电话）构建出关键数据集，基于此展开四阶段闭环系统性研究。如图 1 所示，首先，预处理众包标注，采用置信学习方法[9]有效清理其中的偶然错误。其次，仔细分析真实诈骗用户的表层行为及深层动机，发现两者存在强烈关联：诈骗行为几乎总是受（潜在）暴利驱动才会发生，其通信会话行为明显存在组织化的互通性、时间维度的高频性、空间维度的离散性，是有组织、按计划、成规模的；诈骗攻击特征与诈骗用户的经济收益高度相关。因此，本文提出基于攻击经济学、合理分析诈骗电话时空特征的风险收益建模方法来提取关键特征。随后，为弥补诈骗用户在总用户中占比小的不足，使用基于类别不平衡采样的机器学习分类算法训练模型。最后，考虑到预测错误会严重影响用户体验（反之，少量预测疏漏并不影响用户体验），对预测结果采用基于特殊化统计模型的两阶段验证策略。整体方案大规模使用后取得了很好的实际效果：6 个月内虚商诈骗用户比例降低至 0.023‰，远低于基础运营商在充分信息前提下所达到的 0.1‰；相应地，12321 平台投诉量下降 90%，公安部涉案号码数量下降 75%。

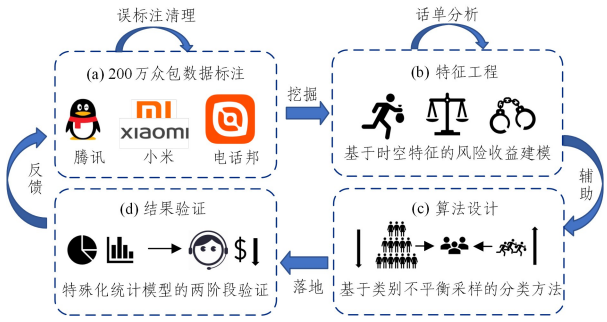


图 1　虚商诈骗检测方法整体架构

Fig. 1　Overall architecture of fraud detection for MVNO

本文主要的科学发现和创新总结如下：

（1）数据标注方面。仔细观察多来源众包标注数据可以发现，部分标注明显不正确（比如将未发生任何主动呼叫的号码标记为诈骗），并且无标注数据中也存在诈骗用户。为此，采用针对性的置信学习（Confident Learning，CL）方法，根据分类器的预测概率分布、结合数据驱动的置信度选择来过滤不可靠标注。同时，将已验证准确无误的标注数据子集和带噪声的标注数据混合处理，观察到处理后数据集的标注准确性接近 100%。

（2）特征工程方面。尽管虚商只能获取用户通信会话层数据，但通过关联分析这些数据的外在表象和诈骗用户的内在动机发现，诈骗电话的发生并不随机或随意，而是由组织利益驱动的，符合经济学规律。基于此观察，细致量化诈骗电话的时空特征并建立风险收益模型，提取出能够有效甄别诈骗用户的关键特征，并定性和定量证明，利用这些特征几乎可以彻底排除所有理性的逐利诈骗者。

（3）算法设计方面。由于虚商用户中诈骗用户占比很小，机器学习训练出的分类模型受多数类别的正常用户样本影响较大，导致边界样本预测不准确。为平衡样本类别，采用过采样与欠采样相结合的少数类别合成-托梅克联系对联合采样法（Synthetic Minority Oversampling Technique-Tomek link undersampling，SMOTETomek）[10]，对多数类别样本（普通用户）使用 Tomek link 方法[11]滤除分类边界的噪声和冗余样本，对少数类别样本（诈骗用户）采用 Borderline-SMOTE 算法[12]生成符合诈骗用户时空特征的近似样本。使用重采样后的样本结合平衡权重的随机森林分类算法，达到了 92% 的召回率和 90% 的精度，召回率显著优于相关工作，精度（因尽量提升召回率）并不理想，但可以通过谨慎的结果验证有效消除负面效果。

（4）结果验证方面。考虑到虚商需要对疑似诈骗用户执行停机销户和拉黑处理，对算法预测的疑似诈骗用户需谨慎对待，防止因误判伤害正常用户。本文权衡可行性和操作成本，设计出基于特殊化统计模型的两阶段客服验证策略，在有效验证的同时最小化验证成本。

在避免上述方案被攻击者破解的前提下，研究相关代码和样本数据已部分开源[1)]，其中所有数据已做脱敏处理，确保不泄露用户隐私或小米移动的商业机密。

1) https://mvno-optimization.github.io

本文第2节综述电信诈骗检测领域的代表性研究工作，特别是与虚商场景相近和基于通信会话特征的检测工作；第3节介绍标注数据的来源、标签清洗和观察分析；第4节讲解基于攻击经济学的特征工程、类别不平衡机器学习算法设计和模型效果；第5节说明针对预测结果的两阶段客服验证策略的设计及其统计学基础；第6节讨论虚商诈骗检测和治理的未来发展方向。

2 相关工作

相比基础运营商，虚商由于缺乏关键数据，对诈骗用户的检测与监管需要更智能和适配的检测技术与策略。考虑到虚商可以利用的数据资源，本节回顾了近年来国内外的数据驱动的诈骗检测相关研究工作，涉及不同场景、维度、规模的异构数据集。随着通信技术的飞速发展，反诈骗技术和诈骗技巧都是与时俱进的，且在疫情常态化时代，诈骗用户的行为模式也有所变更，因此需尽可能调研新的诈骗检测方法。

2.1 中小规模研究

在电信行业存在巨头垄断的市场背景下，运营商对其关键商业数据（如话单数据）严格保密，学术社区可以获取到的用于研究的数据往往比较有限，因此基于中小规模采样数据的诈骗检测研究较多，适合面向企业未来应用的前瞻性探索和评估[13-23]。利用商业数据的研究，主流方法是基于统计机器学习算法的模型检测。2017年，Subudhi等[13]提出了一种采用模糊C均值聚类来发现移动电信网络中诈骗事件的创新方法，基于106个用户的9个月的通信数据，其准确率达93.09%，召回率达95%。2018年，Li等[14]从近8000个用户的6个月的呼叫详细记录（Call Detail Record，CDR）数据集中提取特征并用支持向量机（Support Vector Machine，SVM）算法对用户分类；采用综合考虑诈骗者行为的有限状态机（Finite State Machine，FSM）筛除分类后的可疑用户，其准确率达93.56%，召回率为89.22%。同年，Chouiekh等[15]根据从运营商获取的仅300名用户的真实CDR数据集，设计深度卷积神经网络（Deep Convolutional Neural Network，DCNN）检测诈骗者，准确率达到82%。同年，Liu等[16]对银行电信诈骗问题设计了基于马尔可夫网络的检测方法，在模拟和真实的银行交易数据上达到了0.682的F1分数。近几年基于机器学习算法的小规模诈骗检测研究的准确率也在逐步提高[17-20]。2020年，Lu等[17]针对运营商的超3万个用户的数据，采用自适应合成采样算法（Adaptive Synthetic Sampling，ADASYN）预处理原本不平衡的数据集，然后利用随机森林算法基于新数据集训练模型来防止过拟合，其准确率为92.5%，召回率为87.6%。2022年，Xu等[20]基于电话文本数据设计了带有注意力机制的双向长短时记忆神经网络来识别诈骗电话，在小规模数据集上达到了94.7%的准确率。尽管上述工作实验结果较好，但由于实验数据样本数太小，其在真实工业部署下的可行性和有效性并未经过实践检验。

2.2 大规模研究

尽管中小规模研究可以一定程度评估方法的适配性，但无法取代大规模数据研究的代表性和说服力[24]。此类研究需要运营商的数据支持，同时需要电信从业者提供经验性指导。如果缺乏运营商的支持，则需要较高的人力和时间成本来自行采集相关数据，因此基于大规模数据的诈骗检测研究数量比较有限。2018年，Min等[25]对13万次通话的信令数据集通过聚类来挖掘隐藏用户特征，采用主成分分析来对特征降维，使用网格搜索确定参数，提出了基于K均值的诈骗检测系统，但对该系统的实际表现仅有人工启发式评估，没有和真实标注对比来验证效果。2019年，Liu等[26]提出了一种基于注意力的图表示学习模型，他们采集特定区域的超12万个用户一周的CDR数据集，基于此训练的模型的Micro F1分数达到了0.9，但由于其数据是自行采集的，相比运营商监测数据缺乏权威性且时间跨度较短，因此实际效果存在偏差。随着近几年疫情趋于常态化，出现了新的诈骗特征与相应的大规模检测研究[27-32]。2021年，Terzi等[27]设计了基于通过时变特征表达的用户的行为偏差的诈骗检测模型，在近41万用户的CDR数据集上进行了评估，但准确率仅为86%。2022年，Chadysas等[28]使用立陶宛的虚商Moremins公司的近300万次通话的CDR数据集，进行了基于特殊化指标的离群值分析，但该研究仅为相关特征确定了启发式判断阈值，阈值和方法的通用性未经过工业场景下的实践检验；同年，Jiang等[29]根据众包采集的11万用户两周的CDR数据集进行了时间序列建模，发现了数据中的关键模式并提出了结合注意力机制与霍克斯增强序列模型的检测方法，达到了97.1%的高精度，但检测的关键指标之一的召回率仅为85.7%。相比上述工作，我们的数据规模大（1 529 103个用户）、数据来源权威可靠（小米移动虚商用户CDR数据和12321平台及基础运营商提供的多来源众包标签）、采集时间长（2020年7月，跨度为一个月），并结合领域知识、攻击经济学分析、统计机器学习算法等进行了全面的科学研究，产生的整体方案已经在和小米移动合作的工业场景下大规模部署且具有真实效益。

3 标注数据集

本节首先介绍标注数据集的来源和维度，然后说明对噪声标签的清洗和效果，最后介绍对数据集多维度特征的观察、分析与发现。

3.1 数据集说明

通过和小米移动虚商合作，我们获取了超150万用户的CDR数据集，以及每条话单记录对应的标注。数据集包括1 529 103个样本，采集时间为2020年7月一整月。小米移动和基础运营商合作，获取到了多来源的诈骗用户标签（包括12321平台用户举报和基础运营商收集的来源于多个企业的众包标注数据）。多个不同来源的标签组成繁杂，为便于后续的观察分析与算法学习，我们将带有与诈骗弱相关标签（如骚扰电话）的样本排除，并把不同类别的诈骗标签（如网购诈骗、理财诈骗等）统一，尝试通过虚商用户的CDR数据来进行有效检测。标签的简化和统一，一方面有助于后续的模型训练，另一方面是因为虚商无法获取仅基础运营商才能采集的用户信令数据，且用户通话内容为隐私内容同样无法获取，导致无法对更具体、细分的诈骗标签进行甄别。

考虑到电信用户的CDR数据维度数量庞大、种类多样，我们与小米移动的数据工程师合作，结合先验信息对所有用户的呼叫详细记录数据的维度进行了启发式过滤，提取了需要深入分析的41个诈骗关联特征，如表1所列。根据诈骗用户的行为特点，我们对特征维度进行了简单的分类。首先，诈骗用户为了最大化期望收益，会充分利用时间高频率拨打诈骗电话，在时间上特征体现在通话频率、通话时长、通话间隔等方面；而为了增加成功机会，诈骗者会向不同类型、不同地区的大量用户发起呼叫尝试诈骗，在空间上特征体现在呼叫的不同号码和地区的数量等方面；此外，根据调研[33-34]，我们得知诈骗电话存在明显的团伙作案现象，可以对其呼叫利用的基站相关数据、呼叫号码的漫游地和归属地相关特征等进行分析。除上述特征外我们还对每条话单记录计算了作为主叫平均每个号码的呼叫次数，考虑到用户大多会周期性地或频繁地与自己的常用联系人（如家人、朋友、工作伙伴等）通话，而诈骗者对已尝试并失败的用户不会再次呼叫，其大量的对不同用户的呼叫尝试次数会使得该特征趋近于1，与一般用户体现出差异。另外考虑到很多诈骗者利用老年人的身份信息来购买电话卡拨打诈骗电话，因此年龄也是我们的关注角度之一。

表1　初步抽取的特征维度表

Table 1　Initial feature dimensions

特征类别	维度
时间相关类	主/被叫呼叫总次数，主/被叫日均呼叫次数（均值和25%/50%/75%分位数），主/被叫平均呼叫时长（均值和25%/50%/75%分位数），主/被叫呼叫间隔（均值和25%/50%/75%分位数），首话单时间和激活时间间隔天数
空间相关类	主/被叫呼叫号码数，主叫呼叫号码的归属省份/城市数量，主叫使用的基站数量，主叫利用最多的基站的呼叫次数占比，主叫呼叫数最多的号码所在地的呼叫占比，主叫漫游地和对方同城呼叫次数占比，主叫漫游地和对方同城号码数占比，主叫漫游地与号码归属地属同一省的呼叫占比，主叫漫游地与号码归属地属同一城市的呼叫占比
其他	年龄，主（被）叫总呼叫数除以主（被）叫呼叫号码数量

3.2　噪声标签的清洗

我们的数据集标签是与基础运营商合作获取的众包标签，来源繁杂。为了保证标签的正确性，我们对数据集的标签进行仔细观察和分析发现，存在很多错误标签，如网约车司机号码被标注为诈骗电话。一方面是因为标注是多来源的，源头的可靠性不同，如12321平台标签是官方提供的高可靠标注，而基础运营商提供的企业/用户标注可能存在错误标签；另一方面是因为此类用户的通话对象广泛，易被用户误标注。

为了防止潜在的错误标注影响后续特征挖掘和分类模型的训练，我们调研了标签带噪数据集的处理方法和机器学习算法，对比了先验启发式过滤、置信学习等方法的效果。在从全部数据中选出一部分验证为标签无误的数据子集后，据此来评估上述算法处理后数据集标签的正确性。

首先将对数据的多视图观察分析与小米移动的防诈骗经验结合，设计了基于数据的先验启发式过滤方法。具体来说，我们对诈骗号码区别于正常用户的一些关键特征（如平均每天呼叫次数/号码数/城市数等）根据数据统计和先验知识设定了条件阈值。若标注用户在这些特征上未达到诈骗号码的条件阈值却被标记诈骗，或者超过条件阈值却未被标为诈骗，我们会观察这部分用户的历史通话行为，对行为异常的用户同时以短信或公众号等方式推送问卷调查，尝试了解其近期通话行为异常的原因，据此来确认其标签的正确性。但这种方式在实际应用中局限性极大，首先，不同特征的条件阈值设定相当困难，对于有不同通话习惯的用户，合适的条件阈值往往存在不可忽略的差别，因此经验性的启发式条件阈值过滤方法的准确率低、适用面窄、扩展性差；其次，以短信或公众号的方式能够有效触达的用户极为有限，仅有极少量的用户会对问卷给出真实反馈。因此，基于数据的启发式过滤方法不仅设计难度大，且效率低下、实际效果差。

在深入调研带噪标注数据处理方法后我们发现，Northcutt等[9]提出的置信学习方法和我们的场景适配性较好。置信学习是一种基于噪声数据处理和筛除的方法，它基于概率阈值对噪声估计以及对样本进行置信度排名，通过识别和表征数据集中的错误标注来提升数据集整体的标签正确性。在虚商的大规模众包数据场景下，置信学习的自动化标注去噪方法的理论基础和执行效率都明显优于人工启发式方法。置信学习有3个主要步骤：首先估计噪声标签和真实标签的联合分布；然后根据置信度排名过滤低置信度标注样本；最后修改过滤后数据集样本的类别权重，重训练分类模型。具体来说，设带有噪声的标注数据集包括n_c个类别的N个样本，令众包标注的标签为$\tilde{y}$，正确标签为y^*，尽管我们无法判断哪些记录的标签为y^*，但可以通过交叉验证的方法对概率进行估计。首先用一个分类器对数据集进行训练和分类，用交叉验证计算样本i在类别j下的概率$P(i,j)$；然后计算数据集中每个标注类别j下的平均概率P_j并将其作为置信度阈值；最后用交叉验证得到样本i的正确标签y_i^*，如式(1)所示：

$$y_i^* = \underset{j}{\operatorname{argmax}}\ P(i,j), P(i,j) \geqslant P_j \tag{1}$$

据此可计算$n_c \times n_c$的计数矩阵$\boldsymbol{M}_{\tilde{y},y^*}$。矩阵元素为在类别$j$下的概率超过置信度阈值$P_j$的样本数。举例来说，$\boldsymbol{M}_{\tilde{y}=c_1,y^*=c_2}=k$表示众包标签为$c_1$而其正确标签为$c_2$的样本数为$k$。为使计数矩阵元素的和与数据集众包标注样本数相同，令$|Num_{\tilde{y}=i}|$为众包标签$\tilde{y}=i$的样本数，按式(2)变换矩阵。

$$\widetilde{\boldsymbol{M}}_{\tilde{y}=i,y^*=j} = \frac{\boldsymbol{M}_{\tilde{y}=i,y^*=j}}{\sum\limits_{j\in\{1,2,\cdots,n_c\}} \boldsymbol{M}_{\tilde{y}=i,y^*=j}} \times |Num_{\tilde{y}=i}| \tag{2}$$

最后根据式(2)估计众包标注标签$\tilde{y}$和y^*的联合分布

$Q_{\tilde{y},y^*}$，如式(3)所示：

$$Q_{\tilde{y}=i,y^*=j}=\frac{\tilde{\boldsymbol{M}}_{\tilde{y}=i,y^*=j}}{\sum_{i\in\{1,2,\cdots,n_c\},j\in\{1,2,\cdots,n_c\}}\boldsymbol{M}_{\tilde{y}=i,y^*=j}} \tag{3}$$

得到联合分布后，有 5 种可选方法过滤错误标注样本。1)选取符合$\tilde{y}_i\neq\arg\max_{j\in\{1,2,\cdots,n_c\}}P(i,j)$这个条件的样本进行过滤，也就是选取最大概率对应的标签与众包标签不一致的样本进行过滤；2)筛除在建立计数矩阵时在非对角位置的样本；3)对每个类别的样本按概率 $P(i,j)$从低到高排序，从前往后过滤掉 $N\times\sum_{j\in\{1,2,\cdots,n_c\},j\neq i}Q_{\tilde{y}=i,y^*=j}$ 个样本；4)将计数矩阵的非对角元素按间隔从大到小排序，选前 $N\times Q_{\tilde{y},y^*}$ 个样本过滤；5)将 3)和 4)组合应用。在诈骗标注场景下，标注的可靠性对数据质量和后续模型效果影响很大，而 1)和 3)方法会删除机器学习模型无法确定类别的样本(即概率等于置信阈值的样本)，4)方法包含了方法 3)，因此这 3 种方法在我们的场景下不可靠，故选方法 2)确定数据集。

3.3 数据集的多维度观察发现

在清理了数据集的错误标签和样本后，我们针对 3.1 节中抽取的特征维度进行了细致的观察、挖掘和分析，发现在直觉上与诈骗者行踪相关的基站特征与漫游地/归属地特征方面，诈骗用户与正常用户并没有体现出明显差异；而时间相关特征与呼叫的空间覆盖等特征则与直觉相符，诈骗者的此类特征与正常用户明显不同。本节对数据挖掘中诈骗者体现明显差异的维度进行了分析。

为了最大化经济效益，诈骗者一般会在其号码未被检出的有效期内尽可能多且快地拨打电话实施诈骗，这也在我们的数据中得到了一定程度的证实。如图 2 所示，实曲线和虚曲线分别为标注诈骗用户和无标注用户的平均每天主叫次数 CDF 曲线。显然，虚曲线在实曲线之上，且虚曲线在横坐标很小时有较大转折，说明绝大多数标注诈骗用户的呼叫频率远高于无标注用户。具体来说，无标注用户作为主叫平均每天呼叫次数的最大值为 99，均值为 2，中值仅为 1。在网络通信技术高速发展的时代，基于蜂窝网络的网络电话/语音/视频流逐步取代了传统通信方式，故日均呼叫次数少是当前用户的正常特征；部分正常用户因为职业原因也会大量呼叫，这部分用户需结合多维度特征综合分析。而标注诈骗用户最大值高达 289 次，均值 24 次，中值 11 次，显著高于无标注用户。

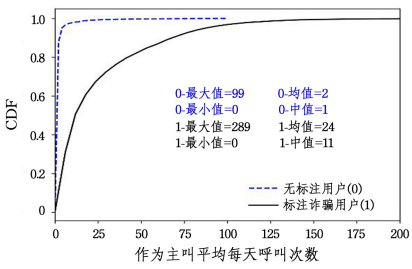


图 2 两类用户作为主叫平均每天呼叫次数 CDF

Fig. 2 CDF of the average numbers of outgoing calls

此外，小部分标注诈骗用户(约 20%)的通话频率与正常用户相仿，根据我们和警方咨询得到的信息，这是因为诈骗团伙大量购买虚商电话卡，通过周期性切换电话卡呼叫来规避模型或基于规则的检测，延长每张电话卡的存活时间以节约成本。因此对于团伙作案的诈骗用户，需要结合其表层行为和深层动机进行多维度特征甄别。

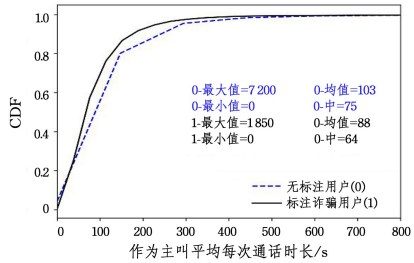


图 3 两类用户作为主叫平均每次通话时长 CDF

Fig. 3 CDF of the average durations of outgoing calls

时间维度上，除频率外还需考虑两个关键特征：平均呼叫时长以及间隔。图 3 和图 4 分别给出了两类用户的平均呼叫时长分布以及平均呼叫间隔分布。图 3 表明，标注诈骗用户曲线在无标注用户上方，即诈骗用户平均通话时长比正常用户短。具体而言，无标注用户平均呼叫时长最大值高达 2 h，而标注诈骗用户最大值为 0.5 h；标注诈骗用户均值也比无标注用户少 15 s。

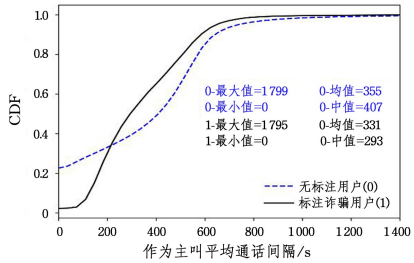


图 4 两类用户作为主叫平均通话间隔 CDF

Fig. 4 CDF of the Average intervals between consecutive outgoing calls

在平均呼叫间隔方面，图 4 中的曲线是根据先验阈值把用户的长通话间隔(如夜晚睡眠时间)滤除后统计绘制的，其中无标注用户有超过 20%呼叫间隔为 0，是因为这部分虚商用户使用的服务以流量为主，没有通话记录。排除掉这部分用户的影响，可以看到标注诈骗用户曲线大体上在无标注用户之上。尽管两类用户的平均呼叫间隔最大值很接近，但均值相差 25s，诈骗用户的中值更是远低于无标注用户(相差 114 s)，可见诈骗用户普遍呼叫间隔较短。

除时间维度特征外，诈骗用户的空间维度特征也与正常用户有明显差异。虚商能获取的诈骗用户空间维度特征比较有限，仅能从其呼叫所利用的基站、号码归属地、号码漫游地等角度分析。首先我们从用户呼叫所利用的基站角度观察分析。如图 5 所示，在“作为主叫利用最多的基站的呼叫次数占比”这一维度上(简称 BSRate)，无标注用户的 CDF 曲线明显位于标注诈骗用户的曲线上方，表明标注诈骗用户利用相同基站的呼叫比例更高。无标注用户的 BSRate 均值仅为 30%，中值为 19%；相比之下，标注诈骗用户的均值高达 59%，中值为 62%，明显偏高。另一方面，图 5 中 CDF 曲线的形状一定程度上反映了两类用户的通话特征，两条曲线间的

空隙表明，标注诈骗用户利用相同基站的呼叫比例明显高于无标注用户；无标注用户曲线斜率逐渐降低，是因为大多数正常用户的通话不会集中利用相同的基站，也就是呼叫地点相对分散；而标注诈骗用户斜率逐渐增加则说明诈骗用户通话位置比较固定，实曲线尾部的长转折也说明有超过 20% 的标注诈骗用户利用最多的基站通话占比在 95% 以上，进一步佐证了上述发现。

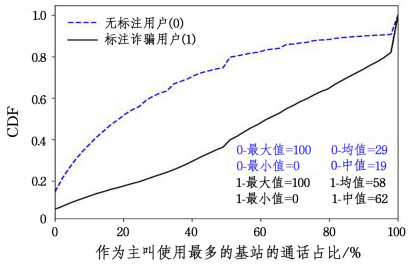


图 5　两类用户作为主叫使用最多的基站通话占比 CDF

Fig. 5　CDF of call proportions of the most involved base station for outgoing calls

从攻击经济学的角度来讲，为了使实际收益达到期望，诈骗用户会在有限的号码存活时间内尽可能地向不同目标发起呼叫，因此空间上的呼叫号码数也是我们的主要观察目标。如图 6 所示，无标注用户曲线明显位于标注诈骗用户曲线之上，且标注诈骗用户的几项统计值（最值、均值、中值）数倍于无标注用户；此外，仅约 1% 的无标注用户呼叫的不同号码数量大于 500，而标注诈骗用户则有约 42%。尽管有超过半数的标注诈骗用户呼叫的不同号码数量小于 500，且两条曲线从起点开始的斜率均较大，但从图像上可以看出在平均呼叫号码数方面标注诈骗用户明显多于无标注用户。对于部分标注诈骗用户呼叫号码数较少的现象，我们推测这和前述的诈骗团伙周期性切换电话卡进行呼叫的情况相关；另外从两条曲线斜率陡降的转折点上也可以看出，诈骗用户对呼叫目标广撒网的特征更加明显。数据的特征表现一定程度上证明了我们对诈骗用户行为的推测的正确性。

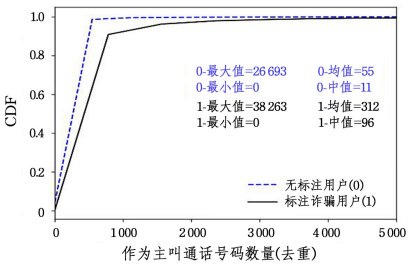


图 6　两类用户作为主叫呼叫的不同号码数 CDF

Fig. 6　CDF of the numbers of phone numbers in outgoing calls

考虑到公安部门提供的诈骗团伙作案的信息，我们还观察了标注诈骗用户呼叫的号码归属地的分布来从侧面进行分析。图 7 为两类用户作为主叫呼叫号码的归属城市数 CDF 图，观察可以发现标注诈骗用户的呼叫目标以非同城号码为主，可能不存在目标集中于某城市来规避检测的情况；该曲线还反映出无标注正常用户大多数情况下的呼叫目标是同城用户。

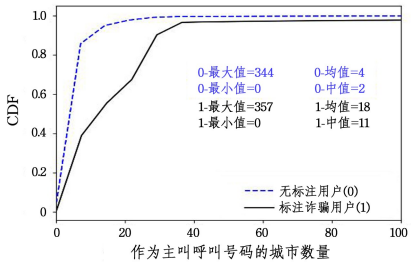


图 7　两类用户作为主叫呼叫号码归属城市数 CDF

Fig. 7　CDF of the numbers of cities to which outgoing calls belong

此外，还需排除诈骗用户省份内作案的可能（即仅向某省份下的城市分派呼叫作案）。我们分析了两类用户作为主叫呼叫号码的归属省份数 CDF，如图 8 所示。可以看出，标注诈骗用户在该维度上显著区别于无标注用户，标注诈骗用户均值是无标注用户的 4 倍，中值是无标注用户 5 倍。这也排除了诈骗者省份内作案的可能，证实了诈骗者效率和利益至上的趋利本质。

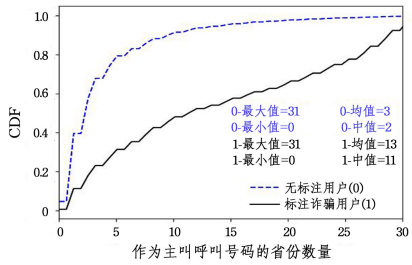


图 8　两类用户作为主叫呼叫号码的归属省份数 CDF

Fig. 8　CDF of the numbers of provinces to which outgoing calls belong

本文还从年龄维度上进行了观察分析。据调研，很多诈骗者会使用老年人的身份信息购买电话卡，于是我们分析了两类用户的年龄 CDF 分布，如图 9 所示。可以看到，标注诈骗用户曲线在 60 岁附近有明显的转折，即高于 60 岁的比例较高，而无标注用户曲线则比较平滑。这一定程度上证实了前文的信息。

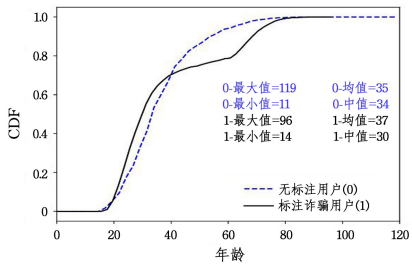


图 9　两类用户年龄 CDF

Fig. 9　CDF of user ages

对于计算得到的特征“作为主叫平均每个号码的呼叫次数”，从诈骗者的角度，他们不会多次向相同号码发起呼叫，而正常用户则会向其社交圈内的用户频繁呼叫，故我们对该特征分布进行了分析。图 10 给出了两类用户在该维度上的 CDF 分布，图中标注诈骗用户曲线在无标注用户上方，且统计值显著低于后者；另外，80% 的标注诈骗用户在一个月内平均每个被叫号码的呼叫次数小于 5，而对应的无标注用户则

仅有不足14%在一个月内该特征的值小于5;实曲线从起点开始的斜率也远大于虚曲线,说明诈骗者对平均每个被叫号码的拨打次数远少于正常用户,印证了我们的猜想。

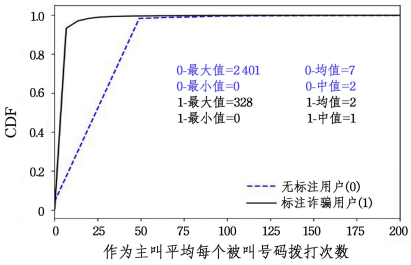


图10 两类用户作为主叫平均每个被叫号码拨打次数CDF

Fig. 10 CDF of the average numbers of calls to each called phone number

4 基于攻击经济学的虚商诈骗检测

在确保数据集标签的正确性后,需要结合诈骗用户最大化期望收益的动机和行为特点,提取可以准确甄别诈骗者的关键特征。本节首先阐述如何对虚商的带标签话单特征数据集进行基于攻击经济学的风险收益建模,结合相关性分析提取关键特征;然后对基于机器学习算法的检测模型的设计和效果进行说明。

4.1 基于风险收益建模分析的特征工程

在确保数据集标注的质量后,我们对诈骗用户的时空特征进行了深入细致的观察分析。根据上文中的观察可以发现,我们过滤了无效特征,并尝试根据保留的特征,对诈骗用户进行风险收益建模分析,以提取可以有效检测诈骗用户的关键特征。根据小米移动虚商从业者和警方的相关资料,电信诈骗用户的普遍行为模式是购买大量SIM卡,并在每张卡被检测和处理前最大化其期望收益。因此可以考虑基于话单数据对每个诈骗号码的平均存活时间进行建模,进而计算每个诈骗号的期望收益。为在存活时间内尽可能地提高期望收益,诈骗者会在时间特征(如呼叫频率、通话时长等)和空间特征(如号码离散度等)上与正常用户存在明显差异。我们从攻击经济学角度选取了相关话单特征来建模诈骗号的存活时间和期望收益。具体来讲,令诈骗号存活的天数为D,作为主叫平均每天通话次数为N,平均每天呼叫的不同号码数为K,平均通话时长为t(单位为s)。据此可知N/K为该号码平均每天对每个不同被叫号码的平均呼叫次数。我们的目标是使用上述话单特征N,K,t对函数进行建模。根据先验知识,D与N和K成反比,与t成正比,与N/K相关联。此外,N与K趋于无穷大时,D趋于0,因此反比例函数是合理适配的函数模型,初步建模如式(4)所示:

$$D=\frac{w_1\times t+b_1}{w_2\times\frac{N}{K}+w_3\times K+b_2} \tag{4}$$

其中w_1,w_2,w_3为权重参数,b_1和b_2为偏移量。

使用诈骗标注数据集对式(4)拟合后可得式(5):

$$D=\frac{5.59\times t+1\,373.08}{0.221\times\frac{N}{K}+6.96\times K+28.93} \tag{5}$$

为验证式(5)的正确性与合理性,我们对其进行了R方检验[35],检验值为0.8,说明拟合效果较好。此外拟合参数本身也呈现出一些特征:对于诈骗用户来说,呼叫的高频与离散的性质突出,t值小、K值大、N/K趋近于1,因此t与K的权重较大,而高K值导致分母较大。为使拟合函数逼近真实存活时间,分子的偏移量相应取值较大,一定程度上说明了拟合函数的合理性。随后令诈骗号的总期望收益为P,令诈骗号的每个目标用户的期望收益为E,则可知P满足式(6)。

$$P=K\times D\times E \tag{6}$$

因此式(6)结合式(5)可以得到P的函数建模。为最大化期望收益P,需要确定自变量N,K,t的值域。我们结合标注数据集的统计结果和工业应用时的启发式设置,确定了3个变量的值域,分别为$N\in[1,350]$,$K\in[1,350]$,$t\in[1,600]$,且需要满足$N\geqslant K$。在上述值域限定下,可以取得的最大期望收益$P_{\max}=671.15E$,对应的参数取值为$N=350$,$K=350$,$t=600$。我们还统计了标注诈骗号码在这3项特征上的平均值($N=35$,$K=34$,$t=95$)并据此计算期望收益,得到结果$P_{\text{avg}}=244E$。此外,若诈骗用户想在时空特征上模仿正常用户以规避检测,对应正常用户在上述特征上的统计均值$N=1$,$K=1$,$t=120$,代入式(5)和式(6)可得$P_{\text{normal}}=57E$,仅为平均期望收益的23%。因此若诈骗用户希望通过在这些维度上模仿正常用户来规避检测,则很难获得高收益,因为只有时空维度的呼叫频率足够高,才能逼近理论期望收益。如此低的攻击频率和期望收益是逐利的诈骗者无法接受的。故可以一定程度上证明,利用这些特征几乎可以彻底排除所有理性的逐利诈骗者,可以认为诈骗用户不会在我们所提取的全部话单特征上与正常用户相似。

此外,为了在统计层面观察各项话单相关特征与诈骗标签的关联,我们还对所有话单特征进行了统计相关性分析。具体来说,我们将每一个维度的话单特征以及标签各构建为一个向量,并分别计算了每一个话单特征向量和标签向量的斯皮尔曼等级相关系数(Spearman Rank Correlation Coefficient,SRCC)[36]。该相关系数是衡量变量之间依赖程度的非参数指标。对于排序(升序或降序均可)后的特征向量$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$和标签向量$\boldsymbol{Y}=\{y_1,y_2,\cdots,y_n\}$,斯皮尔曼等级相关系数的计算式如式(7)所示:

$$SRCC=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{7}$$

根据式(7)可以计算得到话单特征的斯皮尔曼等级相关系数绝对值排序,如图11所示,该图展示了相关性系数的绝对值最高的8个特征(其余特征的相关性系数过低,且在控制变量的实验中我们观察到其余特征无法提升模型效果)。其中相关系数绝对值最高(为0.39)的特征为“作为主叫平均每天呼叫次数”,可见呼叫频率是鉴别诈骗用户的重要参考指标之一。

我们将这8个特征与风险收益建模中提取的相关特征取并集,得到了最终的特征集合,包含9个特征,分别是作为主叫平均每天呼叫次数、作为主叫呼叫的不同归属省份

数量、作为主叫呼叫的不同归属城市数量、作为主叫呼叫的不同号码数量、首话单时间和激活时间间隔天数、作为主叫利用最频繁基站的呼叫占比、作为主叫呼叫最多的归属城市的呼叫占比、平均每个被叫号码呼叫次数以及作为主叫平均呼叫时长，用于诈骗检测模型训练。可以看出这两组特征重合度较高，一定程度证明了特征提取的有效性。

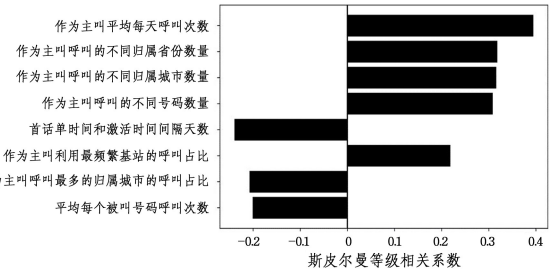


图 11　话单特征的斯皮尔曼等级相关系数 Top 8

Fig. 11　Top 8 Spearman rank correlation coefficients of CDR features

4.2　基于类别不平衡数据的诈骗检测模型

通过基于攻击经济学的风险收益建模结合统计相关性分析，我们确定了可以有效甄别诈骗用户的特征，并据此构建了用于训练机器学习模型的数据集。尽管虚商是电信诈骗用户的重灾区，但考虑到电信技术的普及及其用户规模，实际上诈骗用户在比例上要远远低于正常电信用户，因此数据集呈现出类别不平衡的现象，而直接应用类别不平衡的数据集训练会导致产生的模型对分类边界附近的样本误判率很高[37]。在初期的研究中，我们通过实验确证了这一点：我们直接应用类别不平衡数据集来进行算法设计和模型训练，发现效果并不理想（召回率和精度低于 80%），误判用户大多位于分类边界附近，即部分特征和正常用户较为相近，因此数据集类别不平衡是模型效果不佳的主要原因之一。为此需要采用重采样方法以及平衡类别权重的机器学习方法。

在数据采样方面我们考虑虚商话单特征和样本分布，基于对诈骗团伙规避检测的隐性特征观察（很多诈骗标注用户接近分类边界），针对性地采用了过采样与欠采样结合的 SMOTETomek 采样法。该方法将过采样方法 Borderline-SMOTE 和欠采样方法 Tomek link 相结合。Borderline-SMOTE 方法利用分类边界的少数类别样本来生成新样本，改善样本类别分布。具体来说，我们的数据集 S 中包括少数类别（标注为诈骗）样本集 P 和多数类别（标注为正常）样本集 Q，P 和 Q 中的样本数分别为 n_P 和 n_Q，对于每个少数类别样本 $\boldsymbol{P}_i(i\in[1,n_P])$，计算其在数据集中的 k 个最近邻，在这 k 个最近邻中类别为 P 的样本数为 n_{kP}，若 $n_{kP}=0$，则可以认为该少数类别样本是噪声样本；若 $n_{kP}\leqslant k/2$，则认为该样本为危险样本；若 $n_{kP}\geqslant k/2$，则认为该样本为安全样本。

图 12 给出了 $k=5$ 时样本分布的例子，其中圆形样本为多数类别样本，星形样本为少数类别样本。可以观察到少数类别的噪声样本比较突兀，因为周围大多是多数类别样本，因此被视作噪声；安全样本由于其最近邻样本大多为同样类别的少数类别样本，故不容易被模型误判；而危险样本则恰好处于分类边界附近，且其最近邻样本中多数类别样本居多，因此易被误判。实际生成样本时，设危险样本集为 DS，其样本数为 n_{DS}，则 $DS=\{\widetilde{\boldsymbol{P}_1},\widetilde{\boldsymbol{P}_2},\cdots,\widetilde{\boldsymbol{P}_{n_{DS}}}\}\in P$，我们对 DS 中每个危险样本选择其在 P 中的 t 个最近邻，从中选一个计算该样本与其的距离 d，根据式(8)生成新样本。

$$\boldsymbol{P}_{\text{new}}=\widetilde{\boldsymbol{P}_i}+d\times\delta,\delta\in[0,1] \tag{8}$$

其中，δ 是 0 到 1 之间的随机数，$\boldsymbol{P}_{\text{new}}$ 为生成的新少数类别样本，据此循环执行生成过程，直至总计生成 $t\times d$ 个新少数类别样本。接下来就可以用 Tomek link 方法对加入生成的新样本后的数据集进行欠采样来进一步平衡类别。具体来讲，给定两个有不同标记（诈骗和正常）的样本 $\boldsymbol{a}$ 和 $\boldsymbol{b}$，每个样本是一个 9 维特征向量 $\{a_1,a_2,\cdots,a_9\}$ 和 $\{b_1,b_2,\cdots,b_9\}$，特征维度对应于前文所述的 9 个特征，则令样本 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 之间的欧几里得距离表示为 $\delta(\boldsymbol{a},\boldsymbol{b})$，如式(9)所示：

$$\delta(\boldsymbol{a},\boldsymbol{b})=\sqrt{\sum_{i=1}^{9}(a_i-b_i)^2} \tag{9}$$

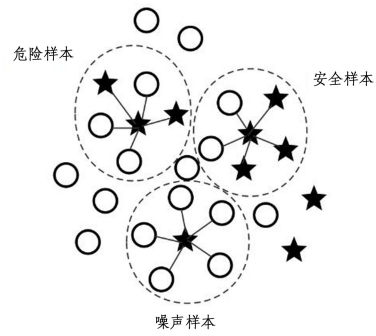


图 12　Borderline-SMOTE 算法的样本定义（$k=5$）

Fig. 12　Definition of samples in Borderline-SMOTE($k=5$)

若不存在样本 $\boldsymbol{c}$ 使 $\delta(\boldsymbol{a},\boldsymbol{c})<\delta(\boldsymbol{a},\boldsymbol{b})$ 或 $\delta(\boldsymbol{b},\boldsymbol{c})<\delta(\boldsymbol{b},\boldsymbol{a})$，则将样本对 $(\boldsymbol{a},\boldsymbol{b})$ 称为一个 Tomek link，所有组成 Tomek link 的样本被称为边界样本，欠采样过程则是将 Tomek link 中的多数类别样本筛除来平衡类别。综上，通过合理地结合 Borderline-SMOTE 过采样和 Tomek link 欠采样方法，得到的集成采样方法 SMOTETomek 可以有效平衡诈骗标注数据集中的类别，提升模型预测的召回率和精度。

我们对主流的机器学习算法（支持向量机[38]、多层感知机[39]、梯度提升决策树[40]、平衡分包[41]和平衡随机森林[42]等）进行了参数搜索和效果对比分析，同时在实验时为算法平衡了数据集类别的权重，结果如表 2 所列。不同算法采用了网格搜索得到的最优参数设置。对比可知，整体来看，平衡随机森林有着最好的表现，召回率为 92%，精度为 90%，相应的 F1 分数为 0.91。这是因为权重平衡的树形分类器可以有效对基于攻击经济学所选取的关键特征进行重要性排序和划分。表 3 列出了对平衡随机森林的参数调整的效果对比，可以明显看出划分最小样本数为 10 时模型的精度明显高于该参数值为 50 时的精度，说明诈骗用户特征不呈现明显的相似和聚集性质；此外，子树数目越多模型的表现和泛化能力越好，在该数据集上一样成立，且对可以看出使用 Gini 系数[43]作为划分标准比信息熵[44]（Entropy）效果更好。平衡分包算法（基分类器为梯度提升决策树）尽管精度极高（99.3%），但

召回率很低(73.8%)。单独的梯度提升决策树也有一定效果(精度为83.1%,召回率为77.1%),但综合评估其F1分数相比利用其为基分类器的集成算法(平衡分包和平衡随机森林)效果略差。此外尽管我们对多层感知机、支持向量机等方法进行了长期的参数搜索和实验,但它们的最佳效果都难以支持实际应用。

表2 机器学习算法效果对比

Table 2 Performance comparison of ML algorithms

机器学习算法	精度/%	召回率/%	F1分数
支持向量机	84.9	66.2	0.74
多层感知机	79.3	76.7	0.78
梯度提升决策树	83.1	77.1	0.80
平衡分包	99.3	73.8	0.85
平衡随机森林	90.0	92.4	0.91

表3 不同参数设置下平衡随机森林效果对比

Table 3 Performance comparison of balanced random forest with different parameter settings

划分标准	子树数目	划分最小样本数	精度/%	召回率/%
Entropy	50	10	72.9	82.0
Entropy	50	50	50.0	86.7
Entropy	100	10	69.1	82.2
Entropy	100	50	52.9	87.8
Gini	50	10	81.0	85.5
Gini	50	50	51.2	86.7
Gini	100	10	90.0	92.4
Gini	100	50	52.4	86.5

5 基于特殊统计模型的两阶段客服验证

在将我们设计的基于攻击经济学的诈骗检测模型部署于小米移动后,统计发现模型平均每月可以检出约5000个诈骗号码。尽管我们的方法可以有效检测诈骗用户,但为了保证模型的高召回率,其精度无法达到100%,这意味着模型预测必然会将部分正常用户误判为诈骗用户。虚商需要谨慎且准确地界定这种情况,并恰当地处理,因为诈骗用户的误判会使这部分虚商用户体验严重受损,进而影响虚商声誉,不利于虚商的运营发展。基于和小米移动的合作调研,我们发现现有的基于机器学习的诈骗检测模型在实际应用中大多采用基于用户先验知识的启发式规则来直接验证,尽管该方法可验证大部分模型预测结果,但仍有小部分预测置信度低的用户难以准确验证,因此需探索更科学有效的验证策略。

为解决上述问题,我们和小米移动配合,通过其他来源(12321官方举报、公安部举报、基础运营商众包标注等)的标注以及先验规则判定,可以明确90%的模型预测诈骗号码的真实标签。对于这部分号码我们会进行二次实名和严格审查;而对于无法明确的10%存疑号码(每月约500个),考虑到误判用户的体验而不能采用严格的二次实名和审查,于是我们设计了一套基于特殊化统计模型的两阶段客服验证策略,可以在不影响这部分用户体验的前提下,既有效验证其预测标签的正确性,又最大化节约虚商的验证成本。

对于10%存疑号码,直觉上的验证方案是直接通过电话、短信或其他触达方式来对其进行回访。对诈骗存疑用户的回访可能会出现以下情况:

(1)被访用户无反馈。若被访用户确实为诈骗者,则无反馈的可能性极高。然而根据小米移动的数据,即使是正常用户对于短信/电话/公众号问卷等触达方式的反馈率也较低,因此这种情况很难准确判断。

(2)被访用户有反馈,但不真实。若被访用户确实为诈骗用户且有反馈,则必然不真实。但大多数客服工作人员不具备通过话术判断用户是否为真实诈骗者的技巧,因此这种情形也无法准确判断。

考虑到上述原因,我们从存疑用户的历史呼叫对象中随机选取2个进行回访。一方面是考虑到只回访1个对象的参考价值不高,且有可能回访到诈骗者的同伙;另一方面选取更多对象回访则虚商客服成本过高。尽管采用这种回访方式有一定可能会回访到诈骗者的同伙,但可以通过小概率原理证明这种情况的概率微乎其微。具体来讲,我们通过统计发现,平均每个标注的真实诈骗用户呼叫的不同号码数量为575,而标注的真实诈骗用户互相之间通话的不同号码数平均为7个,因此设回访某真实诈骗用户,且恰好回访到其同伙的次数为$n(n\in[0,2])$,则可以计算出$P(n=0)=97.6\%$,$P(n=1)=2.4\%$,$P(n=2)=0.01\%$。因此我们回访2个对象恰好回访到其同伙的期望次数$E(n)=0.02$,非常接近0。根据小概率原理[45],小于5%或1%概率的事件在一次实验中不会发生,因此随机选择两个存疑用户的历史呼叫对象进行回访在理论上是有效的。

在此基础上,我们还需考虑虚商的验证成本。首先小米移动在触达用户方面,每条短信成本约为$cost_m=0.05$RMB,每通电话成本(将客服人力成本考虑在内)约为$cost_c=2.08$RMB。可见电话回访的成本远高于短信回访。因此我们考虑通过合理设计两阶段(分别使用两种回访方式)回访方法,即先通过短信触达,再对无反馈用户采用电话回访来减少成本。这种方法需要了解短信推送的点击率,我们观察小米移动的数据发现普通的短信推送点击率仅为2.2%,但令人惊讶的是在短信内加入用户流量/话费余量提醒的相关信息后点击率提升到了8.5%。此外,考虑到短信成本远远低于拨打电话的成本,因此可以考虑多次执行短信推送,结合电话回访,建立成本最优化的回访模型,故需要根据统计计算确定发送短信的次数。

通过对小米移动用户接收多次短信推送的点击率数据序列的观察和分析发现,点击率序列呈指数下降。具体来说,设初始点击率(接收第1条短信的点击率)为$r_1=8.5\%$,则接收第k条短信的点击率如式(10)所示:

$$r_k=r_1\times(0.5)^{k-1} \tag{10}$$

根据该分布,我们以1条短信为粒度推导这个过程,每月需要验证的号码数$N=500$,每条短信的成本为$cost_m$,若发送第一条短信后即对无反馈用户电话回访,则期望成本EC_1的计算如式(11)所示:

$$EC_1=(cost_m+(1-r_1)cost_c)N=976.6\text{RMB} \tag{11}$$

若发送第一轮短信后,对无反馈用户再发送一轮短信,然后对剩余的无反馈用户进行电话回访,则期望成本EC_2的计算式如式(12)所示:

$$EC_2=(cost_m+(1-r_1)(cost_m+(1-r_2)cost_c))N=959\text{RMB} \tag{12}$$

可见期望成本减少了 2%，以此类推可以计算得到发送 3 轮短信后再电话回访的期望成本 $EC_3=961.5$ RMB。该期望成本相比发送两轮短信时不减反增，而由于短信点击率指数级下降，会无限接近于 0，继续增加推送轮数只会增加期望成本。因此我们确定发送两轮短信后再对无反馈用户进行电话回访验证可以在确保成功验证的前提下尽可能降低验证成本（每月为 959RMB）。

结合上述验证方法的诈骗检测方案在小米移动的真实工业场景下部署后，经过跟踪测试，取得了很好的实际效果：6 个月内虚商诈骗用户比例降低为原来的 1/16（从 0.37‰降至 0.023‰），远低于基础运营商在充分信息前提下达到的0.1‰。相应地，12321 平台投诉量下降 90%，公安部涉案号码数量下降 75%，证明了所提方案的有效性。

结束语　本文基于拥有超 200 万客户的大型且具有代表性的轻型虚商小米移动，对移动虚拟运营生态系统发展的关键痛点——诈骗用户的检测和处理进行了深入研究。研究揭示出虚商诈骗用户的行为几乎总是受（潜在）暴利驱动才会发生，这也是其通信会话行为呈现时间维度的高频性、空间维度的离散性的原因。基于对诈骗用户通信会话特征的深度洞察，我们提出了基于攻击经济学的风险收益建模方法来提取关键特征，采用了符合虚商应用场景的采样和机器学习方法来训练检测模型，并对模型预测的存疑用户设计了基于特殊化统计模型的两阶段验证策略来验证预测结果并在确保成功验证的前提下尽可能降低虚商成本。研究成果在小米移动大规模部署中效果显著，为电信行业的净化做出了真实有效的贡献。

尽管所提方法能够大幅降低虚商的诈骗用户比例，但仍存在仅靠话单特征难以甄别的漏网之鱼，对这类攻击者的检测则需要基础运营商的更详细的数据支撑，有待于虚商与基础运营商的深化合作。另一方面，虚商体量和规模的不断发展，不仅可以升级活体认证能力，与手机供应商甚至制造商的深入合作还可以获取更细粒度的设备相关的数据（如设备型号价格等），与其他维度关联分析能进一步提高诈骗检测准确率，是未来的发展方向之一。

参 考 文 献

[1] DAHLMAN E, MILDH G, PARKVALL S, et al. 5G evolution and beyond[J]. IEICE Transactions on Communications, 2021, 104(9): 984-991.

[2] CHAN P P K, LIU W, CHEN D, et al. Face liveness detection using a flash against 2D spoofing attack[J]. IEEE Transactions on Information Forensics and Security, 2017, 13(2): 521-534.

[3] PAN G, SUN L, WU Z, et al. Monocular camera-based face liveness detection by combining eyeblink and scene context[J]. Telecommunication Systems, 2011, 47(3): 215-225.

[4] DALE K, SUNKAVALLI K, JOHNSON M K, et al. Video face replacement[C]// Proceedings of the ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH). 2011: 1-10.

[5] XIAO A, LIU Y, LI Y, et al. An in-depth study of commercial MVNO: Measurement and optimization[C]// Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys). 2019: 457-468.

[6] LI Y, ZHENG J, LI Z, et al. Understanding the ecosystem and addressing the fundamental concerns of commercial MVNO[J]. IEEE/ACM Transactions on Networking, 2020, 28(3): 1364-1377.

[7] GOPAL R K, MEHER S K. A rule-based approach for anomaly detection in subscriber usage pattern[C]// Proceedings of the World Academy Science, Engineering Technology (WASET). 2007: 396-399.

[8] ZHOU C, LIN Z. Study on fraud detection of telecom industry based on rough set[C]// Proceedings of the Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2018: 15-19.

[9] NORTHCUTT C, JIANG L, CHUANG I. Confident learning: estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research, 2021, 70(1): 1373-1411.

[10] WANG Z, WU C, ZHENG K, et al. SMOTETomek-based resampling for personality recognition[J]. IEEE Access, 2019, 7(1): 129678-129689.

[11] DEVI D, BISWAS S K, PURKAVASTHA B, et al. Redundancy-driven modified Tomek-link based undersampling: a solution to class imba-lance[J]. Pattern Recognition Letters, 2017, 93(C): 3-12.

[12] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// Proceedings of the International Conference on Intelligent Computing (ICIC). Springer, 2005: 878-887.

[13] SUBUDHI S, PANIGRAHI S. Use of possibilistic fuzzy c-means clustering for telecom fraud detection[C]// Proceedings of the International Conference on Computational Intelligence in Data Mining (CIDM). Springer, 2017, 10: 633-641.

[14] LI R, ZHANG Y, TUO Y, et al. A novel method for detecting telecom fraud user[C]// Proceedings of the International Conference on Information Systems Engineering (ICISE). IEEE, 2018: 46-50.

[15] CHOUIEKH A, HAJ E H I E. Convnets for fraud detection analysis[J]. Procedia Computer Science, 2018, 127(C): 133-138.

[16] LIU X, WANG X G. Probabilistic graphical model based approach for bank telecommunication fraud detection[J]. Computer Science, 2018, 45(7): 122-128.

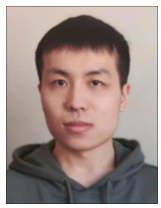
[17] LU C, LIN S, LIU X, et al. Telecom fraud identification based on ADASYN and random forest[C]// Proceedings of the International Conference on Computer and Communication Systems (ICCCS). IEEE, 2020: 447-452.

[18] NI P, YU W. A victim-based framework for telecom fraud analysis: a Bayesian network model[J]. Computational Intelligence and Neuroscience, 2022, 2022: 1-13.

[19] HU X, CHEN H, LIU S, et al. BTG: a bridge to graph machine learning in telecommunications fraud detection[J]. Future Generation Computer Systems, 2022, 137: 274-287.

[20] XU H K, JIANG T T, LI X, et al. BiLSTM network fraud phone recognition based on attention mechanism[J]. Computer Systems and Applications, 2022, 31(3): 326-332.

[21] KARUNATHILAKA A. Fraud detection on international direct dial calls[M]//Diss. 2021.
[22] KASHIR M, BASHIR S. Machine learning techniques for sim box fraud detection[C]// Proceedings of the International Conference on Communication Technologies (ComTech). IEEE, 2019: 4-8.
[23] ZHOU G M, CHEN G X, ZHOU Y Z. Research on telecom fraud user behavior based on CDR analysis[J]. Information Security and Communication Privacy, 2015, 11(1): 114-118.
[24] LI Y, LIN H, LI Z, et al. A nationwide study on cellular reliability: measurement, analysis, and enhancements[C]// Proceedings of the 2021 ACM SIGCOMM 2021 Conference. 2021: 597-609.
[25] MIN X, LIN R. K-means algorithm: fraud detection based on signaling data[C]// Proceedings of the IEEE World Congress on Services. IEEE, 2018: 21-22.
[26] LIU M, LIAO J, WANG J, et al. AGRM: attention-based graph representation model for telecom fraud detection[C]// Proceedings of the International Conference of Communications(ICC). IEEE, 2019: 1-6.
[27] TERZI D S, SAGIROGLU Ş, KILINC H. Telecom fraud detection with big data analytics[J]. International Journal of Data-Science, 2021, 6(3): 191-204.
[28] CHADYSAS V, BUGAJEV A, KRIAUZIENE R, et al. Outlier analysis for telecom fraud detection[C]// Proceedings of the International Baltic Conference on Digital Business and Intelligent Systems(CCIS). Cham: Springer, 2022: 219-231.
[29] JIANG Y, LIU G, WU J, et al. Telecom fraud detection via Hawkes-enhanced sequence model[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(5): 5311-5324.
[30] KRASIC I, ČELAR S. Telecom fraud detection with machine learning on imbalanced dataset[C]// Proceedings of 2022 International Conference on Software, Telecommunications and Computer Networks(SoftCOM). IEEE, 2022: 1-6.
[31] ZHANG J J, TANG Y C, JI S Y. A telecom fraud identification method based on graph neural network[J]. Application of Electronic Technique, 2021, 47(6): 25-29.
[32] YANG J K, XIA W C. Fraud call identification based on user behavior analysis[J]. Computer Systems and Applications, 2021, 30(8): 311-316.
[33] LIU G, GUO J, ZUO Y, et al. Fraud detection via behavioral sequence embedding[J]. Knowledge and Information Systems, 2020, 62(7): 2685-2708.
[34] FAN Y R, YANG T, KONG H F, et al. Calling features mining method of telecom fraud based on knowledge graph[J]. Computer Applications and Software, 2019, 36(11): 182-187.
[35] CAMERON A C, WINDMEIJER F A G. An R-squared measure of goodness of fit for some common nonlinear regression models [J]. Journal of Econometrics, 1997, 77(2): 329-342.
[36] ZAR J H. Spearman rank correlation[M]//Encyclopedia of Biostatistics. 2005.
[37] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.
[38] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.
[39] GARDNER M W, DORLING S R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences[J]. Atmospheric Environment, 1998, 32(14/15): 2627-2636.
[40] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001, 29(1): 1189-1232.
[41] HIDO S, KASHIMA H, TAKAHASHI Y. Roughly balanced bagging for imbalanced data[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2009, 2(5/6): 412-426.
[42] AGUSTA Z P. Modified balanced random forest for improving imbalanced data prediction[J]. International Journal of Advances in Intelligent Informatics, 2019, 5(1): 58-65.
[43] NEMBRINI S, KONIG I R, WRIGHT M N. The revival of the Gini importance[J]. Bioinformatics, 2018, 34(21): 3711-3718.
[44] DONG X, QIAN M, JIANG R. Packet classification based on the decision tree with information entropy[J]. The Journal of Supercomputing, 2020, 76(6): 4117-4131.
[45] ZHANG Z, PU P, HAN D, et al. Self-adaptive louvain algorithm: fast and stable community detection algorithm based on the principle of small probability event[J]. Physica A: Statistical Mechanics and its Applications, 2018, 506(1): 975-986.



LI Yang, born in 1996, Ph.D candidate. His main research interests include network measurement and data mining.

LI Zhenhua, born in 1983, Ph.D, associate professor, Ph.D supervisor, is a senior member of China Computer Federation. His main research interests include mobile network measurement and virtualization technology.

(责任编辑:李亚辉)