



机器学习在物联网虚假用户识别中的运用

张溶芳¹, 许丹丹¹, 王元光², 潘思宇¹, 李正茂³

(1. 中国联合网络通信有限公司研究院, 北京 100176;

2. 中国联合网络通信集团有限公司电子商务中心, 北京 100033;

3. 中国联合网络通信有限公司重庆市分公司, 重庆 401121)

摘要: 随着通信技术的发展, 物联网卡和 5G 技术将得到大规模应用, 但存在个别企业利用物联网卡资费便宜、没有实名制等特点从中非法牟利、破坏社会稳定的问题, 不利于行业健康发展。因此如何识别虚假用户成为物联网行业研究的重要课题。主要研究了在实时海量的物联网终端数据中, 如何运用机器学习模型高效地识别疑似虚假用户。具体来看, 通过研究相关数据的特征, 采用基于正样本和未标记样本的半监督式学习模型建立实时监控异常行为的模型, 达到识别物联网行业中潜在虚假用户的目的。本研究成果体现在节约大量人力物力的同时, 可以帮助相关部门、人员及时发现用户的异常行为, 采取相应的措施避免产生较大损失, 具有广泛的行业应用前景。

关键词: 物联网; 半监督式学习模型; 朴素贝叶斯分类器; 随机森林; 支持向量机; SPY 分类器

中图分类号: TP18; TN929.5

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019090

Application of machine learning in the fake user identification of IoT

ZHANG Rongfang¹, XU Dandan¹, WANG Yuanguang², PAN Siyu¹, LI Zhengmao³

1. Research Institute of China United Network Communication Co., Ltd., Beijing 100176, China

2. E-Commerce Center, China United Network Communication Co., Ltd., Beijing 100033, China

3. Chongqing Branch of China United Network Communication Co., Ltd., Chongqing 401121, China

Abstract: With the development of communication technology, IoT cards and 5G technologies will be applied on a large scale. However, there are some companies have taken advantage of the fact that the price of SIM cards of IoT is cheap and the cards do not have real-name registration system. It is harmful to social stability, which is not conducive to the development of IoT industry. So how to identify these fake users has become an important topic in IoT industry. The purpose was to use machine learning models to identify users who have high suspiciousness effectively. By studying the characteristics of relevant data, a semi-supervised learning model based on positive and unlabeled samples was used to establish a real-time abnormal behavior monitoring model to identify potential fake users in the IoT industry users. At the same time, the model greatly enhanced the working efficiency and has saved the manpower physical resources. Also, it can help relevant departments and governments to discover the abnormal behavior of users in time and take corresponding measures to avoid large losses. So, the proposed method really has broad application prospects in the industry.

Key words: IoT, positive and unlabeled semi-supervised model, Naïve Bayesian classifier, random forest, support vector machine, SPY classifier

1 引言

物联网^[1]作为继计算机、互联网之后世界信息产业发展的第三次浪潮，以惊人的速度增长，延伸到包括医疗、交通、教育、物流等的各个领域。物联网卡也随着物联网行业迎来爆发或发展，其主要面向的客户是企业，采用商业模式运营，伴随着网络技术的成熟，物联网卡和 5G 技术在未来移动通信技术中是绝对的主角。但现有小部分企业利用了物联网卡资费便宜、没有实名制等特点，将批量购入的物联网卡拿到市场上非法出售，利用其便宜的价格赚取暴利差价，同时利用未实名制的漏洞进行短信诈骗等行为。中国联通物联网部门经过一段时间数据的积累与观察、实地走访，查出少数几家在市场上非法出售物联网卡的公司，但非法出售物联网卡的公司远不止这些。随着物联网行业的飞速发展，各行各业成千上万的公司都会相继购买物联网卡，若只依靠个人举报、粗略观察数据等方式识别虚假用户，会浪费大量的人力物力。本文通过建立异常行为监控模型来实时识别潜在的虚假用户，帮助相关部门、人员及时发现用户的异常行为，提早采取措施进行制止避免不必要的损失，具有广泛的行业应用前景。

2 物联网大数据能力平台

2.1 平台基本概况

物联网平台主要分为两个部分，如图 1 所示：一部分叫做服务管理器（service manager），用于公司管理企业用户；另外一部分叫做控制中心（control center），用于企业客户管理其联网设备。

在物联网平台上，用户群会被分为 3 种层次结构，如图 2 所示，分别为公司 1 账户（企业）以及客户（分支机构），每种用户具有不同的平台数据视图，并且提供多个具有不同权限的角色。平台同时提供了数据、短信语音等服务。

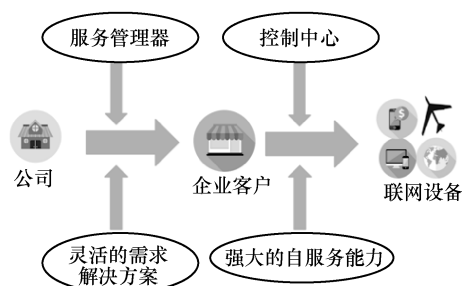


图 1 物联网平台组成部分

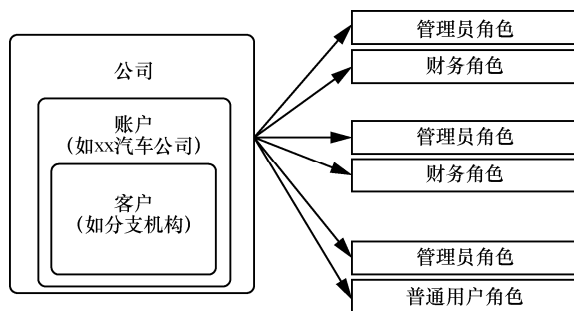


图 2 物联网平台用户和角色

2.2 平台数据服务建设

针对业务需求，物联网平台搭建资源纳管架构 CU-DCOS2.0，上层基于 DAS2.0 完成了能力组件的容器化封装，采取“大数据能力运营平台”+“运营分析门户”的双层架构，实现了从底层数据到上层应用的流程贯通。这个平台提供了 3 个功能，如图 3 所示。第一个功能是大数据能力，在已有服务器上搭建了 CDH Hadoop 集群，提供 Hadoop^[2]、Spark^[3]、Hive^[4]、HBase^[5]等大数据服务；第二个功能是数据分析处理能力，基于 CDH 集群，复用 DAS 组件，搭建可视化的数据处理 Kylin^[6]、Echarts^[7]框架、客户运营数据分析建模平台；第三个功能是容器能力，基于提供的物理机服务器搭建了 CU-DCOS 平台，提供数据库、中间件、DevOps^[8]、监控、负载均衡等服务。

物联网大数据能力平台有三大数据来源，分别是集客系统、订单系统以及 Jasper 公司。Jasper 数据包括经营管理数据，包含 40 多个数据接口，500 多个数据项，每日推送超过 30 GB 的用户行为类数据。集客数据包括增值产品、套餐订购数据、

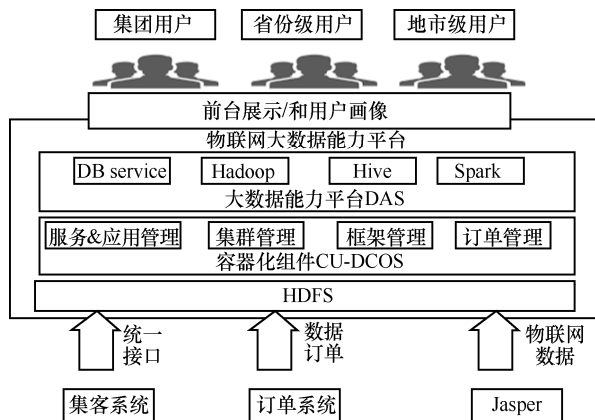


图3 物联网大数据能力平台架构

商机数据。订单系统数据包括卡类订单、订购信息。三大平台数据直接入库 HDFS^[9], 运用 Spark 对于入库文件数据进行预处理, 使用 Python 基于各种场景以及预处理的数据建立模型进行预测, 预测结果可以为用户提供很优质的服务。

3 机器学习理论

机器学习可以分为监督式学习、半监督式学习^[10]、无监督式学习以及强化学习。根据物联网部门提供的数据特征, 即已知少部分虚假用户数据, 未标注的数据当中既有虚假用户的数据又有正常用户的数据, 这使监督学习不可实行, 所以需要通过建立基于正样本和未标记样本的半监督学习模型, 来预测物联网虚假用户群。

3.1 基于正样本和未标记样本的半监督式学习

半监督学习是监督学习与无监督学习相结合的一种学习方法。半监督学习使用大量的未标记数据以及少量标记数据来进行建模。在现实任务中, 未标记样本多、有标记样本少是一个比价普遍现象, 利用好未标记样本来提升模型泛化能力, 就是半监督学习优势所在。半监督学习不仅可以极大地减少人为的工作量, 又可以同时获得准确性较高的结果。在物联网虚假用户识别的案例中, 有大量的未标记数据以及少部分确定的虚假用户数据, 这种数据特征非常适合使用于正样本和未

标记负样本的半监督学习方法进行建模。

在物联网虚假用户识别的案例中, 有大量的未标记数据以及少部分确定的虚假用户数据, 这种数据特征非常适合使用于正样本和未标记负样本的半监督学习方法进行建模, 详细内容请参考文献[11]。

解决这类问题通常有两种方式, 第一种是直接方法, 第二种是两步法。直接方法是仅在正样本上训练。以下模型讲解内容中, P 代表正样本数据, U 代表未知样本数据。下面详细介绍这两种方法的具体内容。

3.1.1 两步法

第一步根据已标注的正样本在未标注样本集中找出可靠的负样本集合, 将正样本和未标记样本问题转化为二分类的问题; 第二步利用正负样本通过迭代训练得到一个二元分类器。理论上已经证明, 如果最大化未标注样本集中负样本的个数, 同时保证正样本被正确分类, 则会得到一个性能不错的分类器。上述步骤中, 找出可靠负样本以及训练二元分类器的方法很多, 下面是对每一步可以使用模型的概述。

两步法第一步使用的模型如下。

(1) 朴素贝叶斯分类器

朴素贝叶斯分类统计的一种分类方式, 是一种利用概率论知识进行分类的算法。假设每个数样本有一个 n 维的特征向量来描述其 n 个属性的值, 即: $X = \{x_1, x_2, \dots, x_n\}$ 。假设有 m 个类别, 分别用 $C_1, C_2, \dots, C_m$ 来表示, 给定一个未知的数据样本 X , 使用朴素贝叶斯分类器将未知样本 X 分配给类别 C_i , 则一定有以下结论:

$$P(C_i | X) > P(C_j | X), 1 \leq j \leq m, i \neq j \quad (1)$$

根据贝叶斯理论可得:

$$P(C | X) = \frac{P(X | C)P(C)}{P(X)} \quad (2)$$

由于 $P(X)$ 对于所有类都是常数, 最大后验概率 $P(C_i | X)$ 可以转化为最大化先验概率

$P(X|C_i)P(C_i)$ 。如果训练数据有很多属性, 计算 $P(X|C_i)$ 的开销会非常大, 因为这是所有属性的联合概率。为此, 通常假设各个属性的取值相互独立, 这样先验概率 $P(x_1|C), P(x_2|C), \dots, P(x_n|C)$ 可以从训练的数据中得到, 计算式如式(3):

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} = \frac{P(C)}{P(X)} \prod_{i=1}^n P(x_i|C) \quad (3)$$

贝叶斯分类器分类依据如式(4):

$$h_{nb}(x) = \arg \max P(C) \prod_{i=1}^n P(x_i|C) \quad (4)$$

利用模型的基本原理以及训练数据的特征, 训练步骤如下:

步骤1 把 P 中的每个样本标记为类别1;

步骤2 把 U 中的每个样本标记为类别-1;

步骤3 使用 P 和 U 训练得到贝叶斯分类器;

步骤4 对 U 中的每个样本使用上述分类器进行分类, 如果分类结果为-1, 则把该样本加入 RN。

模型优点是: 算法数学模型稳定, 分类效果较好; 训练参数少, 模型学习简单。模型的缺点: 贝叶斯分类器算法要求样本之间相互独立, 对于样本之间关联性比较强的样本分类效果比较差; 需要知道先验概率, 且先验概率很多时候取决于假设, 假设的模型可以有很多种, 因此在某些时候会由于假设的先验模型导致预测效果不佳; 需要利用大量样本技术样本类别的概率分布, 因此少量样本分类效果差。模型的详细推导解释请参考文献[12]。

(2) SPY 分类器

SPY 的基本思想是从 P 中划分出一个子集 S 放到 U 中, 从而得到新的正样本集 $P-S$ 和未标识样本集 $U+S$ 。使用 $P-S$ 作为正样本, $U+S$ 作为负样本, 利用迭代的 EM (expectation-maximum) 算法进行分类, 当分类结束后, 利用对那些“间谍”样本的标识, 确定一个参数阈值 th , 再对 U 中的文档进行划分得到可靠的反样本集合 RN。其

中, 从 P 中划分子集 S 的数量比例一般为 15%。算法步骤描述如下:

步骤1 RN 集合置空;

步骤2 从 P 中随机选取子集 S , 得到新的正样本集 $PS=P-S$ 和未标识样本集 $US=U+S$, 记 PS 中各个样本类别为 1, US 各个样本类别为-1;

步骤3 PS 和 US 作为训练集, 用 I-EM 算法训练得到一个贝叶斯分类器;

步骤4 使用子集 S 确定出一个概率阈值 th ;

步骤5 对 US 中的每个样本 x 使用贝叶斯分类器计算其属于正类别的概率 $P(1|x)$, 如果小于阈值概率 th , 则把其加入 RN 集合。

两步法第二步使用的模型如下。

经过第一步可获得可靠的负样本, 基于正样本和可靠的负样本数据, 可以直接使用常规的分类器进行分类。所以第二步主要选择的是分类效果比较好的分类器, 如随机森林树和支持向量机 (support vector machine, SVM)。

(1) 随机森林树

集成学习 (ensemble learning) 是通过构建并结合多个学习器来完成学习任务, 主要包含两类: 一是个体学习器间存在强依赖关系、必须串行生成的序列化方法, 比如提升树 GBDT、XGBOOST 等, 主要思想是 boosting 迭代将弱学习器提升为强学习器; 二是个体学习器之间不存在强依赖关系、可同时生成的并行化方法, 该类的代表是 Bagging。随机森林是 Bagging 的一个拓展变体。

随机森林在以决策树为基础学习器构建 Bagging 集成的基础上, 进一步在决策树的训练过程中引入了随机属性选择。传统决策树在选择划分属性时是在当前节点的属性集合 (假定 m 个属性) 中选择一个最优属性; 而在随机森林中, 对基决策树的每个节点, 先从该节点的属性集合中随机选择一个包含 k 个属性的子集, 然后再从这个子集中选择一个最优属性用于划分。这里的参



数 k 控制了随机性的引入程度：若令 $k=m$ ，则决策树的构建与传统决策树相同；若令 $k=1$ ，则随机选择一个属性用于划分，一般情况下 $k=\text{lb } m$ 。

随机森林还加入了随机属性扰动，这使得最终集成的泛化性能可通过个体学习器之间差异度的增加而进一步提升。随机森林的收敛性与 Bagging 相似，随机森林的起始性能相对较差，特别是只包含一个基础学习器的时候（因为通过引入属性扰动，随机森林中个体学习器的性能往往有所降低）然而随着个体学习器数目的增加随机森林通常会收敛到更低的泛化误差。

在随机森林中，将生成很多的决策树，当在基于某些属性对一个新的对象进行分类判别时，随机森林中的每一棵树都会给出自己的分类选择，并由此进行“投票”，森林整体的输出结果将会是票数最多的分类选项，计算式如式（5）：

$$H(x) = \max_Y \sum_{i=1}^k I(h_i(x) = Y) \quad (5)$$

其中， $H(x)$ 表示随机森林分类结果， $h_i(x)$ 是单个决策树分类结果， Y 表示分类目标， $I(\cdot)$ 为示性函数，通过投票策略哪个分类最多结果就归为哪类，模型的详细推导请参考文献[13]。

（2）支持向量机

支持向量机是形式为 $f(x) = w^T x + b$ 的线性方程组， $w^T x$ 是权向量和输入向量 x 的内积。SVM 用作分类器，如果 $f(x) > 0$ ，则分为 1 类，否则分为 -1 类。SVM 的核心思想是要选出一个超平面将正负样本分开同时最大化最小边界。 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ 是一组训练数据， x_i 代表输入向量， y_i 代表分类标签， $y_i \in \{1, -1\}$ 。所以寻找超平面的问题变成了优化以下问题：

$$\text{最小化: } \frac{1}{2} w^T w \quad (6)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, n \quad (7)$$

为了处理一些因为噪音数据而造成的没有超平面的情况，提出了 Soft-margin SVM 来寻找超平

面的方法，计算式如式（8）：

$$\text{最小化: } \frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \quad (8)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, n \quad (9)$$

$$C \geq 0$$

C 可以控制训练错误的大小。模型的详细解释与推导请参考文献[14]。

（3）Biased SVM

在处理不平衡数据分类的时候，支持向量机遇到的主要问题是分类面会靠近少数类，从而会分错少数类，所以提出了一种新的策略，将正负样本的权重分开对待，对于 PU learning 问题要尽量将正样本分开，因此正样本的权重系数 C 会比较大。以下是一个效果比较好的用于第二阶段的算法，根据 SVM 模型演化而来，求解目标函数为：

$$\min \frac{1}{2} w^T w + C_+ \sum_{i=1 \dots k} \varepsilon_i + C_- \sum_{j=k+1 \dots} \varepsilon_j \quad (10)$$

$$y_i(w^T x + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0 \quad (11)$$

这种将正负样本分开对待的方式使得正确率提升很多，详细推导请参考文献[15]。

3.1.2 直接法

直接法是直接在已知的正样本集数据上进行训练，无需关心未知样本数据。直接法使用的模型是 One-Class SVM，这个模型关心的只有一个类，属于该类就返回结果“是”，不属于就返回结果“不是”，与二分类有很大的差别，在二分类问题中，训练集中由两个类的样本组成，训练出的模型可以区分两个类别；而 One-Class SVM 中的训练样本只有一类，因此训练出的分类器将不属于该类的所有其他样本判为“不是”即可，而不是由于属于另一类才返回的“不是”结果。既然只有一个类，那么可以训练出一个最小的超球面（超球面是指三维以上的空间中的球面，对应的二维空间中就是曲线），将这堆数据全部“包起来”，识别一个新的数据点时，如果这个数据

点落在超球面内, 就属于这个类, 否则不是。具体的公式推导不在此累述, 可参考文献[16]。

3.2 网格搜索

网格搜索法是指定参数值的一种穷举搜索方法, 通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。即将各个参数可能的取值进行排列组合, 列出所有可能的组合结果生成“网格”。然后将各组合用于 SVM 训练, 并使用交叉验证对表现进行评估。在拟合函数尝试了所有的参数组合后, 返回一个合适的分类器, 自动调整至最佳参数组合。详细内容请参考文献[17]。

4 实例建模

基于现有物联网平台 Hadoop 集群上存储的来自于集客系统, 订单系统以及 Jasper 平台物联网数据, 使用 Scala 语言编写的 Spark 数据加工流程提取出重要字段对应的关键数据, 采用基于正样本和未标记样本的半监督分类模型, 训练并获得甄别异常行为的模型参数。根据训练好的模型, 可以对实时采集的物联网终端数据做出及时的判断并将加工结果沉淀到 MySQL 关系型数据中, 并通过 MySQL 关系型数据库中的已有数据信息获取每周、每月、每年的物联网用户异常行为总结, 如图 4 所示。

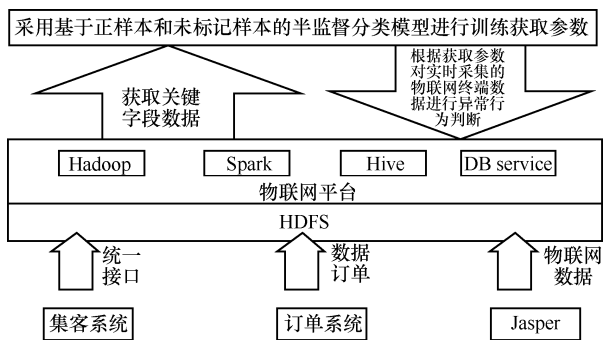


图 4 虚假用户识别模型流程

物联网虚假用户识别模型的完成主要分为 3 个模块: 数据的准备与处理、模型的选择与建立、虚假用户的识别与预测。以下内容会分别对 3 个

模块进行详细的说明。模型建立的流程如图 5 所示。

4.1 数据的准备与处理

基于对物联网行业背景的了解以及数据字典的内容解释, 根据项目需求, 分别选择了用量详情—流量、用量详情—短信、用量详情—通话、账户截图、用户信息这 5 张表进行数据特征的选择。用户详情—流量选择了 7 个字段, 用户详情—短信选择了 5 个字段, 用户详情—语音选择了 7 个字段, 账户截图选择了两个字段, 用户信息选择了两个字段, 详情见表 1。

表 1 数据特征提取

表名	字段
用户详情—流量	ICCID (SIM 卡唯一标识)
	账户 ID
	SIM 卡状态
	数据用量—全部
	数据用量—上行
	数据用量—下行
用户详情—短信	数据用量—全部 (整数)
	ICCID (SIM 卡唯一标识)
	账户 ID
	SIM 卡状态
	发送短信时间
	语音时长—全部
用户详情—语音	语音时长—全部 (整数)
	ICCID (SIM 卡唯一标识)
	账户 ID
	SIM 卡状态
	语音时长—全部
	语音时长—全部 (整数)
账户截图	账户 ID
	操作 ID
用户信息	用户姓名
	WLW_number

数据特征的选择相比于数据量来说偏少, 所以将 7 天的数据合在一起建立具有 51 个维度的数据。以 2018 年 2 月 5—11 日为例进行说明, 使用 Spark 在 Hadoop 集群上分别针对这 7 天的流量、短信、语音进行了统计性分析, 查看其中位数、众数、均值以及缺失值等基本信息。接下来进行数

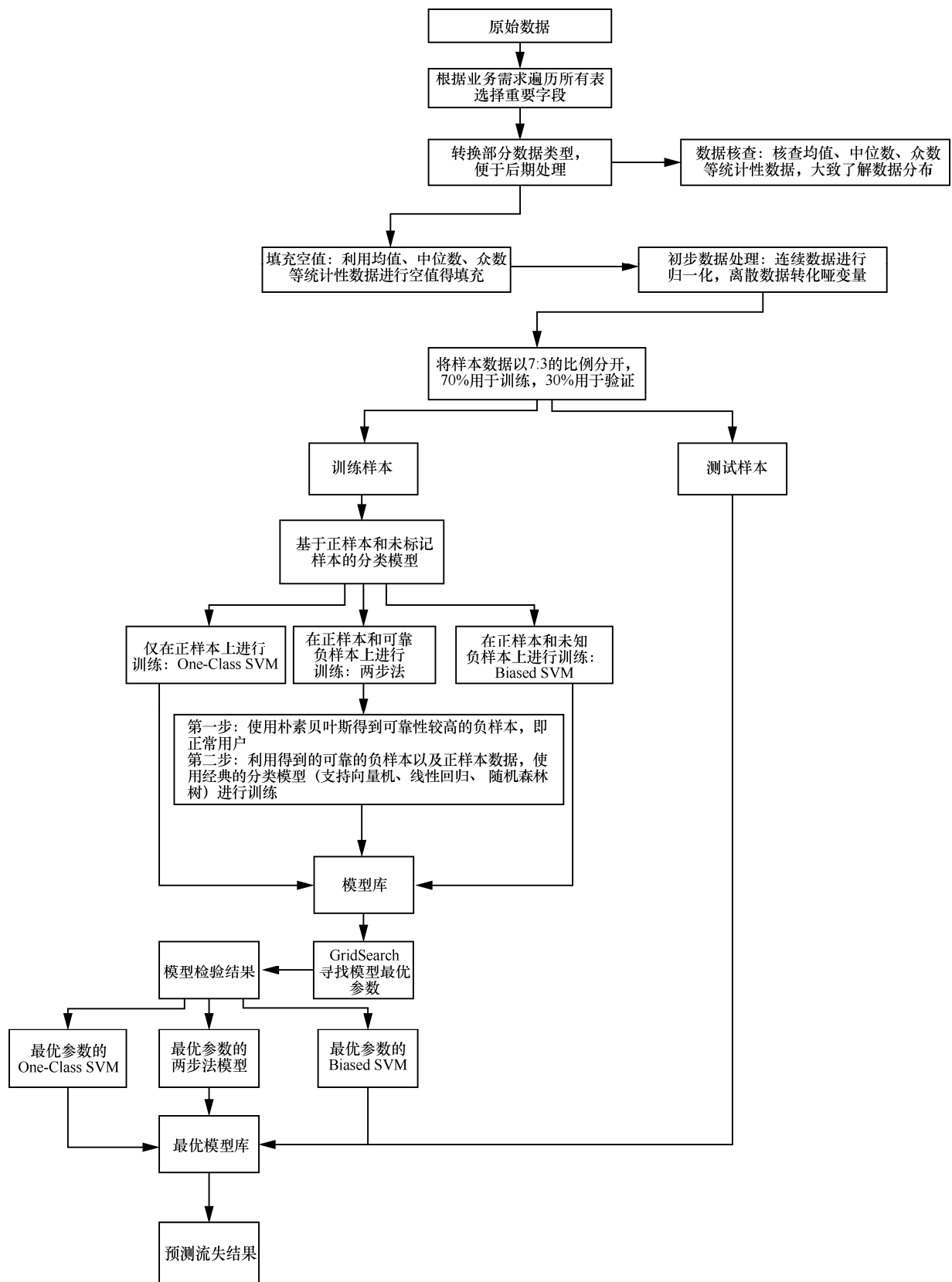


图5 虚假用户识别模型建立流程

据清洗预处理, 过滤掉一些异常数据, 同时使用补零、均值、中位数等方式进行缺失值的填充。对于清理后的数据根据后期模型的需求进行相应处理, 首先把相应字段数据的提取并且分别生成了流量、短信、数据 7 天的数据表。其次根据已知虚假用户的账户 ID 并且结合流量、短信、语音单天的表, 根据 ICCID 建立关联, 可以获得虚假用户所有物联网卡单天的流量、短信、语音的使用情况。短信表进行了预处理, 根据每天收发短信的时间, 对每张物联网卡当天的短信使用量进行了求和。因为大部分物联网卡使用的是流量业务, 所以以流量表为基准左关联语音和短信表, 7 天的数据以第一天为基准进行了内关联, 最终可以得到虚假用户每一张物联网卡 7 天 3 个业务的使用情况。与选择虚假用户的流程一致, 选择除已确定虚假用户以外的账户 ID, 进行数据、短信语音的关联, 获得未知用户的相关数据。

4.2 模型的选择与建立

将预处理的虚假用户数据和未知用户数据进行标签, 打标以后的数据进行随机的排序, 基于交叉验证的方式进行采样划分, 划分比例为 7:3, 其中 70%用于训练, 30%用于测试。在划分好的训练集上进行模型的建立, 本文分别进行了直接方法以及两步法模型的建立, 直接方法使用了 One-Class SVM、Biased SVM 模型, 两步法第一步使用了朴素贝叶斯模型, 第二步使用了随机森林、支持向量机的方式建模, 基于全搜索的方法建立充分的模型库, 在模型库中基于 GridSearch 方法寻找在训练集和测试集上都表现良好的模型最优参数, 进而使得各个模型的表现最优, 并将最优模型进行保存用于后面数据的预测。

4.3 虚假用户预测

将每周新处理的数据进行模型的调用, 输出模型的预测结果。因为已知确定的虚假用户, 对于未知用户并没有确切的推断, 所以模型预测的结果不能按照一般的精确率、召回率、准确率、

F1 得分来进行判断。针对已知数据的特征, 本文采用使用不含负样本的验证集来对模型进行性能的评估。判断准则采用 $\frac{pr}{pr[Y=1]}$ (其中, $pr[Y=1]$ 是正样本的实际概率), 其等价于 $\frac{r^2}{pr[f(x)=1]}$ (其中 $pr[f(x)=1]$ 是一个样本被分为正样本的概率)。其中, r 代表召回率, p 代表精准率, r 可以用过验证集的正样本来计算得到, $pr[f(x)=1]$ 可以由整个验证集计算得到。这个判断准则可以像 F 值一样来判断分类器的性能。

以下是各个模型的预测结果, 根据各个模型的变现情况, 最终选定两步法模型作为最终虚假用户识别模型。两步法中第一步选择了朴素贝叶斯模型, 第二步中选择了随机森林, 具体见表 2。

表 2 模型预测结果

模型	正确率
两步法	朴素贝叶斯+随机森林
	朴素贝叶斯+支持向量机
	朴素贝叶斯+Biased SVM
	SPY 分类器+随机森林
	SPY 分类器+支持向量机
	SPY 分类器+Biased SVM
直接法	One-Class SVM

5 结束语

本文从物联网行业的实际情况出发, 利用已有数据信息, 分析了当前行业虚假用户的特征, 使用机器学习模型建立了物联网虚假用户的识别模型, 可为相关公司提供较可靠的虚假用户识别依据。尽管现有模型已经可以达到一个不错的效果, 但是如果提升预测的精准性, 还有以下几点可以改善: 还需要紧密结合业务, 积极总结各个行业的行为特征, 根据每个行业的特征去提取更多有意义的字段以备后续模型的训练; 尝试更多的分类算法以及深度学习模型来建立适合不同行业特有的模型储



备;可以尝试融合模型,训练出更为精准的虚假用户识别模型;根据后期反馈数据结果,不断完善、提高模型的预测结果。

参考文献:

- [1] 刘楠, 刘露, 张第, 等. 运营商在物联网大数据竞争中的应对思考[J]. 电信科学, 2018, 34(3): 132-137.
LIU N, LIU L, ZHANG D, et al. Reflections on operators' competition in big data of internet of things[J]. Telecommunications Science, 2018, 34(3): 132-137.
- [2] 曹霞, 谢颖华. 基于 Hadoop 的协同过滤并行化算法[J]. 计算机系统应用, 2018(5).
CAO X, XIE Y H. Parallel algorithm of collaborative filtering based on Hadoop[J]. Application of Computer System, 2018(5).
- [3] 胡俊, 胡贤德, 程家兴. 基于 Spark 的大数据混合计算模型[J]. 计算机系统应用, 2015, 24(4): 214-218.
HU J, HU X D, CHENG J X. Big data hybrid computing mode based on Spark[J]. Application of Computer System, 2015, 24(4): 214-218.
- [4] 刘永增, 张晓景, 李先毅. 基于 Hadoop/Hive 的 Web 日志分析系统的设计[J]. 广西大学学报(自然科学版), 2011, 36(S1): 314-317.
LIU Y Z, ZHANG X J, LI X Y. Design of Web log analysis system based on Hadoop/Hive[J]. Journal of Guangxi University (Natural Science Edition), 2011, 36(S1): 314-317.
- [5] 范建永, 龙明, 熊伟. 基于 HBase 的矢量空间数据分布式存储研究[J]. 地理与地理信息科学, 2012, 28(5): 39-42.
FAN J Y, LONG M, XIONG W. Research of vector spatial data distributed storage based on HBase[J]. Geography and Geo-information Science, 2012, 28(5): 39-42.
- [6] 蒋守壮. 基于 Apache Kylin 构建大数据分析平台[M]. 北京: 清华大学出版社, 2016.
JIANG S Z. Construction of large data analysis platform based on Apache Kylin[M]. Beijing: Tsinghua University Press, 2016.
- [7] 王子毅, 张春海. 基于 ECharts 的数据可视化分析组件设计实现[J]. 微型机与应用, 2016, 35(14): 46-48, 51.
WANG Z Y, ZHANG C H. Design and implementation of a data visualization analysis component based on ECharts[J]. Microcomputer its Applications, 2016, 35(14): 46-48, 51.
- [8] 牛晓玲, 吴蕾. DevOps 发展现状研究[J]. 电信网技术, 2017(10).
NIU X L, WU L. Study on DevOps development status[J]. Telecommunication Network Technology, 2017(10).
- [9] 郝树魁. Hadoop HDFS 和 MapReduce 架构浅析[J]. 邮电设计技术, 2012(7): 37-42.
HAO S K. Brief analysis of the architecture of Hadoop HDFS and MapReduce [J]. Designing Techniques of Posts and Telecommunications, 2012(7): 37-42.
- [10] 钟清流, 蔡自兴. 基于支持向量机的渐近式半监督式学习算法[J]. 计算机工程与应用, 2006(25): 19-22.
ZHONG Q L, CAI Z X. Semi-supervised learning algorithm based on SVM and by gradual approach[J]. Application of Computer System, 2006(25): 19-22.
- [11] LIU B, DAI Y, LI X, et al. Building text classifiers using positive and unlabeled examples[C]//2003 IEEE International Conference on Data Mining, Nov 11, 2003, Melbourne, USA. Piscataway: IEEE Press, 2003:179.
- [12] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.
- [13] BREIMAN L. Random forest[J]. Machine Learning, 2001, 45(2): 5-32.
- [14] ADANKON M M, CHERIET M. Support vector machine[J]. Computer Science, 2002, 1(4): 1-28.
- [15] 王和勇, 樊泓坤, 姚正安. SMOTE 和 Biased SVM 相结合的不平衡数据分类[J]. 计算机科学, 2008, 35(5): 174-176.
WANG H Y, FAN H K, YAO Z A. Imbalance data set classification using SMOTE and Biased-SVM[J]. Computer Science, 2008, 35(5): 174-176.
- [16] CHEN Y Q, ZHOU X S, HUANG T S. One-class SVM for learning in image retrieval[C]//International Conference on Image Processing, Oct 7-10, 2001, Thessaloniki, Greece. Piscataway: IEEE Press, 2001: 34-37.
- [17] 王健峰, 张磊, 陈国兴, 等. 基于改进的网格搜索法的 SVM 参数优化[J]. 应用科技, 2012, 39(3): 28-31.
WANG J F, ZHANG L, CHEN G X, et al. A parameter optimization method for an SVM based on improved grid search algorithm[J]. Applied Science and Technology, 2012, 39(3): 28-31.

[作者简介]



张溶芳(1991—), 女, 中国联合网络通信有限公司研究院大数据研究中心工程师, 主要研究方向为数据挖掘、建模、数据可视化等。

许丹丹(1989—), 女, 中国联合网络通信有限公司研究院大数据研究中心工程师, 主要研究方向为数据挖掘、建模、数据可视化。

王元光(1989—), 男, 中国联合网络通信集团有限公司电子商务中心工程师, 主要研究方向为数据挖掘、数据建模等。

潘思宇(1994—), 男, 中国联合网络通信有限公司研究院大数据研究中心工程师, 主要研究方向为数据挖掘、算法研究、数学建模等。

李正茂(1981—), 男, 中国联合网络通信有限公司重庆市分公司市场部主管, 主要研究方向为通信市场产品策略、大数据分析处理等。