DOI: 10.19650/j.cnki.cjsi.J2006814

基于标签传播的涉烟车辆异常检测*

王 贞[1],尤梓荃[2],张锦程[2],甘小莺[2],陶春和[1]
(1. 公安部第三研究所经侦技术研究室 上海 200030; 2. 上海交通大学电子信息与电气工程学院 上海 200240)

摘 要:烟草行业与政府财政收入密切关联,走私假烟不仅会造成国家税收流失,同时也会扰乱市场、危害消费者的健康,如何对涉烟车辆实施有效的监管,对烟草行业的发展有重要意义。针对涉烟车辆的问题,并结合实际采集的车辆数据特征,提出了基于标签传播的涉烟车辆异常检测算法。通过对车辆数据集进行有用特征提取,并采用随机森林算法实现特征选择,在此基础上使用标签传播算法对异常车辆进行分类。结果表明,在历史数据和异常车辆标签较少的情况下,基于标签传播的涉烟车辆异常检测算法能有效的检测出大部分涉烟车辆。在给定数据集中,算法检测出异常点的召回率为 57.7%,远超其他传统机器学习模型。该算法可为运输违禁物品车辆的侦查提供辅助支持。

关键词: 涉烟车辆;异常检测;半监督学习;随机森林;标签传播
中图分类号: TP391 TH701 **文献标识码**: A **国家标准学科分类代码**: 520.60

Anomaly detection of cigarette-smuggling vehicles based on label propagation

Wang Zhen[1], You Ziquan[2], Zhang Jincheng[2], Gan Xiaoying[2], Tao Chunhe[1]
(*1. Research Laboratory of Economic Detection Technology, The Third Research Institute of Ministry of Public Security, Shanghai 200030, China; 2. College of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

Abstract: The tobacco industry is closely related to government revenue. Smuggling of counterfeit cigarettes will not only cause the loss of national tax, but also disrupt the market and endanger consumers' health. How to effectively regulate cigarette-smuggling vehicles is of great significance to the development of the tobacco industry. Aiming at the issue of cigarette-smuggling vehicles, this paper combines the actual collected vehicle data and proposes an anomaly detection algorithm based on label propagation. Firstly, the features of the vehicle data set are extracted. Second, random forest algorithm is adopted to conduct the feature selection. On this basis, label propagation algorithm is utilized to classify the anomaly vehicles. The results show that in the case of less historical data and abnormal vehicle tags, the anomaly detection algorithm of cigarette-smuggling vehicles based on label propagation can effectively detect most cigarette-smuggling vehicles. In the given dataset, the recall rate of the proposed algorithm in detecting anomaly is 57.7%, which outperforms those of other traditional machine learning models. The algorithm can provide auxiliary support for the detection of the vehicles transporting prohibited items.

Keywords: cigarette-smuggling vehicle; anomaly detection; semi-supervise learning; random forest; label propagation

0 引 言

2019 年 1 月至 11 月,烟草行业共查处案值 5 万元以上假烟走私烟案件 9 590 起,同比增长 14.86%;并查获假烟 35.11 万件、走私烟 12.46 万件。一方面,走私假烟有较高的利润,让犯罪分子觉得有利可图,另一方面,执法部门也在不断加大对假烟的打击力度。

而目前执法部门对走私假烟的侦查方式主要依照办案经验,在各主要交通要塞高速卡口采用人工侦别的方式,从车辆特征判断通过车辆是否为涉烟违法。这种方式依赖于大量的人力,且需要稽查人员有较强的个人业

收稿日期:2020-08-24 Received Date:2020-08-24
*基金项目:国家自然科学基金(61672342)项目资助

务素质。若能利用大数据分析中的异常检测技术,从采集出的车辆数据中快速锁定可能的异常涉烟车辆,能大幅提升侦查效率。

异常检测(anomaly detection)是指根据节点特征,在大量数据节点中发现异常节点,避免系统性能的恶化。而这些异常节点的行为特征通常能给我们提供非常重要、关键的信息[1]。如计算机网络中的异常流量可能意味着被黑客攻击的计算机正在向未经授权的目的地发送敏感数据[2]。来自航天飞机传感器的异常读数可能表明航天飞机的某个组件出现故障[3]。信用卡交易数据中的异常可能表明信用卡被盗刷[4]。因此,异常检测已成为近年来在不同研究和应用领域的一个重要方向。

主要的异常检测方法可分为,基于分类的异常检测方法,基于最近邻的异常检测方法和基于聚类的异常检测方法。基于分类的异常检测方法中,Stefano 等[5]利用正常数据训练神经网络,学习出正常数据的行为模式,从而区分出正常点与异常点。Scholkopf 等[6]用一类支持向量机(one-class SVM)学习正常点与异常点的判别边界来区分异常点。基于最邻近的异常检测方法中,Ramaswamy 等[7]将其第 k 个最近邻居的距离作为异常分数。Breunig 等[8]采用欧式距离估计周围密度并作为一个异常指标。基于聚类的异常检测方法中,He 等[9]使用聚类来确定数据中的密集区域,然后对每个聚类进行密度估计。Daniel 等[10]使用谱聚类来初步分类,然后对每个类别进行核密度估计,最后用统计方法识别每个类别中的异常点。近年来,出现了不少基于图的异常检测方法,Ding 等[11]提出了 encoder-decoder 结构,利用图卷积网络作为 encoder,再通过 decoder 恢复信息,对比两者的信息,得到每个节点的异常分数。Liang 等[12]提出的基于半监督学习的 SEANO 算法,考虑了异常点的影响,对于每个节点,同时融合自身信息与邻居信息,并通过预测类标签和节点上下文来平滑异常点的影响。

半监督学习结合了监督学习与无监督学习的优点,能取得较高的准确率和具有良好的泛化能力,且不需要依靠大量标签,非常适合小数据集学习任务。小数据集具有数据量小,标签少等特点。基于大量数据与标签的监督学习算法在小数据集上容易导致过拟合现象。同时,基于无监督学习的异常检测算法存在准确率低的缺陷。然而,标签传播(label propagation,LP)算法作为一种基于图的半监督学习算法,具有运算量小,标签依赖程度低的特点。它利用样本间的关系建立完全图模型,令一个节点的标签按其与其它节点的相似度进行传递,从而完成对样本的分类。不需要大量标签就能利用未标记数据的内在结构,分布规律和临近数据的标记,即可预测和传播未标记数据的标签,然后合并到标记的数据集中。Vercruyssen 等[13]基于标签传播的方法,对商户的用水量异常情况进行检测。Bhatia 等[14]则通过标签传播对社交网络中的节点进行聚类,并将没有分配标签的节点标记为异常来实现异常点检测。邓凯旋等[15]提出了一种新的改进标签传播算法,利用节点重要性分析标签传播算法中的标签传播能力,并制定新的标签更新策略。王诗玉等[16]将标签传播应用于社交网络重叠社区发现中,并利用网络中的极大团作为标签初始化的基本单位,有效的减少了标签冗余并提高了稳定性。

因此,本文采用半监督学习的标签传播算法,利用已掌握的少量标签样本,识别出有相似异常行为的异常点。本文将该算法应用在某地区一次交通稽查车辆数据中,通过特征工程与标签传播算法,成功挖掘出异常涉烟车辆。经专业人员的核查反馈,验证了本文方法的有效性。

文章章节安排如下。第 1 节详细介绍系统模型,说明了如何进行特征提取和特征选择,并通过标签传播算法利用已有标签节点来对无标签节点进行预测。在第 2 节,利用真实数据集验证了算法的有效性。最后在第 3 节进行了总结与未来展望。

1 理论分析

由于实际交通稽查数据具有数据量小,标签少等小数据集的特点,本文提出了一种基于随机森林和标签传播算法的异常检测算法。如图 1 所示,整个系统分为 3 个部分:特征提取,特征选择和标签传播。首先从输入数据中提取出特征,通过随机森林对特征进行筛选,最后再用标签传播算法利用已知标签节点信息找出异常节点。本节会分别对这几个模块进行论述。

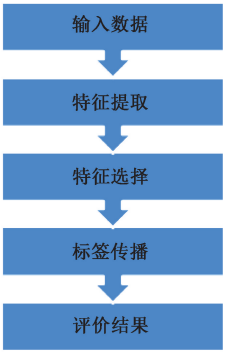


图 1 系统流程图

Fig. 1 Flow chart of the system

1.1 特征提取

假设在一个观察区域边缘有两个观测点,当车辆驶入和驶出该区域时,两个观测点会记录下相关的车辆信息,形成一个 $N \times M'$ 维的小数据集,其中 N 表示车辆数,M' 表示原始数据库特征维数。根据车辆驶入的方向可

以将这两个观测点划分为入观测点和出观测点,其中入观测点记录了车辆从其它地区驶入该区域的信息,出观测点记录了车辆从该区域驶离时所记录的信息。数据集数据中包括车辆出入收费站时的时间,重量,车辆轴数(即车轮对数)等多维特征。下文用节点 $\boldsymbol{x}_i(1 \leqslant i \leqslant N)$ 表示车辆 i。

首先从时间维度上对数据集进行特征提取,需要提取关于每个 $\boldsymbol{x}_i$ 的时间特征向量。用 $f_{ij}(1 \leqslant i \leqslant N, 1 \leqslant j \leqslant 24)$ 表示节点 $\boldsymbol{x}_i$ 观察期间内,在每一天第 j 小时驶入收费站的频率。$t_{ij}(1 \leqslant i \leqslant N, 1 \leqslant j \leqslant 24)$ 表示节点 $\boldsymbol{x}_i$ 观察期间内,在每一天第 j 小时驶入收费站的总次数。则 $f_{ij} = \dfrac{t_{ij}}{\sum_{r=1}^{24} t_{ir}}$,由此可得到关于每个 $\boldsymbol{x}_i$ 的24维时间特征向量$[f_{i1}, f_{i2}, \cdots, f_{i24}]$。

其次,从自身属性的维度上提取特征。令 $l_i(1 \leqslant i \leqslant N)$ 表示节点 $\boldsymbol{x}_i$ 的车辆轴数,$w_{i\_in}(1 \leqslant i \leqslant N)$ 表示节点 $\boldsymbol{x}_i$ 驶入观测点时的重量,$w_{i\_out}(1 \leqslant i \leqslant N)$ 表示节点 $\boldsymbol{x}_i$ 驶出观测点时的重量。$w_{i\_al\_in} = w_{i\_in}/l_i$ 表示 x_i 驶入收费站时的单轴重,$w_{i\_al\_out} = w_{i\_out}/l_i$ 表示驶出收费站时的单轴重。本文同时将 $\boldsymbol{x}_i$ 驶入驶出时单轴重的变化情况作为另一维特征。令 $w_i = w_{i\_al\_in} - w_{i\_al\_out}$ 表示节点 x_i 驶入驶出时单轴重的变化情况。

提取的特征如表 1 所示。

表 1　特征变量说明

Table 1　Explanation of the feature variables

特征	说明
$f_{ij}(1 \leqslant i \leqslant N, 1 \leqslant j \leqslant 24)$	$\boldsymbol{x}_i$ 在观察期间内,在每一天第 j 小时驶入收费站的频率
$w_{i\_al\_in}$	$\boldsymbol{x}_i$ 驶入的单轴重
$w_{i\_al\_out}$	$\boldsymbol{x}_i$ 驶出的单轴重
w_i	$\boldsymbol{x}_i$ 驶入与驶出的单轴重差

由此,得到一个 M 维的特征矩阵 $\boldsymbol{X}' \in \boldsymbol{R}^M$,其中 $M \in Z^+$。

1.2　基于随机森林的特征选择

针对研究对象 $\boldsymbol{X}'$,特征选择指从已有的 $M(M \in Z^+)$ 个特征(Feature)中选择 $d(d \in Z^+ \text{ 且 } d < M)$ 个特征使得系统的特定指标最优化。特征选择之后,有效降低了数据集的复杂度与学习任务的难度,减少学习算法的运行时间,提升模型效率,增加模型可解释性。本文采用随机森林[17]中基于基尼指数的特征重要性评估算法[18-19],对所有特征的重要程度从高到低进行排序,选择其中排名最高的 $d(d \in Z^+ \text{ 且 } d < M)$ 种特征。

本文使用 CART(classification and regression tree)决策树[16]构建的随机森林实现特征选择。用 *VIM*(variable importance measures)表示变量重要性,用已经训练好的随机森林模型对特征重要性评分。若结点 m 选择特征 $F_s(1 \leqslant s \leqslant M)$ 为划分属性,则 F_s 在决策树结点 m 上的重要性,即树结点 m 分枝前后的 Gini 值变化量,如式(1)所示,其中,Gini(l) 和 Gini(r) 分别表示分枝后两个新的树结点的 Gini 指数。

$$VIM_{sm} = \text{Gini}(m) - \text{Gini}(l) - \text{Gini}(r) \tag{1}$$

如果在决策树 q 中,选择特征 F_s 为划分属性的结点集合为 H,那么 F_s 在第 q 棵树的重要性为:

$$VIM_{qs} = \sum_{m \in H} VIM_{sm} \tag{2}$$

随机森林中共有 n 棵树,那么

$$VIM_s = \sum_{i=1}^{n} VIM_{is} \tag{3}$$

最后,将所有求得的特征重要性评分做归一化:

$$VIM_s = \frac{VIM_s}{\sum_{i=1}^{c} VIM_i} \tag{4}$$

从而得到所有特征的重要性评分,将其排序并选择其中排名最高的 $d(d \in Z^+ \text{ 且 } d < M)$ 种特征。至此,可以得到一个 d 维的特征矩阵 $\boldsymbol{X} \in \boldsymbol{R}^d$。

1.3　标签传播算法

给定 N 个数据节点的数据集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \cdots, \boldsymbol{x}_N\} \in \boldsymbol{R}^d$,标签集合 $\boldsymbol{\mathcal{L}} = \{1, 2, 3, \cdots, C\}$,其中标签类别数 C 已知。定义前 l 个点的集合 $X_L = \{\boldsymbol{x}_i\}_{i=1}^{l}$ 为有标签数据集,其标签集合 Y_L 已知。剩下的点组成的集合 $X_U = \{x_i\}_{l+1}^{N}$ 为无标签数据集,其标签集合 Y_U 未知。用 u 表示 X_U 中数据节点的个数,即 u = $|X_U|$。则 $l + u = N$,通常 $l \ll u$。算法的目的是根据已知的 X 和 Y_L 来对 Y_U 进行预测,从而得到数据中每个样本的标签信息。

首先根据 $\boldsymbol{X}$,将所有数据作为节点构造出完全连通图,定义 $d_{ii'}$ 为节点 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_{i'}$ 之间的距离,可根据 $\boldsymbol{x}$ 的不同(如连续值还是离散值)而采用欧式距离、马氏距离或是其他距离,本文采用欧氏距离。定义 $w_{ii'}$ 为节点 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_{i'}$ 之间的边的权重,两点间距离越近,边的权重越大。同时权重 $w_{ii'}$ 的值受超参数 σ 影响。节点与节点之间的边的权重由式(5)所示。

$$w_{ii'} = \exp\left(-\frac{d_{ii'}^2}{\sigma^2}\right) \tag{5}$$

定义一个概率传递矩阵 $T \in R^{(l+u)\times(l+u)}$,让 $T_{ii'}$ 表示标签信息从节点 $\boldsymbol{x}_{i'}$ 传播到 $\boldsymbol{x}_i$ 的概率:

$$T_{ii'} = P(x_{i'} \to x_i) = \frac{w_{ii'}}{\sum_{k=0}^{l+u} w_{ki'}} \tag{6}$$

同时定义一个标签矩阵 $\boldsymbol{Y} \in R^{(l+u)\times C}$,其中的每个元

素 $Y_{i,c_i}=\delta(y_i,c_i)$ 表示节点 $\boldsymbol{x}_i$ 被标注为类别 $c_i\in\{1,2,\cdots,C\}$ 的概率。经过一定次数的迭代以后,标签矩阵 $\boldsymbol{Y}$ 收敛,从而能够得到每个样本所属的类别,得到分类的结果。

标签传播算法过程如下:

1)所有节点传播标签一步:

$$\boldsymbol{Y}\leftarrow\alpha T\boldsymbol{Y}+(1-\alpha)\boldsymbol{Y}_{original} \tag{7}$$

2)夹逼标注数据,重复步骤 1)直到 $\boldsymbol{Y}$ 收敛。

$\boldsymbol{Y}_{original}$ 为初始的标签矩阵。$\alpha\in[0,1]$,为夹紧系数,代表一个点对它的邻居的标签的接受程度。当 $\alpha=0$ 时,表示保持原始标签,不接受邻居的标签;$\alpha=1$ 时表示从邻居处接受标签,忽略原始标签。步骤 2)可以使得节点标签的类别分布集中在给定的类别。

1.4 检测结果评价准则

异常检测的结果可以用混淆矩阵来表示,本文将异常数据类别定义为正类,正常数据类别定义为负类。混淆矩阵的定义如表 2 所示。为了对比不同算法的分类预测结果,本文用召回率(Recall)作为评价指标,具体形式如下:

$$Recall=\frac{TP}{TP+FN} \tag{8}$$

表 2 混淆矩阵

Table 2 Confusion matrix

实际类别	预测为正类	预测为负类
正类	正确正类(true pasitire,TP)	错误正类(false positive,FP)
负类	错误负类(false negative,FN)	正确负类(true negative,TN)

2 实 验

2.1 数据集及实验设置

1)数据集:本次实验利用某地区一次稽查非法运载违禁物品的交通稽查行动的真实数据进行实验。该数据集包括了 2017 年 10 月 1 日~12 月 31 日内所有车辆往返某区域的两个收费站 A,B 时的时间以及收费站所记录的车辆的型号、重量、轴数等信息。其中 A 收费站包含 106 422 个车牌号的 229 976 车辆信息,B 收费站包含 81 320 个车牌号的 1 842 965 条车辆信息。最终整理出 48 078 条在两收费站均有记录的车辆信息与 63 条标签信息,其中正常样本 29 个,异常样本 34 个,同时具有数据量小,标签少等小数据集的特点。

2)对比算法:(1)孤立森林[20]。孤立森林算法是一种基于树的无监督算法,通过随机划分特征与特征值孤立异常数据点,实现异常检测。(2)局部异常因子[21](local outlier factor)。局部异常因子是一种基于密度的无监督算法。通过比较每个点与其领域点的密度来判断该点是否为异常点,若该点的密度越低,则越有可能被认定为异常点。(3)one-class SVM[22]。One-class SVM 是一种经典的无监督算法。通过学习一个超平面将零点与所有数据点在特征空间中分开,从而获取特征空间中数据的概率密度区域。当测试点落在训练数据点区域时,被认定为正常点,落在其他区域为异常点。

3)实验环境:本文实验在 Windows10 操作系统上运行,采用 python 作为算法实现语言,Anaconda4.2 作为集成开发环境。对比算法采用 python 语言的 sklearen 库中的版本。标签传播算法中的夹紧系数 α. 取 0.5。孤立森林算法中,每棵树包含的样本数为 300,异常点比例为 0.001 7。LOF 算法中,每个点的邻居数为 2 异常比例为 0.001 7。one-class svm 中 gamma 取 0.5,异常比例为 0.001 7。

2.2 数据特征分析

对数据集从时间,单轴重,单轴重变化等维度进行分析与特征提取。经过数据清洗与分析,得到以下发现:

1)时间维度特征上。异常车辆的驶入时间分布与正常车辆的驶入时间分布有明显差别,正常车辆驶入时间主要集中在上午 8 点与下午 5 点左右,而异常车辆驶入时间主要集中在夜间。而在车辆驶出时间上,异常车辆与正常车辆的驶出时间相似。这是因为异常车辆在满载违禁物品驶入该区域时,为了逃避检查会选择稽查力度宽松的夜间出行,而驶出时由于车上没有载有违禁物品,因此在驶出时间的选择上与正常车辆的选择相似。

2)单轴重维度特征上。异常车辆会满载违禁物品驶入该区域,使得异常车辆的驶入单轴重主要处于区间 800~1 200 kg 中,与正常车辆的驶入单轴重会有明显区别。同时,由于异常车辆会空载驶出该区域,因此异常车辆驶出时的单轴重也较为集中于某个区间,同样与正常车辆有明显差别。

3)单轴重变化维度特征上。异常车辆在驶入驶出时单轴重的变化主要在 200~400 kg 间,与正常车辆的单轴重变化也有明显差别。

从上面的 3 个分析中可以看出,本文的特征提取是成功有效的。经过提取后的特征向量能明显区分正常车辆与异常车辆,保证了之后的算法识别结果的有效性。

用少量样本训练随机森林。基于上述随机森林算法对特征重要性进行排序,结果如图 2 所示,这里对数据特征进行了脱敏处理,特征名称用“V”+阿拉伯数字代替。最终根据经验值挑出其中重要性程度最高的 9 个特征。其中脱敏后特征与原特征映射关系如表 3 所示。

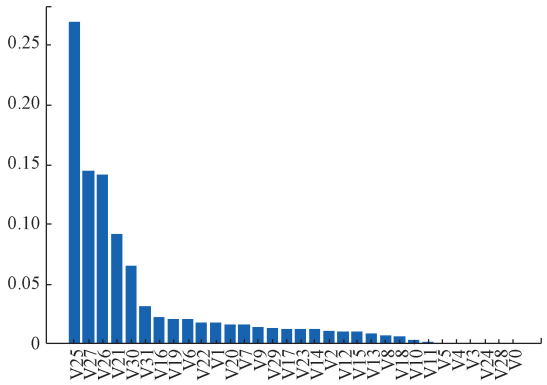


图 2　特征重要程度排序

Fig. 2　Feature importance ranking

表 3　特征映射关系

Table 3　The feature mapping relationship

脱敏后特征	原特征
V1-V24	时间特征向量
V26-V31	驶入驶出单轴重区间向量
V0, V25	轴数与单轴重变化特征

2.3　实验结果

实验结果如表 4 所示。已有异常样本 34 个,正常样本 29 个,用 8 个异常样本,14 个正常样本作为初始标签去预测其他未标记节点。对本次任务来说,异常节点才是检测目标,因此只关注异常类别的实验结果。实验重复了 10 次,每次推荐出异常分数排名前 80 的异常节点,其中带有标记的异常节点平均值为 23,减去作为种子节点的 8 个异常样本,每次检测出异常节点的平均值是 15 个,$\text{Recall}=\frac{23-8}{34-8}=57.7\%$。

表 4　实验结果混淆矩阵

Table 4　The experiment result confusion matrix

		预测	
		0	1
实际	0	46 987	57
	1	11	23

同时,为探究标签传播中超参数 α 值对结果的影响,我们设置不同的 α 值进行实验。结果如表 5 所示。

这是因为 α 值代表一个点对它的邻居的标签的接受程度,α 越小,受原始无标签邻居影响越小,受原始有标签邻居影响越大。从而当一个点周围有很多被标为异常的初始无标签邻居时,只要有一个被标为正常的初始有标签邻居,该点也会被标为正常,从而降低了召回率。

表 5　变量 α 值对召回率影响

Table 5　The influence of variable α on the recall rate

α	召回率/%
0.5	67.6
0.1	67.6
0.01	64.7
0.001	55.9

此外,考虑到交通数据标签的稀有性,还尝试了在完全无监督学习下的其它算法的性能,以期望能在无标签的状态下也能取得一定的结果。3 种对比算法均拟推荐出前 80 个异常点。结果如表 6 所示。

表 6　不同算法结果对比

Table 6　Comparison results of different algorithms

算法	检测出异常点数目	召回率/%
孤立森林	10	35.3
局部异常因子(LOF)	5	14.7
One-class svm	3	8.8
本文提出算法	15	57.7

从结果可以看出,采用无监督学习的孤立森林的效果在 3 种对比算法中效果最好,检测出 10 个异常点,召回率 35.3%,但远不及本文所提出的算法模型。但值得注意的是,这是在无标签的情况下取得的结果,仍然能识别出 10 个异常节点。在无标签的情况下,检测无法达到较高的召回率与准确率,希望通过进行异常检测来缩小稽查范围,提升稽查效率。因此,在无标签的情况下,可以考虑采用孤立森林来进行异常检测。

由于部分异常节点的特征向量与正常节点的类似,因而未被识别出。但从最终检测结果来看,该算法能够根据已掌握的少量信息,从海量数据中快速、准确的发现异常节点,缩小识别范围,证明了算法在小数据集异常检测上的有效性。

3　结　　论

本文提出了一种基于标签传播的涉烟车辆异常检测算法,设计了一套有效的特征提取和特征选择的步骤,并新颖的将标签传播算法应用到异常检测当中。该算法能够有效的在采集的车辆信息中锁定可能的异常涉烟车辆。本文的实验结果最终证明了我们算法的有效性,且性能远超当前先进的无监督算法中的孤立森林算法。虽然由于部分异常节点与正常节点类似而未能识别出来,

但在召回率等指标上是令人满意的。而未来工作则是在此算法模型上继续完善,希望能识别出与正常节点类似的异常节点,进一步提高召回率等性能指标,同时,我们会考虑将算法的使用场景进行拓展,把算法应用到其他的走私稽查中。

参考文献

[1] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: A survey[J]. ACM Computing Surveys, 2009, 41(3), DOI: 10.1145/1541880.1541882.

[2] KUMAR V. Parallel and distributed computing for cybersecurity[J]. IEEE Distributed Systems Online, 2005, 6(10), DOI: 10.1109/mdso.2005.53.

[3] FUJIMAKI R, YAIRI T, MACHIDA K. An approach to spacecraft anomaly detection problem using kernel feature space[C]. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005:21-24.

[4] ALESKEROV E, FREISLEBEN B, RAO B. CARDWATCH: A neural network based database mining system for credit card fraud detection[M]. Cardwatch: A Neural Network Based Database Mining System for Credit Card Fraud Detection, 1997.

[5] STEFANO C D, SANSONE C, VENTO M. To reject or not to reject: That is the question-an answer in case of neural classifiers[J]. IEEE Press, 2000, DOI:10.1109/5326.827457.

[6] SCHöLKOPF, BERNHARD, PLATT, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7):1443-1471.

[7] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]. Proceedings of the 2000 ACMSIGMOD International Conference on Management of date, 2000:427-438.

[8] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[C]. Proceedings of the 2000 ACMSIGMOD International Conference on Management of Date, 2000:93-104.

[9] HE Z, XU X, DENG S. Discovering cluster-based local outliers[J]. Pattern Recognition Letters, 2003, 24(9-10):1641-1650.

[10] DANIEL D, BORIS P, ANDRES M, et al. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach[C]. ACM SIGKDD, 2018.

[11] DING K, LI J, BHANUSHALI R, et al. Deep anomaly detection on attributed networks[C]. Proceedings of the 2019 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2019: 594-602.

[12] LIANG J, JACOBS P, SUN J, et al. Semi-supervised embedding in attributed networks with outliers[C]. Proceedings of the 2018 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2018: 153-161.

[13] VERCRUYSSEN V, MEERT W, VERBRUGGEN G, et al. Semi-supervised anomaly detection with an application to water analytics[C]. 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018.

[14] BHATIA V, SANEJA B, RANI R. INGC: Graph clustering & outlier detection algorithm using label propagation[C]. International Conference on Machine Learning & Data Science, IEEE Computer Society, 2017.

[15] 邓凯旋, 陈鸿昶, 黄瑞阳. 基于标签传播能力的改进 LPA 算法[J]. 计算机工程, 2018, 44(3):60-64.
DENG K X, CHEN H CH, HUANG R Y. Improved LPA algorithm based on tag propagation ability[J]. Computer Engineering, 2018, 44(3):60-64.

[16] 王诗玉, 许国艳, 石水倩. 基于极大团的标签传播重叠社区发现算法[J]. 电子测量技术, 2018, 41(2): 45-49.
WANG SH Y, XU G Y, SHI SH Q. A tag propagation overlapping community discovery algorithm based on extremely large clusters[J]. Electronic Measurement Technology, 2018, 41(2):45-49.

[17] CUTLER A, CUTLER D R, STEVENS J R. Random forests[J]. Machine Learning, 2004, 45(1):157-176.

[18] 杨凯,侯艳,李康. 随机森林变量重要性评分及其研究进展[EB/OL]. [2015-07-23]. http://www.paper.edu.c/releasepaper/content/201507.
YANG K, HOU Y, LI K. Random forest variable importance score and its research progress[EB/OL]. [2015-07-23]. http://www.paper.edu.c/releasepaper/content/201507.

[19] GOLDSTEIN B A, POLLEY E C, BRIGGS FB. Random

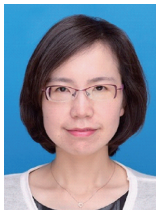
forests for genetic association studies [J]. Stat Appl Genet Mol Biol, 2011,10(1):32.

[20] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]. Data Mining, 2008.

[21] BREUNIG MM, KRIEGEL HP, NG RT, et al. LOF: Identifying density-based local outliers[C]. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA: ACM Press, 2000: 93-104.

[22] SCHLKOPF B, WILLIAMSON R C, SMOLA A J, et al. Support vector method for novelty detection [C]. Advances in Neural Information Processing Systems 12, NIPS Conference, 1999.

作者简介



王贞,2003 年于兰州理工大学获得学士学位,2006 年于兰州理工大学获得硕士学位,现为公安部第三研究所助理研究员,主要研究方向为机器学习、数据分析。

E-mail: 13671846997@ ecid. ga

Wang Zhen received her bachelor degree in 2003 and master degree in 2006 both from Lanzhou University of Technology. She is an assistant researcher in The Third Research Institute of Ministry of Public Security now. Her main research direction includes machine learning and data analysis.

甘小莺(通信作者),2000 年于上海交通大学获学士学位,2005 年于上海交通大学获得工学博士学位。现为上海交通大学电子工程系副教授,主要研究方向为无线网络、网络大数据、网络经济学。

E-mail: ganxiaoying@ sjtu. edu. cn

Gan Xiaoying (Corresponding author) received her bachelor and doctor degrees both from Shanghai Jiao Tong University in 2000 and 2005, respectively. She is currently an associate professor in Department of Electronic Engineering, Shanghai Jiao Tong University. Her main research direction includes wireless networks, network big data and network economics.