经济管理

乔浪超，王进录，高宝红，等. 基于时空数据特征的寄递涉烟犯罪分析方法[J]. 中国烟草学报，2023, 29(1). QIAO Langchao, WANG Jinlu, GAO Baohong, et al. Express-related counterfeit cigarette criminality analysis based on spatio-temporal data features[J]. Acta Tabacaria Sinica, 2023, 29(1). doi:10.16472/j.chinatobacco.2021.166

基于时空数据特征的寄递涉烟犯罪分析方法

乔浪超[1]，王进录[1]，高宝红[1]，杨新刚[1]，
冯文涛[1]，许荣垚[2]，卫毅然[2]，刘伟[2*]

1 中国烟草总公司陕西省公司，专卖监督管理处，陕西西安雁南四路 19 号 710061；
2 西安邮电大学，计算机学院，陕西西安西长安街 618 号 710121

摘 要：【目的】使用大数据和人工智能技术研究基于寄递大数据的“互联网+寄递”新型涉烟犯罪分析方法。【方法】使用中文分词技术对寄递大数据进行预处理。提出了“寄递时空模式”新概念并计算其时域和频域统计量作为时空特征。使用特征选择和降维方法计算时空特征集合中的优选特征，并比较不同分类器算法结合优选特征构建的涉烟犯罪分析模型的性能。【结果】（1）提出的时空特征具有区分涉烟和不涉烟寄递数据的能力。随机森林和 GBDT 分类器整体性能最好，在准确率、阳性和阴性预测值等指标上均达到 0.94 以上。（2）基于优选特征建立的分析模型可以取得和初始特征模型接近的预测结果，优选特征数据储存量仅为原始特征数据的 40%。（3）CFS 特征选择方法选出的优选特征对涉烟预测模型结果的可解释性提供了依据。（4）初步实验表明本文方法可满足寄递涉烟分析的实时性要求。【结论】基于“寄递时空模式”计算的时空特征结合分类器可区分涉烟和不涉烟寄递数据。

关键词：寄递涉烟犯罪；寄递时空模式；时间序列分析；特征选择和降维；机器学习

近年随着我国寄递行业和电子商务产业的飞速发展，以移动社交平台为依托，寄递假烟、走私烟或单人单次多于 2 条真烟（400 支）等违反我国现行烟草专卖管理条例的涉烟包裹，并借助电子结算平台进行实时结算的“互联网+寄递”新型涉烟犯罪行为已成为损害烟草行业发展的重要犯罪方式。在寄递环节查获的假、私卷烟数量逐年上升，对我国卷烟销售乃至整个卷烟生态都带来了巨大冲击[1-2]。

寄递涉烟犯罪具有实施速度快、覆盖范围广、隐蔽性强等特点[3]，现有专卖监管方式无法对其进行有效打击。烟草专卖管理部门迫切需要新方法和新工具来打击寄递涉烟犯罪。近年来大数据和人工智能技术在社交媒体[4]、金融[5]、医疗健康[6]、智能交通[7]、多媒体[8-9]和公共卫生[10-11]等多个应用领域内取得了成功应用。寄递涉烟数据是典型的大数据，可以使用大数据和人工智能技术分析寄递涉烟犯罪行为数据特性、挖掘案件数据价值继而精准打击寄递涉烟犯罪。

根据文献调研结果，使用大数据和人工智能技术分析寄递涉烟犯罪的工作还未见报道。已有相关研究可分为三个方面：一是对寄递网络进行研究、对寄递大数据进行统计分析或结合数据挖掘技术解决寄递过程中的优化问题；二是利用寄递大数据进行异常行为检测；三是公安和烟草专卖部门人员从管理角度探讨了互联网环境下违禁品寄递犯罪问题。

谭旭[12]基于寄递数据分析了快递网络的拓扑特性和交通流特性并发现了寄递网络的三级拓扑结构划分和节点度分布的幂律特性；任思源等[13]基于寄递大数据提出一种城市画像系统。作者使用寄递频次、时间、地址、物品 4 个指标分别对城市中不同社会群体与城市区域的寄递数据进行分析与画像；刘二超[14]研究了利用寄递数据和道路信息的快递服务点选址问题。Li 等[15]介绍了基于大数据技术的京东智能寄递系统的架

作者简介：乔浪超（1984—），硕士，烟草专卖信息化建设，Tel：029-85466535，Email：176214379@qq.com
通讯作者：刘伟（1975—），Tel：029-88166055，Email：liuwei@xupt.edu.cn
收稿日期：2021-08-26；**网络出版日期：**2022-04-20

构与应用。

郝晟[16]将寄递数据与侦查业务需求结合，构建数据仓库并采用聚类算法分析犯罪嫌疑人隐藏寄递特征。李万彪等[17]提出一种基于关系数据模型的犯罪网络发现方法用于挖掘犯罪团伙关系。文杰锋[18]利用配送距离和路径不同研究了快递包裹配送中的异常检测问题。

郭小伟[19]总结了我国寄递违禁品犯罪特点与现状，并研究了防控对策。王宁[20]研究了寄递贩毒犯罪特点及侦查对策。国内公安和烟草专卖管理部门针对互联网环境下的涉烟犯罪问题也进行了探讨[1-3,21-23]。这些工作分析了互联网涉烟犯罪特点，从管理层面研究了打击涉烟犯罪、提升烟草专卖管理水平的思路和方法。

综上，尽管已有研究取得了一定成果，但仍存在不足：（1）“互联网+寄递”新型涉烟犯罪借助寄递网络和互联网等信息技术，必须有针对性地使用最新计算机技术才能有效打击。研究寄递涉烟犯罪的公安和烟草专卖部门人员并未从计算机技术层面对这一问题进行研究[1-3,21-23]。（2）尽管有一些工作涉及到寄递大数据犯罪分析[16-17]，但寄递涉烟大数据有其自身特性（寄递涉烟犯罪面向个人零售终端和消费者，具有条数少、频次多、覆盖广、总量大等特点），和其他寄递违禁品（如毒品等）不同。必须针对寄递涉烟数据自身特点进行研究。

本文使用大数据和人工智能技术研究并提出一种寄递涉烟犯罪分析方法。该方法可以发现寄递环节中违反我国现行烟草专卖管理条例的涉烟包裹，从而辅助烟草专卖管理部门精准打击寄递涉烟犯罪。该方法首先使用中文分词技术对寄递大数据进行预处理。其次通过时间轴校正统计了发、收件地址在历史时间段内涉烟和不涉烟包裹频度数据。将该频度数据视为一维时间序列，计算其时域和频域统计量作为时空特征。之后通过特征选择和降维算法计算优选特征。最后使用分类器算法结合优选特征建立寄递涉烟犯罪分析模型以预测未知寄递包裹的涉烟可能性。初步实验结果证实了本文方法的可行性和有效性。

1　方法

1.1　框架

图 1 给出了本文方法框架图。该方法分为两个阶段：离线训练阶段和在线测试阶段。前者使用已知涉烟和不涉烟寄递数据训练分类器模型。后者使用训练好的分类器模型对未知寄递数据的涉烟可能性进行判断。由图 1 可知，这两个阶段均包含了数据预处理和时空特征计算环节。前者使用中文分词技术对原始寄递数据进行整合及结构化处理。后者依据本文 1.3 节方法计算时空特征。为了计算时空特征中的优选特征，在离线训练阶段还使用了特征选择和降维方法。图 1 建立的寄递涉烟犯罪分析模型是一个动态闭环迭代模型。模型预测结果经烟草专卖人员现场验证后可加入训练数据集合再次训练模型，从而依据寄递涉烟数据的变化不断完善模型，提升预测精度。

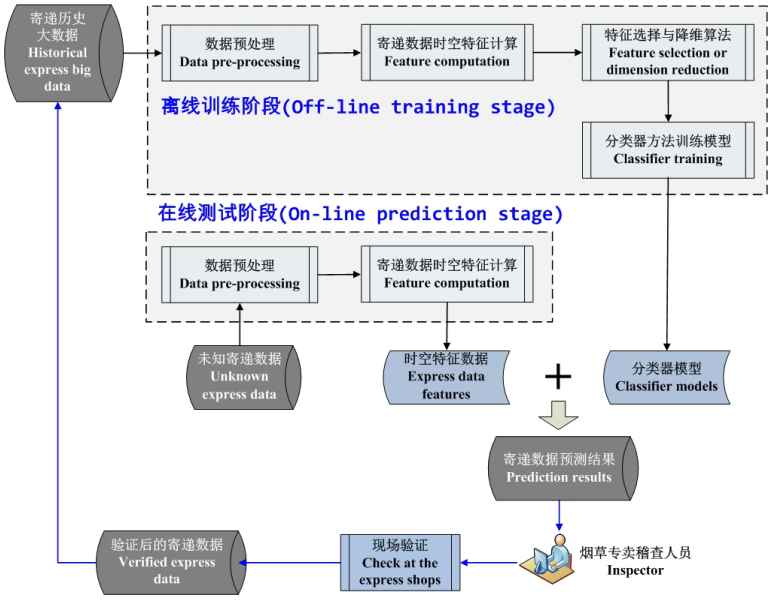


图 1　本文方法框架图

Fig.1 Flowchart of the proposed method

1.2 数据预处理

本文实验中获取的原始寄递数据存在数据缺失，且大量数据以自然语言形式呈现，如“2020-09-12 16:07:08:【广东省汕头市金平区金平东公司】已收件取件人：林*（189*****486）”（隐私数据用*号表示）。为了得到寄递数据中的有效信息（如收发地址等），必须对这些数据进行预处理。本文使用中文分词技术[24-25]对初始寄递数据进行自动分析以得到收发件人、收发件地址等关键数据。具体过程如下：

Step 1：使用基于词典的中文结巴分词[25]方法对数据进行初步中文分词。结巴分词自带的系统词库并非针对寄递数据，无法准确切分寄递地址数据。如“广东省”被切分为“广东”和“省”。

Step 2：为了处理 ***Step 1*** 中错误的分词结果，本文使用自定义词库，即将全国所有省、市、县的名称构建为词库（如辽宁省、西安市、雁塔区）。使用自定义词库对 ***Step 1*** 中的结果进行整合，进一步精确分词，以提高切分地址数据的正确率。

Step 3：以自然语言形式呈现的原始寄递数据中会包含诸如“：【 】”等符号，这些符号会保留在 ***Step 1*** 和 ***Step 2*** 处理结果中。为了过滤掉这些无用信息，本文还构建了包含这些符号的停用词库，使最终分词结果只保留快递地址、姓名、联系电话等有效数据。

Step 4：将分析结果结构化，并进行数据缺失判断和自动填充（如根据 ***Step 2*** 中的自定义词库填充“省-市-县（区）”等地理位置信息）。将结果存储到数据库中供后续使用。

1.3 寄递数据时空特征计算

寄递是快递公司收取寄件人快件，在规定时间内运送至指定收件地点并由收件人签收的活动。寄递大数据具有“时空特性”，即涉烟寄递数据和地理位置相关且带有时间标记。图 2 统计了 2014 年至 2020 年全国 311,777 例涉烟和 396,199 例不涉烟寄递记录在发件时间 00-23 时段（以 1 h 为间隔）时的频度图，横轴上半部为涉烟寄递记录频度图，下半部为同一时段不涉烟寄递记录频度图）。由图 2 可知，涉烟寄递和不涉烟寄递具有不同的“时间特性”。

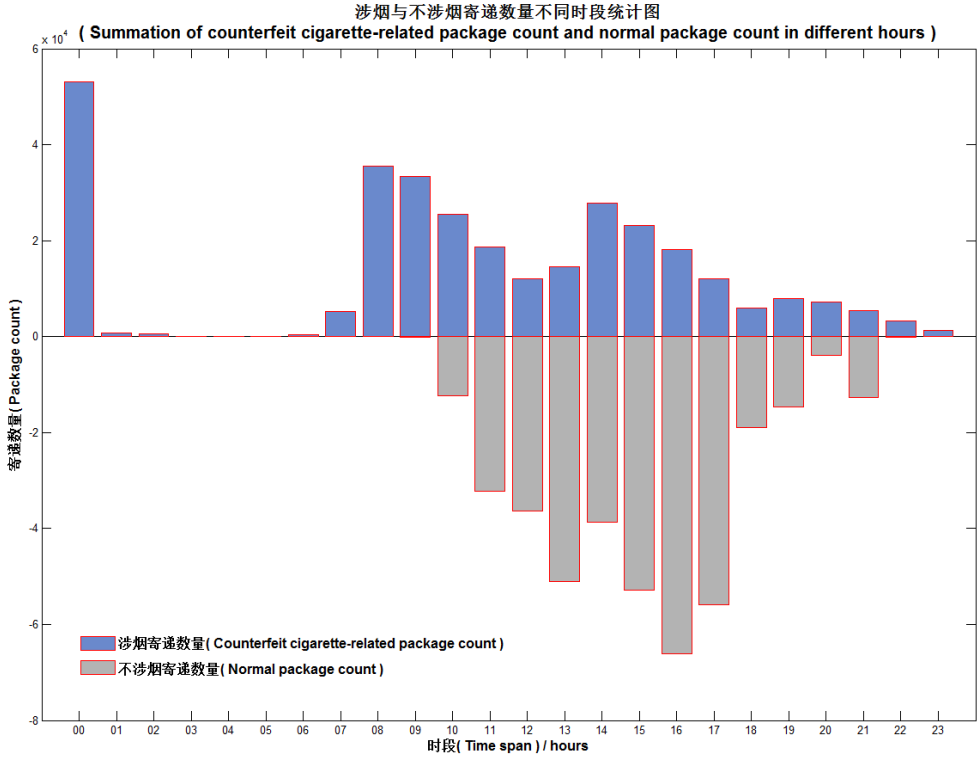


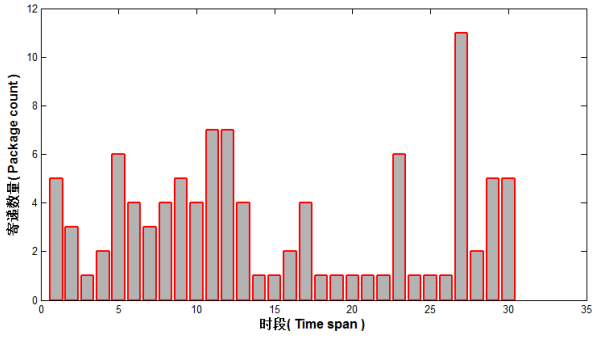
图 2 不同时间段的涉烟与不涉烟包裹寄递统计图

Fig.2 Summation of counterfeit cigarette-related package count and normal package count in different hours

为了正确区分涉烟寄递和不涉烟寄递，必须计算可以描述涉烟寄递数据和不涉烟寄递数据特性的数字量化值，即特征。寄递行为包括发件和收件，分别包含发件地址、发件人姓名、发件人电话号码、发件时

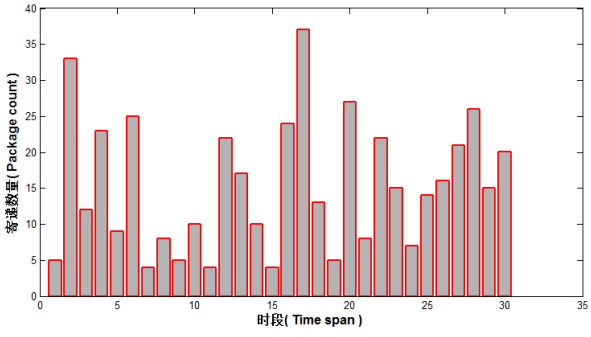
间、收件地址、收件人姓名、收件人电话号码、收件时间等寄递属性。在建立涉烟分析模型时这些属性并非都能作为“特征”使用：（1）由于保护个人隐私，许多寄递记录的姓名和电话号码无法获取；（2）涉烟犯罪嫌疑人为逃避打击，往往在发件时使用假姓名和假电话号码；（3）必须在收件人收件之前预测寄递数据是否涉烟，因此收件时间无法获取。

发、收件地址涉烟包裹的历史寄递模式可用于描述涉烟犯罪。分析寄递大数据可知发、收件地址在给定时段内涉烟和不涉烟寄递记录的寄递频度不同。如果将历史时间段内和发、收件地址相关的包裹寄递频度视为一维时间序列，那么该时间序列可作为发、收件地址涉烟与否的重要依据。该时间序列即为本文提出的“寄递时空模式”新概念，它是包含地址和时间信息的时空数据。图 3(a)-(d)分别表示国内某发、收件地址的“寄递时空模式”。显然，用信号处理中的时间序列分析方法计算“寄递时空模式”所得的时空特征便可用于判断未知寄递数据是否涉烟。



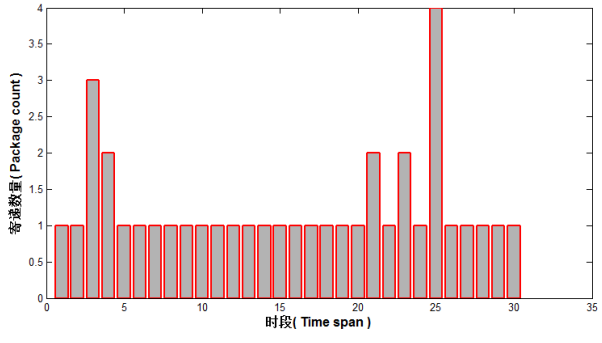
（a）发件地址涉烟包裹寄递频度

(a) Counterfeit cigarette-related package delivery frequency of delivery address



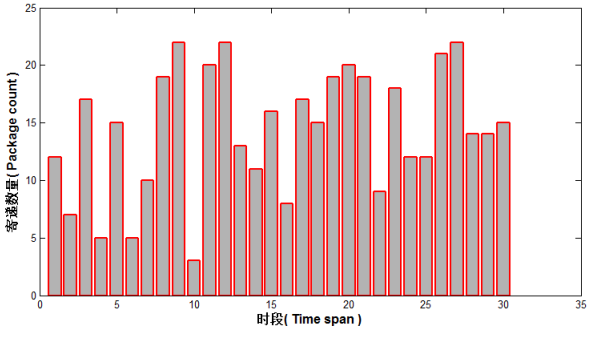
（b）发件地址不涉烟包裹寄递频度

(b) Normal package delivery frequency of delivery address



（c）收件地址涉烟包裹寄递频度

(c) Counterfeit cigarette-related package delivery frequency of receiving address



（d）收件地址不涉烟包裹寄递频度

(d) Normal package delivery frequency of receiving address

图 3　寄递时空模式示例

Fig.3 Illustration of spatio-temporal pattern of delivery and receiving address

发、收地址涉烟寄递行为并非是连续不断的。如果某个地址一年前涉烟，最近不涉烟，那么当前该地址收发涉烟包裹的可能性会降低。本文采用如下方法处理：以涉烟模式为例，假定当前时间为 T_0，某个收、发件地址距离 T_0 最近的涉烟包裹寄递时间为 t，定义权重系数 $\omega=\dfrac{1}{1+(T_0-t)}$，该系数称为时间轴校正系数，即距离 T_0 越近的时间系数越大，越远的时间系数越小。然后以 T_0 为基准，统计距离 T_0 的 N 个时间点上的涉烟包裹频度，可形成包含 N 个数值的时间序列。

假定时间序列为 $x=(x_1,x_2,\ldots,x_n)$，本文使用表 1 所示的 17 维时间序列特征用来描述发、收件地址的寄递模式。其中基本统计特征、时域统计特征、复杂性特征和频域统计特征是广泛使用的时间序列特征算法，寄递频度统计特征是针对寄递行为特点设计的特征算法。

表 1 寄递数据时空特征集合

Tab.1 Spatio-temporal features set of express data

组别（Group）	序号（Index）	说明（Descrption）
基本统计特征	1-3	时间序列的均值、标准差、中位数
寄递频度统计特征	4-11	寄递频度最大值、最小值、最大值第 1 次出现位置的百分比、最大值最后 1 次出现位置的百分比、寄递记录总数、最大值和最小值区间内的寄递记录总数、寄递记录数超过 1 次的频度、寄递记录数大于均值的频度
时域统计特征	12-14	一阶差分绝对和、MAD、香农熵
复杂性特征	15	1D-C0 复杂性
频域统计特征	16-17	傅里叶变换序列的均值、标准差

表 1 中部分特征算法说明如下：

一阶差分绝对和：

$$ASC=\sum_{i=1}^{n-1}|x_{i+1}-x_i| \quad (1)$$

MAD（Mean Absolute Deviation）：其中 μ 为时间序列均值。

$$MAD=\frac{1}{n}\sum_{i=1}^{n-1}|x_i-\mu| \quad (2)$$

香农熵：

$$H=-\sum_{i=1}^{n}p_i\log p_i\text{，其中 }p_i=x_i/\sum_{i=1}^{n}x_i \quad (3)$$

1D-C0 复杂性：1D-C0 复杂性是文献[26-27]提出的度量一维时间序列复杂性程度的方法。其基本思想是将时间序列分解为规则成分和非规则成分两部分，1D-C0 复杂性定义为非规则成分在时间序列中所占的比例。具体算法如下：

令 $\{f(k),k=0,1,\ldots,N-1\}$ 是一个长度为 N 的时间序列，$\{F_N(j),j=0,1,\ldots,N-1\}$ 为其傅立叶变换序列，定义均方值为 $G_N=\frac{1}{N}\sum_{j=0}^{n-1}|F_N(j)|^2$。以该均方值为阈值，令：

$$\widetilde{F}_N(j)=\begin{cases}F_N(j), & if\ |F_N(j)|^2>G_N\\ 0, & if\ |FN(j)|2\le GN\end{cases} \quad (4)$$

然后对 $\left\{\widetilde{F}_N(j),j=0,1,...,N-1\right\}$ 做傅立叶逆变换得到的序列为 $\left\{\widetilde{f}(k),k=0,1,...,N-1\right\}$，定义 1D-C0 复杂度为：

$$C_0=\frac{\sum_{k=0}^{N-1}|f(k)-\widetilde{f}(k)|^2}{\sum_{k=0}^{N-1}|f(k)|^2} \quad (5)$$

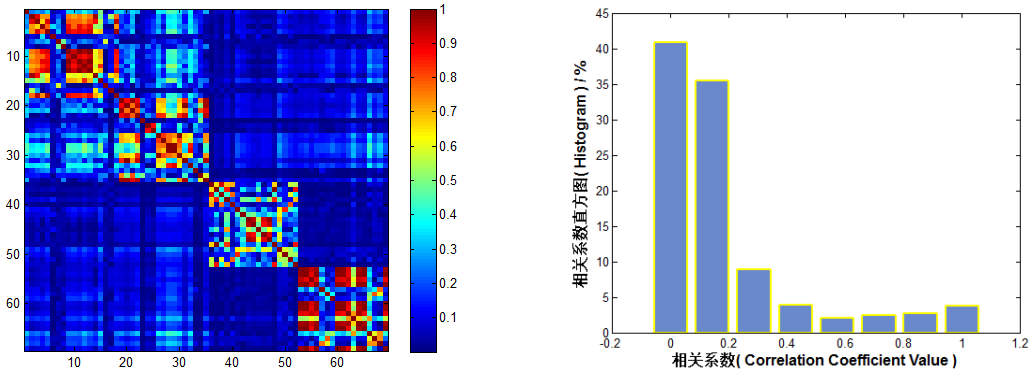
假定某寄递数据的发、收件地址分别为 A 和 B，发件时间为 t（格式为“年-月-日-时段”），根据 A、B 和 t 的值可以得到 t 之前 A、B 的寄递时空模式。按照表 1 计算上述模式的 17 维时间序列特征可得：V_{Ay}（A 涉烟模式）、V_{An}（A 不涉烟模式）、V_{By}（B 涉烟模式）和 V_{Bn}（B 不涉烟模式）。融合上述特征得到 69 维向量 $V=(t,V_{Ay},V_{An},V_{By},V_{Bn})$作为该寄递数据的时空特征并用于后续机器学习模型的训练和测试。

1.4 时空特征分析与选择

本文对表 1 计算所得的 69 维时空特征进行了相关性分析。从实验数据（见本文第 2 节）中的涉烟和不涉烟样本中分别随机挑选了 30%的样本进行分析，可得相关性系数热力图与直方图如图 4 所示。图 4（a）中的热力图蓝色表示低相关，红色表示高相关，颜色越深相关性越大。由图 4 可知，发件地址的涉烟时空模式特征和收件地址的不涉烟时空模式特征存在一些相关性较大的特征。将所有相关系数分为 8 柄，统计其直方图得图 4（b）。由图 4（b）可知，约有 75%的系数值在[0, 0.25]之间，这表示计算所得的时空特征可以较好描述寄递数据特性。

为了计算优选特征，根据图 4（b），本文选择总特征数的 40%（28 维特征）进行实验。使用的特征选择算法为 CFS（Correlation-based Feature Selection）[28]，这是一种计算速度快且较为有效的特征选择算法，适于处理寄递大数据。特征选择实验进行了 1000 次，每次运行时使用 Hold-Out 方法[29]随机选择数据集中 70%的样本进行计算。实验结果如图 5 所示。图中纵坐标为单个特征选取次数，对应的特征序号标注在柱状图上。图 5 按照特征选取次数降序排列。由图 5 可知，序号为 36，即收件地址涉烟时空模式均值特征（见表 1）被选取次数最多（940 次）。序号为 33，即发件地址不涉烟时空模式 1D-C0 复杂性特征（见表 1）被选

取次数次之（902 次）。说明这些特征适于描述寄递数据。表 2 给出了优选特征计算结果。为便于描述，表 2 将表 1 中各个特征组分别编号为 A～E 组，每个组内特征从 1 开始顺序编号，如 A1～A3 即为“基本统计特征”中的时间序列均值、标准差和中位数特征。由表 2 可知，一些特征在多个组别中均被选择，如 B1（寄递频度最大值）、B4（寄递频度最大值最后 1 次出现位置的百分比）、D1（1D-C0 复杂性）等特征，这表明设计的这些特征具有描述不同类型寄递记录的能力，这是构建寄递涉烟犯罪分析模型的理论和实验依据。



（a）热力图(Heat map)　　（b）直方图(Histogram)

图 4 时空特征相关性分析

Fig.4 Spatio-temporal features correlation analysis

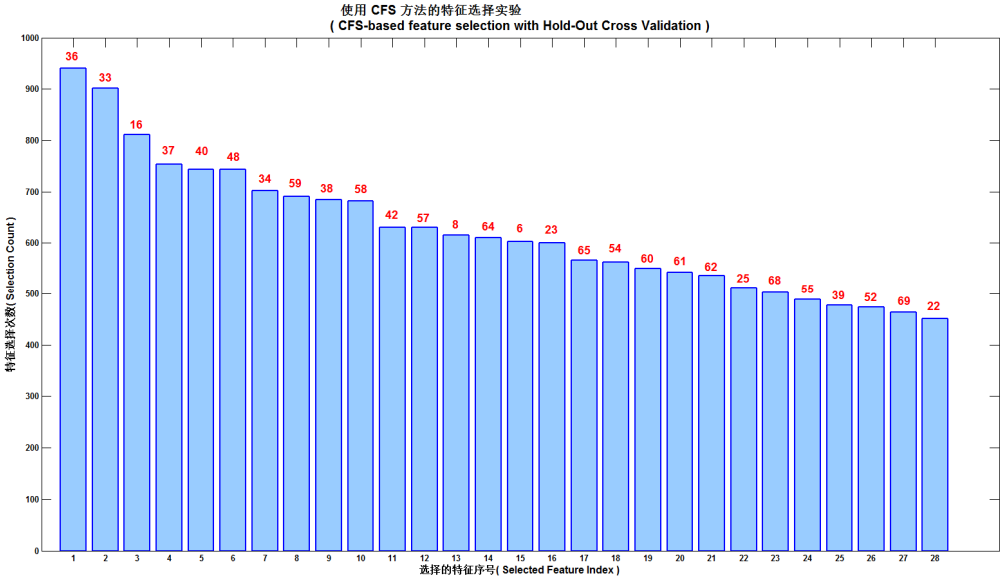


图 5 基于 CFS 算法的特征选择

Fig.5 Feature selection based on CFS(correlation-based feature selection) method

表 2 优选时空特征集合

Tab.2 Optimal spatio-temporal features set after feature selection

组别（Group）	序号（Index）	说明（Description）
发件地址涉烟时空模式	6、8、16	B2、B4、D1
发件地址不涉烟时空模式	22-23、25、33-34	B1-B2、B4、D1、E1
收件地址涉烟时空模式	36-40、42、48、52	A1-A3、B1-B2、B4、C2、E2
收件地址不涉烟时空模式	54-55、57-62、64-65、68-69	A2-A3、B1-B6、B8、C1、E1-E2

1.5 分类器

由图 1 可知，计算寄递大数据的时空特征后，可使用分类器模型和时空特征一起建立寄递涉烟分析模型。寄递记录数据量巨大且涉烟研判必须实时，本文实验中使用了几种计算速度快且广泛使用的传统分类器方法，包括随机森林、逻辑回归和 GBDT（Gradient Boosting Decision Tree）[29]。实验中还使用了近年来在时间序列分析方面广泛使用的深度学习网络 LSTM（Long Short-Term Memory）[30]进行对比研究。

2 结果

2.1 实验数据、环境及参数

实验中使用的寄递数据来源于各快递公司和提供互联网服务的第三方公司，包含采集自 2014 年 1 月至 2021 年 3 月，经过烟草专卖管理部门验证的原始涉烟寄递样本 185,430 条和不涉烟寄递样本 210,298 条。其中涉烟寄递样本是指寄递假烟、走私烟或多于 2 条（400 支）真烟的包裹；不涉烟寄递样本是指没有寄递卷烟或单人单次寄递 2 条（400 支）以内真烟的包裹。原始数据存在数据缺失等问题，经本文第 1 节算法处理后得到满足要求的涉烟样本 3077 个，不涉烟样本 2851 个。

本文实验中所有程序均在配置为 Intel® Core® i9-10900X 十核 64 位处理器 CPU、64G 内存、3080Ti GPU，Windows® 10 中文版（64 位系统）操作系统的服务器上运行。程序使用 Matlab® R2017b 和 Python3.6.2 编写。

表 3 给出了实验中使用的各个分类器设置的主要参数。

表 3 实验中设置的分类器主要参数

Tab.3 Main parameters of different classifiers used in the experiments

分类器名称（Classifier）	分类器参数（Classifier parameters）
随机森林	最大迭代次数 100、用均方差值选择特征
逻辑回归	惩罚项为 L2 范数、用拟牛顿法 LBFGS 作为优化函数、最大迭代次数 100
GBDT	损失函数为均方差、学习率 0.1、弱学习器最大迭代次数 100、树的最大深度 3
LSTM	隐含层数为 10、训练次数 100、堆叠层数 1

2.2 实验结果

判断寄递数据是否涉烟和医学中判断个体是否患病从评价指标上来说是完全类似的。本文实验中使用医学领域中的评价指标来度量模型性能。定义“阳性”为寄递数据涉烟，“阴性”为不涉烟。定义 *TP*、*FP*、*TN*、*FN* 为本文方法预测的真阳性样本数（预测样本涉烟，实际涉烟）、假阳性样本数（预测样本涉烟，实际不涉烟）、真阴性样本数（预测样本不涉烟，实际不涉烟）和假阴性样本数（预测样本不涉烟，实际涉烟）。定义评价指标准确率（Accuracy）、阳性预测值（Positive Predictive Value，PPV）和阴性预测值（Negative Predictive Value，NPV）如公式(6)-(8)所示。显然这些指标值越大表明模型性能越好。

$$Accuracy=(TP+TN)/(TP+FN+FP+TN) \quad (6)$$

$$PPV=TP/(TP+EP) \quad (7)$$

$$NPV=TN/(TN+FN) \quad (8)$$

实验中使用 Hold-Out 方法将整个数据集按照 6:3:1 的比例划分为训练集、测试集和验证集，训练并通过测试的分类器模型在验证集上计算其指标[29]。表 4 给出了使用初始 69 维特征进行预测的实验结果，表中加粗、加下划线指标为最佳结果。

表 5 给出了使用 1.4 节 CFS 算法选取的 28 维优选特征进行预测的实验结果。

表 4 实验结果 I

Tab.4 Experimental results I

分类器名称（Classifier）	准确率（Accuracy）	阳性预测值（PPV）	阴性预测值（NPV）
随机森林	<u>0.9752</u>	0.9744	<u>0.9761</u>
逻辑回归	0.9668	0.9659	0.9678
GBDT	0.9718	<u>0.9752</u>	0.9681
LSTM	0.9448	0.9319	0.9589

表 5 实验结果 II
Tab.5 Experimental results II

分类器名称（Classifier）	准确率（Accuracy）	阳性预测值（PPV）	阴性预测值（NPV）
随机森林	0.9430	0.9440	0.9416
逻辑回归	0.9364	0.9269	0.9504
GBDT	0.9467	0.9386	0.9599
LSTM	0.9364	0.9257	0.9522

为了和 CFS 特征选择算法进行对比，本文还使用了 PCA（Principal Component Analysis）算法[29]对特征数据进行降维，使用降维后的优选特征进行寄递数据涉烟判断。PCA 将初始的 69 维特征降为 28 维，表 6 为实验结果。

表 6 实验结果 III
Tab.6 Experimental results III

分类器名称（Classifier）	准确率（Accuracy）	阳性预测值（PPV）	阴性预测值（NPV）
随机森林	0.9623	0.9472	0.9802
逻辑回归	0.9488	0.9339	0.9666
GBDT	0.9606	0.9452	0.9790
LSTM	0.9516	0.9397	0.9651

通过寄递渠道运送假、私卷烟 2～3 天就能够到达目的地。烟草专卖管理部门必须在 2～3 天内完成对寄递涉烟包裹的发现、跟踪和打击等一系列工作。寄递涉烟分析模型的响应速度必须是实时的。本文计算了上述分类器模型分析寄递数据的平均处理时间（模型训练是离线过程，实验中只计算了模型在线预测时间），如图 6 所示。实验中分类器模型处理时间都非常短，为了能直观对比各种模型的计算时间，图 6 使用了对数坐标（由于时间值小于 1，图 6 纵轴值有负数）。图 6 横轴表示分别预测 200、500、1000、1500 和 2000 条寄递记录，纵轴表示各个模型的对应预测平均时间的对数值。

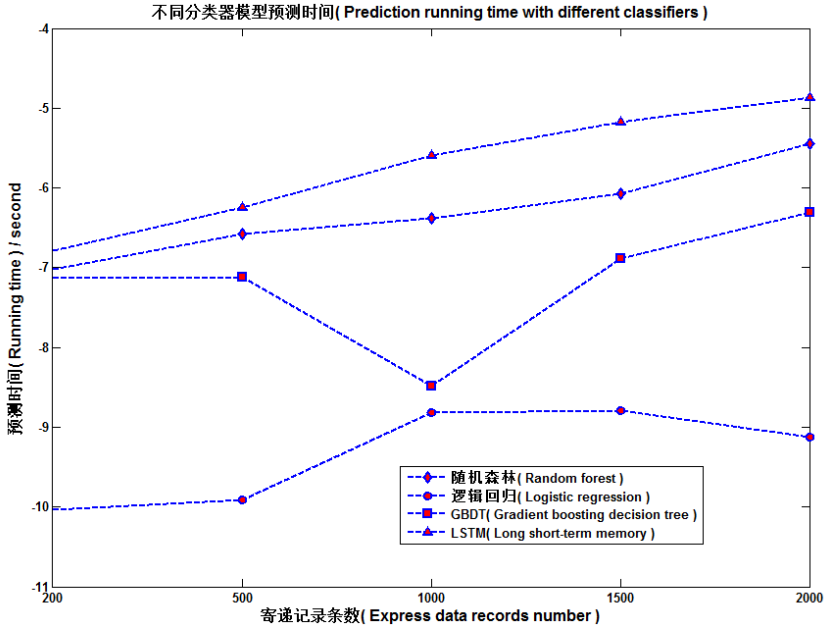


图 6 不同分类器的预测时间
Fig.6 Prediction running time with different classifiers

3 讨论

针对第 2 节的实验结果讨论如下：

由表 4 可知，使用的 4 种分类器方法都取得了较好的实验结果，所有评价指标值都在 0.9 以上。随机森林在 Accuracy 和 NPV 这两个指标上的结果最好，略高于次佳的 GBDT。在 PPV 指标上，GBDT 结果最佳，随机森林次之，略低于 GBDT。这说明本文 1.4 节中提出的时空特征具有较好描述涉烟和不涉烟寄递数据特性的能力。随机森林在判定包裹涉烟与否的正确率及包裹是否不涉烟（NPV）的正确率上表现较好。GBDT 在判定包裹是否真正涉烟（PPV）的正确率上表现较好。逻辑回归和 LSTM 的 3 个指标值和上述 2 个方法相比低约 1%～3%左右。LSTM 在本次实验中表现不佳，表明深度网络在寄递大数据涉烟预测性能方面仍需进一步挖掘。随机森林和 GBDT 整体表现能力最佳，说明集成多棵决策树的学习模型更适于分析寄递涉烟数据。

由表 5 可知，随机森林和 GBDT 仍然取得最佳预测效果。表 5 中最好结果较表 4 低了 2%～3%，这说明 CFS 算法选取的特征集合中丢失了一些可以区分涉烟和不涉烟寄递数据的信息。但需要说明的是，原始特征向量为 69 维，特征选择后的向量为 28 维。储存选择后特征数据空间仅为原始数据的 40%，这意味着使用更少的特征数据就可以取得相近的预测结果。这个结论对构建基于寄递大数据的涉烟犯罪分析系统具有应用价值。此外 CFS 方法可选出具体特征，这为模型预测结果的可解释性提供了依据（见 1.4 节）。

由表 6 可知，随机森林取得最佳分类效果，相较表 4 的最好结果 Accuracy 值低了 1%，PPV 值低了 3%，但 NPV 值略高。相较表 5 的最好结果 Accuracy 和 NPV 值高了 1%，PPV 值相当。表 6 说明 PCA 降维方法略好于 CFS 方法。但 PCA 的结果是在整个数据集上取得的，其计算量要远大于 CFS 方法。此外，在对模型预测结果的可解释性方面 PCA 方法不如 CFS 方法。和 CFS 方法类似，PCA 方法也可以减少特征数据储存量。

综合表 4-表 6 的实验结果可知，本文提出的时空特征能够较好地区分涉烟和不涉烟寄递数据。基于决策树集成的随机森林和 GBDT 模型更适于预测寄递数据。

由图 6 可知，所有分类器模型的预测时间都很短，且随着寄递记录数目的增加增长缓慢。图 6 表明本文方法构建的预测模型可以满足寄递涉烟分析的实时性要求。

4 结论

本文的贡献和结论主要有：（1）将大数据和人工智能技术用于寄递涉烟犯罪分析中，在已有文献中还未见类似工作报道。提出了“寄递时空模式”新概念并据此计算了寄递大数据的时空特征。结合优选特征计算和分类器方法构建的寄递涉烟预测模型在研判精度及实时性方面取得了初步成果，证实了本文方法的可行性和有效性。本文工作对推进烟草行业专卖执法数字化管理，维护国家和消费者利益具有一定应用价值。

（2）实验结果证实基于时空特征和随机森林、GBDT 等方法建立的模型实验效果最好。基于 CFS 特征选择和 PCA 特征降维建立的模型可以取得和初始特征接近的实验结果，但特征数据储存量大为降低。CFS 特征选择方法可以挑选出具体特征，这为模型预测结果的可解释性提供了依据。

（3）实验表明本文方法构建的预测模型可以满足寄递涉烟分析的实时性要求。

尽管取得了一些成果，本文工作仍具一定局限性：一是本文实验中使用的数据集容量较小，需要在更大的数据集上验证本文方法；二是寄递数据集合呈现不平衡性，即涉烟包裹数量要远小于不涉烟包裹数量。后续工作在算法设计上需要考虑这一特点；三是后续工作中应进一步研究深度学习方法在寄递大数据涉烟犯罪分析中的应用。

参考文献

[1] 卜心农. “互联网+”新商业模式下涉烟违法犯罪的现状及对策[J]. 发展，2017, 1:87-88.
BU Xinnong. Status analysis and countermeasure study of tobacco related crimes with novel Internet+ business model[J]. Developing, 2017, 1:87-88.

[2] 高磊. 浅析“互联网+”新模式下涉烟违法犯罪的现状及对策[R]. 中国烟草学会 2016 年度优秀论文汇编：专卖管理主题，2016.
GAO Lei. Brief status analysis and countermeasure study of tobacco related crimes with novel Internet+ model[R]. 2016 Excellent papers compilation of Chinese Tobacco Association:

special issue on monopoly administration, 2016.

[3] 黄祥铭. 制售假烟犯罪案件侦查研究[D]. 中国人民公安大学，2019.
HUANG Xiangmin. Research on the investigation of the crimes of producing and selling counterfeit cigarettes[D]. Dissetation of People's Public Security University of China, 2019.

[4] Paul P V, Monica K, Trishanka M. A survey on big data analytics using social media data[C]. 2017 Innovations in Power and Advanced Computing Technologies, Vellore, India, 21-22 April, 2017.

[5] Li Q, Chen Y, Wang J. Web media and stock markets: a survey and future directions from a big data perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(2):381-399.

[6] Bahri S, Zoghlami N, Abed M. Big data for healthcare: a survey[J]. IEEE Access, 2020, 7: 7397-7408.

[7] Zhu L, Yu F R, Wang Y G. Big data analytics in intelligent transportation systems: a survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1):383-398.

[8] Cun Y L, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521: 436-444.

[9] Greenspan H, Ginneken B V, Summers R M. Deep learning in medical imaging: overview and future promise of an exciting new technique[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1153-1159.

[10] Shi F, Wang J, Shi J. Review of artificial intelligence techniques in imaging data acquisitionm, segmentation and diagnosis for covid-19[J]. IEEE Reviews in Biomedical Engineering, 2020, 14:4-15.

[11] Pham Q V, Nguyen D C, The T H. Artificial intelligence and big data for coronavirus pandemic: a survey on the state-of-the-arts[J]. IEEE Access, 2020, 8: 130820-130839.

[12] 谭旭. 基于物流数据的快递网络分析与建模[D]. 浙江大学，2015.
TAN Xu. Analyzing and modeling express shipping service network based on logistics data[D]. Dissetation of Zhejiang University, 2015.

[13] 任思源，郭斌，张曼. 寄递大数据城市画像[J]. 浙江大学学报（工学版），2019, 53(9): 1779-1787.
REN Siyuan, GUO Bin, ZHANG Man. Urban profiling using express big data[J]. Journal of Zhejiang University (Engineering Science), 2019, 53(9): 1779-1787.

[14] 刘二超. 快递服务便利店选址问题研究[D]. 清华大学，2014.
LIU Erchao. Location planning for express service convenience store[D]. Dissetation of Tsinghua University, 2014.

[15] Li W G, Zhong X Y. Research on the application of smart logistics system based on big data: taking jingdong logistics as an example[C]. 2021 IEEE International Conference on Artificial Intelligence and Industrial Design, Guangzhou, China, 28-30 May, 2021.

[16] 郝晟. 面向侦查的快递数据分析挖掘系统[D]. 天津大学，2014.
HAO Sheng. Express data mining system for technical investigation[D]. Dissetation of Tianjin University, 2014.

[17] 李万彪，余志，龚峻峰. 基于关系数据模型的犯罪网络挖掘研究[J]. 中山大学学报（自然科学版），2014, 53(5):1-7.
LI Wanbiao, YU Zhi, GONG Junfeng. An approach of crime network analysis based on association data model[J]. Journal of Sun Yat-sen University, 2014, 53(5): 1-7.

[18] 文杰锋. 快递物流配送异常检测方法研究[D]. 重庆邮电大学，2016.
WEN Jiefeng. Research on outlier detection algorithm for express logistics[D]. Dissetation of Chongqing University of Posts and Telecommunications, 2016.

[19] 郭小伟. 我国利用快递运输违禁品犯罪研究[D]. 中国人民公安大学，2019.
GUO Xiaowei. Research on the crime of using express delivery to transport contraband in China[D]. Dissetation of People's Public Security University of China, 2019.

[20] 王宁. 寄递渠道贩毒案件侦查研究[D]. 中国人民公安大学, 2020.
WANG Ning. Research on the investigation of drug trafficking cases through delivery channels[D]. Dissetation of People's Public Security University of China, 2020.

[21] 徐志成，严超，陶忆南. 浅谈构建物流运输业涉烟犯罪网络监管体系[R]. 中国烟草学会 2016 年度优秀论文汇编：专卖管理主题，2016.
XU Zhicheng, YAN Chao, TAO Yinan. Discussion on the construction of the network monitoring system of tobacco related crimes in logistics industry[R]. 2016 Excellent papers compilation of Chinese Tobacco Association:special issue on monopoly administration, 2016.

[22] 曾超. 涉烟犯罪问题研究[D]. 中南大学, 2012.
ZENG Chao. A study on the crime involving tobacco[D]. Dissetation of Central South University, 2012.

[23] 刘言. "互联网+"背景下烟草专卖管理研究[D]. 华中师范大学, 2018.
LIU Yan. Research on tobacco monopoly administration in context of Internet+[D]. Dissetation of Central China Normal University, 2018.

[24] Goldberg Y. 基于深度学习的自然语言处理[M]. 车万翔等译. 北京：机械工业出版社, 2020:57-67.
Goldberg Y. Deep learning-based Natural Language processing[M]. Beijing: China Machine Press, 2020:57-67.

[25] 结巴分词[N/OL]. Github 网站. (2018-07-23)[2021-08-16]. https://github.com/fxsjy/jieba.
Jieba word segmentation[N/OL]. Github website. (2018-07-23) [2021-08-16]. https://github.com/fxsjy/jieba.

[26] 陈芳，顾凡及，徐京华. 一种新的人脑信息传输复杂性的研究[J]. 生物物理学报，1998, 14(3):508-512.
CHEN Fang, GU Fanji, XU Jinghua. A new measurement of complexity for studying EEG mutual information[J]. Acta biophysica sinica, 1998, 14(3):508-512.

[27] 沈恩华. 脑电的复杂度分析[D]. 复旦大学，2004.
SHEN Enhua. Complexity analysis of EEG signals[D]. Dissetation of Fudan University, 2004.

[28] Hall M A. Correlation-based feature selection for machine learning[D]. Dissetation of The University of Waikato, 1999.

[29] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.
ZHOU Zhihua. Machine learning[M]. Beijing:Tsinghua University Press, 2016.

Express-related counterfeit cigarette criminality analysis based on spatio-temporal data features

QIAO Langchao[1], WANG Jinlu[1], GAO Baohong[1], YANG Xingang[1], FENG Wentao[1], XU Rongyao[2], WEI Yiran[2], LIU Wei[2*]

1 Division of Monopoly Administration, Shaanxi Branch of China National Tobacco Corporation, Xi'an 710061, China;
2 School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Abstract: [**Background**] This study aims to study the express-related counterfeit cigarette criminality based on big data and artificial intelligence technology. [**Methods**] In the pre-processing stage, Chinese word segmentation method was adopted to process the original data. Then a novel concept named "spatio-temporal pattern of delivery and receiving address" was presented, which is actually time series data established based express package delivery and receiving frequency data within a time span. Spatio-temporal data features can be computed based on spatio-temporal pattern by using time and frequency domain statistics. Next, a CFS (Correlation-based feature selection) or PCA (Principal component analysis) algorithm was applied for the initial spatio-temporal feature pool to determine an optimal feature cluster. Then, the express-related counterfeit cigarette criminality analysis model was trained and optimized and the performance of models using different classifiers was compared. [**Results**] (1) All four classifier models including random forest, logistic regression, gradient boosting decision tree and long short-term memory deep neural network applied in the experiments achieved encouraging experimental results with satisfactory accuracy, PPV and NPV, which implied the proposed spatio-temporal data features has the ability to discriminate the cigarette-related from normal express data. Decision tree-based classifier models like random forest and gradient boosting decision tree classifier yielded the highest accuracy, PPV and NPV, which were all greater than 0.94. (2) Prediction models with optimal feature cluster determined by CFS (Correlation-based feature selection) or PCA(Principal component analysis) algorithm all exhibited slightly lower performance than that of initial feature pool. The storage space of optimal feature cluster accounted for only 40 percent of the initial feature pool. (3) CFS method utilized in the experiments can pick out optimal feature cluster from initial feature pool, which supports the interpretability for prediction results generated by the model. (4) Preliminary experimental results showed that the proposed prediction model can meet the real-time requirements of express-related counterfeit cigarette criminality analysis. [**Conclusion**] Classifiers in cooperated with the spatio-temporal data features computed based on "the spatio-temporal pattern of delivery and receiving address" can discriminate counterfeit cigarette-related express packages from normal express packages.

Keywords: express-related counterfeit cigarette criminality; spatio-temporal pattern of package delivery and receiving; time series analysis; feature selection and reduction; machine learning

***Corresponding author. Email:** liuwei@xupt.edu.cn