基于随机森林和SVR的阿片类药物危机分析

李 军,赵 佳,赵 宸

(阜阳师范大学 计算机与信息工程学院,安徽 阜阳 236041)

摘 要:阿片类药物危机严重威胁着社会的稳定与发展,美国作为危机发生的严重地带,当地政府已下达了针对阿片类药物危机的紧急戒严。依据由NFLIS提供的2010-2017年统计美国五个州的药品事件与人口分布情况,由层次化分析,建立随机森林模型计算特征重要性,随后再通过回归模型,对美国部分城市未来的受影响程度作预测。在回归模型部分,分别使用了深度学习中的Seq2Seq与滑动窗口机制的SVR两种方法分析,通过实验结果对比二者在该问题上的优劣,对五个州内城市2017年的受影响程度作预测。结果表明模型对其中的四个州的预测效果较好。

关键词:随机森林;阿片类药物危机;SVR;Seq2Seq

中图分类号:O22;TP391 **文献标识码:**A **文章编号:**1004-4329(2019)04-009-05

DOI:10.14096/j.cnki.cn34-1069/n/1004-4329(2019)04-009-05

Opioid crisis analysis based on random forest and SVR with sliding window

LI Jun, ZHAO Jia, ZHAO Chen

(School of Computer and Information Engineering, Fuyang Normal University, Fuyang Anhui 236037*, China)*

Abstract: Opioid crisis is a serious threat to social stability and development. As a serious area of the crisis, the local government has issued an emergency martial law against the opioid crisis in the United States. According to the statistics of drug incidents and population distribution in five states provided by NFLIS from 2010 to 2017, a stochastic forest model was established to calculate the importance of characteristics based on hierarchical analysis, and then a regression model was used to predict the future impact of some cities in the United States. In the part of regression model, Seq2Seq in depth learning and SVR in sliding window mechanism are used to analyze two methods respectively. Through the experimental results, the advantages and disadvantages of two methods on this issue are compared, and the impact degree of five cities in the state in 2017 is predicted. The results show that the model has a good prediction effect for four of the states.

Key words: random forest; opioid crisis; regression model; sliding window

阿片类药物是作用于阿片受体以产生吗啡类作用的物质,医学上主要用于缓解病痛、麻醉等。长期使用阿片类药物会导致大脑中的区域结构和功能的发生变化,停止使用会产生焦虑不安。严重可能致命,因此该类药物的使用是受管制的[1-4]。在美国1999年前,有86%阿片类药物的使用者用于非癌症疼痛治疗;2002-2013年,海洛因的滥用导致死亡人数增加了286%,其中约有80%的海洛因使用者在转向使用海洛因之前误用处方类阿片药物。据统计,仅2013年全球有2 800万

收稿日期:2019-05-30

基金项目:安徽省教育厅重点项目(KJ2019A0542);大学生创客实验室(2018CKJH01);安徽省质量工程项目(2018jyxm0507);阜阳师范学院创新团队项目(XDHXTD201703,XDHXTD201709)资助。

作者简介:李 军(1981-),男,硕士,讲师,研究方向:数学建模、计算机应用。

~3 800 万人非法使用阿片类药物(占 15~65 岁之间全球人口的 0.6%~0.8%)。根据美国疾病控制和预防中心的数据显示,阿片类药物的滥用已经导致美国历史上最为严重的药物过量使用,并于 2014 年将该问题列入五大公共卫生挑战之一。截至 2015 年,阿片类药物的娱乐用量和成瘾率上升的原因是阿片类药物的处方药的过度使用和廉价的非法海洛因。2013 年芬太尼等合成阿片类药物相关的死亡人数增加,其中 2016 年芬太尼及相关药物死亡人数就超过 2 万人。1999~2017 年统计数据显示,非法使用阿片类药物的群体中,男性比女性占有更大的人数比例,18~25 岁的年龄群体且更容易接触到阿片类药品。

阿片类药物的滥用严重威胁着社会的和谐稳定,采取科学合理的分析方法可以极大地遏制危机的影响范围与影响程度。在现有数据的基础上,本文通过基于机器学习的方法[5-6],通过评价指标,挖掘事件的潜在影响因子,预测未来的趋势,为后续的决策提供重要的参考标准。

1 预处理

分析的数据集统计了美国 Ohio, Kentucky, West Virginia, Virginia, Pennsylvania 五个州和州内 462 个城市的数据,具体城市数如表 1。数据集分别是:(i)统计 2010-2017 年间,城市上报的不同种类药品的事件报道数目与 5 个州当年总共的药品事件报道数目,共计 24 063 条记录;(ii)统计了 462 个城市 2010-2016 年间的人口普查数据,合计有 2 272 个样本数,149 个人口特征项。

表 1 五个州的城市数量统计

州名	城市数
Ohio	87
Kentucky	55
West Virginia	67
Virginia	133
Pennsylvania	120

数据集里 2010-2012 年的人口普查数据中,有 8 个特征项是缺失的,选择直接删除这些特征项。对于有少部分城市的部分特征项有数据缺失,使用 K 近邻的方法填补缺省值[7]。以各城市的人口统计数据为样本集合,设第 i 个有缺省特征项的城市为 X_i,则第 i 个城市缺省特征项的集合为 S_i,记无缺失特征项城市的集合为 P,没有缺省特征项的城市 j 为 Y_j,计算样本 i 与 j 的欧式距离 $D_{i,j}$

$$D_{i,j}=\sqrt{\sum_{k\in Q-S}(X_{i,k}-Y_{j,k})^2} \tag{1}$$

由(1)式,选择与缺失样本最近的前 K 个样本,构成集合 Q,按均值填补缺省值

$$X_{i,S_i}=\frac{1}{K}\sum_{k=1}^{K}\sum_{q\in Q}C_{k,q} \tag{2}$$

在人口普查数据中,各特征项取值范围相差过大,先对数据预先做了 Z-SCORE 标准化的处理。为了在后续使数据更加直观地展示,又收集了当地城市的经纬度坐标信息。

影响阿片类药物事件的因素有很多,在已拥有的数据与相关背景的基础上,将影响因子归类为三种,分别是药品流行度,当地政策和人口分布。同时将第 i 个城市第 j 年内上报的药品事件数作为衡量该城市当年的受危机影响程度。

1.1 阿片类药品流行度分析

对于不同的城市来说,由于存在着差异性,反映在抵抗外界药物危机入侵的抵抗力不同与控制城市内部药品危机传播的能力也不同。假设把每个城市比作一个独立个体及一定区域内的城市比作群体,需要分析个体和群体两个方面药品流行度。在对 2010-2017 年各城市药品事件数的可视化过程中,发现各州各城市的变化趋于一个稳定的状态。据此,忽略五个州相互间的影响,建立药品流行度指标来作为城市受外界干扰程度的量化标准。

阿片类药品分成处方药与毒品两大部分。通过对五个州 2010-2017 年不同种类药品的事件数进行汇总统计发现,Heroin, Oxycodone, Fentanyl 的药品事件占据了 80%左右。其中 Heroin 划分为毒品一类,占了整个事件的 50%;而 Oxycodone 和 Fentanyl 划分为处方药一类,使用数据可视化进一步分析发现,Oxycodone 事件处于一个逐年下降的趋势,Fentanyl 逐年上升。综合考虑,将 Fentanyl 作为主要的处方用药,Heroin 作为主要的毒品。数据分析发现各州之间的药品事件总数有着显著的差距,因此一个城市的药品流行度是以所在州为衡量标准。通过上述分析,建立药品流行度计算公式(3)

$$d_{i,j}=\frac{\alpha \mathrm{H}_{i,j}+\beta \mathrm{F}_{i,j}+(1-\alpha-\beta)Z_{i,j}}{N_{i,j}} \tag{3}$$

其中:$d_{i,j}$ 表示第 i 个城市第 j 年的药品流行度,该城市 Heroin 药品事件数为 $\mathrm{H}_{i,j}$,Fentanyl 药品事件

数为 $F_{i,j}$,除 Heroin 和 Fentanyl 的药品事件总数为 $Z_{i,j}$;$N_{i,j}$ 表示第 i 个城市所在州的第 j 年药品;α,β 为参数,分别设定成 0.30 和 0.50。

1.2　当地政策分析

对于不同的城市来说,可变因素较多,加上数据集中相关信息的较少,直接从政策的角度分析是比较困难,而从政策所带来的效应分析是一种可行的方案。本文从不同的人口分布出发,实施不同的政策会在不同的人口组成特征上有所体现。例如,若当地城市政策对外国移民不太"友好",则本地的外国人口所占比例会较低。假设在无政策干扰的情况下人口组成特征信息矩阵为 X,表示经过政策的干扰后表现出的人口组成特征信息矩阵 $\hat{X}$,则有

$$\hat{X}=f(X,\theta) \tag{4}$$

因此,将当地政策问题转化成对人口普查数据信息的特征选择。

1.3　人口分布分析

不同地区的人口分布数据中特征数目较大,直接使用回归模型预测很容易导致过拟合现象。如果使用 PCA(principal components analysis)对数据进行降维[8-9],带来的问题是数据的可解释性较差,实际应用的参考价值也较低。集成学习方法在数据挖掘领域有着非常广泛的应用,随机森林作为 bagging 的一个变体,以基于 Gini 指数[10]的特征选择,可以同时获得降维和从分类任务中消除噪声的双重效益。

在对特征进行重要性分析过程中,直接将所有的特征一次性传入随机森林模型会产生一个问题:即一些重要性程度不大的组成特征可能会携带很多隐含的政策信息,同时各个特征间的差异性并不明显。因此这里将特征分为两个层次,即为一级层次因素和二级层次因素。一级层次因素作为上级层次,包含有多个二级层次因素。此处的模型是单独地对每个一级层次因素进行建模,分别计算所有一级层次因素下二级层次因素的特征重要性,随后再对所有的一级层次因素建立随机森林模型,计算特征重要性。

使用 CART 作为随机森林的基学习器。CART 决策树使用 Gini 指数来选择划分属性。设当前样本集合 D 中第 m 类样本所占比例为 $p_k(k=1,2,\cdots,|y|)$,属性 a 有 V 个可能的取值 $\{a^1,a^2,\cdots,a^v\}$,则属性 a 的 Gini 指数为

$$\text{Gini\_ind ex}(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\text{Gini}(D^v) \tag{5}$$

其中,

$$\text{Gini}(D)=1-\sum_{k=1}^{|y|}p_k^2 \tag{6}$$

在候选属性集合 A 中,选择使划分后 Gini 指数最小的属性作为最优划分属性,即

$$a^*=\arg\min \text{Gin\_ind ex}_{(D,a)}(a\in A) \tag{7}$$

训练的过程中,将 2010-2016 年的人口普查数据作为自变量,当年各城市的药品事件数作为因变量,设置集成 CART 数目为 80,得到各二级层次因素的特征重要性如图 1 所示,与一级层次因素的特征重要性。

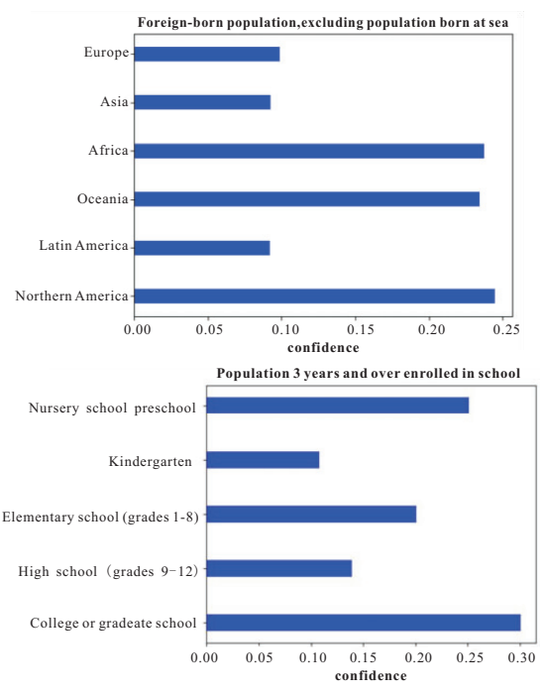

图 1　部分二级层次因素的特征重要性

特征方面选取的原则是:1)在一级层次选取占比最大的前三个及该一级层次下所有的二级层次因素;2)其它一级选取二级层次因素选取特征重要性占比最大的前 1~3 个。基于此,最后选择了 25 个特征用于模型训练。

2　预测模型

预测方面,由于问题是一个时间序列的问题,因此采用 Seq2Seq 模型[11]和带有滑动平均机制下的 SVR(support vector regression),通过模型的特点与实验的结果对问题作进一步的分析与结论。

2.1　基于 Seq2Seq 模型分析

Seq2Seq 是一中的 RNN,在自然语言处理上

取得了巨大的成功。本文将该模型用于对数值的预测,使用 GRU 作为基本的单元,网络模型如图 2 所示。

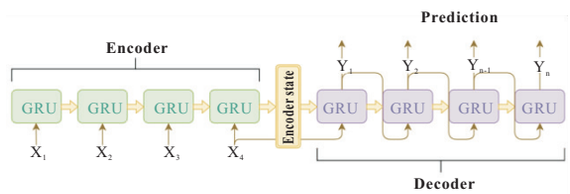

图 2 Seq2Seq 模型

将 2010-2016 年人口分布选择出的 25 个特征与阿片类药物的流行度,各城市每年的药品事件数统一成新的数据集。其中选取 412 个城市用于训练,50 个用于测试。使用了 dropout 正则化方法,损失函数选择了 MSE,在 100 个隐层下进行训练。经过多次尝试(包括微调网络模型结构与调整 dropout 的概率),由于数据量的不足导致训练结果始终不是很好,并且实际的预测结果也与实际偏差较大,如图 3 所示。

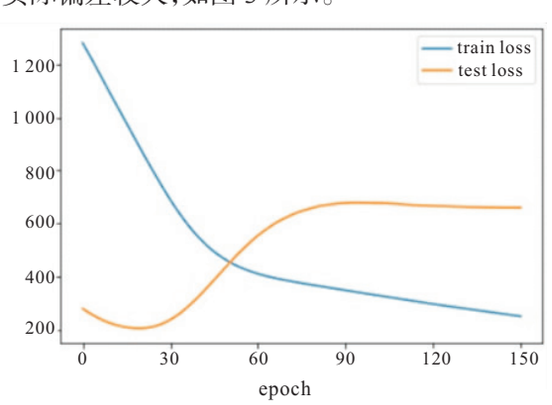

图 3 Seq2Seq 训练过程的 Loss 下降曲线

2.2 基于滑动窗口 SVR 模型分析

对于一般形式下的 SVR[12-13],不能用在对时间序列型数据的预测。基于滑动窗口的思想,对 SVR 的训练过程进行改进,增加约束条件,使之用于预测时间序列型数据。

SVR 模型的优化目标如(8)式:

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i)$$
$$s.t.\begin{cases} f(x_i) - y_i \leqslant \varepsilon + \xi_i, \\ y_i - f(x_i) \leqslant \varepsilon + \hat{\xi}_i, \\ \xi_i \geqslant 0, \hat{\xi}_i \geqslant 0, i = 1,2,\cdots,m \end{cases} \tag{8}$$

其中:w 为超平面法向量;C 为正则化参数;$\xi_i, \hat{\xi}_i$ 为松弛变量;ε 为最大容忍误差。

针对该问题,将每年的数据作为一个基本的单位,为了预测第 T 年的数据,设滑动窗口大小为 w,滑动步长 s。变量间的约束满足(9),训练算法。

$$s.t.\begin{cases} w \leqslant T, \\ s \geqslant w \ , \\ \dfrac{(T-w)}{s} = c, c \in \mathrm{N}^+ \end{cases} \tag{9}$$

使用高斯核函数,在 N=1,S=1,以参数 C=0.001,γ=0.01 建立模型。

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{10}$$

关于训练数据,在之前人口分布所选特征与药品流行度的基础上,又加上滑动窗口内各年内的药品事件数作为另一项或几项特征,通过训练 2010-2016 年的数据,预测 2017 年药品事件数。统计了各州的结果如表 2。

表 2 五个州药品事件数结果统计

州名	预测值	真实值	相对误差	平均误差
Ohio	81 676	110 081	0.258 0	326.49
Kentucky	25 701	28 766	0.106 6	25.54
West Virginia	6 648	3 597	0.848 2	55.96
Virginia	30 759	36 962	0.167 8	46.64
Pennsylvania	58 940	68 751	0.142 7	146.44

2.3 结果分析与模型评价

滑动窗口 SVR 模型在五个州的预测结果里,Kentucky,Pennsylvania,Virginia 的预测结果较好,Ohio 由于药品事件发生的基数较大,导致实际误差要高,但相对误差较低。而对 West Virginia 州的预测结果较差。进一步地,通过抽取 West Virginia 的部分异常数据发现,在 2010-2017 年间,West Virginia 部分城市的药品事件数目呈现出一种异常的变化趋势,该模型对这种异常的变化难以估测。根据城市地理位置信息筛选,发现预测与真实偏离较大的这些城市的坐标大多集中于与 Ohio 和 Pennsylvania 的边界城市地区。在实际数

据中，Ohio, Pennsylvania 的药品事件基数较大，药品流动性更大，因此认为是 West Virginia 受到 Ohio，Pennsylvania 过多的影响。然而模型弱化了州与州之间的关系，导致对 West Virginia 部分城市的预测结果较差。又因为模型的局限性，如果要预测某一年的数据，需要知道这一年之前几年的实际结果，导致模型的持久性较差。

Seq2Seq 模型的一些优势可以弥补模型二的一些局限性。在之前的实验中，测试集的训练曲线在训练的早期下降至最低点，这也说明了数据量限制了模型的性能的提升，事实结果是我们很难进一步地获取更多相关的数据。

3　小结

本文在随机森林了和 SVR 的基础上建立模型。实验从数据的角度进行建模分析，挖掘出不同的人口特征项对阿片类药物危机的影响程度。并且发现深度学习方法在处理该类问题上有着很大的潜在优势，但受限于数据量的匮乏，很难进一步提升模型的性能。因此又提出了滑动窗口机制的 SVR 模型，对美国 Ohio, West Virginia, Pennsylvania, Virginia, Kentucky 五个州的 2017 年药品事件数进行预测。实验结果说明了要控制阿片类危机需要对影响力较大的部分人口实施相应的策略，同时对于 West Virginia 需要加强对 Pennsylvania, Ohio 的边界的监督与管理。

参考文献：

[1]　曲直，刘志民 . 世界卫生组织在国家、地区和全球层面对阿片类镇痛药消耗和需求的比较[J]. 中国药物依赖性杂志，2012，21(2)：152-158.

[2]　美国修订阿片类药品的风险管理计划[J]. 中国药物评价，2018，35(5)：375.

[3]　加拿大修订阿片类药品说明书以减少滥用危害[J]. 中国医药导刊 .2018，20(2)：108.

[4]　药物警戒快讯[J]. 中国药物警戒，2016，13(6)：382-384.

[5]　郑大刚 . 基于深度学习的人脸识别研究及其实现[D]. 南京：南京理工大学，2018.

[6]　肖 坚 . 基于随机森林的不平衡数据分类方法研究[D]. 哈尔滨：哈尔滨工业大学，2013.

[7]　姜大光，孙贺娟，易军凯 . 基于距离的相似最近邻搜索算法研究[J]. 北京：北京化工大学学报(自然科学版)，2017，44(5)：96-98.

[8]　刘会河，徐维超，刘舜 . 基于 SVM 的降维方法在三类 ROC 分析中的应用[J]. 计算机与现代化，2016，20(7)：50-54.

[9]　吴晓婷，闫德勤 . 数据降维方法分析与研究[J]. 计算机应用研究，2009，26(8)：2832-2835.

[10]　胡美春，田大钢 . 基于修正参数简化标准的 ID3 改进算法[J]. 计算机与数字工程，2015，43(7)：1182-1186.

[11]　李潇，闵华松，林云汉 . 一种用于 CBR 推理机的案例学习算法研究[J]. 计算机应用研究，2018，35(12)：3689-3693.

[12]　任海军，孙瑞志，刘广利 . 基于 AR_SVR 模型的时间序列预测算法的研究[J]. 计算机工程与设计，2010，31(2)：421-424.

[13]　孙甲琦 . 基于机器学习的遥测系统故障预测技术研究[D]. 北京：中国航天科技集团公司第一研究院，2017.