重慶醫科大學
Chongqing Medical University

本科生毕业论文

论文题目　机器学习模型在2010-2016年美国五个州的阿片类药物使用量预测的应用

作者姓名　杜兴兴

指导老师姓名　曾庆

（职称、单位名称）　教授、卫生统计与信息管理教研室

年级、专业　2016级应用统计学

论文答辩年月　2020年6月

正文目录

机器学习模型在2010-2016年美国五个州的阿片类药物使用量预测的应用

摘要

目的：评估不同机器学习算法预测阿片类药物使用量的预测效果；研究社会、经济因素对阿片类药物使用影响，探索阿片类药物成瘾的成因。

方法：对2010-2016年美国俄亥俄州、肯塔基州、西弗吉尼亚州、弗吉尼亚州和宾夕法尼亚州的阿片类药物进行描述性分析，提取社会、经济属性，通过K最近邻、决策树、随机森林、支持向量机、人工神经网络和Logistics回归6种机器学习算法建立模型，评价和比较不同算法的优劣并进行预测。

结果：报告的阿片类药物共有56种，各类阿片类的分布主要集中在海洛因、氢考酮、氢可酮这三类，分别占比为53.24%、19.79%、8.57%；支持向量机模型的精确度>0.8，而人工神经网络模型的精确度>0.7。

结论：支持向量机模型和人工神经网络模型在预测效果中的性能都比较理想；且对比与传统的预测模型而言，支持向量机和神经网络模型容易调整参数，能够生成预测效能更好的模型。

关键词：阿片类药物 可视化 机器学习 预测模型

Application of Machine Learning Models to Predict 0pioids Use in Five States from 2010 to 2016

Abstract

Objectives：To evaluate the predictive effectiveness of different machine learning algorithms in predicting opioid use. To study the influence of social and economic factors on opioid use and explore the causes of opioid addiction.

Methods：From 2010 to 2016, Kentucky, West Virginia, Virginia, Ohio and Pennsylvania opioids for descriptive analysis, to extract the social, economic attribute, by K Nearest Neighbor, Decision Tree, and Random Forest, Support Vector Machine, Artificial Neural Network and Logistics 6 kinds of machine learning algorithms to establish regression model, evaluation and comparing the pros and cons of different algorithm and prediction.

Results：There were 56 kinds of opioid drugs reported, and the distribution of various opioids was mainly concentrated in heroin, hydrocodone and hydrocodone, accounting for 53.24%, 19.79% and 8.57% respectively. Accuracy of Support Vector Machine model > 0.8, and the accuracy of Artificial Neural Network model > 0.7.

Conclusion：The performance of Support Vector Machine model and Artificial Neural Network model in predicting effect is ideal. Compared with the traditional prediction models, Support Vector Machines and Neural Network models can easily adjust the parameters and generate better prediction models.

Key words: Opioids; Visualization; Machine learning; Prediction model

1. 前言

1.1. 研究背景和意义

研究表明，全球超过 1/5 的人正在遭受疼痛的折磨，其中半数患者的疼痛程度达到中度或重度水平，65 岁以上的老年人慢性疼痛的发生率更是超过 50%。在疼痛折磨着人类的同时，也产生了各种各样的疼痛缓解药物。其中作为有效缓解中、重度疼痛的阿片类药物，虽然许多人需要阿片类药物来控制他们的慢性和严重的疼痛，但这些治疗的一个常见后果是滥用、成瘾和升级到更恶劣的药物[1]。阿片类药物是一种麻醉性止痛药，包括非法毒品海洛因，他们从罂粟花中提取，或人工合成，结构与其他阿片类药物相似[2]。一些阿片类处方药的例子是吗啡、氢考酮、双氢吗啡酮和芬太尼。虽然这类药物中几乎每一种都可以用来治疗慢性疼痛，但阿片类药物的使用已经远远超出了处方药的范围，而且已经成为一种流行病[3]。

据统计，仅 2013 年全球就有 2800 万~3800 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[4]。阿片类药物大体可以为三大类：非合成阿片类药物:可待因、吗啡、鸦片；半合成阿片类药物:氢可酮、羟可酮、丁丙诺啡、海洛因；合成阿片类药物:芬太尼、布托啡诺、美沙酮、丙氧芬。2013 年以来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[5]。

在过去几十年里，阿片类药物滥用不仅仅是美国存在的问题，包括中国、美国、欧洲在内的世界大部分国家和地区都存在该问题。因此，应对阿片类药物滥用是一个全球

性的公共卫生问题。这种药物成瘾对社会的核心稳定发展构成了严重的威胁[6]。

1.2. 研究现状

目前关于此类问题预测的方法主要以传统的 Logistic 回归模型（Logistic Regression）为主。但传统模型在使用条件上较为苛刻，需要考虑数据分布是否合适、变量之间的共线性和交互作用等多种问题，特别是对多变量的问题尤为明显。因此，应用传统回归模型进行预测阿片类药物使用情况具有一定的局限性[7]。

而作为最近几十年才兴起的机器学习算法在人工智能、生物医学、遗传基因等领域大放异彩。相比于传统的统计方法而言，机器学习算法能够有效克服共线性、多变量、交互作用、数据分布未知等众多问题。利用机器学习算法进行预测研究的思路是通过研究历史数据抓取事务的本质特征，以模型或算法为代表的呈现方式，实现分类、预测、回归拟合等分析行为[8]。

机器学习可以分为有监督学习、无监督学习、半监督学习三种类别。有监督学习是有结果变量的一种监督学习方法，通过已有的一部分输入数据与输出数据之间的关系，生成一个函数，将输入映射到合适的输出。无监督学习在学习时并不知道其分类结果是否正确，亦即没有受到监督式增强(告诉它何种学习是正确的)。半监督学习是近年来机器学习中一个备受关注的内容，其基本思想是利用数据分布上的模型假设，建立学习器对未标签样本进行标签。

目前广泛应用的机器学习算法包括神经网络、K 近邻法、支持向量机、决策树算法、随机森林等，这些算法已经广泛应用与工程学、建筑学等领域，却很少有研究将这些算法应用在公共卫生领域，为了更好地评估这些算法是否能够有效地预测阿片类药物使用量以及寻找具有最好分类效果的分类算法，本研究比较了 6 种机器学习算法在阿片类药物使用量的分类预测效能。

本研究通过多种机器学习算法预测阿片类药物使用量来评价和比较不同的机器学习算法，最终确定最佳的模型。

1.3. 主要研究思路

对不同地区、不同类型的阿片类药物进行描述性分析。通过数据可视化来描述药物使用情况，包括阿片类药物使用的数量和地域趋势。

提取强相关性的社会、经济属性。由于所分析的与阿片类药物使用相关的各种社会经济属性太多，且考虑到过拟合问题和简化模型，因此不适合简单的将所有社会、经济因素加入模型中。通过计算所有变量在各个年份与阿片类药物的相关系数，将2010-2016 年相关系数均大于 0.5 的变量纳入模型中。

将整理后的数据按照 7:3 的比例划分为训练集和测试集，通过机器学习算法建立模型，利用多种机器学习算法来预测阿片类药物使用量。将强相关性的社会、经济属性纳入到不同的机器学习算法中，具体包括如下算法：K-临近算法、支持向量机、决策树、神经网络、随机森林、Logistic 回归。

评价和比较不同算法的优劣。通过不用方法的机器学习模型的预测结果的错误率、F1 评分、支持度、K 折训练集等性能指标评价和比较不同模型的优劣，以期进一步完善机器学习方法在阿片类药物使用量预测分析的统计学模型。

2. 方法

2.1. 数据说明

2.1.1. 数据来源

数据来源于美国国家法医实验室信息系统（https://www.nflis.deadiversion.usdoj.gov）与美国人口统计局（https://www.commerce.gov）。数据包括美国 7 年(2010-2016 年)462 个县的共计 24062 条阿片类药物鉴定计数，以及各年份各县的 704 个一系类常见的社会经济因素。这 462 个县主要来自于这五个州：俄亥俄州、肯塔基州、西弗吉尼亚州、弗吉尼亚州和宾夕法尼亚州。

2.1.2. 分析说明

采用 Office Excel2016 整理数据，通过 Python3.6 软件对数据进行进一步分析，使用到的程序包有 os、zipfile、numpy、pandas、seaborn、matplotlib、sklearn。主要使用到的统计分析方法有数据可视化、K 近邻算法、决策树算法、支持向量机算法、随机

森林算法、神经网络算法、逻辑回归算法。

2.1.3. **数据预处理**

收集到的原始数据包括2010年-2016年共24062条观测数据。其中在2010年、2011年、2012年均为600个特征变量；在2013年为612个特征变量；2014年、2015年、2016年均为612个特征变量。由于数据的复杂性已经部分数据存在缺失等情况，且不同年份的特征变量可能存在不同，不便于直接对数据进行统计分析，因此分析前对数据进行了一定的整理，具体如下：

（1）无效数据处理。删除只在特定年份测量的因素，因为多年的趋势将较少出现，更容易受到潜在的异常值影响；删除所有县数据不完整的因素，由于可能存在由于数据缺失而不明显的隐藏趋势而使模型的效果产生一定程度的误差。 处理之后还剩 19646条观测数据，704个特征变量。

（2）缺失值处理。收集到得数据某些字段值为空得情况很多，一般有三种处理方法：删除记录、数据填补、空值处理。填补缺失值方法有：人工填补、均数填补、中位数或众数填补、多重填补、使用最接近的样本值填补等。不同情况变量的缺失值本研究采用了不同的处理方式[14]。当某一条观测数据缺失项大于总项的 65%时，由于缺失信息较大，故选择删除记录；考虑到不偏离原数据的总体分布，故在删除记录之后用列均值对缺失数据进行填补。处理之后还剩9395条观测数据，704个特征变量。

（3）离散化处理。变量的离散化就是将连续变量的值域划分成为若干个离散的区间，然后用不同符合代表观测值落在每个区间的属性值[15]。由于阿片类数据报告量的极差较大，且为了得到更好的分类预测结果，本次实证研究将各县的“阿片类药物使用报告量”离散化成6个水平：0例[1]、1-9例、10-99例、100-499例、500-999例、1000-4999例、5000例以上。

（4）变量选取处理。由于社会经济属性较多，所以选择与应变量有较强的关联性

[1] 某些州县在2010-2016年间为0例和大于0例的情况都存在，故将0例也划分为1类。

的变量有助于模型的可解释性和精简性。因此通过计算各个自变量与应变量的相关系数来选择强相关属性。考虑到模型的广泛性，选择了 2010 年-2016 年各个年份的社会经济属性与阿片类药物报告量的相关系数均大于 0.5 的所有变量，共有 38 个变量入选。

（5）归一化处理。不同的变量有不同的量纲，比如人口、家庭户数、收入等，数值间的差别比较大，归一化就是为了消除变量之间量纲的影响[17]。归一化就是将数据映射到某一固定区间，一般为（0,1）或（-1,1）。主要是为了数据处理方便提出来的，把有量纲表达式变成无量纲表达式[18]。归一化公式：

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$

考虑到量纲不同问题，故本文对除结局变量外的所有变量均用归一化方法进行处理。处理之后共计还剩 9395 条有效观测数据，38 个特征变量。

（6）数据集划分处理。按照合成类阿片药、非合成类阿片药、半合成阿片药分层（具体药物类别划分方式见 3.1.2 节），将上述处理之后数据按照 7:3 的比例划分为训练集和测试集。得到合成类阿片药的训练集共 2242 条观测，测试集共 961 条观测；非合成类阿片药的训练集共 2092 条观测，测试集共 897 条观测；半合成类阿片药的训练集共 2242 条观测，测试集共 961 条观测。

2.2. 模型方法

2.2.1. K 最近邻法

K 最近邻法（KNN，K-Nearest Neighbor）是数据挖掘分类技术中最简单的方法之一[19]。根据合适的距离函数计算测试集样本中的每一个观测数据和所有的训练集样本的距离，选择与训练集样本距离最小的 K 个样本作为测试集样本的 K 个最近邻，最后根据测试集样本的 K 个最近邻判断测试集样本的类别，通常 K 是不大于 20 的整数。依赖于训练数据集和 K 的取值，输出结果可能会有不同[20]。

本文使用的是 sklearn 包中的 KNeighborsClassifier 方法实现 K 最近邻算法分类，对三种阿片类药物报告量预测做 KNN 分类。

2.2.2. 决策树

决策树（DT，Decision Tree）的算法核心体现着特征变量与结果变量之间的某种映射关系，是一个常见的预测模型。决策树中的节点代表对象，分叉路径代表属性值，叶节点表示从根节点到该节点所经历的路径所表示的对象的值。但是决策树拟合的过程容易出现过拟合现象，目前常用处理过拟合的主要手段便是通过剪枝。虽然通过决策树的剪枝可以降低模型过拟合问题，但是同时也可能会发生模型欠拟合问题。所以必须通过合适的剪枝才能得到效果最优的决策树模型。

本文使用的是 sklearn 包中的 DecisionTreeClassifier 方法实现决策树分类，对三种阿片类药物报告量预测建立决策树模型。

2.2.3. 随机森林

随机森林（RF，Random Forest）是集成学习中的一种较为常用的算法。一个随机森林模型由多颗决策树模型构成，且模型的最终输出结果由每一颗决策树共同决定（常见为个别树输出的类别的众数而定）。如果树的深度越深则更容易学到规则复杂的模型[22]。随机森林本质上是一种较为特殊的 bagging 方法，它将决策树用作 bagging 中的基模型，实际上相当于对于样本和特征都进行了采样，所以随机森林可以很好的避免过拟合问题。同时随机森林由于综合了若干个决策树的基模型，所以在预测效果上由于单个的决策树模型。

在 Python 中本文使用的是 sklearn 包中的 RandomForestClassifier 方法实现随机森林分类，对不同类的阿片类药物报告量预测做随机森林分类模型。

2.2.4. 支持向量机

支持向量机（SVM，Support Vector Machine）是一种二分类的监督学习算法，即可用于分类问题也可以用于回归问题。基本思想是通过求解能够正确划分训练集样本的最佳超平面，以便在不同类的数据点之间进行正确的分类。SVM 算法中有许多不同类型的内核可用于创建这种更高维的空间，例如线性，多项式，Sigmoid 和径向基函数。SVM

的主要优势有如下几点：在高维空间的情况下 SVM 算法处理任然有效；如果维度高于采样数量的情况一下依然有效。SVM 主要有以下缺点：如果特征的数量大大多于采样的数量，为了避免过拟合，合理的正则化变得至关重要；除此之外 SVM 算法并不能直接提供概率估计结果[23]。

在 Python 中本文使用的是 sklearn 包中的 SVM 方法实现支持向量机分类，对不同类的阿片类药物报告量预测做支持向量机模型分类。

2.2.5. 神经网络

人工神经网络（ANNs，Artificial Neural Networks）起源于 1940-1950 年，使通过模仿生物神经的行为特征，形成一种具有学习、联想、记忆和模式识别的人工系统，称为人工神经网络[24]。人工神经网络具有自学习和自适应的能力，可以通过预先提供的一批相互对应的输入输出数据，分析两者的内在关系和规律，最终通过这些规律形成一个复杂的非线性系统函数，这种学习分析过程被称作“训练”。神经元的每一个输入连接都有突触连接强度，用一个连接权值来表示，即将产生的信号通过连接强度放大，每一个输入量都对应有一个相关联的权重[25]。激活函数具有如下性质：可微性，非线性、单调性、输出值于输入值相差不会很大。

在 Python 中本文使用的是 sklearn 包中的 MLPClassifier 方法实现支持人工神经网络分类，对不同类的阿片类药物报告量预测做人工神经网络模型分类。

2.2.6. Logistic 回归

Logistic 回归（LR，Logistic Regression）模型是一种对数概率回归的机器学习算法[26]。作为一种流行病学多元分析方法，被广泛应用于探索二元应变量于影响因素之间关系的研究，例如，疾病诊断、经济预测等[27]。LR 是一种广义的线性模型，因此，需要假设应变量服从伯努利分布，因变量服从高斯分布。Logistic 回归进行预测的步骤如下：首先划分原始数据集为训练集和测试集；然后基于训练集建立逻辑 LR 模型；通过建立的模型计算测试集的预测输出结果；计算损失函数，对比预测结果和实测结果，获

得最佳参数模型。

在 Python 中本文使用的是 sklearn 包中的 LogisticClassifier 方法实现支持逻辑回归分类，对不同类的阿片类药物报告量预测做逻辑回归模型分类。

2.3. 评估方法

2.3.1. 精准率

精准率（Precision）又叫查准率，对于给定测试集的某一类别，分类模型预测正确的比例（混淆矩阵的列维）。主要是针对预测结果而言，在预测为正样本的结果中，我们有多少把握可以预测正确。

$$Precision = \frac{TP}{TP + FP}$$

2.3.2. 召回率

召回率（Recall）又叫查全率，它是针对原样本而言的，在真实值是正例的所有结果中，模型预测正确的比例。召回率越高，代表实际负样本被预测出来的概率越高。

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

2.3.3. F1 评分

由于精确率与召回率通常是此消彼长的，很难兼得，在大规模数据集合中相互制约。而 F1 评分是最常见去同时考虑这两个指标。当 a=1 时，F 值（$F = \frac{2Precision*Recall}{2(Precision+Recall)}$）便是权重因子，代表精确率和召回率的权重相同。

$$F = \frac{(a^2 + 1) * Precision * Recall}{a^2 * (Precision + Recall)}$$

2.3.4. 支持度

支持度（Support）表示同时包含 A 和 B 事务占所有事务的比例。也就是规则前后同时在数据集中出现的比率。

$$\text{Support} = \text{P(A\&B)}$$

2.3.5. K 折交叉验证

K 折交叉验证（K-fold cross-validation），将训练集分割为 K 个样本，随机选择其中

k-1 个样本作为训练集来训练模型，将剩余的 1 个样本作为测试集来验证模型的效果。将这个过程重复 k 次，每个子样本验证一次，将 k 次的结果通过平均的方式综合为一个单一指标进行评估。其中当 k=10 时较为常用，称为 10 次交叉验证。

在 python 中本文使用的是 sklearn 包中的 model_selection 方法实现 k 折交叉验证，对本文所使用得模型进行 10 折交叉验证各个模型的优劣。

3. 结果

3.1. 数据描述

3.1.1. 现状分析

（1）总体阿片类药物的使用情况。为了观察总体数据的分布情况，通过对所有数据的描述性分析得到

图 1 和 图 2。从图中可以看出，报告的阿片类药物共有 56 种，各类阿片类的分布主要集中在海洛因、氢考酮、氢可酮这三类，分别占比为 53.24%、19.79%、8.57%，海洛因的报告量达到 25000 多例，氢考酮的报告量为 10000 多例，氢可酮的报告量为 4000 多例；大部分类型的阿片类药物都小于 0.01%；而较为常见的可卡因、吗啡等占比并没有想象中的那么高，比例分别为 0.73%、1.88%。

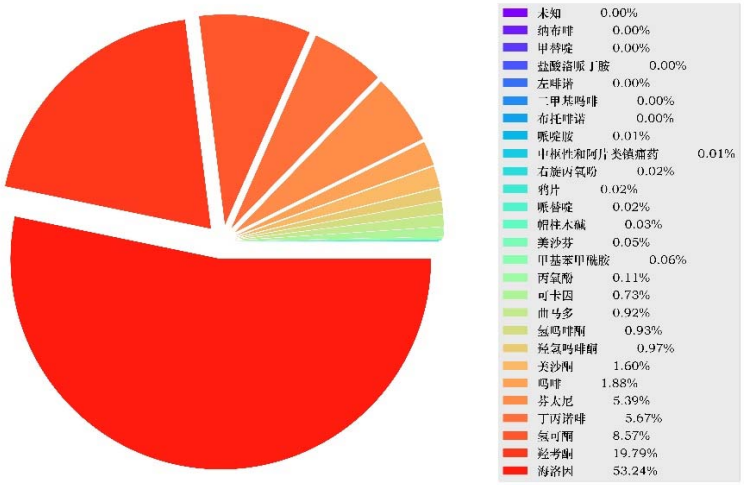

图 1　　各类阿片类药物报告比例饼图

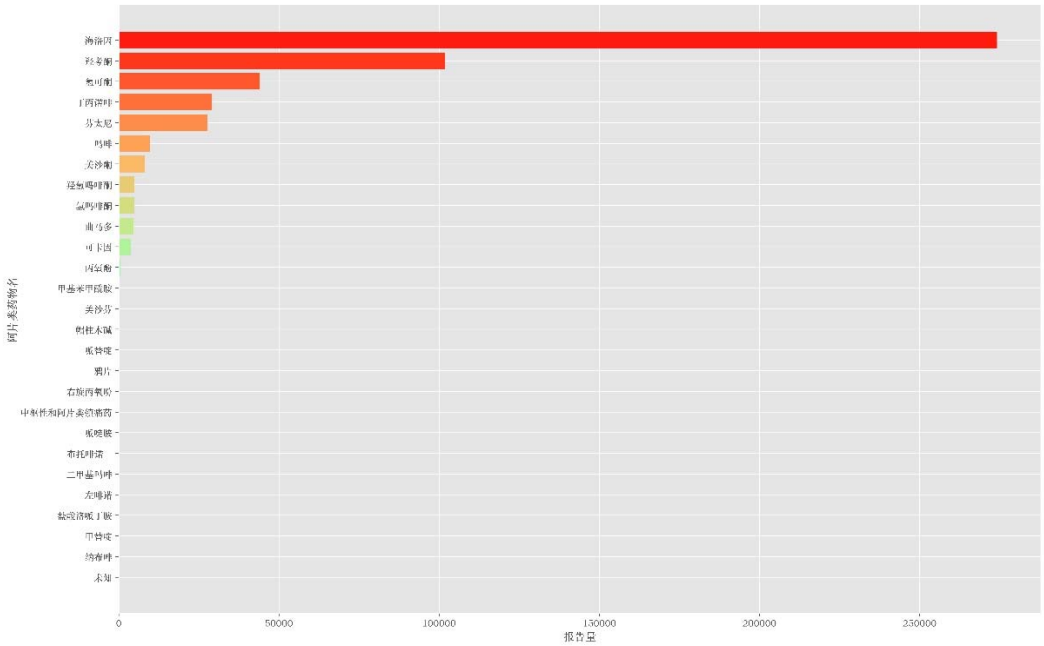

图 2　　各类阿片类药物报告量分布直方图

（2）不同地区阿片类药物的使用情况。在得到总体药物使用的分布情况后，进一步分析不同地区的药物使用情况。根据 2010-2016 年不同州的不同种类报告量绘制了热

力图，见图 3。可以发现，虽然不同年份的报告量有一定的差异，但是在类别上展出来的差异更加明显，肯塔基州主要为氢考酮、海洛因、氢可酮阿片药，俄亥俄州主要为海洛因、氢考酮、丁丙诺啡阿片药，宾夕法尼亚州主要为海洛因、氢考酮、丁丙诺啡阿片药，弗吉尼亚州主要为氢可酮、氢考酮、海洛因阿片药，西弗吉尼亚州主要为氢考酮、氢可酮、海洛因阿片药；可以从图上发现图形模式大致呈现 3 个板块，第一个板块主要以海洛因阿片药为主；第二板块主要以氢考酮为主，第三板块主要以丁丙诺啡为主。基于这三大板块类别便可以将原本 56 种阿片类药物重新划分为三大类。

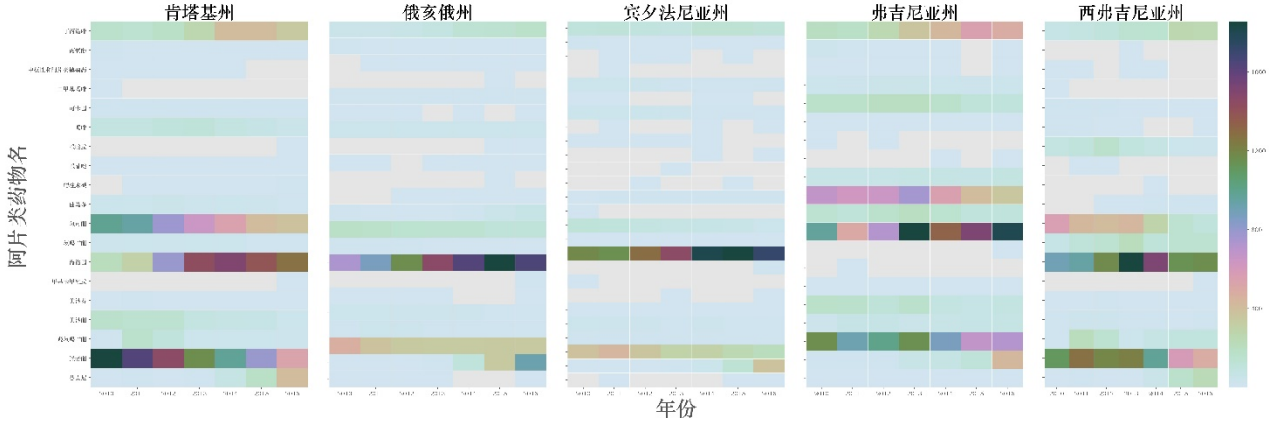

图 3　　各个州各种阿片类药物报告量的热力图

3.1.2. 药物种类划分

根据阿片类药物是否为合成药物将其划分为三大类：合成类阿片药、非合成类阿片药、半合成类阿片药。基于其报告量（见表 1）对 2010-2016 年各个州的报告量绘制折线图，从图 4 可以看出不同州的不同类的阿片药报告量在 2010-2016 年呈现不同趋势，肯塔基州合成阿片药在 2010 年报告量为峰值 6208 例，半合成类阿片药在 2014 年达到峰值 7151 例；俄亥俄州半合成阿片药在 2015 年达到峰值 26674 例，而合成阿片药报告量在 2013 年最低为 5060 例，2016 年最高为 17347 例；宾夕法尼亚州的半合成阿片药的报告量大致也呈现一个先增后减的趋势，在 2015 年达到峰值 20799 例；弗吉尼亚州的半合成阿片药在 2011 年和 2013 年分别出现一个波谷和波峰；西弗吉尼亚州的半合成阿片药在 2013 年出现一个波峰，合成类阿片药在整体上呈现一个递减趋势；非合成阿片药变化趋势不明显；而这五个州的总体上而言，半合成阿片类药物呈现一个先增后

减的趋势，而合成类阿片药为先减后增的趋势，非合成类阿片药变化趋势相对不明显。

可以发现，五个州阿片药物使用量趋势近似相同，合成类阿片类药物的变化趋势与半合成类阿片类药物的变化趋势相反，非合成类阿片药变化趋势不大。

表 1　　2010-2016 年五大州的 3 类阿片类药物报告量[2]

药物种类	年份	肯塔基州	俄亥俄州	宾夕法尼亚州	弗吉尼亚州	弗吉尼亚州
半合成阿片类药物	2010 年	3952	12153	14450	4381	1535
	2011 年	4068	13775	14100	3092	1473
	2012 年	5132	17244	15019	3862	1689
	2013 年	6809	21238	16437	7062	2444
	2014 年	7151	23914	20493	5460	1942
	2015 年	6383	26674	20799	5842	1535
	2016 年	5829	24561	19155	6271	1528
合成阿片类药物						
	2010 年	6208	6928	4689	3811	1249
	2011 年	5917	5999	5207	3189	1672
	2012 年	5224	5284	4383	3390	1521
	2013 年	3961	5060	3537	4004	1500
	2014 年	3628	6391	4030	3140	1251
	2015 年	3266	9843	4474	2613	994
	2016 年	3089	17347	6654	3596	988
非合成阿片类药物						
	2010 年	293	626	675	493	106
	2011 年	299	551	680	466	126
	2012 年	366	615	557	579	166
	2013 年	378	547	435	609	102
	2014 年	302	555	381	437	87
	2015 年	216	610	378	355	42
	2016 年	175	562	355	328	31

[2] 数据通过汇总原始数据得来，本研究只用在绘图；

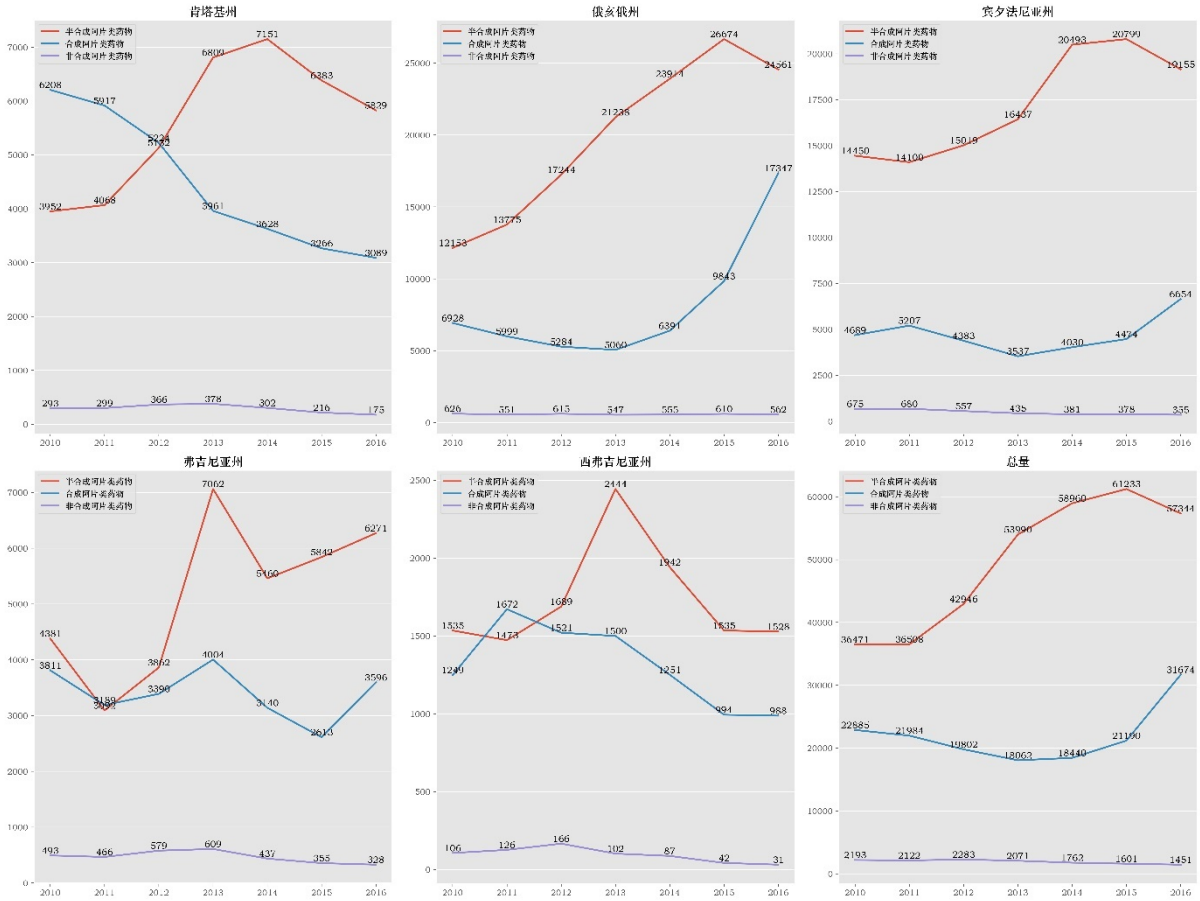

图 4　　2010-2016 年五大州的 3 类阿片类药物报告量折线图

由探索性分析可知，五个州在不同的阿片类药物有着近似相同的变化趋势，半合成阿片类药物呈现一个先增后减的趋势，而合成类阿片药为先减后增的趋势，非合成类阿片药变化趋势相对不明显。因此，为了简化模型和提升模型的泛化能力，将本研究所使用到的机器学习模型应用到测试集的不同类别阿片类药物报告量预测，具体的各个影响因素详见附录 2，通过不同模型的预测结果评估模型优劣。

3.2. 应用模型

将 K 近邻模型、决策树、随机森林、支持向量机、神经网络、Logistic 回归 6 种模型应用于合成类阿片药，得到表 2 所示的混淆矩阵。

表 2　　6 种模型预测合成阿片类药物的混淆矩阵

模型方法	测试集实际分类	预测结果					
		0 人	1-9 人	10-99 人	100-499 人	500-999 人	1000-4999 人
K 近邻法							
	0 人	31	51	15	4	0	0
	1-9 人	44	150	103	4	0	0

	10-99 人	12	84	354	24	0	0
	100-499 人	0	5	40	37	0	0
	500-999 人	0	0	0	1	1	0
	1000-4999 人	0	0	0	1	0	0
决策树							
	0 人	40	46	13	2	0	0
	1-9 人	49	140	106	6	0	0
	10-99 人	24	99	311	40	0	0
	100-499 人	3	3	31	42	3	0
	500-999 人	0	0	0	0	2	0
	1000-4999 人	0	0	0	0	1	0
随机森林							
	0 人	35	43	23	0	0	0
	1-9 人	32	180	88	1	0	0
	10-99 人	5	95	367	7	0	0
	100-499 人	0	3	38	39	2	0
	500-999 人	0	0	0	0	2	0
	1000-4999 人	0	0	0	0	1	0
支持向量机							
	0 人	0	73	27	1	0	0
	1-9 人	0	178	123	0	0	0
	10-99 人	0	80	388	6	0	0
	100-499 人	0	2	60	20	0	0
	500-999 人	0	0	0	1	1	0
	1000-4999 人	0	0	0	1	0	0
神经网络							
	0 人	22	52	23	4	0	0
	1-9 人	22	160	115	4	0	0
	10-99 人	2	71	379	22	0	0
	100-499 人	0	1	43	38	0	0
	500-999 人	0	0	0	2	0	0
	1000-4999 人	0	0	0	1	0	0
Logistic 回归							
	0 人	0	65	35	1	0	0
	1-9 人	1	173	126	1	0	0
	10-99 人	0	68	398	8	0	0
	100-499 人	0	1	57	24	0	0
	500-999 人	0	0	0	0	2	0
	1000-4999 人	0	0	0	1	0	0

将 K 近邻模型、决策树、随机森林、支持向量机、神经网络、Logistic 回归 6 种模型应用于半合成类阿片药，得到表 3 所示的混淆矩阵。

表 3　6 种模型预测半合成阿片类药物的混淆矩阵

模型方法	实际	预测					
		0 人	1-9 人	10-99 人	100-499 人	500-999 人	1000-4999 人
K 近邻法							
	0 人	22	38	13	1	0	0
	1-9 人	23	127	91	4	0	0
	10-99 人	9	71	331	39	0	0
	100-499 人	1	5	55	91	4	0
	500-999 人	0	0	1	14	5	0
	1000-4999 人	0	0	0	4	3	9
决策树							
	0 人	20	34	18	2	0	0
	1-9 人	29	128	84	4	0	0
	10-99 人	14	93	299	43	1	0
	100-499 人	5	5	44	97	5	0
	500-999 人	0	0	2	7	11	0
	1000-4999 人	0	0	0	1	2	13
随机森林							
	0 人	23	38	10	3	0	0
	1-9 人	24	139	80	2	0	0
	10-99 人	7	85	328	30	0	0
	100-499 人	1	3	55	93	4	0
	500-999 人	0	0	0	14	4	2
	1000-4999 人	0	0	1	1	2	12
支持向量机							
	0 人	0	57	14	3	0	0
	1-9 人	0	150	92	3	0	0
	10-99 人	0	77	339	34	0	0
	100-499 人	0	1	74	80	1	0
	500-999 人	0	0	1	17	1	1
	1000-4999 人	0	0	0	4	1	11
神经网络							
	0 人	0	56	16	2	0	0
	1-9 人	1	143	98	3	0	0
	10-99 人	4	81	319	45	0	1
	100-499 人	1	1	59	95	0	0
	500-999 人	1	0	0	15	0	4
	1000-4999 人	0	0	0	6	0	10
Logistic 回归							
	0 人	0	46	26	2	0	0
	1-9 人	0	90	153	2	0	0
	10-99 人	0	32	389	29	0	0

100-499 人	0	1	87	67	1	0
500-999 人	0	0	1	18	0	1
1000-4999 人	0	0	0	4	0	12

将 K 近邻模型、决策树、随机森林、支持向量机、神经网络、Logistic 回归 6 种模型应用于半合成类阿片药，得到表 4 所示的混淆矩阵。

表 4　6 种模型预测半合成阿片类药物的混淆矩阵

模型方法	实际	预测					
		0 人	1-9 人	10-99 人	100-499 人	500-999 人	1000-4999 人
K 近邻法							
	0 人	132	149	6	0	0	0
	1-9 人	112	376	19	0	0	0
	10-99 人	7	54	40	0	0	0
	100-499 人	0	0	2	0	0	0
	500-999 人	0	0	0	0	0	0
	1000-4999 人	0	0	0	0	0	0
决策树							
	0 人	138	131	18	0	0	0
	1-9 人	150	301	56	0	0	0
	10-99 人	13	37	49	2	0	0
	100-499 人	0	0	0	2	0	0
	500-999 人	0	0	0	0	0	0
	1000-4999 人	0	0	0	0	0	0
随机森林							
	0 人	157	126	4	0	0	0
	1-9 人	132	366	9	0	0	0
	10-99 人	4	55	42	0	0	0
	100-499 人	0	0	0	2	0	0
	500-999 人	0	0	0	0	0	0
	1000-4999 人	0	0	0	0	0	0
支持向量机							
	0 人	65	222	0	0	0	0
	1-9 人	31	476	0	0	0	0
	10-99 人	0	70	31	0	0	0
	100-499 人	0	0	1	1	0	0
	500-999 人	0	0	0	0	0	0
	1000-4999 人	0	0	0	0	0	0
神经网络							
	0 人	131	156	0	0	0	0
	1-9 人	84	419	4	0	0	0
	10-99 人	3	64	34	0	0	0

	100-499 人	0	0	0	2	0	0
	500-999 人	0	0	0	0	0	0
	1000-4999 人	0	0	0	0	0	0
Logistic 回归							
	0 人	104	181	2	0	0	0
	1-9 人	48	453	6	0	0	0
	10-99 人	2	62	37	0	0	0
	100-499 人	0	0	0	2	0	0
	500-999 人	0	0	0	0	0	0
	1000-4999 人	0	0	0	0	0	0

3.3. 模型评估

各个模型的效果采用精确度、召回率、F1-评分、支持度四个指标进行评估。具体结果详见表 5。可以发现，KNN 模型在预测三类阿片类药物的使用量方面的预测性能都一般。决策树模型在预测三类阿片类药物的使用量方面的预测性能都一般。随机森林模型在预测三类阿片类药物的使用量方面的预测性能都一般。支持向量机模型在预测半合成阿片类药物的结果上一般，在预测合成阿片类药物的结果上良好，在预测非合成阿片类药物的结果上较优。神经网络模型在预测合成阿片类药物和半合成阿片类药物的结果上一般，在预测非合成阿片类药物的结果上良好。Logistic 模型在预测三类阿片类药物的结果上都表现良好。

表 5　　3 类阿片类药物报告量的 6 种机器学习模型结果

机器学习算法	药物报告量	合成阿片类				非合成阿片类				半合成阿片类			
		精确率	召回率	F1-评分	支持度	精确率	召回率	F1-评分	支持度	精确率	召回率	F1-评分	支持度
最近邻法													
	0 人	0.31	0.36	0.33	87	0.46	0.53	0.49	251	0.30	0.40	0.34	55
	1-9 人	0.50	0.52	0.51	290	0.74	0.65	0.69	579	0.52	0.53	0.52	241
	10-99 人	0.75	0.69	0.72	512	0.40	0.60	0.48	67	0.74	0.67	0.70	491
	100-499 人	0.45	0.52	0.48	71	0.00	0.00	0.00	0	0.58	0.59	0.59	153
	500-999 人	0.50	1.00	0.67	1	0.00	0.00	0.00	0	0.25	0.42	0.31	12
	1000-4999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.56	1.00	0.72	9
	总体	0.61	0.60	0.60	961	0.64	0.61	0.62	897	0.62	0.61	0.61	961
决策树													
	0 人	0.41	0.41	0.41	101	0.48	0.45	0.47	302	0.23	0.23	0.23	73
	1-9 人	0.47	0.48	0.47	291	0.59	0.62	0.61	480	0.49	0.48	0.48	250

	10-99 人	0.66	0.66	0.66	470	0.47	0.42	0.44	113	0.66	0.67	0.66	442
	100-499 人	0.52	0.46	0.49	94	1.00	1.00	1.00	2	0.63	0.60	0.62	164
	500-999 人	1.00	0.40	0.57	5	0.00	0.00	0.00	0	0.45	0.47	0.46	19
	1000-4999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.75	0.92	0.83	13
	总体	0.56	0.56	0.56	961	0.54	0.54	0.54	897	0.57	0.57	0.57	961
随机森林													
	0 人	0.32	0.49	0.39	65	0.49	0.52	0.51	271	0.26	0.32	0.28	60
	1-9 人	0.57	0.52	0.54	332	0.73	0.66	0.69	561	0.54	0.54	0.54	248
	10-99 人	0.73	0.70	0.72	499	0.44	0.70	0.54	63	0.75	0.68	0.72	494
	100-499 人	0.52	0.70	0.60	61	1.00	1.00	1.00	2	0.58	0.67	0.62	134
	500-999 人	0.50	0.25	0.33	4	0.00	0.00	0.00	0	0.20	0.33	0.25	12
	1000-4999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.81	1.00	0.90	13
	总体	0.63	0.62	0.62	961	0.64	0.62	0.63	897	0.64	0.62	0.63	961
支持向量机													
	0 人	0.00	0.00	0.00	0	0.23	0.68	0.34	96	0.00	0.00	0.00	0
	1-9 人	0.59	0.53	0.56	333	0.94	0.62	0.75	768	0.61	0.53	0.57	285
	10-99 人	0.82	0.65	0.72	598	0.31	0.97	0.47	32	0.75	0.65	0.70	520
	100-499 人	0.24	0.69	0.36	29	0.50	1.00	0.67	1	0.51	0.57	0.54	141
	500-999 人	0.50	1.00	0.67	1	0.00	0.00	0.00	0	0.05	0.33	0.09	3
	1000-4999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.69	0.92	0.79	12
	总体	0.72	0.61	0.66	961	0.84	0.64	0.69	897	0.67	0.60	0.64	961
神经网络													
	0 人	0.22	0.48	0.30	46	0.46	0.60	0.52	218	0.00	0.00	0.00	7
	1-9 人	0.53	0.56	0.55	284	0.83	0.66	0.73	639	0.58	0.51	0.54	281
	10-99 人	0.80	0.68	0.73	560	0.34	0.89	0.49	38	0.71	0.65	0.68	492
	100-499 人	0.46	0.54	0.50	71	1.00	1.00	1.00	2	0.61	0.57	0.59	166
	500-999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0
	1000-4999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.62	0.67	0.65	15
	总体	0.67	0.62	0.64	961	0.72	0.65	0.67	897	0.65	0.59	0.62	961
Logistic 回归													
	0 人	0.00	0.00	0.00	1	0.36	0.68	0.47	154	0.00	0.00	0.00	0
	1-9 人	0.57	0.56	0.57	307	0.89	0.65	0.75	696	0.37	0.53	0.43	169

10-99 人	0.84	0.65	0.73	616	0.37	0.82	0.51	45	0.86	0.59	0.70	656
100-499 人	0.29	0.69	0.41	35	1.00	1.00	1.00	2	0.43	0.55	0.48	122
500-999 人	1.00	1.00	1.00	2	0.00	0.00	0.00	0	0.00	0.00	0.00	1
1000-4999 人	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.75	0.92	0.83	13
总体	0.73	0.62	0.67	961	0.78	0.66	0.69	897	0.72	0.58	0.63	961

通过上述四个指标进行具体的模型效果评价之后，并用 10 折交叉验证法作为参考对模型效果进行进一步评估，为了便于观察对数据结果绘制箱式图，结果如图 5 所示。

对于合成阿片类药物的预测模型而言，支持向量机模型的预测精确度从均值和变异程度而言效果最佳，其次是逻辑模型，虽然在精确度均值上神经网络模型略微由于逻辑回归模型，但是总体的变异而言逻辑回归表现更好。由于决策树模型在过拟合和欠拟合问题上的不稳定性，其预测效果表现最差，而集成了多个决策树基模型的随机森林模型的预测精确度明显优于决策树。K 近邻算法作为最简单的机器学习模型，预测效果表现也算中规中矩，优于决策树和随机森林但是劣于人工神经网络。

对于半合成阿片类药物的预测模型而言，支持向量机模型的预测精确度无论是从均值和变异程度而言效果依然表现最佳，且变异程度相比于合成类更低。其次是逻辑模型和神经网络模型。决策树、随机森林、K 近邻模型相比于合成类的变异程度明显大幅度增大。说明模型预测效果并不稳定。

对于非合成阿片类药物的预测模型而言，仍然是支持向量机模型的预测精确度综合表现最佳，逻辑回归、神经网络的表现次之。而 6 种模型的变异程度相对于另外两类阿片类药物都明显更低，说明各个模型对于非合成类药物的预测效果最具有稳定性，预测结果最可靠。

综上而言， K 近邻法、决策树、随机森林这三个机器学习模型在于传统逻辑回归模型对比不占明显优势。因此本文以 10 折交叉验证方法为评估指标得出最优的模型为支持向量机模型，其次为逻辑回归和神经网络模型，最差为决策树模型。

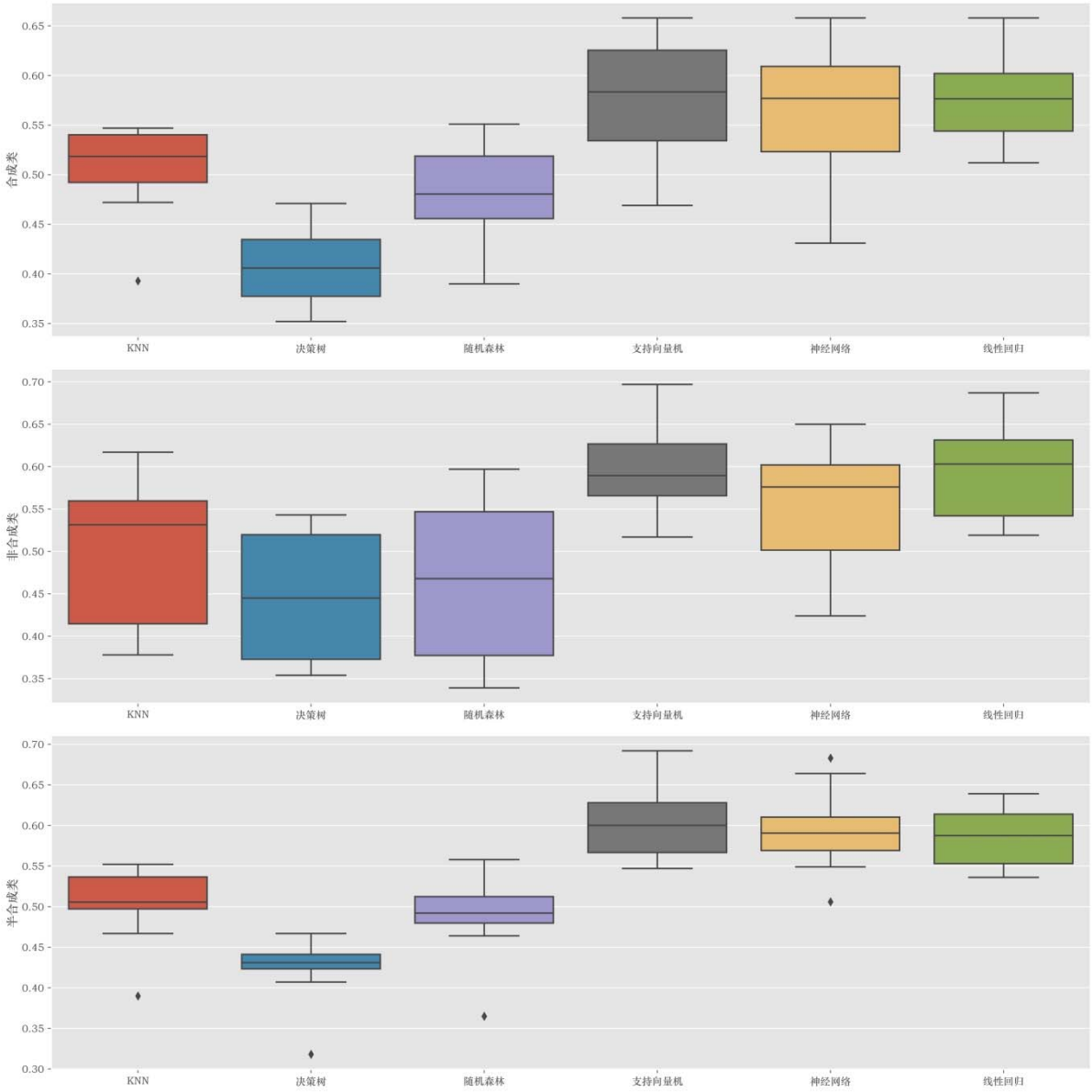

图 5　　6 种机器学习算法的 K 折验证结果箱式图

结论

支持向量机模型和人工神经网络模型在预测效果中的性能都比较理想。支持向量机模型的精确度>0.8，而人工神经网络模型的精确度>0.7。对比与传统的预测模型而言，支持向量机和神经网络模型容易调整参数，能够生成预测效能更好的模型。

对于决策树和随机森林模型，由于随机森林模型本质上是在以决策树为基学习器构建 Bagging 集成的基础上，进一步决策树训练过程中引入了随机属性选择。所以本文也

证实其理论性，发现随机森林模型的预测效能要明显优于决策树模型。随机森林模型能够集成多颗决策树进行预测，客服了单颗决策树泛化能力不足的缺点，有助于模型的外推。同时，集成学习模型也容易调整参数，可以生成预测效能更好的模型，为类似研究提供新思路和方法。KNN 模型作为最简单的机器学习模型之一，在方法思路上具有简洁的优点，但是其实现过程所需计算机空间内存较大，且本文的实证研究发现其预测效能一般。但是相比于传统的统计预测模型而言，容易调整参数和思路简洁是一个明显的优势。传统的 logistics 回归在本实证研究中的预测效能表现虽然不是最佳，但是明显优于决策树和随机森林等模型。但是传统的 logistics 模型的多重共线性和变量间相互关系作用的缺点。

本文从各类阿片类药物使用情况的研究背景出发，分析了不同年份和不同地区阿片类药物报告量的现状，在公共卫生事业越来越重要的情况下，成瘾性药物的研究更应该引起学者们的注意。对阿片类药物使用情况的预测不仅可以让卫生管理人员直观看到社会情况，还可以帮助相关政府部门制定必要的卫生政策。本研究主要工作如下：不同地区阿片类药物报告量、相关社会经济指标数据的采集和清洗；利用数据可视化对阿片类药物报告量的描述性统计分析；相关社会经济指标数据的特征选择；不同类型的阿片类药物报告量的机器学习建模。总结来说，本文能将多种机器学习算法运用到阿片类药物报告量的预测中，更具机器学习模型在阿片类药物报告量数据上的表现，做到分类处理的精准预测。

虽然本文对不同类的阿片类药物报告量预测做到了不错的效果，但是本文的工作还有很多的不足之处，其中包括以下几个方面：虽然收集到 2017 年阿片类药物的报告量，但并没有收集到当年分的社会经济指标数据，导致大量数据失效；且在数据清洗过程中清洗掉大量数据，最终使用到的有效数据只有 9000 多条，相对于机器学习模型而言数据量偏少；社会经济指标未能构造多维，本文可以通过已有变量构造其他变量；使用的算法模型不够新。本文使用到的机器学习算法都是比较传统的机器学习算法，这些算法在一定程度上还有可以优化的空间。比如梯度提升回归树算法、CART 决策树和 C4.5 决策树算法。

参考文献

[1]. 林景怀, 杨明娜与韩凤, 医院住院患者阿片类药物的应用情况. 中国药物经济学, 2019. 14(11): 第32-34+41页.

[2]. 传植, 揭开阿片类药物成瘾之谜. 世界科学, 2019(11): 第13-14页.

[3]. 田野等, 我国2014-2016年阿片类药物使用情况分析. 中国药房, 2019. 30(09): 第1153-1157页.

[4]. 袁莎, 美国阿片类药物危机及中美禁毒合作. 和平与发展, 2019(01): 第101-115+135-136页.

[5]. D, L.等, 家庭医生对自身管理阿片类药物危机的看法. 中国全科医学, 2019. 22(31): 第3804页.

[6]. 邓硕曾北京中医药大学东方医院麻醉科教授, 阿片类药物“双刃”作用能否解, in 健康报. 第 008页.

[7]. 张景奇, 史文宝与纪秀娟, 机器学习在医疗和公共卫生中应用. 中国公共卫生, 2019. 35(10): 第1449-1452页.

[8]. 何巍, 基于机器学习的犯罪预测综述. 科学技术与工程, 2019. 19(36): 第37-43页.

[9]. 程豪, 大数据背景下缺失数据问题及对策. 中国统计, 2019(10): 第72-74页.

[10]. 刘春亚, 基于粗集理论的数据预处理及应用研究, 2003, 重庆大学.

[11]. 尚绍环, 不同批样本归一化处理方法. 电子产品可靠性与环境试验, 2005(04): 第34-35页.

[12]. 高军. 性能退化数据的归一化处理方法. in 中国电子学会可靠性分会第十四届学术年会. 2008. 中国海南海口.

[13]. 窦小凡, KNN算法综述. 通讯世界, 2018(10): 第273-274页.

[14]. 邵珊珊, 基于KNN的分类方法及其应用研究, 2019, 燕山大学.

[15]. 孙明喆, 毕瑶家与孙驰, 改进随机森林算法综述. 现代信息科技, 2019. 3(20): 第28-30页.

[16]. 朱悦, 吴建华与方颖, SVM在冠心病分类预测中的应用研究. 生物医学工程学杂志, 2013. 30(06): 第1180-1185页.

[17]. 许兴阳, 基于BP人工神经网络的医院药库管理系统药品预测模型研究. 科技创新导报, 2008(20): 第169-170页.

[18]. 周济民, 基于神经网络改进的元胞自动机分析——美国阿片类药物滥用情况. 信息系统工程, 2019(11): 第144-145+147页.

[19]. 王凤竹, 稳健的Logistic回归及其应用, 2014, 华北电力大学.

[20]. 黄锦联, 二分类Logistic回归模型的Lq似然估计, 2013, 广西师范大学.

附录

附录 1　2010-2016 年各属性的相关系数

变量名	2010 年相关系数	2011 年相关系数	2012 年相关系数	2013 年相关系数	2014 年相关系数	2015 年相关系数	2016 年相关系数	合计相关系数	均值
HC01_VC11	0.605	0.623	0.616	0.593	0.585	0.621	0.663	0.550	0.615
HC01_VC70	0.615	0.628	0.624	0.589	0.580	0.608	0.641	0.606	0.612
HC01_VC12	0.591	0.609	0.608	0.587	0.591	0.615	0.656	0.558	0.608
HC01_VC13	0.584	0.604	0.601	0.590	0.595	0.619	0.661	0.591	0.608
HC01_VC10	0.571	0.598	0.587	0.595	0.588	0.623	0.663	0.545	0.603
HC01_VC71	0.572	0.588	0.590	0.596	0.582	0.617	0.648	0.596	0.599
HC01_VC66	0.609	0.598	0.604	0.580	0.572	0.595	0.618	0.592	0.597
HC01_VC67	0.589	0.610	0.610	0.575	0.570	0.583	0.634	0.503	0.596
HC01_VC43	0.611	0.628	0.609	0.553	0.563	0.578	0.612	0.552	0.594
HC01_VC65	0.600	0.611	0.607	0.568	0.558	0.582	0.622	0.570	0.593
HC01_VC09	0.581	0.609	0.596	0.566	0.567	0.583	0.617	0.506	0.589
HC01_VC39	0.588	0.607	0.593	0.572	0.554	0.585	0.617	0.553	0.588
HC01_VC62	0.603	0.616	0.614	0.549	0.543	0.572	0.597	0.553	0.585
HC01_VC36	0.595	0.614	0.598	0.545	0.556	0.569	0.602	0.550	0.583
HC01_VC131	0.567	0.591	0.584	0.556	0.569	0.583	0.618	0.581	0.581
HC01_VC29	0.603	0.619	0.587	0.540	0.549	0.565	0.596	0.516	0.580
HC01_VC52	0.611	0.633	0.617	0.531	0.522	0.553	0.588	0.549	0.579
HC01_VC47	0.527	0.559	0.565	0.570	0.584	0.595	0.632	0.560	0.576
HC01_V	0.550	0.575	0.571	0.558	0.570	0.584	0.621	0.571	0.5

C03									76
HC03_VC03	0.550	0.575	0.571	0.558	0.570	0.584	0.621	0.571	0.576
HC01_VC18	0.554	0.574	0.572	0.554	0.572	0.579	0.616	0.572	0.574
HC01_VC31	0.553	0.586	0.579	0.557	0.546	0.576	0.607	0.549	0.572
HC01_VC130	0.552	0.577	0.571	0.547	0.557	0.572	0.606	0.567	0.569
HC01_VC88	0.524	0.554	0.560	0.555	0.573	0.590	0.622	0.571	0.568
HC01_VC26	0.550	0.575	0.571	0.546	0.556	0.571	0.605	0.549	0.568
HC01_VC30	0.563	0.585	0.569	0.547	0.529	0.569	0.596	0.564	0.565
HC01_VC80	0.564	0.584	0.568	0.535	0.546	0.559	0.593	0.520	0.564
HC01_VC89	0.527	0.548	0.545	0.555	0.566	0.578	0.619	0.547	0.563
HC01_VC119	0.527	0.552	0.557	0.547	0.557	0.572	0.606	0.502	0.560
HC01_VC79	0.544	0.565	0.559	0.530	0.541	0.556	0.589	0.549	0.555
HC01_VC77	0.521	0.555	0.549	0.532	0.542	0.562	0.592	0.552	0.550
HC01_VC04	0.521	0.548	0.544	0.530	0.546	0.554	0.588	0.542	0.547
HC01_VC17	0.522	0.549	0.543	0.529	0.540	0.552	0.586	0.539	0.546
HC02_VC03	0.504	0.525	0.534	0.519	0.526	0.519	0.583	0.516	0.530
HC02_VC29	0.523	0.532	0.513	0.517	0.510	0.534	0.567	0.523	0.528
HC02_VC13	0.506	0.535	0.515	0.529	0.520	0.511	0.561	0.514	0.525
HC02_VC04	0.509	0.531	0.515	0.512	0.509	0.516	0.560	0.513	0.522
HC02_VC17	0.509	0.536	0.514	0.503	0.502	0.522	0.558	0.512	0.521

附录 2　变量标签解释

变量	标签
HC01_VC03	按类型划分的住户总数
HC01_VC04	按类型划分的住户---家庭住户(家庭)
HC01_VC09	家庭类型---家庭(家庭)---男性户主，无妻子，家庭
HC01_VC10	按类型划分的家庭---家庭(家庭)---男户主，无妻子，家庭---有 18 岁以下的子女
HC01_VC11	按类型划分的住户---家庭住户(家庭)---女户主，无丈夫，家庭
HC01_VC119	居住一年前---不同的房子在美国
HC01_VC12	按类型划分的住户---家庭住户(家庭)---女户主，没有丈夫在场，家庭---子女不足 18 岁
HC01_VC13	按类型划分的家庭---非家庭家庭
HC01_VC130	出生地点---本地---出生在美国
HC01_VC131	出生地点---在美国本土出生---居住状态
HC01_VC17	按类型划分的住户---名或多于一名十八岁以下的住户
HC01_VC18	按类型划分的住户---65 岁或以上的一名或多名住户
HC01_VC26	关系---户主
HC01_VC29	关系---其他亲属
HC01_VC30	关系---邻居竞争
HC01_VC31	关系---非亲属---未婚伴侣
HC01_VC36	婚姻状况---未婚
HC01_VC39	婚姻状况---丧偶
HC01_VC43	婚姻状况---未婚
HC01_VC47	婚姻状况---离婚
HC01_VC52	生育能力---未婚女性(丧偶、离婚、未婚)
HC01_VC62	祖父母---对孙子孙女负责
HC01_VC65	祖父母---负责孙辈---年负责孙辈---1 年或 2 年
HC01_VC66	祖父母---负责孙子---年负责孙子---3 或 4 年
HC01_VC67	祖父母---负责孙子---年负责孙子---5 年或以上
HC01_VC70	祖父母---他们是女性
HC01_VC71	祖父母---已婚
HC01_VC77	入学---幼稚园
HC01_VC79	入学---高中(9---12 年级)
HC01_VC80	入学---大学或研究生院
HC01_VC88	教育程度---有些大学，没有学位
HC01_VC89	教育程度---副学士学位
HC02_VC03	按类型划分的住户总数
HC02_VC04	按类型划分的住户---家庭住户(家庭)
HC02_VC13	按类型划分的家庭---非家庭家庭
HC02_VC17	按类型划分的住户---名或多于一名十八岁以下的住户
HC02_VC29	关系---其他亲属
HC03_VC03	按类型划分的住户总数
HC03_VC03	按类型划分的住户总数

致谢

弹指一挥间，大学四年的时光即将结束。在重医的四年学习生活是我人生生涯中最美好的记忆。四年的艰苦跋涉，五个月的精心准备，毕业论文终于到了画句号的时候，心头照理该如释重负，但写作过程中常常出现的辗转反侧和力不从心之感挥之不去。论文写作的过程并不轻松。敲完最后一个字符，重新从头细细阅读早已不陌生的文字，我感触颇多。虽然没有什么值得显耀的成果，但对我而言已是宝贵的财富，是无数教诲、关爱和帮助的结果。

我要感谢我的指导老师曾庆老师。曾老师虽然身负教学、科学重任，任抽出时间，指导我们的论文撰写，耳提面命。同时，我要感谢公共卫生与管理学院所有给我授过课的老师，是他们传授给我方方面面的知识，拓宽了我的知识面，培养了我的专业功底，对论文的完成不无裨益。

除此之外感谢我的同学们，从五湖四海来到这个陌生的城市里，是你们和我共同组成一个集体维系着班级体的一个融洽关系。

在论文完成之际，我的心情无法平静，从开始进入课题到论文的顺利完成有多少可敬的师长、同学、朋友给了我无言的帮助，在这里接受我诚挚的谢意！