机器学习模型在阿片类药物使用量预测应用中的评估研究

——关于美国五个州的阿片类药物使用量的实证分析

关键词：机器学习 药物使用率预测 神经网络 SVM 贝叶斯 决策树 随机森林

目录

[1. 前言（绪论） 3](#_Toc32737278)

[1.1. 研究背景和意义 3](#_Toc32737279)

[1.1.1. 研究背景 3](#_Toc32737280)

[1.1.2. 研究意义 4](#_Toc32737281)

[1.2. 国内外研究现状 4](#_Toc32737282)

[1.3. 主要研究思路 5](#_Toc32737283)

[1.3.1. 对不同地区、不同类型的阿片类药物进行描述性分析 5](#_Toc32737284)

[1.3.2. 基于时间序列分析的基本理论建立模型 5](#_Toc32737285)

[1.3.3. 提取强相关性的社会经济属性 5](#_Toc32737286)

[1.3.4. 通过机器学习算法建立模型 5](#_Toc32737287)

[1.3.5. 评价和比较不同算法的优劣 5](#_Toc32737288)

[1.4. 研究方法与内容 6](#_Toc32737289)

[1.4.1. 研究方法 6](#_Toc32737290)

[1.4.2. 研究内容 6](#_Toc32737291)

[1.5. 数据来源与说明 6](#_Toc32737292)

[1.5.1. 数据来源 6](#_Toc32737293)

[1.6. 数据说明 7](#_Toc32737294)

[2. 模型建立准备 7](#_Toc32737295)

[2.1. 现状分析 7](#_Toc32737296)

[2.1.1. 不同类型阿片类药物的使用情况 7](#_Toc32737297)

[2.1.2. 不同地区阿片类药物的使用情况 8](#_Toc32737298)

[2.2. 时间序列趋势 9](#_Toc32737299)

[2.3. 特征属性选择 10](#_Toc32737300)

[3. 机器学习预测模型 10](#_Toc32737301)

[3.1. 神经网络模型 10](#_Toc32737302)

[3.1.1. 方法 10](#_Toc32737303)

[3.1.2. 结果 10](#_Toc32737304)

[3.2. 随机森林模型 10](#_Toc32737305)

[3.2.1. 方法 11](#_Toc32737306)

[3.2.2. 结果 11](#_Toc32737307)

[3.3. 朴素贝叶斯模型 12](#_Toc32737308)

[3.3.1. 方法 12](#_Toc32737309)

[3.3.2. 结果 12](#_Toc32737310)

[3.4. 支持向量机模型 12](#_Toc32737311)

[3.4.1. 方法 12](#_Toc32737312)

[3.4.2. 结果 12](#_Toc32737313)

[3.5. 集成学习模型 13](#_Toc32737314)

[3.5.1. 方法 13](#_Toc32737315)

[3.5.2. 结果 13](#_Toc32737316)

[4. 模型评估 15](#_Toc32737317)

[4.1. 各种算法对比 15](#_Toc32737318)

[4.2. 最优算法的选择 错误!未定义书签。](#_Toc32737319)

[4.3. 模型验证 16](#_Toc32737320)

[5. 结论（讨论） 18](#_Toc32737321)

[5.1. 模型比较结论 18](#_Toc32737322)

[5.2. 社会经济因素结论 18](#_Toc32737323)

[5.3. 政策启示结论 18](#_Toc32737324)

1. 前言（绪论）
   1. 研究背景和意义
      1. 研究背景

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%。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。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。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。

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。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。

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，包括我国、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。

* + 1. 研究意义

在过去二十年中，人类收集，存储，处理数据的能力取得了飞速提升，人类社会的各个角落都积累了大量数据，急需能够有效地对数据进行分析利用的计算机算法，而机器学习恰能顺应时代的迫切需求，因此机器学习这门学科领域很自然地取得巨大发展，受到了广泛的关注。今天，机器学习已经与普通人的生活密切相关。例如在天气预报、能源勘探、环境监测等方面，有效地利用机器学习技术对卫星和传感器数据进行分析，是提高预报和监测准确性的重要途径。

机器学习的研究思路主要是使用计算机模拟人类的学习活动，它是研究计算机识别现有知识、获取新知识、不断改善性能和实现自身完善的方法。这里的学习意味着从数据中学习，它包括有指导学习、无指导学习、半指导学习三种类别。有指导学习是指有结果度量的监督学习过程，希望根据一组特征对结果度量进行预测，例如根据病人的饮食习惯和血糖、血脂来预测糖尿病是否发作。通过学习已知数据集的特征和结果建立起预测模型来预测并度量未知数据的特征和结果。结果度量一般分为定量和定性两种，分别对应与统计学中的回归和分类问题。常见的有指导学习包括：决策树、神经网络、支持向量机、集成学习等。无指导学习是只能观察特征，没有结果度量，此时只能从总体中给出的样本信息对总体做出某些推断以及描述数据是如何组织或聚类的。并不需要某个结果度量或训练集，例如聚类分析、关联规则分析等。半指导学习是近年来机器学习中一个备受关注的内容，已得的观察量中一部分是经由指导者鉴定并加上标识的数据，称为已标识数据，另一部分观察量由于种种原因未能标识，称为未标识数据。通过利用已标识数据去分析并指出适当的一般性规律，再利用此规律去推断得出有关未标识数据的标识。但现在半指导学习方法的性能通常不太稳定，而且半指导学习技术在什么样的条件下能够有效地改善学习性能也不够明确。比较典型的方法便是朴素贝叶斯。

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、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种算法的分类效能。

* 1. 国内外研究现状

国内外研究表明，阿片类药物的滥用是一个多影响因素且影响因素见关系复杂的公共卫生事件。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。

机器学习作为人工智能领域的“内核”，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。对于多重共线性问题，机器学习中的神经网络模型可以很好的解决，而且对于数据的分布几乎没有任何限制。对于变量属性间的交互作用，贝叶斯网络模型可以反应不同因素间的相互关系。在自变量属性众多（如，成百上千个），传统模型不稳定甚至根本无法建立模型，集成学习却可以解决，且模型稳健。而在传染病扩散、毒品传播、疾病预后等多变量、影响因素间相互作用负责的情况下，机器学习算法模型可能显著优于传统预测模型。

目前在结合机器学习算法与阿片类药物滥用的研究中，彭巧娟等学者通过灰色关联度得到阿片类药物使用趋势与社会因素的相关性，利用神经网络进行修正验证。韦可等人通过使用多项式回归的方法，借助支持向量机确定了多项式具体参数，从而预测阿片类药物滥用率。他们所提出的模型虽然具有一定的普适性并能用于预测多个县的药物使用总量，但只是使用了机器学习中一种比较常见和普遍性比较高的算法，并不能比较除不同机器学习算法中的差异优劣。

正是由于阿片类药物使用量的影响因素复杂、使用量差异较大。且自变量之间存在多重共线性和交互作用。传统的预测模型在预测这类问题时受限，为解决这个问题，本研究利用人工智能核心技术的多种机器学习算法来预测阿片类药物使用量。通过评价比较不同的机器学习算法来确定最佳的模型，以此来描述和预测分析阿片类药物使用量问题。

* 1. 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，通过可视化的描述得出初步猜测，以便后续深入的建模分析研究。

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因素加入向量模型中。因此利用相关系数来探索阿片类药物使用与社会经济数据之间的相关性。找出与阿片类具有较强的相关性的社会经济属性。

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（K-nearest neighbor algorithm）、支持向量机(Support Vector Machines)、决策树(Decision Tree)、神经网络（Neural networks）、随机森林（Random Forest）、线性回归（Linear Regression）。对数据按照7:3的比例进行分层划分训练集和测试集。

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。

* 1. 数据来源与说明
     1. 数据来源

数据来源于美国国家法医实验室信息系统（NFLIS）与美国人口统计局（U.S. Census Bureau）。数据包括美国7年(2010-2016年)462个县的麻醉性镇痛药和海洛因的药物鉴定计数，以及各年份各县的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。

* + 1. 数据说明

采用Office Excel2016收集数据，通过Python3.6软件对数据进行进一步分析。由于数据的复杂性已经部分数据存在缺失等情况，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：

删除了没有测量值的因素。

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。

采用均数对缺失数据进行填补。

1. 模型建立准备
   1. 现状分析
      1. 不同类型阿片类药物的使用情况

图片包含 游戏机

描述已自动生成手机屏幕截图

描述已自动生成

* + 1. 不同地区阿片类药物的使用情况

图片包含 游戏机, 床

描述已自动生成

* 1. 划分类别白色的地图

     描述已自动生成
  2. 特征属性选择

1. 机器学习预测模型
   1. K近邻模型
      1. 方法
      2. 结果
   2. 决策树模型
      1. 方法
      2. 结果
   3. 随机森林模型
      1. 方法
      2. 结果
   4. 支持向量机模型
      1. 方法
      2. 结果
   5. 神经网络模型
      1. 方法
      2. 结果
   6. 线性回归模型
      1. 方法
      2. 结果
2. 模型评估与结果
   1. 各种算法对比
   2. 模型验证
   3. 模型结果（影响因素重要性）
3. 结论（讨论）
   1. 模型比较结论
   2. 社会经济因素结论
   3. 政策启示结论