# 附件 3 重庆医科大学本科生毕业论文（设计）开题报告

|  |  |  |  |
| --- | --- | --- | --- |
| **学生姓名** | 杜兴兴 | **学 号** | 2016223592 |
| **学院名称** | 公共卫生与管理学院 | **年级专业** | 2016级应用统计学 |
| **指导教师** | 曾庆 | **职 称** | 教授 |
| **课题名称** | 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 | | |
| **研究目的和意义：**  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%。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、海洛因；半合成阿片类药物:氢可酮、羟可酮、丁丙诺啡；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。  因此，了解和分析阿片类药物的过度使用情况并找出扭转这一趋势的方法刻不容缓。由于阿片类药物在日常生活中很常见，了解周围的人如何将其纳入阿片类药物使用的可能性是很重要的。与周围环境一样，数据支持某些社会经济因素可能与阿片类药物使用增加有关。鉴于这些可能性，最重要的是要找出这些因素是否能预测未来的阿片类药物使用，以及能在多大程度上预测未来的阿片类药物使用，以便制止如此多的受害者阿片类药物成瘾的恶性循环。  本研究从数据出发，通过拟合不同的统计模型来预测和分析不同的阿片类药物在不同的地区的使用情况，用来为政府政策提供信息从而控制危机。同时，通过模型泛化可以扩展到更多相类似的问题上进行预测，使模型具有一定的可扩展性。  **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。  **研究的主要内容：**   1. 各地区、各类阿片类药物使用分布情况； 2. 社会经济因素对阿片类药物分布和扩散的影响强度； 3. 机器学习的不同算法在阿片类药物使用量预测中的比较。   **要解决的主要问题：**   1. 实现多类别的数据可视化； 2. 降低或消除时间效应对模型的影响； 3. 从众多社会经济属性中识别与阿片类药物使用量有强相关的属性； 4. 机器学习算法在python中的实现。   **研究方法和手段：**   1. 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。   1. 基于时间序列分析的基本理论建立模型   考虑到时间序列的效应，为了简化模型，通过建立自回归时间序列模型来消除时间序列在机器学习模型中的影响。考虑到三种阿片类药物的变异特征存在多样性，因此设置三次序列家族来代表三类药物使用的变异。分别对三类阿片类药物建立一个修正的自回归时间序列模型。从时间序列的角度去描述阿片类药物使用的变化过程。   1. 提取强相关性的社会经济属性   考虑到时间序列较短，通过纳入社会经济因素进一步修正时间序列模型，使其更加准确。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。因此利用决策树和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。   1. 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)、朴素贝叶斯算法(naive Bayes)、集成学习（Ensembale Learning）、神经网络（Neural networks）、随机森林。   1. 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、ROC曲线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。最后，综合利用机器学习方法的预测和自回归预测，将平衡两个预测方法权重的问题转化为一个自变量分为为机器学习方法预测值和自回归预测值的线性模型。 | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **研究进度及具体时间安排** | | | | | |
| **起止日期** | **主要研究内容** | | | | **预期结果** |
| 2020.01.01-2020.02.01 | 文献查找、整理、阅读 | | | | 具备较完整的知识体系 |
| 2020.02.02-2020.02.20 | 对数据进行统计分析 | | | | 得出初步模型结果 |
| 2020.02.21-2020.02.29 | 对模型算法进行优化 | | | | 模型得到进一步优化 |
| 2020.03.01-2020.03.05 | 对分析结果进行整理 | | | | 能够对结果进行直接写作 |
| 2020.03.05-2020.03.20 | 完成论文写作 | | | | 基本形成初稿 |
| **课题评价(课题的科学性、可行性、实用价值和理论意义、存在问题和改进意见，是否同意开题)：**  该选题与学生具体实习相结合，选题有一定现实意义，具备完成选题的条件，学生参考了一定的文献，选题对于提高学生的基本知识和技能，对于培养学生的研究能力有益。学生具有一定能力，研究方法和研究计划基本合理，数据能够获得，难度合适，学生能够在预定时间内完成该选题。  同意开题。  **指导教师签名：**  **2020 年 2 月 1 日** | | | | | |
| **开题小组成员**  **(组长请加\*)** | | **姓名** | **职称** | **学科、专业** | **签 名** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **学院毕业论文工作领导小组审查意见:**  **负责人签字:**  **年 月 日** | | | | | |