基于美国毒品犯罪报告数据的分析与挖掘

汤新宇，庞家耀，施卓余，唐麒，张钰铎
北京交通大学 北京-100044 中国

【简介】如今鸦片类药物被广泛应用于各种疗法中的药物。此外生活节奏快，社会竞争激烈，来自各方的压力使人们很可能尝试毒品发泄，并对它们上瘾。因此，毒品地区性泛滥成为了一个世界性的难题，每年都有缉毒警察为此付出宝贵的生命，同时，毒品也会对人们的生命健康造成极大的不利影响。本文以美国经历的一场“鸦片危机”为例，对官方公布的权威数据进行挖掘分析，得到毒品传播的特征以及预防毒品传播和打击有关违法犯罪行为的有效措施。
本文的数据分析方法主要基于数据挖掘、机器学习和统计学等学科或方法，第一步主要采用 K-means 算法进行数据预处理，第二步进行建模分析，主要应用 SVM 与 SVR 技术，第三步运用 Apriori 算法进行关联规则分析，最后通过总结分析得到预防毒品传播的有效措施。
【关键词】毒品、机器学习、数据挖掘

目录

1 项目背景

美国在使用合成和非合成鸦片类药物用以治疗/控制疼痛（合法，处方使用）或娱乐目的（非法，非处方使用）正面临全国性危机。在美国，诸如疾病控制中心（CDC）之类的联邦组织正在为努力挽救生命并预防这种危机持续扩大带来的更一步的负面影响，例如鸦片类药物使用障碍，肝炎和 HIV 等，这将对促进美国

经济增长的重要领域产生影响。如果鸦片类药物危机蔓延到美国人口的所有领域（包括那些受过大学教育的人），那么那些需要专业技能和领域知识的职位将很难得到填补。此外，如果老年人中鸦片类药物成瘾的比例增加，美国需要支付的医疗保健等费用也将大幅度增加。因此，对这场危机的管控和防治势在必行。

2 项目目标

以美国经历的一场“鸦片危机”为例，对官方公布的权威数据进行挖掘分析，得到毒品传播的特征以及预防毒品传播和打击有关违法犯罪行为的有效措施。

3 数据描述

项目基于两个数据集。第一个数据集包含了 2010-2017 年间，美国弗吉尼亚州、西弗吉尼亚州、俄亥俄州、宾夕法尼亚州、肯塔基州五个州及其下属郡县与毒品有关的犯罪案件数。第二个数据集包含了上述地区包括教育程度、年龄组成等在内的人口结构数据。本项目将在第一个数据集的基础上利用数据挖掘的方法探索毒品在五个州之间的传播特征，并对未来毒品滥用发生在不同地区的可能性做出预测，利用第二个数据集探索不同人口结构特征对毒品滥用发生的刺激作用。

4 数据预处理

通过对第一份数据集的简单分析，发现在五个州下辖的共 462 个郡县，有 69 种不同的鸦片类药物相关犯罪报告，但某些县的犯罪报告中有很少或几乎没有部分药品（毒品）的“Drug Reports”（如下表所示），这意味着来自特定县的特定药品的样本是远远不够的。这种样本分布不均的情况，将极有可能导致在进行分析药品传播和预测分析时出现问题。

COUNTY	FIPS_Combined	Substance Name	DrugReports	TotalDrugReportsCounty	TotalDrugReportsState
ACCOMACK	51001	Propoxyphene	1	84	41462
ACCOMACK	51001	Oxycodone	1	84	41462
ACCOMACK	51001	Hydrocodone	6	84	41462
ACCOMACK	51001	Morphine	1	84	41462

为了使分布特征和趋势更加明显，我们使用 K-means 算法对数据进行预处理，将其划分为较大的区域，称为“zone”，每一个“zone”的毒品报告即为该“zone”中的所有县的毒品报告总和。

为了适当地对数据中的地区进行划分，我们应该保证“zone”不要太大或太小。通过多次实验，我们选择了将聚类中线 k 设置为 86，即为最佳情况。从技

术上讲，我们仅使用了基本的 K-means 算法，所有 86 个聚类具有相同的半径，每个聚类的县数大约在 1 到 11 之间。

在聚类之前，我们将美国的“ FIPS_Combined”编码转换为对应的经度和纬度(如下表所示),然后按照地域的距离通过 K-means 算法完成上述任务。最后，我们整合了同一区域内每种药物的所有药物鉴定报告。

County	FIS_Combined	Latitude	Longitude
51001	VA+ACCOMACK	37.74222	-75.6744
39001	OH+ADAMS	38.83989	-83.5052
42001	PA+ADAMS	39.8709	-77.2396
51510	VA+ALEXANDRIA CITY	38.81476	-77.0902
42003	PA+ALLEGHENY	40.45972	-79.976

下表展示了预处理后的数据格式，将数据转换为这种形式后，我们可以从更高层次分析药物传播过程，然后关注几个特殊区域并方便地对其进行深入研究。

YYYY	Zone	substanceName	DrugReportZone	TotalDrugReportZone
2010	0	Propoxyphene	1	84
2010	1	Morphine	9	527
2010	2	Methadone	2	334
2010	3	Heroin	5	427
2010	4	Hydromorphone	5	8500

区域化后的数据如下表所示（部分):

Zone	Latitude	Longitude				
0	37.51973	-75.9775	51001	51115	51131	
1	38.92299	-83.1945	39001	39131	39145	
2	39.96443	-77.2424	42001	42041	42055	42133
3	38.81476	-77.0902	51510			
4	40.21703	-79.846	42003	42051	42125	42129

第二个数据集由第一份数据集中包含地区的各种人口结构特征组成，每种结构特征由口数量、人口百分比两种数据描述方式组成，为消除不同地区人口数量对模型分析结果的影响，在进行数据挖掘时，采用人口百分比的描述方式作为模型输入。

5 模型总结分析

5.1 对比选择

- **多元线性回归**

优势	劣势
估计更有效，更符合实际（相对一维）	**不能很好地拟合非线性数据**
基础、简单、易实现	维度过多导致**过拟合**
根据系数得出唯一结果（低模糊性）	维度过多导致**计算量过大**

采用多元线性回归模型时，需要预先判断变量之间是否是线性关系，即使该模型能够适用于非线性数据但表现效果不佳。其次，本次数据集预处理后仍有100维，如此高纬度的数据拥有极大的数据量，会导致建模计算时间过长，更可能导致过拟合现象的出现，不利于参数的调整与建模结果的分析，故而舍弃此模型。

- **逻辑回归**

不能用逻辑回归去解决**非线性问题**，因为逻辑的决策面是线性的；此外逻辑回归对多重共线性数据较为敏感，很难处理**数据不平衡**的问题；再者逻辑回归的形式非常的简单，很难去拟合数据的真实分布，所以**准确率并不是很高**。此模型依旧不是最佳选择。

- KNN

KNN算法**效率低**，每一次分类或者回归，都要把训练数据和测试数据都算一遍，如果数据量很大的话，需要的算力会很惊人；其次，该算法还对训练数据依赖度特别大（**噪声敏感**），如果我们的训练数据集中，有一两个数据是错误的，但刚好又在我们需要分类的数值的旁边，这样就会直接导致预测的数据的不准确，对训练数据的容错性太差；再者，我们不得不考虑**维数灾难**，KNN对于多维度的数据处理也不是很好。多维空间里看起来相距较近的两个点通过数值表示是可能会很大。KNN依旧不适合我们选择的数据集。

- **决策树**

优势	劣势
易于理解和实现	信息增益**偏向**于那些更多数值的特征
可以处理连续和种类字段	容易**过拟合**
适合高维数据	**忽略属性之间的相关性**

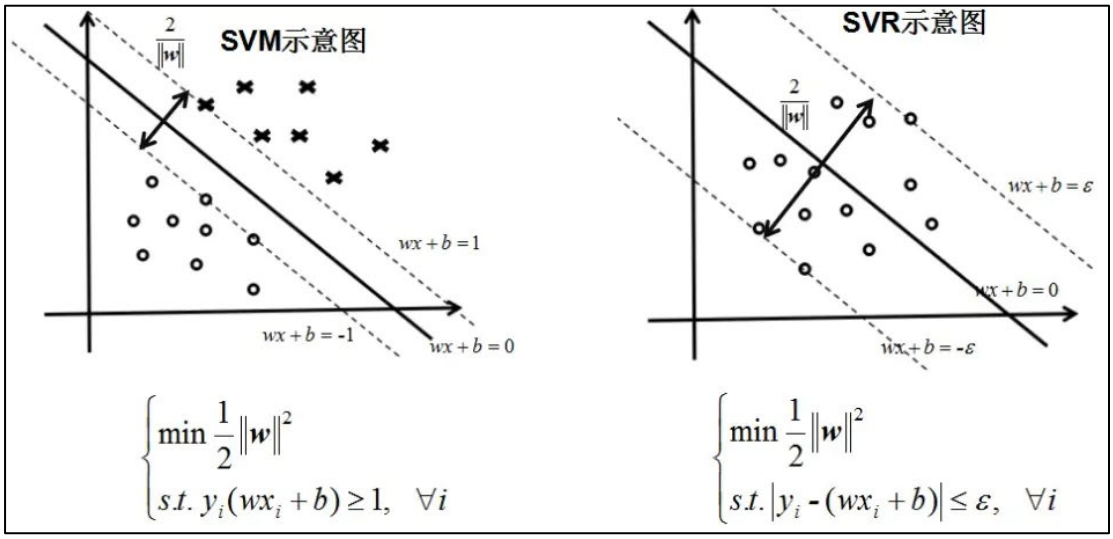
本项目数据集中包含了graphical location, marital status, education level, age distribution等相关性较强的描述特征值，忽略特征值之间的相关性将会是一个愚蠢的决定，故不能选择决策树模型。

- SVM+SVR+关联规则

SVM相对来说更**适用于非线性**数据集，有严格的数学逻辑，在保证**高精度**的条件下能够更有效地处理**高维数据**。最后关联规则的寻找注重了数据集中描述特征值间的相关性，可以更客观易懂地解释模型分析的结果。

SVR回归与SVM分类的区别在于，SVR的样本点最终只有一类，它所寻求的最优超平面不是SVM那样使两类或多类样本点分的“最开”，而是使所有的样本点离着超平面的总偏差最小。（此外，两者的运算速度均较快。）

SVM 是要使到超平面最近的样本点的“距离”最大；
SVR 则是要使到超平面最远的样本点的“距离”最小。



回归就像是寻找一堆数据的**内在的关系**。不论这堆数据有几种类别组成，得到一个公式，拟合这些数据，当给个新的坐标值时，能够求得一个新的值。所以对于 SVR，就是求得一个面或者一个函数，可以把所有数据拟合了，这一点恰巧与我们发掘毒品传播特性从而预防的目标相吻合，至此，这才是我们的最佳选择。

本项目中 SVM 用于分割数据集，将毒品报告较多的县和较少的分割（根据阈值），SVR 用于预测毒品报告将会发生的时间地点，从而预防或打击。

5.2 SVM 评估药物使用阈值

在某些县，与鸦片类药物有关的犯罪案件数量明显高于该州其他地区，这表明该县发生了更为严重的药物使用甚至是毒品滥用的情况。在大多数情况下，异常数据揭示了一个错误或应予以强调。在这里，我们假设一些县已经面临鸦片类药物危机。通过分析数据，我们发现毒品报告较少的县与毒品报告较多的县之间在数量上有较为明显的分界，因此，我们可以使用 SVM 分类器来区分县的类型。

5.2.1 SVM 分类

根据预计的鸦片类药物事件数量，SVM 分类器用于将有陷入鸦片类药物危机风险的县与其他县进行分类。对于二维的数据，SVM 分类器训练了一条称为“超平面”的线以分隔两个不同的类。超平面的计算公式为 $\omega^T x + \gamma = 0$，其中 ω 是超平面的法线向量，x 是平面上的点。有效的超平面满足：

$$\begin{cases} \omega^T \boldsymbol{x}_i + \gamma > 0 & \text{for } y_i = 1 \\ \boldsymbol{\omega}^T \boldsymbol{x}_i + \gamma < 0 & \text{for } y_i = -1 \end{cases}$$

模型中所指的支持向量是最接近超平面的点，它们位于不同集合的边界。

5.2.2 One Class SVM

SVM 分类器是一种有监督的学习模型，但是在没有提供有标签的训练数据的情况下，应使用“无监督的 SVM”，即 One Class SVM。One Class SVM 期望所有不是异常的样本都是正类别，同时它采用一个超球体而不是一个超平面来做划分，该算法在特征空间中获得数据周围的球形边界，期望最小化这个超球体的体积，从而最小化异常点数据的影响。我们首先使用高斯核将数据点投影到特征空间上，然后我们利用以下函数将投影的数据点与原点分离：

$$\begin{cases} \min_{w,\zeta_i,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\zeta_i - \rho \\ s.t.\ \left(w^T\phi(x_i)\right) > \rho - \zeta_i, i = 1,\ldots,n \\ \zeta_i > 0 \end{cases}$$

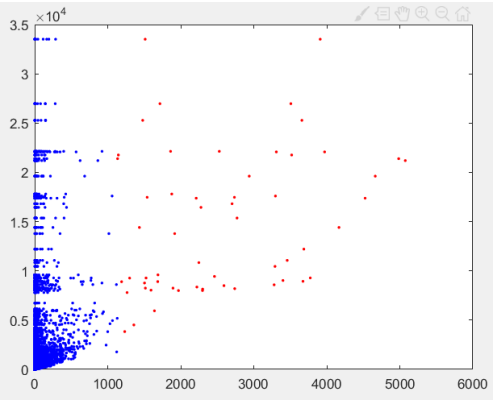
其中 ξ_i 的选取是较为宽松的，$\nu \in (0，1)$，决定了离群值的上限。最终，通过训练得到了一条在特征空间中将原点与数据分隔的线。然后将线和数据转换回未投影的空间，线外即是异常值（可被认为已经发生鸦片危机的数据）。

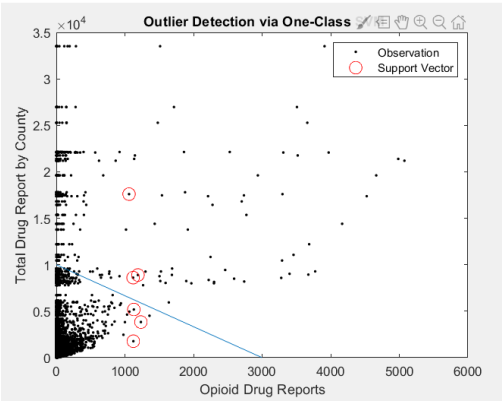
严格来说，One Class SVM 不是一种 outlier detection，而是一种 novelty detection 方法：它的训练集不应该掺杂异常点，因为模型可能会去匹配这些异常点。但在数据维度很高，或者对相关数据分布没有任何假设的情况下，One Class SVM 也可以作为一种很好的 outlier detection 方法。

5.2.3 阈值识别

本项目设计的 SVM 分类器结合了 One Class SVM 和线性 SVM。首先，我们对 2010 年至 2017 年所有的鸦片类药物报告使用 One Class SVM 来确定离群值，与这些离群值关联的县正处于鸦片类药物危机中。然后，我们使用 One Class SVM 获得的标记数据来训练线性 SVM 从而获得用以区分鸦片类药物滥用的阈值。

由于药物滥用的问题不仅与鸦片类药物有关，同时也与其他药物（或毒品）有关，因此，在案件中有大量与鸦片类药物有关的案件，或有大量与药物有关的案件但几乎没有鸦片类药物案件的县都应被视为处于药品滥用危机中。我们将一个县的毒品总报告设为 y 轴，将鸦片类药物的使用设为 x 轴来反映上述情况。同时，我们使用经过训练的模型作为分类器，以预测的药物报告数量作为输入，危险等级作为输出。下图为模型的结果，高亮线即为获得的超平面。





5.3 SVR 预测

为了预测某些药物的危机会在何时何地发生，我们需要对现有数据进行预测分析，回归方法常用来进行解决预测问题。由于支持向量回归(SVR)适合进行多维度的回归且模型已较为成熟，并且不会发生过拟合的问题，因此我们使用 SVR 模型对现有数据进行预测。

5.3.1 SVR 回归

首先创建一个 100×69×6 信息矩阵，以存储每个区域中每种药物的六年药物报告。对于每个区域，我们找到 10 个最相似的区域，每个区域都将提供一个样本，然后使用 SVR 将每个样本作为特定区域特定药物的一个特征来拟合。

基于机器学习的思想，我们将数据分为 3 组，训练集（2010-2015 年有关数据），验证集(2016 年有关数据)和测试集(2017 年有关数据)。我们从训练集中提取特征，通过验证集对参数进行调整， 最后，使用测试集对模型的性能进行评估。

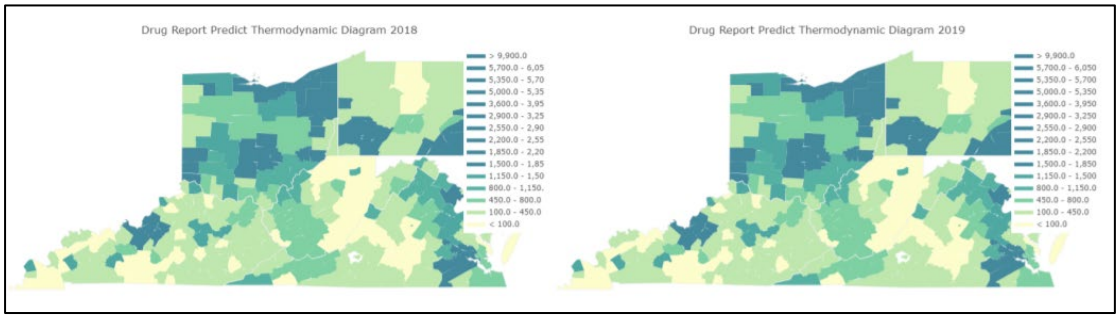
SVR 我们采用 poly 模式，对 poly 的 degree 参数进行不同的参数进行了测试，通过小二乘误差选出参数的最优值，并使用验证集再次训练模型对参数进行调整。在进行实验后，我们发现将参数设置为 3 可取得最好的实验效果。

Degree	1	2	3	4
LSE	47116	39458	28908	48322

由于数据的预测基于我们获得的七年历史数据，为了获得较准确的预测结果，我们使用训练好的模型预测了每个地区在 2018-2019 的毒品数量，并通过 Excel 工具对其进行可视化。

5.3.2 结果分析

从我们的模型得出的结果中，我们可以看到某些地区的情况有所下降，但鸦片事件总数有所增加。鸦片类药物滥用问题在已经面临鸦片类毒品危机的地区变得更加严重。面临鸦片类毒品问题风险的县与开始传播鸦片类毒品的县不一致。然而，在鸦片类毒品开始传播的县周围的县，鸦片类毒品事件的数量往往有所增加。我们估计，远离毒品问题地区的面积正在减少。



因此，我们可以得出结论，鸦片类毒品问题似乎是**区域性**的。除此之外，我们还注意到湖泊和海洋周围地区的鸦片类毒品问题更加严重。这可能是由于这些县的**人口密集**造成的。更多的鸦片类事件发生在更大的人群中，更密集的人群也使得毒品更容易传播。

5.4 关联规则学习

毒品药品的滥用不仅危害着个人的生命健康，也威胁着社会环境的稳定与安全。因此，毒品药品问题的治理和与毒品有关的犯罪问题的管理，不能仅仅从现有态势和未来趋势上入手，更要从根源入手，对各地区的社会经济因素进行分析，挖掘出导致毒品泛滥的因素。

从大规模数据集中寻找物品间的隐含关系被称作关联分析（association analysis）或者关联规则学习（association rule learning）。Apriori 算法是一种基于最小支持度和置信度，从频繁集中发现强关联规则的数据挖掘方法，最初用于从超市大规模交易数据中发掘的产品之间的购买关联。

寻找鸦片药物和数据集中人口结构特征之间的关系也可以使用关联规则来解决。在这里，我们把每个城市想象成一个“购物篮”，每一个人口结构特征都可以被视为商品，同时，我们将第一个数据集中各郡县毒品报告数据添加到“购物篮”中。我们的问题就转化为，当顾客购买哪种商品（具有的人口结构特征）

时，他很有可能购买“鸦片药物”商品。

在数据预处理阶段，我们选择了人口结构特征的百分比描述形式作为模型的输入，这样的数据仍是连续型数据，但本项目设计的关联规则模型需要一个二值化的数据类型作为输入，即 1 代表该“商品”在购物篮中，否则不在购物篮中。为此，我们通过 SoftMax 函数和经验数据对原始数据进行二值化处理，处理后的数据形式如下表所示：

Fators / ID	HC03_VC04	HC03_VC06	HC03_VC07	HC03_VC08	HC03_VC09	...	DRUG
21001	0	1	0	0	0	...	1
21003	0	0	0	0	1	...	0
21005	0	1	0	0	0	...	0
21007	0	0	0	0	1	...	1
...	...	...	...	...	...	...	...

由于数据维度高，直接在整个数据集上查找频繁集和进行关联规则学习是不现实的，因此我们按照人口结构特征的内在含义，将原有的 100 有余的特征划分为 16 个类别，分别为教育程度、年龄组成等，在每个类别分别进行关联规则的发掘。而由于数据体量大，支持度我们设置为 0.05，置信度设置为 0.8，从提升度大于 1 的关联规则中选取后后项(consequents)包含“毒品”商品的规则，结果如下表所示：

	1	2	3	4
HOUSEHOLDS	HC03_VC11	HC03_VC09	HC03_VC12	HC03_VC07
RELATIONSHIP	HC03_VC31	HC03_VC30	HC03_VC25	
MARITAL	HC03_VC35	HC03_VC42	HC03_VC45	HC03_VC38
GRANDPARENTS	HC03_VC67	HC03_VC64	HC03_VC65	
EDUCATIONAL	HC03_VC91	HC03_VC89	HC03_VC88	
RESIDENCE	HC03_VC120	HC03_VC123	HC03_VC122	
BIRTH PLACE	HC03_VC133	HC03_VC128		
LANGUAGE SPOKEN	HC03_VC172	HC03_VC171	HC03_VC173	HC03_VC170
ANCESTRY	HC03_VC198	HC03_VC185	HC03_VC194	HC03_VC182
YEAR OF ENTRY	HC03_VC146	HC03_VC150	HC03_VC144	

进行到这一步，我们筛选除了一部分特征，但我们仍然要确定这些因素是否与毒品滥用具有正相关性。这里我们使用皮尔逊相关系数，由下式给出:

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\mathrm{E}[X]]^2}\sqrt{\mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2}}$$

最后，我们获得了与毒品滥用正相关的关联规则。在对其实际含义进行分析或，我们认为，下列三个人口结构特征与毒品滥用密切相关：

Households with one or more people 65 years and over
Graduate or professional degree
Family households (families) - Female householder, no husband present, family

6 应对策略

6.1 关键因素

根据上述模型的结果，鸦片类毒品危机的三个关键因素是

（1） 单身女性户主的百分比

离婚或伴侣死亡都可能引发女性吸毒或其他心理健康障碍。如果不能得到同伴的支持，女户主将承担更多的生活负担，面临更多的社会压力。此外，女性与男性不同，由于性激素等生理因素，对鸦片类毒品上瘾的可能性较大。

（2）高学历者的百分比

受教育程度高的人不太可能对鸦片类毒品上瘾。尽管他们也可能面临很多工作和社会压力，但他们可以调整自己以适应这种压力，教育可以使人们意识到鸦片类毒品的危险，并帮助他们预防鸦片类毒品成瘾。

（2） 拥有 65 岁及以上一人或多人家庭的百分比

老年人更容易患上身心疾病。药物治疗成为缓解症状和控制疼痛的重要途径。同样，长期的治疗可能会引发老年人同样的上瘾，特别是那些失去孩子或缺乏子女照顾的老年人。

6.2. 策略描述

鸦片类药物一直被认为是一种无毒、价廉、易得的药物，此外鸦片类毒品也极易泛滥，为了减少鸦片类药物的使用，政府可以做几件事。

对于老年人

老年人是影响鸦片类药物滥用问题的关键因素。鸦片类药物广泛应用于医学上减轻疼痛，尤其是老年人。除此之外，医生也不限制给病人提供鸦片类止痛药，有些老年人使用鸦片类止痛药太频繁而上瘾。

政府可以限制医生提供鸦片类药物，或者鼓励制药厂生产廉价、有效的止痛药来替代鸦片类药物。政府可以组织社区养老院帮助这些老人寻找伴侣。共同的兴趣爱好有利于他们的身心健康。

教育水平相关

我们的模型显示，较高的教育水平可能会减少鸦片类毒品的使用。受过教育的人考虑到有可能损害他们的社会地位，倾向于避免使用毒品。

政府可以为公民提供更多的机会进入大学，例如提供奖学金或开办公立大学。在社区开展有关毒品的教育也很重要。一些有关鸦片类毒品知识的宣传口号可以贴在社区的广告牌上，甚至可以贴在鸦片类药物的包装上。

对于单身女性

在医学研究中，女性有时被忽视。有些药物可能对女性有不同的影响，例如，更容易上瘾。除此之外，没有丈夫的女性户主大多承受着更大的压力。他们更可能使用毒品来释放压力。

政府可以监督制药厂在研发过程中更多地考虑女性。政府还可以为没有丈夫的女性房主提供就业机会和福利。社区工作人员要多看望女户主，不仅要关心她

们的身体健康，还要关心她们的心理健康。

其他相关

完善相关法律法规，进行普法宣传，增强特殊地区监管，严厉打击毒品违法犯罪行为。

7 项目总结

本文以美国经历的一场“鸦片危机”为例，对官方公布的权威数据进行挖掘分析，得到毒品传播的特征以及预防毒品传播和打击有关违法犯罪行为的有效措施。虽然本项目主要是以美国数据为背景做挖掘 ，但我国甚至世界各国也可以从中获得一些启发：

（1）可以考虑通过建立数据库，应用我们的改进模型，调查毒品、药品等的传播特征；

（2）发掘毒品药品犯罪或其他犯罪与社会经济因素的关联规则；

（3）分析出这些犯罪的原因与治理方式。

这样就可以做到预防、防治和整治一条龙的社会治理。