#MCM

1906204

一句话简介：显然，这个文章不可能对每个地区的毒品数量变化情况进行灰色预测/时间序列预测，因此，如何利用一个地区与**其附近地区、与其相似地区**的联系进行预测，成为一个重要的问题。

STEP 0： 数据预处理、建模需要考虑的因素

聚类

观察本题的数据可以发现，有许多县的数据在多年中缺失，因此，我们不选择按照县划分并单独预测的做法，而是先使用 K-Means 将多个县聚成一类，成为一个地区。在后文中，我们只对各个地区进行讨论。如果再有数据缺失，那么直接用平均值替换。

两个地区之间相似性的衡量

在衡量这种相似性时，作者考虑了三个因素：

- 两个地区毒品数量的 Pearson 相关系数
- 产生联系的两个地区，地理距离必须在 200Km 之内
- 毒品数量指标：现有毒品数量高的地区和数量低的地区之间可能有不同的增长模式，因此将近几年的毒品数目纳入考虑$$
  Total = \sum_i data_i \nu^{i-1}

  结合第一条和第三条，作者生成了一个相关性指标，但是文中没有说是如何生

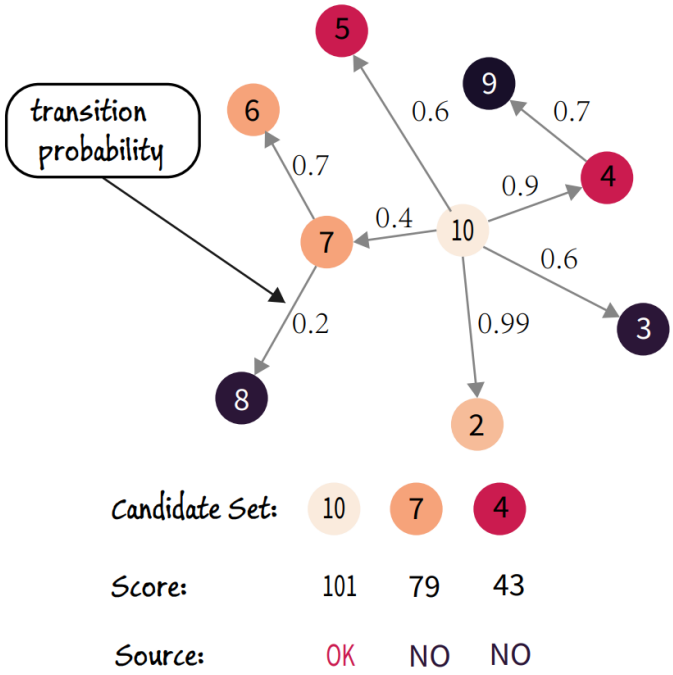
STEP 1 : 第一大问——仅仅考虑毒品数量随着时间的变化

毒品源头的寻找

首先，对于每一个区域，在 200Km 以内的区域内建图：每个点的点权是该区域内吸毒人数，当 A 区域的吸毒人数高于 B，就生成有向边 $A \to B$，边权为 A, B 之间的相关性；其次选取出度最大的五个点作为候选点；最后，从五个候选点出发，依次进行随机游走，每个点向周围转移的概率是归一化的相关性指标。
从随机游走的过程中，可以总结出每个候选点的得分：

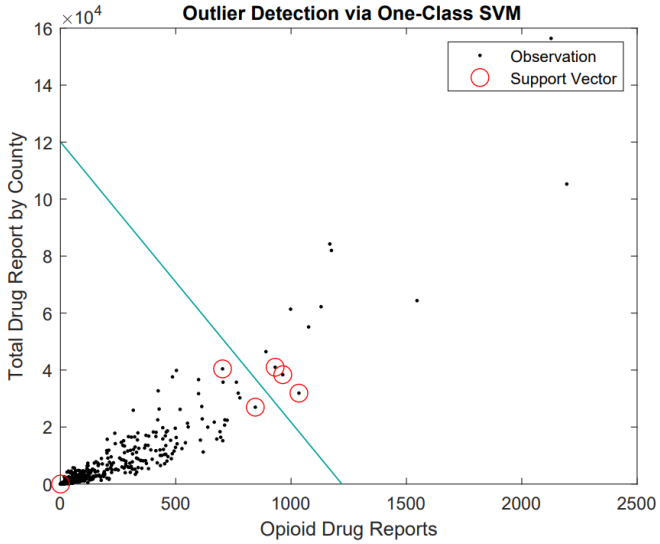
$$Score_i = \sum_j n_j \times W_i \times p_{i->j}$$

其中 n_j 是到达点 j 的总次数， W_i 是源点 i 的人数，$p_{i->j}$ 是从源点 i i 向到达点 j 的转移概率（强烈怀疑这一段有点问题），之后选得分最高的点作为源点即可。



如何划分“受到毒品严重侵袭”的地区

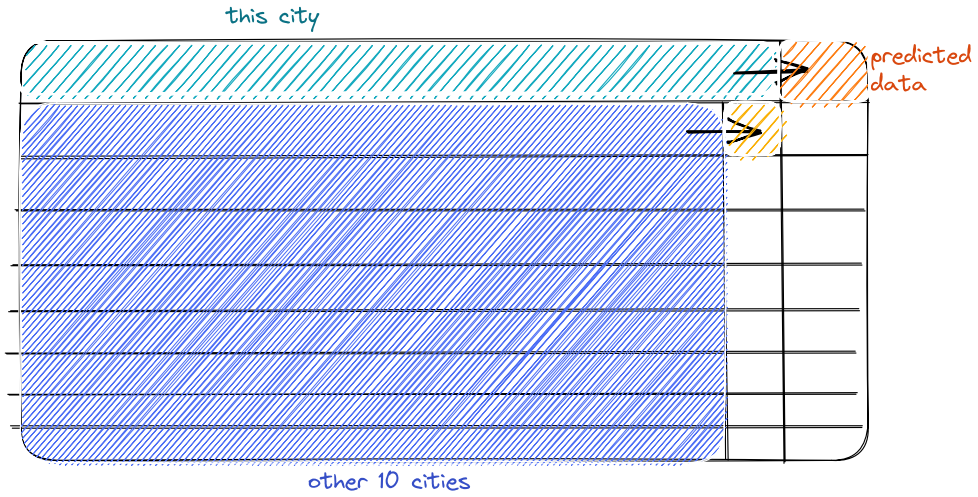
显然这是一个无监督的问题，作者先将其转化为一个有监督的的问题处理。以阿片使用数量、毒品使用总数为两轴，将每个国家标在图上，之后使用一些异常点检测算法（文中使用了 One Class SVM）为离群点打上标记，之后使用线性核函数的 SVM 找出分界线即可。



预测毒品的数量变化

这一部分作者写的比较不明不白，但是其基本思想是使用与某个地区最为相似的十个地区进行预测

我们猜测可能是下面这样



对于用作训练集的十个地区，更高的精确度带来更高的权重

STEP 2：第二大问——考虑经济和政治数据的情况

寻找毒品和其他因素的相关性：关联规则挖掘

由于关联规则挖掘所需要的数据集都是 0-1 数据，直接使用 K-Means 转换一下数据（将数据聚成两类）之后运行标准的关联规则挖掘即可

哪些因素促进了毒品的传播：Pearson 相关性

利用 PCA 给每个城市的毒品传播难易程度打分