**涉毒人员同伙预测模型拟建方案**

# 背景

目前有涉毒前科的多数重点人员中，大部分具有很强的反侦察意识，这些人在犯案时一改以往“一手交钱，一手交货”的传统犯罪方式，而通过选定毒品和毒资的投放地点，差时交易，通过同伙辅助运毒、周转毒资，甚至通过快递、物流等手段运输毒品，导致公安机关人员无法“人赃俱获”，为警务工作增加了难度。因此，目前急需一套行之有效的侦查手段，准确定位同伙嫌疑人的身份，从而能够高效的实施抓捕。

# 需求分析

针对当前在工作中遇到的现象、问题以及现有资源，客户提出了如下需求，整理如下：

1. 以铁路运毒为主要切入点，设计涉毒人员同伙的预测模型，通过该模型能够发现潜在的涉毒人员同伙，缩小嫌疑人范围
2. 在模型构建中需要考虑以下场景：
   1. 主犯与同伙乘坐相同车次列车，但在不同车厢
   2. 主犯与同伙在同一天出行，但乘坐了不同车次列车
   3. 主犯与同伙没有在同一天出行
   4. 主犯与同伙乘坐了不同交通工具

# 可行性分析

本方案主要以数据资源、实际场景两个方面作为分析依据，对方案的可行性做了评估。

## 数据资源

如果需要通过机器学习中有监督学习算法实现模型的训练与构建，那么作为构建预测模型的源数据，为保证模型的精确率达到理想值，需要具备以下特点：

1. 数据维度中需要包含分析对象的特征以及标签，特征指对分析对象特性的描述，例如年龄、文化程度、出行记录、工作情况等；标签是指该分析对象是否为犯案同伙的明确标识。同时有了特征和标签的数据才能通过有监督学习算法训练模型。
2. 数据维度广，数据的维度应覆盖分析对象的大部分特征，包括基本特征（文化程度、婚姻状况、籍贯、民族等）、社会特征（工作情况、社保缴纳记录等）、行为特征（住宿记录、民航铁路客运出行记录、手机通讯记录、手机通讯簿、QQ、微信聊天记录、地图导航记录等）。使用维度不足的数据训练出的模型在现实中预测时往往出现精确率低的情况，最终模型不可用。
3. 数据质量高，数据的质量直接影响训练模型的质量，如果数据中有大量的空值、错值（人为录入错误，格式不统一），则需要花费大量人力物力做数据清洗，且清洗后的数据的可用性也不能保证。

针对目前客户能够提供的数据资源：数据资源中只包括了本市重点人员的铁路近期出行记录以及历史出行记录，而没有与之相对的同伙的出行记录和同伙的明确标识，另外，模型的目的是预测同伙，而数据资源中没有主犯和同伙的关联信息，综上，从数据资源的角度来看，通过有监督学习训练出的模型质量，预测的精确率不能得到保证，模型构建的可行性低。

## 实际场景

根据客户在需求提出的4种场景，分别判断在每种场景下实现建模的可行性，并提出了相应的方案：

1. 在主犯（已知）与同伙（未知）乘坐相同车次列车、不同车厢的场景下：
   1. 在历史铁路出行记录中（一段时间内或者全部记录）找出所有主犯的出行车次（例如2017.01.01的G101车次、2017.02.01的G101车次、2017.02.02的G170车次）
   2. 在历史出行记录中筛选出乘坐过所有相同车次的人，并进行排序，排序方式可以按照每个人的记录总数由大到小排序
   3. 理论上排名越靠前，是同伙的可能性越大（也可根据办案人员经验或其他信息进行排除）
2. 在主犯与同伙同一天出行，但乘坐了不同车次列车的场景下：
   1. 在历史铁路出行记录中找出所有主犯的出行车次
   2. 在历史出行记录中筛选出乘坐过当天所有途径主犯目的地车次的人，并进行排序，排序方式可以按照每个人的记录总数由大到小排序
   3. 理论上排名越靠前，是同伙的可能性越大（也可根据办案人员经验或其他信息进行排除）
3. 在主犯与同伙没有在同一天出行的场景下：
   1. 在历史铁路出行记录中找出所有主犯的出行车次
   2. 在历史出行记录中筛选出与主犯起始和目的地相同的人（比如主犯有如下出行记录：北京南-南京南，南京-上海南，天津-南京。那么找出同样有过x种及以上相同出行路线的人,x可以根据实际情况指定），并进行排序，排序方式可以按照每个人的记录总数由大到小排序
   3. 理论上排名越靠前，是同伙的可能性越大（也可根据办案人员经验或其他信息进行排除）
4. 在主犯与同伙乘坐了不同交通工具的场景下：
   1. 无法实现，需要民航、客运的数据的支持

以上的4个场景和对应的方案中不涉及机器学习的算法，更多的是数据间的关联比对，得到的结果是同伙嫌疑比较大的人员的集合。

## 综合分析

如果想只以机器学习算法训练预测模型，根据目前的数据资源，可行性很低，不建议尝试。如果只通过数据间的关联比对，可以缩小同伙嫌疑人范围，可行性较高，但仍需大量警力对范围内的人进行排查，并且预测精确率无法评估。

# 概要设计

根据可行性分析结果，选择通过关联比对缩小嫌疑人范围的方式可行性较高，可落地的概率较大，而如果客户能够提供更多的数据资源（基本特征（文化程度、婚姻状况、籍贯、民族等）、社会特征（工作情况、社保缴纳记录等）、行为特征（住宿记录、民航铁路客运出行记录、手机通讯记录、手机通讯簿、QQ、微信聊天记录、地图导航记录等）），并且数据质量满足要求，那么通过将关联比对和机器学习算法相结合的方式构建模型则是一种很好的选择。即首先通过关联比对缩小嫌疑人范围，然后通过学习算法再次缩小范围的方式。以下是对以该方式构建模型的总体设计思路（如果提供不了多维度数据，那么以下设计中的机器学习部分将无法实现，但不影响关联比对部分）。

## 总体架构

在标签体系建立的过程中需要对大量警务数据以及其他领域或行业数据进行采集、存储、关联、统计、分析等一系列操作，针对这一需求我们设计了如下架构，主要包括数据采集、数据存储、数据汇聚、数据处理以及数据展示五个部分：



1. **数据采集**：在本地数据仓库与XX数据库（铁路出行历史记录）之间建立安全链接，将模型建模过程中需要的数据定期抽取到本地仓库，实现数据同步，以保证数据良好的时效性；
2. **数据存储**：原始数据有数据量大、数据日益递增的特点，为满足这一要求，使用分布式数据仓库HIVE作为存储容器，存储所有可获取到的数据资源（铁路数据、民航数据、客运数据、旅客住宿数据、涉警行为数据、手机通讯数据等），以及数据处理过程中产生的过程数据；
3. **数据汇聚**：关联各数据源数据，通过身份证号码将人员基本信息与人员基本信息、住宿信息、涉警行为信息等其他信息关联、汇聚，生成维度覆盖更全面的数据源；
4. **数据处理**：通过去重、补全、剔除等方法清洗数据，严格保证数据的规范化、归一化和一致性，为数据分析和数据建模提供干净的数据源，保障数据处理过程顺利进行；此外，我们分别使用基于经验和基于算法的方式构建住户标签体系，二者各展所长，相互补充，协同作业；
5. **数据展示**：将人员基本信息、行为轨迹以及预测结果以更加美观的图表形式进行展示。

## 总体流程



1. 从各数据来源获取铁路出行数据、民航出行数据、客运出行数据、人员基本信息、旅客住宿数据、手机通讯数据等作为原始数据；
2. 通过去重、补全、剔除等方法清洗数据，得到规范化、标准化的干净数据；
3. 获取出行数据，通过关联比对缩小同伙嫌疑人范围，得到关联比对结果；
4. 获取特征数据，随机抽取70%的数据用于构建模型，剩余30%的数据用于评估模型；
5. 通过机器学习算法的方法训练模型；
6. 使用30%的数据对模型的精确率和召回率进行评估，如果评估不通过，则对模型进行调整，修改后重复步骤6；如果评估通过，则可将模型投入到实际场景中进行预测；
7. 将关联比对结果作为模型的输入，再一次缩小嫌疑人范围。

## 模型构建

整体模型由两个模块组成，分别为关联比对模块和机器学习模型产出模块：

关联比对模块的输出为关联比对结果，表现形式为数据表；

机器学习模型产出模块的输出为机器学习模型，表现形式为一段程序，关联比对结果数据表将作为该程序的输入，程序的输出结果则是涉毒人员同伙嫌疑人的名单及其嫌疑概率。

下面分别多两个模块进行具体说明。

### 关联比对模块

#### 输入输出

**输入**为涉毒人员及正常人的铁路历史出行记录，若考虑更多的场景，那么还应提供民航、客运等历史数据；

**输出**为关联比对的结果，每个结果中必须包含个人的身份ID，其他的信息可以根据不同场景有所不同。

#### 流程图



针对不同的场景，需要设计不同的关联比对规则，场景与规则具有一对一的关系，规则中包括了需要使用的哪些出行数据中的哪些信息和关联比对时用的哪些方式方法以及输出什么样的结果。规则制定完成后，将数据按规则转换成结果，此流程相对简单，因此不再赘述。

### 机器学习模型产出模块

#### 输入输出

**输入**为涉毒人员同伙的特征数据；

**输出**为决策树模型。

#### 流程图



#### 主要步骤

相对于关联比对，机器学习模块流程相对繁琐，下面做详细的说明：

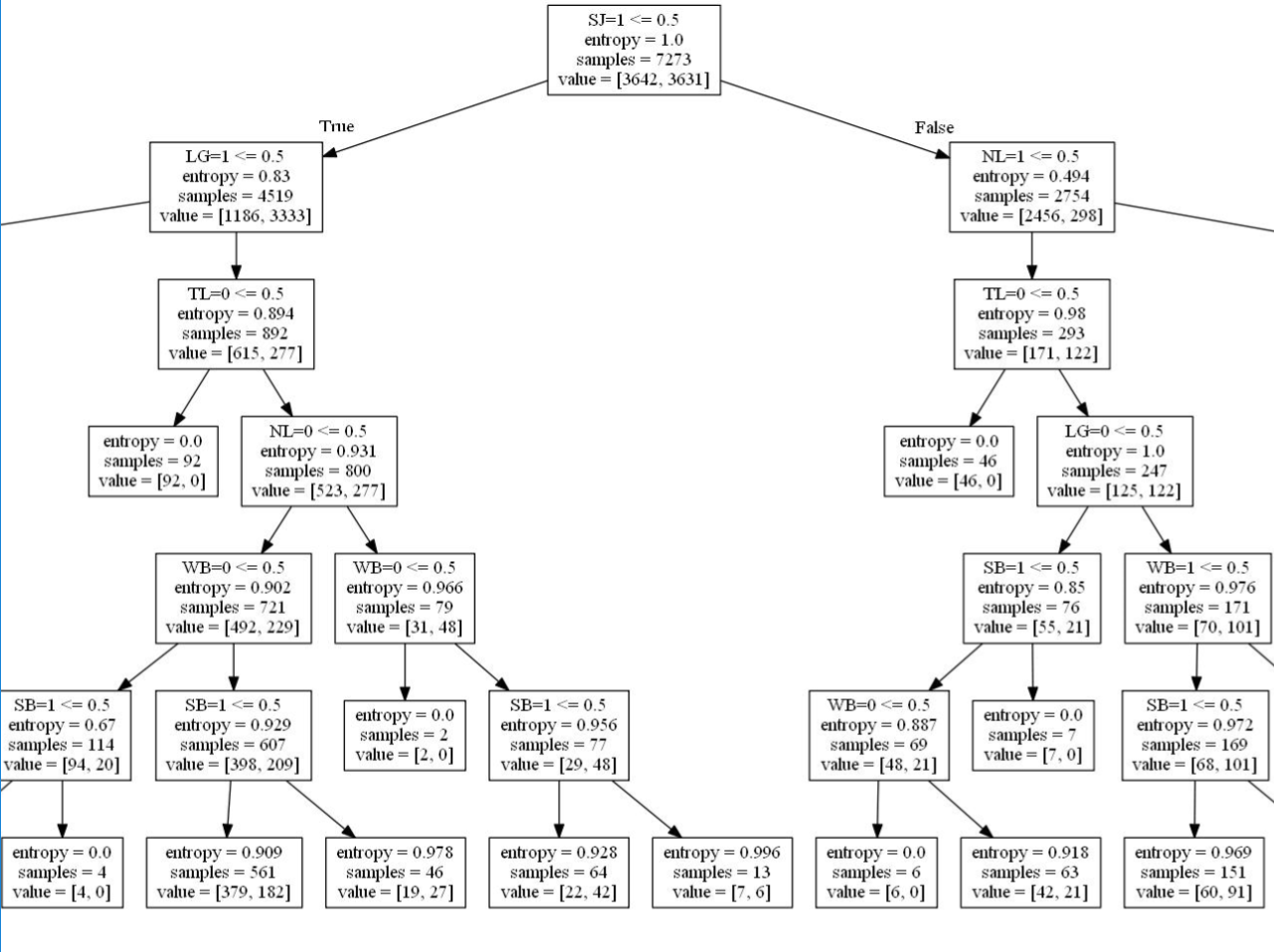
1. 采样：首先获取涉毒同伙人员相关数据，并与旅馆住宿信息、手机通讯录信息、涉警行为信息等数据关联，作为正样本；将非涉毒同伙的正常人员数据作为负样本。但是，由于正常人人数远大于涉毒同伙人员，将会导致正负样本严重不平衡，因此需要在两者特征属性分布基本不变的情况下对正样本采取上采样（即模拟生成和当前稀有样本临近的样本），对负样本进行下采样，（即对负样本聚类，在每个类别中上按比例抽取部分样本）。最终满足正负样本比例1:1。
2. 特征选择：特征主要包括人员基本特征（年龄、文化程度、婚姻状况、籍贯、民族等）、社会特征（工作情况、社保缴纳记录等）、行为特征（住宿记录、手机通讯记录、手机通讯簿、QQ、微信聊天记录、地图导航记录等）
3. 模型选择：初步选择决策树算法，一种基本的分类方法，属于有监督学习，可基于实例特征对实例进行分类。通过决策树算法训练生成的模型可读性强，因为模型就是由数据属性和类别作为树枝节点构成的树，因此在处理案件时办案人员可以很直观地看到模型的构成，熟悉其工作方式，知悉判定为同伙嫌疑人的依据。并且决策树的时间复杂度低，数值不超过其最大深度，所以在利用决策树模型进行预测时速度很快，基于该特点，办案人员在办案过程中能够快速缩小嫌疑人范围，最大程度上加快办案速度，提高办案效率。
4. 模型构建：决策树算法主要有ID3、C4.5以及随机森林等，本方案中选择C4.5算法构建决策树模型，其在ID3算法上做了改进。C4.5使用信息增益率来选择合适的分割节点，信息增益率越大，对分割节点的选择就越合理，信息增益率计算公式如下：





1. 为信息增益
2. 为特征的不纯度
3. 本方案中不纯度用熵值来衡量，公式为：
4. 为特征等于属性时的不纯度
5. 为总样本的不纯度
6. 为总样本在被特征划分后的不纯度
7. 为总样本在被特征划分后的信息增益率
8. 为分割的节点数，一般= 2
9. 表示特征的 属性

假如以年龄（NL）、是否正常缴纳社保（SB）、网吧上网次数（WB）、入住旅馆次数（LG）、是否有涉警记录（SJ）、是否被涉毒人员存为手机联系人（TL）、手机通讯录中是否存有涉毒人员为特征，以是否为涉毒人员同伙为标签，使用总体样本中的70%的样本作为训练集，通过计算分割节点之后和之前的信息增益率来选择最优分割点，最终可以构建形如以下样式的决策树模型：



1. 模型评估：将总体样本剩余30%数据作为测试集，使用决策树模型对该测试集进行预测，将预测结果与实际比对，计算出精确率及召回率，如果精确率和召回率达到要求，则可将决策树模型投入使用；如果没有达到要求，对模型进行优化（剪枝），再次评估，直到达标或确定无法达标为止。

## 模型应用



模型在构建过程中使用的全部是历史数据，模型的输出结果为经过两次筛选的涉毒人员同伙嫌疑人名单，该模型的应用主要表现在如果该名单的人员在本地铁路出行监测系统中出现便立即发出预警，相关部门可立即派遣警力对其实施监控。

# 风险评估

本方案风险点主要在于数据资源的质量，如果数据资源不能够达到在可行性分析中提及的3点要求，那么机器学习模型的精确率、召回率很难达到既定要求（至少大于80%），甚至无法训练模型，如此便导致由关联比对模块得到的结果中，同伙嫌疑人范围仍然很大，对于警务工作的贡献微乎其微。

其次，本方案涉及到的场景有限，实际情况应更加复杂，需要相关领域的专家，或具备相关经验的警务工作人员提供支持，主要表现在在关联比对规则的制定上，规则的好坏直接对最终结果产生影响，如果首次筛选就遗漏目标，导致目标漏网，或者力度不够，导致目标范围过大，都会产生不利结果。