江苏省数据创新大赛

创意方案

**方案名称：** 公寓楼住宅实战管理

**参赛单位：** 科技信息化处 广陵分局

**参赛人员：** 陈焱 曹丹 时菁森

# 建设背景

公寓式住宅租期灵活，在发展中逐渐形成了人员结构复杂、流动性强、治安问题突出等特点。传统的警务工作模式已经无法适应公寓楼管理的需求。针对这一现状，我们以智能门禁系统建设为契机，创新公寓楼住宅实有人口管控手段。首先通过对各类警务数据综合分析得出特殊群体的特征规律及行为模式,据此构建公寓楼住户标签体系。将门禁系统实时收集到的大量鲜活、有效的数据，与公寓楼住户的标签体系相结合应用于公安实战，在实战中对标签体系进行动态化的调整，便于公安机关在大量的实有人口中快速精准发现重点人员及服务对象，针对不同对象，采取相应的打击处理及帮扶措施，在预判重点人员及服务民生等方面进行实践。

# 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**：在本地数据仓库与警务基础平台数据库、情报平台数据库、蛛网数据库之间建立安全链接，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，我们使用分布式数据仓库HIVE作为存储容器，存储所有抽取得到的数据（涉毒人员数据、嫌疑人员数据、网吧上网数据、民航铁路数据、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出行信息、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

如何构建公寓楼住户标签由不同的业务场景决定，以打击防范住户中隐性涉毒人员为业务场景，则需要通过隐性涉毒人员的各种特征构建隐性涉毒人员标签；以服务民生、关爱孤寡老人为业务场景，则需要通过孤寡老人的各种特征构建孤寡老人标签。为了能够适应更多的业务场景，使标签的构建过程与业务场景松耦合，设计了如下流程：



1. 从各数据来源抽取涉毒人员数据、嫌疑人员数据、网吧上网数据、民航铁路数据、旅客住宿数据、涉警行为数据、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. 通过身份证号码将人员基本信息与人员出行信息、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关联数据；
4. 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. 通过基于工作经验或基于决策树分类算法的方法构建模型，生成标签；
6. 使用30%的数据对模型的准确率、精确率以及召回率进行评估，如果评估不通过，则对模型进行修改，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. 通过模型预测并验证预测结果的准确性，根据大量预测结果准确性评估模型、优化模型。

# 住户标签体系构建及应用

## 基于工作经验的标签构建

住户标签是住户多个特征的集合，根据不同住户的基础属性、活动轨迹、特殊行为三个维度给住户打上不同的标签指标。其中基础属性包括人员的年龄层次、同住人员、是否有违法犯罪记录及社保缴纳情况等，活动轨迹包括上网、住宿记录及公寓楼进出记录等，特殊行为包括涉警行为和特定通联行为等。从各类数据资源中挖掘分析出特定人群的特殊规律，以此设定具体的指标参数。

以构建涉毒人员标签为例：

首先分析涉毒人员的特征。我们以2017年首次因涉毒被打击处理的共计644人作为隐性涉毒人员样本（因涉毒人员被抓获成显性人员后，几乎不会使用本人的身份证登记住宿、上网等，因而用显性涉毒人员作为样本不具代表性），研究其2016年全年的行为轨迹，再与同时间段内扬州市所有人历史轨迹作对比，分析得出以下几点差异较大的规律性特征。

### 基础属性特征

#### 年龄结构

抽取的644名隐性涉毒人员样本中，年龄集中在18-47周岁之间的共计592人，占比92%。

#### 社保缴纳记录

抽取的644名隐性涉毒人员中仅有58人在2016年正常缴纳社保，无社保缴纳记录及未连续缴纳社保的人员占比91%。

#### 涉毒前科

涉毒前科人员：将扬州市历史上涉毒人员在除去2017年首次因涉毒被查处的人员后，对其因复吸被公安机关查获的情况进行分析，得出因复吸被查获的涉毒前科人员占比为48%。而在实际工作中，更多复吸人员并未被公安机关查获，所以涉毒人员的实际复吸率要远高于48%。

### 活动轨迹特征

#### 登记旅馆入住时间

我市2016年全年登记旅馆住宿共计6384027人次，经分析入住时间集中在12时-20时之间，在20时-次日4时之间入住的人数呈明显下降趋势。对比644名隐性涉毒人员在2016年的旅馆住宿记录，我们发现上述涉毒人员的登记入住时间在22时-次日3时之间达到峰值，明显有异于常人。

#### 网吧上网下线时间

我市2016年全年网吧上网人员共计20256512人次，下线在4时-8时之间处于低谷。对比644名隐性涉毒人员在2016年的网吧下线记录，我们发现上述涉毒人员的下线时间在6时-8时之间达到峰值，明显有异于常人。

#### 公寓楼进出门时间

由隐性涉毒人员网吧上网、旅馆住宿的活动轨迹特征可以分析出，涉毒人员在夜间的活跃度远高于常人。可将夜间2时—6时进出公寓楼作为一项指标参数。

### 特殊行为特征

#### 涉警行为

644名隐性涉毒人员在2016年共计涉警记录561人次，人均0.87次涉警记录；而扬州市同年龄段的1273849名实有人口中，人均涉警0.25次。

#### 特定通联行为

抽取2017年7月以来蛛网库内涉毒人员211人，提取这211名涉毒人员的通讯录号码，对通讯录中能够比中身份的人员进行比对。

①被这211名涉毒人员中的一名涉毒人员存为关系人的3558人，这3558人中因涉毒被公安机关打击处理过的共420人，占比11.8%；

②被这211名涉毒人员中的两名涉毒人员存为关系人的873人，这873人因涉毒被公安机关打击处理过的共145人，占比16.6%；

③被这211名涉毒人员中的三名涉毒人员存为关系人的194人，这194人因涉毒被公安机关打击处理过的共49人，占比25.3%；

④被这211名涉毒人员中的四名涉毒人员存为关系人的84人，这84人因涉毒被公安机关打击处理过的共45人，占比53.6%；

⑤被这211名涉毒人员中的五名涉毒人员存为关系人的21人，其中3名警察，12人因涉毒曾被公安机关打击处理，其他6人有涉毒以外的其他违法犯罪记录。分析得出除警察外的18人，违法犯罪人员占比100%，涉毒人员占比66.7%；

⑥被这211名涉毒人员中的六名涉毒人员存为关系人的5人，这5人均因涉毒被公安机关打击处理过，占比100%。

在测算中，我们仅是以比中显性涉毒人员为标准计算占比，而在实际情况中，必定有部分涉毒人员并未被公安机关查处过，所以在涉毒人员比中率中，我们所测算出的结果为最小值。以上数据不难看出，被越多涉毒人员存为手机联系人的，其涉毒的可能性越高，当被五个或者五个以上涉毒人员存为手机联系人的，在排除我们公安机关民警的情况下，基本可以判定其违法嫌疑属性。

## 标签应用

仍以涉毒人员标签为例，前面构建的七项标签指标中，涉毒前科属性因比中率较高，可直接定为高风险涉毒人员，结合前科人员进出公寓门禁记录，适时上门核查是否有违法犯罪行为；年龄属性作为基础条件，故将其它五项特征作为标签指标，分类如下：

|  |  |  |
| --- | --- | --- |
| **代码** | **标签指标** | **指标参数** |
| A | 被多名显性涉毒人员存为手机联系人 | 被3名显性涉毒人员存为手机联系人 |
| B | 是否正常缴纳社保 | 未正常缴纳 |
| C | 网吧下线时间 | 6时—8时 |
| D | 入住旅馆时间 | 22时—次日3时 |
| E | 是否涉警人员 | 是 |

为提高公安工作效率，最大程度地避免警力浪费，我们依据蛛网抽取的211名涉毒人员通讯录，将其中被三个涉毒人员存为手机联系人的194人设为样本A，其比中的显性涉毒人员共49人，比中涉毒人员机率至少25.3%。用其他指标与样本A进行排列组合后，验证出比中显性涉毒人员机率的浮动情况如下：

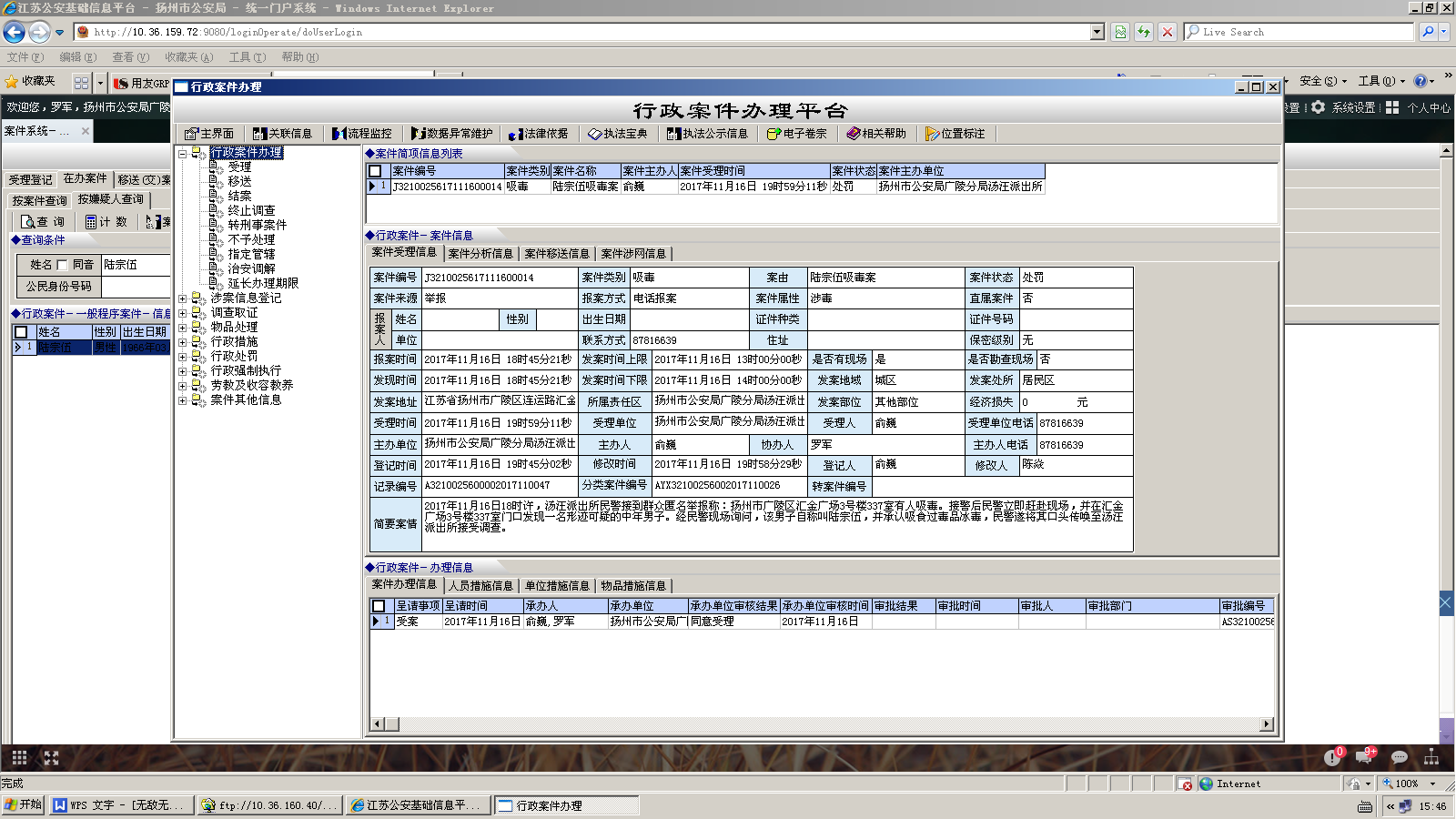
|  |  |  |  |
| --- | --- | --- | --- |
| **组合方式** | **比中显性涉毒比例** | **组合方式** | **比中显性涉毒比例** |
| 样本A | 25.3% | A+C | 16% ↓ |
| A+B | 27.9% ↑ | A+B+C | 17%↓ |
| A+D | 27.3% ↑ | A+C+D | 16.3%↓ |
| A+E | 27.1% ↑ | A+C+E | 15.3%↓ |
| A+B+D | 29.8% ↑ | A+B+C+D | 10%↓ |
| A+B+E | 29% ↑ | A+B+C+E | 16.2%↓ |
| A+D+E | 26.8% ↑ | A+C+D+E | 15.3%↓ |
| A+B+D+E | 28.5%↑ | A+B+C+D+E | 15.7%↓ |

组合验证结果显示：

1、组合指标的种类个数不能作为比中的依据，并非符合越多的指标种类比中机率越高；

2、活动轨迹维度方面的指标要分开组合。（比如网吧下线时间和入住旅馆时间两个指标要分开使用，因上述两个行为同时发生的可能性很小，比中机率会明显下降）

3、以涉毒为例：使用A+B、A+D、A+E、A+B+D、A+B+E、A+D+E、A+B+D+E作为组合后比中率上升，可以将符合上述组合的住户判定为具备高危涉毒风险人员（其他组合类型在测算中比中率下降，说明组合欠缺科学性）。一旦判定为具备高危涉毒风险人员，我们可将公寓楼进出记录作为公安机关上门核查的时间依据，提高查处工作的针对性和有效性。

基于上述标签体系，我们于今年11月在汇金广场成功抓获二名涉毒人员。

**由此，我们可以总结出其他类型犯罪人员均可按照上述模式，先行分析特征规律，设定各指标参数，通过指标间的组合验证得出高危风险人员，构建相应的住户标签及相关指标、参数。通过案例可以看出，通过住户标签应用，可以有效实现预判嫌疑人与打击犯罪的目标，便于公安机关采取准确有效的打击、管理措施，提升公安工作的洞察能力。**

同时，在服务民生方面，标签构建及应用也可发挥显著作用。以构建独居老人标签为例，标签维度在基础属性上需满足70周岁以上且独居，行为轨迹上需同时满足两天及以上无进出门禁记录、无铁路民航旅客记录、无宾馆入住记录。我们对分析得出的上述标签用户，会同社区、物业及时与老人的子女联系或上门核查老人的具体情况，防止意外事件的发生。

|  |  |  |
| --- | --- | --- |
| **序号** | **标签指标** | **指标参数** |
| A | 公寓登记入住人员情况 | 70岁以上，独居 |
| B | 门禁记录 | 48小时未进出 |
| C | 民航、铁路记录 | 未发现民航、铁路记录 |
| D | 宾馆入住记录 | 未发现宾馆入住记录 |

## 基于决策树分类算法的标签构建

决策树算法是一种基本的分类方法，属于有监督学习，它是基于实例特征对实例进行分类的过程，我们也可以认为决策树就是很多if-then的规则集合。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。另外，虽然单个决策树模型在多业务场景下的预测效果不是很好，但是多个（几十甚至上百）单树模型可以根据实际需要进行组合，构成随机森林，从而扩大处理范围，以适应更多更广的业务场景。

下面将以涉毒人员标签为例，分别阐述采样、特征选择、构建模型的过程。

### 采样

关联数据库中嫌疑人信息表和案件信息表，得到2017年所有隐性涉毒人员相关数据，并将该数据与旅馆住宿信息、手机通讯录信息、涉警行为信息等数据关联，作为正样本；将未登记在嫌疑人信息表和违法信息犯罪表中的正常人员数据作为负样本。然而，对于现有数据而言，正负样本严重不平衡，即正常人占比远大于隐性吸毒人员的占比，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），最终使正负样本趋于平衡。

### 特征选择

样本数据中除了包括姓名、性别以及身份证号码等基础特征外，通过关联数据库中其他数据信息，如旅店住宿信息、涉警行为信息、手机通讯录信息等，得到更多不同维度的特征值数据，主要特征包括：

1. 年龄：通过人员基本信息获得年龄特征；
2. 性别：通过人员基本信息得到性别特征；
3. 是否常驻居民：通过关联人员基本信息得到是否常驻居民特征；
4. 住店入住时间：通过关联旅店住宿信息得到入住时间；
5. 住店时长：通过关联旅店住宿信息得到住店时长；
6. 手机联系人数量：通过关联手机通讯录信息得到手机联系人总数；
7. 手机中是否存有显性涉毒人员手机号码：通过关联手机通讯录信息得到手机中存有显性涉毒人员手机号码数量；
8. 是否被显性涉毒人员存为联系人：通过关联手机通讯录信息得到被显性涉毒人员存为联系人的人数；
9. 住户用电量：通过关联住户信息及电表信息得到住户从入住期间平均用电量；
10. 住户用水量：通过关联住户信息及水表信息得到住户从入住期间平均用水量；
11. 住户燃气使用量：通过关联住户信息及燃气使用信息得到住户从入住期间平均煤气使用量；
12. 网吧上线时间：通过关联网吧上网信息得到网吧上线时间；
13. 网吧下线时间：通过关联网吧上网信息得到网吧下线时间；
14. 网吧上网时长：通过关联网吧上网信息得到网吧上网时长；
15. 2017年内乘坐飞机次数：通过关联民航信息得到本年度通过航空出行次数；
16. 2017年内乘坐火车次数：通过关联铁路信息得到本年度通过火车出行次数；
17. 2017年内乘坐客运次数：通过关联客运信息得到本年度通过汽车出行次数；
18. 2017年内出行总次数：通过关联民航、铁路、客运信息得到本年度出行总次数。

### 构建

决策树算法主要有ID3、C4.5以及随机森林等，由于根据住户信息判断其是否为隐性吸毒人员这一场景相对单一，可简单划分为单分类问题，所以本方案中选择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

本方案中不纯度用熵值来衡量，公式为：



1. 为特征等于属性时的不纯度
2. 为总样本的不纯度
3. 为总样本在被特征划分后的不纯度
4. 为总样本在被特征划分后的信息增益率
5. 为分割的节点数，一般= 2
6. 表示特征的 属性

以年龄（A）、是否有涉警行为（B）、手机联系人中是否有显性涉毒人员（C）、是否被显性涉毒人员存为联系人（D）为特征、以是否为隐性涉毒人员（X）为标签为例，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，最终可以构建如下决策树模型：



使用上述决策树模型进行预测：当有新住户登记时，我们可以采集信息得到如下数据：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | 姓名 | 年龄 | 涉警行为 | 联系人中存有  显性涉毒人员 | 被显性涉毒人员存为联系人 | 隐性涉毒人员 |
| 1 | 张三 | >18 | 否 | 是 | 是 | ？ |

将各特征值代入模型(C=yes,D=yes,B=no,A>18)后得到X=yes，即该住户被标记为隐性涉毒人员，应该重点防范。

# 住户标签生命周期管理

在公寓楼住户标签设计应用过程中，我们将住户标签生命周期管理概括为“一线两环”。“一线”是标签生命线，包括“目标群体特征归纳—标签生产—标签应用—应用效果反馈—优化标签”等环节，“精准度”和“科学性”要求贯穿于整条生命线。在生产环节，将业务数据中挖掘出的潜在规律作为标签生产的依据，在应用环节，根据实际应用中的现实情况即时优化标签参数，在优化环节，根据活跃程度对已标签的目标群体进行动态化监测，对已逐渐丧失活跃度的目标予以标签撤销。“两环”是“标签设计-发现数据问题-促进数据整改”和“标签应用-应用效果反馈-优化标签”两个应用闭环。一方面，在利用数据挖掘技术设计标签的过程中，通过数据探索，分析数据分布特征、缺失值、离群点等数据状况，将发现的源头数据问题反馈到业务系统，促进问题数据的检查、修正，提高数据质量；另一方面通过标签应用反馈到标签库，可以保证标签库动态更新，改进标签规则，优化标签应用策略。在此基础上，通过标签推送功能向民警精确推送数据结果，支撑服务公安实战业务。