犯罪情报分析中的数据挖掘应用①

陈 鹏[1], 瞿 珂[2], 胡啸峰[1]

[1](中国人民公安大学 警务信息工程学院, 北京 100038)
[2](北京市公安局情报信息中心, 北京 100034)

摘 要: 本文基于公安业务中的治安防控原理, 构建了面向情报分析和决策指挥的犯罪情报数据挖掘框架. 首先, 对案事件数据库进行预处理和空间编码的基础上得到标准化的案件信息数据, 随后, 利用聚类分析、关联分析和分类分析中的相关方法可得到治安案件的时空风险、重点人特征和作案手段特征等信息. 通过对北京市实际盗窃案件数据进行挖掘, 证明了数据挖掘技术能够很好的应用于犯罪情报的分析.

关键词: 情报分析; 时空分析; 关联分析; 数据挖掘

Application of Data Mining in Criminal Intelligence Analysis

CHEN Peng[1], QU Ke[2], HU Xiao-Feng[1]

[1](Policing Information Engineering Institute, People's Public Security University of China, Beijing 100038, China)
[2](Information Center, Beijing Municipal Public Security Bureau, Beijing 100034, China)

Abstract: This paper builds a framework about the crime data mining of on intelligence analysis and decision command based on the principle of prevention and control in public security. First, we can get the standardized crime information according to the preprocessing of crime database and space encoding. Using the related methods used in cluster analysis, classification and association analysis, we can get such information as the spatial-temporal risk distribution of crime, targeted people features and modus-operandi. Finally, by mining the data of actual theft cases in Beijing, it is proved that data mining methods could play significant role in crime intelligence analysis.

Key words: intelligence analysis; space-time analysis; associate analysis; data mining

前言

随着公安信息化的发展, 公安部门掌握的各类信息也开始呈现出海量化增长的态势, 而这也为公安部门开展犯罪情报分析、探索犯罪活动的基本规律, 进而为犯罪的精确打击、治安防控与高效的智能指挥决策提供了坚实的基础. 近几年, 相关学者在犯罪情报分析领域开展了一系列的工作, 利用数据挖掘、数据可视化等技术来进行犯罪活动规律的探索. 例如金光、熊允发等将数据挖掘决策树应用于公安情报分析, 提出了犯罪风险的预测方法模型[1,2]; 余先虎, 许阳泉, 包晔, 叶文箐等将数据挖掘中的关联分析方法应用于刑事犯罪分析, 实现了犯罪要素的频繁项提取[3-6]; 李代超, 吴文浩等则从时空可视化角度对犯罪行为的时空规律进行了探索[7,8]. 然而综合现有的工作来看, 目前开展的研究更多的关注于特点方法的应用, 缺乏与公安业务尤其是公安情报实战化需求的紧密结合, 这就导致了现有的基于数据挖掘、数据分析方法的犯罪情报研究仅仅是一种片面的探索, 而没有从公安业务的角度来体系化、整体化的应用数据挖掘工具.

基于此, 本文针对目前公安部门在治安防控行动中的业务特点, 提出了案件情报分析中的分析框架, 并以北京市 2011 年盗窃类案件作为案例, 运用相关数据挖掘方法进行了犯罪情报分析.

1 公安信息化中的案件信息特点及分析框架

当前, 公安部门在案件信息的管理上主要是采取

① 基金项目:国家"十二五"科技支撑计划项目(2015BAK12B03);中国人民公安大学基本科研业务费项目(2016JKF01211)
收稿时间:2016-06-03;收到修改稿时间:2016-07-07 [doi: 10.15888/j.cnki.csa.005609]

案件数据库的方式. 在案件的归类与整理过程中会对案件信息按照相应的标准字段进行录入和存储管理, 目前公安部门的案件数据库中包含的主要字段有:

(1) 案件编号: 即案件 ID 号, 用于在案件数据库中标识案件的唯一性;

(2) 发案时间: 案件发生的具体时间, 为一标准的 12 位数字编码, 其格式为: yy/mm/dd/hh/mm/ss;

(3) 发案地址: 案件发生的具体地点, 一般精确到街道门牌号;

(4) 作案部位: 案件发生的地段类型, 如居民小区、公路、商场等;

(5) 作案手段: 指在案件发生过程中嫌疑人所采用的作案手段和工具;

(6) 嫌疑人身份信息: 一般为嫌疑人的身份证号.

公安部门的治安防控业务类型主要包括街面巡逻、卡口盘查和社区防范. 其中街面巡逻主要在特定时段对案发较高的地段进行重点巡逻, 因此在情报分析上应以聚类分析为主, 采取包括时空可视化、时间异常点分析、空间热点挖掘、空间聚类分析和时空风险分析等方法来发现案件的时空热点. 而卡口盘查则相对更加精细化, 主要在重要路口及重点部位对街面行人进行盘查, 以杜绝或震慑可能发生的犯罪活动, 因此在工作中应当侧重于对作案手段、部位等特征的频繁项提取, 例如对案件信息中的作案时段、作案部位和作案手段可通过频率统计[9]、决策树分析[2]、关联分析等[3,4]来发现作案时段、作案部位和作案时段之间存在的联系. 社区防范则主要依托于社区警务, 来对社区中出现的一些异常人员和重点人员进行登记、访问等, 防止其可能出现的一些犯罪行为, 因此在重点人员的筛查和识别上应采取分类等方法, 利用支持向量机、贝叶斯分类等方法根据人员的标签信息如年龄、学历、户籍地、前科记录等来对社区中的普通居民和具有犯罪动机的可疑人员进行筛查.

基于公安部门的业务特点, 本文构建了面向治安防控的情报数据挖掘分析框架. 该框架由两部分组成. 首先, 需要对案件数据库中的案件信息进行清洗, 剔除缺陷数据, 并对案件地址信息进行空间编码; 其次, 对标准化的案件信息分别利用聚类、分类和关联等方法来挖掘案件的时空风险范围、重点异常人员群体和典型作案特征等信息.

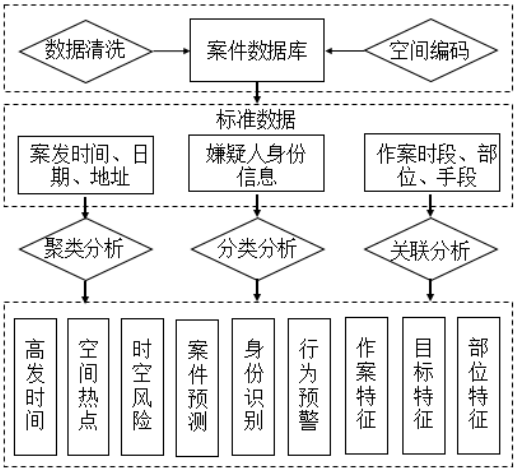


图 1 犯罪情报挖掘分析框架

2 实例分析

2.1 数据来源

本文研究的案件信息来源于北京市公安局盗窃电动车案件信息库, 本文提取了该信息库中 2011 年的案件信息. 该案件信息库中共有案件记录 1000 余条, 其中案件字段有案件编号、发案日期、发案时间、案件状态、破案日期、案发地址、作案手段、作案部位等. 其中发案日期和发案时间提供了案件的时间信息; 案发地址提供了案件的空间位置信息; 作案手段则对物品被盗过程中的作案手法进行了描述, 具体可分为“盗车”、“开锁”、“剪拉”等六类; 作案部位则详细描述了物品被盗的环境特征, 包括“居民区”、“商业区”、“广场”、“车站”等十三类; 案件状态则包括了“立案”、“破案”和“受理”等. 其中对于案件状态显示为“破案”的记录则包含有全部的字段信息, 而显示为“破案”和“受理”的案件记录则仅包含有部分信息. 为了有效的分析案件数据, 达到系统性提取知识要素的目的, 本文以“破案”为标签进行数据检索, 并删除重复记录和信息缺失的部分数据, 最终得到有效案件记录共 363 条.

2.2 时空聚集性分析

对北京市盗窃案件进行时空风险分析, 首先对其进行时间特征分析, 发现案件在时间维度上存在有明显的聚集区. 其中案件高发时段位于中午 12:00 前后和下午 17:00 至晚 21:00 以及凌晨 0:00. 其次, 对案件的空间分布特征进行分析, 如图 3 中所示. 图 3 为案件坐标归一化后的位置分布, 从中可以看到在空间层面上案件的分布存在着若干个明显的聚类团簇.

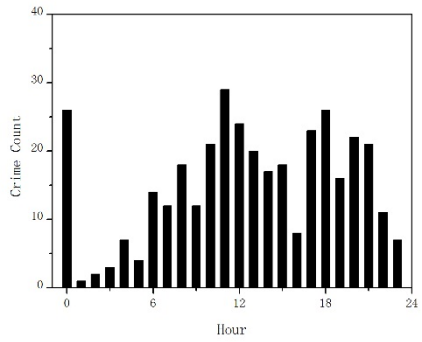


图 2 案件时间分析

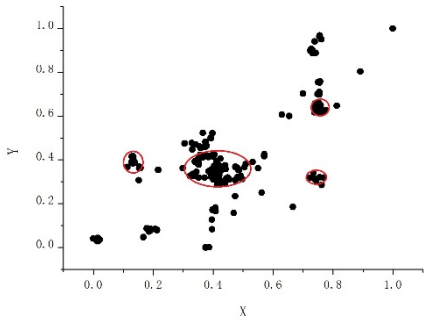


图 3 案件空间分析

1) 对 N 个事件 $S_{t,v}$, 分别利用公式(1)和(2)计算每个事件 $S^i_{t,v}$ 至其他 N-1 个事件的时间距离 δt_{ij} 和空间距离 δl_{ij}, (j=1,2,...,N,$j\neq i$);

$$\delta t_{ij} = t_i - t_j \tag{1}$$

$$\delta l_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{2}$$

2) 设定不同的时间与空间临界值 ΔT_k 与 ΔL_k(k=0,1,...,m), 根据事件 $S^i_{t,v}$ 与事件 $S^j_{t,v}$ 的时间距离 δt_{ij} 与空间距离 δl_{ij} 计算位于不同时空临界值范围内的事件对 Γij(i=0,1,...N; j=0,1,...,N,$j\neq i$)的数量, 最终形成一个 $m\times m$ 的矩阵 φ;

3) 采用蒙特卡罗仿真方法进行检验, 即先假定事件之间的时间距离与空间距离呈相互独立状态, 在保持事件的空间信息不变的基础上, 随机重排时间信息并按步骤(1)-步骤(2)重新统计矩阵 φ 中各要素的数量, 然后计算结果的置信度 p, p=1-n_e/(n_s+1). 其中 n_e 为对应时间临界值 ΔT_k 与空间临界值 ΔL_k 范围内实际事件对数量大于模拟事件对数量的次数, n_s 为蒙特卡洛模拟的次数. 模拟次数越多, 则结果的置信度越高.

利用该方法, 对盗窃案件的时空聚集性进行分析. 从表 1 中结果可见, 盗窃案件具有明显的时空聚集性, 即案发位置周边 100 米范围内, 在案发后接下来的一周时间内再次发案的可能性要大大超出随机发案的概率, 并且不同位置的发案风险是不同的, 其中在原案发位置再次发案的风险达到了 779%, 而在 100 米范围内的发案风险则下降到了 315%. 由此可见, 2011 年发生于北京市的盗窃电动车案件有着时空聚集性的特点, 时空风险半径分别达到了 7 天和 100 米, 并且随着案发后时间的延续和空间距离的延伸而不断下降.

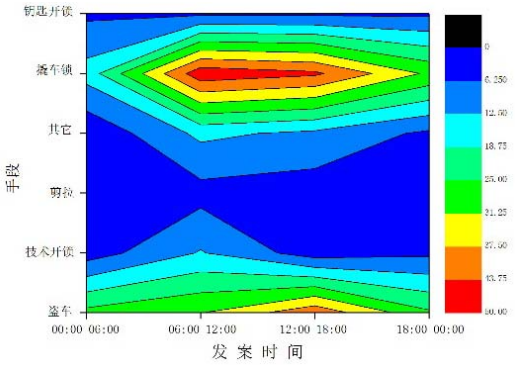
表 1 盗窃案件的时空聚集性分析结果

		时间距离				
		(0,7]	(7,14]	(14,21]	(21,28]	(28,35]
空间距离	0	**7.79**	3.52	2.36	1.72	1.34
	0-100m	**3.15**	2.27	0.77	2.32	0.77
	100-200m	2.60	1.30	2.62	1.34	2.15
	200-300m	1.57	1.63	0.00	1.13	2.32
	300-400m	3.28	2.80	1.20	1.67	0.86
	400-500m	**3.70**	**3.92**	**3.46**	**3.72**	2.24
	500m 上	0.96	0.98	0.99	0.99	1.00

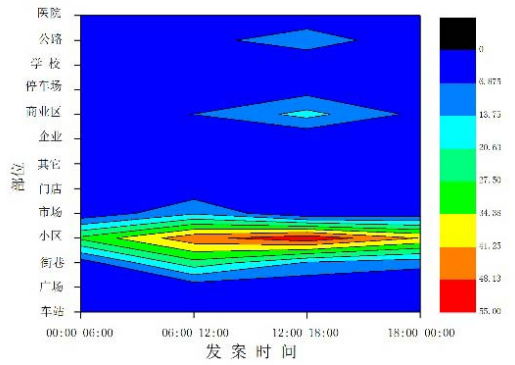
2.3 关联规则分析

在北京市盗窃类案件的数据中, 作案手段分为了六类(盗车、技术开锁、剪拉、撬车锁、钥匙开锁、其它手段), 作案部位分为了十三类(医院、公路、门店、学校、停车场、商业区、企业、市场、小区、街巷、广场、车站、其它), 发案时间则精确到了时段. 然而从案件本身的时间性来看, 报警时间与实际案发时间往往存在着一定的滞后, 即失主在物品被盗一段时间后才会发现并选择报警, 因此, 以案件信息库中的发案时间作为实际案发时间会对结果带来一定的偏差. 为此, 本文调整了时间分析的尺度, 将犯罪时间划分为了上午(06:00-12:00)、下午(12:00-18:00)、前夜(18:00-24:00)、后夜(24:00-06:00)等四个时段, 从而减小了时间尺度过细对分析结果精确度的影响.

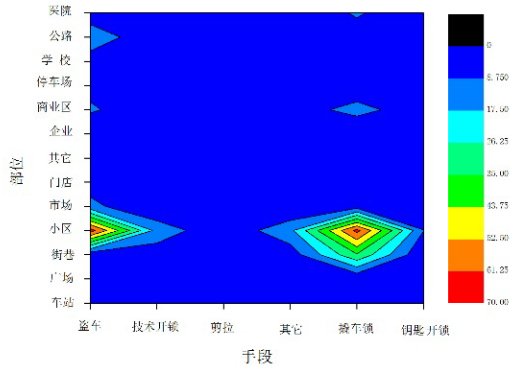
首先, 以作案时间、作案手段和作案部位作为数据维度, 分析不同维度下的作案规律. 图 2 分别给出了不同数据维度下案件数量的热力图, 从图中可以看出, 在盗窃类案件中, 采用撬锁和盗车两种手段的案件较多, 并且前者主要集中发生在白天时段, 而后者则主要集中在下午时段(图 4(a)); 在作案部位上, 发生在小区的案件较多, 并且作案时间主要集中在正午前至傍晚时段(图 4(b)), 而作案手段也基本上表现为撬锁和盗窃两种手段(图 4(c)).



(a)发案时间-作案手段



(b)发案时间-作案部位



(c)作案手段-作案部位

图 4　不同数据维度的发案频率分布热力图

对盗窃类案件信息采用关联分析方法进行分析，分析方法采用经典的 Apriori 算法，该算法的原理如下:

设 $I=\{i_1,i_2,...,i_d\}$ 为全部项的集合，任务相关的数据 D 是数据库事务的集合，其中每个事务 T 是项的集合，使得 $T\in I$. 通过对数据库进行扫描，累计每个项的计数，并收集满足最小支持度的项，找出频繁 1 项集的集合，记为 L_1，然后以 L_1 为基础再次进行数据库扫描寻找频繁 2-项集 L_2，继而是 L_3，依次循环，直到不能再找到频繁 k 项集为止. 确定频繁项后，通过计算最小支持度和最大置信度来确定相应的关联规则. 设 A、B 均为 T 的一个项集并当且仅当 $A\in T$、$B\in T$，$A\cap B=\varnothing$，关联规则表示为 $A\rightarrow B$，则最小支持度为 D 中事务包含 $A\cup B$ 的百分比，表示为概率 $P(A\cup B)$；置信度为 D 中包含 A 的事务同时也包含 B 的事务的百分比，表示为 $P(B/A)$. 同时满足最小支持度阈值(*min_sup*)和最小置信度阈值(*min_conf*)的规则称为强规则[5,6].

本文将最小支持度初步设定为 1%，最小置信度设定为 50%，利用数据挖掘软件 Weka 进行关联规则挖掘，每次通过调整最小支持度和最小置信度的阈值得到分析结果，并将挖掘得到的关联规则结果进行排序，将重复出现次数较多的规则提取出来. 最终发现重复出现的一共有两条强规则.

规则 1. (发案时间="上午")∧(作案部位="街巷")→(作案手段="撬车锁")(4%, 67%)，该规则表明在上午发生的盗窃类案件中有 4%的案件发生在街巷并且作案手段为撬车锁；而发案时间是在上午(06:00-12:00)并且作案部位是在街巷的盗窃电动车案件中，有 67%的案件采用的作案手段是撬锁;

规则 2. (发案时间="前夜")∧(作案手段="盗车")→(作案部位="小区")(4.6%, 60%)，该规则表明发生在前夜的盗窃类案件中有 4.6%的案件采用的作案手段是盗车并且发生在居民小区内；发案时间是在前夜(18:00-24:00)并且作案手段为盗车的盗窃类案件中，有 60%的案件发生在居民小区.

3　结论

作为情报引导警务的重要组成部分，公安情报分析对合理的引导警务决策、分配警力资源、提升防控打击效能具有十分重要的作用. 本文从公安业务特点出发，构建了面向治安防控的公安情报数据挖掘框架，并以北京市 2011 年盗窃案件数据为例进行了分析. 本研究对于建立以面向公安实战化的情报分析模型与分析体系具有一定的实际意义，对于拓展数据挖掘的应用范围和领域具有较大的借鉴作用.

参考文献

1 金光,钱家麒,钱江波,黄蔚民.基于数据挖掘决策树的犯罪风险预测模型.计算机工程,2003,29(9):183–185.

2 熊允发.公安情报分析中决策树方法的应用.中国人民公安大学学报(自然科学版),2008,(1):48–50.

3 余先虎.犯罪行为关联分析研究.宁波工程学院学报,2013,25(3):36–40.
4 许阳泉.改进型 Apriori 算法在犯罪关联分析中的应用.软件导刊,2013,12(11):68–70.
5 包晔.关联分析技术在刑事犯罪分析中的应用.数学的实践与认识,2011,41(20):149–154.
6 叶文箐,吴升.基于加权时空关联规则的公交扒窃犯罪模式识别.地球信息科学,2014,16(4):537–544.
7 李代超,吴升.面向不同主题的犯罪大数据可视化分析.地球信息科学,2014,16(5):735–745.
8 吴文浩,吴升.多时间尺度密度聚类算法的案事件分析应用.地球信息科学,2015,17(7):837–845.
9 陆娟,汤国安,蒋平.犯罪均值频率—一种犯罪时间分布的测度指标.中国人民公安大学学报(社会科学版),2012,(3):152–156.
10 Knox G. Epidemiology of childhood leukaemia in Northumberland and Durham. British Journal of Preventive & Social Medicine, 1964, 18: 17–24.
11 Townsley M, Johnson SD, Ratcliffe JH. Space time dynamics of insurgent activity in Iraq. Security Journal, 2008, 21: 139–146.
12 Ratcliffe JH, Rengert GF. Near-repeat patterns in Philadelphia shootings. Security Journal, 2008, 21: 58–76.