《文本挖掘的“智慧公安”应用的研究》

摘要：

为升级城市公安系统，打好智慧公安的创新战，提高案件侦破率，简化案件串并，便于案件信息录入，该论文结合文本挖掘技术，提出一个公安犯罪案件信息挖掘模型，这个模型分为案件要素提取框架和案件自动分类框架。论文中运用正则表达式、Jieba 分词、支持向量机分类 (SVM)，针对性的提取公安案件的特点信息点，进而便于公安机关的案件串并。

关键词：文本挖掘、SVM、jieba 分类、智慧公安

《search on the application of "intelligent public security" based on Text Mining》

Abstract:

In order to upgrade the urban public security system, improve the case detection rate, simplify the case series and parallel, and facilitate the case information entry, this paper combines the text mining technology, This paper proposes an information mining model of public security crime cases, which is divided into two parts: the framework of case elements extraction and the framework of case automatic classification. In this paper, regular expression, Jieba segmentation, support vector machine classification (SVM) are used to extract the characteristic information points of public security cases, so as to facilitate the public security organs' case concatenation.

Key words:Text mining、SVM、jieba classification、Intelligent public security

目录

一、研究智慧公安的意义

近些年来，随着互联网的飞速发张与广泛应用，案件信息以每年百万条的速度增长，但是依靠传统的案件处理方式，效率不高并且很难挖掘出有价值的信息与线索，因此公安机关必须加强对各类信息的全面整合、综合分析和预警监测，不断提高搜集情报、侦查破案、处理重大警情的能力。智慧公安有必要紧跟互联网时代，创新公安应用。智慧公安利用“物联网”技术进行身份、车牌、人脸、手机、指纹和声音等信息录入采集，传输至平台，与公安已有数据资源碰撞。进行人员管控，大数据深层挖掘和智能研判应用，实现对人员的全方位、立体式管控，提高社会治安防控水平，提高民众社会治安满意度。目前公安系统中传统的信息管理正面临着重大的变革，对健全的公安管理体系的需求日益强烈，就公安基层工作而言，意义深远。

基于文本挖掘，该论文的挖掘目标是借用大数据的技术，让公安系统简化案件文本的筛选和信息点提取，便于公安案件的串并，聚类，和案件系统的录入等。为警务人员减少繁琐的工作量，让公安系统更加智能化，打破传统警务系统的瓶颈和局限，为城市的社会公安领域增添多一道防线。

二、模型框架概述

文本挖掘又被称为文本发现，研究者可以基于文本挖掘来深入了解不同的领域，而在具体的流程上涉及到许多方面的内容，如特征值提取、文本分词、文本分类、文本聚类……在该论文中，我们基于文本挖掘，构建了一个公安犯罪案件信息挖掘模型。模型分成了两个框架，案件要素提取框架和案件自动分类框架。框架结构图如图 1 所示。(【1】)

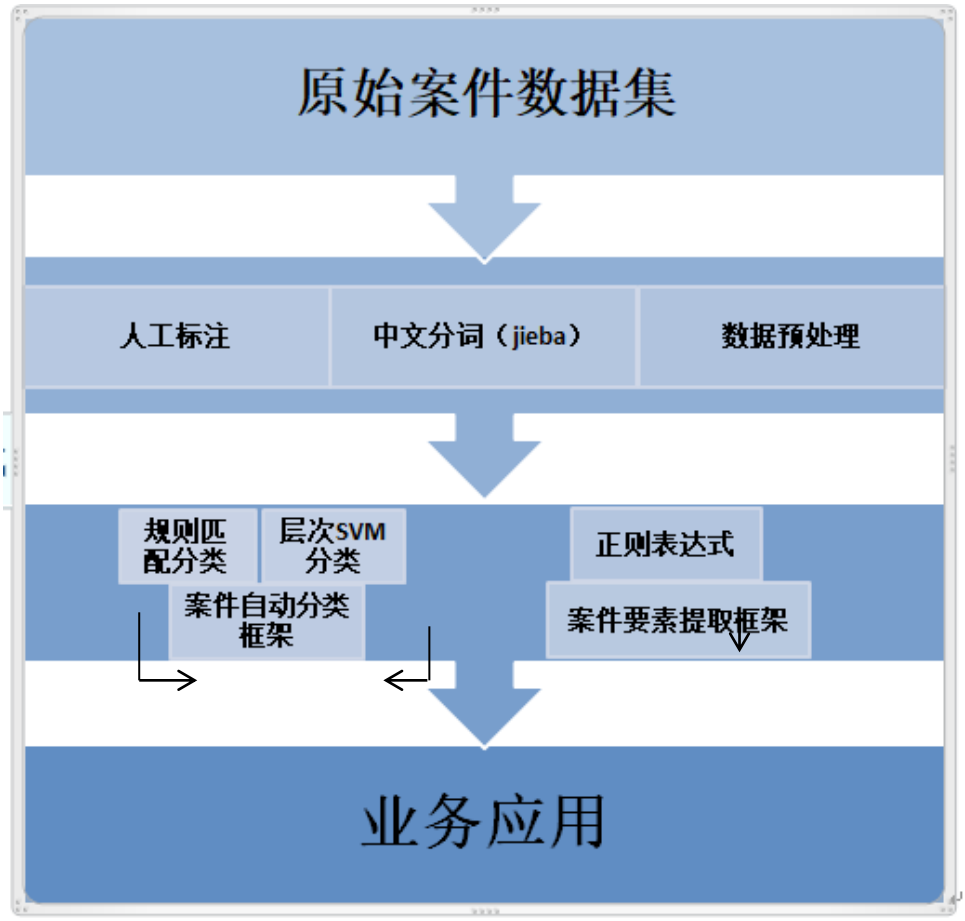

图 1 公安犯罪案件信息挖掘模型框架结构图

三、提取案件要素

（1）提取案件要素的基本内容

所谓提取案件要素，就是在案件文本中提取关键的信息点。提取的案件信息点的结构更短小，内容更丰富。在公安案件中，这关键的信息点包括作案时间，作案地点，作案对象，作案类型，电话号码，作案结构……（【2】）经过文本挖掘之后，公安部门可以根据提取到的信息点对案情进行深一步的了解和处理。例如“2020 年 2 月 14 日，上海市公安局经侦总队会同松江公安分局，联合市场监管部门成功侦破疫情期间本市首起特大生产、销售假劣口罩案件，抓获犯罪嫌疑人邹某等 6 人，捣毁制假窝点 1 处，现场缴获假冒品牌口罩 10 余万枚，以及大

量制假原材料和制假工具，涉案金额 100 余万元。此案也是疫情期间本市首起特大制售假劣口罩案（【3】）。”在本案件中，所提取的涉案时间为“2020 年 2 月 14 日”，作案内容为“生产和销售假劣口罩，捣毁制假窝点 1 处，缴获假冒品牌口罩 10 余万枚、大量制假原材料和制假工具”，作案嫌疑人为“邹某等 6 人”，涉案金额为“100 余万元”。在上述提取的信息点中，有一些信息如“生产和销售假劣口罩” 、“大量制假原材料和制假工具”，对案件的串并非常有帮助，在补充案件串并方面十分重要。

我们在文本过滤这一块，用的就是正则表达式。用于检索、快速匹配、替换某些符合某个模式的文本和数据。在现实生活中，例如公安领域，警察们常常需要提取储存介质的手机号、银行卡号、文件名、URL 网址等数据，或是针对图片、音视频的数据分析，正则表达式就可以达到快速匹配相关数据的高要求（【4】）。编程时要格外注意正则表达式的容错性，尽量全面的考虑案件文本的各项数据的提取规则，根据规则进行对应的编码修改，整合成最后的正则表达式。用最后得到的正则表达式去提取案件文本相对应的信息点。信息点提取例子如下表 1 所示。

案件文本语句	匹配提取的信息点
1、一名醉汉拿着菜刀在河西市场砍伤 3 名路人后，驾车离去，根据监控，车牌号码为粤 A3XXXC	粤 A3XXXC（车牌号）
2、一位妇女来派出所报案称自己的 24 岁的女儿已经失踪两天了（黄*苏，女，4408755637090****）	4408755637090****(身份证号码)
3、一男子在金银首饰店抢劫价值 500000 元的金项链，还抢走了店员价值 2000 的苹果手机一台	
502000 元（涉案金额）	

表 1 信息点提取例子

（2）案件信息点提取的应用

1、在公安日常编辑的文件中简化插入信息录入，简化查找案件信息

根据正则表达式，基于特定的字符串，查找特定文字、删除字符或短文本、指定更换文本内容或字符。

2、利于案件串并

串并案件就是将案情性质相近或相同，有规律、作案手段相似等案件进行分析对比，而基于文本挖掘得到的信息点，通过信息点的分词，聚类，将同组的案件进行串并，便于公安机关侦查，提高案件侦破律。

3、在电子储存媒介或图片集合中发现并提取相关信息点

根据正则表达式，快速匹配目标文件，例如图片的特点.jpg 数据。

四、案件文本挖掘预处理

在案件分类之前 ，需要做好案件文本挖掘预处理工作，文本挖掘一般包含中文分词处理、向量化、特征降维等几个过程，本文中的文本挖掘预处理的具体流程如下图 2 所示。

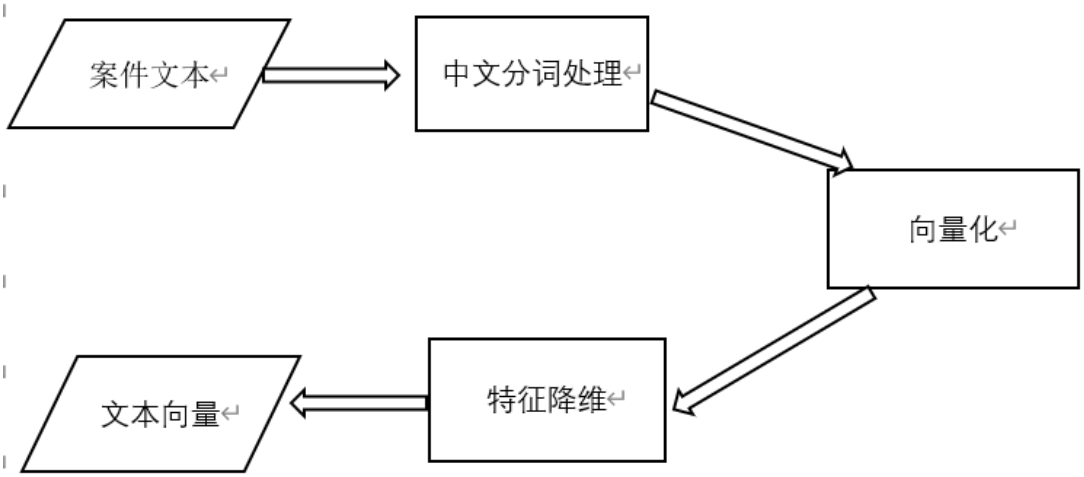

图 2 文本挖掘预处理流程图

（1）中文分词处理

中文分词处理是能将中文文本中的词语正确分开的一种技术，是文本挖掘预处理的基础。本文采用分词效果较佳和运行速度较快的“Jieba”分词。“Jieba”支持自定义词典（支持载入新的词典或者更新自带的词典），这个特点有助于对案件问的进行分词处理，具有实际应用的价值，案件文本中包含了不少的公安领域专业词汇和地区、道路等名称词汇，这些词汇有很大的用处，可将这些特殊词汇加入自定义词典中同时也将区分度差的高频词汇加入到停用词词典中，可大大提升分词的效果。

（2）向量化

将案件文本转换成计算机能够理解的表示形式。本文采用向量空间模型(Vector Space Model，VSM)表示案件文本。该模型的主要思想是：把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量问的间的相似性。对 于所有的文档类和未知文档 ，都可以用此空间中的词条向量（$T1$，$W1$，$T2$，$W2$,…，Tn，Wn）来表示(其中Ti为特征项；Wi为Ti的权重 ）。一般需要构造一个评价函数来表示词条权重，其计算的唯一准则就是要最大限度地区别不同文档。本文采用$TF-IDF$方法来计算特征权重，$TF-IDF$权重是向量空间模型中应用最多的一种权重计算方法，它以词语作为文本的特征项，每个特征项的权重由 TF 权值和 IDF 权值两个部分组成。

（3）特征降维

特征降维就是用来减少维度，去除过拟合现象的方法。特征降维分为特征选择和特征抽取两种。通过降维后生成的特征集合的每一个元素具有更强的代表性，降维能减少计算资源的耗费。本文采用的特征降维的方法：建立停用词集合，分词过程中筛去停用词；建立公安领域的同义词词典，经过同义词替换减少特征集合的维度；在分词过程中，对分词的结果进行词性标注，然后筛去词性为人名的词。(【5】)

五、案件分类

案件分类功能模块的工作流程如下图 3 所示：

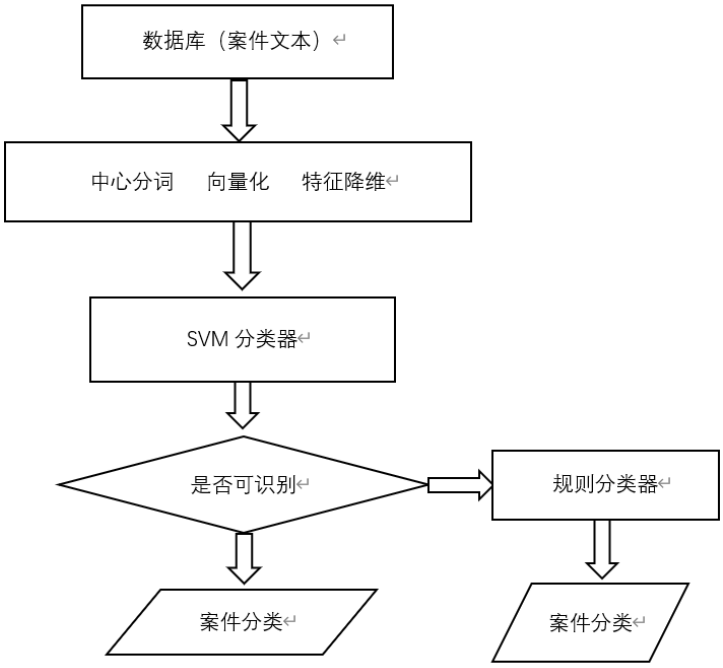

图 3 案件分类流程图

（1）基于 SVM 的文本分类

算法支持向量机（ *SVM* ）是建立在统计学习理论基础上发展而来的一种机器学习方法，它基于小样本学习，结构风险最小化等统计学习原理，将原始数据集合压缩到支持向量集合，学习得到分类决策函数。(【6】) 其具有坚实的理论依据和成功 的实践经验，在许多领域（如手写数字识别，物体识别和文本分类）得到应用，此方法的基本思想是构建一个超平面作为决策平面，使正负模式之前的间距最大。现实 中，每天被录入的案件种类繁多 ，常见的有交通事故、电信诈骗、人室盗窃等 ，比较少见的有吸毒赌博、集资诈骗等。因此，本文根据总体 案件文本类别数量不均衡的特征，采用了自动调整类别权重平衡模式 ，类别权重值与该类出现的频率成反比。设当前样本总数为 n1，案件类别共有 n2 类 ，属于类别 X 的案件出现次数为 $count(x)$，别 X 的权重值计算公式为

$$TF\left(X,d\right)=\frac{count\left(x,d\right)}{size\left(d\right)} \tag{I}$$

$$IDF=log\left(\frac{n1}{docs\left(x,d\right)}\right) \tag{II}$$

$$\begin{aligned} &TF-IDF(q,d) \\ &=sum\left\{i=1..k\,|\,TF-IDF\left(X\left[i\right],d\right)\right\} \\ &=sum\left\{i=1..k\,|\,TF\left(X\left[i\right],d\right)*IDF\left(X\left[i\right]\right)\right\} \end{aligned} \tag{III}$$

其中 d 表示文档， *count* 表示出现次数， *size* 表示文档的总词量， *docs* 表示文件数

（2）规则分类与 SVM 分类结合

在 SVM 分类过程中，分类器的置信度是一个值得重视的参量。在决策过程中，对测试样本分别计算各个子分类器的决策函数值 ，并选取分类器决策函数值最大所对应 的类别作为 测试样本 的预测类别。多采用以决策函数值作为衡量置信度大小的标准，在预测时，记录了每条测试样本的每个 SVM 分类器的决策函数值。本文经过分析研究和实验，对满足以下情况的案件样本拒识：

1、各个分类器的决策函数值均为负数；

2、出现三个及以上的分类器的决策函数值为正数；

3、仅一个分类器的决策函数值为正数，但其值很小，小于 1；

4、出现两个分类器的决策函数值为正数，且数值很接近，相对平均偏差≤5%；

5、通过对分类器的判决结果进行基于决策函数的置信度评估，拒识置信度水平相对较低的决策结果，接受置信度水平较高的决策结果。对于被拒识的案件，本文调用规则匹配分类器确定其类别。

规则匹配分类器是依据产生式规则的思想，建立事实数据库并设计规则库 ，基于现有的规则库示推理过程和行为。基于规则的分类器所产生的规则集的两个重要性质：⑴互斥规则 ：如果规则集 R 中不存在两条规则被同一条记录触发，

则称规则集 R 中的规则是互斥的。这个性质确保每条记录至多被 R 中的一条规则覆盖。⑵穷举规则： 如果对属性值的任意组合，R 中都存在一条规则加以覆盖，则称规则集 R 具有穷举覆盖。这个性质确保每一条记录都至少被 R 中的一条规则覆盖。如果规则集不是互斥的，那么一条记录可能被多条规则覆盖，这些规则的预测可能会相互冲突，那么可以采用有序规则和无序规则解决处理。

本文的规则匹配分类器是一个应用规则库(含 875 条规则，可进行增删改操作)，利用逻辑关系匹配的方法检验案件文本信息的工具。规则库有多个属性列，分别为序号列，关键词列，排斥词列，类别名称列，上级类别列。规则以 *IF…THEN…* 的形式出现，*IF* 所带的是前件(条件)，*THEN* 所带的是后件（结论 ），多个条件是通过逻辑运算 *AND*，*OR*，*NOT* 组合成复合条件 ，当完全满足条件才能推出对应的结论。

当给定一个案件，规则匹配分类器整体的匹配分类过程如图 4 所示：如果遍历了所有规则仍没有匹配成功，就说明对该案件分类失败。

规则匹配分类器依赖于人工经验积累编写而成的规则库，适用于识别出现频率低、具备明显特征词的案件 ，如“纠纷”、 “ 举报”、“涉毒”等类别的案件，对于逻辑关系复杂的案件类别容易产生错误，而且分类速度会很慢，因为规则库中规则数量较多，需要对每一条待分类案件进行顺序遍历规则直到匹配符合，匹配每一条规则还需迭代各个关键词和排斥词，所以，单条案件分类速度远慢于支持向量机分类的速度。综合各种优劣因素本文决定把支持向量机作为主要的分类方法，规则分类为次要方法对案件进行分类，这样便保证了案件分类速度和分类准确率 。

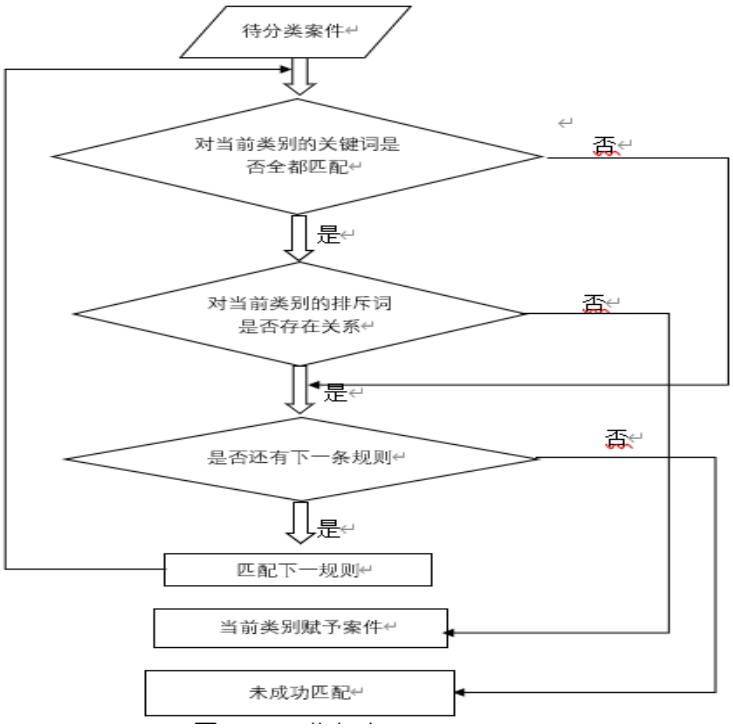

图 4 匹配分类过程

六、实验与评估

（1）提取信息的应用与评估

本文采用 pycharm3.7 作为数据挖掘和分析的工具，此实验以 2020 年 2 月 03 日至 2020 年 2 月 07 日的 1026 条案件样本作为实验数据，下面来对信息提取的功能进行测试。

评价的指标是：实验结果要素 k 被正确提取的比例 Pk

$$Pk = \frac{ap}{(ap+bp+cp)} \qquad (IV)$$

式中：ap 表示的是要素 k 被正确提取的案件个数，bp 表示的是要素 k 被错误提取的案件个数，cp 表示的是要素 k 存在但没有被提取的案件个数。各信息被正确提取率如下图 5：

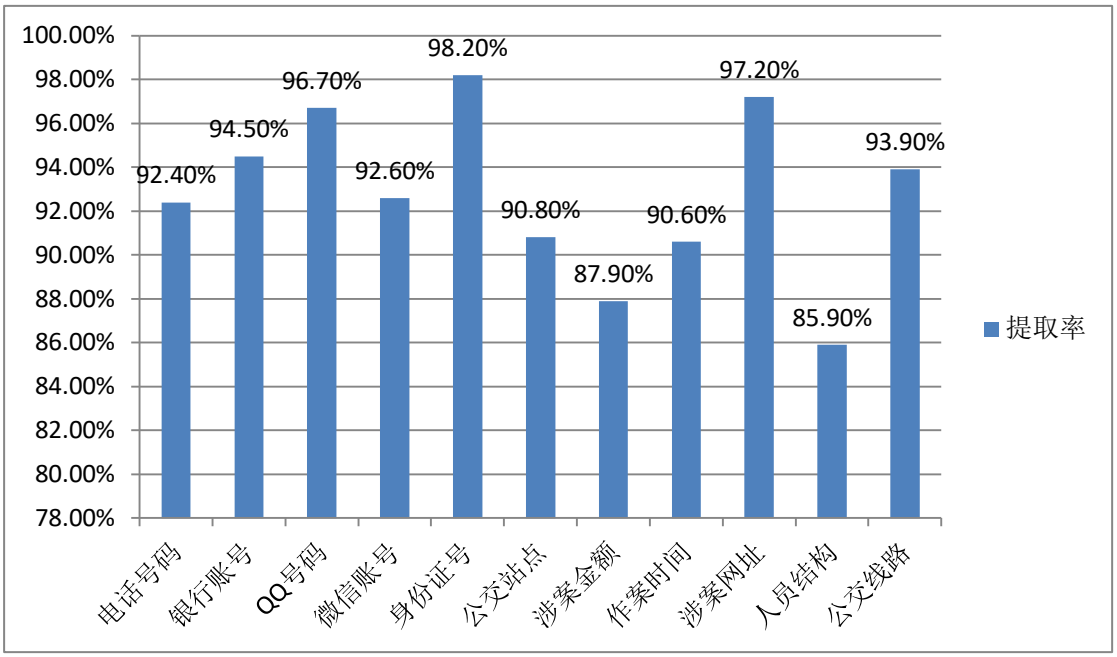

图 5 正确提取率

评估信息提取的效果

1. 电话号码、QQ 号码、微信账号、银行卡账号的正确提取率基本全面、完整。这四个信息的正确提取率都达到了 92.6%以上，但是这些信息没有办法识别是来自嫌疑人还是受害者，也不知道这些号码和账号是否是注册者本人在使用，后续这部分的研究将借助语义分析的方法改进。(【7】)

2. 涉案金额提取基本准确，不过有些案件的涉案金额会包含案件实际发生的金额和非实际发生的金额。下面来解释一下这两者的区别：实际发生的金额如："受害者受害过程中向诈骗犯转了 1 万元"，非实际发生的金额如："对方提出要受害者转账 15600 元，受害者转了 1 万元后发现自己被骗没有转到剩下的金额"，但其中的 15600 也会被提取出来，这种情况下提取的是非有效金额 。"

3. 涉案时间基本可以提取到，但是目前还不能具体提取到报案时间和案件发生时间，需要做进一步的改进。

（2）案件分类的应用与评估

本案件自动分类框架是我们模型的核心内容，可以识别不同性质的案件、随时在公安系统里导入、更改、删除信息或样本，还得到不同的分类器……论文中研究的对象以公安事件的抢劫、电子诈骗、盗窃等涉及金额的案件为主，通过分类框架的规则分类器，我们通过提取关键词，用关键词来判定案件类型，对于一

些非财产的案件我们通过代码编程拒以识别。

在实验中，以准确率作为评价指标的定义公式如下，其中，sp 代表正确的正比例的个数 ，lp 表示错误的正比例的个数。：

$$Precision = \frac{sp}{(sp+lp)} \quad (2)$$

在论文的实验里，我们采用十折交叉验证法（10-fold cross validation），就是将数据集分成十份，轮流将其中 9 份做训练、1 份做验证，10 次的结果的均值作为对算法精度的估计，一般还需要进行多次十折交叉验证求均值（【8】），例如：10 次 10 折交叉验证，以求更精确一点。这个方法的优势在于：同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次。

结合十折交叉验证法，对普通 SVM 分类，层次 SVM 分类 ，规则与层次 SVM 结合三个的分类效果进行对比，结果如下图 6 所示：

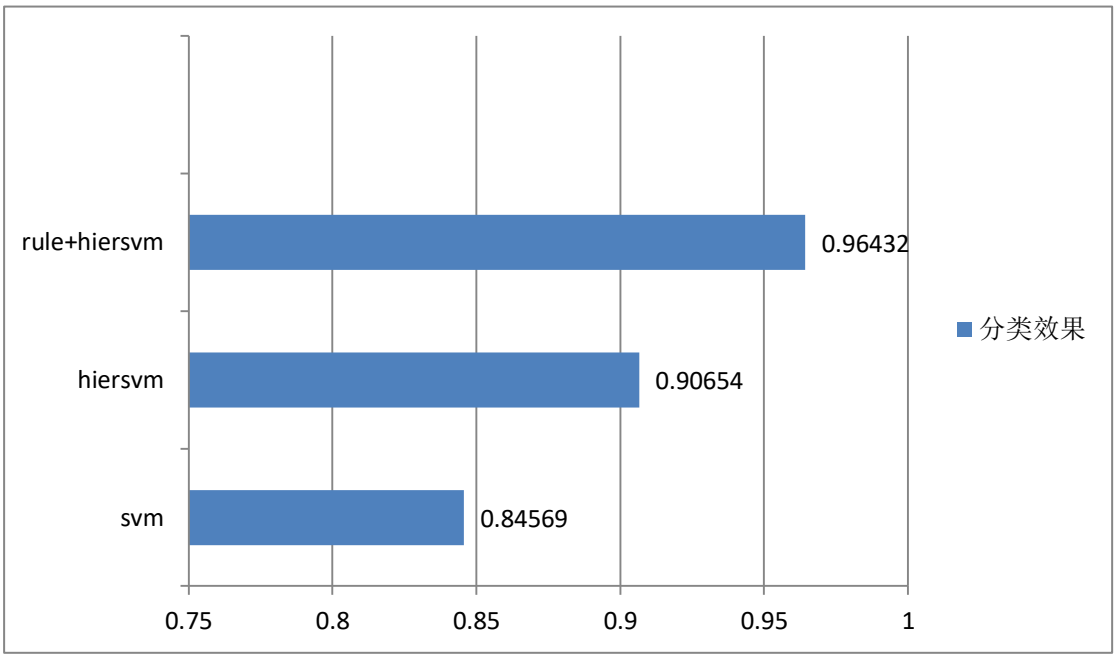

图 6 分类器交叉验证实验结果

根据图 6，论文得到 SVM 分类器均采用了以 $TF-IDF$ 方法计算词条权重，其中，IDF 为逆文本频率的简称，将其作为度量来评估一个词语是否具有普遍重要性. 将停用词阻挡，将人名和同义词进行替换，把错误样本的惩罚因子设为 1。

从图 6 中的数据可以看出层次 SVM 分类器比普通 SVM 分类器的分类准确率提高了 6.085% ，规则与层次 SVM 结合的分类相对于层次 SVM 分类高出了

5.778%，从实验结果来看，数量较少的非财产案件更适合采用规则匹配分类。由于本实验运用的是比较简单的层次结构，（【9】）在理想理论上来说 ，更复杂有层次的实验数据会让层次分类器的分类效果更明显，在现在的科学基础上，规则与层次 *SVM* 结合的分类效果相对更优 。

七、模型猜想

论文中构建的公安犯罪案件信息挖掘模型可应用于社会治安 ，通过案件串并来预测威胁社会安全案件的趋势，减轻公安系统工作人员的工作负担，为人们创造更加安全的社会环境。

八、参考文献

【1】贾俊凯. 公安业务文本信息挖掘的研究与实现[D].东华大学,2011.

【2】魏文燕,吕鑫,高琰.文本挖掘技术在公安领域案件分析中的应用[J].湖南警察学院学报,2017,29(03):98-104.

【3】严打口罩制假售假、野生动物非法售卖 上海警方侦破涉疫案件 130 起 http://www.shanghai.gov.cn/nw2/nw2314/nw2315/nw31406/u21aw1439675.html

【4】李军建.正则表达式在公安业务数据分析中的作用[J].网络安全技术与应用,2015(12):85+87.

【5】赵晖. 支持向量机分类方法及其在文本分类中的应用研究[D].大连理工大学,2006.

【6】赵行. SVM 分类器置信度的研究【 D】 . 北京： 北京邮电大学，2010.

【7】宁琳 .一种基于句法规则的文本挖掘技术的设计Ⅱ】_现代情报，2016(2)：140—144.

【8】十折交叉验证法的深入理解 https://blog.csdn.net/jp_zhou256/article/details/85248578?locationNum=9&fps=1

【9】CSDN 博主「jp_zhou256」的原创文章 https://blog.csdn.net/jp_zhou256/java/article/details/85248578