**硕士学位论文**

**文本挖掘技术在公安领域案件分析中的应用**

Times New Roman Application of text mining technology in the field of public security

学科专业 一级学科名称（三号宋体）

学科方向 二级学科或三级学科名称（三号宋体）

作者姓名 魏文燕

指导教师 高琰教授

中 南 大 学 （小三号宋体）

年 月 （小三号宋体，填阿拉伯数字）

学位论文原创性声明

本人郑重声明，所呈交的学位论文是本人在指导教师指导下进行的研究工作及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他教育机构的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

作者签名： 日期： 年 月 日

学位论文版权使用授权书

本学位论文作者和指导教师完全了解中南大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交学位论文的复印件和电子版；本人允许本学位论文被查阅和借阅；学校可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用复印、缩印或其它手段保存和汇编本学位论文。

保密论文待解密后适应本声明。

作者签名： 指导教师签名

日期： 年 月 日 日期： 年 月 日

（填阿拉伯数字） （填阿拉伯数字）

样例8 中文摘要示例

**文本挖掘技术在公安领域案件分析中的应用**

**摘要：**文本挖掘是一项从大量的、非结构化、语义丰富的文本中获取人们需要或感兴趣的信息的过程，其核心技术有文本预处理、聚类、分类、信息抽取、信息检索等。为辅助民警办案，提高搜集情报、侦查破案的能力，结合公安领域案件文本数据的特点，本文将文本挖掘技术应用于公安领域案件的分析处理中，提出了一个基于文本挖掘技术的案件信息处理框架。框架主要研究对象为民警笔录的案件情况描述的文本，以大量保存在数据库中的案件文本作为实验数据，实现案件分类、案件信息要素提取的目的。本文的研究内容总结如下：

（1）针对案件类别之间存在的结构关系，构建了层次SVM分类模型，称为hierSVM。

（2）提出了hierSVM与规则分类器相结合的分类策略。

（3）利用正则表达式提取案件信息要素。

（4）识别银行账号信息的属性，为作案人员或受害人员。

图X幅，表X个，参考文献X篇（四号宋体）

**关键词（3-8个, 四号黑体）：**…………；…………；…………（四号宋体）

**分类号：（1-2个, 四号黑体）：**…………；…………（四号宋体）

样例9 英文摘要示例

**Title （三号Times New Roman字体）**

**Abstract** （四号**Times New Roman**字体，加粗）：英文摘要内容（四号Times New Roman字体）**………………………………………………………**

**……………………………………………………………………………………………**

**……………………………………………………………………………………………**

**Keywords (3-8，**四号**Times New Roman**字体，加粗 **): …………；…………；…………**（四号Times New Roman字体）

**Classification（1-2，**四号**Times New Roman**字体，加粗**）: …………；…………；…………**（四号Times New Roman字体）

样例10

**目 录（三号黑体）**

1 （第1章）绪论………………………………………………………………1

1.1 （第1章第1节）题名…………………………………………………3

1.1.1（第1章第1节第1小节）题名…………………… ……………7

1.1.2（第1章第1节第2小节）题名…………………………………15

2 （第2章）题名…………………………………………………… ………21

2.1 （第2章第1节）题名…………………………………………… …25

2.1.1（第2章第1节第1小节）题名…………………… …… ……27

5 （第5章）结论……………………………………………… ……………72

参考文献………………………………………………………………… ……91

附录1……………………………………………………………………………95

附录2……………………………………………………………………………98

攻读学位期间主要研究成果………………………… ………… …… ……101

致谢……………………………………………………………… ……………117

（注：除“目录”两字外，其余字体均为小四号宋体）

第一章 引言

1.1 研究背景

随着社会的发展和信息科学的广泛应用，各行各业都保存了许多与业务相关的数据资料，公安领域也不例外。近年来，随着公安信息化建设的推进，大批软件的开发和应用，大量公安信息进入数据库，形成综合信息查询平台，在公安业务查询和公安信息服务等工作中得到了广泛应用。如计算机的广泛应用将沉睡多年的基础数据、台帐翻新变成活的信息，并通过信息共享，有效地提高了整个公安战斗力，较大地实现了科技强警的目的。然而，这些第一资源未能充分有效地得到开发，案件数据的处理手段基本还停留在查询、基础统计的层面上，返回一些较为浅显的信息。因此民警面对大量的数据，仅能依靠自己的洞察力和判断获取深入的信息。如何让这些系统充分发挥作用，真正实现科技强警的战略目标，为社会发展和经济建设保驾护航显得尤为重要。

数据挖掘技术是从大规模的数据集中抽取或挖掘隐含的、实现未知的、潜在有用的信息和知识的有效方法和途径。很多领域都借助数据挖掘技术得到成功的应用结果，如医疗、金融业、零售业、电商、教育机构等。应用程度较为广泛的数据挖掘技术有数据清洗、聚类、分类、预测、关联分析和推荐等。数据挖掘技术应用的数据对象类型有结构化表格、非结构化的文本、图像、视频等。对于公安领域，可以利用文本数据挖掘技术，从大量的由记述性语句构成的非结构化案件文本中，挖掘出破案线索、发现犯罪趋势等，使信息化转化为战斗力，提高公安部门的决策能力。

文本挖掘技术是一个对具有丰富语义的文本进行分析从而理解其所包含的内容和意义的过程，已经成为数据挖掘中一个日益流行而重要的研究领域。（袁军鹏, 朱东华, 李毅,等. 文本挖掘技术研究进展[J]. 计算机应用研究, 2006, 23(2):1-4.）其涵盖多种技术，包括信息抽取、信息检索、自然语言处理和数据挖掘技术，主要研究对象有源于网络的网页文本、新闻文稿、购物评价、电影评价等，和源于业务数据库中的金融报表、医疗报告、生物资料等，这些对象具有共同属性：非结构化或者半结构化。目前，已经有许多应用研究取得了很好的结果，如：在生物学中的应用（朱俊, 殷建平, 赵志恒,等. 基于文本挖掘的精子发生各阶段的相关基因/蛋白名称提取[J]. 计算机研究与发展, 2014, 51(6):1352-1358.），在网络舆情信息分析中的应用（李金海, 何有世, 熊强. 基于大数据技术的网络舆情文本挖掘研究[J]. 情报杂志, 2014(10).），在社会媒体中的应用（任彬, 车万翔, 刘挺. 基于依存句法分析的社会媒体文本挖掘方法——以饮食习惯特色分析为例[J]. 中文信息学报, 2014, 28(6):208-215.），在商业市场中的应用（Nassirtoussi A K, Aghabozorgi S, Wah T Y, et al. Text mining for market prediction: A systematic review[J]. Expert Systems with Applications, 2014, 41(16):7653-7670.）。

面对海量的案情文本信息，文本挖掘技术是非常有用的技术，它能从这些海量的案情文本信息中挖掘中隐藏的、对公安业务人员有用的信息（程春惠. 公安犯罪案件文本挖掘关键技术研究[D]. 浙江大学计算机学院 浙江大学, 2010.）。借助该技术可以完成以下工作：采用机器学习的方法（如朴素贝叶斯、决策树、SVM）对案件进行自动分类，提高案件研判的精度；对案件进行聚类，将相似的案件聚集后分析，有助于发现新型作案方式和案件串并关系；抽取案件文本中的关键信息，如人名、地名、嫌疑人的作案电话、账号和涉嫌诈骗案的网址。因此，实施科技强警，掌握和运用人工智能手段，加快工作速度，严密工作精度，能够使公安机关进行更加高效的管理决策，从而大大提高工作效率，提高战斗力。

1.2 国内外研究现状

公安内部传统的数据库信息处理技术已处于相对落后的地位，案情研判精度和破案速度与国家倡导的智慧城市之间的间隙日益突出，人工智能成为缩小距离的关键手段。因此，公安系统所积累的大量业务数据，和公安部门迫切的实际需求吸引着一大波专家学者对此进行深入研究、试验，不断改进算法、推陈出新。在国内，江苏大学的耿丽娟对公安情报数据分类和关联规则进行了重点研究（耿丽娟. Hadoop下基于分类和关联规则的公安数据挖掘研究[D]. 江苏大学, 2014.），将训练数据集构建成树状层次结构，并对原有的KNN分类算法进行改进，采用差分和动态增加类别的方法克服数据集下类域交叉的问题，有效防止了最近邻和次近邻的误判发生，此算法于Hadoop平台下实现，取得分类精度和速度性能双优的结果；他还提出了基于规则加权的多数据集时态关联规则算法，根据时态因素将各个子数据库中的规则进行加权并合成得到最终的全局时态关联规则，算法在犯罪预测的实验说明其不仅适用于多个不同数据库，还在很大程度上将无意义的关联规则缩减，证明了其具备现实意义。

电子科技大学华东，结合结合浙江省公安机关的信息化技术应用，以及在实际犯罪分子管理工作中存在的问题，提出了基于数据挖掘技术的犯罪人员关联分析系统（华东. 犯罪人员关联系统的设计与实现[D]. 电子科技大学, 2013.），该系统实现的主要功能模块有：犯罪人员信息的维护、人员身份核实确认、列控申请、在控访查、在线分派、在线审批以及分析统计等，系统上线后能收集浙江省公安机关下各派出所的犯罪行为信息和犯罪人员信息并进行预处理，通过大量地深入分析犯罪信息，挖掘出这些信息之中隐含着的关联和规律，为侦破案件提供科学的决策依据；除此之外，系统可以对案件和犯罪人员的特征进行归类统计，提取相同或相似的特征，当类似案件出现时，系统可以直接定位到这些案件，对涉案人员和嫌疑人群进行监控，从而可以为串并案件和相似案件提供帮助，有利于破案。

湖南警察学院的唐德权、史伟奇为了正确识别多模态特征融合所表达的多媒体视频，将图像、音频、视频等特征与其伴随的随机噪声融合在一起，形成多模态特征融合的隐马尔科夫链HMC分词处理算法（唐德权, 史伟奇. 公安网络视频监控系统数据挖掘改进算法研究[J]. 警察技术, 2013(6):34-37.），之后针对关键帧和重要特征的处理建立图数据转换和多媒体特征库，末了，采用稠密子图数据方法辨识多媒体场景中的元语义；他们的实验结果证明，其提出的算法准确率比已有的Baseline算法准确率高，能够将多模态特征融合的多媒体视频更加完整地表示，从而可以识别相似的视频场景，并在此基础上开发出了系统原型，成果地说明了模型的正确性和可行性。

S Anuar，A Selamat，R Sallehuddin（Anuar S, Selamat A, Sallehuddin R. Hybrid Particle Swarm Optimization Feature Selection for Crime Classification[M]// New Trends in Intelligent Information and Database Systems. 2015:101-110.）在他们的研究中提出一个混合犯罪分类模型，通过结合人工神经网络（ANN），粒子群优化（PSO）和灰色关系分析（GRA），以确定导致特定的罪行和分类犯罪的重要特征分为三个不同类别；PSO作为特征选择方法，通过选择最重要的特征来减小数据集的维度，数据集维度的减少可以降低复杂性，从而缩短ANN对犯罪数据集进行分类的运行时间；GRA用于对特定犯罪的选定特征进行排名，从而可视化所选择的犯罪属性的重要性。该实验在社区和犯罪数据集上进行，然后，将PSO特征选择的结果与诸如进化算法（EA）和遗传算法（GA）的其它特征选择方法进行比较，评估每个特征选择方法的分类性能，实验中发现：与EA和GA相比，PSO选择较少的特征；分类性能结果表明，与ANN和GA与ANN的组合相比，PSO与ANN的组合产生较小的误差并缩短运行时间。

R Iqbal，MAA Murad，A Mustapha，PHS Panahy，N Khanahmadliravi（Iqbal R, Murad M A A, Mustapha A, et al. An Experimental Study of Classification Algorithms for Crime Prediction[J]. Indian Journal of Science & Technology, 2013, 6(3):4219-4225.）在研究中，将分类应用于犯罪数据集以预测美利坚合众国不同州的“犯罪类别”，研究中使用的是真实的犯罪数据集，是从1990年美国人口普查的社会经济数据，1990年美国LEMAS调查的执法数据和1995年联邦调查局UCR的犯罪数据中收集的。他们比较了两种不同的分类算法，即朴素贝叶斯和决策树，用于预测美国不同州的“犯罪类别”， 实验结果表明，决策树算法执行Nave贝叶斯算法，在美国不同州的“犯罪类别”预测中达到83.9519％的准确度。

B Cavadas，P Branco，S Pereira（Cavadas B, Branco P, Pereira S. Crime Prediction Using Regression and Resources Optimization[M]// Progress in Artificial Intelligence. Springer International Publishing, 2015:513-524.）认为暴力犯罪是影响社会生活质量和经济发展的众所周知的社会问题。因此，它的预测是执法机构的重要任务，且由于预算的限制，资源的优化是非常重要的，在他们的工作中处理两个方面：预测和优化。其剑走偏锋，一改先前学者将犯罪预测作为分类任务的做法，提出使用回归预测每100k人口的暴力犯罪，并通过整数线性规划公式优化警察的分配，同时考虑到先前的预测。对于涉及数据集的高维度的挑战，使用Pearson相关系数应用层次聚类分析。此步骤删除与目标变量较少相关的一部分特征。然后，应用随机森林（RF）载波来基于均方误差中的影响来计算剩余特征重要性。此外，使用smoteR算法预处理数据，以改进对最关键值的预测：极值。其实验表明，在应用SmoteR算法处理稀有极值之后，随机森林比其他评估学习方法表现更好。根据国家报告数据，南部各州被预测为具有最严重的暴力犯罪率。因此，这些州成为在优化期间分配更多警察的州。他们的研究作出了双重贡献，首先，这是第一次以这回归预测每100k人口的暴力犯罪的方式解决这样的问题；除了预测，还提出了一个整数线性规划公式，用于优化各州的警察分布，这种分配方式考虑了国家的犯罪严重性，人口，密度和预算。

1.3 本文的研究内容

本文研究的出发点是：民警笔录案件时，需要填写许多信息，比如报案日期、案件名称、发案区划、办案单位、案件状态、案件类型、案件类别、简要案情等，其中简要案情是由民警听取报案人的描述，以文字的形式录入数据库的，包含的案情信息最多，也是最具破案价值和参考意义的。案件类别这一栏的信息常常填错，但案件类别本身是一个非常重要的信息，错误的案件类别不利于案件研判，也使案件统计不能得到正确的结果，如果案件分类准确还可用于同类案件集中分析，有助于获得正确的决策。另外简要案情里包含的信息要素有涉案时间、涉案金额、电话号码、QQ号码、银行账号、网站网址、人员结构、公交路线等，如果可以将这些信息准确提取出来，可以为串并案分析，发现重大犯罪行为和发现团伙作案提供线索。

本文研究的目标：通过调研公安领域分析犯罪行为的关键技术，选择合适的文本挖掘技术为协助警务人员研判、分析案件和给予串并案决策信息。

本文研究的关键技术：本文主要围绕两个主要案件分析技术，一是案件自动分类，且达到一定的分类精度；二是案件信息要素提取，将有助于破案的信息要素如涉案网址、手机号码、银行账号等信息抽取出来，其中还针对银行账号，提出了一种识别银行账号所属（作案人员/受害人员）的方法。

本文研究的创新点：

（1）将案件分类模型设计为层次分类结构，当对一个案件样本进行分类时，局部分类器自上而下对其分类。

（2）将规则分类器引入案件分类模型中，对于置信度低于最低限度的分类结果，拒绝该分类结果，并将案件交于规则分类器分类。

（3）利用正则表达式匹配的方法，提取案件描述文本中的信息要素，提取的要素有：作案时间、涉案金额、涉案网址、涉案银行账号、电话号码、QQ号码、交通工具和作案人员结构。

（4）提出根据案件文本的词性序列识别信息要素所属（属于作案人员或受案人员）的方法，以银行账号为例试验并取得结果。

1.4 本文的结构安排

第一章：绪论。本章叙述了本论文的研究背景，调研了国内外犯罪行为的研究现状，研究目的和意义，论文的组织结构安排等。

第二章：相关理论知识简述。简述了支持向量机模型，多类分类，层次结构分类和正则表达式的理论知识。

第三章：案件分类。本章叙述了基于支持向量机的层次分类模型的建模方法，和案件进行分类的流程；为了改善分类的精度，根据案件的实际情况提取一系列规则，设计出规则分类器和层次分类器相结合的方案，将置信度低的分类结果拒绝，交于规则分类器进行分类；并对实验进行了对比分析，

第四章：案件信息要素提取。本章阐述了应用正则表达式匹配的方法，提取案件描述文本中重要信息要素的方法，并根据提取的要素进行了串并案模拟；获取案件文本的词性序列，并根据词性序列来判断所获取的要素的属性（属于作案人员或受害人员），最后以银行账号为例并给出了实验分析结果。

第五章：结论和展望。

第二章 相关理论知识简述

2.1 支持向量机

支持向量机（Support Vector Machine, SVM）是一种有监督的机器学习技术，最早是由Vapnik(1979)在其科研工作过程中提出来并由Boser、Guyou和Vapnik在1992年进一步开发完善。自从SVM应用于模式分类，SVM凭借其出色的性能和优于其他分类器的表现逐渐风靡，至今仍受追捧。这一节用于回顾SVM的原理。

假设给定一个训练集，其中的每个样本都可以归到线性可分的两个类别中，SVM的目标就是找到最大化两个类之间的分离间距的最佳超平面。设定线性可分的训练集为，其中输入样本的输出类别为。分离两类的决策面超平面可以定义为:

(1)

其中是表示超平面的法线的可调整权重向量，表示输入的维度空间，表示偏差。

假设样本数据点满足如下的限定公式：

(2)

(3)

支持向量是与分离超平面最接近的点，并且位于由公式（2）和（3）等式情况定义的超平面上。所谓间距（margin)，即两个类别的支持向量之间的距离，可以被定义为两个超平面到原点的垂直距离之间的差：

margin (4)

显然，使间隔最大化等同于使加权向量的欧几里德范数最小化。受益于凸优化的优点，该问题可以被公式化如下：

(5)

为了处理不可分数据，有必要在公式（2）和公式（3）中引进松弛变量，。由于 与到间距边界的误分距离成正比，所以在目标函数中应包括正惩罚项，以防止产生过多的错误分类，然后目标函数现在可以定义为：

(6)

为了使用拉格朗日乘数解决问题，构建拉格朗日函数的原始形式。将原始拉格朗日变换为相对于拉格朗日乘数最大化的对应的双目标函数。

分类问题中输入和输出之间的函数可能不是线性的，输入空间被映射到高维特征空间以寻找与输出的线性关系，允许在高维特征空间中创建最佳超平面。应用内核技巧，非线性函数执行内积 ，消除需要明确地在高维空间中进行的计算。通常使用的非线性核，以及将在本研究中使用的非线性核是高斯径向基函数（GRBF）核：

(7)

其中代表函数的传播并需要被优化。

拉格朗日的双重形式最大化可以写为：

s.t. (8)

其中是拉格朗日乘子。

通过Karush-Kuhn-Tucker互补条件确定最佳偏差，允许使用决策面将测试点分类如下：

其中,是最优拉格朗日乘子。

2.2 多类SVM分类

最初SVM是被设计为二类分类器。虽然将多分类问题作为单一的“全局”优化问题的解决方法存在，但是在计算上比解决多个二分类问题昂贵得多。因此，许多将多分类问题分解为多个二分类问题并使用SVM作为分类器解决的技术被提了出来。下面介绍几个常用的技术：

（1）一对多策略

对于具有N个类别的分类问题（N>2)，需要构建N个二类SVM分类器，第i个SVM分类器的训练过程为：将标记类别为i类别的样本视作正面样本，而其余非i类的样本均视为负面样本，从而转化为二类分类问题训练分类器，如此获得N个二类SVM分类器。在分类预测阶段，当输入一个测试样本，这N个分类器均对其进行预测，最终输出分数最大的分类器取胜，将其类别赋予测试样本。此为一对多策略，该策略的缺点在于当训练样本集规模很大时它的训练复杂度很高，因为这N个分类器都需要利用全部样本进行训练。

（2）一对一策略

这个策略需要利用N个类别的所有两两组合，构建N(N-1)/2个二类分类器。每个分类器的是由某一类别作正面类别，另一类别作负面类别训练而成。在分类预测阶段，结合这些分类器进行投票，并采用Max Wins算法：选择得到票数最多的类别作为测试样本的类别。优点在于，每个一对一分类器的训练样本数量较少，仅仅是所有训练集中的两种类别的样本集合，较低的采样数量导致较小的非线性，导致训练的时间较短；该方法的缺点是每个测试样本必须被呈现给大量的分类器，N(N-1)/2个，这导致测试速度较慢，特别是当问题中的类别数量很大时。

2.3 层次分类

许多现实世界域需要系统自动将对象组织成已知的分类法。 例如，新闻网站或一般的新闻服务需要将最新文章分类为网站的部分和子部分。这类学习任务通常被称为层次分类，它与多类别学习有所不同：1）它的整个类别集合具有层次性结构，通常以树形结构定义描述；2）每个样本对象必须标记一组与层次结构一致的类：如果一个样本属于某个类别，那它必定属于这个类别的任何祖先。

层次分类算法的目标：学习一个能准确预测一组类别模型，需要注意这些子集通常具有多于一个的元素，并且它们被赋予子树结构，这些子树可能有多个分支，在这种情况下，我们说在标签中有多路径，子树可能不会在树叶上结束，也就是说它们包含部分路径。

大致来说，文献中可用的算法可以分为两组：采用局部观点的算法，以及从全局角度学习模型的算法。局部算法使用不同的方法学习层次结构的每个节点的模型，然后通过在自顶向下的过程中评估局部分类器来获得对象的层次分类，直到模型不能将该节点包括在其附加的类上。另一组全局角度算法认为：层次分类可以被视作一个整体任务而不是一系列局部任务。

由于用于层次分类的局部算法具有模块化的有点，它们可以在并行平台中直接实现，以获得非常快速的学习方法；而且它们很简单，可以用一组SVM构建，只需要做出一些适应性调整；此外，还可以使用可用于调整二分类SVM的已知技术来提高分类器的整体性能。

下面介绍一下采用局部观点的层次分类算法，这也是本文研究中应用的算法。

在层次分类中，首先给定一个类别集合，然后根据已知的分类法则构建一个具有*r*个节点的树 （每个节点代表一种类别）。实际中可以由许多树构成的森林 开始，但是必须为每棵树增加一个人工定义的根节点来连接整个类别的集合，因此下文只建立一颗树来表现类别的层次结构。训练过程先定义一个训练集，其中每一个样本是由输入空间下的向量和输出空间下的向量组成的条目描述，其中将每一个输出解释为一个类别全集的子集当且仅当样本属于类别的时候。假设类别全集遵从如下的结构：

(10)

其中代表节点（类别）的祖先集合（包括类别）。

学习这类任务的直接方法可能在于学习一系列模型，其中每一个模型对应树的一个节点（类别）。然后预测时，条目被分配给所有满足条件的类别。但是这个过程可能导致与不一致的预测结果，为了避免这种情况，必须遵从自上而下的预测过程，如此，条目能被分配给类别如果他先前已经被分配给了类别的父母类别，这样，样本不会仅仅被分配给唯一类别就停止，而会继续自动分配给子孙类别。

下面讨论将层次分类学习任务分解为多个局部二分类器。如果模型将从二类分类任务中进行学习，我们必须指定必须使用的训练示例集合。基本上，当尝试学习类别的模型时，训练集有三个不同的选项：

—— 考虑的所有条目，然后同多分类学习一样可选一对多（one-vs-rest）策略，选取正面样本集合*(POS)*和负面样本集合*(NEG)*为：

(11)

—— 学习区分属于类别或者的任何兄弟姐妹的那些例子，定义为：

(12)

—— 与上一条的选项相同的想法，但是将仅考虑那些根据自顶向下预测策略，被分配给的父节点的样本。基于假设所有样本在任何情况下都属于根类别。因此，在符号中：

（13）

显然，第三选择可能比其他两个理由更充分，因为它在学习和预测阶段期间遵循类似的过程。另一方面，考虑到其他选项可以并行计算所有必要的模型，它产生比其他选项更慢的学习阶段，而第三个选项必须等待父模型来调整数据集。注意，训练速度最快的选择是第二个，因为训练数据集较小并且允许最大程度的并行性。

2.4 正则表达式简述

正则表达式（regular expression）最早是由一位名叫Stephen Kleene的数学家在二十世纪五十年代提出来的，最初的用途是用于描述“正则集”，这是几位神经生理学家所研究的模式。之后的五十年里，正则表达式的用途范围慢慢扩展，在信息抽取、数据结构化等处理字符串的业务上起着十分重要的作用，直至今日，比如应用于HTML文档的特定信息提取（唐惠丽, 郑小妹. 正则表达式的研究及在Web中的应用[J]. 计算机技术与发展, 2013(2):82-84.），应用于网络安全问题（张树壮, 罗浩, 方滨兴. 面向网络安全的正则表达式匹配技术[J]. 软件学报, 2011, 22(8):1838-1854.），从中医文献中抽取特定信息（王志飞, 李晓君, 郭霞珍,等. 正则表达式在中医文献研究中的应用初探[J]. 中国中医药信息杂志, 2010, 17(3):98-99.）结合语义关系对中医疾病知识进行逻辑化抽取（朱玲, 朱彦, 杨峰. 基于中医疾病相关语义关系的正则表达式及知识抽取研究[J]. 世界科学技术-中医药现代化, 2016, 18(8):1241-1250.）。可见正则表达式是一项能够解决许多实际问题的计算机技术。

正则表达式本质上来说是符合某种规则的表达式，可以将其理解为一种对文字进行模糊匹配的语言。正则表达式用一些称为元字符的特殊符号来代表具有某类特征的字符组合并且指定匹配次数，包含元字符的文本不再是具体的文本内容，而是作为一种文本模式，可用于匹配与该模式相符合的所有子文本（文本内的字符串）（Goyvaerts, Levithan J, Steven. Regular expressions cookbook =[M]. 东南大学出版社, 2013.）。利用正则表达式可以快速的分析文本内容，并快速定位到符合模式的字符串。

将正则表达式引入程序语言中，可以实现以下功能：

（1）测试字符串的特定模式，验证所输入内容是否有效；

（2）使用正则表达式表示文本中的某些特定字符，然后将其进行替换、删除等操作；

（3）在文本中利用正则表达式搜索符合模式的字符串，即从文本中提取出目标字符串。

第三章 案件分类

本章介绍案件分类模块的基本流程以及剖析了案件分类中涉及的关键技术。主要包含对案件文本的特征分析、案件类别的结构分析、用于案件文本分词方法的调整与改进、案件的特征选取等一系列文本预处理技术；以及对传统的支持向量机分类进行结构方面的改进、引入规则分类器辅助支持向量机分类。

3.1 案件文本的特征分析

用于本文研究的数据样本来自于湖南省长沙市公安局的情报管理数据库，2013年8月至2015年10月之间的部分案件记录，本文主要研究的部分是由民警笔录的案情简要描述，其基本结构为：开头先记录报警时间和报警人的姓名、联系电话等，然后再记录有报案人员描述的受害经历、受害结果，或者是目击情况、投诉等，最后可能添加有办案民警的观点或案件的当前处理步骤。例如：“2015年7月20日15时至17时之间，张三与李四在某某咖啡西餐厅内被一大约30岁的女性，以借手机打电话为由，骗走四张银行卡内现金5100元，一台白色某某品牌某型号手机（1580\*\*\*），该手机于2015年6月以1000余元购买，现价值约1000元。共计损失约6100元。上述事实有张三、李四的陈诉、现场勘验材料等证据证实。”

经过对大量案情简要文本的内容分析，得出以下结论：

（1）案情简要文本的篇幅一般较短，字数在100至300之间，最常见的篇幅为150字左右；

（2）内容属描述性事件，偏口语化，多见短句，最常见的语句结构为主谓宾形式；

（3）包含多种信息要素，如电信诈骗案中长包含作案人员的银行卡账号信息，电话号码等，其他信息要素有：损失金额、人员描述、身份证号、公交线路、公交站点等。

3.2 文本分类的基本流程

本章主要是将文本分类技术应用于案件，首先简述一下通常的文本分类，其包括的过程有：数据清理、中文分词及预处理、特征选择、文本向量化、算法分类和性能评估。文本分类的基本流程图如图×所示：





图中，文本分类模型先由训练样本集进行训练从而获得分类所需的知识，模型构建之后就能将此模型用于对未知的文本进行分类。

3.3 案件文本特征表示

3.3.1中文分词

获取案件语料数据之后，首先需要做好案件文本处理工作，要将案件文本表示成计算机能够理解与处理的数字形式，之后才能进行进一步分析处理。因此文本表示是案件分类的第一任务。

对案件进行中文分词是将案件文本特征表示的前提。本文采用分词效果和运行性能都较为优秀的“Jieba”分词。“Jieba” 分词是一款目前应用广泛、口碑较好的分词工具，其基于如下算法：

* 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图；
* 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合；
* 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

“Jieba”可以允许用户自行添加自定义词典和停用词词典，这两个特性对于案件文本的分词，具有实用价值。

本文对“Jieba”分词工具进行了一下的改进：

1、案件文本中包含不少公安领域专业词汇和地区、道路等名称词汇，将这些特殊词汇添加至自定义词典，提高分词的准确率；

2、根据词频，将区分度差的高频词汇放入停用词典，有助于提升分类的效果。

3.3.2特征表示

这一步是真正将案件文本转换成计算机能够理解的表示形式的过程。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：将每一文档都映射为由一组规范化正交词条矢量张成的向量空间中的一个点。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*为特征向量词条；*Wi*为*Ti* 的权重）[8]。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。传统的特征项的权重计算方法有TF/IDF表示方法，词频表示法，布尔型表示方法等，本文采用TF/IDF权重计算方法。

3.3.2特征降维

经过以上两个步骤得到的文本特征向量的维度通常都很大，这种特性往往会降低分类的学习效率和性能质量，因此需要采取特征降维的手段降低向量的维度。

特征降维有两种方法。一类称为特征选择（Term Selection），另一类称为特征抽取（Term Extraction）。特征选择是选择部分最有区分类别能力的特征，文档频率(document frequency，简称DF)、卡方统计（chi-squared stats，简称chi2）、信息增益(informationgain，简称IG)和互信息(mutualin formation，简称MI)等特征选择方法在文本分类中广泛应用（徐燕, 李锦涛, 王斌,等. 基于区分类别能力的高性能特征选择方法[J]. 软件学报, 2008, 19(1):82-89.）。特征抽取是指将高维度的特征经过某个函数映射至低维度作为新的特征，常用的特征抽取算法有主成分分析（Principal Component Analysis，简称PCA）。降维后生成的特征集合的每一个元素具有更强的代表性，维数的减少意味着耗费更少的计算资源[9]。

然而不管DF、chi2等特征选择方法还是PCA等特征抽取方法，都需要指定目标特征个数n，而不同n的取值，将会对分类效果产生很大的影响，如果选择n值不当反而会降低分类效果；如果n设为定值，将局限分类器的泛化能力。因此，本文并不考虑大幅度缩减特征维度，而是以如下的特定方法进行适当的特征降维：

1. 根据词频建立停用词集合，分词过程中筛去停用词；
2. 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；
3. 在分词过程中，对分词的结果进行词性标注，筛去词性为人名的词。

3.4 基于SVM的文本分类

上述过程完成之后，就可以根据案件样本和类别标签构建分类模型，然后对测试案件样本进行自动分类，即指训练好的分类器对给定的未知类别案件文本，自动将其归为某一类别。常用的文本分类算法有k-近邻方法、朴素贝叶斯算法、支持向量机方法、决策树方法、神经网络等。（扩充？）本文选用的支持向量机方法进行分类，其主要思想是：假设给定一个训练集，其中的每个样本都可以归到线性可分的两个类别中，SVM的目标就是找到最大化两个类之间的分离间距的最佳超平面，使得正负模式之间的间距最大。

3.4.1 案件类别的特点

为获得较好的分类效果，对组成训练样本集合的案件类别集合进行了分析，其特点有：

（1）各类别的样本数目比例不均衡；

（2）存在类别层次结构。

实际中，每天被录入的案件种类繁多，常见的有电信诈骗、入室盗窃等，比较少见的有吸毒赌博、集资诈骗等。因此，对于总体案件文本类别数量不均衡的特征，本文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n\_samples ，案件类别共有n\_classes类，属于类别y的案件出现次数为count(y)，类别y的权重值计算公式为：

对于案件类别存在的层次结构，需构造层次分类结构对分类模型进行优化，具体在下一节分析。

3.4.2 层次分类结构

通过分析案件文本、案件类别和省厅案件细分化文件，发现案件的类别与类别之前存在层次关系，如下给出具体情况说明：

（1）如某一条关于在道路上实施抢劫财物的案件，不同的警务人员给定的类别就可能不同，其可被分为“抢劫”或“拦路抢劫”，实际上，拦路抢劫属于抢劫的一个类别，换而言之，“抢劫”是上级类别，“拦路抢劫”属于“抢劫”的下级类别。

（2）一种类别存在于某几种类别的相近程度都较高，而与其他的类别相近程度都较低的情况，如“入室盗窃”、“办公室盗窃”、“门面盗窃”、“网吧盗窃”等案件类别相互近似程度较高，可形成一个集体，而“入室盗窃”、“办公室盗窃”与“信用卡诈骗”之间的近似程度比较低，不适合构成同一个集体。

鉴于这两种情况，本文合理地将分类器设计为具有层次结构SVM分类器（HierSVM），自顶向下，逐层分类。采用双层分类结构（类似于两层树结构），训练时，先将所有样本以上级类别作为标签训练上级分类器（根分类器），然后根据各个上级的案件样本子集，以下级类别作为标签训练出每个下级分类器（子分类器）。当给定一个案件样本，首先经过上级分类器获得上级类别，再经下级分类器获得下级类别（最终类别），如图×。



基于层次结构的支持向量机分类器，对于样本类别之间具有层次关系的样本集，可以提高分类准确率。比如在案件样本中，“车扒”和“扒窃”均为样本类别名称，实际上“车扒”是属于“扒窃”的一种类型（在公交车上作案），如果将“车扒”和“扒窃”按同级类别进行分类，“车扒”类的分类准确率仅为0.855；而将“扒窃”作为上级类别，“车扒”作为“扒窃”的下一级类别，“车扒”类的分类准确率为0.975。

在上级分类层中，案件类别差异较大，而类别数目较少，使得特征降维后的特征向量的类别区分能力强，从原理上说，支持向量机的判决函数只与支持向量有关，如果支持向量差异明显，分类间隔的宽度就比较大，从而分类准确率提高。在上级分类的准确率得到充分保障基础上，从下层各个下级类别的案件样本子集中抽取出区分能力强的特征词，构成新的特征向量。各个下级类别的分类器都有属于自己的特征向量，能减少不相关类别的信息干扰，这有利于寻找到最优分类决策面。

通过训练集生成各个支持向量机分类器模型后，便可以进行案件分类，分类工作的基本步骤如下：

1. 利用“jieba”分词对案件进行分词，剔除词性为人名的词，然后经过同义词替换。
2. 假设上级分类器的特征集为，采用TF/IDF权重计算方法，将该案件向量化，得到它的特征向量，然后利用上级分类器预测其上级类别，假设得到类别*L1*。
3. 根据上级类别*L1*，调用对应*L1*类别的下级分类器，假设该下级分类器的特征集为，采用TF/IDF权重计算方法向量化该案件得到新的特征向量，最后得到由该下级分类器预测出的最终案件类别。

3.5 规则分类与HierSVM分类结合

SVM基于结构风险最小化等统计学习原理，将测试样本集合压缩为支持向量集合，学习得到分类器的决策函数（decision function）。在分类过程中，分类器的置信度是一个值得重视的参量，多采用以决策函数值作为衡量置信度大小的标准[13]，决策函数值越大说明分类结果越可靠。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。本文经过实验和分析研究，在预测时，记录了每条测试样本的每个叶子SVM分类器的决策函数值，对符合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. 仅一个分类器的决策函数值为正数，但其值很小，小于1；
  3. 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%；
  4. 出现三个及以上的分类器的决策函数值为正数。

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。在所采用的确定性推理中，规则库中的所有规则、事实数据库中的所有事实和推导出来的结论都是正确的，它们要么成立要么不成立[14]。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。例如，规则库中有这样一条规则：IF（被盗EXIST AND 卧室EXIST）AND NOT（酒店EXIST OR 旅馆EXIST）THEN（类别名称=入室盗窃，上级类别=盗窃），此规则翻译成书面语即为：如果该条案件文本中包含有“被盗”和“卧室”，并且不包含“酒店”或者“旅馆”，那么这条案件被分为“入室盗窃”类别，上级类别为“盗窃”。

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5所示：当出现遍历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，而且由于规则库中规则数量较多，对每一条待分类案件需要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，所以分类速度较慢，单条案件分类速度远慢于支持向量机分类的速度。因而本文采用以支持向量机分类为主，规则分类为辅的方法对案件进行分类，同时保证了分类速度和分类准确率。本案件分类模块可以随时导入、添加训练样本，获得新的分类器模型，使系统满足随时改善（提高分类准确率）的要求，也可以识别新类别的案件。

3.6 实验与分析

本文研究对象为盗窃、诈骗等涉及财产的财产案件，样本集中多为财产案件，如“入室盗窃”、“信用卡盗窃”等，少量为非财产案件，如“纠纷”、“举报”等。样本是由长沙市公安局提供的，随机抽选的2013年1月至2015年5月之间21025条案件记录，总共涉及33个案件类别。下文的实验均采用十折交叉验证法，将10次的结果取平均值作为客观结果。

实验以准确率作为评价指标，定义公式如下：

式中：*tp*是被正确地划分为正例的个数，*fp*是被错误地划分为正例的个数。

实验1：验证本章提出的对分词工具的改进和特征降维的效果。

|  |  |  |
| --- | --- | --- |
| 序号 | 实验对象 | 十折平均准确率 |
| 1 | 原始SVM分类 | 0.783169850 |
| 2 | 改进分词后的SVM分类 | 0.821373257 |
| 3 | 同义词替换后的SVM分类 | 0.800206504 |
| 4 | 删去人名后的SVM分类 | 0.794011357 |
| 5 | 改进分词+同义词替换+删去人名后的SVM分类 | 0.838409912 |

从表3.\*的实验结果可以看出，将分词工具进行专业词汇的补充和增加停用词词典后，对SVM案件分类的分类准确率的提高贡献了3.82百分点，用同义词替换进行降维的方法将分类准确率提高了1.7百分点，单独使用删去人名的的效果略差，仅将分类准确率提高了1.1百分点，将三项改进结合后的SVM分类准确率提高了5.54百分点，说明本文对分词工具的改进和特征降维的方法是有效的。

实验2：验证自动调整类别权重平衡模式的效果

本实验在上一个实验的基础上，对经分词改善和特征降维的SVM分类器引入自动调整类别权重平衡模式，在同一个样本集中进行实验。

|  |  |  |
| --- | --- | --- |
| 序号 | 实验对象 | 十折平均准确率 |
| 1 | 参照SVM分类 | 0.838409912 |
| 2 | 自动调整类别权重平衡模式的SVM分类 | 0.860512479 |

从表中可以看出，引入自动调整类别权重平衡模式的分类效果有所提升，准确率提高了约2.2百分点。

实验3：根据所提出的Rule+HierSVM算法，构建了具有层次结构的SVM分类器，并对不满足置信度标准的分类结果予以拒识，并由规则分类器给出案件类别。下面对普通SVM分类（SVM），层次SVM分类(HierSVM)、规则与层次SVM结合的分类(Rule+HierSVM)的分类效果进行了比较，结果如下：

|  |  |  |
| --- | --- | --- |
| 序号 | 实验对象 | 十折平均准确率 |
| 1 | 普通SVM分类 | 0.838409912 |
| 2 | 层次SVM分类(HierSVM) | 0.929272070 |
| 3 | 规则与层次SVM结合的分类(Rule+HierSVM) | 0.957666494 |

上表数据是对训练集进行交叉验证实验所得到的结果。其中SVM分类器均采用了以TF/IDF方法计算词条权重，过滤了停用词、人名和同义词替换，核函数为线性核函数，对错误样本的惩罚因子设为1。

普通SVM分类器和层次SVM分类器的不同之处为层次SVM分类器具有层次关系结构。从图表中数据可以看到，层次SVM分类器相对于普通SVM分类器，分类准确率提高了约9.1个百分点。本实验数据的层次结构比较简单，理论上来说，具有更复杂层次的实验样本数据更能凸显层次分类器的分类效果。而规则与层次SVM结合的分类又比单纯的层次SVM分类高出约2.84个百分点，说明对于数量较少的非财产案件采用规则匹配分类更为合理。总体而言，在这些分类器中，规则与层次SVM结合的分类效果达到最优。

案件分类可应用于服务治安态势分析，预测各类案件的发生趋势，辅助警务人员决策，还能支持原有公安系统更好地打击犯罪、提高响应能力和作战能力。

第四章 案件信息要素提取

案件要素抽取是提取案件中有用的信息点，或者说有利于破案的线索。根据公安方面提出的要求，需要提取目标信息点为作案时间、涉案金额、身份证号、涉案网址、涉案银行账号、电话号码、QQ号码、交通工具、作案人员结构、公交线路、公交站点。作案时间可用于统计某个时间段内的总体或某类犯罪案件发生频率，涉案金额主要应用于统计犯罪案件造成的经济损失，其他的信息点均为重要的破案线索，利用线索的相关性，可对案件做进一步处理，如串案、并案。其中，提取的作案时间如“2016年1月1日”，“2016年10月”，格式化为 “2016/01/01”、“2016/10”；涉案金额描述形式多样，如“被偷现金500元，苹果手机一台，现价值约为2000元，共计损失2500元”，所提取涉案金额为2500，为了保证后期统计的正确性，所以将分散的涉案金额进行合计或只提取合计损失金额，最终只保留一个总金额；提取的交通工具信息点的描述形如“一辆银白色轿车”、“无牌黑色面包车”；作案人员的描述形如“一个中年男子”、“两名男子和一名女子”等，提取并修整为“□男□女”格式，如“2男1女”。

这部分应用正则表达式匹配方法。正则表达式在文本匹配和文本抽取方面有着强大的功能，在实际工作中如匹配用户邮箱，手机号码，抽取网页内容和网络安全检测等领域有着广泛的应用[6]。对于案件文本，虽然是以半结构化文本的形式记录在库的，但其内容本身结构并不复杂，比较单一，符合时间、地点、人物、事件此类简单的语法结构。难点在于必须考虑正则表达式的容错性，因此，在充分研究分析案件文本数据之后，详尽罗列出各项信息抽取的规则，并测试修改，最终整合成正则表达式。将正则表达式匹配案件文本中的语句，抽取相应的信息。示例如下表所示：

2.2 案件要素抽取的主要应用

1. 简化案件信息录入。自动填充相关信息项，使警务人员免于重复冗余的信息录入；

2. 支持多维度分析统计案件。在作案时间、涉案金额、身份证号、银行账号、电话号码、QQ号码、涉案网址、人员结构、交通工具、公交线路、公交站点这11个维度下对案件进行综合统计分析，有助于案件材料的归纳整理工作；

3. 自动串并案件。在原导入的案件数据的基础上，返回“串并编号”和“串并要素”。则具有相同“串并编号”的案件即为系统自动串并的案件，“串并要素”即为同组串并案件具有的相同要素信息（涉案网址、涉案银行账号、电话号码、QQ号码）。警务人员通过“串并编号”的最大值即可掌握串并案件的总组数，快速了解正在活动的犯罪团伙数量；同一“串并编号”的数量反映出每组案件的规模，从而能够科学安排警力优先针对社会影响较大的多发性案件开展侦查；通过相同的“串并编号”可将相关案件线索串联起来，更有效的侦破打击违法犯罪。所抽取出的案件要素为串并案件提供依据，串并案的意义不仅在于为侦查活动提供情报信息，还在于能够增加案件信息（特别是近年来流行的电信诈骗、微信诈骗等），促使公安机关受理案件，及时发现此类案件新的诈骗特点，是侦查人员全面了解案情，及时调整主动侦查的方向[7]。

本节对上述系统各模块的功能进行了测试和评估。本系统平台为Windows8.1 64 bit操作系统，实验程序所用的编写语言为Python2.7.9，数据库为MS Excel2013，程序界面简洁友好，简单易用。

4.1 信息抽取评估与应用

本部分实验以2015年10月01日至2015年10月05日的8031条案件样本作为实验数据，对信息抽取功能进行了测试。

实验结果要素*i*被正确抽取的比例*Pi*为评价指标：

······2

式中：*rp*为要素*i*被正确提取的案件个数，*wp*为要素*i*被错误提取的案件个数，*up*为要素*i*存在但未被提取的案件个数。

信息抽取效果如图6所示。

信息抽取效果评估：

1. 金额提取基本准确，但有些案件的金额涉及案件实际发生金额和非实际发生金额，实际发生金额如“被盗一部手机，现价值2000元”，非实际发生金额如“对方打来电话说自己中了五千元大奖”，其中的“五千元”也会被提取出来，这种情况下提取的是非有效金额。

2. 身份证号码、电话号码、手机号码、QQ号码提取基本全面、完整，但

如上文提到暂时还不能实现区分号码所属为受害者还是嫌疑人，后续研究将借助语义分析的方法改进[15]。

3. 案件中出现的时间基本都可以提取到，目前不能区分将报案时间和案件发生时间，需要进一步的改进。

第五章 结论与展望