湖南大学

硕士学位论文

数据挖掘技术在公安犯罪行为分析中的应用研究

姓名：刘华

申请学位级别：硕士

专业：计算机技术

指导教师：陈湘涛;谭皓

20080615

摘 要

数据挖掘技术在公安工作中的研究与应用尚处于起步阶段，许多公安业务信息系统还停留在初级处理水平，缺乏综合性的开发应用，智能化的分析研判，科学性的决策预警；缺乏对数据由微观到宏观的加工能力，由宏观数据到微观数据的问题发现手段。如何利用数据挖掘技术挖掘和提取潜藏在大量业务数据中具备关联性的规律趋势，提高公安执法效率与快速反应能力、及时预防和打击犯罪行为，为警务决策提供支持服务，是本文研究的重点。

本文主要工作如下：

针对传统Apriori算法在公安犯罪行为分析中忽略了不同项间不同重要性的问题，提出了改进的加权关联规则挖掘算法WARMA。改进算法设计了加权关联规则模型，通过提出的k-支持期望概念来排除不可能成为加权频繁k-项集的候选集子集，解决了加权关联规则挖掘中加权频繁项集的子集可以不是加权频繁项集的问题。实验及分析表明：算法WARMA有着较少的候选集和较短的执行时间，能更有效的发现重大犯罪行为。

针对传统Apriori算法在公安犯罪行为分析中无法及时敏锐的发现某些新型犯罪行为的问题，提出了基于FUP的改进关联规则增量式更新算法SFUP，SFUP算法通过用敏感参数衡量对新项目的重视情况，改进了一般增量式更新算法产生频繁项集的过程。实验及分析表明：改进算法在越大数据量下优势越明显，能更有效的发现新型犯罪行为，在挖掘过程中显示了良好的空间和时间性能。

针对传统 ID3 算法在公安犯罪行为分析中存在的问题，提出了一种基于先验参数的改进算法 BID3，BID3 算法通过增加先验参数 B，加强了对重要属性的标注，使得决策树减少了对取值较多的属性的依赖性，尽可能减少大数据掩盖小数据的现象发生。实验及分析表明：在处理越大规模数据集的决策树构造过程中，BID3 算法在效率和性能上比 ID3 算法有更大的优越性。

在结合公安犯罪行为分析实际的基础上，进行了基于决策树算法的犯罪行为分析原型系统设计与实验，提出了功能需求与系统流程图，介绍了原型系统模块构成和实现。实验及分析表明：运用数据挖掘技术对公安信息数据库中的海量数据进行挖掘处理，发现趋势规律，从而快速准确的辅助警务决策，在公安工作中具有重要的现实意义。

关键词：数据挖掘；算法；权值参数；敏感参数；先验参数

Abstract

Data mining technologies have been more in-depth researched in many fields,but just have elementary been researched and applied to the public security work,many public security information systems stills stay at a low level, lack the synthetic exploiture and application,lack the intelligence analysis estimation,lack the scientific decision prevision and alarm, lack the the ability of process from microcosmic to macroscopical in data,the means of discover a problem from microcosmic to macroscopical in data.This thesis focses on how to apply data mining technologies to mining and extract the associable rule and trend in the a great deal of data information,in order to enhance the executive efficiency and quickly respond ability,prevention and fight crime in time,and provide the support service for the decision-making in the police affair.

The thesis primary work as follows:

The thesis contrapose the conventional apriori algorithm to exist the problem in public security crime analyze: ignore the problem of different item and different importance, put forward the reformative weighted association rules algorithm WARMA. the reformative algorithm design the weighted association rules model, Expeled to become impossible the weighted frequent k-item of the candidate subset through using k-support expectation, resolve the problem that subset of the weighted frequent item can not be the weighted frequent. experiment and analyse indicate: algorithm WARMA has the less candidate subset and the shorter runtime,can more effectively discover important crime.

The thesis contrapose the conventional apriori algorithm to exist the problem in public security crime analyze:can not duly observantly discover the new crime, put forward the reformative association rules incremental updating algorithm SFUP base on the algorithm FUP. algorithm SFUP scale the recognition circs for new item through using the the sensitivity parameter, improve the ecumenical incremental updating algorithm come into being frequent item process. experiment and analyse indicate: the reformative algorithm more big data quantity more advantage, can effectively discover new crime,and showed good space and time capability in the process of mining, consumedly heighten the all mining efficiency.

The thesis contrapose the conventional ID3 algorithm to exist the problem in public security crime analyze: the one is the according to the minimum principle of the entropy value , the attribute to be listed by ID3 algorithm to should be judged the first , but not so important in realistic; Two is when the majority attribute data

quantity are bigger, the separately attribute data quantity are lesser, easily appear the big data to flood the fraction, thus lose some correspond the important judgment. put forward the improved BID3 algorithm base on the transcendental parameter, the BID3 algorithm improve the attribute choice standard of the conventional ID3 algorithm, intensify the important attribute attention,through the increased transcendental parameter B,so as to the decision tree reduce to the dependence which takes the value more attribute, reduce to appear the phenomenon of a big data to cover up fraction possibly. experiment and analyse indicate: in the decision tree structure process of the more large-scale data gather, BID3 algorithm more advantage compare ID3 algorithm on the efficiency and the function.

The thesis combine the actua public security crime analyze, expansion the design and experiment of criminal analyze prototype system base on the decision tree, put forward the function requirement and system flow chart, introduce the prototype system module structure and realization. experiment and analyse indicate:it is important realism meaning in the public security work that mining the amount of sea data in the public security information database,using data mining technology, discover regulation trend, assist the decision-making in the police affair fleetly true.

Key words: data mining; Algorithms; weight parameter; sensitivity parameter; transcendental parameter

插图索引

附表索引

湖 南 大 学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：刘平　　日期：2008年 9月 4日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南大学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，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密□，在_____年解密后适用本授权书。

2、不保密☑。

（请在以上相应方框内打“√”）

作者签名：刘平　　日期：2008年 9月 4日

导师签名：陈湘清　　日期：2008年 9月 4日

第1章 绪 论

1.1 研究背景

随着信息技术的高速发展，数据库应用的规模、范围和深度不断扩大，特别是数据仓库以及Web等新型数据源的日益普及，数据量呈指数上升，人们面临着快速扩张的数据海洋。但是，由于数据库技术作为一种基本的信息存储和管理方式，仍然以联机事务处理(OLTP)为核心应用[1]，缺少对决策、分析、预测等高级功能的支持机制，因此导致了“数据爆炸但知识贫乏”。

面对这一挑战，数据挖掘(Data Mining)技术应运而生，并显示出强大的生命力。数据挖掘使数据处理技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势。通过数据挖掘，有价值的知识、规则、或更高层次的信息就能从数据库的相关数据集合中抽取出来，这样一来，就把人们对数据的应用，从低层次的查询操作，提高到为各级经营决策者提供决策支持。

经过十几年的研究和实践，数据挖掘已经融合了数据库技术、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算、数据可视化等最新技术的研究成果，形成独具特色的研究分支[2]。像其它新技术的发展历程一样，数据挖掘也必须经过概念提出、概念接受、广泛研究和探索、逐步应用和大量应用等阶段。从目前的现状看，大部分学者认为数据挖掘的研究仍然处于广泛研究和探索阶段。一方面，数据挖掘的概念已经被广泛接受。在理论上，一批具有挑战性和前瞻性的问题被提出，吸引越来越多的研究者。数据挖掘的概念从20世纪80年代被提出后，其经济价值已经显现出来，而且被众多商业厂家所推崇，形成初步的市场。另一方面，数据挖掘系统研制仍有许多问题需要研究和探索。目前，数据挖掘的研究正向着更加深入的方向发展，数据挖掘被称为未来信息处理的骨干技术之一。

当前，随着我国的改革开放和现代化建设的不断深入，社会的政治、经济、科技得到迅猛发展。与此同时，违法犯罪的动态化、组织化、职业化和高智能化、高科技化趋向越来越明显，带有时代特征的新的犯罪形式和犯罪手段(如网络犯罪、洗钱等)不断出现，杀人、绑架、投毒等大案要案的发案率不断提高，现代违法犯罪行为进入了一个高发期和提速期，违法犯罪的深刻变化对公安警务改革、执法效率、犯罪控制、防范战略和侦查谋略等提出了严峻挑战。

公安信息化建设实现了打击违法犯罪工作和现代科学技术的最佳结合点，其

核心在于利用先进的科学技术对信息资源进行系统有效的管理，实现公安信息的“大集中、大整合、高共享”。公安信息化建设是加快警务决策、提高执法效率、更有效的打击与制止犯罪等的重要手段之一。加快公安信息化建设，将信息资源转化为破案力，是实现打击违法犯罪工作跨越式发展的重要途径，更是现实斗争的需要。

随着公安信息化建设的进一步深入，公安工作中的许多业务信息系统逐步从微机个人数据库移植到大型数据库，但对信息的处理还基本停留在增、删、改、查询、统计等传统功能上，缺乏智能化的分析功能。可以说其事务性功能已经基本完善，但系统在信息分析、研判等方面的功能不完善，无法充分满足业务监控预警、决策分析的需要，缺乏对数据由微观到宏观的加工能力，缺乏由宏观数据到微观数据的问题发现手段，利用数据挖掘技术挖掘和提取潜藏在大量业务数据后面具备关联性的规律趋势等基本是空白，在公安的违法犯罪分析研究领域，更是鲜见有效应用。

1.2 国内外研究现状

近年来，国内外各研究机构纷纷展开了对数据挖掘技术的研究和探索工作，极大的推动了数据挖掘技术的迅猛发展。

1.2.1 国外研究现状

从目前研究的现状来看，关于知识发现和数据挖掘的研究，国外起步早，从数据库中发现知识是上世纪八十年代末开始的。KDD一词是在1989年8月于美国底特律市召开的第一届KDD国际学术会议上正式形成的。刚开始每两年召开一次国际KDD学术会议，1993年后每一年召开一次国际KDD学术会议。1995年在加拿大召开了第一届知识发现和数据挖掘国际学术会议。由于把数据库中的“数据”形象地比喻成矿床，“数据挖掘”一词很快流传开来。到目前为止，由美国人工智能协会主办的KDD国际研讨会已召开了十多次，规模由原来的专题讨论会发展到国际学术大会，人数由二三十人到超过千人，论文收录数量也迅速增加，研究重点也从发现方法逐渐转向系统应用直到转向大规模综合系统的开发，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。其他内容的专题会议也把数据挖掘和知识发现列为议题之一，成为当前计算机科学界的一大热点。

世界上研究数据挖掘的组织、机构或大学很多[3]。比较著名的如卡内基梅隆大学（机器制造DM、多媒体数据库DM、互联网DM三个研究中心）、斯坦福大学、麻省理工学院。著名研究机构如：ACM(ACM Special Interest Group on Knowledge Discovery in Data and Data Mining)、KDNet(the European Knowledge Discovery Network of Excellence)。

数据挖掘算法在实际数据挖掘系统中得到了很好的应用。美国斯坦福大学智能数据库系统实验室开发出了大量的商用数据挖掘系统，如DBMiner挖掘系统，该系统包含了许多先进的挖掘算法，并有很多特点：用户无需具有高级的统计知识和培训即可使用该软件；挖掘的知识类型多种多样：从关联规则、序列模式（Sequence Pattern）到发现驱动（Discovery-Driven）的分类等；因采用了许多先进的研究成果，该产品的速度是其同类竞争者的 20 倍；此外该系统可以在多种平台上运行，并与许多主流数据库系统（如SQL-Server，Oracle等）结合紧密；同时还引入了在线分析挖掘技术，使得系统更能发挥数据仓库的分析优势。

IBM的Almaden实验室所进行的Quest项目同样也是数据挖掘研究领域中的佼佼者，该项研究包含了对关联规则、序列模式、分类及时间序列聚类（Time Series Clustering）的研究，其代表性的产品有：DB2 Intelligent或Miner for Data，该产品是在IBM DB2平台下的系统，当然也有Windows NT下的类似产品。

此外，美国宾西法尼亚大学的数据挖掘研究小组也在这些方面取得了显著成果，其主要研究包括：利用注释和文本对数以百万计的文章进行聚类和分析；从多家医院的病人数据库中发现可以提高医疗质量和降低医疗费用的模式；在构建一个模型中选择合适的变量；基于DNA序列预测基因模式等等。

目前世界上比较知名的数据库公司，如Oracle，Sybase等都已在不同程度上将数据挖掘的有关技术结合到其对应的数据库产品中，使得大型数据库的功能向智能化的方向迈进了重要的一步。

总之，美国、欧洲及日本等发达国家在这方面的研究投入巨大，主要的研究成果也集中于这些国家的研究机构和大学。

1.2.2 国内研究现状

与国外相比，国内对数据挖掘的研究稍晚，没有形成整体力量，直到1993年国家自然科学基金才首次支持该领域的研究项目，到上世纪的90年代中后期，初步形成了知识发现和数据挖掘的基本框架。自90年代中期一批研究成果（学术论文）逐渐发表在《计算机学报》、《计算机研究与发展》、《软件学报》、《人工智能与模式识别》等刊物上，研究重点也正在从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。但是基本上还是以学术研究为主，实际应用上处于起步阶段。

所涉及的研究领域一般集中于学习算法的研究、数据挖掘的实际应用以及数据挖掘理论方面的研究。国内从事数据挖掘研究的人员主要在大学，也有部分在研究所或公司。例如：北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究；北京大学也在开展对数据立方体代数的研究；华中科技大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展

了对关联规则挖掘算法的优化和改造；南京大学、四川大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及 Web数据挖掘。

目前国内进行的大多数研究项目是由政府资助进行的，如国家自然科学基金、863计划等。具体的研究项目有中科院计算机研究所的智能信息处理重点实验室研制开发的多策略数据挖掘平台MSMiner系统，此系统集成了关联规则挖掘算法；复旦大学研制开发的ARMiner系统，该系统采用的关联规则挖掘算法是基于Apriori的改进算法。虽然已经取得了相当的成功，但目前在处理极大规模的数据时，如何提高算法效率，如何提供一种与用户交互的方法以及如何将用户的领域知识结合在其中等都是尚待解决的问题。

1.3 研究目的与意义

面对世界多极化、经济全球化、社会信息化给维护国家安全和社会稳定的公安工作带来的巨大挑战，新世纪新阶段公安机关要坚持信息主导警务原则，全面实施科技强警战略，积极推动信息技术在公安工作中的广泛运用，大力提高公安工作的科技含量，切实加快公安工作的信息化步伐，有力的带动和促进公安工作现代化、正规化建设，有效的提升公安整体战斗力。

本文正是基于这一现状，希望通过应用数据挖掘技术在公安犯罪行为分析方面进行有益的探索，充分有效的对公安信息化系统中的海量数据进行挖掘，找出隐藏在大量信息数据背后有价值的规律和结论，把信息资源优势转化为现实破案能力，满足公安行业特殊性要求，以提高公安执法效率与快速反应能力，及时的预防和打击违法犯罪行为，并为警务决策提供支持服务，真正做到“打击犯罪、预防为主”，对公安工作起到积极的推动作用，充分体现科技强警的重要现实意义，这是当前公安信息化建设的进一步发展方向，也是“科技强警”战略的具体体现。

1.4 研究内容

(1)提出了Apriori算法在公安违法犯罪行为中的应用和存在的两个问题，即不同项间的重要性问题和对新项目的敏感性问题，并针对问题提出了加权关联规则挖掘算法和基于敏感参数的增量式更新改进算法。

(2)实验分析了Apriori算法存在的不同项间的重要性问题，提出了基于权值参数的加权关联规则模型和k-支持期望概念，进行了算法描述，通过实验分析和性能测试，显示在相同加权支持度阈值条件下，加权关联规则算法产生的频繁项集和执行时间均小于Apriori算法，证实加权关联规则算法能更有效的发现重大犯罪行为。

(3)实验分析了Apriori算法在公安工作中无法及时敏锐的发现某些新型犯罪

行为，讨论了一般增量式更新算法的基本思想和具体执行过程，针对一般增量式更新算法发现不了新颖的、潜在有用的模式的不足，提出了用敏感参数衡量关联规则挖掘算法对新项目的重视情况，然后从敏感性和时间效率出发对增量式更新算法进行了改进，并通过实验对改进算法和原算法做了分析和比较，最后对改进算法的性能做了分析。其优点是能较好地发现新增数据中的新模式，在挖掘过程中显示了良好的空间和时间性能，并具有较高的敏感性。

(4)指出了传统ID3算法在公安犯罪分析中存在的问题：一是按照使熵值最小的原则,被ID3算法列为应该首先判断的属性,在现实情况中却并不那么重要；二是当大多数属性数据量较大，个别属性数据量较小时，容易出现大数据淹盖小数据的现象，从而失去一些相应重要的判断。

(5)实验分析了ID3算法存在的问题，提出了基于先验参数的BID3算法,进行了实验分析和性能测试，从生成决策树的形态和生成决策树的时间对ID3算法和BID3算法构造的决策树进行了对比分析，并通过大量数据测试，证明了在处理越大规模数据集的决策树构造过程中，BID3算法在效率和性能上比ID3算法有更大的优越性，验证了BID3算法在公安实际工作中更具有较强的现实意义，同时进行了基于决策树算法的犯罪行为分析实验系统的设计，提出了功能需求与系统流程图，介绍了系统模块构成、系统演示界面。

1.5 组织结构

全文分为四章，主要内容如下：

第 1 章概述了数据挖掘技术的产生背景、国内外研究现状、论文的研究背景、研究目的及意义、论文的组织结构等。

第 2 章概述了数据挖掘技术、公安信息化和数据挖掘在行业管理中的应用。

第 3 章实验分析了 Apriori 算法在公安违法犯罪行为中存在的两个问题，提出了基于权值参数的加权关联规则挖掘算法和基于敏感参数的增量式更新改进算法，并进行了算法描述、实验分析、性能测试。

第 4 章实验分析了 ID3 算法在公安犯罪行为分析中存在的问题，提出来基于先验参数的 BID3 算法，进行了实验分析和性能测试，同时进行了基于决策树算法的犯罪行为分析原型系统的设计与实验。

第 2 章 数据挖掘与公安信息化

2.1 数据挖掘

2.1.1 数据挖掘定义

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

数据挖掘，也称数据库中知识发现(Knowledge Discovery in Database)，是一种用于从大型数据库或数据仓库中提取隐藏的预测性信息的新技术。它能识别新颖有效的知识，开采出潜在效用的模式，找出最有价值的信息，指导商业行为或辅助科学研究。数据挖掘的基本过程和主要步骤如图 2.1 所示：

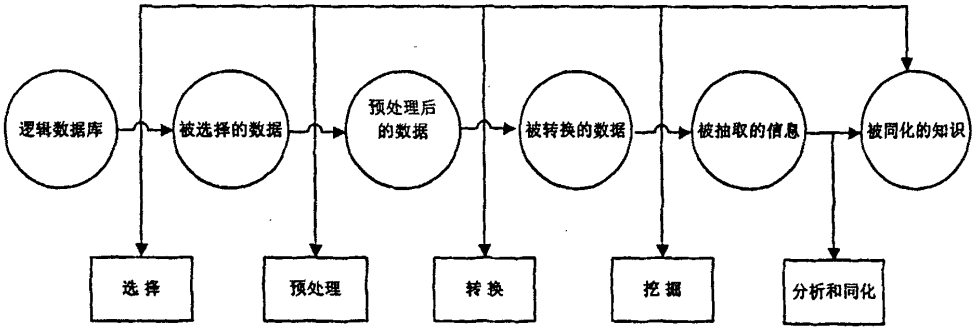

图 2.1 数据挖掘的基本过程和主要步骤

1. 数据选择

全面而丰富的数据是数据挖掘的前提，给需要解决的问题一个明确的定义，认清数据挖掘的目的并选择合适的数据是非常关键的。所选择的数据可以来自现有的数据库系统或者数据仓库。

2. 数据预处理

由于一些不确定因素导致采集到的数据可能存在瑕疵或者不一致性，甚至出现部分数据缺失的情况，因此对数据的整理是必须的。通过数据整理，可以提高研究数据的质量，为下一步数据挖掘的顺利开展做好了准备。

3. 数据的转换

将数据转换成一个分析模型，这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

4. 数据挖掘

利用各种数据挖掘方法对所得到的经过转换的数据进行挖掘，在这个过程中可以对所选择的算法进行改进和完善，所有模式推导与知识的获取工作均由挖掘

算法来实现。

5. 对数据分析和同化

在对挖掘结果进行解释和评估的基础上，需要对结果进行细致而深入的分析，并将分析得到的知识集成到业务信息系统的组织结构中去，这是一个同化的过程，决策者可以依据这些知识来进行科学决策。

2.1.2 数据挖掘任务

数据挖掘的任务一般可以分为两类：描述和预测。描述性挖掘刻划数据库中数据的一般特性，预测性挖掘在当前数据上进行推断，以进行预测。数据挖掘可以发现的知识类型如下[4]：

1.概念/类描述：特征化和区分

概念或类的描述称为概念/类描述，它可以通过数据特征化和数据区分方法实现。数据特征化是目标类数据的一般特征或特性的汇总。数据特征化的输出可以用多种形式提供，包括饼图、条图、曲线、多维数据立方体、含交叉表的多维表。结果描述也可以用概化关系或规则形式提供；数据区分是将目标类对象的一般特性与一个或多个对比类对象的一般特性比较。用于数据区分的方法与用于数据特征化的类似。数据区分的输出类似于数据特征化，但它应该包括比较度量，帮助区分目标类和对比类。用规则表示的区分描述称为区分规则。

2.关联规则

关联分析发现关联规则，关联分析广泛用于购物篮或事务数据分析。有一个关联规则的例子就是“90%的顾客在购买面包和黄油的同时也会购买牛奶”，其直观意义为顾客在购买某些商品的时候有多大倾向会购买另外一些商品。关联规则的一般形式为 $X_1 \wedge \cdots \wedge X_n \Rightarrow Y[C,S]$，表示由 $X_1 \wedge \cdots \wedge X_n$ 可以预测 Y，其置信度为 C，支持度为 S。

3.分类

数据分类实际上就是从数据库中发现数据对象共性，并将数据对象分成不同几类的一个过程。首先是对训练数据进行分析，使用数据的某些特征属性，给出每个类的准确描述(即分类规则)，然后使用这些描述，对数据库中的其它数据进行分类。实际上，分类是一个二步过程，第一步，建立一个模型，描述指定的数据类集或概念集；第二步，使用模型进行分类。模型的建立是基于对训练数据集的分析。模型可以用多种形式表示，如分类规则、判定树、数学公式或神经网络等。例如，通过训练数据获得了下列规则。

IF age=“31…40” AND income=high THEN credit_ rating=excellent

用户可先用测试数据评估分类规则的准确率。如果准确率是可以接受的，则该规则可以用于对新的数据对象进行分类。

4.聚类

将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程称为聚类。对象根据最大化类内的相似性和最小化类间的相似性的原则进行聚类或分组。所形成的每个簇可以作为一个对象类，由它可以导出规则，在许多应用中，可以将一个簇中的数据对象作为一个整体来对待。

聚类与分类不同，聚类分析的输入数据集是一组未标记的对象，也就是说，此时输入的对象还没有被进行任何分类。聚类的目的是根据一定的规则，合理地进行分组或聚类，并用显式或隐式的方法描述不同的类别。所依赖的这些规则是由聚类分析方法定义的。由于分析可以采用不同的方法，所以对于相同的数据集合可能有不同的划分。

5.孤立点分析

数据库中可能包括一些数据对象，它们与数据的一般行为或模型不一致。这些数据对象被称作孤立点。大部分数据挖掘方法将孤立点视为噪声或异常而丢弃。然而，在某些应用中(如欺骗检测)，罕见的事件可能比正常出现的事件更有趣，需要进行孤立点数据分析。孤立点分析也称作孤立点挖掘。

孤立点可以使用统计实验检测来发现。假定有一个数据分布或概率模型，并使用距离度量，到所有聚类的距离很远的对象被视为孤立点。此外，孤立点也可以使用基于偏差的方法来发现，通过考察若干对象主要特征上的差别来识别孤立点。

6.演变分析

演变分析描述行为随时间变化的对象的规律或趋势，比如，时间序列数据分析(如:先购买房屋，随后购买装饰材料，接下来购买家具等等)，并对其建模。

2.1.3 数据挖掘方法

数据挖掘的方法大部分来自于机器学习、人工智能、统计学等领域，他们分别从不同的角度进行数据挖掘，常用的可分为下列几种[5]：

1．决策树方法

利用树形结构来表示决策集合，这些决策集合通过对数据集的分类产生规则。它利用信息论中的信息增益寻找数据库中具有最大信息量的字段，建立决策树的一个节点，再根据字段的不同取值建立树的分支；在每个分支子集中重复建树的下层节点和分支，即可建立决策树。国际上最有影响和最早的决策树方法是由Quinlan研制的ID3方法，后人又发展了其它的决策树方法[6]。

决策树方法的最大优点是直观，其缺点是随着数据复杂性的提高，分支数将增加，管理的难度越来越大。此外，该方法也存在数据的缺值处理问题。

2．规则归纳方法

通过统计方法归纳、提取有价值的 if- then 规则。规则归纳的技术在数据挖掘中被广泛使用，其中以关联规则挖掘的研究开展得较为积极和深入[7]。

3．神经网络方法

从结构上模拟生物神经网络，以模型和学习规则为基础，建立三大类多种神经网络模型:前馈式网络、反馈式网络、自组织网络。这是一种通过训练来学习的非线性预测模型，可以完成分类、聚类、特征挖掘等多种数据挖掘任务。

神经网络[8]的最大优点是它能对复杂问题进行精确的预测。但也存在神经网络难于理解、神经网络易受训练过度的影响、神经网络的训练时间较长等不足。

4．遗传算法

模拟生物进化过程的算法，由繁殖(选择)、交叉(重组)、变异(突变)三个基本算子组成。为了应用遗传算法，需要将数据挖掘任务表达为一种搜索问题，从而发挥遗传算法的优化搜索能力。

遗传算法[9]擅长于数据聚类，能够解决其它技术难以解决的问题。遗传算法通常与神经网络结合起来使用，以在较高的层次上提高模型的可理解性。

5．模糊集

模糊集[10]是表示和处理不确定性数据的重要方法。模糊集不仅可以处理不完全数据、噪声或不精确数据，而且在开发数据的不确定性模型方面是很有用的，它能提供比传统方法更灵巧、更平滑的性能。

6．粗糙集(Rough Set)方法

Rough 集[11]理论是一种处理含糊和不精确性问题的新型数学工具。它特别适合于数据简化、数据相关性的发现、发现数据意义、发现数据的相似或差别、发现数据模式、数据的近似分类等，近年来已被成功地应用在数据挖掘和知识发现研究领域中。与模糊集一样，它是一种处理数据不确定性的数学工具，常与规则归纳、分类和聚类方法结合使用，很少单独使用。

7．可视化技术

将信息模式、数据的关联或趋势等以直观的图形方式表示，决策者可以通过可视化技术[12]交互地分析数据关系。可视化数据分析技术拓宽了传统的图表功能，使用户对数据的剖析更清楚。

8．统计方法

统计方法是从事物外在数量上的表现去推断该事物可能的规律性。通常是先通过统计从其数量表现上分析出一些线索，然后提出一定的假说或学说，再作进一步深入的理论研究。当理论研究得出一些结论时，往往还需要在实践中加以验证。统计方法的优点是精确、易理解，并且已广泛使用。其缺点是很难有效使用。常见的统计方法有回归分析[13]、判别分析[14]、聚类分析[15]、探索性分析[16]等。目前流行的统计软件有 SAS 和 SPSS。

2.1.4 数据挖掘的应用领域

数据挖掘在很多重要领域，尤其是在金融、投资、保险、交通、零售等应用领域发挥了积极促进作用，能为企业作出前瞻性的、基于知识的决策参考意见。在北美，数据挖掘技术已经成功应用于社会生活的方方面面，根据麻省理工学院的《科技评论》评估，“数据挖掘技术”是对未来人类产生重大影响的十大新兴技术之一。随着应用研究的不断深入，数据挖掘将不再是一门高深的理论，将成为一个极具价值的实用技术[17]。

1．数据挖掘在金融业的应用。

在银行业，数据挖掘主要用于信用欺诈的建模和预测、风险评估、趋势分析、收益分析以及辅助直销活动。在金融市场，已将神经网络用于股票价格预测、购买权交易、债券等级评定、资产组合管理、商品价格预测、合并和买进以及金融危机预测等方面。

2．数据挖掘在市场营销业的应用。

数据挖掘在市场营销中的应用可分为两类:数据库市场营销和购物篮分析。前者的任务是通过交互查询、数据分割和模型预测等方法来选择有潜力的顾客以便向他们推销产品。后者的任务是分析市场销售数据(如 POS 数据库)以识别顾客的购买行为模式，从而帮助确定商店货架的布局，促进商品的销售。数据挖掘工具可以用于进行商品销售预测、商品价格分析、零售点的选择等。

3．数据挖掘在生物医学上的应用。

数据挖掘在生物医学上的应用主要集中于分子生物学，尤其是基因工程的研究。它在分子生物学上的工作可分为两种：一是从各种生物体的 DNA 序列中定位出具有某种功能的基因串；二是在基因数据库中搜索与某种具有高阶结构或功能的蛋白质相似的高阶结构序列。

4．数据挖掘在科学研究中的应用。

数据挖掘对高科技的研究是必不可少的。高科技研究的特点就是探索人类未知的秘密，而这正是数据挖掘的特长所在，借助数据挖掘技术从大量的、漫无头绪而且真伪难辨的科学数据和资料中要提炼出对人类有用的信息。

5．数据挖掘在社会科学研究领域的应用前景。

社会科学的特点是从历史看未来，如从社会发展的历史进程中得出社会发展的规律，预测社会发展的趋势；从人类发展的进程和人类的社会行为的变化中寻求对人类行为规律的答案，从而应用于对各种各样的社会问题的求解。数据挖掘在从历史数据中进行规律的发现方面，也有其独到的作用。

6．数据挖掘在其它一些领域的应用。

数据挖掘技术应用于公安司法工作，有利于案件调查、案例分析、犯罪监控

等等，还可用于犯罪行为特征的分析，这也是本论文重点研究和描述的课题。在工业部门，数据挖掘技术可用于进行故障诊断、生产过程优化等。

2.1.5 数据挖掘研究的动态发展趋势

数据挖掘是一个新兴的研究领域，许多问题还有待于研究，目前的研究方向包括下列几个方面[18]：

(1)算法效率和可伸缩性。目前，Gb 和 Tb 规模的数据库也已经在使用，Pb 规模的数据库正在出现。为了保证高效率，运用到大型数据库中的数据挖掘算法应该是高度可伸缩的。

(2)处理不同类型的数据和数据源。数据库中将包含大量复杂的数据类型。如结构化的数据，复杂的数据对象，混合文本，多媒体数据，时空数据，事务数据及历史数据等，甚至出现新的数据库模型。因此，保证数据挖掘系统能有效地处理此类数据库中的数据是至关重要的。

(3)数据挖掘系统的交互性。数据挖掘中操作者的适当参与能加速数据挖掘过程。准确而直观地描述挖掘结果和友好而高效的用户界面一直是研究的重要课题。

(4)Web 挖掘。由于 Web 上存在大量信息，并且 Web 在当今社会扮演越来越重要的角色，因此，Web 挖掘将成为数据挖掘中一个重要和繁荣的子领域。

(5)数据挖掘中的隐私保护与信息安全。随着计算机网络的日益普及，研究数据挖掘可能导致的非法数据入侵是实际应用中亟待解决的问题之一。

(6)数据挖掘语言的标准化。标准的数据挖掘语言或有关方面的标准化工作将有助于数据挖掘系统的研究和开发，有利于用户学习和使用数据挖掘系统。

(7)数据挖掘结果的可用性、确定性及可表达性。所发现的知识需精确地描述数据库的内容，并对己明确的应用是有用的。非精确的结果需借助于不确定性来表达，以相似的规则或多个规则来描述。噪声及应去除的数据在数据挖掘系统中应仔细处理。对发现的知识如何表达是一个系统性的研究项目。

(8)可视化数据挖掘。是从大量数据中发现知识的有效途径。系统研究和开发可视化数据挖掘技术将有助于推进数据挖掘作为数据分析的基本工具。

2.2 公安信息化

公安信息化的基本内涵是运用信息及通信技术打破行政机关的组织界限，使得人们可以从不同渠道获取政府信息及服务，其核心目的在于利用最先进的电子化方式对信息资源进行系统有效的管理，实现公安信息共享。公安信息化是社会信息化的组成部分，其体系主要由信息资源、信息网络、信息技术、信息技术应用、信息化人才队伍、信息化运行标准规范等六方面要素构成。公安信息化是新世纪的社会发展趋势，从技术角度讲，公安信息化就是执法数字化、警务数字化。

2.2.1 公安工作与信息化

进入新世纪，新时期，受经济、科技迅猛发展态势和国际国内犯罪格局变动影响，以及各种消极因素综合作用，现代违法犯罪进入一个全面提速的新阶段。犯罪智能化构筑了犯罪加速变化的基础。网络速度提升了犯罪速度。犯罪的内部分化与反向运动助推犯罪速度变化，在大部分犯罪提速发展的同时，某些现存的犯罪现象演化减缓，甚至停滞、消失。如一部分传统的犯罪虽然表现出继续存在的可能性，但其范围大大缩小，频率大大降低，危害程度也会越来越有限。传统犯罪萎缩与新型犯罪崛起形成强烈反差。

犯罪速度化对社会的影响是巨大的，也对公安警务工作提出了严峻的挑战。犯罪提速增加了治安严峻性，反映出社会关注的“热点”和整治难点。特别是新型犯罪在时间、空间内的快速运行、传播，直接威胁群众的安全感，尤其是暴力犯罪增多、增速对社会安全心理冲击极大，强化人们对治安形势恶性化的主观认识。犯罪速度构成社会公众安全感的重要指标。犯罪快速流动表明，客观上存在打击不力，不足以形成威慑、阻吓犯罪的氛围和存在治安防控网络疏漏，不足以控制犯罪的问题，当犯罪速度成为现代犯罪重要现象，并对社会治安产生综合影响力时，充分评估犯罪速度因素趋势预测、犯罪走势判断的准确性、客观性，有利于增加组织现代警务活动和开展反犯罪斗争的科学性、针对性和有效性[19]。

当前，公安侦查工作中破案方式单一，破案渠道少，侦查手段不全，侦破效率不高的状况依然存在。传统的破案模式是发案以后，侦查员通过现场勘察、案情分析、制定措施寻找发现犯罪分子[20]。

这种侦查模式可以说是一种静态侦查模式。它的弊病在于：从侦查学角度看，这是一种反应性的侦查模式，是在发生了刑事案件以后才进行的侦查，具有很大的滞后性；从思维角度看，它侧重于从案件出发寻找相应犯罪嫌疑人这一正向思维，具有明显的被动性；从方法手段上看，死守排队摸底、挖地三尺、三板斧的体力型侦查活动，具有明显局限性；从效率上看，因为沿用人海战术，依靠人力密集型劳动，效率很低。因此，这是一种以人力劳动为主，科技含量低的粗放型破案方式，不能适应市场经济条件下工作的需要。而公安信息化可以实现积极主动以信息网络技术为核心的动态侦查模式。

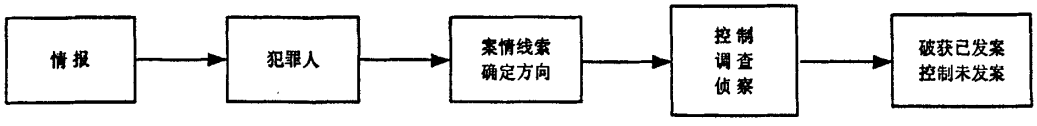

图 2.2 动态侦察模式

图 2.2 所示的动态侦查方式使信息资源被最大限度地运用，将信息资源优势转化为现实破案力，尤其能把一些预谋性案件扼杀于摇篮之中，真正做到“打击

犯罪，预防为主”，改变过去围着案子跑的局面，形成具有时代特征的侦查破案的新局面[21]。公安机关开展的“指纹会战”、“网上追逃”，一些刑侦部门添置了心理测试仪等仪器设施，这就极大地增强了公安机关的侦破能力。利用犯罪信息网络侦查破案，已成为现代刑事侦查中不可缺少的手段，在公安工作中发挥出越来越显著的作用。

以“金盾工程”为载体，以“电子警务”为核心，以实现公安工作信息化为目标的数字化公安建设，是顺应社会和历史潮流，适应现代犯罪发展和犯罪速度变化，以高科技手段打击高科技犯罪的一场新的警务革命。公安工作信息化找到了公安工作与现代科学技术的最佳结合点，加快公安信息化建设，将信息资源转化为破案力，是培育新的破案增长点的必由之路，是实现公安工作跨越式发展的重要途径，更是现实斗争的需要。

2.2.2 金盾工程

金盾工程是公安信息化建设的重要工程，是公安部门利用现代的信息通信技术，增强统一指挥、快速反应、协调作战、打击犯罪的能力，是提高公安工作效率和侦察破案水平，以适应我国在现代经济和社会条件下实现动态管理和打击犯罪的需要，是实现科技强警目标的重要举措[22]。2003 年 9 月 2 日召开的全国“金盾工程”工作会议，决定“金盾工程”正式启动。

1．金盾工程的主要内容

建设公安业务综合数据通信网。实现全国公安广域网和各级公安机关局域网的联网；建设全国各级公安机关无线通信专网和公安卫星通信专网。实现网络基本服务，开展会议电视、远程电视教学、视频传输、移动数据终端、公安专用传呼等业务应用。

建设全国公安综合信息系统。开发、建设和完善国家违法犯罪信息中心(CCIC)、各类公安业务应用(刑侦、经侦、治安、监管、边防、外管、消防、交通管理、计算机安全监察、禁毒、警用装备等)信息系统。

建设公安网络安全保障体系。核心是保证信息的安全，保障信息的机密性、完整性、抗否认和可用性。保证网络 24 小时不间断可靠运行。在公安信息专网上实现安全认证(CA)、对访问的有效控制、应用系统的安全、信息的安全的传输和系统容错备份。对秘密信息的传输，采用 VPN 技术，加密传输。

建设和完善公安各级指挥调度系统。加速信息、通信、指挥系统的技术建设，，实现多媒体信息通信，满足统一指挥、快速反应、协同作战的要求。

建设全国公共信息网络安全监控中心。提高对网上有害信息的监控和发现能力、提高对计算机犯罪的现场勘查取证和电子数据鉴定等案件的侦查能力，提高信息网络安全监督管理能力，确保网络运行安全和网上信息安全。

制定与完善公安信息化标准和规范体系。保障各地建设互联互通，保障信息系统之间信息共享，保障信息安全。

2. 公安信息系统结构

“金盾工程”中公安信息系统的内部结构自上而下分为访问接入层、应用核心层、数据访问层。三层既相对独立、又相互联系，下层为上层提供支持、上层是下层操作的来源[23]。如图 2.3 所示

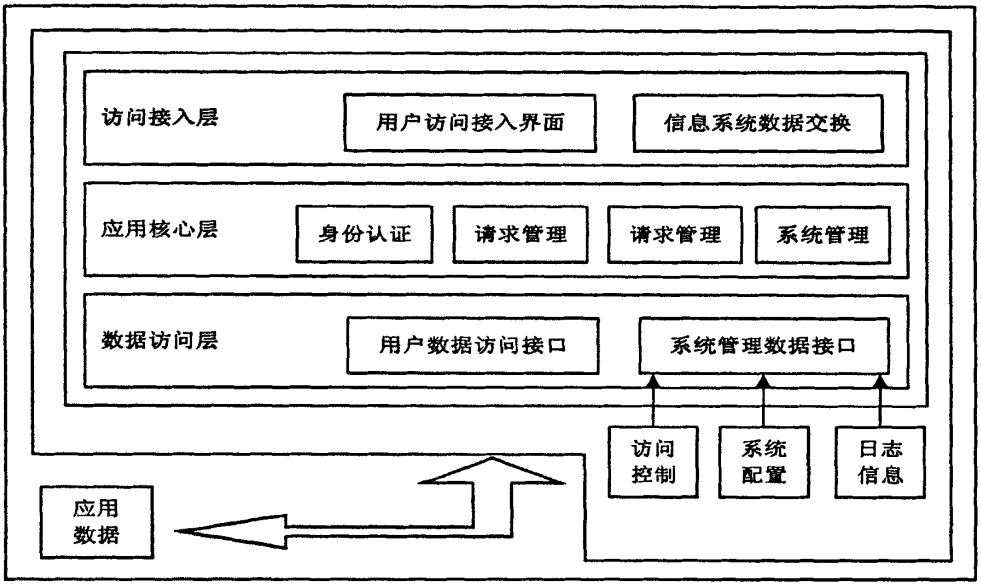

图 2.3 “金盾工程”内部结构

“金盾工程”中，公安信息系统外部结构是指与外部的接口，包括与其它信息系统的接口、与用户的接口、与数据库系统的接口、与身份认证中心的接口、与请求管理中心的接口等，如图 2.4 所示。

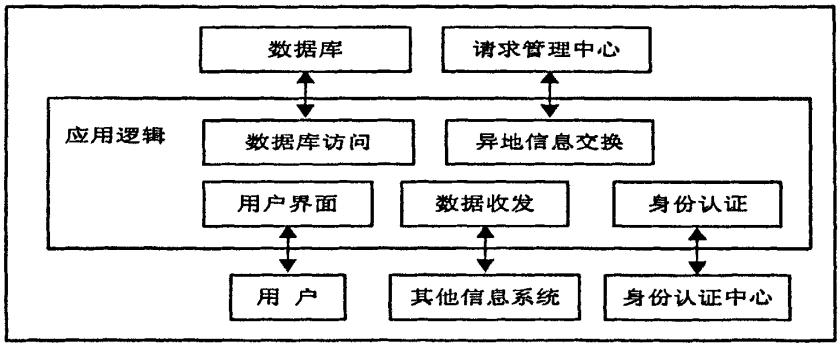

图 2.4 公安信息系统外部结构

2.2.3 我省公安信息化现状

目前，我省公安信息化网络已初具规模，基本完成公安二级、三级和接入网络的建设，各级公安机关局域网已全面开通，全省公安网注册计算机已达 24600 多台，注册服务器达到近 2000 台。全省已建信息系统 46 个，其中一类系统 22 个，

二类系统 9 个，其他系统 15 个，数据量记录总数 1.7 亿条，各类系统存储的数据累计容量已经超过 10T。各类应用业务系统在公安工作中发挥了重要作用，如人口信息系统、车驾管信息系统、出入境人员信息系统、综合信息查询系统等。已建的应用系统涉及到大部分的警种，覆盖了我省公安的主要业务，使部分业务实现了流程化信息管理，为公安各项业务活动发挥了重要作用，也大大提高了我省公安执法水平和快速反应能力。

但我省公安信息化应用客观上也存在一些问题：一是应用面不广。没用形成全警采集、全警应用、全警共享的公安信息化应用格局；二是应用标准不统一。在各业务系统中标准不统一，为数据共享交换设置了障碍；三是部分信息系统设计开发水平不高，功能不完善；四是部分信息系统数据质量不高；五是信息共享与综合利用水平不高；六是缺乏整体规划，发展不平衡；七是缺乏高端应用。基本没有运用数据挖掘技术建设公安情报信息系统等专题应用、研判分析、决策预警等高端应用，提高信息资源的综合开发利用，实现辅助领导决策的等作用。

2.3 数据挖掘在公安工作中的应用

公安信息化系统功能可分为四个层次，其中下层是上层的基础，上层是下层功能的提升和扩展[24] [25]。

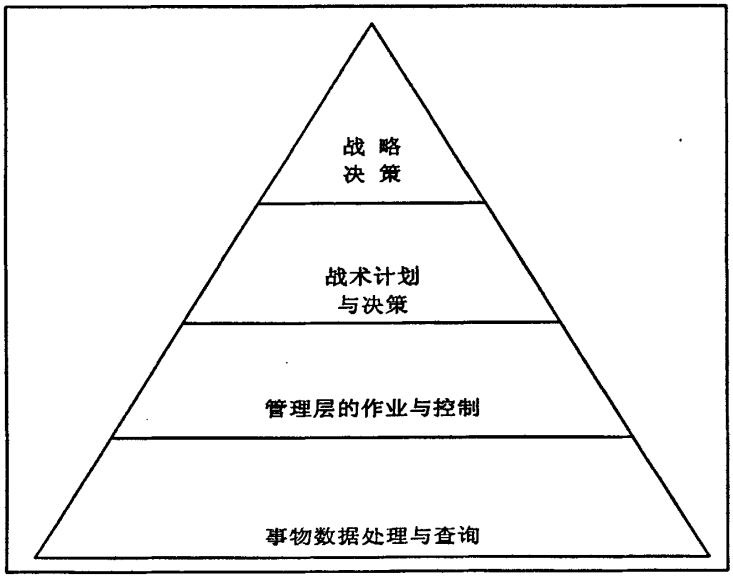

图 2.5　公安信息系统功能

如图 2.5 所示，一层是事务数据处理与查询。二层是管理与控制。为实现整体管理目标，对各种管理信息进行系统、综合处理，辅助各级工作人员进行管理。三层是战术计划与决策。指调用当前和历史信息，使用各种数学模型和方法进行模拟或预测，为领导层决策提供所需数据、信息和背景资料。并对各种方案进行评价和选优，通过人机对话进行分析、比较和判断，为正确决策提供帮助。四层

是战略决策与计划。在战术决策基础上，综合利用管理利学、运筹学、控制论等科学领域的成果，对长远发展目标进行决策。目前，公安信息化体系发挥的作用主要集中在功能表的第一层，第二层相对较少，第三层和第四层则基本没有。

数据挖掘技术的提出正是解决了第三层和第四层的工作，在海量的数据中提取所需要的知识，为科学的进行决策提供了必要的帮助。国外执法系统也在这方面进行了有效的应用，FBI(美国联邦调查局)利用数据仓库技术建立了 NCIC(国家犯罪信息中心)，并通过数据挖掘技术对人员犯罪、调查取证、情报分析以及案件预警等方面进行了卓有成效的工作。

正是由于公安行业的特殊性，应用数据挖掘技术对信息化系统中的海量数据进行挖掘，要以快速准确辅助警务决策、指导警务工作、提高执法效率，这是公安信息化的未来发展方向。公安信息化，除了要满足公安日常业务处理查询的需要，必须要向智能化信息系统发展，以适应未来形势的需要，提高公安快速响应能力与作战能力，更好的保护国家和人民生命财产安全，维护社会长治久安的局面，为国家经济建设服务。

2.4 数据挖掘在公安工作中的展望

随着“金盾工程”的全面启动，我国电子警务建设已经初步成型：信息基础建设已形成一定规模，公安业务管理信息系统建设和应用逐渐普及。在预防和打击犯罪活动方面，建立起在逃人员信息资源库、违法犯罪人员信息资源库、被盗抢汽车信息资源库等多个一级资源库，DNA 数据库、无名尸体和失踪人员数据库等二级资源库。人口管理信息系统常住人口已经超过 12 亿，260 个一类边防检查口岸全面实现了计算机管理。在交通管理方面，全国 400 个车管所全部实行了计算机管理，已存储 4000 多万辆机动车和 7000 多万名驾驶员的信息，并实现全国联网浏览查询；公安综合信息查询应用初见成效，公安机关信息化应用已经全面铺开。公安部信息中心信息量超过 21000 条，总访问量超过 6810 万人次，日访问量超过 17 万人。

数据挖掘技术可以公安应用领域的以下几个方面取得进展和突破[26]：

(1)对违法犯罪行为的分析研究。犯罪行为分析本身是一门涉及面极广的学科，渗透了法学、心理学、行为学等多门学科，需要相当的专门知识，其本身现在还处在不断探索研究的阶段[27]。利用计算机数据挖掘技术，在拟定算法下对大量的犯罪行为记录进行分析，从而发现犯罪的规律、趋势，了解不同犯罪行为之间的关联，以及何种状态会诱发何种犯罪行为。这也是公安司法等相关领域所迫切需要的，具有相当重要的现实意义。本文正是基于此方面的应用研究工作。

(2)对交通管理的决策。交通管理是城市管理的重要问题，传统的手工劳动式的交通指挥已经不能适应实际工作的需要，应用数据挖掘和 OLAP 技术，实时监测

路面状况和交通流量，及时制定对策，有效疏导交通阻塞，是未来交通管理的重要方向。

(3)对警力安排的决策。如何科学、合理地安排警力，在治安状况多变的情况下，既保证社会秩序，又不浪费警力，也是今后工作的重点之一。

(4)消防调度决策。消防工作具有很强的时间性，其调度具有极强的科学性，在人员、车辆配置、水源安排、最佳路线选择等方面都具有大量的信息可供挖掘。

(5)人口管理。可对常住人口、外来人口、暂住人口等的居住时间、性别、年龄、文化程度、从事职业等方面进行统计分析及关联分析,以预防和打击犯罪。

数据挖掘技术作为一门新兴科学，将其有效应用于公安犯罪分析中是公安工作现实斗争的需要，论文在这方面做了一定的工作，但公安工作是项很复杂的工作，算法实际应用到公安工作并实现警务决策、指导警务工作实际还有一定距离。

2.5 本章小结

本章首先介绍了数据挖掘的定义、任务、方法、应用领域、动态发展趋势，然后介绍了当前公安工作现状、公安信息化建设现状、公安信息化建设的重要工程－金盾工程的主要内容和系统结构，我省公安信息化现状，最后介绍了数据挖掘在公安工作中的应用和展望。

第 3 章　违法犯罪行为的关联性分析

3.1 引言

关联规则挖掘是数据挖掘中最活跃的研究方向之一。最早是由 Agrawal 等人于 1993 年针对购物篮分析(Basket Analysis)问题提出的，其目的是为了发现事务数据库(Transaction Database)中不同商品之间的联系规则，通过对这些数据的智能分析，可以获得有关顾客购买模式的一般性规则，可以用来指导商家科学的安排进货、库存以及货架设计。除了购物篮分析外，关联规则挖掘还可以应用到其他领域，如医疗诊断、生物信息学、网页挖掘和科学数据分析。有一个比较经典的例子就是“90%的顾客在购买面包和黄油的同时也会购买牛奶”，其直观意义是顾客在购买某些商品的时候，有多大倾向会购买另外一些商品[28]。

后来，很多的研究者对关联规则的挖掘问题进行了大量的研究。他们的工作涉及关联规则的挖掘理论的探索、原有算法的改进和新算法的设计、并行关联规则挖掘以及数量关联规则挖掘等问题。许多学者在提高挖掘规则算法的效率、适应性、可用性以及应用推广等方面进行了广泛的研究。

3.2 关联规则及 Apriori 算法

3.2.1 关联规则基本概念

设 $I=\{i_1, i_2, \cdots, i_n\}$ 是项的集合。设任务相关的数据 D 为数据库事务的集合，其中每个事务 T 是项的集合，使得 $T \subseteq I$。每一个事务有一个标识符，称作 TID。设 A 是一个项集，事务 T 包含 A 当且仅当 $A \subseteq T$。关联规则是形如 $A \Rightarrow B$ 的蕴涵式，其中 $A \subset I$，$B \subset I$，并且 $A \cap B = \varnothing$。

项的集合称项集(itemset)。包含 k 个项的项集称为 k-项集。项集的出现频率是包含项集的事务数，简称项集的频率、支持计数或计数。如果项集满足最小支持度，则称它为频繁项集(frequent itemset)。频繁 k-项集的集合通常记作 L_K。

关联规则的强度可以用支持度(Support)和置信度(Confidence)度量[29]。支持度确定规则可以用于给定数据集的频繁程度，反映发现规则的有用性；置信度确定 $A \cup B$ 在包含 A 的事务中出现的频繁程度，反映发现规则的确定性。记作：

$$\text{support}(A \Rightarrow B) = P(A \cup B) = \frac{\text{sup port\_count}(A \cup B)}{|D|} \tag{3.1}$$

$$\text{confidence}(A \Rightarrow B) = P(A|B) = \frac{\text{sup port\_count}(A \cup B)}{\text{sup port\_count}(A)} \tag{3.2}$$

给定一个事务集 D，挖掘关联规则问题就是产生支持度和置信度分别大于用户

给定的最小支持度(min_sup)和最小置信度(min_conf)的关联规则[30]。

3.2.2 关联规则的种类

1．按照规则中处理的变量的类别，关联规则可以分为布尔型和数值型。

布尔型关联规则处理的值都是离散的、种类化的[31]，它显示了这些变量之间的关系；而数值型关联规则可以和多维关联或多层关联规则结合起来，对数值型字段进行处理，将其进行动态的分割，或者直接对原始的数据进行处理，当然数值型关联规则中也可以包含种类变量。

例如：国籍＝“中国”⇒语言＝“汉语”，是布尔型关联规则；国籍＝“中国”⇒avg(寿命)=59，涉及的年龄是数值类型，所以是一个数值型关联规则。

2．按照规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

在单层的关联规则中，把所有的变量都当成是同一个层次上的，没有考虑到现实的数据是具有多个不同的层次的；而在多层的关联规则中，对数据的多层性已经进行了充分的考虑，并把这种多层性应用于规则的发现过程。

例如：居住在中国的中国人⇒说汉语，是一个具体细节数据的单层关联规则；中国人⇒说汉语，是一个较高层次和细节层次之间的多层关联规则。

3．基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。

在单维的关联规则中，只涉及到数据的一个维，如用户购买的物品；而在多维的关联规则中，要处理的数据将会涉及多个维。即：单维关联规则是处理单个属性中的一些关系；多维关联规则是处理各个属性之间的某些关系。

例如：咖啡⇒咖啡伴侣，这条规则只涉及到用户购买的物品；国籍=“中国”⇒语言=“汉语”，这条规则就涉及到两个字段的信息，是两个维上的一条关联规则[14]。

3.2.3 关联规则 Apriori 算法思想

Apriori 算法是由 Agrawal 提出的关联规则挖掘的重要经典算法[32]。它使用基于支持度的剪枝技术，系统的控制候选项集指数增长。它的重要理论基础是：频繁项集的所有非空子集都必须也是频繁的、非频繁项集的超集也是非频繁的。

Apriori 算法是一个基于两阶段频集思想的方法，它将关联规则挖掘算法的设计分解为两个子问题：

(1)找到所有频繁项集：是发现满足最小支持度阈值的所有项集，这些项集称为频繁项集。

(2)由频繁项集产生强关联规则：从上一步发现的频繁项集中提取所有高置信度的规则，这些规则称作强规则(strong rule)。

Apriori 是一种宽度优先算法，通过对数据库 D 的多次扫描来发现所有的频繁项目集，在每一次扫描中只考虑具有同一长度(即项目集中所含项目的个数)的所

有项目集。在第一次扫描中，Apriori 算法计算 D 中所有单项目的支持度，生成所有长度为 1 的频繁项目集。在后续的每一次扫描中，首先以 k-1 次扫描所生成的所有频繁项目集为基础产生新的侯选项目集，然后扫描数据库 D，计算这些侯选项目集的支持度，删除其支持度低于用户给定最小支持度的项目集，最后，生成所有长度为 k 的频繁项目集。重复上述过程直到再也发现不了新的频繁项目集为止。

3.2.4 Apriori 算法的性能瓶颈问题

Apriori 作为经典的频繁项目集生成算法，在数据挖掘中具有重要的里程碑的作用。但是随着研究的深入，它的缺点也暴露出来。Apriori 算法有两个致命的性能瓶颈[33]。

1．多次扫描事务数据库，需要很大的 I/O 负载。

对每次 k 循环，候选集 C_k 中的每个元素都必须通过扫描数据库一次来验证其是否加入 L_k。假如一个频繁大项目集包含 10 个项，那么就至少需要扫描事务数据库 10 遍。

2．可能产生庞大的候选集。

由 L_{k-1} 产生 k-候选集 C_k 是指数增长的，例如 10^4 个 1-频繁项目集就有可能产生接近 10^7 个元素的 2-候选集。如此大的候选集对时间和主存空间都是一种挑战。

为了提高 Apriori 算法的效率，出现了一系列的改进算法：

(1)基于 Hash 的项集计数：如果一个 k-项集在 hash-tree 的路径上的一个计数值低于阈值，那它本身也不可能是频繁的。

(2)减少事务记录：不包含任何频繁 k-项集的事务也不可能包含任何(k+1)-的频繁集。

(3)分割：一个项集要想在整个数据库中是频繁的，那么它至少在数据库的一个分割上是频繁的。

(4)采样：在给定数据的子集上挖掘，使用小的支持度+完整性验证方法[34]。

(5)动态项集计数：在添加一个新的候选集之前，先估计一下是不是它的所有子集都是频繁的。

这些算法由于引入了划分(Partitioning)[35]、散列(Hashing)[36]、抽样(Sampling)[37]等方法，在一定程度上改善了算法的适应性和效率。

3.3 关联规则挖掘在违法犯罪行为应用中存在的问题

3.3.1 在违法犯罪行为中的应用

关联分析是为了挖掘出隐藏在数据间的相互关系，从一组给定的数据项以及交易集合中，分析出数据项集在交易集合中出现的频度关系。关联规则 $A \Rightarrow B$ 的解释为“满足 A 中条件的数据库元组多半也满足 B 中条件”。从大量的违法犯罪涉嫌

人员记录中发现犯罪程度和客观因素的关联关系，可有助于许多政策的制定。通过发现引发犯罪的不同因素之间的联系以及哪些因素频繁起作用，这种关联的发现可以帮助有关部门制定政策，如加强义务教育，法制教育，职业教育等，提高公民的素质，并让众多人拥有一技之长，增强其谋生能力，以此降低违法犯罪率。

公安业务信息系统中存有大量犯罪记录，在拟定的算法下对大量的犯罪行为案例记录进行分析，从而发现犯罪的规律、趋势，了解不同犯罪行为之间的关联，以及何种状态会诱发何种犯罪行为等等。从大量案例记录的挖掘中找出犯罪趋势，对于预防犯罪，加强对重点地区进行重点监控有着相当重要的意义[38]。

违法案件中的几种关系：

(1)人与人的关系：犯罪嫌疑人与同案犯的关系，组织者与被组织者、领导者与一般成员等；犯罪嫌疑人与被害人的关系。

(2)人与事的关系：犯罪嫌疑人与发生在案件整个过程之中的一些事件的关系，该犯罪嫌疑人对事件的影响等。注意应该分清组织责任与个人责任。

(3)人与物的关系：犯罪嫌疑人与涉案相关物证的关系，需要弄清楚物属关系，物品在案件中所起到的作用等等。

(4)事与物的关系：案件事实与物证的关系。

3.3.2 存在的问题

由于公安工作的特殊性，Apriori算法即使进行了优化，但在违法犯罪行为分析中存在的一些固有的缺陷还是无法克服的，无法有效满足公安执法理念和警务创新的新要求，主要表现在以下两方面：

1．忽略了不同项具有不同重要性问题。

传统Apriori算法仅仅考虑了被分析的项在数据库中的出现频率，而没有考虑不同的项具有不同的重要性问题，只是简单地将所有项目都视为具有同等价值。这个问题的存在就常常会导致那些具有重要价值，但是出现频率却相对较小的项被忽略。如果采用传统Apriori算法进行公安犯罪行为分析时，那些情节特别严重，社会危害特别大的重大犯罪行为将由于出现频率小而被忽略，这就降低了公安犯罪行为分析结果的有效性，如果以这种分析结果去指导日常警务工作，必然会使警务工作偏离正确的方向，会给工作中留下一些重大的隐患。

2．无法有效解决对新项目的敏感性问题。

减少或增加项目会使项目间的关联发生变化，从而产生新的关联规则。但传统的Apriori算法没考虑到这个问题，即使近期大批增加新的项目，在求各个项目集的支持度时，都是以整个数据库的犯罪行为记录总数作为基数。因此，极有可能出现新的项目在新数据集中频繁出现并产生新的关联规则，但在整个犯罪行为数据库中却是非频繁集，也不能产生最新的关联规则，这与数据挖掘的目的不符，

也发现不了新颖的、潜在有用的模式。Apriori算法在进行公安犯罪行为分析时，将无法及时、敏锐的发现那些新型犯罪行为，必然会造成预防、打击不力，客观上使新的犯罪行为在时间、空间上扩散，形成热点、焦点、难点。

3.4 一种改进的加权关联规则挖掘算法 WARMA

传统Apriori算法在挖掘频繁项集时，实际上存在两大前提假设：一是假设数据库中各项目具有相同的性质和功能，即重要性相同；二是假设数据库中各项目的分配是均匀的，即出现频率相同或相似，即数据库中的各项目以平等一致的方式处理。然而，在实际生活中，不同的项目间往往存在重要性差异和分配不均匀性。如在零售业中，市场经理或许想在挖掘关联规则时，对某些特定的产品(如利润比较高或正在促销的商品)给予更多关注，而对其他的产品则相对不十分关注。

实验1：某商场的事务集D如表3.1所示，I_1、I_2、I_3、I_4，I_5分别表示电池、高档耳机、高档影碟机、随身听、CD片。设min_sup=40%，min_conf=60%，按照传统Apriori算法可以得出表3.2所示的结果。

表3.1 某商场的事务集D

TID	项集
T_1	I_1, I_2, I_5
T_2	I_2, I_4
T_3	I_1, I_4
T_4	I_3, I_5
T_5	I_2, I_4, I_5
T_6	I_2, I_4
T_7	I_1, I_3, I_5
T_8	I_2, I_4
T_9	I_2, I_5
T_{10}	I_4, I_5

表3.2 用传统Apriori算法生成的关联规则

规 则	支持度	置信度
$I_2 \Rightarrow I_4$	40%	66.7%
$I_5 \Rightarrow I_3$	20%	33.3%

从表3.2可以看出，如果把高档耳机和随身听摆放在一起，将有利于销售。但实际上，高档影碟机的利润要远远大于其他商品，为了追求利润最大化，商场经理们显然会对高档影碟机的销售更感兴趣。实验1却挖掘不到任何与高档影碟机有关的信息，这样的关联规则挖掘是不令人满意的。传统Apriori算法处理时，就可能把一些重要性的关联规则忽略，这是因为I_1、I_2、I_3、I_4的重要程度不同。

同样，在公安工作中，那些犯罪情节特别严重，社会影响特别大、给国家和人民生命财产安全造成重大损失的犯罪行为显然需要予以特别关注。但是，在实际生活中，按照一般规律，这些重大的犯罪行为(如跨区域重大犯罪、恐怖主义、涉枪、涉毒、黑恶势力犯罪等)出现的频率往往远远小于那些一般性的普通犯罪行为(如盗窃、赌博、诈骗等)，采用传统Apriori算法进行犯罪行为分析时，这类犯罪行为将可能被忽略掉。因此，在公安犯罪行为中，我们要本着重要性原则，对不同性质的犯罪行为进行科学的分类管理与分析，通过引入了权值的概念，我们提出了基于权值参数的加权关联规则挖掘算法WARMA (Weight Association Rules Mining Algorithm)。

3.4.1 加权关联规则模型

设$I=\{i_1, i_2, \cdots i_m\}$是由不同项目组成的集合，为表征项目的重要性，给每一个项目i_j，赋以权值w_j，其中$0\leqslant w_j\leqslant 1$，$j\in\{1,2,\cdots m\}$[39]。

定义1 关联规则$X\Rightarrow Y$的加权支持度为：

$$w\_sup(X\Rightarrow Y)=\left(\sum_{i_j\in X\cup Y} w_j\right)(Sup(X\cup Y)) \tag{3.3}$$

定义2 如果k-项目集的加权支持度不低于用户给定的最低加权支持度阈值w_minsup，则称此k-项目集为频繁k-项目集。否则就称为非频繁项目集。

$$\left(\sum_{i_j\in X\cup Y} w_j\right)(Sup(X\cup Y))\geqslant w\_minsup \tag{3.4}$$

由定义可以看出，在加权关联规则的挖掘中关键是如何在权值和支持度之间找到一种平衡。如果在计算过程中将权值和支持度分开，那么结果只能找到那些同时满足最小支持度和最小权值的项目集。

3.4.2 权值参数的设定

权值是项目集重要程度的一种度量，权值参数设定的好坏直接影响到挖掘的效果。对权值参数，设定人除了基于自身素质和对项目的主观关注度外，还要考虑以下因素：

(1)如果某个项目集很重要，比如它正在进行促销或者它的利润很高，即使没有很多的顾客购买这些商品，对于经理来说它仍然可能是一个有意义的项目集；

(2)如果某个项目集并不太重要，但它却很流行，以至于在经常性的出现，那么它也可以被认为是有意义的项目集。

为了平衡权值和支持度，我们用权值之和与支持度的乘积来得到加权支持度。然而这样定义就出现了两个问题：一是根据定义可知，即使每一个项目的权值都很小，但当项目集中的项目个数越来越多时，整个项目集的权值也会变得很大。这个包含很多项的规则就是有意义的。二是传统的Apriori 算法已不再适用。

加权关联规则模型通过引入权值参数，用不同的权值来衡量各种犯罪行为的重要性，从而挖掘出重大犯罪行为，有效的解决了用传统Apriori算法进行公安犯罪行为分析时，这些重大犯罪行为由于出现频率小，将被忽略掉的问题，具有较好的实际应用性。

3.4.3 算法 WARMA 的基本思想

由于在加权关联规则挖掘方面，频繁项目集的含义发生了根本变化，频繁项目集的子集未必是频繁的，原有的Apriori算法已不再适用，需设计新的算法对加权关联规则进行挖掘。

设$T=\{t_1, t_2, \cdots, t_n\}$是一个交易数据库，$t_j$表示T的第j个交易或第j个记录，$t_j[i_k]$表示属性$i_k$在第j个记录上的值，$t_j[i_k]$取值为0或1[40]。

定义3 对于任一k-项目集X，如果X是频繁的，那么其支持数Count(X)应满足：

$$\mathrm{Count}(X) \geqslant \frac{w\_\min\sup\times|T|}{\sum_{i_j\in X} w_j} \tag{3.5}$$

设项目集Y为长度为q(q<k)，且$Y\subset X$，项目集(I－Y)中k-q个权重最大的项目为$i_{r1}, i_{r2}, \cdots, i_{r(k-q)}$，则对于任意项目$Z\supseteq Y$有：

$$\sum_{i_j\in Z} w_j \geqslant \sum_{i_j\in Y} w_j + \sum_{j=1}^{k-q} w_{rj} \tag{3.6}$$

令

$$W(Y, k) = \sum_{i_j\in Y} w_j \cdots + \sum_{j=1}^{k-q} w_{rj} \tag{3.7}$$

其中第一部分为q-项目集Y的权重之和，第二部分为(k-q)个最大权重之和。

定义4 对于任何一个包含项目集Y的k-项目集X而言。如果X为频繁k-项目集，则：

$$\mathrm{Count}(X) \geqslant \left\lceil \frac{w\_\min\sup\times|T|}{W(Y,k)} \right\rceil \tag{3.8}$$

由于X为频繁k-项目集，因此Count（X）$\geqslant \frac{w\_\min\sup\times|T|}{\sum_{i_j\in X} w_j}$。

因为$X\supseteq Y$，所以$\sum_{i_j\in X} w_j = \sum_{i_j\in Y} w_j + \sum_{i_j\in(X-Y)} w_j \leqslant \sum_{i_j\in Y} w_j + \sum_{j=1}^{k-q} w_{rj} = W(Y,k)$，从而有式(3.8)。

令$B(Y, k)=\left\lceil \frac{w_\min\sup\times|T|}{W(Y,k)} \right\rceil$（向上取整），并称B(Y，k)为Y的k-支持期望。

定义5 对于任何包含Y的k-项目集X而言，要想成为频繁项目集，则数据库中支持其的记录数Count(X)都必须不小于B(Y，k)。

3.4.4 算法 WARMA 描述

输入：事务数据库D；最小加权支持度阈值w_minsup；最小加权置信度阈值w_minconf；权值w_j，其中$0 \leqslant w_j \leqslant 1$，$j \in \{1,2,\cdots m\}$。

输出：加权关联规则。

(1) size=scan(D);
(2) L=∅; /L用来存放频繁项目集/
(3) for(i=1;i<=size;i++) do begin
(4) C_i=∅; /C_i为候选频繁项目集/
(5) L_i=∅; /L_i用来存放频繁k-项目集/
(6) end;
(7) for each transaction t do
(8) C_1=Count(D,W);
(9) for(k=2;k<=size;k++) do begin
(10) C_k=Apriori-Gen(C_{k-1});
(11) (C_k, L_k)=Check(C_k, D);
(12) L=L∪L_k
(13) end;
(14) 加权关联规则集＝Rules-Gen(L)

算法说明如下：

(1)scan(D)：该子程序的参数为事务数据库D，其功能是发现事务数据库D中的频繁项目集的最大可能长度，并返回该数值。

(2)Count(D,W)：该子程序累计1-项目集的支持数，计算每个1-项目集的k-支持期望，然后收集其支持数不低于k-支持期望的1-项目集，形成C_1。

(3)Check(C_k, D)：该子程序检查遍历交易数据库D，更新C_k中所有候选项目集的支持数。通过类似修剪步骤的方法，删除那些不满足所有可能频繁项目集支持期望的候选项目集，剩余的候选项目集均保存在C_k中。然后，再检查各项目集的加权支持，从中挑选出频繁k-项目集L_k。

(4)Rules-Gen(L)：根据L中大的频繁项目集生成符合最小加权置信度阈值的关联规则，方法与Apriori算法是一样[41]。

3.4.5 算法 WARMA 实验及分析

3.4.5.1 算法WARMA实验数据

实验2：犯罪信息事务D及各项权值如表3.3所示，项集$I=\{I_1, I_2, I_3, I_4\}$={抢劫，吸毒，盗窃，杀人}。设w_minsup=40%(加权支持度阈值)。

表3.3 犯罪信息事务集D及各项权值分配

TID	项集	项目	犯罪行为	权值
T_1	I_1, I_2, I_3, I_4	I_1	抢劫	0.5
T_2	I_2	I_2	吸毒	0.1
T_3	I_1, I_2	I_3	盗窃	0.3
T_4	I_3	I_4	杀人	0.9
T_5	I_1, I_2, I_3, I_4			
T_6	I_1, I_2			
T_7	I_2			
T_8	I_1			
T_9	I_1, I_2, I_3, I_4			
T_{10}	I_2			

用传统Apriori算法可以发现$\{I_1 \Rightarrow I_2\}$，即{抢劫⇒吸毒}，尽管在公安工作中有一定的实际应用性，但在实际中更关注的是与I_4项的关联规则。

下面用加权关联规则挖掘算法进行挖掘规则。

(1)scan(D)。取最大长度作为频繁项目集的最大可能长度。即：size=4。

(2)Count(D,W)。令$C_1=\{I_1, I_2, I_3, I_4\}$，扫描事务数据库D一次，分别求得$C_1$中各项目集的k-支持期望，如表3.4所示。则$C_1=\{I_1, I_2, I_3, I_4\}$，$L_1=\varnothing$。

表3.4 各1-项目集的k-支持期望

项目集	支持数	加权支持度	k		
			2	3	4
$\{I_1\}$	6	0.3	$B(\{I_1\},2)=3$	$B(\{I_1\},3)=3$	$B(\{I_1\},4)=3$
$\{I_2\}$	8	0.08	$B(\{I_2\},2)=4$	$B(\{I_2\},3)=3$	$B(\{I_2\},4)=3$
$\{I_3\}$	4	0.12	$B(\{I_3\},2)=4$	$B(\{I_3\},3)=3$	$B(\{I_3\},4)=3$
$\{I_4\}$	3	0.27	$B(\{I_4\},2)=3$	$B(\{I_4\},3)=3$	$B(\{I_4\},4)=3$

表3.5 各2-项目集的k-支持期望

项目集	支持数	加权支持度	k	
			3	4
$\{I_1, I_2\}$	5	0.3	$B(\{I_1, I_2\},3)=3$	$B(\{I_1, I_2\},4)=3$
$\{I_1, I_3\}$	3	0.24	$B(\{I_1, I_3\},3)=3$	$B(\{I_1, I_3\},4)=3$
$\{I_1, I_4\}$	3	0.42	$B(\{I_1, I_4\},3)=3$	$B(\{I_1, I_4\},4)=3$
$\{I_2, I_3\}$	3	0.12	$B(\{I_2, I_3\},3)=4$	$B(\{I_2, I_3\},4)=3$
$\{I_2, I_4\}$	3	0.3	$B(\{I_2, I_4\},3)=4$	$B(\{I_2, I_4\},4)=3$
$\{I_3, I_4\}$	3	0.36	$B(\{I_3, I_4\},3)=3$	$B(\{I_3, I_4\},4)=3$

(3)Apriori-Gen(C_1)。生成候选频繁2-项目集$C_2=\{\{I_1, I_2\}, \{I_1, I_3\}, \{I_1, I_4\},$

$\{I_2, I_3\}, \{I_2, I_4\}, \{I_3, I_4\}\}$，如表3.5所示。则$C_2 = \{\{I_1, I_2\}, \{I_1, I_3\}, \{I_1, I_4\}, \{I_2, I_3\}, \{I_2, I_4\}, \{I_3, I_4\}\}, L_2=\{\{I_1, I_4\}\}$。

(4)同理，Apriori-Gen(C_2)，Apriori-Gen(C_3)。通过计算可得：$L_3=\{\{I_1, I_2, I_4\}, \{I_1, I_3, I_4\}\}$，$L_4=\{\{I_1, I_2, I_3, I_4\}\}$。

(5)Rules-Gen(L)。$L=\bigcup L_k$中生成关联规则如下（置信度为60%）：

$I_1 \wedge I_2 \Rightarrow I_4$	conf=60%;	$I_1 \wedge I_4 \Rightarrow I_2$	conf=100%;
$I_2 \wedge I_4 \Rightarrow I_1$	conf=100%;	$I_4 \Rightarrow I_1 \wedge I_2$	conf=100%;
$I_1 \wedge I_3 \Rightarrow I_4$	conf=100%;	$I_1 \wedge I_4 \Rightarrow I_3$	conf=100%;
$I_3 \wedge I_4 \Rightarrow I_1$	conf=100%;	$I_3 \Rightarrow I_1 \wedge I_4$	conf=75%;
$I_4 \Rightarrow I_1 \wedge I_3$	conf=100%;	$I_3 \Rightarrow I_1 \wedge I_2 \wedge I_4$	conf=75%;
$I_4 \Rightarrow I_1 \wedge I_2 \wedge I_3$	conf=100%;	$I_1 \wedge I_2 \Rightarrow I_3 \wedge I_4$	conf=60%;
$I_1 \wedge I_3 \Rightarrow I_2 \wedge I_4$	conf=100%;	$I_1 \wedge I_4 \Rightarrow I_2 \wedge I_3$	conf=100%;
$I_2 \wedge I_3 \Rightarrow I_1 \wedge I_4$	conf=100%;	$I_2 \wedge I_4 \Rightarrow I_1 \wedge I_3$	conf=100%;
$I_3 \wedge I_4 \Rightarrow I_1 \wedge I_2$	conf=100%;	$I_1 \wedge I_2 \wedge I_3 \Rightarrow I_4$	conf=100%;
$I_2 \wedge I_3 \wedge I_4 \Rightarrow I_1$	conf=100%;	$I_3 \wedge I_4 \wedge I_1 \Rightarrow I_2$	conf=100%;
$I_1 \wedge I_2 \wedge I_4 \Rightarrow I_3$	conf=100%;		

加权关联规则挖掘算法的主要思想也是从小到大产生频繁项目集。但由于频繁项目集的子集不一定是频繁项目集，因而不能象Apriori_gen那样简单地从频繁(k-1)-项目集中生成候选k-项目集。本算法通过提出k-支持期望来排除那些不可能成为加权频繁k-项集的候选集的子集，而且使得那些不是加权频繁项集，但是可能是其他加权频繁项集的子集的候选项得以保留，保证了算法的正确性，解决了加权关联规则挖掘中加权频繁项集的子集可以不是加权频繁项集的问题。

3.4.5.2 算法WARMA性能测试及分析

为进一步验证的性能，在运行WinXP的P4-2.0G，内存512M、硬盘80G的台式电脑上进行了性能测试，软件开发环境为VFP 6.0。

表3.6 WARMA实验数据

事务数	WARMA算法执行时间(s)	FP-growth算法执行时间(s)	Apriori算法执行时间(s)
825	0.05	0.05	0.32
8312	0.22	0.28	2.47
83628	1.02	1.92	15.78
422305	4.68	8.18	75.33
825025	11.35	17.20	140.46

测试从两个方面进行了性能对比：一是算法WARMA和Apriori、FP-growth在不

同数据量下的执行时间，实验数据与实验结果如表3.6所示；二是算法WARMA和Apriori、FP-growth在事务数据库包含100个项目，50次交易，相同支持度阈值下产生频繁项集所使用的时间。

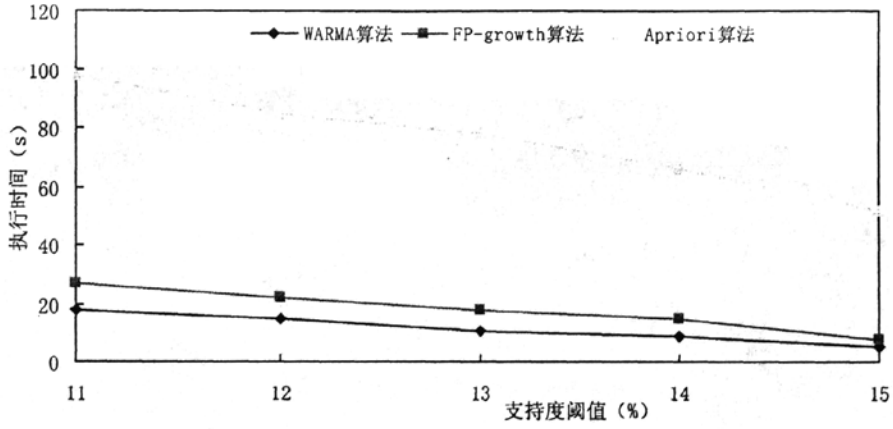

图3.1 不同支持度阈值下生成频繁项集的执行时间

从表3.6的实验结果来看，WARMA算法的速度是Apriori算法的10倍左右，比FP-growth算法稍快。从图3.1可以看出，在相同支持度阈值条件下，WARMA算法的执行时间小于FP-growth算法，远大于Apriori算法。这是因为在加权关联规则算法中为每个项目集设置一个计数器（即k值），只计算项目集的k-支持期望，并将与此项目集的支持数上限比较即可，可以防止多余候选项目集的生成。

3.5 基于 FUP 的改进关联规则增量式更新算法

随着社会的进步、时代的发展，一些原始的、传统的犯罪行为逐渐减少甚至消亡，而一些带有时代特征的新型的犯罪行为(如洗钱、传销、网络诈骗、网络色情、网络赌博等高科技、高智能型犯罪等)不断的出现，所以犯罪行为数据库中的项目就会不断地被更新。利用关联规则挖掘算法对犯罪行为进行分析，必须要能及时、敏锐的掌握与关注那些不断出现的新型犯罪行为，也就是说要能及时发现最新出现的频繁犯罪行为集，并能产生最新的关联规则，并及时将这些最新的关联规则用于指导公安日常警务工作，以便能及时采取措施，防范于未然，将预防与控制点前移，增强工作的主动性，提高执法效率与快速反应能力，加强犯罪控制，建立防范战略与侦查谋略，更有效的、更及时的预防与打击犯罪行为。

增量式更新算法能充分利用已挖掘出的知识来提高挖掘效率，是数据挖掘高效算法研究中一个主要方向。在众多的增量式更新算法中，D.W.Cheung等提出的FUP算法[42]是最典型、最有效和最实用的算法之一，但FUP算法存在对新增项目不敏感的缺点和不足，本节提出一种基于敏感参数的改进关联规则增量式更新算法。

3.5.1 算法 FUP 的分析

由于交易数据库中的数据是不断增加的，这必然会引起关联规则发生变化，如果采用Apriori算法对更新后的数据库进行挖掘，由于数据库规模会越来越大，这样Apriori算法的挖掘效率会变得越来越低。增量式更新算法就是针对这个问题提出的[43]。设原有交易数据库中的数据集记为D(称旧数据集)，新增加数据集记为d(称新数据集)，则当前事务数据库为(D+d)。假设已经采用Apriori算法获得数据集D的频繁项目集是L(D)，则FUP算法的基本思想是：

(1)利用Apriori算法生成新事务数据集d的频繁项目集L(d)，比较L(d)和L(D)，找出其中相同部分，将相同部分放入当前事务数据库(D+d)的频繁项目集中。

(2)对于t∈(L(D)-L(d))的频繁项集，如果t∈L(D)且t∉L(d)，扫描d得到t在d中的支持度$support_d$，根据D中已经求得的$support_D$，求出t在(D+d)中的支持度

$$support_{(D+d)}=\frac{sup\,port_d|d|+sup\,port_D|D|}{|d|+|D|} \tag{3.9}$$

如果$support_{(D+d)}\geq min_sup$，则把t放入当前事务数据库(D+d)的频繁项目集中，否则t不是频繁项目集。

(3)同上一步类似，对于t∈(L(d)-L(D))的频繁项目集，如果t∈L(d)且t∉L(D)，则扫描D得到t在D中的支持度$support_D$，再根据d中已经求得的支持度$support_d$，求出t在(D+d)中的支持度$support_{(D+d)}$。

如果$support_{(D+d)}\geq min_sup$，则把t放入当前事务数据库(D+d)的频繁项目集中，否则t不是频繁项目集[44]。

实验3：设min_sup=0.4，D为旧数据集，d为新增数据集。

表3.7 数据集D、d及其频繁项目集L(D)、L(d)

TID (D)	项	项 L(D)	支持度 L(D)	TID (d)	项	项 L(d)	支持度 L(d)
1	I_1, I_4	I_1	6/10	1	I_1, I_2, I_5	I_1	4/6
2	I_1, I_2	I_2	7/10	2	I_1, I_2, I_3	I_2	4/6
3	I_2, I_3	I_3	5/10	3	I_1, I_4, I_5	I_5	4/6
4	I_1, I_2	I_2, I_3	5/10	4	I_2, I_3, I_5	I_2, I_5	3/6
5	I_1, I_2, I_3			5	I_1		
6	I_2, I_3			6	I_2, I_5		
7	I_1, I_4						
8	I_1						
9	$I_2, I_3,$						
10	I_2, I_3, I_6						

(1) 通过Apriori算法从数据集D和d中分别获得频繁项目集L(D)和L(d)。

(2)对于L(D)和L(d)中相同的频繁项目集,直接加入到数据集(D+d)的频繁项目集L(D+d)中；对于D的频繁项目集，而非d的频繁项目集，则要扫描新数据集d，如果在(D+d)中还是频繁项目集，则这些频繁项目集加入到L(D+d)，否则放弃D中的这些频繁项目集；同样对于d中的频繁项目集，类似要再次扫描D，才能确定是否要加入到L(D+d)中。

表3.8 生成频繁项目L(D+d)的过程

L(D)∩L(d)		L(D)-L(d)		L(d)-L(D)		L(D+d)	
项	支持度	项	支持度	项	支持度	项	支持度
I_1	10/16	I_3	7/16	—	—	I_1	10/16
I_2	11/16	I_2, I_3	7/16			I_2	11/16
						I_3	7/16
						I_2, I_3	7/16

FUP算法还存在缺点和不足:

1.算法必须耗费大量时间处理规模巨大的候选项目集;

2.算法必须多次重复扫描数据库，对候选项目集进行模式匹配，代价很大;

3.算法对新增项目不敏感。

3.5.2 基于敏感参数的改进算法 SFUP

一类犯罪问题总是有先期动态的，通过对新型犯罪行为的挖掘分析，可以反映出整体犯罪的发展趋势，有利于在犯罪速度变化的决战中，抢占先机，加快犯罪控制应对措施变革，有利于在一种新型犯罪刚起，或在其传播过程中抢占制高点，有针对性地作防控决策。如犯罪速度变化促成的建立网上禁毒监控机制的决策就是超前的。另一方面，有利于建立犯罪综合治理的长效机制，采取综合性有效措施加大犯罪成本，降低司法成本，用最小的投入，获取最大的产出是遏制犯罪的有效决策思想。通过分析、追踪犯罪速度快慢变化轨迹，提高犯罪预测能力，建立预警机制，是争取主动权的有效途径。

1. 传统算法的敏感性分析。

传统算法在计算各个项目集的支持度都是以整个犯罪记录数据库的总记录数为基数计算的，即对于新项目的支持度计算基数为(|D|+|d|)，由于|D|>>|d|，所以新项目出现的次数/(|D|+|d|)<<新项目出现的次数/|d|,因此，极可能出现新项目在新数据集d中频繁出现，且有可能产生新的关联规则，但在整个数据库(|D|+|d|)中却非频繁的情况。如果采用传统的算法，则频繁发生的包含新项目的项目集经常会作为非频繁项目集，这样就发现不了新颖的、潜在有用的模式。

传统算法无法有效解决对新项目的敏感性，无法及时敏锐的发现那些新型犯罪行为。这样将可能会导致客观上打击不力，不足以形成威慑、阻吓犯罪的氛围和存在治安防控网络疏漏，不足以控制犯罪等问题，造成其在更大的时间、空间的扩散，给国家和人民群众生命财产安全造成更大的隐患。对传统算法进行进一步的优化和改进，加强对新犯罪行为的敏感性，从而进一步提高犯罪因素趋势预测和犯罪走势判断的准确性、客观性。

2．改进算法的研究对象。

设原有犯罪记录数据库中的数据记录集记为D，新增加的数据集记为d，则整个犯罪记录数据库为(D+d)。为了衡量关联规则挖掘算法对新增项目的敏感性，首先来分析新增项目的概念。如果新增加的项目在新的数据集d中的1-项目集仍然为非频繁项目集，那么新增加的项目对发现频繁项目集的结果没有影响，所以只有那些新增加的支持度大于或等于最小支持度的项目是我们要讨论的对象。

定义6 设交易数据库中在旧的数据集D的基础上，增加了新数据集d，若d中的一个项目newtiem满足：

$$support_d(\{newitem\}) - support_D(\{newitem\}) \geq min\_sup \qquad (3.10)$$

则称项目newitem为新项目。其中$support_D(\{newitem\})$为1-项目集{newitem}在旧数据集D中的支持度，$support_d(\{newitem\})$为1-项目集{newitem}在新数据集d中的支持度。

定义7 利用关联规则挖掘算法所得到的频繁项目集中，如果$\{newitem\} \in L_1(D+d)$，那么我们就称该关联规则挖掘算法对新项目newitem是敏感的，也就是说该关联规则挖掘算法对新项目newetim具有敏感度，否则不具有敏感度。其中，$L_1(D+d)$是整个数据库的频繁1-项目集[45]。

3．关联规则增量式更新的几种情况。

根据实际应用需求，关联规则的更新问题可以分为以下几种情况：

(1)事务数据库不变，最小支持度发生变化时，关联规则的高效更新问题；

(2)最小支持度不变，一个事务数据集d_1添加到事务数据库D中去时，如何生成最新事务数据库$(D \cup d_1)$中的关联规则的问题；

(3)最小支持度不变，从事务数据库D中删除一个事务数据集$d_2(d_2 \in D)$后，如何高效地生成事务数据库$(D-d_2)$中的关联规则的问题。

这三种情况是关联规则更新问题的基础和核心。在本节中，结合第二种情况提出了一个基于敏感参数改进的关联规则增量式更新算法，该算法主要是改进频繁项目集的发现过程，它根据以前发现的频繁项目集和新增的数据动态地更新原来发现的频繁项目集。数据库总是按时间先后顺序不断地积累，在新的算法中把数据按时间先后顺序进行划分，新的关联规则的获得是依靠在新的时间段内增加的数据集及在这之前发现的频繁项目集，而不需考虑在这之前的所有数据集，增

强了对新项目的敏感性。从空间上来说，可以不需存储以前的数据集；从时间上来说，不需扫描数据库中旧的数据集，所以发现频繁项目集的时间也大大节省了。这样增强了对新项目的敏感性，提高了数据挖掘的效率，并且具有很好的合理性。

因此，基于敏感参数的改进关联规则增量式更新算法——SFUP，加强了对新型犯罪行为的敏感性，无疑将会对公安警务工作起到很好的指导作用，将能更及时、更有效的预防和打击这些新型犯罪行为。

3.5.3 SFUP 算法基本思想

引入敏感参数 α (1≤ α ≤∞)，在D中发现频繁项集的过程中，保留那些支持度≥(最小支持度/ α)的频繁项集，当数据库增加新的数据集时，只考虑以前产生的支持度≥(最小支持度/ α)的频繁项集和当前新增加的数据集，由于数据库的规模在不断地增大，而项目的增加却相对稳定，扫描支持度≥(最小支持度/ α)的频繁项集的时间比扫描整个旧数据集的时间要短得多，从而大大提高了关联规则挖掘的效率。

(1)用Apriori算法获得数据集D的支持度≥(最小支持度/ α)的频繁项集 $L_1(D)$；

(2)对于项集I，I∈ $L_1(D)$，根据新数据集d和$L_1(D)$，加入到(D+d)的支持度≥(最小支持度/ α)的频繁项集$L_1(D+d)$中；

(3)遍历新数据集d，用Apriori算法计算d中的支持度≥(最小支持度/ α)的频繁项集$L_1(d)$；

(4)对于项集I'，I' ∈ $L_1(d)$，且I' ∉ $L_1(D+d)$，则把I' 加入到$L_1(D+d)$；

(5)用Apriori算法在$L_1(D+d)$中找出支持度≥最小支持度的频繁项集，即L(D+d)。

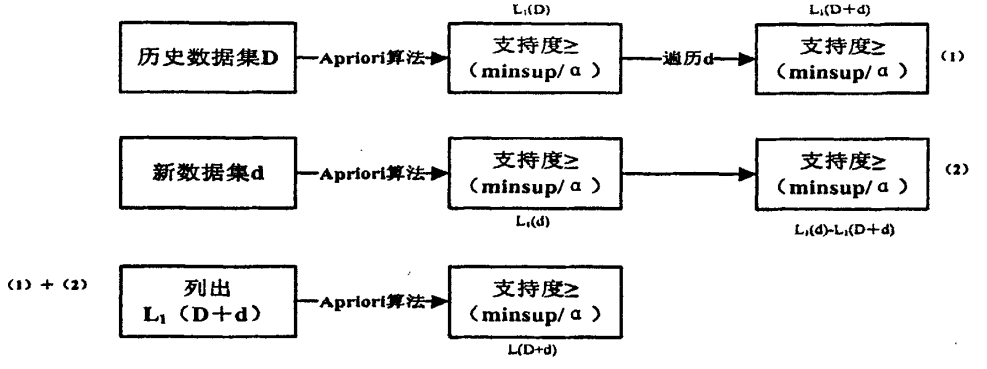

图3.2 算法SFUP基本思想图

3.5.4 敏感参数的含义

在改进算法中，α 的取值有不同的含义：α 的值越大，那么(最小支持度/ α)

就越小，旧数据集中得到的≥(最小支持度/α)的频繁项目集就保留得越多，也就是说旧数据的权值增大；如果α的值越小，那么(最小支持度/α)就越大，旧数据中得到的≥(最小支持度/α)的频繁项目集就保留得越少，也就是说旧数据的权重减小；当α=1时，意味着对增加新数据后的数据集发现频繁项目集时，完全不考虑旧数据集中支持度＜最小支持度的所有项目集。α的取值根据具体的情况选取。

3.5.5 算法 SFUP 描述

输入：参数α(1≤α≤∞)，旧数据集D和新数据集d，支持度阈值$\frac{\min sup}{\alpha}$

输出：(D+d)中的频繁项目集L(D+d)

(1) Apriori(D, $\frac{\min sup}{\alpha}$)；

(2) Apriori(D+d, $\frac{\min sup}{\alpha}$)；

(3) For each transactions $t_d \in L(D)$;

(4) if $t_d \in L(D+d)$

(5) $support_d(t) = support_{D+d}(t)$;

(6) $L(D+d) = \{t_d | support_d(t) \geq minsup\}$

3.5.6 算法 SFUP 实验及分析

3.5.6.1 算法SFUP实验数据

实验4：设最小支持度min_sup为0.4，α=2，则(min_sup/α)=0.2，如表3.9所示，D为旧数据集，d为新数据集。$\{I_1, I_2, I_3 I_4, I_5, I_6\}$分别对应违法犯罪行为{盗窃、赌博、伤人、爆炸、洗钱、间谍}。

(1)通过Apriori算法得到D中支持度≥min_sup/α的频繁项目集$L_1(D)$。项集I，$I \in L_1(D)$，对于数据集D和d，$Sup(I) = \frac{I.Count(D) + I.Count(D)}{|D| + |d|}$，把$Sup(I) \geq min_sup/\alpha$的项目集加入$L_1(D+d)$中。如表3.10所示。

(2)通过Apriori算法得到d中支持度≥min_sup/α的频繁项目集$L_1(d)$。对于项集I'，如果$I' \in L_1(d)$，且$I' \notin L_1(D+d)$，则生成$L_1(d) - L_1(D+d)$。如表3.11所示。

(3)把$L_1(d) - L_1(D+d)$中的项加入到$L_1(D+d)$中，得到更新后的$L_1(D+d)$。如表3.12所示。

(4)用Apriori算法在$L_1(D+d)$中找出支持度≥最小支持度的频繁项集L(D+d)。如表3.13所示。

表3.9 旧数据集D和新数据集d

TID(D)	项	TID(d)	项
1	I_1, I_4	1	I_1, I_2, I_5
2	I_1, I_2	2	I_1, I_2, I_3
3	I_2, I_3	3	I_1, I_4, I_5
4	I_1, I_2	4	I_2, I_3, I_5
5	I_1, I_2, I_3	5	I_1
6	I_2, I_3	6	I_2, I_5
7	I_1, I_4		
8	I_1		
9	$I_2, I_3,$		
10	I_2, I_3, I_6		

表3.10 频繁项目集$L_1(D)$和$L_1(D+d)$

$L_1(D)$项	支持度	$L_1(D+d)$项	支持度
I_1	6/10	I_1	10/16
I_2	7/10	I_2	11/16
I_3	5/10	I_3	7/16
I_4	2/10	I_1, I_2	5/16
I_1, I_2	3/10	I_2, I_3	7/16
I_1, I_4	2/10		
I_2, I_3	5/10		

表3.11 项集$L_1(d)$和$L_1(d)-L_1(D+d)$

$L_1(d)$项	支持度	$(L_1(d)-L_1(D+d))$项	支持度
I_1	4/6	I_5	4/6
I_2	4/6	I_1, I_5	2/6
I_3	2/6	I_2, I_5	3/6
I_5	4/6		
I_1, I_2	2/6		
I_1, I_5	2/6		
I_2, I_3	2/6		
I_2, I_5	3/6		

表3.12 更新后的项集$L_1(D+d)$

$L_1(D+d)$项	支持度
I_1	10/16
I_2	11/16
I_3	7/16
I_1, I_2	5/16
I_2, I_3	7/16
I_5	4/6
I_1, I_5	2/6
I_2, I_5	3/6

表3.13 生成的频繁项集$L(D+d)$

$L(D+d)$项	支持度
I_1	10/16
I_2	11/16
I_3	7/16
I_2, I_3	7/16
I_5	4/6
I_2, I_5	3/6

在基于敏感参数的改进算法中，对于增加新的数据集d后，如果数据集d中的项目集I在D中为非频繁项目集，则以Count(I)/|d|作为其支持度。以实验4中的新旧数据集D和d为例，用FUP算法和SFUP算法生成的频繁1-项集如表3.14所示。

表3.14 算法FUP和SFUP生成的频繁1-项集

增量式更新算法		改进算法	
$L_1(D+d)$项	支持度	$L_1(D+d)$项	支持度
I_1	10/16	I_1	10/16
I_2	11/16	I_2	11/16
I_3	7/16	I_3	7/16
		I_5	4/6

$support_d(\{I_5\})- support_D(\{I_5\})=4/6\geqslant 0.4=min\_sup$，因此，$\{I_5\}$是增加新数据集d后出现的新项目，由表3.14可以看出，在增量式更新算法中，项目集$\{I_5\}$是非频繁，在基于敏感参数的改进算法中项目集$\{I_5\}$是频繁的。因此，增量式更新算法对项目集$\{I_5\}$不敏感，而改进算法对项目集$\{I_5\}$是敏感的。

3.5.6.2 算法SFUP性能测试

为了进一步验证改进算法的性能，在运行WinXP的P4 2.0G，内存512M、硬盘80G进行了算法测试。软件开发环境为VFP6.0。

测试从两个方面进行了性能对比：

一是算法SFUP和Apriori、FUP在不同数据量下的执行时间；

二是算法SFUP和Apriori、FUP在不同数据库和不同支持度下执行时间的比较，测试事物数据库D包含事务记录数67557，数据库d包含事务记录数8124，总项数均为120，D中项最大长度为23，d中项最大长度为43。设最小支持度为40%，置信度为60%，其中SFUP的敏感参数 α 为2。

表3.15 SFUP实验数据

事务数	SFUP算法 执行时间(s)	FUP算法 执行时间(s)	Apriori算法 执行时间(s)
2110	2.12	3.56	10.38
8124	18.22	25.68	90.56
16889	31.46	55.75	250.92
33778	58.37	75.69	405.81
67557	99.26	138.85	780.38

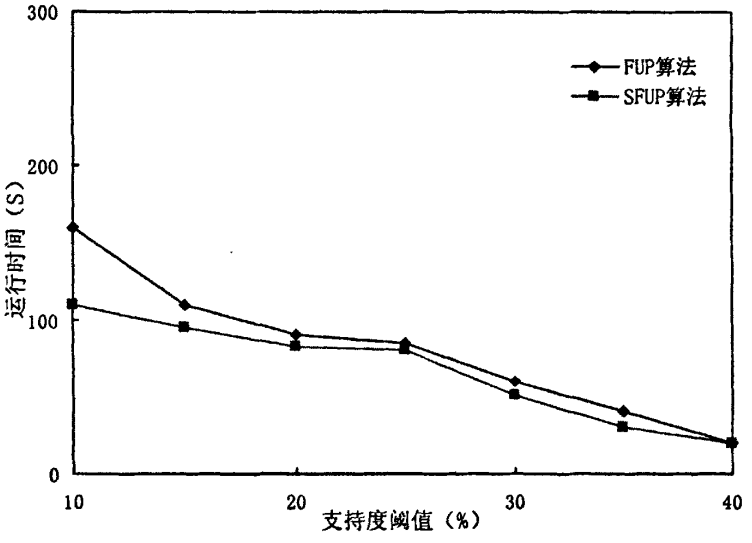

图3.3 算法FUP和SFUP在数据库D上的运行时间比较

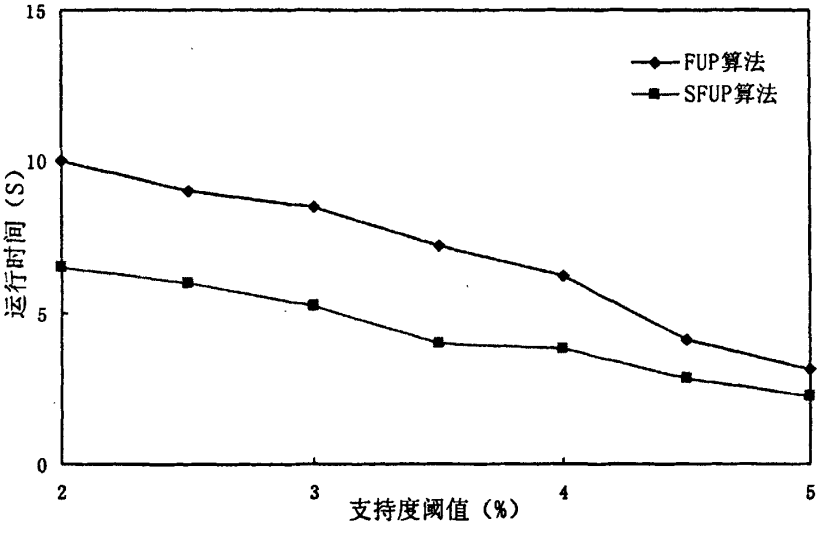

图3.4 算法FUP和SFUP在数据库d上的运行时间比较

从表3.15的实验结果可以看出，SFUP算法在越大数据量下优势越明显。图3.3和图3.4所示的是不同数据库和在不同的支持度下改进算法与FUP算法挖掘时间的比较。当最小支持度阈值减小时，产生的频繁项目集数量增多。当支持度阈值减小时改进算法的运行时间曲线增长平缓的多，说明其计算性能相对稳定。

SFUP算法主要是改进了产生频繁项目集的过程，它根据以前发现的频繁项目集和事务数据库中新增的数据动态的更新原来发现的频繁项集。这样从空间上来说，可以不需要存储以前的数据集，从而节省了消耗的内存；从时间上来说，因为只考虑原来得到的频繁项目集，而不需扫描交易数据库总的数据集，所以发现频繁项目集的时间大大节省了，从而提高了整个关联规则挖掘的效率。但是对于怎样的敏感参数α才能得到最佳的时间复杂性和空间复杂性还有待进一步研究。

3.6 本章小结

本章概述了关联规则的基本概念、种类，Apriori算法思想及其性能瓶颈问题。

本章针对传统Apriori算法在公安犯罪行为分析中忽略了不同项间不同重要性的问题，提出了一种改进的加权关联规则挖掘算法WARMA，改进算法设计了加权关联规则模型，通过提出的k-支持期望概念来排除不可能成为加权频繁k-项集的候选集子集，解决了加权关联规则挖掘中加权频繁项集的子集可以不是加权频繁项集的问题。实验及分析表明：算法WARMA在不同数据量下的执行时间远小于Apriori，在相同加权支持度阈值条件下，加权关联规则算法产生的频繁项集和执行时间均小于Apriori算法，证实WARMA算法能更有效的发现重大犯罪行为。

本章针对传统Apriori算法在公安犯罪行为分析中无法及时敏锐的发现某些新型犯罪行为的问题，提出了基于FUP的改进关联规则增量式更新算法SFUP，SFUP算法通过用敏感参数衡量对新项目的重视情况，从敏感性和时间效率出发，改进了一般增量式更新算法产生频繁项集的过程。实验及分析表明：改进算法SFUP在越大数据量下优势越明显，能较好地发现新增数据中的新模式，具有较高的敏感性，能更有效的发现新型犯罪行为，同时在挖掘过程中显示了良好的空间和时间性能，大大提高了整个挖掘效率。

第 4 章　基于决策树的违法犯罪行为分析

4.1 决策树分类算法

近年来，涌现出了大量适合于不同应用的分类算法。如基于归纳学习的决策树分类、基于向量空间模型的 k-近邻分类、神经网络、基于统计学习理论的支持向量机(Support Vector Machine，SUM)分类和源于概率统计的贝叶斯分类等。

决策树是以实例为基础的归纳学习算法，是一种用于预测模型的算法，它从一组无次序、无规则的元组中推理出决策树表示形式的分类规则。它采用自顶向下的递归方式，利用树形结构来表示决策集合，这些决策集合通过对数据集的分类产生规则。通常包括树的生成和树的剪枝。树的一个内结点表示对一个属性的测试，其分支表示测试的结果，叶结点代表一个类别，在建树的过程中，使用剪枝来剪去数据中的噪声和孤立点，从而提高在未知数据上分类的准确性。从根节点到叶节点的一条路径就对应着一条合取规则，整个决策树就对应着一组析取表达式规则。

当前国际上有影响的示例学习方法是 Quinlan 于 1986 年提出的 ID3 算法。在 ID3 算法的基础上，1993 年 Quinlan 又提出了 C4.5 算法。为了适应处理大规模数据集的需要，后来研究人员又提出了若干改进的算法，其中 SLIQ 和 SPRINT 是比较有代表性的两个算法。

4.1.1 ID3 算法基本原理

ID3 算法的前身是 CLS(concept learning system)算法。其基本算法是贪心算法，它是基于信息熵的决策树分类算法，根据属性集的取值选择实例的类别。

ID3 算法的核心是:在决策树各级节点上选择属性时，用信息增益作为属性的选择标准，以使得在每一个非叶节点进行测试时，能获得关于被测试记录最大的类别信息[46]。

其具体方法是：检测所有的属性，选择信息增益最大的属性产生决策树节点，由该属性的不同取值建立分支，再对各分支的子集递归调用该方法建立决策树节点的分支，直到所有子集仅包含同一类别的数据为止，最后得到一棵决策树，它可以用来对新的样本进行分类。实际上，能正确分类训练数据集的决策树不止一棵，用 ID3 算法能得到节点数最少的决策树。

其基本原理是：设 $H=F_1\times F_2\times\cdots\times F_n$ 是 n 维有穷向量空间，其中 F_j 是有穷离散符号集，H 中的元素 $h=<v_1, v_2, \cdots v_n>$ 叫做例子，其 $v_j\in F_j$, $j=1, 2, \cdots, n$。设 PE 和 NE 是 E 的两个例子集，分别叫做正例集和反例集[38]。

假设向量空间 H 中的正例集 PE 和反例集 NE 的大小分别为 p 和 n，ID3 算法基

于下列两个假设：

(1)在向量空间 H 上的一棵正确决策树对任意例子的分类概率同 H 中正反例的概率一致；

(2)一棵决策树能对一例子作出正确类别判断所需要的信息量为：

$$I(p,n)=-\frac{p}{p+n}\log_2\frac{p}{p+n}-\frac{n}{p+n}\log_2\frac{n}{p+n} \tag{4.1}$$

如果以属性 A 作为决策树的根，A 具有 v 个值$\{v_1, v_2, \cdots v_n\}$，它将 H 分为 v 个子集$\{H_1, H_2, \cdots, H_v\}$，假设 H_i 中含有 p_i 个正例和 n_i 个反例，子集 H_i 的信息熵为 $I(p_i, n_i)$，以属性 A 为根分类后的信息熵为：

$$H(A)=\sum_{i=1}^{v}\frac{p_i-n_i}{p+n}\ I(p_i, n_i) \tag{4.2}$$

因此，以 A 为根的信息增益是：

$$gain(A)=I(p,n)-H(A) \tag{4.3}$$

ID3 选择使 gain(A)最大(即 H(A)最小)的属性 A^*作为根结点，对 A^*的不同取值对应的 H 的 v 个子集 H_i 递归调用上述过程，生成 A*的子结点 $B_1, B_2, \cdots, B_v$。

ID3 算法检验所有的特征，选择信息增益(互信息)最大的特征 A 产生决策树节点，由该特征的不同取值建立分枝，对各分枝的实例子集递归，用该方法建立决策树节点和分枝，直到某一子集中的例子属于同一类。

ID3 算法利用互信息量最大的特征建立决策树，使决策树节点数最小，识别例子准确率高。

4.1.2 ID3 算法基本概念

设 S 是 s 个数据样本的集合。假定类标号属性具有 m 个不同值，定义 m 个不同类 C_i (i=1，…，m)。设 s_i 是类 C_i 中的样本数。对一个给定的样本分类所需的期望信息由下式给出[47]：

$$I(s_1, s_2, \cdots, s_m)=-\sum_{i=1}^{m}p_i\log_2(p_i) \tag{4.4}$$

其中 $p_i=s_i/s$，是任意样本属于 C_i 的概率。注意，对数函数以 2 为底，其原因是信息用二进位编码。

设属性 A 具有 v 个不同值$\{a_1, a_2, \cdots a_v\}$。可以用属性 A 将 S 划分为 v 个子集$\{S_1, S_2, \cdots, S_v\}$；其中，$S_j$ 中的样本在属性 A 上具有相同的值 a_j(j=1, 2…，v)。设 S_{ij} 是子集 S_j 中类 C_i 的样本数。由 A 划分成子集的熵或信息期望由下式给出：

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+s_{2j}+\ldots+s_{mj}}{s}I(s_{1j}, s_{2j}, \ldots, s_{mj}) \tag{4.5}$$

熵值越小，子集划分的纯度越高。对于给定的子集 S_j，其信息期望为：

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \tag{4.6}$$

其中 $P_{ij}=s_{ij}/s_j$，是 S_j 中样本属于 C_i 的概率。

在属性 A 上分枝将获得的信息增益是：

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \tag{4.7}$$

决策树算法计算每个属性的信息增益，并从中挑选出信息增益最大的属性作为给定集合 S 的测试属性并由此产生相应的分支节点。所产生的节点被标记为相应属性，并根据这一属性的不同取值分别产生决策树的分支，每一分支代表一个被划分的样本子集，并据此划分样本。

4.1.3 ID3 算法优劣分析

(1)在 ID3 算法的假设空间包含所有的决策树，它是关于现有属性的有限离散值函数的一个完整空间，避免了假设空间可能不包含目标函数的风险。

(2)当遍历决策树空间时，ID3 仅维护单一的当前假设，失去了表示所有一致假设所带来的优势。

(3)ID3 算法在搜索过程中不进行回溯，每当在树的某一层次选择了一个属性进行测试，它不会再回溯重新考虑这个选择，易受无回溯的爬山搜索中的常见风险影响，收敛到局部最优的答案，而不是全局最优的。

(4)ID3 算法在搜索的每一步都使用当前的所有训练样例，以统计为基础决定怎样精化当前的假设，大大降低了对个别训练样例错误的敏感性。

(5)ID3 算法只能处理离散值的属性。

(6)信息增益度量存在一个内在偏置，它偏袒具有较多值的属性。

(7)ID3 算法增长树的每一个分支的深度，直到恰好能对训练样例完美地分类。

4.1.4 其他分类算法简述

1. C4.5 算法。

C4.5 算法分两个阶段。第一阶段：根据信息增益最大的标准选择某个属性对训练集进行划分，递归调用直到每个划分中的所有样本属于同一类；第二阶段：对建立的树进行剪枝，即剪去建立在噪声数据上的分支。C4.5 算法对 ID3 算法进行了改进：用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；在树构造过程中进行剪枝；能够完成对连续属性的离散化处理；能够对不完整数据进行处理。

C4.5 算法与其它分类算法比较有产生的分类规则易于理解，准确率较高的优点。但在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法效率低。C4.5 适合于能够驻留于内存的数据集，当训练集大得无法在内存容纳时，程序无法运行[48]。

2. SLIQ 算法(supervised learning in quest)

SLIQ 算法在决策树的构造过程中采用了“预排序”和“宽度优先策略”两种技术。所谓预排序，就是针对每个属性的取值，把所有的记录按照从小到大的顺序进行排序，以消除在决策树的每个节点对数据集进行的排序。所谓宽度优先策略，即在决策树的每一层只需对每个属性列表扫描一次，就可以为当前决策树中每个叶子节点找到最优分裂标准。

SLIQ 算法能够处理比 C4.5 大得多的训练集，在一定范围内具有良好的随记录个数和属性个数增长的可伸缩性。但仍然存在一些缺点：在一定程度上限制了可以处理的数据集的大小；不可能达到随记录数目增长的线性可伸缩性。

3. SPRINT 算法(scalable parallelizable inductionof decision trees)

该算法改进了决策树算法的数据结构，去掉了在 SLIQ 中需要驻留于内存的类别列表，将它的类别列表合并到每个属性列表中。这样，在遍历每个属性列表寻找当前节点的最优分裂标准时，不必参照其他信息，将对节点的分裂表现在对属性列表的分裂，即将每个属性列表分成两个，分别存放属于各个节点的记录。

该算法的优点是在寻找每个节点的最优分裂标准时变得更简单。其缺点是对非分裂属性的属性列表进行分裂变得很困难。解决的办法是对分裂属性进行分裂时用哈希表记录下每个记录属于哪个个孩子节点。由于哈希表的大小与训练集的大小成正比，当训练集很大时，哈希表可能无法在内存容纳，此时分裂只能分批执行，这使得 SPRINT 算法的可伸缩性仍然不是很好[49]。

4.2 传统 ID3 算法在犯罪行为分析应用中存在的问题

公安犯罪行为分析是一门涉及面极广的交叉性的新型学科，渗透了法学、心理学、行为学等多门学科，需要相当的专门知识，其本身现在还处在不断探索研究的阶段。数据仓库和数据挖掘技术提出、形成时间不长，在国内得到具体应用的时间较短，实际应用领域较少，目前成功的例子不多，且基本局限在金融、保险、商业等领域，而在犯罪行为分析领域，更是鲜见有效应用。本文希望通过对违法犯罪人员的数据进行实例分析，以决策树理论为基础，初步建立一个犯罪预警模型。

4.2.1 ID3 算法的应用

下面选取了少量样本数据，如表 4.1 所示，仅包含了登记在案的违法犯罪人员的部分项目内容，由此应用决策树 ID3 算法进行挖掘知识。

1. 根据 ID3 算法，对根结点进行分类。

由表 4.1，样本集合的属性为“犯罪程度”，有两个取值{较轻，严重}，即 M=2，设 C1 对应“较轻”类别，C2 对应“严重”类别，则 C1=9，C2=6。为计算

每个属性的信息增益，由公式(4.4)可得分类所需的期望信息：

表 4.1 部分犯罪记录表

序号	年 龄	经济状况	文化程度	有无正当职业	社会关系有无犯罪记录	是否有特 长	是否常驻人口	犯 罪程 度
1	20-30	中等	初中	无	无	有	是	较轻
2	>40	差	小学	无	有	有	是	严重
3	30-40	差	初中	无	有	有	是	严重
4	20-30	中等	高中	有	无	无	是	较轻
5	>40	差	小学	有	有	无	否	严重
6	30-40	差	初中	有	有	无	是	较轻
7	20-30	中等	初中	无	有	有	否	较轻
8	20-30	差	高中	无	无	有	否	严重
9	30-40	中等	高中	有	无	无	是	较轻
10	20-30	中等	初中	有	有	无	是	较轻
11	20-30	差	高中	无	有	有	否	严重
12	>40	差	初中	无	无	无	是	较轻
13	20-30	差	高中	有	无	无	否	较轻
14	20-30	差	初中	有	有	无	否	严重
15	20-30	中等	高中	有	有	无	是	较轻

$$I(s_1, s_2)=I(9,\ 6)=-\frac{9}{15}\log_2\frac{9}{15}-\frac{6}{15}\log_2\frac{6}{15}=0.9709$$

由公式(4.5)、(4.6)计算各属性的熵：

A1＝“年龄”：20－30、30－40、>40

$$E(A1)=\frac{9}{15}\left(-\frac{6}{9}\log_2\frac{6}{9}-\frac{3}{9}\log_2\frac{3}{9}\right)+\frac{3}{15}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{3}{15}\left(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}\right)=0.9183$$

……

由公式(4.7)，则 A1 上分支的信息为：

Gain(A1)=I(9,6)-E(A1)=0.0526

使用同样的方法得出各属性的信息熵和分枝信息如表(4.2)所示。

依据算法，取 Gain(A)最大值为根节点，即现按“经济状况”属性分类。再继续用上述算法求解下一层各属性的熵：

表 4.2 基于根结点分类的各属性信息熵和分枝信息表

属性	年龄	经济状况	文化程度	有无职业	社会关系有无犯罪记录	是否有特长	是否常驻人口
E(A)	0.9183	0.5510	0.7701	0.8925	0.8546	0.8258	0.8258
Gain(A)	0.0526	0.4199	0.2008	0.0784	0.1162	0.1451	0.1451

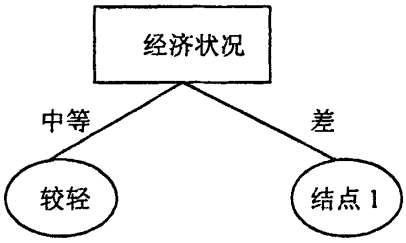

图 4.1 基于 ID3 算法的根结点分类决策树

2. 对图 4.1 中的叶结点 1 进行分类。

表 4.3 基于叶结点 1 的犯罪记录表

序号	年龄	经济状况	文化程度	有无正当职业	社会关系有无犯罪记录	是否有特长	是否常驻人口	犯罪程度
2	>40	差	小学	无	有	有	是	严重
3	30-40	差	初中	无	有	有	是	严重
5	>40	差	小学	有	有	无	否	严重
6	30-40	差	初中	有	有	无	是	较轻
8	20-30	差	高中	无	无	有	否	严重
11	20-30	差	高中	无	有	有	否	严重
12	>40	差	初中	无	无	无	是	较轻
13	20-30	差	高中	有	无	无	否	较轻
14	20-30	差	初中	有	有	无	否	严重

由表 4.3，样本集合的属性为“犯罪程度”，有两个取值{较轻，严重}，即 M=2，设 C_1 对应“较轻”类别，C_2 对应“严重”类别，则 $C_1=3$，$C_2=6$。为计算每个属性的信息增益，由公式(4.4)可得分类所需的期望信息：

$$I(s_1, s_2)=I(3, 6)=-\frac{3}{9}\log_2\frac{3}{9}-\frac{6}{9}\log_2\frac{6}{9}=0.9183$$

由公式(4.5)、(4.6)计算各属性的熵：

A1=“年龄”：20－30、30－40、>40

$$E(A1)=\frac{4}{9}(-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4})+\frac{2}{9}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})+\frac{3}{9}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})=0.8889$$

……

由公式(4.7),则 A1 上分枝的信息为:

Gain(A1)=I(3,6)-E(A1)=0.0294

使用同样的方法得出各属性的信息熵和分枝信息如表 4.4 所示。

表 4.4 基于叶结点 1 分类的各属性信息熵和分枝信息表

属性	年龄	文化程度	有无职业	社会关系有无犯罪记录	是否有特长	是否常驻人口
E(A)	0.8889	0.7505	0.851	0.7394	0.5394	0.851
Gain(A)	0.0294	0.1678	0.0673	0.1789	0.3789	0.0673

依据算法，取 Gain(A)最大值为根节点，即现按“是否有特长”属性分类。

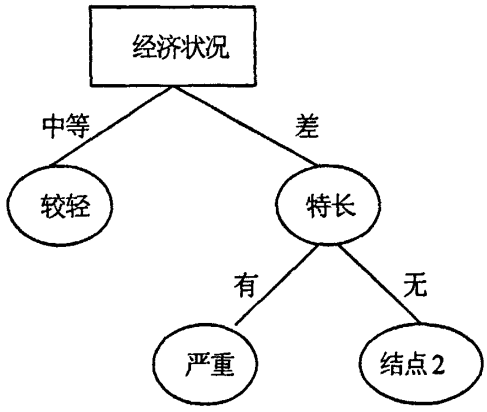

图 4.2 基于 ID3 算法的叶结点 1 分类图

以此类推，得到图 4.3 的 ID3 算法决策树。

3. 传统 ID3 算法的规则归纳。

根据生成的最终决策树从根结点到叶结点的路径及数据集所包含记录的多少，可以得出如下分类规则：当经济状况中等时，犯罪程度较轻；当经济状况差，有特长的，犯罪程度严重，而无特长且社会关系中无犯罪记录的，犯罪程度较轻，如社会关系中有犯罪记录且不是常驻人口的，犯罪程度严重，是常驻人口的，犯罪程度较轻。

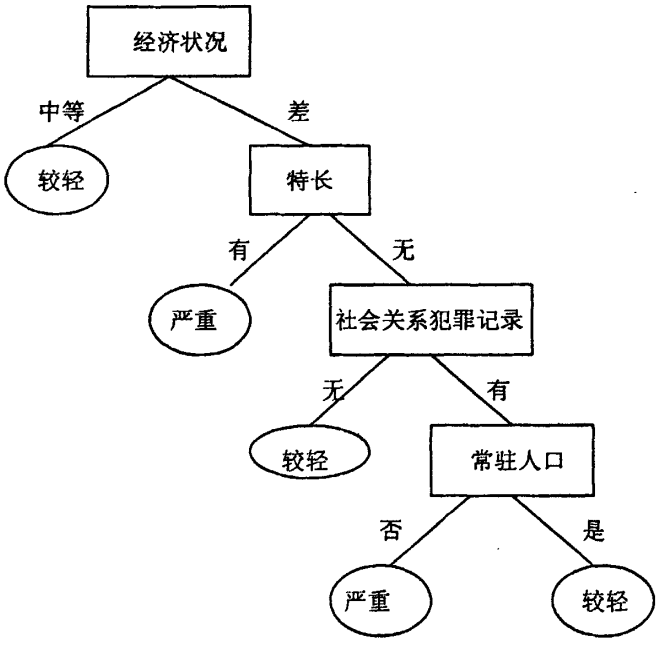

图 4.3 基于 ID3 算法生成的犯罪记录决策树

可以如下表示：

(1) IF(经济状况＝中等) THEN (犯罪程度＝较轻)；

(2) IF(经济状况=差 AND 特长＝有) OR (经济状况＝差 AND 特长＝无 AND 社会关系犯罪记录＝有 AND 常驻人口＝否) THEN (犯罪程度＝严重)；

(3) IF(经济状况=差 AND 特长＝无 AND 社会关系犯罪记录＝无) OR (经济状况＝差 AND 特长＝无 AND 社会关系犯罪记录＝有 AND 常驻人口＝是) THEN (犯罪程度＝较轻)；

由于在实验中的样本数量和项目还不够多，分析程度和分类知识的获取还不够理想，可信度也还相对不够，还只能称作一个简单的粗层次的分类分析模型[41]。

4.2.2 存在的问题

根据在实际工作中的犯罪情况统计分析，可以得出经济状况在犯罪程度分析中并不是一个很重要的因素，同时生成的规则与工作实际也存在一定的偏差，但依据实验确实得出了结论，这就表明该算法在公安工作的应用中存在了一定的偏差。由实验证明，传统的 ID3 算法在公安工作中存在以下缺点：

(1) ID3 选择属性 A 作为测试属性的原则是，A 使得 E(A)最小。研究表明，该算法往往偏向于选择取值较多的属性，而属性较多的属性却不总是最优的属性，即按照使熵值最小的原则被 ID3 算法列为应该首先判断的属性在现实情况中却并不那么重要。这种因素可以由专家、数据统计及生活常识来判断，即在分类后是

否支持先前假定的规则。

(2)当大多数属性数据量较大，个别属性数据量较小时，容易出现大数据掩盖小数据的现象。从而失去一些相应重要的判断。

4.3 一种基于先验参数的改进算法 BID3

4.3.1 基于先验参数的 BID3 算法描述

针对ID3算法偏向于选择取值较多但在实际公安工作中对分类意义并不大的属性作为测试属性，以及容易淹没小的重要数据等缺点，我们引入先验参数B对传统ID3算法进行改进，改进后的算法称为BID3算法。

BID3算法是针对规则生成方法即属性选择标准算法进行了改进。通过对添加先验参数B，加强了重要属性的标注，降低非重要属性的标注，使生成决策树时数量少的数据元组不会被淹没，最终使决策树减少了对取值较多的属性的依赖性，从而尽可能地减少大数据掩盖小数据的现象发生[50]。

基于先验参数B的BID3算法对一个给定的样本分类所需的期望信息公式为：

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{4.8}$$

其中 $p_i=s_i/s$，是任意样本属于 C_i 的概率。

由属性 A 划分成子集的熵或信息期望公式为：

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \ldots + s_{mj}}{s} I(s_{1j}, s_{2j}, \ldots, s_{mj}) \tag{4.9}$$

熵值越小，子集划分的纯度越高。对于给定的子集 S_j，其信息期望为：

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} (p_{ij} + B)\log_2(p_{ij}) \tag{4.10}$$

其中 $P_{ij}=s_{ij}/s_j$，是 S_j 中样本属于 C_i 的概率。

在属性 A 上分枝将获得的信息增益是：

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \tag{4.11}$$

在公式(4.10)中，先验参数B的取值在[0，1]之间，其大小由决策者根据先验知识或领域知识测试后给出。

改进的BID3算法就是把公式(4.10)作为测试属性的选择标准来构造决策树，当公式(4.10)中B＝0时，则回归传统的ID3算法。实际应用中可以首先用ID3算法构造决策树，如果结果中出现了取值少的重要属性比取值多的非重要属性离根结点的距离远的情况，则可用改进后的BD3算法重新构造决策树进行规则提取。

与某些ID3的改进算法相似，BID3算法只是在传统ID3算法的基础上做了一点微调而形成的，如果其参数取值为0则又回归为传统的ID3算法，但BID3算法在解

决领域问题中大数据量掩盖小数据量的重要性方面具有一定的优势。但以BID3算法生成的树基本上都会比用ID3算法所生成的树更大一些，增加了计算量和时间复杂性，同时由决策者根据先验知识确定先验参数的大小也增加了算法的随机性。

4.3.2 先验参数的确定

给定0≤B≤1，其大小由公安工作中的决策者根据公安领域先验知识来确定。它是一个模糊的概念，通常指关于某一事务的先验知识，包括领域知识和专家建议，具体到决策树学习中，则是指在决策树训练过程中除了用于生成和修改决策树的实验集之外的所有影响决策树规则生成和选择的因素，例如实例和规则的表示及转换语言、规则生成和修改所采用的方法、数据冗余及噪音处理等。如果实例集空间很大，则利用先验参数B可以减小搜索空间提高学习效率。

在使用先验参数B的时候应注意，先验参数B的确定是根据公安工作的先验知识或相应领域专家知识进行测试，要符合实际情况。另外，决策中的属性如果有许多先验知识或领域知识，要根据实际情况选择参数，不宜做太多的选择，可以逐步进行，否则会因人为因素影响决策效果。

4.3.3 实验及分析

由传统ID3算法在公安工作中存在问题分析可知，需要降低经济状况在分类中的重要性，相对提高其他属性在分类中的重要性。针对表4.1，设B(经济状况)＝0.35，根据BID3算法及公式(4.8)，来生成基于先验参数的BID3算法决策树。

1．根据 BID3 算法，对根结点进行分类。

由于 B(经济状况)＝0.35，其他属性的先验参数 B＝0，则只需要重新计算 E(经济状况)，Gain(经济状况)，其他值不变。根据公式(4.5)、(4.8)：

$$E(\text{经济状况})=(\frac{6}{15}+0.35)(0)+(\frac{9}{15}+0.35)(-\frac{3}{9}\log_2\frac{3}{9}-\frac{6}{9}\log_2\frac{6}{9})=$$

0.8724。

Gain(经济状况)＝0.0985

比较表 4.2，Gain(文化程度)最大，即按“文化程度”分类。

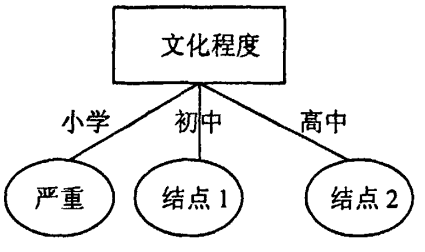

图 4.4 基于 BID3 算法的根结点分类决策树

2. 对图 4.4 中的叶结点 1 进行分类。

表 4.5 基于叶结点 1 的犯罪记录表

序号	年龄	经济状况	文化程度	有无正当职业	社会关系有无犯罪记录	是否有特长	是否常驻人口	犯罪程度
1	20-30	中等	初中	无	无	有	是	较轻
3	30-40	差	初中	无	有	有	是	严重
6	30-40	差	初中	有	有	无	是	较轻
7	20-30	中等	初中	无	有	有	否	较轻
10	20-30	中等	初中	有	有	无	是	较轻
12	>40	差	初中	无	无	无	是	较轻
14	20-30	差	初中	有	有	无	否	严重

由表 4.5，样本集合的属性为“犯罪程度”，有两个取值{较轻，严重}，即 M=2，设 C_1 对应“较轻”类别，C_2 对应“严重”类别，则 $C_1=5$，$C_2=2$。为计算每个属性的信息增益，由公式(4.4)可得分类所需的期望信息：

$$I(5,2)=-\frac{5}{7}\log_2\frac{5}{7}-\frac{2}{7}\log_2\frac{2}{7}=0.8631$$

由公式(4.5)、(4.8)计算各属性的熵：

A1=“年龄”： 20－30、30－40、>40

$$E(A1)=\frac{4}{7}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)+\frac{2}{7}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)+0=0.7493$$

A2=“经济状况”：中、差

$$E(A2)=\left(\frac{4}{7}+0.35\right)\left(-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}\right)+0)=0.9214$$

A3=“有无职业”：无、有

$$E(A3)=\frac{4}{7}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)+\frac{3}{7}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)=0.8572$$

A4=“社会关系有无犯罪记录”：有、无

$$E(A4)=\frac{5}{7}\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\right)+0=0.6935$$

依据 BID3 算法及公式可得如图 4.5 的决策树：

3. 对图 4.4 中的叶结点 2 进行分类。

表 4.6 基于叶结点 2 的犯罪记录表

序号	年龄	经济状况	文化程度	有无正当职业	社会关系有无犯罪记录	是否有特长	是否常驻人口	犯罪程度
4	20-30	中等	高中	有	无	无	是	较轻
8	20-30	差	高中	无	无	有	否	严重
9	30-40	中等	高中	有	无	无	是	较轻
11	20-30	差	高中	无	有	有	否	严重
13	20-30	差	高中	有	无	无	否	较轻
15	20-30	中等	高中	有	有	无	是	较轻

依据 BID3 算法及公式可得如图 4.6 的最终决策树：

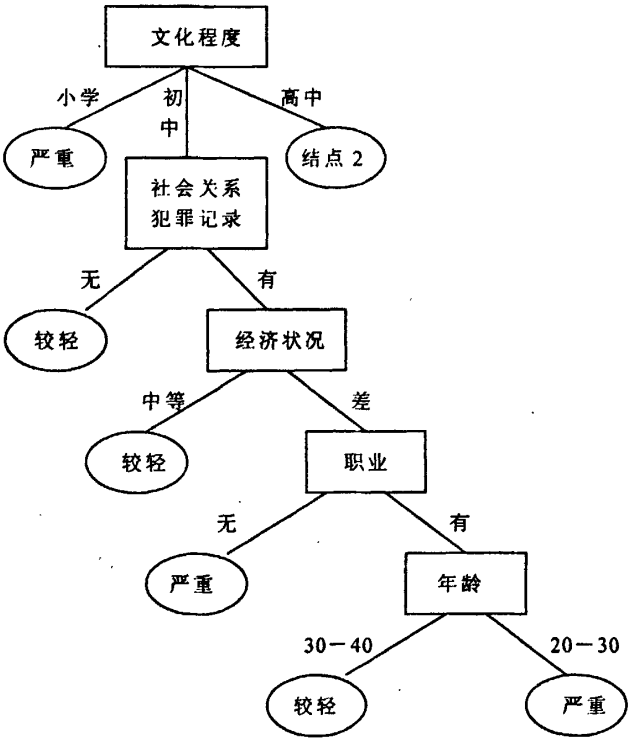

图 4.5 基于 BID3 算法的叶结点 1 分类图

4. 规则归纳。

通过基于 BID3 算法生成的犯罪记录决策树，我们可以看出，文化程度小学的，犯罪程度严重；文化程度高中的，有正当职业的犯罪程度较轻，无正当职业的犯罪程度严重；文化程度初中的情况较为复杂，社会关系中无犯罪记录的和社会关系中有犯罪记录但经济状况中等的，犯罪程度较轻，如经济状况差同时无职业的，犯罪程度严重，如经济状况差同时有职业，年龄在 30－40 的犯罪程度较轻，年龄在 20－30 的犯罪程度严重。 可以描述如下：

（1）IF（文化程度=高中 AND 职业=有）OR（文化程度=初中 AND 社会关系犯罪记录=无）OR （文化程度=初中 AND 社会关系犯罪记录=有 AND 经济状况＝中等)OR（文化程度=初中 AND 社会关系犯罪记录=有 AND 经济状况＝差 AND 职业＝有 AND 年龄＝30－40)

Then（犯罪程度=较轻)；

（2）IF(文化程度=小学）OR（文化程度=高中 AND 职业=无）OR（文化程度=初中 AND 社会关系犯罪记录=有 AND 经济状况＝差 AND 职业＝无）OR（文化程度=初中 AND 社会关系犯罪记录=有 AND 经济状况＝差 AND 职业＝有 AND 年龄＝20－30)

Then（犯罪程度=严重)。

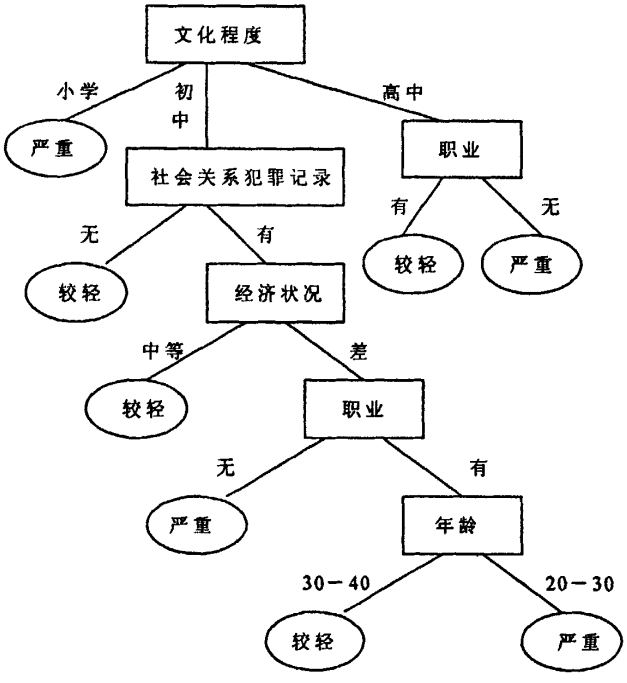

图 4.6 基于 BID3 算法生成的犯罪记录决策树

BID3 算法是在 ID3 算法基础上加入先验参数构成的，它依靠先验知识或领域知识人工增强重要属性在分类决策中提供的信息量，相应减少非重要属性的信息量，在一定程度上优化了 ID3 算法，减少其对取值较多的属性的依赖性，从而改善分类规则和结果，在没有必要的情况下，BID3 算法还可以回归为 ID3 算法。

但 BID3 算法生成的决策树与 ID3 算法生成的有很大的区别，得到的分类规则也不同，有一些也与实际不一致，这与数据包含噪音且数据集又太小有关。如果在大数据量(几十万条)的基础上，采集尽可能多的分类项目(几十项)，并结合犯罪分析的专业知识，从作案对象、手段特点、作案工具、作案时间和场所、专长等特征具体加以分析，这将是一个具有更为现实意义的分析实验[51]。

4.3.4 算法性能测试

为了进一步验证改进算法的性能，在运行 WinXP 的 P4-2.0G，512M 台式电脑上进行了性能测试，软件开发环境为 VFP 6.0。通过测试，分别从以下几个方面对 ID3 算法和 BID3 算法构造的决策树进行如下的对比分析：

1．生成决策树的形态

经过测试分析，发现在不同规模的数据集中，利用 BID3 算法构造的决策树不但可以加快决策树的生长，而且在“根结点数量”、“叶结点数量”、“树的高度”、“规则数量”等几个方面有明显的简化，可以得到结构简单的决策树，便于从中挖掘出易于理解的规则信息。

2．生成决策树的时间

为了证明 BID3 算法有更高的构造效率，分别以不同规模的数据集，利用 ID3 算法和 BID3 算法进行 12 次计算时间的测试，取 12 次计算时间的平均值作为算法构造决策树花费的计算时间。然后通过上述实验数据，对比分析 ID3 算法和 ID3 改进算法在构造决策树花费的计算时间上的差异程度。其中节省时间率＝(BID3 所用平均时间－ID3 所用平均时间)/ ID3 所用平均时间。

表 4.7　ID3 算法和 BID3 算法构造决策树所用的计算时间

记录数量 n	ID3 所用平均时间(ms)	BID3 所用平均时间(ms)	节省时间(ms)	节省时间率(%)
14	562.4	561.6	0.8	0.14
100	573	572	1	0.18
500	633	631	2	0.32
1000	689	687	2	0.29
2000	820	817	3	0.37
5000	1205	1201	4	0.33
8000	1589	1584	5	0.31
10000	1984	1971	13	0.66
15792	2789	2769	20	0.72

从表 4.7 中可以看出，在不同规模的数据集中，BID3 算法构造决策树虽然计算先验知识度参数需要一定的时间开销，但随着记录数据的增大，所花费的时间开销会随着决策树的快速生长得到一定程度的弥补，这充分说明使用 BID3 算法能够以更高的效率构造决策树。

根据表 4.7 中 ID3 算法和 BID3 算法构造决策树所用的计算时间差异，得如图 4.7、图 4.8 和图 4.9 的结果。

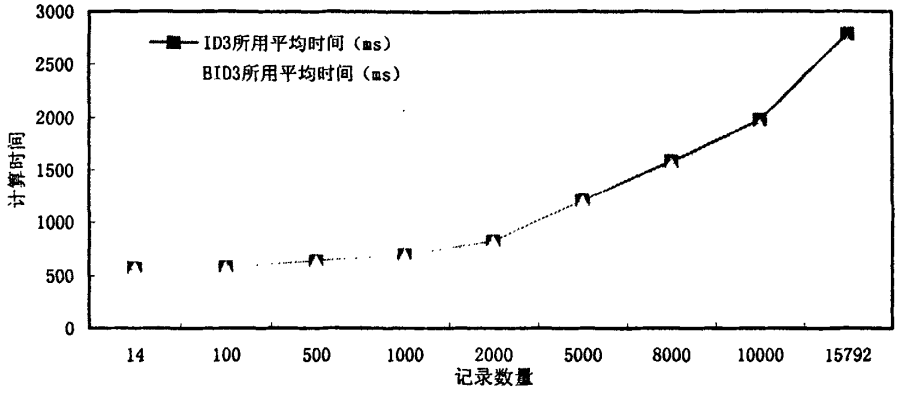

图 4.7 算法 ID3 和 BID3 构造决策树所用时间的对比

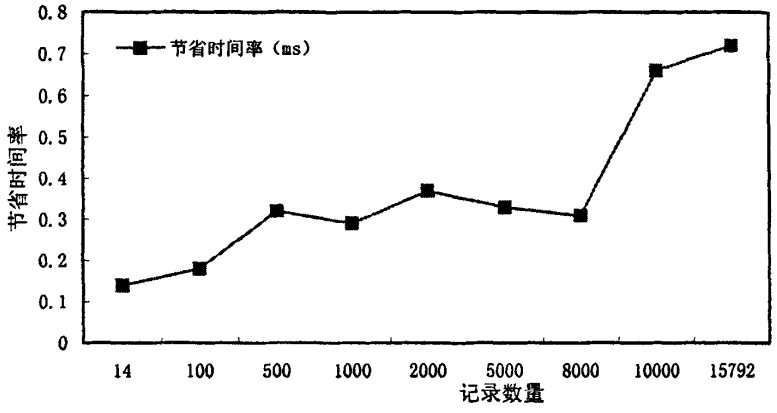

图 4.8 算法 BID3 节省时间率随数据集变化趋势

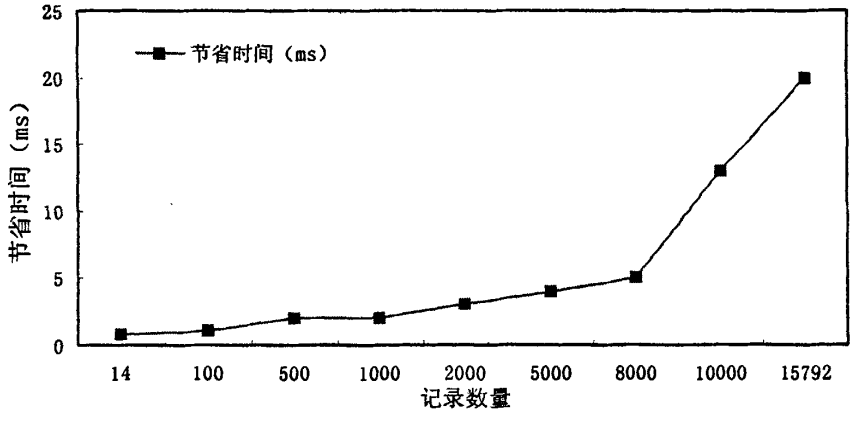

图 4.9 算法 BID3 节省时间随数据集变化趋势

从图 4.7 和图 4.8 中可以看出，在构造决策树的过程中，随着数据集规模的增大，BID3 算法与 ID3 算法相比，所节省的时间也增多了，BID3 算法的高效性越明显。从图 4.8 和图 4.9 中可以看出，在处理越大规模数据集的决策树构造过程中，BID3 算法在效率和性能上比 ID3 算法有更大的优越性。

实验中两个方案的比较表明，该算法是可行的和有效的，利用 BID3 算法来选择决策树分支取值，不但可以加快决策树的生长，而且最重要的是可以得到结构好的决策树，便于从中挖掘好的规则信息。特别是在使用决策树算法来挖掘的数据越多，算法的效率和性能就越好，算法的优越性就越明显。

由于算法改进需要在原来的信息熵计算中加入人为的参与，因此在使用先验知识参数的时候必须注意以下几个方面[52]：

(1)决策者比较关注的属性，可根据先验知识或领域知识进行测试，选择符合实际情况的参数。

(2)当大多数属性数据量较大，个别属性数据量较小，而人们对这些属性重要性认识不足时，有必要设置这些属性的先验知识参数，使其不会出现大数据掩盖小数据的现象。

(3)决策中的属性如果有许多先验知识或领域知识，可根据实际情况选择先验参数，但不宜做太多选择，可逐步进行，否则会因人为因素过多影响决策效果。

4.4 基于决策树算法的犯罪行为分析原型系统的设计与实验

基于决策树算法的犯罪行为分析系统的设计目标是在公安信息数据库的基础上，运用数据挖掘技术，对数据库中的信息进行分析、统计、挖掘等加工处理，从中发现趋势、找到发案规律，并根据这些规律制定相应的处理方案，为各级领导、管理部门和侦察民警分析、发现、掌握公安业务的动向与规律提供及时、准确的各种统计信息，为破案提供辅助决策手段和科学依据[53]。

4.4.1 功能需求与系统流程图

实验系统采用了 C/S 架构（客户/服务器模式），服务器端采用 SQL Server2000 数据库管理系统，然后基于在运行 WinXP 的 P4 2.0G，内存 512M、硬盘 80G 的台式电脑，开发平台为 Delphi 平台上进行了实验系统的设计开发。

实验系统将实现分别以 ID3 算法和 BID3 算法对公安违法犯罪人员记录数据库进行数据挖掘和规则提取，系统具有数据预处理功能，能解决部分数据挖掘问题。

实验系统流程图见图 4.10。服务器端管理从其他数据源获得的不同格式的数据，如文本格式、Oracle 数据、Dbase 格式、Access 格式、Excel 格式等，对这些数据进行筛选、清理、冗余检查、格式转换等预处理准备工作，经过选择、预处理后的业务数据集以 SQL 格式存入待挖掘数据库中，即完成数据预处理功能，包含数据获取、数据取样、数据筛选、数据转换等四个子功能。

客户端作为程序主题部分，负责确定挖掘主题，即定义挖掘目标，设置因变量属性，选择相应决策树算法(如 BID3 算法、ID3 算法等)，生成决策树，提取规则，即完成数据挖掘功能，包含数据挖掘库管理、数据挖掘过程、规则库管理、

规则评价等四个子功能。然后用业务数据集进行检测，如规则合理即可输出，否则循环此过程，直至生成满意的决策树，得到符合事实的规则，即完成数据评价功能，包含结果处理和结果评价等两个子功能。

4.4.2 原型系统模块构成

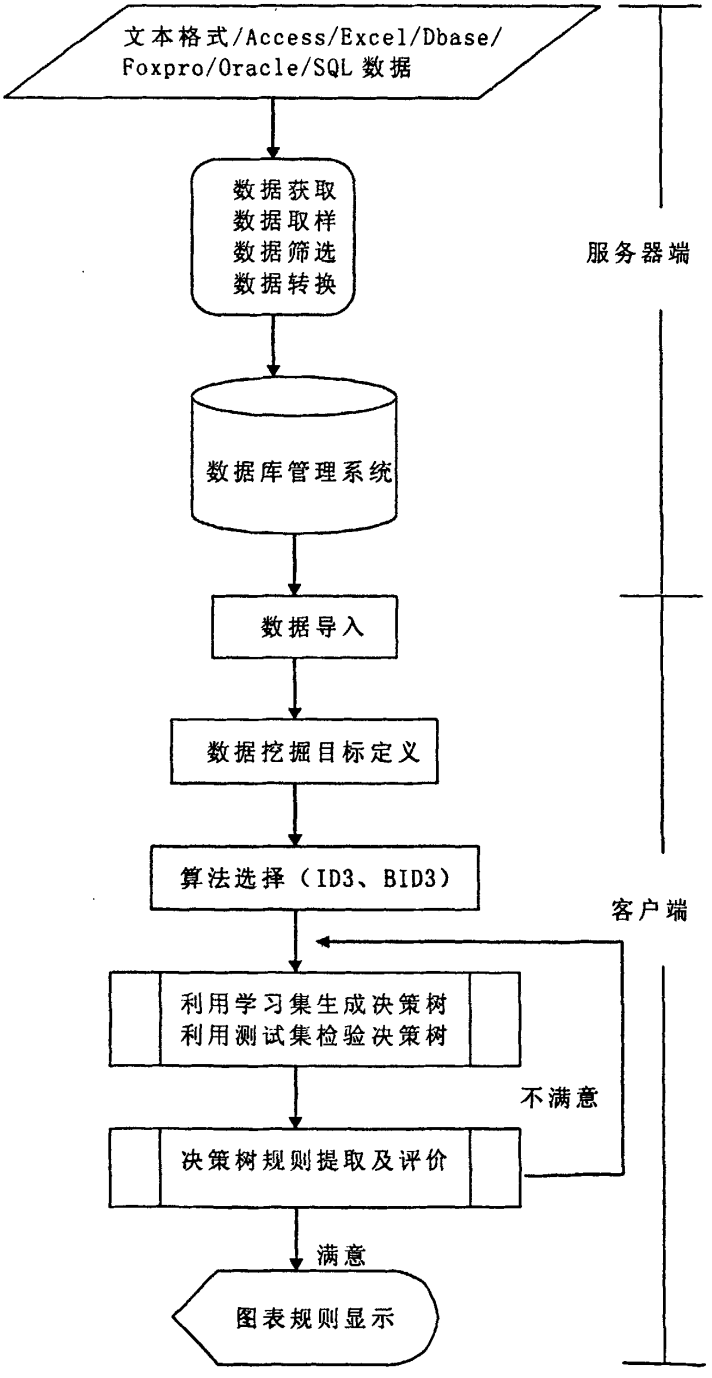

图 4.10 系统流程图

根据系统所要实现的功能，可以把系统分为三个大模块：数据预处理模块、数据挖掘模块、数据评价模块[54]。如图 4.11，其主要功能模块的说明如下：

1．数据预处理模块。

预处理是负责对待挖掘的数据源作必要的准备，对来自关系数据库、多维数据库、数据仓库或文件系统的数据进行转化，对于大数据集，可以通过数据采样减少处理的数据量，然后利用数据清理等手段清除脏数据，同时可以利用概念层次树对原始数据进行必要的抽象，最终将数据处理成数据挖掘算法可以处理的数据[55]。

包含数据获取、数据取样、数据筛选、数据转换四个子模块。数据获取指定要使用的数据类型、名称及位置；数据取样对获取的数据从中取样，取样的方式有很多种，根据数据挖掘的目的灵活选用；数据筛选通过对数据筛选筛选掉不希望包括进来的观测值；数据转换将某一个数据进行某种转换操作，然后将转换后的值作为新的变量存放在样本数据库中，而转换的目的是为了把数据和将来要建立的模型拟合得更好。

2.数据挖掘模块

包含数据挖掘库管理、数据挖掘过程、规则库管理、规则评价四个子模块[56]。

数据挖掘库管理主要完成对经过预处理的数据的日常数据管理工作，提供在数据挖掘过程中按要求要提取数据的功能；

数据挖掘过程使用 ID3 或 BID3 算法对数据进行挖掘，它可以使用索引、并行、删减分支等技术提高挖掘效率。挖掘后产生的规则存入规则库。可以分以下几步骤：

一是数据挖掘前的数据准备。完成数据的提取，进行一些处理，为数据挖掘创造一个良好的挖掘环境。二是数据挖掘过程。完成使用改进 ID3 算法对数据进行挖掘的过程。三是规则库管理。主要完成对数据挖掘产生的规则的存储、导入、导出、查询、备份等日常的数据管理工作，提供按要求要提取规则数据的功能。四是规则评价。主要完成规则库里产生的规则的测试评价功能；

3．数据评价模块

数据评价用一种通用的数据挖掘评价的架构来比较不同模型的效果；预报各种不同类型分析工具的结果。在进行各种比较和预报的评价之后，给出一系列标准的图表，供用户进行定量评价。将数据挖掘模块存储在挖掘库中的结果以可视化的形式表示出来。对于挖掘结果，用户可以进行评价结果的信任程度，并将其存储在数据挖掘库中，供以后挖掘使用。

在具体的系统实现数据挖掘过程中需要循环调用以上模块，直至获得满意的决策信息为止。实验系统主要的程序实现是基于上述数据挖掘模块和决策树规则提取模块[57]。

4.4.3 原型系统实现

系统用户界面如图 4.12 所示。包括两个功能模块：预处理、分类。预处理模块：选择、修改所操作的数据。分类模块：训练、测试用于分类的学习策略。界面介绍如下：

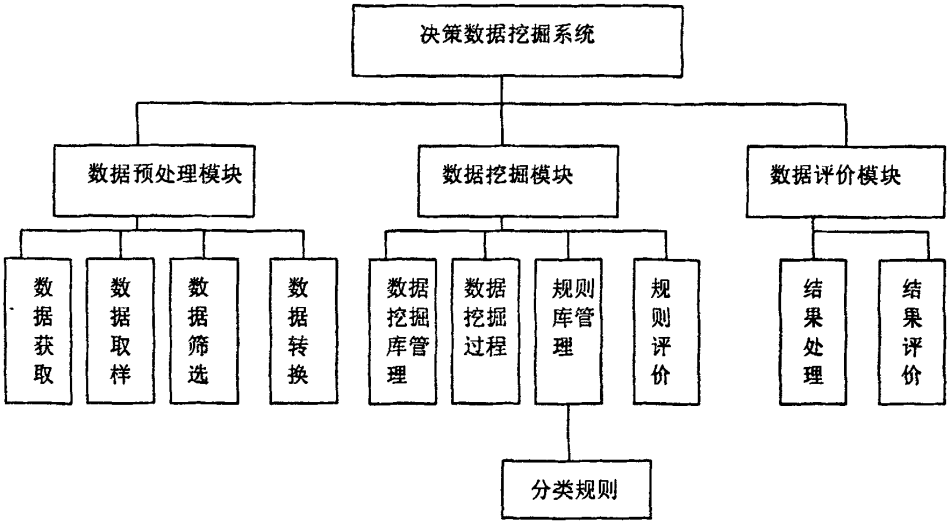

图 4.11 系统功能模块图

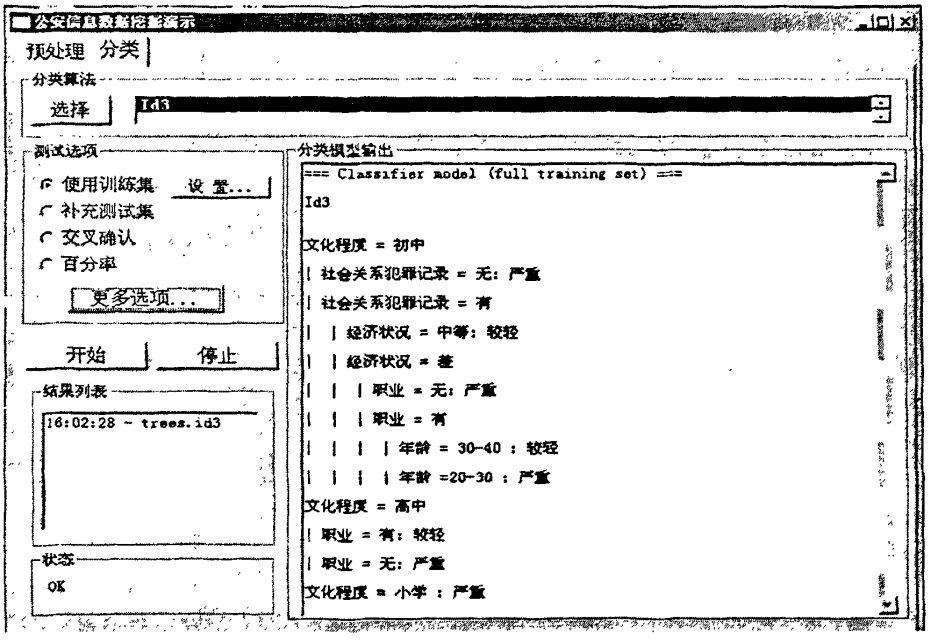

图 4.12 系统演示界面

(1)选择分类方法。“选择”按钮可以用来选择系统中的分类方法。

(2)测试选项框。应用选定的分类方法所得出的结果将根据所设置的测试选项

进行测试。测试有如下四种模式：使用训练集：用来评估分类方法的好坏，即采用这种分类方法对所训练的记录预分类；补充测试集：用来评估分类方法的好坏，采用这种分类方法对一组从数据文件读入的记录进行预先分类；交叉确认：通过交叉验证来评估；百分率：通过预测用于测试的数据的百分比来评估。

(3)分类模型输出框。分类方法输出区域可方便用户浏览结果。

(4)结果列表框。结果列表框将生成含运行时间和算法的结果列表。

4.5 在公安实际系统中的应用信息

在我省已建设的公安22个一类系统，9个二类系统，15个其他系统中，应用数据挖掘技术的几乎没有，但现正全面升级改造，投入试运行的《出入境管理大集中系统》（以下简称“系统”）引入了数据挖掘的分类技术来挖掘有益的信息。

由于我省居民出国境后从事非法活动的数量在全国排名靠前，因此，对这类可能犯罪人员尤其是在港澳犯罪人员进行分类预测就具有相当重要的现实意义。

系统引入了数据挖掘的分类技术，在拟定算法下对大量已遣返的在港澳犯罪人员记录进行详细分析，生成可能犯罪的规律、趋势，及何种状态会诱发何种犯罪行为等，建立了一个赴港澳人员犯罪风险预测模块，并对申请赴港澳人员生成犯罪风险预测提示。

我省居民在港澳的违法犯罪行为一般有三大类，即非法就业、卖淫、刑事犯罪。系统通过对大量录入数据的决策树分析，发现：有正规工作单位且持有一次旅游签证进入港澳的人员基本无犯罪倾向；无正规工作单位且为40岁以下年轻女性且学历较低且一年内变换方式（或护照或港澳通行证，或旅游或商务签证等）多次赴港澳的人员中近70%有在港澳卖淫倾向；无正规工作单位且有违法记录且年龄在50岁以下且为旅游签证进入港澳的人员中有近30%有在港澳刑事犯罪倾向；无正规工作单位且变换方式多次赴港澳达到长期停留目的且年龄在50岁以下且学历不高的男性中近40%有在港澳非法就业犯罪倾向。

通过系统生成的犯罪风险预测提示，审批民警能对可能的犯罪人员的港澳申请进行认真严格的审查，及时发现问题并不予批准其赴港澳申请。系统试运行的几个月以来，公安出入境管理部门根据犯罪风险预测提示，有效的阻止了潜在犯罪倾向人员的赴港澳申请，对无犯罪风险预测提示的人员热情积极服务，充分履行了公安部门严格执法、热情服务的工作职责，同时，我省在港澳从事非法活动的全国排名逐月连续下降。

4.6 本章小结

本章介绍了决策树ID3算法的基本原理、基本概念、优劣分析，简述了C4.5、

SLIQ、SPRINT 等分类算法原理。

本章针对传统 ID3 算法在公安犯罪行为分析中存在的问题：一是按照使熵值最小的原则,被 ID3 算法列为应该首先判断的属性,在现实情况中却并不那么重要；二是当大多数属性数据量较大，个别属性数据量较小时，容易出现大数据淹盖小数据的现象，从而失去一些相应重要的判断。提出了一种基于先验参数的改进算法 BID3，BID3 算法对传统 ID3 算法的属性选择标准进行了改进，通过增加先验参数 B，加强了对重要属性的标注，使得决策树减少了对取值较多的属性的依赖性，尽可能减少大数据掩盖小数据的现象发生。通过生成决策树的形态和生成决策树的时间等实验及分析表明：在处理越大规模数据集的决策树构造过程中，BID3 算法在效率和性能上比 ID3 算法有更大的优越性。

本章在结合公安犯罪行为分析实际的基础上，进行了基于决策树算法的犯罪行为分析原型系统设计与实验，提出了功能需求与系统流程图，介绍了原型系统模块构成和实现，实验分析表明：运用数据挖掘技术对公安信息数据库中的信息进行分析、统计、挖掘等加工处理，从中发现趋势，找到发案规律，为警务决策提供支持服务，在公安工作中具有重要的现实意义。

结 论

当前，我国正处于人民内部矛盾凸显、刑事犯罪高发、对敌斗争复杂的时期。公安机关维护社会稳定、保障人民安居乐业的主要职能不断加重，严重的刑事犯罪和社会治安问题日益严峻，犯罪的动态化、组织化、职业化和智能化的趋向越来越明显，新的犯罪形式和犯罪手段不断出现，严重危害了人民群众的生命财产安全，严重影响了国家安全和社会稳定。如何通过科技手段深入挖掘警务信息资源的效能，将预防与控制点前移，以增强警务工作的主动性，提高执法效率与快速反应能力、及时的预防与打击犯罪行为，成为公安工作中急需解决的问题。

本文正是着眼于数据挖掘技术在公安工作中的应用研究，着重介绍了数据挖掘关联规则与决策树算法在公安工作的犯罪行为分析中的应用和存在的问题，根据公安工作的特殊性提出了改进算法，设计了基于决策树算法的犯罪行为分析系统。

论文所做的主要工作和创新点：

（1）将数据挖掘技术引入公安领域，介绍了全国、全省公安工作与信息化发展现状，阐述了数据挖掘技术在公安工作中应用的重要性和必要性。

（2）针对传统Apriori算法在公安犯罪行为分析中忽略了不同项间不同重要性的问题，提出了一种改进的加权关联规则挖掘算法WARMA，改进算法设计了加权关联规则模型，通过提出的k-支持期望概念来排除不可能成为加权频繁k-项集的候选集子集，解决了加权关联规则挖掘中加权频繁项集的子集可以不是加权频繁项集的问题。实验及分析表明：算法WARMA在不同数据量下的执行时间远小于Apriori，在相同加权支持度阈值条件下，加权关联规则算法产生的频繁项集和执行时间均小于Apriori算法，证实WARMA算法能更有效的发现重大犯罪行为。

（3）针对传统Apriori算法在公安犯罪行为分析中无法及时敏锐的发现某些新型犯罪行为的问题，提出了基于FUP的改进关联规则增量式更新算法SFUP，SFUP算法通过用敏感参数衡量对新项目的重视情况，从敏感性和时间效率出发，改进了一般增量式更新算法产生频繁项集的过程。实验及分析表明：改进算法SFUP在越大数据量下优势越明显，能较好地发现新增数据中的新模式，具有较高的敏感性，能更有效的发现新型犯罪行为，同时在挖掘过程中显示了良好的空间和时间性能，大大提高了整个挖掘效率。

（4）针对传统 ID3 算法在公安犯罪行为分析中存在的问题：一是按照使熵值最小的原则,被 ID3 算法列为应该首先判断的属性,在现实情况中却并不那么重要；

二是当大多数属性数据量较大，个别属性数据量较小时，容易出现大数据淹盖小数据的现象，从而失去一些相应重要的判断。提出了一种基于先验参数的改进算法 BID3，BID3 算法对传统 ID3 算法的属性选择标准进行了改进，通过增加先验参数 B，加强了对重要属性的标注，使得决策树减少了对取值较多的属性的依赖性，尽可能减少大数据掩盖小数据的现象发生。通过生成决策树的形态和生成决策树的时间等实验及分析表明：在处理越大规模数据集的决策树构造过程中，BID3 算法在效率和性能上比 ID3 算法有更大的优越性。

（5）在结合公安犯罪行为分析实际的基础上，进行了基于决策树算法的犯罪行为分析原型系统设计与实验，提出了功能需求与系统流程图，介绍了原型系统模块构成和实现，实验分析表明：运用数据挖掘技术对公安信息数据库中的信息进行分析、统计、挖掘等加工处理，从中发现趋势，找到发案规律，为警务决策提供支持服务，在公安工作中具有重要的现实意义。

本文运用数据挖掘关联规则、决策树技术在公安犯罪行为分析中进行了一些应用和研究，得到了一些具有实用价值的初步结果，但由于自身水平和客观条件限制，研究工作还不完善，还存在问题，需要进一步研究和努力。

(1) 本文提出的数据挖掘改进算法还是一种探索性质的实验，仅仅是提出了改进思想和理论步骤，尽管有一定的应用，但效果与实际中存在一定差距,在公安实际应用中还有待于进一步改进和完善，对于整个公安信息数据挖掘系统真正实施的架构、体系还需要进一步深入研究。

(2)尽管开展了基于犯罪行为分析的决策树原型系统的设计与实验，但还有许多不足之处，如数据表中的离散化数据缺少一定的灵活性，不能够完全反映复杂类型数据的信息，需要进一步改进；其次数据是基于内存而不是磁盘或磁盘组，也未能实现动态的加载数据，另外程序还有待于进一步优化以提高程序运行速度。

(3)继续补充公安业务相关数据，继续致力于对引发违法犯罪嫌疑人犯罪程度的客观属性的探究，只有具有完整准确的属性才能实现对更多问题的分析和决策；继续补充嫌疑人记录，只有海量数据，才能使得出的模型更加接近真实。

(4)聚类分析、序列模式、OLAP 和可视化技术等在公安业务数据中同样起着非常重要的挖掘作用，还需进一步的研究。

参考文献

[1] Paolo Giudici.实用数据挖掘. 袁方，王煜，王丽娟等译.第一版.北京:电子工业出版社，2004，1-3

[2] David Hand，Heikki Mannila，Padhraic Smyth.数据挖掘原理.张银奎，廖丽，宋俊等译.第一版.北京：机械工业出版社，2003，1-3

[3] Y.Liu,S.Y.Sung and H.xiong.A cubic-wise balance approach for privacy preservation in data cubes. Ingormation Sciences. May 2006, Vol. 176, Issue, 9:1215-1240

[4] Jiawei Han，Micheline Kamber.数据挖掘概念与技术.范明,孟小峰等译.第一版.北京：机械工业出版社，2001，14-18

[5] 朱玉全，杨鹤标，孙蕾.数据挖掘技术.第一版.南京：东南大学出版社，2006，7-9

[6] R.Agarwal, C.Aggarwal, and V.V.V.Prasad. A tree projection algorithm for generation of frequent item sets. In Journal of Parallel and Distributed Computing(Special Issue on High Performance Data Mining),2000，23-26

[7] J.Han, J.Pei, and Y.Yin. Mining frequent patterns without candidate generation. In Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD' 00),Dallas,TX, May 2000，1-12

[8] Bryant,Bradley T.,Kulkarni,Arun D Decision trees for data mining Intelligent Engineering Systems Through Artificial Neural Networks,2002,(12):56-58

[9] Kwasnicka,Halina.Doczekalski,Marcin Data mining generation and visualization of dicision trees Systems Science 2003,(3):35-37

[10] Hu,Hui.Golosinski,Tad S Modeling failure pattern of a mining truck with a decision tree algorithm Mineral Resources Engineering 2004,(2):44-47

[11] Vapnik V N.The Nature of Statistical Learning Theory.New York: Springer-Verlag，2005，33-39

[12] Tong X，Qi L.On the convergence of a trust-region method for solving constrained nonlinear equations with degenerate solutions.Journal of Optimization Theory and Applications,2004,123(1):187-211

[13] Lian Yan.Predicting customer behavior in telecommunication

[J]. Intelligent systems, IEEE, VOL19, Issue2, March-April 2004:50-58

[14] Lian Yan. Predicting customer behavior in telecommunication [J]. Intelligent systems, IEEE, VoI19, Issue2, March-April 2004:50-58

[15] Ruggieri S. Efficient C4. 5. IEEE Transactions Knowledge and Data Engineering, 2002, 14(2):43-44

[16] K.T'hearling. An Introduction to Data Mining:Discovering hidden value in your data warehouse, 2004, 75-79

[17] 毛国君，段立娟，王实等.数据挖掘原理与算法.第一版.北京:清华大学出版社，2005，29-32

[18] 罗可.数据库中数据挖掘理论方法及应用研究.[湖南大学博士学位论文].长沙：湖南大学电气与信息工程学院，2004，15-16

[19] 张辉.数据挖掘技术及其在刑侦工作中的应用.信息技术与信息化，2005，(4):111－113

[20] 丁丽萍.论信息挖掘在计算机犯罪侦查技术中的应用.公安大学学报(自然科学版)，2002，29(3):47-50

[21] 郑永红.犯罪信息工作中的数据挖掘技术.广东公安科技，2005，(1):39-41

[22] 李润森.全国公安信息化工程－金盾工程.计算机安全，2002，(1):7-111

[23] 张亮.“金盾工程”中公安信息系统结构.警察技术，2002，(3):13-15

[24] 张蕾华.试论公安工作信息化的必然性.山西警官高等专科学校学报，2003，11(4):17-18

[25] 关莅.公安信息化建设的战略思考.浙江高等专科学校学报，2004，(5):25-28

[26] 金光，钱家麟，黄蔚民.公安业务信息数据挖掘研究.警察技术，2003，(4):7-9

[27] 梅传强.犯罪心理学.第二版.北京：中国法制出版社，2007，1-4

[28] 肖劲松，林子禹，毛超.关联规则在零售商业的应用.计算机工程，2004，30(3):189-190

[29] 杜鹢.数据挖掘中关联规则的研究与应用.[中国人民解放军理工大学博士学位论文].南京：中国人民解放军理工大学军事通信学院，2000，8-10

[30] Pand-Ning, Tan Michael Steinbach, Vipin Kumar. 数据挖掘导论.范明，范宏建等译.第一版.北京:人民邮电出版社，2006，202-207

[31] 潘福铮.数据挖掘中的关联规则.湖北大学学报(自然科学版)，2002，(4)：25-29

[32] 刘以安，羊斌.关联规则挖掘中对Apriori算法的一种改进研究.计算机应用，2007，(2):168-170

[33] 刘星沙，谭利球，熊拥军.关联规则挖掘算法及其应用研究.计算机工程与科学，2007，(1):87-89

[34] 张有承，梁颖红.数据挖掘技术及其应用研究.科技信息，2007，(35):610

[35] S.Foucaud, A.Zanichelli, B.Garilli, M.Scodeggio and P.Franzetti. VIPGI and Elise 3D:Reducing VIMOS-IFU data and searching for emission line sources in data cubes. New Astronomy Reviews, In Press, Corrected Proof, Available online. 19 April 2006

[36] N.Jukic and S.Nestorov. Comprehensive daa warehouse exploration with qualified association-rule mining. Decision Support Systems, In Press, Corrected Proof, Available online. 11 August 2005

[37] F.Berzal, I.Blanco, D.Sanchez, J.M.Serrano and M.A.Vila. A definition for fuzzy approximate dependences. Fuzzy Sets and Systems. January 2005, Vol.149, Issue 1:105-129

[38] 许军.基于公安信息的数据挖掘应用研究.[南京工业大学硕士学位论文].南京：南京工业大学计算机应用技术系，2006，38-39

[39] 周明辉，洪伟.基于权值的数据挖掘算法。杭州电子工业学院学报，2004，24(1):91-94

[40] 朱孝宇，王理冬，汪光阳.一种改进的 Apriori 挖掘关联规则算法.计算机技术与发展，2006，16(12):89-90

[41] 张智军，方颖，许云涛.基于 Apriori 算法的水平加权关联规则挖掘.计算机工程与应用，2003，(14):197-199

[42] 杨学兵，安红梅.一种高效的关联规则增量式更新式算法.计算机技术与发展，2007，(1):114-116

[43] 陈爱萍.关联规则增量算法.电脑知识与技术，2005，(36):20-22

[44] 伊卫国，卫金茂，王名扬.基于项目集加权的增量关联规则算法研究.计算机工程与应用.2004，(34):192-194

[45] 蒙韧，苏毅娟，朱晓峰等.数据挖掘中的增量式关联规则更新算法.广西科学院学报，2006，22(2):125-128

[46] 翟俊海，张素芳，王熙照.ID3 算法的理论基础.兰州大学学报(自然科学版),2007,(1):71-74

[47] 钱江波，陈珺，屠一波等.犯罪分析决策树的构造方法.警察技术，2004，(2):9-11

[48] 刘红岩，陈剑，陈国青.数据挖掘中的数据分类算法综述.清华大学学报(自然科学版)，2002,(6):18-21

[49] 宾宁,李宏,陈松乔.基于 SPRINT 分类算法的异构分布式数据挖掘研究.计算机测量与控制,2005,(1):79-81

[50] 吴俊.数据挖掘技术在公安出入境管理中的应用研究.[合肥工业大学硕士学

位论文].合肥：合肥工业大学工商管理学院，2006，40-44

[51] 金光，刘士荣，李荣茜等.数据挖掘技术在犯罪行为分析中的应用.宁波大学学报(理工版)，2002，15(1):56－58

[52] 郭景峰等.决策树算法的并行性研究.计算机工程，2002，28 (8)：85-86

[53] 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):191-193

[54] 黄建设，姚奇富.数据挖掘技术在犯罪行为分析中的应用.浙江工商职业技术学院学报，2005，4(3):45-47

[55] 刘光明.ID3 算法的研究及在以政府决策为主题的挖掘系统中的应用.[南昌大学硕士学位论文].南昌：南昌大学计算机中心，2006，38-41

[56] 梅世军.刑事案件决策支持系统研究与实现.[同济大学硕士学位论文].上海：同济大学软件学院,2007,41-48

[57] 成文丽.基于决策树的数据挖掘算法的技术研究.[太原理工大学硕士学位论文].太原：太原理工大学，2003，75-76

附录 A 攻读学位期间发表的论文和参加的项目

攻读学位期间发表的论文：

[1] 陈湘涛，刘华. 数据挖掘技术在公安工作中的应用研究.中国科技信息，2007年,(3)：94-95

攻读学位期间参加的项目：

[1] 湖南省警务综合应用平台，公安部“金盾工程”二期重点建设项目，2006—2008 年。

致 谢

在本人攻读工程硕士学位期间，一直得到导师陈湘涛副教授的细心关怀和辛勤指导，在我论文的选题、开题、学术研究、论文撰写的每一个阶段都凝聚着陈教授的心血，每次在他不厌其烦指导我修改论文时，我总会庆幸自己遇到了一位博学多才、认真负责的导师。他高深的学术造诣、严谨的治学态度、高度的责任感和诲人不倦的高尚师德使我受益匪浅、终身难忘。在此谨向我敬爱的导师表示最崇高的谢意！

感谢计算机与通信学院所有教导过我的各位老师，感谢他们给我专业上的细心指导，为我拓展专业知识，扩展视野奠定了坚实基础。

感谢计算机与通信学院学科办的各位老师，感谢他们给我的关心和帮助。

感谢师兄、师弟、师妹们以及诸位同学，他们在学习上给我很大的帮助和支持，他们的学习态度和精神非常值得我学习。

最后感谢我的父母、妻子和朋友，无论我遇到什么样的困难和挫折，他们都始终是我的坚强后盾，给我安慰并赋予我勇气和力量，使我不断前进。

感谢各位老师和专家百忙之中对本文的审阅和提出的宝贵意见！

数据挖掘技术在公安犯罪行为分析中的应用研究

作者： 刘华
学位授予单位： 湖南大学
被引用次数： 1次

本文读者也读过(9条)

1. 吴俊 数据挖掘技术在公安出入境管理中的应用研究[学位论文]2006
2. 杨心智 基于数据仓库技术的公安综合信息分析系统设计与实现[学位论文]2005
3. 杜威.毛莉.彭晗.彭建新 基于决策树挖掘算法的犯罪行为研究[期刊论文]-广东公安科技2010,18(4)
4. 徐伟.张军 关联规则在犯罪行为分析中的应用研究[会议论文]-2008
5. 李刚 数据挖掘技术在侦破网络犯罪中的应用[期刊论文]-西北民族大学学报（自然科学版）2006,27(4)
6. 张予 数据挖掘技术在高危人员犯罪信息挖掘的应用研究[学位论文]2009
7. 许军 基于公安信息的数据挖掘应用研究[学位论文]2006
8. 黄建设.姚奇富.HUANG Jian-she.YAO Qi-fu 数据挖掘技术在犯罪行为分析中的应用[期刊论文]-浙江工商职业技术学院学报2005,4(3)
9. 杨彦维 基于数据挖掘的犯罪行为分析应用研究[学位论文]2007

引证文献(1条)

1. 唐德权.张悦.贺永恒.肖自红 基于图数据挖掘算法的犯罪规律研究及应用[期刊论文]-计算机技术与发展2011(11)