Computer Science and Application 计算机科学与应用, 2019, 9(2), 339-350
Published Online February 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.92039

Hans汉斯

Comparison of Various Classification Techniques for Ethnic Minority Crime Data Mining Based on Bayes and Decision Tree

Shaobing Wu[1,2], Changmei Wang[2*]

[1]Institute of Information Security, Yunnan Police College, Kunming Yunnan
[2]Solar Energy Institute, Yunnan Normal University, Kunming Yunnan
Email: *823804919@qq.com, 1063093199@qq.com

Received: Jan. 28th, 2019; accepted: Feb. 6th, 2019; published: Feb. 13th, 2019

Abstract

Considering the slow speed and inconvenient operating of crime data mining method, we adopted three methods such as the use of Minitab, Bayes and decision trees to mine the data of a national crime area respectively. Then we compared the complexity, difficulty and accuracy of these three methods in mining crime law through concrete examples. In order to simplify calculation and facilitate practice, we designed program interface by C#. Experimental results showed that Bayesian approach performs better in predictive accuracy than Minitab and decision trees.

Keywords

Ethnic Minority Crime Data Mining, Bayes, Decision Tree, Minitab, Classification

一种基于贝叶斯和决策树的少数民族犯罪数据挖掘方法比较研究

吴绍兵[1,2]，王昌梅[2*]

[1]云南警官学院信息网络安全学院，云南 昆明
[2]云南师范大学太阳能研究所，云南 昆明
Email: *823804919@qq.com, 1063093199@qq.com

收稿日期：2019年1月28日；录用日期：2019年2月6日；发布日期：2019年2月13日

*通讯作者。

文章引用: 吴绍兵, 王昌梅. 一种基于贝叶斯和决策树的少数民族犯罪数据挖掘方法比较研究[J]. 计算机科学与应用, 2019, 9(2): 339-350. DOI: 10.12677/csa.2019.92039

摘　要

鉴于犯罪数据挖掘方法应用的计算速度慢、操作不便性等问题，采用了Minitab、贝叶斯和决策树三种方法分别对某民族地区的犯罪数据进行挖掘研究，通过具体实例比较了三种方法挖掘犯罪数据规律的复杂性、难易性和准确性。为了方便实用，简化计算，采用C#设计了程序界面。实验研究结果表明，通过采用贝叶斯的方法预测准确性优于Minitab和决策树。

关键词

少数民族犯罪数据挖掘，贝叶斯，决策树，Minitab，分类

1. 引言

我国正处于国际国内形势严峻、人民内部矛盾凸显、刑事犯罪高发的时期，刑侦部门打击刑事犯罪、保护人民生命财产、维护社会稳定的任务更加繁重而艰巨。随着计算技术、网络技术、通信技术和 Internet 技术的发展，公安行业积累了大量的业务数据，这些数据和由此而产生的信息、情报、知识是侦查破案、打击犯罪、保持社会稳定和维护治安秩序的关键，是公安行业的最大财富。本文针对已有方法存在的问题，提出了将 Minitab、贝叶斯和决策树方法进行比较研究，比较了三种方法的犯罪挖掘复杂性、预测准确度。并对模型结果进行了可视化编程实现。

2. 相关工作

一些学者，在犯罪网络分析与犯罪数据挖掘的研究中，提出了犯罪数据挖掘的一些方法，为犯罪数据挖掘的理论发展做出了重大贡献。数据挖掘应用于犯罪集团或恐怖组织社会网络分析是一种新兴的研究方法，国内外在分析犯罪和恐怖组织之间通信行为方面的研究工作亟待深入[1]，为了挖掘犯罪网络的核心成员，乔少杰等[1]提出了一种基于社会网络分析挖掘犯罪组织核心成员的算法 CNKM (Crime Network Key Member Mining)，通过使用合适的通信数据流分析工具，模拟和分析社会网络中具有不同个性特征个体的通信规律，并检测异常通信行为，进而达到预测犯罪/恐怖活动的目的，为我们进行犯罪数据挖掘提供了一个思路和参考。金光等[2]用数据挖掘中的决策树技术对犯罪行为进行分析，给出了一个较为成功的挖掘思路和模式，得出了一个犯罪风险预测模型。黄建设等[3]利用数据挖掘技术对犯罪行为进行分析，得出了数据挖掘技术在犯罪行为分析中的应用方案。李万彪等[4]通过为期半年的通话与转账数据作为数据源进行实验，利用社会网络分析手段进行犯罪网络分析，分析数据特点，建立关系数据模型，从各类数据资源中发掘犯罪团伙信息。杨莉莉等[5]通过对犯罪组织重点人员的判定以及犯罪组织成员间关系的分析，提出了基于社会网络的犯罪组织关系挖掘方法。G. C. Oatley 等[6]研究了匹配和犯罪预测技术在智能系统中的应用。R. William Adderley [7]在其博士论文中，利用数据挖掘技术进行犯罪趋势分析，并给出了犯罪数据挖掘方法的一个框架。吴绍兵[8]采用贝叶斯网络和 EM 算法来分析影响某地区刑事犯罪的影响因素，给出了影响因素模型，并由该模型得出，影响该地区的刑事犯罪的因素依次是人的因素、环境因素、犯罪类型因素、犯罪位置因素和交通因素等。

以上对犯罪数据挖掘的研究，大多是针对犯罪数据挖掘的算法，进行理论的研究，而实现和可视化应用研究的较少，本文针对已有方法存在的问题，提出了将贝叶斯、Mintab 和决策树三种方法结合的犯罪数据挖掘方法，实验结果表明，该贝叶斯方法的挖掘效果优于 Minitab 和决策树。

3. 数据挖掘方法

3.1. 数据挖掘概述

数据挖掘(Data Mining, DM)是指从大量数据(包括文本)中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。数据挖掘的目的就从数据中“淘金”，就是从数据中获取智能的过程[9] [10]。

3.2. 数据挖掘方法

3.2.1. 数据预处理

数据预处理就是数据中包括的不一致性数据、缺失数据、重复数据、不合理数据、虚假数据、异常数据、逻辑错误数据、非结构化数据、半结构化数据等变为计算机可处理的结构化数据。主要包括抽样、降维和去噪[11]三个阶段。文献[12]从公共数据库的角度提出数据清洗应尽量满足如下要求：

① 尽量放松清洗规则，保证数据的原样性；

② 在清洗过程中，只能作数据映射，不能修改用户数据，不对错误数据进行纠正；

③ 清洗的数据要保存。

1) 抽样

抽样是数据挖掘从大数据集中选择相关数据子集的主要技术。它用于预处理和最终解释步骤中。之所以使用抽样是因为处理全部数据集的计算开销太大。抽样的关键是发现具有整个原始数据集代表性的子集，其具有与整个数据集大概类似的兴趣属性。最简单的抽样技术是随机抽样，任意物品被选中的概率相同。

2) 降维

随着大数据时代的到来，不仅有定义高维空间特征的数据集，也有在空间中信息非常稀疏的数据集。稀疏和维度灾难是大数据挖掘中反复出现的问题。即使在最简单的背景下，也很可能会有成千上万条记录的行和列稀疏矩阵，其中大部分值是零。因此降低维度就自然而然的了。在大数据挖掘中，最相关的降维算法有两个：主成分分析(PCA)和奇异值分解(SVD)。

3) 去噪

数据挖掘中采集的数据可能会有各种噪声，如缺失数据，或者是异常数据。去噪的目的是在最大化信息量时去除掉不必要的影响。

3.2.2. 数据分析

数据分析是从大量数据中发现有趣、有价值的信息，是整个数据挖掘过程的核心部分。而数据分析的成功与否不仅与研究者所运用的一系列理论、算法有关，还与数据分析所涉及的具体领域有着重要的关系。

3.2.3. 结果评价与展示

数据分析的目的是为了让用户能够从中获得有用的信息并对结果进行评价。其难点在于用户并不具备相应的专业知识，因此需要更加具体与形象化的表示，例如可视化技术。这也正是大数据环境下数据挖掘中的一个研究热点与难点。大数据处理的基本流程是数据挖掘思想在大数据环境中的具体表现，因

此有很多的相似之处[13]。

3.3. 数据挖掘算法

3.3.1. Minitab

根据百度百科记载，Minitab 软件是现代质量管理统计的领先者，全球六西格玛实施的共同语言，以无可比拟的强大功能和简易的可视化操作深受广大质量学者和统计专家的青睐。Minitab 1972 年成立于美国的宾夕法尼亚州州立大学。

目前从新兴企业到全球 500 强公司大都选择 Minitab 作为改善流程、提高质量的首选软件，它是现代质量管理统计的领先者，是六西格玛管理中必需的统计分析软件，更是持续质量改进的良好工具软件，它的核心功能就是进行数据分析、图形分析以及趋势预测。本文使用 Minitab 16.0 版作为分析辅助工具，主要使用里面的回归分析模块。

3.3.2. 贝叶斯犯罪挖掘算法

$P(X|C_i)$被称为先验概率，是根据以往经验获得的在已知状态C_i的情况下X_i的分布。对于企业创新风险。先验概率就是在以往经历的失败或者成功的创新活动中，各风险指标的分布情况。在风险指标为离散值的情况下，先验概率可以通过统计历史数据的方法得到。

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{\sum_{j=1}^{n} P(X|C_j)P(C_j)} \tag{1}$$

贝叶斯定理：由总体信息($P(X|C_i)$)、先验信息($P(C_i)$)和样本信息($P(X)$)得出，后验概率($P(C_i|X)$)。

贝叶斯犯罪挖掘算法步骤：

第一步：已知总体信息、样本信息和先验信息概率

第二步：利用贝叶斯公式转换成后验概率

第三步：根据后验概率大小进行决策分类

3.3.3. 决策树

决策树[11] [14]是以树结构形式对目标属性(或类)进行分类的分类器。要分类的观察数据(或物品)是由属性及其目标值组成的。树的节点可以是：1) 决策节点，在这些节点中一个简单属性值被测试来决定应用哪一个子树；2) 叶子节点指示目标属性的值。

4. 基于 Minitab、贝叶斯与决策树的犯罪数据挖掘应用研究

根据上节所述的大数据挖掘相关方法，为简化计算，采用随机抽样的方法，从我们收集的某民族地区的犯罪数据中抽样选取如表 1 的数据作为分析研究的样本数据。

4.1. 基于 Minitab 对犯罪数据的挖掘

Minitab 软件是由 1972 年成立于美国宾夕法尼亚大学的 Minitab Inc.公司开发的。目前从新兴企业到全球500强公司大都选择Minitab作为改善流程、提高质量的首选软件，它是现代质量管理统计的领先者，是六西格玛管理中必需的统计分析软件，更是持续质量改进的良好工具软件，它的核心功能就是进行数据分析、图形分析以及趋势预测。本文使用 Minitab 16.0 版作为分析辅助工具，主要使用里面的回归分析模块。

Table 1. Sample data of a national police crime analysis
表 1. 某民族公安犯罪行为分析样本数据

年龄	经济状况	文化程度	正当职业	犯罪记录	特长	常驻人口	犯罪程度
20~30	中等	初中	无	无	有	是	较轻
>40	差	小学	无	有	有	是	严重
30~40	差	初中	无	有	有	是	严重
20~30	中等	高中	有	无	无	是	较轻
>40	差	小学	有	有	无	否	严重
30~40	差	初中	有	有	无	是	较轻
20~30	中等	初中	无	有	有	否	较轻
20~30	差	高中	无	无	有	否	严重
30~40	中等	高中	有	无	无	是	较轻
20~30	中等	初中	有	有	无	是	严重
20~30	差	高中	无	有	有	否	严重
>40	差	初中	无	无	无	是	较轻
20~30	差	高中	有	无	无	否	较轻
20~30	差	高中	有	无	无	否	较轻
20~30	中等	高中	有	有	无	是	较轻

基于单因素试验设计，对优化工艺条件进行了一系列的试验，在最陡爬坡试验的基础上，根据Box-Benhnken 的中心组合试验设计原理。利用 Design-Expert8.0.6 进行响应面法优化。影响某民族地区犯罪行为的 7 个主要因素：年龄(x_1)，经济状况(x_2)，文化程度(x_3)，正当职业(x_4)，犯罪记录(x_5)，特长(x_6)，常驻人口(x_7)，考察目标为犯罪程度(y)，试验因素水平安排以及根据以上水平编码设计试验表格并检测响应值结果见表 2。

Table 2. Crime data coding table for a certain ethnic area
表 2. 某民族地区犯罪数据编码表

年龄	经济状况	文化程度	职业	犯罪记录	特长	常驻	犯罪程度
x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
3	2	1	1	1	2	2	2
2	2	2	1	1	2	1	1
1	1	2	2	1	1	2	1
1	2	3	2	2	1	2	2
2	1	3	1	2	2	1	1
1	1	2	1	1	2	1	2
1	2	3	2	1	1	2	2
2	2	2	2	2	2	1	1
1	2	3	1	2	2	2	1

对表 2 数据进行二次多元回归拟合，可求出影响因素的一次效应、二次效应以及交互效应的关联方程并可绘制出响应面图。该模型通过二阶经验模型对变量的响应行为进行表征，即：

$$Y=\beta_0+\sum_{i=1}^{7}\beta_i x_i+\sum_{i=1}^{7}\beta_{ii}x_i^2+\sum_{i=1}^{2}\sum_{j=1}^{7}\beta_{ij}x_i x_j \tag{2}$$

式中：y 代表响应值；β_0、β_i、β_{ii} 表示偏移项、线性偏移和二阶偏移系数；β_{ij} 是交互效应系数；x_i ——各因素的编码值。以 x_1，x_2，x_3，x_4，x_5，x_6，x_7 为自变量，y 为响应值，利用 Minitab 16.0 进行响应曲面回归分析得如表 3：y 的回归系数和表 4：y 的方差分析：

Table 3. Estimated regression coefficients for y
表 3. y 的估计回归系数

项	系数	系数标准误	T	P
常量	6.50000	0.000000	*	*
x_1	0.50000	0.000000	*	*
x_2	−1.00000	0.000000	*	*
x_3	−1.00000	0.000000	*	*
x_4	−3.00000	0.000000	*	*
x_5	−0.00000	0.000000	*	*
x_6	0.00000	0.000000	*	*
x_7	−0.50000	0.000000	*	*
$x_1 * x_1$	0.00000	0.000000	*	*
$x_1 * x_2$	−1.00000	0.000000	*	*
$x_1 * x_4$	−0.00000	0.000000	*	*
$x_1 * x_7$	0.50000	0.000000	*	*
$x_2 * x_4$	2.00000	0.000000	*	*

$S=0, \text{PRESS}=*$

$R-Sq=100.00\%, R-Sq(\text{预测})=*\%R-Sq(\text{调整})=100.00\%$

Table 4. Analysis of variance of y
表 4. y 的方差分析

来源	自由度	Seq SS	Adj SS	Adj MS	F	P
回归	12	3.60000	3.60000	0.300000	*	*
线性	7	1.64712	2.12073	0.302961	*	*
x_1	1	0.20417	0.01481	0.014815	*	*
x_2	1	0.35986	0.01538	0.015385	*	*
x_3	1	0.52484	0.50000	0.500000	*	*
x_4	1	0.07723	0.40909	0.409091	*	*
x_5	1	0.14417	0.00000	0.000000	*	*
x_6	1	0.33619	0.00000	0.000000	*	*
x_7	1	0.00065	0.02128	0.021277	*	*
平方	1	0.17982	0.00000	0.000000	*	*
$x_1 * x_1$	1	0.17982	0.00000	0.000000	*	*
交互作用	4	1.77306	1.77306	0.443265	*	*
$x_1 * x_2$	1	0.18865	0.06452	0.064516	*	*
$x_1 * x_4$	1	0.21153	0.00000	0.000000	*	*
$x_1 * x_7$	1	0.57288	0.06667	0.066667	*	*

Continued

$x_2 * x_4$	1	0.80000	0.80000	0.800000	*	*
残差误差	2	0.00000	0.00000	0.000000		
纯误差	2	0.00000	0.00000	0.000000		
合计	14	3.60000				

从而，得响应面方程得：

$$y = 6.5 + 0.5x_1 - x_2 - x_3 - 3x_4 - 0.5x_7 - x_1x_2 + 0.5x_1x_7 + 2x_2x_4 \quad (3)$$

最后，由(3)，利用 Minitab 工具得出的模型，采用 C#设计实现了如下可视化界面(图 1，图 2)。

Figure 1. Crime prediction based on Minitab
图 1. 基于 Minitab 的犯罪预测

Figure 2. Bayesian-based crime prediction
图 2. 基于贝叶斯的犯罪预测

4.2. 基于贝叶斯方法的犯罪数据挖掘应用

由上表可知，将对犯罪程度进行分类的过程归结为典型的两类问题，根据以上的分类规则则可判定任意一个犯罪行为，是较轻，还是严重。

$C = (C_1, C_2)$

其中：C_1 表示，犯罪行为的犯罪程度较轻，C_2 表示犯罪行为的犯罪程度较重。选取上表中的 10 个数据构成的表 5 作为训练数据，另外的数据作为测试数据。

由表 5 得到：$P(C_1=\text{较轻})=0.6$，$P(C_2=\text{严重})=0.4$；

对未知样本 $X=$(年龄为 20~30，经济状况 = 中等，文化程度 = 初中，正当职业 = 无，犯罪记录 = 无，特长 = 有，常驻人口 = 是)，进行分类；

P(年龄 = 20~30|犯罪程度 = 较轻) = 0.5；P(年龄 = 20~30|犯罪程度 = 较重) = 0.74；

P(经济状况 = 中等|犯罪程度 = 较轻) = 0.333；P(经济状况 = 中等|犯罪程度 = 较重) = 0.25；

P(文化程度 = 初中|犯罪程度 = 较轻) = 0.501；P(文化程度 = 初中|犯罪程度 = 较重) = 0.25；

P(正当职业 = 无|犯罪程度 = 较轻) = 0.333；P(正当职业 = 无|犯罪程度 = 较重) = 0.5；

P(犯罪记录 = 无|犯罪程度 = 较轻) = 0.667；P(犯罪记录 = 无|犯罪程度 = 较重) = 0.25；

P(特长 = 有|犯罪程度 = 较轻) = 0.167；P(特长 = 有|犯罪程度 = 较重) = 0.5；

P(常驻人口 = 是|犯罪程度 = 较轻) = 0.5；P(常驻人口 = 是|犯罪程度 = 较重) = 0.25；

利用以上结果，根据公式(1)分母相同，可以不用计算，只比较分子即可，

$$\text{记 } S=\sum_{j=1}^{n}P(X|C_j)P(C_j) \tag{4}$$

可得；

$$P(C_1|X)=\frac{P(X|C_1)P(C_1)}{\sum_{j=1}^{n}P(X|C_j)P(C_j)}=P(X|C_1)P(C_1)/S \tag{5}$$

Table 5. Analysis of criminal behavior in a certain ethnic area
表 5. 某民族地区犯罪行为分析表

年龄	经济状况	文化程度	正当职业	犯罪记录	特长	常驻人口	犯罪程度
>40	差	小学	有	有	无	否	严重
30~40	差	初中	有	有	无	是	较轻
20~30	中等	初中	无	有	有	否	较轻
20~30	差	高中	无	无	有	否	严重
30~40	中等	高中	有	无	无	是	较轻
20~30	中等	初中	有	有	无	是	严重
20~30	差	高中	无	有	有	否	严重
>40	差	初中	无	无	无	是	较轻
20~30	差	高中	有	无	无	否	较轻

所以，$P(C_1|X)=0.5\times0.333\times0.5\times0.333\times0.667\times0.167\times0.5\times0.6=0.000992639$

$P(C_2|X)=0.75\times0.25\times0.25\times0.5\times0.25\times0.5\times0.25\times0.4=0.000292968$

$P(C_1|X)>P(C_2|X)$

分类结果为 C_1，犯罪程度较轻，对照表，是吻合的。

为了方便使用，我们在基于表 5 的情况下，计算出了各条件概率如表 6。最后，利用贝叶斯理论，采用 C#语言程序设计实现了如图 2 所示的界面。

Table 6. Criminal behavior analysis sample data conditional probability table
表 6. 犯罪行为分析样本数据条件概率表

犯罪程度	年龄			经济状况		文化程度		
	20~30	30~40	>40	中等	差	小学	初中	高中
较轻	0.5	0.333	0.167	0.333	0.667	0.001	0.501	0.498
较重	0.74	0.01	0.25	0.25	0.75	0.25	0.25	0.5
犯罪程度	正当职业		犯罪记录		特长		常驻人口	
	有	无	有	无	有	无	是	否
较轻	0.333	0.667	0.333	0.67	0.167	0.833	0.5	0.5
较重	0.5	0.5	0.75	0.25	0.5	0.5	0.25	0.75

4.3. 基于决策树算法的犯罪数据挖掘应用

4.3.1. 选取挖掘目标和采集数据

笔者选择某犯罪行为过程分析数据，希望通过数据挖掘，从中分析出：影响犯罪程度的因素，并得出犯罪程度较轻和犯罪程度严重的分类规则。即，挖掘出犯罪程度。从而为社会治安综合治理以及人员的管控提供决策依据。

4.3.2. 决策树算法的数据挖掘过程

1) 决策树模型的分类原理

a) 决策树的枝干和叶节点是如何生成的？

b) 节点的选取

i) 划分节点时，模型所需的信息量最小

ii) 能生成一个相对简洁的树

c) 基于信息熵的原理

熵是自信息量的数学期望，它是一种信息的度量，属性变化越多，熵越大。

2) 类别熵

根据熵的定义，把类别看作随机变量，并给出相应的计算熵的公式(6)：

$$\mathrm{H}(\mathrm{C})=-\sum_{i=1}^{k}P_{ci}I_{ci}=-\sum_{i=1}^{k}P_{ci}\log_2^{P_{ci}} \tag{6}$$

由本文的数据集，C_1：较轻；C_2：较重；

$P(C_1)=0.6, P(C_2)=0.4$，

$H(C)=H(C_1)+H(C_2)=\left(-0.6\log_2^{0.6}\right)+\left(-0.4\log_2^{0.4}\right)=0.306495+0.366516=0.670311$

3) 条件熵

根据条件熵的定义，及如下公式(7)，可得出第一个变量 X_1：年龄的条件熵。

$$H\left(C\middle|X_1\right)=-\sum_{i=1}^{k}P_{X1i}H\left(C\middle|X_{i1}\right) \tag{7}$$

由数据表可知，$P\left(X_{11}\right)=P$(年龄为 20~30) = 0.5，$P\left(X_{12}\right)=P$(年龄为 30~40) = 0.3，$P\left(X_{13}\right)=P$(年龄 40) = 0.2；

所以，$H\left(C\middle|X_{11}\right)=\left(-0.8\log_2^{0.8}\right)+\left(-0.2\log_2^{0.2}\right)=0.328876+0.178515=0.507391$

$H\left(C\middle|X_{12}\right)=\left(-0.667\log_0^{0.667}\right)+\left(-0.333\log_2^{0.333}\right)=0.270112+0.366171=0.636283$

$H(C|X_{13})=0$

因而，$H(C|\text{年龄})=\sum_{i=1}^{3}P_{X1i}H(C|X_{1i})=0.5\times0.507391+0.3\times0.636383+0.2\times0=0.44458$

4) 信息增益

由类别熵减去条件熵，得到该属性的信息增益。最后根据所有属性的信息增益来确定根节点和其他枝干节点。信息增益由下面的公式(8)给出：

$$\text{Gain}(X)=H(C)-H(C|X) \tag{8}$$

于是，对于上面的年龄属性的信息增益，可计算得：

$\text{Gain}(\text{年龄})=H(C)-H(C|\text{年龄})=0.670311-0.44458=0.225102$

同理，可根据条件熵，得出其他属性的信息增益为：

$\text{Gain}(\text{年龄})=H(C)-H(C|\text{年龄})=0.670311-0.44458=0.225102$

$\text{Gain}(\text{经济状况})=H(C)-H(C|\text{经济状况})=0.670311-0.253696=0.416615$

$\text{Gain}(\text{文化程度})=H(C)-H(C|\text{文化程度})=0.670311-0.4498688=0.2204422$

$\text{Gain}(\text{正当职业})=H(C)-H(C|\text{正当职业})=0.670311-0.588851=0.08146$

$\text{Gain}(\text{犯罪记录})=H(C)-H(C|\text{犯罪记录})=0.670311-0.6408226=0.0294884$

$\text{Gain}(\text{特长})=H(C)-H(C|\text{特长})=0.670311-0.588851=0.08146$

$\text{Gain}(\text{常住人口})=H(C)-H(C|\text{常住人口})=0.670311-0.6098566=0.0604544$

5) 决策树的生成

通过上面计算的信息增益，可得如下的信息增益表 7。

Table 7. Information gain table with economic status
表 7. 含经济状况的信息增益表

属性	年龄	经济状况	文化程度	正当职业	犯罪记录	特长	常驻人口
增益值	0.225102	0.416615	0.220442	0.08146	0.0294884	0.08146	0.0609544

其中，信息增益值最大的属性为经济状况，于是，决策树的根节点，就是经济状况。

去掉经济状况属性，用上面的方法，得到信息增益表为表 8。

Table 8. Information gain table without economic status
表 8. 不含经济状况的信息增益表

属性	年龄	文化程度	正当职业	犯罪记录	特长	常驻人口
增益值	0.230135	0.230135	0.230135	0.0575222	0.230135	0.3633674

其中，信息增益值最大的属性为常驻人口，于是，决策树子树的根节点，就是常驻人口。

重复此步骤，可生成决策树，如图 3。

4.3.3. 决策树挖掘规则应用

根据犯罪数据挖掘决策树，可以直接提取出犯罪预测分类规则：

1) IF 经济状况 = 中等，THEN 犯罪程度 = 较轻；

2) IF 经济状况 = 差 and 常驻人口 = 否，THEN 犯罪程度 = 严重；

3) IF 经济状况 = 差 and 常驻人口 = 是 and 正当职业 = 有，THEN 犯罪程度 = 较轻；

4) IF 经济状况 = 差 and 常驻人口 = 是 and 正当职业 = 无，THEN 犯罪程度 = 严重；

利用决策树方法挖掘犯罪防控的有用信息，根据分类规则找出与研究对象有关联的信息，方便警务

工作者科学决策，并作出相应的警务模式改革。

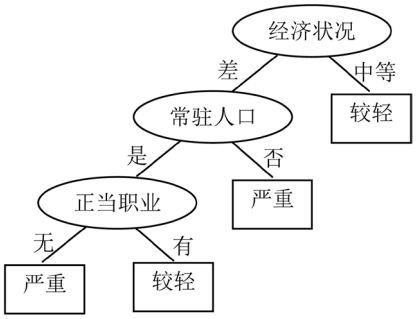

Figure 3. Crime prediction decision tree
图 3. 犯罪预测决策树

4.4. 三种犯罪数据挖掘方法的分析比较

本节，我们采用 Minitab、贝叶斯和决策树三种方法对犯罪数据进行挖掘。Minitab 方法在进行编码的情况下，由软件工具自动计算生成犯罪预测模型，方法直观，计算量小。贝叶斯方法由于要计算先验概率、条件概率及后验概率，使得计算量比 Minitab 大一些。而决策树方法由于要计算总熵、条件熵、信息增益，全部属性的信息增益算完后，再比较大小，才产生决策树的根节点，依次类推，直到生成一棵决策树。通过对三种方法进行预测准确性的测试，得出预测准确性见表 9。

Table 9. Comparison of crime data mining in three methods
表 9. 三种方法的犯罪数据挖掘比较情况表

	复杂性	难易性	计算速度	编程实现难度	预测准确性
Minitab	简单	易	快	低	79.7%
贝叶斯	中	中	中	中	86.3%
决策树	复杂	难	慢	高	82.9%

5. 结论

分类方法是数据挖掘的重要方法之一，其应用领域非常广泛。将基于 Minitab、贝叶斯和决策树理论的分类器应用于犯罪数据挖掘领域，通过对人的初步判断，可对该人的犯罪风险程度进行准确地分类，从而可实现对不同的人员实施不同的管理策略，促进社会的和谐发展，人民的安居乐业。本文采用 Minitab、贝叶斯和决策树三种方法对犯罪数据进行了挖掘。并对 Minitab 和贝叶斯方法得出的模型进行了可视化编程实现。决策树方法的编程实现以及三种方法的融合研究，加强抽样、降维、去噪等预处理技术的研究，提高预测准确性等，将是下一步工作的目标。

基金项目

国家社会科学基金项目(13CFX038)；云南省教育厅科学研究基金项目(2013C188)；云南警官学院教育教学改革项目(2018YJJGB01)。

参考文献

[1] 乔少杰, 唐常杰, 彭京, 等. 基于个性特征仿真邮件分析系统挖掘犯罪网络核心[J]. 计算机学报, 2008, 31(10): 1795-1803.

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

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

[4] 李万彪, 余志, 龚峻峰, 陈锐祥. 基于关系数据模型的犯罪网络挖掘研究[J]. 中山大学学报(自然科学版), 2014, 53(5): 1-7.

[5] 杨莉莉, 杨永川. 基于社会网络的犯罪组织关系挖掘[J]. 计算机工程, 2009, 35(15): 91-93.

[6] Oatley, G.C., Zeleznikow, J. and Ewart, B.W. (2004) Matching and Predicting Crimes. In: Macintosh, A., Ellis, R. and Allen, T., Eds., *Applications and Innovations in Intelligent Systems XII in Proceedings of AI*2004, *The Twenty-Fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence*, Springer, London, 19-32.

[7] Adderley, R.W. (2007) The Use of Data Mining Techniques in Crime Trend Analysis and Offender Profiling. Ph.D. Thesis, University of Wolverhampton, Wolverhampton.

[8] 吴绍兵. 基于贝叶斯网络的刑事犯罪影响因素研究[J]. 计算机与数字工程, 2012, 227(11): 108-111.

[9] 张良均, 陈俊德, 刘名军, 陈荣. 数据挖掘实例分析[M], 北京: 机械工业出版社, 2015.

[10] Han, J.W., Kamber, M. and Pei, J. (2011) Data Mining: Concepts Andtechniques. 3rd Edition, Elsevier Science, Burlington.

[11] Ricci, F., Rokach, L., Shapira, B., Kantor, P.B., 编. 推荐系统: 技术、评估及高效算法[M]. 李艳民, 胡聪, 吴宾, 王雪丽, 等, 译. 北京: 机械工业出版社, 2015: 33-35.

[12] Liu, W.Q. (2014) Modeling Data Quality Control System for Chinese Public Database and Its Empirical Analysis. *Science China: Information Science*, **44**, 835-856.

[13] 米允龙, 米春桥, 刘文奇. 海量数据挖掘过程相关技术研究进展[J]. 计算机科学与探索, 2015, 9(6): 641-659.

[14] Mladenic, D. (1999) Text-Learning and Related Intelligent Agents: A Survey. *IEEE Intelligent System*, **14**, 44-54. https://doi.org/10.1109/5254.784084