浙江理工大学学报(社会科学版),第38卷,第1期,2017年2月
Journal of Zhejiang Sci-Tech University (Social Sciences)
Vol. 38, No. 1, Feb. 2017

DOI:10.3969/j.issn.1673-3851.2017.02.009

基于关联规则的再犯罪特征分析

冯卓慧,冯前进

(浙江警官职业学院信息技术与管理系,杭州 310018)

摘　要:为有效提高罪犯教育改造质量,将基于关联规则的数据挖掘方法引入到再犯罪特征的分析之中,以某监狱罪犯数据库的刑释人员为样本数据,采用Johnson约简算法和Apriori关联算法对其进行属性约简处理和关联规则分析。结果表明:罪名、年龄、文化程度和刑期之间具有较强的正相关,即犯有盗窃罪前科、年龄小、文化程度低、刑期短是再犯罪的主要特征。该方法能揭示潜在的再犯罪规律,对刑罚机关具有参考价值,使其教育改造工作更具有针对性。

关键词:关联规则;再犯罪;数据挖掘;属性约简;Apriori

中图分类号:TP274　　**文献标志码**:A　　**文章编号**:1673-3851 (2017) 01-0057-04

数据挖掘(data mining)是数据库知识发现过程中的一个步骤,是指从大量的数据中通过算法搜索隐藏于其中信息的过程。作为一种通用技术,数据挖掘可以用于任何类型的数据,目前已在国内各个行业和领域都展开了广泛的应用,但在监狱管理中还处于探索和起步阶段。再犯罪既重新违法犯罪,是因犯罪被处以监禁刑或社区矫正,再犯被处以监禁刑或社区矫正的犯罪[1]。一直以来,再犯罪都是世界各国普遍面临的一个社会安全问题,再犯罪也以其特殊的危害性成为我国刑罚机关工作的重点。根据调查,浙江省某监狱2011年重新犯罪人员占全年入监人数的比例高达29.3%[2]。这反映了监狱的教育改造任务还未圆满完成,罪犯改造质量有待提高。要降低重新违法犯罪率,必须先找出刑释解教人员再犯罪的潜在规律和特征因素,才能较为准确的进行预测。根据这一问题,国内很多研究人员对其进行过研究,但主要都是从主观经验或简单统计来进行分析[3-6]。因此,应用数据挖掘技术,对罪犯数据库进行科学有效的分析,从中发现一些潜在的客观规律,有助于提高再犯罪预测的准确性,为刑罚机关在教育改造工作中提供决策支持。

一、关联规则

(一)理论介绍

所谓关联,反映的是一个属性和其他属性之间依赖或关联的知识。例如,在查找英文词典时,可以发现有两个英文单词都能形容关联的含义。一个是相关性relevance,另一个是关联性association,两者都可以用来描述事物之间的关联程度。关联规则能够揭示数据之间的相互关系,而这种关系并没有在数据中直接表现出来[7-8]。

可以通过以下几个参数来描述一个关联规则的属性:

a)支持度($support$)

支持度是D中事务包含$A \cup B$的百分比。项集A的支持度可以定义为:

$$support(A)=\frac{count(A)}{count(B)}.$$

b)置信度($confidence$)

置信度是D中包含A的事务同时也包含B的事务的百分比。计算公式为:

$$confidence(A \rightarrow B)=\frac{support(A,B)}{support(A)}.$$

收稿日期:2016-05-18　　网络出版日期:2017-01-03

基金项目:浙江省教育厅一般科研项目(Y201533854)

作者简介:冯卓慧(1979-),男,湖北荆州人,讲师,硕士,主要从事数据库、数据挖掘、信息安全方面的研究。

c)提升度($lift$)

提升度是置信度与支持度的比值。描述项集 A 的出现对项集 B 的出现有多大的影响,计算公式为:

$$lift(A\rightarrow B)=\frac{confidence(A\rightarrow B)}{support(A)}.$$

当同时满足最小支持度阈值和最小置信度阈值(阈值可根据挖掘需要由用户设定),则认为这些规则是强规则。

(二)Apriori 算法

Apriori 算法是 1994 年由 RakeshAgrawal 和 RamakrishnanSrikant 提出的一种频繁项集挖掘算法,该算法通过前一次找到的频繁项来生成本次的频繁项,目的在于从一个数据集中找出各项之间的关系。该算法的核心思想是:利用 1-频繁项集生成 2-频繁项集,然后根据 2-频繁项集生成 3-频繁项集,以此类推,直到生成所有的频繁项集,最后从频繁项集中找出符合条件的关联规则[9-10]。Apriori 算法通过多次扫描数据库,每次利用候选频繁项集来产生新的频繁项集;算法简单易懂,具有较好的扩展性,但是每次生成候选频繁项集都要扫描数据库,且产生侯选频繁项集时循环生成的组合过多,导致效率低下。

二、分析方法

(一)数据处理

首先检查数据的完整性。一般情况下,罪犯数据库中的缺失值较少,对于这少数部分缺失值可以通过查阅档案补全或直接忽略。其次,将数据中的属性值进行离散化和归类处理。此外,罪犯数据库中基本的属性分类就多达 20 多项,且某些属性值分类也较多,例如罪名(根据《刑法修正九》及两高《关于执行<中华人民共和国刑法>确定罪名的补充规定(六)》的规定,罪名多达 468 个),其中大部分属性值出现频率很小,若直接使用 Apriori 算法进行处理将非常耗时。因此,采用属性约简事先把出现频率低的属性值排除在外,可以有效减少数据集,提高效率。

(二)属性约简

属性约简是在保证系统本身分类能力不变的前提下,删除其中冗余的属性,保留起决定作用的核心属性,它是粗糙集理论中最重要的一个部分。通过知识约简,导出问题的决策或分类规则,其对于研究关联规则的知识发现有着极其重要的意义。粗糙集约简算法主要有基于信息熵、基于可辨识矩阵、基于遗传算法和基于 Johnson 算法等算法。其中基于 Johnson 的约简算法可由用户来决定属性权重,且该算法得出的约简组合只有一组,相比于其他方法而言更加直观。其基本思想是选取出现频率最大的属性加入约简组合,若一个属性出现的频率越高,则它的可分辨能力就越强[11-13]。算法基本步骤如下:

a)令 $S=\{U,R,V,f\}$;$R=C\cup D$,其中 C 和 D 分别为条件属性和决策属性 $C=\cup a_i$,$i=1,2,\cdots,n$;

b)计算可分辨矩阵 M,$M=\{m_{ij}:m_{ij}\neq\varnothing\}$;

c)计算属性 a_i 在 M 中出现的频率 $Count_{ai}(M)$;

d)将 $Max(Count_{ai}(M))$ 的属性记为 a,约简 $RED=RED\cup\{a\}$;

e)删除 M 中所包含全部 a 属性;

f)如果 $M\neq\varnothing$,则转到步骤 c;否则计算结束。

(三)规则提取

将处理好的样本数据进行导入后并转换至矩阵,设置相关的阈值参数,使用 Apriori 算法进行挖掘,最后对生成的规则进行评价。

相关的部分 Python 代码如下:

```
import pandas as pd
from apriori import * #导入函数
import time #导入时间库用来计算用时
filename='crimedata.csv' #导入样本数据
data= pd.read_csv(filename, header = None, dtype=object)
print(u'\n 转换原始数据至矩阵…')
cvt= lambda x : pd.series(1, index = x[pd.notnull(x)]) #转换矩阵的过渡函数
c=map(cvt, data.as_matrix()) #用 map 方式执行
data=pd.DataFrame(c).fillna(0) #实现矩阵转换
del c #删除中间变量节省内存
support=0.3 #最小支持度
confidence=0.8 #最小置信度
lift=1.0 #最小提升度
starttime=time.clock() #计时开始
findrules(data, support, confidence, lift) #开始搜索关联规则
endtime=time.clock() #计时结束
print(u'\n 搜索完成,用时:%0.2f 秒 ' %(end-start))
```

三、实例分析

(一)分析过程

本实例以某男犯监狱 2011 年至 2013 年期间的刑释人员为样本。首先,从基础数据中选取{民族、文化程度、籍贯、婚否、罪名、刑期、出监年龄}作为条件属性,是否再犯作为决策属性。部分条件属性进行了合并(如出监年龄=刑期止日-出生日期)或删除(如删除逮捕机关、逮捕日期、判决机关、判决日期等不相关或相似的属性)。其次,将数据集中的属性值进行离散化和归类处理。例如将文化程度离散化为{1:小学及以下,2:初中,3:高中(中职),4:专科(高职),5:本科及以上};罪名是通过编码来记录的,本身是离散化的,但是为了便于分析还是需要经过一些归类处理(例如:4002、400201、400202、400203分别表示“盗窃”、“惯窃”、“盗窃未遂”和“盗窃预备”,归类后统一用 4002 表示);刑期离散化为{1:36个月以下,2:36～72 个月,3:72～108 个月,4:108个月以上};第一次犯罪出监年龄离散化为{1:25 周岁以下,2:25～35 周岁,3:35～45 周岁,4:45 周岁以上}。

将离散后的数据导入 Rosette 并使用 Johnson 约简算法进行属性约简,得到的约简 R={文化程度、罪名、刑期和出监年龄},分别记为 A, B, C, D。最终的样本数据如表 1 所示。

表 1　经过处理后的部分样本数据

A	B	C	D
1	4002	1	1
3	3003	1	1
1	5701	2	1
1	4002	1	1
2	4003	1	1
2	5123	1	1
1	4001	1	1
2	4002	1	1
1	4002	1	2
3	3003	1	3

(二)分析结果

当 Apriori 算法的最小支持度设置为 0.3,最小置信度设置为 0.8,最小提升度为 1,规则条数为 13,结果如表 2(编码 4002 为盗窃罪)。

表 2　生产的规则结果 1

Id	规则	支持度	置信度	提升度
1	(4002—C1)—A1	0.4226	0.9516	2.1505
2	(D1—4002—C1)—A1	0.3049	0.9505	2.9684
3	4002—A1	0.5147	0.9498	1.7517
4	D1—A1	0.6015	0.9488	1.4974
5	(D1—4002)—A1	0.3441	0.9480	2.6151
6	(D1—C1)—A1	0.4981	0.9455	1.7961
7	C1—A1	0.6973	0.9342	1.2519
8	(D1—4002—A1)—C1	0.3049	0.8859	2.5782
9	(D1—4002)—C1	0.3207	0.8835	2.4372
10	D1—C1	0.5267	0.8309	1.3113
11	(D1—A1)—C1	0.4981	0.8281	1.3774
12	(4002—A1)—C1	0.4226	0.8211	1.5965
13	4002—C1	0.4422	0.8161	1.5051

第 1 条规则解释为:犯有盗窃罪的,刑期短(3年以内),且文化程度低(小学及以下)的罪犯人数占总样本数据的 42%;同时这些刑期短的盗窃犯中,95%的文化程度为小学及以下;提升度(lift)表明该规则比发现同时满足前两项条件(犯有盗窃罪和文化程度低)的情况下更常见。

此外,将最小支持度设置为 0.3,最小置信度设置为 0.6,最小提升度为 1,对罪犯的所有罪名进行单独挖掘,发现结果如表 3 所示。

表 3　生产的规则结果 2

Id	规则	支持度	置信度	提升度
1	4002—4002(*)	0.3642	0.7006	1.2911
2	4002(*)—4002	0.3642	0.6140	1.0324

注:括号内“*”号表示犯罪前科记录中的罪名编号。

结果显示,本次犯罪和犯罪前科记录中都包含盗窃罪的占总样本数据的 36%;在本次犯有盗窃罪的罪犯中,前科记录也包含盗窃罪的占 70%。前科记录中含有盗窃罪的,本次又犯有盗窃罪的占 61%。

(三)规则评价

在分析中使用提升度作为相关性度量后,筛选出了所有的正相关规则,如表 2 中的规则 1,lift=2.1505 表明该规则比发现同时满足前两项条件(犯有盗窃罪和文化程度低)的情况下更常见;同样表 3 中表明这两条规则都不是偶然,它比发现只犯有一次盗窃罪更普遍。此外,表 2 中发现具有一些包含关系的规则,这些规则的后继(RHS)相同,而先导(LHS)为子集关系的规则,如规则 2 的 LHS 包含了规则 1、3、4、5、6、7 的 LHS,且规则 2 的提升度大

于其他规则的提升度,这类规则可以全部合并成规则 2。最后,根据结果可以看出,文化程度低(小学及以下)、刑期短(36 个月以下)、年龄小(25 周岁以下)以及之前犯有盗窃罪的是再犯罪的主要特征,上述结果对之前一些犯罪原因分析的相关文献研究结论进行了进一步的佐证。表 3 表明,盗窃罪具有其多发性和频繁性。针对盗窃的犯罪特点,刑法机关应加强对罪犯的知识技能训练和心理矫治,采用一套科学客观的循证矫正方法来预防和降低实施盗窃犯罪行为再次发生的可能性,提高其回归社会的信心和适应能力[14-15]。

四、结 语

本文基于关联规则研究了再犯罪的特征因素。从分析结果得出,再犯罪因素主要与罪犯的文化程度、罪名、刑期和年龄相关,基本上与之前在主观和经验上来进行分析的文献结果相吻合。目前,我国监狱机关都已经建立罪犯数据库,但是该数据库还是主要用于基本的信息存储和报表统计,缺少对分析和预测等高级功能的支持机制。未来,可以针对罪犯数据库,采用多种挖掘算法进行更深入的挖掘,找出历史数据之间的潜在联系,促使监狱机关做出相应的决策,制订一系列行之有效的改造方法,从而提高罪犯再教育的成功率。

参考文献:

[1] 曾赟. 论再犯罪危险的审查判断标准[J]. 清华法学,2012,8(1):64-77.

[2] 张崇脉. 我国重新犯罪研究的内容分析:以期刊论文为样本[J]. 预防青少年犯罪研究,2015(6):12-22.

[3] 潘开元,李仲林. 北京市监狱管理局在押累犯犯罪原因及矫正对策[J]. 中国司法,2006(4):34-37.

[4] 贺志明,秦志斌,胡赫. 刑释人员重新犯罪的原因分析[J]. 湖南工业职业技术学院学报,2012,12(2):59-61.

[5] 刘宏斌. 当前我国盗窃犯罪的现状及治理[J]. 中国人民公安大学学报(社会科学版),2011,27(4):118-122.

[6] 赵军. 我国犯罪预测及其研究的现状、问题与发展趋势:对"中国知网"的内容分析[J]. 湖南大学学报(社会科学版),2011,25(3):155-160.

[7] 佚名. 关联规则算法概述[J]. 通信企业管理,2005(9):76-77.

[8] 邓灵评. 基于数据挖掘的犯罪行为分析及系统实现[D]. 成都:成都西南交通大学,2014.

[9] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]//Proceeding of the 20th VLDB Conferece. Santiago: Chile, 1994,1215:487-499.

[10] HAN JW, KAMBER M. 数据挖掘:概念与技术[M]. 第 3 版. 北京:机械工业出版社,2012.

[11] PAWLAK Z. Rough Sets: Theoretical Aspects of Reasoning about Data [M]. Dordrecht: Kluwer Academic Publishers,1992.

[12] 徐章艳,杨炳儒,宋威,等. 几种不同属性约简的比较研究[J]. 小型微型计算机系统,2008,29(5):848-853.

[13] 唐卫国,张涛,罗奕,等. 粗糙集属性约简算法综述[J]. 大众科技,2015,17(11):17-19.

[14] 曾赟. 中国监狱罪犯教育改造质量评估研究[J]. 中国法学,2013(3):149-162.

[15] 王锐,丁平,刘伟兵. 关于盗窃惯犯人格特征评估与犯罪对策的研究[J]. 辽宁警专学报,2010(3):76-78.

Analysis on the Features of Recidivism Based on Association Rules

FENG Zhuohui, *FENG Qianjin*

(Department of Information Management, Zhejiang Police Vocational Academy, Hangzhou 310018, China)

Abstract: In order to improve the quality of education reform of criminals effectively, data mining based on association rules was utilized to explore the features of recidivism. Johnson reduction algorithm and Apriori algorithm were applied to analyze partial sample data of released prisoners in the criminal database of a prison. The analysis results show that the charge, age and cultural degree are positively related to the term of penalty. In other words, criminal record of larceny, younger age, low cultural degree and short term of penalty are main features of recidivism. The method of this article can reveal the potential recidivist. Therefore it has reference value for penalty organs and is one of useful tools to improve the quality of correction in prison.

Key words: association rules; recidivism; data mining; attribute reduction; Apriori

(责任编辑:陈和榜)