·实践与应用·

doi：10.15908/j.cnki.cist.2016.03.017

基于 Apriori 算法的重新犯罪关联规则挖掘

汤毅平

（中国电子科技集团公司第二十八研究所 南京 210007）

摘 要：以犯罪人员信息关联关系为重新犯罪数据分析切入点，构建了犯罪次数与犯罪人员信息的数据挖掘模型，以寻找重新犯罪的潜在规律。提出了使用 Apriori 算法对犯罪人员数据进行关联规则挖掘的方法。最后，通过试验验证了其可行性。该方法已成功应用于警用系统。

关键词：重新犯罪；数据挖掘；关联规则；Apriori 算法；R 语言

中图分类号：TP301 **文献标识码**：A **文章编号**：1674-909X(2016)03-0091-05

Recidivism Association Rule Mining Based on Apriori Algorithm

TANG Yiping

(The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China)

Abstract: An association relationship of the crime personnel information is regarded as a cut-in point for recidivism data analysis. A data mining model for the arrest times and crime personnel data is constructed to explore the latency recidivism rule. The method for association rule mining based on Apriori algorithm on the crime personnel data is proposed. Finally, experimental results verify the feasibility of the method. The method has been successfully applied in the police systems.

Key words: recidivism; data mining; association rule; Apriori algorithm; R language

0 引 言

犯罪预测是犯罪学中重要组成部分，犯罪预测需建立在犯罪学知识基础上，而犯罪规律揭示本身就是犯罪预测[1]。公共安全领域中实战需求包括治安防控、反恐维稳、情报研判和案情侦破等。近年来，随着我国社会经济不断发展，犯罪呈高发态势，刑释人员重新犯罪率从原来维持在 8%以下逐渐突破了两位数，有的甚至高达 20%以上，不断攀升[2-3]。

重新犯罪是世界各国普遍面临的复杂社会问题。重新犯罪指原先曾有构成犯罪的严重违法行为的人，无论司法机关是否追究其刑事责任或受到刑罚处罚，后来又有构成犯罪的严重违法行为。重新犯罪率是国际上确认的了解重新犯罪现状的重要指标，指某年度原罪犯被判处刑罚、刑罚执行完毕再次犯罪并被判处刑罚的人数与同年度刑罚执行完毕和赦免人员总数的比例。

我国在重新犯罪领域研究已取得很多进展。例如，文献[2,4]分别对当前中国重新犯罪的状况进行了分析和预测；文献[5]评估了重新犯罪人员的心理危险性；文献[6-7]对未成年群体的重新犯罪现状进行了调查和分析；文献[8]调查了河南省郑州女子监狱女性重新犯罪的情况；文献[9]调查了安徽省女子监狱女性重新犯罪的情况。上述文献从多个角度分析了重新犯罪现象，但研究范围仍集中在犯罪学和法律学的调查分析和理论探究层面，给出的意见和对策均相对抽象，与具体实务工作相去甚远。在当前研究文献中，未涉及使用数据挖掘方法研究重新犯罪问题。

收稿日期：2015-11-18

1 关联规则基本概念

关联分析主要用于在大型数据集中寻找数据间关联关系。这种关联关系以频繁项集和关联规则 2 种形式存在。频繁项集指一起频繁出现的项目的集合。关联规则是假定在 2 个项集间存在强烈的关系，它是表示 2 个项集间的关联度或相关性的规则。关联规则常以 $A \Rightarrow B$ 的形式表示，其中 A 与 B 是 2 个互斥的项集。常用的关联规则度量包括支持度(support)和置信度(confidence)。

1.1 支持度和置信度

设 $I=\{i_1, i_2, \cdots, i_m\}$ 为数据项集合，D 为数据交易库集合，其中每个交易 T 是一个数据项非空子集，即 $T \subset I$；每个交易均包含一个识别编号 TID。设 A 为一个数据项集合，B 为另一个数据项集合，当且仅当 $A \subseteq T$ 时称交易 T 包含 A。一个关联规则具有“$A \subset I, B \subset I$，且 $A \cap B = \varnothing$”形式的蕴含式；规则 $A \Rightarrow B$ 在交易数据集 D 中成立，且具有支持度 s 和置信度 c。这也意味着交易数据集 D 中有 s 比例的交易 T 包含 $A \cup B$ 数据项：

$$\text{support}(A \Rightarrow B) = P(A \cup B) \tag{1}$$

且交易数据集 D 中有 c 比例的交易 T 满足“若包含 A 就包含 B”：

$$\text{confidence}(A \Rightarrow B) = P(B \mid A) = P(A \cup B)/P(A) \tag{2}$$

1.2 频繁项集和关联规则

如果设置取值范围为[0，100%]的最小支持度阈值 min_sup 和最小置信度阈值 min_conf，则当且仅当项集 A 的支持度不小于给定的最小支持度阈值 min_sup，即 $\text{support}(A) \geqslant \text{min\_sup}$ 时，该项集称为频繁项集。若项集 A 与项集 B 的关联支持度 $\text{support}(A \Rightarrow B) \geqslant \text{min\_sup}$ 时，则称 A 与 B 是关联的。

关联规则要求在项集中找到同时符合以下条件的规则：$\text{support}(A \Rightarrow B) > \text{min\_sup}$ 且 $\text{confidence}(A \Rightarrow B) > \text{min\_conf}$。同时满足最小支持度阈值和最小置信度阈值的关联规则称为强规则。支持度和置信度是量化关联分析成功与否的重要指标。

2 关联规则挖掘技术

2.1 数据预处理

在数据挖掘算法工程应用中，需得到准确的模型和高质量数据。数据预处理技术对于数据挖掘至关重要。常见数据预处理过程包括数据清洗、数据变换、数据集成和数据规约等。数据清洗指对目标数据样本进行消除无关数据、填写缺失值、光滑噪声数据以及识别和删除离群点等操作。数据变换指将数据表现形式变换为另一种等价形式，使数据更加规范化，从而提高数据挖掘效率。数据变换主要对数据进行降维处理，消除无效属性，找到真正有用的特征属性，并对属性类型进行正确判断，以便后期数据挖掘的处理和计算。数据集成主要指合并目标数据样本存在于不同数据库间而引起的数据冲突以及不一致等问题的处理操作。数据规约指在挖掘目标的有用特征以及对数据自身内容理解基础上，最大限度地对目标数据样本进行精简处理的过程，主要包括数据样本的参考属性选择和数据抽样的数据处理。

2.2 Apriori 算法

2.2.1 简 介

Apriori 在拉丁文中的意思是“from before，来自从前”。Apriori 算法是数据挖掘领域重要的十大算法之一[10]，该算法由 R. Agrawal 和 R. Srikant 于 1994 年提出[11]，它是一种广度优先的逐层搜索算法，通过对事务计数找出频繁项集，再从中推导出关联规则。实际过程分为 2 个阶段：1) 识别所有满足最小支持度阈值的项集；2) 根据满足最小置信度阈值的这些项集来创建规则。Apriori 算法流程如图 1 所示。

2.2.2 数据要求

Apriori 算法是一种单维、单层和布尔关联规则的数据挖掘算法。Apriori 算法在挖掘关联规则时，只能处理分类型变量，无法处理数值型变量。分类变量指变量值是定性的，表现为互不相容的类别或属性，通常表现为二值、符号和顺序变量 3 种类型。一个二值变量仅取 0 或 1，其中 0 表明(变量所表示的)状态不存在，1 表明相应的状态存在。符号变量是二值变量的推广，可以对 2 个以上状态进行描述。如颜色变量为一个符号变量，可以表示 5 种状态，即红、绿、篮、粉红和黄色。离散顺序变量与符号变量相似，不同点是(对应 N 个状态的)N 个顺序值具有一定顺序含义。

2.2.3 注意事项

Apriori 算法的数据存储要求以项集方式存在，

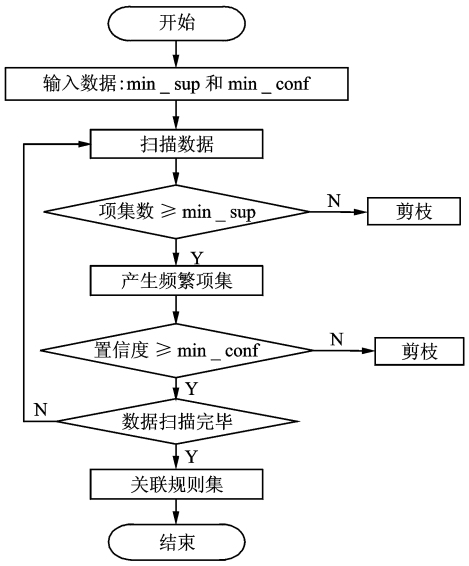

图 1　Apriori 算法流程

项集存储方式有事务表和事实表 2 种。在使用关联规则进行建模时，用户需预先设定好最小支持度和最小置信度。关联规则挖掘算法目标是从事务数据集中找出满足最小支持度和最小置信度的强关联规则，强关联规则对应的项集必定是频繁项集，频繁项集导出的关联规则的置信度又可由频繁项集的支持度算出（根据置信度的定义），根据频繁项集和最小置信度最终产生关联规则。满足最小支持度和最小置信度的关联规则形成了关联规则集。关联规则建模基本模型如图 2 所示。

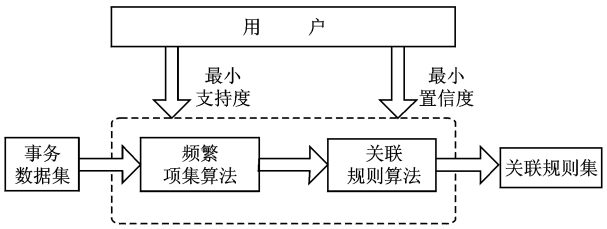

图 2　关联规则建模基本模型

2.3　主要技术与创新点

本文主要研究犯罪人员信息中有应用价值的规律挖掘。犯罪人员信息分为基础和犯罪信息。基础信息包括性别、出生日期、婚姻状况和文化程度等；犯罪信息包括犯罪人员类别和犯罪次数等特征属性。本文采用数据挖掘方法定量分析重新犯罪数据，针对数据复杂多变的特性，通过收集和处理数据，分析和归纳问题，将犯罪数据转化成数据挖掘问题，得出重要相关因子，开创性提出利用 Apriori 算法挖掘犯罪人员数据中隐含的规律，利用犯罪人员信息的 6 类属性推断具有哪些特征的犯罪人员更具有再次犯罪的可能性，揭示重新犯罪问题内在规律，从而辅助办案人员明确破案方向，探求预防控制的方法和措施，达到降低重新犯罪率的目的。

在工程应用方面，数据挖掘的大部分算法本身已经成熟，如何将需解决的实际问题转化成数学问题、采用何种方法进行分析挖掘成为问题的难点。本文为重新犯罪领域的规律挖掘问题提供了一种可行的解决方案，并成功应用于警用系统工程，取得了较好效果。本文采用 R 语言描述整个方法的设计和分析过程。

3　关联规则挖掘在犯罪人员数据中应用

基于现实犯罪人员数据的敏感性，采用 R 语言[12]仿真生成了 50 000 条犯罪数据，保留了与真实数据相同的特征属性。通过考虑犯罪人员的特征属性与犯罪次数的关系，尝试挖掘其中潜在规律，从而推测犯罪人员再次犯罪时可能具有的特征。

3.1　数据预处理

数据分析前需先对数据进行预处理，具体步骤如下：

1）数据选择：根据文献[13]中方法对犯罪数据性质进行分类，从犯罪人员数据中选择了与分析主题相关的 6 项数据：性别、出生日期、婚姻状况、文化程度、犯罪人员类别和犯罪次数。本文考虑上述 6 项数据的关系。

2）数据抽取：从数据库中抽取需分析的犯罪人员数据，存储至数据挖掘的数据表。

3）数据清洗：数据中如有缺失值，则认为不可分析。为使分析结果有意义，有 2 种方案：（1）在分析数据前删除这些缺失值；（2）用平均值或中位数等值填入缺失部分。本文选用方案（1）。

4）数据变换：在数据分析中将分类变量分为无序和有序 2 种。将上述特征属性按以下方式转化为适合分析的格式：

（1）将性别（Gender）转化成二值分类变量：male 表明男，female 表明女；

（2）将年龄（Age）转化为从小到大的有序分类变量：少年（juvenile）＜青年（youth）＜ 壮年（middle aged1）＜ 中年（middle aged2）＜ 老年（old age）；

（3）将婚姻状况（Marriage）转化成无序分类变量：已婚（married）、未婚（unmarried）、离异（devoiced）和其他（others），它们之间同等存在，无先后关系；

（4）将文化程度（Education）转化成从小到大

的有序分类变量：文盲(illiteracy)＜小学(elementary)＜初中(junior)＜高中(senior)＜大学(graduate)＜研究生(postgraduate)；

(5) 将犯罪人员类别(Criminal_record)转化为无序分类变量，可将犯罪人员类别划分为：刑满释放人员(x_1)，在逃人员(x_2)，涉毒人员(x_3)，涉稳人员(x_4)，涉恐人员(x_5)，肇事肇祸精神病人(x_6)，它们之间无大小关系；

(6) 将犯罪次数(Crime_times)转化为无序分类变量，有 1 次犯罪记录显示为 No，有 2 次以上犯罪记录显示为 Yes。

3.2 挖掘过程

本文使用 R 语言中的 arules 包来进行关联规则挖掘。由 R 语言生成 50 000 条数据。

为了直观呈现上述数据，用 R 语言命令生成描绘相对频繁项集的柱状图，如图 3 所示，其中横坐标表明在本次挖掘过程中所有项集的可能值，纵坐标表明每个项集的可能值在数据中相对比例。

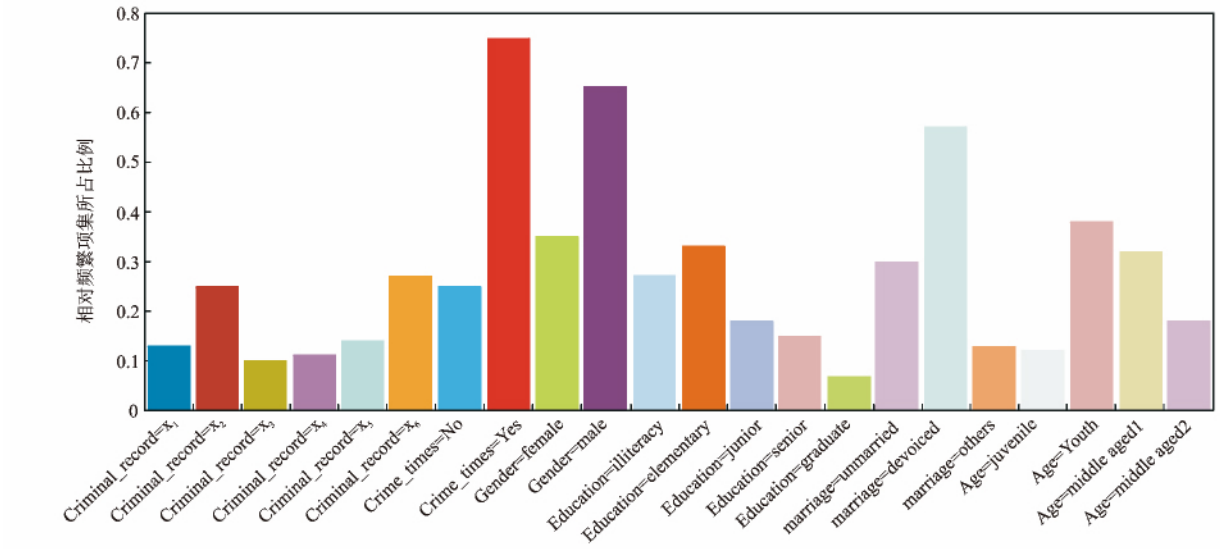

图 3 相对频繁项集柱状图

调用 Apriori 算法，计算满足 2 次以上犯罪记录条件的关联规则，为防止产生过多规则，设置了条件和排序来限制关联规则数量，本文取支持度为 1%，置信度为 60%，关联规则的长度≤5，关联规则表如表 1 所示。

表 1 关联规则表

编号	关联规则		支持度	置信度
1	{Criminal_record=x_2, Gender=male, Education=elementary, Age=middle aged1}	⇒ {Crime_times=Yes}	0.010 2	0.805
2	{Criminal_record=x_3, Age=middle aged2 }	⇒ {Crime_times=Yes}	0.012 6	0.778
3	{Criminal_record=$x1$, Gender=male, Education=junior}	⇒ {Crime_times=Yes}	0.012 6	0.767
⋮				

3.3 关联规则解读

关联规则需进行合理解读。解读表 1 中关联规则，得到以下结论(基于仿真数据的解读结果不能作为现实破案参考)：

1) 具有小学文化的壮年和男性在逃人员，再次犯罪可能性较大，支持度为 1.02%，置信度为 80.5%；

2) 中年涉毒人员，再次犯罪可能性大，支持度为 1.26%，置信度为 77.8%；

3) 具有初中文化的男性刑满释放人员，再次犯罪可能性大，支持度为 1.26%，置信度为 76.7%。

类似规律很多，数据分析人员需对计算出来的关联规则进一步筛选，以便找出最符合预测需求的规则。

4 结束语

本文从公共安全的实战需求出发，基于犯罪学知识，提出了将 Apriori 算法方法应用于重新犯罪规律挖掘的方案，通过仿真犯罪数据详细介绍了 Apriori 算法应用。本文方法已成功应用于实际开发的警用系统，根据实际数据计算出的规则符合人们经验性结论，并且比经验性结论更具体，对辅助治安防控具有现实意义。

需要注意的是，不同于办案人员的经验和知识，

由于关联分析得到的规则完全基于数据，数据的优劣程度决定了所得规则的好坏，因此在数据预处理阶段，确保数据的真实性和高质量是得到正确关联规则的前提。作为一种无监督的学习算法，关联分析能够从无任何关于模式的先验知识的大型数据库中提取知识。关联规则在解决大数据问题上具有广阔前景。后续研究将继续跟踪数据挖掘的最新进展，结合公共安全的实战需求，运用更多的数据挖掘方法来解决公共安全领域实际问题。

参考文献(References)：

[1] 张远煌. 犯罪学[M]. 北京：中国人民大学出版社，2011.

[2] 李均仁. 中国重新犯罪研究[M]. 北京：法律出版社，1992.

[3] 王志强. 重新犯罪实证研究[J]. 中国人民公安大学学报(社会科学版)，2010(5)：38-50.
WANG Zhiqiang. A positive study of the repeat an offense[J]. Journal of Chinese People's Public Security University(Social Sciences Edition)，2010(5)：38-50. (in Chinese)

[4] 丛梅. 我国重新犯罪现状与发展趋势研究[J]. 社会工作(学术版)，2011(12)：91-93.
CONG Mei. Study on the present situation and tendency of recidivism in China[J]. Journal of Social Work，2011(12)：91-93. (in Chinese)

[5] 顾宏翔，李伟，富国徽. 罪犯重新犯罪心理危险性评估分析研究[J]. 江苏警官学院学报，2013，28(2)：52-56.
GU Hongxiang，LI Wei，FU Guohui. Psychological risk assessment of the reoffending criminals[J]. Journal of Jiangsu Police Officer College，2013，28(2)：52-56. (in Chinese)

[6] 丛梅. 未成年人重新犯罪实证研究[J]. 河南警察学院学报，2011，20(5)：21-27.
CONG Mei. Empirical research on juvenile recidivism[J]. Journal of Henan Police College，2011，20(5)：21-27. (in Chinese)

[7] 朱妙，李振武，张世欣. 关于上海市未成年人重新犯罪情况的调研报告[J]. 上海公安高等专科学校学报，2014，24(3)：31-39.
ZHU Miao，LI Zhenwu，ZHANG Shixin. Investigation report on the circumstances of minors of recommitting crime of Shanghai municipality[J]. Journal of Shanghai Police College，2014，24(3)：31-39. (in Chinese)

[8] 谷世清，鲁德浩，王凤芹，等. 对刑释女性重新犯罪的调查与思考：以河南省郑州女子监狱为例[J]. 决策探索(下半月)，2011(7 下)：34.
GU Shiqing，LU Dehao，WANG Fengqin，et al. Investigation and consideration of the female recidivism[J]. Policy Research & Exploration，2011(7 下)：34. (in Chinese)

[9] 马西艳. 女性重新犯罪问题调查[D]. 合肥：安徽大学，2013.

[10] WU X，KUMAR V. The top ten algorithms in data mining[M]. [S. l.]：Chapman & Hall/CRC，2009.

[11] AGRAWAL R，SRIKANT R. Fast algorithms for mining association rules in large databases[C]// Proceedings of the 20th International Conference on Very Large Data Bases. Santiago，Chile：[s. n.]，1994.

[12] R Core Team. R：a language and environment for statistical computing[CP/OL]. Vienna：R Foundation For Statistical Computing. [2015-11-05]. http：// www. R-project. org/.

[13] 朱明. 犯罪数据的性质与整理[J]. 湖北警官学院学报，2012(12)：131-133.
ZHU Ming. Properties and arrangement of crime data[J]. Journal of Hubei University of Police，2012(12)：131-133. (in Chinese)

作者简介：

汤毅平，男(1981—)，博士，工程师，研究方向为模型选择、因果推论和数据挖掘。

(本文编辑：李素华)