昆明学院学报　2018，40（6）：96～99，116　ISSN 1674－5639　CN 53－1211/G4
Journal of Kunming University

大数据背景下 Visual FoxPro 实现关联规则挖掘算法研究

李佐军
（滇西科技师范学院 信息工程学院，云南 临沧 677000）

摘要：随着信息技术的快速发展，各行业都在使用数字化设备采集和管理用户活动信息，且信息量呈现出爆炸式增长，当今世界也随之进入了大数据时代．在大数据背景下，许多管理信息系统仍在使用 Visual FoxPro 收集和管理数据，此外，尚有许多平台收集的数据没有被利用，而处于闲置状态．因此，为把这些闲置的数据利用起来为各行业服务，对使用 Visual FoxPro 实现关联规则挖掘算法进行分析研究．结果显示，自编程序实现的关联规则挖掘结果与手工推导结果完全一致，从而验证了算法的正确性．

关键词：大数据；关联规则；Visual FoxPro；Apriori 算法
中图分类号：TP311.3　**文献标识码**：A　**文章编号**：1674－5639（2018）06－0096－04
DOI：10.14091/j.cnki.kmxyxb.2018.06.019

Research on Visual FoxPro Implementation of Association Rules Mining Algorithm in Big Data Background

LI Zuojun
(College of Information Science & Engneering, Dianxi Science and Technology Normal University, Lincang, Yunnan, China　677000)

Abstract: With the development of computer and information technology, all industries are using digital devices to collect and manage user activity information and the amount of information collected increases rapidly, along with which the world enters the "big data" era. In the context of the big data era, Visual FoxPro is still used by many management information systems to collect and manage the data which are not put into use but in an idle state. In order to utilize the idle data for the management and development of various industries, Visual FoxPro is used to implement the association rule mining algorithm. The results of association rules mining implemented by self-programming are completely consistent with the results of manual derivation, which verifies the correctness of the algorithm.
Key words: Big Data; association rules; Visual FoxPro; Apriori algorithm

2008 年 9 月 Nature 杂志首次推出大数据（Big Data）专刊之后，引发了学术界对大数据的广泛关注，与此同时，各大研究机构及学术期刊也纷纷推出了自己的大数据研究计划和学术专栏[1]．但是大数据作为信息社会发展的一个新生事物，尚处在逐渐被人们认识及应用的过程，且各界对大数据的解释各有侧重，无统一的定义．麦肯锡全球研究所指出："大数据是规模超过了传统数据库软件获取、存储等能力的数据集，不是大于某个特定数据量的数据．因为随着科学技术的发展，大数据标准也会提高，不同行业也会有变化"[2]．因此，大数据不是一个准确的术语，而是对各种数据永不休止积聚的一种表现．IBM 公司认为，大数据具有"4V"特征，即数量庞大（Volume）、种类繁多（Variety）、变化速度快（Velocity）和真实性高（Veracity）[3]．

20 世纪 90 年代以来，作为大数据处理手段之一的数据挖掘技术得到了迅速发展，其定义也曾多次改动．而数据挖掘是一门综合性、多学科交叉的学科，涉及机器学、数理统计分析、神经网络学、数据库研究、模式分析、粗糙集和模糊数学等相关技术[4]．此外，数据挖掘要经过数据

收稿日期：2018－07－29
作者简介：李佐军（1980—），男，云南凤庆人，副教授，硕士，主要从事数据挖掘及大数据应用技术研究．

采集、数据预处理、数据分析、结果表示等一系列过程[5]，如图 1 所示.

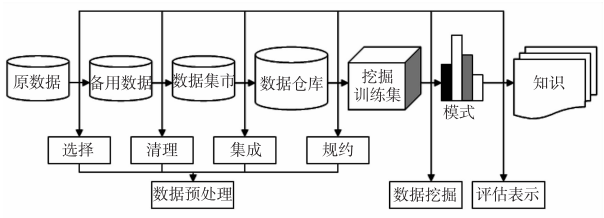

图1　数挖掘的基本过程

在大数据应用盛行的今天，由于许多大数据处理技术是在数据挖掘基础上发展起来的，因此数据挖掘仍是大数据处理的重要技术之一，换言之，研究数据挖掘技术可为更好地使用大数据打下基础. Visual FoxPro 是美国微软公司推出的数据库管理和可视化编程工具，由于其功能强大和容易使用，因此应用越来越广泛. 目前，许多领域运行的管理信息系统仍然将 Visual FoxPro 作为数据管理平台，此外，还有许多平台收集的数据没有被利用起来，成为了闲置的大数据，因此，研究 Visual FoxPro 实现关联规则挖掘算法就是为了更好地利用这些闲置的大数据，并通过关联规则挖掘算法找出它们的发展规律，以便为各行业管理和发展服务.

1　关联规则挖掘简述

1.1　关联规则挖掘的基本概念

关联规则挖掘是由 Agrawal 等人在 1993 年提出的数据挖掘技术，其目的是在数据库中发现两个或者两个以上数据项的关联关系，它是一种实用的、简单的分析规则方法，主要用于查找海量数据中的数据相关联性，发现某些事物同时出现的概率[6]. 用“支持度（Sup）”和“可信度（Conf）”两个阈值来删除无用或作用很小的规则. 为进一步探讨关联规则挖掘算法，下面介绍其相关定义[7]：

定义 1（k-项集）　设 $I=\{I_1, I_2, \cdots, I_k, \cdots, I_m\}$ 是 m 个不同项目或属性的集合，每个 $I_k(k=1, 2, \cdots, m)$ 叫做数据项（Item），数据项的集合 I 叫做 k-项集（k-Itemset）.

定义 2（事务）　指项集 I 上的子集，即 $T\subseteq I$，用 TID 表示每个事务，许多不同的事务构成全体事务集 D（也称事务数据库）.

定义 3（关联规则）　如果项集 X 在某个事务中出现，则必然会使项集 Y 亦在同个事务中出现，表示为 R：$X\Rightarrow Y$，其中 $X\subset I$，$Y\subset I$ 且 $X\cap Y=\varphi$，X 是先决条件，Y 是结果.

定义 4（支持度计数和支持度）　对于关联规则 R：$X\Rightarrow Y$，项集 X 的事务数叫做项集 X 的支持度计数，使用表达式 $count(X)$ 表示数据项集 X 的支持度计数，用表达式 $\sup port(X\Rightarrow Y)=P(X\cup Y)=\frac{count(X)}{N}\times 100\%$ 表示项集 X 的支持度，N 表示事务总数. 支持度反映项集 X 在所有事务集中出现的频率.

定义 5［最小支持度（支持度阈值）与频繁项集）］　在关联规则（$X\Rightarrow Y$）中，最小支持度（min_sup）就是产生关联规则要求数据项必须满足的最小支持阈值，是项集在挖掘统计分析下的最低重要程度. 频繁项集就是支持度大于最小支持度值的数据项集，将所有的频繁 k-项集记作 L_k，反之则是非频繁项集.

定义 6（置信度）　在关联规则（$X\Rightarrow Y$）中，将确定 Y 在包含 X 的事务中出现频繁程度叫做关联规则（$X\Rightarrow Y$）的置信度，用公式 Confidence $(X\Rightarrow Y)=P(Y\mid X)=\frac{\sup port(X\cup Y)}{\sup port(X)}=\frac{\sup port-count(X\cup Y)}{\sup port-count(X)}\times 100\%$ 计算.

定义 7［最小置信度（置信度阈值）］　在关联规则（$X\Rightarrow Y$）中，关联规则必须满足的最小可信度值，称为最小置信度（min_conf），其是关联规则最低可靠程度.

1.2　关联规则挖掘的基本算法：Apriori 算法

Apriori 算法是最经典的关联规则挖掘算法，其首先挖出所有的频繁项集，再由频繁项集产生关联规则. Apriori 算法是一种阶段频繁思想的算法，其核心是采用递推计算方法找出频繁项集的过程，利用频繁项集 L_{k-1} 来生成频繁项集 L_k，包含连接与剪枝两个步骤[8].

1）连接步：为了找出频繁项集 L_k，通过 L_{k-1} 与自己相连接生成候选 k-项集的集合 C_k，$L_i[j]$ 表示 L_i 的第 j 项.

2）剪枝步：候选项集 C_k 是项集 L_k 的超集，所有频繁 k-项集都在候选项集 C_k 内，扫描事务数据库 D，计算候选项在候选项集 C_k 中出现次数，最终根据最小支持度计算确定频繁项集 L_k.

1.3　关联规则挖掘算法案例分析

为进一步理解 Apriori 算法精髓，举例计算出

表 1 事务数据库 D 的频繁项集.

表 1 事务数据库 D

TID	项 ID 的列表
T100	I1，I2，I5
T200	I2，I4
T300	I2，I3
T400	I1，I2，I4
T500	I1，I3
T600	I2，I3
T700	I1，I3
T800	I1，I2，I3，I5
T900	I1，I2，I3

1）第 1 次迭代，扫描事务数据库 D 中的每个项生成候选 1-项集 C_1，并统计出每个事务项出现的次数，假设最小支持计数阈值为 2（即是 $min_sup = 2/9 \approx 22\%$），筛选出频繁 1-项集 L_1.

2）把频繁 1-项集 L_1 与自己相连接生成候选 2-项集C_2，扫描事务数据库 D 中的每个事务，统计出 C_2 中每个候选项集的支持计数，筛选出当最小支持计数阈值为 2（即是 $min_ sup = 2/9 \approx 22\%$）时的频繁 2-项集 L_2.

3）采用同样方法，生成候选 3-项集 C_3，并筛选出频繁 3-项集 L_3.

4）频繁 3-项集 L_3 与自己连接生成候选 4-项集 C_4，扫描事务数据库 D 筛选出 $C_4 = \varphi$，到此算法终止，找出了如图 2 所示的所有频繁项集.

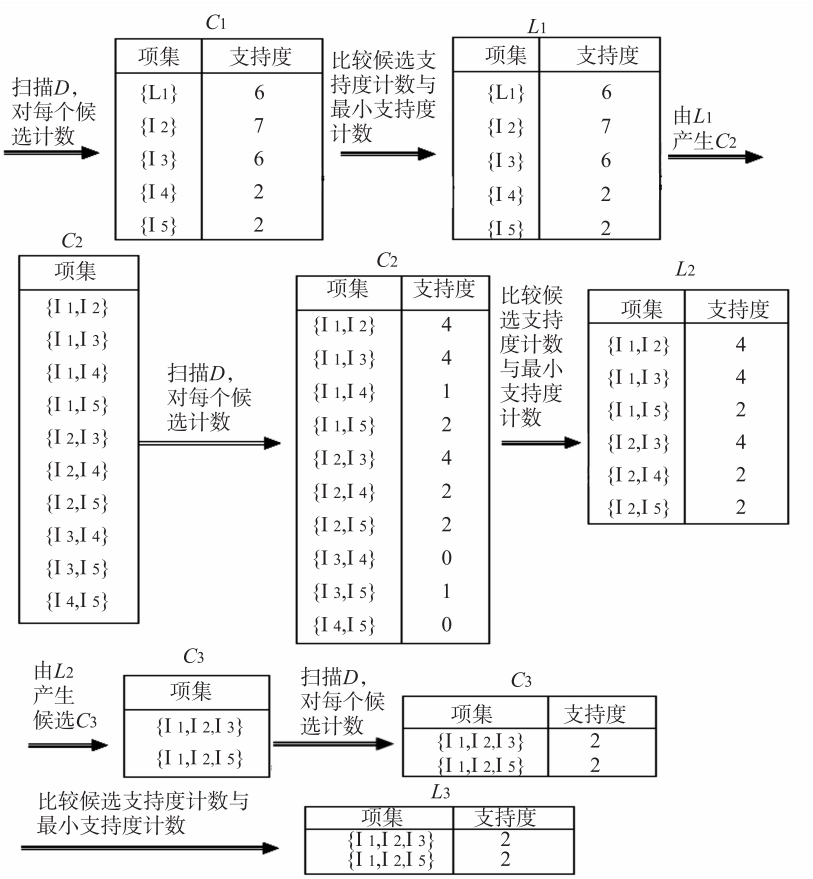

图2 候选项集和频繁项集的产生过程

2 Visual FoxPro 实现关联规则挖掘算法

2.1 关联规则挖掘算法基本流程

根据“1.2 关联规则挖掘的基本算法”可知，关联规则挖掘分为两个步骤：第 1 步生成候选项集，第 2 步筛选出频繁项集. 算法基本流程如图 3 所示.

2.2 Visual FoxPro 实现关联规则挖掘算法及正确性验证

根据图 3 使用 Visual FoxPro 语言实现“1.3 关联规则挖掘算法案例分析”的候选项集和频繁项集的产生过程，首先，将表 1 事务数据库 D 转换

为表 2 关系数据表，其中“Y”表示对应事务出现，否则为不出现.

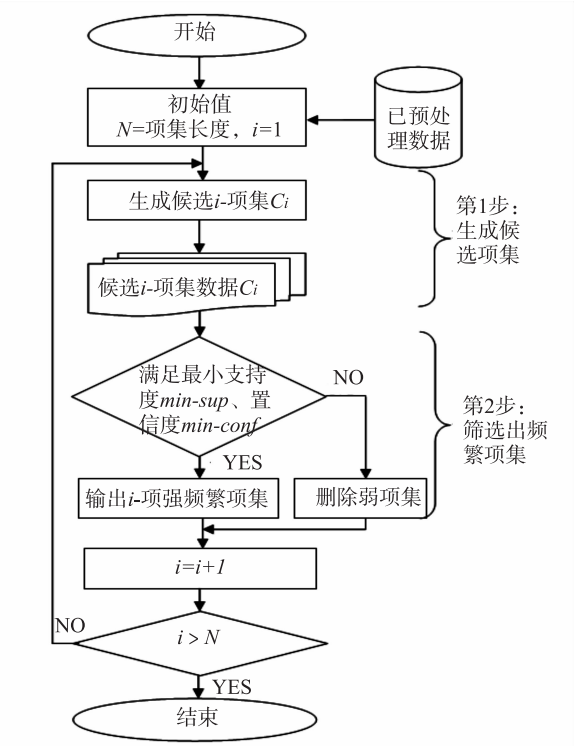

图3　关联挖掘算法流程

表 2　事务关系数据库结构

TID	I1	I2	I3	I4	I5
T100	Y	Y			Y
T200		Y		Y	
T300		Y	Y		
T400	Y	Y		Y	
T500	Y		Y		
T600		Y	Y		
T700	Y		Y		
T800	Y	Y	Y		Y
T900	Y	Y	Y		

其次，将表 2 事务关系数据库使用 Visual FoxPro实现，同时编写程序计算频繁项集，具体见算法 1，运行结果如图 4 所示.

算法 1　Visual FoxPro 实现关联规则挖掘算法伪代码

```
** 初始化
AFieldcount = 项集长度
Min_sup = 2 && 最小支持度计算
nCount = 0 && 用于记录项集条数
FOR i = 1 TO AFieldcount
    stri = stri + 'i 项集{' + Fieldai + '}' && 生成候选 i 项集 Ci
    S = (SELECT COUNT (TID) FROM DATA WHERE DATA. &Fieldai = " Y") && 计算候选 i 项集的支持度
  IF S > = Min_sup && 赛选出频繁项集
    stri = stri + 'i 项集 {' + Fieldai + '}' + space (2) + '支持度计数:' + S
    nCount = nCount + 1
  ENDIF
ENDFOR
** 显示项集
thisform. edit1. value = stri && 显示项集
thisform. label6. caption = '共' + nCount + '个项集!' && 显示项集条数
```

图4　计算频繁项集

最后，从图 4 运行结果与图 2 推导结果比较可看出，当最小支持度计算为 2 时，自编程序计算结果与手工推导结果完全一致，由此可知，自编程序实现的关联规则挖掘算法是正确的.

3　结语

随着信息技术和通信技术的飞速发展，人类进入了大数据时代，大数据时代的“数据量”“数据类型”“数据变化速度”都呈现出爆炸式增长，这些爆炸式增长的数据有些是结构化或可转化为结构化的数据，因此，可采用关联规则挖掘对其进行分析研究. 此外，在大数据技术盛行的今天，传统关联规则挖掘方法可以从看起来并无关联的海量数据项集中找出具有潜在关系的两个或多个数据项，这对于大数据技术应用发展来说，无疑具有十分重要的意义.

（下转第 116 页）

联网、云计算、人工智能等技术正在重塑课堂教学模式，形成了一种多元立体化的优质教学资源体系格局，为城乡优质教育资源共享的数字鸿沟找到一个有效的解决方法，同时促进了城乡优质教育资源的均衡发展及教育公平，实现“人人皆学、处处能学、时时可学”的学习型社会.

［参考文献］

［1］中华人民国家中长期教育改革和发展规划纲要领导小组办公室. 国家中长期教育改革和发展规划纲要：2010—2020 年［M］. 北京：人民教育出版社，2010.

［2］教育部. 教育部关于印发《教育信息化“十三五”规划》的通知［J］. 苏州教育信息化，2016（3）：7－11.

［3］云南省人民政府办公厅. 关于印发云南省教育事业发展“十三五”规划的通知［J］. 云南教育（视界时政版），2017（5）：8－29.

［4］罗锋，蔡丹，夏文忠，等. 云南义务教育均衡发展现状刍析［J］. 昭通学院学报，2017（5）：118－124.

［5］魏丽丽. 浅析我国农村义务教育存在的若干问题及其对策［J］. 赤子（上中旬），2015（12）：122.

［6］谢敏.“互联网＋教育”背景下的教学智慧研究［D］. 长春：吉林大学，2016.

［7］张文青. 西安优质教育资源共享共建有效路径研究［D］. 西安：陕西师范大学，2016.

［8］汪学均. 视频互动同步课堂教学模式研究［J］. 中国电化教育，2017（4）：122－128.

［9］张天扬. 扩大中小学优质数字教育资源覆盖面的调查研究［J］. 黑龙江教育学院学报，2015（11）：87－88.

［10］刘艳.“互联网＋教育”理念下教学智慧影响因素分析［J］. 读与写（教育教学刊），2016（12）：282－283.

［11］杨彦军. 2016 年信息化教育软件产业发展现状述评［J］. 中国信息技术教育，2016（23）：12－14.

［12］任友群，郑旭东，吴旻瑜. 深度推进信息技术与教育的融合创新：《教育信息化“十三五”规划》（2016）解读［J］. 现代远程教育研究，2016（5）：3－9.

（上接第 99 页）

［参考文献］

［1］李泊溪. 大数据与生产力［J］. 经济研究参考，2014（10）：14－20.

［2］周吉，吕志坚. 大数据研究综述：信息处理的视角［C］//第 27 届全国计算机信息管理学术研讨会论文集. 北京：中国科学技术情报学会，2013.

［3］包达尔罕，罗敏，刘智伟，等. 大数据及其应用技术探讨［J］. 内蒙古民族大学学报（自然科学版），2014（2）：156－158.

［4］刘雨露. 数据挖掘在高校学生管理决策中的应用模式分析［J］. 成都信息工程学院学报，2006（3）：373－377.

［5］魏顺平. 教育数据挖掘：现状与趋势［C］//第 13 届全国计算机模拟与信息技术学术会议. 北京：中国优选法统筹法与经济数学研究会，2011.

［6］彭慧伶，刘发升. 关联规则挖掘与分类规则挖掘的比较研究［J］. 计算机与现代化，2006（7）：56－58.

［7］HAN J W，KAMBER M. 数据挖掘概念与技术［M］. 范明，孟小峰，译. 北京：机械工业出版社，2008：151－152.

［8］朱庆. 关联规则挖掘算法的研究［D］. 乌鲁木齐：新疆农业大学，2010.