**基于Apriori算法的学生成绩分析**

**在课程关联性的应用研究[[1]](#footnote-1)**

**陈喜华1 ，黄海宁1 ，黄沛杰2**

**（1广州工商学院 广东 广州 510850；**

**2华南农业大学 广东 广州 510642）**

**摘要：**利用Apriori关联规则算法，并借助SPSS Modeler软件，对学生成绩进行深层次分析，挖掘出课程之间的关联性，并对挖掘结果进行详细的分析和说明，这对今后设置课程、制定教学计划、修订人才培养方案、开展教学活动以及学生学习等具有重大的现实与指导意义。

**关键词：**关联规则；Apriori算法；SPSS Modeler；成绩分析；课程关联性

**中图分类号：**TP399 **文献标识码：** **文章编号：**

# 一、引言

随着高等教育普及率越来越高，在校学生规模越来越大，各个高校的教务管理系统也积累了大量的成绩数据，面对这些海量数据，很多教学管理人员还停留在对数据进行简单的增、删、改、查等操作，不能对隐含在数据背后有用的知识及信息进行有效的提取与分析。如何从这些学生成绩中发现有用的信息，找出课程之间的关联性，关联规则便能有效解决这一问题。

# 二、关联规则概述

关联规则（Association Rule）就是从大量数据中寻找项集之间存在的关联性[1]，它是最常用的数据挖掘方法之一。对事务数据库中的数据进行分析，就是要挖掘出满足最低条件支持度（min sup）阈值与最小规则置信度（min con）阈值要求的所有关联规则[2]。常用的关联规则算法有Apriori算法。

## （一）关联规则的概念

$设I由一个m个项目组成的集合I=\left\{i\_{1}，i\_{2}，\cdots ，i\_{m}\right\}$，设与任务相关的数据集D是数据库事务的集合，$其中每个事务T是项的集合， 使得A⊆I$，每一个事务都有唯一的标识符TID，$设A是一个项集，事务T包含A，当且仅当A⊆I$，关联规则是形如$A=>B\left[s，c\right]的蕴涵式，其中$，

$A⊂I，B⊂I，并且A∩B=Ø，则规则A=>B$在事务集D中成立[3]。s表示关联规则的支持度（Support），表示数据集D中同时包含A与B的交集，它是用户规定的关联规则必须满足的min sup；c表示关联规则的置信度（Confidence），表示数据集D中包含事务A的同时也包含B的百分比，它是用户规定的Association Rule必须满足的min con[4]。min sup与min con的公式如（1）-（3）所示。

$$ Support(A=>B)=P\left（A∪B\right） （1）$$

$$ Confidence(A=>B)= P(A|B)=\frac{P(A∪B)}{P(A)} （2）$$

$$ Confidence(A=>B)= P(A|B)=\frac{ Support(A∪B)}{Support(A)} （3）$$

同时满足min sup阈值与min con阈值的规则称为强规则[5]。项的集合称为项集（Item set）。含有k项的集合称为k项集，项集出现的频率是指包含项集的事务数，简称项集的频率、支持计数或计数[6, 7]。对于I及D，T中所有满足用户指定的min sup的项目集，即Confidence(A)≥min con的I的非空子集，称为频繁项目集或者大项目集。在频繁项目集中挑选出所有不被其他元素包含的频繁项目集称为最大频繁项目集[8]。

关联规则挖掘主要分为两个步骤：首先，找出频繁项集。需要满足$Support count(A)\geq min sup$。其次，由频繁项集产生强关联规则。需要满足$Support count(A)\geq min sup$，且$Confidence(A)\geq min con$[9]。

## （二）关联规则的分类

关联规则根据其变量类型划分，可以分为Boolean关联规则与Numerical关联规则。根据其数据的维数划分，可以分为Single Dimensional关联规则与Multidimensional关联规则。根据其数据的抽象层次划分，可以分为Single-Level关联规则与Multilevel关联规则。

## （三）Apriori算法的思想

Apriori算法是一种基于Boolean关联规则挖掘算法，用来挖掘数据库中各数据项之间的频繁模式[9]，即各数据项之间的关联性，其基本思想是利用先验知识逐层搜索的迭代方式来获取频繁项集，由第k-项集计算第(k+1)-项集。首先找出频繁1-项集，记为$L\_{1}，然后利用L\_{1}产生候选项集C\_{2}$，再计算出2-项集$L\_{2}$，以此类推，经过多次循环和迭代，直到无法找到更多的频繁k-$项集L\_{k}为止$[10]。

## （四）Apriori算法的流程

根据关联规则的步骤可知，实现Apriori算法，主要包含以下两个步骤。其流程图如图1所示。

N

Y

开始

扫描数据库，得到候选1-项集

计算支持计数，得到频繁1-项集

令k=1

对频繁k-项集进行自连接、剪枝操作，产生候选频繁(k+1)-项集

计算支持计数，得到频繁(k+1)-项集

频繁(k+1)-项集为空

得到频繁项集集合

结束

k++

图1 Apriori算法查找频繁项集流程图

第一，连接操作：首先扫描D，得到候选1-项集，再计算各项集的支持计数，确定频繁1-$项集L\_{1}$，接着利用1-项集进行自连接，得到频繁2-$项集L\_{2}$。依次类推，由$L\_{k-1}，最终找到频繁项集L\_{k}。$

第二，剪枝操作：$L\_{k-1}$与自身连接产生的候选项集记为$C\_{k}，C\_{k}是频繁项集L\_{k}的超集$，即$L\_{k}⊆C\_{k}。扫描D，计算C\_{k}$中各项集的支持计数，并确定所有支持计数不小于min sup阈值的候选项集为频繁项集$L\_{k}$[9]。

## （五）Apriori算法的性质

由于Apriori算法在迭代时需要重复扫描整个D，为了提高算法的执行效率，运用到以下三个重要性质。

性质1 一个项集的支持度不大于其子集的支持度[9]。

性质2 由频繁项集产生的所有非空子集也是频繁的[11]。

性质3 任何非频繁项集的超集一定也是非频繁项集[12]。

## （六）Apriori算法的优缺点

Apriori算法的优点是算法简单、容易实现，但存在以下缺点：

（1）频繁扫描数据库。在Apriori算法中，每一次迭代计算候选项集的支持计数，不管事务是否含有候选项集，都要完整扫描一次D。D中的事务越多，算法执行的时间就越长。

（2）容易产生大量的候选项集[13]。应用Apriori算法生成候选项集C过程之中没有及时删除不符合min sup阈值的项，导致由$L\_{k-1}$产生k-候选项集$C\_{k}$不断增长，延长了扫描D的时间。

# 三、基于关联规则的学生成绩分析

为了挖掘课程之间的关联性，本文利用关联规则的Apriori算法来分析学生成绩，整个实施过程分为方案设计、数据准备、模型建立、挖掘结果与分析四个部分。

## （一）方案设计

从学校教务系统导出2015级全体会计学本科专业学生（总共20个班，合计1270人）入学以来三个学期，9门专业必修课程成绩，导出数据源类型为Excel工作表，选择Apriori算法，IBM SPSS Modeler 14.1挖掘工具，挖掘课程之间的关联性，即哪些课程对学生成绩影响较大，哪些课程应先开设，哪些课程适合后开设。

## （二）数据准备

1.数据选取

本文选取广州GS学院会计系2015级会计学专业20个行政班级，共计1270名学生入学以来三个学期所有课程期末考试成绩作为研究对象，数据都是从该校教务管理系统导出。

2.数据集成与清理

（1）删除不相关数据。由于导出的原始数据包括任选课程、课程学分、课程性质、学分绩点、平均分和总分等信息，这些不是本次成绩挖掘分析的考虑因素，所以将其删除。

（2）合并类似课程。将类似课程合并，再采用求平均值的方法进行取值（四舍五入取整法）。需要合并的课程信息如表1所示。经过合并，剩下14门课程。

|  |
| --- |
| 表1 合并课程信息 |

|  |  |
| --- | --- |
| 合并前课程名称 | 合并后课程名称 |
| 大学英语Ⅰ、大学英语Ⅱ | 大学英语 |
| 大学语文Ⅰ、大学语文Ⅱ | 大学语文 |
| 中级财务会计Ⅰ、中级财务会计Ⅱ | 中级财务会计 |
| 大学体育Ⅰ、大学体育Ⅱ、大学体育Ⅲ | 大学体育 |
| 宏观经济学、微观经济学 | 经济学 |
| 微积分Ⅰ、微积分Ⅱ | 微积分 |
| 思想道德与法律基础、马克思主义基本原理、中国近现代史纲要 | 德育 |
| 创新创业基础、大学生职业发展与就业指导 | 创业与就业 |

（3）数据清理。由于原始数据出现部分学生的课程成绩被标注为“缓考”、“缺考”、“取消考试资格”和“舞弊”等特殊情况，此次在不减少研究数据，又能避免成绩存在太大误差的情况下，统一将此部分学生的成绩填充为所在班级的课程平均分。

3.数据转换

由于Apriori算法仅支持布尔型数据的关联分析，为了降低成绩数据的复杂度，首先需要将学生成绩中文本格式类型的数据统一转换成数值型数据，再将数值型数据转换成离散的Boolean类型数据，因为Apriori算法是基于Boolean Association Rule挖掘算法，本身并不支持连续型的数据。本例采用的方法是将学生成绩高于所在班级该课程的平均分记为l，低于所在班级该课程的平均分记为0，其目的是减少不同班级由不同教师评定成绩时带有主观因素造成的误差。

## （三）模型建立

利用Apriori算法，在SPSS Modeler中建立关联分析模型，如图2所示。



图2 Apriori关联规则算法挖掘模型

图2是针对本次挖掘目标所设计的流。“2015级会计学专业学生成绩.xls”作为源文件节点，经过字段过滤、类型设置，与模型建立连接。因为Apriori算法模型中一共用到14个属性字段，因此，称为14字段。最后，将运行后得到的模型和“网络”图形节点连接到流中相应位置，实现结果的多样性。

具体操作步骤和主要参数设置如下：

1.打开SPSS Modeler软件，添加“Excel”节点，导入准备好的数据文件，并设置所有数值型字段的“标准小数位置”为0，运行之后可得到相应的事实表。需要注意的是，在导入文件之前，必须关闭学生成绩工作簿，否则将提示“无法读取文件列名”。

2.添加“过滤”节点，过滤掉序号、学号、姓名、班级、性别等不相关的字段。

3.添加“类型”节点。所有字段的“测量”设置为“标志”，“角色”设置为“两者”。两者是指既可作为输入数据又可作为输出数据。最后“读取值”。

4.在数据流编辑区添加“Apriori”模型节点，并与“类型”节点建立连接。双击“Apriori”节点，在“字段”选项卡，选择“使用定制设置”，把“后项”与“前项”都设置为“全选”。在“模型”选项卡中设置min sup=30，min con=87%，“最大前项数”为10，如图3所示。注意：此过程需反复调整参数，以便得到合适的参数，最终才能获得相对准确的结果。按照此处的设置，可以得到12条规则。



图3 设置“模型”选项卡

5.为了直观看到统计结果之间的关系，在“类型”节点后添加“网络”节点，在“图”选项卡中，字段选择“全选”，并勾选“仅显示真值标志”；在“选项”选项卡中将“可显示的最大链接数”设为25，如图4所示。其他默认设置。



图4 设置“选项”选项卡

## （四）挖掘结果与分析

1.挖掘结果

在图3中，单击“运行”按钮，生成挖掘模型“14字段”，并与“类型”节点自动建立连接。双击模型“14字段”，可以得到挖掘结果，如图5所示，按照上面的设置，此处一共生成12条关联规则。



图5 Apriori挖掘模型

在图4中，单击“运行”按钮，得到“14字段：绝对值的网状图”，如图6所示。



图6 14字段：绝对值的网状图

2.结果分析

根据图6的网状图，可以发现不同课程之间的连线有多有少，有粗有细，连线越多，说明与其他课程的联系就越密切；连线越粗，说明课程之间的影响程度就越大。对于那些与其他课程有较多联系的课程，应作为基础课程，较早开设，如：法规与职业道德、中级财务会计、经济学、税法、会计学、创业与就业。对于彼此之间相互影响程度较大的课程，应重点考虑哪门课程作为先导课程，哪门课程作为后续课程，如：中级财务会计和会计学，中级财务会计和经济学，中级财务会计和法规与职业道德，法规与职业道德和税法。

根据图5的挖掘结果，并结合对图6网状图的分析，可以得到：

（1）会计学、法规与职业道德、经济学、税法能够促进中级财务会计的学习，说明中级财务会计应在修读完前面四门课程基础上再开设。统计学、微积分等课程虽能促进中级财务会计的学习，但影响并不大。

（2）法规与职业道德能促进经济学、税法、会计学的学习，因此，应考虑优先开设。

（3）经济学和税法，经济学和会计学之间是互为促进的关系，可以同时开设。

（4）在模型中没有挖掘出大学语文和创业与就业之间的规则，但网状图却显示它们存在密切的关联，可能是数据在集成与清理、数据转换时产生了误差，又或者它们达不到“最低条件支持度”和“最小规则置信度”。

（5）在模型中没有挖掘出“计算机应用基础=>会计信息系统”这一规则，说明这两门课程虽然都与计算机操作相关，但关系并不密切。

（6）挖掘模型和网状图均没有挖掘出计算机应用基础、大学英语、大学体育和会计信息系统，可能它们与其他课程的关联程度极小，而且，前三门课程均是通识基础课，独立性较强，因此，也属正常。

通过对以上结果分析，可以明确课程之间的关联性以及重要性，挖掘结果能给学校教学管理人员设置课程、制定教学计划以及修订人才培养方案提供决策支持。

3.误差分析

（1）原始数据有限，规模也不是很大。

（2）学生的综合成绩=平时成绩（30%）+中考成绩（30%）+末考成绩（40%），教师对学生的态度会影响到平时成绩的高低，学生在不同时间段的学习效果也可能影响到中考成绩和末考成绩。

（3）原始学生成绩在清理、集成、转换、离散化过程中，产生了误差。

4.有效性分析

通过借助SPSS Modeler软件，运用Apriori算法对学生成绩进行分析，能够挖掘出隐藏在成绩之中有价值的知识与信息，找出课程之间的关联性，并自动计算出课程之间的支持度与置信度，有效地度量了课程之间的影响程度，虽然结果可能存在一定误差，但给教学管理人员分析课程之间的关联性提供了一种思路，也为其决策提供了参考依据。

# 四、总结

本文首先介绍了Association Rule的概念与分类，接着对Apriori算法进行研究，最后借助SPSS Modeler软件，结合Apriori关联规则算法，从学生成绩中挖掘出课程之间的关联性，并且对挖掘结果进行详细的分析与说明，为今后教学管理人员开展相关教学管理工作提供决策支持。

Study on the Application of Students' Grades Analysis Based on Apriori Algorithm
in Curriculum Relevance

CHEN Xihua**1**, HUANG Haining**1**, HUANG Peijie**2**

（1.Guangzhou College of Technology and Business, Guangzhou, Guangdong 510850, China; 2.South China Agricultural University, Guangzhou, Guanghdong 610642, China）

**Abstract****:** By using the association rule algorithm —— Apriori and using the SPSS Modeler software, the author makes a deep analysis of the students' grades, excavates the relevance between courses, and analyzes and explains the excavation results in detail, which is of practical and instructive significance to set up course, make teaching plan, revise talent training program as well as carry out teaching activities and students' learning in the future.

**Key words:** association rule; Apriori algorithm; SPSS Modeler; grades analysis; curriculum relevance

(责任编辑：陈小全)

**参考文献**

[1] 马志新. 频繁项集挖掘问题的研究[D]. 兰州大学, 2005：16-18.

[2] 张素兰. 一种基于事务压缩的关联规则优化算法[J]. 计算机工程与设计, 2006,27(18):3450-3453.

[3] 陈小健, 宋承继. 数据挖掘技术在科研项目量化与评价中的应用[J]. 电子设计工程, 2015，(12):37-40.

[4] 王立亚, 张春英, 刘保相. 带参数区间关联规则挖掘算法与应用[J]. 计算机科学与探索, 2016，(11):1546-1554.

[5] 温海波. 近似频繁集合强亲密关联性的矩阵挖掘方法[J]. 安徽建筑工业学院学报：自然科学版, 2013，(03):56-60.

[6] 黄龙军, 段隆振, 章志明. 一种基于上三角项集矩阵的频繁项集挖掘算法[J]. 计算机应用研究, 2006，(11):25-26.

[7] 叶晓波. 一种基于二进制编码的频繁项集查找算法[J]. 楚雄师范学院学报, 2009，(03):13-19.

[8] 胡维华, 冯伟. 基于分解事务矩阵的关联规则挖掘算法[J]. 计算机应用, 2014，(S2):113-116.

[9] 严的兵. 基于数据挖掘的学生成绩分析[D]. 安徽大学, 2011：5-8.

[10] 任伟建, 于博文. 基于矩阵约简的Apriori算法改进[J]. 计算机与现代化, 2015，(9):1-5.

[11] 郭云峰, 张集祥. 一种基于位向量的关联规则挖掘算法[C]. 2008年中国信息技术与应用学术论坛,成都:西南信息中心,2008.

[12] 陈波, 王乐, 董鹏. 挖掘最大频繁项集的事务集迭代算法[J]. 计算机工程与应用, 2009,(06):141-144.

[13] 谢川, 孙逢啸, 倪世宏. 一种基于矩阵的Apriori改进算法[J]. 计算机仿真, 2013,(08):245-249.

1. 收稿日期：2018年03月30日

作者简介：陈喜华（1985-），男，广东潮州人，网络工程师、助理研究员，硕士学位，研究方向：教育教学管理，计算机应用

黄海宁（1991-），女，广东广州人，硕士学位，研究方向:教育教学管理，学科教学

黄沛杰（1980-），男，广东潮州人，通讯作者，副教授，博士，研究方向：人工智能、自然语言处理、口语对话系统

基金项目：2017年广州工商学院校级质量工程--高等教育教学改革项目研究成果之一。（编号：ZL20171138）

广东省教育研究院民办教育研究基地（广州工商学院）2017年科研课题《民办高校学科、专业、课程、教材建设与教学改革研究》(编号：GMJ201707)。 [↑](#footnote-ref-1)