doi:10.16104/j.issn.1673-1891.2019.02.022

关联规则兴趣度挖掘在选课中的应用探讨

李佐军
(滇西科技师范学院信息工程学院,云南 临沧 677000)

摘要:高校信息化的选课系统积累了大量"闲置"的选课数据,如何把这些"闲置"数据利用起来为高校服务成了教学管理人员需要解决的问题。为了解决"闲置"选课数据的问题,对关联规则兴趣度挖掘在选课中的应用开展了分析讨论。使用Visual FoxPro语言编写了选课数据分析软件,并对选课数据进行挖掘分析,找出不同专业类型的学生对不同类型课程的偏好,为教师指导学生选课具有重要意义。

关键词:关联规则;兴趣度;闲置数据;选课
中图分类号:TP311.13 **文献标志码**:A **文章编号**:1673-1891(2019)02-0103-03

Discussion on the Application of Association Rules' Interestingness Data Mining in Selective Courses Management

LI Zuojun
(School of Information Science & Engneering, West Yunnan University, Lincang, Yunnan 677000)

Abstract: The selective course system of colleges and universities has accumulated a large number of "idle" course selection data, and how to use these "idle" data in service of colleges and universities has become a problem that teaching management needs to address. To solve the "idle" course selection data problem, the application of the association rules' interestingness data mining in the course selection is analyzed and discussed. The course selection data analysis software is written in Visual FoxPro language, and the course selection data are analyzed to find out the preferences of different types of students for different types of courses, which can be very helpful when teachers guide students to the selection of courses.

Keywords: association rules; interestingness; idle data; selection of courses

随着信息技术的发展和高校招生规模的扩大,高校都采用了信息化选课系统,这些系统在运行中产生了大量的选课历史数据,使用传统的查询、统计等方法无法获取这些历史数据内在价值,造成这些数据长期处在"闲置"状态。数据挖掘技术正好能解决"闲置"数据的问题,通过数据挖掘技术分析信息化选课系统产生的闲置历史数据,找出专业与选修课之间的关系,为学校调整课程开设、设置课程的学分等提供理论指导。使用选课率等信息做出判断,指导学生选择有利于自己整体素质提高的互补课程,也可为解决师资、教室等资源合理配置提供依据[1]。

1 数据挖掘和关联规则

1.1 数据挖掘概念

自从二十世纪九十年代以来,数据挖掘相关技术取得了飞速发展,其概念也更新了多次。数据挖掘是一个多学科交织的、综合性学科,内容包含数学、统计学、数据库、人工智能等[2]。数据挖掘分数据采集、预处理、挖掘和评估表示四个步骤,数挖掘基本步骤如图1所示。数据挖掘常用算法为:关联规则算法、决策树方法、人工神经网络算法、遗传算法、粗糙集方法、模糊论方法、贝叶斯模型算法等[3]。

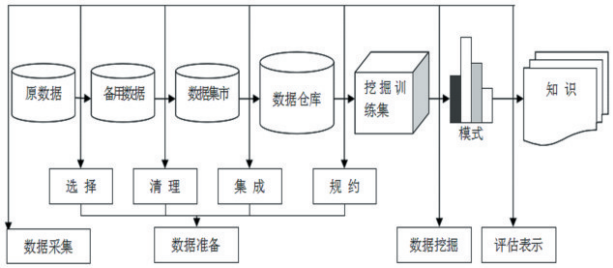


图1 数挖掘基本步骤

1.2 关联规则的定义

Agrawal等人在一九九三年提出了关联规则挖

收稿日期:2018-12-31
作者简介:李佐军(1980—),男,云南凤庆人,副教授,硕士,研究方向:数据挖掘及大数据应用技术。

掘技术[4]。关联规则挖掘是为了在数据库中找出数据项的关联程度而提出的,是实用性强、处理简单的规则产生方式,应用于发现大量数据集中的数据间关联性和相关性,用来描述事务中属性间同时发生的规律和模式[5]。

设$I=\{I_1,I_2,I_3,\cdots\cdots I_m\}$是m个不同项目或属性的集合,事务$T$(Transaction)是项集$I$上的子集,即$T\subseteq I$,且每个事务T都有一个唯一的标识TID,不同事务组成全体事务集D[6]。若X、Y是项集I的子集,X在某事务中存在,则必然会使Y亦在这个事务中存在,则关联规则表示为R:$X\Rightarrow Y$,其中$X\subset I$,$Y\subset I$且$X\cap Y=\varnothing$,X称为先决条件,Y称为结果。关联规则表示项集X中的项目出现时,项集Y中的项目也同时出现的概率。

关联规则用"支持度(Sup)"和"可信度(Conf)"两个参数衡量挖掘结果的使用价值,通过设置它们的阈值来删除无用或作用很小的规则。用(1)式来计算关联规则(R:$X\Rightarrow Y$)支持度S($X\Rightarrow Y$),支持度表示项集X在所有事务集中出现的频率;用(2)式来计算关联规则(R:$X\Rightarrow Y$)置信度C($X\Rightarrow Y$),置信度表示项集Y中的项目在项集X中出现的概率[7]。

$$S(X\Rightarrow Y)=\mathrm{P}(X\cup Y)=\frac{\mathrm{Count}(X)}{\mathrm{N}}\times 100\% \quad (1)$$

$$\mathrm{C}(X\Rightarrow Y)=\mathrm{P}(Y|X)=\frac{\mathrm{S}(X\Rightarrow Y)}{S(X)}=\frac{Count(X\cup Y)}{Count(X)}100\% \quad (2)$$

1.3 Apriori算法

Apriori算法是关联规则挖掘技术的最基本算法。第一步,它找出事务数据库中的全部频繁项集;第二步,根据设定阈值生成强关联规则。Apriori算法的基本思想是基于频繁项集生成过程,使用递推方式推演出频繁项集,使用频繁项集L_{k-1}产生频繁项集L_k,其过程分为连接和剪枝两步[8]。

第一步,连接:使用项集L_{k-1}与自己相连接生成候选k项集的集合C_k。

第二步,剪枝:L_k是候选k项集C_k的子集,扫描事务数据库D,确定候选项集C_k中候选项的个数,最终确定频繁项集L_k。

1.4 关联规则兴趣度的引入

在传统关联规则算法中,只引用支持度与置信度来衡量挖掘结果的价值,这会导致关联规则的误差比较大。因此,引入兴趣度来提高挖掘结果的准确性。李永立等人在文献[9]提出了一种关于兴趣度的模型,把兴趣度定义为:如果$I=\{i_1,i_2,\cdots,i_n\}$为项集,D是全体事务集,则关联规则X⇒Y的兴趣度表示为(3)式。兴趣度值越大,则关联规则的使用价值越高[10]。

$$I(X\Rightarrow Y)=\frac{\frac{Count(X\cup Y)}{Count(X)}-\frac{Count(Y)}{N}}{\sqrt{\frac{\frac{Count(Y)}{N}\times\left(1-\frac{Count(Y)}{N}\right)}{N}}} \quad (3)$$

2 关联规则挖掘实现

2.1 数据准备

选取滇西科技师范学院2012年到2018年的学生选课数据作为挖掘分析对象,对选课相关数据进行合法性不断、缺失值处理、归类等预处理后,数据量约16 000条。每条选课信息包含学号、专业类(教育学类、工商管理类、中国语言文学类、计算机类、植物生产类、马克思主义理论类、法学类、体育学类、经济学类等)、选修课类型(自然科学类选修课、工程技术类选修课、社会科学类选修课、人文艺术类选修课、经济管理类选修课)字段。预处理后的选课信息表结构如表1所示,表1中"Y"表示该名同学选修了此类选修课程,空的表示没有选修此类课程(表2同)。

表1 选课信息表结构

学 号	专业类	自然科学类选修课	工程技术类选修课	社会科学类选修课	人文艺术类选修课	经济管理类选修课
0101110241	教育学类	Y		Y	Y	
0103110110	服务业管理类		Y	Y		Y
0501110127	计算机类		Y	Y	Y	
……	……	……	……	……	……	……

2.2 算法实现

选课关系挖掘工具使用Visual FoxPro 6.0作为后台数据库管理工具和系统开发平台。

2.2.1 数据库实现

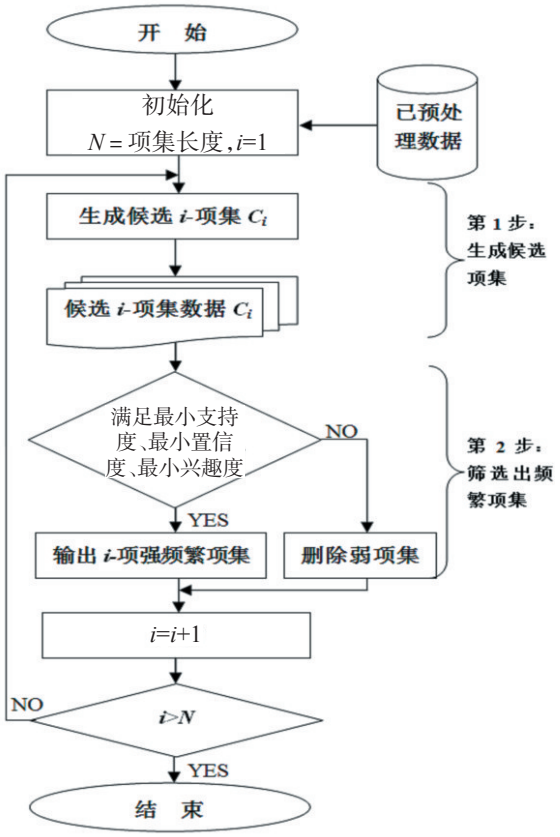
选课数据表包含学号、专业类、选修课类型(自然科学类选修课、工程技术类选修课、社会科学类选修课、人文艺术类选修课、经济管理类选修课)7个字段,其数据结构见表2所示。

表2 选课信息表结构

序号	字段名	类型	宽度	备 注
1	学号	字符型	12	
2	专业类	字符型	16	教育学类、计算机类等
3	自然科学类选修课	字符型	1	Y或空
4	工程技术类选修课	字符型	1	Y或空
5	社会科学类选修课	字符型	1	Y或空
6	人文艺术类选修课	字符型	1	Y或空
7	经济管理类选修课	字符型	1	Y或空

2.2.2 算法实现

选课关系挖掘软件的分析对象是学生选课数据，以“属性A→属性B 支持度：S% 置信度：C% 兴趣度：XQD”格式生成规则。使用Visual FoxPro编程语言实现基于兴趣度的关联规则算法，算法流程如图2所示。



3 挖掘结果分析

3.1 挖掘结果显示

根据预先设定的最小支持度阈值(2%)、最小置信度阈值(20%)和最小兴趣度阈值为5，以形如“属性A→属性B 支持度：S% 置信度：C% 兴趣度：XQD”的式子显示满足条件的规则。挖掘界面及结果如图3所示。

3.2 挖掘结果分析及应用

由选课关系挖掘结果图(图3)可知，当支持度阈值、置信度阈值、兴趣度阈值分别为2%、20%、5

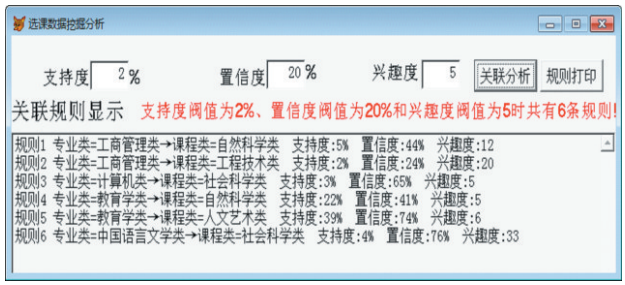


图3 选课关系挖掘界面及结果显示图

时，选课数据挖掘分析系统共产生6条规则，做进一步分析，可得出以下结论：

第一，从规则1-2可看出，专业为工商管理类的学生倾向于选修自然科学类和工程技术类课程，说明工商管理类的学生对自然科学和工程技术相关知识比较感兴趣。因此，学校在进行课程设置时，多设置与自然科学和工程技术相关选修课供专业为工商管理类的学生选择。

第二，从规则3可看出，专业为计算机类的学生倾向于选修社会科学类的课程，说明计算机类的学生除了专业课外，还对社会科学类相关知识感兴趣。所以在进行计算机类专业课程设置时，适当增加社会科学类课程。

第三，从规则4-5可看出，专业为教育类的学生倾向于选修自然科学和人文艺术类的课程。所以应对教育类专业的学生开设更多的自然科学和人文艺术类的课程。

第四，从规则6可看出，专业为中国语言文学类的学生主要选修社会科学类的课程。因此，学校负责选课指导的教师要引导中国语言文学类学生选修除了社会科学类的课程，补充学生知识结构的全面性。

4 结语

文章对关联规则挖掘在选课中的应用进行研究探索，引入了兴趣度阈值，提高了挖掘结果利用价值。使用Visual FoxPro语言编写的选课数据分析软件对学生“专业类—选修课程类”数据进行分析，找出不同类型专业学生对不同类型课程的选择偏好，为教师指导学生选课提供了依据，对学校选课管理具有重要意义。

参考文献：

[1] 王博,刘庆刚,张琴.数据挖掘在选课系统中的应用[J].计算机与数字工程,2011,39(5):83-86.

[2] 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.

[3] 张国荣.基于粗糙集的数据挖掘算法研究与应用[D].兰州:西北师范大学,2011.

[4] 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.

(下转第119页)

学习态度,变被动为主动,更多地参与到课堂之中,转变了期末考试前突击复习的做法,更多精力投入到平时的学习之中。通过软件操作的强化训练与统计调研报告的接合,学生对统计学课程知识点掌握得更牢固,实践能力也得到明显提升,基本达到改革预期目标[5]。另一方面,在改革过程中发现,虽然平时考核内容和形式都进行了改革,但大多数都是以小组形式开展,容易造成学生"搭便车"现象,这也在一定程度上影响了部分学生参加平时各项教学活动的积极性。可考虑从3个方面入手加以改进:首先,加强过程性考核监控力度。由于重平时考核这种过程性考核,所以需要加强过程性考核监控力度,确保每一位同学真正参与到小组活动当中;其次,在小组考核的基础上,增加小组内成员间的互评,对小组成员的不同贡献进行区分,增加每一位学生的参与度;最后,加强考后的沟通。考试的最终目的是让学生能力得到提升,因此考后的试卷分析不仅是老师的工作,学生也应参与其中,应加强考后的师生沟通,将教学过程延长至考后的分析与总结。

参考文献:

[1] 曹玲玲,陈沛然.构建应用技术型高校统计学课程考核评价指标体系的探索[J].黑龙江畜牧兽医,2016(11):248-251.
[2] 方磊,齐瑞,孙萍萍.基于形成性评价的《传统康复方法学》PBL教学考核研究[J].时珍国医国药,2016(8):1999-2001.
[3] 常乐.翻转课堂教学模式过程考核设计——以"动态网页设计"课程为例[J].无线互联科技,2017(9):85-86.
[4] 梁旭华.过程考核评价体系的建立——以制药设备与车间设计课程为例[J].价值工程,2018(1):221-222.
[5] 刘静.医学统计学课堂教学的多元化改革探索[J].中国卫生统计,2017(2):150-152.

(责任编辑:蒋召雪)

(上接第105页)

[5] 张丽.关联规则挖掘研究[J].赤峰学院学报(自然科学版),2009,25(5)17-18.
[6] 陈晓云,胡运发.N个最频繁项集挖掘算法[J].模式识别与人工智能,2007,20(4):512-518.
[7] JIAWEI H,MICHELINE K著,范明,孟小峰译.数据挖掘概念与技术[M].北京:机械工业出版社,2008:151-152.
[8] 叶强,李一军.基于支持度-显著度的关联规则分类方法研究[C].南京:中国系统工程学会,2005.
[9] 李永立,吴冲,王崑声.一种新的关联规则兴趣度度量方法[J].情报科学,2011,30(5):503-507.
[10] 李佐军.大数据时代下关联规则兴趣度挖掘在就业分析中的应用[J].软件工程,2018,21(11):25-27.

(责任编辑:曲继鹏)