基于关联规则 Apriori 算法的学生成绩分析

Analysis of Students´ Achievement Based on Association Rule Apriori Algorithm

王成勇 WANG Cheng-yong

(华北电力大学数理学院,北京 102206)

(Mathematics and Physics Department, North China Electric Power University, Beijing 102206, China)

摘要:关联规则挖掘是数据挖掘领域研究的热点问题,其中 Apriori 算法是经典的关联规则算法。将关联规则 Apriori 算法应用到学生成绩分析中,挖掘出课程与课程之间的相互关系,寻找各方面影响学生成绩的因素,发现隐藏在数据背后有价值的信息,从而为学生选课和教师教学以及教学管理工作等提供辅助性的建议与决策。

Abstract: Association rule mining is a hot topic in the field of data mining. Apriori algorithm is a classical association rule algorithm. This paper applies the association rule apriori algorithm to analyze student achievement data, digs out the relationship between the course and the curriculum, finds out the factors that affect the student achievement in all aspects, and finds the hidden information behind the data, so as to provide supplementary advice and decision-making for student course selection, teacher teaching and teaching management.

关键词:学生成绩分析 数据挖掘;关联规则技术;Apriori 算法

Key words: student achievement analysis; data mining; association rule technique; Apriori algorithm

中图分类号:TP311.1 **文献标识码**:A **文章编号**:1006-4311(2018)05-0171-03

DOI:10.14018/j.cnki.cn13-1085/n.2018.05.068

0 引言

近年来随着信息技术的飞速发展,数据资源变得越来越丰富,在高校的教学管理系统中存储了大量的学生成绩数据信息,但由于缺乏必要的技术手段,因此只能对这些数据信息进行简单的统计、备份和查询。隐藏在这些大量成绩数据背后的信息不能得到有效的利用,不利于人才的培养和教学质量的提高,因而迫切需要有更新的技术方法对这些数据进行处理分析。

关联规则挖掘[1-2]就是一门从历史数据集中发现隐含模式,从海量数据集中发现潜在的有价值信息的技术方法,它反映了一个事件与其他事件直接依赖或关联的知识,这几年已经成为数据挖掘技术研究领域的热门话题[3]。本文运用关联规则 Apriori 算法挖掘学生成绩数据,可以挖掘出课程与课程之间的相互关系、影响学生成绩的因素等一些有价值的信息,这些信息可为教学及管理工作提供支持性的建议,同时也为更加合理的制定人才培养方案和提高教育教学质量提供科学依据。

1 关联规则基本理论

设有事务数据库 $D=\{T_1,T_2,\cdots,T_n\}$,T_j $(j=1,2,\cdots,n)$ 称为事务 T,构成 T 的元素 i_k $(k=1,2,\cdots,p)$ 被称为项,设 D 中所有项的集合为 $I=(i_1,i_2,\cdots,i_m)$,则有 $T\subseteq I$。

定义 1 项集和频繁项集:假设 $A=\{i_1,i_2,\cdots,i_t\}$ $(1\leq t\leq m)$,那么 A 称为 D 中的一个项集,而且为 t 项集。项集 A 的支持度也就是 D 中包含 A 的事务在 D 的全部事务中所占的比例,即

$$\text{Support}(A)=\frac{|\{T:A\subseteq T,T\in D\}|}{|D|}=P(A)$$

如果项集 A 的支持度大于或者等于最小支持度阈值 min_Support,即 $\text{Support}(A)\geq\text{min\_Support}$,那么项集 A 称为 D 中的频繁项集。

定义 2 关联规则:关联规则是形如 $A\Rightarrow B$ 的形式,A 和 B 都是 D 中的项集,并且满足 $A\cap B=\varnothing$。项集 A 称为关联规则的条件,项集 B 称为关联规则的结论。

定义 3 支持度和置信度:关联规则 $A\Rightarrow B$ 的支持度就是同时包含 A 和 B 的事务在 D 的所有事务中所占的比例,即项集 $A\cup B$ 的支持度

$$\text{Support}(A\Rightarrow B)=\frac{|\{T:A\cup B\subseteq T,T\in D\}|}{|D|}\text{Support}(A\cup B)$$

$$=P(A\cup B)$$

关联规则 $A\Rightarrow B$ 的置信度就是指同时包含 A 和 B 的事务在全部包含 A 的事务中所占的比例

$$\text{Confidence}(A\Rightarrow B)=\frac{|\{T:A\cup B\subseteq T,T\in D\}|}{|\{T:A\subseteq T,T\in D\}|}$$

$$=\frac{\text{Support}(A\Rightarrow B)}{\text{Support}(A)}=P(B|A)$$

定义 4 强关联规则:假设有关联规则 $A\Rightarrow B$,它的支持度和置信度分别满足最小支持度阈值和最小置信度阈值,那么就称该规则是强关联规则。

基于以上定义,关联规则挖掘可以描述为从给定的事务数据库中,通过合适的挖掘算法,找出满足最小支持度阈值和最小置信度阈值的全部强关联规则。

2 关联规则挖掘算法

2.1 寻找频繁项目集

在对学生成绩数据进行关联规则分析时,这里采用了 Apriori 算法来寻找全部的频繁项目集。Apriori 算法是一种重要的关联规则挖掘算法,它使用了一种被称为逐层搜索的迭代算法,k-项集用于搜索 (k+1)-项集。首先需要扫描事物数据库,累积每个项的计数,然后收集满足最小支持度的项,从而找出频繁 1-项目集的集合 L_1。L_1 用于寻找频繁 2-项目集的集合 L_2,而 L_2 用于寻找频繁 3-项目集的集合 L_3,如此下去,直至不能找到频繁 k-项目集 L_k 为止[4]。

运用频繁 k-项集用于搜索 (k+1)-项集是 Apriori 算

作者简介:王成勇(1991-),男,满族,河北承德人,华北电力大学数理学院硕士研究生,研究方向为数据挖掘、电力市场稳定性。

法的核心，该步骤分为连接步和剪枝步：

①连接步骤：为了寻找 L_k，在 k（k>1）次扫描数据库时，通过 L_{k-1} 与自身连接产生候选 k-项集的集合 C_k。

②剪枝步骤：由于 C_k 是 L_k 的超集，即 C_k 的成员可能是也可能不是频繁的。需要扫描全部的事务数据库，确定 C_k 中每个候选的计数，判断是否大于或者等于最小支持度计数，如果是，那么便认为该候选是频繁的。为了压缩 C_k，可以运用 Apriori 性质：任何一个频繁项集的全部非空子集也一定是频繁的，若某个候选的非空子集不是频繁的，那么该候选项集肯定也不是频繁的，从而可以将其从 C_k 中删去。

Apriori 算法描述如下[5-6]：

输入：数据库 D；最小支持度 min_Support

输出：D 中的频繁项目集 L

方法：

L_1=find_frequent_1-itemsets(D);;

for(k=2;$L_{k-1}\neq\Phi$;k++){

C_k=apriori_gen(L_{k-1},min_Support)

for each transaction t ∈ D{

C_t=subset(C_k,t);

for each candidate c ∈ C_t

c.count++;

}

L_k={c ∈ C_k|c.count≥min_Support}

}

return L=$U_k L_k$

2.2 生成强关联规则

对于上面得到的每个频繁项目集 L，生成强关联规则的步骤如下：

①生成 L 的所有非空子集；

②对于 L 的每个非空子集 S，令 R=L-S。

如果有

$$\frac{\text{Support}(S\cup R)}{\text{Support}(S)}\geq \text{Min\_Confidence}$$

即 S⇒R 满足最小置信度阈值，那么输出关联规则 S⇒R。又因为这个规则是从频繁项目集 L 中生成的，因此一定满足最小支持度阈值，所以这个规则为强关联规则。根据上面的两个步骤，就可以得出事物数据库 D 的全部强关联规则。

3 应用 Apriori 算法分析学生成绩

3.1 挖掘目标与流程

关联规则挖掘必须具有针对性，也就是说挖掘目标要明确，本文希望通过对学生成绩数据信息进行研究，找到满足最小支持度和最小置信度的强关联规则，挖掘出课程与课程之间的相互关系，并期望以此结果来指导教育教学工作。其中关联规则挖掘的具体过程如图 1 所示。

3.2 数据采集

关联规则挖掘需要丰富的数据信息作为基础。本研究选取学生成绩数据库中 8 门专业课程作为研究对象，选取 1000 条数据，用以挖掘课程之间的关联性。学生成绩信息数据如表 1 所示。其中 Xuehao 为学号，A~H 分别代表 8 门课程。

3.3 数据的处理

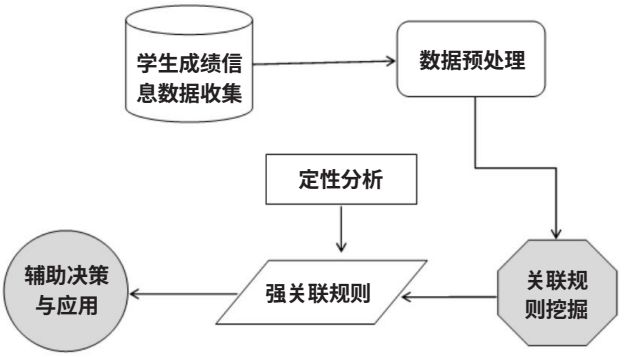

图 1 关联规则挖掘流程图

表 1 学生成绩数据表

Xuehao	A	B	C	D	E	F	G	H
001	88	92	63	76	69	52	88	75
002	67	54	60	79	73	77	80	64
003	90	93	81	67	85	82	69	81
004	86	62	90	73	92	78	71	91
005	76	82	92	80	57	66	74	80
…	…	…	…	…	…	…	…	…

通过对原始数据进行简单的泛化处理，可以得到更加丰富的数据信息[7-8]。在这部分将对成绩数据进行离散化，成绩达到 90 分及以上的代表“优秀”、成绩在 80 分（包括 80 分）到 90 分之间的代表“良好”、成绩在 70 分（包括 70 分）到 80 分之间的代表“中等”、成绩在 60 分（包括 60 分）到 70 分之间的代表“及格”、成绩在 60 分以下的代表“不及格”，其中“优秀、良好、中等、及格、不及格”分别用数字“1、2、3、4、5”表示，离散化后的数据如表 2 所示。

表 2 转换后的学生成绩数据表

Xuehao	A	B	C	D	E	F	G	H
001	A_2	B_1	C_4	D_3	E_4	F_5	G_2	H_3
002	A_4	B_5	C_4	D_3	E_3	F_3	G_2	H_4
003	A_1	B_1	C_2	D_4	E_2	F_2	G_4	H_2
004	A_2	B_4	C_1	D_3	E_1	F_3	G_3	H_1
005	A_3	B_2	C_1	D_2	E_5	F_4	G_3	H_2
…	…	…	…	…	…	…	…	…

3.4 挖掘关联规则

这一步的关键是选择恰当的关联规则挖掘算法对数据进行分析处理。这里采用关联规则 Apriori 算法对离散化后的学生成绩数据信息进行挖掘。设定最小支持度为 25%、最小置信度为 60%。运行关联规则 Apriori 算法程序后，得到的部分实验结果如表 3 所示。

表 3 部分关联规则结果表

编号	规则	支持度（%）	置信度（%）
1	B_1=>G_1	43	79
2	B_4=>G_5	40	75
3	$A_2\wedge C_2$=>F_2	38	87
4	D_1=>H_1	32	76
5	H_1=>E_1	27	70
…	…	…	…

3.5 结果分析

对于挖掘得到的强关联规则，需要对结果进行分析。根据表 3 可知，规则 1 和 2 说明了学好 B 课程对于学好 G 课程有着重要的影响，在安排课程的时候，要将 B 课程排

基于解释结构模型的 IE 专业本科毕业设计考评体系研究

Research on Evaluation System of IE Professional Undergraduate Graduation Design Based on Interpretative Structure Modeling

刘冠权 LIU Guan-quan;谢亚雯 XIE Ya-wen;陈艳 CHEN Yan
(青岛理工大学管理学院,青岛 266520)
(School of Management,Qingdao University of Science and Technology,Qingdao 266520,China)

摘要:本科毕业设计作为学生巩固专业知识和提升综合能力的关键实践环节,但在实践中,却忽视了毕业设计的过程特性和综合特性。运用解释结构模型的方法构建应用型 IE 专业本科毕业设计评价指标体系的基础上,采用层次分析法确定了毕业设计各个环节的权重,构建了对全过程进行综合考核的评价体系,使得答辩成绩更加公平可靠,从而从根本上提升毕业设计的质量。

Abstract: Graduation design is a key practical step for student to consolidate profesional knowledge and enhance the comprehensive ability. But in practice, the essential procedural and comprehensive characteristics of graduation design are ignored. Based on the evaluation index system of the applied IE professional graduation design by using the method of Interpretative Structual Modeling (ISM) , and using the Analytic Hierarchy process (AHP) to determine the weight of all sections of graduation. An evaluation system has been built which covers the comprehensive assessment of whole process. Evaluation system makes the defense results more fair and reliable, so as to improve the quality of graduation design fundamentally.

关键词:毕业设计;解释结构模型;层次分析法;考核评价体系

Key words: graduation design;ISM;AHP;Evaluation System

中图分类号:G642.477 **文献标识码**:A **文章编号**:1006-4311(2018)05-0173-04

DOI:10.14018/j.cnki.cn13-1085/n.2018.05.069

0 引言

工业工程是一门对生产系统进行设计、改善和实施的应用型学科,随着我国经济转型升级、创新驱动、发展先进制造业和现代服务业的双重需求[1],对 IE 人才的能力也有了新的需求,要求不仅要掌握多学科的专业知识,也要具备一定的综合性技能,同时富有开拓精神和创新能力的复合型、应用型人才[2]。毕业设计(论文)环节是高校教学计划中的一个重要环节,也是最后一个环节,是对学生四年学习内容的综合应用与升华,在培养学生实践能力和创新精神的同时,也是衡量学校教学质量的重要方面[3]。很多学生不愿意花时间和精力好好完成毕业设计,更多的是依靠网络,抄袭之风盛行,在思想上对毕业设计不够重视,认为毕

基金项目:高校教改项目(MX4-052)。

作者简介:刘冠权(1971-),男,陕西榆林人,青岛理工大学管理学院副教授,博士,硕士生导师,研究方向为现代工业工程理论与应用。

在前面,同时教师在教学过程中要督促学生学好 B 课程。

规则 3 说明如果 A 课程和 C 课程学的好,那么 F 课程也就学的好一些。从规则 3 的置信度来分析,其置信度为 87%,说明 A、C 课程与 F 课程的关联程度比较强。在课程的设置方面,A、C 课程需要排在 F 课程的前面。

规则 4 和 5 说明了 D、E、H 三门课程关联比较紧密,并且 D 课程是最关键的,教师在讲解时要仔细讲解,让学生打好基础。从表 3 中还可以得出,这三门课程的开课顺序应该为 D、H、E,同时尽量要将课程安排在连续的三个学期。其它规则的分析方法也是如此,决策者可以根据具体的实际情况借鉴参考。

4 结论

关联规则挖掘技术是一种非常有用的技术工具,可以广泛的应用于教学管理过程中,它能够挖掘出学生各门课程成绩之间的影响程度,找到教学中各方面影响学生学习成绩的因素,发现隐藏在成绩背后的潜在规律,帮助我们更好地了解课程的设置顺序以及课时安排是否科学合理,从而为提高学校的教学管理和人才培养质量起到积极的促进作用。

参考文献:

[1]梁循.数据挖掘算法与应用[M].北京大学出版社,2006.

[2]Liu J, Liu B, Liu J. Association Rule Mining Algorithm Based On Fuzzy Association Rules Lattice and Apriori[J]. Journal of Convergence Information Technology, 2013, 8(8):399-406.

[3]Chen W, JiaNan. Teaching analysis based on association rule mining[C]// Conference Anthology, IEEE. IEEE, 2013:1-3.

[4]韩天鹏.关联规则挖掘算法研究及其应用[D].中南民族大学,2008.

[5]Cheng M, Xu K, Gong X. Research on audit log association rule mining based on improved Apriori algorithm [C]// IEEE International Conference on Big Data Analysis. IEEE, 2016:1-7.

[6]Yang Q. The Application of Apriori Algorithm in the Analysis of Excel Skill Test Results [J]. Guide of Science & Education, 2013.

[7]李忠哗,王凤利,何丕廉,等.关联规则挖掘在课程相关分析中的应用[J].河北农业大学学报,2010,33(3):116-119.

[8]黄秋勇.基于关联规则挖掘的课程设置合理性分析[J].智能计算机与应用,2010(5):57-59.