ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 219－230
DOI:10.16337/j.1004-9037.2015.01.022
© 2015 by Journal of Data Acquisition and Processing
http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于互信息的教育数据矩阵加权正负关联模式发现

余　如[1,2]　黄名选[3]　黄丽霞[4]

(1.广西教育学院文化传播学系,南宁,530023;2.广西教育学院党政办,南宁,530023;3.广西教育学院科研处,南宁,530023;4.广西教育学院教务处,南宁,530023)

摘　要:本文将互信息模型引入教育数据关联模式挖掘,提出一种基于互信息的教育数据矩阵加权正负关联模式挖掘算法,给出与其相关的定理及其证明。本文算法克服了现有挖掘算法的缺陷,考虑了教育数据项集在学生信息数据库中具有的权值,采用新的正负关联模式评价标准,挖掘出更接近实际情况的正负关联模式。通过关联模式分析,发现教育数据中潜在有用的教育、教学规律和教育发展趋势,为教育管理、教育决策和教学改革提供科学的依据。以真实的教育数据作为实验数据测试集,实验结果表明,本文算法有效,在教育信息化数据处理与分析中具有重要的应用价值。

关键词:教育数据挖掘;关联规则;负关联规则;矩阵加权模式;互信息

中图分类号:TP391　　**文献标志码**:A

Discovery of Matrix-Weighted Positive and Negative Association Patterns from Educational Data Based on Mutual Information

Yu Ru[1,2], Huang Mingxuan[3], Huang Lixia[4]

(1. Department of Culture and Transmission, Guangxi College of Education, Nanning, 530023, China; 2. Office of Administrative Management, Guangxi College of Education, Nanning, 530023, China; 3. Office of Scientific Research Management, Guangxi College of Education, Nanning, 530023, China; 4. Teaching Affairs Office, Guangxi College of Education, Nanning, 530023, China)

Abstract: The mutual information model is introduced into the educational data association patterns mining. A new mining algorithm of the matrix-weighted positive and negative association patterns from educational data is presented based on mutual information model, and the related theorems and their proof are given. The algorithm overcomes the defects of the existing algorithms for weighted association patterns. It pays special attention to the various weights of the itemset in database, and also uses a new evaluation standard of positive and negative association patterns. Hence the positive and negative association patterns obtained from the educational data get closer to reality. Analysis on these patterns shows that, the potential educational and teaching rules, as well as educational development trend are discovered, providing a scientific basis for management, decision-making and teaching reform in education. Experiment results on real educational data demonstrate that the proposed algorithm is effective and reliable, with important potential value in the educational data processing and analyzing.

Key words: educational data mining; association rule; negative association rule; matrix-weighted pattern; mutual information

---

基金项目:国家自然科学基金(61262028)资助项目;广西自然科学基金(2012GXNSFAA053235)资助项目;广西教育厅科研(2013LX236,201203YB225)资助项目;广西高校优秀人才计划(桂教人[2011]40号)资助项目。
收稿日期:2013-11-09;**修订日期**:2013-12-24

引 言

教育数据挖掘是数据挖掘研究的一个重要应用研究分支,涉及教育学、计算机科学和统计学等多学科的研究领域,是一种从教育数据中挖掘那些事先未知的、具有很高应用价值的教育模式和教育知识的过程。近年来,随着教育信息化的迅猛发展,教育数据急剧增多,如何从这些海量的教育数据中挖掘出那些潜在的、有用的教育教学关联模式,一直是国内外学者关注和研究的课题。不同学者从不同的角度和方法对其进行研究,取得了显著的研究成果。现有的教育数据关联模式挖掘研究主要集中在以下 3 个方面:

(1)教育数据正负关联模式挖掘技术:这是传统的关联模式挖掘技术在教育信息化领域的应用。该技术只考虑项目在数据库中出现的频度,将各个项目按平等一致的方式处理。其典型挖掘算法是 Apriori 算法[1]和 FP_Growth 算法[2]以及 Wu 等[3]提出的正负关联规则挖掘算法,此外,汤春明等[4]提出了数据流闭合频繁项集挖掘算法,以及 Xu 等[5]提出能反映周期性变化规律的关联规则,拓展了数据挖掘的研究,取得了良好的挖掘效果。当前,教育数据关联模式挖掘对象主要是学校的课程、学生计划、课程成绩等教学环境数据,挖掘其数据间的相关性和学生行为模式,为教务管理、课程体系设计等提供决策支持。Kumar 等[6]系统地阐述了教育数据挖掘以及教育模式的重要性和必要性。Chaudhary 等[7]指出数据挖掘技术能有助于教育数据的数学分析。Pandey 等[8]指出教育数据挖掘技术能发现一些优秀教师的教学素养。Pal[9]和 Baradwaj[10]使用关联模式挖掘技术对学生期末考试成绩分类,有助于教师对差生给予更多的重视和辅导,控制和减少了辍学率。文献[11,12]改进传统挖掘算法,挖掘课程关联模式,并进行课程相关性分析。Borkar 等[13]采用传统挖掘算法对教学数据进行挖掘,通过模式分析预测学生后续学习以及毕业情况。传统关联模式挖掘的缺陷是:只考虑课程的选修频度,没有考虑课程之间具有不同的重要性,更没有考虑课程教学效果(即课程成绩)。

(2) 教育数据加权正负关联模式挖掘技术:该技术克服了传统关联模式挖掘的缺陷,给课程赋予的权值,以体现课程之间不同的重要性。例如,物理课程在理科有很高的地位,应赋予较高的权值,历史课程在文科也应赋予高权值。其典型挖掘算法有 Cai 等[14]提出的 MINWAL 算法和 Jiang 等[15]提出的加权正负关联规则挖掘算法。2012 年以来,教育数据加权关联模式挖掘得以关注和研究。文献[16,17]等从不同的角度提出了教育数据加权关联规则挖掘算法,均取得了较好的挖掘效果。当前,该技术的缺陷是:没有考虑学生课程教学效果(即课程成绩)。

(3)教育数据矩阵加权关联模式挖掘技术:把项目权值随着事务记录不同而变化的数据称为矩阵加权数据,也称完全加权数据,教育数据属于矩阵加权数据。其典型的算法是 KWEstimate 算法[18]和 MWARM 算法[19]。现有算法有效地解决了矩阵加权正关联模式挖掘问题,但没有解决矩阵加权负关联模式的挖掘。随着教育数据挖掘研究的不断深入,矩阵加权正负关联模式挖掘技术在教育信息化数据关联分析中具有重要的研究价值。当前,教育数据矩阵加权关联模式挖掘技术还鲜有报道。

针对上述问题,本文将互信息模型引入教育数据矩阵加权正负关联模式挖掘,提出一种新的基于互信息的教育数据矩阵加权正负关联模式发现算法。该算法充分考虑课程选修关联和重视课程教学效果,采用互信息模型衡量教育数据矩阵加权正负关联模式,取得了良好的挖掘效果。实验结果表明,本文算法有效,在教育信息化数据处理与分析中具有重要的应用价值。

1 基本概念及相关定理

典型的教育数据是学校的教务数据,例如,课程考试成绩数据等。把课程当作项目,课程考试成绩当作项目权重。设 $SD=\{s_1,s_2,\cdots,s_n\}$是学生信息数据库(Student database,SD),$s_i(1\leqslant i\leqslant n)$表示 SD 中的第 i 个学生记录,Course$=\{c_1,c_2,\cdots,c_m\}$表示所选修的课程 (course)项集,$c_j(1\leqslant j\leqslant m)$表示第 j 个

课程项目,$w[s_i][c_j]$ ($1\leqslant i\leqslant n$, $1\leqslant j\leqslant m$))表示第 i 个学生 s_i 的第 j 门课程 c_j 的成绩权值,如果课程 c_j 没有成绩,其成绩权值为 0。设 $C_1=\{c_1,c_2,\cdots,c_{m1}\}$ ($m_1<m$),$C_2=\{c_1,c_2,\cdots,c_{m2}\}$ ($m_2<m$),$C_1\subset$ Course,$C_2\subset$ Course,$C_1\cap C_2=\varnothing$,参照传统的支持度和置信度概念,给出如下基本定义。

定义 1　教育数据矩阵加权正负关联模式。教育数据矩阵加权正负关联模式指的是课程项目集(C_1, C_2)的矩阵加权正关联规则模式($C_1\rightarrow C_2$)和负关联规则模式($\neg C_1\rightarrow C_2$,$C_1\rightarrow\neg C_2$,$\neg C_1\rightarrow\neg C_2$)。

定义 2　教育数据矩阵加权支持度(Educational data matrix-weighted support,edmwsup)。参照文献[18]的完全加权支持度定义,教育数据矩阵加权项集 C 支持度为

$$\mathrm{edmwsup}(C)=\frac{w_c}{n\times k} \tag{1}$$

式中:$w_c=\sum_{s_i\in(CD)}\sum_{C_j\in C}w[s_i][c_j]$,$k$ 为项集 C 的项目个数,n 是数据库 SD 的记录总数。

设最小支持度阈值为 minsup,则教育数据矩阵加权项集 C 为频繁项集当且仅当 edmwsup(C)$\geqslant$ minsup。教育数据矩阵加权正负关联模式支持度的计算公式为

$$\mathrm{edmwsup}(\neg C)=1-\mathrm{edmwsup}(C) \tag{2}$$

$$\mathrm{edmwsup}(C_1\rightarrow C_2)=\mathrm{edmwsup}(C_1\cup C_2) \tag{3}$$

$$\mathrm{edmwsup}(C_1\rightarrow\neg C_2)=\mathrm{edmwsup}(C_1)-\mathrm{edmwsup}(C_1\cup C_2) \tag{4}$$

$$\mathrm{edmwsup}(\neg C_1\rightarrow C_2)=\mathrm{edmwsup}(C_2)-\mathrm{edmwsup}(C_1\cup C_2) \tag{5}$$

$$\mathrm{edmwsup}(\neg C_1\rightarrow\neg C_2)=1-\mathrm{edmwsup}(C_1)-\mathrm{edmwsup}(C_2)+\mathrm{edmwsup}(C_1\cup C_2) \tag{6}$$

定义 3　教育数据矩阵加权置信度(Educational data matrix-weighted confidence,edmwconf)。教育数据矩阵加权正负关联模式置信度的计算公式为

$$\mathrm{edmwconf}(C_1\rightarrow C_2)=\frac{\mathrm{edmwsup}(C_1\cup C_2)}{\mathrm{edmwsup}(C_1)} \tag{7}$$

$$\mathrm{edmwconf}(C_1\rightarrow\neg C_2)=1-\mathrm{edmwconf}(C_1\rightarrow C_2) \tag{8}$$

$$\mathrm{edmwconf}(\neg C_1\rightarrow C_2)=\frac{\mathrm{edmwsup}(\neg C_1\cup C_2)}{\mathrm{edmwsup}(\neg C_1)} \tag{9}$$

$$\mathrm{edmwconf}(\neg C_1\rightarrow\neg C_2)=\frac{\mathrm{edmwsup}(\neg C_1\cup\neg C_2)}{\mathrm{edmwsup}(\neg C_1)} \tag{10}$$

定义 4　教育数据矩阵加权强正负关联模式　当 C_1 和 C_2 满足下列 2 个条件时,关联模式 $C_1\rightarrow C_2$,$C_1\rightarrow\neg C_2$,$\neg C_1\rightarrow C_2$,$\neg C_1\rightarrow\neg C_2$ 称为教育数据矩阵加权强正负关联模式:(1) C_1 和 C_2 是频繁项集;(2)$C_1\rightarrow C_2$,$C_1\rightarrow\neg C_2$,$\neg C_1\rightarrow C_2$,$\neg C_1\rightarrow\neg C_2$ 的矩阵加权支持度大于或等于 minsup,置信度不小于最小置信度阈值(minconf)。

定义 5　教育数据矩阵加权互信息(Educational data matrix-weighted mutual information,edmwMI)。教育数据矩阵加权项集(C_1,C_2)互信息(edmwMI(C_1,C_2))是用来衡量两个矩阵加权项集 C_1 和 C_2 的相关程度,其计算公式为

$$\mathrm{edmwMI}(C_1,C_2)=\log\frac{\mathrm{edmwsup}(C_1\cup C_2)}{\mathrm{edmwsup}(C_1)\times\mathrm{edmwsup}(C_2)} \tag{11}$$

根据互信息性质,教育数据矩阵加权项集互信息有 3 种情况:①若 edmwMI(C_1,C_2)$>$0,则 C_1 和 C_2 成正相关,反之亦然;②若 edmwMI(C_1,C_2)$<$0,则 C_1 和 C_2 成负相关,反之亦然;③若 edmwMI(C_1,C_2)$=$0,则 C_1 和 C_2 无相关。

定理 1　若 edmwMI(C_1, C_2)$>$0,则 edmwMI($\neg C_1$, $\neg C_2$)$>$0,edmwMI(C_1, $\neg C_2$)$<$0,edmwMI($\neg C_1$, C_2)$<$0,反之亦然。

证明:(1)edmwMI(C_1,C_2)>0⇒edmwMI($\neg C_1$,$\neg C_2$)>0。

由 edmwMI(C_1,C_2)>0 及式(12)得

$$\frac{\text{edmwsup}(C_1 \cup C_2)}{\text{edmwsup}(C_1) \times \text{edmwsup}(C_2)} > 1 \Rightarrow \text{edmwsup}(C_1 \cup C_2) > \text{edmwsup}(C_1) \times \text{edmwsup}(C_2) \quad (12)$$

由式(1)代入式(12)可得

$$\frac{w_{c_1 \cup c_2}}{k_{12}} > \frac{w_{c_1} \times w_{c_2}}{n \times k_1 \times k_2} \quad (13)$$

式中:k_{12},k_1,k_2 分别表示项集($C_1 \cup C_2$),C_1,C_2 的项目个数,$k_{12} \geqslant 2$,$k_1 \geqslant 1$,$k_2 \geqslant 1$。

由式(1,6)和式(11)可得

$$\text{edmwMI}(\neg C_1, \neg C_2) = \log \frac{\rho + \dfrac{w_{c_1 \cup c_2}}{k_{12}}}{\rho + \dfrac{w_{c_1} \times w_{c_2}}{n \times k_1 \times k_2}} \quad (14)$$

式中 $\rho = n - \dfrac{w_{c_1}}{k_1} - \dfrac{w_{c_2}}{k_2}$。

因为 n 是 SD 的记录总数,其值都是比较大的正数,故

$$\rho \geqslant 0 \quad (15)$$

由式(13,14)以及式(15)可得

$$\text{edmwMI}(\neg C_1, \neg C_2) > 0$$

(2) edmwMI(C_1,C_2)>0⇒edmwMI(C_1,$\neg C_2$)<0。

由式(4)和式(11)可得

$$\text{edmwMI}(C_1, \neg C_2) = \log \frac{\text{edmwsup}(C_1) - \text{edmwsup}(C_1 \cup C_2)}{\text{edmwsup}(C_1) - \text{edmwsup}(C_1) \times \text{edmwsup}(C_2)} \quad (16)$$

由 edmwsup(C_1)>0,edmwsup(C_2)>0,edmwsup($C_1 \cup C_2$)>0 以及式(12)得到

$$\frac{\text{edmwsup}(C_1) - \text{edmwsup}(C_1 \cup C_2)}{\text{edmwsup}(C_1) - \text{edmwsup}(C_1) \times \text{edmwsup}(C_2)} < 1 \quad (17)$$

由式(16,17)得到 edmwMI(C_1,$\neg C_2$)<0。

同理可证 edmwMI(C_1,C_2)>0⇒edmwMI($\neg C_1$,C_2)<0。

(2)edmwMI(C_1,$\neg C_2$)<0⇒edmwMI(C_1,C_2)>0。

$$\text{edmwMI}(C_1, \neg C_2) < 0 \Rightarrow \log \frac{\text{edmwsup}(C_1 \cup \neg C_2)}{\text{edmwsup}(C_1) \times \text{edmwsup}(\neg C_2)} < 0$$

$$\Rightarrow \frac{\text{edmwsup}(C_1) - \text{edmwsup}(C_1 \cup C_2)}{\text{edmwsup}(C_1) - \text{edmwsup}(C_1) \times \text{edmwsup}(C_2)} < 1$$

$$\Rightarrow \text{edmwsup}(C_1 \cup C_2) > \text{edmwsup}(C_1) \times \text{edmwsup}(C_2)$$

$$\Rightarrow \frac{\text{edmwsup}(C_1 \cup C_2)}{\text{edmwsup}(C_1) \times \text{edmwsup}(C_2)} > 1$$

$$\Rightarrow \log \frac{\text{edmwsup}(C_1 \cup C_2)}{\text{edmwsup}(C_1) \times \text{edmwsup}(C_2)} > 0 \Rightarrow \text{edmwMI}(C_1, C_2) > 0$$

同理可以证明 edmwMI($\neg C_1$,$\neg C_2$)>0 或 edmwMI($\neg C_1$,C_2)<0 ⇒ edmwMI(C_1,C_2)>0。

证毕。

定 理 2 若 edmwMI(C_1,C_2)<0,则 edmwMI($\neg C_1$, $\neg C_2$)<0, edmwMI(C_1, $\neg C_2$)>0, edmwMI($\neg C_1$, C_2)>0,反之亦然。

证明方法类似定理 1 的证明。

2　基于互信息的教育数据矩阵加权正负关联模式挖掘算法

2.1　基本思想

基于互信息的教育数据矩阵加权正负关联模式挖掘分为 3 个阶段:(1)教育数据预处理阶段:对教育数据进行预处理,构建矩阵加权学生信息数据库(SD)和课程项目库;(2)教育数据矩阵加权频繁项集和负项集生成阶段:首先从 SD 和课程项目库中挖掘频繁 1_项集和负 1-项集;然后,从 i-项集($i\geqslant 2$)开始,候选 i-项集 C_i 由 C_{i-1} 进行 Apriori 连接[1]得到,根据文献[19]定理 1,从 C_i 中提取负 i-项集,再从余下的候选 i-项集 C_i 中挖掘频繁 i_项集和负 i-项集,计算包含 i-项集的($i+1$)_权值阈值,循环上述操作,直到 C_i 为空集即结束挖掘;(3)教育数据矩阵加权强正负关联模式产生阶段:采用支持度-置信度-互信息架构,根据定理 1 和定理 2,从频繁项集和负项集挖掘矩阵加权强正负关联模式。

2.2　支持度-置信度-互信息架构的模式评价标准

传统关联规则模式评价是基于支持度-置信度架构,其缺点是会产生相互矛盾的模式以及一些无效、虚假和冗余的模式。传统的正负关联规则模式的评价是基于支持度-置信度-相关度[15]架构,避免了相互矛盾的模式出现,但还会出现一些无效的模式。本文提出将支持度-置信度-互信息架构作为教育数据矩阵加权正负关联模式的评价标准,为关联规则挖掘研究提供新的思路,获得了良好的挖掘效果。互信息是信息论的一个概念,用来衡量两个事件的相关程度,在教育数据挖掘研究中,互信息能很好地区分正负关联规则模式。具体的评价策略是:对于矩阵加权项集(C_1,C_2),当 C_1 和 C_1 的支持度不小于 minsup 时,若 edmwMI(C_1,C_2)>0,并且 $C_1\rightarrow C_2$ 和$\neg C_1\rightarrow\neg C_2$ 的支持度和置信度都不小于其相应的阈值,则得到强关联模式 $C_1\rightarrow C_2$ 和$\neg C_1\rightarrow\neg C_2$;同理,若 edmwMI($C_1$,$C_2$)$<0$,可挖掘出强负关联模式 $C_1\rightarrow\neg C_2$ 和$\neg C_1\rightarrow C_2$。

2.3　算法描述

输入:SD 为学生信息数据库,minsup 为最小支持度阈值, minconf 为最小置信度阈值。

输出:教育数据矩阵加权强正负关联模式。

Begin

步骤 1:教育数据预处理,构建矩阵加权学生信息数据库(SD)和课程项目库,挖掘矩阵加权频繁 1_项集和负 1-项集,计算包含 1-项集的 2_权值阈值[19];

步骤 2:从 i-项集($i\geqslant 2$)开始,重复如下操作,直到候选 i-项集 C_i 为空集才结束挖掘。

(1) C_{i-1} 进行 Ariori 连接[1]得到候选 i-项集 C_i;

(2) 根据文献[19]定理 1,从 C_i 中提取负 i-项集;

(3) 计算余下 C_i 的支持度,若其大于或等于 minsup,则得到频繁 i_项集,否则,得到负 i-项集;

(4) 计算包含 C_i 的($i+1$)_权值阈值。

步骤 3:对于频繁 i_项集和负 i-项集,循环下列操作,直到频繁 i_项集和负 i-项集集合为空。

(1) 对于矩阵加权频繁项集 (C_1,C_2),在 C_1 和 C_2 的支持度$\geqslant$minsup 情况下,若 edmwMI(C_1,C_2)>0,并且 $C_1\rightarrow C_2$ 和$\neg C_1\rightarrow\neg C_2$ 的支持度和置信度分别不小于 minsup 和 minconf,则挖掘出矩阵加权强正关联规则 $C_1\rightarrow C_2$ 和强负关联规则$\neg C_1\rightarrow\neg C_2$;若 edmwMI($C_1$,$C_2$)$<0$ 并且 $C_1\rightarrow\neg C_2$ 和$\neg C_1\rightarrow C_2$ 的支持度和置信度分别不小于 minsup 和 minconf,则挖掘出强负关联规则 $C_1\rightarrow\neg C_2$ 和$\neg C_1\rightarrow C_2$。

(2) 对于矩阵加权负项集 (C_1,C_2),在 C_1 和 C_2 的支持度$\geqslant$minsup 情况下,若 edmwMI(C_1,C_2)>0,并且$\neg C_1\rightarrow\neg C_2$ 的支持度和置信度分别不小于 minsup 和 minconf,则挖掘出矩阵加权强负关联规则$\neg C_1\rightarrow\neg C_2$;若 edmwMI($C_1$,$C_2$)$<0$,并且 $C_1\rightarrow\neg C_2$ 和$\neg C_1\rightarrow C_2$ 的支持度和置信度分别大于或等于 minsup 和 minconf,则挖掘出强负关联规则 $C_1\rightarrow\neg C_2$ 和$\neg C_1\rightarrow C_2$。

步骤 4:输出教育数据矩阵加权强正负关联模式。

End.

3 实验及结果分析

3.1 实验数据

为了验证本文算法的有效性,选择来自高校教务真实的课程考试成绩数据为实验数据测试集。测试集是历届毕业生在校学习成绩,共 2 000 个学生记录,课程科目共 121 门。将课程成绩权值规范化为[0,1]的范围(即将课程成绩除以 100),对课程项目名称统一编号,如"英语翻译"课程用编号 I1 表示,"基础英语"课程用编号 I2 表示,"英语朗读技巧"课程用编号 I3 表示,构建学生信息数据库和课程项目库。

3.2 实验结果及分析

将文献[3]的项无加权正负关联规则挖掘算法(记为算法 1)作为对比算法,分别从支持度阈值变化、置信度阈值变化等 2 个方面对本文算法 (记为算法 2)和算法 1 的挖掘性能进行实验比较和分析。

3.2.1 教育数据项集模式数量比较

支持度阈值变化情况下,取项目个数为 20,2 种算法挖掘出的教育数据频繁项集和负项集模式数量比较如表 1 所示。

表 1 在不同支持度阈值下频繁项集和负项集数量比较

Table 1 Number comparison of frequent itemsets and negative itemsets at different support

minsup	频繁项集		负项集	
	算法 1	算法 2	算法 1	算法 2
0.1	369	369	0	0
0.2	369	229	0	124
0.3	24	24	20	329
0.4	24	20	20	333
0.5	24	1	20	352
0.6	1	1	19	352
合计	811	644(−20.59%)	79	1 490

表 1 表明,本文算法(即算法 2)所挖掘出的矩阵加权频繁项集数量比对比算法挖掘的无加权频繁项集数量少 20.59%,而挖掘出的矩阵加权负项集比对比算法挖掘的无加权负项集数量多了近 18 倍。

3.2.2 教育数据正负关联规则模式数量比较

(1) 支持度阈值变化情况下,置信度阈值取值为 0.1,项目个数为 20,2 种算法挖掘出的教育数据正负关联规则模式数量比较如表 2 和表 3 所示。

表 2 在不同支持度阈值下正负关联规则数量比较

Table 2 Number comparison of positive and negative association rules at different support thresholds

minsup	$C_1 \rightarrow C_2$		$C_1 \rightarrow \neg C_2$	
	算法 1	算法 2	算法 1	算法 2
0.20	2 088	1 912	2 088	2 136
0.22	1 908	1 448	1 952	2 136
0.24	1 360	104	1 952	2 136
0.26	1 258	76	1 952	2 136
0.28	658	76	1 952	2 136
0.30	52	76	1 952	2 136
合计	7 324	3 692(−49.59%)	11 848	12 816(+8.17%)

表 3　在不同支持度阈值下负关联规则数量比较

Table 3　Number comparison of negative association rules at different support

minsup	$C_1 \rightarrow C_2$		$C_1 \rightarrow \neg C_2$	
	算法 1	算法 2	算法 1	算法 2
0.20	42	0	42	0
0.22	42	0	42	0
0.24	26	0	26	0
0.26	21	0	21	0
0.28	0	0	0	0
0.30	0	0	0	0
合计	131	0	131	0

从表 3 可知,支持度阈值变化情况下,对比算法能挖掘出负规则模式$\neg C_1 \rightarrow C_2$ 和 $C_1 \rightarrow \neg C_2$,而本文算法没有挖掘出这类负模式。

(2)置信度阈值的取值从 0.1 到 0.9,支持度阈值为 0.2,项目个数为 20,2 种算法挖掘出的教育数据正负关联规则模式数量比较如表 4 和表 5 所示。

表 4　在不同置信度阈值下正负关联规则数量比较

Table 4　Number comparison of positive and negative association rules at different confidence

minsup	$C_1 \rightarrow C_2$		$C_1 \rightarrow \neg C_2$	
	算法 1	算法 2	算法 1	算法 2
0.1	2 088	1 912	2 088	2 136
0.2	2 088	1 912	2 088	2 136
0.3	2 088	1 907	2 027	2 136
0.4	2 027	1 900	2 027	2 124
0.5	1 790	1 810	2 020	2 124
0.6	1 571	1 482	2 020	2 120
0.7	1 564	1 345	1 564	2 030
0.8	1 564	1 345	1 564	1 608
0.9	1 564	1 341	1 564	1 640
合计	16 344	14 954(−8.50%)	16 962	18 054(+6.43%)

从表 4 可知,支持度阈值不变情况下,随着置信度的增加,本文算法和对比算法挖掘出的 $C_1 \rightarrow C_2$ 和 $\neg C_1 \rightarrow \neg C_2$ 数量逐渐减少。本文算法挖掘出的矩阵加权 $C_1 \rightarrow C_2$ 数量比对比算法的无加权正关联规则数量少 8.5%,而其矩阵加权$\neg C_1 \rightarrow \neg C_2$ 数量比对比算法的多 6.43%。

表 5　在不同置信度阈值下负关联规则数量比较

Table 5　Number comparison of negative association rules at different confidence

minconf	$C_1 \rightarrow C_2$		$C_1 \rightarrow \neg C_2$	
	算法 1	算法 2	算法 1	算法 2
0.1	42	0	42	0
0.2	42	0	42	0
0.3	42	0	42	0
0.4	0	0	42	0
0.5	0	0	42	0
0.6	0	0	0	0
0.7	0	0	0	0
0.8	0	0	0	0
0.9	0	0	0	0
合计	126	0	210	0

表 5 表明,在不同置信度阈值下,本文算法也没有挖掘出 $C_1 \rightarrow \neg C_2$ 和 $\neg C_1 \rightarrow C_2$ 模式,而对比算法能挖掘出这类负模式。

3.2.3 教育数据挖掘时间效率比较

(1) 支持度阈值变化情况下教育数据项集模式挖掘时间比较。项目个数为 20,支持度阈值取值 0.1 到 0.5,2 种算法挖掘教育数据频繁项集和负项集的时间比较如表 6 所示。

表 6 在不同支持度阈值下项集挖掘时间比较 s

Table 6 Comparison of time taken for mining itemsets at different support

minsup	算法 1	算法 2
0.1	2 151.650	3 642.460
0.2	2 144.489	2 129.170
0.3	90.137	2 124.309
0.4	89.606	2 120.93
0.5	89.498	2 120.206
合计	4 565.38	12 137.08 (+165.85%)

表 6 表明,支持度阈值变化情况下,本文算法挖掘教育数据项集的时间比对比算法的多165.85%,说明本文算法挖掘项集的时间效率比对比算法的低。

(2)支持度阈值变化情况下教育数据正负关联规则挖掘时间比较。置信度为 0.1,项目个数为 20,支持度阈值取值 0.2 到 0.3,2 种算法的教育数据正负关联规则挖掘时间比较如表 7 所示。

表 7 在不同支持度阈值下正负关联规则挖掘时间比较 s

Table 7 Comparison of time taken for mining positive and negative association rules at different support thresholds

minsup	算法 1	算法 2
0.20	308.475	257.765
0.22	400.343	252.252
0.24	300.363	248.201
0.26	300.84	246.965
0.28	300.209	250.557
0.30	296.417	247.762
合计	1 906.647	1 503.502 (−21.14%)

表 7 表明,支持度变化情况下,本文算法的正负关联规则挖掘时间比对比算法的少 21.14%,表明本文算法的正负关联规则模式挖掘时间效率比对比算法的高。

(3)置信度阈值变化情况下教育数据正负关联规则挖掘时间比较。支持度为 0.2,项目个数为 20,置信度阈值取值 0.1 到 0.9,本文算法和对比算法的教育数据正负关联规则挖掘时间比较如表 8 所示。

表 8 表明,置信度变化情况下,本文算法的正负关联规则挖掘时间比对比算法的少 16.87%,表明本文算法挖掘正负关联规则模式的时间效率比对比算法的高。

表 8　在不同置信度阈值下正负关联规则挖掘时间比较　s

Table 8　Comparison of time taken for mining positive and negative association rules at different confidence thresholds

minconf	算法 1	算法 2
0.1	308.475	257.765
0.2	306.410	254.301
0.3	305.714	252.300
0.4	306.386	254.343
0.5	303.828	253.844
0.6	302.828	253.157
0.7	302.705	253.343
0.8	302.542	250.07
0.9	301.127	248.476
合计	2 740.015	2 277.599 (−16.87%)

3.2.4　教育数据正负关联模式实例分析

本节对 2 种算法挖掘出来的教育数据正负关联规则模式进行合理性分析,列举了 2 种算法挖掘出的部分正负关联规则模式实例,如表 9 和表 10 所示。

表 9　本文算法挖掘的矩阵加权正负关联规则实例

Table 9　Examples of matrix-weighted positive and negative association rules mined by the algorithm proposed in this paper

序号	矩阵加权正负关联模式实例	支持度
1	英汉翻译→英文报刊阅读	0.685 4
2	英语语法→(英文报刊阅读,商务英语阅读)	0.418 3
3	(英语朗读技巧、英语语音)→英文报刊阅读	0.650 0
4	¬商务英语听力→¬会场培训英语	0.625 1
5	¬商务英语听力→¬剑桥商务英语	0.618 0
6	¬商务英语听力→¬(英语口语、国际贸易英语)	0.615 1
⋮	⋮	⋮

表 9 列举了本文算法挖掘的矩阵加权正负关联模式实例。通过模式分析发现,所列举的正负关联模式与现实情况很接近,是合理的、有效的模式。例如,学好《英语语法》《英汉翻译》《英语朗读技巧》《英语语音》等课程,能促进《英文报刊阅读》课程的学习与掌握(序号 1,2 和 3 模式),它们之间是正相关关联;没有学好《商务英语听力》,也很难学好《会场培训英语》《剑桥商务英语》《英语口语》和《国际贸易英语》

表 10　对比算法挖掘的无加权正负关联规则实例

Table 10　Examples of unweighted positive and negative association rules mined by the contrastive algorithm

序号	无加权正负关联模式实例	支持度
1	旅游概论(旅游地理	0.21
2	英语听力Ⅰ→旅游概论	0.21
3	¬旅游英语→综合英语Ⅰ	0.26
4	¬商务英语听力Ⅰ→综合英语Ⅰ	0.26
5	¬综合英语Ⅰ→(英文报刊阅读,网络英语)	0.24
6	旅游英语→¬综合英语Ⅰ	0.24
⋮	⋮	⋮

课程(序号 4,5 和 6 模式)。

表 10 列举的是对比算法挖掘的部分无加权正负关联规则模式。对其模式分析后发现,对比算法挖掘的正负关联模式中有些与现实情况不尽相符,存在一些不合理的、无效的模式,特别是挖掘出的负关联规则模式 $C_1 \rightarrow \neg C_2$ 和 $\neg C_1 \rightarrow C_2$ 对于课程成绩数据关联分析来说意义不是很大。例如,序号 1 模式表明学好了《旅游概论》课程可以有助于学好《旅游地理》课程,与现实基本相符,是一条有效模式,但是序号 2 模式的前件和后件似乎关系不是很大,互不受影响,与现实基本不符,是无效模式;学不好《旅游英语》课程或者《商务英语听力》课程就可以学好《综合英语Ⅰ》(序号 3,4 模式),学不好《综合英语Ⅰ》就可以学好《英文报刊阅读》和《网络英语》(序号 5 模式),学好课程《旅游英语》就学不好《综合英语Ⅰ》课程(序号 6 模式),这些模式都不合情理,应该是不合理模式。

3.2.5 实验结果分析

综上所述,本文算法是有效的,与现有无加权正负模式挖掘算法比较,具有以下特点:

(1) 本文算法所挖掘出的矩阵加权频繁项集和正关联规则模式 $C_1 \rightarrow C_2$ 数量比对比算法挖掘的无加权模式数量少,而挖掘出的矩阵加权负项集和负关联模式 $\neg C_1 \rightarrow \neg C_2$ 数量比对比算法挖掘的无加权模式数量多。

(2) 本文算法的正负关联规则模式挖掘时间效率比对比算法的高。

(3) 本文算法没有挖掘出形如 $C_1 \rightarrow \neg C_2$ 和 $\neg C_1 \rightarrow C_2$ 的负关联规则模式,对比算法能挖掘出这类负模式。而这类负模式对于课程成绩数据关联的分析意义不是很大。

(4)与基于项集频度挖掘的对比算法不同,本文算法是基于学生成绩权值的挖掘,能挖掘出客观反映教学效果的矩阵加权课程关联模式,通过模式分析后得到的教育、教学模式应该更客观、更合理,更接近现实情况。

主要原因分析如下:对比算法只考虑项目出现频度,不考虑课程的教学效果(即考试成绩),因而,其挖掘的模式仅能反映课程的选修关联。由于项目选修频度都比较高,对比算法挖掘的频繁项集数量就比较多,导致挖掘正负关联模式的时间效率降低。本文算法重视和考虑了课程成绩,所挖掘的正负关联模式能反映课程教学效果的关联,更能接近现实情况。另外,本文采用支持度-置信度-互信息架构评价矩阵加权正负关联模式,能避免更多无效、无趣的关联模式出现,频繁项集和正关联规则模式($C_1 \rightarrow C_2$)数量变少,减少了其挖掘正负关联模式的时间,从而提高了关联规则的挖掘效率。

实验结果也表明,本文算法还存在一些不足:其挖掘项集的时间效率比对比算法的低。主要原因是:本文算法是基于权值挖掘的,在挖掘项集时,以项集权值为运算对象的计算量比较大,计算时间开销多。而对比算法是基于频度挖掘的,以项集频度为运算对象,其计算量不如本文算法的多,因而挖掘项集的时间开销变少了。

4 结束语

教育系统的变革和发展得益于信息技术的迅猛发展以及教育信息化建设的推进,由此积累起来的海量教育电子数据具有不可估量的潜在价值,已经引起国内外教育界和计算机领域学者的极大关注和研究。针对现有教育数据关联模式挖掘存在的不足,本文将互信息模型引入教育数据正负关联模式挖掘,提出一种新的教育数据正负关联模式评价标准以及基于互信息的教育数据矩阵加权正负关联模式挖掘算法。该算法采用支持度-置信度-互信息架构作为衡量教育数据矩阵加权正负关联模式的评价标准,不仅考虑了课程项目的频度,更重视课程考试成绩,获得了良好的挖掘效果。通过关联模式分析,可

以发现教育领域潜在的教学教育规律及其发展趋势。实验结果表明,本文提出的挖掘算法是有效的。下一步的研究是扩大实验数据量,完善研究方法,从算法设计的层面研究如何减少挖掘矩阵加权项集的计算量开销,提高挖掘项集时间效率,开发一个实用的教务数据关联模式挖掘与分析系统,服务于学校的教学活动。

参考文献:

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database[C]// Proceeding of 1993 ACM SIGMOD International Conference on Management of Data. N Y, USA: ACM,1993: 207-216.

[2] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, USA: ACM, 2000:1-12.

[3] Wu Xindong, Zhang Chengqi, Zhang Shichao. Efficient mining of both positive and negative association rules[J]. ACM Transactions on Information Systems, 2004,22(3): 381-405.

[4] 汤春明,王培义,曲英涛. 在线挖掘数据流闭合频繁项集 CMNL-SW 算法[J]. 数据采集与处理,2012,27(4):508-513.
Tang Chunming, Wang Peiyi, Qu Yingtao. CMNL-SW algorithm on onl ine mining closed frequent itemsets over data stream [J]. Journal of Data Acquisition and Processing, 2012,27(4):508-513.

[5] Xu Min, Jin Yuanping, Zhu Wujia, et al. Mining cyclic generalized association rules[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2002,19(1):98-102.

[6] Kumar D V, Chadha A. An empirical study of the applications of data mining techniques in higher education[J]. International Journal of Advanced Computer Science and Applications (IJACSA), 2011,2(3):80-84.

[7] Chaudhary M P, Gupta K, Jijja A, et al. Mathematical analysis of data mining in higher education[J]. International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004), 2011,2(3):345-353.

[8] Pandey U K, Pal S. Mining data to find adept teachers in dealing with students[J]. International Journal of Intelligent Systems and Applications, 2012(3):27-33.

[9] Pal S. Mining educational data to reduce dropout rates of engineering students[J]. International Journal of Information Engineering and Electronic Business, 2012(2):1-7.

[10] Baradwaj B K, Pal S. Mining educational data to analyze students performance[J] . International Journal of Advanced Computer Science and Applications(IJACSA), 2011,2(6):63-69.

[11] 董辉. 基于兴趣度的高职课程关联规则挖掘[J]. 吉首大学学报:自然科学版,2012,33(3):41-46.
Dong Hui. Association rule mining based on the interestingness about vocational college courses[J]. Journal of Jishou University:Natural Science Edition, 2012,33(3):41-46.

[12] 李忠哗, 王凤利, 何丕廉. 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.
Li Zhonghua, Wang Fengli, He Pilian. Application of association rule of data mining in course relativity analysis[J]. Journal of Agricultural University of Hebei, 2010,33(3):116-119.

[13] Borkar S, Rajeswari K. Predicting students academic performance using education data mining[J]. International Journal of Computer Science and Mobile Computing(IJCSMC), 2013, 2(7):273-279.

[14] Cai C H, Da A, Fu W C, et al. Mining association rules with weighted items [C]// Proceedings of the 1998 International Symposium on Database Engineering & Applications. Washington, DC, USA: IEEE Computer Society, 1998: 68-77.

[15] Jiang He, ZhaoYuanyuan. Mining positive and negative association rules with weighted items[C]// Proceedings of 2008 International Symposium on Distributed Computing and Application for Business Engineering and Science(DCABES2008). Beijing, China: Publishing House of Electronics Industry ,2008: 450-454.

[16] 刘建炜,张颖. 基于加权关联规则算法的学生成绩数据挖掘研究[J]. 福建教育学院学报,2012(3):123-125.
Liu Jianwei, Zhang Ying. Research of student achievement data mining based on the weighted association rules algorithm [J]. Journal of Fujian College of Education, 2012(3):123-125.

[17] 陈世保,徐峰,吴国凤. 基于难度系数的加权关联规则在试卷评估中的应用[J]. 井冈山大学学报:自然科学版,2013,34(1):

70-74.

Chen Shibao, Xu Feng, Wu Guofeng. Weighted association rules based on the coefficient of difficulty in the assessment of papers[J]. Journal of Jinggangshan University: Natural Science Edition, 2013, 34(1): 70-74.

[18] 谭义红,林亚平. 向量空间模型中完全加权关联规则的挖掘[J]. 计算机工程与应用,2003(13):208-211.

Tan Yihong, Lin Yaping. Mining all-weighted association rules from vector space model[J]. Computer Engineering and Applications, 2003, 39(13): 208-211.

[19] 黄名选,严小卫,张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报,2009,20(7):1854-1865.

Huang Mingxuan, Yan Xiaowei, Zhang Shichao. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining[J]. Journal of Software, 2009, 20(7): 1854-1865.

作者简介:**余如**(1974-),女,主管护师,研究方向:教育信息化信息处理与挖掘,E-mail:yuru721@163.com;**黄名选**(1966-),男,教授,研究方向:文本挖掘、教育数据挖掘与信息检索;**黄丽霞**(1983-),女,讲师,研究方向:教育数据挖掘。