关联规则在学生CET4成绩中的应用*

陈 伟[1],程黄金[2]
(1.淮南联合大学 计算机系,安徽 淮南 232038;
2.淮南联合大学 基础部,安徽 淮南 232038)

摘 要:关联规则是数据挖掘的主要技术之一,是描述数据库中一组数据项之间的某种潜在关系的规则。以学生CET4成绩数据为研究对象,运用关联规则挖掘算法Apriori算法,找出学生CET4成绩中听力、阅读、写作、综合测试四部分成绩之间的关系,以及这四部分成绩与总分之间的关系。

关键词:关联规则;Apriori算法;频繁项集;数据挖掘;CET4
中图分类号:TP311 **文献标识码**:A

Application of association rule in students' CET4 scores

CHEN Wei[1], CHENG Huang Jin[2]
(1.Dept. of Computer, Huainan Union University,Huainan 232038, China;
2.Foundamental Department of Huaina Union University, Huainan 232038, China)

Abstract: Association rule is one of the key technologies of data mining. It describes the potential relations of a group of data items in database. In this paper, we study the students' CET4 scores, and use association rule mine algorithm-Apriori algorithm to find the relationship of each items of students' CET4 scores, namely, listening, reading, writing and synthes test and the relationship between these four items and the total scores.

Key words: association rule; Apriori algorithm; frequent itemsets; data mining; CET4

数据挖掘技术应用于教学管理中的主要方法是根据现有信息系统中的数据,挖掘高校教学管理工作中平常看不见、也无从知道的规律,以此来提高管理效率,帮助教师改进现有的教学方式和方法,从而增强高校的竞争优势。关联规则形式简洁,易于解释和理解,并可以有效地捕捉数据间的重要关系。

1 关联规则的原理

一般地,关联规则挖掘是指从一个大型的数据集中发现有趣的关联或相关关系,即从数据集中识别出频繁出现的属性值集,也称为频繁项集(简称频繁集),然后再利用这些频繁集创建描述关联规则的过程[1]。

关联规则可形式化定义为:

设 $I=\{i_1,i_2,\cdots,i_m\}$是由 m 个不同的项组成的集合。给定一个事务数据库 D,其中每一个事务 T 是 I 中一组项的集合,即 $T\subseteq I$,T 有一个唯一的标示符 TID。若项集 $X\subseteq I$ 且 $X\subseteq T$,则事务 T 包含项集 X。

关联规则是形如 $X=>Y$ 的蕴涵式,其中 $X\subseteq T$,$Y\subseteq T$,并且 $X\cap Y=\Phi$,如果 D 中事务包含 $X\cup Y$ 的百分比为 S,则称 S 为关联规则 $X=>Y$ 的支持度,它是概率 $P(X\cup Y)$;如果 D 中包含 X 的事务同时也包含 Y 的百分比为 C,则称 C 为关联规则 $X=>Y$ 的置信度,它是条件概率 $P(Y/X)$。习惯上将关联规则表示为 $X=>Y(S\%,C\%)$。

关联规则的发现任务或问题就是在事务数据库 D 中找出所具有用户给定的最小支持度阈值(min_ sup)和最小置信度阈值(min_cof)的关联规则,即这些关联规则的支持度和置信度分别不小于最小支持度和最小置信度,

* 基金项目:2007年安徽省高校青年教师资助计划题目(项目批号:2007jq1197)

这样，每条被挖掘出来的关联规则就可以用一个蕴涵式(X=>Y)和两个阈值(最小支持度 min_sup 和最小置信度 min_cof)表示[1]。

2 关联规则挖掘算法

2.1 Apriori 算法：使用候选项集找频繁项集

Apriori 算法通过对数据库 D 的多次扫描来发现所有的频繁项目集。在第一次扫描数据库时，对项集 I 中的每一个数据项计算其支持度，确定出满足最小支持度的频繁 1 项集的集合 L_1，然后，L_1 用于找频繁 2 项集的集合 L_2，如此下去……在后续的第 k 次扫描中，首先以 k–1 次扫描中所发现的含 k–1 个元素的频繁项集的集合 L_{k-1} 为基础，生成所有新的候选项目集 C_K（Candidate Itemsets），即潜在的频繁项目集，然后扫描数据库 D，计算这些候选项目集的支持度，最后从候选集 C_K 中确定出满足最小支持度的频繁 k 项集的集合 L_k，并将 L_k 作为下一次扫描的基础。重复上述过程直到再也发现不了新的频繁项目集为止。

2.2 由频繁项集产生关联规则

找出了所有的频繁项集，由它们产生强关联规则就很方便了(强关联规则满足最小支持度和最小置信度)。对于置信度，公式为：$\text{confidence}(X=>Y)=P(Y|X)=\dfrac{\text{support\_count}(X\cup Y)}{\text{support\_count}(X)}$，其中 $\text{support\_count}(X\cup Y)$ 是包含项集 $X\cup Y$ 的事务数。$\text{support\_count}(X)$ 是包含项集 X 的事务数。关联规则产生如下：

对于任意一个频繁项集 L 和 L 的任何非空子集 $S\subseteq L$，如果比率 $\dfrac{\text{support}(L)}{\text{support}(S)}\geq \min\_\ con$，则生成关联规则 $R:S\Rightarrow(L-S)$，且该规则的置信度和支持度分别为：$\text{confidence}(R)=\dfrac{\text{support}(L)}{\text{support}(S)}$，$\text{support}(R)=\text{support}(L)$[2]。

3 关联规则的应用

以下以 Visual Foxpro6.0 为工具进行讨论。

3.1 数据预处理

对现有的学生 CET4 成绩进行数据预处理(Data preprocessing)，包括 2 个步骤:数据清理(Data Clearing)和数据变换(Data Transformation)。

（1）数据清理：对表中的原始数据进行数据清理，消除一些冗余数据，消除噪声数据，消除重复记录。很多学生的 CET4 成绩数据都为 0,通过调查知道这些数据缺失的原因是学生未参加考试，把这些数据都从数据库表中删除。数据清理后的如图 1 所示。

（2）数据变换：将数据转换成适合于挖掘的形式。由于学生 CET4 成绩是以数字的形式给出的，不利于数据挖掘的进行，因此需对听力、阅读、写作、综合测试 4 项的连续属性值进行离散化处理，即转换为优秀、良好、中、及格、不及格 5 个等级。因为 CET4 的分值分配为：总分 710，听力 249，阅读 249，写作 142，综合测试 70，所以要把分数换算为百分制。如分数高于 85 为“优”，介于 80～85 之间为“良”，70～80 之间为“中”，60～70 之间为“及格”，60 分以下为“不及格”。“不及格”、“及格”、“中”、“良”、“优”设定为 1、2、3、4、

成绩表1

院系名称	专业名称	年级	学号	姓名	性别	总分	听力分数	阅读分数	写作分数	综合分数
机电系	电气自动化技术	05	200512101	丁凤娇	女	365	131	127	68	39
机电系	电气自动化技术	05	200512102	钱丹丹	女	342	107	130	68	37
机电系	电气自动化技术	05	200512103	王吉娜	女	291	95	98	59	39
机电系	电气自动化技术	05	200512105	万蕾	女	350	105	127	77	41
机电系	电气自动化技术	05	200512106	徐宣宣	女	351	116	121	75	39
机电系	电气自动化技术	05	200512107	胡晓慧	女	400	140	146	75	39
机电系	电气自动化技术	05	200512108	徐艳	女	311	110	92	68	41
机电系	电气自动化技术	05	200512110	邵婷	女	342	112	134	64	32
机电系	电气自动化技术	05	200512112	马茹	女	334	125	113	55	41
机电系	电气自动化技术	05	200512113	龚宁	女	423	120	163	86	54
机电系	电气自动化技术	05	200512114	冯雪	女	444	178	161	66	39
机电系	电气自动化技术	05	200512117	郭超群	男	362	142	96	81	43
机电系	电气自动化技术	05	200512118	徐冰玉	男	427	185	115	88	39
机电系	电气自动化技术	05	200512126	刘建成	男	379	127	127	88	37
机电系	电气自动化技术	05	200512128	韩娜飞	男	375	110	161	59	45
机电系	电气自动化技术	05	200512135	章成君	男	340	114	117	68	41
机电系	电气自动化技术	05	200512136	杨一	男	340	110	134	55	41
机电系	电气自动化技术	05	200512138	陈思源	男	367	135	127	70	35
机电系	电气自动化技术	05	200512139	王虎	男	354	105	144	66	39
机电系	电气自动化技术	05	200512142	王国龙	男	327	107	127	61	32
机电系	电气自动化技术	05	200512143	张斗文	男	308	105	111	53	39
机电系	电气自动化技术	05	200512203	林燕	女	371	129	146	59	37
机电系	电气自动化技术	05	200512204	刘芳芳	女	352	114	138	68	32
机电系	电气自动化技术	05	200512205	何满平	女	365	125	142	59	39
机电系	电气自动化技术	05	200512206	汪兰兰	女	369	116	138	70	45
机电系	电气自动化技术	05	200512207	李金玲	女	317	103	123	59	32
机电系	电气自动化技术	05	200512210	王菊香	女	402	125	155	81	41
机电系	电气自动化技术	05	200512211	于小云	女	373	122	142	70	39
机电系	电气自动化技术	05	200512213	王菁	女	412	131	142	79	60
机电系	电气自动化技术	05	200512214	刘拍颖	女	336	114	113	68	41
机电系	电气自动化技术	05	200512218	邵南	男	392	142	138	77	35
机电系	电气自动化技术	05	200512219	方玉胡	男	369	133	146	55	35
机电系	电气自动化技术	05	200512220	王中令	男	290	105	100	55	30

图 1 成绩视图 1

成绩表

学号	A	B	C	D	E
200512101	a1	b1	c1	d1	e1
200512102	a1	b1	c1	d1	e1
200512103	a1	b1	c1	d1	e1
200512105	a1	b1	c1	d1	e1
200512106	a1	b1	c1	d1	e1
200512107	a1	b1	c1	d1	e1
200512108	a1	b1	c1	d1	e1
200512110	a1	b1	c1	d1	e1
200512112	a1	b1	c1	d1	e1
200512113	a2	b1	c2	d2	e3
200512114	a2	b3	c2	d1	e1
200512117	a1	b1	c1	d1	e2
200512118	a2	b3	c1	d2	e1
200512126	a1	b1	c1	d2	e1
200512128	a1	b1	c2	d1	e2
200512135	a1	b1	c1	d1	e1
200512136	a1	b1	c1	d1	e1
200512138	a1	b1	c1	d1	e1
200512139	a1	b1	c1	d1	e1
200512142	a1	b1	c1	d1	e1
200512143	a1	b1	c1	d1	e1
200512203	a1	b1	c1	d1	e1
200512204	a1	b1	c1	d1	e1
200512205	a1	b1	c1	d1	e1
200512206	a1	b1	c1	d1	e2
200512207	a1	b1	c1	d1	e1
200512210	a1	b1	c2	d1	e1
200512211	a1	b1	c1	d1	e1
200512213	a1	b1	c1	d1	e5
200512214	a1	b1	c1	d1	e1
200512218	a1	b1	c1	d1	e1
200512219	a1	b1	c1	d1	e1

图 2 成绩视图 2

5；用“A”代表总分，“B”代表听力分数，“C”代表阅读分数，“D”代表写作分数，“E”代表综合测试分数；将除了学号的所有字段都改为字符型。数据变换后如图2所示，总计1 814条记录。

3.2 设计思路

3.2.1 求解频繁项集

图2中的学生成绩表.DBF为本文要研究的事务数据库，它有6个字段，均为字符型。求解频繁项集步骤如下：

(1) 建立一个项目数据表ITEM.DBF,该表中有1个字段，字段名为A，数据类型为字符型，用于存放CET4成绩中每个组成部分的所有分数段的表达值，该表中每条记录代表一种表达值，表中的记录数就是表达值形式的数目。该数据表中的记录升序排列，分别为a1、a2、a3、b1、b2、b3、c1、c2、c3、d1、d2、d3、d4、e1、e2、e3、e4、e5。

(2) 建立6个空数据表FRENQ1、FRENQ2、FRENQ3、FRENQ4、FRENQ5、FRENQ6，分别用来存放1、2、3、4、5、6频繁项集和它们的支持度计数。其中FRENQ1中有2个字段A、SUP，FRENQ2有3个字段A、B、SUP，FRENQ3有4个字段A、B、C、SUP，FRENQ4有5个字段A、B、C、D、SUP，FRENQ5有6个字段A、B、C、D、E、SUP，FRENQ6有7个字段A、B、C、D、E、F、SUP，只有SUP为数值型，其余的数据类型均为字符型。

(3) 利用成绩表产生一个辅助数据表ITEM1，该表中只有一个字段ITEMSET，数据类型为字符型，记录数与成绩数据表相同,数据为成绩表中的A+B+C+D+E的值。

(4) 在求每个频繁项目集时，分2步进行：第1步产生候选项，第2步生成频繁项目集。具体过程如下:首先，扫描ITEM表中每一条记录，对应在ITEM1.DBF求出所有的长度为1的该候选项的支持度，如果支持度大于给定的最小支持度,就把它存入FRENQ1.DBF中，直至把ITEM中的记录数扫描完为止。随后,利用FRENQ1.DBF产生长度为2的候选项,扫描ITEM1.DBF求出所有长为2的该候选项集的支持度，如果支持度大于给定的最小支持度，就存入FRENQ2.DBF中,直至扫描完FRENQ1.DBF中的记录为止。其余的以此类推,直到求出所有的频繁项目集。若发现某频繁项集的数目为零，则停止计算。最后，输出所有项目的频繁集。在该程序中依然运用了Apriori算法的性质：如果一个项集是频繁的,则它的所有子集也是频繁的[3-6]。

设定最小支持度为0.04，支持度计数为73，产生了79个频繁项集。实验结果如图3所示。

3.2.2 提取关联规则

从已经产生的频繁项集中确定它们的子集，然后根据关联规则的挖掘算法原理，假设最小置信度为30%，由程序得出350个关联规则。部分试验结果如图4所示。

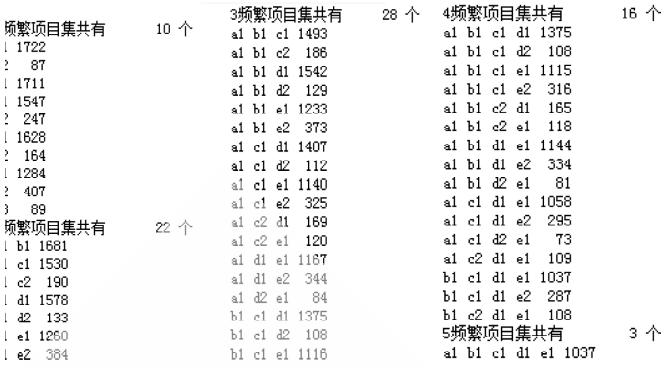

图3 求解频繁项集实验结果

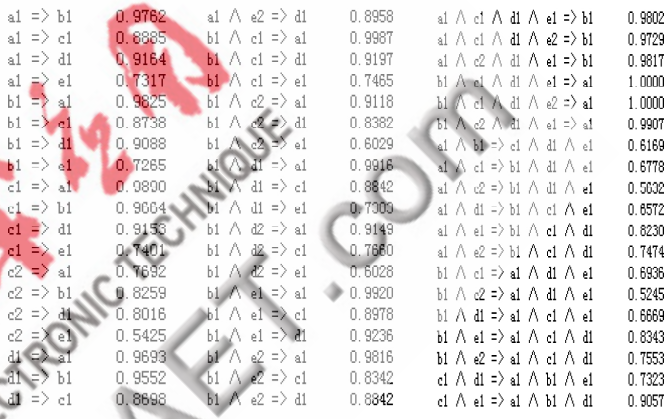

图4 提取关联规则实验结果

4 结果分析

(1) 听力、写作与总分之间的关系是双向的，即听力或写作分较低，总分一般也较低；反之，总分较低，听力或写作也较低。因为对于学生而言，一般学习听力、写作的主动性较差，而这两种题型也是一般学生考试中最棘手的题型。

(2) 阅读、综合测试和总分之间的关系主要表现为单向，即阅读、综合测试分较低，总分极有可能较低，但反之未必。这是由于CET4中的阅读和综合测试两项的分值比例相对较大引起的。

(3) 任意两项（或两项以上）得分较低，总分都较低。其中，听力和阅读是影响总分最大的两个因素。

(4) 在听力、阅读、写作和综合测试四项中，综合测试题得分与其他三项得分的关系相对较小；而听力和写作则与阅读和综合测试的关系比较紧密。

(5) 从与总分的关系，以及与其余单项的关系来看，听力、阅读、写作和综合测试四项中，听力是最突出的。

最后得出这样一个结论：在日常教学中应进一步强调听力题的重要地位，进一步加强听力的训练。

关联规则的应用很广泛。本文根据关联规则的挖掘过程，对学生CET4成绩的各个部分进行了分析。利

(下转第15页)

图 9 未压缩的“壳”处理文件(加壳)后

从图 9 观察得出，使用未压缩壳处理 calc.exe 后文件大小为 207 KB，而使用压缩壳处理后却变为了 155 KB，达到了预期的效果。

防病毒保护壳的优点为它既可以发现已知病毒又可以发现未知病毒，在一定程度上起到保护软件不被非法修改、提醒用户及时查杀病毒等作用。

缺点是病毒感染并非文件相关信息(路径名、文件名、大小)改变的惟一的非他性原因，有可能是正常程序引起的，所以，该防病毒保护“壳”会出现误报警的情况。另外，考虑到病毒的多样性，对于出现“可疑病毒”的情况，尚未进行相应处理。

参考文献

[1] 陈健伟，朱梅. 计算机病毒与反病毒技术研究[J]. 电子与电信，2006, 12(34).

[2] 张桂勇，陈芳琼. APIforWindows2000/XP 详解[M]. 北京：清华大学出版社，2003.

[3] 杨华民,梁水.Delphi 函数参考大全[M]. 北京：人民邮电出版社，2006.

(收稿日期：2008-11-30)

(上接第 9 页)

发送的报文通过转发服务器转发到现场仪表中，现场仪表根据报文中的指令，返回远程 Modbus 仪表数据报文，如图 7 所示。

图 7 系统测试报文

返回的 Modbus 报文中包含了仪表采集的现场数据，可以根据这些数据进行分析，也可以把数据保存在企业现场仪表数据库中，满足企业运行的分析、决策。

通过以上对数据交互管理平台 Modbus 协议的严格测试表明：数据交互管理对 Modbus 协议能够及时快速地响应，能够响应多客户机的访问，响应时间能够在项目要求的范围内，响应数据无错误。多台客户机可以同时对数据交互管理平台进行访问，数据交互管理平台能够及时响应多台客户机的访问。

参考文献

[1] 刘震，徐学洲. 一种基于多级分布式管理的数据采集软件模型[J]. 现代电子技术，2003，26(19)：75-77，80.

[2] 汪奇，宋煜华. 基于B/S结构的数字视频监控系统的设计与实现[J]. 计算机工程，2006，32(19)：251-252，272.

[3] 李善平，刘文峰，王焕龙. Linux与嵌入式系统[M]. 北京：清华大学出版社，2003.

[4] 陈赜. ARM9嵌入式技术及Linux高级实践教程[M]. 北京：北京航空航天大学出版社，2005.

[5] 邹思轶. 嵌入式Linux设计与应用[M]. 北京：清华大学出版社，2002.

(收稿日期：2008-11-25)

(上接第 12 页)

用 Apriori 算法，借助于计算机，可以对于海量数据进行分析，从而可以进行更为全面和客观的预测与决策。分析的结果将会对某门课程的教学提供大量有用的信息，从而指导我们的教学。

参考文献

[1] 陈文伟，黄金才. 数据仓库与数据挖掘[M]. 北京：人民邮电出版社，2004.

[2] 韩家炜. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2000.

[3] 齐晓峰. 数据挖掘技术在学生成绩管理中的应用研究[D]. 阜新：辽宁工程技术大学，2006.

[4] 赵辉. 数据挖掘技术在学生成绩分析中的研究及应用[D].大连：大连海事大学，2007.

[5] 陆楠. 关联规则的挖掘及其算法的研究[D]. 长春：吉林大学，2007

[6] 罗可，吴建华，吴杰. 一种用 Visual Foxpro 求频繁项目集的方法[J]. 计算机工程，2001(5).

(收稿日期：2008-11-17)