数据挖掘技术在课程相关性中的应用研究

陈 熔

(江苏畜牧兽医职业技术学院,江苏 泰州 225300)

【摘 要】通过数据挖掘技术对教学管理数据库中的学生成绩进行分析,找出各课程之间的隐藏关系,得到一些合理、可靠的课程关联规则,并根据这些规则进行课程的合理设置。

【关键词】数据挖掘;关联规则;课程相关性

【中图分类号】TP311.13 【文献标识码】B 【文章编号】1673-1891(2007)02-0067-03

1 引言

随着基于校园网络教学管理系统中学生成绩信息的急剧增长,现在的教务管理人员很难再像以前一样找出规律进行决策。特别是还有一些难以察觉的隐含信息,比如一门课程设置后对其他后续课程的影响,以及前、后续课程设置的顺序对教学效果的影响等等。因此必须借助于相应的工具进行数据规律及模式的发现,为决策提供决策。数据挖掘技术可以用于从大量的数据中发现隐藏于其后的规律或数据间的关系,它通常采用机器自动识别的方式,不需要更多的人工干预。采用数据挖掘技术,可以为用户的决策分析提供智能的、自动化的辅助手段。

2 数据挖掘概述

2.1 数据挖掘的基本概念

数据挖掘(Data Mining,DM)就是对观测到的数据集(通常是非常庞大的)进行分析,目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据。数据挖掘技术可以帮助人们从数据库、特别是数据仓库的相关数据集中提取出所感兴趣的知识、规则或更高层次的信息,并可以帮助人们从不同程度上去分析它们,从而可以更加有效地利用数据库或数据仓库中的数据。数据挖掘技术不仅可以用于描述过去数据的发展过程,还可以进一步预测未来趋势。这些信息是关于数据的整体特征的描述及对发展趋势的预测,在决策生成中具有重要的参考价值,从而可以很好的支持人们的决策。

2.2 数据挖掘的任务

数据挖掘的任务是发现知识,主要包括以下几类知识的发现:广义型的知识,反映同类事务共性的知识;特征型知识,反映事物各方面特征的知识;差异性知识,反映不同事物之间属性差别的知识;关联型知识,反映事物之间依赖或关联的知识;预测性知识,根据历史和当前的数据推测未来的数据;偏离型知识,揭示事物偏离常规现象。

2.3 数据挖掘的关联规则

关联规则挖掘是寻找数据项中的有趣联系,决定哪些事情将一起发生。更确切的说,关联规则通过量化的数字描述物品甲的出现对物品乙的出现有多大的影响。

设 $I=\{i_1,i_2,\cdots,i_n\}$ 是项的集合,其中的元素称为项,S 为 T 的集合,这里 T 是项的集合,并且 $T\subseteq I$。如果 $X\subseteq T$,那么称 T 包含 X。

一个关联规则是形如 $X\Rightarrow Y$ 的蕴涵式,这里 $X\subseteq I, Y\subseteq I$,并且 $X\cap Y=\phi$。规则 $X\Rightarrow Y$ 在集合 S 中的支持度(support)是 S 集中包含 X 和 Y 的数与所有项数之比,记为 support$(X\Rightarrow Y)$,即:support$(X\Rightarrow Y)=\{T:X\cup Y\Rightarrow T, T\in S\}/S$

规则 $X\Rightarrow Y$ 的可信度是指包含 X 和 Y 的数与包含 X 的数之比,记为 confidence$(X\Rightarrow Y)$,即:confidence$(X\Rightarrow Y)=\{T:X\cup Y\subseteq T, T\in S\}/\{T:X\subseteq T, T\in S\}$

关联规则挖掘的任务是:给定一个集 S,求出所

收稿日期:2007-04-06

作者简介:陈熔(1975-),男,江苏泰州人,讲师,研究方向:数据库技术和信息安全。

有满足最小支持度和最小可信度的关联规则。

3 数据挖掘技术在课程相关性的应用

3.1 数据的预处理

学校教学管理系统中的原始数据是不能直接应用于数据挖掘的，表1是某学校教务处学生成绩数据库，显然，这些信息中存在很多无用的冗余数据，而且这些无法进行数据挖掘，必须进行数据的预处理。

表1 学生成绩数据库

学号	姓名	班级代码	学年	学期	课程代码	课程名称	学分	成绩	补考1	补考2	……
*01	张*	2004*	03－04	1	0601012	数据结构	4	86			
*02	王*	2004*	03－04	1	0601012	数据结构	4	78			
*03	陈*	2004*	03－04	1	0601012	数据结构	4	95			
*04	叶*	2004*	03－04	1	0601012	数据结构	4	60			

数据预处理通常有两种方法，一种是采用横向结构，另一种是采用纵向结构。一般教学管理数据库的学生成绩均是采用纵向结构，为不破坏原有的数据库结构，我们决定采用纵向结构，且如将学生成绩转换成横向结构会花费较多的时间且在系统其他应用时会产生连接上的冗余、效率低等缺点，故放弃此方法。预处理后的表以每一个学生作为一个事务，该事务包含此学生的所有数据，如表2：

表2 处理后的学生成绩表

学号(xh)	课程代码(kch)	成绩(cj) ……
2004013101	0601011	75
2004013101	0601012	87
2004013101	0601013	84
2004013102	0601011	60

3.2 数据挖掘算法

最经典的关联规则挖掘算法是Apriori算法，其思想是利用已知的高频数据项集推导其他高频数据项集。Apriori算法是一种宽度优先算法，算法步骤如下：

①在第一次扫描中，Apriori算法计算D中所有单个项目的支持度，生成所有长度为1的1——频繁项集的集合L_1。

②如果L_{k-1}已生成，现在可用它来生成L_k。若有两个L_{k-1}，如果其前面的L_{k-2}相同，而最后一项不同，则将这样的两个L_{k-1}进行连接后得到候选k——项集的集合Ck。

③对候选k——项集C_k进行剪枝，从C_k中删除所有$(k-1)$——子集不全包含在L_{k-1}中的项集。

④扫描数据库事务D，对于其中的每一个事务，如果它包含C_k中的候选项集c，则将c的计数值加1（在扫描开始时，初始值为0）。扫描C_k，计算这些候选项集的支持度，删除其支持度低于用户给定的最小支持度的项集，最后，生成所有长度为k的频繁项集L_k。

⑤重复步骤②到④，直到L_k为空。

⑥对L_1到L_k取并集即为最终的频繁集L。

Apriori方法在由候选频繁项目集确定频繁项目集时只需扫描一遍数据库即可得到所需结果，但由于我们的数据库采用纵向结构，每个事务的数据分布在许多条纪录中，故我们改进算法，为侯选频繁项目集的每一个项目，逐遍扫描事务库，以得到所需数据，

```
L1 = find－frequent_l_itemsets(D)；
For(k = 2；Lk－1≠Φ，k + +){
    Ck = apriori_gen(Lk－1，min_sup)；
    for each c∈Ck {          // scan D for counts
        c. count + +；
    }
    Lk = {c∈Ck | c. count≥min_sup}
}
return L = ∪kLk；
procedure has_infrequent_subset(c: candidate k－
itemset；Lk－1: frequent(k－1)－itemset)；
for each (k－1)－subset s of c
if s ∉ Lk－1 then
        return TRUE；
```

return false；

3.3 挖掘结果

根据以上算法对学校教学管理数据库中相关专业的学生进行了测试，并设置最小支持度为 0.2，最小置信度为 0.5，得出先学 数据结构》对学习《语言》是有好处的，学习 数据结构》成绩优的同学再选学《语言》及格可能性大。支持度为 0.21，可信度达 0.59。可得出学习顺序建议：数据结构》⇒《语言》。

4 结束语

本文在分析了数据挖掘技术在课程相关性研究中应用的可行性之后，提出了通过关联规则进行课程相关性研究的实施方案，并对一个专业的部分课程的数据进行了挖掘，证明了通过数据挖掘关联规则对课程相关研究的实际意义，这将为学生在课程学习中进行有关的决策提供一定的帮助和参考。

参考文献：

[1]李敏．数据挖掘在辅助决策系统的应用研究[J]，微计算机信息，2004，20（5）：96－97.

[2]丁知斌，袁方．基于数据仓库的数据挖掘技术在高校学生成绩分析中的应用[J]，河北大学成人教育学院学报，2004，6（4）：19－21.

[3]欧阳辉，王员根，陈启买．关联规则在教务管理中的应用[J]，现代计算机，2006（9）：101－103.

[4]郑晓栋．数据挖掘在厦门大学研究生成绩系统中的研究与应用[J]，福建电脑，2005（1）：88－89.

[5]曲守宁，董彩云，徐德军，吴桐．关联规则算法研究及其在教学系统中的应用[J]，计算机系统应用，2005（4）：20－23.

[6]夏火松．数据仓库与数据挖掘技术[M]．北京：科学出版社，2005.

Research on Application of Data Mining about Correlation among the Courses

CHEN Rong

(Jiangsu Animal Husbandry & Veterinary College, Taizhou, Jiangsu 225300)

Abstract: Analyzing Student's achievement utilizing data mining to administer to educational manegement DB, relationship conceaing finding out between every curriculum , it is rightful to obtain some dependable association rules, moreover on the basis of these regulations carries on the rightful installation of courses.

Key words: Data mining; Association rules; Correlation among the courses

（责任编辑：张荣萍）

（上接 63 页）

Abstract: This paper introduces a set of design scheme about high－spread clone for plants in constructing modern agriculture by using MCU technique , sensor technique and modern physical agriculture technique . We studied and developed a set of high－speed clone system for plants by applying technique such as bioreactor, plant voice composes , high voltage electric field , modern illumination and so on. This system has achieved a higher economic and social effect.

Key words: MCU；Sensor；Plant voice composes；High voltage；Modern illumination

（责任编辑：张荣萍）