关联规则研究及在远程教育考试系统中的应用

郑春香,韩承双
(安庆师范学院 计算机与信息学院,安徽 安庆 246011)

摘 要:研究了关联规则分类算法,应用关联规则Apriori算法,对远程教育考试系统数据样本进行数据分析,从分析的结果中发现有价值的数据模式,寻找其中存在的关系和规则,可以为教学和考试环节发挥调节、控制、指导作用,为远程教育管理提供合理、科学的决策支持。以分类关联规则挖掘为主线,研究了数据挖掘流程中数据预处理技术、分类关联规则挖掘建模及实施应用等过程的实现。实验结果表明,该分类应用系统实现了对考试数据的自动分类,具有较好的分类运算速度。

关键词:Apriori算法;关联规则;远程教育;考试系统
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)08-0186-03

Research on Association Rule Mining and Application of Long-Distance Education System

ZHENG Chun-xiang, HAN Cheng-shuang
(Computers Department of Anqing Teachers College, Anqing 246011, China)

Abstract: Focuses on the issues of the research of association rules in classification algorithm, use association rules of Apriori algorithm and analyze long-distance education information. Find out valuable data and existent relations and rules, and then improve the teaching and provide reasonable and scientific guidance for the policy-making of the long-distance education. Based on the tool of classification association rules mining, several techniques in different phases of data mining process composed of multivariable supervised discretization algorithm, and classification association rules based modeling methods are investigated in this dissertation. According to the experiment result of classification, this program has a good capability of automatic classification and high running speed.

Key words: Apriori algorithm; association rule; long-distance education; exam system

0 引 言

远程教育作为教育的重要领域,为人才战略和终身教育发挥了重要作用。

考试是教学过程的重要组成部分,是对教与学活动过程进行考察的有效手段。目前大多数学校都建立了考试和题库系统,应用关联规则技术对远程教育考试系统数据样本进行分析,找出数据中有价值的模式和规则,可以为教学和考试环节发挥调节、控制、指导作用,为远程教育管理决策提供依据。因此,基于关联规则技术在远程教育考试系统中的应用研究有着重要的意义。

收稿日期:2008-12-10;**修回日期**:2009-03-04
基金项目:安徽省自然科学研究项目(KJ2007A124ZC);安徽省安庆市重点科技项目(20081208)
作者简介:郑春香(1973-),女,湖北红安人,硕士,讲师,研究方向为数据库知识发现、计算机应用。

1 关联规则算法研究

1.1 基本概念和问题描述

R.Agrawal等人在1993年提出了关联规则的概念和模型[1],关联规则的基本概念和问题描述如下:

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是 m 个不同项目的一个集合,每个 $i_k(k=1,2,\cdots,m)$ 称为数据项(Item),数据项的集合 I 称为数据项集(Item-set),简称为项集。其元素个数称为数据项集的长度,长度为 k 的数据项集简称为k-项集(k-Item set)。

事物 T(Transaction)是数据项集 I 上的一个子集,即 $T\subseteq I$,每个事务均有一个惟一的标识符TD与之相联,不同事务的全体构成了全体事务集 D(即事务数据库)。

设 $X\subset I$ 为数据项集,B 为事务库 D 中包含 X 的事务的数量,A 为事务库 D 中包含的所有事务的数量,则数据项集 X 的支持度(Support)定义为:

$$\text{Support}(X)=B/A$$

项集 X 的支持度 Support(X) 描述了项集 X 的重要性。

最小支持度(Minimum Support) 表示发现关联规则要求数据项必须满足的最小支持阈值,记为 minsup。支持度大于最小支持度的数据项集称为频繁项集,简称频集,反之,称为非频繁项集。

关联规则(Association Rule) 可以表示为[2]:

$R: X \Rightarrow Y$

其中,$X \subset I$, $Y \subset I$,并且 $X \cap Y = \varnothing$,它表示如果项集 X 在某一事务中出现,则必然会导致项目集 Y 也会在同一事务中出现。

对于关联规则 $R: X \Rightarrow Y$,其中,$X \subset I$, $Y \subset I$,并且 $X \cap Y = \varnothing$,规则 R 的置信度(Confidence) 定义为:

Confidence(R) = Support($X \cup Y$)/Support(x)

规则的置信度描述了规则的可靠程度。

最小置信度(Minimum Confidence) 表示关联规则所必须满足的最小可信度,记为 Minconf,它表示关联规则的最低可靠性。

1.2 Apriori 频集算法

Apriori 算法是由 R. Agrawal 等人在 1994 年提出来的关联规则挖掘的一个经典算法[3,4],后来的许多算法都是基于该算法的思想。该算法利用了两个基本性质:

(1)任何频集的子集必定是频集;

(2)任何非频繁项集的超集必定是非频繁项集。挖掘关联规则主要包含以下两个步骤[5]:

1)发现所有的频繁项集。根据定义,这些项集的频度至少应等于最小支持度(预先设置)。

2)由频繁项集产生强关联规则。根据定义,这些规则必须满足最小支持度和最小置信度阈值。

Apriori 算法的基本思路是重复扫描数据库,其核心程序描述如下:

L1 = {large 1 - item sets};

For($k = 2; L_{k-1} = \varnothing; k++$) do begin

C_k = apriori_ gen(L_{k-1})//新的候选集

For all transactions $t \in D$ do begin

C_i = subset(c_k, t)//事务 t 中包含的候选集

For all candidates $c \in C_t$

c. count ++;

end

$L_k = \{c \in C_k \mid c.\text{count} \geqslant \text{minsup}\}$

EndAnswer = $U_k L_k$;

算法首先扫描一遍数据库计算各个 1 - 项集的支持度,从而得到频繁 1 - 项集 L_1,然后采用迭代方式,逐步找出频繁 2 - 项集,3 - 项集,…,直至有某个 r 值使得 L_r 为空,算法终止。

2 网上考试系统分析与设计

网上考试系统是远程教育系统建设的重要组成部分,网上考试系统要管理考生认证登录,自动组卷,自动计时考试、提交考试数据、自动评卷判分等诸多过程。本方案以远程教育考试的应用需求为设计目标,以数据库驱动网站的概念作为设计基础,本着实用性、先进性、可扩充性、开放性、安全性、经济性的原则进行设计。系统采用 Web 方式构建,分角色进行管理,根据远程教育考试事务范围将系统进行分块构建,其中主要包含:登录模块、题库模块、组卷模块、考试模块、评卷模块等模块内容。

考试系统设计目标如图 1 所示。

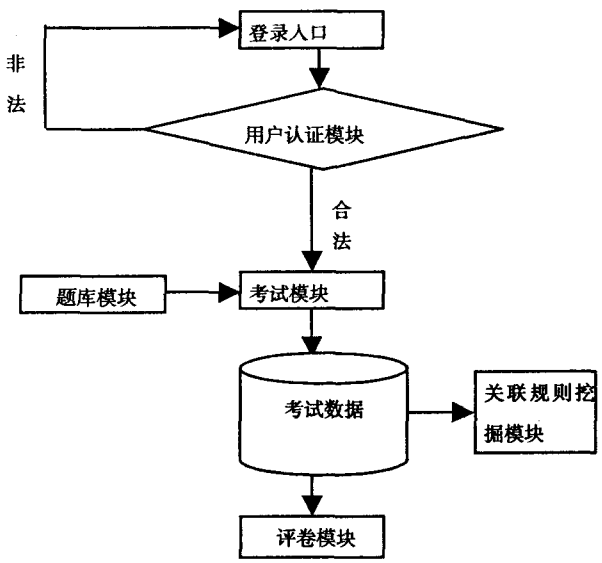

图 1　考试系统设计目标

3 关联规则技术在考试系统中的应用

目前考试系统中积累了大量的考试数据,可以开发数据挖掘模块对考试系统数据源中的数据进行分析,发现有价值的数据模式,为远程教育教学过程、教学内容的改进提供依据。

3.1 系统模型设计

根据文中研究方向,设计系统模型,主要环节有:收集数据、数据预处理模块、数据准备、程序运行、查看结果等,如图 2 所示。

3.2 数据准备

数据准备包括数据的选择、抽取,数据的转换,数据的预处理[6,7]。用于数据挖掘的源数据,可以使用 Microsoft SQL Server 2000 或者 Oracle 来建立数据仓库。考虑到目前拥有以及未来可能拥有的数据量,选择了 Microsoft SQL Server 2000。由于获得的样本数

据有的存储在 Microsoft SQL Server 2000 数据库中，有的存储在 Microsoft Access 2000 数据库中，因此，需要将 Microsoft Access 2000 数据库中的数据导入到 Microsoft SQL Server 2000 数据库中。

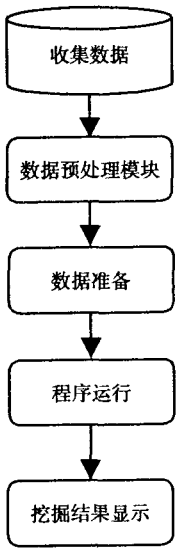

图 2 基于关联规则的考试数据挖掘模型

3.3 利用 Apriori 算法进行挖掘

利用 Apriori 关联规则挖掘算法[8]，输入：事务数据库 D 数据表和最小支持度阈值 minsup，输出：事务数据库 D 中的频繁项集 L 数据表。

以一个二维关系数据表作为事务数据表（见表 1），一维表示题目，这里共有 8 道题，二维表示考生记录，分别使用 $u_1, u_2, \cdots$ 标识，事务数据库计数为 10。

表 1 事务数据库 D

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
U_1	8	15	12	10	17	9	17	32
U_2	11	12	10	15	23	7	15	30
U_3	12	17	9	13	21	7	14	27
U_4	17	19	16	17	20	11	18	2
U_5	12	14	15	13	19	9	20	18
…	…	…	…	…	…	…	…	…
U_{10}	15	14	11	16	19	10	17	18

3.4 模式分析

根据给定的最小置信度 minconf，求出强关联规则，表 2 中给出部分给定 minsup 和 minconf 条件下挖掘出的强关联规则表。

可以对 Apriori 算法挖掘出来的频繁项集进行分析，找出一些有用的模式和规则。例如，实验中，考生事务最小支持度 minsup = 30%，最小置信度 minconf = 90%，有 30% 的考生如果第 1、第 4 题出错，那么第 6 题出错的可能性有 90%。可以分析，第 1、第 4、第 6 题考察的知识点有较强的关联性，可以将分析结果反映给相关的职能部门，对教学和考试内容的开展起到一定的辅助决策作用。

表 2 强关联规则表

编号	minsup	minconf	强关联规则
1	30%	80%	$L_1 \Rightarrow l_6$
2	30%	90%	$L_1 l_4 \Rightarrow l_6, l_4 l_6 \Rightarrow l_1$
3	50%	60%	$L_2 \Rightarrow l_3$
…	…	…	…

4 结束语

文中主要研究了数据挖掘中的关联规则技术及其在远程教育考试中的应用问题。以远程教育考试数据为模型，按照数据挖掘的基本步骤，利用关联规则 Apriori 算法，使用已有的远程考试经验数据分析提取规则，为以后的教学和考核提供合理的、科学的决策支持。

参考文献：

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database[C]//Proceedings of the ACM SIGMOD conference on management of data. New York: ACM, 1993: 207 - 216.

[2] 刘独玉，杨晋浩，钟守铭. 关联规则挖掘研究综述[J]. 成都大学学报，2006(3): 54 - 58.

[3] Agrawal R, SriKant R. Fast algorithms for mining association for mining association rules[C]//Proceedings of the 20th international Conference on very large database. [s. l.]: Morgan kaufman Pub inc, 1994: 487 - 499.

[4] 何中胜，庄燕滨. 基于 Apriori & Fp - growth 的频繁项集发现算法[J]. 计算机技术与发展，2008，18(7): 44 - 46.

[5] Han Jiawei, Kamber M. Data mining - concepts and techniques [M]. Sam Francisco: Morgan Kaufmann publishers, 2001: 232 - 236.

[6] 陈文庆，许 棠. 关联规则挖掘 Apriori 算法的改进与实现[J]. 微机发展（现更名：计算机技术与发展），2005，15(8): 155 - 157.

[7] 杨键兵. 数据挖掘中关联规则的改进算法及其实现[J]. 微计算机信息，2006(7): 195 - 197.

[8] 刘林东，曾小宁. Apriori 算法在网上考试系统中的应用[J]. 广东教育学院学报，2005(5): 103 - 108.