

分类号:

单位代码: 10019

密 级:

学 号: s030999

中国农业大学

学位论文

数据挖掘技术在教学信息管理中的应用研究

The Application and Research of Data Mining in Teaching Information Management

研 究 生: 李湘军

指 导 教 师: 黄燕 副教授

合 作 指 导 教 师:

申请学位门类级别: 工学硕士

专 业 名 称: 计算机应用技术

研 究 方 向: 网络技术、数据挖掘技术

所 在 学 院: 信息与电气工程学院

2006 年 5 月

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国农业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：

时间：

年 月 日

关于论文使用授权的说明

本人完全了解中国农业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。同意中国农业大学可以用不同方式在不同媒体上发表、传播学位论文的全部或部分内容。

(保密的学位论文在解密后应遵守此协议)

研究生签名：

时间：

年 月 日

导师签名：

时间：

年 月 日

摘要

随着高校的不断扩招，积累了越来越多的历史教学数据，管理和检索这些数据变得越来越困难。如何合理有效地利用这些数据中隐藏的信息，更好地为高校的教学、科研和管理工作服务，是现在和今后教育教学领域研究的重要课题。

数据挖掘是一种从大量数据中提取有用信息的技术，是当前计算机科学研究的前沿领域。利用数据挖掘技术能够从大量的教学信息中发现诸多教育因素间潜在的相互作用和影响关系，从而探索出合理的数据挖掘模型和有效的教学方式。本文运用关联分析、决策树、回归分析等方法建立了学生选课特征关联模型，课程间成绩影响关联模型，公共课影响因素决策树模型等教学信息挖掘模型。

OLAM 技术结合了数据挖掘和 OLAP 技术的优势，能快速有效地对数据进行挖掘，是数据库与数据挖掘领域发展的重要方向。本文探讨了基于 Web 的 OLAM 系统的体系架构与实现，并应用于教学信息挖掘，获取隐藏的信息，为教学决策提供辅助和指导。

关键词：数据挖掘，OLAP，OLAM，B/S

Abstract

As the enrolment of colleges enlarging, there are more and more convenient teaching data, which are more and more difficult to manage and search. It has been an important task for teaching research field now and later, how to make use of the information in them for teaching, research and management.

Data Mining has been an active area of computer science research recently, which is a technique to extract interesting information from a large amount of data. The mutual possible relations among many educational factors can be discovered by data mining technology, so that the reasonable model and effective teaching way could be found. This paper creates some models such as association model of choosing courses, association model of score, tree model of required courses score factors, by using association, decision tree, and regression methods.

OLAM technology can provide fast and effective data mining service for teaching information with both virtues of Data mining and OLAP. This is an important direction in which the database and data mining are going. This paper discusses the architecture and realization of web-based OLAM teaching information system, and gets hidden information by applying in teaching information mining, which provides assistance and guidance for decision-making in teaching.

Key words: Data mining, OLAP, OLAM, B/S

目 录

第一章 绪论	1
1.1. 研究目的和意义	1
1.2. 国内外研究现状	1
1.3. 研究内容和方法	2
1.4. 论文的结构	3
第二章 数据挖掘技术	4
2.1 数据挖掘概述	4
2.2 关联规则	6
2.3 决策树	9
2.4 回归分析	10
第三章 在教学信息分析中的应用	11
3.1 教学信息数据仓库	11
3.2 教学信息模型的建立	15
3.3 公共课影响因素决策树模型	16
3.4 课程间成绩影响关联模型	22
3.5 学生选课特征关联模型	25
3.6 必修课成绩回归模型	28
第四章 基于 WEB 的数据挖掘系统	31
4.1 系统的设计	31
4.2 系统的实现	32
第五章 结论与建议	42
参考文献	43
致谢	46
个人简介	47

第一章 绪论

1.1. 研究目的和意义

目前,各高校已普遍实现用数据库管理系统对教学信息进行管理。目前的数据库系统虽然能较好的实现数据的录入、查询和统计等功能,但无法支持对数据背后重要信息的挖掘。1999 年实施高校扩招以来,我国各高校招生规模逐年扩大,学生信息也急剧增加,这带来了一系列问题:一方面,对大量的学生信息进行有效地分类、组织和检索越来越困难;另一方面,积累了大量的历史数据,而其中蕴含着宝贵的教育信息。如何在海量数据中得到可靠实用的信息,并将其灵活地指导教学活动,保证教学质量,成为各高校面临的一大挑战。

数据仓库是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集,可以用以支持企业或组织的决策分析处理[1]。建立在数据仓库基础上的联机分析处理(OLAP)技术能够根据分析人员的要求快速灵活地进行海量数据的复杂查询处理,并且以一种直观易懂的形式将查询结果提供给决策制定人,以便他们准确掌握的相关数据。

数据挖掘(Data Mining)是一个利用各种分析方法和分析工具在大规模海量数据中建立模型和发现数据间关系的技术,这些模型和关系可以用来做出决策和预测[2]。数据挖掘技术和 OLAP 技术已成功地应用于商业、工业等领域,但在教育领域的应用还刚刚起步。数据挖掘技术能通过丰富的分析方法,从教学历史数据中获取隐藏的教学信息,为教学决策提供辅助和指导。

而开发基于 Web 的教育信息数据挖掘系统,能为教学信息挖掘提供一个操作方便且统一的用户界面,使其更加快捷便利地享受强大的数据挖掘服务。

OLAM(On-Line Analytical Mining,联机分析挖掘)技术是一种结合了 OLAP(On-Line Analytical Processing,联机分析处理)技术和 Data Mining(数据挖掘)技术两者优势的数据挖掘领域的新技术,能高效地从数据库中对存储的海量数据进行抽取、处理和多方面分析,发现隐藏的关系和规律,从而辅助决策和预测。而基于 Web 的 OLAM 技术,能使用户只需通过 Web 浏览器就能跨空间地享受数据挖掘服务。

将结合了 Data Mining 和 OLAP 技术的 OLAM 技术应用于教育领域,并把挖掘模型用于高校教学评估和预测,是改善高校面临局面的一种有益的尝试。

通过本课题得到的教育数据挖掘模型,将为数据挖掘在教育领域进一步的应用积累了经验。将 OLAM 技术引入教育领域,并应用于高校教学评估和预测,有助于改进高校教学工作。而开发基于 Web 的 OLAM 教育挖掘系统,也是此技术在教学信息管理领域的一次尝试。

1.2. 国内外研究现状

1989 年在第十一届国际联合人工智能学术会议上正式提出数据挖掘(Data Mining)的概念[3]。数据挖掘(DM)是一个利用各种分析方法和分析工具在大规模海量数据中建立模型和发现数据间关系的过程,这些模型和关系可以用来做出决策和预测。数据挖掘的主要功能包括关联分

析、聚类分析、概念描述以及分类和预测功能等。经过十几年的发展,数据挖掘已广泛应用于电信[4][5][6]、保险[7][8][9]、电力[10][11]、医疗[12][13]等行业。

1993年 E.F.Codd 提出联机分析处理技术(OLAP)[14]。联机分析技术可使企业数据人员、企业经理及企业其他管理人员通过对企业信息的多种可能的观察角度进行快速、一致和交互性的存取,以获得对信息的深入理解。OLAP 技术有两个主要的特点:一、在线性(On_Line),表现为对用户请求的快速响应和交互式操作,它的实现是由客户机/服务器体系结构完成的;二、多维分析(Multi_Analysis),系统能够提供对数据分析的多维视图和分析,包括对层次维和多层次维的支持。基于数据仓库的 OLAP 技术研究及应用方兴未艾[15][16][17][18][19][20]。

OLAP 是验证型分析工具,由用户驱动;数据挖掘是挖掘型分析工具,由数据驱动。OLAP 的分析结果可以数据挖掘提供指导,而在数据挖掘的结果中进行 OLAP 分析能拓展其分析深度。两者相辅相成,结合起来能提供更加优质的数据分析和功能决策。1997年 J.W.HAN 提出了联机分析挖掘(OLAM)的概念[21]。

OLAM 建立在多维数据库和 OLAP 的基础之上,因此基于超立方体的高性能挖掘算法应是其核心所在。在文献[22]中提出了基于一维数组的高效数据立方体,并由它构建一种 HOLAP,在其基础上提出了关联规则的挖掘算法。这种 HOLAP 实现了快速性和灵活性的平衡,同时也为数据挖掘提供了较好的数据空间。

目前已形成了一些 OLAM 的模型。文献[23]中提出的基于影响域(Influence Domain)的 OLAM 模型,将数据挖掘和 OLAP 技术结合在统一的框架之中,使得数据挖掘能够在数据库或数据仓库的不同部位或抽象级别上进行。影响域是一种广义的数据立方体。立方体上计算的是聚合

(Aggregation),而影响域上计算的是蕴涵(Implication),即数据中隐藏的模式。影响域同立方体一样具有属性和值,不同点在于它具有置信度(Confidence)。立方体将维映射至置信度。因此影响域更适合于 OLAM 分析。目前 OLAM 和 OLAP 都基于客户机/服务器模式,基于 Web 的 OLAM 是 Web 数据库技术下一步发展的目标[24]。国外已有研究将 OLAM 技术应用于管理和挖掘临床信息[25]。

目前,数据挖掘技术在教育方面的应用还处于起步阶段,2004年10月在美国拉斯维加斯召开的第7届 Data Mining Technology Conference 会议首次将 Data Mining in Education 纳入大会议题。传统对教育数据的分析,主要使用如 SAS 编程等方法进行简单的统计分析[26][27]。近年来才开始将 OLAP 和 DM 应用于教育领域[28][29][30]。随着研究的深入,OLAP 和数据挖掘技术在教育领域一定会得到广泛的应用,并发挥巨大的作用。

1.3. 研究内容和方法

教学信息数据库中含有大量的历史数据,其中蕴藏着各种规律和丰富的参考信息,将这些规律和信息揭示并展现出来,应用于实际教育教学中,能极大地促进教育的发展。

本课题结合中国农业大学信息与电气工程学院的实际教学数据,建立教学信息数据仓库,利用开发的数据挖掘系统和评估与预测模型,对教学数据进行挖掘。

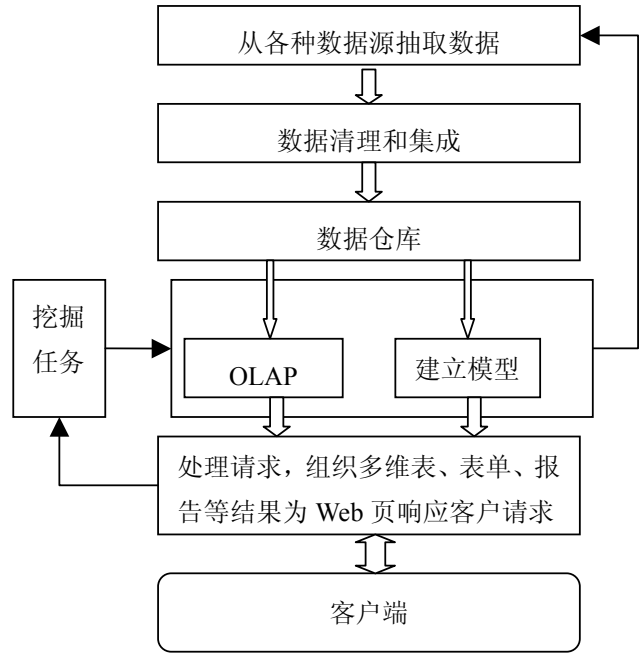


图 1-1、技术路线

研究中将在深入研究教学评估指标和体系的基础上，探讨建立教学评估与预测模型的方法，以及为教育领域建立基于 Web 的 OLAM 系统的可行性，为在新形势下推进高校教育改革提供参考和借鉴。

1.4. 论文的结构

第一章是绪论部分。简要介绍了数据挖掘和 OLAP 的相关概念、研究现状和热点，以及本文研究的内容和方法。

第二章介绍了数据挖掘技术的理论基础，重点论述了研究中用到的相关方法，包括关联规则、决策树、回归分析等常用数据挖掘方法。

第三章教学信息数据仓库的建立与 OLAP 的应用，并详细介绍了在此基础上运用数据挖掘的方法建立的公共课影响因素决策树模型，学生选课特征关联模型，课程间成绩影响关联模型等教学信息挖掘模型。

第四章基于 Web 的 OLAM 系统的设计与实现。详细介绍了基于 Web 的 OLAM 系统的体系结构，以及针对实际教学需要设计的模型与实现的方法。

第五章总结和展望。总结数据挖掘的方法及应用，展望了今后的工作方向。

第二章 数据挖掘技术

2.1 数据挖掘概述

2.1.1. 引言

随着信息技术的快速发展,人们要面对越来越庞大的数据。加州伯克利分校研究人员研究表明,全球信息生产量在 1999 年到 2002 年 3 年间以平均每年 30%左右的速度递增,2002 年中达到五万亿兆字节,这足以填满 50 万座美国国会图书馆[31]。“信息爆炸”时代的到来,使得要从数量庞大、纷繁复杂的数据中寻找有价值信息的使用者深感头疼。人们的目光转向了数据挖掘(Data Mining)技术。

一般说来,数据挖掘(DM)是一个利用各种分析方法和分析工具在大规模海量数据中建立模型和发现数据间关系的过程,这些模型和关系可以用来做出决策和预测。例如:超市分析交易数据,可以安排货架上货物摆布,以提高销售;信用卡公司分析信用卡历史数据,判断哪些人有风险,哪些没有;广告公司通过分析人们购买模式,估计他们的收入和孩子数目,作为潜在的市场信息;税务局则可分析不同团体交所得税的记录,发现异常模型和趋势。数据挖掘还有其他叫法如数据挖掘和知识发现(DMKD)、数据库中知识发现(KDD)、数据融合(Data Fusion)等等,但在产业界和研究界更加流行数据挖掘和数据库中知识发现的叫法。

数据挖掘涉及多种学科领域,包括数据库、人工智能、数理统计、神经网络、可视化、并行计算等。在电子数据处理的初期,人们就曾试图通过机器学习等领域的方法来实现自动决策支持,但收效不大。后来随着神经网络技术的形成和发展,人们的注意力又转向知识工程。80 年代人们又在新神经网络理论的指导下,重新回到机器学习的方法上,并将其成果应用于处理大型商业数据库。KDD(数据库知识发现),即数据挖掘就于 20 世纪 80 年代后期出现了[3]。数据挖掘技术的发展紧随着数据库技术的发展。60 年代,数据库技术还处于数据收集和静态数据访问阶段,以后逐渐演化到复杂的数据库系统。70 年代以后,又从层次和网络数据库发展到关系数据库。80 年代中期以来,数据仓库(Ware House)由于其面向主题、集成性、时变性和非易失性的特点,已成为数据分析和联机分析处理的重要平台[32],这为数据挖掘的蓬勃发展奠定了基础。

数据挖掘从 1989 年第十一届国际联合人工智能学术会议上正式提出以来,学术界就没有中断过对它的研究。国际 KDD 组委会于 1995 年把专题讨论会更名为国际会议,深入地探讨发现方法和系统应用,1997 年第 3 届 KDD 国际学术大会上还进行了数据挖掘工具的竞赛评奖活动,2003 年 ACM-SIGKDD 在华盛顿组织了第 9 届知识发现与数据挖掘国际会议(KDD'03)。数据挖掘在学术界和工业界的影响越来越大,本世纪将会继续成为计算机科学界的热点。

2.1.2. 数据挖掘的功能

数据挖掘通过预测未来趋势及行为,做出预测性的、基于知识的决策。数据挖掘的目标是从数据库中发现隐含的、有意义的知识,按其功能可分为以下几类:

1. 关联分析

关联分析能寻找到数据库中大量数据的相关联系,常用的两种技术为关联规则和序列模式。关联规则可用于如分析客户在超市买牙刷的同时又买牙膏的可能性;序列模式分析则如买了电脑的顾客会在三个月内买杀毒软件。

2. 聚类

聚类就是将数据对象分组为多个类或簇,使得在同一个簇中的对象之间具有较高的相似度,而在不同簇中的对象差别很大。聚类增强了人们对客观现实的认识,是概念描述和偏差分析的先决条件。聚类分析有广泛的应用,包括市场或客户分割、生物学研究、空间数据分析等方面,但 Jessica Lin 等人新的研究认为对时间序列流的聚类是毫无意义的[33]。

3. 自动预测趋势和行为

数据挖掘通过对数据库中的数据进行分类和预测,可以自动地提出描述重要数据类的模型或预测未来的数据趋势。这在商业界的应用很广,包括信誉证实、选择购物和性能预测等。一个典型的例子是市场预测问题,数据挖掘利用原有的销售记录来预测新推出的产品的销售情况等。

4. 概念描述

数据库中存在着丰富的数据,但人们总希望能以简洁的描述形式来描述汇集的数据集。概念描述就是对某类对象的内涵进行描述并概括出这类对象的有关特征。概念描述分为特征性描述(characterization)和区别性描述(discrimination),前者描述某类对象的共同特征,后者描述不同类对象之间的区别。生成区别性描述的方法很多,如决策树方法、遗传算法等。

5. 偏差检测

数据库中的数据常有一些异常记录,从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识,如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。这常用于金融银行业中检测欺诈行为,或市场分析中分析特殊消费者的消费习惯。

2.1.3. 数据挖掘的过程

从广义的数据挖掘的定义而言,典型的数据挖掘系统由以下六部分组成: 1.数据库、数据仓库或其他类型的信息库。2.数据库或数据仓库服务器。3.数据挖掘引擎。4.知识库。5.模式评估。6.图形用户界面。图 1-2 显示出了典型的数据挖掘系统的结构。



图 2-1、数据挖掘系统的结构

确切地说这里指的是数据库知识发现（KDD）的过程，数据挖掘被看作整个过程的一个关键步骤。数据挖掘专家 Jiawei Han 将知识发现的步骤概括为[34]:

1. 数据清理（消除噪声或不一致数据）。
2. 数据集成（多种数据源可以组合在一起）。
3. 数据选择（从数据库中检索与分析任务相关的数据）。
4. 数据变换（数据变换或统一成适合挖掘的形式）。
5. 数据挖掘（基本步骤，使用智能方法提取数据模式）。
6. 模式评估（根据某种兴趣度度量，识别表示知识的真正有趣的模式）。
7. 知识发现（使用可视化和知识表示技术，向用户提供挖掘的知识）。

2.2 关联规则

关联规则挖掘是数据挖掘中最常用的挖掘方法之一。关联规则用于从海量数据中发现一个事物与其他事物间的相互关联性或相互依赖性。

关联规则是由 Agrawal、Imielinski、Swami 首先提出的，是数据挖掘研究的重要方法之一，而文中提到的“规则”则指运用某种挖掘方法挖掘得到的数据内在的规则（rules）。

设 $I=\{i_1, i_2, \dots, i_m\}$ 是项的集合。设任务相关的数据 D 是数据库事务的集合，其中每个事务 T 是项的集合，使得 $T \subseteq I$ 。每一个事务有一个标识符，记作 TID。

[定义 1]: 给定一个记录集 $D = \{(I, D, T)\}$, $T \subseteq I$, 如果 $A \subseteq I$, $B \subseteq I$, 且 $A \cap B = \emptyset$, 则称 $A \Rightarrow B$ 为关联规则。关联规则 $A \Rightarrow B$ 解释为: 满足 A 中条件的数据库元组多半也满足 B 中的条件。 $A \Rightarrow B$ 在 D 中具有支持度 s 和置信度 c , 其中 s 是事务集 D 中包含 $A \cup B$ (即 A 和 B 二者) 的百分比, c 是 D 中包含 A 事务同时也包含 B 事务的百分比。即:

$$\text{support}(A \Rightarrow B) = P(A \cup B)$$

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

[定义 2]同时满足最小支持度(min_sup) 和最小置信度(min_conf) 的规则称作强关联规则。

[定义 3]: 项的集合称为项集(itemset)。包含 k 个项的项集称为 k -项集。如果项集满足最小支持度阈值, 则称它为频繁项集(frequent itemset)。满足最小支持度 s 的项集称为频繁 k -项集。

由以上定义的关联规则的挖掘步骤分为两步[34]: 1、找出所有频繁项集。就是找出数据库事务集合中满足最小支持度的项集。2、由频繁项集产生强关联规则。这些强关联规则必须满足最小支持度和最小置信度。

找频繁项集的基本算法是 Agrawal 提出的 Apriori 算法[35], 这是目前最具影响力的数据挖掘算法。这种算法的基本思想是使用逐层搜索技术探查 Apriori 性质 (频繁项集的所有非空子集都必须是频繁的): 经过第 k 次迭代($k > 1$), 它根据频繁 k -项集, 形成频繁($k+1$)-项集候选, 并扫描数据库一次, 找出完整的频繁($k+1$)-项集 L_{k+1} 。

具体算法如下:

输入:事务数据库 D , 最小支持度阈值 min_sup

```

输出:  $D$  中频繁项集  $L_i$  ( $i=1,2,\dots,k$ )
 $L_1 = \text{find\_frequent\_1-itemsets}(D)$ 
for ( $k = 2$ ;  $L_{k-1} \neq \emptyset$ ;  $k++$ )
{
     $C_k = \text{apriori\_gen}(L_{k-1}, \text{min\_sup})$ ;
    for each transaction  $t \in D$ 
    { // scan  $D$  for counts
         $Ct = \text{Subset}(C_k, t)$ ; // get the subset  $s$  of  $t$  that are candidates
        for each candidate  $c \in Ct$ 
             $c.\text{count}++$ ;
    }
     $L_k = \{c \in C_k \mid c.\text{count} \geq \text{min\_sup}\}$ 
}
return  $L = \bigcup_k L_k$ ;

procedure apriori_gen( $L_{k-1}$ :frequent ( $k-1$ )-itemsets; min_sup: minimum support threshold)
    for each itemset  $I1 \in L_{k-1}$ 
        for each itemset  $I2 \in L_{k-1}$ 
            if ( $I1[1] = I2[1] \wedge I1[2] = I2[2] \wedge \dots \wedge I1[k-2] = I2[k-2] \wedge I1[k-1] = I2[k-1]$ ) then {
                 $c = I1 \cup I2$ ; // join step: generate candidates
                if has_infrequent_subset( $c, L_{k-1}$ ) then
                    delete  $c$ ;
                else add  $c$  to  $C_k$ ;
            }
    return  $C_k$ ;

procedure has_infrequent_subset( $c$ :candidate  $k$ -itemset;  $L_{k-1}$ :frequent ( $k-1$ )-itemset)
    // use prior knowledge
    for each ( $k-1$ )-subset  $s$  of  $c$ 
        if  $s \notin L_{k-1}$  then
            return TRUE;
    return FALSE
    
```

在实际应用中,针对不同的挖掘任务和应用领域,用户关注的规则是不同的。然而随着数据仓库中数据的不断增加,关联规则挖掘得到的规则越来越多,最终用户将面对堆积如山的规则,而有用的规则往往淹没在用户不感兴趣的规则的海洋中。因此,如何从大量的规则中选取用户关注的规则,是人们在数据挖掘实际应用中经常面临的问题,也是目前数据挖掘应用研究的一个重要内容。

在实际应用中,针对不同的挖掘任务和应用领域,用户关注的规则是不同的。然而随着数据仓库中数据的不断增加,关联规则挖掘得到的规则越来越多,最终用户将面对堆积如山的规则,

而有用的规则往往淹没在用户不感兴趣的规则的海洋中。因此，如何从大量的规则中选取用户关注的规则，是人们在数据挖掘实际应用中经常面临的问题，也是目前数据挖掘应用研究的一个重要内容。

在整个数据挖掘过程中实施一定的约束，就能大大提高挖掘的效率，同时获得用户关注的规则。关联规则挖掘中主要的约束包括：

- 1) 数据约束：指定任务相关的数据集、变量名。
- 2) 兴趣度约束：指定规则兴趣度阈值或统计度量，如支持度和置信度。
- 3) 规则约束：指定规则的形式，如规则的最大项数，规则前件与后件要关注的指定谓词的最大或最小个数，或属性、属性值和聚集之间的联系等。
- 4) 先验知识约束：通过领域知识对规则进行度量。

基于上述约束的关联规则挖掘，在用户提供的各种约束指导下进行挖掘，从而得到用户感兴趣的规则。

目前，约束关联规则挖掘方法的研究主要集中在兴趣度约束方面。

支持度与置信度：大部分的关联规则挖掘都使用支持度—置信度框架。支持度反映了发现规则的有用性，置信度由 Agrawal 提出[36]，反映了挖掘规则的确定性。最小支持度表示一组事务集在统计意义上需要满足的最低程度；最小置信度反映了规则的最低可靠度。在支持度—置信度框架下，置信度只是给定了前件与后件 A、B 的条件概率的估计，并不度量它们之间蕴涵的实际强度。

作用度：作用度 (lift) 是规则的置信度与期望置信度的比值。使用最小支持度和置信度阈值排除了一些无趣的规则，但强关联规则也可能是无趣的并可能引起误导。引入作用度能有效地过滤掉误导的强关联规则。从以下推导可以看出，作用度反映了规则前件与后件的相关性。

$$\text{lift}(A \Rightarrow B) = \text{confidence} / \text{expected confidence} = P(B|A) / P(B) = P(A \cup B) / (P(A)P(B)) = \text{corr}(A, B)$$

因此，如果 $\text{lift} < 1$ ，则 A 和 B 的出现是负相关的；如果 $\text{lift} > 1$ ，则 A 和 B 的出现是正相关；如果 $\text{lift} = 1$ ，则 A 和 B 的出现相互独立，没有相关性。只有 lift 大于 1 的规则才是有趣的，其他的强关联规则可以认为是误导的。规则的统计独立性在 Piatesket Shapiro 的文献[37]中有深入研究。

J-Measure 函数：J-Measure 函数的定义为：

$$J(B; A) = P(A) \left[P(B|A) \log \frac{P(B|A)}{P(B)} + (1 - P(B|A)) \log \frac{1 - P(B|A)}{1 - P(B)} \right]$$

J-Measure 函数由 Symth 提出[38]，并作为规则的重要性度量引入规则归纳算法 ITRULE。它采用对数计算，反映的是信息量，受频度的影响较小。J-Measure 函数考虑了规则的前件 A 和后件 B 的概率分布的相似程度，以及用 A 的出现概率作为前件的简洁性的度量(一般地，A 的长度越小， $P(A)$ 越大)，是比较好的兴趣度度量。

此外，Toivonen[39]也进行了相关约束关联规则的研究，他提出了根据规则的后件对挖掘出的关联规则集合进行分组，选取分组的覆盖集合(Cover Rules) 作为所关注的规则等。考虑到挖掘出的规则最终还是要由用户来确定是否有趣，这种判断主观性较强，因此有些文献[40][41]也研究将先验知识等因素应用于约束关联规则的挖掘方法。

2.3 决策树

决策树是一个类似流程图的树型结构，其中树的每个内部节点代表对一个属性的测试，其分支代表测试的每个结果，而树的每个叶节点则代表一个类别，树的根节点，即树的最高层节点，是整个决策树的开始。

决策树很擅长处理非数值型数据，这与神经网络只能处理数值型数据比起来，就免去了很多数据预处理工作，并且训练决策树的时间远远低于训练神经网络的时间[42]。决策树主要是基于数据的属性值进行归纳分类，常用于量的参数，而且解释性很好。对于复杂问题可以使用层次方法“IF-THEN”规则进行分类。决策树的流行算法有 Quinlan 提出 ID3 和 C4.5 方法，Kass 提出的 CHAID 方法[43]，Breiman 等人提出的 CART 方法[44]等，可收缩性好的算法有 SLIQ 和 SPRINT，可以处理分类属性和连续属性问题。

CHAID(Chi-Square Automatic Interaction Detector，卡方自动交互检测)是一种快速多维树型统计算法。CHAID 在每次分割时利用卡方检验(Chi-Square Test)来计算节点中类别的属性值，以属性值大小来决定决策树是否继续生长，不必作修剪树的动作。其过程是：

将分类指标与结果变量进行交叉分类，产生一系列二维分类表，分别计算二维分类表的 X^2 值，以产生 P 值最小的二维列表的变量为最佳的初始分类变量，然后在此基础上继续分类，直到 P 大于设定的有统计意义的 α 值时为止。CHAID 自动地把数据分成互斥的、无遗漏的组群，但只适用于类别型数据。

CART(Classification and Regression Trees，分类回归树)是由 Leo Breiman、Jerome Friedman、Richard Olshen 和 Charles Stone 于 1984 年提出的一种数据分类和预测算法。其算法思想是：

假设集合 T 包含 N 个类别的纪录，那么其 Gini 指数为：

$$Gini(t) = 1 - \sum_{j=1}^N [P(j|t)]^2$$

式中：P(j|t)为类别 j 在 t 节点处的相对频率。当 Gini(t)最小为 0 时，即在此节点处所有记录都属于同一类别，表示能得到最大的有用信息；当此节点中的所有记录对于类别字段来说是均匀分布时，Gini(t)最大，表示能得到最小的有用信息。如果集合分成 k 个部分，那么进行这个分割的 Gini 就是

$$Ginisplit(T) = \sum_{i=1}^k \frac{n_i}{n} Gini(t)$$

式中：k 是子节点的个数， n_i 是在子节点 i 处的记录数，n 是在节点 P 处的记录数。

CART 对于每个属性都要遍历所有可以的分割方法后，若能提供最小的 $Ginisplit$ ，就被选择作为此节点处分裂的标准，对于根节点和子节点都一样。CART 算法得到的决策树每个节点有两个分支，这种树也称为二叉树。

2.4 回归分析

线性回归是研究一个应变变量与一个或多个自变量之间的直线依赖关系的统计分析方法。它主要基于最小二乘法原理的无偏估计。最小二乘法的基本假设是残差的平方和为最小。为满足最小二乘法的这个基本假设，若未来的拟合直线方程为： $y = \alpha + \beta x$ ，其中 α, β 是回归系数，则利用最小二乘法求解回归系数的方法如下：

设给定了 n 个样本，它们的坐标分别为： $(x_1, y_1), (x_2, y_2), (x_3, y_3) \dots (x_n, y_n)$ 。可求得：

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n, \bar{y} = (y_1 + y_2 + \dots + y_n) / n \quad (1)$$

则：

$$\beta = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

$$\alpha = \bar{y} - \beta \bar{x} \quad (3)$$

根据自变量个数的不同，回归分析可分为一元线性回归和多元线性回归两种类型。

一元线性回归分析是最简单的回归形式，研究一个自变量和一个应变变量之间是否存在某种线性关系的统计学方法。

多元线性回归是一元线性回归的扩展，研究多个自变量和一个因变量之间是否存在某种线性关系，应用范围更广。

第三章 在教学信息分析中的应用

3.1 教学信息数据仓库

3.1.1. 数据仓库及 OLAP 技术

数据仓库是一个面向主题的、集成的、时变的、非易失的数据集合，用来支持管理人员的决策。数据仓库的四个特征具体指：

1、面向主题的：指数据仓库是面向应用的，有明确的应用主题。因此需要排除对于决策无用的数据，提供特定主题的简明视图。

2、集成的：指各部门操作数据的集成，不是数据的简单堆积，而是要按照统一的规范进行清理、转换后再装载到数据仓库中。

3、时变的：指数据仓库中的数据从时间角度提供信息，需要隐式或显式的时间维度。

4、非易失的：指数据仓库只提供数据的初始化装入和数据访问，不进行操作。

数据仓库主要有三方面的作用：

首先，数据仓库支持多维分析(Multi-Dimensional Analysis)。多维分析是通过把一个实体的多项重要的属性定义为多个维度，使得用户能方便地汇总数据集，简化了数据的分析处理逻辑，并能对不同维度值的数据进行比较，而维度则表示了对信息的不同理解角度，例如，时间和地区是经常采用的维度。应用多维分析可以在一个查询中对不同的数据进行纵向或横向的比较，这在决策工程中非常有用。

其次，数据仓库提供了标准的报表和图表功能，其中的数据来源于不同的多个事务处理系统，因此，数据仓库的报表和图表是关于整个企业集成信息的报表和图表。这些功能是对传统的联机事务处理(OLTP)的扩充，但在数据仓库中，数据是经过汇总归纳的，保证了报表和图表反映的是整个企业的一致信息。

第三，数据仓库是数据挖掘(Data Mining)技术的关键基础。数据挖掘技术要在已有数据中识别数据的模式，以帮助用户理解现有的信息，并在已有信息的基础上，对未来的状况做出预测。当然，也可以把数据从数据仓库中转到数据挖掘库或者数据集中再进行数据挖掘。

联机分析处理(On Line Analytical Processing, 简称 OLAP)技术实现了解释模型和思考模型。主要功能是深入了解事务，并做出总结性分析，以可视化的方式呈现给用户。OLAP 有两大特性：多维性和直接面向用户。它将决策支持系统带入了更高的层次。该分析处理技术从企业的数据集中收集信息，并运用了数学运算和数据处理技术。它一般以数据仓库为基础对数据进行多维化和预综合分析，构建面向分析的多维数据模型，再使用多维分析方法从多个不同角度对多维数据进行分析、比较，找出它们之间的内在联系。联机分析处理使分析活动从方法驱动转向了数据驱动，并使分析方法和数据结构实现了分离。

3.1.2. 教学信息数据仓库的设计

数据仓库的模型现在常用的有星形模型、雪花模型、多维数据模型、面向对象模型。在数据仓库中，依据所选定的主题、所要存储的数据内容、支持数据仓库的系统环境、对象间的关系来决定使用那种模型。

星型模型的中心是一个事实表，一个典型的事实表包括外键和可能被度量的事实，可以由成千上万个行组成。与之相连的对象称为维表。每个维表通过主键和外键与事实表相关联，但是维表之间通常没有关联。维表包含可用于 SQL 查找标准的数据属性，一般比较小。一个简单的星型模型由一个事实表和若干个维表组成。复杂的星型模型包含多个事实表和维表。星型模型中事实表到维表的联系是通过事实表的外键参照维表的主键来建立关联而实现的。在每张维表中除了包含每一维的主键外，还有说明该维的一些其他属性字段。维表记录了维的层次关系。

星型模型主要有如下优点：在星型模型中进行的复杂查询，可以直接通过各维的层次比较、上钻、下钻等操作完成，大大减少用户的查询响应时间；大量的商业智能工具(BI)都支持星型模型；星型模型既可以被用在简单的数据集也可以被应用在巨型数据仓库上。

选取电气、电信、电子和计算机专业 4 个专业 99 年到 04 年的本科生信息共 51 个数据集，其中电气 21 张，电信 6 张，电子 12 张，计算机 12 张。涉及学生学号、姓名、选课时间、课程号、课程名称、课程学分、学时以及成绩等信息。这里根据当前教学信息的特点，采用星型模型（图 3-1）。

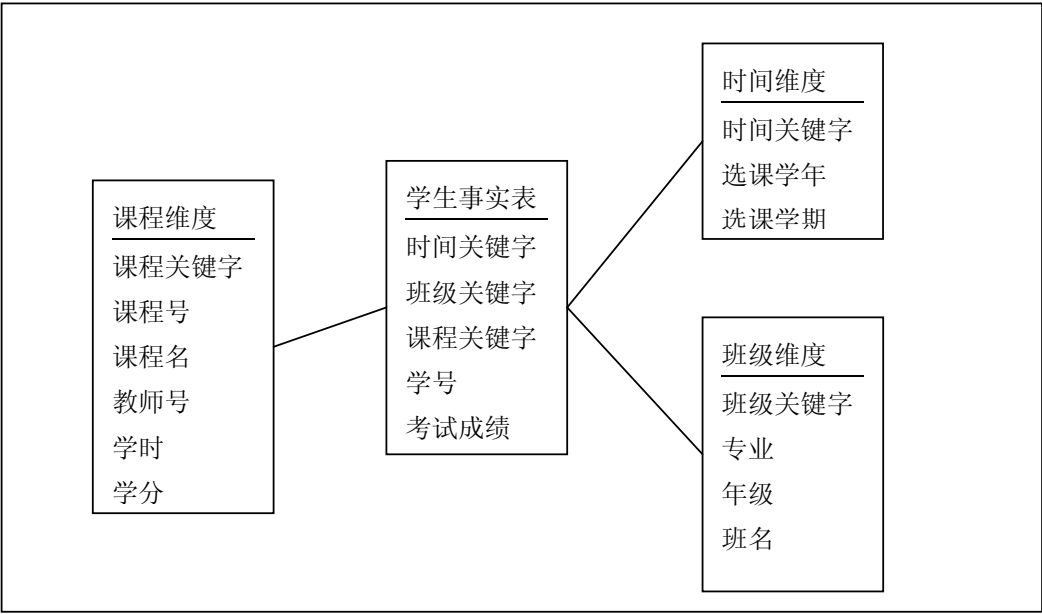


图 3-1、星型维度模型

3.1.3. 教学信息数据仓库的实现

为建立多维数据库，需要根据情况增加一些维字段，如根据选课时间 xnxq 增加字段选课年 xyear 和选课学期 xterm 作为时间维，再增加字段专业 zy、年级 nj、班名 bm 以备下面建立班级维。处理完成后的数据集 Mddata.MDDB_input1，共有 72897 条记录，13 个字段。

采用 SAS 中的 EIS/OLAP Application Builder 建立教学信息数据仓库。

首先为 MDDB 注册。需要在注册表 (Repository) 中为一个或多个变量赋予 CATEGORY 属性，再为一个或多个变量赋予 ANALYSIS 属性。具体方法是：

在 EIS Main Menu 窗体中的“Metabase”选择 File->New 建新的注册表 (Repository)。然后在“Table”中添加源数据集，这里是前面整理好的数据集 Mddata.Mddb_input1。接着需要在“Table”中的“Columns”中设置一个分析列，这里为变量考试成绩 kscj 添加分析属性 (Analysis)。最后就可在右侧的“Attributes”中为该数据集添加维信息了。这里我们为多维数据库设计了两个维度，时间维度 time hierarchy (xyear,xterm) 和班级维度 class hierarchy (kch,zy,nj,bm)，见图 3-2。

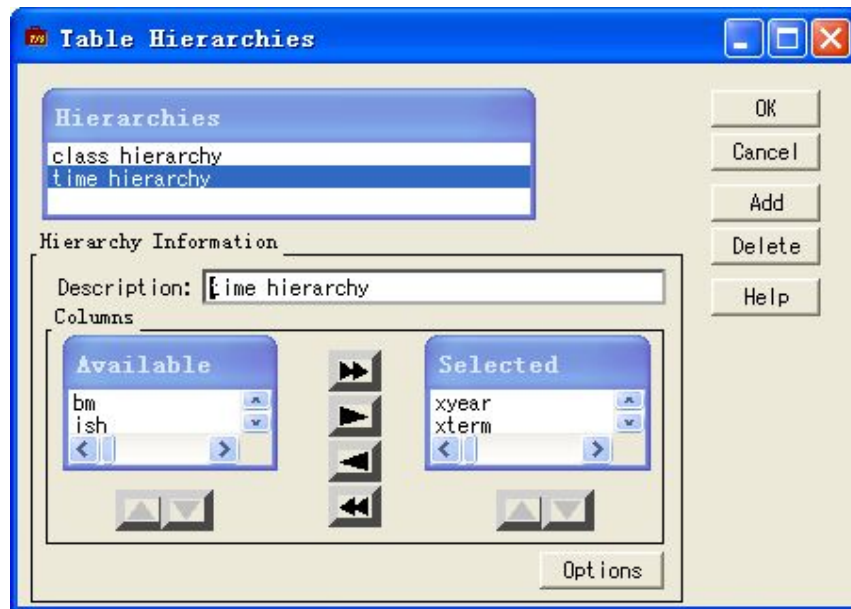


图 3-2、建立维度

然后就可以创建 MDDB 了。

首先要建立一个 Application Database。在“Build EIS”窗体中选择 File->New，输入所要建的 Application Database 的名称及描述，这里命名为 Student。

然后在“Build EIS”窗体中点击“Add”，在“Object Databases”中选择“Data Access”再在“Objects”中选择“Multidimensional database”，为建立一个 MDDB 输入名称描述路径等信息。还要在“MDDB”处为多维数据库选择存储路径，在“Table”处选择前面注册的数据集 Mddata.Mddb_input1，选取前面在“Metabase”中建好的 dimension 和 analysis columns，见图 3-3。点击“Create”后，就建好了学生信息多维数据库。

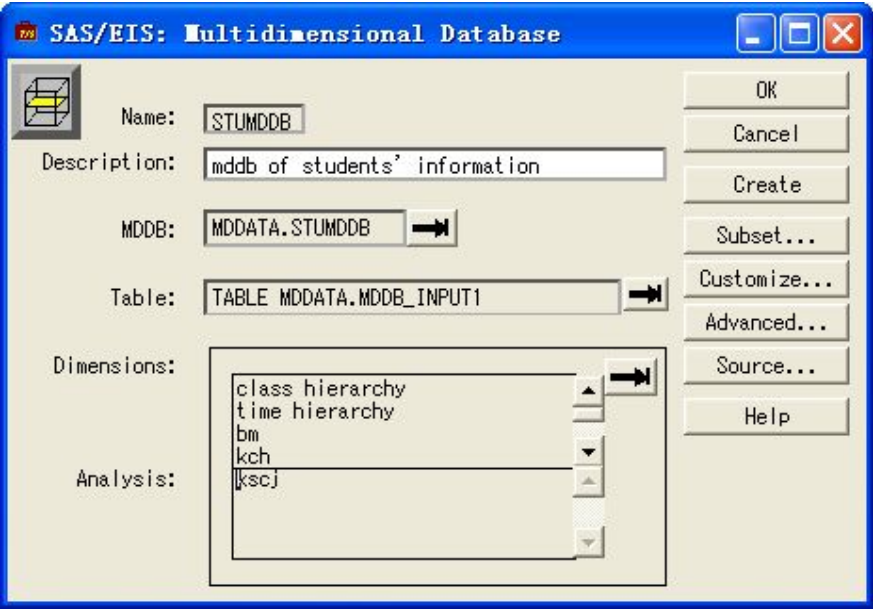


图 3-3、创建多维数据库

建立多维报告（Multidimensional Report）后（图 3-4），可以对多维数据立方体进行上卷、下钻、切片、切块和旋转等操作，以多种可视化形式展现多维数据。

The screenshot shows the 'report of clee student mddb' window. It displays a table with the following data:

		kscj		
zy	xyear	Total Number of Nonmissing Values	Sum	Uncorrected Sum of Square
dq	2004	4501	350572.90	28156115.19
	2003	6435	481321.30	37543509.85
	2002	5600	431539.20	34033875.04
	2001	3131	237116.30	18461228.85
	2000	1459	107046.80	8067276.24
	1999	64	4326.00	308510
dx		14	927.00	65785
	2004	1328	105579.50	8698977.75
	2003	3514	275051.80	22230451.06
	2002	2531	199876.80	16177848.12
	2001	1328	100206.20	7789817.24
dz	2000	10	576.00	42308
	2004	1871	143221.00	11363261
	2003	5959	459242.70	37366330.41
	2002	6062	476656.50	38352410.65
	2001	4199	330500.80	26670305.76
	2000	2451	190580.20	15252432.92

图 3-4、多维数据库展现

上卷(roll up)，通过某一维的概念分层向上将低层次的细节数据概括到高层次的汇总数据，或者减少维数；

下钻(drill down)，是上卷的逆操作，从汇总数据深入到细节数据进行观察或引入新维。

切片(Slice)，在给定的数据立方体的某一维上选定一组维成员，形成一个子立方体。切片能

对超过四个维的数据空间结构舍弃一些观察角度，而在两个维上集中精力观察、分析数据，从而大大降低分析的难度。

切块(Dice)，在给定的数据立方体上对两个或多个维执行选择，定义子立方体。也就是限制多维数组的某一维的取值区间。切片可以看成是切块在这个区间只有一个维成员时的特例。

旋转(Pivot)，是一种目视操作，转动数据的视角，提供数据的替代表示。它可能是交换行和列，也可能是把某一个行维移到列维中去，或用页面之外的一个维替换页面内的某一个维。

教学信息数据仓库的和 OLAP 技术的应用，客观地显示了查询者想要了解的众多教学因素的分析 and 汇总报表，为后面的数据挖掘工作提供线索和参考依据。数据挖掘在此基础上能够进一步利用各种统计分析方法对数据进行再分析，以获得更深入的理解以帮助教学决策者寻找规律和内在信息。

3.2 教学信息模型的建立

本文将在教学信息数据仓库的基础上，使用 SAS/Enterprise Miner 对教学信息进行数据挖掘，并建立各种数据挖掘模型。

SAS/Enterprise Miner 封装并集成了创建数据源、数据采样、数据划分、变量转换、数据研究及预处理、数据过滤、建模、模型评估和决策等模块，为用户提供了一个拖拉式操作环境，使数据挖掘过程变得易于实现。它的可视化操作可引导用户按 SEMMA 原则方便地进行数据挖掘。

SEMMA 代表采样(Sample)、探索(Explore)、修改(Modify)、建模(Model)和评估(Assess) [45]。使用 SEMMA 方式，测试模型能被保存以重复使用或经过导出集成到其他程序中。这使得 Enterprise Miner 可作为客户所采用的任何挖掘迭代方法的一部分。SEMMA 挖掘过程具体如下：

(1) 采样

从大型数据集中抽取足以包含企业重要信息的可靠的样本。但样本大小也应考虑对执行速度的影响，不能大到无法处理。可以使用采样节点对数据集进行随机采样、分层采样和聚集采样，然后用数据划分节点把数据划分为练习、测试和有效数据集。对大型数据库进行采样能显著地缩短建模练习的时间。

(2) 探索

对数据进行探索，搜索预期的关系和非预期的趋势和异常，以获得解释和想法。这可以帮助精简数据挖掘的步骤。如果使用可视化的探索仍无法清楚地揭示其趋势，那就通过因子分析、相关分析和聚类等统计学手段来探索。

(3) 修改

用创建、选取和转换变量的方法修改数据，使重点放在模型选取的过程上。经过探索阶段的发现，我们可能会发现需要引入或删减一些变量。因为数据挖掘是一个动态的、交互式的过程，因此当有新的信息可用时你需要更新数据挖掘的方法或模型。

(4) 建模

这是数据挖掘过程的关键。根据数据集的特征和要实现的目标，选择回归分析、决策树、神经网络等方法建模。

（5）评估

评估上述数据挖掘过程得到的结果和模型的效用和可靠性，在能可靠预测所需结果的模型中通过比较选取效果最好的模型，或决定是否重新进行数据挖掘过程。

Enterprise Miner 运用 SEMMA 组织方式，通过连接资料节点和程序节点的方式建构可视化的数据流程图，能方便地实现应用探索性统计和可视化技术，选择和转换重要的预测变量，为变量建模预测结果以及确定模型的准确性等工作。

3.3 公共课影响因素决策树模型

学生的考试成绩是教学质量的重要评价指标之一，这受到多种因素的影响，包括学生学习情况、课程设置、学习氛围等因素。教学改革中的重要一环就是通过对课程设置的改革来促进教学质量。通过研究影响公共课程成绩的相关因素，找出进行课程设置的改革重点，为教学改革提供参考。本部分重点通过运用数据挖掘技术中的决策树算法对《高等数学》和《马克思主义哲学原理》课程相关信息进行建模分析，挖掘影响公共课程成绩的相关因素，探讨数据挖掘技术在教育教学领域中的应用。这里将考察纵跨 99 级到 04 级的电气、电信、电子和计算机共 4 个专业的《高等数学》和《马克思主义哲学原理》课程成绩的分类情况。

本研究采用 SAS V8.2 软件中的 Enterprise Miner（简称 EM）模块进行数据挖掘。SAS EM 在 SAS 数据仓库和数据挖掘方法论的基础之上，采用图形化界面、菜单驱动方式，为用户提供了一个数据挖掘的集成环境，集成了数据获取工具、数据抽样工具、数据筛选工具、数据变量转化工具、数据挖掘数据库、数据挖掘方法等。SAS EM 为建立决策树提供了数据剖析工具、决策树浏览工具（决策树基本内容和统计值的汇总表、决策树的导航浏览器、决策树的图形显示、决策树的评价图表）、数据挖掘评价工具等，并且支持 CHAID 和 CART 软件包。我们运用这两种决策树算法来分别建立决策树模型。

（一）数据处理

首先，提取出各专业中《高等数学》课程成绩的相关记录。《高等数学》课程在全院系均为必修课，字段课程属性 kcsx 不具有分类功能，因此不提取。提取的数据集字段包括课程号 kch，学分 xf，学时 xs，教师号 jsh，考试成绩 kscj，我们增加了一个表示专业的字段 zy，以区分不同的专业。然后，合并各数据集为一个数据集，作为决策树分析的输入数据源，共有 1827 条记录，6 个字段。

面向属性的归纳方法（AOI）使用概念分层，通过以高层概念替换低层概念训练数据。AOI 通过删除或概化具有大量不同值的属性，用来进行一些预相关分析[46]。

一方面进行属性选择和删除不相关的属性。在进行数据挖掘工作时将一些不相关的属性去掉，既能提高挖掘效率，又能提高分类器的泛化能力。因此在前期数据预处理时删除这些不相关的属性。

另一方面进行属性概化。这里主要是对考试成绩 kscj 进行概化。由于考试成绩来自不同时间，因此考试难易程度和评分标准均不相同，最好的概化方法是针对每次的考试不同的成绩分布情况进行概化。但事实上，历史记录并不显示哪些成绩是同一次考试的，这种方法显然不现实。考虑

到《高等数学》课程是公共必修课，每届的考试难度及评价标准相对比较稳定，所以可以对所有成绩一起进行分析。考试成绩 kscj 的分析情况见图 3-5 和图 3-6。

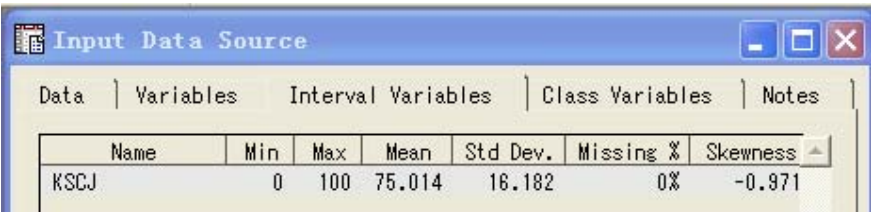


图 3-5、Input Data Source 节点

观察图 8 所示的分析结果发现，成绩并不是以 60 分为均值成正态分布，而是高出 15 分左右，从侧面反映了这门课考试的题目难易程度。将考试成绩映射到一个有序概念域 $D=\{bad,past,good\}$ ，具体方法是在 Transform Variable 节点中对考试成绩变量 kscj 利用 Bucket 进行转换，如图 3-6，结果见表 3-1。

表 3-1 概化成绩对应表

精确成绩的范围	模糊成绩
kscj<60	01:low-60
60=<kscj<80	02:60-80
80=<kscj	03:80-high

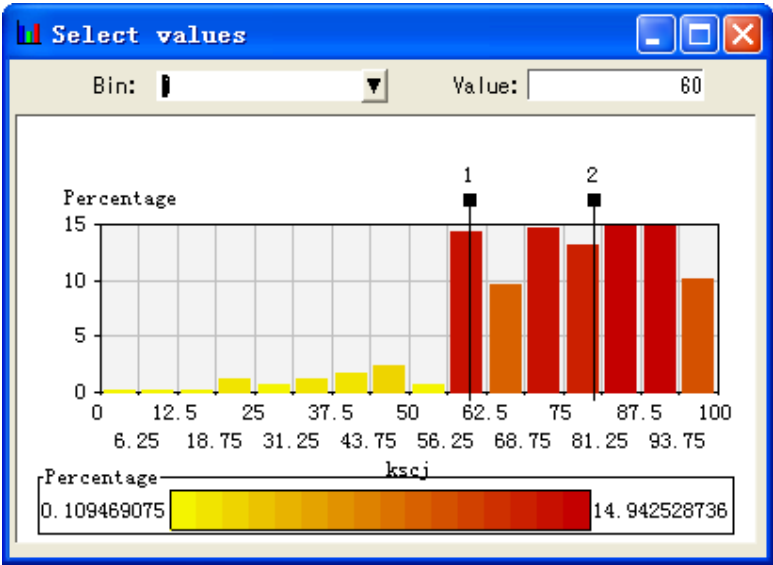
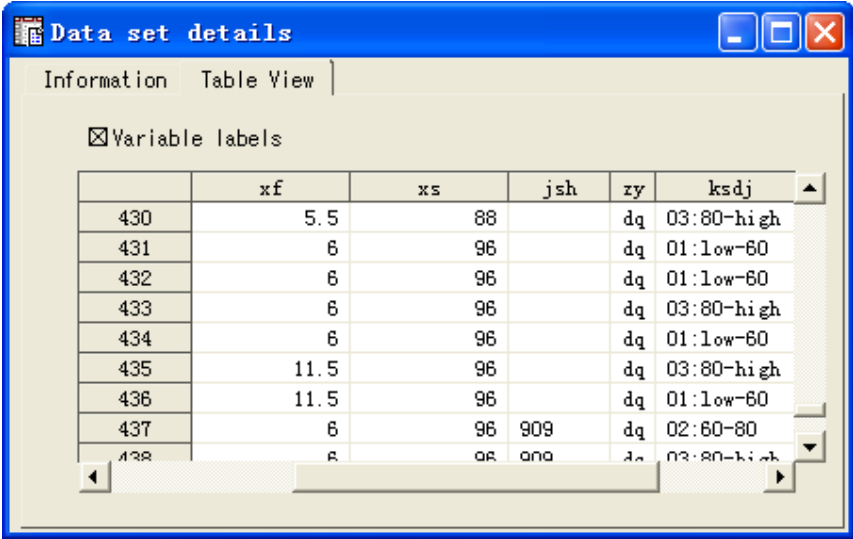


图 3-6、Transform Variable 节点

经过处理后的数据集见图 3-7。最后保留专业 zy，学分 xf，学时 xs，教师号 jsh 及概化后的考试成绩 ksdj。



	xf	xs	jsh	zy	ksdj
430	5.5	88		dq	03:80-high
431	6	96		dq	01:low-60
432	6	96		dq	01:low-60
433	6	96		dq	03:80-high
434	6	96		dq	01:low-60
435	11.5	96		dq	03:80-high
436	11.5	96		dq	01:low-60
437	6	96	909	dq	02:60-80
438	6	96	909	dq	03:80-high

图 3-7、处理后的数据集

（二）建立决策树模型

在 SAS/EM 中按照图 3-8 添加节点建立决策树模型，其中添加的两个 Tree 节点分别运用 CHAID 和 CART 算法进行分类。

对 CHAID 的 Tree 节点，设置分割标准为卡方检验，显著水平为 0.05；为避免自动剪枝，设置模型评估度量为 Total leaf impurity(Gini Index)；为了强制进行启发式搜索，设置全面分割搜索中的最大尝试次数为 0；设置 p-value adjustment 为 Kass,并指定在选择了分枝数之后应用该指标。

对 CART 的 Tree 节点，设置分割标准为 Gini 约简（Gini reduction），节点的最大分枝数为 2；为避免自动剪枝，模型评估度量也设置为 Total leaf impurity(Gini Index)；设置 Sub-Tree 为 Best assessment value；而全面分割搜索中的最大尝试次数设置为 5000。

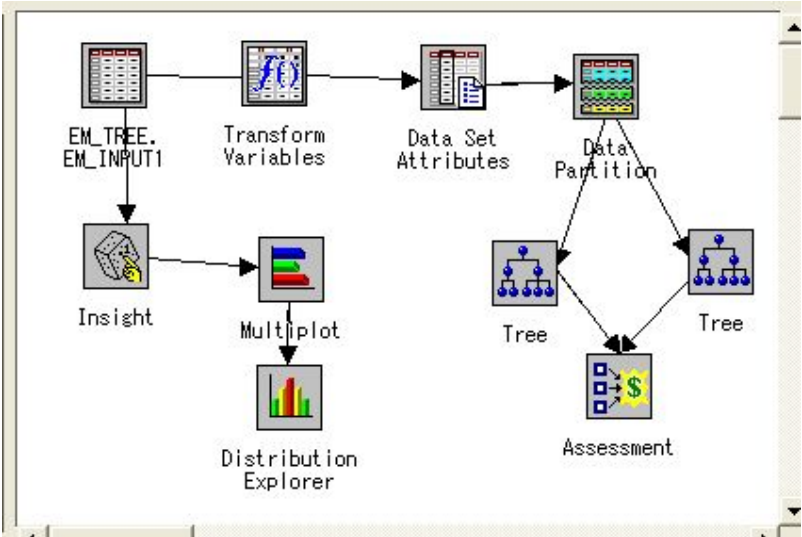


图 3-8、建立决策树模型

CHAID 和 CART 分别是使用卡方检验和 Gini 指数作为属性分割方法的代表算法。就类别的种类数和给定的某个属性的值的个数而言，对于信息增益，Gini 指数使属性中含有大量的有用信息，但随着类别种类数的增加，Gini 指数值减少，这对于属性选择来讲并不理想[47]；而对于卡

方检验, 当类别的种类增加时, X^2 值并没有明显的变化趋势[7], 而信息增益的值随着类别种类的增加而明显增加。而且 X^2 统计并不区分属性值中所含有的信息量的多少。

所分析的数据中属性的值的个数并不大, 因此不存在上述问题, 研究中更多关注的是这两种方法哪种更适合用于建立此模型的问题。

(三) 结果

通过 Assessment 节点可以对两个模型进行对比评价。从积累响应度 (Cumulative Response) 来看, 20%到 35%的百分数, CHAID 模型好于 CART 模型; 20%以下及 35%以上的百分数, CART 模型好于 CHAID 模型。综合而言 CART 模型适用范围更大些, 因此对于本研究所用的学生成绩数据, 使用 CART 模型比 CHAID 模型更合适。所以下面将重点分析 CART 模型中的决策结果。

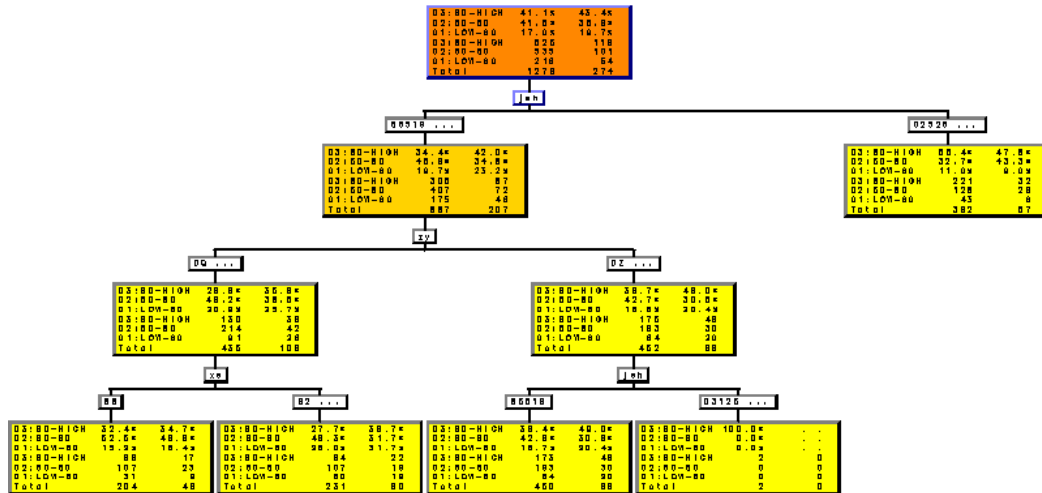


图 3-9、CART 决策树图

从 CART 模型决策树图 (图 3-9) 可见, 树的第一层是按照教师号进行分枝的, 因此得到结果为: 影响“高等数学”课程成绩的最主要因素是授课教师。从第一个节点我们还注意到, 总体的课程成绩比例为“优秀: 一般: 不及格=41.1%: 41.8%: 17.0%”。另外, 我们从模型中得到各属性的重要程度分别为: importance(z)=0.9003, importance(xs)=0.6257, 而 importance(xf)=0.3973。

表 3-2 CART 决策树规则

组合名称	03:80-HI	02:60-80	01:LOW-6
教师 =02526/02536/03035/03128/77507/ 82580/86019/89023/89523/909	56.4% (221)	32.7% (128)	11.0% (43)
教师=86519/90534 专业=电气/电信 学时= 92/96	27.7% (64)	46.3% (107)	26.0% (60)
教师=86519/90534 专业=电气/电信 学时= 88 学分=5.5	35.6%(48)	51.1%(69)	13.3%(18)
教师=86519/90534 专业=电气/电信 学时= 88 学分=6	26.1%(18)	55.1%(38)	18.8%(13)
教师=86519 专业=电子/计算机	38.4% (173)	42.9% (193)	18.7% (84)
教师=03126/86039 专业=电子/计算 机 学分=2.5	100.0% (2)	0.0% (0)	0.0% (0)

该决策树的部分规则如表 3-2 所示，详细解读如下：

规则 1 显示教师号为 02526/02536/03035/03128/77507/82580/86019/89023/89523/909 教课的“高等数学”课程成绩中达到优秀比例高于总体，达到 56.4%。各层次成绩比例为“优秀：一般：不及格=56.4%：32.7%：11.0%”；相对于总体的课程成绩比例为“优秀：一般：不及格=41.1%：41.8%：17.0%”略好一些。由此推断剩下两位教师号为 86519/90534 的教师所教的学生成绩情况低于平均。

从规则 2、3、4，我们也注意到，这两位教师教授的电气和电信专业的考试成绩，无论学时和学分怎样设置，优秀率都低于总体。

另外，规则 5 显示，教师号为 86519 的教师教授的电子和计算机专业的“高等数学”考试成绩与总体情况相当。

规则 6 显示，教师号为 03126/86039 的教师教授的电子和计算机专业，学分为 2.5 分的“高等数学”考试成绩优秀率为 100%。但由于这里选修 2.5 学分的该课的学生仅有 2 人，成绩均优秀，所以“优秀率为 100%”。这种情况的样本数太少，此规则不具有统计意义，体现更多的是随机性。

下面我们又用决策树模型对《马克思主义哲学原理》课程成绩进行了数据挖掘。

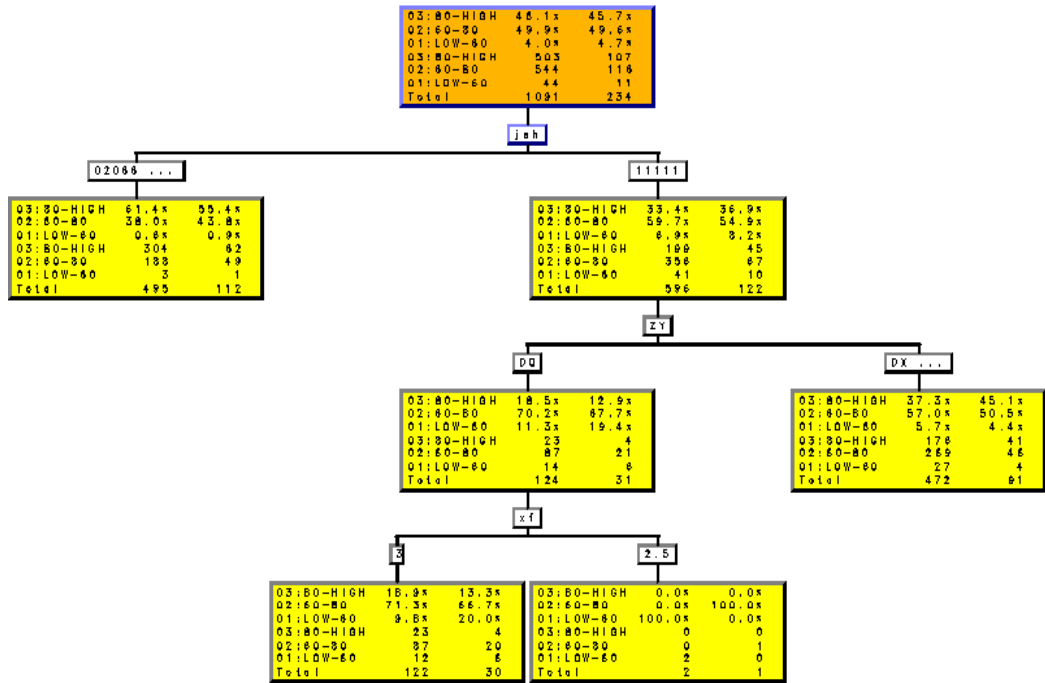


图 3-10、CHAID 模型决策树图

进行算法评估时，发现这里 CHAID 模型更合适些。从积累响应度（Cumulative Response）来看，在 10%以下及 30%以上的范围上，CHAID 模型高于 CART 模型；而 20%到 30%的范围内，CART 模型高于 CHAID 模型。综合而言 CHAID 模型适用范围更大些，因此对于本研究所用的学生成绩数据，使用 CHAID 模型比 CART 模型更合适。所以下面将重点分析 CHAID 模型中的决策结果。

该模型对《马克思主义哲学原理》课程成绩相关记录进行了相同的挖掘，类似对《高等数学》课程决策树图的分析，发现影响《马克思主义哲学原理》课程成绩的因素依次为授课教师、专业、学分、学时，importance(zy)=0.5793,importance(xf)=0.1075,而 importance(xs)=0，所以学时（xs）属性被舍弃。说明学时设置的长短对于《马克思主义哲学原理》课程成绩的高低并无明显影响。由此可见影响《马克思主义哲学原理》课程成绩的主要因素是授课教师和专业。

表 3-3 CHAID 决策树规则

组合名称	03:80-HI	02:60-80	01:LOW-6
教师02066/03129/78534/WP0400	61.4%（304）	38.0%（133）	0.6%（3）
教师11111 电气/电子/计算机	37.3%（176）	57.0%（269）	5.7%（27）
教师11111 电气 学分=3	18.9%（23）	71.3%（87）	9.8%（12）
教师11111 电气 学分=2.5	0.0%（0）	0.0%（0）	100.0%（2）

该决策树的部分规则列于表 3-3，从中我们可以得出如下信息：

由教师号分别为“02066/03129/78534/WP0400”的四位老师所教的《马克思主义哲学原理》课程的成绩达到优秀的比例明显高于其它教师，达 61.4%，不及格率仅为 0.6%。而教师号为“11111”所教的班级的学生成绩达到优秀的仅为 37.3%，不及格率高达 5.7%。

由教师号为“11111”所教的、专业为“电气/电子/计算机”的三个专业的学生《马克思主义哲学原理》课程成绩分布为“优秀：一般：不及格=37.3%:57.0%:5.7%”，而总数据集中课程成绩为“优秀：一般：不及格=50%:45%:4.2%”（见图 3-10 节点 1），可见不同教师所授课班级（大班）的成绩存在明显差异。对于这种结果，有必要进一步了解、分析造成这种情况的更具体的原因，从而对今后的教学工作加以引导。

表 3-3 中的后两条规则显示，电气专业中由教师“11111”教的学生，其《马克思主义哲学原理》课程成绩为一般及优秀的学生主要集中在选修的课程学分设置为 3 分中，而只有 2 人选修了该科为 2.5 学分的课，可见课程学分设置的高低对学生学习的积极性有很大的影响。由于其中选修 2.5 学分的该课的学生仅有 2 人，成绩均较差，所以“不及格率为 100%”。这种情况的样本数太少，此规则不具有统计意义，体现更多的是随机性。

本研究采用决策树分类法中的 CHAID 算法和 CART 算法对《高等数学》和《马克思主义哲学原理》课程成绩进行分类的方法，快速有效地挖掘出一些有指导意义的规则结果：对公共课程成绩影响最大的因素是授课教师和专业，其次是课程学分的设置，而学时的长短影响因课程特点不同而异。该模型应用于其他公共课程成绩的影响因素分析也取得了良好效果。该结论说明，任课教师的教学水平及教学效果是提高教学质量和学生成绩的重要因素之一，也是今后学校在教学改革中的重点工作。对于不同的数据，所采用的挖掘方法也不同，本研究采用模型评估方法来决定选择何种决策树分类法。教育教学信息量大，值得挖掘的信息还很多，本研究仅在此方面做了一些探索性研究，希望能与同行在这方面进行交流与勾通，共同推动教育教学改革的发展。

3.4 课程间成绩影响关联模型

学习是一个积累的过程，以前学习的相关内容对后面的学习有一定的影响，因此，某些课程的成绩好坏很大程度上影响另一门相关课程的成绩好坏。课程间成绩影响关联模型将分析信息与电气学院 03 级电子专业和计算机专业本科生选课信息中课程间的影响。

由于使用的关联规则挖掘方法是针对事务数据集进行操作的，且挖掘任务是挖掘学生选课关联规则，所以数据源选取电子专业 03 级和计算机专业 03 级的选课数据表，电子专业 03 级 2 个班的选课记录共有 1950 条记录，计算机专业 03 级 2 个班的选课记录共有 2282 条记录，合起来共 4232 条选课记录，变量包括 xh（学号）、kch（课程号）、kcm（课程名）等。

首先，需要将不同专业的选课记录合为一个数据集，并且将成绩按等级划分。对计数变量先将其离散化，如对成绩 $kscj < 60$ ，表示为不及格 bad， $60 \leq kscj < 80$ ，表示为通过 passing， $80 \leq kscj$ ，表示为优良 good。然后，对分类变量按其取值分为几个属性-值，如 0100001_good 和 08110010_good 等。之所以采用课程号 kch 而不是课程名 kcm，是因为尽管不同的学生可能都选修了相同名字的 课程，但不同课程号的同名课程开设的时间考试的内容等都不尽相同，因此不能简单的认为可以用相同的评价标准来分析，而课程号可以区分这些课程。

```
data paper.ass3z;
set em_courc.dz031_etl em_courc.dz032_etl em_courc.jsj031_etl em_courc.jsj032_etl;
keep xh kch kscj zy dj; //zy 为新增加的变量表示“专业”，dj 表示成绩“等级”
```

```
if _N_ <= 1950 then zy = ' dz' ; //电子专业
else zy = ' jsj' ; //计算机
if abs(kscj) < 60 then dj = 'bad';
else if ((60 <= abs(kscj)) & (abs(kscj) < 80)) then dj = 'passing';
else dj = 'good';
run;
//将 kch,dj 合并，并分为若干个属性
data paper.ass2z;
set paper.ass3z;
keep xh kcdj;
format kcdj $15.;
kcdj = compress(kch) || '_' || compress(dj);
run;
```

经过处理过的数据集 ass2z，如图 3-11

VIEWTABLE: Paper.Ass2z		
	xh	kcdj
1	030802123	0100001_goo
2	020802327	0100007_goo
3	020802327	0101001_goo
4	020802327	0101002_pas
5	020802327	0101003_pas
6	020802327	0101004_pas
7	020802327	0101017_goo
8	030802105	0101022_pas
9	020802327	0101058_pas
10	030802102	0104005_goo
11	030802125	02130095_goo
12	030802104	02130095_goo
13	030802118	0302021_goo
14	030802127	03130147_goo
15	020802327	0401003_bad
16	020802327	0402001_pas
17	020802327	0402027_bad
18	020802327	0403002_pas
19	020802327	0403004_pas
20	020802327	0403018_pas

图 3-11、处理后的数据

对上述任务数据集进行关联规则挖掘（图 3-12），设置学号 xh 为 id 变量，课程成绩等级 kcdj 为 target 变量，设置 min_sup=0.1，min_con=0.6，最大项集数为 3。

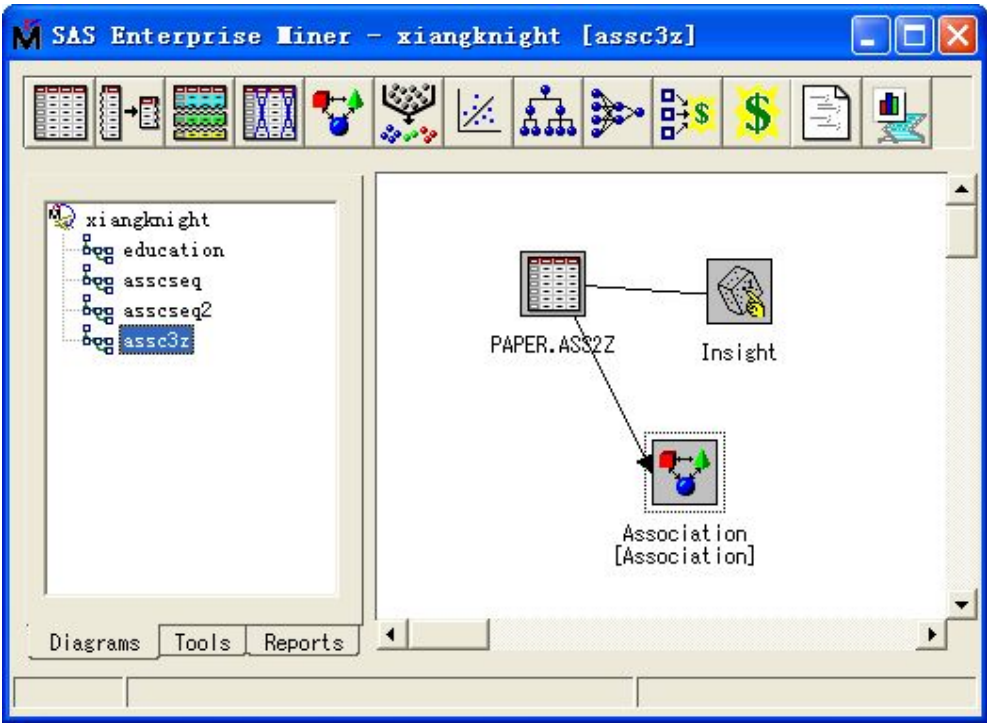


图 3-12、课程间成绩影响关联模型

最后挖掘出的部分关联规则如表 3-4 所示。

表 3-4 部分规则列表

No	Relations	Lift	Support (%)	Confidence (%)	Count	Rules
39	2	1.15	42.75	95.16	59.00	12110160_goo ==>16110030_goo
134	2	1.66	32.61	81.82	45.00	08110010_goo ==> 08110730_goo
136	2	2.30	31.88	73.33	44.00	07140010_goo ==> 83110150_goo
...

规则 136 值得关注，07140010 代表“金工实习（B）”，83110150 代表“工程制图基础”，这两门课程都是电子专业的必修课。“金工实习（B）”达到优良的电子专业的学生“工程制图基础”达到优良的可能性为 73.33%，有 31.88%的选课记录支持该结论，且 lift=2.30 表明“工程制图基础”达到优良的可能性是原来的 2.3 倍。可以认为学好“金工实习（B）”课程对学好“工程制图基础”有很大的影响。这条规则在很大程度上反驳了实际教学中认为“金工实习（B）”课程不重要的观点，为今后重视该门课程提供量化依据。

规则 39 中，12110160 代表“毛泽东思想概论”，16110030 代表“体育（C）”，这两门课程都是电子和计算机专业学生必修的课程，且课程号一致，具有普遍意义。由规则可见，“毛泽东思想概论”成绩良好的学生“体育（C）”成绩良好的可能性高达 95.16%，有 42.75%的选课记录支

持该结论。可以认为“注重德育的学生同时也非常注意自身的体育学习”。

规则 134 中，08110010 代表“C 语言程序设计”，08110730 代表“面向对象程序设计”。规则表明，“C 语言程序设计”成绩优良的学生“面向对象程序设计”成绩优良的可能性为 81.82%，有 32.61%的选课记录支持该结论。根据常识，有“C 语言程序设计”基础对于学习“面向对象程序设计”课程是有好处的。

该模型显示了课程间成绩的影响关系，揭示出某些课程的成绩好坏很大程度上影响另一门课程的成绩好坏，由此根据这些规则可以调整课程的开设顺序和重视程度，也可以为学生选修某些课程提供参考。

3.5 学生选课特征关联模型

该模型通过对基于约束的关联规则挖掘方法的分析和研究，结合实际学生选课信息，提出了适合的约束条件来剪除无兴趣规则，并挖掘出部分课程间的有趣规则,为今后的教学课程设置提供了参考。

由于使用的关联规则挖掘方法是针对事务数据集进行操作的，且挖掘任务是挖掘学生选课关联规则，所以只选取计算机系 xx 级学生选课数据集市为数据源，共 2282 条选课记录，变量包括 xh（学号）、kch（课程号）、kcm（课程名）等，见表 3-5。

表 3-5 选课数据集市样例

xh	xnxq	kch	kcm	xf	xs	kcsx	kscj
030805124	2003-2004	0100001	市场营销	2	32	任选	95
030805101	2003-2004	08110010	C 语言程序设计	4	64	必修	80
030805126					

SAS 的 EM 采用 Apriori 算法利用 k-项集来探索(k+1)-项集。首先找出频繁 1-项集的集合，该集合记作 L₁，L₁用于找频繁 2-项集的集合 L₂，而 L₂用于找 L₃，如此下去，直到不能找到频繁 k-项集为止。然后再根据预先设定的最小支持度和置信度产生规则。在 SAS 的 EM 中主要设置三个参数：产生规则的项集最大数目(items)、最小支持度(min_sup)和最小置信度(min_conf)。输出结果显示满足要求的所有规则以及每条规则的作用度 (lift)、支持度(support)、置信度(confidence)等信息。

对上述任务数据集进行关联规则挖掘（图 3-13），设置 min_sup=0.1，min_con=0.6，最大项集数为 3，最后挖掘出的关联规则共有 31484 条。通过观察，在得到的极大数量的关联规则中有很多是支持度和置信度都很高的强关联规则。根据先验知识，每一个专业都有一些课程是必修课，几乎所有的学生都需要选择这些课程，这导致最后相当数量的关联规则的支持度和置信度很高而实际意义不大。

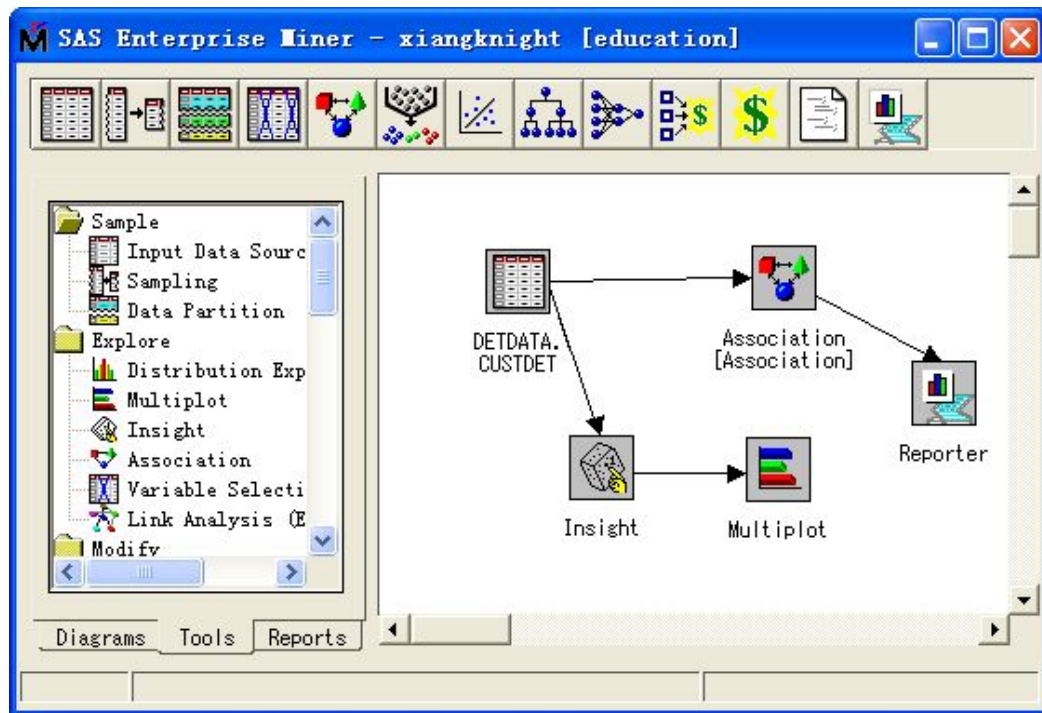


图 3-13、学生选课特征关联模型

所以，需要利用约束关联规则的方法，提取研究者所关注的规则。在 Enterprise Miner 中可以通过 Subset Table 设置支持度、置信度以及作用度等兴趣度约束并指定规则前后件等规则约束方法对规则进行剪除。例如，从图 3-14 可以看出，挖掘出的关联规则的支持度集中在 86.28% 以下，对于支持度更大的规则根据前面的先验知识分析可知，大部分应该是无趣子集，所以设置支持度的约束范围为 $9.72\% \leq \text{support} \leq 86.28\%$ 。同理，根据其他兴趣度的统计情况，再设置 $60\% \leq \text{confidence}$, $1.35 \leq \text{lift} \leq 10.29$ ，最后得到的关联规则有 142 条（部分结果见表 3-6）。此外，如果研究者希望关注选择哪些课程对选择某门课程有影响，则可以设置规则的后件为该门课程，进行规则约束挖掘。例如我们想了解哪些课程对选择“网页设计基础”课程有影响，则设置后件为“网页设计基础”，经过挖掘我们得到了 31 条规则。在这 31 条规则中，发现了一些有趣的规则，见表 3-6 中提取到的 R2 规则：如果选修了“中国茶文化”课程，则选修“网页设计基础”课程的可能性是 87.5%，有 9.72% 的选课记录支持该结论，根据 $\text{lift}=2.03$ 可知：选修“网页设计基础”课的可能性提高到原来的 2.03 倍。对于此结论我们可以理解为：该校 xx 级计算机专业的学生中对“中国茶文化”感兴趣的学生同时也更愿意学习网页设计方面的课程。常识使我们想到，这两门课程都具有文化艺术的特点，说明“中国茶文化”课程与“网页设计基础”课程具有相近的性质，也说明该校 xx 级计算机专业的学生中有 9.72% 的学生对文化艺术感兴趣。教学课程安排时可以考虑通过课程设置和对学生的引导，来激发学生学习该门课程的兴趣。

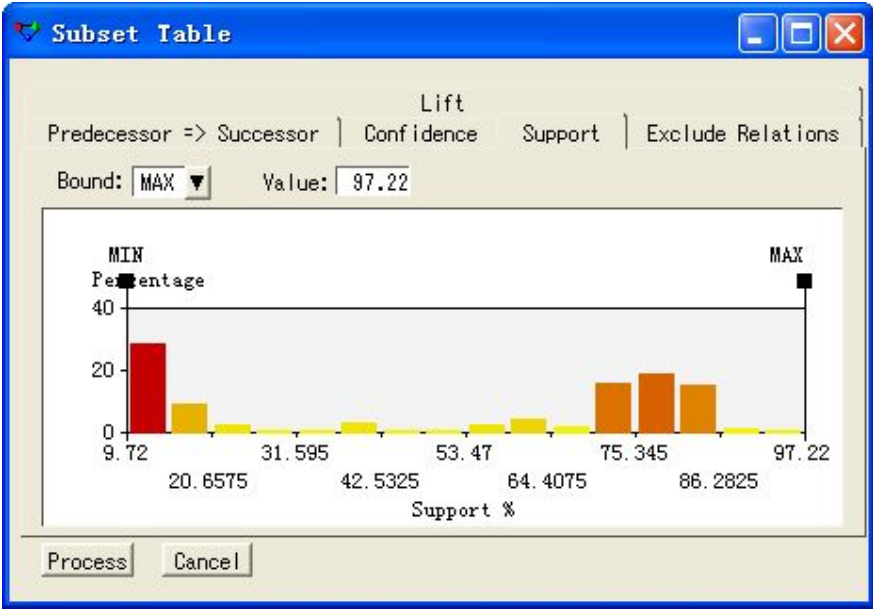


图 3-14、支持度的设置

表 3-6 约束后的部分规则列表

No	Relations	Lift	Support (%)	Confidence (%)	Count	Rules
1	2	1.45	11.11	88.89	8.00	线性代数（A） ==> 大学英语（一）
2	2	2.03	9.72	87.50	7.00	中国茶文化 ==> 网页设计基础
...
9	3	1.38	19.44	100.00	14.00	口语（辅修） ==> 形势与政策 & 大学英语（二）

关联规则挖掘能向用户揭示数据库中一组对象与另一组对象之间存在的内在关联关系，但挖掘出的规则条数成千上万，使得用户无所适从。若想提高挖掘结果的准确性，合理运用约束关联规则挖掘方法并结合要挖掘的集合内容给出合理的支持度、置信度以及作用度，可以帮助研究者得到有效的结果。即在运用关联规则进行数据挖掘时，合理运用约束条件，能够有效地解决这类问题。

本研究使用不同的约束条件，对学生选课信息进行了研究，取得了有效的结果。说明通过对教学信息进行约束关联数据挖掘，可能挖掘出一些有意义的规则，对于学校进行教学课程设置及改革能够提供有意义的参考和借鉴。

3.6 必修课成绩回归模型

SAS 软件提供了前向选择法、后向选择、逐步回归分析法、最大相关法、最小相关法、全部引入法等 9 种建立回归方程的方法。

其中，前向选择法，模型开始时没有候选因素，然后逐个加入与目标显著相关的因素，直到没有满足显著水平的因素或达到满足停止的标准。这种算法适用于有大量候选因素的情况。

各个专业中的必修课的成绩间存在着一定的关系，尤其是那些课程学习存在较强依赖的，教学决策者希望通过建立必修课成绩的回归模型，了解学生目前学习情况，以及预测未来考试成绩，掌握学习发展趋势。

这里我们重点对前面课程间成绩影响关联模型发掘出前后成绩有重要影响的课程进行回归分析。以对信管专业 021 班成绩记录进行回归分析为例，经过处理后得到的数据集 Em_reg.xg021Em 共有 29 条记录，165 个变量。

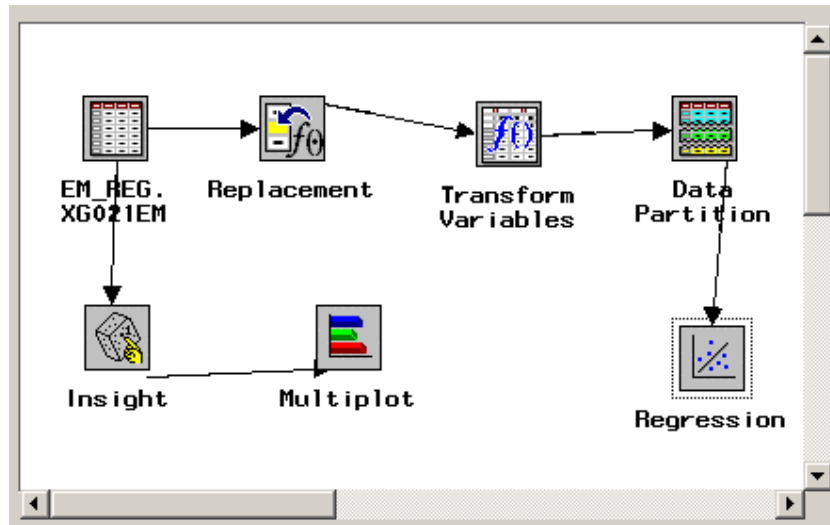


图 3-15、必修课成绩回归模型

添加 Input Data Source 节点。将缺失值严重的变量的模型角色设置为 rejected，保留了 38 个变量，其中设置 kch198123 的变量（代表《系统工程课》，为必修课，开课时间 2004-2005-1）为 target 角色，其余设置为 input 角色。

添加 Data Partition 节点。Method 采用简单随机抽样法（simple random），设置输入数据集的 80% 为训练数据集（training data set），20% 为验证数据集（validation data set）。

添加 Regression 节点。回归分析类型选为线性回归（Linear），Link function 为 Identity。对 coding class variables 的方法有 Deviation 和 GLM，默认选 Deviation。这里不禁止截距（即不选中 suppress intercept）。分析方法选择前向回归分析（Forward），Criteria 选择 none。显著度接受默认值，分别为 entry=0.5, stay=0.5。

运行后，结果为：

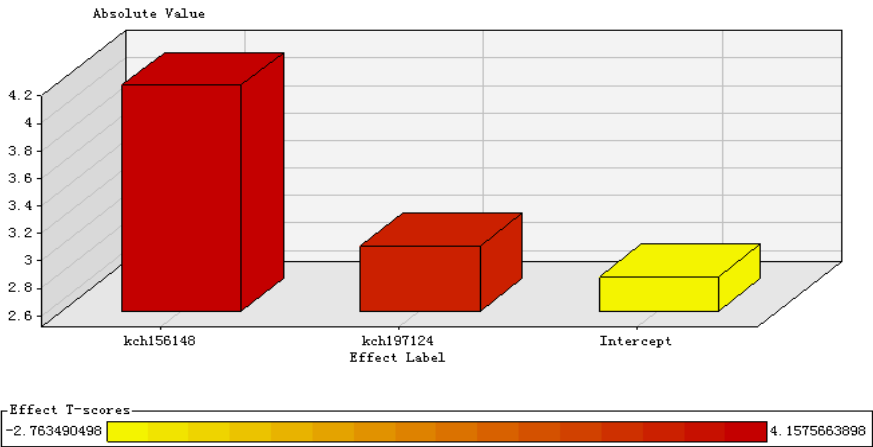


图 3-16、estimate

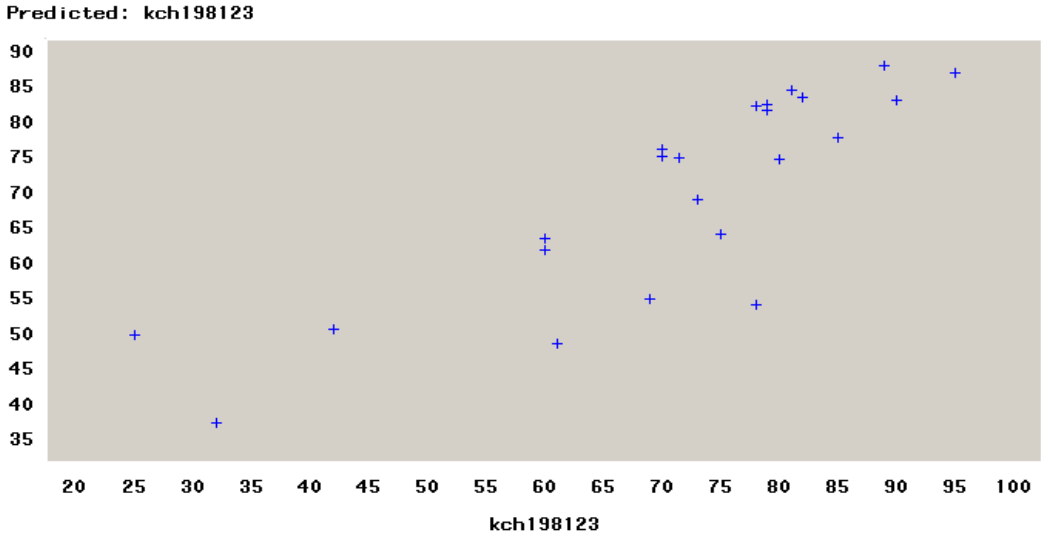


图 3-17、Plot

图 3-17 中反映回归模型中预测值与实际值的符合度。如果符合度好的，则回归模型是有用的。散点集中在中分线说明符合度好，即模型较符合。

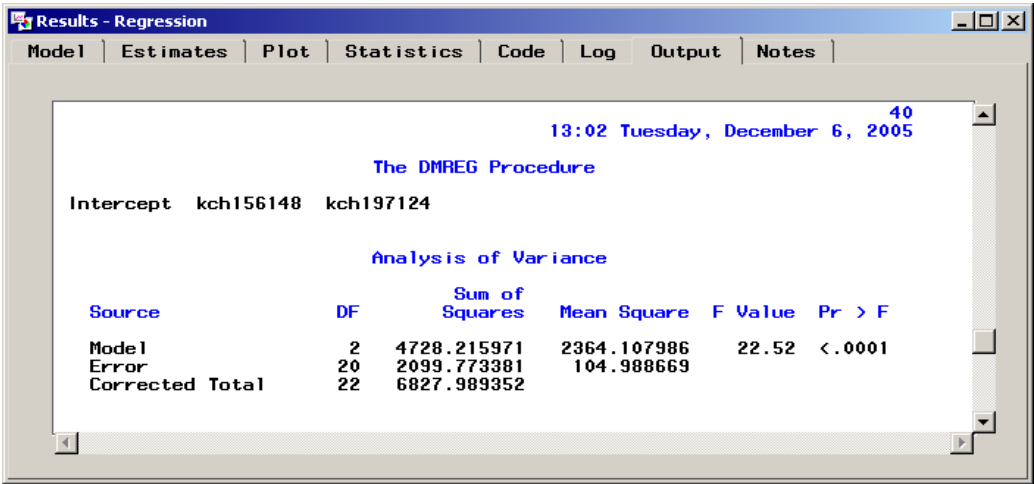


图 3-18、模型信息

图 3-18 用于判断模型的统计学意义，由 $P < 0.05$ 知此模型具有统计学意义。

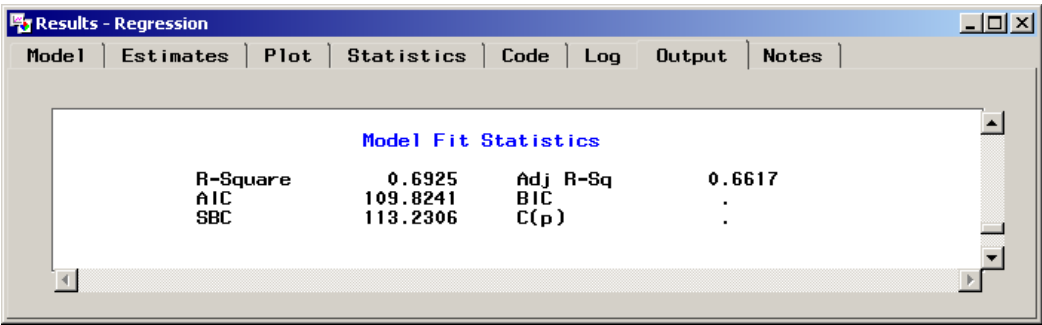


图 3-19、模型拟合的统计量

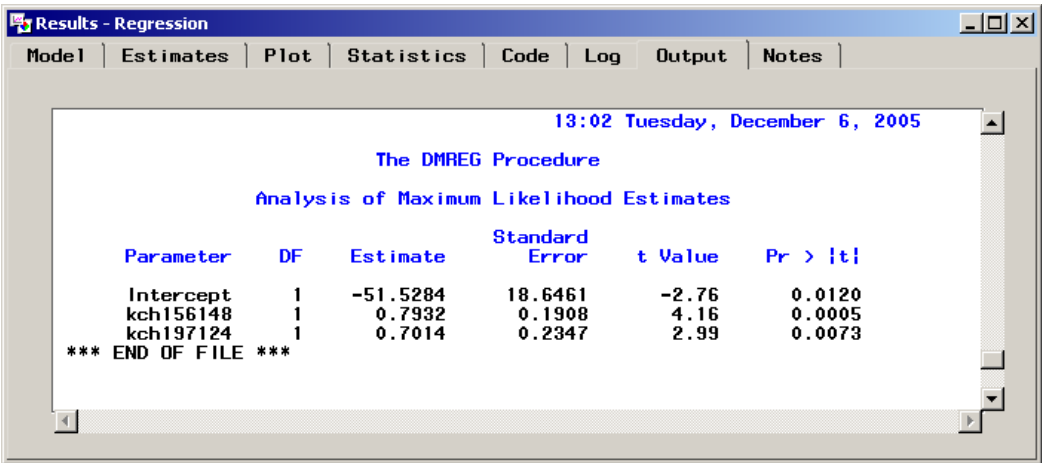


图 3-20、模型的参数估计

图 3-20 分别给出截距项和自变量回归系数等的估计值以及对应的假设检验结果。 $P > |t|$ 的值均小于 0.05，具有高显著度。综合上述结果，得到的回归方程为：

$$\text{系统工程成绩} = \text{大学数学成绩} \times 0.7932 + \text{地学基础} \times 0.7014 - 51.5284$$

课程间成绩影响关联模型从课程间关联的程度来发现课程成绩的关系，而成绩回归模型能进一步的对这些课程成绩挖掘出明确的回归方程。这两种方法相结合，能详细地显示相关课程间的成绩关系。这对教学管理掌控课程间影响关系有重要参考价值。

第四章 基于 Web 的数据挖掘系统

4.1 系统的设计

4.1.1. 系统的特征

OLAM 建立在多维数据库和 OLAP 技术的基础之上,可以对数据立方体进行上卷、下钻、切片、切块、旋转等操作,结合可视化技术能为用户提供丰富的数据探索功能。同时,由于使用的数据是经过清理、集成和过滤等预处理得到的数据仓库或数据立方体,因此能够高质量地进行各种数据挖掘活动。基于 Web 的 OLAM 更具有开放性,能为多用户提供一致的前端展示工具,而且其数据集存储于服务器数据仓库中,可有效地保护数据和随时对信息做出决策。

传统的数据挖掘系统主要基于 C/S (Client/Server) 结构,具有系统通讯开销小和能充分利用两端硬件环境的优点。但 B/S(Browser/Server)结构的数据挖掘系统的优势更加突出:

(1) 节约投资。B/S 软件在初期一次投入成本后,一般不需要再投入。

(2) 简化工作。B/S 软件安装在服务器端即可解决问题,需要更改时,只需要调整服务器端。C/S 结构的软件则需要安装在客户端,调整时要涉及每台客户机。

(3) 数据安全。B/S 结构下的数据集中存在于中央数据库中,可有效的保护数据,并可随时掌握企业信息做出决策。C/S 结构下需要在各地分别安装区域级服务器,一旦出问题,威胁数据安全。

(4) 网络限制。B/S 结构适用于各种网络。C/S 结构则仅适用于局域网内用户或宽带用户。

基于上述优点,目前各种数据挖掘软件都在积极地向 B/S 结构发展。B/S 体系结构使得用户通过浏览器可以访问多个应用服务器,形成点到多点、多点到多点的体系结构模式。由于客户端使用单一的 HTML 语言,因而简化了客户端系统的管理和使用,使系统的管理和维护集中在服务器端,而较大的带宽使得 B/S 体系结构具有较强的交互处理能力。

4.1.2. 系统的体系结构

基于 Web 的 OLAM 系统的体系结构如图 4-1 所示。

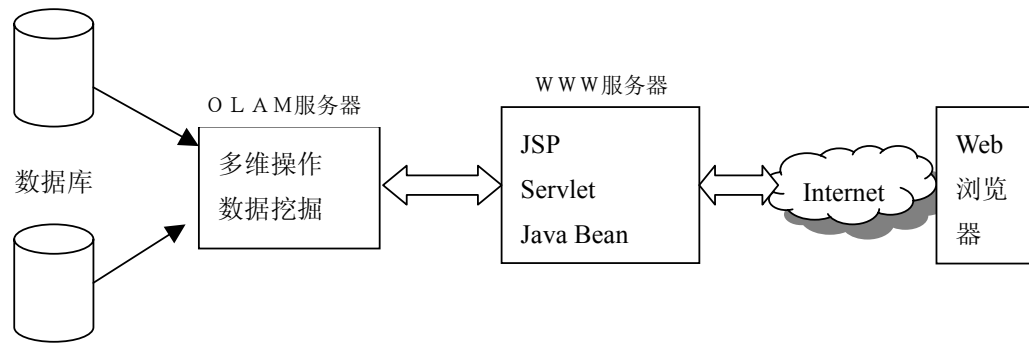


图 4-1、系统结构图

从图中可以看出，OLAM 系统为用户在线进行多维分析和数据挖掘提供接口。首先，用户通过 HTML 文件中的表单提出数据分析挖掘请求，并提交给 WWW 服务器；其次，WWW 服务器端调用相应的应用程序，并根据需要激活 OLAM 服务程序。OLAM 服务器引擎根据用户请求，在元数据的指导下，将立方体操作译为 SQL 请求，并交给 DBMS(DWMS)执行。最后，系统将数据挖掘分析的结果以可视化的形式返回给用户端的 Web 浏览器。

4.2 系统的实现

4.2.1. 采用的技术与服务

1、SAS/IT 技术

主要使用 SAS/IT 中的两种服务：

IOM(the Integrated Object Model)，提供分布式对象接口。IOM 允许用户使用工业标准语言、编程工具和通信协议来开发客户程序访问 IOM 服务器上的服务。通过 IOM 桥通信协议，不同的用户可以透明地连接到多平台上的 IOM 服务器。

SAS Foundation Services，是核心体系服务的集合。Java 程序员可以使用它编写分布式应用程序同 SAS 平台进行集成。此服务提供对 IOM 服务器的客户连接、动态服务探索、用户授权、概要管理、会话上下文管理、元数据与内容存储库访问、活动日志、事件管理、信息发布和存储过程执行。

2、SAS Java Components

SAS Java Components 中拥有大量的模型和组件，其中包含基于 JSP/Servlet 的可视化组件和图形组件、基于 Swing 的组件、通过 JDBC 和 Information Maps 访问关系数据的模型、通过 Information 访问 OLAP 数据的模型、以及各种 Utility 类。

3、SAS Management Console

SAS Management Console 是 Java 应用程序，可使用单个界面执行创建和维护跨平台的集成环境所需的管理任务，而不必为计算环境中的每一个应用程序提供单独的管理界面。通过 SAS

Management Console 可以管理服务器定义、逻辑库定义、用户定义、资源访问控制、元数据储存库、SAS 许可、作业预定和 XML 映射。

SAS Management Console 通过为每一个计算资源或控制创建和维护元数据定义来实现其功能。这些元数据定义存储在 SAS 元数据服务器的元数据储存库中，其他应用程序可以使用元数据储存库中的这些元数据。

4、SAS Information Map Studio

SAS Information Map Studio 是一个可以创建和管理 SAS 信息映射（关于你的物理数据的业务元数据）的应用程序。它提供了一个图形化用户界面，使得用户可以创建和浏览信息映射。信息映射是对物理数据源的界面友好的元数据定义，能让用户通过查询一个数据仓库来满足特定的业务需要。一个信息映射包含数据项和过滤器。数据项对应一个物理数据字段或者计算。过滤器包含子集数据的标准。SAS Information Map Studio 也允许设置不同的属性来控制查询产生，查询执行，以及数据访问。

5、SAS AppDev Studio

SAS AppDev Studio 为开发瘦客户端和胖客户端的商业智能应用程序提供了单一界面，可运行在 MS Windows 98/2000/NT 平台。它支持服务器端和客户端的 Web 标准。

SAS AppDev Studio 中提供的服务包括：

AppDev Studio Middleware Server (MWS) 可以使多客户通过成为 funneling 的过程使用相同的 SAS session。

The Integrated Object Model (IOM) spawner 可以使你使用 SAS Integrated Object Model 连接远程对象。

IOM Server 用于 Windows95 和 Windows98 开发环境。

The AppDev Studio Service Manager 可以使你在 launch configured AppDev Studio 服务时，来指定你想启动哪个服务。

A Java Web server 可以使你开发测试和共享服务器端 Java 技术，如 JavaServer pages 和 servlets。完成 webAF 工程可能需要使用一个 Java Web 服务器。

SAS/CONNECT Spawner 可以使你使用 SAS/CONNECT 连接远程的 SAS 主机。spawner 程序能对通过网络传递的用户 id 和密码进行加密。

SAS/SHARE Data Server 可以使两个或多个客户同时写相同的 SAS 文件。

webAF 软件是集成可视化开发环境，能快速建立 java 应用程序，java 小应用程序，Web 应用程序和类，通过拖拽面向对象的接口以减少编程量。通过专用的 Java 类能方便地访问 SAS 软件，透明访问 SAS/AF 对象，访问表和 MDDb，访问 SAS 计算功能。webAF 具有 Jakarta ANT 支持，内嵌有 Web 服务器 Apache 和应用服务器 Tomcat，所开发的应用程序可以部署到任何支持 J2EE 标准的应用服务器上，如 IBM WebSphere 和 BEA Weblogic。

4.2.2. 数据源的设置

由于数据挖掘系统不同于传统的数据管理系统，所使用的数据不仅包括关系数据，还包括 OLAP 数据，并且需要针对海量数据访问提供快速响应，因此设置适当的数据源形式至关重要。

本系统中由 SAS 应用服务器集中对数据进行采集、集成、清理和管理，形成 OLAM 分析所需的数据集和数据立方体。通过 IOM 访问这些数据的方式包括 JDBC 和信息映射两种。通过信息映射不仅可以对常规的关系型数据进行访问，还能对联机分析处理所要求的多维数据进行操作。通过此方式，服务器端可以根据需要来决定用户访问的权限，并根据领域知识提供特定数据集合的过滤器，方便用户访问。

本系统的数据源设置步骤为，先在 SAS Management Console 中建立 SAS 逻辑库，然后利用 SAS Information Map Studio 将数据表和立方体映射为信息映射，供开发的数据挖掘系统访问。

首先，使用 SAS Management Console 时，应用程序需要通过元数据配置文件连接至元数据服务器。元数据配置文件指定要连接的元数据服务器和元数据储存库，以及连接时使用的用户凭证。

然后使用“数据逻辑库管理器”定义逻辑库。“数据逻辑库管理器”具有管理存储在“SAS 开放式元数据储存库”中的逻辑库和数据库模式定义的功能。其他功能还包括：定义新的逻辑库和数据库模式、更改逻辑库和模式的属性以及更改逻辑库和服务器之间的分配。

定义新逻辑库：

- 1、选择“数据逻辑库管理器”下的“SAS 逻辑库”文件夹。
- 2、从工具栏、弹出菜单或“操作”菜单中选择“新建逻辑库”
- 3、“新建逻辑库向导”启动。该向导可用于定义新的 SAS 逻辑库。
- 4、您可以定义的逻辑库类型取决于可用的逻辑库资源模板。如果您需要定义的逻辑库类型未在向导中列出，您必须加载适用的资源模板。
- 5、“新建逻辑库向导”会引导您完成定义逻辑库的整个过程，包括输入逻辑库的名称、逻辑库引用名、引擎和路径（不是所有类型的逻辑库都需要引擎和路径）。您还可以指定逻辑库的高级选项，包括访问级别、编码和平台专有的选项。

这里我们建立逻辑库“webdata”，引擎选用 Base，路径设为要导入的数据表的所在路径。

然后导入数据表，这里要导入课题中的数据仓库和数据挖掘模型相关的数据表。右击“webdata”->“导入表”，启动“连接 SAS”窗体

连接 SAS 服务器后，会显示前面设置路径下的所有数据表，用户可以根据需要选择要导入的数据表。

定义完逻辑库后，就可以在 SAS Information Map Studio 中映射这些数据表了，见图 4-2。

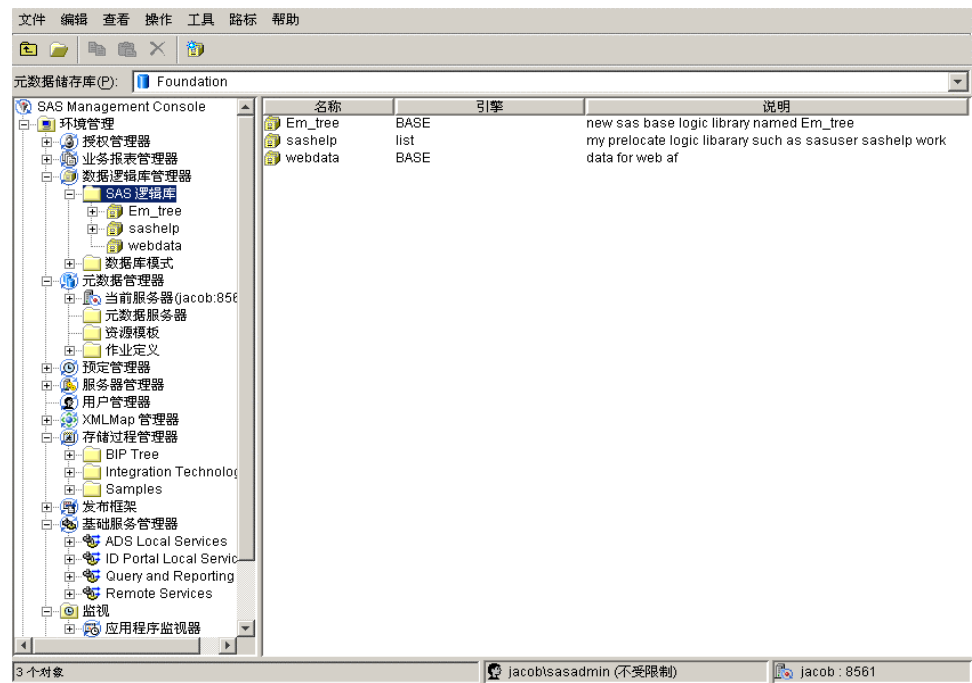


图 4-2、数据逻辑库管理器

在 SAS Information Map Studio 中创建新的信息映射，先从菜单选择“插入”->“表”或“插入”->“立方体”。

点击“确定”后，“物理数据”显示刚才选择的数据表及字段，然后选择需要映射的字段到“信息映射”。

保存，此信息映射将被保存到 SAS Management Console 中定义的存储库“Foundation”下。

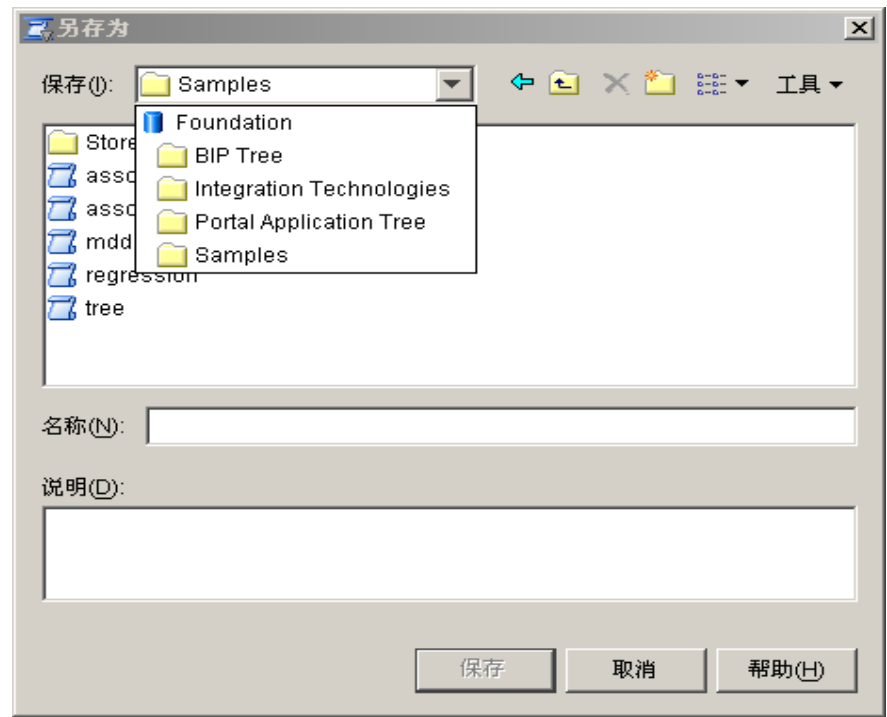


图 4-3、建立信息映射

此外，还可以根据需要创建一些过滤器，使得用户可以方便的查看自己关心的数据子集。这里创建了一个多维数据库中专业为计算机的数据子集。过滤器建立后，会被保存在信息映射中。

以后用 WEB AF 开发的 Web 应用程序就可以通过 Information Map 访问这些数据了。

4.2.3. Web 应用程序开发

Model-View-Controller(MVC)设计模式最初是在 20 世纪 70 年代由施乐（Xerox）Palo Alto 研究中心(PARC, Palo Alto Research Center)被提出的。MVC 模式最先被用来在第一代基于视窗的计算机上管理 GUI 和用户交互[49]。近些年开始被推荐应用在 J2EE 平台上。

本 Web 应用程序采用 MVC 设计模式开发，如图 4-4。

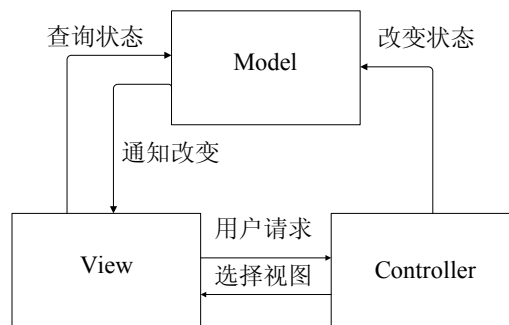


图 4-4、MVC 设计模式

Model 组件表示应用程序的数据，并包括这些数据的访问和修改得业务规则，负责在后台存储或远程系统中维护数据。本系统采用 Visual Data Explorer 组件作为 Model 组件，这是一个 JavaBean，负责与 SAS 服务器中数据源连接。这里数据源采用前面定义的信息映射。

View 组件是用户看到并与之交互的界面，主要负责从模型访问数据指定如何表示数据，并当模型改变时，维护表示的一致性。这里采用 Jsp 页面履行 View 的职责。

Controller 组件定义应用程序的行为，解释用户动作，并把它映射为模型执行的过程。它负责模型和视图之间的交互，控制对用户输入的响应方式和流程。这里用 Servlet 作为 Controller 组件，用于处理控制和设置业务逻辑。它主要包括两个动作：一方面将用户的请求分发到相应的模型，另一方面将模型的改变及时反映在视图上。

Web 应用开发可以利用 SAS/BI 提供的完备的开发包，即节省开发成本又能提供丰富的功能，如 InfoMapViewControllerServlet.java 在该系统中负责控制作用，需要引入 com.sas.actionprovider 包、com.sas.iquery 包、com.sas.services 包、com.sas.servlet 包等，具体引用类如下：

```

import java.io.*;
import javax.servlet.*;
import javax.servlet.http.*;
import com.sas.actionprovider.HttpAction;
import com.sas.actionprovider.HttpActionProvider;
import com.sas.actionprovider.support.ActionProviderSupportTypes;
    
```

```

import
com.sas.actionprovider.support.remotefileselector.HttpRemoteFileSelectorTableViewSupport;
import com.sas.iquery.metadata.IntelligentQueryMetadataService;
import com.sas.iquery.metadata.IntelligentQueryMetadataServiceFactory;
import com.sas.services.discovery.DiscoveryService;
import com.sas.services.discovery.DiscoveryServiceInterface;
import com.sas.services.discovery.ServiceTemplate;
import com.sas.services.information.RepositoryInterface;
import com.sas.services.session.SessionContextInterface;
import com.sas.services.session.SessionServiceInterface;
import com.sas.services.user.UserContextInterface;
import com.sas.services.user.UserServiceInterface;
import com.sas.servlet.tbeans.remotefileselector2.html.InformationServicesSearch;
import com.sas.servlet.tbeans.remotefileselector2.html.InformationServicesSelector;
import com.sas.servlet.tbeans.remotefileselector2.RemoteFileSelectorKeysInterface;
import
com.sas.swing.models.remotefileselector2.information.services.BaseInformationServicesModel;
import
com.sas.swing.models.remotefileselector2.information.services.InformationServicesNavigationModel;
import com.sas.web.keys.ComponentKeys;
import java.util.List;
import java.util.ArrayList;
import listeners.ExamplesSessionBindingListener;

```

必须对一些属性文件进行配置，如 `sas_metadata_source_omr.properties` 定义了元数据和本地基础服务的信息，Web 应用程序会从该文件读取设置的各种信息。部分设置如下：

```

software_component=ADS Local Services
deployment_group_1=BIP Core Services

omr_host=JACOB
omr_port=8561
omr_repository=Foundation
omr_user=sasadmin
omr_password=admin

```

`FoundationServicesContextListener.java` 监听类会在应用程序启动时通过读取 `sas_metadata_source_omr.properties` 文件部署本地基础服务。

web.xml 包含了元数据用户信息和应用程序配置信息。部分设置如下：

```
<!-- Context parameter that specifies the local SAS Foundation Services
      metadata source properties file. -->
<context-param>
  <param-name>local.sas.foundation.services</param-name>
  <param-value>/WEB-INF/conf/sas_metadata_source_omr.properties</param-value>
</context-param>

<!-- Declare ServletContextListener to deploy and destroy local
      SAS Foundation Services -->
<listener>
  <listener-class>listeners.FoundationServicesContextListener</listener-class>
</listener>

<!-- Declare SAS MethodInvocationServlet -->
<servlet>
  <servlet-name>MethodInvocationServlet</servlet-name>
  <servlet-class>com.sas.servlet.util.MethodInvocationServlet</servlet-class>
</servlet>

<!-- Declare SAS SelectorServlet -->
<servlet>
  <servlet-name>SelectorServlet</servlet-name>
  <servlet-class>com.sas.servlet.tbeans.dataselectors.SelectorServlet</servlet-class>
</servlet>

<!-- Declare SAS StreamContentServlet -->
<servlet>
  <servlet-name>StreamContentServlet</servlet-name>
  <servlet-class>com.sas.servlet.util.StreamContentServlet</servlet-class>
</servlet>

<!-- Declare SAS VisualDataExplorerServlet -->
<servlet>
  <servlet-name>VisualDataExplorerServlet</servlet-name>
  <servlet-class>com.sas.servlet.util.VisualDataExplorerServlet</servlet-class>
</servlet>

<!-- Declare InfoMapViewControllerServlet -->
<servlet>
  <servlet-name>InfoMapViewControllerServlet</servlet-name>
```

```

        <servlet-class>servlets.InfoMapViewControllerServlet</servlet-class>
        <!-- Servlet init parameter 1 -->
        <init-param>
            <param-name>metadata-domain</param-name>
            <param-value>DefaultAuth</param-value>
        </init-param>
        <!-- Servlet init parameter 2 -->
        <init-param>
            <param-name>metadata-userid</param-name>
            <param-value>sasguest</param-value>
        </init-param>
        <!-- Servlet init parameter 3 -->
        <init-param>
            <param-name>metadata-password</param-name>
            <param-value>guest</param-value>
        </init-param>
    </servlet>

    <!-- Standard Action Servlet Mapping -->
    <servlet-mapping>
        <servlet-name>action</servlet-name>
        <url-pattern>*.do</url-pattern>
    </servlet-mapping>

    <!-- Standard MethodInvocationServlet mapping -->
    <servlet-mapping>
        <servlet-name>MethodInvocationServlet</servlet-name>
        <url-pattern>/MethodInvocationServlet</url-pattern>
    </servlet-mapping>

    <!-- Standard SelectorServlet mapping -->
    <servlet-mapping>
        <servlet-name>SelectorServlet</servlet-name>
        <url-pattern>/SelectorServlet</url-pattern>
    </servlet-mapping>

    <!-- Standard StreamContentServlet mapping -->
    <servlet-mapping>
        <servlet-name>StreamContentServlet</servlet-name>
        <url-pattern>/StreamContentServlet</url-pattern>
    </servlet-mapping>

```

```
</servlet-mapping>
<!-- Standard VisualDataExplorerServlet mapping -->
<servlet-mapping>
    <servlet-name>VisualDataExplorerServlet</servlet-name>
    <url-pattern>/VisualDataExplorerServlet</url-pattern>
</servlet-mapping>

<!-- Standard InfoMapViewControllerServlet mapping -->
<servlet-mapping>
    <servlet-name>InfoMapViewControllerServlet</servlet-name>
    <url-pattern>/InfoMapView</url-pattern>
</servlet-mapping>

<!-- The Usual Welcome File List -->
<welcome-file-list>
    <welcome-file>index.jsp</welcome-file>
</welcome-file-list>
</web-app>
```

编译整个项目后,需要启动 Java Web server,点击“工具”->“服务”->“启动 Java Web server”。最后就可以从浏览器中执行 Web 应用程序了。

本研究开发的 Web 应用程序通过 SAS/IT 中 IOM(the Integrated Object Model)分布式对象接口同 OLAM 服务器进行交互。OLAM 服务器收到 Web 应用程序转发的用户数据挖掘请求后,通过 SAS 基础服务与 SAS 应用服务器进行通信,对 SAS 数据集和数据立方体进行探索、展示和数据挖掘。

4.2.4. 基于 Web 的数据挖掘系统的应用

本论文通过阐述对我校信息与电气工程学院的学生信息库中学号为“0101004”的公共课的挖掘方法和结果,对研究的基于 Web 的 OLAM 系统进行应用说明。挖掘的数据包括 99 级到 04 级电气、电信、电子和计算机 4 个专业的记录,涉及学生学号、姓名、选课时间、课程号、课程名称、课程学分、学时以及成绩等信息。

(1) 可视化探索

该系统能够通过表和图形的方式展示信息映射数据,包括散点图、直方图、条线图、曲线图、饼状图、彩色映射表等,提供多方位的教学信息的探索。例如,图 4-5 在线展示了电气专业某班学生该门课程的成绩,其中横坐标为各个学生的学号,纵坐标为该课程的考试成绩。从图中我们可以清楚看出,该门课程考试成绩主要介于 60 与 90 分之间。经过对各班该门课程的可视化探索后,根据教学经验和挖掘需要,对考试成绩进行属性概化,映射到一个有序概念域

$D=\{\text{bad,past,good}\}$ 。对该门课程考试成绩的可视化探索,为提交数据挖掘请求提供了参考。

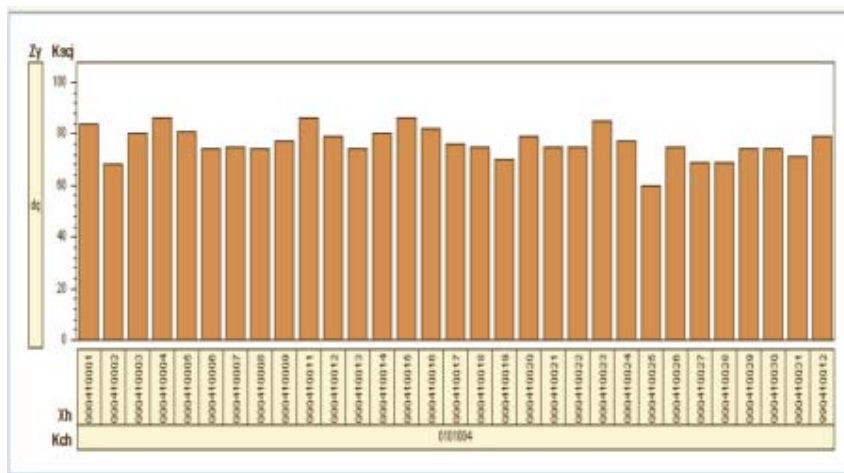


图 4-5、学生学习成绩直方图

(2) 数据挖掘分析

通过此数据挖掘系统进行在线数据挖掘分析时，首先通过 Web 应用程序浏览 SAS 应用服务器上存储的相关信息映射数据集，选取专业 zy，学分 xf，学时 xs，教师号 jsh 及概化后的考试成绩 ksdj 等字段，选择数据挖掘方法并设置目标字段，然后提交表单。Controller 组件接收到用户请求后，通过 IOM 接口将挖掘任务转交给 SAS/IT 服务器，再调度 SAS 应用服务器运行挖掘功能；Model 组件负责维护用户选取的字段信息及数据，并与 SAS/IT 服务器保持通信。SAS 应用服务器运算完毕后，SAS/IT 服务器将结果返回给 Web 应用程序，由 View 组件将结果以可视化的形式呈现给用户。

OLAM 技术结合了数据挖掘和 OLAP 技术的数据多维性、分析在线性、挖掘多样性等优点,是数据挖掘领域的一个重要发展方向。基于 Web 的 OLAM 系统能够为用户提供友好、直观、多样的数据挖掘方法。该系统应用于教学信息管理分析,方便了教学管理者和教师在线对教学信息进行数据挖掘,同时方便更多的相关人员在线浏览挖掘结果,对于更有效地管理和监督教学质量,公平、公正地评价相关课程提供了有效的帮助,能够有效地促进教学质量的提高和改革。

第五章 结论与建议

数据挖掘技术在教学领域的研究应用是数据挖掘领域研究的方向，本论文的主要工作与贡献如下：

- 1、教学质量评估长期处于难于量化的状况，这给教学管理带来极大不便，将数据挖掘技术应用于教育教学领域是教学研究的一个新热点。
- 2、本研究探讨了运用 SAS/Enterprise Miner 软件按照 SEMMA 方式建立各类教学信息挖掘模型的方法，所得模型可重复用于教学历史数据的分析，对量化教学质量评估提供可靠参考。
- 3、探讨了基于 Web 的 OLAM 系统的体系架构与实现，并应用于教学信息挖掘，获取隐藏的信息，为教学决策提供辅助和指导。

本文对基于 Web 的 OLAM 技术进行了一些探讨性的研究和尝试，还存在一些不完善的地方，需要进一步的研究和完善，主要表现在如下方面：

- 1、教学历史数据数量庞大而复杂，要得到非常合适的数据挖掘模型需要不断地实验和行业知识指导。
- 2、由于 SAS 软件提供的用于 web 开发的接口有限，目前只能开发简单的 web 访问程序，有待进一步完善。

整合 OLAP 和数据挖掘技术的 OLAM 技术是该领域发展的趋势，随着研究的深入和开发商提供的软件接口的完善，在教学领域一定会得到广泛应用。

参考文献

- [1]W.H.Inmon 著, 王志海等译, 数据仓库, 北京: 机械工业出版社, 2000.5, P20-22;
- [2]李湘军, 数据挖掘研究的综述, 中国科技论文在线,
<http://www.paper.edu.cn/process/download.jsp?file=200405-2;>
- [3]数据挖掘资料汇编, 数据挖掘讨论组, <http://datamining.126.com/> 2002.12;
- [4]高延铭.数据挖掘在通信行业 CRM 中的应用研究: [硕士学位论文], 青岛: 中国海洋大学, 2003;
- [5]周宇.基于数据挖掘的电信领域客户流失分析: [硕士学位论文], 北京: 北京邮电大学, 2003;
- [6]贾琳 李明, 基于数据挖掘的电信客户流失模型的建立与实现, 计算机工程与应用, 2004 年 04 期;
- [7]马国厚.基于数据仓库的保险决策支持系统: [硕士学位论文], 武汉: 武汉水利电力大学, 2000;
- [8]宋向平.数据挖掘技术在车险 CRM 中的应用研究: [硕士学位论文], 杭州: 浙江大学, 2003;
- [9]陈永强 胡雷芳, 数据挖掘技术在人寿保险 CRM 系统中的应用研究, 成组技术与生产现代化, 2004 年 01 期;
- [10]范习辉.数据挖掘在电力系统警报信息处理中的应用研究: [博士学位论文], 武汉: 华中科技大学, 2003;
- [11] 李扬 王治华 卢毅, SAS 在电力负荷特性分析及预测方面的应用, 中国电力, 2002 年 06 期;
- [12]于长春 贺佳等, 数据挖掘技术在肝癌术后预测分析中的应用初探, 第二军医大学学报, 2003 年 11 期;
- [13]朱蔚恒.数据挖掘在医疗中的应用: [硕士学位论文], 广州: 中山大学, 2003;
- [14]王珊, 数据仓库与联机分析处理, 北京: 科学出版社, 1998;
- [15]张维东.数据仓库、OLAP 及数据挖掘技术的研究与设计: [博士学位论文], 上海: 同济大学, 2000;
- [16]陈长清.数据仓库与联机分析处理技术研究: [博士学位论文], 武汉: 华中科技大学, 2002;
- [17]嵇晓.数据仓库工程方法论的研究及在一个大型钢铁企业中的实践: [博士学位论文], 沈阳: 东北大学, 2002;
- [18]龚晶.数据仓库技术在销售决策支持系统中的应用研究: [硕士学位论文], 武汉: 武汉大学, 2004;
- [19]徐建锋.基于数据仓库的分类分析研究: [硕士学位论文], 青岛: 青岛大学, 2003;
- [20]杨士哲.基于数据仓库的决策支持系统的研究与开发: [硕士学位论文], 杭州: 浙江大学, 2002;
- [21]Jiawei Han, OLAP Mining: An Integration of OLAP with Data Mining, IFIP1997, Chapman&Hall Press;
- [22]曹菊光.联机分析挖掘处理技术(OLAM)的研究: [博士学位论文], 杭州: 浙江大学, 2001;
- [23]石磊 石云, OLAP 与数据挖掘一体化模型的分析与讨论, 小型微型计算机系统, 2000 年 11 期;
- [24]刘夫涛 张雷 艾波, OLAM 以及基于 Web 的 OLAM, 计算机工程与应用, 2000 年 09 期;
- [25]Han, H Song, IY Hu, XH Prestrud, A Brennan, MF Brooks, AD, Managing and mining clinical

- outcomes, DATABASE SYSTEMS FOR ADVANCED APPLICATIONS, 2004;
- [26]徐茂祖 张桂花, 教育测量学基础, 北京: 中国铁道出版社, 1995.10;
- [27]黄晓红 张维和, 教育评价定量分析方法的研究, 中国轻工教育, 2004 年 02 期;
- [28]瞿斌 王战军, 联机分析处理技术在研究生教育评估中的应用, 科学学研究, 2003 年 06 期;
- [29]王志敏, OLAP 和数据挖掘技术在 CET-4 和 CET-6 模拟预测系统中的应用, 上海应用技术学院学报, 2004.6;
- [30]邢涛.数据挖掘在高校学生管理系统中的应用研究: [硕士学位论文], 北京: 北京航空航天大学, 2004;
- [31] 姚志刚, 央视国际 <http://www.cctv.com/news/science/20031030/101465.shtml>, 2003.10.30
- [32] W.H.Inmon 著, 王志海等译, 数据仓库, 机械工业出版社, 2000.5, P20-22;
- [33] Jessica Lin, Eamonn Keogh, Wagner Truppel, Clustering of Streaming Time Series is Meaningless, DMKD'03, 2003.6;
- [34] Jiawei Han, Micheline Kamber 著, 范明, 孟小峰等译, 数据挖掘: 概念与技术.机械工业出版社, 2001.8;
- [35] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]. In Proc 1994 Int Conf Very Large Data Bases, Santiago, Chile, 1994, P487-499;
- [36] Agrawal R, Mannila H, Srikant R et al. Fast discovery of association rules. In: Fayyad M, Piatetsky-Shapiro G, Smyth Peds. Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI/MIT Press, 1996.307- 328;
- [37] G. Piatetsky-Shapiro, editor. Notes of AAAI' 91 Workshop Knowledge Discovery in Databases(KDD' 91). Anaheim, CA, July 1991;
- [38] Symth P, Goodman R M. An information theoretic approach to rule induction from databases. IEEE Trans on Knowledge and Data Engineering, 1992, 4 (4) : 301-316
- [39] Toivonen H, Klemettinen M, Ronkainen P et al. Pruning and grouping discovered association rules. In: Mlnet Workshop on Statistics, Machine Learning and Discovery in Database, Gete, Greece, 1995.47- 52;
- [40] 程继华, 郭建生, 施鹏飞, 挖掘所关注规则的多策略方法研究, 计算机学报, 2000 年 01 期;
- [41] 郝雷, 王咏, 盛焕桦, 通过先验知识挖掘更有意义的关联规则, 计算机仿真, 2005 年 03 期;
- [42] John Durkin, 蔡竞峰, 蔡自兴.决策树技术及其当前研究方向, 控制工程, 2005 年 01 期.
- [43] G.V.Kass. An exploratory techniques for investigating large quantities of categorical data. Applied Statistics, 29:119-127, 1980.
- [44] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Monterey, CA: Wadsworth International Group, 1984.
- [45] SAS Institute Ltd. SAS 产品白皮书, 2000.8;
- [46] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han. Generalization and decision tree induction: Efficient classification in data mining. In Proc. 1997 Int. Workshop Research Issues on Data Engineering (RIDE'97), Pages 111-120, Birmingham, England, Apr. 1997.
- [47] Kononenko I. On Basis in Estimating Multi-Valued Attributes [A]. Proc. 14th International Joint

Conference on Artificial Intelligence (IJCAI'95)[C].Montreal ,Canada:[s.n.],1995.1034-1040.

[48]WHITE A P,LIU WEI ZHONG.Bias in Information-Based Measures in Decision Tree Induction.Machine Learning,1994,15(3):299-319.

[49][美]James Turner ,Kevin Bedell 著.Struts KICK START.北京： 电子工业出版社， 2004;

致谢

硕士论文的完成首先要感谢我的导师黄燕副教授，黄老师广博的知识、丰富的见识、严谨的治学态度使我受益匪浅，我的硕士论文从选题、论证、实验到完成，都倾注了导师大量的心血。她的悉心指导使我学到了许许多多有用的知识。我深深感到自己在这几年取得的每一个进步，都离不开导师的培养、关心和鼓励。

在论文期间，得到信息与电气工程学院计算机系吴平教授及各位老师和同学们的热心帮助和大力支持，同时，也离不开为该课题提供数据和指导的院办刘红岩老师及其它老师的协作与帮助，在此一并表示最诚挚的谢意。

最后，谨以本文献给我的父母和家人，多年的求学生涯，是他们给了我精神和物质上的帮助和支持，他们的殷切希望和鼓励是我不懈奋斗的动力，他们对我的支持和理解推动着我勇往直前，希望本文能够使他们多年操劳的心得到一丝慰藉。

李湘军

2006年5月于中国农业大学

个人简介

李湘军：男（1981-），河南原阳人，中国农业大学信息与电气工程学院计算机系计算机应用技术专业硕士研究生，主要研究方向数据仓库、数据挖掘技术。

在校期间发表论文：

- 1、李湘军 刘洪岩 吴平 黄燕，基于 Web 的 OLAM 系统的实现与应用，计算机应用研究[J]，2006.10；
- 2、李湘军 黄燕，基于约束的关联挖掘在教学信息中的应用研究，科技广场（信息科学版）[J]，2005.6；