分 类 号：TP311　　单位代码：10183
研究生学号：201253R278　　密　　级：公　　开

吉 林 大 学

硕士学位论文

（专业学位）

基于数据挖掘技术的学生成绩分析系统的设计与实现

Design and Realization of Analysis System of Students' Scores Based on Data Mining Technology

作 者 姓 名：马丹
类　　别：工程硕士
领域（方向）：软件工程
指 导 教 师：车翔久　教授
培 养 单 位：计算机科学与技术学院

2015 年 6 月

吉林大学博士(或硕士)学位论文原创性声明

本人郑重声明：所呈交学位论文，是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：2015 年 6 月 16 日

基于数据挖掘技术的学生成绩分析系统的设计与实现

Design and Implementation of Analysis System Of Students' Scores Based on Data Mining Technology

作者姓名：马丹

专业名称：软件工程

指导教师：车翔久 教授

学位类别：软件工程硕士

答辩日期：2015 年5 月3 0 日

摘　　要

基于数据挖掘技术的学生成绩分析系统的设计与实现

随着大数据成为网络信息行业的重点词汇，在教育领域应用数据挖掘技术来为教育革命提供动力成为可能。数据挖掘技术对于中小学生以及高等院校学生的学习行为各类表现、学习的成绩以及毕业后的职业规划都可以提供有价值的信息，比如可以改善学生的学习方法、帮助学生发现在作业或考试中存在的一些经常被忽略但却很重要信息、为学生具体学科的学习提供个性化服务、及时发现学生潜在的辍学风险等。

教育数据挖掘 EDM 是一个新兴领域，是为备满足日益增长的教育普遍评价的需要。EDM 侧重于收集、归档和分析学生学习相关的数据并进行评估.在学校对学生的教育中存在很多较为明显的数据，比如学生的入学率、报到率、上课考勤率、辍学率、奖学金分布比率等，当然最重要的是学生各科成绩分数的数据。进一步具体到课堂中也存在类似的数据，比如学生回答问题的准确率、提交的作业的正确率、课堂发言次数、师生互动时间、学生回答问题的平均时长等，这些数据按照数据挖掘技术的流程经过专业的收集、预处理、统计、分类和分析后，可以成为对学生多方位表现预测的依据。

本文设计的成绩分析系统主要是利用数据挖掘算法设计模型来分析教育数据，所要实现的目标有三点：

第一，预测新入学学生是否能在第一学期考试通过。如果预测说一个学生倾向于在考试中失败，那么可以建议学生在考试之前采取额外的努力，提高自己成绩并帮助他通过考试。分类方法包括类似决策树、贝叶斯网络可以实现，本文采用的是 CART 和 C4.5，ID3 决策树算法。

第二，利用数据挖掘技术选择有重修风险的学生。对于高校学生来说，当考试分数低于重修分数的时候需要在下学期开学初参加重修辅导班，并重新考试。这不仅增加了老师的教学负担，也给了学生带来不必要的负担，在数据挖掘技术的帮助下，我们能够更准确的选择有针对性的学生。在本文中，将采用一个基于数据挖掘的方法选择该类的学生。方法的关键是采用基于得分的关联规则技术，该方法具有非常不错的效果，优于传统方法。

第三，通过学生的成绩分析学生之间关系、课程之间关系和影响。在教育中，学生的分数是一个非常重要的定量评价指示器，可以客观地反映教育的影响，是一个重要的科学决策依据。因此，分析和研究学生的分数是非常重要的。学生成绩数据库中的数据呈指数级增长。传统的查询和简单的统计分析无法满足分析的需要，不能为教学捕捉有

用的信息。本文利用聚类算法和决策树挖掘学生的分数，通过分析学生-课程-成绩之间的关系可以获得一些教学和管理有价值的信息。

本文的系统测试采用 SAS Enterprise Miner 平台，它设计为被初学者和有经验的用户使用,它的 GUI 界面是数据流驱动的，且它易于理解和使用。它允许一个分析者通过构造一个使用链接连接数据结点和处理结点的可视数据流图建造一个模型。另外，此界面允许把处理结点直接插入到数据流中。

本文最后利用 SAS Enterprise Miner 平台对本文设计的数据挖掘模型进行测试，包括输入数据源、数据分区点检查、变量缺失值替代、交互分组、决策树模型、聚类分析模型等，测试效果表明成绩分析功能能够实现。

关键词：

成绩预测、学业预警、成绩分析、聚类、关联规则、SAS

Abstract

Design and Implementation of Analysis System

Of Students' Scores Based on Data Mining Technology

Big data become the focus of the network information industry vocabulary, the application of data mining technology in the field of education for the education revolution become possible. Data mining technology in primary and middle school students and college students' learning behaviors of all kinds of performance, learning achievement and career plans after graduation can provide valuable information, such as to improve the students' learning methods, to help students find in some of the homework or exams often ignored but very important information, and it provides personalized service for students in specific subjects, timely finds students potential risk of dropping out of school.

Education data mining is an emerging field, EDM is prepared to meet the growing demand for education universal evaluation. EDM is focused on collecting, archiving and analyzing the data of student studies. There is much more obvious in the data in the school, such as the student's enrollment, registration rate, class attendance rate, dropout rates, scholarship distribution ratio, etc., of course the student score data results in all the subjects is the most important. There is a similar data specific to the classroom, such as the students answer the questions of accuracy, submit homework correctly, a number of and interaction between teachers and students class time, the average time of answer questions. These data can become the basis for students with all-round performance prediction after process through professional collection, pretreatment, statistics, classification and analysis according to the data mining,.

Performance analysis system mainly uses the data mining algorithm design model to analyze the education data. There are three goals:

First, it predicts whether the new entrance students pass through the first semester exam. If it predicts a student tend to fail in the exam, and then you can suggest the students before the exam to take extra effort, to improve their performance and help him to pass the exam. Classification methods including similar decision tree, Bayesian network, this paper USES a CART and C4.5, ID3 decision tree algorithm.

Second, it selects students who have rebuilt risk with data mining technology. College students will rebuild the remedial class and test again when the examination score is lower

than it needs to be rebuilt scores at the beginning of next term. This not only increases the teacher's teaching burden, also gave the students bring unnecessary burden, with the help of the data mining technology; we can more accurately choose targeted students. In this article, will use a data mining based method to choose the class of students. Method is the key technology of association rules based on the score, the effect of this method is very good, is better than traditional methods.

Third, it finds the relationship between student performance analysis of the students, the relationship between curriculum and influence. In education, students' scores of quantitative evaluation is a very important indicator, can objectively reflect the influence of the education, is an important scientific decision-making basis. Therefore, analysis and research on the students' score is very important. The traditional query and simple statistics analysis can't meet the needs of analysis; can not capture useful information for the teaching. This paper, by using clustering algorithm and decision tree mining students' scores, through the analysis of the relationship between the students course - can get some teaching and management of the valuable information.

In this paper, the system test use SAS Enterprise Miner platform, it is designed for beginners and experienced users, its GUI interface is driven data flow, and it is easy to understand and use. It allows an analysis, by constructing a use link nodes and deal with the data of visual data flow graph to build a model. In addition, the interface allows the nodes directly inserted into the data stream processing.

Finally, the paper use SAS Enterprise Miner platform testing data mining model is designed in this paper, including the input data source, data partition point inspection, variable grouping missing value alternative, interaction, and the decision tree model, the cluster analysis model, such as test results show that performance analysis functions can be achieved.

Keywords:

Performance prediction, academic warning, performance analysis, clustering, association rules, SAS

目　　录

第 1 章 绪论

1.1 课题研究背景和意义

传统的学校教育经过一代又一代的口传身受，逐渐形成了一种经验式的教学，老师们更重视的是自己因为对学生重要的一些影响学习的因素，比如老师认为课堂学习很重要，那么就会在提问问题、课堂互动等方面做的多一些。而目前教育领域对于教育教学方式的审核或考评，也是依据某种经验或固定模式在进行，比如在课堂授课过程中，老师能否顺利推进教学环节、提出的问题是否符合知识点、设置的互动环节是否有效等。在这过程中，学生只是作为课堂中的体验者，很多时间被忽视的，至于学生的反应真正是如何的，没有一个完整的、强大的数据集来支撑以供分析和证明[1]。

而数据挖掘技术可以从技术角度以大量数据做依托对学生的表现做充分分析，分析学生的课堂需求和具体课程具体时刻的具体表现，这样的数据就会变得活起来，老师也可以通过这样的数据分析了解学生，进而改变自己的教学行为，改变自己的教学思维，因此说，数据挖掘技术对于教育领域有着重要的作用。

一般来说，数据是指通过科学的实验、充分的检验和统计等方式所得到的，可以用于一些决策、验证、技术设计和研究等工作。通过对这样的数据做全面、准确的测试、收集、分类和存储，在经过严格的统计分析和其他一些检验方法，可以从这些数据中得到具有一定说服力的结论。

在学校对学生的教育中存在很多较为明显的数据，比如学生的入学率、报到率、上课考勤率、辍学率、奖学金分布比率等，当然最重要的是学生各科成绩分数的数据。进一步具体到课堂中也存在类似的数据，比如学生回答问题的准确率、提交的作业的正确率、课堂发言次数、师生互动时间、学生回答问题的平均时长等，这些数据按照数据挖掘技术的流程经过专业的收集、预处理、统计、分类和分析后，可以成为对学生多方位表现预测的依据[2]。

随着大数据的应用越来越广泛，逐渐被各行各业所接受，教育领域同样存在着大量数据等待数据挖掘技术进行分析，从而为教育改革提供参考意见。教育数据包括大学及中学学生在生活、学习、社会实践等多方面，这些都能提供有意义的信息息，利用这些信息，可以帮助学生改变学习方式、找出在学习中存在的容易被忽视的东西，从而给具体学生提供有针对性的指导，及时发出学业预警，防止学生重修或辍学。在这方面，美国教育部已经建立了用于数据统计和挖掘分析的数据存储库，随着我国教育方式的多样化，教育数据将会更加被重视，数据挖掘技术也在教育领域得到发展。

教育数据在多年的积累当中形成了大量的结构化和非结构化的数据。因为我国推行的教育信息化建设，很多学校都建立教务管理系统，存储了大量的结构化的数据，包括了学生的基本信息、学生的历次成绩、学生的奖惩情况等。当然，也有很多的非结构化数据对于分析学生是非常有用的，比如调查问卷获得的数据、学生出勤率、学生家庭状态等。只有对尽量全面的数据进行充分，才能对学生有着更全面的了解，做出的预测的准确度才能越高[3]。

另外作为数据挖掘技术基础的数据收集工作对于它们未来的使用是非常重要要。所收集的数据需要标准化数据或转换为标准化，以方便对数据进行仔细分析，这意味着要用直观的方法对收集的数据进行分类。目前数据收集的主要来源为高校自身的教务管理系统中的考试成绩及其他相关数据。也可以根据具体需要从学生的试卷或其他文字材料中收集。

总之，利用数据挖掘技术对学生进行全方位的分析，可以为学生提供个性服务，帮助学生形成个性学习方法，同时对于成绩较差有着重修或辍学危险的学生提供预警。对于教师来说，也可以结合分析结果，改进自己的教学方式，帮助学校的管理者拿出更好地管理模式，帮助学生在学习、实践、生活方面提高。

1.2 国内外研究现状

1.2.1 国外研究现状

美国的公共教育开始将数据分析技术应用到教学当中，为更好地推动美国教育部2012 年投入两亿美元开发教育数据计划，以运用数据分析来完善教育教学工作。其中的学习分析系统就是用于理解学生个性化学习模式，它通过数据挖掘、训练数据和功能模块的框架模式运作。这个学习分析系统能够发现很多教师从前没有关注的信息，用于帮助学生能够更好的更多的更准确的学习。比如如果一个学生成绩不好，分析其中的影响因素可能包括周边环境、身体原因、家庭原因等。再比如如果学生考试不及格，是否能够表示他没有完全掌握学习的内容？利用数据挖掘的分析功能，可以向教师提供更多更有用的信息 [5]。

没有教育数据领域有很多大公司的身影，比如 IBM。还有一些新兴的公司专注于此，比如 Civitas Learning，这家公司利用数据挖掘技术中的机器学习方式，对提高学生成绩做出预测分析。公司的立足之本是在教育领域构建的庞大的教育数据库，这些海量数据包括了上百万的学生数据和几百万条相关记录。数据中包括了学生的考试成绩、重修辅

导率、出勤率、辍学率等。通过帮助各大学做数据分析，公司帮助大学学生了解成绩不佳的主要原因, 帮助学生及时调整学习方式或选修课程, 对有辍学风险的学生提出预警。

加拿大的教育科技公司 Desire to Learn 利用数据挖掘技术面向搞笑的学生，他们的业务重点是根据学生的历史成绩数据及学生的社会行为、家庭状况等，来分析和预测学生未来的课程成绩。这家公司产品分析过加拿大和美国上千万学生，利用公司提供的管理系统，也是数据收集系统，与学生在线交流、完成一些测试、监控学生阅读书籍等。通过持续不断的分析，系统不仅仅为学生，也为教师提供了大量信息，帮助教师及时找出学生成绩不佳的影响因素，并做出调整和改进[6]。

1.2.2 国内研究现状

以 2002 年作为标志，我们国内对于教育数据挖掘技术的理解与处理水平，至少与国际先进水准有着 13 年的差距。举例来说，国内一个学生读完 9 年制义务教育产生的可供分析的量化数据基本不会超过 10KB，包括个人与家庭基本信息，学校与教师相关信息，各门各科的考试成绩，身高体重等生理数据，读书馆与体育馆的使用记录，医疗信息与保险信息等。这样的数据量，一台较高配置的普通家庭电脑，初级的 EXCEL 软件就能进行 5000 名以下学生量的统计分析工作。目前随着人们对于“ 数据 ”的理解更为深入了，许多我们曾经并没有重视的，或者缺乏技术与方法去收集的信息，现在都可以作为“ 数据 ”进行记录与分析了[7]。

传统的数据分析技术更关注的是学生群体的表现水平或学业水平、身体素质发展状况、对教学及管理的满意程度等，这些数据的采集来源可以周期性或阶段性获得，不需要即时跟踪。而当前的数据挖掘技术在数据的来源和应用方向上都有所不同。它更注的是学生个体更加细微的表现，以学生课上表现为例，所需要收集的信息有学生翻开书的时间、对哪些话表示出关注、一道题需要多久完成、主动回答问题次数等，这些都是学生个性化表现方式。同时，这些数据贯穿于学生的整个学习和生活，时时刻刻都在发生，充分体现了学生与教师之间、学生与学生之间的关系。可以这样说，对这类数据的挖掘能够更准确的找出学生在学习和生活中发生问题的答案。

目前国内实践方面也在加强探索：

1）对待学生的评估开始全方面关注，深入了解学生成绩背后的原因。比如两个学生的成绩相同，并且多很优异，按照传统方式来说，可以对这样的学生不用给予建议。而进过全面数据评估，将会发现一个学生成绩优异是因为逻辑分析能力较强，那么这样的学生存在较好的发展潜力；另一个学生成绩优异的原因仅仅因为记忆力出色，这样问题是在低年级可以凭借记忆力获得好成绩，而到了高年级，面对需要更多理解分析力的课程时，他将会出现问题。因此需要及时提出辅导，以帮助他及早弥补差距 [8]。

2）增加教师在课堂上收集数据的过程。通过收集学生在课堂的细微表现，比如学生的发言质量、主动回答次数、作业完成情况等，可以在某个阶段汇总数据，形成学生课堂学习的数据流程模型，进而分析学生发展所需要提高的因素，同时也能过促进教师改进教学方式，活跃课堂，或是对学生有针对性的辅导。举例来说，一节课分三个部分，第一第二部分的学习过程学生能够紧跟教师节奏，而在第三部分，学生明显反应时间过长，这就需要对这一部分的教学内容以及教师对于这部分内容的讲解需要改进，帮助学生更好地消化和理解。

3）收集学生校外信息加以数据挖掘，理清学生课外因素对于学习的影响。这些信息的收集可以通过与学生家长的沟通或数据记录，进而掌握学生的课外活动情况，比如课余时间学生所读的书的类型、喜欢玩耍的地方、家庭其他成员与学生之间的互动等。这些信息是有价值的。对这些数据建立数据库并加以分析，可以更加全面的了解学生，并对其课外活动安排为家长提出参考建议。

1.3 本文研究内容

本文设计的成绩分析系统主要是利用数据挖掘算法设计模型来分析教育数据，所要实现的目标有三点：

1）预测新入学学生是否能在第一学期考试通过。如果预测说一个学生倾向于在考试中失败，那么可以建议学生在考试之前采取额外的努力，提高自己成绩并帮助他通过考试。分类方法包括类似决策树、贝叶斯网络可以实现，本文采用的是 CART 和 C4.5，ID3 决策树算法。

2）利用数据挖掘技术选择有重修风险的学生。对于高校学生来说，当考试分数低于重修分数的时候需要在下学期开学初参加重修辅导班，并重新考试。这不仅增加了老师的教学负担，也给了学生带来不必要的负担，在数据挖掘技术的帮助下，我们能够更准确的选择有针对性的学生。在本文中，将采用一个基于数据挖掘的方法选择该类的学生。方法的关键是采用基于得分的关联规则技术，该方法具有非常不错的效果，优于传统方法。

3）通过学生的成绩分析学生之间关系、课程之间关系和影响。在教育中，学生的分数是一个非常重要的定量评价指示器，可以客观地反映教育的影响，是一个重要的科学决策依据。因此，分析和研究学生的分数是非常重要的。学生成绩数据库中的数据呈指数级增长。传统的查询和简单的统计分析无法满足分析的需要，不能为教学捕捉有用的信息。本文利用聚类算法和决策树挖掘学生的分数，通过分析学生-课程-成绩之间的关系可以获得一些教学和管理有价值的信息。

1.4 论文组织结构

第一章绪论。

第二章教育数据挖掘 EDM。本章介绍了教育数据挖掘 EDM 的数据收集来源和类别，数据挖掘的一些关键技术：K 聚类、层次聚类、光谱聚类、关联规则等。同时以 EDM 在国外实践的几个案例说明教育数据挖掘的有效性、实用性。

第三章系统需求分析。本章从功能需求、数据需求和技术需求几方面分析了实现成绩分析系统所需要做的准备工作。

第四章系统设计与实现。本章介绍了三个功能的设计与实现：C4.5 决策树算法根据学生过去几年积累的数据可以有效的预测新生的第一学期考试成绩表现；利用改进的关联规则挖掘技术有效预测有重修风险的学生；使用聚类算法和决策树来全面分析学生的得分，进而得出学生之间的关系、学生与课程之间的关系和课程之间的影响。

第五章系统测试。本章利用 SAS Enterprise Miner 平台对本文设计的数据挖掘模型进行测试，包括输入数据源、数据分区点检查、变量缺失值替代、交互分组、决策树模型、聚类分析模型等。

第六章总结。

第 2 章 教育数据挖掘 EDM

教育数据挖掘 EDM 是一个新兴领域, 是为备满足日益增长的教育普遍评价的需要。EDM 侧重于收集、归档和分析学生学习相关的数据并进行评估。EDM 是一个非常新的和非常小的学术领域。如同所有的新领域，EDM 与现有的学科有很多新的重叠。许多人都是从 EDM 研究智能教学系统（ITS），可以获得大量的教育数据。EEDM 的分析研究往往是在心理与教育统计学相关技术。相比几十年教育运用统计方法，EDM 利用数据挖掘和机器学习的研究结果将彻底改变过往研究。数据挖掘通常利用巨大的数据集和关联的研究，采用高效的算法寻找意义的数据。EDM 项目有大量的数据，包括数以千计数据集和成千上万的记录，这只是一样普遍使用的数据集数十或数百人的记录。通用机器学习研究，特别是无监督或半监督学习，更直接影响 EDM[9]。

大多数 EDM（教育数据挖掘）项目的结构可以分为三个部分：收集、归档和分析。集合是指用于记录的工具和辅导系统相关信息，学生在线测验成绩，或一个智能辅导系统事件。归档是存储和浏览收集数据的过程。对于分数数据, 这是一个相对较小的问题, 但对于大量生成的数据呈现的可能是一项重要的任务。分析工作采用机器学习和数据挖掘工具, 试图对收集到的数据获得关于学生的学习更深了解, 以及发现之间的关系问题, 并由此定量的理解认知过程并深化发展。这三个任务具有一定的复杂性和重要性，同时这三个任务可以解决任何 EDM（教育数据挖掘）项目。

2.1 教育统计数据

教育数据研究包括许多不同的领域，包括从认知科学的统计数据。在 EDM（教育数据挖掘）比较接近的是心理和认知心理学。

2.1.1 心理测验学

定量分支是最适用于 EDM 的心理指标。心理测验学可以追溯到至少 19 世纪, Francis Galton 是第一个关注潜在的人类思维和能力相关知识的人。这些都是很难量化的值，没有直接的方法来衡量它们，没有隐式的单位或维度等测量。

心理测验学有几种主要理论影响了当前的 EDM 研究。最有影响力的是经典测试理论(CTT)和项目反应理论(IRT)。

一、经典测试理论

经典测试理论的早期心理测验学适用于 EDM。二十世纪，心理学家和数学家斯皮尔曼寻求智力支持他的理论。他指出，在许多情况下学生个人成绩测试问题是高度相关，并提议的一个最早的智力理论，称为 G 理论，学生智慧被视为一个维特征。这一心理学教育理论在一些地区今天仍在使用。然而，斯皮尔曼维认为一些数据根本不匹配该模型。转向数学，利用当时的技术计算测量值之间的相关性，有了共同的因素分析来理解大量的数据，发现数据的潜在因素。CFA 假设 f 因素的线性模型可以描述底层系统的行为，学生正确回答问题的能力被描述为：

$$M_{i,j} = \mu_j + \sum_{k=1}^{f} W_{i,k} H_{k,j} + \varepsilon$$

其中 $M_{i,j}$ 是学生 i 在项目 j 所得的分数，μ_j 是一个隐藏变量用来描述项目的难度，W 是一个矩阵，描述每个学生对于每个因素的能力，H 矩阵描述每个项目的每个因素，ε 是分数测量的噪声数据。由此产生的因素是一组向量生成一些向量输入矩阵的子空间[10]。

CFA 是第一个心理测验学定量理论，包括测试错误的概念，认为一个学生的分数在考试并不是绝对的知识测量，而是一个随机的知识状态。这概念已经催生了一系列心理测验学统计估计方法。是这些 CFA 和相关的方法统称为经典测试理论。

用于 EDM 的重要概念是可靠性、内部一致性和有效性。可靠性是衡量产生的分数。一个常见的方法，讨论可靠性是两次试验法。可靠性无法准确测量，实践中有许多近似的技术工作，包括替代形式的测试与稍有差异的问题，或更常见的 split-half 相关性。

内部一致性度量的程度在问题测量相同的特征。内部一致性是高度相关的可靠性。Split-half 关联，这主要是视为一个可靠性统计，是一种减少的量表的阿尔法统计内部一致性的公约数[11]。

有效性是一个更加困难的概念，一般采用一个估计成绩实际测量工具。这往往依赖外部比较和验证，通常是难以量化。

二、项目反应理论(IRT)

项目反应理论 IRT 需要学生在一个单一的课程表现，因此没有直接适用评估项目。项目反应理论 IRT 为项目建模和开发一个简单的估计丢失的得分数据将会有一定的帮助。项目反应理论 IRT 参数化分布给定学生能力的概率水平 θ，通常这些特性曲线表示两个参数逻辑函数：

$$P(\theta) = \frac{1}{1 + e^{\alpha(\theta - \beta)}}$$

这里 β 表示项目的难度，α 描述项目信息，描述一个能力低于 β 的学生能回答正确的可能性有多大，以及一个能力高于 β 的学生能回答错误的可能性有多大[12]。

项目反应理论 IRT 是强大的, 因为它提供了一个关于项目难度和学生能力的定量尺度问题。由于项目反应理论 IRT 量化困难，能力在一个绝对标度下，它适用于无关困难的问题。高级新手学生将导致低的问题平均分数，但还会跟踪部分特性曲线上升。回答同一个问题的能力强的学生将位于曲线较高部分，根据项目反应理论 IRT，提取参数曲线在这两种情况下是相同的[13]。

2.1.2 认知理论

认知科学家研究注意、知觉、记忆、语言等学习上的细节问题。心理学作为一个实验科学最早出现在 1870 年代, 认知科学与实验心理学是作为一个整体的, 在 1950 年代, 认知心理学进一步专业化。认知心理学最适用于 EDM 的是那些重叠的学习和教育。这些主要是理论思维的组织和检索信息。EDM 与认知理论相关的重点在于符号系统和理论。一般认为，一组基于离散符号的逻辑操作的规则集和符号可能创建一个类似人工智能。大多数现代的研究主要集中在概率推理,

认知理论知识表示在不同的形式：程序性技能，声明事实和琐事，以及一个概念类似于计算机处理器寄存器标志性记忆，代表那些碎片的信息。主要功能作为框架，代表着知识和它的结构。

即使没有定量的适用性，这些结果对教育学习也是至关重要的过程，但在更广泛的社区教育仍然遗憾的是未知的。SOAP 是另一种常见的认知体系结构，类似于行为的架构，对于不同类型的知识它有不同的表示方法[14]。

下面的公式提供了一种定量关系管理的认知任务的关系：

$$t = X\alpha^{n}$$

这是其最常见的形式，规定反应时间在一个特定的任务指数试验的数量，这意味着实践提高了性能。经过多次调查，这是普遍在测试对象和附近发现任务。进一步研究表明，它实际上是一个人类学习认知架构结果(实际)，比如应用技巧，认识到当应用技巧或检索一个特定的事实表现出同样的关系。通过提高准确性可以构造一个简单的数学模型来量化这一结果。

2.1.3 Q 矩阵

Q 矩阵理论捕获底层的是一个方法因素和能力，产生一个学生成绩矩阵。在最简单的情况下，Q 矩阵应用于二进制 m×n 评分矩阵。Q 矩阵操作这个矩阵以及假设的潜在

能力或因素 t。每个问题是假设有一个二进制与每个 t 的关系因素。每个学生同样假定有一个二进制能力。因此 Q 矩阵将输入矩阵分解为一个二进制 m×t 矩阵的学生的能力这些因素, 以及一个二进制 t×n 矩阵的相关性问题。一个学生假设正确解答一个问题当且仅当与这个问题相关的因素是自己能力的一个子集[15]。

大多数 Q 矩阵利用梯度下降的重建误差矩阵来找到最优模型以适应数据。给定一个初始随机排列，该算法找到单个学生进入切换矩阵或问题矩阵以复制输入矩阵。作为一个离散优化问题与搜索空间增长的 O(2t*(m + n))，这是一个计算密集型的解决方案。一个线性规划公式可以减少运行时间，提高解决方案质量。

2.2 机器学习和数据挖掘技术

2.2.1 聚类

聚类关注对象组织在某些方面是相似的，相似的性质取决于所使用的聚类算法，进而反映了底层模型。聚类是一种无监督学习应用程序，不需要训练数据校准问题。聚类本质上是一个复杂的主题，没有绝对的正确性。如果一组人给出一个散点图，并被要求聚类的数据点，存在着不确定性的协议。给定一组点与已知的标签，可以调查数据集准确的聚类算法。一致性是要求移动点在同一聚类，移动聚类应该与重新聚类时有相同的输出。这些限制制定必要的权衡在所有的聚类算法[16]。

一、k-means

k-means 算法(MacQueen，1967)是最简单常见的聚类算法，对于各种类型问题表现相当不错。k-means 算法假设有 k 聚类中心，点不一定是输入定义的一部分，每个点做 k 聚类。如果聚类中心是已知的，聚类很简单：输入数据中的每个点被分配给最近的中心一些指标。这一指标通常是欧几里得距离，高维可以使用其他一些指标数据。

为了从数据中心确定聚类的位置，k-means 进行简单迭代：从输入作为聚类中心随机选择 k 点, 然后分配在到最近的聚类中, 每个点与新的聚类中心计算以发现每个聚类。这一过程持续进行直到没有点改变作业，或直到中心只是一个微不足道的数量。这并不一定是最优的聚类中心，所以随机重新选择是常见的。

该 k-均值算法有一些缺点。最常被引用的其中之一是确定 k，这通常必须事先了解数据；指定 K 后分区的簇的数量问题；每个维度没有预先确定的规模，这相当于每个维度的一个额外的参数，降低了该算法通用性和创新性[17]。

二、光谱聚类

光谱聚类(费舍尔和波兰，2005)通常使用线性代数实现，可以解释为一个相对简单的技术。输入数据 $x_1, x_2, ..., x_n$，形成完整的、加权和完全连接图，边缘点之间 i 和 j 给出权重相等的相似点。一个光谱聚类到 k 聚类是通过找到最低权重削减图。

在实践中这通常是通过生成 n×n 相似矩阵 S，$S_{i,j}$ 是 x_i, x_j 相似性度量，通常为欧几里得距离。K 从这个矩阵中提取的特征向量与最大特征值有关[18]。

三、层次聚类

为了克服需要一个参数代表聚类和允许的数量丰富的条件下，一些算法产生分层聚类。在这个方案中，聚类算法的输出不是一个数据点分配给聚类，而是相对相似的系统树图代表每一个输入。

编码每个聚类的数量，每一项单独聚类项目集中在一起。通过搜索系统树图的水平聚类数量得到数量相同的子树图，很容易发现这些聚类系统树图。一个示例系统树图代表的相似性如图 2.1 所示。

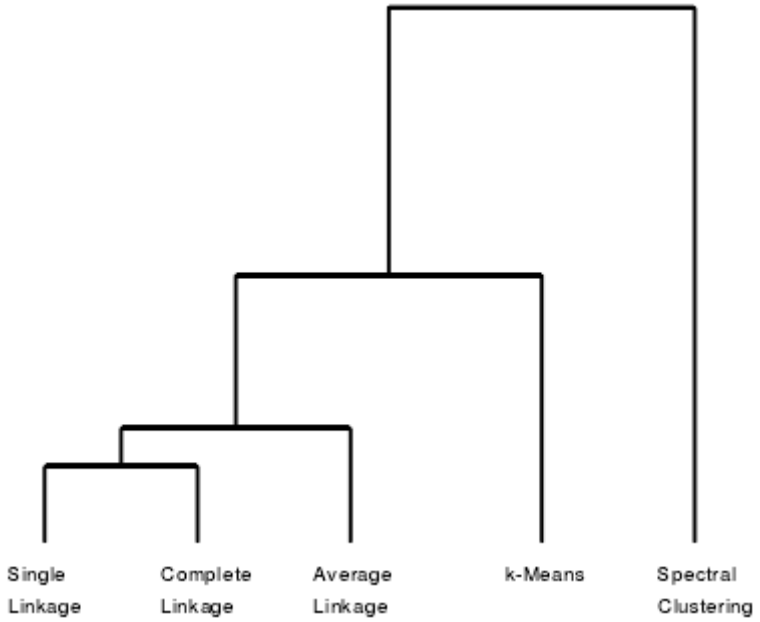

图 2.1　系统树图的相似性聚类算法

层次聚类算法通常会凝聚的，也就是说开始所有项目分配给自己的聚类，然后附近的集群是按照一定的距离度量的合并。在分层或凝聚聚类存在三种常用的技术：Single Linkage、Complete Linkage、Average Linkage[19]。

在 Single Linkage，所有项目最初分配给自己的聚类。在每个步骤中，最小的两个聚类之间的最小距离不同聚类标识和有关的物品。发现两个点不是在同一个聚类，计算之间的最小距离后合并这两个点的聚类。在构建系统树图时，不同的仅仅是这两个点之间的距离，导致附近(strongly-clustered)指向显示相似度高，而遥远的(weakly-clustered)点显示低相似性。

Complete Linkage 采用相反的方法联系。按照最小最大点距离合并两簇。每对点必须紧密结合在一起以便联系。这创造了更多的严格的分组集群，这往往造成集群形如长链。

Average Linkage 使用聚类的欧几里得距离划分，发现最小的聚类中心之间的距离。聚类属性算法依赖于界定的系统树图用于确定聚类。如果聚类停止，没有聚类在一些固定的距离，这些算法因尺度不变性失败。如果这个距离阈值作为一个函数的两个点之间的最大距离输入，可以违反一致性[20]。

2.2.2 降维

目前 EDM（教育数据挖掘）研究中的另一类算法使用数字降维算法。这些可以在许多方面适用有损压缩，解释方差的输入数据，找到基础因素生成数据。

一般公式为 X=UH

1）奇异值分解

奇异值分解是输入数据维数降低最好的方法。由此产生的因素是输入的子空间的正交基向量空间。确定因素按照重要性通过寻找的方向最高的方差的输入数据，直到所需的许多因素已经被提取。理论上元素 U 和 H 范围（-∞，+∞）。奇异值分解通常被认为是最好的工具用于降维，尤其是当不了解底层数据生成模型时。

2）独立分量分析

独立分量分析类似于奇异值分解，它确定了一组跨越的子空间的输入向量空间向量。独立分量分析不是局限于识别正交向量，独立分量分析是特别适合的情况是非正交的输入的数据是一个线性组合。

独立分量分析在教育领域一些重要的含义。如果底层因素解释为“ 主题 ”，然后 X 在每个话题分解成 U 和 H，意味着 U 表示学生能力，H 是每个主题相关性问题。

通过正交不需要主题，主题可以发现有一些重叠。例如，不认为能力与链表是完全正交的能力与二进制搜索树。两者都是递归的离散结构。

3）非负矩阵分解

另一种降维的方法是非负矩阵分解。它不限制所得到的分解的几何性质，非负矩阵分解限制数值属性, U 和 H 限制非负, 这在教育领域比较直观, 因为没有负相关的话题, , 学生没有消极的能力。非负矩阵分解通过梯度解决与随机重新启动。它是一个连续不断的优化问题，最优方向来更新 U 和 H，更新达到一个最小值[21]。

2.2.3 关联规则

数据挖掘技术在商业领域最伟大的应用是发现关联规则，也被称为市场购物篮分析，关联规则发现最好的理解通过无处不在的超市。

假设每一个注册超市收据存储在一个中央数据库。信息存储在数据库，列出所有的项目一起购买(因此”市场篮子”)。给定一个大型数据库的事务日志，提出以下问题：哪些项目可以一起卖? 对非显而易见的项目配对，最常见的出现在交易让店主整理项目知识最大化的冲动性购买行为。典型的例子是啤酒与尿布：在一个假设的交易数据库中发现啤酒和尿布有一起购买倾向。它使营销人员意识到利用放置一个啤酒并显示附近的尿布，可以增加购买啤酒的顾客同时选购尿布[22]。

2.3 EDM 在国外高等教育实践

数据挖掘是一种强大的工具，通过数据挖掘，一所大学可以预测 85%的准确率，学生会或不会毕业。大学可以使用这些信息来集中注意力帮助这些学生摆脱危险。数据挖掘依靠几种基本方法：分类、评估、可视化。分类识别关系和集群，并研究对象的分离。使用规则归纳算法处理分类结果，如“存在”或“辍学”，和“转移”或“留下来。估计包括预测函数或可能性和处理连续的结果变量。可视化使用交互式图形展示诱导规则和分数。可视化可用于描述数学的三维地理位置坐标。例如，高等教育机构可以使用分类综合分析学生特点，或者使用估计预测各种结果的可能性，如可转移性、持久性、保留和课程的成功。

分类和评估使用无监督或监管建模技术。无监督数据挖掘用于未知情况下，特定的组织或模式。例如，在学生课程数据库，鲜为人知的课程通常被作为一个群体，或相关联的课程类型的学生类型。无监督数据挖掘是常用的第一个研究模式和搜索以前隐藏的模式[23]。

监督数据挖掘是使用一个已知的结果的记录。例如，一个毕业的数据库包含完成学业的学生，以及那些辍学记录。监督数据挖掘用于研究两组的学术行为，意图的行为模式与学术的历史和其他记录信息。

所谓的“机器学习”使用人工智能感应规则并划分模式，分析师可以适用于新数据。一旦一个模型表现良好，分析师可以给另一个学生团体，如新生，应用学到的信息到新组毕业预测的可能性。所有的这些步骤自动化形成迅速准确的估计，相比传统的节约时间和资源。

以下三个案例研究说明高等教育数据挖掘的关键作用。

一、实践研究：创造有意义的学习结果

问题：“ 学校知道学生什么?这个案例研究演示了 EDM 使用无监督数据挖掘可以建立学习成果类型学为学生使用。

一个典型的高等院校招生 4000 人, 传统上会将学生划分为 “ 职业教育指导 ”或“ 基本技能升级 ” 等类别。然而，这些分类是基于学生的初始入学教育目标。而这些都是包容的分类，不帮助说明每个学生类型之间的区别[24]。

解决方案：为 4000 名学生建立适当的类型，研究人员使用 Two Step 和 k - means 两个强大聚类算法。首先应用上述算法确定通用分组，结果喜忧参半。聚类之间的边界是不清楚和分散，甚至在抵抗反复测试后数据集，以及疑似异常值情况下似乎不属于任何分类，结果并没有显著提高。采用一种替代方法，看着教育成果结合长度的研究。定义教育成果说起来容易做起来难。测试数据挖掘的领域知识，没有绝对正确或错误的类型。删除离群值或将它们添加到一个特定的聚类, Two Step 算法产生以下聚类 :“ 转移 ”、“ 职业学生 ”、“ 学生基本技能 ”，“ 学生混合的结果 ” 和 “ 辍学。k - means 验证这些聚类。引入年限元素给每个聚类的新维度，转移一些完成学业很快学生。

结果：数据挖掘结合学生人口统计数据和其他信息，使大学能够提高其理解学生类型。某些年轻的学生花费更多时间和精力在学分较高毕业较快的课程上。经过重新聚类后，学院开始把学生划分出更多的类，描述学生快速积累学分的类别，花了相当的时间长度毕业的类别等。类型学很重要，因为它们超越传统分析识别同质群体的学生，从而增加预测建模算法的准确性。即使数据挖掘项目结束时发现的适当的类型学，新发现的模式和关系帮助教育者和管理者更好的满足不同的学生团体的需求[25]。

二、学术计划和干预转学预测

问题：这个案例研究展示高等教育解决一个棘手的问题：如何准确地预测学术成果以便及时学术干预。当机构使用数据挖掘预测哪些学生是最危险，机构可以防止学生失败，甚至在学生意识到之前。

超过一半的社区学院学生以四年制大学作为他们的目标。由于学术困难，许多人需要很长时间才能决定转学或不转学。虽然发现学生转学是困难的，但是考虑到允许社区学院和大学匹配他们的数据，这意味着数据挖掘研究者可以链接社区学院的学术行为到他们的转学结果。

解决方案：使用转学的结果数据，分析建立了一个数据集，其中包含学生的一般性转学聚类。数据集分成测试数据集和验证数据集，使用专有的随机化方法。结果变量被转移。其他变量，如人口、课程单元累积，金融援助，要分析预测没有逐步测试的意义。数据挖掘是非常宽容的变量在数据交互和非线性关系。监督数据挖掘是适当的方法，因此，分析师同时利用神经网络和规则归纳算法，也是为了对比和比较预测精度。

结果：数据挖掘使大学能够准确识别好的转学的候选人。广泛的机器学习后，神经网络算法，神经网络，预测精度为 72%，规则归纳算法，C5.0 C&RT，有 80%的预测精度。模型对测试数据集和产生了类似的结果，表明他们了解了数据模式[26]。

三、预测校友的承诺

问题：对于一个典型的超过 20000 人的院校，大学毕业生人数可以十倍其入学人数。大多数大学定期发送邮件向校友们，即使校友无法回应。这些邮件通常每年要花费 100000 美元。本案例研究表明，数据挖掘可以帮助大学集中在最有可能做出承诺的校友。

解决方案：通常很难确定邮件是否直接影响毕业生的数量和价值承诺。考虑到相同的类型的邮件，一个校友可能定期贡献而另一个不是。增加了存在离群值的混乱程度，如校友意外大量贡献。

在图 2.2 中，图表显示使用数据挖掘来确定校友邮件接收者和直接邮寄到所有校友。曲线是最佳回报率(校友贡献)预测的数据挖掘。45 度线是如果收到了邮件整个预测结果。在这种情况下表明，当邮件到达 30%的预测的数据挖掘是响应，80%的人会回应与承诺。如果整个人数收到邮件，只有 40%的人会有什么样的反应。如果每一个基点= 2500 美元，储蓄=(70% * 2500 美元)-(30% * 2500 美元)= 175000 - 75000 = 100000 美元。没有数据挖掘，它将花费 100000 美元到达 80%校友。

下面的图 2 显示了数据挖掘可能带来的校友承诺好处。

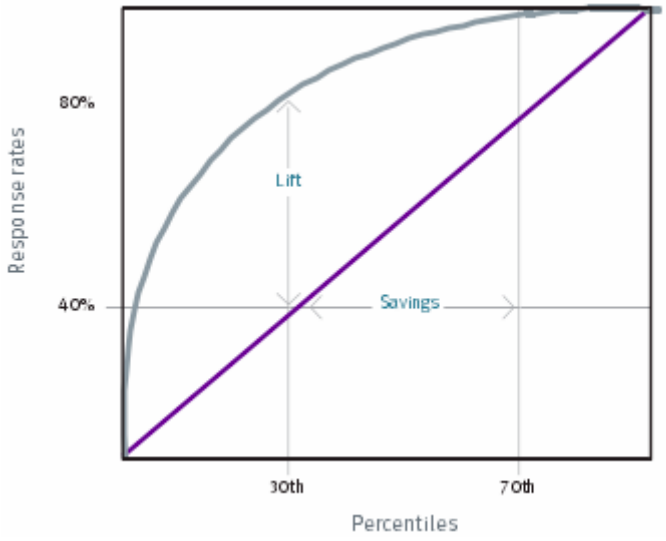

图 2.2　数据挖掘增益图

结果：使用数据挖掘，学院发现了一个方法，使其邮件更有效，提高校友承诺，减少邮寄费用。这是最好的描述使用一个概念叫做“ 升力。“ 如果 20%的校友回应学院的承诺要求，大学应该把精力集中在那些 20%。如果数据挖掘可以快速识别潜在的捐赠者的比率 2 - 4，然后大学可以通过邮件取得成果达到校友人数的 40%，从而节约大量时间和金钱。

数据挖掘是一个功能强大的分析工具，使教育机构更好地分配资源和员工，主动管理学生的成果，提高学校发展的有效性。能够发现隐藏的模式在大型数据库，社区学院和大学可以建立模型，预测的精度很高。通过作用于这些预测模型，教育机构可以有效地解决问题。

2.4 本章小结

本章介绍了教育数据挖掘 EDM 的数据收集来源和类别, 数据挖掘的一些关键技术：K 聚类、层次聚类、光谱聚类、关联规则等。同时以 EDM 在国外实践的几个案例说明教育数据挖掘的有效性、实用性。

第 3 章 系统需求分析

3.1 功能需求

教育数据挖掘是一种新兴的技术，相关的教育领域数据可以应用于数据挖掘。教育数据挖掘使用了很多技术，如决策树、神经网络、聚类分析、关联规则、贝叶斯等。使用这些技术可以产生各种各样的知识发现。发现的知识可以用于对学生执行特定的预测、改变传统课堂教学模式、利用检测异常值的结果发现学生的变化等等。通过分析学生在学习、生活中扮演的角色，来为学生有可能的改变及早提出建议和帮助。

本文成绩分析系统所需要实现的功能有以下几点：

1）预测新入学学生是否能在新学年第一学期考试通过。如果预测说一个学生倾向于在考试中失败，那么可以建议学生在考试之前采取额外的努力，提高自己成绩并帮助他通过考试。

为实现这个目标，数据挖掘需要做到如下几方面：

a)收集预测变量的数据来源,

b)识别影响学生学习行为和职业生涯的不同的因素。

c)使用数据挖掘分类技术构建预测模型。

d)验证工程开发模型。

2）利用数据挖掘技术选择有重修风险的学生。对于高校学生来说，当考试分数低于重修分数的时候需要在下学期开学初参加重修辅导班，并重新考试。这不仅增加了老师的教学负担，也给了学生带来不必要的负担，在数据挖掘技术的帮助下，我们能够更准确的选择有针对性的学生。在本文中，将采用一个基于数据挖掘的方法选择该类的学生。方法的关键是采用基于得分的关联规则技术，该方法具有非常不错的效果，优于传统方法。

3）通过学生的成绩分析学生之间关系、课程之间关系和影响。在教育中，学生的分数是一个非常重要的定量评价指示器，可以客观地反映教育的影响，是一个重要的科学决策依据。因此，分析和研究学生的分数是非常重要的。学生成绩数据库中的数据呈指数级增长。传统的查询和简单的统计分析无法满足分析的需要，不能为教学捕捉有用的信息。本文利用聚类算法和决策树挖掘学生的分数，通过分析学生-课程-成绩之间的关系可以获得一些教学和管理有价值的信息，如学生的关系，课程之间的关系和课程之间的影响。

3.2 数据需求

数据往往来自不同来源的信息，并结合这些不同的来源。一个典型的数据集有数千的记录。记录可能代表如个别客户实体，一个特定的交易，或某一个家庭。数据集的变量包含特定信息，如人口信息，历史分数，或每个课程信息。这些信息是如何使用取决于感兴趣的研究问题。当谈论不同类型的数据，考虑每个变量的测量水平。

1）区间变量的均值（或平均）是有意义的，如平均成绩。

2）分类：由一组水平的变量，比如性别（男性或女）或成绩划分（优异，正常，较差）。一般来说，如果变量不连续（即平均没有意义，如平均性别），分类数据可以按多种方法分类。

根据挖掘的目的考虑这些子组的分类变量：

1）对于数据集来说一个变量有相同的观测值。

2）二进制变量只有两个可能的值。

3）名义变量有两个以上的值，但是值的水平没有隐含的秩序。

4）序数变量有两个以上的值，值有一个隐含的顺序。

注：序数变量可以作为名义变量，然而，名义变量不能作为序数变量，因为没有隐含的排序。缺失值不包括在计数。为了获得一个有意义的分析，必须建立相应的数据集为每个变量指定正确的值。

3.3 技术需求

预测建模技术，能够确定是否一组输入变量在预测一些结果变量是有用的。例如，高校管理机构可以尝试确定学生的家庭情况和高中成绩（输入变量）有助于预测学生是否能够通过新学年第一学期的考试（结果变量）。

区分输入变量与结果变量，需要为数据集的每个变量确定角色。通过使用目标模型识别结果变量的作用，利用输入模式识别输入变量的作用。模型角色的例子包括性别，姓名，家庭住址和家庭状况。如果想排除一些变量，需要利用模型确定这些变量的作用。

预测建模技术，需要一个或多个结果变量。预测的结果根据每个技术的标准不同而不同，如准确性最大化或利润最大化。本文使用多个预测建模技术，包括通过关联规则，决策树，聚类分析。这些技术能够预测二进制，名义的，顺序，或连续结果变量输入变量的任何组合。能够识别在一个数据集潜在的模式。

1）聚类分析：这种分析基于一组输入变量试图找到数据的自然分组。分组后，观察集群，使用输入变量来描述每一个组。当集群已被确定和解释，可以决定是否每个集群独立。

2）关联分析：分析确定产品或服务的组合，对于本文的成绩分析系统就是要确定课程和成绩之间的关系、课程与课程之间的关系等。

3.4 本章小结

本章从功能需求、数据需求和技术需求几方面分析了实现成绩分析系统所需要做的准备工作。

第 4 章 系统设计与实现

4.1 预测高校新生第一学期考试表现

教育数据库中存储的数据量迅速增加。这些数据库包含隐藏的信息可以提高学生的表现。教育数据挖掘是用于研究教育领域的数据挖掘隐藏的知识。分类方法包括类似决策树、贝叶斯网络等应用于教育数据，预测学生在考试的表现。这个预测将有助于识别较弱的学生，帮助他们取得更好的成绩。CART 和 C4.5，ID3 决策树算法应用工科新生的数据来预测他们在第一学期期末考试的表现。决策树预测的结果可以帮助一定数量的学生提前预警并很可能通过考试或在下学期提升成绩。在期末考试结果出来后获得标记的学生被送入系统结果分析下一个会话。比较分析的结果预测较弱的学生需要提高和拿出改善的结果。

4.1.1 决策树

决策树是一种类似流程图的树结构，每个内部节点用矩形表示，叶节点用椭圆表示。所有内部节点有两个或两个以上的子节点。所有内部节点包含多个分裂，用于测试一个表达式的属性的价值。从内部节点到子节点用不同的测试结果标记。每一个叶子节点都有一个与之关联的类标签。

决策树通常用于获得信息，决策树开始根节点是为用户采取有目标的行动。从这个节点，用户将递归地根据每个节点决定树学习算法。最终结果是一个决策树每个分支代表一个可能的场景的决定和它的结果。

这三个常用的决策树学习算法是：ID3、C4.5 和 CART。

1）ID3 算法

ID3 迭代二分是 1986 年昆兰罗斯引入决策树算法，它是基于搜索算法。包括两个阶段：构建树和树修剪。ID3 使用信息增益测量选择分裂属性。它只接受分类属性建立一个树模型。当数据存在噪声时它不能给出准确的结果，这需要数据预处理技术来去除噪声数据。

构建决策树，计算信息增益每一个属性并选择最高的属性信息增益指定作为根节点。标签属性作为一个根节点，属性的值表示为弧。对所有可能结果实例测试，以检查是否在相同类下降与否。如果所有的实例是在同一个类下降，节点代表单一类名称，否则选择分裂属性分类的实例。

可以使用 ID3 利用离散化或直接算法处理连续属性，通过考虑值寻找最佳分裂点，这需要设置一个阈值的属性值。ID3 不支持修剪。

2）C4.5 算法

C4.5 算法处理分类和连续属性来构建一个决定树。为了处理连续属性，C4.5 分裂基于所选择的属性值为两个分区等所有高于阈值的值。它还处理缺失属性值。C4.5 使用增益比作为一个属性选择测量构建决策树。当有许多结果一个属性的值，它消除了偏见信息增益。首先，计算每个属性的获得率。根节点的属性获得比是最大的。C4.5 使用悲观修剪去除不必要的决策树分支来提高分类的准确性。

3）CART

CART 代表分类和回归树，CART 处理分类和连续属性构建决策树。它处理缺失值。CART 使用基尼系数作为属性选择度量构建决策树。与 ID3 和 C4.5 算法不同，CART 产生二元分裂。因此，它生成二叉树。基尼指数衡量不使用概率假设。CART 使用成本修剪的复杂性改善不可靠的分支准确性。

4.1.2 数据准备

预测一个学生需要大量的学习相关参数需要考虑。包括所有的预测模型个人、社会、心理和其他环境变量的有效预测，这些能够反映学生们的表现。与学生有关的背景数据、擅长某类能力，完成考试的能力时间等也将在预测性能扮演一个角色。

数据准备工作研究中使用的数据集是来自于学校数据库、权威机构所做的相关数据调查和一些研究工作，另外也包括调查问卷所获得的信息。

在下面的数据表中只选择了数据挖掘所需要的字段，收集到的数据报名表由学生入学的时候填写。包括他们的人口统计数据(类别、性别等)、性能数据(历次考试成绩、入学成绩、校内社团表现等等)，地址和联系电话。大部分的属性显示学生的过去的表现。其中的一些信息从数据库中提取的变量。所有的预测和响应变量的数据库在下表中给出参考。

表 4.1 学生相关的变量（变量值用汉字说明）

变量	变量描述	变量值
branch	学生课程分支	CS 、IT、 ME
sex	学生性别	男、女
HSG	学生高中教育表现，以六科成绩计算。	O:90~100 A:80~89 B:70~79 C:60~69

		D:50~59 E:40~49 F:<40
SSG	学生初中教育表现，以五科成绩计算。	同上
Atype	入学类型	正常高考、保送、体育加分、少数民族加分等
lan	语言类别	英语、日语、俄语、韩语
Loc	居住位置	城市、乡镇、农村、山里
hos	是否住宿	Y、N
Fsize	家庭人口数量	1、2、3、>3
Fstat	家庭状态	正常、离异、单亲、孤儿
FAin	家庭收入	贫困、低收入、正常、高收入
Fqual	父亲教育程度	无、小学、初中、高中、专科、本科、研究生、博士
Mqual	母亲教育程度	无、小学、初中、高中、专科、本科、研究生、博士
Focc	父亲职业	服务业、商业、机关、其他
Mocc	母亲职业	服务业、商业、机关、其他
Result	新学期考试结果	优异、通过、失败

域值的一些变量被定义如下：

Branch：课程分支，计算机科学工程类 CS、信息技术类 IT、机械工程类 ME。

HSG：取学生在高中期间六科综合成绩，语文、数学、物理、化学、政治、生物。

Atype：确定学生入学类型。除了已经列出的之外，还应包括军警烈士子女等其他情况。

Result：通过表示全部学科考试及格，优异为获得奖学金学生，失败为有至少一科没有及格的学生。

4.1.3 决策树算法模型

WEKA 是开源软件，实现了一个收集并广泛应用于机器学习算法数据挖掘应用程序。根据上面的数据创建 engg.arff 文件。这个文件被载入 WEKA explorer。分类面板允

许用户分类和应用回归算法生成的数据集，以估计生成的预测模型的准确性，并可视化错误的预测或模型本身。在 WEKA 有 16 个如 ID3 决策树算法、J48 ADT 等可实现。我们选择用于分类的算法是 ID3、C4.5 和 CART。在“ 测试选项 ”下，选择 10 倍交叉验证作为我们的评估方法。因为没有单独的评估数据集，生成模型的准确性这是必要的。模型生成的决策树，这些预测模型提供预测新方法用于评测学生。

三个决策树作为预测模型，从学生获得由三个机器学习的数据集算法：ID3 决策树算法、C4.5 决策树算法和 CART 算法。

图 4.1、4.2 和 4.3 显示了生成的规则的 ID3、C4.5 和 CART。

```
SSG = A
| Focc = Service: Pass
| Focc = Business
| | Sex = M: Promoted
| | Sex = F: Pass
| Focc = NA: null
| Focc = Retaired: null
| Focc = Agri: Pass
SSG = B
| HSG = A
| | Med = English: Pass
| | Med = Hindi: Promoted
| HSG = O: Pass
| HSG = B
| | FAIn = Medium
| | | Atype = UPSEE: Pass
| | | Atype = Direct: Promoted
| | FAIn = High: Fail
| | FAIn = Poor: Fail
| | FAIn = BPL: null
| HSG = C: Promoted
| HSG = E: null
| HSG = D: Promoted
| HSG = F: null
SSG = C
| Fqual = PG
| | Branch = CSE: Pass
| | Branch = IT: Promoted
| | Branch = ME: null
| Fqual = UG
| | HSG = A
| | | Branch = CSE: Promoted
| | | Branch = IT: null
| | | Branch = ME: Pass
| | HSG = O: null
| | HSG = B: Pass
| | HSG = C
| | | Sex = M: Pass
| | | Sex = F: Promoted
| | HSG = E: Promoted
| | HSG = D: Promoted
| | HSG = F: null
| Fqual = Ph.D.: Promoted
| Fqual = Secondary: Fail
| Fqual = Elementary
| | Branch = CSE: null
| | Branch = IT: Pass
| | Branch = ME: Promoted
```

图 4.1　ID3 算法

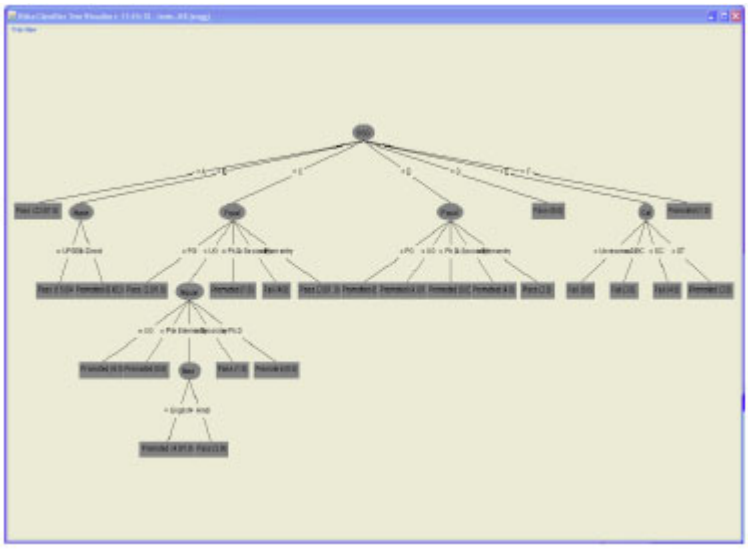

图 4.2　C4.5 决策树算法

SSG=(O)|(A): Pass(26.0/1.0)
SSG!=(O)|(A)
| Focc=(Service)|(Business)
| | FAIn=(Poor)|(High)
| | | Lloc=(Village)|(Town)|(Tahseel):
Promoted(11.0/0.0)
| | | Lloc!=(Village)|(Town)|(Tahseel):
Fail(4.0/1.0)
| | FAIn!=(Poor)|(High)
| | | Mocc=(Service)
| | | | Branch=(ME): Pass(2.0/0.0)
| | | | Branch!=(ME): Promoted(7.0/0.0)
| | | Mocc!=(Service)
| | | | HSG=(O)|(B)|(A)|(E)|(F):
Pass(15.0/1.0)
| | | | HSG!=(O)|(B)|(A)|(E)|(F)
| | | | | Sex=(F): Promoted(2.0/0.0)
| | | | | Sex!=(F): Pass(2.0/1.0)
| Focc!=(Service)|(Business)
| | HSG=(A): Promoted(3.0/0.0)
| | HSG!=(A)
| | | Sex=(F): Promoted(2.0/1.0)
| | | Sex!=(F): Fail(9.0/2.0)

图 4.3　CART 算法

4.1.4　决策树分析

ID3 算法以信息增益作为决策树分类属性标准，计算公式为：信息增益=期望信息值-信息熵，也可以表示为原有的信息需求与信息的信息需求之间的差。

$$Gain(A) = Info(D) - Info_A(D)$$

原有的信息需求基于类比例，新的信息需求基于对属性 A 的重新划分。一般选择最高的增益属性 A 作为决策树结点的分裂属性。

按照表 4.1 的数据变量，选择计算出“ 性别 ”、“ 课程分枝 ”、“ HSG ”、“ SSG ”、“ 居住位置 ”、“ 母亲教育程度 ”、“ 考试结果 ” 的信息增益值，如下表 4.2 所示：

表 4.2　各属性期望信息值和信息增益值

属性	期望 info	增益 Gain
性别	0.901	0.093
课程分支	0.875	0.089
HSG	0.932	0.114
SSG	0.914	0.102
居住位置	0.822	0.041
母亲教育程度	0.886	0.035
考试结果	0.803	0.147

上表中考试结果的增益值最大，将考试结果作为分裂属性。

可以得出这样的规则：

1）性别（女）∧课程分支（CS）∧HSG（O）⇒考试结果（优异）

这条规则表明新入学到计算机科学工程的女生，如果高中平均成绩为 90~100，在大学第一学期的考试中将取得优异的成绩。分析原因为高中成绩优异的女生在新入学的半年里能够保持高中学习的习惯，并且计算机科学类课程基础较多，不会有较难理解的专业课，因此考试结果获得优异的可能性较大。

2）性别（男）∧HSG（B）∧母亲教育程度（本科）⇒考试结果（通过）

这条规则表明新入学的男生，高中平均成绩为 70~79，母亲教育程度为本科，在新学年第一学期的考试中成绩通过。分析原因为高中成绩良好的男生虽然会受到一些外界影响，由于母亲的教育经历的影响，督促学生考试通过的可能性较大。

3）HSG（C）∧SSG（C）∧课程分支（ME）∧居住位置（农村）⇒考试结果（失败）

这条规则表明高中和初中的平均成绩为 60~69，居住在农村，选择的机械工程类课程的学生在新学年的考试中面临失败的风险。分析原因是学生居住在农村，由于基础教育资源稍差，并且初中和高中的平均成绩一般，表明没有打小较好的基础，存在偏科的情况，大学英语将是一个关卡，尤其机械工程类课程难度较大，学生考试失败的可能性

较大。

4.1.5 算法分类器比较

表 4.2 显示了 ID3、C4.5 和 CART 分类算法应用于上面的数据集使用 10 倍交叉验证的精度如下：

表 4.3 分类器精度

算法	正确分类实例	错误分类实例	用时(秒)
ID3	63.142%	26.658%	0.004
C4.5	68.254%	31.225%	0.03
CART	63.154%	38.147%	0.09

表 4.2 显示，相比其他两种算法 C4.5 正确分类实例精度最高为 68.254%。ID3 和 CART 算法还处于一个可接受的准确性水平。另外在分类器训练数据构建模型所用时间上来看，ID3 具有一定的优势。

下面的表 4.4 显示了三个机器学习生产预测模型的准确性。

表 4.4 分类器预测结果准确率

算法	Result	TP(真正元比率)	FP(假正元比率)
ID3	通过	0.714	0.188
	提高	0.625	0.231
	失败	0.784	0.064
C4.5	通过	0.745	0.207
	提高	0.517	0.211
	失败	0.786	0.091
CART	通过	0.809	0.347
	提高	0.31	0.15
	失败	0.641	0.112

数据挖掘分类技术帮助研究人员有效地通过知识发现数据，决策树算法生产分类规则，比其他分类方法更容易解释。通过对常用的决策树分类器进行了研究和实验，找到最好的分类器用于预测工科学生在第一学年考试表现。从分类器预测准确度很明显的发现 ID3 和 C4.5 决策树方法模型成功识别的学生可能会考试失败。这些学生可以考虑适当的咨询等提高他们的考试结果。机器学习算法 C4.5 决策树算法根据学生过去几年积累的数据可以有效的预测模型。结果表明，我们可以通过应用预测模型预测新生的第一学期考试成绩表现。

4.2 发现有潜在重修风险学生

本节关注的任务是选择有潜在重修风险的学生。对于高校学生来说，当考试分数低于重修分数的时候需要在下学期开学初参加重修辅导班，并重新考试。这不仅增加了老师的教学负担，也给了学生带来不必要的负担，在数据挖掘技术的帮助下，我们能够更准确的选择有针对性的学生。

4.2.1 设计目标

设计目标分为以下两个步骤来解决这个问题：

1）识别潜在的薄弱的学生。

2）为薄弱学生选择重修风险较大的课程。

第一步可以被看作是一个分类问题。然而事实并非如此，使用分类器预测谁一定会在考试中重修的精度太低。例如，使用普通程度的结果，分类系统 C4.5 只能识别弱势学生的一半， CBA 系统只能识别三分之二薄弱的学生。

从我们的经验和实验发现，得分方法是更合适的，而不是给每个学生分配一个明确的类(“ 弱 ”或“ 不弱 ”学生)，评分模型分配一个概率估计每个学生表达的可能性。然后灵活地选择一个学生的具体学科作为预警。

虽然已经存在的基于决策树和评分方法贝叶斯模型，一般来说本文基于得分关联的方法优于这些传统方法。新方法基于关联规则，我们需要通过引入关联规则的概念挖掘。

4.2.2 关联规则设计

关联规则是一个重要的类存在于规律数据库。典型应用关联规则是市场篮子分析，分析客户购买的商品是如何关联的。一个例子关联规则如下，奶酪→啤酒(sup= 10%, conf = 80%)这条规则说，10%的顾客会购买奶酪和啤酒在一起，那些购买奶酪还买啤酒的概率为 80%。关联规则挖掘模型可以表示如下： $I=\{i_1,i_2,.....,i_m\}$是一组项目，D 是一组事务(数据库)，每个事务 d 是一组项目， $d \subseteq I$ ，关联规则表示为 $X \to Y, X \in I, Y \in I, X \cap Y = \phi$，规则 X→Y 在事务集 D 的置信度为 c，表示在 D 事务集中 c % 支持 X 还支持 Y。给定一组事务 D(数据库)的问题挖掘关联规则是所有关联规则发现支持度和置信度大于或等于用户-指定最小支持度和最小置信度。

一、需要解决的问题

使用关联规则进行评分，我们需要解决许多问题。

1、关系表。

应用程序中的挖掘关联规则，需要从一个关系挖掘关联规则表(而不是一组事务)，因为我们的任务使用这张表格的数据。对关联规则挖掘这种类型的数据，需要每个数值属性离散化后间隔关联规则挖掘，只需要分类值或物品。离散化后，可以将每个数据(记录)数据集作为一组(属性，值)对和一个类标签。一个(属性，值)是一项，每一个数据成为一个事务。现有的关联规则挖掘算法可以应用于数据集。

在传统的关联规则挖掘中，任何项目都可以出现在等式或规则的左侧或右侧。对于得分，我们有一个固定的两个类的类属性。因此，我们感兴趣的规则使用一个类等式。也就是说，关联规则的形式：$X \to C_i$，C_i是一个类的类属性，X 是其他属性的一个项目集。感兴趣的类称为积极的类(如，“薄弱”学生类)，其他类称为负类(如，“不弱”学生类)。规则需要满足最小支持度。

如使用现有的关联规则挖掘算法，这种类型的规则很简单，只要找出满足最小支持度的规则就可以：$<item_1, item_2, ..., item_k, C_i>$，$C_i$事先是固定的。在挖掘过程中，每次迭代添加一个新条目到项目集。也就是说，在第一次迭代发现的所有项集为$<item_1, C_i>$，在第二次迭代中，发现项目集为$<item_1, item_2, C_i>$，以此类推。

2、最小支持度和最小置信度

传统的关联规则挖掘使用一个最小支持度和一个最小置信度。这不适合我们的任务，因为班级分配给我们的数据可能很不平衡。首先讨论最小支持度的问题：

1）如果设置得太大，我们不可能发现涉及少数类规则，而这通常是我们感兴趣的类。

2）为了让关联规则涉及少数类，设置最小支持度非常低的话。这可能导致组合过多，因为多数类可能有太多的规则，大多数都符合很多条件但覆盖很少的数据。这些规则没有预测价值，也导致增加了执行时间。

单一的最小置信度也会造成问题。例如，在一个数据库，只有 5%弱学生和 95%的不弱学生，如果我们设置最小置信度为 96%，我们可能不会找到任何“弱”类的规则，因为它以如此高的置信度在这个数据库包含可靠的“弱”类的规则是不可能的。如果我们设定一个较低的置信度，比如 50%，我们会发现 50%-95%之间存在很多规则，其中很多规则是无意义的。

3、解决方案

我们通过对不同的类使用不同的最小支持度和最小置信度来解决这个问题，我们只需要用户指定一个总的最小的支持度 t_minsup，然后根据类分布数据分发给每个类如下：

$$\min\sup(C_i) = t\_\min\sup \times \frac{f(C_i)}{|D|}$$

$f(C_i)$ 是训练数据中 C_i 类实例数量，D 是训练数据中所有实例的数量。使用这个公式可以给频繁(消极的)类更高的最小支持度，给罕见的(积极的)类较低的最小支持度。这将确保产生足够的积极的类规则，不会产生太多毫无意义的负类规则。

对于最小置信度，我们也使用下面的公式来为每个类分配最小置信度。

$$\min conf(C_i) = \frac{f(C_i)}{|D|}$$

这个公式的意义在于不会产生 C_i 类规则置信度小于最小置信度, 因为这些规则没有任何意义。

4、对得分使用关联规则的问题

关联规则挖掘关键的特性是它的完整性，即它的目标是发现数据的规则。可以利用规则设计好的得分函数。存在的问题是经常有很多可以应用的规则，不同的规则可能会有不同信息, 其中许多甚至相互矛盾。例如, 一个规则可以说数据为积极的类概率为 0.9, 而另一个规则说它属于消极类概率为 0.9，问题是应该相信哪条规则。对于传统分类系统，这不是一个问题因为通常只有一个答案。

二、使用关联规则用于学生得分

从训练数据生成的规则可以用于取得新的(或测试)数据。由于每个规则有附加的支持度和置信度，因此容易设计一个模式用于得分数据。利用关联规则设计最好的评分方法是非常困难的，因为有无数的可能。下面描述一个既有效且高效的启发式技术。

需要达到的目标时：

1）对于有许多符合置信度规则的积极类包含数据的情况下，数据应该分配高分数。

2）当消极类数据有很多置信度规则符合的情况下，对这些消极类数据分配较低的分数。

因此，我们假设以下总体得分函数，这是一个加权平均考虑上面的信息(支持度信息在权重中)。下面的公式表明给定一个数据情况下，数据的得分情况。分数 S 的值在 0 与 1 之间。

$$S = \frac{\sum_{i \in POS} W^{i}_{positive} \times conf^{i} + \sum_{j \in NEG} W^{j}_{negative} \times conf^{j}_{positive}}{\sum_{i \in POS} W^{i}_{positive} + \sum_{j \in NEG} W^{j}_{negative}}$$

POS 覆盖数据实例的积极类项集，NEG 是覆盖数据实例的消极类项集，$conf^{i}$ 是原始的积极类规则置信度。$W^{i}_{positive}$ 是积极类规则 i 的权重，$W^{j}_{negative}$ 是消极类规则 j 的权重，$conf^{j}_{positive}$ 是消极类规则 j 转换为积极类规则的置信度。

下面的问题是需要解决权重的问题，这里包括消极类权重和积极类权重。可用的信息是支持度和置信度。经过大量实验，发现将二者结合起来是比较理想的。积极类权重公式如下：

$$W^{i}_{positive} = conf^{i} \times \sup^{i}$$

$conf^{i}$ 和 $\sup^{i}$ 是积极类原始的置信度和支持度，下面公式为消极类权重。

$$W^{j}_{negative} = \frac{conf^{j} \times \sup^{j}}{k}$$

$conf^{j}$ 和 $\sup^{j}$ 是消极类原始的置信度和支持度。k 是一个用于不断减少消极类规则影响值(这些规则通常有很高的支持度和置信度)。通过实验来确定 k，这里确定 k = 3，这个值是执行表现最好的。

要注意是计算消极类规则权重时，不将消极类规则转换成积极类规则并使用积极类的支持度和置信度。相反，仍然使用小积累原始的的支持度和置信度。这有助于我们实现上述两种目标。

最后，当出现多个具有相同的分数数据情况时，用下面一个优先级(值是 P)公式计算：

$$P = \frac{\sum_{i \in POS} sup^{i} - \sum_{j \in NEG} sup^{j}}{|POS| + |NEG|}$$

| POS |和| NEG |表示积极类和消极类各自规则的数量。这个公式使用规则的支持度计算优先级。基本上，获得较高优先级的是具有高支持度的积极类和低支持度的消极类规则，当数据不满足任何规则分配 S = 0，P = 0。

4.2.3 关联规则分析

利用关联规则对在校学生的成绩进行数据挖掘，旨在从中选出具有重修风险的学生，因此从记录中随机选择 50 条左右，选择的属性有性别、高中成绩、作业完成率、出勤率、课程类别、职务、课外活动、期末成绩共八类。

一、离散化数据

使用关联规则前对数据做离散化，即将不同的属性继续分类，可以减少数据处理工作量并方便分析。

A.性别属性离散化：A1(男)、A2(女)。

B.高中成绩数据离散：B1（优异)、B2（良好)、B3（普通）

C.作业完成率离散化：C1（较好)、C2（一般)、C3（较差）

D.出勤率离散化：D1（较好)、D2（一般)、D3（较差）

E.课程类别离散化：E1（专业课)、E2（公共基础课)、E3（选修课）

F.职务离散化：F1（学生会及社团)、F2（班级干部)、F3（学生）

G.课外活动：课外活动的项比较多，这里大概划分为三类，G1（网络游戏)、G2（网络聊天、电影、购物等休闲)、G3（户外活动）

H.期末成绩：因为只是挖掘具有重修风险的学生，重修分数点为 45 分，因此分为两类 H1（大于等于 45)、H2（小于 45）.

下表 4.5 为部分数据样本。

表 4.5　部分数据样本

编号	性别	高中成绩	作业完成率	出勤率	课程类别	职务	课外活动	期末成绩
1	A1	B2	C1	D1	E1	F3	G3	H1
2	A1	B3	C3	D3	E2	F1	G1	H2
3	A1	B1	C2	D2	E1	F2	G2	H1
4	A1	B2	C3	D3	E3	F3	G1	H2
5	A1	B3	C3	D2	E1	F1	G1	H1
6	A1	B3	C3	D3	E2	F3	G1	H1
7	A1	B2	C1	D1	E1	F3	G3	H1
8	A2	B1	C1	D1	E3	F2	G2	H1
9	A2	B3	C2	D1	E1	F3	G3	H1
10	A2	B2	C3	D2	E2	F1	G2	H2

为使得数据挖掘获得有益的信息，设定的最小支持度为 14%，最小置信度为 80%，根据上述计算，最小频繁计数为 7，获得的频繁项集如下表 4.6 所示。

表 4.6　数据频繁项集

编号	作业完成率	出勤率	课程类别	课外活动	期末成绩	频繁计数
1	C3	D3	E1	G1	H2	7
2	C3	D2	E1	G2	H2	6
3	C3	D1	E2	G1	H2	5

找出大于最小支持度的频繁项集计算该频繁项集的最小置信度，公式如下：

$$confidence(A \Rightarrow B) = P(A|B) = \frac{\sup port\_count(A \cup B)}{\sup port\_count(A)}$$

如果该值大于设置的最小置信度，则视为强关联规则。以期末成绩为目标属性，得到的关联规则为：

1） $C3 \wedge D3 \wedge E1 \wedge G1 \Rightarrow H2$， $confidence = \frac{7}{7} = 100\%$

2） $C3 \wedge D2 \wedge E1 \wedge G2 \Rightarrow H2$， $confidence = \frac{6}{7} = 85.7\%$

3） $C3 \wedge D1 \wedge E2 \wedge G1 \Rightarrow H2$， $confidence = \frac{5}{7} = 71.4\%$

根据最下置信度为 80%去掉第三条规则，分析满足条件的前两条规则。

1）规则 1 分析：作业完成率 C3（较差）、出勤率 D3（较差）、课程类别 E1（专业课）、课外活动 G1（网络游戏），这些学生的期末成绩 H2（低于 45 分）。分析原因可以很清楚的看出提高课程成绩的保障是作业完成率和出勤率，课外的网络游戏活动必然对学生的精力有较大影响，另外由于专业课的难度较大，因此学生重修的风险较大。

2）规则 2 分析：作业完成率 C3（较差）、出勤率 D2（一般）、课程类别 E1（专业课）、课外活动 G2（网络休闲类），这些学生的期末成绩 H2（低于 45 分）。除了规则 1 中已经分析过的原因之外，发现出勤率提高一些对整体成绩帮助并不大，没有一个好的出勤率必然无法保证作业的完成率，也必然会由于专业课的难度而无法深入的理解内容，因此课程重修的风险较大。

根据上述频繁项集，以课程为目标属性的话，同样可以得到两条规则，基本分析类似不再赘述。

4.2.4　算法技术比较

现在比较结果，这里采用两种评分技术，一个计分法是基于 C4.5 决策树系统，另一种是基于天真贝叶斯技术。模型构建所需的训练数据来自于前四年学生各课程的考试

成绩。根据各算法不同，分组边界和重修风险课程对于每个组可能会有所不同。表 4.6 显示了系统基于不同标准的测试结果。

表 4.6　基于不同标准测试结果

选择标准	PE	PS	Recall
传统模式	142/269=52.79%	42/112=37.5%	42/45=93.3%
本文关联规则	162/264=61.36%	41/87=47.13%	41/45=91.1%
天真贝叶斯	137/261=52.5%	40/92=43.5%	40/45=88.9%
C4.5	136/234=58.1%	36/77=46.8%	36/45=80%

从这个表我们可以观察如下：

检查 PE 测量显示，利用本文关联规则方法，教师花费在重修班(264)小于传统选择方法(269)，而花在提高有重修风险的学生的努力从 142 提高到 162。这是非常重要的，因为花更多的努力在实际较弱的学生身上。

检查 PS 测量显示，传统选择方法需要总共有 112 名学生去参加重修课程，本文方法为 87 名。这意味着通过应用数据挖掘技术能够减少学生重修人数。这是一个很大的改进。

检查召回率 Recall 显示，与传统方法基本一致。

本文方法也比其他两种评分方法（天真贝叶斯和 C4.5）有一定的优势。查看下表 4.7，这是三种方法各自对学生评分后排名的情况。

表 4.7　三种方法评分排名对照表

	本文关联规则	C4.5	天真贝叶斯
1	15	11	12
2	10	9	9
3	8	0	8
4	3	2	4
5	5	3	4
6	0	1	1
7	2	2	1
8	2	4	2
9	0	5	3
10	0	2	1

上述评分表是通过训练数据构建模型后，对学生测试数据评分情况，表 4.6 显示了三个不同的评分系统对于有重修风险的学生的评分情况。显然，我们看到本文计分法优于其他两种，评分范围较大，同时最好的选择薄弱的学生，包括三个 0 分学生，即是说本文方法比其他两种方法能够挖掘出更多的薄弱学生。

在本节中，设计的目标是选择潜在的薄弱的学生，提前预警其重修风险。使用数据挖掘技术达到更好的结果，减少了学生和老师的负担。本节所使用的关联规则是在传统的方法上加以改进。最大的改进之处是改变传统的只有一个最小支持度和一个最小置信度，而是针对不同的分类规则设计公式来分配不同的最小置信度和最小支持度，并针对消极类和积极类设置不同的规则权重，从而改进了因数据不平衡产生的一些问题，比如预测不准确等。

4.3 挖掘影响学生成绩因素

4.3.1 设计目标

近年来，许多研究人员已经逐渐改变使用统计技术计算学生的分数与数据挖掘技术，而改用数据挖掘技术，采用粗糙集理论调查学生分数基本数据库，并提取分类规则理解影响学生成绩的因素。分类规则通常有平时成绩、课堂上和效率、课后学习时间等，这些主要因素影响整体分数。在本节中使用 ID3 决策树算法，分析学生的分数影响的关联因素，包括学生社会活动、娱乐活动、甚至英语分数。最后结论是道德、智力和身体彼此相辅相成。使用 Apriori 关联规则分析大学学生的得分，发现一些课程与后续专业课程的成绩密切相关。传统方法的主要缺点是把学生的分数排序并分区，在此基础上进行聚类。然后使用了 k -means 集群算法获得学生分数排名分区。

本节利用聚类算法和决策树挖掘学生的分数，通过分析学生-课程-成绩之间的关系可以获得一些教学和管理有价值的信息，如学生的关系，课程之间的关系和课程之间的影响。

4.3.2 分析学生分数挖掘设计

数据挖掘的主要流程如图 4.4 所示。

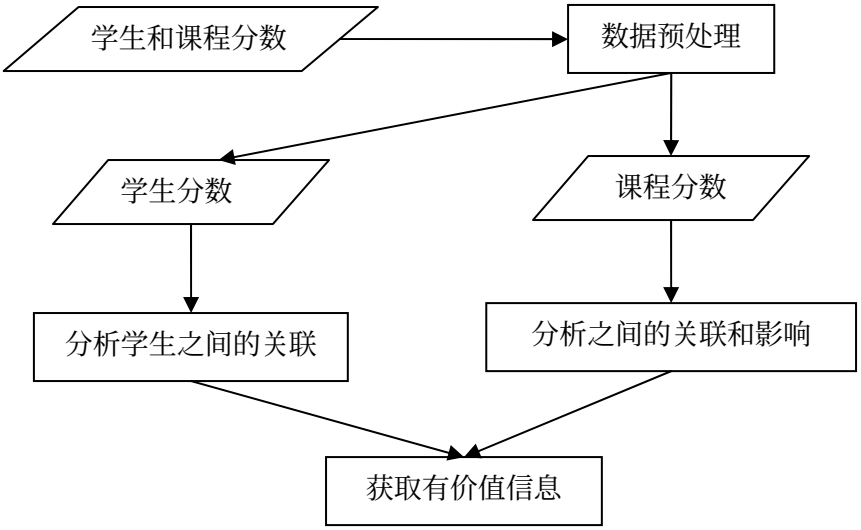

图 4.4　挖掘挖掘主要流程

定义 1：学生和课程的分数数据集。让 $D=\{d_1,d_2,...,d_n\}$ 表示学生的分数数据集。d_i 代表第 i 个学生信息，包括学生基本信息 s_i (学院、系、姓名、学号等)，课程信息集 c_i 包含学生学习的所有课程信息，整体评级包括课程的数量，所需学分，已完成学分和平均数量)。假设 m 代表课程的总数，所以 $c_i=\{c_{i1},c_{i2},.....,c_{im}\}$，$c_{ij}$ 代表了第 j 个课程信息(包括课程编号、课程名称、学分、教师、考试分数，结果构成等)重要因素。

定义 2：学生的分数。假设 $score_{ij}$ 表示第 i 个学生的第 j 个课程得分。学生分数表达式：

$$SCORE=(score_{i1},score_{i2},.....,score_{im})_{i=1}^{n}$$

定义 3：课程的分数。通过反相，我们获得的分数表达式：

$$CSCORE=(\csc ore_{i1},\csc ore_{i2},.....,\csc ore_{in})_{i=1}^{m}$$

学生成绩分析的过程有以下三个步骤。

1）预处理原始数据集 D 获得学生得分 SCORE。

2）利用聚类算法分析学生之间的关系。

首先，划分学生分数 SCORE 获得排序。通过聚类课程分数 CSCORE，我们可以避免错误产生线性变换多维评估分数，进而获得更好的分数。直接聚类不能区分重要的学位课程，学分高的课程几乎是最基本的理论课程，实践课程学分较低。为了更好的挖掘学生关系，需要区分理论基础和实践能力。本节使用归一化法，设置学分权重分配每个学生得分 SCORE。归一化法公式如下：

$$\frac{x_i * W_i}{\sqrt{\sum_{i=0}^{N}(x_i * W_i)^2}}$$

x_i是第 i+1 课程考试分数，W_i是第 i+1 课程学分，N 代表课程总数。设置平均属性缺失值。这样可以知道谁的基础理论知识更好，谁的实践能力更好。

3）分别利用聚类算法和决策树算法分析之间课程之间的关系和影响。

把所有的课程分为三个部分：公共基础课程、专业课程和选修课程课程。如果学生不知道课程的困难程度，他们无法有效地安排他们时间。如果学校不知道选修课程的广泛程度，他们无法有效地设置选修课程。为了能够深入理解课程，本节使用聚类算法分析课程和课程之间的关系。安排现有的课程也是关注点，如果不能有效地安排课程，它将对某些专业课程有一些负面影响。为了发现课程之间的影响以便更有效地安排这些课程，采用决策树算法分析课程分数。

4.3.3 学生分数分析挖掘实现

选择某高校计算机工程系学生的分数,这个数据集包含四个年级106个学生的成绩。使用聚类算法和 J48 决策树分析数据。

一、基于聚类给学生排序

有 167 个课程纳入课程信息，大多数是公共选修课程专业选修课程。每个学生选择的课程不同于其他人，每个学生只能选择一些课程，我们不能直接集群学生得分矩阵，因为学生考试分数包含许多缺失值。选择最基本的课程和 30 个专业课程作为学生的考试分数属性。对考试分数缺失的记录，他们考试结果将被记录为“？”此外，没有被学生选择的课程，它们的考试结果也被记为”?这些“？”标记将被平均值所取代。

利用 EM(Expectation-maximization 最大期望)聚类算法对数据集聚类，参数为默认。结果得到四个集群。结果如图 4.5 所示。

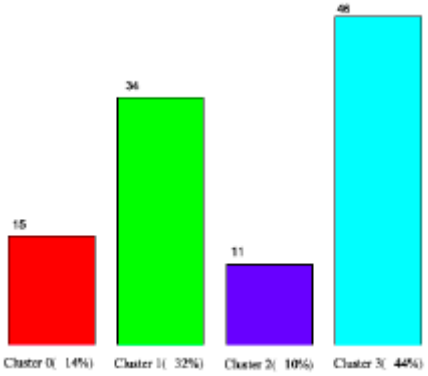

图 4.5 EM 聚类结果

四个集群的分数曲线如下图 4.6 所示。

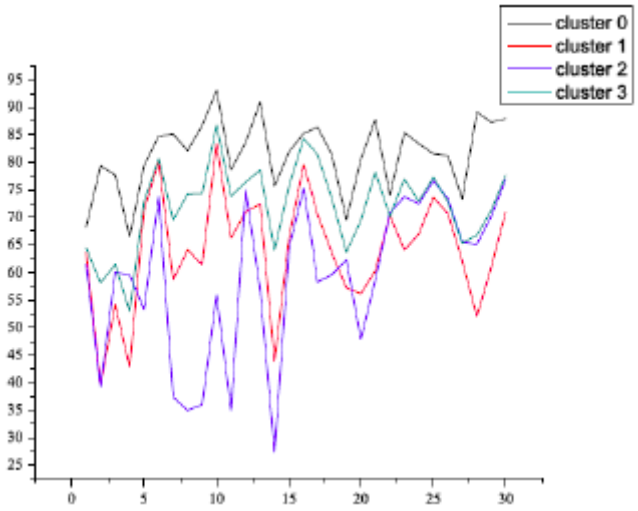

图 4.6　四个集群的分数曲线

从图 4.6 中可以发现，集群 0 代表分数优异的学生，而集群 3 是分数较好的学生。集群 1 和集群 2 中的学生都表现糟糕, 集群 2 学生分数在一定程度上甚至比集群 1 更差。根据原始数据集，我们可以得出这样的结论，集群 0 代表的是 11 年入学的成绩优异的学生，集群 3 代表是分数较好的 11 入学的学生，集群 1 代表分数较差的 11 年入学的学生，集群 2 代表分数较差的 10 年入学的学生。我们可以把集群 2 作为中继器。

二、数据标准化后基于聚类分析学生之间的关系

学生考试分数数据标准化后，可用 EM 聚类算法(默认参数)重新聚类数据集，获得 3 个集群，如图 4.7 所示。

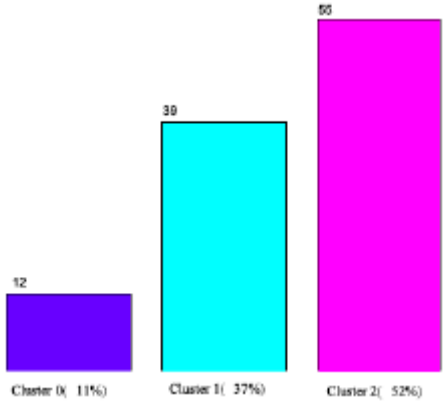

图 4.7　二次聚类结果

三个集群的分数曲线如下图 4.8 所示。

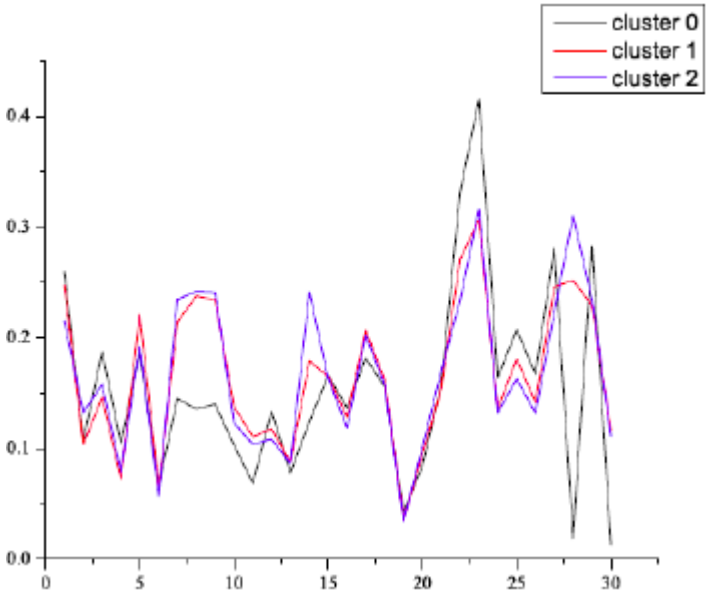

图 4.8　三个聚类的分数曲线

根据原始数据集和图 4.8，我们可以发现，集群 0 代表分数较差的学生，事实上在集群 0 有 11 个中继器，而剩下的人的分数也很糟糕，，如果他不改变自己也可能会成为一个中继器。中继器选择的一些课程已停止，新的学生不能选择。因此，许多中继器的分数与 11 年入学的学生分数之间没有可比性。例如，一个学生以前学习大学物理时成了一个中继器，他的得分是 68。对那些 10 入学的学生来说可能是很低分数，它对 11 年入学的学生来说可能是较好的分数。这包含许多因素，如不同时期教师已经改变了。这样我们也可以知道为什么集群 0 的一些分数比集群 1 和集群 2 还要高。经过上述分析可确定集群 0 代表中继器。

集群 1 的曲线和集群 2 是密切相关的，但是他们并不能代表优秀或好学生。为了直接显示两个曲线之间的差异，我们标记集群 1 的项目为 1，标记集群 2 的项目为-1。对比图如下图 4.9 所示。

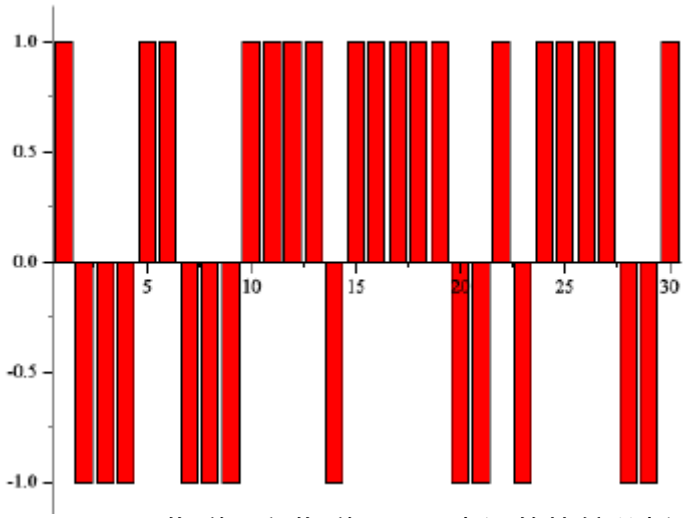

图 4.9　集群 1 和集群 2 项目之间的比较分析

根据原始数据和图 4.9 的对比分析，可以得到结论如下：

集群 1 的优势课程是大学英语 3、大学英语 4、电子实践，计算机科学与技术导论、数据库应用程序、数据结构、模拟电子课程设计、人工智能、电子商务、面向对象 c++ 编程、计算机算法设计与分析、物理实验、大学英语 1、工程化学、大学计算机基础、工程制图、大学英语 2、编程实践。

集群 2 的优势课程是哲学、大学物理、技术经济学、基本数字电路、离散数学、数据结构、计算机组织原则、线性代数、统计理论和方法、科学与工程 1、高等数学 1、高等数学 2，C 语言。

集群 1 的优势课程是英语、专业选修课程或其他注重实践能力课程(如物理实验、工程制图等)。然而，集群 2 的优势课程主要是基础或专业理论课程。因此，集群 1 学生有更好的实践能力，集群 2 学生有更好的基础理论知识。两个集群的成绩没有明显差异，他们分别代表不同能力。例如，管理专业学生在人际互动方面表现比计算机系的学生好，而计算机系的学生在技能方面表现更好一些。因此，集群 1 和集群的基础上，管理者可以帮助他们发展他们的优势，弥补他们的弱点。

三、基于聚类分析课程之间的关系

转换学生分数数据后，我们利用这部分学生的课程分数作为数据集，课程分数包括 167 记录。然后利用 EM 聚类算法聚类课程分数，得到 4 个集群，如图 4.10 所示。

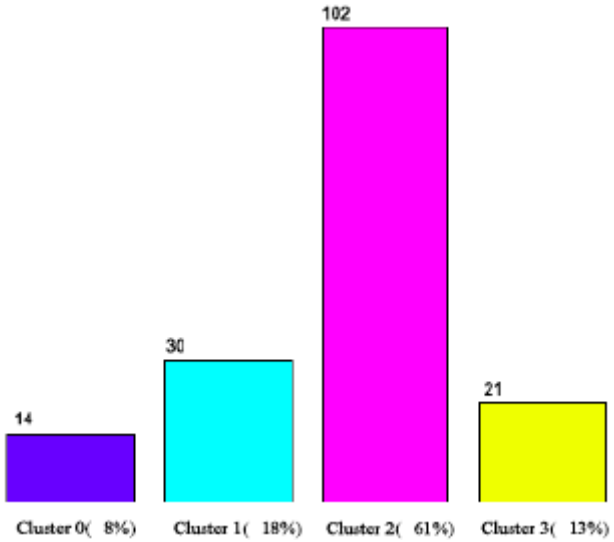

图 4.10　EM 聚类结果

集群 3 与集群 0 课程分数曲线对照图如下图 4.11 所示。

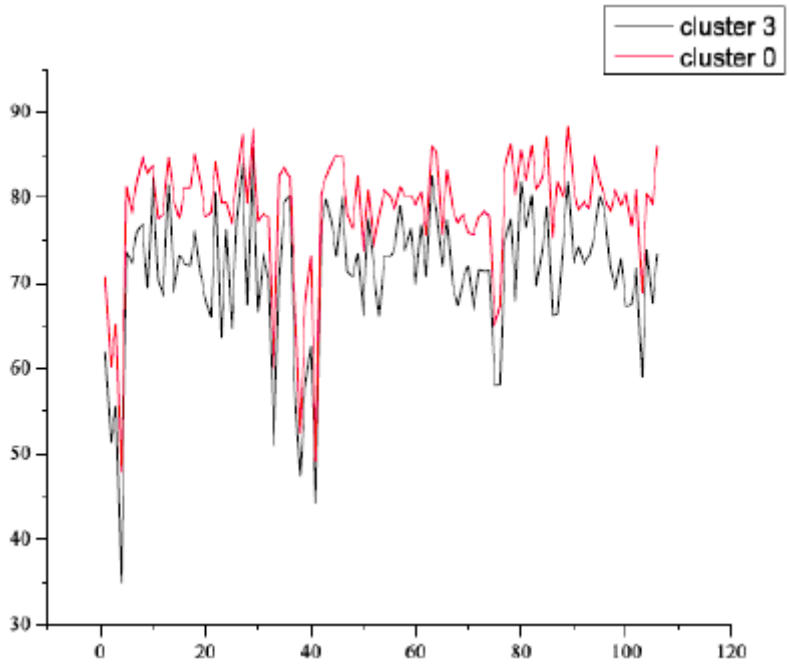

图 4.11　集群 3 与集群 0 课程分数对比分析

集群 3 与集群 2 课程分数曲线对照图如下图 4.12 所示。

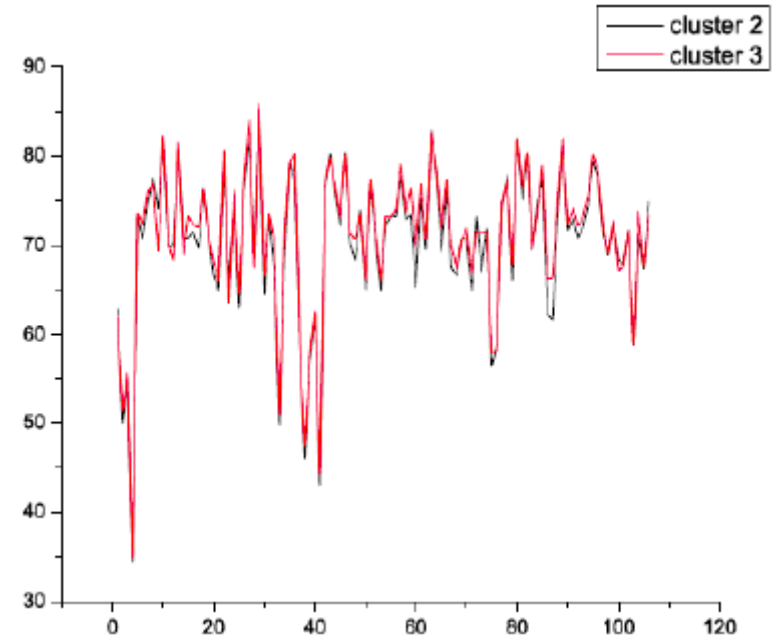

图 4.12　集群 3 与集群 2 课程分数对比分析

集群 1 与集群 2 课程分数曲线对照图如下图 4.13 所示。

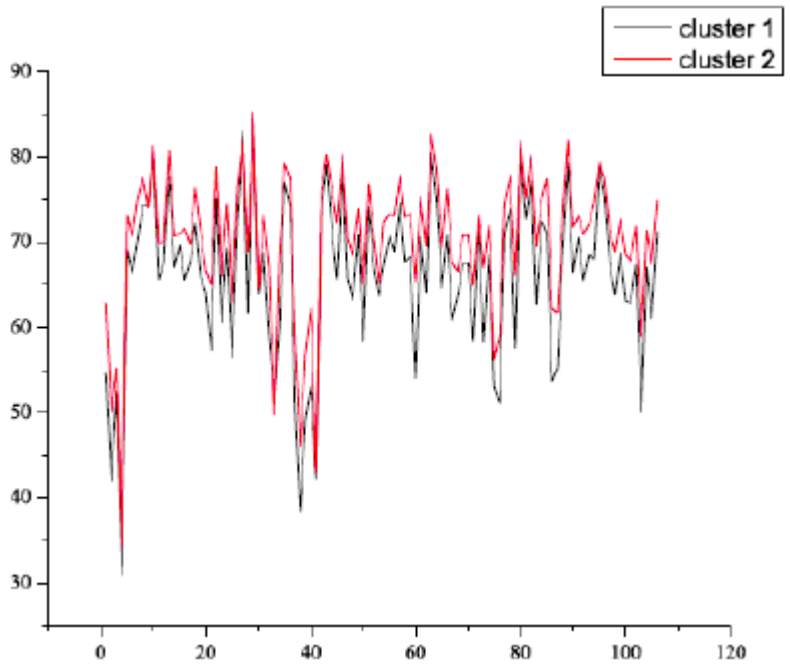

图 4.13　集群 1 与集群 2 课程分数对比分析

根据图 4.11、图 4.12 和图 4.13 对照分析，可以发现集群 0 的课程分数比较优异，集群 3 课程分数较好，集群 1 课程分数较差，集群 2 课程分数比集群 1 稍好一点。

根据原始数据集，我们可以获得一个结论和一些有用的信息。集群 0 高分课程几乎都是简单的公共课程。集群 1 的课程是艰难的公共课程，很多课程是专业课程。因此，集群 1 课程更重要。集群 2 课程是一些学生选择的选修课程。集群 3 课程是比较普遍的选修课，可以反映学生的兴趣。

根据这些有用的信息，学生可以知道哪些课程是重要的，哪些是容易的，进而能够有效地安排时间。同时，大学可以顾虑学生的利益，有效设置选修课程。

四、基于决策树算法分析课程之间影响

选择数据结构课程作为分析对象，使用 J48 决策树分析数据结构。因为有太多的属性，如图 4.14 所示，我们只画出决策树的大概结构。

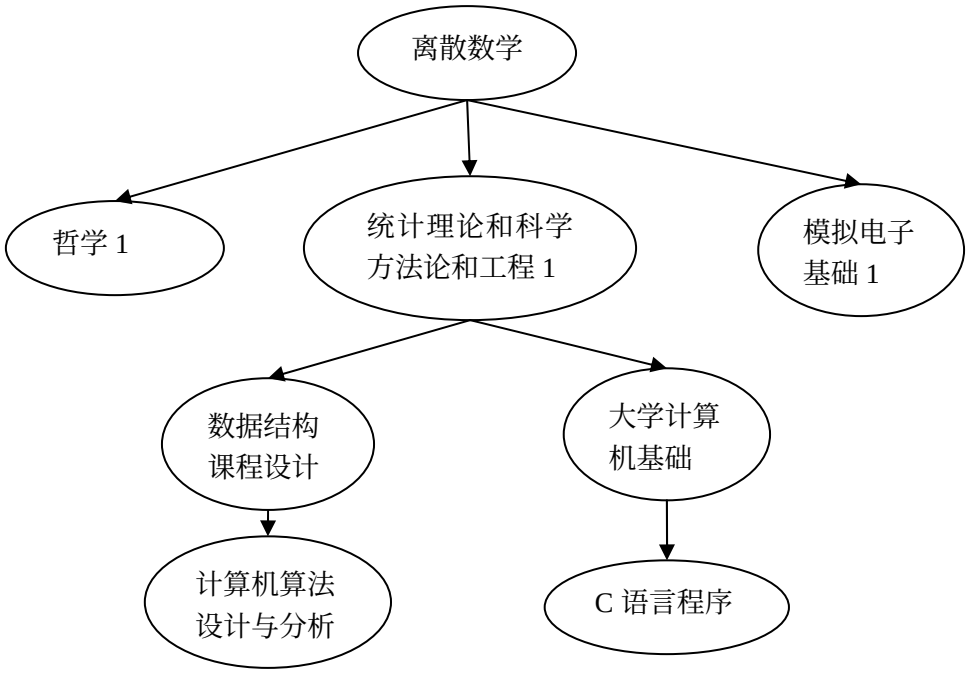

图 4.14　数据结构决策树

如图 4.14 所示，离散数学对于数据结构是非常重要的课程。这是因为图论等离散数学知识是对学习数据结构是有用的。学生一般在第二学期学习离散数学和数据结构。考虑离散数学的重要性，应该在数据结构课程之前开离散数学课程。

4.3.4　结论

使用聚类算法和决策树来全面分析学生的得分，并获得一些规则有利于教学和管理。由于数据集很小和不完整，无法进一步分析学生的分数，还需要引进社会网络分析学生和学生之间的关系，或学生和老师之间关系，从而发现影响学生成绩的因素。

4.4　本章小结

本章介绍了三个功能的设计与实现：C4.5 决策树算法根据学生过去几年积累的数据可以有效的预测新生的第一学期考试成绩表现；利用改进的关联规则挖掘技术有效预测有重修风险的学生；使用聚类算法和决策树来全面分析学生的得分，进而得出学生之间的关系、学生与课程之间的关系和课程之间的影响。

第 5 章 系统测试

5.1 测试平台 SAS Enterprise Miner

SAS Enterprise Miner 的 GUI 界面是数据流驱动的，易于理解和使用。它允许一个分析者通过构造一个使用链接连接数据结点和处理结点的可视数据流图建造一个模型。另外，此界面允许把处理结点直接插入到数据流中。SAS Enterprise Miner 运行在客户/服务器上。在客户/服务器模式下，可以把服务器配置成一个数据服务器、计算服务器或两者的综合。SAS Enterprise Miner 可以执行数据访问、操纵和预处理，数据界面贯穿于 SAS 数据集, 数据也能通过标准 SAS 数据程序访问 RDBMS 和 PC 格式数据的 ACCESS。对 Oracle、Informix、Sybase 和 DB2RDBMS 的支持是通过 ACCESS 来实现。

5.2 创建模型

5.2.1 输入数据源

指定输入数据输入数据源节点，双击或右键单击该节点并选择打开。单击“选择”以选择数据集。另外，可以键入数据集名称。在视图的数据集库中，选择定义的库列表 sampsio。选择设置在从数据集列表 sampsio.hmeq 数据，选择 OK。出现输入数据源对话框。

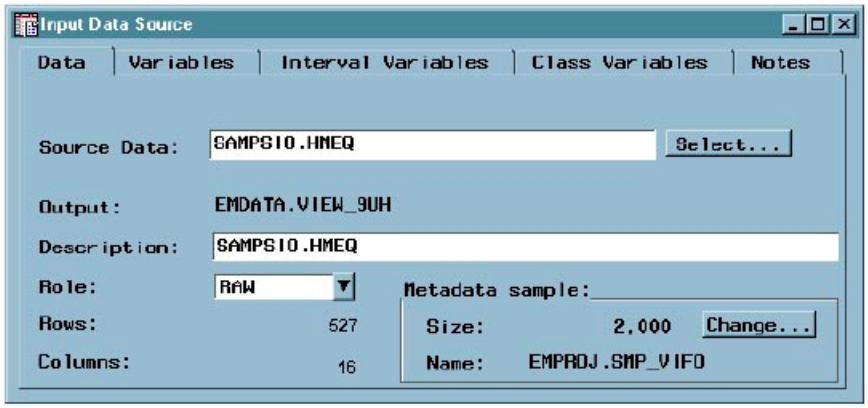

图 5.1 输入数据源

观察到该数据集有 527 个观测（行）和 16 个变量（列）。源数据集的名称是 sampsio.hmeq，size2000 表示一个元数据样本大小可以为 2000。

5.2.2 数据分区节点检查

数据挖掘需要的输入数据样本分为训练，验证和测试数据集。默认情况下，可以简单随机抽样的应用。也可以产生分层的样品或实施之前实施用户定义的样品：

- 简单随机抽样简单随机选择。在每一个观测数据设置要选择相同的概率。
- 分层抽样选择分层和当时使用的选项分层选项卡设置层。
- 用户定义采样，选择用户定义的和当时使用的选项、用户定义的标签来识别变量的数据集的标识分区。

选项卡的左下角，能指定一个随机种子初始化采样过程。在计算机程序随机通常开始某些类型的种子。如果使用的设置相同的种子相同的数据（除了种子= 0）在不同的流量，会得到相同的分区，不同的分区产生潜在的不同结果。

图 5.2　随机种子初始化

选项卡的右侧可以指定配置训练、验证和测试数据的百分比。百分比必须总计达100%。

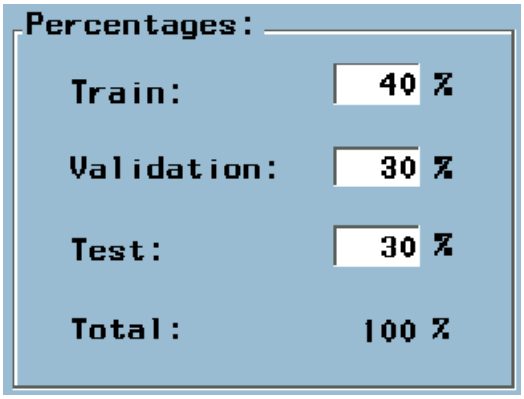

图 5.3　配置训练、验证、测试数据百分

5.2.3 变量缺失值替代

建立一个回归模型或神经网络模型，对所使用的所有训练数据集的观测是必要的，找到缺失值并替代。决策树直接处理缺失值，而回归和神经网络模型忽略所有不完整观察（变量有一个缺失值或多个）。这对建立在相同的一组观测值更适合比较模型，所以进行更换之前，可以将从决策树得到的模型与拟合回归或神经网络模型结果进行比较。

默认情况下，数据挖掘使用的样本训练数据集的选择值替换。

- 观察有缺失值为区间变量，缺失值用相应的变量样本的平均值取代。
- 观察有缺失值的二进制，转为名义或有序变量，

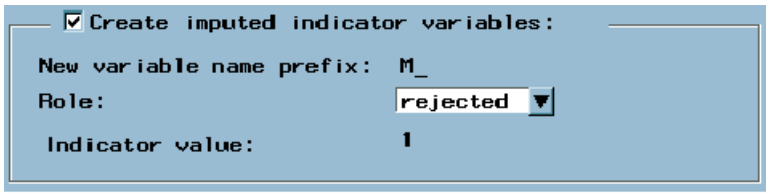

图 5.4　变量缺失值替代

5.2.4　交互式分组

添加一个互动分组节点连接到数据划分节点。在此设置窗口指定该节点使用的设置执行自动分组。互动分组节点以应用信息价值或基尼评分标准评价指标预测功率。可以选择自动提交复选框。此外，分组统计到 SAS 数据集的选择出口分组，并指定一个有效的 SAS 数据集名称统计集中。

冻结复选框可以防止用户修改分组被覆盖的自动分组结果。这里选择自动提交复选框，选择信息的价值提交准则，并修改提交的值为 0.3。同时，选择出口分组统计提交复选框。

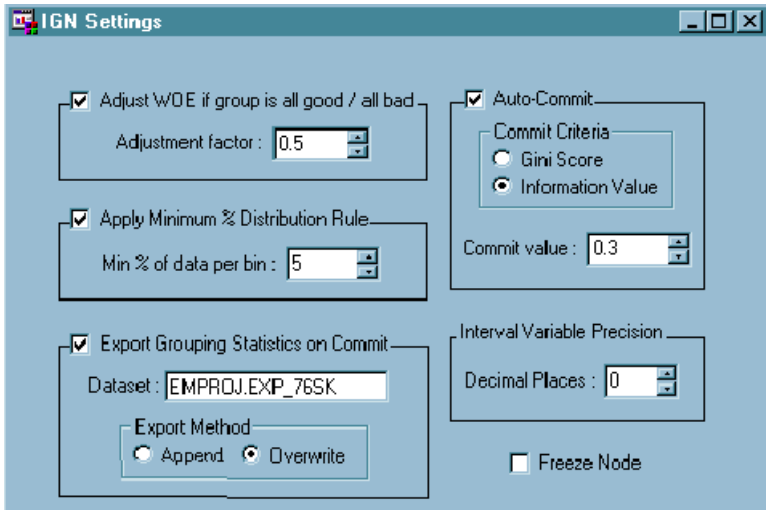

图 5.5　交互式分组设置窗口

互动分组设置进行自动分组，一个分组变时提交其信息价值或基尼系数大于节点指定设置价值 0.3 时，单击“是”查看结果，输出窗口选项卡提交类型，和信息的价值和基尼系数的值了保持现状。有信息的值小于 0.3 的变量被设置为“否”。

Keep	Commit Type	Auto-Grouped Infoval	Committed Infoval	Auto-Grouped Gini	Committed Gini	Variable
YES	AUTO	0.32282	0.32282	31.48315	31.48315	BRANCH
NO		0.13905	.	20.77339	.	SEX
YES	AUTO	1.9323	1.9323	67.21572	67.21572	HSG
YES	AUTO	0.58	0.58	33.40933	33.40933	SSG
YES	AUTO	0.36493	0.36493	22.39867	22.39867	ATYPE
NO		0.10552	.	16.93698	.	LAN
NO		0.23764	.	23.9369	.	LOC
NO		0.21016	.	24.27859	.	HOS
NO		0.21576	.	23.52207	.	FSIZE
NO		0.0149	.	5.90057	.	FSTAT
YES	AUTO	0.74527	0.74527	38.29417	38.29417	RESULT
NO		0.10181	.	17.10195	.	FAIN

图 5.6　自动分组输出结果

5.2.5　决策树模型评估

训练和验证决策树模型可以通过评估表和评估图显示，大的决策树需要考虑充分配合，以及是否过度拟合的问题。如果用于训练和验证数据的分类类似的所有子树，是在训练数据中存在的，默认情况下，用于验证和最少数量的评估值最高的子树叶被选择。在这个例子中，目标变量是二进制差。默认情况下，该误判率作为一个双目标模型评估措施。6～13 子树误分类的最小值率（0.121365）验证，选择 6 叶子树，分类矩阵包括每个级别的分类预测目标变量，其中缺席训练数据的 12%正确分类。正常训练数据的 77%正确分类。决策树图显示节点的统计、分离变量和拆分规则。顶部节点代表整个训练数据集。默认情况下，节点的颜色是成正比的，线宽度的比例比表明一个分支训练观测数根节点。

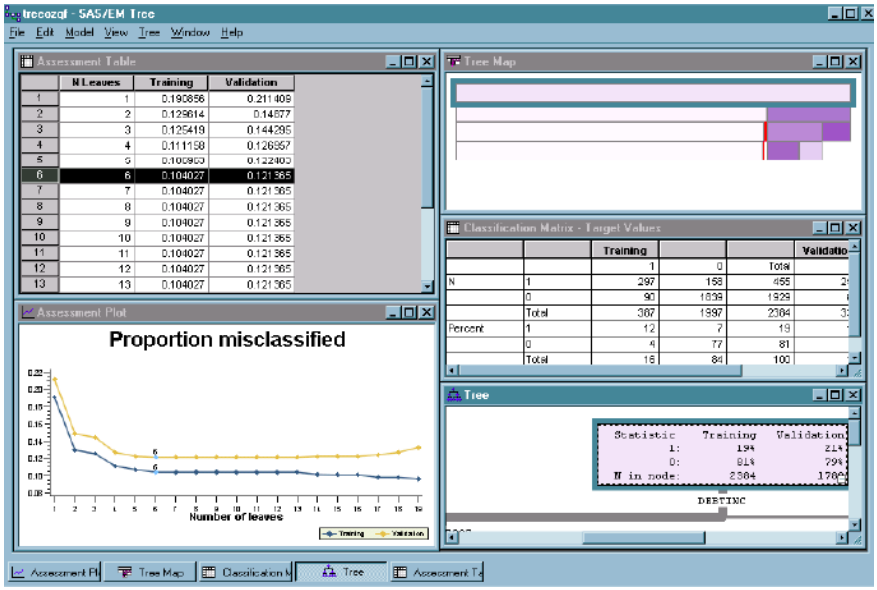

图 5.7　决策树模型评估图

5.3 学生分数聚类分析测试

利用 SAS 聚类分析测试第四章 4.3 学生分数分析，统计相似的学生。需要注意的是在这个例子中没有响应变量。只识别不同的学生群体。聚类分析常被称为监督分类因为它尝试预测组或类的成员，为一个特定的响应变量分类。聚类另一方面被称为无监督的分类，因为它确定组或类的所有输入变量中的数据。这些组、群被分配数量；然而，聚类数不能使用评估团簇之间的距离。

设置输入数据源节点：

1）打开输入数据源节点。

2）选择数据集。

3）模型设定角色名称 ID。

4）探索和描述统计分布。

5）选择区间变量标签，选择类变量标签和观察有没有缺失值。

设置群集节点：

1）打开群集节点。变量标签是活跃，使用一个标准化的选项。

2）选择标准差。

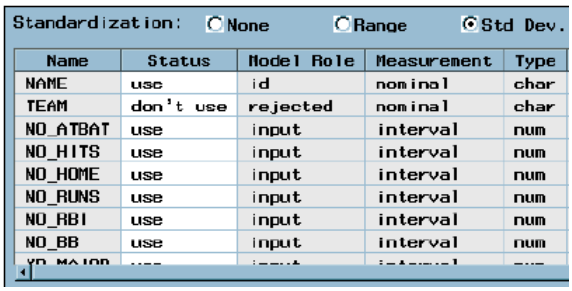

图 5.8 设置群集节点

默认情况下，群集节点使用立方聚类准则（CCC），可以更改默认的样本选择数据标签。指定一个不同的数字初始聚类选择的选择标准，算法集群剩余的簇，直到聚类为一个集群。

设置簇的数目以限制集群的数量：

1）打开群集节点。

2）选择集群标签。

3）选择簇的节数的选择标准。

4）簇的最大数目设置为 10。

5）单击“ 确定 ”。关闭集群节点，保存更改时提示。

聚类形成的三个集群如下图所示：

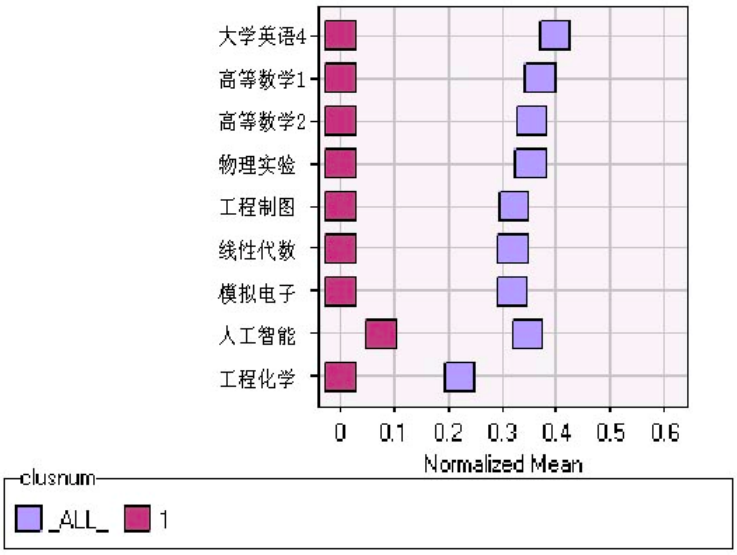

图 5.9　聚类集群 1

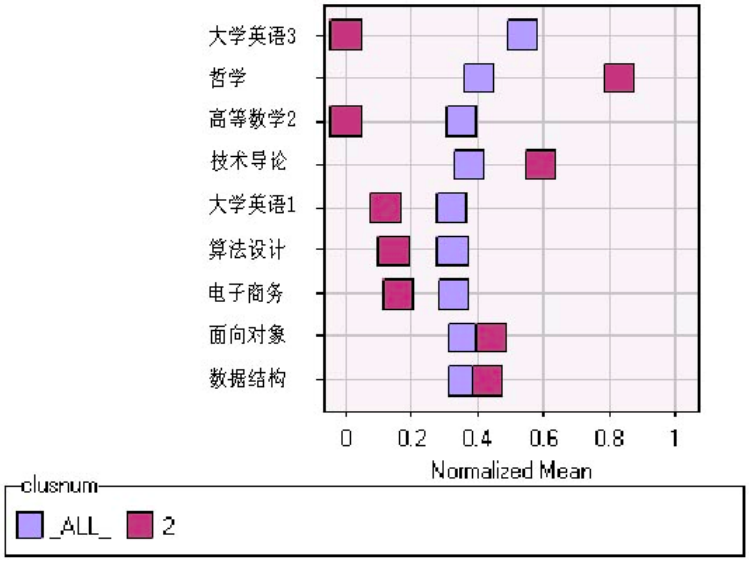

图 5.10　聚类集群 2

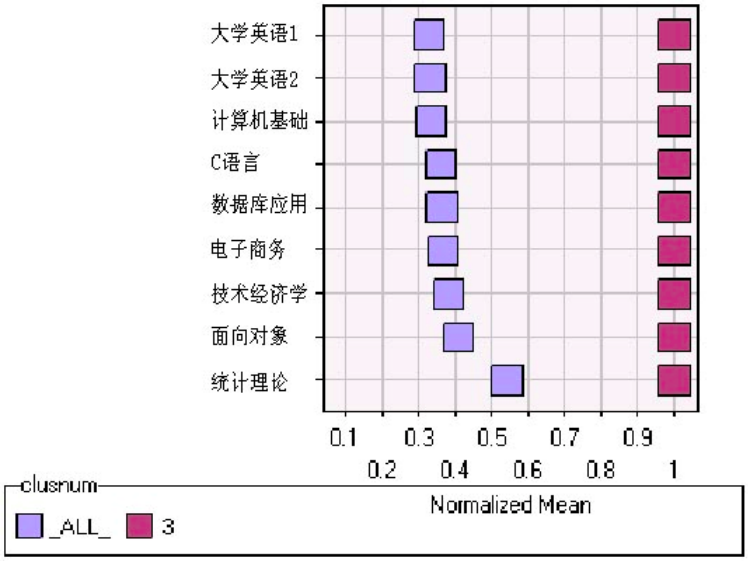

图 5.11　聚类集群 3

集群 3 学生的成绩比较优异，所对应的课程基本为简单的公共基础课。集群 2 学生的成绩较好，成绩好点的课程基本为选修课。集群 3 学生的成绩较差，多为比较困难的公共基础课。

5.4 本章小结

本章利用 SAS Enterprise Miner 平台对本文设计的数据挖掘模型进行测试，包括输入数据源、数据分区点检查、变量缺失值替代、交互分组、决策树模型、聚类分析模型等。

第 6 章 总结

随着大数据成为网络信息行业的重点词汇，在教育领域应用数据挖掘技术来为教育革命提供动力成为可能。数据挖掘技术对于中小学生以及高等院校学生的学习行为各类表现、学习的成绩以及毕业后的职业规划都可以提供有价值的信息，比如可以改善学生的学习方法、帮助学生发现在作业或考试中存在的一些经常被忽略但却很重要信息、为学生具体学科的学习提供个性化服务、及时发现学生潜在的辍学风险等。

教育数据挖掘 EDM 是一个新兴领域, 是为备满足日益增长的教育普遍评价的需要。EDM 侧重于收集、归档和分析学生学习相关的数据并进行评估.在学校对学生的教育中存在很多较为明显的数据，比如学生的入学率、报到率、上课考勤率、辍学率、奖学金分布比率等，当然最重要的是学生各科成绩分数的数据。进一步具体到课堂中也存在类似的数据，比如学生回答问题的准确率、提交的作业的正确率、课堂发言次数、师生互动时间、学生回答问题的平均时长等，这些数据按照数据挖掘技术的流程经过专业的收集、预处理、统计、分类和分析后，可以成为对学生多方位表现预测的依据。

本文设计的成绩分析系统主要是利用数据挖掘算法设计模型来分析教育数据，所要实现的目标有三点：

1）预测新入学学生是否能在第一学期考试通过。如果预测说一个学生倾向于在考试中失败，那么可以建议学生在考试之前采取额外的努力，提高自己成绩并帮助他通过考试。分类方法包括类似决策树、贝叶斯网络可以实现，本文采用的是 CART 和 C4.5, ID3 决策树算法。

2）利用数据挖掘技术选择有重修风险的学生。对于高校学生来说，当考试分数低于重修分数的时候需要在下学期开学初参加重修辅导班，并重新考试。这不仅增加了老师的教学负担，也给了学生带来不必要的负担，在数据挖掘技术的帮助下，我们能够更准确的选择有针对性的学生。在本文中，将采用一个基于数据挖掘的方法选择该类的学生。方法的关键是采用基于得分的关联规则技术，该方法具有非常不错的效果，优于传统方法。

3）通过学生的成绩分析学生之间关系、课程之间关系和影响。在教育中，学生的分数是一个非常重要的定量评价指示器，可以客观地反映教育的影响，是一个重要的科学决策依据。因此，分析和研究学生的分数是非常重要的。学生成绩数据库中的数据呈指数级增长。传统的查询和简单的统计分析无法满足分析的需要，不能为教学捕捉有用的信息。本文利用聚类算法和决策树挖掘学生的分数，通过分析学生-课程-成绩之间的关系可以获得一些教学和管理有价值的信息。

本文的系统测试采用 SAS Enterprise Miner 平台，它设计为被初学者和有经验的用

户使用,它的 GUI 界面是数据流驱动的，且它易于理解和使用。它允许一个分析者通过构造一个使用链接连接数据结点和处理结点的可视数据流图建造一个模型。另外，此界面允许把处理结点直接插入到数据流中。

本文最后利用 SAS Enterprise Miner 平台对本文设计的数据挖掘模型进行测试，包括输入数据源、数据分区点检查、变量缺失值替代、交互分组、决策树模型、聚类分析模型等，测试效果表明成绩分析功能能够实现。

作者简介

马丹，女，1983 年 3 月 27 日出生于辽宁省桓仁县，
2008 年 7 月毕业于沈阳师范大学，汉语言文学专业。
2009 年 4 月于辽宁建筑职业学院档案馆工作至今

参考文献

[1]毛国君.数据挖掘原理与算法（第二版）[M].北京:清华大学出版社.2007.10

[2]朱玉全.数据挖掘技术[M].南京:东南大学出版社.2006.11

[3]董琳.数据挖掘实用机器学习技术[M].北京:机械工业出版社.2006.8

[4]张治斌,王艳萍.数据挖掘技术在数字化校园中的应用[J].现代计算机.2006.5

[5]吴东升,数据挖掘在高校教学及学生学习评价中的应用空间分析[J].电脑知识与技术.2006

[6]刘红岩,陈剑著.数据挖掘中的数据分类算法综述[J].清华大学学报(自然科学版).2002.04

[7]罗海蛟.数据挖掘中分类算法的研究及其应用[J].微机发展，2003，13

[8]毛国君,数据挖掘的概念、系统结构和方法[M]. 北京:机械工业出版社，2002

[9]邵峰晶,于忠清.数据挖掘原理与算法[M]. 北京:中国水利水电出版社，2003

[10]陈志泊,数据仓库与数据挖掘[M]. 北京:清华大学出版社,2009.5

[11]ZhaoHui Tang.数据挖掘原理与应用——SQL Server 2005[M]. 北京:清华大学出版社，2007.1

[12]陈黎,王敏.聚类分析的方法[J].计算机工程与应用，2002，6

[13]高新波. 模糊聚类分析及其应用[J].西安:西安电子科技大学，2004.1

[14]David Hand,Heikki. Mannila,Padluaie.Principles of data milling[J].The MIT Press.2001

[15] Donald Feinberg, Mark A. Beyer. Magic Auadrant for Data Warehouse Database Management Systems. 2007.

[16] Cliff Longman, CTO, Kalido. Warehouse Lifecycle Management Concepts and Principles .2008

[17] J. R. Quinlan.Introduction of decision tree[J]. Journal of Machine learning， 1986：81-106.

[18] J. R. Quinlan.C4.5: Programs for Machine Learning [J], Morgan Kaufmann Publishers, Inc, 1992.

[19] Alaa el-Halees. Mining students data to analyze e-Learning behavior: A Case Study, 2009.

[20] J. Han and M. Kamber. Data Mining: Concepts and Techniques[J] Morgan Kaufmann, 2000.

[21] B.K. Bharadwaj and S. Pal. Data Mining: A prediction for performance improvement using classification[J], International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140,2011.

[22] U . K. Pandey, and S. Pal. Data Mining: A prediction of performer or underperformer using classification[J]. (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[23] S. T. Hijazi, and R. S. M. M. Naqvi. Factors affecting student sperformance: A Case of Private Colleges[J]. Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[24] Z. N. Khan. Scholastic achievement of higher secondary students in science stream[J] Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87,2005.

[25] Z. J. Kovacic. Early prediction of student success: Mining student enrollment data[J], Proceedings of Informing Science & IT Education Conference ， 2010.

[26] Galit.et.al Examining online learning processes based on log files analysis: [J].Research, Reflection and Innovations in Integrating ICT in Education 2007.

致　　谢

在研究生论文完成之际，我特别要感谢我的导师车翔久教授，在我研究生生涯中起到很重要的作用，车老师严谨的治学态度和谦逊的为人作风给我留下深刻印象。对于本篇论文，车老师也给予了很多有意义的重要建议和修改意见。当每次遇到困难求助到车老师，都能得到耐心细致的解答，正是在这样无私的帮助下，我才能顺利完成研究生论文的撰写工作。

另外，我还要感谢三年研究生学习期间授课的老师们，你们丰富的知识、灵活的教学手段都是我受益匪浅。

最后感谢我的家人和朋友，在你们的支持下，我才能顺利的完成研究生学习任务。