单位代码 10006

学 号

分类号

密 级 秘 密

****

毕业设计(论文)

学生成绩相关性分析与系统设计实现

|  |  |
| --- | --- |
| 院（系）名称 | 计算机学院 |
| 专业名称 | 计算机科学与技术 |
| 学生姓名 | 谢何涛 |
| 指导教师 | 傅翠娇 |

2016年6月

学生成绩相关性分析与系统设计实现 谢何涛 北京航空航天大学

北京航空航天大学

**本科生毕业设计（论文）任务书**

Ⅰ、毕业设计（论文）：

学生成绩相关性分析与系统设计实现

Ⅱ、毕业设计（论文）使用的原始资料（数据）及设计技术要求：

原始数据来源于傅翠娇老师记录的13级14级非计算机专业学生上计算机思维导论课程成绩

设计技术要求：使用WEKA进行相关性分析，使用Java进行系统设计实现

Ⅲ、毕业设计（论文）工作内容：

1．原始数据预处理

2．数据挖掘

3. 知识表示和结论

4. 系统需求分析

5. 系统设计实现

Ⅳ、主要参考资料：

[1] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J].Journal of Computer Research and Development, 2013, 50(1):146-169.

[2] 涂新莉，刘波，林伟伟. 大数据研究综述[J].Application Research of Computers, Vol.31 No.6 Jun. 2014.

[3] 韩家炜.数据挖掘：概念与技术[M]. 北京：机械工业出版社，2012：8-83

[4] 冯丽霞. 大学生学习成绩影响因素研究[J].China Electric Power Education, 2013, (13).

[5] guoxiang Liu, fengxia Yang. The Application of Data mining technology in college teaching[D], 沧州：沧州职业学院.2012.

[6] Jian Wang, Zhubing Lu, Weihua Wu, Yuzhou Li. The Application of Data Mining Technology Based on Teaching Information[D]. 重庆：西南大学, 2012.

[7] Chakrit Pong-inwong, Wararat Rungworawut. Teaching Evaluation Using Data Mining on Moodle LMS Forum[A].

[8] Azauri S. Figueira, Amanda S. Lino, Adriano Del Pino Lino. Educational Data Mining to Track Students Performance on Teaching Learning Environment LabSQL[A].

[9] 扈玉婷. 大数据背景下数据挖掘技术在大学英语听力教学效果分析中的应用[J]. 校园英语, 2015(3).

[10] 李建中，刘显敏. 大数据的一个重要方面：数据可用性[J].计算机研究与发展,2013,50-(6).

计算机 学院（系） 计算机科学与技术 专业类 120612 班

学生 谢何涛

毕业设计（论文）时间： 2016 年 1月 8 日至 2016 年 6 月 6 日

答辩时间： 2016 年 6 月 6 日

成 绩：

指导教师： 傅翠娇

兼职教师或答疑教师（并指出所负责部分）：

系（教研室） 主任（签字）：

注：任务书应该附在已完成的毕业设计（论文）的首页。

**本人声明**

我声明，本论文及其研究工作是由本人在导师指导下独立完成的，在完成论文时所利用的一切资料均已在参考文献中列出。

作者：谢何涛

签字：

时间：2016年 6 月

学生成绩相关性分析与系统设计实现

学 生：谢何涛

指导老师：傅翠娇

摘要

近年来随着高校不断扩招，学校规模迅速增大，同时计算机技术和网络技术的飞速发展促进了模式的改变，各高校纷纷利用计算机、校园网构建综合管理系统，提高了学校的综合管理效率。但目前，各高校管理系统仅提供查询、存储等功能，没有充分发挥对于海量数据的充分利用。针对这种现象，本文详细研究了通过得到的有关学生的各方面信息，利用数据挖掘工具，分析出相关数据中影响学生成绩的因素，得出一些初步结论，然后在得出结论的基础上，设计并实现学生成绩分析系统，为老师教学工作服务。本文主要研究的是北航的一门公共基础课程，计算机思维导论。通过对计算机思维导论的研究，希望能够发现一些潜在的影响因素，帮助老师改善课程。其他学科也可以参照本文的研究过程和采用的研究方法，进行分析研究。

**关键词：** 数据挖掘，计算机思维导论，成绩分析系统

The correlation analysis and system design and system implementation of student achievements

Author: Xie Hetao

Tutor: Fu Cuijiao

Abstract

With the enrollment expansion of universities and the increasing of school scale in recent years, at the same time, the rapid development of computer technology and network technology promoted the change of the pattern, universities are using the computer and campus net to build integrated management system, improving the comprehensive efficiency of management of the school. But the management system of universities only provides query function and storage function, the system do not make full use of massive data which is related to students. In order to change that phenomenon, this paper analyses the relation data to find the factors affecting the student achievements by analyzing the information relation to the students in detail. This paper has got some preliminary conclusions, and not only designed but also implemented the system which can serve teachers’ teaching job. This paper mainly studied the introduction to computer thinking which is a public course in Beihang university. By studying the introduction to computer thinking, this paper hopes to find some potential factors in order to help teachers improve this course. Other courses can also refer to the research process and the same analysis methods to be analyzed.

**Key words:** data mining, the introduction to computer thinking, the achievement analysis system

**目录**

[1 绪论 1](#_Toc452026635)

[1.1 研究背景与意义 1](#_Toc452026636)

[1.2 研究现状 2](#_Toc452026637)

[1.2.1 1.2.1国内研究现状 2](#_Toc452026638)

[1.2.2 1.2.2国外研究现状（675） 2](#_Toc452026639)

[1.3 研究目标： 4](#_Toc452026640)

[1.4 论文结构安排：（308） 4](#_Toc452026641)

[2 数据挖掘介绍 5](#_Toc452026642)

[2.1 数据挖掘技术 5](#_Toc452026643)

[2.2 数据挖掘工具 5](#_Toc452026644)

[2.3 数据挖掘算法 6](#_Toc452026645)

[2.3.1 2.3.1C4.5算法 6](#_Toc452026646)

[2.3.2 2.3.2 Apriori算法（1000） 10](#_Toc452026647)

[3 成绩相关性分析 12](#_Toc452026648)

[3.1 数据预处理（1000） 12](#_Toc452026649)

[3.1.1 3.1.1 数据清理 12](#_Toc452026650)

[3.1.2 3.1.2 数据集成 12](#_Toc452026651)

[3.1.3 3.1.3 数据选择 13](#_Toc452026652)

[3.1.4 3.1.4 数据变换 13](#_Toc452026653)

[3.2 数据挖掘（1000） 14](#_Toc452026654)

[3.3 知识表示及结论（1000） 20](#_Toc452026655)

[3.4 本章小结 25](#_Toc452026656)

[4 成绩分析系统的设计与实现 25](#_Toc452026657)

[4.1 系统需求分析（2000） 25](#_Toc452026658)

[4.1.1 4.1.1 功能需求 25](#_Toc452026659)

[4.1.2 4.1.2 数据需求 26](#_Toc452026660)

[4.1.3 4.1.3 技术需求 27](#_Toc452026661)

[4.2 系统设计与实现（2000） 27](#_Toc452026662)

[4.2.1 4.2.1 系统设计 27](#_Toc452026663)

[4.2.2 4.2.2 系统实现 30](#_Toc452026664)

[4.3 本章小结 35](#_Toc452026665)

[5 总结 35](#_Toc452026666)

[5.1 工作总结 35](#_Toc452026667)

[5.2 不足与改进 36](#_Toc452026668)

[致谢 37](#_Toc452026669)

[参考文献 37](#_Toc452026670)

# **绪论**

## **研究背景与意义**

近年来，计算机的快速发展，尤其是互联网+时代的来临，人们每天接触的数据规模不断增大，大数据也成为一个新兴的产业迅速发展起来。大数据指的是数据量过于庞大以至于超过了任何一个计算机处理能力的范围，大数据的典型特征是4V（Volume, Variety, Velocity, Value/Veracity），即规模性、多样性、高速性、价值性或者真实性[1]。目前大数据在商业智能、互联网、咨询与服务以及金融业、通信、医疗服务等行业显现，并产生了巨大的产业空间和社会价值 [2]。

大数据研究主要集中在如何进行大数据处理，分析以及管理技术和软件应用上。大数据处理主要包括数据清理，数据集成，数据选择，数据变换，数据挖掘，模式评估，知识表示这七个部分组成，数据挖掘是大数据处理中的一个重要环节[3]。

数据挖掘主要是通过分析大量数据，去除噪音数据，留下有用的数据，然后分析这些数据的特点，发现其中的一些规律。目前，数据挖掘的工具有很多，WEKA（Waikato Environment for Knowledge Analysis, 全名怀卡托智能分析环境），Rapid Miner，PMML（Predictive Model Markup Language）等。通过将数据挖掘技术应用到老师教学工作中， 可以方便老师更好的开展教学工作[4]。

随着近年来高校的不断扩招，学校规模迅速增大，同时计算机技术和网络技术的飞速发展促进了教学模式的改变，各高校纷纷利用校园网、计算机构建学生管理系统，提高了学校对学生进行综合管理的效率。但是，目前，各高校教务系统仅仅为学生和老师提供了查询、存储等功能，并没有为老师提供一个比较好的分析系统，对海量的学生数据进行分析，帮助老师进行教学工作，提升教学质量。通过数据挖掘技术，老师可以对学生进行全方位的分析，结合分析结果，改进自己的教学方式以及课程设置，帮助学校更好的管理，教学，帮助学生更好的提高学习效果和对知识的掌握。因此，本次研究希望通过对北航一门公共基础专业课的研究分析，能够设计出一个专门针对这门课程的分析系统。

本次研究的主要对象是北航一门公共基础课程，计算思维导论，通过老师提供的学生数据，首先利用WEKA软件进行分析，找出其中对该课程成绩有影响的因素，然后，设计出一个专门针对这门课程的学生成绩分析系统。通过导入学生数据，对学生成绩有影响的因素进行单独分析，并可视化处理，展现给老师，借此来帮助老师更好的了解影响学生学习的各种因素，方便老师做出教学上的调整。

## **研究现状**

### **1.2.1 国内研究现状**

国内对于影响学生成绩的研究取得过一些基础性的理论成果，北京联合大学讲师冯丽霞老师通过问卷调查的方式研究了北京某高校1~3年级大学生，发放问卷800多份，通过自编问卷来考察大学生学生成绩的影响因素，其中共30个题目，六个维度，包括学习动机，学习环境，学习兴趣，学习压力和困难，学习能力和学习态度等，采用5等级李克特量表（Likert scale），通过对回收的问卷进行统计发现，这六个维度都会影响学生成绩，但是，对于不同的学科，不同的维度的影响程度是不一样的[5]。沧州师范学院的两名学生提出利用数据挖掘技术对学生的个人信息（如年龄，性别，智力，认知能力等），考试结果以及出勤率的分析来了解学生的学习状况从而评估学生的各方面素质，为教师安排课程教学提供参考依据，并通过对各课程之间关联程度的研究来合理安排课程[6]。吉林大学的马丹硕士研究了利用数据挖掘算法设计模型来分析教育数据，预测学生考试是否通过，依次来提醒学生进行准备工作，但是没有涉及到分析学生的成绩信息，给老师提供一些课程上的建议[7]。西南大学应用技术学院研究了学生各个课程之间的关联程度，主要运用了先验关联算法，找出学生的课程学习中是否有某种先后的关系，希望通过尝试性的实验来为老师的课程安排工作能够提供一个科学的选择方法，但是，对于其他的关联算法，如决策树分析，聚类分析，统计分析，并没有做类似研究[8]。

总的来说，国内对于这方面的研究取得过不错的进展，对于本次学生成绩的分析以及系统的设计实现都提供了不错的思路。

### **1.2.2 国外研究现状（675 ）**

国外发达国家由于科技的发达程度以及计算机的普及程度要高于我国，因此对于数据挖掘以及数据挖掘在教育上的应用要比我国更广。对于这方面的研究也要更加细致一点。

泰国孔敬大学通过对LMS（Learning Managed System）论坛上的学生评价数据进行挖掘处理，通过比较ID3算法、Best First Tree、贝叶斯学习算法进行比较分析，发现ID3算法在分析学生评价里面的感情色彩有更高的正确率[9]。有的大学通过数据挖掘在远程学习课程中常见的存储大量的数据，详细记录互动开发学生的，帮助教师监管教学过程，基于获得数据来了解学生的学习环境，帮助老师改善教学过程中的纪律[10]。

## **研究目标：**

本文希望利用数据挖掘工具WEKA，首先分析学生成绩，找到影响学生成绩的因素，然后，设计并实现学生成绩分析系统。该系统主要分成三个部分，第一部分是对学生成绩的各个因素进行分析并可视化处理，便于老师了解教学情况，其中主要涉及到的算法是C4.5算法，还用到了统计分析技术。第二部分是对学生的评价信息进行提取，筛选出有用的信息，用条形图、折线图、饼图等形式将评价信息简化处理，展现给老师，方便老师调整教学计划，第三部分是帮助老师进行教学统计，计算工作，老师通过导入数据，设置相关参数，即可获得想要的学生最终总成绩，减少老师的工作量，让老师能有更多的时间来帮助和回答学生遇到的问题和困难。

## **论文结构安排**

本文结构主要分成以下五个部分

第一章为绪论，主要介绍研究背景及意义，国内外研究现状以及研究目标。这部分主要涉及的是前期的资料查找工作以及研究目标的确定，为后面的工作提供了理论支持和研究思路、方向。

第二章为相关技术介绍，主要详细介绍了数据挖掘技术，数据挖掘工具以及相关挖掘算法，为后面的分析和实现部分提供理论基础，方便读者了解这方面的一些理论知识。

第三章为成绩相关性分析，该部分详细讲述了如何利用WEKA进行数据挖掘工作以及通过分析所得到的一些结论，为后面的系统设计实现部分提供了理论支持和设计思路。

第四章为系统设计与实现，主要介绍了学生成绩分析系统的系统需求，系统的设计和实现。

第五部分为结论，主要是对本论文前面所有工作的总结以及系统所存在的缺陷的阐述和解决办法。

# **数据挖掘介绍**

## **数据挖掘技术**

数据挖掘是指从大量数据中发现潜在规律、提取出有用的知识。19世纪80年代数据挖掘技术开始崭露头角，到90年代末，数据挖掘技术已经得到了广泛的应用，以数据挖掘为核心的商业智能（BI）已经成为众多行业中的重要组成部分。目前数据挖掘技术在金融风险预测、质量分析、产品产量、基因工程研究、Internet站点访问模式发现等许多领域得到了成功的应用。

数据挖掘是知识发现过程的一个基本步骤，知识发现过程分为七个步骤，分别是数据清理，数据集成，数据选择，数据变换，数据挖掘，模式评估，知识表示。数据清理主要目的是填充缺省值，识别并消除异常值，纠正不一致数据。数据集成是指将分散的数据收集整理到一起去，便于统一的进行管理，分析等操作。数据选择是指从大量数据中选择出所要分析的数据，以便减少数据量。数据变换是指将数据变换成易于处理的形式，如将学生百分制的成绩变换成等级制，这样能够有效减少数据处理的时间以及发现一些更普遍的规律。模式评估根据某种兴趣度度量，识别提供知识的真正有趣的模式。知识表示是将得到的结论进行可视化处理，将结论以图像的形式呈现给用户，便于用户理解。

## **数据挖掘工具**

WEKA是一款基于JAVA的集合了大量能承担数据挖掘人物的机器学习算法，可通过API进行扩展的开源的数据挖掘软件，它是由新西兰的怀卡托大学开发出来的。WEKA包含四个部分，分别是Explorer, Experimenter, KnowledgeFlow, Simple CLI，Explorer是主要的功能部件，它包含数据预处理，分类决策树分析，聚类分析，关联规则分析，属性选择和可视化，通过Explorer部件能够完成对数据的挖掘处理工作，Experimenter部件主要作用是用来比较两种以上算法对于同一项数据挖掘工作的准确度和相关性，找出适合某项数据挖掘工作的最好的算法。KnowledgeFlow主要是为了简化数据挖掘过程，它包含一个面板，在该面板上添加部件，将各部件按顺序连接，构成一个程序流程图，直接导出文件后可以得到最终分析结果。Simple CLI的作用类似于控制台，通过输入命令来达到操作的目的，如果对于WEKA的指令比较熟悉的话可以通过Simple CLI进行操作。WEKA采用的数据格式为ARFF（Attribute-Relation File Format）文件，对于Excel文件可以先转换成CSV文件，然后通过WEKA转换成ARFF文件。

## **数据挖掘算法**

### **2.3.1 C4.5算法**

C4.5算法属于机器学习算法中的一种分类决策树算法，分类决策树算法是一种监督学习的算法，就是通过给定的一堆样本，每个样本都有一组属性和类属性（即要研究的属性），然后通过学习得到一个分类器，通过这个分类器可以对新出现但无法确定类属性的对象进行分类预测。分类决策树算法实际上是建立一个map的过程，这个map相当于一个预测模型，C4.5算法是其中比较经典的算法。

C4.5算法本质上不是一种算法，而是一系列的一套算法，C4.5算法构造分类树的步骤如下图：

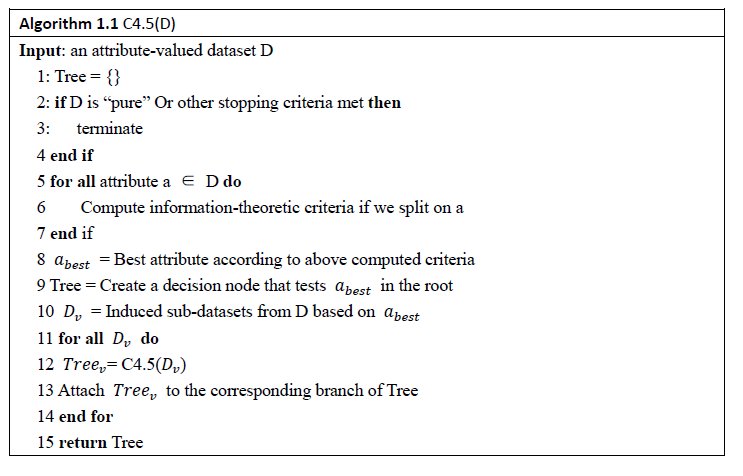


图 2.1

C4.5算法总体思路是从样本中所有属性中每次找出其中的最优属性进行分类，递归，直到分出最终结果或者满足某种条件为止。

下面通过一个数据案例介绍一下C4.5算法的执行过程。数据集见下图

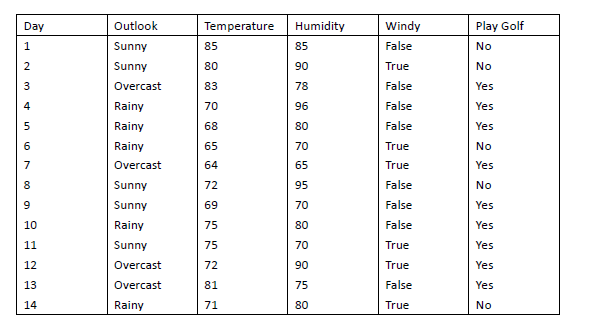


图 2.2

首先，计算类属性的信息熵， 信息熵反映的是一个系统有序化的程度，信息熵越高，系统越稳定，反之，则越混乱。类属性信息熵计算公式如下：

Info (D) =-

该组数据中类属性为Play Golf，该属性的信息熵为

Info (Play Golf) =-9/14\*-5/14\*=0.94

然后，分别计算得到其他属性的信息熵，其他属性信息熵的计算公式如下：

(D)=

Outlook属性的信息熵为

Info (Outlook) =5/14\*(-2/5\*-3/5\*) + 4/14\*(-)

+ 5/14\*(-2/5\*-3/5\*) = 0.694

其他属性的信息熵计算方法相同，信息熵依次为

Info (Temperature) =0.911

Info (Humidity) =0.789

Info (Windy) =0.892

计算完信息熵以后计算每个属性的信息增益值，信息增益值是ID3算法采用的进行决策分类的依据，信息增益值的计算公式为

Gain (A) = info(D)- (D)

分别计算得到每个属性的信息增益值

Gain (Outlook) = Info (Play Golf)-Info (Outlook) =0.94-0.694=0.246

Gain (Temperature) = Info (Play Golf)-Info (Temperature) =0.94-0.911=0.029

Gain (Humidity) = Info (Play Golf)-Info (Humidity) =0.94-0.789=0.151

Gain (Windy) = Info (Play Golf)-Info (Windy) =0.94-0.892=0.048

然后计算每个属性的分裂信息度量，分裂信息度量是用来衡量属性分裂数据的广度和均匀程度的，计算公式如下

SplitInformation (A) =-

分别计算得到每个属性的分裂信息度量

SplitInformation (Outlook) =-5/14\*-5/14\*-4/14\*=1.5774

SplitInformation (Temperature) =1.5567

SplitInformation (Humidity) =1.0

SplitInformation (Windy) =0.9852

最后计算出每个属性的信息增益率，ID3算法使用的是信息增益率，而C4.5算法使用的是信息增益率，选用信息增益率能够克服属性选择过程中因选择属性值多的属性而带来的问题。信息增益率的计算公式为

GainRatio (A) =

分别计算得到每个属性的信息增益率

GainRatio (Outlook) ==0.246/1.5774=0.156

GainRatio (Temperature) ==0.029/1.5567=0.018

GainRatio (Humidity) ==0.151/1.0=0.151

GainRatio (Windy) ==0.048/0.9852=0.048

比较属性的信息增益率的大小，选择信息增益率最大的属性进行第一次分类。然后采用递归的算法进行下一次分类，直到符合要求后停止分类。

确定决策树停止递归分类过程由两种方式，一种是当某节点的分支所包含的样本都属于同一类时，停止分类。另一种方式是当分支所包含的样本数量小于给定阈值时，停止分类，一般采用第二种情况，此时分支的类别为分支中样本数占大多数的类别。

由此，C4.5算法生成了决策树，但此时的决策树不是最终的决策树，此时生成的决策树的树叶节点所包含的样本基本上属于同一个类别，这种决策树看起来十分完美，但是仅限于对于这组测试用例，如果测试用例里包含了一些错误的数据，那么这棵树也错误的接受了，这种现象叫做“过拟合”现象，过拟合的决策没有普适性，可是对于测试样例很合适，但是对于其他的实例就不适用了，因此，需要在此基础上进行决策树剪枝，得到简化的决策树。决策树剪枝有两种方法，一种方法是用一个额外的测试数据集来检验这个决策树，尝试着用将非叶子节点替换成叶子节点，然后比较两者的准确程度，如果替换后的准确程度高就采用替换后的结果，用bottom-up的方式遍历所有子节点，直到所有的子树都有最优的正确率为止。第二种方法是采用悲观剪枝的思路，依据测试用例中的样本误差率作为是否剪枝依据，误差率为该分类中正确的样本数除以总样本数，一般情况下用叶子节点代替分支节点误差率肯定会增大，因此为了达到剪枝的目的，要在给子树的误判计算加上一个惩罚因子，这样，分支和叶子节点就能在公平的条件下进行比较，在剪枝后判断子树的误差率是否在一个给定的范围内，如果在，就表明剪枝成功。否则，不能剪枝，然后依次对其他子树进行悲观剪枝，完成决策树的剪枝工作。一般情况下，由于第一种方法需要额外的测试用例集，因此，大多数情况下采用悲观剪枝的方法进行决策树剪枝。

上述数据最终经过剪枝后生成的决策树如下图。



图 2.3

### **2.3.2 Apriori算法（1000）**

Apriori算法即为先验关联规则算法，该算法也是一种数据挖掘算法，用于发现大量数据中各属性之间存在的规律关系。Apriori算法中主要的参数有两个，一个是置信度，一个是支持度，设A={}是一个项集，T={}是事务T所包含的A中元素的集合，S={}是事务S中所包含A中的元素的集合，则

support(T=>S)=P(T∪S)

confidence (T=>S)=P(S|T)

其中P(T∪S)表示事件T和事件S同时发生的概率，P(S|T)表示在事件T发生的情况下事件S发生的概率，先验关联规则以支持度和置信度为选择标准，同时满足最小支持度和最小置信度的规则称为强规则。

Apriori算法的挖掘过程分两步：

找出所有频繁项集，频繁项集必须大于等于最小支持度；

由频繁项集产生强关联规则，强关联规则中的每个项集都必须满足最小支持度，并且每条规则要满足最小置信度。

下面通过一个案例来说明一下Apriori算法的执行过程，数据如下图：

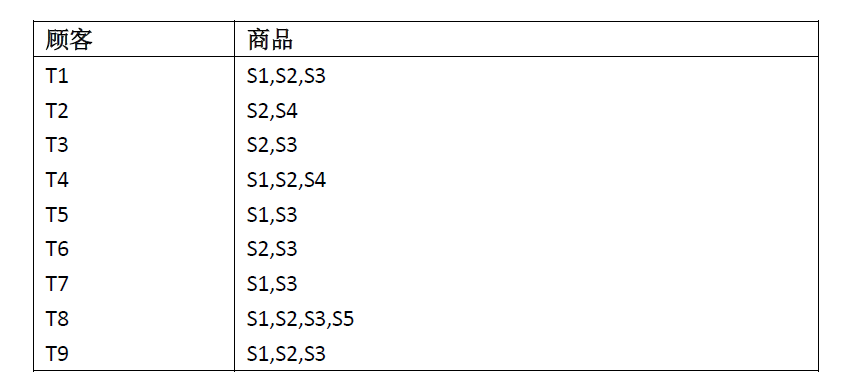
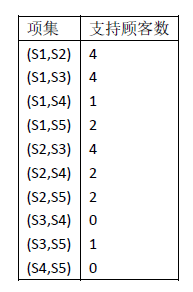
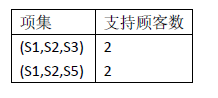


图 2.4

假设最小顾客支持度为2，min\_support=2/9=22%。

首先找出1-项频繁项集合L1，L1={S1,S2,S3,S4,S5},由于单项没有任何规律，因此，继续查找2-项频繁项集。

查找2-项频繁项集前，先由L1产生候选2-项频繁项集C2，C2如图2.4所示。

C2L2C3(L3)

图2.4

由于最小支持度为2，所以C2中支持数小于2的都不符合要求，因此除去支持数小于等于2的项集，得到2-项频繁集L2。

由2-项频繁项集继续产生候选3-项集，依次类推，得到3-项频繁项集，在由3-项频繁项集产生4-项频繁项集过程中，由于(S2,S3,S5)不是3-项频繁向集，所以不存在4-项频繁项集。至此，所有符合要求的频繁项集全部产生。

假设最小置信度为0.7，则符合要求的频繁项集置信度如下：

(S1,S2,S5) S1∩S2=>S5 confidence=2/4=50%

S1∩S5=>S2 confidence=2/2=100%

S2∩S5=>S1 confidence=2/2=100%

S1=>S2∩S5 confidence=2/6=33%

S2=>S1∩S5 confidence=2/7=29%

S5=>S1∩S2 confidence=2/2=100%

其中满足最小置信度的规则为2,3,6，即同时购买商品S1和S5的顾客100%购买S2，同时购买S2和S5的顾客购买S1的概率为100%，购买S5的顾客100%同时购买S1和S2。

# **成绩相关性分析**

## **数据预处理（1000）**

本文研究的对象是北航计算机思维导论这门课程的学生成绩，计算思维导论是一门面向大学一年级非计算机专业学生的课程，针对没有或者很少有编程经验的学生，课程选择简单易学而又功能强大的Pathon语言为载体，使学生建立程序设计的信心。该课程包含44个学时，其中理论授课26学时，实验18小时，该课程是一门理论与实验相结合的课程。因此，本文所涉及的数据来源于计算思维导论课程相关数据，主要由傅翠娇老师提供，原始数据如下图所示。

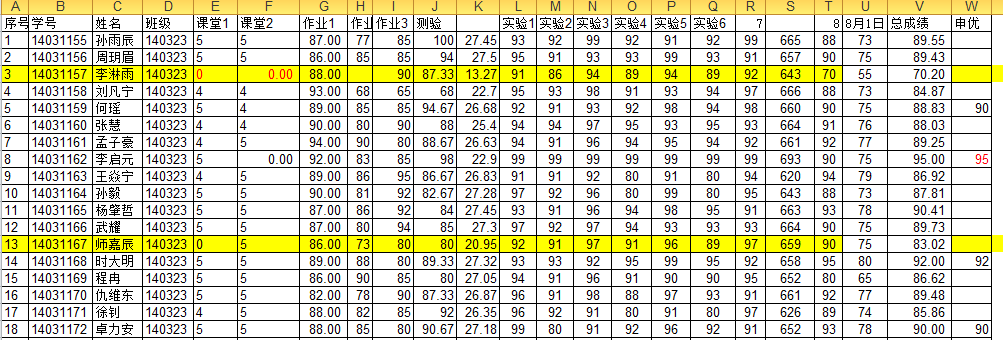


图3.1

数据主要包括学生学号，姓名，班级，课堂成绩，作业成绩，实验成绩，测验成绩以及总成绩，数据量约为800个。

### **3.1.1 数据清理**

原始数据中存在少量缺省数值的数据，对于缺省值的数据，采用的办法通常有忽略该元组、人工填写缺省值、使用属性平均值代替等，在本文中采取的办法是忽略该元组，这样可以减少不确定缺省值所带来的分析误差。

除了缺省值数据外，还有一些数据是老师用于统计最终成绩而添加的数据，如上图中的没有属性名称的数据，这些数据属于噪音数据，要将其清理掉。

### **3.1.2 数据集成**

本次分析中原始数据来自多个不同的Excel 表格，并且数据的属性不完全相同，因此，在进行数据分析前需要将数据集成到少数几个表格中，在所收集到的数据中，一部分数据主要包含平时成绩，期中成绩，期末成绩，实验成绩和总成绩，一部分数据包含学生专业，所在班级和期末成绩，还有一部分包含学生所在班级，课堂成绩，作业成绩，实验成绩，测验成绩以及总成绩，据此，可以将收集到的数据集成到三张Excel表格当中去，根据要研究的不同方面选着不同的数据。

### **3.1.3 数据选择**

原始数据中包含学生的学号、姓名、所在班级、所学专业、大作业、每次实验成绩，作业成绩，测验成绩等信息，包含的信息比较全面，但是，在分析研究的过程中并不是所有的信息都能够用的到。在所有分析中，学生的学号，姓名对于所要研究的内容均没有帮助，因此可以将此类信息删除。在分析学生所在院系对学生成绩的影响时，除了所学专业是有用的信息外，其他属性都属于无用信息，都要予以删除。综合分析中，学生的课堂成绩，测验成绩，实验成绩以及作业成绩都是相关因素，所以需要对这些信息进行保留，对于班级，专业等信息都要予以删除。

### **3.1.4 数据变换**

本文采用的数据挖掘工具是WEKA开源软件，由于WEKA不支持Excel格式的数据，因此需要将Excel表格进行转换，先转换成CSV文件，在通过WEKA转换成ARFF文件，ARFF文件是最适合WEKA进行分析的文件格式，如下图所示

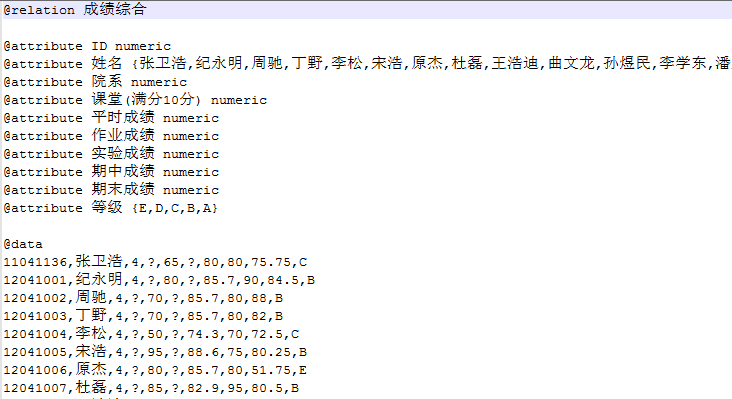


图 3.2

WEKA支持四种数据类型，分别是numeric（数值型），nominal（分类型），string（字符串型），data（日期和时间型）。Excel文件转换成ARFF文件时，数据属性默认为numeric属性，而在进行先验关联规则分析时，WEKA只支持对nominal型的数据进行分析，因此在分析之间要进行数据属性的转换，将numeric型转换成nominal型。在对学生成绩进行分析时，由于学生成绩采用的百分制计算的，在进行分析时可能不会得出十分普遍的规律，因此在分析前对学生成绩进行了处理，将学生成绩分成5个等级，分别用A,B,C,D,E表示，代表优（>=90），良（>=80），中(>=70)，及格(>=60)，差(<60)。

综合分析时，学生课堂成绩有两次成绩，如果每次成绩都要考虑进来的话，将不会得出任何规律，因此需要对数据进行变换处理，两次课堂成绩选择用平均数的形式进行代替，而课堂成绩采用的是10分制。学生实验成绩共有8次成绩，进行综合分析时，实验成绩仅作为一项考虑，因此需要将8次实验成绩进行加权平均，得出一个最终的总的实验成绩。为了方便分析，百分制的实验成绩同样采用5级形式表示。最终得到了一个简化的易于分析的数据集，如下图所示：



图3.3

## **数据挖掘**

数据处理完成后，就可以利用WEKA进行数据挖掘的工作了，在WEKA上进行的分析工作主要是综合分析和各单项因素的分析。

WEKA中的比较常用的算法有C4.5算法，朴素贝叶斯算法，Apriori算法，聚类分析算法以及线性回归分析。综合分析过程中，考虑过使用朴素贝叶斯算法进行数据挖掘，但是对比C4.5算法和朴素贝叶斯算法后，发现C4.5算法的预测正确率要高于朴素贝叶斯算法，因此使用了C4.5算法。下图是使用两种算法后的对比。

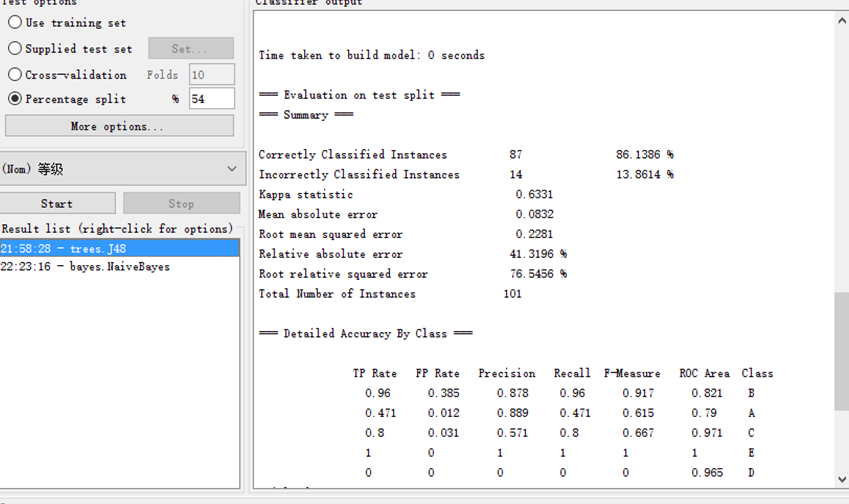


图3.4

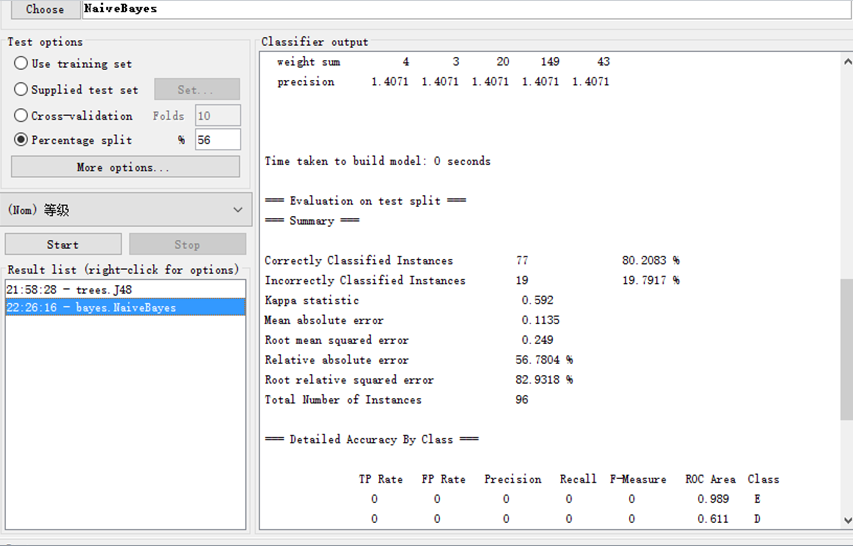


图3.6

从上图可以看出，朴素贝叶斯算法的预测正确率为80.2%，而C4.5算法的预测正确率是86.1%。平均绝对误差（Mean absolute error），均方根误差（Root mean squared error）这些参数中，C4.5都要比朴素贝叶斯算法低。平均绝对误差反应的而是预测值误差的实际情况，越小表示预测越准确，均方根误差是用来衡量观测值与真实值之间的偏差，越小表示偏差越小。下图是C4.5算法分析后的结果：

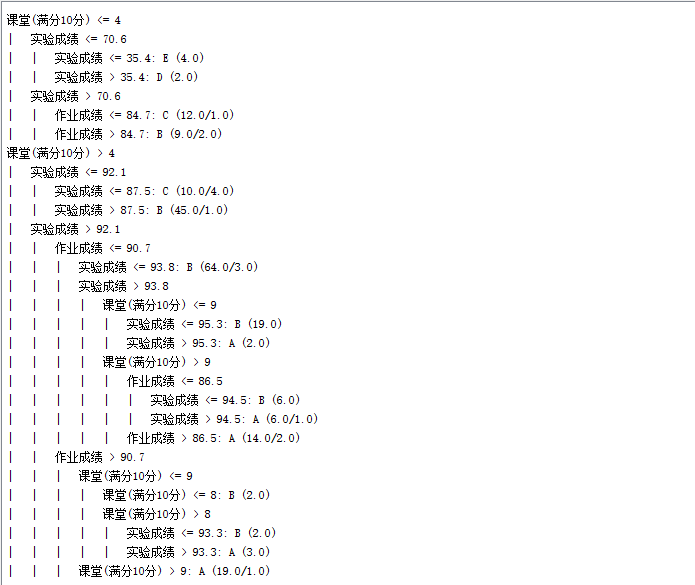


图3.7

这是一棵分类决策树，这棵决策树最先进行分类的属性是课堂成绩，首先将学生课堂成绩按照小于等于4分和大于4分分成两个分支，在小于等于4分的分支里，又按照实验成绩进行分类，分成小于等于70.6分和大于70.6分两个分支，在实验成绩小于等于70.6分时，又对实验成绩进行分支，分支线是35.4分，当实验成绩小于等于35.4分时预测学生成绩为E，其中测试样本数是4个，4个全部预测正确，当实验成绩大于35.4分时，预测学生成绩为D，其中测试样本数是2个，2个全部预测正确。课堂成绩小于等于4分，实验成绩大于70.6分时，分类树又按照作业成绩进行分类，分为小于等于84.7和大于84.7两个叶子，在小于等于84.7时，有13个测试样本数，其中12个预测正确，1个错误，在大于84.7的叶子中，有11个测试样本，9个预测正确，2个预测错误。

当课堂成绩大于4分时，按照实验成绩进行分支，以92.1分为分界线，实验成绩小于等于92.1分时，又按照实验成绩分支，分界线为87.5分，小于等于87.5分时，共14个测试样本，10个预测正确，大于87.5分时，共46个测试样本，45个正确。当实验成绩大于92.1分时，按作业成绩进行分类，分界线为90.7分，小于等于90.7分时，按照实验成绩分成两类，分界线为93.8，小于等于93.8时，测试样本用67个，64个预测正确。大于93.8分时，又依次按照课堂成绩，实验成绩，作业成绩进行分类。课堂成绩大于4实验成绩大于92.1作业成绩大于90.7时，按照课堂成绩，实验成绩依次分类，最终得到上述的决策树，上述决策树的叶节点包含的样本数最少为2。

最终得到的决策树中叶子节点有16个，树大小即树中节点数为31。

综合分析完成后，又进行了单个属性的分析

首先分析了学生课堂表现与总成绩的关系，使用Apriori算法进行分析，得出如下图所示结果

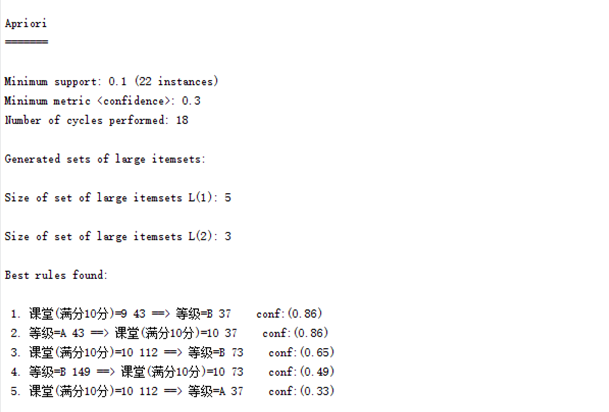


图3.8

由上图可知，符合最小支持度0.1的1-项频繁集有5个，2-项频繁集为3个，其中满足最小置信度0.3的强规则有5个，分别如下

课堂（满分10分）=9 => 等级（学生成绩）=B confidence=37/43=0.86

等级（学生成绩）=A => 课堂（满分10分）=10 confidence=37/43=0.86

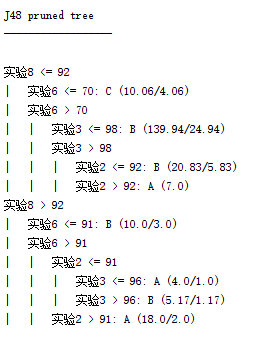
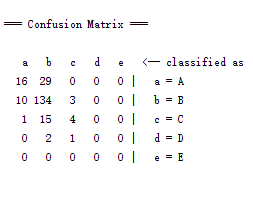
课堂（满分10分）=10 => 等级（学生成绩）=B confidence=73/112=0.65

等级（学生成绩）=B => 课堂（满分10分）=10 confidence=73/149=0.49

课堂（满分10分）=10 => 等级（学生成绩）=A confidence=37/112=0.33

其中对于本次分析有帮助的强规则为第一个和第三个，课堂成绩为9分的学生中有86%的学生成绩会取得B，课堂成绩为10分的学生中有65的学生成绩会取得A。其他的要么置信度过低，不具备参考价值，要么是通过总成绩预测课堂成绩，不是本文分析的目的，所以舍弃。

然后对学生实验成绩进行了分析，学生总共有8次实验成绩，采用的分析算法是C4.5算法。开始分析时，默认叶子节点实例数最小值为2，得到的分类决策树过于繁琐，有20个叶子，39个节点，并且预测正确率比较低，后来，经过反复测试和修改，发现当叶子中实例数为4时，得到的分类决策树包含8个叶子，共15个节点，而且预测正确率有了明显的提高，得到的分析结果如下图3.2（5）a所示：

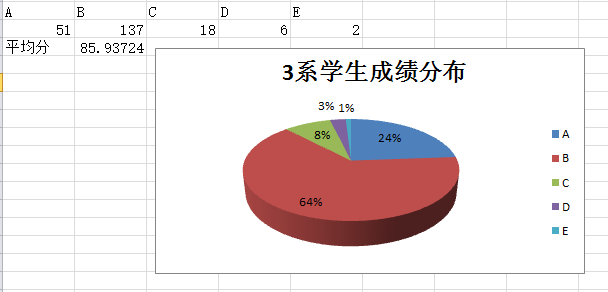
ab

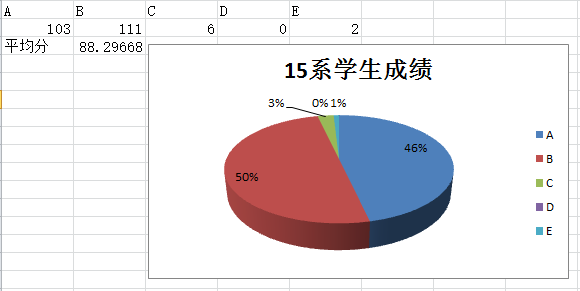
图3.9

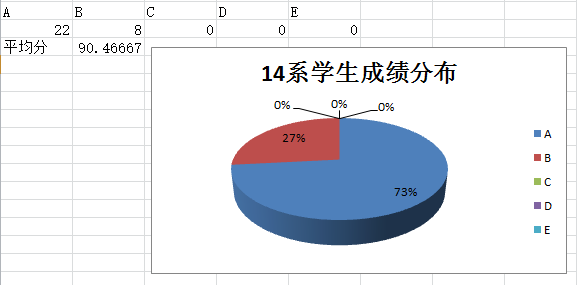
生成的决策树首先按实验8的成绩进行分类，以92分为分界线，小于等于92分时，按照实验六进行分类，分界线为70分，实验6成绩小于70分时，预测结果为C，实验6成绩大于70分时，按实验3成绩进行分支，以98分为分界线，小于98分时，预测结果为B，实验成大于98时，按照实验2的成绩进行分支，分界线为92分，，小于92分时，预测为B，大于92分时，预测为A。实验8成绩大于92的分支中，以实验6成绩为分支，分界线为91，小于91时，预测成绩为B，大于91分时，以实验2成绩为分支，分界线是91分，大于91分时，预测成绩为A，小于91分时，按实验3进行分支，分界线为96分，大于96分时，预测成绩为B，小于96分时，预测结果为A。

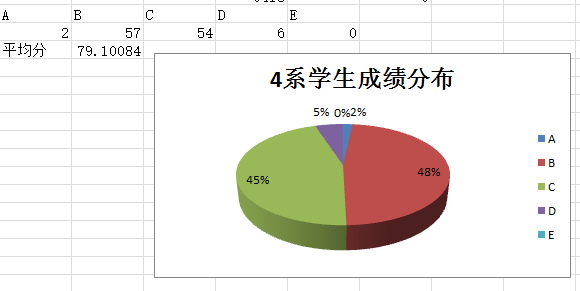
上图3.2（5）b为详细结果的2阶矩阵，x轴（classified as）表示预测结果，y轴表示实际结果，其中预测结果为A的实际结果为16个为A，10个为B，1个为C，预测结果为B的实际结果为29个A，134个B，15个C，2个D，预测结果为C的实际结果为3个C，4个C，1个D，D和E的预测结果和实际结果都为0.

最后分析了一下学生所在院系是否对学生成绩有影响，分析学生院系影响时，采用C4.5算法进行分析时，由于各院系学生成绩都有五个不同的等级，所以并不适合分析，同样Apriori算法，线性回归等WEKA里面的数据挖掘算法都不适合对此进行分析，综合考虑以后，决定采用数学上的统计分析方法，将统计的结果用条形图，折线图，饼图等形式展现出来。下图是统计分析的结果：









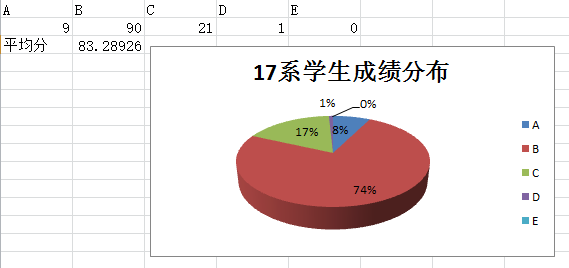


图3.10

上图中，3系学生平均成绩为85.9分，成绩为A的占24%，成绩为B的占64%，成绩为E的占1%，15系学生平均成绩为88.3分，成绩为A的占46%，成绩为B的占50%，成绩为E的占1%，14系学生平均成绩为90.5分，成绩为A的占73%，成绩为B的占27%，成绩为E的占0，4系学生平均成绩为79.1分，成绩为A的占2%，成绩为B的占48%，成绩为E的占0，17系学生平均成绩为83.3分，成绩为A的占8%，成绩为B的占74%，成绩为E的占0.

## **知识表示及结论（1000）**

数据挖掘工作结束后，需要对数据进行知识表示，知识表示是使用可视化和知识表示技术，向用户提供挖掘到的信息，方便用户分析得出结论。

综合分析得到的结果用可视化的决策树表示如下图

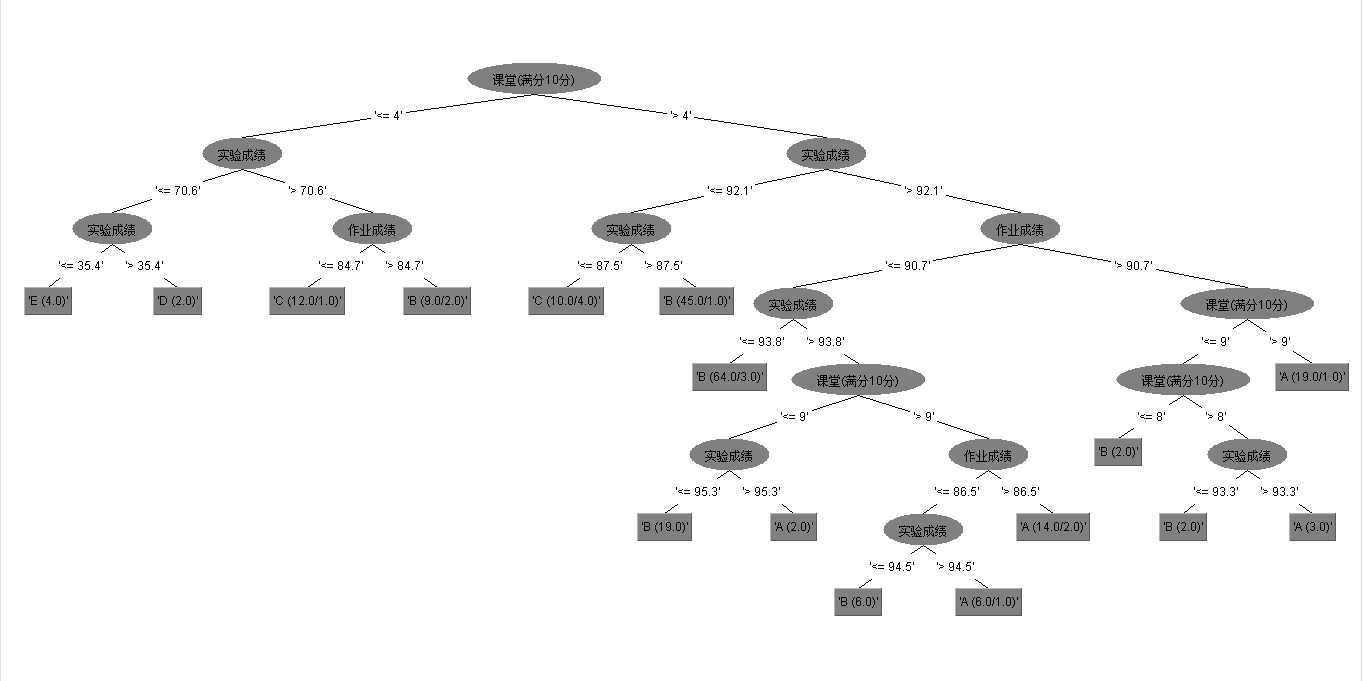


图3.11决策树图

上述决策树是上节综合分析中采用C4.5算法挖掘出的决策树的可视化图形表示，相比较文字而言，图形更能使我们发现一些明显的规律。C4.5算法进行决策树分支的依据是信息增益率，信息增益率反映的是该属性能反映类属性结果的程度，信息增益率越高，说明该属性越能反映出类属性，从另一方面来说，说明该属性与类属性的相关程度越高。从上图可以看出，课堂，实验成绩以及作业成绩对于学生成绩的影响程度依次递减。

课堂成绩分析过程中，将分析结果可视化如下图所示：

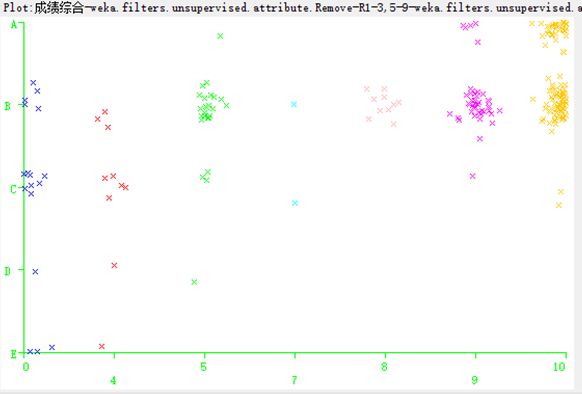


图3.12课堂成绩

该可视化图形为散点图，横坐标是学生的课堂成绩，纵坐标是学生的总成绩等级，从散点图中可以看出，学生课堂成绩为9分和10分的学生中99%的学生总成绩为A和B，总成绩等级为E的学生课堂成绩都低于5分，从散点图中可以看出，学生课堂成绩和总成绩基本成线性相关关系，课堂成绩为0分的学生成绩大多在C（包含C）以下，低于8分的学生成绩大多在B（包含B）以下，课堂成绩在8分以上的，学生成绩大多在B以上。

实验成绩的决策树可视化图形表示如下图：

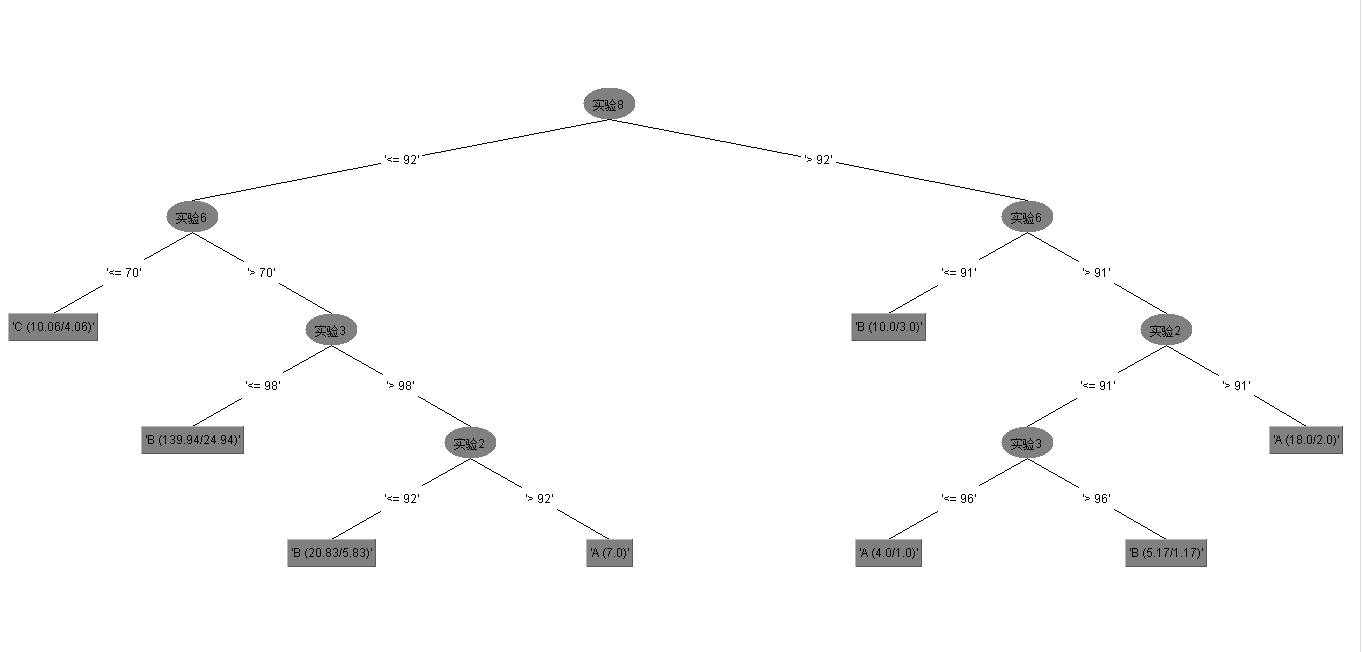
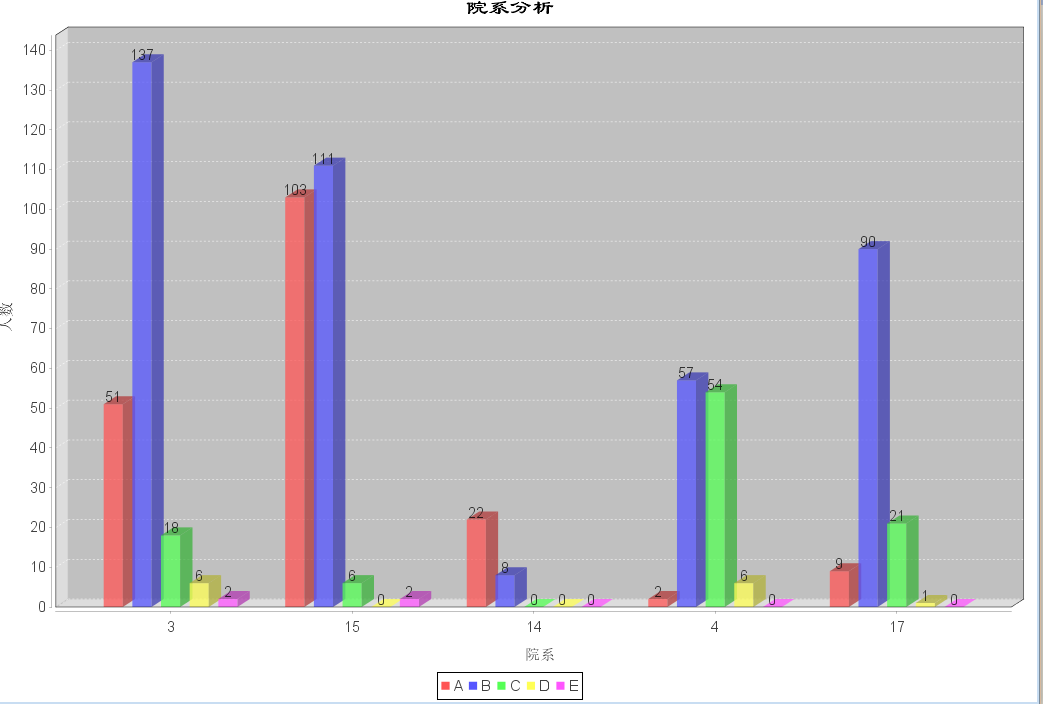


图3.13实验成绩

上图反映了学生8次实验成绩与总成绩的相关程度，从图中可以看出，实验8对学生成绩的影响最大，实验6其次，实验3和实验2影响较小，而其他的实验1，实验4，实验5，实验7与学生总成绩的相关程度最小。其中，由于实验8成绩在总成绩占比比较大，因此影响最大与此有部分关系，但是其他七次实验所占的比重都是相同的，不过与成绩的相关程度却不一样，因此，可以认为，实验8，实验6，实验3，实验2更能反映出学生对于这门课程的掌握程度，实验1，实验4，实验5，实验7可能在内容设计上不太合理，例如，在分析过程中发现，实验1学生的成绩全部在90分以上，说明实验1的难度设计不合理，在下次上机实验时，老师可以根据分析结果相应的调整一下实验内容，尽量使每次实验课程都能够与教学内容更紧密的联系在一起。

院系分析使用的是数理统计分析方法，在数据挖掘过程中，展示的是单个院系的学生成绩分布情况，这样对比起来的效果不是很明显，在知识表示阶段，本文采用了将学生院系信息放在一张图表上表示出来，得到的效果十分明显，如下图所示：



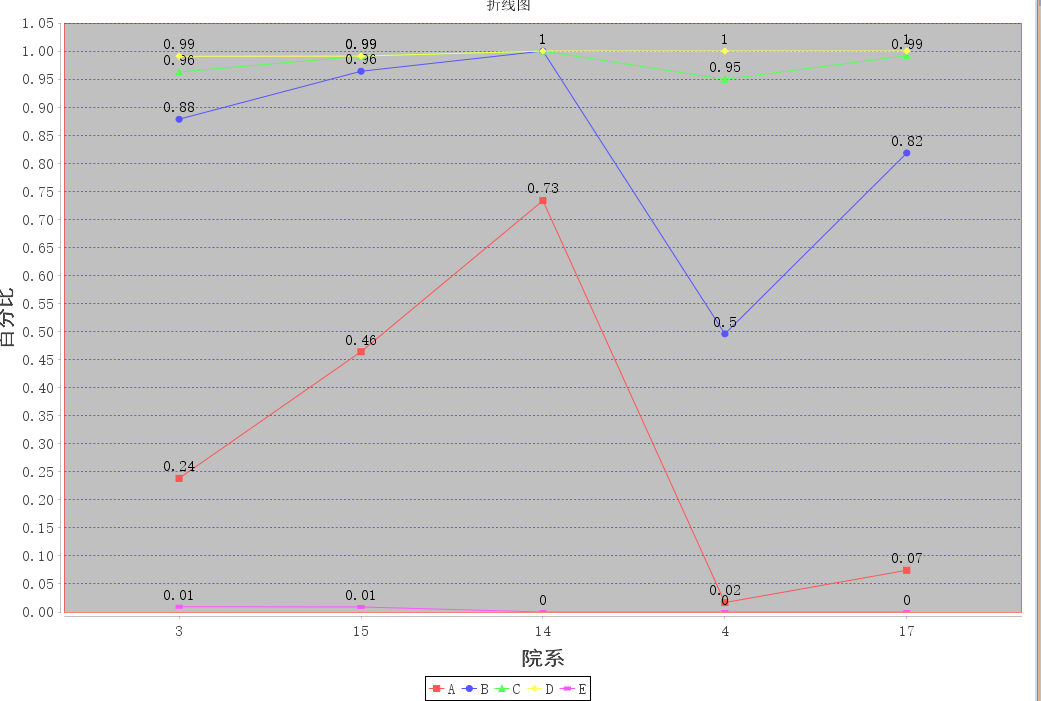


图3.14院系分析

上述折线图中，红色表示学生成绩为A的人数占该院系总人数百分比，蓝色表示学生成绩为A和B的人数占该院系总人数的百分比，绿色表示学生成绩为A,B和C的人数占该院系总人数的百分比，黄色表示学生成绩为A，B，C和D的人数占该院系总人数的百分比，紫色表示学生成绩为E的人数占该院系总人数的百分比，从折线图可以看出，14系学生成绩在所有系中是最好的，4系学生成绩在所有系中是最差的。从条形图可以看出，4系，14系，17系学生成绩没有不及格现象，虽然14系样本数较少，但是14系学生成绩全部在B以上，4系学生成绩为A的人数很少，这些现象都表明学生所在院系对于学生成绩是有影响的。虽然这门课程是面向大一所有非计算机专业的学生进行授课的，但是不同院系学生所学到的基础课程不一样，所掌握的知识面也不一样，对于计算思维导论这门课程，每个系的学生的学习能力也不一样。因此，老师在开设这门课程时，可以考虑不同的院系开始不同的课程难度，上机实验和作业的难度都可以做出相应调整。

综合以上分析，本文可以得出以下几方面的结论:

从大的方面讲，直接影响学生总成绩的因素有学生的课堂表现，学生的作业成绩和学生的实验成绩，其中学生的课堂成绩与学生成绩的相关程度最强，其次是学生的实验成绩，最后是学生的作业成绩。

学生课堂成绩与学生总成绩基本呈线性相关关系，学生的课堂表现越好，学生的总成绩也就越好，学生课堂成绩在9分以上的学生99%的总成绩都是B以上，即80分以上。

学生的上机实验中，比较重要的实验是实验8，实验6，实验2和实验3，这几次实验设计的难度比较合理，内容比较科学，与学生总成绩有很强的联系，而实验1,实验4，实验5，实验7可能在实验内容方面设计的不够合理，尤其是实验1在难度上设计的不科学，学生成绩普遍在90分以上。

间接影响学生成绩的可能因素有学生所在院系，学生性别，学生的家庭背景等等其他的一些因素，本文研究中由于缺乏相关的数据，因此只研究了学生所在院系对学生成绩的影响。分析发现，学生所在院系对学生成绩有比较明显的影响，14系学生在学习这门课程时，能够很快理解掌握老师所讲的知识。4系学生对于这门课程的学习能力就相对要差一点，因此，老师在教授不同院系的学生时，可以稍微改变下不同学院学生的课程难度。例如，将14系学生的课程难度提高一点，将4系学生的课程难度降低一点。由于每个学院的学生在学习这门课程的时候教授方法，课程要求都是一样的，唯一造成存在如此明显的差异的情况就是学生所在学院所学的基础课程不同的缘故。

## **本章小结**

本章详细介绍了分析学生成绩影响因素方面的几个重要过程，主要包括数据预处理，数据，数据挖掘，知识表示及结论，在数据预处理过程中，主要介绍了数据清理，数据集成，数据变换几个方面，在数据挖掘部分，主要介绍了对影响学生成绩因素从几个不同的方面进行挖掘分析的过程。知识表示及结论部分，通过将挖掘到的信息用可视化的方式展现给用户，方便用户对结果进行分析，得出一些普遍性的结论。

本章主要利用WEKA这款开源软件研究了影响计算机思维导论这门课程学生成绩的因素，得出了一些初步的结论，本章的分析主要是为了下一章的学生成绩分析系统服务的，希望通过研究所发现的一些规律对于下一章成绩系统的设计能够有所帮助。在本章的分析研究中发现，C4.5算法对于分析多种因素对学生成绩造成影响的情况时准确率比较高，对于分析单个因素对学生总成绩的影响时，可以考虑使用数理统计方面的知识。

# **成绩分析系统的设计与实现**

## **系统需求分析（2000）**

计算机技术的快速发展和应用使得人们的生活变得十分的便捷，随着高校的扩招和计算机技术的应用，现有高校都建设有学生成绩管理系统，负责统计学生的各种个人信息，学生成绩，学生可以通过学校教务系统进行选课，老师评价，成绩查询，老师通过教务系统可以给学生布置作业，评价学生作业等，但是教务系统对于其中大量的学生成绩数据并没有进一步的进行研究分析，造成了大量有用数据资源的浪费。为了改善这种现象，本文希望能够设计一种学生成绩分析系统，这个系统主要用于北航计算机思维导论这门课程的学生成绩分析处理工作，帮助老师进行教学方面的改进工作，指出老师教学工作中存在的不足，提高学生的成绩。

### **4.1.1 功能需求**

成绩分析系统需要完成帮助老师统计学生成绩信息，对学生成绩进行详细分析，提取出学生评价中的有用信息等三个主要功能。

第一个功能是帮助老师统计学生成绩信息，计算机思维导论是一门理论与上机实验相结合的一门课程，这门课程的总成绩是由学生的课堂成绩，作业成绩，实验成绩以及考试成绩乘以一定的比例得到的，因此老师在得到学生的这些数据以后还要进行一些处理才能得到最终的总成绩，为了减少老师的工作量，希望在学生成绩分析系统中能够帮助老师实现总成绩的统计工作，老师需要导入要统计的Excel 表格，然后输入各属性的比例参数，就能够得到最终的填入了总成绩信息的学生成绩表。

第二个功能是对学生的成绩进行分析，这个部分包括三个小功能，第一个是利用C4.5算法对学生成绩进行全面的分析工作，最后生成一个决策树。第二个小功能是对影响学生成绩的单项因素进行数理统计分析，包含院系影响和实验影响。第三个小功能是对老师想了解的其他属性进行分析研究，用老师希望的可视化方式展现出来。

第三个功能是从学生的评价中挖掘出对课程有用的改进建议，给老师提供一个参考的价值。该功能是在分析学生评价中的信息对学生成绩是否有影响的时候产生的一个评价信息统计工具，虽然没能从评价中发现对学生成绩有影响的因素，但是在统计过程中发现学生对课程提出了许多有用的建议，因此决定将这部分功能加到系统中，希望老师能够清楚的看到大多数学生对于课程改进的意见，帮助老师对课程做出调整做参考的作用。

### **4.1.2 数据需求**

成绩分析系统的数据来源于老师所收集到的数据，数据用Excel表格的形式进行处理，

第一个功能模块中所包含的数据应该包括所有与这门课程总成绩相关的学生的其他方面的成绩，在Excel表格中如果出现了公式，系统将无法识别这些数据，因此需要将表格中的公式全部转化成数值形式。并且，系统对于缺省值的数据无法处理，对于一些没有实验成绩或者作业成绩的学生信息希望能够自动用0代替。

第二个功能模块中第一部分所包含的信息应该是所有老师想要分析的与学生总成绩相关的因素，而对于一些不必要的噪音数据应该予以删除，以免对分析结果造成影响第第二部分所包含的信息中只要有要分析的属性信息就可以了，没用的信息不会对分析结果造成影响，所以可以不用理会，但是缺省值数据一定要补上0。第三部分的数据需求跟第二部分的要求一样，只要包含所要分析的属性和总成绩即可，缺省值补上0。第二个功能模块的数据格式也是Excel表格文件。

第三个功能模块中学生评价信息是以Word文档的形式读取的，里面包含了学生对于这门课程的评价信息。评价信息要求内容真实，评价可以为空，并且评价信息是中文形式。

### **4.1.3 技术需求**

系统的编程语言是Java，使用Java语言地主要原因是考虑到了Java具有很强的可移植性，能够在各种操作系统上运行，并且产生同样的执行结果。而且Java语言是一种大家所熟悉的语言，能够引用各种软件包，编程和调试都比较方便。

系统的功能实现部分用到了C4.5算法，即分类决策树分析算法，在实现C4.5算法的过程中，需要对连续性的数据和离散型的数据进行分开处理，对于离散型的数据，进行分支处理时之间按照样本属性本身的取值个数进行计算。对于连续型的属性Am，如果某节点该分支样本数量为Ms，则首先将该分支样本按数值大小依次排列，然后按照离散属性对样本进行分支，共有Ms-1种分支方法，对每种分支方法按照离散属性计算信息增益率，然后比较信息增益率大小，找到增益率最大的分支，则该连续型属性按照最佳信息增益率的分支进行分支。

系统的数理统计可视化展示部分用到的Java里包含的一个开放的图表绘制类库JFreeChart，使用JFreeChart可以生成柱状图，折线图，饼图，散点图，时序图等各种图表。

系统的数据录入部分引入了Apache POI 开放源码函数库，该函数库支持Java对Microsoft Office各种格式的文档进行读写操作，其中HSSF提供读写Excel表格的功能，HWPF提供读写Word文档的功能。

系统的界面开发使用的Swing开发工具包，它以抽象窗口工具包（AWT）为基础是跨平台应用程序可以使用任何可插拔的外观风格。Swing主要包含三个基本构造块，标签，按钮和文本字段。

## **系统设计与实现（2000）**

### **4.2.1 系统设计**

成绩统计模块设计只包含一个部分，打开成绩统计模块部分界面左上方显示一些关于成绩统计模块操作的说明，界面中央是输入文件的文本框和一些按钮，按钮主要是文件查找按钮，参数输入按钮，以及统计按钮，参数输入按钮是用来输入与总成绩直接相关的学生各项成绩的系数，点击统计按钮以后系统直接通过公式计算将信息填入导入的Excel表格当中。

成绩分析模块的设计分成三个部分，分别是综合分析，单因素分析，其他因素分析。

综合分析部分使用的是C4.5算法，界面如下图所示：

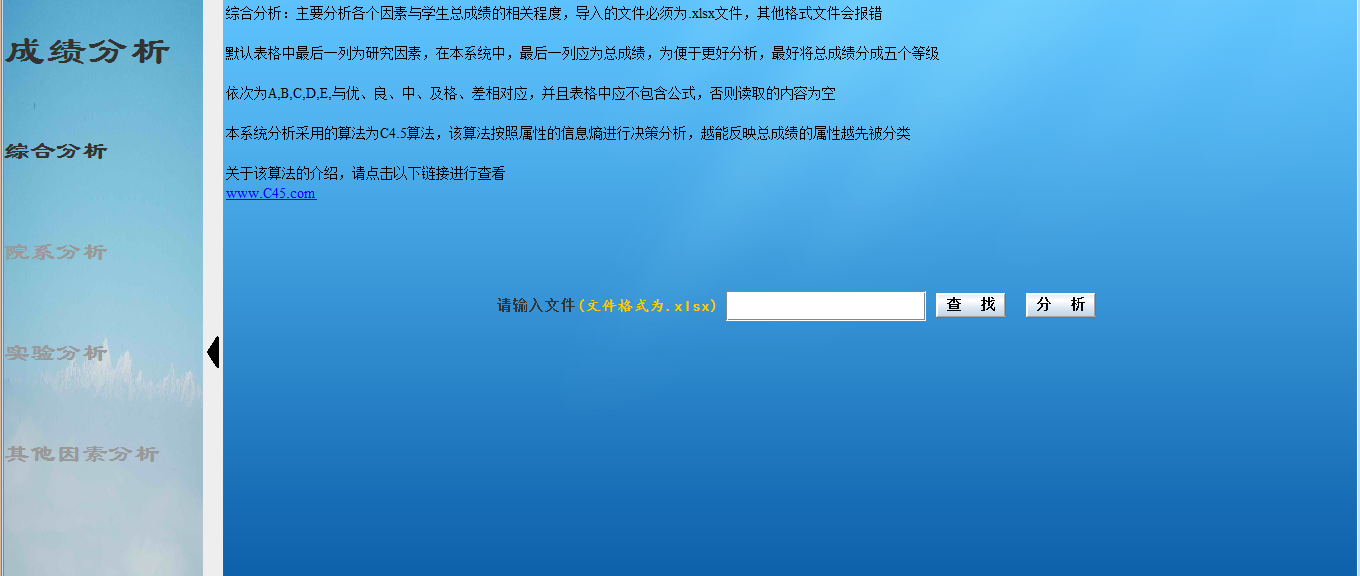


图 4.1

综合分析界面中，首先显示的是输入数据的初始界面，该界面左上方详细介绍了如何进行综合分析及要注意的一些事项，还有算法详细介绍的超链接。在界面的中央是输入文件路径的选项，包括查找按钮和分析按钮，文本框中可以选择手动输入也可以点击查找按钮调用Windows自带的查询框图，点击查询按钮，效果如下图所示：

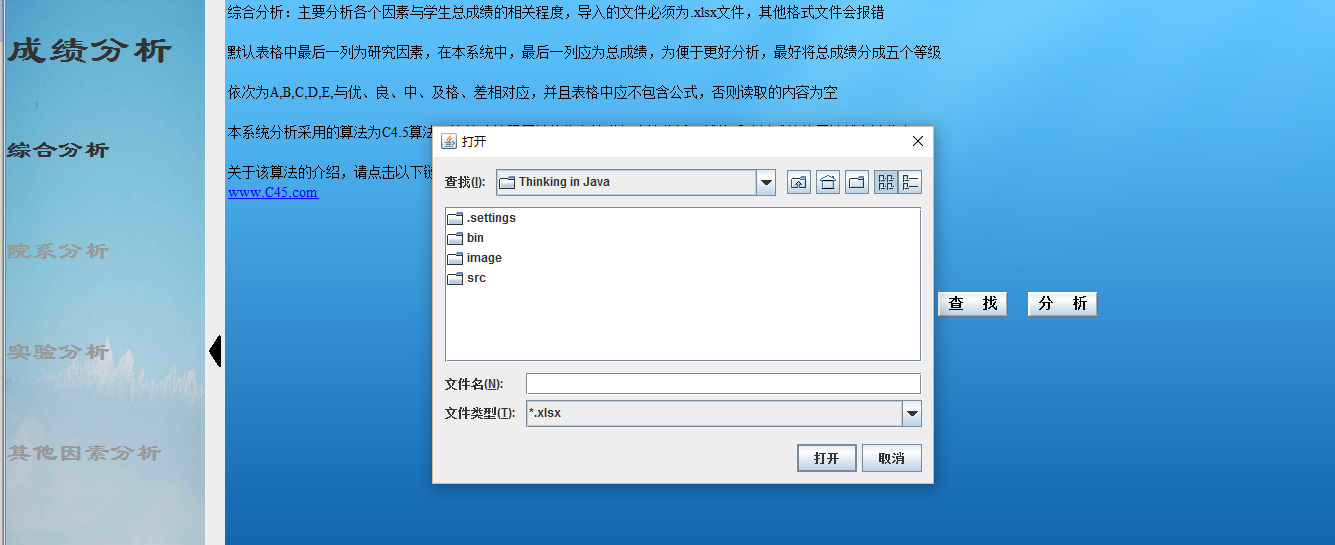


图 4.2

如果输入的文件路径不正确时，会提示文件不存在，并清空文本框内容，如果输入的文件类型正确，并且存在，则系统会自动进行分析，并给出分析的决策树，如下图所示：

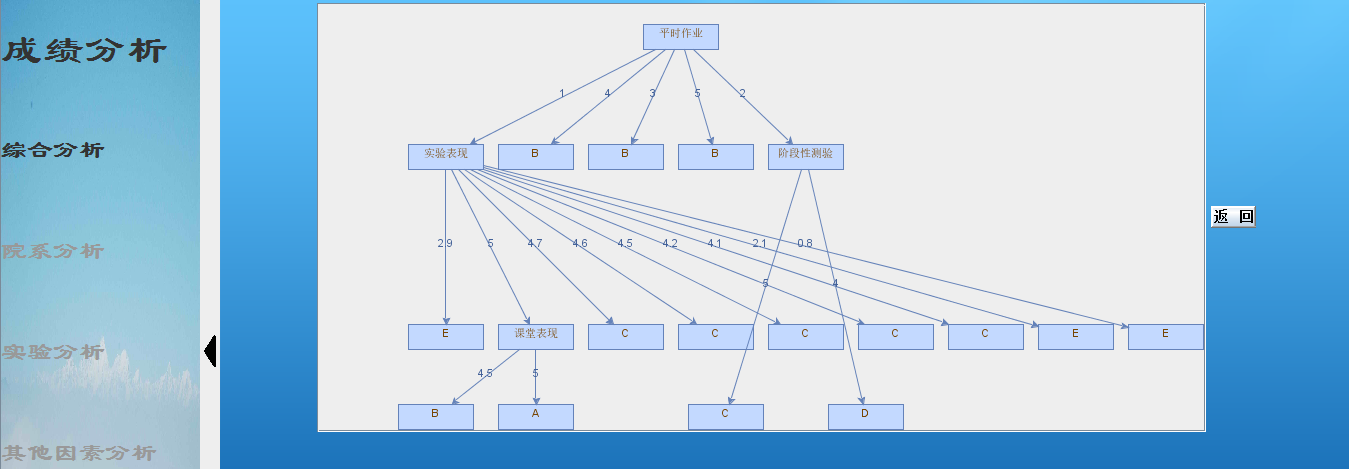


图4.3

分析按钮点击后会显示到上图中的界面，在该界面正中央是分类决策树，界面右侧有个返回按钮，点击返回按钮会退到上一个界面，重新进行数据选择和分析。

成绩分析第二部分是单因素分析部分，该部分用于分析已知的对学生成绩有影响的单个因素，主要包含院系和实验两项组成，由于两部分的实现原理相同，因此在此通过院系这一部分来说明一下单因素分析部分的设计。

首先是单因素分析的显示界面，如下图所示：

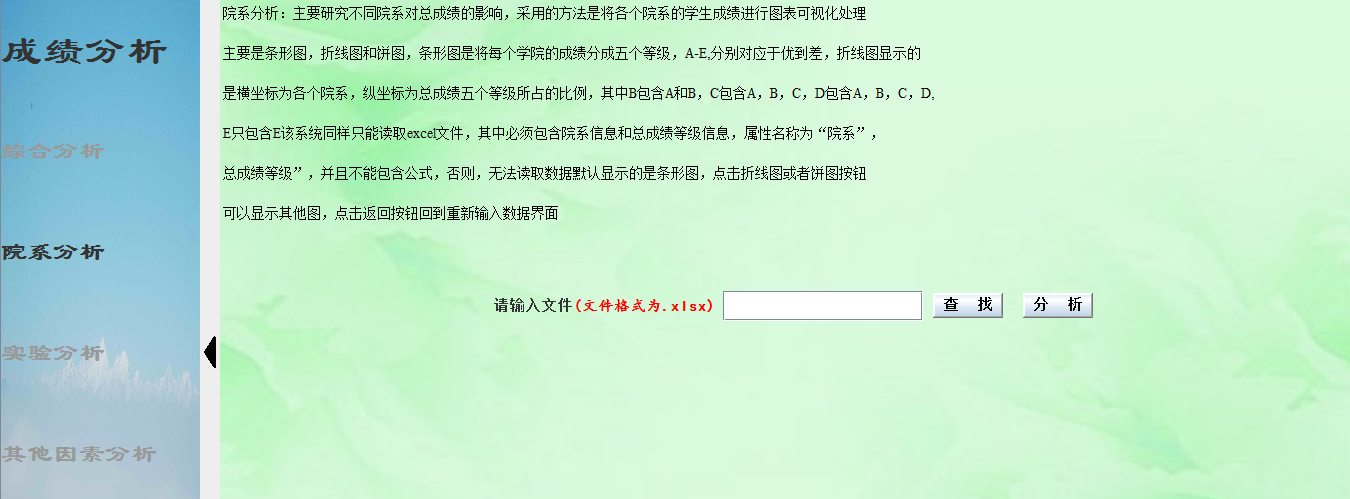


图 4.4

该界面与综合分析显示的界面一样，在这里就不在详细介绍。当输入文件并点击分析按钮以后，出现如下图所示的界面：

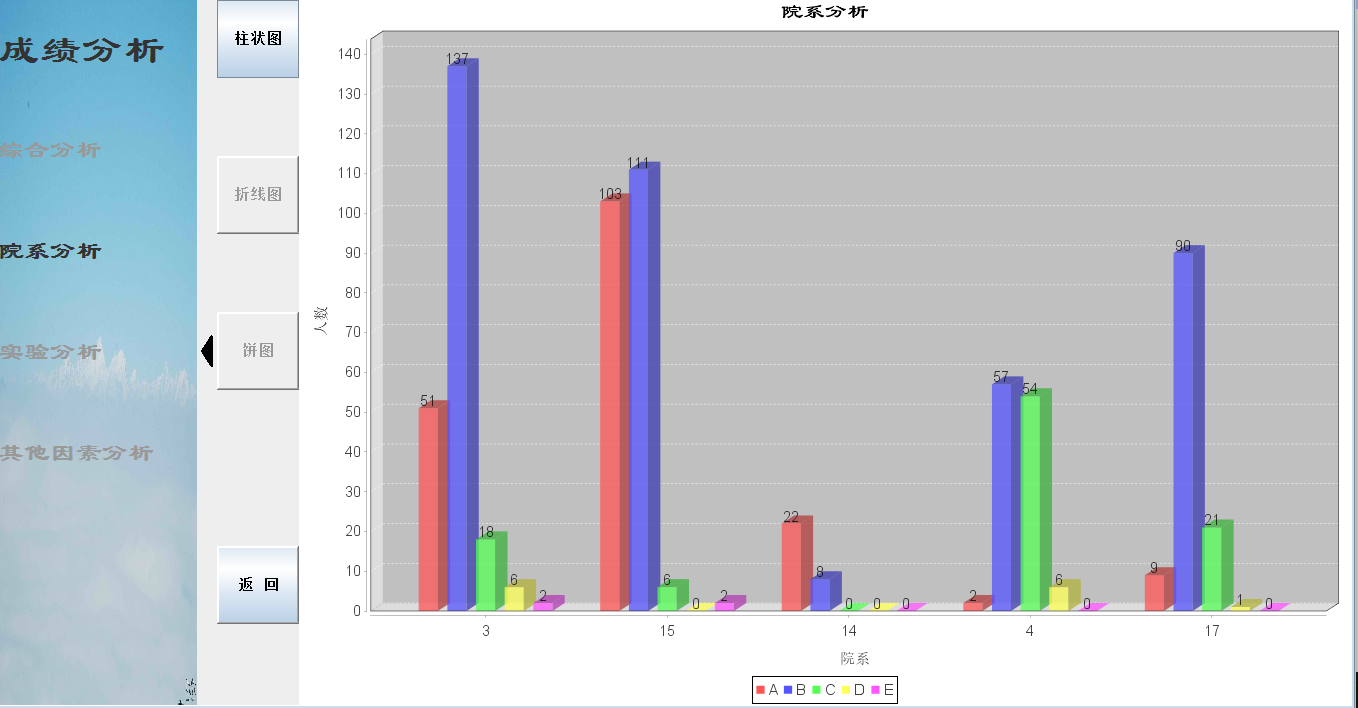


图 4.5

院系部分分析时使用的是数理统计工具，因此该界面左侧会显示三种用于数理统计的统计图，分别是柱状图，折线图，饼图。默认显示的柱状图，点击不同的按钮会显示不同的图形，在左下方有个返回按钮，同样点击返回按钮可以返回到数据输入界面。

成绩分析第三部分是其他因素分析，该部分主要用于研究其他未在本次分析中出现的因素对学生成绩的影响，该部分的显示界面与第一、二部分的显示界面是基本相同的，不同的地方在于多了一个属性选择按钮，在输入数据以后，点击属性选择按钮会弹出一个对话框，在对话框中填入要分析的属性，然后点击分析按钮会进行到功能分析界面，功能分析界面左侧包含众多的数据挖掘方法，选择相应的分析算法后会在右侧界面显示分析后的结果，该部分的设计是为了弥补第一部分和第二部分的不足。

学生评价模块包含一个部分，与上述两个部分的初始显示界面相同，学生评价模块的初始显示也是左上角显示操作相关的注意事项，界面中央是数据输入区域，唯一不同的是数据的格式在学生评价模块中为Word文件，在输入文件点击分析按钮后，会进入到学生评价信息的统计界面，点击分析按钮后系统会自动调用数据收集算法对信息进行收集。显示界面中有条形图，折线图和饼图3种显示方式。学生评价模块的设计是为了从学生的评价中提取出一些有关课程改进的建议，方便老师对课程进行一些调整。

### **4.2.2 系统实现**

系统的实现采用的Java语言来编写的，整个系统放在一个JFrame里面，JFrame上采用的是JTabbedPane拆分窗格格式，将JFrame分成三个部分，分别是学生成绩统计，学生成绩分析，学生评价统计。

学生成绩统计部分初始界面的部分代码如下：



图4.6

图中显示的是学生成绩统计界面中的所有部件，RemoteCall类是调用Windows自带的文件查询窗口，部分代码如下：

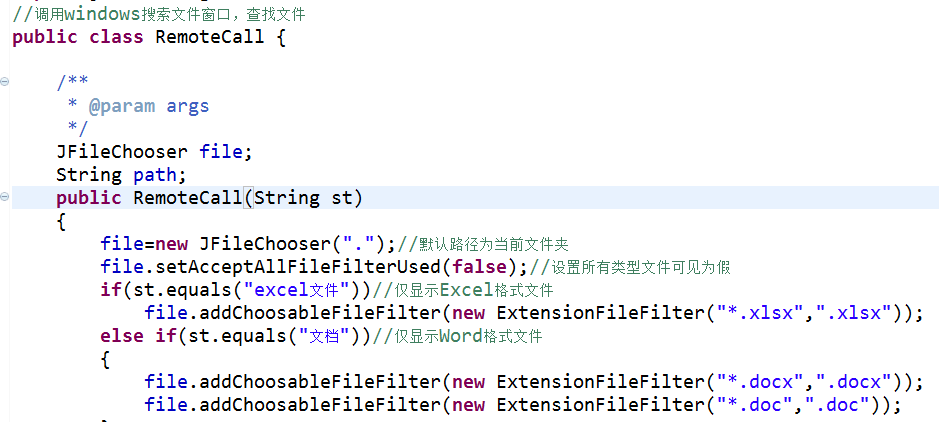


图4.7

上图中，文件查找对话框默认的文件路径是本程序所在的路径，当要求输入的文件格式为Excel文件时，查找框只会显示Excel文件和文件夹，当要求输入的文件格式为Word文件时，查找框只会显示Word文件和文件夹。

统计模块读取了Excel 文件后，会弹出设置参数的对话框，然后就会对学生成绩进行统计并导出原Excel表格，这里比较简单就不再详细描述了。

成绩分析部分界面的实现代码如下：



图4.8

成绩分析部分面板分成左右两个部分，左边是四个标签，表示要分析的内容，右边是显示界面，为卡片布局，共4个卡片，每个卡片又是单独的卡片布局，用来显示初始的数据输入界面和结果显示界面。分析结果显示界面又分成左右两个部分，左边是结果的显示方式，右边是结果显示，成绩分析布局的实现涉及两个主要功能的实现，一个是C4.5算法的实现，一个是统计绘图。

C4.5算法的实现分成两个大部分，第一个是信息增益率的计算，信息增益率的计算法分成三小步，分别是计算信息熵，信息增益，分裂信息度量和计算信息增益率，计算信息熵的部分代码如下：

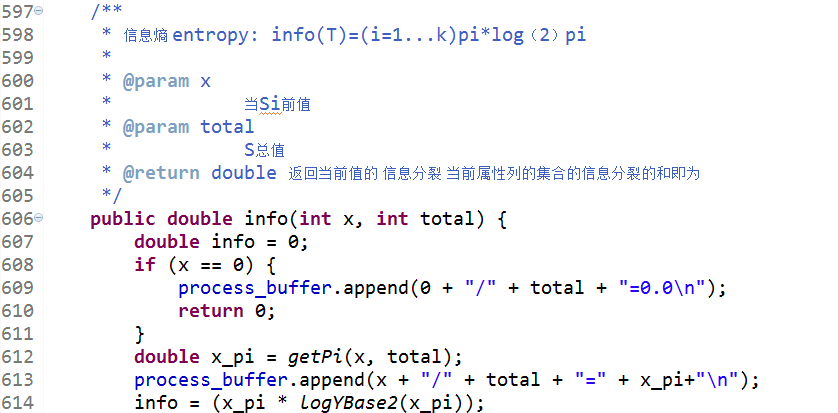


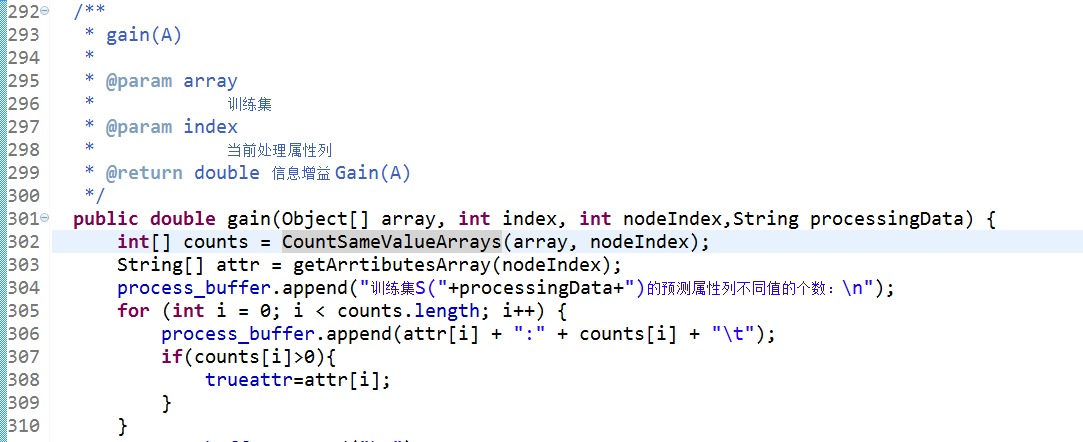
图4.9

计算信息熵的公式为：

Info (D) =-

其中D为要计算的属性的信息熵，表示该属性某一分类值占所有样本数的比值。

计算信息增益值的部分代码如下：

图4.10

计算信息增益值的公式为：

Gain (A) = info(D)- (D)

其中info(D)为类属性的信息熵， (D)为属性A的信息熵。

计算分裂信息度量的部分代码如下：

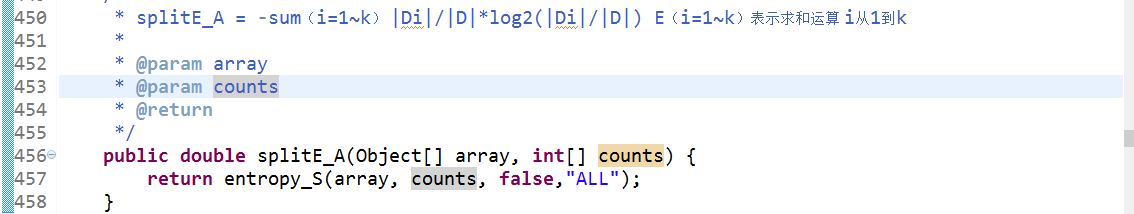


图4.11

计算分裂信息度量的公式为：

SplitInformation (A) =-

其中表示属性p中分类占总样本数的比值。

计算信息增益率的部分代码如下：

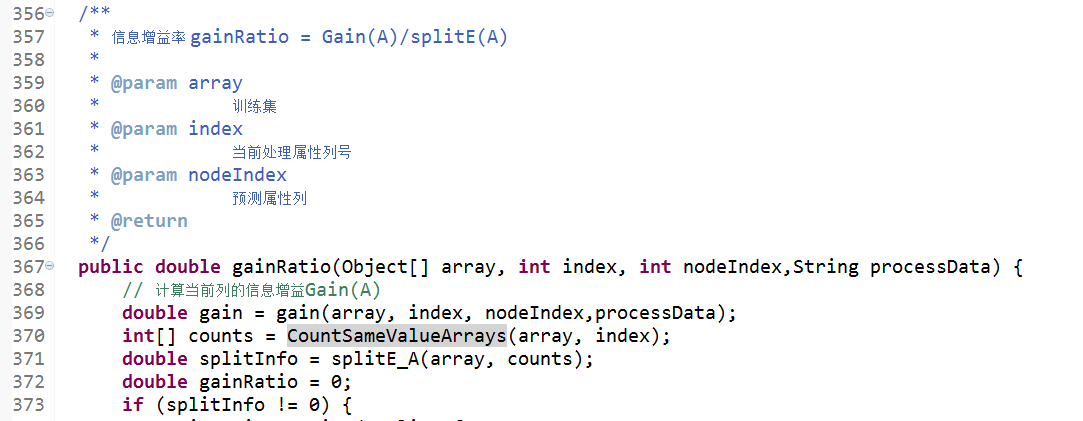


图4.12

计算信息增益率的公式为：

GainRatio (A) =

其中为属性A的信息增益值，是属性A的分裂信息度量。

第二部分是决策树的绘制，部分代码如下：

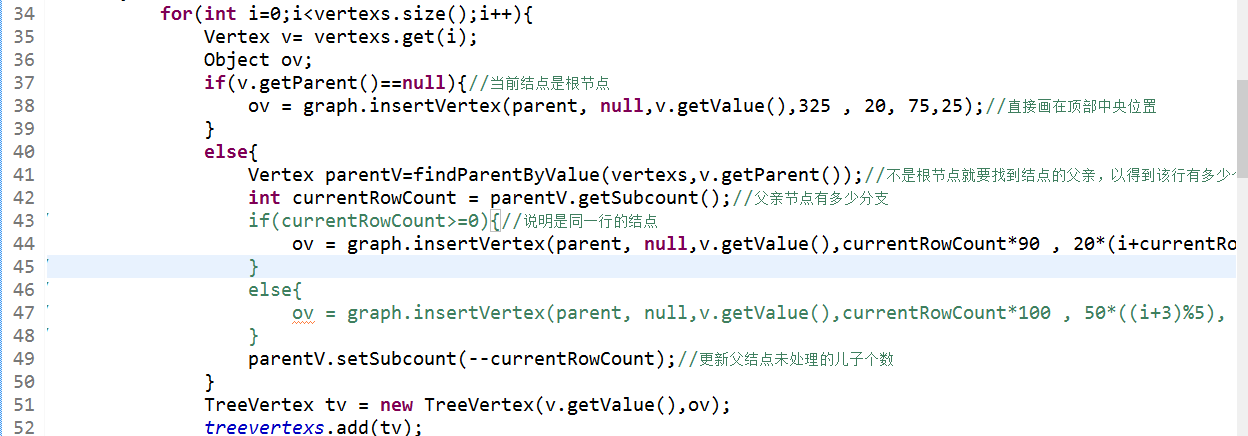


图4.13

实现思路是先从根节点开始画起，将根节点画在界面顶部中央，然后找根节点的左分支，画左分支时先得到父节点的节点数，然后循环从左往右画，画完后再画分支节点的子节点，按照上述步骤进行递归，直到所有节点画完为止。

成绩分析部分的第二大功能是统计绘图，统计绘图分两大步骤进行，第一步是导入有用数据。部分代码如下图：

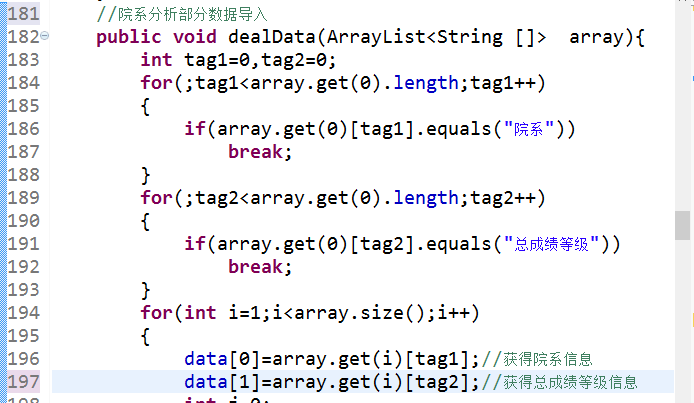


图4.14

从Excel表格中读取的数据存入到一个动态数组中，然后再从动态数组中读取该功能部分需要的数据，进行相关统计。

第二步是统计图的绘制，部分代码如下：



图4.15

绘图过程中首先要对统计图的字体进行设置，否则会出现中文显示乱码的现象，然后相应的统计图要填入相应的数据，最后是进行图形绘制，图形的绘制引入了一个开放的图标绘制类库JFreeChart，这个库中包含各种统计图的生成，直接引入即可。

学生评价部分采用的是卡片布局，包含数据输入界面和结果显示界面。数据输入界面和结果显示界面与前面的数据输入界面布局相同，实现方式相同。

## **本章小结**

本章主要介绍了两方面的内容，一方面介绍了成绩分析系统的需求分析，另一方面介绍了系统的设计与实现过程。

成绩分析系统的需求分析主要包括功能需求，数据需求和技术需求。功能需求部分说明了成绩分析系统所要实现的功能以及实现这些功能的目的。数据需求部分说明了系统所要求的数据来源，数据格式以及数据内容，要求所分析的数据必须真实有效，不能存在空缺值。技术需求部分说明了成绩分析系统中所涉及到的一些比较重要的算法和工具，介绍了编写系统的语言。

成绩分析系统的设计与实现过程主要介绍了系统的三个模块的主要设计与实现过程。其中包括每个模块的布局，每个模块的操作过程，以及每个模块实现过程中比较重要的一些代码和说明。

本章是本文中的工程实现部分，主要的是关于一些系统实际编写过程中的大体框架，本章系统设计与实现的基础是第三章的数据挖掘分析所得出的结论。本章在一定程度上将第三章的内容进行了整理汇总，然后设计并实现成一个可以为老师教学工作服务的系统。

# **总结**

## **工作总结**

随着计算机的快速发展，互联网的普及，大数据离人们的生活越来越近。目前，全国高校为了方便管理学生教学工作，都建立了自己的一套完整教务系统，方便老师进行教学工作，也方便学生进行个人信息的查询工作。但是，目前的教务系统提供的仅仅是查询，存储等工作，对于教务系统中庞大的学生各方面的信息并没有充分利用。因此，本文希望通过建立一个分析学生成绩的系统，希望能够充分利用教务系统中的庞大信息，帮助教师改善教学工作。

本文从4个方面对学生成绩分析系统进行了系统的描述，首先介绍了国内外数据挖掘方面的发展和数据挖掘技术应用在学校成绩管理方面的使用，通过阅读了大量的资料和各种文献，发现国内和国外对于这些方面都有研究。国外的学者们研究了课程之间的相互联系，发现大学生的课程之间存在学习的先后顺序关系，还有的学者研究了学生课程论坛中的信息，从学生的交流信息中发现学生的社交能力和认知能力以及交流能力有着线性相关关系。国内部分高校认识到数据挖掘的重要性，因此仿照外国高校尝试开设了数据挖掘课程，有些学者用数据挖掘技术分析了大学生的性别与英语成绩的好坏是否有关系以及存在怎样的关系。总的来说，国外和国内对于数据挖掘的研究以及数据挖掘在教育方面的应用都取得过不错的进展，这也为本文的研究提供了理论基础和研究思路。

其次，本文介绍了本次研究过程中数据挖掘方面的相关知识，包括了解了数据挖掘的意义和步骤，本文进行数据挖掘的工具以及数据挖掘过程中使用到的一些算法。本文采用的数据挖掘工具是WEKA，WEKA是一款强大的数据挖掘工作平台，集成了大量的数据娃聚算法和数据处理的工具。本文设计到的数据挖掘算法是C4.5算法和Apriori算法。前者能够利用现有样本集找出不同属性对类属性的信息反映情况，可以用来发现哪种属性最能影响类属性，后者可以用来发现两者或者多者属性之间的关系，都适合挖掘出影响学生成绩的因素。

第三部分对整个数据挖掘工作进行了介绍，包括数据预处理，数据挖掘以及知识表示和最后的结论。数据预处理部分对原数据进行了数据集成，数据变换，数据选择等方面的处理，主要是为了清除噪音数据，方便分析。数据挖掘部分介绍了进行数据分析的操作过程，得到了哪些结果，知识表示部分将数据挖掘的结果用可视化的方式展现出来，方便得出最后的结论。

第四部分是系统的设计实现部分，主要包括系统的需求分析和系统的设计实现部分。系统的需求分析部分介绍了系统的功能需求，系统的数据需求和系统的技术需求，为系统的设计与实现提供了技术支持和理论支持。系统的设计实现部分对整个系统的设计实现进行了详细说明，分别介绍了每个功能模块的整体结构和实现方式。

总体上说，本文的工作主要有两个部分，一部分是分析工作，通过收集的数据分析影响因素，得出一些结论，另一部分是系统设计实现工作，根据分析得出的结论，设计成绩系统的功能模块，使成绩分析系统具有这门课程的独特的特色。

## **不足与改进**

本文在分析成绩和系统设计方面都取得了一定的成果，但是整个过程中仍有不足之处，主要包括以下几个方面。

一是收集到的信息不足，由于各种因素的影响，没能在上学期结束前将设计的问卷调查发放到学生手中进行数据采集，导致得到的数据属性比较少。可以进行分析的方面也偏少，因此希望以后老师在进行成绩分析时能够设计一张问卷调查表，在学期期末发放给学生，让学生进行填写，然后用收集到的信息来进行分析，得到的结论会更加的全面。

二是系统的设计不够完善，系统的功能可以更加完善一点，本系统可以跟学生的信息统计系统相连接起来，把学生的信息统计和信息分析直接联合起来，一方面方便老师对学生成绩的统计分析，一方面也能使得系统的功能更加完善。

# **致谢**

转眼间，从去年12月份的毕设开题答辩，到今年3月的中期答辩，到现在的毕设答辩，过去了有半年的时间了。在这半年年时间里，我非常感谢我的导师傅翠娇老师，她给我提供了很多的帮助，从最初的选导师过程中，老师为我想了很长时间的毕设题目，问我对哪方面感兴趣，到开题答辩时期，老师帮助我筛选了好多篇中英文相关文献，在中期答辩的时候，老师督促我安排好自己的毕设进程，帮助我对中期答辩的ppt和文档进行批改。在中期答辩结束以后，老师每周五都会给我们召开毕设工作会议，让我们说一下自己的进度以及在工作过程中遇到的问题，积极和我们一起探讨解决，还给我们安排了下一周的工作量。在这里，真心感谢傅翠娇老师，感谢傅翠娇老师对我工作的认真督促和积极指导，我是一个喜欢拖延的人，如果不是老师积极的监督我进行毕设工作，我想我可能要等到最后才开始做毕设。真心感谢傅老师的付出，谢谢。

其次，我要感谢大学四年里的各位老师，如果没有老师们传授的各方面的知识，我今天就不会有能力来开展我的毕设工作，谢谢各位老师。

最后，我要感谢陪我一起坚持上自习的小伙伴们，没有你们的陪伴、鼓励和监督，我也是不能够完成毕设工作的，感谢每一个陪伴我走下去的同学，谢谢你们。

# **参考文献**

[1] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J].Journal of Computer Research and Development, 2013, 50(1):146-169.

[2] 涂新莉，刘波，林伟伟. 大数据研究综述[J].Application Research of Computers, Vol.31 No.6 Jun. 2014.

[3] 韩家炜.数据挖掘：概念与技术[M]. 北京：机械工业出版社，2012：8-83

[4] 冯丽霞. 大学生学习成绩影响因素研究[J].China Electric Power Education, 2013, (13)

[5] guoxiang Liu, fengxia Yang. The Application of Data mining technology in college teaching[D], 沧州：沧州职业学院.2012.

[6] Jian Wang, Zhubing Lu, Weihua Wu, Yuzhou Li. The Application of Data Mining Technology Based on Teaching Information[D]. 重庆：西南大学, 2012.

[7] Chakrit Pong-inwong, Wararat Rungworawut. Teaching Evaluation Using Data Mining on Moodle LMS Forum[A].

[8] Azauri S. Figueira, Amanda S. Lino, Adriano Del Pino Lino. Educational Data Mining to Track Students Performance on Teaching Learning Environment LabSQL[A].

[9] 扈玉婷. 大数据背景下数据挖掘技术在大学英语听力教学效果分析中的应用[J]. 校园英语, 2015(3).

[10] 李建中，刘显敏. 大数据的一个重要方面：数据可用性[J].计算机研究与发展,2013,50-(6).

[11] 李静燕. 关联规则挖掘算法研究[D], 西安：西安电子科技大学, 2008.

[12] 王珊, 王会举, 覃雄派, 周烜. 架构大数据：挑战、现状与展望[J]. 计算机学报, 2011-10, 34(10).

[13] 任磊, 杜一, 马帅, 张小龙, 戴国忠. 大数据可视分析综述[J]. 软件学报, 2014:25(9).

[14] Alvaro Figueira, Joanne Laranjeiro. Work in Progress – iGraphs for Characterization of Online Communities[A].

[15] Elena Nesterova, Ricardo Ulloa Azpeitia. Assessment of learning activities in discussion forums online[A].

[16] G.Alan Wang, Harry Jiannan Wang, Jiexun Li, Weiguo Fan. Mining Knowledge Sharing Processes in Online Discussion Forums.

[17] Liu Fang, Chen Pu. Hierarchical clustering based teaching reform courses examination data analysis approach applied in China Open University system[A].