《课程实践报告》

关联规则在学生成绩管理中的应用研究

成员： 楚梦颖 学号： 31617002

成员： 吕梦天 学号： 31617011

大连理工大学

Dalian University of Technology

# 项目概述

目前，在高校的管理中，对学生的管理是其重要的组成部分。但由于我国大众化高等教育的普及，高校连年扩招，学生数量的激增，给高校教学工作带来了诸多新的问题。学生进入大学是否就等于进入“保险箱”?随着大学考试制度的规范化、严格化，经常有些学生会因为多门课程成绩的不及格而导致留级、退学的后果。然而，许多对学生的管理工作落到了刚参加工作的年轻辅导员的身上，因为知识的缺乏、经验的不足使他们对学生的管理方面显得力不从心，这势必会对学校的整体管理质量产生负面影响。如何科学地研究分析包括学生成绩在内的各个教学环节中的大量的数据信息，从中获取知识，继而科学地指导教学、管理工作显得至关重要。

作为教学管理部门，如果能在这些后果出现之前，得知前期哪些课程学的不好会影响后期课程的学习并有可能导致留级、退学，以便对这些学生提出警告，加强督学，避免最终留级、退学的发生，势必将有助于学生成绩提升，从而体现出对学生的人性化教育。这就不得不提到数据挖掘技术。

数据挖掘作为一种热门的技术，能够从大量的数据中通过特定的算法搜索出隐藏于其中的信息。然而，数据挖掘技术在商业、金融业以及企业的生产、市场营销等方面都得到了广泛的应用，而在教育领域应用相对较少。由于院校的不断规模化发展，造成学生成绩数据量不断激增，而大部分院校对这些巨量成绩数据的处理基本上都还只是做一些较为简单的处理和分析，比如在阅卷、登分、上报、公布、登记并分类保管后，成绩处理工作即告结束，而并没有从中进行认真深入分析。展望未来几年后高校的发展，其趋势将会从数量型转向质量型、从外延发展模式转向内涵发展模式，无疑将导致院校间竞争更加激烈。因此，提高教学质量作为学校内涵发展和建设的中心任务，其重要性和艰巨性今后都将会进一步增强，而学生成绩又是作为判别教学质量优劣的一个重要依据，所以大力运用新技术、新方法来加强学生成绩的深入剖析，及时发现更多的潜在而有用的模式和关系并对症下药，从而促进教学质量的显著提高，诸如此类工作的开展和强化对任何一所高校而言，目前都无疑将比以往任何时候显得更为迫切和重要。

所以，各院校目前尽管都积累了大量学生成绩数据，但对这些成绩数据的使用基本上还停留在一些诸如排序、筛选、简单汇总等初级处理阶段，缺乏对其加以深入挖掘分析。在这些巨量成绩数据的背后肯定隐藏着一些对教学决策和改革有重要参考价值的规则，而这些与学校实际情况相符的规则、模式的发现和应用无疑会给以始终围绕提高教学质量为中心的学校教学管理、决策及改革等各项工作的开展提供重要指导意义。

所以，本文希望通过对大连理工大学软件学院2012级本科生必修课成绩的分析，挖掘出一些对教学决策和改革有重要参考价值的规则。也希望我们的研究能够根据不同的需求，以不同的方式，为不同的用户提供有用的信息。

(1)教学管理者：运用学生成绩分析应用对学生考试成绩进行分析，从而直观的了解到现阶段学校的教学状况，还可以了解到一个年级的学风状况，有助于分析导致这种现象产生的原因，进而规避此现象的发生。

(2)教师：根据对学生考试成绩进行分析的结果，有助于了解学生掌握知识的状况，对老师起到一个反馈的作用，使老师调整教学模式，提高学生对学习过程的兴趣度，从而改善高校教学质量。

(3)学生：根据自己现有的学习成绩，可以预测出在后续学习过程中的弱势学科，学生可以根据预测结果，对以后的弱势学科有针对性的学习，避免考试不及格现象的发生。

# 项目研究

课程之间存在着一定的相关性。本节通过对大连理工大学软件学院2012级本科生必修课成绩的分析，运用Apriori算法挖掘出相互有影响的课程间的关联规则，从而能够发现导致留级、退学的那些课程，进而指导学生合理选课和安排对不同科目的学习计划。并希望将这些信息呈现给教学工作者，从而为教学部门提供决策支持信息，促使更好地开展教学工作，提高教学质量。

## 2.1 数据整理

我们以教务公布的大连理工大学软件学院2012级本科生前三年必修课成绩为数据源，共计562条数据如图1。

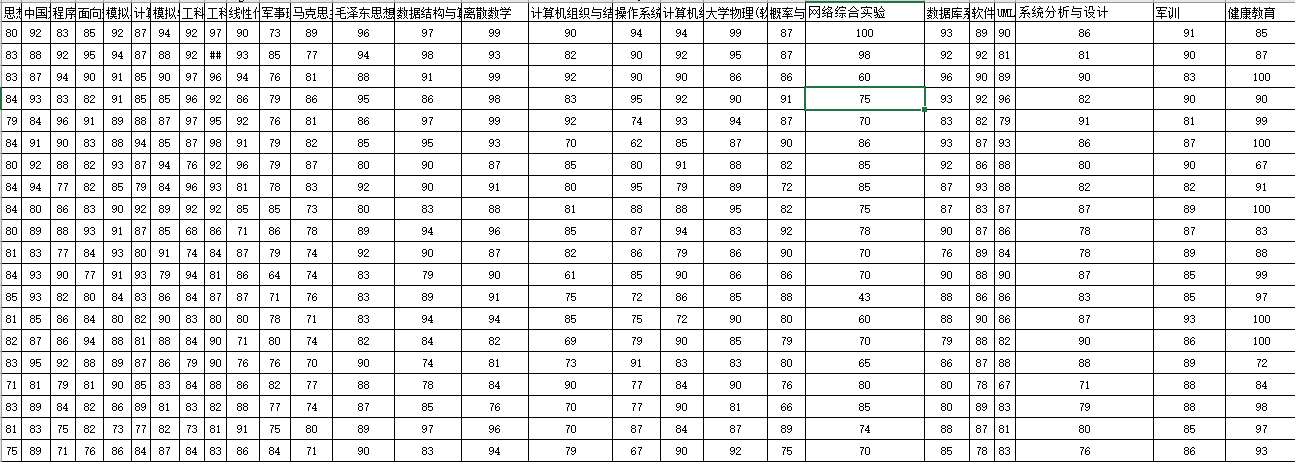


图1 软件学院2012级本科生前三年必修课成绩片段

通过分析发现有如下一些特点：

①成绩评定标准相同，均为百分制。

②所公布的成绩为学生一考成绩。

③成绩单中按照不同的专业（网络工程，日语强化，软件工程）显示成绩，各专业学生所修必修课不尽相同。

④存在成绩缺少，即存在空缺值。从教务老师处得知，缺少成绩主要是因为留级、退学、错学期选课或考试等原因造成的。

针对以上分析的数据特点，我们采用了以下的数据处理方法：

（1）各专业学生课程不尽相同

考虑到不同专业必修课大部分是相同的，所以我们只保留了公共课程，删除了非公共课程，如日语课，英语课，编译原理，高级C语言及网络编程技术等，只研究公共的27门课程。

（2）成绩缺失

结合学生考试实际，缺失数据不能直接填充0值。而处理数据缺失的常用方法有均值填充法，预测法，删除缺失等，我们采用了删除缺失的方法，保留了541条有效数据。

（3）课程名称对应关系

为了方便后序操作，我们把课程名称转化为了字母表示，对应规则如下表：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 思想道德修养与法律基础 | 中国近现代史纲要 | 程序设计基础与C程序设计 | 面向对象方法与C++程序设 | 模拟与数字电路 | 计算机系统组装与设置 | 模拟与数字电路实验 | 工科数学分析基础1 | 工科数学分析基础2 | 线性代数与解析几何 |
| A | B | C | D | E | F | G | H | I | J |
| 军事理论 | 马克思主义基本原理 | 毛泽东思想和中国特色社会主义理论体系概论 | 数据结构与算法 | 离散数学 | 计算机组织与结构 | 操作系统 | 计算机组织与结构实验 | 大学物理 | 概率与统计A |
| K | L | M | N | O | P | Q | R | S | T |
| 网络综合实验 | 数据库系统 | 软件工程 | UML | 系统分析与设计 |  |  |  |  |  |
| U | V | W | X | Y |  |  |  |  |  |

表1 课程名称与字母的对应关系

（4）数据离散化

数据离散化是指将连续属性值转换为有限的分立值的过程。现实生活中的 成绩数据是连续型的，但连续的取值是没有必要的，例如“成绩为92”与“成绩为93”在一般情况下近似的可以视为同一情形，如(成绩：优)。

对于成绩数据的离散化，可以有以下方法：

①固定区间法：如对于百分制，我们将小于60分的成绩设定为“1”，[60， 70]设定为“2”，[70，80]设定为“3”，[80，90]设定为“4”，90分以上设定为 “5”，则对于每个成绩数据，我们都可以得到它的对应离散值。

②等分区间法：对于所选取的成绩数据源，我们找出它的最大值和最小值， 按照所需的区间数目对其进行等分，从而得到成绩的离散区间。如成绩最低分是50，最高分是100，若划分为5个区间，则[50，60]设 定为“l”，[60，70）设定为“2”，[70，80）设定为“3”，[80，90）设定 为“4”，[90，100]设定为“5”，这样对于每个学生的入校成绩，我们都能得到它的对应离散值。

③排序量化法：在实际教学考试中，由于每门课试题的难易程度并不相同，或者老师的评分习惯有所差异，这样单纯从成绩上来进行离散划分就不太合适，在这种情况下，可以采用排序量化法。这种方法要求在离散化之前，先对成绩数据进行排序。然后按照实际的要求，对全部数据进行比例划分。例如，对于100个数据排序后，我们将前10％的成绩数据设定为 “1”，将接下来的40％设定为“2”，再将接下来的40％设定为“3”，将最后的10％设定为“4”，这样，对于每个位置的成绩数据都有一个离散值与之对应。

我们采用的是第三种方法。我们将541人的27门课的成绩按照1:3:1划分为三等份。由于存在较多并列的分数，最终将[89,100]设定为“3”，[71,88]设定为“2”，[0,70]设定为“1”。本文选取大连理工大学软件学院2012级本科生前三年必修课成绩来挖掘关联规则，学生成绩数据片段如图2：

经过上述步骤的转化，可以得到转化后的学生成绩数据片段如图3：

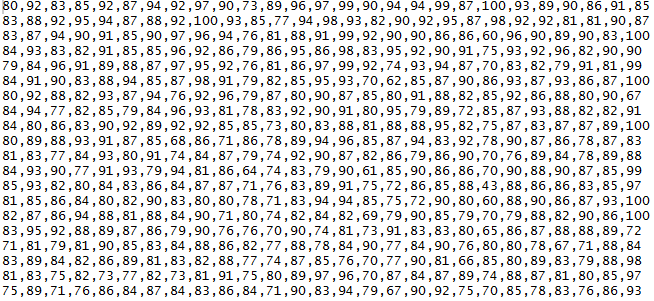


图2转化前的学生成绩数据片段

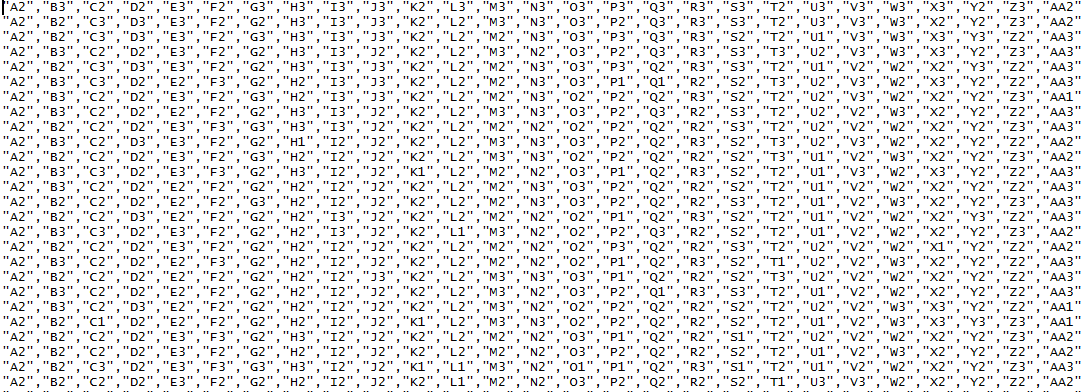


图3 转化后的学生成绩数据片段

数据处理的核心代码如下：

makeLevel <- function(score = 60,ch = "A"){

if(score <= 70)

return (paste(ch,"1",sep=""))

if(score > 70 && score < 89)

return (paste(ch,"2",sep=""))

if(score >= 89)

return (paste(ch,"3",sep=""))

return (0)

}

achivement<-read.csv(getFilePath("achivement"),header=FALSE,fileEncoding="utf-8",encoding="utf-8")

size <- length(achivement)

result <- do.call(rbind,lapply(achivement[,1],makeLevel,numConvertToColName(1)))

for(col in 2:size){

wdata <- do.call(rbind,lapply(achivement[,col],makeLevel,numConvertToColName(col)))

result <- cbind(result,wdata)

}

write.csv(result,getFilePath("achivement2"),row.names=TRUE,fileEncoding = "utf-8")

## 2.2 数据挖掘

关联规则是数据挖掘诸多功能中的一种，也是目前最为重要和应用最广泛 的数据挖掘方法之一。本节将运用Apriori算法挖掘出相互有影响的课程间的关联规则，从而能够发现导致留级、退学的那些课程，进而指导学生合理选课和安排对不同科目的学习计划。

### 2.2.1 分析整体概况

挖掘关联规则之前我们先做了一个简单的统计，以便了解整体概况。首先我们先针对每一个科目的平均成绩做了一个统计图,如图4;在图中我们可以看出各科的平均成绩均在75分以上，说明我校学生的学习水平平均比较高。

然后我们针对各科成绩在的三个等级中的分布绘制了统计图，如图5；其中蓝色表示成绩在[89,100]区间的个数，绿色表示成绩在[71,88] 区间的个数，红色表示成绩在[0,70] 区间的个数。在图5中我们可以看出课程AA和课程Z的高分人数比较多。通过表1可以看出这两门课分别是军训和健康教育，可以得出这两门课的高分普遍偏多。

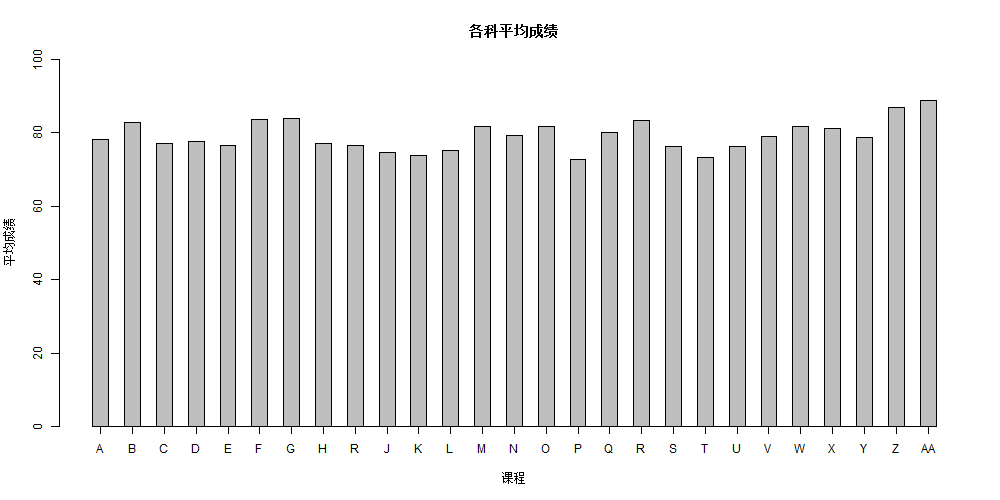


图4 各科平均成绩直方图

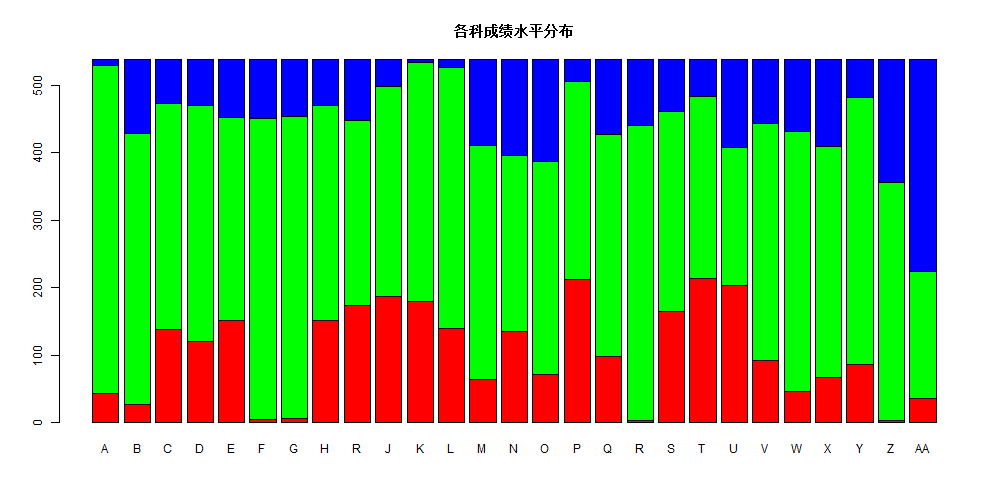


图5 各科成绩水平分布直方图

### 2.2.2 挖掘关联规则

关联规则是数据挖掘诸多功能中的一种，也是目前最为重要和应用最广泛 的数据挖掘方法之一。关联规则挖掘是发现大量数据中项集之间有趣的关联或相关联系。典型的关联规则挖掘是购物篮分析。由于关联规则挖掘可以发现用传统的人工智能和统计方法无法发现的项与项或属性与属性间的关联规律，因此具有重要的研究价值。同时它也满足了人们从大规模数据存储中获取知识的迫切需求。另外，作为数据挖掘中的一个基本任务，关联规则又被用于其他挖掘任务当中，例如数据建模、分类及决策支持等。尽管关联规则挖掘起源于商业上对市场购物篮进行分析的问题，但是随着研究的不断深入，其基本模型在多角度得到了扩充。关联规则挖掘技术的应用领域也越来越广泛。

本节我们主要采用Apriori算法来挖掘分析。该算法是一种最有影响的挖掘布尔关联规则频繁项集的算法，其核心思想是通过候选集生成和情节的向下封闭检测两个阶段来挖掘频繁项集。而且算法已经被广泛的应用到商业、网络安全等各个领域。

挖掘关联规则的步骤大体可以由一个两步的过程来描述：

①找出所有的频繁项集。即找出所有那些支持度大于事先给定的支持度阈 值的项集。

②在找出的频繁项集的基础上产生强关联规则。即产生那些支持度和置信 度分别大于或等于事先给定的支持度阈值和置信度阈值的关联规则。

在上述两个步骤中，第二个步骤相对要容易一些，因为它只需要在己经找出的频繁项集的基础上列出所有可能的关联规则，然后用置信度阈值来衡量这些关联规则。而第一个步骤是挖掘关联规则的关键步骤，需要大量的I/O操作，挖掘关联规则的总体性能由第一个步骤决定。

衡量关联规则价值的指标：

①支持度support：就是该项集出现的次数除以总的记录数。支持度的意义在于度量项集在整个事务集中出现的频次。我们在发现规则的时候，希望关注频次高的项集。

②置信度confidence：项集{X，Y}同时出现的次数占项集{X}出现次数的比例。发生X的条件下，又发生Y的概率。计算方法如下：

http://img.blog.csdn.net/20150420154115423?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvZ2p3YW5nMTk4Mw==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

③提升度lift:度量项集{X}和项集{Y}的独立性。即，Lift(X->Y)= 1 表明 {X}，{Y}相互独立。如果X 独立于Y ，P(XY)=P(X)\*P(Y)。如果该值=1,说明两个条件没有任何关联,如果<1,说明A条件(或者说A事件的发生)与B事件是相斥的。

设置不同的支持度，置信度阈值，得到的关联规则数目是不同的。为了直观得到关联规则数目随支持度，置信度变化的情况，我们针对置信度阈值分别为0.3，0.7，0.8，支持度阈值分别是0.35，0.4，0.45，0.5，0.55，0.6，0.65，0.7进行了试验，得到了关联规则数目随支持度，置信度的变化情况，绘制了图6。核心代码如下：

getItemsCount <- function(support,confidence=0.3,minlen=2){

w <- getRules(support,confidence,minlen)

writeData(w,paste("support",support,sep=""))

return (length(w))

}

getDataSourceToPaintLineGraph <- function(){

sRange <- c(0.3,0.35,0.4,0.45,0.5,0.55,0.6,0.65,0.7)

result <- c()

for(r in sRange){

result <- append(result,c(getItemsCount(r)))

}

paintLineGraph(y=result,x=sRange)

}

paintLineGraph <- function(y,x){

png(filename=paste(getImgPath(),"关联规则数目对比图.png",sep=""),width=1000,height=500)

plot(x = x,type="b",y=y)

dev.off()}

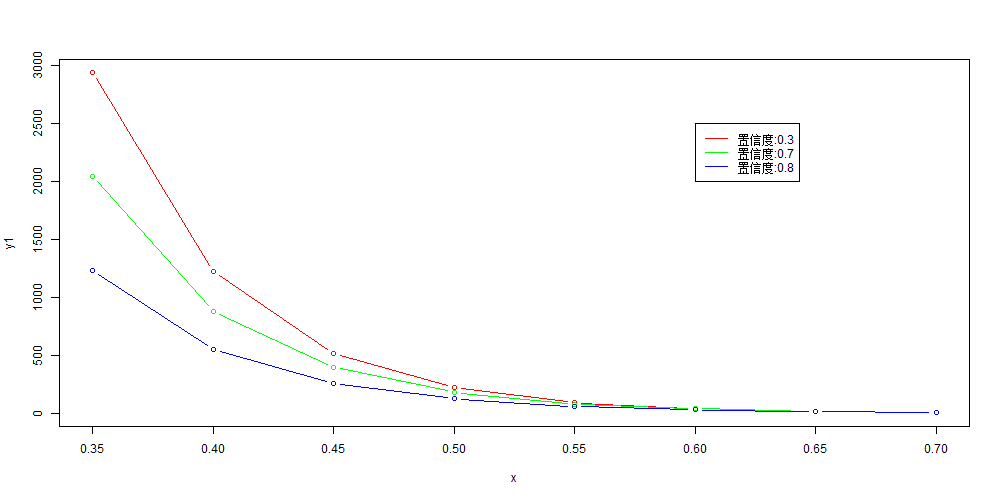


图6 关联规则数目随支持度、置信度阈值调整的变化趋势图

之后，我们选定置信度阈值为0.3和支持度阈值为0.5，得到了226条关联规则。我们使用R的arules包提供的apriori方法进行关联规则的发现，并根据发掘出的关联规则绘制了Scatter Plot，如图7，Grouped Matrix，如图8，以及Graph，如图9。相关代码如下：

dealWithRules <- function(support=0.5,confidence=0.3,minlen=2){

rules <- getRules(support,confidence,minlen)

writeData(rules,"achivement\_rules")

paintScatterPlot(rules,"Scatter\_rules")

paintGroupedMatrix(rules,"Matrix\_rules")

paintGraph(rules,"Graph\_rules")

}

getRules <- function(support=0.5,confidence=0.3,minlen=2){

wdata <- getDataSource()

achivementRules <- apriori(wdata,parameter = list(support=support,

confidence=confidence,minlen=minlen) )

ordered\_achivementrules <- sort(achivementRules, by="lift")

return (ordered\_achivementrules)

}

writeData <- function(wdata,filename){

write(wdata,file=getFilePath(filename),sep=",",quote=TRUE,row.names=FALSE)

}

#'绘制Scatter Plot

paintScatterPlot <- function(wdata,filename){

png(filename=paste(getImgPath(),filename,".png",sep=""),width=1000,height=500)

plot(wdata)

dev.off()

}

#'绘制Grouped Matrix

paintGroupedMatrix <- function(wdata,filename){

png(filename=paste(getImgPath(),filename,".png",sep=""),width=1000,height=500)

plot(wdata,method="grouped")

dev.off()

}

#'绘制graph

paintGraph <- function(wdata,filename){

png(filename=paste(getImgPath(),filename,".png",sep=""),width=1000,height=500)

plot(wdata,method="graph",shading = "lift")

dev.off()

}

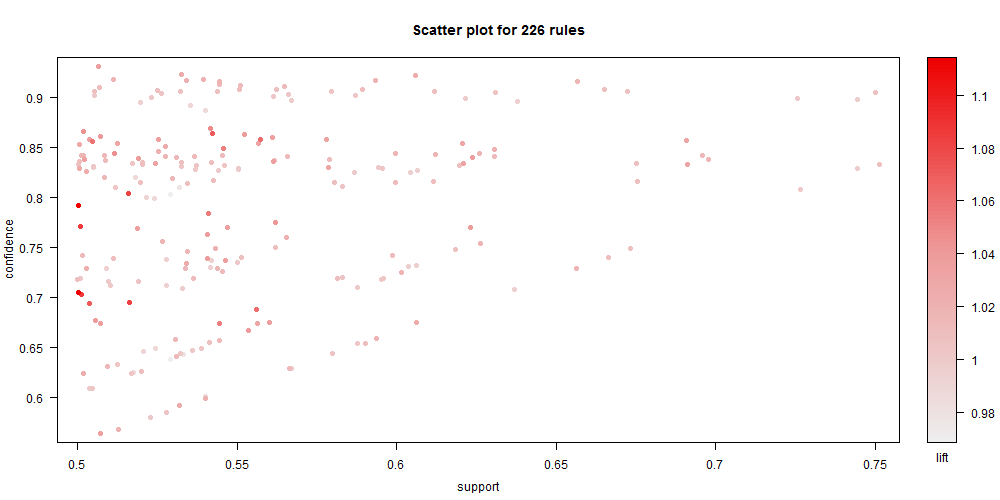


图7

图7的散点图表示了关联规则的分布规律，支持度主要集中在0.5-0.6之间，置信度集中在0.6-0.9之间，其中每一个点的颜色的深浅表示了该条规则的提升度的大小。

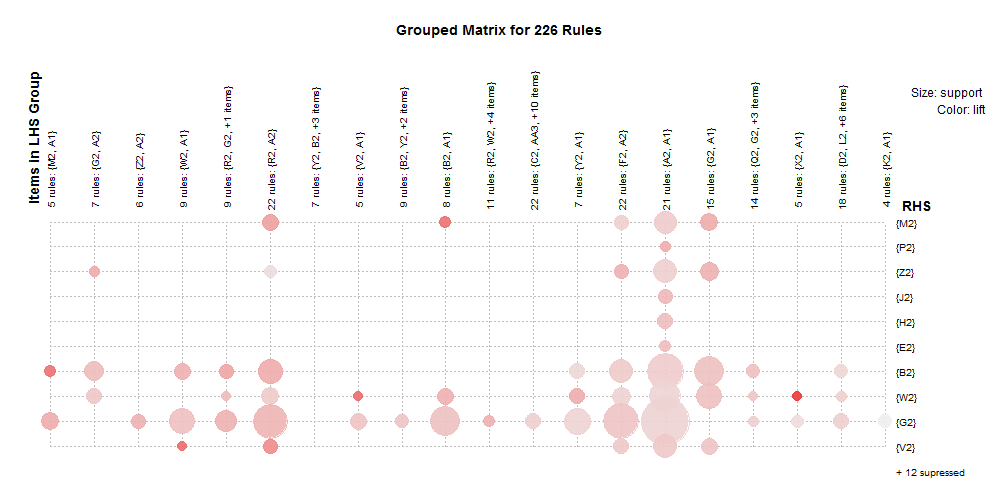


图8

在图8这个矩阵图中，每一个气泡就是一个关联规则，上面的就是该规则的LHS，右侧的就是RHS，其中气泡的大小代表了支持度的大小，颜色的深浅表示了提升度的大小，

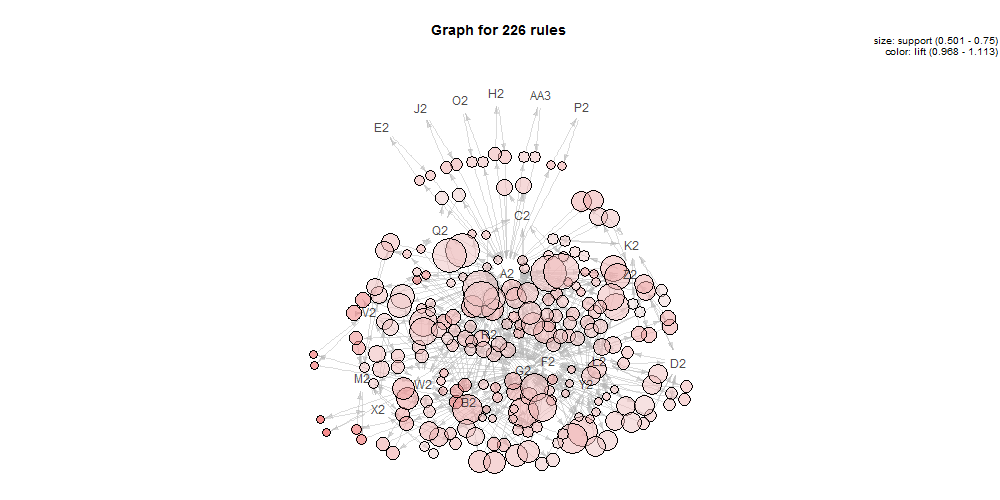


图9

在图9中，每一个圈的入度就是关联规则的LHS，圈的大小表示关联规则的支持度，颜色的深浅表示的是提升度的大小。

## 2.3 数据分析

在考试成绩分析这一主题中，我们想分析课程之间的关联关系，这样将通过数据挖掘找出各门课程之间的影响关系。 设定最小支持度为：0．5，最小置信度为：O．3，则共得 到98个频繁项集，225条关联规则，这里我们仅选取部分 结果如表2,进行解释：

|  |  |  |  |
| --- | --- | --- | --- |
| **关联规则** | **支持度** | **置信度** | **提升度** |
| {机组实验2} => {机组1} | 0.345656192236599 | 0.4279176201373 | 1.09199732308622 |
| {数电模电实验2} => {机组1} | 0.349353049907579 | 0.420935412026726 | 1.07417951842669 |
| {机组实验2} => {网络综合实验1} | 0.310536044362292 | 0.384439359267735 | 1.02454036139825 |
| {C++2} => {C2} | 0.304990757855823 | 0.767441860465116 | 1.23567275747508 |
| {大学物理2} => {概率与统计2} | 0.32902033271719 | 0.601351351351351 | 1.20048369402613 |

表2 部分关联规则

通过以上关联规则，我们可以发现：C++比较不错的同学，他的C语言的成绩也不错；机组实验成绩一般的同学和数电模电实验成绩一般的同学，其机组成绩也不会太好；机组实验成绩一般的同学，他的网络综合实验成绩也比较一般；大学物理课程比较好的同学，他的概率与统计成绩也会比较不错.

通过这些关联规则的发掘，可以根据单个学生现有成绩，挖掘出学生个体后续所修课程可能出现的结果。通过对学生成绩进行分析，学生可以查看各科成绩之间关系，通过输入某些课程的考试成绩，将这些考试成绩作为关联规则的前件，依据规则库中的规则，为学生提供后续课程的预期成绩，以使该学生在以后的学习过程中，了解哪些科目是其弱势课程，从而巩固加强对薄弱科目的学习。

同时，教师及教学管理者可以分析在出现某种结果的情况下，是什么原因导致该问题的出现，抓取所有可以导致该后件出现的前件，抓取出该类学生，可以在以后的教学中进行更良好的分班，有利于老师根据学生的不同学习程度教授课程，达到因材施教的效果，从而有助于提升教学质量。

# 项目总结

将数据挖掘技术应用于学校的教学、管理中，对提高学校教学管理水平有很好的指导作用，而且采用先进技术对考试过程和教学环节中产生的数据进行多层次、多角度的分析，利用分析结果辅助教学决策是保证教学质量、提高学生 素质、保障学生就业的必然要求。

数据挖掘作为种新兴的数据分析技术，其研究取得了令人瞩目的成就。已经成功地应用到了许多领域。目前，还没有将其应用到教育领域的成功实例出现，而在教育信息化的大趋势下，数据挖掘技术应用于教育领域内的数据分析，必然会有十分广阔的前景。本文分析了由于我国大众化高等教育的普及，高校学生数量的激增，给高校教学工作所带来的诸多影响教学质量的问题，提出了将数据挖掘技术应用于高校教学研究中的一种新思路，并提出了依照这一新思路的实施方案。 本文着重研究数据挖掘关联规则算法，通过对常用算法的比较研究，总结了它们的长处和不足。本文利用Apriori算法实现了关联规则在学生成绩分析中 的应用。在挖掘之前，对数据库进行了数据清理、数据转换、数据消减等数据 预处理，处理了空缺数据、将连续值属性离散化，为进一步挖掘打好基础。

通过本文的研究，初步实现了数据挖掘技术在高校教学中的应用。但是，仍然存在着许多问题需进一步研究：

1)在数据预处理方面还不够完善，还需要依靠其它数据库工具人工完成。

2)课程开课的先后顺序需要手动去筛选。