****

**数据挖掘课程实验报告**

学 院： 金融学院

课程名称： 数据挖掘

教学班级： A01

姓 名： 李茂源

学 号： 0172841

实 验 报 告

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程名称 | | 数据挖掘及财经应用 | | 教学班级 | | A01 | | 指导老师 | | 江腾蛟 |
| 学号 | 0172841 | | 姓名 | | 李茂源 | | 行政班级 | | 金融172 | |
| 实验项目 | | 实验1：数据挖掘流程初探—基于Drug数据集 | | | | | | | | |
| 组员名单 | | 独立完成 | | | | | | | | |
| 实验类型 | | ■操作性实验 □验证性实验 □综合性实验 | | | | | | | | |
| 实验地点 | |  | | | 实验日期 | | | 可写起止时间段 | | |

# 实验目的和要求：

**目的：**

(1)初识数据挖掘与数据分析；

(2)发现数据挖掘与数据分析的趣味性。

**实验数据集介绍：**

数据文件(DRUG.TXT)以文本文件形式存放.

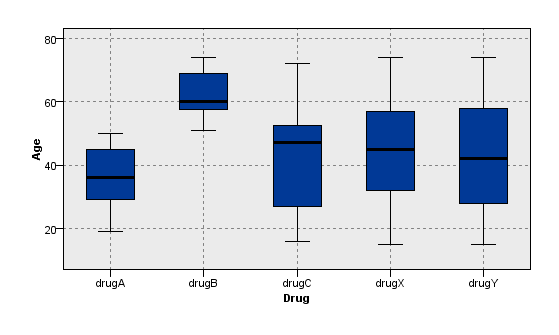
这是一份关于药物研究的数据。大批患有同种疾病的不同病人，服用五种药物中的一种（Drug A，Drug B，Drug C，Drug X和Drug Y）之后都取得了同样的治疗效果。现需要利用数据挖掘发现以往药物处方适用的规律，给出不同的临床特征病人更适合服用哪种药物的建议，为未来医生开具处方提供参考。

案例数据是随机挑选的部分病人服用药物前的基本临床检查数据，包括：病人的年龄（Age）、性别（Sex）、血压(BP）、胆固醇（Cholesterol）、唾液中钠元素（Na）、钾元素（K）含量和服用药物（Drug）。

**实验要求：**

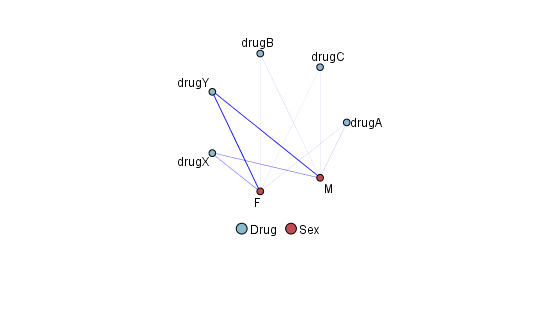
1. 读入数据源文件，输出表以查看具体数据内容并将表名重命名为Drug；
2. 添加类型节点，读取值并将Drug的角色设为目标，其余属性角色均为输入；
3. 添加数据审核节点，观察基本统计量与高级统计量下审核内容的区别，观察质量标签下各个选项的含义；
4. 分别观察各个输入变量与目标变量之间的相关性；

（4.1）Age与Drug



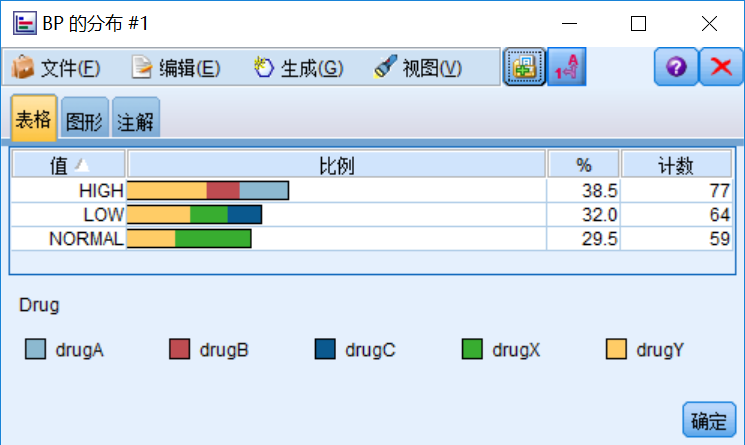
上图，你有什么发现？

（4.2）Sex与Drug



这个图呢？

（4.3）BP与Drug



这个图呢？

（4.4）……

1. 利用特征选择进行特征的重要性判定。
2. 建模，利用C5.0构建分类树并解读模型结果；比较特征选择与C5.0对特征重要性的判断是否一致。
3. 生成新属性Na/K
4. 分析Na/K与Drug的关系
5. 过滤节点去掉Na和K属性，因为已生成Na/K新属性。
6. 再次类型节点，读取值，设置输入与目标角色。
7. 利用特征选择进行特征的重要性判定。
8. 建模，利用C5.0构建分类树并解读模型结果；比较特征选择与C5.0对特征重要性的判断是否一致。
9. 利用分析节点，对比前后两个模型的正确率。

说明：提交一份尽你目前所能的详尽分析报告。

By the way,既然是分析报告，图文并茂肯定是最好的，对每张图都要有编号、有解读！

# 实验过程

## 数据审核观察数据

对drug字段进行数据审核如图所示

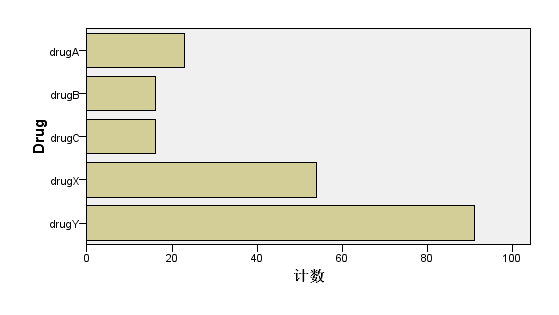


图 ‑1

如图2-1所示：

* 使用X,Y药物的人数显著多于使用A,B,C药物的人数。

## 分析相关性

### 分析age和drug的关系

首先对age字段绘制直方图，使用drug作为交叠字段分析

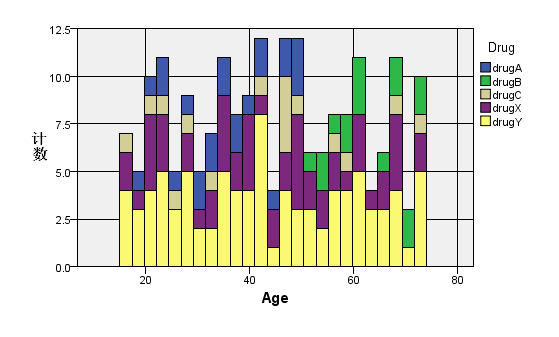


图 ‑2

如图2-2所示：

* 使用drugB的人都为在50岁以上的人
* 使用drugA的人都为50岁以下的人
* drugX和drugY对与所有年龄的人都适用。

为了直观表现年龄和药物的分布关系，绘制箱图进行分析

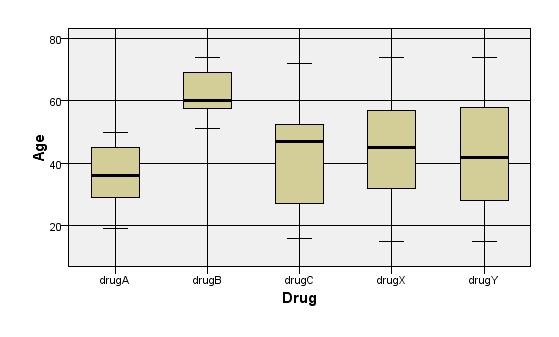


图 ‑3

如图2-3所示：

使用drugB的人，基本分布在高年龄。

为了进一步说明年龄和drug的相关性，使用C5.0分析，并对建模结果进行分析

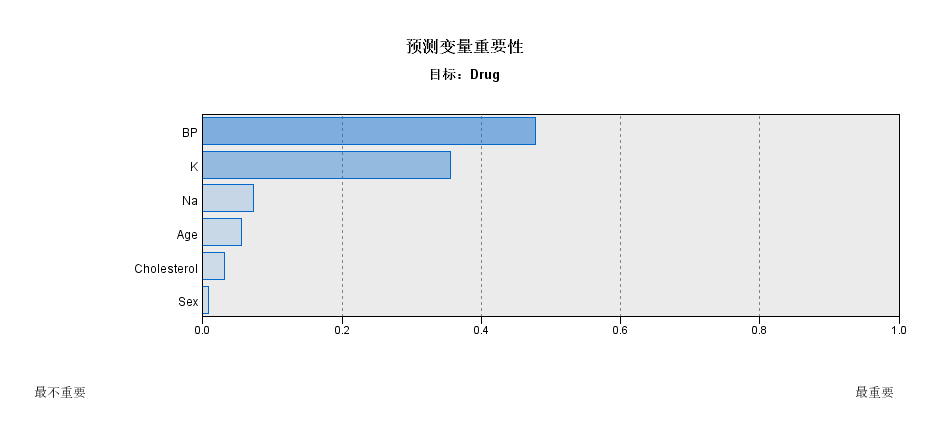


图 ‑4

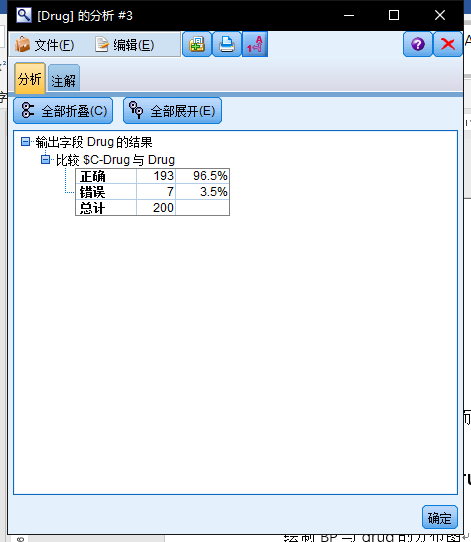


图 ‑5

如图2-5所示：

重要性最大的不是age，而是BP，其次是K。

### 分析BP与drug

绘制BP与drug的分布图

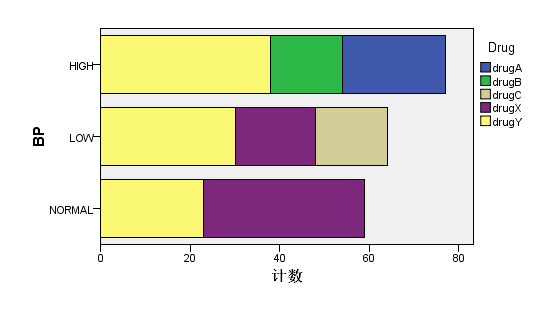


图 ‑6

如图2-6所示：

* drugY对于任何血压的人都适合
* drugB和drugA都用在血压高的人群
* drugC都用在低血压的人
* drugX只用在低血压和血液正常的人群
* drugX用在正常血压人群中，效果比较显著

用矩阵来分析BP和drug可以得到

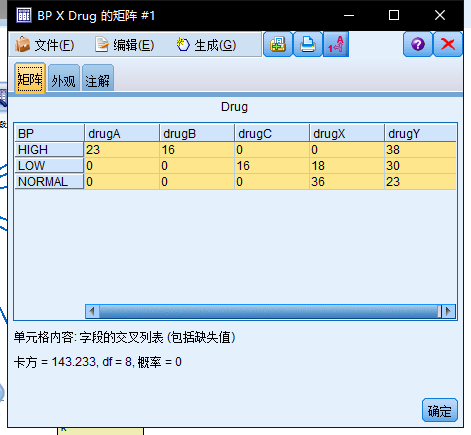


图 ‑7

图2-7说明药物和BP相关性显著

### 对K和drug进行分析

绘制直方图可得

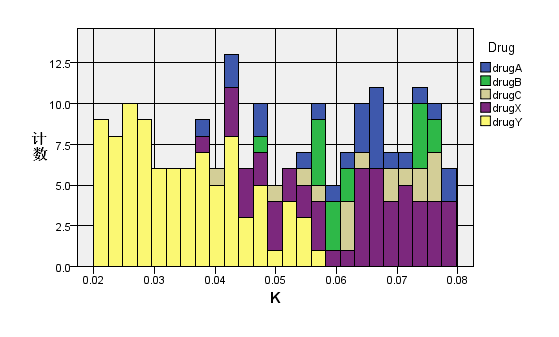


图 ‑8

如图2-8所示：

* 当K的含量少于0.038的时候，全部用drugY.

### Cholesterol和drug关系分析

利用矩阵输出Cholesterol和drug

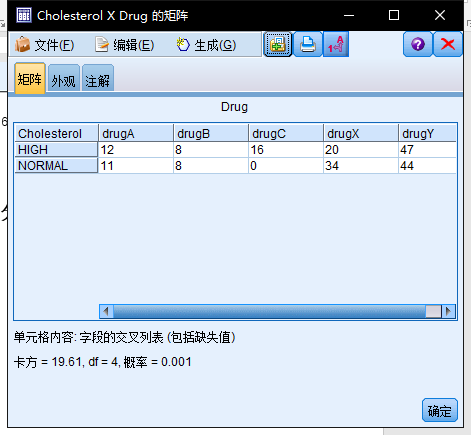


图 ‑9

图2-9说明drug和Cholesterol不相关的概率仅为0.001，否定原假设，drug和Cholesterol相关

进一步绘制直方图进行分析

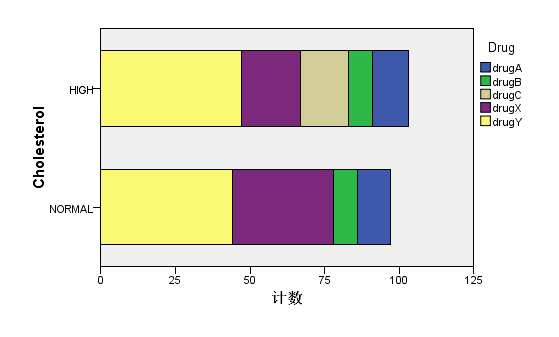


图 ‑10

如图2-10所示：

* 仅仅在高胆固醇时候才可用drugC

### Na和K对drug分析

先以x为Na，以Y为K，交叠字段用drug绘制散点图

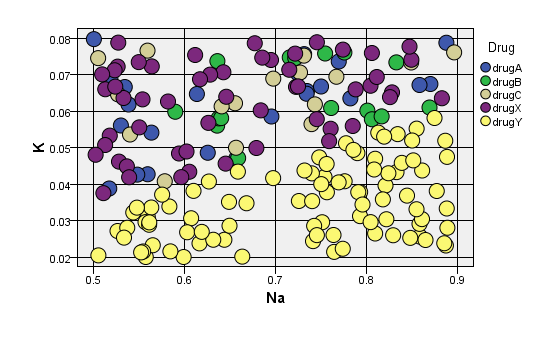


图 ‑11

导出字段Na/K，并利用直方图显示

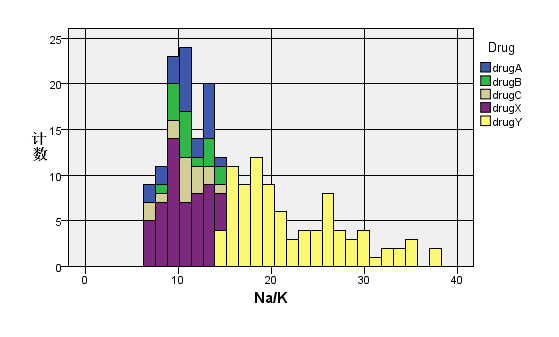


图 ‑12

如图2-12所示：

* Na/K的比例大于15.8时候，都用drugY

## 建立模型

通过分析，发现drug的使用与K,Na/k,BP,age,Cholesterol都有关，对这些量用C5.0模型进行建模，得到如下结果：

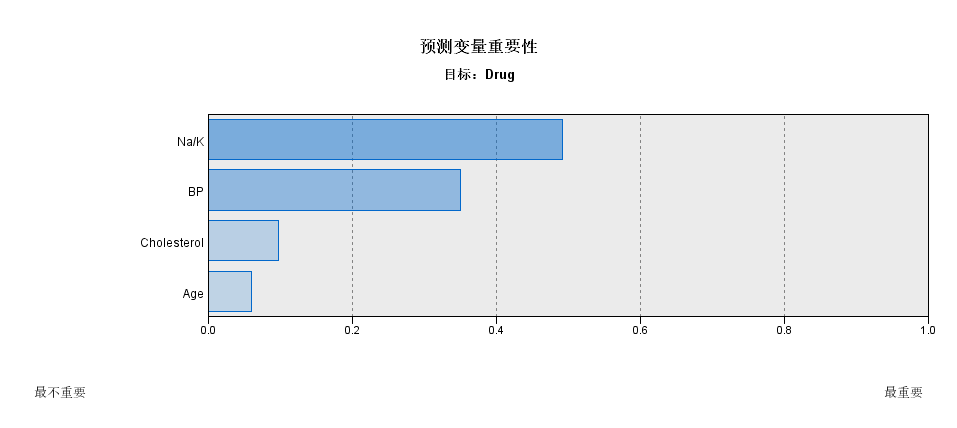


图 ‑13

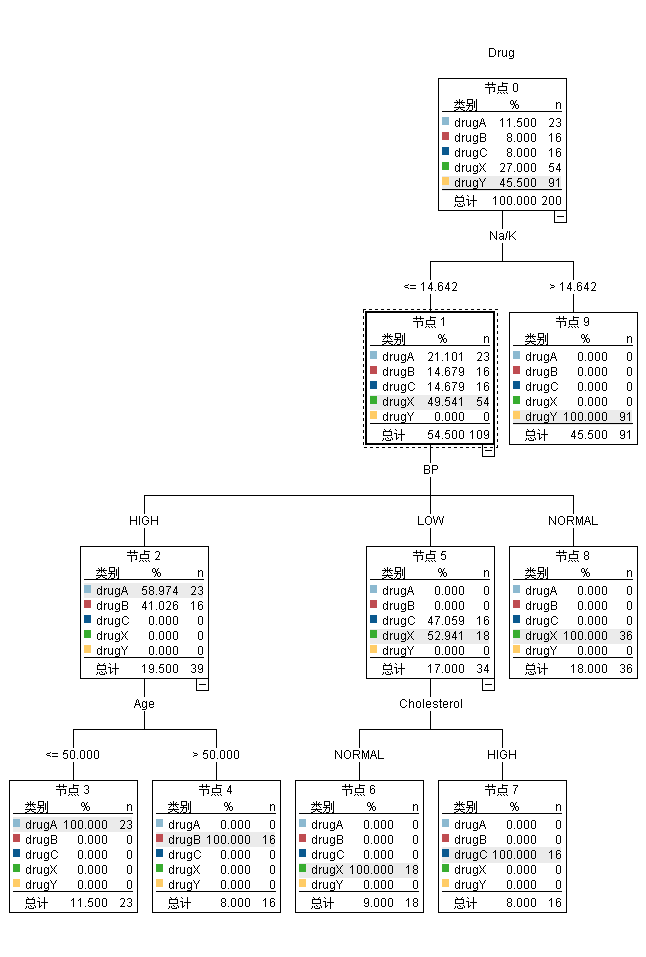


图 ‑14

对模型进行分析得出：

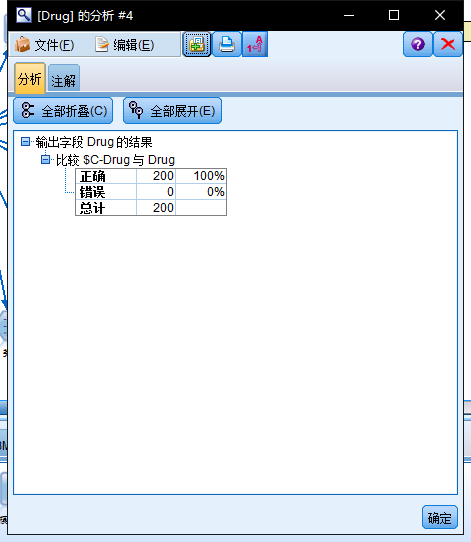


图 ‑15

正确提高到100%，存在样本全部正确。可能存在过拟合，但是由于是学习数据，数据不可清洗，也不存在抽样问题，不知道如何解决。

# 问题反馈与收获（3.1实验过程中遇到的问题和解决办法; 3.2 通过此次实验操作有哪些收获，包括操作技巧、心得总结等均可记录下来。）

## 问题与解决办法

* 刚上手软件不太熟练，还需要反复练习才可以解决这个问题。
* 对于分析方法掌握不够，理论知识不够完善，完成报告时候，还需要不断查阅理论资料。
* 对数据的敏感性不够，找不出合理的分析方法。
* 报告逻辑一开始不是很清洗，需要反复修改

## 发现与收获

* 学会一些软件基本操作。
* 理论知识有所提升，比如懂得如何看矩阵，如何利用c5.0决策。

# 指导教师评语及成绩：

评语：

成绩： 指导教师签名：

批阅日期：月 日