****

**学生实验报告**

学 院： 信息管理学院

课程名称： 数据挖掘

教学班级： B01

姓 名： 李茂源

学 号： 0172841

实 验 报 告

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程名称 | | 数据挖掘及财经 | | 教学班级 | | B01 | | 指导老师 | | 江腾蛟 |
| 学号 |  | | 姓名 | |  | | 行政班级 | |  | |
| 实验项目 | | 分类预测的综合案例—基于Titanic数据集 | | | | | | | | |
| 组员名单 | | 2-3人一组 | | | | | | | | |
| 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 | | |

目 录

[实 验 报 告 2](#_Toc4963263)

[目 录 3](#_Toc4963264)

[1 实验目的和要求： 4](#_Toc4963265)

[2 实验过程（记录实验步骤、分析实验结果） 6](#_Toc4963266)

[2.1 数据预处理 6](#_Toc4963267)

[2.1.1 了解数据 6](#_Toc4963268)

[2.1.2 数据清洗 7](#_Toc4963269)

[2.2 各字段分析 9](#_Toc4963270)

[2.2.1 age和survived 9](#_Toc4963271)

[2.2.2 sex和survived 10](#_Toc4963272)

[2.2.3 Pclass与survived 11](#_Toc4963273)

[2.2.4 Pclass和Fare 11](#_Toc4963274)

[2.2.5 Cabin字段探索 12](#_Toc4963275)

[3 问题反馈与收获（实验过程中遇到的问题、自己的解决办法，对没解决的问题，自己的思路。） 13](#_Toc4963276)

[3.1 问题与解决办法 13](#_Toc4963277)

[3.2 发现与收获 14](#_Toc4963278)

[四、指导教师评语及成绩： 14](#_Toc4963279)

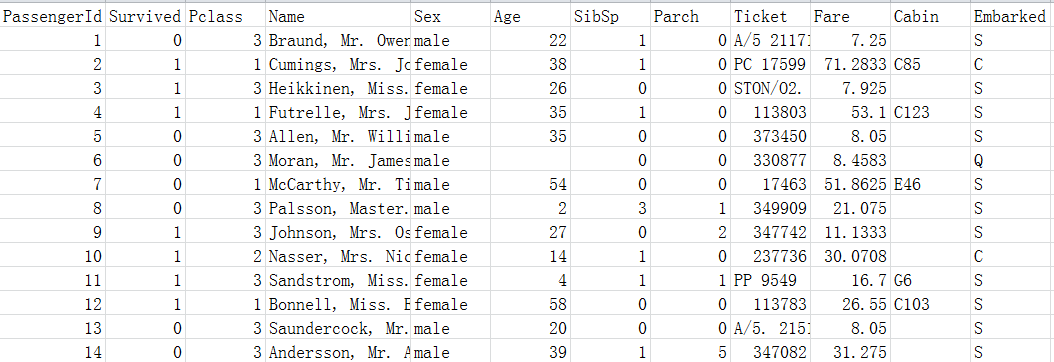
# 实验目的和要求：

1. 实验目的

泰坦尼克号事故相信大家都知道，这次的实验任务就跟这次事故的数据集有关。

泰坦尼克号问题之背景，就是那个大家都熟悉的“Jack and Rose”的故事，豪华游艇沉没了，大家都惊恐逃生，可是救生艇的数量有限，无法人人都有，副船长发话了“lady and kid first！”，但最后哪些人获救了呢？我们收集到的训练和测试数据是一些乘客的个人信息以及存活状况，要尝试根据它生成合适的模型并预测其他人的存活状况。这是一个二分类问题。

下表是泰坦尼克号船上人员数据，表中包含12个字段信息，其意义如下：



PassengerId => 乘客ID

Pclass =>乘客等级(1/2/3等舱位)（属性代表船舱等级，1-一等舱，2-二等舱，3-三等舱，从一定程度上反应了这个乘客经济情况和社会地位。）

Name =>乘客姓名

Sex =>性别

Age =>年龄

SibSp =>堂兄弟/妹个数

Parch =>父母与小孩个数

Ticket =>船票信息（字母与数字具体代表什么信息，需要猜测分析判断）

Fare =>票价

Cabin =>客舱

Embarked =>登船港口

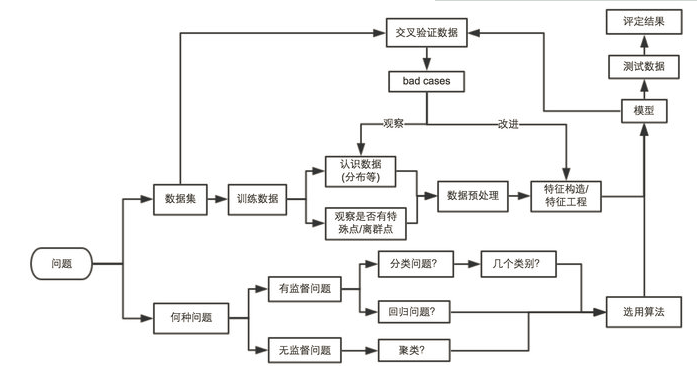
Survived=>乘客是否获救

（友情提示：这是竞赛组织方提供的原始数据集，小心处理，有可能你在读入数据集时就会抓狂！）

请根据titanic\_train.csv数据集，利用多种分类模型(至少三个)预测乘客是否获救与幸存，并比较各个模型的性能，并能在预测时根据模型融合给出最佳预测结果。报告中要体现详细的分析过程和分析理由，并截取相应的实验步骤与实验结果图。

1. 实验要求

机器学习一般工作流程如下图所示。



**1）分析问题性质，**这是一个分类还是聚类问题，若是分类问题，分析属性集和目标类，选定可适用的算法有哪些？（也可根据算法调整属性集中的属性类型，如连续变离散等，解决问题的方法不止一个。）

**2）理解数据和认识数据，完成数据的初步探索。**

(1) 缺失值处理：针对age、Carbin和Emarked三个属性上的缺失值根据属性类型和数据理解进行相应的缺失值处理，给出理由；

(2) 观察与统计乘客的各个属性与获救之间的关系。下面初步给出了5条可能存在的相关性，除此之外，同学们还可以列举其他相应属性与获救之间的相关性分析。

A.乘客中获救人数与遇难人数比例；

B.乘客各个等级的分布状况以及与获救的关系；

C.遇难与获救人员年龄、性别的关系；

D.乘客各个登船口岸的人数分布以及与获救之间的关系；

E.堂兄弟/妹，孩子/父母有几人，对是否获救的影响。

**3）根据1）步和2）步对数据的观察及打算采取的分类算法，对属性进行处理生成特征属性（生成适合模型算法需要的数据表）。**

可能包括的操作有：

(1)属性变换（如定性属性、定量属性间的变换）

(2)属性类型变换（如连续类型变分类类型、数据的规范化等）

(3)特征生成（根据现有属性导出新属性、分类类型转换为二元标志类型等）

(4)特征选择与主成分分析

**4) 构建分类模型，并进行模型的性能比较。**

注意：没有绝对的机器学习算法优劣，也没有绝对的高性能的机器学习算法，只有在特定的场景、数据和特征下更合适的机器学习算法。因此构建各个分类模型，重在分析其适应性。

**5）模型融合。**

模型融合的大意就是：当我们手头上有一堆在同一份数据集上训练得到的分类器(比如C5.0，C&R树, 贝叶斯网络, SVM，神经网络等)，我们让他们都分别去做判定，然后对结果做投票统计，取票数最多的结果为最后结果。模型融合可以比较好地缓解训练过程中产生的过拟合问题，从而对于结果的准确度提升有一定的帮助。在单个模型上也可以做融合，这就是Bagging和boosting技术。

（友情提示：这是kaggle比赛中的Titanic预测生存情况，大家可以网上查找相关的资料，借鉴前人的经验会让你的工作更有头绪和轻松些！）

# 实验过程（记录实验步骤、分析实验结果）

## 数据预处理

### 了解数据

首先对数据进行大致了解和分析，进行数据审核，发现有缺失值，等会处理。



图 ‑1

### 数据清洗

#### 年龄缺失值填充

##### 使用中位数填充

因为数据中有很多缺失值，比较重要的有age，缺失了117例。尝试用statistic节点对age进行输出分析

|  |  |
| --- | --- |
| 计数 | 714 |
| 平均值 | 29.679 |
| 最小值 | 0 |
| 最大值 | 80 |
| 范围 | 80 |
| 方差 | 211.309 |
| 标准差 | 14.536 |
| 平均值标准误差 | 0.544 |
| 中位数 | 28 |

图 ‑2

如图2-2所示：

titanic号上年龄集中分布在30左右，范围跨度从0到80

暂时使用中位数进行填充

##### 使用决策树进行填充

年龄与称呼肯定有关系，比如Master主要是14岁一下的男孩的称呼，可以稍微寻找一下相关性，使用正则表达式提取称呼得到Call字段，然后利用直方图查看两者之间的关系。

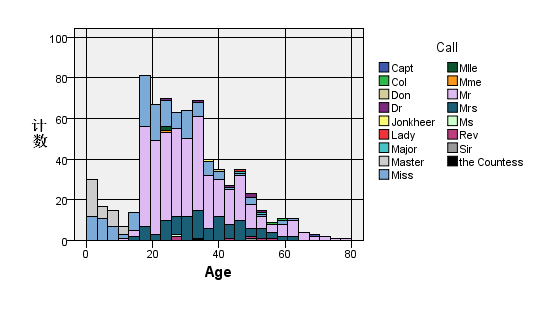


图 ‑3

如图2-3所示：

* Master分布主要是在14岁以下。
* 14岁以后男性以Mr称呼比较多。
* Miss和Mrs的分布范围很广，或许还需要用Pclass和Parch结合进行分析

只用称呼来判断年龄显然远远不够，需要结合Pclass和Parch这两个字段来确定年龄所在的区间，首先把年龄进行等宽分箱，以13为区间长度分为6箱，使用过滤器筛掉一些无用节点，做出C5.0决策树。

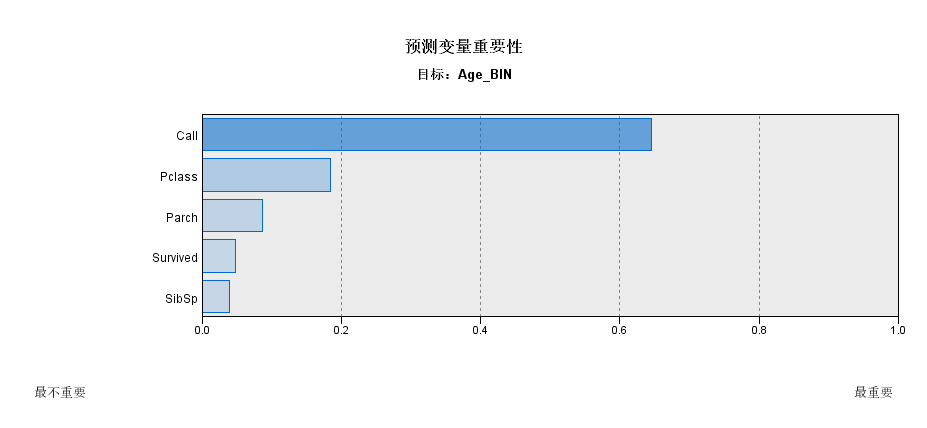


图 ‑4

如图2-4所示：

* + 这里重要性第一是称呼Call,重要性为0.65，其次是Pclass，再来就是Parch。使用分析节点对这个决策树进行分析，发现正确率仅仅有45.%，所以次方法还不能使用。放弃此方法。

## 各字段分析

### age和survived

使用直方图，我们可以直观看到年龄与生存之间的关系

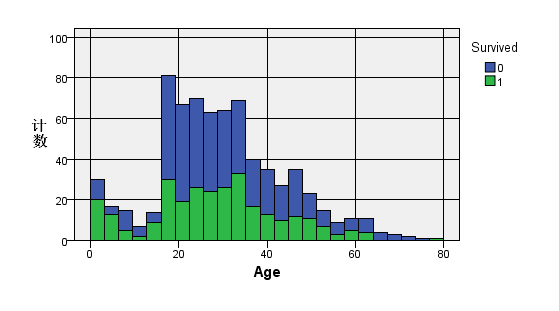


图 ‑5

如图2-5所示：

* 查资料知道泰坦尼克号事件中生存者更少，所一1代表生存而0代表死亡
* 62岁以上老人无人生还
* 18岁以下，也就是未成年人生还比例较高

### sex和survived

资料中有提到，船长说，让小孩和女人先上，可以对性别分布与是否生存进行分析

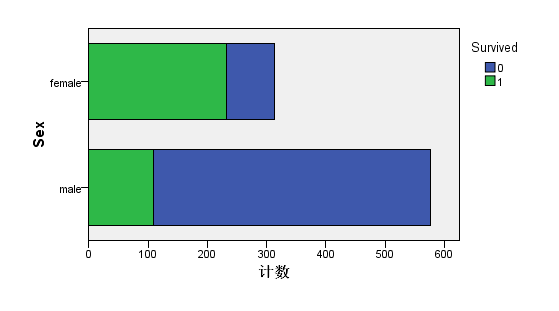


图 ‑6

如图2-6所示：

* 明显可以看出女性生还的比例显著高于男性生还的比例，能否存活与性别有很强的相关性

### Pclass与survived

在当时英国资本主义时期，等级制度森严，有可能存在等级高的先上救生艇的情况，使用矩阵和分布图对这两个字段进行分析

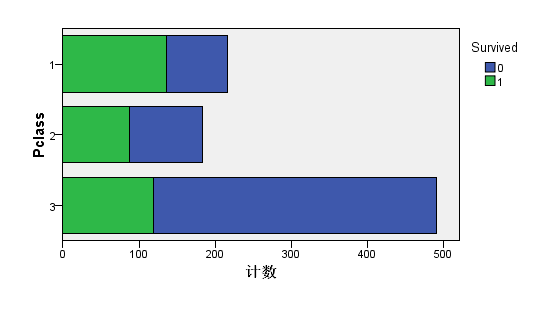


图 ‑7

如图2-7所示：

* Pclass为1和2的存活比例显著高于3的比例，可以考虑Pclass为1是地位较高的人
* 使用矩阵同样能得到两个字段有相关性

### Pclass和Fare

Fare是一个连续字段，在之后有可能要进行分箱处理，票价和等级可能会有一定关系，可以用来分箱或者预测，现在使用直方图寻找一下Pclass和Fare的相关性

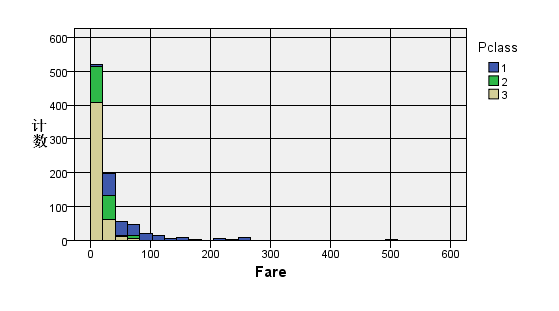


图 ‑8

如图2-8所示：

* + 这里的票价越高，明显Pclass为1的乘客比例越大，而票价低的明显等级3的比例大，进一步验证Pclass为1的乘客地位更高

### Cabin字段探索

Cabin是缺失值非常多的字段，进一步挖掘原因，可能是有些没有舱位的可能是散舱，也就是说有Cabin属性的人很有可能是地位较高的人，所以我们导出一个新字段CabinH，值1代表有舱位，值0代表没有舱位，将图标准化，探索一下CabinH和Pclass的关系。

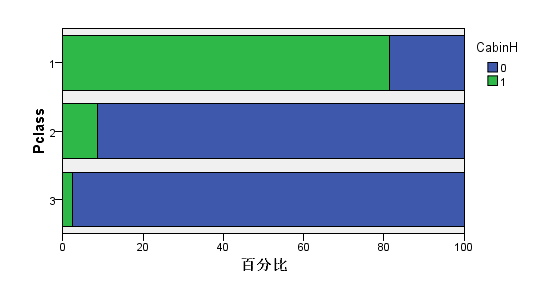


图 ‑9

如图2-9所示：

* 明显Pclass为1的人群含有舱位的比例更大，也就是说，有舱位的人，很有可能地位更高，有更高的存活率

所以进一步分析CabinH与Survived的关系，使用分布图标准化后得到Cabin和Survived的关系。

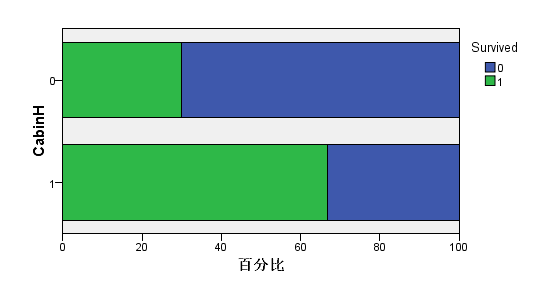


图 ‑10

如图2-10所示：

* 有舱位的人存活率明显高于无舱位的，所以以后建模要将这个变量考虑进去。

# 问题反馈与收获（实验过程中遇到的问题、自己的解决办法，对没解决的问题，自己的思路。）

## 问题与解决办法

Q：称呼Call字段难以导出，难以解决。

A：使用excel的正则表达式接口，编写VBA代码，写出正则表达式的函数，使用正则表达式提取出称呼。

Q：年龄分箱处理不会。

A：找到了分级字段，可以对年龄进行分箱。

Q：分箱的时候不知道用多少宽度得到的准确性高。

A：一个一个实验，采取相关性最高的进行分箱。

Q：不知道如何利用软件对字段进行填充。

A：查阅资料，找出填充字段，选择填充条件，就可以对字段进行填充。

## 发现与收获

* + - 知道了如何分箱，如何进行字段填充，熟练了软件。
      * 各类图形的使用明显更加熟练，能够很好对数据属性进行把握，掌握好数据类型。
    - 在作图的时候，选择按照颜色标准化，可以按照百分比显示
      * 使用导出节点的空值表示就是””。和EXCEL是一样的，可以使用节点导出CabinH

四、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