****

**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：

学 号：

实 验 报 告

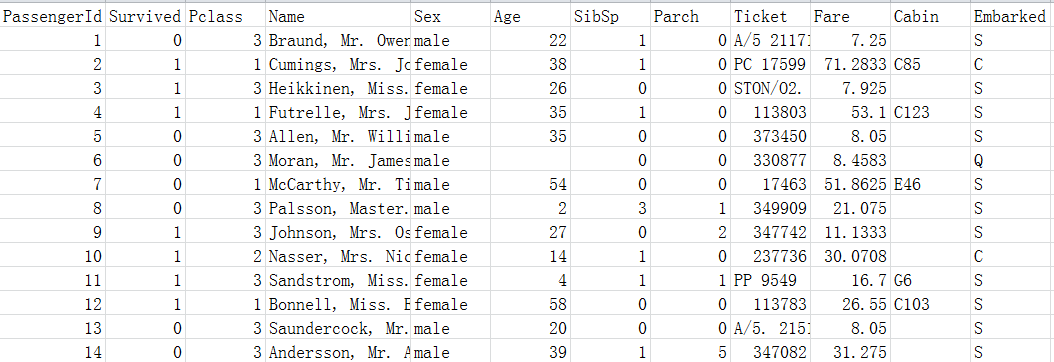
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程名称 | | 数据挖掘及财经 | | 教学班级 | | B01 | | 指导老师 | | 江腾蛟 |
| 学号 |  | | 姓名 | |  | | 行政班级 | |  | |
| 实验项目 | | 分类预测的综合案例—基于Titanic数据集 | | | | | | | | |
| 组员名单 | | 2-3人一组 | | | | | | | | |
| 实验类型 | | □操作性实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

1. 实验目的和要求：
2. 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

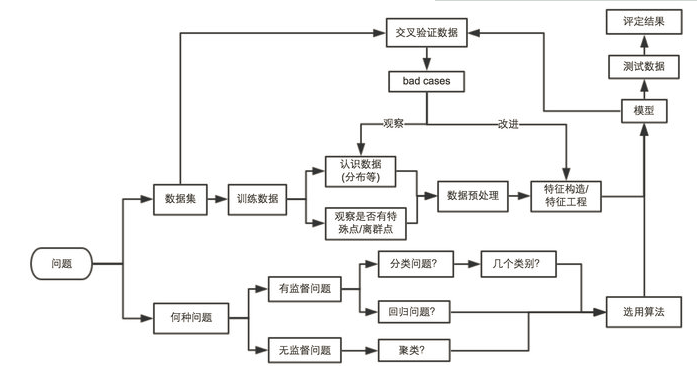
Survived=>乘客是否获救

（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！）

二、实验过程（记录实验步骤、分析实验结果）

2.1 XXXX（按数据挖掘过程分步骤描述）

2.2 XXXX

三、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。）

3.1 问题与解决办法

3.2 发现与收获

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