关联规则挖掘

崔绿叶 2120150981

**1、数据集的选择及处理**

本次实验选择的数据集为titanic数据集，数据来源为<https://www.kaggle.com/c/titanic/data>。数据集内包含PassengerId，Survived，Pclass，Name，Sex，Age，SibSp，Parch，Ticket，Fare，Cabin，Embarked十二项，分别代表旅客ID，是否存活，客舱类型，姓名，性别，年龄，孩子数目，父母数目，船票号，费用，客舱号，上船口。实验只选取了其中的6种类型的数据：Survived，Pclass，Sex，Age，SibSp，Embarked；并使用Excel对数据做了以下处理：

（a）Survived类型中，使用“survived”代替数据1，使用“dead”代替数据0。

（b）Pclass类型中，使用“first”代替数据1，使用“second”代替数据2，使用“third”代替数据3。

（c）Age类型中，年龄小于10的为“kid”，大于10小于20的为“teenager”，大于20小于40的为“youth”，大于40小于60的为“middle-ager”，大于60的为“elder”。

（d）SibSp类型中，有后代的为“hasSpring”，没有后代的为“noSpring”。

处理后的数据如下图所示：



**2、找出频繁项集**

本次实验使用R语言，找出频繁项集的代码如下：



可以看到频繁项集的前十名，如下图：



从结果中我们可以看到船上大部分人都没有后代，65%的人为男性，撞冰山之后的死亡人数大约为六成等。{dead, third}的支持度为0.4175，我们可以合理地猜测三等舱的位置在事故发生时并不利于逃生，或者在救助的过程中一二等舱的乘客有优先上救生船的优势。

**3、导出关联规则**

导出关联规则的代码如下：



关联规则按支持度排序的前十名如下图：



关联规则按支持度排序的前十名如下图：



从中可以看出如果大于60岁的死者一定为男性，可以合理地猜测大于60岁的女性会被优先救助。二等舱的儿童一定会存活则可能说明儿童有优先上救生船的权利，但对三等舱的儿童则没有类似规则，说明救助的过程中等级因素也很重要。

**4、对规则进行评价**

这里使用lift指标对规则进行评价，代码如下：



关联规则按lift指标排序的前十名如下图：



从中可以得到比较有用的规则。如，二等舱存活下来的男性有一半概率小于10岁等。

**5、可视化**

由于matlab和java中没有可视化的包，所以采用R实现，使用扩展包arulesViz。



**图8 散点图**

从图中可以看出，高lift对应低support。



**图9 散点图**

从图中看出，很多规则都有高lift。



**图10 泡泡图**



**图11 平行坐标图**