**Titanic存活数据关联规则挖掘**

1. **数据描述**

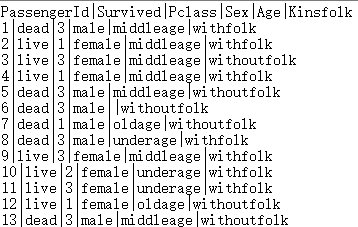
原始数据下载自<https://www.kaggle.com/c/titanic/data>。原始数据（train.csv）包含PassengerId，Survived，Pclass，Name，Sex，Age，SibSp，Parch，Ticket，Fare，Cabin，Embarked这12个字段。Pclass代表座舱级别(分为一等，二等和三等)。SibSp代表是否有兄妹和配偶，Parch代表是否有父母和孩子，Embarked为上船的港口。

数据共有891条记录。

1. **数据预处理**

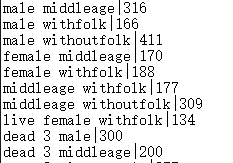
将数据处理成可用于关联数据挖掘的形式。其中Name，Ticket,Fare，Cabin以及Embarked属于乘客的一些个人信息，没有多大的挖掘价值。剩余字段中，SibSp和Parch表示亲属的个数。因此，将其合并成一个字段Kinsfolk(withfolk，withoutfolk)，用于表示是否有亲属，其余字段PassengerId，Survived，Pclass，Sex，Age保留。PassengerId用来表示乘客，训练集中共891名；Survived表示乘客是否存活，有live和dead两个属性，表示存活和死亡；Pclass表示乘客座舱等级，分为1、2、3三等；Sex有famale和male，Age根据年龄分为三段，18岁以下为underage，19-45岁为middleage，45岁以上为oldage。

共六个字段：



1. **频繁项集**

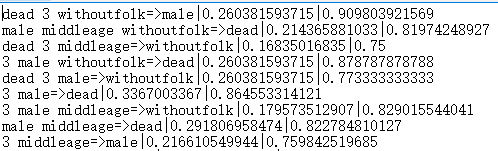
实验中测试了多组支持度和置信度，选择了其中两组minSup=0.15,minConf=0.75和minSup=0.20,minConf=0.75，获得的频繁项集列表存放在freItemSetList15\_75.txt和freItemSetList20\_75.txt中。



数据中后面的数字代表该频繁项出现的频度。

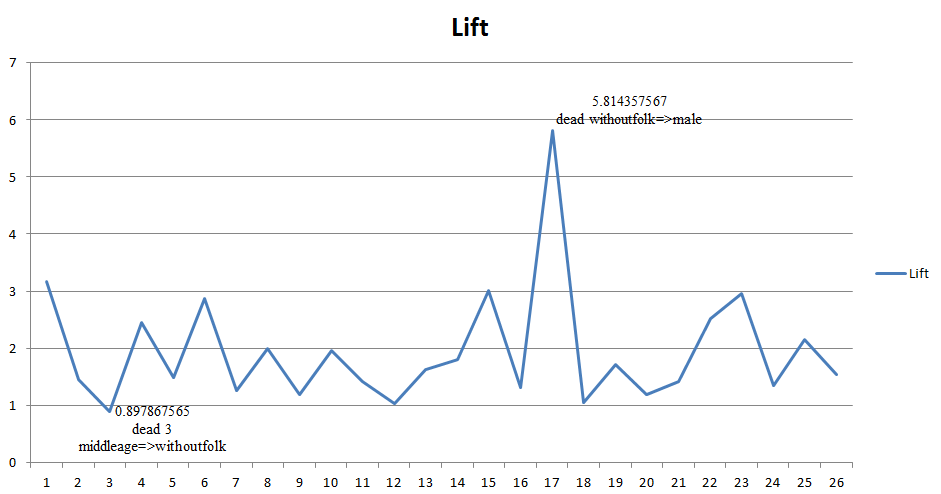
1. **关联规则（去除冗余后）**

生成关联规则时，置信度不低于0.75。获得的频繁项规则存放在freItemSetRules15\_75.txt和freItemSetRules20\_75.txt中。



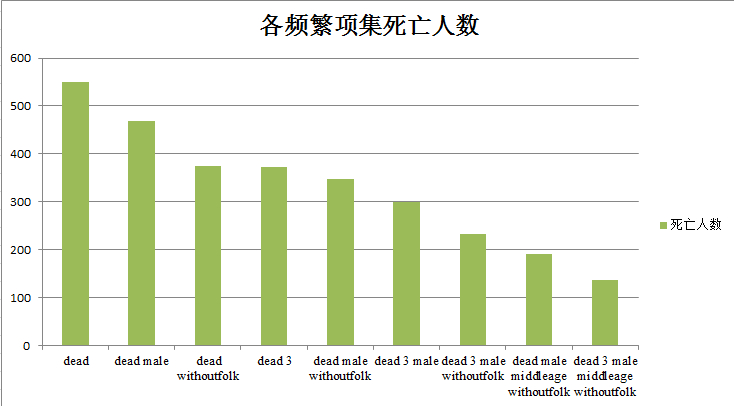
其中第二列数字表示该支持度，第三列表示置信度。

1. **规则评价（Lift）**

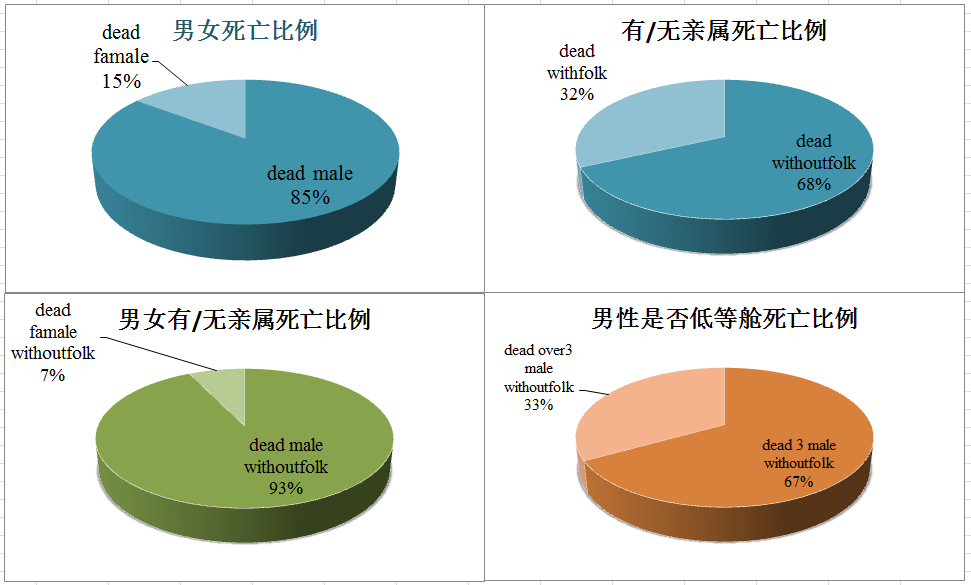


用Lift值对生成的规则进行评价，其中Lift值最低的规则为dead 3 middleage=>withoutfolk，为0.897868，也是唯一一个小于1的规则，说明该规则与规则后键呈负相关。最高Lift值的规则是dead withoutfolk=>male，为5.81436，说明该规则较为合理。

1. **可视化分析**

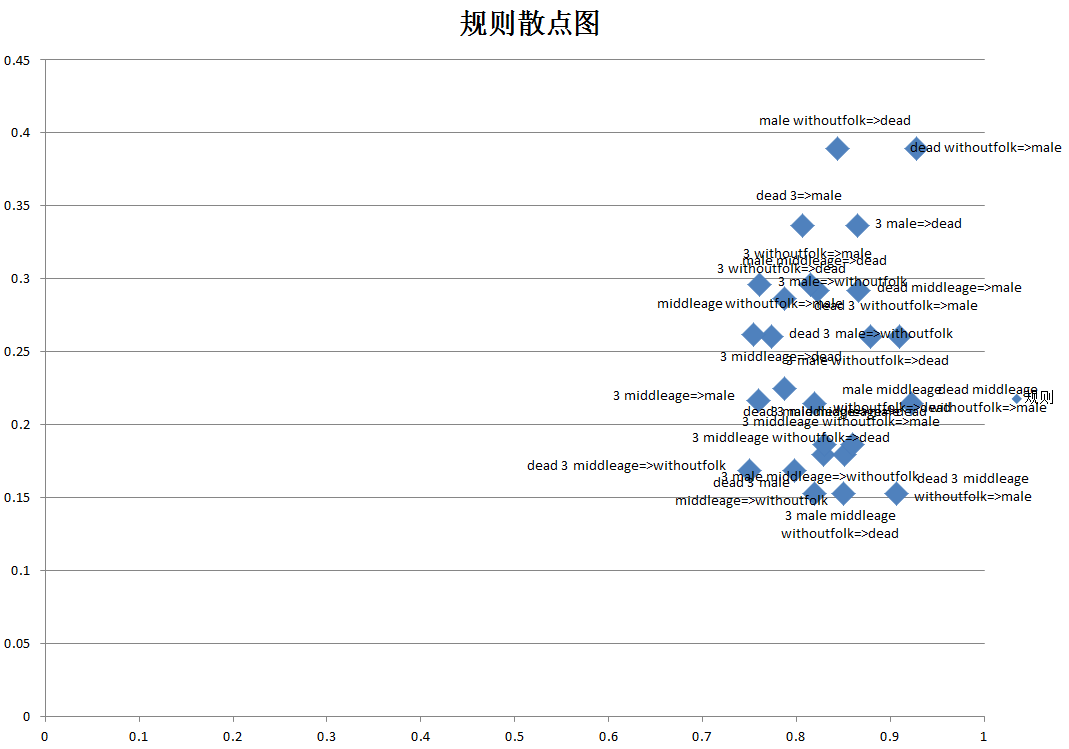


在此次实验中发现了比较有趣的现象，就是单身中青年穷男性遇难的可能性更高，为了证实这个结论，统计了一下几个比例。



首先死亡的单身男性占比高达93%，而又是低等舱(一定程度上说明穷。。。)的死亡比例63%。这些比例都证明了该事实。

而且，3 male middleage withoutfolk=>dead该规则的置信度高达0.85(支持度为0.153).说明买三等舱的中青年单身男性有85%的遇难了。。。



以上为规则散点图，选取了支持度为15%，置信度为75%，所以整体靠右。