数据挖掘课程作业之关联规则挖掘

计算机学院-2120151056-于畅泳

一.　实验要求

1.对数据集进行处理，转换成适合关联规则挖掘的形式；

2.找出频繁项集；

3.导出关联规则，计算其支持度和置信度；

4.去除冗余的规则；

5.对规则进行评价，可使用 Lift，也可以使用教材中所提及的其它指标；

6.使用可视化技术，如散点图、平行坐标、泡泡图等，对规则进行展示。

二.　数据说明

从以下 2 个数据集中任选一个：

UCI 的”急性炎症”数据集

Titanic 存活数据

本次实验采用数据集“Titanic 存活数据”，使用网站 https://www.kaggle.com/c/titanic/data 中的 train.csv 作为原始数据。

数据包含以下内容：

survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class

	(1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation
	(C = Cherbourg; Q = Queenstown; S = Southampton)

二. 分析报告及程序

实验环境：Microsoft Windows 10 家庭中文版 (64 位)

Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz(2592 MHz)

16.00 GB 内存

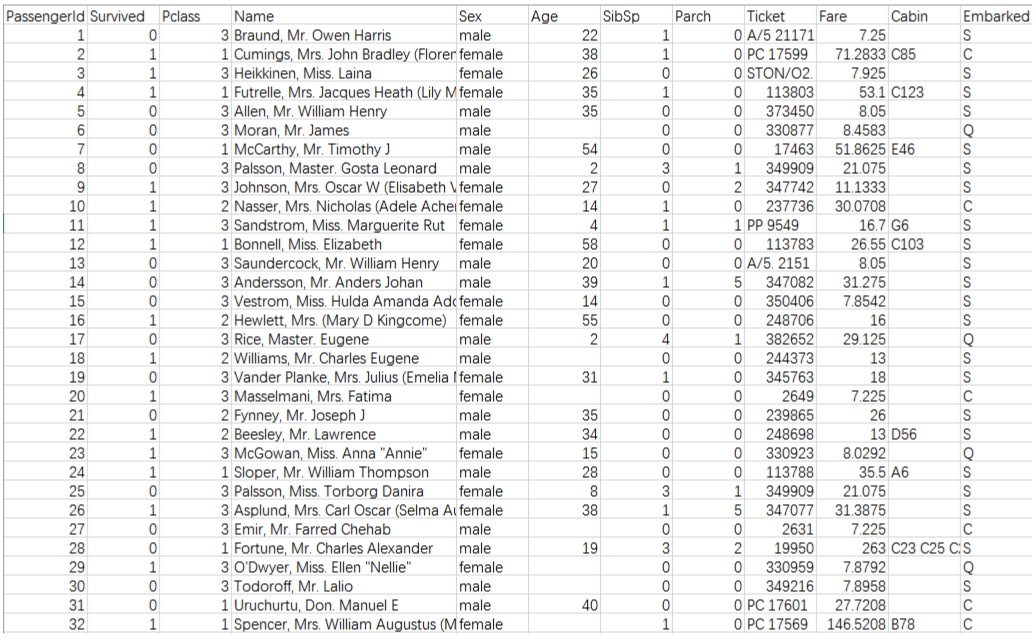
实验软件：pyCharm 5.0.2 (python 3.5.1)

SQL

实验内容：

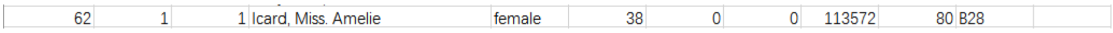
1. 数据集的选取及处理

数据集包含部分内容如下：

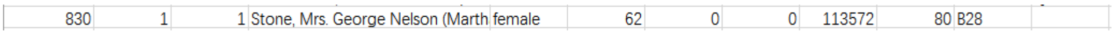


PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Floren	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth V	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achei	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Ad	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia I	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma A	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208		C
32	1	1	Spencer, Mrs. William Augustus (M	female		1	0	PC 17569	146.5208	B78	C

将数据转换成适合关联规则挖掘的形式，我们只留取survived，Pclass，Sex， Embarked四项数据，同时删除有缺失Embarked数据的第62行和第830行。

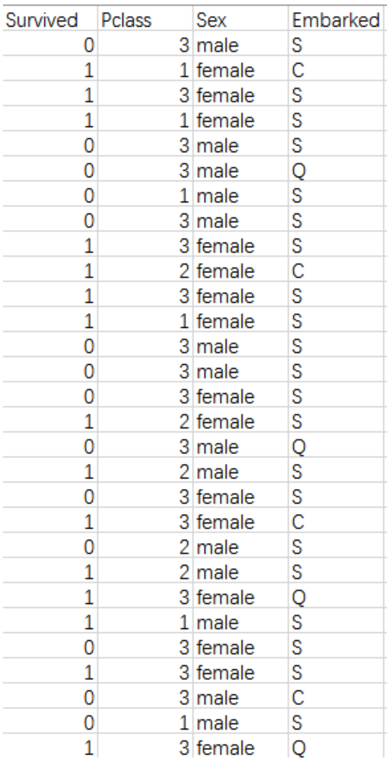


62	1	1	Icard, Miss. Amelie	female	38	0	0	113572	80	B28	



830	1	1	Stone, Mrs. George Nelson (Marth	female	62	0	0	113572	80	B28	

得到的数据如下（部分）：



Survived	Pclass	Sex	Embarked
0	3	male	S
1	1	female	C
1	3	female	S
1	1	female	S
0	3	male	S
0	3	male	Q
0	1	male	S
0	3	male	S
1	3	female	S
1	2	female	C
1	3	female	S
1	1	female	S
0	3	male	S
0	3	male	S
0	3	female	S
1	2	female	S
0	3	male	Q
1	2	male	S
0	3	female	S
1	3	female	C
0	2	male	S
1	2	male	S
1	3	female	Q
1	1	male	S
0	3	female	S
1	3	female	S
0	3	male	C
0	1	male	S
1	3	female	Q

2. 找出频繁项集，并计算支持度

所有项集（Items）：所有项目的集合。定义为：I。

项集（Itemset）：同时出现的项的集合。

支持度（Support）：

定义为 supp(X) = occur(X) / count(D) = P(X)。

置信度（Confidence/Strength）：

定义为 conf(X->Y) = supp(X ∪ Y) / supp(X) = P(Y|X)。

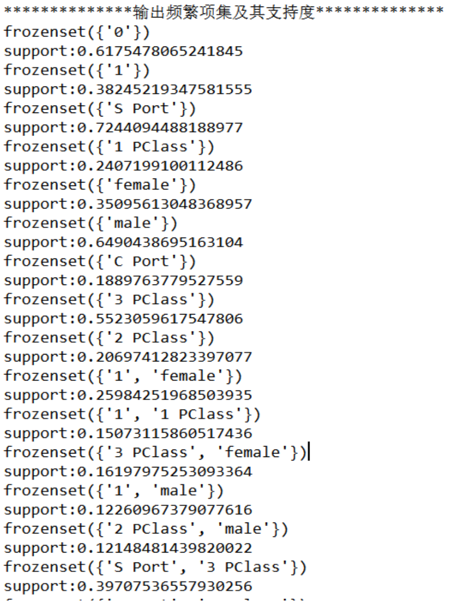
频繁集（Frequent itemset）：

支持度大于等于特定的最小支持度（Minimum Support/minsup）的项集。表示为 L[k]。注意，频繁集的子集一定是频繁集。

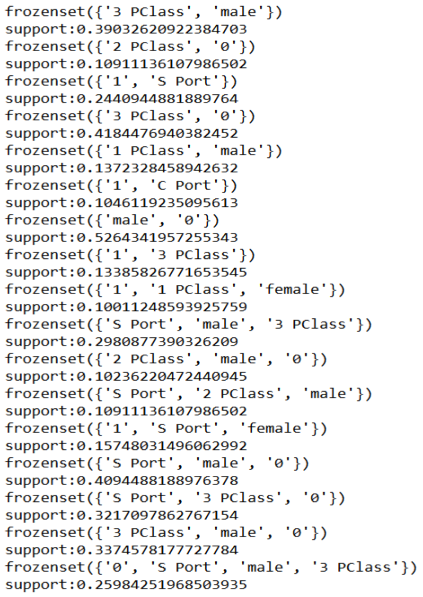
提升比率（提升度 Lift）：

定义为 lift(X -> Y) = lift(Y -> X) = conf(X -> Y)/supp(Y) = conf(Y -> X)/supp(X) = P(X and Y)/(P(X)P(Y))

通过设置支持度的大小（0.1），选择出频数满足条件的项，作为频繁项。计算满足条件频繁项的支持度。



```
**************输出频繁项集及其支持度**************
frozenset({'0'})
support:0.6175478065241845
frozenset({'1'})
support:0.38245219347581555
frozenset({'S Port'})
support:0.7244094488188977
frozenset({'1 PClass'})
support:0.2407199100112486
frozenset({'female'})
support:0.35095613048368957
frozenset({'male'})
support:0.6490438695163104
frozenset({'C Port'})
support:0.1889763779527559
frozenset({'3 PClass'})
support:0.5523059617547806
frozenset({'2 PClass'})
support:0.20697412823397077
frozenset({'1', 'female'})
support:0.25984251968503935
frozenset({'1', '1 PClass'})
support:0.15073115860517436
frozenset({'3 PClass', 'female'})
support:0.16197975253093364
frozenset({'1', 'male'})
support:0.12260967379077616
frozenset({'2 PClass', 'male'})
support:0.12148481439820022
frozenset({'S Port', '3 PClass'})
support:0.39707536557930256
```

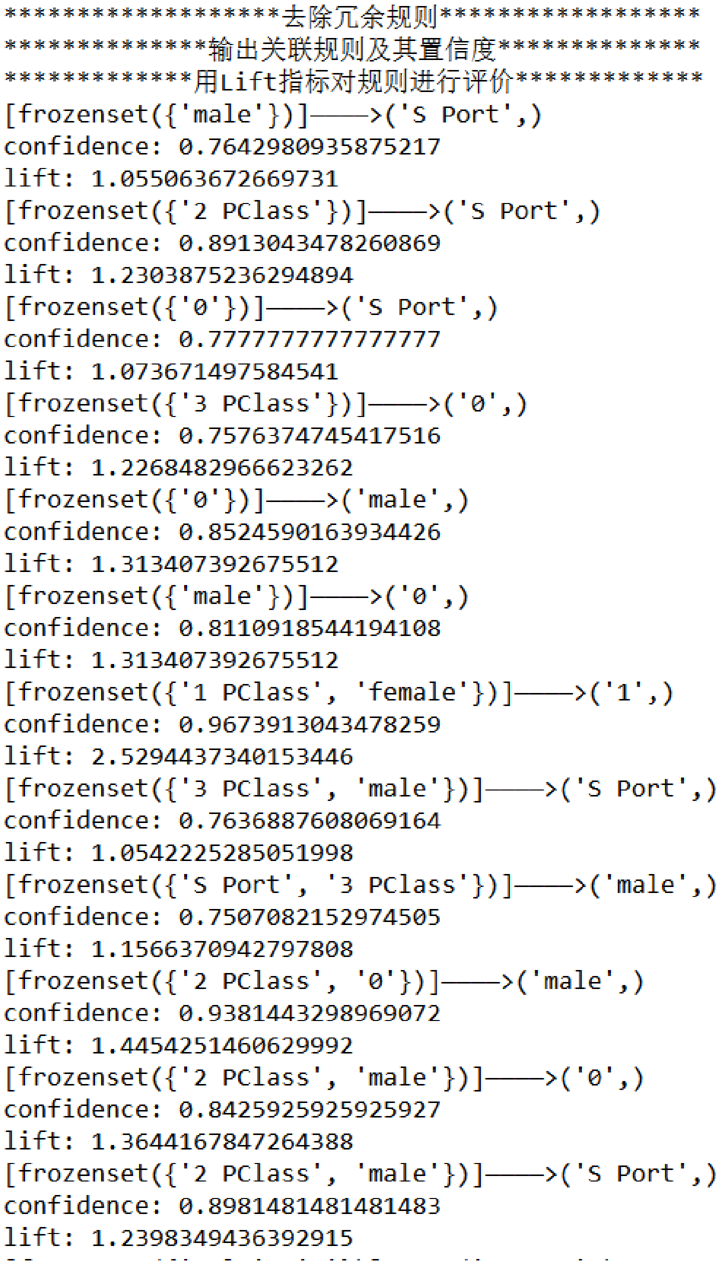


```
frozenset({'3 PClass', 'male'})
support:0.39032620922384703
frozenset({'2 PClass', '0'})
support:0.10911136107986502
frozenset({'1', 'S Port'})
support:0.2440944881889764
frozenset({'3 PClass', '0'})
support:0.4184476940382452
frozenset({'1 PClass', 'male'})
support:0.1372328458942632
frozenset({'1', 'C Port'})
support:0.1046119235095613
frozenset({'male', '0'})
support:0.5264341957255343
frozenset({'1', '3 PClass'})
support:0.13385826771653545
frozenset({'1', '1 PClass', 'female'})
support:0.10011248593925759
frozenset({'S Port', 'male', '3 PClass'})
support:0.2980877390326209
frozenset({'2 PClass', 'male', '0'})
support:0.10236220472440945
frozenset({'S Port', '2 PClass', 'male'})
support:0.10911136107986502
frozenset({'1', 'S Port', 'female'})
support:0.15748031496062992
frozenset({'S Port', 'male', '0'})
support:0.4094488188976378
frozenset({'S Port', '3 PClass', '0'})
support:0.3217097862767154
frozenset({'3 PClass', 'male', '0'})
support:0.3374578177727784
frozenset({'0', 'S Port', 'male', '3 PClass'})
support:0.25984251968503935
```

我们设置的最小支持度为0.1，以最后一个四项集为例，可以看出，在 S 港口登船坐三等舱的男性很大概率没能从灾难中存活下来。

3. 利用 Lift 指标评价关联规则，并计算置信度

由于置信度的缺陷在于该度量忽略了规则后续中项集的支持度，高置信度的规则有时可能出现误导。为解决这个问题的一种办法是使用称作提升度的 Lift 度量来评价关联规则。

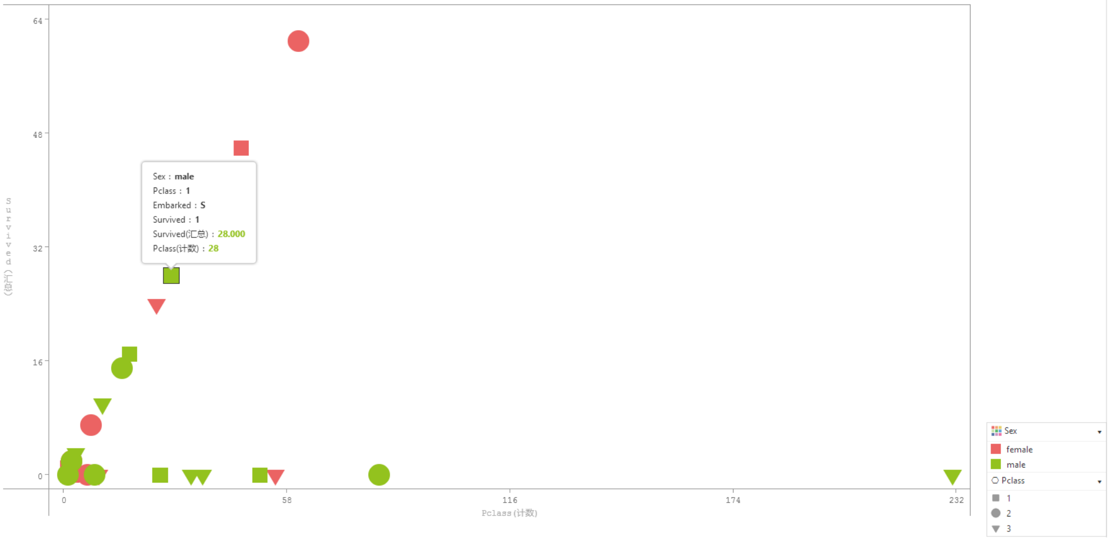


```
*******************去除冗余规则******************
**************输出关联规则及其置信度**************
*************用Lift指标对规则进行评价*************
[frozenset({'male'})]———>('S Port',)
confidence: 0.7642980935875217
lift: 1.055063672669731
[frozenset({'2 PClass'})]———>('S Port',)
confidence: 0.8913043478260869
lift: 1.2303875236294894
[frozenset({'0'})]———>('S Port',)
confidence: 0.7777777777777777
lift: 1.073671497584541
[frozenset({'3 PClass'})]———>('0',)
confidence: 0.7576374745417516
lift: 1.2268482966623262
[frozenset({'0'})]———>('male',)
confidence: 0.8524590163934426
lift: 1.313407392675512
[frozenset({'male'})]———>('0',)
confidence: 0.8110918544194108
lift: 1.313407392675512
[frozenset({'1 PClass', 'female'})]———>('1',)
confidence: 0.9673913043478259
lift: 2.5294437340153446
[frozenset({'3 PClass', 'male'})]———>('S Port',)
confidence: 0.7636887608069164
lift: 1.0542225285051998
[frozenset({'S Port', '3 PClass'})]———>('male',)
confidence: 0.7507082152974505
lift: 1.1566370942797808
[frozenset({'2 PClass', '0'})]———>('male',)
confidence: 0.9381443298969072
lift: 1.4454251460629992
[frozenset({'2 PClass', 'male'})]———>('0',)
confidence: 0.8425925925925927
lift: 1.3644167847264388
[frozenset({'2 PClass', 'male'})]———>('S Port',)
confidence: 0.8981481481481483
lift: 1.2398349436392915
```

可以得到各关联规则的置信度以及用 Lift 指标评价的结果。

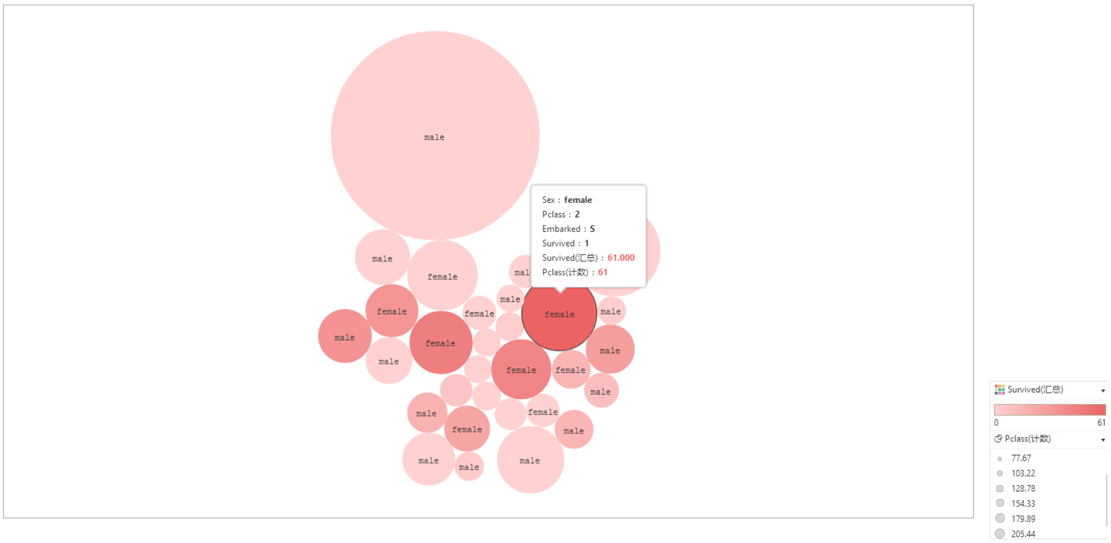
4. 使用可视化技术规则进行展示

散点图展示：



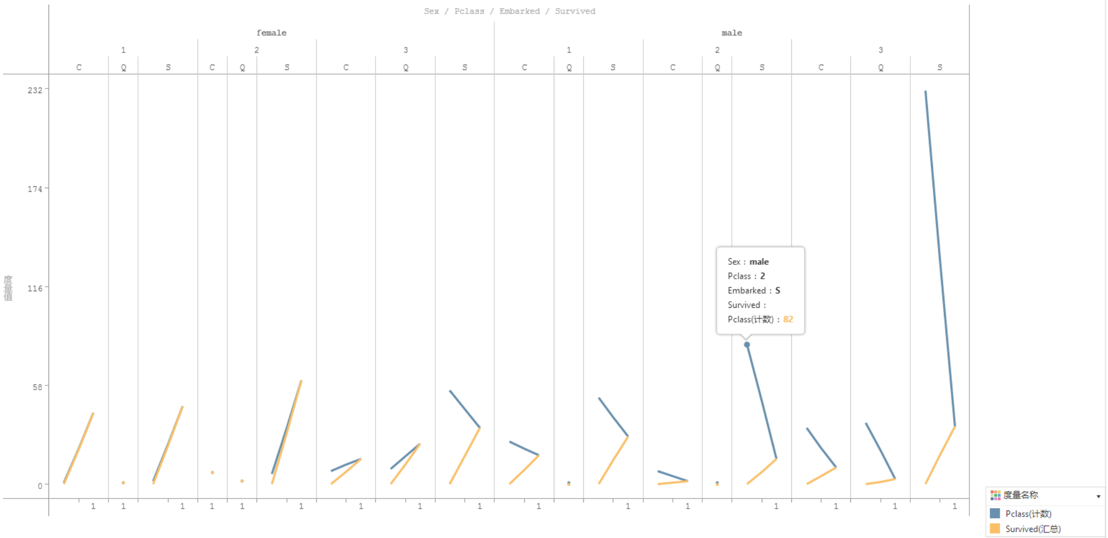
横坐标为船舱等级，纵坐标为是否存活。从中可以看出，男性（或女性）在 S 港口（或 C， Q 港口）登船坐一等（或二，三等）舱的存活（或死亡）人数。

气泡图展示：



如图，气泡越大，则乘坐该等级船舱（Pclass）的人数越多，颜色越深，则存活人数（survived）越多。

平行坐标展示：



从图中可以清晰地看出，分别依照每个属性的值进行展示的平行坐标图。首先，按着性别分类展示，再按船舱等级细分，接下来按着乘船港口分类展示。

5. 总结

通过完成数据关联规则分析的作业，我学习了 python 的使用，对于数据分析有了进一步的了解，掌握了频繁项集，支持度，置信度等概念，熟知了 Lift 评价指标，学习了散点图，平行坐标，气泡图等可视化技术，为今后进行更深层的数据挖掘工作做了准备。

附：实验结果保存在“挖掘结果及评价.txt”中，数据分析程序为 data_mining.py，可视化结果图保存在文件夹“可视化关联规则结果”中。