Apriori 算法求频繁项集与关联规则

学号：PB21000224　　姓名：陈鸿绪　　日期：6.1

一． 特征选择与数据预处理

在实验三中，我们已经找出了与 REPEAT 列存在强或弱相关性的若干属性的集合 S：

$$S=\{"ST127Q01TA","ST127Q02TA","EC033Q01NA","EC150Q07WA"$$

$$"EC150Q06WA","ST005Q01TA","ST007Q01TA","ST001D01T","GRADE"\}$$

由于实验要求对 ST127Q01A、ST127Q02TA、ST127Q02TA 进行合并处理，所以基于实验 3 的结果利用逻辑运算和四则运算构造如下 5 个新特征值：

$$"my\_repeat"=!('ST127Q01TA'+'ST127Q02TA'==2)$$

$$"feature\_plan"=!(('EC033Q01NA'+'EC150Q07WA'+'EC150Q06WA')<=4)$$

$$"m\_d\_education"=\left(\frac{2\times'ST005Q01TA'^2+'ST007Q01TA'^2}{3}\right)^{\frac{1}{2}}>=2.5$$

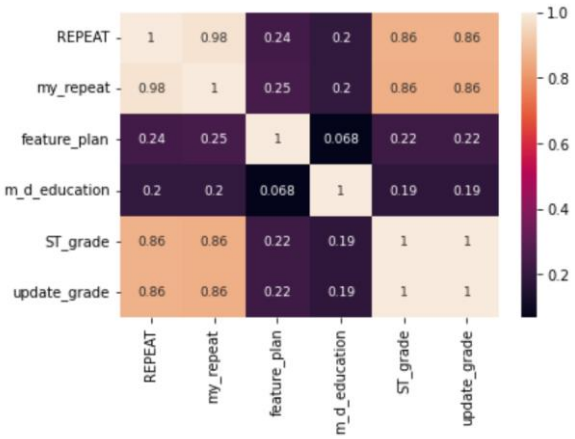
$$"ST\_grade"=!('ST001D01T'>=10)$$

$$"update\_grade"=!('GRADE'>=0)$$

将读入 CSV 文件得到的 DataFrame 数据对象取出属性名在属性集合 S 中的列，考察其各列数据的缺失度：

CNTRYID	0.000000
ST127Q01TA	0.050289
ST127Q02TA	0.038719
EC033Q01NA	0.285731
EC150Q07WA	0.297586
EC150Q06WA	0.298345
ST005Q01TA	0.022145
ST007Q01TA	0.039904
ST001D01T	0.000000
GRADE	0.000000
REPEAT	0.001755

发现整体数据集缺失情况并不严重，考虑采取前后值填充方法进行数据缺失填充预处理。再利用所构造特征的函数建立新特征加入该 DataFrame 对象，并删除其他特征，保留新特征与 REPEAT 特征。考察新特征与 REPEAT 特征之间的相关性：



由相关性可见：my_repeat、ST_grade、update_grade 明显是强相关，m_d_education、feature_plan 明显为弱相关。

二． 求频繁项集与关联规则的算法设计

1. Apriori 算法设计

在实验中取支持度、置信度阈值分别为 0.6，0.15。L 为频繁项集目标存储列表，LN 为过程中剔除项集的存储列表，support 为支持度列表，dict_为记录频繁 1-项集对应的值记录的字典。下面的算法即为程序中 Apriori 函数对应的算法思想。

a. K=1，对支持度为 0.6，寻找支持度大于 0.6 的频繁 1 项集，将其加入 L，支持度纳入 support 列表中，将每个 1-项集对应的值记录在字典 dict_中。

b. 设 K=k，对之前求出的频繁(k-1)-项集两两求并。如果求并后项集大小大于 k，则不予考虑，继续下一组并；若求并后项集大小恰好为 k，接下来考察 K=k-1 时任何被剔除的(k-1)-项集（即在 LN 中的(k-1)-项集）是否在求并得到的项集中，若存在一个被剔除的项集在求并得到的项集中，则将其纳入 LN 中，继续下一组并，若所有的被剔除(k-1)-项集均不在其中，则将该频繁 k-项集纳入 L 中、支持度纳入 support 列表中。继续下一组并。K 自增 1 。

c. 若得到频繁 k 项集只有一个或者没有，则跳出函数，否则返回继续执行 b 。

经过 Apriori 函数后，L 中即为得到的频繁项集，然而这只是建立在支持度之上，还需考虑置信度的阈值。

2. 计算置信度等以及基于置信度剪枝

在实验附加中，还要求计算 lift、PS、和 ψ-coefficient 等值，我们在求置信度

confidence 的过程中事实上可以顺带求出以上附加要求。可以在用 DataFrame 的 value_counts 方法求出各个比例从而求得 L 中的频繁项集所对应的各种指标，根据置信度 confidence 的阈值 0.15，将 L 中置信度 confidence 阈值小于 0.15 的项集剪枝即可。经过上述过程后，即可得到频繁项集与关联规则。

三． 频繁项集与关联规则的结果与分析

应用以上算法和函数，只考察 REPEAT 为 1 的情形，对构造的 5 个新特征进行频繁项集与关联规则的求解。得到频繁 1-项集对应的值的字典 dict_如下：

antecedents	value
'my_repeat'	1
'feature_plan'	0
'ST_grade'	1
'update_grade	1

对应频繁项集的值与支持度：

antecedents	consequentts	support
'my_repeat'	'REPEAT'=1	0.97187
'feature_plan'	'REPEAT'=1	0.62599
'ST_grade'	'REPEAT'=1	0.93101
'update_grade	'REPEAT'=1	0.93101
'update_grade', 'my_repeat'	'REPEAT'=1	0.91298
'update_grade', 'my_repeat'	'REPEAT'=1	0.91298
'update_grade', 'ST_grade'	'REPEAT'=1	0.93101
'update_grade', 'ST_grade', 'my_repeat'	'REPEAT'=1	0.91298

对应频繁项集的值与置信度等其他各种指标：

antecedents	consequentts	confidence	lift	PS	ψ-coefficient
'my_repeat'	'REPEAT'=1	0.99366	4.81101	0.15901	0.97829
'feature_plan'	'REPEAT'=1	0.15951	0.77232	-0.03811	-0.24026
'ST_grade'	'REPEAT'=1	0.85811	4.15470	0.14601	0.86496
'update_grade'	'REPEAT'=1	0.85811	4.15470	0.14601	0.86496
'ST_grade', 'my_repeat'	'REPEAT'=1	0.99824	4.83319	0.14955	0.94379
'update_grade', 'my_repeat'	'REPEAT'=1	0.99824	4.83319	0.14955	0.94379
'update_grade', 'ST_grade'	'REPEAT'=1	0.85811	4.15471	0.14601	0.86496
'update_grade', 'ST_grade', 'my_repeat'	'REPEAT'=1	0.99824	4.83319	0.14955	0.94379

经过分析，我们可以发现：

1. 在以上频繁项集中，最特殊的为 feature_plan 的频繁项集，该特征所得到的关联规则支

持度只有 0.6 左右，置信度大概在 0.16 的水平，而其他频繁项集的支持度 0.9 左右、置信度也在 0.9 左右。甚至如果考察 lift 值、PS 值、ψ-coefficient 值，会发现 lift 值小于 1，PS 值和 ψ-coefficient 值为负数，这三个值明显显示出这个关联规则并不优。但是我们可以分析 feature_plan 与 REPEAT 列的相关性，由之前的热图可以发现两者是明显的弱相关，发生此类情况也是在情理之中。

2. 考察其中 update_grade 对应的关联规则与 ST_grade 对应的关联规则，发现虽然两者是通过不同的特征变换而来的，但是两者各种指标几乎一样，两者存在在一个项集的时候指标也是与单个存在时等价。可以合理推测两者应该完全线性相关，对 REPEAT=1 的关联规则发挥相同作用。
3. 从图中可以明显发现，当本次实验中频繁 k-项集包含频繁 m-项集的时候，频繁 k-项集的支持度不大于频繁 m-项集，而置信度则是不小于频繁 m-项集。在本次实验结果中，lift、PS、ψ-coefficient 与置信度保持相同的单调性，所以可以猜测结论：频繁 k-项集包含频繁 m-项集的时，频繁 k-项集的 lift、PS、ψ-coefficient 值是不小于频繁 m-项集的。所以在评价频繁项集和关联规则的时候不能只考虑单个指标，结合实际情况，多个指标结合考虑才会更优。
4. 综合上述结论和结果取值，可以发现本次结果那些包含了强相关的关联规则和频繁项集基本可以算得上优的关联规则和频繁项集。置信度与支持度等指标均达到了一个较优较平衡的值。

四． 不同 CNTRYID（单个国家）的关联规则与频繁项集

将所有数据进行关于 CNTRYID 的划分与切割，相同的 CNTRYID 划分在同一个 DataFrame 中，总共划分出 6 个 DataFrame，然而其中有一个只有一条数据，所以对这个数据块不予考虑。这样只需要对剩下 5 个 DataFrame 依次进行类似上面的操作，得到每个数据分块关联规则和频繁项集，具体数据如下所示。比较不同数据块的关联规则与频繁项集，发现按 CNTRYID 来分的数据得到关联规则和频繁项集与总数据集所得到的几乎都不一样。比如 CNTRYID=484，频繁项集多达 28 个，而原数据集得到的只有 8 个，例如其中 `['feature_plan', 'my_repeat', 'm_d_education', 'REPEAT']`是全部数据集处理后并没有的频繁项集。这其实辛普森悖论的具体体现。

频繁项集：支持度 support

CNTRYID = 724 ：

```
['my_repeat', 'REPEAT']: 1.0
['ST_grade', 'REPEAT']: 0.9589825119236884
['update_grade', 'REPEAT']: 0.9589825119236884
['ST_grade', 'my_repeat', 'REPEAT']: 0.9589825119236884
['update_grade', 'my_repeat', 'REPEAT']: 0.9589825119236884
['update_grade', 'ST_grade', 'REPEAT']: 0.9589825119236884
['update_grade', 'ST_grade', 'my_repeat', 'REPEAT']: 0.958982511923
6884
```

CNTRYID=484 ：

```
['my_repeat', 'REPEAT']: 0.9296875
['ST_grade', 'REPEAT']: 0.880859375
['update_grade', 'REPEAT'],: 0.880859375
['feature_plan', 'my_repeat', 'REPEAT']: 0.9296875
['my_repeat', 'm_d_education', 'REPEAT']: 0.658203125
['ST_grade', 'my_repeat', 'REPEAT']: 0.880859375
['update_grade', 'my_repeat', 'REPEAT']: 0.880859375
['ST_grade', 'feature_plan', 'REPEAT']: 0.880859375
['update_grade', 'feature_plan', 'REPEAT']: 0.880859375
['ST_grade', 'm_d_education', 'REPEAT']: 0.62890625
['update_grade', 'm_d_education', 'REPEAT']: 0.62890625
['update_grade', 'ST_grade', 'REPEAT']: 0.880859375
['feature_plan', 'my_repeat', 'm_d_education', 'REPEAT']: 0.6582031
['ST_grade', 'feature_plan', 'my_repeat', 'REPEAT']: 0.880859375
['update_grade', 'feature_plan', 'my_repeat', 'REPEAT']: 0.88085937
['ST_grade', 'my_repeat', 'm_d_education', 'REPEAT']: 0.62890625
['update_grade', 'my_repeat', 'm_d_education', 'REPEAT']: 0.6289062
['update_grade', 'ST_grade', 'my_repeat', 'REPEAT']: 0.880859375
['ST_grade', 'feature_plan', 'm_d_education', 'REPEAT']: 0.62890625
['update_grade', 'feature_plan', 'm_d_education', 'REPEAT']: 0.6289
['update_grade', 'ST_grade', 'feature_plan', 'REPEAT']: 0.880859375
['update_grade', 'ST_grade', 'm_d_education', 'REPEAT']: 0.62890625
['feature_plan', 'my_repeat', 'ST_grade', 'm_d_education', 'REPEAT
']: 0.62890625
['update_grade', 'feature_plan', 'my_repeat', 'm_d_education', 'RE
PEAT']: 0.62890625
['update_grade', 'feature_plan', 'my_repeat', 'ST_grade', 'REPEAT
']: 0.880859375
```

['update_grade', 'my_repeat', 'ST_grade', 'm_d_education', 'REPEAT']: 0.62890625
['update_grade', 'feature_plan', 'ST_grade', 'm_d_education', 'REPEAT']: 0.62890625
['my_repeat', 'm_d_education', 'update_grade', 'feature_plan', 'ST_grade', 'REPEAT']: 0.62890625

CNTRYID=152：

['my_repeat', 'REPEAT']: 0.7929125138427464
['feature_plan', 'REPEAT']: 1.0
['ST_grade', 'REPEAT']: 0.867109634551495
['update_grade', 'REPEAT']: 0.867109634551495
['feature_plan', 'my_repeat', 'REPEAT']: 0.7929125138427464
['ST_grade', 'my_repeat', 'REPEAT']: 0.6932447397563677
['update_grade', 'my_repeat', 'REPEAT']: 0.6932447397563677
['ST_grade', 'feature_plan', 'REPEAT']: 0.867109634551495
['update_grade', 'feature_plan', 'REPEAT']: 0.867109634551495
['update_grade', 'ST_grade', 'REPEAT']: 0.867109634551495
['ST_grade', 'feature_plan', 'my_repeat', 'REPEAT']: 0.69324473975
['update_grade', 'feature_plan', 'my_repeat', 'REPEAT']: 0.69324473
['update_grade', 'ST_grade', 'my_repeat', 'REPEAT']: 0.693244739756
['update_grade', 'ST_grade', 'feature_plan', 'REPEAT']: 0.867109635
['update_grade', 'feature_plan', 'my_repeat', 'ST_grade', 'REPEAT']: 0.6932447397563677

CNTRYID=591：

['my_repeat', 'REPEAT']: 0.9799599198396793
['feature_plan', 'REPEAT']: 0.6603206412825652
['m_d_education', 'REPEAT']: 0.6803607214428857
['ST_grade', 'REPEAT']: 0.843687374749499
['update_grade', 'REPEAT']: 0.843687374749499
['feature_plan', 'my_repeat', 'REPEAT']: 0.6482965931863728
['my_repeat', 'm_d_education', 'REPEAT']: 0.6683366733466933
['ST_grade', 'my_repeat', 'REPEAT']: 0.843687374749499
['update_grade', 'my_repeat', 'REPEAT']: 0.843687374749499
['ST_grade', 'm_d_education', 'REPEAT']: 0.6022044088176353
['update_grade', 'm_d_education', 'REPEAT']: 0.6022044088176353
['update_grade', 'ST_grade', 'REPEAT']: 0.843687374749499
['ST_grade', 'my_repeat', 'm_d_education', 'REPEAT']: 0.60220440882
['update_grade', 'my_repeat', 'm_d_education', 'REPEAT']: 0.6022044
['update_grade', 'ST_grade', 'my_repeat', 'REPEAT']: 0.843687374749
['update_grade', 'ST_grade', 'm_d_education', 'REPEAT']: 0.60220441

['update_grade', 'my_repeat', 'ST_grade', 'm_d_education', 'REPEAT']: 0.6022044088176353

CNTRYID=214:

['my_repeat', 'REPEAT']: 0.75
['feature_plan', 'my_repeat', 'REPEAT']: 0.75
['my_repeat', 'm_d_education', 'REPEAT']: 0.625
['feature_plan', 'my_repeat', 'm_d_education', 'REPEAT']: 0.625

[参 考 文 献]

1.《数据挖掘导论》：参考了 Apriori 的具体伪代码算法