频繁项集与关联规则挖掘

吴韬略*

中国科学技术大学 大数据学院，合肥

【摘　要】 本次实验实现 FP-growth 算法，以支持度和置信度作为阈值寻找频繁项集，挖掘关联规则。并计算了对应的 lift，PS，coefficient 值。

【关键词】 **频繁项集**，**关联规则**，FP-growth

1　数据预处理

数据预处理步骤同实验三，选取与 REPEAT 列相关系数最大的五个特征：

$$['ST001D01T','ISCEDL','ST005Q01TA','ST013Q01TA','EVERE0-1']$$

其中，$'EVERE0-1'$ 是实验三中自己构建的特征。定义如下：

$$['EVERE'] = ['ST127Q01TA'] + ['ST127Q02TA'] + ['ST127Q03TA']$$

$$['EVERE0-1'] = \begin{cases} 0, ['EVERE'] < column['EVERE'].mean \\ 1, ['EVERE'] \geqslant column['EVERE'].mean \end{cases}$$

然后查询 codebook 将选取的特征名换为直观表示其含义的名字：

$$'EVERE0-1' : 'ever-repeated'$$

$$'ST001D01T' : 'international-Grade'$$

$$'ISCEDL' : 'ISCEDlevel'$$

$$'ST005Q01TA' : 'mother-school'$$

$$'ST013Q01TA' : 'books-inroom'$$

$$'REPEAT' : 'REPEAT'$$

2　频繁项集

选择 FP-growth 算法[1]，结果如图1：

```
{   frozenset({'REPEAT1.0'}): 1.0,
    frozenset({'REPEAT1.0', 'ever-repeated1'}): 1.0,
    frozenset({'REPEAT1.0', 'ISCED level2.0995'}): 0.8850442579606851,
    frozenset({'REPEAT1.0', 'ISCED level2.0995', 'ever-repeated1'}): 0.8850442579606851,
    frozenset({'REPEAT1.0', 'international-Grade8.5'}): 0.7535348890677089,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ever-repeated1'}): 0.7535348890677089,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995'}): 0.6777790550638004,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995', 'ever-repeated1'}): 0.6777790550638004}
```

图 1: 频繁项集: 左边为频繁项集，右边为支持度

说明：由于数据预处理使用了分箱均值去噪和均值填充空值，故结果可能与原来的整数值有一定偏差，不过可以看到偏差并不大（因为在实验四中没有进行规范化处理），所以不影响本实验。

分析如下：

(1) $ever-repeated1$ 作为一个相关系数为 1 的特征，自然会与 $REPEAT$ 保持一致，复读的同学之前学习生涯中肯定复读过，这很合理。

(2) $ISCED\,level$ 指的是受教育水平，2.0995 对应 $ISCED\,level\,2$ 左右，这是一个中间值。可得结论：受教育水平低的人读书时间短，复读的机会较少，当然也可能是因为他们在学习遇到困难时更倾向于辍学而不是复读；受教育水平高的人一般智商较高，虽然读书时间较长，可是其凭借自身天赋也能减少复读的概率；因此，处于中间的同学复读概率最大，其所在项集支持度也最高。

(3) $international-Grade$ 指的是学生国际成绩，8.5 对应 $Grade$ 8 到 $Grade$ 9 之间，属于中等偏下的成绩。这是一个有迷惑性的结果，初步看会以为是和 2 中 $ISCED\ level$ 2.0995 一样的结论。其实不然，通过 codebook 可以发现比该成绩区间低的成绩 $Grade$ 7 仅占 0.75%，所以其支持度小是由于其基数少而不是复读概率低。因此，我们有结论：成绩越低的学生复读的概率越大，支持度越高，这与学校挂科留级等制度相吻合。

3 关联规则

FP-growth 算法的关联规则发掘与 Apriori 算法是一样的：从顶向下，一步步生成候选集合再裁剪即可。结果如图2：

	antecedent	confidence	Lift	PS	coeficient
0	(ever-repeated1)	1.000000	4.839867	0.163927	1.000000
1	(ISCED level2.0995)	0.276128	1.336423	0.046033	0.240403
2	(ISCED level2.0995, ever-repeated1)	1.000000	4.839867	0.145082	0.926994
3	(international-Grade8.5)	0.849203	4.110031	0.117812	0.751995
4	(international-Grade8.5, ever-repeated1)	1.000000	4.839867	0.123524	0.841478
5	(ISCED level2.0995, international-Grade8.5)	0.878689	4.252735	0.107111	0.722769
6	(ISCED level2.0995, international-Grade8.5, ever-repeated1)	1.000000	4.839867	0.111106	0.790763

图 2: 关联规则

分析如下：

(1) $ever-repeated1$ 仍然高居榜首，结论同1。在此不再重复。

(2) $ISCED\ level\ 2.0995$ 单独作为前件已经无法满足置信度要求，必须与其余二者结合才可以满足要求，这也说明它本身和 $REPEAT$ 的关联性不是很强，并且与其余二者结合后反而降低了部分衡量指标，比如 $(international-Grade8.5, ISCED\ level2.0995, ever-repeated1)$ 比 $(international-Grade8.5, ever-repeated1)$ 的 $PS\ coeficient$ 值要低。但是其与 $international-Grade8.5$ 集合时却显著提高了 $antecedent\ confidence$ 这两个指标，考虑到正常情况下不会有 $ever-repeated$ 这么 bug 的特征存在，本实验还是认为 $ISCED\ level2.0995$ 是一个优良的适合组合的特征。

(3) $international-Grade8.5$ 与 $REPEAT$ 关联性较强，但是其与 $ever-repeated1$ 集合的效果反而不如 $(ISCED\ level2.0995, ever-repeated1)$，推测得到这可能是一个不太适合组合的特征。然瑕不掩瑜，同样的，考虑到正常情况下不会有 $ever-repeated$ 这么 bug 的特征存在，$international-Grade8.5$ 是我们正常情况下能找到的最佳单特征了。

4 辛普森悖论

将数据集分成不同国家的数据集，再次进行上述处理得到结果如图3。注：由于要寻找辛普森悖论现象，本节取支持度阈值为 0.4，置信度阈值为 0.6。

分析如下：

(1) 不同国家的频繁项集和关联规则差异还是比较大的。如 Chile 中 $ISCED\ level2.9$ 替代 $ISCED\ level2.0995$ 出现在了关联规则中，Dominican Republic 中 $books-inroom1.4975$ 出现在了关联规则中。

(2) 存在辛普森现象：在各个国家的关联规则中，$(books-inroom1.4975, international-Grade8.5)$ 的置信度都大于 $(ISCED\ level2.0995, international-Grade8.5)$ 的置信度，但是在上一节中我们知

国家152:

```
[   frozenset({'REPEAT1.0'}): 1.0,
    frozenset({'REPEAT1.0', 'ever-repeated1'}): 1.0,
    frozenset({'ISCED level2.9', 'REPEAT1.0'}): 0.8636363636363636,
    frozenset({'ISCED level2.9', 'REPEAT1.0', 'ever-repeated1'}): 0.8636363636363636,
    frozenset({'REPEAT1.0', 'international-Grade8.5'}): 0.7305986696230599,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ever-repeated1'}): 0.7305986696230599,
    frozenset({'ISCED level2.9', 'REPEAT1.0', 'international-Grade8.5'}): 0.7305986696230599,
    frozenset({'ISCED level2.9', 'REPEAT1.0', 'international-Grade8.5', 'ever-repeated1'}): 0.7305986696230599,
    frozenset({'REPEAT1.0', 'books-inroom1.4975'}): 0.6751662971175166,
    frozenset({'REPEAT1.0', 'books-inroom1.4975', 'ever-repeated1'}): 0.6751662971175166]
```

	antecedent	confidence	Lift	PS	coeficient
0	(ever-repeated1)	1.000000	6.279379	0.133890	1.000000
1	(ISCED level2.9, ever-repeated1)	1.000000	6.279379	0.115633	0.917546
2	(international-Grade8.5)	0.653122	4.101200	0.087979	0.628383
3	(international-Grade8.5, ever-repeated1)	1.000000	6.279379	0.097820	0.833743
4	(ISCED level2.9, international-Grade8.5)	0.653122	4.101200	0.087979	0.628383
5	(ISCED level2.9, international-Grade8.5, ever-repeated1)	1.000000	6.279379	0.097820	0.833743
6	(books-inroom1.4975)	0.216803	1.361389	0.028542	0.156011
7	(ever-repeated1, books-inroom1.4975)	1.000000	6.279379	0.090398	0.797516

(a) 国家：Chile

国家188:

```
[]
```

	antecedent	confidence	Lift	PS	coeficient

(b) 国家：Costa Rica

国家214:

```
[   frozenset({'REPEAT1.0'}): 1.0,
    frozenset({'REPEAT1.0', 'ever-repeated1'}): 1.0,
    frozenset({'international-Grade9.5', 'REPEAT1.0'}): 0.75,
    frozenset({'international-Grade9.5', 'REPEAT1.0', 'ever-repeated1'}): 0.75,
    frozenset({'ISCED level2.9', 'REPEAT1.0'}): 0.75,
    frozenset({'ISCED level2.9', 'REPEAT1.0', 'ever-repeated1'}): 0.75,
    frozenset({'international-Grade9.5', 'ISCED level2.9', 'REPEAT1.0'}): 0.75,
    frozenset({'international-Grade9.5', 'ISCED level2.9', 'REPEAT1.0', 'ever-repeated1'}): 0.75,
    frozenset({'REPEAT1.0', 'books-inroom1.4975'}): 0.75,
    frozenset({'REPEAT1.0', 'books-inroom1.4975', 'ever-repeated1'}): 0.75,
    frozenset({'international-Grade9.5', 'REPEAT1.0', 'books-inroom1.4975'}): 0.625,
    frozenset({'international-Grade9.5', 'REPEAT1.0', 'books-inroom1.4975', 'ever-repeated1'}): 0.625,
    frozenset({'ISCED level2.9', 'REPEAT1.0', 'books-inroom1.4975'}): 0.625,
    frozenset({'ISCED level2.9', 'REPEAT1.0', 'books-inroom1.4975', 'ever-repeated1'}): 0.625,
    frozenset({'international-Grade9.5', 'ISCED level2.9', 'REPEAT1.0', 'books-inroom1.4975'}): 0.625,
    frozenset({'international-Grade9.5', 'books-inroom1.4975', 'ISCED level2.9', 'REPEAT1.0', 'ever-repeated1'}): 0.625]
```

	antecedent	confidence	Lift	PS	coeficient
0	(ever-repeated1)	1.0	13.0	0.071006	1.000000
1	(international-Grade9.5, ever-repeated1)	1.0	13.0	0.053254	0.857143
2	(ISCED level2.9, ever-repeated1)	1.0	13.0	0.053254	0.857143
3	(international-Grade9.5, ISCED level2.9, ever-repeated1)	1.0	13.0	0.053254	0.857143
4	(ever-repeated1, books-inroom1.4975)	1.0	13.0	0.053254	0.857143
5	(international-Grade9.5, ever-repeated1, books-inroom1.4975)	1.0	13.0	0.044379	0.778499
6	(ISCED level2.9, ever-repeated1, books-inroom1.4975)	1.0	13.0	0.044379	0.778499
7	(international-Grade9.5, ISCED level2.9, ever-repeated1, books-inroom1.4975)	1.0	13.0	0.044379	0.778499

(c) 国家：Dominican Republic

国家484:

```
[   frozenset({'REPEAT1.0'}): 1.0,
    frozenset({'REPEAT1.0', 'ever-repeated1'}): 1.0,
    frozenset({'REPEAT1.0', 'ISCED level2.0995'}): 0.8825831702544031,
    frozenset({'REPEAT1.0', 'ISCED level2.0995', 'ever-repeated1'}): 0.8825831702544031,
    frozenset({'REPEAT1.0', 'books-inroom1.4975'}): 0.7318982387475538,
    frozenset({'REPEAT1.0', 'books-inroom1.4975', 'ever-repeated1'}): 0.7318982387475538,
    frozenset({'REPEAT1.0', 'international-Grade8.5'}): 0.7005870841487280,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ever-repeated1'}): 0.7005870841487280,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995'}): 0.7005870841487280,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995', 'ever-repeated1'}): 0.7005870841487280,
    frozenset({'REPEAT1.0', 'books-inroom1.4975', 'ISCED level2.0995'}): 0.6575342465753424,
    frozenset({'books-inroom1.4975', 'REPEAT1.0', 'ISCED level2.0995', 'ever-repeated1'}): 0.6575342465753424]
```

	antecedent	confidence	Lift	PS	coeficient
0	(ever-repeated1)	1.000000	11.403131	0.080005	1.000000
1	(ISCED level2.0995)	0.622928	7.103332	0.066502	0.712757
2	(ISCED level2.0995, ever-repeated1)	1.000000	11.403131	0.070611	0.934202
3	(ever-repeated1, books-inroom1.4975)	1.000000	11.403131	0.058555	0.844695
4	(international-Grade8.5)	0.588816	6.714344	0.052288	0.604703
5	(international-Grade8.5, ever-repeated1)	1.000000	11.403131	0.056050	0.825220
6	(ISCED level2.0995, international-Grade8.5)	0.588816	6.714344	0.052288	0.604703
7	(ISCED level2.0995, international-Grade8.5, ever-repeated1)	1.000000	11.403131	0.056050	0.825220
8	(ISCED level2.0995, books-inroom1.4975)	0.702929	8.015590	0.050469	0.650219
9	(ISCED level2.0995, ever-repeated1, books-inroom1.4975)	1.000000	11.403131	0.052606	0.797859

(d) 国家：Mexico

国家591:

```
[   frozenset({'REPEAT1.0'}): 1.0,
    frozenset({'REPEAT1.0', 'ever-repeated1'}): 1.0,
    frozenset({'REPEAT1.0', 'ISCED level2.0995'}): 0.8442871587462083,
    frozenset({'REPEAT1.0', 'ISCED level2.0995', 'ever-repeated1'}): 0.8442871587462083,
    frozenset({'REPEAT1.0', 'books-inroom1.4975'}): 0.7522750252780587,
    frozenset({'REPEAT1.0', 'books-inroom1.4975', 'ever-repeated1'}): 0.7522750252780587,
    frozenset({'REPEAT1.0', 'books-inroom1.4975', 'ISCED level2.0995'}): 0.6471183013144590,
    frozenset({'books-inroom1.4975', 'REPEAT1.0', 'ISCED level2.0995', 'ever-repeated1'}): 0.6471183013144590,
    frozenset({'REPEAT1.0', 'international-Grade8.5'}): 0.6117290192113246,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ever-repeated1'}): 0.6117290192113246,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995'}): 0.6117290192113246,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995', 'ever-repeated1'}): 0.6117290192113246]
```

	antecedent	confidence	Lift	PS	coeficient
0	(ever-repeated1)	1.000000	4.614762	0.169739	1.000000
1	(ISCED level2.0995)	0.764652	3.528687	0.131106	0.745893
2	(ISCED level2.0995, ever-repeated1)	1.000000	4.614762	0.143308	0.899678
3	(books-inroom1.4975)	0.234700	1.083086	0.012505	0.065900
4	(ever-repeated1, books-inroom1.4975)	1.000000	4.614762	0.127690	0.839063
5	(ISCED level2.0995, books-inroom1.4975)	0.765550	3.532832	0.100535	0.630858
6	(ISCED level2.0995, ever-repeated1, books-inroom1.4975)	1.000000	4.614762	0.109841	0.767831
7	(international-Grade8.5)	0.732446	3.380062	0.093341	0.588462
8	(international-Grade8.5, ever-repeated1)	1.000000	4.614762	0.103834	0.743233
9	(ISCED level2.0995, international-Grade8.5)	0.732446	3.380062	0.093341	0.588462
10	(ISCED level2.0995, international-Grade8.5, ever-repeated1)	1.000000	4.614762	0.103834	0.743233

(e) 国家：Panama

国家724:

```
[   frozenset({'REPEAT1.0'}): 1.0,
    frozenset({'REPEAT1.0', 'ever-repeated1'}): 1.0,
    frozenset({'REPEAT1.0', 'ISCED level2.0995'}): 0.9998409922086182,
    frozenset({'REPEAT1.0', 'ISCED level2.0995', 'ever-repeated1'}): 0.9998409922086182,
    frozenset({'REPEAT1.0', 'international-Grade8.5'}): 0.7843854348863094,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ever-repeated1'}): 0.7843854348863094,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995'}): 0.7843854348863094,
    frozenset({'REPEAT1.0', 'international-Grade8.5', 'ISCED level2.0995', 'ever-repeated1'}): 0.7843854348863094}
```

	antecedent	confidence	Lift	PS	coeficient
0	(ever-repeated1)	1.000000	4.124980	0.183655	1.000000
1	(ISCED level2.0995)	0.242527	1.000419	0.000102	0.009865
2	(ISCED level2.0995, ever-repeated1)	1.000000	4.124980	0.183626	0.999895
3	(international-Grade8.5)	0.935876	3.860468	0.140898	0.817108
4	(international-Grade8.5, ever-repeated1)	1.000000	4.124980	0.144057	0.856597
5	(ISCED level2.0995, international-Grade8.5)	0.935876	3.860468	0.140898	0.817108
6	(ISCED level2.0995, international-Grade8.5, ever-repeated1)	1.000000	4.124980	0.144057	0.856597

(f) 国家：Spain

图 3: 不同国家的频繁项集和关联规则

道，对于整个数据集，$(ISCED\ \ level2.0995, international - Grade8.5)$ 的置信度是大于 $(books - inroom1.4975, international-Grade8.5)$ 的置信度的。寻找原因，应是 $(ISCED\ level2.0995, international-Grade8.5)$ 在 Spain 中置信度很高，而且 $(ISCED\ level2.0995, international-Grade8.5, REPEAT1.0)$ 在 Spain 基数较大，数量较多导致的。

参考文献

[1] Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pages 1–5, 2005.

[2] Liu Qi. Prof. qiliu’s slides on ad2023. `http://staff.ustc.edu.cn/~qiliuql/files/AD2023/4.1.pdf`.

[3] Liu Qi. Prof. qiliu’s slides on ad2023. `http://staff.ustc.edu.cn/~qiliuql/files/AD2023/4.2.pdf`.