数据分析与数据挖掘第三次作业第二大题 – a问分析文档

数据分析与数据挖掘

DAM COURSE, SPRING 2018

BY

1552674 李　源

同濟大學
TONGJI UNIVERSITY

Tongji University
School of Software Engineering

1 代码运行结果

不同特征的生成时间如下：

```
TYPE.1 count/ratio - count
0:00:07.979611
************************
TYPE.1 count/ratio - penetration
0:00:03.572952
************************
TYPE.1 count/ratio - product diversity
0:00:03.223754
************************
TYPE.2 AGG feature - brand/category/item AGG
0:00:15.392799
************************
TYPE.2 AGG feature - user AGG
0:00:04.405343
************************
TYPE.4 complex feature - repeat feature
0:00:07.842378
************************
```

2 分析讨论

这里我针对我选取的特征和生成特征的方法进行讨论。

2.1 选取的特征

我最初选取的所有特征汇总如下：（注意，每一个 monthly 分组计算了三次，因为每次 monthly 相关的特征是将 3 个月的数据一起放进去的。）

特征种类	特征含义	分组方法	特征个数
TYPE.1 count/ratio - count	购买/被购买的次数	UI、monthly 分组	3
		UI、whole 分组	1
		UB、monthly 分组	3
		UB、whole 分组	1
		UC、monthly 分组	3
		UC、whole 分组	1
	购买/被购买的金额	UI、monthly 分组	3
		UI、whole 分组	1
TYPE.1 count/ratio - product diversity	购买的 I 的数量	U、monthly 分组	3
		U、whole 分组	1
	购买的 B 的数量	U、monthly 分组	3
		U、whole 分组	1
	购买的 C 的数量	U、monthly 分组	3
		U、whole 分组	1
TYPE.1 count/ratio - penetration	购买过的 U 的数量	I、monthly 分组	3
		I、whole 分组	1
	购买过的 U 的数量	B、monthly 分组	3
		B、whole 分组	1
	购买过的 U 的数量	C、monthly 分组	3
		C、whole 分组	1

特征种类	特征含义	分组方法	特征个数
TYPE.2 AGG feature - month AGG	AGG	monthly 分组	36
TYPE.2 AGG feature - user AGG	购买过的 U 的AGG	I 分组	4
		B 分组	4
		C 分组	4
TYPE.2 AGG feature - B/C/I AGG	I 次数的AGG	U 分组	4
	I 金额的AGG	U 分组	4
	B 次数的AGG	U 分组	4
	B 金额的AGG	U 分组	4
	C 次数的AGG	U 分组	4
	C 金额的AGG	U 分组	4
TYPE.4 complex feature - trend	trend	monthly 分组	36
TYPE.4 complex feature - repeat feature	重复购买者	I 分组	1
		B 分组	1
		C 分组	1

选取特征依据我主要是参考了如下两个因素:

- 其分组方式生成的 label 对，是否是在我之后的预测当中需要的，比如 b 问中是 vipno - pluno 这样的一对，那么我在提取特征的时候，会将相关的特征都提取出来。
- 这种特征是否有类似的特征可以替代。比如在上面的表格中我列出的，都是与购买次数相关的特征，而不是与购买天数相关的。因为我考虑这两种特征在一定程度上是相似的，因此我在抽取特征时，只选择其中一种，并且进行对比。

这里我是抽取出了我考虑之后想用到的数据，在每一问中，我仍会使用 leave-out auc 的方法，对这些特征进行进一步地筛选。具体的描述可以见 b 问文档。

2.2 特征选取方法

具体代码可见 a.py 文件。这里简单描述一下方法，我主要用到的是 python 中的字典这个数据结构。

我是先按照分组方法，比如 按照 type1 和 type2 分组（这里的 type 是指的 U、I、B、C），则先根据这两个出现过的 type1 和 type2 两两组合，生成字典 的 key，然后在原始数据中去寻找，将对应的 value 填补上去。

之后，我再将 value 不存在的数据删去，而对于存在 value 的数据，则是按 “type1-type2-value” 的格式存储。如果只有一个 type 进行分组，方法是类似的。

耗时在前文有描述，基本上生成一种类型的数据，大概耗时在几秒中的样子，当然这也要考虑到该类型中特征的数量多少。