聚类分析

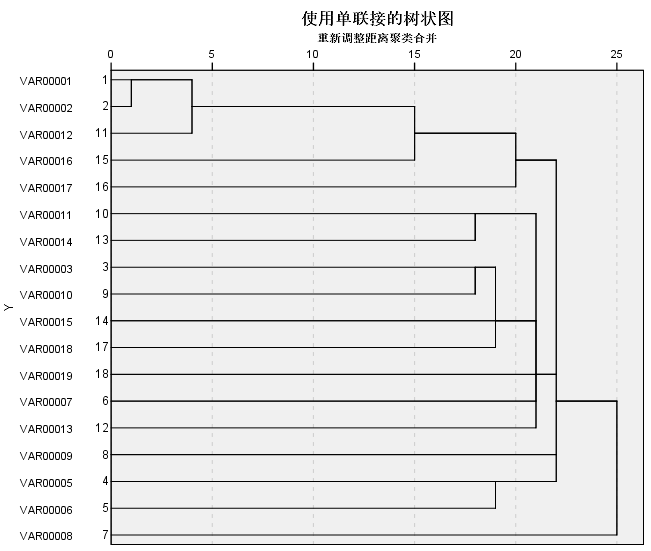
聚类分析是指标降维的一种，主要目的是将很多指标进行聚类，聚类和分类不同，区别是：分类是已经知道有哪些类别，然后将各个指标或者变量进行分类。

聚类则是不知道有哪些类别，根据一定的规则进行聚类。

例如Q型聚类分析（样本聚类）是，第一次的时候计算各个样本（一个样本是一类）之间的距离（这个距离可以是绝对距离，也可以是欧几里得距离等等，常用的是Minkowski距离），将距离最小的两个聚成一类，这个时候就少了一类，然后针对新的N个类重新进行聚类（对于刚才由两个类合并的那个类则可以根据一定的规则进行转化，这个规则包括最短举例法，最长距离法，重心法，类平均法，离差平方和法等等，），重新聚类后又少了一类，循环进行，一直到还有一类聚类结束。

那么到底聚成多少类合适呢？这个由于评判标准不同，也不好说多少类合适。但是在每一次聚类后都会有一个指标，观察这个指标，如果这个指标突然变化，就可以认为聚类到这里就可以了。

R型聚类和Q型聚类类似，R型聚类叫做变量聚类，因为是变量所以一个变量有很多数据，这个时候可以根据各个变量之间的相关性系数（就像Q型聚类的“距离”）确定。聚类分析建议使用spss进行，操作比较简单。还可以直接生成聚类图。点击上方的“分析”，有一个“分类”一般使用的是系统聚类。然后的操作就和主成分那些差不多了，只不过需要选择“方法”“绘制”个案（也可以叫做样本）还是变量。然后就出来结果了，看结果就行了。



结果图，变量1和变量2大概是在“1”左右形成一类，然后变量1+2大概和变量12在“4”左右形成一类。因为橙色的线在聚类过程中，聚类的指标变化比较大，可以认为到这里聚类结束。（看图法，比较直观，但是不推荐）

也可以看下面的聚类表，系数那一列，第二阶和第三阶对应的系数变化相比其他的比较明显，那么可以认为到这里聚类结束。但是这样聚类效果不好，再往下观察，寻找合适的阶数，确定聚类什么时候结束比较合适（我这里的数据是我随机生成的，聚类效果不好，一般的数据聚类效果会比较好。）具体聚成几类，需综合考虑题目背景与要求还有聚类的结果“系数”等

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **聚类表** | | | | | | |
| 阶 | 群集组合 | | 系数 | 首次出现阶群集 | | 下一阶 |
| 群集 1 | 群集 2 | 群集 1 | 群集 2 |
| 1 | 1 | 2 | 123.984 | 0 | 0 | 2 |
| 2 | 1 | 11 | 127.201 | 1 | 0 | 3 |
| 3 | 1 | 15 | 137.055 | 2 | 0 | 9 |
| 4 | 3 | 9 | 140.043 | 0 | 0 | 6 |
| 5 | 10 | 13 | 140.314 | 0 | 0 | 10 |
| 6 | 3 | 14 | 140.833 | 4 | 0 | 8 |
| 7 | 4 | 5 | 141.046 | 0 | 0 | 16 |
| 8 | 3 | 17 | 141.053 | 6 | 0 | 10 |
| 9 | 1 | 16 | 141.915 | 3 | 0 | 14 |
| 10 | 3 | 10 | 142.562 | 8 | 5 | 11 |
| 11 | 3 | 18 | 142.780 | 10 | 0 | 12 |
| 12 | 3 | 6 | 143.108 | 11 | 0 | 13 |
| 13 | 3 | 12 | 143.122 | 12 | 0 | 14 |
| 14 | 1 | 3 | 143.346 | 9 | 13 | 15 |
| 15 | 1 | 8 | 143.611 | 14 | 0 | 16 |
| 16 | 1 | 4 | 144.167 | 15 | 7 | 17 |
| 17 | 1 | 7 | 147.010 | 16 | 0 | 0 |

关于聚类分析的MATLAB，具体的用法自己可以看《matlab统计分析与应用 40个案例分析》