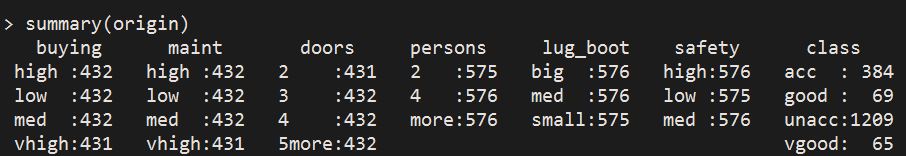
**購車偏好度：關聯分析與預測**

1. 資料型態



一共1727筆資料物件和7個屬性。

* 屬性資料
  + 費用

|  |  |
| --- | --- |
| Buying 價格 | vhigh、high、med、low |
| Maint 維修費用 | vhigh、high、med、low |

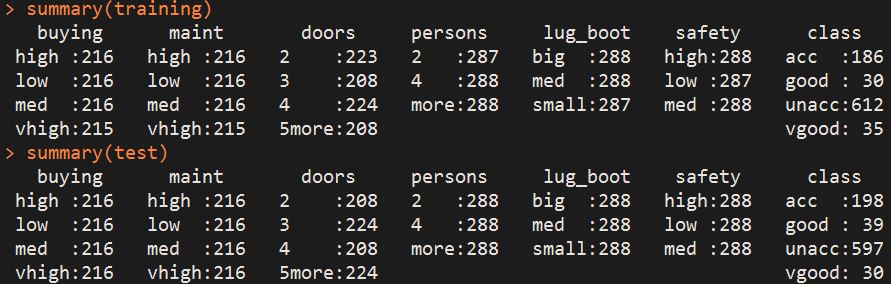
* 車子屬性

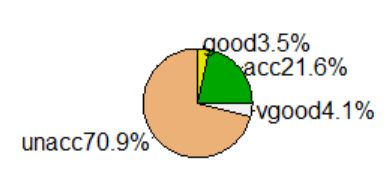
|  |  |
| --- | --- |
| doors車門數量 | 5以上、4、3、2 |
| persons承載人數 | 5以上、4、2 |
| lug\_boot行李箱尺寸 | big、med、small |
| safety安全性 | high、med、low |

* 分類

|  |  |
| --- | --- |
| class接受度 | vgood、good、acc、unacc |

由於資料龐大，足夠將資料分成training和test data，故將原data切成兩份存檔。並且兩份資料散布情況相同。



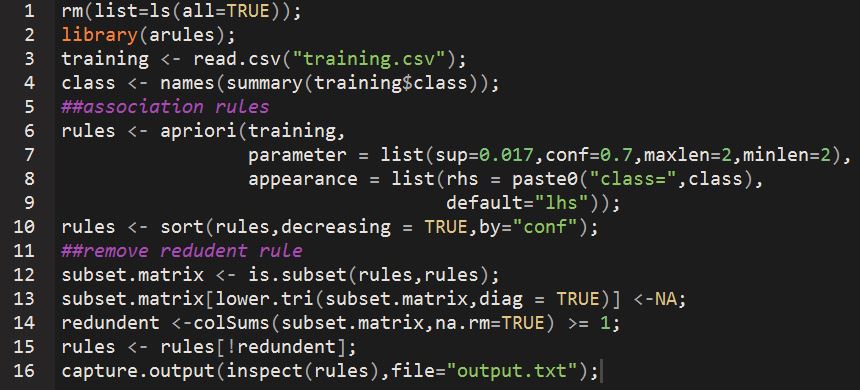


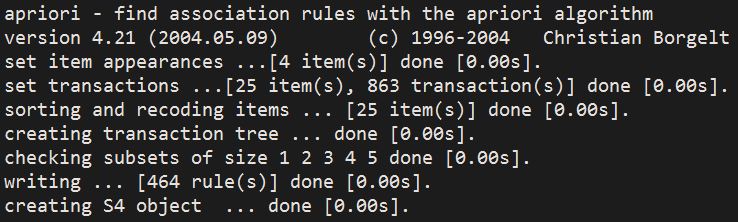
可以看出四個類別的比例不平均，unacceptable的比例遠遠超過其他三者。在探索關聯規則時關於unacceptable理論上會比較多。

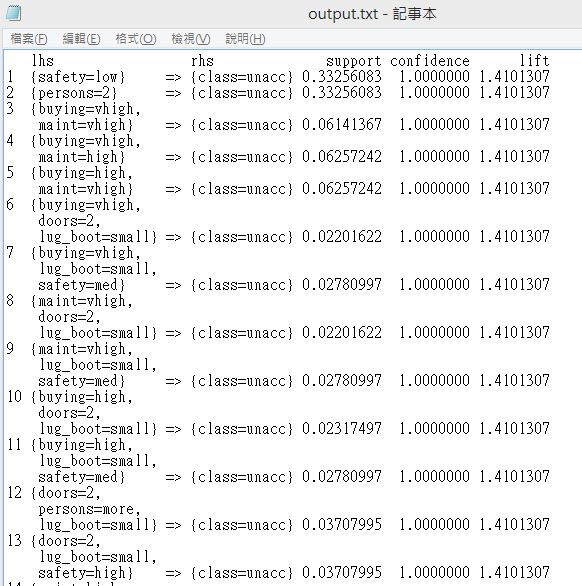
1. 關聯分析

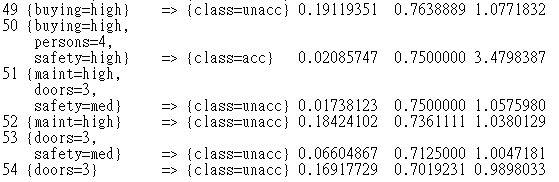
* 第一組參數 support：0.017（約15人）、confidence：0.7

由於組合太多，先分析confidence較高的，而且與分類相關的規則，並在過程中就把冗於規則移除，最後依照confidence高低排序。









總共產生464條規則，刪去冗餘規則後剩餘54條規則，其中53條與unacceptable有關，只有1條是關於acceptable。

其中safety=low和person=2在資料中出現的比率和safety=low→class=unacc、person=2→class=unacc的support相等，可以明確表示兩者為強規則。

對good、vgood來說，support=0.017的門檻太高，完全無法產生相關規則；但對unacc來說，support=0.017太低，取出太多規則。

* 第二組參數