Statistical analysis for bigdata:

Term Paper

麵包店交易資料分析與關聯規則應用

610721204／陳克威／2019-06-28

GitHub Repository URL:

<https://github.com/D1034181036/BigData_TermPaper>

1. 資料集：

此資料集來自於Kaggle上的Transactions from a bakery，是一家麵包店的交易資料，總共有4個欄位分別為日期、時間、交易編號、交易商品，如表1所示，日期從2016/10/30至2017/04/09，總共有21293筆(row)交易商品數量。

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Time | Transaction | Item |
| 2016/10/30 | 09:58:11 | 1 | Bread |
| 2016/10/30 | 10:05:34 | 2 | Scandinavian |
| 2016/10/30 | 10:05:34 | 2 | Scandinavian |
| 2016/10/30 | 10:07:57 | 3 | Hot chocolate |
| 2016/10/30 | 10:07:57 | 3 | Jam |
| 2016/10/30 | 10:07:57 | 3 | Cookies |
| 2016/10/30 | 10:08:41 | 4 | Muffin |

表1　資料集中的前4筆交易資料

1. 資料前處理：

首先我們使用R語言讀取csv資料檔，從原始資料中（資料清理前）能觀察到以下幾個特徵及問題：

(1)在交易資料中，Transaction編號為1至9684，但從資料讀取後發現實際為9531筆交易，可以發現有部份交易編號不見的情形。

(2)在交易商品中有一欄為"NONE"，因為我們不清楚這項商品代表的意義，因此將其刪除，刪除後的資料為9465筆。

(4)在商品[Ella's Kitchen Pouches]與[Valentine's card]中含有單引號，在使用arules套件中的read.transactions指令讀取資料時會有問題，在這裡我們提出兩種解決方式，一種方法為直接將其符號移除，另一種方法是先使用read.csv讀取進R語言，接著再將其改為transactions的格式。

我們將資料清理前／後做比較，如表2所示，現在前處理已經完成，我們可以開始進一步分析這份交易資料了。

|  |  |  |
| --- | --- | --- |
|  | 資料清理前 | 資料清理後 |
| transactions | 9531 | 9465 |
| items | 95 | 94 |

表2　資料清理前／後之比較

1. 資料分析：

首先我們將商品依照銷售次數列出，如圖1所示，可以看出咖啡與麵包是店裡的主力商品。

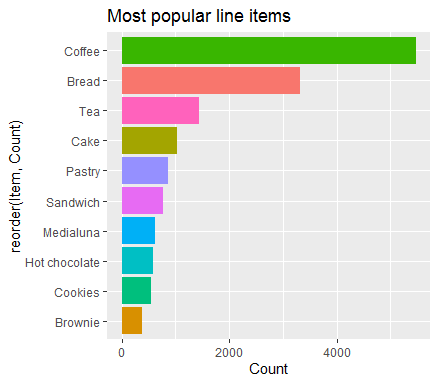


圖1　商品銷售次數

接著我們觀察每日的銷售數，如圖2所示，週末的時候通常生意最好。

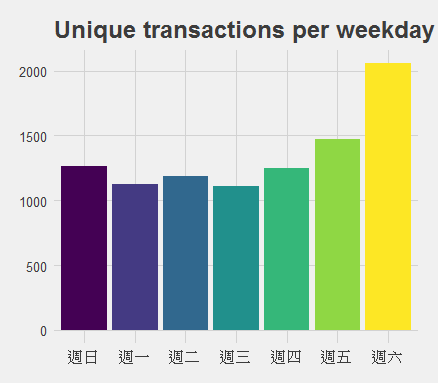
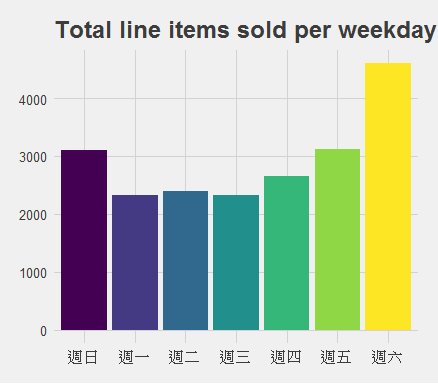


圖2　每日商品銷售數(左)、每日交易數(右)

我們繼續觀察每小時的銷售數，如圖3所示，可以看出大約早上8點開始營業，中午與下午時生意特別好，營業時間大約到晚上6點。

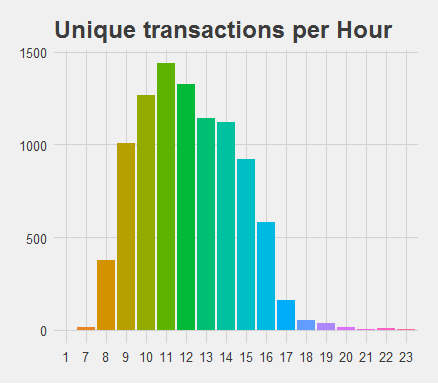
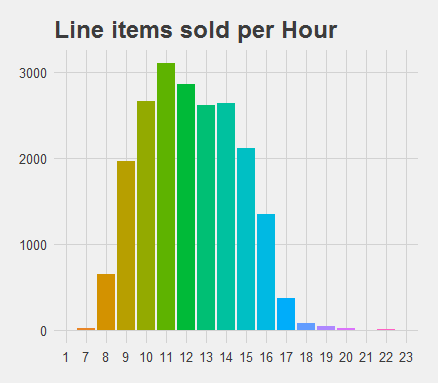


圖3　每小時商品銷售數(左)、每小時交易數(右)

我們可以觀察每筆交易購買的商品數，如表3所示，顧客到店裡平均會買2項商品。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 1 | 1 | 2 | 1.995 | 3 | 10 |

表3　每筆交易商品數

1. 頻繁項目集與關聯規則分析：

我們使用Agrawal&Srikant所提出的Apriori演算法[3]來找出Frequent Itemsets，此方法相對簡單，在R語言中可以使用arules套件來實做。

我們將min\_support設定為0.01，min\_length設定為2，並且不考慮confidence，這代表該itemsets必須至少在9465筆交易中出現94次，且該itemset至少要有兩項商品以上。

這裡我們將Frequent Itemsets列出並且按照support排序，如圖4所示，看起來較多的是常見的食物與咖啡組合。

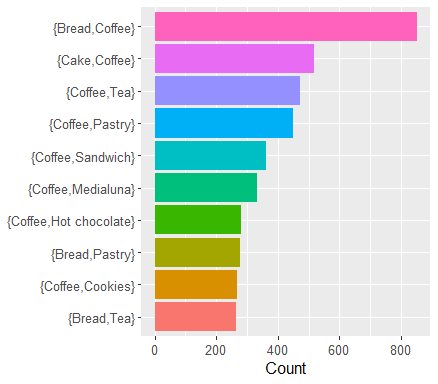


圖4　前十項頻繁項目集

為了找出有趣的關聯規則，我們將min\_support設定為0.01，min\_confidence設定為0.5，min\_length設定為2，以lift做排序找出前十項規則，如表4所示，大部分還是食物與咖啡的組合。

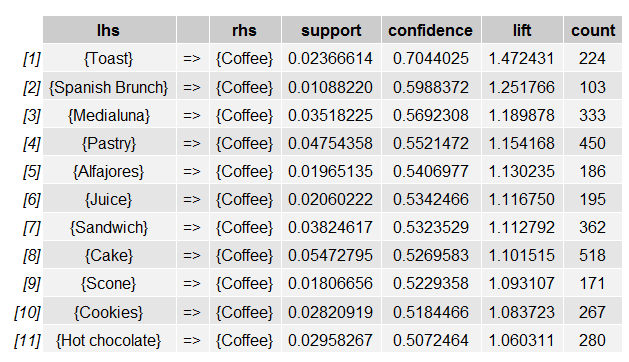


表4　前十項lift值的關聯規則(min\_sup=0.01)

我們將min\_support降低至0.0005，找找看有哪些有趣的規則，如表5所示，買明信片的人也會買Tshirt？這裡的樣本數太少了，實際應用的價值可能不是很大。

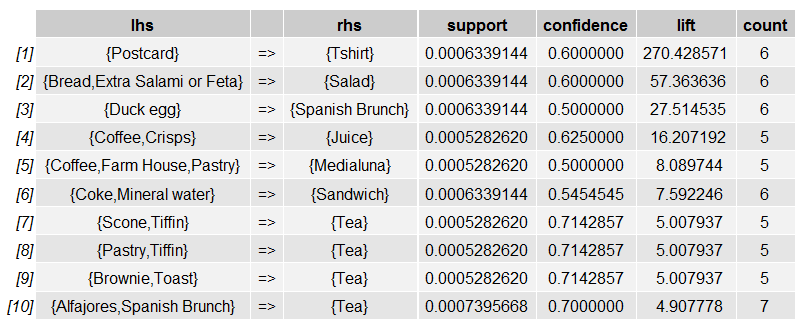


表5　前十項lift值的關聯規則(min\_sup=0.0005)

在這份資料中有趣的關聯規則較少，我們認為可能是商品種類比較單一，資料總數也不是很多。

1. 補充：分類器的基本應用：

也許我們可以利用分類器來預測顧客是否會購買某項商品，在結帳時可以做商品推薦，我們嘗試利用Naive Bayes Classifiers來預測顧客是否會購買麵包，在已知顧客購物籃中的物品的情況下，我們利用貝氏定理算出所有P(Bread=0|Items)以及P(Bread=1|Items)的機率，利用這些數據建立模型後，就能利用模型來預測顧客是否會購買麵包。

我們將資料分為80%訓練資料以及20%測試資料，實驗結果如表6所示。

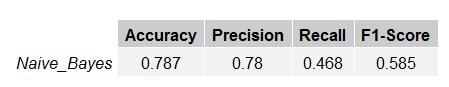


表6　 使用Naive Bayes Classifier預測顧客是否購買麵包

1. 參考資料：

[1] Sulman Sarwar. (2018-11-13) Transactions from a bakery (Dataset) <https://www.kaggle.com/sulmansarwar/transactions-from-a-bakery>

(last access:2019/06/28)

[2] Edward Yu. (2018-11-17) Bakery sales data exploration.

<https://www.kaggle.com/tastycanofmalk/bakery-sales-data-exploration>

(last access:2019/06/28)

[3] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).