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

2. 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

3. 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。

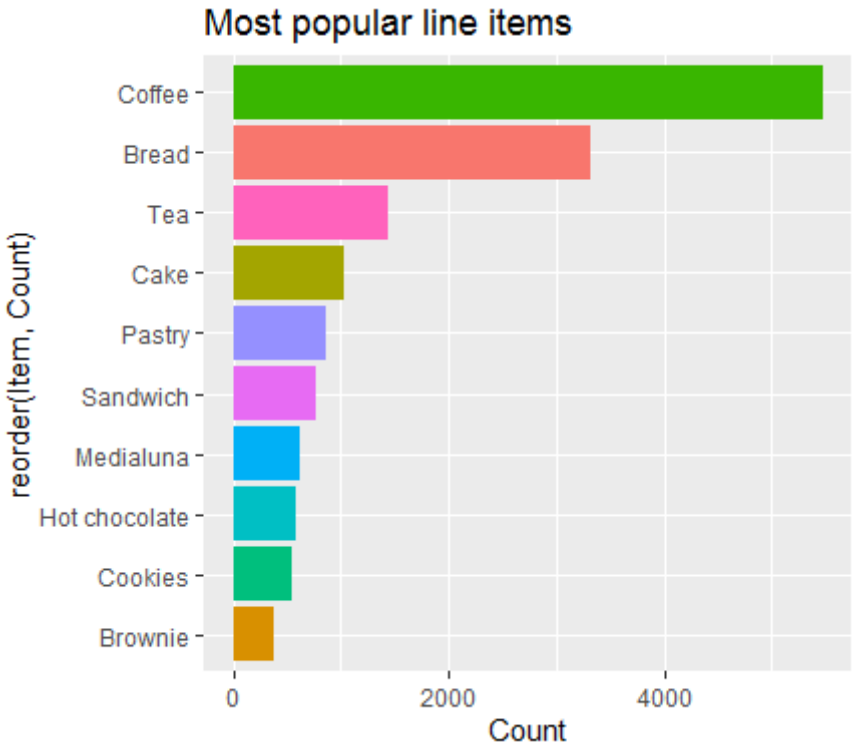

圖 1　商品銷售次數

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。

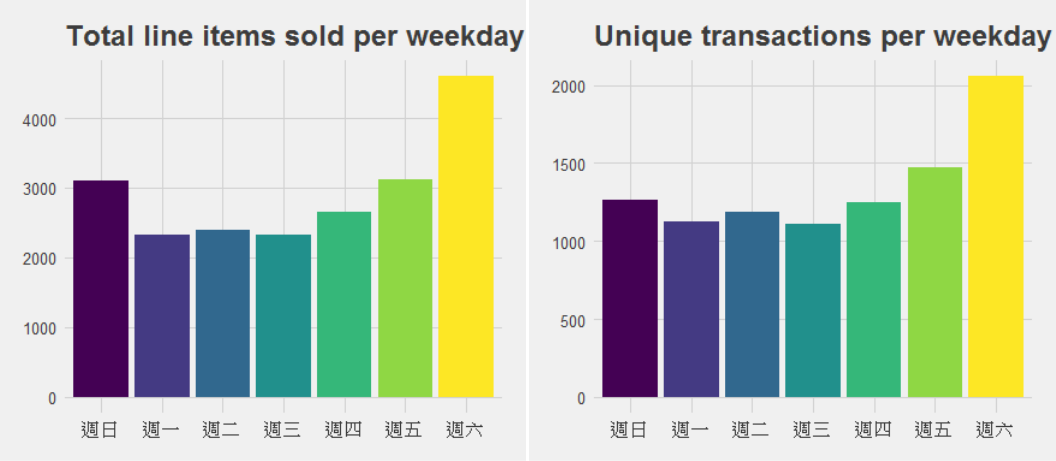

圖 2　每日商品銷售數(左)、每日交易數(右)

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。

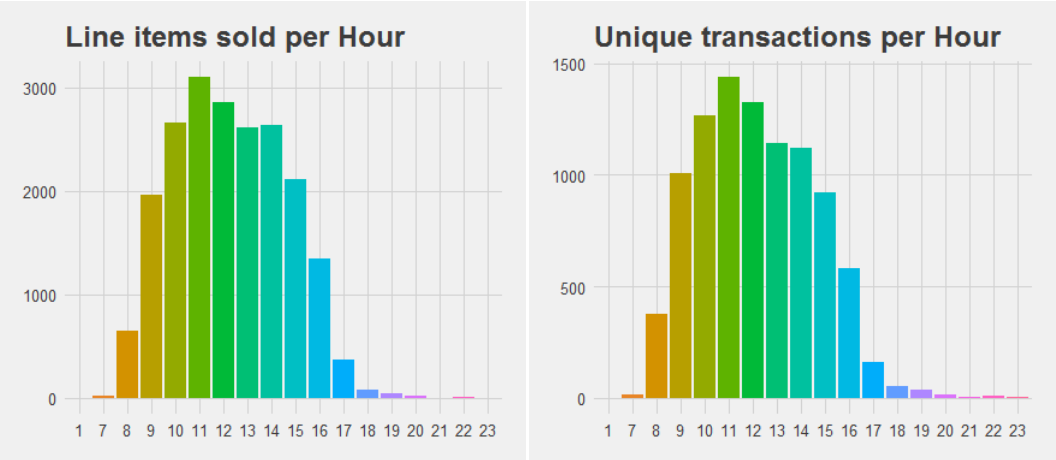

圖 3　每小時商品銷售數(左)、每小時交易數(右)

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

4. 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。

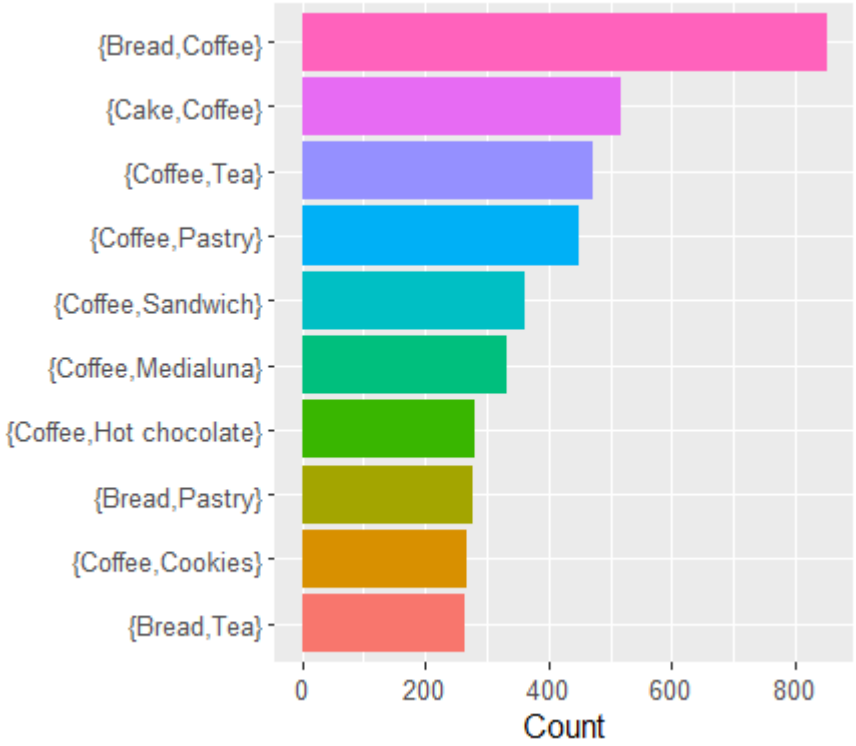

圖 4　前十項頻繁項目集

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。

	lhs		rhs	support	confidence	lift	count
[1]	{Toast}	=>	{Coffee}	0.02366614	0.7044025	1.472431	224
[2]	{Spanish Brunch}	=>	{Coffee}	0.01088220	0.5988372	1.251766	103
[3]	{Medialuna}	=>	{Coffee}	0.03518225	0.5692308	1.189878	333
[4]	{Pastry}	=>	{Coffee}	0.04754358	0.5521472	1.154168	450
[5]	{Alfajores}	=>	{Coffee}	0.01965135	0.5406977	1.130235	186
[6]	{Juice}	=>	{Coffee}	0.02060222	0.5342466	1.116750	195
[7]	{Sandwich}	=>	{Coffee}	0.03824617	0.5323529	1.112792	362
[8]	{Cake}	=>	{Coffee}	0.05472795	0.5269583	1.101515	518
[9]	{Scone}	=>	{Coffee}	0.01806656	0.5229358	1.093107	171
[10]	{Cookies}	=>	{Coffee}	0.02820919	0.5184466	1.083723	267
[11]	{Hot chocolate}	=>	{Coffee}	0.02958267	0.5072464	1.060311	280

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。

	lhs		rhs	support	confidence	lift	count
[1]	{Postcard}	=>	{Tshirt}	0.0006339144	0.6000000	270.428571	6
[2]	{Bread,Extra Salami or Feta}	=>	{Salad}	0.0006339144	0.6000000	57.363636	6
[3]	{Duck egg}	=>	{Spanish Brunch}	0.0006339144	0.5000000	27.514535	6
[4]	{Coffee,Crisps}	=>	{Juice}	0.0005282620	0.6250000	16.207192	5
[5]	{Coffee,Farm House,Pastry}	=>	{Medialuna}	0.0005282620	0.5000000	8.089744	5
[6]	{Coke,Mineral water}	=>	{Sandwich}	0.0006339144	0.5454545	7.592246	6
[7]	{Scone,Tiffin}	=>	{Tea}	0.0005282620	0.7142857	5.007937	5
[8]	{Pastry,Tiffin}	=>	{Tea}	0.0005282620	0.7142857	5.007937	5
[9]	{Brownie,Toast}	=>	{Tea}	0.0005282620	0.7142857	5.007937	5
[10]	{Alfajores,Spanish Brunch}	=>	{Tea}	0.0007395668	0.7000000	4.907778	7

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。

5. 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。

	Accuracy	Precision	Recall	F1-Score
Naive_Bayes	0.787	0.78	0.468	0.585

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

6. 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).