**IBM 資料產生器**

經過**反覆實驗**後，本研究猜測IBM產生的資料可能為**公司商品交易**的資料，第一直行為*交易ID*，第二直行為*商品名稱*(item)，未整合的資料如表1所示，整合後的資料如表2所示，將表2中IitemList的資料進行資料分析，找出其中的規則或模型

IBM產生器有個***隱藏限制***，就是***nitems要大於tlen***，假設參數nitems從10個item種類中抓資料，但參數tlen一筆交易一次可以產生20種item，那程式就會error(電腦無限當機，多試幾次才發現關鍵)

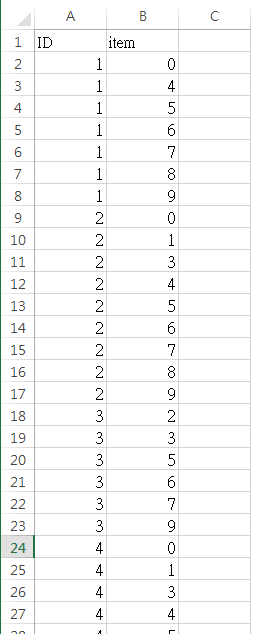
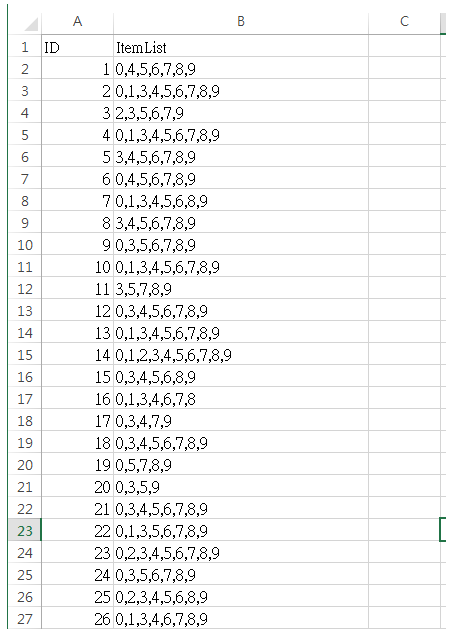
 

表1 表2

**標準版** IBM lit -ascii -ntrans 10 -tlen 4 -nitems 0.01 > DMsample1.txt

交易總數7554(將資料整合後)

項目總數30910(未將資料整合前)

**大量版** IBM lit -ascii -ntrans 200 -tlen 4 -nitems 0.01 > DMsample2.txt

交易總數151898

項目總數622452

**多種版**IBM lit -ascii -ntrans 10 -tlen 4 -nitems 0.1 > DMsample3.txt

交易總數7632

項目總數39190

**多項版**IBM lit -ascii -ntrans 10 -tlen 10 -nitems 0.1 > DMsample4.txt

交易總數9850

項目總數97482

**超大版**IBM lit -ascii -ntrans 100 -tlen 10 -nitems 0.1 > DMsample5.txt

交易總數98261

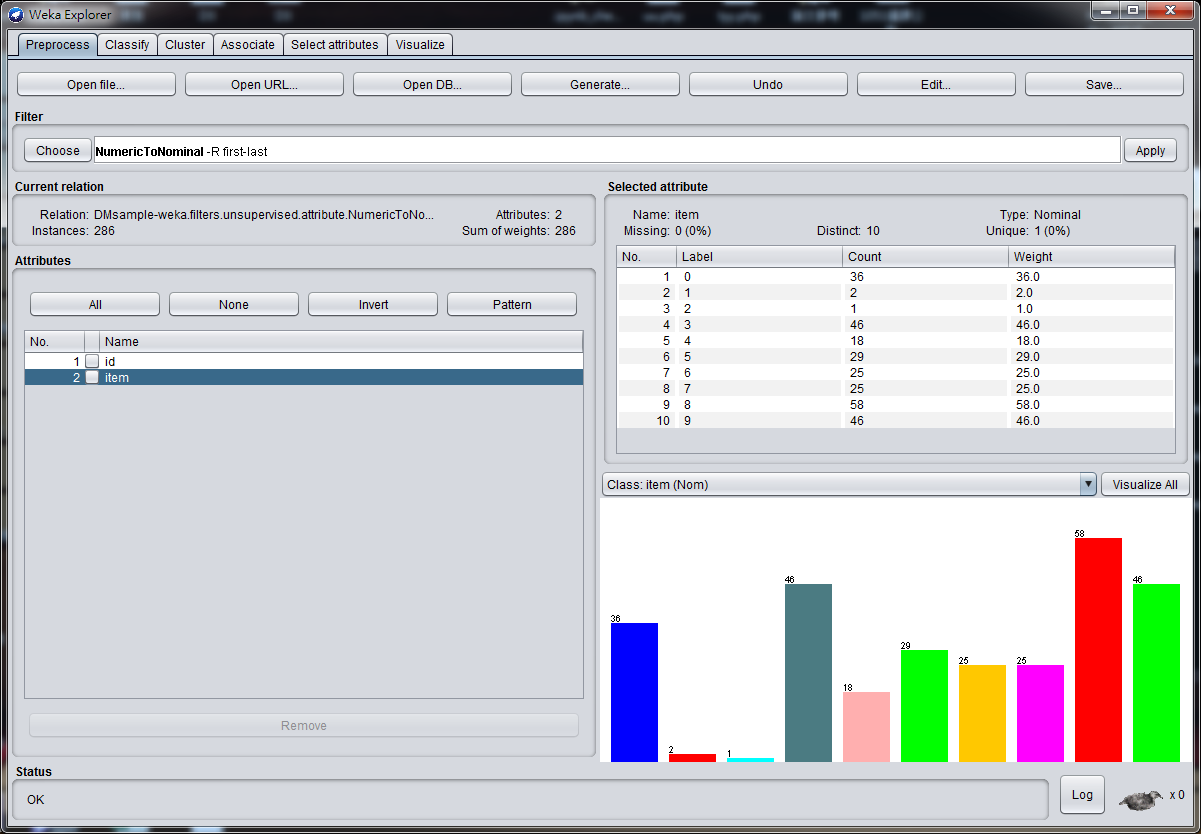
項目總數968221

**小型版** IBM lit -ascii -ntrans 0.1 -tlen 4 -nitems 0.01 > DMsample0.txt

交易總數77

項目總數287

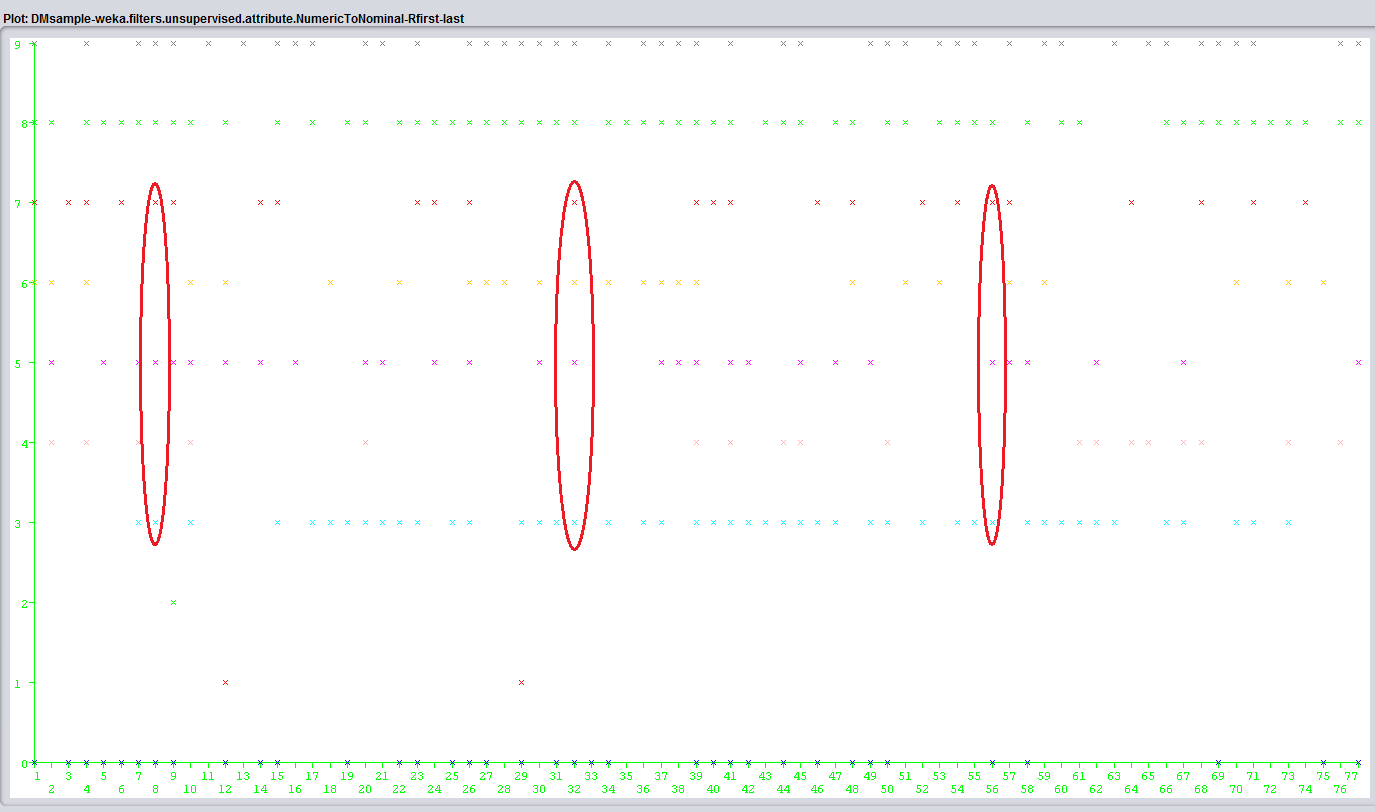
將**小型版**DMsample.txt放入weka分析，這邊要先用Numeric To Nominal 做前處理(Filter)，不能用數值型資料做分析，之後可以看到item種類有10個的**圖表，如圖1所示**

****

**圖1**

進行**Visualize**可看出**item分布狀況**，如圖2所示，X軸為交易ID(IBM的參數ntrans)，Y軸為item的10個種類(IBM的參數nitems)，每個交易ID平均有幾個item為每一直行的個數(IBM的參數tlen)，而我們需要去找所有交易(每一直行)互相比較後所**蘊含的模型或規則**，類似如圖2中用**紅圈括起來的3個圈**，但在weka中，要進行Association rule分析必須每個項目都是屬性attribute，所以目前的data(DMsample.txt)**只能用來圖示化顯示**，必須將資料做***格式轉換***後才能在weka和本實驗中用的R語言進行分析。

*提醒:圖2中的紅圈只是模擬舉例，不是實驗結果!*

****

**圖2**

**陣列轉檔**

Python3 :

conda create –n py3 python=3 anaconda

activate py3

ipython notebook

檔案python3-convert

或 用**IBM-Quest-Data-Converter**提供的功能去做**資料格式轉換**，轉換前的格式如圖3，轉換後的格式如表3

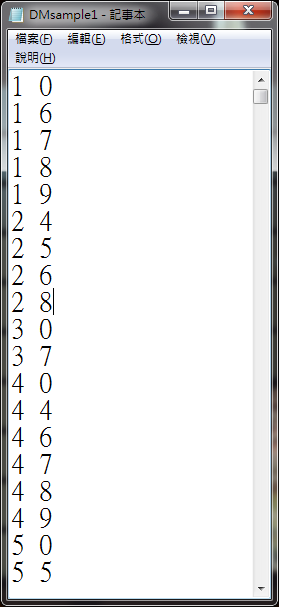
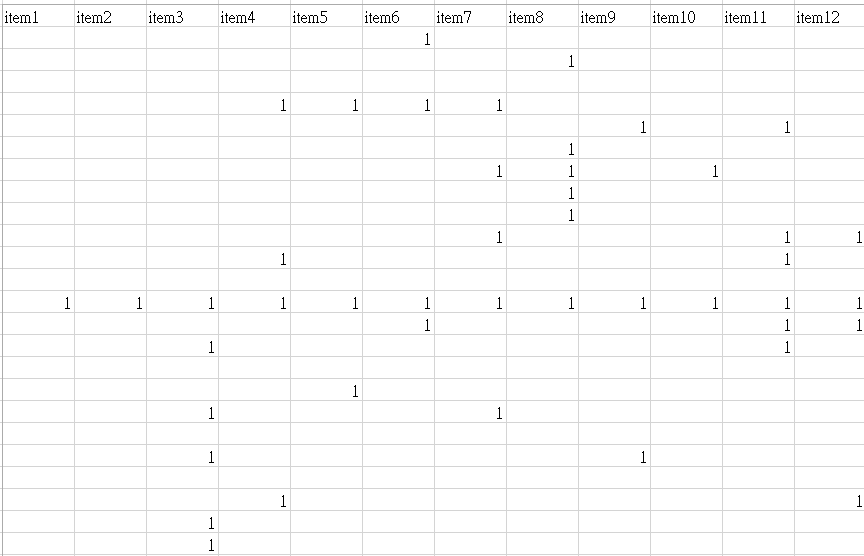
 

圖3 表3

表3中將每個item種類列出來，從item1到item10，每筆交易ID用1與0表示。

**之後產生檔案TDMsample1**、**TDMsample2、TDMsample3…，表示調整後資料格式**

**R語言Apriori分析**

***找包含至少10個rule的結果!***

**標準版**

**sup 0.1 conf 0.79**

***以conf由大到小***

lhs rhs support confidence lift

[1] {V3,V5,V10} => {V8} 0.1114641 0.8111753 1.097156

[2] {V5,V9,V10} => {V8} 0.1080222 0.8079208 1.092754

[3] {V3,V7,V10} => {V8} 0.1108022 0.8032630 1.086454

[4] {V3,V5,V7} => {V8} 0.1040508 0.8028601 1.085910

[5] {V4,V10} => {V8} 0.1045804 0.8004053 1.082589

**[6] {V3,V6,V9} => {V8} 0.1306593 0.7985437 1.080071**

[7] {V6,V10} => {V8} 0.1168917 0.7983725 1.079840

**[8] {V5,V10} => {V8} 0.1481334 0.7964413 1.077228**

**[9] {V3,V5,V9} => {V8} 0.1703733 0.7939543 1.073864**

[10] {V7,V9,V10} => {V8} 0.1086841 0.7917068 1.070824

**維持sup水準，conf降到0.79才找到10筆rule**

**嘗試把conf降低，sup拉高，但等找到rule時 conf已經低於0.5**

**買item5、3、10的客戶，有0.81的可能性也會買item8**

**買item5、9、10的客戶，有0.8的可能性也會買item8**

**買item3、7、10的客戶，有0.8的可能性也會買item8**

**買item5、3、7的客戶，有0.8的可能性也會買item8**

**買item4、10的客戶，有0.8的可能性也會買item8**

**紅色部分為較優規則，sup與conf都高**

set of 10 rules

rule length distribution (lhs + rhs):sizes

3 4

3 7

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.00 3.25 4.00 3.70 4.00 4.00

summary of quality measures:

support confidence lift

Min. :0.1041 Min. :0.7917 Min. :1.071

1st Qu.:0.1082 1st Qu.:0.7969 1st Qu.:1.078

Median :0.1111 Median :0.7995 Median :1.081

Mean :0.1214 Mean :0.8005 Mean :1.083

3rd Qu.:0.1272 3rd Qu.:0.8032 3rd Qu.:1.086

Max. :0.1704 Max. :0.8112 Max. :1.097

mining info:

data ntransactions support confidence

beer 7554 0.1 0.79

**----------------------------------**

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen

0.79 0.1 1 none FALSE TRUE 5 0.1 1 10

target ext

rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 755

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[10 item(s), 7554 transaction(s)] done [0.00s].

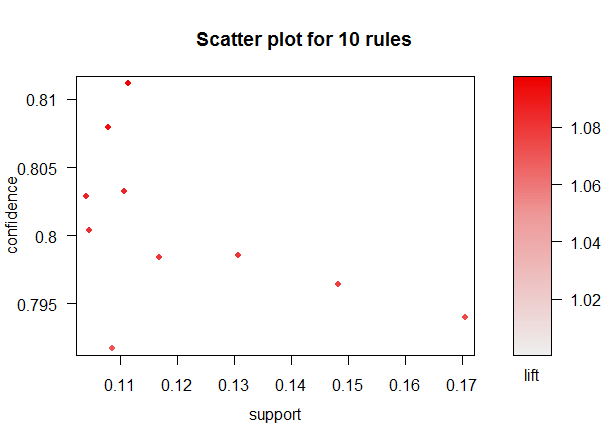
sorting and recoding items ... [8 item(s)] done [0.00s].

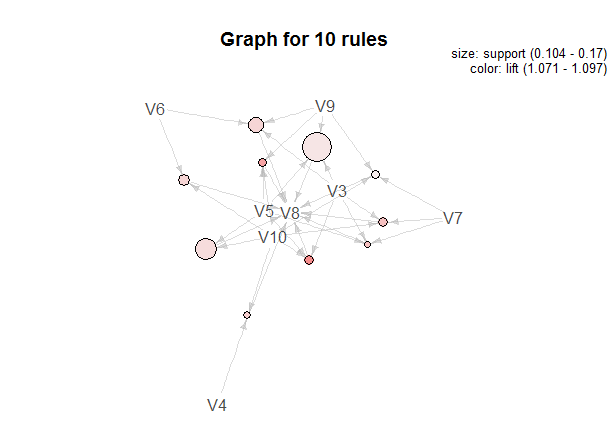
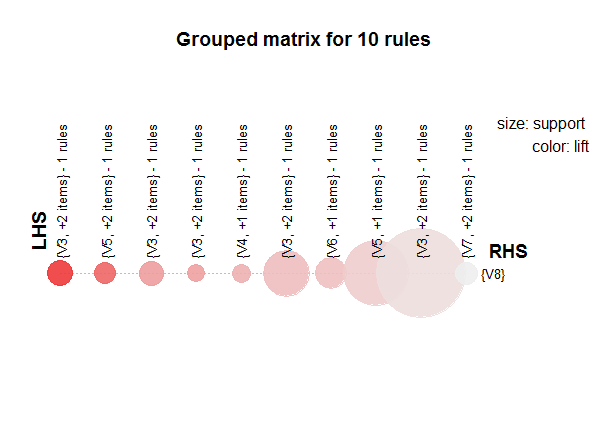
creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 4 5 done [0.00s].

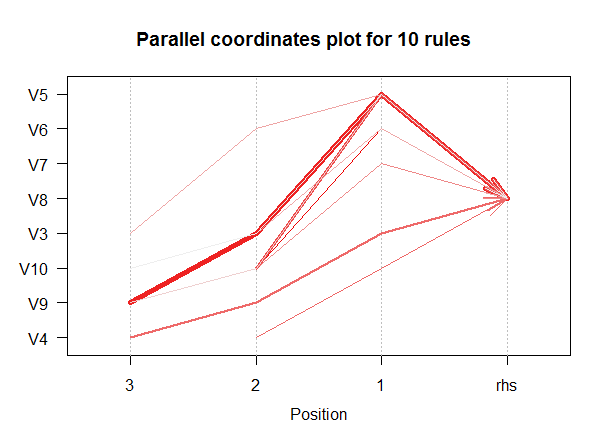
writing ... [10 rule(s)] done [0.00s].

creating S4 object ... done [0.00s].

****

****

**a=1**

****

**大量版**

**sup 0.1 conf 0.798**

***以conf由大到小***

lhs rhs support confidence lift

[1] {V3,V7,V10} => {V8} 0.1131878 0.8106082 1.092574

[2] {V4,V10} => {V8} 0.1084083 0.8101446 1.091949

[3] {V3,V5,V7} => {V8} 0.1079672 0.8091573 1.090618

[4] {V3,V5,V10} => {V8} 0.1134248 0.8080386 1.089110

[5] {V6,V10} => {V8} 0.1170193 0.8057205 1.085986

[6] {V7,V9,V10} => {V8} 0.1072825 0.8027982 1.082047

[7] {V5,V7,V9} => {V8} 0.1034510 0.8021849 1.081220

**[8] {V3,V4,V9} => {V8} 0.1238989 0.8007148 1.079239**

[9] {V5,V9,V10} => {V8} 0.1078684 0.7995023 1.077605

**[10] {V3,V6,V9} => {V8} 0.1308839 0.7988187 1.076683**

**買item7、3、10的客戶，有0.81的可能性也會買item8**

**買item4、10的客戶，有0.81的可能性也會買item8**

**買item3、7、5的客戶，有0.81的可能性也會買item8**

**買item5、3、10的客戶，有0.8的可能性也會買item8**

**買item6、10的客戶，有0.8的可能性也會買item8**

-------------------------------------------

**維持sup水準，conf降到0.798才找到10筆rule**

**嘗試把conf降低，sup拉高，但等找到rule時 conf已經低於0.5**

**Sup 0.5 conf 0.5**

lhs rhs support confidence lift

[4] {V3} => {V8} 0.5268733 0.7516906 1.013162

[5] {V8} => {V3} 0.5268733 0.7101431 1.013162

---------------------------------------------------------

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen

0.798 0.1 1 none FALSE TRUE 5 0.1 1 10

target ext

rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 15189

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[10 item(s), 151898 transaction(s)] done [0.03s].

sorting and recoding items ... [8 item(s)] done [0.00s].

creating transaction tree ... done [0.06s].

checking subsets of size 1 2 3 4 5 done [0.00s].

writing ... [10 rule(s)] done [0.00s].

creating S4 object ... done [0.00s].

**----------------------------------**

set of 10 rules

rule length distribution (lhs + rhs):sizes

3 4

2 8

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.0 4.0 4.0 3.8 4.0 4.0

summary of quality measures:

support confidence lift

Min. :0.1035 Min. :0.7988 Min. :1.077

1st Qu.:0.1079 1st Qu.:0.8011 1st Qu.:1.080

Median :0.1108 Median :0.8043 Median :1.084

Mean :0.1133 Mean :0.8048 Mean :1.085

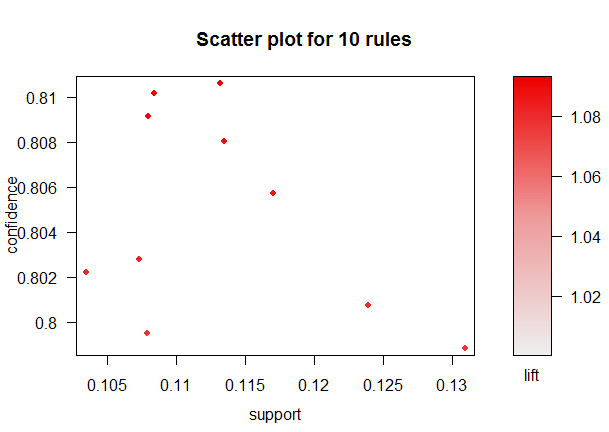
3rd Qu.:0.1161 3rd Qu.:0.8089 3rd Qu.:1.090

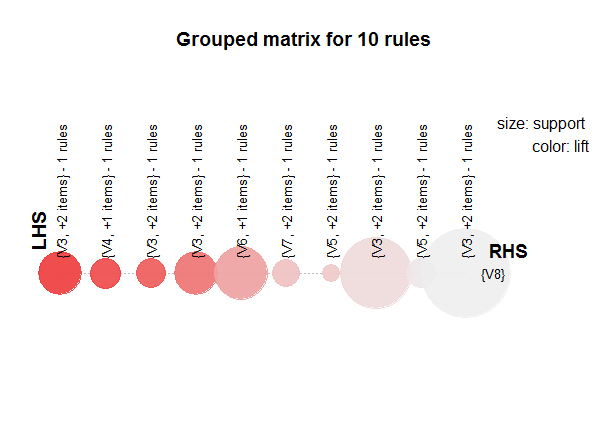
Max. :0.1309 Max. :0.8106 Max. :1.093

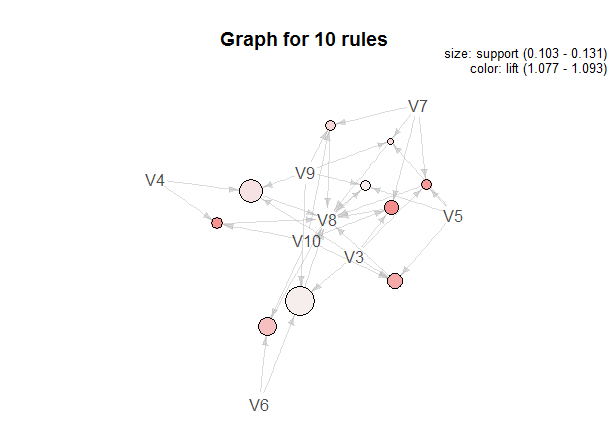
mining info:

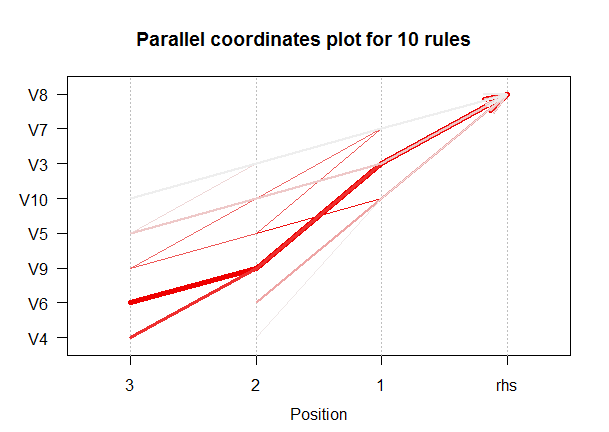
data ntransactions support confidence

beer 151898 0.1 0.798

****

****

****

****

**多種版**

**sup 0.0017 conf 0.74**

***以conf由大到小***

lhs rhs support confidence lift

[1] {V26,V38,V83} => {V63} 0.001965409 0.9375000 4.485893

[2] {V47,V49,V63} => {V87} 0.001703354 0.9285714 4.656279

[3] {V26,V63,V83} => {V38} 0.001965409 0.8823529 4.113694

[4] {V47,V49,V87} => {V63} 0.001703354 0.8666667 4.146959

**[5] {V4,V26} => {V63} 0.002489518 0.8260870 3.952787**

[6] {V86,V88} => {V69} 0.001703354 0.8125000 4.799536

**[7] {V9,V26,V69} => {V63} 0.002882600 0.7857143 3.759606**

[8] {V56} => {V3} 0.002096436 0.7619048 7.314286

**[9] {V9,V26,V63} => {V69} 0.002882600 0.7586207 4.481264**

**買item26、38、83的客戶，有0.93的可能性也會買item8**

**買item47、49、63的客戶，有0.92的可能性也會買item8**

**買item26、63、83的客戶，有0.88的可能性也會買item8**

**買item87、49、47的客戶，有0.86的可能性也會買item8**

**買item4、26的客戶，有0.82的可能性也會買item8**

**----------------------------------**

set of 9 rules

rule length distribution (lhs + rhs):sizes

2 3 4

1 2 6

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.000 3.000 4.000 3.556 4.000 4.000

summary of quality measures:

support confidence lift

Min. :0.001703 Min. :0.7586 Min. :3.760

1st Qu.:0.001703 1st Qu.:0.7857 1st Qu.:4.114

Median :0.001965 Median :0.8261 Median :4.481

Mean :0.002155 Mean :0.8400 Mean :4.634

3rd Qu.:0.002490 3rd Qu.:0.8824 3rd Qu.:4.656

Max. :0.002883 Max. :0.9375 Max. :7.314

mining info:

data ntransactions support confidence

beer 7632 0.0017 0.74

**----------------------------------**

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen

0.74 0.1 1 none FALSE TRUE 5 0.0017 1 10

target ext

rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 12

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[99 item(s), 7632 transaction(s)] done [0.00s].

sorting and recoding items ... [97 item(s)] done [0.00s].

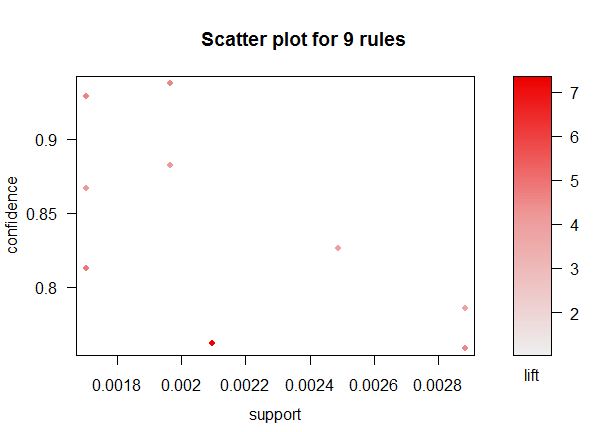
creating transaction tree ... done [0.01s].

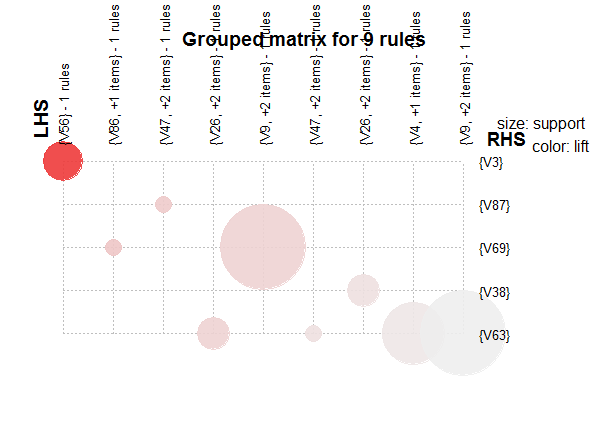
checking subsets of size 1 2 3 4 done [0.00s].

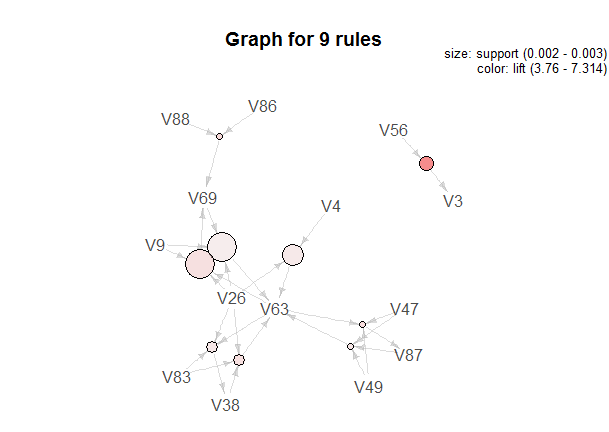
writing ... [9 rule(s)] done [0.00s].

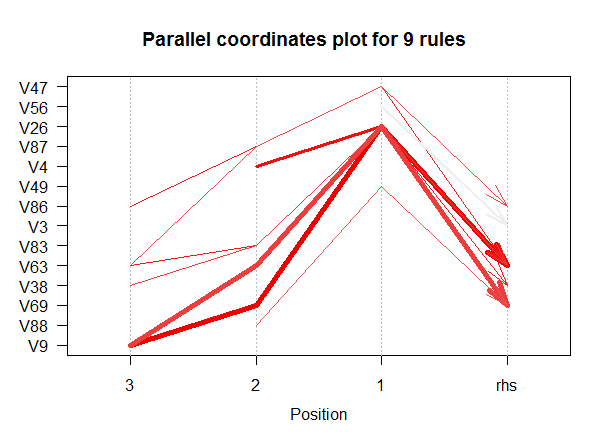
creating S4 object ... done [0.00s].

**----------------------------------**

****

****

****

****

**多項版**

**sup 0.00253 conf 0.9**

***以conf由大到小***

lhs rhs support confidence lift

[1] {V29,V56} => {V3} 0.002639594 0.9629630 4.940201

[2] {V29,V48,V79} => {V3} 0.002538071 0.9615385 4.932893

[3] {V28,V29,V49,V69} => {V38} 0.002538071 0.9615385 2.512242

**[4] {V48,V56} => {V3} 0.002842640 0.9333333 4.788194**

[5] {V28,V48,V49,V69,V87} => {V38} 0.002741117 0.9310345 2.432544

[6] {V3,V29,V79} => {V48} 0.002538071 0.9259259 3.890943

[7] {V48,V49,V87,V90} => {V38} 0.002538071 0.9259259 2.419196

[8] {V69,V87,V90,V97} => {V38} 0.002538071 0.9259259 2.419196

**[9] {V48,V73,V90} => {V38} 0.003350254 0.9166667 2.395004**

[10] {V48,V69,V73,V90} => {V38} 0.002741117 0.9000000 2.351459

[11] {V38,V69,V73,V90} => {V48} 0.002741117 0.9000000 3.781997

**----------------------------------**

set of 11 rules

rule length distribution (lhs + rhs):sizes

3 4 5 6

2 3 5 1

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.000 4.000 5.000 4.455 5.000 6.000

summary of quality measures:

support confidence lift

Min. :0.002538 Min. :0.9000 Min. :2.351

1st Qu.:0.002538 1st Qu.:0.9213 1st Qu.:2.419

Median :0.002640 Median :0.9259 Median :2.512

Mean :0.002704 Mean :0.9314 Mean :3.351

3rd Qu.:0.002741 3rd Qu.:0.9474 3rd Qu.:4.340

Max. :0.003350 Max. :0.9630 Max. :4.940

mining info:

data ntransactions support confidence

beer 9850 0.00253 0.9

**----------------------------------**

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen

0.9 0.1 1 none FALSE TRUE 5 0.00253 1 10

target ext

rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 24

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[99 item(s), 9850 transaction(s)] done [0.00s].

sorting and recoding items ... [98 item(s)] done [0.00s].

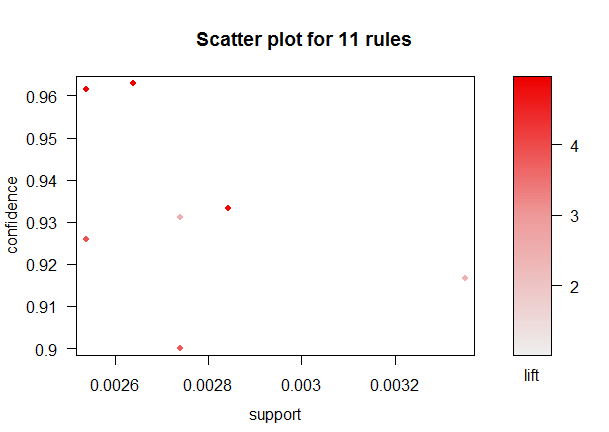
creating transaction tree ... done [0.00s].

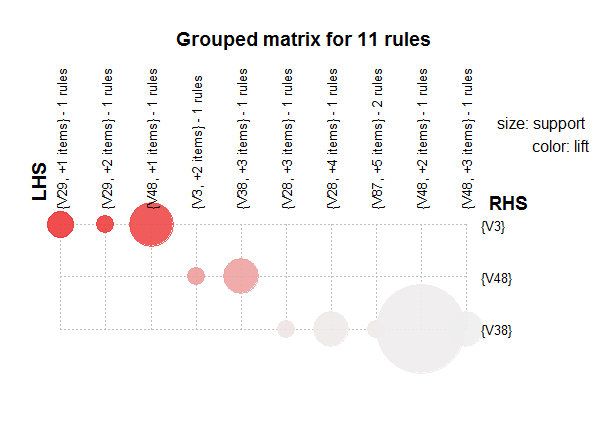
checking subsets of size 1 2 3 4 5 6 done [0.06s].

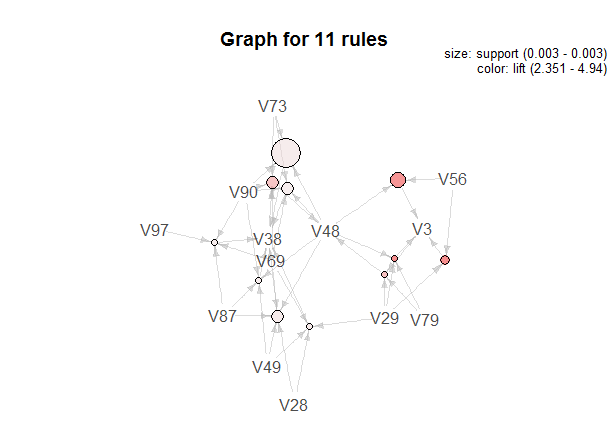
writing ... [11 rule(s)] done [0.00s].

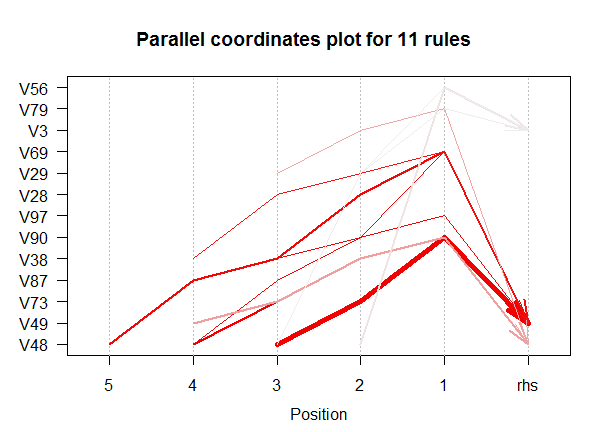
creating S4 object ... done [0.00s].

**----------------------------------**

****

****

****

****

**超大版**

**sup 0.002 conf 0.82**

***以conf由大到小***

lhs rhs support confidence lift

**[1] {V56,V89} => {V3} 0.002208404 0.9475983 4.826704**

**[2] {V29,V56} => {V3} 0.002228758 0.9358974 4.767105**

**[3] {V48,V56} => {V3} 0.002432298 0.8917910 4.542444**

[4] {V56,V89} => {V78} 0.002015042 0.8646288 5.989798

[5] {V56,V89} => {V48} 0.002015042 0.8646288 3.629188

[6] {V39,V67,V71,V78} => {V63} 0.002025219 0.8540773 2.286716

[7] {V78,V79,V89} => {V3} 0.002045573 0.8516949 4.338209

[8] {V29,V48,V79} => {V3} 0.002015042 0.8497854 4.328483

[9] {V35,V86,V88} => {V69} 0.002045573 0.8375000 2.671696

[10] {V28,V29,V48,V73} => {V38} 0.002045573 0.8237705 2.182793

**----------------------------------**

set of 10 rules

rule length distribution (lhs + rhs):sizes

3 4 5

5 3 2

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.0 3.0 3.5 3.7 4.0 5.0

summary of quality measures:

support confidence lift

Min. :0.002015 Min. :0.8238 Min. :2.183

1st Qu.:0.002018 1st Qu.:0.8503 1st Qu.:2.911

Median :0.002046 Median :0.8594 Median :4.333

Mean :0.002108 Mean :0.8721 Mean :3.956

3rd Qu.:0.002168 3rd Qu.:0.8850 3rd Qu.:4.711

Max. :0.002432 Max. :0.9476 Max. :5.990

mining info:

data ntransactions support confidence

beer 98261 0.002 0.82

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen

0.82 0.1 1 none FALSE TRUE 5 0.002 1 10

target ext

rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 196

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[99 item(s), 98261 transaction(s)] done [0.04s].

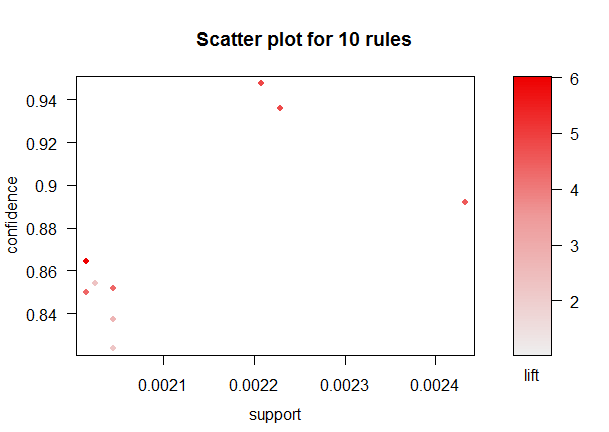
sorting and recoding items ... [98 item(s)] done [0.01s].

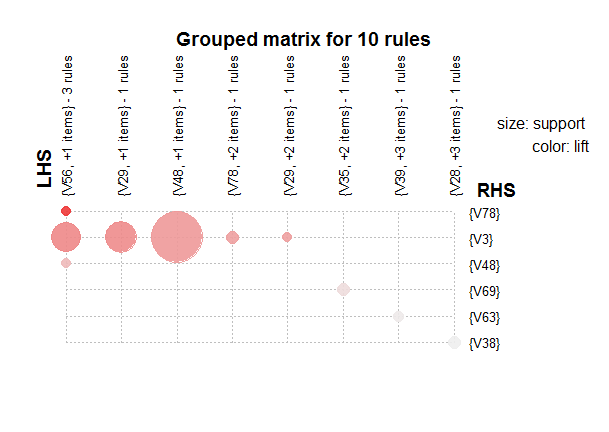
creating transaction tree ... done [0.05s].

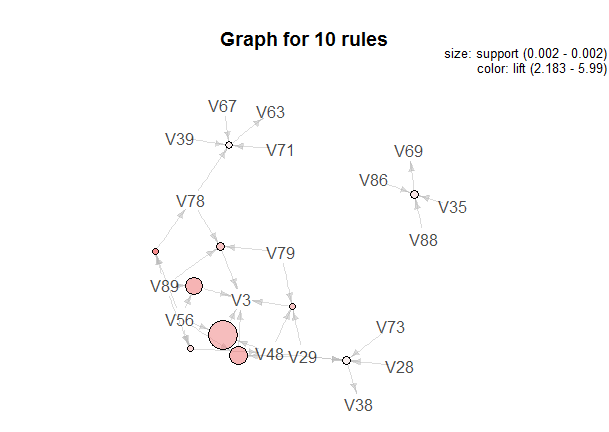
checking subsets of size 1 2 3 4 5 6 done [0.40s].

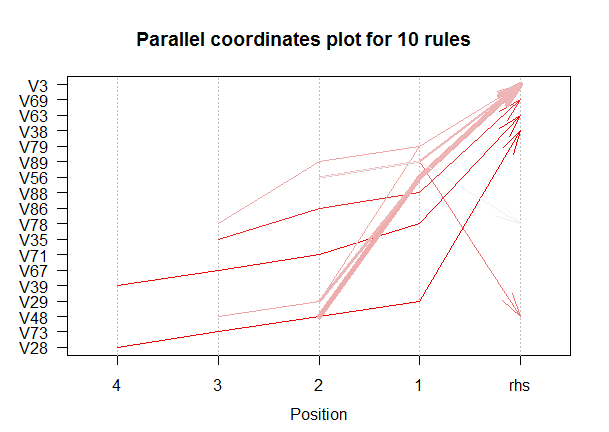
writing ... [10 rule(s)] done [0.01s].

creating S4 object ... done [0.01s].





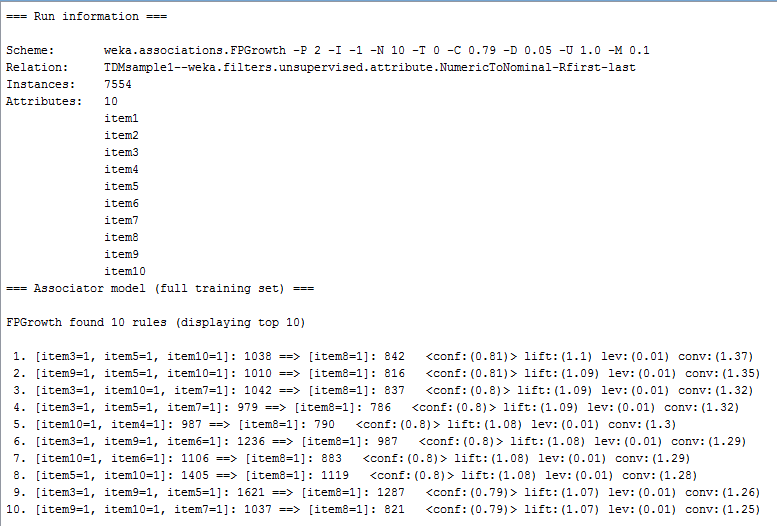




**Weka分析**

**標準版TDMsample1 FPGrowth 分析 sup 0.1 conf 0.79**

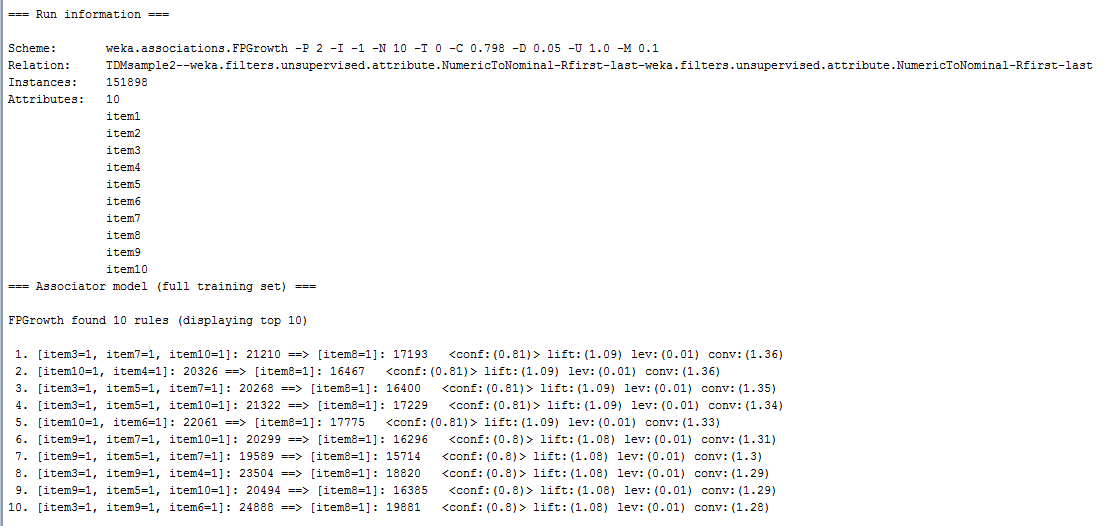
**維持sup水準，conf降到0.79才找到10筆rule**



**----------------------------------**

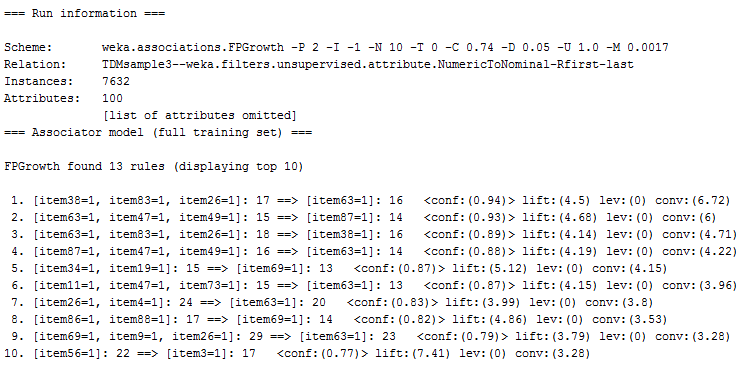
**大量版TDMsample2 FPGrowth 分析 sup 0.1 conf 0.798**

**維持sup水準，conf降到0.798才找到10筆rule**

****

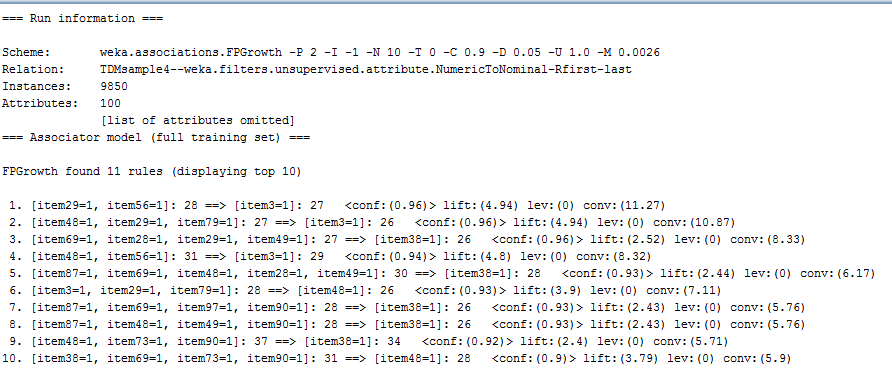
**----------------------------------**

**多種版TDMsample3 FPGrowth 分析 sup 0.0017 conf 0.74**

****

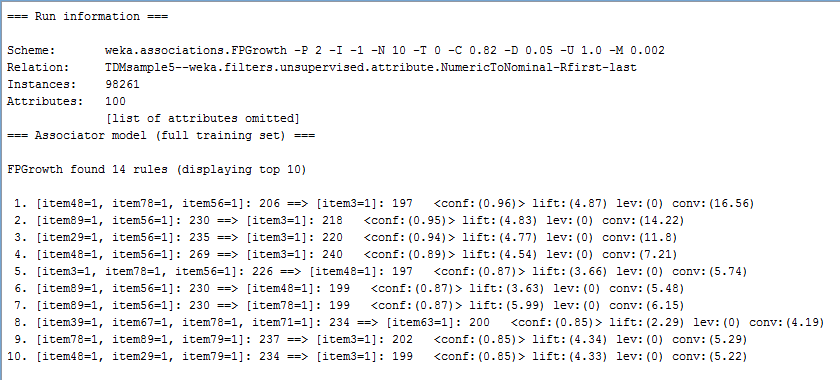
**----------------------------------**

**多項版TDMsample4 FPGrowth 分析 sup 0.0026 conf 0.9**

****

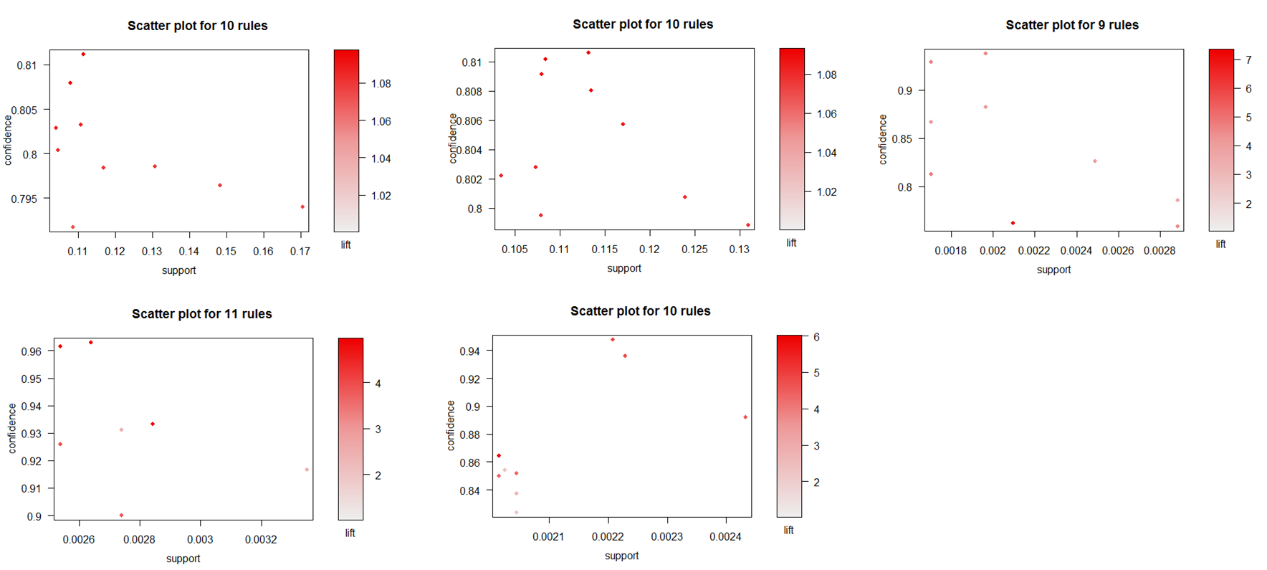
**----------------------------------**

**超大版TDMsample5 FPGrowth 分析 sup 0.002 conf 0.82**

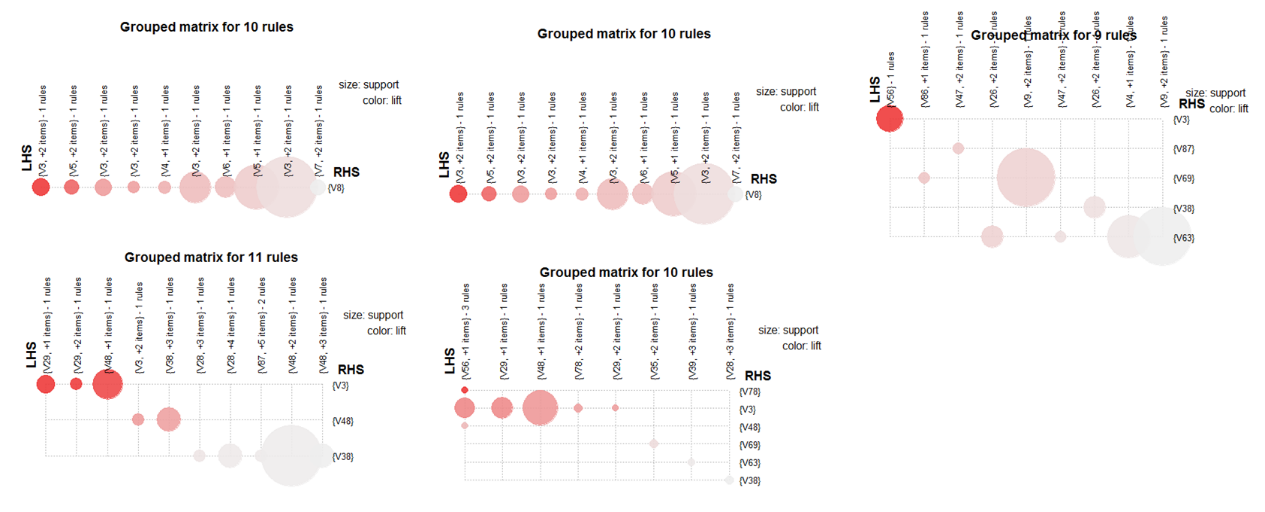
****

**數據很大時，min sup必須減低，不然找不到規則!**

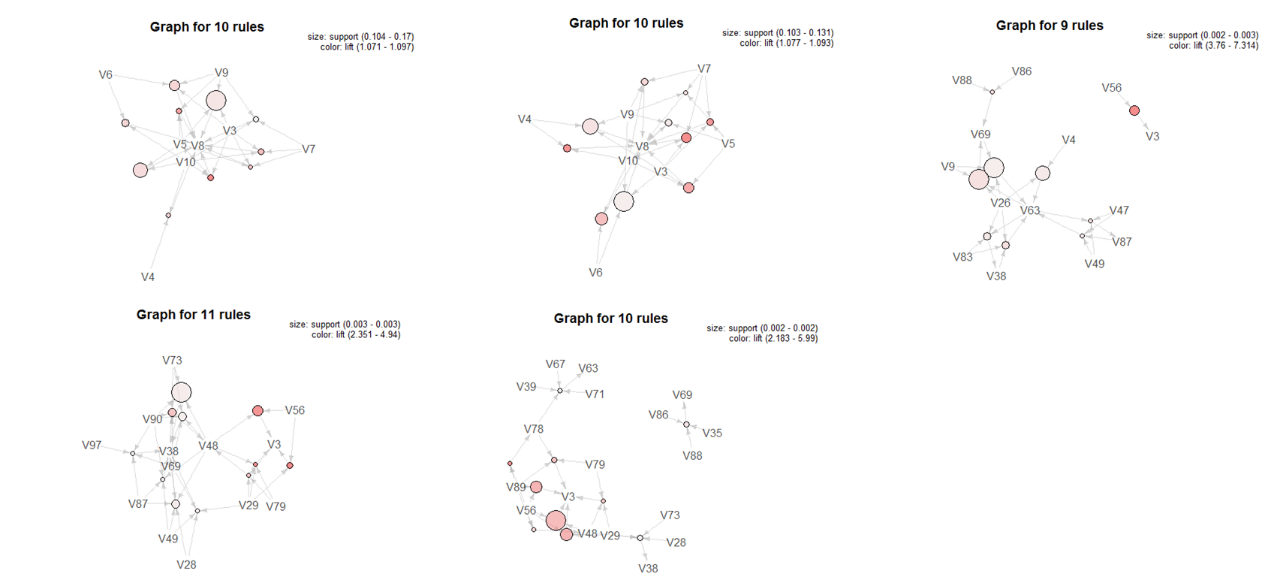
**程式碼的結果與weka得到的rule相似度非常高!**

****

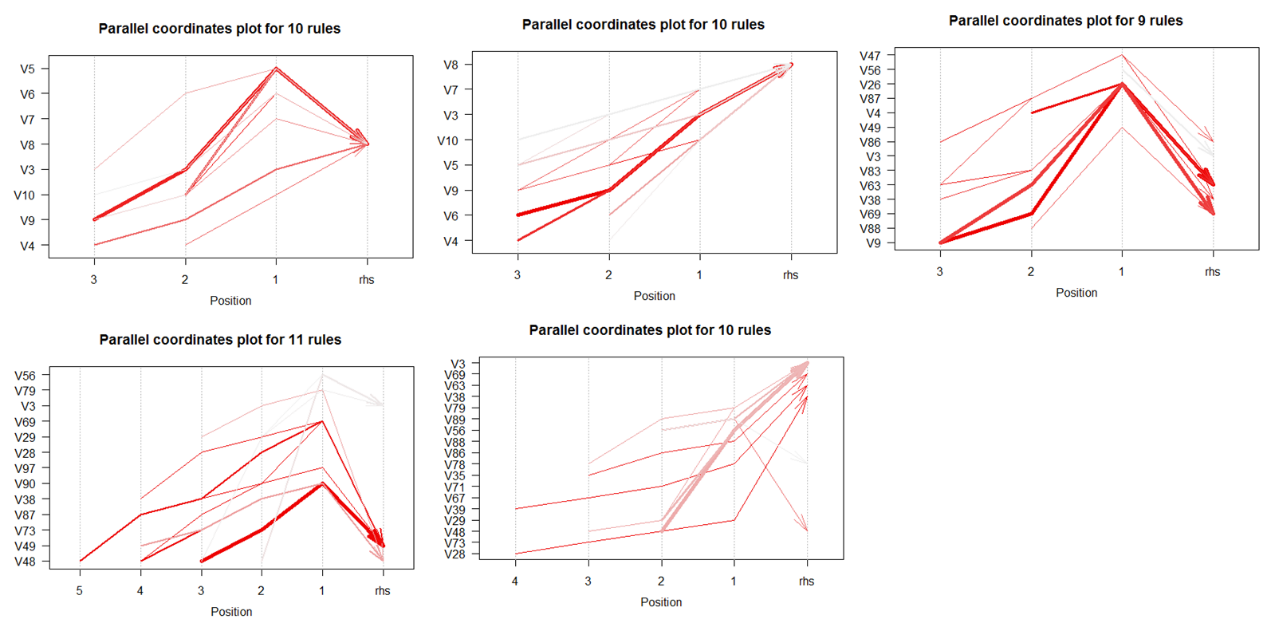
**標準版sup靠左，conf平均分布，大量版sup和conf都快接近平均分布，多種版sup和conf數值都提升，lift暴增，多項版sup和conf又增加，lift下降，超大版的規則很多分布在左下角，規則彼此間的差距在sup、conf、lift中都更加明顯**

****

**標準版許多item規則都指向V8，大量版變化不大，多種版將種類分散了，多項版簡化了種類的影響力，超大版則有點像標準版的數據放大化**

****

**標準版規則彼此之間的關聯性強且集中，大量版有幾個規則特色被放大、更清楚，差異性更加凸顯，多種版出現數個新規則，且關聯性成各占一方的情勢，多項版增加了多種版規則與規則間的連結，讓規則間有更多指向性，互相影響，超大版因大量數據，使得規則間的特色與關係已超清楚的方式凸顯出來**

****

**標準版rule在1時出現轉折，最後匯集，大量版趨勢穩定向上，多種版在1時出現轉折，且彼此匯集點不同，多項版在1時仍有轉折，但彼此最後匯集，超大版類似大量版穩定朝上匯集，但有一個例外**