ECT_HW3

2019

第一大題

第一大題

對 CreditCardPromotion進行 Association Rule , 並使用 Apriori 演算法,設定 confidence = 0.9、minimum support = 0.2 , 並回答以下問題：

第一大題(a)-題目

用 Weka 軟體：

(a)請嘗試著修改 CreditCardPromotion.arff 的欄位與上圖相同,使其可以執行 Association Rule, 請說明使用的方法以及解釋原來的檔案不能執行的原因? (10%)

第一大題(a)-解答

- **方法一**：透過文字編輯器開啟，並將資料修改為右方圖片所示
- **方法二**：在Weka利用前處理

NumericToNominal的方法將欄位數值轉換成對應的nominal數值，再用文字編輯器修改其中的值。

- 原因：Aprioir演算法要求其處理的資料欄位皆為Nominal

Relation: CreditCardPromotion

No.	1: Income Range	2: Magazine Promotion	3: Credit Card Insurance	4: **Sex**
	Nominal	Nominal	Nominal	Nominal
1	40-50000	Yes	No	Male
2	40-50000	No	No	Male
3	20-30000	Yes	Yes	Male
4	30-40000	Yes	Yes	Male
5	20-30000	No	No	Male
6	30-40000	Yes	No	Male
7	30-40000	Yes	No	Female
8	50-60000	Yes	No	Female
9	20-30000	No	No	Female
10	30-40000	Yes	No	Female

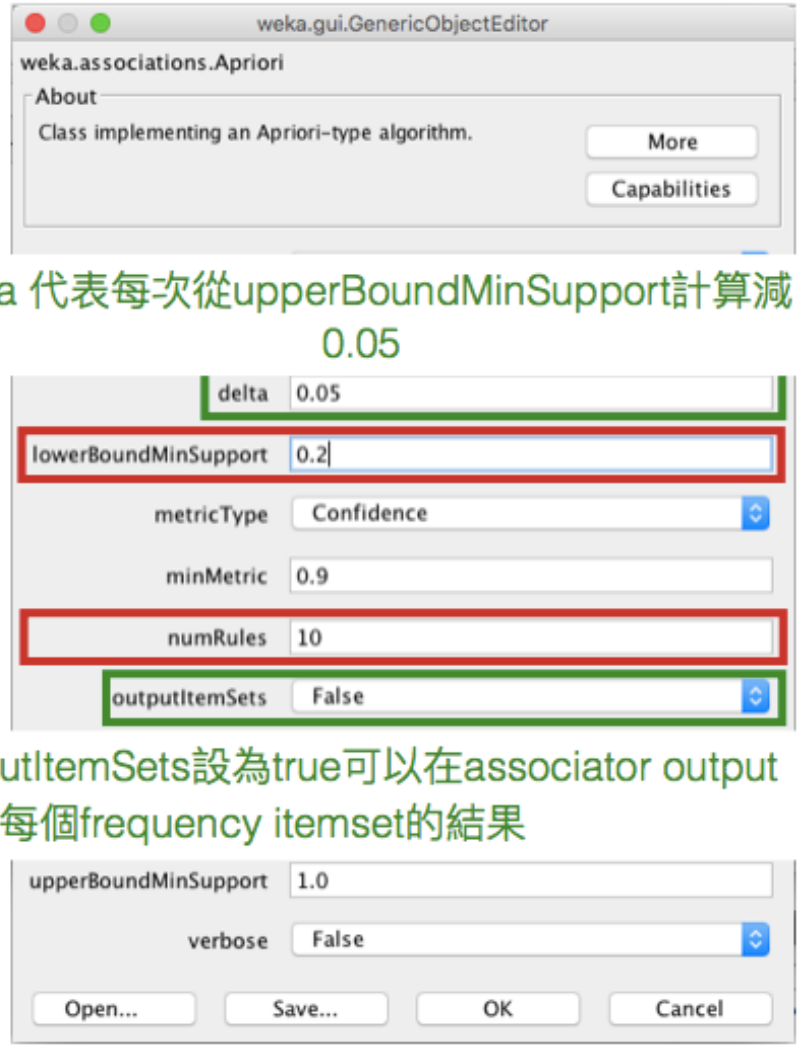
第一大題(b)-題目

(b)請將 numRule 設成5和10,其各別執行後的 Minimum support 為何,請比較兩者並說明造成其差異的原因。(15%)

第一大題(b)-解答

```
Apriori                                  Apriori
=======                                  =======

Minimum support: 0.25 (3 instances)      Minimum support: 0.35 (3 instances)
Minimum metric <confidence>: 0.9         Minimum metric <confidence>: 0.9
Number of cycles performed: 15           Number of cycles performed: 13
```



delta 代表每次從upperBoundMinSupport計算減0.05

outputItemSets設為true可以在associator output看到每個frequency itemset的結果

- 規則數設為10的Minimum support值0.25小於規則數設為5的Minimum support值0.35。其原因在於想要找尋的rule數較多，必須放寬每次篩選通過的數目，所以Minimum support數值才會比較低，使找到的規則更容易進入下一階段的篩選，最後找到的rule總數也會比較多，反之5條rules，則篩選通過的數目不需要那麼多，門檻就可以拉高。Minimum support數值才會相對高。

第一大題(c)-題目

(c)將 numRule 設成10, 列出前5條rule(15%)

第一大題(c)-解答

```
Best rules found:

 1. Income Range=30-40000 4 ==> Magazine Promotion=Yes 4    conf:(1)
 2. Sex=Female 4 ==> Credit Card Insurance=No 4    conf:(1)
 3. Magazine Promotion=No 3 ==> Credit Card Insurance=No 3    conf:(1)
 4. Income Range=30-40000 Credit Card Insurance=No 3 ==> Magazine Promotion=Yes 3
 5. Magazine Promotion=Yes Sex=Female 3 ==> Credit Card Insurance=No 3    conf:(1)
 6. Income Range=40-50000 2 ==> Credit Card Insurance=No 2    conf:(1)
 7. Income Range=40-50000 2 ==> Sex=Male 2    conf:(1)
 8. Credit Card Insurance=Yes 2 ==> Magazine Promotion=Yes 2    conf:(1)
 9. Credit Card Insurance=Yes 2 ==> Sex=Male 2    conf:(1)
10. Income Range=20-30000 Credit Card Insurance=No 2 ==> Magazine Promotion=No 2
```

從預設結果可得到10條Confidence皆為1的結果

並從中列出五條規則：

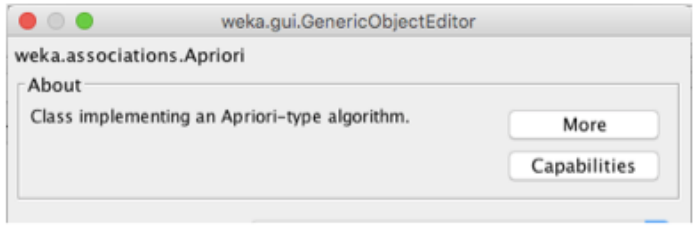
1. if Income Range=30-40000 then Magazine Promotion=Yes
2. if Sex=Female then Credit Card Insurance=No
3. if Magazine Promotion=No then Credit Card Insurance=No
4. if Income Range=30-40000 and Credit Card Insurance=No then Magazine Promotion=Yes
5. if Magazine Promotion=Yes and Sex=Female then Credit Card Insurance=No

第一大題(d)-題目

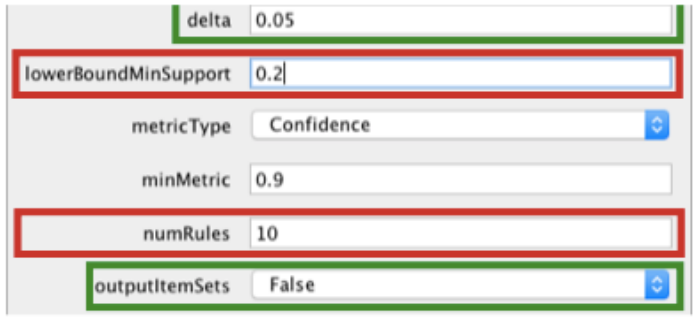
(d)如何在 Associator output 產生 Itemset, 請截圖說明並附上 Itemset 結果。(15%)

第一大題(d)-解答

呈現frequency itemset方式



delta 代表每次從upperBoundMinSupport計算減0.05



outputItemSets設為true可以在associator output看到每個frequency itemset的結果


```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Large Itemsets L(1):
Income Range=20-30000 3
Income Range=30-40000 4
Income Range=40-50000 2
Magazine Promotion=Yes 7
Magazine Promotion=No 3
Credit Card Insurance=Yes 2
Credit Card Insurance=No 8
Sex=Male 6
Sex=Female 4

Size of set of large itemsets L(2): 18

Large Itemsets L(2):
Income Range=20-30000 Magazine Promotion=No 2
Income Range=20-30000 Credit Card Insurance=No 2
Income Range=20-30000 Sex=Male 2
Income Range=30-40000 Magazine Promotion=Yes 4
Income Range=30-40000 Credit Card Insurance=No 3
Income Range=30-40000 Sex=Male 2
Income Range=30-40000 Sex=Female 2
Income Range=40-50000 Credit Card Insurance=No 2
Income Range=40-50000 Sex=Male 2
Magazine Promotion=Yes Credit Card Insurance=Yes 2
Magazine Promotion=Yes Credit Card Insurance=No 5
Magazine Promotion=Yes Sex=Male 4
Magazine Promotion=Yes Sex=Female 3
Magazine Promotion=No Credit Card Insurance=No 3
Magazine Promotion=No Sex=Male 2
Credit Card Insurance=Yes Sex=Male 2
Credit Card Insurance=No Sex=Male 4
Credit Card Insurance=No Sex=Female 4

Size of set of large itemsets L(3): 10

Large Itemsets L(3):
Income Range=20-30000 Magazine Promotion=No Credit Card Insurance=No 2
Income Range=30-40000 Magazine Promotion=Yes Credit Card Insurance=No 3
Income Range=30-40000 Magazine Promotion=Yes Sex=Male 2
Income Range=30-40000 Magazine Promotion=Yes Sex=Female 2
Income Range=30-40000 Credit Card Insurance=No Sex=Female 2
Income Range=40-50000 Credit Card Insurance=No Sex=Male 2
Magazine Promotion=Yes Credit Card Insurance=Yes Sex=Male 2
Magazine Promotion=Yes Credit Card Insurance=No Sex=Male 2
Magazine Promotion=Yes Credit Card Insurance=No Sex=Female 3
Magazine Promotion=No Credit Card Insurance=No Sex=Male 2

Size of set of large itemsets L(4): 1

Large Itemsets L(4):
Income Range=30-40000 Magazine Promotion=Yes Credit Card Insurance=No Sex=Female 2
```

第一大題(e)-題目

用 Python ：

(e)將已修改過的CreditCardPromotion.arff轉成csv檔，使用 Apriori 演算法進行分析,設定 confidence = 0.9、minimum support = 0.2，過程中對所有重要程式步驟進行截圖並加以說明，越詳盡越好。(15%)

第一大題(e)-解答

- 讀取資料集，轉成list

```
with open('CreditCardPromotion_v1.csv', 'r') as csvfile:
  data = csv.reader(csvfile)
  data_list = list(data)

print(data_list)
```

- Apriori參數設定，並將結果匯出CSV檔

```
result=(list(apriori(data_list, min_support=0.2, min_confidence=0.9)))
df=pd.DataFrame(result)
df.to_csv("appriori_homework.csv")
print(df.head(10))
```

- 結果顯示

```
              items   support                                ordered_statistics
0  (Yes, 30-40000)  0.222222  [((30-40000), (Yes), 1.0, 2.571428571428571)]
1     (Female, No)  0.222222                  [((Female), (No), 1.0, 2.25)]
```

第一大題(f)-題目

(f)調整apriori( )內的參數，產生與(c) 小題一樣的結果，截圖並加以說明(15%)

第一大題(f)-解答

- Apriori參數設定，調整信賴度

```
result=(list(apriori(data_list, min_confidence=1)))
```

- 將結果排序

```
print(df.sort_values(by='support',ascending=False))
```

```
                          items   support  \
0               (30-40000, Yes)  0.222222
3                  (Female, No)  0.222222
7           (30-40000, Yes, No)  0.166667
9             (Female, Yes, No)  0.166667
1              (Male, 40-50000)  0.111111
2                (40-50000, No)  0.111111
4        (Female, 30-40000, No)  0.111111
5       (Female, 30-40000, Yes)  0.111111
6         (Male, 30-40000, Yes)  0.111111
8          (Male, 40-50000, No)  0.111111
10  (Female, 30-40000, Yes, No)  0.111111

                                ordered_statistics
0       [((30-40000), (Yes), 1.0, 2.571428571428571)]
3                      [((Female), (No), 1.0, 2.25)]
7   [((30-40000, No), (Yes), 1.0, 2.571428571428571)]
9                  [((Female, Yes), (No), 1.0, 2.25)]
1                    [((40-50000), (Male), 1.0, 3.0)]
2                    [((40-50000), (No), 1.0, 2.25)]
4             [((Female, 30-40000), (No), 1.0, 2.25)]
5   [((Female, 30-40000), (Yes), 1.0, 2.5714285714...
6   [((Male, 30-40000), (Yes), 1.0, 2.571428571428...
8   [((Male, 40-50000), (No), 1.0, 2.25), ((40-500...
10  [((Female, 30-40000, No), (Yes), 1.0, 2.571428...
```

第一大題(g)-題目

(e)請自己計算 (記錄在 Word 上或手算拍照附圖皆可),並與 (d)小題結果做驗證。(15%)

第一大題(g)-解答

minimum題目假設為0.2，表示每個frequency itemsets要大於或等於10*0.2=2 才會成立

1 itemset:

(Income-Range=40-50000 2)
(Income-Range=30-40000 4)
~~(Income-Range=50-60000 1)~~
(Income-Range=20-30000 3)
(Magazine-Prom=Yes 7)
(Magazine-Prom=No 3)
(Credit-Card-Ins=Yes 2)
(Credit-Card-Ins=No 8)
(Sex=Male 6)
(Sex=Female 4)

第二大題(g)-解答

2 itemset:

~~(Income-Range=40-50000, Magazine-Prom=Yes 1)~~
~~(Income-Range=40-50000, Magazine-Prom=No 1)~~
(Income-Range=30-40000, Magazine-Prom=Yes 5)
~~(Income-Range=30-40000, Magazine-Prom=No 1)~~
~~(Income-Range=20-30000, Magazine-Prom=Yes 1)~~
(Income-Range=20-30000, Magazine-Prom=No 2)
~~(Income-Range=40-50000, Credit-Card=Yes 1)~~
(Income-Range=40-50000, Credit-Card=No 2)
(Income-Range=30-40000, Credit-Card=Yes 2)
(Income-Range=30-40000, Credit-Card=No 2)
~~(Income-Range=20-30000, Credit-Card=Yes 1)~~
(Income-Range=20-30000, Credit-Card=No 2)
(Magazine-Prom=Yes, Sex=Male 4)
(Magazine-Prom=Yes, Sex=Female 3)
(Magazine-Prom=Yes, Credit-Card=Yes 2)
(Magazine-Prom=Yes, Credit-Card=No 5)
~~(Magazine-Prom=No, Credit-Card=Yes 1)~~
(Magazine-Prom=No, Credit-Card=No 3)
~~(Income-Range=20-30000, Sex=Male 1)~~
~~(Income-Range=20-30000, Sex=Female 1)~~
(Income-Range=40-50000, Sex=Male 3)
~~(Income-Range=40-50000, Sex=Female 1)~~
(Income-Range=30-40000, Sex=Male 3)
(Income-Range=30-40000, Sex=Female 2)
(Magazine-Prom=No, Sex=Male 2)
~~(Magazine-Prom=No, Sex=Female 1)~~
(Credit-Card=Yes, Sex=Male 2)
(Credit-Card=No, Sex=Male 4)
(Credit-Card=No, Sex=Female 4)

計算到此就可以去找尋符合大於Confidence的Rule了，當然也可以繼續尋找3 itemset的部分…

第二大題(g)-解答

3itemset:

(Income Range=20-30000, Credit-Card=No Magazine Prom=No 2)
(Income Range=30-40000, Credit-Card=Yes Magazine Prom=No 3)
(Income Range=30-40000, Sex=Male, Magazine Prom=Yes 2)
(Income Range=30-40000, Sex=Female, Magazine Prom=Yes 2)
(Income Range=30-40000, Sex=Male, Credit-Card=No 2)
(Income Range=40-50000, Sex=Female, Credit-Card=No 2)
(Magazine Prom=Yes, Credit-Card=Yes, Sex=Male 2)
(Magazine Prom=Yes, Credit-Card=No, Sex=Male 2)
(Magazine Prom=Yes, Credit-Card=No, Sex=Female 3)
(Magazine Prom=No, Credit-Card=No, Sex=Male 2)

Rule:

1. IF Income-Range=30-40000 THEN Magazine-Prom=Yes 5 , Conf=5/5=1
2. IF Sex=Female THEN Creadit-Card=No 4 , Conf=4/4=1
3. IF Magazine Prom=No THEN Credit Card=NO 3, Conf=3/3=1
4. IF Magazine Prom=Yes , Sex=Female 3 THEN Credit Card=No 3, Conf=3/3=1
5. IF Income Range=30-40000 , Credit Card=No THEN Magazine Prom=Yes 3, Conf=3/3=1