

Lab02-Mining Frequent Itemsets and Association Rules

Data Mining -Term I /2020-2021

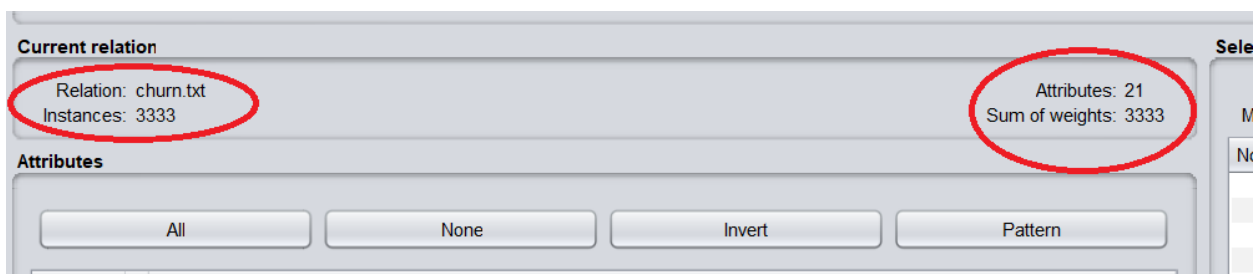
Thông tin thành viên

Họ Tên	MSSV
Trần Ngọc Tịnh	18120597
Nguyễn Ngọc Năng Toàn	18120600

I Báo cáo Data

1 Nhận biết Dữ liệu ban đầu thông qua file Mô tả dữ liệu

Tập dữ liệu churn.txt có 21 thuộc tính & 3333 mẫu



+State: Categorical .Mô tả tên 50 tiểu bang và quận của Columbia

+Account length: Integer-valued .Tài khoản đã hoạt động được bao lâu

+Area code: Categorical . Mã vùng

+Phone number: Categorical . Số điện thoại của khách hàng -> Đại diện cho ID của khách hàng

+International Plan: Dichotomous Categorical. Phân loại nhị phân có 2 giá trị của thuộc tính “ Có” Và “Không ” tham gia

+VoiceMail Plan: Dichotomous Categorical. Phân loại nhị phân có 2 giá trị của thuộc tính “ Có” Và “Không ” tham gia

+Number of voice mail messages: Integer-Valued . Số lượng thư thoại

+Total day minutes: Continuous . Tổng số phút khách hàng sử dụng trong ngày

+Total day calls: Integer-valued . Tổng số cuộc gọi trong ngày

+Total day charge: Continuous . Tổng số phí trong ngày

+Total evening minutes: Continuous . Tổng số phút khách hàng sử dụng trong tối

+Total evening calls: Integer-valued . Tổng số cuộc gọi trong tối

+Total evening charge: Continuous . Tổng số phí trong tối

+Total night minutes: Tổng số phút khách hàng sử dụng ban đêm

+Total night calls: Integer-valued . Tổng số cuộc gọi ban đêm

+Total night charge: Continuous . Tổng số phí sử dụng ban đêm

+Total international minutes: Tổng số phút khách hàng sử dụng liên lạc quốc tế

+Total international calls: Tổng số cuộc gọi quốc tế

+Total international charge: Tổng số phí cuộc gọi quốc tế

+Number of calls to customer service: Integer-valued . Số cuộc gọi đến dịch vụ khách hàng

2 **Phân tích các thuộc tính của dữ liệu để Tiền Xử Lý dữ liệu**

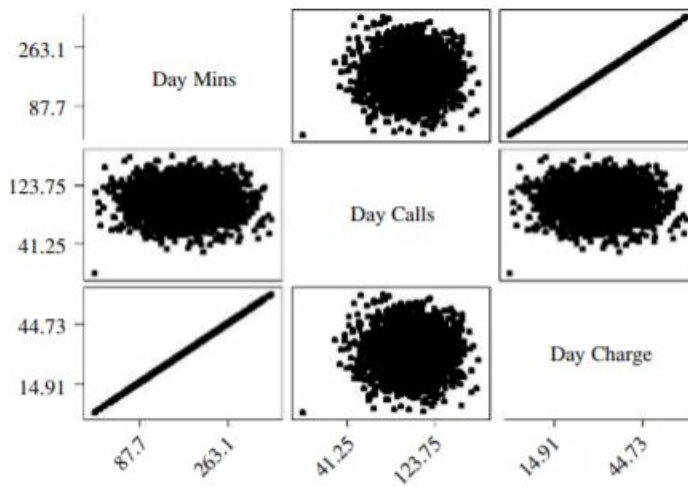


Figure 3.2 Matrix plot of day minutes, day calls, and day charge.

Đồ thị biểu thị mối quan hệ giữa 3 thuộc tính Day Mins , Day Calls , Day Charge
Dựa vào đồ thị ta thấy không có bất kì mối quan hệ nào giữa số phút trong ngày và cuộc gọi trong ngày hoặc giữa các cuộc gọi trong ngày với phí trong ngày

Mặt khác : Có một mối quan hệ tuyến tính giữa số phút trong ngày và phí trong ngày

$$\text{Phí} = 0,000613 + 0.17 * \text{Phút}$$

⇒ Vì phí trong ngày tương quan với số phút trong ngày ,chúng ta sẽ loại bỏ 1 trong 2 biến (ở đây em loại bỏ thuộc tính Day Charge). Các thuộc tính tương tự như tối , đêm , quốc tế cũng tương tự nên tiếp tục loại bỏ thuộc tính Eve Eve Charge', 'Night Charge', 'Intl Charge'. Chúng ta đã loại được 4 thuộc tính

Statistics of [15...]	
File	Edit
Generate	?
Collapse All	Expand All
Account Length	Statistics
Mean	101.065
Min	1.000
Max	243.000
Standard Deviation	39.822
Median	101.000
Voice Mail Messages	Statistics
Mean	8.099
Min	0.000
Max	51.000
Standard Deviation	13.688
Median	0.000
Day Minutes	Statistics
Mean	179.775
Min	0.000
Max	350.800
Standard Deviation	54.467
Median	179.400
Day Calls	Statistics
Mean	100.436
Min	0.000
Max	165.000
Standard Deviation	20.069
Median	101.000
Day Charge	Statistics
Mean	30.562
Min	0.000
Max	59.640
Standard Deviation	9.259
Median	30.500

Statistics of [15...]	
File	Edit
Generate	?
Collapse All	Expand All
Night Charge	Statistics
Mean	9.039
Min	1.040
Max	17.770
Standard Deviation	2.276
Median	9.050
International Minutes	Statistics
Mean	10.237
Min	0.000
Max	20.000
Standard Deviation	2.792
Median	10.300
International Calls	Statistics
Mean	4.479
Min	0.000
Max	20.000
Standard Deviation	2.461
Median	4.000
International Charge	Statistics
Mean	2.765
Min	0.000
Max	5.400
Standard Deviation	0.754
Median	2.780
Customer Service Calls	Statistics
Mean	1.563
Min	0.000
Max	9.000
Standard Deviation	1.315
Median	1.000

Các số liệu tóm tắt các thuộc tính Numeric như
Mean,Min,Max,SD,Median

Từ bảng ta có thể hình dung sơ lược dải phân bố của các thuộc tính Numeric

Vd:

- Thời gian tài khoản dài nhất là 234 và thấp nhất là 1, Thời gian trung bình của 3333 thuộc tính là 101

-Tin nhắn thư thoại trong ngày ít nhất là 0, nhiều nhất là 51. Đặc biệt có giá trị Median=0. Nghĩa là có ít nhất một nửa các khách hàng có thư thoại là 0

-Các thuộc tính như Day Minutes, Day Calls, Day Charge, Night Charge, Intl Minutes, Intl Calls, Intl Charge, Customer Service Calls có giá trị Mean khá gần với giá trị Median -> Phân bố các giá trị đều

Tiếp đến chúng ta sẽ phân tích về các thuộc tính có kiểu là Categorical là State và Area Code

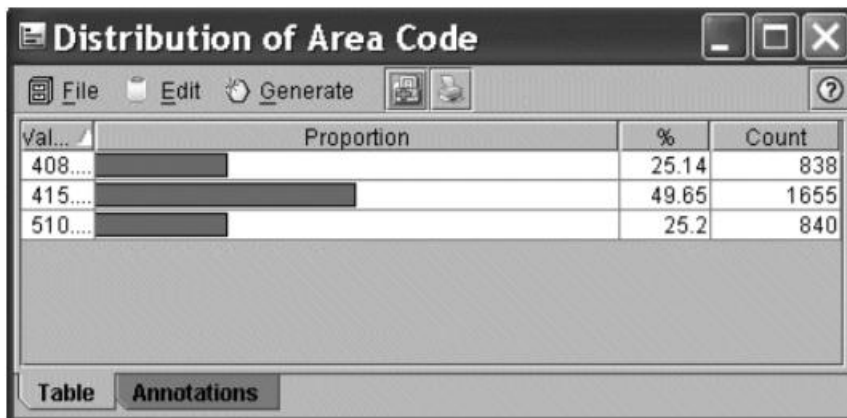


Figure 3.12 Only three area codes for all records.

Có kiểu dữ liệu là số nhưng em lại xếp nó vào Categorical là bởi vì mục đích chính của nó là dùng để phân loại các vùng địa lí. Điều kì lạ ở đây là bộ giá trị của Area Code chỉ gồm có 3 giá trị đó là (408 , 415 , 510) và theo thông tin miêu tả từ file mô tả tệp tin thì 3 giá trị này đều thuộc về California

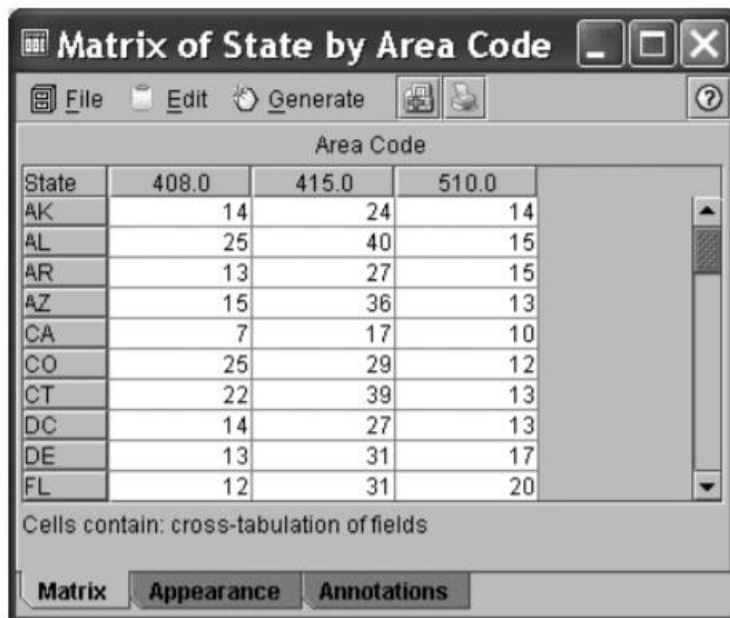


Figure 3.13 Anomaly: three area codes distributed across all 50 states.

Tuy nhiên theo hình ở phía dưới đây ba vùng được phân bố nhiều hơn hoặc ít hơn đồng đều trên tất cả các Tiểu Bang và Đặc khu Columbia, đây có thể là trường chỉ chứa các giá trị xấu. Do đó nếu chúng ta đi xa có thể không giúp làm đầu vào cho các mô hình tiếp theo -> Có thể gây ra lỗi

⇒ Kết Luận : Quyết định loại bỏ 2 thuộc tính State (Mô tả 50 tiểu bang) , Area Code vì đây là 2 thuộc tính không bình thường như đã phân tích ở trên

Các thuộc tính Numeric đã cho thì em thấy việc sử dụng phân cấp Binning là hợp lý nhất (also called banding)

+ Account Length

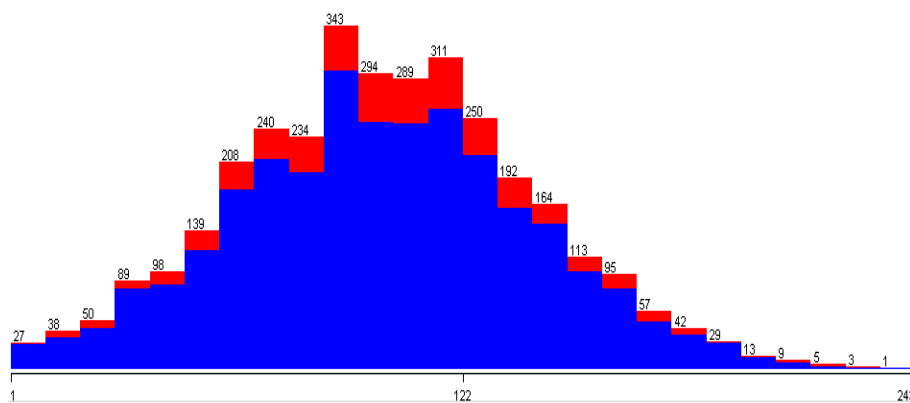
Statistic	Value
Minimum	1
Maximum	243
Mean	101.065
StdDev	39.822

Class: Churn? (Nom)

Visualize All

Nhìn vào sự phân bố thì chúng ta có thể thấy giá trị Min là 1, Max là 243 và Mean là 101,065. Đồ thị có sự phân bố khá đều nên em quyết định gom các thuộc tính này về 3 giá trị đó là :

- Short (0->100)
- Middle Length (100->100)
- Long (200->250)



Dữ liệu sau khi được Binning

Name: Account Length		Type: Nominal	
Missing: 1 (0%)		Unique: 0 (0%)	
Distinct: 3			
No.	Label	Count	Weight
1	Long	354	354.0
2	Middle_Length	296	296.0
3	Short	349	349.0

Class: Churn? (Nom)

Visualize All

Nhận xét : Dữ liệu khá đồng đều

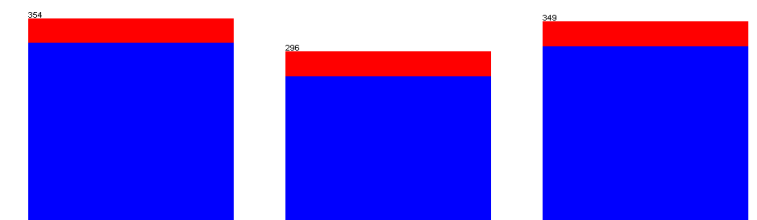
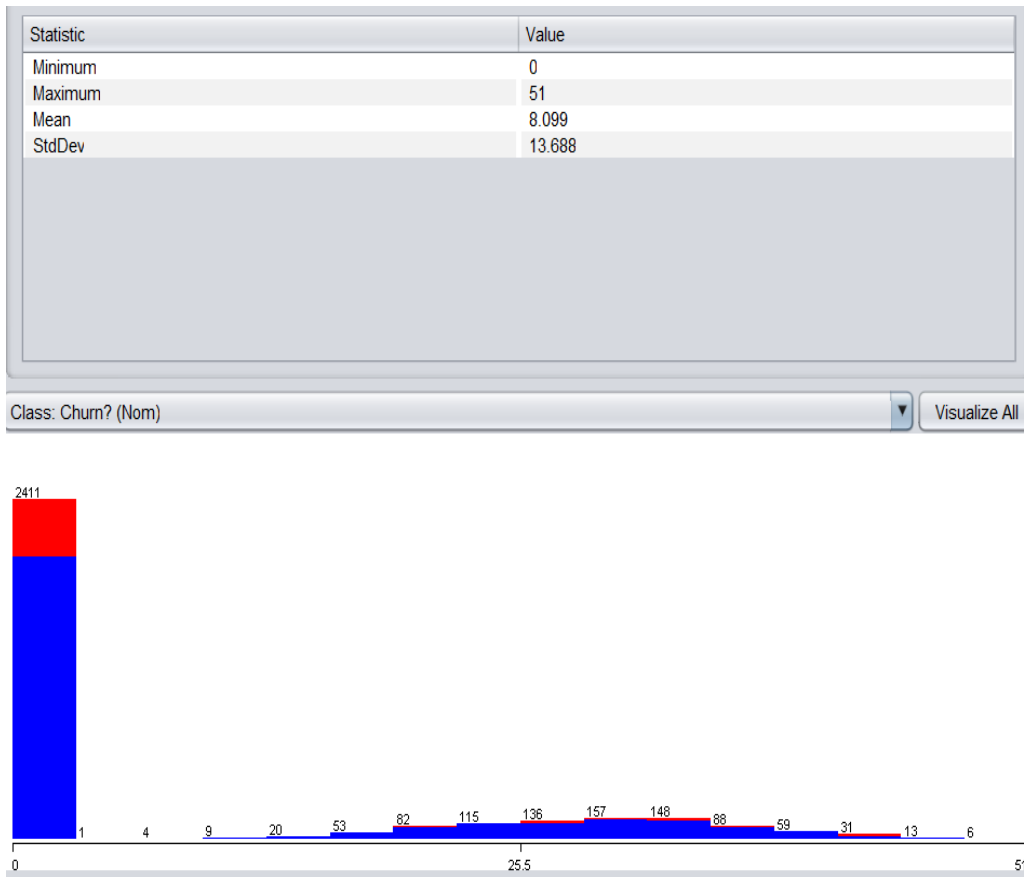


Figure Sau Khi Phân Cấp

+ VMail Message :



Nhìn vào sự phân bố ta có thể thấy giá trị 0 chiếm đa số. Từ đây có thể nhận xét rất ít khách hàng nhả tin. Giá trị Max cũng chỉ là 51. Đồ thị phân bố không đồng đều.

Quyết định gom đồ thị này về 3 giá trị
 + None (0->13)
 + Middle (13 -> 23)
 + Large (23->53)

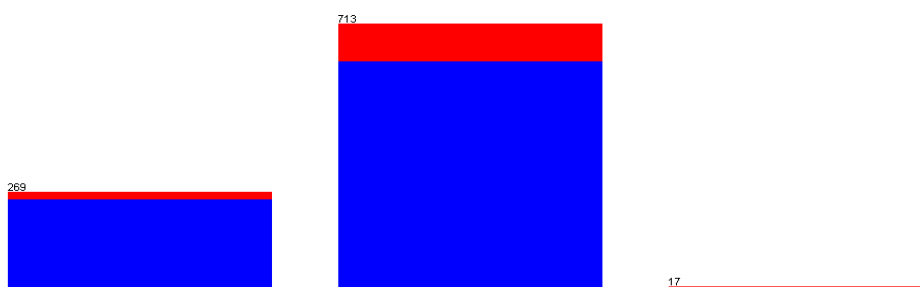
Dữ liệu thuộc tính VMail Message sau khi được Binning

No.	Label	Count	Weight
1	Middle(13->23)	269	269.0
2	None	713	713.0
3	Large(23->51)	17	17.0

Class: Churn? (Nom)

Visualize All

Nhận xét : Dữ liệu tin nhắn đa số nằm ở None & Middle



Các thuộc tính còn lại như Day Mins , Day Calls , Eve Mins , Eve Calls , Night Mins , Night Calls , Intl Mins , Intl Calls , CustSery Calls như đã phân tích ở trên là các kiểu dữ liệu phân bố khá bình thường và ổn định nên mỗi kiểu dữ liệu em sẽ dùng rang(min,max) và chia làm 4 phần -> 4 cấp độ khác nhau.

Ví dụ

Selected attribute	
Name: Day Mins Missing: 0 (0%)	Distinct: 1667 Type: Numeric Unique: 770 (23%)
Statistic	Value
Minimum	0
Maximum	350.8
Mean	179.775
StdDev	54.467

Min : 0
Max : 350
-> Chia làm 4 phần
0 -> 100
100 ->200
200->300
300->400

Trước khi Binning

Selected attribute			
Name: Day Mins Missing: 0 (0%)		Distinct: 4	Type: Nominal Unique: 0 (0%)
No.	Label	Count	Weight
1	200->300	364	364.0
2	100->200	549	549.0
3	300->400	17	17.0
4	0->100	70	70.0

Sau Khi Binning

II Code

Cú pháp để chạy code `python3 main.py churn.txt data.txt 2000`

Trong đó : main.py : Tên file python

churn.txt: File Input

data.txt : File Output

2000 : Có ý nghĩa là sẽ lấy tập dữ liệu con gồm 2000 dòng đầu tiên

Trong phần Code em sử dụng

+ Thư viện Pandas

- Pandas dùng để đọc , ghi dữ liệu
- Binning các thuộc tính

+ 3 Hàm

```
- def pre_process(df): Dùng để tiền xử lí ( Loại bỏ các thuộc tính không cần thiết )
- def hierarchies(df): Phân cấp các thuộc tính
- def main(fileinput,output,nrows): Đọc, ghi, xử lí dữ liệu
```

1 Hàm pre_process(df)

```
def pre_process(df):
    """
    :param dataset: Data Frame
    :return: Data Frame
    """
    Remove=['State', 'Phone', 'Area Code', 'Day Charge', 'Eve Charge', 'Night Charge', 'Intl Charge']
    df.drop(Remove,axis=1,inplace=True)
    return df
```

Hàm Pre_Process nhận tham số là 1 Data Frame và trả về một Data Frame

Hàm sẽ giúp chúng ta xóa các cột không có tác động vào Churn mà chúng ta đã phân tích ở trên thông qua Method drop.

2 Hàm hierarchies(df):


```
range = [0, 100, 200, 250]
grade_rank = pd.cut(df['Account Length'], range, right=False, labels=['Short', 'Middle_Length', 'Long'])
df['Account Length'] = grade_rank
```

Hàm hierarchies nhận tham số là 1 Data Frame và trả về một Data Frame

Hàm sẽ giúp chúng ta phân cấp (Binning) thông qua Method cut của thư viện Pandas

Chúng ta sẽ gán các giá trị trong mảng Range để lưu các đoạn giá trị và đặt các đoạn đó vào Labels

3 Hàm main(fileinput,output,nrows)

Hàm main sẽ nhận vào 3 tham số

fileinput : là file chúng ta chọn dùng để thực hiện.Ở đây là churn.txt

output : là file chúng ta chọn để xuất dữ liệu sau khi tiền xử lí và phân cấp

nrows : là số nguyên dương, có ý nghĩa chọn ra nrows hàng đầu tiên để xử lí (tập con của churn.txt)

III EXPERIMENT

1 Mục đích của việc phân tích dữ liệu

Dùng để làm sạch , chuyển đổi , mô hình hóa dữ liệu để khám phá thông tin hữu ích cho việc đưa ra quyết định kinh doanh . Trích xuất thông tin hữu ích từ dữ liệu và đưa ra quyết định dựa trên phân tích dữ liệu(cụ thể là quyết định CHURN là TRUE hay FALSE)

Ví dụ cụ thể : Nếu doanh nghiệp chậm phát triển hoặc thì bạn phải nhìn lại những sai sót , từ đó lập kế hoạch mà không lặp lại những sai lầm đó . Ngay cả khi doanh nghiệp của bạn đang phát triển , bạn sẽ mong muốn cho doanh nghiệp của bạn phát triển hơn nữa . Tất cả những gì bạn cần làm là phân tích dữ liệu kinh doanh và quy trình kinh doanh của bạn để đưa ra những chuyển lược mới . Việc phân tích kinh doanh giúp bạn

- +Dự đoán xu hướng và hành vi của khách hàng
- +Phân tích,giải thích và cung cấp dữ liệu có ý nghĩa
- +Tăng năng suất kinh doanh
- +Thúc đẩy quá trình ra quyết định hiệu quả

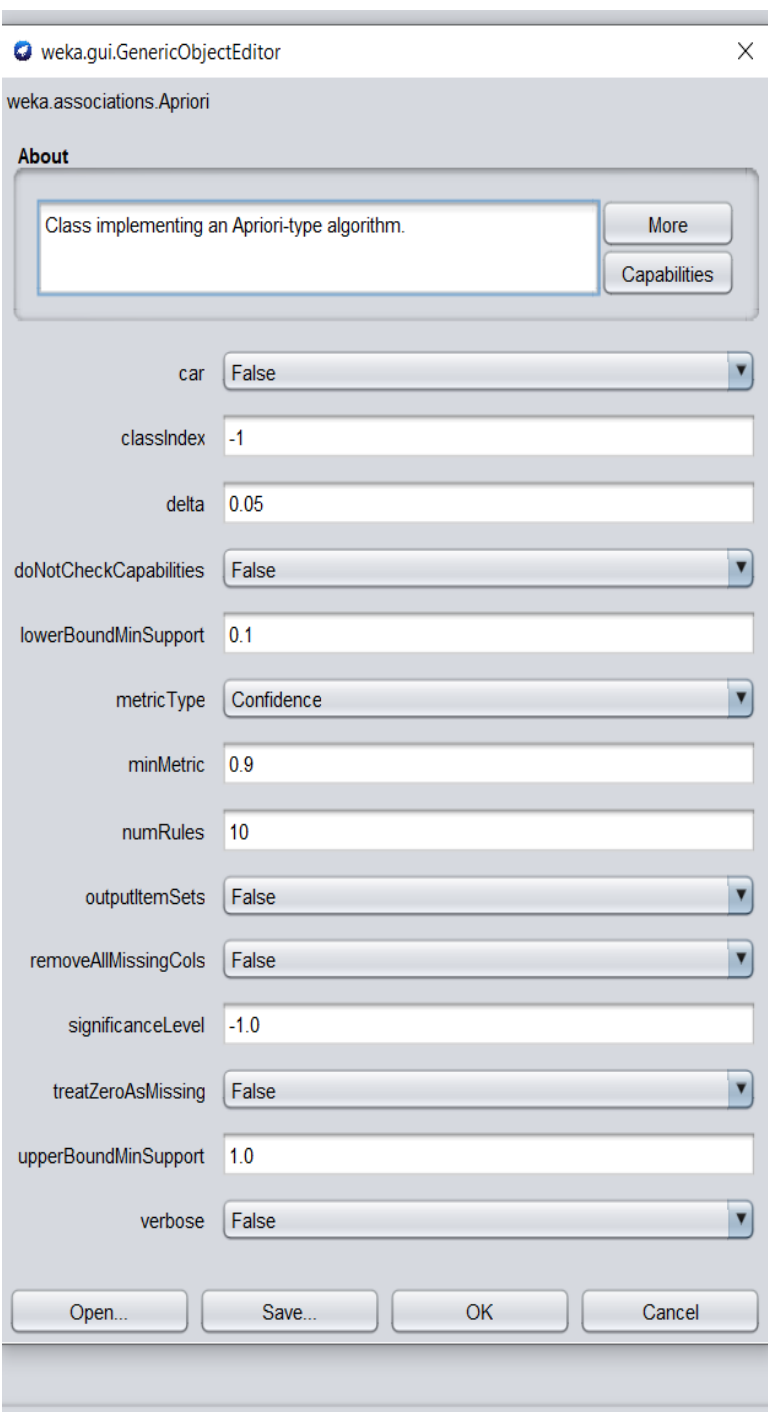
Tham khảo tại <https://blog.tomorrowmarketers.org/data-analysis-la-gi-va-ung-dung-trong-kinh-doanh-nhu-the-nao/>

2 Thử nghiệm với thuật toán Apiori

+Thể hiện dữ liệu :Tập dữ liệu có sẵn được thầy cung cấp churn.txt là một tập dữ liệu khá đầy đủ và chi tiết,không có các khoảng trống hoặc lỗi.

+Phương pháp tiền xử lí : Do tập dữ liệu khá là sạch nên em chỉ dùng các phương pháp xóa các dữ liệu dư thừa và phân cấp cho dữ liệu để thuận tiện cho việc chuẩn đoán sau này .

+Tham số của hệ thống , hệ số , độ đo ứng dụng



weka.gui.GenericObjectEditor

weka.associations.Apriori

About

Class implementing an Apriori-type algorithm.

More

Capabilities

car False

classIndex -1

delta 0.05

doNotCheckCapabilities False

lowerBoundMinSupport 0.1

metricType Confidence

minMetric 0.9

numRules 10

outputItemSets False

removeAllMissingCols False

significanceLevel -1.0

treatZeroAsMissing False

upperBoundMinSupport 1.0

verbose False

Open... Save... OK Cancel

Car : Nếu các quy tắc kết hợp lớp được bật(True) sẽ được khai thác thay vì các quy tắc kết hợp (chung).

classIndex: Chỉ mục của thuộc tính lớp. Nếu được đặt thành -1, thuộc tính cuối cùng được coi là thuộc tính lớp.

delta: Lặp lại giảm hỗ trợ bởi yếu tố này. Giảm hỗ trợ cho đến khi đạt được mức hỗ trợ tối thiểu hoặc số lượng quy tắc yêu cầu đã được tạo.

doNotCheckCapabilities: Lặp lại giảm hỗ trợ bởi yếu tố này. Giảm hỗ trợ cho đến khi đạt được mức hỗ trợ tối thiểu hoặc số lượng quy tắc yêu cầu đã được tạo.

lowerBoundMinSupport : Giới hạn support tối thiểu

metricType : Đặt loại chỉ số để xếp hạng các quy tắc. Độ tin cậy là tỷ lệ của các ví dụ được bao phủ bởi tiền đề cũng được bao phủ bởi hệ quả (Chỉ có thể khai thác các quy tắc kết hợp lớp bằng cách sử dụng độ tin cậy). Mức tăng là độ tin cậy chia cho tỷ lệ của tất cả các ví dụ được bao hàm bởi hệ quả. Đây là thước đo tầm quan trọng của hiệp hội độc lập với hỗ trợ. Đòn bẩy là tỷ lệ các ví dụ bổ sung được bao hàm bởi cả tiền đề và hệ quả trên những điều mong đợi nếu tiền đề và hệ quả độc lập với nhau. Tổng số ví dụ mà điều này đại diện được trình bày trong ngoặc đơn sau đòn bẩy. Niềm tin là một thước đo khác để đánh giá sự độc lập. Xác suất được đưa ra bởi P (tiền đề) P (! Hệ quả) / P (tiền đề,! Hệ quả).

minMetric : Mức tối thiểu mà luật được chấp nhận

numRules : Số luật sẽ tìm

outputItemSets : Xuất ra màn hình các ItemSets

removeAllMissingCols : Xóa tất cả các cột bị thiếu dữ liệu

significanceLevel : Mức độ đáng kể. Kiểm tra mức độ quan trọng (chỉ số confident).

treatZeroAsMissing : Nếu được bật, số không (nghĩa là giá trị đầu tiên của giá trị danh nghĩa) được xử lý giống như một giá trị bị thiếu

upperBoundMinSupport : Giới hạn trên cho mức hỗ trợ tối thiểu. Bắt đầu giảm dần hỗ trợ tối thiểu từ giá trị này

verbose : Nếu được bật, thuật toán sẽ được chạy ở chế độ tiết

+Kết quả sau khi chạy thuật toán Apriori với Weka

```
CustServ Calls
Churn?
=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.85 (850 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Best rules found:

1. CustServ Calls=0->5 971 ==> Intl Mins=0->100 971    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. Int'l Plan=no 900 ==> Intl Mins=0->100 900    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. Int'l Plan=no CustServ Calls=0->5 874 ==> Intl Mins=0->100 874    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. Churn?=False. 871 ==> Intl Mins=0->100 871    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. CustServ Calls=0->5 Churn?=False. 858 ==> Intl Mins=0->100 858    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. Churn?=False. 871 ==> CustServ Calls=0->5 858    <conf:(0.99)> lift:(1.01) lev:(0.01) [12] conv:(1.8)
7. Intl Mins=0->100 Churn?=False. 871 ==> CustServ Calls=0->5 858    <conf:(0.99)> lift:(1.01) lev:(0.01) [12] conv:(1.8)
8. Churn?=False. 871 ==> Intl Mins=0->100 CustServ Calls=0->5 858    <conf:(0.99)> lift:(1.01) lev:(0.01) [12] conv:(1.8)
9. Int'l Plan=no 900 ==> CustServ Calls=0->5 874    <conf:(0.97)> lift:(1) lev:(0) [0] conv:(0.97)
10. Int'l Plan=no Intl Mins=0->100 900 ==> CustServ Calls=0->5 874    <conf:(0.97)> lift:(1) lev:(0) [0] conv:(0.97)
```

+ Sau khi chạy chương trình ta thu được

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Kết quả thu được các tập phổ biến

+ Ngoài ra còn thu được 10 tập luật phổ biến ở dưới

Best rules found:

```
1. Int'l Plan=no Day Mins=100->200 CustServ Calls=0->5 486 ==> Churn?=False. 463    conf:(0.95)
2. Int'l Plan=no Day Mins=100->200 Intl Mins=0->100 CustServ Calls=0->5 486 ==> Churn?=False. 463    conf:(0.95)
3. Int'l Plan=no Day Mins=100->200 502 ==> Churn?=False. 471    conf:(0.94)
4. Int'l Plan=no Day Mins=100->200 Intl Mins=0->100 502 ==> Churn?=False. 471    conf:(0.94)
5. Day Mins=100->200 CustServ Calls=0->5 531 ==> Churn?=False. 492    conf:(0.93)
6. Day Mins=100->200 Intl Mins=0->100 CustServ Calls=0->5 531 ==> Churn?=False. 492    conf:(0.93)
7. Int'l Plan=no CustServ Calls=0->5 874 ==> Churn?=False. 799    conf:(0.91)
8. Int'l Plan=no Intl Mins=0->100 CustServ Calls=0->5 874 ==> Churn?=False. 799    conf:(0.91)
9. Day Mins=100->200 549 ==> Churn?=False. 500    conf:(0.91)
10. Day Mins=100->200 Intl Mins=0->100 549 ==> Churn?=False. 500    conf:(0.91)
```

10 Tập Luật Phổ Biến

III Tóm tắt kết quả

- + Kết quả được đánh giá dựa trên các tập luật phổ biến thu được
- + Tập luật tốt nhất mà em thu được đó là

```
Int'l Plan=no Day Mins=100->200 CustServ Calls=0->5 486 ==> Churn?=False. 463    conf:(0.95)
```

Tập này có ý nghĩa là Nếu kế hoạch quốc tế là No , có Day Mins thuộc khoảng từ 100 đến 200 và CustServ Calls thuộc từ khoảng 0 -> 5 sẽ kết luận được thuộc tính Churn là False với confident là 0.95 và Mini Support bằng 0.85

- + Điểm mạnh và yếu của em trong bài tập này
 - Điểm mạnh : Nắm bắt được cách hoạt động của thuật toán Apiori
 - Điểm yếu : Đọc tài liệu tiếng anh chưa thành thạo dẫn đến khó khăn trong việc xử lí các thuộc tính của tập dữ liệu Churn.txt