

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN



LAB 02:
FREQUENT ITEMSETS
AND
ASSOCIATION RULES

Data Mining - Term I/2020-2021

MỤC LỤC

A. PHÂN CÔNG CÔNG VIỆC	3
B. NỘI DUNG	3
I. DATA:	3
1. Mô tả tập dữ liệu: Churn.txt	3
2. Khái niệm phân cấp:	4
II. CODE	4
1. Mô tả chức năng file:	4
2. Mô tả cú pháp (Command line argument):	4
III. EXPERIMENTS	5
1. Mục đích của việc phân tích dữ liệu:	5
2. Tiến hành thí nghiệm:	6
3. Phân tích kết quả thí nghiệm:	22
4. Phân tích biểu đồ:	30
IV. SUMMARY:	33

A. PHÂN CÔNG CÔNG VIỆC

MSSV	Họ và tên	Công việc	Tỉ lệ
18120027	Nguyễn Thị Thu Hằng	Code chuẩn hoá dữ liệu, phân cấp dữ liệu Phân tích tập dữ liệu gốc và các trường hợp khác	60%
18120178	Phạm Thị Hoài Hiền	Mô tả dữ liệu Phân tích tập dữ liệu đã phân loại Tóm tắt kết quả	40%

B. NỘI DUNG

I. DATA:

1. Mô tả tập dữ liệu: Churn.txt

Churn, còn được gọi là *tiêu hao* (attrition), là một thuật ngữ được sử dụng để chỉ một khách hàng rời bỏ dịch vụ của một công ty để chuyển sang một công ty khác.

“Churn.txt” là bộ dữ liệu bao gồm 21 thuộc tính. Trong đó 20 thuộc tính kèm theo thông tin của 3333 khách hàng và một thuộc tính “Churn?” chỉ báo về việc khách hàng đó có rời bỏ công ty hay không.

STT	Tên thuộc tính	Kiểu giá trị	Mô tả
1	State	Nominal	Tên viết tắt của 50 bang Hoa Kỳ và Quận Columbia (District of Columbia)
2	Account length	Numeric	Thời gian tài khoản hoạt động
3	Area code	Nominal	Mã vùng
4	Phone number	Nominal	Số điện thoại
5	Int'l Plan	Nominal {yes,no}	Khách hàng có sử dụng gói cước quốc tế hay không?
6	VoiceMail Plan	Nominal {yes,no}	Khách hàng có sử dụng gói cước thư thoại hay không?
7	VMail Messages	Numeric	Số lượng thư thoại
8	Day Mins	Numeric	Số phút thuê bao sử dụng vào ban ngày
9	Day Calls	Numeric	Tổng số cuộc gọi vào ban ngày
10	Day Charge	Numeric	Tổng chi phí vào ban ngày

11	Eve Mins	Numeric	Số phút thuê bao sử dụng vào buổi tối
12	Eve Calls	Numeric	Tổng số cuộc gọi vào buổi tối
13	Eve Charge	Numeric	Tổng chi phí vào buổi tối
14	Night Mins	Numeric	Số phút thuê bao sử dụng vào ban đêm
15	Night Calls	Numeric	Tổng số cuộc gọi vào ban đêm
16	Night Charge	Numeric	Tổng chi phí vào ban đêm
17	Intl Mins	Numeric	Số phút thuê bao sử dụng gọi quốc tế
18	Intl Calls	Numeric	Tổng số cuộc gọi quốc tế
19	Intl Charge	Numeric	Tổng chi phí gọi quốc tế
20	CustServ Calls	Numeric	Số lượng cuộc gọi đến dịch vụ khách hàng
21	Churn	Nominal { True, False }	Thuê bao có rời bỏ công ty hay không? (true: có; false: không)

Một số thuộc tính có thể không được xét tới vì nó không ảnh hưởng đến hành vi của người dùng, ví dụ như State, Area Code, Phone (vùng, mã vùng và số điện thoại đều không ảnh hưởng đến hành vi người dùng, không có tính quy luật với các hành vi khác)

2. Khái niệm phân cấp:

Áp dụng Set grouping hierarchy để phân loại các thuộc tính số thành các nhóm:

{low,high}

{low,medium,high}

{very_low, low, medium, high, very_high}

Ta sử dụng công thức $(\max - \min + 1)/N$ (với N là số khoảng cần chia) để tìm khoảng ngăn cách giữa các giỏ (bin), sau đó sắp xếp các giá trị trong mảng vào các giỏ đã chia.

II. CODE

1. Mô tả chức năng file:

+ **Binning.py**: Chia dữ liệu theo khoảng. Gồm các khoảng mặc định {low, high}, {low, medium, high}, {very_low, low, medium, high, very_high}

+ **Normalize.py**: Chuẩn hóa dữ liệu theo kiểu min và z-score

2. Mô tả cú pháp (Command line argument):

```
Argument syntax:
python Binning.py input.csv output.csv -a "FirstAttribute" -a "SecondAttribute" -b /*The Number of Bin*/
Example: python Binning.py churn.csv output.csv -a all -b 3
```

Argument syntax of Binning.py

```
Argument syntax:
python Normalize.py input.csv output.csv -m "Method" -a "FirstAttribute" -a "SecondAttribute"
Example: python Normalize.py churn.csv output.csv -m minmax -a all
```

Argument syntax of Normalize.py

(*) Lưu ý:

+ Đối với thông số -b đại diện cho số khoảng cần chia: ta chỉ chọn {2,3,5} tương ứng với các mức độ phân cấp mong muốn tương ứng:

2 - {low,high}

3 - {low,medium,high}

5 - {very_low, low, medium, high, very_high}

```
parser.add_argument("-b", "--NumOfBin", type=int, choices=[2, 3, 5], help="Number of bins for binning")
```

+ Đối với thông số -m đại diện cho phương thức chuẩn hóa mong muốn ta chỉ chọn 2 phương thức minmax và zscore

```
parser.add_argument("-m", "--Method", choices=["minmax", "zscore"], help="Method of Normalized")
```

+ Đối với thông số -a thì tương ứng với mỗi thuộc tính ta sẽ gọi như sau: **-a “tên thuộc tính”** (nên cho vào ngoặc kép vì có nhiều thuộc tính có khoảng trắng”. Lưu ý nếu muốn gọi tất cả các thuộc tính ta dùng **-a all**

III. EXPERIMENTS

1. Mục đích của việc phân tích dữ liệu:

a) Để hiểu hơn về dữ liệu, nội dung cần biết thêm về dữ liệu

Nội dung cần tìm hiểu	Kết quả tìm hiểu
Tập dữ liệu thuộc lĩnh vực nào?	Viễn thông
Tập dữ liệu bao gồm các thuộc tính như nào? Và ý nghĩa của các giá trị đó tương ứng với thuộc tính đó ra sao?	(Bảng dữ liệu ở mục I.1)
Thuộc tính chính trong tập dữ liệu?	Churn?
Sự tương quan và mối liên hệ của các thuộc tính với nhau? Và với thuộc tính chính?	(Thể hiện qua thí nghiệm)
Các thuộc tính cần lược bỏ để giảm chiều dữ liệu	(Thể hiện qua thí nghiệm)

Các thuộc tính nào có liên quan trực tiếp đến giá trị của thuộc tính mầu chốt? Trong trường hợp này nó thể hiện hành vi của khách hàng như thế nào thì có khả năng khách hàng rời bỏ công ty?	(Thể hiện qua thí nghiệm)
---	---------------------------

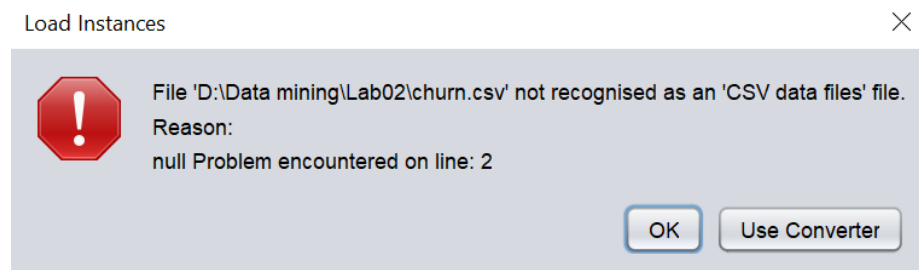
b) Phục vụ cho việc ra quyết định? Vậy bạn cần dựa vào dữ liệu để quyết định điều gì?

- + Cho biết các nguyên nhân dẫn đến việc khách hàng rời bỏ công ty
- + Nâng cao và phát huy thuộc tính nào để giảm bớt số lượng khách hàng rời bỏ công ty?
- + Cải thiện hoặc lược bỏ hoặc tăng giảm chi phí hợp lý các dịch vụ nào để giảm bớt số lượng khách hàng rời bỏ công ty?

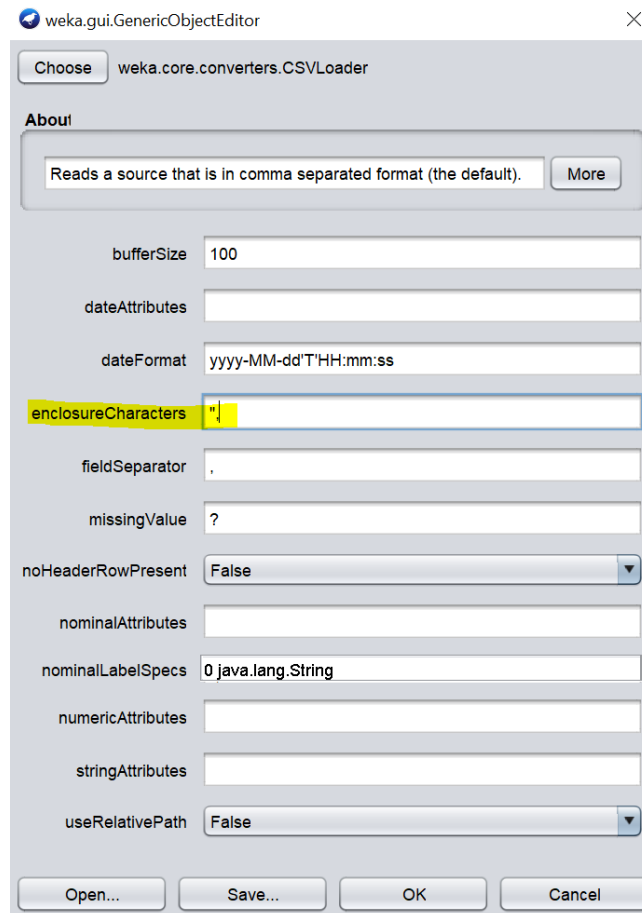
2. Tiến hành thí nghiệm:

a. Mở tập dữ liệu Churn.txt trong Weka:

Ta đổi định dạng file “Churn.txt” thành “Churn.csv” để có thể mở được trong Weka. Tuy nhiên sẽ có lỗi trong lúc dùng Weka để mở file như sau:



Bấm vào nút “Use converter”, sau đó sửa mục *enclosureCharacters* như hình dưới rồi bấm OK



b. Chạy thử nghiệm lần 1:

- ***Tiền xử lí:***

Cần phải thay đổi một số mục như hình để có thể đọc được file churn.txt và chạy thử nghiệm với Apriori. Vì khi sử dụng Apriori thì các thuộc tính phải thuộc Nominal nên phải thêm “first-last” để biến đổi Numeric thành Nominal

weka.gui.GenericObjectEditor

Choose weka.core.converters.CSVLoader

About

Reads a source that is in comma separated format (the default). More

bufferSize 100

dateAttributes

dateFormat yyyy-MM-dd'T'HH:mm:ss

enclosureCharacters ",

fieldSeparator ,

missingValue ?

noHeaderRowPresent False

nominalAttributes first-last

nominalLabelSpecs 0 java.lang.String

numericAttributes

stringAttributes

useRelativePath False

Open... Save... OK Cancel

Chọn tab Association Rules để tiến hành phân tích dữ liệu

Preprocess Classify Cluster Associate Select attributes Visualize

Associator

Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

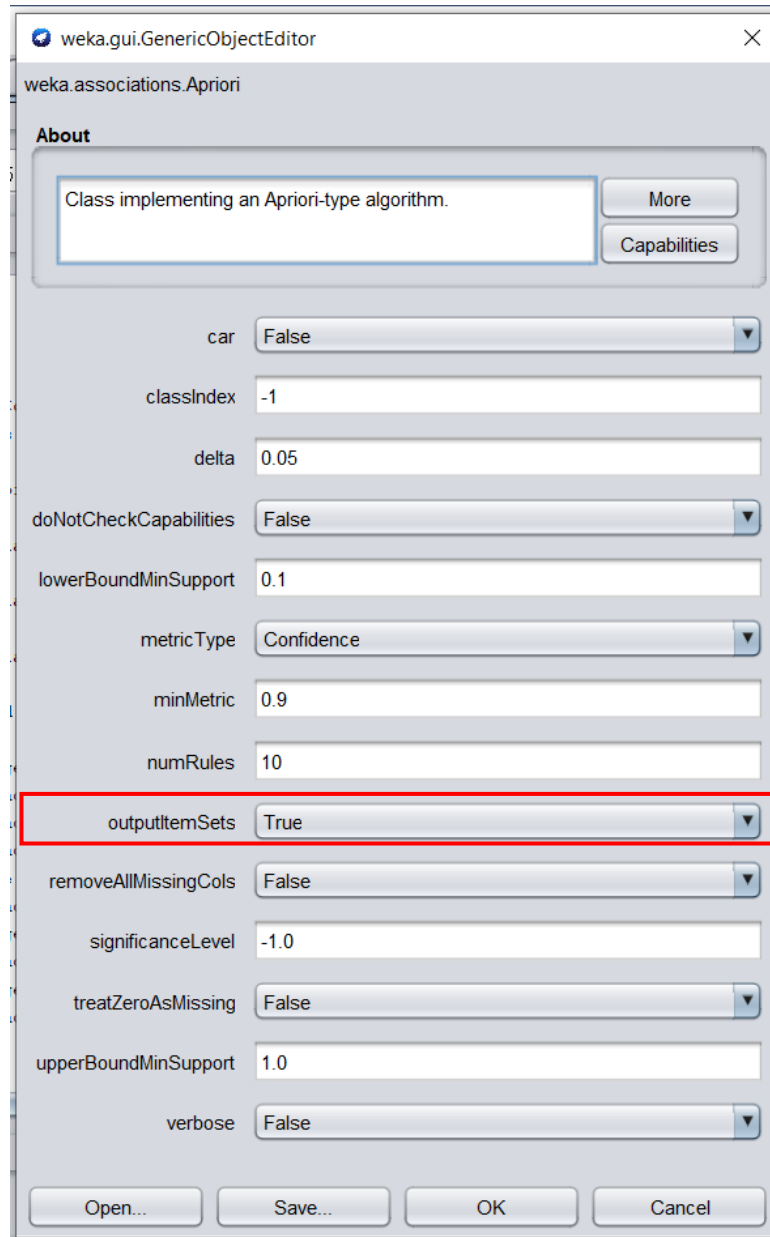
Start Stop

Associator output

Result list (right-click for options)

- **Chạy thử nghiệm:**

Lần 1 ta chạy thử nghiệm tập dữ liệu churn.txt với các thông số mặc định mà Weka cung cấp, tuy nhiên để quan sát ra hơn các Large Itemset ta chọn mode True cho outputItemSets. Chọn Ok, sau đó nhấn chọn Start



Kết quả:

+ Có 2166 mẫu đạt minsup = 0.65

Minimum support: 0.65 (2166 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 7

+ Các Large Itemsets

Generated sets of large itemsets:

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

Large Itemsets L(1):
Int'l Plan=no 3010
VMail Plan=no 2411
VMail Message=0 2411
Churn?=False. 2850
```

```
Size of set of large itemsets L(2): 4

Large Itemsets L(2):
Int'l Plan=no VMail Plan=no 2180
Int'l Plan=no VMail Message=0 2180
Int'l Plan=no Churn?=False. 2664
VMail Plan=no VMail Message=0 2411
```

```
Size of set of large itemsets L(3): 1

Large Itemsets L(3):
Int'l Plan=no VMail Plan=no VMail Message=0 2180
```

- Trong Large itemsets L(2) ta thấy rằng VMail Plan=no VMail Message=0 2411 (điều này đương nhiên vì nếu người dùng không sử dụng gói thư thoại thì sẽ không có số lượng thư thoại), nên ta không cần chú ý các large itemset này về sau
- Int'l Plan=no VMail Plan=no 2180 => 2180 mẫu thể hiện người không dùng gói Quốc tế cũng không dùng thư thoại
- Int'l Plan=no Churn?=False. 2664 => 2664 mẫu thể hiện số lượng người không dùng gói quốc tế thì không rời bỏ công ty

+ Các luật bao gồm:

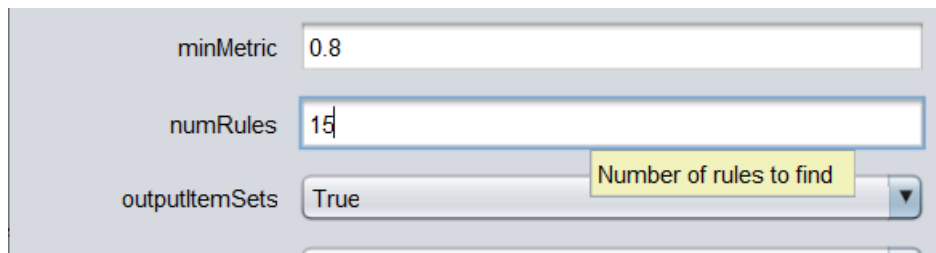
Best rules found:

```
1. VMail Message=0 2411 ==> VMail Plan=no 2411    <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)
2. VMail Plan=no 2411 ==> VMail Message=0 2411    <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)
3. Int'l Plan=no VMail Message=0 2180 ==> VMail Plan=no 2180    <conf:(1)> lift:(1.38) lev:(0.18) [603] conv:(603.05)
4. Int'l Plan=no VMail Plan=no 2180 ==> VMail Message=0 2180    <conf:(1)> lift:(1.38) lev:(0.18) [603] conv:(603.05)
5. Churn?=False. 2850 ==> Int'l Plan=no 2664    <conf:(0.93)> lift:(1.04) lev:(0.03) [90] conv:(1.48)
6. VMail Plan=no 2411 ==> Int'l Plan=no 2180    <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)
7. VMail Message=0 2411 ==> Int'l Plan=no 2180    <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)
8. VMail Plan=no VMail Message=0 2411 ==> Int'l Plan=no 2180    <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)
9. VMail Message=0 2411 ==> Int'l Plan=no VMail Plan=no 2180    <conf:(0.9)> lift:(1.38) lev:(0.18) [603] conv:(3.6)
10. VMail Plan=no 2411 ==> Int'l Plan=no VMail Message=0 2180    <conf:(0.9)> lift:(1.38) lev:(0.18) [603] conv:(3.6)
```

Ta chú ý đến các luật có độ tin cậy ($\langle \text{conf}:(x) \rangle$) lớn và có liên quan đến thuộc tính mẫu chốt Churn? Ta dễ dàng nhận thấy luật thứ 5 có độ tin cậy 93%, nhưng ta không tìm được một luật ở chiều ngược lại. *Tuy nhiên ta có thể kết thấy rõ ràng một điều khách hàng không có xu hướng rời bỏ công ty thì không sử dụng gói Quốc tế và điều này mang độ tin cậy lớn 93% nên ta có thể tạm suy đoán rằng những khách hàng sử dụng gói Quốc tế có xu hướng rời bỏ công ty cao hơn.*

c. Chạy thử nghiệm lần 2:

- Lần 2 ta giảm minMetric (độ tin cậy) còn 80% và tăng số luật cần tìm lên 15. Chọn OK, sau đó nhấn chọn Start



minMetric 0.8

numRules 15

outputItemSets True

Number of rules to find

- Kết quả: Có 2000 mẫu đạt minsup = 0.6

```
Minimum support: 0.6 (2000 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 8
```

- Các large itemsets

Generated sets of large itemsets:

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

Large Itemsets L(1):

Int'l Plan=no 3010

VMail Plan=no 2411

VMail Message=0 2411

Churn?=False. 2850

Size of set of large itemsets L(2): 6

Large Itemsets L(2):

Int'l Plan=no VMail Plan=no 2180

Int'l Plan=no VMail Message=0 2180

Int'l Plan=no Churn?=False. 2664

VMail Plan=no VMail Message=0 2411

VMail Plan=no Churn?=False. 2008

VMail Message=0 Churn?=False. 2008

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

Large Itemsets L(3):

Int'l Plan=no VMail Plan=no VMail Message=0 2180

VMail Plan=no VMail Message=0 Churn?=False. 2008

- Int'l_Plan=no Churn?=False. 2664 xuất hiện lại => đây là một large itemsets ổn định, không phụ thuộc vào độ tin cậy
- VMail_Plan=no Churn?=False. 2008 mới xuất hiện => Có 2008 mẫu thể hiện khách hàng không sử dụng dịch vụ thư thoại thì không rời bỏ công ty.

Các luật:

Best rules found:

```
1. VMail Message=0 2411 ==> VMail Plan=no 2411    <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)
2. VMail Plan=no 2411 ==> VMail Message=0 2411    <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)
3. Int'l Plan=no VMail Message=0 2180 ==> VMail Plan=no 2180    <conf:(1)> lift:(1.38) lev:(0.18) [603] conv:(603.05)
4. Int'l Plan=no VMail Plan=no 2180 ==> VMail Message=0 2180    <conf:(1)> lift:(1.38) lev:(0.18) [603] conv:(603.05)
5. VMail Message=0 Churn?=False. 2008 ==> VMail Plan=no 2008    <conf:(1)> lift:(1.38) lev:(0.17) [555] conv:(555.47)
6. VMail Plan=no Churn?=False. 2008 ==> VMail Message=0 2008    <conf:(1)> lift:(1.38) lev:(0.17) [555] conv:(555.47)
7. Churn?=False. 2850 ==> Int'l Plan=no 2664    <conf:(0.93)> lift:(1.04) lev:(0.03) [90] conv:(1.48)
8. VMail Plan=no 2411 ==> Int'l Plan=no 2180    <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)
9. VMail Message=0 2411 ==> Int'l Plan=no 2180    <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)
10. VMail Plan=no VMail Message=0 2411 ==> Int'l Plan=no 2180    <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)
11. VMail Message=0 2411 ==> Int'l Plan=no VMail Plan=no 2180    <conf:(0.9)> lift:(1.38) lev:(0.18) [603] conv:(3.6)
12. VMail Plan=no 2411 ==> Int'l Plan=no VMail Message=0 2180    <conf:(0.9)> lift:(1.38) lev:(0.18) [603] conv:(3.6)
13. Int'l Plan=no 3010 ==> Churn?=False. 2664    <conf:(0.89)> lift:(1.04) lev:(0.03) [90] conv:(1.26)
14. VMail Plan=no 2411 ==> Churn?=False. 2008    <conf:(0.83)> lift:(0.97) lev:(-0.02) [-53] conv:(0.86)
15. VMail Message=0 2411 ==> Churn?=False. 2008    <conf:(0.83)> lift:(0.97) lev:(-0.02) [-53] conv:(0.86)
```

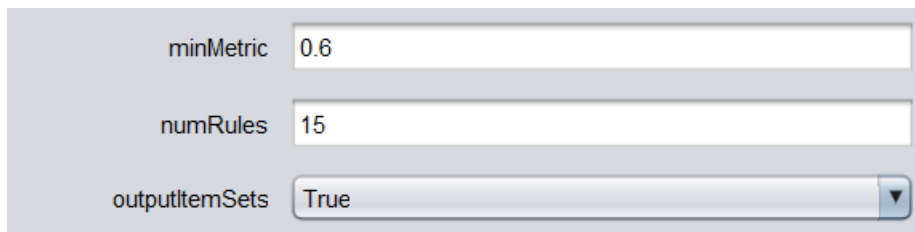
Ta chú ý đến các luật có độ tin cậy ($\langle \text{conf}(\text{x}) \rangle$) lớn và có liên quan đến thuộc tính mẫu chốt Churn?

- Ta dễ dàng thấy được luật số 7 đã xuất hiện ở lần thí nghiệm trước
- Luật $\text{Int'l_Plan=no } 3010 \implies \text{Churn?}=False. 2664 \langle \text{conf}:(0.89) \rangle \text{lift}:(1.04) \text{lev}:(0.03) [90] \text{conv}:(1.26)$. Lần xuất hiện luật theo chiều ngược lại với độ tin cậy 89% mang ý nghĩa những khách hàng không sử dụng gói quốc tế thì khả năng không rời bỏ công ty cao hơn.
- Luật $\text{VMail Plan=no } 2411 \implies \text{Churn?}=False. 2008 \langle \text{conf}:(0.83) \rangle \text{lift}:(0.97) \text{lev}:(-0.02) [-53] \text{conv}:(0.86)$. Xuất hiện luật mới mang ý nghĩa khách hàng không sử dụng dịch vụ hộp thư thoại thì khả năng không rời bỏ khách hàng cao hơn, với độ tin cậy của luật là 83%
- Luật $\text{VMail Message=0 } 2411 \implies \text{Churn?}=False. 2008 \langle \text{conf}:(0.83) \rangle \text{lift}:(0.97) \text{lev}:(-0.02) [-53] \text{conv}:(0.86)$. 2 thuộc tính VMail Plan và VMail Message này được xem như tương đồng như đã giải thích ở thử nghiệm 1.

Từ các luật trên ta còn thấy được các thuê bao không rời bỏ công ty thường là thuê bao không sử dụng dịch vụ thư thoại. Ta có thể tạm suy đoán rằng những khách hàng sử dụng dịch vụ thư thoại thì sẽ rời bỏ công ty cao hơn.

d. Chạy thử nghiệm lần 3:

- Lần 3 ta giảm minMetric (độ tin cậy) còn 60% và tăng số luật cần tìm lên 15. Chọn OK, sau đó nhấn chọn Start



minMetric	0.6
numRules	15
outputItemSets	True

- Kết quả : Không có gì thay đổi so với lần 2

(*) Tổng kết cho 3 lần chạy thử nghiệm đầu tiên:

- Qua 3 lần thử nghiệm ta thấy có một mối quan hệ khá chặt chẽ giữa việc khách hàng có sử dụng gói Quốc tế và gói Thư thoại đối với việc khách hàng có rời bỏ công ty hay không. Cụ thể các luật được dự đoán là:

+ *Những khách hàng sử dụng gói Quốc tế có xu hướng rời bỏ công ty*

+ *Khách hàng không sử dụng dịch vụ thư thoại thì không rời bỏ công ty*

- Tuy nhiên ta nhận thấy được trong số khách hàng sử dụng gói Quốc tế (cụ thể là $3333 - 2664 = 669$ mẫu) chưa được khai thác vấn đề liệu họ có rời bỏ công ty hay không. Tương tự đối với $3333 - 2411 = 922$ mẫu đối với trường hợp khách hàng có sử dụng gói Thư thoại. *(Cần có giải pháp xem xét – Xem các mục sau)*

- Quan sát các thuộc tính được sử dụng trong 3 lần thử nghiệm đầu tiên Weka chỉ xác định được các large itemsets ở các thuộc tính như: Int'l_Plan, VMail_Plan, VMail_Message và Churn? chứ không xét đến các thuộc tính khác như Day_Mins, Day_Calls, Eve_Mins, Eve_Calls,... Lý giải cho điều này: các thuộc tính không được xét đến có kiểu dữ liệu gốc là Numeric (số nguyên hoặc số thực có tính liên tục), nhưng vì ta đã đổi thành Nominal để chạy được thuật toán Apriori nên sự phân bố của chúng trở nên quá rời rạc, không đủ mạnh để hình thành các large itemsets *(Cần có giải pháp xem xét – Xem các mục sau)*

() Phân tích trên tập dữ liệu đã qua phân loại*

() Tiền xử lý dữ liệu bằng cách chia khoảng các thuộc tính có giá trị Numeric sao cho sự phân bố của chúng không quá rời rạc. Tạo điều kiện tham gia hình thành Itemsets. Cụ thể sử dụng khái niệm phân cấp được đề cập ở mục I.2 và code Binning.py với hướng dẫn ở II.1 mục đích phân các thuộc tính Numeric thành Nominal có thể áp dụng Apriori*

() Với phân loại {low, high}*

e. Chạy thử nghiệm lần 4:

- **Tiền xử lý:**

- Ta thực hiện với dữ liệu đã Binning theo 2 cấp độ {low, high}
- Đầu tiên là chạy file python để Binning Data với cú pháp

Ex: *python Binning.py churn.csv churn2.csv -a all -b 2*

- Ta sẽ có Churn2.csv như sau:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	State	Account Li	Area Code	Phone	Int'l Plan	VMail Plan	VMail Mes	Day Mins	Day Calls	Day Charg	Eve Mins	Eve Calls	Eve Charg	Night Min	Night Calls	Night Char	Int'l Mins	Int'l Calls	Int'l Charg
1	KS	High	Low	382-4657	no	yes	High	High	High	High	High	High	High	High	Low	High	High	Low	Hi
2	OH	Low	Low	371-7191	no	yes	High	Low	High	Low	High	High	High	High	Low	High	High	Low	Hi
3	NJ	High	Low	358-1921	no	no	Low	High	High	High	Low	High	Low	Low	High	Low	High	Low	Hi
4	OH	Low	Low	375-9999	yes	no	Low	High	Low	High	Low	High	Low	Low	Low	Low	Low	Low	Lo
5	OK	Low	Low	330-6626	yes	no	Low	Low	High	Low	Low	High	Low	Low	High	Low	High	Low	Hi
6	AL	Low		391-8027	yes	no	Low	High	High	High	High	High	High	Low	High	High	Low	Low	Lo
7	MA	Low		355-9993	no	yes	Low	High	High	High	High	High	High	High	High	High	Low	Low	Lo
8	MO	High	Low	329-9001	yes	no	Low	Low	Low	Low	Low	High	Low	High	Low	High	Low	Low	Lo
9	LA	Low	Low	335-4719	no	no	Low	High	High	High	High	Low	High	High	Low	High	Low	Low	Hi
10	WV	High	Low	330-8173	yes	yes	High	High	High	High	High	High	High	High	Low	High	High	Low	Hi
11	IN	Low	Low	329-6603	no	no	Low	Low	High	Low	High	Low	High	High	High	High	High	Low	Hi
12	RI	Low	Low	344-9403	no	no	Low	High	High	High	Low	High	Low	Low	Low	Low	Low	Low	Hi
13	IA	High	Low	363-1107	no	no	Low	Low	High	Low	Low	Low	Low	Low	High	Low	High	Low	Hi
14	MT	Low		394-8006	no	no	Low	Low	High	Low	High	Low	High	Low	High	Low	High	Low	Hi
15	IA	Low	Low	366-9238	no	no	Low	Low	Low	Low	High	Low	High	Low	Low	High	High	Low	Hi
16	NY	High	Low	351-7269	no	no	Low	High	Low	High	High	High	High	Low	High	Low	Low	Low	Lo
17	ID	Low	Low	350-8884	no	yes	High	High	High	High	High	High	High	Low	Low	Low	High	Low	Hi
18	VT	Low		386-2923	no	no	Low	High	High	High	High	High	High	Low	High	Low	Low	Low	Lo
19	VA	Low		256-3003	no	yes	High	High	High	High	High	High	High	Low	High	Low	High	Low	Lo

- Đưa file này vào Weka để tiến hành phân tích.
- **Chạy thử nghiệm:**
- Ta cho minMetric (độ tin cậy) là 90%, số luật là 10. Ta có kết quả như sau, có 2500 mẫu đạt minsup:

Minimum support: 0.75 (2500 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Large Itemsets L(1):

Int'l Plan=no 3010

VMail Message=Low 2658

Day Calls=High 2762

Eve Calls=High 2625

Intl Calls=Low 3205

Intl Charge=High 2617

CustServ Calls=Low 3066

Churn?=False. 2850

Size of set of large itemsets L(2): 12

Large Itemsets L(2):

Int'l Plan=no Day Calls=High 2504

Int'l Plan=no Intl Calls=Low 2897

Int'l Plan=no CustServ Calls=Low 2771

Int'l Plan=no Churn?=False. 2664

VMail Message=Low Intl Calls=Low 2559

Day Calls=High Intl Calls=Low 2655

Day Calls=High CustServ Calls=Low 2540

Eve Calls=High Intl Calls=Low 2527

Intl Calls=Low Intl Charge=High 2523

Intl Calls=Low CustServ Calls=Low 2944

Intl Calls=Low Churn?=False. 2741

CustServ Calls=Low Churn?=False. 2721

Size of set of large itemsets L(3): 4

Large Itemsets L(3):

Int'l Plan=no Intl Calls=Low CustServ Calls=Low 2664

Int'l Plan=no Intl Calls=Low Churn?=False. 2564

Int'l Plan=no CustServ Calls=Low Churn?=False. 2544

Intl Calls=Low CustServ Calls=Low Churn?=False. 2614

- Các luật nhận được là:

Best rules found:

```
1. Intl Charge=High 2617 ==> Intl Calls=Low 2523    <conf:(0.96)> lift:(1) lev:(0) [6] conv:(1.06)
2. VMail Message=Low 2658 ==> Intl Calls=Low 2559    <conf:(0.96)> lift:(1) lev:(0) [3] conv:(1.02)
3. Eve Calls=High 2625 ==> Intl Calls=Low 2527    <conf:(0.96)> lift:(1) lev:(0) [2] conv:(1.02)
4. Int'l Plan=no Churn?=False. 2664 ==> Intl Calls=Low 2564    <conf:(0.96)> lift:(1) lev:(0) [2] conv:(1.01)
5. Int'l Plan=no 3010 ==> Intl Calls=Low 2897    <conf:(0.96)> lift:(1) lev:(0) [2] conv:(1.01)
6. Churn?=False. 2850 ==> Intl Calls=Low 2741    <conf:(0.96)> lift:(1) lev:(0) [0] conv:(1)
7. Int'l Plan=no CustServ Calls=Low 2771 ==> Intl Calls=Low 2664    <conf:(0.96)> lift:(1) lev:(-0) [0] conv:(0.99)
8. Day Calls=High 2762 ==> Intl Calls=Low 2655    <conf:(0.96)> lift:(1) lev:(-0) [0] conv:(0.98)
9. CustServ Calls=Low Churn?=False. 2721 ==> Intl Calls=Low 2614    <conf:(0.96)> lift:(1) lev:(-0) [-2] conv:(0.97)
10. CustServ Calls=Low 3066 ==> Intl Calls=Low 2944    <conf:(0.96)> lift:(1) lev:(-0) [-4] conv:(0.96)
```

- Các thuộc tính khác đã được xét đến nhiều hơn, tuy nhiên không có luật nào có độ tin cậy là 100%.
- Xét luật thứ 6, ta thấy trong số các khách hàng không có xu hướng rời bỏ công ty (2850 mẫu) thì số cuộc gọi quốc tế của họ thấp (2741 mẫu). Tuy nhiên chưa có luật ngược lại. *Ta tạm đoán rằng khách hàng có số cuộc gọi quốc tế thấp thì sẽ không có xu hướng rời bỏ công ty.* Điều này cũng phù hợp với các luật được đưa ra ở các thí nghiệm trước (*khách hàng không có xu hướng rời bỏ công ty thì không sử dụng gói Quốc tế*)
- Xét luật thứ 9, ta thấy có thêm một điều kiện bổ sung cho luật thứ 6, đó là *nếu cuộc gọi tới dịch vụ khách hàng thấp và không có xu hướng rời bỏ công ty thì sẽ có số cuộc gọi quốc tế thấp.* Ta tạm để dự đoán này sang một bên để xem xét tiếp những thí nghiệm tiếp theo.
- Xét luật thứ 3 và 8, ta thấy nếu có *số cuộc gọi vào ban ngày và buổi tối cao thì số cuộc gọi quốc tế thấp* (độ tin cậy 0.96)

f. Chạy thử nghiệm lần 5:

Sử dụng bước tiền xử lí giống như thí nghiệm 4, lần này ta giảm minMetric còn 80% và tăng số luật lên 15. Có 2500 mẫu đạt minsup

Minimum support: 0.75 (2500 instances)

Minimum metric <confidence>: 0.8

Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Large Itemsets L(1):
Int'l Plan=no 3010
VMail Message=Low 2658
Day Calls=High 2762
Eve Calls=High 2625
Intl Calls=Low 3205
Intl Charge=High 2617
CustServ Calls=Low 3066
Churn?=False. 2850

Size of set of large itemsets L(2): 12

Large Itemsets L(2):
Int'l Plan=no Day Calls=High 2504
Int'l Plan=no Intl Calls=Low 2897
Int'l Plan=no CustServ Calls=Low 2771
Int'l Plan=no Churn?=False. 2664
VMail Message=Low Intl Calls=Low 2559
Day Calls=High Intl Calls=Low 2655
Day Calls=High CustServ Calls=Low 2540
Eve Calls=High Intl Calls=Low 2527
Intl Calls=Low Intl Charge=High 2523
Intl Calls=Low CustServ Calls=Low 2944
Intl Calls=Low Churn?=False. 2741
CustServ Calls=Low Churn?=False. 2721

Size of set of large itemsets L(3): 4

Large Itemsets L(3):
Int'l Plan=no Intl Calls=Low CustServ Calls=Low 2664
Int'l Plan=no Intl Calls=Low Churn?=False. 2564
Int'l Plan=no CustServ Calls=Low Churn?=False. 2544
Intl Calls=Low CustServ Calls=Low Churn?=False. 2614

- Trong L(2), ta thấy Int'l Plan=no Int'l Calls=Low 2897 (điều này đương nhiên vì nếu người dùng không sử dụng gói quốc tế thì sẽ có số cuộc gọi quốc tế thấp), nên ta không cần chú ý các large itemset này về sau

Ta có các luật như sau:

Best rules found:

```
1. Intl Charge=High 2617 ==> Intl Calls=Low 2523 <conf:(0.96)> lift:(1) lev:(0) [6] conv:(1.06)
2. VMail Message=Low 2658 ==> Intl Calls=Low 2559 <conf:(0.96)> lift:(1) lev:(0) [3] conv:(1.02)
3. Eve Calls=High 2625 ==> Intl Calls=Low 2527 <conf:(0.96)> lift:(1) lev:(0) [2] conv:(1.02)
4. Int'l Plan=no Churn?=False. 2664 ==> Intl Calls=Low 2564 <conf:(0.96)> lift:(1) lev:(0) [2] conv:(1.01)
5. Int'l Plan=no 3010 ==> Intl Calls=Low 2897 <conf:(0.96)> lift:(1) lev:(0) [2] conv:(1.01)
6. Churn?=False. 2850 ==> Intl Calls=Low 2741 <conf:(0.96)> lift:(1) lev:(0) [0] conv:(1)
7. Int'l Plan=no CustServ Calls=Low 2771 ==> Intl Calls=Low 2664 <conf:(0.96)> lift:(1) lev:(-0) [0] conv:(0.99)
8. Day Calls=High 2762 ==> Intl Calls=Low 2655 <conf:(0.96)> lift:(1) lev:(-0) [0] conv:(0.98)
9. CustServ Calls=Low Churn?=False. 2721 ==> Intl Calls=Low 2614 <conf:(0.96)> lift:(1) lev:(-0) [-2] conv:(0.97)
10. CustServ Calls=Low 3066 ==> Intl Calls=Low 2944 <conf:(0.96)> lift:(1) lev:(-0) [-4] conv:(0.96)
11. Int'l Plan=no Churn?=False. 2664 ==> CustServ Calls=Low 2544 <conf:(0.95)> lift:(1.04) lev:(0.03) [93] conv:(1.76)
12. Churn?=False. 2850 ==> CustServ Calls=Low 2721 <conf:(0.95)> lift:(1.04) lev:(0.03) [99] conv:(1.76)
13. Intl Calls=Low Churn?=False. 2741 ==> CustServ Calls=Low 2614 <conf:(0.95)> lift:(1.04) lev:(0.03) [92] conv:(1.72)
14. Intl Calls=Low Churn?=False. 2741 ==> Int'l Plan=no 2564 <conf:(0.94)> lift:(1.04) lev:(0.03) [88] conv:(1.49)
15. CustServ Calls=Low Churn?=False. 2721 ==> Int'l Plan=no 2544 <conf:(0.93)> lift:(1.04) lev:(0.03) [86] conv:(1.48)
```

- Luật thứ 6 xuất hiện lại như trong thí nghiệm 4 và ta vẫn chưa thấy có chiều ngược lại.
- Luật số 3 và 8 tiếp tục xuất hiện lại với độ tin cậy 96%
- Xét luật số 12, ta thấy khách hàng không có xu hướng rời bỏ công ty (2850 mẫu) thì số cuộc gọi tới dịch vụ khách hàng thấp (2721 mẫu) với độ tin cậy là 95%. Tuy nhiên cũng chưa có luật ngược lại nên ta tạm dự đoán rằng *số cuộc gọi tới dịch vụ khách hàng thấp thì không có xu hướng rời bỏ công ty.*

(*) Với phân loại {low, medium, high}

g. Chạy thử nghiệm lần 6:

- **Tiền xử lí:**

Lần này ta sẽ chạy Binning theo 3 cấp độ {low, medium, high}

Đầu tiên là chạy file python để Binning Data với cú pháp:

Ex: python Binning.py churn.csv churn3.csv -a all -b 3

Đưa file churn2.csv vào Weka để phân tích.

- **Chạy thử nghiệm:**

Cho minMetric là 90% và số luật là 10, ta có kết quả sau, có 2166 mẫu đạt minsup = 65%

```
Minimum support: 0.65 (2166 instances)
```

```
Minimum metric <confidence>: 0.9
```

```
Number of cycles performed: 7
```

```
Size of set of large itemsets L(2): 15
```

```
Large Itemsets L(2):
```

```
Area Code=Low Int'l Plan=no 2272  
Int'l Plan=no VMail Plan=no 2180  
Int'l Plan=no VMail Message=Low 2221  
Int'l Plan=no Eve Mins=Medium 2194  
Int'l Plan=no Night Mins=Medium 2363  
Int'l Plan=no Night Calls=Medium 2296  
Int'l Plan=no Intl Calls=Low 2200  
Int'l Plan=no Intl Charge=High 2286  
Int'l Plan=no CustServ Calls=Low 2380  
Int'l Plan=no Churn?=False. 2664  
VMail Plan=no VMail Message=Low 2411  
Day Mins=Medium Day Charge=Medium 2282  
Night Mins=Medium Churn?=False. 2212  
Night Calls=Medium Churn?=False. 2183  
CustServ Calls=Low Churn?=False. 2336
```

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

```
Large Itemsets L(3):
```

```
Int'l Plan=no VMail Plan=no VMail Message=Low 2180  
Int'l Plan=no CustServ Calls=Low Churn?=False. 2182
```

Các luật tìm được:

Best rules found:

1. VMail Plan=no 2411 ==> VMail Message=Low 2411 <conf:(1)> lift:(1.36) lev:(0.19) [632] conv:(632.95)
2. Int'l Plan=no VMail Plan=no 2180 ==> VMail Message=Low 2180 <conf:(1)> lift:(1.36) lev:(0.17) [572] conv:(572.31)
3. Int'l Plan=no VMail Message=Low 2221 ==> VMail Plan=no 2180 <conf:(0.98)> lift:(1.36) lev:(0.17) [573] conv:(14.63)
4. VMail Message=Low 2458 ==> VMail Plan=no 2411 <conf:(0.98)> lift:(1.36) lev:(0.19) [632] conv:(14.17)
5. Day Charge=Medium 2327 ==> Day Mins=Medium 2282 <conf:(0.98)> lift:(1.38) lev:(0.19) [623] conv:(14.52)
6. Day Mins=Medium 2376 ==> Day Charge=Medium 2282 <conf:(0.96)> lift:(1.38) lev:(0.19) [623] conv:(7.55)
7. Churn?=False. 2850 ==> Int'l Plan=no 2664 <conf:(0.93)> lift:(1.04) lev:(0.03) [90] conv:(1.48)
8. CustServ Calls=Low Churn?=False. 2336 ==> Int'l Plan=no 2182 <conf:(0.93)> lift:(1.03) lev:(0.02) [72] conv:(1.46)
9. Int'l Plan=no CustServ Calls=Low 2380 ==> Churn?=False. 2182 <conf:(0.92)> lift:(1.07) lev:(0.04) [146] conv:(1.73)
10. Area Code=Low 2493 ==> Int'l Plan=no 2272 <conf:(0.91)> lift:(1.01) lev:(0.01) [20] conv:(1.09)

- Chú ý luật số 9, ta thấy nếu khách hàng không có dịch vụ gọi quốc tế và số cuộc gọi tới dịch vụ khách hàng thấp thì sẽ không có xu hướng rời khỏi công ty (độ tin cậy 92%). Luật này có thể xem là luật ngược của luật số 12 trong thí nghiệm 5 (*khách hàng không có xu hướng rời khỏi công ty thì sẽ có số cuộc gọi tới dịch vụ khách hàng thấp*). Vậy dự đoán *số cuộc gọi tới dịch vụ khách hàng thấp thì không có xu hướng rời bỏ công ty* là đúng.

- Giảm độ tin cậy minMetric xuống 0.8, ta cũng thu được kết quả tương tự.

(*) Với phân loại {very_low, low, medium, high, very_high}

h. Chạy thử nghiệm lần 7:

- **Tiền xử lí:**

Lần này ta sẽ chạy Binning theo 5 cấp độ {very_low, low, medium, high, very_high}

Đầu tiên là chạy file python để Binning Data với cú pháp:

Ex: `python Binning.py churn.csv churn5.csv -a all -b 5`

Đưa file churn5.csv vào Weka để phân tích.

- **Chạy thử nghiệm:**

Cho minMetric là 90% và số luật là 15, ta có kết quả sau, có 2333 mẫu đạt minsup = 70%

Minimum support: 0.7 (2333 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 6

```
Size of set of large itemsets L(2): 7
```

```
Large Itemsets L(2):
```

```
Area Code=VeryLow Intl Charge=VeryHigh 2449
```

```
Int'l Plan=no Intl Charge=VeryHigh 2957
```

```
Int'l Plan=no Churn?=False. 2664
```

```
VMail Plan=no VMail Message=VeryLow 2411
```

```
VMail Plan=no Intl Charge=VeryHigh 2368
```

```
VMail Message=VeryLow Intl Charge=VeryHigh 2373
```

```
Intl Charge=VeryHigh Churn?=False. 2796
```

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

```
Large Itemsets L(3):
```

```
Int'l Plan=no Intl Charge=VeryHigh Churn?=False. 2614
```

```
VMail Plan=no VMail Message=VeryLow Intl Charge=VeryHigh 2368
```

- Xét L(2), ta thấy Int'l Plan=no Intl Charge=VeryHigh 2957, tức là nếu khách hàng không sử dụng dịch vụ gọi quốc tế thì phí gọi quốc tế cao (2957 mẫu). Đây là điều hiển nhiên nên ta sẽ không xét large itemsets này về sau.

Các luật tìm được là:

Best rules found:

```
1. VMail Plan=no 2411 ==> VMail Message=VeryLow 2411 <conf:(1)> lift:(1.38) lev:(0.2) [663] conv:(663.33)
2. VMail Plan=no Intl Charge=VeryHigh 2368 ==> VMail Message=VeryLow 2368 <conf:(1)> lift:(1.38) lev:(0.2) [651] conv:(651.5)
3. VMail Message=VeryLow 2416 ==> VMail Plan=no 2411 <conf:(1)> lift:(1.38) lev:(0.2) [663] conv:(111.39)
4. VMail Message=VeryLow Intl Charge=VeryHigh 2373 ==> VMail Plan=no 2368 <conf:(1)> lift:(1.38) lev:(0.2) [651] conv:(109.41)
5. Int'l Plan=no 3010 ==> Intl Charge=VeryHigh 2957 <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.99)
6. Area Code=VeryLow 2493 ==> Intl Charge=VeryHigh 2449 <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.98)
7. VMail Message=VeryLow 2416 ==> Intl Charge=VeryHigh 2373 <conf:(0.98)> lift:(1) lev:(-0) [0] conv:(0.97)
8. VMail Plan=no 2411 ==> Intl Charge=VeryHigh 2368 <conf:(0.98)> lift:(1) lev:(-0) [0] conv:(0.97)
9. VMail Plan=no VMail Message=VeryLow 2411 ==> Intl Charge=VeryHigh 2368 <conf:(0.98)> lift:(1) lev:(-0) [0] conv:(0.97)
10. VMail Plan=no 2411 ==> VMail Message=VeryLow Intl Charge=VeryHigh 2368 <conf:(0.98)> lift:(1.38) lev:(0.2) [651] conv:(15.78)
11. Int'l Plan=no Churn?=False. 2664 ==> Intl Charge=VeryHigh 2614 <conf:(0.98)> lift:(1) lev:(-0) [-2] conv:(0.92)
12. Churn?=False. 2850 ==> Intl Charge=VeryHigh 2796 <conf:(0.98)> lift:(1) lev:(-0) [-3] conv:(0.92)
13. VMail Message=VeryLow 2416 ==> VMail Plan=no Intl Charge=VeryHigh 2368 <conf:(0.98)> lift:(1.38) lev:(0.2) [651] conv:(14.28)
14. Intl Charge=VeryHigh Churn?=False. 2796 ==> Int'l Plan=no 2614 <conf:(0.93)> lift:(1.04) lev:(0.03) [88] conv:(1.48)
15. Churn?=False. 2850 ==> Int'l Plan=no 2664 <conf:(0.93)> lift:(1.04) lev:(0.03) [90] conv:(1.48)
```

- Xét luật số 15, ta thấy nội dung trùng với luật đã xét ở thí nghiệm 1, 2
- Các luật còn lại không có gì đáng chú ý

(*) Tổng kết cho 3 lần chạy thử nghiệm 4, 5, 6 với phương pháp Binning:

Qua 3 lần thử nghiệm ta đã có thêm các mối quan hệ về thuộc tính CustServ Calls (cuộc gọi về dịch vụ khách hàng); làm rõ thêm mối liên hệ về số cuộc gọi quốc tế với xu hướng rời bỏ công ty. Cụ thể là:

+ Số cuộc gọi tới dịch vụ khách hàng thấp thì không có xu hướng rời bỏ công ty

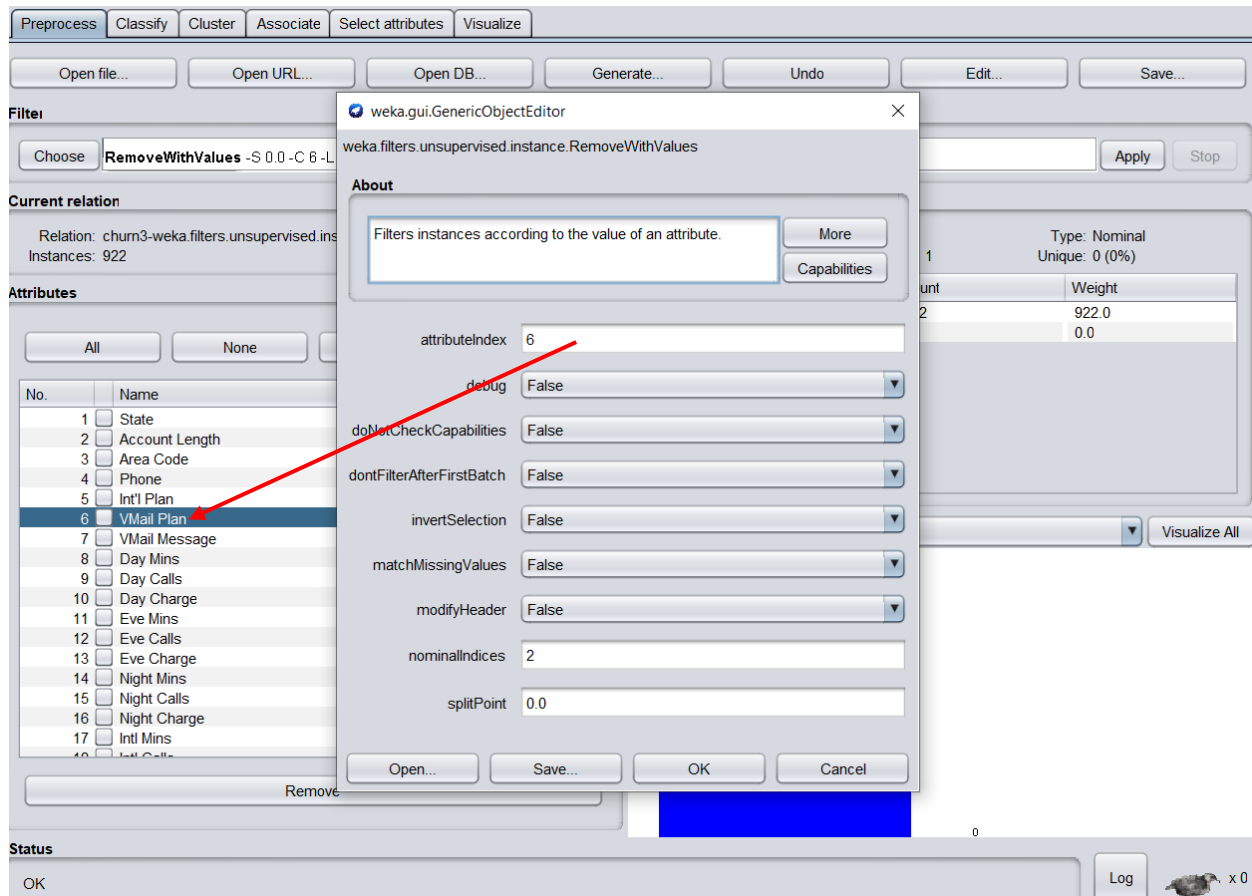
Tuy nhiên các vấn đề còn lại ở lần tổng kết đầu tiên (không xét đến các thuộc tính khác như Day_Mins, Day_Calls, Eve_Mins, Eve_Calls,...) vẫn chưa được giải quyết.

=> Cần tìm các phương pháp xử lý thích hợp hơn

3. Phân tích kết quả thí nghiệm:

a) Phân tích mối liên hệ giữa hành vi sử dụng thư thoại và rời bỏ công ty

- Trong phần tổng kết 3 lần chạy thử nghiệm đầu tiên ta thấy có đến 922 thuê bao có sử dụng dịch vụ thư thoại nhưng chưa được khai thác liệu nhóm này có xu hướng rời bỏ công ty ra sao. Trong phần phân tích này chọn giải pháp là lược bỏ toàn bộ các thuê bao không sử dụng dịch vụ thư thoại bằng các sử dụng Filter **RemoveWithValues (unsupervised)** của Weka. Ta sẽ thực hiện khảo sát trên cách phân loại {low, medium, high}

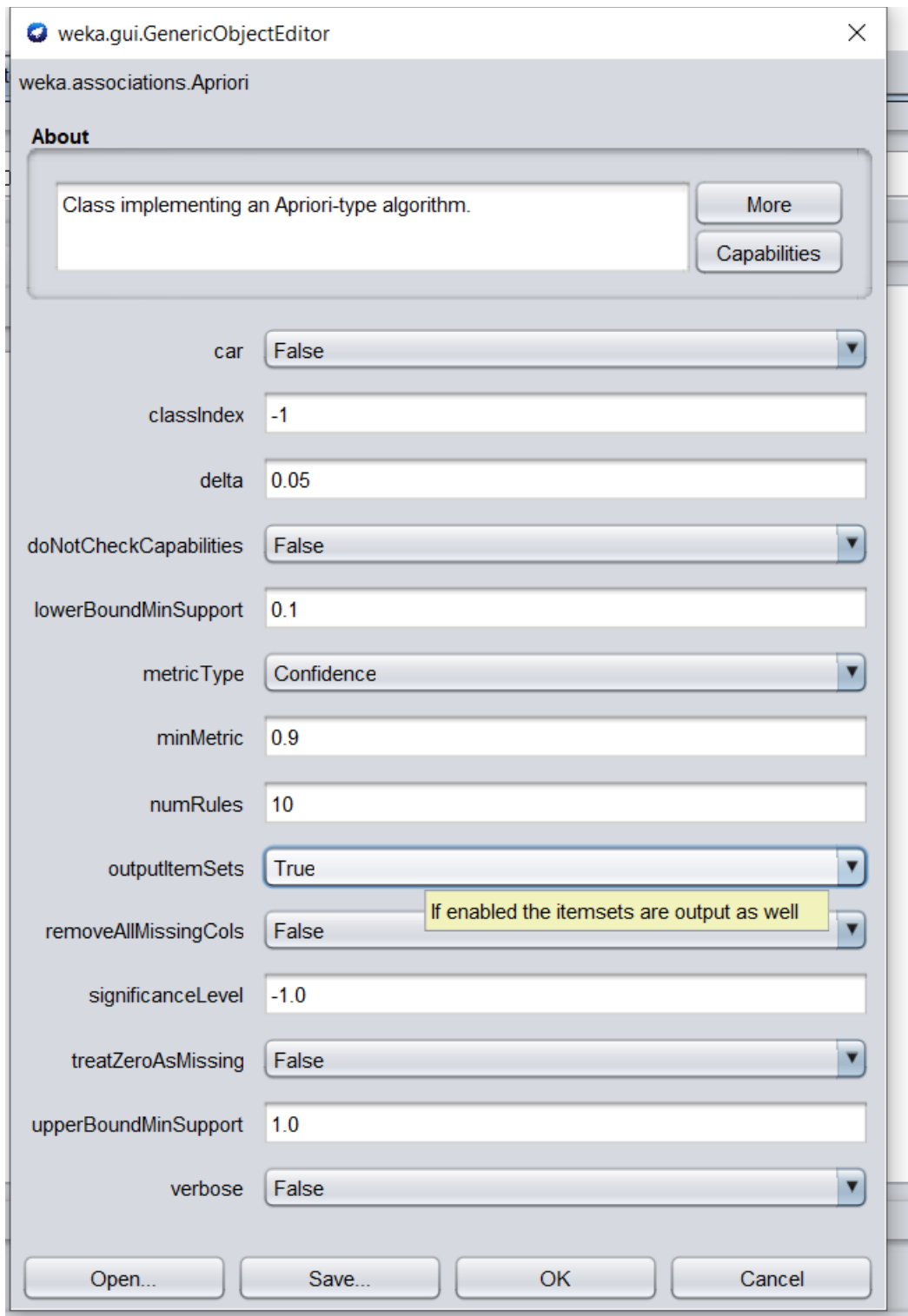


- Sau khi áp dụng bộ lọc, chỉ còn lại 922 mẫu dữ liệu có sử dụng thư thoại

Selected attribute			
Name: VMail Plan		Type: Nominal	
Missing: 0 (0%)		Distinct: 1	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	yes	922	922.0
2	no	0	0.0

(*) Chạy thử nghiệm lần 1 (thí nghiệm số 8):

- Sử dụng các thông số mặc định của Weka (nếu muốn xem xét cái *large Itemset* thì phải chọn mode *True* tại *outputItemSets*)



- Kết quả: Có 784 mẫu đạt minsup = 0.85

```
Minimum support: 0.85 (784 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3
```


- Các large Itemsets:

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Large Itemsets L(1):

Int'l Plan=no 830

VMail Plan=yes 922

Churn?=False 842

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

Large Itemsets L(2):

Int'l Plan=no VMail Plan=yes 830

Int'l Plan=no Churn?=False 786

VMail Plan=yes Churn?=False 842

Size of set of large itemsets L(3): 1

Large Itemsets L(3):

Int'l Plan=no VMail Plan=yes Churn?=False 786

Cả L1, L2, L3 đều xuất hiện Vmail Plan=yes.

Trong đó:

+ Mail_Plan=yes Churn=False. 842 => minsup của itemset này là 0.91 (842/922)

+ Intl_Plan=no VMail_Plan=yes Churn=False. 786 => minsup của itemsets này là 0.85 (786/922)

- Các luật:

Best rules found:

```
1. Churn?=False 842 ==> VMail Plan=yes 842    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. Int'l Plan=no 830 ==> VMail Plan=yes 830    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. Int'l Plan=no Churn?=False 786 ==> VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. Int'l Plan=no 830 ==> Churn?=False 786    <conf:(0.95)> lift:(1.04) lev:(0.03) [28] conv:(1.6)
5. Int'l Plan=no VMail Plan=yes 830 ==> Churn?=False 786    <conf:(0.95)> lift:(1.04) lev:(0.03) [28] conv:(1.6)
6. Int'l Plan=no 830 ==> VMail Plan=yes Churn?=False 786    <conf:(0.95)> lift:(1.04) lev:(0.03) [28] conv:(1.6)
7. Churn?=False 842 ==> Int'l Plan=no 786    <conf:(0.93)> lift:(1.04) lev:(0.03) [28] conv:(1.47)
8. VMail Plan=yes Churn?=False 842 ==> Int'l Plan=no 786    <conf:(0.93)> lift:(1.04) lev:(0.03) [28] conv:(1.47)
9. Churn?=False 842 ==> Int'l Plan=no VMail Plan=yes 786    <conf:(0.93)> lift:(1.04) lev:(0.03) [28] conv:(1.47)
10. VMail Plan=yes 922 ==> Churn?=False 842    <conf:(0.91)> lift:(1) lev:(0) [0] conv:(0.99)
```

Quan sát bộ luật ta chú ý các luật có thuộc tính mẫu chốt “Churn?”

+ Luật 1. Churn?=False 842 ==> VMail Plan=yes 842 <conf:(1)> lift:(1) lev:(0) [0] conv:(0) và Luật 10. VMail_Plan=yes 922 ==> Churn=False.

842 <conf:(0.91)> lift:(1) lev:(0) [0] conv:(0.99) => Ta thấy 2 luật này cung cấp một thông tin là các thuê bao có sử dụng thư thoại thì không có xu hướng rời bỏ công ty với các độ tin cậy cao là 100% và 91%

+ Ta có thể suy đoán hầu hết các khách hàng có sử dụng dịch vụ thư thoại thì có xu hướng rời bỏ công ty rất thấp. Bên cạnh đó ta cũng có 2 suy đoán khác cần được làm rõ cho (922 – 842 = 80 khách hàng còn lại chưa phân tích) là:

- Khách hàng sử dụng gói thư thoại và gói quốc tế
- Khách hàng chỉ sử dụng gói thư thoại không sử dụng gói quốc tế

(*) Chạy thử nghiệm lần 2 (thí nghiệm số 9):

- Thay đổi thông số như hình dưới và chạy thử nghiệm lại:



minMetric	0.9
numRules	20
outputItemSets	True

- Kết quả: Có 645 mẫu đạt minsup = 0.7
- Các largeItemsets:

Size of set of large itemsets L(2): 17

Large Itemsets L(2):

Int'l Plan=no VMail Plan=yes 830

Int'l Plan=no Night Mins=Medium 655

Int'l Plan=no CustServ Calls=Low 679

Int'l Plan=no Churn?=False 786

VMail Plan=yes Day Mins=Medium 667

VMail Plan=yes Day Charge=Medium 657

VMail Plan=yes Eve Mins=Medium 665

VMail Plan=yes Eve Calls=Medium 667

VMail Plan=yes Night Mins=Medium 722

VMail Plan=yes Night Calls=Medium 702

VMail Plan=yes Intl Calls=Low 673

VMail Plan=yes Intl Charge=High 693

VMail Plan=yes CustServ Calls=Low 752

VMail Plan=yes Churn?=False 842

Day Mins=Medium Day Charge=Medium 646

Night Mins=Medium Churn?=False 661

CustServ Calls=Low Churn?=False 709

Các itemsets này cho thấy rằng khách hàng gọi điện vào buổi tối, ban đêm và ban ngày ở mức trung bình thì thường sử dụng dịch vụ thư thoại

- Các Luật:

+ Luật 16. VMail_Plan=yes CustServCalls=low 752 ==> Churn=False. 709
<conf:(0.94)> lift:(1.03) lev:(0.02) [22] conv:(1.48) và Luật 17.
CustServCalls=low 752 ==> VMail_Plan=yes Churn=False. 709
<conf:(0.94)> lift:(1.03) lev:(0.02) [22] conv:(1.48). Ta thấy 2 luật này cho
ta thấy rằng *các khách hàng có sử dụng dịch vụ thư thoại thì cũng ít gọi đến
chăm sóc khách hàng*

**(*) Tổng kết cho 2 lần chạy thử nghiệm cho phân tích mối liên hệ giữa
hành vi sử dụng thư thoại và rời bỏ công ty**

- Nhóm khách hàng sử dụng thư thoại có xu hướng rời bỏ công ty thấp và
xu hướng gọi điện không quá nhiều vào buổi tối và ban đêm, và còn ít sử
dụng dịch vụ chăm sóc khách hàng

- Có 2 xu hướng đáng chú ý:

- *Khách hàng có sử dụng gói Quốc tế thì có xu hướng rời bỏ công ty cao*
- *Khách hàng có sử dụng dịch vụ Thư thoại có xu hướng rời bỏ công ty
thấp.*

**b) Phân tích trường hợp “Khách hàng chỉ sử dụng gói thư thoại không
sử dụng gói quốc tế”**

- Ta sử dụng bộ lọc *RemoveWithValues (unsupervised)* của Weka để lọc
ra một tập dữ liệu trong đó Intl_Plan=no và VMail_Plan=yes => được
một tập dữ liệu gồm có 830 mẫu

Selected attribute			
Name: Int'l Plan		Type: Nominal	
Missing: 0 (0%)		Unique: 0 (0%)	
		Distinct: 1	
No.	Label	Count	Weight
1	no	830	830.0
2	yes	0	0.0

(*) Chạy thử nghiệm lần 1 (thí nghiệm số 10):

- Sử dụng các thông số mặc định
- Các luật:

Best rules found:

```
1. VMail Plan=yes 830 ==> Int'l Plan=no 830    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. Int'l Plan=no 830 ==> VMail Plan=yes 830    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. Churn?=False 786 ==> Int'l Plan=no 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. Churn?=False 786 ==> VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. VMail Plan=yes Churn?=False 786 ==> Int'l Plan=no 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. Int'l Plan=no Churn?=False 786 ==> VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. Churn?=False 786 ==> Int'l Plan=no VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. Int'l Plan=no 830 ==> Churn?=False 786    <conf:(0.95)> lift:(1) lev:(0) [0] conv:(0.98)
9. VMail Plan=yes 830 ==> Churn?=False 786    <conf:(0.95)> lift:(1) lev:(0) [0] conv:(0.98)
10. Int'l Plan=no VMail Plan=yes 830 ==> Churn?=False 786    <conf:(0.95)> lift:(1) lev:(0) [0] conv:(0.98)
```

Chú ý luật 8, 9, 10, ta thấy các luật không có gì thay đổi so với các dự đoán xu hướng ở các thí nghiệm trước. Luật số 10 cho thấy *khách hàng không sử dụng dịch vụ quốc tế và có sử dụng thư thoại thì không rời bỏ công ty (độ tin cậy 95%)*.

(*) Chạy thử nghiệm lần 2 (thí nghiệm số 11):

- Điều chỉnh số luật thành 20
- Các luật:

Best rules found:

```
1. VMail Plan=yes 830 ==> Int'l Plan=no 830    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. Int'l Plan=no 830 ==> VMail Plan=yes 830    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. Churn?=False 786 ==> Int'l Plan=no 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. Churn?=False 786 ==> VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. VMail Plan=yes Churn?=False 786 ==> Int'l Plan=no 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. Int'l Plan=no Churn?=False 786 ==> VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. Churn?=False 786 ==> Int'l Plan=no VMail Plan=yes 786    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. CustServ Calls=Low 679 ==> Int'l Plan=no 679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
9. CustServ Calls=Low 679 ==> VMail Plan=yes 679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. VMail Plan=yes CustServ Calls=Low 679 ==> Int'l Plan=no 679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
11. Int'l Plan=no CustServ Calls=Low 679 ==> VMail Plan=yes 679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
12. CustServ Calls=Low 679 ==> Int'l Plan=no VMail Plan=yes 679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
13. CustServ Calls=Low Churn?=False 661 ==> Int'l Plan=no 661    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
14. CustServ Calls=Low Churn?=False 661 ==> VMail Plan=yes 661    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
15. VMail Plan=yes CustServ Calls=Low Churn?=False 661 ==> Int'l Plan=no 661    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
16. Int'l Plan=no CustServ Calls=Low Churn?=False 661 ==> VMail Plan=yes 661    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
17. CustServ Calls=Low Churn?=False 661 ==> Int'l Plan=no VMail Plan=yes 661    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
18. Night Mins=Medium 655 ==> Int'l Plan=no 655    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
19. Night Mins=Medium 655 ==> VMail Plan=yes 655    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
20. VMail Plan=yes Night Mins=Medium 655 ==> Int'l Plan=no 655    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

Quan 2 lần chạy thử nghiệm ta nhận thấy các suy đoán ta được ra ngày càng được minh chứng tính đúng đắn rõ hơn.

c) Phân tích trường hợp “Khách hàng sử dụng gói thư thoại và gói quốc tế”

- Ta sử dụng bộ lọc *RemoveWithValues (unsupervised)* của Weka để lọc ra một tập dữ liệu trong đó Intl_Plan=yes và VMail_Plan=yes => được một tập dữ liệu gồm có 92 mẫu

Selected attribute			
Name: Int'l Plan		Type: Nominal	
Missing: 0 (0%)		Unique: 0 (0%)	
		Distinct: 1	
No.	Label	Count	Weight
1	no	0	0.0
2	yes	92	92.0

(*) Chạy thử nghiệm lần 1 (thí nghiệm số 12):

- Với thông số mặc định của Weka
- Kết quả

Best rules found:

```

1. VMail Plan=yes 92 ==> Int'l Plan=yes 92    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. Int'l Plan=yes 92 ==> VMail Plan=yes 92    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. Intl Charge=High 76 ==> Int'l Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. Intl Charge=High 76 ==> VMail Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. VMail Plan=yes Intl Charge=High 76 ==> Int'l Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. Int'l Plan=yes Intl Charge=High 76 ==> VMail Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. Intl Charge=High 76 ==> Int'l Plan=yes VMail Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. CustServ Calls=Low 73 ==> Int'l Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
9. CustServ Calls=Low 73 ==> VMail Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. VMail Plan=yes CustServ Calls=Low 73 ==> Int'l Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

```

⇒ Chưa thể rút ra kết luận cho thuộc tính mẫu chốt

(*) Chạy thử nghiệm lần 2 (thí nghiệm số 13):

- Với số lượng luật tăng lên 20
- Kết quả:

Best rules found:

```

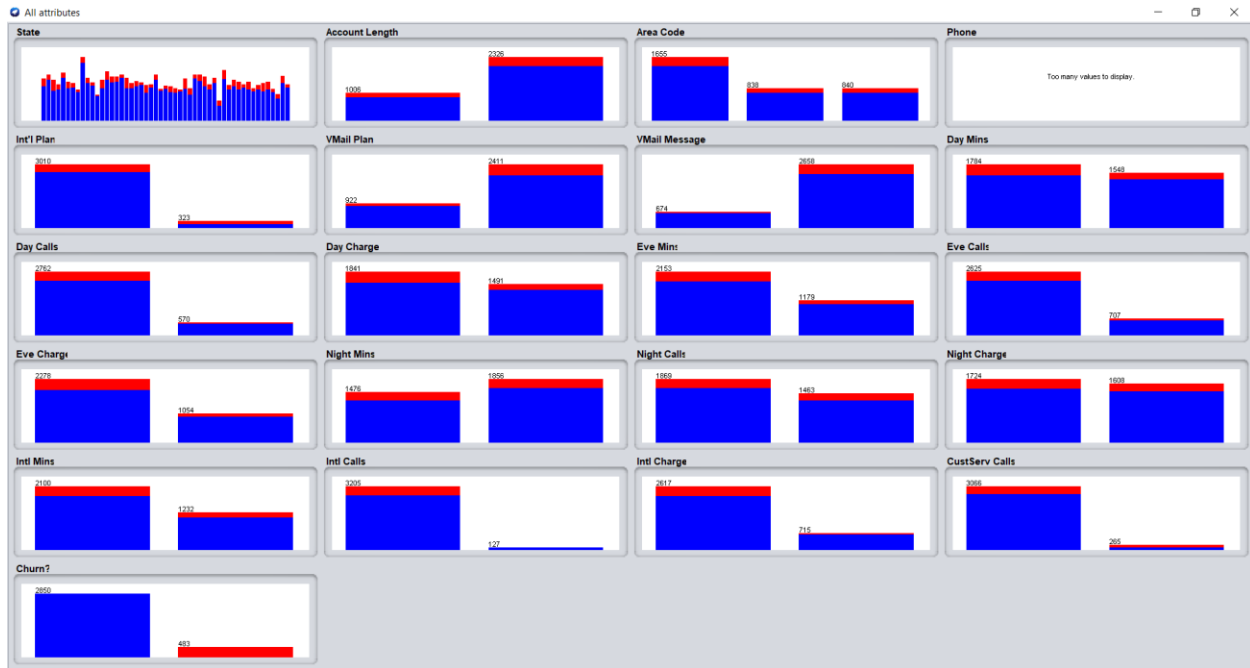
1. VMail Plan=yes 92 ==> Int'l Plan=yes 92    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. Int'l Plan=yes 92 ==> VMail Plan=yes 92    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. Intl Charge=High 76 ==> Int'l Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. Intl Charge=High 76 ==> VMail Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. VMail Plan=yes Intl Charge=High 76 ==> Int'l Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. Int'l Plan=yes Intl Charge=High 76 ==> VMail Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. Intl Charge=High 76 ==> Int'l Plan=yes VMail Plan=yes 76   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. CustServ Calls=Low 73 ==> Int'l Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
9. CustServ Calls=Low 73 ==> VMail Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. VMail Plan=yes CustServ Calls=Low 73 ==> Int'l Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
11. Int'l Plan=yes CustServ Calls=Low 73 ==> VMail Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
12. CustServ Calls=Low 73 ==> Int'l Plan=yes VMail Plan=yes 73   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
13. Night Calls=Medium 71 ==> Int'l Plan=yes 71   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
14. Night Calls=Medium 71 ==> VMail Plan=yes 71   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
15. VMail Plan=yes Night Calls=Medium 71 ==> Int'l Plan=yes 71   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
16. Int'l Plan=yes Night Calls=Medium 71 ==> VMail Plan=yes 71   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
17. Night Calls=Medium 71 ==> Int'l Plan=yes VMail Plan=yes 71   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
18. Day Mins=Medium 68 ==> Int'l Plan=yes 68   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
19. Day Charge=Medium 68 ==> Int'l Plan=yes 68   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
20. Eve Calls=Medium 68 ==> Int'l Plan=yes 68   <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

```

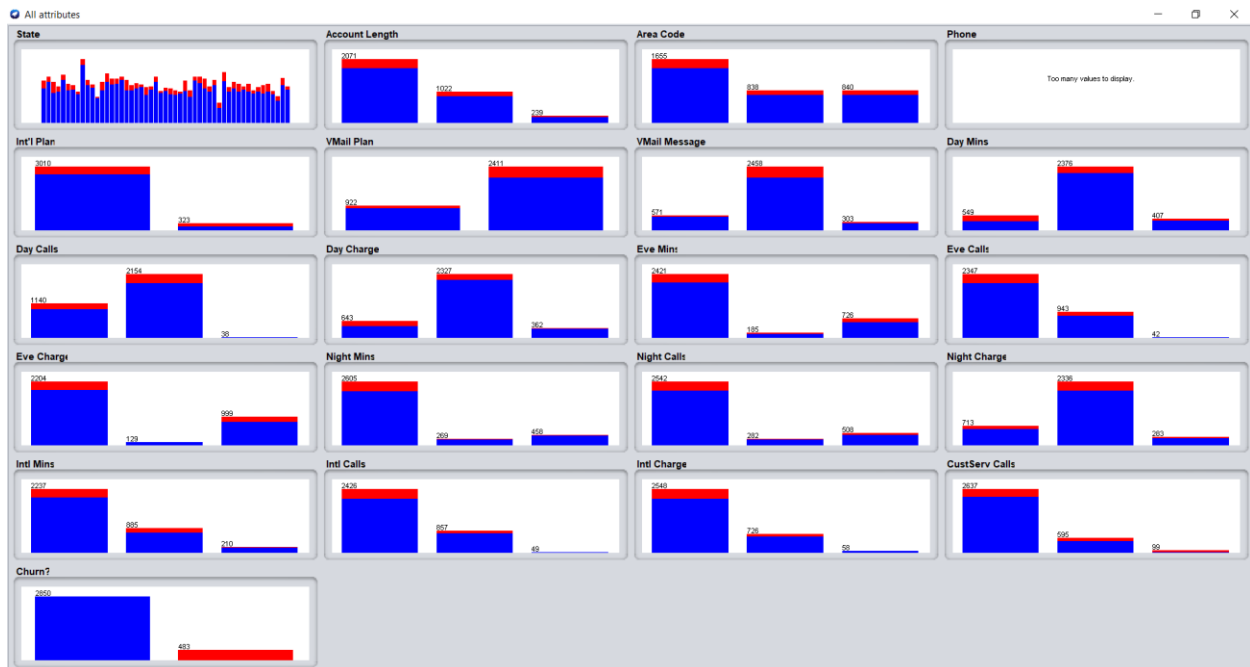
⇒ Chưa thể rút ra kết luận cho thuộc tính mẫu chốt

4. Phân tích biểu đồ:

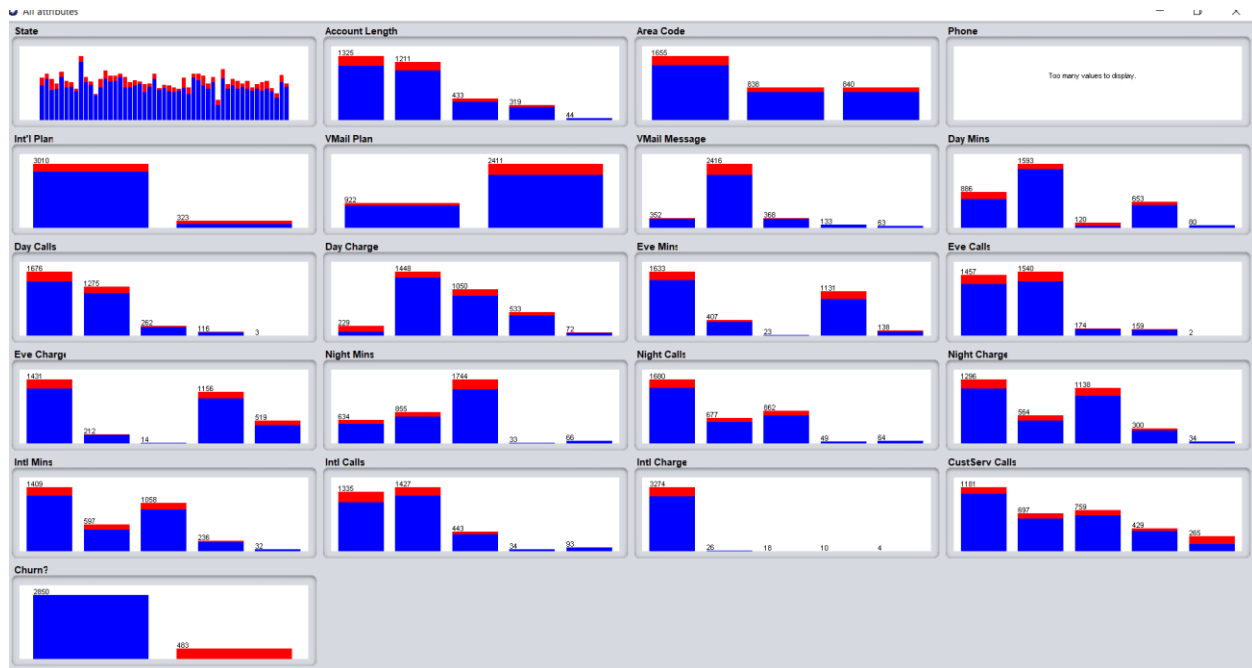
(*) Với phân cấp {low, high}



(*) Với phân cấp {low, medium, high}



(*) Với phân cấp {verylow, low, medium, high, veryhigh}



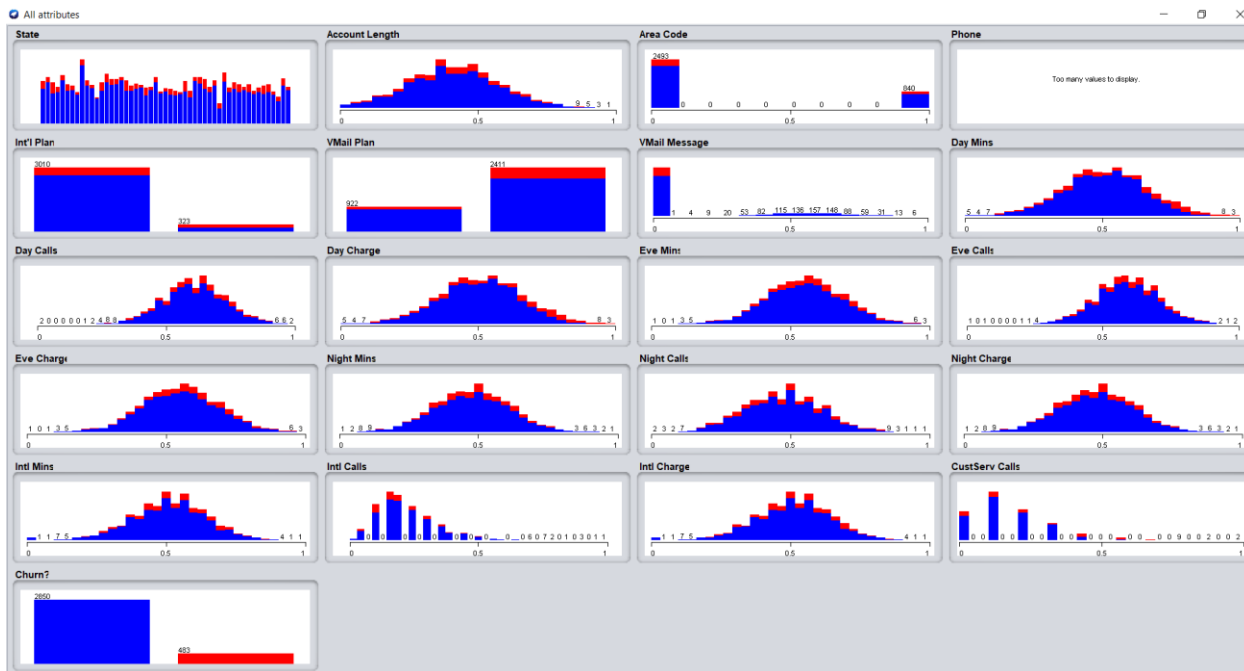
=> Thông qua ba biểu đồ Histogram của 3 phân cấp ta khó mà có thể biết được dữ liệu bị có bị lệch trái hay lệch phải hay không nên thử quan sát các biểu đồ của dữ liệu sau khi đã được chuẩn hóa.

Sử dụng file python để Normalize Data với cú pháp

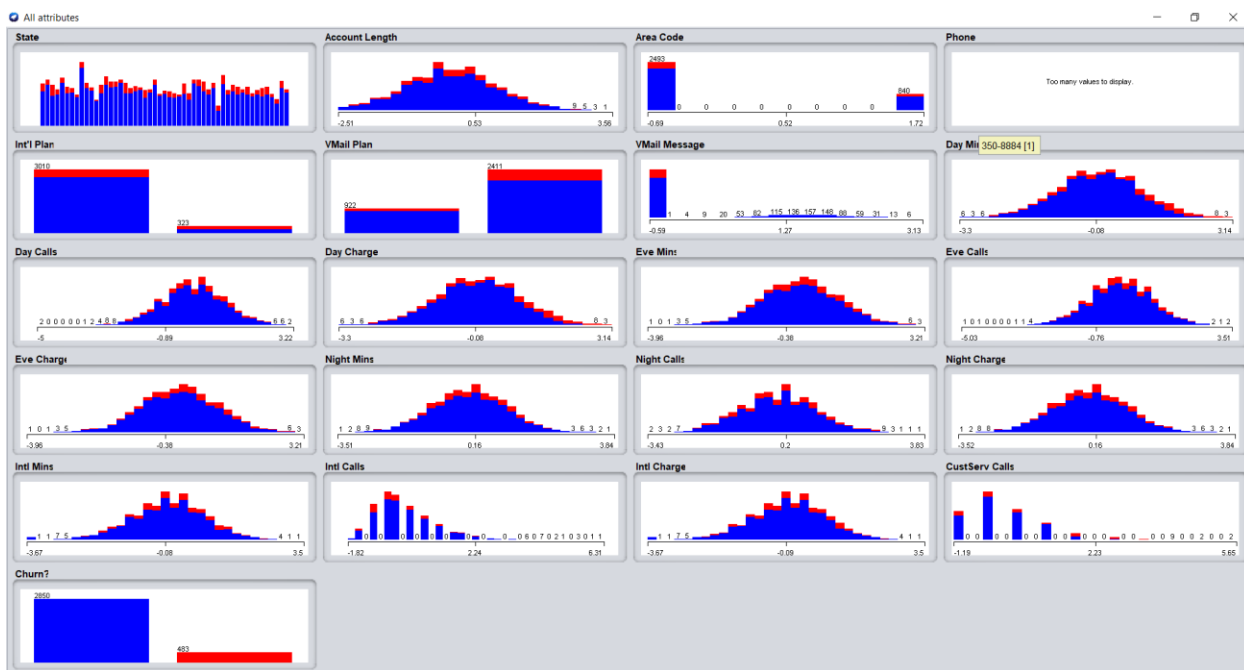
Ex: `python Normalize.py churn.csv churn_no1.csv -m minmax -a all`

`python Normalize.py churn.csv churn_no2.csv -m zscore -a all`

(*) **Chuẩn hóa minmax**



(*) Chuẩn hóa Zscore



- ⇒ Có thể thấy được dữ liệu không bị lệch phải hay lệch trái
- ⇒ Tần suất gọi điện và hành vi rời bỏ công ty hầu như không có luật kết hợp. Quan sát các histogram, ta thấy dù là gọi vào thời gian nào hoặc có gọi quốc tế hay không cũng không ảnh hưởng nhiều đến việc rời bỏ công ty.
- ⇒ Các thuộc tính State, Account Length, Area Code, Phone tương tự cũng không ảnh hưởng đến việc khách hàng có rời bỏ công ty hay không.

IV. SUMMARY:

Qua các thí nghiệm, ta rút ra được các kết luận như sau:

- *Khách hàng không sử dụng dịch vụ quốc tế và có sử dụng thư thoại thì không rời bỏ công ty.*
- *Khách hàng có sử dụng gói Quốc tế thì có xu hướng rời bỏ công ty cao*
- *Các khách hàng có sử dụng dịch vụ thư thoại thì cũng ít gọi đến chăm sóc khách hàng*
- *Số cuộc gọi tới dịch vụ khách hàng thấp thì không có xu hướng rời bỏ công ty*

Qua đây công ty có thể có điều chỉnh như sau:

- Cải thiện dịch vụ gói Quốc tế để làm giảm số lượng khách hàng bỏ đi
- Giữ nguyên & phát triển dịch vụ thư thoại để giữ khách hàng ổn định

Điểm mạnh: sử dụng các phương pháp tiền xử lí khác nhau để khai thác dữ liệu tốt nhất

Điểm yếu: chưa khai thác dữ liệu triệt để, không xem xét được mối tương quan giữa các tập thuộc tính như Day Calls, Night Calls, Eve Calls, Account Length,...