**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

****

**LAB 02:**

**FREQUENT ITEMSETS**

**AND**

**ASSOCIATION RULES**

1. **PHÂN CÔNG CÔNG VIỆC**
2. **NỘI DUNG**
3. **DATA:**
4. **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 | Norminal | 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 | Norminal | Mã vùng |
| 4 | Phone number | Norminal | Số điện thoại |
| 5 | Int’l Plan | Norminal {yes,no} | Khách hành có sử dụng gói cước quốc tế hay không? |
| 6 | VoiceMail Plan | Norminal {yes,no} | Khách hành có sử dụng gói cước thư thoại hay không? |
| 7 | VMail Messages | Numeric | Số lượng thư thoại |
| 8 | Day Mins | Numeric | Thời gian thuê bao sử dụng vào ban ngày |
| 9 | Day Calls | Numeric | Tổng số cuốc gọi và ban ngày |
| 10 |  |  |  |
| 11 |  |  |  |
| 12 |  |  |  |
| 13 |  |  |  |
| 14 |  |  |  |
| 15 |  |  |  |
| 16 |  |  |  |
| 17 |  |  |  |
| 18 |  |  |  |
| 19 |  |  |  |
| 20 |  |  |  |
| 21 | Churn | Norminal  {True, False} | Thuê bao có rời bỏ công ty hay không? |

1. **Khái niệm phân cấp:**

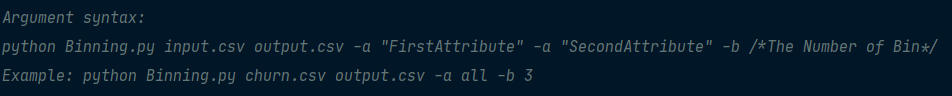
Áp dụng Set grouping hierarchy để phân loại các thuộc tính số thành các nhóm:  
 {low,high}  
 {low,medium,high}  
 {very\_low, low, medium, high, very\_high}

1. **CODE**
2. **Mô tả chức năng file:**

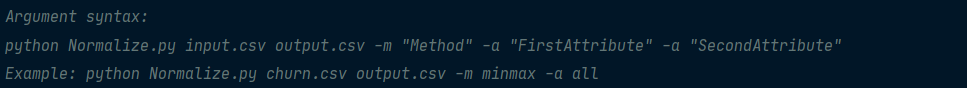
**+ Binning.py:** Chia dữ liệu theo khoảng. Gồm các khoảng mặc định **{low,high}, {low,medium,high}, {very\_low,low,medium,high,very\_high}**

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

1. **Mô tả cú pháp (Command line argument):**

****

*Argument syntax of Binning.py*

****

*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}**

****

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



+ 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**

1. **EXPERIMENTS**
2. **Mục đích của việc phân tích dữ liệu:**
3. **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 mấu 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)* |

1. **Phục vụ cho việc ra quyết định? Vậy bạn cần dựa vào dữ liệu để quyết định điều gì?**

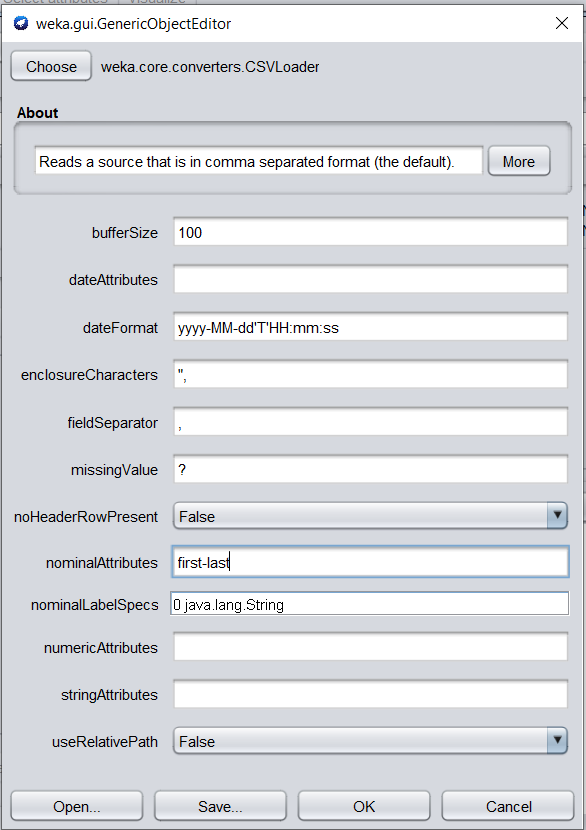
+ Cho biết các nguyên nhân dẫn đến việc khách hàng rời bỏ công ty

+ Nâng cao và phát huy thuộc tính nào để giảm bớt số lượng khách hàng rời bỏ công ty?

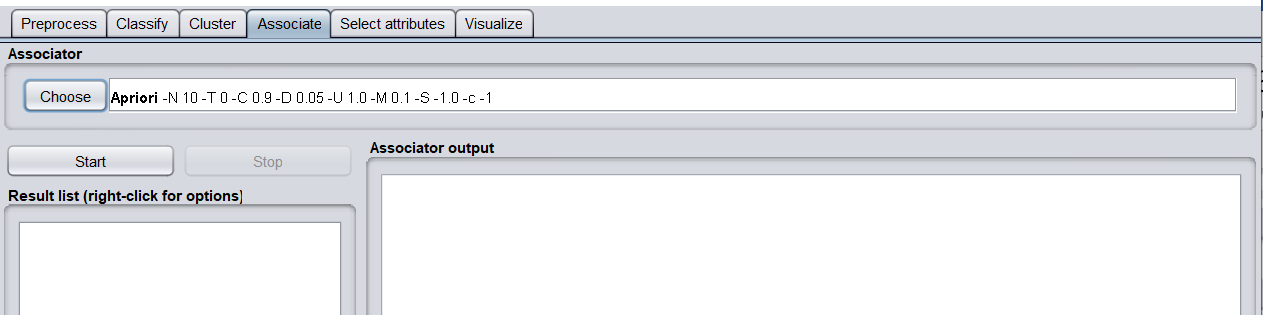
+ Cải thiện hoặc lượt bỏ hoăc tăng giảm chi phí hợp lý các dịch vụ nào để giảm bớt số lượng khách hàng rời bỏ công ty?

1. **Tiến hành thí nghiệm:**
2. **Mở tập dữ liệu Churn.txt trong Weka**

* 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 Norminal nên phải thêm “first-last” để biến đổi Numeric thành Norminal

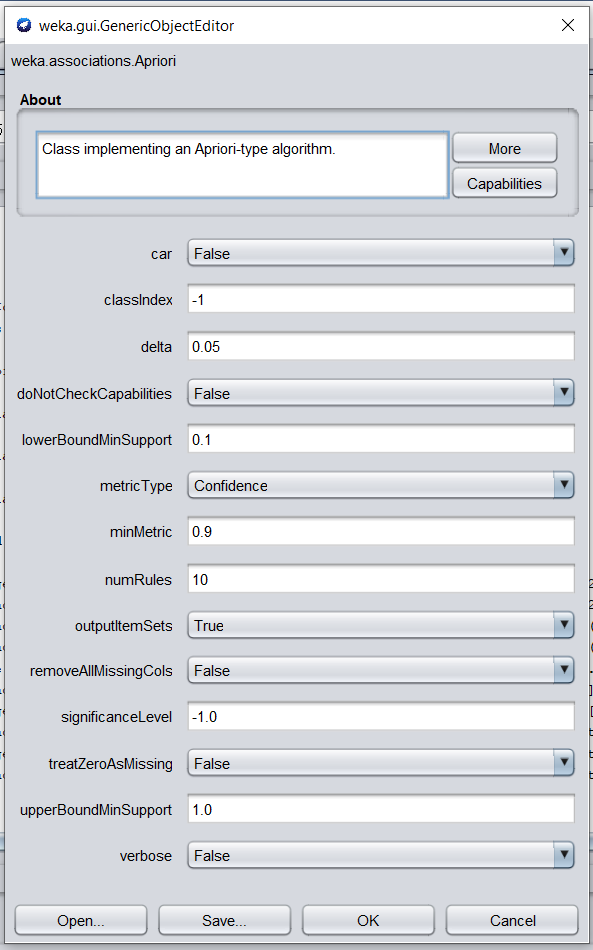


* Chọn tab Association Rules để tiếp hành phân tích dữ liệu

****

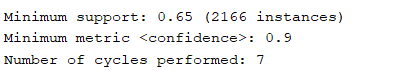
1. **Chạy thử nghiệm lần 1:**

* 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

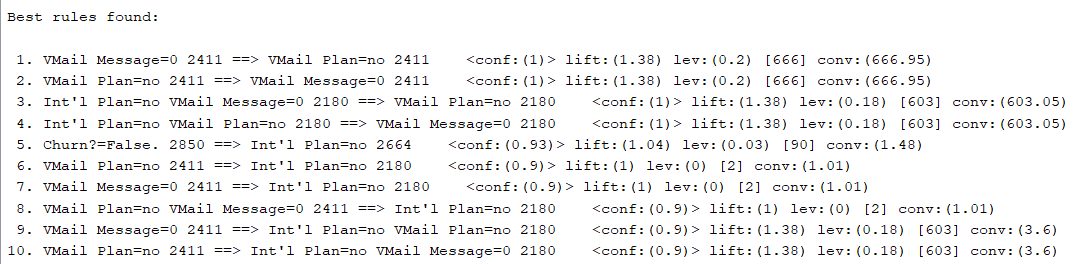
****

**+** Các Large Itemsets



* Trong Large itemsets L(2) ta thấy rằng VMail Plan=no VMail Message=0 2411(điều này đương nhiên), 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

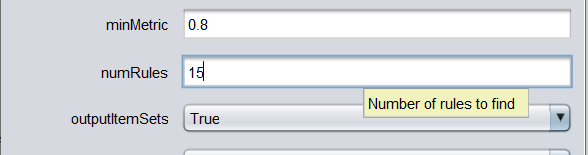
+ Các luật bao gồm:



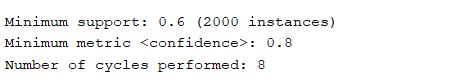
Ta chú ý đến các luật có độ tin cậy ( < <conf:(x)> > ) lớn và có liên quan đến thuộc tính mấu chốt Churn? Ta dễ dàng nhận thấy luật thứ 5 có độ tin cậy 93%, nhưng ta không tìm được một luật ở chiều ngược lại. *Tuy nhiên ta có thể kết thấy rõ ràng một điều khách hàng không có xu hướng rời bỏ công ty thì 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.*

1. **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



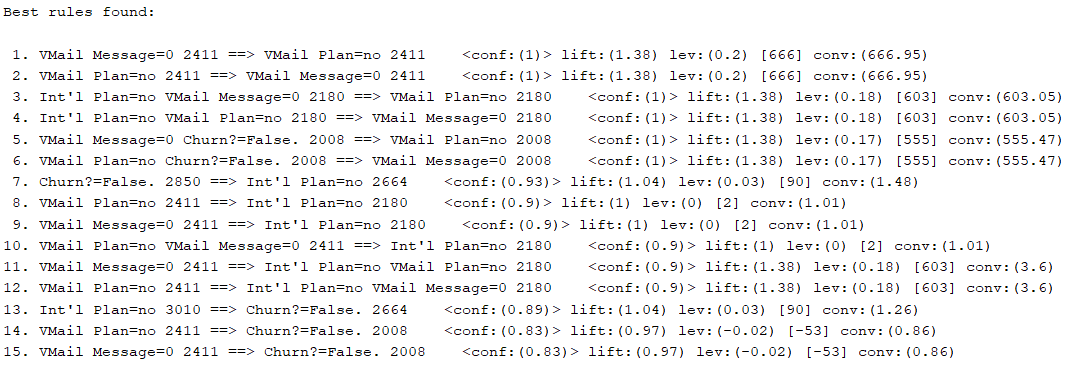
* Kết quả: Có 2000 mẫu đạt minsup = 0.6



* Các large itemsets



* 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:



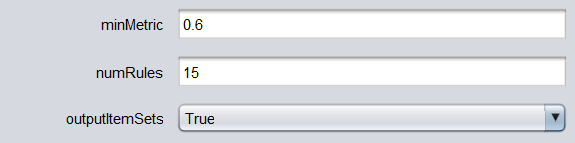
Ta chú ý đến các luật có độ tin cậy ( < <conf:(x)> > ) lớn và có liên quan đến thuộc tính mấu chốt Churn?

* Ta dễ dàng thấy được luật số 7 đã xuất hiện ở lần thí nghiệm trước
* Luật *Int'l\_Plan=no 3010 ==> Churn?=False. 2664 <conf:(0.89)> lift:(1.04) lev:(0.03) [90] 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 *VMail Plan=no 2411 ==> Churn?=False. 2008 <conf:(0.83)> lift:(0.97) lev:(-0.02) [-53] 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 *VMail Message=0 2411 ==> Churn?=False. 2008 <conf:(0.83)> lift:(0.97) lev:(-0.02) [-53] conv:(0.86).* Như đã giải thích 2 thuộc tính này được xem như tương đồng như đã giải thích ở trên.

*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.*

1. **Chạy thử nghiệm lần 3:**

* 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



* 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?

- Tuy nhiên ta nhận thấy được trong số khách hàng sử dụng gói Quốc tế (cụ thể là 3333 – 2664 = 669 mẫu) chưa được khái thác vấn đề liệu hó 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)*

- 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)*

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

*(\*) Tiền xử lí 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 Norminal có thể áp dụng Apriori*

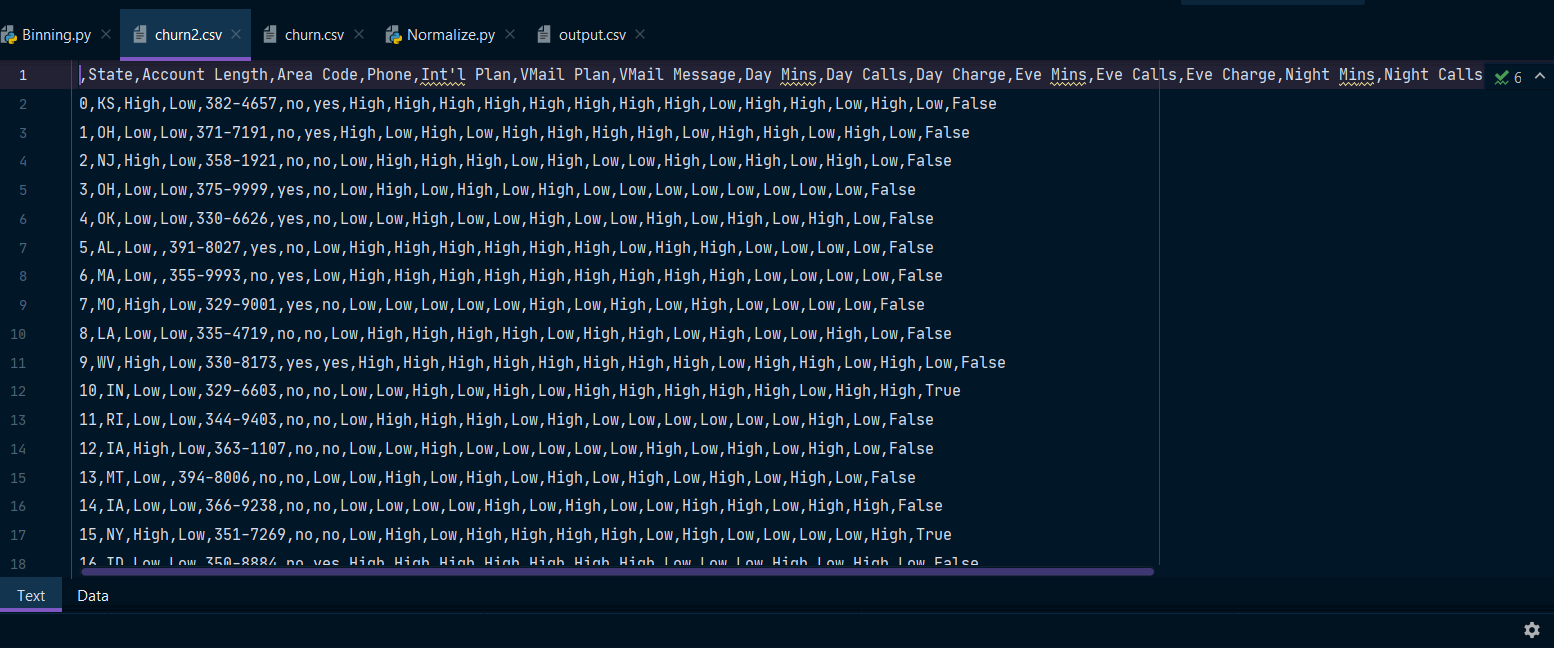
***(\*) Với phân loại {low, high}***

1. **Chạy thử nghiệm lần 4:**

* Ta thực hiện với dữ liệu đã Binning theo 2 cấp độ
* 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:

****

1. **Chạy thử nghiệm lần 5:**
2. **Chạy thử nghiệm lần 6:**

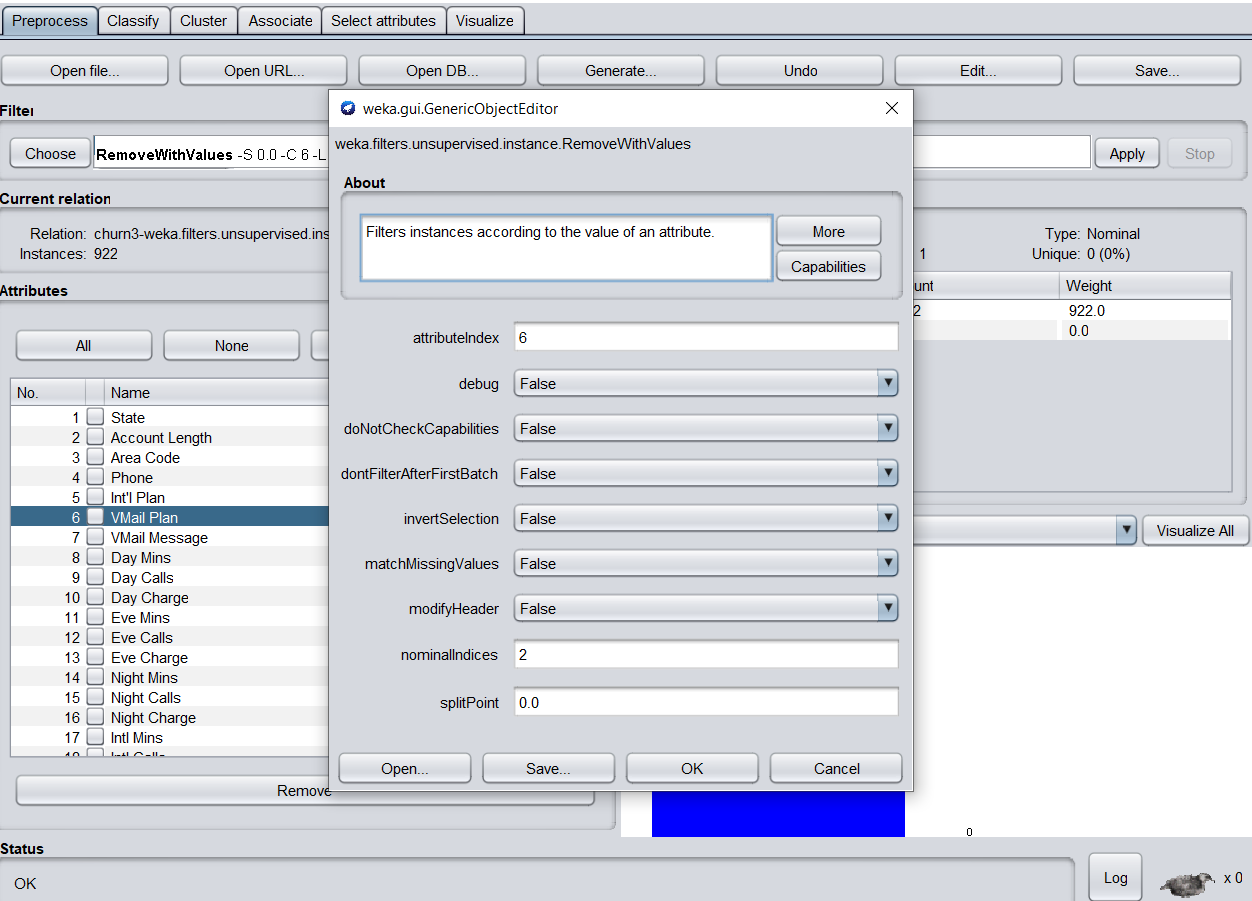
***(\*) Với phân loại {low, medium, high}***

1. **Chạy thử nghiệm lần 7:**
2. **Chạy thử nghiệm lần 8:**
3. **Chạy thử nghiệm lần 9:**

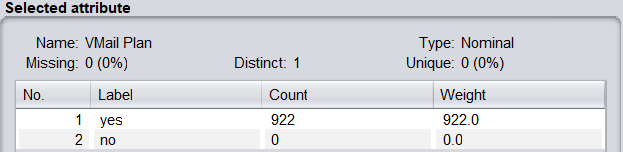
***(\*) Với phân loại {verylow, low, medium, high, veryhigh}***

1. **Chạy thử nghiệm lần 10:**
2. **Chạy thử nghiệm lần 11:**
3. **Chạy thử nghiệm lần 12:**
4. **Phân tích kết quả thí nghiệm:**
5. **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}

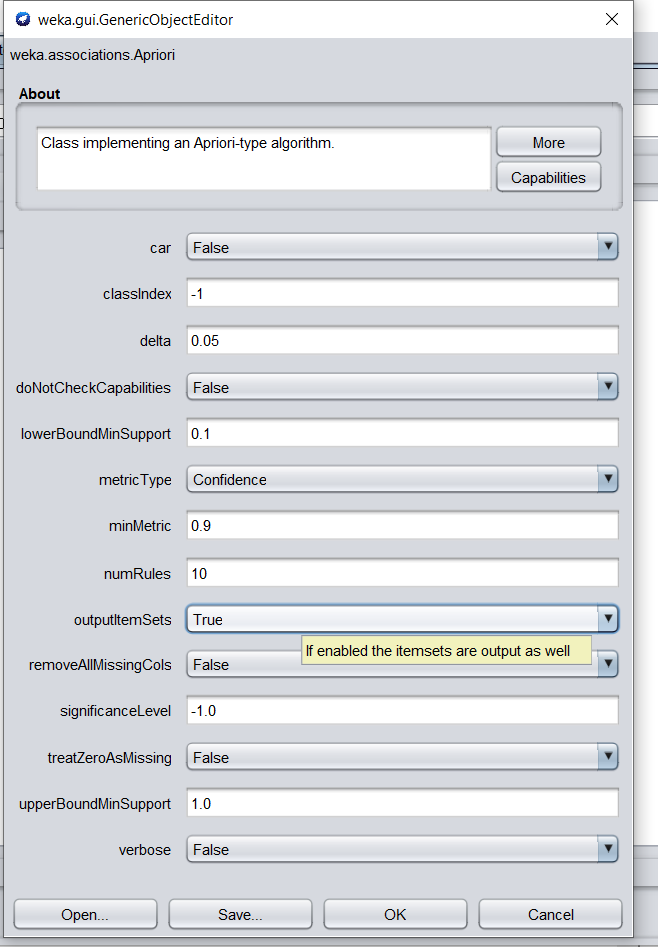


* Sau khi áp dụng bộ lọc, chỉ còn lại 922 mẫu dữ liệu có sử dụng thư thoại

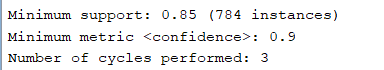


(\*) **Chạy thử nghiệm lần 1:**

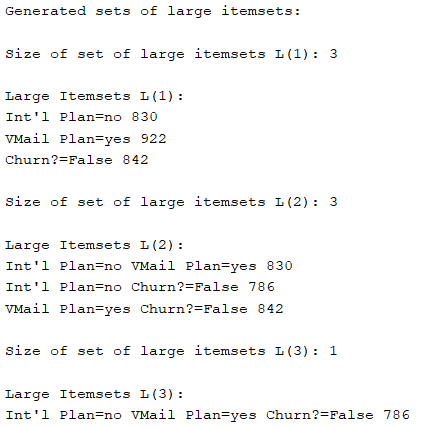
* Sử dụng các thông số mặc định của Weka ( *nếu muốn xem xét cái large Itemset thì phải chọn mode True tại outputItemSets* )



* Kết quả: Có 784 mẫu đạt minsup = 0.85



* Các large Itemsets:



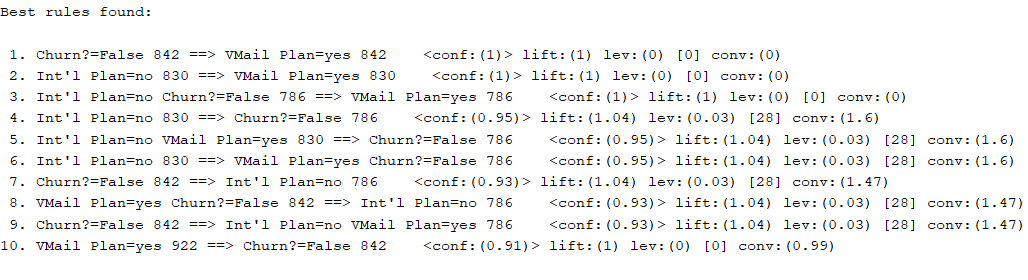
Cả L1, L2, L3 đều xuất hiện Vmail Plan=yes.

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:



Quan sát bộ luật ta chú ý các luật có thuộc tính mấu 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 cùng 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

+ Đến đây là đã có 2 suy đoán có xu hương trái ngược nhau:

* Khách hàng sử dụng gói Quốc tế thường có xu hướng rời bỏ công ty
* Khách hàng có sử dụng gói Thư thoại thường không có xu hướng rời bỏ công ty

(\*) **Chạy thử nghiệm lần 2:**

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



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



Các itemsets này cho thấy rằng khách hàng gọi điện vào buổi tối, ban đêm và gọi quốc tế 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 tuối và ban đêm, và còn ít sử dụng dịch vụ chăm sóc khách hàng

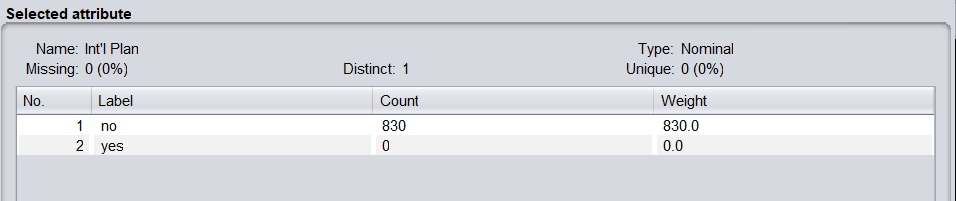
- Có 2 xu hướng đáng chú ý:

\* 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.

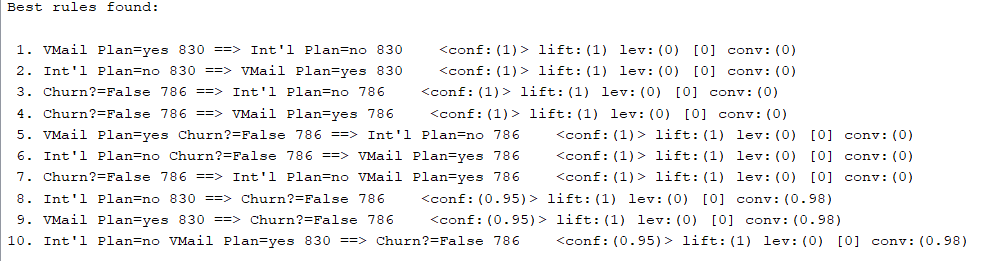
1. **Phân thích trường hợp “Khách hàng chỉ sử dụng gói thư thoại không sử dụng gói quốc tế”**

* 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



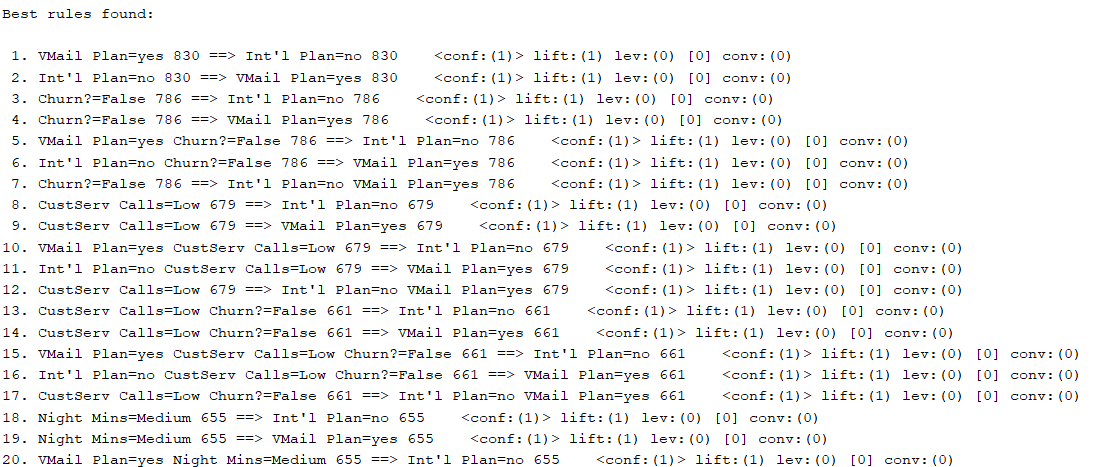
**(\*) Chạy thử nghiệm lần 1:**

* Sử dụng các thông số mắc định
* Các luật:

****

(\*) Chạy thử nghiệm lần 2:

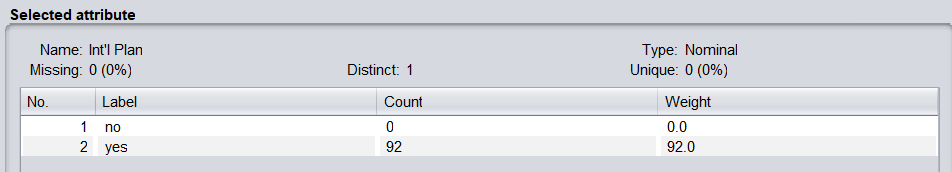
* Điều chỉnh số luật thành 20
* Các luật:



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

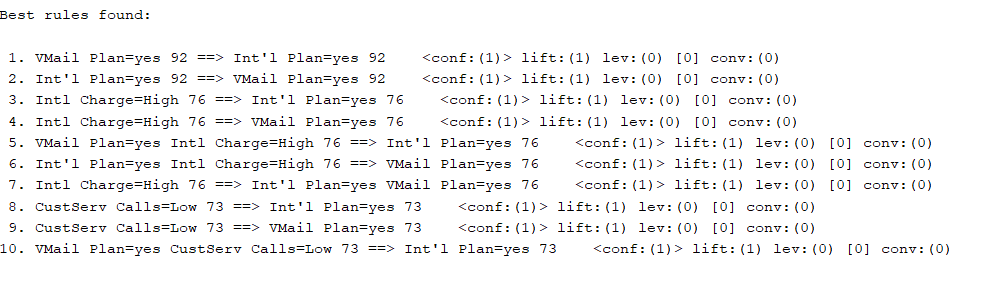
1. **Phân tích trường hợp “Khách hàng sử dụng gói thư thoại và gói quốc tế”**

* 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

****

**(\*) Chạy thử nghiệm lần 1:**

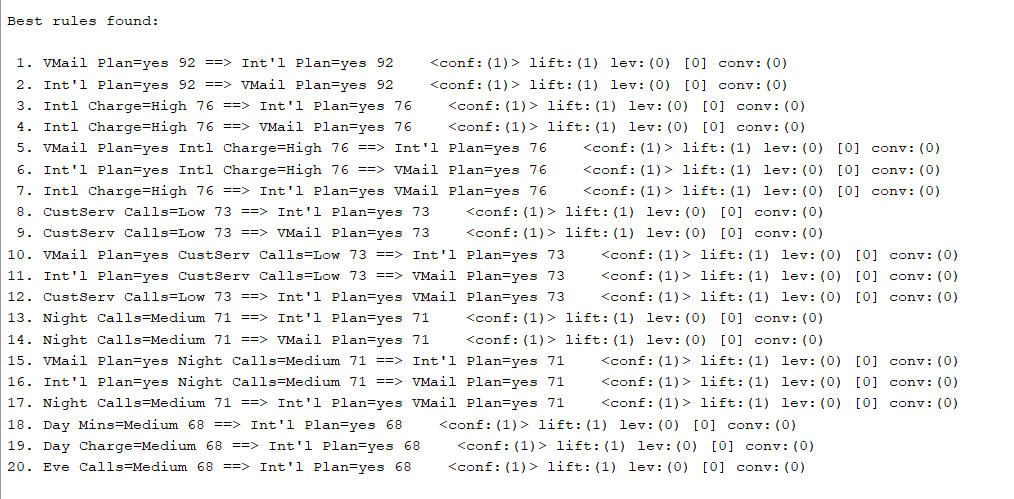
* Với thông số mặc định của Weka
* Kết quả

****

* *Chưa thể rút ra kết luận cho thuộc tính mấu chốt*

**(\*) Chạy thử nghiệm lần 2:**

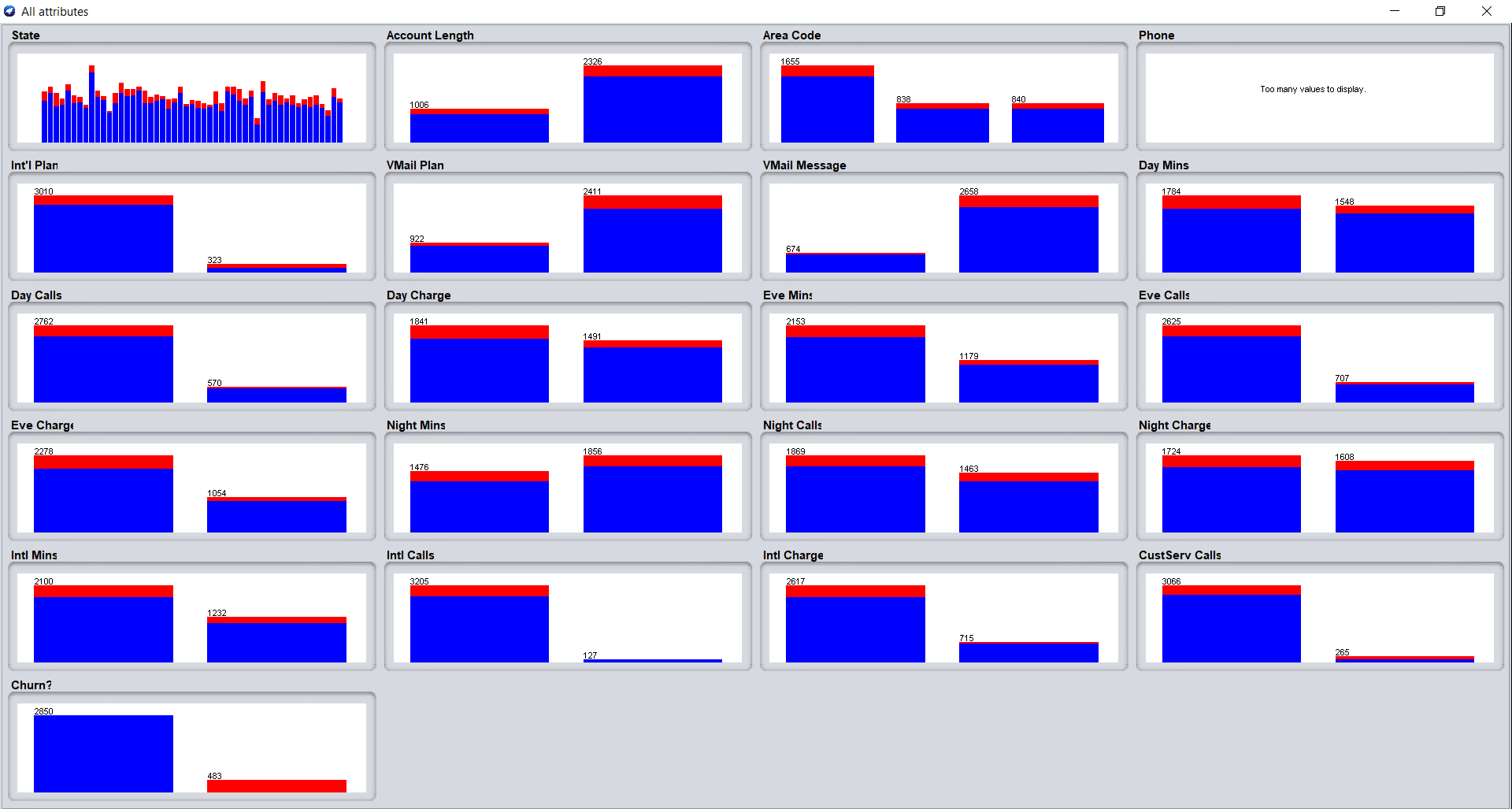
* Với số lượng luật tăng lên 20
* Kết quả:



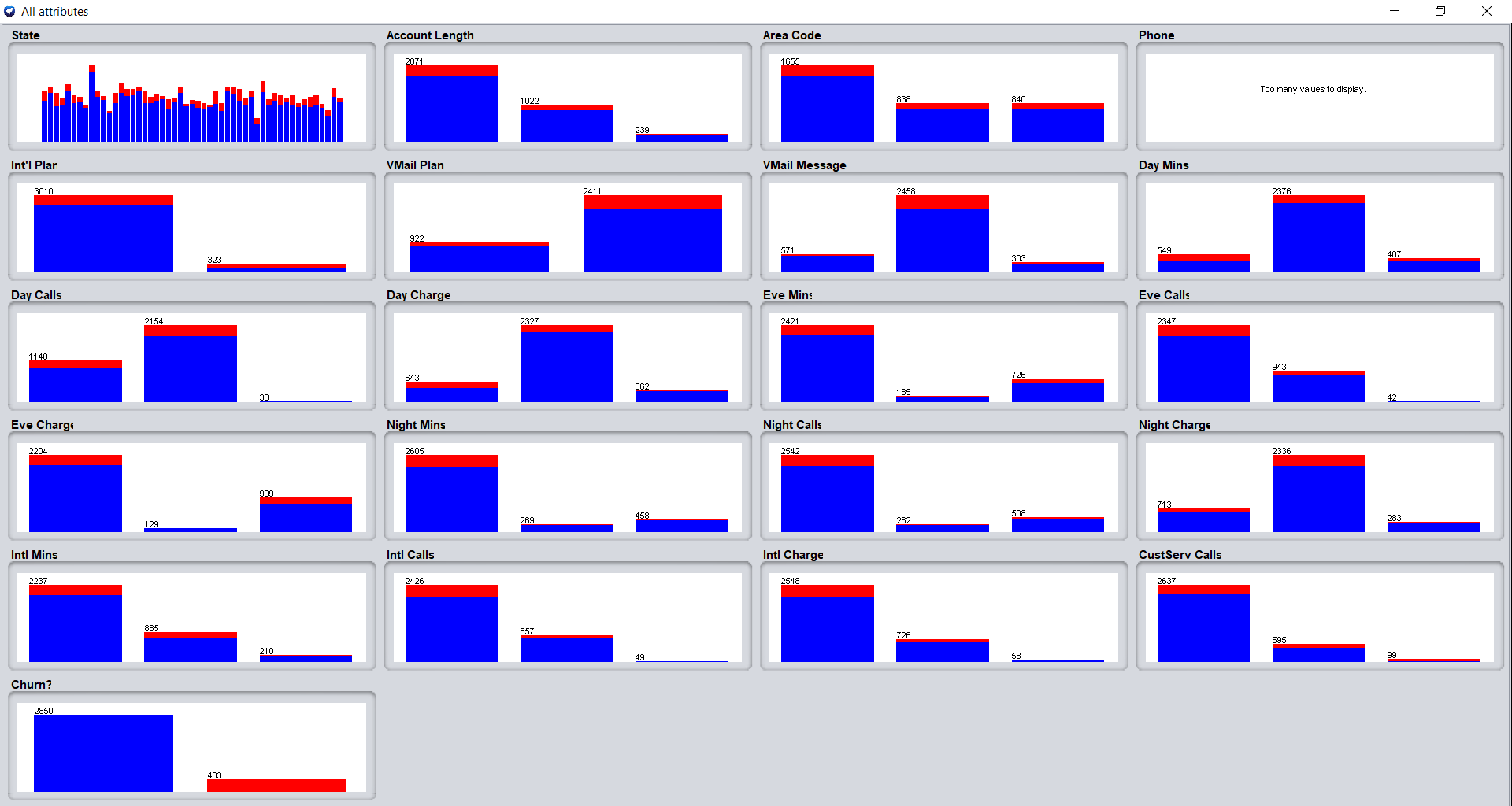
* *Chưa thể rút ra kết luận cho thuộc tính mấu chốt*

1. **Phân tích biểu đồ:**

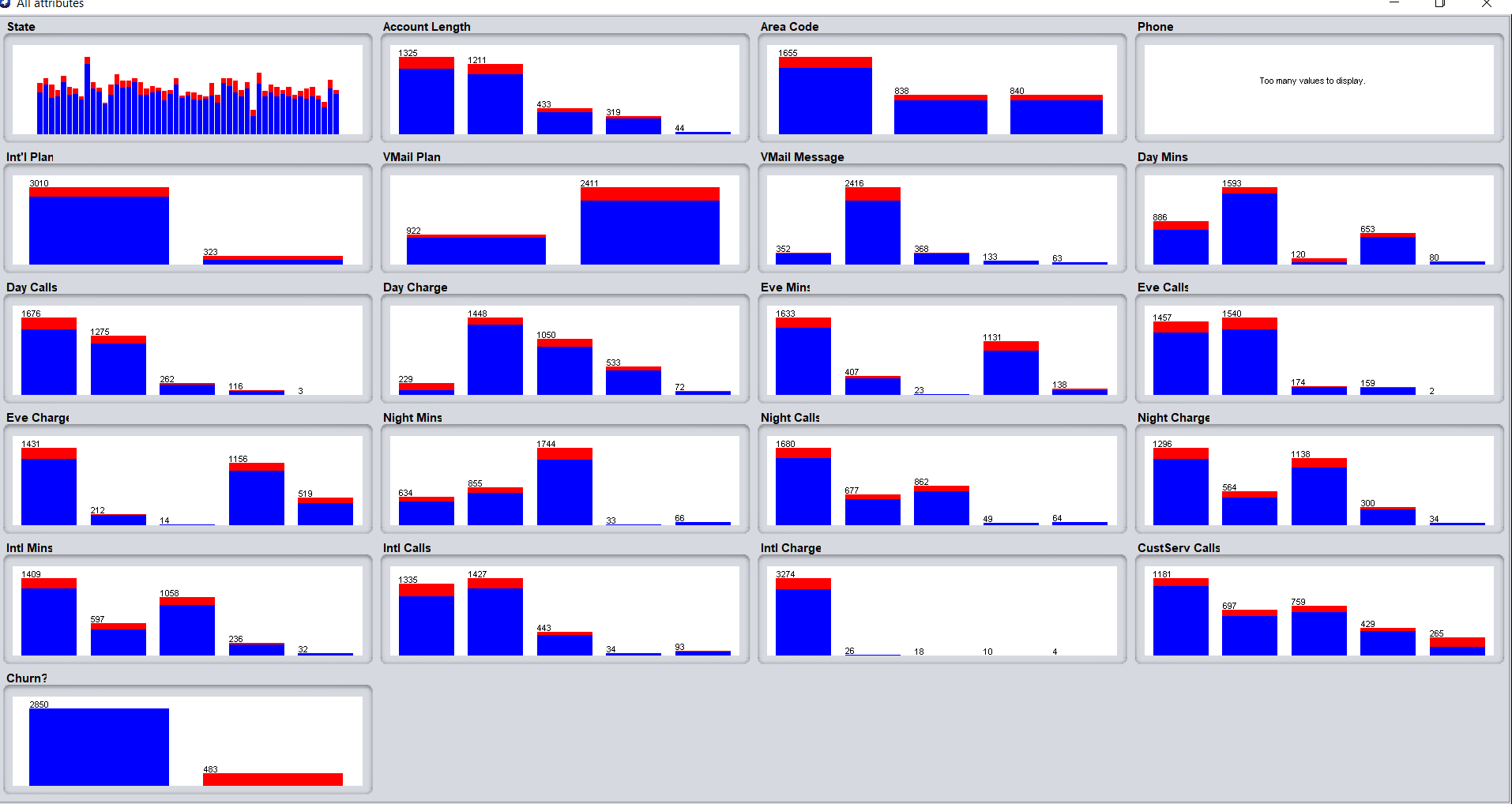
**(\*) Với phân cấp {low, high}**

****

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

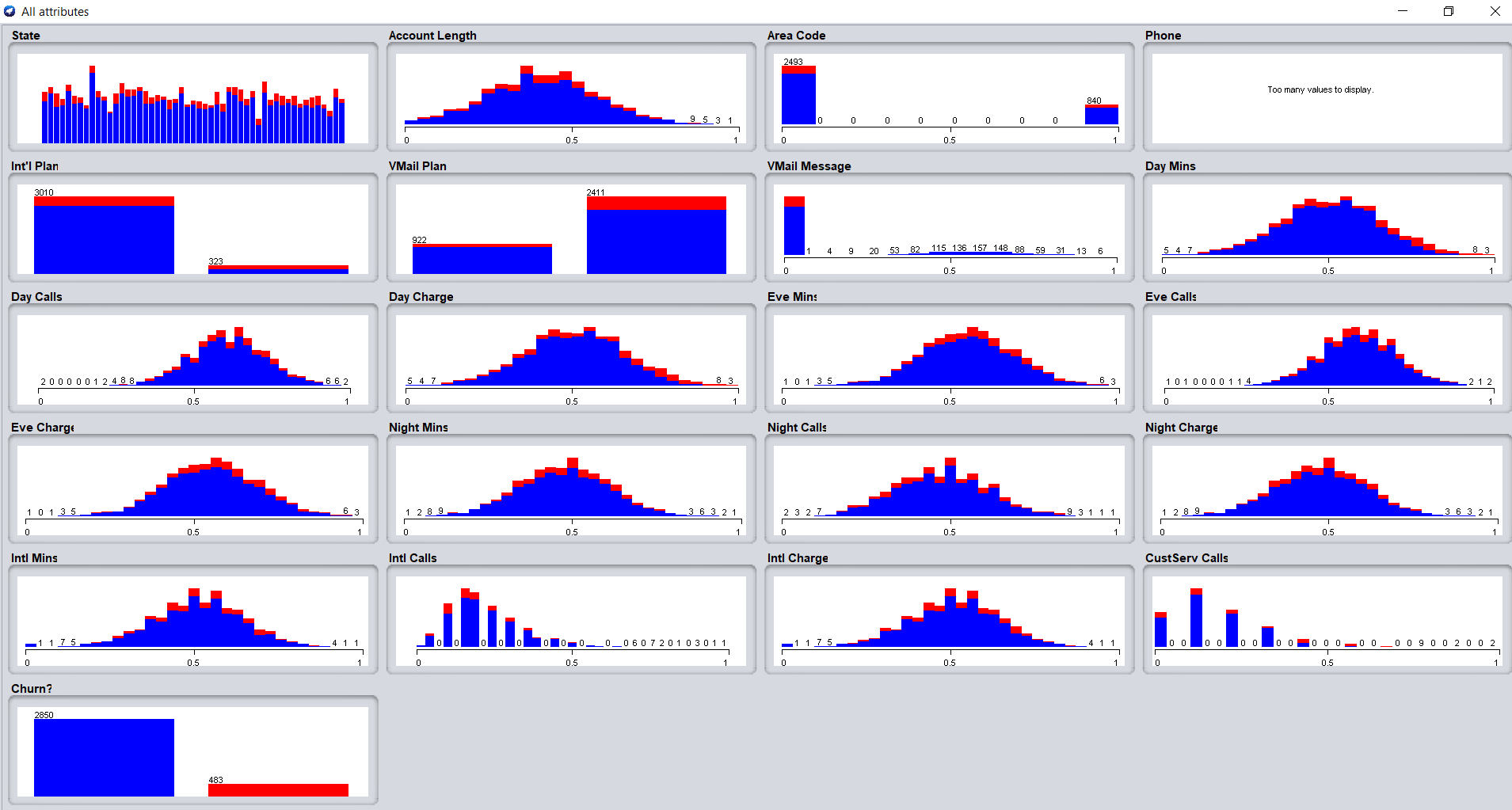
****

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

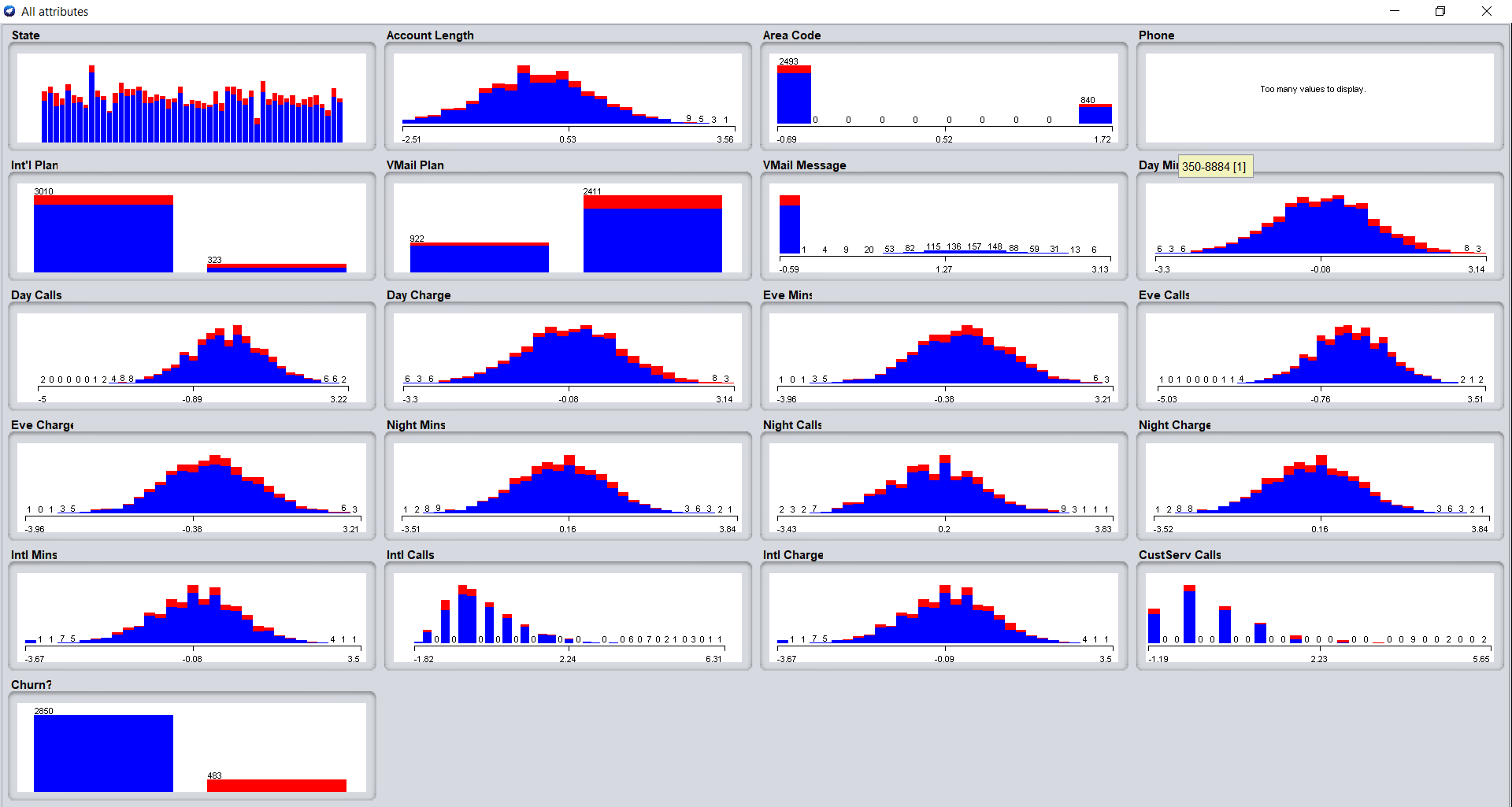
****

*=>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.

*(\*) 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 hướng đến việc khách hàng có rời bỏ công ty hay không.*

1. **SUMARY**