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

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**



**LAB 2 - Frequent itemsets and association rules**

**MÔN: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**| Giáo viên hướng dẫn |**

**Thầy: Nguyễn Ngọc Đức**

**Sinh viên thực hiện:**

**CAO TẤT CƯỜNG – 18120296**

**HÀ VĂN DUY – 18120339**

**Chuyên ngành: Khoa học máy tính**

Thành phố Hồ Chí Minh – 2020

# MỤC LỤC

[MỤC LỤC 2](#_Toc56328356)

[Phần I: Thông tin chung 3](#_Toc56328357)

[Phần II: Data 4](#_Toc56328358)

[2.1. Mô tả tập dữ liệu 4](#_Toc56328359)

[2.2. Các khái niệm phân cấp (concept hierarchies) 4](#_Toc56328360)

[2.2.1 Schema hierarchy (lược đồ phân cấp) 5](#_Toc56328361)

[2.2.2 Set-grouping hierarchy 5](#_Toc56328362)

[2.2.3 Operation-derived hierarchy 6](#_Toc56328363)

[2.2.4 Rule-based hierarchy 6](#_Toc56328364)

[Phần III: Code 8](#_Toc56328365)

[3.1. Tiền xử lý dữ liệu 8](#_Toc56328366)

[3.2.1. Quan sát tập dữ liệu 8](#_Toc56328367)

[3.2.2. Loại bỏ các thuộc tính không cần thiết hoặc có tương quan với nhau 8](#_Toc56328368)

[2.2.3. Rời rạc hóa các thuộc tính có kiểu numeric 12](#_Toc56328369)

[2.2.4 Tiền xử lý dữ liệu nâng cao đối với module Weka không có sẵn 15](#_Toc56328370)

[3.2. Khai thác luật kết hợp với thuật toán Apriori 15](#_Toc56328371)

[Phần IV: Experiments 19](#_Toc56328372)

[4.1. Tại sao phải phân tích dữ liệu 19](#_Toc56328373)

[4.2. Khai thác luật 19](#_Toc56328374)

[4.2.1. Khai thác luật dựa trên các hệ số khác nhau 19](#_Toc56328375)

[4.2.1. Thử nghiệm lại với góc nhìn khác của dữ liệu 21](#_Toc56328376)

[Phần V: Tóm tắt kết quả 28](#_Toc56328377)

[5.1. Đánh giá kết quả 28](#_Toc56328378)

[5.2. Tập luật tốt thu được 28](#_Toc56328379)

[5.3. Đánh giá nhóm 29](#_Toc56328380)

[5.3.1 Ưu điểm 29](#_Toc56328381)

[5.3.2 Nhược điểm 29](#_Toc56328382)

[PHỤ LỤC 30](#_Toc56328383)

[TÀI LIỆU THAM KHẢO 31](#_Toc56328384)

# Phần I: Thông tin chung

Danh sách thành viên:

* Cao Tất Cường 18120296
* Hà Văn Duy 18120339

Mục tiêu của bài tập: khai thác luật kết hợp trên dữ liệu với nhiều thuộc tính.

Cấu trúc thư mục nộp bài:

* File báo cáo: 18120296\_18120339.docx, 18120296\_18120339.pdf
* Dữ liệu đã qua tiền xử lý: churn\_preprocessing.arff, churn\_Set-grouping\_hierarchy.csv, churn\_set-grouping\_hierarchy.arff, churn\_intl-plan\_vmail-plan.arff, churn\_All-Mins.arff, churn\_All-Mins\_CustServCalls.arff
* Mã nguồn phần tiền xử lý nâng cao: Set-grouping\_hierarchy.ipynb

# Phần II: Data

## 2.1. Mô tả tập dữ liệu

Chuẩn bị: Tập dữ liệu *churn* (cũng được gọi là *attrition*), là một thuật ngữ để chỉ ra việc một khách hàng rời bỏ dịch vụ của một công ty, doanh nghiệp ban đầu để chuyển sang sử dụng dịch vụ một công ty khác [1].

Tập dữ liệu chứa 20 biến giá trị thông tin về 3333 khách hàng, cùng với lớp về việc khách hàng đó có rời bỏ dịch vụ của công ty hay không.

Các biến như sau:

* State: categorical, cho 50 bang và Đặc khu Columbia
* Account length: numeric, tài khoản đã hoạt động trong bao lâu
* Area code: categorical
* Phone number: bản chất là một đại diện cho ID khách hàng
* International Plan: dichotomous categorical, yes or no
* VoiceMail Plan: dichotomous categorical, yes or no
* Number of voice mail messages: numeric
* Total day minutes: numeric, sô phút khách hàng sử dụng dịch vụ trong ngày
* Total day calls: numeric, sô cuộc gọi khách hàng sử dụng dịch vụ trong ngày
* Total day charge: numeric, phí dựa vào Total day minutes và Total day calls
* Total evening minutes: numeric, sô phút khách hàng sử dụng dịch vụ trong ngày
* Total evening calls: numeric, sô cuộc gọi khách hàng sử dụng dịch vụ trong buổi tối
* Total evening charge: numeric, phí dựa vào Total evening minutes và Total evening calls
* Total night minutes: numeric, sô phút khách hàng sử dụng dịch vụ trong khuya
* Total night calls: numeric, sô cuộc gọi khách hàng sử dụng dịch vụ trong khuya
* Total night charge: numeric, phí dựa vào Total night minutes và Total night calls
* Total international minutes: numeric, sô phút khách hàng sử dụng dịch vụ gọi quốc tế
* Total international calls: numeric, sô cuộc gọi khách hàng sử dụng dịch vụ gọi quốc tế Total international charge: numeric, phí dựa vào Total international minutes và Total international calls

## 2.2. Các khái niệm phân cấp (concept hierarchies)

Phân cấp được mô tả như là một chuỗi ánh xạ từ một tập hợp các khái niệm cấp thấp đến một tập các khái niệm cấp cao hơn, đồng thời cũng tổng quát hơn.

Hệ thống phân cấp khái niệm có thể được sử dụng để giảm dữ liệu bằng cách thu thập và thay thế các khái niệm cấp thấp (ví dụ như tuổi của một người, là một biến liên tục) bằng các khái niệm cấp cao hơn (còn trẻ, trung niên, người già..., là một thuộc tính rời rạc).

Với khái niệm phân cấp, ta có thể linh hoạt quan sát dữ liệu dưới nhiều cấp bậc, mức độ tổng quát khác nhau. Hơn nữa, khi tiến hành khai thác dữ liệu trên một tập giảm (tối thiểu thuộc tính cấp thấp, quan sát ở các thuộc tính cấp cao), số thuộc tính sẽ ít lại, điều đó khiến cho quá trình khai thác sẽ diễn ra nhanh hơn, ưu việt hơn khi hoạt động trên các tập dữ liệu lớn.

Khái niệm phân cấp có thể được chia ra làm 4 loại cơ bản: schema, set-grouping, operation\_derived, rule-based concept hierarchies.

#### 2.2.1 Schema hierarchy (lược đồ phân cấp)

Loại phân cấp này được hình thành ở cấp lược đồ bằng cách xác định thứ tự từng phần để phản ánh mối quan hệ giữa các thuộc tính trong cơ sở dữ liệu­. Ví dụ, một bộ các thuộc tính gồm *số nhà, đường, phường, quận* sẽ tạo thành một lược đồ phân cấp như sau:

số nhà < đường < phường < quận

Trong ví dụ trên, các thuộc tính ở bên trái sẽ có cấp bậc bé hơn các thuộc tính phía bên phải của nó. Với một địa chỉ cụ thể, ví dụ “227 Nguyễn Văn Cừ, phường 4, Quận 5”, ta thấy dữ liệu này đã được biểu diễn theo thứ tự dưới dạng lược đồ phân cấp rồi, vì vậy không cần phải có các ràng buộc nào khác giữa các thuộc tính của tập dữ liệu đối với thuộc tính địa chỉ nữa.

Đối với tập dữ liệu *churn*, ta có thể dùng lược đồ phân cấp để áp dụng lên các thuộc tính *Day Mins, Eve Mins, Night Mins*… bằng cách chia miền giá trị của nó ra làm các khoảng rời rạc, sau đó gán nhãn cho mỗi khoảng rời rạc đó để tạo ra phân cấp khái niệm.

Ví dụ, trong *Day Mins* (số phút gọi ban ngày), ta chia ra làm các khoảng (tương ứng với mỗi khoảng là nhãn):

Level 0: [-Inf .. 50]: ít

Level 1: [51 ... 150]: trung bình

Level 2: [151 ... 250]: nhiều

Level 3: [251 ... Inf]: rất nhiều

Khi đó, lược đồ phân cấp khái niệm sẽ được biểu diễn theo thứ tự sau:

Day Mins (phút): ít < trung bình < nhiều < rất nhiều

#### 2.2.2 Set-grouping hierarchy

Loại phân cấp này được hình thành bằng cách định nghĩa và gom nhóm các nhóm giá trị của thuộc tính có cùng mối quan hệ lại với nhau. Đối với mỗi nhóm, ta vẫn có thể xây dựng thuộc tính phân cấp để dễ dàng quan sát và khai thác dữ liệu hơn.

Việc gom nhóm này được dùng để tinh chỉnh hoặc mở rộng lược đồ phân cấp đã được định nghĩa trước đó.

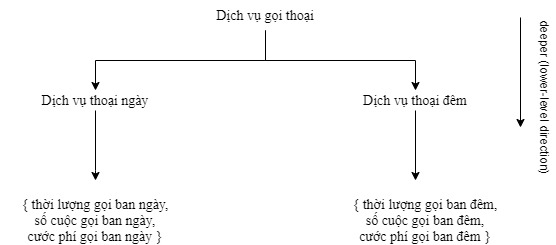
Ví dụ, đối với tập dữ liệu *churn* có chứa các thuộc tính *thời lượng gọi ban ngày, thời lượng gọi ban đêm, số cuộc gọi ban ngày, số cuộc gọi ban đêm, cước phí ban ngày, cước phí ban đêm,* ta có thể gom thành các nhóm như sau:

{ thời lượng gọi ban ngày, số cuộc gọi ban ngày, cước phí ban ngày} > dịch vụ thoại ngày

{ thời lượng gọi ban đêm, số cuộc gọi ban đêm, cước phí ban đêm} > dịch vụ thoại đêm

{ dịch vụ thoại ngày, dịch vụ thoại đêm } > dịch vụ gọi thoại

Khi đó, khái niệm phân cấp sẽ được định nghĩa như sau:



#### 2.2.3 Operation-derived hierarchy

Loại phân cấp này được hình thành dựa trên hoạt động được chỉ định bởi người dùng, các chuyên gia hay hệ thống khai thác dữ liệu. Các thủ tục này có thể bao gồm việc giải mã thông tin, mã hóa và trích xuất các dữ liệu phức tạp từ một cụm trong tập dữ liệu cho trước.

Ví dụ: Cấu trúc một địa chỉ email: [username@department.university.hostname.root](mailto:username@department.university.hostname.root)

Ví dụ cụ thể: 12345678@student.hcmus.edu.vn

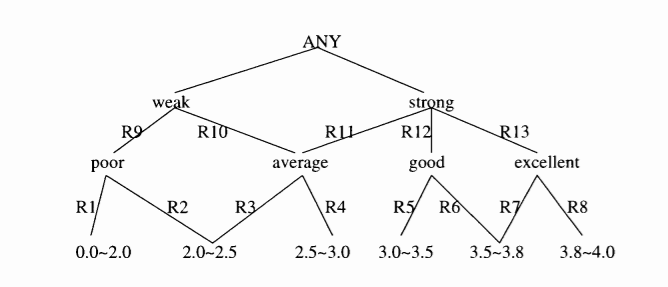
Như vậy, ta có thể trích xuất ra được trình tự các cấp bậc tăng dần như sau: username > department > university > hostname > root.

Đối với tập dữ liệu *churn,* loại hình phân cấp này chưa được định nghĩa bởi chuyên gia hay hệ thống khai thác dữ liệu, vì thế các thuộc tính trong tập đều là tối giản, không thể giải mã và trích xuất được nữa.

#### 2.2.4 Rule-based hierarchy

Đây là một hệ thống phân cấp dựa trên quy tắc toàn bộ hoặc một phần hệ thống phân cấp khái niệm được xác định bởi một bộ quy tắc và được đánh giá tự động dựa trên những dữ liệu hoặc hệ thống sẵn có. Với cách định nghĩa này, mỗi khái niệm sẽ đều có thể được khái quát hóa lên một cấp bậc cao hơn tương ứng khái quát hơn thông qua một điều kiện cụ thể bất kì.

Ví dụ, ta xét việc đánh giá kết quả học tập của một sinh viên đại học thông qua điểm tổng kết (GPA) thang điểm 4. Nếu điểm GPA của sinh viên đó là 3.6 thì ta có thể khái quát ở mức khái niệm lên tức là sinh viên đó thuộc nhóm *suất sắc.* Tuy nhiên, trong một vài trường hợp ta cần phải khái quát hóa một giá trị dựa trên cả một điều kiện cụ thể khác nữa. Cùng là điểm GPA 3.6, sinh viên đó chỉ có thể thuộc nhóm *tốt* nếu là sinh viên tốt nghiệp sau đại học, thay vì *xuất sắc* nếu tốt nghiệp đại học.



Hình trên biểu diễn một hệ thống phân cấp khái niệm dựa trên quy tắc nói về cách đánh giá kết quả của một sinh viên thông qua điểm GPA [2].

Ta xây dựng được một số luật phổ biến dựa trên phân cấp khái niệm trên như sau:

R1: {0.0 ~ 2.0} > Yếu

R2: {2.0 ~ 2.5} và tốt nghiệp sau đại học > Yếu

R3: {2.0 ~ 2.5} và tốt nghiệp đại học > Trung bình

...

Đối với tập dữ liệu *churn,* xét dịch vụ tin nhắn âm thanh (Voice Mail). Ta có thể phân cấp khái niệm VMail Message mô tả tần suất sử dụng dịch vụ tin nhắn âm thanh dựa trên một vài quy tắc như sau:

R1: *VMail Plan* = “no” > không dùng

R2: {0 ~ 25} và *Vmail Plan* = “yes” > dùng ít

R3: {25 trở lên} và *Vmail Plan* = “yes” > dùng nhiều

Ta có thể áp dụng quy tắc đó cho dịch vụ *Int'l Plan* để sau cùng cho ra được một tập các thuộc tính mới mang tính khái quát hơn trước khi đem tập dữ liệu đi khai thác.

Như vậy, một hệ thống phân cấp khái niệm dựa trên quy tắc sẽ được xác định bởi một tập các quy tắc mà trong đó việc đánh giá nó thường liên quan đến dữ liệu trong tập dữ liệu.

# Phần III: Code

## 3.1. Tiền xử lý dữ liệu

#### 3.2.1. Quan sát tập dữ liệu

Trong tập dữ liệu quan sát thấy 3333 mẫu dữ liệu, 21 thuộc tính và 0% bị thiếu giá trị nên sẽ bỏ qua bước xử lý giá trị bị thiếu. Ta cần loại bỏ các thuộc tính không cần thiết hoặc tương quan với nhau và rời rạc hóa những thuộc tính có kiểu numeric.

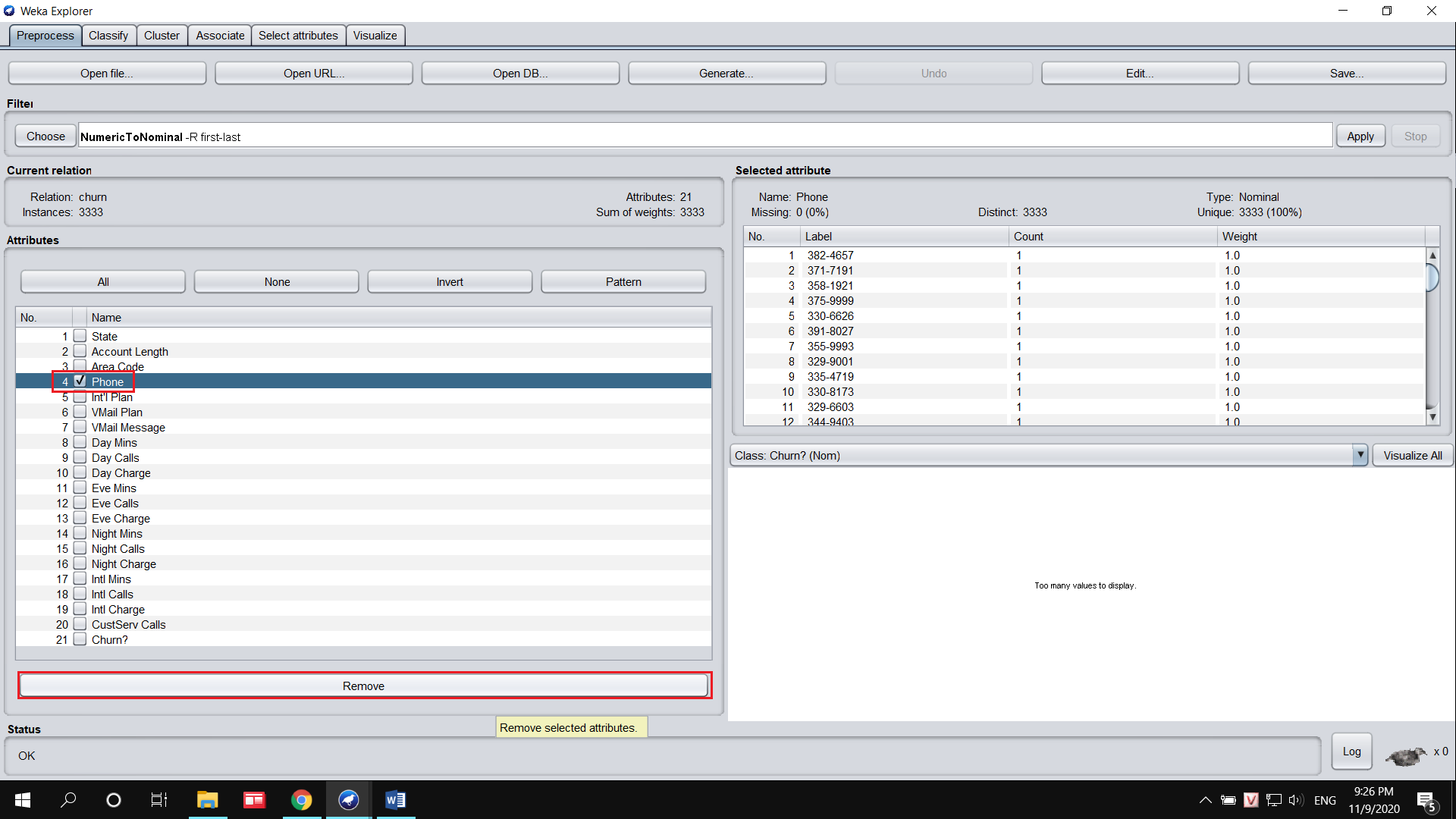
#### 3.2.2. Loại bỏ các thuộc tính không cần thiết hoặc có tương quan với nhau

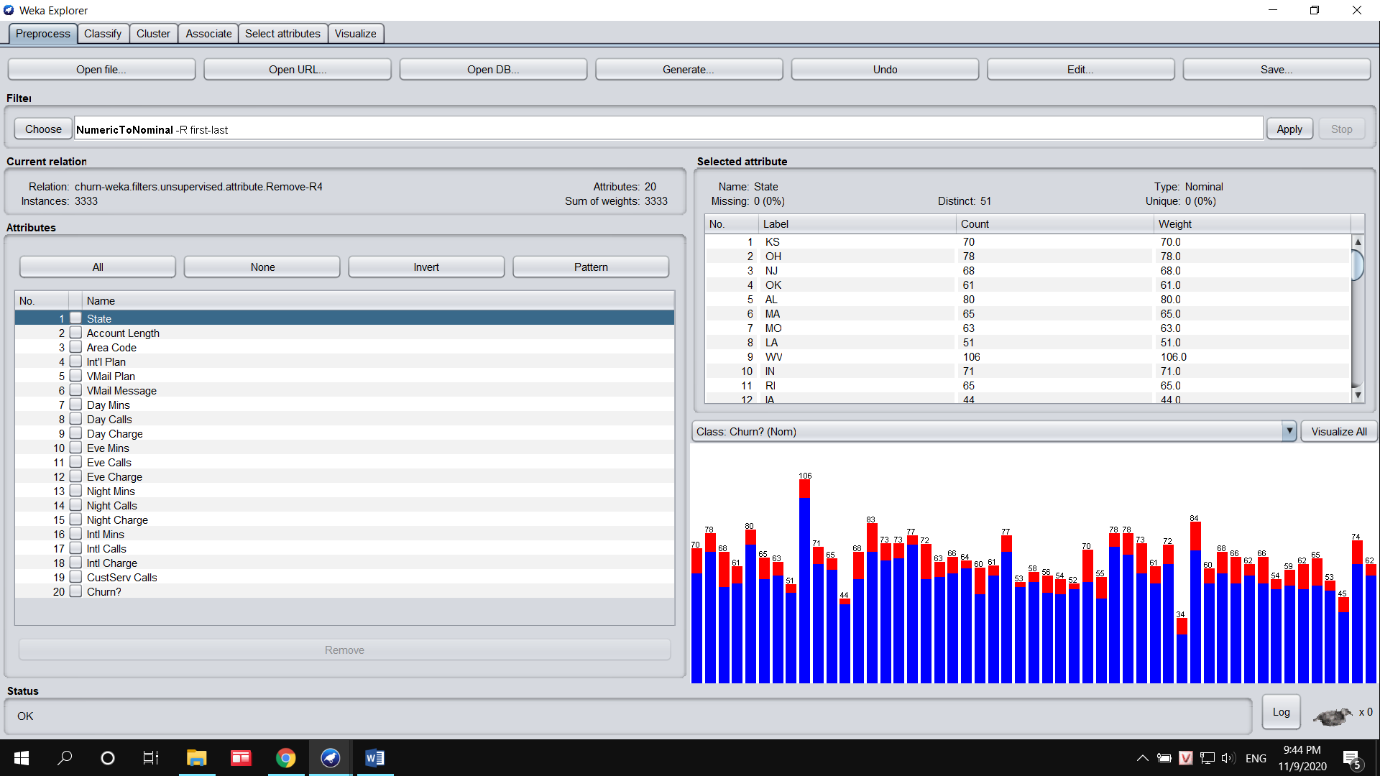
**Mục đích**: Chỉ chọn những thuộc tính cần thiết. Tìm ra tập thuộc tính con nhỏ nhất sao cho phân phối xác suất kết quả của các lớp dữ liệu càng gần với phân phối gốc sử dụng toàn bộ thuộc tính càng tốt. Loại bỏ bớt các thuộc tính không liên quan (*Phone*), bị mâu thuẫn (*State, Area Code*) hoặc các thuộc tính tương đồng hoàn hảo với nhau ta có thể loại bỏ 1 trong 2.

**Tác dụng**: Kích thước của không gian tập dữ liệu được giảm xuống, do đó các thuật toán khai thác dữ liệu có thể tìm ra giải pháp tối ưu toàn bộ tập dữ liệu một cách hiệu quả hơn.

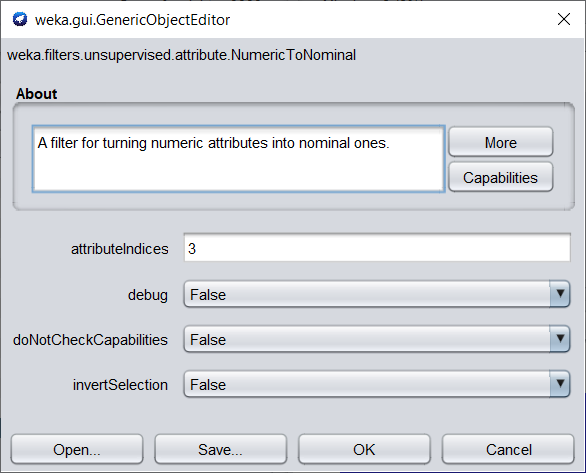
Ta thấy mỗi khách hàng được xác định duy nhất bởi thuộc tính *Phone* nên cần loại bỏ thuộc tính này trước khi thực hiện các bước tiếp theo trong Association Rule Mining.

Cách xóa thuộc tính trong weka ta chọn *Phone* rồi nhấn Remove.

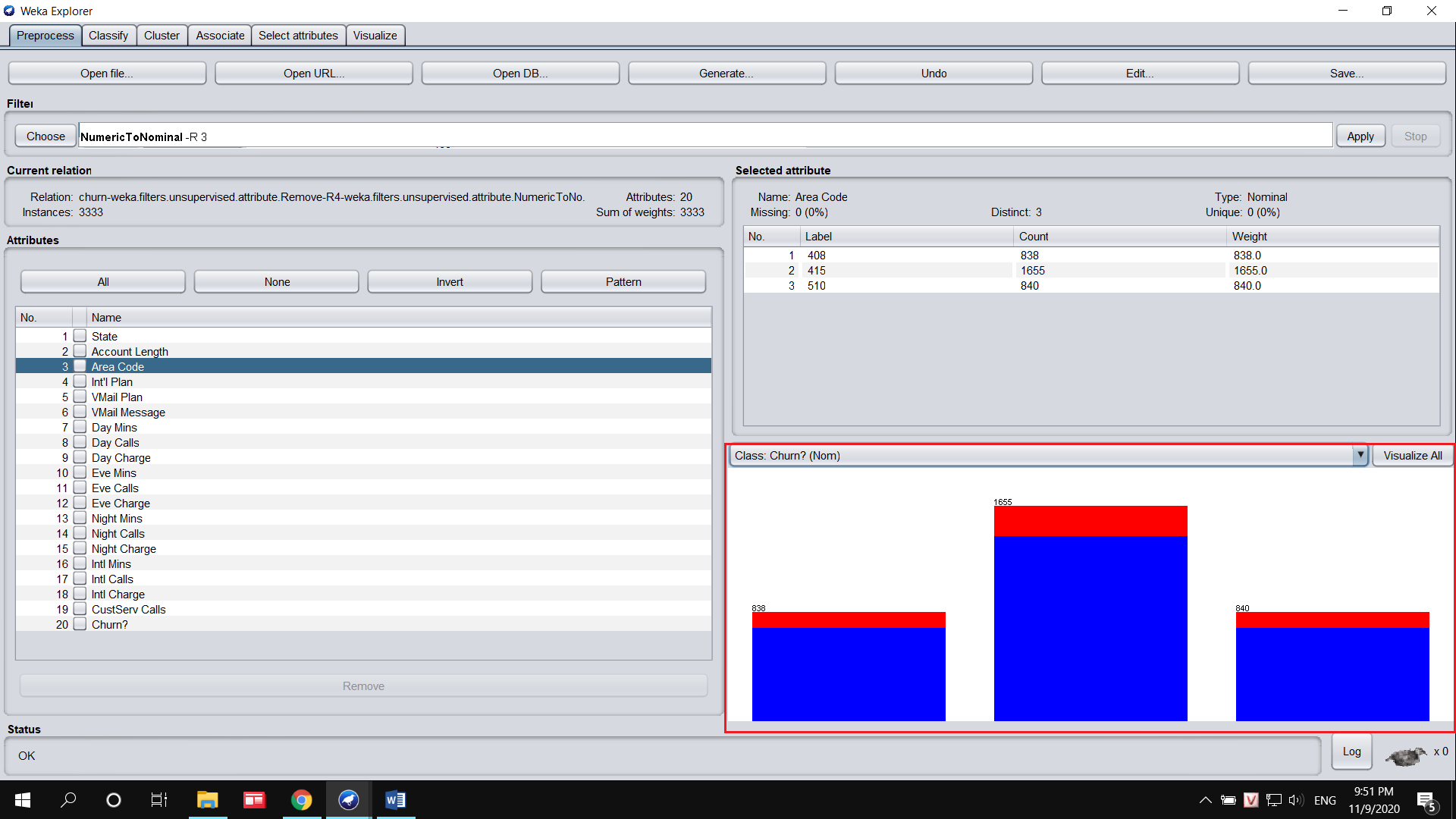


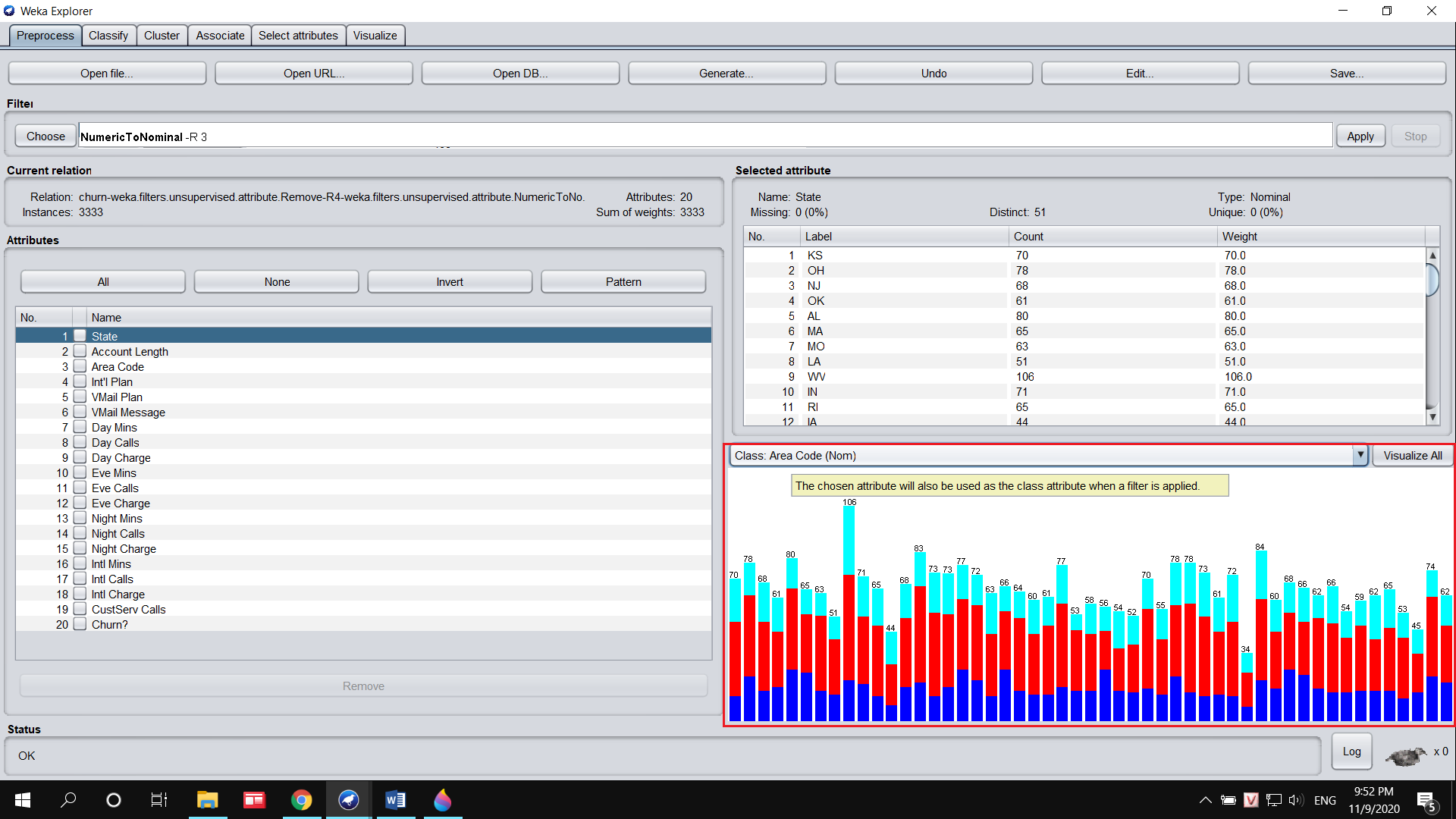
Kết quả của tập dữ liệu sau khi xóa *Phone*:

Mặc dù *Area Code* là kiểu numeric chúng cũng có thể được sử dụng như các biến phân loại, vì chúng có thể phân loại khách hàng. Nhưng thực tế là *Area Code* chỉ chứa ba giá trị khác nhau cho tất cả các bản ghi: 408, 415, 510. Ta có thể chuyển *Area Code* sang kiểu nominal để thấy rõ bằng cách sau:

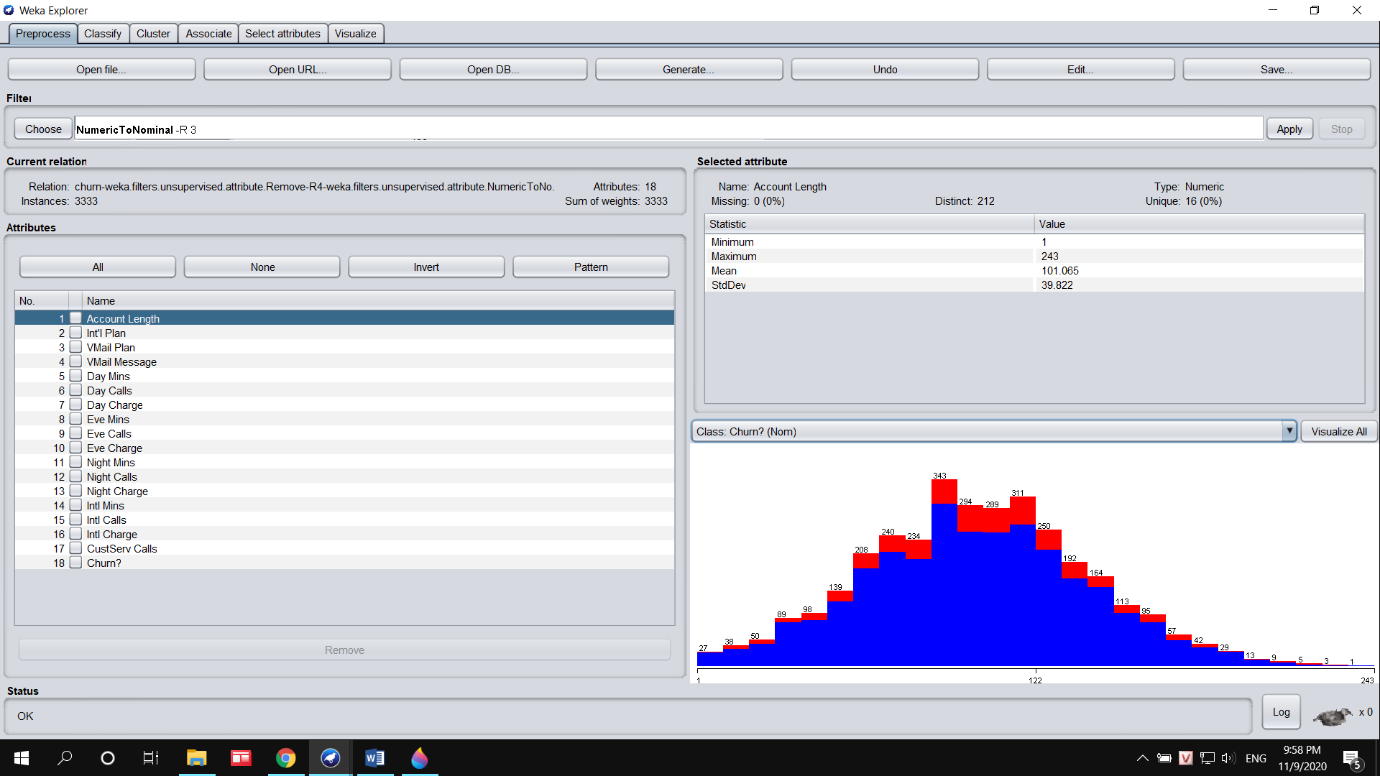
*Trong panel Filter > Choose  > filters >unsupervised > attribute > NumericToNominal*

attributeIndices: 3 (*Area Code*)

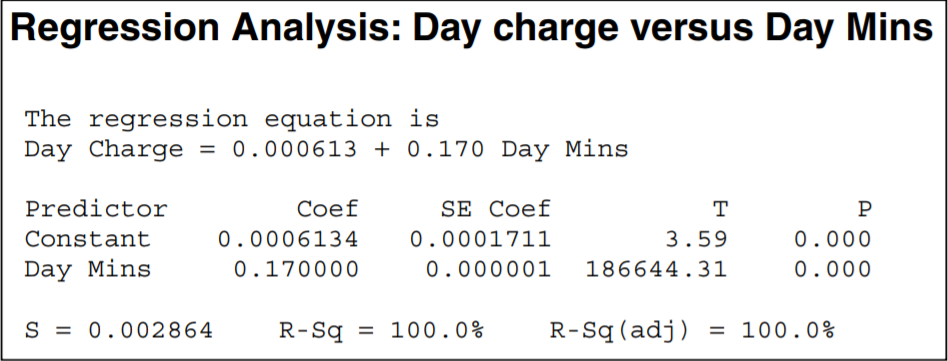
Thuộc tính *Area Code* ở dạng nominal

Sau khi chuyển sang nominal ta thấy cross tabulation giữa *State* và *Area Code* ba mã vùng dường như được phân phối trên tất cả các tiểu bang.

Có thể các chuyên gia tên miền có thể giải thích về điều này, nhưng cũng có thể trường chỉ chứa dữ liệu xấu. Do đó, không đưa nó vào làm đầu vào cho các mô hình khai thác dữ liệu trong giai đoạn tiếp theo. Nên ta sẽ xóa thuộc tính *State* và *Area Code.*

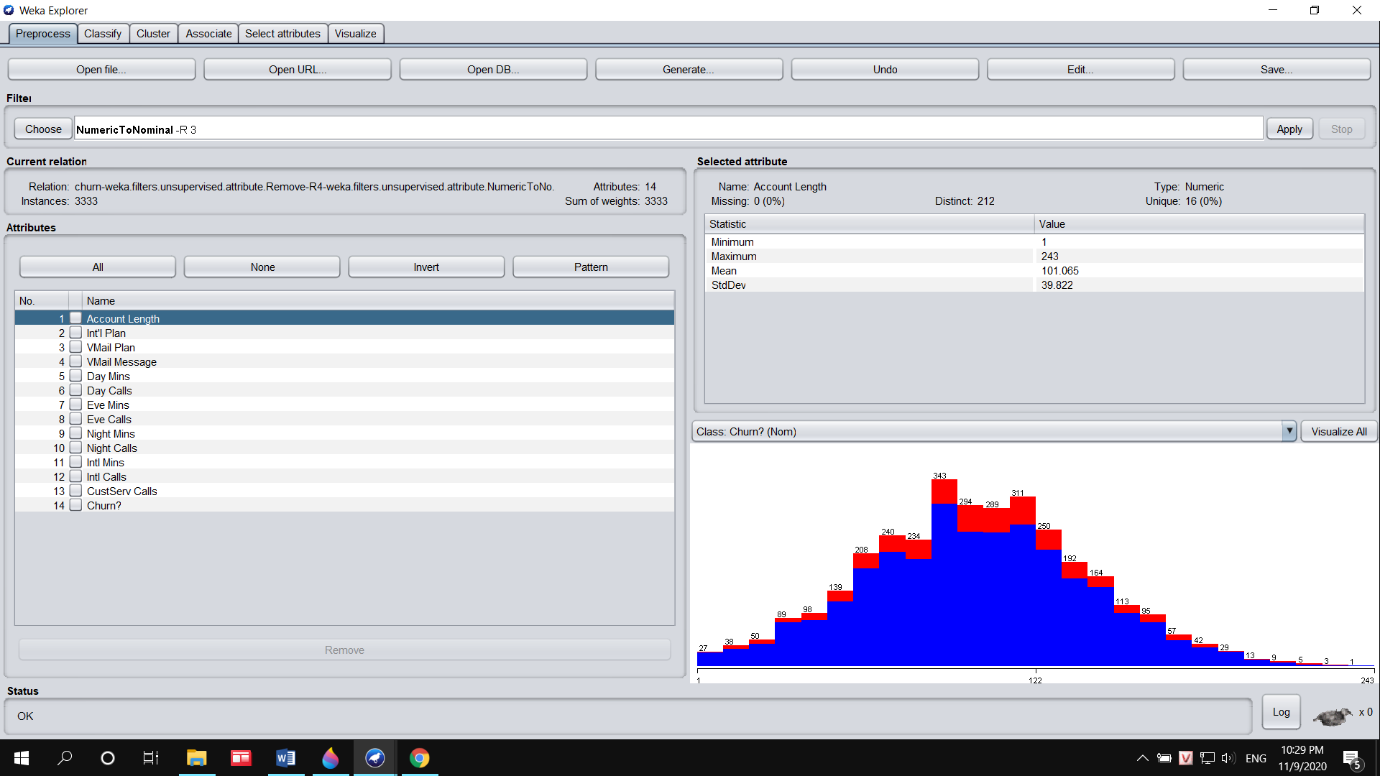
Kết quả sau tập thuộc tính khi xóa 2 thuộc tính trên:

Phân tích hồi quy: Day Charge so với Day Mins cho kết quả như sau:



Ta thấy *Day Charge* = 0.000613 + 0.17 \* *Day Mins* đây là một hàm tuyến tính nên ta kết luận *Day Charge* và *Day Mins* tương quan hoàn hảo với nhau nên ta có thể loại bỏ 1 trong 2 thuộc tính đó. Ta sẽ loại bỏ thuộc tính *Day Charge.*

Tương tự, có các cặp (Eve Charge, Eve Mins), (Night Charge, Night Mins), (Intl Charge, Intl Mins) tương quan hoàn hảo nên loại bỏ thêm các thuộc tính Eve Charge, Night Charge, Intl Charge.

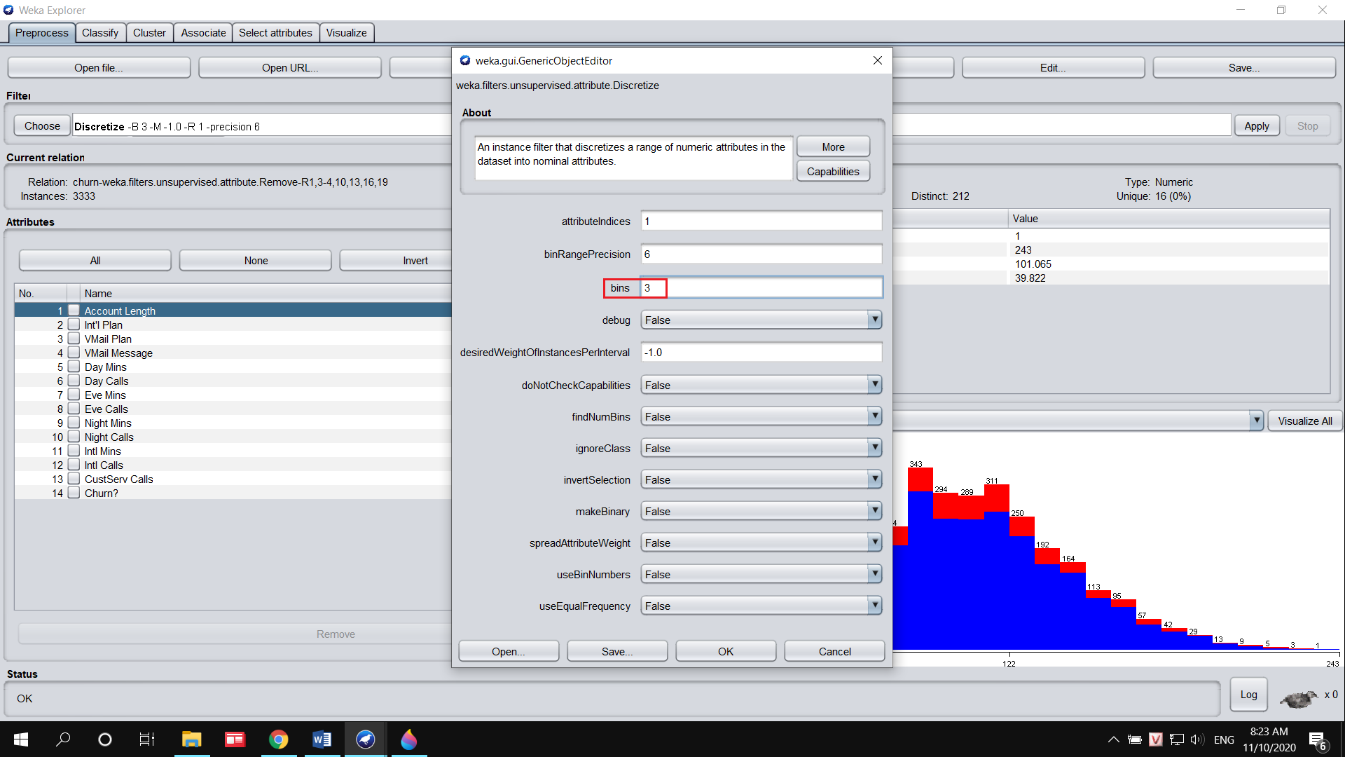
Sau khi loại bỏ các thuộc tính trên ta có tập dữ liệu như sau:

#### 2.2.3. Rời rạc hóa các thuộc tính có kiểu numeric

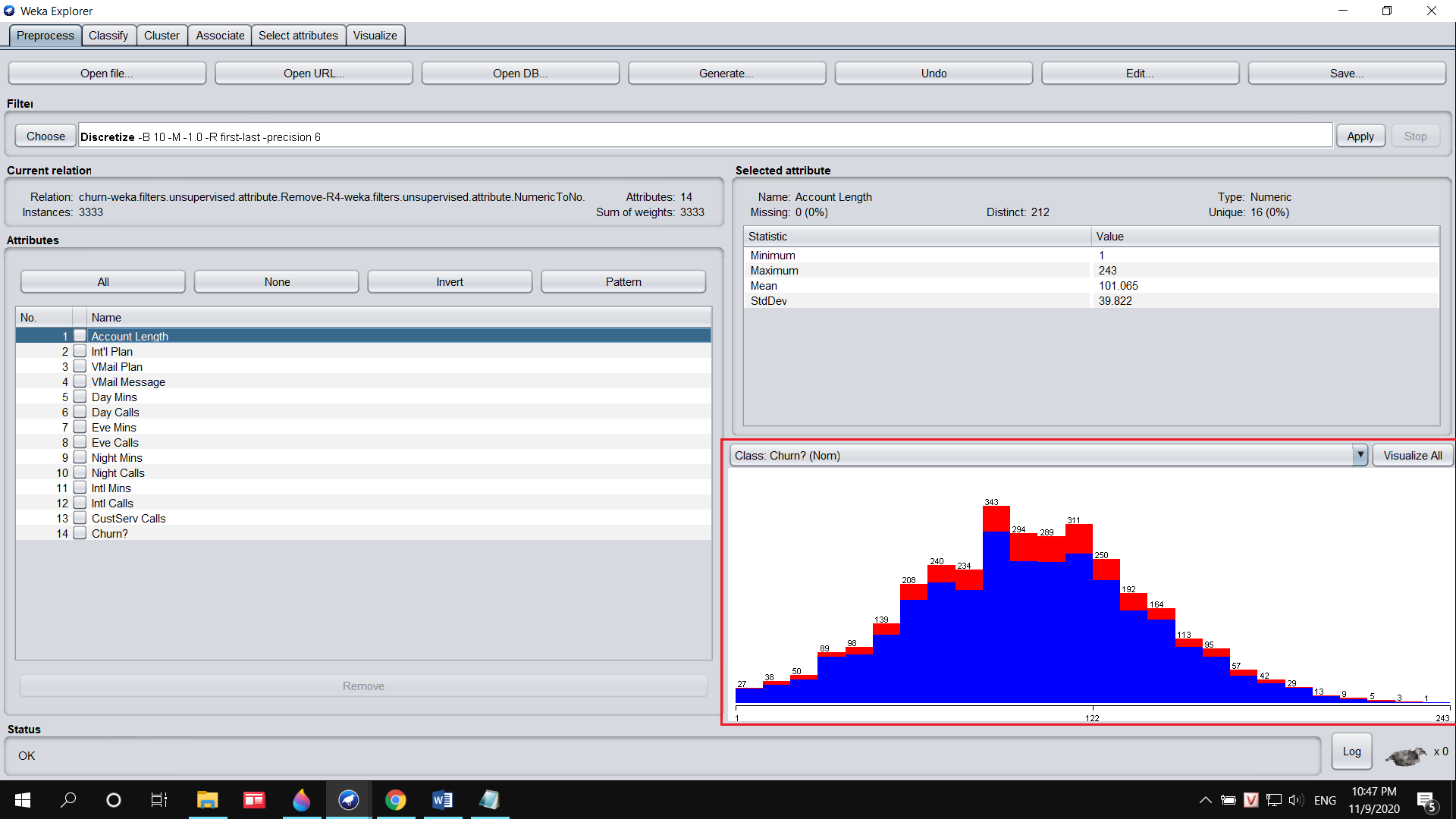
Mục đích: Để giảm số lượng giá trị cho một thuộc tính liên tục cho trước. Muốn sử dụng thuật toán Apriori thì tất cả các thuộc tính trong tập dữ liệu phải là nominal.

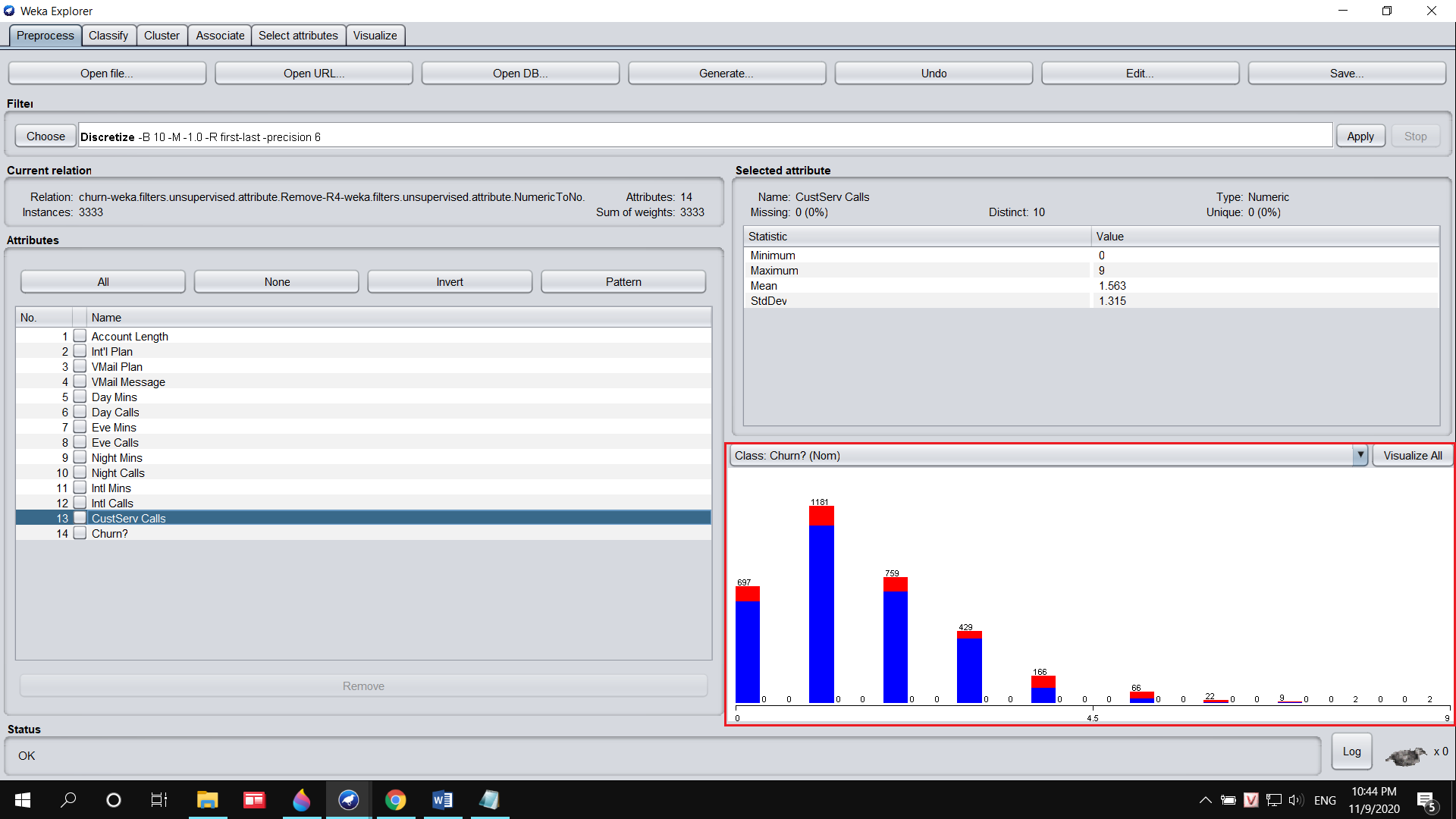
Tác dụng: Tạo ra phân cấp khái niệm. Giảm dữ liệu bằng cách thu thập và thay thế mức độ khái niệm thấp (numeric value for Day Mins) bằng khái niệm mức độ cao (VD: low, medium, high).

Cách rời rạc hóa dữ liệu trong Weka:

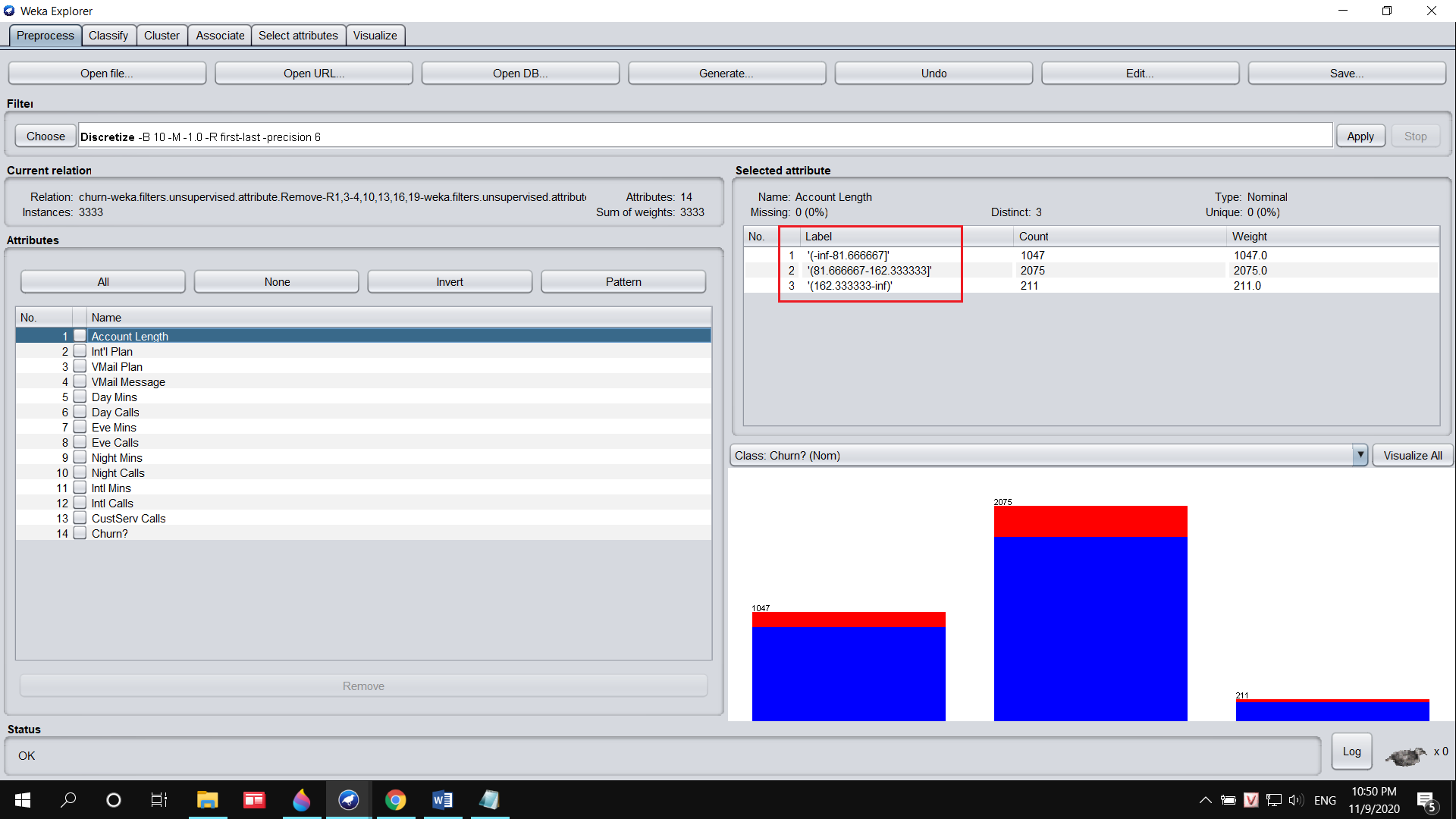
*Trong panel Filter > Choose  > filters >unsupervised > attribute > Discretize*

Với các thuộc tính *Account Length, VMail Message, Day Mins, Day Calls, Eve Mins, Eve Calls, Night Mins, Night Calls, Intl Mins, Intl Calls* ta chia ra 3 khoảng tương ứng với {low, medium, high}.

Với thuộc tính *CustServ Calls* ta chia ra 2 khoảng tương ứng với {low, high} vì ta thấy dữ liệu lệch hẳn về một bên.

**

Sau khi rời rạc hóa dữ liệu thì Weka tự tạo ra các nhãn như sau:

Ta thay các nhãn đó thành các nhãn ta đã định nghĩa ở trên bằng cách sử dụng text editor.

'Account Length' {'(-inf-81.666667]','(81.666667-162.333333]','(162.333333-inf)'} => {low, medium, high}

'VMail Message' {'(-inf-17]','(17-34]','(34-inf)'} => {low, medium, high}

'Day Mins' {'(-inf-116.933333]','(116.933333-233.866667]','(233.866667-inf)'} => {low, medium, high}

'Day Calls' {'(-inf-55]','(55-110]','(110-inf)'} => {low, medium, high}

'Eve Mins' {'(-inf-121.233333]','(121.233333-242.466667]','(242.466667-inf)'} => {low, medium, high}

'Eve Calls' {'\'(-inf-56.666667]\'','\'(56.666667-113.333333]\'','\'(113.333333-inf)\''} => {low, medium, high}

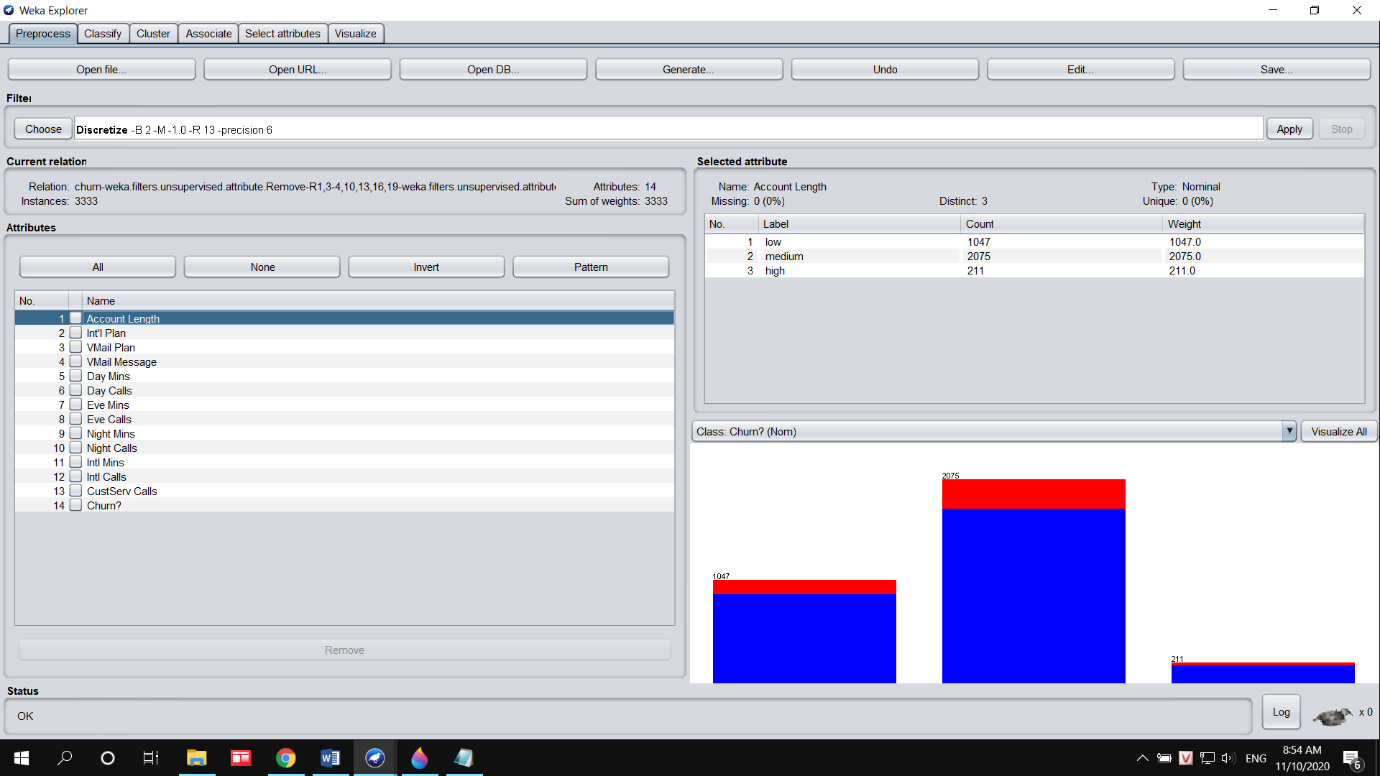
'Night Mins' {'(-inf-147.133333]','(147.133333-271.066667]','(271.066667-inf)'} => {low, medium, high}

'Night Calls' {'(-inf-80.333333]','(80.333333-127.666667]','(127.666667-inf)'} => {low, medium, high}

'Intl Mins' {'(-inf-6.666667]\'','\'(6.666667-13.333333]\'','\'(13.333333-inf)\''} => {low, medium, high}

'Intl Calls' {'(-inf-6.666667]','(6.666667-13.333333]','(13.333333-inf)'} => {low, medium, high}

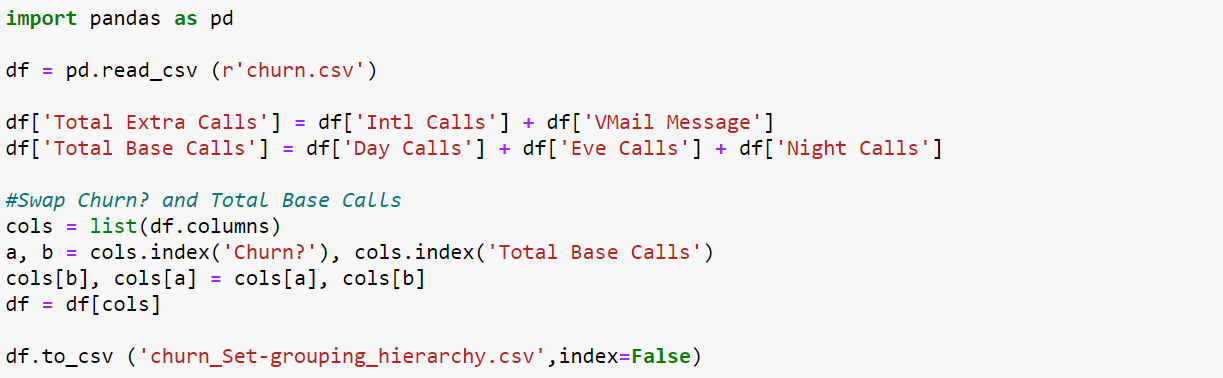
'CustServ Calls' {'(-inf-4.5]','(4.5-inf)'} => {low, high}

Sau khi thay nhãn xong thì có tập dữ liệu như sau:

Tập dữ liệu này được lưu thành file **churn\_preprocessing.arff**.

#### 2.2.4 Tiền xử lý dữ liệu nâng cao đối với module Weka không có sẵn

Sử dụng đoạn code sau để chuẩn bị dữ liệu cho khái niệm phân cấp Set-grouping hierarchy Ta thêm 2 cột Toal Base Calls (tổng của các cột Day Calls, Eve Calls, Night Calls) và Toal Extra Calls (tổng của 2 cột Intl Calls và Vmail Message) vào tập dữ liệu.



Tập dữ liệu này được lưu thành file **churn\_Set-grouping\_hierarchy.csv**. Sau đó, ta tiến hành tiếp tục tiền xử lý tập dữ liệu mới này tương tự như mục **2.2.2** và **2.2.3** trước khi đưa vào khai thác.

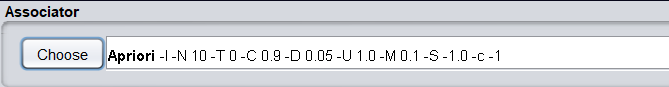
## 3.2. Khai thác luật kết hợp với thuật toán Apriori

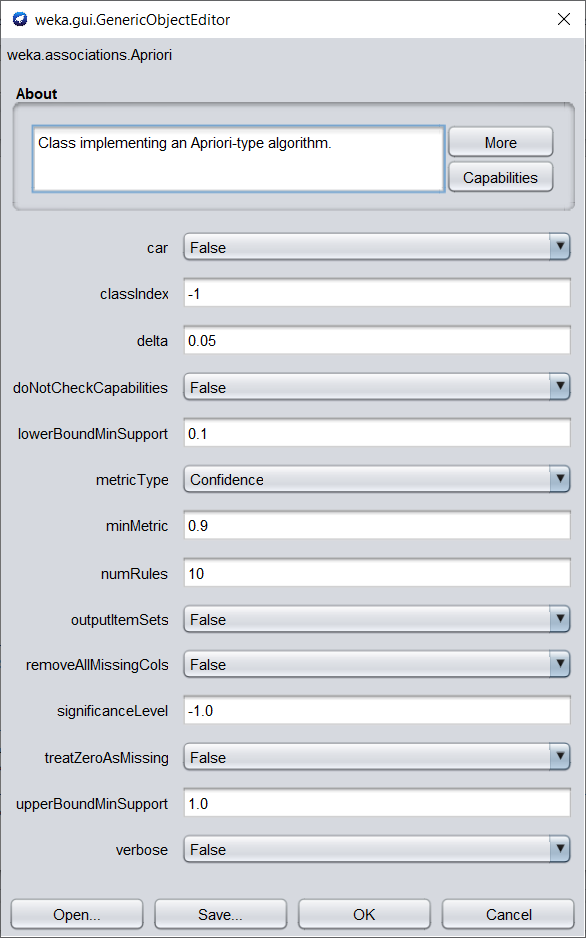
Sau khi tiền xử lý dữ liệu xong thì bắt đầu khai thác luật kết hợp với thuật toán Apriori trong module association rule mining của Weka.

Đặc điểm:

* Chọn độ đo lý thú của luật, chọn giá trị minconf và số lượng luật cần khai thác
* Lặp lại việc giảm minsup cho đến khi tìm được đủ số luật yêu cầu với minconf

*Trong module associate > Choose  > associations > Apriori*





Các tham số chính của thuật toán Apriori:

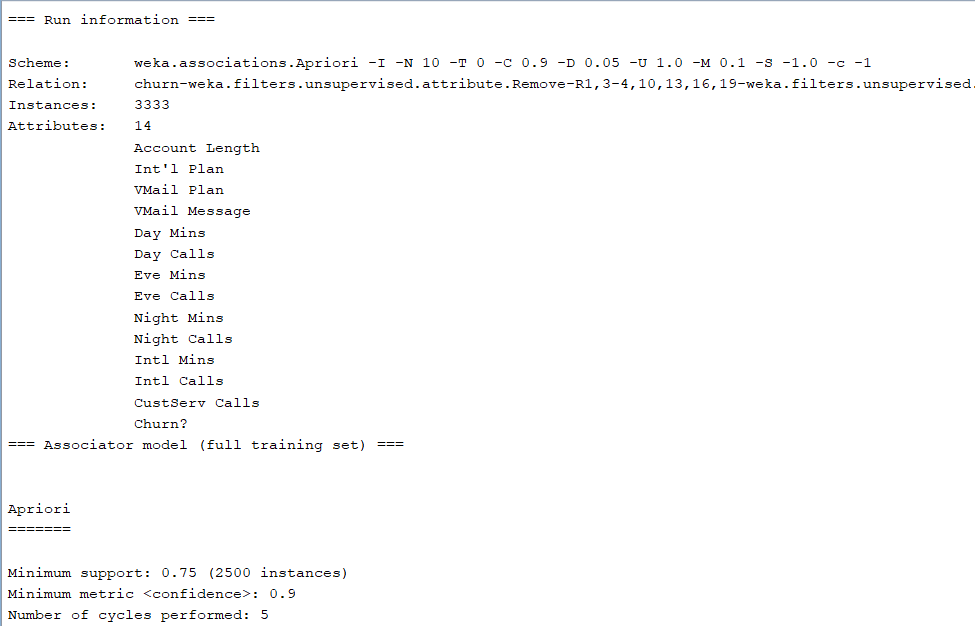
* lowerBoundMinSupport (-M): Chặn dưới minsupport.
* upperBoundMinSupport (-U): Chặn trên minsupport. Bắt đầu giảm từ giá trị này.
* delta (-D): hệ số giảm support khi lặp , giảm support cho đến khi đạt minsupp hay đã phát sinh đủ số luật.
* metric Type(-T): độ đo lý thú của luật bao gồm: Confidence, Lift, Leverage, Conviction.
* minMetric (-C): độ tin cậy nhỏ nhất. Chỉ xét những luật có điểm lớn hơn giá trị này.
* numRules (-N): số luật cần phải tìm.
* outputItemSets (-I): Xuất ra các nội dung các tập hạng mục.

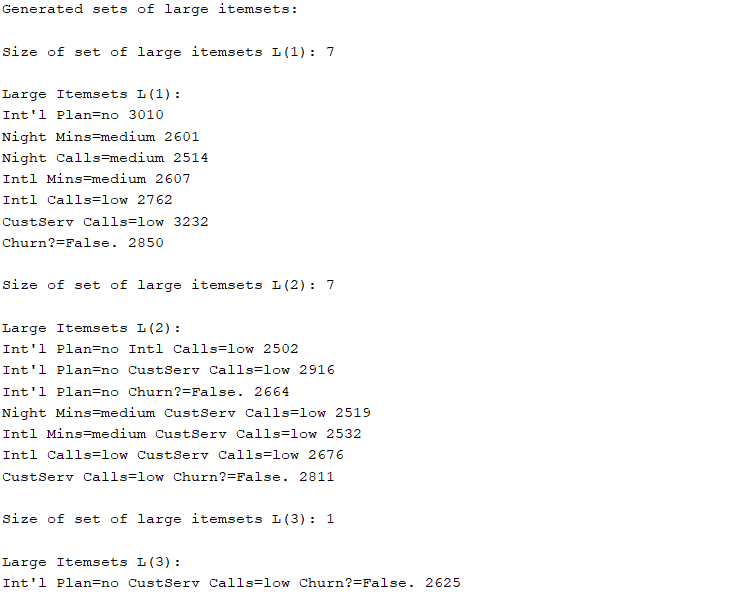
Chạy thuật toán Apriori trên tập dữ liệu với các tham số:

* lowerBoundMinSupport = 0.1
* upperBoundMinSupport
* delta = 0.05
* metric Type = Confidence
* minMetric = 0.9
* numRules = 10
* outputItemSets = True

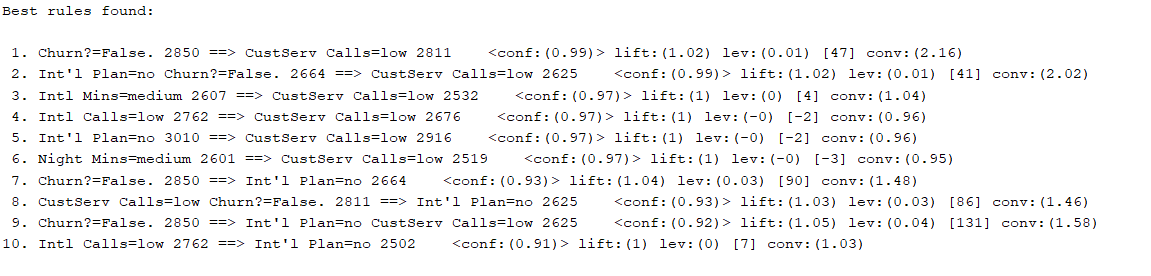
Phân tích kết quả:

Ta có Minsup = 0.75, Minconf = 0.9 thì sẽ sinh ra đủ 10 luật nên thuật toán dừng.



Thông tin chi tiết của các tập phổ biến:

Thông tin chi tiết về các tập phổ biến: 10 luật kết hợp tốt nhất thu được, được sắp xếp theo thứ tự giảm dần giá trị conf:



# Phần IV: Experiments

## 4.1. Tại sao phải phân tích dữ liệu

Phân tích dữ liệu giúp ta hiểu rõ dữ liệu hơn, biết được những yếu tố tác động đến hành vi của khách hàng khi quyết định rời bỏ hoặc tiếp tục sử dụng dịch vụ của doanh nghiệp.

Việc phân tích dữ liệu giúp việc xem và hiểu những gì đang diễn ra nhanh hơn, từ đó doanh nghiệp có thể kịp thời đưa ra các giải pháp phù hợp để giải quyết vấn đề phát sinh ra mà ảnh hưởng đến dữ liệu và doanh thu của doanh nghiệp.

Việc phân tích dữ liệu cũng giúp ta tìm ra được sự liên quan, sự ràng buộc giữa các thuộc tính với nhau, từ đó có thể khái quát hóa các khái niệm để phân lớp các thuộc tính, chỉ ra được các đặc tính riêng của mỗi phân lớp trong miền giá trị của các thuộc tính trong tập dữ liệu.

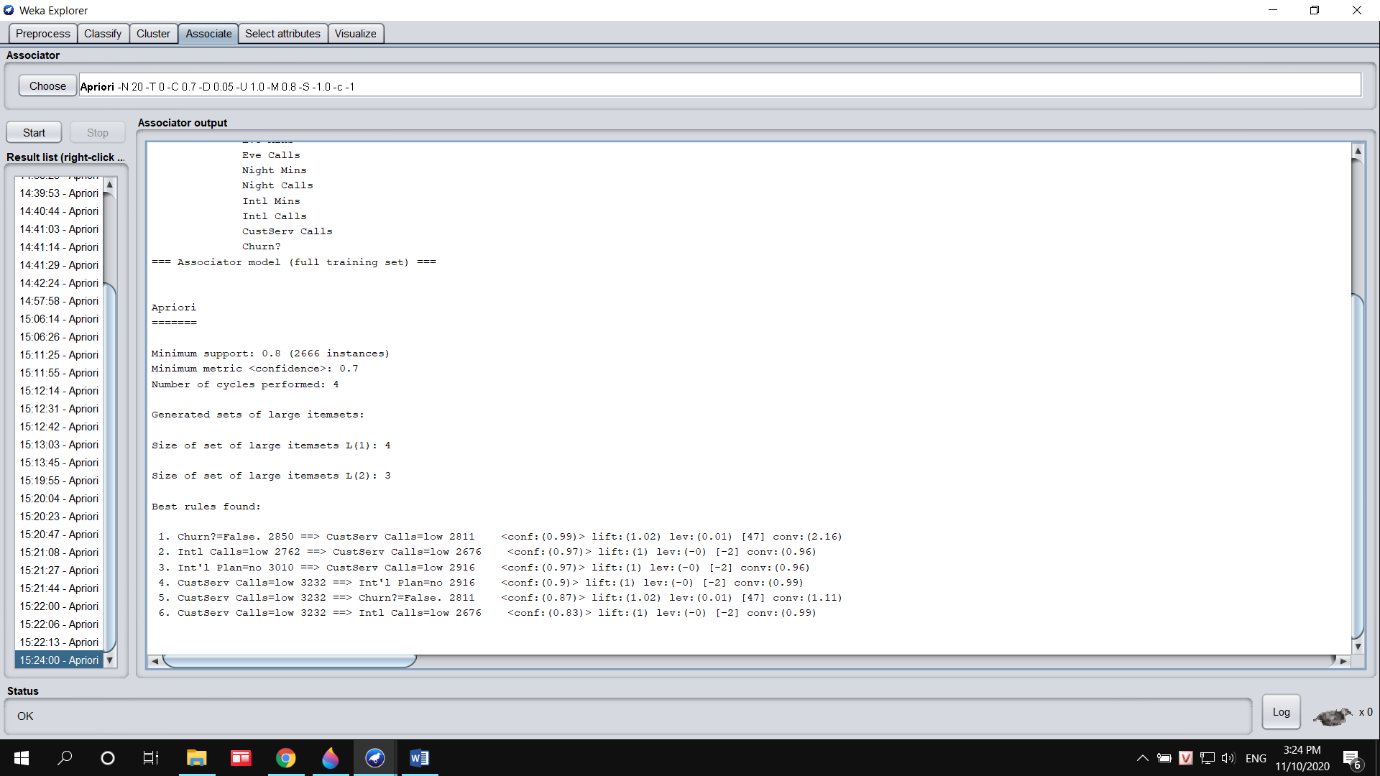
## 4.2. Khai thác luật

#### 4.2.1. Khai thác luật dựa trên các hệ số khác nhau

*Thử nghiệm 1: Dùng tất cả biến của dữ liệu*

Tiền xử lý: Loại bỏ thuộc tính thừa và thuộc tính tương quan, rời rạc hóa các thuộc tính có kiểu numeric (đã mô tả ở phần 2).

Tham số hệ thống cho thuật toán Apriori: Minsup = 0.8, Minconf = 0.7, Metric Type = Confidence, numRules = 20.

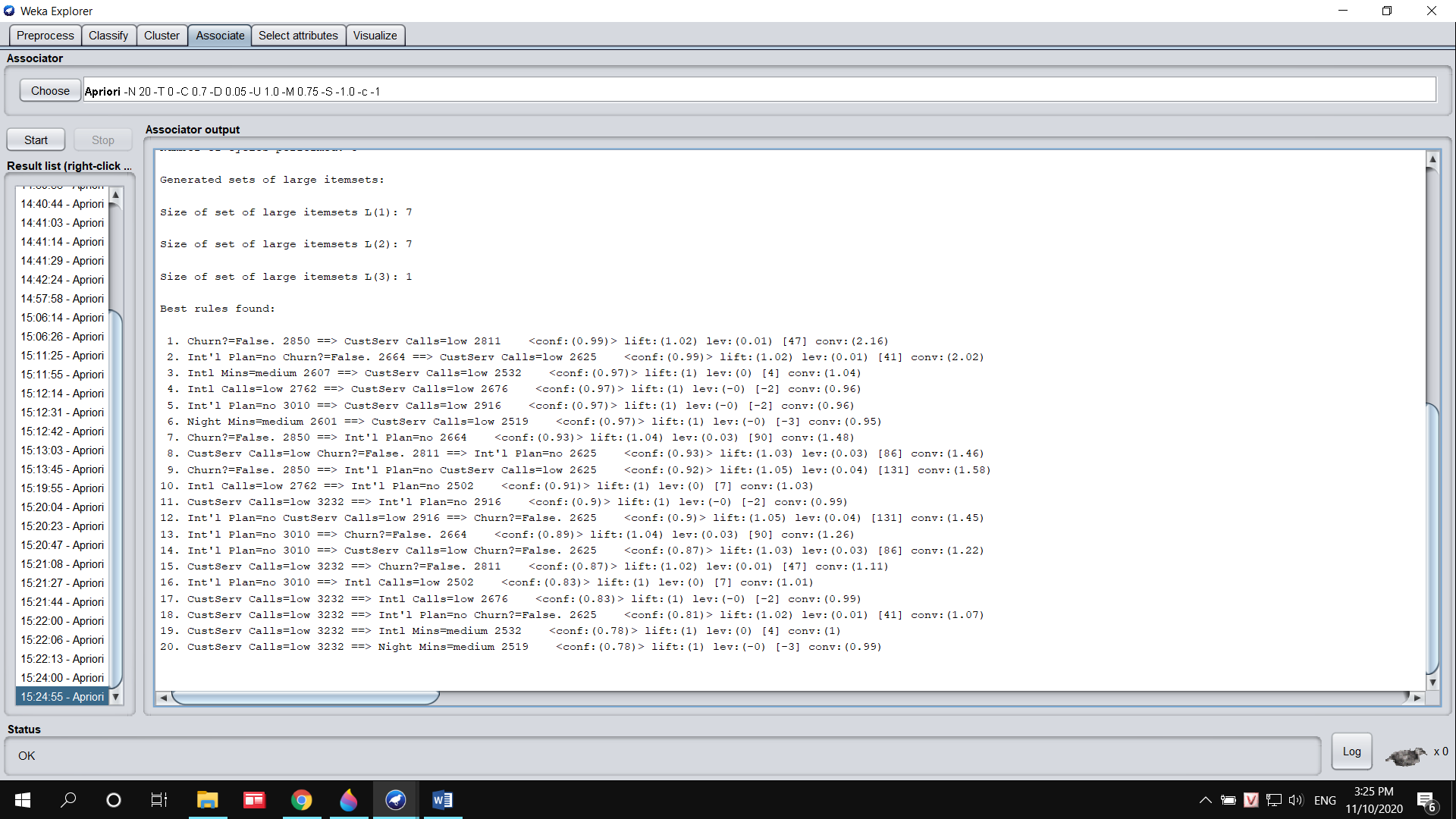
Ý nghĩa của các luật:

1. Khách hàng không rời bỏ dịch vụ (churn?=false) thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
2. Khách hàng có số cuộc gọi quốc tế thấp (Intl Calls=low) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
3. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
4. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không sử dụng dịch vụ gọi quốc tế (Int'l Plan=no)
5. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không rời bỏ dịch vụ (churn?=false)
6. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì có số cuộc gọi quốc tế thấp (Intl Calls=low)

*Thử nghiệm 2: Dùng tất cả biến của dữ liệu, tuy nhiên giảm độ hỗ trợ cho luật*

Tiền xử lý: Loại bỏ thuộc tính thừa và thuộc tính tương quan, rời rạc hóa các thuộc tính có kiểu numeric (đã mô tả ở phần 2).

Tham số hệ thống cho thuật toán Apriori: Minsup = 0.75, Minconf = 0.7, Metric Type = Confidence, numRules = 20.



Ý nghĩa các luật:

1. Khách hàng không rời bỏ dịch vụ (churn?=false) thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
2. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) và không rời bỏ dịch vụ (churn=false?) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
3. Khách hàng có số phút gọi quốc tế trung bình (Intl Mins=medium) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
4. Khách hàng có số cuộc gọi quốc tế thấp (Intl Calls=low) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
5. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
6. Khách hàng có số phút gọi buổi đêm khuya trung bình (Night Mins=medium) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
7. Khách hàng không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
8. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) và không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
9. Khách hàng không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no) và cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
10. Khách hàng có số cuộc gọi quốc tế thấp (Intl Calls=low) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
11. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
12. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) và có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không rời bỏ dịch vụ (churn?=false)
13. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì không rời bỏ dịch vụ (churn?=false)
14. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) và không rời bỏ dịch vụ (churn?=false)
15. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không rời bỏ dịch vụ (churn?=false)
16. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì có số cuộc gọi quốc tế thấp (Intl Calls=low)
17. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì có số cuộc gọi quốc tế thấp (Intl Calls=low)
18. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no) và không rời bỏ dịch vụ (churn?=false)
19. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì có số phút gọi quốc tế trung bình (Intl Mins=medium)
20. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì có số phút gọi buổi đêm khuya trung bình (Night Mins=medium)

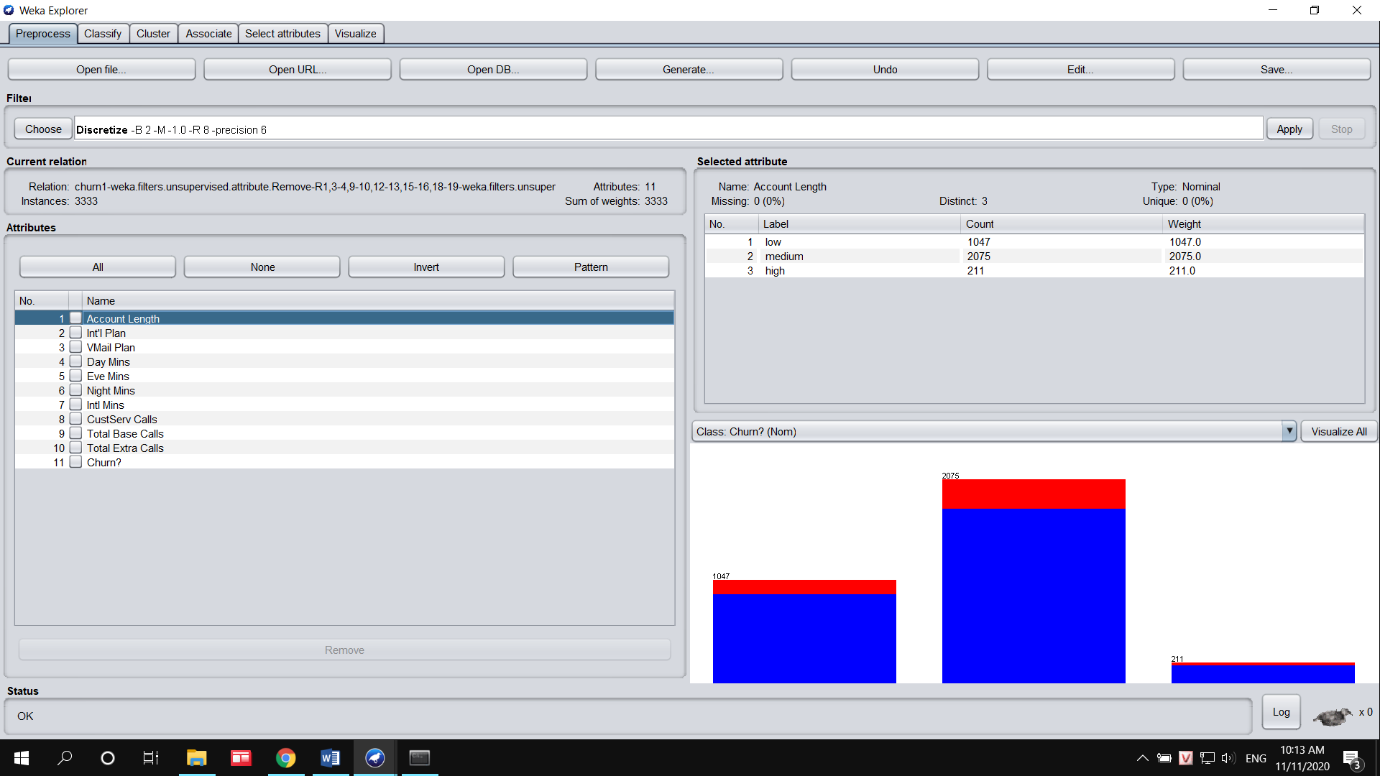
#### 4.2.1. Thử nghiệm lại với góc nhìn khác của dữ liệu

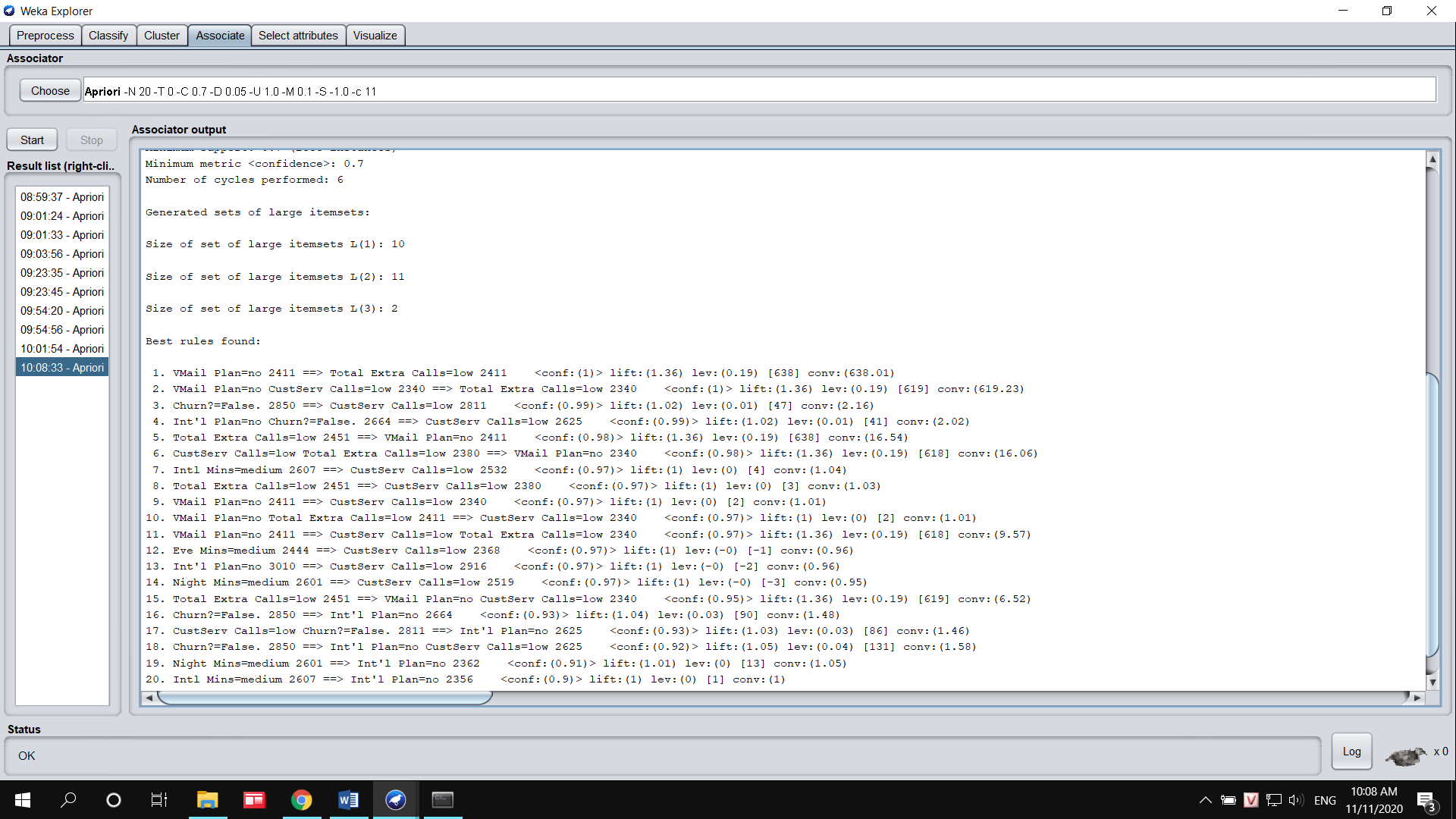
**Tổng quát hóa hóa dữ liệu dựa trên khái niệm phân cấp Set-grouping hierarchy**

Ta tiến hành chia các dịch vụ gọi thoại ra thành 2 nhóm:

1. Gói dịch vụ cơ bản: bao gồm các dịch vụ gọi thoại ban ngày, dịch vụ gọi thoại buổi chiều và dịch vụ gọi thoại buổi tối.
2. Gói dịch vụ bổ sung: bao gồm gói dịch vụ gọi âm thanh và gói dịch vụ gọi quốc tế.

Từ 2 nhóm trên, ta có thể tổng quát hóa khái niệm các thuộc tính của tập dữ liệu về 2 thuộc tính là gói dịch vụ cơ bản và gói dịch vụ bố sung. Đối với mỗi thuộc tính, ta có thể dùng một đại lượng chung để biểu diễn giá trị của mỗi khách hàng đó là tổng số *cuộc gọi/tin nhắn* cho mỗi gói dịch vụ. Sau đó, ta tiến hành khai thác tập luật trên tập dữ liệu mới này. Đương nhiên, tập luật khai thác được từ tập dữ liệu này sẽ mang tính khái quát hơn tập ban đầu.

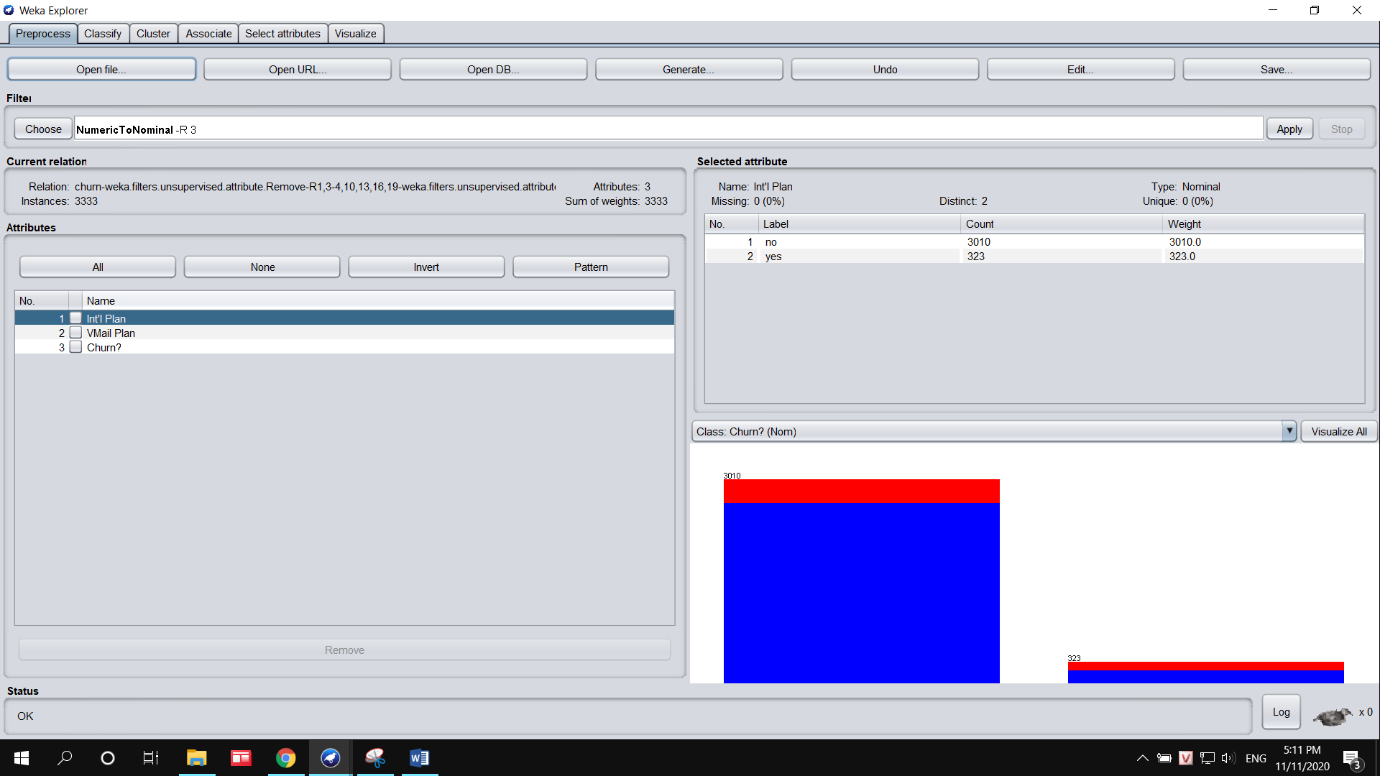
Ta thêm 2 cột Toal Base Calls (tổng của các cột Day Calls, Eve Calls, Night Calls) và Toal Extra Calls (tổng của 2 cột Intl Calls và Vmail Message) vào tập dữ liệu. Đồng thời xóa các cột Day Calls, Eve Calls, Night Calls, Intl Calls và Vmail Message và thực hiện các bước tiền xử lý như đã trình bày ở phần Code (rời rạc hóa dữ liệu). Tập dữ liệu lưu ở **churn\_set-grouping\_hierarchy.arff**:

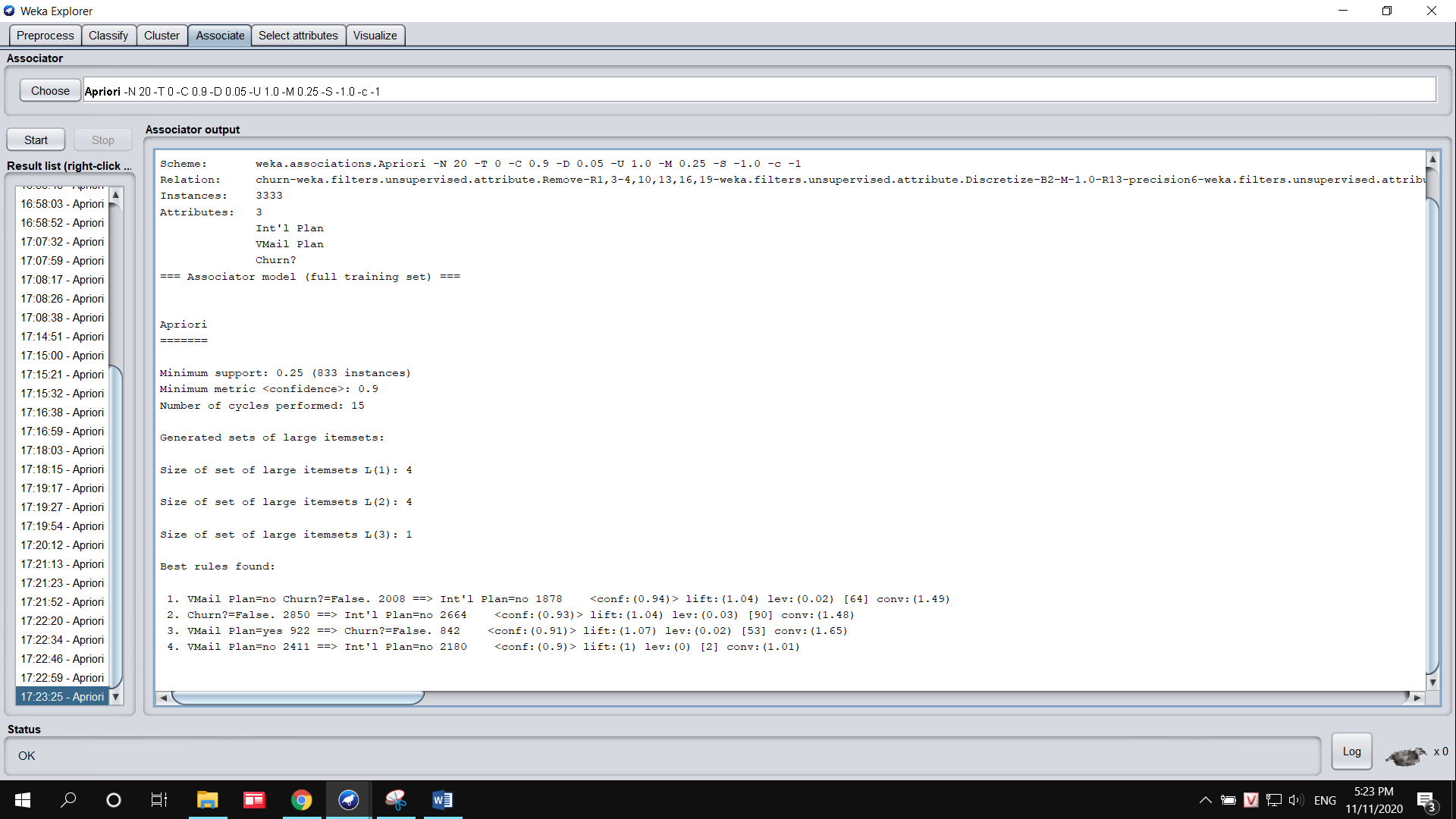
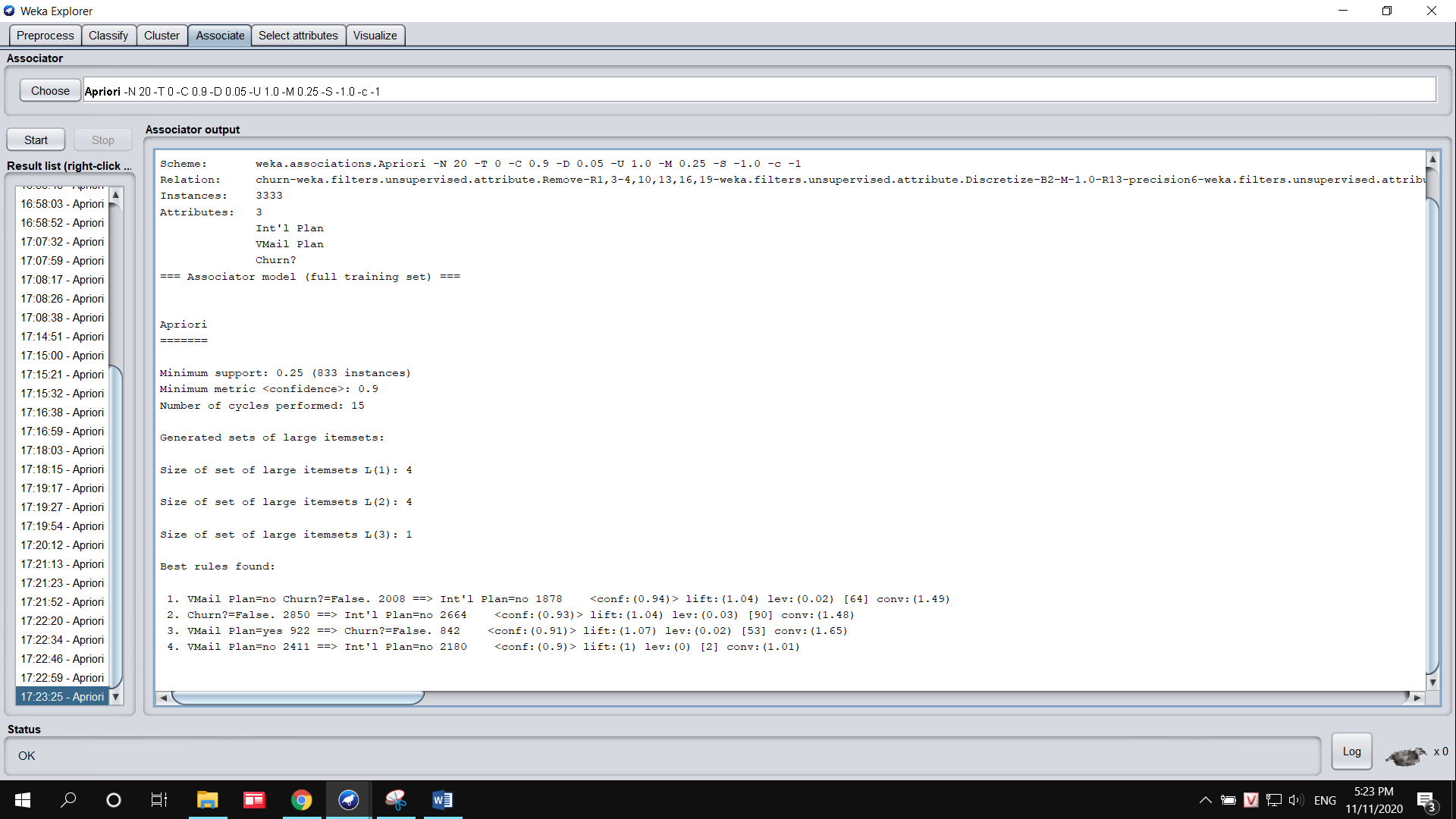
Sử dụng thuật toán Apriori để khai thác tập dữ liệu trên với Minsup = 0.7, Minconf = 0,7 Metric Type = Confidence, numRules = 20 ta khai thác được các luật như sau:

Ý nghĩa của các luật:

1. Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) thì có số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low)
2. Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) và cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì có số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low)
3. Khách hàng không rời bỏ dịch vụ (churn?=false) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
4. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) và không rời bỏ dịch vụ (churn?=false) thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
5. Khách hàng có số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low) thì không sử dụng gói cước Voice Mail (Vmail Plan=no)
6. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) và số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low) thì không sử dụng gói cước Voice Mail (Vmail Plan=no)
7. Khách hàng có số phút gọi quốc tế trung bình (Intl Mins=medium) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
8. Khách hàng có số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
9. Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
10. Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) và số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
11. Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) và số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low)
12. Khách hàng số phút gọi buổi tối trung bình (Eve Mins=medium) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
13. Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
14. Khách hàng số phút gọi buổi đêm khuya trung bình (Night Mins=medium) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
15. Khách hàng có số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low) thì không sử dụng gói cước Voice Mail (Vmail Plan=no) và cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
16. Khách hàng không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
17. Khách hàng có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) và không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
18. Khách hàng không rời bỏ dịch vụ (churn?=false) thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) và không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
19. Khách hàng số phút gọi buổi đêm khuya trung bình (Night Mins=medium) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
20. Khách hàng có số phút gọi quốc tế trung bình (Intl Mins=medium) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)

**Cụ thể hóa dữ liệu dựa trên lựa chọn các phần khác nhau của dữ liệu.**

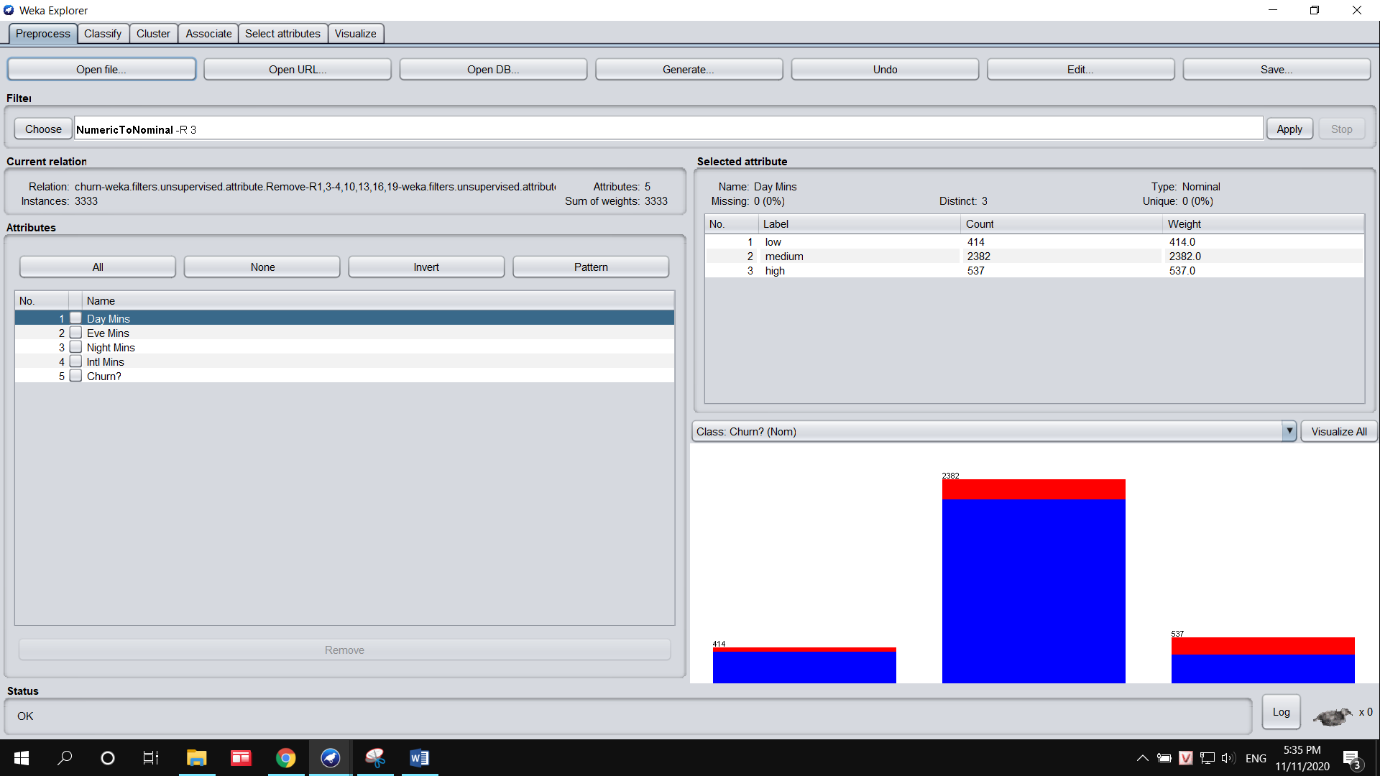
Ta muốn tìm ra mối quan hệ giữa các gói dịch vụ (Int’l Plan, Vmail Plan) với Churn. Ta có tập dữ liệu với 3 thuộc tính trên lưu ở file **churn\_intl-plan\_vmail-plan.arff**.

Sử dụng thuật toán Apriori để khai thác tập dữ liệu trên với Minsup = 0.25, Minconf = 0,9 Metric Type = Confidence, numRules = 20 ta khai thác được các luật như sau:

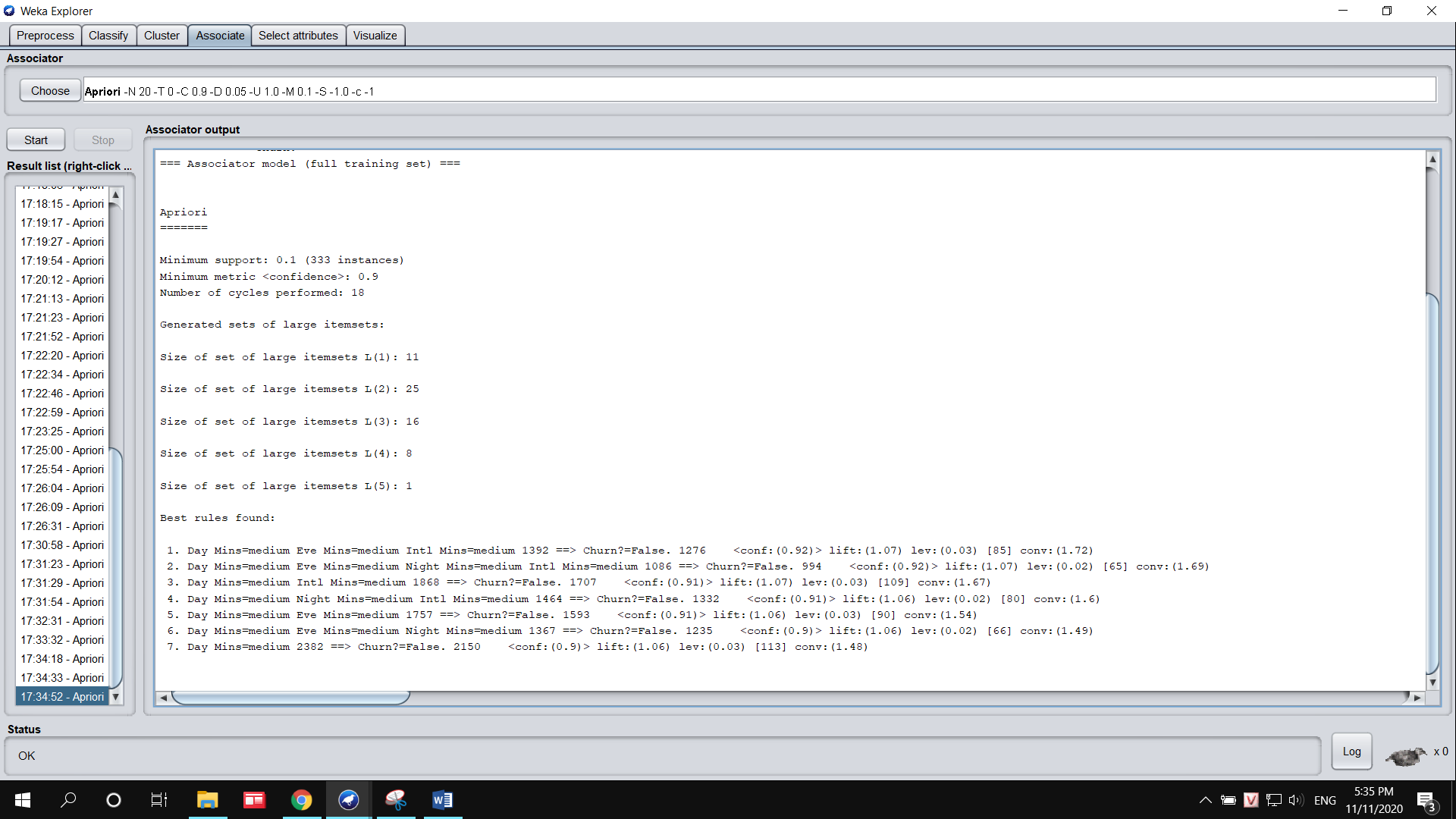
Ý nghĩa các luật:

* Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) và không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
* Khách hàng không rời bỏ dịch vụ (churn?=false) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
* Khách hàng sử dụng gói cước Voice Mail (Vmail Plan=yes) thì không rời bỏ dịch vụ (churn?=false)
* Khách hàng không sử dụng gói cước Voice Mail (Vmail Plan=no) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)

Ta muốn tìm ra mối quan hệ giữa số phút gọi (Day Mins, Eve Mins, Night Mins, Intl Mins) với Churn. Ta có tập dữ liệu sau với 5 thuộc tính trên lưu ở file **churn\_All-Mins.arff**.



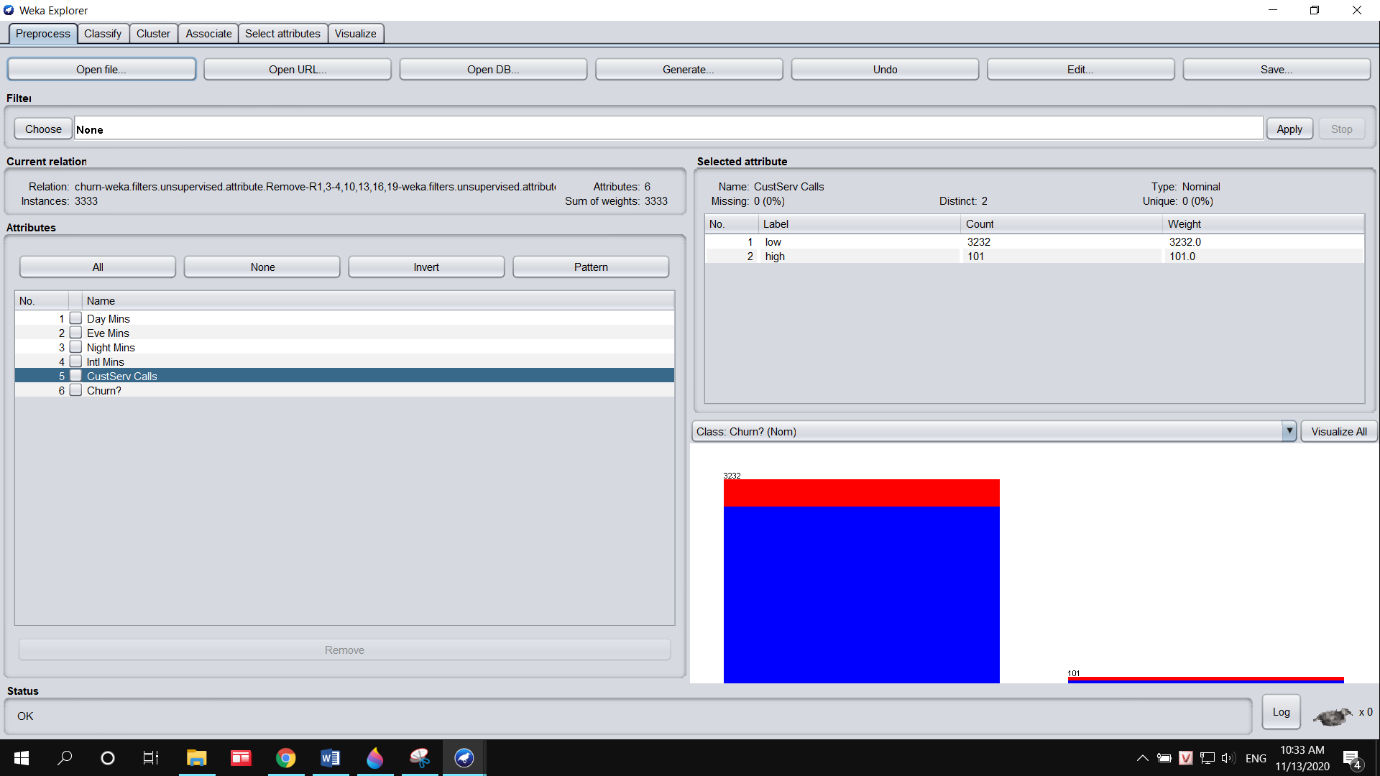
Sử dụng thuật toán Apriori để khai thác tập dữ liệu trên với Minsup = 0.1, Minconf = 0,9 Metric Type = Confidence, numRules = 20 ta khai thác được các luật như sau:

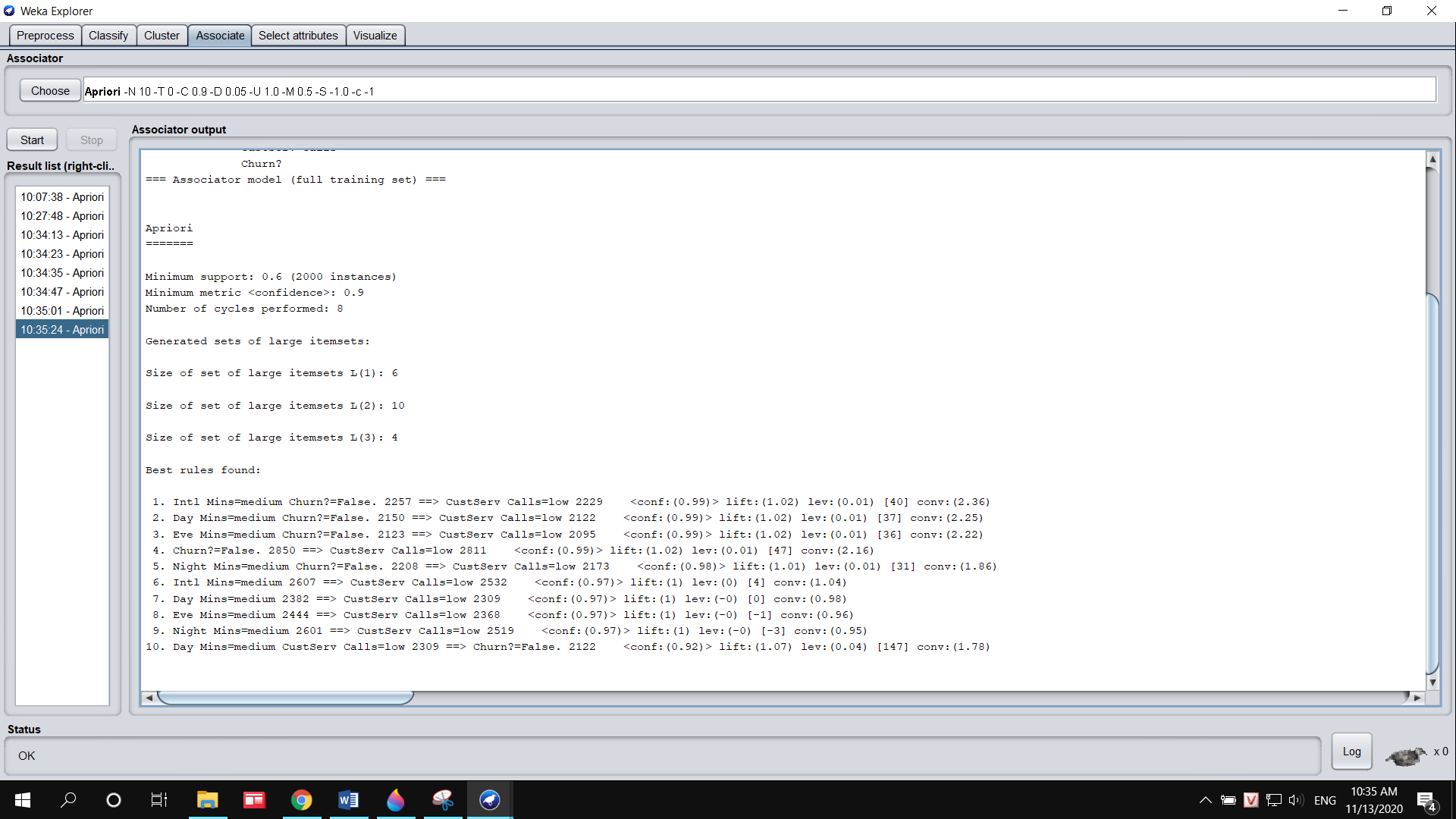


Vì các luật có nghĩa tương đương nhau nên luật cuối cùng là:

Khách hàng có số phút gọi ban ngày, đêm, khuya, quốc tế trung bình thì không rời bỏ dịch vụ (Churn?=False)

Ta muốn tìm ra mối quan hệ giữa số phút gọi (Day Mins, Eve Mins, Night Mins, Intl Mins, CustServ Calls) với Churn. Ta có tập dữ liệu sau với 6 thuộc tính trên lưu ở file **churn\_All-Mins\_CustServCalls.arff**.



Sử dụng thuật toán Apriori để khai thác tập dữ liệu trên với Minsup = 0.6, Minconf = 0,9 Metric Type = Confidence, numRules = 10 ta khai thác được các luật như sau:

Ý nghĩa luật đặc trưng nhất:

1. Khách hàng có số phút gọi ban ngày, đêm, khuya, quốc tế trung bình thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
2. Khách hàng có số phút gọi ban ngày trung bình (Day Mins=medium) và có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low) thì không rời bỏ dịch vụ (Churn?=False)

# Phần V: Tóm tắt kết quả

## 5.1. Đánh giá kết quả

Ta có một số phương pháp để đánh giá kết quả:

* *Dựa trên tính chính xác của các luật trả về trên tập dữ liệu đầu vào.* Bằng cách điều chỉnh các thông số của thuật toán khai phá luật như minsup, minconf, ta dễ dàng lựa chọn và đánh giá được độ tin cậy và tỉ lệ xuất hiện của mỗi luật trong các giao tác trên tập dữ liệu.
* *Dựa trên khoảng cách giữa các luật trả về.* Một bộ luật được gọi là đặc trưng của tập dữ liệu nếu tần suất xuất hiện của nó nhiều hơn đáng kể so với những luật còn lại. Đối với những luật như vậy, tỉ lệ nó là một luật có ích trong việc ra quyết định là rất cao.
* *Dựa trên kích thước của tập dữ liệu.* Các thuật toán khai phá luật kết hợp được dùng trong việc khai thác dữ liệu đều chỉ chạy trên một tập các giao tác ở quá khứ. Nó tốt hơn là chỉ nên dùng để tìm ra một mẫu (*pattern*) để khi có sự xuất hiện lại của mẫu này ở trong tương lai, ta sẽ dự đoán được cách thức mà mẫu này diễn ra (Ví dụ: phải đến mùa đông thì khách hàng mới mua kèm lò sưởi với chăn điện). Khi đó, nếu tập dữ liệu các giao tác của ta càng lớn, các luật thu được khi khai thác sẽ càng khái quát, là một mẫu (*pattern*) lớn khiến nó trở nên phù hợp hơn cho việc ra quyết định.

Trong khuôn khổ đồ án này, nhóm thực hiện đánh giá dựa trên tính chính xác của các luật trả về trên tập dữ liệu đầu vào và dựa trên khoảng cách giữa các tập trả về. Mục tiêu là tìm ra được các luật khách quan và tốt nhất có thể để phục vụ cho việc đưa ra quyết định của doanh nghiệp.

## 5.2. Tập luật tốt thu được

Nhóm muốn tìm ra được tập luật dẫn đến Churn = False để phục vụ cho việc đưa ra kết luận giúp tăng lợi nhuận của công ty

* Khách hàng có số cuộc gọi dịch vụ thấp (CustServ Calls=low) thì tỷ lệ rời bỏ dịch vụ (Churn?=False)
* Khách hàng có gói cước Voice Mail (Vmail Plan=yes) thì tỷ lệ rời bỏ dịch vụ thấp (Churn?=False)
* Khách hàng không rời bỏ dịch vụ (Churn?=False) thì không sử dụng gói cước gọi quốc tế (Int'l Plan=no)
* Khách hàng không sử dụng gói cước gọi quốc tế (Int'l Plan=no) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
* Khách hàng có số phút gọi ban ngày, đêm, khuya, quốc tế trung bình thì không rời bỏ dịch vụ (Churn?=False)
* Khách hàng có số cuộc gọi dịch vụ bổ sung thấp (Total Extra Calls=low) thì cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)
* Khách hàng có số phút gọi ban ngày, đêm, khuya, quốc tế trung bình thì có cuộc gọi dịch vụ khách hàng thấp (CustServ Calls=low)

## 5.3. Đánh giá nhóm

#### 5.3.1 Ưu điểm

Sử dụng thành thạo Weka trong việc khai thác luật kết hợp.

Phân tích, đánh giá sự quan trọng của các thuộc tính trên tập dữ liệu.

Thực hiện các biện pháp tiền xử lý, tiền xử lý nâng cao trên tập dữ liệu dựa trên phân cấp khái niệm để có nhiều cái nhìn khách quan hơn khi khai thác dữ liệu.

Khai thác các tập luật kết hợp với thuật toán Apriori bằng nhiều tham số khác nhau để chỉ ra các luật kết hợp thu được là đáng tin cậy.

#### 5.3.2 Nhược điểm

Các luật khai thác được còn nhiều, có thể dư thừa, không ảnh hưởng nhiều đến việc ra quyết định của doanh nghiệp.

# PHỤ LỤC

# TÀI LIỆU THAM KHẢO

[1] <http://ce.sharif.edu/courses/85-86/1/ce925/assignments/files/assignDir4/Churn.pdf>

[2] <http://summit.sfu.ca/system/files/iritems1/7984/b18914287.pdf>