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

BÁO CÁO ĐỒ ÁN 2
KHAI THÁC TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP

Bộ môn: Khai thác dữ liệu và ứng dụng
Giảng viên hướng dẫn: Nguyễn Ngọc Đức

2020 - 2021

MỤC LỤC

I. THÔNG TIN NHÓM

Họ và tên	MSSV
Lê Hoàng Phương Nhi	18120496
Lê Thị Như Quỳnh	18120530

- Mức độ hoàn thành đồ án: 100%

II. BÁO CÁO ĐỒ ÁN

1. Data

1.1. Mô tả tập dữ liệu

- Tập dữ liệu gồm 21 thuộc tính và 3333 mẫu. Đưa ra thông tin về khách hàng cùng với dấu hiệu khách hàng đó có bỏ công ty hay không

Thuộc tính	
State	50 tiểu bang và quận huyện ở Columbia
Account Length	Thời gian hoạt động của tài khoản (tài khoản đã hoạt động bao lâu)
Area Code	Mã vùng
Phone	Kiểu như ID khách hàng
Int'l Plan	Khách hàng có sử dụng dịch vụ quốc tế không, 2 giá trị yes or no
VMail Plan	Khách hàng có sử dụng dịch vụ VoiceMail không, 2 giá trị yes or no
Vmail Message	Số tin nhắn thoại
Day Mins	Số phút khách hàng sử dụng dịch vụ trong ngày
Day Calls	Số cuộc gọi khách hàng thực hiện trong ngày
Day Charge	Phí ngày
Eve Mins	Số phút khách hàng sử dụng dịch vụ vào buổi tối
Eve Calls	Số cuộc gọi khách hàng thực hiện vào buổi tối
Eve Charge	Phí buổi tối
Night Mins	Số phút khách hàng sử dụng dịch vụ vào buổi đêm
Night Calls	Số cuộc gọi khách hàng thực hiện vào ban đêm

Night Charge	Phí buổi đêm
Intl Mins	Số phút khách hàng sử dụng dịch vụ để thực hiện các cuộc gọi quốc tế
Intl Calls	Số cuộc gọi quốc tế khách hàng thực hiện
Intl Charge	Phí quốc tế
CustServ Calls	Số cuộc gọi đến dịch vụ khách hàng
Churn?	Có bỏ công ty hay không?

- Các thuộc tính số (numeric): Account Length, Area Code, Vmail Message, Day Mins, Day Calls, Day Charge, Eve Mins, Eve Calls, Eve Charge, Night Mins, Night Calls, Night Charge, Intl Mins. Intl Calls, Intl Charge, CustServ Calls
- Các thuộc tính rời rạc (nominal): State, Phone, Int'l Plan, Vmail Plan, Churn?
- Chi tiết về các thuộc tính số (numeric):

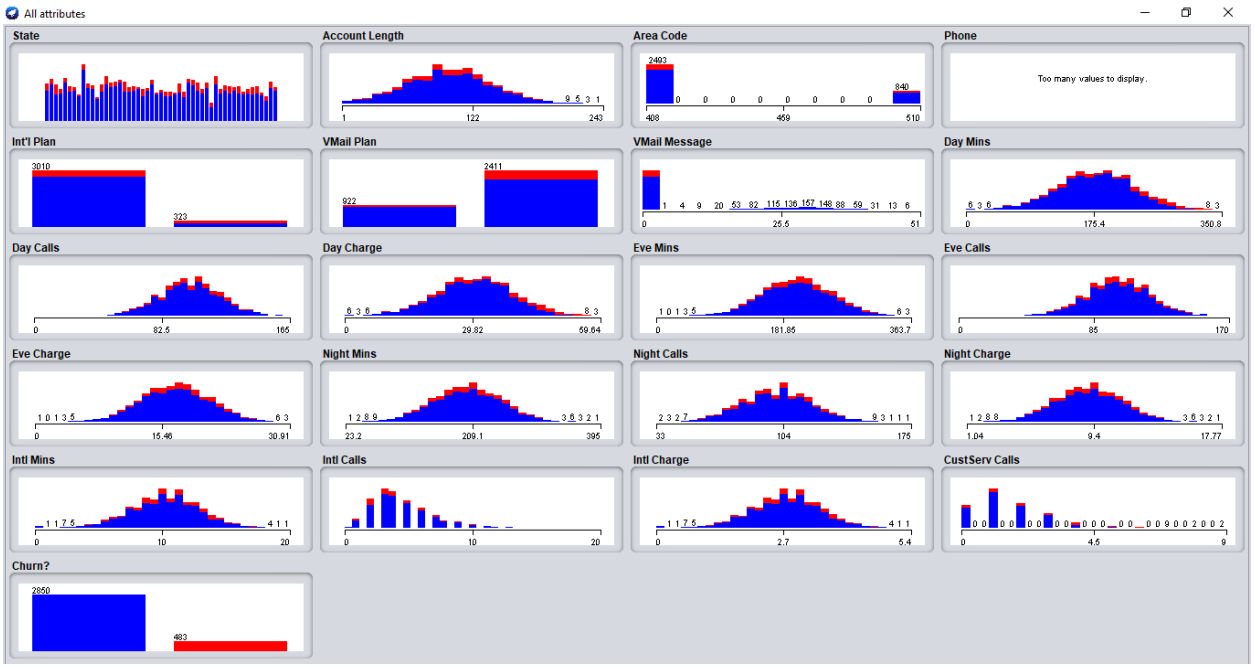
Thuộc tính	Minimum	Maximum	Mean	StdDev
Account Length	1	243	101.065	39.822
Area Code	408	510	437.182	42.371
Vmail Message	0	51	8.099	13.688
Day Mins	0	350.8	179.775	54.467
Day Calls	0	165	100.436	20.069
Day Charge	0	59.64	30.562	9.259
Eve Mins	0	363.7	200.98	50.714
Eve Calls	0	170	100.114	19.923
Eve Charge	0	30.91	17.084	4.311
Night Mins	23.2	395	200.872	50.574
Night Calls	33	175	100.108	19.569
Night Charge	1.04	17.77	9.039	2.276
Intl Mins	0	20	10.237	2.792
Intl Calls	0	20	4.479	2.461
Intl Charge	0	5.4	2.765	0.754
CustServ Calls	0	9	1.563	1.315

- Đồ thị biểu diễn sự phân bố giá trị của các thuộc tính:

+ Đối với các thuộc tính dạng số (numeric), đồ thị này sẽ là histogram: chia miền giá trị [min, max] thành nhiều miền con [ai,bi] với kích thước xấp xỉ nhau. Ứng với mỗi miền con, ta đếm số lượng mẫu có giá trị thuộc tính nằm trong miền. Cuối cùng, biểu diễn dưới dạng đồ thị cột.

+ Đối với các thuộc tính rời rạc (nominal), ứng với mỗi giá trị của thuộc tính, ta đếm số lượng mẫu có giá trị đó. Tương tự như các thuộc tính dạng số, biểu diễn thành đồ thị cột. Nhìn vào đồ thị, với mỗi giá trị của thuộc tính, ta biết được số lượng mẫu với có giá trị đó và biết tương đối trong đó có bao nhiêu mẫu có nhãn tương ứng.

Dưới đây là đồ thị phân bố giá trị của thuộc tính theo nhãn là Churn



- Dựa vào các đồ thị phân tán giữa các thuộc tính ta thấy được có các cặp thuộc tính tương quan: (Day Mins, Day Charge), (Eve Mins, Eve Charge), (Night Mins, Night Charge), (Intl Mins, Intl Charge).

1.2. Các khái niệm phân cấp

Ở đây phân cấp sau khi đã chuẩn hóa dữ liệu bằng phương pháp Standadize

- Thuộc tính Account Length:

  Phân thành 3 cấp:

  - ✓ 1: (-inf – -0.487]
  - ✓ 2: (-0.487 – 1.539]
  - ✓ 3: (-1.539 – inf)

- Thuộc tính Vmail Message:

  Phân thành 3 cấp:

  - ✓ 1: (-inf – 0.65]
  - ✓ 2: (0.65 – 1.892]
  - ✓ 3: (1.892 – inf)

- Thuộc tính Day Mins:

  Phân thành 3 cấp:

  - ✓ 1: (-inf – -1.154]
  - ✓ 2: (-1.154 – 0.993]
  - ✓ 3: (0.993 – inf)

- Thuộc tính Day Calls:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – -2.264]
  - ✓ 2: (-2.264 – 0.477]
  - ✓ 3: (0.477– inf)

- Thuộc tính Eve Mins:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – -1.572]
  - ✓ 2: (-1.572 – 0.818]
  - ✓ 3: (0.818 – inf)

- Thuộc tính Eve Calls:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – -2.181]
  - ✓ 2: (-2.181 – 0.664]
  - ✓ 3: (0.664 – inf)

- Thuộc tính Night Mins:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – -1.063]
  - ✓ 2: (-1.063 – 1.388]
  - ✓ 3: (1.388 – inf)

- Thuộc tính Night Calls:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – -1.011]
  - ✓ 2: (-1.011 – 1.408]
  - ✓ 3: (1.408 – inf)

- Thuộc tính Intl Mins:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – -1.279]
  - ✓ 2: (-1.279 – 1.109]
  - ✓ 3: (1.109 – inf)

- Thuộc tính Intl Calls:
  Phân thành 3 cấp:
  - ✓ 1: (-inf – 0.889]
  - ✓ 2: (0.889 – 3.597]
  - ✓ 3: (3.597 – inf)

- Thuộc tính CustServ Calls:

Phân thành 3 cấp:

- ✓ 1: (-inf – 1.092]
- ✓ 2: (1.092 – 3.373]
- ✓ 3: (3.373 – inf)

2. Experiments

2.1. Mục đích phân tích dữ liệu

- Việc phân tích dữ liệu nhằm hiểu rõ hơn về dữ liệu. Phân tích dữ liệu nhằm khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định.
- Việc phân tích dữ liệu có phục vụ ra quyết định. Những quyết định được đưa ra dựa vào dữ liệu:

+ Mô tả: có khả năng đặc trưng hóa các thuộc tính chung của dữ liệu được khái phá
+ Dự đoán: có khả năng suy luận từ dữ liệu hiện có để dự đoán

- Phục vụ các tác vụ:

+ Phân lớp: phân loại những quan sát/thể hiện khác nhau thành những lớp cho trước (phân lớp là có giám sát).
+ Dự đoán:

- ✓ Tương tự phân lớp, ngoại trừ phân loại dựa theo một số gái trị dự đoán và ước lượng tương lai.
- ✓ Trong sự đoán, dữ liệu lịch sử được sử dụng để xây dựng một mô hình giải thích thái độ quan sát hiện tại. Mô hình có thể được áp dụng đến một thể hiện mới để dự đoán hành vi tương lại hay dự đoán giá trị tương lại của một số biến bị thiếu

+ Nhóm quan hệ:

- ✓ Xác định items nào thường đi cùng với nhau
- ✓ Thường đề cập như là phân tích rõ thị trường.

+ Gom nhóm: giống như phân lớp, gom nhóm sẽ tổ chức dữ liệu thành các lớp
+ Tổng quát hóa và phân biệt hóa: là việc tổng hợp những đặc điểm chung của các đối tượng trong cùng một lớp mục tiêu

2.2. Các thử nghiệm

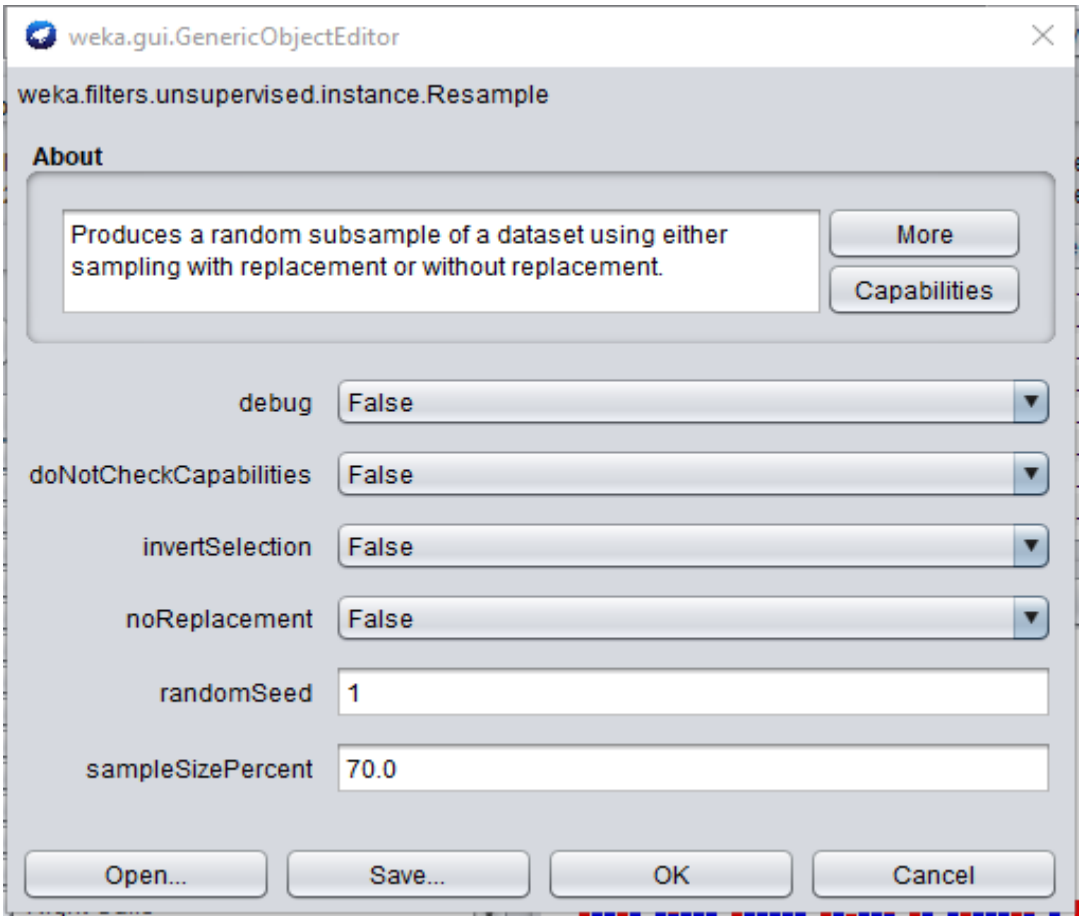
**** Thử nghiệm 1:***

- Thể hiện dữ liệu: Sử dụng dữ liệu churn_experiment1.arff trong mục data. Đây là tập dữ liệu gốc, chỉ qua 2 bước xử lý đơn giản để có thể sử dụng được thuật toán khai thác dữ liệu.
- Các phương pháp tiền xử lý:

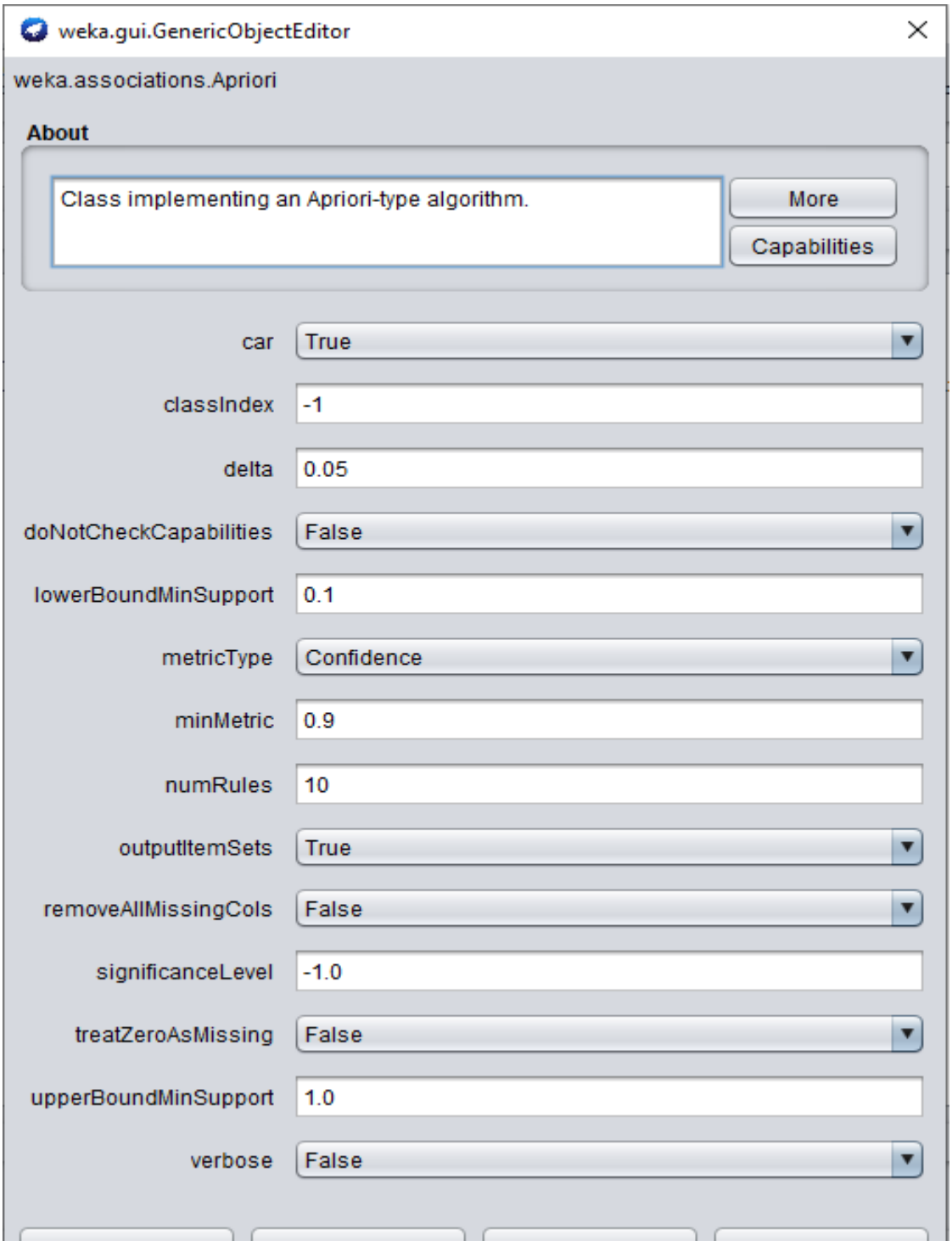
+ Chuyển đổi dữ liệu: Đổi các thuộc tính numeric thành nominal để có thể sử dụng thuật toán Apriori khai thuật luật kết hợp:

Trong Filter -> Choose -> weka -> filters -> unsupervised -> attribute -> numerictonominal -> apply

+ Rút gọn dữ liệu: Các cơ sở dữ liệu của chúng ta rất lớn, không thể thao tác trực tiếp được. Các kỹ thuật rút gọn dữ liệu được áp dụng để tiền xử lý dữ liệu:
Trong Filter -> Choose -> weka -> filters -> unsupervised -> instance -> resample và chỉnh các thông số như sau:



Sau đó nhấn apply
- Trong đó:
+ noReplacement: True nghĩa là không có thay thế ngược lại là có. Ở đây chọn False nghĩa là chọn ngẫu nhiên có thay thế.
+ sampleSizePercent: kích cỡ tập dữ liệu con cần lấy. Ở đây là 70 tức là lấy 70% từ tập gốc.
- Tham số của hệ thống, hệ số, độ đo sử dụng: Sử dụng thuật toán Apriori để khai thác luật kết hợp.
Trong tab Associate -> Choose -> Apriori và thay đổi các thông số như sau:



Sau đó nhấn Start

Trong đó:

+ car: khai phá luật kết hợp phân lớp. Ở đây bằng True tức là khai phá luật kết hợp phân lớp

+ classindex: index của lớp dùng trong trường hợp car = true, -1 ở đây là lớp cuối cùng

+ lowerBoundMinSupport: cận dưới độ hỗ trợ tối thiểu. Nên chọn từ 0.1 trở lên. Ở đây, dùng 0.1
+ metricType: dạng thang đo độ tin cậy của thuật giải. Ở đây ta sử dụng: Confidence
+ minMetric: số điểm tối thiểu chấp nhận được của thang đo (tức là độ tin cậy). Ở đây là 0.9
+ numRules: số luật cần tìm. Ở đây là 10
+ outputitemSets: hiển thị tập dữ liệu. Ở đây dùng True để có thể quan sát tập dữ liệu lấy được
+ removeAllMissingCols: loại bỏ các cột không chứa giá trị
+ significanceLevel: mức ý nghĩa, chỉ hoạt động với metric type là Confidence
+ treatZeroAsMissing: loại bỏ giá trị đầu tiên mỗi row
+ upperBoundMinSupport: cận trên độ hỗ trợ tối thiểu
+ verbose: chạy chế độ hiển thị chi tiết quá trình
- Phương pháp hậu xử lý: Sử dụng các cách tối ưu luật đã học. Vì để chọn ra được tập luật tốt hơn (không có các luật thừa, các luật là luật đủ, luật không mâu thuẫn với nhau, không có suy diễn bắc cầu). Cách tối ưu:
+ Loại bỏ các tiền đề luật không cần thiết.
+ Loại bỏ các luật thừa, luật không cần thiết.
- Kết quả thực nghiệm và ý nghĩa:
+ Các luật thu được:

```
Best rules found:

 1. Int'l Plan=no VMail Plan=yes 542 ==> Churn?=False. 521    conf:(0.96)
 2. Area Code=415 Int'l Plan=no VMail Plan=yes 277 ==> Churn?=False. 265    conf:(0.96)
 3. Area Code=415 VMail Plan=yes 311 ==> Churn?=False. 290    conf:(0.93)
 4. VMail Plan=yes 608 ==> Churn?=False. 563    conf:(0.93)
 5. Int'l Plan=no CustServ Calls=1 714 ==> Churn?=False. 661    conf:(0.93)
 6. Int'l Plan=no CustServ Calls=0 473 ==> Churn?=False. 432    conf:(0.91)
 7. Int'l Plan=no CustServ Calls=3 286 ==> Churn?=False. 261    conf:(0.91)
 8. Int'l Plan=no VMail Plan=no CustServ Calls=1 517 ==> Churn?=False. 468    conf:(0.91)
 9. Int'l Plan=no VMail Message=0 CustServ Calls=1 517 ==> Churn?=False. 468    conf:(0.91)
10. Int'l Plan=no VMail Plan=no VMail Message=0 CustServ Calls=1 517 ==> Churn?=False. 468    conf:(0.91)
```

Nhận thấy có vài luật bị dư thừa nên ta tiến hành tối ưu luật:

- ✓ Quan sát 4 luật đầu tiên ta thấy chỉ cần Vmail Plan = yes -> Churn = False nên ta sẽ loại bỏ 3 luật đầu tiên, giữ lại luật thứ 4.
- ✓ Quan sát 4 luật 5, 8, 9, 10 ta thấy luật thứ 8, 9, 10 bị dư thừa -> bỏ

→ Tập luật thu được:

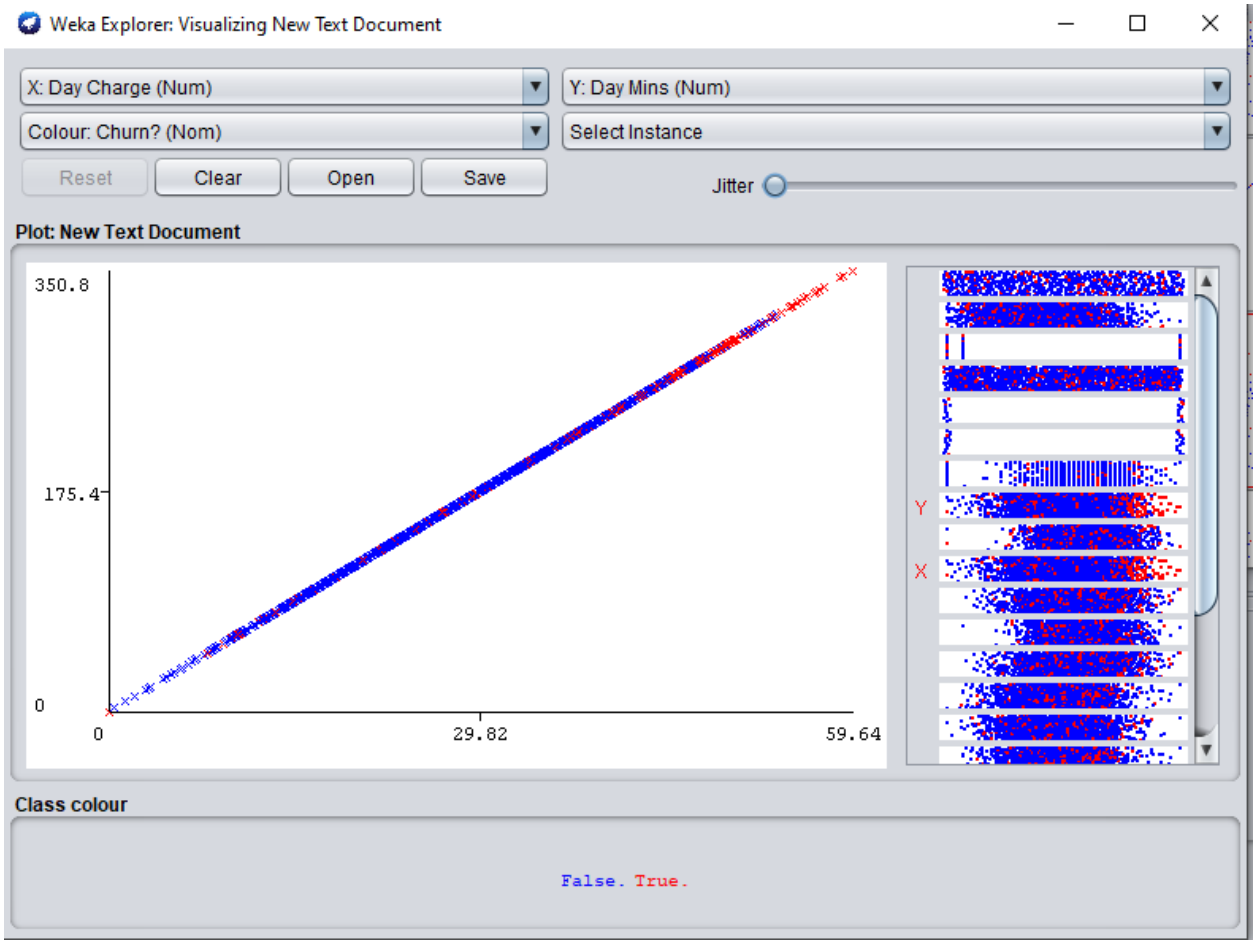
- ✓ VMail Plan=yes ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls=1 ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls=0 ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls=3 ==> Churn?=False

**** Thử nghiệm 2:***

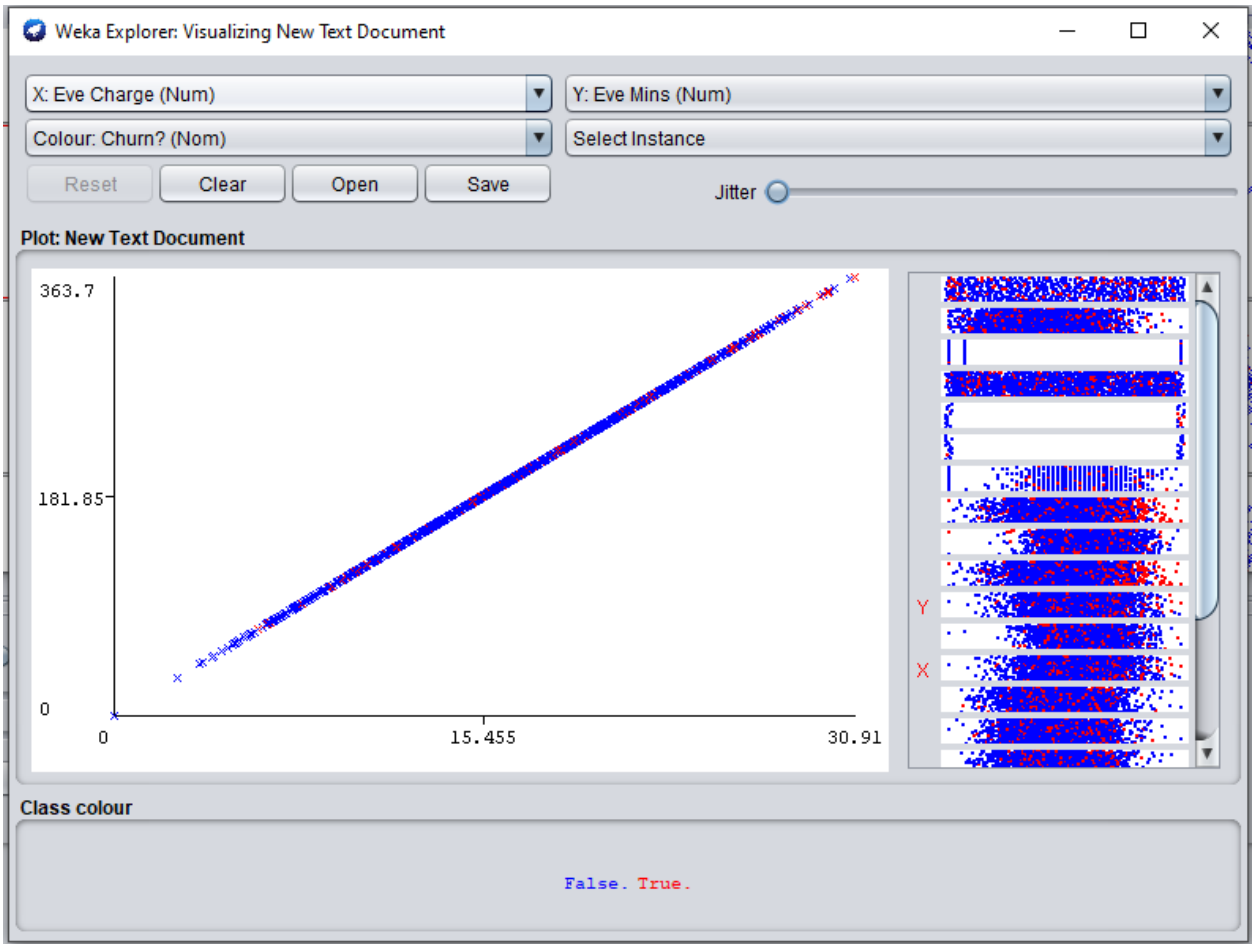
- Thể hiện dữ liệu: Sử dụng dữ liệu churn_experiment2.arff trong mục data. Đây là tập dữ liệu đã qua vài bước tiền xử lý (sẽ được nêu bên dưới).
- Các phương pháp tiền xử lý:

+ Chọn lọc thuộc tính: Ở bước này ta sẽ loại bỏ 1 vài thuộc tính không cần thiết.

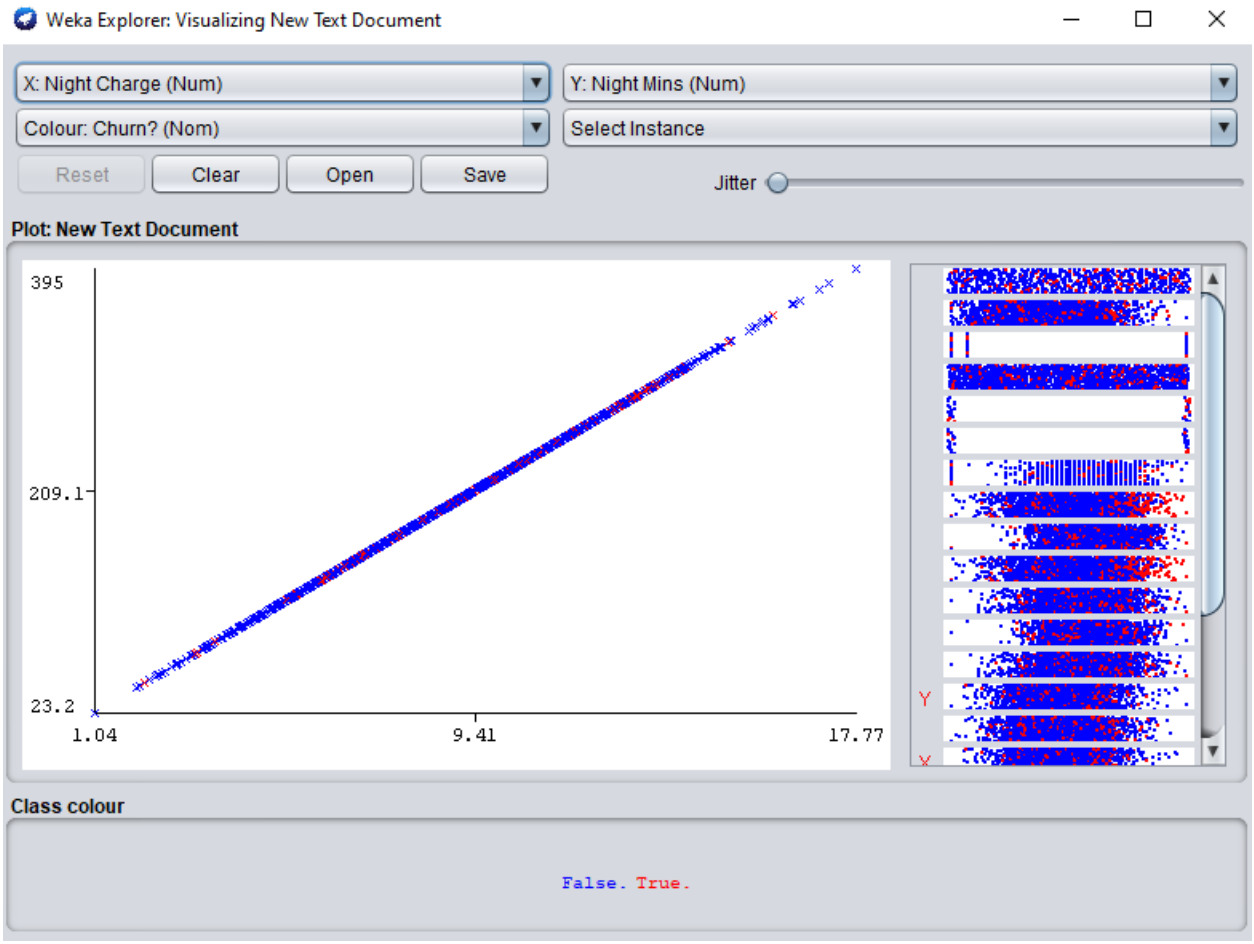
- ✓ Xem xét mối quan hệ giữa các cặp thuộc tính: Day Mins – Day Charge, Eve Mins – Eve Charge, Night Mins – Night Charge, Intl Mins – Intl Charge. Ta thấy các thuộc tính Charge là hàm của các thuộc tính Mins và có thể biểu diễn dưới dạng các đường hồi quy tuyến tính bậc nhất.
  Day Mins – Day Charge:



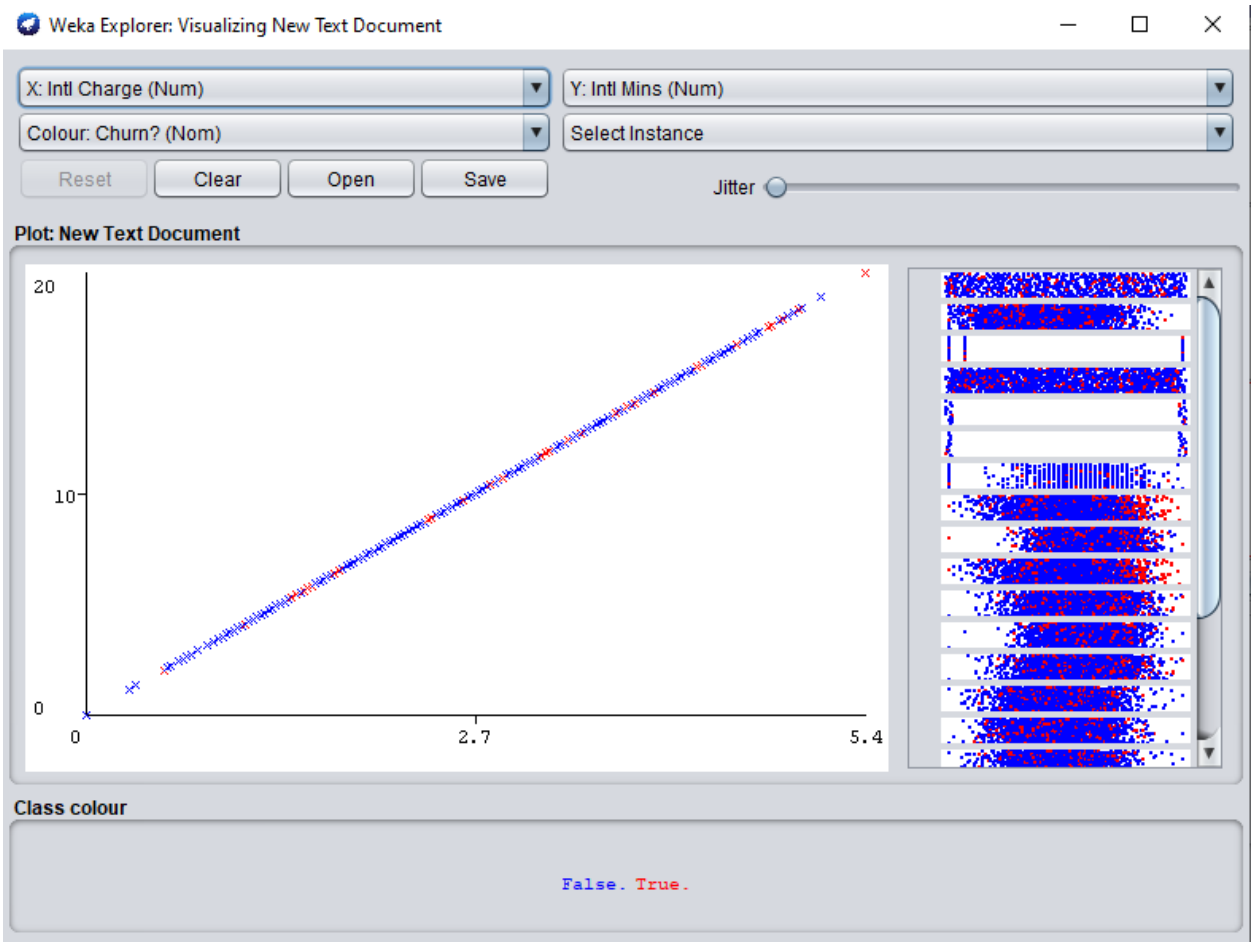
Eve Mins – Eve Charge:



Night Mins – Night Charge:



Intl Mins – Intl Charge:



→ Loại bỏ các thuộc tính Charge để kết quả phân tích được tốt hơn.

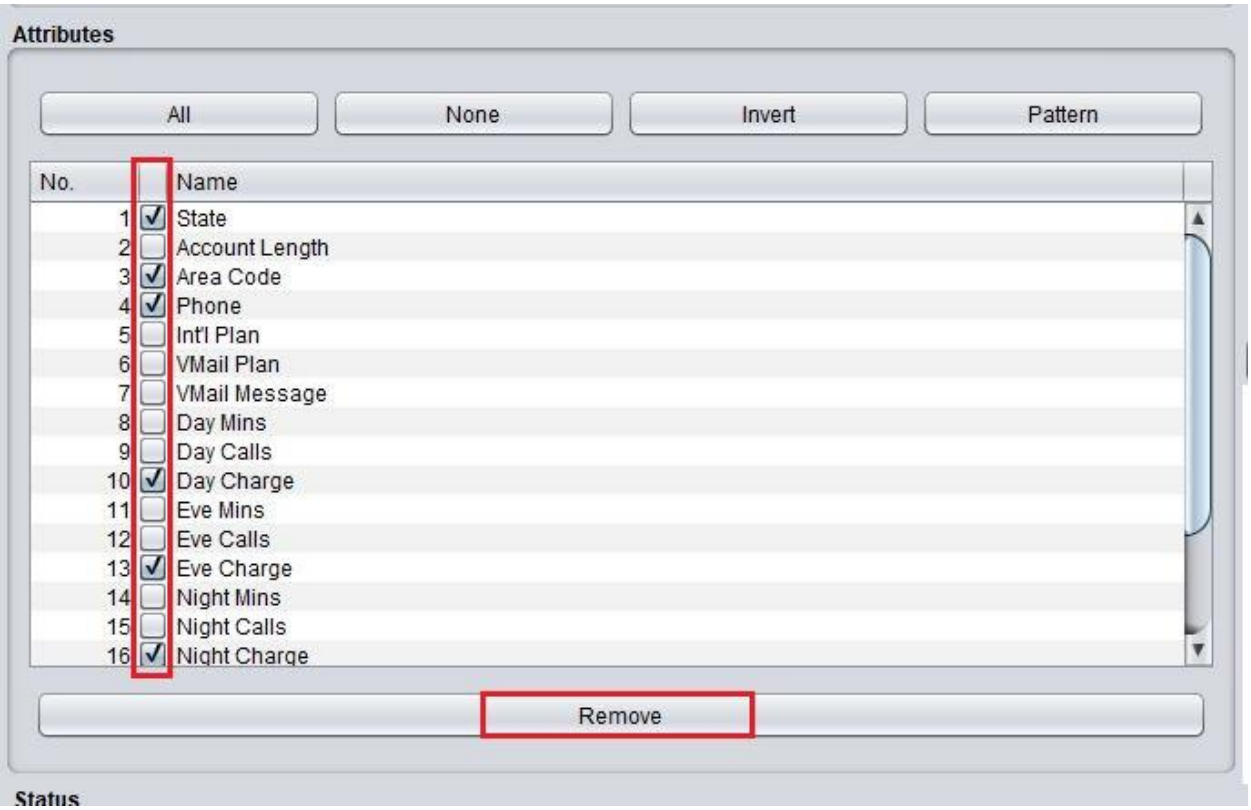
✓ Loại bỏ những thuộc tính dị thường:

- Thuộc tính Area Code: Chỉ có 3 giá trị nhưng đều là ở California
- Thuộc tính State: 3 mã Area code được sử dụng cho cả 50 bang

✓ Loại bỏ thuộc tính Phone vì thuộc tính này như là ID khách hàng.

→ Loại bỏ được 7 thuộc tính.

→ Trong tab Preprocess tích chọn 7 thuộc tính đã nêu trên ở Attributes sau đó nhấn Remove để loại bỏ.

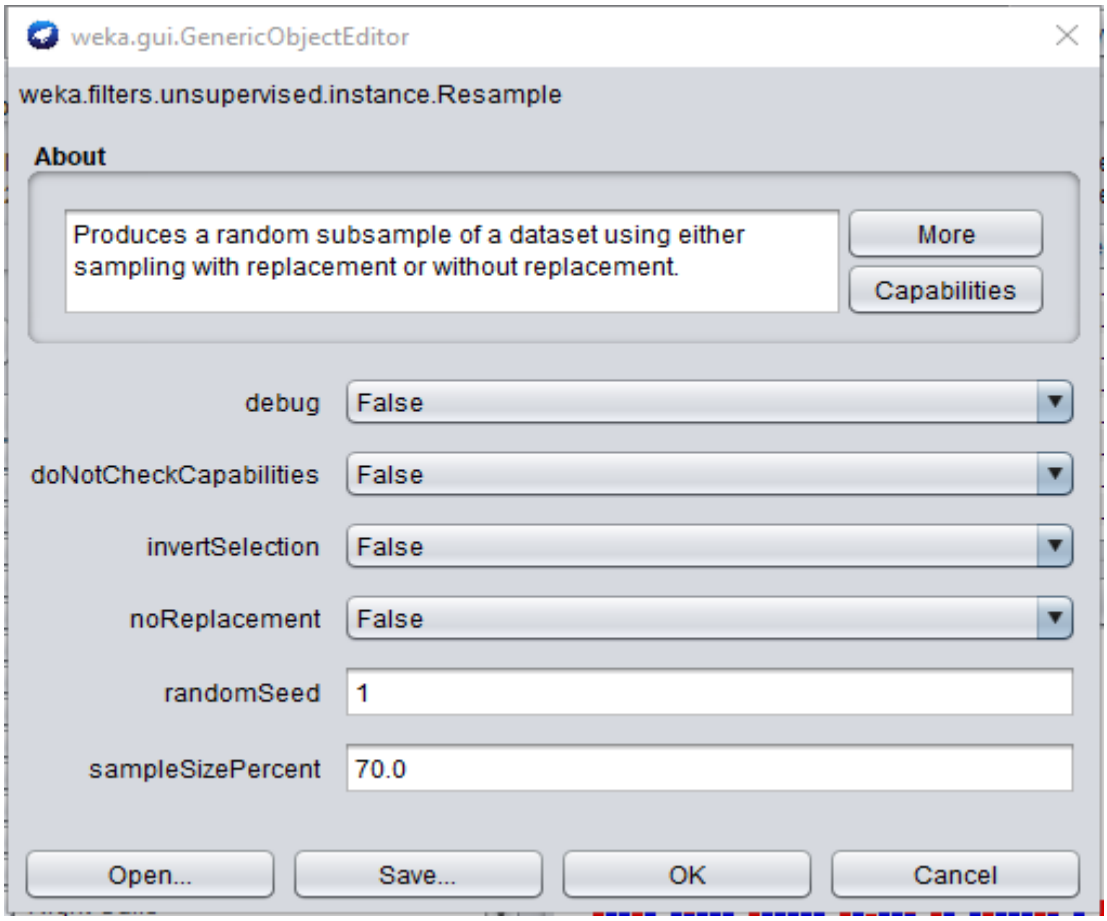


+ Chuyển đổi dữ liệu: Đổi các thuộc tính numeric thành nominal để có thể sử dụng thuật toán Apriori khai thuật luật kết hợp:

Trong Filter -> Choose -> weka -> filters -> unsupervised -> attribute -> numerictonominal -> apply

+ Rút gọn dữ liệu: Các cơ sở dữ liệu của chúng ta rất lớn, không thể thao tác trực tiếp được. Các kỹ thuật rút gọn dữ liệu được áp dụng để tiền xử lý dữ liệu:

Trong Filter -> Choose -> weka -> filters -> unsupervised -> instance -> resample và chỉnh các thông số như sau:



Sau đó nhấn apply

- Trong đó:

+ noReplacement: True nghĩa là không có thay thế ngược lại là có. Ở đây chọn False nghĩa là chọn ngẫu nhiên có thay thế.

+ sampleSizePercent: kích cỡ tập dữ liệu con cần lấy. Ở đây là 70 tức là lấy 70% từ tập gốc.

- Tham số của hệ thống, hệ số, độ đo sử dụng: Sử dụng thuật toán Apriori để khai thác luật kết hợp. (Giống ở thử nghiệm 1)

- Phương pháp hậu xử lý: Sử dụng các cách tối ưu luật đã học. Vì để chọn ra được tập luật tốt hơn (không có các luật thừa, các luật là luật đủ, luật không mâu thuẫn với nhau, không có suy diễn bắc cầu). Cách tối ưu:

+ Loại bỏ các tiền đề luật không cần thiết.

+ Loại bỏ các luật thừa, luật không cần thiết.

- Kết quả thực nghiệm và ý nghĩa:

+ Các luật thu được:

```
Best rules found:

 1. Int'l Plan=no VMail Plan=yes 542 ==> Churn?=False. 521    conf:(0.96)
 2. VMail Plan=yes 608 ==> Churn?=False. 563    conf:(0.93)
 3. Int'l Plan=no CustServ Calls=1 714 ==> Churn?=False. 661    conf:(0.93)
 4. Int'l Plan=no CustServ Calls=0 473 ==> Churn?=False. 432    conf:(0.91)
 5. Int'l Plan=no CustServ Calls=3 286 ==> Churn?=False. 261    conf:(0.91)
 6. Int'l Plan=no VMail Plan=no CustServ Calls=1 517 ==> Churn?=False. 468    conf:(0.91)
 7. Int'l Plan=no VMail Message=0 CustServ Calls=1 517 ==> Churn?=False. 468    conf:(0.91)
 8. Int'l Plan=no VMail Plan=no VMail Message=0 CustServ Calls=1 517 ==> Churn?=False. 468    conf:(0.91)
 9. Int'l Plan=no CustServ Calls=2 464 ==> Churn?=False. 419    conf:(0.9)
10. CustServ Calls=3 315 ==> Churn?=False. 284    conf:(0.9)
```

Nhận thấy có vài luật bị dư thừa nên ta tiến hành tối ưu luật:

- ✓ Quan sát 2 luật đầu tiên ta thấy chỉ cần Vmail Plan = yes -> Churn = False nên ta sẽ loại bỏ luật đầu tiên, giữ lại luật thứ 2.
- ✓ Quan sát 4 luật 3, 6, 7, 8 ta thấy luật thứ 6, 7, 8 bị dư thừa -> bỏ
- ✓ Quan sát 2 luật 5, 10 ta thấy luật thứ 5 bị dư thừa -> bỏ

→ Tập luật thu được:

- ✓ VMail Plan=yes ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls=1 ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls=0 ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls=2 ==> Churn?=False
- ✓ CustServ Calls=3 ==> Churn?=False

**** Thử nghiệm 3:***

- Thể hiện dữ liệu: Sử dụng dữ liệu churn_experiment3.arff trong mục data. Đây là tập dữ liệu đã qua các bước tiền xử lý (sẽ được nêu ở bên dưới).
- Các phương pháp tiền xử lý:

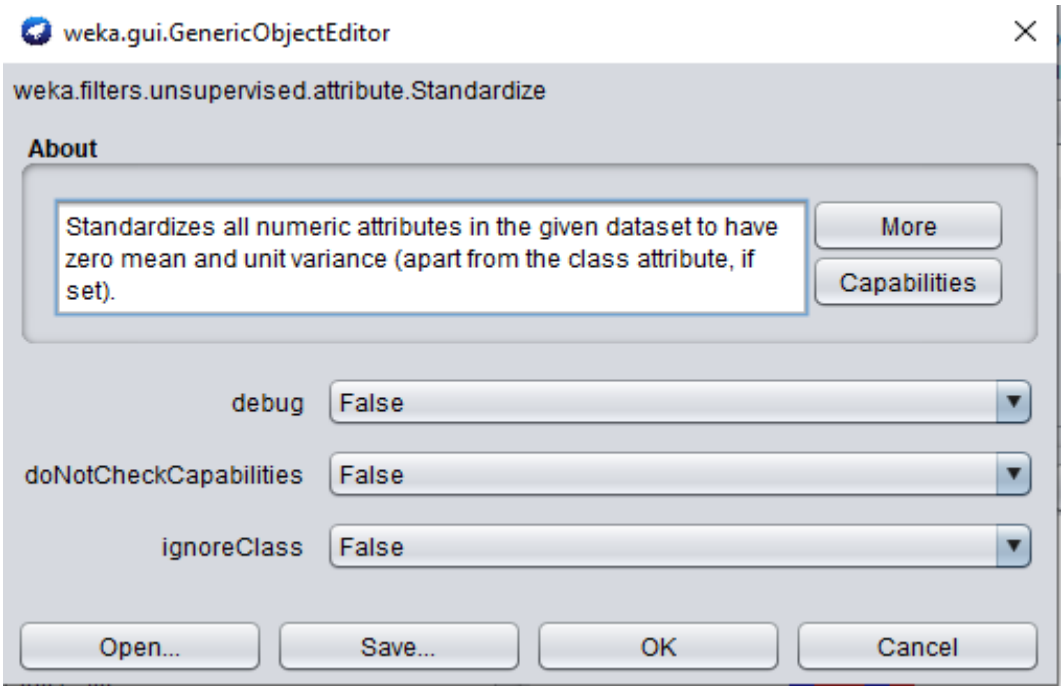
+ Chọn lọc thuộc tính: Ở bước này ta sẽ loại bỏ 1 vài thuộc tính không cần thiết. (Đã nêu ở trên thử nghiệm 2).

+ Chuẩn hóa dữ liệu: Ở bước này ta sẽ chuẩn hóa dữ liệu bằng bộ lọc Standardize. Dùng để thay đổi giá trị của các thuộc tính số sao cho chúng có giá trị trung bình là 0 và độ lệch chuẩn là 1. Cụ thể, với mỗi thuộc tính, ta tính giá trị trung bình μ và độ lệch chuẩn σ, sau đó mỗi giá trị x sẽ được thay thế bằng:

$$X = \frac{x - \mu}{\sigma}$$

Bộ lọc này giả định rằng dữ liệu có phân phối chuẩn. Một điểm mạnh của lựa chọn này là nó sẽ không quá bị ảnh hưởng bởi giá trị nhiễu.

Trong tab Preprocessing, ở khung Filter -> Choose -> weka -> filters -> unsupervised -> attribute -> standardize. Và chỉnh các thông số như sau:

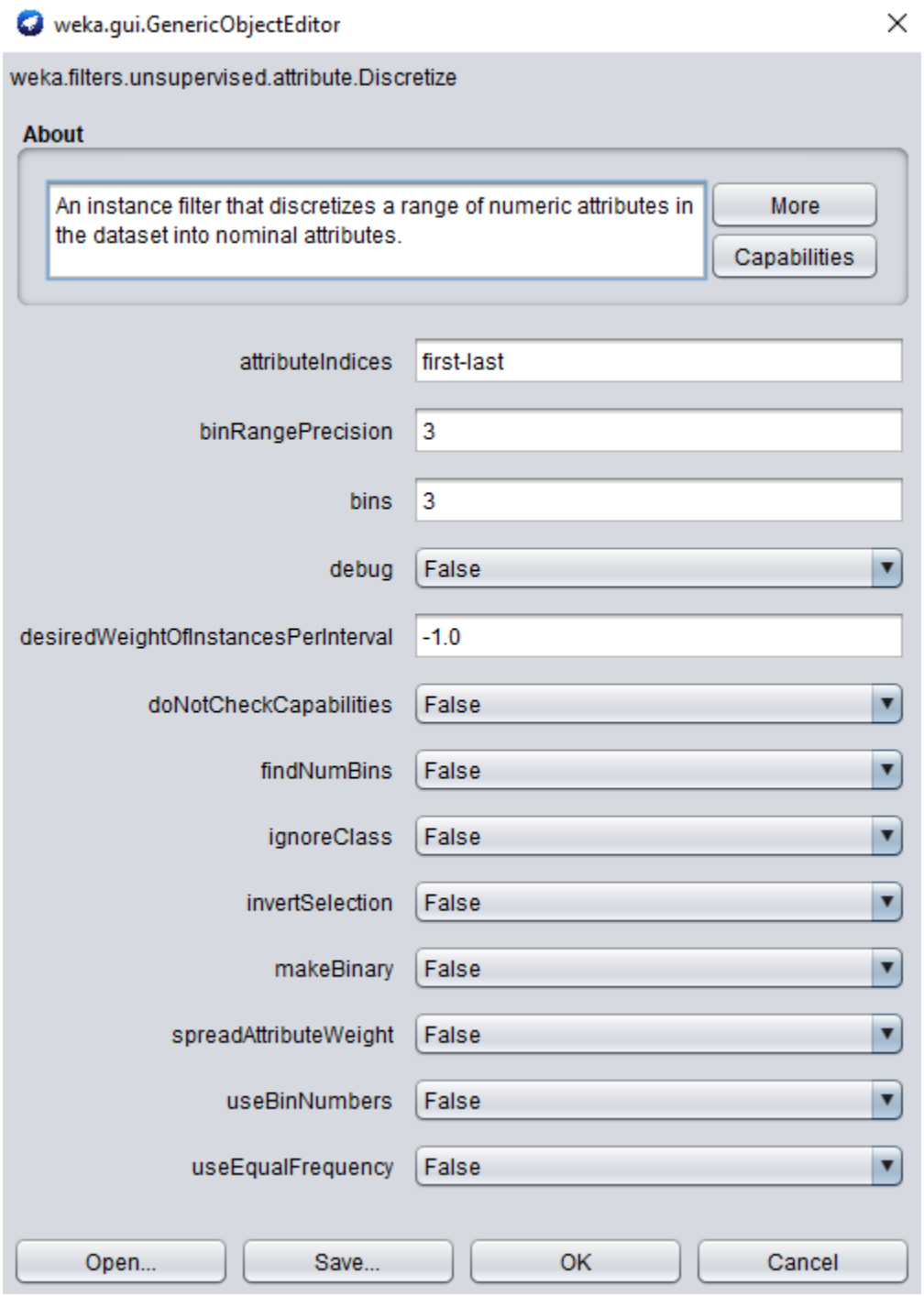


Sau đó nhấn apply

+ Rút gọn dữ liệu: Các cơ sở dữ liệu của chúng ta rất lớn, không thể thao tác trực tiếp được. Các kỹ thuật rút gọn dữ liệu được áp dụng để tiền xử lý dữ liệu: Rời rạc hóa thuộc tính numeric thành nominal và lấy tập con ngẫu nhiên của dữ liệu.

- ✓ Trong Weka, dùng bộ lọc Discretize - bộ lọc dùng để rời rạc hóa các thuộc tính numeric thành nominal. Việc rời rạc đơn giản bằng cách chia giỏ (bining), sắp xếp và chia dữ liệu vào các giỏ có cùng độ rộng. Chia vùng giá trị thành N khoảng cùng kích thước, độ rộng của từng khoảng = (max – min)/N

Trong tab Preprocessing, ở khung Filter -> Choose -> weka -> filters -> unsupervised -> attribute -> Discretize. Và chỉnh các thông số như sau:



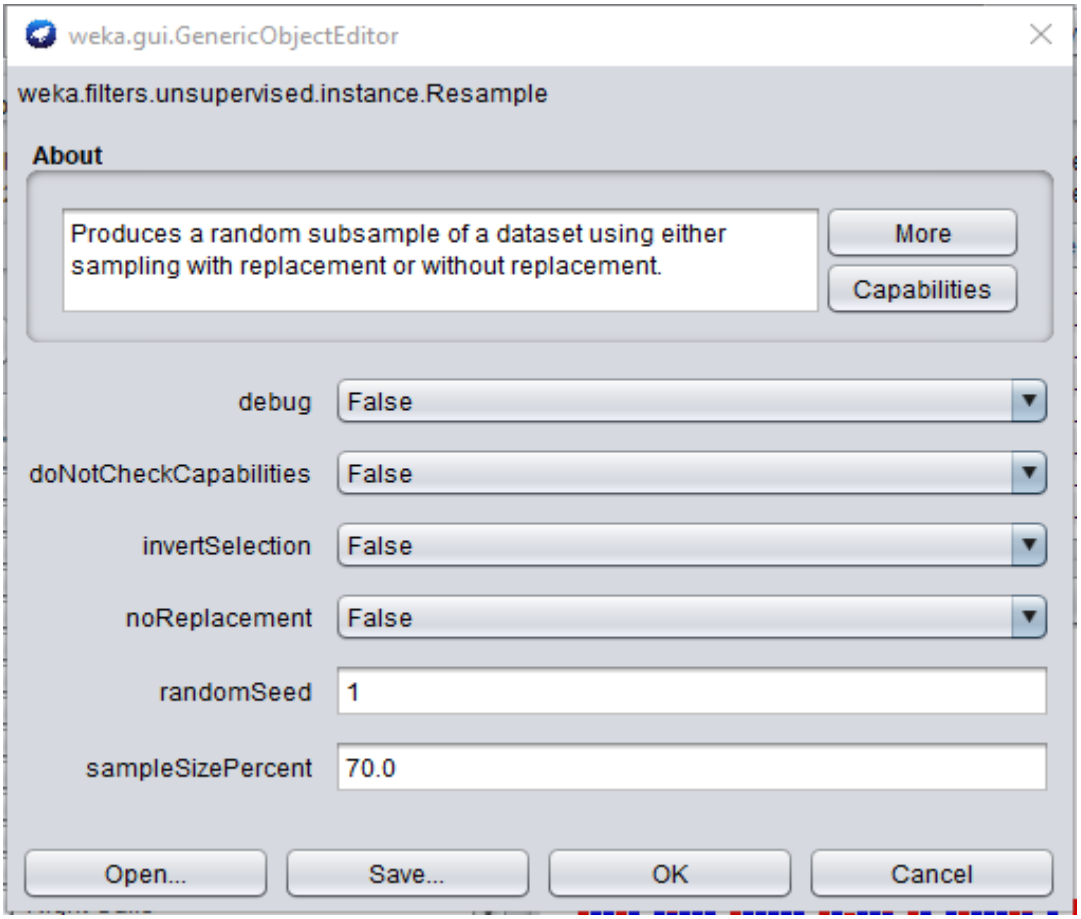
Sau đó nhấn apply

Trong đó:

+ attributeIndices: những thuộc tính được lựa chọn để áp dụng bộ lọc
+ binRangePrecision: số chữ số phần thập phân
+ bins: số khoảng cần chia

- ✓ Trong Weka để chọn tập con ngẫu nhiên của dữ liệu ta sẽ dùng bộ lọc resample

Trong Filter -> Choose -> weka -> filters -> unsupervised -> instance -> resample và chỉnh các thông số như sau:



Sau đó nhấn apply

Trong đó:

+ noReplacement: True nghĩa là không có thay thế ngược lại là có. Ở đây chọn False nghĩa là chọn ngẫu nhiên có thay thế.

+ sampleSizePercent: kích cỡ tập dữ liệu con cần lấy. Ở đây là 70 tức là lấy 70% từ tập gốc.

- Tham số của hệ thống, hệ số, độ đo sử dụng: Sử dụng thuật toán Apriori để khai thác luật kết hợp. (Giống ở thử nghiệm 1).

- Phương pháp hậu xử lý: Sử dụng các cách tối ưu luật đã học. Vì để chọn ra được tập luật tốt hơn (không có các luật thừa, các luật là luật đủ, luật không mâu thuẫn với nhau, không có suy diễn bắc cầu). Cách tối ưu:

+ Loại bỏ các tiền đề luật không cần thiết.

+ Loại bỏ các luật thừa, luật không cần thiết.

- Kết quả thực nghiệm và ý nghĩa:

+ Các luật thu được:

```
Best rules found:

 1. Int'l Plan=no Day Mins='(-1.154-0.993]' CustServ Calls='(-inf-1.092]' 1363 ==> Churn?=False. 1315    conf:(0.96)
 2. Int'l Plan=no Eve Mins='(-1.572-0.818]' CustServ Calls='(-inf-1.092]' 1404 ==> Churn?=False. 1315    conf:(0.94)
 3. Day Mins='(-1.154-0.993]' CustServ Calls='(-inf-1.092]' 1495 ==> Churn?=False. 1397    conf:(0.93)
 4. Int'l Plan=no Day Mins='(-1.154-0.993]' 1491 ==> Churn?=False. 1390    conf:(0.93)
 5. Int'l Plan=no Night Calls='(-1.011-1.408]' CustServ Calls='(-inf-1.092]' 1484 ==> Churn?=False. 1365    conf:(0.92)
 6. Int'l Plan=no CustServ Calls='(-inf-1.092]' 1937 ==> Churn?=False. 1773    conf:(0.92)
 7. Int'l Plan=no Intl Calls='(-inf-0.889]' CustServ Calls='(-inf-1.092]' 1579 ==> Churn?=False. 1444    conf:(0.91)
 8. Int'l Plan=no Intl Mins='(-1.279-1.109]' CustServ Calls='(-inf-1.092]' 1502 ==> Churn?=False. 1373    conf:(0.91)
 9. Int'l Plan=no Eve Calls='(-2.181-0.664]' CustServ Calls='(-inf-1.092]' 1440 ==> Churn?=False. 1314    conf:(0.91)
10. Eve Mins='(-1.572-0.818]' CustServ Calls='(-inf-1.092]' 1541 ==> Churn?=False. 1406    conf:(0.91)
```

Nhận thấy có vài luật bị dư thừa nên ta tiến hành tối ưu luật:

- ✓ Quan sát các luật: 1, 2, 5, 6, 7, 8, 9 ta thấy các luật 1, 2, 5, 7, 8, 9 dư thừa -> bỏ

→ Tập luật thu được:

- ✓ Day Mins='(-1.154-0.993]' CustServ Calls='(-inf-1.092]' ==> Churn?=False
- ✓ Int'l Plan=no Day Mins='(-1.154-0.993]' ==> Churn?=False
- ✓ Int'l Plan=no CustServ Calls='(-inf-1.092]' ==> Churn?=False
- ✓ Eve Mins='(-1.572-0.818]' CustServ Calls='(-inf-1.092]' ==> Churn?=False

3. Tóm tắt kết quả

**** Cách đánh giá kết quả và tiêu chí đánh giá kết quả:***

Chọn kết quả của thử nghiệm 3 vì tập dữ liệu này đã được tiền xử lý -> cải thiện chất lượng dữ liệu -> cải thiện chất lượng của kết quả khai phá.

**** Tập luật tốt nhất đạt được và mô tả:***

- Day Mins='(-1.154-0.993]' CustServ Calls='(-inf-1.092]' ==> Churn?=False

→ Các khách hàng có Day Mins nằm trong khoảng (-1.154-0.993] và có số CustServ Calls ≤ 1.092 sẽ không bỏ công ty

- Int'l Plan=no Day Mins='(-1.154-0.993]' ==> Churn?=False

→ Các khách hàng không đăng ký dịch vụ Int'l Plan và có Day Mins nằm trong khoảng (-1.154-0.993] sẽ không bỏ công ty

- Int'l Plan=no CustServ Calls='(-inf-1.092]' ==> Churn?=False

→ Các khách hàng không đăng ký dịch vụ Int'l Plan có số CustServ Calls ≤ 1.092 sẽ không bỏ công ty

- Eve Mins='(-1.572-0.818]' CustServ Calls='(-inf-1.092]' ==> Churn?=False

→ Các khách hàng có Eve Mins nằm trong khoảng (-1.572-0.818] và có số CustServ Calls ≤ 1.092 sẽ không bỏ công ty

**** Điểm mạnh và điểm yếu trong bài tập này:***

- Điểm mạnh: Có khả năng tìm kiếm được nhiều tài liệu tham khảo.
- Điểm yếu: Khả năng đọc hiểu tài liệu tiếng Anh còn kém.

III. TÀI LIỆU THAM KHẢO

- Slide bài giảng
- Tài liệu thầy gửi
- Text book Data Mining Concepts and Techniques
- https://www.slideshare.net/HoQuangThanh/la-chn-thuc-tnh-v-khai-ph-lut-kt-hp-trn-weka
- https://github.com/vltanh/hcmus-DataMining
- https://medium.com/@karim_ouda/tutorial-document-classification-using-weka-aa98d5edb6fa
- https://fracpete.github.io/python-weka-wrapper/api.html#classifiers
- http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/
- https://www.slideshare.net/Nilesh_raghav/frequent-itemset-mining-methods
- https://www.youtube.com/watch?v=MekiVjFgchQ&t=491s