Trường đại học Khoa học Tự nhiên – đại học Quốc gia thành phố Hồ Chí Minh

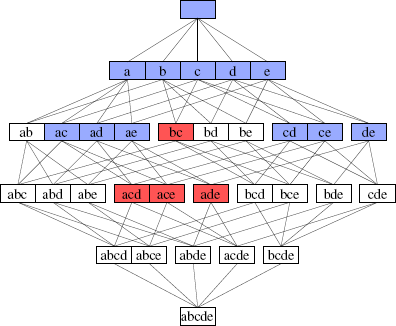
Khoa Công nghệ thông tin

---🙡🙣---

**Báo cáo thực hành**

Tập phổ biến và

khai thác luật kết hợp



Giáo viên hướng dẫn: **Nguyễn Ngọc Đức**

Lớp: **Khai thác dữ liệu và ứng dụng – CQ2017/21**

Nhóm 21: **Nguyễn Đăng Anh Thi 1612645**

**Nguyễn Huỳnh Xuân Mai 1712091**

**Âu Dương Tấn Sang 1712145**

**I. Tổng quan:**

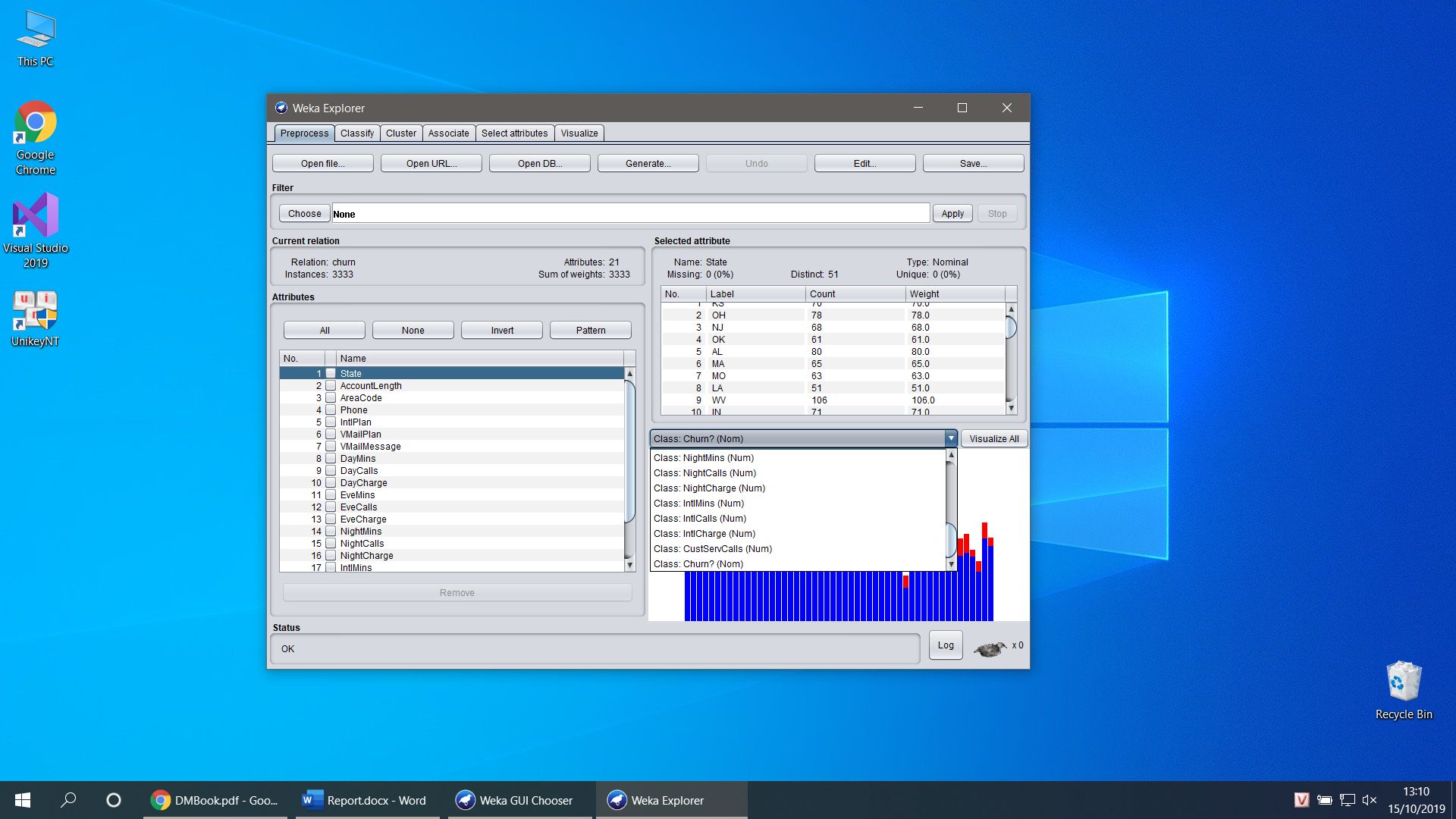
Khai thác luật kết hợp là một trong những phương thức hàng đầu để nhận dạng mẫu và khai thác tri thức từ các tập dữ liệu đủ lớn. Mục tiêu của bài thực hành này là áp dụng các thuật toán khai thác luật kết hợp (được phần mềm Weka cung cấp) để giải quyết các vấn đề thực sự trong cuộc sống.

Tình huống đặt ra là một công ty viễn thông bỗng dưng có một lượng lớn khách hành ngưng sử dụng dịch vụ nhưng lại không rõ lí do vì sao. Bộ phận khai thác dữ liệu phải tìm hiểu vấn đề này thông qua việc khai thác luật kết hợp trên tập dữ liệu của công ty được cung cấp ở đường dẫn dưới đây.

Dữ liệu: <http://ce.sharif.edu/courses/8586/1/ce925/assignments/files/assignDir2/churn.txt>

**II. Mô tả và tiền xử lý dữ liệu:**

Chuyển tập dữ liệu trên thành dạng .csv (hoặc .arff) và mở bằng Weka.



1. Mô tả dữ liệu

Tổng quan, tập dữ liệu có 21 thuộc tính và 3333 thể hiện. Dưới đây là tóm tắt sơ lược

• State: nominal, mã bang

• AccountLength: nominal, thời gian hoạt động của khách hàng

• AreaCode: norminal, mã vùng

• Phone: nominal, số điện thoại

• IntlPlan: nominal, có hoặc không đăng kí InternationalCall

• VMailPlan: nominal, có hoặc không đăng kí VoiceMail

• VMailMessage: numeric, số lượng VoiceMail

• DayMins, EveMins, NightMins, IntlMins: numeric, thời gian gọi

• DayCalls, EveCalls, NightCalls, IntlCalls: numeric, số cuộc gọi

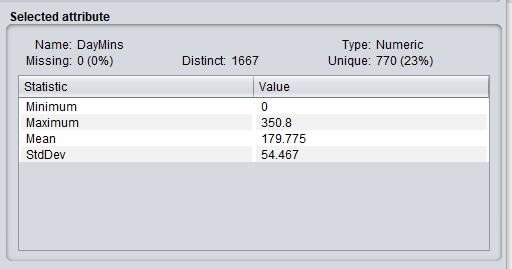
• DayCharge, EveCharge, NightCharge, IntlCharge: numeric, cước phí

• CustServCalls: numeric, số cuộc gọi đến CusomerService

• Churn: nominal, khách hàng có hoặc không ngưng sử dụng

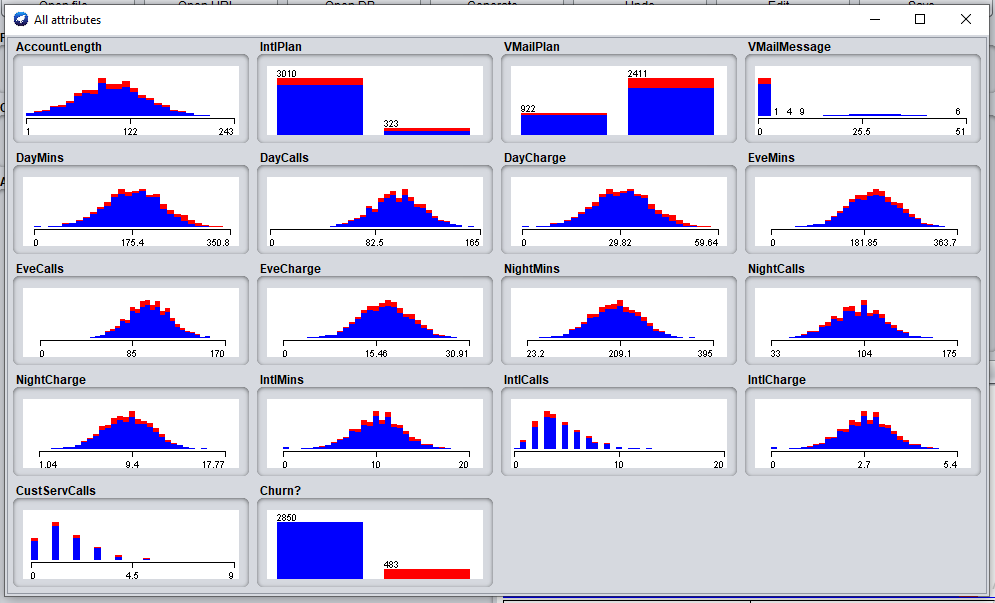
Các thuộc tính định danh như AreaCode, Phone, hay State có thể lược bỏ bớt.

Do không có giá trị nào bị khuyết, dễ dàng tính được các độ đo như trung bình mẫu, độ lệch chuẩn,…cho thuộc tính numeric và và các độ đo thống kê khác cho thuộc tính norminal. Trong đó, thuộc tính IntlPlan, VMailPlan và Churn là thuộc tính phân loại nhị phân (yes/no).



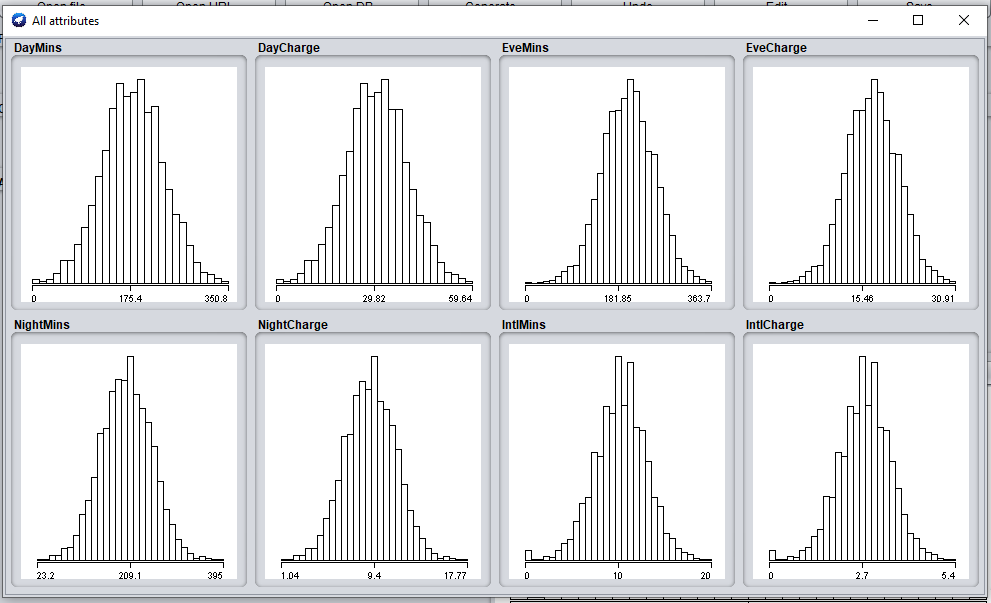
2. Thống kê

Rút ra được nhận xét: các thuộc tính DayMins, EveMins và NightMins có phân bố khá rời rạc (độ lệch chuẩn rất cao), các thuộc tính VmailMessage, IntlCharge và CustServCalls bị lệch so với phân phối chuẩn.



3. Phân phối dữ liệu

Hơn nữa, thuộc tính Charge gần như thay đổi tuyến tính với Mins, nên có thể lược đi một trong 2 loại thuộc tính này. Ở đây, ta sẽ bỏ đi thuộc tính cước phí.



4. Mins vs Charge

Tiếp theo, phải lọc bớt các thể hiện dữ liệu ngoại lai (outlier). Có nhiều phương pháp khác nhau cho thao tác này. Ở đây, nhóm sử dụng module chuẩn hóa Z đã cài đặt ở lab01 để xác định các ngoại lai này.

Đặt file churn.csv và preprocess.py vào cùng thư mục, sau đó mở command line tại thư mục này và nhập lệnh sau:

python3 preprocess.py --input churn.txt --output normalized.csv --task zScoreNorm --prop prop1 prop2 …

Trong đó prop1, prop2,…là các thuộc tính muốn chuẩn hóa. Sau đó chọn ngưỡng threshold để loại bỏ các ngoại lai, giữ lại các thể hiện có giá trị nằm trong khoảng [-threshold, threshold]:

python3 preprocess.py --input normalized.csv --output removed.csv --task removeOutlier --threshold thresholdValue --prop prop1 prop2 …

**II. Khai thác luật kết hợp:**

Với mỗi nhu cầu khác nhau, cách nhìn dữ liệu cũng khác nhau, sinh ra có khái niệm phân cấp dữ liệu. Có nhiều cách phân cấp cho một thuộc tính dữ liệu nào đó, chẳng hạn:

• Địa chỉ: quốc gia > bang > thành phố > …

• Tuổi: thanh niên < trung niên < lão niên < …

• …

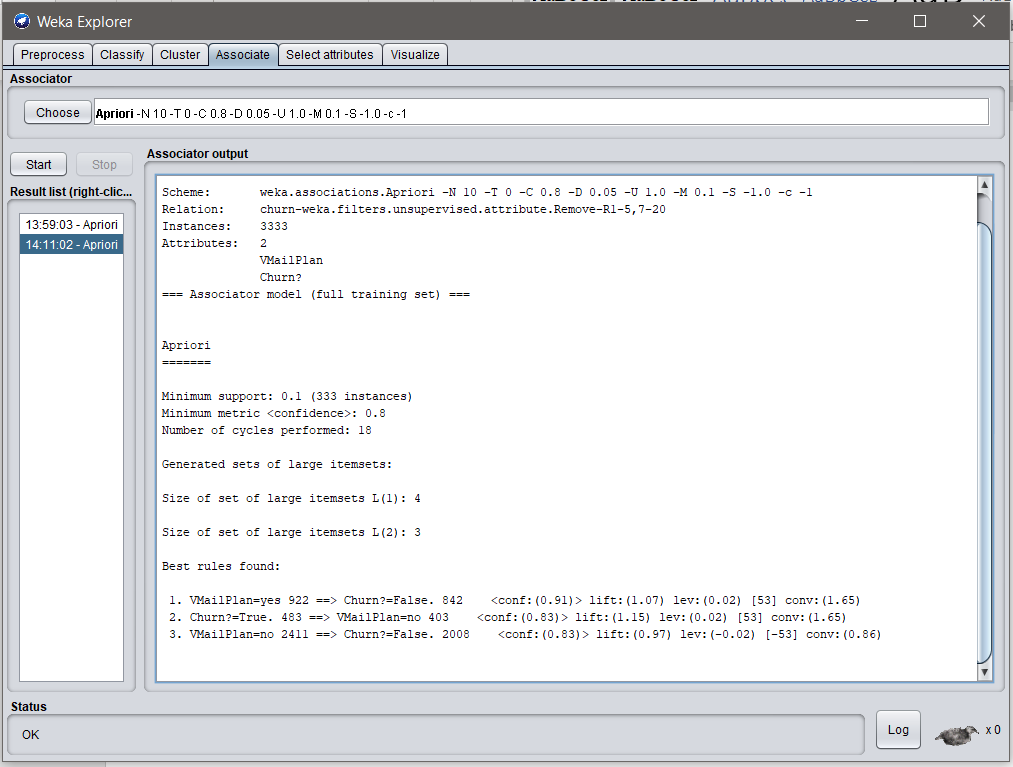
Việc phân cấp như vậy sẽ chia tập dữ liệu thành nhiều nhóm, mỗi nhóm có thể phát sinh các luật kết hợp thỏa mãn được yêu cầu của người phân tích dữ liệu.

Dưới đây là một số thử nghiệm, giúp trả lời những câu hỏi cốt yếu khi công ty xảy ra vấn đề với việc khách hàng ngưng sử dụng dịch vụ do công ty cung cấp.

**a. Tìm nguyên nhân khách hàng không muốn tiếp tục sử dụng:**

Phần lớn lí do khiến khách hàng cảm thấy không hài lòng chính là chất lượng của dịch vụ nào đó. Chính vì thế, phải xem xét sự tác động của các dịch vụ đến quyết định của khách hàng. Đó có thể là các tác động tích cực hoặc tiêu cực, từ đó công ty có thể đưa ra phương án kích cầu tiêu dùng hợp lí.

Thực hiện lọc ra thuộc tính VMailPlan và Churn, mở tab Associate và chọn thuật toán Apriori, nhập lowerBoundMinSupport là 0.1, minMetric (metric là Confidence) là 0.8, rồi bấm Start. Thu được kết quả như sau:



6. VMailPlan

Kết quả trên cho thấy, Apriori khai thác được 3 luật với minSup = 0.1 và minCon = 0.8:

• VMailPlan = yes => Churn = False <conf: 0.91>

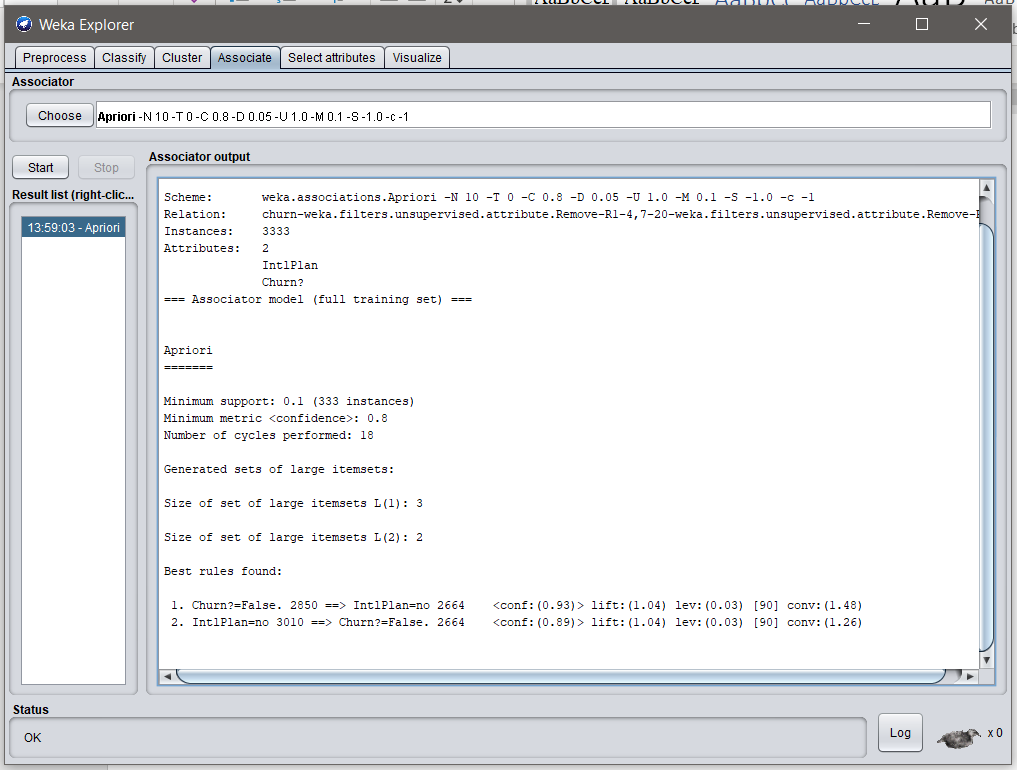
• Churn = True => VMailPlan = no <conf: 0.83>

• VMailPlan = no => Churn = False <conf: 0.83>

Có nghĩa rằng: 91% khách hàng đăng kí dịch vụ VoiceMall tiếp tục sử dụng, và những khách hàng không sử dụng nữa thì 83% là do không đăng kí VoiceMail.

***Vậy, đa phần khách hàng ngưng sử dụng là không đăng kí dịch vụ VoiceCall (83%), cần khuyến khích khách hàng sử dụng dịch vụ này***.

Tiếp theo, thử nghiệm tương tự trên thuộc tính IntlPlan, thu được kết quả:



7. IntlPlan

Thuật toán Apriori thu được 2 luật kết hợp:

• Churn = False => IntlPlan = no <conf: 0.93>

• IntlPlan = no => Churn = False <conf: 0.89>

Nghĩa là: 93% khách hàng đang sử dụng là không đăng kí InternationalCall và 89% khách hàng không đăng kí dịch vụ InternationalCall vẫn đang sử dụng. ***Vậy nên chưa thể kết luận việc ngưng sử dụng là do dịch vụ InternationalCall.***

**b. Tìm những khách hàng có khả năng sắp ngưng sử dụng (phục vụ cho tác vụ phân lớp):**

Tìm câu trả lời cho vấn đề này khó khăn hơn việc tìm dịch vụ tốt/chưa tốt như ở trên.