

- Analyse the association rules mined from the data set. What are their confidence and support values?

* Minimum support: 0.15 (2 instances)
* Minimum metric <confidence>: 0.9

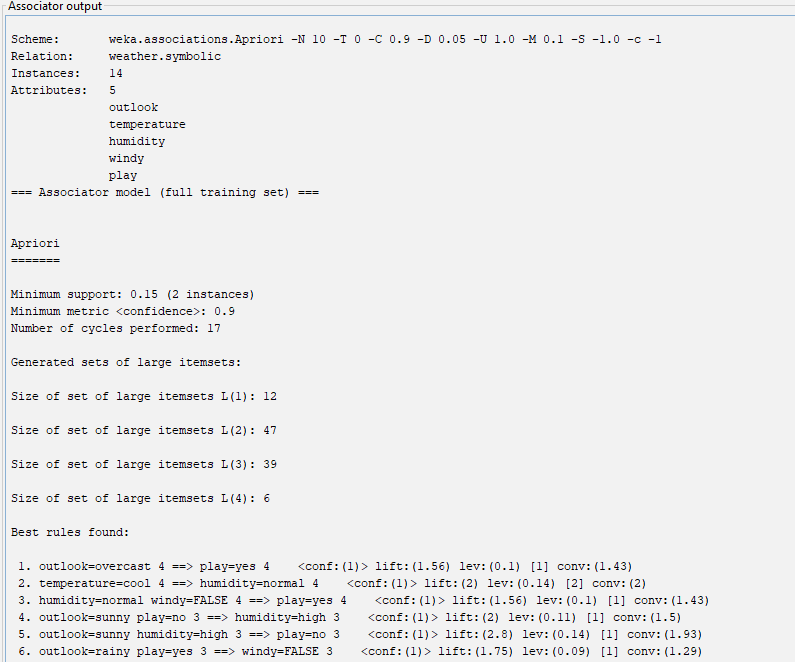
- Examine the number of large itemsets.

* Luật 1: milk=T butter=F ==> beer=F
  + Confidence: 1 (luôn đúng 100% khi "milk = T" và "butter = F", "beer = F")
  + Support: 4/15 = 0.27 (luật này xảy ra trong 4 lần)
* Luật 4: butter=F beer=T ==> milk=F
  + Confidence: 1 (luôn đúng 100%)
  + Support: 3/15 = 0.2

- Try changing different values for the following parameters \lowerBoundMinSupport " (min threshold for support), \minMetric" (min threshold for confidence). As you change these parameter values what do you notice about the rules that are found by the associator? Note that the parameter \numRules" limits the maximum number of rules that the associator looks for, you can try changing this value.

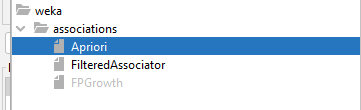
* Số vòng lặp: 16 vòng lặp.
* Ngưỡng hỗ trợ tối thiểu: 0.2 (cần ít nhất 3 lần xuất hiện trong tập dữ liệu).
* Ngưỡng tin cậy tối thiểu: 0.5 (luật phải đúng ít nhất 50% thời gian).
* Số lượng itemsets lớn:
  + L(1): 8 tập hợp mục
  + L(2): 18 tập hợp mục
  + L(3): 9 tập hợp mục
* Luật 1: milk=T butter=F ==> beer=F
  + Confidence: 1 (luôn đúng 100%)
  + Lift: 1.67 (tăng xác suất xảy ra của beer=F khi milk=T butter=F)
* Luật 6: butter=T ==> bread=T
  + Confidence: 0.86 (86% đúng khi butter=T)
  + Lift: 1.29

Exercise 2: Find association rules in the weather.nominal dataset by Weka explorer



a. Load the data in Weka Explorer. Select the Associate tab. How many different associations rule mining algorithms are available?

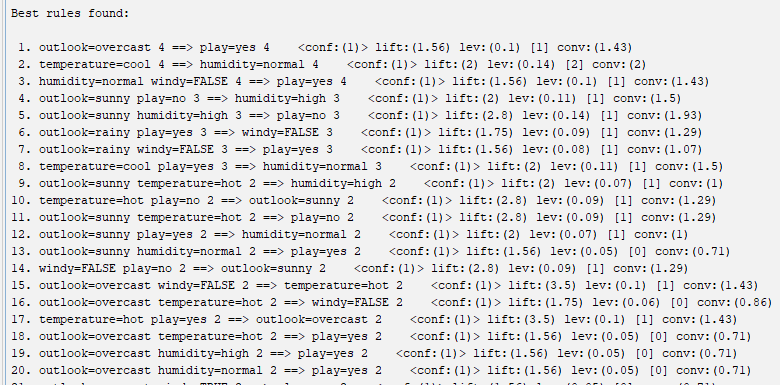
Chỉ xuất hiệ được 3 thuật toán khai thác quy tắc liên kết.



b. Choose Apriori algorithm with the following parameters (which you can select by clicking on the chosen algorithm: support threshold = 15% (lowerBoundMinSupport = 0.15), confidence threshold = 90% (metricType = confidence, minMetric = 0.9), number of rules = 50 (numRules = 50). After starting the algorithm, how many rules do you find? Could you use the regular weather dataset to get the results? Explain why.

* Với các tham số đã chọn, thuật toán Apriori tìm được 50 luật.
* Có thể, nhưng tập dữ liệu thời tiết quá nhỏ (chỉ 14 mẫu), nên các luật tìm được có thể không phản ánh chính xác thực tế và dễ bị quá khớp. Tập dữ liệu lớn hơn sẽ cho kết quả đáng tin cậy hơn.

c. Paste a screenshot of the Explorer window showing at least the first 20 rules



d. Based on the output, what is the support for this item set?

outlook = rainy humidity = normal windy = FALSE play = yes

* Dựa trên tập các mục lớn L(4), support của tập mục outlook = rainy, humidity = normal, windy = FALSE, play = yes 2. Điều này có nghĩa là tập hợp này xuất hiện trong 2 trường hợp, tương ứng với 14.3% (2/14) số lần xuất hiện trong toàn bộ dữ liệu.

e. Apriori algorithm generates association rules from frequent itemsets. How many itemsets of size 4 were found? Which rule(s) have been generated from itemset of size 4 (temperature=mild, windy=false, play=yes, outlook=rainy)? List their numbers in the list of rules.

Trong Large Itemsets L(4), có tổng cộng 6 itemsets với kích thước 4 đã được tìm thấy. Các itemsets này bao gồm 4 thuộc tính, ví dụ như: outlook=sunny temperature=hot humidity=high play=no.

f. What is the total number of possible rules for the weather data for each combination of values in the following table?

| Minimun confidence | Minimum support | Number of rules |
| --- | --- | --- |
| 0.9 | 0.3 | 3 |
| 0.9 | 0.2 | 8 |
| 0.9 | 0.1 | 336 |
| 0.8 | 0.3 | 5 |
| 0.8 | 0.2 | 10 |
| 0.8 | 0.1 | 338 |
| 0.7 | 0.3 | 6 |
| 0.7 | 0.2 | 17 |
| 0.7 | 0.1 | 345 |

Exercise 3: The 'database' below has four transactions:

Trans\_id Itemlist

T1 K, A, D, B

T2 D, A C, E, B

T3 C, A, B, E

T4 B, A, D

a. Use manually the Apriori algorithm to find all frequent itemsets and all association rules with minimum support is 60% and the minimum confidence is 80%.   
Frequent Itemsets:

1. Size of set of large itemsets L(1): 4 itemsets có kích thước 1.
2. Size of set of large itemsets L(2): 5 itemsets có kích thước 2.
3. Size of set of large itemsets L(3): 2 itemsets có kích thước 3.

### Association Rules:

Đây là danh sách các luật được tìm thấy với confidence 100% (conf:1):

1. B=TRUE → A=TRUE (4 instances)
   * Confidence: 1 (100%)
   * Lift: 1
2. A=TRUE → B=TRUE (4 instances)
   * Confidence: 1 (100%)
   * Lift: 1
3. K=FALSE → A=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
4. K=FALSE → B=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
5. D=TRUE → A=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
6. D=TRUE → B=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
7. K=FALSE, B=TRUE → A=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
8. K=FALSE, A=TRUE → B=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
9. K=FALSE → A=TRUE, B=TRUE (3 instances)
   * Confidence: 1 (100%)
   * Lift: 1
10. D=TRUE, B=TRUE → A=TRUE (3 instances)
    * Confidence: 1 (100%)
    * Lift: 1

b. What association rules can be found in this set using WEKA, if the minimum support is 60% and the minimum confidence is 80%.

* Tất cả các luật được tạo ra đều có confidence là 100% và lift là 1. Điều này có nghĩa là các tập mục này luôn xảy ra cùng nhau trong các giao dịch đã cho.
* Do các ràng buộc về support (60%) và confidence (80%) khá cao, các luật tìm thấy đều xảy ra với xác suất cao trong dữ liệu.

c. Compare the association rules output from Apriori in weka and manually.

1. Quy tắc liên kết từ WEKA:

Bạn đặt mức hỗ trợ tối thiểu là 60% và mức độ tin cậy tối thiểu là 80%.

Các quy tắc được tìm thấy khi sử dụng WEKA với 4 giao dịch (trường hợp) và 6 thuộc tính (mục) như sau:

Quy tắc:

B=TRUE → A=TRUE (Độ tin cậy: 100%)

A=TRUE → B=TRUE (Độ tin cậy: 100%)

K=FALSE → A=TRUE (Độ tin cậy: 100%)

K=FALSE → B=TRUE (Độ tin cậy: 100%)

D=TRUE → A=TRUE (Độ tin cậy: 100%)

D=TRUE → B=TRUE (Độ tin cậy: 100%)

K=FALSE, B=TRUE → A=TRUE (Độ tin cậy: 100%)

K=FALSE, A=TRUE → B=TRUE (Độ tin cậy: 100%)

K=FALSE → A=ĐÚNG, B=ĐÚNG (Độ tin cậy: 100%)

D=ĐÚNG, B=ĐÚNG → A=ĐÚNG (Độ tin cậy: 100%)

2. Tính toán thủ công các quy tắc liên kết (Thuật toán Apriori):

các bước sau:

Bước 1: Tính toán hỗ trợ cho các tập mục

Đầu tiên,i xác định tất cả các tập mục thường xuyên đáp ứng hỗ trợ tối thiểu là 60% (tức là xuất hiện trong ít nhất 3 trong số 4 giao dịch).

T1: {K, A, D, B}

T2: {D, A, C, E, B}

T3: {C, A, B, E}

T4: {B, A, D}

Tần suất tập hợp mục:

K: 1 (25%)

A: 4 (100%)

B: 4 (100%)

D: 3 (75%)

C: 2 (50%)

E: 2 (50%)

Theo tần suất, chỉ các tập hợp mục xuất hiện trong 3 giao dịch trở lên (≥60%) mới được coi là thường xuyên:

Tập ​​hợp 1 mục thường xuyên: {A, B, D}

Đối với tập hợp 2 mục:

(A, B): 4 (100%)

(A, D): 3 (75%)

(B, D): 3 (75%)

Đối với tập 3 mục:

(A, B, D): 3 (75%)

Bước 2: Tạo quy tắc liên kết

tạo quy tắc từ các tập mục thường xuyên:

Quy tắc từ (A, B):

A → B (Hỗ trợ = 100%, Độ tin cậy = 100%)

B → A (Hỗ trợ = 100%, Độ tin cậy = 100%)

Quy tắc từ (A, D):

A → D (Hỗ trợ = 75%, Độ tin cậy = 75%)

D → A (Hỗ trợ = 75%, Độ tin cậy = 100%)

Quy tắc từ (B, D):

B → D (Hỗ trợ = 75%, Độ tin cậy = 75%)

D → B (Hỗ trợ = 75%, Độ tin cậy = 100%)

Quy tắc từ (A, B, D):

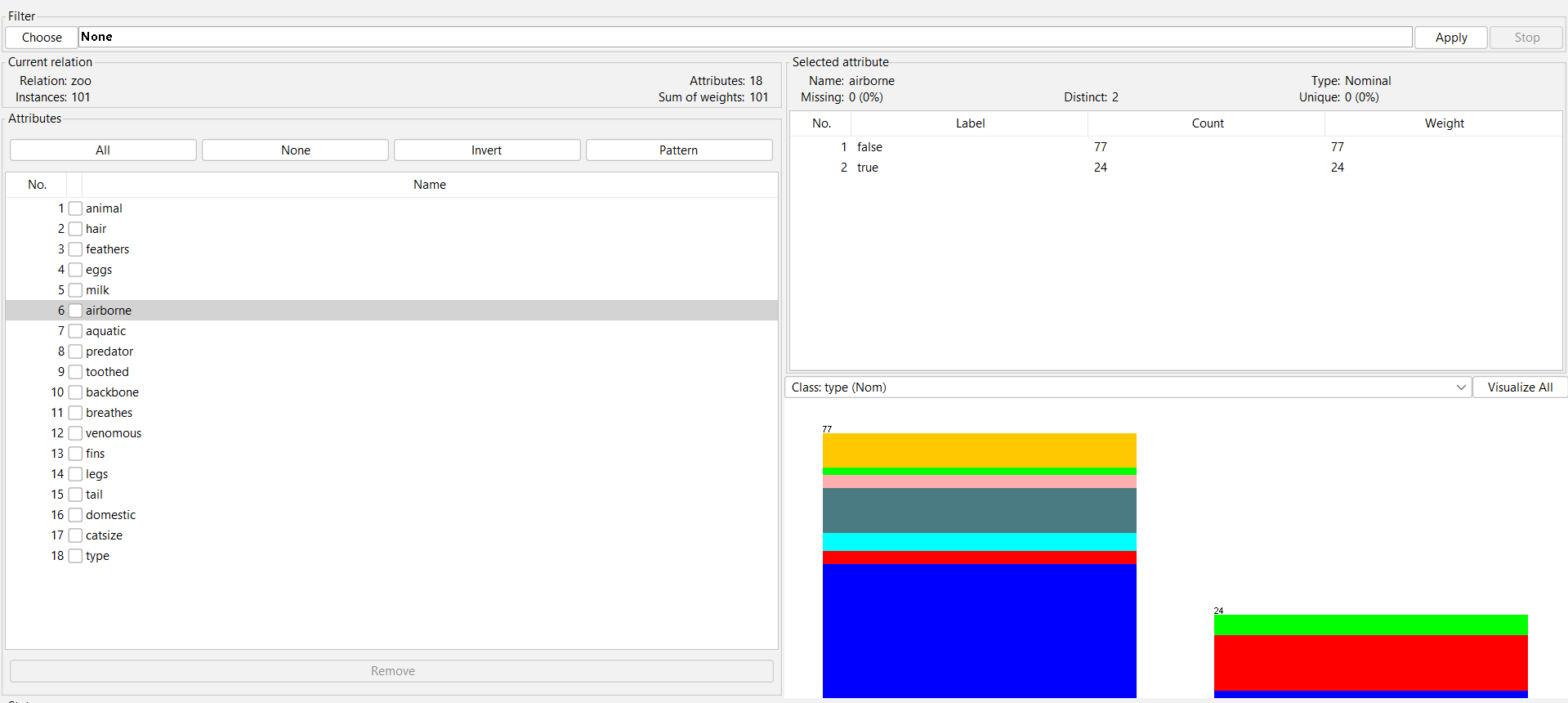
A, B → D (Hỗ trợ = 75%, Độ tin cậy = 75%)

A, D → B (Hỗ trợ = 75%, Độ tin cậy = 100%)

B, D → A (Hỗ trợ = 75%, Độ tin cậy = 100%)

Exercise 4:

* Download the zoo dataset and load it into Weka.



* Examine the attributes. Are all attributes nominal?

Thuộc tính nominal:

1. nominalanimalnominal: `{aardvark, antelope, bass, bear, ...}` (tên động vật là nominal)

2. nominalhairnominal: `{false, true}` (nhị phân nominal)

3. nominalfeathersnominal: `{false, true}` (nhị phân nominal)

4. nominaleggsnominal: `{false, true}` (nhị phân nominal)

5. nominalmilknominal: `{false, true}` (nhị phân nominal)

6. nominalairbornenominal: `{false, true}` (nhị phân nominal)

7. nominalaquaticnominal: `{false, true}` (nhị phân nominal)

8. nominalpredatornominal: `{false, true}` (nhị phân nominal)

9. nominaltoothednominal: `{false, true}` (nhị phân nominal)

10. nominalbackbonenominal: `{false, true}` (nhị phân nominal)

11. nominalbreathesnominal: `{false, true}` (nhị phân nominal)

12. nominalvenomousnominal: `{false, true}` (nhị phân nominal)

13. nominalfinsnominal: `{false, true}` (nhị phân nominal)

14. nominaltailnominal: `{false, true}` (nhị phân nominal)

15. nominaldomesticnominal: `{false, true}` (nhị phân nominal)

16. nominalcatsizenominal: `{false, true}` (nhị phân nominal)

17. nominaltypenominal: `{mammal, bird, reptile, fish, amphibian, insect, invertebrate}` (nominal)

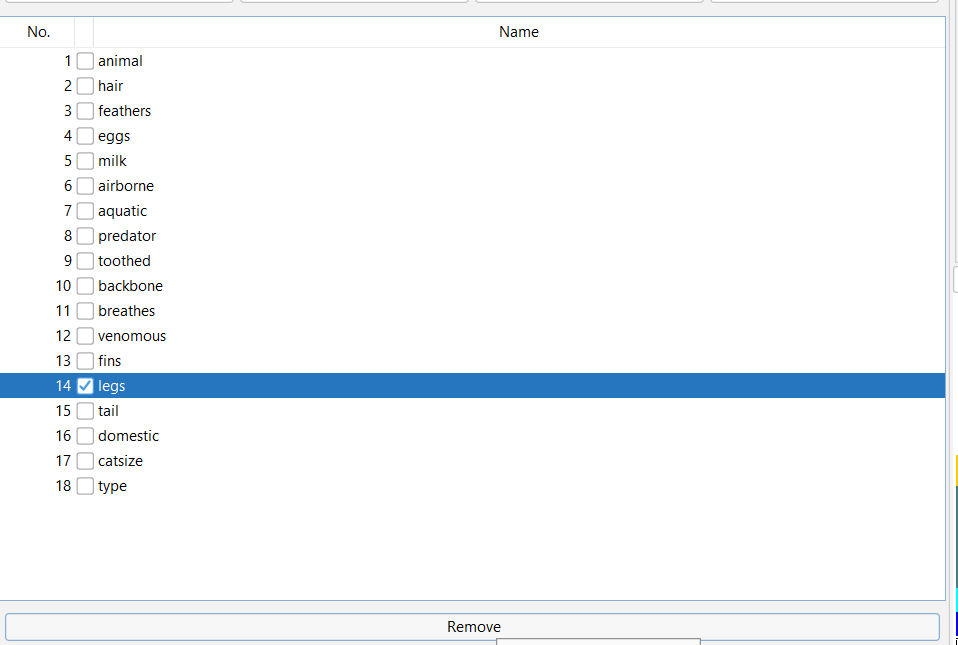
Thuộc tính không phải nominal (dữ liệu số):

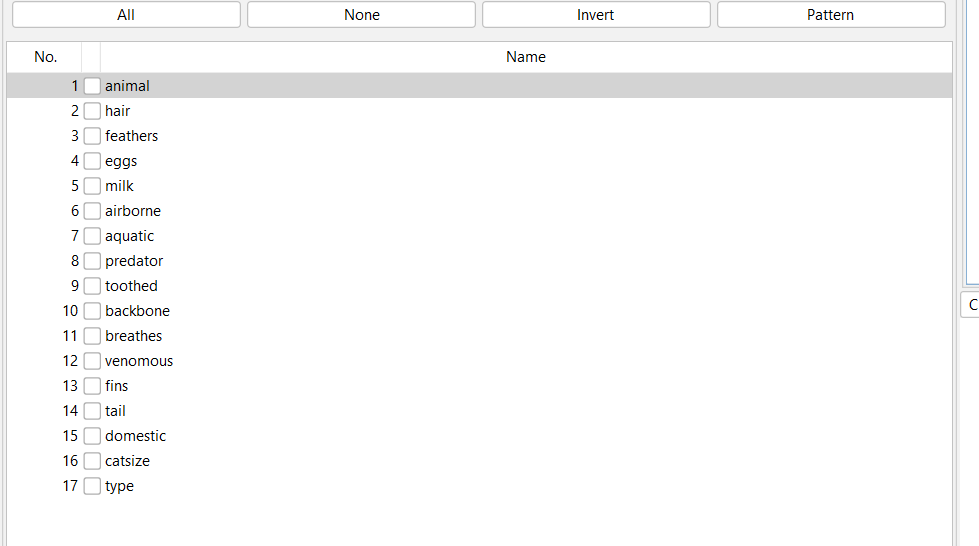
legsnominal: `INTEGER [0,9]` (số nguyên, số học, vì nó biểu thị số lượng chân)

.

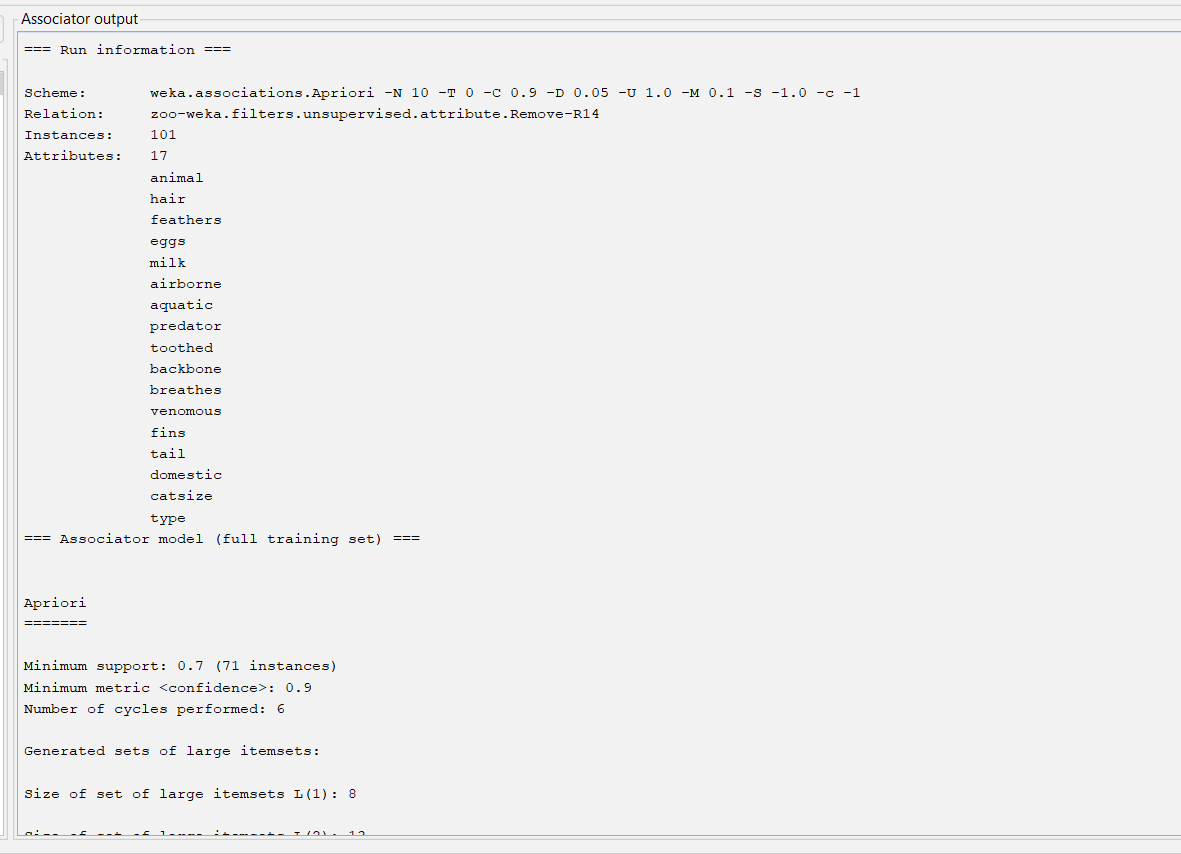
* Preprocess data to make sure the dataset can be used directly with Apriori algorithm.

Sử dụng chức năng remove xóa dữ liệu của cột leg để có thể tiền xử lý dữ liệu để đảm bảo tập dữ liệu có thể được sử dụng trực tiếp với thuật toán Apriori.

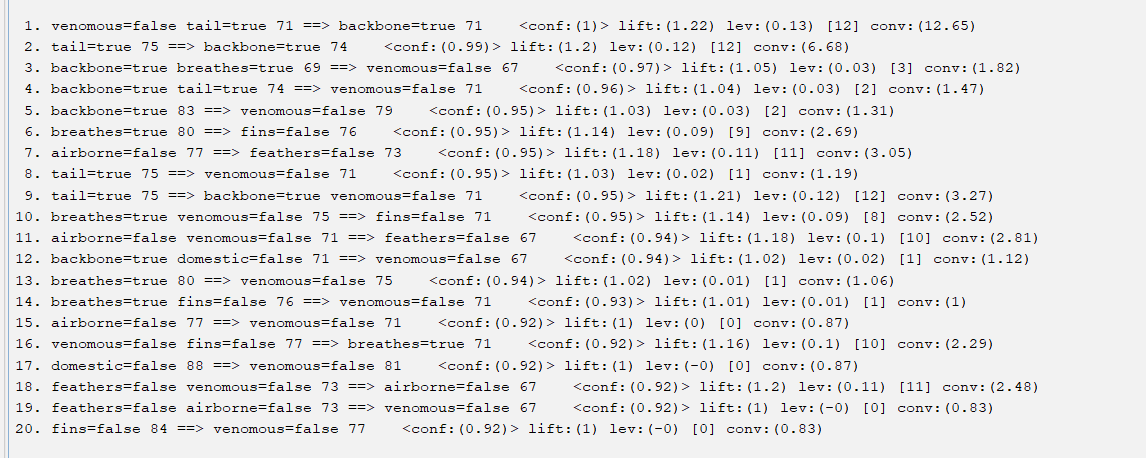


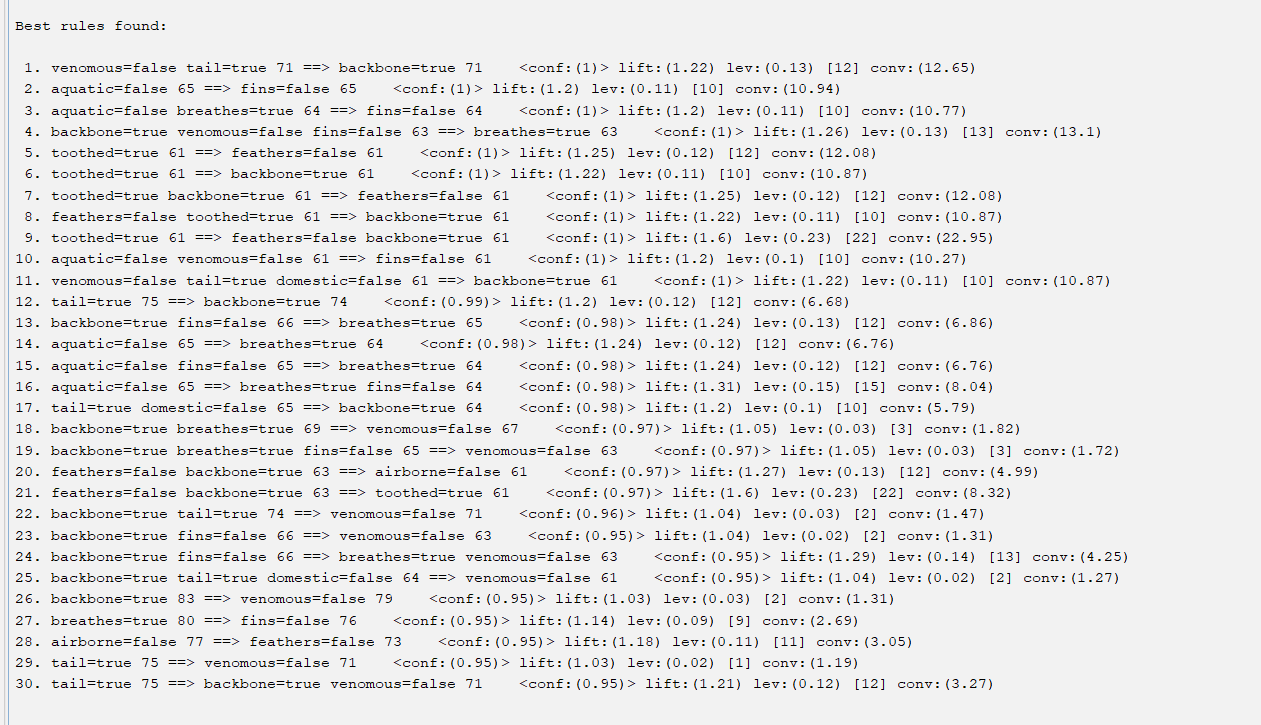


* Use the Apriori algorithm with the default parameters. Record the generated rules.

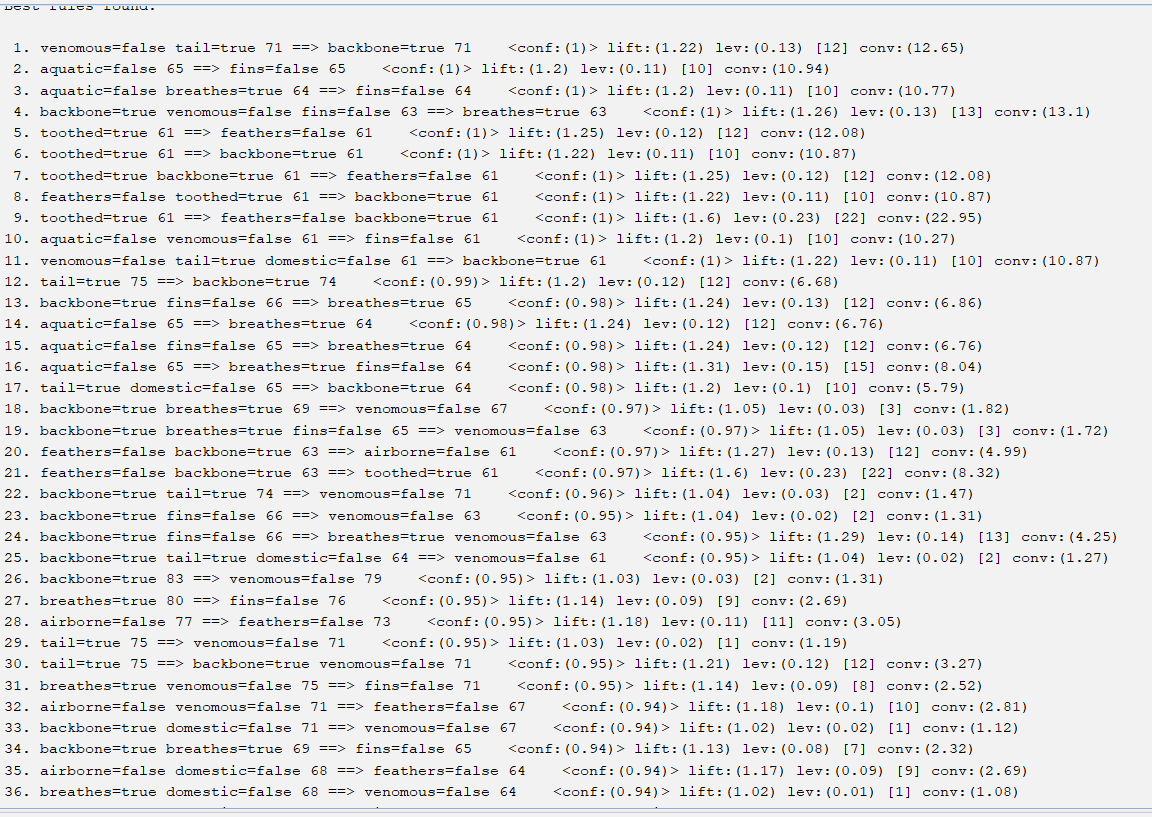


* Vary the number of rules generated (20, 30, ...). Record how many rules have to be generated before generating a rule containing type=mammal.

20 rule

30 rule 

40 rule



Sau 20, 30, 40, 50, .. rule không tìm thầy thuộc tính type = mammal

* Vary the maximum support until a rule containing type=mammal is the top rule generated. Record the maximum support needed.
* Select one generated rule that was interesting to you. Why was it interesting? What does it mean?

Hãy xem Quy tắc 4:

Quy tắc 4: backbone=true venomous=false fins=false 63 ==> breathes=true 63

Tại sao nó thú vị:

Bộ thuộc tính toàn diện: Quy tắc này kết hợp nhiều thuộc tính—backbone=true, venomous=false, fins=false—để dự đoán một thuộc tính khác, breathes=true. Nó nắm bắt được hồ sơ tương đối chi tiết của loài động vật đang được mô tả.

Độ tin cậy cao: Quy tắc có độ tin cậy cao (conf:(1)) và giá trị nâng đáng kể, cho thấy mối quan hệ chặt chẽ và cụ thể giữa các thuộc tính.

* Check its confidence and support - are they high enough?

### Đánh giá Mức độ Tin cậy và Hỗ trợ

* Mức độ tin cậy (Confidence): 1 (hoàn toàn tin cậy)
* Hỗ trợ (Support): 63 (số lượng trường hợp khá lớn)