

## BÁO CÁO TÌM HIỂU PHẦN MỀM WEKA

Nguyễn Thành Tân: 1512491

Đỗ Thành Nhơn: 1512387

### I. Tìm hiểu Weka

#### 1. Thông tin chung

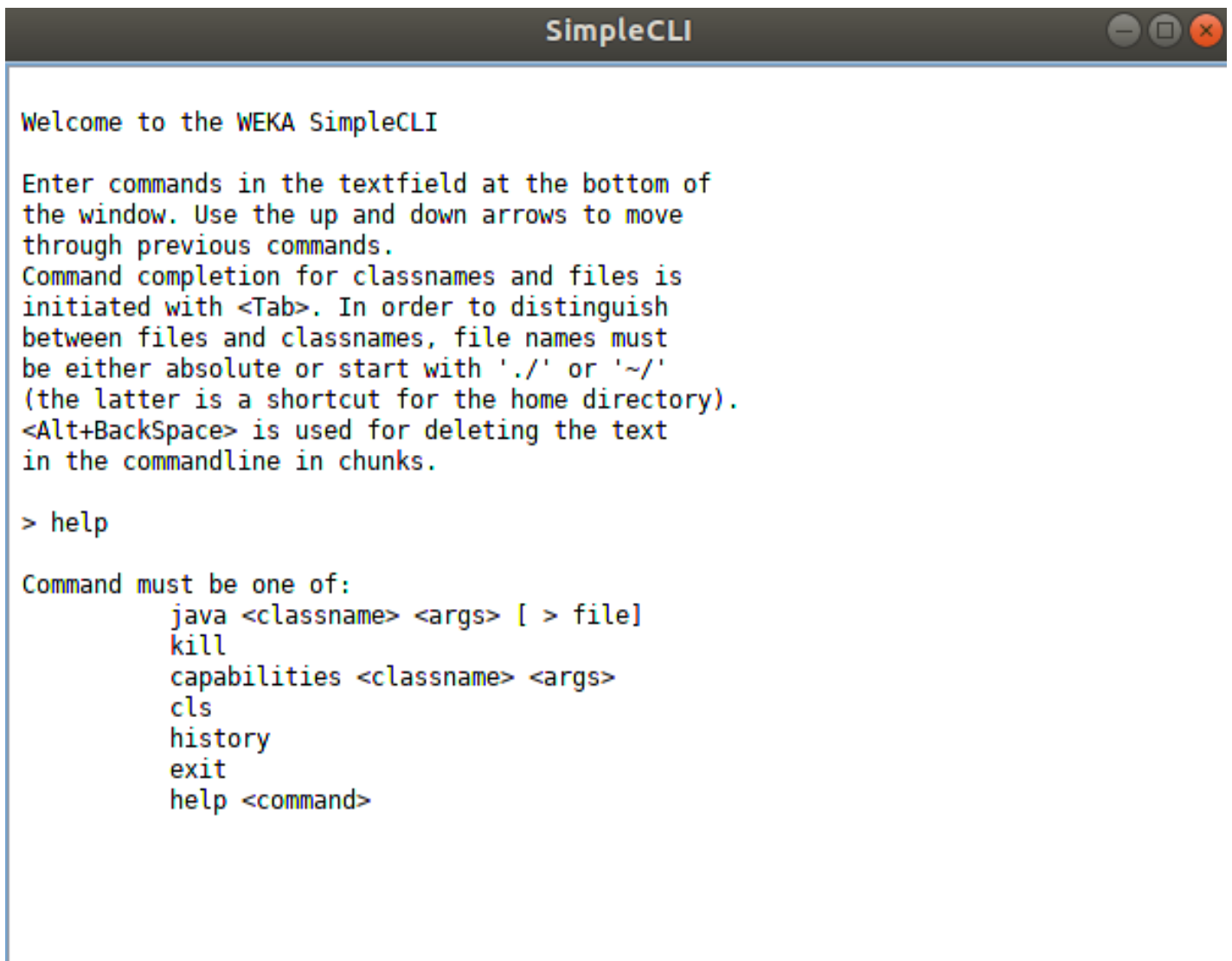
Weka là một phần mềm mã nguồn mở được viết trên ngôn ngữ java gồm các chức năng:

- Phân loại
  - Gom nhóm
  - Luật quan hệ
- #### 2. Interface
- Màn hình GUI chính:



## TÌM HIỂU WEKA

- Các command line cơ bản:



The screenshot shows a window titled "SimpleCLI" with standard Windows window controls (minimize, maximize, close). The window contains the following text:

```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or '~/ '
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
    java <classname> <args> [ > file]
    kill
    capabilities <classname> <args>
    cls
    history
    exit
    help <command>
```

## TÌM HIỂU WEKA

- Explorer (nơi thực hiện đa phần công việc)
  - Tiền xử lý dữ liệu

**Preprocess** | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

**Filter**  
Choose **None** | Apply

**Current relation**  
Relation: zoo | Instances: 101 | Attributes: 18 | Sum of weights: 101

**Attributes**  
All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> animal
2	<input type="checkbox"/> hair
3	<input type="checkbox"/> feathers
4	<input type="checkbox"/> eggs
5	<input type="checkbox"/> milk
6	<input type="checkbox"/> airborne
7	<input type="checkbox"/> aquatic
8	<input type="checkbox"/> predator

Remove

**Selected attribute**  
Name: animal | Missing: 0 (0%) | Distinct: 100 | Type: Nominal | Unique: 99 (98%)

No.	Label	Count	Weight
1	aardvark	1	1.0
2	antelope	1	1.0
3	bass	1	1.0
4	bear	1	1.0
5	boar	1	1.0

Class: type (Nom) | Visualize All

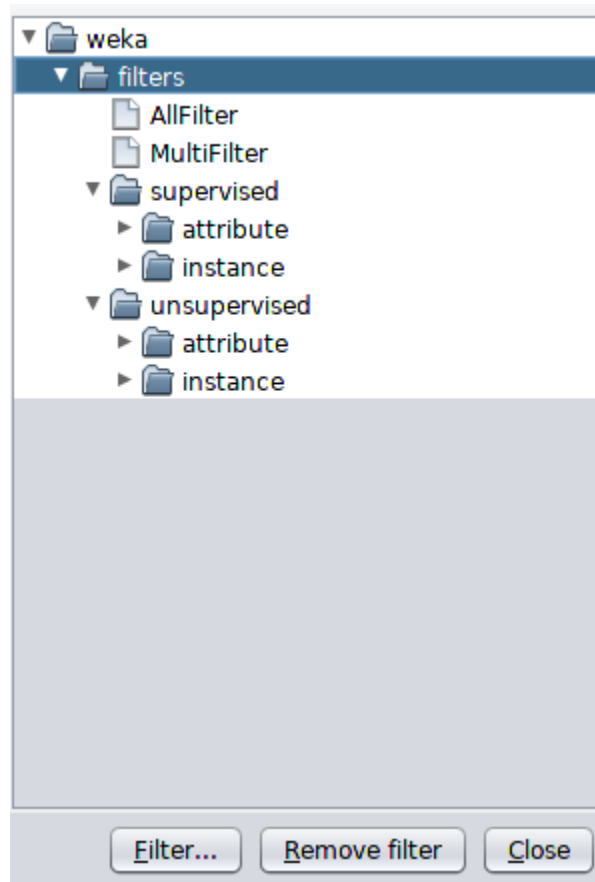
**Status**  
OK | Log | x 0

Load data bằng file, url,...

Sau khi edit data có thể save lại data mới.

## TÌM HIỂU WEKA

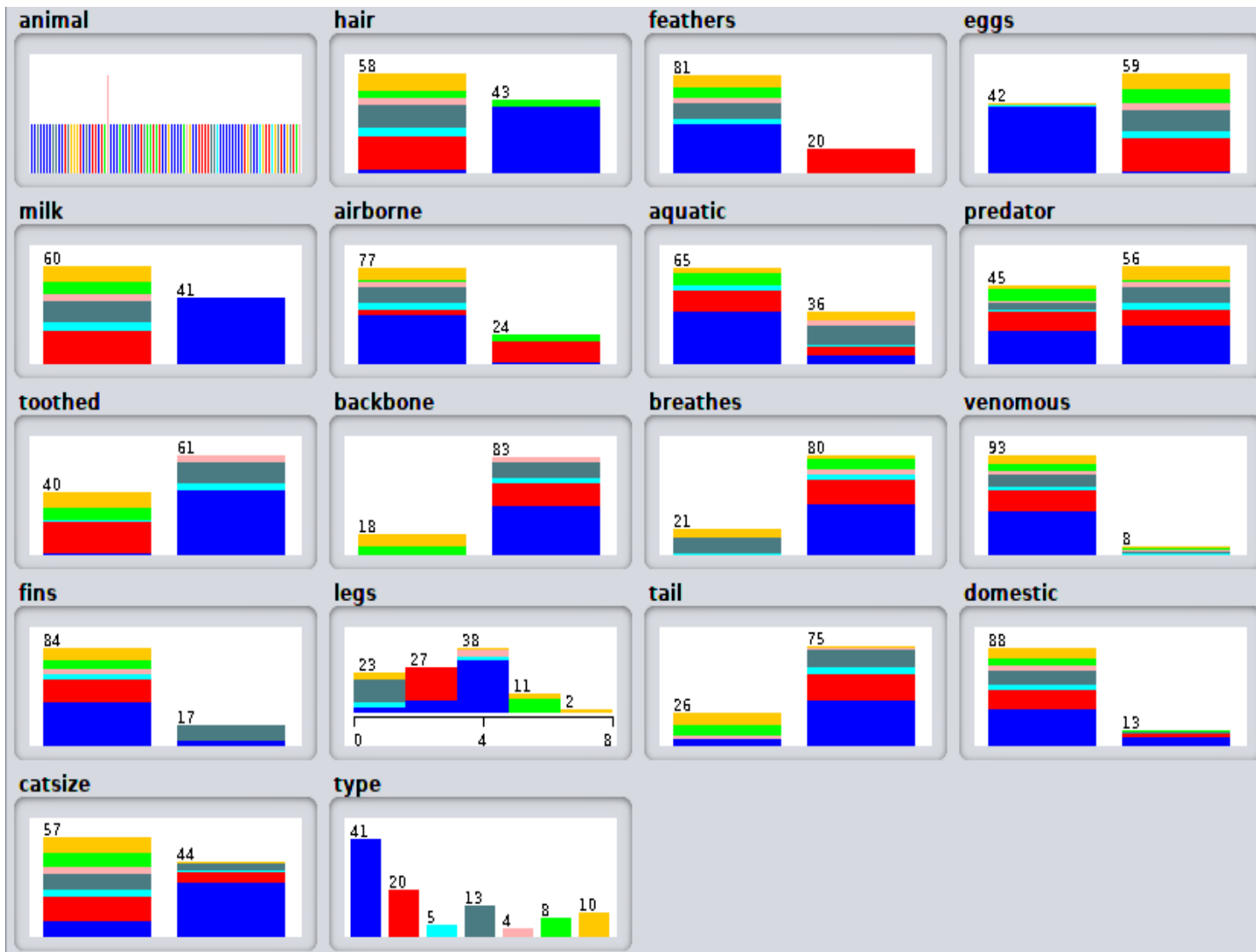
### ➤ Filter



Ta có thể áp dụng filter lên dataset để loại bớt dữ liệu không cần thiết hoặc chọn dữ liệu cần thiết.

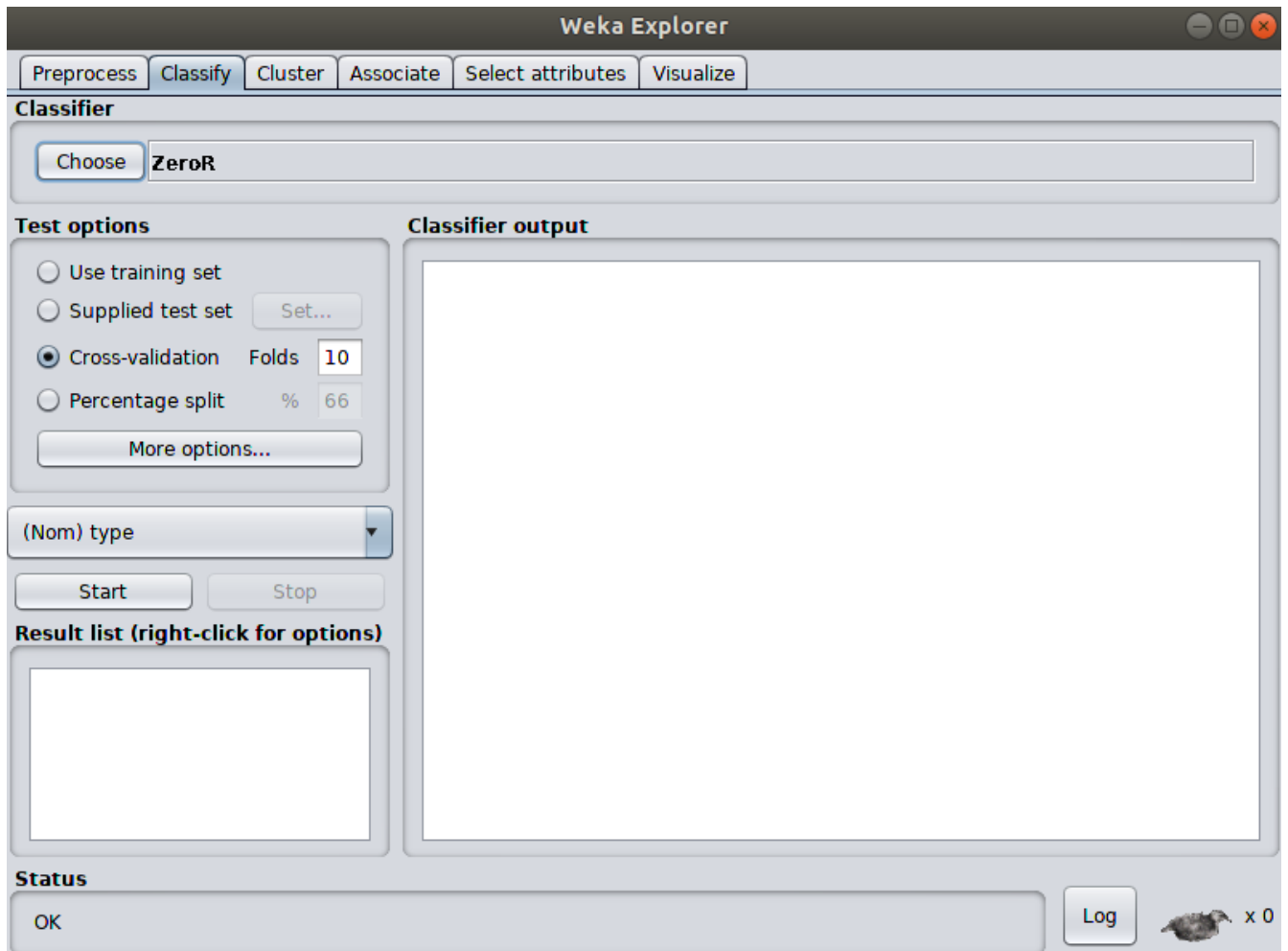
# TÌM HIỂU WEKA

- Visualization  
Visualize all để xem trực quan về toàn bộ thuộc tính của dữ liệu.

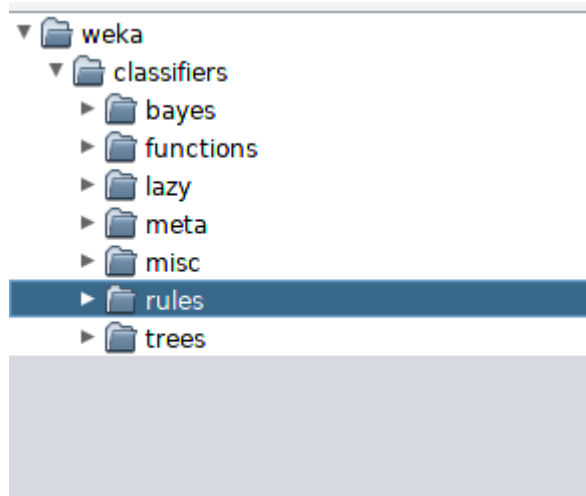


# TÌM HIỂU WEKA

## ➤ Phân lớp

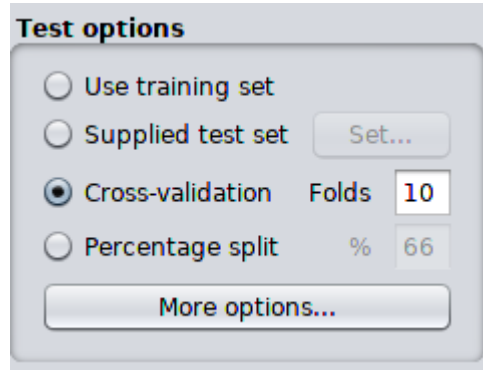


Choose: chọn thuật toán



## TÌM HIỂU WEKA

Test option:



Trong đó “Use training set” là dùng dữ liệu train để test, “Supplied test set” là người dùng cung cấp dữ liệu test, “Cross-validation” là chia dữ liệu train thành nhiều folds và giữ lại 1 fold để test trong lúc train, “Percentage split” chia tỉ lệ dữ liệu train là bao nhiêu %.

Output khi chạy thuật toán J48 (cải tiến ID3) trên tập dữ liệu Zoo.arff:

Các thông tin cơ bản bao gồm thuật toán, dataset, số lượng, thuộc tính, test option.

Kết quả cây quyết định (các thông số của cây như độ chính xác, F-measure,...).

```
=== Stratified cross-validation ===  
=== Summary ===
```

Correctly Classified Instances	93	92.0792 %
Incorrectly Classified Instances	8	7.9208 %
Kappa statistic	0.8955	
Mean absolute error	0.0225	
Root mean squared error	0.14	
Relative absolute error	10.2478 %	
Root relative squared error	42.4398 %	
Total Number of Instances	101	

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.600	0.010	0.750	0.600	0.667	0.656	0.793	0.420
	1.000	0.011	0.929	1.000	0.963	0.958	0.994	0.929
	0.750	0.000	1.000	0.750	0.857	0.862	0.872	0.760
	0.625	0.032	0.625	0.625	0.625	0.593	0.920	0.677
	0.800	0.033	0.727	0.800	0.762	0.735	0.986	0.812
Weighted Avg.	0.921	0.008	0.922	0.921	0.920	0.914	0.976	0.908

```
Size of the tree : 17
```

Confusion matrix cho biết các sample nào đúng, sai.

## TÌM HIỂU WEKA

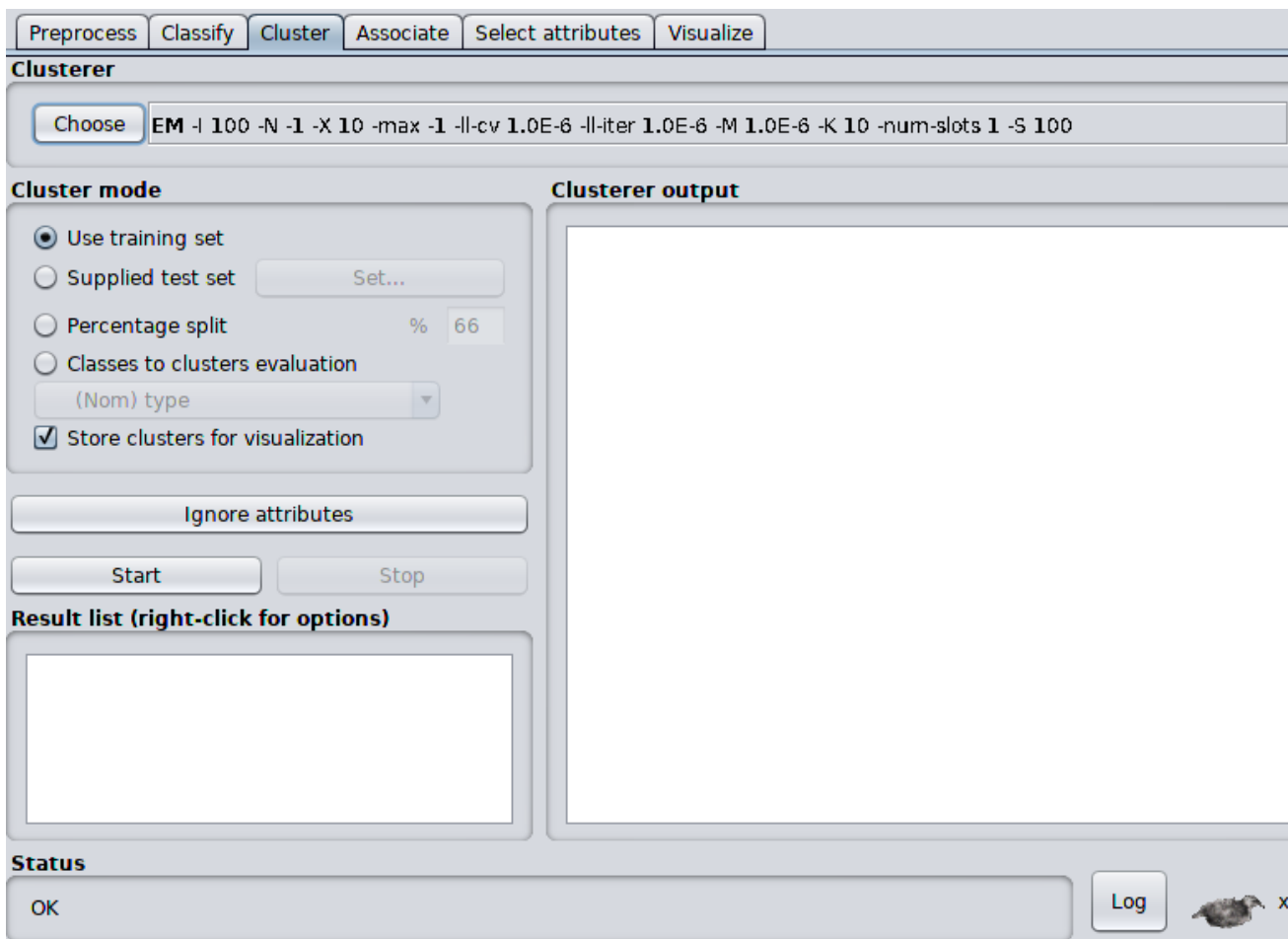
=== Confusion Matrix ===

	a	b	c	d	e	f	g		<-- classified as
41	0	0	0	0	0	0	0		a = mammal
0	20	0	0	0	0	0	0		b = bird
0	0	3	1	0	1	0	0		c = reptile
0	0	0	13	0	0	0	0		d = fish
0	0	1	0	3	0	0	0		e = amphibian
0	0	0	0	0	0	5	3		f = insect
0	0	0	0	0	0	2	8		g = invertebrate

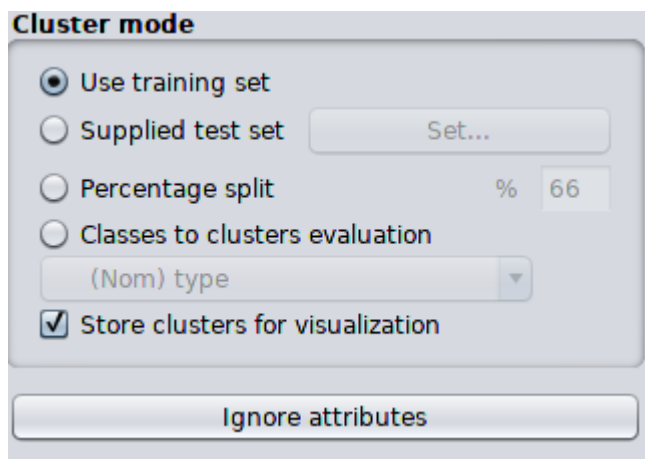


## TÌM HIỂU WEKA

### ➤ Gom nhóm

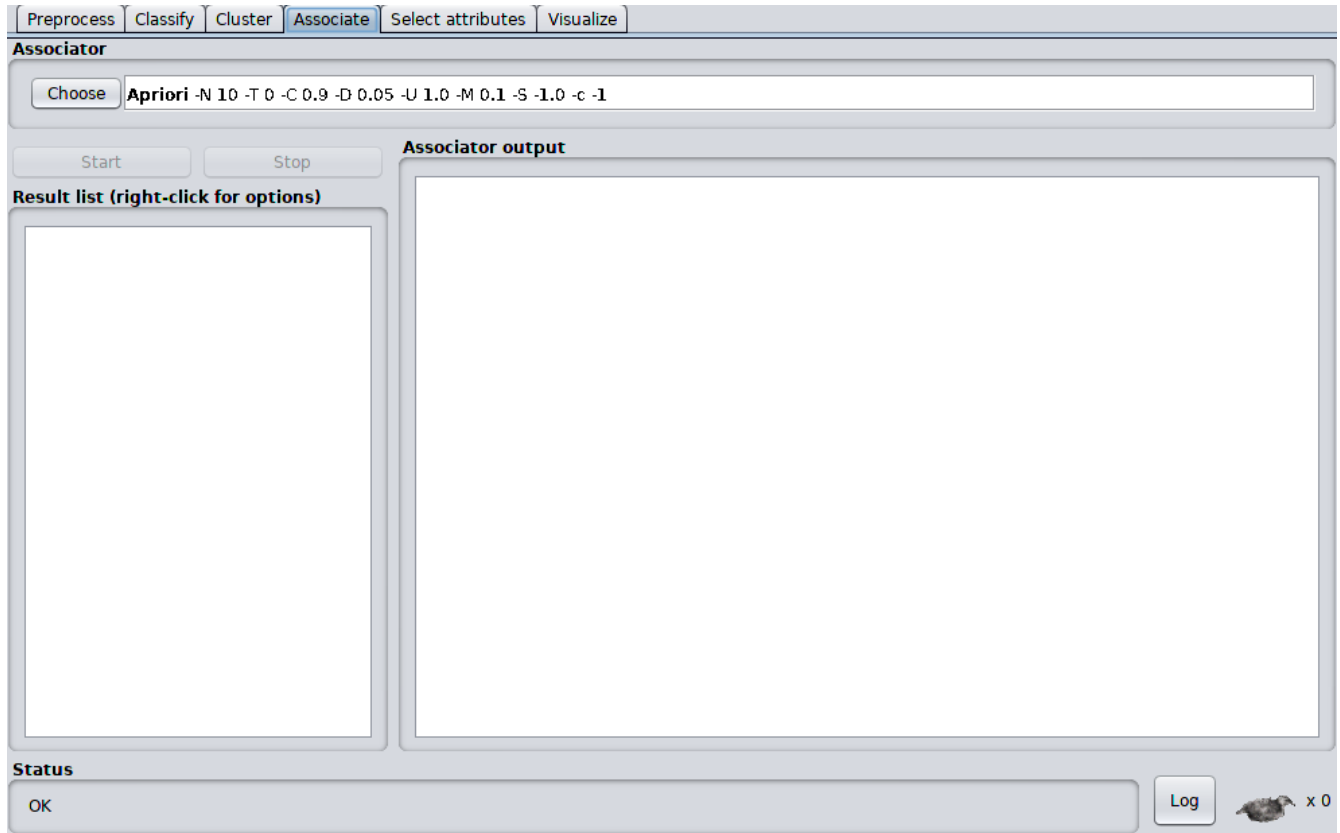


Cách sử dụng tương tự như phân lớp. Điểm khác biệt duy nhất là có thêm tính năng Ignore attributes để loại bớt những thuộc tính không dùng trong thuật toán gom nhóm.



## TÌM HIỂU WEKA

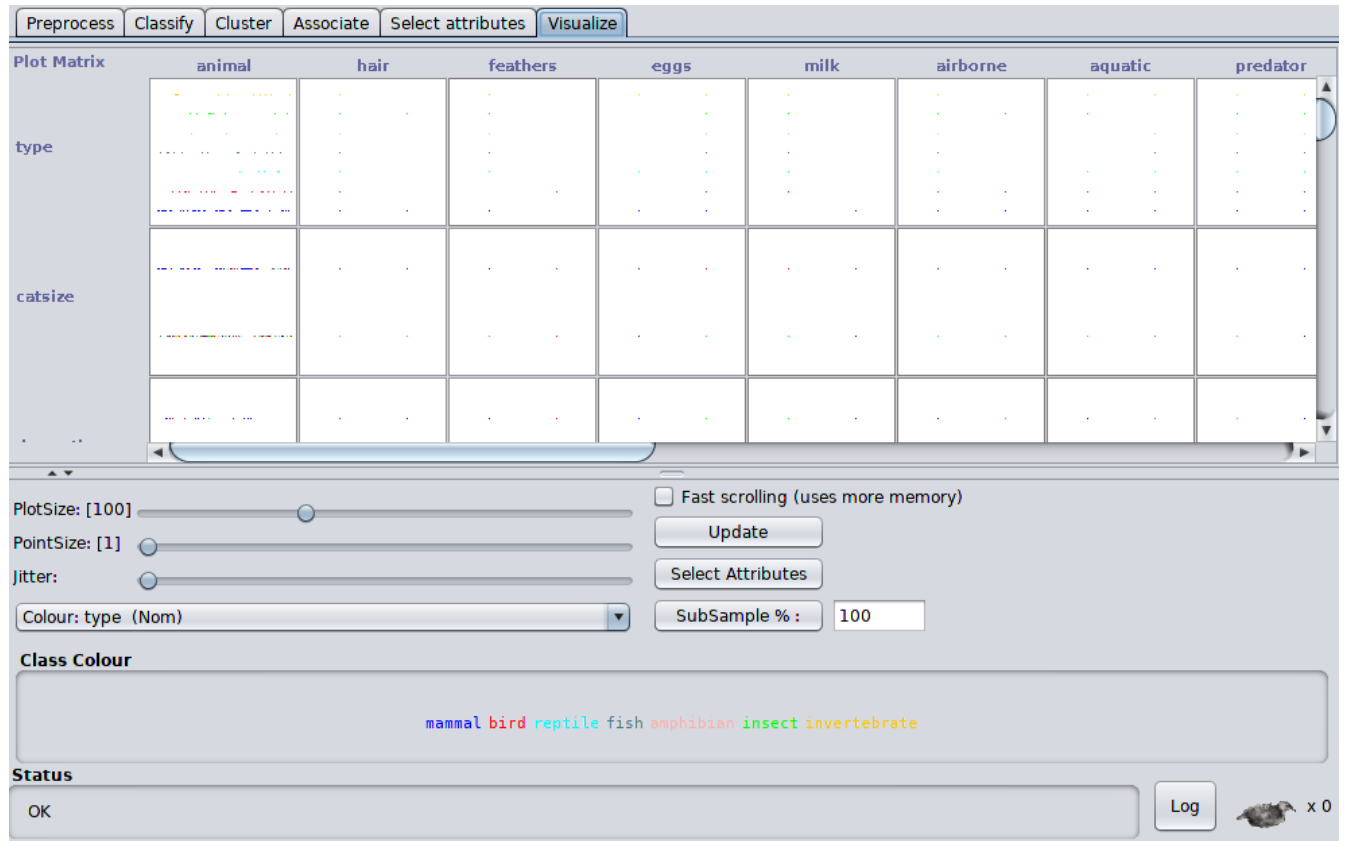
- Associate rule



Cách sử dụng tương tự như classify và cluster, chỉ cần load dữ liệu, chọn thuật toán và chạy.

## TÌM HIỂU WEKA

- Mục visualize chi tiết về dữ liệu (plot matrix)  
Ta có thể tùy chọn các options để explore dữ liệu sao cho thuận tiện nhất.



## II. THAO TÁC TRÊN DỮ LIỆU ZOO.arff

1. Dữ liệu

Số mẫu: 101

Thuộc tính: 18

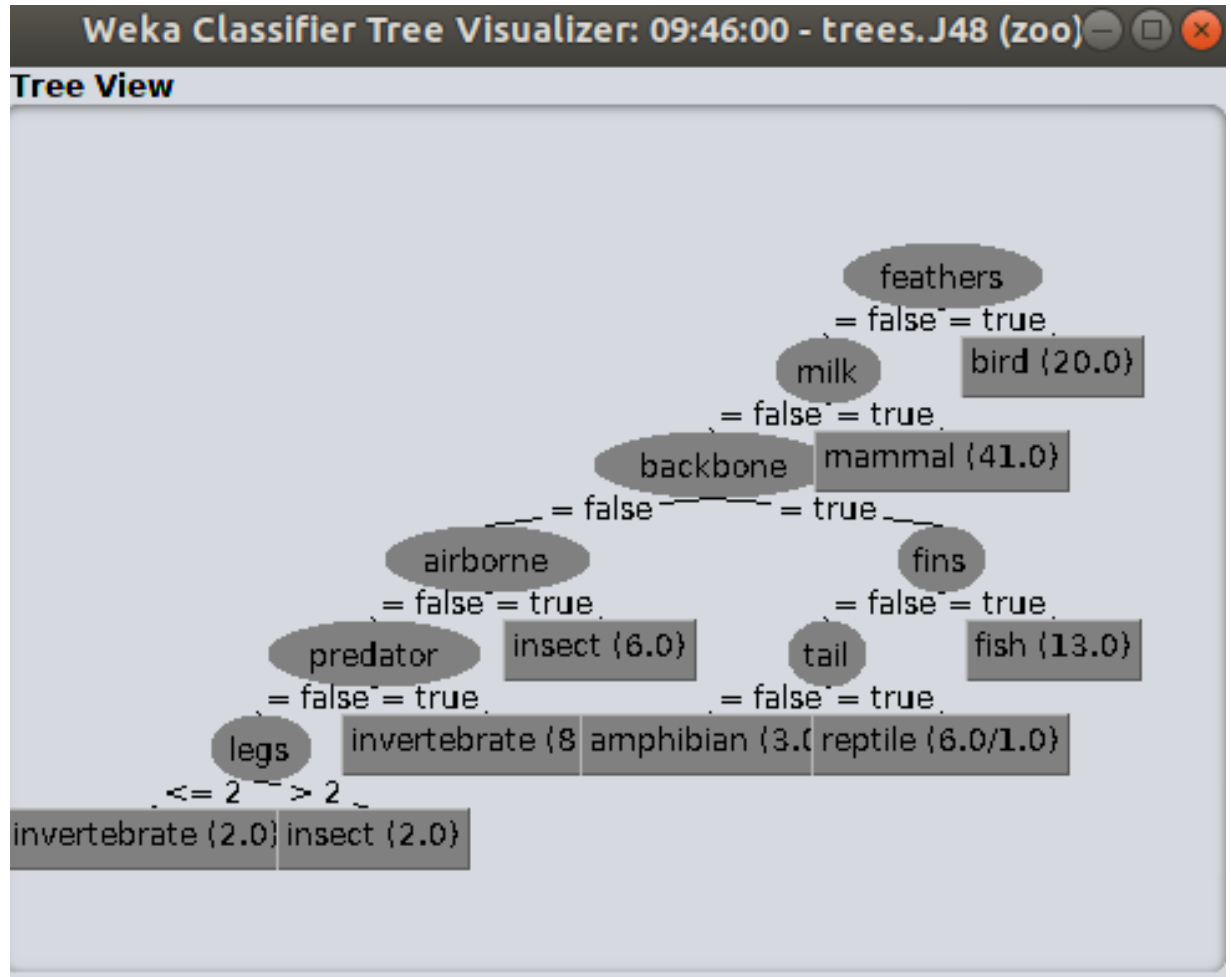
- |                 |                                         |
|-----------------|-----------------------------------------|
| 1. animal name: | Unique for each instance                |
| 2. hair         | Boolean                                 |
| 3. feathers     | Boolean                                 |
| 4. eggs         | Boolean                                 |
| 5. milk         | Boolean                                 |
| 6. airborne     | Boolean                                 |
| 7. aquatic      | Boolean                                 |
| 8. predator     | Boolean                                 |
| 9. toothed      | Boolean                                 |
| 10. backbone    | Boolean                                 |
| 11. breathes    | Boolean                                 |
| 12. venomous    | Boolean                                 |
| 13. fins        | Boolean                                 |
| 14. legs        | Numeric (set of values: {0,2,4,5,6,8})  |
| 15. tail        | Boolean                                 |
| 16. domestic    | Boolean                                 |
| 17. catsize     | Boolean                                 |
| 18. type        | Numeric (integer values in range [1,7]) |

Đặt lại tên cho phân lớp:

@ATTRIBUTE type { mammal, bird, reptile, fish, amphibian, insect, invertebrate }

## TÌM HIỂU WEKA

### 2. Cây đã sinh:



## TÌM HIỂU WEKA

### 3. Kết quả trên 5 sample:

=== Re-evaluation on test set ===

User supplied test set

Relation: zoo

Instances: unknown (yet). Reading incrementally

Attributes: 18

=== Predictions on user test set ===

inst#	actual	predicted	error	prediction
1	1:?	1:mammal		1
2	1:?	2:bird		1
3	1:?	3:reptile		0.833
4	1:?	4:fish		1
5	1:?	3:reptile		0.833

Predicted là model dự đoán với độ tự tin tương ứng là 1,1,0.833,1,0.833.