Assignment1 (Decision Tree) for INF 552

2016 Fall
Yi Ren 5527 2051 70

Part 1, Implementation of Decision Tree

1. language: Java 1.8

2. tool: eclipse 4.5.1

3. data structure: (1).ArrayList<ArrayList<String>>, (2).ArrayList<String>, (3).Tree, (4).String[], (5).Queue<TreeNode<String>>, (6).HashMap<key, value>

4. logic: the overall logic is build a tree with three element in the tree node: data(String, store the attribute), children(ArrayList<TreeNode>), and the value(ArrayList<String>, store branch) that corresponds to the children. To build the tree, I use five functions: builtDecisionTree, getUsefulData, bestAttributeToSplit, decisionTree, getAttributeValueList.

   In order to build the tree, I need two kinds of data, one is the records of each line in dt-data.txt, it is ArrayList<ArrayList<String>>, another is the attribute list of the dt-data.txt, it is ArrayList<String>, that is what getData(filePath) and getAttribute(filePath) do.

   To calculate Information Gain, I need to get Entropy, than calculate Information Gain, that is what calulateEntropy(data) and calulateInfoGain(data , indexofAttribute) do.

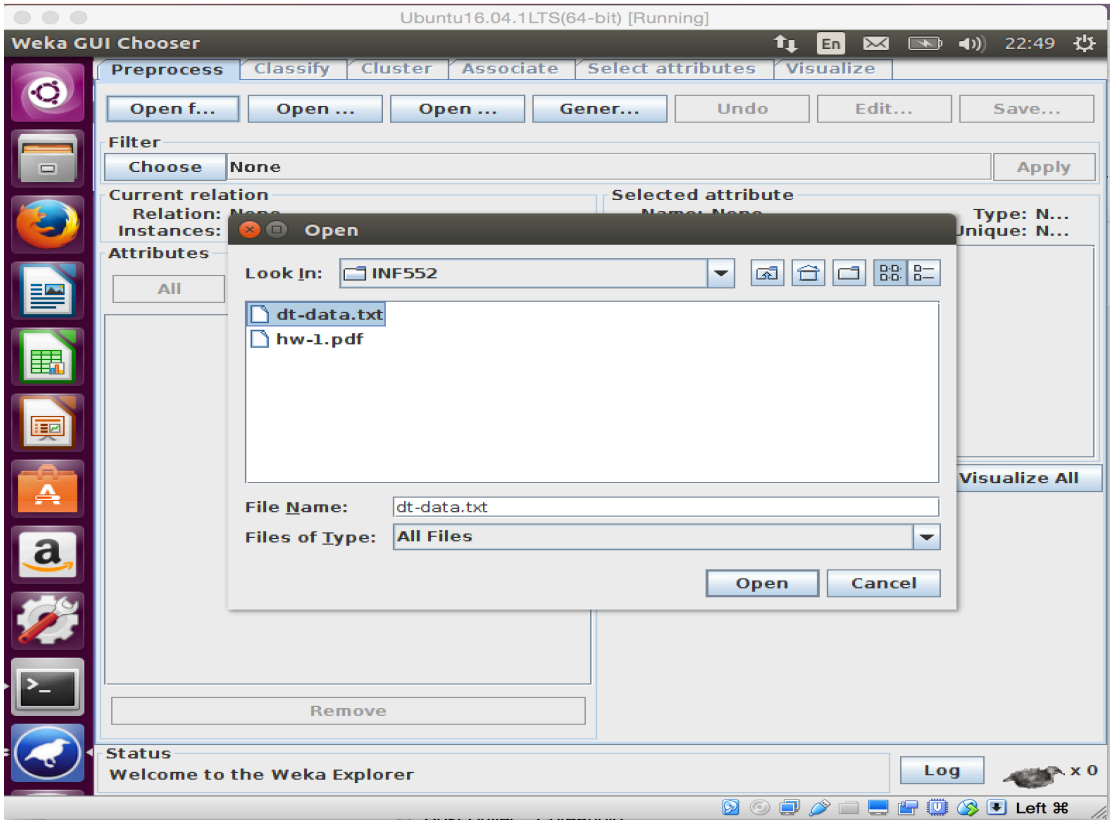
Part 2: Software Familiarization
I find that some people suggest weka to implement decision tree on the Internet, so I install weka on the Ubuntu system, and learn how to use it. I feel weka is very convenient to use , especially for the new users. And the interface of Weka is user friendly, and the visualization is great. Weka provide some decision tree algorithm, such as in this picture, the algorithm is J48. Also it is smart that my program because it can produce a simple model and have some noise tolerance.

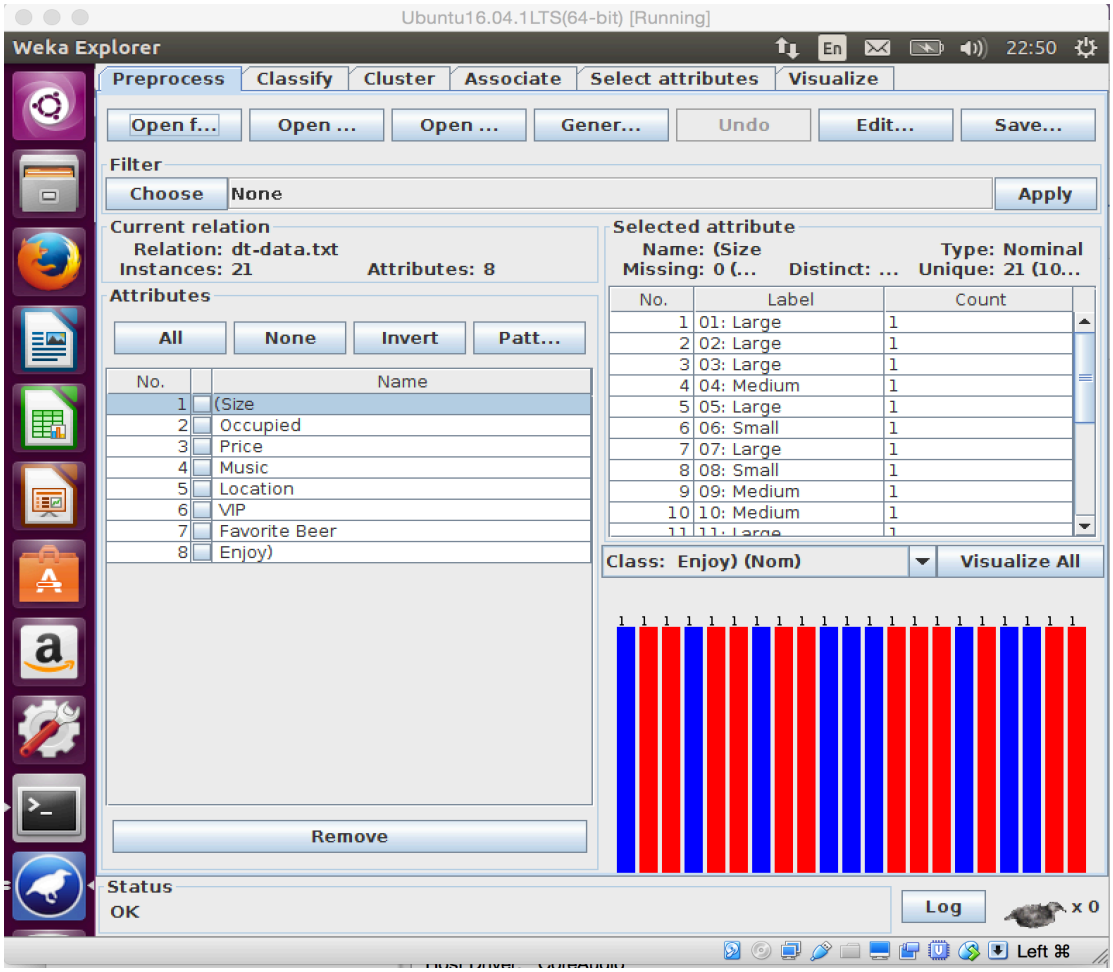
Install Weka on Ubuntu:

```
sudo apt-get install weka
```

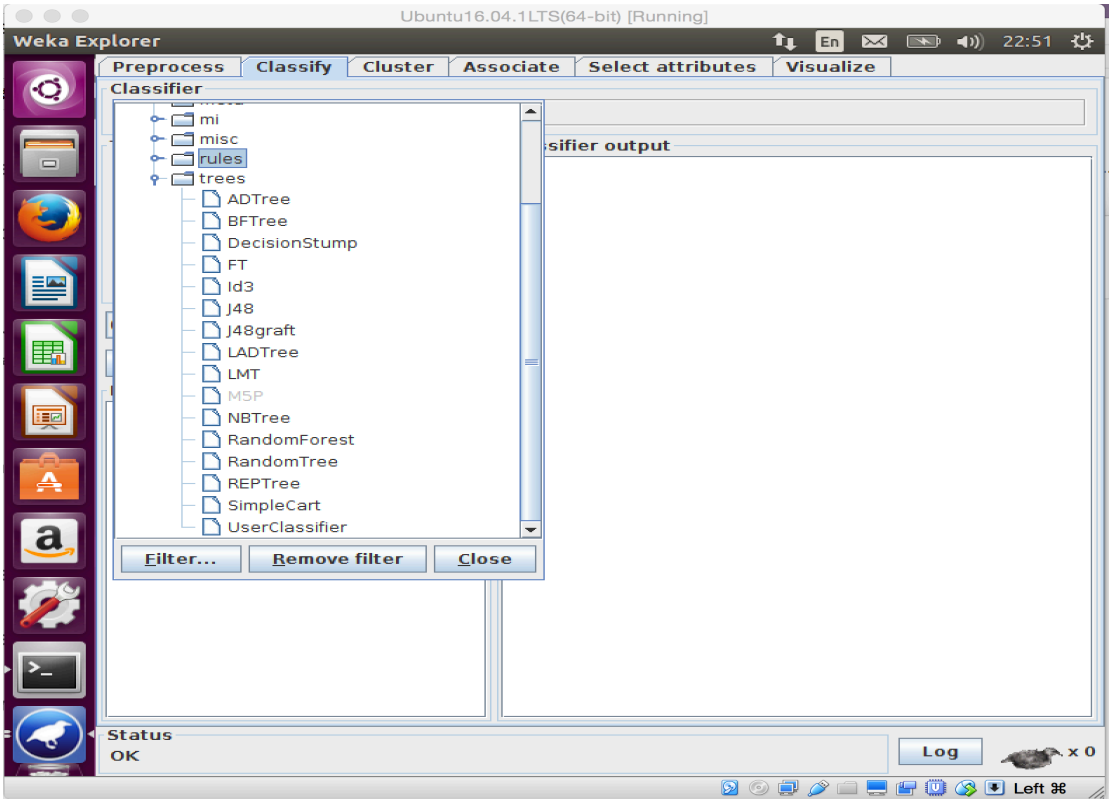
How to use:

(1). upload data:

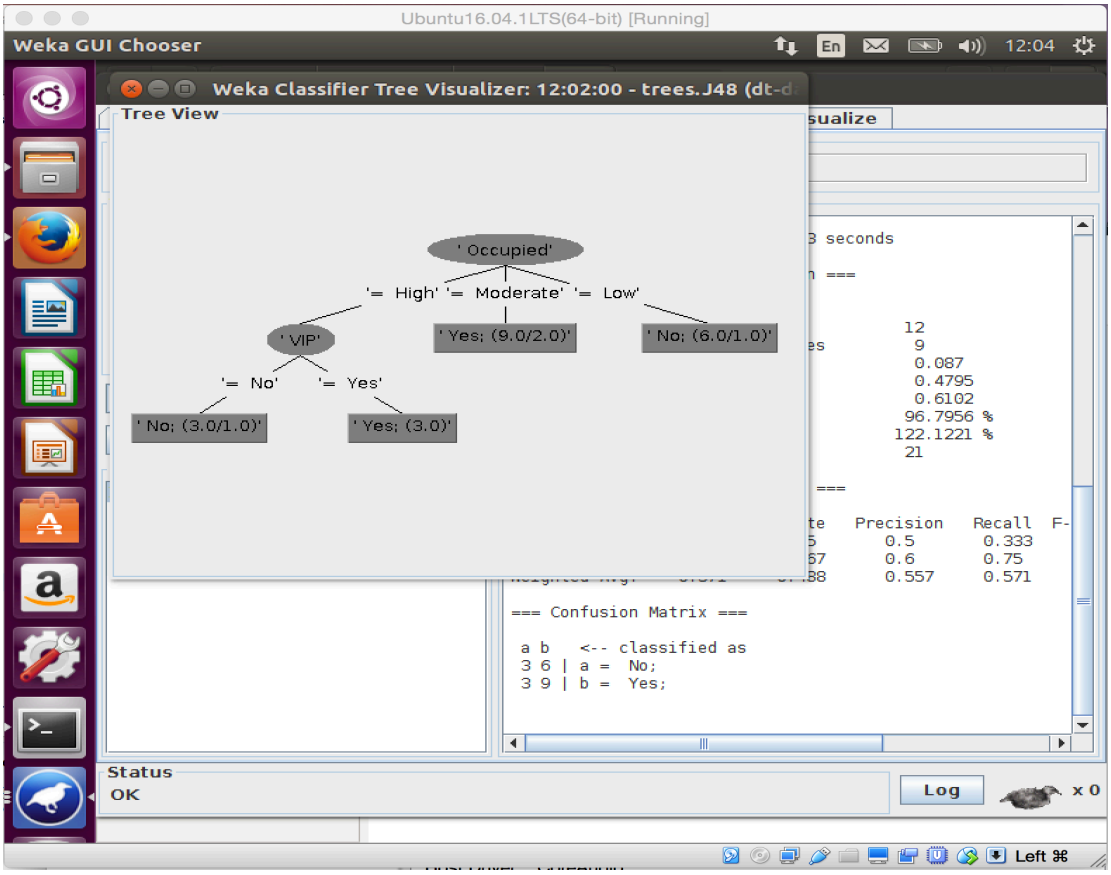




(2). Choose decision tree algorithm:



This is the result of dt-data.txt on weka(I run this library several times so the time on picture is different):



Part 3: Applications

On the Internet, I find an interesting and useful application of decision tree: use decision tree to classify the star/cosmic-ray.

Scientists applied decision tree learning to distinguishing stars and cosmic rays in images from Hubble Space Telescope. They use 20 features to label each star or cosmic ray. 2211 pre-classified images were used as training data, and 2282 pre-classified images was used to measure the performance of decision. The result is great, which means the accuracy is over 95%.