**220CT - Data and Information Retrieval**

**Intelligent Systems for Big Data Lab 2. - Data mining tools**

**Task 1** Identify at least two data mining tools not covered in the lectures and produce a script and basic slides for a video to review each of these. Your script should give the strengths and weaknesses of the tools, what features they have and how they compared to other data mining tools. We are looking for about two minutes on each tool. Explain your script to the person next to you for feedback. After the session you might consider creating the videos reviewing the two data mining tools.

**Task 2** Using a data mining tool.

**Running WEKA**

Click the Mylaunch on the desktop.

Search for WEKA.

Install WEKA.

Then the first interface window appears: Weka **GUI Chooser**.



**Download the data**

From moodle download, unzip and extract the iris.arff and credit.arff data. This iris data is features for different types of iris and the credit data includes features to determine if a person is credit worthy.

**We will explore how we can preprocess data in the future sessions.**

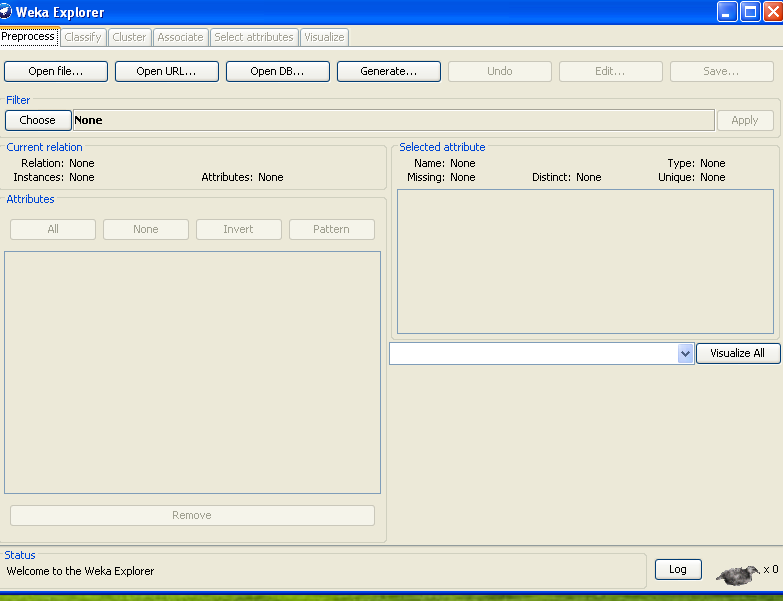
**Lets get started**

First we need to look at the data that we are hoping that the machine learning approach will ‘learn’ and classify.

*Click on the explorer button.*

The explorer window opens.

As the name implies this is where you can look at the data.



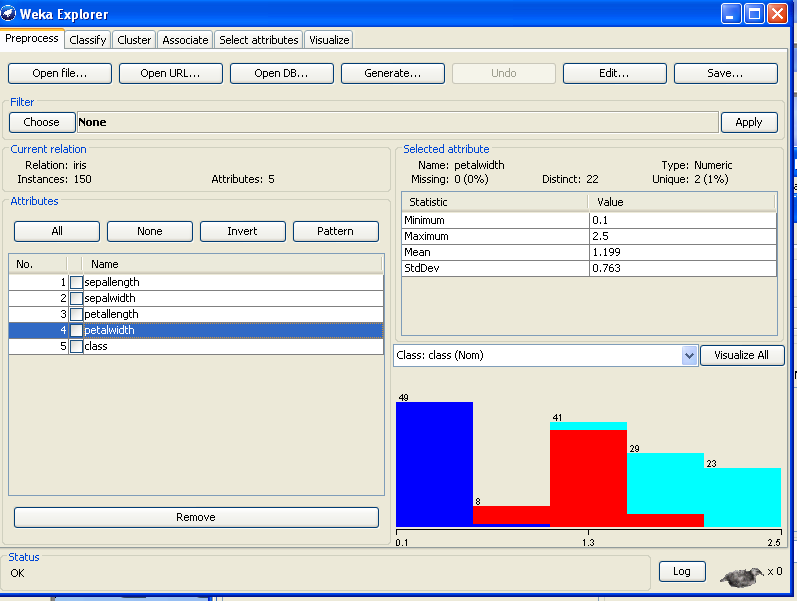
*Press the open file button*

*Go to the directory containing the iris.arff file*

Weka supports various formats including ..arff and .csv. These are simply data files.

*Click on iris.arff*

The iris problem is famous in machine learning. It is typically used when someone invents a new algorithm as it is a good benchmark data. It uses 4 attributes to classify what type of iris the flower is.



*Click on the different attributes*.

The different colours are the different classes of iris. You should see that it is not easy to split the data into the three iris classes.

**Activity**

What are the three different types of iris?

Answer:

In the selected attribute panel there is information on the data. This is valuable as it can give you if you have any extremely high values that might prevent your machine learning approach from learning the data.

**Activity**

Now open the credit.arff.

What are some of the attributes used to determine credit worthiness and what are the classes?

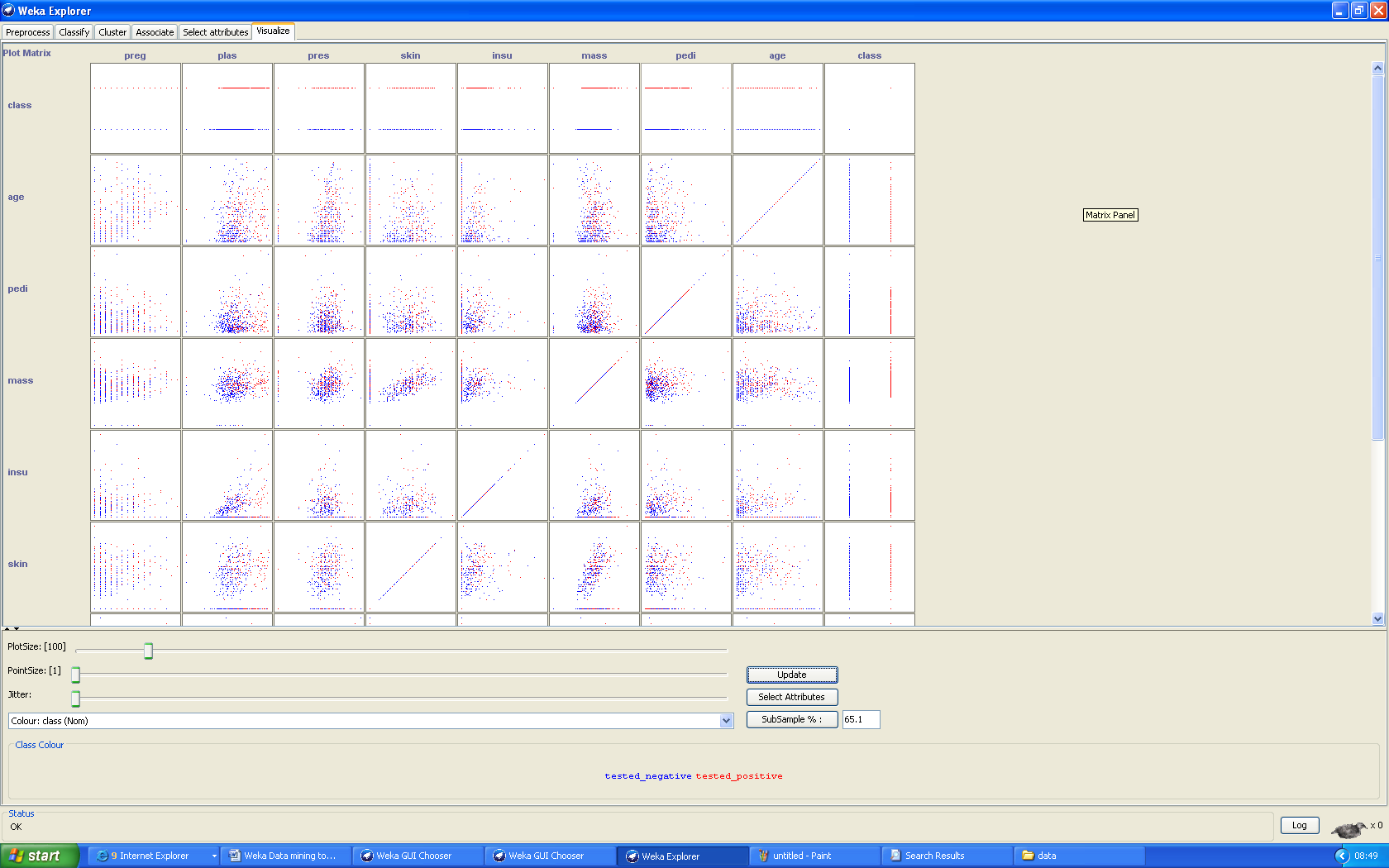
Produce a screenshot of the bar-chart of the values for the credit\_history attribute.

Answer:

**WEKA Visualization**

*For the isis data click on the visualize tab.*

This plots the attribute against attribute. You can see if there is any relationship between these attributes.



Click on a panel and you will see a bigger version of the graph.

**Activity**

For the credit.arff data use the visualisation tab. Produce a screenshot for plot of age against credit amount. Does there appear to be a relationship between these two features?

Answer:

Once you have got a feeling for the data, move on to using machine learning tools to learn and classify the data.

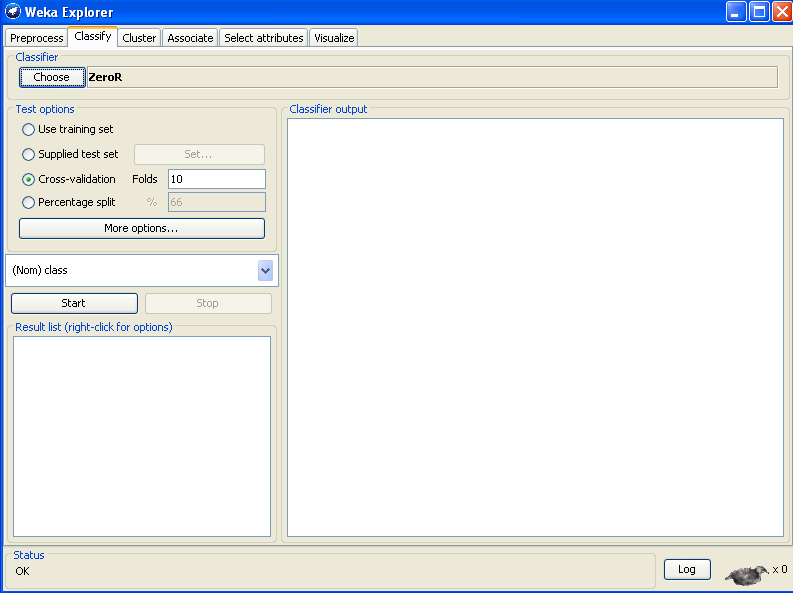
**WEKA Intelligent data mining techniques**

We will now look at how we can get the machine learning approaches to learn data and recognise samples it as not seen before.

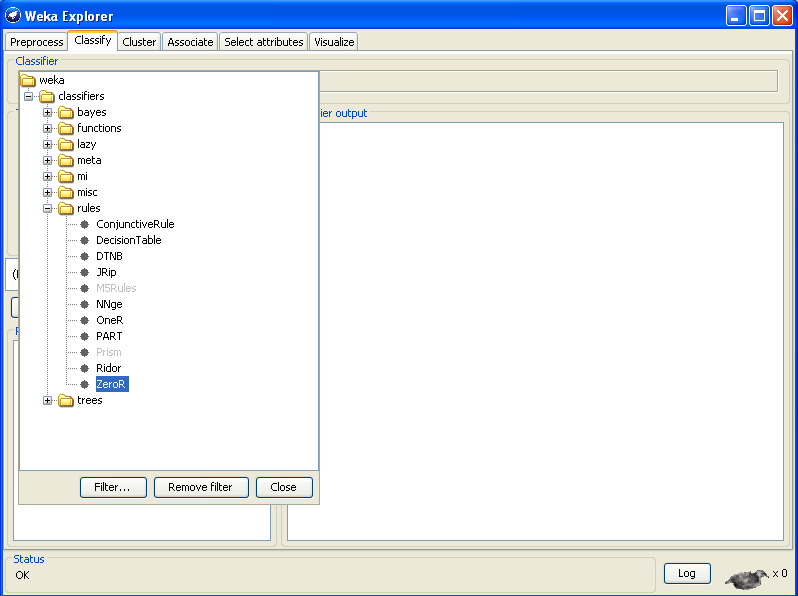
Ensure you have opened the iris data iris.arff.

*Select the classify tab*

*Click the choose button*



There are numerous classifiers to choose from. Each is backed by a specific way of classifying the data. The best approach is very much depends on the data you are using.



**Decision Tree**

*Select trees and J48*

This is a decision tree algorithm.

*Select under test option cross-validation 10 folds split.*

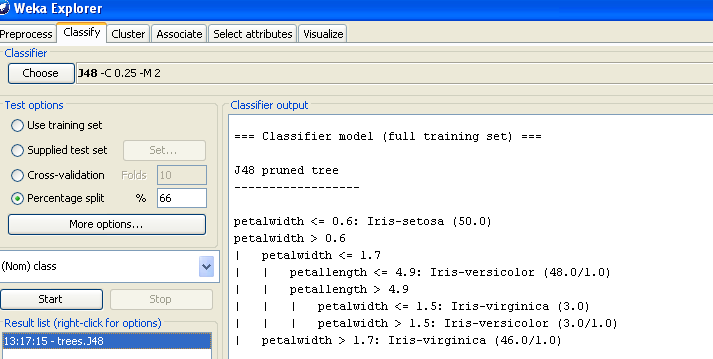
The 10 folds split approach takes the data split it into 10 chunks and uses 9 chunks to train the network and the 10th chunk to test it. It repeats this 10 times with a different chunk for testing.

*Click start*

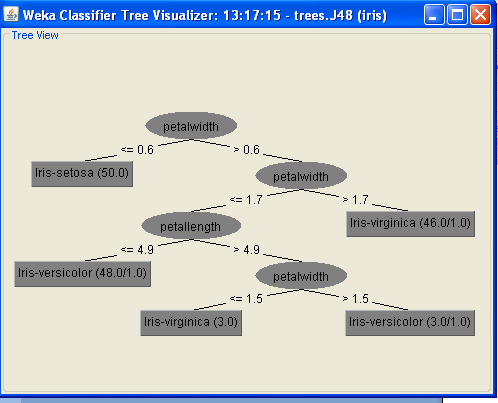
The results appear in the table on screen.

How well did the tree algorithm do in classifying iris?

*Right click on the result list and select visualize tree*



A tree is created that shows the rules that it used to classify the iris into different types.



**Neural Networks**

Data mining would not be complete without neural networks. In this example we will use the multilayer perceptron neural network (MLP) to classify the iris data.

**Activity**

Do some research on the web on what a neural network and is and what a multilayer perceptron is. Keep it simple!!

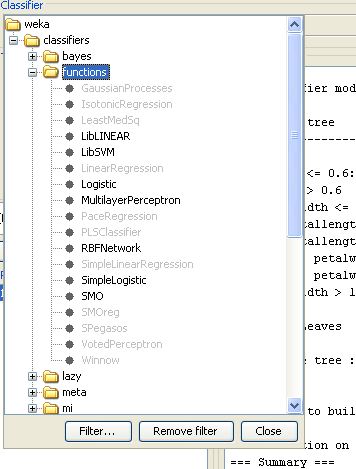
Ensure you have opened the iris data iris.arff.

*Select the classify tab*

*Click the choose button*

*Select Functions*

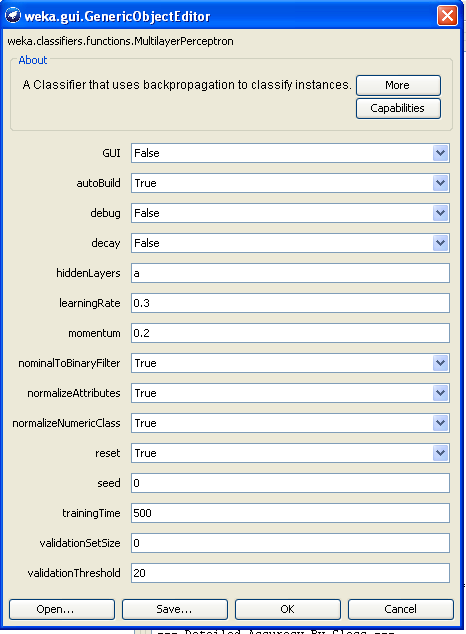
*Select MultilayerPerceptron*



*In the choose line you will see the various parameters used by the multilayer perceptron such as training time.*

*Right click on the choose line and select show properties and you can see each of these parameters.*

*Just press OK at the moment*



*Select cross validation 10-folds in test options*

*Press start*

**Activity**

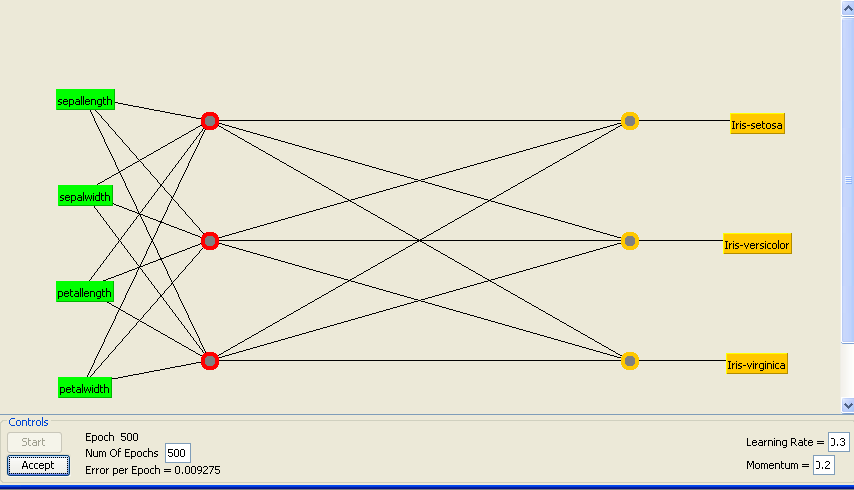
What results did this neural network give? Take a look at the confusion matrix. How many of the test samples for each type of iris does it get right and how many wrong?

Answer:

*In properties try changing GUI to true.*

*Press start*

You should see the structure of the current neural network. The red units are the units in the hidden layer this is where the learning takes places. The yellow units are the outputs. A value is created here with the highest being the class that the sample is classified as.



**Activity**

With the credit data create a MLP neural network (leave all parameters unchanged).

Create a screenshot of the network’s performance. Describe in a couple of sentences how successful the MLP is in classifying unseen samples of people’s credit worthiness. Look at the confusion matrix to see how many it gets wrong or right.

Currently in the hidden layer the value is a, which sets the number of units in the hidden layer to the (number of attributes+number of classes)/2. Try using the value 8 instead. Also reduce the training time to 200. Press start and create a screen shot showing the performance of this network. Was it better or worse than before? How does changing these parameters influence the performance of the neural network?

Answer:

**Clustering**

Clustering is the process by which the machine learning approach positions like data in the same location on a map. The advantage of clustering approaches is the do not need to be told the correct classes they work it out themselves.

*Ensure the iris.arff data is open*

*Select the cluster tab*

*Select the EM clustering technique*

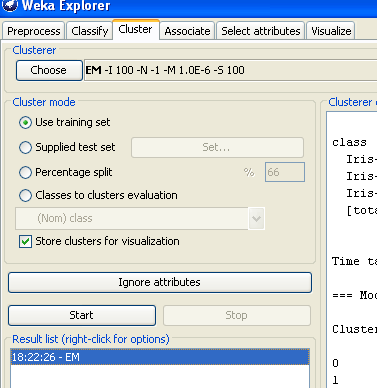
*Click start*

**Activity**

How many clusters does it create? Do they match the different iris?

Answer:

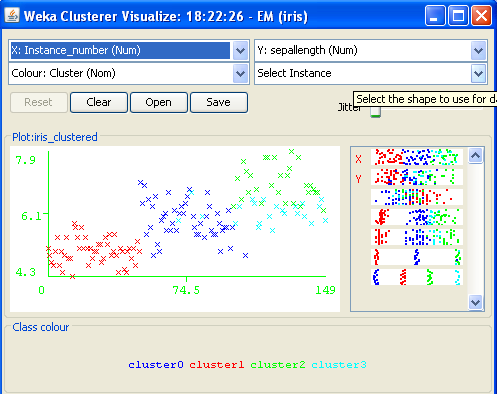
*Under results list right click on the EM*



*Select visualize clusters assigned*

It shows the clusters that have been created for the iris data.

If you double click on one of the crosses it shows you what the original data point was.



**Activity**

Try this clustering approach with the credit.arff data.

Create a screenshot showing the visualisation of the clusters. Does the clustering method create two clear groups that match whether a person is or is not credit worthy.

Answer: