**WEKA Machine Learning:** Running Basic Algorithms

**Aim:** To show how to run several algorithms on datasets to get an idea of how many types of data we have. We shall use 3 algorithms commonly used to get a first feel of data. We also look at basic data cleaning.

**Data files needed:** iris.arff and ionosphere.arff

**Algorithms explored:** ZeroR, 1R, SVM to determine key attributes.

1. **ZeroR:**

Open the file in the Pre-process tab.

In the Classify tab with the Choose button select Rules >

Classifiers > ZeroR

In the confusion matrix we get a feel that there are 2 categories

=== Confusion Matrix == a b <-- classified as

6 2 | a = A 8 0 | b = B

1. **OneR (1R)**

Open the file in the Pre-process tab.

In the Classify tab with the Choose button select Rules > Classifiers > OneR (1R)

In the confusion matrix we get a feel that there are 3 categories

=== Confusion Matrix ===

a b c <-- classified as 50 0 0 | a = Iris-setosa

0 44 6 | b = Iris-versicolor

1. 6 44 | c = Iris-virginica
2. **SVM (Support Vector Machine)**

Open the file in the Preprocess tab.

◦ Examine the data with the **Edit** button in **Pre-process**

You will see a table of the data. The columns are:

No. = the number of the row of the data

Sepal and petal measurements which are the 4 rows of data

class = the type of flower so we can train the system to categorise flower types

The Selected Attribute shows: there is 0 data missing values, 35 rows of data, 9 unique items of data, mean average of each attribute and the range of measurements (minimum and maximum).

Below the attributes area we can see a coloured graph which indicates how many types there may be.

To run a VSM (vector space machine) called SMO in Weka:

In the Classify tab with the Choose button select

◦ functions > SMO

▪ click in the command line next to the Choose button and change

· filterType – No

· click in Kernel exponent and set it to 2 (to force Weka to use an SVM)

◦ then press Start

We see in the confusion matrix that there are 3 types of iris flower a, b, c.

=== Confusion Matrix ===

a b c <-- classified as

50 0 0 | a = Iris-setosa

0 47 3 | b = Iris-versicolor

0 3 47 | c = Iris-virginica

**Using Filters to Prepare/Clean Data**

WEKA implements pre-processing of data by means of the editor (as seen above) and Filters. We list some of the main filters. Filters are selected in the Pre-processing tab by the Filters button. Use the ionosphere dataset.

Filters can make very large datasets smaller in order for them to processed on less powerful systems. Or they can randomise the order of the data for better machine learning (even adding noise to data).

* **To reduce the dataset size use:** Filter > Supervised> Instance > Resample and select e.g. 50%
* **To merge data ranges e.g. income into low/medium/high:** Filter> Supervised> Attribute> Discretize
* **To reorder data for better processing**: Filter> Unsupervised> Reorder
* **To add noise to improve some algorithms**: Filter> Unsupervised > AddNoise
* To automatically reduce attributes: Filter> Unsupervised > PrincipalComponents

Filters also standardize the ranges of data.

• **Normalise data to -1 to +1:** Unsupervised> Standardise

Filters can also pick out a subset of features to process to make processing more efficient.

• Removing attributes: Filter> Unsupervised> Remove and indicate column e.g. 1 and inverse

Filters can fill in missing values:

* With the ionosphere freshly loaded use edit to select some values in a column and delete them
* You will see a number of missing values in the attribute window
* Now we will replace these values automatically with a filter: Filter> Unsupervised> Attribute > ReplaceMissingValues
* Then go back to edit and see what values have replaced the missing ones