An Introduction to Weka

Objectives

- An introduction to the WEKA Explorer environment
- Creating an ARFF file and reading it into WEKA

Tasks

The WEKA GUI Chooser window is used to launch WEKA's graphical environments. WEKA Explorer is an environment for exploring data with WEKA. In this lab, we will be focusing on creating an ARFF file and reading it into WEKA, and using the WEKA Explorer.

ARFF File

Attribute-Relation File Format (ARFF) is a file format recognized by WEKA. An ARFF file typically has a *.arff* extension and contains two sections – a Header section and a Data section.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. An example header on the standard Cricketer's dataset looks like this:

```
% 1. Title: Cricketer's Data
%
% 2. Sources:
%      (a) Creator: DM Instructor
%      (b) Data Source: STAR CRICKET
%      (c) Date: January, 2012
%
@RELATION cricketers

@ATTRIBUTE name                STRING
@ATTRIBUTE matches             NUMERIC
@ATTRIBUTE runs                NUMERIC
@ATTRIBUTE wickets             NUMERIC
@ATTRIBUTE catches             NUMERIC
@ATTRIBUTE battingaverage      NUMERIC
@ATTRIBUTE class {Batsman,Bowler,Wicketkeeper,Allrounder}
```

The **Data** of the ARFF file looks like the following:

```
@DATA
Sachin,420,20000,150,165,47.61904762,Batsman
Rahul,350,9854,32,351,28.15428571,Wicketkeeper
Anil,380,3521,468,164,9.265789474,Bowler
Dhoni,165,8564,0,265,51.9030303,Wicketkeeper
```

```
Kohli,82,3698,70,65,45.09756098,Allrounder
Ishant,114,865,165,68,7.587719298,Bowler
Zaheer,265,2569,320,87,9.694339623,Bowler
Sehwag,260,9685,98,106,37.25,Batsman
```

Lines that begin with a % are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive.

Details of ARFF File

The ARFF Header Section

The ARFF Header section of the file contains the relation declaration and attribute declarations.

The @relation Declaration

The relation name is defined as the first line in the ARFF file. The format is:

```
@relation <relation-name>
```

where <relation-name> is a string. The string must be quoted if the name includes spaces.

The @attribute Declarations

Attribute declarations take the form of an orderd sequence of **@attribute** statements. Each attribute in the data set has its own **@attribute** statement which uniquely defines the name of that attribute and it's data type. The order the attributes are declared indicates the column position in the data section of the file. For example, if an attribute is the third one declared then Weka expects that all that attributes values will be found in the third comma delimited column.

The format for the **@attribute** statement is:

```
@attribute <attribute-name> <datatype>
```

where the *<attribute-name>* must start with an alphabetic character. If spaces are to be included in the name then the entire name must be quoted.

The <*datatype*> can be any of the four types currently (version 3.2.1) supported by Weka:

- numeric
- <nominal-specification>
- string
- date [<date-format>]

where <nominal-specification> and <date-format> are defined below. The keywords **numeric**, **string** and **date** are case insensitive.

Numeric attributes

Numeric attributes can be real or integer numbers.

Nominal attributes

Nominal values are defined by providing an <nominal-specification> listing the possible values: {<nominal-name1>, <nominal-name2>, <nominal-name3>, ...}

For example, the class value of the cricketers dataset can be defined as follows:

```
@ATTRIBUTE class {Batsman,Bowler,Wicketkeeper,Allrounder}
```

Values that contain spaces must be quoted.

String attributes

String attributes allow us to create attributes containing arbitrary textual values. This is very useful in text-mining applications, as we can create datasets with string attributes, then write Weka Filters to manipulate strings (like StringToWordVectorFilter). String attributes are declared as follows:

```
@ATTRIBUTE name    string
```

Date attributes

Date attribute declarations take the form:

```
@attribute <name> date [<date-format>]
```

where <name> is the name for the attribute and <date-format> is an optional string specifying how date values should be parsed and printed (this is the same format used by SimpleDateFormat). The default format string accepts the ISO-8601 combined date and time format: "yyyy-MM-dd'T'HH:mm:ss".

Dates must be specified in the data section as the corresponding string representations of the date/time (see example below).

ARFF Data Section

The ARFF Data section of the file contains the data declaration line and the actual instance lines.

The @data Declaration

The **@data** declaration is a single line denoting the start of the data segment in the file. The format is:

```
@data
```

The instance data

Each instance is represented on a single line, with carriage returns denoting the end of the instance.

Attribute values for each instance are delimited by commas. They must appear in the order that they were declared in the header section (i.e. the data corresponding to the nth **@attribute** declaration is always the nth field of the attribute).

Missing values are represented by a single question mark, as in:

```
@data
Anil,380,3521,468,164,9.265789474,Bowler
```

Values of string and nominal attributes are case sensitive, and any that contain space must be quoted, as follows:

```
@relation CricketersAge

@attribute name string
@attribute age string

@data
'Sachin Tendulkar',37
'Anil Khumble',42
'Rahul Dravid',39
```

Dates must be specified in the data section using the string representation specified in the attribute declaration. For example:

```
@RELATION Timestamps

@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"

@DATA
"2001-04-03 12:12:12"
"2001-05-03 12:59:55"
```

Sparse ARFF files

Sparse ARFF files are very similar to ARFF files, but data with value 0 are not be explicitly represented.

Sparse ARFF files have the same header (i.e **@relation** and **@attribute** tags) but the data section is different. Instead of representing each value in order, like this:

```
@data
0, X, 0, Y, "class A"
0, 0, W, 0, "class B"
```

the non-zero attributes are explicitly identified by attribute number and their value stated, like this:

```
@data
{1 X, 3 Y, 4 "class A"}
{2 W, 4 "class B"}
```

Each instance is surrounded by curly braces, and the format for each entry is: <index> <space> <value> where index is the attribute index (starting from 0).

Note that the omitted values in a sparse instance are **0**, they are not "missing" values! If a value is unknown, you must explicitly represent it with a question mark (?).

TASK TO DO

1. Copy the data given in the file *TestData.doc*, to an Excel sheet.
2. Save the data set as CSV format.
3. Open it with a word processor and format it according to the ARFF specifications. Save as *data1.arff*.
4. It should exactly like the sample data mentioned in page 1 and 2.

The WEKA Explorer

Section Tabs
At the very top of the window, just below the title bar, is a row of tabs. When the Explorer is first started only the first tab is active. The tabs are as follows: Pre-process, Classify, Cluster, Associate, Select Attributes, and Visualize.

Status Box
The status box appears at the very bottom of the window. It displays messages that keep you informed about what's going on.

Opening files
The first button at the top of the preprocess section ***Open File*** enables us to load data into WEKA. Clicking that button brings up a dialogue box allowing you to browse for the data file on the local file system. Using the Open File button, read in the ARFF file you already created in this lab.

The Current Relation
Once the data has been loaded, the Preprocess panel shows a variety of information. The **Current Relation** box displays three entities – the name of the relation, the number of attributes in the data, and the number of instances in the data.

Attributes
Below the **Current Relation** box is a box titled **Attributes**. There are three buttons and beneath them is a list of attributes in the current relation. The three buttons – **All**, **None**, and **Invert** can be used to select desired attributes from the list.

When you click on different rows in the list of attributes, the fields change in the box to the right titled **Selected Attribute**. This box displays the characteristics of the currently highlighted attribute, namely – **Name**, **Type**, **Missing**, **Distinct**, and **Unique**.

Below these is a list showing more information about the values stored in this attribute, which differ depending on its type. For instance, if the attribute is numeric, the list gives four statistics describing the distribution of value in the data – the minimum, maximum, mean, and standard deviation. And below these is a colored histogram, color-coded according to the attribute chosen as the ***Class*** using the box above the histogram. Note that only nominal ***Class*** attributes will result in a color-coding. After pressing the Visualize All button, histograms for all the attributes are shown in a separate window

Desired attributes can be removed by using the **Remove** button below the list of attributes. This can be undone by clicking the Undo button which is located in the top-right corner of the **Preprocess** panel. The **Edit** button next to it can be used to modify your data manually in a dataset editor.