Department of
Computer Science and Engineering

Title: Introduction to WEKA

Data Mining Lab

CSE 424

Green University of Bangladesh

1 Objective(s)

- To understand WEKA tool usage to analysis training data.

2 WEKA Introduction

In order to experiment with the application, first we will run the program and we can see the window like figure 1. Now we will choose "Explorer" option and the next window in figure 2 will open where the data set needs to be presented to WEKA in a format the program understands. There are rules for the type of data that WEKA will accept and three options for loading data into the program.

- Open File- allows for the user to select files residing on the local machine or recorded medium
- Open URL- provides a mechanism to locate a file or data source from a different location specified by the user
- Open Database- allows the user to retrieve files or data from a database source provided by the user

Figure 1: WEKA GUI

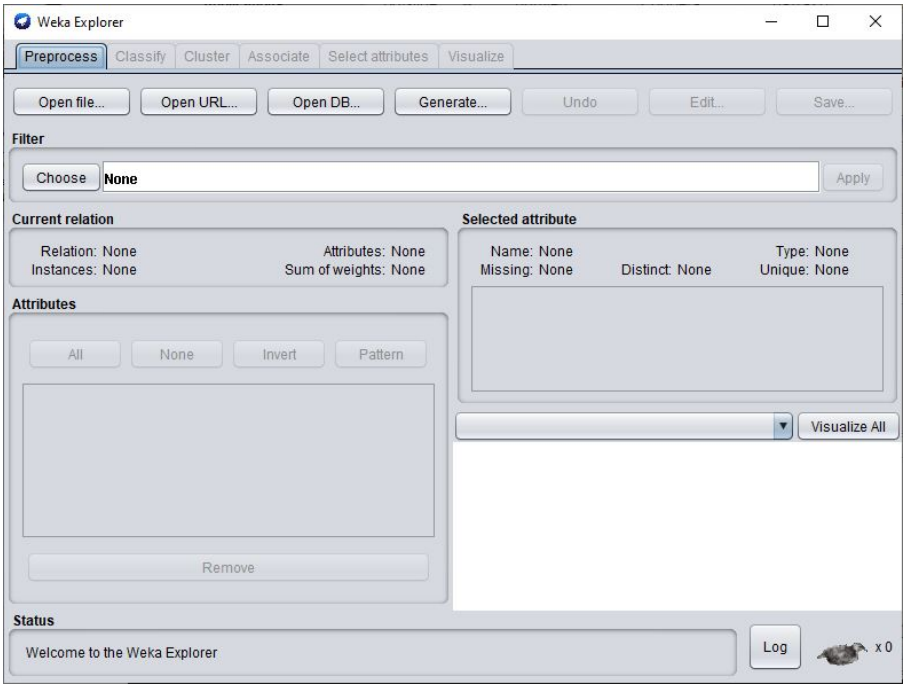

Figure 2: WEKA Explorer

2.1 Pre-processing

At the very beginning lets open a data file by the explorer form local machine. Usually after installation the default data files are located in C:\Program Files\Weka-3-8\data, if the WEKA installed version is Weka 3.8. Figure 3 is the opening of "*weather.numeric.arff*" using WEKA explorer and figure 4 is representing the contents of the file. The visual representation of the attributes and their counts are showing in figure 5 after opening the file. You can select different attributes by using cursor and see different visualization and count values of instances present in the data file.

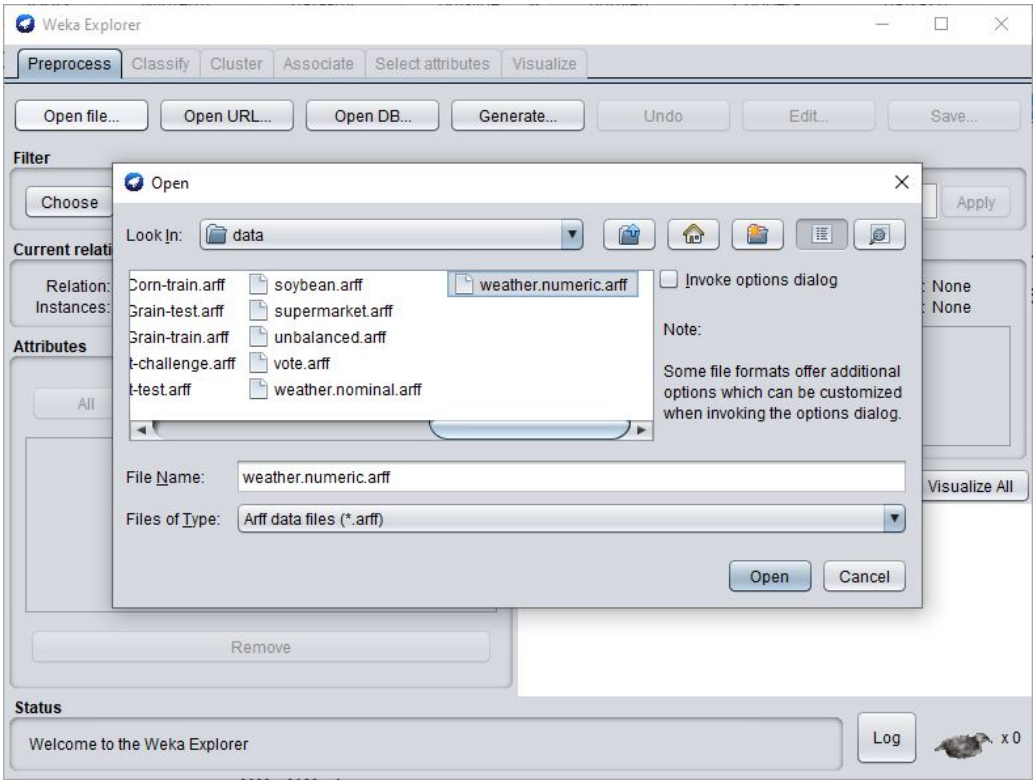

Figure 3: Opening "weather.numeric.arff" file

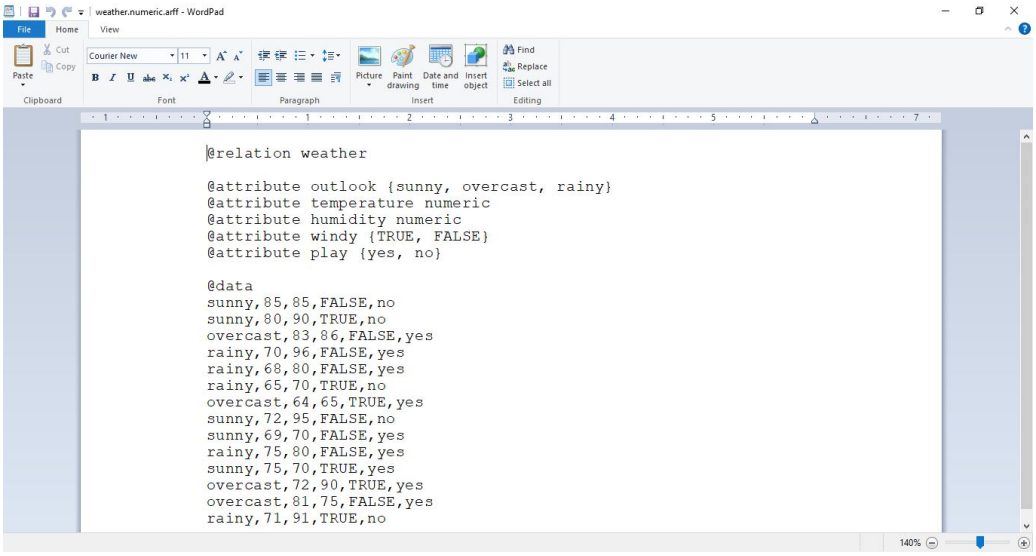

Figure 4: Contents of "weather.numeric.arff" file

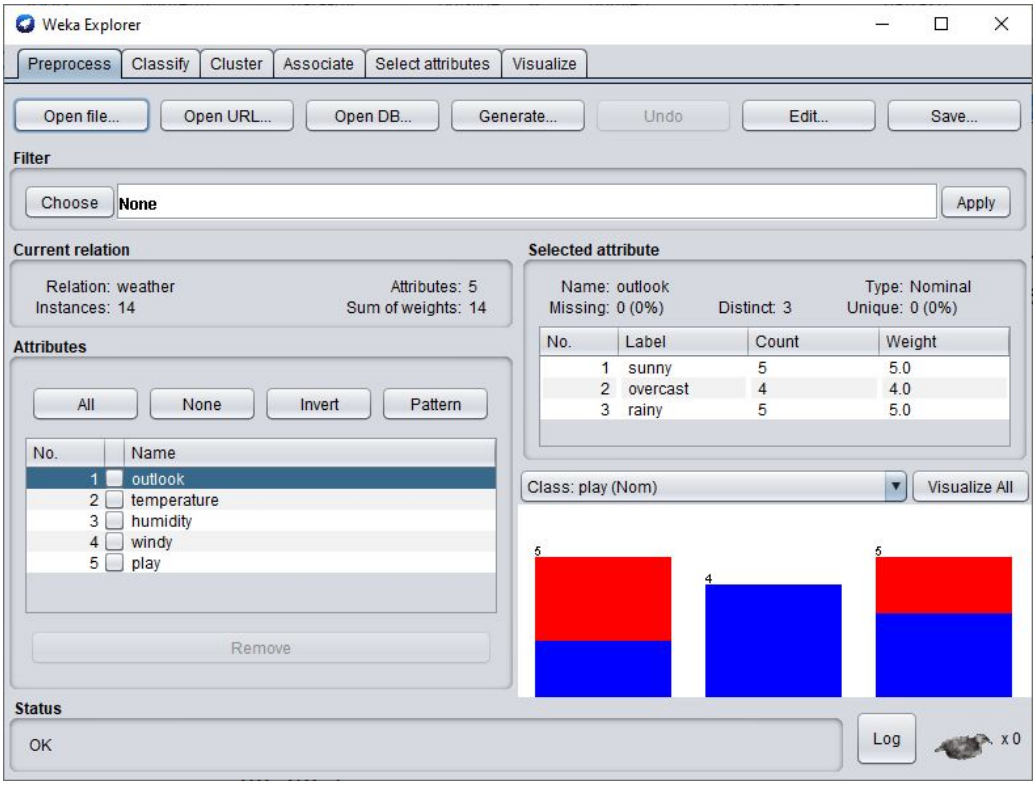

Figure 5: Visual representation of the attributes of "weather.numeric.arff" file

2.2 Classification

The user has the option of applying many different algorithms to the data set in order to produce a representation of information. The best approach is to independently apply a mixture of the available choices and see what yields something close to the desired results. The Classify tab is where the user selects the classifier choices. Following output subsection shows some of the categories.

2.2.1 Output

```
1  === Run information ===
2
3  Scheme:       weka.classifiers.rules.ZeroR
4  Relation:     weather
5  Instances:    14
6  Attributes:   5
7                outlook
8                temperature
9                humidity
10               windy
11               play
12 Test mode:    10-fold cross-validation
13
14 === Classifier model (full training set) ===
15
16 ZeroR predicts class value: yes
17
18 Time taken to build model: 0 seconds
19
20 === Stratified cross-validation ===
21 === Summary ===
22
```

```
Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0
Mean absolute error                      0.4762
Root mean squared error                  0.4934
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC
                    Area  PRC Area  Class
                 1.000    1.000    0.643      1.000    0.783      0.000    0.178
                          0.555     yes
                 0.000    0.000    0.000      0.000    0.000      0.000    0.178
                          0.318     no
Weighted Avg.    0.643    0.643    0.413      0.643    0.503      0.000    0.178
         0.470

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 5 0 | b = no
```

2.3 Clustering

The Cluster tab opens the process that is used to identify commonalities or clusters of occurrences within the data set and produce information for the user to analyze. There are a few options within the cluster window that are similar to those described in the Classify tab. These options are: use training set, supplied test set and percentage split. The fourth option is classes to cluster evaluation, which compares how well the data compares with a pre-assigned class within the data. While in cluster mode, users have the option of ignoring some of the attributes from the data set. This can be useful if there are specific attributes causing the results to be out of range, or for large data sets. Following output subsection shows the Cluster window and some of its options.

2.3.1 Output

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-
    iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data

=== Clustering model (full training set) ===

EM
==
Number of clusters selected by cross validation: 1
Number of iterations performed: 2
```

```
                Cluster
Attribute            0
                    (1)
=======================
outlook
  sunny               6
  overcast            5
  rainy               6
  [total]            17
temperature
  mean          73.5714
  std. dev.      6.3326

humidity
  mean          81.6429
  std. dev.      9.9111

windy
  TRUE                7
  FALSE               9
  [total]            16
play
  yes                10
  no                  6
  [total]            16

Time taken to build model (full training data) : 0.13 seconds

=== Model and evaluation on training set ===

Clustered Instances
0       14 (100%)

Log likelihood: -9.4063
```

2.4 Association

The associate tab opens a window to select the options for associations within the data set. The user selects one of the choices and presses start to yield the results. There are few options for this window and one of the most popular, Apriori, is shown in output subsection given below.

2.4.1 Output

```
=== Run information ===

Scheme:       weka.associations.FilteredAssociator -F "weka.filters.MultiFilter
    -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues \"" -c -1 -W
    weka.associations.Apriori -- -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
    -c -1
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
```

2.5 Selecting Attributes

The next tab is used to select the specific attributes used for the calculation process. By default all of the available attributes are used in the evaluation of the data set. If the user wanted to exclude certain categories of the data they would deselect those specific choices from the list in the cluster window. This is useful if some of the attributes are of a different form such as alphanumeric data that could alter the results. The software searches through the selected attributes to decide which of them will best fit the desired calculation. To perform this, the user has to select two options, an attribute evaluator and a search method. Once this is done the program evaluates the data based on the subset of the attributes, then it performs the necessary search for commonality with the date. Figure 8 shows the opinions of attribute evaluation.

2.5.1 Output

```
1  === Run information ===
2
3  Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
4  Search:       weka.attributeSelection.BestFirst -D 1 -N 5
5  Relation:     weather
6  Instances:    14
7  Attributes:   5
8                outlook
9                temperature
10               humidity
11               windy
12               play
13 Evaluation mode:    evaluate on all training data
14
15 === Attribute Selection on all input data ===
16
17 Search Method:
18   Best first.
19   Start set: no attributes
20   Search direction: forward
21   Stale search after 5 node expansions
22   Total number of subsets evaluated: 11
23   Merit of best subset found:    0.196
24
25 Attribute Subset Evaluator (supervised, Class (nominal): 5 play):
26   CFS Subset Evaluator
27   Including locally predictive attributes
28
29 Selected attributes: 1,4 : 2
30                       outlook
31                       windy
```

2.6 Visualization

The last tab in the window is the visualization tab which is shown in figure 6. Using the other tabs in the program, calculations and comparisons have occurred on the data set. Selections of attributes and methods of manipulation have been chosen. The final piece of the puzzle is looking at the information that has been derived throughout the process. The user can now actually see the data displayed in a two dimensional representation of the information. The first screen that the user sees when they select the visualization option is a matrix of plots representing the different attributes within the data set plotted against the other attributes. If a lot of attributes are selected, there is a scroll bar to view all of the produced plots. The user can select a specific plot from the matrix to analyze its contents in a larger, popup window. A grid pattern of the plots allows the

user to select the attribute positioning to their liking for better understanding. Once a specific plot has been selected, the user can change the attributes from one view to another.

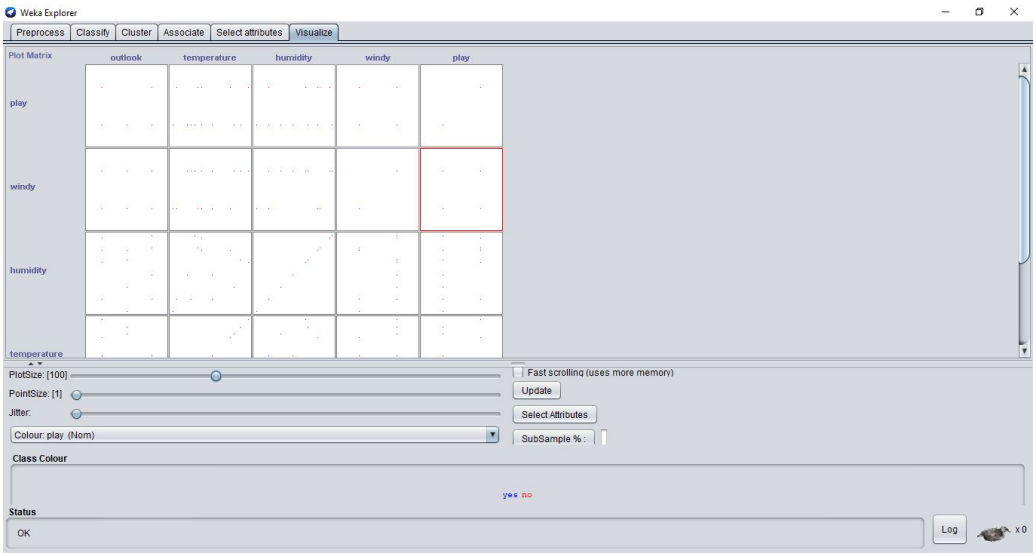

Figure 6: Visualization tab

3 Discussion & Conclusion

Based on the focused objective(s) to understand the use of WEKA tool and practice classification, clustering and attribute selection etc. The additional lab exercise will increase confidence towards the fulfilment of the objectives(s).

4 Lab Exercise

- Open new data file and configure WEKA for that particular file. Now run the above mentioned methods in section 2 observe the results and generate inner reflection.
- Lab report will be given in the next lab after other experiment on WEKA or different topic of Data Mining and Machine Learning.

5 Policy

Copying from internet, classmate, seniors, or from any other source is strongly prohibited. 100% marks will be *deducted* if any such copying is detected for lab exercise.