Experiment 1: Installation of WEKA Tool

Aim: A. Investigation the Application interfaces of the Weka tool. Introduction:

Introduction

Weka (pronounced to rhyme with Mecca) is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques
- Ease of use due to its graphical user interfaces.

Description:

Open the program. Once the program has been loaded on the user's machine it is opened by navigating to the program's start option and that will depend on the users operating system. Figure 1.1 is an example of the initial opening screen on a computer.

There are four options available on this initial screen:

Fig: 1.1 Weka GUI

1. Explorer - the graphical interface used to conduct experimentation on raw data After clicking the Explorer button the weka explorer interface appears.

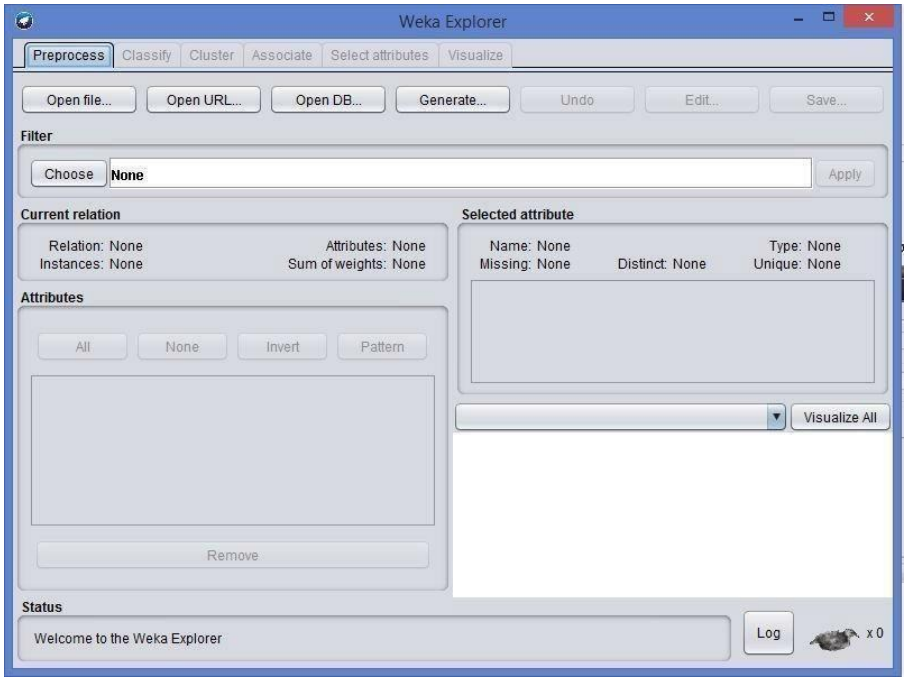
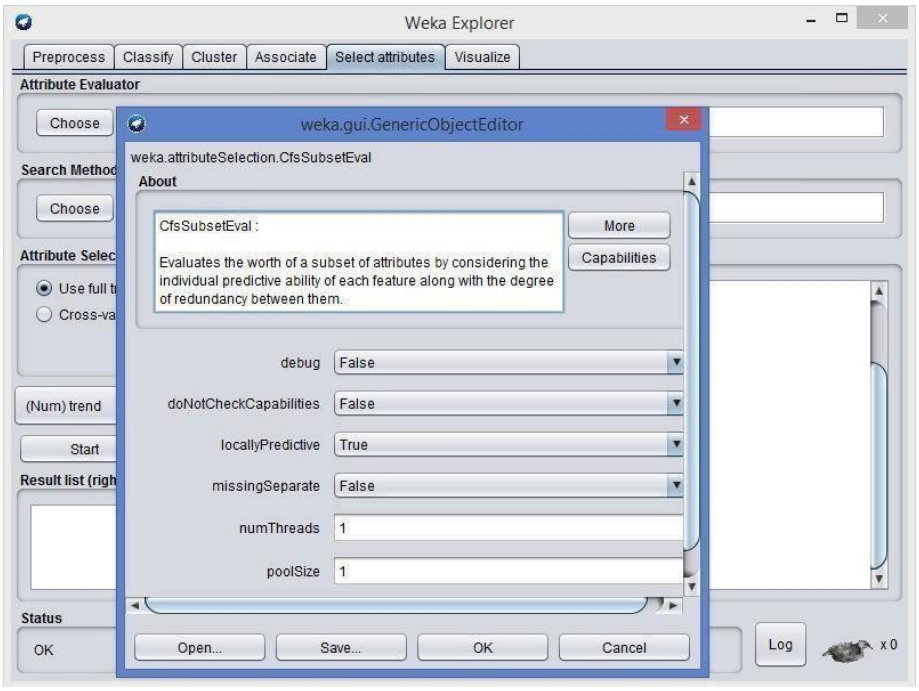
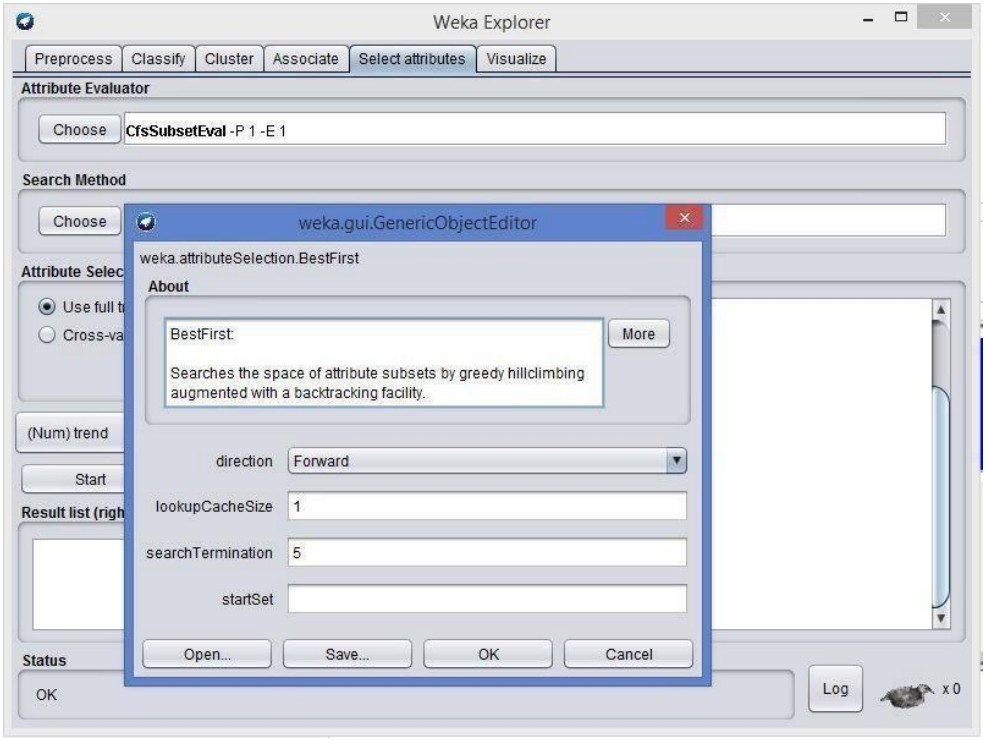


Fig: 1.2 Pre-processor


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
weka.gui.GenericObjectEditor
weka.attributeSelection.CfsSubsetEval
About
CfsSubsetEval :
Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
More
Capabilities
debug
False
doNotCheckCapabilities
False
locallyPredictive
True
missingSeparate
False
numThreads
1
poolSize
1
Open...
Save...
OK
Cancel
Search Method
Choose
Attribute Selec
Use full t
Cross-va
(Num) trend
Start
Result list (righ
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval -P 1 -E 1
Search Method
Choose
weka.gui.GenericObjectEditor
weka.attributeSelection.BestFirst
About
BestFirst:
Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.
More
direction
Forward
lookupCacheSize
1
searchTermination
5
startSet
Open...
Save...
OK
Cancel
Attribute Selec
Use full t
Cross-va
(Num) trend
Start
Result list (righ
Status
OK
Log
x 0

Inside the weka explorer window there are six tabs:

1. **Preprocess-** used to choose the data file to be used by the application.

   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 user

2. **Classify-** used to test and train different learning schemes on the preprocessed data file under experimentation

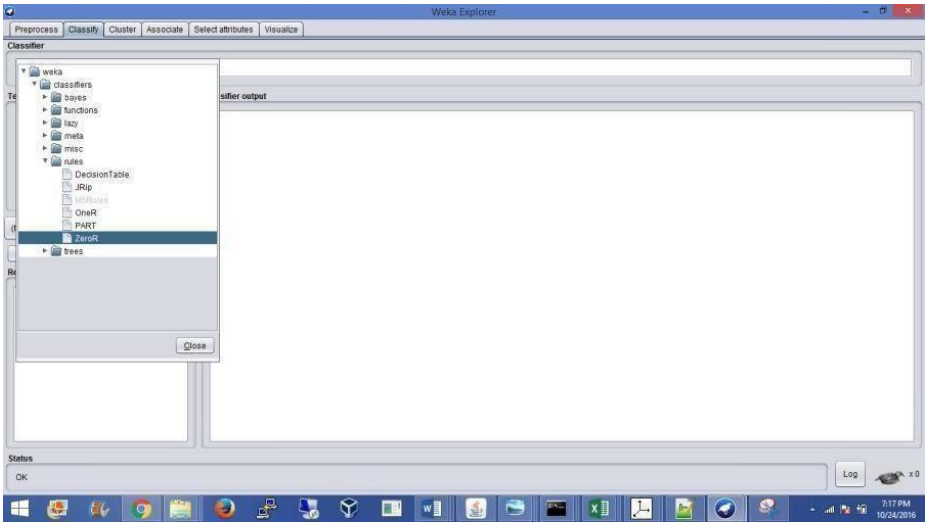


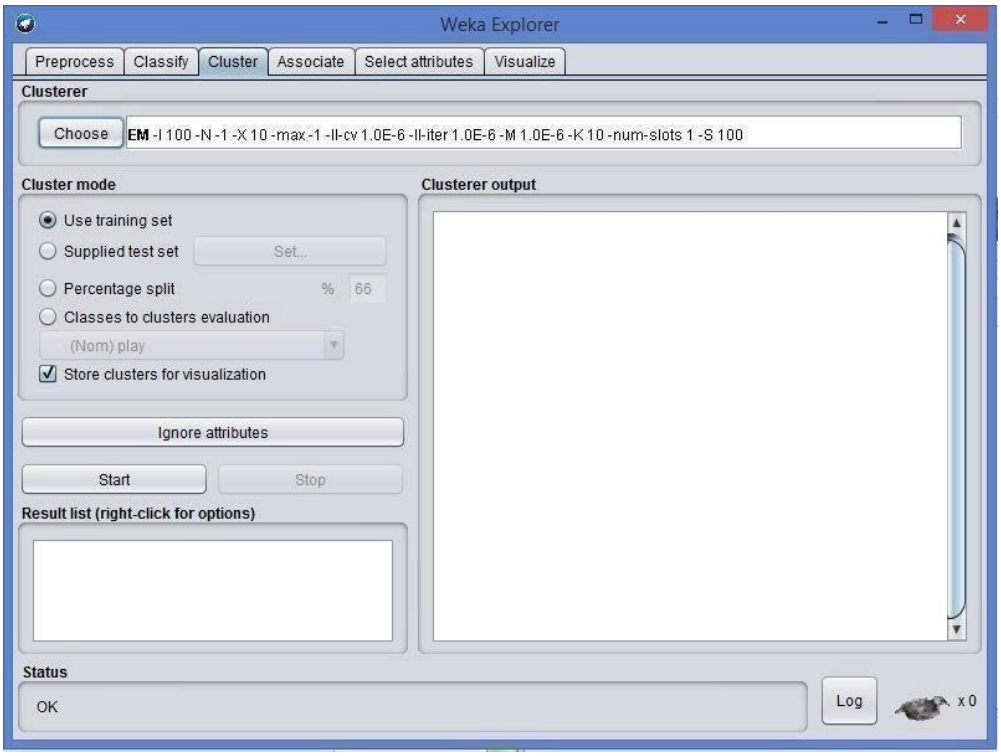
Fig: 1.3 choosing Zero set from classify

Again there are several options to be selected inside of the classify tab. Test option gives the user the choice of using four different test mode scenarios on the data set.

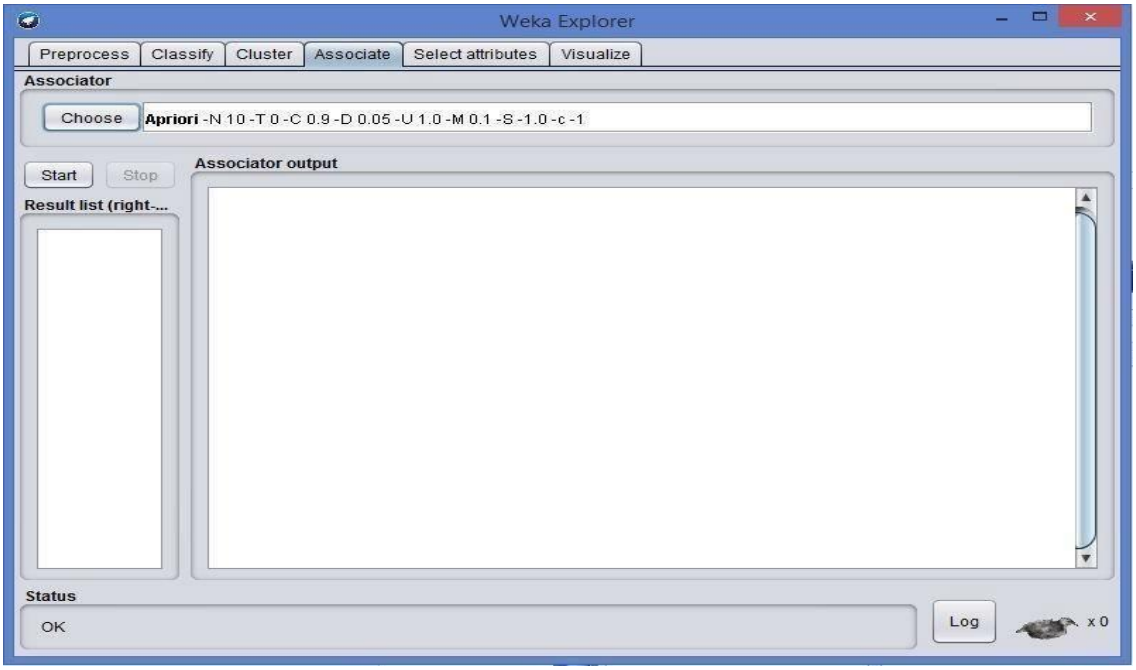
1. Use training set
2. Supplied training set
3. Cross validation
4. Split percentage

3. **Cluster-** used to apply different tools that identify clusters within the data file.

The Cluster tab opens the process that is used to identify commonalties or clusters of occurrences within the data set and produce information for the user to analyze.



4. **Association-** used to apply different rules to the data file that identify association within the data. The associate tab opens a window to select the options for associations within thedataset.



5. **Select attributes-**used to apply different rules to reveal changes based on selected attributes inclusion or exclusion from the experiment

6. **Visualize-** used to see what the various manipulation produced on the data set in a 2D format, in scatter plot and bar graph output.

2. Experimenter - this option allows users to conduct different experimental variations on data sets and perform statistical manipulation. The Weka Experiment Environment enables the user to create, run, modify, and analyze experiments in a more convenient manner than is possible when processing the schemes individually. For example, the user can create an experiment that runs several schemes against a series of datasets and then analyze the results to determine if one of the schemes is (statistically) better than the other schemes.

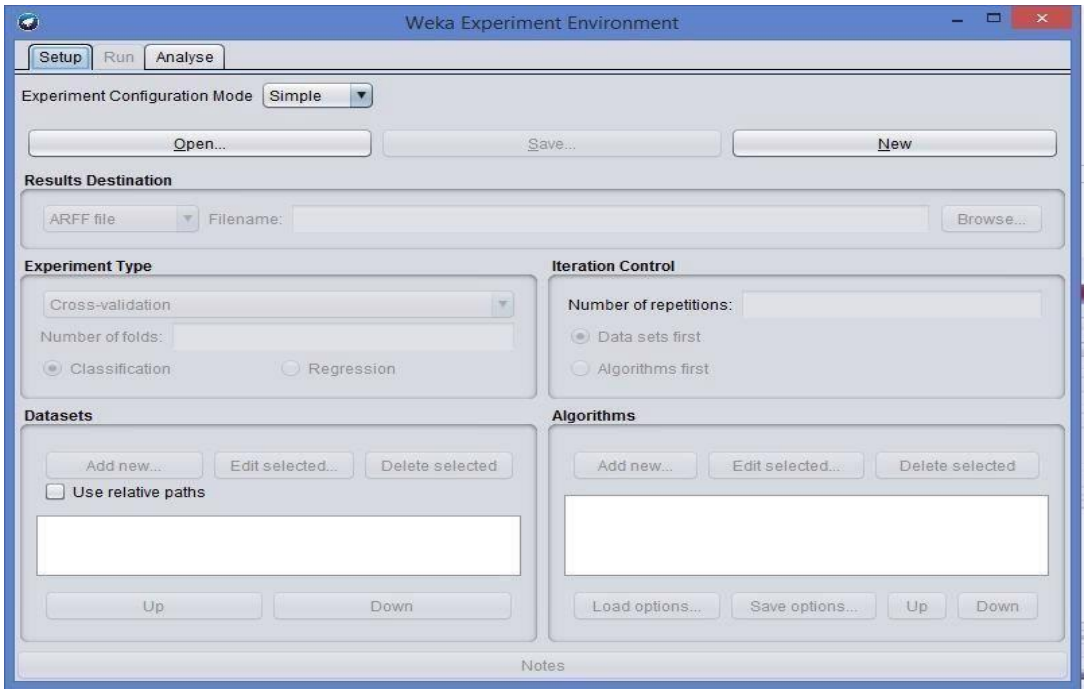


Fig: 1.6 Weka experiment

Results destination: ARFF file, CSV file, JDBC database.

Experiment type: Cross-validation (default), Train/Test Percentage Split (data randomized).

Iteration control: Number of repetitions, Data sets first/Algorithms first.

Algorithms: filters

3. Knowledge Flow -basically the same functionality as Explorer with drag and drop functionality. The advantage of this option is that it supports incremental learning from previous results

4. Simple CLI - provides users without a graphic interface option the ability to execute commands from a terminal window.

b. Explore the default datasets in weka tool.

Click the "***Open file…***" button to open a data set and double click on the "***data***" directory.

Weka provides a number of small common machine learning datasets that you can use to practice on.

Select the "***iris.arff***" file to load the Iris dataset.

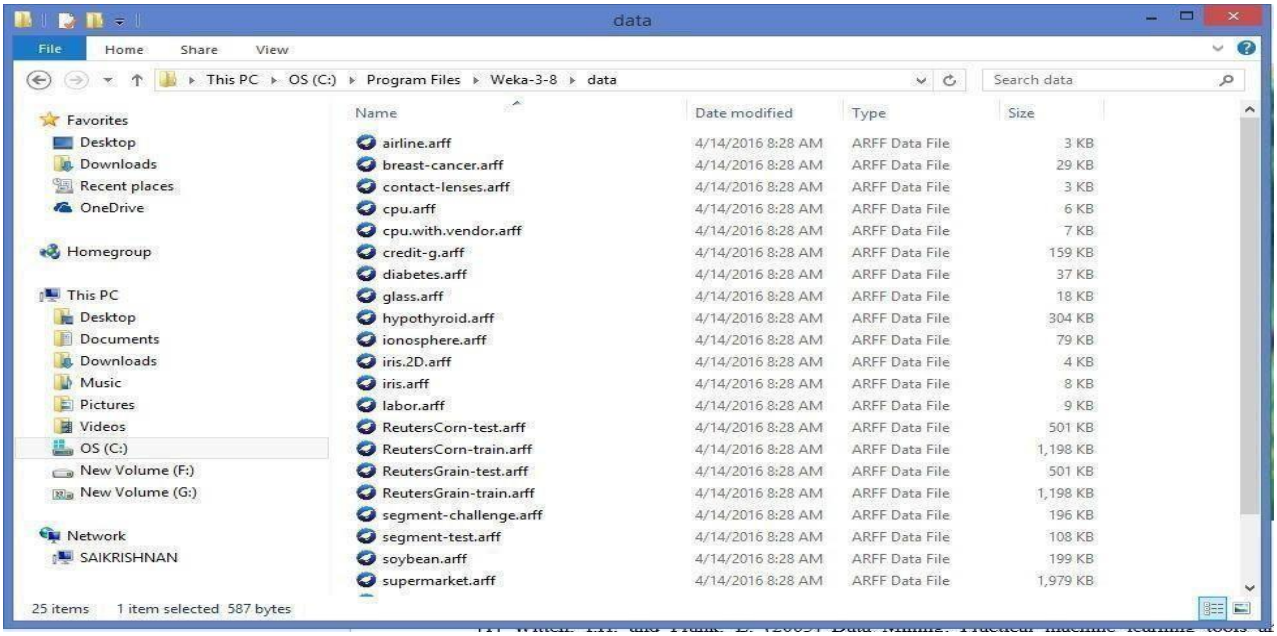


Fig: 1.7 Different Data Sets in weka

References:

[1] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools andtechniques. 2nd edition Morgan Kaufmann, San Francisco.
[2] Ross Quinlan (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.
[3] CVS–http://weka.sourceforge.net/wiki/index.php/CVS
[4] Weka Doc–http://weka.sourceforge.net/wekadoc/

Exercise:

1. **Normalize the data using min-max normalization. Save Screenshots and make word file or pdf file. Submit it with your RollNo_lab2.**