**ASSIGNMENT 1**

**Name: Jain Samkitkumar Hasmukhlal**

**Roll No.: 20 (BEIT)**

**WEKA**

**THEORY:**

**Introduction:**

* **Introduction of WEKA**

Weka is open source software under the GNU General Public License. System is developed at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment for Knowledge Analysis. The software is freely available at http://www.cs.waikato.ac.nz/ml/weka. The system is written using object oriented language Java. There are several different levels at which Weka can be used. Weka provides implementations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering and association rule extraction.

**Concept/Working:**

* **Steps to download and configure the WEKA**

Download Weka (the stable version) from http://www.cs.waikato.ac.nz/ml/weka/

– Choose a self-extracting executable (including Java VM)

– (If you are interested in modifying/extending weka there is a developer

version that includes the source code)

After download is completed, run the self extracting file to install Weka, and use

the default set-ups.

* **Working of WEKA**

The general working steps are given below by considering the example of Hierarchical clustering.

1. Select a dataset for example iris.arff.

2. Select option Cluster

3. Choose cluster type: Hierarchial Cluster

4. Select cluster mode: Training Set.

5. Click on Start.

* **Features of WEKA**

Main features of Weka include:

* 49 data preprocessing tools
* 76 classification/regression algorithms
* 8 clustering algorithms
* 15 attribute/subset evaluators + 10 search algorithms for feature selection.
* 3 algorithms for finding association rules
* 3 graphical user interfaces

– “The Explorer” (exploratory data analysis)

– “The Experimenter” (experimental environment)

– “The KnowledgeFlow” (new process model inspired interface)

**Advantages/Disadvantages:**

* **Advantages of WEKA.**

Free availability:

* Under the GNU General Public License
  + Portability
  + Fully implemented in the Java programming language and thus runs
* on almost any modern computing platforms
  + Windows, Mac OS X and Linux
  + Comprehensive collection of data preprocessing and modeling
* techniques
  + Supports standard data mining tasks: data preprocessing, clustering,
* classification, regression, visualization, and feature selection .
  + Easy to use GUI
  + Provides access to SQL databases
  + Using Java Database Connectivity and can process the result
* returned by a database query.
* The obvious advantage of a package like Weka is that a whole range of data preparation, feature selection and data mining algorithms are integrated. This means that only one data format is needed, and trying out and comparing different approaches becomes really easy. The package also comes with a GUI, which should make it easier to use.
* **Disadvantages of WEKA.**
* Sequence modeling is not covered by the algorithms included in the Weka distribution.
* Not capable of multi-relational data mining.
* Memory bound.
* Do not implement the newest techniques. For example the MLP implemented has a very basic training algorithm (backprop with momentum), and the SVM only uses polynomial kernels, and does not support numeric estimation. Therefore, it will be necessary to combine WEKA with some of the other tools like Netlab or SVM\_torch.
* Though the software is for free: the documentation for the GUI is quite limited.
* Limited scaling. For difficult tasks on large datasets, the running time can become quite long, and java sometimes gives an OutOfMemory error. This problem can be reduced by using the '-mxx' option when calling java, where x is memory size (eg '50m'). For large datasets it will always be necessary to reduce the size to be able to work within reasonable time limits.
* The GUI does not implement all the possible options. Things that could be very useful, like scoring of a test set, are not provided in the GUI, but can be called from the command line interface. So sometimes it will be necessary to switch between GUI and command line.

**Any Software or Hardware Used:**

* **Hardware or software required for WEKA.**

Hardware:

* 4GB RAM

Software:

* Java
* 64-bit / 32-bits versions of Windows.
* 64-bit / 32-bits Linux

**INPUT/OUTPUT:**

**Input:**

Dataset with the attributes or features may or may not be trained. (.ARFF, .CSV, C4.5 and binary)

**Output:**

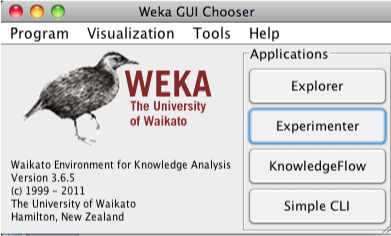
Graphical Representation of analyzed data from Dataset in case of R.

We get the accuracy of successful classification in weka.

**OPERATIONAL STEPS REQUIRED:**

* **Steps to operate WEKA**

Choose “WEKA 3.7.x” from Programs. The first interface that appears looks like the one below.



Click on “Explorer”

To load a data set from library eg. ‘weather.arff.’ So file click on “Open File” and browse the path for ‘weather.arff’.

Select the attributes.

Under the Classify tab, click ‘Choose’ and select a classifier from the drop-down menu. E.g.: ‘Decision Stump’



Once, a classifier is chosen, select percentage split and leave it with its default values. The default ratio is 66% for training and 34% for testing.

Click ‘Start’ to train and test the classifier.

**LATEST TRENDS:**

* **For WEKA**

Weka 3.7.7

**APPLICATIONS:**

* **Application of WEKA:**

The WEKA system has been applied successfully in a variety of areas including the areas of agriculture, machine learning research and education.

**LIMITATIONS:**

* **Limitations of WEKA**

GUI is not as well documented.

2 different Modules cannot be combined (ex. modules for both PCA and clustering without writing a Java Code).

The Weka GUI provides several built-in 'visualization' panels but these are very limited.

Manipulation of data sets is not easy in Weka

**CONCLUSION:**

Downloaded the open source softwares R-base, RStudio and WEKA. Studied the distinct features and functionality of both the software platforms. Found WEKA easier to learn but there are some limitations in case of Graphical Representations, Modifying the dataset etc. R is difficult to learn for novice but its Graphical Representation is better than WEKA.

**REFERENCES:**

[1] [**www.cs.waikato.ac.nz/ml/weka/**](http://www.cs.waikato.ac.nz/ml/weka/)

[2][**https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf**](https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf)