13 / 09 / 23 FAMILIARIZATION OF THE WEKA TOOLKIT 2

Aim : To familiarize

Download, Installation of WEKA data mining tool kit
. Understand the features of WEKA toolkit such as Explorer, knowledge flow interface, Experimenter, command line interface
. Navigate the options available in WEKA

Theory

WEKA (Waikato Environment for knowledge Analysis) is an open source software that provides tools for data preprocessing, implementation of several ML algorithms and visualization tools. It can be used to develop ML techniques and apply them to real world data mining problems. This is fully developed in Java language and provides access to SQL databases using Java Database Connectivity (JDBC).

>> Weka functionalities

The various functionalities include many stages in dealing with big data

i) Raw data collection
ii) Data Pre processing to change the raw data
iii) Application of ML algorithms that is specific to the application, parameters
iv) Output visualization to inspect the data
v) Selection of best ML model specific to the application

>> Installation of WEKA

Visit WEKA's official website and download the installation file and run it to complete the installation

>> Launching WEKA application

From the screen displayed (GUI application) choose how to run the application. The five different types are

i) Explorer ii) Experimenter iii) knowledge Flow iv) Workbench v) Simple CLI

>> WEKA explorer

The different types of ML tabs that can be seen are

i) Pre process ii) Classify iii) Cluster iv) Associate v) Visualize
vi) Select attributes

Under these tabs there are pre implemented ML algorithms:

i) Pre process Tab: Select data file, process it and make it fit for applying the various ML algorithms

ii) Classify Tab: Several supervised and unsupervised algorithms like linear Regression, support Vector machines, Decision Trees, Random Forest, Naive Bayes etc may be applied.

iii) Cluster Tab: Some clustering algorithms are k-means, filtered, hierarchical and so on.

iv) Associate Tab: This includes Apriori, Filtered Associator, FPGrowth

v) Select attributes Tab: Allows feature selection based on algorithms such as ClassifierSubsetEval, Principal Components etc

vi) Visualize Tab: Allows to visualize the processed data for analysis

20/09/2023 DATA PRE-PROCESSING 4

Aim: To understand and perform the following operations

- Study the arff file format
- Explore the available data sets in WEKA
- Load the dataset (Breast Cancer) and observe the following:
  a. list the attribute names and types
  b. Number of records in each data set
  c Identify the class attribute (if any)
  d. Perform preprocessing (min 3)
  e. Plot Histogram

Theory

>> Loading Data

This can be done from the following sources -

i) Local File System

Under the ML tab, click on 'Open file' button. A directory navigator window opens up through which we can navigate to the desired folder. WEKA installation comes with some example databases (C:\Program Files\WEKA-3-8-6\data). The contents of the file would be loaded in the WEKA environment

ii) Web

On clicking 'Open URL...' button which opens up a pop up box. Type any URL where data is stored. The Explorer will load the data from the remote site.

iii) DB

On clicking 'Open DB...' button, a window opens up where we can set the connection string to the input database, set up the query for data selection, process the query and load selected records in WEKA

>> File Formats

- The different types of files supported by WEKA includes arff, arff.gz, bsi, csv, dat, data, json, json.gz, libsvm, m, names, xrff, xrff.gz. The default type is arff.
- Arff format (Attribute-Relation file format)
  → This contains 2 sections - header and data where the header describes the attribute types and and data section contains a comma separated list of data.

>> Exploring datasets in WEKA

- Open the bsuat cancer dataset using 'open file...' option.
- current relation sub window
  → It shows the name of the loaded dataset
  → There are 14 instances (no. of rows)
  → The table contains 5 attributes (5 fields)
- Attributes sub window
  → This appears on the left side and displays various fields in database.
  → The weather database contains 5 fields - outlook, temperature, humidity, windy and play.
- Selected attributes sub window

On selecting an attribute from list, further details can be displayed on the right side. For example, in the temperature attribute the following can be observed

→ The name and type of attribute
→ The type here is Nominal
→ The number of missing values is 0
→ There are 3 distinct values with no unique value
→ The table underneath shows the nominal values - hot, mild etc
→ It shows count, weight in terms of % for each nominal value

The following can be observed

- The @relation tag defines the name of DB
- The @attribute tag defines the attributes
- The @data tag starts the list of data rows separated by comma
- The attribute can take nominal values
  @attribute outlook {sunny, overcast, rainy}
- The attribute can take real values
  @attribute temperature real
- A target, class variable can also be a set
  @attribute play {yes, no}

- Visualization of attributes

  This can be seen at the bottom of the window, on clicking the 'Visualize All' button.

Data preprocessing is a data mining technique to transform raw data into an efficient format. The processed data is then fed to different algorithms for analysis. The various steps involved in data preprocessing are

i) Data Cleaning : It involves handling of missing data, noisy data etc

a. Missing Data : It can be handled in various ways

→ Ignore the tuples (suitable for large data)

→ Fill missing value using mean or most probable value

b. Noisy data : These are the meaningless data that can't be interpreted by machines and is generated due to faulty data collection, data entry errors etc.

→ Binning method (works on sorted data).

→ Regression (Linear or multiple)

→ clustering which help recognize outliers.

ii) Data Transformation : This involves

a. Normalization (scale data values in a specified range)

b. Attribute selection (new attributes are constructed)

c. Discretization (Replaces raw values by interval or conceptual levels)

d. Concept hierarchy generation (Converts from lower to higher hierarchy).

iii) Data reduction : This aims to increase the storage efficiency and reduce data storage and analysis costs

a. Data Cube Aggregation (for construction of data cube)

b. Attribute subset selection (Discards attributes other than highly relevant data)

c. Numerosity reduction (enable storing of data models rather than data)

e. Dimensionality reduction (By encoding mechanisms like wavelet and PCA).

>> Data cleaning using WEKA

i) Replace Missing values - with mode, means from training data

ii) Remove With Values - filter instances according to the value of attributes

iii) InterQuartile Range - Detects outliers and extreme values

iv) Discretize - converts a range of numeric attributes to nominal attributes

v) Normalize - Normalizes all numeric values in the given dataset

vi) Nominal To Binary - Converts all nominal attributes to binary numeric attributes

vii) Numeric To Nominal - Converts numeric to nominal attributes

viii) Remove - Removes a range of attributes

ix) RemoveByName - Removes based on a regular expression matched but will not remove the class attribute

x) RenameAttribute - Used for renaming attributes.

Conclusion