Experiment - 2

Aim: Explore Machine learning tool "WEKA"

→ Explore WEKA Data Mining / Machine Learning Toolkit.

→ Downloding and/or installation of WEKA data mining toolkit.

→ understand the features of WEKA toolkit such as Explorer, Knowledge Flow interface, Experimenter, command-line interface.

→ Navigate the options available in the WEKA (ex. select attributes panel, preprocess panel, classify panel etc.)

→ Study the arff file format Explore the available data sets in WEKA. Load a data set (ex. Weather dataset etc)

→ Load each dataset and observe the following:

1. List the attribute names and they types
2. Number of records in each dataset
3. Identify the class attribute (if any)
4. plot Histogram.
5. Determine the number of records for each class.
6. visulize the data in various dimensions.

Objectives:

Data Ware housing is a technique of gathering and Analyzing data from many sources to get Valuable business insights. Typically a data ware house integrates and analyzes business data from many sources Data Ware housing is a vital Component of business intelligence.

Preproccsser:

The data that is collected from the field contains many unwanted things that leads to wrong analysis. Thus, the data must be preprocessed to meet the requirements of the type of analysis you are seeking. This is the done in the preprocessing module

classifiers:

classifiers in WEKA are the models for predicting nominal & numeric quantities. The learning schemes available in WEKA include decision trees and lists, instance - based classifiers, classifiers include bagging, boosting, stacking, error - correcting output codes and locally weighted learning.

WEKA:

WEKA (Waikota environment for knowledge Analysis) is a popular site of Machine learning software written in Java developed at the university of Waikota. New zealand. Weka is free software available under the GNO General Public License.

Weka - an open source software provides tools for data preprocessing implementation of several Machine learning Algorithms and visualization tools so that you can develop machine learning techniques and apply them to real - world data mining problems.

Explorer: It is an environment for exploring data explorer Consists of several tools. They are :-

→ preprocess:

It is the first step in machine learning is to proprocess the data. It is used to select the data file, process it and make it fit for applying the various machine learning Algorithms.

→ classify:

The classify tab provides you several machine learning algorithms for the classification of your data. Such as linearRegression, logistic Regression.

→ cluster:

under the cluster tab there are several clustering Algorithm provided - such as Simple k means, filtered cluster, Hierarchical cluster.

→ Associate:

under the Associate tab you would find Apriori filtered Associator and FP Growth.

→ Select Attributes Tab:

Select Attributes allows you feature selections based on several algorithms such as classifier. subset eval, principal component.

→ Visualize Tab:

The Visualize option allows you to visualize your processed data for analysis.

Simple CLI :

It provides a simple Command-line interface and allows direct execution of Weka Commands.

Experimenter :

It is an environment for performing experiments and Conducting statistical tests between learning schemes.

Knowledge Flow :

It is a Java-Beans based interface for setting up and running machine Learning experiments

Trees J48 classifier :

It is an Algorithm to generate a decision tree that is generated by C4.5. It is also known as statistical classifier. For decision tree classification, we need a database.

Weather nominal :

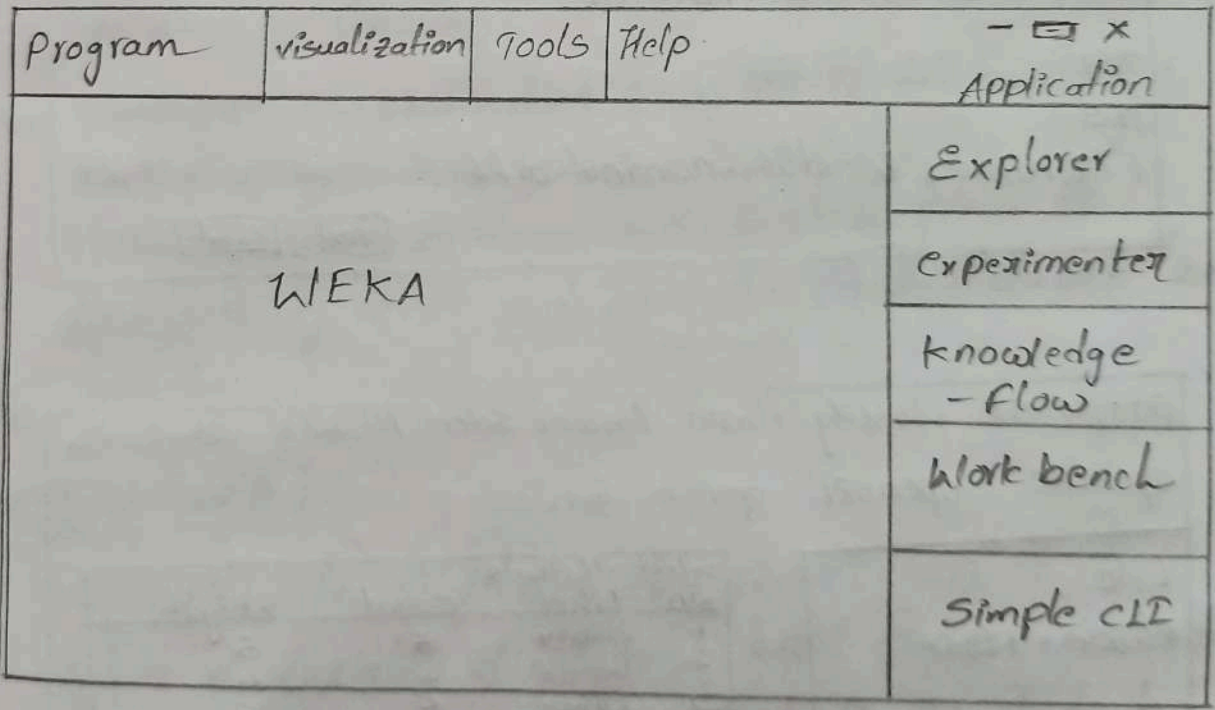
In Weka, attributes Can be nominal or numeric. The value of a nomial attribute is represented by a word: Sunny, overcast and rainy for the outlook attribute: yes and no for the play attribute.

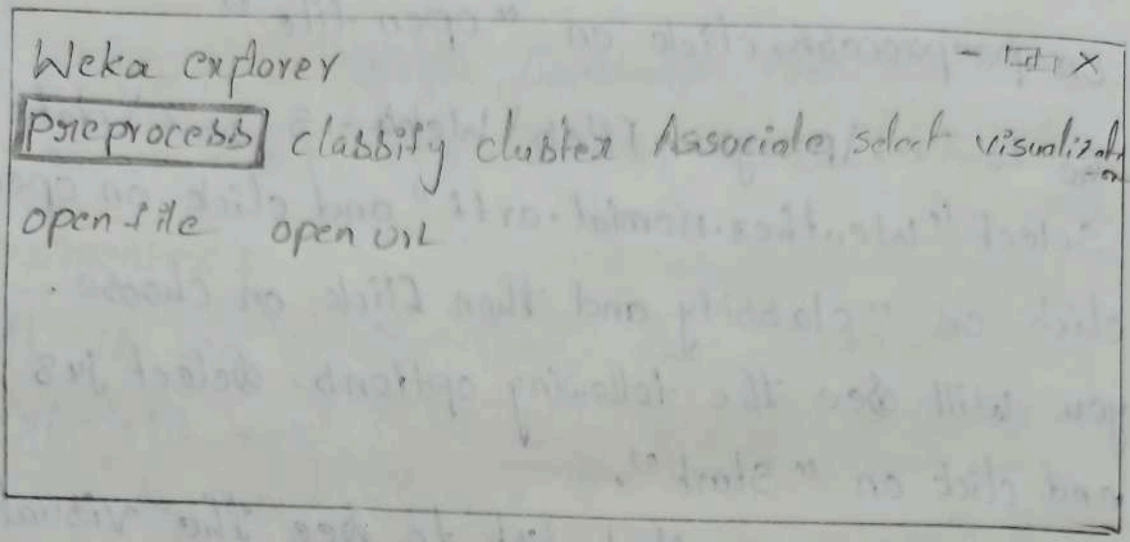
Steps Required :

1. open WEKA you can see 5 tabs on the right side of the application. They are: explorer, experimentor, knowledge Flow, Work Bench, simple CLI

2. click on "Explorer"
3. on preprocess. click on "open file"
4. Go to "c:\program files\weka-3-8-6\data", Select "weather.nomial.arff" and click on open.
5. click on "classify and then click on choose.
6. you will see the following options. select j48 and click on "start".
7. click on the resulted list to see the visual
8. click on the resulted list and click on visualize tree option.

Outcome of the experiment:

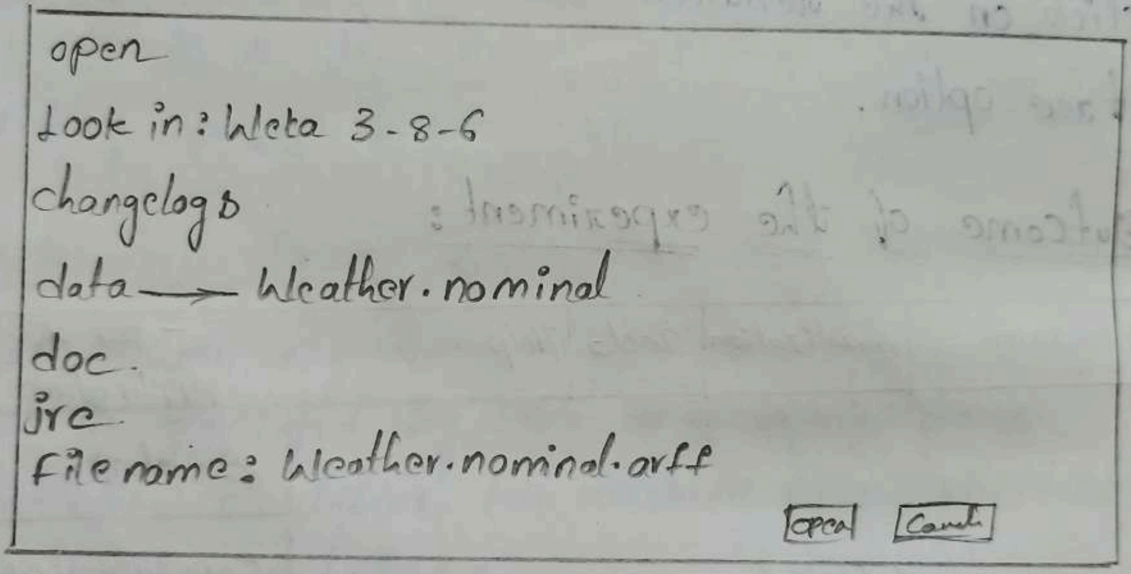




Weka explorer

Preprocess classify cluster Associate select visualization

open file open url



open

Look in: Weka 3-8-6

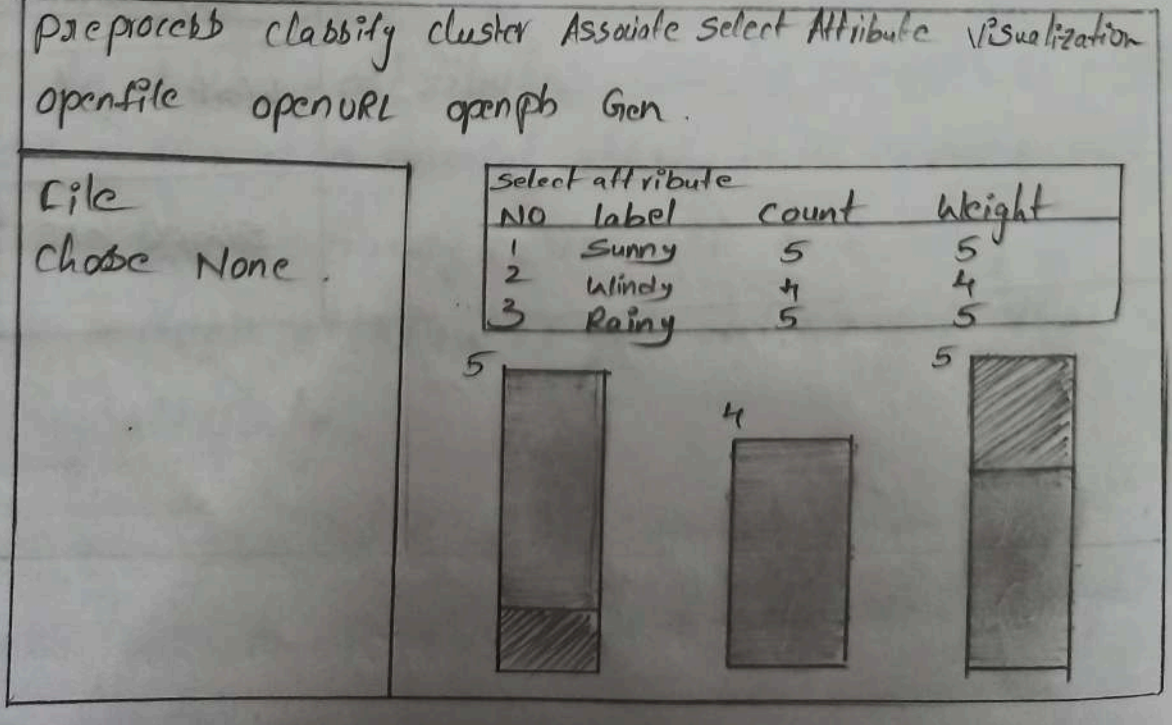
changelogs

data → Weather.nominal

doc.

jre

File name: Weather.nominal.arff



Preprocess classify cluster Assoiate select Attribute Visualization

openfile openURL openDB Gen.

File

Chose None.

select attribute

NO	label	count	Weight
1	Sunny	5	5
2	Windy	4	4
3	Rainy	5	5

Experiment - 4

Aim: Demonstrate performing classification on data Sets

→ Load each dataset into Weka and run Id3, J48 classification algorithm. Study the classifier output. Compute entropy values, kappa statistic.

→ Extract if-then rules from the decision tree generated by the classifier, observe the Confusion matrix.

→ Load each dataset into Weka and perform Naive-bayes and K-NN classifiers for each dataset, and classification and K-Nearest Neighbour classification. Interpret the results obtained.

→ Plot RoC Curves.

→ Compare classification results of ID3, J48, Naive-Bayes and k-NN classifiers for each dataset and deduce which classifier is performing best and poor for each dataset and justify.

Objectives:

The ultimate Objective of classification is to relate a variable of interest with Observed variables. The actual variable of interest is meant to be of "Qualitative" type. The algorithm required for performing the classification is known as the classifier.

Zero R :-

→ Zero R is the simplest classification method which relies on the largest and ignores all predictors

→ Zero R classifier simply predicts the majority Category.

→ Although there is no predictability power in ZeroR it is useful for determining a baseline performance as a benchmark for other classification methods.

one R :

→ This method is used in the sequential learning Algorithm for leaving the rules.

→ It returns a single rule that covers at least some examples.

→ However, what makes it really powerful is its ability to create relations among the attributes given. Hence covering a larger hypothesis space.

Explorer :

It is an environment for exploring data.

Simple CLI :

It provides a simple Command-line Interface and allows direct execution of Weka Commands.

Experimenter:

It is an environment for performing experiment and Conducting statistical tests between learning schemes.

knowledge flow:

It is a Java-Beans based interface to setting up and running machine learning experiments.

Preprocess:

It is the first step in machine learning to preprocess the data. It is used to select the data file preprocessing and make it fit for applying the various machine learning Algorithms.

Classify:

The classify tab provides you several machine learning algorithms for the classification of your data. such as linear-regression, Logistic Regression.

Test options:

Before you run the classification algorithm, you need to set test options. set test options in the Test options box. The test options that available Now are:-

1) use training set: evaluates the classifier on how well it predicts the class of the instances it was trained on.

2) Supplied test set: evaluates the classifier on how well it predicts the class of a set of instances loaded from a file. clicking on the "set..." button brings up a dialog allowing you to choose the file to test on.

3) Cross Validation:
evaluates the classifier by cross-validation, using the number of flolds that are entered in the 'Folds' text field.

4) percentage split:
evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing the amount of data held out depends on the value entered in the '%' field.

Steps Required:

1. open weka you can see 5 tabs on the right side of the application. These are explorer, experimentor, knowledge flow, work bench, Simple CLI.

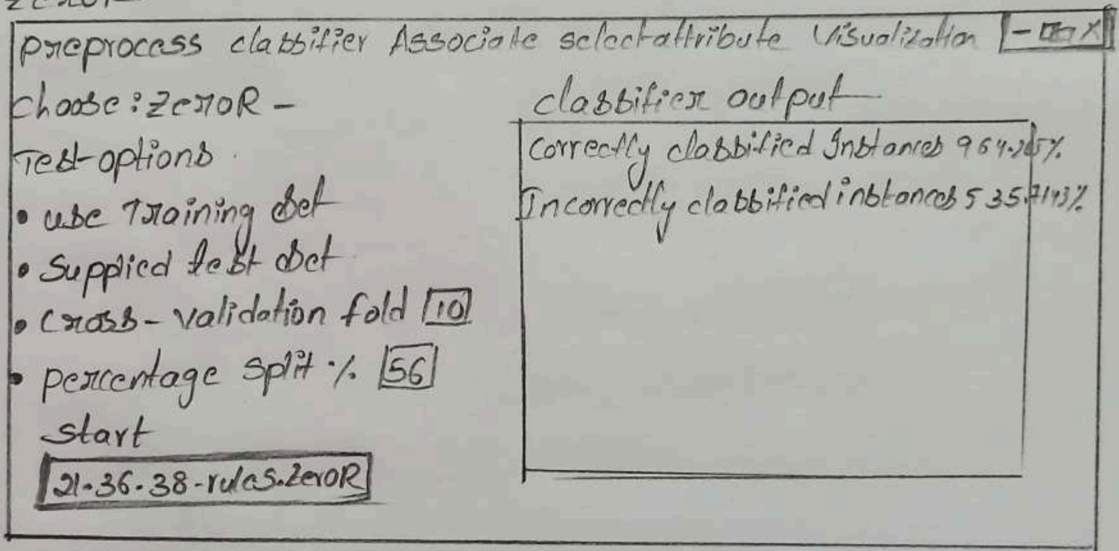
2. click on 'explorer'.

3. you can see classify tab click on the classify button.

4. you Can observe choose test options etc.
5. In test option you can see Cross - Validation folds. set it as 10.
6. Right click on choose option, Then Select the ZeroR algorithm or one R algorithm.
7. Click Start button.
8. ZeroR algorithm or one R algorithm will execute and it gives the Output.

Output:

ZeroR



preprocess classifier Associate selectattribute Visualization

choose: ZeroR -

Test options.

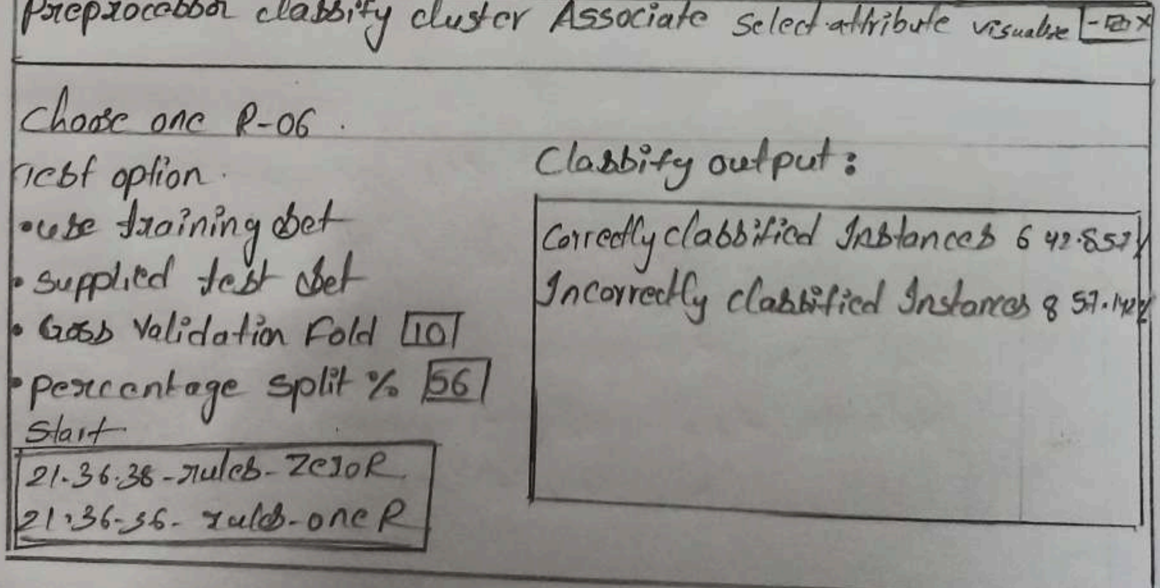
- use Training set
- Supplied test set
- cross - validation fold 10
- percentage split % 56

start

21-36-38-rules.ZeroR

classifier output

Correctly classified Instances 9 64.285%
Incorrectly classified instances 5 35.7143%



Preprocessor classify cluster Associate select attribute visualise

choose one R-06.

Test option.

- use training set
- supplied test set
- Cross Validation Fold 10
- percentage split % 56

Start

21.36.38 - rules - ZeroR
21.36-36 - rules - oneR

Classify output:

Correctly classified Instances 6 42.857%
Incorrectly classified Instances 8 57.142%