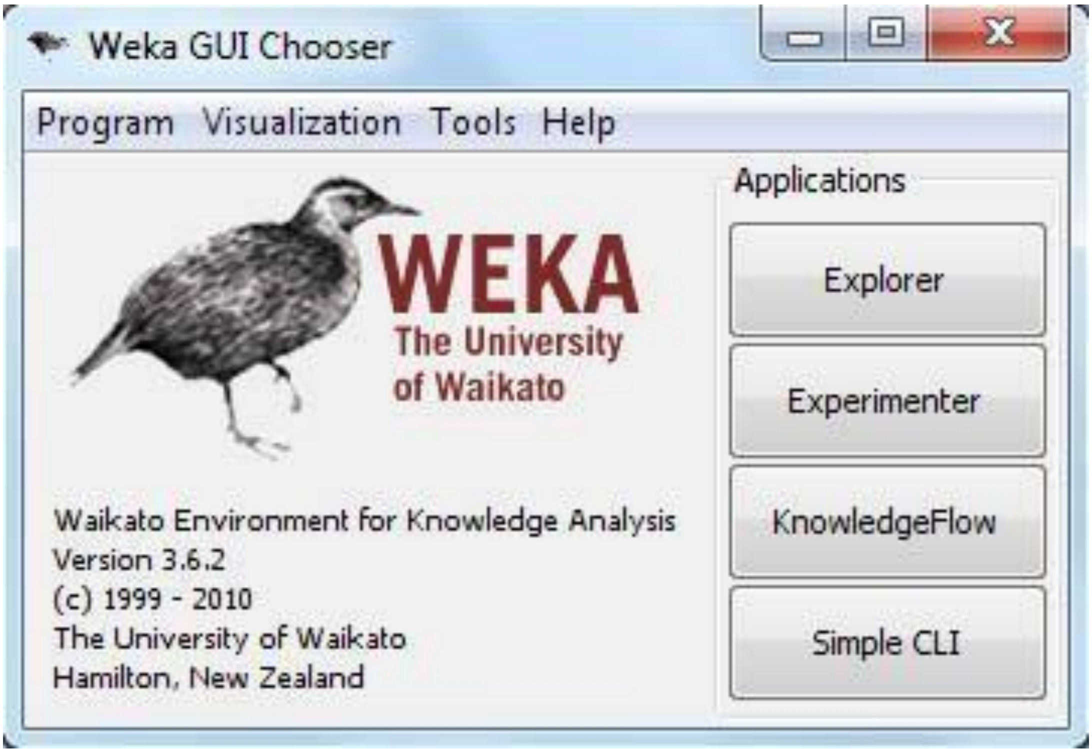
Experiment No -5

Aim: Using open-source tools (Weka) to implement Association Mining Algorithm

Theroy:

The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.



Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

The buttons can be used to start the following applications:

Explorer: An environment for exploring data with WEKA

Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.

Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

SimpleCLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

Explorer:

At the very top of the window, just below the title bar, is a row of tabs. The tabs are as follows:

Preprocess. Choose and modify the data being acted on.

Classify. Train and test learning schemes that classify or perform regression.

Cluster. Learn clusters for the data.

Associate. Learn association rules for the data.

Select attributes. Select the most relevant attributes in the data.

Visualize. View an interactive 2D plot of the data.

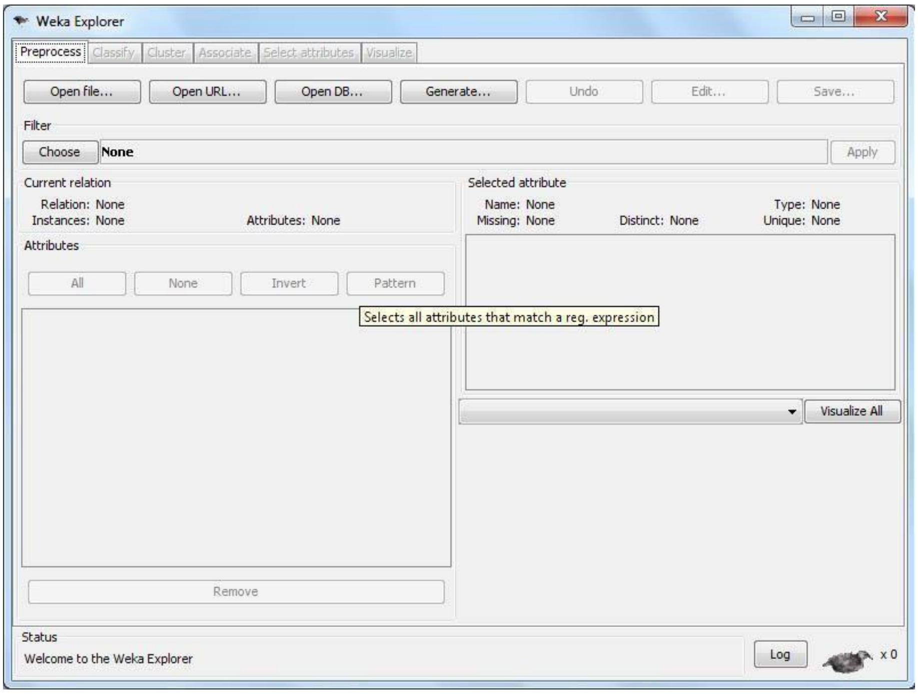
Loading Data:

The first four buttons at the top of the preprocess section enable you to load data into WEKA:

Open file. .... Brings up a dialog box allowing you to browse for the data file on the local file system.

Open URL. ... Asks for a Uniform Resource Locator address for where the data is stored.

Open DB..... Reads data from a database. (Note that to make this work you might have to edit the file in weka/experiment/DatabaseUtils.props.)



Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Selects all attributes that match a reg. expression

Welcome to the Weka Explorer

The Current Relation:

Once some data has been loaded, the Preprocess panel shows a variety of information. The Current relation box (the “current relation” is the currently loaded data, which can be interpreted as a single relational table in database terminology) has three entries:

Relation. The name of the relation, as given in the file it was loaded from. Filters (described below) modify the name of a relation.

Instances. The number of instances (data points/records) in the data.

Attributes. The number of attributes (features) in the data.
When you click on different rows in the list of attributes, the fields change in the box to the right titled Selected attribute. This box displays the characteristics of the currently highlighted attribute in the list:

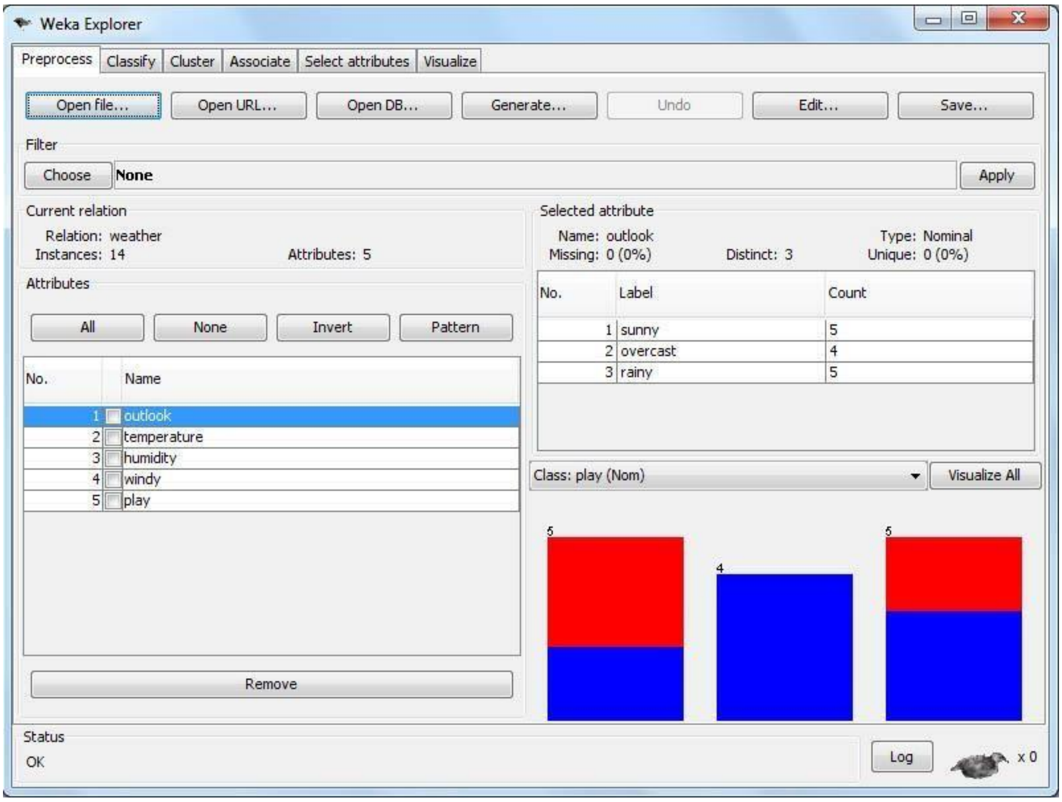
Name. The name of the attribute, the same as that given in the attribute list.

Type. The type of attribute, most commonly Nominal or Numeric.

Missing. The number (and percentage) of instances in the data for which this attribute is missing (unspecified).

Distinct. The number of different values that the data contains for this attribute.

Unique. The number (and percentage) of instances in the data having a value for this attribute that no other instances have.



Filter
None

Current relation
Relation: weather
Instances: 14 Attributes: 5

Attributes

No.	Name
1	□ outlook
2	□ temperature
3	□ humidity
4	□ windy
5	□ play

Selected attribute
Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom)

Status
OK

Below these statistics is a list showing more information about the values stored in this attribute, which differ depending on its type. If the attribute is nominal, the list consists of each possible value for the attribute along with the number of instances that have that value. If the attribute is numeric, the list gives four statistics describing the distribution of values in the data—the minimum, maximum, mean and standard deviation. And below these statistics there is a coloured histogram, color-coded according to the attribute chosen as the Class using the box above the histogram. (This box will bring up a drop-down list of available selections when clicked.) Note that only nominal Class attributes will result in a color-coding. Finally, after pressing the Visualize All button, histograms for all the attributes in the data are shown in a separate window. Returning to the attribute list, to begin with all the tick boxes are unticked. They can be toggled on/off by clicking on them individually. The four buttons above can also be used to change the selection:

All. All boxes are ticked.

None. All boxes are cleared (unticked).

Invert. Boxes that are ticked become unticked and vice versa.

Pattern. Enables the user to select attributes based on a Perl 5 Regular Expression. E.g., .* id selects all attributes which name ends with id.

Association Rule:

Name: - WEKA Associations Apriori

Options

- **car** -- If enabled class association rules are mined instead of (general) association rules.

- **classIndex** -- Index of the class attribute. If set to -1, the last attribute is taken as class attribute.

- **delta** -- Iteratively decrease support by this factor. Reduces support until min support is reached or required number of rules has been generated.

- **lowerBoundMinSupport** -- Lower bound for minimum support.

- **metricType** -- Set the type of metric by which to rank rules. Confidence is the proportion of the examples covered by the premise that are also covered by the consequence (Class association rules can only be mined using confidence). Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support. Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other. The total number of examples that this represents is presented in brackets following the leverage. Conviction is another measure of departure from independence.

- **minMetric** -- Minimum metric score. Consider only rules with scores higher than this
- value.

- **numRules** -- Number of rules to find.

- **removeAllMissingCols** -- Remove columns with all missing values.

- **significanceLevel** -- Significance level. Significance test (confidence metric only).

- **upperBoundMinSupport** -- Upper bound for minimum support. Start iteratively decreasing minimum support from this value.

No. of Rules: 10

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A false -c -1
Relation: contact-lenses
Instances: 40
Attributes: 5
age
spectacle-prescrip
astigmatism
tear-prod-rate
contact-lenses

=== Associator model (full training set) ===

=== Apriori ===

Minimum support: 0.15
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 40

Size of set of large itemsets L(3): 18

Size of set of large itemsets L(4): 3

Best rules found:

1. contact-lenses=soft 12 ==> tear-prod-rate=normal 12 conf:(1)
2. astigmatism=no contact-lenses=soft 11 ==> tear-prod-rate=normal 11 conf:(1)
3. contact-lenses=hard 8 ==> tear-prod-rate=normal 8 conf:(1)
4. spectacle-prescrip=hypermetrope tear-prod-rate=reduced 8 ==> contact-lenses=none 8
5. Conf :(1)
6. age = presbyopic tear-prod-rate=reduced 7 ==> contact-lenses=none 7 conf:(1)
7. spectacle-prescrip=hypermetrope contact-lenses=soft 7 ==> tear-prod-rate=normal 7
8. Conf : (1)
9. 7. astigmatism = tear-prod-rate=reduced 7 ==> contact-lenses=none 7 conf:(1)
10. astigmatism = yes contact-lenses=hard 7 ==> tear-prod-rate=normal 7 conf:(1)
11. age = young contact-lenses=soft 6 ==> astigmatism=no 6 conf:(1)
12. age = contact-lenses=soft 6 ==> tear-prod-rate=normal 6 conf:(1)

No. Of Rules: 3

```
=== Run information ===

Scheme: weka.associations.Apriori -N 3 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A
false -c -1
Relation: contact-lenses
Instances: 40
Attributes: 5
age
spectacle-prescrip
astigmatism
tear-prod-rate
contact-lenses


=== Associator model (full training set) ===

=== Apriori ===

Minimum support: 0.25
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 1

Best rules found:

 1. contact-lenses=soft 12 ==> tear-prod-rate=normal 12 conf:(1)
 2. astigmatism=no contact-lenses=soft 11 ==> tear-prod-rate=normal 11
 3. tear-prod-rate=reduced 17 ==> contact-lenses=none 16 conf:(0.94)
    conf:(1)
```

Output:

```
=== Run information ===

Scheme:       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:     supermarket
Instances:     4627
Attributes:    217
              [list of attributes omitted]
=== Associator model (full training set) ===


====Apriori=======

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1
```

Best rules found:

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
4. biscuits=t fruit=and vegetables=t total=high 815 ==> bread and cake=t 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
9. frozen foods=t fruit=and vegetables=t total=high 834 ==> bread and cake=t 757 <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

Conclusion:

Thus, we learned and successfully implemented the Apriori algorithm in weka open source in the iris dataset. We can easily visualize the association in above dataset