

Classification rules and descriptive induction

12/11/2008

Voting dataset
Iris dataset

Voting dataset

- 435 instances
- 16 attributes
 - 16 nominal attributes
 - 0 numeric attributes
- No target variable
- No missing values

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open fil... | Open U... | Open D... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation
Relation: voting
Instances: 435 Attributes: 16

Selected attribute
Name: handicapped-infants Type: No...
Missi... 12 (... Distinct: ... Unique: 0 (0...)

Label	Count
n	236
y	187

Attributes: All | None | Invert

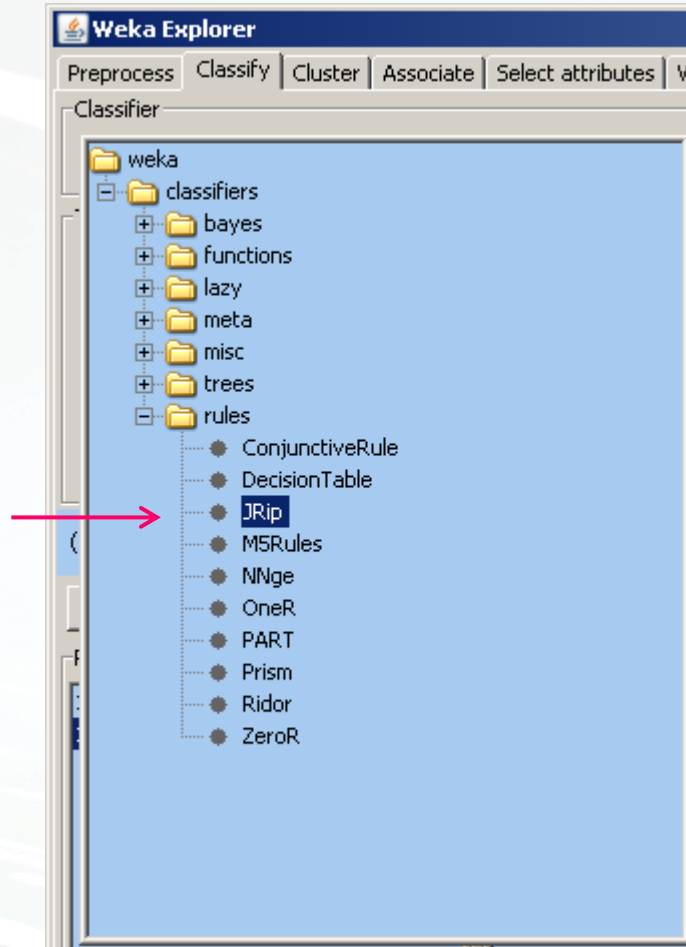
No.	Name
<input checked="" type="checkbox"/>	1 handicapped-infants
<input type="checkbox"/>	2 water-project-cost-sharing
<input type="checkbox"/>	3 adoption-of-the-budget-resolution
<input type="checkbox"/>	4 physician-fee-freeze
<input type="checkbox"/>	5 el-salvador-aid
<input type="checkbox"/>	6 religious-groups-in-schools
<input type="checkbox"/>	7 anti-satellite-test-ban
<input type="checkbox"/>	8 aid-to-nicaraguan-contras
<input type="checkbox"/>	9 mx-missile
<input type="checkbox"/>	10 immigration
<input type="checkbox"/>	11 synfuels-corporation-cutback
<input type="checkbox"/>	12 education-spending
<input type="checkbox"/>	13 superfund-right-to-sue
<input type="checkbox"/>	14 crime
<input type="checkbox"/>	15 duty-free-exports
<input type="checkbox"/>	16 party

Remove

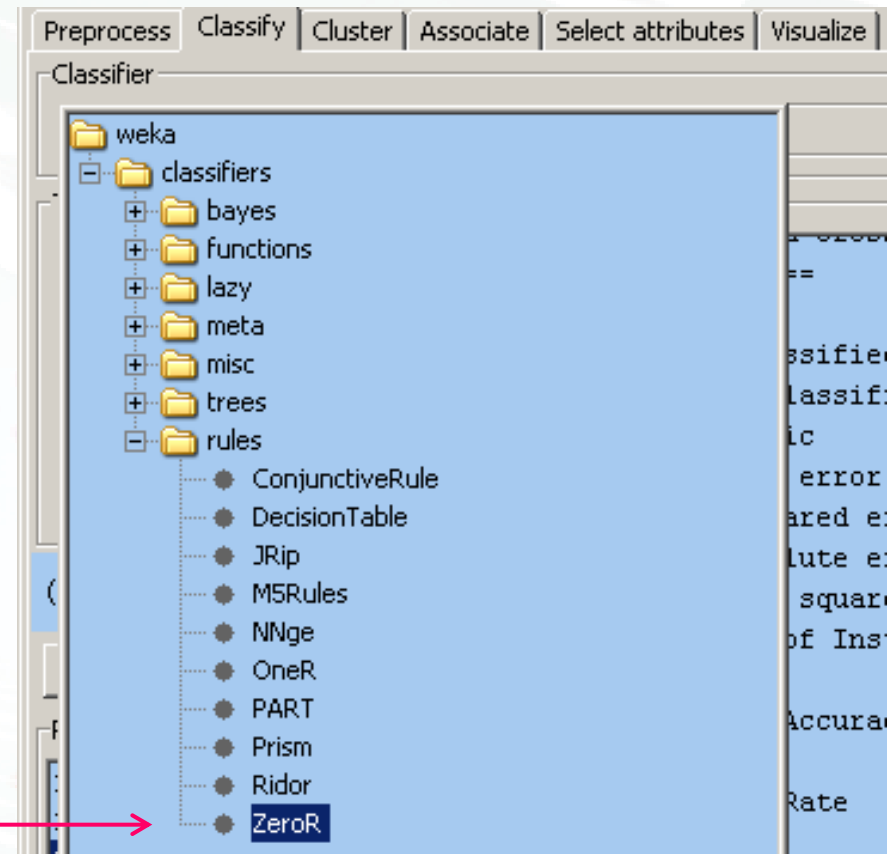
Class: party (Nom) Visualize All

Status: OK Log x 0

Classification rules: Weka → classifiers → rules → JRip



Baseline classifier: Weka → classifiers → rules → ZeroR



Association rules

Weka → associations → Apriori

1

2

The screenshot shows the Weka Explorer interface with the 'Associate' tab selected. The 'Associator' dropdown is set to 'Apriori' with parameters: -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0. The 'Start' button is highlighted. The 'Associator output' window displays the following text:

```

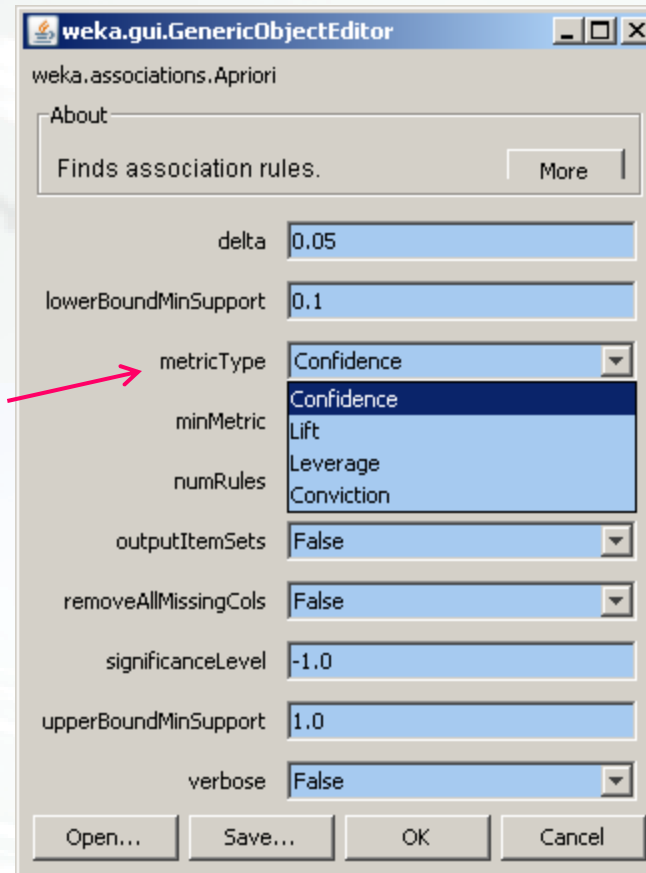
Size of set of large itemsets L(4): 1

Best rules found:

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> party=democrat 219   conf: (1)
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> party=democrat 198   conf: (1)
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> party=democrat 210   conf: (1)
4. physician-fee-freeze=n education-spending=n 202 ==> party=democrat 201   conf: (1)
5. physician-fee-freeze=n 247 ==> party=democrat 245   conf: (0.99)
6. el-salvador-aid=n party=democrat 200 ==> aid-to-nicaraguan-contras=y 197   conf: (0.99)
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204   conf: (0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y party=democrat 203 ==> physician-fee-freeze=n 203   conf: (0.98)
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> party=democrat 197   conf: (0.97)
10. aid-to-nicaraguan-contras=y party=democrat 218 ==> physician-fee-freeze=n 210   conf: (0.96)
    
```

The 'Result list' on the left shows '16:25:34 - Apriori' as the selected result.

Quality of association rules



Compare classification and association rules

- Purpose
- Format
- Quality measure
- Ruleset / List of rules
- Exhaustiveness of algorithms

Compare classification and association rules

- Purpose
 - Classification rules: classification
 - Association rules: exploratory data analysis, descriptive induction
- Format
 - Both in the format $X \rightarrow Y$
 - Classification rules: Y is a pair “target variable=class”
 - Association rules: both X and Y are itemsets \cong conjunctions of attribute-value pairs
- Quality measure
 - Classification rules: classification accuracy of the ruleset, precision, rule accuracy, weighted relative accuracy
 - Association rules: support, confidence, lift, leverage, conviction
- Ruleset / List of rules
 - Classification rules: can be both: unordered sets of rules or ordered list of rules
 - Association rules: unordered set of rules
- Exhaustiveness of algorithms
 - Classification rules: heuristic algorithms
 - Association rules: exhaustive algorithms, guarantee the optimal results

Iris dataset

- 150 instances
- 4 attributes
 - 0 nominal attributes
 - 4 numeric attributes
- Nominal target variable
 - 3 values:
 - Iris-setosa (30%)
 - Iris-versicolor (30%)
 - Iris-virginica (30%)
- No missing values

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Ope... | Ope... | Ope... | Undo | Edit... | Sav...

Filter: Choose **None** Apply

Current relation: Relation: iris, Instances: 150, Attributes: 5

Selected attribute: Name: iris, Type: N..., Missing: ..., Distinct: Unique: 0...

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Class: iris (Nom) Visualize All

50 50 50

Status: OK Log x 0

Clustering

Weka → clusters → SimpleKMeans

1

2

3

The screenshot displays the Weka Explorer interface. The 'Clusterer' list on the left contains several algorithms, with 'SimpleKMeans' selected. The 'Clusterer output' window shows the following text:

```

Clusterer output
Number of iterations: 6
Within cluster sum of squared errors: 6.99811400482676

Cluster centroids:

```

The 'weka.gui.GenericObjectEditor' dialog is open, showing the configuration for 'weka.clusterers.SimpleKMeans'. The 'About' section contains the text: 'Cluster data using the k means algorithm'. The 'numClusters' field is set to 3, and the 'seed' field is set to 10. The dialog also includes 'Open...', 'Save...', 'OK', and 'Cancel' buttons.

The status bar at the bottom of the Weka Explorer window shows 'OK'.

Clustering visualization

The screenshot displays the Weka Clusterer Visualize window for a SimpleKMeans model. The main window on the left shows the 'Cluster:' tab with the 'Store clusters for visualization' checkbox checked. A context menu is open over the 'Result list' table, with 'Visualize cluster assignments' selected. The main visualization window on the right shows a scatter plot of 'sepal length (Num)' vs 'petal width (Num)'. The plot contains three clusters of data points, color-coded as red, blue, and green. A 'Jitter' slider is visible above the plot. To the right of the plot is a legend for 'Class colour' with three entries: 'cluster0' (red), 'cluster1' (blue), and 'cluster2' (green). Below the legend is a small table showing the distribution of points for each cluster across the X and Y axes.

Cluster	X: sepal length (Num)	Y: petal width (Num)
cluster0	4.3 - 6.1	0.1 - 1.3
cluster1	4.3 - 6.1	1.3 - 2.5
cluster2	6.1 - 7.9	1.3 - 2.5