

Data Mining file

Jatin Singh Kushwaha
23/SCA/BSC.IT/004

1. Demonstration of preprocessing on dataset employee.arff.

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active. The dataset 'employee.arff' is loaded, showing 11 instances and 3 attributes. The 'performance' attribute is selected as the class.

Current relation:
Relation: employee
Instances: 11
Attributes: 3
Sum of weights: 11

Attributes:
1 ☒ age
2 ☒ salary
3 ☒ performance

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

No.	Label	Count	Weight
1	good	3	3
2	avg	4	4
3	poor	4	4

Class: performance (Nom)

Visualize All

The bar chart shows the distribution of the 'performance' class. The 'good' class (blue) has a count of 3. The 'avg' class (red) has a count of 4. The 'poor' class (cyan) has a count of 4.

Status: OK

2. Demonstration of preprocessing on dataset employee.arff.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Associator

Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop

Result list (right-click for ...)

09:10:20 - Apriori

Associator 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: employee
Instances: 11
Attributes: 3
  age
  salary
  performance

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

Apriori
=====

Minimum support: 0.1 (1 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17
Size of set of large itemsets L(2): 25
Size of set of large itemsets L(3): 11

Best rules found:

1. age=27 2 ==> performance=poor 2 <conf:(1)> lift:(2.75) lev:(0.12) [1] conv:(1.27)
2. age=29 2 ==> performance=avg 2 <conf:(1)> lift:(2.75) lev:(0.12) [1] conv:(1.27)
3. age=30 2 ==> performance=avg 2 <conf:(1)> lift:(2.75) lev:(0.12) [1] conv:(1.27)
```

Status OK

Log x 0

```
4. age=48 2 ==> performance=good 2 <conf:(1)> lift:(3.67) lev:(0.13) [1] conv:(1.45)
5. salary=17k 2 ==> performance=poor 2 <conf:(1)> lift:(2.75) lev:(0.12) [1] conv:(1.27)
6. salary=20k 2 ==> performance=avg 2 <conf:(1)> lift:(2.75) lev:(0.12) [1] conv:(1.27)
7. salary=25k 2 ==> performance=avg 2 <conf:(1)> lift:(2.75) lev:(0.12) [1] conv:(1.27)
8. salary=32k 2 ==> performance=good 2 <conf:(1)> lift:(3.67) lev:(0.13) [1] conv:(1.45)
9. salary=10k 1 ==> age=25 1 <conf:(1)> lift:(11) lev:(0.08) [0] conv:(0.91)
10. age=25 1 ==> salary=10k 1 <conf:(1)> lift:(11) lev:(0.08) [0] conv:(0.91)
```

Status OK

Log x 0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Plot Matrix

age salary performance

performance

salary

age

PlotSize: [100]

PointSize: [1]

Jitter:

Colour: performance (Nom)

Class Colour

good avg poor

Status OK

Log x 0

3. Demonstration of classification rule process on dataset employee.arff using naïve Bayes algorithm.

The screenshot shows the Weka Explorer interface with the Naive Bayes classifier selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section displays the following information:

```
=== Run information ===  
Scheme: weka.classifiers.bayes.NaiveBayes  
Relation: employee  
Instances: 11  
Attributes: 3  
age  
salary  
performance  
Test mode: 10-fold cross-validation  
=== Classifier model (full training set) ===  
Naive Bayes Classifier  
Attribute Class  
good avg poor  
(0.29) (0.36) (0.36)  
-----  
age  
25 1.0 1.0 2.0  
27 1.0 1.0 3.0  
28 1.0 1.0 2.0  
29 1.0 3.0 1.0  
30 1.0 3.0 1.0  
35 2.0 1.0 1.0  
48 3.0 1.0 1.0  
[total] 10.0 11.0 11.0  
salary  
10k 1.0 1.0 2.0  
15k 1.0 1.0 2.0
```

The status bar at the bottom shows 'Status OK' and a 'Log' button.

The screenshot shows the Weka Explorer interface with the Naive Bayes classifier selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section displays the following information:

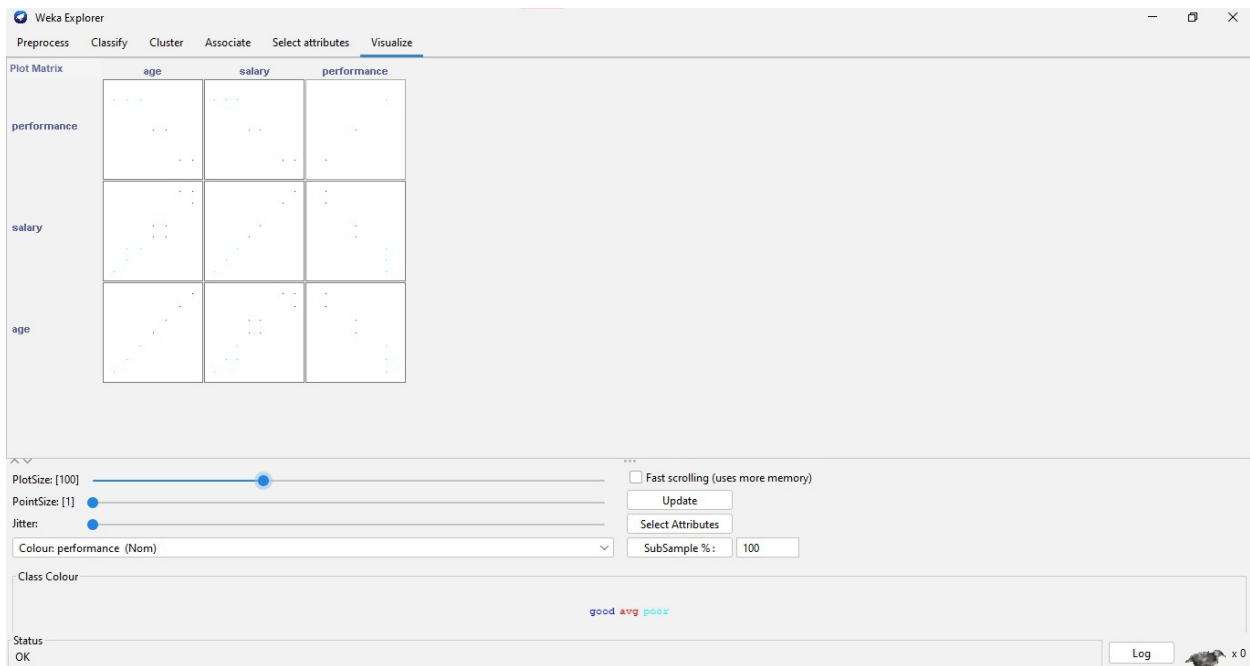
```
17k 1.0 1.0 3.0  
20k 1.0 3.0 1.0  
25k 1.0 3.0 1.0  
30k 1.0 1.0 1.0  
35k 2.0 1.0 1.0  
32k 3.0 1.0 1.0  
[total] 11.0 12.0 12.0  
Time taken to build model: 0 seconds  
=== Stratified cross-validation ===  
=== Summary ===  
Correctly Classified Instances 10 90.9091 %  
Incorrectly Classified Instances 1 9.0909 %  
Kappa statistic 0.8625  
Mean absolute error 0.2899  
Root mean squared error 0.3171  
Relative absolute error 61.3111 %  
Root relative squared error 63.0158 %  
Total Number of Instances 11  
=== Detailed Accuracy By Class ===  
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class  
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 good  
1.000 0.143 0.800 1.000 0.889 0.828 1.000 1.000 avg  
0.750 0.000 1.000 0.750 0.857 0.810 1.000 1.000 poor  
Weighted Avg. 0.909 0.052 0.927 0.909 0.908 0.868 1.000 1.000  
=== Confusion Matrix ===
```

The status bar at the bottom shows 'Status OK' and a 'Log' button.

The screenshot shows the Weka Explorer interface with the Naive Bayes classifier selected. The 'Classifier output' section displays the following information:

```
=== Confusion Matrix ===  
a b c <-- classified as  
3 0 0 | a = good  
0 4 0 | b = avg  
0 1 3 | c = poor
```

The status bar at the bottom shows 'Status OK' and a 'Log' button.



4. Demonstration of clustering rule process on dataset Employee.arff using simple k- means.

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'SimpleKMeans' algorithm is chosen with the following command: `-init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10`. The 'Cluster mode' section has 'Use training set' selected, 'Ignore attributes' is empty, and 'Store clusters for visualization' is checked. The 'Cluster output' pane displays the following information:

```
=== Run information ===
Scheme:      weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:    employee
Instances:   11
Attributes:  3
              age
              salary
              performance
Test mode:   evaluate on training data

=== Clustering model (full training set) ===

kMeans
=====

Number of iterations: 2
Within cluster sum of squared errors: 17.0

Initial starting points (random):

Cluster 0: 30,25k,avg
Cluster 1: 25,10k,poor

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute      Full Data      Cluster#
              (11.0)      (7.0)      (4.0)
age             27             29             27
salary          17k            20k            17k
performance    avg             avg             poor

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      7 ( 64%)
1      4 ( 36%)
```

The screenshot shows the 'Visualize' tab in Weka Explorer. A scatter plot matrix is displayed for the attributes 'age', 'salary', and 'performance'. The matrix shows the relationships between these variables for the 11 instances in the dataset. The 'age' vs 'salary' plot shows a positive correlation, while 'age' vs 'performance' and 'salary' vs 'performance' show more scattered data points. The 'Class Colour' is set to 'performance (Nom)' with a legend showing 'good' (blue), 'avg' (green), and 'poor' (red).

This screenshot is identical to the previous one, showing the 'Visualize' tab in Weka Explorer with the scatter plot matrix for 'age', 'salary', and 'performance'. The 'Class Colour' is set to 'performance (Nom)' with a legend showing 'good' (blue), 'avg' (green), and 'poor' (red).