

## CSC4008 Assignment3

118010045 Cui Yuncong

### 1. Load weather.nominal.arff, and answer the following questions. (10')

#### 1) How many instances, attributes in this dataset?

Instances: 14

Attributes: 5

Current relation	
Relation: weather.symbolic	Attributes: 5
Instances: 14	Sum of weights: 14

#### 2) The number of the distinct label of the attribute: temperature.

Number: 3

Selected attribute			
Name: outlook		Type: Nominal	
Missing: 0 (0%)		Unique: 0 (0%)	
		Distinct: 3	
No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

#### 3) Use the filter to remove the value: high in the humidity attribute from the dataset.

Choose the RemoveWithValues filter.

Fill in the attributeIndex with 3, fill in nominalIndex with 1.

Apply.

Filter

Choose: RemoveWithValues -S 0.0 -C 3 -L 1 Apply Stop

Current relation

Relation: weather.symbolic  
Instances: 14

Attributes: 5  
Sum of weights: 14

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input checked="" type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

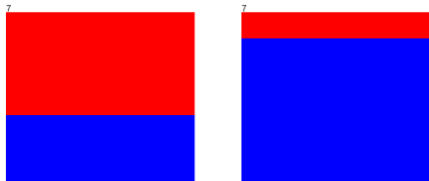
Remove

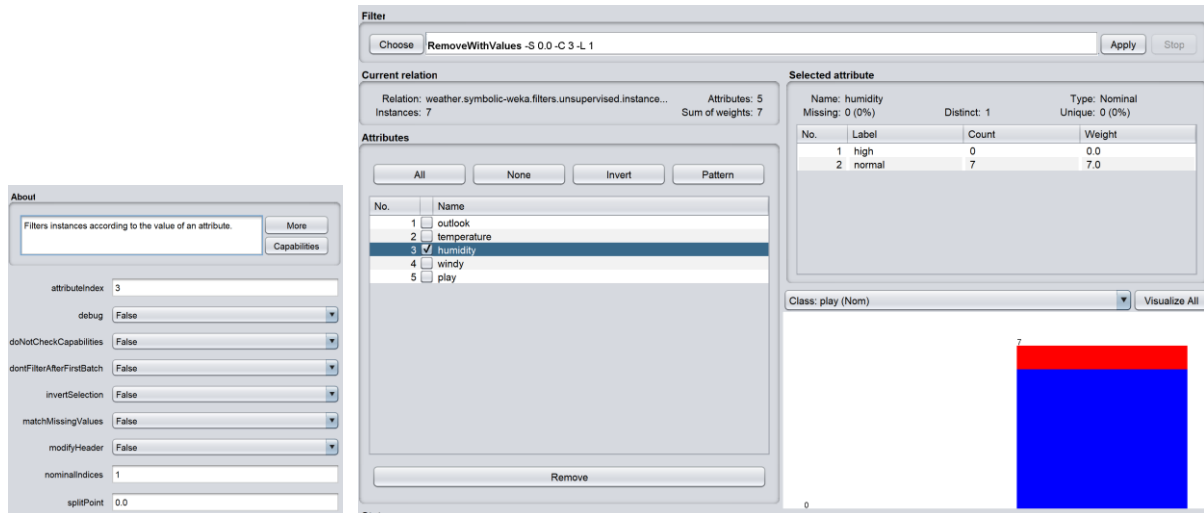
Selected attribute

Name: humidity  
Missing: 0 (0%)  
Distinct: 2  
Type: Nominal  
Unique: 0 (0%)

No.	Label	Count	Weight
1	high	7	7.0
2	normal	7	7.0

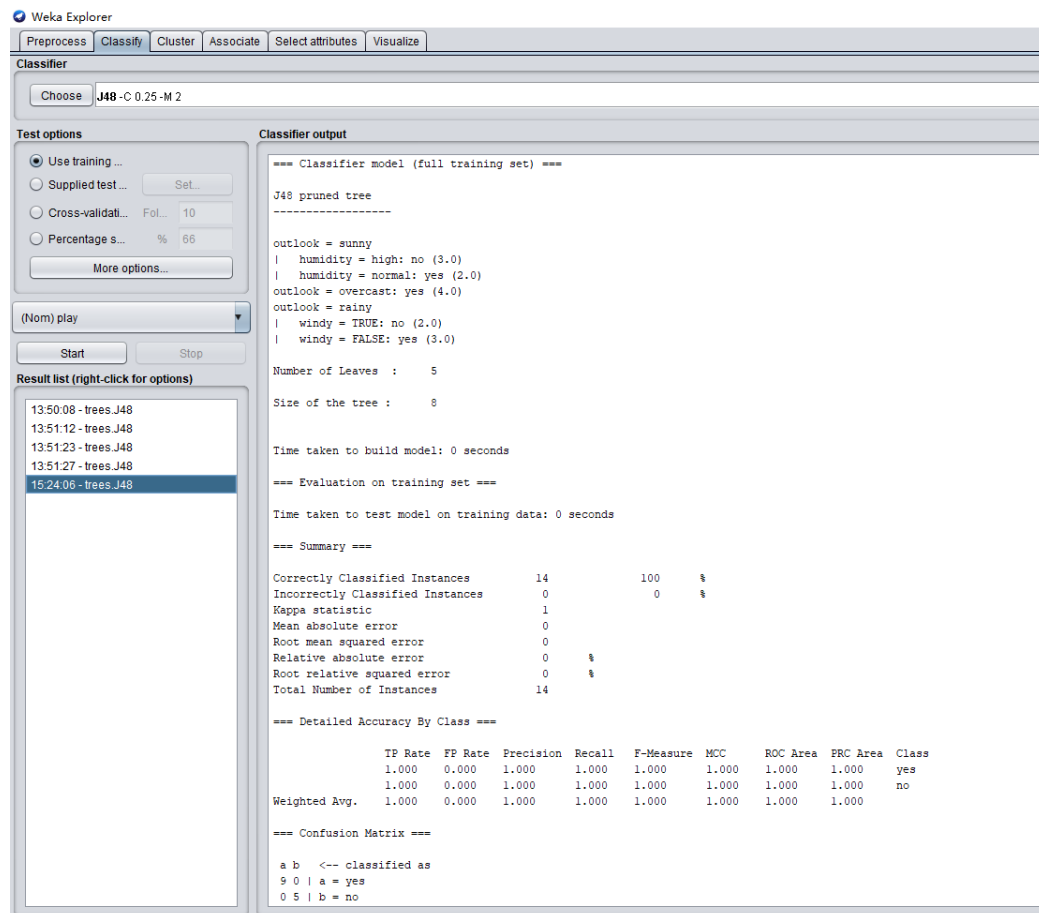
Class: play (Nom) Visualize All

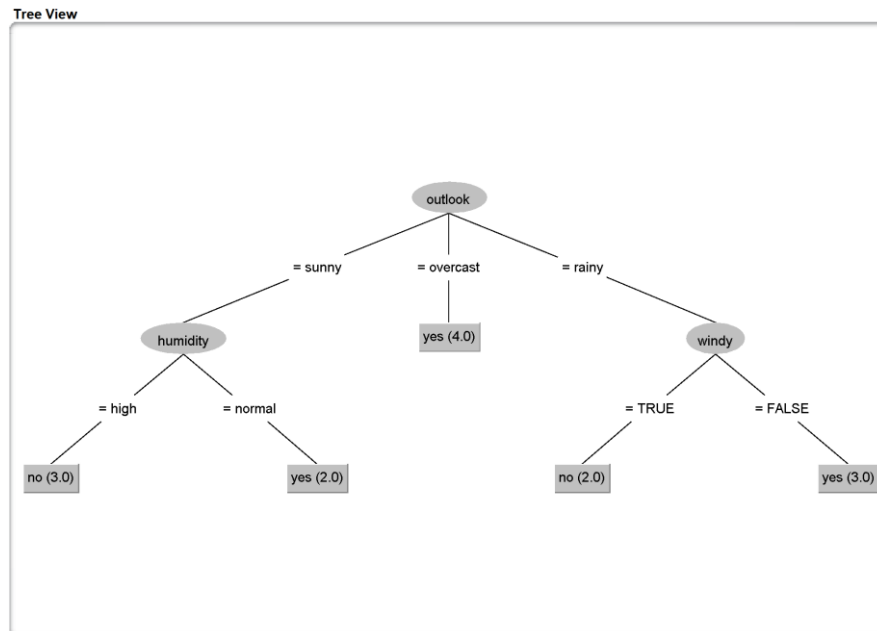




4) Please build a J48 decision tree to classify the dataset and visualize the tree in a figure.

Choose the J48 decision tree.





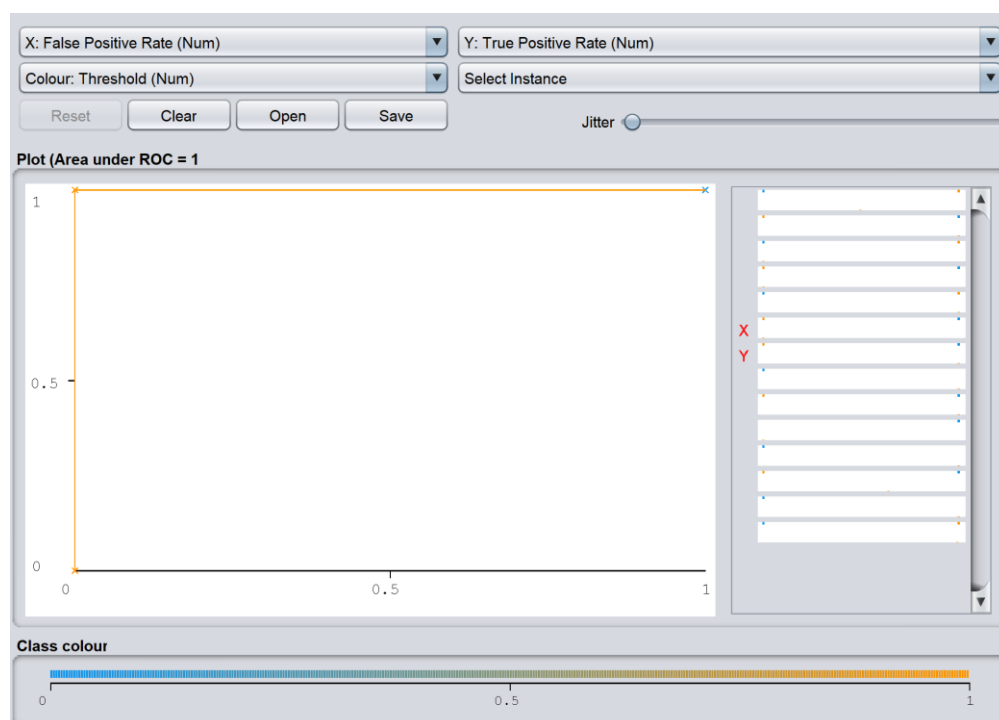
**5) Analyze the performance of J48 (including but not limited to TPR, FPR, ROC, AUC).**

$N = 14$     $TP = 9$     $FN = 0$     $FP = 0$     $TN = 5$

$TPR = TP / (TP + FN) = 1$     $FPR = FP / (FP + TN) = 0$

$AUC = 1$

The higher the TPR is, the better the classifier. The lower the FPR is, the better the classifier.  $TPR = 1$  and  $FPR = 0$  means J48 is the perfect classifier for this dataset. The AUC value is the area covered by the ROC curve. Obviously, the larger the AUC, the better the classifier.  $AUC = 1$  means J48 is the perfect classifier for this dataset.



## 2. Load HR-Employee-Attrition.csv dataset. (10')

### 1) Implement attribute selection.

Evaluator: CfsSubsetEval

Search Method: BestFirst

N = 12

These attributes are Age, BusinessTravel, EnvironmentSatisfaction, JobInvolvement, JobLevel, MonthlyIncome, OverTime, StockOptionLevel, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsWithCurrManager

The screenshot shows the WEKA Attribute Evaluator window. The 'Evaluator' is set to 'CfsSubsetEval -P 1 -E 1' and the 'Search Method' is 'BestFirst -D 1 -N 5'. The 'Attribute Selection Mode' is set to 'Use full training set' with 'Folds' set to 10 and 'Seed' set to 1. The target attribute is '(Nom) Attrition'. The 'Start' button is highlighted. The 'Result list' on the left shows a list of attribute subsets and their performance metrics. The 'Attribute selection output' on the right displays the search process details and the selected attributes.

**Attribute Evaluator**

Choose **CfsSubsetEval -P 1 -E 1**

**Search Method**

Choose **BestFirst -D 1 -N 5**

**Attribute Selection Mode**

☒ Use full training set  
☐ Cross-validation Folds   
Seed

(Nom) Attrition

Start Stop

**Result list (right-click for options)**

- 13:13:15 - Ranker + ClassifierAttribute
- 13:13:25 - Ranker + ClassifierAttribute
- 13:13:28 - Ranker + ClassifierAttribute
- 13:13:31 - Ranker + ClassifierAttribute
- 13:13:34 - Ranker + ClassifierAttribute
- 13:13:37 - Ranker + ClassifierAttribute
- 13:13:41 - Ranker + ClassifierAttribute
- 13:13:53 - GreedyStepwise + CfsSubs
- 13:13:59 - GreedyStepwise + CfsSubs
- 13:53:39 - Ranker + ClassifierAttribute
- 13:54:48 - Ranker + ClassifierAttribute
- 13:55:14 - Ranker + ClassifierAttribute
- 13:55:23 - Ranker + ClassifierAttribute
- 15:19:49 - BestFirst + CfsSubsetEval
- 15:20:11 - BestFirst + CfsSubsetEval

**Attribute selection output**

```
=== Attribute Selection on all input data ===  
  
Search Method:  
  Best first.  
  Start set: no attributes  
  Search direction: forward  
  Stale search after 5 node expansions  
  Total number of subsets evaluated: 430  
  Merit of best subset found: 0.094  
  
Attribute Subset Evaluator (supervised, Class (nominal): 2 Attrition):  
  CFS Subset Evaluator  
  Including locally predictive attributes  
  
Selected attributes: 1,3,11,14,15,19,23,28,29,31,32,35 : 12  
  Age  
  BusinessTravel  
  EnvironmentSatisfaction  
  JobInvolvement  
  JobLevel  
  MonthlyIncome  
  OverTime  
  StockOptionLevel  
  TotalWorkingYears  
  WorkLifeBalance  
  YearsAtCompany  
  YearsWithCurrManager
```

## 2) Using these N attributes to build a classifier.

The RandomForest is chosen among these 3 classifiers (RandomForest, J48, NaiveBayes). Because its TPR = 1 and FPR = 0, which means the accuracy is 100%.

Classifier: RandomForest

The screenshot shows the Weka Classifier window with the RandomForest classifier selected. The 'Test options' section on the left has 'Use training set' selected. The 'Classifier output' section on the right displays the following information:

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
```

Time taken to build model: 0.48 seconds

--- Evaluation on training set ---

Time taken to test model on training data: 0.09 seconds

--- Summary ---

Metric	Value	Percentage
Correctly Classified Instances	1470	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0798	
Root mean squared error	0.1244	
Relative absolute error	29.4683 %	
Root relative squared error	33.817 %	
Total Number of Instances	1470	

--- Detailed Accuracy By Class ---

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Yes
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	No

--- Confusion Matrix ---

a \ b	classified as
237	0   a = Yes
0	1233   b = No

Classifier: J48

The screenshot shows the Weka Classifier window with the J48 classifier selected. The 'Test options' section on the left has 'Use training set' selected. The 'Classifier output' section on the right displays the following information:

```
weka.classifiers.trees.J48 --C 0.25 -M 2
```

Time taken to build model: 0.02 seconds

--- Evaluation on training set ---

Time taken to test model on training data: 0.02 seconds

--- Summary ---

Metric	Value	Percentage
Correctly Classified Instances	1343	91.3605 %
Incorrectly Classified Instances	127	8.6395 %
Kappa statistic	0.6126	
Mean absolute error	0.152	
Root mean squared error	0.2757	
Relative absolute error	56.1412 %	
Root relative squared error	74.9706 %	
Total Number of Instances	1470	

--- Detailed Accuracy By Class ---

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.515	0.010	0.910	0.515	0.658	0.645	0.806	0.649	Yes
	0.990	0.485	0.914	0.990	0.951	0.645	0.806	0.933	No

--- Confusion Matrix ---

a \ b	classified as
122	115   a = Yes
12	1221   b = No

## Classifier: NaiveBayes

**Classifier**

Choose **AttributeSelectedClassifier** -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W weka.classifiers.bayes.NaiveBayes

**Test options**

☒ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☐ Percentage split % 66  
More options...

(Nom) Attrition

Start Stop

**Result list (right-click for options)**

- 13:50:08 - trees.J48
- 13:51:12 - trees.J48
- 13:51:23 - trees.J48
- 13:51:27 - trees.J48
- 15:24:06 - trees.J48
- 16:56:33 - trees.J48
- 16:57:21 - meta.AttributeSelectedClassifier
- 16:57:29 - meta.AttributeSelectedClassifier
- 16:57:32 - meta.AttributeSelectedClassifier
- 16:57:34 - meta.AttributeSelectedClassifier
- 17:15:12 - trees.J48
- 17:15:54 - meta.AttributeSelectedClassifier
- 17:22:10 - meta.AttributeSelectedClassifier
- 17:23:40 - meta.AttributeSelectedClassifier
- 17:24:28 - meta.AttributeSelectedClassifier

**Classifier output**

```
Time taken to build model: 0.01 seconds

--- Evaluation on training set ---

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances      1215      82.6531 %
Incorrectly Classified Instances    255      17.3469 %
Kappa statistic                    0.4062
Mean absolute error                0.251
Root mean squared error            0.3612
Relative absolute error            92.6964 %
Root relative squared error        98.2285 %
Total Number of Instances          1470

--- Detailed Accuracy By Class ---

              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
Weighted Avg.   0.827   0.388   0.841     0.827   0.833     0.409   0.764    0.924    Yes
                  0.827   0.388   0.841     0.827   0.833     0.409   0.764    0.858    No

=== Confusion Matrix ===

  a  b  <-- classified as
133 104 | a = Yes
151 1082 | b = No
```

### 3) Compare different test options (Cross-validation and Percentage split), and explain which one is the best.

The cross-validation is the best. The method, cross validation, takes full advantage of all the samples. The percentage split only uses part of the dataset. As a result, the accuracy of cross-validation is better than the percentage split.

Percentage split (66%):

**Classifier**

Choose **AttributeSelectedClassifier** -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W weka.classifiers.trees.RandomTree -- -K 0 -M 1.0

**Test options**

☐ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☒ Percentage split % 66  
More options...

(Nom) Attrition

Start Stop

**Result list (right-click for options)**

- 13:50:08 - trees.J48
- 13:51:12 - trees.J48
- 13:51:23 - trees.J48
- 13:51:27 - trees.J48
- 15:24:06 - trees.J48
- 16:56:33 - trees.J48
- 16:57:21 - meta.AttributeSelectedClassifier
- 16:57:29 - meta.AttributeSelectedClassifier
- 16:57:32 - meta.AttributeSelectedClassifier
- 16:57:34 - meta.AttributeSelectedClassifier
- 17:15:12 - trees.J48
- 17:15:54 - meta.AttributeSelectedClassifier
- 17:22:10 - meta.AttributeSelectedClassifier
- 17:23:40 - meta.AttributeSelectedClassifier
- 17:24:28 - meta.AttributeSelectedClassifier
- 17:30:05 - meta.AttributeSelectedClassifier

**Classifier output**

```
Size of the tree : 608

Time taken to build model: 0.01 seconds

--- Evaluation on test split ---

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances      383      76.6 %
Incorrectly Classified Instances    117      23.4 %
Kappa statistic                    0.2177
Mean absolute error                0.232
Root mean squared error            0.4806
Relative absolute error            84.058 %
Root relative squared error        126.0038 %
Total Number of Instances          500

--- Detailed Accuracy By Class ---

              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
Weighted Avg.   0.766   0.541   0.773     0.766   0.769     0.218   0.613    0.750    Yes
                  0.766   0.541   0.773     0.766   0.769     0.218   0.613    0.858    No

=== Confusion Matrix ===

  a  b  <-- classified as
 33  55 | a = Yes
 62 350 | b = No
```

Cross-validation (10 folds):

Classifier

Choose
AttributeSelectedClassifier -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W weka.classifiers.trees.RandomTree --K 0 -M 1.0 -V

Test options

☐ Use training set  
☐ Supplied test set Set...  
☒ Cross-validation Folds 10  
☐ Percentage split % 66  
More options...

(Nom) Attrition

Start Stop

Result list (right-click for options)

17:35:53 - meta.AttributeSelectedClassifier  
17:36:04 - meta.AttributeSelectedClassifier  
17:36:14 - meta.AttributeSelectedClassifier  
17:36:20 - meta.AttributeSelectedClassifier  
17:37:50 - meta.AttributeSelectedClassifier

Classifier output

```

| | | | | EnvironmentSatisfaction >= 3.5 : No (15/0)
| | | YearsAtCompany >= 38.5 : Yes (1/0)

Size of the tree : 608

Time taken to build model: 0 seconds

--- Stratified cross-validation ---
=== Summary ===

Correctly Classified Instances      1142           77.6871 %
Incorrectly Classified Instances    328           22.3129 %
Kappa statistic                    0.1942
Mean absolute error                 0.2231
Root mean squared error             0.4724
Relative absolute error              82.393 %
Root relative squared error         128.4505 %
Total Number of Instances          1470

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.338   0.139   0.319     0.338     0.328     0.194   0.599    0.214    Yes
                0.861   0.662   0.871     0.861     0.866     0.194   0.599    0.867    No
Weighted Avg.   0.777   0.578   0.782     0.777     0.779     0.194   0.599    0.762

=== Confusion Matrix ===

      a   b  <-- classified as
      80 157 |  a = Yes
      171 1062 |  b = No

```