Jinyang Li
A20317851
09/16/2015

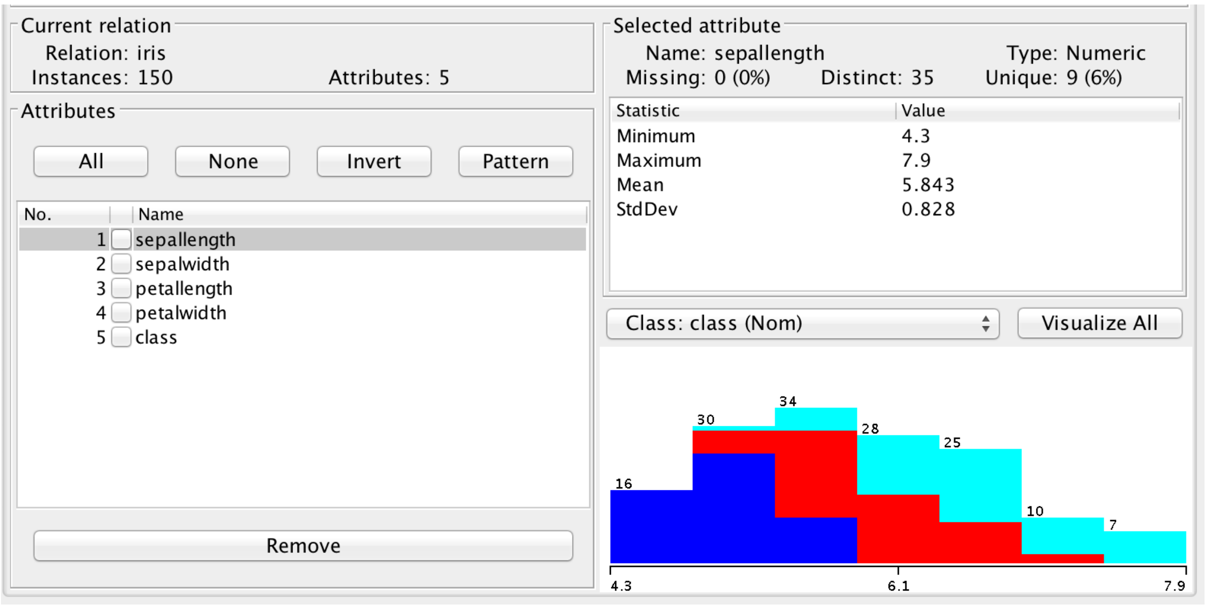
Special data sets analyze report

CS422 HW-02

• Object Overview

In this report, the following data sets provided by Weka were used.

Iris



Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

No.	Name
1	□ sepallength
2	□ sepalwidth
3	□ petallength
4	□ petalwidth
5	□ class

Selected attribute

Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

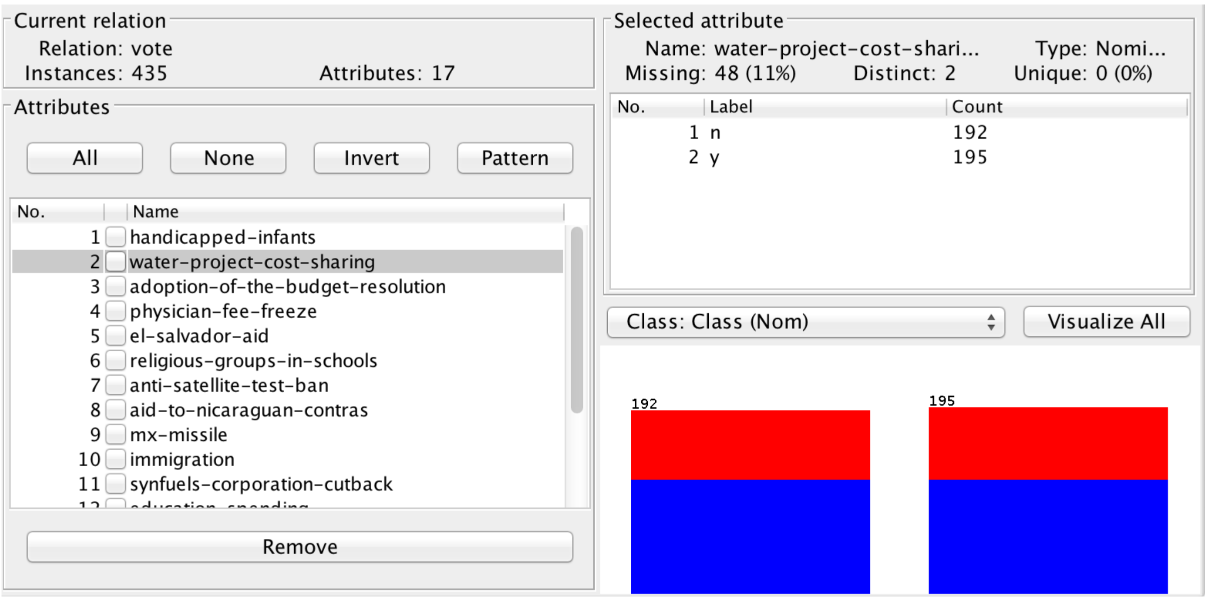
Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

	Min	Max	Mean	StdDev
Sepal length	4.3	7.9	5.84	0.83
Sepal width	2	4.4	3.05	0.43
petal length	1	6.9	3.76	1.76
petal width	0.1	2.5	1.20	0.76
Class	Setosa, Versicolour, Virginica			

the number of attributes: 5

vote



Current relation

Relation: vote

Instances: 435 Attributes: 17

Attributes

No.	Name
1	□ handicapped-infants
2	□ water-project-cost-sharing
3	□ adoption-of-the-budget-resolution
4	□ physician-fee-freeze
5	□ el-salvador-aid
6	□ religious-groups-in-schools
7	□ anti-satellite-test-ban
8	□ aid-to-nicaraguan-contras
9	□ mx-missile
10	□ immigration
11	□ synfuels-corporation-cutback
12	□ education-spending

Selected attribute

Name: water-project-cost-shari... Type: Nomi...

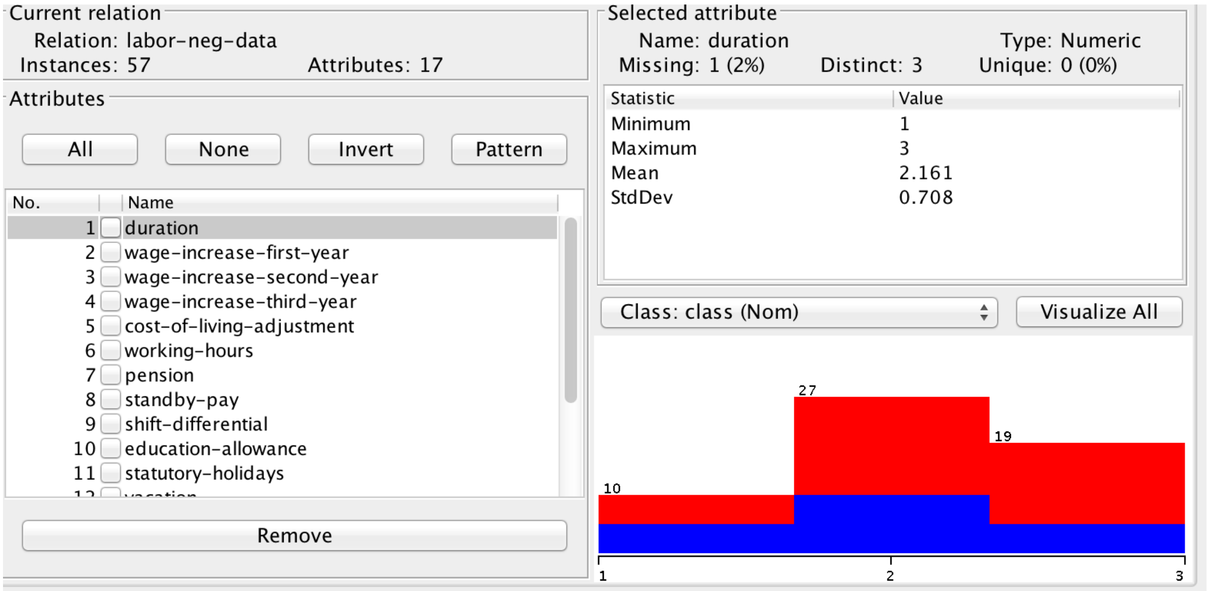
Missing: 48 (11%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	n	192
2	y	195

Class: Class (Nom)

For this data set, the number of attributes is 17, 16 of the attributes consist of the No, Label and the count, which is called nominal type. They are made of distinct results of simple yes or no survey. In class attribute, this data set has democrat and republican labels.

labor



Current relation

Relation: labor-neg-data
Instances: 57 Attributes: 17

Attributes

No.	Name
1	□ duration
2	□ wage-increase-first-year
3	□ wage-increase-second-year
4	□ wage-increase-third-year
5	□ cost-of-living-adjustment
6	□ working-hours
7	□ pension
8	□ standby-pay
9	□ shift-differential
10	□ education-allowance
11	□ statutory-holidays
12	□ vacation

Selected attribute

Name: duration Type: Numeric
Missing: 1 (2%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	3
Mean	2.161
StdDev	0.708

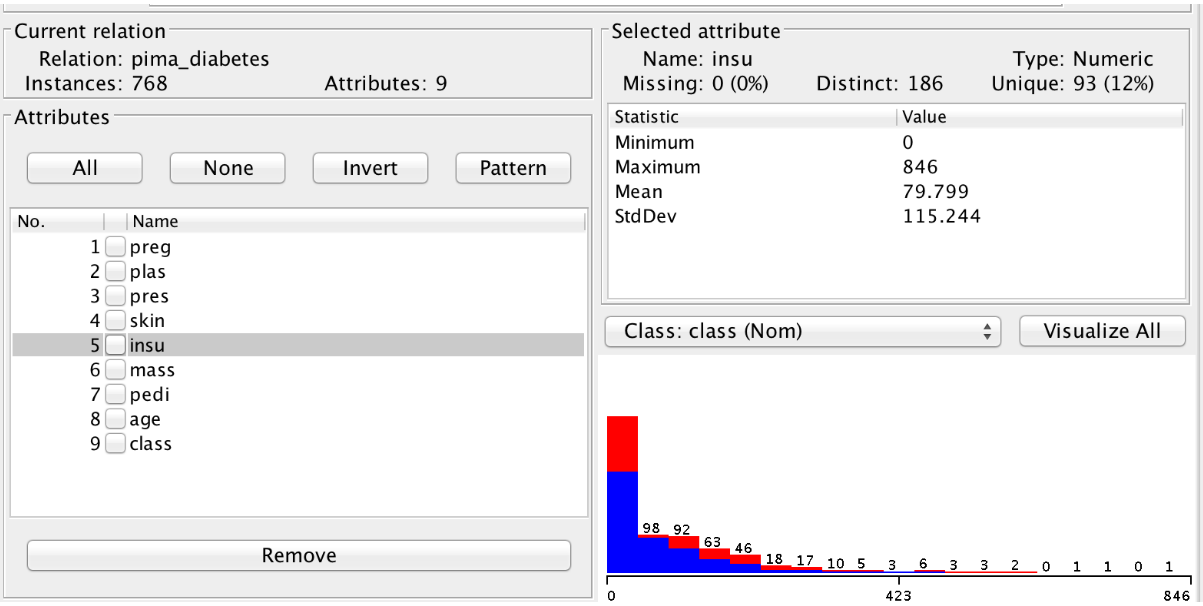
	min	max	mean	StdDev
duration	1	3	2.161	0.708
wage-increase-first year	2	7	3.804	1.371
wage-increase-second-year	2	7	3.972	1.164
wage-increase-third-year	2	5.1	3.913	1.304
working-hours	27	40	38.039	2.506
standby-pay	2	14	7.444	5.028
shift-differential	0	25	4.871	4.544
statutory-holidays	9	15	11.094	1.26
class	bad 20, good 37			

	label	count
cost-of-living-adjustment	**{none, tcf, tc}**	**{22, 8, 7}**
pension	**{none, ret_allw, empl_contr}**	**{11, 4, 12}**
education-allowance	**{yes, no}**	**{10, 12}**
vacation	**{below,_average, average, generous}**	**{18, 17, 16}**
longterm-disability-assistance	**{yes, no}**	**{20, 8}**
contribution-to-dental-plan	**{none, half, full}**	**{9, 15, 13}**
bereavement-assistance	**{yes, no}**	**{27, 3}**
contribution-to-health-plan	**{none, half, full}**	**{8, 9, 20}**

* label {A, B, C} has the compatible value of count {1, 2, 3}

As above, this data set has 17 attributes, 8 of them are numeric, 9 of them are nominal.

Diabetes



Current relation
Relation: pima_diabetes
Instances: 768 Attributes: 9

Attributes

No.	Name
1	□ preg
2	□ plas
3	□ pres
4	□ skin
5	□ insu
6	□ mass
7	□ pedi
8	□ age
9	□ class

Selected attribute
Name: insu Type: Numeric
Missing: 0 (0%) Distinct: 186 Unique: 93 (12%)

Statistic	Value
Minimum	0
Maximum	846
Mean	79.799
StdDev	115.244

Class: class (Nom)

	min	max	mean	StdDev
preg	0	17	3.845	3.37
plas	0	199	120.895	31.973
pres	0	122	69.105	19.356
skin	0	99	20.536	15.952
insu	0	846	79.799	115.244
imass	0	67.1	31.993	7.884
pedi	0.078	2.42	0.472	0.331
age	21	81	33.241	11.76
class	{tested_negative, tested_positive}		{500, 268}	

This data set has 9 attributes.

• Experimental Method

The 2 decision tree algorithms were used in this experiment.

SimpleCart - A basic algorithm used for training set.

Parameters :

-S 1 // The random number seed to be used.

-M 2.0 //The minimal number of observations at the terminal nodes

-N 5 //The number of folds in the internal cross-validation

-C 1.0 //The percentage of the training set size

DecisionStump

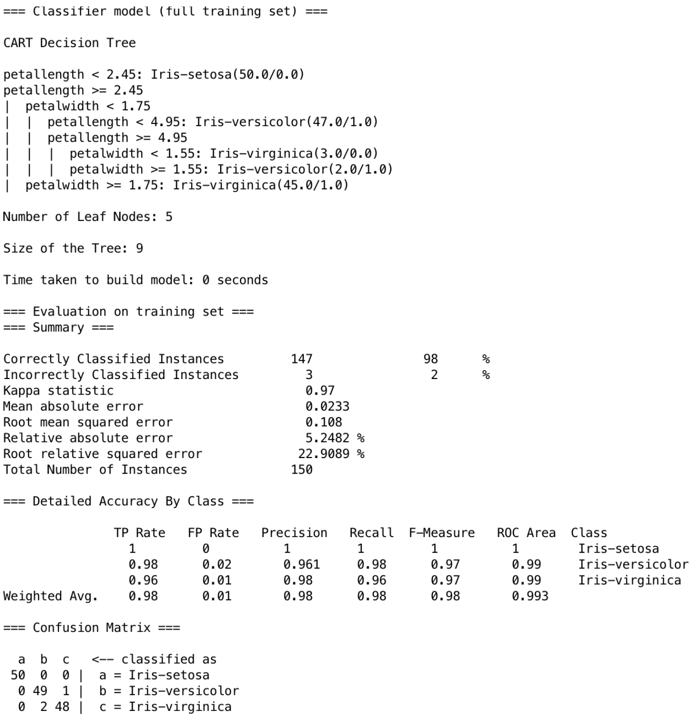
No parameters for this algorithm in Weka.

• Experimental Process

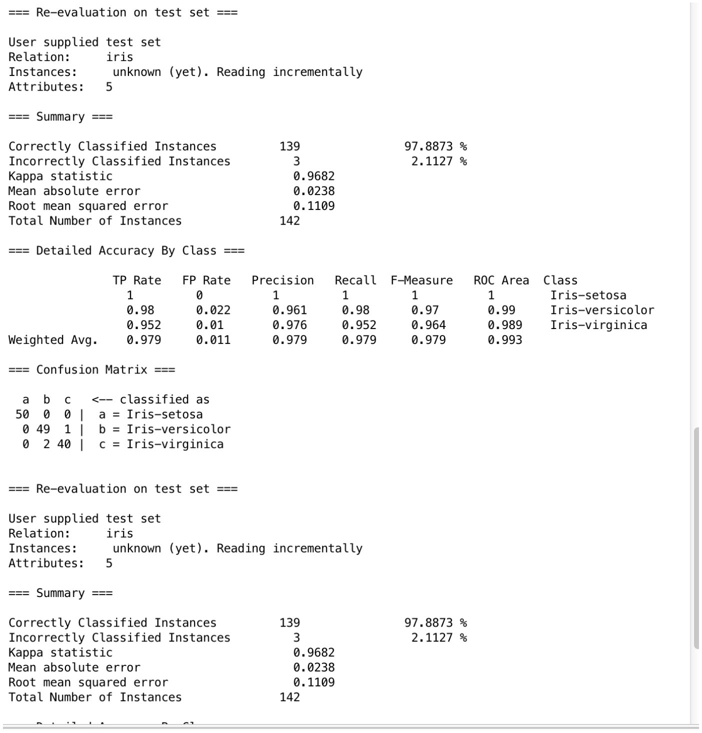
The four data sets were classified by two algorithms separately with training set and test set.

Iris

simpleCart - Traing set test-set-with **delete** 5 instances



```
=== Classifier model (full training set) ===

CART Decision Tree

petallength < 2.45: Iris-setosa(50.0/0.0)
petallength >= 2.45
|  petalwidth < 1.75
|  |  petallength < 4.95: Iris-versicolor(47.0/1.0)
|  |  petallength >= 4.95
|  |  |  petalwidth < 1.55: Iris-virginica(3.0/0.0)
|  |  |  petalwidth >= 1.55: Iris-versicolor(2.0/1.0)
|  petalwidth >= 1.75: Iris-virginica(45.0/1.0)

Number of Leaf Nodes: 5

Size of the Tree: 9

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         147               98      %
Incorrectly Classified Instances         3                2      %
Kappa statistic                          0.97
Mean absolute error                      0.0233
Root mean squared error                  0.108
Relative absolute error                  5.2482 %
Root relative squared error             22.9089 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1        1           1        Iris-setosa
                 0.98      0.02       0.961     0.98     0.97        0.99     Iris-versicolor
                 0.96      0.01       0.98      0.96     0.97        0.99     Iris-virginica
Weighted Avg.    0.98      0.01       0.98      0.98     0.98        0.993

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```



```
=== Re-evaluation on test set ===

User supplied test set
Relation:     iris
Instances:     unknown (yet). Reading incrementally
Attributes:   5

=== Summary ===

Correctly Classified Instances         139               97.8873 %
Incorrectly Classified Instances         3                2.1127 %
Kappa statistic                          0.9682
Mean absolute error                      0.0238
Root mean squared error                  0.1109
Total Number of Instances              142

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1        1           1        Iris-setosa
                 0.98      0.022      0.961     0.98     0.97        0.99     Iris-versicolor
                 0.952     0.01       0.976     0.952    0.964       0.989    Iris-virginica
Weighted Avg.    0.979     0.011      0.979     0.979    0.979       0.993

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  2 40 |  c = Iris-virginica


=== Re-evaluation on test set ===

User supplied test set
Relation:     iris
Instances:     unknown (yet). Reading incrementally
Attributes:   5

=== Summary ===

Correctly Classified Instances         139               97.8873 %
Incorrectly Classified Instances         3                2.1127 %
Kappa statistic                          0.9682
Mean absolute error                      0.0238
Root mean squared error                  0.1109
Total Number of Instances              142
```

There are differences between the training set and the test set. But, it doesn't matter since the accuracy Avg. is very similar, the same result when use decision stump to do the same process. The **size of the tree & the number of the leaves** are the same for both algorithms from results.

When change the parameters, (only available on Simple cart)

—modify the M from 2.0 to 4.0

the number of leaf nodes change from 5 to 4, the size of the tree change from 9 to 7

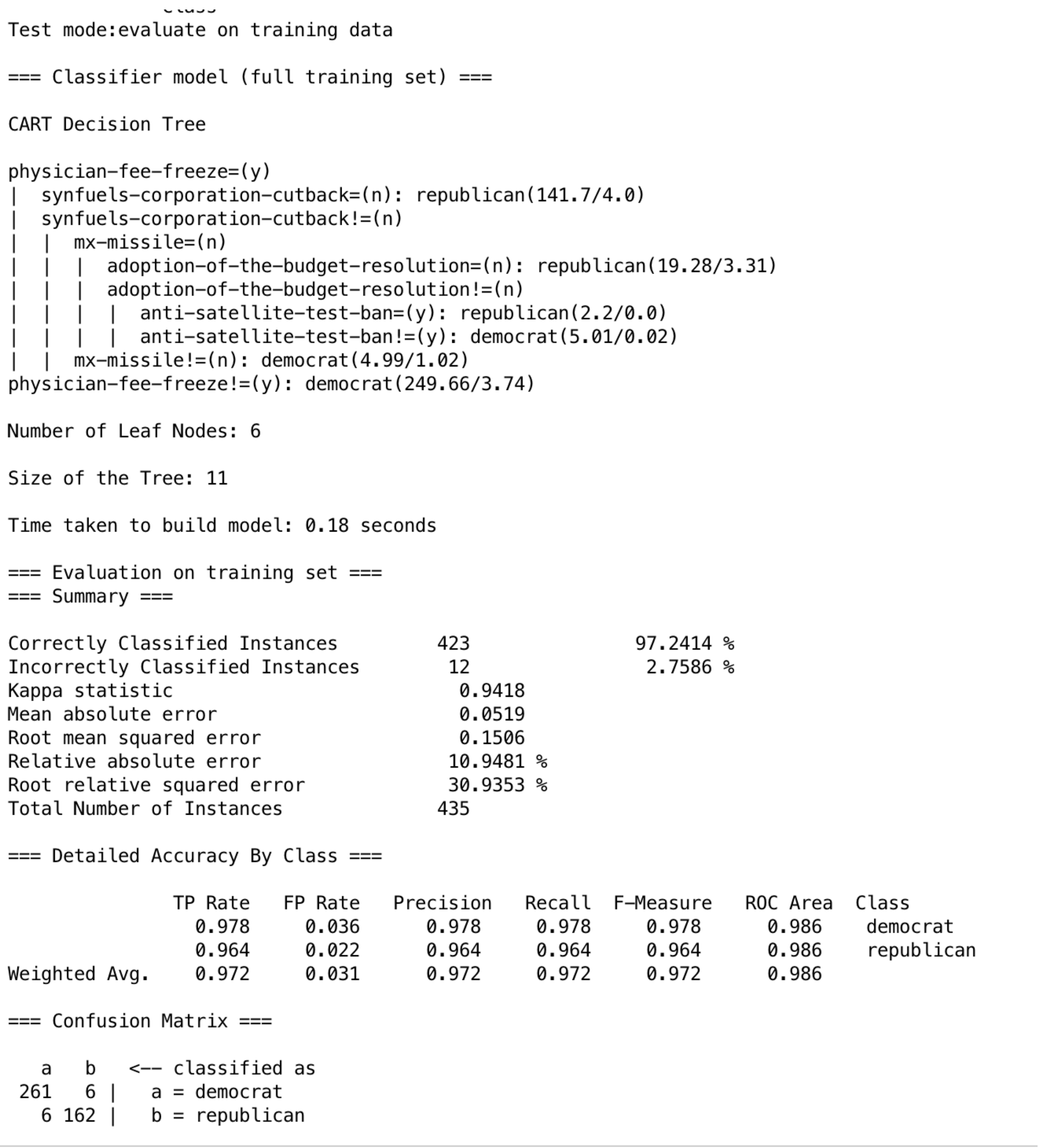
—modify the M from 2.0 to 4.0

the number of leaf nodes change from 5 to 3, the size of the tree change from 9 to 5

The accuracy was dropped down when change the minimal number of observation of terminal trees.

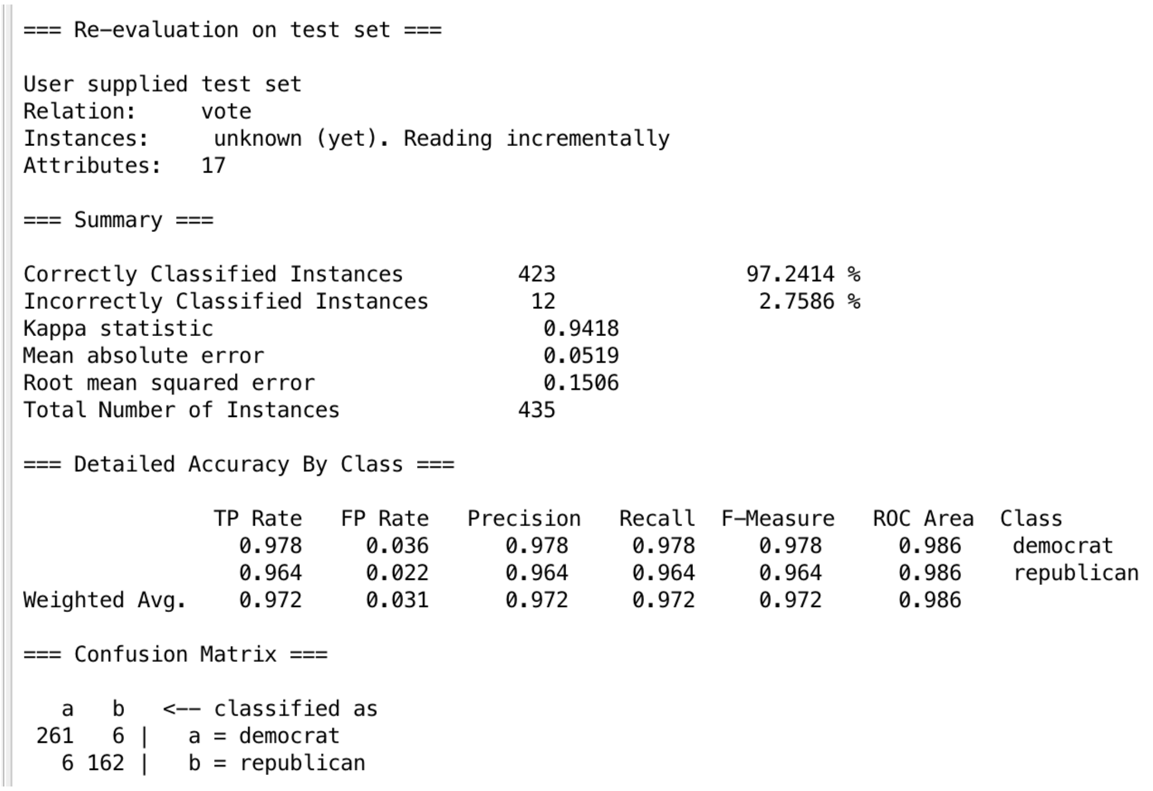
The accuracy also decrease when change the '-s' and '-C 'parameters.

Vote



```
Test mode:evaluate on training data

=== Classifier model (full training set) ===

CART Decision Tree

physician-fee-freeze=(y)
|  synfuels-corporation-cutback=(n): republican(141.7/4.0)
|  synfuels-corporation-cutback!=(n)
|  |  mx-missile=(n)
|  |  |  adoption-of-the-budget-resolution=(n): republican(19.28/3.31)
|  |  |  adoption-of-the-budget-resolution!=(n)
|  |  |  |  anti-satellite-test-ban=(y): republican(2.2/0.0)
|  |  |  |  anti-satellite-test-ban!=(y): democrat(5.01/0.02)
|  |  mx-missile!=(n): democrat(4.99/1.02)
physician-fee-freeze!=(y): democrat(249.66/3.74)

Number of Leaf Nodes: 6

Size of the Tree: 11

Time taken to build model: 0.18 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         423               97.2414 %
Incorrectly Classified Instances        12                2.7586 %
Kappa statistic                          0.9418
Mean absolute error                      0.0519
Root mean squared error                  0.1506
Relative absolute error                 10.9481 %
Root relative squared error             30.9353 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.978     0.036      0.978     0.978     0.978      0.986    democrat
                 0.964     0.022      0.964     0.964     0.964      0.986    republican
Weighted Avg.    0.972     0.031      0.972     0.972     0.972      0.986

=== Confusion Matrix ===

   a   b   <-- classified as
 261   6 |   a = democrat
   6 162 |   b = republican
```

Training set



```
=== Re-evaluation on test set ===

User supplied test set
Relation:     vote
Instances:     unknown (yet). Reading incrementally
Attributes:   17

=== Summary ===

Correctly Classified Instances         423               97.2414 %
Incorrectly Classified Instances        12                2.7586 %
Kappa statistic                          0.9418
Mean absolute error                      0.0519
Root mean squared error                  0.1506
Total Number of Instances              435

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.978     0.036      0.978     0.978     0.978      0.986    democrat
                 0.964     0.022      0.964     0.964     0.964      0.986    republican
Weighted Avg.    0.972     0.031      0.972     0.972     0.972      0.986

=== Confusion Matrix ===

   a   b   <-- classified as
 261   6 |   a = democrat
   6 162 |   b = republican
```

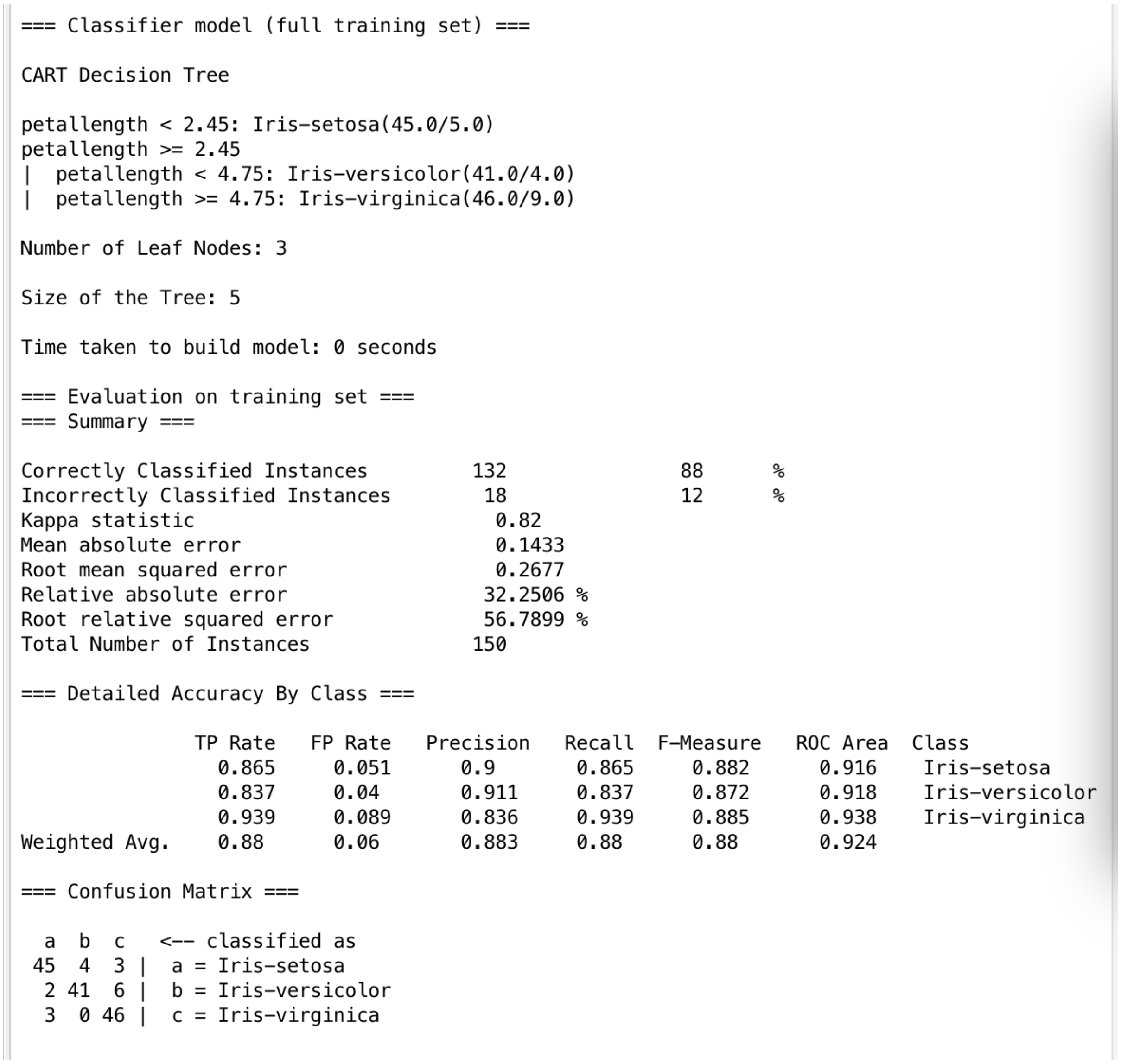
Test set

The same accuracy without modify data set.

For other 2 data sets, when do the same process, the result regarding accuracy changes is the same as Iris and vote data sets.

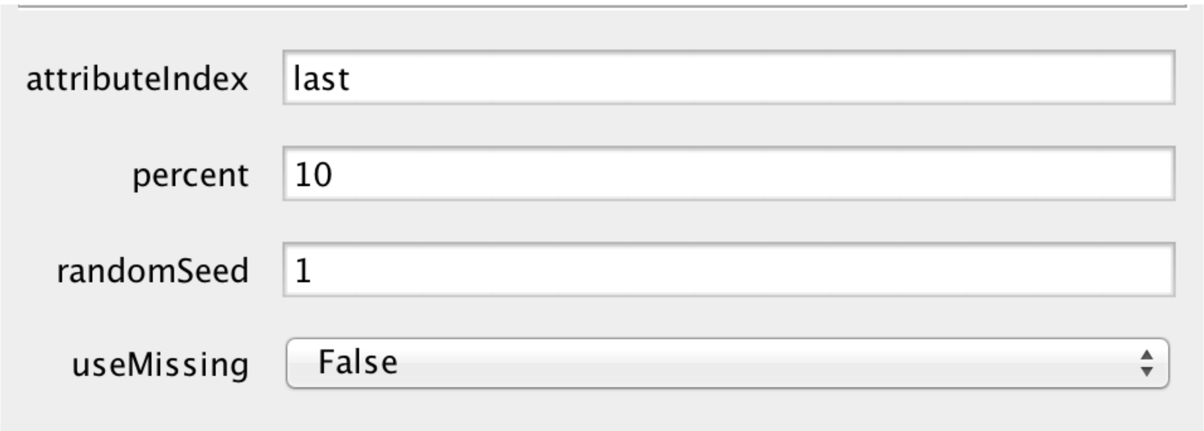
- Introduce noise

When apply the noise to iris data set and using simple cart algorithm



```
=== Classifier model (full training set) ===

CART Decision Tree

petallength < 2.45: Iris-setosa(45.0/5.0)
petallength >= 2.45
|  petallength < 4.75: Iris-versicolor(41.0/4.0)
|  petallength >= 4.75: Iris-virginica(46.0/9.0)

Number of Leaf Nodes: 3

Size of the Tree: 5

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         132               88      %
Incorrectly Classified Instances        18               12      %
Kappa statistic                          0.82
Mean absolute error                      0.1433
Root mean squared error                  0.2677
Relative absolute error                 32.2506 %
Root relative squared error             56.7899 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.865     0.051      0.9       0.865     0.882      0.916    Iris-setosa
                 0.837     0.04       0.911     0.837     0.872      0.918    Iris-versicolor
                 0.939     0.089      0.836     0.939     0.885      0.938    Iris-virginica
Weighted Avg.    0.88      0.06       0.883     0.88      0.88       0.924

=== Confusion Matrix ===

  a  b  c   <-- classified as
 45  4  3 |  a = Iris-setosa
  2 41  6 |  b = Iris-versicolor
  3  0 46 |  c = Iris-virginica
```

As the result showing above with addNoise() option below:



attributeIndex: last

percent: 10

randomSeed: 1

useMissing: False

Clearly, with more noise data, the accuracy of correctly classified instances was drop down to 88%. Also, the number of leaf nodes decrease to 3. The errors increase large with noise.

When add missing value, there’s no large change to accuracy.

• Class distribution

For iris, the class distributions are:

setOsa, versiColor, virginica

When replace the class setosa to virginca of some objects in data set, the number of correct classified objects was decreased. Apply same step to vote, accuracy also decreased.

This means the class distribution do effect the result of experiment.

Cross-validation data sets	training tree size	test reesize
iris	9	9
vote	11	11
labor	3	3
diabetes	5	5

We do have the same distribution.

• Conclusion

In weka, the specific classier parameters would performance great with default options when analyzing the data sets provided by weka. Through modify the options to make the accuracy to increase is the most important step when using weka to analyzing data sets that from the real world.