

ECT_HW8

2019

第一大題

- 利用 Weka 對 DeerHunter.arff 進行前處理，依序完成以下步驟及問題：
 - Replace Missing Value，需列出補上的值為何 (10%)
 - Outlier Detection& Remove (10%)
 - Attribute Selection，請篩選出適合的屬性 (10%)

Replace Missing Value

以平均值代替

10: gender	11: age	12: hunter
Numeric	Numeric	Numeric
1.0	18.0	11
1.0	18.0	8
1.0	18.0	6
1.0	18.0	10
1.0	18.0	5
1.0	19.0	8
1.0	19.0	9
1.0		6
1.0	19.0	8
1.0	19.0	2
1.0	19.0	3
1.0	19.0	11

married	9: income	10: gender
Numeric	Numeric	Numeric
1.0	15000.0	
1.0	5000.0	
1.0	40000.0	
1.0	15000.0	
0.0	40000.0	
0.0	15000.0	
0.0	15000.0	
0.0		
1.0	22500.0	
1.0	15000.0	

Name: age	Type: Numeric
Missing: 5 (0%)	Distinct: 68
	Unique: 3 (0%)
Statistic	Value
Minimum	18
Maximum	90
Mean	38.316
StdDev	12.875

Selected attribute	
Name: income	Type: Numeric
Missing: 5 (0%)	Distinct: 7
	Unique: 0 (0%)
Statistic	Value
Minimum	5000
Maximum	85000
Mean	37678.394
StdDev	20096.855

10: gender	11: age
Numeric	Numeric
1.0	18.0
1.0	18.0
1.0	18.0
1.0	18.0
1.0	18.0
1.0	19.0
1.0	19.0
1.0	38.316
1.0	18.0

9: income	10: gender
Numeric	Numeric
15000.0	
5000.0	
40000.0	
15000.0	
40000.0	
15000.0	
15000.0	
15000.0	
37678.394	
22500.0	
15000.0	

Filter

ReplaceMissingValues

Current relation

Relation: DeerHunter_missing-weka.filters.unsupervi...
Instances: 6059

Attributes: 21
Sum of weights: 6059

Attributes

Outlier Detection & Remove

Current relation

Relation: DeerHunter_missing-weka.filters.unsupervi...	Attributes: 21
Instances: 6059	Sum of weights: 6059

Attributes

Choose **RemoveMisclassified** -W "weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4" -C -1 -F 0 -T 0.1 -I 0 **Apply** **Stop**

Current relation

Relation: DeerHunter_missing-weka.filters.unsupervi...	Attributes: 21
Instances: 6059	Sum of weights: 6059

Attributes

Selected attribute

Name: gender	Distinct: 2	Type: Numeric
Missing: 0 (0%)		Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.936
StdDev	0.244

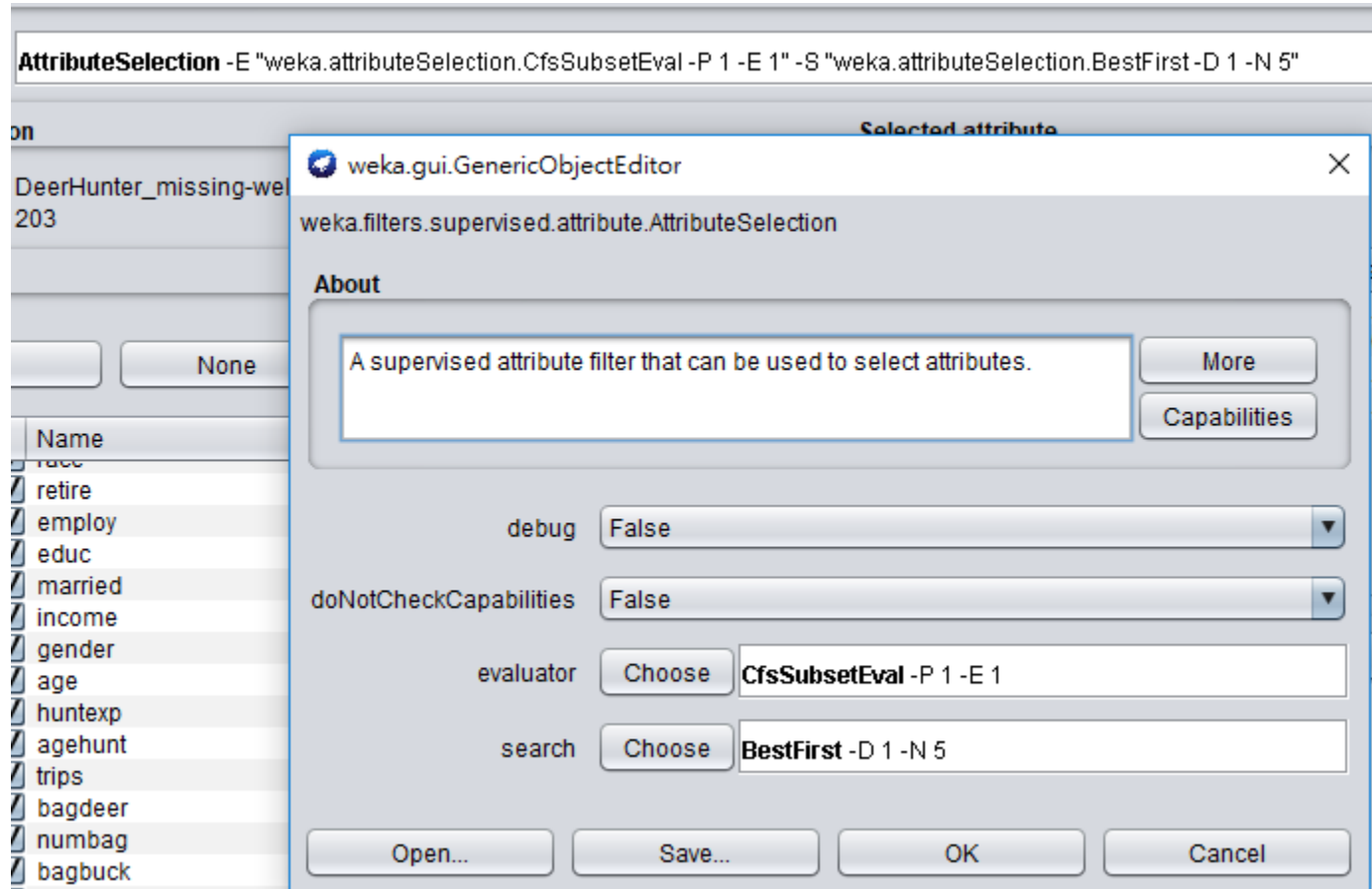
Attributes

All **None** **Invert** **Pattern**

Current relation

Relation: DeerHunter_missing-weka.filters.unsupervi...	Attributes: 21
Instances: 203	Sum of weights: 203

Attribute Selection



Attributes: 21
Sum of weights: 203



Attributes: 15
Sum of weights: 203

第二大題

- 請用weka對BreastCancer.csv對目標diagnosis進行Ensemble learning並與未使用的結果進行比較(請列出重要過程及適當說明)：
 - 以10 Folds cross-validation進行J48分類(5%)
 - 以10 Folds cross-validation進行Bagging分類並選擇J48 classifier進行分類(10%)
 - 以10 Folds cross-validation進行Bagging分類並選擇Randomforest進行分類(10%)
 - 以10 Folds cross-validation進行AdaBoost分類並選擇DecisionStump進行分類(10%)

J48分類

The screenshot shows the Weka GUI with the J48 classifier selected. The 'Test options' panel on the left shows 'Cross-validation' is selected with 10 folds. The 'Classifier output' panel on the right displays the results of the stratified cross-validation.

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %
-

(Nom) diagnosis

Result list (right-click for options)

Time	Model
21:32:24	trees.J48

Classifier output

Time taken to build model: 0.04 seconds

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

Correctly Classified Instances	530	93.1459 %
Incorrectly Classified Instances	39	6.8541 %
Kappa statistic	0.8541	
Mean absolute error	0.0741	
Root mean squared error	0.2574	
Relative absolute error	15.8345 %	
Root relative squared error	53.2317 %	
Total Number of Instances	569	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
	0.920	0.062	0.899	0.920	0.909	0.854
	0.938	0.080	0.952	0.938	0.945	0.854

Bagging J48

Choose **Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Test options

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

More options...

(Nom) diagnosis ▼

Start Stop

Result list (right-click for options)

21:32:24 - trees.J48
21:33:06 - meta.Bagging

Classifier output

```
symmetry_worst
fractal_dimension_worst
Test mode: 10-fold cross-validation

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

Bagging with 10 iterations and base learner

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

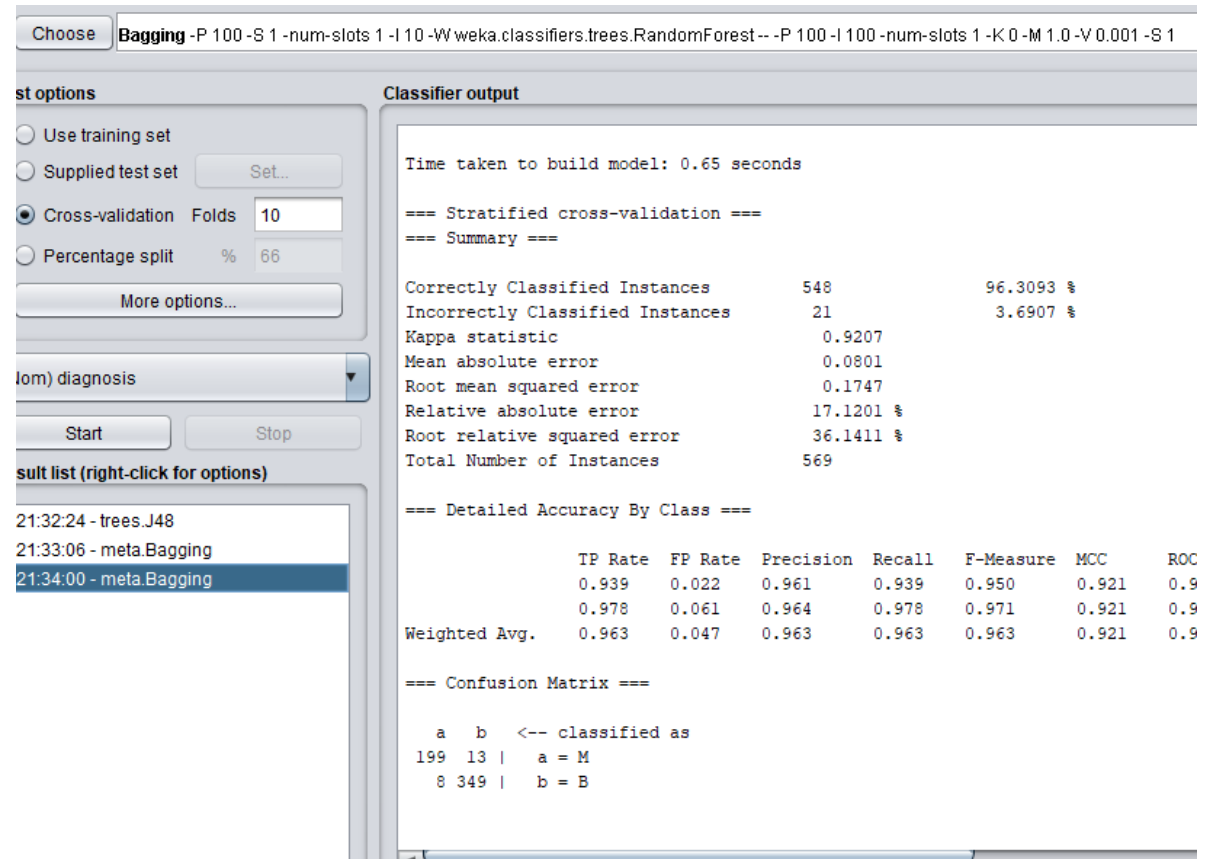
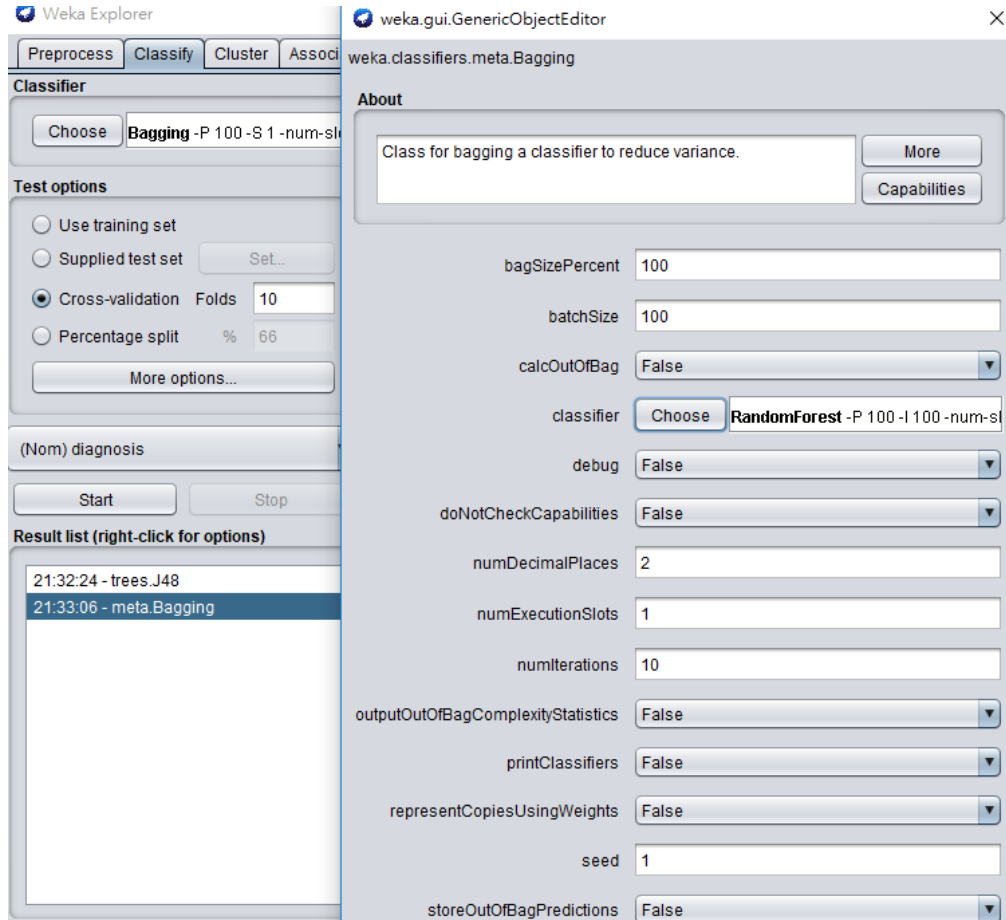
Time taken to build model: 0.05 seconds

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

Correctly Classified Instances      541      95.0791 %
Incorrectly Classified Instances    28       4.9209 %
Kappa statistic                    0.8947
Mean absolute error                 0.0771
Root mean squared error            0.1973
Relative absolute error            16.4792 %
Root relative squared error        40.8019 %
Total Number of Instances          569

=== Detailed Accuracy By Class ===
```


Bagging Randomforest



AdaBoost DecisionStump

The screenshot shows the Weka GUI with the AdaBoostM1 classifier selected. The 'Test options' panel on the left is configured for cross-validation with 10 folds. The 'Classifier output' panel on the right displays the results of the cross-validation, including a summary of performance metrics and a detailed accuracy breakdown by class.

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %
-

(Nom) diagnosis

Result list (right-click for options)

- 21:32:24 - trees.J48
- 21:33:06 - meta.Bagging
- 21:34:00 - meta.Bagging
- 21:34:35 - meta.AdaBoostM1

Classifier output

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	538	94.5518 %
Incorrectly Classified Instances	31	5.4482 %
Kappa statistic	0.8834	
Mean absolute error	0.0578	
Root mean squared error	0.1967	
Relative absolute error	12.3539 %	
Root relative squared error	40.6882 %	
Total Number of Instances	569	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MC
	0.925	0.042	0.929	0.925	0.927	0.
	0.958	0.075	0.955	0.958	0.957	0.
Weighted Avg.	0.946	0.063	0.945	0.946	0.945	0.

=== Confusion Matrix ===

a b <-- classified as

196	16	a = M
15	342	b = B

第三大題

- 請用python對BreastCancer.csv對目標diagnosis進行Ensemble learning並與未使用的結果進行比較(請列出重要過程及適當說明)：
 - 以10 Folds cross-validation進行DecisionTreeClassifier分類(5%)
 - 以10 Folds cross-validation進行Bagging , n_estimators=10分類(10%)
 - 以10 Folds cross-validation進行RandomForest分類(10%)
 - 以10 Folds cross-validation進行AdaBoost, n_estimators=10分類(10%)

Import/data processing

```
In [1]: import pandas as pd
        from sklearn import preprocessing
        from sklearn.model_selection import cross_val_score
        from sklearn.model_selection import train_test_split
        from sklearn.ensemble import BaggingClassifier
        from sklearn.ensemble import AdaBoostClassifier
        from sklearn import metrics
        from sklearn import tree
        from sklearn.ensemble import RandomForestClassifier
```

```
In [2]: data = pd.read_csv('BreastCancer.csv')
```

```
In [3]: data.dropna()
        label=data.iloc[:,1]
```

```
In [4]: data = data.drop('diagnosis',axis=1)
```

```
In [5]: le = preprocessing.LabelEncoder()
        encodedlabel = le.fit_transform(label)
```

```
In [6]: train_X, test_X, train_y, test_y = train_test_split(data, encodedlabel, test_size = 0.3)
```

DecisionTree

```
clf = tree.DecisionTreeClassifier(criterion = 'entropy',max_depth=3,max_leaf_nodes = 4)

scores = cross_val_score(clf, data, encodedlabel, cv=10)
scores.mean()
```

0.9016744447325209

Bagging

```
bagging = BaggingClassifier(n_estimators = 10)

scores = cross_val_score(bagging, data, encodedlabel, cv=10)
scores.mean()
```

0.9509938639702705

RandomForest

```
Rndclf = RandomForestClassifier(n_estimators=10)

scores = cross_val_score(Rndclf, data, encodedlabel, cv=10)
scores.mean()
```

0.9684480598046841

AdaBoost

```
adaBoost = AdaBoostClassifier(n_estimators = 100)

scores = cross_val_score(adaBoost, data, encodedlabel, cv=10)
scores.mean()
```

0.9702326938034741

python cross-validation

- sklearn.model_selection. cross_val_score

```
scores = cross_val_score(分類方法, 屬性, 目標, fold量)  
scores.mean()
```

- sklearn.ensemble(base_estimator(**default=decision tree**))
 - BaggingClassifier (n_estimators)
 - AdaBoostClassifier(n_estimators)