

Pass Tas10 4.1: Attac10 Classification using Naïve Bayes Algorithm

“Naïve Bayes” classification:

Weka Explorer | Dragon Center

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

Classifier

Choose | **NaiveBayes**

Test options

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

(Nom) class

Start | Stop

Result list (right-click for options)

- 18.02.07 - bayes NaiveBayes
- 18.03.32 - bayes NaiveBayes

Classifier output

```

precision              0.01      0.01

Time taken to build model: 0.67 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 1.23 seconds

=== Summary ===

Correctly Classified Instances   113902          90.4178 %
Incorrectly Classified Instances 12071          9.5822 %
Kappa statistic                 0.8067
Mean absolute error             0.0962
Root mean squared error         0.3054
Relative absolute error         19.3417 %
Root relative squared error     61.2231 %
Total Number of Instances      125973

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
          0.937   0.134   0.890   0.937   0.913   0.808   0.967   0.964   normal
          0.866   0.063   0.923   0.866   0.894   0.808   0.965   0.949   anomaly
Weighted Avg.   0.904   0.101   0.905   0.904   0.904   0.808   0.966   0.957

=== Confusion Matrix ===

  a    b  <-- classified as
63106 4237 |  a = normal
 7834 50796 |  b = anomaly
  
```

10-fold cross-validation:

Weka Explorer

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

Classifier

Choose | **NaiveBayes**

Test options

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

(Nom) class

Start | Stop

Result list (right-click for options)

- 18.02.07 - bayes NaiveBayes
- 18.03.32 - bayes NaiveBayes

Classifier output

```

mean              0.0447      0.207
std. dev.         0.1922      0.4034
weight sum        67343      58630
precision         0.01      0.01

Time taken to build model: 0.6 seconds

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

Correctly Classified Instances   113058          90.3829 %
Incorrectly Classified Instances 12115          9.6171 %
Kappa statistic                 0.806
Mean absolute error             0.0965
Root mean squared error         0.3058
Relative absolute error         19.3947 %
Root relative squared error     61.3067 %
Total Number of Instances      125973

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
          0.936   0.134   0.890   0.936   0.912   0.807   0.967   0.964   normal
          0.866   0.064   0.922   0.866   0.893   0.807   0.965   0.949   anomaly
Weighted Avg.   0.904   0.101   0.905   0.904   0.904   0.807   0.966   0.957

=== Confusion Matrix ===

  a    b  <-- classified as
63060 4283 |  a = normal
 7832 50798 |  b = anomaly
  
```

Using test dataset:

Test options

- ☐ Use training set
- ☒ Supplied test set
- ☐ Cross-validation Folds: 10
- ☐ Percentage split %: 66
-

(Nom) class:

Result list (right-click for options)

- 18.02.07 - bayes.NaiveBayes
- 18.03.32 - bayes.NaiveBayes
- 18.06.41 - bayes.NaiveBayes

Classifier output

```

precision                0.01      0.01

Time taken to build model: 0.53 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.28 seconds

=== Summary ===

Correctly Classified Instances   17160      76.1178 %
Incorrectly Classified Instances  5384      23.8822 %
Kappa statistic                  0.5365
Mean absolute error              0.2307
Root mean squared error          0.4862
Relative absolute error          47.2824 %
Root relative squared error      96.1029 %
Total Number of Instances       22544

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MDC     ROC Area  FRC Area  Class
          0.931   0.367   0.657     0.931   0.771     0.572   0.895    0.844    normal
          0.633   0.069   0.924     0.633   0.751     0.572   0.917    0.911    anomaly
Weighted Avg.   0.761   0.198   0.809     0.761   0.759     0.572   0.908    0.882

=== Confusion Matrix ===

  a  b  <-- classified as
9041 670 |  a = normal
4714 8119 |  b = anomaly
  
```

Comparing the results between 10-fold cross-validation and the one obtained using the test dataset. Using the confusion matrix to explain the results:

1. The correctly classified instance of 10 fold cross-validation is higher [90.3%] compared to that of the test data set [76.1%].

2. Confusion matrix of 10 cross-validation

=== Confusion Matrix ===

```

a  b  <-- classified as
63060 4283 |  a = normal
7832 50798 |  b = anomaly
  
```

3. Confusion matrix of the test data set

=== Confusion Matrix ===

```

a  b  <-- classified as
9041 670 |  a = normal
4714 8119 |  b = anomaly
  
```

As observed from both of these confusion matrices the TP, TN is considerably higher than FP and FN comparatively between these matrices 2 & 3 which are directly proportionate to the precision and accuracy score.

10 fold cross-validation:

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 10-1 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Reference:

[1]. <https://www.openml.org/a/estimation-procedures/1>