

Credit Task 4.2: Intrusion Detection using Supervised Learning Techniques

“Naïve Bayes” classification:

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

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 ===

Metric	Value	Percentage
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
Weighted Avg.	0.904	0.101	0.905	0.904	0.904	0.808	0.966	0.957	normal
									anomaly

=== Confusion Matrix ===

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

10-fold cross-validation:

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

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 ===

Metric	Value	Percentage
Correctly Classified Instances	113858	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
Weighted Avg.	0.904	0.101	0.905	0.904	0.904	0.807	0.966	0.957	normal
									anomaly

=== Confusion Matrix ===

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

Using test dataset:

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
- 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      76.1176 %
Incorrectly Classified Instances    5384      23.8822 %
Kappa statistic                    0.5365
Mean absolute error                 0.2397
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  MCC   ROC Area  PRC Area  Class
          0.531   0.367   0.657    0.921   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.

6.)

For Train dataset:

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
lazy.IBK	1	1	1	1	1	1
trees.DecisionStump	0.92	0.079	0.922	0.922	0.922	0.922
trees.RandomTree	1.0	1.0	1.0	1.0	1.0	1.0
bayes.BayesNet	0.972	0.031	0.973	0.972	0.972	0.998
rules.OneR	0.964	0.032	0.966	0.964	0.964	0.966

For Test dataset

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
lazy.IBK	0.794	0.165	0.841	0.794	0.792	0.814
trees.DecisionStump	0.800	0.162	0.841	0.800	0.799	0.819
trees.RandomTree	0.814	0.160	0.837	0.814	0.814	0.827
bayes.BayesNet	0.744	0.200	0.822	0.744	0.739	0.945
rules.OneR	0.814	0.151	0.851	0.814	0.814	0.831

Resample at data size of 20%:

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit

Filter: Choose **Resample -B 0.0 -S 1 -Z 20.0**

Current relation: Relation: KDDTrain-weka filters supervised instance Resample-B0.0-S1-Z20.0
Instances: 25194 Attributes: 42 Sum of weights: 25194

Attributes

All None Invert Pattern

No.	Name
20	is_outgoing_email
21	is_host_login
22	is_guest_login
23	count
24	srv_count
25	error_rate
26	srv_error_rate
27	error_rate
28	srv_error_rate
29	same_srv_rate
30	diff_srv_rate
31	srv_diff_host_rate
32	dst_host_count
33	dst_host_srv_count
34	dst_host_same_srv_rate
35	dst_host_diff_srv_rate
36	dst_host_same_src_port_rate
37	dst_host_srv_diff_host_rate
38	dst_host_error_rate
39	dst_host_srv_error_rate
40	dst_host_error_rate
41	dst_host_srv_error_rate
42	class

Remove

Selected attribute

No.	Label	Count	Weight
1	normal	13468	13468.0
2	anomaly	11726	11726.0

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

Class: class (Nom) Visualize All

Visualized view showing two colored squares representing the classes: normal (blue) and anomaly (red).

SVM classifier (SMO) using POLY:

Weka Explorer

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

Choose **SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V 1 -W 1 -K 'weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01' -calibrator 'weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4'**

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)

- 20:24:42 - lazyJk
- 20:55:40 - trees.DecisionStump
- 20:55:59 - trees.RandomTree
- 20:58:41 - bayes.BayesNet
- 21:09:42 - rules.OneR
- 21:22:00 - lazyJk
- 21:25:56 - trees.DecisionStump
- 21:26:14 - trees.RandomTree
- 21:26:31 - bayes.BayesNet
- 21:26:50 - rules.OneR
- 21:42:34 - functions.SMO**
- 21:44:16 - functions.SMO

Classifier output

Time taken to build model: 34.83 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.13 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	16810	74.5653 %
Incorrectly Classified Instances	5734	25.4347 %
Kappa statistic	0.5077	
Mean absolute error	0.2543	
Root mean squared error	0.5043	
Relative absolute error	50.387 %	
Root relative squared error	95.6801 %	
Total Number of Instances	22544	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
Weighted Avg.	0.525	0.390	0.642	0.925	0.758	0.546	0.767	0.626	normal
	0.610	0.075	0.915	0.610	0.732	0.546	0.767	0.760	anomaly

=== Confusion Matrix ===

	a	b	<-- classified as
8982 729	a = normal		
5005 7828	b = anomaly		

SVM classifier (SMO) using RBF:

Classifier
Choose: SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.RBFGaussianKernel -C 250007 -G 0.01" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

Test options
☐ Use training set
☒ Supplied test set
☐ Cross-validation Folds: 10
☐ Percentage split %: 66
 More options...
 (Nom) class
 Start Stop

Result list (right-click for options)
 20:24:42 - lazyJk
 20:55:40 - trees.DecisionStump
 20:55:59 - trees.RandomTree
 20:58:41 - bayes.BayesNet
 21:09:42 - rules.OneR
 21:22:00 - lazyJk
 21:25:56 - trees.DecisionStump
 21:28:14 - trees.RandomTree
 21:28:31 - bayes.BayesNet
 21:28:50 - rules.OneR
 21:42:34 - functions.SMO
 21:44:16 - functions.SMO

Classifier output

```

Time taken to build model: 118.07 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 14.68 seconds

=== Summary ===

Correctly Classified Instances      16414           72.8087 %
Incorrectly Classified Instances    6130           27.1913 %
Kappa statistic                    0.4762
Mean absolute error                 0.2719
Root mean squared error             0.5215
Relative absolute error             53.8668 %
Root relative squared error         109.0647 %
Total Number of Instances          22544

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  PRC Area  Class
          -----  -
0.925   0.421   0.624   0.925   0.746   0.520   0.752   0.610   normal
0.579   0.075   0.911   0.579   0.708   0.520   0.752   0.767   anomaly
Weighted Avg.   0.728   0.224   0.788   0.728   0.724   0.520   0.752   0.699

=== Confusion Matrix ===

  a  b  <-- classified as
8986 725 |  a = normal
5405 7428 |  b = anomaly
  
```

Confusion matrix and computation time of SVM classifier (SMO) using POLY:

=== Confusion Matrix ===

```

a  b  <-- classified as
8982 729 |  a = normal
5005 7828 |  b = anomaly
  
```

Time taken to build model: 34.83 seconds

Time taken to test model on supplied test set: 0.13 seconds

Confusion matrix and computation time of SVM classifier (SMO) using POLY:

=== Confusion Matrix ===

```

a  b  <-- classified as
8986 725 |  a = normal
5405 7428 |  b = anomaly
  
```

Time taken to build model: 118.07 seconds

Time taken to test model on supplied test set: 14.68 seconds

Even though the confusion matrix between the two aren't showing much difference there is surely a notable difference in the computation time taken by both on the same machine.