Data Mining and Decision Systems

08338

Assessed Coursework

Data Mining of Legacy Data

Stage 3. Classifier Performance

Student Number: 201303310

PDF Due: 2pm 30 November 2015 via E-Bridge

Monday, 30 November 2015

# Classifier Performance

Classifier Performance on the given data (**Data\_Base**) and cleaned data (e.g. **Data\_Clean2**) is described in Table 1 below, based on the worksheet **Performance**.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Data** | **RMSE** | **Accuracy** | **TP** | **FP** | **TN** | **FN** | **Sensitivity** | **Specificity** |
| J48 | Data\_Base | 0.1638 | 93.94 | 377 | 27 | 584 | 35 | 0.9150 | 0.9558 |
| J48 | Data\_RiskKnown | 0.2232 | 94.61 | 379 | 23 | 586 | 32 | 0.9221 | 0.9622 |
| J48 | Data\_Clean0 | 0.2253 | 94.38 | 376 | 24 | 582 | 33 | 0.9193 | 0.9604 |
| J48 | Data\_Clean1 | 0.1793 | 96.50 | 382 | 17 | 583 | 18 | 0.9550 | 0.9717 |
|  |  |  |  |  |  |  |  |  |  |
| J48 | Data\_Clean2 | 0.1793 | 96.50 | 382 | 17 | 583 | 18 | 0.9550 | 0.9717 |
| NaiveBayes | Data\_Clean2 | 0.1917 | 94.90 | 366 | 17 | 583 | 34 | 0.9150 | 0.9717 |
| SMO | Data\_Clean2 | 0.1844 | 96.60 | 384 | 18 | 582 | 16 | 0.9600 | 0.9700 |
| JRip | Data\_Clean2 | 0.1710 | 97.00 | 384 | 14 | 586 | 16 | 0.9600 | 0.9767 |
| Ridor | Data\_Clean2 | 0.1761 | 96.90 | 376 | 7 | 593 | 24 | 0.9400 | 0.9883 |
| NNge | Data\_Clean2 | 0.1342 | 98.20 | 391 | 9 | 591 | 9 | 0.9775 | 0.9850 |
| PART | Data\_Clean2 | 0.1383 | 97.90 | 392 | 13 | 587 | 8 | 0.9800 | 0.9783 |

*Table1. Summary Table of Classifier Performance on the ACW data*

To find the classifier with the greatest performance for the data set provided, there is first a test to decipher which cleaned data set is the best performing. The classifier ‘J48’ is used to test. As seen on Table1, ‘Data\_Clean2’ is shown to have the greatest accuracy, sensitivity and specificity. However the root mean squared error (RMSE) is lower than the data sets ‘Data\_RiskKnown’ and ‘Data\_Clean0’. This is possibly due to 5 duplicated records being removed between data sets ‘Data\_Clean0’ and ‘Data\_Clean1’, yet it is still greater than that of the base data set and consequently ‘Data\_Clean2’ will be used with the other classifiers.

The classifier with the highest RMSE is ‘NaiveBayes’, however it is not the best performing due to lacking accuracy, sensitivity and specificity. There is another classifier (‘NNge’) with the greatest accuracy, sensitivity and specificity but has the lowest RMSE value. The classifier which is possibly the best performing is ‘SMO’ – it has the second highest RMSE, good midrange accuracy, sensitivity and specificity. However, the difference between the classifiers is at best small, leading to a range of classifiers being used as the best possible route.