Nicholas Louis Brown

Professor Wallace

IST 707

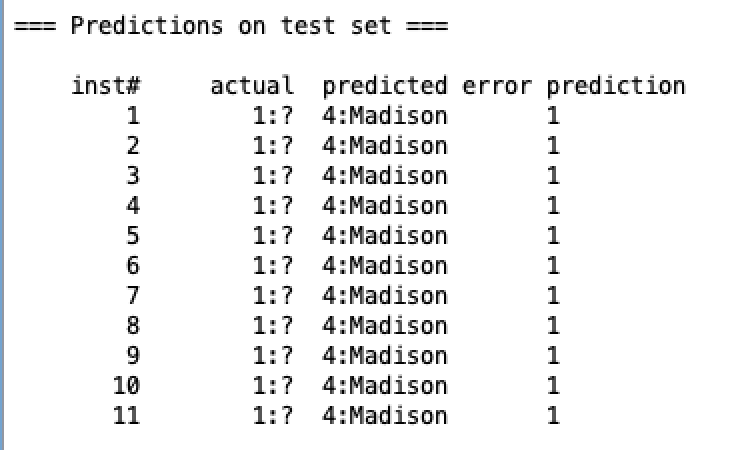
October 2, 2018

Author Attribution Using J48 Decision Tree

**Executive Summary:**

The federalist papers are a series of papers with a disputed history: no one is sure whether Alexander Hamilton, James Madison, or John Jay authored the documents. Using clustering techniques last week, we concluded that Madison had authored the mystery documents. This report focuses on using a different technique, the J48 decision tree model, to predict the authors of the 11 mystery documents. The process for doing this type of analysis required three separate steps: first I had to split the data into two sets, one containing the known works (our training set) and the other containing unknown works (what we are going to use the model to test). After the data was split I created multiple models with varying settings using the training set. For my final model I decided on the J48 decision tree using 3-fold cross validation. The accuracy and error of the model are reported in the table below. The final step in solving this mystery was to actually apply our mystery documents to the model we made. This involved testing our model using the mystery data we separated in step one. The final conclusion made by the decision tree (shown in the screenshot below) was that James Madison wrote all 11 of the disputed papers. This is the same conclusion that our clustering analysis presented.

**J48 Prediction Results:**



**J48 Model Accuracy:**

|  |  |  |
| --- | --- | --- |
| Correctly Classified Instances | 63 | 85.1351 % |
| Incorrectly Classified Instances | 11 | 14.8649 % |

**Methodology:**

**Preprocessing:**

Preprocessing of the data was done by hand using a text editing application. First, the 11 rows of data without a known author was removed and placed into a separate csv document. The remaining data was kept in a csv file labeled ‘training’. To ensure the algorithm worked with all the data I removed the file name variable from both sets of data as this was not useful for predicting who wrote what.

**Building the Model– J48 Algorithm:**

The algorithm used to build the decision tree was the J48 algorithm. The parameters for the model involved settings the confidenceFactor to 0.25 and the minimum number of instances to 2. Much of the model tuning occurred by modifying either the cross-validation settings or the percent-split settings. For the first few models I tried using a percent split ranging from 60% to 80%, while the accuracy of the models was high I concluded that they could not be trusted because the size of the dataset was too small. To better test the accuracy of the model I chose to do k-folds validation. I settled on a 3-fold validation mostly because this would leave sizable chunks left to both test and train the model.

**Testing the data:**

Once our model was built we used the Weka feature which allowed us to test the model with a new dataset. For our new data we chose the 11 mystery author sets intending for each one to be assigned to its best matching author. The results of the test are in the screenshot above.

**Conclusion:**

While the J48 decision tree model was able to predict the authors with a computed confidence unlike the k-means clustering algorithm, it is important to distinguish between the strengths and weaknesses of the two approaches. The k-means algorithm gave us a good idea and visualization of the authors where our decision tree we can follow and backup with numbers. I would hesitate to attribute authorship to Madison had both the clustering and decision tree models not reached the same conclusion. The decision tree is another tool in the data scientist toolkit which can help to reinforce or reject our hypotheses.