(Support, c

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

onfidence) online offline accuracy

# Lab 3: Classification

# Introduction

The purpose of this lab was to perform Association Rule mining on the Reuters news articles dataset seen in previous labs. After the rules were created, they were to be used to classify the dataset. The results for efficiency, and efficacy were compared to those seen in Lab 3 with the K Nearest Neighbors, Naïve Bayesian, and Decision Tree classifiers. Additionally, a discussion on the effects of input parameters for Association Rule mining will be discussed.

# Preprocessing

There were only minor surface level changes in preprocessing between this lab and the previous one on classifying. The point of this lab was to compare the two processes, so the feature vectors used were kept very similar. They formatting of the actual vectors changed in order to use apriori

# Classification

Offline Training Time

The average time in milliseconds it took to fit the classifier to one tuple of data. The total time for the fitting was divided by the number of tuples in the training set.

Online Training Time

The average time in milliseconds it took the classifier to predict the label of one tuple of data. The total time for predicting all of the testing labels was divided by the total number of tuples in the testing set.

Accuracy

# Comparing Results

As a reminder of the last lab, and for purposes of comparison I will reintroduce the results from the KNN, Naïve Bayesian, and Decision Tree classifiers. Since the best training split for each of these was unanimously 66/33, I used the same for the Association Rule classification. The comparison can be seen in figure 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Offline Time per tuple (ms) | Online Time per tuple (ms) | Accuracy |
| KNN | 0.096 | 9.963 | 0.7123 |
| Bayesian | 0.0169 | 0.6918 | 0.6145 |
| D-Tree | 0.2062 | 0.0031 | 0.7041 |

**Figure 1**: KNN, Naïve Bayesian, Decision Tree, and Association Rule classifier results on full Reuters dataset with a training split of 66/33.

Training split

For each of the three algorithms, the training split seemed to have small effect on the accuracy of the classification. There was the consistent trend of 66/33 being the best split as far as accuracy was concerned, where going above or below that would decrease the accuracy slightly. Accuracy and Precision

Precision and accuracy measures were almost the same for both K Nearest Neighbors and Decision Tree classifiers in every split. For Naïve Bayesian the split was slightly more significant, with the precision being higher than accuracy by about 4% in every case. K Nearest Neighbors and Decision Tree classifiers produced the highest accuracy, with around 71% each, while Bayes only correctly classified about 61% correctly.

# Conclusion

From the results above, there are a couple of key takeaways. The most scalable algorithm was the Decision Tree classifier, which once fit to data can predict labels for data in very little time due to its logarithmic time complexity based tree structure. The most accurate classifier was the K Nearest Neighbors, which produced a maximum accuracy of 71.23%. This, however, was the least scalable of the three algorithms tested, and also only marginally outperformed the Decision Tree classifier. Therefore, for this dataset, and others like it, Decision Tree classification seems to be the overall best of the three algorithms.