Sydney Strzempko
Empirical/Programming Assignment 1
Prof. Roni Khardon
COMP 135 Tufts University

Report on Results

Generated Results

Note: Also refer to graph folder in generated code

Chart A

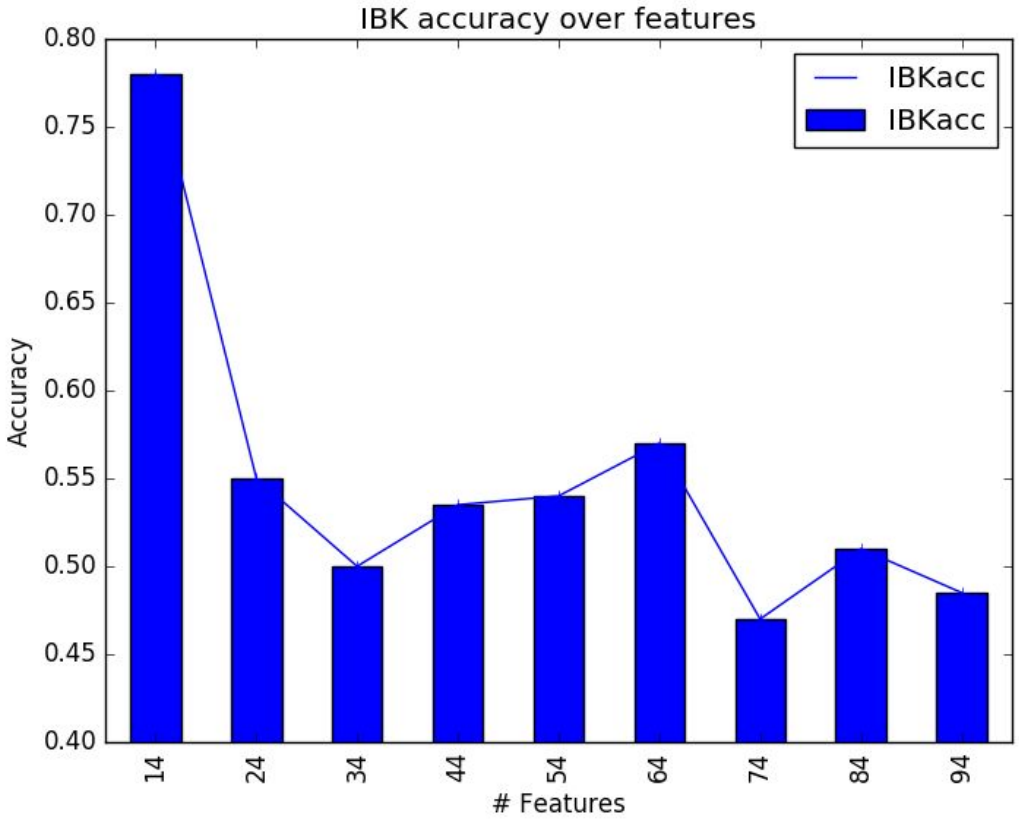


Chart B,C below

Chart B

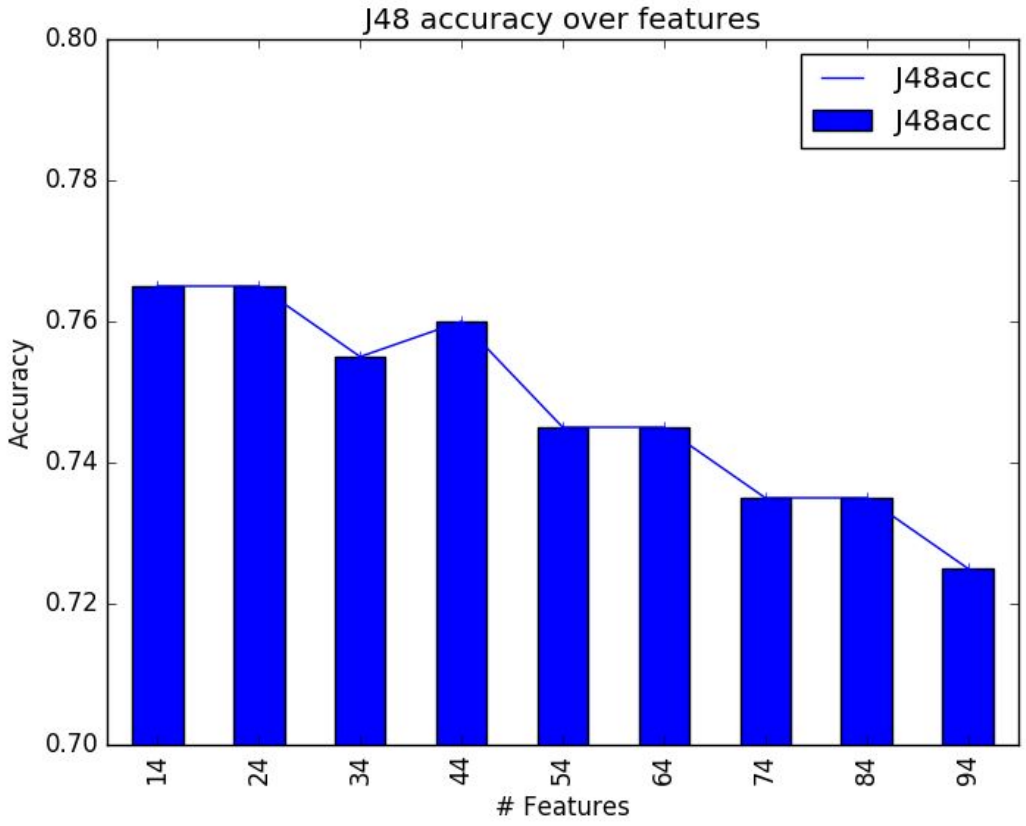
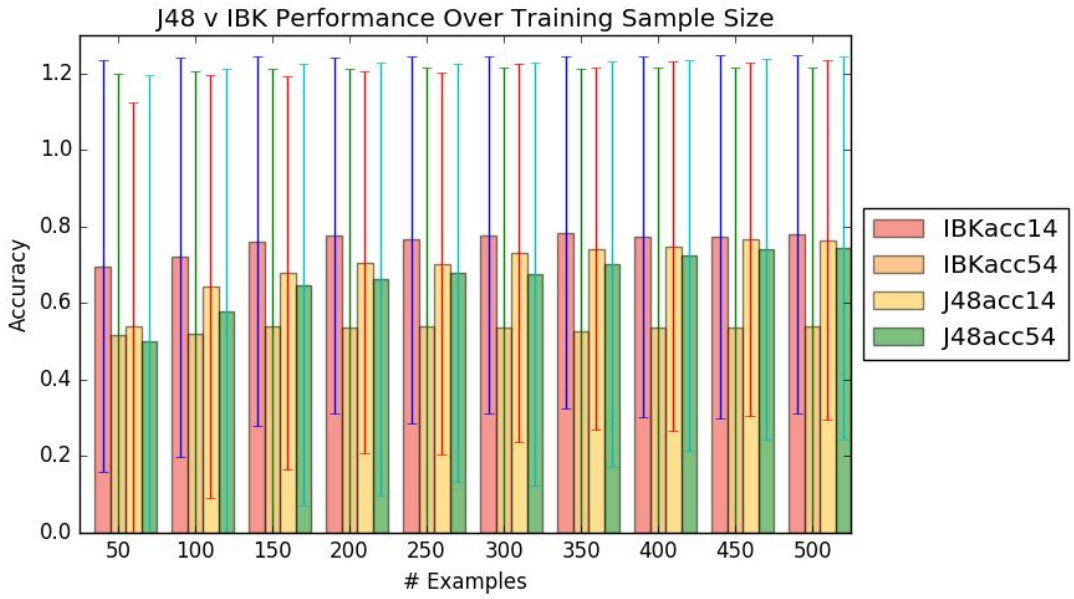


Chart C



Discussion & Observation

In experimenting with machine learning algorithms, performance can be measured over a variety of criterion, but even more so, over a variety of differences in the sets provided and the quantities of documents in those sets. The two algorithms examined in this assignment were provided by the weka Machine Learning library; IBk and J48. In examining performance, we focused on two variations in sets provided: first, the number of features present in each document in the train and corresponding test sets, and then, the number of examples provided in each document in the train set.

As both Chart A and Chart B show, accuracy in classifying test sets is generally easier with a smaller number of features in each document. This makes sense especially in the J48 learning tree algorithm case, which under the covers of the weka implementation allows for a smaller tree to be built, reducing the chance of overfitting on the test set. In the case of the IBk algorithm, the thread is less clear; although the 14-feature sets consistently overperformed in multiple trials (and in the final chart above), the thread is lost a little in accuracy over an increasing number of features. The wide range in accuracy could indicate that the IBk algorithm provided by weka is incredibly sensitive to the number of features provided in the train and test sets. Overall, for 14 or less features, the IBk algorithm would be preferable in machine learning, but for any more features, the J48 algorithm outperforms the IBk.

As Chart C shows, accuracy in classifying a fixed test set in two fixed feature categories (14- and 54-) increases with a greater number of documents in the examples provided. As expected based on our previous observations, the IBk algorithm outperforms the J48 algorithm over the 14-feature sets, but the J48 over 14- and 54- feature sets consistently outperform the IBk over the 54-feature sets. Accuracy overall from a range of 50-500 documents in each train set improves. This is expected as allowing the weka classifiers a greater volume of train documents before performing classification on a test set would allow the algorithm to be better-tuned as it runs with more iterations in the training phase than another instance of the algorithm. Although overfitting could be a concern, the reality of the small range of train documents provided gets rid of that concern and lends itself to the gradual increase in accuracy over greater number of training examples over all algorithms across both features. However, compared to the wide range in accuracy generated by differing feature sizes above, I would argue that ultimately the size of the train set, if bound within a reasonable range, does not impact IBk or J48 algorithm sensitivity nearly so much as differing feature size with fixed test sets.