Accuracy Report: Group 12

Background:

Random Forests within Machine Learning are currently the most accurate algorithm yet to be developed ([berkeley.edu](http://berkeley.edu)). However, there may still be room for improvement. Although the best way to improve accuracy would be to use more data, this is not always available and may decrease efficiency. The solution to this would be to use a given data set to the most extent possible. Currently, approximately 1/3 of data is going unused during Random Forest algorithm development. This is known as out-of-bag incidences and a reduction of these would lead to greater accuracy (Robnik-Sikonja).

Our Proposed Solution:

We propose the following:

Instead of simply randomly choosing small data sets from the larger group of data in order to develop the multiple trees, we divide the data in a more organized yet still mostly random fashion in an attempt to utilize *all* possible data.

The Details:

First, we set aside a random set of data for later testing in order to verify that our algorithm is not overly-optimistic. Then, the remaining set of data, rather than being randomly split into smaller sets, is split in a systematic way in order to utilize data as much as possible.

In order to do this, the desired number of trees would be defined, and it would be determined how many data sets could be defined between said number of trees:

*Number of data sets/number of desired trees = number of data sets per tree*

Then, each tree would be randomly assigned the calculated number of data sets necessary. When a data set was assigned to a tree, it would then not be able to be utilized by any other tree. This process would continue until all data sets were assigned to a tree, and thus, *all* data would be used, cutting out the out-of-bag incidences.

In the incidence in which the number of data sets divided by the number of trees would not produce a whole integer number, each data set could be assigned a certain number of times (to a few trees instead of just one) until each set had been assigned an equal number of times and each tree was defined by an equal number of data sets.

Conclusions:

* + - * Although this method does decrease the independency of each tree, this slight decrease in the random nature of the sets should not effect the outcomes due to the use of the mode method in determining the final outcome from the aggregated trees.
      * The benefits and theoretical increase in accuracy due to this method would outweigh any independency lost.
      * Out-of-bag incidences would be removed, thus increasing accuracy.
      * The use of all data would allow for deeper analyzation and better machine learning without a need for significantly larger data sets.

Citations:

Boulicaut, Jean-François. Machine learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004: proceedings. Berlin: Springer-Verlag, 2004. Print.

Breiman, Leo; Cutler, Adele. “Random Forests.” <https://www.stat.berkeley.edu/~breiman/RandomForests>