John Hancock

Advanced Data Mining and Machine Learning CAP 6778

September $26^{th}$, 2014

# Review of Random Forest: A Reliable Tool For Patient Response Prediction

The two main ideas in, "Random Forest: A Reliable Tool For Patient Response Prediction" are:

- show how Random Forest works well regardless of the feature selection technique we use
- As an ensemble learner it can outperform single learners

Random forest is a classifier, it is an ensemble of decision trees.

In [1] we are using random forest for patient response prediction. The data in [1] is bioinformatic data for cancer patients' response to a drug to treat their cancer. The response is patient's response to a treatment. In this case the treatment is the drug bortezomib. Bortezomib is a cancer treatment. Sometimes the treatment can kill the patient, so it would be good to be able to predict the patients response to the treatment before administering it.

Responses can be:

- positive: complete response, partial response, and minimal response; the patient has a full recovery or shows some signs of getting better
- negative: no change or progressive disease; the patient stays the same or continues to get worse.

The dataset is bioinformatic information, so we have a large number of attributes - over 22,000.

The results will show that with random forest we do not have to worry about feature selection. [1] explains random forest builds *numTrees* decision trees with *numFeatures* attributes. The attributes a tree will use are randomly chosen. We use the class chosen by the largest number of trees as classifier selection. [1]says the optimum number of trees is 100, but we think we would have to use an odd number of trees for binary classification in order to avoid a tie vote.

The authors of [1]used the dataset because it is focused - it covers patients with the same disease and the same treatment applied to the disease, and it has an interesting scale, 169 patients and over 22,000 attributes for each instance (patient).

Experiment details: 7 classifiers: random forest (RFT), 5 Nearest Neighbors (5NN), naive bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), and Multi-Layer Perceptron MPP). 19 feature rankers are used: 7 common, 11 threshold-based feature selection techniques (TBFS), one unusual signal to noise (SNR), and 14 different feature subset sizes from 5 to 1000.

The results show random forest can work well with no matter what feature selection technique we use.

Classification is performed using 5 fold cross validation.

[1] does analysis on two classification schemes: response vs. no response (R vs NR), and response vs. progressive disease (R vs PD).

The authors of [1]tested 63,840 models.

The paper has tables of top results:

- R vs PD with TBFS filters: 18 rankers, 6 learners classifiers random forest does better than average of 5 other learners for all the different kinds of filters, RFT always does better than other 5 classifiers

- R vs NR with TBFS filters: same result, RFT always does better

- R vs. PD with top 1000 features: same result, RFT always does better
- R vs. NR with top 1000 features: same result, RFT always does better.

[1] then does statistical tests on results to find significance of experimental factors for results: z-test, anova, Tukey HSD.

The tests show:

- 1000 features not significantly different from using top number of features
- choice of ranker marginal not important for random forest results
- RFT significantly better in hsd r vs pd for best number of features, hsd for 1000 features
- I see what looks like some overlap for RFT, MLP, NB for HSD with R vs. NR for best number of features or top 1000 features.

The conclusions we can draw are: random forest is best classifier for patient response prediction for the Mulligan et. al data set, ranker is not important, no difference for using top 1000 features vs. best number of features.

Future work: random forest requires more data, but patient response data is not public domain

Note: paper has a typo - mb should be nb for naive bayes in feature rankers.

Finally, remember random forest is an ensemble learner, this is why it does better than single classifiers.

# References

[1] T. M. Khoshgoftaar, D. Dittman, R. Wald, A.Napolitano, "Random Forest: A Reliable Tool For Patient Response Prediction." Florida Atlantic University, Bocat Raton FL 33431.