Predicting Gene Function Using SVMs and Bayesian Networks

Laura Miron, Benoit Pit-Claudel
{lmiron, bpithia}@stanford.edu

Before running the SVM on all go nodes, we perform a parameter search on the most represented GO node (GO:0045944, positive regulation of transcription by RNA polymerase II).

Below, graphs comparing kernels (linear and rbf with different y values) and penalty parameters C of the error term.

• Very good accuracy to be expected since the proportion of positive examples for each classifier is very low
• Most important metric should be recall: a false positive should be compensated by the Bayesian network, whereas the false negative could have more impact
• Average accuracy for all nodes over 97.7%

Due to bugs in the library pgmpy, we are currently unable to make inferences on the full net of proteins obtained through [5]. Microarray data has one or more missing columns for each example, which we complete using KNN, and 161 float features

Labeling the examples total, mix of floats and Booleans.

Methods

• In our final classifier, we train one 10-Ensemble SVM per gene ontology node, using a linear kernel and C = 1.0
• One challenge in gene prediction is the small number of examples overall, and in particular the small number of positive examples for each GO node; we therefore use bootstrapped samples with replacement to train the classifiers

Conclusions

• Improved accuracy on individual SVM classifiers compared to Barutcuoglu et al. [1]
• Where Barutcuoglu et al had better results with C=∞ and using the rbf kernel, we obtained better results with a linear kernel.
• Refinements in GO classification between 2006 and today might explain better results.
• Their bootstrapping procedure is effective for dealing with very few positive examples and possibly uncertain examples
• Useful for newly discovered species where little data is available

References


Figure 1. Saccharomyces cerevisiae [yeast] (2)

Figure 2. Results of parameter search on linear kernel and GO:0045944

Figure 3. Hierarchical relationship between GO nodes, Barutcuoglu et al.

Figure 4. Individual svm accuracy vs. bayes net accuracy for selected GO nodes

Figure 5. Bayes net structure [Barutcuoglu]