Writeup for Challange 2

CIHC Team

September 22, 2010

Preparation

In order to have a balanced representation of the data among our training set, we used the following normalization procedure:
For each TF:

1. Take logarithm of the signal mean (call it log signal mean).
2. Pick 2000 random numbers from the uniform distribution between min. and max. log signal mean.
3. For each of these numbers, take the sequence with the closest log signal mean.
4. Erase duplicates.

This gave us normalized training sets of sizes between 807 (for TF 53) and 1685 (for TF 6).

The training sets for each TF where then divided randomly into to equally large sets – a training set for our classifiers a test set. The blender was trained on this test set.

Classifiers

n-gram spectra

We performed 4-gram and 5-gram spectral analysis using the Random Forest [1] algorithm. Attribute selection was made according to the Boruta [3] algorithm. $4^5 = 1024$ attributes while operating on approx. 1000 training examples was however too many to use Boruta. In order to avoid this problem, we considered just 128 5-grams whose importance estimate by the random forest was highest. Boruta did not reject approx. 50 4-grams and 50 5-grams, which where then used to train another forests.

All four forests, for each TF, where used by the blender.

RankMotif++

We used RankMotif++ [2] to find motifs of length 8 in our training set. For each TF we run 4 initializations and chose result with best likelihood. MAST (from The MEME Suite) was used to find motifs in sequences from test set. As a classifier for blender we took smallest p-value of finding the motif in each sequence.

Blending

Random Forest algorithm was used as a blender, with 1000 voting trees.

References

[1] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[2] Xiaoyu Chen, Timothy R Hughes, and Quaid Morris. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics (Oxford, England)*, 23(13):i72–9, July 2007.

[3] Miron B Kursa and Witold R Rudnicki. Feature Selection with the Boruta Package. *Journal Of Statistical Software*, 36(11), 2010.