ALFREDO BAIONE
mat. s279328

Protein functions classification report

From the Uniprot database 1771 proteins were selected, all of them with the following characteristics:

- popular organism: human;
- protein existence: protein level;
- annotation score: 5.

Functions were chosen according to two main criteria: parsimony and interpretability. Thus, the following protein functions were enrolled:

- “Absorption": ability to break down other proteins;
- "Catalytic activity": ability to increase the rate at which chemical reactions occur;
- "Cofactor": assisting enzymatic catalysis and preserving the structural integrity of proteins;
- "DNA binding": control of transcription and translation, DNA repair, splicing, apoptosis and mediating stress responses;
- "Activity regulation": regulation of enzyme activity;
- "Kinetics": the ability to associate with other proteins;
- "Pathway": the presence inside a large network of chemical reactions.

1. Preprocessing

First of all, every observation with NA entries for every functional variable was removed from the dataset as well as every functional variable with NA entries for every observation. It was also checked that every protein matched only one functional variable, in order to address a single-label classification task. Unfortunately, there were only two observations for the functional variable “Kinetics” and so they were also removed from the dataset with their corresponding variable.
125 observations and 8 variables remained, 5 of which related to protein functions ("Catalytic activity", "Cofactor", "DNA binding", "Activity regulation" and "Pathway").
Finally, a 5-level (one for each protein function in the dataset) factorial variable to be predicted, named “Function”, was created and the variable “Sequence” has been transformed into a numerical one, due to computational reasons.

2. Processing

In the early phase of processing, the dataset was divided into training and test set, possibly preserving the fraction of data (0.7) assigned to every class of the “Function” variable.
Every training session was conducted using a leave-one-out cross validation.
4 models were trained: a naive bayes classifier, a decision tree, a random forest and a neural network.

After having predicted the five possible outcomes with these classifiers, accuracies, macro-averaged recalls and macro-averaged precisions were computed and compared one to each other for every prediction obtained.
In the end, confusion matrices of the four models were plotted, in order to visualize and better estimate the final results.

3. Comments

Protein functions classification is, without any doubt, a very difficult task. In this particular case, probably, the main issues encountered were three:

i. the fact that after the preprocessing phase only 125 observations remained;
ii. the lack of specific information regarding the functional variables chosen;
iii. the imbalance in the number of observations for each "Function" level.

Actually, more than 4 models were trained and tested, but in the end they all seemed to produce , more or less, the same results. The highest level of accuracy (0.54) was obtained with the random forest model, that was also the only one capable of predicting correctly at least one outcome of two very small classes (i.e. "Cofactor" and "DNA binding").