

Programming3 big data

Supervised machine learning to predict protein function on InterPROscan datasets



xth of july

Daan Steur

Hanze university of applied sciences

# Notice: Assignment not finished, only a base. Will deliver the actual code and report after summer.

# Introduction

In the world of microbiology, exploration of new species and the functions of those species can be rather time consuming. Let say a new species is discovered and you want to know the specific protein functions inside that species. To get to that answer, the bacteria need to be grown, induced, cells need to be extracted, protein needs to be extracted and purified, that protein eventually needs to be sequenced and you in the end you align its DNA against tools like BLAST, to see if any useful results pop up. How wonderfully refreshing it would be to simply this by a large number of steps. You grow the bacteria, you sequence the bacteria, you run a model over that data and a list potential active proteins and the functions of those proteins gets printed onto a csv. The problem however is not the lack of data, the problem is the lack of accurate predictive models. In this report an attempt is made use supervised learning techniques to generate an accurate model to predict the protein function based on feature characteristics.

# Materials & method

The creation of the proposed supervised models comes in 4 steps, the loading in of the InterPROscan data in an efficient matter, the pre-processing of that data, the training of the models, and the evaluation of the model

#### Loading the data

For the loading of the data the spark module was utilized

#### Pre-processing of data

The data has a couple criteria to obey by for starters the interpro number should cover more than 90% of the protein sequence and it needs to cover the largest length of the sequence.

#### Training the models

Three models were tested including random forest, logistic regression and Naïve bayes, these models were choices for there ability to work best on big data and this is a multi-classification problem, especially Naive Bayes classifier performs better when the assumption of independence is true than other models, such as logistic regression, and requires fewer training data. That is way logistic regression functions as a sort of baseline model. On the opposite side random forest has a more complicated calculation process and will require longer, but will it outperform Naïve bayes.

#### Evaluation of the models

The models are evaluated based on the accuracy of the prediction and the processing time.

# Results

After trying several models, including random forest, logistic regression and Naïve bayes,

Logistic regression :

Naïve bayes :

Random forest :

(Some plot displaying the difference)

# Conclusion

Model x seems to perform best, with an accuracy of x, that is x% better than model y and also x% better than model z

In conclusion supervised learning can be used to predict the protein function, but to a certain extent, long processing time, in-efficient code and other such factors can have influence on the process in its whole. Future improvements are to try an different set of models, to find the optimum model for the job. Furthermore a closer looks most be given to the processing frame of the data, yes spark is very useful, but it has its limitation, but given an extension into a larger scale, there might be better alternatives for data processing. But nevertheless shown here is that given some time, a good prediction can be made as to what a protein function is based on relatively simple characteristics. ultimately paving the road to faster function diagnostics.

**Assignment**

Keep the report short 2 pages max

Short description of the project and what you do

How you processed the data

Give results

Conclusions.

But all projects can be done in dask or spark.

**4. Project 1**

**4.1. Introduction**

The first project you can choose is a continuation of the InterPROscan dataset you worked on for Assignment 5. You now know the structure and type of data that is involved. What I would ike you to do is to develop scikit-learn machine learning models to predict the function of the proteins in the dataset. I'm not going to mandate any particular model, but you should use your knowledge from DS3 to determine good candidate methods and try a couple, and report on your Given the size of the data, you should train your machine models in parallel! Dask is an obvious candidate for this, but you can also use any of the other frameworks we discussed in the course.

The "function" of the protein which is the "class" your model should predict is defined as the InterPRO number which covers:

1. >90% of the protein's sequence
2. Covers the largest length of the sequence

The "features" that you use to predict that class are the *other*, smaller, InterPRO annotations. Take care to remove "noise" from the dataset, ie. lines from the TSV file which *don't* contain InterPRO numbers, proteins which *don't* have a large feature (according to the criteria above) etc. Be sure to note in your report how you processed the data and why you made the choices you did!

Your script should, given one or more of these TSV files with InterPROscan annotations, produce models (Pickle them to save them!) which predict protein function, and a file of training-data on which you trained them.