Random subsampling techniques for sea bass mortality prediction

Giovanni Gaio, Simone Moretti

June 6, 2025

1 Introduction

The analysis of genomes to predict the effects of disease. But given the size of genomes it is not always easy to understand the genes that have an phenotype related to the disease effects.

2 Methodology

In this work we tried to evaluate the effectiveness of random subsampling of genes in improving predictions of mortality. We always subsampled selecting a subset of genes in the whole dataset.

The idea is to help our estimator to not get lost in the $\sim 10^6$ genes, but allowing it to work with only a smaller number of genes at a time. Hopefully in this way our model will give better results.

Once we subsampled the dataset, we used the `XGBoost`[1] library to construct a predictor in a fast and easy manner.

Uniform subsampling. The first and simplest thing we tried was to randomly and uniformly subsample the genes on the entire genome. We selected a given fraction of the genes, keeping or discarding each gene with fixed and uniform probability.

Uniform subsampling on chromosomes. A second possibility is to uniformly sample a fixed fraction of genes on each chromosome.

Subsampling using annotations. Some genes have been linked to specific organs, tissues or functions. This information can be used to select and sample genes

3 Results

4 Conclusions

References

[1] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.