Università degli Studi di Padova

Random subsampling techniques for sea bass mortality prediction

Giovanni Gaio, Simone Moretti

July 21, 2025

Overview

- Motivation: Identifying impactful SNPs in sea bass mortality
- Dataset: Genomic SNP data, mortality outcomes, and annotations
- Method: Subsampling techniques with XGBoost
- Results: Accuracy/F1 vs. subsampling rate
- Conclusion: Subsampling preserves predictive power

SNPs and Sea Bass Mortality

- VNN is a widespread lethal disease in sea life.
- SNPs may increase or decrease the chance of death.
- Goal: Predict mortality based on SNP profiles.

Challenges with Genomic Data

- Each fish: over 6 million SNP positions.
- Sample size: only 990 sea bass individuals.
- Traditional models overfit due to data dimensionality.

Machine Learning Approach

- Use XGBoost classifier for mortality prediction.
- Need to reduce feature space: apply subsampling.
- Evaluate performance on subsampled datasets.

Pipeline Overview

SNP Dataset Structure

- 990 rows (fish), each with 6,072,853 SNP features.
- SNP values: 0 (no mutation), 1 (heterozygous), 2 (homozygous alt).
- Each fish is paired with a mortality label.

Annotation Metadata

- Annotations include function: Promoter, Enhancer, Open Chromatin.
- Tissue number (0–25) indicates location-specific relevance.

Uniform Subsampling

- Randomly sample a fixed proportion p of all SNPs.
- Simple but may cause imbalance across chromosomes.

Per-Chromosome Subsampling

- Ensures balanced representation from each chromosome.
- Randomly sample same number of SNPs per chromosome.

Annotation-Based Subsampling

- Filter SNPs by biological annotation.
- Then apply uniform subsampling to relevant regions.

Subsampling Strategy Comparison

- Trade-offs in simplicity, biological interpretability, and balance.
- Aim: maximize predictive power while reducing dimensionality.

XGBoost Training

- Training/testing split is fixed before subsampling.
- Trained multiple times per method/rate to average performance.

Control of Randomness

- XGBoost random seed fixed.
- Subsampling is the only random step.

Subsampling Ratios

- Subsampled with multiple p values (e.g. 0.01, 0.05, 0.1, 0.2).
- Trained model for each combination.

Results: Genome-wide Subsampling

- Accuracy and F1 show mostly flat trend across p values.
- Slight performance drop at very low rates ($p < 0.01$).

Results: Annotated Regions

- Scores consistent across function (e.g., Promoter vs Enhancer).
- Subsampling by tissue shows minor fluctuations.

Interpretation of Results

- Model performance largely independent of subsampling rate.
- Low p reduces predictive quality — not surprising.
- Subsampling improves speed without sacrificing accuracy.

Conclusions

- Random SNP subsampling retains model effectiveness.
- No strong trend between rate and accuracy (outside extremes).
- Enables faster, scalable experimentation for genomic prediction.

Thank You / Questions?

Questions?