Machine Learning methods for genome enabled prediction of complex traits: benchmarking and robustness to marker elimination.

J. Elenter[1], G. Etchebarne[1], I. Hounie[1], F. Lecumberry[1], M.I. Fariello[1]

1. Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

Introduction

Genome enabled phenotype prediction consists of predicting an individual's physical characteristics from its genotype and environment. Accurate trait prediction plays an important role in fields such as medicine, agriculture and animal breeding. Moreover, sequencing techniques are as powerful and cheap as ever. Still, developing high-performing, general predictors remains an unsolved task. Population structure, complexity of traits, number of samples and SNPs all present significant differences between datasets, thus affecting model performance.

In this study, we evaluate and compare the performance of four different algorithms on plant and animal datasets. These algorithms are **Ridge regression**, **Gradient Boosting Machines** (GBM), **Random Forests** (RF), and **Support Vector Machines** (SVM). We outperform the state of the art results in each dataset through an exhaustive randomized hyperparameter search. This result shows the importance of proper hyperparameter optimization.

Furthermore, we assess the impact of marker encoding by comparing **additive** and **one-hot** approaches. In addition, we evaluate the importance of marker density by **eliminating random marker subsets** from the genotype matrix. We conclude that **all models present a negligible loss in performance** until a very high portion (~95%) of them are missing.

Data description

We evaluate the models in two plant and two animal datasets: yeast yield, wheat yield, Holstein cattle milk yield, and German bulls sire conception rate.

The **yeast** dataset contains 1008 individuals with 11,623 binary markers each. Each individual was measured in 46 different environments. Yeast is an haploid organism.

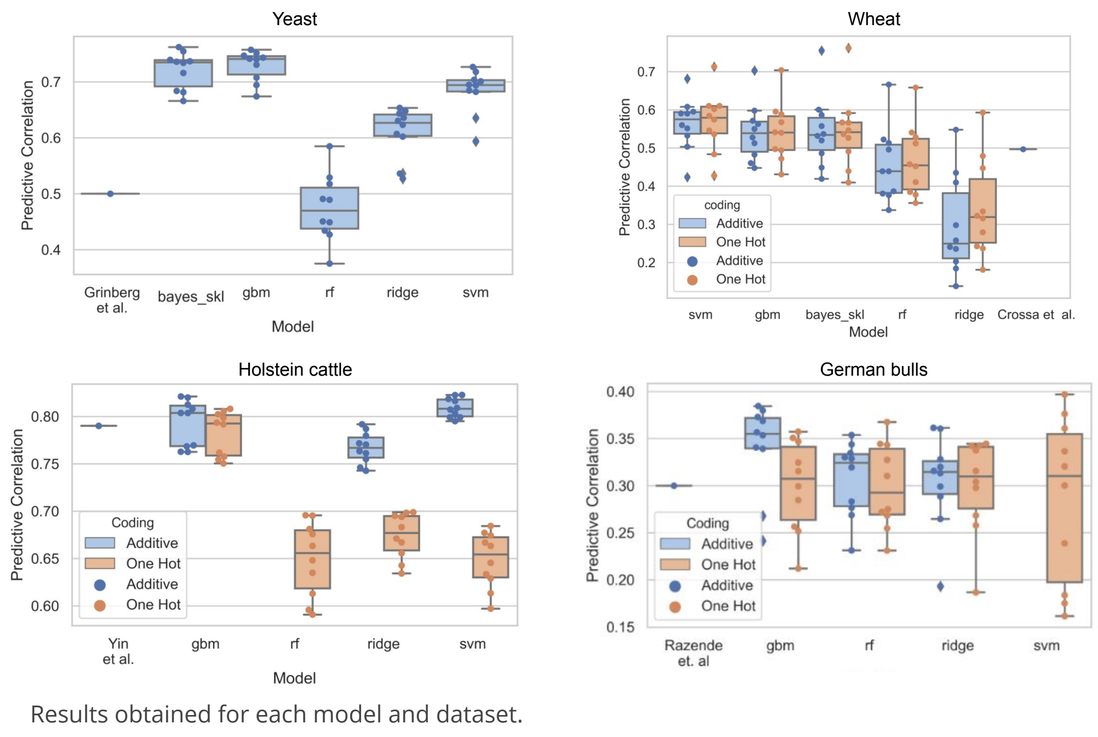
The **wheat** dataset is the smallest one with 599 individuals and 1,447 markers and 4 different environments. Markers with an allele frequency lower than 0.05 or greater than 0.95 were removed, which resulted in 1279 total markers.

The **Holstein cattle milk yield** dataset consists of 5,024 bulls with 42,551 markers, measured in a single environment.

Lastly, the **German bulls sire conception rate** dataset is made of 1,569 bulls, featuring 107,371 markers each. Markers that mapped to sex chromosomes, had minor allelic frequencies below 1%, or had a call rate of less than 90% were removed, resulting in a total of 95,434 markers.

Results

The results are showcased in the charts below. We only show one environment per dataset for clarity purposes. It is important to note that the best performing model varies depending on the environment in which the phenotype was measured. For each dataset, we compare with the best results available in the literature [1, 2, 3, 4]. Each point represents a different train/test split.



Results obtained for each model and dataset.

We consistently outperform previous state of the art results via an exhaustive, randomized random hyperparameter search. Even simple models such as Ridge regression achieve competitive performance when optimized properly.

Notice the high variability between different splits and the presence of major outliers, particularly in the German bulls dataset. This phenomena is mainly caused by the incredibly low sample to dimensionality ratio. In the German bulls dataset, this ratio is 1569 / 95,434 = 0.016.

Acknowledgements

This work was partially funded by project ANII FSDA 1-2018-1-154364. The experiments presented in this work were carried out using ClusterUy (site: https://cluster.uy)..

References

[1] Grinberg, Nastasiya F et al. "An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat." Machine learning vol. 109,2 (2020): 251-277.

[2] F. M. Rezende, J. P. Nani, and F. Pe´nagaricano, "Genomic prediction of bull fertility in us jersey dairy cattle," Journal of dairy science, vol. 102, no. 4, 2019.

[3] J. Crossa, P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Cer´on-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li et al., "Genomic prediction in cimmyt maize and wheat breeding programs," Heredity, vol. 112, no. 1, pp. 48–60, 2014.

[4]Z. Zhang, M. Erbe, J. He, U. Ober, N. Gao, H. Zhang, H. Simianer, and J. Li, "Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix," G3: Genes, Genomes, Genetics, vol. 5, no. 4.

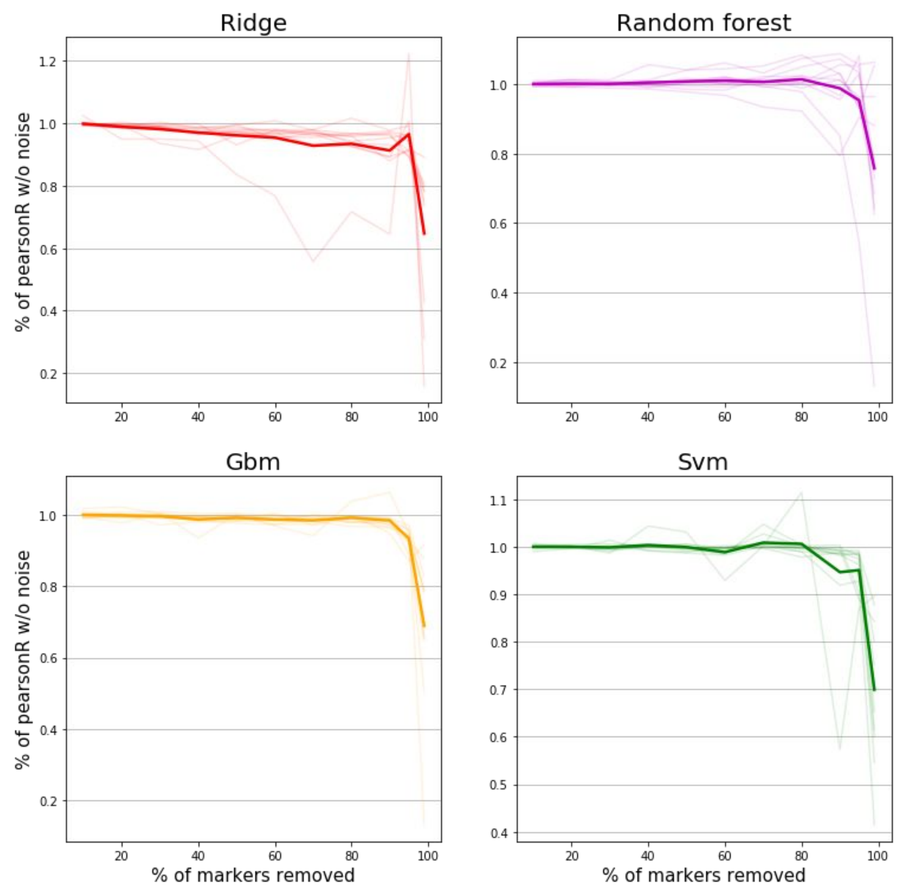
4,

Additive versus One-hot encoding

Additive encoding achieved consistently better performance in all three diploid datasets. This is an interesting result since markers are categorical variables in nature. This performance drop may be caused by the duplication in dimensionality as well as the loss of the one-to-one correspondence between variables and biological markers. Moreover, we hypothesize that the **linkage disequilibrium** phenomenon further impacts the model's ability to recognize important markers when the genotype is one-hot encoded, since there's twice as much highly correlated columns.

Robustness to random marker elimination

This experiment consists of randomly selecting a portion of the genotype's markers and deleting them (same markers for all individuals, with an uniform distribution). Sequencing techniques are as potent as ever, enabling ever growing sampling rates. Although denser sampling may seem beneficial at first glance (more information is better than less information), linkage disequilibrium means many of this markers are redundant. Moreover, the increase in the input's dimensionality should be matched with more sequenced individuals, which is not always the case. This experiment aims to evaluate the extent to which denser sampling impacts the model's performance. The proportion of deleted markers ranged from 10 all the way up to 99 percent.



Results of the missing markers experiment. The experiment was ran for the yeast dataset, testing each model across 10 different splits in 12 of the 46 environments. Each split's relative change in performance (w.r.t. the uncontaminated dataset) is averaged across all environments. The softer lines represents each split's individual results, while the bolder one is the mean across all splits.

The plots above show surprising results: in the yeast dataset, performance remains unaffected up until almost all markers (95 percent and above) are deleted, after which it rapidly declines. This may be explained by two factors: high linkage disequilibrium and low amount of quantitative trait loci (QTL). From a signal processing point of view, the former means that chunks from the genome (input signal) are highly correlated and thus, most markers are redundant. Consequently, removing these markers has little impact on predictive accuracy.

Conclusions and main takeaways

There's no single best model that maintains a consistent performance across all datasets or environments. However, gradient boosting machines and support vector machines tend to produce the most consistent, high-performing results. Model ensembling is a promising alternative that could yield more robust results.

Models in this particular task are very sensitive towards hyperparameter tuning. Regardless of the model, exhaustive hyperparameter searches are worth their high computational cost. The difference between a poorly and a well tuned model can be extremely high. To this end, more sophisticated searching techniques (such as bayesian approaches) may yield even better performance.

Regarding marker encoding, the classical additive approach remains better than its one-hot counterpart. Other techniques such as target encoding, which keeps the best of both worlds (same dimensionality while treating markers as categorical variables), are worth exploring.

Lastly, although sequencing techniques are allowing more markers to be sampled, this increase shows diminishing returns. These results highlight the impact of the number of samples over sample density in this particular task.