

Exploration of machine learning methods for genomic selection in cattle

Auteur : Hervers, Florent

Promoteur(s) : Geurts, Pierre; Druet, Tom

Faculté : Faculté des Sciences appliquées

Diplôme : Master en ingénieur civil en informatique, à finalité spécialisée en "intelligent systems"

Année académique : 2024-2025

URI/URL : <https://github.com/Florent-Hervers/master-thesis-code>; <http://hdl.handle.net/2268.2/23245>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Exploration of machine learning methods for genomic selection in cattle

Florent Hervers

Academic year: 2024 - 2025

Master of Science in Computer Science Engineering

Thesis supervisors: Pierre Geurts and Tom Druet

Genomic selection is a method developed to help breeders select the best parents to have a progeny that exhibits good targeted characteristics. This selection is performed based on the genotype of the animals. The field of genomic selection consists of designing a model that can be trained based on a training population that is able to predict the chosen phenotypes based only on the genotypes.

Machine learning is a field that has become more and more popular during the last decade. These algorithms are used to build complex models based on a large amount of data. These methods have provided very powerful models in a large variety of fields, from image segmentation to the processing of natural language.

As more and more animals are genotyped, the amount of available data becomes big enough to be able to consider machine learning methods. Some architectures trained on cattle datasets were proposed in the literature, but there are still many possible algorithms to be evaluated. In this thesis, we will explore a wide range of machine learning methods from the linear Ridge regression or XGBoost to more complex neural networks, like convolutional neural networks or transformers. The objective is to find models that perform better than the state-of-the-art GBLUP linear model, and perform some experiments on the models to better understand how to apply the machine learning methods for genomic selection.

From all the models we have trained, no model had better performance than the GBLUP models, but some models, the MLP, the ridge regression, and the SVM, reached similar performances. The experiments also show interesting results. The usage of maximum or average pooling layers seems to decrease the performance of the convolutional neural network. Models trained to predict several phenotypes at the same time do reach the same performance as the models trained on a single phenotype, which can reduce the time required to develop a model. Indeed, only a single architecture would have to be tuned instead of one for every phenotype. The impact of the size of the training set was studied. It shows that using more data could increase the performance of the different architectures. Finally, some suggestions are proposed for future research on the usage of machine learning methods for genomic selection.