Feature selection and intelligent livestock management

Alsahaf, Ahmad

DOI:
10.33612/diss.145238079

Summary

Computational animal breeding relies on genetic-statistical models that are aimed at estimating breeding values, which in turn are used to rank animals based on their genetic potential. Modern livestock production systems, however, collect large amounts of data throughout the life of an animal that are not directly suited for those statistical models, such as periodic phenotype and environmental observations. In this thesis, we explore the potential of exploiting that additional data to improve future phenotype prediction in livestock using machine learning methods. To that effect, we provide two examples. In chapter 2, we show that random forest regression outperforms linear models in predicting slaughter age in pigs based on a mix of phenotypic, genetic, and pedigree information. The predictions are made before the start of the fattening phase of the pigs and could therefore be used to assign them to uniform groups based on their forecasted growth rates. In the same application, we additionally demonstrate that machine learning-derived feature importance analysis can give a breakdown of the components responsible for the observed phenotypic performance, thereby providing an alternative to model-based fixed and random effect estimation.

In chapter 3, we use a combination of RGB-D computer vision and supervised learning to estimate the muscularity of live pigs. The objective of that application is to replace the subjective assessments given by a human operator at the farm, and consequently reduce stress on the animals. Results show that the proposed system can accurately mimic the assessment patterns of the human operator. In short, the two proposed examples provide more evidence - in accordance with recent findings in literature - that livestock breeding and management practices could be improved through data-driven modelling.

In chapter 4 of the thesis, we propose a novel wrapper feature selection algorithm based on tree ensembles and boosting. Feature selection is a sub-discipline of

machine learning concerned with discriminating relevant input features from irrelevant and redundant ones. With the increase of dataset dimensions, and a rising interest in model transparency, reliable feature selection becomes increasingly useful. Datasets in the omics fields in particular, like those used in molecular breeding and in personalized medicine, can benefit significantly from feature selection. In those fields of study, datasets often contain large numbers of features and small numbers of samples, which poses a challenge to machine learning methods that require large numbers of samples to learn their parameters, like deep neural networks.

The proposed algorithm, which we call FeatBoost, uses an iterative process of boosting, or sample re-weighting, and model evaluations to select features that are relevant and not redundant to each other. We evaluate the performance of the algorithm against a number of benchmarks, including ReliefF, a filter-based selection method, and two alternative tree ensemble based methods, Boruta and XGBoost-derived feature ranking. FeatBoost outperforms the competing methods on most of the tested datasets.