DOI: 10.1111/jbg.12263

EDITORIAL

WILEY Journal of Animal Breeding and Genetics

Animal Breeding learning from machine learning

Big data are the current buzzword and machine learning (ML) is its oracle. As is well known, ML comprises a wide variety of techniques to identify patterns or to predict characteristics in large data sets. For a good reason, ML is permeating the whole society and industry. One of the most popular recent breakthroughs of ML is a software, named AlphaGo, defeating world professional players at Go game (Silver et al., 2016 *Nature* 529:484–489) or the unsupervised machine recognition of cats from YouTube videos. In a less altruistic vein, ML is used to personalize the commercials you observe in your browser or to automatize reading and translating.

Can Animal Breeding methods be inspired or improved by modern ML technology? ML and breeding share important objectives such as prediction and, unsurprisingly, several works have applied ML algorithms to genomic prediction (e.g., review González-Recio et al. 2014, *Livest Prod Sci* 166:217–231). Further, the data sets in Animal Breeding has traditionally been bigger than in many of its contemporary biological sciences, and developing efficient algorithms have been an important and relevant activity in our field. Animal Breeding does use big data sets and statistical techniques that fall within the ML scope, so Animal Breeding is machine learning, or at least a subset of the machine learning area. However, I want to point out several relevant differences between them, at least conceptually.

First, the modern concept of "big data" refers not only to size but also to heterogeneity. Typical conventional breeding data sets are very simple (phenotypes, pedigree and, now, markers) and easily fit into relational databases, whereas modern biological information is predominantly unstructured. For example, current genomic data deluge comes in the form of numerous disparate, often huge data sets, for example, SNP and gene annotation, metabolic pathways, expression data sets, methylation and protein motifs. In this framework, annotation initiatives in livestock such as the FAANG project (Andersson et al. 2015, *Genom Biol* 16:57) could potentially be of great value to validate the accurateness of many of these data sets in livestock, since current information often relies on model or human species.

Heterogeneous genomic information is starting to be used in several biomedicine areas (http://ieeexplore.ieee.org/document/7347331/), such as the improvement in medical diagnosis in humans, but has only been superficially used in animal breeding. My conjecture is that, if all this information could be used to improve reliability of genetic predictions, it would also help to better understand the biology behind the phenotypes of interest. Traditional breeding methods seem incapable to fully utilize this kind of information, whereas some ML tools, such as deep neural networks, look well suited for the task. Much research remains to be done in this area though. Initially, ML tools could be simply used to assign priors to each SNP (say to give larger weights to non-synonymous and regulatory SNPs at high frequency than to the rest of SNPs) and then utilize these priors in the genomic prediction method of choice. Nevertheless, a more satisfactory solution would be to combine unstructured and structured data for prediction purposes. A limitation for this preferred approach is that some genomic or metabolic data are expensive to collect and are measured in only a few individuals.

The second difference between fields is that Animal Breeding theory is overwhelmingly ruled by the linear model paradigm. In stark contrast, ML comprises a wide diversity of tools: classification and regression trees, random forests, kernel-based methods (support vector machines, kernel ridge regression), neural networks (multilayer perceptron, convolutional neural networks, deep belief networks) and so on. The traditional, sometimes conservative, animal breeder is puzzled by such a baroque catalogue! Some studies (e.g., González-Recio et al., op. cit.) generally show that the performance of the different methods can be relatively similar, whereas hyperparameter values can have a strong impact even when using the same method. The main reason for so many ML tools–it seems to me–is pragmatic. While breeding methods are developed conditional on the model, ML aims at finding the most efficient algorithm/method and little or no care is given to the notion of "model". Even if some ML approaches are based on linear models, many other tools (e.g., trees and neural networks) are basically model-free since what matters at the end of the day is that they predict satisfactorily when the number of variables vastly exceeds that of observations.

As the Animal Breeding theory stems from the infinitesimal, additive paradigm, typical genomic selection methods (Bayes-alphabet like) fit nicely and have a clear genetic interpretation, at least in terms of variance components. In contrast, ML results are typically difficult to interpret in terms of SNP direct effects or of variance components.

This is a serious limitation for some of these tools to be understood, let alone to be adopted by the breeding community. Note, however, that we may be over-interpreting genetic signals even in Bayes-family models, because the effect of the prior never vanishes when $p > n$ (Gianola, 2013, *Genetics* 194:3573–3596). Individual SNP effects and hence variance component estimates are in fact by-products of the chosen prior.

In addition to using ML methods to incorporate unstructured information in genomic prediction, I envisage other topics where ML is promising. One of them is detecting selective sweeps or more generally to infer demographic history of a population. One of the limitations of current methods is the difficulty of finding tests that are uniformly powerful across selective and demographic scenarios. A proposed approach has been to develop tests that combine multiple summary statistics (e.g., Grossman et al. 2010, *Science* 327:883–886). Yet, ML methods such as ensemble methods offer a more appealing and less constrained approach to deal with multitude of predictors that can be individually "weak" (i.e., poor performing) but that are able to produce a good joint predictor (e.g., Schrider & Kern, 2016, *Plos Genet* e1005928). ML concepts can also be adapted to improve upon classical inference methods such as Approximate Bayesian Computation (ABC, Sheehan & Song, 2016, *Plos Comput Biol* e1004845).

But the one player who can benefit most from ML is the animal breeding industry itself. ML offers powerful tools to deal with heterogeneous data that industry regularly or irregularly collects over the years. ML can be used to produce improved predictors of genetic merit, discover unknown relationships between phenotypes or between phenotypes and environmental variables, to monitor competitiveness–to name but a few applications.

To conclude, ML is a collection of diverse technologies that will have an increasing impact on breeding as unstructured genomic and phenotype data expand. ML has its limitations though. Among those are the difficulty of interpreting ML algorithms in genetic terms and the required vast amounts of data to learn well—and these data sets may be currently missing in our field. Traditional, highly parameterized Animal Breeding methods seem more frugal in data requirements. In the review of the Madison International Conference on Quantitative Genetics (Simianer & Sorensen, 2016 *J Anim Breed Genet* 133:249–250), it was quoted Richard Lewontin as having said: "quantitative genetics [. . .] is an attempt to produce knowledge by a systematization of ignorance." Had he known machine learning . . .

ACKNOWLEDGEMENT

Thanks very much to Oscar González-Recio for his insightful comments.

M. Pérez-Enciso
ICREA – Centre for Research in Agricultural Genomics, Barcelona, Spain
Email: miguel.perez@uab.es