Assigning pigs to uniform target weight groups using machine learning

A. Alsahaf[1], G. Azzopardi[1], B. Ducro[2], E. Hanenberg[3], R. F. Veerkamp[2], N. Petkov[1]

[1] University of Groningen, Johann Bernoulli Institute of Mathematics and Computer Science, P.O. Box 407, 9700 AK Groningen, The Netherlands
[2] Wageningen University & Research, P.O. Box 338, 6708PB Wageningen, The Netherlands
[3] Topigs Norsvin Research Center, Beuningen, The Netherlands

Summary

A standard practice at pig farms is to assign finisher pigs to groups based on their live weight measurements or based on visual inspection of their sizes. As an alternative, we used machine learning classification, namely the random forest algorithm, for assigning finisher pigs to groups for the purpose of increasing body weight uniformity in each group. Instead of relying solely on weight measurements, random forest enabled us to combine weight measurements with other phenotypes and genetic data (in the form of EBV's). We found that using random forest with the combination of phenotypic and genetic data achieves the lowest classification error (0.3409) in 10-fold cross-validation, followed by random forest with phenotypic and genetic data separately (0.3460 and 0.4591), then standard assignment based on birth weight (0.5611), and finally standard assignment based on the weight at the start of the finishing phase (0.7015).

Keywords: machine learning, random forest, pig breeding

Introduction

Variation in bodyweight has a big impact on the farming of pigs. Feed costs, drug dosages, farm management, and procurement plans are affected by the weights of the pigs being handled, and the uniformity (or lack thereof) of those weights. For instance, if a group of pigs in a finishing pen contains slow growers; those pigs must be retained in the pen until they reach market weight before the pen can be cleared to receive a new group. Therefore, a good estimate of each pig's growth performance can greatly improve the efficiency at pig farms and breeding facilities.

The purpose of accurate pig growth prediction is the ability to assign pigs at the farm to groups that will be uniform in weight at a target age, or groups that will reach a target weight at a uniform age. The standard practice of assigning finisher pigs to pens is based on past and current weight measurements of the pigs; or more frequently is done through visual inspection alone.

As with other animals, pig growth is a complex phenomenon that is influenced by many factors, including sex, age, weight history, feed intake, genetics, health, sow and litter characteristics, and farm conditions (Apichottanakul *et al.*, 2012). Therefore, it is not effective to isolate one, or too few of these factors, as predictors or proxies of future weight or growth.

The machine learning approach differs from traditional statistical analysis in that it emphasizes prediction accuracy of the models rather than the fit of the data to predetermined statistical models or structures (Breiman, 2001b), therefore allowing the inclusion of heterogeneous data types without hypotheses on which distributions generate them.

In animal science literature, machine learning methods have been used for predicting growth in farmed shrimps (Yu *et al.*, 2006), broilers (Roush *et al.*, 2006) and pigs (Apichottanakul *et al.*, 2011). Other notable uses of machine learning in animal science, specifically the use of the random forest algorithm, include identifying additive and epistatic genes associated with residual feed intake in dairy cattle (Yao *et al.*, 2013), identifying geographic patterns of different pig production systems (Thanapongtharm *et al.*, 2016), and predicting the insemination outcome of dairy cattle (Shahinfar *et al.*, 2014).

In this study, we use machine learning, namely the random forest classification algorithm (Breiman, 2001a) to combine the predictive power of both genetic and phenotypic predictors. In doing so, we aim to decrease the classification error of the following task: assigning each pig to one of three groups, based on the age it reaches a target weight of 120 kg.

Materials and methods

Data

The dataset used in this study was provided by Topigs-Norsvin. It consisted of features of purebred pigs that were born within a 4-year span in three farms. The features comprised different information about each pig from birth up until the start of the finishing phase such as birth weight, sex, and gestation length. These features form the input matrix , where is the number of pigs, and =28 is the number of features. We distinguish the phenotypic feature matrix from the genetic one by denoting them and respectively (, while denotes the complete feature matrix that includes both phenotypic and genetic data. A list of all features is given in the appendix.

The standardized age at 120 kilograms, being a proxy of a pig's growth potential near slaughter age, was used as the output . For classification, a discretized version of is created by labelling the lowest third of the pigs with respect to the value of (128 to 174 days) as "fast growers", i.e. the pigs that reach the target weight fastest or at the youngest age; and like so the middle third (175 to 190 days) as "average growers", and the final third (190 to 265 days) as "slow growers".

Classification methods

Random forest classification.

The random forest algorithm is a tree-based ensemble learning method. In machine learning, ensemble methods are those that combine weak regression or classification models to obtain a model that is stronger than all of its constituents. In the case of random forest, the aggregated base models are decision-tree predictors. The algorithm uses bagging (Breiman, 1996), as well as random sampling from the feature space at each node of a tree to create a “forest” of diverse tree predictors, which leads to a reduction of variance compared to an individual tree, and a reduction of over-fitting and sensitivity to changes in data.

Random forest for classification works as follows: i) Drawing bootstrapped sub-samples (random sampling with replacement) from the training set to grow classification trees; ii) Sampling variables form the feature matrix at each splitting node in each tree, and selecting the best split in each node until each tree is fully grown or a stopping criterion is met; iii) Computing the final prediction as the majority vote of predictions. In this paper, we

use the following parameters for the algorithm: , (rounded), and the stopping criterion is to stop splitting a node if the number of samples in it is less than .

Random forest provides an internal measure of feature importance, which can be utilized to interpret the resulting models, namely, to know which features are most relevant to the output. This feature importance measure is derived from accumulating the splitting scores for each variable. In this study, we use this measure to rank the features relative to each other. Then, we reevaluate the classification model using only the topmost ranking features. We implemented random forest using the Scikit-learn module in Python (Pedregosa *et al.*, 2011).

Standard pig assignment strategies.

The standard assignment strategies we present here describe simple rules that a pig farmer may implement without the use of computational tools. This can be done by relying on one of the available weight measurements: birth weight and the weight at the start of the finishing phase. Using the latter as an example, a farmer can group the heaviest third of her herd into a pen or a group of pens designated for the pigs that will reach the target weight fastest. Similarly, she places the average and lightest thirds of her herd into designated pens. This corresponds to two separate assignment strategies, one for each of the available weight measurements.

Results

Classification results

For each of the classification strategies, 10-fold cross-validation is implemented, and the average classification errors on the validation folds are presented in Table 1.

Table 1. Classification error (in 10-fold cross-validation) for the standard assignment strategies based on birth weight , and weight at the start of finishing ; and random forest with phenotypic features , genetic features , and all features . The baseline of 0.67 is the error made when the assignment is arbitrary without taking into account any available information.

	Standard assignment		Random forest		
Class					
Baseline	0.6700	0.6700	0.6700	0.6700	0.6700
Fast growers	0.5301	0.6900	0.2667	0.3763	0.2694
Average growers	0.6535	0.6907	0.4789	0.6108	0.4732
Slow growers	0.4997	0.7237	0.2925	0.3902	0.2803
Total	0.5611	0.7015	0.3460	0.4591	0.3409

Feature ranking

Figure 1 gives the ranking of features when random forest is run on the entire feature matrix , using all samples. To take into account the inherent randomness in the algorithm, we show the feature ranking results from the training folds of the 10-fold cross validation in the appendix.

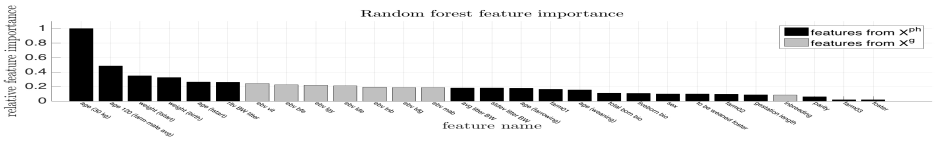

Figure 1. Relative importance of each feature from random forest, normalized so that the most important feature has an importance of 1.

Discussion and conclusion

The classification comparison shows a clear advantage of random forest over the standard pig assignment strategies that we proposed in this study, which were meant to mimic standard farm practices. That being said, the standard strategy based on birth weight still resulted in a much more uniform grouping than random (classification error), with a classification error of 0.5611; making it a viable and easy solution for this problem, if birth weight measurements were available to the farmer. On the other hand, assignment based the start of finishing weight, which would be the latest weight measurement at the moment of the assignment decision, seems to perform no better than a random assignment.

Using random forest, the phenotypic features result in a good classification with an error of 0.3460. The addition of genetic features (estimated breeding values) reduces the error to 0.3409. When the experiments are repeated with the top five ranking features, the resulting error is 0.3593, whereas the top ten features result in an error of 0.3442, close to that achieved with all the features.

Compared to other machine learning methods, like neural networks or support vector machines, random forest has a simpler model structure, making it easier to interpret by potential end users of this application. Moreover, random forest, being based on decision trees, is able to deal with heterogeneous data without the need of normalization. Nevertheless, it would be valuable in future work to make a comprehensive comparison between different machine learning classification methods for this application.

In conclusion, machine learning classification, random forest in this case, can assist pig farmers and breeders in achieving groups that are more uniform in weight by taking advantage of available data, a lot of which is relevant to the weight phenotype, but whose potential is untapped with traditional methods.

List of References

Apichottanakul, A., Pathumnakul, S., and Piewthongngam, K., 2012. The role of pig size prediction in supply chain planning. biosystems engineering 113 (3): 298-307.

Breiman L., 1996. Bagging predictors. Machine learning 24(2): 123-40.

Breiman, L., 2001a. Random forests. Machine learning 45 (1): 5-32.

Breiman, L., 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16 (3): 199-231.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,

Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12 (Oct): 2825-2830

Roush, W., Dozier, and W., Branton, S., 2006. Comparison of gompertz and neural network models of broiler growth. Poultry Science 85 (4): 794-797.

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P., and Weigel, K., 2014. Prediction of insemination outcomes in holstein dairy cattle using alternative machine learning algorithms, Journal of dairy science 97 (2): 731-742.

Thanapongtharm, W., Linard, C., Chinson, P., Kasemsuwan, S., Visser, M., Gaughan, A. E., Epprech, M., Robinson, T. P., and Gilbert, M., 2016. Spatial analysis and characteristics of pig farming in Thailand. BMC Veterinary Research 12 (1): 218.

Yao, C., Spurlock, D.M., Armentano, L.E., Page, C.D., VandeHaar, M.J., Bickhart, D.M. and Weigel, K.A., 2013. Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. Journal of dairy science 96 (10): 6716-6729.

Yu, R., Leung, P., and Bienfang, P., 2006. Predicting shrimp growth: artificial neural network versus nonlinear regression models. Aquacultural Engineering 34 (1): 26-32.

Acknowledgments

The authors want to acknowledge Topigs Norsvin for providing the data and the Netherlands Organisation of Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin for their financial support.

Appendix

Full list of features

In Table A.1, we include the full list of features in feature matrix , with their descriptions, types, ranges, means, and standard deviations whenever applicable

Table A.1.: Full list of features in feature matrix X and the output Y.
Table legend: (Ph.) phenotypic feature, (G.) genetic feature. (Num.) numerical feature. (Cat.) categorical feature.

		Type				
Feature name	Description (unit)	Ph./G.	Num./Cat.	Range	Mean	Std. dev
parity	Parity number of biological mother	Ph.	Num.	1 - 13	2.73	1.63
inbreeding	Inbreeding coefficient	G.	Num.	0 - 0.26	0.0178	0.0180
weight (birth)	Weight at birth (g)	Ph.	Num.	330 - 3250	1380	298
age (30 kg)	Age at 30 kg (days)	Ph.	Num.	48.9 - 115.3	76.44	8.09
weight (tstart)	Weight at the start of the finishing phase (kg)	Ph.	Num.	15 - 50	31.21	7.07
stdev litter BW	Std. deviation in birth weight in biological litter	Ph.	Num.	0 - 1036	279.26	80.31
avg litter BW	Average birth weight in biological litter	Ph.	Num.	600 - 2740	1299.28	211.61
rltv BW litter	Relative birth weight of animal compared to littermates	Ph.	Num.	-1080 - 1160	80.79	230.57
to be weaned foster	Number of piglets to be weaned by the foster mother	Ph.	Num.	0 - 38	13.59	2.89

liveborn bio	Number of born alive piglets in the biological litter	Ph.	Num.	1 - 28	14.23	3.28
total born bio	Number total born piglets in the biological litter	Ph.	Num.	1 - 30	15.53	3.44
age (weaning)	Age at weaning (days)	Ph.	Num.	1 - 63	23.99	4.57
gestation length	Gestation length of biological dam	Ph.	Num.	108 - 123	115.18	1.59
ebv lgy	Breeding value for sow longevity (parent average).	G.	Num.	-0.79 - 1.12	0.05	0.24
ebv vit	Breeding value for piglet vitality [current EBV].	G.	Num.	-11.9 - 12.6	0.14	3.17
ebv bfe	Breeding value for back fat thickness [parent average].	G.	Num.	-3.69 - 2.4	-0.28	0.89
ebv lde	Breeding value for loin depth thickness [parent average].	G.	Num.	-4.83 - 5.98	0.52	1.55
ebv tnb	Breeding value for total number born piglets [parent average].	G.	Num.	-2.25 - 2.69	-0.04	0.59
ebv mab	Breeding value for mothering ability [parent	G.	Num.	-6.58 - 4.90	0.08	1.39

	average].					
age 120 (farm-mate avg)	Age at 120 kg of farm-line-sex mates in last 3 months (days).	G.	Num.	156 - 202	182.19	10.97
age (farrowing)	Age of biological mother at farrowing (days).	Ph.	Num.	313 - 2119	616.48	243.88
ebv tdg	Breeding value for daily gain (calculated by quarter).	G.	Num.	31.22 - 39.79	35.21	1.45
sex	Female or male.	Ph.	Cat.	-	-	-
farm01	Farm of birth – farm 01.	Ph.	Cat.	-	-	-
farm02	Farm of birth – farm 02.	Ph.	Cat.	-	-	-
farm03	Farm of birth – farm 03.	Ph.	Cat.	-	-	-
foster	Fostered by biological or foster dam.	Ph.	Cat.	-	-	-
age (tstart)	Age at the start of the finishing phase.	Ph.	Num.	39 - 168	77.54	11.44
age (120 kg)	Standardized age at 120 kg, used as output variable after discretization.	Ph.	Num.	120.30 - 265.60	182.97	18.48

Feature ranking with multiple runs of random forest

Due to the randomness of random forest, the internal feature importance measure and corresponding feature ranking are commonly derived from multiple runs of the algorithm with different sub-samples of the data in order to ensure the robustness of the ranking. For that purpose, we include the feature ranking derived from running the algorithm on the training subsets of the 10-fold cross-validation.

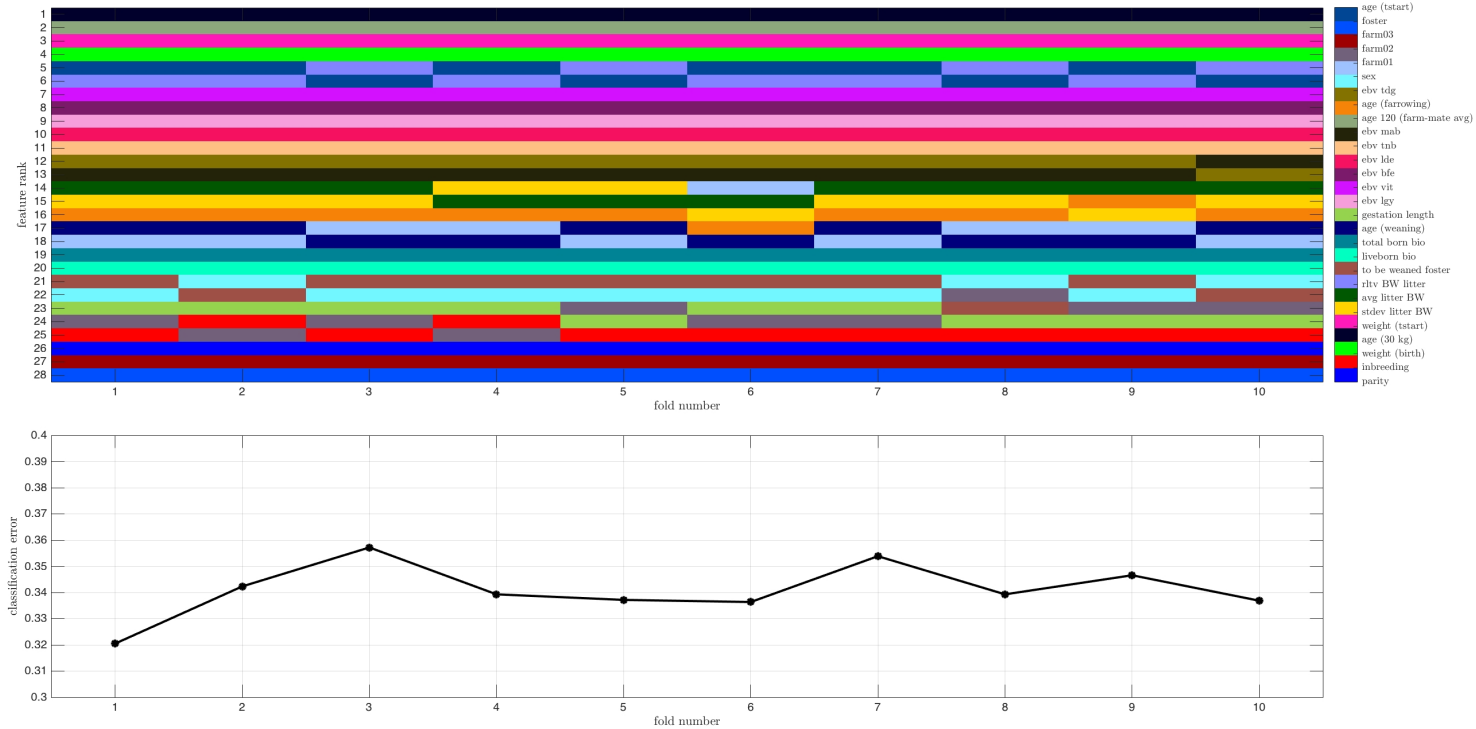

Figure A.1. Top: Feature rank derived from random forest implemented on the training subsets of 10-fold cross-validation. Bottom: The classification error on the corresponding validation subsets.