Predicting the polled genotype of ungenotyped animals using machine learning methods

A.B. Gjuvsland[1*]

[1] Geno, Storhamargata 44, 2317, Hamar, Norway; [*]arne.gjuvsland@geno.no

Abstract

Selection for polled in Norwegian Red cattle started 30 years ago and intensified the last 5 years. The SNP chip includes the (celtic) polled mutation, making selection very accurate for genotyped animals. However, ~85% of calves are ungenotyped and phenotypic horn status has been used to predict the polled genotype. This prediction has low accuracy (r^2=0.69) which limits genetic progress. To improve predictions, a broad set of variables related to horns were extracted from the herd registration system. Data for ~134,000 genotyped animals was used to train and compare different machine learning methods for predicting polled genotype. Gradient boosting machines performed best, followed by random forests, deep neural networks, and generalized linear models. The best gradient boosting machine gave a large improvement in accuracy (r^2=0.87). We have implemented this in the routine evaluations and expect increased accuracy of selection for ungenotyped animals and faster increase in polled allele frequency.

Introduction

The main genetic determinant of horns in cattle is the Polled locus, with the polled allele (*P*) dominant over horned (*p*). However, the genetics of horns is more complex (Grobler *et al.* 2021). Heterozygous polled (*Pp*) animals sometimes develop loose horn-like structures called scurs, and the inheritance and phenotypic expression of scurs is dependent on sex and breed. A number of scurs loci have been mapped on different chromosomes in different breeds (Gehrke *et al.* 2002) and in Norwegian Red cattle (NR), a scurs locus has been identified on BTA5 (Gjuvsland *et al.* 2019)

Historically, most NR were horned, but the celtic *P* allele (Medugorac *et al.* 2012) segregated in the population. In 1990 when selection for polled started the *P*-frequency was 2-3%. With genomics came new possibilities, a haplotype test was developed and later the causal mutation was added to the chip. When large-scale cow genotyping started in 2017, polled was included in the index (current weight 5%). For new-born calfs the *P*-frequency is now ~20%. For genotyped animals the number of P alleles is weighted into the index. However, ~85% of NR calfs are not genotyped and for these the predicted number of *P* alleles is used. Farmers register the phenotypic horn status for new-born calfs, and in each category the mean number of *P*-alleles among genotyped animals has been used for the ungenotyped animals. However, the accuracy of prediction is low, limiting the genetic progress for polled.

The machine learning ecosystem has developed rapidly in the last decade, with methodological breakthroughs and striking applications in many fields. The speed and breadth of development makes it challenging to identify the best algorithm and software for the problem at hand. Automated machine learning (AutoML) aims at automatizing the process of training models, tuning hyperparameters and comparing a wide range of algorithms as much as possible. Several powerful and easy-to-use AutoML packages are now available (Feurer *et al.* 2021, Ledell and Poirier 2020).

The goal of this work was to make a better predictor of polled genotype for ungenotyped animals. The approach taken here to achieve this is to use a wider set of relevant registrations and test a range of machine learning algorithms.

Materials & Methods

Dataset. Records from 1,154,254 animals with published breeding values were extracted from the database used for routine evaluations, 134,155 of the animals were genotyped. Each animal in the data set was characterized by birthyear, gender and the variables in Table 1.

Table 1. Records extracted for all animals in the dataset.

Variable	Description	Values
Pnum	Number of *P* alleles for the animal. NA if not genotyped.	0, 1, 2, NA
Hornbud	Horn bud registered by farmer at birth	Horned, Polled, Unknown
Dehorned	Dehorning registered by veterinarian	0, 1
Dehorned_age	Age (days) at dehorning	Integer
Pnum_sire, Pnum_dam, Pnum_mgs, Pnum_smgd	Number of *P* alleles for sire, dam, maternal grandsire and sire of maternal granddam of the animal. NA if relative is unknown or ungenotyped.	0, 1, 2, NA
Hornbud_dam	*Hornbud* for dam of animal	Horned, Polled, Unknown
Dehorned_dam	*Dehorned* for dam of animal	0,1
Snum_sire	Number of scurs alleles for sire of animal	0,1,2,NA

Cow features: For cows additional variables were computed: number of calfs (*Noff*), total number of *P* alleles for sires of the calves (*Npnum_sires*), number of calves registered as phenotypically "Polled" (*Npolloff*) and number of dehorned calves (*Ndhoff*). Two additional variables were derived as: *off_score1=Npolloff-Npnum_sires* and *off_score2=Ndhoff/Noff*. *Herd features:* Registration quality varies between herds and some herd variables were defined in an attempt to account for this: within-herd correlation between being registered as horned and being dehorned (*corr_HDH*), within-herd correlation between not being registered as polled and being dehorned (*corr_NPDH*), proportion of herd registered as horned (*rate_H*), proportion of herd not registered as polled (*rate_NP*) and the number of animals (*N_herd*).

Prediction method 1 Registered horn status, category-means. The first method for predicting genetic horn status (*Pnum*) for ungenotyped animals was based on the registered phenotypic hornstatus (*Hornbud*). Within each phenotypic category (Horned, Polled, Unknown) the mean number of *P-alleles* for genotyped animals was used as the predicted number *of P*-alleles for all the ungenotyped animals. This method was used in the routine runs for NR until 2021.

Prediction method 2 Genotypes of parents or male ancestors. Here we computed *Pnum_ped,* the expected number of *P* alleles inherited from the parents (eq. 1) or male ancestors (eq.2, if dam is not genotyped) of an animal. Equation 1 was used for animals with genotyped dams, and equation 2 for animals with ungenotyped dams.

$$Pnum\_ped = 0.5*Pnum\_sire + 0.5*Pnum\_dam \quad (1)$$

$$Pnum\_ped = 0.5*Pnum\_sire + 0.25*Pnum\_mgs + 0.125*Pnum\_smgd \quad (2)$$

Prediction method 3 Automatic machine learning with all variabless. Here we used automatic machine learning to predict genetic horn status (*Pnum*) from all the other variables. We used the AutoML functionality (LeDell and Poirier, 2020) in the H2O (http://h2o.ai) R-package v3.32.1.1. The following machine learning algorithms were included: generalized linear models with elastic net regularisation, distributed random forests, gradient boosting machines, fully-connected multi-layer artificial neural network and ensemble models. The automl function was run on 20 CPUs for 4 hours and 10-fold cross-validation was used for tuning of hyperparameters and model comparison.

Results and Discussion

Distribution of horn bud and dehorning registrations. Phenotypic registrations of horn status (*Hornbud)* and dehorning correspond fairly well with polled genotype. The majority of *pp* animals are registered as horned, while most *Pp* and *PP* are registered as polled (Figure1). For both *pp* and *Pp* animals there are small, but clear sex-differences in registered horn status, with male calves being registered as horned more often than females. When it comes to dehorning there is a substantial sex-difference for *Pp* animals, with 21.8% of the males and 3.5% of the females being dehorned.

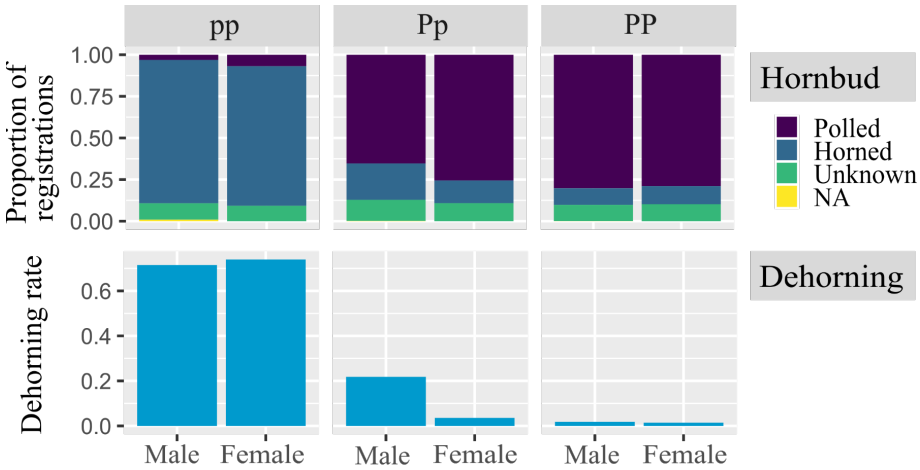

Figure 1 Horn phenotypes conditional on sex and polled genotype for 127K genotyped NR calves (92K females / 35K males) born 2005 or later.

Automatic machine learning. Gradient boosting machines (GBM) outperformed the other methods, while random forests, deep learning networks and generalized linear models had quite similar performance (Table 1). For many applications with complex data types like images, deep learning networks are superior, but for tabular data gradient boosting is often better (Shwarz-Ziv and Armon 2022). In many machine learning use cases an ensemble of all models outperform the best performing single model, but here there is practically no difference between the ensemble and the best GBM. GBMs are complex, non-linear models, but "variable importance" allows for some interpretability. The best GBM ranks *Hornbud* as the most important variable, followed by *Pnum_sire, Dehorned* and *Hornbud_dam.*

Table 1. AutoML leaderboard.

Best model per algorithm	RMSE
Ensemble of all models	0.2959
Gradient boosting machine	0.2960
Distributed random forest	0.3208
Deep learning[1]	0.3336
Generalized linear model[2]	0.3385

[1] Fully connected multilayer neural networks
[2] With elastic net regularization

Comparison of prediction methods. Cross-validation (10-fold) on genotyped animals was used to evaluate and compare the prediction methods. *Pnum* and the predictions are numeric so we used correlation, rather than classification-metrics, to measure accuracy. Prediction of number of *P* alleles by registered horn status (Figure 2 red line) separates (on average) *pp* animals from *Pp* and *PP*, but it is unable to separate *Pp* from *PP* due to dominance. The accuracy, measured as the correlation to the true genotype (black line), is fairly low (cross-validated r^2=0.69). Using the expected number of *P* alleles from the genotype of parents or male ancestors (green line) separates *pp* from *Pp* and *Pp* from *PP*, but the accuracy goes down (r^2=0.61). Prediction with

gradient boosting (blue line) separates well between the three polled genotypes and the accuracy increases substantially (r^2=0.87).

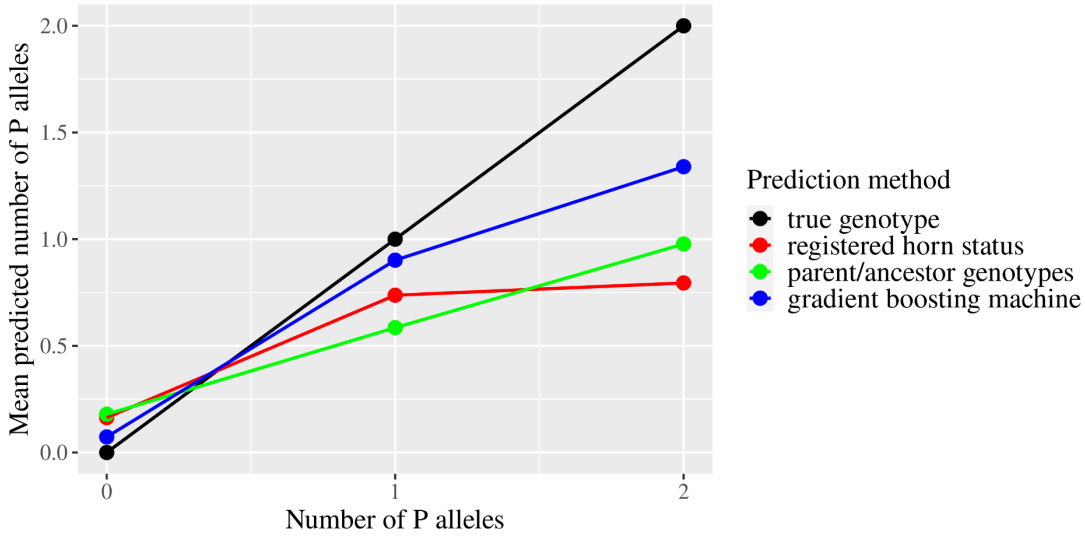

Figure 2. Mean predictions of polled genotype for genotyped animals . True number of *P*-alleles (x-axis) versus the mean predicted number (y-axis) when the genotype is masked in cross-validation.

Conclusions and implementation: Machine learning with a broad set of variables gives a substantial increase in a prediction accuracy compared to prediction based on registered horn status. The prediction models are trained and validated on genotyped animals, however the practical use in the breeding programme is for predicting polled genotypes of ungenotyped animal. The value of the predictions therefore relies on the genotyped animals being representative for the whole population. We expect the genotyped animals to be representative in many aspects, but when it comes to things like unknown parents and pedigree errors there are systematic differences and the prediction accuracy for ungenotyped animals will be lower than observed here. However, considering the low numbers of pedigree errors and unknown parents in the NR population we expect the predictions to be quite accurate for most animals and have implemented the gradient boosting method in our routine evaluations. This will give more accurate selection for polled in particular on the female side where many breeding animals in the herds and dams of potential selection candidates are ungenotyped.

References

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2021). *arXiv preprint arXiv:2007.04074*. https://arxiv.org/abs/2007.04074v2

Gehrke L.J., Capitan A., Scheper C. *et al.* (2020) *Gen Sel Evol* 52:6. https://doi:10.1186/s12711-020-0525-z

Gjuvsland, A.B., Storlien H., and Larsgard A.G. (2019) Proc. of 70th EAAP, Ghent, Belgium.

Grobler R., van Marle-Köster E. and Visser C. (2021) *Livestock Sci*ence 247:104479. https://doi:10.1016/j.livsci.2021.104479

LeDell E.A. and Poirier S. (2020) 7th ICML Workshop on Automated Machine Learning. https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.

Medugorac I., Seichter D., Graf A., *et al.* (2012) PLoS ONE 7(6):e39477. https://doi:10.1371/journal.pone.0039477

Shwartz-Ziv, R. and Armon, A. (2022) Information Fusion, 81:84-90. https://doi.org/10.1016/j.inffus.2021.11.011