A COMPARATIVE STUDY OF MACHINE LEARNING METHODS TO PREDICT AVERAGE DAILY GAIN FROM SINGLE NUCLEOTIDE POLYMORPHISMS

Piles M.[1]*, Tusell L.[2], Velasco-Galilea M[1]., Ballester M[1]., Sánchez J.P.[1]

[1]Animal Breeding and Genetics Program, Institute of Agriculture and Food Research and Technology (IRTA), Torre Marimon s/n, 08140 Caldes de Montbui, Barcelona, Spain
[2]GenPhySE, Université de Toulouse, INRAE, F-31326 Castanet-Tolosan, France
*Corresponding author: miriam.piles@irta.es

ABSTRACT

This study compares the accuracy of prediction of total genetic effects, i.e. additive and non-additive genetic effects, of average daily gain (**ADG**) from single-nucleotide polymorphisms (**SNPs**) using two machine learning (**ML**) algorithms, i.e. Elastic Net and Support Vector Machine, and a genome-enabled best linear unbiased prediction model (**GBLUP**) as benchmark. The target examples were 439 ADG records which were previously adjusted for environmental systematic and random effects. After quality control and selection of one SNP per linkage group, the retained 14,713 SNPs were ranked using their importance measure for predicting the adjusted ADG records. Then, different subsets with increasing number of the most informative SNPs (50, 100, 200, 300, 500 and all most informative SNPs) were used as variables for predicting adjusted ADG records either by using radial basis function SVM or ENET. Optimal hyperparameters for the two algorithm were tuned using nested resampling. The predictive performance of each ML algorithm and the GBLUP was evaluated as the median of the Spearman correlation (**SC**) across the 30 testing sets originated from a 6-fold cross-validation repeated 5 times. The best predictive performance and repeatable results were obtained with a subset of 100 SNPs and using ENET with a median SC of 0.26 and an interquartile range of 0.07. Predictive ability was null when using all available SNPs either using ENET, SVM or GBLUP. The selected subset of 100 SNPs that have been identified could be potentially used in selection to boost genetic progress of ADG.

Key words: Support Vector Machine, Elastic Net, prediction, growth, genome selection

INTRODUCTION

The availability of high-density panels of molecular markers in rabbit makes possible the use of genomic selection in this species. However, its efficacy and economic interest for the improvement of expensive and difficult to measure traits needs to be assessed. Marker-based models for genetic selection have shown their superiority over pedigree-based models for predicting complex traits in many species (Hayes et al., 2009; de los Campos et al., 2009). Most of the applications use additive linear regression models. However, prediction accuracy could be even further improved by using models and procedures that are able to capture and integrate other sources of non-additive genetic variation such as dominance or epistasis even when the number of records is much smaller than the number of parameters. Machine learning (**ML**) algorithms can capture complex relationships between predictor variables and target traits. They have substantial computational demands and risk of overfitting the training data. However, when they are applied within a resampling strategy to predict or classify an output, it is feasible to obtain an optimal parameterization of the prediction model and an assessment of the generalizability of the results. Among them, Support Vector Machine (**SVM**; Vapnik et al., 1999) and Elastic Net (**ENET**; Zou and Hastie, 2005) are, respectively, non-linear and linear method, which have shown good performances in both classification and regression problems (Long et al., 2011; Zou and Hastie, 2005).

The aim of this study is to assess the accuracy of prediction of total genetic effects, i.e. additive and non-additive genetic effects, of average daily gain (**ADG**) from single-nucleotide polymorphisms (**SNPs**) using ML algorithms.

MATERIALS AND METHODS

Animals and Data. Animals come from the Caldes line selected for growth rate during the fattening period (32-60d). They were bred in 5 batches in two farms and under two feeding regimens: *ad libitum* or restricted to 75% of the *ad libitum* feed intake. Animals were weighted once per week and ADG was computed for each animal as the regression coefficient of body weight on age at recording using the lm() function of the "stats" R package. ADG records were adjusted for systematic and random environmental factors with the function lmer() of the R package "lme4". Systematic factors resulted from the combination of the farm with batch, feeding regimen, food type, body size at weaning, parity order and litter size. Random factors included box and litter. Outlier records within combination of systematic effects were removed. Finally, adjusted records were centered and standardized. A total of 439 records remained for the analyses. The DNA extraction was carried out from liver samples of 439 growing rabbits using the kit NucleoSpin Tissue (250prep) (Macherey-Nagel). DNA extracts were sent to an Affymetrix platform to conduct genotyping using the Axiom Rabbit Genotyping Array "Axiom_OrCunSNP" (Thermo Fisher Scientific), which includes 199,692 variants. Only 161,830 variants were segregating in our population and, after retaining the SNPs mapped in autosomes in the OryCun2.0 assembly and applying standard quality control criteria, 114,604 SNPs were retained. Quality control criteria comprised retaining animals having at least 90% of SNPs correctly genotyped, SNPs with less than 5% missing genotype data and SNPs with a MAF higher than 5%. The linkage disequilibrium decay pattern from our population was estimated and used to retain one SNP per linkage group resulting in 14,713 SNPs kept for further analyses.

Statistical Analysis. In a first step, SNPs were ranked using their importance measure for predicting the adjusted ADG records in an unbiased random forest algorithm based on conditional inference (Strobl et al., 2007). Then, different subsets with increasing number of the most informative SNPs (50, 100, 200, 300, 500 and all most informative SNPs) were used as variables for predicting adjusted ADG records either by using radial basis function SVM or ENET. Support vector regression is an application of SVM methodology (Vapnik, 1995) which minimizes a regularized loss function (the insensitive-loss function). Performance of SVM is very sensitive to the values of two main hyper-parameters: the "cost parameter" ("C"), which is a trade-off between model complexity and training error; and the "gamma" parameter from the Gaussian function inside the kernel. Both hyper-parameters were simultaneously tuned in nested resampling. Elastic net (Zou and Hastie, 2005) is a regression method that combines in a mixture the penalty approach of ridge regression ($\lambda_1 \times \left[\sum_{j=1}^{p} \beta_j^2\right]$) and the penalty approach of lasso ($\lambda_2 \times \left[\sum_{j=1}^{p} |\beta_j|\right]$) as: $\lambda \times \left((1-\alpha) \times \left[\sum_{j=1}^{p} |\beta_j|\right] + \alpha \times \left[\sum_{j=1}^{p} \beta_j^2\right]\right)$, with λ being the coefficient for the ENET penalty, $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$ and the β's being the regression coefficients of the multiple regression. When the number of predictors is much higher than the number of data ($p >> n$), ENET allows more than n predictor variables to be selected out of p candidates. In addition, it can select groups of correlated variables, as it could be the case of genes sharing the same biological pathway. The value of α hyper-parameter was tuned in nested resampling, while the optimal λ parameter was internally found by cross-validation.

Both ML algorithms were implemented using the R package "mlr" (Bichl et al., 2016) which allowed to compare results from the two algorithms under the same conditions and to find the optimal hyperparameters for each algorithm. The ”e1071” R package and the function “cv.glmnet” from the “glmnet” R package were used via “mlr ” for the SVM and ENET analyses, respectively. The implementation of nested resampling for hyperparameter tuning and validation consisted of 2 nested resampling loops. In an outer resampling loop, a 6-fold cross-validation repeated 5 times was performed resulting in 30 pairs of training/testing sets. On each outer training set, hyper-parameter tuning was done executing an inner resampling loop that consisted in a 6-fold cross-validation repeated 2 times. Therefore, there was one set of selected hyper-parameters for each outer training set. The performance criterion used to select the best hyper-parameter set was the coefficient of determination (i.e. R-squared). Then the learner was fitted on each training set using the selected hyperparameters and its performance was evaluated on the corresponding testing set.

The predictive performance of a genome-enabled best linear unbiased prediction model (**GBLUP)** was used as a benchmark as it has been widely used for prediction of genomic breeding values (de los Campos et al., 2009). In GBLUP, adjusted ADG phenotypes are regressed on additive genomic effects, or genomic breeding values $\mathbf{u} = \{u_i\}$ (for *i*=1,…, *n* individuals) that are assumed to be normally distributed $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$, where σ_u^2 is the additive genomic variance, and $\mathbf{G}$ is the genomic relationship matrix (VanRaden, 2008). GBLUP was implemented using the R package “BGLR” (Pérez & de los Campos, 2014) and its predictive ability was assessed using the same training/testing sets used in the ML methods.

For all methods, the median (**Md**) and the interquartile range (**iqr**) of the Spearman correlation (**SC**) between the predicted and observed adjusted phenotypes of the 30 testing sets was used to assess predictive performance.

RESULTS AND DISCUSSION

Figure 1 shows boxplots of the SC between predicted and observed adjusted phenotypes obtained in the 30 testing datasets using SVM and ENET with the different subsets of SNPs. The best predictive performance and repeatable results were obtained with a subset of 100 SNPs and using either ENET or SVM being the Md (iqr) 0.26 (0.07) and 0.23 (0.09) for ENET and SVM, respectively. These figures can be considered expectable predictive performances given the low genetic determinism estimated for this trait in the same population. Thus, using data from the same experiment, the posterior means of heritability for ADG under restricted and *ad libitum* feeding were estimated to be 0.08 (SD = 0.02) and 0.21 (SD = 0.05), respectively. The fact that best predictive performance is obtained with a subset with the 100 most informative SNPs for the target trait in both ML algorithms indicates the high importance of performing feature selection for prediction purposes, especially when the number of features is high and the number of training examples available is limited. In this study, feature selection allowed reducing the number of predictor variables to a small number, which possibly avoids redundant information while reducing parameter dimensionality and computation time. From the point of view of selection, it could allow to genotype candidates with a low density SNP-chip, reducing genotyping costs.

Elastic net performed well with a reduced number of SNPs. However, it failed to fit a model when the number of features was equal or larger than 200. For example, it failed for 28 out of 30 training/test pairs of sets using 500 SNPs and in all the cases when all SNPs were used. This result was not expected since this method involves regularization through two penalization terms. Radial basis function SVM enables modeling nonlinear relationships between the phenotype and the SNPs. In this study, SVM did not outperform the predictive performance already obtained with ENET, possibly because most of the genetic determinism of ADG is of additive nature.

The Md (iqr) of the SC between observed and predicted values for GBLUP was -0.04 (0.15), which indicates no ability to predict adjusted ADG records when all SNPs are used as predictors in a linear model. Same results were obtained with SVM using all predictors available. : -0.05 (0.13).

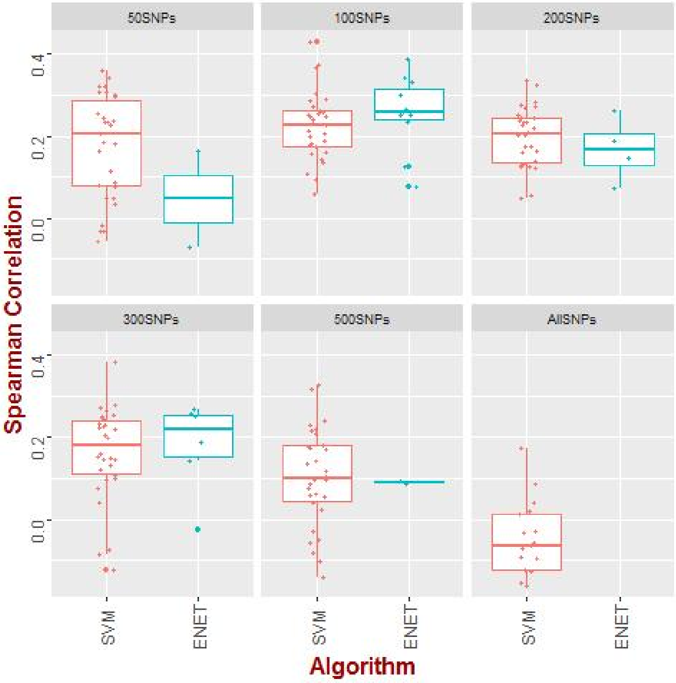

Figure 1: Boxplots of the Spearman correlation between predicted and observed data obtained in the testing datasets using radial basis function Suport Vector Machine (SVM) and Elastic Net (ENET) using different subsets with increasing number of predictors (50, 100, 200, 300, 500 and all of the 14,731 SNPs).

CONCLUSIONS

This is the first time that ML algorithms have been used to predict rabbit phenotypes from SNP genotypes. A low prediction performance was obtained with both SVM and ENET with a subset of 100 SNPs selected using a random forest algorithm whereas predictive ability was null when using all available SNPs either using ENET, SVM or GBLUP. The selected subset of SNPs that have been identified could be potentially used in selection for ADG.

ACKNOWLEDGEMENTS

This research was supported by the Feed-a-Gene Project funded by the European's Union H2020 Programme under grant agreement EU 633531.

REFERENCES

Perez P., de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package, Genetics, 198, 483-495.

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G., Jones Z.M. 2016 mlr: Machine Learning in R. J Mach Learn Res.,17:1-5.

Strobl .C, Boulesteix A.L., Zeileis A., Hothorn T. 2007.Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics;8:25.

Vapnik VN. 1999. The nature of statistical learning theory. 2nd ed. New York: Springer-Verlag.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol., 67, 301-20.

VanRaden PM 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.