A COMPARATIVE STUDY OF SUPPORT VECTOR MACHINE AND GBLUP 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 radial basis function Support Vector Machine, (**SVM**) and a genome-enabled best linear unbiased prediction model (**GBLUP**) as benchmark. The target examples were 425 ADG records which were previously adjusted for environmental systematic effects. After quality control and selection of one SNP per linkage group 14,710 SNPs were retained. A nested resampling was implemented. For analysis with SVM, in each training set of an outer 6-fold cross-validation, SNPs were first ranked using their rank correlation with the adjusted ADG records. Then, hyper-parameter tuning was performed using an inner 6-fold cross-validation in each training set for different learner configurations including as predictor variables different subsets with increasing number (50, 100, 200, 300, 500, 1000) of the best ranked SNPs and a set with all of them. Finally, prediction performance was evaluated in the outer testing sets using the median of the Spearman's correlation (**SC**) between predicted and adjusted phenotypes. Same pairs of training/test sets were used for prediction of adjusted ADG records using GBLUP. The best prediction performance was obtained with SVM with a subset of 1000 SNPs. In this case, the median (**Md**) of SC was 0.34 with an interquartile range (**IQR**) of 0.20 for this parameter. When prediction was performed using GBLUP with all SNPs, the Md of the SC was 0.28 with an IQR of 0.12. The selected subset of SNPs that have been identified could be potentially used in selection to boost genetic progress of ADG.

Key words: Support Vector Machine, machine learning, 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., 2013). 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) is a non-linear method which have shown good performances in both classification and regression problems (Long et al., 2011).
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 machine learning 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 environmental factors with the function lm() of the R package "stats". Systematic factors resulted from the combination of the farm with batch, feeding regimen, food type, body size at weaning, parity order and litter size. Outlier records within combination of systematic effects were removed. Finally, adjusted records were centered and standardized. A total of 425 records remained for the analyses. The DNA extraction was carried out from liver samples of the same 425 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 14,710 SNPs kept for further analyses.

Statistical Analysis. A nested resampling was implemented. In each training set of an outer 6-fold cross-validation, ranking of SNPs based on the rank correlation between the adjusted ADG records and the SNP was first established. Then, tuning of the hyper-parameters of the radial basis function SVM was performed also within each training set using an inner 6-fold cross-validation. This was done for different configurations of the learner, which included as predictor variables different subsets with increasing number of the most correlated SNPs (50, 100, 200, 300, 500 and 1000) and a set with all of them. Finally, model was fitted on the entire outer training set and the prediction performance was evaluated on the corresponding outer testing sets.
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 evaluating all possible combinations of the tested values (C = 0.0001, 0.01, 0.1,1; sigma = 0.005, 0.05, 0.5, 5). The performance criterion used to select the best hyper-parameter set was the mean squared error. SVM was implemented using the ”e1071” R package within the "mlr" R package (Bichl et al., 2016) which allowed to find the optimal hyperparameters and compare results across learner configurations.
The prediction 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., 2013). 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). A Bayesian GBLUP was implemented using the BLUPF90 family programs (Misztal, 2002) and its predictive ability was assessed using the same training/testing sets used for the machine learning analysis.

For both methods, the median (**Md**) and the interquartile range (**IQR**) of the Spearman correlation (**SC**) between the predicted and observed adjusted phenotypes of the 6 testing sets was used to assess prediction performance.

RESULTS AND DISCUSSION

Figure 1 shows boxplots of the SC between predicted and observed adjusted phenotypes obtained in the 6 testing datasets using SVM with the different subsets of SNPs and the GBLUP using the 14710 SNPs. The best prediction performance was obtained with SVM using a subset of 1000 SNPs being the Md (IQR) 0.34 (0.20). These figures can be considered quite good prediction 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 (Piles et al., 2017).

The fact that best predictive performance is obtained with a subset with the 1000 most informative SNPs and that similar results are obtained with just 500 out of 14710 SNPs, 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 to a small value the number of predictor variables, which possibly avoids redundant information while reducing parameter dimensionality and computation time. From the point of view of selection, this identified subset of SNPs could allow to genotype candidates with a low density SNP-chip, reducing genotyping costs.

The Md (IQR) of the SC between observed and predicted values for GBLUP was 0.28 (0.12), which indicates a slightly lower ability to predict adjusted ADG records when all SNPs are used to account for genomic relationships among individuals in a linear model. Radial basis function SVM enables modeling nonlinear relationships between the phenotype and the SNPs. In this study, SVM slightly outperformed the predictive performance obtained with GBLUP, possibly because most of the genetic determinism of ADG is of additive nature.

CONCLUSIONS

This is the first time that a ML algorithm has been used to predict rabbit phenotypes from SNP genotypes. A good prediction performance was obtained with a subset of just 500 SNPs selected on the basis of the rank correlation between SNP and the adjusted records. The prediction performance of SVM slightly outperformed that of GBLUP, which used all SNPs information. 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.

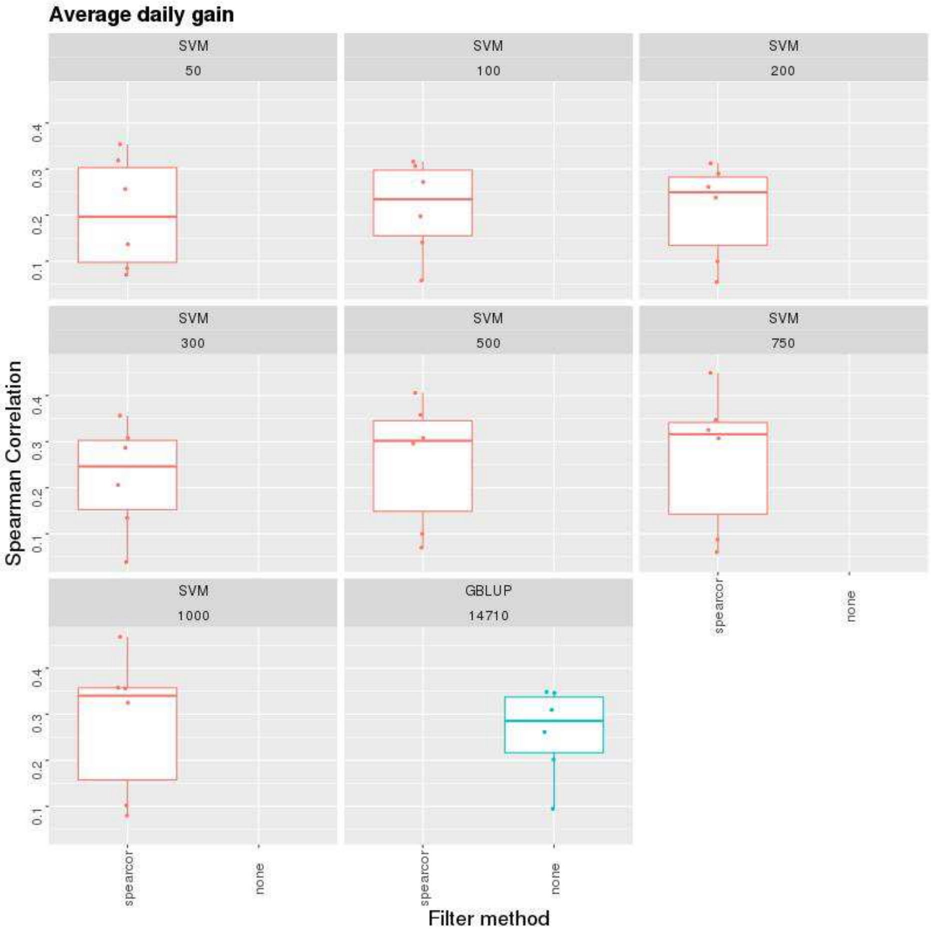

Figure 1: Boxplots of the Spearman correlation between predicted and observed data obtained in the testing datasets using radial basis function Suport Vector Machine (SVM) and genomic BLUP (GBLUP) using different subsets with increasing number of predictors (50, 100, 200, 300, 500 and all of the 14,731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

REFERENCES

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.*

de los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D., Calus M.P.L. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics 193, 327-345*

Hayes B., Bowman P., Chamberlain A., Verbyla K., Goddard M. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Select. Evol. 41, 51.*

Long N., Gianola D., Rosa G.J., Weigel K.A. 2011. Application of support vector regression to genome-assisted prediction of quantitative traits, TAG. *Theor. Appl. Genet. 123, 1065-1074.*

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T., Lee D.H. BLUPF90 and related programs (BGF90). 2002. *In: Proc. 7th World Congress on Genetics Applied to Livestock Production: 19–23 August 2002, Montpellier, 2002*

Piles M., David I., Ramon J., Canario L., Rafel O., Pascual M., Ragab M., Sánchez J.P. 2017. Interaction of direct and social genetic effects with feeding regime in growing rabbits, *Genet. Select. Evol. 49,58.*

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

VanRaden P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91, 4414–4423.*