Genomic Evaluation using Machine Learning Algorithms in the Spanish Holstein Population

J. A. Jiménez-Montero[1], O. González-Recio[2], R. Alenda[1] and J. Pena[3]
[1]Departamento de Producción Animal, E.T.S.I. Agrónomos – Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain
[2] Departamento de Mejora Genética Animal, INIA, Crta. La Coruña km. 7.5, 28040 Madrid, Spain
*[3]Departamento Técnico CONAFE, Ctra. de Andalucía, km. 23,600. 28340 Valdemoro (Madrid). Spai*n
e-mail: **joseantonio.jimenez.montero@upm.es**

Abstract

The aim of this study was to validate the recorded data and genome-assisted evaluation models for the Spanish Holstein population as an initial step towards the first official national genomic evaluation.
Preliminary national genomic evaluation for production and type traits in Holstein Friesian bulls in Spain were tested using both the Spanish reference population (ESP), composed by 2,115 progeny tested bulls, and the Eurogenomics population (EG), composed by 22,247 progeny tested bulls. Four different traits currently included in the Spanish genetic evaluation were used: milk yield (MY), fat yield (FY), protein yield (PY), and udder depth (UD).

Two different genomic evaluation methodologies, Bayesian-Lasso (B-Lasso) and a machine learning algorithm: Random-Boosting (R-Boost) were compared to traditional pedigree index (PI).
The predictive ability was measured in terms of correlations, mean square error (MSE) and regression coefficients between progeny proofs and direct genomic values (DGV) in the validation set. Genomic evaluations were more accurate than the traditional pedigree index. The increment in Pearson correlation between observed and predicted response depended on the trait, but the EG population provided greater accuracy than ESP at predicting future progeny performance, as expected.

The methodologies implemented showed similar results. B-Lasso showed higher Pearson correlations for MY (0.590 vs 0.572), FY (0.655vs 0.649) and PY (0.583vs 0.545), whereas R-Boost showed larger values for UD (0.584 vs 0.562).

Genomic predictions from R-Boost resulted in 4.03% lower predictive mean square errors than B-Lasso. R-Boost showed smaller MSE for MY, PY and UD, whereas B-Lasso was preferred for FY in terms of MSE.

R-Boost showed regression coefficients more close to 1 than B-Lasso.

The response to different methodologies of genomic evaluation was within the range of values expected for a population of a similar size. The methods that presented higher Pearson correlation also showed larger MSE. This should be considered in model comparison study deciding the method with better predictive ability.

Keywords: genome-assisted evaluation, machine learning, predictive ability, model comparison

Introduction

Over the last decade the Spanish breeding program has provided competitive bulls for the national and international markets due to a robust milk-recording scheme and an efficient organization. Special care has been taken in recording morphologic traits. GS has revolutionized dairy cattle breeding since 2008. Taking advantage of this technology is necessary to maintain the program's viability.

Different approaches are currently used for estimating genomic values. It is important to evaluate the performance of diverse methodologies and to identify the methodology

that has a higher predictive accuracy for routine Genomic Selection (GS) evaluations in a given population. Machine learning methods are an interesting alternative for dealing with these situations (Long *et al.,* 2007). Machine learning methods usually compare equal or favorable to Bayesian regression models (e.g., Moser *et al.,* 2009; González-Recio and Forni, 2011). These non-parametric methods can be implemented in both regressions on markers (e.g., Boosting) and/or building a (co)variance structures such as RKHS (Gianola *et al.,* 2006). The boosting algorithm is one of the most appealing machine learning methods for dealing with genomic-assisted evaluation problems and provide higher accuracies and lower biases than other methods (González-Recio *et al.,* 2010). A more efficient estimation of DGVs in dairy cattle can be obtained through some modifications on the algorithm; this modified algorithm, called Random-Boosting (R-Boost) was described by González-Recio (Personal communication)

The aim of this study was to check the predictive ability of two different methods for genomic evaluations in the Spanish Holstein population (ESP) as an initial step towards the first official national genomic evaluation.

Materials and Methods

Genotyped Bulls

A total of 1797 sires progeny tested in Spain were included as Spanish reference population.

When the Eurogenomics reference population was included the population size increased to 22,300 genotyped animals. All animals were genotyped with either v1 or v2 of the Illumina BovineSNP50 Beadchip.

Phenotypes

The January 2009 progeny proofs for milk yield (MY), fat yield (FY), protein yield (PY), and udder depth (UD), were used as a dependent variable. The production and type data were collected between 1980 and 2008. Production data existed for 1,414,347 daughters of genotyped bulls, while 969,567 of them had also type record available.

SNP Editing

SNPs with >5% incidence of missing genotypes across individuals, and SNPs with a minimum allele frequency (MAF) lower than 5% were discarded, leaving 39,714 SNPs for testing.

Only common SNPs in v.1 and v.2 versions of the 50K Illumina Beadchip were used with the Eurogenomics (EG) population, X chromosome was also discarded. After editing a total of 36971 SNPs were retained in this case.

Training and Validation Data Sets

The respective training sets were comprised of bulls born before 2005 with reliability higher than 75% in the January 2009 MACE genetic evaluation (1576 and 1562 in the ESP for production and type, respectively, and 14494 and 14306 in EG for production and type, respectively). The December 2011 progeny proofs were used as benchmark predicted response for sires born in or after 2005 as a testing set. Only sires with reliability higher than 75% in their December 2011 progeny proofs were included (221 and 196 in the ESP for production and type, respectively, and 3306 and 268 in EG for production and type, respectively).

Genomic Evaluation Model

The general structure for the models in linear form is

$$\mathbf{y} = \mu\mathbf{1}_n + \sum_j \mathbf{X}_j\mathbf{g}_j + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypic records, μ is the overall mean, $\mathbf{1}_n$ is a vector of n ones, $\sum_j$ is a summation over all markers, $\mathbf{g}_j$ is the vector of the effects for each marker effects, $\mathbf{X}_j$ is a design matrix of genotype codes and $\mathbf{e}$ is a vector of residuals.

Method 1 Bayesian-Lasso

The Bayesian counterpart of the LASSO model (Park and Casella 2008; de los Campos *et al.*, 2009) was used to estimate the SNP coefficients in the training population. A single chain of Gibbs sampling was run using 25,000 iterations and a burn-in period of 15,000.

Method 2 Random-Boosting

The boosting algorithm is a machine learning technique that combines different predictors and some shrinkage factor (Friedman, 2001). Boosting iteratively adds basis functions such that each addition further reduces the selected loss function (Hastie *et al.*, 2009). In this study, the ordinary least square estimation was chosen as basis function and was successively applied to the residuals of the previous estimation in a sequential manner. The MSE of the prediction was used as the loss function to minimize. R-Boost is a modification of the original algorithm that proposes to sample *mtry* covariates at random out of the *p* SNPs at each iteration, and select the SNP among the *mtry* that minimizes the given loss function.

The R-Boost algorithm would flow as follows:

(Initialization): Given data $\Psi = (\mathbf{y}, \mathbf{X})$, let the prediction of phenotypes be $\hat{F}_0 = \hat{\mu}$.

Then, for *m* in {1 to *M*}, with *M* being large proceed as:

Step 1. Draw *mtry* out of *p* covariates from the original training set to construct a reduced training covariate matrix $\Psi^{(b)} = (\mathbf{y}, \mathbf{X}_{mtry})$ to train the algorithm in iteration *m*.

Step 2. Calculate the loss function *L* $\left(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, mtry_m)\right)$ for all *mtry* SNPs and select that minimizing $\sum_{i=1}^{n} L(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, mtry_m))$ in the tuning set at iteration *m*, with $h(y_i; \mathbf{x}_i, mtry_m)$ being the prediction of the observation *i* in the tuning set using the learned parameters or coefficients of $h(\cdot)$ on the SNP $mtry_m$. These parameters or coefficients are learned using the training set as in the original algorithm.

Step 3. Updated predictions at iteration *m* in the form $F_m(\mathbf{x}_i) = F_{m-1}(\mathbf{x}_i) + v \cdot h(y_i; \mathbf{x}_i, mtry_m)$ with *v* being some shrinkage factor, e.g. *v*=0.10.

Step 4. Update the residuals to be used in the next iteration as $y_i = y_i - F_m(\mathbf{x}_i)$.

Repeat steps 1 to 4 a large number of times (*M*).

In this case, *v* was set to 0.10 for the production traits and 0.20 for UD and the percentage of SNPs selected at each iteration (*mtry*) was set to 0.05, 0.01, 0.05, and 0.05 for MD, FD, PD, and UD, respectively.

The main advantage of this approach is that the covariates (SNPs) are randomly sampled to compute the loss function, thereby decreasing the computation time while maintaining similar or better predictive ability than the original Gradient Boosting.

Based on the performance of the algorithm, the EG population was evaluated with R-Boost.

Criterion for Comparisons

DGVs were predicted for sires in the testing set. The accuracy of the genomic evaluation was computed as the Pearson correlation between the predicted DGV and the December 2011 progeny proofs. The pedigree index for sires in the testing set was used as benchmark. It was calculated as 50% of the sires EBV, +25% of the maternal grand sires EBV +12.5% of the maternal great-grand sires EBV. Finally, the MSE of the predictions and regression coeficients were also calculated.

Results and Discussion

Quality Control

For the ESP population, after filtering, the distribution of MAF was nearly uniformly

distributed with a mean of 0.28. The average distance of adjacent SNPs was 0.06 Mb. The remaining SNPs had a heterozygosity of 0.286%. The linkage disequilibrium, measured as r^2, between adjacent SNPs was 0.24. All of those values were in the range of previously reported values for other Holstein populations (Wiggans *et al.,* 2009; Banos and Coffey, 2010). The descriptors of the genomic structure of the population used in this study showed that the Spanish population is similar to other Holstein dairy cattle populations, as expected.

Accuracy

The accuracy and MSE of the estimated DGVs for each approach (B-Lasso and R-Boost) are shown in Table 1 for the ESP population. Traditional PI accuracies ranged from 0.386 to 0.460. The predicted DGVs showed higher accuracies than the PIs for all considered traits, with an average increment of 41%, ranging from 24% for UD to 59% for FY. Similar results have been previously reported in other studies using Holstein populations (VanRaden *et al.,* 2009; Moser *et al.,* 2010). Consequently, the selection of young animals based on genomic values may be preferable to selection based on traditional pedigree information as expected.

B-Lasso showed slightly higher Pearson correlations for MY (0.590), FY (0.655) and PY (0.583), whereas R-Boost showed larger values for UD (0.584).

MSE

Predicted MSE for each trait and method using the ESP population are shown in Table 1. B-Lasso showed higher (4%) MSE averaged across traits in comparison to R-Boost. In particular for UD, B-Lasso showed a 16% higher MSE than R-Boost. The machine learning algorithm was preferred for MY PY and UD while B-Lasso resulted in lower MSE for FY.

The differences between the methods were more remarkable in terms of estimated MSE than of accuracy. The estimated MSE for MY, and PY were larger for B-Lasso despite the fact that this method showed larger Pearson correlation. In a previous study, Verbyla *et al.* (2009) showed similar MSE when Bayesian approaches are compared with Genomic BLUP. Their results showed similar but still larger MSE than results of the Spanish population for FY and PY. The reason of these differences could be related to their smaller reference population size (1098 progeny tested bulls). The different methodologies, including non-parametric, implemented in this study showed similar predictive ability, although the best method was trait dependant. Further research is needed to determine the relationship between the kind of trait and the most suitable method for evaluation. B-Lasso showed to be preferable in terms of Pearson correlation. However, the methods that presented the highest Pearson correlation also showed large MSE. This should be considered in the model comparison when deciding the method with better predictive ability.

Regression coefficients

Regression coeficients ranged from 0.61 for the B-Lasso (UD) to 1.06 for the R-Boost (FY). With respect to unity, the R-Boost estimates showed closer estimates for MY(0.84), PY(0.85) and UD (0.82). B-Lasso had low coefficients for MY (0.72), PY (0.70) and UD (0.61).

Reference population size

In table 2, accuracy, MSE and regression coefficients of the estimated DGVs for R-Boost approach is shown for the EG population. Accuracies were similar to the ones obtained for a GBLUP approach with 15% poligenic effect but regression coefficients were much more favourable for R-Boost (Gonzalez-Recio, personal communication).

In figure 1, the gain in accuracy from comparing genomic estimates using the ESP reference population with the inclusion of the EG reference population.

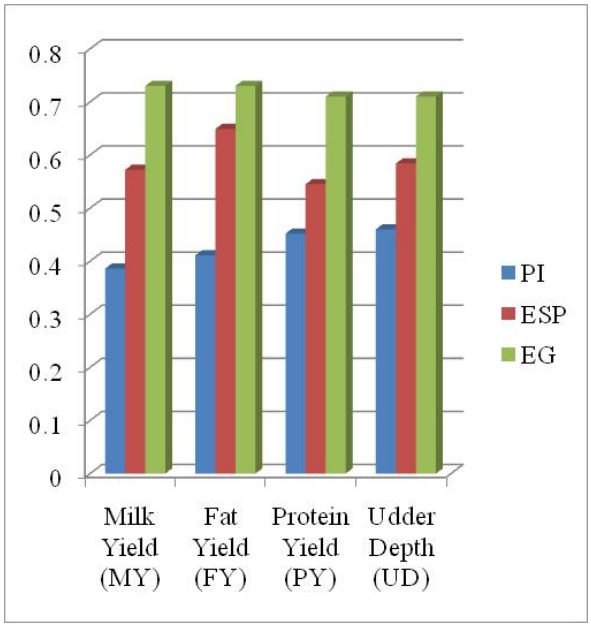

Figure 1. Accuracy of pedigree index (PI) and genomic estimates using the Random Boosting algorithm on both the Spanish (ESP) and Eurogenomics (EUR) reference populations.

Increments in accuracy due to a larger reference population ranged between 0.08 Pearson correlations units for FY to 0.17 for PY, resulting in predictions that are in average, 23% and 70% more accurate than those resulted with the national reference population and the traditional pedigree index respectively.

Conclusions

With the aim of improving the selection efficiency in both IA centers and commercial farms, GS has been implemented in the ESP breeding program. Identification of superior animals is therefore expected to be more accurate and feasible at younger ages than was previously possible. Research will continue on the reported traits and will be extended to the remaining traits included in the Spanish genomic evaluations.

The collaboration within the EG consortium, which includes a reference population with over 20,000 progeny-tested bulls, substantially increased the accuracy of genomic evaluations in the Spanish genome-assisted evaluations.

Acknowledgments

The authors acknowledge to CONAFE and EUROGENOMICS for providing genotypes and phenotypes, to IA Centers, and Farmers for providing biological samples used in this study, to “Dirección General de Producciones y Mercados Agrarios” and “Laboratorio Central de Veterinaria del Ministerio de Agricultura, Alimentación y Medio Ambiente” for support on the genotyping process, and funds from the project CDTI-P080250866 UPM and INIA-CC10-046,.

References

Banos, G. & Coffey, M.P. 2010. Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. *J. Dairy Sci. 93*, 2775–2778.

De los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. & Cotes, J.M. 2009. Posterior predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics 182*, 375–385.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Statist. 29*, 1189–1232.

Gianola, D., Fernando, R.L. & Stella, A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics. 173*, 1761-1776.

González-Recio, O., Weigel, K.A., Gianola, D., Naya, H. & Rosa, G.J.M. 2010. **L2-**Boosting algorithm applied to high-dimensional problems in genomic selection. *Genet.Res. (Camb). 92:3*, 227-37.

González-Recio, O. & Forni, S. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genet. Sel. Evol. 43:7*.

Hastie, T.J., Tibshirani, R. & Friedman, J. 2009. *The elements of statistical learning.* 2nd ed. Springer. New York, NY.

Long, N., Gianola, D., Rosa, G.J.M., Weigel, S. & Avendano, S. 2007. Machine learning

classification procedure for selecting SNPs in genomic selection: application to early mortality inbroilers. *J. Anim. Breed. Genet. 124,* 377–389.

Moser, G., Tier, B., Crump, R.E., Khatkar, M.S. & Raadsma, H.W. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol. 41:56.*

Moser, G., Khatkar, M.S., Hayes, B. & Raadsma,V. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genet. Sel. Evol.42:37.*

Park, T. & Casella, G. 2008. The Bayesian Lasso. *J. Amer. Stat. Soc. 103,* 681–686.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sontegard, T.S.G., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92,* 16–24.

Verbyla, K.L., Hayes, B.J., Bowman, P.J. & Goddard, M.E. 2009. Accuracy of genomic selection using stochastic variable selection in Australian Holstein Friesian dairy cattle. *Genet.Res. (Camb). 91,* 307-311.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Schenkel, F.S. & Van Tassell, C.P. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *J. Dairy Sci. 92,* 3431–3436.

Table 1. Accuracy, measured as Pearson correlation, mean square error (MSE) and regression coefficients for the genomic predictions of two different evaluation methodologies indexed using the Spanish reference population for four traits of economic interest in dairy cattle.

Methods[1]	Milk Yield (MY)	Fat Yield (FY)	Protein Yield (PY)	Udder Depth (UD)
Accuracy				
P. Index	0.386	0.411	0.452	0.460
B-Lasso	**0.590**	**0.655**	**0.583**	0.562
R-Boost	0.572	0.649	0.545	**0.584**
MSE				
B-Lasso	172328.60	**273.47**	143.38	0.69
R-Boost	**167063.40**	282.36	**141.84**	**0.58**
Regression coefficient				
B-Lasso	0.72	**0.89**	0.70	0.61
R-Boost	**0.84**	1.06	**0.85**	**0.82**

In bold: The preferred method within trait and comparison criteria.
[1]Methods: P. Index (Traditional pedigree index), B-Lasso (Bayesian Lasso), and R-Boost (Random Boosting)

Table 2. Accuracy, measured as Pearson correlation, mean square error (MSE) and regression coefficients for the genomic predictions of R-Boosting methodology using the EuroGenomics reference population for four traits of economic interest in dairy cattle.

Criteria	Milk Yield (MY)	Fat Yield (FY)	Protein Yield (PY)	Udder Depth (UD)
Accuracy	0.73	0.73	0.71	0.71
MSE	174786	290	175	0.69
Regression coefficient	0.97	0.98	0.96	0.93