EVALUACIÓN GENÉTICA (QUE NO GENÓMICA) MEDIANTE BOOSTING

Casellas, J.
Grup de Recerca en Remugants. Departament de Ciència Animal i dels Aliments. Universitat Autònoma de Barcelona, 08193 Bellaterra. joaquim.casellas@uab.cat

INTRODUCCIÓN

Desde que se derivaron los modelos BLUP (*best linear unbiased prediction*) a mediados del siglo pasado (Henderson 1950), esta herramienta analítica se erigió como la piedra filosofal de la mejora genética animal durante décadas. Los modelos BLUP se centran en la predicción del mérito (o demérito) genético del ganado para los caracteres productivos evaluados, corrigiendo también para otros efectos sistemáticos y aleatorios. Aunque la precisión de los valores genéticos se incrementa sustancialmente al incorporar información genómica a los modelos de evaluación (Meuwissen *et al*. 2001), los modelos BLUP clásicos aún resultan esenciales en la mayoría de los esquemas de selección dónde el genotipado masivo excede, de largo, el presupuesto disponible. Es por ello que resulta especialmente relevante la posibilidad de evaluar alternativas metodológicas al BLUP, con el objetivo de incrementar la precisión de los valores genéticos obtenidos a partir únicamente de la información fenotípica y genealógica.

El algoritmo de *boosting* ha sido recientemente propuesto como una alternativa no paramétrica en el campo de la selección genómica (González-Recio *et al*. 2010), proporcionando precisiones elevadas e incluso sesgos menores. No obstante, su utilidad dentro del ámbito de la evaluación genética (que no genómica) nunca ha sido evaluada. Es por ello que el objetivo principal de este trabajo fue adaptar el algoritmo de *boosting* a un marco de evaluación genética de reproductores, comparando sus resultados con los obtenidos mediante un modelo mixto lineal estándar.

MATERIAL Y MÉTODOS

El *boosting* es una técnica de aprendizaje automático propuesta por Schapire (1990) y Freund (1990); combina predictores moderadamente imprecisos (conocidos como *weak learners*) con el objetivo de generar una norma de predicción con mayor habilidad de predicción que cualquiera de los *weak learners* por separado. Este tipo de aproximaciones fueron desarrolladas originariamente para problemas de clasificación (Freund 1990; Schapire 1990), aunque pronto se generalizaron a los algoritmos de regresión (Friedman 2001), y recientemente realizaron las primeras incursiones en el campo de la mejora genética animal (Bühlmann 2006; González-Recio *et al*. 2010).

Este trabajo se centra en el algoritmo L_2-*boosting* (Freund & Schapire 1996), el cual evalúa el grado de incorrección de las predicciones a partir de la diferencia cuadrática media entre los datos reales y predichos. Además, se propone el siguiente *weak learner*,

$$y_j = \beta_i a_{ij} + e_j,$$

donde y_j es el registro fenotípico del individuo j (pre-corregido para las demás fuentes ambientales de variación), β_i es el coeficiente de regresión inherente al mérito genético del individuo i, a_{ij} representa el coeficiente de parentesco aditivo entre los individuos i y j, y e_j corresponde al residuo. Para una base de datos con s individuos, n de los cuales disponiendo de información fenotípica, el L_2-*boosting* se basa en tres pasos sucesivos que se repiten de manera iterativa hasta alcanzar un determinado criterio de convergencia. Durante el primer paso, el *weak learner* se aplica a cada individuo por separado, proporcionando coeficientes de regresión específicos para cada uno de ellos. El segundo paso requiere de la computación de residuos aplicando la fórmula

$$\mathbf{r}_i = \mathbf{y} - \Sigma_{j=1,n} [v(\beta_i a_{ij})],$$

asumiendo que $\mathbf{r}_i$ es el vector de residuos al aplicar el *weak learner* sobre el individuo i, $\mathbf{y}$ es el vector de fenotipos resultantes después de las correcciones efectuadas en las iteraciones previas (ver el tercer y último paso del procedimiento), y v es un parámetro de contracción fijado a 0,01 en el caso que nos ocupa. El cuadrado medio del error (CME) para el i-éssimo coeficiente de regresión se obtiene de la siguiente expresión,

$$\text{CME}_i = \Sigma_{j=1,n} r_i^2$$

y caracteriza el ajuste del *weak learner* para el individuo *i*. Este paso proporciona un total de *s* estimas de CME_i distintas, evaluando la contribución de cada individuo incluido en la genealogía. El tercer y último paso implica la selección del individuo que minimiza el CME (asumimos que es el individuo *k*), y los datos fenotípicos de actualizan aplicando $\mathbf{y} = \mathbf{r}_k$.

Para evaluar el comportamiento predictivo del L_2-*boosting* se simularon poblaciones de 2.000 individuos con 10 generaciones no solapadas de 10 machos y 190 hembras bajo apareamiento aleatorio. Se simuló un registro fenotípico para todos los individuos, incluyendo la influencia de dos efectos fijos (2 y 10 niveles respectivamente; efectos obtenidos de una distribución uniforme entre -1 y 1) y el efecto genético aditivo aunque bajo tres escenarios distintos (h^2=0,1, 0,2 y 0,4; σ_e^2=1). Dado que la convergencia del procedimiento se evaluó mediante validación cruzada, se seleccionó un 10% de los individuos nacidos en las últimas cinco generaciones y la población se subdividió en el set de entrenamiento (90%) y el de ajuste (10%). El procedimiento de L_2-*boosting* se aplicaba sobre el set de entrenamiento, mientras que el CME se calculaba en el set de ajuste al final de cada iteración. El algoritmo finalizaba cuando la reducción en el CME después de cada iteración era inferior a 10^{-5}.

Los resultados se compararon con los obtenidos bajo un modelo BLUP estándar resuelto mediante inferencia Bayesiana. En este caso, se realizó el análisis descartando el fenotipo del 10% de individuos seleccionados en la últimas 5 generaciones, y se calculo el CME resultante sobre el set de ajuste al finalizar el análisis. En ambos casos se calculó también la diferencia cuadrática media entre los valores genéticos simulados y predichos para los individuos del set de ajuste. En total se realizaron 10 réplicas para cada escenario de heredabilidad.

RESULTADOS Y DISCUSSIÓN

Las propiedades de predicción del L_2-*boosting* se resumen en las figuras 1 y 2 de este trabajo; a pesar que se restringen a las réplicas con h^2=0,1 por limitaciones de espacio, los demás escenarios siguieron comportamientos muy parecidos. Aunque, en promedio, el CME alcanzó valores ligeramente inferiores para el L_2-*boosting* (1,133 ± 0,035) que para el BLUP clásico (1,128 ± 0,034), no se observó una patrón constante en la figura 1, sugiriéndose que ambas aproximaciones se comportaban de manera muy parecida y que las ligeras ventajas obtenidas se debían muy probablemente al azar. De manera parecida, la diferencia cuadrática media entre los valores mejorantes simulados y predichos fue muy similar bajo ambas aproximaciones y sin mostrar una tendencia clara a favor de una o la otra.

En conclusión, el L_2-*boosting* se sugiere como una alternativa interesante a los modelos BLUP para la evaluación genética cuando existen individuos sin información fenotípica; esta situación se da en la mayoría de sistemas productivos ganaderos al tener que seleccionar la reposición antes incluso de que aporten sus primeras producciones (p.ej., prolificidad en cerdas y conejas, producción de leche en rumiantes, calidad de la carne). En cualquier caso, representa un punto de partida muy interesante para el futuro desarrollo y evaluación de otros *weak learners*, con el objetivo de mejorar esta aproximación no paramétrica que, ya inicialmente, aporta una habilidad de predicción comparable a la de los modelos BLUP.

REFERENCIAS BIBLIOGRÁFICAS

• Bühlmann, P. 2006. Ann. Stat. 34: 559-583 • Freund, Y. 1990. Proc. 3th Ann. Workshop Comput. Learning Theory, Rochester, NY, 202-216 • Freund, Y. & Schapire, R. E. 1996. Proc. 13th Int. Conf. Machine Learning, San Francisco, CA, 148-156 • Friedman, J. H. 2001. Ann. Stat. 29: 1189-1232 • González-Recio, O., Weigel, K. A., Gianola, D., Naya, H. & Rosa, G. J. M. 2010. Genet. Res. 92: 227-237 • Henderson, C. R. 1950. Ann. Math. Statist. 28: 276-290 • Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. 2001. Genetics 157: 1819-1829 • Schapire, R. 1990. Mach. Learn. 5: 197-227 •

Agradecimientos: Esta investigación se enmarca dentro del proyecto AGL2010-21176/GAN financiado por el Ministerio de Economía y Competitividad; el contrato de J. Casellas se vincula al programa “Ramon y Cajal” (RYC-2009-04049).

Figura 1. *Cuadrado medio del error para el 10% de individuos asignados al set de ajuste (su fenotipo no contribuye al análisis) y obtenido mediante BLUP clásico (barras blancas) o boosting (barras negras).*

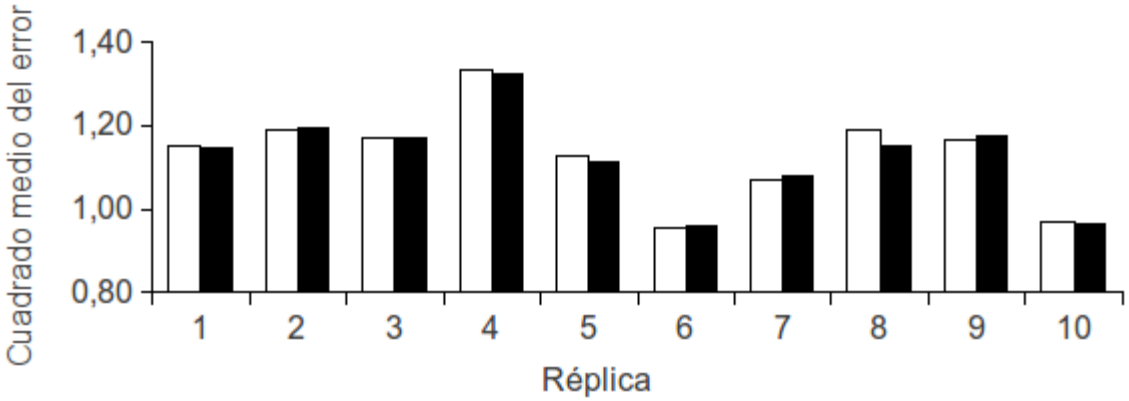
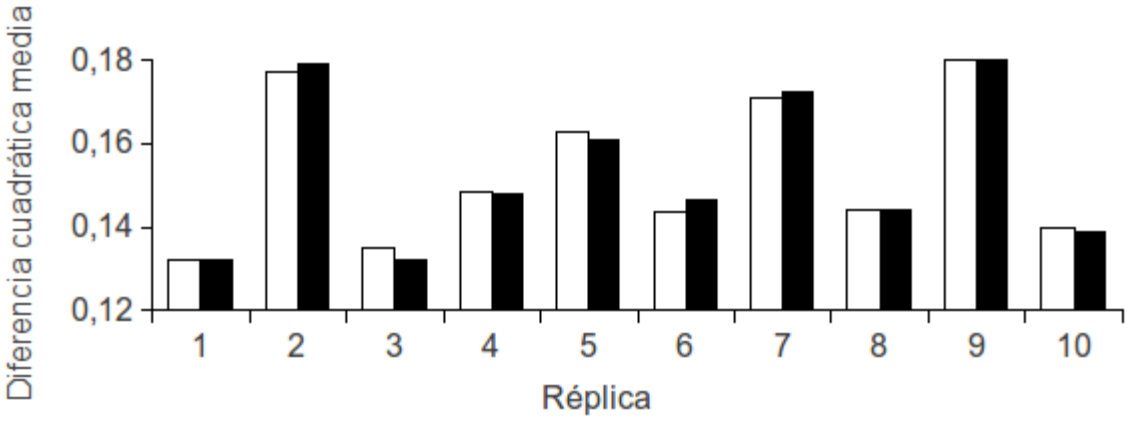


Figura 2. *Diferencia cuadrática media entre los valores mejorantes simulados y predichos para el 10% de individuos asignados al set de ajuste (su fenotipo no contribuye al análisis) y obtenido mediante BLUP clásico (barras blancas) o boosting (barras negras).*



GENETIC EVALUATION BY BOOSTING

ABSTRACT: This research focuses on the adaptation of the boosting algorithm to perform genetic evaluation of livestock populations. Boosting combines rough and moderately inaccurate predictors known as “weak learners” into a prediction rule with potentially greater predictive ability than that of any of the individual weak learners. This approach was evaluated on simulated data sets and compared with mixed linear animal models solved through a standard Bayesian approach. Prediction abilities of both boosting and BLUP approaches were similar, without a clear pattern favoring one or the other method. In a similar way, mean square differences between simulated and predicted breeding values were almost equal, and slight advantages were randomly obtained by the boosting or the BLUP approach depending on the analyzed data set. This suggested that boosting must be viewed as an appealing alternative when selection decisions must be taken on individuals without phenotypic data.

Keywords: artificial intelligence, accuracy, boosting, genetic evaluation