Evaluación de diferentes métodos de predicción del RFI a partir del genotipo utilizando algoritmos de aprendizaje automático

M.Mora[1], P.González [2], JR.Quevedo[2], E.Montañes[2], L.Tusell[1], M.Piles[1]

[1]Instituto de Investigación y Tecnología Agroalimentaria (IRTA), IRTA Torre Marimon s/n, 08140, Caldes de Montbui, España
monica.mora@irta.cat (Correspondencia a Mónica Mora)
[2]Centro de Inteligencia Artificial, Universidad de Oviedo, 33204, Gijón, España

Resumen

Si se dispone del genotipo, la predicción de la eficiencia alimentaria de un individuo puede hacerse a edades muy tempranas y sin la necesidad de las medidas de los caracteres implicados. En este estudio se probaron distintos métodos de aprendizaje automático para la predicción del consumo del pienso residual (RFI) de cerdos en crecimiento a partir de 45610 SNPs. Los datos correspondían a 5828 animales con medidas de consumo (FI) y crecimiento (ADG) diarios, espesor de la grasa dorsal y peso metabólico. El objetivo fue comparar la calidad de la predicción del RFI obtenida (*i*) utilizando únicamente los SNPs como variables predictoras; (*ii*) obteniendo predicciones simultáneas de FI, ADG, BF, MW en un modelo de regresión múltiple (*Multioutput*) para calcular el RFI usando las variables predichas a partir de los SNPs y (*iii*) utilizando las predicciones individuales de FI, ADG, BF, MW y los SNPs como variables predictoras del RFI (*Staking*). Los algoritmos utilizados para implementar los 3 modelos fueron *Random Forest* (RF) y *Support Vector Regression* (SVR), excepto para *Multioutput* en el que solo se implementó RF. Se aplicó un muestreo anidado con una validación cruzada (CV) externa en 10 partes. La optimización de parámetros fue llevada a cabo mediante una búsqueda en rejilla (*grid search*) utilizando una CV de 3 partes. La evaluación de la calidad de los modelos se efectuó con distintas métricas de regresión como el error absoluto y relativo, medidas de ranking como la correlación de Spearman y dos medidas de utilidad. Estas últimas miden la calidad de la clasificación en base a la pertenencia al grupo del 10% de los mejores candidatos a la selección (*TopRank*). Si un individuo se clasifica en el *TopRank* real su pérdida es 0 y si no hay dos variantes: *i*) la pérdida es igual a 1 (*TopRank01*) y *ii*) la pérdida es la distancia entre la posición predicha y la real dividida entre el número de candidatos a la selección (*TopRankDistance*). Este esquema se implementó utilizando como variables predictoras diferentes subconjuntos con un número creciente de los SNPs más informativos identificados con RF (de 200 a 3000). Los resultados muestran que las predicciones obtenidas con RF fueron siempre mejores que las obtenidas con SVR. Por otra parte, la predicción directa del RFI resultó ser la mejor estrategia para todas las métricas, con la que, por ejemplo, se obtuvo un error relativo de 0,97, una correlación de *Spearman* de 0,24 y una *TopRank01* y *TopRankDistance* de 0,83 y 0,34 utilizando RF y un subconjunto de los 1200 SNPs más informativos. La utilización de métodos *multioutput* o *Staking* no supuso una mejora en ningún caso. Con ellos el error relativo fue de 0,98 y 0,99, y la correlación de Spearman de 0,20 y 0,18, respectivamente. La *TopRankDistance* fue similar y alrededor de 0,36 para el *Multioutput* y *Staking*.

Palabras claves: eficiencia alimentaria, predicción, multioutput, SNPs, Staking