BOOSTING AS A MARKER SELECTION STRATEGY FOR GENOMIC PREDICTION WITH ARTIFICIAL NEURAL NETWORKS

XIII International Symposium on Genetics and Breeding, 13ª edição, de 25/10/2022 a 27/10/2022
ISBN dos Anais: 978-65-5465-014-4

BARRETO; Cynthia Aparecida Valiati [1], CELERI; Maurício de Oliveira [2], BARBOSA; Wagner Faria [3], LIMA; Leísa Pires [4], SILVEIRA; Lucas Souza da [5], NASCIMENTO; Ana Carolina Campana [6], AZEVEDO; Camila Ferreira [7], NASCIMENTO; Moysés [8]

RESUMO

Artificial neural network (ANN) is a non-parametric tool that has been used for prediction of genomic estimated breeding values (GEBVs). However, ANNs demand high computational costs to fit the prediction models, which may limit the use of this methodology. The purpose of this study is to present a two-step genomic prediction approach. In the first step, an indirect marker selection was performed by boosting by selecting 5% and 20% of the most important markers. After that, in the second step, the selected markers were used as input variables in the ANN model. The results obtained were compared with ANN without marker selection. The ANN topology used was two hidden layers with a number of neurons ranging from one to 20, with a logistic activation function and considering the backpropagation algorithm. The data were simulated considering the mean dominance degree equal to zero and genetic architecture with 100 QTLs controlling the trait and heritability 0.2. To evaluate the fit, a 5-fold cross-validation was performed, with 800 observations used for training and 200 observations for validation. Marker selection was performed in each training set. The highest accuracy (0.98±0.00) was observed when indirect selection of the top 5% markers was performed, followed by selecting the top 20% markers (0.96±0.00) and the network without marker selection (0.85±0.00). This is partly due to the reduction of the search space which optimizes the learning process and improves the model's predictive ability. The processing time ranged from 3.55s to 38.59s for prediction for the fit by selecting the top 5% markers and for the ANN without prior marker selection, respectively. The marker selection generated a sparse set of markers covering a large part of the genome. The results obtained suggest that boosting is an efficient methodology for marker selection and the two-step prediction model stood out as an alternative to solve the problem of the high computational cost of ANNs.

PALAVRAS-CHAVE: genomic selection, markers selection, machine learning

[1] Universidade Federal de Viçosa – Departamento de Biologia Geral, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com
[2] Universidade Federal de Viçosa – Departamento de Estatística, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com
[3] Universidade Federal de Viçosa – Departamento de Estatística, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com
[4] Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais – Campus Muriaé, Avenida Coronel Monteiro de Castro, 550 – 36884-036 – Muriaé, MG – Brasil, cynthiavaliatibarreto
[5] Universidade Federal de Viçosa – Departamento de Estatística, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com
[6] Universidade Federal de Viçosa – Departamento de Estatística, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com
[7] Universidade Federal de Viçosa – Departamento de Estatística, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com
[8] Universidade Federal de Viçosa – Departamento de Estatística, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG – Brasil, cynthiavaliatibarreto@gmail.com