RANDOM FOREST AND DISCRIMINANT ANALYSIS BY PARTIAL LEAST SQUARES ON NIR DATA FOR PHENOTYPIC CLASSIFICATION OF SUGARCANE CLONES

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

ANDRADE; Andréa Carla Bastos [1], DIAS; Cristina Silva [2], PETERNELLI; Luiz Alexandre [3]

RESUMO

The growing demand for biomass for power generation and second-generation ethanol has driven the selection of sugarcane cultivars with higher fiber and apparent sucrose levels. In this sense, it is crucial to seek classification methods combined with near-infrared spectroscopy (NIR) to facilitate the desired selection. The present work aims to compare two models on NIR spectroscopy data to evaluate the predictive performance in sugarcane clones. A set of NIR data composed of 460 samples was used, classified as high and low levels of fiber (FIB) and apparent sucrose (PC). To evaluate the classification methods, the latent variable parameters (nVL) and number of predictors (m) were chosen for PLS-DA and RF, respectively. K-fold cross-validation was used to evaluate the classification methods with k=10 in the calibration set samples to help choose the values of the parameters associated with each model. The parameters chosen were those that presented a lower estimate of the classification error in the cross-validation. It was possible to obtain the confusion matrix for each model and calculate the corresponding classification errors, sensitivity, and specificity. We used some statistical tests to compare the methods and verify which performs best for classifying the clones based on their NIR spectra. The results obtained in comparing the methods indicated that PLS-DA differed from RF for the classification of properties %PC and %FIB ($p < 0.05$). We also evaluated these methods' classification errors, sensitivity, and specificity. PLS-DA was more satisfactory for all these parameters than RF since the formers presented lower classification errors and higher sensitivity and specificity values. Therefore, these methods can be considered helpful in classifying the NIR spectroscopy data used in this work. For a more accurate analysis, it would be interesting to test these methods in other NIR datasets and compare them with the results obtained in this work.

PALAVRAS-CHAVE: Near infrared spectroscopy, Sugarcane selection, Supervised learning

[1] Universidade Federal de Viçosa, andrea.andrade@ufv.br
[2] Universidade Federal de Viçosa, cristinadias360@gmail.com
[3] Universidade Federal de Viçosa, peternelli@ufv.br