doi:10.20396/revpibic.v0i0.id

XXVI Congresso de Iniciação Científica Unicamp
17 a 19 de outubro Campinas | Brasil

LEARNING CURVES FOR THE MODELING OF SUGAR CANE PRODUCTIVITY

Vitor H. Nisieimon*, Luiz H. A. Rodrigues, Felipe F. Bocca, Matheus Ferraciolli.

Abstract
Predicting the final yield of a crop is one of the most important aspects of a mill's agricultural planning. However, numerous factors must be considered to ensure a realistic forecast. Data mining techniques are capable of generating models that predict these values by relating a large amount of data. In this project, we studied learning curves, a tool used in the analysis of a model's performance according to the amount of data available. In an analysis of a database for a sugarcane production, we compared three different modeling techniques, suitable for regression models in the prediction of the final productivity.

Key words:
Machine Learning, Learning Curves, Sugarcane

Introduction

With the advent of Agriculture 4.0, one of the main sectors of Brazilian agriculture, the sugar-energy sector, has increasingly sought the implantation of technologies, mainly sensors for the monitoring of production. Such equipment, however, generates an immense amount of data, making it difficult to handle the information generated for analysis of the productive cycle as well as productivity.

In OLIVEIRA, BOCCA and RODRIGUES (2017)[1] several data mining techniques were used for modeling with 2 years of production data in order to predict the final productivity of sugarcane, which presents itself as an alternative with great potential for an analysis with a large amount of data.

In the performance analysis of the models, however, the adequacy of the data volume to the proposed modeling was not verified. In PERLINCH, PROOST and SIMONOFF (2003)[2] this analysis was made by using learning curves comparing two different techniques and, as a result, it was realized that increasing the amount of data in a modeling is not always meant for good performance.

In this work, therefore, we analyzed the modeling of a database for a sugarcane production to predict the final productivity, through the learning curves in order to verify the adequacy of techniques and models for the volume of data available.

Results and Discussion

Statistical software R was used in the development of all steps of the project. We used three different techniques: Support Vector Machine (SVM), Gradient Boosting Machine (GBM) and Random Forest (RF), which are suitable for the proposed regression problem.

By increasing the amount of data in the training of the model from 20% to 100% of the available data, by 5%, and evaluating the error in the same test set, we obtained the learning curves presented in Figure 1.

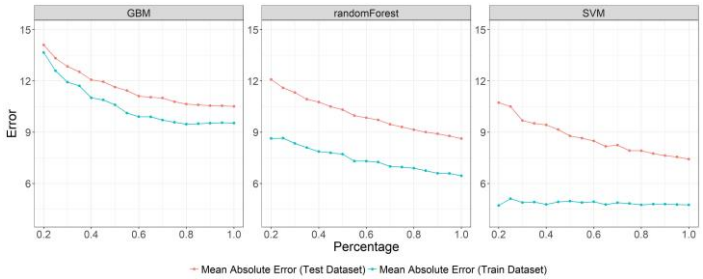

Figure 1. Learning curve for the three techniques used.

Analyzing the curves, therefore, it is perceived that, as in PERLICH, PROVOST and SIMONOFF (2003)[2], the curves begin to stabilize in a certain error value demonstrating, thus, the inefficiency in increasing the amount of data in the modeling from there. It occurs when approximately 80% of the data available to the modeling are used for the three techniques. In addition, it is still possible to analyze an important characteristic of the models: the trade-off between bias and variance. The bias can be obtained through the error dimension presented in the curves, while the variance can be obtained through the distance between the blue and red curves. This shows that the GBM presented the highest bias and the lowest variance, while the SVM presented the lowest bias and the highest variance and the randomForest remained between the previous two.

Conclusions

Since in the sugarcane production scenario it is common to find several different conditions, from mills that have high volume of historical data to those in opposite situation, evaluating and comparing the performances between the modeling techniques is crucial to choose the one that best fits a particular scenario and then to be able to support decision-makers.

[1] OLIVEIRA, M. P. G. DE; BOCCA, F. F.; RODRIGUES, L. H. A. From spreadsheets to sugar content modeling: A data mining approach. Computers and Electronics in Agriculture, v. 132, p. 14–20, 2017.

[2] PERLICH, C.; PROVOST, F.; SIMONOFF, J. S. Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. Journal of machine learning research: JMLR, v. 4, p. 211–255, 2003.