Computers and Electronics in Agriculture 128 (2016) 67–76

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling

Felipe F. Bocca, Luiz Henrique Antunes Rodrigues *

School of Agricultural Engineering, University of Campinas, Brazil

ARTICLE INFO

Article history:
Received 23 November 2015
Received in revised form 16 August 2016
Accepted 18 August 2016
Available online 29 August 2016

Keywords:
Artificial neural networks
Boosted regression trees
Machine learning
Random forest
Regression trees
Support vector machines

ABSTRACT

Crop yield models can assist decision makers within any agro-industrial supply chain, even with regard to decisions that are unrelated to the crop production. Considering the characteristics of the mechanisms and data related to yield, data mining techniques are suitable candidates for modelling. The use of these techniques within a context with feature engineering, feature selection, and proper tuning can further improve performance beyond a simple replacement of multiple linear regression. To evaluate the impact of the different steps in the mentioned context, we evaluated sugarcane (*Saccharum* spp.) yield modelling with data obtained from a sugarcane mill. For a combination of six techniques, tuning, feature selection, and feature engineering, leading to 66 combinations, we assessed final model performance. Average performance across combinations resulted in a mean absolute error (MAE) of 6.42 Mg ha^{-1}. Using different techniques led to a range of MAE from 4.57 to 8.80 Mg ha^{-1} on average. The best and worst performances for an individual model were MAEs of 4.11 and 9.00 Mg ha^{-1}. Models with lower performance were close to simply predicting yield from the average yield for each number of cuts (MAE of 9.86 Mg ha^{-1}). Tuning and feature engineering reduced the MAE on average by 1.17 and 0.64 Mg ha^{-1}, respectively. Feature selection removed nearly 40% of the features but increased the MAE by 0.19 Mg ha^{-1}. The performance of models was improved by simple strategies such as decomposing weather attributes and detailing fertilisation. Evaluation of feature importance provided by the RReliefF feature selection algorithm was used to explain the performance gains. If empirical models are needed, they will rely on using advanced techniques, but they will need proper algorithm tuning and feature engineering to extract most of the information from datasets. Based on the results, we recommend following the presented workflow for the development of yield models.

1. Introduction

Empirical sugarcane (*Saccharum* spp.) yield models, as for any crop, can help several decision makers. Crop models coupled with numerical weather prediction models may benefit whole value chains in agricultural production if developed as part of an interdisciplinary and goal-oriented approach (Stone and Meinke, 2005). Considering the goal of decision making, models can be simply a functional relation between crop inputs and outputs. The use of such models should be preferred over using averaged information for planning purposes (Higgins et al., 2007), and estimations should address the stochastic nature of those estimations over a deterministic approach (Ahumada and Villalobos, 2009). Usually, neither aspects are addressed in sugarcane production planning (Bocca et al., 2015).

Modelling agricultural outputs as a function of its inputs is often referred as empirical modelling, in opposition to the mechanistic/process-based models such as APSIM (Keating et al., 2003), DSSAT (Jones et al., 2003), CROPSYST (Stöckle et al., 2003), WOFOST (Van Ittersum et al., 2003) or EPIC (Brisson et al., 2003), in which different processes related to the plant development are simulated for describing plant growth. Often the two approaches are compared concerning their aims: empirical models are used just to provide a straight answer to a specific need, usually at a specific location, and mechanistic/process-based models are used to explain and simulate crop development. Examples of process-based models for sugarcane can be viewed in Marin and Jones (2014) and Stokes et al. (2016). Empirical models are limited to the scope of the data used in their creation, and process-based models are limited to the scope where their processes are valid, which is often considered a broader scope in comparison to empirical based models. Regarding data requirements, empirical models

* Corresponding author.
E-mail address: lique@feagri.unicamp.br (L.H.A. Rodrigues).

http://dx.doi.org/10.1016/j.compag.2016.08.015
0168-1699/

are often considered alternatives to the more specific data requirements of mechanistic models.

Considering the typical data available in sugarcane mills (Lawes and Lawn, 2005), the use of data mining techniques for empirical modelling should be a natural decision: data mining techniques are robust to noise, auto-correlation of features and can work with different data types. These characteristics describe most of the agricultural, soil, and weather data and their relations. While these techniques bring some new capabilities, their application should happen in a context where the model training/induction is only one step. Examples of the full context are the academic formulation of Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996) or the industrially focused Cross Industry Process for Data Mining – CRISP-DM – and the "Sample, Explore, Modify, Model, Assess" - SEMMA. Considering the analogy between these methodologies (Azevedo and Santos, 2008), this paper will focus on the KDD terminology and context. For academic publications, steps[1] like "understanding the domain", "understanding the goal", "creating the dataset", and "data preprocessing" are implicit or the main objective. But other steps of KDD might not be receiving the proper attention. This paper will focus on two aspects of the fourth step, "data reduction and projection", along with the methodology of data mining (the sixth step), specifically:

- Feature selection (FS): feature importance evaluation and selection.
- Feature engineering: the creation of features derived from original features.
- Model tuning: searching for algorithm hyperparameters for better performance.

To contextualise the effects of such practices in yield modelling, 13 papers were selected as a result of a non-comprehensive review, focusing on papers in which machine learning/data mining techniques were applied to yield modelling in the following context: data related to the production is used to predict final yield. Analysing the techniques used in the papers (Table 1), most (10/13) used Artificial Neural Networks (ANN), often (4/13) including Regression Trees (RT). Some papers (4/13) perform comparisons with Multiple Linear Regression (MLR), or some related model such as regression with higher order polynomials (Alvarez, 2009) or Generalised Linear Models (GLM; Zheng et al., 2009). ANN always performed better than GLM or MLR, which is expected, given the higher modelling capacity of ANNs. RT was also better than MLR and GLM, which is also expected. Other techniques were only used in two papers, namely Random Forest (RF), Support Vector Machines (SVM), Nearest Neighbour (kNN), M5-Prime, and Radial Basis Function (RBF) network. Ruß (2009) showed a better result for the SVM technique, followed closely by RBF. Since Gonzalez-Sanchez et al. (2014) did not perform tuning consistently across techniques, their results concerning which technique is better are not useful. As Macià and Bernadó-Mansilla (2014) pointed out, for benchmarking, all of the algorithms should be optimised for a fair comparison. The authors also reinforce the "no free lunch" theorem in machine learning for such comparisons, as the superiority of an algorithm in a set of problems cannot be generalised. At most, conclusions such as those reached by Fernández-Delgado et al. (2014) regarding classification problems (discrete or categorical labels) can lead to rules of thumb about techniques to prioritise evaluations. From the analysis of a large body of datasets, they show that RF is most likely the best classifier with no statistical difference to SVM with Gaussian Kernel. From the perspective of families of classifiers, the best families presented by the authors are, in descending order, RF, SVM, ANN, Boosting Ensembles, and C5. From these results, if model performance is the only goal, the best practice is, therefore, to evaluate several techniques.

Table 1
Techniques used in reviewed papers. The technique with the best performance in the paper is marked with an asterisk.

Paper	Technique
Kaul et al. (2005)	ANN
Park et al. (2005)	ANN*, RT, GLM
Zhang et al. (2005)	RT
Zhang et al. (2006)	RT*, MLR
Ji et al. (2007)	ANN*, MLR
Alvarez (2009)	ANN*, MLR (second order)
Ruß (2009)	ANN, RBF, RT, SVM*
Zheng et al. (2009)	RT*, GLM
Dai et al. (2011)	ANN
Gonzalez-Sanchez et al. (2014)	ANN, MLR, RT, SVM, kNN*, M5P*
Haghverdi et al. (2014)	ANN*, RT
Matsumura et al. (2014)	ANN*, MLR
Thuankaewsing et al. (2015)	ANN

Data mining techniques were applied to model yield on different scales from district scale (e.g. Alvarez, 2009; Matsumura et al., 2014) to precision agriculture scale (Ruß, 2009) (see Table 2). Most of the 13 studies focused on a single crop, and the data used for modelling came from experiments (6). Automated FS (for every algorithm) was only used by Alvarez (2009) with the conventional approach of stepwise regression for FS. Zhang et al. (2005) performed a Principal Components Analysis to transform data and used stepwise FS for MLR. Explicitly manual FS was performed by five out of thirteen papers, while one paper (Gonzalez-Sanchez et al., 2014) performed an exhaustive search. The FS effect can be negatively biased in this context. In most experiments conducted, researchers collect data that are supposedly related to the phenomenon of interest, given resource and/or time constraints on the collection and analysis of data. The oriented collection of data means that these kinds of datasets have only pre-approved features. In the context of automated data collection, use of new sensors, evaluation of different measures or timing, FS can have a greater impact. In these cases, FS can enhance model quality by discarding bogus features or simply decreasing the model and computational complexity by keeping the most important features, with one example being Ruß and Kruse (2010). Haghverdi et al. (2014) performed sensitivity analysis to analyse the features of neural networks but did not use this information to select features. Examples of features engineered are the different aggregations for weather data in Ji et al. (2007), Zhang et al. (2006), Kaul et al. (2005) and Zhang et al. (2005) and fertilisation in Park et al. (2005) and Matsumura et al. (2014). In the mentioned papers, feature engineering is evident and justified by agronomic knowledge, but no paper presented an evaluation of feature importance before deciding on the inclusion of created features.

Regarding the tuning procedures, most of the papers that used ANN performed tuning for the neural network, often changing the number of nodes and seeds for initialization. Overall (10/13), the papers performed some tuning (Table 2). Zhang et al. (2005) used a suggestion for tree parameters based on the dataset size. Gonzalez-Sanchez et al. (2014) used different approaches for each algorithm, from setting k equal to 5 in kNN and using topology found in previous studies for ANN to manual testing and the values from Ruß (2009) for SVM cost with a linear kernel. One paper (Haghverdi et al., 2014) only tuned the ANN and not the RT.

[1] The KDD steps presented by Fayyad et al. (1996) are: 1 – developing an understanding of the application domain and understanding the goal, 2 – creating the target data set, 3 – data cleaning and preprocessing, 4 – data reduction and projection, 5 – matching the goals of the KDD process (step 1) to a particular data-mining method, 6 – exploratory analysis and model and hypothesis selection (model's parameters decisions are made in this step), 7 – data mining, 8 – interpreting mined patterns and 9 – acting upon discovered knowledge.