Chemometrics and Intelligent Laboratory Systems 181 (2018) 52–63

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Which regression method to use? Making informed decisions in "data-rich/knowledge poor" scenarios – The Predictive Analytics Comparison framework (PAC)

Ricardo Rendall, Marco S. Reis *

CIEPQPF, Department of Chemical Engineering, University of Coimbra, Rua Sílvio Lima, 3030-790, Coimbra, Portugal

ARTICLE INFO

Keywords:
Predictive analytics
Variable selection
Penalized regression methods
Latent variables methods
Tree-based ensembles methods
Manufacturing 4.0
Big data

ABSTRACT

In the big data and Manufacturing 4.0 era, there is a growing interest in using advanced analytical platforms to develop predictive modeling approaches that take advantage of the wealthy of data available. Typically, practitioners have their own favorite methods to address the modeling task, as a result of their technical background, past experience or software available, among other possible reasons. However, the importance of this task in the future justifies and requires more informed decisions about the predictive solution to adopt. Therefore, a wider variety of methods should be considered and assessed before taking the final decision. Having passed through this process many times and in different application scenarios (chemical industry, biofuels, drink and food, shipping industry, etc.), the authors developed a software framework that is able to speed up the selection process, while securing a rigorous and robust assessment: the Predictive Analytics Comparison framework (PAC). PAC is a systematic and robust framework for model screening and development that was developed in Matlab, but its implementation can be carried out on other software platforms. It comprises four essential blocks: i) Analytics Domain; ii) Data Domain; iii) Comparison Engine; iv) Results Report. PAC was developed for the case of a single response variable, but can be extended to multiple responses by considering each one separately. Some case studies will be presented in this article in order to illustrate PAC's efficiency and robustness for problem-specific methods screening, in the absence of prior knowledge. For instance, the analysis of a real world dataset reveals that, even when addressing the same predictive problem and using the same response variable, the best modeling approach may not be the one foreseen *a priori* and may not even be always the same when different predictor sets are used. With an increasing frequency, situations like these raise considerable challenges to practitioners, underlining the importance of having a tool such as PAC to assist them in making more informed decisions and to benefit from the availability of data in Manufacturing 4.0 environments.

1. Introduction

With the emergence of Manufacturing 4.0, advanced predictive analytics, and regression methods in particular [1–3], have been attracting considerable interest in many areas of science and in different application contexts, such as market analysis [4–6], manufacturing [7,8], food and beverage [9–11], pharmaceutical [12–14], petrochemical and chemical [15,16], etc., due to the growing availability of data collected from fast and informative process sensors as well as large databases that facilitate their storage, integration and retrieval. In this context, research on predictive methods has been driven by the need to develop suitable techniques equipped with the necessary methodological, algorithmic and computational components that allow them to cope with the prevalent characteristics observed in the collected datasets, such as high-dimensionality [17,18], collinearity [19,20], sparsity [21], nonlinearity [22], non-stationarity [23], missing data [24,25], among others. This effort led to the proliferation of a large number of methods and variants, spread through the vast technical literature, making it very difficult for practitioners to decide which methods best suit their particular application scenarios. Prior knowledge could be useful for selecting a suitable set of regression methods to adopt, but most often the particularities of each case study and the lack of more detailed information make it impossible to rule out other methods from the pool of candidates. Therefore, the prevalent and most realistic, honest, and

* Corresponding author.
E-mail address: marco@eq.uc.pt (M.S. Reis).

https://doi.org/10.1016/j.chemolab.2018.08.004
Received 21 August 2017; Received in revised form 3 August 2018; Accepted 12 August 2018
Available online 17 August 2018

unbiased perspective one often is forced to accept is the lack of absolute certainty about the best class of predictive methodologies to use (not to speak, the best method to use). We call this a "data-rich/knowledge poor" scenario, given the availability of data resources and, simultaneously, the lack of consistent information on how to derive the best predictive models from them.

In this context, comparison studies are unavoidable and represent a reliable way to test and select regression approaches that are eligible for predicting the response variable of interest. However, these studies take considerable time to carry out and require resources of knowledge, software and time that many users do not have at their disposal or simply cannot afford. Even for the few cases where they were conducted, they still present limitations in the number of methods tested (usually less than 5 [26], leaving some classes absent from analysis [27,28]) as well as in the way the comparison is done (namely in the accuracy and robustness of the approach and metrics used).

Therefore, it is the purpose of this work to put forward a platform for problem-specific methods screening, called, Predictive Analytics Comparison framework (PAC), that is able to speed up the selection process, while securing a rigorous and robust assessment of the methods under analysis and a proper use of the data available. This framework was developed and tested by the authors in different contexts (chemical industry, biofuels, drink and food, shipping industry, etc.), leading to consistent results that improved the base predictive solution adopted, with a very short implementation time. PAC was designed to help practitioners identifying the approaches with higher performance potential for their particular applications, using a structured, rigorous and informative methodology: the methodology is structured, because it is composed by four integrated components (see below); it is rigorous, because the comparison is conducted with a state of the art double cross-validation method that generates information for conducting formal statistical hypothesis tests which finally lead to a sound assessment of the methods' relative performances; finally, it is informative, because PAC will not only provide a report with results about the hierarchy of methods that best suit the application under analysis, but also present which predictors are more relevant in each class of methods, contributing to enrich the knowledge about the problem under analysis, among other interpretational information on the structure of data.

More specifically, PAC is composed by four components (Fig. 1): i) Analytics Domain; ii) Data Domain; iii) Comparison Engine; iv) Results Report. The Analytics block encompasses a rich variety of predictive approaches to be scanned in each application context under analysis. It establishes the analytics domain of assessment or comparison. The variety of methods was carefully considered by the authors. They are segmented in four classes: variable selection, penalized regression, latent variable, and tree-based ensemble methods. Each class of methods has different *a priori* assumptions regarding the data generating mechanism and their suitability depends on data features such as the level of sparsity (in a sparse problem, only a few variables have predictive value), collinearity (existence of associations among regressors), modularity (presence of block-wise structure in the regressors) and the underlying relationship between predictors and response variable (if it is linear or some non-linearity is present).

The Data Domain regards the dataset that will be used to conduct the comparison study and that determines the inference-basis of the study. It should be carefully considered because the results will be critically dependent on what is inserted into this component. In predictive problems, attention should be paid to the existence of clustered data, multiple processes/phenomena superimposed, transcription errors, outliers, signal to noise ratio in the response, etc. – this module is subject to the well-known GIGO principle of computer science ("garbage in garbage out"), which impacts the entire PAC framework.

The Comparison Engine performs a robust assessment of the predictive capabilities of each representative in the Analytics Domain, using the Data Domain as the inference basis. The computations are conducted in such a way as to potentiate an optimized generation of performance metrics in the Results block. The performance of each regression method is assessed by the root mean squared error of double cross-validation ($RMSE^{dcv}$). Pairwise comparisons are conducted with resort to formal statistical hypothesis testing, in order to incorporate the variability of results in the analysis.

The last block of the PAC framework is the Results Report, where the final Key Performance Indicators (KPI) for the methods under analysis are provided, as well as additional information for interpreting the model, according to the nature of each class of methods (e.g., suitable measures of predictors' importance). Furthermore, one can also make inferences regarding the structure of the dataset, namely its sparsity and collinearity levels based on the profile of important predictor variables and the relative performance of the different methods.

The remaining of this paper is organized as follows. In Section 2, the Predictive Analytics Comparison framework (PAC) is described and details are given about its four components. In Section 3 we illustrate the application of PAC with different datasets, including simulated and real world scenarios. These results are further discussed in Section 4 and Section 5 closes the article with a summary of the main features of the proposed framework and suggestions for future work.

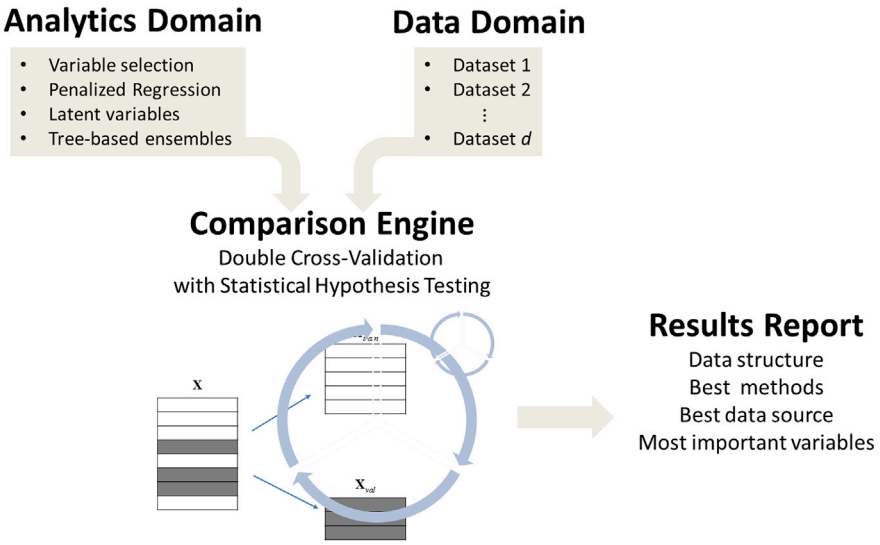

Fig. 1. The PAC framework and its modules.