PREDICTIVE MODELING WITH HIGH-DIMENSIONAL INDUSTRIAL DATA

REIS, Marco S.[1]

[1] *CIEPQPF – Department of Chemical Engineering, University of Coimbra, Polo II – Rua Sílvio Lima, Coimbra, Portugal, marco@eq.uc.pt*

Abstract. The development of data-driven predictive models has been gaining importance in Industry. These models provide the way to obtain reliable estimates of output variables at lower sampling rates (related to product quality or other properties) based on a set predictor variables which are usually easier to measure and less expensive. Applications range from soft sensors, to process diagnosis, quality prediction, process control and optimization.

In this presentation, the problem of developing predictive models from high-dimensional industrial and laboratory data is addressed. The topics of sparsity and collinearity are discussed and solutions to handle them described. A comparison framework is also proposed in order to assess the performance of a rich variety of predictive analytic tools, compare them and provide guidelines for choosing a suitable methodology in a given application scenario.

Results obtained suggest that matching the structure of the system with that of the predictive model often lead to improved accuracy and interpretation. Therefore, all sources of information – data-driven and process background – should be called upon during model development.

Keywords. Sparsity, collineairity, variable selection, latent variable regression, penalized regression, tree-based methods

1 Introduction

Sparsity and collinearity are two pervasive characteristics found in industrial and laboratory data that pose relevant challenges in the development of predictive approaches for relating process variables (X) with continuous (quality-related) variables (Y). Applications where these problems arise include soft-sensor development, process monitoring, control and optimization. Given the importance of developing these data-driven models, a rich variety of regression methods has been proposed in the literature, each one of them with different prior assumptions regarding the nature of X, Y and their underlying relationship. Their performances depend, to a large extent, on the particular structure of data and whether or not it matches the assumptions made by the methods. For instance, data generated by a spectroscopic analysis is predominantly collinear with a latent variable structure and a regression method such as partial least squares (PLS) is in principle more suitable. On the other hand, Design of Experiments data tends to be uncorrelated and sparse, and variable selection techniques are expected to perform better.

In practice, a suitable regression method may be selected using knowledge regarding the data structure together with some inspection of the data generating mechanism. However, very often this information is very limited and the choice is not straightforward. In this scenario, practitioners are often guided by personal preferences regarding their favorite predictive methods. In this work, we apply a comparison framework in order to assess and compare the performance of different regression methods and obtain guidelines for methods that should be selected in a given practical situation.

2 Methods

The predictive methods contemplated in this study cover a wide range of prior assumptions regarding the data generating mechanism and were grouped into four different classes: variable selection methods (Andersen & Bro, 2010), penalized regression methods (Hesterberg, Choi, Meier, & Fraley, 2008), latent variables methods (Jackson, 2005) and ensemble methods (Dietterich, 2000).

In variable selection methods, a subset of predictor variables is selected, based on a specific criterion. In this class of methods, three different approaches were tested: forward stepwise regression (FSR), best subset (BS) and a genetic algorithm approach (GA).

In the class of penalized regression methods, the model's coefficients are obtained by minimizing the squared residuals with a penalty in their magnitude because when the dataset has a high degree of collinearity, small variations in the training data can have a significant impact on the value of the coefficients. Four approaches were contemplated: ridge regression (RR), LASSO, elastic nets (EN) and support vector regression (SVR).

The methods referred above assume that only some individual variables may influence the observed response (sparsity). On the other hand, latent variable methods assume that the observed variability in both X and Y arises from a few unobservable underlying variables (also known as latent variables), which can be estimated from linear combinations of the original variables. In this class of methods, principal component regression (PCR), PCR with forward stepwise (PCR_FS) and PLS were tested.

The last class of approaches tested contains the tree-based ensemble methods, which include: bagging of regression trees (RT), random regression forests (RF) and boosting of regression trees (BT).

In order to assess the performance of the all the methods, a Monte Carlo Cross-Validation framework was developed to estimate the distribution of their prediction errors under testing conditions (quantified by the root mean square error of prediction, under independent testing, RMSEP). The procedure consists of a double cross-validation scheme, and it was used to select the appropriate model formalism and to assess the significance of the observed difference, both from a statistical and practical sense. The training results in each Monte Carlo run provide also useful insights regarding the variables' importance for prediction as well as in the identification of subsets of important predictors.

3 Results

The comparison framework was applied to three case studies in order to demonstrate the importance of choosing the adequate predictive approach for a given regression problem. The first case study is based on simulated data from a sparse correlation model (data set 1): 100 observations from 20 predictor variables were simulated. The second case study is also based on simulated data but from a latent variable model structure with 4 latent variables which govern the observed variability in both X and Y. Since this type of problems is usually characterized by a high number of variables, 100 prediction variables and 500 observations were generated for the latent variable model (data set 2). The last case study concerns a real world application, where data was collected to develop predictive approaches for wine age prediction (Pereira, Reis, Saraiva, & Marques, 2011). 52 samples of Madeira wine (from the same grape variety, Malvasia) were analysed by 3 different analytical techniques, each one providing a set of predictor variables: high-performance liquid chromatography (HPLC), gas chromatography-mass spectroscopy (GC-MS) and UV-vis spectroscopy (data set 3). These variables were used to develop predictive models for the wine age.

Table 2 summarizes the results obtained. For the first case study, FSR presented the best results and was able to select the most important predictors, followed by EN and BS. Thus, the class of variable selection methods is adequate for this data structure since they have the appropriate prior assumptions (only GA presented poor performance, perhaps due to the limited sample size and

some tendency to overfitting). As for case study 2 (latent variable), the results obtained indicate that latent variable methods have a good performance, with PCR and PLS having the second and third best overall performance, respectively. EN presented the best performance but the difference was not statistically significant when compared to PCR and PLS. Again, matching the models prior assumptions and the data generating mechanism provides a more parsimonious description of the data and improved predictions. Finally, for the third case study, BT was the best method for the HPLC measurements (which was the analytical source leading to the best predictive performance). This result point out that, sparsity and collinearity are not the only features present in this data set, but also nonlinearity, a characteristic that BT can handle better than the other predictive approaches tested, which are based on linear modelling frameworks.

Table 1. Best methods for each data set.

Data Set	Best Methods	R^2 for test set
1	FSR	0.92
2	EN ; PCR ; PLS	0.95 ; 0.95 ; 0.95
3*	BT ; RR ; EN	0.99 ; 0.96 ; 0.96

* Results only for the analytical source leading to the best predictive performance (HPLS).

4 Conclusions

In this presentation, we have applied and compared the performance of well-known representatives from four groups of methodologies that are often seen as good solutions to handle high-dimensional data sets. We have considered three distinct application scenarios, with different characteristics of sparsity and collinearity. The methods' performance was assessed by their prediction errors in independently generated test sets, in the scope of a double cross-validation scheme. The results obtained demonstrate the importance of matching the model prior assumptions to the data generating mechanism. In the first case study, a correlation model was simulated and good results were obtained with variable selection methods and penalized regression methods. The second case study was based on data from a latent variable model and latent variables methods had an advantage over the others. Finally, the third case study concerned a practical application of wine age prediction using 3 different analytical procedures. The results showed that the best method depends on the analytical procedure used since each one extracts different types of information, but the best combination of measurement source/predictive method was obtained for HPLC/BT. This brings forward important information about the types of components that are relevant for monitoring wine ageing and the nature of their relationship with the wine aging time (non-linear, in this case). This study highlights the importance of using all sources of information available about the problem under analysis and to avoid relying in favorite methods for prediction, because their success may be highly problem-dependent.

References

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics, 24*(11-12), 728-737.

Dietterich, T. G. (2000). Ensemble methods in machine learning *Multiple classifier systems* (pp. 1-15): Springer.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys, 2*, 61-93.

Jackson, J. E. (2005). *A user's guide to principal components* (Vol. 587): John Wiley & Sons.

Pereira, A. C., Reis, M. S., Saraiva, P. M., & Marques, J. C. (2011). Development of a fast and reliable method for long- and short-term wine age prediction. *Talanta, 86*, 293-304.