Chemometrics and Intelligent Laboratory Systems 127 (2013) 7–16

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables

Marco S. Reis *

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

ARTICLE INFO

Article history:
Received 29 January 2013
Received in revised form 26 April 2013
Accepted 12 May 2013
Available online 23 May 2013

Keywords:
Network-Induced Supervised Learning
Partial correlation
Clustering
Generalized topological overlap measure
Interpretation
Partial least squares

ABSTRACT

Current classification and regression methodologies are strongly focused on maximizing prediction accuracy. Interpretation is usually relegated to a second stage, after model estimation, where its parameters and related quantities are scrutinized for relevant information regarding the process and phenomena under analysis. Network-Induced Supervised Learning (NI-SL) is a recently proposed framework that balances the goals of prediction accuracy and interpretation [1], by adopting a modelling formalism that matches more closely the dependency structure of variables in current complex systems. This framework computes interpretable features that are incorporated in the final model, which effectively constrain the predictive space to be used. However, this restriction does not compromise prediction ability, which quite often is enhanced. Both classification and regression problems can be handled. Four widely different real world datasets were used to illustrate the main features claimed for the NI-SL framework.

1. Introduction

Quantitative models are pervasive in chemical and related sciences, being required in a broad range of applications, either tacitly, as in classification, process monitoring or fault detection, where they are usually encoded in the method's structure and parameters estimated from reference data, or in an explicit way, such as in process optimization, advanced control and product/process design, where a model must be entirely specified before such tasks can be implemented. In this context, a wide spectrum of modelling approaches can be adopted, ranging from those relying on a priori knowledge about the specific processes and phenomena going on, where models are derived from the application of the fundamental laws of nature (conservation of mass, energy and momentum), to pure data-driven approaches able to "infer" or "induce" process knowledge from data available in abundance, as happens when the predictive space of the problem under analysis is densely covered by reliable observations.

However, in practice, the analyst is often confronted with situations where the amount of a priori knowledge available is limited and, given the number of variables involved, the predictive space is also not densely populated with observations (a common consequence of the well-known "curse of dimensionality"). In these scenarios, empirical modelling approaches emerge as adequate solutions, by combining elements of the two extreme paradigms: they use some data to develop models, but the wide "gaps" in the multidimensional space are filled using the model structure postulated, based on previous information about the process and/or resulting from successive model refinements and accumulated experience. The way empirical modelling frameworks are currently developed and implemented falls into two possible categories. On one side, one finds approaches that consider each variable one-at-a-time and the model is built in a stagewise fashion. This process may be entirely sequential or involving iterations, but the distinguishing feature is the consideration of a single variable in each step. Examples include the several methodologies for constructing ordinary least squares (OLS) models (forward addition, backward removal, forward/backward stepwise) [2], classification and regression trees (CART) [3], k-nearest neighbour classification and regression methods (k-NN) [4]. On the other side, we find multivariate methods that consider simultaneously all variables involved. Even though in the end they will weight each variable differently, all variables are considered together in the analysis of the problem. Examples include partial least squares (PLS) [5–11], principal component regression (PCR) [9,10], partial least squares for discriminant analysis (PLS-DA) [12], soft independent modelling by class analogy (SIMCA) [13], linear discriminant analysis classifiers (LDA) [14,15], etc. However, neither of these two paradigms for the development of empirical models match the underlying structure of variables found in complex systems, irrespectively of their artificial (e.g., industrial facilities) or natural origin (cells, tissues, cultures of microorganisms, etc.). It is well known today that complex systems present high levels of *modularity*, *hierarchy* and *specialization* [16–19]. Systems' input variables do not operate altogether at the same time, nor do they act in an entirely isolated fashion. They are

* Tel.: +351 239 798 700; fax: +351 239 798 703.
E-mail address: marco@eq.uc.pt.

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.05.007

organized in modules, with a certain degree of specialization, that operate in the scope of one or several relevant functions in the system. The modules or cluster of variables may sometimes work simultaneously, in a synergistic way, or some of them may be silent under certain circumstances. In this context, what the real nature of systems shows, is that the basic modelling elements should be modules or clusters of variables, instead of isolated variables or the whole set of them. Thus, empirical modelling frameworks should adapt to this reality, in order to enable the development of mathematical descriptions that are closer to the fundamental nature of complex systems. Therefore, one of the features of the methodology described in this article is to construct and handle clusters of functionally related variables, instead of isolated variables or variates containing all variables under analysis (a variate is a linear combination of variables).

But developing an empirical modelling framework that better matches the systems' inner mechanism is not the only issue to be improved in current empirical methods. Another problem arises from the full priority attributed by these methods, to the maximization of prediction ability, leaving model interpretation concerns to a subsequent stage, after the model is established and validated. Examples include OLS (maximization of quality of fit), PCA (maximization of explained variation), PLS (maximization of prediction ability) and LDA or LQA (maximization of true classification rate). The tacit premise is, apparently, "the best we can predict, the best we will be able to explain". However this is often not the case. Excessive focus on prediction usually leads to situations where all degrees of freedom available are used to maximize the amount of explained variability of the response, resulting in complex and ambiguous combinations of variables that raise many difficulties to interpretational queries. Of course, the interpretation difficulties do not constitute a serious problem when the goal of the analysis is strictly centred on the fitting or prediction ability of the model, as happens for instance in calibration and soft-sensor applications. However, the current challenges for analysts and engineers involve, with an increasing frequency, the analysis and operation of progressively more complex systems where the central task is more often focused in interpreting the nature of the relationships between all the variables involved, their interaction and specific roles in the process, than on producing accurate estimates for some system properties or output variables. For instance, the goal can be to gain insight in the way the systems work for the purposes of process improvement or development of new products, in which case the information about the structure of relationships involved can be of great value for defining the next sequence of experiments. Examples of applications where this scenario can be found, include (but are not restricted to): Quality by Design in the pharmaceutical sector (seeking a suitable design space where more efforts can then be devoted in order to find a proper formulation solution), cosmetic and food products, analysis of biosystems (gene regulation, proteomics, metabolomics), analysis of formulated products, or products with complex matrices (wine, paints), and reduction of fluctuations in chemical processing industries. This problem is also found in other scenarios involving the analysis of complex systems, such as the discovery of mechanisms for complex chemical reactions, inference of the molecular origin of a disease, maximization of the throughput from metabolic reactions, etc. In all these cases, the focus in prediction ability is overtaken by the need to collect information about the structure of the system and to know which modules have an active role on the phenomena under study. In this context, Network-Induced Supervised Learning (NI-SL) addresses from the very onset of the analysis, the issue of improving the interpretational value of the results [1]. This is done by building, in the first stage, modules of variables that potentially share the same function. These modules or clusters are then used to construct the final model, whilst keeping their integrity. Therefore, in the end, it is possible to analyse which modules are playing an active role in driving the variability of the response. It is also quite easy to extract the way variables interact in the scope of each module that was found relevant, by analysing their associated weights in the selected variates.

With these two goals in mind (matching complex systems underlying mechanism and balancing interpretation and prediction accuracy), this article presents the NI-SL approach, as follows. In the next section, we describe the basic stages of the framework, and refer how the method is implemented in each one of its stages. Then, in the third section, the results achieved from the implementation of NI-SL in four real world case studies from widely different scenarios, are presented. Two of them involve classification problems (addressed with Network-Induced Classification, NI-C) and the other two, regression problems (where we apply Network-Induced Regression, NI-R). In the final section, we conclude by summarizing the main contributions of this work, and point out some interesting areas of future research in the continuation of the developments reported here.

2. Methods

In this section, we introduce the empirical modelling framework and describe its modules and methods employed. Being a supervised learning framework, it addresses both classification and regression problems. We assume that a suitable training set is available, [**X**,*y*], where **X** represents the $(n \times m)$ matrix with m variables disposed in columns, side-by-side, containing n observations, and y is the $(n \times 1)$ vector of the response variable, which can be either quantitative or categorical depending on the type of problem addressed. The NI-SL framework consists of two stages (Fig. 1). In the first stage, clusters of variables potentially involved in the same function are formed, by analysing the amount of unique information shared by pairs of variables as a measure of their direct interaction. This will be evaluated through the computation of partial correlations and the whole process is robustified by a neighbourhood analysis of all pairs of variables in question. The methodology followed is called Network-Induced Clustering. In the second stage, the basic modelling elements formed in Stage I – the clusters of functionally related variables – , are processed, selected and combined, in order to derive the required classification or regression model, leading to Network-Induced Classification (NI-C) or

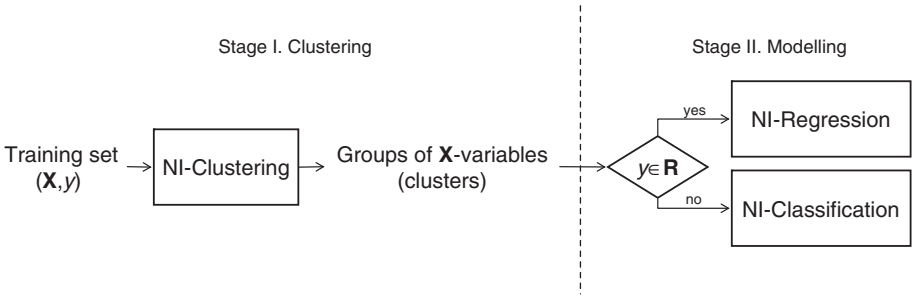

Fig. 1. The modular and stagewise structure of the empirical modelling framework NI-SL.