Chemometrics and Intelligent Laboratory Systems 157 (2016) 67–77

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A hybrid input variable selection method for building soft sensor from correlated process variables

Md Musfiqur Rahman, Syed Ahmad Imtiaz*, Kelly Hawboldt

Faculty of Engineering and Applied Science, Memorial University of Newfoundland, 240 Prince Philip Dr, St. John's, NL A1B 3X5, Canada

ARTICLE INFO

Article history:
Received 24 June 2015
Received in revised form 20 June 2016
Accepted 21 June 2016
Available online 2 July 2016

Keywords:
Variable selection
Soft sensor
Retrospective Taguchi method
Backward elimination
Support vector regression (SVR)

ABSTRACT

Selection of the most relevant input variables for an inferential predictor is important for good prediction ability. A hybrid variable selection method is proposed for selecting input variables for support vector regression (SVR) model. The proposed method combines Taguchi's experimental design method with backward elimination method to select the most relevant variables from a large set of process variables. Taguchi's design of experiment (DoE) method was used to screen variables, as process variables are highly correlated this poses difficulty to fill in the design matrix of Taguchi's DoE method. The proposed method makes several modifications to Taguchi's method to deal with this problem. Subsequently backward elimination method was used to select the final set of input variables. The efficacy of the proposed methodology is demonstrated on an industrial case study.

1. Introduction

Soft sensors are widely used in process industries for monitoring and control purposes. Selection of the appropriate input variables for soft sensor has many potential benefits including data visualization, reduced training time, improved utilization time, and better prediction performance. Many diverse techniques are used to select inputs for soft sensors which fall under three groups: wrapper methods, filter methods, and embedded methods [10]. Variable selection step is an integral part of embedded methods, on the other hand wrapper and filter methods are optional steps in the overall methodology. Wrapper methods use the target model in selecting the input variables. Several subsets of input variables are used to build the model, and the set of variables with the best prediction capability is typically selected. For complex models computational load can be high and the search can become intractable. Often optimization algorithms are used to expedite the search. Filter methods, on the other hand, either use an optional output from the model or utilize an indirect estimator to measure the prediction ability of the selected subsets and thereby, the target model has no influence on the variable selection. When an indirect estimator is used, these methods can be used as a separate preprocessing step to any model structure, and give additional confirmation on the selection of variables. Filter methods can also provide insight into the range and quality of the data set which are important information for building a model as well as in application of a soft sensor on-line. Recently a filter method called retrospective Taguchi method was used for selecting input variables for a SVR based soft sensor [20]. Retrospective Taguchi method is based on Taguchi's experimental design approach that utilizes stored data from data historian to design the experiments. It is a powerful statistical design approach applied for improving quality in products and processes by reducing variability in the process [2, 12, 22]. Taguchi method uses design of experiment (DoE) and assumes that factors are independent of each other which is restrictive for dealing with process data. Due to material recycle and heat integration measured process variables are highly correlated. Also, data are corrupted by measurement noise, and often variables are operated in a narrow range of operation which make it difficult to populate Taguchi's design matrix using historical process data. In this paper we specifically address these application issues of Taguchi method. We combined Taguchi's method with a wrapper method, i.e. backward elimination, and developed a hybrid method to better handle the correlated data set. The core of the method is grouping correlated variables based on their correlation matrix, applying retrospective Taguchi method to find important groups, and finally, eliminating least contributing variables from the selected groups using backward elimination approach. We applied the proposed methodology for selecting input variables for a SVR based soft sensor. The performance of the proposed method is compared

* Corresponding author.
E-mail address: simtiaz@mun.ca (S. Imtiaz).

http://dx.doi.org/10.1016/j.chemolab.2016.06.015
0169-7439/

with variable importance in projection (VIP) method which is one of the suggested methods when correlation or collinearity is present in input data [5].

The rest of the paper is divided into the following sections. In Section 2 we provide a brief review of input variable section methods for soft sensors. The preliminaries of Taguchi's variable selection method are covered in Section 3. Section 4 describes the detailed methodology to build inferential predictor for Terephthalic Acid process with specific focus on the methodology of variable selection from a set of correlated variables. Section 5 describes the industrial case study and results from application of the proposed methodology to the industrial data set. Finally, in Section 6 key conclusions and contributions are summarized.

2. Literature review

Selection of input variables is an important problem in many areas including, regression, soft sensor development, experimental, and simulation studies. Partial least squares (PLS) has been the preferred method for regression in many areas including process data analysis and chemometrics, as such many variable selection methods have emerged for PLS model. We focus our review on variable selection methods for PLS and SVR, also discuss the relationship between variable selection methods for soft sensors and other closely related areas.

2.1. Input variable selection methods for PLS model

In the context of process data analysis VIP-PLS is widely used for selecting important input variables for soft sensor. VIP is one of the early methods for selecting a combination of variables that give good prediction performance and is embedded in Matlab PLS Toolbox. VIP values provide a combined measure of contribution of independent variables in X block in describing the dependent variable in Y block. The total contribution of each variable is calculated by adding the fractional contribution of each variable to the variances explained by each PLS components. The fractional contribution of each variable is calculated based upon the weights of regression model. A VIP value smaller than 1 indicates a non-important variable which can be ignored [1]. VIP also delivers good performance when multi-collinearity is present in the data set [5].

A variety of search strategies, for example, backward variable elimination (BVE), regularized elimination procedure, genetic algorithm (GA) has been used for selecting input variables for PLS model [8, 18]. GA combined with PLS regression (GA-PLS) is a preferred method in process industries for developing soft sensor [14]. Through this selection process after several generations a set of high performing variables evolves. GA-PLS is embedded in AspenIQ, which is Aspen Technology's inferential predictor building module. However, in a multivariate framework the input variables can cancel each other's effect and GA-PLS does not always select variables that are the most important from a process knowledge point of view.

PLS regression is also widely used in chemometrics for data analysis. PLS combined with competitive adaptive re-weighted sampling (CARS) was used for near infrared (NIR) data analysis. Normalized absolute weights of PLS coefficients are used for weighting the variables, subsequently several sets of variables are created based on re-weighted sampling. Finally, variables with the least RMSECV are selected [15]. A modified version of CARS called stability competitive adaptive re-weighted sampling (SCARS) uses a stability index, defined as the absolute value of regression coefficient divided by its standard deviation for re-weighting the variables. SCARS usually gives fewer number of informative variables compared to CARS [29]. Latent projective graph (LPG) is another simple yet effective method for near infrared spectral analysis. The method is based on the assumption that collinear wavelengths in the calibration spectra may have the same contribution to the modeling. As such these variables are redundant in building models. The variables located at the inflections of an LPG are found to be informative for the quantitative models. The method lead to parsimonious models for NIR data when used with PLS and other kind of models [23]. Centner et al. [4] proposed uninformative variable elimination in PLS (UVE-PLS) where artificial noise variables are introduced, all variables having lower contribution than noise variables are eliminated. The procedure is repeated several times until the selection criterion is met [4]. A similar method, RT-PLS was developed based on Fisher's randomization test (RT) to select informative wavelength for NIR spectroscopy. A regular PLS model and a number of random PLS models are constructed by randomizing the dependent variable. The regression coefficients of the regular model are compared against the respective regression coefficients of all of the random models. Coefficients which are bigger than the random models most of the time are retained, while the other are eliminated [28]. A comprehensive review on variable selection methods for PLS can be found in [17] and the references within.

2.2. Input variable selection methods for SVR model

Compared to PLS limited research has been done on selection of variables for SVR models. Bi et al. [3] showed that in case of linear SVR where l_1-norm is minimized, the minimization drives sufficient number of weights to zero, and variable selection is a non-iterative process. Thus, l_1-norm linear SVR can be used to select variables similar to any filter method, subsequently a non-linear l_1-norm SVM can be used for prediction purpose [3]. SVR is a computationally intensive method, computational complexity of SVR model increases quadratically with the number of variables. An exhaustive search for input variables can quickly become intractable. Greedy search strategies namely, forward selection and backward elimination has been used for minimizing the computational load. Guyon and Elisseeff [10] proposed SVM-RFE, based on backward elimination. Variables are removed sequentially one at a time and their effects on the weights of the SVR model are observed. Ranking of each variable is done based upon the relative change in weights, variable with maximum relative sensitivity is removed [10]. Several variations of this selection criterion are also reported in literature [21]. These various methods for variable selection were mainly applied for unraveling information from gene expressions. Selection of input variables from large set of process variables for building SVR based soft sensor received less attention. In our previous work we used Taguchi's DoE method for selecting input variables for SVR model [20]. The developed method was applied to build inferential predictor for TA plant. In this current research we further develop this method to deal with highly correlated set of process data.

2.3. Relationship to other areas of research

Selection of input variables is also important research topic in other closely related areas, for example, DoE and simulation studies. In DoE, instead of choosing the experiments or scenarios arbitrarily, experiments are chosen following design matrices suggested by sampling techniques, for example, orthogonal array, or space filling designs, such as Latin hypercube sampling. Significant work has been done in this area by statisticians, and simulation analysts. Comprehensive discussions on DoE methods and relationship between these diverse areas is given in [13, 24]. There are significant differences between selection of variables for soft sensors and selection of variables in simulation studies. DoE based methods work best at the early design stage or when experiments are conducted on simulation model where inputs can be freely chosen. In process systems, conducting experiments is usually either