Chemometrics and Intelligent Laboratory Systems 135 (2014) 90–109

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Industrial PLS model variable selection using moving window variable importance in projection

Bo Lu [a], Ivan Castillo [b,*], Leo Chiang [b], Thomas F. Edgar [a]

[a] McKetta Department of Chemical Engineering, The University of Texas at Austin, Austin, TX, United States
[b] Analytical Technology Center, The Dow Chemical Company, Freeport, TX, United States

ARTICLE INFO

Article history:
Received 10 December 2013
Received in revised form 28 February 2014
Accepted 31 March 2014
Available online 8 April 2014

Keywords:
PLS regression
Variable selection
Model reduction
Partial least squares
Multivariate statistics
Process monitoring
Soft sensors
Inferential sensors

ABSTRACT

Soft sensors (or inferential sensors) have been demonstrated to be an effective solution for monitoring quality performance and control applications in the chemical industry. One of the key issues during the development of soft sensor models is the selection of relevant variables from a large array of measurements. A subset of variables that are selected based on first principles and statistical correlations eases the model development process. The resulting model will perform better and will be easier to maintain during the deployment stage. In the current literature, data-driven variable selection methods have been investigated within the context of spectroscopic data and bioinformatics. In these studies, the variable selection methods assume that the inherent correlation in the entire data set remains fixed. This is not the case in common industrial processes. In this paper, existing variable selection methods based on partial least squares (PLS) will first be evaluated. Second, we will present a new approach called moving window variable importance in projection (MW-VIP) to target the selection of correlations present in segments or small clusters. Finally, a set of new evaluation criteria will be presented along with industrial data set modeling results to demonstrate the effectiveness of our proposed approach.

1. Introduction

Continuous improvements in energy efficiency, safety, plant reliability and up-time have become performance drivers for chemical engineering industries, where multivariate process monitoring, model-based control and plant wide optimization strategies play important roles [1]. In most cases, high quality, reliable and timely measurements of key process variables and quality variables are needed. Yet, most quality variables such as concentration or purity require lab analysis that introduces time delay and consequently reduces their effectiveness in real-time monitoring. A possible solution is to use on-line analyzers, but they can be costly to calibrate and maintain over the long-term. Soft sensors are a possible alternative that utilize existing process measurements to predict key quality variables. Two main approaches of soft sensor development are data-driven and fundamental model-driven [2]. Soft sensors are able to provide real-time quality predictions that are much faster in response time, and thus can be extended to provide real-time monitoring, fault diagnosis, and advanced control.

Data driven soft sensors apply multivariate statistics and machine learning techniques to find empirical correlations between process variables and quality variables. Partial least squares (PLS) and principal component analysis (PCA) are two popular techniques for finding these correlations. There are many publications related to the application of these methods for soft sensor and multivariate monitoring [3–6]. More recently, reduced order dynamic PLS based soft sensors have been developed for the monitoring of processes experiencing large transport delays [7]. PLS regression in particular is suited to deal with high-dimensional data in the presence of colinearity. In practice, performance of the regression models could often be improved when a subset of highly relevant variables is used instead of the whole training data set [8]. The reduced models are more resilient to measurement noise and are often more interpretable. Thus, a successful variable selection procedure will improve the interpretation and identification of the underlying process conditions.

The popularity of PLS methods has also generated interest in PLS variable selection techniques. Mehmood et al. [9] showed that the number of publications in the field of PLS modeling and related methods has increased exponentially since 1988. The field of application outlined in his review ranged from gene selection data to gene expression data, quantitative structure–activity relationship (QSAR) descriptor selection, spectroscopy wavelength selection, and bio-marker selection. Kalivas and Sutter provided another review of variable selection in the field of QSAR descriptor selection [10]. Saeys et al. reviewed popular selection techniques in the field of bioinformatics [11]. While being quite comprehensive in the methods reviewed, these reviews did not cover PLS

* Corresponding author.
E-mail address: castillo2@dow.com (I. Castillo).

http://dx.doi.org/10.1016/j.chemolab.2014.03.020
0169-7439/

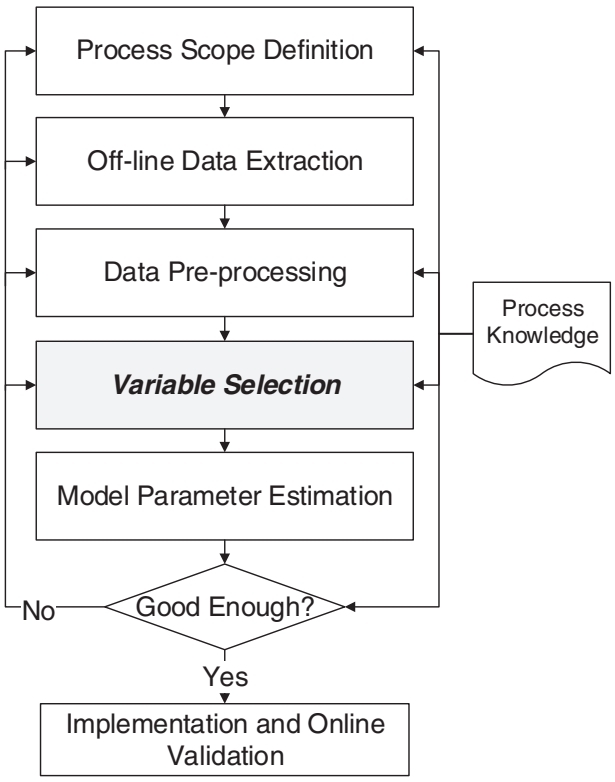

Fig. 1. Overview of the soft sensor modeling process.

variable selection in the context of industrial process data that involves multiple operating modes. Since multivariate statistical models are heavily influenced by the properties of the underlying data, the variable selection methods suited for bio-marker selection are likely to be inappropriate for process variable selection. As a result, the goal of this paper is to evaluate the existing methods for variable selection of industrial process data, and to present an improved variable selection method appropriate for industrial processes with multiple operating modes.

The structure of this paper is organized as follows. First, we evaluate existing variable selection methods; several representative techniques will be implemented and assessed. Second, we will present our improvements to the current methods to address their shortfalls. Last, we will present evaluation results using a set of model-free criteria that help in the assessment of variable selection performance.

2. Background

2.1. Data-driven PLS and variable selection methods

An in-depth introduction of PLS methods is available in [12–14], and thus these methods are not discussed in detail here. The soft sensor model development process typically consists of data gathering, preprocessing, variable selection, model development, and model validation. Implementation details of each step will vary depending on the specific application. Kaldec and Sliskovic provided comprehensive reviews of the soft sensor development process for interested readers [2, 15]. For industrial processes, the model development work flow can be summarized in Fig. 1. Defining the proper model scope, applying the right pre-processing steps and performing model validation are critical steps in addition to variable selection. Expert knowledge and first principles-based understanding of the process are useful aids in prescreening variables, transforming non-linear variables and validating models. Soft sensor modeling practitioners should maximize process knowledge integration to give physical significance to the resulting PLS models.

There are many existing variable selection methods specific to PLS regression. Mehmood et al. showed that the number of publications in PLS related work has been growing exponentially since the 1980s [9]. A general observation made from this body of research is that most variable selection methods calculate a variable importance ranking metric and then apply this metric in the subsequent steps to find the optimal variables. Mehmood et al. and Saeys et al. suggested categorizing these methods based on the mechanism of variable ranking and selection into three types — filter, wrapper, and embedded [9,11]. Fig. 2

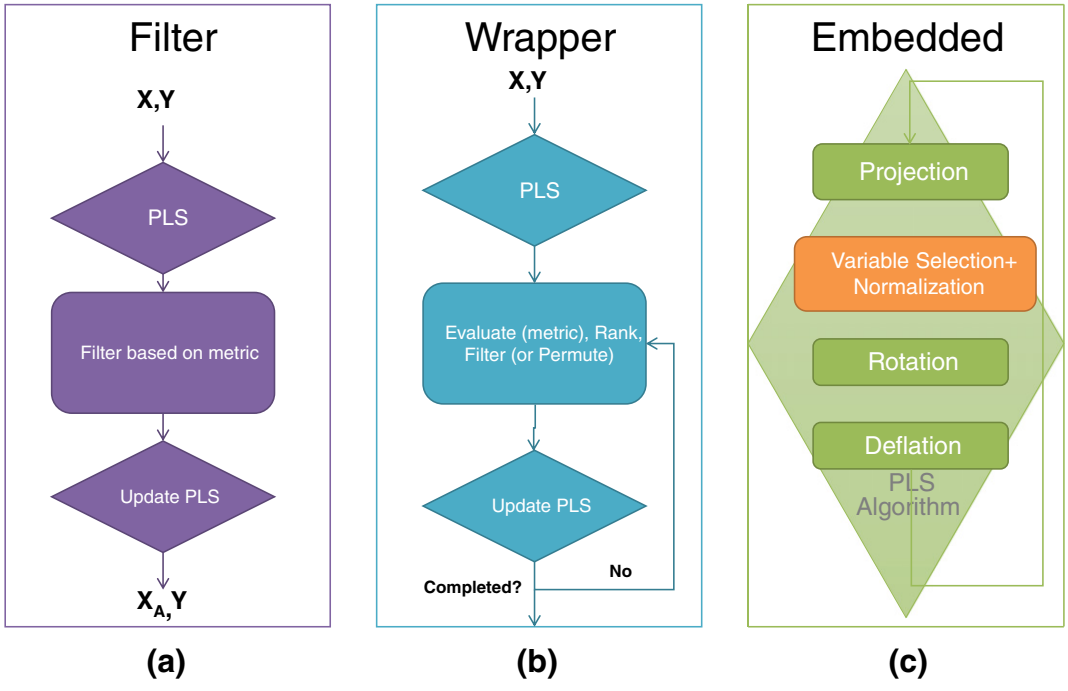

Fig. 2. Three categories of PLS variable selection methods — (a) filter, (b) wrapper and (c) embedded.