Chemometrics and Intelligent Laboratory Systems 167 (2017) 85–95

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A variable selection method for soft sensor development through mixed integer quadratic programming

Weiyu Jian [a], Lingyu Zhu [b], Zuhua Xu [a], Xi Chen [a,*]

[a] *State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China*
[b] *College of Chemical Engineering, Zhejiang University of Technology, Hangzhou, 310014, Zhejiang, China*

ARTICLE INFO

Keywords:
Soft sensor
Variable selection
Mixed integer quadratic programming
Bayesian information criterion

ABSTRACT

Soft sensors are widely employed in industry to predict quality variables, which are difficult to measure online, by using secondary variables. To build an accurate soft sensor, a proper variable selection is critical. In this project, a method of selecting the optimal secondary variables for a soft sensor model is proposed. It is formulated as a nested optimization problem. In each iteration, a mixed integer quadratic programming (MIQP) is conducted with the Bayesian information criterion (BIC) to estimate the prediction error. A warm start (WS) technique is developed to speed up the convergence. The proposed method is evaluated using a number of instances from the UCI Machine Learning Repository. The computational results demonstrate that this method is well suited for finding the best variable subsets. The method is successfully applied to build soft sensors for an industrial distillation column. The results show that the proposed method can effectively select feature variables that will improve the model prediction performance and reduce the model complexity. Comparisons with other methods, including the traditional partial least square technique, are also presented.

1. Introduction

With the development of production technology, modern chemical processes have become increasingly stricter in terms of product quality. To properly control the product quality, it is necessary to measure the product quality online. Unfortunately, there are a large number of cases where quality variables cannot be detected online by conventional hardware sensors for economic or technical reasons. The soft sensor technique, therefore, becomes an attractive approach to address this problem. The soft sensor technique utilizes easily measurable process variables, the so-called secondary variables, to estimate primary variables on-line by constructing a mathematical relationship between secondary variables and quality variables [1–3].

It is well known that a proper modeling method is essential for developing a soft sensor with good performance. Techniques such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), artificial neural networks (ANN) and support vector regression (SVR) have been well reported [4–11] for the development of soft sensor models. However, the effectiveness and maturity of a soft sensor model are likely to be achieved only if those secondary variables that are most closely related to the primary variables are employed. With increasingly more industrial data becoming available in chemical processes, it is challenging to determine what data should be used in the soft sensor model. Furthermore, the inappropriate selection of secondary variables in a model may lead to many problems such as difficulties in model explanation and parameter estimation. Moreover, redundant variables in a model may lead to overfitting and therefore poor prediction for new data not used in the model training. Therefore, variable selection, though not as well studied as modeling techniques, is also critical for developing soft sensors. With a proper variable selection, we can not only improve model predictive performance but also simplify the model complexity and ease the model interpretation.

The variable selection in a soft sensor can be regarded as the best variable subset selection, which has been proved to be an NP-hard problem. Guyon et al. [12] categorized variable subset selection methods into three methods: filter, wrapper and embedded methods. Filter methods select a variable subset based on a ranking criterion. A commonly used filter method is the correlation criterion, which computes the importance of every variable independently by comparing the correlation between every variable and the dependent variable [13]. However, the filter methods always lead to the selection of a redundant subset. The same model performance could be achieved with a smaller

[*] Corresponding author.
E-mail address: xi_chen@zju.edu.cn (X. Chen).

http://dx.doi.org/10.1016/j.chemolab.2017.05.011
Received 17 January 2017; Received in revised form 8 April 2017; Accepted 8 May 2017
Available online 25 May 2017

subset of variables selected by wrapper methods or embedded methods. Wrapper methods evaluate the subsets based on the model prediction performance on a validation set and search the space of possible variable subsets according to the predefined search algorithm until satisfactory prediction performance is achieved [14]. Commonly used search algorithms include genetic algorithms and simulated annealing algorithms. These algorithms are often criticized for their massive amounts of computation. Thus, heuristic search strategies have been devised. Among them, the stepwise regression method is a well-known method due to its computational advantages. The stepwise regression method repeats forward selection and backward elimination until a stopping criterion is satisfied. However, this method sacrifices prediction performance; it cannot find an optimal variable subset. In contrast to wrapper methods, embedded methods can simultaneously select variable subsets during model construction [15]. Embedded methods optimize a two-part objective function with a goodness-of-fit (GOF) term and a penalty term for a large number of variables directly. Typical methods include the least absolute shrinkage and selection operator (LASSO) [16], the smoothly clipped absolute deviation (SCAD) penalty [17], the minimax concave penalty (MCP) [18], and mixed integer programming (MIP). The LASSO technique uses a squared objective function that is penalized by a function of the magnitude of the regression coefficients to perform variable selection. However, it results in far less accurate solutions, likely due to the highly-correlated variables. SCAD can produce sparse set of solution; but the coefficients are biased for large coefficients. Similarly, MCP is also biased though it is fast and continuous. The MIP method has the potential to select the best variables and is unbiased. The MIP has been widely used in optimal design and operation in process systems engineering [19–21]. For variable selection, however, only a few publications have been reported that formulate the subset selection problem as a mixed integer quadratic programming (MIQP) problem by minimizing the sum of squared model deviations. In particular, Bertsimas et al. [22] developed a tailored branch-and-bound procedure to solve the MIQP problem. Konno et al. [23] employed the mean absolute model deviation as an objective function and formulated the variable selection problem as a mixed integer linear programming (MILP) problem, which is easier to solve than the MIQP problem. However, both the MIQP and MILP methods need to predetermine the number of selected variables before solving them. Otherwise, most or all the candidate variables will eventually be included in the model because the more complicated the model is, the more accurately it fits during training. Therefore, it is not appropriate to use the sum of squared model deviation or mean absolute model deviation as the GOF measure; it is necessary to find a more effective GOF measure to estimate the true prediction error of variable subsets.

Some classical GOF measures, such as the Akaike information criterion (AIC) [24], Bayesian information criterion (BIC) [25] and Mallows' C_p [26], have been proposed for estimating the true prediction error. Emet et al. [27] studied model structure selection for minimizing the AIC. This problem is computationally intractable due to its nonlinearity and nonconvexity. They noted that, if the residual variance is known or predefined, the model structure selection problem for minimizing the AIC can be reduced to an MIQP problem with a convex and nonlinear objective function. However, this conflicts with the formal definition of the AIC, which requires the residual variance to be the maximum likelihood estimation. Miyashiro et al. [28] proposed an MIQP formulation for subset selection using Mallows' C_p as the objective function. This method enables one to find the best subset of variables in terms of Mallows' C_p. However, it may also result in redundant variables because the penalty term in Mallows' C_p is not sufficiently large.

In this paper, an efficient and computationally tractable method is proposed for selecting the best secondary variables in soft sensor development. To improve the model prediction ability while maintaining adequate model simplicity, the variable selection task is conducted through a set of optimization problems. In each step, the preset subset dimension is incremented by one, and an MIQP problem is solved to select the best variable subset. Then, the BIC is applied to estimate the model true prediction error and determine the termination of the selected variable subset. Examples from UCI datasets are used to evaluate the proposed method. An application to develop the soft sensor for an industrial distillation column composition is also presented.

2. Multiple linear regression model and BIC

Multivariate statistical methods [4] are widely used to derive regression models such as the MLR model, PCR model and PLS model. Among them, the MLR model has attracted numerous studies in variable selection, especially normal linear regression models, due to their analytical expression convenience.

MLR model attempts to model the relationship between independent variables and a dependent variable by fitting a linear equation to given data. Given n samples $(x_{i1}, x_{i2}, \cdots x_{ip}; y_i)$, for $i = 1, 2, \cdots n$, x_{ij} $(j = 1, 2, \cdots, p)$ are p candidate independent variables, and y_i is a dependent variable. A MLR model is constructed for predicting the output y as follows:

$$\mathrm{y} = b + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + \varepsilon \quad \varepsilon \sim N_n\left(0, \sigma^2 I\right) \tag{1}$$

where ε is a prediction residual, which is random, independent, and identically distributed with zero mean and unknown variance σ^2, and b and a_j $(j = 1, 2, \cdots, p)$ are p+ 1 unknown parameters to be estimated.

For convenience of explanation, we rewrite the model (1) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon} \tag{2}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{a} = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

The variable selection problem is to determine a subset of variables $\{x_1, x_2, \cdots x_k\}$, $k < p$ from all the candidate independent variables. There are a total of $2^p - 1$ possible different combinations. Therefore, it is necessary to evaluate each subset regression model using some GOF measures and select the variables closest to the true model.

For a prediction model, the major task is to predict unknown data. The quality of the established model should be evaluated according to its generalization performance. Therefore, the prediction error for test data should be of concern instead of the error for the training data when evaluating a prediction model.

It is well known that a model is always highly fitted to the training data compared to the validation data. Therefore, the expected error for the validation data will be higher than the error for the training data. As the model complexity increases, the training error always decreases. However, simultaneously, overfitting can easily occur if we simply focus on the training error. Although it is impossible to measure the true prediction error precisely, there are several methods to estimate the expected validation error with good accuracy. An obvious way to estimate the prediction error is to calculate the model complexity and then add it to the model training error. For a linear model, the model complexity represents the number of variables in the model. It is well known that a generalized information criterion (GIC), such as the AIC, the BIC and Mallows' C_p, works in this way to estimate the true prediction error as follows:

$$GIC = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2 + \gamma p \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector, p is the number of