Chemometrics and Intelligent Laboratory Systems 125 (2013) 153–165

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Multiway elastic net (MEN) for final product quality prediction and quality-related analysis of batch processes

Chih-Chiun Chiu, Yuan Yao *

Department of Chemical Engineering, National Tsing Hua University, Hsinchu, 31003, Taiwan, ROC

ARTICLE INFO

Article history:
Received 10 March 2013
Received in revised form 7 April 2013
Accepted 11 April 2013
Available online 19 April 2013

Keywords:
Batch process
Quality prediction
Process analysis
Variable selection
Elastic net

ABSTRACT

In batch processes, the final product quality is determined by the trajectories of the process variables throughout each batch. Consequently, there are two important issues that should be considered in the quality-related modeling. First, the process variable trajectories usually contribute to the final product quality cumulatively along the operation time within each batch. Such effect is named as the cumulative effect. Second, each process variable may have different impacts on the product quality at different time intervals, which is denoted as the time-varying effect. In order to model both two effects reasonably, a multiway elastic net (MEN) method is proposed in this paper. Accordingly, a quality prediction and process analysis scheme is presented. MEN integrates variable selection and regression in batch process modeling, where the regression coefficients are regularized in an automatic manner. With proper data pre-treatment, MEN can provide both accurate prediction and good interpretation. For online prediction, a future data estimation approach is proposed based on the *k*-nearest neighbor technique. The application of the proposed scheme to an injection molding process shows that MEN is not only effective in the online quality prediction but also enhance the understanding of the process.

1. Introduction

In order to meet the demands of the ever-changing market, batch processes have been widely applied in modern industry, for their flexibility in the production of high-value-added products. The requirement of consistent and high final product quality inspires the quality-related research of batch processes. Due to the lack of online measurement equipment, final product quality, in most cases, can only be measured at the end of each batch run. Therefore, online quality prediction based on multivariate statistical methods has attracted great research attention, and various types of regression models have been proposed for modeling batch processes, e.g. [1–4].

Different criteria can be adopted for evaluating the quality of a regression model. Typically, the following two aspects are of great importance [5]: (a) the prediction accuracy on future data, and (b) the interpretation of the model. For a process model, prediction accuracy is important but not the only thing one should look at. The model parameters should be interpretable and consistent with process knowledge. In batch processes, the final product quality is usually determined by the dynamic trajectories of the process variables in a cumulative way throughout each batch. Such effect can be denoted as the cumulative effect. In the meantime, batch processes have time-varying characteristics in nature, implying that each process variable may make different contributions to the final product quality at different time intervals in a batch run. In this paper, such an effect is named as the time-varying effect. To achieve good quality prediction and process understanding, both two effects should be taken into consideration in the process modeling procedure.

Multiway partial least squares (MPLS) [1] is the most popular multivariate statistical method for batch process quality prediction, which treats the entire batch data as a single object in process modeling. Hence, the cumulative effect is automatically described. However, during the modeling steps, MPLS does not consider the time-varying effect. Especially, because of the existence of the measurement noise, the MPLS regression coefficients are seldom equal to zero, even if the corresponding variables or time intervals contain no information for quality prediction. A natural question is whether it is possible to weight the predictors according to their importance before or during regression modeling. Various types of multiblock PLS methods [6–8] may be adopted to solve this problem by assigning different weights to different blocks before building the regression model. Unfortunately, the block division requires prior process knowledge that may be insufficient in practice. Moreover, the multiblock PLS methods do not provide a way for the calculation of the weights. Chu et al. [3] integrated the bootstrapping technique with a stepwise variable selection method to eliminate unimportant predictors from the original data set. However, their method leads to discontinued variable selection along the operation time. Such selection results do not have a clear physical meaning and are difficult to interpret. In addition, the online prediction results were not shown in their paper.

* Corresponding author. Tel.: +886 3 5713690; fax: +886 3 5715408.
E-mail address: yyao@mx.nthu.edu.tw (Y. Yao).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.04.006

More recently, Lu and Gao [2] proposed a stage-based PLS method. Although this method has several attractive features, there are still shortcomings. Stage-based PLS divides the operation stages into two types: critical-to-quality stages and non-critical-to-quality stages. Only the data in the former stages are utilized in online quality prediction. Therefore, in the viewpoint of variable weighting, stage-based PLS assigns unit weight to all predictors in critical-to-quality stages and zero weight to all predictors in non-critical-to-quality stages. The relative importance of different variables and different time intervals within each stage is not revealed. Moreover, for online quality prediction, stage-based PLS mainly utilizes process information at each single time interval, while the information about the cumulative effect is less well reflected. In 2007, a variable-weighted PLS (VW-PLS) method [9] was developed to choose variable weights using optimization technique, but such method is not suited for batch process modeling. In a batch process model, there may be thousands of predictors, causing heavy computation burden in solving the optimization problem. Furthermore, VW-PLS has no means to avoid over-optimization that leads to unreasonable weighting results and poor generalization capability of the model.

It is well known that variable selection can enhance the predictability of a regression model and make the model more parsimonious. Most typically, stepwise selection procedures have been widely utilized. However, as discussed in [10], these methods have several drawbacks: (a) the theoretical properties of these methods are difficult to understand; (b) the estimates are extremely variable; and (c) the computational costs are high when the number of predictors is large. In the last two decades, regression regularization methods have been developed to carry out parameter estimation and variable selection simultaneously. A representative method is the least absolute shrinkage and selection operator (lasso) [11] that imposes a bound on the L_1 norm of the coefficients. This results in shrinkage of regression coefficients. Several coefficients may become identically zero. In a sense, lasso assigns a weight to each predictor variable reflecting its contribution to the outcome. Such feature is desired in the quality-related modeling of batch processes, to reflect the time-varying effect on the final product quality. However, lasso suffers from two major limitations [5]. First, if the number of predictor variables (M) is larger than the number of observations (N), lasso selects at most N variables. Second, if there are groups of highly correlated predictors, lasso tends to arbitrarily select only one from each group and creates models difficult to interpret. Both the situations are common in batch process data. Therefore, lasso is not a suitable method for batch process modeling. Instead, another regularization technique named as elastic net [5] does not have such limitations. Unlike lasso, elastic net encourages a grouping effect, where strongly correlated predictor variables tend to be in or out of the model together. Meanwhile, elastic net is particularly useful in the case of $M > N$.

Enlightened by the regularization techniques, a multiway elastic net (MEN) method is proposed in this paper for online quality prediction and quality-related analysis of batch processes. On one hand, MEN inherits the properties from elastic net, which enables this method to regularize the regression coefficients in a batch process model. As a result, although MEN does not consider the time-varying effect explicitly by assigning different weights to the predictors before model building, the related information is automatically built into the model through coefficient regularization during modeling. On the other hand, based on proper unfolding and normalization of the batch process data, MEN also models the cumulative effect well. For online prediction, a future data estimation approach is proposed based on the k-nearest neighbor (kNN) technique [12]. Comparing to the conventional estimation methods, the kNN estimation performs much better. By adopting MEN, not only the quality prediction accuracy but also the interpretation of regression coefficients is improved. Process knowledge can be explored through model analysis.

The paper is organized as follows. In Section 2, the MPLS and stage-based PLS methods are reviewed and discussed in details to show the motivations of this paper. Then the multiway elastic net method is proposed in Section 3, including data pre-treatment, regression modeling for batch processes, future data estimation, and quality-related process analysis. In Section 4, the proposed method is applied to an injection molding process for illustrating its good feasibility in both product quality prediction and process knowledge mining. Finally, conclusions are drawn in Section 5.

2. Existing methods and motivations

2.1. MPLS and stage-based PLS

For predicting the quality of batch process final products, MPLS [1] is possibly the best-known statistical methods, which extends the application of partial least squares (PLS) [13] to batch process data. As most regression methods, PLS deals with two-dimensional data matrices. However, the historical process data collected from a batch process are usually represented by a three-dimensional data matrix $\underline{\mathbf{X}}(I \times J \times K)$, where I is the number of total batches in the historical dataset, J is the number of process variables, and K is the number of total sampling time intervals in a batch. Therefore, MPLS unfolds $\mathbf{X}$ to a two-dimensional data matrix $\mathbf{X}_b(I \times JK)$, by keeping the dimension of batches and merging variable and time dimensions. Each row of the unfolded matrix $\mathbf{X}_b$ contains all data within a batch, while each column corresponds to the measurement value of a particular variable at a certain time interval. Then, normalization can be performed on this two-way matrix to remove the influences of the units and the measurement ranges of different variables, where each column of $\mathbf{X}_b$ is mean-centered and scaled to unit variance. The formula is as below:

$$\tilde{x}_{n,m} = \frac{x_{n,m} - \bar{x}_m}{s_m} \quad (n = 1, \cdots, I; m = 1, \cdots, JK), \tag{1}$$

where n is the row index of $\mathbf{X}_b$, m is the column index, $x_{n,m}$ is the (n, m)-th entry of $\mathbf{X}_b$, $\bar{x}_m$ is the mean value of the m-th column, s_m is the standard deviation of the m-th column, and $\tilde{x}_{n,m}$ is the (n,m)-th entry in the normalized matrix. Such unfolding and normalization are named as batch-wise unfolding and normalization. The normalized matrix is then regressed with the normalized vector $\mathbf{y}$ of the final product quality to establish the prediction model. Since MPLS regards the entire batch data as a single object, estimation of future measurements is needed in online quality prediction.

Another typical statistical method for batch process quality prediction is stage-based PLS [2] that was proposed based on the following findings. (1) Many batch processes have multiple operation stages (also called phases), where different stages may have different influences on the product quality. (2) A particular type of product quality may be determined in some particular stages and by some particular process variables. Stage-based PLS normalizes batch process data in a similar way as in MPLS. Then, the normalized data matrix is split to K number of time-slice matrices $\mathbf{X}^k(I \times J)$, where $k = 1, \cdots, K$, and I, J, K are the indices of batch, variable and time, respectively. By regressing each $\mathbf{X}^k$ to the normalized quality data vector $\mathbf{y}$, the time-slice PLS models are built and the corresponding time-slice regression coefficients are calculated. These coefficients are divided into several groups utilizing the k-means clustering algorithm [14]. Associating the clustering results with operation time information, the stage division results are achieved. The stage models are then created as an average of all time-slice PLS models in each stage. The multiple coefficient of determination, R^2 [15], is used to evaluate the fitness of each stage model and reveal the importance of each time interval in determining the product quality. If the average value of R^2 in a certain stage is larger than a threshold value, this stage is defined as