Chapter 27
Key Process Variable Identification for Quality Classification Based on PLSR Model and Wrapper Feature Selection

Wen-meng Tian, Zhen He, and Wei Yan

Abstract In modern manufacturing, hundreds of process variables are collected, and it is usually difficult to identify the most informative ones. Partial Least Square Regression provides an efficient way to evaluate each variable, but it cannot evaluate any variable subset as a whole. In the paper, a new framework of key process variable identification is proposed. It combines PLSR model and wrapper feature selection to firstly assess every variable individually and then the top variables in groups. Five datasets are tested, and the average classification accuracy is higher and the key process variables identified are less than the available approaches.

Keywords Classification • PLS • Variable Selection • Wrapper

Introduction

In modern manufacturing process, high dimensional process data play crucial parts in quality monitoring and diagnosis. Usually, the process variables are noisy and redundant, making it almost impossible to predict the quality effectively. Thus, identifying best "predictors" for quality classification is critical for process modeling, monitoring, and control (Su et al. 2006).

Partial Least Squares Regression (PLSR) is a well established statistical model, and it has lots of advantages as follows (Kettaneha et al. 2005). (1) It can deal with high multicollinearity between variables; (2) It requires smaller sample size than regular Multiple Linear Regression (MLR); (3) its parameters can be used to analyze variable importance. Also, PLSR methodology is helpful to identify best

Supported by National Natural Science Foundation of China (No.70931004, 70802043).

W.-m. Tian (✉) • Z. He • W. Yan
College of Management & Economics, Tianjin University, Tianjin, China
e-mail: megtiantju@gmail.com

R. Dou (ed.), *Proceedings of 2012 3rd International Asia Conference on Industrial Engineering and Management Innovation(IEMI2012)*,
DOI 10.1007/978-3-642-33012-4_27,

variables when combined with some other methods, such as data mining and feature selection (Anzanello et al. 2009, 2012). However, the available methods suffer from some serious problems. One is that correlations between variables are not taken into consideration. That makes their methods simple and flexible, while not effective for classification so as to evaluate each possible subset of variables.

On the other hand, Wrapper feature selection is widely used as a preprocessing technique to high dimensional datasets to find a best variable subset which has a good capability of classification (Kohavi and John 1997; Inza et al. 2004). Lots of wrapper algorithms have been developed to solve variable selection problems in microarray analysis, text classification, and industrial processes.

In the paper, a new variable selection methodology is proposed based on PLSR and wrapper feature selection techniques, making it easy to evaluate the variable subsets as a whole, instead of just calculating an importance index one by one.

The rest of the paper is organized as follows. A detailed introduction of the proposed methodology is presented in section "Methodology". Next, section "Proposed Framework" provides a framework of the proposed method. Then, section "Results" compares the experimental results with some other newly published methods. Last, the method is summarized and some future research topics are proposed in section "Conclusion".

Methodology

PLSR: Model Structure and Parameters

Partial Least Squares Regression (PLSR) model is not only widely used to model linear relationships between variables and responses, but also effective to deal with multicollinearity between the multiple variables or responses. Furthermore, it is highly tolerant to small sample sizes.

The most popular algorithms to implement PLSR are SIMPLS (Jong 1993) and NIPALS (Gerladi and Kowalski 1986). Both can be easily performed with the PLS functions in Statistical Toolbox of Matlab2011a.

The PLSR model can be developed from a training dataset of two matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$, which demonstrate N observations in K process variables and M final quality responses, respectively. In the model, a small number of components, $\boldsymbol{T}$ and $\boldsymbol{U}$, are extracted from original $\boldsymbol{X}$ and $\boldsymbol{Y}$. In fact, $\boldsymbol{T}$ and $\boldsymbol{U}$ are linear combinations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and they are often called "X-scores" and "Y-scores". Formulas are shown below, where $\boldsymbol{W}$ and $\boldsymbol{Q}$ are original matrices' weights on extracted components.

$$T = XW \tag{27.1}$$

$$U = YQ \tag{27.2}$$

Therefore, $\boldsymbol{T}$ and $\boldsymbol{U}$ could be good predictors of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and the residual matrices $\boldsymbol{E}$ and $\boldsymbol{G}$ should be very small. In Eqs. (27.3) and (27.4), $\boldsymbol{P}$ and $\boldsymbol{C}$ are loading matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

$$X = TP' + E \tag{27.3}$$

$$Y = UC' + G \tag{27.4}$$

Then, $\boldsymbol{T}$ can be good predictors of $\boldsymbol{U}$. See Eq. (27.5), $\boldsymbol{H}$ is the residual matrix, which should be "small enough" to ensure prediction accuracy.

$$U = TD + H \tag{27.5}$$

Giving the Eqs. (27.1) and (27.2), Eq. (27.5) can be rewritten as a multiple regression model, and $\boldsymbol{F}$ is the residual matrix.

$$Y = XB + F \tag{27.6}$$

The data structures in the model can be clearly shown in Fig. 27.1, which employed the structure in Hoskuldsson (1988) with slight modifications to make it more understandable. More mathematical details in PLSR model can be found in literature (Wold et al. 2001).

Variable Importance on Projection Based on PLSR model

Variable Importance on Projection (VIP) is firstly defined in Gerladi and Kowalski (1986). It is a weighted summary of the variable's importance for the response matrix $\boldsymbol{Y}$, and it can be easily obtained by the formula below.

$$VIP_k = \sqrt{\frac{K \sum_a {w_{ka}^*}^2 \times SSY_a}{SSY_T}} \tag{27.7}$$

Where

$$SSY_a = \frac{1}{M} \sum_{m=1}^{M} \frac{Cov(y_m, t_a)^2}{Var(y_m) Var(t_a)} \tag{27.8}$$

$$SSY_T = \sum_{a=1}^{A} SSY_a \tag{27.9}$$

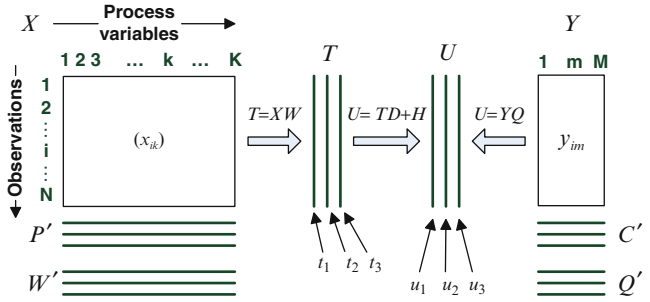

Fig. 27.1 The structure of partial least square regression

$$W^* = W(P'W)^{-1} \tag{27.10}$$

Here w^*_{ka} is the transformed x_k's weights of component a

Though there is another version of VIP in Anzanello et al. (2009), Eq. (27.7) is the most popular one. Both versions illustrate the same relationship between $\boldsymbol{Y}$ and $\boldsymbol{T}$.

Wrapper Feature Selection

Feature selection (FS) is a commonly used data preprocessing technique to identify most informative features (also called variables) for a better classification model (Guyon and Elisseeff 2003). There are two mainstream methods of feature selection, filter and wrapper.

Filter is to calculate a certain evaluation index, such as Information Gain (Hall and Holmes 2003) and Symmetrical Uncertainty (Yu and Liu 2003), for each variable, and then eliminate the variables with a "low value". Wrapper, on the other hand, is to search the feature space to find a variable subset with optimized classification accuracy based on a certain classifier. Compared with filters, wrappers can find a more satisfactory subset for a certain classifier (Hua et al. 2009). As our objective of variable selection is to build a more reliable, yet less complicated classification model to predict the quality of products, wrapper method can achieve better results due to its inherent advantage.

There are three key elements in a wrapper method, a search engine to generate variable subsets, a learning algorithm, and an evaluation criterion. The implementation framework is shown in Fig. 27.2. More detailed information about wrapper feature selection can be found in the literature (Kohavi and John 1997).

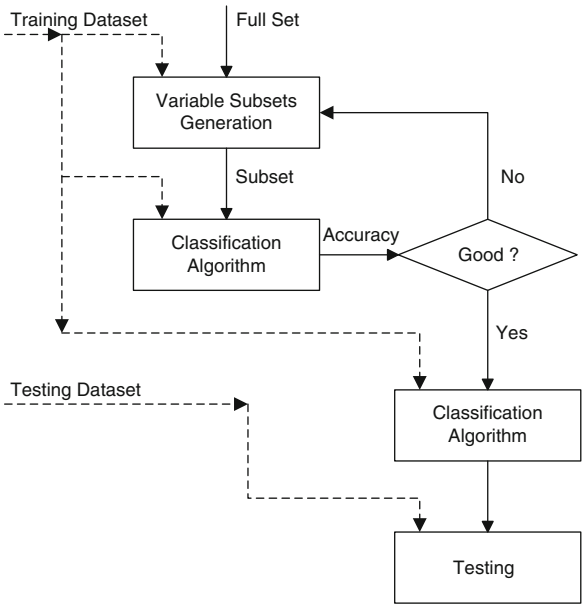

Fig. 27.2 The structure of wrapper feature selection

Proposed Framework

The framework of key process variable identification includes the following 4 steps. Step 1 is data preparation; Step 2 is to construct a PLSR model, and obtain the VIP values of the process variables; Step 3 implements a wrapper feature selection with the SFFS search engine and KNN classifier to evaluate the performance of each feature subset. At last, a testing step will be performed in Step 4, using the reduced testing dataset to predict the goodness of the variable selection.

Step 1: Split the original dataset into two exclusive subsets

The original dataset is separated into training and testing datasets by an appropriate proportion, say, 3:2 or 4:1. Also, as the process responses are continuous, some cut-off value of the responses should be obtained to meet the needs of classification. In this case, the proportion of different response classes should be almost the same in the training and testing datasets.

Step 2: Construct a PLSR model with the training dataset, and calculate VIP for each process variable

The process data should firstly be normalized or the data analysis would be affected by different scales of different variables. NIPALS algorithm can be implemented in the Statistical Toolbox of Matlab2011a. In the meantime, the parameter matrices can be calculated, and the VIP index of each variable should be obtained from Eq. (27.7). Also, the variable index should be reordered according to the descending order of VIP.

Step 3: Apply Wrapper feature selection to the training dataset to search for the optimal subset

Wrapper feature selection is to implement a heuristic search in the state space of the variables to find an optimal (or suboptimal) subset with the best performance of a certain learning algorithm. In this paper, Sequential Floating Forward Selection (SFFS) (Pudil et al. 1994) is our choice of the search engine, and K-Nearest Neighbor classification (Aha et al. 1991) is used to evaluate the subsets in each iteration step. The detailed algorithm of SFFS could be found in 16, and it is chosen for this good capability in jumping out of local optimum.

The KNN classification should be our first choice of learning algorithm for it is easy to understand, efficient for computation, and has only one parameter k to set in building the model (Anzanello et al. 2012). Furthermore, the appropriate value of k can be obtained from cross validation. The distance is defined as Euclidean distance for it is widely used in instance-based classification rules. By majority voting of the k nearest training samples, the class label of the testing sample can be predicted efficiently (Aha et al. 1991).

Step 4: Classify the testing dataset with the optimal subset of variables obtained in Step 3

Based on the optimal variable subset obtained from Step 3, the testing dataset can be examined to compute the classification accuracy. In this case, the parameter k and Euclidean distance are applied so that the model for testing is identical with the result of wrapper feature selection.

Results

To justify the effectiveness of the proposed framework, five datasets from real industry are used. Also, the proposed PLSR-Wrapper framework is compared with the method in literature (Anzanello et al. 2009) and the PLSR model in literature (Gauchi and Chagnon 2001). Both of the two methods above have been applied to these five datasets, and the testing result can be easily employed for a fair comparison.

All the datasets, namely ADPN, LATEX, OXY, SPIRA, and PAPER, are from chemistry industry. They are real process data from production of nylon, latex, titanium dioxide, antibiotics, and paper recycling, respectively. The samples of each datasets are categorized into two classes with a proper cut-off value of the final response for each dataset. The values come from (Anzanello et al. 2012; Gauchi and Chagnon 2001).

The PLSR model is applied to the five datasets just as it is in Anzanello et al. (2009), and the VIP index for each variable is calculated. The variable index is reordered in the descending order of VIP so that it can be used in the wrapper feature selection method.

Table 27.1 Comparison of available and proposed approaches

	All variables used		After variable selection CCR		Percent of variables retained	
Datasets	KNN CCR	PLS CCR	Method in [3]	PLSR-Wrapper	Method in [3]	PLSR-Wrapper
ADPN	0.78	0.86	0.87	0.86	8.0%	11.0%
LATEX	0.78	0.83	0.87	0.77	7.7%	7.7%
PAPER	0.81	0.59	0.83	0.90	18.5%	3.7%
OXY	0.73	0.73	0.73	0.90	6.3%	2.1%
SPIRA	0.86	0.83	0.9	0.93	4.2%	6.3%
Average	0.792	0.77	0.84	0.872	8.9%	6.15%

Then, the wrapper feature selection algorithm with a VIP-defined order is performed. The summary of the performance of the proposed framework and some previous methods are shown in Table 27.1. In the table, two variable subset performances, classification accuracy and percent of retained variables of different methods are compared.

In Table 27.1, the first column is the name of the datasets; the second and third column record the Classification Correct Rate (CCR) with all variables to construct the KNN and PLS model; the next two demonstrate the CCR of variable selection method in Anzanello et al. (2009) and in this paper, respectively; the last two show percent of retained variables of both methods. It is indicated in the table that both variable selection approaches can retain a small percent of variables while obtaining a higher CCR. Comparing with the approach in Anzanello et al. (2009), the proposed method is more effective in dimension reduction and it can obtain more accurate classification models as well.

Conclusion

A new framework combining PLSR model and wrapper feature selection is proposed to identify the key process variables in high dimensional process data for quality classification. VIP is applied to determine the relative importance of variables, and to generate a reordered variable sequence. A wrapper feature selection method is employed, with SFFS to heuristically search the space and KNN to evaluate the variable subsets in each iteration step.

The framework has been applied to five widely tested datasets in chemistry industry, and the experimental results indicate that the proposed method can construct a classification model with better average performance with a smaller percent of retained variables.

Future research includes how to determine which classification algorithm or variable selection method should be applied for different datasets, for the result also indicates that the proposed framework is not equally effective to all the five datasets.

Acknowledgment We would like to express our gratitude to Prof. Jean-Pierre Gauchi for providing the datasets of ADPN, LATEX, OXY, and SPIRA; and to Prof. Svante Wold for providing the PAPER dataset and some helpful advice about PLSR model. We also thank Dr. Michel J. Anzanello and Prof. Susan L. Albin for their supportive advice and encouragement during the algorithm testing.

References

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6:37–66

Anzanello MJ, Albin SL, Chaovalitwongse WA (2009) Selecting the best variables for classifying production batches into two quality levels. Chemom Intell Lab Syst 97:111–117

Anzanello MJ, Albin SL, Chaovalitwongse WA (2012) Multicriteria variable selection for classification of production batches. Eur J Oper Res 218:97–105

Gauchi J, Chagnon P (2001) Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemom Intell Lab Syst 58:171–193

Gerladi P, Kowalski BR (1986) Partial least squares regression: a tutorial. Anal Chim Acta 185:1–17

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hall MA, Holmes G (2003) Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 15(3):1437–1447

Hoskuldsson A (1988) PLS regression methods. J Chemom 2:211–228

Hua J, Tembe WD, Dougherty ER (2009) Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognit 42:409–424

Inza I, Larranaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 31:91–103

Jong S (1993) SIMPLS: an alternative approach to partial least squares regression. Chemom Intell Lab Syst 18:251–263

Kettaneha N, Berglundb A, Wold S (2005) PCA and PLS with very large data sets. Comput Stat Data Anal 48:69–85

Kohavi R, John GH (1997) Wrappers for feature selection. Artif Intell 97:273–324

Pudil P, Novovicova J, Kittler J (1994) Floating search methods in feature selection. Pattern Recognit Lett 15:1119–1125

Su C, Chen L, Chiang T (2006) A neural network based information granulation approach to shorten the cellular phone test process. Comput Ind 57:412–423

Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the twentieth international conference on machine learning (ICML-2003), Washington, DC