Yongsheng QI[1], Yongting LI[1], Zhicheng SUN[1]

Inner Mongolia University of Technology (1)

A novel stage-based KPLS-PLS monitoring and quality prediction approach for batch processes

***Abstract**. A novel KPLS-PLS batch monitoring and quality prediction approach based on fuzzy clustering soft-partition is proposed to solve the stage-transition monitoring and prediction problem in multistage batch processes. The proposed method calculates firstly similarity indices between different time-slice data matrices of batch processes, then phase division algorithm is designed by fuzzy clustering based on the similarity index, following by a fuzzy membership grade transition identification step. By setting a series of KPLS and PLS models with time-varying covariance structures for transitions and steady phases, it reflects objectively the diversity of transitional characteristics, capture the nonlinear relationships among process variables of the transition and can monitor and predict batch processes more accurately and efficiently. The superiority of the proposed method is illustrated by applying it to industrial application of fed-batch penicillin fermentation process. The results clearly demonstrate the effectiveness and feasibility of the proposed method*

Streszczenie:** Zaproponowano nową metodę KPLS (kernel partial least squers) – PLS monitorowania i przewidywania wieloetapowych procesów wsadowych. Metoda oparta została o klastrowanie rozmyte, pozwala na wykrycie przejść między etapami i dokładniejsze przewidywanie przebiegu procesu przez uniknięcie wpływu nieliniowości. Wyższość proponowanej metody zilustrowano wykorzystując ją do badania przemysłowego procesu fermentacji wsadu pożywki penicyliny. **Nowa etapowa metoda KPLS – PLS badania monitoringu i przewidywania jakości procesów wsadowych

Keywords: Batch Monitoring, Multiphase, Partial Least Squares, Kernel Partial Least Squares.
Słowa kluczowe: Monitoring wsadu, Wielofazowość, Metoda uogólnionych najmniejszych kwadratów

Introduction

Batch or semi-batch processes have been utilized to produce high-value-added products in the biological, pharmaceutical, food, semi-conductor industries. In order to get higher productivity, it is necessary to ensure that condition of batch processes remains closely fixed around a pre-specified trajectory. The common natures of multi-phase, time-varying, finite duration and batch-to-batch variations make batch processes more difficult to control than continuous processes. Hence, proper process monitoring and quality prediction is important to not only quality improvement but also process safety. Multivariate statistical methods based on multi-way partial least squares (MPLS) proposed by Nomikos and MacGregor [1] have been widely employed in batch process monitoring and quality prediction.

In industries, many batch processes are carried out in a sequence of steps, which are called multiphase batch processes [2]. Different phases may have different process natures. Traditional MPLS method takes the entire batch data as a single object in modeling. Therefore, the unique process correlation information of different phases is not reflected. This not only makes difficulties on understanding of process nature, but also affects monitoring efficiency. Considering that multiple phases with transitions from phase to phase are important characteristics of many batch processes, it is desirable to develop stage-based models. Then each model represents a specific phase and focuses on the local behaviors of the batch processes. In recent years, different phase-division methods have been proposed and different modeling methods have been developed that take the phase effects into consideration. Kosanovich et al developed two-stage MPLS models to analyze the phase-specific nature of a two-phase jacketed exothermic batch chemical reactor [3]. Zhao investigated multiple PLS models for different operating modes based on metrics in the form of principal angles to measure the similarities of any two models, which have been applied to continuous processes [4]. Lu *et al* developed a stage-based PLS modeling method based on the fact that changes in the process correlation may relate to its stages [5].

However, the above methods are all strict stage partition algorithms, which neglect the stage-to-stage transiting characteristics. The misclassification may occur at the beginning and end of each stage, because the hard-partition methods in dealing with patterns between two neighboring clusters, and may lead to false alarm and missing alarm in on-line monitoring due to batch variation. Moreover, stage nature can be reflected by the process variable correlations, and the nonlinear characteristics of process are often related with the process stage closely. Stage-based multiple PLS models is a partial linearization method for the nonlinear batch process in essence. However, transition between phase to phase has more nonlinear characteristics compared to steady phase, due to transitions from phase to phase are important characteristics and the corresponding phases in different batch processes are uneven lengths. To monitor and predict batch processes more accurately and efficiently, the nonlinear features of transition are needed to be considered carefully.

In the present article, we present a novel KPLS-PLS batch monitoring and quality prediction approach based on fuzzy clustering soft-transition. The method proposed here solves the above problems. This paper is organized as follows. First, the similarity index, the MPLS and KPLS monitoring method are outlined and a KPLS-PLS algorithm is suggested. Then, the superiority of the proposed monitoring method over traditional MPLS is illustrated by applying to the industrial penicillin fermentation process. Finally, conclusions are given.

Materials and methods

Similarity index

The concept of similarity or dissimilarity is often used for classifying a set of data. For example, the dissimilarity between two classes is measured and the two classes with the smallest degree of dissimilarity are combined to generate a new class. To evaluate the difference between distributions of data sets, a classification method based on the Karhunen-Loeve(KL) expansion is used in this work, which is a well-known technique for feature extraction or dimensionality reduction in pattern recognition.

Considering the following two data sets with the same number of columns and arbitrary rows, $X_1 \in \Re^{M\times J}$ and $X_2 \in \Re^{N\times J}$, the dissimilarity index D is defined as follows[6]

(1) $$D = dist(X_1, X_2) = \frac{4}{J}\sum_{j=1}^{J}(\lambda_1^j - 0.5)^2 = \frac{4}{J}\sum_{j=1}^{J}(\lambda_2^j - 0.5)^2$$

Where, λ_j is the eigenvalue of the covariance matrix of the transformed matrix obtained from one data set X_i, which is defined as $\sqrt{(N_i - 1)/(N_1 + N_2 - 2)}X_i \mathrm{P}_0 \Lambda^{-1/2}$, where $P_0{}^T S P_0 = \Lambda$(Λ is a diagonal matrix whose elements are eigenvalues of S) and S=1/(N_1+N_2-2)($X_1{}^T X_1$+$X_2{}^T X_2$). J is the number of process variables. The dissimilarity index, D, has been shown to change between zero and one. Here, each column of X_i is assumed to be mean-centered and scaled.

MPLS.

The dataset of a batch process can be arranged as a three-way array. A batch run has J input variables (j = 1,2,...,J) and M output variables (m= 1,2, ... , M) measured k times (k = 1,2,...,K) throughout the batch. Data of the same form exist for each of the I batch runs (i = 1,2,...,I) stored in a historical database. The three-way array $X(I\times J\times K)$ can be unfolded in three ways, which give rise to the different two-dimensional matrices .In this paper, the process measurements array X is unfolded to $X(KI\times J)$ by preserving the variable direction. As a result, MPLS decomposes $X(KI\times J)$ and $\mathrm{Y}(KI\times M)$ matrices with mean zero into the form.

(2) $$X = TP^T + E; \quad Y = UQ^T + F$$

Where $T(KI\times R)$ and $U(KI\times R)$ are the scores, $P(J\times R)$ and $Q(M\times R)$ the loadings and $E(KI\times J)$ and $F(KI\times M)$ are the residuals. This approach has the merits that it does not require estimation of future missing values. The unfolded X matrix is then related to the quality variable using PLS.

(3) $$Y = XC^{PLS} + V$$

Where C^{PLS} is an inner coefficient vector of J×1 and V is a residual vector of KI×1.Details on the PLS algorithms can be found in Ref[7].

For on-line monitoring using MPLS, two statistics, T^2 and the squared prediction error (SPE), are generally calculated. T^2 is defined as follows and its confidence limits can be obtained from the indicated F-distribution

(4) $$T_k^2 = t_{new,k}\Lambda_k^{-1}t_{new,k}{}^T \sim \frac{R(I^2 -1)}{I(I-R)}F_{R,I\text{-}R,\alpha}$$

where $t_{new,k}(R\times 1)$ are the scores of the new batch at time k, Λ_k is the covariance matrix of T_k calculated during development of the model. In the proposed approach, we use time-varying covariance Λ_k to replace fixed covariance Λ at each time during a batch, which is considered to incorporate the major dynamic characteristics of the batch process and can obtain better monitoring performance [8]. The SPE is defined as the sum of the squares of the errors at time k. The confidence limit of SPE can be obtained from the following χ^2 distribution.

(5) $$SPE_k = e_{new,k}e_{new,k}^T \sim (v_k / 2m_k)\chi^2_{2m_k^2/v_k}; \qquad e_{new,k} = x_{new,k}(I - PP^T)$$

Where $x_{new,k}(J\times 1)$ is new batch data at time k, m_k and v_k are the mean and variance of the SPE at time k obtained for the dataset used to develop the model.

KPLS.

In general, the PLS can be effectively performed only on a set of observations that vary linearly. When the variations are nonlinear, the data can be mapped into a higher dimensional space in which they vary linearly. According to Cover's theorem, the nonlinear data structure in the input space is more likely to be linear after high-dimensional nonlinear mapping. This higher-dimensional linear space is referred to as the feature space. KPLS is formulated in this feature space to extend linear PLS to its nonlinear kernel form.

First, consider a nonlinear transformation of the input variables x_i, i =1,. . .,n into feature space F: $x_k \in R^{m\rightarrow}\Phi\in F$. where it is assumed that $\sum_{k=1}^{N}\Phi(x_k) = 0$, and φ is a nonlinear mapping function that projects the input vectors from the input space to F. Through the introduction of the kernel trick, $\Phi(x_i)^T\Phi(x_i)$=$K(x_i,x_j)$, one can avoid both performing explicit nonlinear mappings and computing dot products in the feature space. The matrix of the regression coefficient *B* in the KPLS algorithm will have the form

(6) $$B = \Phi^T U(T^T KU)^{-1}T^T Y$$

As a result, when the number of test data is n_t(1,. . ., n_t), the predictions on training data and test data can be made as follows, respectively

(7) $$\hat{Y} = \Phi B = KU(T^T KU)^{-1}T^T Y; \qquad \hat{Y}_t = \Phi_t B = K_t U(T^T KU)^{-1}T^T Y$$

Where, Φ_t is the matrix of the mapped test points and is K_t the (n_t×n) test matrix whose elements are K_{ij}=$K(x_i, x_j)$, where x_i is the i-th test vector and x_j is the j-th training vector. The detailed description of KPLS algorithm can be found in the work by Kim[9].

Stage-based KPLS-PLS algorithm

In this section, a similarity-based phase division and transition identification method is proposed. The procedure is plotted in Fig.1 and the detailed description is given below.

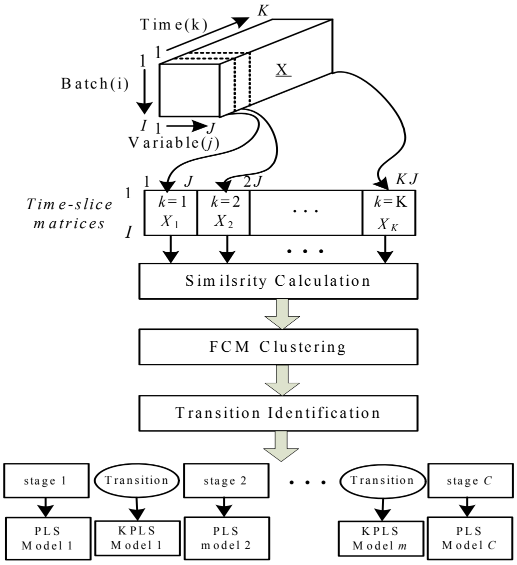

Fig. 1. Illustration of fuzzy clustering soft-transition KPLS-PLS algorithm

Similarity-based phase division and transition identification

(1)Unfold three-way batch data matrix to $X(I\times JK)$, and normalize the unfolded data matrix.(2) Reorganize the normalized data into a three-way data matrix $X(I\times J\times K)$ and split it into *K* number of time-slice data matrices $X_k(I\times J)$ (k = 1,2,3,...,K).(3) Calculate similarity index D_i(i=1,2,...,K) between time-slice data matrices X_i and X_j(j=1,2,…,K and i≠j) as follow

(8) $$D_i(k) = dist(X_i, X_j) \begin{cases} k = j & j = 1,2,\cdots,i-1 \\ k = j-1 & j = i+1,\cdots,K \end{cases}$$

(4) Cluster the K number of vectors D_i into C clusters with fuzzy C-means clustering algorithm.(5) At each cluster (phase), the values of the maximum membership grade at each sample time are plotted on an univariate control plot, in order to detect the outliers. The successive outliers occur at the beginning and end of each cluster are identified as the points in transitions and removed from each cluster.

Then, the remained part of each cluster is the range of steady phase.

Transition identification makes use of univariate statistical monitoring method to identify the transitions as outliers. The reason of doing so is that in each steady phase, the values of similarity index between time-slice data matrices should be similar, and their distribution can be approximately regarded as normal distribution. In contrary, the similarity index between the time-slice data matrices in transitions and steady phase could be quite different. Thus, the transition samples can be detected as outliers with the univariate control plots at the beginning and end of each phase.

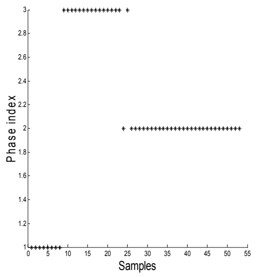

Fig.2. Phase division result

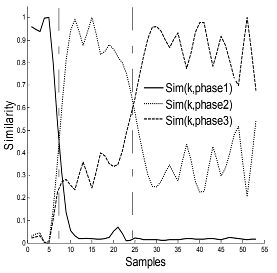

Fig.3.Process nature

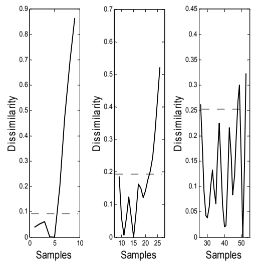

Fig.4.Transition ranges identification

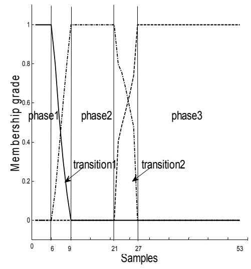

Fig.5.Sketch map

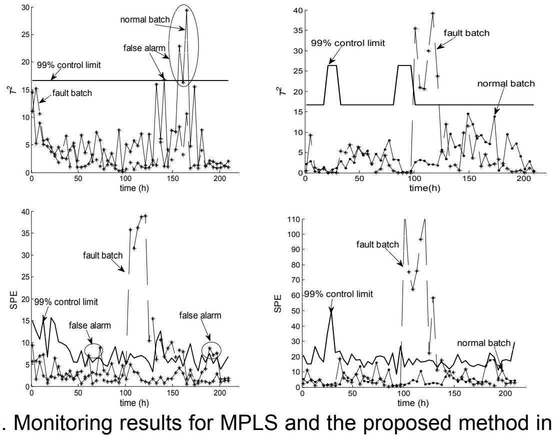

Fig.6. Monitoring results for MPLS and the proposed method in two test batches

Phase modelling and transitions modelling

After the identification of the ranges of steady phases and transitions, the procedure of modeling is given as following.(1) At each steady phase, perform improved PLS on the steady phase data matrix $X_c(n_cI\times J)$ and $Y_c(n_cI\times M)$, where n_c is the number of samples in steady phase *c*.(2) At each transition, perform KPLS on the transition data matrix based on variable-wise unfolding $X_s(n_sI\times J)$ and $Y_s(n_sI\times M)$, here n_s is the number of samples in transition *s*.(3) A cross-validation method based on the predicted residual error sum of squares (PRESS) was used to select the number of latent variables for the improved PLS and KPLS models.(4) Calculate the confidence limits of T^2 and SPE statistics at each time k (k = 1,2,...,K) based on corresponding steady phase or transition model.

On-line monitoring and quality prediction

(1) For new sampling data at time k, $x_{new,k}(1\times J)$, normalize it using the same mean and standard deviation obtained from the modeling procedure.(2) According to process time, project the new data $x_{new,k}$ onto the corresponding steady phase or transition model to calculate the latent vectors $t_{new,k}$.(3) According to $t_{new,k}$, calculate the T^2 and SPE statistics based on corresponding model.(4) Monitor the T^2 and SPE statistics of test data whether T^2 or SPE exceeds its confidence limit calculated in the modeling procedure.(5) Predict quality vector $y(1\times M)$ using PLS regression model according to Eq.3 or KPLS regression model according to Eq.7 in the corresponding steady phase or transition. Go back to step1 and process the next measurement.

Case studies

In this section, the proposed method was applied to the monitoring an industrial penicillin fermentation process. In our study, the model has been developed by carefully selecting 24 good batched from the history data that reflect the normal desired operation of the fermentation. The 9 variables considered for monitoring include: Temperature, pH, Air flow, Agitator current, Culture volume, Sugar feed rate, Ammonia feed rate, Phenyl acetic acid feed rate, Ammonium sulphate feed rate. The duration of each batch was about 212h. The sampling interval is 4h. According to the proposed method, the whole fermentation process was automatically divided into three steady phases which cover the sampling intervals 1~6, 10~21 and 28~54, and two transitions which include the sampling intervals 7~9 and 22~27.

The phase division results are plotted in Fig.2. Using the proposed partition method, the real fed-batch phase is divided into two main stages and the whole process is divided into three primary stages as well as corresponding transition regions. It is clear that the phase division is consistent with the process nature. In penicillin cultivation process, process nature changes with operation time. Such changes can be indicated with the trends of the similarity values Sim(k, phase c), as shown in Fig.3. Sim(k, c) is the membership grade between the *k*th time-slice data matrix and the *c*th cluster by FCM given. The values of Sim(k,c) changes gradually with the process operation. It becomes larger when the process approaching to phase *c*. During phase *c*, Sim(k,c) keeps large values which indicate high similarities between the time-slice data matrices and the current cluster-center. When process operates far away from phase *c*, the similarities become small again. The gradual changes of Sim(k,c) values at the beginnings and ends of phases confirm the existence of transitions from phase to phase.

The univariate statistical process control plots are utilized to identify the transition ranges. For each phase, the values of dissimilarity 1-Sim(k,c) ($k\in$c) are plotted in Fig.4. As introduced in Section 3.1, the successive outliers at the beginning and the end of each phase indicate the transition ranges. After transition identification, each steady phase and transition range are modeled with the method described in Section 3.2. The membership grades are plotted in Fig.5. Again, the transition attributes from phase to phase are shown clearly. The models that constructed using traditional MPLS and the proposed method are then tested against monitoring of two different operating states batches.

Fig.6 shows the monitoring results for two batches, a normal and a fault batch, using two methods. To normal batch, in T^2 and SPE monitoring charts of the MPLS (see fig.7(a)), false alarm rates are 5.7% and 7.5% respectively, which significantly deviates from the 1% set value of control limit. In contrast, no false alarm in either SPE or T^2 monitoring charts of the proposed method, and it achieves a successful monitoring to the normal batch. To fault batch (in the fermentation medium, the air flow of fermentation process is rapidly reduced by some mechanical failure, but due to the operator's adjustment, the process restore to normal operating condition, the final penicillin titer is acceptable.), the T^2 and SPE values of the proposed method increase beyond control limit at time 100h. The detection time is faster than that of MPLS by 4h. However, the T^2 charts of MPLS can not provide good detection, and does not give any alarm in the abnormal batch. In addition, after the fault is eliminated, MPLS method can not quickly shake off the impact of the failure, leading to the false monitoring results. The reason is that MPLS ignores the multiphase characteristics and takes the batch as whole, but the proposed method compensates the shortcoming of traditional MPLS. Through comparison of two methods, we can draw the conclusion that the proposed method more objectively assesses the process of production and the effort of operators.

Fig.7 shows quality prediction results using two methods for a normal batch. In addition, choose root mean square error (RMSE) as accuracy evaluation criteria, as presented in Table 1. RMSE can be calculated as follows

(9) $$RMSE = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(\hat{y}_i - y_i)^2}$$

Where y_i is actual quality value, $\hat{y}_i$ is predictive quality value, p is sampling time. As can be seen from the Fig.9 and Table.1, KPLS-PLS clearly outperforms MPLS in predicting penicillin titer of the batch. Moreover, to original MPLS, prediction results of phase 1 and phase 2 (0~100h) is much closer to the actual values than those of other phases. It indicates that a single model may accurately describe only one or several of the phase characteristics, which often leads to a larger prediction error under the other phases. So, the phase-based approaches are intuitively well suited for multi-phase batch process monitoring and quality prediction.

Conclusions

Multiphase/multistage batch processes widely exist in industrial applications. These processes can be divided into several phases based on variable correlation structure changes. In many cases, processes operate from one phase to another through gradual transitions. In this work, a novel KPLS-PLS batch monitoring and quality prediction approach based on fuzzy clustering soft-transition is proposed taking the process features of multiple phases and gradual transitions into consideration. The proposed method reflects objectively the diversity of transitional characteristics, extract the nonlinear relationships among process variables of transition and can preferably solve the stage-transition monitoring problem in multistage batch processes. The proposed method was applied to the industrial penicillin fermentation process, the results indicate that the proposed method is feasible and shows good efficiency.

Table 1. Comparison of RMSE of Two Methods

	MPLS	KPLS-PLS
RMSE	0.1034	0.0188

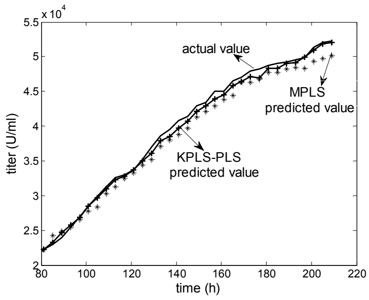

Fig.7. comparison of estimation results of Penicillin titer

Acknowlegements

This work was supported in part by NSFC under Grant Nos. 61174109 and the Science Fund of Inner Mongolia University of Technology Grant ZS0201037.

REFERENCES

[1] Nomikos P., MacGregor J. F.. Multivariate SPC charts for monitoring batch process, *Technometrics*, 37(1995), No. 1, 41-59.

[2] Undey C., Tatara E.. Intelligent real-time performance monitoring and quality prediction for batch/fed-batch cultivations, *Journal of Bitechnology*, 108(2004), 61-77.

[3] Kosanovich K. A., Piovoso M. J., Dahl K. S.. Multi-Way PCA Applied to an Industrial Batch *Process, In Proceedings of the American Control Conference*, Baltimore, MD, June, (1994)

[4] Zhao S. J., Zhang J., Xu Y. M.. Performance monitoring of processes with multiple operating modes through multiple PLS models, *Journal of Process Control*, 16(2006), No. 7, 763-772.

[5] Lu N., Gao F.. Stage-based process analysis and quality prediction for batch processes, *Industrial and Engineering Chemistry Research*, 44 (2005), No. 10, 3547-3555.

[6] Zhao C., Wang F., Gao F.. Adaptive Monitoring Method for Batch Processes Based on Phase Dissimilarity Updating with Limited Modeling Data, *Industrial and Engineering Chemistry Research*, 46(2007), No. 14, 4943-4953.

[7] Malthouse E. C., Tamhane A. C., R. Mah S. H.. Nonlinear partial least squares, *Computers and Chemical Engineering*, 21(1997), 875–890.

[8] Lee J.M., C. K. Yoo, et al. Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis, *Journal of Biotechnology*, 110(2004), No. 2, 119-136.

[9] Kim K., Lee J. M., Lee I. B.. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction, *Chemometrics and Intelligent Laboratory Systems*, 79(2005), 22–30.

***Authors**: Ph.D. Yongsheng QI, School of Electric Power, Inner Mongolia University of Technology, Huhhot 010051, China. E-mail: qyslyt@163.com; Ph.D.Yongting LI, School of Electric Power, Inner Mongolia University of Technology, Huhhot 010051, China. E-mail: liyongting1234@163.com.*