CHEMICAL ENGINEERING TRANSACTIONS
VOL. 61, 2017
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Petar S Varbanov, Rongxin Su, Hon Loong Lam, Xia Liu, Jiří J Klemeš
Copyright © 2017, AIDIC Servizi S.r.l.
ISBN 978-88-95608-51-8; **ISSN** 2283-9216
DOI: 10.3303/CET1761236

A Novel Training Sample Selection Approach for Near-Infrared Spectroscopy Model and Its Industrial Application

Kaixun He[a,*], Yiran Li[b], Kai Wang[c]

[a]College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590 China
[b]School of Information and Control Engineering, China University of mining and technology, Xuzhou 221116 China
[c]East China University of Science and Technology, Shanghai 200237, China
kaixunhe@sdust.edu.cn

Near-infrared (NIR) spectroscopy has been widely applied for the real-time measurements of quality variables, which plays an important role in process control, monitoring and optimization. Since the prediction accuracy of NIR model strongly depends on the structure of training samples, it is important to optimize the process of training samples selection. Therefore, in the present work, a cross validation based approach which combined with kmeans++ algorithm is developed for this optimization. Based on the results, an efficient adaptive multi-model approach can be developed. During online application, according to the similarity distance between query sample and sub-models, the optimal sub-model can be selected and the high-performance predictions can be achieved. The usefulness and superiority of the proposed method is demonstrated and compared with other modeling algorithms in a real-world gasoline blending process in China.

1. Introduction

In process industry, key product quality should be measured accurately and timely in order to produce high-quality products (Bakirov et al., 2017). However, traditional lab analyses are expensive, time consuming, and introduce a significant time delay to the optimal control system. During recent years, near-infrared (NIR) spectroscopy has been widely employed as an online process analytical tool (PAT) to address these issues. The dominant advantage of this method is its ability to provide estimation results much more rapidly with little or no sample preparation (Mei et al., 2016).

By using NIR-based analytical tool, difficult-to-measure key properties are estimated by the NIR spectra using statistical or machine learning techniques based on Beer's law. He et al. (2015) has reported this application in online gasoline blending process and a dual updating strategy was adopted to improve the accuracy of NIR model. In this strategy, Local weighted strategy was used in sampling intervals and recursive method is adopted when new reference samples become available. Obviously, for its application, the key step is to establish NIR quantitative calibration model. Based on the estimated properties, online optimal control can be carried out.

Due to the ability to deal with co-linearity as well as high dimensionality, principal component regression (PCA) and partial least-squares regression (PLS) have long been widely adopted (Quiñones et al., 2014). For the properties which have nonlinear relationship with NIR absorbance, nonlinear PLS, artificial neural networks (ANNs), the support vector machine based regression method (SVR) and Gaussian process regression (GPR) are used (Balabin and Lomakina, 2011). All these mentioned methods are static and global based strategy, which have been adopted as useful methods for online prediction in the last decades. Nevertheless, the static based models cannot always function well due to changes of process raw materials, process fouling, and etc. (Kadlec et al., 2011). To cope with such issue, various adaptive strategies have been proposed, such as recursive or incremental based algorithm, moving window strategy, Just In Time Learning (JITL), local weighted regression (LW) and etc.

The typical representatives of incremental methods is recursive partial least-squares (RPLS), which expands training dataset by adding every available new sample. When the sampling operation is continuously, it can update the original model and capture the new variation of the process. However, since the reference properties

Please cite this article as: He K., Li Y., Wang K., 2017, A novel training sample selection approach for near-infrared spectroscopy model and its industrial application, Chemical Engineering Transactions, 61, 1429-1434 DOI:10.3303/CET1761236

have to be analysed offline, sampling interval is long and not uniform generally. Therefore, the model cannot be updated timely and does not deliver satisfactory predictions in real world application. The moving window approach abandons old data while new samples are acquired. Hence, it has the similar issues as recursive algorithms. Recently, JITL and LW strategies gained popularity because of their ability to deal with nonlinearity as well as abrupt changes. Due to the updating process does not depend on the new sample's reference information, both of them can adjust the calibration model to capture the current state of process timely. However, the number of local training samples is difficult to determine. More samples be included will lead to a large online computation load, while fewer samples lead to a deterioration of the performance (Ge and Song, 2010). Additionally, training samples are only selected based on similarity distance, in this manner, the dependent variable information and process knowledge are not taken into consideration arbitrarily. Hence, with an unsuitable similarity criterion, it is possible to construct a training dataset which leads to a larger prediction error. All of these features limit its application in the real-world industry.

As discussed above, for the process characteristic with large sampling interval, strong nonlinearity, and multi operation condition, it is not possible to achieve an adequate exactitude in the predictions using just one model for a wider or the entire range. As we know, LW strategy is a suitable approach in practical. But, in order to get a high-performance model, the training samples should be carefully selected.

Motivated by these issues, this paper intends to develop a novel supervised training sample selection method and establish a local weighted partial least square (LWPLS) NIR model for online prediction. The proposed method includes two steps: (1) offline process: the main task of this procedure is to divide the original dataset into several sub sets using kmeans++ algorithm and then optimize each sub set based on cross-validation. (2) online process: this procedure selects the optimal sub-dataset for each query sample and build the corresponding local model using local weighted strategy. Based on the two steps, the most relevant model can be determined for each new sample and the good prediction performance can be desired. To verify the effectiveness of the novel strategy, an industry case study is provided.

2. Theory and algorithm

Historical NIR data are sampled from multiple operating processes, which are characterized by inherent nonlinearity and shifting dynamics. Hence, it is difficult to select appropriate training dataset and construct an accurate NIR model for a specific process. In practical, this procedure is mainly depended on the process knowledge of experienced engineers, however, it is time-consuming. Additionally, for online application, the nonlinear and adaptive forms of NIR model should be adopted to cope with nonlinearity and time variance issues. This motivated us to explore a new sample selection and model updating approach named 'k-means and cross validation based local weighted PLS (KmCv-LWPLS), the details of the proposed method is presented in the following.

2.1 Dataset partition strategy

The procedure of our proposed method is summarized below.

Step 1. Determine the number of partitions k . In this work, the parameter k is determined by trial and error.

Step 2. Pre−process the training dataset $\{\mathbf{X}^{train}, \mathbf{Y}^{train}\}$, and extract feature components by PCA, then we can get the low-dimensional input variables $\mathbf{X}^{pca}$,

Step 3. Carry out k-means++ algorithm, get the initial clustering labels tag_j and clustering centers u_j ,

Step 4. According to the acquired label vector tag_j, the original dataset $\{\mathbf{X}^{train}, \mathbf{Y}^{train}\}$ is divided as

$$\{\mathbf{X}^{train}, \mathbf{Y}^{train}\} = \{(\mathbf{X}sub, \mathbf{Y}sub)_1, ..., (\mathbf{X}sub, \mathbf{Y}sub)_j, ..., (\mathbf{X}sub, \mathbf{Y}sub)_k\} \quad (1)$$

where $j = 1, 2, ..., k$, and then establish PLS model for each sub-dataset,

Step 5. Detect the boundary points of each sub-dataset,

Step 6. Holdout all the boundary points, and establish PLS model $subM_j$ for each new sub-dataset $subD_j$,

Step 7. Calculate the square error SE_b for all the boundary points using each $subM_j$,

$$SE_b = (y_b - \hat{y}_b)^2 \quad (2)$$

where y_b is laboratory analysis value of the boundary sample x_b and $\hat{y}_b$ denotes its predicted value.

Step 8. Allocate the boundary points to $subD_j$ which give the minimum prediction error.

To detect the boundary points of each sub-dataset (Step 5), a leave-one-out cross validation method is adopted. The basic idea of the proposed strategy is to evaluate the contribution of each training sample and sort them. Hereby, the points which deteriorate the performance of NIR model can be detected. The details of this method for each sub-dataset $(\mathbf{X}sub, \mathbf{Y}sub)_j$ are presented as follows:

1: For each si, holdout $(x, y)_{si}$ and build PLS model using the remaining samples. Where $si < N_j$ and N_j is the sample number of sub-dataset $(\mathbf{X}sub, \mathbf{Y}sub)_j$,

2: Calculate the prediction value $\hat{y}_{si}$ using the established PLS model,

3: Carry out leave-one-out cross validation algorithm to get the prediction value of the remaining samples $\hat{y}_{sj}$ (where $sj = 1, 2, ..., N_j$ and $sj \neq si$),

4: Calculate the root-mean-square error (RMSE),

$$RMSE_{si}^{j} = \sqrt{\frac{(y_{si} - \hat{y}_{si})^2 + \sum_{sj \neq si}^{sj=1\sim N_j} (y_{sj} - \hat{y}_{sj})^2}{N_j}} \tag{3}$$

5: Sort $(x, y)_{si}$ according to the ascending order of $RMSE_{si}^{j}$,

6: Holdout the former sk ($sk = 1, 2, ...$) samples and establish PLS model using the remaining data,

7: Calculate $RMSE_{sk}^{j}$ using the method described in Step 2- Step 4,

8: Increase the value of sk, and repeat Step 6-7 until $RMSE_{sk}^{j}$ reaches the minimum value,

9: The samples $\{(x, y)_1, (x, y)_2, ..., (x, y)_{sk}\}_j$ are denoted as boundary points and the new sub-dataset is denoted as $subD_j$.

2.2 Model selection and updating strategy

According to the procedure mentioned in Section 2.1, the optimal partition of the original dataset can be obtained. Then, for each query sample, the optimal sub-dataset should be selected and the corresponding sub-model can be established. In this study, the Euclidean distance $d_{q,k}$ between the query sample x_q and $\bar{x}_{sub}^{k}$ are calculated to detect the optimal sub-model.
Here,

$$d_{q,k} = \sqrt{(x_q - \bar{x}_{sub}^{k})(x_q - \bar{x}_{sub}^{k})^T} \tag{4}$$

And $\bar{x}_{sub}^{k}$ denotes the mean value of sub-dataset k. The sub-dataset with the minimum $d_{q,k}$ will be adopted.

As long as we can get the optimal sub-dataset, a prediction model can be established using PCR, PLS and etc. In Section 2.1, PLS is adopted to build NIR model due to its simplicity. However, as mentioned before, PLS is a static, global and linear based method, it may lead to inaccurate estimations in some local regions and its robustness is often jeopardized by process variations. Thereby, adaptive modelling and updating methods are necessary. Consider the uneven and low frequency sampling of reference data as well as the large time delay of its lab analysis, the traditional bias updating and recursive algorithms are not available in practical. Both of LW and JITL methods are widely applied to address such issues. LW method weights training sample x_i according to the similarity distance between x_i and x_q. While, JITL selects new samples from historical dataset

to build a new model for every query sample x_q. In this way, the predictive accuracy may drop due to changes of training samples and the number of training samples. Compared with JITL, LW does not change the structure of the original training dataset, thereby enabling improved the stability of NIR model within the updating process. Hence, in this research, local weighted method is adopted. The procedure of the adopted LW method has been mentioned in Section 2.2 and the weight evaluates the similarity between x_q and x_i is calculated as follows:

$$\omega_i = \exp(-\frac{\sum_{d=1}^{m}(x_{i,d} - x_{q,d})^2}{std(\sum_{d=1}^{m}(x_{i,d} - x_{q,d})^2) \times \alpha}) \quad (5)$$

Here α is a localization parameter. When α is small, the similarity decreases steeply, otherwise, it changes gradually (Kim et al., 2013).

3. Case study

In this section, one case study is provided to validate the practicability of our proposed method. The dataset was obtained from the real-world gasoline blending process. Four modelling approaches, namely, PLS, LWPLS, k-means based local weighted PLS (Km-LWPLS) and JITL are applied to NIR model development for comparison. Precisely, both the root mean square error (RMSE) and coefficient of determination (R^2) are defined as follows for quantity comparisons of different algorithms. The model which gives the lowest RMSE and the highest R^2 is considered best.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (7)$$

The four comparing algorithms investigated are as follows:
(1) PLS: A global PLS model is established. Its structure remains unchanged during the whole process.
(2) LWPLS: For every query sample x_q, different weights are assigned to the training samples based on ω_i, and then a new local PLS model is trained to predict the output.
(3) Km-LWPLS: In this approach, the original training dataset is divided into k clusters by k-means++ algorithm. Then, we can get k sub-models. For each query sample x_q, the optimal sub-model is selected and LWPLS is carried out to give the predicted values.
(4) JITL: The JITL method can establish local model for each query sample based on similarity distance. In this study, the similarity ω_i is used for the sake of simplicity. Local training samples were selected according to the equation $\omega_i > \gamma$, where γ is the similarity threshold. In addition, PLS algorithm is used to build a local model.

3.1 Gasoline Blending Process

Gasoline blending is a crucial unit operation in the gasoline industry. It is the final step before gasoline product be delivered (He et al., 2016). In this study, NIR model is adopted to predict the research octane number (RON) which is the key property of gasoline. A total of 312 samples have been collected from daily process records and the corresponding laboratory analysis. The spectra range was restricted to 1,100 nm to 1,300 nm, each NIR sample consists of 201 wavelength variables. Reference values of RON were measured using standard ASTM testing methodologies. In addition, for proprietary reasons, the property values (RON) were normalized between -1 and 1. The original samples were divided into training and testing dataset, randomly: 175 samples were utilized for training and the remaining samples were used as testing dataset. Then, the training dataset was segregated into 3 clusters using kmeans++ algorithm. The parameters of all the methods are tabulated in Table 1, and the comparison results of all the methods are listed in Table 2. According to the RMSE and R^2,

PLS gives considerably higher error than the other methods. It illustrates that this method cannot capture the change of the process well.

Table 1: Optimal parameters of each algorithm

Method	k	α	γ
PLS	_	_	_
LWPLS	_	0.01	_
Km-LWPLS	3	0.01	_
JITL	_	_	0.001
KmCV-LWPLS	3	0.01	_

As a result, in an industry application, the PLS model need to be updated frequently, which is very time-consuming. Therefore, it is not suitable for online prediction. The proposed KmCv-LWPLS has the lowest RMSE and the highest R^2. It indicates that this method can able to estimate the future data (testing data) effectively. The Km-LWPLS approach gives better performance than JITL, and the results of JITL algorithm are better than that of LWPLS. These clearly show that a global, static and linear model cannot function well when the processes are characterized with nonlinearity and time-varying, while updating and multi-model strategy such as LWPLS, JITL, Km-LWPLS, KmCv-LWPLS and etc. can improve prediction accuracy. Although the training samples of LWPLS are the same as that of PLS, LWPLS gives better results. This indicates that local weighted strategy enables PLS to account for nonlinearity as well as the time-varying issues. However, the high-level samples in the training dataset influence the performance hugely, and lead to large prediction error. Hence, the performance of this method is worse than JITL and KmCv-LWPLS.

The conventional JITL based method selects local training samples based on the Euclidean distance and establishes local model for each query sample. Based on this, for each new sample, the high level points are not included in the local model. Hence, JITL approach performs better than LWPLS. However, as shown in Figure 1, for some abrupt change, the prediction error is large. One reason for this phenomenon seems to be that the information of the objective variable y and process knowledge are not taken into consideration when select the training samples. As a result, for these query samples, JITL approach cannot obtain an optimal training dataset and leads to poor performance. In addition, the number of similar samples is changed each time, which lead to the instability of the prediction accuracy. For example, modelling information is insufficient with fewer samples, and the high-level points may be included and jeopardize model performance if more samples are chosen.

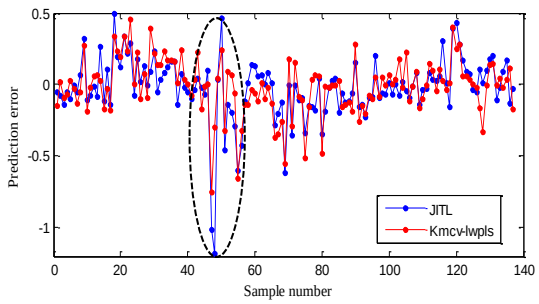

Figure 1: Error results of JITL and KmCv-LWPLS

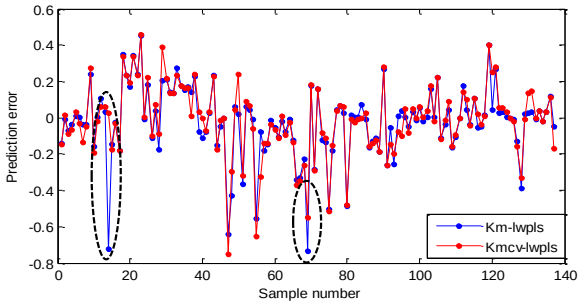

Figure 2: Error curve of Km-LWPLS and KmCv-LWPLS

Compared to JITL, Km-LWPLS and KmCv-LWPLS build local models offline based on the historical data. For online application, the optimal local model is selected based on similarity criterion. This strategy remains the structure of each sub-model stability and eliminates the influence of high level data. Hence, as showed in Table 2, both of Km-LWPLS and KmCv-LWPLS are more effective than JITL, LWPLS and PLS methods. According to the results, the proposed KmCv-LWPLS improves the RMSE and R^2 in comparison with Km-LWPLS approach. In addition, the only difference of Km-LWPLS and KmCv-LWPLS is the strategy of dataset partition. These clearly show the importance to optimize clustering process when a multi-model be established. Besides, as illustrated in Figure 2, the improved strategy can handle abrupt change more effective and gives a smaller error. As analysed previously, the multi-model based strategy is suitable for the nonlinearity, multi-operation

process. Additionally, it is important to optimal the dataset partition process combined with process knowledge and dependent variable information. Typically, the cross-validation strategy proposed in this paper is effective. Besides, for the model updating, new sampling points can improve the performance, however, it is difficult to implement because of the large sampling interval and low sampling frequency. Therefore, local weighted approach is more practical than the new sample based updating strategy. Since the proposed method can take full use of the advantage of LW and cross validation, it can provide the best results.

Table 2: Model performance of each algorithm

Method	RMSE	R^2
PLS	0.3252	0.9255
LWPLS	0.2336	0.9616
JITL	0.2221	0.9652
Km-LWPLS	0.1986	0.9722
KmCV-LWPLS	0.1955	0.9731

4. Conclusions

This paper expounds the importance of modelling algorithms for NIR system and points out that the traditional modelling strategies are insufficient to establish an effective NIR model, especially for the industrial process characteristic with high nonlinearity and large sampling interval. In current work, we propose a cross validation based multi-model modelling strategy to handle this issue. Through applications to a real industry data set, it is demonstrated that the proposed KmCv-LWPLS algorithm generally outperforms the global based, local weighted based and JITL based methods. In order to reduce the computation complexity, a further modification of the proposed method could be taken into account.

Acknowledgments

We would like to acknowledge financial support for this work from Shandong University of Science and Technology and the financial support for this work from Shandong Provincial Natural Science Foundation, China (ZR2017BF026, ZR2017PF002).

References

Bakirov R., Gabrys B., Fay, D., 2017, Multiple adaptive mechanisms for data-driven soft sensors, Computers & Chemical Engineering, 96, 42-54

Balabin R. M., Lomakina, E. I., 2011, Support vector machine regression (SVR/LS-SVM)--an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst, 136, 1703-1712

Ge Z., Song Z., 2010, A comparative study of just-in-time-learning based methods for online soft sensor modeling, Chemometrics and Intelligent Laboratory Systems, 104, 306-317

He K. , Qian F., Cheng H., Du W., 2015, A novel adaptive algorithm with near-infrared spectroscopy and its application in online gasoline blending processes, Chemometrics and Intelligent Laboratory Systems, 140, 117-125

He K., Qian F., Cheng H., Du W., 2016, Improved integrated optimization method of gasoline blend planning and real-time blend recipes, Industrial & Engineering Chemistry Research, 55, 4632-4645

Kadlec P., Grbić R., Gabrys B., 2011, Review of adaptation mechanisms for data-driven soft sensors. Computers & Chemical Engineering, 35, 1-24

Kim S., Kano M., Hasebe S., Takinami A., Seki T., 2013, Long-Term Industrial Applications of Inferential Control Based on Just-In-Time Soft-Sensors: Economical Impact and Challenges, Industrial & Engineering Chemistry Research, 52, 12346-12356

Mei Q.-P., Li T.-F., Yao L.-Z., Huang D., Yang Y.-L., 2016, Study of an adaptable calibration model of near-infrared spectra based on KF-PLS, Chemometrics and Intelligent Laboratory Systems, 157, 152-161

Quiñones L., Velazquez C., Obregon L., 2014, A novel multiple linear multivariate NIR calibration model-based strategy for in-line monitoring of continuous mixing, AIChE Journal, 60, 3123-3132