Causal-assisted Sequence Segmentation and Its Soft Sensing Application for Multiphase Industrial Processes

Yimeng He,a Le Yao,b Xiangyin Kong,a Xinmin Zhang,a,\* Zhihuan Song,a Manabu Kanoc

a The State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

b The School of Mathematics, Hangzhou Normal University, Hangzhou 311121, China

c Department of Systems Science, Kyoto University, Kyoto 606-8501, Japan

xinminzhang@zju.edu.cn

Abstract

The multiphase characteristics of industrial processes pose challenges to industrial big data modeling. Conventional soft sensing models often overlook process dynamics and struggle to handle transient behaviours like phase transitions. To solve these problems, this article proposes a causal-assisted segmentation (CAS) model. The CAS model first uncovers the local dynamic characteristics of causal relationships among variables and splits the sequence according to the abrupt changes of causal mechanisms during phase transitions, where the consistency of causal mechanisms in the time dimension is examined by the proposed causal similarity. Then, the causal relationships in each phase are represented as a temporal causal graph (TCG). Furthermore, a soft sensing model called the temporal-causal graph convolutional network (TC-GCN) is established by transferring the time-extended data and the weighted adjacency matrix of the TCG to the graph convolutional layer. The effectiveness of the proposed CAS and TC-GCN models was verified through a penicillin fermentation process. Experimental results demonstrate that the breakpoints discovered by the CAS are consistent with the reaction mechanisms. Furthermore, the TC-GCN significantly improves the prediction accuracy.

**Keywords**: Multiphase process, causal discovery, soft sensor, time series segmentation, graph convolutional network.

* 1. Introduction

Many industrial processes exhibit multiphase characteristics, like some batch processes. Here, the multiphase characteristics indicate that due to the dynamic nature of processes or the influence of time-varying factors, relationships between variables or other process characteristics change even within a single operating batch (Yao and Gao, 2009). Some commonly used multivariate statistical control methods, such as principal component analysis (PCA) (Albazzaz and Wang, 2006), treat the data of one batch as a whole, ignoring the variations of process characteristics. This affects the monitoring and prediction capabilities of the models (Zhang et al., 2015).

To achieve a better understanding of multiphase processes, plenty of phase segmentation approaches have been proposed to segment the whole process into different phases. Especially, data-driven segmentation methods are widely applied to overcome the lack of process knowledge. Camacho et al. (2006) used slice PCA to extract variable correlation information, and then clustered the fragments. Wang et al. (2019) proposed a method that segments the sequence based on the predictive ability of the prediction model.

However, the models mentioned above all have certain limitations. For example, some models do not guarantee the continuity in the time dimension. Some do not consider the interpretability of the segmentation results. In other words, the methods that divide phases based on the process correlation information may lead to divisions that are not exactly the same as reaction phases. To solve these problems, the proposed casual-assisted segmentation (CAS) model combines causal discovery with predictive models, and segments the sequence by detecting the changes of causal mechanisms. It ensures continuity in the time dimension while enhancing the interpretability of the segmentation results. Additionally, the extracted temporal causal graph can be used to assist the development of the quality prediction model.

* 1. Preliminary of temporal causal discovery

This section introduces an original temporal causal discovery model, NTS-NOTEARS (Sun et al, 2023). It is used for the development of the proposed CAS model. NTS-NOTEARS consists of 1D convolutional neural networks (CNNs), which are designed to discover the instantaneous and lagged variable dependencies of multivariable time series.

Suppose that the time series includes *d* variables, there are *d* CNNs jointly trained, where predicts the expectation of the target variable at each time stamp , given the preceding and instantaneous input variables:

|  |  |
| --- | --- |
|  | (1) |

where  denotes the parents of  that are determined by the CNN weights (see next paragraph). Here *K* is the hyperparameter denoting the maximum time lag, so the input of  consists of all preceding variables up to the time stamp *t-K* (denoted as ) and all variables at the same time stamp *t* other than  (denoted as ).

To estimate the dependency strength of the edge between  and , the kernel weights of  are transformed to the elements of the weighted adjacency matrix *W*. Let  represent the *m* kernel weights corresponding to the input  in , the transform equation is as follows:

|  |  |
| --- | --- |
|  for  | (2) |

* 1. Causal-assisted segmentation model

The proposed causal-assisted segmentation (CAS) model identifies the breakpoints where the phase changes. The CAS model uses the mentioned NTS-NOTEARS model to detect the time stamps where the preceding causal mechanisms no longer fit the current samples. For each phase, the causal mechanisms are extracted from the initial window of time, and the samples are deemed to be from a new phase if they are not predictable by using these causal mechanisms.

Specifically, there is a time series . The procedure of the CAS model for the phase division is as follows:

Step 1. *Initialization*. The samples in a time window of length *h* are used to discover the causal mechanisms of the *p*-th phase, which is denoted as **** where .  is the breakpoint between the (*p*-1)-th phase and the *p*-th phase, as well as the starting point of the *p*-th phase.  is normalized to have the zero mean and unit standard deviation, denoted as . The mean vector of  is denoted as .

Step 2. *Model Training*. The NTS-NOTEARS model is trained using the normalized window data . Through the training process, the temporal causal graph  and training loss of the *p*-th phase can be obtained as follows:

|  |  |
| --- | --- |
|  | (3) |

where  is the predictive vector of , and each element of is the predictive value obtained from the CNN mentioned in Eq. (1).

Step 3. *Phase Extension*. The CAS model extends the phase in the time step of *w*, until it finds the breakpoint where the CNNs of the NTS-NOTEARS are not able to predict the current samples. At the breakpoint, the former causal mechanisms are no longer able to adapt to the current samples and a new phase is supposed to start.

The similarity distance is defined to determine whether the current samples correspond to the former phase or not. The similarity distance includes the causal similarity distance and the stable similarity distance.

The causal similarity distance measures the difference in causal dependencies between the initial window and the current window, which is formulated as the testing loss for the current samples:

|  |  |
| --- | --- |
|  | (4) |

where **** denotes the current samples, and *n* is the number of moving steps.  means the normalized current samples, and  denotes the predictive value of  from the NTS-NOTEARS model.

The stable similarity distance is used to evaluate the distance of stable states between the initial window and the current window, which is represented as the Manhattan distance between the mean vectors of the two data windows:

|  |  |
| --- | --- |
|  | (5) |

where  denotes the mean vector of , and  denotes the mean vector of .

The similarity distance comprehensively considers  and . The coefficient  is set to balance the two items (generally, scaling the two distances to a similar range is recommended). Therefore, the similarity distance is formulated as follows:

|  |  |
| --- | --- |
|  | (6) |

when  is up to the threshold, the time stamp is regarded as the breakpoint between the (*p*-1)-th phase and the *p*-th phase, which is denoted as . The threshold of  is defined as follows:

|  |  |
| --- | --- |
|  | (7) |

where  indicates the limit of , while  indicates the limit of .

Step 4. *Repetition*. Step 1-3 are supposed to be repeated until the number of breakpoints reaches the maximum  or the length of the remaining sequence is less than the minimum .

* 1. Temporal-causal graph convolutional network

Using the proposed CAS model, the time series is divided into different phases, and the temporal causal graph for each phase is obtained. To implement the quality prediction task, a new soft sensing model called temporal-causal graph convolutional network (TC-GCN) is constructed for each phase. The TC-GCN model integrates the temporal and spatial features by transferring the sequence data and the weighted adjacency matrix of the temporal causal graph to the graph convolutional (GC) layer.

In order to match the dimension of the temporal causal graph, the phase data is extended to , each sample  consists of , and *n* is the number of samples in this phase.

The GC layer can be written as:

|  |  |
| --- | --- |
|  | (8) |

where  is the weighted adjacency matrix of the temporal causal graph ,  is the encoder parameter matrix which is obtained through the training.  is the matrix square root of , and  is the degree matrix of the graph, given as follows:

|  |  |
| --- | --- |
|  | (9) |

The input of the following multiple perceptron (MLP) is given as:

|  |  |
| --- | --- |
|  | (10) |

where  is the trainable weight matrix.

* 1. Case study

In this section, the penicillin fed-batch fermentation process was used to validate the efficiency of the proposed CAS and TC-GCN models. The experiment was implemented on PenSim v2.0 platform. The total simulation time is 400 hours. The sample rate is 401 /h. The collected training and testing data both consist of 16000 samples. The two data sets are collected in two simulations in the normal state, with different setting values of the initial substrate concentrate (15.0 g/L and 14.5 g/L respectively). As the process knowledge, 0 ~ 45 h is the stage of biomass accumulation. 45~400 h belongs to the continuous penicillin production stage. The variable description of this process is listed in the Table 1. The dissolved oxygen concentration is taken as the quality variable. In the segmentation process, the initial window length *h* is set to 1000.

Table 1. The variable description of the fermentation process

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var. | Description | Var. | Description | Var. | Description |
| X1 | Aeration rate | X6 | Dissolved oxygen concentration | X11 | PH |
| X2 | Agitator power | X7 | Biomass concentration | X12 | Reactor temperature |
| X3 | Substrate feed rate | X8 | Penicillin concentration | X13 | Generated heat |
| X4 | Substrate feed temperature  | X9 | Culture medium volume | X14 | Base flow rate |
| X5 | Substrate concentrate | X10 | CO2 concentration | X15 | Cold water flow rate |

The maximum time lag *K* is set to 3. The number of the convolutional kernels *m* is set to 160. The coefficient  is 10. The two scale hyperparameters  and are respectively 2 and 1.2. The CAS model segments the continuous 16000 testing samples into four phases: the obtained breakpoints are 1900, 2900, and 6000; the corresponding time stamps are respectively 47.5 h, 72.5 h, and 150 h. Figure 1 shows the segmentation results of three models: Gaussian mixture model (GMM) (Ariba et al., 2023), Greedy Gaussian segmentation (GGS) (Hallac et al., 2019), and the CAS model. As shown in Figure 1, compared with the true first phase-transition time stamp 40 h, the first breakpoints found by the three methods are 47.5 h (CAS), 53.6 h (GMM), and 53.6 h (GGS) respectively. It also can be seen that CAS clearly identifies the transition phase, which is located between the breakpoints 1900 and 2900, as shown by the solid line in Figure 1. However, the breakpoints found by GMM and GGS are more susceptible to the mean value.



Figure 1. The segmentation results

After the segmentation, the TC-GCN models are established respectively using all the training data and the data in each divided phase. The numbers of hidden neurons in the two GC layers are 512 and 256 respectively. The numbers of hidden neurons for the multiple perceptron are 512 and 128 respectively. The batch size is 128. The learning rate is 0.001. The number of epochs is set to 2000. Table 2 displays the root mean square errors (RMSEs) of different models using various data sets. The TC-GCN models use adjacency matrices without weights, while the TC-GCN(W) models use the weighted adjacency matrices.

According to the results in Table 2, it can be seen that the soft sensing models trained using different phase data obtained lower RMSEs compared to training using all data. It indicates that the phase segmentation by CAS and the sample matching based on the similarity distance can improve the predictive accuracy of the soft sensing model. The results also display that the addition of adjacency matrices to the soft sensor model can enhance the predictive accuracy of testing samples at each phase, and the improvement is more significant when using weighted adjacency matrices.

Table 2. The RMSEs of different models using various data sets

|  |  |
| --- | --- |
| Using all the sequence data | Using the data in each divided phase |
| \ | MLP | TC-GCN | TC-GCN(W) |  \ | MLP | TC-GCN | TC-GCN(W) |
| Phase 1 | 0.174 | 0.148 | 0.168 | Phase 1 | 0.148 | 0.100 | **0.070** |
| Phase 2 | 0.219 | 0.266 | 0.306 | Phase 2 | 0.223 | 0.202 | **0.196** |
| Phase 3 | 0.151 | 0.185 | 0.164 | Phase 3 | 0.118 | 0.089 | **0.082** |
| Phase 4 | 0.143 | 0.127 | 0.105 | Phase 4 | 0.034 | 0.027 | **0.027** |
| Average | 0.154 | 0.154 | 0.146 | Average | 0.096 | 0.076 | **0.069** |

* 1. Conclusion

In this work, a new phase segmentation model called causal-assisted segmentation (CAS) is first proposed, and then a temporal-causal graph convolutional network (TC-GCN) is designed for soft sensing modeling. The CAS model first discovers the local dynamic characteristics of causal relationships among variables and detects the breakpoints by identifying the abrupt changes of the causal mechanisms during phase transitions. The similarity distance is designed to measure the difference in the causal mechanisms of two different phases. For each phase, a TC-GCN model is established using the weighted adjacency matrix of the temporal causal graph. The verification experiments were conducted on the penicillin fermentation industrial process. The breakpoints found by the CAS are closer to the time stamps of phase transition. The designed TC-GCN model shows a significant improvement in prediction accuracy.

References

Y. Yao, F. Gao, 2009, A survey on multistage/multiphase statistical modeling methods for batch processes, Annual Reviews in Control, 33, 2, 172-183.

H. Albazzaz, X. Wang, 2006, Multivariate statistical batch process monitoring using dynamic independent component analysis, Computer aided chemical engineering, 21, 1341-1346.

X. Zhang, Y. Li, M. Kano, 2015, Quality prediction in complex batch processes with just-in-time learning model based on non-Gaussian dissimilarity measure, Industrial & Engineering Chemistry Research, 54, 31, 7694-7705.

J. Camacho, J. Picó, 2006, Multi-phase principal component analysis for batch processes modelling, Chemometrics and Intelligent Laboratory Systems, 81, 2, 127-136.

K. Wang, L. Rippon L, J. Chen, Z. Song, R.B. Gopaluni, 2019, Data-driven dynamic modeling and online monitoring for multiphase and multimode batch processes with uneven batch durations, Industrial & Engineering Chemistry Research, 58, 30, 13628-13641.

X. Sun, O. Schulte, G. Liu, P. Paupart, 2023, NTS-NOTEARS: Learning Nonparametric DBNs With Prior Knowledge, PMLR, 1942-1964.

H. Ariba, P. Vanabelle, S. Benaly, H. Thomas, C.R. André, G. Leonard, 2021, Application of data science to study fluorine losses in the phosphate industry, Computer Aided Chemical Engineering, 50, 1059-1065.

D. Hallac, P. Nystrup, S. Boyd, 2019, Greedy Gaussian segmentation of multivariate time series, Advances in Data Analysis and Classification, 13, 3, 727-751.