Self-attention Mechanism based Dynamic Fault Diagnosis and Classification for Chemical Processes

Shuai Chen¹, Lin Luo¹, Qilei Xia¹, and Lunjie Wang¹

¹ School of Information and Control Engineering, Liaoning Petrochemical University, Fushun 113001, Liaoning, P. R. China

Email: luolin@lnpu.edu.cn

Abstract. A dynamic fault detection and diagnosis technique based on deep encoder-decoder network with self-attention mechanism is proposed in this paper. Although traditional encoder-decoder networks exhibit capability in extracting the temporal dependencies, the architecture encodes the input sequence into a fixed-length internal representation. This limits the performance of these networks, especially when considering relatively long input sequences. The self-attention mechanism is used to weight the local feature vectors and retain the correlation between the local information of the signal and the process operation state, so as to extract the effective feature vectors. The extracted features are then fed into the bidirectional encoder-decoder network. The resulting deep network is not only generalizing the importance of local temporal feature, but it also allows the interpretable feature representation and classification simultaneously. The experiments on the benchmark Tennessee Eastman process show that the proposed model has better diagnostic performance on receiver operating characteristic (ROC) and precise-recall (PR) curves than the classical diagnostic method based on the long-short term memory (LSTM) network with convolution layers.

Keywords: dynamic fault detection, deep learning, attention mechanism.

1. Introduction

The rapid development of information science, automation technology, and modern instrumentation has placed greater demands on the design, analysis, manufacturing, and management of complex industrial processes [1-3]. Maintaining the safe and reliable operation of complex industrial processes is a common goal for petroleum and chemical industries. Regular maintenance and management of equipment is therefore essential not only to ensure the safety of people and the production environment, but also to communicate important information about the quality of the product at the right time. In the actual production process, it is difficult to establish accurate process and fault models due to the large scale of the process and many uncertainties, which limits the scope of application of quantitative and qualitative methods [4-6].

The data-driven approach uses the normal and the fault history data stored in the distribution control system to build the corresponding models for process monitoring. The approach relies only on measurement data and is particularly suitable for complex process industries. With the development of machine learning, deep learning has been introduced into the process monitoring. It has been demonstrated that these models have better ability to model complex relationships compared to shallow models [7-9]. For example, Luo et al. [2] proposed an adaptive monitoring strategy with a tensor factorization deep network. They extracted fault-sensitive characteristics with the tensor representations, which enable efficient cross-layer knowledge. Wu et al. [8] investigated a fault diagnosis model based on a deep convolutional neural network for chemical process. Both the spatial
and temporal features are dynamically extracted by the convolutional layers. Zhao et al. [9] exploited the long short-term memory (LSTM) to build a dynamic monitoring strategy for the Tennessee Eastman (TE) process.

In real industrial processes, dynamic characteristics are an inherent and unavoidable problem in process monitoring due to the presence of feedback systems and random voices. For the long-term real-time applications, the information on the process dynamics is important and should be preserved during the model construction. Although the dynamics are suitably captured by the existing LSTMs using recurrent feedback, the local information are hardly incorporated into the later sub-models. To improve the generalization capability, the local temporal dependencies should be preserved across different time steps. Therefore, designing an additional layer structure with the dynamic properties of the process is the key to more reasonable and effective control of safe operation of dynamic systems.

Motivated by these, this paper proposes a deep encoder-decoder network with the self-attention mechanism for the application of online fault classification. In the proposed deep network, the dynamical system with the multi-level feature extracted by the convolutional layers is instantiated as a bidirectional recurrent neural network (RNN). Additional layer is designed by using a self-attention mechanism which is utilized to preserve the local process dynamics. The resulting network is an end-to-end trainable model by stacking the multiple layers. Experimental results on the Tennessee Eastman (TE) benchmark process show that the proposed network achieves the promising performance compared with the state-of-the-art model.

2. Deep encoder-decoder network

In recent years, RNNs have been widely used in various fields due to the capacity of bidirectional propagation, feedback and association between layers. It exists a recurrent structure compared to the feedforward neural network. The input of each neuron is not only the output of the previous one but also the output information of the neighbouring neurons in the same layer. Although RNNs are capable of capturing temporal correlations, the error gradients propagating by the layers have the problem with vanishing gradients. To address this problem, a bidirectional LSTM network is utilized as an encoder-decoder representation, which is as shown in figure 1.

Assume that a multivariate time series with $N$ samples and $D$ dimensions can be defined as $X_k \in \mathbb{R}^{D \times T}$, $k = 1, \cdots, N$, where contains a sequence of $\Delta t$ sampling points. A typical LSTM networks is to generate an associated sequence of outputs $y_{A_T}$ by 3 gates and a memory cell. Specifically,

$$
\begin{align*}
g_t^u &= \sigma(W^u h_{t-1} + U^u x_t) \\
g_t^f &= \sigma(W^f h_{t-1} + U^f x_t) \\
g_t^o &= \sigma(W^o h_{t-1} + U^o x_t) \\
g_t^c &= \tan(W^c h_{t-1} + U^c x_t) \\
m_t &= g_t^f \odot m_{t-1} + g_t^o \odot g_t^c \\
m_t &= \tan(g_t^o \odot m_t)
\end{align*}
$$

(1)

where $h_t$ is a sequence of hidden states, $W^u$, $W^f$, $W^o$ and $W^c$ are the weight matrices of input gate, forget gate, output gate and memory cell respectively. $\sigma(\cdot)$ is the rectified linear unit (ReLU) function. $\odot$ is the elementwise multiplication. The LSTM network architecture is illustrated in figure 1(a). Additionally, the spatial dependencies provide insight into different types of faults, i.e., some abnormal events influence only one subsystem, whereas other ones influence the network. In order to extract the spatial correlations from a long sequence, a temporal convolutional layer is used in the proposed models. The resulting structure of the convolutional bidirectional encoder-decoder representation network is shown in figure 1(b) which consists of an encoding network and a forecasting network.
3. Self-attention mechanism for sequential fault diagnosis

The dynamic features extracted using the encoder-decoder model are sufficient to characterize the process dynamics for the stage segmentation. However, the limitation is that the encoder has to compress the whole sequence of information into a fixed length vector. It is difficult to represent the information of the whole sequence for the vector, meanwhile the information carried by the first input can be diluted by the later input. In order to improve the prediction accuracy, we introduce the self-attention mechanism to the encoder-decoder model. The structure of the model is shown in figure 2.

Specifically, a vector generated from the sequence of the hidden states $c_t$ is obtained by a weighted sum of these states $h_k$, $k = 1, \ldots, T$, at position $k$,

$$c_t = \sum_k a_{t,k} h_k$$  \hspace{1cm} (2)

where $a_{t,k}$ is the weight of each hidden state, which can be given as,

$$a_{t,k} = \frac{\exp(e_{t,k})}{\sum_{k=1}^{T} e_{t,k}}$$  \hspace{1cm} (3)

where the alignment model $e_{t,k}$ is learned by the following equation,

$$e_{t,k} = \sigma(W_a h_{t,k} x_k)$$  \hspace{1cm} (4)

$$h_{t,k} = \tanh(x_t^T W_t + x_k^T W_t)$$  \hspace{1cm} (5)

where $W_a$ and $W_t$ are learnable weights.

4. Case study on benchmark process

In this section, the performance evaluation of the proposed method is carried on the TE process, which is a benchmark process for the process modelling and monitoring. Experiments are conducted on an
4.1. Process description
The TE process consists of five main process units, which are a stripper reboiler, a recovery compressor, a flash separator, an exothermic two-phase reactor and a condenser. For the reactants, the vapours generated by the separator are recovered to the reactor by a recovery compressor. Meanwhile, a portion of the recovery stream needs to be discharged to prevent the aggregation of the by-products and inert components. The condensed components from the separator are pumped to the stripper column for separation, and the remaining reactants from the stripper are fed into the recovery stream and eventually returned to the reactor.

It should note that the control strategies applied to the benchmark differ in their ability to overcome faults accordingly, thus leading to differences in the results [10]. Here we choose the control strategy from the literature [11]. There are a total of 52 measurement variables, including 22 process measurement variables, 18 component measurement variables, and 12 operational variables. The training dataset contains 500 sample points under normal operating condition. The test dataset contains 960 samples, and each dataset corresponds to a fault mode. The simulation platform can set 21 types of failure modes, including step, random variation, slow drift, valve sticking, valve jamming and unknown failure.

4.2. Parameter configuration and evaluation metrics
For the layer structure, 1-D convolutional layer was designed to build the diagnostic model since the raw data are in 1-D format. The encoder has 2 convolutional layers and 2 pooling layers, where the size of convolutional filters and kernels are set to 64 and 1, respectively. The size of max-pooling in the pooling layers is 1. The activation function in the convolutional layers uses the ReLU function. The output dimension of the signal is reduced by convolution and pooling operations. The samples are then processed by flatten operation as the feature extraction of the whole network.

The performance on the multi-class classification should be considered with a total of 21 fault modes in the TE process. For multi-class tasks, the quality of the overall classification is typically averaged over all classes in many possible ways. For the binary classification case, the ROC curve can be used as a continuous variable indicator to describe the sensitivity and specificity of the diagnostic model. The precise-recall curve can be used to further determine the classification performance of the model if there is a disparity in the ratio of normal to faulty samples. For the multi-classification problem, two metrics are utilized to assess the overall performance, including micro-averaging and macro-averaging metrics.

4.3. Fault classification and performance analysis
In order to evaluate the performance of LSTM network with convolution layer (LSTM-CNN) and the multiplicative attention network (MAN), a series of experiments are then conducted in the fault detection and classification of the multivariate TE sequential data. There are 14 representative process status including in the experiments, such as the normal case, IDVs (1)-(2), (4)-(8), (11), (13)-(14), (17)-(19). The ROC curve of each method for the fault modes are demonstrated in figure 3. It can see that both of the deep learning methods provide similar performances for the most of IDVs. It is clearly observed from the micro- and macro-averaging area under curve (AUC) of ROC curve that MAN provides a better quality of the overall performance. The micro- and macro-average of AUC using the LSTM-CNN model are both 0.99, while the values corresponding to the MAN model are 1. To further analyse the recognition capability of the model, the evaluation using the PR curve is illustrated in figure 4. As shown in figure 4(b), the PR curve with the MAN model is closer to the upper right corner of the coordinate axis compared to the PR curve shown in figure 4(a). Moreover, the micro- and macro-average values of the AUC in the PR curves using the LSTM-CNN model are respectively 0.96 and 0.94, while the values corresponding to the MAN model are 0.98 and 0.97. The overall performances using the MAN model for all fault modes are better than those corresponding to the LSTM-CNN. Moreover, the F1 scores are investigated to balance the precision and the recall of
the fault classification, which are shown in Figure 5. It can be seen that the F1 score of MAN is around 0.9 with the different faults, which means that it is suitable for the multi-fault classification tasks.

![Figure 3. ROC curves of (a) LSTM-CNN and (b) MAN over the different fault modes. The micro- and macro-averaging ROC curves are denoted by dashed pink and navy-blue lines, respectively.](image)

![Figure 4. P-R curves of (a) LSTM-CNN and (b) MAN over the different fault modes. The micro- and macro-averaging P-R curves are denoted by dashed pink and navy-blue lines, respectively. The horizontal line (red with dashes) represents the random performance level.](image)

![Figure 5. F1 scores of LSTM-CNN and MAN over the different fault modes.](image)

5. Conclusion
In the field of chemical fault diagnosis, the methods based on the shallow structure still have some drawbacks on the temporal feature extraction although they have improved in the accuracy of
diagnosis compared with traditional methods. In order to overcome the drawback, this paper constructs a chemical process fault diagnosis model based on attention mechanism. The experiments on the TE process show that the fault diagnosis method based on attention network has better feature extraction and higher diagnosis accuracy compared with LSTM-CNN. In the future work, the layers can be designed to extract the spatial features and to allow its application for online scenario.

6. References

[1] Jiang Q, Yan X. Learning deep correlated representations for nonlinear process monitoring, IEEE Transactions on Industrial Informatics, 2019, 15(12): 6200-6209.

[2] Luo L, Xie L, Su H Y. Deep learning with tensor factorization layers for sequential fault diagnosis and industrial process monitoring, IEEE Access, 2020, 8: 105494-105506.

[3] Luo L, Xie L, Su H Y. Robust mixture Bayesian latent variable regression with structural sparsity and application to inferential sensing of quality variables. Industrial & Engineering Chemistry Research, 2020, 59(50): 21822 – 21840.

[4] Qin S J. Survey on data-driven industrial process monitoring and diagnosis, Annual Review Control, 2012, 36(2): 220-234.

[5] Luo L, Su H Y, Ban L. Independent component analysis - based sparse autoencoder in the application of fault diagnosis, In: Proceeding of the 11th World Congress on Intelligent Control and Automation, China: Shenyang, 2014, 1378-1382.

[6] Heo S, Lee J, H. Fault detection and classification using artificial neural networks. IFAC-PapersOnLine, 2018, 51(18): 470 - 475.

[7] Zhang Z P, Zhao J S. A deep belief network based fault diagnosis model for complex chemical processes, Computers & Chemical Engineering, 2017, 107(5): 395-407.

[8] Wu H, Zhao J. Deep convolutional neural network model based chemical process fault diagnosis. Computers & Chemical Engineering, 2018, 115(12): 185-197.

[9] Zhao H, Sun S, Jin B. Sequential Fault Diagnosis Based on LSTM Neural Network. IEEE Access, 2018, 6: 12929-12939.

[10] Yin S, Ding S, Haghani A, Hao H, Zhang P. A comparison study of basic data- driven fault diagnosis and process monitoring methods on the benchmark tennessee-eastman process. Journal of Process Control, 2012, 22(9): 1567-1581.

[11] Larsson T, Hestetun K, Hovland E, Skogestad S. Self-optimizing control of a large- scale plant: the tennessee eastman process. Industrial & Engineering Chemistry Research, 2001, 40(22): 4889-4901.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grants No. 61703191), the Foundation of Liaoning Educational Committee (No. L2017LQN028) and the Scientific Research Foundation of Liaoning Petrochemical University (No. 2017XJJ-012).