Recurrent Pattern Aware Multiphase Flow Soft Sensor Model

Luyao Wang*
College of Control Science and Engineering, Zhejiang University, Hangzhou, China
*Corresponding author: Luyaowang@zju.edu.cn

Abstract. Given data captured from multiphase flow industrial process which has non-stationary and dynamic characteristics, it’s difficult to predict the dominant variables precisely via traditional soft sensor models. In this paper, we propose RP-LSTM (Recurrent Pattern Aware Long Short Term Memory) model, a soft sensing method that takes both dynamic of time-sequence and pattern mining of industrial process into account. In this way, we can fully utilize limit model capacity, and enhance model’s knowledge on recurrent data stream patterns. Experiments on real data set demonstrate that RP-LSTM out-performs traditional sequence deep learning models by a significant margin.

1. Introduction
With the continuous development of automation technologies, modern industry has step into the stage of fully automated production. In the real industrial production scene, there are some important process variables that cannot be directly detected, or the price of instrumentation is expensive and the maintenance is difficult. The emergence of soft sensing technology provides an effective way to solve the above problems. The core idea of soft sensor is to use variables that can be easily measured (auxiliary variables) to establish a model to predict variables that are difficult to measure (dominant variables).

The industrial production process is continuous in time, that is, the data with similar sampling time has strong time series correlation. Therefore, many studies have proposed dynamic soft sensor models according to the characteristics of time series data. In order to facilitate the study of the relationship between the past and the present moment, the soft sensor model based on Recurrent Neural Network (RNN) is introduced. However, due to the problem of vanishing gradient, RNN is unable to effectively consider the influence of long-term timing relationship. Later on, the emergence of Long Short Term Memory (LSTM) network has effectively overcome the vanishing gradient problem of RNN. Through the introduction of cell state, LSTM can extract both long-term and short-term time series correlation characteristics at the same time. Yuan et al. [1] proposed a supervised LSTM (SLSTM) for soft sensor, which can extract nonlinear dynamic features related to dominant variables more pertinently, and is suitable for continuous processes with sequential and nonlinear characteristics. Based on the LSTM model, the attention mechanism can also be introduced to make the model focus on the samples that are more helpful for prediction and improve the soft sensor performance. [2]

The above dynamic soft sensor models pay more attention to how to effectively extract the dynamic features that can reflect the temporal relationship of the process, but all ignore the deeply hidden process patterns. However, in the actual industrial production process, the working conditions of the equipment in two different time periods may share some similarity, so do the evolving trends of various process variables. Different time periods have the possibility of being in a same regime which means a pattern
with unique process characteristics compared with other periods. For example, the whole process may go through regime 0, 1, 2, 1, 0, 2 sequentially. If the recurrent patterns in the industrial process can be obtained, we can pertinently reinforce the training of the time periods with the same patterns. When predicting the dominant variables in a certain period, the soft sensor model reinforced with the same pattern is adopted. Compared with the traditional method, this method of targeted intensive training can deeply mine the pattern information of the process, so that the effect and interpretability of soft sensor can be improved. In the data mining area, there has been an ever-increasing interest in mining time series. Li et al. [3] developed DynaMMo, a scalable algorithm for co-evolving time sequences with missing values. Based on a linear dynamical system, DynaMMo is able to segment a data sequence. Yasuko Matsubara et al. [4] proposed Autoplait, a fully automatic mining algorithm for co-evolving time sequences. As an effective pattern mining method, Autoplait is parameter-free and requires no user intervention, no prior training, and no parameter tuning.

In view of the above methods, this paper contributes in the following aspects:

The information of recurrent patterns is introduced into the soft sensor of three-phase flow process for the first time to fully utilize the process characteristics.

The combination of Recurrent Pattern and LSTM takes both process features and time correlation into account, which improves the performance of soft sensor.

2. Related Works

The traditional soft sensor modeling methods are mainly based on mechanism analysis, regression analysis and artificial intelligence. The soft sensor based on mechanism analysis looks for the mathematical relationships between variables by studying the mechanical motion law of the fluid and the mechanism analysis of various physical and chemical factors in the industrial process. For complex industrial processes, the inherent mechanism is usually not understood, which make it difficult to establish a soft sensor model only based on mechanism analysis. Regression analysis uses the least square method to find the empirical formula of the correlation between the process variables. Typical methods include Principal Component Regression (PCR), Partial Least Square Regression (PLSR), Independent Component Regression (ICR), Gaussian Process Regression (GPR) and so on. Artificial Neural Network (ANN) is a dynamic model of distributed parallel information processing algorithm structure that imitates the behavior characteristics of human brain neural network. The soft sensor based on ANN can be applied to systems with high nonlinearity and uncertainty, and has the ability of self-learning, associative storage, and high-speed search for optimal solutions. The common ANN forms are Back Propagation (BP), Radial Basis Function (RBF), Counter Propagation Networks (CPN) and so on.

According to the dynamic characteristics of the industrial process, recently there is an explosion of dynamic soft sensor models generated for taking the time series correlation into account. Dynamic soft sensing methods can be divided into four categories: autoregression, dynamic matrix extension, state space and deep learning.

Kruger et al. [5] combined autoregressive model with Principal Component Analysis (PCA) to introduce time series information into PCA model and achieved good results in dynamic process monitoring. With the increasing complexity of the industrial process and more and more nonlinear links between variables, the soft sensor model based on autoregression cannot well describe the real dynamic relationship between samples. Kano et al. [6] established a soft sensor model on the dynamic augmented matrix by using the Partial Least Square (PLS) model and obtained good experimental results on the actual distillation column data set. The method of augmented matrix does describe the dynamic characteristics to some extent, but the augmented matrix often has a large dimension, which will affect the computational efficiency of the model and introduce noise.

Shang et al. [7] constructed the probability slow feature between process samples as dynamic feature description through representation learning, which improved the prediction accuracy of the soft sensor model. Chen et al. [8] designed a switched linear dynamic system model by using the method of Gaussian filtering, which effectively improved the detection rate of new faults in the model. Based on the state space soft sensor model, the randomness of process dynamics is well described by probability,
but in most cases, the dynamic process is described linearly, and the reusability of the model is not strong.

To simultaneously consider the dynamic and nonlinear features of industrial process, Ke et al. [9] proposed the Long Short Term Memory (LSTM) network to develop soft sensor models. Yuan et al. [10] proposed the supervised LSTM (SLSTM) models by integrating the label information to achieve superior performance on the data set of penicillin fermentation process. The introduction of attention mechanism helps LSTM focus on the samples that are beneficial for prediction and improve the soft sensor performance. [11]

3. Optimization Model
In this section, we describe the two basic building blocks: LSTM and Autoplait, followed by our proposed model: RP-LSTM.

3.1. LSTM

![Fig 1. Structure of the LSTM unit.][12]

The LSTM network is a variant of the standard RNN. The detailed structure of the basic LSTM unit is shown in Figure 1. The LSTM unit contains three gate controllers, namely the input, forget, and output gates. The three gates are mainly used to determine what information should be remembered. The LSTM network realizes temporal memory through the gate switching to prevent the gradient vanishing.

3.2. Autoplait

The inner-most loop of Autoplait algorithm is named as “CutPointSerch”, a loop to find cut points that are used to generate segments of two regimes. After that, the inner loop “RegimeSplit” works to estimate optimal parameters of the segmentation of two regime. Finally, the out loop “Autoplait” searched for the best number of regimes. The concrete steps of Autoplait is illustrated in Algorithm 1, where $Cost_t(X; C)$ refers to the total cost of $X$ given set of parameters $C$. 

---

[12]: https://www.example.com
Table 1. The concrete steps of Autoplait

| Algorithm 1 Autoplait(X) [13] |
|--------------------------------|
| 1:                            |
| Input: Bundle of time series X |
| 2:                            |
| Output: Complete set of parameters C |
| 3:                            |
| (a) Number of segments, m     |
| 4:                            |
| (b) Number of regimes, r      |
| 5:                            |
| (c) Segment set, S = {s1,...,sm} |
| 6:                            |
| (d) Segment membership, F = {f1,...,fm} |
| 7:                            |
| (e) Model parameters, Θ = {θ1, . . . , θr; ∆} |
| 8:                            |
| Q = ∅; S = ∅; m = 0; r = 0; S0 = {1, n}; m0 = 1; |
| 9:                            |
| /* Q: stack for number of segments, segment set, regime */ |
| 10:                           |
| θ0 = BaumWelch (X[|S0|]); /* Estimate model θ0 of S0 */ |
| 11:                           |
| Push an entry {m0, S0, θ0} into Q; |
| 12:                           |
| while stack Q ≠ ∅ do          |
| 13:                           |
| Pop an entry {m0, S0, θ0} from Q; |
| 14:                           |
| /* Try to refine a regime */   |
| 15:                           |
| {m1,m2,S1,S2,θ1,θ2;∆} = RegimeSplit (X[S0]); |
| 16:                           |
| /* Compare single regime θ0 v.s. regime pair θ1 and θ2 */ |
| 17:                           |
| if CostT(X; S0, θ0) > CostT(X; S1,S2, θ1,θ2) then |
| 18:                           |
| /* Regime pair win, split regime */ |
| 19:                           |
| Push entries {m1, S1, θ1}, {m2, S2, θ2} into Q; |
| 20:                           |
| else                          |
| /* Single regime win, no more split. leave it out of the stack */ |
| 21:                           |
| S = S\{S0}; Θ = Θ\{θ0}; r = r + 1; |
| 22:                           |
| Update ∆rr;                   |
| 23:                           |
| fi = r (i = m + 1,...,m0); m= m + m0; |
| 24:                           |
| end if                        |
| 25:                           |
| end while                     |
| 26:                           |
| return C = {m, r, S, F, Θ}    |

3.3. RP-LSTM

Fig 2. A running example of RP-LSTM. Samples in training set and test set, and the predicted values are represented by squares. Circles illustrates different regimes. LSTM0, LSTM1, LSTM2 are models strengthened by training data that belong to regime 0, 1, 2, respectively. The dotted lines of different colors represent the targeted prediction of different LSTM models.

Firstly, recurrent patterns are mined by pattern mining methods. Considering that the true values of dominant variables in test set are unknown in the real industrial scene of soft sensing, the pattern mining methods are supposed to work on the basis of the values of auxiliary variables at each sampling time. Autoplait is applied to divide the entire industrial process into several regimes. Different regimes represent the distinguishing characteristics of the variation tendencies of the process variables at the corresponding time periods. Secondly, separate LSTM models are generated based on the reinforced
training of data at the corresponding time periods. Separate LSTM models are applied to make the targeted prediction on the test set. Finally, the prediction results of each model are combined in chronological order to get a complete prediction of the dominant variables. The LSTM units are responsible for dealing with the temporal correlation between samples, while Recurrent Pattern is responsible for dealing with the distinctions in process characteristics between time periods.

Figure 2 shows a running example of RP-LSTM, the training set sequentially goes through regime 0, 1, 2 that are marked by blue, yellow, pink colors, while the test set goes through regime 1, 0, 2 in chronological order. Here, three patterns repeat twice throughout the whole data set, once in the training set and once in the test set. This is a simple and ideal situation designed to facilitate the presentation. In the real data set, there are likely to be regimes that appear only in the training set or the test set. In the former case, the models enhanced according to regimes are redundant for the test set, but insufficient for the latter. Model redundancy can be simply solved by not using these models. In the case of insufficient models, degradation can be carried out, that is, the corresponding time periods in the test set can be predicted by a basic LSTM model that has not been enhanced by any regimes. In the actual experiments, if only a few samples in the test set fail to find the corresponding models, for the sake of convenience, these small samples can be merged into the regimes of the adjacent time periods.

4. Experiments

4.1. Dataset and Selection of Dominant Variables

| Number | Location | Physical Parameter | Unit |
|--------|----------|--------------------|------|
| 1      | PT312    | Air delivery pressure | MPa  |
| 2      | PT401    | Pressure in the bottom of the riser | MPa  |
| 3      | PT408    | Pressure in top of the riser | MPa  |
| 4      | PT403    | **Pressure in top separator** | MPa  |
| 5      | PT501    | Pressure in three-phase separator | MPa  |
| 6      | PT408    | Diff. pressure (PT401-PT408) | MPa  |
| 7      | PT403    | Differential pressure over VC404 | MPa  |
| 8      | FT305    | Flow rate input air | Sm³/s |
| 9      | FT104    | Flow rate input water | kg/s  |
| 10     | FT407    | Flow rate top riser | kg/s  |
| 11     | LI405    | **Level top separator** | m     |
| 12     | FT406    | Flow rate top separator output | kg/s  |
| 13     | FT407    | Density top riser | kg/m³ |
| 14     | FT406    | Density top separator output | kg/m³ |
| 15     | FT104    | Density water input | kg/m³ |
| 16     | FT407    | Temperature top riser | ºC    |
| 17     | FT406    | **Temperature top separator output** | ºC    |
| 18     | FT104    | Temperature water input | ºC    |
| 19     | LI504    | Level gas-liquid 3 phase separator | %    |
| 20     | VC501    | Position of valve VC501 | %    |
| 21     | VC302    | Position of valve VC302 | %    |
| 22     | VC101    | Position of valve VC101 | %    |
| 23     | PO1      | Water pump current | A     |
| 24     | PT417    | Pressure in mixture zone 2" line | MPa  |
The data is captured from the Three-phase Flow Facility at Cranfield University, which is managed by a Fieldbus based supervisory, control and data acquisition (SCADA) system provided by Emerson Process Management. The variables include 24 process variables from different locations on the three-phase flow system (see Table 1) and two process inputs (air and water flow rate set point). All the data was sampled at a sampling rate of 1 Hz. Three data sets (T_1, T_2 and T_3) were acquired from the system with different set points of air and water flows under normal operating conditions.

The T_2 dataset is selected for our soft sensor experiments. In the three-phase flow process, 9825 samples were simulated in T_2 data set. Among them, the first 6878 samples are adopted as the training set, and other 2947 samples are taken as the test set. In order to focus on the working conditions of the top separator, variable 4, 11, 17 are selected as the dominant variables which record the pressure, level and temperature in the top separator respectively. Other 21 process variables are used as the auxiliary variables to estimate the dominant variables.

4.2. Performance Comparison

To compare the performance, RNN, LSTM, and RP-LSTM models are developed to predict the three dominant variables in the three-phase flow process. The Mean Squared Error (MSE) between the true value and the predicted value is used to measure the effectiveness of different models. The static soft sensor models, such as Random Forest, are also tried in the experiment, but its comprehensive performance is worse than that of RNN and LSTM.

Table II shows the prediction MSE of variable 4, 11, 17 on the test set through RNN, LSTM, and RP-LSTM models. Figure 3 further shows the detailed prediction errors of variable 4 on each sample from the test set, respectively. It is obvious that the proposed RP-LSTM indeed reduces the prediction error and improve the effectiveness of soft sensor. As shown in Figure 4, Autoplait divides the whole data set into four different regimes (0, 1, 2, 3), and the test set goes though the regime 0, 3, 2 in turn. Different time periods in test set are predicted by different LSTM models intensively trained with the corresponding period in the training set.

| Model     | Variables’ Number | 4     | 11    | 17    |
|-----------|-------------------|-------|-------|-------|
| RNN       |                   | 4.816 × 10^{-7} | 5.558 × 10^{-4} | 1.928 × 10^{-2} |
| LSTM      |                   | 4.194 × 10^{-7} | 5.121 × 10^{-4} | 1.033 × 10^{-2} |
| RP-LSTM   |                   | 5.502 × 10^{-8} | 2.563 × 10^{-4} | 4.759 × 10^{-3} |

Fig 3. Comparison of the prediction error of variable 4 by RNN, LSTM, RP-LSTM.
Fig 4. Prediction of variable 17 by RP-LSTM on test set. Regimes mined by Autoplait are illustrated in red, yellow and green respectively.

Fig 5. Prediction of three dominant variables by LSTM and RP-LSTM on test set. V4, V11 and V17 refers to the variable 4, 11, 17 respectively. Figure 5 visually show the prediction of three dominant variables by LSTM and RP-LSTM models. After targeted reinforcement based on the recurrent patterns, RP-LSTM shows more superior prediction ability in the peaks and troughs where the variation trends of variables changed dramatically.

The better prediction performance of RP-LSTM on those peaks and troughs is well explainable. Considering the non-stationarity features of the industrial process of three-phase flow, the change trends of variables are inclined to mutate, which makes it difficult for simple LSTM model to predict the entire test set precisely. As a strong pattern mining method, Autoplait helps soft sensor models to be aware of the mutation of the process characteristics and to make targeted predictions. Hence, RP-LSTM is more
suitable to describe the non-stationary behaviors of the dominant variables and improve the effectiveness of soft sensor model largely.

5. Conclusions
The proposed RP-LSTM takes both the time correlation between samples and the distinctions of process characteristics between time periods into account. With the aid of reinforced knowledge on recurrent patterns, the performance of soft sensor models indeed improves. The effectiveness of the proposed model is verified by the real data of three-phase flow equipment. To further improve the prediction performance, the future work considers combining recurrent patterns with more advanced soft sensor models, such as deep sequence neural networks with attention mechanism.

References
[1] Yuan X F, Li L, Wang Y L. Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network[J]. IEEE Transactions on Industrial Informatics, 2020, 16(5): 3168-3176.
[2] Yuan X F, Li L, Shardt Y, et al. Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development[J]. IEEE Transactions on Industrial Electronics, 2020: 1-1.
[3] Li J, McCann N, Pollard C. Faloutsos. Dynammo: Mining and summarization of coevolving sequences with missing values. In KDD, 2009.
[4] Matsubara Y, Sakurai Y, Faloutsos C. AutoPlait: Automatic mining of co-evolving time sequences[J]. ACM, 2014.
[5] Kruger U, Zhou Y, Irwin G W. Improved principal component monitoring of large-scale processes[J]. Journal of Process Control, 2004, 14(8): 879-888.
[6] Kano M, Miyazaki K, Hasebe S, et al. Inferential control system of distillation compositions using dynamic partial least squares regression[J]. Journal of process control, 2000, 10(2-3): 157-166.
[7] Shang C, Huang B, Yang F, et al. Probabilistic slow feature analysis-based representation learning from massive process data for soft sensor modeling[J]. AIChE Jornal, 2015, 61(12): 4126-4139.
[8] Chen X, Ge Z. Switching LDS-based approach for process fault detection and classification[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 146: 169-178.
[9] Tsinghua W, K, Huang D, Fan Y, et al. Soft sensor development and applications based on LSTM in deep neural networks[C]// 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2017.
[10] Yuan X F, Li L, Wang Y L. Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network[J]. IEEE Transactions on Industrial Informatics, 2020, 16(5): 3168-3176.
[11] Yuan X F, Li L, Shardt Y, et al. Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development[J]. IEEE Transactions on Industrial Electronics, 2020: 1-1.
[12] Yuan X F, Li L, Wang Y L. Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network[J]. IEEE Transactions on Industrial Informatics, 2020, 16(5): 3168-3176.
[13] Matsubara Y, Sakurai Y, Faloutsos C. AutoPlait: Automatic mining of co-evolving time sequences[J]. ACM, 2014.
[14] Ruiz-Carcel C, Cao Y, D Mba, et al. Statistical process monitoring of a multiphase flow facility[J]. Control Engineering Practice, 2015, 42(sep.):74-88.