Semi-supervised Soft Sensor Modeling Method for Crude Furfuryl Alcohol Distillation Process Based on Cosine Similarity-Discount

L H Li¹,* and Y S Dai¹

¹College of Information and Control Engineering, China University of Petroleum (East China), Shandong Qingdao, 266580, China

* Lilonghao_upc@qq.com

Abstract. In view of the dynamic time variation and multirate data problems in crude furfuryl alcohol distillation process, too few samples of modeling data caused poor generalization ability of soft sensor model and reduced the accuracy of soft sensor model prediction. To solve this problem, this paper proposes a semi-supervised soft sensor modeling method based on cosine similarity-discount (SSSMI-COSD). By calculating the cosine similarity between labeled and unlabeled samples in the same time interval, and combining with the proposed constraints, the data clustering between labeled and unlabeled samples is realized. In addition, in order to avoid the ill-conditioned problem caused by high-dimensional input variables, the clustered data are fused by discount factor value (DFV). An actual data of a crude furfuryl alcohol distillation process (CFADP) simulation experiment were carried out. The results show that the proposed SSSMI-COSD method can effectively improve the soft sensor model prediction accuracy for a crude furfuryl alcohol distillation process.

1. Introduction

The distillation process is a process of separating a complex mixture into a single chemical product. The distillation process is a process of separating a complex mixture into a single chemical product with relatively high purity through heat and mass transfer. The principle is to make use of the different relative volatility of each component of the mixture, through evaporation and condensation at a certain temperature and pressure, so that the light components in the liquid phase and the recombinant components in the gas phase can be transferred to each other, thus realizing the separation of components [1]. For distillation column, its characteristics are affected not only by its own internal factors, but also by external factors, resulting in dynamic time-varying distillation process [2]. Because the labeled sample data need to be obtained by utilizing people's professional knowledge and experience, it leads to the difficulty of obtaining valuable labeled sample information. Thus, the classifier with supervised classification based on only a small number of labeled samples does not have good generalization ability [3]. Unlabeled samples are usually collected by various sensor transducers on DCS system, which is characterized by real-time acquisition. Comparing with manual analysis, it is easier to get and larger in quantity [4], and include more sample information [5]. Semi-supervised ensemble learning can further improve the performance of soft sensor model on the basis of ensemble learning by utilizing a large number of unlabeled samples [6].
Aiming at the problem of semi-supervised soft sensing, numerous scientific researchers have studied it. Yuan[7] proposes a nonlinear semi-active soft sensing modeling method based on semi-supervised weighted probabilistic principal component regression, to dispose of unequal data sets with only a small amount of tagged data. Luo[8] aims at the low prediction accuracy of reservoir parameters, combining LSSVM with semi-supervised learning, establishing a semi-supervised regression SLSSVM model. Ferreira[9] applies a semi-supervised soft sensor model to CO training, and improves the quality of unlabeled data by eliminating irrelevant features by feature ranking method. Jin[10] aims at most of the current chemical process soft sensor models do not use labeled samples, proposes a semi-supervised limit learning machine applied to the modeling of chemical process soft sensor.

To solve the problem of few labeled samples and poor generalization ability of soft sensor model in crude furfuryl alcohol distillation process, this paper proposes a semi-supervised soft sensor modeling method based on cosine similarity-discount. This method aiming at the problem of data multiplicity in crude furfuryl alcohol distillation process, by calculating cosine similarity between labeled and unlabeled samples in the same time interval, analyzes the semblance between labeled and unlabeled samples. Based on the definition of constraints, the data clustering of unlabeled samples is realized. The generalization ability of soft sensor model has been improved. In order to avoid the ill-conditioned problem caused by high-dimensional input variables, the clustered data are fused by discount factor value. A new data set of soft sensor modeling has been established. An actual data of a crude furfuryl alcohol distillation process (CFADP) simulation experiment were carried out. The results show that compared with the traditional soft sensor modeling method, the SSSMI-COSD method has better fitting performance and effectively improves the soft sensor model prediction accuracy for a crude furfuryl alcohol distillation process.

2. Data Sample Marking and Fusion Method Based on Cosine Similarity-Discount

Cosine similarity[11,12] is a method of calculating relevance. This method calculates the angle cosine between two vectors as a measure of similarity. The greater the cosine similarity, the greater the similarity between the two vectors; the smaller the similarity between the two vectors, the smaller the similarity between them[13].

The cosine similarity formula between variable \( X=(x_1, x_2, \ldots, x_n) \) and \( Y=(y_1, y_2, \ldots, y_n) \) is:

\[
\cos \langle X, Y \rangle = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\left( \sum_{i=1}^{n} x_i^2 \right)^{1/2} \left( \sum_{i=1}^{n} y_i^2 \right)^{1/2}} \tag{1}
\]

Where \( x_i \) and \( y_i \) represent the \( i \)th element values of variables \( X \) and \( Y \).

The cosine similarity values are \( C_j, 1 \leq j \leq p \).

In order to improve the data quality of samples of soft sensor modeling and eliminate the influence of abnormal data on the data quality, this paper proposes the numerical constraints of cosine similarity of clustering data to achieve high-quality clustering of unlabeled sample data with constraints.

If the cosine similarity between unlabeled sample data and labeled sample data is greater than the threshold value \( \xi \), it is considered that the current unlabeled sample data can be labeled. Otherwise, the unlabeled sample data can’t be labeled. The formula is expressed as follows:

\[
\begin{cases}
  C_j \geq \xi; & u_j, 1 \leq i \leq n; 1 \leq j \leq p \text{ labeled} \\
  C_j < \xi; & u_j, 1 \leq i \leq n; 1 \leq j \leq p \text{ unlabeled}
\end{cases} \tag{2}
\]

Because the length of clustered data can’t be determined, the value of the discount factor in the discount method[14] can be adjusted dynamically. At the same time, the function of new data can be enhanced, the influence of old data can be reduced[15], and the quality of input sample data with data multiplicity problem can be effectively improved.
At the same time, in order to ensure the integrity of clustering sample data in the same time period, a numerical calculation method of discount factor and constraints are proposed in this paper. Assuming that the sample size is \( P \) in the same time period after clustering, discount factor values are \( \lambda_1, \lambda_2, \ldots, \lambda_p \).

Then the formula for calculating the discount factor values is:

\[
\lambda_1 + \lambda_2 + \cdots + \lambda_p = 1
\]  

(3)

The constraint is as follows:

\[
\lambda_1 > \lambda_2 > \cdots > \lambda_p
\]  

(4)

The time span between the new labeled sample data with discount factor value \( \lambda_i \) and the original labeled sample data is the largest.

The sample set \( S \) of semi-supervised soft sensor modeling data is:

\[
S = \left\{ v_j(t_k) \mid k = 1, 2, \ldots, M; j = 1, 2, \ldots, N \right\}
\]  

(5)

Where \( v_j(t_k) \) is the \( j \)th input variables of time \( t_k \), \( y(t_k) \) is the output variables of time \( t_k \), \( t_k (k = 1, 2, \ldots, M) \) represents the sampling time of \( M \) sample points output by the system.

Soft sensor model is established based on modeling sample data set \( S \).

3. Soft Sensor Modeling Based on LSSVM

LSSVM[16] has better calculation speed, convergence accuracy and generalization performance. It is more suitable for crude furfuryl alcohol distillation process with small sample size.

The model of LSSVM[17] is:

\[
y(x) = \omega^{\top} \phi(x) + b
\]  

(6)

Where \( \phi(\bullet) \) is a nonlinear transformation function. \( \omega \) is an adjustable weight vector. \( b \) is an offset.

The objective function[18] is:

\[
\min J(\omega, \xi) = \frac{1}{2} \omega^{\top} \omega + \frac{C}{2} \sum_{i=1}^{l} \xi_i^{\top} \xi_i
\]  

(7)

subject to

\[
y_i = \omega^{\top} \phi(x_i) + b + \xi_i \quad (i = 1, 2, \ldots, l)
\]  

(8)

Where \( x_i, y_i \) are the corresponding input and output vectors. \( \xi_i \) is the difference between the system output value and actual value. \( C \) is a regularization parameter.

Because of the good universality of radial basis kernel function[19], the radial basis function is chosen.

\[
k(x, x') = \exp \left\{-\frac{\|x-x'\|^2}{2\sigma^2}\right\}, \quad \sigma > 0
\]  

(9)

4. Simulation and Analysis of Crude Furfuryl Alcohol Distillation Process

The principle diagram of distillation process in a crude furfuryl alcohol production enterprise is shown in Figure 1.
The two-column distillation process shown in Figure 1. Because of the purity quality index value of the product needs to be obtained through laboratory tests, and has a certain lag, it is necessary to establish a soft-sensing model. Soft sensor modeling parameters are shown in Table 1:

### Table 1. Descriptions of modeling parameters in DP SSM.

| Modeling parameter | Description                  |
|--------------------|------------------------------|
| $u_1$              | T1001 Distillation column bottom temperature |
| $u_2$              | T1001 Distillation column middle temperature |
| $u_3$              | T1002 Distillation column bottom temperature |
| $u_4$              | T1002 Distillation column middle temperature |
| $y$                | Product purity               |

The simulation experiment data used in this paper are from DCS system of a furfuryl alcohol production enterprise. The auxiliary variable sampling period is 1 h, and the dominant variable sampling period is 12 h.

The soft sensor model based on traditional labeled sample data adopts the following model structure:

$$ \hat{y}_L (t_k) = f_{LDP} (u_j (t_j)), j = 1, 2, 3, 4 \quad (10) $$

According to the characteristics of crude furfuryl alcohol distillation process and related expert knowledge, the model structure based on semi-supervised clustering fusion data is as follows:

$$ \hat{y}_S (t_k) = f_{SDP} (v_j (t_k)), j = 1, 2, 3, 4 \quad (11) $$

$$ v_j (t_k) = f_{DU} \begin{bmatrix} u_j (t_{k-n}) \\ \vdots \\ u_j (t_k) \end{bmatrix} \quad (12) $$
Where $y(t_k)$ is the purity obtained in the laboratory at time $t_k$, $u_j(t_{k,a})$ represents the sampling value at the time $t_{k,a}$ of the $j$th auxiliary variable, $v_j(t_k)$ represents a discounted fusion value of $u_j(t_{k,a})$. $\hat{y}(t_k)$ is the predicted value of soft sensor model.

According to the cosine similarity value of unlabeled sample data, the clustering constraint $\xi = 0.98$ is setted in this paper.

The parameters of LSSVM soft sensor modeling method are set to: $C = 30, \sigma^2 = 3$.

In this paper, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used to evaluate the fitting and prediction performance of regression analysis.

The formula of RMSE is:

$$RMSE = \left( \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \right)^{1/2}$$  \hspace{1cm} (13)

The formula of MAE is:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$ \hspace{1cm} (14)

Where $N$ is the total number of samples, $y_i$ is the predicted output value, and $y_i$ is the actual value.

The training results of soft sensor model based on labeled sample data and semi-supervised fusion data are shown in Figure 2 and Figure 3.

Compared with Figure 2 and Figure 3, the training curve shows some deviations. However, the training curve based on semi-supervised fusion data is closer to the actual data and better than the training curve based on labeled sample data.

**Table 2. Comparison of fitting performance of different training samples.**

| Training samples                | RMSE  | MAE  |
|--------------------------------|-------|------|
| Labeled Samples                | 0.0331| 0.0256|
| Semi-supervised Fusion Samples | 0.0291| 0.0219|

The prediction results of soft sensor model test data based on labeled sample data and semi-supervised fusion data are shown in Figure 4 and Figure 5.
Compared with Figure 4 and Figure 5, the data prediction curves show some deviations, but the prediction curves based on semi-supervised fusion data are more compact and uniform than those based on labeled sample data. The soft sensor model based on semi-supervised fusion data improved RMSE and MAE by 7.48% and 51.94% respectively compared with labeled sample data.

Table 3. Comparison of predictive performance of different test samples.

| Training samples                  | RMSE  | MAE   |
|-----------------------------------|-------|-------|
| Labeled Samples                  | 0.1003| 0.0799|
| Semi-supervised Fusion            | 0.0928| 0.0384|

Comparing the LSSVM fitting curve of Figure 2 and Figure 3 with the RMSE and MAE evaluation index values of Table 2, the following conclusions can be drawn: although the training curve has some deviation, the semi-supervised dynamic fusion sample data fitting curve is better than the labeled sample data fitting curve, and RMSE, MAE evaluation index values have been improved.

Comparing the LSSVM prediction curve of Figure 4 and Figure 5 with the RMSE and MAE evaluation index values of Table 3, the following conclusions can be drawn: Although the prediction curve has some deviation, the semi-supervised dynamic fusion sample data prediction curve is better than the labeled sample data prediction curve, and RMSE, MAE evaluation index values have improved.

Through the simulation modeling process and test data prediction based on the actual industrial data of crude furfurfuryl alcohol distillation process, and comparing the corresponding evaluation index values, combined with the above analysis, the following conclusions can be drawn:

The soft sensor model based on cosine similarity and discount weighted semi-supervised clustering fusion data is superior to the soft sensor model based on labeled sample data in fitting degree and prediction accuracy.

5. Conclusion
In this paper, the crude furfuryl alcohol distillation process is taken as the research background. The simulation and modeling of the actual data of the distillation process show that the proposed semi-supervised soft sensing modeling method based on cosine similarity-discount can realize semi-supervised clustering and fusion of the unlabeled sample data of crude furfurfuryl alcohol distillation process. The semi-supervised data fusion can effectively improve the fitting degree of the soft sensor...
model and the accuracy of data prediction through the simulation experiment of establishing the soft sensor model.

Acknowledgments
Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

References
[1] Wu X C and Li J T 2005 J. Kun. Uni. Sci. Tec. 51-4
[2] Li D 2016 Liaoning. Uni. Tec.
[3] Shang N L, Wang X L, Shen K X, Lu Y L, Ma X P and Lan Y H 2015 Com. APP. Soft. 162-6+179
[4] Zhang J and Yang H Z 2013 Com. App. Che. 1453-6
[5] Wang W, Xi J, Chong A and Li L 2017 IEEE. Tra. Hum. Sys. 650-60
[6] Ge Z, Huang B and Song Z 2015 J. Che. 793-804
[7] Yuan X, Ge Z, Huang B, Song Z and Wang Y 2016 IEEE. Tra. Ind. Inf. 532-41
[8] Luo W P, Li H Q and Shi N 2016 App. Geo. 406-15
[9] Ferreira V, Souza F and Rui A 2017 IEEE. Int. Con. Ind. Inf. 257-62
[10] Jin X 2015 Zhejiang. Uni. Tec.
[11] Kalhori H, Alamdari M and Lin Y 2018 Mea. 648-57
[12] Hernandez A and Garcia N 2016 IEEE. Lat. Ame. Tra. 2857-61
[13] Geng Z Q, Hu H X and Han Y M 2017 J. Bei. Uni. Che. Tec. 87-94
[14] Fan J and Wang G 2017 Opt. Meth. Soft. 139-49
[15] Song K, Wang H and Li P 2004 J. Che. Ind. Eng. 942-6
[16] Suykens J and Johan A K 2002 Int. J. Circ. Theor. Appl. 605-15
[17] Li C X, Ding X D and Zheng X F 2017 J. Vib. Sho. 52-8
[18] Suykens J, Brabanter J and Lukas L 2002 Neurocom. 85-105
[19] Roushangar K, Saghebian S M and Mouaze D 2017 Int. J. Sed. Rese. 51-4