CHEMICAL ENGINEERING TRANSACTIONS

VOL. 103, 2023

A publication of

AIDIC

The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Petar S. Varbanov, Panos Seferlis, Yee Van Fan, Athanasios Papadopoulos
Copyright © 2023, AIDIC Servizi S.r.l.
ISBN 979-12-81206-02-1; **ISSN** 2283-9216

DOI: 10.3303/CET23103011

A Conditional Entropy Based Feature Selection for Soft Sensor Development in Chemical Processes

Cheng Ji[a], Fangyuan Ma[b], Jingde Wang, Wei Sun[a,*]

[a]College of Chemical Engineering, Beijing University of Chemical Technology, 100029, North Third Ring Road 15, Chaoyang District, Beijing, China
[b]Center of process monitoring and data analysis, Wuxi Research Institute of Applied Technologies, Tsinghua University, 214072 Wuxi, China
sunwei@mail.buct.edu.cn

Data-driven soft sensors have been extensively studied in the process industry to provide an accurate online estimation of quality-related variables with easy-to-measure variables. For chemical processes with massive process variables, the performance of soft sensor models could be significantly improved by variable selection because part of these measurements is redundant or independent of quality-related variables. Generally, the variable selection is achieved by ranking process variables in order of their importance to the quality-related variables by correlation analysis. However, considering that correlation analysis methods are relative measures of variable dependence, the determination of the final variable set is quite subjective because there are several user-defined parameters. To overcome this limitation, a conditional entropy-based feature selection method is proposed. Considering that information entropy measures the degree of system chaos, the proposed method is based on the idea that the quality-related variable can be fully estimated if its information entropy is reduced to 0 by a set of optimal variables. Independent variables are first sorted by mutual information, then the conditional entropy of the quality-related variable is calculated iteratively until the result is close to 0, which indicates that the quality-related variable can be fully estimated by the currently selected variables. The final variable set is determined by further excluding the redundant variables according to information gain. The effectiveness and superiority of the method are validated through a case study on an industrial naphtha cracking furnace.

1. Introduction

With the high-speed development of information technology in the process industry, sensors have been widely applied to acquire measurements for process monitoring and control. Based on these online measurements, soft sensors can be established to provide estimations on critical process variables which reflect the state of process operation but are hard to measure (Kadlec et al., 2009). Especially for chemical processes, online measurements of quality-related indicators are generally not available, and can only be obtained through offline analysis, which is expensive and will lead to delays in real-time response. Therefore, data-driven soft sensors have attracted significant attention to reduce costs in the chemical industry.

The most popular data-driven soft sensors can be summarized as Principal Component Regression (PCR), Partial Least Squares (PLS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN) (Souza et al., 2016). All these methods are designed to extract key features from historical data and find a mapping between the features and the quality-related variable. To ensure that the mapping relationship is reliable, several factors should be considered to avoid overfitting. In addition to the sample size of the training dataset, the probability of model overfitting also increases with the increase of dimension. It has been proven that the soft sensor performance can be improved by selecting vital variables that are highly correlated with the quality-related, rather than using all available ones (Wang et al., 2015). Therefore, a series of feature selection methods are proposed for variable selection in soft sensors.

Mutual Information (MI) has been widely applied to sort variables in order of their importance to the quality-related variable. MI can be used to identify both linear and nonlinear variable correlation, but it is limited in evaluating variable dependence because there is no upper limit in MI. Although normalized MI is further

Paper Received: 08 April 2023; Revised: 10 August 2023; Accepted: 30 August 2023
Please cite this article as: Ji C., Ma F., Wan J., Sun W., 2023, A Conditional Entropy Based Feature Selection for Soft Sensor Development in Chemical Processes, Chemical Engineering Transactions, 103, 61-66 DOI:10.3303/CET23103011

proposed to scale MI to between 0 and 1 (Estevez et al., 2009), a user-defined threshold is still required to determine whether a variable could be selected. In a previous study, it is suggested to employ the Monte Carlo method to obtain a threshold based on MI of random simulation data (Ji et al., 2023). However, the redundancy of the candidate process variables has not been considered. Regarding this issue, it is suggested that the feature selection should be conducted based on the analysis of relevance and redundancy. The max-relevance and min-redundancy criterion has been adopted in several recently proposed MI-based feature selection approaches (Che et al., 2017). For example, a feature selection method based on joint MI maximization is proposed, in which conditional MI is employed to measure the importance of the candidate variable given a pre-selected feature subset (Bennasar et al., 2015). Similarly, a max-relevance and min-redundancy method based on neighborhood MI is proposed (Xiao et al., 2019), and a maximum dynamic relevancy minimum redundancy-based feature selection algorithm is also proposed (Yin et al., 2023). In this way, the redundancy of the candidate variables can be considered, but the desired number of features is still a subjective parameter that needs to be defined by users. To obtain a unified criterion for variable selection, a conditional entropy-based feature selection method is proposed in this work. The main idea of the proposed method is that the quality-related variable can be fully estimated if its information entropy is reduced to 0 by a set of independent variables. Therefore, the primal information entropy of the quality-related variable is first calculated, then candidate variables are sorted by MI and included continuously until the conditional entropy of the quality-related variable is reduced to 0. The conditional entropy is derived in this work for the consideration of two forms of information redundancy. Based on the proposed strategy, the variable set can be determined to provide an accurate estimation of the quality-related variable with minimum redundancy and no user-defined parameters. Meanwhile, the importance of the selected variables is also obtained according to their respective information gain. The effectiveness of the proposed method is validated through a prediction task on the composition of the product ethylene in an industrial naphtha cracking furnace.

2. Preliminaries

In this section, preliminary knowledge including information theory and SVR is briefly introduced as the basis of the proposed method.

2.1 Information theory

Information theory is a branch of applied mathematics developed by combining probability theory and mathematical statistics. Its beginnings can be traced back to the definition of information entropy,

$$H(X) = -\sum_{i=1}^{n} p(x_i)\, log(p(x_i)) \tag{1}$$

where X is a random variable with n samples, and $p(x_i)$ denotes the probability distribution of the ith sample. Information entropy is a measurement that could represent the uncertainty of the random variable. The larger the information entropy values, the higher the uncertainty of the variable, and the less information the data contain. In soft sensors, independent variables are selected to reduce the uncertainty of the quality-related variable. Generally, the selected variables should be highly correlated with the quality-related variable. MI has been commonly used to measure the dependence between a pair of variables,

$$I(X,Y) = \sum_{x}\sum_{y} p(x,y)\, log(\frac{p(x,y)}{p(x)p(y)}) \tag{2}$$

where $p(x,y)$ is joint probability distribution. If X is independent of Y, the MI is equal to 0. Comparatively, the MI will be large when there is a strong correlation between X and Y. However, a limitation arises in evaluating variable correlation because there is no upper limit of MI. Several solutions such as normalized MI and the Monte Carlo method have been proposed to handle this issue. A threshold of MI can be obtained to determine whether a variable can be selected. Another issue to be considered is the redundancy among candidate variables. Conditional MI is proposed to calculate MI between a candidate variable and the quality-related variable given a set of selected variables,

$$I(X,Y|Z) = \sum_{x}\sum_{y}\sum_{z} p(x,y,z)\, log(\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \tag{3}$$

where $I(X,Y|Z)$ denotes MI between X and Y given Z. On the base of above studies, a unified criterion for variable selection in soft sensor development is derived and will be analyzed in Section 3.

2.2 Support vector regression

SVR is a well-known machine learning method for regression tasks, which shows superior performance in nonlinear processes with a small sample. Given a training dataset $\boldsymbol{x}$, SVR aims to find a regression function that can fit all training samples,

$$f(x) = \mathbf{w}^T \Phi(\mathbf{x}) + b \tag{4}$$

where $\mathbf{w}$ is a coefficient vector in feature space, $\Phi(\mathbf{x})$ is a kernel function to map the input into a high-dimension feature space, and the b is the intercept. The $\mathbf{w}$ and b can be determined by solving the optimization problem referred in previous research. The introduction of the kernel trick enables SVR to be applied in nonlinear processes because data tend to be linearly separable in high-dimension space. Moreover, SVR also performs well in small sample datasets because it is based on statistical learning. In this work, SVR is adopted to verify the effectiveness of the proposed feature selection method for two reasons. First, chemical processes are characterized by high process nonlinearity. On the other hand, the sampling frequency of the quality-related variable is relatively low, which means the number of available samples is limited for training the regression model.

3. The proposed conditional entropy-based feature selection

In this section, the proposed conditional entropy-based feature selection strategy is derived, in which two forms of redundancy are discussed and an objective criterion for selecting independent variables in soft sensors is provided.

3.1 Conditional entropy-based feature selection strategy

As mentioned before, the information contained in a process variable can be measured by information entropy. If a set of independent variables can be selected to reduce the information entropy of the dependent variable to 0, then this target variable can be fully estimated. Based on this idea, a conditional entropy-based feature selection strategy is proposed by continuously including candidate independent variables to calculate the conditional entropy of the dependent variable iteratively. The final selected variable set can be determined when the conditional entropy is reduced to near 0 in the latest iteration.
For the first iteration, the conditional entropy can be calculated as follows,

$$H(Y|X_1) = H(Y) - I(X_1, Y) \tag{5}$$

where Y is the quality-related variable to be estimated, and X_1 is the first candidate variable. For more than two variables, the redundancy among candidate variables has to be considered during the calculation of condition entropy. As shown in Figure 1, the two forms of redundancy are displayed in red and blue. Generally, the redundancy between X_1 and X_2 is represented as the MI between each other. However, regarding the regression task in this work, only the form of redundancy in red should be considered because the other form of redundancy in blue does not contribute to any information gain to the quality-related variable Y. Therefore, the conditional entropy in the proposed method should be calculated as follows,

$$H(Y|X_1, X_2) = H(Y) - I(X_1, Y) - I(X_2, Y) + I(X_1, X_2) - I(X_1, X_2|Y) \tag{6}$$

where $I(X_1, X_2|Y)$ denotes the conditional MI between X_1 and X_2 given Y. It represents the amount of information in blue in Figure 1. Similarly, the conditional entropy of n variables can be derived as follows,

$$H(Y|X_\mathrm{i}, \mathrm{i} = 1,2, \ldots, \mathrm{n}) = H(Y) - \sum_{i=1}^{n} I(X_\mathrm{i}, Y) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(I(X_\mathrm{i}, X_\mathrm{j}) - I(X_\mathrm{i}, X_\mathrm{j}|Y) \right) \tag{7}$$

The iteration terminates when the conditional entropy is closed to 0, and current X_i is the selected variable set used to estimate Y. The degree of importance of these variables can also be obtained according to the information gain each one contributes.

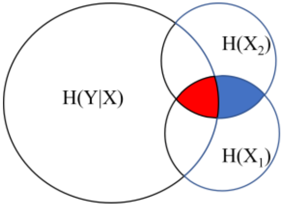

Figure 1: Two forms of redundancy in calculating conditional entropy

3.2 The implementation procedures of the proposed feature selection method

According to the previous analysis, the overall procedures of the proposed variable selection strategy can be summarized as follows,
Step 1: The initial information entropy of the quality-related variable to be estimated is calculated.
Step 2: Candidate independent variables are sorted according to MI.
Step 3: Candidate variables are included in the order of MI to calculate the conditional entropy of the quality-related variable iteratively.
Step 4: The variable set is determined when the conditional entropy is reduced to near zero.
Step 5: The importance of each variable is further analysed according to the information gain each variable contributes to the quality-related variable.
It is worth noting that there is no user-defined parameter in the proposed method, which means it provides an objective criterion for selecting independent variables in soft sensors. Moreover, the redundancy among candidate variables is also considered by analysing the two forms of information redundancy, based on which the calculation of condition entropy is derived. The effectiveness of the proposed method will be verified with SVR in the next section.

4. Case study

In this section, an industrial naphtha cracking process to produce ethylene is adopted as a case study to verify the proposed method.

4.1 Process and data description

The diagram of key equipment, the naphtha cracking furnace is shown in Figure 2, in which a total of 63 process variables are measured online, including flow rate, temperature, and pressure. A detailed description of these variables can be referred to in our previous study(Ji et al., 2020). Ethylene is a key product of this process and it plays an important role in acquiring its real-time composition. However, the composition of the product is hard to measure online, and can only be obtained through offline analysis. As a result, the sampling period of ethylene composition is 16 minutes, which is much longer than the 1 minute for other variables. Therefore, it is of great significance to apply soft sensors to provide a real-time estimation of the ethylene composition using easy-to-measure variables.
Because of the long sampling period of the ethylene composition, available labeled samples are limited. In this work, a total of 1,000 samples are divided into a training dataset, and the other 140 samples are used as the test dataset to verify the performance of the model. For the 63 available process variables, some of them are redundant or not associated with the ethylene composition. The model tends to be overfitting if all of these variables are adopted. Therefore, the proposed conditional entropy-based feature selection method is employed to select important variables and eliminate redundancy.

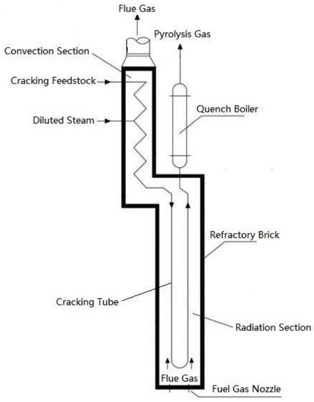

Figure 2: Diagram of the naphtha cracking furnace used in this work

4.2 Results analysis and discussion

As shown in the variable selection results in Table 1, the initial information entropy of the ethylene composition is about 5.4, which is calculated using Eq.(1). Candidate variables are continuously included in the order of MI, and the conditional entropy defined in Eq.(7) is calculated. It can be shown that the conditional entropy drops to almost 0 at the 15th iteration, indicating that the ethylene composition can be estimated by the current variable set. Finally, a total of 14 variables are selected by the proposed method and their importance can be referred to the information gain.

Table 1: A summary of the variable selection result by the proposed method

Iteration No.	Selected variables	Information entropy	Information gain
1	-	5.399737	-
2	(1)	4.821703	0.578034
3	(1,2)	4.306121	0.515582
4	(1,2,56)	4.203153	0.102968
5	(1,2,56,58)	3.983944	0.219209
6	(1,2,56,58,20)	3.778017	0.205921
7	(1,2,56,58,20,21)	3.702920	0.075097
8	(1,2,56,58,20,21,61)	2.794307	0.908613
9	(1,2,56,58,20,21,61,54)	2.195488	0.598819
10	(1,2,56,58,20,21,61,54,17)	2.243739	-0.048251
11	(1,2,56,58,20,21,61,54,17,18)	2.310592	-0.066853
12	(1,2,56,58,20,21,61,54,17,18,22)	2.726768	-0.416176
13	(1,2,56,58,20,21,61,54,17,18,22,55)	1.537778	1.1889899
14	(1,2,56,58,20,21,61,54,17,18,22,55,19)	1.504173	0.0336050
15	(1,2,56,58,20,21,61,54,17,18,22,55,19,57)	0.146901	1.3572720

To verify the effectiveness of the proposed method, SVR models are established using 2 to 63 variables in order of MI. Then the models are tested through the test dataset and their performance is evaluated through Mean Absolute Error (MAE). According to the results shown in Figure 3, the MAE first decreases with the increase of the number of variables because more information is included to estimate the ethylene composition. However, the MAE increases when the number of variables is larger than 17, indicating that the following variables are redundant. The model reaches the best performance when the number of variables is 15, which is consistent with the variable selection results in Table 1.

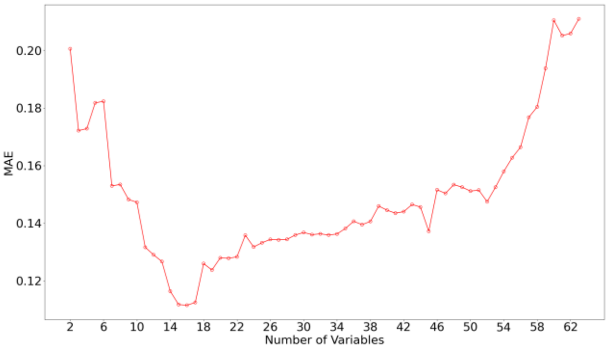

Figure 3: The prediction accuracy of SVR using 2 to 63 variables

The prediction results using selected variables and all available variables are displayed in Figure 4(a). The model established by selected variables performs an accurate estimation of the ethylene composition, while the model established by all available variables does not perform the best. The reason lies in the overfitting of the model caused by the redundancy among candidate variables. Moreover, the importance of each selected variable can be evaluated simultaneously by the information gain. According to Table 1, the information gain of variables 56, 21, 17, 18, 22, and 19 is small, indicating that these variables are not important. To further prove it, prediction performance without an important variable and all 6 unimportant variables is displayed in Figure 4(b). The results show that the MAE of the model without unimportant variables is 0.122, while the MAE of the model without an important variable is 0.256. It can be concluded that the model maintains an accurate prediction without all these 6 unimportant variables, while the prediction accuracy of the model decreases significantly after an important variable is excluded. The results demonstrate the effectiveness of the proposed variable selection strategy in selecting key variables and eliminating redundancy in soft sensor development.

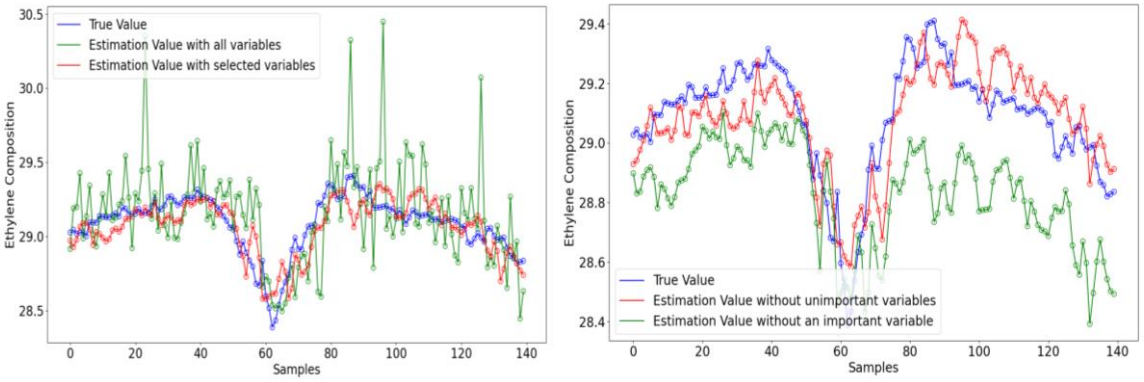

Figure 4: (a) Prediction performance with selected variables and all variables (b) Prediction performance without important variables and unimportant variables

5. Conclusions

In this work, a conditional entropy-based feature selection method is proposed to provide an objective criterion for selecting independent variables in soft sensors. According to the results in a case study on the industrial naphtha cracking process, it can be concluded that redundancy contained in process variables could lead to the overfitting of the soft sensor model. The proposed method addresses this issue by introducing information gain to measure the importance and redundancy of variables. Based on the idea that a variable can be fully estimated if its information entropy can be reduced to 0 given a set of independent variables, the conditional entropy of the target variable is then adopted as an indicator to obtain the optimal variable set. The effectiveness of the proposed strategy is verified through the prediction of the ethylene composition. Compared with existing feature selection methods, the proposed method does not require any user-defined parameters, indicating that it can also be adapted to soft sensor development in many other industrial processes. However, the redundancy variables, which may also be useful, are excluded together with the irrelevant variables in this work. Future works could focus on the use of these redundancy terms.

Acknowledgments

The authors gratefully acknowledge the following institutions for support: The National Natural Science Foundation of China (Grant No. 22278018).

References

Bennasar M., Hicks Y., Setchi R., 2015, Feature selection using Joint Mutual Information Maximisation. Expert Syst. Appl., 42, (22), 8520-8532.

Che J., Yang Y., Li L., Bai X., Zhang S., Deng C., 2017, Maximum relevance minimum common redundancy feature selection for nonlinear data. Information Sciences, 409-410, 68-86.

Estevez P. A., Tesmer M., Perez C. A., Zurada J. M., 2009, Normalized Mutual Information Feature Selection, IEEE Trans. Neural Networks, 20, (2), 189-201.

Ji C., Ma F., Wang J., Sun W., 2023, Profitability related industrial-scale batch processes monitoring via deep learning based soft sensor development. Comput. Chem. Eng., 170.

Ji C., Ma F., Zhu X., Wang J., Sun W., 2020, Fault Propagation Path Inference in a Complex Chemical Process Based on Time-delayed Mutual Information Analysis. Computer Aided Chemical Engineering, 48, 30th European Symposium on Computer Aided Process Engineering, 1165-1170.

Kadlec P., Gabrys B., Strandt S., 2009, Data-driven soft sensors in the process industry. Comput. Chem. Eng., 33, 795-814.

Souza F.A.A., Araújo R., Mendes J., 2016, Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst., 152, 69-79.

Wang Z.X., He Q.P., Wang J., 2015, Comparison of variable selection methods for PLS-based soft sensor modeling. J. Process Control, 26, 56-72.

Xiao L., Wang C., Dong Y., Wang J., 2019, A novel sub-models selection algorithm based on max-relevance and min-redundancy neighborhood mutual information. Information Sciences, 486, 310-339.

Yin K., Zhai J., Xie A., Zhu J., 2023, Feature selection using max dynamic relevancy and min redundancy. Pattern Analysis and Applications, 26, 631–643.