Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(3):1164-1169

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A method of variables selection for soft sensor based on distributed mutual information

Xia Yuan[1*], Huizhong Yang[1] and Nam Sun Wang[2]

[1]*Key Laboratory of the Ministry of Education for Advanced Control in Light Industry Process, Jiangnan University, China*
[2]*Department of Chemical & Biomolecular, University of Maryland, College Park, MD, 20742, US*

ABSTRACT

A method of distributed mutual information is proposed for selecting secondary variables in a soft sensor. The mutual information between the predicted primary variable and the secondary variables is obtained by estimating the probability distribution of every secondary variable and the predicted variable. This information indirectly reflects the linear or nonlinear correlations between the predicted variable and the secondary variables. A threshold value is obtained by t-test approach as a criterion to judge the correlation of variables. Subsequently, the variables whose mutual information is greater than the threshold value are further screened to be selected as the relevant variables or to be discarded as weakly relevant variables. Finally, a soft sensor model is built based on the support vector machine algorithm with the selected secondary variables.

Key words: Soft sensor modeling; distributed; mutual information; support vector machine

---

INTRODUCTION

Data-driven soft sensors have been developed and implementedfor a long time. They gained on popularity with the increasing availabilityof recorded data in the process industries and availabilityof computational power to process the data. The collected data,also referred to as historical data, can be exploited by statisticaland machine learning techniques to obtain additional informationthat can be used to make decisions towards more efficient andsafe process operation. This kind of information can, for instance,be an instant prediction of the variables that are related to theproduct quality, which can be achieved using online predictionsoft sensors, or the estimation of current process state, which can beachieved using process monitoring and fault detection soft sensors[1-2]. However,this task is not trivial because historical data are often data richbut information poor (Dong &McAvoy, 1996) and therefore themodel building on its basis is a challenging task.

The first generation of data-driven soft sensors relied on offlinemodelling using the recorded historical data. In such a case, thecollected historical recordings are used for the model identification.This step may for instance include the identification of optimalweights of an Artificial Neural Network(ANN) or principal components of a Principal Component Analysis(PCA)-based soft sensor[3-4]. However, in order to guarantee the success of the offlinesoft sensors, there are several conditions that have to be fulfilled.Most critically, the historical data has to contain all possible futurestates and conditions of the process. This includes not only thestates in which the process can be operated but also states relatedto environmental changes, changes of the process input materials,etc. Even if the collected historical data contains all the requiredprocess states, another difficulty is to select a model type, and itsparameters, in such a way that the model can comprehend all thedifferent conditions. This results in high model complexity, whichin turn demands and a large number of historical data for the modeldevelopment.

A prominent issue in the process industry is the accurate online measurement of important quality parameters to ensure product quality, smooth continuous unit operation, and maximum production capacity. As production processes involve a large number of complex physical interactions, chemical reactions, and conversion and transfer of materials and energy, existing hardware cannot directly yield accurate on-line measurement of quality parameters. Instead, soft measurement technique estimates primary variables indirectly by mapping the relationship between readily measured variables and the primary variables that cannot be directly measured according to a set of optimization rules.

The first step in the implementation of soft measurement technology is the selection of auxiliary variables. The auxiliary variables are selected mainly based on sensitivity, accuracy, specificity, and robustness. The impor tance of variable selection has been mentioned in the literature [5-6]. A well-selected set of auxiliary variables can effectively overcome the dimensionality problem and improve the model's validity. One way to select auxiliary variables is to identify a set of measurable auxiliary variables that affect primary variables based on a mechanistic analysis. An alternative way is based on a statistical analysis of the correlation in the sample data to exclude irrelevant variables. Variable selection methods based on PLS and PCA have been reported in the literature [7]. Other variable selection methods are based on mutual informationcalculated indirectly from the entropy and conditional entropy [8]. Another approach estimates joint probability density function (multi-dimensional Gaussian distribution) to determine the input variables of a neural network according to mutual information[9].

In this paper, the number of auxiliary variables is determined by mutual information. An example of estimating the concentration of phenol with a soft sensor model illustrates how to calculate the mutual information between auxiliary variables and primary variables from a set of samples. Subsequently, a threshold is determined by the t-test method for judging the correlation between the primary variable and every auxiliary variable[10-11], and the variables meeting the threshold criterion areselected. Finally, a soft sensor model of phenol is refined by removing redundant variables.

MUTUALINFORMATIONAND T-TEST

Mutual information is defined as the amount of information of a random variable that is contained in another random variable. The mutual information of variables between X and Y is definedas[12]:

$$I(X;Y)=\iint m(x,y)\log\frac{m(x,y)}{m_X(x)m_Y(y)}dxdy \quad (1)$$

where $m(x,y)$describes the joint probability density of X and Y, and $m_X(x)$and $m_Y(y)$ respectively describe the marginal probability density of X and Y. A high value of the mutual information means variable Y contains more information on variable X, thus, a greater correlation between the two variables. Therefore,mutual information can be used to select auxiliary variables in a soft senor.

In the case of unknown variance σ^2, estimation and hypothesis testing for the mean value of a normal population is commonly carried out with t-statistic.(Assumethe sample $I \sim N(\mu,\sigma^2)$ (for mean value μ and variance σ^2),thesample$[I_1,I_2,...,I_n]$ of general statistic, the statistics $T=\frac{\bar{I}-m}{S/\sqrt{n}}\sim t(n-1)$ is constructed with the mean value $\bar{I}$ and variance S^2.Given the problem $H_0:m=I_0,H_1:m\neq I_0$ (where I_0 describes the correlation between the primary variable and other random variables) and under the testlevel a, the probability of an unlikely event is given by $P\left\{\left|\frac{\bar{I}-m}{S/\sqrt{n}}\right|\geq t_{\frac{a}{2}}(n-1)\right\}=a$ and the rejection domain is $\left(-\infty,-t_{\frac{a}{2}}(n-1)\right]\cup\left[t_{\frac{a}{2}}(n-1),+\infty\right)$. The hypothesis is rejected when the statistics value belongs to the rejection domain; otherwise, the hypothesis is accepted.

Since t-test is used here to find the maximum mean value $I_{0\max}$ of mutual information for accepting the nullhypothesis[13], this is a recursive process for H_0.Assume the mutual information between the primary variable

[1]This work is supported by National Nature Science Foundation of China(61273070)

and a set of normally distributed random variables is I_i. Tests are randomly repeated n times, the mean value and variance of correlation are $\bar{I} = \sum_{i=1}^{n} I_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(I_i - \bar{I})^2$ .Given a test level α, $t_{\alpha/2}(n)$ can be found from a table. Finally

$$I_{0\max} = \bar{I} + t_{a/2}(n)\sqrt{S^2/n} \quad (2)$$

is obtained from the formula $|\frac{\bar{I} - I_{0\max}}{\sqrt{S^2/n}}| = t_{a/2}(n)$. $I_{0\max}$ initiallyreflects the highest mutual information between the primary variable and unrelated variables. In order to ensure certain redundancy, the threshold $\delta_1 (d_1 = d_1 I_{0\max})$ is determined finally, $d_1 \in [1.0, 1.1]$.

VARIABLE SELECTION ALGORITHM OF DISTRIBUTED MUTUALINFORMATION

John et al. divide input space into three categories, i.e., strongly relevant variables, weakly relevant variables, and irrelevantvariables[14]. The optimal variable subset should not contain irrelevant variables. Because inclusion of weakly relevant variables leads to variable redundancy, the optimal variable subset should contain only strongly relevant variables but exclude weakly relevant variables and irrelevant variables.

Considering the correlation between input variable X and output variable Y, the mutual informationI(X;Y) ofvariables X and Y represents the amount of information that Y contains. If the input variable X_i meets:

$$I(X_i;Y) \geq \delta_1 \quad (3)$$

where δ_1 is a correlation threshold. It indicates that X_i contains a certain amount of information of Y, i.e., X_i is a relevant variable of Y. If the input variable X_i does not meet the above formula (3), the variable X_i contains only negligible information on Y or none at all, and X_i is regarded as an irrelevant variable and is excluded. Thus, the variable subset F consists of both strongly relevant variables and weakly relevant variables that satisfyEqn 3.It is necessary that the subset F is further screened because there may exit redundancy in the weakly relevant variables. Redundant variables should be eliminated from the subset F to maintain the correlation between the subset F and output variable Y.

Mutual information refers to the correlation between two sets of events, and it is difficult to estimate mutual information in a high dimensional case. In order to simplify calculation, a subset of variables is replaced by a single variable to measure the redundancy between a subset of variables and output variable Y. This article draws lessons from the algorithm of min-redundancy max-relevance (MRMR) [15], which uses the mean value of mutual information to evaluate the degree of redundancy. In other words, the mutual information between the output Y and the subset F,I(F;Y), is evaluated as the mean value of the mutual information between Y and each of the element of the subset F, as shown in the following equation :

$$I(F;Y) = \frac{1}{|F|}\sum_{F_i \subset F} I(F_i;Y) \quad (4)$$

where F_i is an element of the subset F.

In this paper, redundant variables are removed via a backward method. When deleting an input variable X_i, if the mutual information between the output Y and subset F meets the redundancy constraints given inEqn (5), X_i is a redundant variable; otherwise, it is a relevant variable.

$$I(F;Y) - I(F - X_i;Y) < \delta_2 I(F;Y) \quad (5)$$

where δ_2 is a given threshold of redundancy, $\delta_2 \in [0,1]$

The specific steps in the variable selection algorithm of distributed mutual information are as follows:

1) The mutual information of variables $I_i(X_i;Y), i = 1,2,3,...,n$ can be calculated from Eqn (1).

2) $I_{0\max}$ isobtained from Eqn (2) based on t-test. Pick a relevance threshold δ_1 and select all input variables X_i ($i = 1, 2, ..., m$)that meet the correlation condition (3). Arrange X_i in a descending order in accordance with the

values of mutual information to structure the subset F.Let $M = F$.

3) The mutual information between a subset of the remaining input variables and the output variable, $I(M - M_i; Y)$,is calculated by Eqn (4)after each variable $M_i (i = 1,2,..., m)$ is removed in turn.

4) The variable M_i meets Eqn (5), it is redundant and excluded; otherwise, M_i is retained.

5) Repeat Steps 3) and 4) until each variable has been removed in turn. Then the algorithm stops.

There are two adjustable parameters in the above algorithm, the relevance threshold δ_1 and the redundancy threshold δ_2 . δ_1 is determined by $I_{0\max}$ which is obtained from the t-test. The higher the δ_1 value is, the higher the requirement of the correlation between variables, and subsequently a less number of variablesis selected in the first step. The higher the δ_2 value is,the higher the requirement of redundancy of variables selected, and subsequently a less number of the variables in the optimal subset M.

Casestudy

The example is taken from a BisphenolA(BPA)plant where a soft sensor estimates on-line the concentration of phenol in a crystallizer tower C303. The material exiting melting pot V304 is recycled back to crystallizer tower C303 for crystallizing again. The six auxiliary variables determined through an initial analysis are the three physically measured variables (temperature, level, and mass flow of crystal tower C303) and the three estimated variables (concentration of phenol and BPA in the V304 outlet and concentration of BPA-24 in the C303 outlet). Finally the concentration of phenol in C303 is estimated with a Support Vector Machine(SVM) model. In order to select the auxiliary variables related to the primary variable (the concentration of phenol in crystallizer unit C303), 150 groups of data were retrieved from the production site, with 100 groups as the model training set and the rest as the test set. The relevance and redundancy between the primary variable and auxiliary variables in the training data set were analyzed with the proposed method of distributed mutual information. A support vector machine model was established with the optimal subset of measurements and verified with the test data set.

The mutual information of each auxiliary variable and the primary variable is shown in Table 1. The mutual information in a descending order is: output BPA of V304> output phenol of V304> material flow of C303> level of C303> temperature of C303> output BPA-24 of C303.

Table 1.Mutual information between auxiliary variables and phenol

Variable	Mutual information I
Outlet BPA of V304	0.1275
Outlet phenol of V304	0.1176
Material flow of C303	0.0313
Level of C303	0.0075
Temperature of C303	0.0032
Outlet BPA-24 of C303	0.0005

First, it is necessary to determine a threshold for selecting variables of relevance. In order to determine the critical value of rejection domain where the relevance of the primary variable and auxiliary variables falls, a group of random sample set following a normal distribution between 0-1 is generated with a computer. The mutual information between the primary variable and them is calculated, and the process is repeated 30 times to obtain the sample set for the t-test. Subsequently, according to the t-test, $t_{0.025}(n-1) = t_{0.025}29 = 2.0452$ is obtained from the table for hypothesis testing (30 times at testlevel $a = 0.05$). Based on Eqn(2), $\overline{I} = 0.0288$ and $I_{0\max} = 0.0291$ was calculated. The threshold coefficient d_1 is determined as1.05 and δ_1 as 0.0306. After an initially screening, outlet phenol and BPA ofV304 and material flow of C303 are retained. Finally, taking $\delta_2 = 0.4$ ensures a higher requirement of redundancy in order to eliminate redundant variables, and outlet BPA and phenol of V304 are retained according to thejudgment formula of redundancy condition (5). The mean relative error of training and testing are got by the simulation of support vector machine model after deleting variables. Assuming the actual value is y and the estimated value is y', MRE is defined as:

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y - y'}{y}\right| \times 100\%$$

The simulation results are shown in Table 2.

The above table indicates that the most auxiliary variables lead to the highest training accuracy but the worst testing accuracy.After the first screening step, the auxiliary variables are outlet BPA of V304, outlet phenol of V304, and mass flow of C303, and the testing accuracy of the model improved. After the redundancy step, only outlet phenol of V304 and outlet BPA of V304 are retained, and the testing error is the least.Therefore, retaining more auxiliary variables increases modelcomplexity but cannot improve testing accuracy. Considering both relevanceand redundancy between auxiliary variables and the primary variable lowers the final testing error. An analysis of the process indicates that each component of C303 has been relatively stable; therefore, the concentration of phenol mainly depends on the components of V304. Applying our method of variable selection for soft sensor based on distributedmutual information led to retaining the outlet phenol of V304 and outlet BPA of V304, which in turn yielded more accurate estimation of phenol in C303. The test results are shown in Figure 1.

Table 2.Training and testing after deleting variables

Auxiliary variables in order to remove	The remaining number of auxiliary variables	Training （%）	Testing （%）
		MRE	MRE
Retain all variables	6	0.42	1.72
Variables after the first step	3	1.45	1.51
The rest of variables	2	1.28	1.33

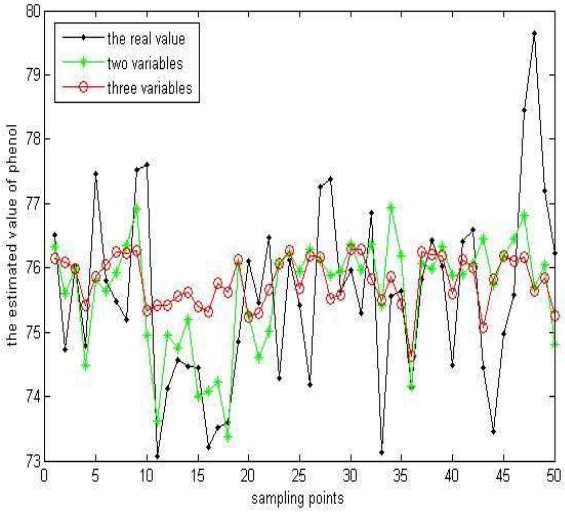

FIGURE 1. The comparison of test results

.

CONCLUSION

The method of variable selection for soft sensor based on distributed mutual information has some advantages, well reflecting linear or nonlinear relevance and redundancy between the primary variable and auxiliary variables. The simulation results show that this method effectively eliminates irrelevant and redundant variables, reduces model complexity, and improves the estimation accuracy and generalization capability of the model.

REFERENCES

[1] Fortuna, L., Graziani, S., &Xibilia, M. *IEEEInstrumentation & Measurement Magazine*,**2005**, 8(4), 26–34.
[2] Fortuna, L., Graziani, S., & Xibilia, M. G. *Control Engineering Practice*,**2005**, 13(4),499–508.
[3] Qin, S. J. *Computersand Chemical Engineering*,**1998**, 22(4–5), 503–514.
[4] Raennar, S., MacGregor, J. F., & Wold, S. *Chemometrics and Intelligent Laboratory Systems*,**1998**, 41(1), 73–81.
[5] Q. X. Zhu and N. Lang, *Control Engineering*, **2011**, 18(3): 0388-0392. (In Chinese）
[6] X. Z. Wang, *Control and Decision*, **2010**, 25(10): 1589-1592. (In Chinese)
[7] R. L. Liu, W. Q. Chen, H. Y. Su, *Journal of Nanjing University of Posts and Telecommunications*, **2006**, 26(1): 76-79. (In Chinese)
[8] C F Tsai, *Expert Systems with Application*, **2006**, 31(4): 808-825.
[9] B. Pang, X. Sun, Ch.W. Ye, K.J. Chen, *Computer Modelling and New Technologies*, **2013**, 17(3), 87–92
[10] L. Kang, G. Qian, *Computer Modelling and New Technologies*, **2013**, 17(3), 20–26
[11] A Sharma, *Journal of Hydrology*, **2000**, 239(1-4): 232～239.
[12] S G Brown, A new perspective for information theoretic feature selection. In: *Proceedings of the 12th*

International Conference on Artificial Intelligence and Statistics. Florida, USA: JMLR, **2009**.49−56 .
[13] H. Z. Yang, J. Zhang, H. F. Tao, *Control Engineering*, **2012**, 19(4): 0562-0565. (In Chinese)
[14] A B Barrett, L Barnett, A K Seth. *Physical Review E*, **2010**,81(4): 041907J
[15] H. C. Peng, F. H. Long, C. Ding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2005**, 27(8): 1226−1238