CHEMICAL ENGINEERING TRANSACTIONS
VOL. 61, 2017
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Petar S Varbanov, Rongxin Su, Hon Loong Lam, Xia Liu, Jiří J Klemeš

ISBN 978-88-95608-51-8; **ISSN** 2283-9216
DOI: 10.3303/CET1761288

Application of Variable Selection Method Based on Genetic Algorithm in Marine Enzyme Fermentation

Guohai Liu, Peisuo Yang, Yuhan Ding*, Congli Mei, Yonghong Huang, Xianglin Zhu

Jiangsu University, Zhenjiang, 212013, Jiangsu, China
yhding@ujs.edu.cn

Genetic algorithm（GA）, a global searching method, is applied to select the variables in soft sensing of marine enzyme fermentation. Compared with traditional methods of MIV and PCA, the optimal variables selected by GA have clear physical meaning and fewer numbers on the basis of variable selection frequency. After that, a soft sensing model based on BP neural network is established and a BP-GA soft sensing model is realized. Soft sensing results of enzyme activity show that, BP-GA model can provide better non-linear fitting ability and higher soft sensing accuracy, compared with the models of BP, BP-MIV and BP-PCA.

1. Introduction

Biological fermentation technology is one of the key technologies to strengthen the national power and promote the economy development. It plays an extremely important role in enriching food types, guaranteeing food safety and improving living quality. With the continuous expansion of the fermentation industry scale, the requirement of on-line measurement technology is increasing day by day. It is of great significance to realize the real-time measurement of key biochemical parameters of the fermentation processes. At present, some key biochemical parameters (such as enzyme activity in marine enzyme fermentation) are difficult to be measured on-line due to technical and economic limitations. The soft sensing technology is an effective way to solve the above problem (Chen et al., 2017).

Marine enzyme is a kind of typical microorganism fermentation, which is nonlinear, multi-variable and strong coupling (Huang et al., 2013). As is known to all, if the soft sensing model in marine enzyme fermentation process has massive input variables and these variables are redundant, it will make the establishment of the model need longer training time. And the redundant variables will also affect the soft sensing accuracy simultaneously. Therefore, to obtain a model with a good soft sensing ability, variable selection is extremely important. At present, the frequently used variable selection methods are based on sensitivity, MIV (mean impact value), PCA (principal component analysis) and so on. However, method based on sensitivity ignores the interaction between variables impacts the output result. It also lacks stability for sensitivity coefficients (Cai et al., 2008). Method based on MIV is seriously dependent on the network and the result will get worse if the network is not properly designed (Lu et al., 2011). The disadvantage of PCA method is that the meaning of each characteristics dimension of principal component is not as clear as that of original inputs, and non-principle components with small variance value may also contain important information because of sample differences, which may influent the follow-up data processing if they are discarded due to descending dimension (Llie et al., 2017).

Genetic algorithm (GA) is a computational optimization method by simulating the evolutionary process of Darwin's genetic selection and natural selection, which was proposed by Holland et al. (1975). GA has the following advantages: (1) It uses the selection, crossover, mutation and other operations to generate new individuals, so it can easily expand the search scope, which guarantees that the optimization results obtained by searching are global optimal solutions; (2) It is a kind of intelligent search algorithm, which take advantages of the fitness function to achieve the real value gradually; (3) It improves the speed of the search because a

Please cite this article as: Liu G., Yang P., Ding Y., Mei C., Huang Y., Zhu X., 2017, Application of variable selection method based on genetic algorithm in marine enzyme fermentation, Chemical Engineering Transactions, 61, 1741-1746 DOI:10.3303/CET1761288

set of individuals are calculated with iteration at the same time. Consequently, GA has been widely used in various fields (Pappu et al., 2017).

In this study, GA is used to select optimal variables to simplify the BP soft sensing model and reduce computational complexity. Compared with the traditional methods of MIV and PCA, GA method can obtain the least variable number and make the physical meaning of new variables clearer. The BP-GA model is then utilized to soft sensing the value of marine enzyme activity online. Simulation results confirm that the BP-GA model has the advantages of good stability and high soft sensing accuracy, compared with the BP, BP-MIV, and BP-PCA models.

2. Genetic algorithm and BP neural network

2.1 Genetic algorithm

Genetic algorithm (GA) is a kind of artificial intelligence optimization method with the function of highly nonlinear mapping, self-adaption and self-organizing for global optimization (Chu et al., 2001). It abstractly describes the process of evolution as 3 operators: selection, cross-over and mutation. Firstly, the GA encodes a solution vector of a problem to strings. Secondly, the fitness function is used to solve the fitness degree of each individual in the population. According to the survival of the fittest, individuals who adapt to the environment are selected and new offspring are generated through genetic operators (Xu et al., 2011). Finally, the most suitable individuals can be obtained, which is the optimal solution of the problem after several generations of evolution.

2.2 BP neural network

BP neural network is a kind of computing structure by simulating the structure and function of biological neuron system. Like neural systems, BP neural networks contain many nodes, which are distributed hierarchically and are not connected to each other in the same layer, but nodes between layers are interrelated. BP neural network modeling is a data-based approach that can approximate real models in the absence of knowledge of internal mechanisms and uncertainties. The model can be continually revised by learning.

The commonly expression of the BP neural network with single hidden layer is:

$$g(x,\theta)=\beta_o+\sum_{j=1}^{q}F(\omega_{j0}+\sum_{k=1}^{p}\omega_{ji}x_i)\beta_j \tag{1}$$

where the expression for x is $x=[x_1\ \ x_2\ ...\ x_p]^T$; p is the number of input variables (auxiliary variables); β_i $(i=0,1...,q)$ is the connection weight from the hidden layer to the output layer; $\omega_{ji}=(\omega_{j0},\omega_{j1}...,\omega_{jp})$ is the connection weight from the input layer to the hidden layer.

3. Experimental research

In this study, Pa040523 (Chi et al., 2006) strains are the object of study which are isolated and obtained from Bohai and the Yellow Sea. Pa040523 strain is a kind of marine bacteria producing marine enzyme. The living marine creature and its production is compatible with high salt, high pressure, low temperature, low light, low light, and so on. The cold-active enzymes produced by psychrophiles and psychrotrophs have the following 3 characteristics: low temperature and high catalytic efficiency; high structural flexibility; thermal instability. So, it is more advantageous than the middle temperature enzyme and high temperature enzyme in the application.

Marine microorganisms are very strict to the temperature of growth environment, only in a limited temperature, pH value, time, amount of inoculation, ventilation and other range of growth, which has the minimum, the most appropriate and the maximum critical value. Through the preliminary analysis of bacterial culture experiments, the optimum pH value, temperature, time, inoculation amount and aeration rate of the marine enzyme producing strain are 5.5, 12 °C, 72 h, 7 % and 170 mL. Under the optimum fermentation conditions, the enzyme activity in the fermentation tank of 50 L strain will reach the highest.

The fermentation tank should be sterilized before using. The time of sterilization is about 30 min where the sterilization temperature was set to 121 °C and sterilization pressure is set to 0.11 MPa. Through material feeding system, the right amount of dextrin, soybean oil, alcohol, ammonia and other nutrients are poured into the fermentation tank. The fermentation will last for about 72 h under the condition of 12 °C, pH value being about 5.5.

Different kinds of sensors, such temperature sensor, pH sensor, dissolved oxygen sensor, tank pressure sensor, air flow sensor, speed sensor, are connected to fermentation tank, which can dynamically monitor it in

real time.In a fermentation period, the activity of an enzyme was tested at intervals of 4 h. a sample of 5 min interval is obtained by polynomial difference, and formed the output sample. 4 batches of experimental data are collected, among which three batches are used as training samples, and the other one batch is used as the test samples.

4. Soft sensing of BP-GA model for marine enzyme fermentation

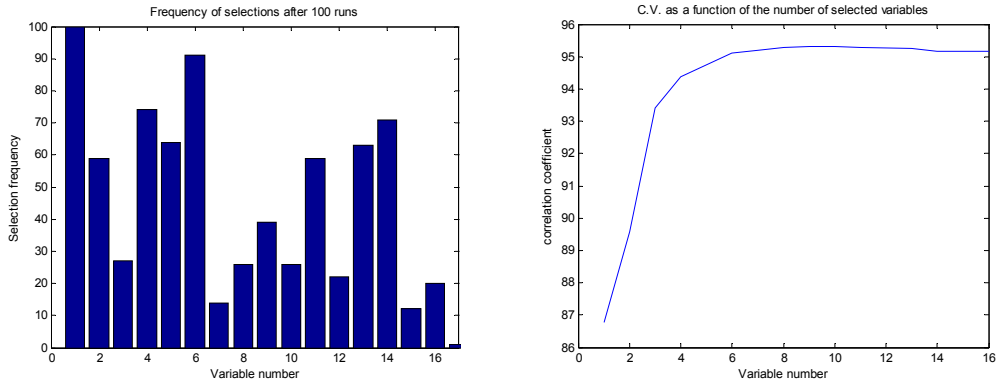
4.1 Variable optimization

The parameter settings of the GA are shown in Table 1.

Table 1: Parameters setting of GA

Main parameters	Maximum iterations	Population size	Crossover probability	Mutation probability
value	100	30	0.5	0.01

The variables with pretreated raw data are selected by GA. The variable select result is as shown in Figure 1.



(a) Variable selection frequency *(b) The trend of correlation coefficient with different variables*

Figure 1: The result of GA variable selection

Figure 1(a) is the probability map after 100 iterations and Figure 1(b) shows the change trend of the correlation coefficient with the number of selected variables. To make the variables easy to choose and to have a higher contribution rates, the variable selecting criteria a^* is set to 0.98. The cumulative contribution rates of the selected input variables satisfy $a_r = \sum_{i=1}^{s} \alpha_r \geq a^*$ $(s < n)$, where α_r are the relative contribution rates of the auxiliary variables to the dominant variable. The selected s variables can be considered important variables and the remaining $n - s$ variables are eliminated because they have less impact on the output. The comparison results of different methods are shown in Table 2.

Table 2: Comparison results of different variable selection methods

Variable selection methods	Original number	Optimized number	Time (s)
MIV	17	13	8.35
PCA	17	10	4.46
GA	17	9	9.57

As is shown in Table 2, GA method filters out the number of 9 variables, the least variable number, less than the number of MIV, 13, and PCA, 10. The time of selection variables of PCA is 4.46 s, time of MIV is 8.35 s and time of GA is 9.57 s. Although PCA is relatively small in time, it is acceptable for these three methods to be of the same temporal dimension. In general, GA variable selection method can simplify the model, lowered the computational complexity and reduce the possibility of introducing noise.

4.2 Soft sensing results of BP-GA model

In this study, a three layer feed-forward BP neural network with 9-15-1 structure is applied. The activating function of hidden layer called "Sigmoid" is $f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ and activating function of output layer is linear function $f(x)=x$. The network is trained 100 times with BP training algorithm until the error is reduced to a very small number 10^{-4}, and then the trained BP neural network parameters are preserved. The flowchart of the GA-BP method is as shown in Figure 2.

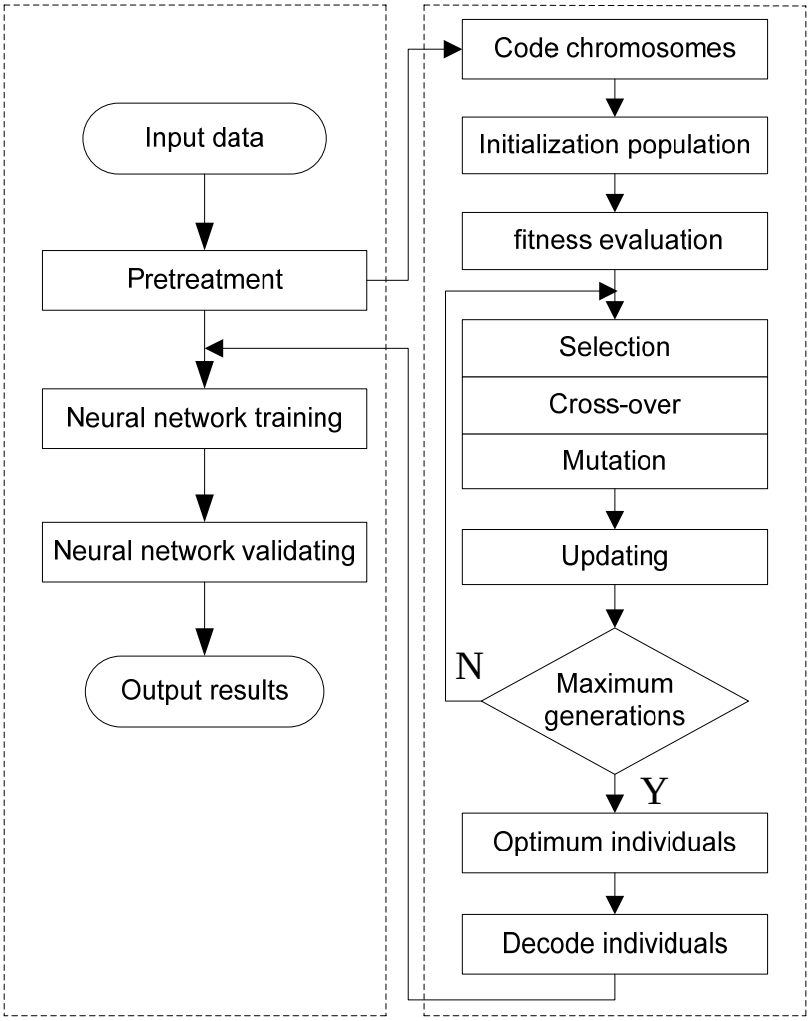


Figure 2: The flowchart of GA-BP method

To show the superiority of the method proposed in this study, BP-GA model is compared with other soft sensing models of BP, BP-MIV and BP-PCA. The soft sensing results are shown in Figure 3.
It can be clearly seen from Figure 3 that the fitness between the real values and soft sensing values is preferable, which shows that the soft sensing effect is satisfactory. Compared with the BP, BP-MIV and BP-PCA models, the soft sensing results of the BP-GA model are closer to the true values.
To compare the performance of soft sensing models, RMSE (root mean square error) is applied, which is calculated as Eq(2).

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(y_{ss,i}-y_{real,i})^{2}}{n}} \tag{2}$$

where $y_{ss,i}$ represents soft sensing values, $y_{real,i}$ is real values in the test data set, and n is the number of test data.

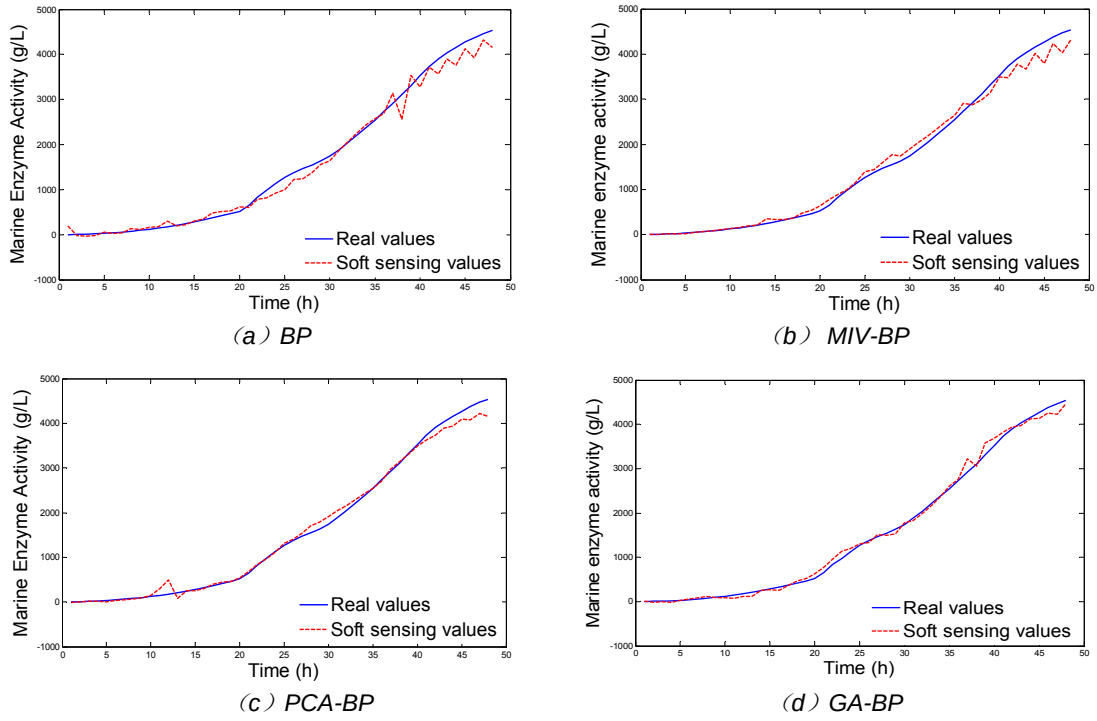


Figure 3: Comparison of different models with soft sensing results

The RMSE of the 4 soft sensing models are shown in Table 3.

Table 3: Soft sensing error of different models

Evaluation index (unit)	BP	BP-MIV	BP-PCA	BP-GA
RMSE (g/L)	177.3674	148.1806	124.1008	87.5861

From the Table 3, we can clearly find that the RMSE of BP model is the largest, which is 177.3674 g/L. The RMSE of BP-MIV model and BP-PCA model is relatively small, while BP-GA model has the smallest RMSE, which is only 84.4051 g/L. Therefore, the BP-GA soft sensing model proposed in this study can greatly improve the precision of soft sensing of enzyme activity and also have higher credibility.

Because GA is a stochastic method, the test set must be simulated several times by using the above BP-GA model. 5 times of the best soft sensing accuracy is recorded in Table 4.

Table 4 shows that the errors of these 5 times are almost same, whose average value is 87.5861 g/L and its optimal value is 83.4835 g/L. This illustrates that the stochasticity of GA method has little influence on the soft sensing results and the BP-GA model shows good stability.

Table 4: Soft sensing accuracy of different times of BP-GA

Times	1	2	3	4	5	Average value	Optimal value
RMSE (g/L)	89.0923	88.4651	83.4835	86.7842	90.1054	87.5861	83.4835

5. Conclusions

In marine enzyme fermentation process, the enzyme activity is related to many auxiliary variables and these variables are somehow related and redundant, so it is feasible to establish a soft sensing model and it is necessary to select variables and eliminate the redundancy as well. In this paper, a GA variable selecting method is proposed. Compared with traditional methods of MIV and PCA, GA selects the least number of variables and the selected variables have a clearer physical meaning. Then, a GA-BP soft sensing model is

established and applied to soft sensing the enzyme activity. The experimental results show that compared with traditional models of BP, BP-MIV and BP-PCA, the soft sensing result of BP-GA model has been greatly improved and the RMSE has been greatly reduced.

Acknowledgments

This work is supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD [2011]6), Open Research Foundation of Key Laboratory of Modern Agricultural Equipment and Technology in Jiangsu University(NZ201301), and Natural Science Foundation of Jiangsu Province of China (BK20130531, BK20151345).

References

Cai Y., Xing Y., Hu D., 2008, On sensitivity analysis, Journal of Beijing Normal University, 44(1), 9-16.

Chen X., Chen X., She J., Wu M., 2017, A hybrid just-in-time soft sensor for carbon efficiency of iron ore sintering process based on feature extraction of cross-sectional frames at discharge end, Journal of Process Control, 54, 14-24.

Chi N., Zhang Q., Wang X., Dou S., Zhang X., 2006, Study on fermentation conditions of a marine low temperature acid protease high-production strain from pseudomonas alcaligenes, Microbiology China, 33(2), 106-108.

Chu X., Yuan H., Wang Y., Lu W., 2001, Variable selection for partial least squares modeling by genetic algorithms, Chinese Journal of Analytical Chemistry, 29(4), 437-442.

Holland J., 1975, Adaptation in natural and artificial systems, Ann Arbor, MI, USA: University of Michigan Press.

Huang Y., Sun L., Sun Y., Zhu X., 2013, Soft sensor modeling based on biological variables of marine, Information & Control, 42(4), 506-510.

Llie A., Scarisoareanu M., Morjan L., Dutu E., Badiceanu M., Mihailescu L., 2017, Principal component analysis of raman spectra for TiO2 nanoparticle characterization, Applied Surface Science, 417, 93-103.

Lu Y., Wang W., 2011, Variable selection of financial distress prediction-the SVM method based on mean impact value, Systems Engineering, 29(8), 73-78.

Pappu S., Gummadi S., 2017, Artificial neural network and regression coupled genetic algorithm to optimize parameters for enhanced xylitol production by debaryomyces nepalensis in bioreactor, Biochemical Engineering Journal, 120, 136-145.

Xu G., 2011, Research for construction and application of PCA-GA-SVM model, Journal of Quantitative & Technical Economics, 2, 135-147.