CHEMICAL ENGINEERING TRANSACTIONS

VOL. 51, 2016

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Tichun Wang, Hongyang Zhang, Lei Tian
Copyright © 2016, AIDIC Servizi S.r.l.,
**ISBN** 978-88-95608-43-3; **ISSN** 2283-9216

DOI: 10.3303/CET1651224

# Vector Machines Regression Applied in Penicillin Fermentation Process Control

Lei Wang*[a], Qian Feng[b]

[a]Ministry of Organization, North China University of Science and Technology, Tangshan 063009, China
[b]College of Mechanical Engineering, North China University of Science and Technology, Tangshan 063009, China
anshiqi2008@163.com

According to the characteristics of biochemical processes, this paper studied the small samples of the data process. Small amount of data can be quickly measured variable modeling. Introduces the support vector machine (SVM) modeling as small sample theory, the least squares support vector machine (LS-SVM) algorithm based on improved SVM applied to the typical fermented penicillin biochemical processes in the past. Online prediction model only off-line testing of important process variables based on the simulation results simulation platform show that the method is only by learning the few batches of sample data. Establish a penicillin product concentration, cell concentration and substrate concentration. This paper analyzes the characteristics of the penicillin fermentation process, and the use of SVM method for modeling, give potency relationship between the factors and its influence. Through experiments, the influence of SVM model parameters to adjust performance. Through a variety of models from the established field data can be found, SVM is better than ANN modeling.

## 1. Introduction

Nowadays, with the advent development of bio-engineering era and plant cell culture technology, fermentation industry catches more and more attention by the scientific community (Yimam-Seid, 2003)). Technical and economic indicators of China fermentation industry are behind the international advanced level (Yu, 2012). Line detection of the fermentation process, real time control and optimization of the process flow plus is an effective way to improve the overall level of fermentation, but fermentation process of biomass key variable (cell concentration, product concentration and substrate concentration), due to biological limit level sensor technology development, has not been solved, leading to advanced control algorithms and optimization strategy can only stay in the theoretical study (Gu, 2015). In recent years, starting from the inferential control, and gradually formed a soft-sensing technology to solve this problem, the variables are online estimate (Yu, 2012). The basic idea of the soft measurement technology is important for those who are difficult to measure variables cannot be measured oath or transcript (called dominant variable), select a group of leading variables related measurable variables (called secondary auxiliary variables or variable), by construct a mathematical relationship to extrapolate and estimate the dominant variables in software instead of hardware (sensors) function (Yuan, 2014).

Existing models established fermentation process can be divided into three categories: white box model and gray box model, black-box model (Jia, 2012). White-box model to determine a priori knowledge of the structure and mechanism of the model, this model requires the mechanism of process of dynamic characteristics, transmission characteristics and biochemical characteristics have a better understanding (Yuan and Mori, 2014). It is generally from a smart balance, monod equation, arrhenius equation and so starting the establishment of kinetics-based, reflecting the opening mechanism model between biomass and measurable variables (Schmidberger and Ni, 2015). For white-box model, although a clear physical meaning of each parameter, gives the relationship between biomass and testability auxiliary variables, however, the soft measurement model intelligent balance, monod equation type (Campbel and Dom, 2003). Based on the established, since bacteria in the fermentation process growth on the theoretical treatment reaction is greatly simplified, resulting model cannot reflect the true nature of the reaction of microbial growth, poor adaptability:

Please cite this article as: Wang L., Feng Q., 2016, Vector machines regression applied in penicillin fermentation process control, Chemical Engineering Transactions, 51, 1339-1344 DOI:10.3303/CET1651224

microbial fermentation process and is now still lack sufficient understanding of many unknown microbial reactions can not direct modeling, thus establishing direct effective mechanism model fermentation process difficult.

Introduces the support vector machine (SVM) modeling small sample theory, the least squares support vector machine (LS-SVM) This fast algorithm based on improved SVM applied to the typical fermented penicillin biochemical processes in the past. Online prediction model only off-line testing of important process variables based on the simulation results simulation platform show that the method is only by learning the few batches of sample data, the establishment of a penicillin product concentration, cell concentration and substrate concentration . This paper analyzes the characteristics of the penicillin fermentation process, and the use of SVM method for modeling, give potency relationship between the factors and its influence. Through experiments, the influence of SVM model parameters to adjust performance. And through a variety of models from the established field data can be found, SVM is better than ANN modeling.

## 2. SVM regression estimation theory

### 2.1 Linear regression based on support vector machine

There n samples: $(x_1, y_1)$, $(x_2, y_2)$,…$(x_n, y_n)$, in it $y_i \in R$, $i=1,2,...,n$, A function $f = w\cdot x + b$ fitting n samples, And assume that all samples can be in accuracy ε (called insensitive loss function) without error fitting, namely:

$$\begin{cases} y_i - w\cdot x_i - b \le \varepsilon \\ w\cdot x_i + b - y_i \le \varepsilon \end{cases} \quad i=1,2,...,n \tag{1}$$

Taking into account the fitting error, by introducing a relaxation factor $\xi_i \ge 0, \xi_i^* \ge 0$ and $\xi_i \xi_i^* = 0$ relaxation of the constraints, then (1) is converted to the formula:

$$\begin{cases} y_i - w\cdot x_i - b \le \varepsilon + \xi_i \\ w\cdot x_i + b - y_i \le \varepsilon + \xi_i^* \end{cases} \quad i=1,2,...,n \tag{2}$$

Based on support vector machine linear regression method is to make it $\frac{1}{2}\|w\|^2 + C\cdot\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)^j$ minimize, where C is the penalty coefficient. Using dual theory, establish Lagrange equation. In conditions $\sum_{i=1}^{n}\left(a_i - a_i^*\right) = 0 \quad a_i \ge 0, a_i^* \le C, i=1,2,...,n$ by maximizing the objective function:

$$\begin{aligned} W\left(a,a^*\right) &= -\varepsilon\sum_{i=1}^{n}\left(a_i - a_i^*\right) + \\ &\sum_{i=1}^{n} y_i\left(a_i - a_i^*\right) + \frac{1}{2}\sum_{i,j=1}^{n}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right)\left(x_i x_j\right) \end{aligned} \tag{3}$$

On only a portion of a formula solution is not equal to zero, corresponding to the sample called support vectors. Corresponding regression function is:

$$\begin{aligned} f\left(x\right) &= w\cdot x + b \\ &= \sum_{i=1}^{n}\left(a_i - a_i^*\right)\left(x_i x_j\right) + b^* \end{aligned} \tag{4}$$

### 2.2 SVM regression estimation theory

The topology SVM support vector may decide to avoid the neural network topology requires trial and error experience limitations, is considered as the best theory for small sample classification and regression problems. Therefore, in pattern recognition, regression estimation, probability density function estimation, etc. are widely used and achieved good results. But its application in the industrial sector is relatively small, in the application of biochemical processes is even rarer.

First, consider the use of a linear function f(x)=<w, x>+b to fit the sample data set {$x_i$, $y_i$} (i=1,2,…,n; $x_i \in R^d$) and linear insensitive loss function, such as (1) the formula, where x is the independent variable, y is the dependent variable, ε Representative for controlling the fitting accuracy is not sensitive parameters.

$$\begin{aligned} & L\left[y, f(x,a)\right] \\ & = L\left[y - f(x,a)\right] \\ & = \begin{cases} 0 & \left|y - f(x,a)\right| \le \varepsilon \\ \left|y - f(x,a)\right| - \varepsilon & others \end{cases} \end{aligned} \tag{5}$$

SVM is to look at the formula (1) constraints best fit hyperplane that minimize 1 / 2‖w‖ so that all training data can be used in precision ε at fitting a linear function, where y is a scalar xi and w is d-dimensional vector. And the introduction of (non-negative) slack variables ξi, ξ * i and slack variables for controlling role in the objective function of (positive) penalty coefficient C, C representative of the larger error ε beyond the penalties stronger. SVM control function sets the complexity of the method is to return to function most flat, so regression modeling will be converted to the target as shown in functional constraint minimization problem in the formula.

$$\min \Phi\left(x, \xi_i, \xi_i^*\right) = \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)^j \tag{6}$$

Lagrange multiplier method for solving the quadratic programming problems with linear inequality constraints can be obtained dual optimization problem as represented by the formula:

$$\begin{aligned} & \max W\left(\alpha, \alpha^*\right) \\ & = -\varepsilon \sum_{i=1}^{n}\left(\alpha_i^* + \alpha_i\right) + \sum_{i=1}^{n} y_i\left(\alpha_i^* - \alpha_i\right) \\ & -\frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i^* + \alpha_i\right)\left(\alpha_j^* + \alpha_j\right)\left(\left\langle x_i, x_j\right\rangle\right) \end{aligned} \tag{7}$$

Regression function:

$$f(x) = \sum_{i=1}^{n}\left(\alpha_i^* + \alpha_i\right)K\left(x_i, x\right) + b \tag{8}$$

Common linear kernel function kernel function, polynomial kernel function, radial basis function (RBF), multilayer perceptron kernel of four, using a common kernel function of RBF paper form.

## 3. Experiments and results

### 3.1 Penicillin fermentation process modeling

Data using batch, each batch represents a complete fermentation process. The first batch of data as training data 185, divided into 55 samples. The first batch of 191 data as the test data, is divided into 54 samples. To reflect the fermentation process is nonlinear, radial basis function, so the main parameters of SVM insensitive coefficient ε, penalty coefficient C and the width of the coefficient σ. Using the mean relative error of each sample point $MER = \frac{1}{n}\sum_{i=1}^{n}\left(\left|y_i - \hat{y}_i\right| / y_i\right)$ Model training error and prediction error, n is the number of samples, yi are actual and estimated value. Model structure for the $y(t+\Delta t) = f\left(y(t), t, x_1(t), \ldots, x_8(t)\right)$ y (t) representative of the time potency, t representative of the fermentation time, x1-x8 representing other eight input model variables; y (t+Δt) representative of the model output, i.e., time t of potency.

In Matlab software, select the parameter σ = 10000, C = 80000, ε = 300, the results of the first batch of 185 training data in Figure 1, the number of support vectors is 29, the mean relative error of the training is 0.011299. The mean prediction relative error is 0.0092825. As can be seen from Figure 2, the effect is very good modeling, model SVM method established in the test data show higher performance on the training data.

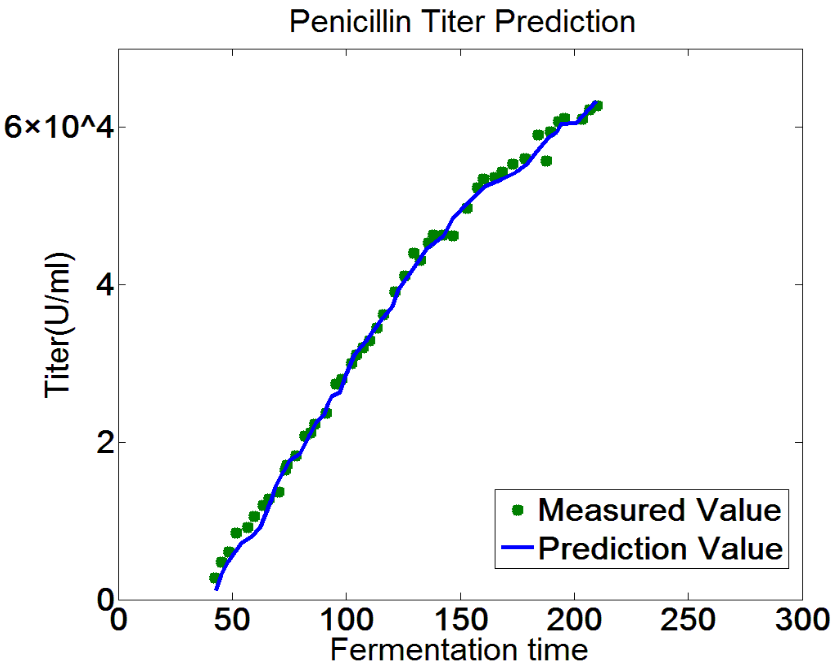

*Figure 1: Data estimated results*

### 3.2 Online prediction model based Pensim simulation platform

Penicillin fermentation process is more typical biochemical processes, due to the strong nonlinear fermentation process, when the variability and uncertainty, and the current lack of such product concentration, cell concentration and substrate concentration on-line detection equipment, a serious impact on the critical process variables the fermentation process of effective control and optimization, thus establishing online forecasting models for these key process variables has important practical significance. It should be noted that, although the article is a penicillin fermentation process, for example the typical biochemical research, but below the proposed method is fully applicable to the modeling of other biochemical processes.

In this paper, used for modeling data from Pensim simulation platform, the software kernel uses Bajpai mechanism model based on improved Birol model on this platform can achieve a series of simple simulation penicillin fermentation process, studies have indicated that the practical simulation platform and effectiveness. Typical non-structural model as Bajpai models, Birol model is mainly based on experimental data and other Pirt Bajpai model was improved. In this model, considering not only the pH value, temperature, air flow, stirring on power, acceleration rate bottoms stream and other control variables bacteria and penicillin production, and the cell growth, carbon dioxide, penicillin production, substrate consumption, the heat of reaction and other factors are also included in the comprehensive model to go, so they can comprehensively reflect the penicillin fermentation process.

To carry out the first batch i predict, before then i -1 sample batch by batch was sliced, and arranged in order according to the batch-moment view three-dimensional form of penicillin fermentation process data array, in Figure 2 as a number of training samples obtained prediction models and to predict the i-th batch.

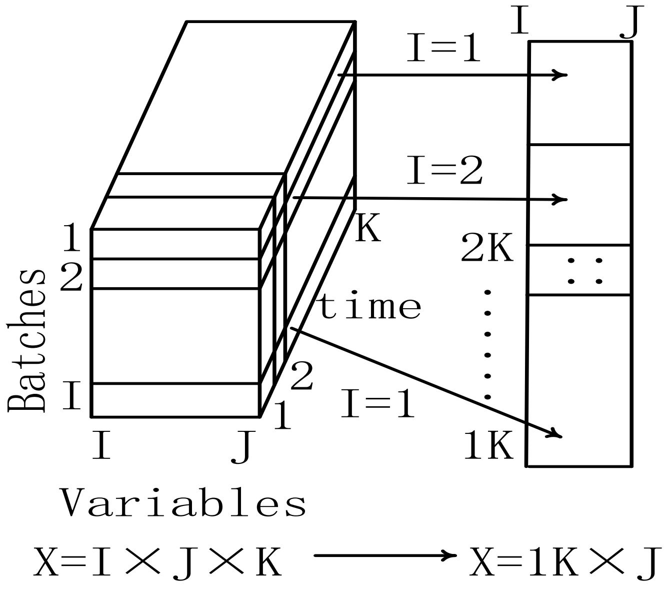

*Figure 2: Unfolding of batch process data by preserving variable direction*

### 3.3 Effect parameter adjustment of SVM modeling

Different penalty coefficient C, insensitive and width coefficient ε σ coefficient calculation training error and the prediction error, the error is calculated using the MER, as shown in Table 1.

*Table 1: The effect of parameters on Model quality*

| Penalty Coefficient | Insensitive Coefficient | Width Coefficient | Training Error | Prediction Error | Support Vector |
|---|---|---|---|---|---|
| 1e8 | 300 | 10000 | 0.0093124 | 0.018549 | 29 |
| 80000 | 1000 | 10000 | 0.021673 | 0.019959 | 7 |
| 80000 | 300 | 3e5 | 0.22905 | 0.24056 | 52 |
| 80000 | 300 | 50000 | 0.019477 | 0.018003 | 36 |
| 80000 | 300 | 10000 | 0.011299 | 0.0092825 | 29 |
| 80000 | 50 | 10000 | 0.010111 | 0.0097281 | 48 |
| 80000 | 20 | 10000 | 0.0097822 | 0.0099574 | 52 |
| 80000 | 200 | 8000 | 0.011886 | 0.0095004 | 32 |
| 80000 | 300 | 5000 | 0.0091384 | 0.012209 | 31 |
| 80000 | 300 | 2000 | 0.010919 | 0.038832 | 38 |
| 80000 | 300 | 1000 | 0.011552 | 0.11254 | 51 |
| 5000 | 300 | 10000 | 0.048549 | 0.049298 | 33 |
| 1000 | 500 | 10000 | 0.27273 | 0.28433 | 45 |

By analyzing the data in Table 1, C the greater the more stringent punishment for the error, which is fitting for high precision, making training difficult, very time-consuming. Thus, as C increases, the fitting error and the prediction error will be reduced when C increased to a certain extent, the fitting error stabilized, when C is too large, there will be "over-fitting" phenomenon, this time instead, the prediction error will increase.

## 4. Conclusion

Optimal control requires reliable, it can reflect the changes of the model parameters and their relationship in the process. Due to the lack of non-linear fermentation process, time-varying and biosensors, as well as among the parameters associated with severe, classical systems theory can hardly establish an appropriate model for the complex fermentation process. Online prediction model only off-line testing of important process variables based on the simulation results simulation platform show that the method is only by learning the few batches of sample data, the establishment of a penicillin product concentration, cell concentration and substrate concentration . This paper analyzes the characteristics of the penicillin fermentation process, and the use of SVM method for modeling, give potency relationship between the factors and its influence. Through experiments, the influence of SVM model parameters to adjust performance. And through a variety of models from the established field data can be found, SVM is better than ANN modeling.

### Acknowledgments

This work is supported by the Science and technology research Youth Fund Project of Hebei Province: Internal model control based on support vector machine in the process of biological fermentation(No. QN2016254).

### Reference

Campbel C.S., Maglio P.P., Cozzi A., Dom B., 2003, Expertise identification using email communications. In: The proceeding of the 12th International conference on information and knowledge management, New Orleans, LA, pp 528–531. doi:10.1145/956863.956965

Dom B., Eiron I., Cozzi A., Zhang Y., 2003, Graph-based ranking algorithms for email expertise analysis. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, New York, NY, ACM Press, pp 42–48. doi:10.1145/882082.882093

Gu B., Pan F., 2015, A Soft Sensor Modelling of Biomass Concentration during Fermentation using Accurate Incremental Online v-Support Vector Regression Learning Algorithm. American Journal of Biochemistry & Biotechnology, 11(3), 149.

Jia M., Xu H., Liu X., Wang N., 2012, The optimization of the kind and parameters of kernel function in KPCA for process monitoring, Computers & Chemical Engineering, 46, 94-104.

Mori J., Yu J., 2014, Quality relevant nonlinear batch process performance monitoring using a kernel based multiway non-Gaussian latent subspace projection approach. Journal of Process Control, 24(1), 57-71.

Ni C., Yan X., 2015, Elman Neural Networks with Sensitivity Pruning for Modeling Fed-Batch Fermentation Processes. Journal of chemical engineering of Japan, 48(3), 230-237.

Schmidberger T., Posch C., Sasse A., Gülch C., Huber R., 2015, Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance- based modeling. Biotechnology progress, 31(4), 1119-1127.

Yimam-Seid D, Kobsa A, 2003, Expert finding systems for organizations: problem and domain analysis and the DEMOIR approach. J Organ Comput Electron Commer 13:1–24. doi:10.1207/S15327744JOCE1301_1.

Yu J., 2012, A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses, Computers & Chemical Engineering, 41, 134-144.

Yu J., 2012, Multiway Gaussian mixture model based adaptive kernel partial least squares regression method for soft sensor estimation and reliable quality prediction of nonlinear multiphase batch processes. Industrial & Engineering Chemistry Research, 51(40), 13227-13237.

Yuan X., Ge Z., Zhang H., Song Z., Wang P., 2014, August, Soft Sensor for Multiphase and Multimode Processes Based on Gaussian Mixture Regression. In World Congress (Vol. 19, No. 1, pp. 1067-1072).

Yuan X., Ge Z., Song Z., 2014, Locally weighted kernel principal component regression model for soft sensing of nonlinear time-variant processes, Industrial & Engineering Chemistry Research, 53(35), 13736-13749.

Yuan X., Ge Z., Song Z., 2014, Soft sensor model development in multiphase/multimode processes based on Gaussian mixture regression, Chemometrics and Intelligent Laboratory Systems, 138, 97-109.