World Academy of Science, Engineering and Technology
International Journal of Agricultural and Biosystems Engineering
Vol:8, No:5, 2014

Bioprocess Optimization Based On Relevance Vector Regression Models and Evolutionary Programming Technique

R. Simutis, V. Galvanauskas, D. Levisauskas, J. Repsyte

Open Science Index, Agricultural and Biosystems Engineering Vol:8, No:5, 2014 publications.waset.org/9998131/pdf

***Abstract*—**This paper proposes a bioprocess optimization procedure based on Relevance Vector Regression models and evolutionary programming technique. Relevance Vector Regression scheme allows developing a compact and stable data-based process model avoiding time-consuming modeling expenses. The model building and process optimization procedure could be done in a half-automated way and repeated after every new cultivation run. The proposed technique was tested in a simulated mammalian cell cultivation process. The obtained results are promising and could be attractive for optimization of industrial bioprocesses.

***Keywords*—**Bioprocess optimization, Evolutionary programming, Relevance Vector Regression.

I. Introduction

IN the recent years, application of mathematical models for development and optimization of industrial biotechnological processes has attracted a lot of attention. Usually, in the first stages of process development only limited amount of experimental data is available. Consequently, the determination of rational process operation modes and optimization of the time profiles for manipulated process variables should be carried out using traditional mechanistic models [1]-[3] or more sophisticated hybrid models (combination of mechanistic models and nonlinear black-box models) [4]-[7]. Unfortunately, development of such models requires relatively high-skilled personnel and takes a lot of time. With the accumulation of experimental data, an application of pure data-based models and optimization procedures can become very attractive. In such cases, it is assumed that the company has accumulated sufficient amount of experimental data about the process and seeks to use it appropriately for improving of existing processes. Nowadays the data-based modeling and optimization procedures allow carrying out the process optimization tasks in a half-automated way. The diversity of data-based modeling techniques for modeling of bioprocess monitoring is very broad [8]. However, recently more and more attention is drawn by the process models based on Relevance Vector Regression (RVR) technique, where the parameters of regression model is estimated based on Bayesian inference [9], [10]. Relevance vector regression technique allows developing compact and stable data-based process model avoiding time-consuming modeling expenses. When a data-based model is created, the evolutionary programming optimization procedures could be applied to obtain the time profiles for manipulated process variables (e. g., substrate feeding rate, temperature, and pH profiles) during the cultivation process. This paper describes the essential steps for bioprocess optimization based on RVR models and evolutionary programming technique. The authors start with explaining the idea of using the RVR technique for creating of data-based model and with the application of evolutionary programming technique for process optimization. Then they present the application of proposed procedure for optimization of simulated mammalian cell cultivation process. Finally the authors discuss the efficiency of the proposed procedure and provide recommendations for application of this procedure for real cultivation processes.

R.Simutis,V. Galvanauskas, D. Levisauskas, and J. Repsyte are with Automation Department, Kaunas University of Technology, Kaunas, Lithuania (phone: +370-37-300261; e-mail: Rimvydas.Simutis@ktu.lt, Vytautas.Galvanauskas@ktu.lt,Donatas.Levisauskas@ktu.lt,Jolanta.Repsyte @ktu.lt).

II. Material and Methods

A. Relevance Vector Regression Technique

Relevance Vector Regression technique is a Bayesian sparse kernel technique for regression that uses Bayesian inference to obtain parameters of the regression model [9], [10]. When designing a data-based model for biotechnological process, the objective is to find an underlying functional model $y(\mathbf{x})$ that estimates output values, given input vector $\mathbf{x}$. Such model $y(\mathbf{x})$ could be of the following form:

$$y(\mathbf{x},\mathbf{w}) = \sum_{i=1}^{M} w_i \Phi_i(\mathbf{x}) = w^T \Phi(\mathbf{x}) \quad (1)$$

This model is linear in the parameters and has a number of analytic advantages. Nevertheless, by choosing the basis functions $\boldsymbol{\Phi_m(x)}$ to be nonlinear, $y(\mathbf{x},\mathbf{w})$ will be nonlinear too. In particular, the basis functions often are given by kernels, with one kernel associated with each of the data points from the training set. This type of model is very flexible, and if statistical complexity of the model is appropriately managed, it can be very effectively applied to build efficient data-based models for various bioprocesses. Usually the RVR models are employed to present a static correlation between input and output variables. If the input vector is extended by historical values of input and output variables, RVR model could be used to simulate the nonlinear dynamic of the complicated processes. It is important to note that RVR model is very

Open Science Index, Agricultural and Biosystems Engineering Vol:8, No:5, 2014 publications.waset.org/9998131/pdf

suitable for processes with high-dimension input vector. For such processes, the traditional nonlinear regression models and feed-forward artificial neural networks usually have poor generalization properties and could not be used for process optimization tasks. In this application, the authors used radial basis functions kernels for RVR model. Recently the sparse Bayesian methodology has been designed to estimate the regression parameters w_i very efficiently [10]. As a result, a Relevance Vector Regression model derived by this methodology comprises only few non-zero parameters w_i and the final process model incorporates a compact set of basis functions. For development of data-based RVR model, the authors used SparseBayes software package (MATLAB environment) [11] for the discussed application.

B. Evolutionary Programming Technique for Process Optimization

Mathematical models can describe a biochemical process only with limited accuracy. Hence, the choice of tools for optimizing time profiles of manipulated variables calculated from the model equations must be adapted to the accuracy of the models. In particular, the random search and evolutionary programming algorithms could give good possibilities for such optimization. The general open-loop optimization problem addressed here is to determine optimal time profiles of the critical control variables (e. g., substrate feed profile, pH profile, temperature profile) during the cultivation in such a way that a predefined objective function or performance index is maximized. In the proposed approach, optimal time profiles for manipulated control variables are generated using feed-forward artificial neural networks. The weights of the artificial neural networks are adapted using evolutionary programming technique [12]. The basic steps of this technique can be summarized as follows:

- Start the development of optimal time profiles for every manipulated variable by choosing a population of feed-forward artificial neural networks with random weights and generate the time profile for manipulated variables;
- Calculate the performance index for each member of the population using obtained time profile of manipulated variable and dynamic process model, which is developed using RVR technique. Keep the best half of the artificial neural network population and delete the rest.
- For each of the selected artificial neural networks, produce an offspring-ANN by mutation of weights;
- Use this new population of ANNs as the next starting point for optimization and repeat the optimization procedure again, until the solution (process performance index) improves or computation time reaches the allowed limit.

By searching for optimal time profiles, a separate ANN for every manipulated process variable should be used and the weights of all ANN's must be updated in a parallel way.

C. Bioprocess Model

For testing of the proposed modeling and optimization procedure, the authors developed a simplified model for fed-batch cultivation process of CHO mammalian cells, based on assumptions provided in [13]. The concentration of viable cells X_v in the bioreactor was modeled using the following simplified equation:

$$\frac{dX_v}{dt} = \mu X_v - \frac{F}{W} X_v$$

$$\frac{dW}{dt} = F \tag{2}$$

where μ is specific growth rate (1/h), W is reactor weight (kg), $F = F_{glc} + F_{gln}$ are glucose and glutamine feed rates (kg/h), respectively. It was assumed that glucose and glutamine do not accumulate significantly in cultivation medium and therefore specific growth rate of cells depends directly on glucose and glutamine feed rates. An insufficient supply of glucose and glutamine will limit the cell growth. On the other hand, an overflow of supplied substrates will trigger the forming of metabolic by-products, which can decrease the cell growth. These effects can be approximated using the following equation:

$$\mu = \mu_{opt} \frac{1}{1+t_1} \frac{1}{1+t_2} - k_d \,, \tag{3}$$

$$t_1 = s\left(\frac{S_{glc}F_{glc}}{WX_v} - k_{glc}\right)^2, t_2 = s\left(\frac{S_{gln}F_{gln}}{WX_v} - k_{gln}\right)^2$$

where μ_{opt} is specific growth rate value when growing conditions are optimal for product formation, k_{glc} and k_{gln} are glucose and glutamine specific feed rates, which provide optimal growing conditions for cells, S_{glc} and S_{gln} are glucose and glutamine concentrations in feeds, s –sensitivity factor for not adjusted feeds. As follows from (3), the optimal profiles for glucose and glutamine feeds during the cultivation can be estimated using equations:

$$F_{glc} = k_{glc} \frac{WX_v}{S_{glc}}, F_{gln} = k_{gln} \frac{WX_v}{S_{gln}}. \tag{4}$$

It is clear that the optimal time profiles of manipulated variables are not known *a priori* in real processes and they must be found using the proposed modeling and optimization technique. To imitate such modeling and optimization procedure, the designed simplified CHO cell growth model was used to generate 40 different data sets by varying the glucose and glutamine feed rates randomly around their optimal value (±30% variation range from the optimal value at every measurement point).The parameter values for simplified process model are given in Table I.

TABLE I
PARAMETER VALUES FOR SIMPLIFIED PROCESS MODEL

Parameter	Value
k_{glc}	0.02 (g/ h *10^9cells)
k_{gln}	0.003(g/(h *10^9cells)
μ_{opt}	0.024 (1/h)
k_d	0.004(1/h)
s	5000

Open Science Index, Agricultural and Biosystems Engineering Vol:8, No:5, 2014 publications.waset.org/9998131/pdf

Fed-batch CHO cell growth process was simulated within time range t=36-144 h., start cell concentration was X_v=0.7 (10^9cells/kg), and bioreactor start-weight was W=1.0 kg. Glucose concentration in the feed was S_{glc}= 15 g/kg, and glutamine concentration in the feed was S_{gln} =5 g/kg.

For development of RVR model, 30 data sets were used for model identification and 10 data sets for model testing. The measurement data (viable cell concentration, glucose and glutamine feed rates) were collected every 12 hour, cell concentration data were corrupted by the white Gaussian noise (Mean=0, STD=0.03X_v). Typical cultivation runs used for RVM model development are shown in Fig. 1.

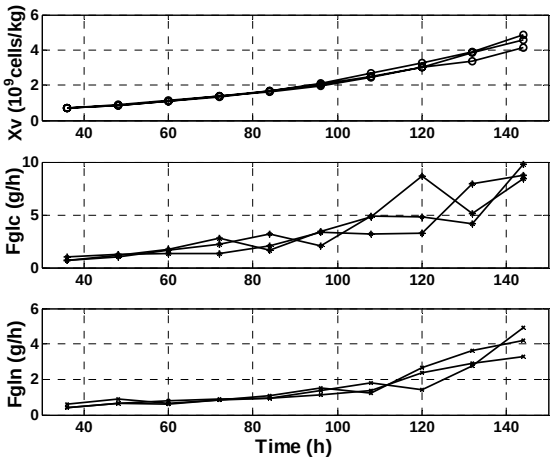


Fig. 1 Datasets from typical cultivation runs

III. Results and Discussion

RVR model was developed using the data from 30 cultivation runs. Input vector $\boldsymbol{x}$ included the delayed cell concentration data, delayed glucose and glutamine feed rates, total amount of used glucose $\sum glc$ and glutamine $\sum gln$, and cultivation time $t(k)$. The output of the RVR model y was the actual cell concentration $X_v(k)$.

$$\boldsymbol{x} = [X_v(k-2), X_v(k-1), F_{glc}(k-2), F_{glc}(k-1), F_{gln}(k-1), F_{gln}(k-2), \sum glc(k-1), \sum gln(k-1), t(k)],$$
$$y(\boldsymbol{x}) = X_v(k) \quad (5)$$

Radial basis functions kernels were used for the RVR model. The quality of the RVM model was evaluated by calculating the *MAPE* (mean absolute percentage error) between the simulated and real cell concentration values and by determination of correlation coefficient between them. The meta-parameters of RVR algorithm were adapted, based on model quality in the validation sets. Fig. 2 shows the correlation between predicted and real cell concentration values obtained for the validation sets.

The correlation coefficient for all validation sets was approximately R=0.99 and *MAPE* value for validation sets was: *MAPE*=4.3%. Thus, the model validation test showed relatively good modeling quality of RVR model. Fig. 3 shows typical simulation results for two validation sets.

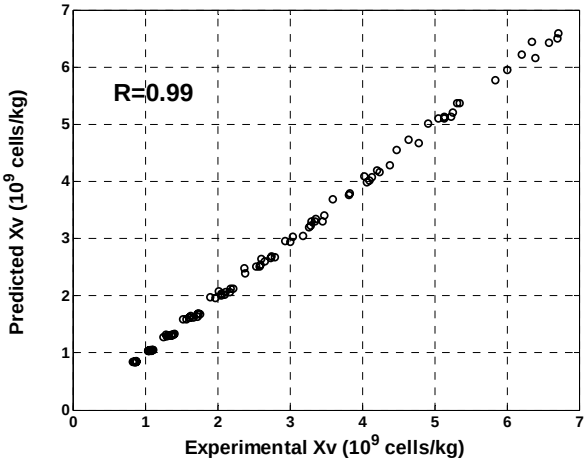


Fig. 2 Correlation between real and predicted cell concentration values

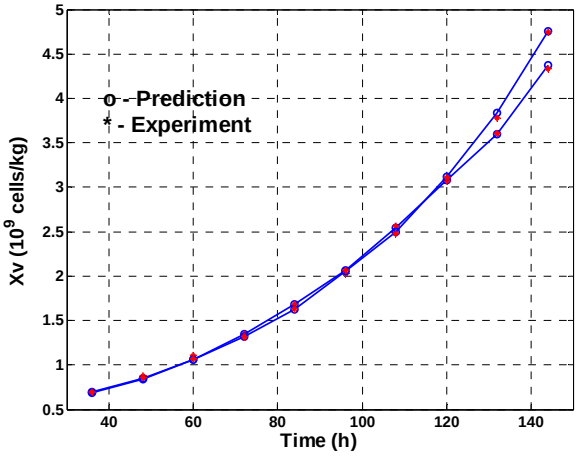


Fig. 3 Typical prediction results for two validation sets

Nevertheless, because the proposed model identification procedure is based on one-step prediction method, the real quality of the RVR model could be tested only by applying the developed model for process optimization task. For that purpose, 10 feed forward artificial neural networks (ANN) with randomly generated weights $\boldsymbol{w_1}$,$\boldsymbol{w_2}$were created to determine optimal glucose and glutamine feed profiles for analyzed process according the following equation:

$$F_{glc}(k) = ANN_{glc}(k, \boldsymbol{w}_1) + F_1(k),$$
$$F_{gln}(k) = ANN_{gln}(k, \boldsymbol{w}_2) + F_2(k) \quad (6)$$

where $ANN_{glc}(k,\boldsymbol{w_1})$ and $ANN_{gln}(k,\boldsymbol{w_2})$ glucose and glutamine feed rates at the interval k obtained from the ANN, $F_1(k)$and $F_2(k)$ are average glucose and glutamine feed rates at the interval k estimated from all data sets. The ANNs used in the study have one hidden layer with 5 hyperbolic tangent activation neurons and the output layer with linear activation functions. The single input variable for the ANNs was the cultivation time $t(k)$, and the output variable was glucose or glutamine feed rate. The above described evolutionary

programming procedure was used for tuning the ANNs' weights [w_1,w_2] to maximize the viable cell concentration at the end of cultivation.

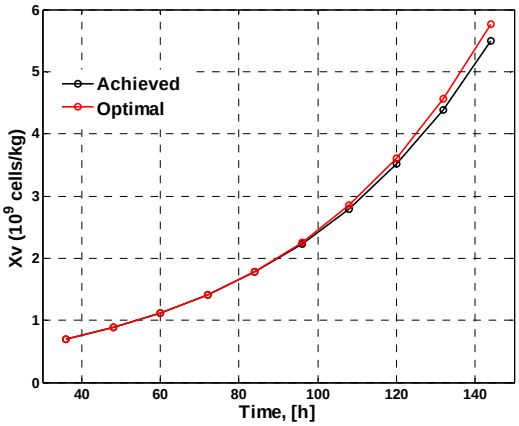


Fig. 4 Optimal vs achieved viable cell concentration

During the optimization, the objective function (viable cell concentration at the end of cultivation) achieves the value $Xv(k_f)$= 5.52(10^9cells/kg) for analysed case, which is close to the maximal cell concentration $X_v(k_f)$= 5.76(10^9cells/kg) when the glucose and glutamine feed rates are optimally controlled (estimated according to (4)) (Fig. 4). Also, the obtained glucose and feed rate profiles are close to the optimal ones (Fig. 5).

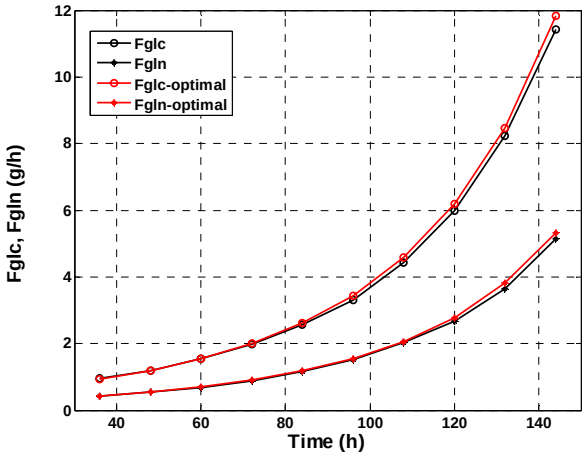


Fig. 5 The obtained and optimal feed rate profiles

Thus, the proposed data-based modeling and optimization procedure allowed achieving process performance, which was close to the optimal one. The modeling/optimization procedure hasn't used any a priori knowledge about the analyzed process and was based only on the provided experimental data.

IV. Conclusion

Data-based process modeling and optimization technique is attractive for improvement of already existing bioprocesses. The proposed technique includes the creation of a process model using Relevance Vector Regressions method and employing the obtained model for process optimization using evolutionary programming technique. The process modeling/optimization steps could be implemented in a half-automatic way and can be easily repeated after every new cultivation run. The proposed technique was tested in a simulated mammalian cell cultivation process for maximization of viable cell concentration at the end of cultivation. However, the same procedure could be applied for maximization/minimization of other objective functions (e. g., product concentration or amount of total product at the end of cultivation). The obtained results are promising and the proposed technique could be attractive for practical applications to improve already operating processes. Nevertheless, in order to apply this technique for real process optimization, a big amount of experimental data is required. Experimental data from approximately 30-50 cultivation runs could be a good basis to start applying the proposed technique.

Acknowledgment

This research was funded by a grant (No.MIP-056/2013) from the Research Council of Lithuania.

References

[1] J. H. Nielsen, J. Villadsen, G. Lidén, *Bioreaction Engineering Principles*, Kluwer Academic/Plenum Publishers, New York, 2003.
[2] B. Xu, M. Jahic, S.-O. Enfors, "Modeling of Overflow Metabolism in Batch and Fed-Batch Cultures of Escherichia coli," *Biotechnology Progress 15*, 1999, pp.81-90.
[3] B. Sonnleitner, "Measurement, monitoring, modelling and control," in*Basic Biotechnology*, 3rd ed., C. Ratledge, B. Kristiansen (eds). Cambridge University Press, 2006, pp. 251-270.
[4] M. L. Thompson, M. A. Kramer, "Modeling chemical processes using prior knowledge and neural networks," *AIChE Journal 40 (8)*, 1994, pp. 1328-1340.
[5] J. Schubert, R. Simutis, M. Dors, I. Havlik, A. Lübbert, "Bioprocess optimization and control: Application of hybrid modelling," *Journal of Biotechnology 35*, 1994, pp. 51-68.
[6] S. Gnoth, R. Simutis, A. Lübbert, "Selective expression of the soluble product fraction in Escherichia coli cultures employed in recombinant protein production processes," *Applied Microbiology and Biotechnology 87 (6)*, 2010, pp. 2047-2058.
[7] J. Peres, R. Oliveira, S. F. de Azevedo, "Bioprocess hybrid parametric/nonparametric modelling based on the concept of mixture of experts," *Biochemical Engineering Journal 39 (1)*, 2008, pp. 190-206.
[8] P. Kadlec, B. Gabrys, S. Strandt,"Data-driven soft sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, 2009, pp. 795–814.
[9] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, 2006.
[10] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research 1*, 2001, pp. 211–244.
[11] SparseBayes Version 2.0 software package for Matlab, 2013, http://www.relevancevector.com.
[12] D. J. Fogel, "Evolutionary Computation: toward a new philosophy of machine intelligence," *IEEE Press*, New York, 1995.
[13] M. Aehle, K. Bork, S. Schaepe, A. Kuprijanov, R. Horstkorte, R.Simutis, A. Lübbert, "Increasing batch-to-batch reproducibility of CHO-cell cultures using a model predictive control approach," *Cytotechnology*. Dordrecht : Springer. ISSN 0920-9069, 2012, Vol. 64, iss. 6, pp. 623-634.

Open Science Index, Agricultural and Biosystems Engineering Vol:8, No:5, 2014 publications.waset.org/9998131/pdf