CHEMICAL ENGINEERING TRANSACTIONS
VOL. 35, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
www.aidic.it/cet
Guest Editors: Petar Varbanov, Jiří Klemeš, Panos Seferlis, Athanasios I. Papadopoulos, Spyros Voutetakis
Copyright © 2013, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-26-6; **ISSN** 1974-9791
DOI: 10.3303/CET1335160

A SVM Gray-Box Model for a Solid Substrate Fermentation Process

Gonzalo Acuña[a,*], Jennifer González[a] , Millaray Curilem[b] , Francisco Cubillos[c],

[a]Universidad de Santiago de Chile, USACH, Departamento de Ingeniería Informática , Av. Ecuador 3659, Santiago, Chile.
[b]Universidad de la Frontera, UFRO, Departamento de Ingeniería Eléctrica, Av. Francisco Salazar 01145, Temuco, Chile.
[c]Universidad de Santiago de Chile, USACH, Departamento de Ingeniería Química , Av. Ecuador 3659, Santiago, Chile.
gonzalo.acuna@usach.cl

Gray-Box models (GBM) which combine a priori knowledge of a process –e.g. first principle equations- with a black-box modeling technique are useful when some parameters of the first-principle model –normally time-variant parameters like the specific kinetics of some bioprocesses- cannot be easily determined. In this case the black-box part of the GBM can be used to model the influence of input and state variables on the evolution of those parameters. The most commonly used black-box technique for GBM is Artificial Neural Networks (ANN). However Support Vector Machine (SVM) has shown its usefulness by improving over the performance of different supervised learning methods, either as classification models or as regression models. In this paper, a kind of SVM –the Least-Square Support Vector Machine (LS-SVM)- are used to develop a GBM for a solid-substrate fermentation (SSF) batch process, the growth of the filamentous fungus *Gibberella fujikuroi*. SSF are well known as low water consumption processes, therefore reducing liquid effluent treatment costs. They can also use agricultural wastes as substrates. Although these advantages lack of adequate models attempts to better exploit SSF processes at an industrial level. The aim of the present work is then to build a GBM to simultaneously estimate the specific growth kinetics and the specific production kinetics. Good results confirm that SVM can be effectively used for developing GBM for SSF processes.

1. Introduction

Given the increasing complexity of industrial processes, the construction of suitable models for control, optimization and monitoring is not an easy task, because it requires great knowledge of the process itself.
A good alternative is the use of data-driven models or black-box models. Neural networks are one of the preferred tools for building this kind of models. However, when a priori knowledge of the process is available gray-box models (GBM) arise as a very good modeling alternative. They combine a priori knowledge of a process –e.g. first principle equations- with a black-box modeling technique and are useful when some parameters of the first-principle model –normally time-variant parameters like the specific kinetics of some bioprocesses- cannot be easily determined. In this case the black-box part of the GBM can be used to model the influence of input and state variables on the evolution of those time-variant parameters.
The most commonly used black-box technique for GBM is Artificial Neural Networks (ANN). Despite a number of successful results achieved with ANN (Wu et al., 2012; Barrios et al., 2011) there still remain unsolved a number of key issues such as: difficulty of choosing the number of hidden nodes, the overfitting problem, the existence of local minima solution, poor generalization capabilities and so on.
Support Vector Machine (SVM) has shown its usefulness by improving over the performance of different supervised learning methods, either as classification models or as regression models (Curilem et al., 2011). The SVM has many advantages such as good generalization performance, fewer free parameters

Please cite this article as: Acuna G., Gonzalez J., Curilem M., Cubillos F., 2013, A svm gray-box model for a solid substrate fermentation process, Chemical Engineering Transactions, 35, 961-966 DOI:10.3303/CET1335160

to be adjusted and a convex optimization problem to be solved (non-existence of local minima solutions) (Scholkopf et al, 2000).
This paper makes a contribution by applying a kind of SVM –the Least-Square Support Vector Machine, LS-SVM (Suykens et al., 2002) - to develop a GBM for a solid-substrate fermentation (SSF) batch process, the growth of the filamentous fungus *Gibberella fujikuroi*. SSF are well known as low water consumption processes, therefore reducing liquid effluent treatment costs. They can also use agricultural wastes as substrates. However lack of adequate models attempts to better exploit SSF processes at an industrial level. The aim then is to build a SVM-GBM to simultaneously estimate the specific growth kinetics and the specific production kinetics of the SSF process while assuring a low error of the output variables of the model.

2. Materials and Methods

2.1 Data

Data used for the development of all models were generated by integrating the phenomenological model included in Section 3.
A 5 % amplitude noise was added in order to have more realistic conditions. Data were divided into a training set (700 points) a test set (300 points) and a validation set (300 points). The latter is not used during the training process and it is used to evaluate performance of all models.

2.2 Gray-box models

GBM are the combination of first principle or phenomenological models with a black-box technique. Depending on the way the information is shared between the model and the black-box the GBM can be classified as parallel or series model (Thompson and Kramer, 1994). In the series model (Figure 1), the one used in this paper, the black-box determines difficult-to-model or difficult-to-obtain phenomenological model parameter values which are used by the first principle model to obtain the process variables values. In the parallel configuration, the black-box learns how to compensate the errors produced by using only the phenomenological model. They were not conceived to obtain the unknown parameter values which are the aim of this work and this is why they were not considered.

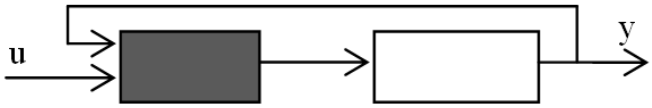

Figure1: Series GBM: The black box brings the parameters to the phenomenological model (white box) in order to obtain the outputs.

Concerning the GBM series configuration Acuña et al. (1999) distinguished between two methods of training. The first one corresponds to direct training (Figure 2(a)), which uses the error originated at the output of the black-box for training (Acuña and Pinto, 2006). The second method is indirect training (Figure 2(b)), which uses the error originated at the model's output for training purposes (Cruz and Acuña, 2010).

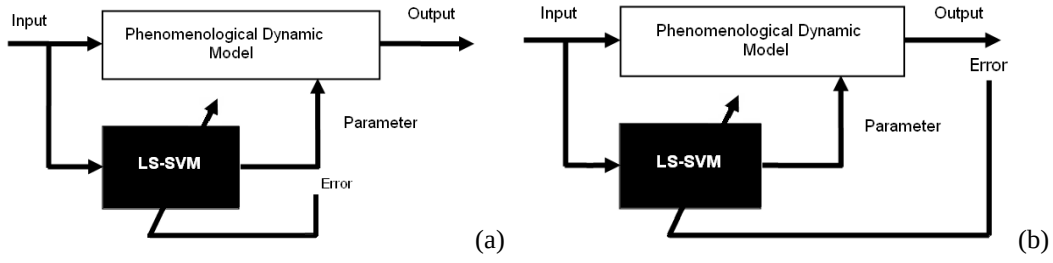

Figure 2: (a) GBM with direct training when the black-box is LS-SVM. (b) GBM with indirect training when the black-box is LS-SVM.

Most part of GBM applications used ANN as the black-box tool. In this paper an ANN a LS-SVM is chosen as the black-box instead of the ANN and the direct training method is used.

2.3 Least-Square Support Vector Machine (LS-SVM)

LS-SVM arises as an alternative to the SVM approach developed by Vapnik (1995). They are used for classification and regression. The difference with ε-insensitive SVM is that it solves a more simple linear system instead of a quadratic programming optimization problem, also requiring less training parameters. LS-SVM was proposed by Suykens et al. (2002).

Given a training set $\{x_i, y_i\}_{i=1}^{n}, x \in \Re^n, y \in \Re$ where n is the number of samples in the input space the SVM model for regression takes the form:

$$f(x) = w^T \varphi(x) + b \tag{1}$$

Where $\varphi(x)$, maps the input space to a feature space of higher dimension in case the problem to be solved is non-linear. The aim is to minimize a error loss function represented by the following quadratic equation:

$$L(x, y, f) = |y - f(x)|^2 \tag{2}$$

For LS-SVM the optimization problem can be stated as:

$$\min_{w,b,e} \; J(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{n} e_i^2$$
$$\text{subject to } \; y_i = w^T \varphi(x_i) + b + e_i, \quad \forall i = 1,...n \tag{3}$$

Where *J* is a loss function which depends on the weighting vector w and the value b, which are the parameters of the nonlinear approximator, $\gamma > 0$ is a regularization factor and e_i corresponds to the residuals for the i[th] sample ($e=\{e_1,e_2,\ldots, e_n\}$.

This optimization problem is solved using the Lagrange method. Partial derivatives are calculated with respect to all the variables and equated to zero, obtaining optimization conditions. After replacement of *w* and *e* and using the Kernel function yields a system of linear equations. This becomes the model being finally used for the LS-SVM:

$$y(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{4}$$

Where, α_i and b are the solution to the linear system, *y(x)* is the regression function and $K(x,x_i)$ is the kernel function that allows calculation of the dot product without requiring the explicit knowledge of $\varphi(x)$.

2.4 Error indices

The error indices used to evaluate the performance of the developed GBMs are those included in Table 1:

Table 1: Performance indices

Index of Agreement:	Root Mean Squared:	Residual Standard Deviation:
$IA = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(\lvert o_i' \rvert + \lvert p_i' \rvert)^2}$	$RMS = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i)^2}}$	$RSD = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{n}}$

with $\mathbf{o_i}$: Observed values or actual values at time i; $\mathbf{p_i}$: predicted values or estimated values over time i; $\mathbf{p_i'} = \mathbf{p_i} - \mathbf{o_m}$; $\mathbf{o_i'} = \mathbf{o_i} - \mathbf{o_m}$ .Where $\mathbf{o_m}$ corresponds to the mean value of the observed values to the total number n of data.

It is considered that RMS and RSD values lower than 0.1 are acceptable. In the case of AI acceptable values are above 0.9.

3. SSF Process

Growth of *Gibberella fujikuroi* and production of Giberelic Acid (GA_3) has been mainly done using submerged fermentation. However it has been demonstrated that solid state fermentation has many advantages over submerged fermentation, especially concerning lower energy consumption, greater yields, lesser environmental impact and so on (Rangaswamy, 2012). For Chilean agriculture production of

GA_3 is very important because the cost of GA_3 represents almost 30% of the cost of production of a kind of exportable grapes (Gelmi et al., 2002).
The mathematical model for the SSF process (Gelmi et al., 2002) consists of eight differential equations describing the evolution of: measured biomass concentration (Xmeasu), living biomass (X), urea (U), Intermediate Nitrogen (NI), Glucose (S), Gibberelic Acid production (GA), carbon dioxide production (CO_2) and oxygen consumption (O_2). μ and β correspond to the specific growth rate and the specific production rate respectively. Only for simulation purposes it is considered that parameters μ and β followed a Monod model as described in Gelmi et al. (2002). We will assume that these are the difficult to obtain time-variant parameters that we will aim to estimate.
The other model parameters different than μ and β were identified on the basis of some specific experiments and their values can be obtained from Gelmi et al. (2002) for controlled temperature and water activity conditions (T=25 ºC, Aw=0.992).
The series GBM for the SSF process is represented in Figure 3. The only assumption made concerning parameters μ and β is that they depend on Intermediate Nitrogen evolution.

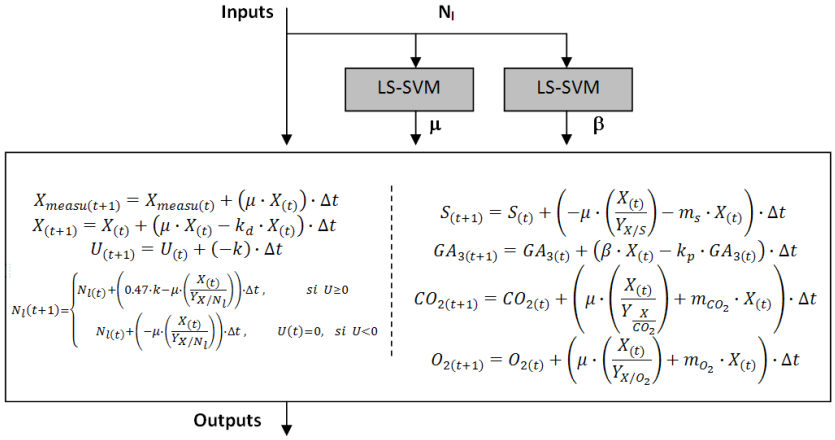

Figure 3: SSF GBM: the black-boxes bring the specific kinetics parameters μ and β to the phenomenological SSF model in order to get the outputs (CO_2 and O_2)

The initial values used for simulation of data are: Xmeasu(0)=0; X(0)=4e-3; U(0)=4e-3; NI(0)= 0.5e-4; S(0)=4e-3; GA3(0)=0; CO2(0)=0; O2(0)=0; and Δt=0.1.

4. Results and Discussion

LS-SVM Lab V. 1.7 Toolbox (Suykens et al., 2010) was used for the development of the black-box part of the GBM. A Radial Basis Function Kernel was utilized. The tuning parameters used were $\sigma^2 = 0.0001$ and a regularization factor $\gamma = 2^{30}$. These parameters were found after analysis of the error for the test set using the exhaustive method also included in the already mentioned Toolbox.
Model Predictive Output (MPO) consisting in multiple step ahead predictions of the state variables and hence of the estimated parameters only from their initial values was used in order to test the adequacy of the already developed models. Results shown in Figure 4 exhibit very good adequacy of the predicted outputs (O_2 and CO_2) compared to the noisy data of the validation set. This is very interesting because training of the black-boxes –the LS-SVMs in this case- when a direct training method is used consists in minimizing the error at the output of the black-boxes not the error at the output of the model.

Table 2: Performance índices for parameters and outputs for a MPO prediction for the validation data set

	μ	*β*	CO_2	O_2
IA	1.0	1.0	1.0	1.0
RMS	1.9e-2	4.7e-3	2.9e-2	3.0e-2
RSD	1.0e-3	2.0e-6	4.1e-3	1.9e-3

On the other hand it is not surprising that MPO predictions of the time-variant parameters μ and β for the validation data set follow the evolution of both parameters with very good accuracy as it can be seen in Figure 5 and in Table 1, also with with IA=1.0 and very low RMS and RSD indices.

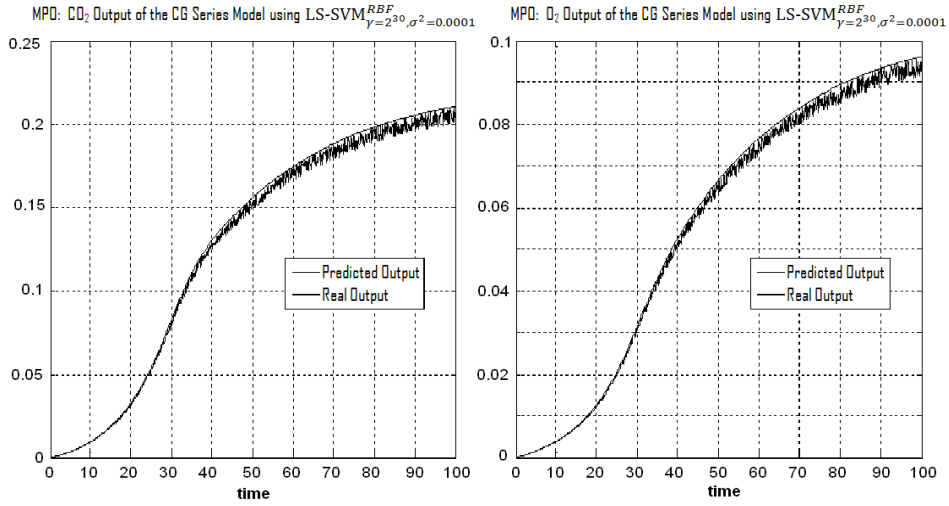

Figure 4: Simulated (Real) and predicted data for the output variables CO_2 and O_2

Performance indices of Table 2 only confirms the exhibited results of Figure 4 with IA=1.0 and very low RMS and RSD indices.

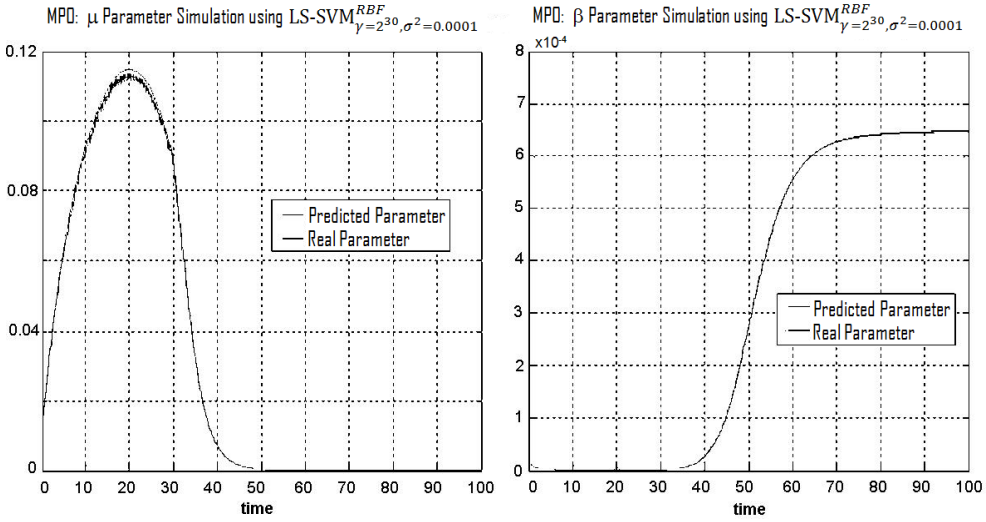

Figure 5: Simulated (Real) and predicted data for the time-variant parameters μ and β

In fact future research work will be focused in developing an indirect method to develop the GBM. This is because using the output error for training allows not working with the supposed unknown parameters values. Then only minimizing the output error of the GBM will bring the correct parameter values.

5. Conclusions

Grey-Box Models constitute a good alternative for those real world processes for which the available a priori knowledge is incomplete, for example in a variety of industrial processes. Indeed they can be used, as it is the case of this work, as time-variant parameters estimators, like the specifics kinetics parameters μ and β. Regarding the good results obtained LS-SVM proved to be appropriate tools to be used as the black-box part of the GBM. Future research work will be focused in using an indirect method to train GBM

with LS-SVM thus allowing finding the time-variant parameters evolution by minimizing the error of the process ouputs which are normally the only variables for which real-time measurements are available.

Acknowledgments

This work was partially supported by DICYT-USACH Grant 061219AL and the Dirección de Investigación, Universidad de La Frontera.

References

Acuña G., F. Cubillos, J. Thibault, E. Latrille, 1999, Comparison of methods for training grey-box neural models, Computers. Chem Engng., 23:S561-S564.

Acuña, G, Pinto, E., 2006, Development of a Matlab Toolbox for the design of grey-box neural models, International Journal of Computers, Communications and Control, IJCCC, I, 7-14 .

Barrios, J.A., Torres-Alvarado, M., Cavazos, A., Leduc, L., 2011, Neural and Neural Gray-Box modeling for entry temperature prediction in a hot strip mill, Journal of Materials Engineering and Performance, 20, 1128-1139.

Cruz, F., Acuña, G., 2010, Indirect training with error backpropagation in Gray-Box Neural Model: application to a chemical process, IEEE Proceedings of the XXIX International Conference of the Chilean Computer Science Society, 265-269, DOI 10.1109/SCCC.2010.

Curilem, M., Acuña, G., Cubillos, F. and Vhymeister, E., 2011, Neural networks and support vector machine models applied to energy consumption optimization in semiautogenous grinding, Chemical Engineering Transactions, 25: 761-766, DOI: 10.3303/CET1125127

Gelmi, C., Perez-Correa, R., Agosin, E., 2002, Modelling *Gibberella fujikuroi* growth and GA_3 production on solid-state fermentation, Process Biochemistry, 37, 1033-1040.

Rangaswamy, V., 2012, Improved production of Gibberelic Acid by *Fusarium moniliforme*, Journal of Microbiology Research, 2, 51-55.

Suykens, J.A.K, Van Gestel, T., De Brabanter., De Moor, B., Vandewalle, J, 2002, Least Square Support Vector Machines, World Scientific Pub. Co., Singapore, ISBN 981-238-151-1

Suykens J., De Brabanter K., Karsmakers P., Ojeda F., Alzate C., De Brabanter J., Pelckmans K., De Moor B., Vandewalle J., 2010, LS-SVM Lab Toolbox User´s guide, version1.7, available in www.esat.kuleuven.be/sista/lssvmlab

Schölkopf, B., Smola, A., Williamson, R.C., Bartlett, P.L., 2000, New support vector algorithms, Neural Computation, 12, 1207-1245.

Thompson M. L., Kramer M. A., 1994, Modeling Chemical Processes Using Prior Knowledge and Neural Networks, AIChE Journal, 48, 1328-1340.

Vapnik, V. 1995, The Nature of Statistical Learning Theory, Springer Verlag.

Wu, H., Cao, L., Wang, J., 2012, Gray-box modeling and control of polymer weight distribution using orthogonal polynomial neural networks, Journal of Process Control, 22, 1624-1636.