Biochemical Engineering Journal 85 (2014) 89–100

Contents lists available at ScienceDirect

Biochemical Engineering Journal

journal homepage: www.elsevier.com/locate/bej

Regular Article

Soft-sensor development for biochemical systems using genetic programming

Suraj Sharma, Sanjeev S. Tambe*

Artificial Intelligence Systems Group, Chemical Engineering and Process Development Division, National Chemical Laboratory, Dr. Homi Bhabha Road, Pune 411008, India

ARTICLE INFO

Article history:
Received 20 August 2013
Received in revised form 5 February 2014
Accepted 6 February 2014
Available online 15 February 2014

Keywords:
Artificial intelligence
Batch processing
Bioprocess monitoring
Lipase
Modeling
Sensors

ABSTRACT

Soft-sensors are software based process monitoring systems/models. In real-time they estimate those process variables, which are difficult to measure online or whose measurement by analytical procedures is tedious and time-consuming. In this study, the *genetic programming* (GP), an artificial intelligence based data-driven modeling formalism, has been introduced for the development of soft-sensors for biochemical processes. The novelty of the GP is that given example input–output data, it searches and optimizes both the form (structure) and parameters of an appropriate linear/nonlinear data-fitting model. In this study, GP-based soft-sensors have been developed for two bioprocesses, namely extracellular production of lipase enzyme and bacterial production of poly(3-hydroxybutyrate-*co*-3-hydroxyvalerate) copolymer. While in case study-I, the soft-sensor predicts the time-dependent *lipase activity* (U/ml), in case study-II it predicts the amount of *accumulated polyhydroxyalkanoates* (% dcw). The prediction and generalization performance of the GP-based soft-sensors was compared with the corresponding multi-layer perceptron (MLP) neural network and support vector regression (SVR) based soft-sensors. This comparison indicates that in the first case study the GP-based soft-sensor with the training and test set correlation coefficient (root-mean-squared-error) magnitudes of >0.96 (≈0.962 U/ml) has clearly outperformed the two other soft-sensors. In case study-II involving bacterial copolymer production, the GP and SVR based soft-sensors have performed equally well (correlation coefficient ≈ 0.98) while the MLP based soft-sensor's performance was relatively inferior (correlation coefficient ≈ 0.94).

1. Introduction

In today's industrial world, various types of sensors are needed to provide speedy and reliable measurements of a wide variety of process and/or product related variables and parameters. Sensors are sophisticated devices used in detecting and producing a measurable response to a change in, for example, chemical, physical, electrical, biochemical or optical state of a system. These measurements assist process operators and engineers in: (i) knowing the current state of the process, (ii) controlling and monitoring of the process, (iii) detecting and diagnosing any abnormal process behavior timely, (iv) taking corrective actions in the event of an abnormal process behavior, and (v) optimizing the performance of the process with a view to minimize costs and/or improve its efficiency. In many instances, however, an appropriate hardware-based robust sensor for measuring a process variable is either unavailable or the alternative analytical procedure is time-consuming and tedious. In such situations, the alternative of developing a soft-sensor should be explored. A soft-sensor is a software module capable of estimating a process variable in real-time. This module comprises a mathematical model, which makes use of the available quantitative knowledge regarding other process variables and parameters to estimate the magnitude of the chosen variable. The available information pertaining to other variables could be in the form of sensor measurements and/or mathematical models.

The challenges involved in the soft-sensor development for biochemical processes are the same as encountered in their modeling, optimization and control; the notable ones are as given below.

- Bioprocesses are characterized by their complex dynamics, such as inverse response, dead time and strong nonlinearities. These stem primarily from their main driving force, namely, microorganisms (cells), which are very sensitive to any variations in

Abbreviations: AI, artificial intelligence; ANNs, artificial neural networks; *R*, correlation coefficient; R^2, variance; GP, genetic programming; MLP, multilayer perceptron; ML, machine learning; RMSE, root-mean-squared-error; SVR, support vector regression.

* Corresponding author. Tel.: +91 20 25902156; fax: +91 20 25893041.
E-mail address: ss.tambe@ncl.res.in (S.S. Tambe).

http://dx.doi.org/10.1016/j.bej.2014.02.007
1369-703X/© 2014 Elsevier B.V. All rights reserved.

the reaction environment (e.g., temperature, substrate concentration, pH, among others) [1].

- An important class of bioprocesses, i.e., batch fermentation, commonly evolves through three stages, namely *lag*, *exponential* and *stationary* stages. The factors that influence the behavior of micro-organisms vary in each stage, owing to which the batch fermentation system exhibits different nonlinear characteristics in different stages. As a result, a global soft-sensor model for batch fermentation leads to complicated structure with limited prediction accuracy [2].
- Crucial biochemical variables and/or parameters are hard to measure online in bioprocesses such as batch fermentation.
- In bioprocesses involving induced cultures, there exists a variation in the morphology, energy metabolism and macroscopic composition of the cells; hence quantification of "biomass" or similar variables is not straightforward [3].

Since 1970–1980s the cost of computer-based instrumentation lowered significantly, and the concept of soft-sensor gained ground in the process estimation and inferential controls [4], bioprocess monitoring [5,6], control of nonlinear bioprocesses [7], biological wastewater treatment [8], melt index prediction [9], etc. For developing a soft-sensor, two principal approaches are *phenomenological* and *empirical* modeling. The former approach is employed when the detailed knowledge about the physico-chemical phenomena (kinetics, mass transfer, thermodynamics, etc.) underlying the process is available. Very often, gaining this knowledge itself becomes a tedious and costly task owing to the complex nature of the process and the extensive experimentation involved in collecting the necessary data. These difficulties make the phenomenological modeling route to soft-sensor development impractical. In such a situation, empirical modeling can be resorted to for the development of a soft-sensor.

There exist three commonly utilized methodologies for developing empirical models, namely regression analysis, artificial neural networks (ANNs) and support vector regression (SVR). For a pre-specified data-fitting function, the linear/nonlinear regression estimates the magnitudes of the function parameters that fit the given input–output data. Since many chemical and biochemical processes exhibit nonlinear behavior choosing an appropriate data-fitting function from a large number of possible alternatives becomes a daunting task. Despite expending a huge effort in guessing and testing different nonlinear data-fitting functions, there is no guarantee that a well-fitting function can indeed be secured in a finite number of trials. The other two empirical modeling formalisms, viz. ANNs and SVR, overcome the difficulties associated with the regression analysis since they do not require specification of the exact form of the data fitting function. Accordingly, ANN and SVR formalisms have been exploited in the development of soft-sensors and related applications including control of a distillation process [10], fed-batch reactor operation [11], and hybrid modeling of fermentation process [12]. Although these are potent nonlinear function approximation methods with a wide applicability, the ANNs and to some extent the SVR generate "black box" models whose structure and parameters do not provide any insight into the phenomena underlying the process being modeled.

In the present study, an artificial intelligence (AI) based exclusively data-driven modeling paradigm known as *genetic programming* (GP) [13] has been proposed for developing soft-sensor models for biochemical processes. Given multiple input–single output (MISO) data, the novelty of the GP formalism lies in its ability to search and optimize the form as also parameters of an appropriate linear/nonlinear data-fitting function. Despite its novelty and attractive properties, the GP formalism has not been explored widely for data-driven modeling applications in chemical and biochemical sciences/engineering to the same extent as other exclusively data-driven modeling methods, namely ANNs and SVR. In one of the soft-sensor applications involving the GP formalism, Kordon et al. [14] developed a soft-sensor for the emission estimation in one of the Dow Chemical Company plants in Freeport, TX. In this study, the soft-sensor was developed by integrating three computational intelligence approaches, namely, GP, analytical neural networks, and support vector machines. A rigorous literature survey indicates that the present study is the first one, wherein the GP formalism has been utilized for the development of soft-sensors for biochemical processes. The efficacy of the GP-based soft-sensors for biochemical processes has been demonstrated by conducting two case studies involving microorganism assisted extracellular production of lipase and production of bacterial poly(3-hydroxybutyrate-*co*-3-hydroxyvalerate) copolymer. In these case studies, multiple input–single output (MISO) example data sets have been utilized in searching and optimizing the functional form (structure) as also parameters of the MISO data-fitting functions (soft-sensors). While in the first case study, the soft-sensor predicts the time-dependent *lipase activity* (U/ml), in the second case study it predicts the amount of *accumulated polyhydroxyalkanoates* (% dcw). The prediction accuracy and generalization capability of the GP-based soft-sensors have been compared with those developed using the ANN and SVR formalisms.

This paper is structured as follows. Section 2 provides a detailed description of the GP formalism and its implementation. The commonly used feed-forward artificial neural network, namely *multilayer perceptron* (MLP) and the machine learning based SVR formalism have been described in sections three and four, respectively. The two case studies wherein the GP-based soft-sensor models have been developed for two biochemical systems are presented in Section 5. This section also provides results of the comparison of the GP, MLP and SVR based soft-sensor models pertaining to the two biochemical systems. Finally, Section 6 summarizes the principal findings of the study.

2. Genetic programming (GP)

In its original form, the GP formalism was proposed as a method for automatically generating computer programs that perform predefined tasks [13]. It is an extension of the *Genetic algorithm* (GA) formalism [15]. Given an objective function, the GA efficiently searches and optimizes the values of the decision variables that would maximize or minimize the function. Similar to the GA, the GP is founded on the Darwinian principles of *natural selection* and *reproduction*. Accordingly, the GP implementation uses simplified analogs of the naturally occurring genetic operations namely, *crossover* and *mutation*. There exist a number of schemes for implementing the genetic programming methodology, such as the *tree-structured* GP, *linear* GP, *gene expression programming*, *multi expression programming*, *grammatical evolution*, *Cartesian* GP and *stack-based* GP. Among these the tree-structured GP forms the most commonly employed GP-implementation.

Apart from automatically generating computer codes that execute pre-specified tasks, the GP can also be used for an attractive and novel application known as *symbolic regression*. Given an example data set comprising independent (predictor/causal) and dependent (response) variables, the GP-based symbolic regression is capable of searching and optimizing the form (structure) and associated parameters of a suitable linear/nonlinear mathematical model that fits the data—or at least an approximation to these. A notable feature of the GP-based symbolic regression is that unlike ANNs and SVR, it makes no assumptions about the form of the data-fitting function. The unique benefits of the symbolic regression include a human insight into and interpretability of the obtained models, identification of the key variables and