Applied Soft Computing 12 (2012) 3237–3245

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Genetic fuzzy system for data-driven soft sensors design

Jérôme Mendes, Francisco Souza, Rui Araújo*, Nuno Gonçalves

Institute for Systems and Robotics (ISR-UC), Department of Electrical and Computer Engineering (DEEC-UC), University of Coimbra, Pólo II, PT-3030-290 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 12 November 2010
Received in revised form 9 April 2011
Accepted 20 May 2012
Available online 6 June 2012

Keywords:
Soft sensors
Takagi–Sugeno fuzzy model
Genetic algorithms
Hierarchical genetic fuzzy systems

ABSTRACT

This paper proposes a new method for soft sensors (SS) design for industrial applications based on a Takagi–Sugeno (T–S) fuzzy model. The learning of the T–S model is performed from input/output data to approximate unknown nonlinear processes by a coevolationary genetic algorithm (GA). The proposed method is an automatic tool for SS design since it does not require any prior knowledge concerning the structure (e.g. the number of rules) and the database (e.g. antecedent fuzzy sets) of the T–S fuzzy model, and concerning the selection of the adequate input variables and their respective time delays for the prediction setting. The GA approach is composed by five hierarchical levels and has the global goal of maximizing the prediction accuracy. The first level consists in the selection of the set of input variables and respective delays for the T–S fuzzy model. The second level considers the encoding of the membership functions. The individual rules are defined at the third level, the population of the set of rules is treated in fourth level, and a population of fuzzy systems is handled at the fifth level. To validate and demonstrate the performance and effectiveness of the proposed algorithm, it is applied on two prediction problems. The first is the Box–Jenkins benchmark problem, and the second is the estimation of the flour concentration in the effluent of a real-world wastewater treatment system. Simulation results are presented showing that the developed evolving T–S fuzzy model can identify the nonlinear systems satisfactorily with appropriate input variables and delay selection and a reasonable number of rules. The proposed methodology is able to design all the parts of the T–S fuzzy prediction model. Moreover, presented comparison results indicate that the proposed method outperforms other previously proposed methods for the design of prediction models, including methods previously proposed for the design of T–S models.

1. Introduction

Data-driven soft sensors (DDSS) are inferential models that use on-line available sensor measures for on-line estimation of variables which cannot be automatically measured at all, or can only be measured at high cost, sporadically, or with large time delays (e.g. laboratory analysis). These models are based on measurements which are recorded and provided as historical data. The models themselves are empirical predictive models. They are valuable tools to many industrial applications such as refineries, pulp and paper mills, wastewater treatment systems, just to give a few examples [1].

The development of DDSS can be divided into four main stages: (I) data collection, and selection of historical data; (II) data preprocessing; (III) model selection, training and validation; and (IV) soft sensor maintenance. In the first stage, the data for training and evaluation of the model is selected. The usual steps in preprocessing are the handling of missing data, outliers detection, selection of relevant variables (i.e. feature selection), detecting the delays between the particular variables, and handling of drifting data [2]. One problem in the preprocessing step is that it requires a large amount of manual work and expert knowledge about the underlying process. Next, the model selection, training and validation phase is one of most important in soft sensors development, requiring the correct selection of the model, so that it can correctly reproduce the target variable. The last step is SS maintenance, where the goal is to maintain good SS response even in the presence of process variations, or some data change.

DDSS are built based on empirical observations of the process. Fuzzy models are useful to model systems when they cannot be defined in precise mathematical terms from physical laws. Fuzzy modeling of systems has been worked in many scientific researches. Takagi and Sugeno (T–S) have proposed a search algorithm for a fuzzy controller and generalized their research to fuzzy identification [3]. T–S fuzzy models are suitable to model a large class of nonlinear systems and have gained much popularity because of their rule consequent structure which is a mathematical

* Corresponding author. Tel.: +351 239796276; fax: +351 239406672.
E-mail addresses: jermendes@isr.uc.pt (J. Mendes), fasouza@isr.uc.pt (F. Souza), rui@isr.uc.pt (R. Araújo), nunogon@isr.uc.pt (N. Gonçalves).

1568-4946/$ – see front matter
http://dx.doi.org/10.1016/j.asoc.2012.05.009

function. To design the T–S fuzzy model, the global operation of the nonlinear system can be accurately approximated into several local affine models. T–S fuzzy models are a powerful tool to implement the prediction structure of soft sensors.

GAs have proved to be useful in solving a variety of search and optimization problems. This motivates that GAs might be a useful soft computing technique for designing SS based on T–S fuzzy models. Genetic algorithms also have been used for learning classifier systems [4], and nonlinear system identification [5–10]. In [4], it is proposed a GAs approach to generate optimal fuzzy rules in a classification setting that continuously monitors system states for automatically detecting faults on a HVAC system. In [8], a genetic algorithms based multi-objective optimization technique was utilized in the training process of a feedforward neural network, using noisy data from an industrial iron blast furnace. In [11] an approach is proposed for a dynamic creation and evolving of a first order Takagi–Sugeno fuzzy system model. In [12] it is investigated a technique for modeling and identification of a new dynamic NARX fuzzy model by means of genetic algorithms. The paper proposed the use of a modified genetic algorithm combined with the predictive capability of NARX Takagi–Sugeno model for generating the dynamic NARX T–S fuzzy model. In [13] a fuzzy wavelet neural network model that uses a GA approach for adjusting parameters, is introduced for function approximation and nonlinear dynamic system identification from input–output pairs. [14] proposed a methodology for automatically extracting T–S fuzzy models from data using particle swarm optimization. The structures and parameters of the fuzzy models are encoded into particles and evolve together so that the optimal structure and parameters can be achieved simultaneously. In [15], a technique for the modeling of nonlinear control processes using fuzzy modeling approach based on the Takagi–Sugeno fuzzy model with a combination of a genetic algorithm and the recursive least squares method is proposed.

The above mentioned steps (II) and (III) are essential for correct DDSS development, being directly related to the selection of input variables and the respective time lags and to model identification. The selection of the most adequate input variables and the respective time delays is crucial since the use of the correct variables with the correct delays can lead to better prediction accuracy because they can contain more information about the output than incorrect variables and/or variables with incorrect delays. Some studies have used techniques based on variance, such as principal component analysis (PCA) for variable selection [16]. These methods are designed for linear models, so they cannot be the best choice for nonlinear modeling. In the works [5,6], input variables and respective time delays are automatically selected jointly with the learning of the prediction model. The methods of [4,7–10], have the limitation of not being able to perform automatic selection of variables and delays: pre-selection is performed. Pre-selection may be performed completely from human knowledge (e.g. knowledge of the real model, such as in [7,10,12–15,17]) or using some auxiliary criteria that does not take advantage of taking into account the prediction model being learned, such as correlation coefficients, Kohonen maps and Lipschitz quotients [9], regularity criterion [11,18] or analysis of "fuzzy curves" [19]. Some approaches have the limitation of not performing the selection of the time delays of input variables (e.g. [4,8]). While [5,7–10,17–19] learn fuzzy models, Ref. [6] learns NARX models. However, fuzzy models have a linguistic interpretation which is an important desirable characteristic for human users/operators of the SS, such as in industrial applications.

An approach using methods for both nonlinear variable selection and learning T–S fuzzy models was proposed by [20] and later by [9]. In [20], it is introduced a hierarchical evolutionary approach to optimize the parameters of Takagi–Sugeno fuzzy systems, where the selection of the variables is performed completely from human knowledge, such as knowledge about the real model. The problems addressed are function approximation and pattern classification. As an evolution or improvement of [20], in the work [9], it was proposed to add to [20] a mechanism for pre-selection of the variables by an auxiliary criteria. The proposed method is addressed for soft sensors applications. It uses T–S fuzzy models learned from available input/output data by means of a coevolutionary GA and a neuro-based technique. The soft sensor design is carried out in two steps. First, the input variables of the fuzzy model are pre-selected from the variables of the dynamical process by means of correlation coefficients, Kohonen maps and Lipschitz quotients. Such selection procedure considers nonlinear relations among the input and output variables. Second, a hierarchical GA is used to identify the fuzzy model itself. The input variable selection approach proposed by [9] has some drawbacks. First, the selection of the number of neurons in the Kohonen maps is not automatically performed. Second, variables and delays selection is not jointly performed with the learning of the fuzzy model (pre-selection is performed), which precludes the global optimization of the prediction setting. Finally, the selection of input variables is not accompanied with the selection of the respective time delays. The later shortcoming can bring low-accuracy results because a variable with the correct delay can contain more information about the output than a variable with an incorrect delay.

This paper proposes a novel methodology for soft sensors development. The proposed method is an automatic tool for SS design, with the following main characteristics: (1) it automatically performs the optimization of the variable and delay selection jointly with the learning and optimization of the system model without the need for any prior human knowledge, (2) the T–S fuzzy model structure is constructed just according to the data characteristics, and (3) it is optimized by means of GAs. This work has been inspired by [9]. However, it will jointly optimize a larger number of components of the prediction setting when compared with in [9]. A hierarchical genetic algorithm (HGA) will be used to optimize a large set of parameters encoded at five different levels to design the T–S fuzzy model. When more complex design decisions involving a large number of parameters must be made, a global formulation of the problem representing all the parameters in just one optimization level can be inadequate. It is well known that computation, search, and optimization problems become more difficult to solve when the dimensionality of the state-space increases. In many cases, this problem is known as the curse of dimensionality. To tackle this issue, in this paper the global problem is divided into various optimization levels, where the genetic evolution (optimization) of each level is performed separately, but is influenced by the current populations and optimization states of all the levels. HGAs make it possible to have different layers optimizing different parts of the T–S model, and facilitate the human interpretation of the optimization structure. The main advancement of is this work in comparison with [9] is the addition of a new hierarchical level responsible for the selection of variables and delays. The hierarchical genetic fuzzy system is constituted by five levels. In the first level, the input variables and respective delays are chosen with the goal of attaining the highest possible prediction accuracy of the T–S fuzzy model. The selection of variables and delays is performed jointly with the learning of the fuzzy model, which increases the global optimization performance. The second level encodes the membership functions. The individual rules are defined at the third level. The population of the set of rules is defined in the fourth level, and a population of fuzzy systems is treated at the fifth level. The least squares method is used to determine the parameters of the rule consequents. Levels II–V were based on [9].

To validate and demonstrate the performance and effectiveness of the proposed algorithm, it is applied on two prediction problems. The first is the Box–Jenkins benchmark problem, and the