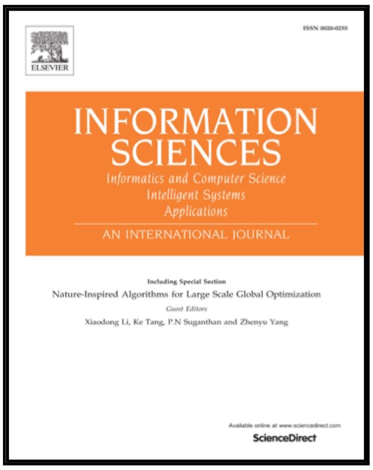
Accepted Manuscript



A Robust Model Structure Selection Method for Small Sample Size and Multiple Datasets Problems

Yuanlin Gu , Hua -Liang Wei

PII: S0020-0255(18)30262-7
DOI: 10.1016/j.ins.2018.04.007
Reference: INS 13550

To appear in: *Information Sciences*

Received date: 19 October 2017
Revised date: 12 March 2018
Accepted date: 2 April 2018

A Robust Model Structure Selection Method for Small Sample Size and Multiple Datasets Problems

Yuanlin Gu, Hua-Liang Wei*

(Corresponding author: w.hualiang@sheffield.ac.uk)

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, United Kingdom

Abstract: In model identification, the existence of uncertainty normally generates negative impact on the accuracy and performance of the identified models, especially when the size of data used is rather small. With a small data set, least squares estimates are biased, the resulting models may not be reliable for further analysis and future use. This study introduces a novel robust model structure selection method for model identification. The proposed method can successfully reduce the model structure uncertainty and therefore improve the model performances. Case studies on simulation data and real data are presented to illustrate how the proposed metric works for robust model identification.

Keyword: nonlinear systems; systems identification; model uncertainty; model structure detection

1. Introduction

The procedure of model identification includes several steps including data collection and processing, selection of model representation, model structure detection and selection, model parameter estimation, and model validity test [25]. A wide variety of model types have been developed for nonlinear input-output system identification, modelling and control, for example, nonlinear autoregressive with exogenous inputs (NARMAX) model [15], neural networks [13,16,20,26], Bayesian network [19], fuzzy model [11,27,36,37,39], wavelet models [5,8,31,38] and so on. Among these, the NARMAX model is one of the most commonly used model types for many real-world applications including ecological systems [22], environmental systems [3], space weather [1,10,34], medicine [4], societal [18] and neurophysiological [21] sciences, etc.

Broadly speaking, data based modelling approaches can be categorized into two groups: parametric and nonparametric. Nonparametric methods are those that do not make strong assumptions about the form of the mapping functions (that map the model "input" variables to the model "output" variables). Most existing artificial neural networks are nonparametric approaches. In [24] it is stated that "Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don’t want to worry too much about choosing just the right features” (p.757). One of the advantages of neural networks is that in general they can achieve relatively higher performances in dealing with complicated data modelling problems defined in high dimensional space. However, the model structure of most neural networks is very complicated and cannot be simply written down. In addition, neural networks models often involve a large number of variables and take a long time for training. General neural networks models cannot provide a transparent model structure, where the significance of individual variables and the role of their interactions are invisible. Moreover, the implementation of some nonparametric approaches for example Bayesian networks normally would need a huge number of samples. In comparison with neural networks models, parametric NARX models use a nonlinear polynomial structure and often only need a small number of effective model terms to describe the system. It can be achieved by selecting a number of most important model terms by an orthogonal forward regression (OFR) algorithm [14,33], so that it generally only requires a relatively small number of input and output data points [6,30]. In many applications (e.g. [3],[4]), where the main objective of the modelling tasks is not only to predict future behavior, but also reveal and understand which model variables are most important and how the candidate variables interactively affect the system behavior, parametric models are usually become a first choice.

Under some specific conditions and assumptions, most existing model identification methods work well and can provide sufficiently reliable models for most applications. However, in many cases where there is modelling uncertainty (e.g. in data, model form and structure, parameters, noise level, etc.), the identified models may lack reliability and thus less useful. This is particularly true when the available data set is small. This study focuses on parametric models and aims to answer the following challenging question. Given a small set of experimental data of a system, how to build a model that best represents the underlying system dynamics hidden in the data? Most data modelling approaches can