Contents lists available at SciVerse ScienceDirect

Control Engineering Practice

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

Probabilistic characterisation of model error using Gaussian mixture model—With application to Charpy impact energy prediction for alloy steel

Yong Yao Yang*, Mahdi Mahfouf, George Panoutsos

Institute for Microstructural and Mechanical Process Engineering: The University of Sheffield (IMMPETUS), Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield S1 3JD, UK

ARTICLE INFO

Article history:
Received 4 February 2011
Accepted 2 October 2011
Available online 28 October 2011

Keywords:
Prediction error
Probabilistic reasoning
Gaussian mixture model
Hybrid modelling
Artificial neural network

ABSTRACT

A novel approach to characterise the model prediction errors using a Gaussian mixture model is proposed. The motivation for this work lies behind many data models that are developed through prediction error minimisation with the assumption of a normal noise distribution. When the noise is non-normal, which may often be the case in complicated data modelling scenarios, the model prediction errors may contain rich information, which can be further exploited for model refinement and improvement. The key contents presented in this paper include: choosing the relevant variables to form the error data, optimising the number of Gaussian components required for the error data modelling, and fitting the Gaussian mixture parameters using an expectation-maximisation algorithm. Application of the proposed method for further model improvement, within the framework of hybrid deterministic/stochastic modelling, is also discussed. Preliminary results on the real industrial Charpy impact energy data for heat-treated steels show its effectiveness for model error characterisation, and the potential for model performance improvement in terms of prediction accuracy as well as providing accurate prediction confidence intervals.

1. Introduction

Data-driven modelling has gathered much pace and popularity due to the rapid growth in computing power and the availability of extensive data and various information in modern industrial processes. It has been developed with contributions from artificial intelligence, data mining, knowledge discovery in databases, computational intelligence, machine learning, intelligent data analysis, soft computing, and pattern recognition. There are often overlaps among those different terminologies due to the nature of parallel development related to data modelling in different disciplines. Common data modelling approaches include multivariate regression, artificial neural networks, and adaptive neuron-fuzzy systems, to only mention a few here (Abbod, Zhu, Linkens, Sellars, & Mahfouf, 2006; Mahfouf, Jamei, Linkens, & Tenner, 2008). While in statistical regression the regressor structure and the type of input–output function need to be predefined, other intelligent approaches, such as artificial neural network, need little pre-assumptions and are more flexible due to their capacity of being a universal approximator. The modelling strategies and algorithms have also evolved towards dealing with nonlinear and high-dimensional input/output mappings. However, many of the data models are derived based on some type of error minimisation, under the explicit or implicit assumption that the modelling errors follow a normal distribution. In reality, the process to be modelled is often very complex, suffering from different random disturbances, various measurement scatters, and non-measurable (hidden) inputs. Hence, the assumption of a normal error distribution may not be valid, leading to sub-optimal model predictions.

In this paper, a new modelling strategy aimed at exploiting the rich information hidden behind the prediction error data using a Gaussian mixture model (GMM) is proposed. GMM has found extensive applications in speaker verification, colour image detection, and other similar pattern classification and cluster analysis problems (He, Pan, & Lin, 2006; Huang & Chau, 2008; Kinnunen, Saastamoinen, Hautamäki, Vinni, & Fränti, 2009), due to its advantage of well-known properties, analytically tractability, and the existence of elegant expectation maximisation algorithms. Chen, Morris, and Martin (2006) investigated infinite GMM for probability density estimation and its application in statistical process monitoring. However, the application of GMM for error data characterisation has, to the authors' best knowledge, not yet been exploited. We believe that this is a promising research topic, as the GMM structure has the ability to approximate any reasonably behaved continuous probability density function (pdf) provided that enough Gaussian components are included. Hence, it can be employed to model complex non-normal model error distribution. The main motivation here is to

* Corresponding author.
E-mail addresses: y.y.yang@sheffield.ac.uk, yongyyang@gmail.com (Y.Y. Yang).

0967-0661/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.conengprac.2011.10.001

develop a GMM for non-normal model error characterisation, and to use the GMM to further improve the original data model in a complimentary way through model fusion. Alternatively, the error GMM can be employed to pursue data model validation and redevelopment, through hypothesis testing or data model refinement guided by the information extracted from the error GMM.

The concept of using prediction errors (also known as residuals) for model refinement is not new. Cooks and Weisberg (1982) had looked the role residuals can play in regression, while Oliveira and Pedrycz (2007) conducted analysis on the residuals to validate the assumptions of normality and homoscedasticity in their fuzzy clustering. Mauricio (2008) discussed the computing and use of residuals in time serials modelling, with focus on using residuals to reduce model inaccuracies. The majority of residual analysis research aims at hypothesis testing in order to confirm or reject the initial assumptions for modelling, and if rejected how revised assumptions are to be postulated for progressive model development. Some statistical data models are developed based on the relaxed i.i.d assumption of the observed data without imposing the restriction of normal distributions (Vapnik, 1998), and approaches to validate the assumptions based on the concept of model complexity and data nature (Kantz and Schreiber, 2003). The error data have been exploited in quite a different way in this paper, with the random noises involved in the modelling being viewed as complex, neither a normal distribution nor i.i.d. The residuals, together with other relevant input variables, are used to elicit a probabilistic model in the form of a GMM, so that the complicated probabilistic behaviour of the prediction error can be exploited in order to improve the data model through hybrid deterministic/stochastic modelling. As this type of modelling philosophy is relatively new, concepts, issues and rationales relating to the error data modelling using GMM, as well as how such an error GMM can be harmonised with the original data model, are also being investigated. The remainder of the paper is organised as follows: Section 2 introduces a typical data modelling scenario concerning the input–output relationship, together with an example of artificial neural network data modelling using a synthetic data set. Section 3 outlines the error data characterisation based on a GMM framework and situations where such an error probabilistic modelling should be most beneficial. Key techniques and algorithms for the error GMM are then illustrated using the same synthetic data introduced in Section 2, together with some general guidance for error GMM implementation and discussions. In Section 4, a case study of the error GMM on the Charpy impact energy data extracted from an industrial database containing heat-treated steels is presented. The resultant error GMM is then exploited to improve the associated neural network data model for Charpy impact energy prediction, with output correction based on the conditional error means. In addition, reliable confidence intervals for the model predictions can be calculated based on the conditional error standard deviations derived from the error GMM. Section 5 concludes the paper with discussions, remarks, and suggestions for future research.

2. Data modelling using neural network

It is assumed that no sufficient physical insight of the process is available, so a physical based model or first principle model cannot be formulated. In data driven modelling, the first task should involve the choice of a model form (often belongs to families that are known to have good flexibility and have been "successful in the past") based on the nature of the available model data. Common data-driven model paradigms include, but not limited to, statistical regression, neural networks and adaptive fuzzy systems (Bishop, 2006; Chen & Linkens, 2001). Recently, data-driven models have been constructed through Genetic Programming and evolutionary algorithm based approaches (Brezocnik, Buchmeister, & Gusel, 2011; Kovacic, 2009), and are increasingly becoming a competitor of artificial neural networks as far as input–output mapping is concerned. The primary objective of the data modelling here is to provide good predictions of the outputs for new inputs, rather than to find the true input–output relationship, which is often impossible for complicated processes.

The general form of the data model can be expressed in the following mathematical form:

$$\begin{aligned} &y = g(\boldsymbol{x},\boldsymbol{\theta}) + \xi \in R \\ &\boldsymbol{x} = [x_1,x_2\ldots x_n]^T \in R^n \\ &\boldsymbol{\theta} = [\theta_1,\theta_2\ldots\theta_l]^T \in R^l \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is a n-dimensional input vector, y is the output variable, $g(\boldsymbol{x},\boldsymbol{\theta})$ is a nonlinear function with a l-dimensional parameter vector $\boldsymbol{\theta}$, ξ is the model error, which accounts for all random noises and un-modelled errors on the output, and the superscript T represents the transpose of the corresponding vector or matrix.

Eq. (1) represents a typical nonlinear static process, but can be generalised for dynamic processes by augmenting the input vector with the past inputs and past outputs. It can also be extend to multiple-inputs multiple-outputs (MIMO) systems by introducing a vector function and a vector output to replace the corresponding scalar components. It can even cope with time-varying system, via the mechanism of defining the parameter vector $\boldsymbol{\theta}$ being a function of time t. Of course, the corresponding time-varying modelling techniques, such as time-recession window, need to be adopted and they are beyond the scope of this paper.

The main source of information for data-driven modelling is primarily from the available input–output data, although knowledge about the system to be modelled is beneficial and should be used if available in choosing the model paradigm and model structure. The available model data can be arranged into input matrix $\boldsymbol{X}$ and output vector $\boldsymbol{Y}$, as given by the following equation:

$$\begin{aligned} &\boldsymbol{X} = [\boldsymbol{x}(1),\boldsymbol{x}(2),\ldots,\boldsymbol{x}(N)]^T = \begin{bmatrix} x_1(1) & x_2(1) & \ldots & x_n(1) \\ x_1(2) & x_2(2) & \ldots & x_n(2) \\ \ldots & \ldots & \ldots & \ldots \\ x_1(N) & x_2(N) & \ldots & x_n(N) \end{bmatrix} \\ &\boldsymbol{Y} = [y(1),y(2),\ldots,y(N)]^T \end{aligned} \tag{2}$$

where N is the total number of process data available for modelling. The input matrix $\boldsymbol{X}$ is arranged in such a way that each row represents a record consisting of all the input variables at a specific measurement, while the columns in $\boldsymbol{X}$ represent the collection of measurements for the corresponding inputs.

The central task in data-drive modelling is to identify the model parameter vector $\boldsymbol{\theta}$ from the available process data $(\boldsymbol{X},\boldsymbol{Y})$ under the selected model framework. Usually this is achieved by some kind of optimisation for a specified performance criterion. A commonly used performance criterion is the root mean square error (RMSE), defined by

$$RMSE = \sqrt{\sum_{k=1}^{N}((y(k) - \widehat{y}(k))^2/N} \tag{3}$$

where $\widehat{y}(k) = g(\boldsymbol{x}(k),\ \widehat{\boldsymbol{\theta}})$ is the output prediction of the model parameterised by $\widehat{\boldsymbol{\theta}}$. The role of the data model, in the perspective of error data probabilistic characterisation, is to generate the prediction errors for the available model data. Any model framework, be it multivariate regression, neural network (Bishop, 1995;