ELSEVIER

Fuel Processing Technology 88 (2007) 483–491

FUEL PROCESSING TECHNOLOGY

www.elsevier.com/locate/fuproc

Predictive modeling of mercury speciation in combustion flue gases using GMDH-based abductive networks

R.E. Abdel-Aal *

Department of Computer Engineering, King Fahd University of Petroleum and Minerals, P. O. Box 1759, Dhahran 31261, Saudi Arabia

Received 29 June 2006; received in revised form 6 December 2006; accepted 12 December 2006

Abstract

Modeling mercury speciation is an important requirement for estimating harmful emissions from coal-fired power plants and developing strategies to reduce them. First-principle models based on chemical, kinetic, and thermodynamic aspects exist, but these are complex and difficult to develop. The use of modern data-based machine learning techniques has been recently introduced, including neural networks. Here we propose an alternative approach using abductive networks based on the group method of data handling (GMDH) algorithm, with the advantages of simplified and more automated model synthesis, automatic selection of significant inputs, and more transparent input–output model relationships. Models were developed for predicting three types of mercury speciation (elemental, oxidized, and particulate) using a small dataset containing six inputs parameters on the composition of the coal used and boiler operating conditions. Prediction performance compares favourably with neural network models developed using the same dataset, with correlation coefficients as high as 0.97 for training data. Network committees (ensembles) are proposed as a means of improving prediction accuracy, and suggestions are made for future work to further improve performance.

Index terms: Mercury speciation; Flue gases; Boiler emissions; Predictive modeling; Inferential emission monitoring; Soft sensors; Abductive networks; GMDH algorithm; Neural networks; Network committees; Network ensembles

1. Introduction

Determining mercury speciation is an important requirement for estimating mercury emission from combustion flue gases and the efficiency of control measures to reduce it. Major mercury sources from human activities are coal-fired electric utility boilers, where speciation depends on the operating conditions, including the type of coal used and flue gas temperature and composition. Mercury compounds from combustion sources consist mainly of gaseous elemental mercury (Hg^0), gaseous oxidized mercury (Hg^{2+}), and particle-bound mercury (Hg_p) [1]. Theoretical first-principle approaches have been used to study mercury speciation, including kinetic modeling [2] and thermodynamic equilibrium calculations [3]. Theoretical models rely on knowledge of the processes involved, which are often quite complex and highly nonlinear and therefore are difficult to describe accurately. Model development can also be expensive and time consuming. In addition, the above methods suffer from limited accuracy due to the lack of accurate rate constants of reaction mechanisms and to uncertainties caused by model assumptions and simplifications and incomplete understanding of mercury science [4]. Recently, data-based modeling using machine learning techniques, such as neural networks, fuzzy logic, and genetic algorithms, has become a popular approach for solving complex nonlinear problems without requiring exhaustive theoretical knowledge of the phenomenon being modeled. Such approaches depend primarily on experimental input–output data on the process, which are usually readily available in large quantities, rather than accurate theoretical knowledge. A model for the phenomenon considered is developed through training on input–output process data in the form of an adequate number of solved examples. Once synthesized, the model can be used to perform fast predictions of outputs corresponding to new cases previously unseen during training. The method offers a number of advantages over conventional approaches, including increased tolerance to noise and uncertainty, reduced need for knowledge on the modeled phenomenon, and the relative ease of developing and updating the model. In the last few years, neural networks have

* Tel.: +966 3 860 4320; fax: +966 3 860 3059.
E-mail address: radwan@kfupm.edu.sa.

0378-3820/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.fuproc.2006.12.005

formed the basis of many soft (inferential) sensors for monitoring pollutant emissions [5–9]. Such sensors offer a cost-effective and reliable alternative to expensive online analyzers in many areas of application, including mercury monitoring. Tsiros and Dimopoulos have used neural networks and other statistical and machine learning techniques to model atmospheric emission of gaseous soil mercury and identify critical factors for controlling such emissions [10]. Neural networks were used together with optical fiber chemical sensors for monitoring mercury and other heavy metals in aquatic samples [11]. Neural networks were used for modeling a spectrophotometric kinetic system and optimizing the experimental conditions for measuring traces of mercury in water [12]. Applications of neural networks to modeling mercury speciation in flue gases are relatively scarce in the literature, and only the work by Jensen et al. [13] could be cited at the time of writing.

In general, the neural network approach suffers from a number of limitations, including difficulty in determining optimum network topology and training parameters [14]. There are many choices to be made in determining numerous critical design parameters with little guidance available [15], and designers often resort to trial and error approaches which can be tedious and time consuming [16,17]. Such design parameters include the number and size of the hidden layers, the type of neuron transfer functions for the various layers, the learning rate and momentum coefficient, and training stopping criteria to avoid overfitting and ensure adequate generalization with new data. Another limitation is the black box nature of neural network models that give little insight into the modeled relationship and the relative significance of various inputs, thus providing poor explanation facilities [18]. The acceptability of, and confidence in, automated prediction tools in areas such as electric load forecasting, pollution control and medical diagnosis is related to their transparency and their ability to justify results to human operators, experts and decision makers [19]. To overcome such limitations, we propose using self-organized abductive networks [20] based on the group method of data handling (GMDH) learning algorithm [21,22] as an alternative machine learning approach to modeling and estimating mercury speciation in the flue gasses of coal-fired power plants. We have previously used this approach in several weather prediction applications including modeling and forecasting the minimum [23] and maximum [24] daily temperatures and the hourly temperature profile [25]. Compared to neural networks, abductive networks offer the advantages of faster model development requiring little or no user intervention, faster convergence during model synthesis without the problem of getting stuck in local minima, automatic selection of effective input variables, and automatic configuration of the model structure [14]. With the model represented as a hierarchy of polynomial expressions, resulting analytical model relationships can provide insight into the modeled phenomena, highlight contributions of various inputs, and allow comparison with previously used empirical or statistical models. The technique automatically avoids overfitting by using a proven regularization criterion based on penalizing model complexity [22] without requiring a dedicated validation dataset during training, as is the case with many neural network paradigms.

Following a brief description of abductive networks and the underlying GMDH learning algorithm in Section 2, the mercury speciation dataset used in this study is described in Section 3. In Section 4, abductive network models for the three types of mercury speciation are described and their performance on both the training and evaluation sets is analyzed and compared with neural network results reported in the literature for the same dataset. Single (monolithic) abductive models of various levels of model complexity are presented. Modular network committees (ensembles) are also introduced as a means of improving prediction accuracy beyond that obtained with the monolithic models. Section 5 includes conclusions and suggestions for future work.

2. GMDH and AIM abductive networks

AIM (abductory inductive mechanism) [26] is a supervised inductive machine learning tool for automatically synthesizing abductive network models from a database of inputs and outputs representing a training set of solved examples. As a GMDH algorithm, the tool can automatically synthesize adequate models that embody the inherent structure of complex and highly nonlinear systems. Automation of model synthesis not only lessens the burden on the analyst but also safeguards the model generated against influence by human biases and misjudgments. The GMDH approach is a formalized paradigm for iterated (multi-phase) polynomial regression capable of producing a high-degree polynomial model in effective predictors. The process is 'evolutionary' in nature, using initially simple (myopic) regression relationships to derive more accurate representations in the next iteration. To prevent exponential growth and limit model complexity, the algorithm selects only relationships having good predicting powers within each phase. Iteration is stopped when the new generation regression equations start to have poorer prediction performance than those of the previous generation, at which point the model starts to become overspecialized and therefore unlikely to perform well with new data. The algorithm has three main elements: representation, selection, and stopping. It applies abduction heuristics for making decisions concerning some or all of these three aspects.

To illustrate these steps for the classical GMDH approach, consider an estimation database of n_e observations (rows) and $m+1$ columns for m independent variables (x_1, x_2,..., x_m) and one dependent variable y. In the first iteration we assume that our predictors are the actual input variables. The initial rough prediction equations are derived by taking each pair of input variables (x_i, x_j; $i,j=1, 2,..., m$) together with the output y and computing the quadratic regression polynomial [21]:

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_ix_j. \quad (1)$$

Each of the resulting $m(m-1)/2$ polynomials is evaluated using data for the pair of x variables used to generate it, thus producing new estimation variables (z_1, z_2,..., $z_{m(m-1)/2}$) which would be expected to describe y better than the original variables. The resulting z variables are screened according to