Applied Thermal Engineering 143 (2018) 160–171

Contents lists available at ScienceDirect

Applied Thermal Engineering

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

Research Paper

Typical condition library construction for the development of data-driven models in power plants

You Lv[a,*], Carlos E. Romero[b], Tingting Yang[a], Fang Fang[a], Jizhen Liu[a]

[a] State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Beijing 102206, China
[b] Energy Research Center, Lehigh University, Bethlehem, PA 18015, USA

HIGHLIGHTS

- Performance of data-driven models highly depends on the data for model training.
- The typical condition library to collect informative operating data is presented.
- An expression is provided to describe the condition information of operating data.
- Genetic algorithm is used to search operating data with a large information index.
- Samples in the library are used to develop models of a SCR system in power plants.
- The proposed library is effective for data preparation of the model development.

ARTICLE INFO

Keywords:
Data-driven model
Operating data processing
Typical condition selection
Selective catalytic reduction
Thermal power plant

ABSTRACT

The performance of data-driven models to describe processes in thermal power plants highly depends on the operating data that are used for model training. Models developed using representative operating data with rich process information are more likely to produce accurate predictions. This paper presents a concept of typical condition library to collect informative operating data from power plants. First, a quantitative criterion to measure the condition information of operating data is defined by three factors including variation span, distribution status, and redundancy. Additionally, an analytic expression is provided to describe the condition information metric. Then, the genetic algorithm is used to search operating data samples with the largest information for the condition library construction. A numerical case is given to validate the proposed metric and the sample selection strategy. Finally, the typical condition library is applied to develop a model for a selective catalytic reduction (SCR) system in a coal-fired power plant. The results indicate that the SCR model trained using samples in the library can give accurate predictions, and the typical condition library is effective for initial data preparation of model development.

1. Introduction

Data have become an important aspect of industrial production processes. The application of information technology in power plants has allowed real-time recording of operating data in distributed control systems (DCS) and supervisor information systems (SIS) [1]. Detailed operation information of power plants is contained in these data, based on which process models can be developed using statistical and artificial intelligence techniques. Such modeling methods are also known as data-driven models. Compared with first-principle models, data-driven models have the advantage of not requiring precise knowledge about the process mechanism [2]. In addition, operation optimization and condition monitoring can be applied using data-driven models without modifying the equipment, thereby saving time and resources [3]. Thus, data-driven models are valuable for the optimal and secure operation of the power generation process.

When developing data-driven models, a common practice is to select an independent set of samples from the available data to train the model, and the remaining set is used for model validation. Thus, the model performance heavily depends on the characteristics of the data samples used for model training [4–6]. In general, two types of data sources, namely, field test data and actual operating data, are mainly used to establish models in power plants [7]. The field test data are obtained by performing parametric field tests on power plants. In the

* Corresponding author.
E-mail addresses: you.lv@hotmail.com, you.lv@ncepu.edu.cn (Y. Lv).

https://doi.org/10.1016/j.applthermaleng.2018.07.083
Received 5 March 2018; Received in revised form 24 June 2018; Accepted 16 July 2018
Available online 20 July 2018
1359-4311/ © 2018 Elsevier Ltd. All rights reserved.

field test, a list of conditions is given according to test designs, and operating parameters are tuned to enable the operation of the boiler under these conditions. Conducting a field test can be costly because considerable time and effort need to be invested. Data samples acquired from field tests are limited in number but are informative. Field test data usually cover a wide operating range, and they are distributed uniformly with no redundant information. Different from field test data, actual operating data are recorded during daily operation, which makes them easier to obtain. A significant feature of the operating data is being sample-rich but information-poor. Power plant loads usually change depending on the electric regulation in daily operation. Thus, operating data are irregularly distributed in different conditions.

A problem with models developed using operating data is that training samples cannot contain all possible operating conditions if they are randomly selected [8]. The model trained using data samples in certain local conditions can only partially describe the process characteristics. If the operating condition changes, predictions will be inaccurate [9,10]. Considering all the data to train the model requires a heavy computational burden, and information redundancy will exist in the training set, which can reduce the model accuracy. Thus, a feasible solution is to select some informative samples that cover the entire condition range for model training. This paper presents a concept of collecting operating data with rich information into a data set named a typical condition library. The typical condition library is defined as a data set that consists of samples that can represent the characteristics of the entirety of operating conditions. The samples in the typical condition library may be small in size but informative, which indicates that they can contain rich information on operating conditions, that is, condition information. Thus, accurate predictions can be guaranteed if these samples are used to construct data-driven models in power plants.

Statistical and artificial intelligence methods, including artificial neural networks (ANN) and support vector machines (SVM), have provided a theoretical basis for developing data-driven models in the power and energy industry. Kalogirou [11] reviewed the application of artificial intelligence in energy system modeling. A large body of literature has reported the use of operating data to construct models for the prediction of important parameters in power plants, such as boiler temperature [12], drum pressure [13], wall temperature [14], boiler performance [15,16], cooling tower [17], and heat transfer [18]. Theoretically, these models can obtain complete characteristics if the selected training samples are as large as possible. However, condition similarity exists in different periods of the operating data because of load regulation by the grid. Selecting more training data does not necessarily increase the condition information. Instead, this procedure may add information redundancy and model complexity, which will reduce prediction precision and computation efficiency. Furthermore, newly acquired data that contain more useful information will represent a small proportion if the initial training data size is significantly large, which will influence the model update [19].

Thus, some scholars have investigated the influence of the training data on model performance. Tong et al. [20] studied the accuracy of ANN models using different training sets, and comparison results showed that the model trained using a set with uniformly distributed samples had better accuracy than the model trained using simply selected samples. Alam et al. [21] analyzed the effect of training data on the performance of ANN models, and they found that the model developed based on the Latin design performed best in terms of prediction accuracy. Random sampling, stratified sampling, and clustered sampling methods have also been presented to select training data for model development in [22]. The abovementioned methods were mainly applied to chemometrics, spectrum analysis, image processing, and text recognition. Some other methods have been proposed to select training data samples for ANN and SVM learning [23,24], and they mainly focused on the efficiency improvement of specific modeling algorithms instead of the operating data.

Research on training data selection for model development has also been conducted in the field of power and energy. Paudel et al. [25] obtained a representative training set by selecting relevant samples from all the available data, and they used them to develop SVM models for the prediction of energy consumption. In research of [21,26], informative data samples were acquired by conducting field tests on boilers in power plants. The obtained samples were different from the operating data because parameters could not be tuned in accordance with the test design during daily operation. Smrekar et al. [27] considered data selection an important stage of the ANN model development and stated that the training set should cover a complete range of operating conditions with wide variation to train ANN models successfully with a small number of samples. García et al. [9] used a self-organizing concept to select training samples and constructed a model to predict the electric characteristics of the photovoltaic material. Hakeem et al. [28] also pointed that reliable data sets were critically important for the training and testing of ANN models. Some other references utilized the moving window method and the similarity criterion to find samples that were related to the current operating condition, and these samples were then applied to construct local models. When new conditions occurred, the samples were searched again to reconstruct and update these local models [29].

Although studies have been conducted on data-driven models in the power and energy industry, few of them present a systematic description of the selection of informative samples from actual operating data for model training. The main novelty of this study is that it provides a quantitative criterion to measure the condition information contained in the operating data. Based on this criterion, an intelligent optimization algorithm, that is, the genetic algorithm (GA), is applied to search informative samples for constructing a typical condition library, which can provide a data basis for process modeling and condition monitoring. A numerical simulation is given to validate the proposed metric and sample selection method. Furthermore, the typical condition library is applied to construct the model of a selective catalytic reduction (SCR) system in a coal-fired power plant. Training sets and prediction results of SCR models are compared and discussed.

The remainder of the paper is organized as follows. Section 2 presents a theoretical analysis of factors that influence the information quality of operating data. Section 3 provides an expression of condition information metric and describes the construction of typical condition library using GA. In Section 4, a numerical case is given for validating the proposed metric and library. The application of typical condition library to model the SCR system in a power plant is presented in Section 5. Section 6 concludes the paper.

2. Analysis of factors influencing data quality

Without loss of generality, we consider modeling a process with multiple inputs and a single output:

$$y = f(\boldsymbol{x},\theta) \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is independent variables, p is the dimension, $y_i \in \mathbb{R}$ is the predicted variable, and θ is the model parameter vector. The model training process would find the optimal θ to minimize prediction errors. The model performance heavily depends on the property of the training data samples. The performance will be good if the model is trained using informative samples; conversely, the performance will be poor if the model is trained using samples with low information. In collecting high-quality samples into the typical condition library, a quantitative criterion should be introduced to measure the condition information contained in the data samples. Three factors, namely, variation span, distribution status, and redundancy, are considered to define the condition information metric, and the analysis of these factors is detailed below.

Download English Version:

https://daneshyari.com/en/article/7044624

Download Persian Version:

https://daneshyari.com/article/7044624

Daneshyari.com