Energy 147 (2018) 418–427

Contents lists available at ScienceDirect

Energy

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

A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of Ethylene industry

Yan-Lin He [a, b], Ping-Jiang Wang [a, b], Ming-Qing Zhang [a, b], Qun-Xiong Zhu [a, b, *], Yuan Xu [a, b, **]

[a] College of Information Science & Technology, Beijing University of Chemical Technology, Beijing, 100029, China
[b] Engineering Research Center of Intelligent PSE, Ministry of Education of China, Beijing, 100029, China

ARTICLE INFO

Article history:

Keywords:
Energy prediction and analysis
Small data
Virtual samples generation
Nonlinear interpolation
Extreme learning machine

ABSTRACT

An accurate energy prediction and optimization model plays a very important role in the petrochemical industries. Due to the imbalanced and uncompleted characteristics of complex petrochemical small data, it is a big challenge to build accurate prediction and optimization models for energy analysis. In order to solve this problem, a nonlinear interpolation virtual sample generation method integrated with extreme learning machine is proposed. Well virtual input and output variables can be generated through interpolation of the hidden layer outputs of extreme learning machine. The generated virtual samples are put together with the original samples to train models for enhancing accuracy performance. To validate the effectiveness of the proposed nonlinear interpolation virtual sample generation method, a standard function is firstly selected, and then the proposed nonlinear interpolation virtual sample generation method is applied to developing a model of energy analysis for ethylene production systems. Simulation results showed that the prediction accuracy could be significantly improved, which provided helpful guidance for production departments and government to achieve the goal of energy management of petrochemical industries.

1. Introduction

It is well-known that data-driven methods have been widely used to build prediction models in many industries, such as chemical industry, metallurgy industry, and electricity industry. The data-driven methods are used for modeling by learning from the collected structural data. So, data-driven algorithms are data-dependent. In other words, if one wants to build an accurate and reliable data-driven model, sufficient data and a good distribution assumption are two necessary conditions [1]. In the references of [2] and [3], small data problems refer to the case where the number of samples is less than 50 concerning engineering applications or less than 30 regarding academic researches. The whole features of a population are hard to completely be revealed by the small data because of the insufficient information [4]. For example, a FLLS data-driven model is utilized to predict the key variables of complex chemical processes [5]. A least square algorithm is used in FLLS, where numbers of samples are required for the least square algorithm. If the FLLS model is built using a small number of samples, then the FLLS will achieve a bad performance in terms of accuracy and reliability. Small samples play a negative effect on the performance of data-driven models [6]. Although the big data era is coming in many fields, there are still two limitations. One of the limitations is that it usually takes many efforts and consumes much time in sorting and cleaning the big data due to the massive useless and ill-conditional information. The other limitation is that some key variables are different to measure online. In the chemical engineering, the small data problem usually occurs, especially in the early period of the production processes [7,8] and chemical process transitions [9]. An accuracy model plays a very important role in energy analysis. It is a big challenge to build an accurate model

* Corresponding author. College of Information Science & Technology, Beijing University of Chemical Technology, Beijing, 100029, China.
** Corresponding author. College of Information Science & Technology, Beijing University of Chemical Technology, Beijing, 100029, China.
E-mail addresses: zhuqx@mail.buct.edu.cn (Q.-X. Zhu), xuyuan@mail.buct.edu.cn (Y. Xu).

https://doi.org/10.1016/j.energy.2018.01.059
0360-5442/

based on collected data with a small number. The small data problem has caused serious concerns on both industrial circles and scientific researches. Hence, it is an urgent task to well solve the problem of small data.

Related researchers have developed varieties of methods to deal with the small data problem. In summary, there are three kinds of methods for handling the small data problem. The first one is based on the gray prediction model: the AGO of the gray theory is used to deal with the raw samples, like ANGM (1, 1) [10], GBM [11], BGM (1, 1) [2], and so on; the inherent mechanism of the system usually is ignored in the method of gray prediction; as a result, big mistake will occur sometimes. The second one is based on feature extraction: in this method, dimension reduction based methods are used to extract useful attributes and features for enhancing the analytical performance; the feature extraction method is usually applied to the field of medicine where the size of sample set is small but the dimensionality is fairly high, which is well known as large p small n dataset; projection pursuit [12], wrapper approach [13] and t-test selection [14] as feature extraction methods have been successfully applied; however, some important information may be lost through feature extraction, which is bad for the performance. The last one is based on VSG: in this method, the forecasting performance is enhancing by adding newly generated virtual samples.

Among the three state-of-the-art technologies, VSG based methods are the most promising and popular. Based on the data information hidden in the small data, many new virtual samples with useful information can be generated using the VSG technology. With the aid of the newly generated virtual samples, the accuracy of the prediction models can be improved. The idea of VSG was originally presented by Niyogi et al. [15], where virtual samples were generated using the prior information extracted from a given small training data. Then the newly generated virtual data is used for improving the performance of image recognition. The procedure of VSG is equivalent to incorporating the prior knowledge, which has been mathematically proved [15]. Li et al. [16] used the VSG method to acquire the robust dynamic environment knowledge of manufacturing systems. The VSG method was adopted to build up management knowledge in the early manufacturing stage [8]. In reference [17], the learning accuracy was significantly improved with the aid of VSG. MTD [18], GA [19], PSO [1] and Monte Carlo [20] based VSG were also proposed. Although many VSG methods have been proposed, some parameters of these VSG methods are difficult to determine.

For further solve the small sample problem for providing accurate model of energy analysis, an effective NIVSG method integrated with ELM is proposed in this paper. On one hand, the high nonlinearity nature between the process input data and the output data is taken into consideration in the proposed NIVSG method. On the other hand, the proposed NIVSG method integrates with an efficient neural network model for better generating the virtual samples. It is well known that the BPNN is a powerful nonlinear modeling tool and has been widely applied to a lot of engineering fields [21,22]. However, the gradient descent based learning algorithm used in BPNN and some other feedforward neural networks has some inherent flaws, like easily obtaining the local minima, difficulties in determining the initial parameters, consuming much time in the training phase, and so on [23]. To avoid the above drawbacks of the gradient descent based learning algorithm, ELM was proposed [24]. A brand-new learning algorithm with a fast learning speed was used in ELM. The efficient learning algorithm method of ELM has been successfully applied to many other neural networks, like RBFNN [25], FLANN [1], and so on. There are no parameters that need to be tuned in ELM [26]. ELM based methods have been widely used in many fields, like process modeling [27], pattern classification [28], prediction [29], and so on. That is the reason why ELM is adopted in this paper.

The rest of this paper is organized as follows: Section 2 presents a brief overview on extreme learning machine; a detailed description of the proposed NIVSG is given in Section 3; a case study and the simulation results are provided in Section 4, Section 5 contains conclusion.

2. Preliminaries

In this section, brief overviews of extreme learning machine are presented. The standard ELM algorithm consists of an input layer, a hidden layer and an output layer. The structure of ELM is shown in Fig. 1. The input weights between the input layer and the hidden layer can be assigned randomly and the output weights between the hidden layer and the output layer are analytically determined using the least square method. So, ELM is simple to understand and fast to implement.

The learning algorithm of ELM is given as follows. Consider a training dataset with M samples x_i, y_i is provided, in which $x_i = [x_{i1}, x_{i2}, ...x_{il}]^{\mathrm{T}} \in \mathrm{R}^l$ is the input vector, $y_i = [y_{i1}, y_{i2}, ..., y_{im}]^{\mathrm{T}} \in \mathrm{R}^m$ is the output vector. The output of ELM with N_h hidden nodes is shown as follows:

$$y_i = \sum_{j=1}^{N_h} \beta_i f(\omega_i \cdot x_j + b_i) \tag{1}$$

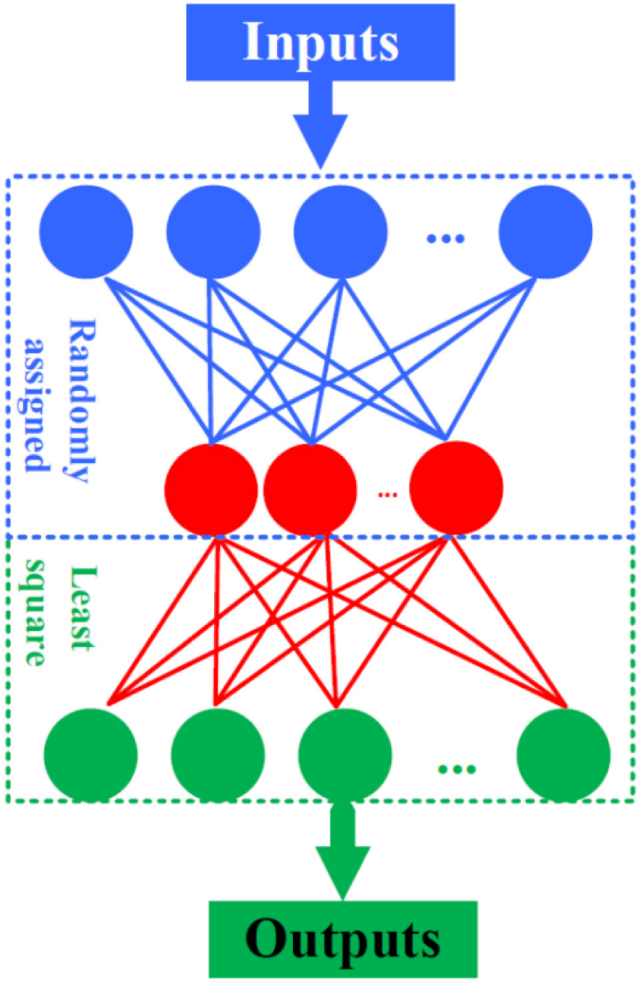

Fig. 1. The structure of extreme learning machine.