Chemometrics and Intelligent Laboratory Systems 171 (2017) 151–160

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# An improved multi-kernel RVM integrated with CEEMD for high-quality intervals prediction construction and its intelligent modeling application

Yuan Xu [a,b], Mingqing Zhang [a,b], Qunxiong Zhu [a,b,*], Yanlin He [a,b,*]

[a] College of Information Science and 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

*Keywords:*
Relevant vector machine
Complementary ensemble empirical mode decomposition
Prediction intervals
Modeling and prediction
Purified Terephthalic acid process

ABSTRACT

Most of existing modeling methods are based on point prediction. However, the accuracy of point prediction cannot meet the actual demand due to existence of high noise, volatility, complexity and irregularity inherent in the chemical process data. In order to solve this problem, a hybrid high-quality prediction intervals (PIs) method integrating complementary ensemble empirical mode decomposition (CEEMD), sample entropy (SE), and improved multi-kernel relevant vector machine (RVM) is proposed in the paper. The proposed PIs method mainly consists of three aspects: Firstly, CEEMD is adopted to decompose the original data into several independent intrinsic mode functions (IMFS), and then SE is used to analyze the complexity of the extracted IMFs to obtain recombinant components; Secondly, an improved multi-kernel RVM (MRVM) is presented to predict recombinant components independently, in which the linear kernel and the Gaussian kernel are combined; Thirdly, the predicted components are aggregated to obtain an ensemble result using another MRVM for constructing the high-quality PIs. To verify the performance of the proposed PIs method, a purified Terephthalic acid (PTA) solvent system is selected. Comparative simulation results demonstrate that the proposed PIs method greatly outperforms on coverage probability and sharpness in all the step predictions.

## 1. Introduction

Most of existing prediction methods are based on the deterministic point prediction. However, due to the non-stationary and high volatility of the time series data collected from process industries, the accuracy of the single point model cannot meet the actual demand. Aiming to solve this issue, a prediction interval (PI) approach is proposed as an essential and potential measure of uncertainty and risk [1,2]. Recently, there are two main methods to study the PI: (1) The upper and lower bounds of the interval are obtained directly by optimizing a function of target, without the prior knowledge of point estimation and data characteristic distribution, such as the lower and upper bounds estimation (LUBE) method [3–6] and double-outputs-based ELM neural network [7,8]. The optimization algorithms are used to optimize the objective function to achieve the narrower intervals bounds. However, the optimization Algorithm itself is prone to the problems of local convergence and computational complexity; (2) Another PI method is constructed by the point prediction error. A deterministic prediction model and special priori assumptions are often required in traditional point-based construction methods, such as Delta [9], Bayesian [10], MVE [11,12], and bootstrap [13,14] methods. However, there exists the problem of massive computational requirements. To solve the problem, statistical methods and artificial intelligence methods are often combined. Statistical methods mainly consider to be used in the time series [15], such as ARIMA method [16,17]. A non-stationary time series is transformed into stationary time series by ARIMA, and then the dependent variable is established only by its lagged value as well as the present value of the random error term. However, the high nonlinear relationship between the input and the output cannot be explained by ARIMA, only applicable to autoregressive model. Artificial intelligent methods are commonly used in PIs area, such as neural network and support vector regression. Due to the high fitting precision of neural network, it is widely used in PI prediction. Bootstrap and ELM are combined in Refs. [18,19], and the upper and lower bounds of the interval are constructed by using normal distribution. In support vector regression (SVR) [20,21], the input space is mapped to high dimension for getting the mapping relationship between output and input by the kernel functions. However, the kernel parameter and penalty factor are difficult to choose, and the calculated time is longer [22]. To solve the problem, least squares-based method (LSSVR) and Bayesian probability-based method [23] are proposed. De

* Corresponding authors. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China.
*E-mail addresses:* zhuqx@mail.buct.edu.cn (Q. Zhu), heyl@mail.buct.edu.cn (Y. He).

https://doi.org/10.1016/j.chemolab.2017.10.019
Received 11 September 2017; Received in revised form 23 October 2017; Accepted 27 October 2017
Available online 31 October 2017

et al. [24] and Cheng et al. [25] proposed a PIs method based on Least Squares Support Vector Regression, where the probability density estimation was used to construct the intervals. He et al. [26] proposed a PIs method based on relevant vector machine (RVM), since the variance was required to compute in the RVM algorithm. So, it is easy to be used in the interval prediction [27,28]. Merely, these methods rely on quantile analysis of the prediction error. Therefore, in order to obtain a high-quality interval, higher point prediction accuracy must be demanded [2,29].

In addition, since the inherent features are contained in the historical time series data and abundant information are provided by the features for point prediction [30]. Some decomposition methods are proposed to extract the internal features of time series. In Ref. [31], artificial neural network (ANN) integrating empirical mode decomposition (EMD) prediction method is proposed, while EMD exists the mode mixing phenomenon. As investigated in Refs. [32,33], ensemble empirical mode decomposition (EEMD) is applied to solve the mode mixing phenomenon of EMD by adding noise, but the added noise is not independent. In Ref. [34], complementary ensemble empirical mode decomposition (CEEMD) method is proposed to reduce the mode mixing effect by adding positive and negative noise, and has better convergence than EEMD. By using CEEMD the time series is decomposed into some intrinsic mode functions (IMFs) and the inherent characteristics are extracted. Then the sample entropy [35] is applied to analyze the complexity of IMFs, reconstruct the IMFs to group noise, cycle, trend components and eliminate the noise component for reaching the purpose of noise reduction. RVM as an improvement of SVR is widely-accepted Bayesian model commonly applied in regression. However, a typical local kernel Gauss kernel is used in the traditional RVM, and its kernel length is a given value. In order to improve these problems, a linear kernel function as a generalization of the better global kernel function is added to RVM, the linear kernel and Gaussian kernel are combined to form a multi-kernel RVM (MRVM) and the kernel density estimation (KDE) method is used to optimize the length of the kernel function. For the previous optimization, the probability density and the distribution function can be obtained by the KDE. MRVM is used to prediction cycle and trend components in this paper. To verify the performance of the proposed PIs method based on CEEMD and MRVM, PTA solvent system is used as case testing.

The remaining parts of this paper are organized as follows: in Section 2, the principles of CEEMD, SE and RVM algorithms are briefly explained; Section 3 presents the proposed PIs model based on CEEMD and MRVM in detail; the evaluation criteria measures are discussed in Section 4; the results and discussions of comparison experiments are shown in Section 5; Section 6 highlights the conclusions of the work.

## 2. Methodology formulations

In this section, the principles of CEEMD, SE and RVM algorithms are briefly introduced.

### 2.1. CEEMD

EMD [30] method can decompose time series into a series of IMFs based on local feature scales, but it exists model mixing phenomenon. To settle model mixing problem and improve the decomposition ability of EMD, an EEMD method is proposed by adding white noise analysis. However, the added noise is not independent. CEEMD method was presented as an improvement based on the EEMD and EMD. The added white noise in CEEMD is a pair of positive and negative to the original time series. In CEEMD, 2 times the white noise ensembles test is added to the EEMD method, so that the decomposition is more adequate and has better performance on the noise elimination and less consumption of extra computational time. The final obtained IMF component is the mean value of the multiple decomposition of the IMFs. The CEEMD is briefly explained as follows:

Step 1: Add a pair of positive and negative white noise into original time series data.

$$\begin{bmatrix} Q_j^+ \\ Q_j^- \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_N \\ n_N \end{bmatrix} \tag{1}$$

where $n_N$ is the added white noise; $Q_j^+$ and $Q_j^-$ are the sum of the original data positive and negative noise, respectively. $N$ is the number of samples, $j$ is the jth sample.

Step 2: Decompose $Q_j^+$ and $Q_j^-$ to obtain a series of independent periodic IMFs by EMD method.

$$\begin{cases} Q_j^+ = \sum_{i=1}^{m} IMF_{ji}^+ \\ Q_j^- = \sum_{i=1}^{m} IMF_{ji}^- \end{cases} \tag{2}$$

where $IMF_{ji}^+$ is the $i$th IMFs after adding the positive white noise, $IMF_{ji}^-$ is the $i$th IMFs after adding the negative white noise. $m$ is the number of IMFs.

Step 3: Repeat step1 and step2 to obtain the corresponding IMF components, and calculate the average of all the IMFs.

$$IMF_j = \frac{1}{2N} \sum_{i=1}^{N} \left( IMF_{ji}^+ + IMF_{ji}^- \right) \tag{3}$$

where $IMF_j$ is the final component obtained by CEEMD decomposition.

### 2.2. SE

*SE* has been widely applied to analyze IMFs component complexity, and it shows an advantage on data length independence in computing the complexity of time series. So, in the paper, SE is used to analyze the complexity of IMFs obtained by CEEMD decomposition. Based on the SE value, the IMFs are reconstructed into the three parts: noise, cycle, and trend, instead of predicting for each IMF. The definition criteria of noise, cycle and trend is: (1) If $SE_{IMFs} - SE_{Original} > W$ (W is a specified width threshold), then the IMFs is defined as the noise; (2) If $SE_{Original} - SE_{IMFs} > W$, then the IMFs is defined as the trend; (3) If $|SE_{Original} - SE_{IMF}| > W$, then the IMFs is defined as the cycle.

For a given sequence data $\{x(i)\}$, the SE value is calculated as follows:

Step1. Time series is composed in order of $M$-dimensional vectors

$$X(i) = [x(i), x(i+1), \ldots, x(i+M-1)], i = 1, 2, \ldots, N-M+1 \tag{4}$$

Step2. The distance between $X_{M(i)}$ and $X_{M(j)}$, $d_M[X(i), X(j)]$ is calculated as follows:

$$d_M[X(i), X(j)] = \max|x(i+K) - x(j+K)| \tag{5}$$

where $K = 0, 1, \ldots, M-1$.

Step3. Define a number $A_i$, when the $d_M[X(i), X(j)]$ is less than or equal to $R$

$$A_i = num\{d_M[X(i), X(j) \le R]\}, i \neq j \tag{6}$$

where $R$ is a given threshold.

The average value is: