Mining Science and Technology (China) 21 (2011) 153–157

Contents lists available at ScienceDirect

Mining Science and Technology (China)

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

Time-series gas prediction model using LS-SVR within a Bayesian framework

Qiao Meiying [a,b,*], Ma Xiaoping [a], Lan Jianyi [c], Wang Ying [a]

[a] School of Information and Electrical Engineering, China University Mining & Technology, Xuzhou 221008, China
[b] School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China
[c] School of Energy Science and Engineering, Henan Polytechnic University, Jiaozuo 454000, China

ARTICLE INFO

Article history:
Received 24 May 2010
Received in revised form
12 June 2010
Accepted 20 December 2010

Keywords:
Bayesian framework
LS-SVR
Time-series
Gas prediction

ABSTRACT

The traditional least squares support vector regression (LS-SVR) model, using cross validation to determine the regularization parameter and kernel parameter, is time-consuming. We propose a Bayesian evidence framework to infer the LS-SVR model parameters. Three levels Bayesian inferences are used to determine the model parameters, regularization hyper-parameters and tune the nuclear parameters by model comparison. On this basis, we established Bayesian LS-SVR time-series gas forecasting models and provide steps for the algorithm. The gas outburst data of a Hebi 10th mine working face is used to validate the model. The optimal embedding dimension and delay time of the time series were obtained by the smallest differential entropy method. Finally, within a MATLAB7.1 environment, we used actual coal gas data to compare the traditional LS-SVR and the Bayesian LS-SVR with LS-SVMlab1.5 Toolbox simulation. The results show that the Bayesian framework of an LS-SVR significantly improves the speed and accuracy of the forecast.

1. Introduction

Gas disasters are well known and predictions of gas outbursts have become of common concern all around the world. In a recent study, Li and Zhang have proposed a reliable method of non-contact continuous dynamic forecasts [1]. However, due to the complexity of geological conditions, random testing of gas data which contains a considerable amount of random noise, has caused accurate prediction of gas emissions to become a worldwide problem. Since neural networks can approximate any nonlinear function, many early warning models of coal and gas outbursts have been established by means of neural networks. For example, Yang et al. proposed an improved differential evolutionary neural network model to predict coal and gas outbursts [2]. Miao et al. proposed a coal and gas outburst prediction model, combining a neural network with the Dempster-Shafter evidence [3]. Sa et al. set the stage for an electromagnetic radiation and neural network, forecasting coal and gas outbursts [4]. However, neural networks are based on the empirical risk minimization principle; hence generalization performance of model is poor. Furthermore, neural network parameters are artificial and established with a great deal of randomness and are not supported by a relatively prefect theory. Wang et al. analyzed dynamic time series (the Q sequence) of the working face of coal and gas outbursts and calculated the maximum Lyapunov index, the second order Renyi entropy and the correlation dimension. They proved that the Q sequence has chaotic fractal features. Their study provided a theoretical basis for a time-series gas outburst prediction model [5]. A new machine learning algorithm-Support Vector Machine (SVM) has been developed in recent years. An SVM is based on a structural risk minimization principle and can therefore be expected to provide a good generalization performance [6]. The LS-SVR is an improved SVM and was proposed by Suykens and Vandewalle in 1999. It has been used in many fields of engineering. LS-SVR retains structural risk minimization principles, but changes inequality constraints into equations [7]. Thus, quadratic programming problems changed into solving a system of equations, improving the optimization speed of the algorithm. However, for LS-SVR models a regularization parameter and a kernel parameter need to be determined. A traditional LS-SVR uses cross-validation to determine these two parameters, but this method is very time-consuming, especially when the sample is large, which makes the LS-SVR less effective for real-time online prediction.

In our investigation, we have applied a Bayesian evidence framework in order to infer the LS-SVR model parameters. The basic idea of this Bayesian framework is to maximize the parameter distribution a posteriori, so the optimal parameter values of LS-SVR are obtained by maximizing parameter distribution a posteriori [8].

* Corresponding author. Tel.: +86 13403999600.
E-mail address: qiaomy@hpu.edu.cn (Q. Meiying).

1674-5264/$ – see front matter
doi:10.1016/j.mstc.2010.12.012

We used three levels of Bayesian inferences to determine the two model parameters, as well as regularization hyper-parameters and tuned the nuclear parameters by model comparisons. The time series of gas outburst data of the Hebi 10th mine working face was used to validate this model. Simultaneously, the optimal embedding dimensions and delay time of time series were obtained by a smallest differential entropy method [9,10]. In the end and within the MATLAB7.1 environment, we used actual coal gas data to validate the feasibility of the algorithm.

2. Basic principle of LS-SVR

Given a sample set $D = (x_i, y_i), i = 1, \cdots, N$, where $x_i \in R^m$ is the i-th input sample data and $y_i \in R$ the i-th output.

Structural risk minimization is achieved by maximizing the classification interval between samples in the feature space, so that the LS-SVR optimization problem can be written as follows [8]:

$$\min_{\omega,b,e} J_1(\omega, e) = \mu E_W + \zeta E_D = \frac{1}{2}\mu\omega^T\omega + \frac{1}{2}\zeta\sum_{i=1}^{N} e_i^2 \tag{1}$$

$$s\cdot t\cdot y_i = \omega^T\varphi(x_i) + b + e_i, (i = 1, 2\cdots N) \tag{2}$$

where

$$E_W = \frac{1}{2}\omega^T\omega \tag{3}$$

$$E_D = \frac{1}{2}\sum_{i=1}^{N} e_i^2 = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \omega^T\varphi(x_i) - b\right)^2 \tag{4}$$

where $\varphi(\cdot) : R^n \to R^{n_f}$ is the nonlinear mapping function (kernel function), which maps samples into the feature space; $\omega \in R^{n_f}$ is the weight vector, $b \in R$ the bias term, $e_i \in R$ are error variables and μ and ζ are the adjustable hyper-parameters. Functions that satisfy Mercer's theorem can be used as kernel functions. We have opted for the RBF kernel function:

$$K(x_i, x) = \varphi(x_i)^T\varphi(x) = \exp\left(-\|x - x_i\|_2^2/\sigma^2\right)$$

It is difficult to solve ω directly, because ω is in the feature space. Therefore, a solution will be obtained in the dual space. The following Lagrange function is introduced:

$$L(\omega, b, e, \alpha) = J(\omega, e) - \sum_{i=1}^{N}\alpha_i\left\{\omega^T\varphi(x_i) + b + e_i - y_i\right\} \tag{5}$$

where α_i is the Lagrange multiplier.

Because this optimization problem satisfies the Kuhn-Tucker conditions, we can obtain Eq. (6):

$$\begin{cases} \dfrac{\partial L(\omega,b,e,\alpha)}{\partial\omega} = 0 \to \omega = \displaystyle\sum_{i=1}^{N}\alpha_i\varphi(x_i) \\ \dfrac{\partial L(\omega,b,e,\alpha)}{\partial b} = 0 \to \displaystyle\sum_{i=1}^{N}\alpha_i = 0 \\ \dfrac{\partial L(\omega,b,e,\alpha)}{\partial e_i} = 0 \to \alpha_i = \gamma e_i \quad i = 1, \cdots, N \\ \dfrac{\partial L(\omega,b,e,\alpha)}{\partial \alpha_i} = 0 \to \omega^T\varphi(x_i) + b + e_i - y_i = 0 \quad i = 1, \cdots, N \end{cases} \tag{6}$$

where $\gamma = \zeta/\mu$

We can obtain the solution of a least squares regression equation by solving the following set of equations:

$$\begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & \Omega + \frac{1}{\gamma}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{7}$$

where $\Omega_{ij} = \varphi(x_i)^T\varphi(x_j) = K(x_i, x_j), i, j = 1, \cdots, N$; $\alpha = [\alpha_1, \cdots, \alpha_N]$, $y = [y_1; \cdots; y_N]$, $\vec{1} = [1; \cdots; 1]$, Thus, we can obtain α and b by solving Eq. (7). The standard LS-SVR model can be drawn as follows:

$$y(x) = \sum_{i=1}^{N}\alpha_i K(x, x_i) + b \tag{8}$$

The γ and σ values must be pre-determined, when we use the LS-SVR with the RBF kernel function if we wish to make predictions.

3. LS-SVR within Bayesian inferences

3.1. Inference of the model parameters ω and b (Level 1)

Given the data points $D = \{(x_i, y_i)\}_{i=1}^N$ and the hyper-parameters μ and ζ of model H (an LS-SVR model with the RBF kernel function), we can obtain the model parameters ω and b by maximizing a posteriori $P(\omega, b|D, \log\mu, \log\zeta, H)$. In the first level of inference, application of Bayes' rule provides for [11]:

$$P(\omega,b|D,\log\mu,\log\zeta,H) = \frac{P(D|\omega,b,\log\mu,\log\zeta,H)P(\omega,b|\log\mu,\log\zeta,H)}{P(D|\log\mu,\log\zeta,H)} \tag{9}$$

i.e:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}},$$

The evidence $P(D|\log\mu, \log\zeta, H)$ is a constant, which will be used at the second inference. If we assume a priori that $P(\omega, b|\log\mu, \log\zeta, H)$ is independent of the hyper-parameter ζ, then:

$$P(\omega, b|\log\mu, \log\zeta, H) = P(\omega, b|\log\mu, H).$$

If we also assume that ω is independent of b and the weight vector ω is a Gaussian distribution, then:

$$P(\omega|\log\mu, H) = \left(\frac{\mu}{2\pi}\right)^{n_f/2}\exp\left(-\frac{\mu}{2}\omega^T\omega\right).$$

A priori, b can be approximated as a Gaussian distribution, i.e:

$$P(b|\log\sigma_b, H) = \left(\frac{\sigma_b^{-2}}{2\pi}\right)^{1/2}\exp\left(-\frac{b^2}{2\sigma_b^2}\right)$$

With $\sigma_b \to \infty$, we can obtain a priori the following distribution:

$$P(\omega, b|\log\mu, H) = \left(\frac{\mu}{2\pi}\right)^{n_f/2}\exp\left(-\frac{\mu}{2}\omega^T\omega\right)\frac{1}{\sqrt{2\pi\sigma_b^2}} \times \exp\left(-\frac{b^2}{2\sigma_b^2}\right) \propto \left(\frac{\mu}{2\pi}\right)^{n_f/2}\exp\left(-\frac{\mu}{2}\omega^T\omega\right) \tag{10}$$

We assume $P(D|\omega, b, \log\mu, \log\zeta, H)$ is independent of μ. Then:

$$P(D|\omega, b, \log\mu, \log\zeta, H) = P(D|\omega, b, \log\zeta, H)$$

We assume the sample data points are independent of each other. It follows that:

$$P(D|\omega, b, \log\mu, \log\zeta, H) = \prod_{i=1}^{N} P(x_i, y_i|\omega, b, \log\zeta, H) \tag{11}$$

Since $P(x_i, y_i|\omega, b, \log\zeta, H) \propto P(e_i|\omega, b, \log\zeta, H)$ and from (2) we know the error $e_i = y_i - (\omega^T\varphi(x_i) + b)$. If we assume the error has a Gaussian distribution, then: