Computational Materials Science 44 (2008) 647–655

Contents lists available at ScienceDirect

Computational Materials Science

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

Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials

S.F. Fang *, M.P. Wang, W.H. Qi, F. Zheng

School of Materials Science and Engineering, Central South University, Changsha 410083, PR China

ARTICLE INFO

Article history:
Received 4 August 2007
Received in revised form 25 April 2008
Accepted 7 May 2008
Available online 26 June 2008

PACS:
81.05.Bx
81.65.Kn
07.05.Mh
07.05.Kf

Keywords:
Atmospheric corrosion
Genetic algorithms (GAs)
Support vector regression (SVR)
Artificial neural network (ANN)
Zinc
Steel

ABSTRACT

A novel methodology hybridizing genetic algorithms (GAs) and support vector regression (SVR) and capable of forecasting atmospheric corrosion of metallic materials such as zinc and steel has been proposed and tested. Available techniques of data mining of the atmospheric corrosion of zinc and steel are used to examine the forecasting capability of the model. In order to improve predictive accuracy and generalization ability, GAs are adopted to automatically determine the optimal hyper-parameters for SVR. The performance of the hybrid model (GAs + SVR = GASVR) and the artificial neural network (ANN) has been compared with the experimental values. The result shows that the hybrid model provides better prediction capability and is therefore considered as a promising alternative method for forecasting atmospheric corrosion of zinc and steel.

1. Introduction

Atmospheric corrosion is an important process causing deterioration of metallic materials exposed to external environments. In order to study the corrosion behaviors, researchers [1–7] have developed many models of corrosive damages and kinetic equations using multiple linear regression techniques. For example, Feliu et al. [1] have found a general equation, $C = At^n$, to describe the corrosion process, where A stands for the corrosion data of the first year, t is the exposure time in years and C is the corrosion data after t years. The linear regression technique was used to estimate A and n in terms of affecting factors. The correlation coefficients for n are 0.44 and 0.62 for steel and zinc, respectively. However, the result predicted by their proposed linear regression model is still unsatisfactory since the model only considered the linear relationship between the affecting factors. Without considering the nonlinear nature of corrosion behavior, the model has been effective only in very restrictive areas including partial affecting factors rather than those environments including additional important factors or other complex interactions. Such a limitation may either be due to the lack of quality in the corrosion data or the oversimplification of the mathematical model.

On the other hand, the artificial neural network (ANN) has been applied to the study of corrosion [8–13]. The ANN method is more applicable for modeling nonlinear and complex systems, which are hard to be described by physical models. Díaz and López [10] applied an ANN model to estimate the damage function of carbon steel as a function of some environmental variables. The ANN numerical model generates better predictions than the classical linear regression. Cai et al. [12] presented a phenomenological model of the atmospheric corrosion which outperformed the multiple linear regression model of Feliu et al. [1]. Pintos et al. [13] presented an ANN model for the prediction of the corrosion rate of carbon steel as a function of relevant metereochemical variables. However, conventional ANNs still suffer from several weaknesses such as the need for a large number of controlling parameters, the difficulty in obtaining stable solutions, the danger of over-fitting and thus the lack of generalization capability.

Recently, support vector machine (SVM) developed by Vapnik [14–15] has been receiving increased attention with remarkable results. The main difference between conventional ANNs and

* Corresponding author. Tel.: +86 731 8830264; fax: +86 731 8876692.
E-mail address: fang757@163.com (S.F. Fang).

0927-0256/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.commatsci.2008.05.010

SVM lies in the risk minimization principle. Conventional ANNs implement the empirical risk minimization (ERM) principle to minimize the error on the training data, while SVM adheres to the Structural Risk Minimization (SRM) principle seeking to set up an upper bound of the generalization error [16]. The term SVM is typically used to describe classification problems with support vector methods. However, with the introduction of ε-insensitive loss function, SVM has been extended to solve nonlinear regression estimation problems, and a regression version of SVM is also called support vector regression (SVR). SVM has also been applied to solve engineering problems concerning pattern recognition, regression estimation, time-series prediction and inverse solution of dynamic systems [17–25]. SVR has achieved great success both in academic and industrial platforms due to its many attractive features and promising generalization performance.

Evolutionary algorithms (EAs) are the common term used for algorithms based on principles of natural evolution. EAs contain genetic algorithms (GAs), evolution strategies, evolutionary programming and genetic programming (GP). EAs have been applied to SVM in two ways [26]: using GP to evolve kernel functions, and using EAs to evolve kernel parameters. Sullivan and Luke [26] applied GP to find the SVM's optimal kernel function. Howley and Madden [27] utilized GP to evolve a suitable kernel for a SVM. Majid et al. [28] developed Optimal Composite Classifier through the combination of SVM classifiers using GP for gender classification problem. Huang et al. [29] proposed a hybrid GA-SVM strategy capable of simultaneously performing feature selection task and model parameters optimization. de Souza and de Carvalho [30] presented an approach which combined SVM and GAs for a multiclass model selection.

EAs have already been applied for SVM model selection and optimization of parameters in other fields, especially like that of pattern recognition. However, the methods through the combination of EAs and SVM have scarcely been used to solve nonlinear regression estimation problems in materials science. As mentioned above, the key to establishing an efficient SVR model is to choose a proper set of hyper-parameters. However, no effective guidelines have ever been put forward. Some of the recommendations are contradictory and confusing. Therefore, real-value genetic algorithms (RGAs) are adopted to automatically determine the optimal hyper-parameters of SVR with the highest predictive accuracy and generalization ability simultaneously.

In this paper, we have proposed SVR-based model by means of the integration of RGAs and SVR to forecast atmospheric corrosion of zinc and steel. In Section 2, we provide a detailed description of SVR and GASVR models. In Section 3, we describe the data source and experimental settings. Then the results of RGA have been analyzed and the parameters for SVR have been optimized, followed by the discussion of the experimental results in Section 4. In addition, we also compare our method with that of ANN. Finally, we present our conclusions in Section 5.

2. Recurrent support vector machines with genetic algorithms

2.1. Support vector regression

The basic concept of support vector regression is to map nonlinearly the original data x into a higher dimensional feature space and solve a linear regression problem in this feature space [14–17]. First we use a linear function to regress the data set $\{x_i, y_i\}$, $i = 1, 2, \ldots, n$, $x_i \in R^n$, $y_i \in R$. The SVM regression function is

$$f(x) = \langle w, x\rangle + b. \tag{1}$$

The regression problem is equivalent to minimize the following regularized risk function:

$$R(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i) - y_i) + \frac{1}{2}||w||^2, \tag{2}$$

where

$$L(f(x) - y) = \begin{cases} \|f(x) - y\| - \varepsilon & \text{for} |f(x) - y| \geqslant \varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Eq. (3) is also called ε-insensitive loss function. This function defines an ε-tube. If the predicted value is within the ε-tube, the loss is zero. If the predicted value is outside the tube, the loss is equal to the magnitude of the difference between the predicted value and the radius ε of the tube. By substituting the ε-insensitive loss function into Eq. (2), the optimization object becomes:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi + \xi_i^*) \tag{4}$$

with the constraints,

$$\text{subject to} \quad \begin{cases} y_i - \langle w, x_i\rangle - b \leqslant \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \leqslant \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geqslant 0 \end{cases} \tag{5}$$

where the constant $C > 0$ stands for the penalty degree of the sample with error exceeding ε. Two positive slack variables ξ and ξ^* represent the distance from actual values to the corresponding boundary values of ε-tube. A dual problem can then be derived by using the optimization method to maximize the function

$$\text{maximize} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i, x_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*), \tag{6}$$

$$\text{subject to} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \text{ and } 0 \leqslant \alpha_i, \alpha_i^* \leqslant C, \tag{7}$$

where α_i and α_i^* are Lagrange multipliers.

The SVM for function fitting obtained by using the above-mentioned maximization function is then given by

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b. \tag{8}$$

Only parts of α_i and α_i^* have non-zero values. These errors of data points on non-zero coefficients are referred to as the support vectors.

As for the nonlinear cases, the solution can be found by mapping the original problems to the linear ones in a characteristic space of high dimension, in which dot product manipulation can be substituted by a kernel function, i.e. $K(x_i, x_j) = \varphi(x_i)\varphi(x_j)$. In this work, the Gaussian radial basis kernel function (RBF) $\exp\left(-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2\right)$ is used in the SVR. Substituting $K(x_i, x_j)$ for $\langle x_i, x_j\rangle$ in Eq. (6) allows us to reformulate the SVM algorithm in a nonlinear paradigm. Finally, we have

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b. \tag{9}$$

2.2. GA-SVR model

Genetic algorithms are optimization and search technique philosophically based on the concepts of biological evolution (natural genetics and natural selection) and Darwin's theory of survival of the best [31–34]. These algorithms are used to solve linear and nonlinear problems by exploring all regions of state space and exploiting potential areas through mutation, crossover and selective operations applied to individuals in the population [31]. The