Least Squares Support Vector Machine for Gas Concentration Forecasting in Coal Mine

Jian Cheng, Jian-Sheng Qian and Yi-Nan Guo,

School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, 221008, China

Summary
Gas concentration, which is a chaotic time series in essence, is a key factor of the coal mine safety. An accurate forecast of gas concentration is required to guarantee safety and has very highly social and economic benefits. Least squares support vector machine (LS-SVM) has been receiving increasing interest in areas ranging from its original application in pattern recognition to other applications such as regression estimation due to its remarkable generalization performance. In this paper, LS-SVM is a promising method for the forecasting of gas concentration because it uses a risk function consisting of the empirical error and a regularized term which is derived from the structural risk minimization principle. The variability in performance of LS-SVM with respect to the free parameters is investigated experimentally. In addition, this study examines the feasibility of applying LS-SVM in gas concentration forecasting by comparing it with the multilayer back-propagation neural network (BPNN) and the regularized radial basis function neural network (RBFNN). The experimental results show that among the three methods, LS-SVM outperforms the BPNN gas concentration forecasting, and there are comparable generalization performance between LS-SVM and RBFNN, but LS-SVM converges faster than the RBFNN. Finally, LS-SVM provides a promising alternative for gas concentration forecasting.

Key words:
Least squares support vector machine, regression estimation, time series, gas concentration, coal mine.

Introduction

The gas concentration, which is a chaotic time series in essence, is a key factor that endangers the produce in coal mine, so the forecasting of gas concentration is very important for mine safety. Chaotic time series forecasting is to approximate the unknown nonlinear functional mapping of a chaotic signal that is gas concentration in this study [1]. However, most practical time series are of nonlinear and chaotic nature that makes conventional, linear forecasting methods inapplicable. Hence, a number of nonlinear forecasting methods have been developed including neural network exceed conventional methods by orders of magnitude in accuracy [2]. The common neural networks are the multilayer back-propagation neural network (BPNN) and the regularized radial basis function neural network (RBFNN). However, neural network itself, based on the empirical risk minimization principle, is characterized by the pitfalls difficult to overcome. When the number of training data set is small, the precision can not be guaranteed; When the number of training data set is large, this method is significantly limited because of the slow rate of learning and poor generalization [3].

Recently, Support vector machine (SVM), as a novel learning machine developed by Vapnik and his coworkers in 1995 [4], has been proposed as a novel technique in time series forecasting [5]. SVM is a new approach of pattern recognition established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error via the kernel functions and the sparsity of the solution [6]. SVM usually achieves higher generalization performance than traditional neural network in solving many machine learning problems. Another key characteristic of SVM is that training SVM is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVM is always unique and globally optimal [7].

SVM has demonstrated its success in chaotic time series analysis. However, little work has been done in coal mine gas concentration forecasting. Least squares support vector machine (LS-SVM) [8], as a new kind of SVM, is easier to use than usual SVM. The objective of this paper is to propose a LS-SVM method for gas concentration forecasting in coal mine.

This paper is structured as follows: Section 2 provides a brief introduction to LS-SVM for regression. The phase space reconstruction and nonlinear function approximation are given in Section 3. Section 4 presents the results and discussions on the experimental validation. Finally, some concluding remarks are drawn in Section 5.

2. LS-SVM for Regression

The regression problem in SVM is formulated and represented as a convex quadratic programming problem [7]. Basically, the SVM estimator maps the inputs into a higher dimensional feature space in which a linear estimator is constructed by minimizing an appropriate cost function. Using Mercer's theorem, the estimator is obtained by solving a finite dimensional quadratic programming problem in the dual space avoiding explicit

Manuscript received June 5, 2006.
Manuscript revised June 25, 2006.

knowledge of the high dimensional mapping and using only the related kernel function. A nonlinear SVM has the advantage of solving regression problems by convex quadratic programming on the one hand. On the other hand, only support vectors, which usually consist of a small number of training data points, are used. How much the SVM formulation may be simplified without losing any of its advantages?

Motivated by the above idea, Suykens [8] originally proposed a modification to the Vapnik's SVM regression formulation. This method, which adopts the least squares linear equations, this method is easier to use than quadratic programming solvers in SVM method. This new kind of SVM is so-called least squares support vector machine (LS-SVM). LS-SVM uses equality constraints instead of inequality constraints as a least square error term instead of the standard error term.

Given a training data set of N samples $\{x_k, y_k\}_{k=1}^N$ with input data $x_k \in R^n$ and output data $y_k \in R$, one considers the following optimization problem in primal weight space:

$$\min\{J(w,e) = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2\} \quad (1)$$

subject to

$$y_k = w^T\varphi(x_k) + b + e_k, k = 1,\cdots,N \quad (2)$$

With $\varphi(\cdot): R^{n_k}$ a function which maps the input space into a higher dimensional feature space, weight vector $w \in R^{n_k}$ in primal weight space, error variables $e_k \in R$ and bias term b. Note that the cost function J consists of a sum squared fitting error and a regularization term, which is also a standard procedure for the training of multilayer perceptrons's and is related to ridge regression. The relative importance of these terms is determined by the positive real constant γ. In the case of noisy data one avoids over-fitting by taking a smaller γ value.

In primal weight space one has the model

$$y(x) = w^T\varphi(x) + b \quad (3)$$

The weight vector w can be infinite dimensional, which makes a calculation of w from Eq. (3) impossible in general.

Therefore, one computes the model in dual space instead of the primal space. One defines the Lagrangian

$$L(w,b,e,\alpha) = J(w,e) - \sum_{k=1}^{N}\alpha_k\{w^T\varphi(x_k) + b + e_k - y_k\} \quad (4)$$

With Lagrangian multiplier $\alpha_k \in R$, called support values. The conditions for optimality are given by

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \to w = \sum_{k=1}^{N}\alpha_k\varphi(x_k) \\ \frac{\partial L}{\partial e_k} = 0 \to \alpha_k = \gamma e_k, k = 1,\cdots,N \\ \frac{\partial L}{\partial b} = 0 \to \sum_{k=1}^{N}\alpha_k = 0 \\ \frac{\partial L}{\partial \alpha_k} = 0 \to w^T\varphi(x_k) + b + e_k - y_k = 0, k = 1,\cdots,N \end{cases} \quad (5)$$

These conditions are similar to standard SVM optimality conditions, expect for the condition $\alpha_k = \gamma \cdot e_k$. At this point one loose the sparseness property in LS-SVM.

After elimination of w, e one obtains the solution

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & \varphi(x_k)^T\varphi(x_k) + \gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (6)$$

Where $y = [y_1,\cdots,y_N]$, $1_v = [1,\cdots,1]$, $\alpha = [\alpha_1,\cdots,\alpha_N]$.

According to the Mercer's condition, there exists a mapping $\varphi(\cdot)$ and an expansion

$$K(x,y) = \sum_i \varphi_i(x)\varphi_i(y), x, y \in R^n \quad (7)$$

If and only if, for any $g(x)$ such that $\int g(x)^2 dx$ is finite, one has

$$\int K(x,y)g(x)g(y)dxdy \geq 0 \quad (8)$$

As a result, one can choose a kernel $K(\cdot,\cdot)$ such that $K(x_k, x_l) = \varphi(x_k)^T\varphi(x_l)$, $k,l = 1,\cdots,N$. The resulting LS-SVM model for function estimation becomes

$$y(x) = \sum_{k=1}^{N}\alpha_k K(x, x_k) + b \quad (9)$$

Where α_k, b are the solution to the linear system, $K(\cdot,\cdot)$ represents the high dimensional feature spaces that is nonlinear mapped from the input space. The LS-SVM approximates the function using the Eq. (9). In this study, the Gaussian function is used as kernel function

$$K(x, x_k) = \exp(-\|x - x_k\|/\sigma^2) \quad (10)$$

Where σ is a positive real constant. Note that in the case of the Gaussian kernel function, one has only two additional tuning parameters σ^2 and γ in Eq. (10) and in Eq. (1) respectively.

3. The Phase Space Reconstruction and Nonlinear Function Approximation

Chaos occurs as a feature of orbits $x(t)$ arising from systems of differential equations of $dx(t)/dt = F(x(t))$ with three or more degrees of freedom or invertible maps of $x(t+1) = F(x(t))$. As a class of observable signals $x(t)$, chaos lies logically between the well-studied domain

of predictable, regular, or quasi-periodic signal and the totally irregular stochastic signals [9]. In many systems the interaction between the underlying physical processes that are responsible for the evolution of system behavior are unknown. In addition, it is seldom that one has information about all the relevant dynamic variables. Instead, one usually tries to construct a multivariate phase space in which the dynamics unfold using system output measurements of a single time series by converting the time series to a multidimensional phase space. In phase space reconstruction, a scalar time series $\{x_t\}$, $t=1,\cdots,N$, with sampling time Δt, is converted to its phase space using the method of delays:

$$X_t = [x_t, x_{t+1}, \cdots, x_{t+(d-1)}] \quad (11)$$

Where $t = 1,2,\cdots, N-(d-1)T/\Delta t$, d is the embedding dimension, and T is the delay time. In other words, phase space reconstruction techniques convert a single scalar time series to a state-vector representation using the embedding dimension (d) and delay time (T). This reconstruction is required for both characterization and forecasting. In this study, LS-SVM will be employed to capture the dynamics depicted in Eq.(11) with the purpose of producing reliable predictions: $X_{t+T} = f(X_t)$,

$$f:(x_t, x_{t+1}, \cdots, x_{t+(d-1)}) \to (x_{t+T}, x_{t+1+T}, \cdots, X_{t+(d-1)+T}) \quad (12)$$

so $\hat{y} = X_{t+(d-1)+T} = g(x_t, x_{t+1}, \cdots, x_{t+(d-1)})$. The training data consists of a d-dimensional vector, $x \in R^d$, and the response or output, $y \in R$. The goal of the learning machine, then, is to estimate an unknown continuous, real-valued function $g(x)$ that is capable of making accurate predictions of an output y, for previously unseen values of x, thus utilizing information about the dynamics of system behavior in the phase space representation to make forecasts of future system states in observation space. T is the prediction step. In this paper, we try applying LS-SVM to estimate the unknown function $g(x)$.

4. Results and Discussion

The research data used in this study is sampled from a gas sensor in coal mine. This study selects 1700 samples from the industrial field, and the phase space is reconstructed, in which d and T are, respectively, fixed at 10 and 5. The first 1000 samples are used for training the model, and 200 samples are used for validating the model, while the remaining 500 samples are used for testing the identified model only. Through the industrial field empirical knowledge and several trials, the phase space parameters are selected as following: the embedding dimension d=4, and the prediction step T=6.

The learning ability and the generalization of the model can be evaluated by the root mean squared error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (13)$$

where N is the number of sample data; y_i and $\hat{y}_i$ are the i-th real sample value and model predicted value. The smaller the values of *RMSE*, the closer the predicted value to the actual value (a smaller value suggests a better predictor).

When applying LS-SVM to gas concentration forecasting, the first thing that needs to be considered is what kernel function is to be used. As the dynamics of gas concentration time series are strongly nonlinear, it is intuitively believed that using nonlinear kernel functions could achieve better performance than the linear kernel. In this investigation, the Gaussian kernel functions trend to give good performance under general smoothness assumptions. The second thing that needs to be considered is what values of the kernel parameters (γ and σ^2) are to be used. As there is no structured way to choose the optimal parameters of LS-SVM, the values of the parameters that produce the best result in the validation set are used for LS-SVM. Fig.1 and Fig.2 give the *RMSE* of LS-SVM at various γ and σ^2 on training set and validation set, respectively.

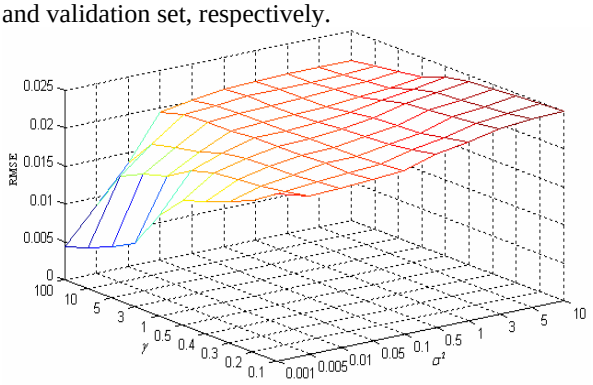

Fig. 1 The behavior of *RMSE* in LS-SVM on the training set.

The Fig.1 shows that the *RMSE* on the training set increases with γ and decreases with σ^2. On the other hand, the *RMSE* on the validation set decreases initially but subsequently increases as σ^2 and γ increase, as shown in Fig.2. This indicates that too small a value of σ^2 (0.001-0.01) and too large a value of γ (10-100) cause LS-SVM to overfit the training data while too large a value of σ^2 (0. 5-10) and too small a value of γ (0.1-0.5) cause LS-SVM to underfit the training data, respectively. An appropriate value for σ^2 would be between 0.01 and 0.5, and an appropriate value for γ would be between 0.5

and 10. In this respect, it can be get that σ^2 and γ play an important role on the generalization performance of LS-SVM, so σ^2 and γ are, respectively, fixed at 0.05 and 5 for following experiments.

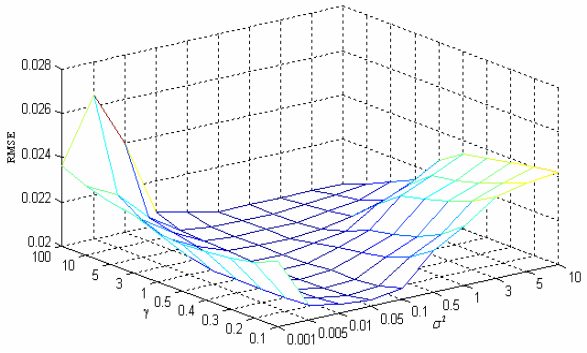

Fig. 2 The behavior of *RMSE* in LS-SVM on the validation set.

A standard three-layer BPNN is used as a benchmark. There are nine nodes in the input layer according to the number of phase space dimension. The output node is equal to 1, and the number of hidden nodes is equal to 10 which is determined based on the validation set. The hidden nodes use the sigmoid transfer function and the output node uses the linear transfer function. The Levenberg-Marquardt learning algorithm is adopted for training the BPNN as it is a kind of improved BP algorithm which has rapid training rate and good performance [2].

In the regularized RBFNN which is also used as the benchmark, the centers, the variances, and the output weights are all adjusted [10]. The number of hidden nodes and the regularization parameters are chosen based on the validation set.

The results are collated and the averages of the best five records obtained in 20 trials on the training and testing data set are given in Table 1. From the table, it can be observed that in all the data points, the largest values of training and testing *RMSE* are in the BFNN. Among the three models, the smallest values of training and testing *RMSE* occurred in LS-SVM, followed by the RBFNN. The training and testing *RMSE* are comparable among the RBFNN and LS-SVM, but CPU time of LS-SVM is 1.594s that is the smallest among three models, meanwhile, the largest CPU time is 25.203s in the RBFNN. The results show that LS-SVM outperform both BPNN and RBFNN with training *RMSE*=0.0187 and testing *RMSE*=0.0194, and the time spent to find the solution is largely less for in LS-SVM than both BPNN and RBFNN with CPU time is equal 1.594s.

Table 1: The results in gas concentration time series

Performance		Models		
		BPNN	RBFNN	LS-SVM
RSME	Training	0.0346	0.0203	0.0187
	Testing	0.0507	0.0215	0.0194
CPU time (s)		6.968	25.203	1.594

Fig.3 and Fig.4 illustrate the predicted and actual values of the LS-SVM in training and testing set, respectively. The solid line is the actual value, and the dotted line is the predicted value of the LS-SVM which has best performance among three models. Obviously, the LS-SVM forecast more closely to actual values than both BPNN and RBFNN models, and there are correspondingly smaller prediction errors in the LS-SVM than the BPNN and RBFNN models, as illustrated in Fig. 5 (The solid and the dotted line are, respectively, the predicted errors of the LS-SVM and the BPNN model), the predicted error in the RBFNN is not shown in Fig.5 as it is close to the predicted error in the LS-SVM.

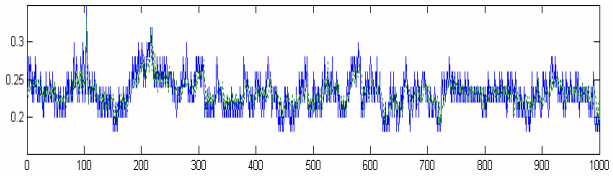

Fig. 3 The predicted and actual values of the LS-SVM in training set.

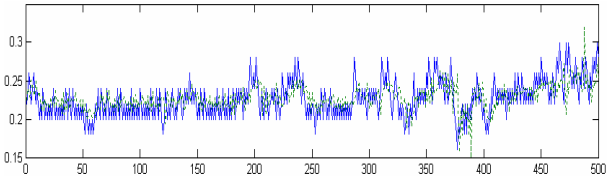

Fig. 4 The predicted and actual values of the LS-SVM in testing set.

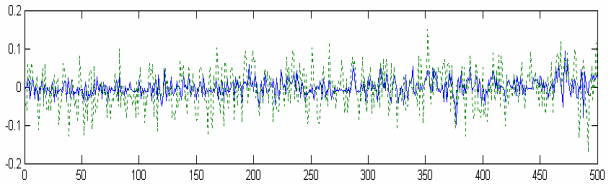

Fig. 5 The prediction errors in the LS-SVM and the BPNN model.

5. Conclusions

The application of LS-SVM in gas concentration forecasting is studied in this paper, and the effect of the value of the kernel parameter (σ^2 and γ) in the LS-SVM was investigated. The experimental result showed that the

prediction performances of LS-SVM are sensitive to the value of these parameters. Thus, it is important to find the optimal value of the parameters.

In addition, this study compared LS-SVM with BPNN and RBFNN. The experimental results show that LS-SVM outperformed BPNN and RBFNN. The results may be attributable to the following facts: On the one hand, LS-SVM implements the structural risk minimization principle which minimizes an upper bound of generalization error rather than minimizes the training error, eventually leading to better generalization performance than the BPNN which implements the the empirical risk minimization principle. On the other hand, BPNN may not converge to global solutions, in the case of LS-SVM, training LS-SVM is equivalent to solving a linearly constrained quadratic programming, and the solution of LS-SVM is always unique, optimal and global. The experiment also shows that there is similar performance between the regularized RBFNN and LS-SVM. The reason lies in the fact that both LS-SVM and the regularized RBFNN minimize the regularized risk function, rather than the empirical risk function as used in the BPNN. Finally, this study concluded that LS-SVM provides a promising alternative for gas concentration forecasting in coal mine.

There will be other research issues which enhance the prediction performance of LS-SVM if they are investigated. The prediction performance may be increased if the optimum parameters of LS-SVM are selected and this remains a very interesting topic for further study. Finally, in this paper only the Gaussian kernel function is investigated, future work needs to explore more useful kernel functions for improving the performance of LS-SVM.

Acknowledgments

This research is supported by National Natural Science Foundation of China under grant 70533050, Postdoctoral Science Foundation of China under grant 2005037225, Postdoctoral Science Foundation of Jiangsu Province under grant [2004]300 and Young Science Foundation of CUMT under grant OC4465.

References

[1] J. D. Farmer and J. J. Sidorowich, Predicting Chaotic Time Series, Phys. Rev. Lett., Vol.59, pp.845-848, 1987

[2] S. Haykin, Neural Networks: a Comprehensive Foundation, Prentice-Hall, New Jersey, 1999

[3] J. Moody and C. Darken, Fast learning in networks of locally tuned processing units, Neural Comput., Vol.1, No.2, pp. 281-294, 1989

[4] V. N. Vapnik, The nature of statistical learning theory, Springer-Verlag, New York, 1995

[5] L. J. Cao, E. H. Francis Tay, Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on neural networks, Vol.14, No.6, pp.1506-1518, 2003

[6] N. Cristianini, J. S. Taylor, An introduction to support vector machines: and other kernel-based learning methods, Cambridge University Press, New York, 2000

[7] V. N. Vapnik, S. E. Golowich, and A. J. Smola, Support vector method for function approximation, regression estimation and signal processing, Adv. Neural Information Processing Syst. 9, pp.281-287, 1996

[8] J. A. K. Suykens, J. Vanderwalle, B. D. Moor, Optimal control by least squares support vector machines, Neural Network, Vol.14, pp.23-35, 2001

[9] T. Asefa, Statistical learning machine concepts and application in water resources management. Ph.D. dissertation, Utah State University, Logan, U. T., 2004

[10] J. Moody and C. Darken, Fast learning in networks of locally tuned processing units, Neural Comput., Vol.1, No.2, pp.281-294, 1989

Jian Cheng received the B.S. and M.S. degrees in China University of Mining and Technology in 1997 and 2003, respectively. Since 2004, he has been a Ph.D candidate in communication and information system at China University of Mining and Technology. The main research objects include machine learning and pattern recognition.

Jian-Sheng Qian is a professor and doctoral supervisor in communication and information system at China University of Mining and Technology. His research interests are in pattern recognition and image processing, optical fiber communication, multimedia system, artificial intelligence and machine learning technologies.

Yi-Nan Guo received the Ph.D. in China University of Mining and Technology in 2003. Since 2004, she has worked in electrical engineering at China University of Mining and Technology as postdoctoral. The main research objects include evolutionary algorithms, multi-agent system and artificial intelligent.