International Journal of Control and Automation
Vol.9, No.2 (2016), pp.219-228
http://dx.doi.org/10.14257/ijca.2016.9.2.21

GA-Support Vector Regression Based Ship Traffic Flow Prediction

Hao Zhang, Yingjie Xiao, Xiangen Bai, XiaojunYang and Liang Chen

Engineering Research Center of the shipping simulation, Ministry of Education, Merchant Marine College, Shanghai Maritime University, Shanghai 201306, China
haozhang@shmtu.edu.cn

Abstract

The observation and forecasting of vessel traffic flow is the foundmental of design for ships' routeing system. An integrated Genetic Algorithm (GA) based Support Vector Machine (SVM) model for vessel traffic flow forecasting with input factors selection procession is presented in this paper. GA based SVM forecasting model is established whose parameters were optimized through genetic algorithms. Finally, the prediction model is used for ningbo-zhoushan port and the prediction result shows that the improved model reflects the actual growth of vessel traffic flow trend more reasonable and effectively.

Keywords*: Intelligent transportation systems; Support Vector Regression; vessel traffic flow; prediction*

1. Introduction

In recent years, ship traffic volume of the entrance channels in most of Chinese ports has been increased rapidly. On one hand, these channels contribute greatly to economic and social development .On the other hand increased volume caused a mass of accidents and loss which requires a higher standard on channel programming and safety management. Forecasting of ship's traffic flow is the science to provide foundation to the channel programming and safety management. Ship traffic volume forecasting is a new domain which combines water traffic engineering with economic forecasting. Due to so many involved factors and indexes, it is more complicated than the general economic forecasting and needs different methods or combinations based on different conditions. Prediction research on ship traffic flow can provide basic data for harbor layout design and management of ship navigation.

The prediction of ship traffic flow is an important fundamental preparation for layout and design of channels as well as management of ship navigation. An intelligent fusion algorithm is applied to ship traffic flow forecasting to remedy the shortcomings in existing ship flow prediction systems, such as low degree of forecasting accuracy and the dependence on experience. Through analyzing the ship flux data gained during the survey in 2007 at Jiagying Bridge over the Changjiang River, the experiment shows that the fusion prediction can forecast multi source data and also reduce the forecast uncertainty so as to increase the accuracy and the robustness of prediction.

Recently, researchers have concentrated on vessels' traffic flow forecasting with ANN because of the following two advantages: one is capability of approximating any nonlinear function and the other is model determination through the learning process. Owing to the outstanding merits of ANN, ANN based model is suitable for solving traffic flow forecasting problem in maritime domain, since it can properly represent the complex nonlinear relationships that exist between load and a series of factors that influence it.

ISSN: 2005-4297 IJCA

A modified combination forecasting model is set up for forecasting ship traffic flow, based on harbor characteristics, ship behavior and historical data of ship traffic flow to remedy shortcomings of present ship traffic flow prediction systems such as low precision and dependence on experience. Through analyzing the nine-year ship flux data of Tianjin harbor, results show that this modified combination forecasting model can reduce uncertainty in prediction so as to enhance precision and stability of the whole forecasting system.

Based on the study of the characteristics of vessel traffic flow, taking the vessel traffic flow control as the ultimate goal, the vessel traffic flow prediction model based on the BP (Back Propagation) neural network was setup. Taking the Yangtze estuary deepwater channel's traffic flow data as training samples, the simulation analysis was carried out. Comparing with the forecasting data and the measured data weighted, it showed that the model is valid for the prediction of vessel traffic flow.

With forecast time span reducing, its non-linearity, time-dependent nature and uncertainty become more and more strong. The orthodox prediction models such as History Average Model, Time-Series Model, Nonparametric Regressive Model, Kalman Filtering Model, Neural Network Model and Combination Prediction Model cann't predict well in effect and precision.

2. Forecasting Models

2.1. Support Vector Regression

Recently SVM which was developed by Vapnik (1995) is one of the methods that are receiving increasing attention with remarkable results. The main difference between ANN and SVM is the principle of risk minimization. While ANN implements empirical risk minimization to minimize the error on the training data, SVM implements the principle of Structural Risk Minimization by constructing an optimal separating hyperplane in the hidden feature space, using quadratic programming to find a unique solution. When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM, and how to set the best kernel parameters. These two problems are crucial, because the feature subset choice influences the appropriate kernel parameters and vice versa (Fro¨hlich and Chapelle, 2003).

Considering a set of training data $\{(x_1, y_1),....,(x_\ell, y_\ell)\}$, where each $x_i \subset R^n$ denotes the input space of the sample and has a corresponding target value $y_i \subset R$ for $i=1,..., l$ where l corresponds to the size of the training data. The idea of the regression problem is to determine a function that can approximate future values accurately.

The generic SVR estimating function takes the form:

$$f(x) = (w \cdot \Phi(x)) + b \qquad (1)$$

where $w \subset R^n$, $b \subset R$ and Φ denotes a non-linear transformation from R^n to high dimensional space. Our goal is to find the value of w and b such that values of x can be determined by minimizing the regression risk:

$$R_{reg}(f) = C\sum_{i=0}^{\ell}\Gamma(f(xi) - yi) + \frac{1}{2}\|w\|^2 \qquad (2)$$

where $\Gamma(\cdot)$ is a cost function, C is a constant and vector w can be written in terms of data points as:

$$w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\Phi(x_i) \qquad (3)$$

By substituting equation (3) into equation (1), the generic equation can be rewritten as:

$$f(x) = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*)(\Phi(x_i) \cdot \Phi(x)) + b$$

$$= \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad (4)$$

In equation (4) the dot product can be replaced with function $k(x_i, x)$, known as the kernel function. Kernel functions enable dot product to be performed in high-dimensional feature space using low dimensional space data input without knowing the transformation Φ. All kernel functions must satisfy Mercer's condition that corresponds to the inner product of some feature space. The radial basis function (RBF) is commonly used as the kernel for regression:

$$k(x_i, x) = \exp\left\{-\gamma \left| x - x_i \right|^2\right\} \qquad (5)$$

Some common kernels are shown in Table 1. In our studies we have experimented with these three kernels.

Table 1.Common Kernel Functions

Kernels	Functions
Linear	$x \cdot y$
Polynomial	$\left[\left(x * x_i\right)+1\right]^d$
RBF	$\exp\left\{-\gamma \left| x - x_i \right|^2\right\}$

The ε-insensitive loss function is the most widely used cost function. The function is in the form:

$$\Gamma(f(x) - y) = \begin{cases} |f(x) - y| - \varepsilon, & for \quad |f(x) - y| \geq \varepsilon \\ 0 & otherwise \end{cases} \qquad (6)$$

By solving the quadratic optimization problem in (7), the regression risk in equation (2) and the ε-insensitive loss function (6) can be minimized:

$$\frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i, x_j) - \sum_{i=1}^{\ell} \alpha_i^* (y_i - \varepsilon) - \alpha_i (y_i + \varepsilon)$$

subject to

$$\sum_{i=1}^{\ell} \alpha_i - \alpha_i^* = 0, \quad \alpha_i, \alpha_i^* \in [0, C] \qquad (7)$$

The Lagrange multipliers, α_i *and* α_i^*, represent solutions to the above quadratic problem that act as forces pushing predictions towards target value y_i. Only the non-zero values of the Lagrange multipliers in equation (7) are useful in forecasting the regression line and are known as support vectors. For all points inside the ε-tube, the Lagrange multipliers equal to zero do not contribute to the regression function. Only if the requirement $|f(x) - y| \geq \varepsilon$ (See Figure 1) is fulfilled, Lagrange multipliers may be non-zero values and used as support vectors.

The constant C introduced in equation (2) determines penalties to estimation errors. A large C assigns higher penalties to errors so that the regression is trained to minimize error with lower generalization while a small C assigns fewer penalties to errors; this allows the

minimization of margin with errors, thus higher generalization ability. If C goes to infinitely large, SVR would not allow the occurrence of any error and result in a complex model, whereas when C goes to zero, the result would tolerate a large amount of errors and the model would be less complex.

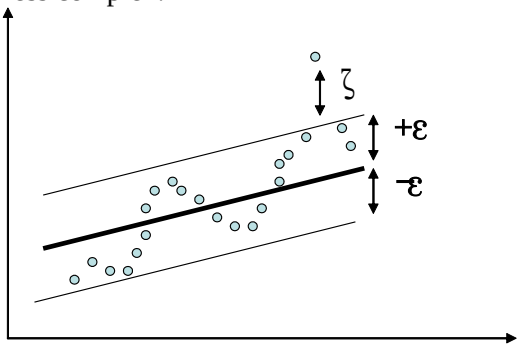

Figure 1. Support Vector Regression to Fit a Tube with Radius ε to the Data and Positive Slack Variables ζ_i Measuring the Points Lying Outside of the Tube

Now, we have solved the value of w in terms of the Lagrange multipliers. For the variable b, it can be computed by applying Karush-Kuhn-Tucker (KKT) conditions which, in this case, implies that the product of the Lagrange multipliers and constrains has to equal zero:

$$\begin{aligned} \alpha_i(\varepsilon + \zeta_i - y_i + (w, x_i) + b) = 0 \\ \alpha_i^*(\varepsilon + \zeta_i^* + y_i - (w, x_i) - b) = 0 \end{aligned} \tag{8}$$

and

$$\begin{aligned} (C - \alpha_i)\zeta_i = 0 \\ (C - \alpha_i^*)\zeta_i^* = 0 \end{aligned} \tag{9}$$

where ζ_i and ζ_i^* are slack variables used to measure errors outside the ε-tube. Since $\alpha_i, \alpha_i^* = 0$ and $\zeta_i^* = 0$ for $\alpha_i^* \in (0, C)$, b can be computed as follows:

$$\begin{aligned} b = y_i - (w, x_i) - \varepsilon \quad for\ \alpha_i \in (0, C) \\ b = y_i - (w, x_i) + \varepsilon \quad for\ \alpha_i^* \in (0, C) \end{aligned} \tag{10}$$

Putting it all together, we can use SVM and SVR without knowing the transformation.

2.2. GA-Based Feature Selection and Parameters Optimization of SVR Models

For support vector machine model, three free parameters (σ, ε and *C*) greatly affect the performance of the SVM models, so the big problem is how to select the values of parameters that will allow good performance. Selecting appropriate values for parameters of SVM plays an important role in the performance of SVM. However, structured methods for selecting parameters are lacking, moreover it is not known beforehand which values are the best for the problem. Optimizing the parameters of SVM is crucial for the best prediction performance. In this paper, the genetic algorithm is adopted to determine free parameters of support vector machine. GA is a directed random search technique which is widely applied in optimization problems where the number of parameters is large and the analytical

solutions are difficult to obtain. GA can help finding the optimal solution over a domain globally.

The main procedures are as follows:

The chromosome design, fitness function, and system architecture for the proposed GA-based feature selection and parameter optimization are described as follows.

2.2.1 Chromosome, Gene and Population Design

GA starts with a set of random solutions called population. The individual solution in population is called chromosome which is represented in encoded form and each chromosome consists of a series of individual structures called genes. In the case of parameters optimizing for SVR, genes correspond to the hyper parameters to be tuned, such as C, s. Consequently, chromosome X is the parameters vector; where xi represents the ith parameter and s denotes the size of the parameters vector to be tuned.

Genetic operators. Based on the principle of natural evolution, new solutions are generated from the previous generation of population. Three basic operators, i.e. selection, crossover and mutation, are involved.

Selection is used to realize the principle of survival of the fittest by keeping and deleting some solutions from the existing population. The solutions selected will be used to form the parent population to generate offspring. Operators crossover and mutation are used to generate new solutions (offspring)from the current solutions(parent population).Crossover is done by exchanging some digits of two individuals and mutation is done by changing some digits of one individual.

2.2.2 Fitness Function

To realize selection, the solutions are selected according to their values of objective function called fitness. In the case of parameters optimizing for SVR, fitness corresponds to the generalization performance. Several measurement indicators could be used here such as MSE, MAPE and ME. In our experiments, we adopt average relative variance to make the results comparable. We can calculate them with following equations:

(Average Relative Variance: ARV), definition:

$$ARV = \frac{\sum_{i=1}^{N}[x(i) - \hat{x}(i)]^2}{\sum_{i=1}^{N}[x(i) - \bar{x}(i)]^2}$$

Among: N -number; $x(i)$ -actual data; $\bar{x}$ -average actual measurement; $\hat{x}(i)$ -prediction.

Normalization

Pn= 2(p-minp)/(maxp-minp)

$$MSE = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here

Step 1. Step1. Encoding
SVM parameters and initialization, establish randomly an initial population of chromosomes. I.e. generate randomly a group SVM parameters and encoding the SVM parameters by adopting certain encoding program, thus forming initial population.

Step 2. Step2. Evaluating fitness

Train SVM through taking each chromosome's gene value as SVM parameters, training set as input and output set. After finishing, the chromosome's fitness value should be evaluated. The fitness value for each chromosome was calculated according to a negative normalized root mean square error (NRMSE) shown as follows:

Where i a and i f represent the actual and forecasting values respectively, and n is the number of forecasting periods.

Step 3. Step3. Selection operation, crossover operation, and mutation operation

Selection is performed to select excellent chromosomes to reproduce. Based on fitness function, chromosomes with higher fitness values are more likely to yield offspring in the next generation by means of the roulette wheel. Crossover is performed randomly to exchange genes between two chromosomes. The mutation operation follows the crossover operation, and determines whether a chromosome should be mutated in the next generation. Offspring replaces the old population and forms a new population in the next generation by the three operations, the evolutionary process proceeds until stop conditions are satisfied.

Among: p is the data sample, Pn is the normalized value, maxp and minp are the max value and min value of the data sample. The data sample are converted to 0 ~ 2.To achieve high forecasting accuracy and speed, it is important to identify the main factors affecting the load. The forecasting is primarily dependent on five explanative variables such as No. of foreign ships in and out of Ningbo port, total foreign trade of Ningbo port, GDP of Ningbo port, the container TEUs and the port cargo throughput. The data from year 1991 to year 2007 is set as training sets and the data of year 2008 and 2009 are set as testing sets. By using the Matlab software, (c,g) are set [1,1000; 0.01,20], population scale is set to 20,selection is 0.08,crossover is 0.8,mutation is 0.05, the max evolution generation is 50.g=0.125,c=11.3137.

To evaluate the resulting the presented model's performance, the root mean square error (RMSE) is calculated for the predicted.

For proposed GA-SVM forecasting model, the RMSE is about 294.75. For BP forecasting model, the corresponding error parameters are about 799.86 and 4.25%. So the RSME obtained by the GA-SVM model are better than the SPO-SVM model and SVM model. In all the GA-SVM model for short term load forecasting has been implemented successfully and it can give satisfactory results.

Table 2. Compare of Different Forcasting Method

Year	RMSE			predicted	actual
	SVM	SPO-SVM	GA-SVM		
2012	87.93%	92.54%	95.47%	896217	875485
2013				1021731	1095874

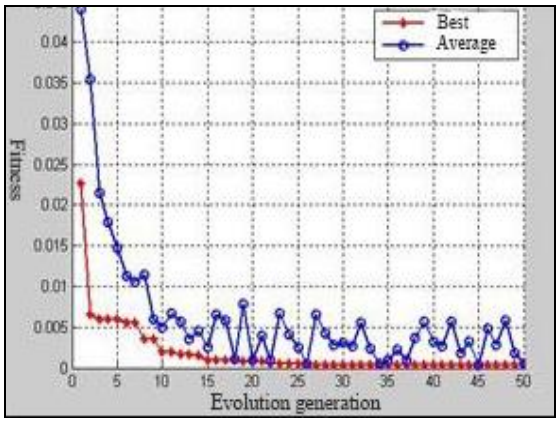

Figure 2. Curve of Fitness

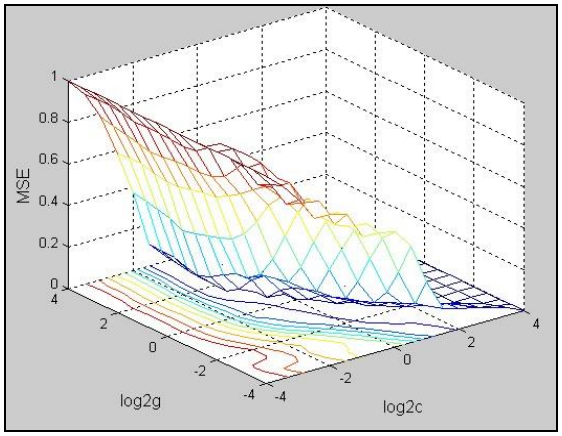

Figure 3. 3-D Contour Map of Parameters G, C and Mse

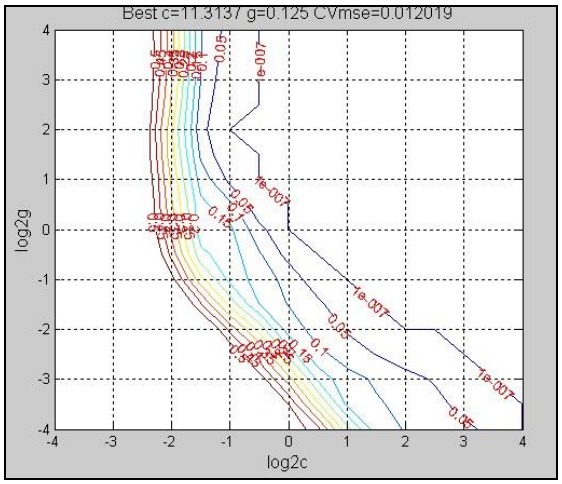

Figure 4. Contour Map of Parameters G, C

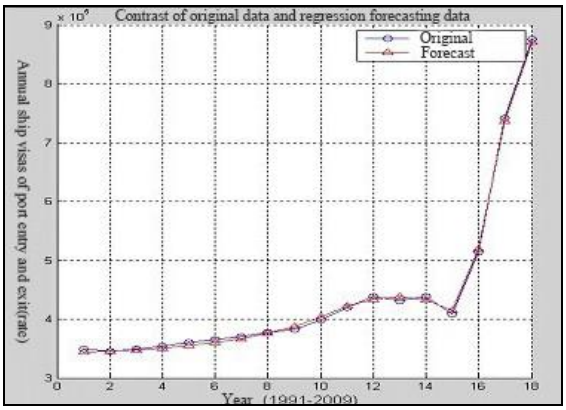

Figure 5. Comparison of Actual Data and Predicted Data

5. Conclusion and Discussion

This paper integrated the SVM and GA techniques to optimize the prediction of vessel's traffic flow. The proposed GA-SVM system has the following characteristics distinct from previous work. First, the proposed system employs No. of foreign ships in and out of Ningbo port, total foreign trade of Ningbo port, GDP of Ningbo port, the container TEUs and the port cargo throughput data from real statistics of Ningbo Maritime Safety Administration (MSA), instead of using simulation data. Second, it adopts GA to adjust the values of parameters of the SVM for enhancing the prediction accuracy. Lastly, this is the first application of SVM for predicting vessels traffic flow to our best knowledge.

After the experiments, we summarized four critical contributions. First, SVM model training process is highly important, since it affects the accuracy of the VMS system. A SVM model is used to improve the prediction accuracy by the kernel function. Consequently SVM model for novelty prediction could achieve lower classification errors compared with the CSFM, and Intelligent Fusion methods. Second, the experiment results show that both GA methods increase the learning performance of SVM model and enhance the prediction accuracy in our proposed system. Finally, we apply empirical study to indicate that the proposed approach can generate appropriate prediction accurately of traffic flow based on the SVM approach. Furthermore, experimental results illustrate that with the same number of input data sets, the proposed method can achieve a slightly better prediction performance compared to those with SVM and PSO-SVM.

Acknowledgements

This work is supported by National Natural Science Foundation of China (51149001).

References

[1] E. Avci, "Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm–support vector machines: HGASVM", Expert Systems with Applications, vol. 36, no. 2, **(2009)**, pp. 1391-1402.
[2] H. Huang, T.H. Tang and Y.X. Jin, "Ship Traffic Flow Forecast System Based on Intelligent Fusion", Navigation of China, vol. 31, no. 4, **(2008)**, pp. 364-367.
[3] J. Lu and X.L. Fang, "Composite Systematic Forecasting Model and M ethod for Vessel Traffic Flow Forecasting", Journal of Dalian Maritime University, vol. 22, no. 2, **(1996)**, pp. 33-35.
[4] Y.C. Tian and J.B. Chen, "Vessel Traffic Flow Prediction Based on BP Neural Network [J]", Ship & Ocean Engineering, vol. 39, no. 1, **(2010)**, pp. 122-125.
[5] S.W. Fei, C.L. Liu and Y.B. Miao, "Support vector machine with genetic algorithm for forecasting of key-gas ratios in oil-immersed transformer", Expert Systems with Applications, vol. 36, no. 3, pp. 6326-6331.

[6] J.T. Jeng, "Hybrid approach of selecting hyperparameters of support vector machine for regression", IEEE Transactions on Systems, Man and Cybernetics Part B, vol. 36, no. 3, **(2006)**, pp. 699-709.
[7] W. He, Z. Wang and H. Jiang, "Model optimizing and feature selecting for support vector regression in time series forecasting", Neurocomputing, vol. 72, no. 1-3, **(2008)**, pp. 600-611.
[8] "A model updating strategy for predicting time series with seasonal patterns", Applied Soft Computing Journal, vol. 10, no. 1, **(2010)**, pp. 276-283.
[9] "GMRVVm-SVR model for financial time series forecasting", Expert Systems with Applications, vol. 37, no. 12, **(2010)**, pp. 7813-7818**.**

Authors

Zhang Hao, He is a PHD from Shanghai Maritime University major in maritime safety, transportation information engineering and control.