

Combinatorial Methodology Study for Network Traffic Prediction

Xianmin Wei

*School of Computer Engineering, Weifang University
5147 Eastern Dongfeng Street, Weifang 261061, China
wfxuweixm@126.com*

Abstract

Network traffic is a typical time-series data with strong lag and aftereffect, for the existence of local optimum, time-consuming and other defects in the method for the currently determining number of lags, this paper presents a combination of network traffic prediction method (GS-GA-LSSVM). At first, using geo-statistics (GS) to quickly determine the optimal lag order of network traffic, then reconstructing network traffic with the lag order, and finally using genetic algorithm (GA) to optimize least squares supporting vector machine (LSSVM) to model predictions for network traffic. Simulation results show that, GS-GA-LSSVM on network traffic prediction accuracy is better than any of the participating models, and which can better reflect the complex dynamics of network traffic disciplines.

Keywords: *Network traffic, Geo-statistics, Least squares supporting vector machine, Genetic algorithm*

1. Introduction

With the great growth of network size and business type, network traffic becomes a nonlinear, time-varying dynamical systems. In order to meet the expanding needs of network traffic, high-speed networks need to be effectively managed, while network management is based on network traffic prediction, high-precision prediction model can not only predict future traffic data, but also can be applied to access control, network bandwidth allocation and congestion control, and many other aspects of the network [1].

For problems in network traffic prediction, scholars have made a lot in-depth study and proposed many network traffic prediction models. Traditional network traffic prediction models include linear regression analysis, gray model and time series, etc.[2,3], these methods are based on linear modeling, but network traffic is a complex nonlinear system, therefore, traditional linear model cannot describe the characteristics of modern changes in network traffic accurately, prediction results are unreliable [4]. In recent years, with the development of nonlinear theory, nonlinear network traffic prediction models have appeared based on neural networks, supporting vector machines [5]. As network traffic is a typical time-series data, which has characteristics of hysteresis and aftereffect, therefore, in network traffic prediction, it is necessary to determine the order of optimal lag network traffic and model parameters. Traditional lag order was determined by linear regression method, it is difficult to find the optimal lag order [6]. Since then some scholars have proposed a number of fixed-order nonlinear methods to improve the prediction performance of the model, but these methods are very time-consuming, and affect the efficiency of network traffic forecast [7]. Based on the statistical theory, supporting vector machine algorithm (least square support vector machine, LSSVM) has solved nonlinear problems, and overcome the defect of local optimum and low operating efficiency of traditional machine learning algorithms, and

become a major network traffic prediction method. Predictable performance of LSSVM is sensitive to the choice of parameters, but so far, because lack of good guidance of LSSVM parameters selection, in practice most of the parameters were empirically determined, due to improper parameter selection network traffic, prediction accuracy is low [8].

For the existing problems in current network traffic prediction, this paper proposes a combination of network traffic prediction algorithm (GS-GA-LSSVM). At first, using geo-statistics (GS) on the analysis of correlation degree between network traffic data, and quickly determining the optimal lag order of network traffic time series; then reconstructing network traffic data according to optimal lag order into LSSVM study, and using genetic algorithms (GA) to optimize the parameters for LSSVM to establish the optimal network traffic prediction model. Finally through concrete examples to verify the performance of the model [15].

The remainder of the paper is organized as follows: Section 2 describes the review of geo-statistics and least square supporting vector machine, while Section 3 presents the detailed network flow forecast model of GS-LSSVM. Then, Section 4 describes the evaluation results and discussion of GS-LSSVM in network traffic prediction, and finally Section 5 provides the conclusions.

2. Review of Geo-statistics and Least Square Supporting Vector Machine

2.1. Geo-statistics(GS)

Geo-statistics is also known as geological statistics, which is a new and gradually formed branch of mathematical statistics, based on a large number of theoretical research by the famous French statistician G. Matheron.

It is based on the regional variables as the basis, by means of variation function, the randomness and structural feature, a science or spatial correlation and dependence of natural phenomena. All with the spatial data structure and randomness, or spatial correlation and dependence, or spatial pattern and variability of relevant research, and these data were optimal unbiased on interpolative estimation or simulation, discrete, the volatility of these data, the application of statistical theory and methods.

Common statistics and classical statistics show that they are based on a large amount of sampling through the analysis of the sample frequency of attribute values or mean variance relationship, distribution and its relevant rules determine the relationship between the spatial distribution pattern. But the biggest characteristic is different from the classical statistics, e.g. statistics considers the size of sample values, the sample location and the distance between samples, and makes up for the shortcomings of classical statistics ignoring the spatial orientation.

The theoretical basis of statistical analysis of assumptions includes regional variables, analysis of variance and spatial estimation.

GS is a statistical approach based on regionalized variable theory, which uses variable function to describe dual characteristics of structure and randomness of variables [9]. For the observation of time series data $Z(x_i), i = 1, 2, \dots, n$, and mutation function value $\gamma(h)$ is calculated as below.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (1)$$

Where $z(x_i)$ and $z(x_i + h)$ are observed values of $z(x)$ at x_i and $x_i + h$.

In the second-order stationary assumptions, the covariance function can be expressed as follows.

$$C(h) = E[Z(x)Z(x+h)] - m^2 \quad (2)$$

For different time interval distance h , according to Eq.(2), $\gamma(h)$ can be calculated as corresponding variance function value, with h as the horizontal axis, $\gamma(h)$ as vertical axis, the variance function curve (Fig.1) can be obtained.

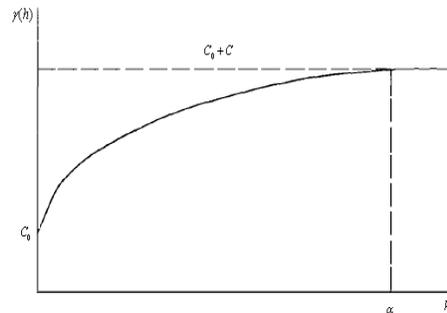


Figure 1. Variance Function Curve

In Figure 1, $\gamma(h)$ increases with the rise of h , when h is beyond a certain distance, $\gamma(h)$ is stable around a limit, where the variation function value is called a base station, represented with the C_0+C , and this distance is called the interval variable range, expressed with α . Y-axis intercept of variance function curve is called area discontinuity value, expressed by C_0 .

2.2. Least Squares Supporting Vector Machines

For a linear function in N dimension space, supporting vector machine, its VC dimension $N+1$. The VC dimension constraints may be greatly reduced to ensure good generalization, even in very high dimensional space can also be smaller than VC dimension set. At the same time, the original problem is converted into a dual problem, the computational complexity is no longer dependent on space dimension, but depends on the number of samples, especially the number of support vectors in each sample. These characteristics effectively dealing with high dimensional problems become possible.

For the nonlinear problem, through the nonlinear transform into linear problem in a high dimensional space and in the transform space.

The optimal classification face, this transformation can be complex, so this idea in general is not easy to achieve. In the dual problem above, both optimizing the classification function involves only the training samples.

The inner product, in the high dimension space, in fact only the inner product operation, and the inner product operation can be realize the function of the original space, we don't even need to know the transformation form. According to the theory of function, only a kernel function satisfies condition, it is corresponding to an inner product in the transform space.

Dimension and complexity of supporting vector machines is independent of the input space, and relies on the number of sample data, therefore, greater the number of samples, more complex and slower computing speed the corresponding quadratic planning problem, which

limits the application range of supporting vector machines [10]. Suykens, et al. proposed the least squares support vector machine (LSSVM) on the basis of standard supporting vector machines, where the loss function in standard supporting vector was set as error squared sum, and the inequality constraints was changed into equality constraints, reducing the pending parameters, and the quadratic planning problem was turned into linear KKK (Karush Kuhn Kucker) solving equations, which reducing the complexity of solving problems, broadening the application space of supporting vector machine.

For the training set, $\{(x_i, y_i)\}, i = 1, 2, \dots, n$, x_i and y_i represents sample input and output, respectively, $x_i \in R^n, y_i \in R^n$, through the nonlinear mapping function $\varphi(\square)$, samples are mapped into a high dimensional feature space, thereby obtaining optimal linear regression function:

$$f(x) = w^T \varphi(x) + b \tag{3}$$

Where, ω is the weight value of the feature vector space, b is the offset.

According to the structural risk minimization principle, LSSVM regression model for solving problem in Eq.(3) is:

$$\begin{aligned} \min \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^n \zeta_i^2 \\ \text{s.t. } y_i - w^T \varphi(x) + b = e_i, (i = 1, 2, \dots, n) \end{aligned} \tag{4}$$

Where, γ is the penalty parameter for balancing training error and model complexity, e_i is the error between the actual value and the regression function.

By introducing Lagrange multipliers, the above constrained optimization problem can be changed into a dual space unconstrained optimization problems, namely:

$$\begin{aligned} L(w, b, \zeta, \alpha) = \min \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^n \zeta_i^2 + \\ \sum_{i=1}^n \alpha_i (w^T \varphi(x) - b + e_i - y_i) \end{aligned} \tag{5}$$

Where, α_i is the Lagrange multiplier.

According to Mercer conditions, the nuclear-defined functions are defined as follows:

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \tag{6}$$

In this study, radial basis function as LSSVM kernel function, radial basis function is:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{7}$$

Finally, LSSVM regression model is:

$$f(x) = \sum_{i=1}^N \alpha_i \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) + b \tag{8}$$

Where, σ represents the width of radial basis auditors.

2.3. Parameter Optimization in Least Squares Supporting Vector Machines

LSSVM generalization performance in practical application depends on the selection of parameters γ and σ to a considerable extent. Error penalty parameter γ is the role of regulating the ratio of learning machines and the confidence range in the determining data subspace, to make the best generalization ability of learning machines, different data subspace has different optimal γ . In determining data subspace, small γ represents the small value of the small empirical error, the small complexity of the learning machine and the larger value of the empirical risk, and vice versa. The former is called "less learning" phenomenon, while the latter is called "over learning". σ insensitive loss function determines the size of the number of supporting vectors. Larger the number σ , less the number of supporting vectors machines, the lower function estimation accuracy; smaller σ , more number of supporting vectors machines, higher function the estimation accuracy. However, less σ is not better, because although the accuracy is improved, the time required for the algorithm becomes longer, so the most reasonable σ must be done to ensure speed and accuracy of LSSVM, only the correct choice of parameters can make LSSVM obtain a good fit regression estimation results [16].

How to choose LSSVM parameters is not yet effective. Usually through cross-validation trial or gradient descent method to solve, because of too many methods or human factors, or requiring continuous conduction function, it is easy to fall into local minima. Genetic Algorithm (GA) is a heuristic search algorithm, which has a strong global search capability, through genetic operators to simulate such as replication, crossover and mutation phenomenon in natural biological process of genetic, on behalf of the population of individuals by merit, and thus ultimately best individual are obtained, so in this study GA is used to optimize the parameters of LSSVM [17].

Basic steps of GA for LSSVM parameter optimization are as follows.

(1) Assume P as the population scale. Initial population $w = (w_1, w_2, \dots, w_p)^T$ randomly generated with P individuals, given a selected range of data, using linear interpolation function to generate a real number vector w_1, w_2, \dots, w_x of individuals w_i in population as a chromosome of GA, in order to obtain high-precision γ and σ values, using real numbers coding method.

(2) Determine the individual's modest function. Given a LSSVM parameter, decoding the chromosome in step (1), to obtain γ and σ values, to input training samples into LSSVM for training, to achieve output value of the model satisfying the accuracy, to treat the training error square sum as the fitness of individual w_i in the population of W.

(3) Using the roulette selection operator, which is based on proportional fitness selection strategy for the population in each generation to choose chromosomes, choosing probability:

$$p_i = f_i / \sum_{i=1}^P f_i, i = 1, 2, \dots, P \quad (9)$$

Where, f_i is the reciprocal of fitness value, P is the population scale.

(4) Due to using real number-coded for individuals, crossover method uses the same real-number crossover approach, the crossover operation at j bit for the k-th gene w_k and the l-th gene w_l are respectively given below:

$$\begin{cases} w_{kj} = w_{kj}(1 - b) + w_{lj}b \\ w_{lj} = w_{lj}(1 - b) + w_{kj}b \end{cases} \quad (10)$$

Where b is a random number between 0 and 1.

(5) Mutation operation, selecting the j -th gene bit of the i -th individual for mutation in Eq.(11).

$$w_{ij} = \begin{cases} w_{ij} + (w_{ij} - w_{\max}) f(g), & r \geq 0.5 \\ w_{ij} + (w_{\min} - w_{ij}) f(g), & r < 0.5 \end{cases} \quad (11)$$

$$f(g) = r_2(1 - g / G_{\max})$$

In Eq. (11), w_{\max} and w_{\min} , respectively, upper and lower bound of genetic value, r is a random number between 0 and 1, r_2 is a random number, g is the current iteration number, G_{\max} is the maximum evolution generation.

(6) Best individual of GA is decomposed of γ and σ of LSSVM, these γ and σ values were LSSVM prediction model, the LSSVM prediction model was trained to predict the optimal value of network traffic in time series.

3. Network Flow Forecast Model of GS-LSSVM

3.1. Idea of GS Combining with LSSVM

During determining lag order of network traffic prediction model based on LSSVM, to describe network traffic data correlation by GS, where the data falls within the scope of the change process α are correlated with time, and with the data points increasing the time interval to reduce the correlation degree, which can quickly determine the optimal lag order based on α value for network traffic prediction model, then according to the optimal lag order to reconstruct network traffic samples, finally, input the reconstructed samples to LSSVM, in order to establish a network traffic prediction model. Figure 2 shows the specific process.

3.2. Data Normalization

Subject to a variety of factors, relatively large changes are in the scope of network traffic, which adversely affects the LSSVM learning speed, to eliminate the adverse effects of improving learning efficiency, network traffic values can be normalized, the normalized formula is as follows:

Wherein x_i is the initial value of the network traffic, x_i' is the normalized flow rate of the network, x_{\max} and x_{\min} are the maximum and minimum values in network streams, respectively.

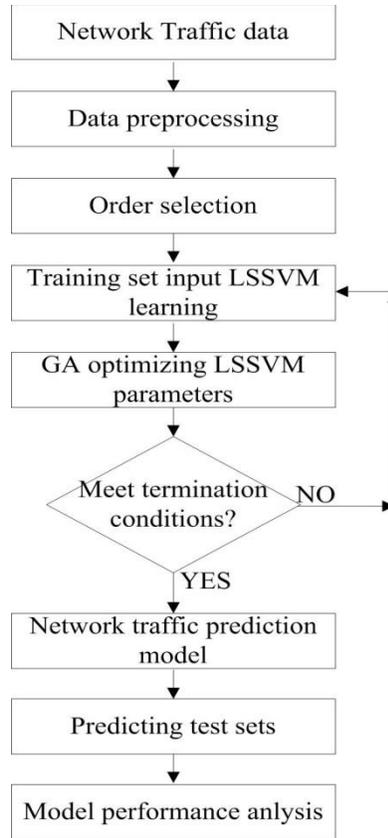


Figure 2. Workflow of Network Traffic Prediction Model

3.3. LSSVM Model Order Determined based on GS

Network traffic has obvious structural characteristics of time, current network traffic usually associates with previous network traffic at multiple time points, data in closer time has strong correlation, and time correlation is weak for distant data, so the network traffic has lag and aftereffect characteristics. Lag order n has great influence on network traffic prediction: even though the too small order can reflect the correlation between data, but which could easily lead to missing data; If the order is too large, it is possible to select more adjacent data, but may contain too much redundant information, and therefore the lag order cannot be too small or too large. Lag order is determined as follows in network traffic prediction model based on GS:

Suppose $z(x_i)$ of $n, i=1,2,\dots,n$, and $(p < n)$ is set to the unchecked point, then take the first $p-1$ samples as the analysis data, take hour as the time interval, *i.e.*, $\Delta h = 1$; distance between the samples is $h_{i,j} = j - i$, to ensure each number sufficiently large for $N(h)$ is, in GS the

analytical requirement specified: $h \leq \frac{1}{2} \max(h_{i,j})$.

According to sample distant h and halve varied function value of $\gamma(h)$ to construct function graph, observing h in largest $\gamma(h)$ value corresponding to the process, *i.e.*, changed process α . According to the known GS principle, samples within the untested point α period have relation with the untested points, *i.e.*, the sample within period α has great impact on unmeasured data, the model lag order is α .

3.4. LSSVM Pre-modeling Network Flow after Reconstruction

To determine the optimal lag order of time series in network traffic by using GS, and according to the optimal order to reconstruct network traffic data; then reconstructed data was divided into two parts of training set and test set, input the training set to LSSVM for learning, using GA to establish optimal network flow model for optimizing LSSVM parameters, using the model established in the last to predict test set to obtain network traffic prediction results, and the predicted results were analyzed.

4. Evaluation Results and Discussion of GS-LSSVM in Network Traffic Prediction

4.1. Data Sources

Experimental data came from network traffic Library, <http://newsfeed.ntcu.net/~news/20011/>, which collected a total of 208 days of network traffic per hour, 4992 data form master node router “Incoming articles” from January 1, 2011, as shown in particular in Figure 3. Figure 3 shows that network traffic has strong nonlinear, time variability. The data is divided into two parts, wherein the last 1000 as the test set, the remaining data as the training set.

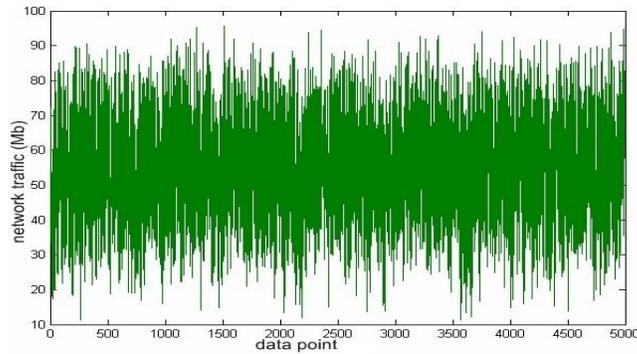


Figure 3. The Original Network Traffic Data

4.2. Evaluation Index and Compared Models

In order to evaluate advantages and disadvantages of GS-GA-LSSVM model performance, choosing GS-LSSVM (LSSVM parameters using grid search) compared with GA-LSSVM (lag order in [11]) and traditional LSSVM (variable order, LSSVM parameters using grid search) algorithm. Model performance index evaluation system is: using Mean squared error (MSE) and Mean absolute percentage error (MAPE). MSE and MAPE are defined as follows:

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n} \quad (12)$$

$$MAPE = \frac{\sum |y_i - \hat{y}_i| / y_i}{n} \times 100 \quad (13)$$

Where y_i is the actual value, \hat{y}_i is the model predicted values and n is the number of test samples [12].

4.3. Model Implementation

Firstly, using GS to have an analysis on network traffic data, and the results of the 100 previous data change process is shown in Figure 4, which shows the optimal lag order of network traffic data is 3[13], which indicates that the former three time network traffic has impact on network traffic of prediction point, which means that the network traffic of the former three points is input to predict current time network traffic of LSSVM[14], using this method to reconstruct network traffic, to form a multi-dimensional network traffic time series.

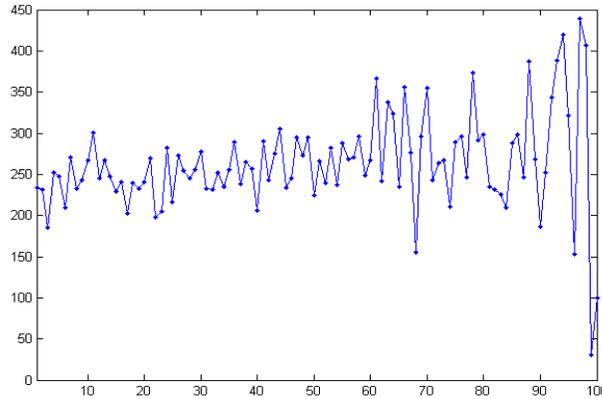


Figure 4. Determining the Optimal Lag Order of Network Traffic

In MATLAB7.0, self-programming and call LSSVM GA toolbox to achieve LSSVM algorithm, kernel function is Gaussian, the training set input to LSSVM for training, and using genetic algorithms to optimize the parameters of LSSVM to obtain the optimal parameters $\gamma = 29.15$ and $\sigma = 0.75$; then adopting the optimal parameters for the training set to re-learn, to establish the optimal network traffic prediction model; Finally, the prediction model to predict the test sample set, the prediction results are obtained in Figure 5.

4.4. Performance Compared with other Models

Prediction results of MAPE and MSE using LSSVM, GS-LSSVM, GA-LSSVM and GS-GA-LSSVM to set of test samples are shown in Table 1.

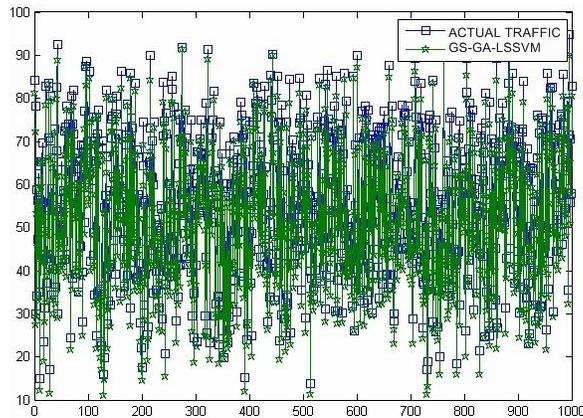


Figure 5. Prediction of GS-GA-LSSVM

Table 1. Prediction Results Compared with Various Models

MODEL	MSE	MAPE(%)
LSSVM	30.83	14.70
GS-LSSVM	21.06	10.85
GA-LSSVM	19.88	10.15
GS-GA-LSSVM	12.19	6.18

The following conclusions are drawn from the analysis of comparison results in Table 1.

(1) Either MSE or MAPE, prediction performance in GA-LSSVM are superior to that in LSSVM, which indicates that GA optimized parameters of LSSVM to find the optimal LSSVM parameters to improve network traffic prediction accuracy.

(2) At the same time, the prediction accuracy in GS-LSSVM is higher than in LSSVM, which indicates that using lag order of GS to quickly identify network traffic can be a good way to reflect the relevance of the data to fully tap the hidden network traffic information between data.

(3) In all forecast models, GS-GA-LSSVM has the maximum prediction accuracy, indicating that when predicting the network traffic time series of nonlinear, time-varying, not only considering the correlation between the data, but taking into account the parameters optimization problem of non-linear prediction model.

It is common from the above that GS-GA-LSSVM combines the advantages of time-series analysis and regression analysis, which can better reflect the data correlation with minimal structural risk, nonlinear, avoiding over-fitting, dimensionality curse, local minimization, excellent promotion ability and other advantages, so that the model prediction has higher performance and better stability.

5. Conclusions

Introduction of GS with data correlation analysis and global search ability of GA into measuring the network traffic, to respectively determine the lag order and optimal parameters by varying process of the model, taking into account the characteristics of nonlinear changes in network traffic, introduction of non-linear predictive ability of the LSSVM model to improve network traffic prediction accuracy. Simulation example shows that compared with the traditional network traffic prediction method, GS-GA-LSSVM improves the prediction accuracy of the network traffic, the model can be extended to the field of nonlinear prediction.

This study is only for a one-dimensional time-series network traffic forecast, but in fact, the network traffic is closely related to a variety of factors, which is a typical multi-dimensional time series, so with extent of the forecast time, GS-GA-LSSVM model prediction accuracy will be affected, therefore, to consider multiple factors and replenish new data to correct forecasting model to improve the prediction accuracy of network traffic, which needs further research.

Acknowledgements

This work is partly supported by Shandong Natural Science Foundation Program (ZR2011FL006), Shandong Science and Technology Development Program (2011YD01044), Shandong Spark Program (2012XH06005), and Weifang municipal Science and Technology Development Program (201301050).

References

- [1] T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification Using Machine Learning", *IEEE Communications Surveys and Tutorials*, vol.10, no. 4, (2008).
- [2] B. Gao, Q. Zhang and Y. Liang, "Self-similar Network Traffic Forecast based on EMD and ARMA", *Communications Journal*, vol. 32, no. 4, (2011).
- [3] T. Lei and Z. Yu, "Wavelet Neural Network model for Network Traffic Prediction", *Journal of Computer Applications*, vol. 26, no. 3, (2006).
- [4] Y. Jin, G. Zhou, "Theoretical Mean-variance Relationship of IP Network Traffic based on ON / OFF Model", *China Ser F-Inf Sci.*, vol. 52, no. 4, (2009).
- [5] S. Wang and Z. Qiu, "Combining Multi-fractal Network Traffic Nonlinear Prediction", *Communications Journal*, vol. 28, no. 2, (2007).
- [6] Y. Miao and W. Yong, "Based on Support Vector Regression Machine Learning Network Traffic Prediction", *Guilin Institute of Technology*, vol. 27, no. 2, (2007).
- [7] C. Xiang, "Time Series Forecasting Model based on Support Vector Machine", Changsha: Hunan Agricultural University, (2012).
- [8] A. Esta, F. Gringolif and L. Salgarelli, "Support Vector Machines for TCP Traffic Classification", *Computer Networks*, vol. 53, no. 14, (2009).
- [9] H.L. Qi, H. Zhao, W.W. Liu and H. B. Zang, "Parameters Optimization and Nonlinearity Analysis of Grating Eddy Current Displacement Sensor Using Neural Network and Genetic Algorithm", *Journal of Zhejiang University Science A*, vol. 10, no. 8, (2009).
- [10] J. Wang and Z. Gao, "Network Traffic Prediction based on RBF Neural Network Modeling", *Computer Engineering and Applications*, vol. 44, no. 13, (2008).
- [11] Z. Yuan, Y. Zhang and J. Xiong, "Multivariate Time Series Analysis and its Application based on SVR in the Agricultural Sciences", *Agricultural Sciences of China*, vol. 41, no. 8, (2008).
- [12] T. Stotzle and H. Hoos, "The MAX-MIN ant System and Local Search for the Traveling Salesman Problem, *IEEE International Conference on Evolutionary Computation and Evolutionary Programming*, Indianapolis, (1997) October 13-19; Seattle, USA.
- [13] B. Bullnheimer, R. F. Hartl and C. C Strauss, "A New Rank-based Version of the Ant System: A Computational Study", *Central European Journal for Operations Research and Economics*, vol. 7, no. 1, (1999).
- [14] O. Cordón, F. Herrera, I. Fernández, Ez de Viana and L Moreno, "A New ACO Model Integrating Evolutionary Computation Concept: The Best-Worst Ant System", *Lecture Notes in Computer Science*.1838, (2000).
- [15] X. Tao, X. Zhou and T. Jeffrey, "Designing Heterogeneous Sensor Networks for Estimating and Predicting Path Travel Time Dynamics: An Information-theoretic Modeling Approach", *Transportation Research Part B: Methodological*, vol. 57, (2013).
- [16] K. Issa, K. Abdallah, A. Faheem and S. Khaled, "Dependable Wireless Sensor Networks for Reliable and Secure Humanitarian Relief Applications", *Ad Hoc Networks, PART A*, vol. 13, (2014).
- [17] L. Christoph, K. Dirk, B. Andreas, S. Holger, B. Nico, V. Dirk, L. Heiko and G. Andreas, "Energy Efficiency of Load-adaptively Operated Telecommunication Networks", *Journal of Lightwave Technology*, vol. 32, no. , (2014).

Author



Xianmin Wei. He received the M. Sc. degree in computer applications from Shandong Science and Technology University (2005). He is currently an associate professor in school of computer engineering at Weifang University, China. He has published over 30 papers and 3 books in professional fields. Since 2011, he has been a member of IEEE-CS, ACM and CCCF, respectively. His fields of research are focused on swarm intelligent, intelligent sensor networks.

