International Journal of Smart Home
Vol. 10, No. 3, (2016), pp.139-148
http://dx.doi.org/10.14257/ijsh.2016.10.3.14

# A Method for Missing Data Recovery of Air Pollutants Monitoring in Henhouse Based on QGSA-SVM

Jinming Liu[1,2], Qiuju Xie[1], Guiyang Liu[1] and Yong Sun[2]

[1]*College of Information Technology, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang 163319, China*
[2]*College of Engineering, Northeast Agricultural University, Harbin Heilongiang 150030, China*
*jinmingliu2008@126.com*

## *Abstract*

*To solve the data missing problem caused by sensor faults during the air pollutants monitoring in henhouse, a method for missing data recovery was proposed based on support vector machine (SVM). Multiple factors that influence monitoring values of the air pollutants in henhouse, such as temporal, spatial and environmental, were considered to established a SVM regression model to estimate the missing data of the air pollutants monitoring. Meanwhile, to obtain better prediction accuracy, regression model parameters were optimized by a novel hybrid optimization algorithm which was combined standard genetic algorithm with quantum genetic strategy and simulated annealing tactics. Taking the data processing of the ammonia (NH3) concentration as an example, the proposed method was tested with the monitoring data of 3 days in a farm. The estimation results of missing data shown that there was a litter error between the estimated data and monitoring data, the maximal relative error was 5.87% (percent), the average relative error was 1.77% (percent). It is verified that this method of missing data recovery is feasible and valid.*

**Keywords:** *Genetic algorithm, quantum genetic, simulated annealing, support vector machine, henhouse, air pollutants monitoring, data recovery*

## 1. Introduction

At present, China's poultry husbandry has been greatly improved in the aspects of variety, nutrition and disease control level, but there are constantly many problems in environment control. Especially, air pollutants including ammonia (NH3), particulate matter (PM), odor, and pathogens, emitted from egg production represent risk to the health of animals, poultry workers and neighbors, and to the global environment [1, 2]. Air pollutants must be accurately measured, and the factors that influence pollutants emissions must be synthetically analyzed. Therefore, a continuous and reliable monitoring air pollutant is very important to controlling and disposing pollutants [3, 4]. In the henhouse, multiplex sensors of pollutants concentration need to be installed for monitoring the pollutants concentrations constantly in order that the air pollutants emissions rules is analyzed. However, the complex environment in henhouse may cause sensors to be shifted or damaged, which causes the deviation or completely error of monitoring data [5, 6]. The missing data of the air pollutants monitoring need to be recovered so that the integrity and accuracy of monitoring data can be ensured. However, multiple factors that influence the air pollutants concentration in henhouse such as temporal, spatial and environmental are interrelated and interacted on each other. The missing data recovery of air pollutants monitoring is a complex nonlinear problem, there is a large error between the estimated value and the actual value by using the linear interpolation method. The neural network algorithm has been adopted to estimate the missing data of air pollutants monitoring, and obtained better estimation results of the missing data. But there are many deficiencies in the neural network itself, such as local minimum value problems and over-learning problems.

Support vector machine (SVM) is a machine learning method with good generalization ability which is based on the statistical learning theory of small sample data and the principle of structural risk minimization [7]. SVM solves the disadvantages of neural network, which is effectively dealt with all kinds of nonlinear problems and widely used for solving various regression prediction problems [8]. The prediction accuracy of SVM depends directly upon its own parameters, so the parameters of SVM have attracted increasing attention on regression prediction problems. For requiring a lot of computation time by using the grid search parameters optimization methods, the relevant scholars have put forward the intelligent optimization algorithm to optimize the parameters of SVM, such as particle swarm optimization (PSO) [9] and genetic algorithm (GA) [10]. Among them, GA has a good ability of robustness and global optimization, so it is fit for solving the complex optimization problems. However, GA still suffers from two drawbacks, premature convergence and poor search ability in the late evolution.

In this paper, a novel hybrid optimization algorithm is proposed by combining standard GA with quantum genetic strategy and simulated annealing tactics which is called quantum genetic simulated annealing algorithm, QGSA for short. In the population evolution of QGSA, the quantum code and genetic evolution are introduced to keep the population diversity of GA in order to avoid the premature convergence, and the fitness function of GA is designed to combine with the temperature parameters of simulated annealing algorithm (SA) to improve search efficiency in the late evolution. A method for missing data recovery of air pollutants monitoring in henhouse is presented based on SVM regression model combined with QGSA (QGSA-SVM for short) which parameters are optimized by using QGSA, and tested by using randomly selected monitoring data of 3 days in a farm.

## 2. SVM Theoretical Basis

The goal of SVM regression is to find a function in order that it can predict the corresponding dependent variable from other independent variable except the samples after trained. Namely, the regression function is sought as follows.

$$f(x) = (w^T x) + b \tag{1}$$

Where, $w$ is the weight, $b$ is the threshold value. The regression function $f(x)$ is to minimize the following objective function.

$$\min\left(\frac{1}{2}\|w\|^2 + c \cdot R_{emp}\right) \tag{2}$$

Where, $c$ is the penalty factor, and $R_{emp}$ is the training error.

The basic idea of SVM nonlinear regression is a nonlinear transforms which is mapped the original nonlinear problem to a linear problem in high dimensional eigenspace. And then, the solution of the original problem is obtained by linear regression in high dimensional eigenspace. The nonlinear transforms are achieved by defining an appropriate inner product function. In the high dimensional eigenspace, the kernel function can be used in place of the inner product operation of linear problems. The frequently used kernel functions include linear kernel, polynomial kernel, radial basis function (RBF) kernel, sigmoid kernel and so on. In the literature [11], the RBF Kernel function SVM prediction model has displayed remarkable advantages, and the prediction accuracy is the highest. Therefore, the RBF kernel function is adopted as the kernel function of the SVM prediction model, and the function expression is shown as follows.

$$K(u,v) = \exp(-\gamma \| u - v \|^2) \tag{3}$$

Where, $\gamma = 1/2\sigma^2$ is a parameter of the kernel function, $\sigma$ is the width parameter, $u$ is any point in space, and $v$ is the center point.

In this paper, the SVM prediction model of missing data recovery is designed and achieved by using the LibSVM software package. The epsilon-support vector regression (epsilon-SVR for short) is adopted as the type of SVM, and RBF Gaussian kernel is adopted as the kernel function. The SVM parameters that need to be optimized include the penalty parameter $C$ , the kernel function parameter $\gamma$ and the insensitive loss function parameter $\varepsilon$ .

## 3. Missing Data Recovery Based on QGSA-SVM

### 3.1. Input-Output Parameters of SVM

Multiple factors that influence the air pollutants concentration in henhouse, such as temporal, spatial and environmental, are synthetically considered to established a multiple input and single output prediction model of SVM to estimate the missing data of the air pollutants monitoring. The missing data of the air pollutants monitoring for a certain time are recovered by using the SVM regression prediction model. The multiple input parameters of SVM prediction model are as follows: First, the monitoring data of the air pollutants concentration at the previous sampling instant of the missing data sampling point; secondly, the change of the air pollutants concentration in neighboring sampling point of adjacent sampling instant; third, the monitoring value of ambient temperature, relative humidity and wind speeds in the missing data sampling point. The single output value is the estimate of the air pollutants concentrate in the missing data sampling points. Training the SVM prediction model through the long time continuous monitoring data, the trained SVM model could become a good estimator of the missing data which contains the nonlinear relationship between input variables and the output variable.

### 3.2. Optimization of SVM Parameters Based on QGSA

While SVM is used for solving regression problems, choosing the appropriate SVM parameters is an important factor affecting the prediction accuracy of SVM. In this paper, the SVM parameters $C$ , $\gamma$ and $\varepsilon$ are optimized based on QGSA combined with K-cross-validation. In the process of optimization, the QGSA is composed of two parts. The first part is the quantum code genetic evolution process (QGA for short), another part is the genetic evolution process combined with simulated annealing tactics (GSA for short). The QGA is responsible for the population initialization and the generation of perturbation solution set, and the genetic evolution of simulated annealing in GSA is completed. The diversity of the population is extended by the perturbation solution of QGSA, which effectively avoids the premature convergence, and QGSA can effectively improve the search efficiency of the algorithm in the late evolution.

**3.2.1. Coding and Population Initialization:** When using QGSA to optimize the SVM parameters, the multi-bit quantum encoding is adopted as the encoding mode of the initial solution. Three SVM parameters $C$ , $\gamma$ and $\varepsilon$ are respectively represented three genes of chromosome, each gene is encoded into $k$ bit quantum encoding. The multi-bit quantum encoding structure of chromosome can be expressed as follows.

$$P = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} & \alpha_{21} & \cdots & \alpha_{2n} & \cdots & \alpha_{m1} & \cdots & \alpha_{mn} \\ \beta_{11} & \cdots & \beta_{1n} & \beta_{21} & \cdots & \beta_{2n} & \cdots & \beta_{m1} & \cdots & \beta_{mn} \end{bmatrix} \tag{4}$$

Where, $(\alpha_{mn},\beta_{mn})$ is a probability amplitude of the quantum states, and $|\alpha_{mn}|^2+|\beta_{mn}|^2=1$ , $m=1,2,3$ , $n=1,2,\cdots,k$ . $(\alpha_{mn},\beta_{mn})$ can be set as $(1/\sqrt{2},1/\sqrt{2})$ when the population is initialized.

In the process of evolution, the multi-bit quantum encoding is still adopted as the encoding mode of QGA, and initial population of QGSA is adopted as initial population of QGA. The binary number encoding is adopted as the encoding method of GSA, each chromosome is encode into the binary sequence $a_1a_2\cdots a_kb_1b_2\cdots b_kc_1c_2\cdots c_k$ , the initial population of GSA is obtained by the initial population of QGSA after a quantum probability collapse. The binary sequence $a_1a_2\cdots a_k$ , $b_1b_2\cdots b_k$ and $c_1c_2\cdots c_k$ are respectively the gene codes of parameter $C$ , $\gamma$ and $\varepsilon$ . The corresponding real decoding formula for each gene can be expressed as follows.

$$f(x)=\left(\sum_{i=1}^{k}w_i\cdot 2^{i-1}\right)\cdot\frac{(U_2-U_1)}{2^k-1}+U_1 \tag{5}$$

Where, $w_i$ is the binary gene bit of optimization parameters in chromosome, $[U_1,U_2]$ is the range of optimization parameters, and $k$ is the length of the single gene of binary code. In this paper, $k=20$, the encoding chromosome length of QGA and GSA is 60 bit.

### 3.2.2. Design of Fitness Function:

The purpose of the SVM prediction model is making the error between actual values and estimated values as small as possible, so the mean square error (MSE) of the K-cross-validation is adopted as the objective function of QGSA. Obviously, the smaller objective function value is the better prediction accuracy of the prediction model. The objective function of QGSA is directly adopted as the objective function of QGA, and the fitness function of GSA is defined as follows.

$$fit(x)=\exp\left(-\frac{f(x)-f_{\min}}{t}\right) \tag{6}$$

Where, $f(x)$ is the objective function value of current chromosome, $f_{\min}$ is the minimum of the objective function in current generation population, and $t$ is the temperature parameter in current generation.

The adjusted fitness function of GSA is designed by combining with the temperature parameter, which makes the algorithm to calculate the little difference of the fitness function value at high temperature (the early evolution), effectively prevent premature convergence by avoiding the whole population filled with the individual good chromosome. The good individual has a relatively large fitness function value at low temperature (the late evolution), more easily could be passed on to the next generation, which improves the convergence speed of the algorithm.

### 3.2.3. Design of Genetic Operation

The genetic evolution of QGSA consists of two parts: the quantum genetic evolution of QGA and the simulated annealing genetic evolution of GSA.

The quantum rotation gate update strategy is adopted as quantum genetic evolution based on literature [12], and ensures that the direction of evolution minimizes the MSE of QGA population. Some excellent individuals are selected as the perturbation solution set in each generation population of QGA.

The genetic operation of GSA includes selection, crossover and mutation operations.

The roulette wheel method is adopted as the selection operation of GSA combined with the optimal maintaining strategy, the single point crossover strategy is adopted as the crossover operation, and the multi-bit mutation is adopted as the mutation operation.

### 3.2.4. GSA Initial Temperature and Annealing Parameters

The initial temperature is defined by using the formula $t_0 = M\delta$ , where $\delta = f_{\max} - f_{\min}$ , $f_{\max}$ and $f_{\min}$ are respectively the maximum and minimum of the objective function value in initialization population. The formula $t_{n+1} = \alpha t_n$ is adopted as annealing function, where $0 < \alpha < 1$ . The speed of annealing operation can be controlled by adjusting the value of $\alpha$ .

### 3.2.5. Steps of Optimizing SVM Parameters by QGSA

The concrete steps that QGSA optimizes the parameters of SVM regression model are as follows:

Step1: According to the population initialization method, $popSize$ chromosomes of multi-bit quantum encoding are generated as the initial population of QGSA, and binary initial population of GSA is obtained by the initial population of QGSA after a quantum probability collapse.

Step2: The real-number values of the parameters $C$ , $\gamma$ and $\varepsilon$ are respectively calculated by decoding each binary chromosome in population of GSA, the values of objective function and fitness function are calculated by combining with K-cross-validation. Set the generation gap to $GGAP$ , $GGAP \times popSize$ new chromosomes are generated based on the fitness function values. A new generation population is generated by substituting the new chromosomes for the chromosomes of the original population in turn which has the minimum fitness value.

Step3: After probability collapse and real decoding for the population of QGA, the objective function MSE of each chromosome is calculated. The perturbation solution set is constructed by selecting $(1 - GGAP) \times popSize$ excellent individuals from the binary population of QGA after collapsing. The perturbation solutions are added to the population of GSA by using the method in Setp2. Then the new population of QGA is generated by using the quantum genetic evolution based on the objective function.

Step4: If the end condition is satisfied, stop, and return the best values of the regression SVM parameters. Otherwise go to Setp2; continue to optimize the SVM parameters.

### 3.3. Prediction of Missing Data by SVM

When completed the SVM parameters optimization based on QGSA, the SVM prediction model is established by the optimized parameters $C$ , $\gamma$ and $\varepsilon$ . After trained the SVM prediction model, the test set is adopted to evaluate the prediction model by the estimated accuracy of the missing monitoring data. However, when using the prediction model for estimation, the sample attribute of air pollutants concentration in the monitoring data at the previous sampling instant should be an estimated value of air pollutants at the previous sampling instant. That is to say, the property values of current moment are predicted by other attributes of current moment combined with the estimation of the previous sampling instant. This is a typical time series prediction problem which needs disposing by some special methods.

To sum up, the algorithm flowchart of the SVM prediction model is given which is used to recover the missing data of the air pollutants monitoring. The algorithm flowchart is shown in Figure 1.

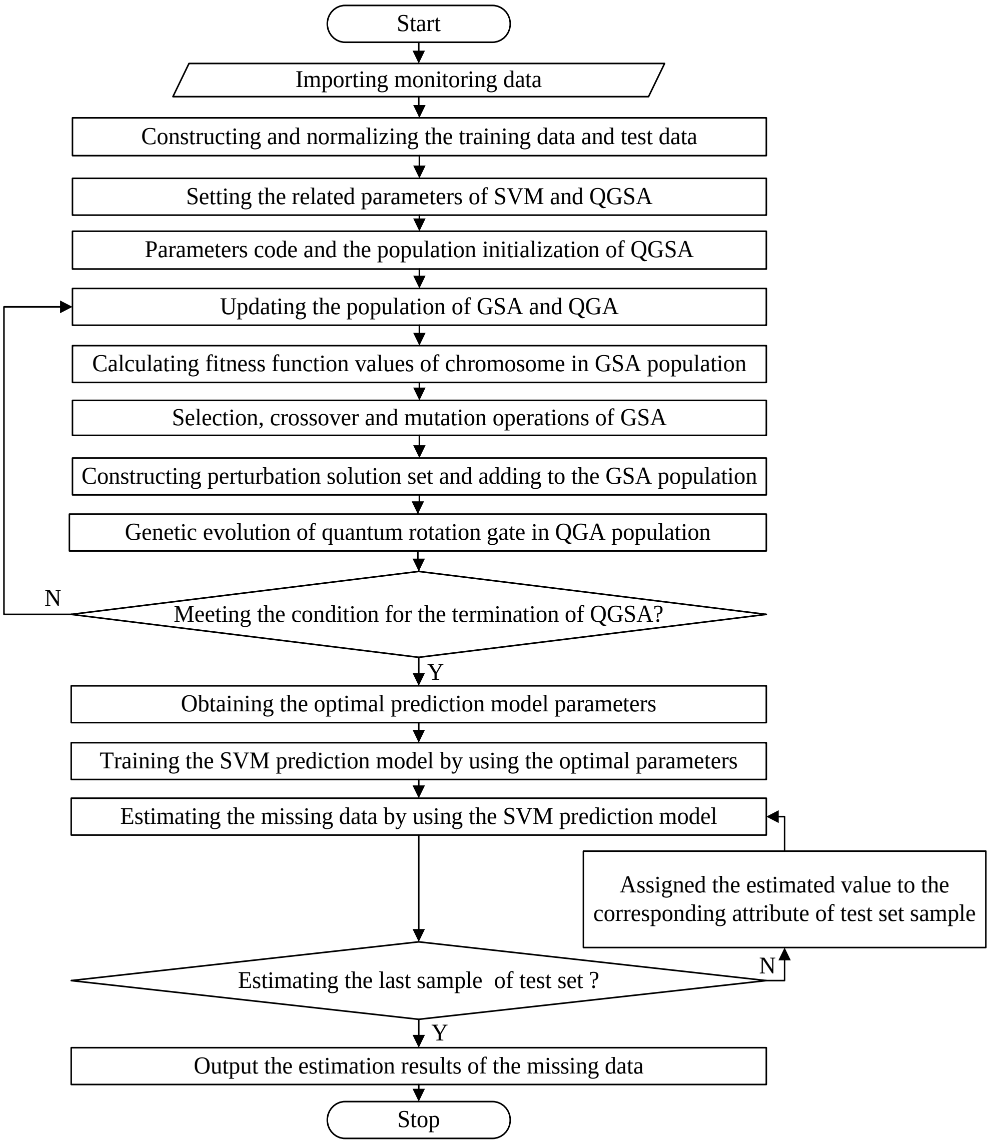
Start
Importing monitoring data
Constructing and normalizing the training data and test data
Setting the related parameters of SVM and QGSA
Parameters code and the population initialization of QGSA
Updating the population of GSA and QGA
Calculating fitness function values of chromosome in GSA population
Selection, crossover and mutation operations of GSA
Constructing perturbation solution set and adding to the GSA population
Genetic evolution of quantum rotation gate in QGA population
N
Meeting the condition for the termination of QGSA?
Y
Obtaining the optimal prediction model parameters
Training the SVM prediction model by using the optimal parameters
Estimating the missing data by using the SVM prediction model
Assigned the estimated value to the corresponding attribute of test set sample
N
Estimating the last sample of test set ?
Y
Output the estimation results of the missing data
Stop

**Figure 1. Flowchart of Missing Data Recovery**

## 4. Experimental Simulation and Analysis

In this paper, the data processing of the $NH_3$ concentration is taken as an example to evaluate the proposed method of the missing data recovery. The monitoring data of 3 days [13, 14] are randomly selected in a farm to test the SVM prediction model. Sampling the $NH_3$ concentration and other related monitoring data every hour, there are 72 groups of data samples in 3 days, taking the front 48 samples as training set, and the remaining 24 samples as the test set. When using the K-cross-validation combined with QGSA to optimize the parameters of SVM prediction model, relative parameters setting include: population size is 20, evolutional generation is 50, initial temperature $M$ is 100, annealing parameter $\alpha$ is 0.8, optimization range of penalty parameter $C$ , kernel function parameter $\gamma$ and insensitive loss function parameter $\varepsilon$ are respectively set as [0,100], [0,100] and [0.001,1], crossover probability is 0.7, mutation probability is $0.7/L$ ( $L$ is the code length of chromosome), and using 5-cross-validation. Through testing for many times, the SVM parameter optimization results of the best prediction model corresponding are obtained as follows:

$C$ =50.9848, $\gamma$ =0.05865, $\varepsilon$ =0.03611. The corresponding MSE is 0.001255. The evolution of the parameter optimization process is shown in Figure 2.

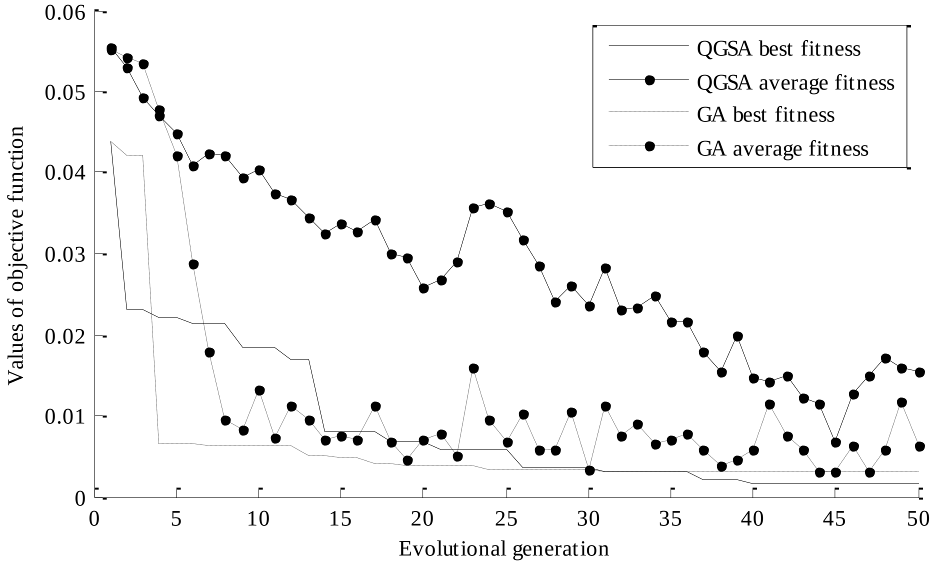

**Figure 2. Optimization Process of Parameters**

The Figure 2 shows that the average objective function value obtained by GA is quickly close to its optimal value, and GA gets local optimums. However, the average objective function value obtained by QGSA is large different from the best value, the average objective function value is relatively smaller in the early evolution, and it is relative larger in the late evolution. The reasons are as follows: Firstly, the perturbation solution set generated by QGA is made up of the good chromosomes, and can effectively improve the convergence speed in early evolution. Secondly, the perturbation solution set has a certain degree of uncertainty, which is based on the quantum probability collapse, and it can also expand the diversity of the population in a certain extent and avoid premature convergence. Lastly, the fitness function is designed combining with the temperature parameter, which can improve the search ability of the algorithm.

When using test set to validate the trained SVM prediction model, aiming at the issue about time series prediction, the relative error is adopted as the evaluation standard to evaluate predicted results instead of MSE. Through testing for many times, the regression results of the best prediction model are calculated as follows: The maximum relative error is 5.87%, the minimum relative error is 0.08%, and the average relative error is 1.77%. The analysis results of the test set are shown in Figure 3.

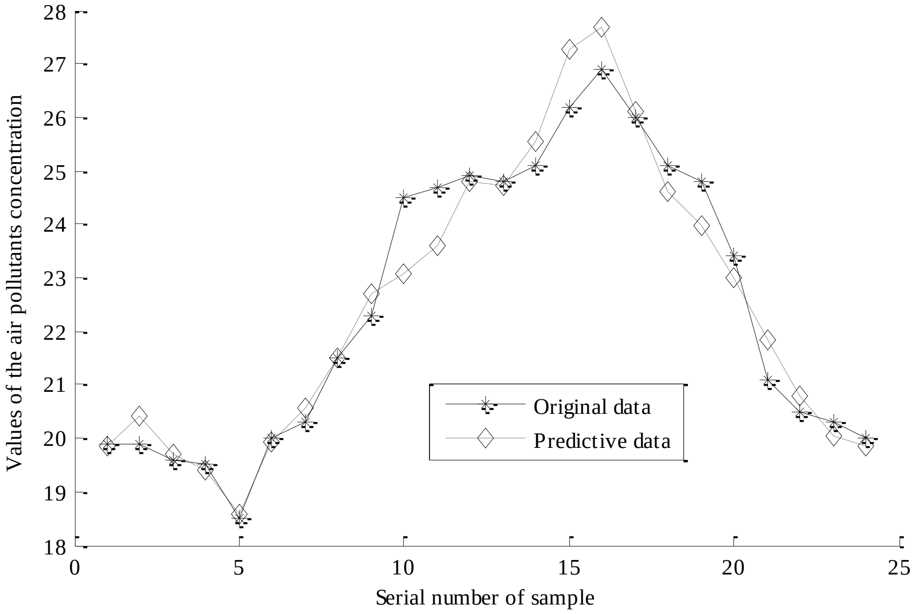

**Figure 3. Regression Results of Test Set**

In order to test the superiority of the proposed method for missing data recovery, the proposed SVM prediction model based on QGSA (QGSA-SVM for short) is contrasted with four kinds of prediction model. The first is the BP neural network prediction model (BP-NN for short), the second is the parameter-optimized SVM prediction model by grid search algorithm (Grid-SVM for short), the third is the parameter-optimized SVM prediction model based on particle swarm optimization algorithm (PSO-SVM for short), and the fourth is the parameter-optimized SVM prediction model based on GA (GA-SVM for short). The software and hardware environments for testing are as follows: Win7 64bit OS, Matlab 2012b, LibSVM-3.1 Toolbox, AMD Athlon(tm) X4 730 CPU, 4GB RAM. Prediction efficiency and performance by using different regression prediction model are shown in Table 1.

**Table 1. Comparison of Prediction Results with Different Models**

| Type of algorithm | Execution time (s) | Maximal relative error (%) | Minimal relative error (%) | Average relative error (%) |
|---|---|---|---|---|
| BP-NN | 40.6 | 5.99 | 0.07 | 3.17 |
| Grid-SVM | 8.1 | 6.95 | 0.01 | 2.54 |
| PSO-SVM | 7.9 | 5.69 | 0.02 | 2.41 |
| GA-SVM | 5.6 | 6.99 | 0.22 | 2.15 |
| QGSA-SVM | 6.1 | 5.87 | 0.08 | 1.77 |

As described in Table 1, the execution time of four SVM prediction models is obviously less than BP neural network, and average relative error of SVM prediction is also less than BP neural network. It is show that the SVM prediction model has certain advantages in solving the problem of regression. Comparing with other SVM prediction models, the maximum relative error of QGSA-SVM is slightly higher than PSO-SVM, and its execution time is slightly more than GA-SVM, but its average relative error is the smallest. The simulation results show that the QGSA-SVM prediction model obtains the best prediction result, realizes the unity of efficiency and performance, and the prediction efficiency and performance of QGSA-SVM is superiors to other SVM prediction models.

## 5. Conclusion

Considering the relationship between the air pollutants concentration in henhouse and a variety of factors such as time, space and environment, a method for missing data recovery is presented based on QGSA-SVM prediction model, which is based on SVM regression model combining GA with quantum genetic strategy and simulated annealing tactics. Comparative test results of estimation data and monitoring data show that the data recovery accuracy of the proposed method is very high, the maximal relative error was 5.87%, and the average relative error was 1.77%. This method enhances the complementarities between sensors, and improves the reliability of the monitoring system. It provides a reliable basis for measuring the air pollutants of the henhouse for a period of time, analyzing emission regularity of air pollutants in henhouse, controlling and disposing the air pollutants.

## Acknowledgments

The work is partially supported by the Youth Science Funds of Heilongjiang Province of China under Grant No.QC2013C065, the Scientific Research Fund of Heilongjiang Provincial Education Department under Grant Nos.12531465, 12541596 and the Science and Technology Planning Project of Heilongjiang Province Land Reclamation Bureau under Grant No.HNK125B-04-06.

## References

[1] A. Heber, T. Lim, J. Ni, P. Tao, A. Schmidt, J. Koziel, S. Hoff, L. Jacobson, Y. Zhang and G. Baughman, Journal of the Air & Waste Management Association, vol. 56, no. 12, **(2006)**.

[2] A. Heber, J. Ni, T. Lim, P. Tao, A. Schmidt, J. Koziel, D. Beasley, S. Hoff, R. Nicolai, L. Jacobson and Y. Zhang, Journal of the Air & Waste Management Association, vol. 56, no. 10, **(2006)**.

[3] R. W. Bottcher, K. M. Keener, R. D. Munilla, C. M. Williams and S. S. Schiffman, "Applied Engineering in Agriculture", vol. 20, no. 3, **(2004)**.

[4] H. Guo, W. Dehod, J. Agnew, J. R. Feddes, C. Laguë and S. Pang, "Transactions of the ASABE", vol. 50, no. 4, **(2007)**.

[5] Y. C. Lo, J. A. Koziel, L. Cai, S. J. Hoff, W. S. Jenks and H. Xin, Journal of Environmental Quality, vol. 37, no. 2, **(2008)**.

[6] L. Jacobson, B. Hetchler, D. Schmidt, R. Nicolai, A. Heber, J. Ni, S. Hoff, J. Koziel, Y. Zhang, D. Beasley and D. Parker, Journal of the Air & Waste Management Association, vol. 56, no. 10, **(2006)**.

[7] V. N. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag Press, New York, **(1995)**.

[8] P. P. Du, Journal of Mining & Safety Engineering, vol. 29, no. 4, **(2012)**.

[9] W. Liu, J. P. Wang, C. H. Liu and T. J. Ying, "Transactions of the Chinese Society of Agricultural Machinery", vol. 43, no. 4, **(2012)**.

[10] W. G. Chen, L. Teng, J. Liu, S. Y. Peng and C. X. Sun, Transactions of China Electrotechnical Society", vol. 29, no. 1, **(2014)**

[11] X. Wang, Z. Q Wang, G. Jin and J. Yang, "Transactions of the Chinese Society of Agricultural Engineering", vol. 30, no. 4, **(2014)**

[12] P. Wu and T. Lin, "Chinese Journal of Scientific Instrument", vol. 35, no. 2, **(2014)**.

[13] V. Aneja, W. Schlesinger, R. Knighton, G. Jennings, D. Niyogi, W. Gillian and C. Duke, "Characterization and Abatement of Air Emissions from Egg Production", Proceedings of the Workshop on Agricultural Air Quality: State of the Science, Potomac USA, **(2006)** June 5-8.

[14] J. M. Liu, Q. J. Xie and Y. Y. Zhang, International Journal of Smart Home, vol. 9, no. 5, **(2015)**.

## Authors

**Jinming Liu** is currently a lecturer at Heilongjiang Bayi Agricultural University. He is a Ph.D. candidate of Northeast Agricultural University. His research work is the application of information technology in agriculture.

**Qiuju Xie** is currently an associate professor at Heilongjiang Bayi Agricultural University. She received a Ph.D. degree in Northeast Agricultural University. Her research work is in the field of information technology of livestock breeding.

**Guiyang Liu** is currently a professor at Heilongjiang Bayi Agricultural University. He received master's degree in Harbin University of Science and Technology. His research work is the application of virtual reality in agriculture.