International Journal of Smart Home
Vol. 9, No. 5 (2015), pp. 175-184
http://dx.doi.org/10.14257/ijsh.2015.9.5.17

A Method for Missing Data Recovery of Waste Gas Monitoring in Animal Building Based on GA-SVM

Jinming Liu, Qiuju Xie and Yuanyuan Zhang

College of Information Technology, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang 163319, China
jinmingliu2008@126.com

Abstract

In order to solve the data missing problem caused by sensor faults during the waste gas monitoring in animal building, a method for missing data recovery was presented based on support vector machine (SVM) combined with genetic algorithm (GA). Multiple factors that influence monitoring values of the waste gas in animal building such as temporal, spatial and environmental, were considered to established a SVM regression prediction model to estimate the missing data of the waste gas monitoring. Meanwhile, to obtain better prediction accuracy, model parameters were optimized by the GA. The data processing of the ammonia (NH3) concentration was taken as an example; monitoring data of 3 days were randomly selected in a farm to test the presented model in this paper. It is shown that there was a very little error between the estimated data and the monitoring data, the maximal relative error was 6.99 % (percent), and the average relative error was 2.15 % (percent). It is an effective method for missing data recovery and a practical way of data processing for waste gas monitoring in animal building.

Keywords: *Genetic algorithm (GA), support vector machine (SVM), animal building, waste gas monitoring, data recovery*

1. Introduction

With the industrialization of livestock and poultry breeding, all kind of harmful gas emitted from animals not only directly affect the health of herdsmen and peoples in neighboring areas, also have significant effect on the healthy growth of livestock and food safety issues [1, 2]. To take effective methods for controlling and disposing the harmful gas, the waste gas concentration must be continuously and reliably monitored in livestock barns and poultry houses, exhaust emissions must be accurately measured, and factors that influence harmful gas emissions must be synthetically analyzed [3, 4]. Therefore, in order to analyze the harmful gas emissions rules, multiplex sensors of harmful gas concentration need to be installed in the animal building, gas concentrations need to be constantly monitored. Meanwhile, all monitoring data must be transmitted to the host computer database and analyzed to establish the emission model of the waste gas. However, the complex environment in animal buildings may cause sensors is shifted or damaged, which causes the deviation or completely error of monitoring data [5, 6]. To ensure the integrity and accuracy of the monitoring data, the missing data of the waste gas monitoring need to be recovered. Multiple factors that influence the waste gas concentration in animal building such as temporal, spatial and environmental, are interrelated and interact on each other. However, there is a large error between the estimated value and the actual value by using the linear interpolation method in this complex nonlinear system. The neural network algorithm has been adopted to estimate the missing data of waste gas monitoring, and obtained better estimation results of the missing data. But there are many deficiencies in the neural network itself, such as local

ISSN: 1975-4094 IJSH

minimum value problems, over-learning problems, and the selection of structure and type.

Support vector machine (SVM) is a machine learning method with good generalization ability which is under 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, 9]. The prediction accuracy of SVM depends directly upon the parameters, so the parameters of SVM have attracted increasing attention in regression prediction problems. The relevant scholars have put forward the intelligent optimization algorithm to optimize the parameters of SVM, such as particle swarm optimization (PSO) [10, 11] and genetic algorithm (GA) [12, 13]. Among them, GA has a good ability of robustness and global optimization, so it is fit for solving the complex optimization problems. In this paper, a method for missing data recovery is presented based on SVM combined with GA in livestock barns and poultry houses, and tested by using randomly selected monitoring data of 3 days in a farm.

2. Materials and Methods

2.1. Theoretical Basis

The basic idea of SVM nonlinear regression is the use of 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.

On the basis of the selected kernel function, the selection of SVM prediction parameters plays an essential in prediction accuracy. It is the most popular method of SVM parameters optimization that the gird search algorithm combines with the cross validation method. However, it is a time-consuming and inefficient performance which needs to be improved. GA is an intelligent global optimization method which has highly nonlinear mapping, adaptive and self-organization properties. According to the error between actual values and prediction values of SVM, GA can efficiently optimize the SVM parameters in the encoded population. Through selection, crossover and mutation operation, SVM model parameters are randomly optimized in the appointed range. After several generations of genetic evolution, obtained the individual with the best fitness can be used as the optimal parameters of SVM prediction model.

2.2. Input-output Parameters of SVM

Multiple factors that influence the waste gas concentration in livestock barns and poultry houses, 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 waste gas monitoring. The missing data of the waste gas 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 waste gas concentration at the previous sampling instant of the missing data sampling point; secondly, the change of the waste gas 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 waste gas 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.

When the monitoring data is missing, the missing data would be estimated by entering the appropriate data into the trained SVM prediction model. However, the input and output data should be normalized before using them to train SVM prediction model. The normalization method of each parameter is defined as follows:

$$y = \frac{(y_{\max} - y_{\min})(x - x_{\min})}{x_{\max} - x_{\min}} + y_{\min} \tag{1}$$

Where y is the post-normalized data, x is pre-normalized data, $x_{\max}$ and $x_{\min}$ are respectively the maximum and minimum of the monitoring data, $[y_{\min}, y_{\max}]$ is the normalization interval. If $x_{\max}$ and $x_{\min}$ are equal, namely all the monitoring data of this parameter are the same, set y equal to $y_{\min}$. Through many simulation tests found that the SVM prediction mode can obtain the best estimation results while the normalization intervals of input variables and output variable are respectively set to [-1, 1] and [0, 1].

2.3. Selection of SVM Kernel Function

SVM kernel functions establish an implicit mapping between the original sample space and the feature space. Its basic idea is that linearly non-separable problem of the original space can be convert into linear separable problems of high dimensional eigenspace. While SVM is solved regression problems, choosing the appropriate kernel function is an important factor affecting the prediction accuracy of SVM. The published paper [14] discussed the selection of kernel functions while SVM was adopted to solve nonlinear multivariate prediction problems. The results of the research show that the SVM prediction model is recognized the most accurate estimation based on the RBF kernel function. The Gaussian kernel function is the most commonly RBF kernel function, which is calculated as follows:

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

Where γ is the kernel parameter, $\gamma = \frac{1}{2\sigma^2}$, u is any point in the space, v is the center point, σ is the width parameter.

2.4. Optimization of SVM Parameters based on GA

In this paper, LibSVM toolbox is used to design the SVM regression prediction model, the epsilon-support vector regression (epsilon-SVR) is adopted as the SVM type, the Gaussian RBF kernel function is employed. Corresponding parameters should be optimized which includes the penalty parameter C, the kernel parameter γ and the insensitive loss function parameter ε.

2.4.1. Coding and Population Initialization: When using GA to optimize the SVM parameters, the encoding method is binary number encoding. Three SVM parameters C, γ and ε are respectively represented three genes of chromosome, each gene is encoded into k bit binary sequence. The structure of chromosome is shown in Figure 1.

a_1	a_2	…	a_k	b_1	b_2	…	b_k	c_1	c_2	…	c_k

Figure 1. Structure of Chromosome

Where the binary sequence $a_1a_2\cdots a_k$ is the gene code of the parameter C, the binary sequence $b_1b_2\cdots b_k$ is the gene code of the parameter γ, the binary sequence $c_1c_2\cdots c_k$ is the gene code of parameter ε. The corresponding real decoding formula for each gene can be expressed as follows:

$$f(x)=\left(\sum_{i=1}^{k} w_i \cdot 2^{i-1}\right)\cdot\frac{(U_2-U_1)}{2^k-1}+U_1 \tag{3}$$

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

When randomly generating initial population, in order to ensure the diversity of the population, the numbers of "1" in each chromosome binary code randomly are generated, and these "1" are randomly distributed into each chromosome.

2.4.2. Design of Fitness Function: In this paper, the SVM parameters are optimized based on GA combined with K-cross-validation, the purpose of the SVM prediction mode 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 fitness function of GA. Obviously, the fitness function value is smaller and prediction mode is better, and the more accurate prediction.

2.4.3. Design of Genetic Operation: Genetic operations of GA include selection, crossover and mutation operations. In order to keep the relatively permanent diversity of population chromosome, the stochastic sampling method is adopted as the selection operation. Meanwhile, the chromosome with the minimum fitness value is directly saved into the next generation population by the best saving tactics. The strategy of crossover operation is the single point crossover, and the strategy of mutation operation is the multiple bit mutation.

2.5. Prediction of Missing Data by SVM

When completing the SVM parameters optimization based on GA, the SVM prediction mode is established by the optimized parameters C, γ and ε. After training the SVM prediction, 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 waste gas concentration in the monitoring data at the previous sampling instant should be an estimated value of waste gas 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 waste gas monitoring. The algorithm flowchart is shown in Figure 2.

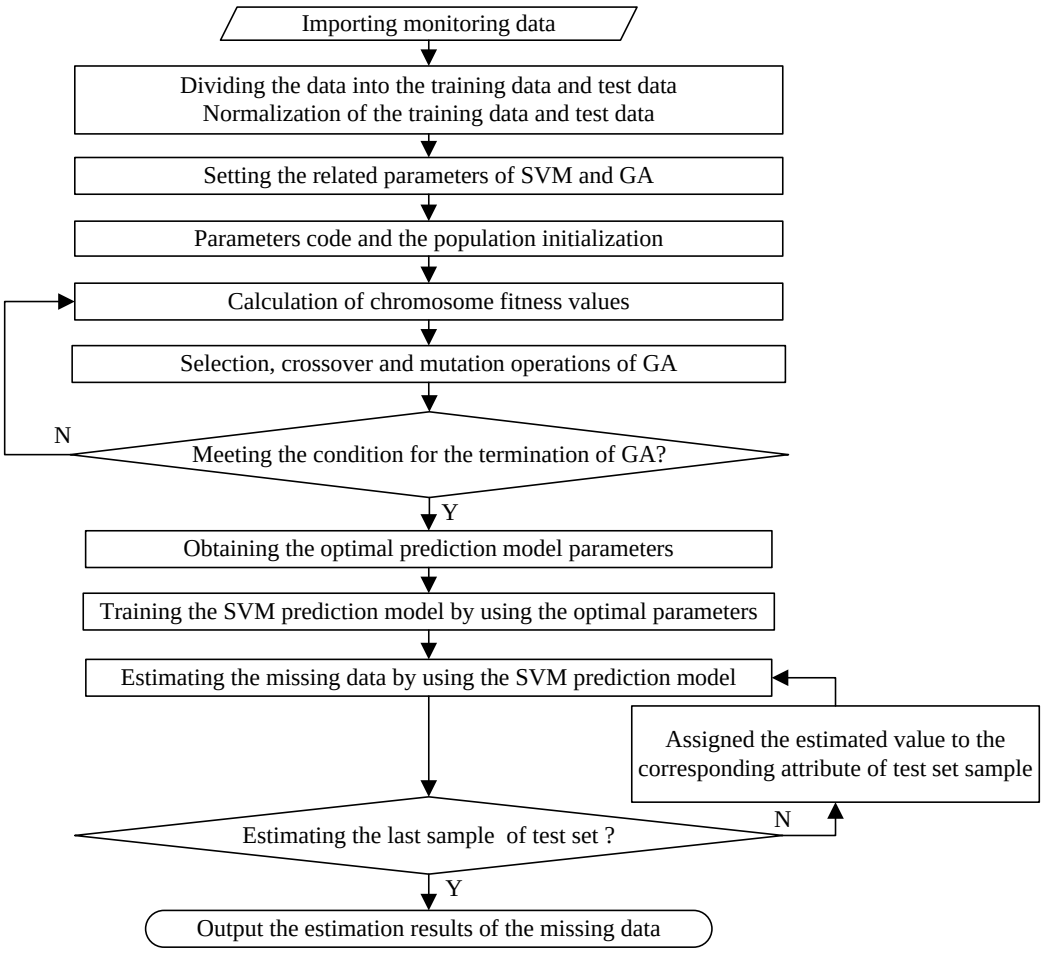

Figure 2. Flowchart of Missing Data Recovery

3. Results and Analysis

3.1. Data Sources

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 are randomly selected in a farm to test the SVM prediction model. Sampling the NH_3 concentration and other related monitoring data every hour, therefore, 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. Part of the sample data after preprocessed are shown in Table 1.

Table 1. Part of Experimental Data

Serial number	NH_3 concentration of the test channel (ppm)	Ambient temperature (℃)	Relative humidity (%)	Wind speeds (m/s)	NH_3 concentration of the test channel at the previous sampling instant (ppm)	Change of NH_3 concentration in adjacent channel (ppm)
1	15.5	19.19	57.56	0.62	20.0	-0.3
2	14.5	19.21	57.59	0.66	15.5	-0.4
3	13.4	19.32	59.76	0.77	14.5	0.2
4	13.3	19.45	59.66	0.70	13.4	9.3

5	13.1	19.80	59.85	0.87	13.3	0.5
6	18.4	20.72	57.87	0.49	13.1	0.3
7	19.1	21.55	57.50	0.48	18.4	4.9
8	19.4	21.76	55.98	0.50	19.1	0.1
9	19.5	21.88	55.86	0.51	19.4	0.5
10	20.7	21.97	55.96	0.40	19.5	-0.1
11	21.1	22.10	55.86	0.38	20.7	4.7
12	22.8	22.25	55.78	0.34	21.1	1.7

3.2. Setting of Related Parameters

When using the K-cross-validation combined with GA to optimize the parameters of SVM prediction model, related parameters setting include: population size is 20, evolutional generation is 50, optimization range of penalty parameter C, kernel function parameter γ and insensitive loss function parameter ε are respectively set as [0,100], [0,100] and [0.001,1], Crossover probability is 0.7, mutation probability is $0.7/Lind$ ($Lind$ 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 =61.6872, γ =0.0545, ε =0.0361. The corresponding MSE is 0.0013. The evolution of the parameter optimization process is shown in Figure 3.

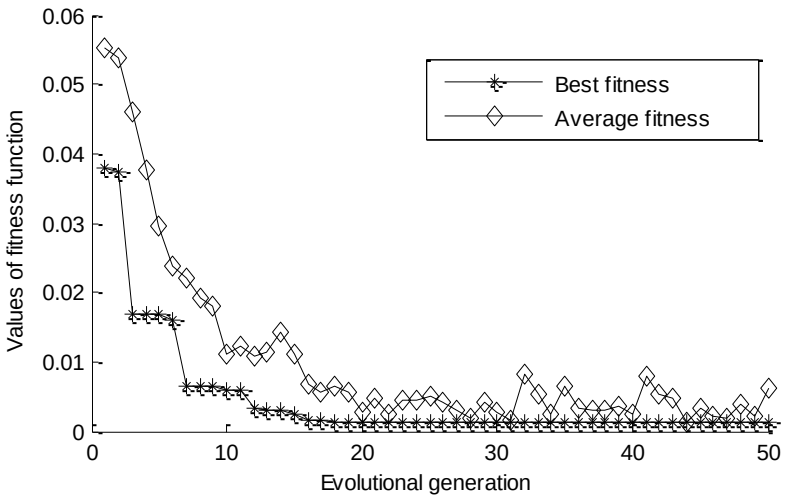

Figure 3. Optimization Process of Parameters

3.3. Analysis of Simulation Results

The SVM prediction model is obtained by taking the parameters C、γ、ε and training set into the training function. Testing the SVM prediction model with the training set, the corresponding MSE is 0.009. The test results are shown in Figure 4.

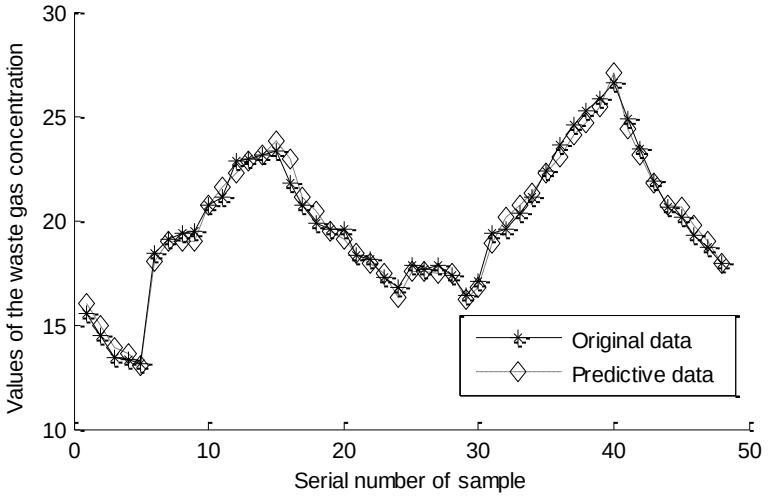

Figure 4. Regression Results of Training Set

The Figure 4 shows that the proposed method can fit the training set very well by the SVM prediction model.

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 6.99%, the minimum relative error is 0.22%, and the average relative error is 2.15%. The regression fitting results of the test set are shown in Figure 5.

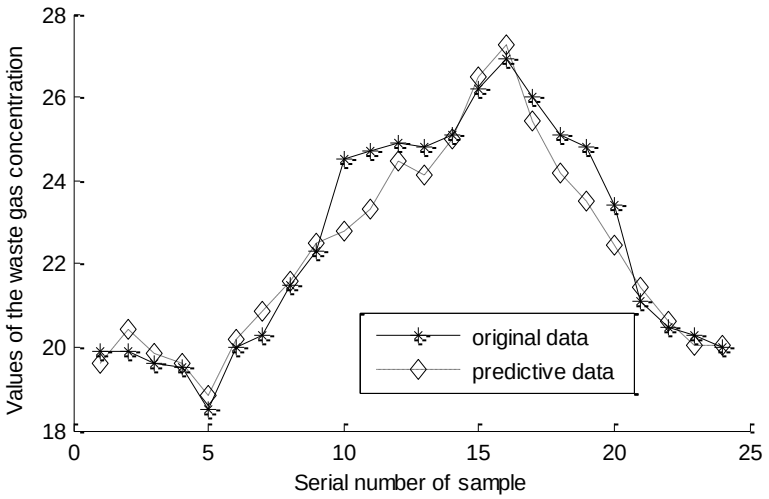

Figure 5. Regression Results of Test Set

In order to test the superiority of the proposed method for missing data recovery, the proposed SVM prediction model based on (GA-SVM for short) is contrasted with three 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), and the third is the parameter-optimized SVM prediction model based on particle swarm optimization algorithm (PSO-SVM for short). The prediction efficiency and performance by using different regression prediction model are shown in Table 2.

Table 2. Comparison of Predict Results with Different Models

Type of algorithm	Execution time (s)	Maximal relative error (%)	Minimal relative error (%)	Average relative error (%)
BP-NN	60	5.99	0.07	3.17
Grid-SVM	9	9.39	0.41	4.89
PSO-SVM	7	8.37	0.47	3.38
GA-SVM	7	6.99	0.22	2.15

As described in Table 2, the execution time of three SVM prediction models is obviously less than BP neural network. Although the maximum and minimum relative error of GA-SVM is slightly higher than BP neural network, its average relative error is minimum. In addition, the prediction efficiency and performance of GA-SVM is superiors to other SVM prediction model. The GA-SVM prediction model obtains the best prediction result, and realizes the unity of efficiency and performance. The results of the PSO-SVM and GSA-SVM are respectively the best prediction result after many times test. When restoring the missing data, only need to save the best prediction model after many times test, and use this model to estimate the missing data subsequently. The simulation results of GA-SVM prediction model show that the optimized results of the parameters of C、γ and ε using GA have a large difference, the corresponding results of the test set regression fitting also have a large difference, the total average relative error is 3.27% in test 100 times, less than the prediction error of three kinds of prediction model of BP neural network, grid-SVM and PSO-SVM. Therefore, the estimated accuracy of GA-SVM prediction model is highest, could satisfy the needs of analyzing the emission regularity of harmful gas in animal building.

4. Discussion

The fitting results of test set have a large difference with testing GA-SVM prediction mode for many times. The reason for this is that the standard GA tends to get local optimums. The parameters optimized results are directly determined by the initial population's random selection and evolution process of GA. Maintaining the diversity of GA population can effectively avoid premature convergence. Increasing the population of GA can extend population diversity in a certain extent, but it will seriously affect the optimization efficiency of GA. Through the way of increasing the crossover and mutation probability can effectively improve the GA population diversity, but the original solution may be completely destroyed because of the excessive crossover and mutation probability. So introducing the idea of adaptive crossover and mutation probability can avoid the precocious problem and accelerate the convergence of GA which adaptively changes the probability of the crossover and mutation. At the same time, other intelligent algorithms such as simulated annealing algorithm can be combined with GA designing a hybrid optimization algorithm which improves the estimated accuracy of the regression prediction model. This is a question and need to be explored in depth.

5. Conclusion

Considering the relationship between the waste gas concentration in animal building and a variety of factors such as time, space and environment, a method for missing data recovery is presented based on support vector machine(SVM) combined with genetic algorithm(GA). This method enhances the complementarity between sensors, and improves the reliability of the monitoring system. The simulation results show that the method of data recovery is feasible and valid. It

provides a reliable basis for measuring the exhaust emissions of the animal building for a period of time, analyzing emission regularity of harmful gas in livestock barns and poultry houses, controlling and disposing the harmful gas.

Acknowledgments

The work is partially supported by the Undergraduate Training Programs for Innovation and Entrepreneurship of Heilongjiang Province of China under Grant No.201310223002 and the Youth Science Funds of Heilongjiang Province of China under Grant No.QC2013C065.

References

[1] A. J. Heber, T. T. Lim, J. Q. Ni, P. C. Tao, A. M. Schmidt, J. A. Koziel, S. J. Hoff, L. D. Jacobson, Y. H. Zhang and G. B. Baughman, "Journal of the Air & Waste Management Association", vol. 12, no. 56, **(2006)**.
[2] A. J. Heber, J. Q. Ni, T. T. Lim, P. C. Tao, A. M. Schmidt, J. A. Koziel, D. B. Beasley, S. J. Hoff, R. E. Nicolai, L. D. Jacobson and Y. H. Zhang, "Journal of the Air & Waste Management Association", vol. 10, no. 56, **(2006)**.
[3] R. W. Bottcher, K. M. Keener, R. D. Munilla, C. M. Williams and S. S. Schiffman, "Applied Engineering in Agriculture", vol. 3, no. 20, **(2004)**.
[4] H. Guo, W. Dehod, J. Agnew, J. R. Feddes, C. Laguë and S. Pang, "Transactions of the ASABE", vol. 4, no. 50, **(2007)**.
[5] Y. C. Lo, J. A. Koziel, L. Cai, S. J. Hoff, W. S. Jenks and H. Xin, "Journal of Environmental Quality", vol. 2, no. 37, **(2008)**.
[6] L. D. Jacobson, B. P. Hetchler, D. R. Schmidt, R. E. Nicolai, A. J. Heber, J. Q. Ni, S. J. Hoff, J. A. Koziel, Y. H. Zhang, D. B. Beasley and D. B. Parker, "Journal of the Air & Waste Management Association", vol. 10, no. 56, **(2006)**.
[7] V. N. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag Press, **(1995)**.
[8] P. P. Du, "Journal of Mining & Safety Engineering", vol. 4, no. 29, **(2012)**.
[9] D. Dai, X. T. Huang, Z. Dai, Y. P. Hao, L. H. Li and C. Fu, "High Voltage Engineering", vol. 11, no. 39, **(2013)**.
[10] W. Liu, J. P. Wang, C. H. Liu and T. J. Ying, "Transactions of the Chinese Society of Agricultural Machinery", vol. 4, no. 43, **(2012)**.
[11] S. G. Ye, Y. Peng and H. C. Zhou, "Journal of Dalian University of Technology", vol. 1, no. 51, **(2011)**.
[12] X. L. Liu, X. S. Zhao, F. Lu and W. B. Sun, "Journal of China Coal Society", vol. 12, no. 27, **(2012)**.
[13] W. G. Chen, L. Teng, J. Liu, S. Y. Peng and C. X. Sun, "Transactions of China Electrotechnical Society", vol. 1, no. 29, **(2014)**.
[14] X. Wang, Z. Q. Wang, G. Jin and J. Yang, "Transactions of the Chinese Society of Agricultural Engineering", vol. 4, no. 30, **(2014)**.

Authors

Jinming Liu, he is currently a lecturer at Heilongjiang Bayi Agricultural University. He received master's degree in Yanshan University. His research work is the application of information technology in agriculture.

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