2013 2nd International Conference on Geological and Environmental Sciences
IPCBEE vol.52 (2013) © (2013) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2013. V52. 17

GA-SVR Prediction of Failure Depth of Coal Seam Floor Based on Small Sample Data

Yueming QI [1+], Xiaoer ZHAO [2], Bin LUO [3] and Menglang LUO [4]

[1,2,3,4]School of Resources and Earth Sciences, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China

Abstract. The failure depth of coal seam floor is an important index to evaluate whether water inrush from coal seam floor will be occurred in the process of coal mining. It is determined by actual measurement or empirical method frequently. The measured results are relatively accurate, but the process is complicated and costs high. As a result, the sample data is little. The empirical method is simple but requires strict application conditions. In this paper, on the basis of previous practical experience and theoretical summary, the Genetic Algorithm-Support Vector Regression (GA-SVR) model is put forward to predict the failure depth of coal seam floor. The optimal SVR parameters are obtained by GA to learn the training samples. GA-SVR model is used to predict the training and testing samples. It shows that the GA method of optimization parameters is reliable. We also can select the optimal parameters by methods of Particle Swarm Optimization (PSO) and Grid Search (GS). The GA-SVR model can predict it much more accurately than PSO-SVR and GS-SVR model.

Keywords: Failure Depth of Coal Seam Floor; Prediction; Support Vector Regression; Genetic Algorithm

1. Introduction

With the increasing depth of coal mining in North China, the major confined water disaster in Ordovician carbonate rock is more and more serious. As Wang et al, poited out,, three zones will be formed on the floor of the coal seam[1]. Determining accurate depth of the three zones, especially the floor failure zone, is the key to predicate whether water inrush will take place on the floor of coal seam.

So far, some field measurement methods are used to determine floor failure depth，such as those using mine electrical profiling, electromagnetic wave, bore sound wave, ultrasound imaging and seismic wave ct [1-2]. Despite of their reliability, large amount of repeated observations will cost a lot of human, material and financial resource. On the other hand, non-measured methods can also be used to predict the floor failure depth,such as the method based on the underground pressure control theory apllied by Shi Longqing [3], regression analysis, by Guan Yingbin [4], $FLAC^{3D}$, by Duan Hongfei [5], and material simulation, by Wang Jisong [6]. However, the physical and mechanical change in the floor strata of coal seam induced by mining is complex and nonlinear . Many factors are involved in it. So, it is necessary to choose the most important factors as evaluation indexes to describe the process and the extent of floor failure accurately. To improve the performance of nonlinear prediction for floor failure depth, Yu Xiaoge [7] establishes prediction model of floor failure depth based on BP neural network. But when the sample size is small, neural network can hardly make good prediction [8].

Generally, it is difficult to get index data and field measurement data. Even if you get it, the number of samples is limited. In this case, we need to seek approriate statistical methods for nonlinear prediction of floor failure based on small sample data. Support Vector Machine (SVM) is applied in this paper to obtain good statistical regularities and generalization ability under the condition of small sample. The structure of

+ Corresponding author. Tel.: +86 13914880319;
E-mail address: (ym_qi@126.com).

SVM is simple, learning speed is fast and prediction precision is high [9-11]. As a kind of SVM, Support Vector Regression (SVR) is used for nonlinear prediction under the condition of small sample data. Using Genetic Algorithm to select parameters for SVR is an effective method. So, this paper proposes the GA-SVR model to predict the failure depth of coal seam floor. Particle Swarm Optimization and Grid Search Algorithm also can be used to select parameters for SVR. Practice shows that prediction precision of GA-SVR model is high and GA is a reliable method of optimizing parameters. Prediction results are more accurate than that of PSO-SVR and GS-SVR model.

2. Support Vector Machine

Support Vector Machine (SVM) was first proposed in 1995 by Cortes and Vapnik, which is suitable for problem with few samples, nonlinear or high-dimension. It could be generalized and applied to machine learning problems, such as function fitting [11-13]. The SVM theory is divided into Support Vector Classification (SVC) and Support Vector Regression (SVR). SVR algorithm is applied to time series prediction, nonlinear modeling and prediction, optimal control and other aspects.

Considering training sample set $\left\{(\mathbf{x_i}, \mathrm{y_i}) \middle| i=1,2,...,l\right\}$. Among them, $\mathbf{x_i} \in \mathbf{X} = R^n$ is an input and $\mathrm{y_i} \in \mathrm{Y} = R$ is a target output. For nonlinear SVR, the idea of using SVM to regress and predict is mapping data to the high dimensional feature space by nonlinear mapping using kernel function and structuring an optimal linear regression function:

$$f(\mathbf{x}) = \mathbf{w} \bullet \phi(\mathbf{x}) + b \tag{1}$$

$$\mathrm{K}(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i}) \bullet \phi(\mathbf{x_j}) \tag{2}$$

Where $\mathbf{w}$ is weight value vector, ϕ is mapping function, $\phi(\mathbf{x})$ is vector in high dimensional feature space, $\mathrm{K}(\mathbf{x_i}, \mathbf{x_j})$ is kernel function.

The standard form of support vector regression is:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \xi^*} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* \\ \text{subject to} \quad & \mathbf{w}^T\phi(x_i) + b - y_i \le \varepsilon + \xi_i, \\ & y_i - \mathbf{w}^T\phi(x_i) - b \le \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \ge 0, i = 1, \cdots, l. \end{aligned} \tag{3}$$

Where C is penalty factor, ε is insensitive loss function, ξ_i and ξ_i^* are the slack variables.

According to Lagrange transformation of formula (3), the dual problem is got, as bellowed.

$$\begin{aligned} \min_{\alpha, \alpha^*} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T Q(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) \\ \text{Subject to} \quad & \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) = 0, 0 \le \alpha_i, \alpha_i^* \le C, i = 1, \cdots, l, \end{aligned} \tag{4}$$

Where $Q_{ij} = K(\mathbf{x_i}, \mathbf{x_j}) \equiv \phi(\mathbf{x_i}) \bullet \phi(\mathbf{x_j})$.

The approximate function is:

$$f(\mathbf{x}) = \sum_{i=1}^{l}(-\alpha_i + \alpha_i^*) K(\mathbf{x_i}, \mathbf{x}) + b \tag{5}$$

The type of algorithm is different as different selection of kernel function. Currently, the commonly used kernel functions are linear kernel function, polynomial kernel function, RBF kernel function and sigmoid kernel function. This paper chooses RBF as the kernel function.

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp\left\{-\sigma\left|\mathbf{x_i} - \mathbf{x_j}\right|^2\right\} \tag{6}$$

Where σ is the width of RBF kernel.

3. GA-SVR prediction model

3.1. Parameters selection of SVR based on GA

The key parameters of SVR, i.e. σ (the width of RBF kernel), C (penalty factor) and ε (insensitive loss function) have a great influence on the accuracy of SVM regression. They are given by experience or test without better ways before. To avoid the blindness and low efficiency of selecting parameters by human, GA is used to seeking for the optimal parameters of SVR considering that GA has strong global search capability and it can simulate natural selection and genetic variation in the process of biological evolution. GA operates iteratively on a population of structures, each one of which represents a candidate solution to the problem, properly encoded as a string of symbols. A randomly generated set of such strings forms the initial population from which the GA starts its search [14].

(1) Parameters encoding

The SVR parameters optimization process is a complex and continuous parameters optimization problem. Encoding is the crux of parameters optimization. It is decided by the character of the problem which kind of encoding method is chosen. GA utilizes binary encoding method, because binary encoding method is simple and easy to implement [14].

(2) Produce initial population and set fitness function

The initial population is produced in the solution space at random, and they are evenly distributed in solution space. The fitness function is defined to be mean square error after the cross-validation based on the training data.

$$F = \text{MSE}=\sqrt{\frac{1}{l}\sum_{i=1}^{l}(y_i - f_i)^2} \tag{7}$$

Where y_i is the actual value of the data set, f_i is predictive value of the data set, and l is the number of test samples.

(3) Genetic operation

Three basic genetic operators guide this search: selection, crossover, and mutation. The genetic search process is iterative: evaluating, selecting, and recombining strings in the population during each iteration until reaching some termination condition.

3.2. Procedure of establishing GA-SVR prediction model

GA-SVR model is established to provide an efficient method to select parameters of SVR [15]. The procedure of establishing GA-SVR prediction model is illustrated in Fig.1.

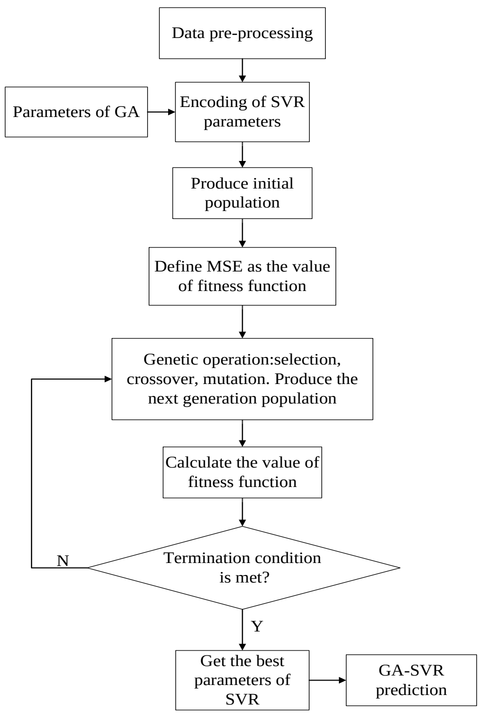

Fig. 1: The procedure of establishing GA-SVR prediction model.

4. Training and test samples

Table 1: Training and test samples.

Node number	Mining depth /m	Dip angle /(°)	Mining thickness /m	Inclined length of working face /m	Resistance to floor failure	Whether there is fault or fracture zone	Failure depth/m
1	123	15	1.1	70	0.2	0	7.00
2	130	15	1.4	135	0.4	0	12.00
3	110	12	1.4	100	0.4	0	10.70
4	148	18	1.8	95	0.8	0	9.00
5	225	14	1.9	130	0.8	0	9.75
6	308	10	1	160	0.6	0	10.50
7	287	10	1	130	0.6	0	9.50
8	300	8	1.8	100	0.4	0	10.00
9	230	10	2.3	120	0.6	0	13.00
10	310	26	1.8	128	0.2	0	16.80
11	259	4	3	160	0.6	0	16.40
12	320	4	5.4	60	0.6	0	9.70
13	520	30	0.94	120	0.6	0	13.00
14	400	9	7.5	34	0.4	0	8.00
15	227	12	3.5	30	0.4	1	7.00
16	900	26	2	200	0.6	0	27.00
17	1000	30	2	200	0.6	0	38.00
18	200	10	1.6	100	0.2	0	8.50
19	375	14	2.4	70	0.6	0	9.70
20	118	18	2.5	80	0.2	0	10.00
21	400	9	4	34	0.4	0	6.00
22	310	26	1.8	128	0.2	1	29.60
23	227	12	3.5	30	0.4	0	3.50
24	320	4	5.4	100	0.6	0	11.70
25	400	9	4	45	0.4	0	6.50
26	327	12	2.4	120	0.6	0	11.70
27	230	26	3.5	180	0.4	0	20.00

4.1. Selection of index factors

There are many factors affecting the floor failure of coal seam. So, based on the previous research results [7], six major factors are selected as prediction indexes in this paper,: ①Mining depth. ②Coal seam dip angle. ③Mining thickness. ④Inclined length of working face. ⑤Resistance to floor failure.⑥Existence of fault or fracture zone in the working face.

4.2. Obtaining and pre-processing of training and test samples

Field data of 27 working faces is selected as training and test samples, of which 24 as training samples and the other 3 as test samples. The data is shown in Tab.1. To avoid attributes in greater numerical ranges dominating those in smaller numerical ranges and numerical difficulties during the calculation [13], the original data should be normalized to the range (0, 1) before establishing model using data. The conversion formula is:

$$x' = (x - x_{\min}) / (x_{\max} - x_{\min}) \tag{8}$$

5. Model prediction

5.1. Selection of model parameters

SVR obtains knowledge from the training samples, and GA is applied to select C, σ, and ε. GA-SVR prediction model is established. The model parameters are optimized in MATLAB, choosing 3-e-SVR and RBF kernel function.

In order to avoid convergence of the population becoming difficult [14], the size of population is set to 20. Because bigger population will enlarge the GA running time and disperse the conformation of population. The maximum evolutionary generation is set to 200. The crossover probability and the mutation probability are 0.4 and 0.01 in this paper. The optimizing spaces of C, σ, and ε are respectively (0, 100), (0, 1000), and (0, 1).

The values of indexes in Tab.1 are normalized to be input vector of training function. The values of failure depth in Tab.1 are normalized to be target vector. After the training is finished, the optimal parameters C, σ, and ε are respectively 98.362, 0.192, and 0.01. The curve of training process is illustrated in Fig. 2.

5.2. Prediction of training and test samples

According to the optimal parameters and training samples, prediction model of failure depth is established through svmtrain function, which is used to make a prediction for the training and test samples. After reverse-normalization, we get the comparison curve between prediction values and training target

vector. Fig.3 shows it. Prediction values of the test samples are in Tab.2.

5.3. Prediction model based on PSO-SVR and GS-SVR

5.3.1 PSO-SVR prediction model

As an intelligent optimization algorithm, PSO possess the character of high convergence rate and global optimization [16]. Using PSO to select the optimal parameters, we can get the fitness curve of optimal parameters as Fig.4. Prediction values of the test samples based on PSO-SVR model are in Tab.2.

5.3.2 GS-SVR prediction model

GS is a common method for parameters selection of nonlinear SVM. Each pair of (C, σ) is tried. The pair is selected with the best cross-validation accuracy rate [17]. The contour map of grid optimization is illustrated in Fig.5. Prediction values of the test samples based on GS-SVR model are in Tab.2.

5.4. Contrastive analysis of the results

The prediction results in Table 2 shows that the maximum absolute error of failure depth of coal seam floor based on GA-SVR model is 0.8 m; maximum relative error is 4.00%. The above data indicates that advantages of GA-SVR are obvious. Prediction results of GA-SVR are more accurate than PSO-SVR model and GS-SVR model. The optimization of SVR parameters based on GA is better than that based on PSO and GS.

Table 2: Prediction values comparison of different models.

Node number	Actual value/m	GA-SVR Prediction value/m	GA-SVR Prediction error/%	PSO-SVR Prediction value /m	PSO-SVR Prediction error /%	GS-SVR Prediction value /m	GS-SVR Prediction error /%
25	6.50	6.72	3.23	6.83	5.07	6.81	4.77
26	11.70	11.29	-3.50	12.64	8.03	12.63	7.95
27	20.00	20.80	4.00	22.84	14.20	22.87	14.35

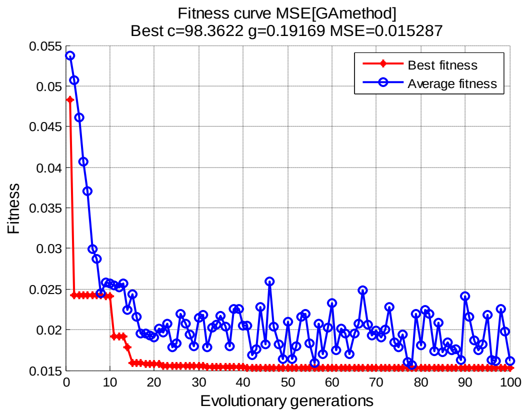

Fig. 2: The fitness curve of optimal parameters by GA.

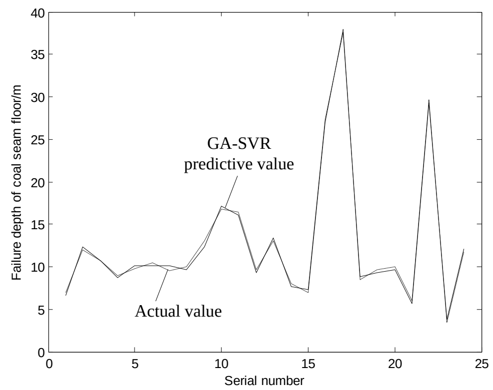

Fig. 3: The comparison curve of prediction values and training target vector.

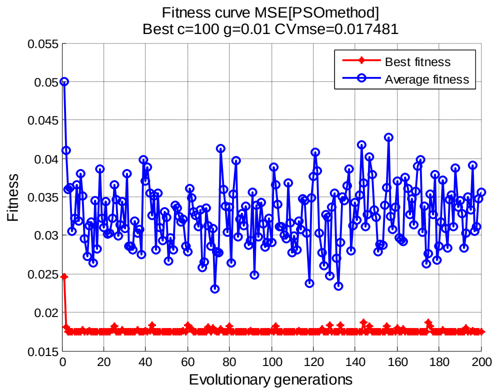

Fig. 4: The fitness curve of optimal parameters by PSO.

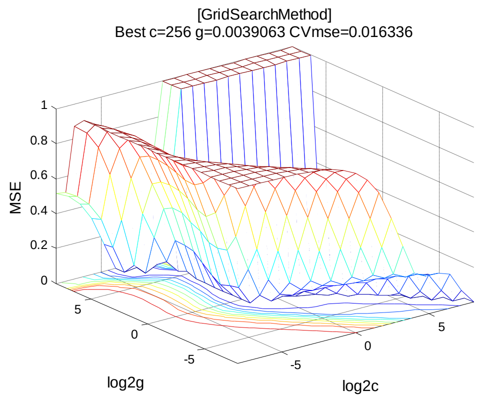

Fig. 5: The contour map of grid optimization.

6. Conclusions

There is nonlinear relationship between failure depth of coal seam floor and its influencing factors. To solve this problem, the basic idea of SVR is that mapping the input data space to a high dimensional feature space by a nonlinear mapping and structuring an optimal linear regression function to determine the failure depth.

It is difficult to get complete index data and field measurement data of floor failure, and the number of samples is limited. SVR can provide good prediction precision even lack of samples as long as the model parameters are suitable. GA-SVR prediction model is a small sample learning method based on the structural risk minimization principle. The results show that it has preferably applicability.

The parameters selection of SVR is important on prediction accuracy of floor failure depth. Thus, we present a new approach to realize parameters selection of SVR in this paper. It uses GA to select best parameters of SVR in global scope automatically. Examples show that parameters selection of GA-SVR is more reasonable and reliable compared with PSO-SVR and GS-SVR, and its prediction accuracy is higher. Therefore, it is more suitable for floor failure prediction.

7. Acknowledgements

We thank for the financial support from the National Natural Science Foundation (41002087), Postdoctor Science Foundation of Jiangsu Province (1001046C) and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

8. References

[1] J. C. Wang, Y. C. Xu, G. M. Xu, et al. Application of mine electric profiling method to detect floor failure depth of coal mining face. *Coal Sci. Tech.* 2010, **38** (1): 97-100.

[2] C. W. Liu, M. Zhang, W. S. Zhao. Determination of the floor failure depth after mining of coal seam using the sonic measuring technique. *Coal Sci. Tech.* 2004, (3): 4-5.

[3] L. Q. Shi, J. Han. Theory and practice of dividing coal mining area floor into four-zone. *Journal Chi. Uni. Min. Tech.* 2005, **24** (1): 16- 23.

[4] Y. B. Guan, H. M. Li, J. C. Lu. Research of No.9 coal seam floor's fracture regularity in xiandewang coal mine. *Journal Chi. Co. Soc.* 2003, **28** (2): 121- 125.

[5] H. F. Duan, Z. Q. Jiang, R. Zhang. Field measurement and simulation research on failure depth of fully mechanized thin coal seam floor in yangcun coal mine. *Journal Chi. Co. Soc.* 2011, **36**(1): 71- 73.

[6] J. S. Wang, Y. B. Guan, S. X. Bao, et al. Application of similar material simulation in research of coal seam floor failure regularity. *Global Geo.* 2006, **25** (1): 86-90.

[7] X. G. Yu, J. Han, L. Q. Shi, et al. Forecast of destroyed floor depth based on BP neural networks. *Journal Chi. Co. Soc.* 2009, **34** (6): 731- 736.

[8] Q. MENG, Y. S. WANG and Y. ZHOU. Prediction of spontaneous combustion in caving zone based on Rough

Set and SVM. *Journal Chi. Co. Soc.* 2010, **35** (12): 2100-2104.

[9] B. Z. Liu, B. Liang. Prediction of seam floor water inrush based on combining principal component analysis and support vector regression. *Coal Geo. Exp.* 2011, **39** (1): 28-31.

[10] Z. G. Yan, H. B. Bai. MMH support vector machines model for recognizing multi-head stream of water inrush in mine. *Chinese Jou. Ro. Mech. Eng.* 2009, **28** (2): 324-330.

[11] N. Y. Deng, Y. J. Tian. *Support Vector Machines: Theory, Algorithms and Development.* SCI Press, 2009.

[12] Vapnik V, ladimir N. *The Nature of Statistical Learning Theory.* Springer2-Verlag, Inc, 2000.

[13] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a Library for Support Vector Machines. *ACM Tran. Intell. Syst. Tech.* 2011, 2 (3).

[14] X. M. GAO, S. F. YANG and Y. HU. Leakage forecasting for water supply network based on GA-SVM model. *Proc. of the 2010 Symposium on Piezoelectricity, Acoustic Waves and Device Applications.* Xiamen: The Acoustical Society of China. 2010, pp. 206-209.

[15] L. M. Li and G. R. Wen. Parameters selection of support vector regression based on genetic algorithm. *Computer Eng. Appl.* 2008, **44** (7): 23−26.

[16] J. Kennedy, R. C. Eberhart. Particle swarm optimization. *IEEE International Conference on Neural Network.* New Yolk. 1995, pp. 1942-1948.

[17] Q. Y. Li, H. Zhou, A. P. Lin, etc. Prediction of ash fusion temperature based on grid search and support vector machine. *Journal Zhe. Uni.* 2011, **45** (12): 2181-2187.