Prediction of Water Inrush from Coal Floor Based on Small Sample Data Mining Technology and Realization Using MATLAB

Li Yang
School of Economy and Management, Anhui University of Science & Technology, Huainan, Anhui, China
E-mail: yangli081003@163.com

Chengcheng Liu
School of Economics and Management, Anhui University of Science & Technology, Huainan, China, 232001
E-mail: yangli081003@163.com

Jing Han
Huainan Vocational & Technical College, Huainan, Anhui, china
E-mail: hanjing623@163.com

Wu Sheng
School of Economy and Management, Anhui University of Science & Technology, Huainan, Anhui, China
Email: wsheng116@163.com

***Abstract*—For there are few samples of water inrush from coal floor, how to dig wider information under the circumstance of limited sample data is to improve the prediction accuracy of water inrush form coal floor. Therefore, prediction model of water inrush is established based on correlation analysis and support vector machine (SVM). Not only does the model simplify the influenced indexes to construct the index system of water inrush from coal floor, but also dig the value of sample data to solve the problem of small sample and nonlinear prediction. Through the empirical analysis, the model using MATLAB can accurately predict whether water inrush from coal floor occurs.**

***Index Terms*—Small Sample; Support Vector Machine; Correlation Analysis; Prediction of Water Inrush; MATLAB**

I. INTRODUCTION

For quite a long time, coal industry in China is still the main energy industry. However, current safety situation of mine production remains very serious because of frequent occurring of water inrush from coal floor accidents, which causes significant losses to the life and property of staff, and seriously influence and restrict coal mine safety production. Only to accurately predict whether water-inrush occurs and take necessary measures ahead of time can ensure coal mine safety production.

For immediately collecting sample data on the water inrush accident scene form coal mine is very difficult, there is typically insufficient problem of negative-class data. Therefore, digging wider information under the circumstance of limited data is the problem need to be solved in small sample method. And SVM is the best theory for the small sample nonlinear classification and prediction. Because the influence factors of water-inrush are many, this paper puts forward water inrush prediction model based on the correlation analysis and SVM. Not only does this model make full use of the advantages of correlation analytic method to optimize the attribute index, reduce the number of the model input variables and improve the convergence speed, but also can deeply dig sample data value, which accurately predicts water-inrush from coal floor.

II. PREDICTION MODEL OF WATER INRUSH FROM COAL FLOOR BASED ON CORRELATION ANALYSIS AND SVM

A. *Correlation Analysis [2, 3]*

Correlation analysis is a common statistical method which studies the degree of correlation of indexes or variables. Correlation coefficient accurately reflects the degree of linear relationship between variables. Commonly, using Pearson correlation coefficient, its computation formula is:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\left(y_i - \bar{y}\right)^2}} \tag{1}$$

where n is the sample size; x_i and y_i are the sample values of variables. In general, $-1 \le r \le 1$, and

doi:10.4304/jsw.7.10.2241-2246

absolute value of r reacts the close degree of correlation between two variables. The larger absolute value means that correlation is more close, and vice versa. $|r|=1$ means perfect correlation, $r=0$ means perfect un-correlation.

In general, t statistics is used as the test statistic of Pearson correlation coefficient, the formula is:

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2)$$

where, t statistics obeys the t distribution of the $n-2$ degree of freedom.

B. SVM PRINCIPLE[4-11]

SVM is based on the principle of VC dimension and structural risk minimization, which is a tool of solving machine learning problem by means of optimization method. The following describes standard C-support vector machine (C-SVC) model.

Given a training set $(x_i, y_i)(i=1,2,\cdots l)$ with input vector $x \in R^n$ and corresponding binary class labels $y \in \{+1,-1\}$, the training set can be linearly divided into two parts using a hyperplane. Usually, the hyperplane may be expressed as follows:

$$w \cdot x_i + b = 0 \quad (3)$$

where, w is weight vector; b is offset.

In the classification, one of the core ideas of SVM is that the model has high generalization ability, which causes solving w and b optimization problem:

$$\min_{\omega,b} \quad \frac{1}{2}\|w\|^2$$
$$s.t. \quad y_i\left((\omega \cdot x_i)+b\right) \geq 1, \quad i=1,\cdots l \quad (4)$$

For the linear inseparable problem, relaxation variable ξ_i and penalty parameter $C>0$ need to be introduced. So, the formula (4) is turned into the following optimization problem:

$$\min_{\omega,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i$$
$$s.t.\begin{cases} y_i\left((w \cdot x_i)+b\right) \geq 1-\xi_i & i=1,\cdots l \\ \xi_i \geq 0, \quad i=1,\cdots l \end{cases} \quad (5)$$

For deriving the dual problem of the original problem (6), Lagrange function is introduced:

$$L(w,b,\xi,\alpha,\beta)=\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\left(y_i\left((w \cdot x_i)+b\right)-1+\xi_i\right)-\sum_{i=1}^{l}\beta_i\xi_i \quad (6)$$

where, $\eta_i^{(*)}$ and $\alpha_i^{(*)} \geq 0$ are the multiplier vectors of Lagrange. So, its corresponding dual problem is:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j (x_i \cdot x_j)\alpha_i\alpha_j - \sum_{j=1}^{l} a_j$$
$$s.t.\begin{cases} \sum_{i=1}^{l} y_i\alpha_i = 0 \\ 0 \leq \alpha_i \leq C, \quad i=1,\cdots l \end{cases} \quad (7)$$

Solving the formula (7) can obtain the optimum values α^* and b^*. So, the optimal classification decision function can be expressed as:

$$f(x)=\mathrm{sgn}((w^* \cdot x)+b^*)=\mathrm{sgn}(\sum_{i=1}^{n} y_i a_i^* (x_i \cdot x)+b^*) \quad (8)$$

For the nonlinear problem, C-SVC can be through a nonlinear mapping function making the original data map to the appropriate high dimensional feature space. And, data can be analyzed and processed in this space. This mapping function is called kernel function $K(x, x^{'})$, which has to meet Mercer theorem. Namely, its corresponding kernel function matrix is the symmetric positive semidefinite matrix. Then, finding the optimal classification hyperplane can correctly separate from two classes of data points as much as possible in this space. So, the dual problem of nonlinear classification problems is:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j K(x_i, x_j)\alpha_i\alpha_j - \sum_{j=1}^{l} a_j$$
$$s.t.\begin{cases} \sum_{i=1}^{l} y_i\alpha_i = 0 \\ 0 \leq \alpha_i \leq C, \quad i=1,\cdots l \end{cases} \quad (9)$$

Therefore, the corresponding optimal classification decision function is:

$$f(x)=sgn(\sum_{i=1}^{N} y_i\alpha_i^* K(x_i, x)+b^*) \quad (10)$$

For the two classification problems, if $f(x)=1$, evaluation object x belongs to the first class; if $f(x)=-1$, evaluation object x belongs to the second class.

In solving practical problems, according to the characteristic of the problem to select the appropriate kernel function can achieve the nonlinear transform of the original data. Common kernel functions are:

(1) Gauss RBF kernel function:

$$K(x, x^{'})=\exp\left(-\left\|x-x^{'}\right\|^2 \big/ \sigma^2\right)$$

(2) Polynomial kernel functions:

$$K(x, x^{'})=\exp\left(-\left\|x-x^{'}\right\|^2 \big/ \sigma^2\right)$$

(3) Multi-layer perceptron kernel function:

$$K(x,y)=\tanh(ky \cdot x+\theta)$$

III. REALIZATION OF C-SVC ALGORITHN USING MATLAB

This paper selects the Gauss RBF kernel function $K(x,x^{'})=\exp\left(-\left\|x-x^{'}\right\|^{2}/\sigma^{2}\right)$ $(\sigma>0)$ as the C-SVC kernel function, and makes use of MATLAB 10.0 to realize the C-SVC algorithm. The main procedure statement is as follows:

```
function [nsv, alpha, b0] = svc(X,Y,ker,C)
   if (nargin <2 | nargin>4) % check correct number of arguments
      help svc
   else
      fprintf('Support Vector Classification\n')
      fprintf('_____________________________\n')
      n = size(X,1);
      if (nargin<4) C=Inf;, end
      if (nargin<3) ker='linear';, end
         epsilon = svtol(C);
      fprintf('Constructing ...\n');
      H = zeros(n,n);
      for i=1:n
         for j=1:n
            H(i,j) = Y(i)*Y(j)*svkernel(ker,X(i,:),X(j,:));
         end
      end
      c = -ones(n,1);
      H = H+1e-10*eye(size(H));
      vlb = zeros(n,1);
         vub = C*ones(n,1);
      x0 = zeros(n,1);
      neqcstr = nobias(ker); % Set the number of equality constraints (1 or 0)
      if neqcstr
         A = Y';, b = 0;
      else
         A = [];, b = [];
      end
      fprintf('Optimising ...\n');
      st = cputime;
      [alpha lambda how] = qp(H, c, A, b, vlb, vub, x0, neqcstr);
      fprintf('Execution time: %4.1f seconds\n',cputime - st);
      fprintf('Status : %s\n',how);
      w2 = alpha'*H*alpha;
      fprintf('|w0|^2     : %f\n',w2);
      fprintf('Margin     : %f\n',2/sqrt(w2));
      fprintf('Sum alpha : %f\n',sum(alpha));
      svi = find( alpha > epsilon);
      nsv = length(svi);
      fprintf('Support Vectors : %d (%3.1f%%)\n',nsv,100*nsv/n);
         b0 = 0;
      if nobias(ker) ~= 0
         svii = find( alpha > epsilon & alpha < (C - epsilon));
         if length(svii) > 0
            b0 =   (1/length(svii))*sum(Y(svii) - H(svii,svi)*alpha(svi).*Y(svii));
         else
            fprintf('No support vectors on margin - cannot compute bias.\n');
         end
      end
   end
```

IV. EMPIRICAL ANALYSIS

A. Influence Factors of WaterInrush form Coal Floor

Happening of water inrush has to do with geology, hydrology geology, mining activities and so on. According to the characteristics of water inrush accident to determine the influence factors of water inrush are: water pressure, aquifer thickness, water-resisting layer thickness, sand rock ratio, mud rock ratio, coal floor elevation, coal seam dip angle, fault throw, distance to the fault, mining height, mining depth and adopt speed. Among them, the pressure of aquifer is the source power which water inrush occurs. The thickness of the aquifer is one of the parameters of aquifer scale and rich water level, which determines the water and the characteristics of water inrush. Water-resisting layer is the restrain condition of water inrush. The thickness of water-resisting layer and lithologic reflects the ability of water-resisting layer how to block water. The coal seam occurrence condition decides mining engineering decorate. Coal floor elevation and the coal seam in mining space determine the space relationship of aquifer, water-resisting layer and geological structures. Fault and fault distance from the gap reflect the influence of the structure. Height and depth of mining, and adopt speed mainly embody the mine pressure distribution and the destruction of coal floor depth. [12-14] In order to study water inrush from coal floor, this paper use 1 for water inrush, -1 for the opposite. Samples are shown in table 1.

TABLEI

SAMPLE SET OF WATER INRUSH

Number	Aquifer thickness(m)	Water pressure(Mpa)	Water-resisting layer thickness (m)	Sand rock ratio (%)	Mud rock ratio (%)	Coal floor elevation (m)	Coal seam dip angle(°)	Fault throw(m)	Distance to the fault(m)	Mining height(m)	Mining depth(m)	Adopt speed (m·d^{-1})	Water inrush
1	520.00	1.01	45.00	2.89	64.36	119.00	14.00	68.00	75.00	8.00	187.50	0.50	1
2	6.17	2.30	46.91	30.89	53.78	-107.28	11.00	1.00	24.00	1.50	291.28	2.00	-1
3	210.00	2.55	50.00	6.70	63.79	-50.00	13.00	4.00	10.00	8.00	412.40	2.00	1

4	6.20	1.50	38.90	25.53	26.59	-169.70	13.50	1.20	7.00	1.50	369.50	2.00	-1
5	7.98	2.01	28.00	1.89	64.36	10.00	18.00	0.60	10.00	8.00	344.00	0.50	1
6	520.00	0.74	65.00	11.68	53.26	148.00	11.00	79.00	63.00	7.50	175.50	0.50	1
7	265.00	1.33	36.38	3.52	64.36	80.00	7.00	0.80	62.00	8.00	218.80	1.50	-1
8	76.56	0.34	32.65	1.89	53.78	-110.70	6.00	22.00	6.00	0.90	230.00	0.50	-1
9	90.00	2.01	28.00	10.69	49.49	-68.00	18.00	0.60	10.00	8.00	130.00	2.00	1
10	6.20	1.78	38.90	28.41	49.49	-199.50	14.00	1.00	5.00	1.50	369.50	2.00	-1
11	520.00	1.91	43.00	2.00	50.28	14.50	11.00	1.50	2.00	8.00	295.40	1.50	1
12	85.00	0.92	33.61	6.70	45.50	-120.20	8.00	0.50	0.00	1.40	110.00	1.00	-1
13	10.58	1.06	27.79	8.93	50.36	95.65	7.00	0.46	21.00	2.00	310.00	1.50	-1
14	520.00	0.69	42.00	8.93	66.55	178.00	12.00	32.00	19.00	2.00	152.00	0.50	1
15	180.50	2.35	50.00	1.89	64.36	-15.00	16.00	10.00	153.00	2.00	369.10	1.00	1
16	6.20	1.45	38.90	28.41	49.49	-165.20	13.50	3.00	27.00	1.50	369.50	2.00	-1
17	120.00	1.82	26.39	15.65	50.36	20.50	12.00	4.00	16.00	0.80	123.00	1.50	1
18	520.00	1.45	46.00	1.89	60.36	76.00	15.00	2.50	9.00	8.00	230.00	0.50	1
19	6.20	1.91	43.11	36.85	50.36	-68.00	8.00	1.10	130.00	1.50	243.00	2.00	-1

B. Application of Correlation Analysis and SVM Model in Predicting WaterInrush form Coal Floor

• Analyzing the indexes with correlation analysis

Due to the many influence factors of water inrush form coal floor, the high correlation factors are chosen in accordance with correlation analysis in order to construct the water inrush index system. Then, the results of using SPSS [15] soft to analyze the indexes are shown in table 2.

TABLE II

CORRELATIONS OF WATER INRUSH FACTORS FORM COAL FLOOR

		Water inrush			Water inrush
Aquifer thickness	Pearson Correlation	.645**	Coal seam dip angle	Pearson Correlation	.212
	Sig.(2-tailed)	.003		Sig.(2-tailed)	.384
	N	19		N	19
Water pressure	Pearson Correlation	.587**	Fault throw	Pearson Correlation	.371
	Sig.(2-tailed)	.008		Sig.(2-tailed)	.118
	N	19		N	19
Water-resisting layer thickness	Pearson Correlation	.604**	Distance to the fault	Pearson Correlation	.063
	Sig.(2-tailed)	.006		Sig.(2-tailed)	.798
	N	19		N	19
Sand rock ratio	Pearson Correlation	-.554*	Mining height	Pearson Correlation	.600**
	Sig.(2-tailed)	.014		Sig.(2-tailed)	.007
	N	19		N	19
Mud rock ratio	Pearson Correlation	.503*	Mining depth	Pearson Correlation	-.197
	Sig.(2-tailed)	.028		Sig.(2-tailed)	.419
	N	19		N	19
Coal floor elevation	Pearson Correlation	.582**	Adopt speed	Pearson Correlation	.231
	Sig.(2-tailed)	.009		Sig.(2-tailed)	.340
	N	19		N	19

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

According to the analysis results in table 2, this paper selects seven indexes of high correlation with water inrush as input of SVM, including: aquifer thickness, water pressure, water-resisting layer thickness, sand rock ratio, mud rock ratio, coal floor elevation and mining height, and whether water inrush as output, in order to construct the water inrush index system. Then, the seven index data are done standardized processing, which is shown in table 3.

TABLE III

STANDARDIZATION SAMPLE DATA

Number	Aquifer thickness(m)	Water pressure(Mpa)	Water-resisting layer thickness(m)	Sand rock ratio (%)	Mud rock ratio (%)	Coal floor elevation(m)	Mining height(m)	Water inrush
1	1.5265	-0.8472	0.5130	-0.8145	1.0518	1.2071	1.1550	1
2	-0.8759	1.2418	0.7101	1.5870	-0.0496	-0.7944	-0.8289	-1
3	0.0771	1.6466	1.0290	-0.4877	0.9925	-0.2877	1.1550	1
4	-0.8757	-0.0537	-0.1164	1.1273	-2.8804	-1.3465	-0.8289	-1
5	-0.8674	0.7722	-1.2413	-0.9002	1.0518	0.2430	1.1550	1
6	1.5265	-1.2844	2.5769	-0.0606	-0.1038	1.4636	1.0024	1
7	0.3343	-0.3290	-0.3765	-0.7604	1.0518	0.8621	1.1550	-1

8	-0.5468	-1.9321	-0.7614	-0.9002	-0.0496	-0.8246	-1.0120	-1
9	-0.4839	0.7722	-1.2413	-0.1455	-0.4963	-0.4469	1.1550	1
10	-0.8757	0.3997	-0.1164	1.3743	-0.4963	-1.6101	-0.8289	-1
11	1.5265	0.6102	0.3066	-0.8908	-0.4140	0.2828	1.1550	1
12	-0.5073	-0.9929	-0.6623	-0.4877	-0.9117	-0.9086	-0.8594	-1
13	-0.8553	-0.7662	-1.2629	-0.2964	-0.4057	1.0005	-0.6763	-1
14	1.5265	-1.3653	0.2035	-0.2964	1.2798	1.7289	-0.6763	1
15	-0.0608	1.3227	1.0290	-0.9002	1.0518	0.0218	-0.6763	1
16	-0.8757	-0.1347	-0.1164	1.3743	-0.4963	-1.3067	-0.8289	-1
17	-0.3437	0.4645	-1.4074	0.2799	-0.4057	0.3358	-1.0425	1
18	1.5265	-0.1347	0.6162	-0.9002	0.6354	0.8267	1.1550	1
19	-0.8757	0.6102	0.3180	2.0981	-0.4057	-0.4469	-0.8289	-1

• Using MATLAB to classify and predict

Selecting 1 to 12 data as the training samples and inspecting the 13 to 16 samples are to predict the values of the 17 to 19 samples in table 3 using matlab10.0. Prediction results are shown in table 4.

TABLEIV

PREDICTION RESULTS

Number	Actual value	Predicted value	Water inrush
17	1	1	Yes
18	1	1	Yes
19	-1	-1	Not

In table 4, the prediction results of C-SVC the prediction results are accord with actual conditions, which suggests that C-SVC has the strong generalization ability and high prediction accuracy.

V. CONCLUSION

According to the basic principle of correlation analysis and SVM, this paper establishes the prediction model of water inrush. The model makes full use of the advantages of SVM and correlation analysis, which not only can deeply dig sample data value to solve the problem of small sample and nonlinear prediction, but also overcome the correlation between the variables and reduce the numbers of the input variables in order to improve the prediction precision and convergence speed. Through empirical analysis that the prediction results are accord with actual conditions, this shows that the model can accurately predict the condition of water inrush from coal floor. And the model can be directly applied to predict whether water inrush from coal floor occurs on the spot, which can provide the theory basis of ensuring coal mine safety production for taking necessary measures ahead of time.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China(71071003), the MOE Project of Youth Foundation of Humanities and Social Science（09YJC630004）, and the Natural Science Foundation key project of Anhui University (KJ2009A59).

REFERENCES

[1] Teng Weiping, Yu Shanxian, Hu Bo, et al. Study of application of regression method based on support vector machine on Zhejiang drought and flood’s forecast at flood season[J]. Journal of Zhejiang University (Science Edition), 2008, 35(3):343.

[2] Zhang Ruenchu. Multivariate statistical analysis [M]. Science press, 2006.

[3] Gou Ke, Gong Hao. Multivariate statistical method and its application [M]. University of electronic science and technology press, 2003.

[4] Deng Naiyang, Tian Yingjie. Support vector machine——theory, algorithm and development [M]. Science press, 2009.

[5] Vapnik V. The Statistical Learning Theory [M]. New York: Jone Wileg, 1998.

[6] Deng Naiyang, Tian Yingjie. A new method for data mining——Support vector machine[M]. Science press, 2005.

[7] Zhang Xuegong. Introduction to Statistical Learning Theory and Support Vector Machines [J]. Acta Automation Sinica, 2000, 26(1): 32-41.

[8] V. David Sanchez A. Advanced support vector machines and kernel methods [J]. Neurocomputing, 2003, 55: 5-20.

[9] Vapnik V. The nature of statistical learning theory (2nded) [M]. Berlin: Springer, l999.

[10] Vapnik．V．N, Golowich．S, A．Smola. Support Vector Method for Function Approximation, Regression Estimation, and Sigal Processing [J]. In: Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press.9:281-287.1997.

[11] Malin Song, Shuhong Wang, Jie Wu, Li Yang. (2011),“A New Space-Time Correlation Coefficient and its Comparison with Moran’s Index on Evaluation”, Management Decision，Vol. 49 No. 9, pp. 1426-43.

[12] Liu Zaipin. Prediction of water inrush from coal seam based on data mining classification technology [D]. China Coal Research Institute, 2008.

[13] Mi Jinke, Qiao Wei, Yue Zuncai, et al. Study on nonlinear prediction of water inrush Coal mine [J]. Energy Technology and Management, 2011(1).

[14] Li Yang, Malin Song. Coal mine safety evaluation with V-fold cross-validation and bp neural network[J]. Journal of Computers, 2010,5(9):1364-1372.

[15] Yan Zhigang, Bai Haibo, zhang Hairong. A novel SVM model for the analysis and prediction of water inrush from coal mine [J]. China Safety Science Journal, 2008(7).

[16] Li Yang, Malin Song. Formation mechanism of green

strategic alliances and its cooperative system for coal-mining eco-industrial parks based on synthetic decision support system[J]. Journal of Computers, 2009,4(11):1109-1117.

[17] Zhang Hongbing, Jia Laixi, Li Lu. SPSS Treasure knowledge [M]. Publishing House of Electronics Industry, 2005.

[18] Stophen J. Chapman. Matlab programming (4th Edition) [M]. Science press, 2011.

Mr. **Li Yang** is a professor in School of Economics and Management, Anhui University of Science & Technology, Huainan, Anhui, China. His major field of study includes decision theory and methods, emergency management and energy economy. (E-mail: yangli081003@163.com)