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 from 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:

\[ r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \]  

where \( n \) is the sample size; \( x_i \) and \( y_i \) are the sample values of variables. In general, \( -1 \leq r \leq 1 \), and
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:

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

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:

$$w \cdot x + b = 0$$

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:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$\n
s.t. $y_i ((w \cdot x_i) + b) \geq 1, \quad i = 1, \cdots l$ \n
For the linear inseparable problem, relaxation variable $\xi_i$ and penalty parameter $C>0$ need to be introduced. So, the formula (4) is turned into the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i$$\n
s.t. $y_i ((w \cdot x_i) + b) \geq 1 - \xi_i, \quad i = 1, \cdots l$ \n
$$\xi_i \geq 0, \quad i = 1, \cdots l$$

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

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

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

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

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

$$f(x) = \text{sgn}((w^* \cdot x) + b^*) = \text{sgn}\left(\sum_{i=1}^{l} \alpha_i^* (x_i \cdot x) + b^*\right)$$

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:

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

Therefore, the corresponding optimal classification decision function is:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{l} \alpha_i^* K(x_i,x) + b^*\right)$$

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:

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

2. Polynomial kernel functions:

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

3. Multi-layer perceptron kernel function:

$$K(x,x) = \text{tanh}(k y \cdot x + \theta)$$
III. REALIZATION OF C-SVC ALGORITHM USING MATLAB

This paper selects the Gauss RBF kernel function \( K(\mathbf{x}, \mathbf{x}') = \exp \left( -\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2} \right) \) (\( \sigma > 0 \)) 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:

```matlab
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 = size(X,1);
if (nargin<4) C=Inf;, end
if (nargin<3) ker='linear';, end
epsilon = svtol(C);
fprintf('Constructing ...
');
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 ...
');
st = cputime;
[alpha lambda how] = qp(H, c, A, b, vlb, vub, x0, neqcstr);
    fprintf('Execution time: %4.1f seconds
',cputime - st);
    fprintf('Status : %s
',how);
w2 = alpha'*H*alpha;
    fprintf('Sum alpha : %f
',sum(alpha));
svi = find( alpha > epsilon);
nsv = length(svi);
    fprintf('Support Vectors : %d
(%.3f%%/%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.
');
        end
    end
end
```

IV. EMPIRICAL ANALYSIS

A. Influence Factors of Water Inrush form Coal Floor

Happening of water inrush has to do with geology, hydrology, 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.

<table>
<thead>
<tr>
<th>Number</th>
<th>Aquifer thickness (m)</th>
<th>Water pressure (MPa)</th>
<th>Water-resisting layer thickness (m)</th>
<th>Sand rock ratio (%)</th>
<th>Mud rock ratio (%)</th>
<th>Coal floor elevation (m)</th>
<th>Coal seam dip angle (°)</th>
<th>Fault throw (m)</th>
<th>Distance to the fault (m)</th>
<th>Mining height (m)</th>
<th>Mining depth (m)</th>
<th>Adopt speed (m·d⁻¹)</th>
<th>Water inrush</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>520.00</td>
<td>1.01</td>
<td>45.00</td>
<td>2.89</td>
<td>64.36</td>
<td>119.00</td>
<td>14.00</td>
<td>68.00</td>
<td>75.00</td>
<td>8.00</td>
<td>187.50</td>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6.17</td>
<td>2.30</td>
<td>46.91</td>
<td>30.89</td>
<td>53.78</td>
<td>-107.28</td>
<td>11.00</td>
<td>1.00</td>
<td>24.00</td>
<td>1.50</td>
<td>291.28</td>
<td>2.00</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>210.00</td>
<td>2.55</td>
<td>50.00</td>
<td>6.70</td>
<td>63.79</td>
<td>-50.00</td>
<td>13.00</td>
<td>4.00</td>
<td>10.00</td>
<td>8.00</td>
<td>412.40</td>
<td>2.00</td>
<td>1</td>
</tr>
</tbody>
</table>
B. Application of Correlation Analysis and SVM Model in Predicting Water Inrush from Coal Floor

Analyzing the indexes with correlation analysis. Due to the many influence factors of water inrush from 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 from Coal Floor

<table>
<thead>
<tr>
<th>Water Inrush</th>
<th>Aquifer thickness</th>
<th>Water Pressure</th>
<th>Water-resisting Layer Thickness</th>
<th>Sand Rock Ratio</th>
<th>Mud Rock Ratio</th>
<th>Coal Floor Elevation</th>
<th>Mining Height</th>
<th>Mining Depth</th>
<th>Adopt Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson Correlation</td>
<td>Sig.(2-tailed)</td>
<td>Pearson Correlation</td>
<td>Sig.(2-tailed)</td>
<td>Pearson Correlation</td>
<td>Sig.(2-tailed)</td>
<td>Pearson Correlation</td>
<td>Sig.(2-tailed)</td>
<td>Pearson Correlation</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Water Inrush</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Number</th>
<th>Aquifer thickness (m)</th>
<th>Water pressure (Mpa)</th>
<th>Water-resisting layer thickness (m)</th>
<th>Sand rock ratio (%)</th>
<th>Mud rock ratio (%)</th>
<th>Coal floor elevation (m)</th>
<th>Mining height (m)</th>
<th>Water inrush</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5265</td>
<td>0.8472</td>
<td>0.5130</td>
<td>-0.8145</td>
<td>1.0518</td>
<td>1.2071</td>
<td>1.1550</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-0.8759</td>
<td>1.2418</td>
<td>0.7101</td>
<td>-1.5870</td>
<td>-0.0496</td>
<td>-0.7944</td>
<td>-0.8289</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>0.0771</td>
<td>1.6466</td>
<td>1.0290</td>
<td>-0.0487</td>
<td>0.9925</td>
<td>-0.2877</td>
<td>-0.1528</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>-0.8757</td>
<td>-0.0537</td>
<td>-0.3164</td>
<td>1.1273</td>
<td>-2.8804</td>
<td>-1.3465</td>
<td>-0.8289</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>-0.8674</td>
<td>0.7722</td>
<td>-1.2413</td>
<td>-0.9002</td>
<td>1.0518</td>
<td>0.2430</td>
<td>1.1550</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1.5265</td>
<td>-1.2844</td>
<td>2.5769</td>
<td>-0.0606</td>
<td>-0.1038</td>
<td>1.4636</td>
<td>1.0024</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.3343</td>
<td>-0.3290</td>
<td>-0.3765</td>
<td>-0.7604</td>
<td>1.0518</td>
<td>0.8621</td>
<td>1.1550</td>
<td>-1</td>
</tr>
</tbody>
</table>
accurately predict the condition of water inrush from coal floor with actual conditions, this shows that the model can accurately predict the condition of water inrush from coal floor.

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, 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.

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.

<table>
<thead>
<tr>
<th>Number</th>
<th>Actual value</th>
<th>Predicted value</th>
<th>Water inrush</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>19</td>
<td>-1</td>
<td>-1</td>
<td>Not</td>
</tr>
</tbody>
</table>

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

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


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)