“2014 ISSST”, 2014 International Symposium on Safety Science and Technology

Research on prediction of gas emission based on self-organizing data mining in coal mines

LI Runqiu\textsuperscript{a,b}, SHI Shiliang\textsuperscript{a,b}, WU Aiyou\textsuperscript{a,b}, LUO Wenke\textsuperscript{a,b}, ZHU Hongping\textsuperscript{c,}\textsuperscript{*}

\textsuperscript{a}School of Energy & Safety Engineering, Hunan University of Science & Technology, Xiangtan 411201, Hunan, China
\textsuperscript{b}Hunan Provincial Key Laboratory of Safe Mining Techniques of Coal Mines, Xiangtan 411201, Hunan, China
\textsuperscript{c}School of Information & Electronics Engineering, Hunan University of Science & Technology, Xiangtan 411201, Hunan, China

Abstract

In order to accurately predicting gas emission in coal mines, the complicated nonlinear characteristics of gas emission was analysis, the prediction method was put forward for gas emission based on self-organizing data mining. It was used the ternary quadratic polynomial for the local function and the original variable was used in each generation, and the minimum deviation principle was used for criteria of model selected. And then, the high-order equation of prediction was established for gas emission by self-organizing data mining method. The fitness relative error of this prediction model was ±0.03\% and predictive relative error was ±1.45\% to gas emission in coal mine. The results show that: self-organizing data mining method can automatically analyze non-linear relation between the gas emission and the factors, and can be establish the explicit high order equation to descript the gas emission laws, and the prediction model has enough prediction accuracy for application of actual engineering in coal mines.

Keywords: gas emission, prediction, self-organizing data mining (SODM), group method of data handling (GMDH), safety, coal mine.

Nomenclature

\begin{tabular}{|l|}
\hline
y & absolute of gas emission (m$^3$/min) \\
\hline
\end{tabular}

* Corresponding author. Tel.: +86-139-7321-7556,
E-mail address: lirq_hn@sina.com
Project 51274100 Supported by the National Natural Science Foundation of China. Project 14C0424 Supported by the Hunan province education department of China.
1. Introduction

It is a fundamental safety work to calculating the gas content in coal seam and forecasting the gas emission from coal and rock in coal mine. In the 1950s, it has been established some measurement methods and processes of gas content in coal bed, and has been put forward and applied the mine statistical method to calculate and predict mine gas content and emission in coal mines in China. In the 1980s, the method of different-source prediction for gas emission also has been proposed and applied. Since then, the method of analogy method, the gas geology mathematical model, velocity method and other prediction methods for gas emission have been put forward and applied. After decades of exploration and study by researcher and scholar, it has been formed the maturity and traditional prediction method and technology for gas emission in coal mines that was suitable the coal seam occurrence conditions in China. All the methods and technologies was offer a scientific basis to design and retrofit for the new and old coal mine. With the expansion of coal mining intensity and production scale, people has deepening understood and grasped the essential feature of mine gas emission system by the development and its application of computer technology and mathematics method and the nonlinear theory. So, many new prediction methods has been to research and application for coal seam gas emission, For example, based on multivariate linear regression to prediction gas emission [1], based on the grey system theory [2], neural network [3], support vector machine (SVM) [4–5], evidence theory [6], chaos theory [7], fractal theory [8], rough set theory [9], and so on methods to prediction gas emission in coal mines. The research and application of these new methods have been to promoting the improvement of the level of research on mine gas emission rule, and improved and raised the level of the coal mine safety production.

Self-Organizing Data Mining Method (SODM) is a series of algorithms to solving complicated and nonlinear systems based on Group Method of Data Handling (GMDH) which proposed by Ivakhnenko academician in Ukraine academy of sciences in 1967. This method can overcome the problems of neural network such as can't explain the results of practical significance, over-fitting, and also have the advantages of using a priori knowledge of system, inductive learning algorithm process. So, it have widely useful[10–13]. In 1980s, the self-organizing data mining method was introduced into China. Then, some scholars has studied and applied the method, such as economic early warning based on self-organizing data mining method [14]; Power demand predict [15]; E-commerce customer prediction [16], etc., but its application range are mainly concentrated in the economic, financial, business, and so on, it was rarely found in study of safety and disaster prevention and control in coal mine.

There have a lots of factors to influence the mine gas emission in coal mines such as gas content, coal and rock mining scale, production process, ventilation mode and others. The subsystem of gas emission has typical characteristics of complex nonlinear system, we still can't know exactly how much variables affecting this subsystem state, and also cannot establish a deterministic system dynamics equation to describe the gas emission [17]. The self-organizing data mining method is a heuristic self-organization method to study complex multivariable system. When in the case of unknown nonlinear systems structure, this method can be established its mathematical model according to a small amount of input and output data. Therefore, this paper attempts to introduce self-organizing data mining method to analysis the coupling relationship between coal mine gas emission factors for the coal mine gas emission prediction research, and to establish the prediction model. It would be to provide new methods for the prevention and control of gas disaster and to improve the level of coal mine disaster control in coal mine.

2. The basic principle of self-organizing data mining

The basic idea of the SODM is using the general iterative algorithm to establish a simple initial model then to gradually build a model of optimal complex according to internal criteria and external criteria by the samples of complex system. The new model structure is more complicated than the original model in each step and still retain certain evolution characteristics of the original system structure. This entire modelling process is completely self-
organizing process [12–13]. The basic principle of SODM are as follows:

For the complex systems, suppose there are \( m \) input variable and an output variable \( X = [x_1, x_2, \ldots, x_m] \), the inputs and output will be exist a relation function which were as follows:

\[
y = f(x_1, x_2, \ldots, x_m)
\]

(1)

The function \( f \) can be expanded into discrete forms by use the Volterra polynomial functional series:

\[
y = b_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m} \sum_{j=1}^{m} b_{ij} x_i x_j + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} b_{ijk} x_i x_j x_k + \cdots
\]

(2)

where, \( b_0, b_i, b_{ij}, b_{ijk} \) is for undetermined coefficients.

Equation (2) is also called the Kolmogorov-Gabor polynomial. In theory, to an arbitrary nonlinear function, it can be approximation through the Equation (2). But with the increase of input variables, the number of items in Equation (2) will increase sharply. In other words, the amount of undetermined coefficients of the model will be increase significantly. At the same time, the required samples will increase sharply, and it also can cause the rapid increase of computing workload and cause some problem about the instability of calculation.

To solve the above problems, the self-organizing data mining method is put forward a way which through the “hierarchical partial implementation” to achieve “full implementation”. It use any two input variable to establish a completely polynomials in the first generation. The completely polynomials is commonly used binary quadratic polynomial completely, such as:

\[
y^{(1)}_{\mu} = f^{(1)}_{\mu}(x_i, x_j) = a^{(1)}_{\mu ij} + a^{(1)}_{\mu i} x_i + a^{(1)}_{\mu j} x_j + a^{(1)}_{\mu ij} x_i x_j + a^{(1)}_{\mu i} x_i^2 + a^{(1)}_{\mu j} x_j^2
\]

(3)

where, \( y^{(1)}_{\mu} \) is the output in first generation, \( p = 1, 2, \ldots, m(m-1)/2 \), \( x_i, x_j \) is the input, \( i, j = 1, 2, \ldots, m \), \( i \neq j \), \( a^{(1)}_{\mu ij}, a^{(1)}_{\mu i}, a^{(1)}_{\mu j}, a^{(1)}_{\mu ij}, a^{(1)}_{\mu i} \) is for undetermined coefficients. The superscript (1) means the first generation.

The Equation (3) is also called the local function of complex system modelling, the undetermined coefficient is generally estimate use sample data by the least square method to make minimum mean square error. There, the minimum sum of mean square error is the internal criteria. So that we can build \( m(m-1)/2 \) local function, and then leave a part of the above “best” local function by some external criteria such as the sum mean square error is minimum. The output of the best local function will be as the inputs in second generation to build a new binary quadratic polynomial. By this way, we can establish the model in the third generation, the fourth generation, etc. such as following:

\[
y^{(r)}_{q} = f^{(r)}_{q}(y^{(r-1)}_{k}, y^{(r-1)}_{l})
\]

(4)

where, superscript \( r \) is for generation, \( q \) is the number of local function in \( r \) generation, \( k \) and \( l \) is the number of local function \( r-1 \) generation, and \( k \neq l \).

Clearly, the self-organizing data mining method can be build a large number of local function in each generation. At the same time, some of best local function was be selected as input in the next generation by the external criteria. The termination rule of whole algorithms is the complexity and optimal principle, which means with the function complexity gradually increases, the model of external criterion value is increase in the firstly and then decreases, when the criterion value is minimum, the function is the optimal complexity model[8].

On the one hand, when using the self-organizing data mining method to modelling, it does not need to set parameters and model forms in advance, and also does not pay attention to whether there is a certain relationship between the input variables, it only by the self-organization to determine the relationship between the input variable and output variable. On the other hands, at the end of the self-organizing data mining method to build model, it may be only reserve the some input variables which influence is bigger to output and rejection some input variables which influence is smaller to output.

3. The self-organizing data mining method of gas emission

The production system of coal mines is a large and complex dynamics system coupled time-space, the gas emission is a complicated nonlinear process [17] which has a lots of influence factors. And the sample data about gas emission are always independent of each other and often show different degrees of nonlinear that obtained from the monitoring in coal mines. So, we can use Kolmogorov-Gabor polynomials to establish and approximate the
unknown prediction model of gas emission according to the self-organizing data mining method.

That the binary quadratic polynomial as local function is the most widely used in the process of modelling by using self-organizing data mining method. To achieve optimal complexity as soon as possible, it have the following changes for the self-organizing data mining modelling method. One is the original variables as next inputs with the output of previous generation to establish local function in the second and later generation in the process. The next changes is the ternary quadratic polynomial was used for the local function because each generation added to the original variable.

For example, to the subsystem of gas emission in coal mines, \( y \) is gas emission as output variables, \( X=[x_1, x_2, \ldots, x_n] \) is the factors affecting gas emission as input variables, specific modelling process and steps are as follows:

Firstly, sample data is be group handled. \( W \), the sample-set of gas emission, is divided into three sets, train-set \( A \), test-set \( B \) and predict-set \( C \), where \( W=A \cup B \cup C \), \( A \cap B=\emptyset \), \( A \cap C=\emptyset \), \( B \cap C=\emptyset \).

Secondly, local function is established between \( y \) and \( X \). The local function is to be established by using the ternary quadratic polynomial according to Equation (5) in the first generation. The undetermined coefficients of Equation (5) is estimated through the least squares method by the train-set \( A \).

\[
y_p^{(1)} = f_p^{(1)}(x_i, x_j, x_k) = a_{0p}^{(1)} + a_{1p}^{(1)} x_i + a_{2p}^{(1)} x_j + a_{3p}^{(1)} x_k + a_{4p}^{(1)} x_i x_j + a_{5p}^{(1)} x_i x_k + a_{6p}^{(1)} x_j x_k + a_{7p}^{(1)} x_i^2 + a_{8p}^{(1)} x_j^2 + a_{9p}^{(1)} x_k^2
\]

(5)

where, \( y_p^{(1)} \) is the output of first generation, \( p=1, 2, \ldots, m(m-1)/2 \), \( x_i, x_j, x_k \) is the inputs, \( i, j, k =1,2,\ldots,m \), \( i \neq j, i \neq k, j \neq k \), \( a_{0p}^{(1)}, a_{1p}^{(1)}, \ldots \) is the undetermined coefficients, the superscript (1) means the first generation.

Thirdly, selection of local function. Some best local function is to be keep on through the appropriate external criterion by use test-set B.

Fourthly, the new local function is to be established. Using the original input variables and the output of selected local function above steps, the new local function is to be established as Equation (5). In the new local function, it must be included output of local function in each step as the input variables.

Fifthly, it repeat steps thirdly and fourthly and will be continuously to create local functions of the subsequent generation until the standards obtain the minimum value of external criterion, then, the optimal complexity model is obtained and the algorithm is terminates.

Finally, to forecast the gas emission. It will be forecast the gas emission prediction-set by using the optimal complexity model.

4. Examples of its application

This is a coal mine with coal and gas outburst belongs to Hunan coal mine enterprise group. The coal seam permeability is low and mine gas reserves is about 400 Mm\(^3\) in this coal mine. The main working bed thickness is average 2.5 m, dip angle of coal seam average is 51\(^\circ\), and thickness of adjacent coal seam is 2.2 m on average and have an average of 27 m to main working seam space. The long of mining workface is about 75 m and advancing speed is 3.5 m/d and mining intensity is about 725 t/d. The average coal seam gas content is about 7.25 m\(^3\)/t and adjacent layer average gas content is about 7.69 m\(^3\)/t. In the process of mining in this coal mine, a lot of factors about safety and disasters have been carried on the comprehensive monitoring to effectively control the gas disasters.

For convenience of analysis, we selected seven factors about gas emission (\( y \), m\(^3\)/min) to study. The seven factors are gas content (\( x_1 \), m\(^3\)/t), coal thickness (\( x_2 \), m), velocity (\( x_3 \), m/d), production (\( x_4 \), t/d), and the gas content of adjacent layer (\( x_5 \), m\(^3\)/t), adjacent layer thickness (\( x_6 \), m), adjacent layer spacing (\( x_7 \), m). The twenty samples is in Table 1 that selected from the monitoring data and it was be used to establish the prediction model for gas emission by the self-organizing data mining method.

This twenty samples were divided into three subsets: train-set, test-test and predict-set. The number of first to sixteenth is as the train-set and test-set, the number of seventeenth to twentieth is as prediction-set. The ahead 16 samples is use “leave one” method to form the train-set and test-set. This means is randomly selected 15 samples from the 16 samples as train-set and the rest one is as test-set. Undetermined coefficients of local function is estimated by least squares. The external criterion is use “minimum deviation rules”, that is to found the all function of minimum variance of fitness between train-set and test-set. Programming with Matlab according to the ahead programs and steps of gas emission self-organizing data mining methods by using train-set and test-set. At the fifth
generation, the model was convergence. The final model and local function in each generation is as Equation (6):

\[
\begin{align*}
    y &= 0.0896 + 1.0215 y_{29}^4 - 0.0411 x_3 - 0.0371 x_5 - 0.0087 x_4 y_{29}^4 - 0.0084 x_3 y_{29}^4 \\
    &+ 0.0126 x_3 x_5 + 0.0051 (y_{29}^4)^2 + 0.0046 (x_5)^2 + 0.0034 (x_3)^2 \\
    y_{29}^4 &= 0.1770 + 1.0186 y_{32}^3 - 0.0711 x_1 - 0.0729 x_3 - 0.0083 x_1 y_{32}^3 - 0.0053 x_3 y_{32}^3 \\
    &+ 0.0146 x_1 x_3 - 0.0022 (y_{32}^3)^2 + 0.0076 (x_1)^2 + 0.0085 (x_3)^2 \\
    y_{32}^3 &= 1.1716 + 0.6194 y_{15}^3 + 0.0772 x_3 - 0.4524 x_2 + 0.0523 x_2 y_{15}^3 + 0.0509 x_3 y_{15}^3 \\
    &+ 0.0090 x_2 x_3 + 0.0057 (y_{15}^3)^2 - 0.00376 (x_2)^2 + 0.0378 (x_3)^2 \\
    y_{15}^3 &= 0.1287 + 4.1132 y_{11}^5 - 3.2843 x_5 + 0.0497 x_4 - 2.2750 x_5^2 - 0.3875 x_4 y_{11}^5 \\
    &+ 0.7383 x_2 x_3 + 1.2152 (x_5)^2 - 0.2012 (x_4)^2 \\
    y_{11} &= -6.5219 - 1.4520 x_4 + 4.9263 x_3 + 1.9977 x_1 - 0.6751 x_3 x_5 + 0.8697 x_3 x_7 \\
    &- 0.9192 x_1 x_3 + 0.0198 (x_7)^2 + 0.0062 (x_5)^2 - 0.0670 (x_4)^2
\end{align*}
\]

Table 1. Gas emission in coal mine.

<table>
<thead>
<tr>
<th>No.</th>
<th>gas emission ((y, \text{m}^3/\text{min}))</th>
<th>gas content ((x_1, \text{m}^3/\text{t}))</th>
<th>coal thickness ((x_2, \text{m}))</th>
<th>velocity ((x_3, \text{m/d}))</th>
<th>production ((x_4, \text{t/d}))</th>
<th>gas content of adjacent layer ((x_5, \text{m}^3/\text{t}))</th>
<th>adjacent layer thickness ((x_6, \text{m}))</th>
<th>adjacent layer spacing ((x_7, \text{m}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.84</td>
<td>6.95</td>
<td>2.5</td>
<td>3.4</td>
<td>725</td>
<td>7.04</td>
<td>2.5</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>4.29</td>
<td>7.37</td>
<td>2.6</td>
<td>3.6</td>
<td>727</td>
<td>7.88</td>
<td>2.1</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>4.71</td>
<td>7.59</td>
<td>2.1</td>
<td>3.5</td>
<td>789</td>
<td>7.63</td>
<td>2.4</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>4.86</td>
<td>7.84</td>
<td>2.5</td>
<td>3.8</td>
<td>768</td>
<td>7.91</td>
<td>2.1</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>3.88</td>
<td>7.36</td>
<td>2.3</td>
<td>3.5</td>
<td>637</td>
<td>8.18</td>
<td>2.2</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>5.03</td>
<td>7.62</td>
<td>2.5</td>
<td>4.2</td>
<td>769</td>
<td>8.09</td>
<td>2.4</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>4.58</td>
<td>7.94</td>
<td>1.9</td>
<td>3.7</td>
<td>692</td>
<td>8.30</td>
<td>2.5</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>4.09</td>
<td>7.28</td>
<td>2.6</td>
<td>3.8</td>
<td>669</td>
<td>7.37</td>
<td>2.2</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>4.40</td>
<td>7.48</td>
<td>2.6</td>
<td>3.3</td>
<td>739</td>
<td>8.03</td>
<td>2.8</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>4.04</td>
<td>7.51</td>
<td>2.3</td>
<td>4.1</td>
<td>658</td>
<td>7.56</td>
<td>2.6</td>
<td>28</td>
</tr>
<tr>
<td>11</td>
<td>4.42</td>
<td>7.02</td>
<td>2.9</td>
<td>3.6</td>
<td>738</td>
<td>7.66</td>
<td>2.0</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>3.22</td>
<td>6.86</td>
<td>2.8</td>
<td>3.8</td>
<td>666</td>
<td>6.92</td>
<td>2.1</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>3.58</td>
<td>7.28</td>
<td>3.6</td>
<td>3.1</td>
<td>684</td>
<td>7.66</td>
<td>2.9</td>
<td>27</td>
</tr>
<tr>
<td>14</td>
<td>4.32</td>
<td>7.94</td>
<td>3.0</td>
<td>3.4</td>
<td>681</td>
<td>8.35</td>
<td>1.7</td>
<td>23</td>
</tr>
<tr>
<td>15</td>
<td>4.74</td>
<td>7.63</td>
<td>2.6</td>
<td>2.9</td>
<td>717</td>
<td>8.37</td>
<td>1.9</td>
<td>25</td>
</tr>
<tr>
<td>16</td>
<td>3.80</td>
<td>6.86</td>
<td>3.0</td>
<td>3.4</td>
<td>779</td>
<td>7.72</td>
<td>2.2</td>
<td>24</td>
</tr>
<tr>
<td>17</td>
<td>4.07</td>
<td>7.02</td>
<td>2.7</td>
<td>3.8</td>
<td>792</td>
<td>7.79</td>
<td>2.1</td>
<td>26</td>
</tr>
<tr>
<td>18</td>
<td>4.23</td>
<td>6.86</td>
<td>2.5</td>
<td>3.4</td>
<td>787</td>
<td>7.37</td>
<td>2.4</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>4.30</td>
<td>7.38</td>
<td>2.6</td>
<td>3.0</td>
<td>768</td>
<td>7.23</td>
<td>2.3</td>
<td>28</td>
</tr>
<tr>
<td>20</td>
<td>4.15</td>
<td>6.78</td>
<td>2.6</td>
<td>3.1</td>
<td>757</td>
<td>6.90</td>
<td>2.3</td>
<td>27</td>
</tr>
</tbody>
</table>

The Fig. 1 is showed the fitting relative error of the 16 samples according to the Equation (6), the relative error of Equation (6) is ±0.03%, the results is conform to the requirements of the modelling with train-set and test-set.

We also known from the Equation (6) that the final model and local function in each generation is only retained the six factors of gas emission which were chosen, and the factors, adjacent layer thickness \((x_6)\), is to be eliminated. From the point of the actual situation of gas emission factors in coal mines, the influence of adjacent layer thickness
in above seven factors is relatively weaker than other factors, the modelling results are basically in line with the actual situation. So, we can use Equation (6) to undertake gas emission prediction.

\[ \text{Fig. 1. Relative error of train-set and test-set.} \]

The Table 2 is showed the prediction of gas emission by using Equation (6). The prediction relative error of above four samples is from -1.44% to 1.44%, and the average absolute relative error of that is 1.25%. This is means the prediction precision was achieved the higher to meet the application of actual practical engineering.

Table 2. Prediction of gas emission in coal mine.

<table>
<thead>
<tr>
<th>No.</th>
<th>actual gas emission (m³/min)</th>
<th>prediction of gas emission (m³/min)</th>
<th>residual error(m³/min)</th>
<th>relative error(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>4.07</td>
<td>4.0118</td>
<td>-0.0582</td>
<td>-1.4304</td>
</tr>
<tr>
<td>18</td>
<td>4.23</td>
<td>4.2695</td>
<td>0.0395</td>
<td>0.9341</td>
</tr>
<tr>
<td>19</td>
<td>4.30</td>
<td>4.2420</td>
<td>-0.0580</td>
<td>-1.3494</td>
</tr>
<tr>
<td>20</td>
<td>4.15</td>
<td>4.2032</td>
<td>0.0532</td>
<td>1.2816</td>
</tr>
</tbody>
</table>

5. Conclusions

(1) According to the nonlinear characteristics of gas emission in coal mines, the gas emission predicting method was put forward based on the self-organizing data mining. This method can be automatic analysis of coupling relationship between the influence factor of gas emission and gas emission which selected six factors from sevens to establish prediction model for gas emission. The result of selection of factors accords with practice of coal mine safety production.

(2) The prediction models of gas emission in coal mines was established based on self-organizing data mining. The sample fitting relative error is ±0.03%, and the relative prediction error is ±1.45% by using the model. The accuracy of prediction model is higher to satisfy the engineering application.

(3) It is difficult to established an equation to exact describe the influence factors of coal mine gas emission because it has a lots of factors to influence the gas emission and nonlinear and complex characteristics is exist in these factors. The self-organizing data mining method can analyse the influence factors of the coupling relationship between the gas emission factors and outputs and can be to establish high order equation to accurately depict the gas emission. That provides a new method for coal mine gas emission prediction research.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.51274100) and supported by the Hunan province education department of China (No.14C0424). Their supports are acknowledged with thanks. We wish to express our sincere appreciations to those who made suggestions for improvement.
References


