Prediction of Coal-Bed Methane Production Based on PCA and SVM

Dongqin He¹, Shun Zhan¹, Pengbo He¹, Han Lu²*, and Xiaoya Chen²
¹ Engineering technology research institute of Petrochina coalbed methane co. LTD, Beijing, China
² School of Science, Southwest Petroleum University, Chengdu, Sichuan, China

*Corresponding author e-mail: luhan_1024@outlook.com

Abstract. Accurate prediction of oil production is difficult because there are so many factors affecting production. In this paper, new screening rules are used for principal component analysis and factor screening, the support vector machine is used to complete subsequent learning and prediction work, and Gaussian regression is selected to determine the dividing line of prediction results. The results show that the model is more accurate in predicting Wells with production below 2000 cubic meters, and more accurate in predicting Wells with production above 2000 cubic meters.

Keywords: Yield Prediction, Principal Component Analysis, Support Vector Machine

1 Introduction
Production prediction, which can not only provide indexes for construction design before fracturing, but also provide references for construction evaluation after fracturing, is a topic of special concern to oilfield. With the development of oil and gas industry, more and more gas wells need to be increased. In the case of various formation factors, logging data and construction schemes, it is of great significance to accurately predict the output of Wells to be constructed. Too many factors have also invisibly increased the difficulty of forecasting. Traditional prediction methods include linear regression, time series, gray theory, etc., and many other prediction methods have been developed later, such as the neural network method, the chaos theory method, etc. Lai Fp et al. used the moving average method to predict oil well production [11]. Sun Jm et al. applied node analysis method to study the productivity prediction of oil Wells [12]. Miao Hp and Wang Hx published "prediction of horizontal well yield after pressure and optimization of fracture number" on "oil drilling and production technology". In the calculation of multi-fracture production prediction, the prediction period is divided into three stages, which greatly reduces the calculation amount [13]. In addition, support vector machine (SVM) is usually used to forecast grain yield. In this paper, the prediction analysis will be combined with various methods to evaluate the model of the coal-bed gas production.
2 Model Building

2.1 Data Preprocessing

The data used in this study are fracturing files, logging files and drainage and production files from an oilfield in Shanxi. Since the main influencing factors cannot be defined precisely before the prediction, most of the parameters provided in fracturing files, logging files and scheduling files are extracted from the original data files, and some additional influencing factors based on field experience are added. The first 2/3 Wells were used as the training set and last 1/3 Wells were used as the prediction verification set for simulation experiments [4]. Different evaluation indexes often have different dimensions and orders of magnitude, which will affect the results of data analysis. We can solve the problem of comparability between indicators by normalizing the data, so as to eliminate the dimensional influence between indicators [8], which is suitable for comprehensive comparison and evaluation [6]. The transformation form is:

\[ x_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

Among them, \( x_i \) represents the \( i^{\text{th}} \) value in the data column to be normalized, \( x_{\text{max}} \) represents the maximum value in the data, \( x_{\text{min}} \) represents the minimum value in the data.

2.2 Principal Component Analysis Model

Principal component analysis index and multivariate minority comprehensive index method, which emphasize the similarity and difference between data, are essentially based on second-order statistical calculation, and represent the main information of original data in the cube through effective transformation from dimensional characteristic components, and the optimal realization of variance. This method attempts to recombine a number of the original indicators with a certain correlation (such as P indicator) and the original indicators are replaced with a set of unrelated comprehensive indicators. The usual mathematical treatment is to take the original P exponential as a linear combination, as a new composite exponential [7]. The most classical method is to represent the variance of characteristic F1, that is, the larger the Var(F1) is, the more information F1 contains. In order to effectively reflect the original information, the existing information of F1 does not need to re-appear in F2, which is required mathematically:

\[ \text{cov}(F_1, F_2) = 0 \]  

Thus, F2 is called the second principal component, and so on, third, fourth... All the way to the \( p^{\text{th}} \) principal component. In this study, the program prepared by MATLAB was used for principal component analysis, and the factors affecting the yield were processed. Project the original data into the principal component analysis space, extract the principal component of the subject evaluation and the principal component of the evaluation according to the contribution rate, calculate the load of each principal component and each evaluation index, and establish the factor selection model based on the principal component analysis.

Due to the complexity and diversity of underground conditions, the principal component analysis method should not copy the original thinking, but should be improved according to different conditions. Through the observation of the data, it is found that the geographical environment itself is an extremely irregular factor, and the depth of coal seam, the length of crack and other factors are all objective factors underground, so the data are quite different. If the variance is chosen as the only criterion for determining the principal component, it is easy to choose factors such as the length of fluctuation as the principal component. To avoid this situation, we introduce a new criterion based on variance, namely gradient criterion. Since the data are discrete points, the gradient matrix of each
factor to the dependent variable can be used to represent the influence ability of each factor to the dependent variable. We use $\|\text{var-diff}\|_2$ to select the first principal component, and then calculate the covariance to find the following principal component.

2.3 Support Vector Machine Model

The basic idea of SVM is that a nonlinear mapping $\Phi$ is used to map data to a high-dimensional feature space, then a regression estimation function is constructed in the high-dimensional feature space, and then the mapping is returned to the original space $[8]$. By defining the appropriate kernel function $K(x_i, x_j)$, the nonlinear transformation is realized. In addition, considering that some samples cannot be separated from the hyperplane and classified correctly, we generally adopt relaxation variables to solve this problem. Therefore, the optimization equation of the support vector regression model can be expressed as:

$$
\max_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \\
\text{s.t.} \begin{align*}
& y_i - f(X_i) \leq \varepsilon + \xi_i, \xi_i \geq 0 \\
& f(X_i) - y_i \leq \varepsilon + \xi_i^*, \xi_i^* \geq 0
\end{align*}
$$

where $\varepsilon$ is a hyperactive parameter that determines the width of the interval boundary. Lagrangian function and dual problem can be obtained by introducing Lagrangian multiplier $\alpha, \alpha^*, \mu, \mu^*$:

$$
L(w, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) - \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} \mu_i^* \xi_i^*
+ \sum_{i=1}^{N} \alpha_i [i \times f(X_i) - y_i - \varepsilon - \xi_i] + \sum_{i=1}^{N} \alpha_i^* [i \times f(X_i) - y_i - \varepsilon - \xi_i^*]
$$

Its dual problem can be expressed as follows:

$$
\max_{\alpha, \alpha^*} \sum_{i=1}^{N} \left[ y_i \alpha_i^* - \alpha_i \right] - \varepsilon (\alpha_i^* + \alpha_i) \\
\text{s.t.} \begin{align*}
& \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0 \\
& 0 \leq \alpha_i, \alpha_i^* \leq C
\end{align*}
$$

By solving the dual problem, we can get the function of SVR:

$$
f(X) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) X_i^T X + b
$$

SVR can obtain nonlinear regression results through kernel function. For the above problems, Gaussian kernel function can be selected, because the Gaussian kernel function only contains an unknown parameter $\sigma$, which is easy to optimize $[8]$. Gaussian kernel function can be expressed as:

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

where $\sigma > 0$. Currently, there are three unknown parameters affecting SVM learning performance, namely $\sigma$ in kernel function, $\varepsilon$ in loss function and $C$ in target function. The performance of
SVM largely depends on the selection of parameters. Therefore, the grid search method is adopted to optimize the parameters \(^{10}\), and the final choice is:

\[
C = 32, \sigma = 0.004, \varepsilon = 0.01
\]  

(8)

3 Yield Prediction

3.1 Factor Screening

In this chapter, the model is applied to the production prediction of a block in Shanxi. Examples of oil well data obtained are shown in the following table:

Table 1. Sample of raw data

| Well no. | Average daily production (m\(^3\)) | Coalbed depth (m) | Coalbed thickness (m) | Halfseam length (m) | Formation pressure (MPa) |
|----------|-----------------------------------|-------------------|-----------------------|----------------------|--------------------------|
| 01#      | 890                               | 1224              | 4                     | 67                   | 7.73                     |
| 02#      | 180                               | 1282              | 4                     | 95.3                 | 12.56                    |
| 03#      | 620                               | 3188              | 4                     | 151.4                | 42.12                    |
| 04#      | 500                               | 1278.25           | 4.5                   | 76.8                 | 12.53                    |
| 05#      | 610                               | 1781              | 8                     | 88.6                 | 11.4                     |
| 06#      | 1040                              | 2021              | 8                     | 83.5                 | 14.4                     |

There were 113 Wells with a total of 32 factors. Due to too many well counts and factors, the above table only shows the data of the first 5 factors of the first 6 Wells in the complete table. All the above data are from the coal bed methane mining site, true and valid. This paper will take this set of data as an example to combine and predict CBM production. We use the method in section 2.1 to normalize the data, and the results are shown in the following table:

Table 2. Sample of normalized data

| Well no. | Average daily production | Coalbed depth | Coalbed thickness | Halfseam length | Formation pressure |
|----------|--------------------------|---------------|-------------------|-----------------|--------------------|
| 01#      | 0.076                    | 0.026         | 0.000             | 0.064           | 0.059              |
| 02#      | 0.002                    | 0.047         | 0.000             | 0.197           | 0.143              |
| 03#      | 0.048                    | 0.726         | 0.000             | 0.461           | 0.653              |
| 04#      | 0.035                    | 0.046         | 0.004             | 0.110           | 0.142              |
| 05#      | 0.047                    | 0.225         | 0.032             | 0.166           | 0.122              |
| 06#      | 0.092                    | 0.310         | 0.032             | 0.142           | 0.174              |

After solving the improved principal component analysis method, six main influencing factors were finally determined, including reservoir thickness, formation pressure, permeability, sand body distribution, natural resistivity and scour zone natural potential. The characteristic values and contribution rates of each factor are shown in the following table:
Table 3. The results of factor screening

| The main factors | Coalbed thickness | Formation pressure | Permeability | SAR  | RXO  | SP  |
|------------------|-------------------|---------------------|--------------|------|------|-----|
| Eigenvalue       | 7.7324            | 3.496166            | 3.204186     | 2.531| 1.892| 1.6395|
| contribution     | 0.2974            | 0.2758              | 0.1642       | 0.089| 0.072| 0.0631|
| Cumulative       | 0.2974            | 0.5732              | 0.7374       | 0.826| 0.899| 0.9623|

3.2 Fitting and Prediction

Linear regression, SVM regression and gaussian regression were used to fit and predict the data by MATLAB. The results are as followed:

The fitting results of linear regression and SVM models are shown in the following table:

Table 4. fitting coefficients and relative errors

| Coefficient | X1    | X2    | X3    | X4    | X5    | X6    | Rela’err |
|-------------|-------|-------|-------|-------|-------|-------|----------|
| Linear      | 0.0023| 0.098 | -0.0035| 0.041 | 0.004 | -0.004| 0.28823  |
In the above table, Rela 'error means the relative error of the fitting results of different models. As can be seen from the above table, the fitting and prediction error of the SVM model is smaller than that of the linear regression model. Moreover, gauss regression provides a dividing line. It can be seen that the prediction accuracy of SVM is higher for the points below the spline, while the prediction error of the points above the spline is larger. All data points are classified and analyzed according to the gaussian regression curve, and the results are shown in the following table:

**Table 5. Accuracy of data points**

| Precision (%) | All data | Above the boundary | Below the boundary |
|---------------|----------|--------------------|-------------------|
| Linear        | 71.18    | 78.74              | 64.43             |
| SVM           | 79.2     | 58.39              | 86.67             |

4 Conclusion

1. PCA can reduce high-dimensional variables to low-dimensional variables, while SVM regression modeling maps low-dimensional nonlinear inputs to high-dimensional linear outputs.
2. The support vector machine (SVM) method is based on the influence factors of the evaluation object. This method gets rid of the complicated evaluation process to some extent and is simpler and more feasible than the traditional evaluation method.
3. The fitting results of different kernel functions are different. For more accurate prediction, we can usually choose a variety of kernel functions to construct the model.
4. The Gaussian function is suitable for classifying discrete nonlinear data points. By fitting the classified data points with different models, more accurate results can be obtained.
5. The accuracy summary shows that the SVM model is suitable for predicting daily average oil production of less than 2000 Wells. In other words, using the SVM model to predict data with results less than 2000 has high reliability.

Acknowledgments

This work was funded by the National Science and Technology Major Project of China (No. 2016ZX05042-003 & No. 2016ZX05065-003)

Reference

[1] Lai Fengpeng, Li Zhiping, Cen fang, et al. Prediction of oil well production using moving average method [J]. Oil well test, 2007, 16(1):15-16.
[2] Xi’an Zhixian information technology co., LTD., sancai journal editing system, manuscript management platform. Study on oil well productivity prediction using node analysis method [J]. Logging technology, 2006, 30(4):350-353.
[3] Miao Heping, Wang Hongxun. Production prediction and fracture number optimization after horizontal well pressure [J]. Oil drilling and production technology, 1992, 14(6):051-56.
[4] Li Honglian, Wang Chunhua, Yuan Baozong, et al. Learning strategies for support vector machines for large-scale training sets [J]. Acta comput arica sinica (05):140-144.
[5] Yu Liping, Ppan Yuntao, Wu Yishan. Study on standardization method of comprehensive evaluation data of academic journals [J]. Books and information, 2009, 53(12):146-149.
[6] Li y, Zeng z x, Zhang m, et al. Application of principal component analysis in multi-index comprehensive evaluation [J]. Journal of Hebei university of technology, 1999, 28(1):94-97. (in Chinese with English abstract)
[7] Lin Haiming, du Zifang. Problems that should be paid attention to in the comprehensive evaluation of principal component analysis [J]. Journal of statistical research, 2013, 30(8):25-31.
[8] Ding Shifei, Qi Bingjuan, Tan Hongyan. Research review of support vector machine theory and algorithm [J]. Journal of university of electronic science and technology (1):4-12.

[9] Qi Hengnian. Research review on support vector machines and their applications [J]. Computer engineering, 2004, 30(10):6-9.

[10] HEPeng, Qiu jianlin, et al. Analysis and research on several network topology search methods [J]. Computer technology and development, 2005, 15(7):17-19.