Research on Pump Inspection Cycle Early Warning Method Based on Big Data

DU Lihong[1*], LIU Yufang[1], CAO Fei[2], LI Fang[1], MIN Guizhi[1], LIU Zhongbo[1], XIA Zhixue [1], and ZHANG Zhanmin[1]

[1] Engineering Technology Research Institute, PetroChina Huabei Oilfield Company, Renqiu 062550, China
[2] Science and Technology Division, PetroChina Huabei Oilfield Company, Renqiu 062550,China

Abstract: At present, the existing indicator diagram can only be used for expost judgment and can not give early warning, and the influencing factors of pump inspection period are nonlinear, multi constrained and multi variable. In this paper, big data machine learning method is used to carry out relevant research. Firstly, around the influencing factors of pump inspection cycle, relevant data are collected and the evaluation index of pump inspection cycle is designed. Then, based on feature engineering technology, the production parameters of oil wells in different pump inspection periods are calculated to form the analysis sample set of pump inspection period. Finally, the early warning model of pump inspection period is established by using machine learning technology. The experimental results show that: the pump inspection cycle early warning model established by stochastic forest algorithm can identify the pump inspection status of single well, and the accuracy rate is about 85%.

1 Introduction

During the operation of the pumping well, the production efficiency of the pumping pump is reduced due to the influence of various factors such as paraffin deposit, sand production and heavy oil production, so it is necessary to solve the problem of down-hole faults of pumping wells by means of workover and pump inspection[1]. There are two ways of pump inspection for workover of pumping wells, one way is to check the pump in a planned way, the other way is to be forced to check the pump, at present, it is mainly the latter method. Even though the indicator diagram diagnosis technology can realize the working condition analysis of the oil well pump, can not give advance warning for some situations of pump inspection such as rod break, pump leak and pipe leak and so on. At the same time, in order to improve the production condition of the pumping well, it is necessary to optimize the working parameters of the pumping well and take maintenance measures, including checking the pump, adjusting the parameters, cleaning and preventing the wax, washing the well, flushing and preventing the sand, cleaning and preventing the scale. Various measures occur alternately in equal or non-equal cycles, which affect oil wells together, resulting in multi-event, non-linear, multi-constraint, multi-variable characteristics of oil well production data. Due to the interaction of various factors, and the complex of the influence degree and relationship, the problem of accurate early warning of pump inspection cycle has not been solved, which directly affects the optimization and implementation of production and operation management. The early warning of pump inspection cycle in pumping wells has become a key problem to be solved urgently in the oil field.

The related research on pump inspection period can be divided into three types. The first type is to extend the pump inspection period from the technological point of view[2-6], such as: taking gas-proof measures for gas-affected wells, adding anti-eccentric wear technology at the bottom of sucker rods; Anti-corrosion coupling of sucker rod with bidirectional protection. The second kind is to optimize the pumping unit system by simulation technology[7-10]. The third type is cut in from single technique or single factor[11-19], such as the effect of vibration load on eccentric wear of Rod and tubing, the effect of water cut and submergence on eccentric wear of Rod and tubing, the effect of swabbing parameter adjustment on pump detection rate of oil well, etc.

The problems and techniques of pump inspection described in the literature mainly focus on the traditional methods of single pump inspection cause, sometimes comprehensive pump inspection strategy and treatment methods, and little research based on big data artificial intelligence method. With the development of data science, machine learning technology provides a new idea and method for intelligent evaluation of petroleum exploration and development. Based on the concept of data-driven, this paper first collects relevant data around the influencing factors of pump inspection cycle, establishes the evaluation index of pump inspection cycle, and carries out the construction of big data of pump inspection cycle. Then, it uses feature engineering technology to calculate the production parameters of oil

* Corresponding author: *cyy_dlh@petrochina.com.cn

wells in different pump inspection cycles, forms the sample set of pump inspection cycle analysis, and uses random forest algorithm to establish the early warning model of pump inspection cycle, so as to achieve the early warning of single well production status.

2 Establishment of evaluation index of pump inspection cycle

The main factors that affect the pump inspection of are fatigue strength, wear of tubing and Rod, corrosion, Wax Scale, leakage, break off, working system of oil well and so on. These factors are ultimately represented in the data. In the era of big data, the dynamic analysis data of these oil and water wells provide the most basic data support for the intelligent analysis and prediction of pump detection cycle. According to the influencing factors of pump inspection cycle and the data status of the project, the evaluation index of pump inspection cycle is designed. It mainly includes production data index and work diagram data index.

Tab.1 Evaluation index of pump inspection period

Index classification	index item	data sources
Production data index	production time、pump diameter、pump depth、stroke、jig frequency、total daily stroke、total times per day、pump delivery、pump efficiency、daily liquid production、daily oil production、daily water production、daily gas production、water-bearing、sand-bearing、gas oil ratio、oil pressure、casing pressure、back pressure、hydrostatic level、dynamic liquid level、up current、downward current、dosing times、well washing times	Well production data
Work diagram data index.	maximum load、minimum load、stroke、jig frequency、C effective、transfer load、balance degree、upstroke、maximum current、maximum down stroke current	well diagram data

Because the production Data and work diagram data belong to time series, and the amount of data is large, it will increase the complexity of the model by directly using time series for machine learning modelling. In this paper, the characteristic parameters such as mean, median, range, variance, standard deviation, coefficient of variation, crest, trough, skewness and Kurtosis are extracted.

Skewness reflects the asymmetry of production index-time curve in pump inspection period. The formula is as follows: where the μ is the mean value, σ is the standard deviation, in the actual calculation, we replace μ with its sample value.

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E\left[(X-\mu)^3\right]}{\left(E\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}} \quad (1)$$

Kurtosis reflects the sharpness of production index-time curve and is a statistic to describe the degree of steepness and slowness of all value distribution. Kurtosis is defined as the fourth standard moment, which is formulated as follows:

$$Kurt[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{E\left[(X-\mu)^4\right]}{\left(E\left[(X-\mu)^2\right]\right)^2} \quad (2)$$

3 Sample construction of pump inspection cycle

If we extract the peak, tough, trend, cycle and other characteristics for each time series index, then a sample of pump inspection cycle will have hundreds of features, high-dimensional data often contain observations of irrelevant or redundant features. Therefore, before modeling, it is necessary to detect the correlation of the feature parameters and delete the redundant feature parameters. Correlation analysis is an important index to quantify the consistency degree of variation among different factors. The degree of correlation between sample factors is quantified using the correlation coefficient, which is a numerical value between [-1,1] and the larger the absolute value of the Corr, the higher the correlation between the different factors was, a negative number indicates a negative correlation (the value of the factor changed in the opposite direction) and a positive number indicates a positive correlation (the value of the factor changed in the same direction) . In this paper, the Pearson Correlation Coefficient is used to calculate the correlation between the parameters. Assuming that there are two variables X and Y, the Pearson Correlation Coefficient between the two variables can be calculated by the following formula:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E\left(Y^2-E^2(Y)\right)}} \quad (3)$$

Where E is the mathematical expectation, Cov is the covariance, and N is the number of variables. In this paper, we delete the parameter with correlation greater than 0.8.

The characteristic parameters of pump inspection cycle data in the whole plant are calculated, and 776 samples are finally generated through vacancy value deletion, correlation analysis, redundant attribute deletion and other methods, as shown in the table below.

Tab.2 Machine learning sample set

production rate	effective production rate	well washing times	dosing times	etc.	warning level
0.943	0.986	1	0	…	Level 4
0.11	0.972	1	3	…	Level 2

0.829	0.993	2	1	…	Level 2
0.945	0.978	2	0	…	Level 4
0.945	0.98	1	1	…	Level 4
0.946	0.981	2	12	…	Level 4
0.811	0.856	0	0	…	Level 4
0.689	0.991	0	0	…	Level 2
0.767	0.779	0	0	…	Level 4
…	…	…	…	…	…

4 Early warning model of pump inspection cycle

Machine Learning Algorithms play an important role in big data analysis, including classification and prediction, clustering, association analysis and hundreds of other algorithms, such as random forest, support vector machine, neural network, etc. Due to the different data distribution characteristics, the applicable algorithms are different among different data sets. How to choose a suitable machine learning algorithm to obtain the best prediction accuracy is a cycle of processing, including repeated selection of models, parameters and analysis of effectiveness.

4.1 Parameter selection

There are many algorithms for parameter selection, the most common of which are the algorithms of random forest feature importance assessment, including MDI (Mean Decrease Impurity Important) and MDA(Mean Decrease Accuracy Important).

(1)Mean Decrease Impurity Important (MDI)

Mean Decrease Impurity Important (MDI): indicates the average reduction of the error per feature. In each split of each tree, the improvement of the split criterion is an important measure of the split variables, which are accumulated separately for each of the trees in the forest. The decision tree divides the node into two sub-nodes according to some rules. Each split is aimed at a feature that minimizes the error. The error can be calculated using mean square error, Keeny purity, information gain, or some other metric set as needed. In sk-learn, the enhancement effect of each split is weighted by the number of samples arriving at the node, and the significance of the feature is normalized.

(2) Mean Decrease Accuracy Important (MDA)

Mean Decrease Accuracy Important (MDA) disrupts the order of the eigenvalues of each feature and measures the effect of changes in the order on the accuracy of the model. This ingenious method uses out-of-pocket data to calculate importance. The OOB data is part of the training set, but is not used to train this particular tree. Calculate the fundamental error using OOB data, then randomly scramble the order for each feature. For the non-important features, the effect of the disorder order on the accuracy of the model is not too great, but for the important features, the disorder order will reduce the accuracy of the model.

After repeated calculations of the importance of the parameters, 35 feature parameters were selected, they are production time _ total cumulative volatility, the daily oil production _total accumulative volatility, total stroke, the back pressure _ total accumulative volatility, daily oil production _ the mean value of cumulative volatility, daily liquid production _ total accumulative volatility, production time _ the mean value of cumulative volatility, back pressure_ total accumulative volatility ,daily total jig frequency _ variance, dynamic liquid level _kurtosis, downward current _variance, water cut _ total accumulative volatility, production time _ kurtosis, water cut _ total accumulative volatility, daily liquid production _ coefficient of variation, deflection of transfer load, C total effective accumulative volatility, pump depth _ skewness, production time _ skewness, water cut _variance, water cut _ skewness, daily oil production _ skewness, jig frequency _ the minimum, daily fluid production _variance, back pressure _ interquartile range, daily oil production _ mode, downward current _ skewness, gas-oil ratio _ skewness, pump efficiency _ skewness, maximum load mode, dynamic fluid level _ the minimum value.

4.2 Model building and evaluation

4.2.1 Data set partitioning

Machine learning sample set is generally devided into traning set and test set, The former is used to train the parameters of the model and the latter is used to test the generalization ability of the model. In the process of data mining, because of various reasons, we can not collect all the sample data, so we need to sample. There are many methods of sampling, such as random sampling, stratified sampling, systematic sampling, cluster sampling and so on. In the course of this project, stratified sampling is used to divide the population into several homogeneous layers, and then random sampling or mechanical sampling is used in each layer. Stratified sampling is characterized by combining scientific grouping with sampling, grouping reduces the effect of the variability of each sampling layer and ensures that the samples are representative enough. Stratified sampling can be divided into general stratified sampling and stratified proportional sampling according to different sampling methods in homogeneous layers. The general stratified sampling is to determine the sample size of each layer according to the sample variability, more samples with large variability, less samples with small variability. Stratified proportional sampling is usually used when the variability of the sample is not known in advance.

Using the method of stratified sampling, 776 sample sets are divided into training sets and test sets according to the ratio of 8:2, of which 620 are training sets and 156 are test sets.

Tab.3 Sample set division

level	Sample number	Training sets	Test sets
Level 1	97	79	18
Level 2	155	128	27

Level 3	204	157	47
Level 4	320	256	64
total	776	620	156

4.2.2 Random forest modeling

Random forest is a classifier ensemble learning algorithm, which has good effect for regression and classification problems. The random forest algorithm is based on the decision Tree Algorithm, which divides the sample space recursively to form a series of recursive IF-THEN rules and organize them in the tree structure according to the data features in the training set. Using the idea of stochastic simulation, N random decision trees are constructed to form "forest", and the final prediction is made by synthesizing the results of the decision trees in "forest". The idea for the random forest algorithm is illustrated in the following figure.

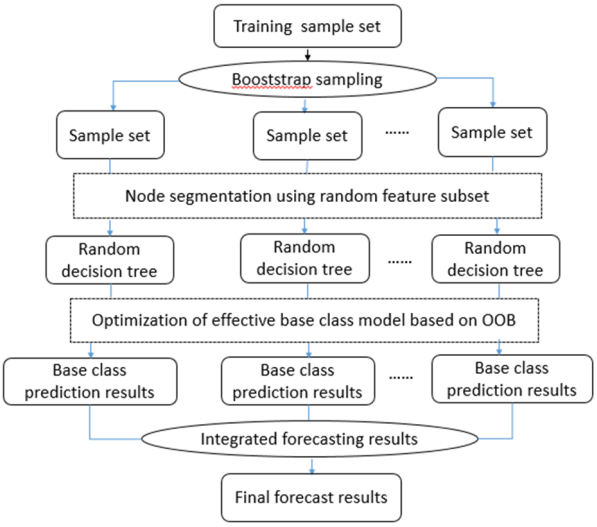
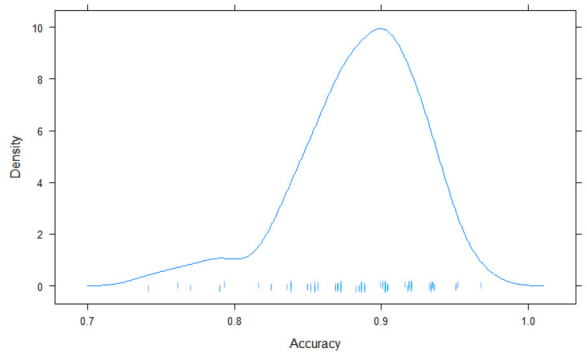


Fig.1 Flow chart of random forest algorithm

The 35 parameters above were used as input characteristics, and the random forest tree algorithm is used to establish the model. Through repeated iterations, the average accuracy of the model is about 0.85, as shown in the figure below.



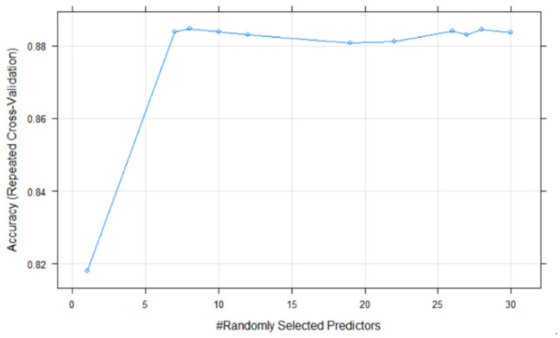


Fig.2 Density distribution of random forest model

Confusion Matrix, also known as error Matrix, is a standard format for precision evaluation. It is expressed in the form of N-row and N-column Matrix, which is mainly used to compare the classification results with the actual measured values. Each column of the confusion Matrix represents the prediction category, the total number of columns represents the number of data predicted for that category, and each row represents the true attribution category of the data, the total number of data per row represents the number of data instances for that category. The confusion Matrix of the model is shown in the following table:

Tab.4 Confusion matrix of random forest model

		True level				accuracy
		Level 1	Level 2	Level 3	Level 4	
Forecast level	Level 1	13	4	0	0	76.4%
	Level 2	5	21	3	1	70%
	Level 3	0	2	43	7	82.6%
	Level 4	0	0	1	56	98.2%
Recall ratio		72.2%	77.8%	91%	87.5%	85.2%

5 Conclusion

This paper presents the application of big data technology in the early warning of pump inspection cycle analysis, including the establishment of evaluation index, sample building technology, machine learning modeling technology, etc. Because oil data is a complex system, which spans a large range of time and space, data quality, data integrity, data noise and data set imbalance, etc. , the construction of machine learning data sets is facing a huge challenge. Because of the limitation of the sample, the sample may not be able to cover all the pump inspection, there will be a phenomenon that the model accuracy is high, but the application accuracy is low. This phenomenon will be gradually improved with the application of the model.

Reference

1. CUI Zhenhua. The Rod Pumping System[M]. Beijing: Petrolrum Technology Press, 1994:2-5.
2. WANG Junqi, CAO Qiang, LI Gang, et al. A New Technology of Extended Pump Detection Period and Its Application[J]. Drilling & Production technology,

2005, 28(02): 52-54, 114.

3. SUN Yongquan. Eccentric Wear of Rod and Tube in Pumping Well and Its Prevention[J]. China Petroleum and Chemical Standard and Quality, 2011, 26(05): 144-146.
4. FANG Yaping, ZHANG Shengyong, ZHANG Zhenlong. Ananlysis of the Factors Affecting the Pump Inspection Period of Pumping Wells and Suggestions for Measures[J]. China Petroleum and Chemical Standard and Quality, 2012, 23(07):168-169.
5. HAN Liang, FENG Aixia, ZHANG Lingshan, et al. Analysis on Causes of Oil Leakage in Block F and Its Prevention Measures[J]. Natural Gas and Oil, 2015(01): 62-64.
6. SUN Zhi, WANG Yan, LI Desheng, et al. Research and Application of Rod and Tube Eccentric Wear Control Technology in Polymer Bearing Pumping Wells[J]. Oil-gasfield Surface Engineering, 2004, 23(09): 1-4.
7. DONG Shiming. Computer Simulation and System Optimization of Dynamic Parameters of Pumping Wells[M]. Beijing: Petroleum Industry Press, 2003: 1-8.
8. ZHOU Chunping. Post Buckling Analysis of Sucker Rod String in Vertical Well[D]. Harbin: Master's Thesis of Solid Mechanics in Harbin Engineering University, 2006: 1-7.
9. WU Xiaodong, WU Jing. The Mathematical Model for Calculating Normal Force in Polymer Solution[J]. Petroleum Drilling Techniques, 2003, 31(06): 5-6.
10. [10] DONG Shimin. Mechanical analysis on causes of worn rod string and tubing of rod pumping wells in the water-flooding oilfield[J]. Acta Petrolei Sinica, 2003, 24(04): 108-112.
11. ZHANG Zhanmin. Reliability Forecast Model of Oil Pumping Wells Checking Period and Actual Application[J]. Value Engineering, 2015, 34(08): 31-32.
12. ZHOU Hao. Shandong Industrial Technology, Prediction Model of Pump Inspection Period Based on BP Neural Network[J]. 2019, 000(018): 76-76.
13. MENG Lingkai. Research on The Prediction Method of Eccentic Wear Life Between Rod And Tube for Flooding Pumping Well[D]. Yanshan University,2016.
14. ZHANG Hong. Reliability prediction method of pump cycle and rod life in pumping well inspection[D]. Yanshan University,2012.
15. GAO Jiangang, SHAO Changxin, XU Jiajun. Investigation into optimal time for pump check of rod pumped wells while pumping[J]. China Petroleum Machinery,2003,31(08): 44-46
16. HAN Xiuting, WANG Xiuling, HOU Yu, et al. The effect of rod vibration load on side attrition of pumping rods[J]. Petroleum Geology & Oilfield Development in Daqing, 2004,23(01):38-41,76.
17. GUO Xiaozhong, LIU Hongju, CUI Yagui, et al. Effect of Water Cut and Submergence Depth on Eccentric Wear of Rods and Tubing[J]. Petroleum Geology & Oilfield Development in Daqing, 2006,25(04):82-84,124.
18. ZHANG Huimin. Application of Three-parameter Weibull Distribution in the Analysis of Reliability Engineering[J]. Mechanical Management and Development, 2009,24(3):59-60
19. KANG Rui, LI Ruiying, WANG Naichao, et al. Reliability and Maintainability Engineering[M]. Beijing: Tsinghua University Press: 53-54.