IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 07, Issue 07 (July. 2017), ||V1|| PP 01-04

Identification of Tight Glutenite Reservoir Lithology Based On Machine Learning

DanCheng[1],Yaqi Zuo[2], Mingrui Chai[1], Qinghan Long[3]， Huixing Wang[4]
1 College of Earth Science, Yangtze University , Wuhan, Hubei, China ;
2 Eleventh oil production plant, Changqing Oilfield Company, Shaanxi, Xi'an, China;
3 School of mathematics, Hebei University of Engineering,Hebei, Handan, China;
4 School of economics, Tianjin University of Finance Economics, Tianjin,China

Abstract:In order to solve the problems of low accuracy and poor generalization ability, a new method to identify lithology by machine learning is proposed.InMahu depression X723 wells Baikouquan formation as an example,first of all, the logging data standardization, eliminate the differences between classes of data; secondly, feature selection, selection of sensitive to lithology logging data; finally, using random forest, BP neural network, C5.0, SVM models. The results show that the machine learning methods on the accuracy of area lithology identification with high rate of up to 87.5%. The method is novel and effective, and can be popularized.

Keywords: identify lithology; Mahu depression; machine learning methods

About First Author: Cheng dan(1991-) , female ,Graduated From Yangtze University College of Technology & Engineering , studying for master,study about sedimentary.

Preface

In recent years, with the need of oil and gas resources strategy and the development of exploration and development technology, tight sandstone reservoir has been paid more and more attention[1].It is a basic and important work to identify lithology with logging data[2]. Because of the complexity of rock types of tight sand conglomerate reservoir, the conventional lithologic identification method is no longer applicable.Machine learning is an effective data mining technology, which has been widely used in many fields in recent years[3].The study area of Mahu sag X723 wells Baikouquan formation belongs to the typical dense sand conglomerate reservoir, in order to accurately identify the reservoir lithology, this paper puts forward a method of machine learning to identify lithology .

I. PRINCIPLE AND REALIZATION

The basic theory of machine learning method is based on a large amount of data, the learning behavior through computer simulation, found to contain the law to predict unknown data, the research is focused on learning and reasoning for specific types of data[4].In the dense sand conglomerate reservoir, there are complex nonlinear relationship between the core data and conventional logging data, and machine learning method has strong nonlinear mapping ability, can according to the known data of learning to identify unknown lithology[5-6].The specific process is as follows:①The well logging data standardization, eliminate the differences between classes of data;②Exception handling;③Using the core and logging data to carry out the feature selection,and select the log data which is sensitive to lithology;④Using a variety of algorithms to establish the prediction model, selectand error analysis, the most effective algorithm.

II. APPLICATION EXAMPLE

2.1 Standardized treatment

The conventional logging data with different dimension, in order to eliminate the impact of data between different classes of analysis results[7], the need for processing original data.There are many methods of data standardization. In view of the fact that the Z-Score standardization is applicable to the situation where the data is the biggest and the minimum is unknown, the application of Z-Score is the most widely used, therefore, the author uses Z-Score to deal with the logging data, as form (1),

$$Z_x=（x-\sigma）/\mu \qquad （1）$$

In the formula, Z_x is the standard data logging calculation; x is the original logging data; Mu is the mean value of each logging data; the sigma is the standard deviation of log data.

2.2 Exception handling

Due to the abnormal value of the analysis results of the model have an important impact, therefore, to explore and eliminate outliers may exist in the data, is the premise and guarantee of establishing the data model to reflect the original appearance of object and the real law[8].The method of outlier diagnosis based on clustering in multi dimension space.The principle is that by clustering and calculating the distance between the sample points and the data group, as well as the distance from the distance to determine the diagnosis of outliers and the cause of abnormal points.

Outlier analysis includes three stages:first, clustering, according to the "closeness" of the samples clustered into several classes; second, computing, which is based on the first stage of clustering on the basis of distance, to calculate the abnormal indexes of all sample points; third, diagnosis, which is based on the second stage of abnormal indexes, determine the abnormal points, and analyzes the cause of abnormal samples, the outlier analysis showed abnormal variables in which direction.

2.3 Preferred variable

Optimization of variables is the key to improve the accuracy of model prediction.No matter how superior the algorithm is, the introduction of the variables that are not correlated well with the predicted targets will have a huge impact on the prediction results due to the amplification effect of the error superposition.Therefore, it is very important to select lithology sensitive log data before building the model.This paper uses the method of feature selection to find the log data which is sensitive to lithology.

The importance of variables can be investigated from two aspects: first, from the variables themselves, the important variables should be more information, that is, the variable value of the larger variables;Second, from the point of view of the correlation between input variables and output variables.There is a difference between the variables of different measurement types.

The input variables for the numerical type, the output variables for the classification type, the use of variance analysis.The input variables are observed variables, and the output variables are control variables.We should analyse if there are significant differences in the mean value of input variables when the output variables are at different levelsIf there is a significant difference in the mean value of the input variables at different levels of the output variables, the correlation between input variables and output variables is stronger.Otherwise the opposite.

The use of SPSS feature selection node to identify lithology sensitive logging data, results show (Figure 1), these logging data such as RI, RT, PORE, RXO, CNL, VCL, AC have the strongestimportance and most sensitive to lithology, so choose these logging data as predictive variables.

Figure 1 the result of feature selection of logging data

2.4 Modeling and error analysis

There are many machine learning algorithms, including time series analysis, multiple regression analysis, decision trees, artificial neural networks, support vector machines, random forests, etc.However, the characteristics of the data vary widely, even if the best algorithm can not be applied to all the prediction

problems, so it is important to compare and optimize a variety of algorithms[9].

A nonlinear relationship between the complex lithology and its related logging data, time series analysis and multiple regression analysis of the linear model cannot well reflect the rules among the data based on the change of nonlinear, low prediction accuracy, limited application, there is no longer applicable[10].

The decision tree is based on the greedy algorithm, the recursive partition set up, including the tree generation and pruning in two stages[11]. In the process of generation, The segmentation method is the key to attribute selection.

According to the different segmentation methods, decision tree can be divided into two categories: first, information theoretic approach, the more representative is the C5.0 algorithm;Second, based on the minimum GINI index method, commonly used CART algorithm.C5.0 has obvious effect in solving classification problems, and has high efficiency, ease of use and strong robustness[7].

Artificial neural network (ANNs) is a mathematical model which can simulate the behavior characteristics of animal neural network andprocess information,strong nonlinear mapping ability.at present, the most widely used neural network model is BP neural network[12].

Support vector machine is the structural risk minimization principle of statistical learning theory based on the selection of optimal hyperplane as the discriminant function, and introduce the kernel function, the linear separable problem can be cast into a high dimensional space into a linear separable problem for data analysis[13].It has obvious advantages in solving small sample, nonlinear and high dimensional recognition problems.

Random forest is an integrated algorithm, which is based on the decision tree as the basic classifier, and combines multiple decision trees to improve the prediction accuracy[14].Compared with other algorithms, the random forest is not sensitive to multicollinearity, the prediction accuracy is high, and it will not be over fitting[15].

In view of the obvious advantages of C5.0, BP neural network, SVM, random forest in solving nonlinear problems, the four methods were used to establish the model and the results were compared.Using SPSS Modeler to establish prediction model (Figure2),the specific operation is as follows: 1.enter clean data; 2. the role of logging data is set as input, and the role of core data is set as the target; 3 .partition, 80%ofthe data set to the training set, 20% of the data set to the test set;4. insert C5.0, BP neural network, SVM and random forest algorithm node.

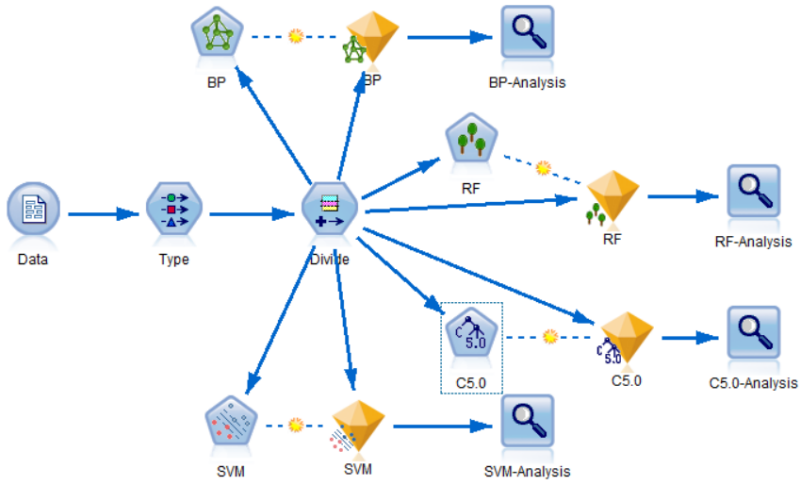

Fig. 2 lithology identification and prediction model

The results show that the prediction accuracy of SVM is the highest, and the prediction accuracy is up to 87.5%.The reason is that the amount of sample data is less and SVM is the best classificationin the case of small sample.

Table 1 prediction accuracy of each algorithm

algorithm	Prediction accuracy
BP	62.50%
RF	50%
C5.0	62.50%
SVM	87.50%

III. CONCLUSION

In this paper, the method of identifying lithology by machine learning is presented, which shows good effect and improves the accuracy of lithology identification.

The effect is better than other methods, can be extended.Comparing with other machine learning methods to predict lithology, support vector machine (SVM) has higher prediction accuracy.

Using support vector machine (SVM) algorithm in logging lithology identification can achieve better results, which can be extended to oil and gas, water, productivity prediction and so on, and has a broad application prospect.

REFERENCE

[1]Larose D T. Data Mining Methods and Models. New York: John Wiley & Sons, Inc., 2005.
[2]Mohaghegh S D. A new methodology for the identification of best practices in the oil and gas industry, using intelligent systems. Journal of Petroleum Science and Engineering, 2005,49(3～4):239～260.
[3]Guangren Shi, Xingxi Zhou, Guangya Zhang, et al. The use of artificial neural network analysis and multiple regression for trap quality evaluation: A case study of the Northern Kuqa Depression of Tarim Basin in western China. Marine and Petroleum Geology, 2004,21(3):411～420.
[4]Tan P , Steinbach M and Kumar V.Introduction to Data Mining .A ddison Wesley , 2005
[5]Li Hao, Liu Shuanglian. Log geological meanings of [M]. Sinopec press, 2015.
[6]Li Hongqi, Meng Zhaoxu, Tan Rahman s, et al. Data mining method in petroleum exploration and Development Research on the application of [J]. oil geophysical prospecting, 2010, 45 (1): 85-91.
[7]Witten I H and Frank E.DataMining:Practical Machine Learning Tools and Techniques .Second E-dition, M o rgan K aufmann , 2005.
[8] Zhou Zhihua. Machine learning [M]. Tsinghua University press, 2016.
[9]SHI Guangren. Prospect of the Application of Data Mining in Petroleum Exploration Databases[J]. China Petroleum Exploration, 2009(0 1):60-64.
[10]LI Hongqi, GUO Haifeng, GUO Haimin, et al. An approach of data mining for evaluation of complex formation using well logs[J]. ACT A PETROLEI SINICA, 2009, 30(4): 542-549.
[11] Mitchell Tomm., CengHuajun, Zhang Yinkui. Machine learning [M]. Mechanical Industry Press, 2003.
[12]Luck M. Elements of Machine Learning, Pat Langley.[J]. Journal of Logic, Language and Information,1998,7(1):103-105.
[13]Dietterich T G. Machine-Learning Research; Four Current Directions[M]. 1997.
[14]Breiman L. Random forests[J]. Machine Learning,2001,45(1):5-32.
[15]Lappalainen H, Miskin J W. Ensemble Learning[M]. Springer London,2000.75-92.