Journal of Petroleum Science and Engineering 156 (2017) 605–615

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: www.elsevier.com/locate/petrol

A comprehensive data mining approach to estimate the rate of penetration: Application of neural network, rule based models and feature ranking

Sajad Eskandarian [a], Peyman Bahrami [b,*], Pezhman Kazemi [c]

[a] Faculty of Mining, Petroleum and Geophysics Engineering, Shahrood University of Technology, Shahrood, Iran
[b] Young Researchers and Elites Club, Science and Research Branch, Islamic Azad University, Tehran, Iran
[c] Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, 43007, Catalonia, Spain

ARTICLE INFO

Keywords:
Rate of penetration
Feature ranking
Random forest
Neural network
Rule extraction
Data mining

ABSTRACT

Rate of Penetration (ROP) estimation is one of the main factors in drilling optimization and minimizing the operation costs. However, ROP depends on many parameters which make its prediction a complex problem. In the presented study, a novel and reliable computational approach for prediction of ROP is proposed. Firstly, *fscaret* package in R environment was implemented to find out the importance and ranking of the inputs parameters. According to the feature ranking technique, weight on bit and mud weight had the highest impact on ROP based on their ranges within this dataset. Also, for developing further models Cubist method was applied to reduce the input vector from 13 to 6 and 4. Then, Random Forest (RF) and Monotone Multi-Layer Perceptron (MON-MLP) models were applied to predict ROP. The goodness of fit for all models were measured by RMSE and R^2 in 10-fold cross validation scheme, and both models showed a reliable accuracy. In order to gain a deeper understanding of the relationships between input parameters and ROP, MON-MLP model with 6 inputs was used to check the effect of weight on bit, mud weight and viscosity. Finally, RF model with 4 variables was used to extract the most important rules from dataset as a transparent model.

1. Introduction

Oil and gas well drilling is one of the most important and expensive part of the oil industry. One of the main factors which has an effect on drilling efficiency and cost is the Rate of Penetration (ROP). It is important to find and optimize the relationship between drilling parameters and ROP in order to reduce cost and increase drilling operation's performance eventually. The ROP can be explained by the progress of a bit in rocks and formations in time units, and feet per minute and/or meters per hour are the commonly used units for that. There are several parameters have an effect on ROP such as bit type, Weight on Bit (WOB), revolution per minute (rpm), mud properties, drilling hydraulic, formation conditions, etc (Walker et al., 1986). Generally, ROP is surveyed in instantaneous and average types. Instantaneous ROP is recorded in restricted time and distance during drilling, like the data that is used in this work, while average ROP is recorded in time periods that drill pipes run in a well.

In the past, several researchers tried to develop models and formulate related parameters to ROP (Bourgoyne and Young, 1974; Warren, 1987; Winters et al., 1987); however, none of them could make an accurate and comprehensive estimation because of the complex and non-linear behavior of parameters with ROP (Ricardo et al., 2007). These difficulties have conducted the recent research toward using intelligent solutions which implement computers as an alternative (Rahimzadeh et al., 2010; Monazami et al., 2012; Edalatkhah et al., 2010).

This study is intended to propose a procedure for predicting the ROP and/or other complex systems in order to reach reliable models. It consists of using feature ranking technique and applying different data mining algorithms to search available design space and build predictive models.

2. Materials and methods

In this work, different computational intelligence methods have been used to model the ROP employing on a dataset containing 226 data points. These data points were collected from 5 different wells of a field located in South-West of Iran using the rotary table, and includes different characteristics of drilling operations, drilling hydraulic, etc. in paying and non-paying zones.

* Corresponding author.
E-mail addresses: s.eskandarian@gmail.com (S. Eskandarian), pymnbahrami@yahoo.com (P. Bahrami), pezhman.kazemi@urv.cat (P. Kazemi).

http://dx.doi.org/10.1016/j.petrol.2017.06.039
Received 17 February 2017; Received in revised form 7 June 2017; Accepted 15 June 2017
Available online 16 June 2017

Table 1
Input variables in feature ranking.

Variable	Range
WOB, klb	6–30
rpm	4–100
Flow Rate, gpm	350–1100
Pump Pressure, psi	1200–3800
Incline	5.1–56
Azimuth	52.41–322
Measured Drilling Depth, m	1020–5186
MW[a], ppg	8.85–17.25
Funnel Viscosity, sec	28–179
Plastic Viscosity, cp	3–97
Yield Point, lb/100sf	4–76
Gel 10s[b], lb/100sf	1–73
Gel 10m[c], lb/100sf	2–143

[a] Mud weight.
[b] Gel strength after 10 s.
[c] Gel strength after 10 min.

2.1. Data mining

Using new technologies in the present era, causes a large volume of data to be created in a rapid pace, thus our success in data interpretation depends on a comprehensive insight within it. Emerging advanced processing systems over the last decade, has allowed us to put away the old time-consuming and weary data analysis methods. Data mining is the term that is used for processes of finding hidden patterns and correlations through data to predict the outcomes, and is comprised of three main scientific fields of statistics, artificial intelligence and machine learning. Statistics and Artificial Intelligence (AI) can be found extensively in the literature (Freedman, 2009; Radermacher, 1991). In machine learning, the computers are used to probe structures within a dataset even if there is not a theory behind the way they look like (Kordon, 2010). In this work, different machine learning approaches have been applied to model the ROP based on various input variables.

2.2. Feature ranking

Dealing with many features in machine learning problems, can cause some difficulties in both the dataset and the derived models. Feature ranking is a technique which finds the importance of features, and can be used to reduce the dimensions of the dataset. Using this technique also has the benefits of shorter training time, ease of interpretation of models, overfitting reduction and lower cost in data collection (James et al., 2013). Depends on the choice of predictive models for feature ranking approach, different results of variable ranking and importance would be achieved (Szlek et al., 2016). Though, it is always a moot point to choose a proper model in feature ranking.

In this work, *fscaret* package as a in R environment is used to reduce the input vector and give stable variable importance (Szlek and Mendyk, 2015; Team, 2015). *fscaret* uses a large number of 103 various models in order to determine the importance of features and rank them. At the end, the importance of variables is averaged through all models to give reliable results, and is scaled from 0 to 100.

In this work, firstly a total number of 21 variables were chosen to estimate the output. Then, 8 constant or nearly constant variables were excluded, and 13 remaining variables were selected for feature ranking. These features are shown in Table 1.

2.3. Model assessment

To evaluate the model's goodness of fit, Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) was used. Additionally, to examine the generalization ability of the models, 10-fold Cross Validation (10CV) approach was performed. The model which has a lower RMSE and higher R^2, can be considered as the best model. RMSE and R^2 are calculated based on Equations (1) and (2).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{prd,i} - y_{act,i}\right)^2}{n}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{prd,i} - y_{act,i}\right)}{\sum_{i=1}^{n}\left(y_{prd,i} - y_m\right)} \quad (2)$$

where y_{act} and y_{prd} are the actual and predicted values; respectively, i is the data record number, y_m is average of the experimental value, and *n* is the total number of records.

It is essential to check the generalization ability of produced models in data mining problems. One of the common methods for model's validation is to split the dataset into training and test sets, and evaluate the performance based on the test set. However, inability of measuring the generalization ability of models is a weak point of this method, and it is possible that accurate models are generated by chance especially in low to medium sized datasets.

One of the techniques to survey how a model can generalize to an independent data set is 10CV (Zhang and Yang, 2015; Bahrami et al., 2016). The data set is split into 10 parts where; at first step, one of the 10 subsets is chosen as the test set, and the other 9 subsets are all considered as the training set. This procedure continues 10 times, and at each step one of the parts is recognized as the test set. Then, the average of RMSE and R^2 is calculated for all 10 trials. 10CV has a lower variance compared to other methods, and can confirm that the model is generalized and not over-fitted. Fig. 1 demonstrates the 10CV procedure.

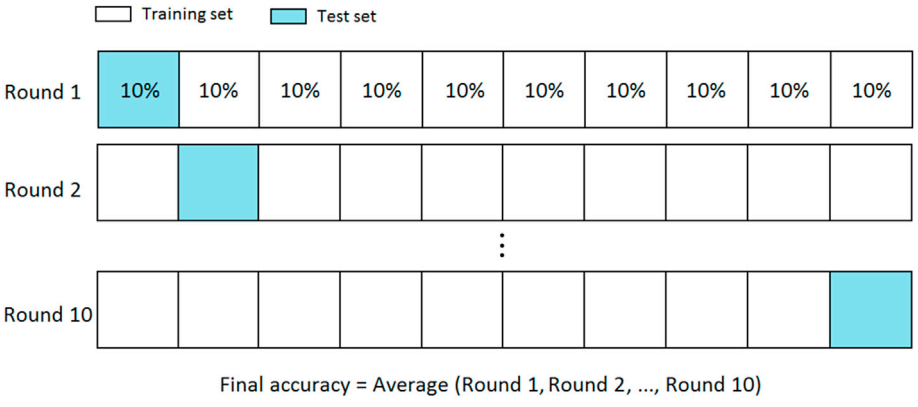

Fig. 1. Schematic of 10CV method.