Intelligent knowledge management for identifying excess water production in oil wells

M. Rabiei & R. Gupta
Department of Mathematics and Statistics,
Curtin University of Technology, Australia

Abstract

In hydrocarbon production, certain amount of water production is inevitable and sometimes even necessary. Problems arise when water rate exceeds the WOR (water/oil ratio) economic level, producing no or little oil with it. A lot of resources are set aside for implementing strategies to effectively manage the production of the excessive water to minimize its environmental and economic impact. Water shutoff technologies are available to effectively manage excess water production; however, their use requires the knowledge of the underlying cause. The conventional diagnostic techniques are only capable of identifying the existence of excess water and cannot pinpoint the exact type and cause of the water production mechanism (WPM). A common industrial practice is to monitor the trend of changes in WOR against time to identify two types of WPMs, namely coning and channelling. However, it has been demonstrated that WOR plots are not general and there are deficiencies in the current usage of these plots. In this paper we present a new technique for diagnosing WPMs. We extracted predictive data points from plots of WOR against the oil recovery factor and collect information on a range of basic reservoir characteristics. This information is processed through tree-based ensemble classifiers. Next we construct a new dataset smeared from the original dataset, and generate a depictive tree for ensemble using a combination of the new and original datasets. To generate the depictive tree we used a new class of tree classifiers called logistic model tree (LMT). Our results show high prediction accuracy rates of at least 93% and easy to implement workflow. Adoption of this methodology would lead to accurate and timely management of water production saving oil and gas companies considerable time and money.

Keywords: water production mechanisms, water/oil ratio, ensemble classifiers.

WIT Transactions on Engineering Sciences, Vol 81, © 2012 WIT Press
www.witpress.com, ISSN 1743-3533 (on-line)
doi:10.2495/PMR120161

1 Introduction

A review of the available literature on the topic of excess water production in oil wells establishes that the industry still lacks a simple, easy to use tool, which takes advantage of all the relevant data and produces accurate and interpretable results [1]. While monitoring the trend of oil and water production data is a commonly used procedure to detect any abnormalities [2], it does not provide a very reliable tool for WPM diagnosis. The diagnosis of WPMs is a very complex task and requires a thorough examination of all the available data. Investigating the nature of the excess water produced into the well involves a multistep process in which, various types of data, which are usually accompanied with uncertainties, are looked in to and analysed. A solution to a better problem diagnosis under uncertainty is to supplement expert knowledge with predictions from mathematical and intelligent computing models.

In this work, we approach the problem of WPM diagnosis as a classification problem and use simulated reservoir models to depict various WPMs. Any ordinary classification problem, involves a learning stage in which a learning dataset, made up of a combination of predictor parameters corresponding to a particular class are fed to a learning algorithm to generate a classification model. The simulated reservoir models are used to build the learning dataset for the classification models. Each WPM case can be described by complex interaction of numerous reservoir parameters leading to different WOR plots, which display the characteristic trends of water and oil production in that WPM. In our innovative approach, we extract a sequence of informative discrete parameters from WOR plots, by recording values of oil recovery factor (RF) corresponding to a range of WOR values. Heuristically, set of such parameters would quantify the trend in the WOR curves and would be effective for discriminating classes of WPMs. We then incorporate the extracted information from these plots together with the knowledge of the reservoir characteristics into a knowledge base for developing tree-based classification models.

A classification tree is a display of the sequence of tests leading to a class label in a classification procedure similar to human decision making process. Prediction accuracy of the tree can be improved by constructing multiple trees [3, 4] on the different subset of data across observations and features. The results of the multiple trees can then be collated to form an ensemble classifier, whose individual predictions are combined in some manner (e.g., voting) to form a final prediction. Interpretability of the ensemble classifier can be improved by procreating new training cases by generating a smeared sample of the original data, and finally generating a single representative tree.

The rest of the paper is organized as follows. First, the generation of the learning dataset is explained. Next, we define the WPMs classification problem and outline the classification models. Finally, we present the results and conclusions.

2 Generation of learning dataset for classification models

Synthetic reservoir models depicting excess water production problems of coning, channelling and gravity segregated flows and associated WOR-RF plots were used to build the learning dataset for the classification models. These models represent coning from bottom water drive, coning from edge water drive, channelling from injection water, channelling from edge water drive and also a complex condition of bottom water drive with baffles in vertical direction (for detailed information on the simulated models please refer to Rabiei [1]). From these base models, various scenarios of wettability with different values of oil viscosity and different degrees of crossflow between layers were simulated to cover a large range of practical situations with excess water production and the associated WOR-RF plots were generated.

On these plots, few points of splits across WOR plot were heuristically identified, so that within each segment the gradient remained constant. For each of the points the sequence of corresponding RF values was recorded. We considered the cut off value point at WOR equal to 40, which represented 97.5% water cut. In view of the small values of RF below WOR=1, only two representative parameters at WOR=0.1 and WOR=0.5 corresponding to water cut values of 9% and 33% respectively were selected from this region. The segment located between WOR=1 and WOR=10, equivalent to water cut values of 50% and 91% respectively, exhibited the most information-rich part of the plot with regards to the RF values. Each RF_{WOR} parameter represents a RF value corresponding to a different level of WOR ranging between 0.1 and 40 (e.g. $RF_{WOR0.1}$ represents the value of RF at WOR equal to 0.1). Figure 1 illustrates the split points and the segments on a sample WOR plot.

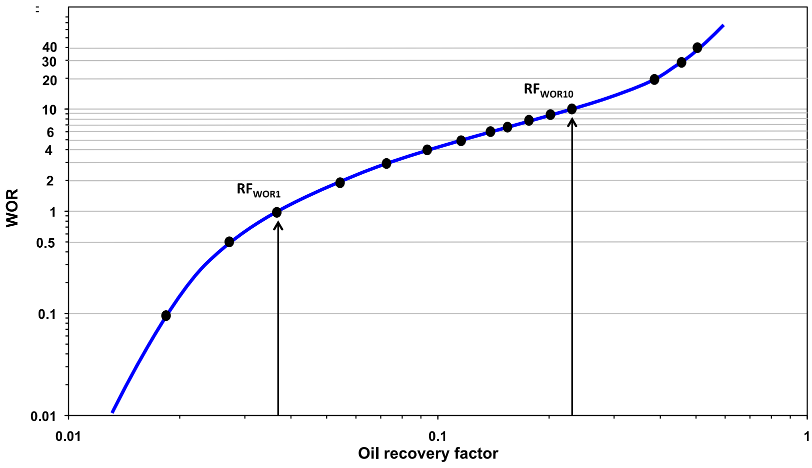

Figure 1: Split points on a sample WOR plot.

In this manner, k cases (k=714) of WPMs were generated where each case was identified by a set of static parameters $(s_1, s_2, ..., s_n, n = 8)$ obtained from simulated reservoir models (Table 1) and dynamic parameters $(D_1, D_2, ..., D_m, m = 15)$ extracted from WOR-RF plots and corresponding WPM type. These cases were stacked in to a matrix (714 × 24) forming the final dataset to be used for classification purpose.

Table 1: Reservoir characteristics selected as input into the classification models.

Parameter	**Abbreviation**
Vertical to horizontal permeability	Kv/Kh
API	API
Wettability	WET
Initial oil flow rate	IOFR
Plateau period for the initial oil flow rate	PP
Drainage area	DA
Aquifer strength (Water/oil volume)	AQWOV
Water injection rate	WIR

The dataset can be represented as $CD = \{(L_{ij}, C_j), i = 1,2,...,8+15, j = 1,2,...,N\}$, where N is the number of cases in the learning dataset, $L_{ij} = (S_{1j}, S_{2j}, ..., S_{nj}, D_{1j}, D_{2j}, ..., D_{mj})$ is the vector of the values of the static and dynamic parameters for the jth case in the learning dataset and C_j is the code for the corresponding WPM (1=*Channelling*, 2=*Coning*, 3=*GravityDominated*, 4=*NoWater*). The cases with WOR values of 0.1 or less were labelled as *NoWater* and were used as control cases. The cases were then randomly sampled to form the learning and validating sets such that both learning and validating datasets had the same proportion of cases from each WPM class. The learning set included two thirds of the cases (N=476) in the dataset used for constructing and training the ensemble models. The remaining cases formed the validating set, used for evaluating and comparing the efficiency of the developed models.

3 Defining the WPMs classification problem

The WPMs classification problem is defined as $I_r(y) = f(s_1, s_2, ..., s_n, d_1, d_2, ..., d_m)$, where $(s_1, s_2, ..., s_n)$ are values of the n static reservoir parameters, $(d_1, d_2, ..., d_m)$ are values of m dynamic parameters extracted from WOR-RF plots and $I_r(y)$ is an indicator parameter taking values of c=$\{1, 2, 3, 4\}$ corresponding to the labels for each classification category of WPMs. The index r=1, 2, …, 14, corresponds to the models at different production stages, corresponding to known WOR values at that point of time.

For this study, we considered two different scenarios of pre and post-water-production and for each scenario, appropriate set of parameters were used accordingly. In the first scenario (r=0), the classifier comprised only the static reservoir parameters. Such a model could be applied before a well starts production to investigate the possible likelihood of a water production problem in the future:

$$\text{Model \#0: } \{I_0(y) = f(s_1, s_2, ..., s_n)\}$$

For the second scenario, both static reservoir parameters and dynamic RFWOR parameters (m=1, 2, …, 15) were employed in order to investigate the interaction between these parameters and the resulted effect on WPM diagnosis. These parameters were sequentially added to generate a separate model for each stage of the water production cycle. This procedure would enable thorough examination of the effect of the extracted dynamic parameters in identifying the WPM. It would also define at which stage of water production cycle, one is more likely to identify the cause of water production more accurately. For this purpose, a separate classification model was implemented for each dynamic parameter, while taking into account the history of WOR trends before that specific production point.

$$\text{Model \#1: } \{I_1(y) = f(s_1, s_2, ..., s_n, d_1, d_2)\}$$
$$\text{Model \#2: } \{I_2(y) = f(s_1, s_2, ..., s_n, d_1, d_2, d_3)\}$$
$$\vdots$$
$$\text{Model \#14: } \{I_{14}(y) = f(s_1, s_2, ..., s_n, d_1, d_2, ..., d_{15})\}$$

The classification models were produced using three popular ensemble classification techniques in data mining, namely, bagging [3], random forest [4] and AdaBoost [5]. In ensemble classification algorithms, the results from several individual classifiers are integrated in some manner (averaging or voting) in an attempt to provide a more accurate prediction. Random forest technique proved to be the best performing algorithm for this study. However, the results of ensemble classifiers are often complex and difficult to analyse. To make the results of these models more appealing and understandable to the end user for predicting and diagnosis of different WPMs in oil fields, we generated a single representing tree called the depictive tree from the ensemble of trees produced by random forest. To develop this depictive tree, firstly, the data smearing technique from Breiman and Shang [6] was used to generate a new dataset consisting of manufactured predictor parameters. Secondly, these manufactured predictor parameters were fed to the selected ensemble classifier (random forest) to predict a WPM type for each set of predictor parameters and form a new case of WPM. This new manufactured dataset was combined with the original dataset in an attempt to retain the latent traits of the original problem and used to generate an easy to comprehend and user-friendly interface using the LMT (logistic model trees) algorithm [7]. LMT combines the linear logistic regression with the classification algorithm to overcome the disadvantages associated with either method.

4 Results and conclusions

The accuracy results and associated kappa value of the developed models are shown in table 2. While these models convey high total accuracy rates, it is also important to investigate their performance on identifying individual problem types as well. The risks and costs associated with wrong diagnosis of a WPM make it more reasonable to choose a model with a lower total accuracy but with acceptable performance in identifying all problem types.

Table 2: Accuracy and kappa value obtained from each model.

Model#	Accuracy	Kappa
0	90%	0.84
1	96%	0.94
2	95%	0.93
3	95%	0.93
4	95%	0.92
5	94%	0.92
6	93%	0.90
7	94%	0.92
8	94%	0.91
9	93%	0.90
10	94%	0.91
11	94%	0.91
12	93%	0.90
13	93%	0.89
14	93%	0.90

Figure 2 shows the associated votes for each case with regards to the model used in a colour-coded bar-plot, where each model in increasing index indicates the progression of well depletion. Each column in the bar plot corresponds to a case from the validating dataset. The rows correspond to the predicted class for each case by each model. Each WPM requires a specific treatment methodology, which usually costs a lot of time and money. Wrong diagnosis or failure to diagnose a problem type can entail costly operations on companies without any success. By assessing the models in sequence, cases that have been consistently misclassified can be identified and further examined to reveal any possible mistakes or abnormalities in those specific WPM cases. It is clear from fig. 2 that number of the misclassified cases in Model#0 is significantly higher than that of Models#(1-14). Especially, when the amount of the produced water is not large, the models perform very well and the number of misclassified cases is limited.

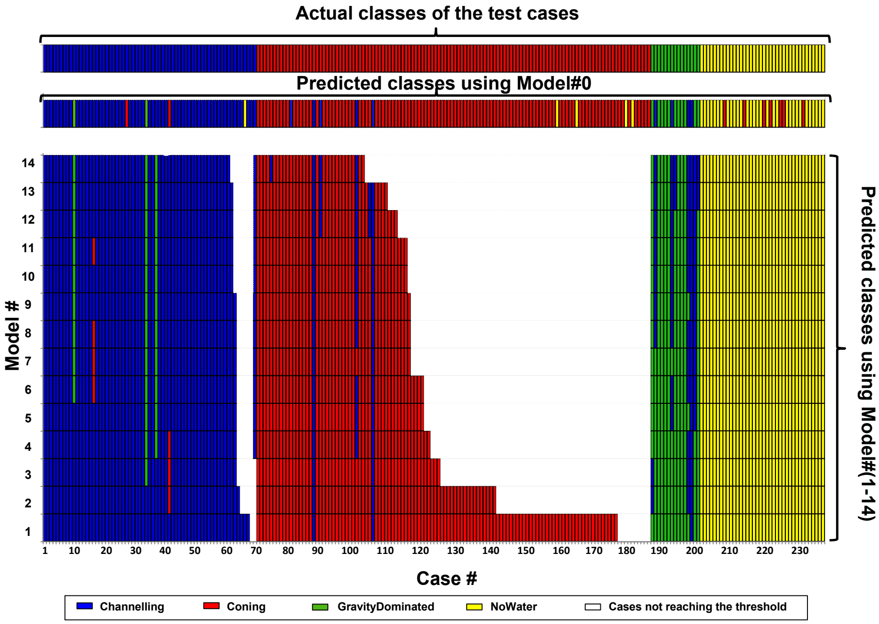

Figure 2: The sequential classification votes allocated to each case using classification models in pre and post-water-production scenarios.

The findings here establish that the applied technique can be successfully used in WPMs diagnostics. We obtained staggering accuracy rates of at least 90% and 93% for the two scenarios, respectively. The models are easy to comprehend for the non-professional end users and are reasonably applicable in situations where water production data are not available.

References

[1] Rabiei, M., *Excess Water Production Diagnosis in Oil Fields Using Ensemble Classifiers,* Ph.D Thesis, Curtin University of Technology, Australia, 2011.

[2] Seright, R. S., Improved Methods for Water Shutoff. *Final Technical Progress Report (U.S. DOE Report DOE/PC/91008-14), U.S. DOE Contract DE-AC22-94PC91008,* BDM-Oklahoma Subcontract G4S60330, 1998.

[3] Breiman, L., Bagging Predictors, *Machine Learning*, **24 (2)**, pp. 123-140, 1996.

[4] Breiman, L., Random Forests, *Machine Learning* **45**, pp. 5-32, 2001.

[5] Freund, Y. and Schapire, R.E., A Decision-Theoretic Generalization Of On-Line Learning And An Application To Boosting, *Journal of Computer and System Sciences,* **55 (1)**, pp. 119-139, 1995.

[6] Breiman, L. and Shang, N., Born again trees, *Technical Report*, Berkeley: Department of Statistics, University of California, 1996.

[7] Landwehr, N., Hall, M. and Frank, E., Logistic Model Trees, *14th European Conference on Machine Learning*, Croatia, 2003.