Feature Selection based on Mutual Information

for Machine learning prediction of Petroleum reservoir properties

Muhammad Aliyu Sulaiman[1], Jane Labadin[2]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,
94300 Kota Samarahan, Sarawak, Malaysia
[1]muhalisu@gmail.com, [2] ljane@pps.unimas.my

***Abstract*— the application of machine learning models such as support vector machine (SVM) and artificial neural networks (ANN) in predicting reservoir properties has been effective in the recent years when compared with the traditional empirical methods. Despite that the machine learning models suffer a lot in the faces of uncertain data which is common characteristics of well log dataset. The reason for uncertainty in well log dataset includes a missing scale, data interpretation and measurement error problems. Feature Selection aimed at selecting feature subset that is relevant to the predicting property. In this paper a feature selection based on mutual information criterion is proposed, the strong point of this method relies on the choice of threshold based on statistically sound criterion for the typical greedy feedforward method of feature selection. Experimental results indicate that the proposed method is capable of improving the performance of the machine learning models in terms of prediction accuracy and reduction in training time.**

Keywords—Machine Learning; Mutual Information; Feature Selection.

I. INTRODUCTION

Well logging is at the heart of the oil and gas exploration, it provides continuous record of rock's formation properties. Reservoir variables are known to be used as input data to a reservoir study. These variables are commonly derived through a number of processes and they are not measured directly from well logging tools. Out of all the reservoir properties, the reservoir porosity and permeability collectively refer to as core logs are of great importance, because accurate prediction of these properties is essential in determining where to drill and if found, how much of oil and gas can be recovered [2]. However, existence of uncertainty in well log dataset affects the optimal performance of machine learning models to predict these properties, in order to address the problem associated with uncertainty in well log dataset as regards to the performance of machine learning models we introduced a feature selection algorithm based on Mutual Information criterion. The choice of mutual information is because of its ability to select features that retain relevant information of the predicting parameter. Moreover, to measure the effectiveness of the proposed study we implemented a machine learning models based on back propagation neural networks. And we used the trained classifiers to test our proposed method in terms of prediction accuracy and training time, by comparing the performance of selected feature subsets with the performance of full feature set.

The rest of the paper is organized as follows: section II is a background of the study and it discusses the overview of well log data and the feature selection methods. Section III, we present Mutual Information hypothesis and formulation of its estimation from the dataset. Section IV presents the proposed feature selection for well log dataset based on greedy feedforward procedure. Experimental studies are detailed in section V, which include experimental setups, results of experiments and discussion.

II. BACKGROUND OF THE STUDY

Literature review provides overview of sources of uncertainty in well log data and how the uncertainty affects optimal performance of machine learning applications to oil and gas predictions. Finally, feature selection algorithms in related studies are reviewed.

A. Overview of uncertainty in well log dataset and how it affeacts accuracy of Machine learning predictors.

Uncertainty information in data is useful information that can be utilized to improve the quality of underlying result. As such, feature with greater uncertainty may not be as important as one which has a lower amount of uncertainty [5].

As earlier on mentioned in the introduction, reservoir variables such as porosity, permeability, water saturation and minerals are known to be used as input data to a reservoir study. And these variables are commonly derived through a number of processes which include acquisition, processing, interpretation and calibration and they are not measured directly from well logging tools. Each of these processes has uncertainty and as such the result of petrophysical data or well logs will equally have uncertainty and limitation [1, 3]. More so it is commonly acknowledged that uncertainty exist at all stages of petroleum exploration [3, 5 & 6]. And not only that, it propagates with each stage since each stage is built on the result from the previous stages.

A hybrid model for predicting Pressure Volume and Temperature (PVT) properties of crude oil is presented in [7], the model was based on the fusion of Type-2 fuzzy logic system (type-2 FLS) and Sensitivity-based linear learning method (SBLLM). The authors categorically recognized the presence of uncertainty in well-log datasets and limitation of SBLLM to generalize when there is uncertainty in dataset. And since Type-2 FLS is known for modeling uncertainty, therefore it is used to improve the prediction ability of