Journal of Natural Gas Science and Engineering 22 (2015) 515–522

Contents lists available at ScienceDirect

Journal of Natural Gas Science and Engineering

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

Investigating the effect of correlation-based feature selection on the performance of support vector machines in reservoir characterization

Kabiru O. Akande [a, *], Taoreed O. Owolabi [b], Sunday O. Olatunji [c]

[a] Electrical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[b] Physics Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[c] Computer Science Department, University of Dammam, Dammam, Saudi Arabia

A R T I C L E I N F O

Article history:
Received 13 October 2014
Received in revised form
12 December 2014
Accepted 5 January 2015
Available online 8 January 2015

Keywords:
Feature selection
Correlation
Support vector machines
Reservoir characterization
Permeability

A B S T R A C T

Permeability is an important property of hydrocarbon reservoir as crude oil lies underneath rock formations with lower permeability and its accurate estimation is paramount to successful oil and gas exploration. In this work, we investigate the effect of feature selection on the generalization performance and predictive capability of support vector machine (SVM) in predicting the permeability of carbonate reservoirs. The feature selection was based on estimating the correlation between the target attribute and each of the available predictors. SVM has been improved through the feature selection approach employed. The uniqueness of this approach is the fact that it employs fewer dataset in improving the performance of the SVM model. The effect of the approach has been investigated using real-industrial datasets obtained during petroleum exploration from five distinct oil wells located in a Middle Eastern oil and gas field. The results from this approach are very promising and suggest a way to improve on the performance of this algorithm and many other computational intelligence methods through systematic selection of the best features thereby reducing the number of features employed.

1. Introduction

Permeability is an indication of the ease with which fluid flows through a material (Darcy, 1856) and in petroleum engineering, it is the ability of porous rock to allow for passage of oil and gas (Sunday Olusanya Olatunji et al., 2014). A porous rock (with pores) is not necessarily permeable rather permeability is an indication of the interconnection of the pore spaces and the ease with which fluid passes through them (Ayan et al., 2001). It is one of the most important flow characterizations of oil and gas reservoir and its knowledge can lead to various important deductions as well as decisions regarding the reservoir under investigation (Tusiani and Shearer, 2007). Its accurate estimation is fundamental to a successful flow characterization of reservoir as well as determining the scale of the medium. Specifically, important information regarding the amount of oil and gas present in a reservoir, what percentage of this oil and gas is recoverable, the fluid saturation distribution (flow rate), estimation of future exploration as well as adequate and proper design of exploration equipment can all be gleaned from permeability estimation. Also, several fundamental issues in petroleum industry can only be resolved with adequate knowledge of permeability estimation (Olatunji et al., 2011). However, it is very difficult to estimate permeability and this has been an area of rigorous research for engineers and practitioners in oil and gas industries. There are many available techniques such as well-log evaluation, core measurement and well testing for permeability estimation (Ahmed et al., 1991). These techniques use well-log and core data obtained from boreholes for permeability estimation. The main tool used in correlating the data is regression analysis which assumes an existence of linear or non-linear relationships between the predictors and target. The predictors in this case are the various properties of rock while the target attribute is the permeability. However, this approach has yielded little success and is far from achieving a good result in accurate prediction of permeability (S.O. Olatunji et al., 2014a,b).

Various techniques used in estimating permeability can be grouped into three major categories. They are empirical, statistical and computational intelligence methods (Olatunji et al., 2011). The empirically determined values are those measured in the laboratory while statistical models make use of multiple regression

* Corresponding author.
E-mail addresses: koakande@gmail.com (K.O. Akande), owolabitaoreedolakunle@gmail.com (T.O. Owolabi), oluolatunji.aadam@gmail.com (S.O. Olatunji).

http://dx.doi.org/10.1016/j.jngse.2015.01.007
1875-5100/

analysis. Computational intelligence methods are various machine learning techniques applied to permeability prediction. We briefly delve into these various methods in order to provide an overview of the development in these fields over the years.

One of the early contributors to this task is Kozeny whose work was extended by Carman (Carrier, 2003). They formulated a formula relating permeability to other measurable rock properties. This relation has its shortcomings in that it is only applicable for uniformly sized spheres and moreover, any physical properties of the rock could only be measured from core analysis with the aid of special equipment (Olatunji et al., 2011). Tixier proposed and established a relationship between permeability and resistivity gradient using empirical model (Tixier, 1949). However, his model only estimates average permeability over the range of area covered by the resistivity gradients. He later developed a simpler model, which is more frequently used, based on the work of Wyllie and Rose (Wyllie and Rose, 2013). Many other models were developed from Kozeny equation as outlined by Pirson (Pirson, 1963).

Multiple regression analysis was employed by Wendt and Sakurai (Wendt et al., 1986) in estimation of permeability though with some drawbacks. Despite boosting the performance of the model in predicting extreme values of permeability using weighted average of high and low values, the model can still be statistically biased, prone to instability and its distribution is narrower than that of the original data set. A lot of advancements has been made in using empirical and statistical models for permeability prediction yet despite all the best efforts, many inaccuracies still persists hence a need for further exploration of computational intelligence which has enjoyed great success in recognizing complex patterns (Olatunji et al., 2011).

Machine learning has revolutionized data analysis for parameter prediction. It enjoys wide ranging successes in diverse engineering problems where it has been applied. Researchers have proffered solutions to many real-life complex tasks with the aid of various machine learning techniques. The techniques have excelled and enjoyed wide acceptance in areas such as medicine (Olatunji and Arif, 2013), manufacturing (Stoneking, 1999), face detection (Osuna et al., 1997), speech-related application (Olatunji et al., 2013, this article is availabe at http://dl.acm.org/citation.cfm?id=1909246), material properties (Taoreed O. Owolabi et al., 2014a,b) and many others. This has spurred revolutionary research into their application in oil and gas industry (Bruce et al., 2000). The field of artificial intelligence is full of techniques used in learning complex patterns between predictors and their targets. A very popular technique that enjoys early success in permeability prediction is the neural network based approach also aptly referred to as virtual measurement technique (Wong et al., 2010). Several works exist that has made use of artificial neural network in predicting permeability of carbonate reservoir from other well log data. A pioneer work in this field is Bruce et al. which comprehensively dealt with the details of permeability estimation from well logs with the aid of Bayesian neural network (Bruce et al., 2000).

Hybrid of fuzzy logic and artificial neural network (ANN) was applied to naturally fractured reservoirs and the authors were able to propose a method for complete description of fractured reservoir (El Ouahed et al., 2005). Permeability of fractured reservoirs is of great importance as they affect the migration of oil and gas. Despite various proposed methods, there is still no established practice that is general and the performance of existing techniques requires improvement (Saemi et al., 2007). Hence, a simple and effective data preprocessing technique has been investigated in this work by systematic selection of the best features from the available pool of features or attributes. The proposed approach is found to have profound effect on improving the performance of SVM as indicated in the achieved results discussed later in this work. Additionally, the performance improvement recorded was achieved using a subset of the available data. This is highly desired as the challenge of Huge Data modeling is presently dominating the computing world. This is due to the fact that industries are being inundated with enormous amount of data and storing all these data and modeling them are very challenging. Therefore, the strategy considered in this work can serve as an effective and efficient preselection method of extracting a sufficiently small subset of available data without compromising performance. In fact, since the strategy employed in this work tends to select the features that are most predictive of a given target attribute, superior performance is recorded in most of the cases considered in addition to reducing the number of features employed.

The choice of SVM is due to its many unique features which include sound mathematical foundation, non-convergence to local minima and accurate generalization and predictive ability when trained on small datasets (Akande et al., 2014). The rest of this paper is organized as follows: Section 2 briefly describes the statistical learning algorithm and correlation-based approach considered in this work. Section 3 gives description of the datasets and details of the experiments. Section 4 discusses the results and their interpretations. Section 5 concludes the paper and suggests recommendation.

2. Computational intelligence method

2.1. Support vector machine

Support vector machine is a tool derived from statistical learning theory for classification and regression tasks (Cortes and Vapnik, 1995). In classification problems, SVM employ the use of optimal separation principle. This principle selects (from among infinite number of linear classifiers) an hyperplane with the maximum margin between linearly separable classes. The optimum separation hyperplane selected is based on minimization of generalization error or on defined upper bound on the error using structural risk minimization. An optimum separation hyperplane is the one with maximum distance from the closest point of the two classes (Cortes and Vapnik, 1995), (Cristianini and Shawe-Taylor, 1999). However, SVM seek an hyperplane with maximum margin as well as minimizes a quantity with direct proportionality to number of misclassification errors for non-separable classes. A tradeoff between number of misclassification errors and maximum margin is chosen for the system using a predefined positive constant. In order to construct linear decision surfaces using SVM, a finite set of variables can be transformed on to a higher dimensional space where linear classification can then be carried out (Cortes and Vapnik, 1995). This technique is applied in extending SVM to regression problems.

The concept of maximum margin in SVM classification is extended to regression problems by defining ε −insensitive loss function which was proposed by Vapnik (Cortes and Vapnik, 1995). This is termed ε −Support Vector Regression (SVR). SVR seeks for all the training data a function having at most epsilon deviation from the actual target vectors. This means that SVR limits the error between its hypothesis and actual values to maximum of epsilon (ensure that it does not exceed this constant) and the function that achieve this has to be as flat as possible. SVR linear decision function can be formulated as:

$$f(\boldsymbol{x}, \propto) = \langle w, x \rangle + b \tag{1}$$

A smaller value for the adjustable parameter yields a better model in general and indicates the flatness of the linear decision function. Thus, the objective function of SVR is the minimization of