Computers & Geosciences 64 (2014) 52–60

Contents lists available at ScienceDirect

Computers & Geosciences

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

Identifying organic-rich Marcellus Shale lithofacies by support vector machine classifier in the Appalachian basin

Guochang Wang[a,*], Timothy R. Carr[b,c], Yiwen Ju[a], Chaofeng Li[d]

[a] College of Earth Science, University of Chinese Academy of Sciences, Beijing 100049, China
[b] Department of Geology and Geography, West Virginia University, Morgantown, WV 26506, USA
[c] National Energy Technology Laboratory, Pittsburgh, PA 15236, USA
[d] School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China

ARTICLE INFO

Article history:
Received 3 October 2013
Received in revised form
3 December 2013
Accepted 7 December 2013
Available online 16 December 2013

Keywords:
Shale lithofacies
Support vector machine
Classification
Marcellus Shale

ABSTRACT

Unconventional shale reservoirs as the result of extremely low matrix permeability, higher potential gas productivity requires not only sufficient gas-in-place, but also a high concentration of brittle minerals (silica and/or carbonate) that is amenable to hydraulic fracturing. Shale lithofacies is primarily defined by mineral composition and organic matter richness, and its representation as a 3-D model has advantages in recognizing productive zones of shale-gas reservoirs, designing horizontal wells and stimulation strategy, and aiding in understanding depositional process of organic-rich shale. A challenging and key step is to effectively recognize shale lithofacies from well conventional logs, where the relationship is very complex and nonlinear. In the recognition of shale lithofacies, the application of support vector machine (SVM), which underlies statistical learning theory and structural risk minimization principle, is superior to the traditional empirical risk minimization principle employed by artificial neural network (ANN). We propose SVM classifier combined with learning algorithms, such as grid searching, genetic algorithm and particle swarm optimization, and various kernel functions the approach to identify Marcellus Shale lithofacies. Compared with ANN classifiers, the experimental results of SVM classifiers showed higher cross-validation accuracy, better stability and less computational time cost. The SVM classifier with radius basis function as kernel worked best as it is trained by particle swarm optimization. The lithofacies predicted using the SVM classifier are used to build a 3-D Marcellus Shale lithofacies model, which assists in identifying higher productive zones, especially with thermal maturity and natural fractures.

1. Introduction

The continuous shale reservoir generally possesses distinct characteristics from conventional sandstone and carbonate reservoirs, especially extremely low matrix permeability and organic matter richness. With regard to modern shale-gas reservoirs (e.g., Barnett Shale, Marcellus Shale, Eagle Ford Shale, etc.), economical gas flow can not be achieved without hydraulic fracturing. The "fracability", largely influenced by mineral composition, has been considered as one of the most important parameters for shale-gas reservoir evaluation. The total organic carbon (TOC) can clearly indicate the gas content due to the 'self-source and self-reservoir' feature of shale reservoirs, even though thermal maturation and the sealing capability also influence gas content. In unconventional shale gas exploration and development the emphasis is often placed solely on organic content, but the productivity of shale reservoirs also is highly dependent on the ability of the rock to respond effectively to hydraulic stimulation (Bowker, 2007; Carr et al., 2013; Wang and Carr, 2012a, 2013). Therefore, instead of using petrographic information and sedimentary structures (Chang et al., 2000, 2002; Hickey and Henk, 2007; Jungmann et al., 2011; Loucks and Ruppel, 2007; Rider, 1996; Saggaf and Nebrija, 2003; Singh, 2008), geologists working in the subsurface tend to define organic shale lithofacies by mineral composition and TOC content (Jonk et al., 2012; Wang and Carr, 2012a, 2013). This approach to define shale lithofacies shows strong relationship with conventional logs and consequently can be predicted in large number of wells with sufficient well logs. Modeling of the 3-D distribution of shale lithofacies on the basis of constraining well data coupled with the information of thermal maturation and natural fracture features can be used to effectively recognize productive zones of shale-gas reservoirs (Wang and Carr, 2013). However, a key challenge to construct reliable shale lithofacies 3-D models is to accurately predict shale lithofacies with available

* Corresponding author. Tel.: +86 10 88256830; fax: +86 10 88256466.
E-mail address: w.guochang@gmail.com (G. Wang).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.12.002

conventional logs and limited core data in order to provide reliable constraining data for geostatistical modeling (Wang and Carr, 2012b).

The fuzzy feature of conventional logs is intrinsic to the prediction of lithofacies due to the wide range of responses for each lithofacies, the different scales between logging data and core data, and the varying borehole environment (Dubois et al., 2007). The fuzzy nature of petrophysical response creates a complex and non-linear relationship between log response and shale lithofacies. At the same time, shale lithofacies recognition is a typical multi-class classification problem (Wang and Carr, 2012b), but most algorithms were developed first for binary classification problems (Dietterich and Bakiri, 1995; Li et al., 2003; Ou and Murphey, 2007). It is not a trivial task to decompose multi-class to binary classification problems, since each method has unique limitations (Allwein et al., 2000; Li et al., 2003). All approaches lead to the difficulties and complexities in predicting shale lithofacies even though several mathematical methods are widely used (Table 1). Discriminant analysis (Berteig et al., 1985; Dubois et al., 2007), fuzzy logic (Cuddy, 2000; Dubois et al., 2007; Olatunji, 2008; Rezaei and Movahed, 2008; Wong et al., 1995), artificial neural network (ANN) (Chang et al., 2000; Dubois et al., 2007; Hu et al., 2005; Liu et al., 1992; Negi et al., 2006; Qi and Carr, 2006) and support vector machine (SVM) (Al-Anazi and Gates, 2010; El-sebakhy et al., 2010) have been discussed in the classification of geosciences. Each method has its strength and weakness (Table 1). We have comprehensively analyzed the function and ability of ANN in predicting Marcellus Shale lithofacies (Wang et al., 2013; Wang and Carr, 2012b). In this paper, another powerful machine learning approach, support vector machine (SVM), is evaluated for shale lithofacies recognition as a significant part of 3-D shale-gas reservoir characterization.

The SVM was developed as the basis of statistical learning theory (Vapnik, 1995), and has become increasingly popular in pattern recognition due to its excellent performance on a wide range of classification problems. Compared to ANN, SVM is firmly grounded in rigorous mathematical analysis, and has been applied successfully to several real work problem, such as text and image classification, handwriting recognition, remote sensing recognition, and geologic facies identification (Cristianini and Shawe-Taylor, 2000). In this paper, we first introduce the important mathematical basis of SVM, and then use it to predict Marcellus Shale lithofacies. The result of SVM approach will be compared with ANN methods for prediction of shale lithofacies.

2. Support vector machine principles

2.1. Binary SVM classifier

Statistical learning theory as proposed by Vapnik (1995) to deal with the problem of finding a predictive function based on small sample of training dataset (e.g., pattern recognition), provides the framework underlying SVM. Similar to supervised learning algorithm of pattern recognition, the objective is to minimize risk

Table 1
The widely used pattern recognition models and their comparison (modified from Kordon (2010) and Wang and Carr (2012a)).

Mathematic model	Strengths	Weaknesses
Discriminant analysis	Simple Strong theory base Powerful for linear classifications	Only good at low complexity No ability to learning Inappropriate for non-linear issues
Fuzzy logic	Captures linguistic ambiguity Computing with words User friendly	Only good at low complexity Difficult to scale up Costly maintenance
Artificial Neural Network	Learns from data Universal approximation Fast development	Black box model Poor extrapolation Maintenance nightmare
Support Vector Machine	Learns from small data records Model complexity control Novelty detection and data condensation	Black box model Difficultly market to a broad class of users due to complex theory Limited infrastructure

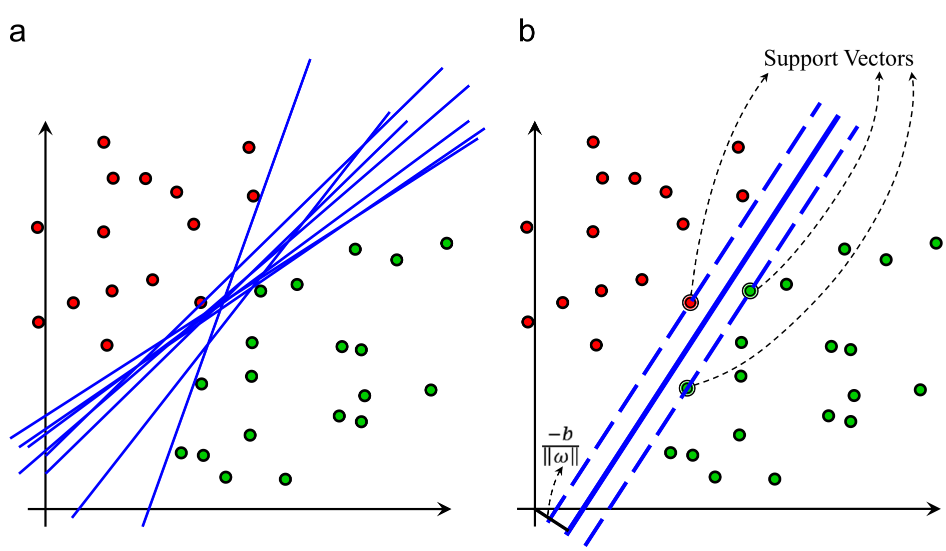

Fig. 1. The large number of linear separating hyperplanes (a) and optimal linear separating hyperplane (the purple solid line) (b) for two classes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)