Facies Classification from well logs using a Scalable SVM based Approach

santosh kumar sahu, sanjai kumar singh, H.K.P Nair, Manoj Ranjan, I S Negi and Rajeev Tandon*
Oil and Natural Gas Corporation Limited
sahu_santosh@ongc.co.in

Keywords

Facies Classification, SVM, Big Data Analytics, Scalable Approach, Bayesian Optimization

Summary

There has been much excitement recently about big data and the dire need for data scientists who possess the ability to extract meaning from it. Geoscientists, meanwhile, have been doing science with voluminous data for years, without needing to brag about how big it is. But now that large, complex data sets should process smartly. As a result, it improves productivity by reducing the computational process. As a result, Big Data Analytics takes a vital role in Oil and Gas Exploration. It provides tools to support structure, unstructured and semi-structured data for analytics. Also, it offers scalable machine learning algorithms for fast processing of data using Machine Learning Approach. It also provides tools to visualize a large amount of data in a practical way that motivates us to implement our model using Scalable Machine Learning Approach. In this work, we describe a scalable machine learning algorithm for facies classification. The algorithm has been designed to work even with a relatively small training set and support to classify a large volume of testing data. A Scalable Support Vector Machine (SVM) approach is implemented and the model is optimized by Bayesian optimization method.

In this experiment, Well Logs used as input data. There are total 13 well's log considered for analysis. The Well data contains 13 attributes out of which 5 attributes are selected based on the feature analysis algorithm. The data is normalized using Min-Max normalization technique, and for SVM Classification it transforms into Sparse Representation for reducing computational time. The model provides a stable classification accuracy with minimum computational time.

Introduction

We live in a digital world where data is increasing rapidly due to advance in technology, mostly the use of sensors (IoT), and ease of Internet Technology. The IoT in embedded devices used in all sectors to accurately gather information, real-time control and operation of the different task without human interventions. As a result, vast amount of data is generated, and it's a big challenge to store, process and analyze the data for decision making. The term "Big Data signify the sheer volume, variety, velocity, and veracity of such data." Big Data (11) is structured, unstructured and semi-structured or heterogeneous. It becomes difficult for computing systems to manage 'Big Data' because of the immense speed and volume at which it is generated. Conventional data management, warehousing, and data analysis system fizzle to analyze the heterogeneous data for not only processing but also storing and retrieving the data. The need to sort, organize, analyze, and systematically offer this critical data leads to the rise of the much-discussed term, Big Data.

As per the characteristics of Big Data it can handle large volume, variety and velocity of data. It motivates us to implement the proposed approach using this technique. Scalable algorithm is able to maintain the same efficiency when the workload grows. As a result, it is not suffering in curse of dimensionality.

In this paper, the scalable machine learning algorithm is used to classify the lithofacies using basic well logs. Facies classification consists in assigning a rock type to a specific sample on the basis of the well logs. It is a crucial task in seismic interpretation because different rocks have different permeability and fluid saturation for a given porosity.

The traditional approaches consist of manually assigning litho-facies by human interpreters and is a very tedious and time-consuming process. Therefore, several alternative methods to the issue of facies classification from well log data have been proposed.

Facies Classification from well logs using a Scalable SVM based Approach

Wolf et. Al. (2) suggested a multivariate statistical approach to identify facies. Busch et al. (3) used analytical procedures to find lithofacies. The first works were based on classical multivariate statistical methods (Wolf et al., 1982; Busch et al., 1987). Later, Wolf et al. (1982); Busch et al. (1987) proposed the use of neural networks for rock classification (Baldwin et al., 1990;

Table 1 provides the details of earlier experiments carried out to determine the lithofacies using statistical and machine learning approaches. As per the literature study, all the methods used in their research is suffering from the curse of dimensionality. In many cases, due to the high volume of data, the predictive models overfits, biased and not able to provide better accuracy. To avoid the curse of dimensionality, a scalable approach is proposed in this experiment that deals high dimension, variety, veracity as well as velocity data. Besides with, the analysis process parallelizes the execution to maintain fault tolerant and the bottleneck of data during implementation. The big advantage is that the program comes to the data for processing. To predict the lithofacies, scalable machine learning approach used to tackle the situations mentioned above. The detail algorithms and data flow described in subsequent sections.

Data Preparation

The data preprocessing is a part of data mining technique that involves transforming raw data into appropriate format as per the learning approach. Several approaches such as data acquisition, data integration, cleaning, feature analysis, feature extraction and data transformation are adopted. In this experiment total twenty well are considered for analysis.

- *Problem Definition*

  $X = \{x_1, x_2, \ldots x_n\}$ set of features and Y is the class label/outcome of each instance of X and £ is the SVM learning algorithm.

  The problem is to find the optimal h using the following equation:

  $h_{optimal}$=optimize(£(X)) where $\in_X$ is minimum and optimize(£(X)) optimize the hyperparameter of the SVM learning algorithm using Bayesian Optimization method.

Data Cleaning

The basic measured Well logs are considered for this experiment. The features are as given below:

NPHI: Measures the effect of interaction with the rocks and fluids on the neutron flux.

DT: Measures the travel time of an elastic wave through the formation.

P-IMP: Measures the acoustic property of the media

GR: Measuring naturally occurring gamma radiation to characterize the rock or sediment in a borehole or drill hole.

RHOB: Measures the bulk density of media

FACIES: Rock type or class to a specific sample

In data cleaning phase, the duplicate samples, samples that contains NaN or NULL values, and outliers are removed. The data is divided into feature sets i.e. X and Class label i.e. Y as per requirement.

As we know for supervised approach, three subsets of data require from learning to deployment process. As a result, we divide the logs as training data (8 well's

Table 1: Review of different approaches used for facies classification

Authors	Year	Methods to determine Lithofacies
Wolf et. al. (2)	1982	Multivariate approach
Busch et al. (3)	1987	Neural Network for rock classification
Baldwin et. al. (4)	1990	Neural Network
Rogers et al. (5)	1992	Neural Network
Bhatt et al. (8)	2002	Modular Neural Network
Marroquín et. al. (8)	2008	Visual Data mining approach
Al-Anazi et. al. (7)	2010	Support Vector Machine (SVM)
Hall et. al. (9)	2016	Machine Learning Approach
Tschannen et. al. (1)	2017	Inception Convolutional Network
Bestagini et al. (6)	2017	Feature Augmentation based classification approach using machine learning

log), validation data (2 well's log) and testing data (3 well's log).

The histogram of training data and class distribution is given in Figure 1 and Figure 2 respectively after data cleaning phase. There are four classes in the training data. Figure 2 shows that the class 3 has the maximum instances and the class 1 is the minimum and the data set is imbalanced. As a result, in the further data preprocessing phase, we should take necessary steps to deal with the imbalanced data. The correlation among the variables including facies is given in Figure 3.

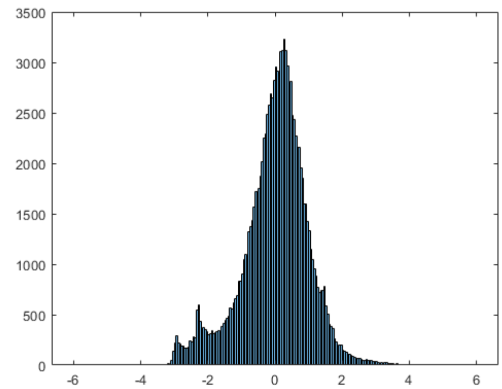

Figure 1: Histogram of feature sets of the training data

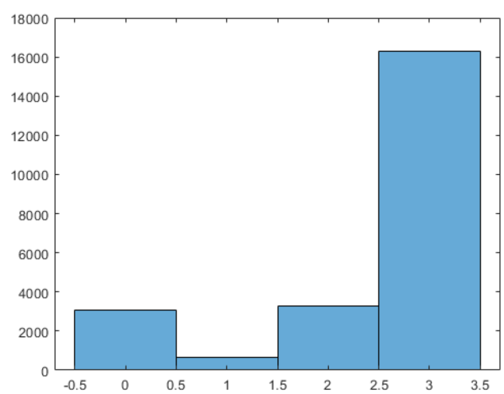

Figure 2: Histogram of class distribution of the training data.

Feature Selection: In machine learning and statistics, feature selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

- simplification of models to make them easier to interpret,
- shorter training times,
- to avoid the curse of dimensionality,
- Enhanced generalization by reducing overfitting.

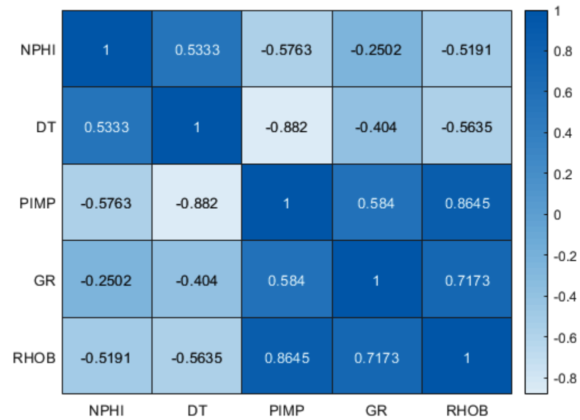

Figure 3: Correlation Matrix of the Training Data

Five feature selection techniques are used to evaluate the features and determine how much the features are contributing to find the facies. The feature sequence is as NPHI, DT, PIMP, GR, & RHOB in this experiment.

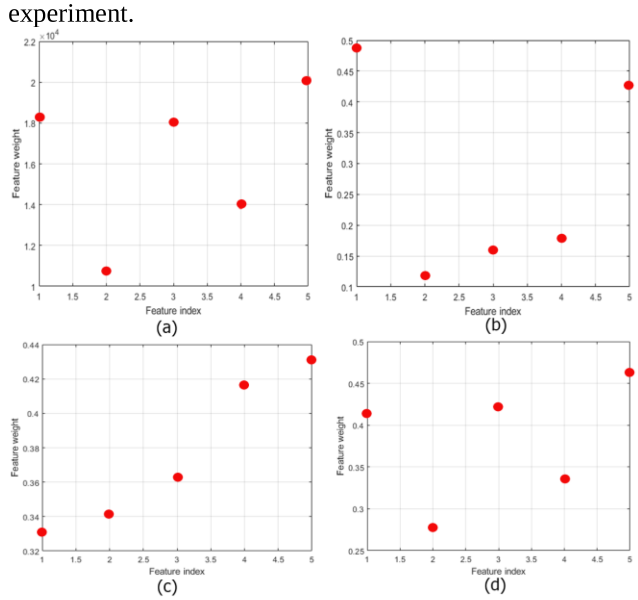

Figure 4: Feature selection using (a) Chi-Square (b) Fisher-score (c) Gini-index (d) Info-Gain approach

Figure 4 shows that all the features are contributing to find facies of a rock and all are relevant as a result, no attribute drop for further analysis.

Data Transformation:

Many machine learning algorithms attempt to find trends in the data by comparing features of data points. However, there is an issue when the features are on drastically different scales.

Table 2: System Specifications

Operating System	Windows 10 & Cent OS
Data Handling	Flat Files
Processing	Matlab, Excel
Visualization	Matlab & Excel
Processor	Intel Xeon
RAM	64GB

The goal of normalization is to make every datapoint have the same scale so each feature is equally important. The following Equation shows the Z-Score normalization of the dataset X.

$$Z-ScoreNorm(X_i)=\frac{X_i-\mu}{\sigma} \quad \ldots \quad Eq-1$$

$$Where\,\sigma=\sqrt{\frac{1}{|X|-1}\sum_{i=1}^{|X|}\left(X_i-\mu\right)^2}$$

$$\mu=\frac{1}{|X|}\sum_{i=1}^{|X|}X_i$$

Experiment Details

The Support Vector Machine (SVM) approach is implemented using MATLAB. The system configuration for this experiment is given in Table 2.

The tall array concept of MATLAB provides the scalability of the model. It can deal with high volume of data and compatible to Apache Hadoop. MATLAB 2018a version is used in this experiment and for visualization and model building. For multiclass classification, OneVsAll SVM model is used in this experiment.

Result and Discussion

The hyperparameters of the SVM is optimized by Bayesian optimization technique. The detailed optimization process of the objective function is given in Figure 5. After several interative process, the hyper parameters values are accepted by the optimization process. Finally, the OnevsAll SVM multiclass approach optimized with the value 986.34 for Box Constraint and 0.30078 for Kernel Scale and the value of the objective function is 0.14807 with our training dataset. The optimal hyperparameter are set to an SVM model and trained using the 10 well log data. The confusion matrix during the training process is given in Figure 6. The training accuracy is nearly 99% and the two class (2 & 3) gives some misclassified instances. Similarly, the ROC is shown in Figure 7 indicate the same as mentioned.

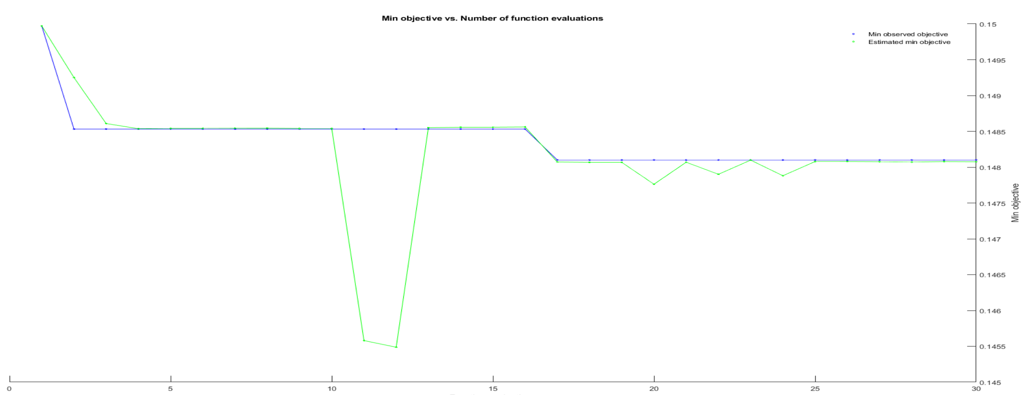

Figure 5: Optimization of objective function of SVM using Bayesian optimization technique

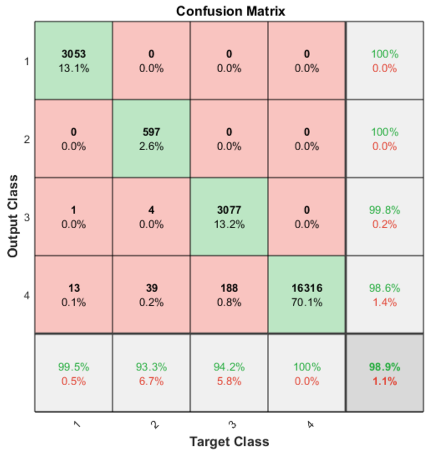

Figure 6 Confusion Matrix during model training process to find training error.

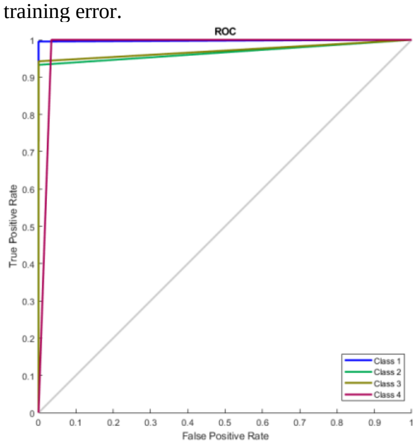

Figure 7: ROC during learning process of the SVM Model

The color code of the confusion matrix gives a clear understanding regarding the performance of the proposed model. The diagonal green color indicates the performance of each class and instance which is correctly classified and the red mixed color showing the class or instances that are not correctly classified. The objective is to achieve 100% in green color section and 0% in red mixed color section. In ideal case the accuracy showing in red mixed color should be 0 which indicates that the model achieves 100% accuracy.

During the testing process the three well's log is used as blind and fed to the model for facies prediction. The confusion matrix of the test result is as given in Figure 8 and ROC in Figure 9. All the instances of class 2 and some samples of class 3 are not classified correctly due to least distribution data of that class. The model is bias to the high distribution classes that should be minimize.

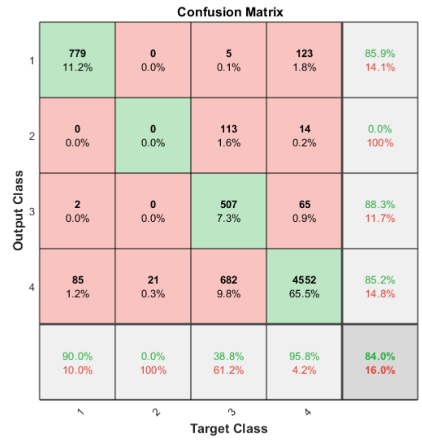

Figure 8: Confusion Matrix during testing process

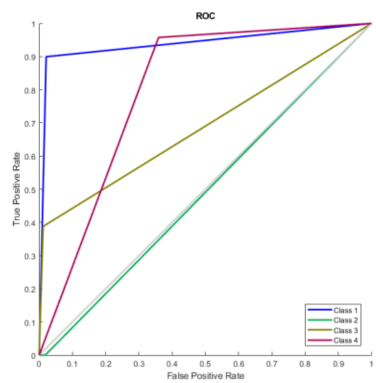

Figure 9: ROC during Testing Process

As per the Confusion matrix the accuracy for the three blind well is 84%. The Class 1 and Class 4 is better classified as compare to Class 2 and Class 3. The Class 2 is lies on the diagonal and its accuracy is about to 50%. The overall performance of the model is satisfactory. Furthermore, we have predicted a blind well whose facies already determined by the domain expert and check the result. The logs of the wells are given in Figure 10. The last two column shows the actual vs predicted facies by the SVM model. The model achieves more closer to the facies which was found by domain expert.

The specific color is assigned to a particular class is showing in ROC curves of Figure 7 and Figure 9. In ideal case the roc of each class touches the top left corner of the graph. During learning process, the all the curves touches nearly the top left corner. In testing process, the class 2 is far away as compare to the other three classes.

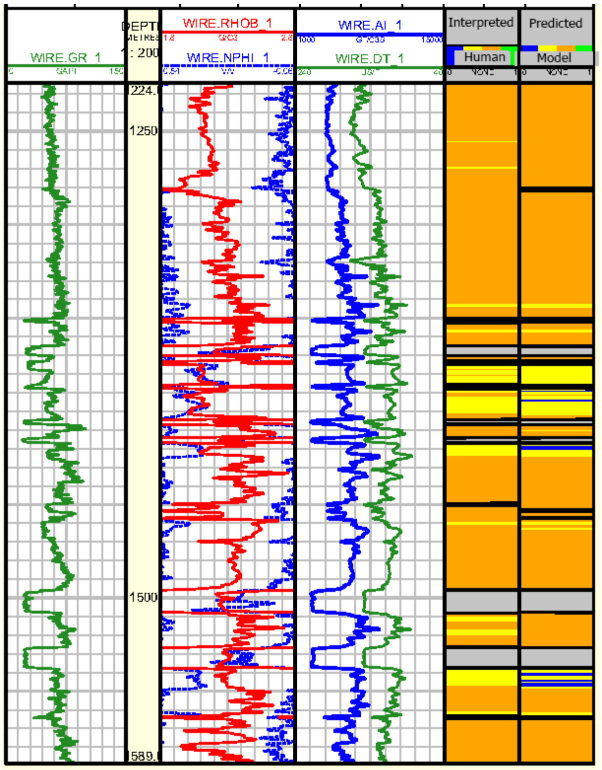

Figure 10: The Actual vs Predicted Facies of a Well

Conclusions

The input data is divided as 60% training data, 20% validation and 20% testing data. We found the training, validation and testing error in term of various performance evaluation parameters and confusion matrix. The predictive models provide a stable output both on low and high volume of data. The final classification accuracy by using both techniques is various from 84% to 90%. Therefore, we conclude that the scalable machine learning algorithms are outperforming in facies classification using well log. It also described the data flow from raw to the processed data format used for predictive analysis.

In our future work, we will apply more sophisticated processes to handle the imbalanced data and try to improve the misclassification rate during the machine learning and testing processes.

References

Tschannen, V., Delescluse, M., Rodriguez, M., & Keuper, J. (2017). Facies classification from well logs using an inception convolutional network.

Wolf, M., & Pelissier-Combescure, J. (1982, January). FACIOLOG-automatic electrofacies determination. In SPWLA 23rd Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

Busch, J. M., Fortney, W. G., & Berry, L. N. (1987). Determination of lithology from well logs by statistical analysis. SPE formation evaluation, 2(04), 412-418.

Baldwin, J. L., Bateman, R. M., & Wheatley, C. L. (1990). Application of a neural network to the problem of mineral identification from well logs. The Log Analyst, 31(05).

Rogers, S. J., Fang, J. H., Karr, C. L., & Stanley, D. A. (1992). Determination of lithology from well logs using a neural network (1). AAPG bulletin, 76(5), 731-739.

Bestagini, P., Lipari, V., & Tubaro, S. (2017). A machine learning approach to facies classification using well logs. In SEG Technical Program Expanded Abstracts 2017 (pp. 2137-2142). Society of Exploration Geophysicists.

Al-Anazi, A., & Gates, I. D. (2010). A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. Engineering Geology, 114(3-4), 267-277.

8. Marroquín, Iván Dimitri, Jean-Jules Brault, and Bruce S. Hart. "A visual data-mining methodology for seismic facies analysis: Part 1—Testing and comparison with other unsupervised clustering methods." Geophysics 74, no. 1 (2008): P1-P11.

9. Bhatt, Alpana, and Hans B. Helle. "Determination of facies from well logs using modular neural networks." Petroleum Geoscience 8, no. 3 (2002): 217-228.

10. Hall, Brendon. "Facies classification using machine learning." The Leading Edge 35, no. 10 (2016): 906-909.

11. DT Editorial Services,"BIG DATA BLACK BOOK", Dreamtech Press, 2016.