Machine Learning: Using Optimized KNN (K-Nearest Neighbors) to Predict the Facies Classifications

Hadyan Pratama[(1)]

[(1)] *Pertamina University (hadyan_pratama@rocketmail.com)*

ABSTRACT

Machine learning is getting more popularity nowadays since it has potential to improve the quality of our life in many fields through data learning and optimization. In a short explanation, machine learning can be seen as non-linear equation optimization that try to predict one output from several input parameters although it may not have any logical relationships. In the world of geoscience work, one challenging problem is lithology estimation from several logs input and/or seismic data. Until today, predicting lithology from seismic data is common and popularly known as seismic inversion or color inversion, where seismic data has been converted from impedance to lithology to help interpreter pick horizons. This study is focusing on an effort to estimate facies or lithology by only looking some well logs data input using a method named K-Nearest Neighbor method as one of many machine learning methods. I have tried to pick a unique K value with best errors for predicting facies classification. The value has been tested at training data with satisfactory results that is better than the results from tradidtional KKN. In the summary, although in this study KNN method successfully estimated the lithology, reasonable and practical results should be considered with several things such as regional geological model, a continuous improved model with newer data once acquired, coring data validation, and etc. Future work that try to expand the capability between seismic data and lithology facies would be interesting to pursue.

KEY WORDS: machine learning, k-nearest neighbor, euclidean distance, facies classification

INTRODUCTION

Machine learning is a research field that is located at the intersection of statistics, mathematics and computer science. The application of machine learning methods can be applied including oil and gas industries. For examples, Waldeland (2018) has tried to use this technique in seismic interpretation while Isebor (2012) tried to use on geological uncertainty analysis for oil production optimization.

One benefit from this method is the process of geological/lithology prediction can be fast and almost automatic with high accuracy to classify facies from well logs data input. Final model can be applied as a reference model to help geoscientist determine the lithology.

Hall (2016)) introduced facies classification problem that can be predicted with machine learning methods Zhang et al. (2017) applied XGBoost algorithm to solve the same challenge. The focus in this work is to use KNN algorithm for facies prediction. The KNN algorithm is probably the simplest machine learning algorithm. Practically, it only need to find the closest data points in the training dataset as its “nearest neighbors”. My objective is to find an optimal number of k that maximizes the accuracy so we can more accurately predict the blind data.

METHODOLOGY

Generally, there are 3 types of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. The application in this paper belongs to the category of supervised learning. This type of algorithm starts with a training of data as variable inputs that is matched with real output data. The training process will continue until the model achieves a satisfied level of accuracy on the training data. Examples of other supervised methods includes: regression, decision tree, random forest, XGBoost, logistic regression etc.

I used the nearest-neighbor method that is sometimes also named as memory-based method. Given a training set m-labeled patterns, the nearest-neighbor procedure decides that a new pattern named as X, belongs to the same category as its closest neighbors in training set. In another word, the k-nearest-neighbor method assigns a new pattern, X, to that category to which the plurality of its k closest neighbors belongs.

The distance metric used in nearest-neighbor method usually is Euclidean distance. This is the distance between two patterns $(x_{11}, x_{12}, x_{13}, \dots, x_{1n})$ and $(x_{21}, x_{22}, x_{23}, \dots, x_{2n})$ is

$$\sqrt{\sum_{j=1}^{n}(x_{1j} - x_{2j})^2} \qquad (1)$$

10.1190/SEGJ2018-139.1

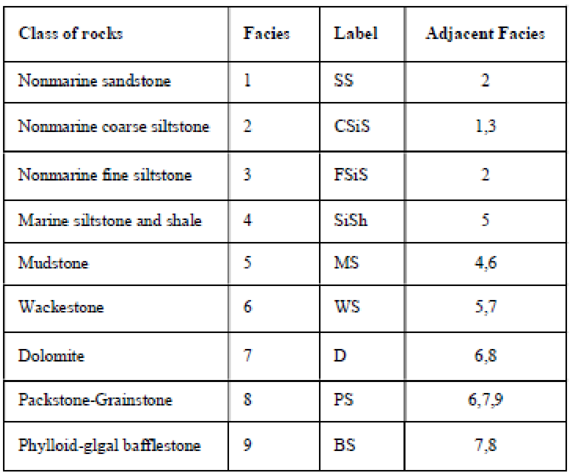
This distance measure is often modified by scaling the features so that the spread of attribute values along each dimension is approximately the same (Zhang, 2005). Further analysis can also be done by putting weight on any preferences input. For example: one may have an estimation though experiences that lithology should be more dependent on a specific data set such as density in one area, can put more weight on density logs so that it will find more closest patterns on density data set

DATA ANALYSIS AND MODEL SELECTION

The data in this work come from the Council Grove gas reservoir in Southwest Kansas. The Panoma Council Grove Field is predominantly a carbonate gas reservoir encompassing 2700 square miles in Southwestern Kansas. This dataset has ten wells (with 4149 examples), consisting of a set of seven predictor variables and a rock facies (class) for each example vector and validation (test) data (830 examples from two wells) having the same seven predictor variables in the feature vector. Facies are based on the examination of cores from nine wells taken vertically at half-foot intervals. Predictor variables include five from the wireline log measurements and two geologic constraining variables that are derived from geologic knowledge. Five wireline log measurements are; Gamma Ray (GR), Resistivity Logging (ILD_Log10), Photoelectric Effect (PE), Neutron-density porosity difference (Delta PHI), and Average neutron-density porosity (PHIND). Two geologic constraining variables are; Nonmarine-marine indicator (NM_M), and Relative position (RELPOS). These are essentially continuous variables sampled at a half-foot sample rate.

The nine discrete facies (classes of rocks), the abbreviated labels, and the corresponding adjacent facies are listed in the following Table 1 (Hall, 2016). The facies gradually blend into one another, and some of the neighboring facies are rather close. Mislabeling within these neighboring is possible to occur.

Table 1. Classification of rock types and facies in this study. Note that I use the same rock type and facies as Hall (2016) .



Class of rocks	Facies	Label	Adjacent Facies
Nonmarine sandstone	1	SS	2
Nonmarine coarse siltstone	2	CSiS	1,3
Nonmarine fine siltstone	3	FSiS	2
Marine siltstone and shale	4	SiSh	5
Mudstone	5	MS	4,6
Wackestone	6	WS	5,7
Dolomite	7	D	6,8
Packstone-Grainstone	8	PS	6,7,9
Phylloid-glgal bafflestone	9	BS	7,8

To perform machine learning modelling with nearest neighbor method, there are some steps that need to be done. First, I examined the data that will be used to train the classifier. The data consists of 5 wireline log measures, 2 indicator variables, and 1 facies label at half foot interval. In machine learning terminology, each log measurement is a feature vector that maps a set of ‘features’ (the log measures) to a class (the facies type). Then, some basic statistical analysis can be generated, for example, the distribution of each classes (Figure 1), and heatmap of features (Figure 2), which produces correlation plot for us to observe relationship between variables. These figures are the initial blocks to explore the data.

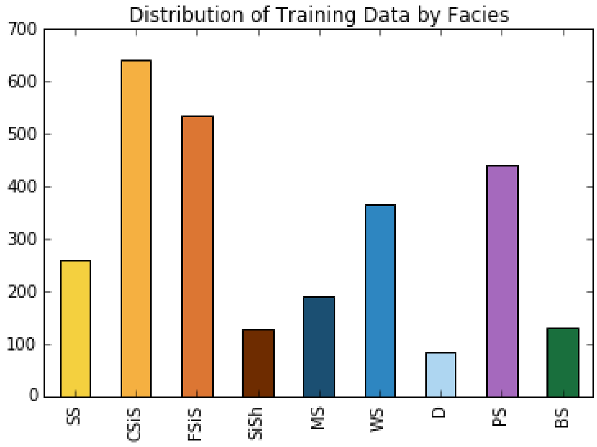


Figure 1. Distribution of facies (Hall, 2016).

10.1190/SEGJ2018-139.1

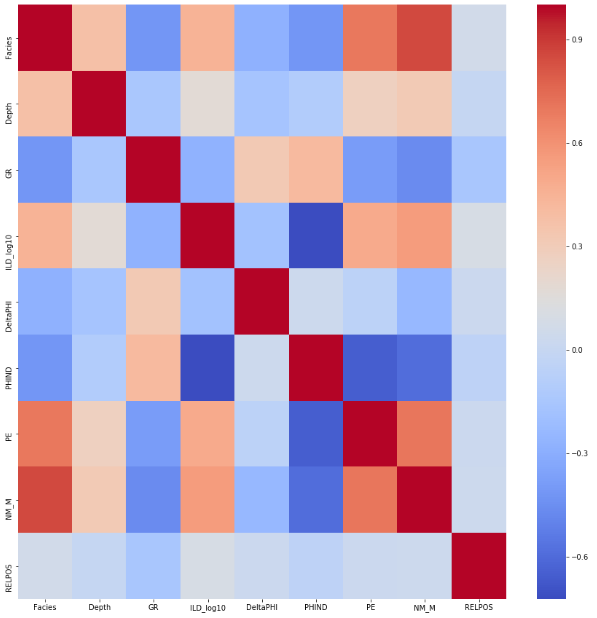


Figure 2. Correlation between variables.

Finally, I built reliable models to predict the Y values (Facies) based on X values (the seven predictor variables).

To choose the best numbers of K, we must know what is the error number from each potential number of K. This step is very different with the traditional KNN that choose k numbers by default and then iterate it. Figure 3 show plot between difference K number and Error rate. For this study K=1 yields the lowest error rate for this training data.

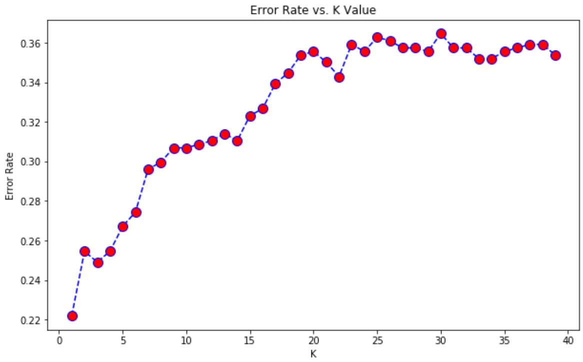


Figure 3. Error rate vs K value.

Once I get potential K value, the process continue to take the K nearest neighbors number and applied to new data point by the distances metric. In this paper I used Euclidean Distance. As we know that before, the Euclidean distance defined as the straight closest line connecting two points and calculate it as the square root of the sum of the squared difference between the elements of two vectors that is relatable with our bidimensional data plane.
From the distances matric, we can calculate the nearest data point from each category to our new data point. Finally, a cross-validation is conducted to access the performance before applying to another two blind well test data. The best accuracy (F1 score) we have so far is 0.78. Comparing to any other contestant in this contest the value of accuracy is around 0.62.

RESULT

As the final result check, I compared true classification in well log data with the predicted ones. Figure 4 show the true well log data name ‘Shrimplin’. Column 1 to 5 shows data input, while column 6 shows real data lithology and the last column show the predicted lithology. In general, the performance is very good through there are some difference where the predicted model show more lithology on a very much thinner scale. An additional filtering on the input logs may help avoid this problem.

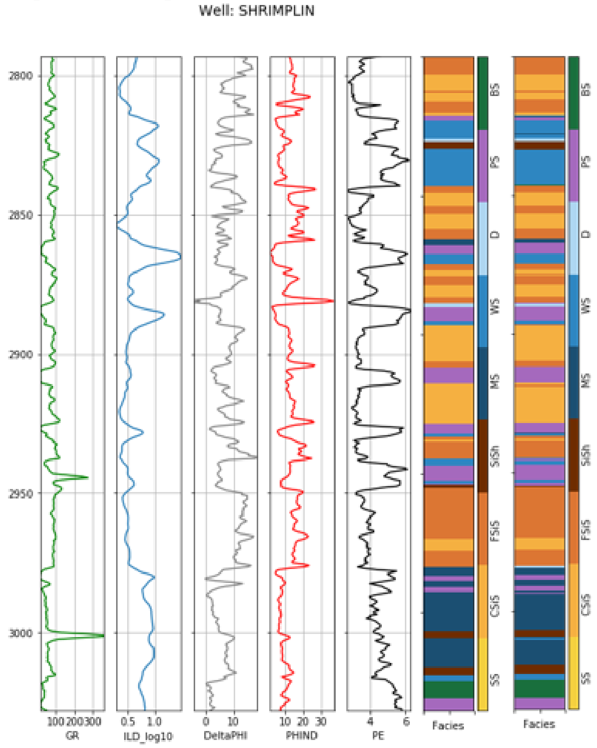


Figure 4. True and predicted ‘Shrimplin’ well log data. True lithology is shown on column 5 and predicted lithology on the last column.

Next, I applied the algorithm to two blind well log data, Stuart and Crawford. The results can be seen on Figure 5. Note that, each region may have different K value, a complete study in necessary to find unique K value for each basin.

10.1190/SEGJ2018-139.1

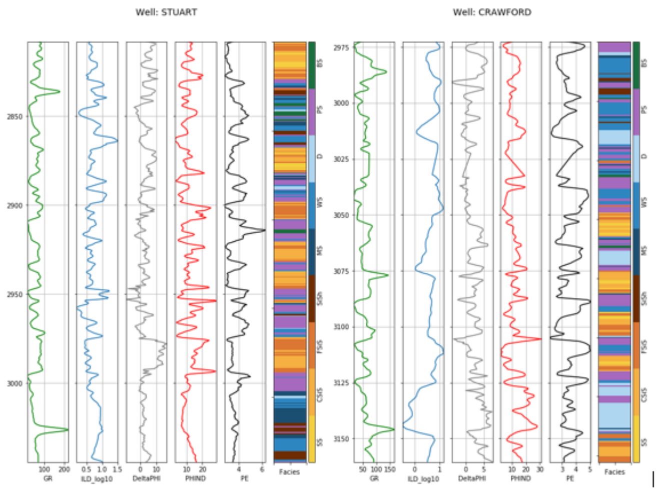


Figure 5. Predicted models on 2 blind wells.

CONCLUSION

I have successfully applied KNN with optimization method to a classification problem in the rock facies. The result can be used as a reference for lithology estimation in the region. However, although the method seems promising, knowledge about one field is necessary to compare the results. Final model should also considered regional geological model. A full automation in the process without human judgment is not recommended.

ACKNOWLEDGEMENTS

The author like to thank Brendon Hall for the Machine Learning contest and github page for tutorial, Mr. Husni Mubarak for sharing his experience and helpful discussions on well log data analytics, and Mr. Sandy K. Suhardja for reviewing.

REFERENCES

Hall, B., 2016. Facies classification using machine learning. The Leading Edge, 10, pp. 906-909.

Isebor, O. J., and Grujic, Ognjen, 2012, Use of machine learning in petroleum production optimization under geological uncertainty. Stanford University.

Müller, A. C., and Guido, Sarah, 2016, Introduction to machine learning with python, O’reilly, pp. 35-44.

Nilsson, N. J., 1998. Introduction to machine learning. Robotics Laboratory Departement of Computer Science Standford University, pp. 70-72.

Waldeland, A. U., Jensen, A. C., Gelius, L., and Solberg, A. H. S., Convolutional neural networks for automated seismic interpretation, The Leading Edge, 37, pp. 529-537.

Zhang, Licheng, and Zhan, Cheng, 2017, Machine learning in rock facies classification: an application of xgboost, SEG Library.

Zhang, M. L., and Zhou Z. H, 2005, A k-nearest neighbor based algorithm for multi-label classification, A k-nearest neighbor based algorithm for multi-label classification.

10.1190/SEGJ2018-139.1