Machine Learning Algorithms for Classification Geology Data from Well Logging

1st Timur Merembayev
IICT
Almaty, Kazakhstan
merembaevt@gmail.com

2nd Rassul Yunussov
Satbayev University
Almaty, Kazakhstan
yunussov@gmail.com

3rd Amirgaliyev Yedilkhan
Suleyman Demirel University
Kaskelen, Kazakhstan
amir_ed@mail.ru

***Abstract*—Machine learning today becomes more and more effective instrument to solve many particular problems, where there are difficulties to apply well known and described math model. In other words - it is a great tool to describe non-linear phenomena. We tried to use this technique to improve existing process of stratigraphy and lithology interpretation and reduce costs on site by applying computer leaded predictions on the basis of existing on-field collected data. Article describes usage of machine learning algorithms for several geology data stratigraphy and lithology boundaries classification based on geophysics logging data for deposits in Kazakhstan. Correct marking of stratigraphy and lithology from geophysics logging data is complex non-linear task. To solve this task we applied several algorithms of machine learning: random forest, logistic regression, gradient boosting (scikit-learn library), k – nearest neighbour (KNN) and XGBoost.**

***Index Terms*—stratigraphy, lithology, classification, machine learning, geophysics logging data**

I. Introduction

The Chu–Sarysu basins of Kazakhstan Fig. 1 is a large artesian basin that was split into two main components following the Pliocene uplift of the Karatau Mountain Range. The basins are filled with thick sandy aquifers capped by impermeable shaly beds. Mineralization, as stacked roll fronts, is hosted by sands of Upper Cretaceous to Palaeocene–Eocene age.

Moyunkum deposit Fig. 2 is a part of the Uvanas – Kanjugan metallogenic zone, where it is controlled by regional redox fronts in permeable zones of 3 Paleogene horizons, from top to bottom: Ikansk, Uyuk and Kanjugan.

Knowledge of stratigraphy and lithology boundaries Fig. 3 is important for hydrogeology, oil and gas and in-situ recovery areas. Stratigraphy boundaries for uranium deposits are provided two main information for detail study: water level horizon, which is important for monitoring of environment situation in underground and defining production reservoir that give bottom and top level for detail geology study. Lithology boundaries are provided main information about geology and detail information about type of geology – permeable or impermeable rocks, it is major data which give to engineer to make decision – to mine ore or not mine. Stratigraphy and lithology are obtained by drilling and geophysics logging. Interpretation of data is based on resistivity and spontaneous measurement.

Well logging interpretation is based on mathematical and physical modelling of the processes under study (solution of direct problems of geophysics), statistical methods (correlation and discrimination), solving systems of non-linear equations petrophysics (inverse problem of geophysics) and some other linear statistical methods.

Depending on the logging signals from the calculated physical and geological parameters - permeability, shaliness and radioactivity of host rocks, water content, salinity of aquifers, etc. Often geophysical raw acquired data have complex non-

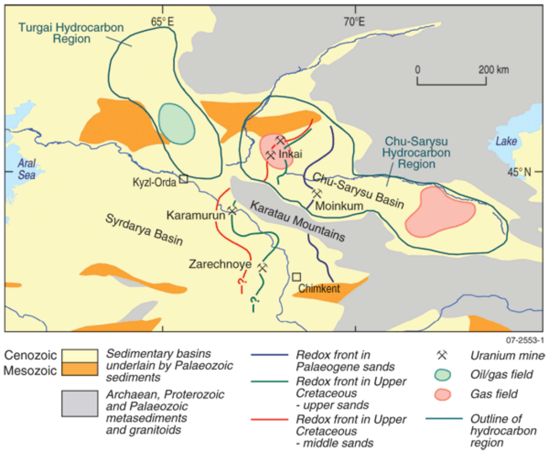
Turgai Hydrocarbon Region
65°E
70°E
0
200 km
Lake
45° N
Aral Sea
Kyzl-Orda
Inkai
Chu-Sarysu Hydrocarbon Region
Chu-Sarysu Basin
Moinkum
Karamurun
Karatau Mountains
Syrdarya Basin
Zarechnoye
Chimkent
07-2553-1
Cenozoic
Mesozoic
Sedimentary basins underlain by Palaeozoic sediments
Archaean, Proterozoic and Palaeozoic metasediments and granitoids
Redox front in Palaeogene sands
Redox front in Upper Cretaceous - upper sands
Redox front in Upper Cretaceous - middle sands
Uranium mine
Oil/gas field
Gas field
Outline of hydrocarbon region

Fig. 1. Regional geology of Chu-Sarysu basins, Southern Kazakhstan.

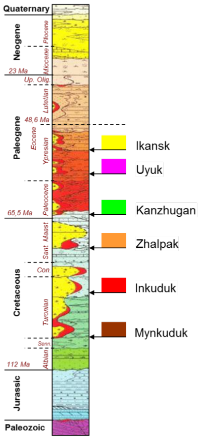

Fig. 2. Stratigraphy levels for Chu-Sarysu basins.

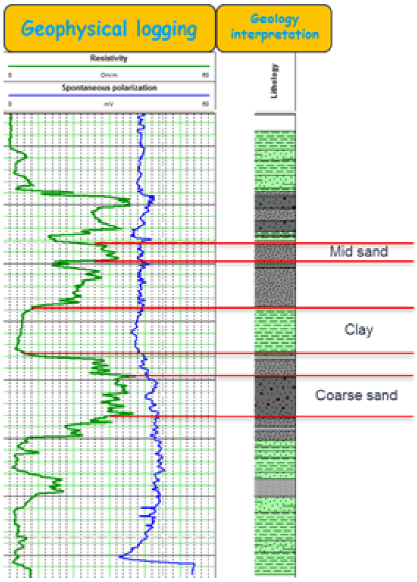

Fig. 3. Interpretation of geophysical logging and creation lithology types.

linear nature. In addition, it is necessary to take into account the geophysics measurement error signal, specificity of the geological environment. All these analytical conditions and parameters lead to the assumption that the solutions for the problems of geophysical techniques will be more easily obtained by using algorithms such as: random forest, logistic regression, gradient boosting and k – nearest neighbour. These methods have the property of adaptability, generalizations, knowledge extraction and modelling of complex non-linear dependencies in the data.

In the last few years machine learning algorithms are explored in different areas and in geology area it is often used. Before for automation process of interpretation geology was used statistical methods [1], fuzzy logic classification [3], Naïve Bayes classifier [4] also was proposed to use artificial neural network (ANN) for rocks classification [5], [6]. Naïve Bayes classifier and ANN are used a lot for lithology classification but algorithms perform poorly on minority lithology data, this research is done in article [7]. In [6] article author used a probabilistic ANN for classification and demonstrate that it is performed better than support vector machine.

In this article we describe using the machine learning algorithms for stratigraphy and lithology interpretation on uranium deposit. This problem is not complicate if we compare with lithology facies classification challenge due to complicate geology on Chu–Sarysu basins. In this article we do not use ANN but apply quite new algorithms of gradient boosting algorithms: XGBoost and gradient boosting (scikit-learn library). We compared an accuracy of algorithms for stratigraphy and lithology datasets.

II. Formalization

Underground stratification of geology is studied through geophysics logging. Significant changing of geology stratification defines boundaries of stratigraphy, interpretation geophysics logging provides rock types, for example clay, sand, organic matter, etc. – lithological facies. Our targets are view – generalization of boundaries; compare accuracy of algorithms for stratigraphy and lithological facies data. To solve it we use next data:

- Resistivity Logging (R) – well logging that measuring electric resistivity of rocks
- Spontaneous potential logging (SP) – well logging that measuring small electric potentials of rocks
- Depth (D) – values of distance between geodesic elevation and down points
- Elevation (El) – values of distance between well head (0 value) and down points

On each depth with these type of data will be defined next 8 classes of stratigraphy Tab. I:

TABLE I
Classes of Stratigraphy

Strati Code	Description
01-Q	Cenozioque-Quaternaire
02-N2	Cenozioque-Quaternaire
03-N1	Cenozioque-Quaternaire
04-P-2-3-im1	Paleogene/ Intoumak
04-P-2-3-im2	Paleogene/ Intoumak
05-P-2-2-ik	Paleogene/ Uyuk+Ikansk/ Ikansk
06-P-2-2-uk-sup	Paleogene/ Uyuk+Ikansk/ Uyuk
09-P-1-2-kn2	Paleogene aquifere/ Kanjougan

On each depth with these type of data will be defined next 6 classes of lithology Tab. II:

TABLE II
Classes of Lithology

Lithology Code	Description
1	Clay
2	Fine sand
3	Medium-grained sand
4	Coarse sand
5	Sandstone
6	Fine-grained sand

Base on this matrix of data (R,SP,D,EL), our goal is to find best machine learning algorithms which will associate the appropriate stratigraphy class according to input parameters values with best accuracy. And check can selected algorithms show same accuracy for lithology data. The training set is from Moinkum deposit and selected 42 wells in area of 2 square kilometre, see Fig. 4. For training selected 75 % (32 wells , 156512 rows) of data and testing 25 % (10 wells, 48491 rows).

III. Data Analysis and Model Selection

In this article we will speak about supervised learning algorithms for classification. During our previous researches we already applied several models and algorithms to improve remote sensing data analysis procedures [8]. We choose several algorithms: regression, random forest, KNN (these algorithms are well described in the book [9]) and XGBoost (it was created and developed by Tianqi Ch. [10]). In this article we will not go in details in process of feature engineering data tuning each algorithm to train data. All these steps of process are described in [11]. We present our comparison analysis of Resistivity log, it is based on manual algorithm of geophysics engineer interpretation logs: to detect slope of changing logging value depend on previous values if slope of curve grow up it means a rock is permeable and if down rock is impermeable as well. In this case we can detect boundaries of stratigraphy levels. To define slope of curve we will use gradients of 1st and 2nd derivatives of resistivity log:

- Gradient of RM curve, first derivative (R') – angle of changing curve is defined a boundary of rock types, if curve grows than rock is permeable, usually type is sand and if curve is down than rock is impermeable, usually type is clay
- Gradient of R curve, second derivative (R'')

Experimental results show how these two features are impact to accuracy during train and test data. Before training the model we need to study existed data: to create statistical distribution of training data Tab. III; to plot a histogram of number of training stratigraphy and lithology data Fig. 5, 6; In geoscience there is cross plots are very common tool for visualization, we create scatter matrix which visualize the variation between features: KS, PS, Grad1, Grad2, STRATI Fig. 7. To look at well logs: KS, PS, stratigraphy and lithology we use tutorial described by A. Amato del Monte's Fig. 8.

TABLE III
STATISTICS OF STRATIGRAPHY AND LITHOLOGY DATA.

STAT	ELEV	DEPTH	PS	KS	STRATI	LITHO
count	156200	156200	156200	156200	156200	31440
mean	-66.1	251.1	33.2	50.9	4.2	1.9
std	141.5	141.2	24.1	89.4	2.7	1.6
min	-335.5	0.9	-93	0	0	0
25%	-187.9	128.9	NaN	8.5	2	0
50%	-65.9	250.9	NaN	13.4	5	2
75%	56.1	373	NaN	30.8	7	3
max	199.1	516.2	163.8	2412.9	7	5

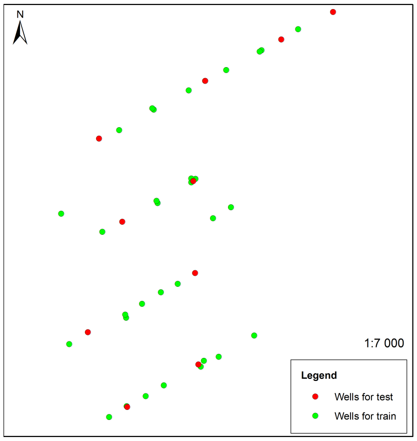

Fig. 4. Location of wells which selected for train and test algorithms

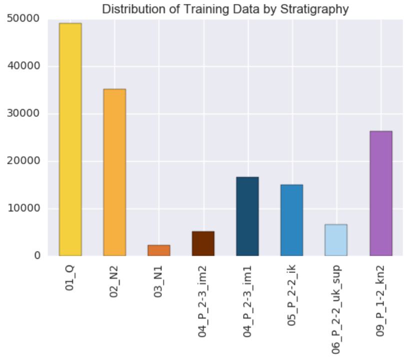

Fig. 5. Histogram of Stratigraphy data for training

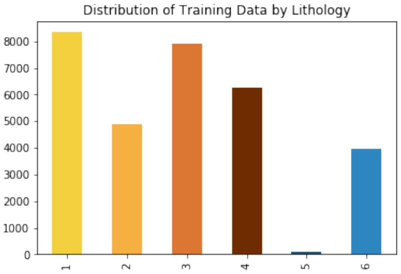

Fig. 6. Histogram of Lithology data for training

IV. Experimental results

During the experiments we have obtained the accuracy of selected algorithms and time performance of each algorithm. To measure the quality of algorithm we have selected metrics, and selected accuracy as a main metric. To validate the results of training we've selected a 3 fold cross – validation strategy. The tuning of algorithm meta-parameters is impor-

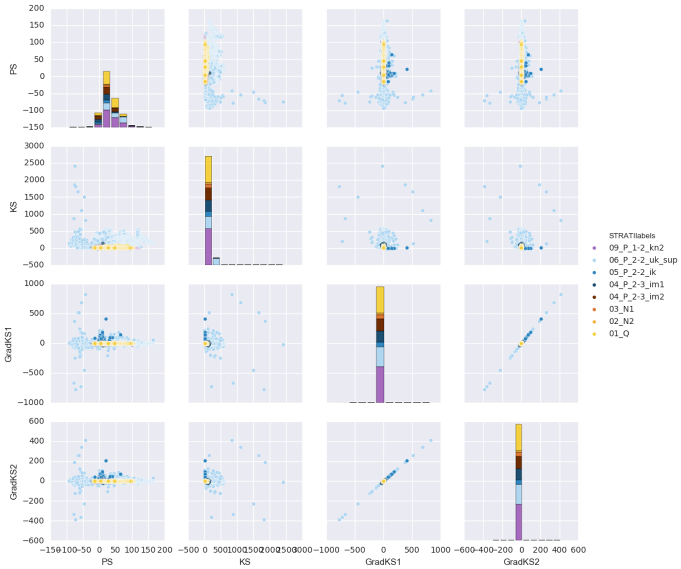

Fig. 7. Scatter matrix for all features

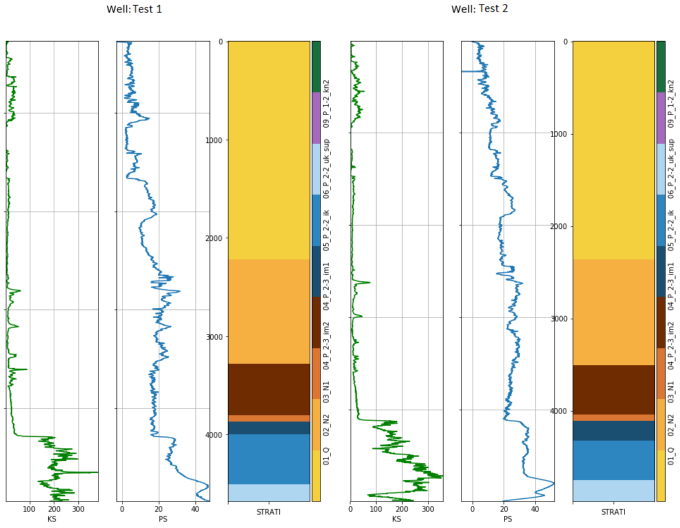

Fig. 8. Logging of two wells

tant process, and we have configured the best parameters for each of the model. To train our models we selected 4 features: (R,SP,EL,D). Another train model is done for 6 features:(R,SP,EL,D,R',R''). We can see that addition of 2 extra features have not significantly improved the accuracy in Tab. IV, V.

TABLE IV
ACCURACY FOR 4 FEATURES TEST STRATIGRAPHY DATA

Algorithm	Accuracy	Time
Gradient Boosting	0.96	2 min
KNeighbors	0.97	1 sec
Random Forest	0.96	2 min
Logistic Regression	0.95	2 min
XGBoost	0.97	2 min

TABLE V
ACCURACY FOR 6 FEATURES TEST STRATIGRAPHY DATA

Algorithm	Accuracy	Time
Gradient Boosting	0.97	3 min
KNeighbors	0.94	2 sec
Random Forest	0.96	6 min
Logistic Regression	0.96	4 min
XGBoost	0.97	4 min

TABLE VI
ACCURACY FOR 4 FEATURES TEST LITHOLOGY DATA

Algorithm	Accuracy	Time
Gradient Boosting	0.58	3 min
KNeighbors	0.53	1 sec
Random Forest	0.66	12 sec
Logistic Regression	0.61	16 sec
XGBoost	0.57	1 min

TABLE VII
ACCURACY FOR 6 FEATURES TEST LITHOLOGY DATA

Algorithm	Accuracy	Time
Gradient Boosting	0.65	6 min
KNeighbors	0.51	1 sec
Random Forest	0.70	15 sec
Logistic Regression	0.68	38 sec
XGBoost	0.65	3 min

On the basis of result we recommend to use XGBoost algorithm for making this type of classification and if you are limited by time you can use K – nearest neighbours, the algorithm has a good training time and shows high accuracy on test data.

V. CONCLUSION

In this article, we used 5 machine learning algorithms to approach stratigraphy classification based on KS, PS data and additional data like gradient (first and second derivative) for uranium deposit. All algorithms showed high accuracy on test stratigraphy data and less for lithology data. This approach helps to create a self-automatic definition of stratigraphy levels and reduces the time for data interpretation. We've successfully used this model on site and it helps to improve the whole business process of uranium mining.

Our hypothesis was that using gradient can help to improve accuracy of model but after testing and comparison of accuracy with 4 and 6 features we have obtained the results that showed not significant improvement of accuracy for stratigraphy. Tab. IV, V Indeed - the accuracy with 4 features was already very good. However for lithology data additional features (gradient) increase the accuracy and it is explained that gradient of curve is defined a boundary of rock types. If compare accuracy of algorithms between stratigraphy and lithology we can see significant difference and accuracy for lithology data is not acceptable Tab. VI, VII.

According to the achieved results, our future work will be dedicated to application of XGBoost and random forest algorithms for facies lithology classification, verify hypothesis of gradient logging for improvement accuracy and apply deep learning algorithm (convolution neural network) for facies lithology classification.

REFERENCES

[1] M. Wolf, J. Pelissier-Combescure, “Faciologautomatic electrofacies determination”. Presented at the SPWLA 23rd Annual Logging Symposium, Society of Petrophysicists and Well-Log Analysts, 1982.

[2] J. Busch, W. Fortney , L. Berry, “Determination of lithology from well logs by statistical analysis”. SPE formation evaluation, pp.412–418, 1987.

[3] B. Z. Hsieh , C. Lewis , Z.S. Lin, “Lithology identification of aquifers from geophysical well logs and fuzzy logic analysis”. Comput. Geosci., 31, pp.263-275, April (3), 2005.

[4] Y. Li , R. Anderson-Sprecher, “Facies identification from well logs: a comparison of discriminant analysis and naïve Bayes classifier”. J. Pet. Sci. Eng., 53, pp.149-157, September (3–4), 2006.

[5] S. J. Rogers , J. H. Fang, C. L. Karr and D. Stanley, “Determination of lithology from well logs using a neural network”. AAPG bulletin, pp.731–739, 1992.

[6] A. Al-Anazi, I.D. Gates, “On the capability of support vector machines to classify lithology from well logs”. Nat. Resour. Res., 19, pp.125-139, June (2), 2010.

[7] M. K. Dubois, G. C. Bohling, S. Chakrabarti, “Comparison of four approaches to a rock facies classification problem”. Comput. Geosci., 33, pp.599-617, May (5), 2007.

[8] Y. Amirgaliyev, A. Mukhamedgaliyev, R. Yunussov, T. Merembayev, “Automated method of 3D plane models creation”. vol. 3, pp.17-22. Kyrgyz Republic National Academy of Science, 2009.

[9] C. M. Bishop, “Pattern recognition and machine learning (information science and statistics)”. Springer-Verlag, 2006.

[10] Ch. Tianqi, C. Guestrin, “XGBoost: A Scalable Tree Boosting System”. pp.785-794, KDD, 2016.

[11] https://github.com/aadm/Wellmagic