AEGC Australasian Exploration Geoscience Conference Brisbane 2023

Unsupervised lithology clustering from well logs, a case study in Ha Lam coalfield, Vietnam

Duy Thong Kieu
Hanoi University of Mining and Geology
kieuduythong@humg.edu.vn

Nguyen Binh Kieu
Petro Explorers Inc.
binhkieu@petroexplorers.com

Duy Phuc Do
Vinacomin - Mining Geology Joint Stock Company (VMG)
duyphucc97@gmail.com

Ngoc Cuong Phi
Vinacomin - Mining Geology Joint Stock Company (VMG)
phingoccuongmdc@gmail.com

SUMMARY

Manual interpretation of massive well log data is time-consuming and prone to human bias. Machine Learning (ML) prediction is expected to be a robust tool to interpret lithology automatically. In this work, we apply unsupervised ML techniques such as K-means and Fuzzy C-Mean to the interpreted wells for lithology clustering. The four clustered dataset are then compared with the experts' facies interpretation to assess the clustering performance and to relabel them (removing human bias). Those labelled dataset will be fed into a supervised model for automating the facies interpretation work in the next phase of the project. The input well logs are Natural Gamma (NG) and Gamma – gamma (GG) logs from the wells in Ha Lam coalfield, Vietnam.

Key words: lithology prediction, unsupervised learning, supervised learning, coalfield.

INTRODUCTION

Lithology identification is essential in coal mining for coal reserves estimation and quality assessment. Lithology can be identified by analysing geophysics logs and/or core samples (Kumar et al., 2022). Manual lithology picking is quite limited due to the need for expertise and its time-consuming process. Additionally, lithology interpretation by experts is subjective, causing inconsistent results among different analysers, thus resulting in uncertainty in reserve calculation and consequently inappropriate development plan of the mine field.

Unsupervised clustering method is applied in the effort to minimize the above risks and defects. It is also used to utilize the experts' interpretation in generating label dataset for the supervised classification process in the next phase of the project. This paper only depicts the first phase of the project, using unsupervised learning to derive label data ((Kumar et al., 2022; Saxena et al., 2017; Singh et al., 2020)).

METHODOLOGY

Unsupervised machine learning technique is widely used for clustering, which does not require labelled data, aiming to discover patterns/insights from the data. In this work, the well logs are clustered to 4 classes which are related to "coal", "sandstone", "shale and siltstone" and "shaly and dirty coal" facies interpreted by experts. This process is applied to all wells in the dataset. The wells, which are facies interpreted by experts, are used to assess the performance of unsupervised model and name the clusters to geology domain. K-means and Fuzzy c-means (FCM) clustering methods are selected for this work (Saxena et al., 2017).

K-means clustering

K-means clustering is one the most common unsupervised learning algorithms. This techniqe aims to divide a data set into K clusters; the items in the same cluster are similar and items in the different clusters are farther apart. In the K-mean algorithm, the clustering process minimizes the distances within a cluster. The advantage of K-means clustering is that it is relatively fast. However, the challenge with K-means clustering is that the number of clusters must be pre-determined and the clustering results require post-analysing; it is not an easy task in many cases.

Fuzzy c-means clustering

Fuzzy c-means clustering is also an unsupervised learning method (Bezdek et al., 1984). It is more natural than hard clustering like K-means in several situations (Kieu et al., 2015; Kitzig et al., 2016). FCM clustering is applied in many areas of geophysics. Dekkers et al. (2014) untilized FCM clustering to identify rock units of distinct rock magnetic properties due to differing geological conditions. FCM can also define pseudo-lithology from geophysical data (Kieu and Kepic, 2020).

Evaluation metrics

Evaluation of classification and cluster versus manual interpretation

C_matrix=[cm_{ij}], where i =1, 2, …, N; j=1, 2, …, N, N is number of lithology

- True Positive (TP): It means the actual value and the predicted values are the same.

  TP(i)=cm(i,i)

- False Negative (FN): This means the actual value is positive, but the model has predicted it as negative. The sum of values of corresponding rows except the TP value.

$$\mathrm{FN(i)} = \left(\sum_{j=1}^{N} cm_{ij}\right) - cm_{ii}$$

- False Positive (FP): This means the actual value is negative, but the model has predicted it as positive. The sum of values of corresponding column except the TP value.

$$\mathrm{FP(i)} = \left(\sum_{j=1}^{N} cm_{ji}\right) - cm_{ii}$$

- True Negative (TN): It means the actual value and the predicted values are the same. The sum of values of all columns and row except the values of that class that we are calculating the values

$$TN(i) = \left(\sum_{i=1}^{N}\sum_{j=1}^{N} cm_{ij}\right) - TP(i) - FP(i) - FN(i)$$

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = TP/(TP+FN)
Specificity =TN/(TN+FP)
Precision = TP/(TP+FP)

A CASE STUDY FROM HA LAM COALFIELD, VIETNAM

Introduction of Ha Lam coalfield and the challenges

On average, there is a huge volume of coal exploration and drilling projects in the Northeast region of Vietnam every year. Therefore, the large number of boreholes and the well-bore geophysical data are often processed by geophysicists. However, the analysis results are subjective, and the analysis process is time-consuming. Therefore, developing a machine learning approach to identify coal seams and lithology from well logs can provide a geophysical stratigraphic column, reducing interpretation time and increasing the accuracy of the results (VGM, 2021).

Data sets

The data set includes natural gamma (NG) and gamma – gamma (GG) logs of ten boreholes**,** the statistical summary of the data is shown in Table 1. The Table 1 and Figure 1 show that the data vary with boreholes. The relationship between GG and NG is nonlinear.

		All Holes	Hole 1	Hole 2	Hole 3	Hole 4	Hole 5	Hole 6	Hole 7	Hole 8	Hole 9	Hole 10
Number of samples		35948	1661	2089	2454	4699	3920	2486	2560	3725	6572	5781
GG	min	2.54	2.98	2.85	2.66	3.10	2.80	2.54	2.79	2.87	2.81	2.67
	max	3.93	3.88	3.88	3.85	3.82	3.92	3.83	3.84	3.83	3.93	3.92
	mean	3.20	3.18	3.18	3.21	3.32	3.28	3.16	3.11	3.13	3.21	3.15
NG	min	0.03	0.54	0.76	0.76	0.18	0.24	1.07	0.03	0.60	0.08	0.10
	max	2.31	2.22	2.22	2.16	2.31	2.23	2.17	2.19	2.18	1.98	2.20
	mean	1.74	1.70	1.85	1.86	1.86	1.75	1.93	1.71	1.80	1.55	1.68

Table 1. Summary well logs information of the ten holes.

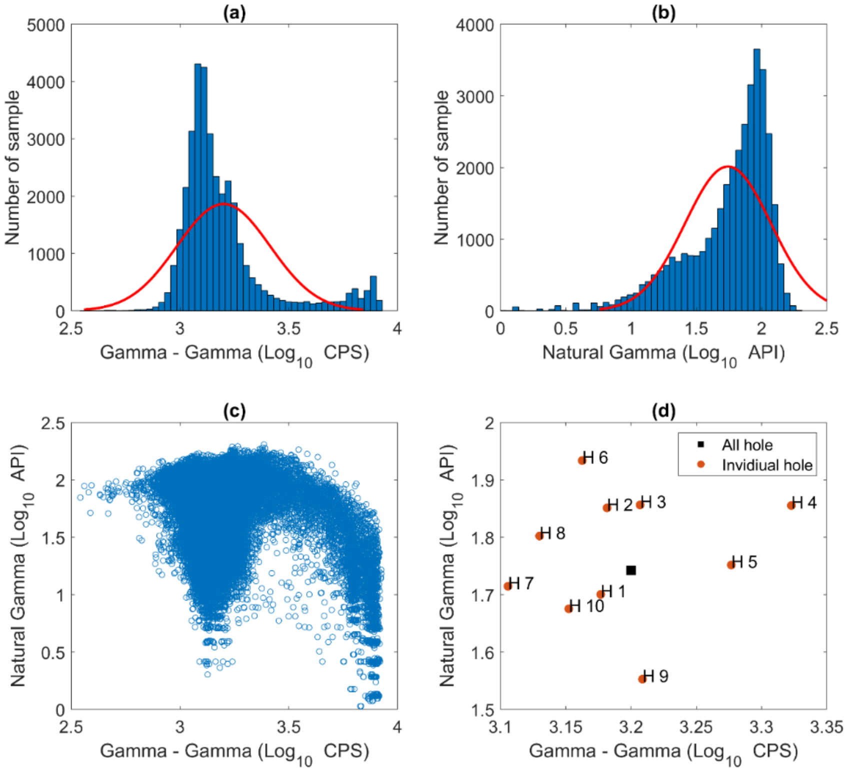

Figure 1.The well logs data of the ten holes. Histogram (bar chars) and distribution function (red lines) of Gamma – Gamma logs (a) and Natural gamma log (b). The cross-plot of all data (c) and mean values (d).

According to expert analysis, these ten boreholes comprise four lithology units: 1- Sandstone; 2- Siltstone and shale; 3- Shaly and Dirty coal; 4- Coal. Based on the distribution of the mean values (Figure 1d), the data are divided into training and testing sets. The training data include wells: 1, 3, 4, 5, 6, 7, 8 and 9. The test set comprises 2 and 10 wells. The statistical information of the data with lithology units is presented in Table 2 and Figure 2.

		Training				Testing			
Lithology code		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Number of samples		15431	9255	1388	2004	4835	2419	194	422
	min	2.54	2.66	2.98	3.05	2.78	2.67	3.02	3.01
GG	max	3.69	3.90	3.89	3.93	3.54	3.87	3.88	3.92
	mean	3.14	3.16	3.46	3.80	3.09	3.16	3.50	3.76
	min	0.30	0.40	0.96	0.03	0.71	0.76	1.18	0.10
NG	max	2.21	2.29	2.31	2.08	2.18	2.22	2.12	2.09
	mean	1.69	1.96	1.84	1.18	1.64	1.96	1.87	1.18

Table 2.Statistics information of the well logs data by training and testing datasets in four lithology units: 1- Sandstone; 2- Siltstone and shale; 3- Shaly and Dirty coal; 4- Coal.

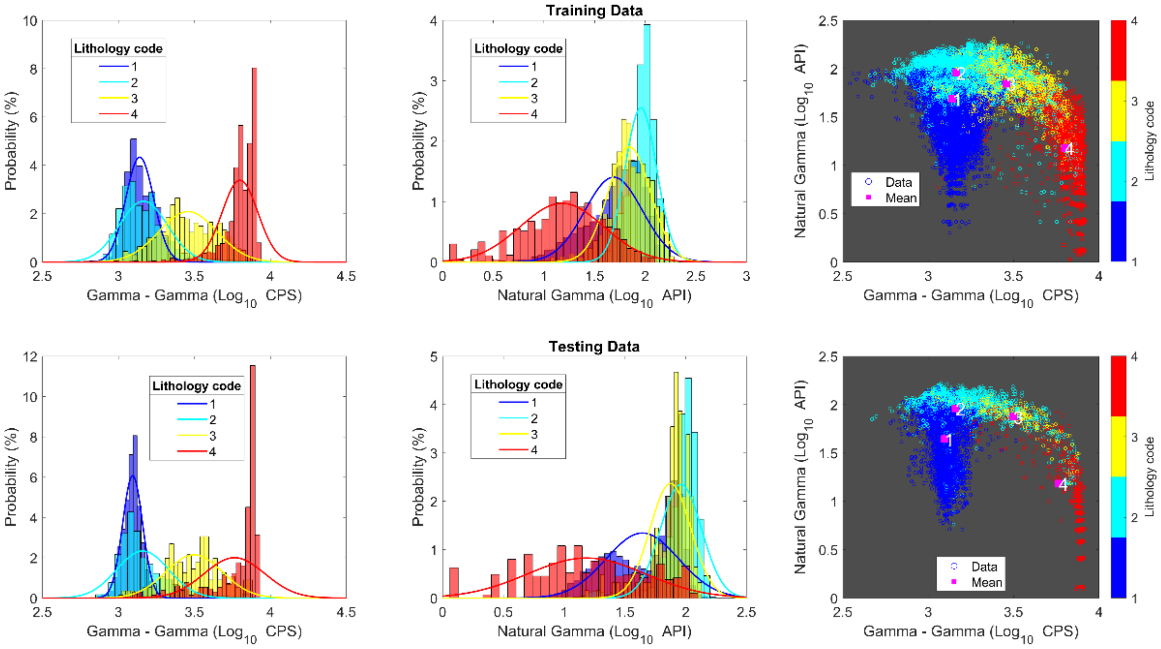

Figure 2. The lithology unit 4 is well separated from other units; unit 1 and 2 are overlap. The unit 1, 2 and 3 are overlap and separate from unit 4. Coal (unit 4) will be easily distinguished

There is a significant overlap among facies by experts' interpretation on both training and test sets, which can come from either the complex rock property changes or experts' facies cut-off inconsistency or even both (Figure 2). The histogram of each cluster using K-means (Figure 3) and FCM (Figure 4) are almost similar the lithological units (Figure 2). In general, the mean values of each lithological units are closer to centre values of K-means (Figure 3c) than cntre values of FCM results (Figure 4below).

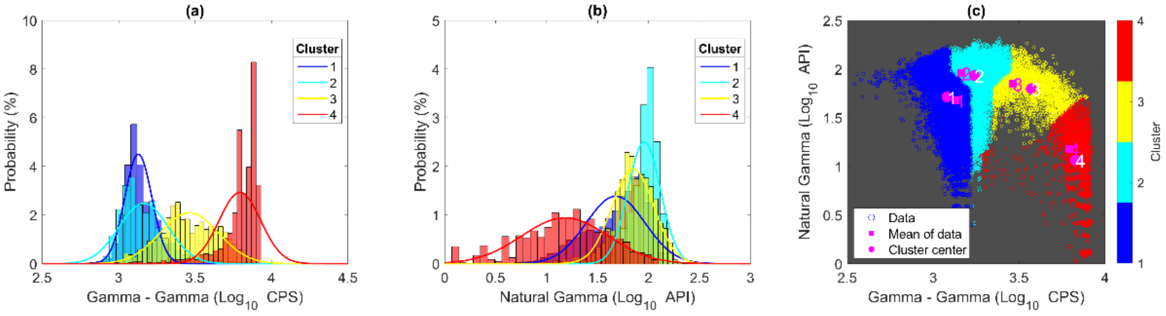

Figure 3. Histogram of GG (a) and NG (b). Cross-plot between GG and NG (c) coloured by four clusters predicted by K-means technique on training set.

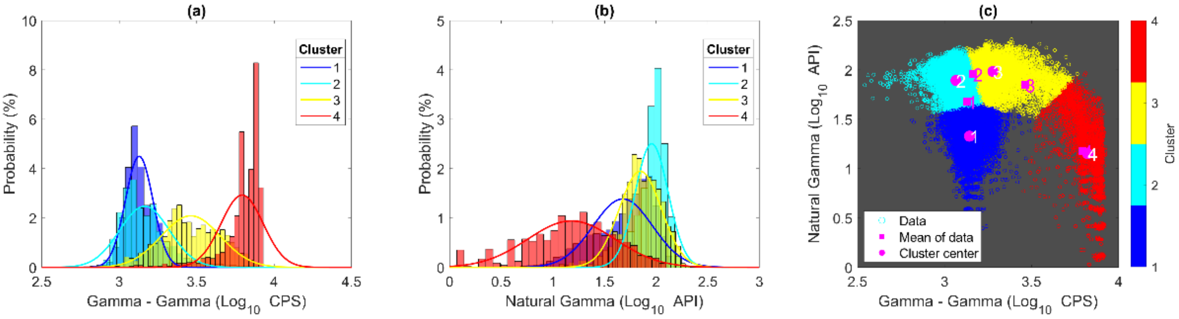

Figure 4. Histogram of GG (a) and NG (b). Cross-plot between GG and NG (c) coloured four clusters predicted by FCM technique on training set.

Cluster	Training							
	Accuracy		Sensitivity		Specificity		Precision	
	K-means	FCM	K-means	FCM	K-means	FCM	K-means	FCM
1	0.68	0.62	0.71	0.32	0.64	0.99	0.71	0.97
2	0.65	0.59	0.46	0.56	0.74	0.61	0.47	0.41
3	0.94	0.73	0.53	0.74	0.96	0.73	0.44	0.13
4	0.98	0.98	0.82	0.93	0.99	0.98	0.9	0.82
	Testing							
1	0.77	0.77	0.91	0.91	0.55	0.55	0.76	0.76
2	0.74	0.74	0.37	0.37	0.9	0.9	0.63	0.63
3	0.96	0.96	0.58	0.58	0.97	0.97	0.33	0.33
4	0.98	0.98	0.72	0.72	0.99	0.99	0.88	0.88

Table 3. Assessment of clustering results for K-means and FCM clustering vs manual interpretation by experts.

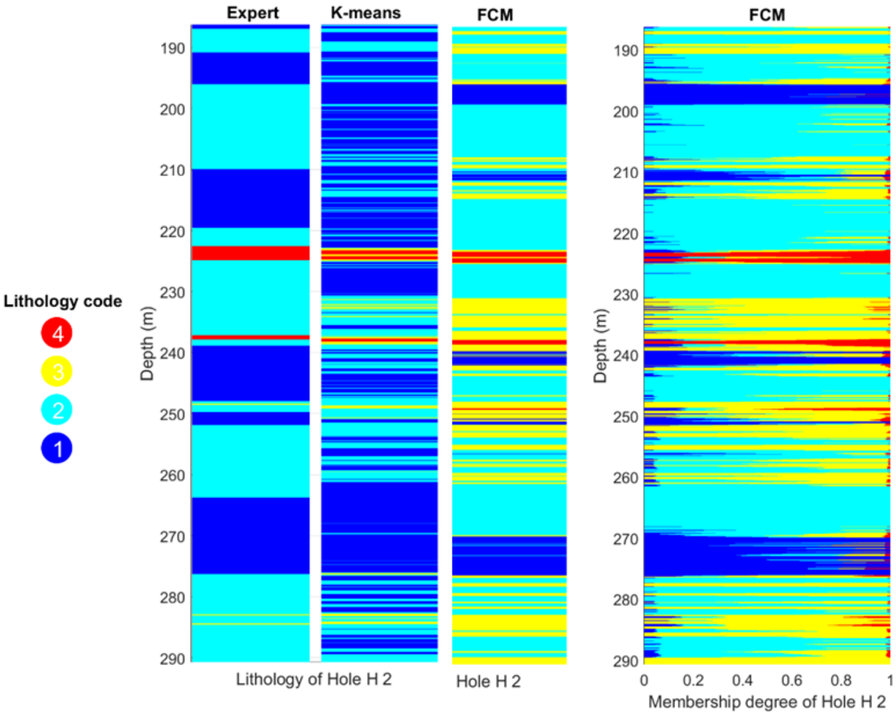

Figure 5. Comparison of expert interpretation vs K-means and FCM clustering techniques. The membership degree of FCM technique shows a better correlation with the sample. The sample can be set in multiple groups as the fact of geology rock can be interpreted to several facies depend on analysers.

Unit 3 and 4 clustering show a perfect performance in both training and test sets (Table 3 and Figure 5). The non-coal facies performance seems poor; it could come from less attention of the interpreter due to non reservoir target. The unsupervised learning results can clearly define the "facies", where FCM show apparently more logical outcome. ML based facies clustering not only reveals more detailed geologic column in non-coal sections compared to experts (Figure 6), but also more precise at the target sections (Figure 6).

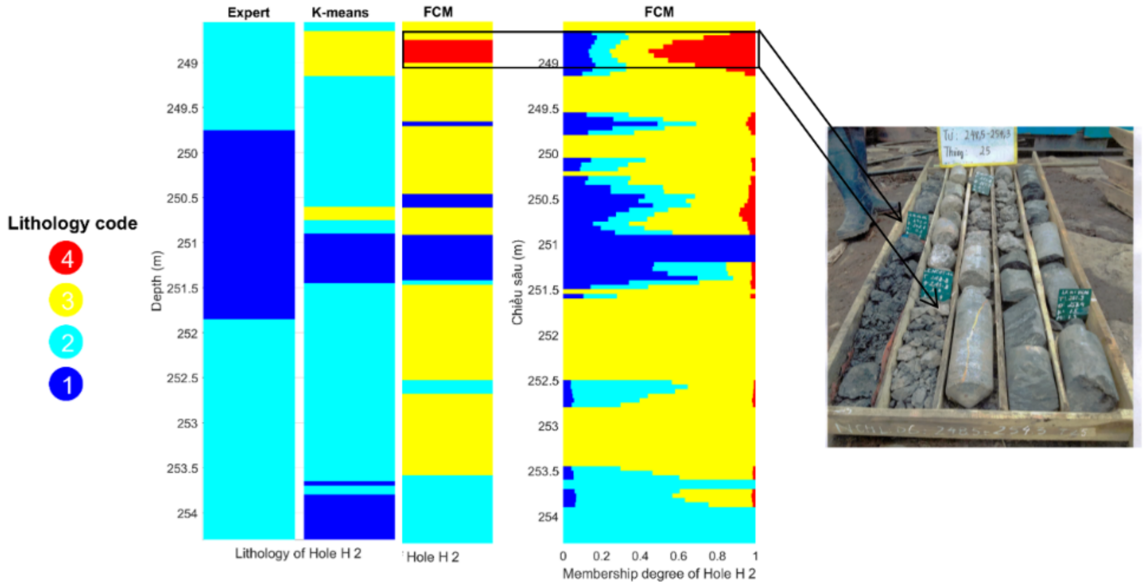

Figure 6. Comparison of expert interpretation vs K-means and FCM clustering techniques. The mark area shows the consistent between the clustering results and drill cores.

CONCLUSION

The K-means and Fuzzy C-means unsupervised predictions show more detailed facies (unit 1, 2) than the expert interpretation, where the unit 4 (the main target) is predicted with the highest accuracy. FCM with membership degree clustering outperforms in predicting coal related facies, especially unit 4 (the main target – coal), it shows very good correlation to the core samples. The detailed and consistent facies interpretation helps to minimize the uncertainty in reserve estimation thus leading to a more effective development plan. The predicted clusters are renamed to geology domain, these become labelled dataset. Lacking labelled data is one of the biggest challenges in applying ML/DL in the mining industry. The labelled dataset will be fed into the supervised learning model for automation process in the next phase of the project.

ACKNOWLEDGMENTS

We would like to thank Vinacomin - Mining Geology Join Stock Company (VMG) for permission to use the data sets.

REFERENCES

Bezdek, J. C., Ehrlich, R., and Full, W., 1984, FCM: The fuzzy c-means clustering algorithm: Computers & Geosciences, v. 10, no. 2-3, p. 191-203.

Dekkers, M. J., Heslop, D., Herrero-Bervera, E., Acton, G., and Krasa, D., 2014, Insights into magmatic processes and hydrothermal alteration of in situ superfast spreading ocean crust at ODP/IODP site 1256 from a cluster analysis of rock magnetic properties: Geochemistry, Geophysics, Geosystems, v. 15, no. 8, p. 3430-3447.

Kieu, D. T., and Kepic, A., 2020, Seismic-impedance inversion with fuzzy clustering constraints: an example from the Carlin Gold District, Nevada, USA: Geophysical Prospecting, v. 68, no. 1 - Cost-Effective and Innovative Mineral Exploration Solutions, p. 103-128.

Kieu, T. D., Kepic, A., and Kitzig, C., 2015, Classification of Geochemical and Petrophysical Data by Using Fuzzy Clustering, 24th International Geophysical Conference and Exhibition, Volume 2015: Perth, Australia, ASEG, p. 1-4.

Kitzig, M. C., Kepic, A., and Kieu, D. T., 2016, Testing cluster analysis on combined petrophysical and geochemical data for rock mass classification: Exploration Geophysics, p. -.

Kumar, T., Seelam, N. K., and Rao, G. S., 2022, Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India: Journal of Applied Geophysics, v. 199, p. 104605.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T., 2017, A review of clustering techniques and developments: Neurocomputing, v. 267, p. 664-681.

Singh, H., Seol, Y., and Myshakin, E. M., 2020, Automated Well-Log Processing and Lithology Classification by Identifying Optimal Features Through Unsupervised and Supervised Machine-Learning Algorithms: SPE Journal, v. 25, no. 05, p. 2778-2800.

Vinacomin - Mining Geology Join Stock Company (VMG), 2021. Ha Lam reserve upgrade report.