Supervised Machine Learning in Electrofacies Classification: A Rough Set Theory Approach

Touhid Mohammad Hossain¹, Junzo Wataada², Maman Hermana², Izzatdin A Aziz¹

¹Department of Computer and Information Sciences, Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Malaysia.
²Department of Geosciences, Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Malaysia

Abstract. Electrofacies were initially introduced for defining a set of recorded log responses in order to characterize a bed and permitted it to be distinguished from the other rock units as an improvement to the traditional use of well logs. Grouping a formation into electrofacies can be used in lithology prediction, reservoir characterization and discrimination. Usually Multivariate statistical analyses, such as principal component analysis ‘PCA’ and cluster analysis are used for this purpose. In this study Extra Tree Classifier (ETC) based feature selection method is used to select the important attributes and three distinctive electrofacies were extracted from the dendrogram plot using the selected attributes. Finally, we proposed a rough set theory (RST) based white box classification approach to extract the pattern of the electrofacies in the form of decision rules which will allow the geosciences researchers to correlate the electrofacies with the lithofacies from the extracted rough set (RS) rules.

1. Introduction
Well log clustering is done is to find the similarities or dissimilarities among the log responses in the multivariate space with the aim of grouping them together into distinct classes known as electrofacies [1] An electrofacies contains a unique log response set, which can characterize the substantial properties of the fluids and rocks contained in the volume recorded from the logging tools. The identified clusters or electrofacies can interpret and reflect the hydrologic, lithologic, and diagenetic characteristics of a uncored well. If we apply the supplementary information such as geological insight or core observations, the identified electrofacies (EF) groups can be calibrated to ensure their interpretable geological meaning and for this electrofacies classification process needs to be explained efficiently.

There is no particular method to define electrofacies. The usual requirements are that they should be defined from a reliable log set of petrophysical measurements and the similarities or dissimilarities among the down-hole intervals needs to be expressed quantitatively from the sample log responses. Most recently several mathematical models have been introduced to automate the task of electrofacies identification and classification. For defining electrofacies the researches principally consist of using principal components analysis, cluster analysis, and discriminant analysis. Several other methods for classifying electrofacies have been proposed in the literature [2-5][10] . Almost all the available commercial software packages for subsurface modeling have electrofacies functions. Unfortunately, explanations about how these functions generate the results are rarely explained, and the procedures run as “black-boxes.” In [6] Feed Forward Neural Network and clustering are used for the
determination of classification of electrofacies which is also a black-box approach. This research proposes a critical “white-box” approach towards classifying and defining the electrofacies constructed from wireline logs by K-means clustering using a RST based Rule Induction method. RST is a machine learning tool that performs granular computation from a vague idea (set) depending on two vivid concepts, which are lower approximations and upper approximations. To perform granular computation, RST requires only the provided data [7]. RST performs by employing a granular understanding of the provided dataset. RST includes numerous advantages [8,9]. The most major advantages are shown here.

- Provides effective algorithms for finding invisible patterns inside the dataset.
- Generates a nominal dataset called the data reduction and, thus, shows the significance of data.
- Generates set of significant decision rules from the objects that are easily understandable and explainable.

2. Methodology:
RST based electrofacies classification model combines of several steps which are described in the subsections below. Figure 1 illustrates the overall workflow.

2.1. Feature Selection:
Ideally, there can be a huge number of log attributes available for electrofacies calculation. Important attributes can be selected depending on response and resolution to the properties of major interest. In our research we used Extra Trees Classifier (ETC) for selecting the important attributes from a set of 28 well log attributes. ETC, also called as Extremely Randomized Trees Classifier is an update to random forest classifier section of ensemble learning technique which combines the outputs of numerous de-correlated decision trees together as a “forest” to calculate it’s classification result [11]. For feature selection each feature is ordered descendingly according to the Gini Importance (a mathematical criterion used in the decision of feature of split) that is recorded during the construction of the forest of each feature and the user can select the prominent k features according to the application criteria. For our experiment we considered k=10 to get the 10 most important attributes.

![Data Flow Diagram of the model](image-url)
that contribute for the 10 lithology decision classes. For feature selection, lithology classes are used as the decision variable because the constructed electrofacies classes will be used for reflecting lithologic characteristics in the extension of the work. The calculated feature importance is shown in figure 2.

![Figure 2. Top 10 important features achieved by ETC.](image)

The primary description of the selected attributes and the summaries are shown in Table 1 below:

| Abbreviation | Gamma Ray | Neutron Porosity | Density Correlation | to Electric Effect | Density Porosity |
|--------------|-----------|-----------------|---------------------|-------------------|-----------------|
| Unit         | .api      | .decp           | .g/cc               | .none             | .decp           |
| Min.         | 46.49     | 0.1053          | -0.0565             | 0.5105            | 0.081           |
| Mean         | 140.85    | 0.2942          | 0.02471             | 1.8684            | 0.254           |
| Max.         | 385.89    | 1.2316          | 0.3244              | 3.7343            | 0.8082          |

| Abbreviation | Conductivity | Caliper | Borehole Volume | Hole Diameter | pressional Sonic |
|--------------|--------------|--------|-----------------|---------------|------------------|
| Unit         | .mma/m       | .in    | .m3             | .in           | .uspf            |
| Min.         | 113.9        | 5.297  | 0.4481          | 5.297         | 60.81            |
| Mean         | 1240.4       | 5.863  | 2.0633          | 5.863         | 101.69           |
| Max.         | 9215.5       | 7.58   | 4.265           | 7.58          | 159.61           |

2.2. Clustering the logs:
The goal of clustering is to classify a dataset into several groups which are externally isolated and internally homogeneous on the basis of a measure of similarity and dissimilarity among the groups. The number of electrofacies classes is arbitrary because electrofacies are empirically defined. The
number of defined electrofacies usually depends on the number of log properties used in the system and the joint characteristics of the statistical distributions of the log readings [17]. It also illustrates the goal of electrofacies classification and the way in where the final categorization will be interpreted and used. In our experiment, 3 electrofacies classes are constructed by non hierarchical k-means clustering algorithm. For clustering the values of the attributes have been standardized to z scores for getting a proper distribution of the dataset. In figure 3 the clusters classes are shown.

2.3. Principal Component Analysis (PCA):
PCA is used to reduce the dimensionality of the data and to summarize and visualize the data effectively without a significant loss of information. Several studies has been done using PCA for related problems and to generate principal component values for electrofacies classification [12,13]. In our research we implemented RST based electrofacies classification using the raw value from the logs and PCA is done to visualize the distribution of the logs according to the electrofacies classes defined by the clustering algorithm. For PCA analysis standardized values of are used. From figure 3 it is visible that the electrofacies classes follow proper disjoint distribution which is projected by the biplot generated by principal component 1 (PC1) and principal component 2 (PC2).

![Figure 3. Three clusters identified from Principal Component Analysis using PC1 and PC2](image)

2.4. Rules Construction using RST Rule Induction algorithm
- In this step, all the selected attributes’ raw values have been discretized into 15 equal length intervals. The discritized intervals will allow RST to make the rules using the value intervals
consisting of different value ranges of the attributes as shown in table 1.

- The whole well log raw dataset is divided into two, training set (70%) and testing set (30%). The training and testing sets consist of 1922 and 825 log samples respectively.
- In this step the training set is used to extract rules using Rule Induction algorithm of RST. The RST rule $r$ can be expressed as,

$$R: (b_{x1} = V_{x1}) \land (b_{x2} = K_{x2}), \land ... \land (b_{xn} = K_{xn}) \rightarrow (d = K)$$

where, $b_{xy}$ and $K_{xy}$ denote conditional attributes and the attribute values respectively. The left-hand side of the rule $R$ is the set of attribute value sets which is the condition part and denoted as cond($R$), and the right hand side of $R$ is referred to the decision part, dec($R$). In short, a RST rule can be denoted as, IF cond($R$) THEN dec($R$).

A set consisting of 70 rules are generated. Some sample rules are provided in table 2 below:

| Rule No. | Rules                                                                 | Support Size | Laplace |
|---------|----------------------------------------------------------------------|--------------|----------|
| 1       | IF DRHO is (-0.0311,-0.00571) and NPHI is (0.18,0.255] and DPHI is (0.226,0.275] THEN Electrofacies is 2. | 135          | .9855    |
| 2       | IF NPHI is (0.18,0.255] and DTC is (93.7,100] THEN Electrofacies is 2. | 137          | .9857    |

2.5. **Model Evaluation:**
In this step the generated rules are applied on the testing set to evaluate the model's performance. The result shows that the model has an error of 5.0442% only the accuracy of the model is 0.949558 or 94.9558%.

3. **Results and Discussion:**
The classification result (94.9558%) indicates that the model is robust for explaining the conditions for the 3 EF groups and the generated rules can be used for explaining and the describing the conditions for the log samples to belong into different electrofacies classes. For example, rule no. 2 in table 2 says, if NPHI (Neutron Porosity) value ranges from .18 deep to .255 deep and DTC (compressional wave travel time) ranges from 93.7 uspf to 100 uspf then the sample belongs to Electrofacies 2. The description of the rules can produce a whitebox model that explains the reasons for a log sample belonging to a particular electrofacies class. Based on the easily understandable and interpretable geological meanings in terms of rules the geoscientists can extract the hydrologic, lithologic, diagenetic or other characteristics according to the need.

4. **Conclusion:**
The main idea of the research was to predict electrofacies accurately which has very significant bearing on other reservoir parameter calculations like permeability, lithology, depositional facies etc. For classifying electrofacies we used a RST based Rule Induction method to identify clusters from unique description of well log responses reflecting minerals and lithofacies from the logged interval. Our results and methodology show that a RST based EF classification will help in doing the accurate prediction of EF classes and the rules can be used for further analysis for determining lithology of other studies. In the extension of this research we will construct a model to define the lithology with the extracted RST rules.
Acknowledgment
The author would like to thank Petroleum Research Fund (PRF), Cost Center 0153AB-A33 and Universiti Teknologi PETRONAS (UTP) for Graduate Research Assistantship (GRA) scheme.

Reference:
[1] Euzen, T., Delamaide, T., Feuchtwanger, and K.D. Kingsmith, Well Log Cluster Analysis: An Innovative Tool for Unconventional Exploration, in CSUG/SPE 137822, Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, Canada, 19–21 October, 2010.
[2] Berteig V et al, Lithofacies prediction from well data, in Transactions of SPWLA 26th logging symposium Paper TT, p25, 1985.
[3] Busch JM, Fortney WG, Berry LN (1987) Determination of lithology from well logs by statistical analysis. SPE Form Eval, vol. 2, issue 4, pp412–418.
[4] Anxionnaz H, Delfiner P, Delhomme JP, Computer-generated corelike descriptions from open-hole logs. Am Assoc Petrol Geol, vol 74, issue 4, pp375–393.
[5] Hernandez–Martinez E et al, Facies recognition using multifractal Hurst analysis: applications to well-log data, Math Geosci, 2013.URL: https://doi.org/10.1007/s1100401394456
[6] Kumar B, Kishore M, Electrofacies classification—a critical approach, in 6th international conference and exposition on petroleum geophysics, pp 822–825, Kolkata, India, 2006.
[7] Songbo Tan, Yuefen Wang and Xueqi Cheng, "Text feature rank in based on rough-set theory", in 415 IEEE/WIC/ACM International Conference on Web Intelligence, pp 659-662, November 2-5, 2007.
[8] Zdzislaw Pawlak, Rough Set Theory for Intelligent Industrial Applications, IEEE Honolulu, HI, USA, 1999.
[9] Touhid Mohammad Hossain, Junzo Watada, Maman Hermana and Hiroshi Sakai, A rough set based rule induction approach to geoscience data, in International Conference of Unconventional Modelling, Simulation & Optimization on Soft Computing and Meta Heuristics, UMSO 2018, Fukuok, Japan, December 2-5, 2018.
[10] Euzen, T, Power MR, “Well log cluster analysis and electrofacies classification: a probabilistic approach for integrating log with mineralogical data.”, Am Assoc Petrol Geol Search Discov 41277, 2014.
[11] Pierre Geurts, Damien Ernst, Louis Wehenkel, “Extremely randomized trees”, Machine Learning vo. 36, issue 1, pp. 3-42, 2006.
[12] Bucheb, J. A. and Evans, H. B., “Some Applications of Method Used in Electrofacies Identification”, The Log Analyst (Jan.- Feb. 1994) 14.
[13] Lim, Jong-Se, Kang, J.M., and Kim, J., “Multivariate Statistical Analysis for Automatic Electrofacies Determination from Well Log Measurements,” paper SPE 38028 presented at the 1997 SPE Asia Pacific Oil and Gas Conference, Kuala Lumpur, Malaysia, 14-16 April, 1997.