2022

UCS prediction by group-based machine learning method

Jimmy Xuekai Li
University of Queensland

Matt Tsang
Anglo American Metallurgical Coal

Stephen Giese
Anglo American Metallurgical Coal

Ruizhi Zhong
University of Queensland

Joan Esterle
University of Queensland

See next page for additional authors

Follow this and additional works at: https://ro.uow.edu.au/coal

Recommended Citation
https://ro.uow.edu.au/coal/841

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
Authors
Jimmy Xuekai Li, Matt Tsang, Stephen Giese, Ruizhi Zhong, Joan Esterle, Claire Pirona, Mojtaba Rajabi, and Zhongwei Chen

This conference paper is available at Research Online: https://ro.uow.edu.au/coal/841
UCS PREDICTION BY GROUP-BASED MACHINE LEARNING METHOD

Jimmy Xuekai Li¹, Matt Tsang², Stephen Giese³, Ruizhi Zhong⁴, Joan Esterle⁵, Claire Pirona⁶, Mojtaba Rajabi⁷ and Zhongwei Chen⁸

ABSTRACT: Rock Uniaxial Compressive Strength (UCS) is an important parameter for almost all aspects of geomechanics and geotechnical designs and analysis. Traditionally, it was determined by laboratory experiments or calculated from the sonic log (e.g., Sonic Transit Time (STT)) using empirical or curve fitting correlations. However, due to the complex mineral composition and heterogeneous porosity in sedimentary rocks, the data distribution pattern of the UCS-STT relation could not be precisely described by the empirical equations with simple fitting curves. To overcome this challenge, machine learning methods have been increasingly adopted in the literature to predict the UCS from the geophysical logging data. The accuracy of the machine learning predictions relies on the quality of training data, and it may vary with rock type. In this paper, the “prediction based on group classification” concept was adopted, and a Group-based Machine Learning (GML) method was introduced to predict the UCS, which has demonstrated better performance than the conventional machine learning methods. The data analysis procedure to achieve high-quality training input data was demonstrated, and the techniques of data cleaning before the training was recommended. The implementation of the GML method requires using unsupervised learning models to classify the group of rock types firstly, and then the UCS values are predicted by the machine learning models trained for different groups; for each group, multiple machine learning models are evaluated. Finally, the previous two steps were integrated for the automatic group-based UCS prediction from the geophysical logs.

INTRODUCTION

The Uniaxial Compressive Strength (UCS) has been widely used in the evaluation of the in-situ rock strength and the structural competence in the mining industry. It is conventionally measured by the laboratory rock compressional experiments, which is time-consuming and expensive. Field tests of underground rock or outcrop exposures such as ball-peen hammer test or point-load test are also frequently used as rough index test of UCS (Mark and Molinda, 2005). In Australian context, UCS is routinely measured in the laboratory or indirectly predicted by using empirical correlation with sonic logging data, either Sonic Transit Time (STT) or compressional wave velocity (Vp) (Butel et al., 2014; Hatherly et al., 2016; Zhou and Guo, 2020). This was particularly beneficial at the planning stage since the UCS estimate was performed in a faster way for the continuous borehole intervals. Summary of various empirical relationships between UCS and sonic data can be found in a number of literature, such as Aladejare et al. (2021) and Miah et al. (2020).

However, the rock strength is not only related to the sonic response in the rock mass but also correlates to the other geophysical properties such as rock density, porosity, and natural radiation that reflects elemental and mineral composition. These rock properties are often measured in the exploration boreholes by routine wireline logging programs. Thus, researchers have been using them as additional

1 Postdoctoral Research Fellow, School of Mechanical and Mining Engineering, The University of Queensland, St Lucia, 4072, Australia. Email: jimmy.li@uq.edu.au
2 Geotechnical Specialist: Modelling, Anglo American, Brisbane, 4000, Australia. Email: matt.tsang@angloamerican.com
3Specialist: Strata, Anglo American, Brisbane, 4000, Australia. Email: stephen.giese@angloamerican.com
4 Research Fellow, Centre for Natural Gas, The University of Queensland, St Lucia, 4072, Australia. Email: r.zhong@uq.edu.au
5 Emeritus Professor, School of Earth and Environmental Sciences, The University of Queensland, St Lucia, 4072, Australia. Email: j.esterle@uq.edu.au
6 PhD researcher, School of Mechanical and Mining Engineering, The University of Queensland, St Lucia, 4072, Australia. Email: c.morton@uq.net.au
7 ARC DECRA Research Fellow, School of Earth and Environmental Sciences, The University of Queensland, St Lucia, 4072, Australia. Email: m.rajabi@uq.edu.au
8 Senior Lecturer, School of Mechanical and Mining Engineering, The University of Queensland, St Lucia, 4072, Australia.
*Corresponding author. Email: zhongwei.chen@uq.edu.au
input parameters besides the sonic data to develop machine learning models to predict the UCS, which attracts increasing attention in recent years. For example, Asadi (2017) used density, penetration rate, porosity and sonic data as input parameters to train an artificial neural network (ANN) for the UCS prediction. Barzegar et al. (2020) used an ensemble tree-based machine learning model to predict the UCS of travertine rocks. Chen and Zhang (2020) developed a physics-constrained Long-Short Term Memory for the prediction of geomechanical properties. Sun et al. (2020) developed hybrid machine learning models to estimate the UCS. Chemmakh (2021) used machine learning methods to predict the UCS and tensile strength of the Bakken field rocks. More recently, Zhong et al. (2021) applied the advanced XGBoost (XGB) machine learning algorithm (Chen and Guestrin, 2016) to predict the UCS and Young’s Modulus by using geophysical logs for Australian coal mines.

Due to the limited quantity of the laboratory UCS data with the corresponding geophysical logs, we contend that the machine learning methods built on insufficient data (hundreds of data points or less) were used in the literature, resulting in the training data not being representative of the actual scenario. Zhong et al. (2021) used 4,000+ data points for the model training. However, the predicted results tended to concentrate on the centre of the data cluster due to the uneven data distribution of the training data. These difficulties impose challenges of applying machine learning methods to the UCS prediction, because it reflects the source sampling bias of a heterogeneous system.

The group before prediction concept has been adopted in the user manual of the GOHFER 3D software, where the sonic data were divided into different lithology categories before being used for Young’s modulus prediction (Barree, 2018). However, lithology might not be a good indicator to distinguish the rock type with different geotechnical properties. For example, the rocks with similar sonic velocity and UCS ratings might belong to different lithologies. The sonic-UCS empirical fitting curves were very similar for the different rock lithologies except for the coal (Barree, 2018).

In this paper, a Group-based Machine Learning (GML) model for UCS prediction was developed and implemented. Grouping data was the first process in the exploratory data analysis phase of a data science project. It divided the data into different groups based on shared characteristics or properties. If data points were closed to each other, then they would be grouped together in the same cluster, and those that were far away would be placed in another cluster. Once the training data were divided into groups, the machine learning models were trained for each group, and the trained models were saved individually for further use.

The STT, Gamma Ray (GR) and formation density were used as the input parameters for the UCS prediction machine learning modelling. An unsupervised learning method (K-means) automatically divided the training data into k groups, where the k referred to the number of centroids assigned in the dataset. A centroid was a location representing the centre of the group. Every data point was allocated to each of the groups by reducing the in-cluster sum of squares. There was generally an abundance of laboratory UCS testing available for Australian collieries that, in many instances, have been correlated with sonic logging from borehole exploration but not often with density and gamma log data. A python script has been developed to extract the STT, GR and density (input data) from the corresponding sample depth of the geophysical logs for the available laboratory UCS (output data). The integration of the input and output data were clustered into groups before performing the machine learning modelling. The result of the GML method showed improvement of the random forest prediction and a significant increase in the accuracy of the XGBoost prediction compared to the conventional non-grouping machine learning methods.

**DATA ANALYSIS AND CLEANING**

Inaccurate and incorrect data could greatly compromise the performance of any machine learning model. Data analysis and cleaning were therefore critically important before machine learning modelling. Some common data cleaning operations included identifying and removing the single appeared values, duplications, NaN (not a number) readings and etc. Many types of errors might exist in a dataset, which required not only the typical data cleaning operation in a standard statistical analysis but also sophisticated methods on many occasions. Domain knowledge might also be required to process the complex datasets of geotechnical and geophysical field and laboratory testing results.

Besides the general data analysis and cleaning operations, some common errors in the training dataset included:
- Tool failure: the downhole wireline logging or logging while drilling tool failure or malfunction might result in zero or extraordinarily high or low readings. This error usually was easy to identify. The missing log data for the intervals might also be predicted by machine learning methods (Zhong et al., 2020).

- Inconsistent unit: this happened when extracting the corresponding geophysical logging data for the laboratory UCS. For example, the unit of STT used in some old logging datasets (LAS files) was μs/ft; however, in the new logging dataset, it was μs/m. It might be identified by visualizing the training data. The best practice was to check the header of the LAS files to confirm if the consistent units were used.

- Invalid laboratory UCS test result: A length to diameter (L/D) ratio of between 2 and 3 was generally recommended for UCS tests with scientific and technical utilities (Hawkins, 1998; Güneyli and Rüşen, 2016). However, there were large numbers of specimens used for UCS laboratory tests. The UCS test result of the unqualified sample should be removed as long as the L/D information was available.

- Inaccurate geophysical logging data: This error was also known as the boundary effect for the logging data (Zhou and Esterle, 2007; Zhou and Esterle, 2008). It happened when the core sample was retrieved from the location closed to the formation boundary/interface. Previous researchers have studied this type of error and suggested that the samples should be located at the centre of uniform (flat) log segments rather than the depths where the log gradient was steep (McNally, 1987; Butel et al., 2014).

The boundary effect on the log readings of the formation closed to the coal seam was demonstrated in Figure 1. The samplings of the downhole logging tools were always from certain depth intervals, which was known as the vertical resolution. The typical vertical resolution of commonly used logging tools: STT (1.4 m), GR (0.3 m), density (0.37 m), neutron (0.61 m) could be found on the tool technical specification sheets (Weatherford, 2021). Therefore, the response of the logging tool for the transition of formation lithology was often in a gradient manner. The gradient transition intervals identified for GR, STT, and density logging data were about 0.4 - 0.45 m (Figure 1). The vertical resolution of the acoustic televiewer scanner (ATV) (< 5 cm) was much smaller than the conventional logging tools as mentioned above so that the ATV image log had a clear and sharp change for the formation boundary.

![Figure 1: The boundary effect on the log readings of the formation closed to the coal seam](image)

Once considering all possible errors, one could further use various data manipulation techniques to process the raw correlation data of UCS and geophysical logging readings. Figure 2 compared the raw data and the clean data across multiple correlation views.

There were 2,355 data points in the clean dataset. The contour of the data density was computed as shown in Figure 3, where the data were unevenly distributed in the STT-UCS cross-plot and heavily concentrated around which STT ≈ 300 μs/m and UCS ≈ 20 MPa.
The data were further subdivided into different lithology and mine sites categories, as demonstrated in Figures 4 and 5, respectively. However, the grouped data clusters heavily overlapped each other, indicating that neither the lithology facies nor the mine sites were good criteria to differentiate the rock types from the geotechnical or geomechanics perspective.
GROUP-BASED MACHINE LEARNING: MODELLING AND IMPLEMENTATION

1. Unsupervised Learning for Data Grouping
To overcome the difficulty of imbalanced data distribution, a Group-based Machine Learning (GML) method was proposed. The group was referred to as a cluster of data points aggregated together because of certain similarities. The K-means method was used to automatically divide the dataset into different groups, where the data points within the same group surrounded a centroid (Hamerly and Elkan, 2003). The number of the centroids of groups was pre-defined as the k value for the target dataset. An example of K-means clustering was demonstrated in Figure 6, where k = 3.
In the modelling stage, 4D data (STT, GR, density, and UCS) and 3D data (STT, GR, and density) were used to perform the automatically clustering. Three cases ($k = 3$, 5, and 7) were tested, as shown in Figure 7. Note that the neighbouring groups might overlap on one dimension, but they were separated in the 3D distribution.

It was found that the coordinates of the centroids had little difference for 4D and 3D data. Thus, 3D data were selected to compute the centroid coordinates and aggregate the dataset into different groups. Figure 8 displays the matrix of the data distribution for $k = 3$, 5, and 7.
2. Group-based Machine Learning Workflow

The machine learning models should be trained for each group that has been identified by unsupervised learning. These groups represented the rock types that had distinguished STT-GR-density properties which might reflect their geotechnical properties.

However, data quantity distribution in different groups might largely vary from one group to another because of the irregular distribution of the correlations of datasets (Figure 9).

For each group of training, several machine learning algorithms were used (step 3 in Figure 11). Random forest (RFR) and XGB had better performance than the support vector machine (SV), ANN, genetic algorithm (GA), polynomial regression (PR) and linear regression (LR).
Figure 10: (a) The train data vs all data; (b) train data were subdivided into 5 groups (k = 5)

A schematic of the GML workflow has been given in Figure 11. For example, 3 RFR sub-models were trained in the GML method for the case k = 3 since there were three groups. To implement the UCS prediction of the GML (k = 3), one needed to classify the group ID (1, 2, or 3) of each input data record (STT-GR-density on the same depth) firstly (step 5 in Figure 11). The group classification required comparing the target data point to each group centroid and selecting the nearest one as its group ID (Figure 6). Then the UCS values were predicted for each group by using the corresponding trained model. Finally, the predicted result from each group was integrated into the overall UCS prediction.

3. Implementation and Recommendation
The results of the GML methods (GML k = 3, 5, 7) for UCS prediction were shown in Figure 12 by following the workflow as in Figure 11. It was shown that the RFR algorithm had a stable performance for all three cases due to the random nature of the RFR algorithm.
On the other hand, it was also noticed that some datasets on the edge of the entire data cluster were hardly predicted accurately, even if the GML method was used. The more groups were used, the more complicated the GML model would be; however, more groups did not guarantee a better result than the GML with fewer groups. Notably, the RFR algorithm had the best performance when $k = 5$ amount all three GML methods. Table 1 has ranked machine learning methods used in the GML ($k = 5$) method by comparing various error metrics.

Table 1: The error metrics of the machine learning methods and the curve fitting method in the GML method ($k = 5$)

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFR</td>
<td>0.89</td>
<td>3.62</td>
<td>2.6</td>
</tr>
<tr>
<td>XGB</td>
<td>0.82</td>
<td>4.42</td>
<td>3.02</td>
</tr>
<tr>
<td>ANN</td>
<td>-0.04</td>
<td>8.95</td>
<td>6.49</td>
</tr>
<tr>
<td>SV</td>
<td>-0.16</td>
<td>9.01</td>
<td>6.37</td>
</tr>
<tr>
<td>Fitting</td>
<td>-0.41</td>
<td>9.3</td>
<td>6.69</td>
</tr>
</tbody>
</table>

The UCS predictions by conventional non-grouping machine learning methods (which could be deemed as GML $k = 1$) and the curve fitting method were reported in Figure 13 for reference. It was shown that most predicted results except the RFR tended to concentrate on the centre of the data cluster.

Figure 12: The prediction results of the GML ($GML\, k = 3, 5, 7$)

Figure 13: The prediction results of the conventional non-grouping machine learning methods

The performance of XGB had a significant improvement when using the GML method, especially with higher group numbers. For example, the $R^2$ error of XGB increased from 0.70 for GML ($k = 3$) to 0.82 for GML ($k = 5$) and 0.86 for GML ($k = 7$). More other error metrics were provided in Table 2.
Table 2: Comparison of the errors of XGB algorithm for conventional non-grouping method and the GML methods (GML k=3, 5, 7)

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Grouping</td>
<td>0.47</td>
<td>6.87</td>
<td>4.97</td>
<td></td>
</tr>
<tr>
<td>GP (k=3)</td>
<td>0.7</td>
<td>5.5</td>
<td>3.88</td>
<td></td>
</tr>
<tr>
<td>GP (k=5)</td>
<td>0.82</td>
<td>4.42</td>
<td>3.02</td>
<td>✓</td>
</tr>
<tr>
<td>GP (k=7)</td>
<td>0.86</td>
<td>3.98</td>
<td>2.71</td>
<td>✓</td>
</tr>
</tbody>
</table>

Therefore, the recommendation was to use the GML method (GML k = 5, 7) with RFR or XGB algorithms for the UCS prediction.

**DISCUSSION: INTERPRETABILITY OF THE MACHINE LEARNING PREDICTIONS**

Although the machine learning method was seen as a black-box model, the users were keen to find a solution to qualitatively interpret the prediction result. It was an important step to understand the underlying mechanics of the correlation and the potential pitfalls during the prediction. LIME (Local Interpretable Model-agnostic Explanations) has been developed by Ribeiro et al. (2016) to assist the developer and model user in understanding the model by perturbing the input data samples and understanding how the predictions changed accordingly.

![LIME Diagram](image)

**Figure 14: Interpret the black-box model by using LIME**

LIME assumed a black-box machine learning model and investigated the relationship between input and output, represented by the model (Figure 14). It attempted to explain predictions for each data sample and evaluate the importance of each input parameter.

![Importance Score](image)

**Figure 15: The importance score of each input parameter for the UCS prediction result**

In one example of UCS prediction using machine learning (Figure 15), the importance score of each input parameter (STT, GR, and density) was computed for the particular predicted UCS value of 29.17 MPa, where the overall predicted UCS ranged from 4.35 to 77.98 MPa. The sign of the importance score was referred to the direction of the correlation. For example, the importance score of STT meant that the higher STT was, the lower UCS would be. For a similar principle, the higher the density, the higher the UCS since the importance score for density was positive. The absolute values of the importance score indicated the impact of that input on the prediction result. LIME has been demonstrated to be a great tool to explain what machine learning algorithms (or models) were doing, which helped us to gain insight into machine learning modelling.
CONCLUSION

Machine learning has been widely used in UCS prediction. The optimization of the machine learning performance relied on the quantity of the training data and data quality. In this paper, a novel Group-based Machine Learning (GML) has been developed for the UCS prediction. Main findings from this study included:

- The data analysis and cleaning process for the geophysical logging data and UCS laboratory data were discussed in this paper. Some common types of error data were suggested.

- The GML was proposed. The data points were classified into groups by using the unsupervised learning (K-means) method. Then the machine learning regression models were built for each group. A workflow of the GML modelling and implementation was developed.

- The GML showed better performance than the conventional non-grouping machine learning models. The GML with intermediate group number (i.e. k = 5 in this paper) might perform better than the low or high k values (i.e. k = 3 or 7).

The GML method might vary depending on the ratio of the train/test data splitting. An optimization process was required to achieve the best model. In sum, the GML method provided a new idea for the machine learning modelling on the data with the imbalanced distribution. The technique developed in this paper might also be used in other kinds of machine learning prediction tasks.

ACKNOWLEDGEMENTS

The authors would like to acknowledge technical and financial support from ACARP and the industry sponsors (Anglo American, Glencore, and BHP).

REFERENCE


