Application of Machine Learning in Prediction of Lithofacies from Well Logs

Yuxiang Tian[1], Alfred Renaud[#] and Guillermo Goldsztein[#]

[1]Clements High School, USA
[#]Advisor

ABSTRACT

As humanity's reliance on mineral resources continuously grows, new technologies need to be implemented in the mining industry to fulfill this demand. A method of using well logging data to predict rock types in the surrounding area using a multiclass classification neural network is discussed as a potential way to increase efficiency. The model achieved an accuracy rate much higher than would be possible through guessing, 70% as compared to 11%, demonstrating the effectiveness of such technology. Potential ways this model can be applied and improved were also discussed.

Introduction

As computers become increasingly more powerful and accessible, their use can gradually begin to spread to various industries previously dominated by human labor, such as the mining industry. Mining is at the center of human technological development (Coates, 1985), and its advancement is essential to the progression of human civilization. With the problem of climate change looming closer and closer, a focus on this industry is now more important than ever. Electric vehicles can use up to six times as many mineral resources as conventional fossil fuel-powered vehicles, and renewable energy technologies, such as wind turbines, can use up to nine times as many mineral resources as a traditional natural gas power plant that generates the same amount of energy (Iea, n.d.). This increased usage drives up demand for certain critical minerals, such as copper, lithium, and nickel, among others, leading to an increase in mining activity.

This increase in demand requires mining companies to pursue more efficient and effective methods of detecting mineral deposits, such as well logging. Well logging is a technology that, though most prominently used in the petroleum industry, has seen extensive use in the mining industry as well. This technology first requires a borehole to be drilled, and then an array of instruments is inserted to measure resistivity, gamma-ray emission, and other properties of the surrounding rock (Encyclopædia Britannica, n.d.). This process produces data that can be used to determine surrounding rock types and mineral concentrations. Well logging allows miners to pinpoint those areas that have the highest mineral concentrations, increasing yield and efficiency, and decreasing the environmental impact of larger mines.

In this article, a neural network is created to interpret this information and determine the surrounding lithofacies, or the specific type of rock that is present in that area through the use of multiclass classification. Neural networks are computer programs designed to mimic how biological neurons function and operate, by first learning the data through the use of a training data set. The model can then be used to predict unknown data by providing incomplete input data (IBM, n.d.). Multiclass classification is a task that neural networks are able to perform, classifying data into multiple separate categories (Likebupt, n.d.). Such a neural network is utilized in the present paper to determine the type of rock that would be present when given a set of data produced through various well-logging tools.

Methods

The data used to train this model comes from Kaggle, a public dataset-sharing website, and consists of 3231 rows. The data within the dataset comes from a series of nine natural gas wells located in Kansas. Although the main goal of this paper is to prove new technologies that can be utilized in mining, the well-logging data from both the mining and petroleum industries can be used interchangeably when developing a neural network as they both provide the same types of information. The dataset provides eight features, or variables, relevant to the prediction of surrounding lithofacies. The features provided and their meanings are as follows:

1. Depth: depth of the tools within the well in meters
2. GR: a measurement of the gamma-ray emission within the surrounding rock in API (American Petroleum Institute) units
3. ILD_log10: a measurement of the electrical resistivity of the surrounding rock, in ohm-meters (Ω·m) transformed to a base-10 logarithmic scale
4. DeltaPHI: porosity index of the surrounding rock
5. PHIND: an average of the neutron and density logs of the surrounding rock
6. PE: a log of the photoelectric absorption factor
7. NM_M: a binary variable indicating whether the region is of marine or nonmarine origin
8. RELPOS: relative position

The dataset also includes nine possible facies, which are included within the dataset as the numerical values 1 through 9, designated as the following:

1. SS: Nonmarine sandstone
2. CSiS: Nonmarine coarse siltstone
3. FSiS: Nonmarine fine siltstone
4. SiSH: Marine siltstone and shale
5. MS: Mudstone
6. WS: Wackestone
7. D: Dolomite
8. PS: Packstone-grainstone
9. BS: Phylloid-algal bafflestone

Facies	Formation	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	NM_M	RELPOS
3	A1 SH	SHRIMPLIN	2793.0	77.45	0.664	9.9	11.915	4.6	1	1.000
3	A1 SH	SHRIMPLIN	2793.5	78.26	0.661	14.2	12.565	4.1	1	0.979
3	A1 SH	SHRIMPLIN	2794.0	79.05	0.658	14.8	13.050	3.6	1	0.957
3	A1 SH	SHRIMPLIN	2794.5	86.10	0.655	13.9	13.115	3.5	1	0.936
3	A1 SH	SHRIMPLIN	2795.0	74.58	0.647	13.5	13.300	3.4	1	0.915

Figure 1. First five lines of the dataset (Meintanis, 2020)

A neural network was implemented that utilized multiclass classification to take in the eight features and output the most likely lithofacies that could correspond with the input data. The current state of this data is unsuitable for such a model, however. The first step to refining the dataset was deleting the unnecessary columns, namely the columns “Formation” and “Well Name” as these pieces of information are not at all relevant to the types of rock present in an area.

Certain columns also displayed information in a way that could confound the neural network and needed to be modified. The column “NM_M” consisted of the values 1 and 2, with 1 corresponding to non-marine and 2 corresponding to marine. Because this section is completely binary, it was renamed to “marine”, 1 was replaced with 0, and 2 was replaced with 1. 0 would therefore mean “false” and 1 would mean “true.”

A process called One-Hot Encoding was then performed on the facies column, which is used to convert categorical data to numerical data (GeeksForGeeks, 2023a). Although the facies column may appear numerical at first, it is important to understand that each number is representative of an individual rock type. In other words, having words instead of numbers in that column would make practically no difference. Because neural networks cannot work with this type of data, nine new columns were created, each labeled with its respective lithofacies and filled with 0. For each row, a 1 would then replace the 0 in the column that corresponded with the value under the facies column. The facies column was then deleted as it was no longer useful.

Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	RELPOS	Marine
2793.0	77.45	0.664	9.900	11.915	4.600	1.000	0
2793.5	78.26	0.661	14.200	12.565	4.100	0.979	0
2794.0	79.05	0.658	14.800	13.050	3.600	0.957	0
2794.5	86.10	0.655	13.900	13.115	3.500	0.936	0
2795.0	74.58	0.647	13.500	13.300	3.400	0.915	0

Figure 2. First five lines of the first half of the dataset post-modification.

SS	CSiS	FSiS	SiSH	MS	WS	D	PS	DS
0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0

Figure 3. First five lines of the second half of the dataset post-modification.

The data was then split into two separate sets, one containing the nine lithofacies, and the other containing all the data that could be used to find said lithofacies. These two sets were then further split into training and validation sets. The training sets would be used to train the model and accounted for approximately 75%

of the data. The validation sets were used to validate the results of the model, displaying errors and allowing for the adjustment of hyperparameters.

The data was then scaled, which is a process in which the features in a dataset are altered so that their values are close to each other. This enhances the neural network's learning abilities as these programs tend to add a greater bias to larger numbers (GeeksForGeeks, 2023b). Because the values in this dataset range from less than 1 to several thousand, feature scaling is essential to ensure the model weighs each number equally.

Results

A neural network was created with two hidden layers, each with 20 nodes, and utilized the rectified linear unit (ReLU) activation function, essential for a multiclass classification program such as this one. The number of hidden layers and nodes was determined through extensive trial and error, where higher accuracy percentages in the validation set were deemed optimal. This was necessary to prevent overfitting, a phenomenon where a model is trained so well on its training set that it performs poorly on anything else. This is highly suboptimal for the purposes of this program. The epochs was set at 500, a value that provided both good runtime and good accuracy.

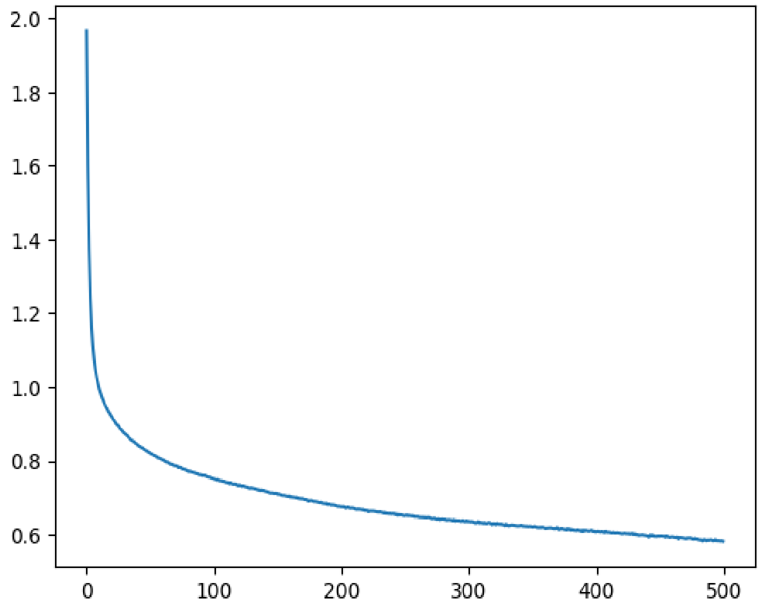

Figure 4. Loss curve of the function, with loss on the y-axis and epochs on the x-axis.

After running the program, the accuracy percentages were output, allowing for the adjustment of the hyperparameters, such as the nodes in each hidden layer and the number of hidden layers, optimizing the program. The final result achieved with this program was an 85% accuracy rate in the validation set and a 70% accuracy rate in the training set. This does suggest some overfitting; however, this overfitting does not compromise the accuracy rates of the training set and was therefore deemed necessary. This level of accuracy is a massive improvement over mere estimation, which would result in an 11% accuracy rate.

Figure 5. Table comparing the neural network approach with the estimation approach in terms of accuracy in both training and validation sets.

	Neural Network Approach	Estimation Approach
Training Set	0.85	0.11
Validation Set	0.70	0.11

Discussion

In this study, the process of creating a multiclass classification neural network focused on predicting lithofacies based on well-logging data was discussed. This included the process of optimizing the dataset for such a neural network, as well as the construction process of the actual network itself. The accuracy rate this program was able to achieve was several times higher than what could have been achieved through guessing, demonstrating a successful algorithm.

Limitations

This study could be better and could benefit greatly from more advanced techniques and a more comprehensive range of data from various localities. In order for such a system to function more effectively in the field, it should also include more lithofacies and ore mineral percentages in said lithofacies in order to be better suited for the mining industry. Ultimately, this study demonstrated the possibilities of such a technology and how it could be applied to a field typically not associated with technological advancement in order to increase both energy and production efficiency.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Coates, D. R. (1985). Geology and Society. In Geology and Society. essay, Chapman and Hall.

Encyclopædia Britannica, inc. (n.d.). *Core sampling*. Encyclopædia Britannica. https://www.britannica.com/technology/core-sampling

GeeksforGeeks. (2023a, July 18). *Feature engineering: Scaling, normalization, and standardization.* GeeksforGeeks. https://www.geeksforgeeks.org/ml-feature-scaling-part-2/

GeeksforGeeks. (2023b, April 18). *One hot encoding in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/

Iea. (n.d.). *Executive summary – the role of critical minerals in Clean Energy Transitions – analysis*. IEA. https://www.iea.org/reports/the-role-of-critical-minerals-in-clean-energy-transitions/executive-summary

Likebupt. (n.d.). *Multiclass neural network: Component reference - azure machine learning*. Multiclass Neural Network: Component Reference - Azure Machine Learning | Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/multiclass-neural-network?view=azureml-api-2

Meintanis, Ioannis. (2020, August 15). *Well log facies dataset*. Kaggle. https://www.kaggle.com/datasets/imeintanis/well-log-facies-dataset?resource=download

What are neural networks?. IBM. (n.d.). https://www.ibm.com/topics/neural-networks