Comparing Machine Learning Methods for Predicting Seismic P-wave Velocity on Global Scale

1,2[0000Š 0002Š 0007Š 630X]
2[0000Š 0001Š 6625Š 4335]
1[0000Š 0001Š 5055Š 0180]
1[0000Š 0003Š 0323Š 8578]

1
2

Abstract.

Keywords: . .

1 Introduction

2 Prediction of Seismic P-wave Velocities

P-wave velocity is one of the physical properties of sub-surface rocks which can help to predict the materials beneath the ocean ”oor. Based on measurements done during di erent drilling projects Dumke et al. [5] gathered the information of 333 borehole logs to form a dataset which is depicted in Figure 1. They have compared the results of prediction using Random Decision Forests (RDF) [2] with hamiltonian functions [6], which were used in the past as a conventional method to compute the average P-wave velocity.

2.1 Characteristics of the Dataset

Our data have been gathered during di erent drilling campaigns performed on a global scale, see Figure 1. The dataset is divided into 10 folds to perform cross validation. The 10 folds are separated from each other based on geographic location. This is an important feature of the dataset to prevent over“tting. Location-wise separation implies that the samples which have been used in training sessions belong to di erent locations than the samples used for the prediction.

We compared the results of our predictions on this dataset. Some key characteristics of this dataset are that,

1. the number of categories for independent variables which are usually referred to as features is large,
2. the amount data with respect to the number of categories on each sample is scarce for each borehole,
3. the trend in data di ers in each correlation of data pairs, and
4. there are many outliers and noise in the dataset.

thickness in this pair plot. Many outliers can be detected. In addition, in each pair of the variables there is clear change in the trend of data distribution.

Fig. 2: Dataset pair plot

2.2 Investigated Machine Learning Methods

To achieve improved results, we have done prediction on seismic p-wave velocity using di erent machine learning methods to assess the changes in prediction accuracy. We have used scikit-learn [9], and keras [1] to implement 3 di erent machine learning methods:

- Support vector regression (SVR): The kernel used in SVR [4] is a radial basis function [11] of degree 8.

- Polynomial regression: In polynomial regression, the degree of the polynomial is set to 2.
- Neural networks (NN): We applied a 2-layer feed forward neural network, each layer consists of 32 neurons. As activation function we used the relu function, which is a widely used activation function for neural networks. In each training iteration, a batch of 50 data samples goes through the model. The training session contained 400 epochs and the learning rate was set to 0.001. We used the RMSProp [3] optimizer for our model.

Dumke et al [5] used a Random Forest Regressor using scikit-learn [9]. They used 1000 decision trees in their study. The number of the predictor variables was 38. Based on their study, using the most important 16 features results in better predictions.

2.3 Prediction Results

We have selected the most important 16 features of the dataset as done before [5]. The importance of the features are calculated using RDF. Feature importance can be de“ned as a score assigned to each category of independent variables with respect to their in”uence on the performance of the prediction model. Thus, the performance of the prediction is divided into 4 categories based on the previous work by Dumke et al [5]. Each category speci“es the level of performance

attempts we are up to investigate methods for data augmentation in our “eld of marine data science.

For reproducibility and reusability, we publish our software open source [10] and the data together with the analytics service OceanTEA [7], as we did in the past [8].

Acknowledgment

The “rst author is funded through the Helmholtz School for Marine Data Science (MarDATA, https : //www.mardata.de/), Grant No. HIDSS-0005.

References

1. François Chollet et al. Keras: The Python deep learning library. ascl, pages ascl… 1806, 2018.
2. Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests for classi“cation, regression, density estimation, manifold learning and semi-supervised learning. Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114, 5(6):12, 2011.