

Corrosion Risk Prediction Based on Offshore Pipeline Service Database and Machine Learning Algorithms

Zhihao Qu¹, Xiaoqi Yue¹, Yi Li², Xiaqiao Li¹, Lihua Hu², Yue Ren¹, Yao Lv¹, Lei Zhang^{1*}

¹Institute for Advanced Materials and Technology, University of Science and Technology
Beijing, Beijing, China

²CNOOC Research Institute, Beijing, China

ABSTRACT

Pipelines are an important part of offshore oil and gas field development facilities and the main means of gathering and transporting offshore oil and gas resources. However, pipelines are subject to deterioration and degradation in the corrosion media. Corrosion risk assessment and prediction is an effective way to avoid leakage of oil and gas field pipelines and facilities, ensure safe operation and save cost. Hence, in this study, a machine learning model with excellent predictive performance was constructed for corrosion rate, to provide an effective mean for processing complex corrosion data and to provide a useful tool for further exploration of submarine pipeline corrosion problems. Meanwhile, a method that can effectively and quickly evaluate the accuracy of corrosion rate prediction model was explored, which can be used as a reference to select the most appropriate and accurate Machine Learning (ML) model based on existing data.

Keywords: Machine Learning, Statistical Indicator, Corrosion, Database, Global Performance Indicator.

INTRODUCTION

Pipelines have been the main transportation pattern of oil and gas because of their safety and economy,¹ which are considered as the lifeline of offshore oil and gas transportation. With the booming development of offshore oil industry, the frequency of pipeline leakage is also increasing.^{2,3} Corrosion is one of the important factors due to some characteristics such as operating environment, service life and transportation medium, etc.,⁴ which damages the integrity of the pipeline and damage the normal operation of pipelines.⁵ Furthermore, leakage accidents caused by pipeline corrosion have occurred all over the world, accounting for 70~90% of total accidents,⁶ which has caused huge economy losses and catastrophic environmental damage. Therefore, it is necessary to predict and evaluate the corrosion degree of subsea pipelines with the available corrosion monitoring data to reduce possible accidents resulting from pipeline corrosion.⁷

In recent years, extensive work has been promoted on pipelines corrosion and a variety of pipeline corrosion prediction models have been established. However, corrosion is a rather complex process, in which there are many factors involved, and there may be relationships among them. As a result, it is difficult for traditional empirical

and semi-empirical models to mirror the interactions among various corrosion factors. Besides, they have limitations as they don't take into account the stochastic nature of the corrosion process.⁸

With the development of computer technology, a number of Machine Learning (ML) algorithm models have been developed and the application of ML in risk prediction and assessment of pipelines has been widely used by researchers due to its advantages in dealing with multi-dimensional, nonlinear and uncertain characteristics.⁹ The ML algorithm model can learn the relationship between inputs and outputs from training data set, and then predict the outputs of any unknown inputs.¹⁰ Thus, using ML algorithm technology, it is possible to capture the interaction between corrosion factors and their impacts on corrosion in the overall sense.

Ren et al.¹¹ applied Back Propagation (BP) neural network to predict the corrosion rate of long-distance natural gas pipelines. A prediction model for the natural gas pipeline internal corrosion rate is established with natural gas pipeline mileage, elevation difference, pipe inclination, pressure, Reynolds number used as an input parameter and the maximum average corrosion rate of the pipeline used as an output parameter. The results elucidate that BP neural network possesses fitting accuracy and forecasting result. Hatami et al.¹² constructed a hybrid model based on Least Square Support Vector Machine (LSSVM) to predict CO₂ corrosion, and Coupled Simulated Annealing (CSA) optimization technique was adapted to optimize the model. The Gaussian Kernel functions were proposed as the desired function which can provide high precision for predicting CO₂ corrosion in oil and gas industries. Ossai¹³ established the time-varying corrosion defect depth growth of corroded pipelines with the historical operating parameters based on machine learning, which is depend on feed-forward Subspace Clustered Neural Network (SSCN) and Particle Swarm Optimization (PSO). Zhi et al.¹⁴ used a hybrid method combining random forest (RF) and Spearman correlation analyses to reduce the dimensionality of the dataset prior to support vector regression (SVR) modelling. Compared to dimensionality reduction by maximal information coefficient (MIC) and principal component analysis (PCA), the SVR model which is reduced dimensionality by RF presented higher accuracy.

Based on the above, ML algorithm models have been widely used to predict corrosion. However, the studies about applications on subsea pipeline corrosion forecast have rarely been reported. Therefore, this study aims to develop a ML model to predict submarine pipeline corrosion and to provide a method that can effectively and quickly evaluate the accuracy of corrosion rate prediction model.

ML MODEL

Support Vector Machine (SVM)

The SVM is considered as a ML model that has the advantages of global optimization, nonlinearity and generalization, as proposed by Vapnik et al. in 1995.¹⁵ By introducing kernel function, the SVM has excellent performance on small sample data set, and

shows certain development prospect in practical engineering field. The basic principle of the SVM is to establish the optimal hyperplane in space by using the principle of structure minimization to obtain the maximum classification interval, so that the classifier can get global optimization in the whole sample space.

Based on the principle of structural risk minimization, the SVM usually adopts the following model to determine the regression function and formulated as below:

$$\min \frac{1}{2} \omega^T \omega + C \sum_{i=1}^l \xi_i + C \sum_{i=1}^l \xi_i^* \quad (1)$$

$$s. t. \begin{cases} \omega^T \phi(x_i) + b - y_i \leq \varepsilon + \xi_i \\ y_i - \omega^T \phi(x_i) - b \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, (i = 1, 2, \dots, l) \end{cases} \quad (2)$$

Where C represents the regularization constant, and l represents the number of samples; ω is the normal vector, which determines the direction of the hyperplane; b is the displacement term, which determines the distance between the hyperplane and the origin; ξ_i and ξ_i^* are the slack variables, and ε is insensitive loss function.

By introducing Lagrange multipliers α_i and α_i^* , the above equation is transformed into the following dual optimization problem:

$$\min \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \quad (3)$$

$$s. t. \begin{cases} e^T (\alpha - \alpha^*) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C, (i = 1, 2, \dots, l) \end{cases} \quad (4)$$

The regression function of the SVM can be obtained by solving the above problems:

$$F(x) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) K(x_i, x) + b \quad (5)$$

$K(x_i, x)$ represents the SVM kernel functions. In this paper, SVM based on Gaussian Kernel is selected to establish corrosion prediction model.

K-Nearest Neighbors (KNN)

The KNN is a basic classification and regression method proposed by Cover and Hart in 1968.¹⁶ The working mechanism of KNN algorithm is to find the k training samples closest to a test sample in the training set based on some distance measurement, and then assign the average value of some attributes of the k "neighbors" to the test sample, so as to realize the prediction of the corresponding attributes of the sample. The key of KNN algorithm is to determine the number of "neighbors" k and the distance measurement method. The commonly used distance measurement formula is as follows:

$$D_1(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (6)$$

Where $D_1(x, y)$ represents the Euclidean distance. In most cases, the Euclidean distance can satisfy the situational requirements without considering other methods.

$$D_2(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (7)$$

$$D_3(x, y) = \sqrt[p]{\sum_{i=1}^n (|x_i - y_i|)^p} \quad (8)$$

Where $D_2(x, y)$ represents the Manhattan distance; Where $D_3(x, y)$ represents the Chebyshev distance.

Random Forest (RF)

RF algorithm is a fast and flexible machine learning algorithm for mining high-dimensional data, which was proposed in 2001¹⁷. Its principle is explained as follows: Based on bootstrap resampling technology, n samples are repeatedly and randomly selected from the original training sample set, so as to generate a new training sample set to train the decision tree. Then K decision trees are generated to form a random forest according to the above steps. The prediction results of new data are determined by the voting scores of all decision trees. The RF algorithm is an improvement of decision tree algorithm, which has good practical performance and high dimensional data set processing ability.

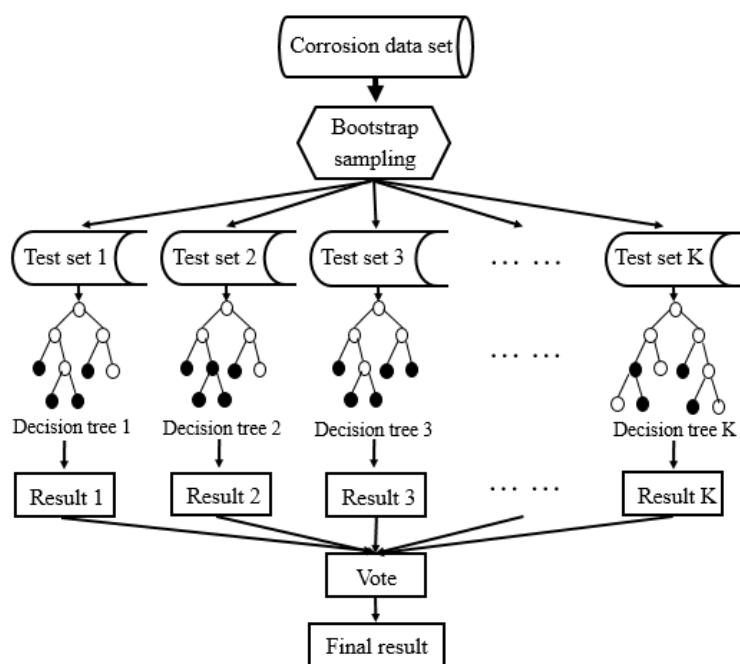


Figure 1: Schematic diagram of the RF model

Extremely randomized Tree (ETR)

Extremely randomized tree is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its prediction result. Each decision tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index)¹⁸. This random sample of features leads to the creation of multiple de-correlated decision trees. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Bootstrap aggregating (BAG)

Bootstrap aggregating is a technique that reduces generalization error by combining multiple models. This algorithm can improve the stability and accuracy of the statistical classifiers and regressors, it can also help the model avoid overfitting.¹⁹ The principle is to train several different models individually, and then get multiple models to vote on the output of the test set. This is a commonly applied strategy in machine learning called model averaging. The technique using this strategy is called ensemble methods (integration methods). It is worth mentioning that dropout is actually a subtle method to achieve bagging at low cost, building different exponentially level models by randomly inactivating some good elements. The difference is that bagging each model is independent, and dropout shares a portion of the parameters between models to reduce the memory required for storage exponential level models.

Multi-Layer Perceptron (MLP)

MLP, also known as Artificial Neural Network, has been a hot topic in Artificial intelligence research since 1980s. Artificial neural network consists of input layer, output layer and hidden layer.²⁰ The input layer is responsible for receiving signals and data from the external world, the output layer realizes the output of system processing results, and the hidden layer is located between the input layer and the output layer.

A perceptron has one or more inputs, bias, activation functions, and a single output. The perceptron takes inputs, multiplies them by some weight, and then passes them to the activation function to produce the output. There are many activation functions to choose from, such as sigmoid function, tanh function, relu function, and so on. The activation function selected in this study was relu function.