Academic Journal of Emerging Technologies
ISSN Print: 3104-4417
ISSN Online: 3104-4425
CD ISSN: 3104-4433
https://doi.org/10.63313/AJET.9030

A Pipe Sticking Prediction Method Based on Random Forest

Jiadong Yang
School of Mechanical Engineering, Xi'an Shiyou University, Xi'an, Shaanxi, 710065, China
Email: yjdd201770@163.com

How to cite this paper: Yang, J. D. (2025). A Pipe Sticking Prediction Method Based on Random Forest. *Academic Journal of Emerging Technologies*, 2(1), 52–59. ISSN Print: 3104-4417; ISSN Online: 3104-4425. https://doi.org/10.63313/AJET.9030
Published: 2025-12-31

Abstract

This study aims to address the challenge of predicting pipe sticking accidents in petroleum drilling engineering. In response to the high blindness of tradi-tional empirical qualitative operations, the difficulty of mathematical models in accurately reflecting drilling laws, and their poor operability, machine learning methods such as neural networks (NNs), support vector machines (SVM), and random forests are introduced for pipe sticking prediction re-search. Adhering to the principles of sample representativeness, category balance, and diversity, 9 drilling parameters including depth, drilling speed, weight on bit (WOB), torque, friction coefficient, outlet temperature, rota-tional speed, inlet flow rate, and outlet flow rate were selected. Drilling data from 30 wells in the Yan'an area were collected. Measurement data within a certain period before pipe sticking were used as pipe sticking samples, which were cross-arranged with normal samples to construct the training dataset. By comparing the performance of SVM (linear kernel, radial basis kernel) and random forest models, the results show that the random forest performs the best. In its confusion matrix, the number of correctly predicted samples for true labels 0 and 1 is 190 and 207 respectively, with only 9 and 5 misclassifi-cations. The accuracy rate reaches 0.966. In summary, this study confirms that the random forest has significant advantages in balancing "accuracy and efficiency" in pipe sticking prediction tasks, providing a reliable early warning model and technical support for safe drilling operations.

Keywords

Random Forest; Stuck Pipe; Petroleum Machinery

1. Introduction

Petroleum and natural gas are the world's most important energy sources and chemical raw materials, and also crucial productive factors and strategic resources in the economic development of contemporary society. The petroleum industry is mainly engaged in the exploration, development, and processing of petroleum and natural gas, and drilling engineering is the primary means for their exploration and

development. During drilling operations, the phenomenon where drilling tools get stuck in the wellbore and cannot move freely due to various reasons is called pipe sticking, such as sand bridge sticking, adhesion sticking, keyway sticking, collapse sticking, dry drilling sticking, hole shrinkage sticking, mud packing sticking, and falling object sticking (1). Although extensive research has been conducted on the handling of pipe sticking accidents at home and abroad, and many treatment methods have been proposed, on-site operations still mainly rely on experience to provide qualitative operating rules to avoid pipe sticking. This cannot eliminate the blindness of operations, nor can it fundamentally reduce the occurrence of pipe sticking accidents or achieve timely prediction of such accidents.

To prevent pipe sticking accidents, technicians have applied various mathematical models to pipe sticking research, which have played a significant role in improving and optimizing drilling processes. However, due to the wide range of fields involved in drilling work, it is difficult for mathematical models to accurately reflect the laws in the drilling process. Meanwhile, mathematical models have shortcomings such as difficult coefficient calculation and uncertain processes, making them inconvenient for practical operation. Therefore, pattern recognition methods need to be adopted to analyze and predict pipe sticking accidents (2).

Neural Networks (NNs) are a pattern recognition technology that simulates biological neural networks for information processing (3). They possess adaptive and self-organizing capabilities, and can perform associative memory, nonlinear, and distributed processing. Neural networks can train and learn from information in different states, making them very suitable for establishing accident prediction models. More importantly, they avoid the difficulties in data analysis and modeling caused by precise mathematical models. Therefore, neural networks can accurately capture the precursor characteristics of pipe sticking by mining the complex nonlinear correlations between drilling parameters, improve the real-time performance and accuracy of pipe sticking prediction, and provide reliable early warnings for safe drilling operations.

2. Model Introduction

2.1. Introduction to Support Vector Machine Algorithm

Support Vector Machine (SVM) was first proposed by Cortes and Vapnik in 1995. It has many unique advantages in solving small-sample, nonlinear, and high-dimensional pattern recognition problems, and can be extended to other machine learning tasks such as function fitting.

The SVM method is based on statistical learning theory's VC dimension and the principle of structural risk minimization. It seeks the optimal trade-off between model complexity and learning ability based on limited sample information to achieve the best generalization ability. The VC dimension reflects the learning ability of a function set; generally speaking, the higher the VC dimension, the more complex

the learning machine.

2.1.1. Linear Classifier

Linear classifiers are the simplest and most effective form of classifiers, and SVM has been developed from the optimal hyperplane in linearly separable cases.

2.1.1.1. Linear Separability

Samples are linearly separable if a linear function can separate them completely and correctly; otherwise, they are non-linearly separable.

2.1.1.2. Optimal Hyperplane

Like **Figure 1**, the square and circular points represent two types of samples. H is the classification line, and H1 and H2 are lines passing through the closest samples of each class and parallel to the classification line. The distance between them, denoted as margin, is called the classification margin. The classification line H is called the optimal classification line if it not only correctly separates the two classes but also maximizes the classification margin. Maximizing the classification margin is essentially a control over generalization ability, which is one of the core ideas of SVM.

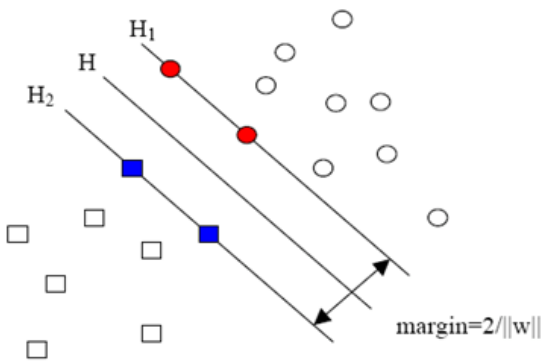

Figure 1. Optimal hyperplane

2.2. Introduction to Random Forest Algorithm

In machine learning, a random forest is a classifier consisting of multiple decision trees, and its output class is determined by the mode of the classes output by individual trees. The base classifiers composing the random forest are called decision trees. The algorithm for inferring random forests was developed by Leo Breiman and Adele Cutler.

2.2.1. Decision Tree Algorithm

Like **Figure 2**, a decision tree can be regarded as a tree-like prediction model, which is a hierarchical structure composed of nodes and directed edges. The tree contains

three types of nodes: root node, internal nodes, and terminal nodes (leaf nodes). A decision tree has only one root node, which is the combination of the entire training set. Each internal node in the tree represents a splitting problem, which partitions the samples arriving at the node according to a specific attribute, dividing the dataset into two or more subsets. Each terminal node (leaf node) is a dataset with a classification label, and each path from the root node to a leaf node of the decision tree forms a class. There are many decision tree algorithms, such as ID3 and CART. These algorithms all adopt a top-down greedy approach: each internal node selects the attribute with the best classification effect for splitting, which can be divided into two or more child nodes. This process continues until the decision tree can accurately classify all training data or all attributes are used up.

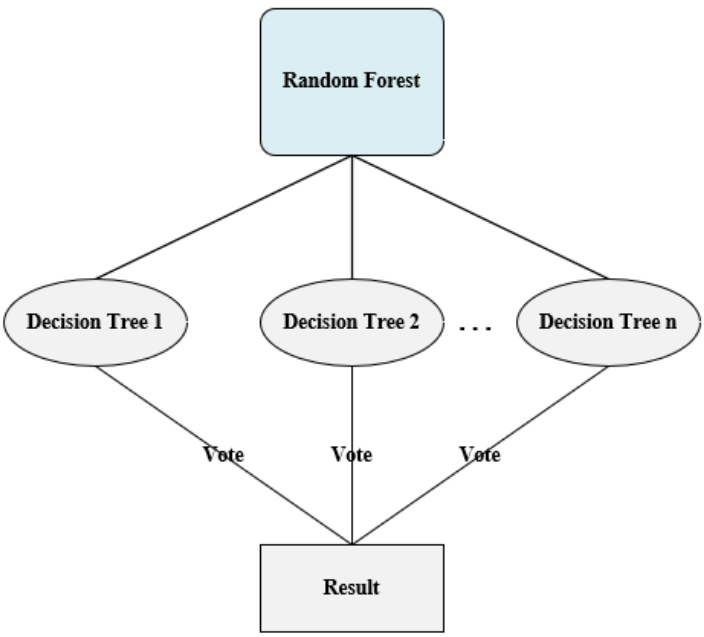

Figure 2. Decision Tree Algorithm

2.2.2. Advantages of Decision Trees

The decision tree method does not require assuming the distribution of prior probabilities. This non-parametric feature gives it better flexibility and robustness. The decision trees or production rules generated by the decision tree method have the characteristics of simple and intuitive structure, easy understanding, and high computational efficiency. The decision tree method can effectively suppress noise in training samples and solve the problem of missing attributes, thus preventing the reduction in accuracy caused by noise in training samples and missing data.

2.2.3. Advantages of Random Forest Algorithm

It can generate highly accurate classifiers for various types of data; It can handle a

large number of input variables; When building the forest, it can internally generate an unbiased estimate of the generalization error; It includes an effective method for estimating missing data, and can maintain accuracy even if a large portion of the data is missing.

3. Dataset Construction

The laws extracted during network training are implied in the samples, so the selected samples must be representative. The balance and diversity of sample categories are two principles for sample selection and organization, meaning that the types of samples should be comprehensive and the number of samples in each category should be basically the same. If samples of a certain category are missing in the network, the network will fail to recognize samples of that category after training. If the number of samples in the network is not balanced, it may lead to over-training of the network on categories with a large number of samples. There are two methods for organizing samples: one is to randomly select samples from the training set for input; the other is to alternately input samples of different categories. The purpose is to avoid the centralized input of samples of the same category, where the matching mapping relationship is trained intensively. When samples of different categories are input, the adjustment of weights tends to the new mapping relationship, which may negate the training results of the previous category. Based on the above considerations, before network training, this study first cross-arranges pipe sticking samples and normal operation samples, and then inputs them into the network for training. The parameters selected in this study are: depth, drilling speed, weight on bit, torque, friction coefficient, outlet temperature, rotational speed, inlet flow rate, and outlet flow rate. Finally, the function to be realized in this study is to predict pipe sticking accidents in advance, so the pipe sticking samples in the training data must be measurement data within a certain period before pipe sticking. In accordance with the above principles, this study collects drilling data from 30 wells in the Yan'an area as the training samples for the pipe sticking prediction network.

3.1. Model Results

The confusion matrix results are shown in the following figures 3; figures 4; figures 5:

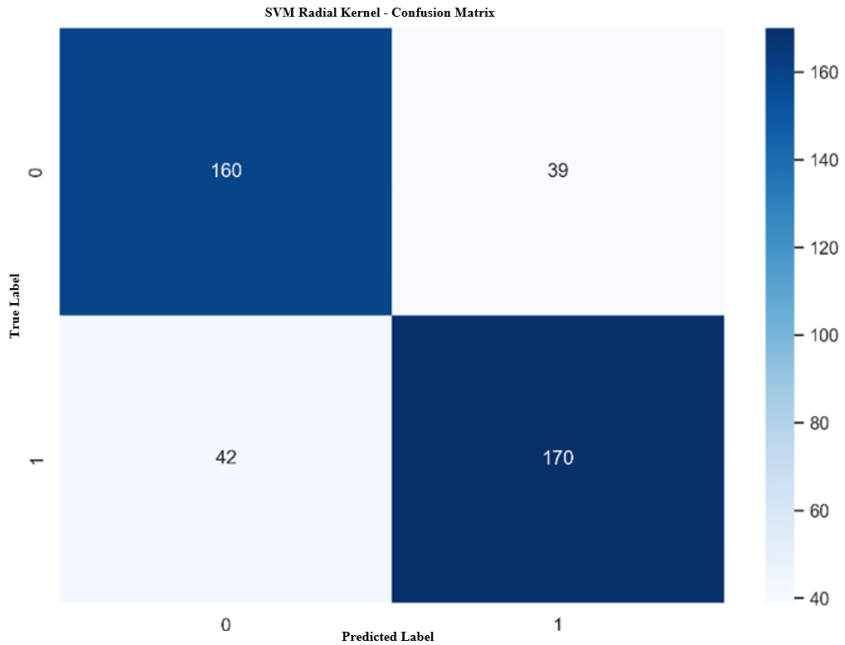

Figure 3. SVM Radial Kernel - Confusion Matrix

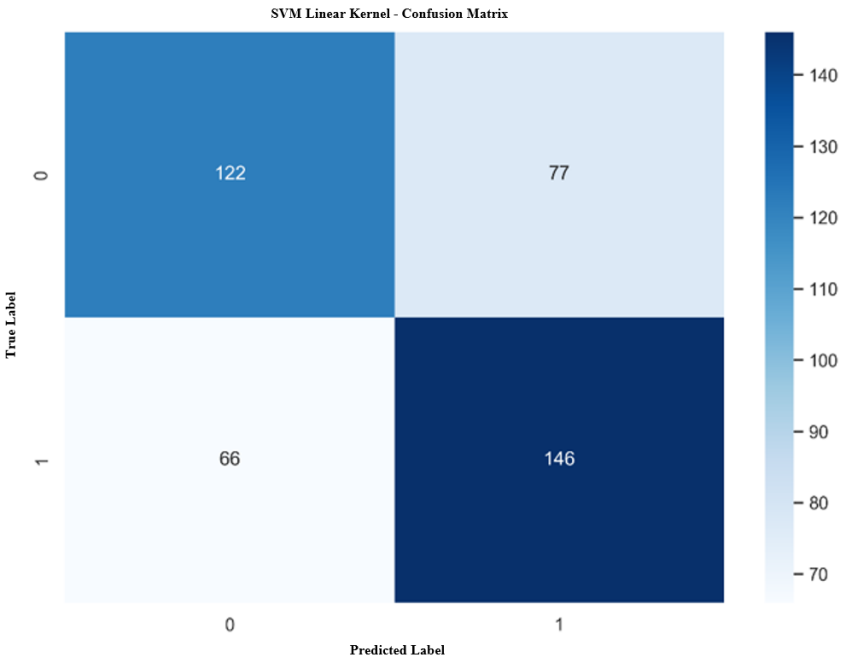

Figure 4. SVM Linear Kernel - Confusion Matrix

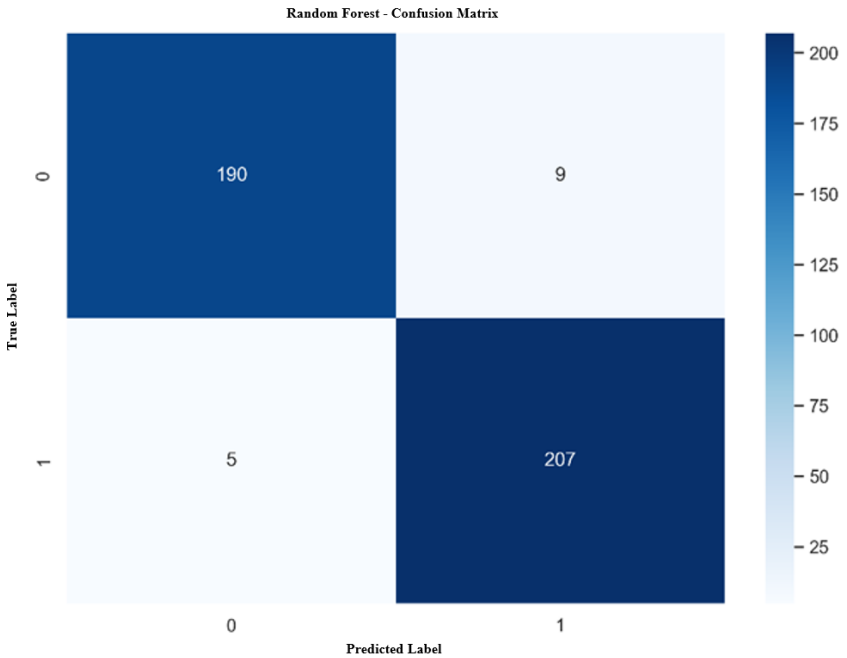

Figure 5. Random Forest - Confusion Matrix

3.2. Model Results Description

This study compares the performance of three classification models—SVM (linear kernel, radial basis kernel) and random forest—in the target task. Based on the confusion matrices and quantitative indicators, the following core conclusions from the table are drawn:

Table 6. Analysis of Feature Factors

Features	Model		
	SVM Linear Kernel	SVM Radial Basis Kernel	Random Forest
Accuracy	0.652	0.803	0.966
Precision	0.655	0.813	0.958
Recall	0.689	0.802	0.976
F1 Score	0.671	0.808	0.967
Training Time	0.052	0.045	0.124
Testing Time	0.002	0.007	0.004

- The SVM radial basis kernel outperforms the linear kernel: In the confusion matrix of the SVM radial basis kernel, the number of correctly predicted samples for true labels 0 and 1 is 160 and 170, with 39 and 42 misclassifications; while the SVM linear kernel only has 122 and 146 correctly predicted samples, with 77 and 66 misclassifications.
- The random forest performs the best: In its confusion matrix, the number of correctly predicted samples for true labels 0 and 1 reaches 190 and 207

respectively, with only 9 misclassifications (true 0 misjudged as 1) and 5 misclassifications (true 1 misjudged as 0). Corresponding to an accuracy rate of 0.966 and an F1 score of 0.967, all indicators are much higher than those of the two SVM models.

3.3. Conclusions

In terms of training and testing time: The training efficiency of the SVM linear kernel (training time 0.052 seconds) and radial basis kernel (training time 0.045 seconds) is slightly higher than that of the random forest (training time 0.124 seconds), but the testing time of the random forest (0.004 seconds) is within a reasonable range among the three. Combined with performance, the random forest has more practical value in the trade-off between "accuracy and efficiency".

Among the SVM models, the nonlinear fitting ability of the radial basis kernel is more suitable for the data distribution, and its performance is better than that of the linear kernel, but overall it is still weaker than the random forest based on the ensemble learning framework. For the classification task of this study, the random forest is the optimal choice: the proportion of misclassified samples in its confusion matrix is extremely low, and it has better classification stability and accuracy.

References

[1] Zhao Chunbo. Systematic Analysis of Factors Affecting Pipe Sticking[J]. Xinjiang Petroleum Science & Technology, 1998, 8(2): 1-6.

[2] Liao Mingyan. Condition Monitoring and Fault Diagnosis of Drilling Processes Based on Integration of Neural Networks and Evidence Theory[J]. Journal of China University of Petroleum, 2007, 31(5): 134-136.

[3] Zhang Liangjun. Tutorial on the Use of Neural Networks[M]. Beijing: China Machine Press, 2007: 26-45.