Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080220
ISSN 2371-8838 Vol. 8 Num. 2

# *Fusion of Improved Polynomial Regression and Random Forest Network Intrusion Detection Model*

**Jie Zhang*, Tao Hong, Xingxiang Lin**

*School of Computer and Artificial Intelligence, Huanghuai University, Zhumadian, Henan, China*
**Corresponding author*

***Keywords:*** Intrusion detection; Polynomial regression; Random Forest

***Abstract:*** With the advent of the era of big data and the continuous development of the Internet, it is becoming more and more important to maintain network information security. Faced with the high frequency of network intrusion, network intrusion detection system has become a key technology to detect network attacks, but improving the accuracy of intrusion detection system is still an urgent problem to be solved. To solve this problem, this paper improves the polynomial regression algorithm and proposes a network intrusion detection model that integrates the improved polynomial regression algorithm and random forest. The model is tested on the NSL-KDD dataset, and the experimental results show that the model improves the detection accuracy and the overall performance is good.

## 1. Introduction

With the rapid development of information technology and the continuous expansion of network scale, network malicious attack technology is complicated and upgraded, and various network information security incidents occur frequently. Network intrusion detection technology has become an important part of network security protection[1].However, due to the late start of intrusion detection technology in China, it is still in the developing stage, and many intrusion detection products are not accurate enough, difficult to accurately control network intrusion behavior, and there are big loopholes[2].Intrusion Detection System (IDS)[3] is a kind of proactive security protection technology. Through real-time monitoring of network,IDS can effectively perceive network attacks and provide response decisions for security managers. Intrusion detection can be regarded as a classification problem, that is, binary or multi-classification judgment of host data and network traffic data. Supervised machine learning technology can effectively classify data into categories, so it is widely used in the field of intrusion detection [4].

Literature [5] proposes an intrusion detection model RF-GBDT, which integrates random forest model for feature conversion and uses gradient lifting decision tree model for classification. This model not only reduces the training time, but also has a higher detection rate and a lower false positive rate, which has significant advantages in solving the multi-classification problem of unbalanced network intrusion detection data. Literature [6] proposes a network intrusion detection model based on random forest, which is superior to the traditional Adaboost method in DR, Accuracy and Precision, and has excellent detection performance and detection stability. Literature [7] introduces a cost-sensitive random forest algorithm based on distributed hierarchical sampling, which can reduce the

influence of data category skew, improve classifier performance, and improve detection efficiency. Literature [8] proposed an improved random forest classifier network intrusion detection method to solve the problem that overfitting occurs in the clustering process of machine learning algorithms, resulting in low accuracy of network intrusion detection. This method divides the data into different clusters by Gaussian mixture model clustering algorithm, trains different random forest classifiers for each cluster, and conducts network intrusion detection through these trained random forest classifiers. Compared with other machine learn the problem of poor detection accuracy of network intrusion detection data set, this paper proposes a network intrusion detection model that integrates improved polynomial regression algorithm and random forest. Compared with traditional machine learning algorithms, this model has better performance in judging and classifying attack behavior.

## 2. Polynomial regression with random forests

### 2.1 Polynomial regression algorithm

Polynomial regression is generally regarded as an extension of linear regression. For polynomial regression, it mainly serves the following purposes:

• To analyze the linear relationship between the independent variable and the dependent variable, and the strength of this relationship.

• Estimate the regression equation and obtain better independent variables, and estimate the dependent variables according to the independent variables.

Polynomial regression uses the Least Square Method (LSM) [9] to find the optimal function match of the data by minimizing the sum of squares of the residuals. To obtain a polynomial regression model that can fit the target data set perfectly by polynomial regression algorithm, the essence is to solve the weight θ of each characteristic independent variable. Linear regression first constructs an optimization function of a convex function and uses least squares and gradient descent to calculate the final fitting parameters.

Polynomial expansion general term:

$$[X_1 + X_2 \cdots X_m]^n = \sum_{n=1}^{n} \frac{n!}{n_1!n_2!\cdots n_m!} \cdot X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m} \tag{1}$$

Polynomial regression formula:

$$\bar{y} = \sum_{m=1}^{m} \lambda \cdot \left[ \frac{n!}{n_1!n_2!\cdots n_m!} \cdot X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m} \right] \tag{2}$$

Predicted value processing:

$$\hat{y} = \left[\bar{y} + 0.5\right] \quad , \quad \hat{y} \in (1,5) \tag{3}$$

### 2.2 Random forest algorithm

General random forest (RF) is a statistical learning theory, which uses bootsrap resampling method to extract multiple samples from the original samples, conduct decision tree modeling for each bootsrap sample, and then combine the predictions of multiple decision trees to obtain the final prediction result through voting [10].

Based on Bagging integration of decision tree based learning, random forest introduces the selection of random attributes in the training process of decision tree. Random forest algorithm is simple, easy to implement, and has low computational overhead, showing strong performance in many realistic tasks [11]. Using random forests to predict whether it's an attack or a normal access. Random forests reduce the risk of overfitting through the combination of multiple decision trees, for

example through average decision trees [12]. The larger the number of decision trees under the random forest algorithm, the better the result of generalization is often.

## 3. Network intrusion detection model combining improved polynomial regression and random forest

Random forest is an ensemble learning method based on decision trees. Its core idea is to combine a single classifier decision tree with problems of overfitting and local convergence into multiple classifier forests [13]. It can effectively reduce the variance of the model and improve the prediction performance. Polynomial regression can capture the nonlinear relationship in the data and further improve the fitting ability of the model. By combining the two, more accurate prediction of intrusion data can be achieved. As a new integrated classifier, random forest has the advantages of fewer training samples, less manual intervention and higher classification accuracy. It can process high-dimensional data and get classification results quickly. The random forest algorithm requires relatively few parameters set by users, and has good robustness compared with other traditional classification methods. The random forest algorithm can automatically select features and select the most important features by constructing decision trees [14]. The polynomial regression model has good interpretability and can intuitively show the relationship between input features and target variables. The combination of polynomial regression and random forest is helpful to improve the interpretability of the model and facilitate the analysis of the model prediction results. Both random forest algorithm and polynomial regression algorithm have good adaptability in dealing with different types of data and problems. Combining the two enables efficient and accurate forecasting in different scenarios.

Random forest has strong robustness to noisy data [15], it uses the integration of voting mechanism and decision tree. This enables the combined algorithm to maintain high predictive performance when dealing with data containing noise.

The network intrusion detection model combining improved polynomial regression and random forest is shown in the figure below.

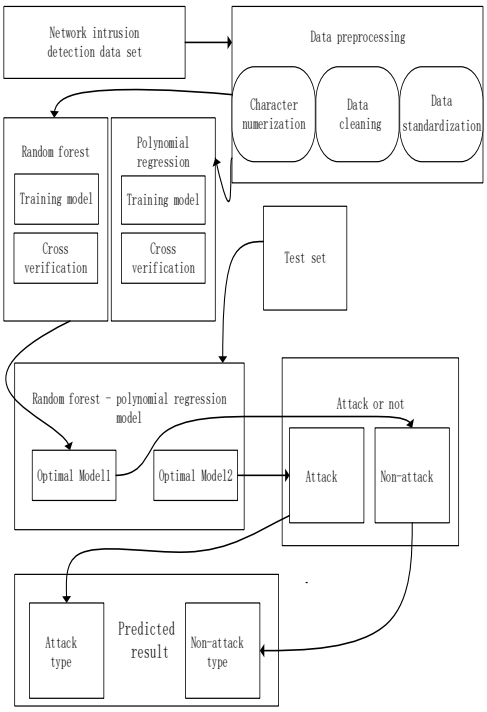



Figure 1: Intrusion detection model based on the combination of random forest and polynomial

The network intrusion detection model combining improved polynomial regression and random forest is shown in Figure 1. The intrusion detection model proposed in this paper mainly consists of two stages: model training stage and model testing stage. The specific process is described as follows:

(1) Model training stage: Data preprocessing is carried out on the original data set, such as numerical characterization of partial character feature labels, cleaning of abnormal data, and coding of feature labels, etc., to ensure the normal progress of the training model data. Model training can be parallel. For the random forest algorithm, we train the model according to the known feature labels, and get the best prediction whether the model is an attack model after cross-checking. Polynomial regression trains the model according to the known labels and cross-checks to get the best model to predict the type of attack.

(2) Model testing stage: The new verification set is tested with the trained optimal fusion model, and a stable model is obtained. The stable model is tested on the test set and the model is evaluated comprehensively.

## 4. Experiment and analysis

### 4.1 Experimental environment and data

The experimental environment was local Python (Pyspark component), the memory was 16GB, the processor was AMD Ryzen 7 4800H-CPU 2.9GHz, and the system was Windos10. Nationnal Secu Labradory Knownledge Discovery Dataset (NSL-KDD) [16] is a data set commonly used in intrusion detection research. The network intrusion detection model in this paper adopted NSL-KDD as the training set and test set of the experiment. There are both continuous data and discrete data in intrusion detection data, and the difference of order of magnitude between different feature attributes in the data is large. NSL-KDD improves the large number of redundant data in KDDCUP99 and its data is too complicated. It can not only reflect the network traffic structure and intrusion, but also modify, extend and copy. In order to build a reasonable data set, it is usually necessary to preprocess and balance the data set [17].

There are 497654 pieces of data in this experiment, and each piece of data has 43 attributes. The data includes one normal behavior and four aggressive behaviors, namely Dos, Porbe, U2R, and R2L. In this paper, the NSL-KDD data set is divided into the training set and the test set of the experiment at the ratio of 8:2. In the experiment, a stable model is obtained by training set, and the ten-fold cross-validation method [18]is used to test the test set, and multiple indexes are used to evaluate the model.

### 4.2 Model evaluation index

Because the sample distribution of NSL-KDD dataset is not balanced [19], if the accuracy rate is used to measure the model's advantages and disadvantages, the effect of the model in the real environment cannot be correctly reflected. In this paper, accuracy, recall rate, accuracy, F1, Macro-F1, MCC (Matthews correlation coefficient) and other commonly used evaluation indicators are used to evaluate the model. The formula of each evaluation indicator is shown in Table 1, and each evaluation indicator is explained as follows:

(1) Accuracy represents the proportion of samples correctly classified by the classifier to the total number of samples.

(2) The recall rate represents the number of categories correctly identified by the classifier divided by the number of categories, reflecting the classifier's recall rate for a category.

(3) The accuracy rate represents the number of a category correctly recognized by the classifier divided by the number of a category recognized by the classifier, reflecting the accuracy rate of the classifier for a category.

(4) The F1 value is the harmonic average of the recall rate and the accuracy rate, and the recall rate and the accuracy rate affect each other. Under normal circumstances, the accuracy rate is high, the recall rate is low; High recall rate, low accuracy rate; If both values are high, F1 is used to measure them.

(5) F1-score is an indicator used to measure the accuracy of binary classification model in statistics, and is used to measure the accuracy of unbalanced data. It takes into account both the accuracy rate of classification model and the recall rate. F1-score can be regarded as a weighted average of model accuracy and recall rate. Its maximum value is 1 and its minimum value is 0. Macro-F1 evaluation is adopted in this paper.

(6) MCC is essentially the correlation coefficient between the observed and predicted binary classifications; It returns a value between -1 and +1. The coefficient +1 indicates perfect prediction, 0 indicates no better than random prediction, and -1 indicates a complete inconsistency between prediction and observation. Statistics are also called phi coefficients. As a correlation coefficient, the Matthews correlation coefficient is the geometric average of the regression coefficients of the problem and its duals, and it is a highly informative fraction for establishing the predictive quality of a binary classifier in a confusion matrix environment.

Table 1: Formula of each evaluation index

| INDEX | FORMULA |
|---|---|
| **ACCURACY** | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| **RECALL** | $Recall=\frac{TP}{TP+FN}$ |
| **PRECISION** | $\frac{TP}{TP+FP}$ |
| **F1** | $\frac{2\cdot Recall\cdot Precision}{Recall+Precision}$ |
| **MCC** | $\frac{TP*TN-FP*FN}{\sqrt{((TP+FP)*(TP+FN)*(FN+TN)*}}$ |

Confusion Matrix, also known as error matrix, is a standard format for precision evaluation, expressed in the matrix form of n rows and n columns. The confusion matrix is often used to calculate indicators to evaluate the validity of the model, as shown in Table 2, which describes four possible scenarios.

True Positive (TP) represents the positive sample predicted by the model to be positive. False Positive (FP) represents the negative sample predicted by the model to be positive; False Negative (FN) represents a positive sample predicted by the model to be negative; True Negative (TN) represents a negative sample predicted by the model to be negative.

Table 2: Confusion matrix

| REAL | Predicted Results | |
|---|---|---|
| | Positive | Negative |
| Positive | (True Positive) | (False Negative) |
| Negative | (False Positive) | (True Negative) |

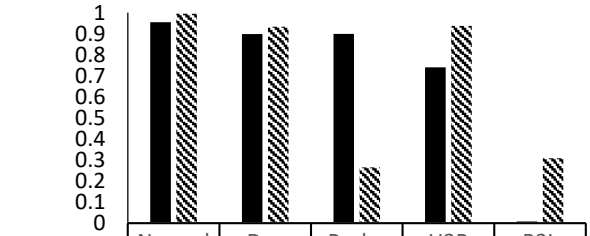
### 4.3 Experimental results and analysis of mixed model

In the experiment, we first preprocessed the raw data to eliminate noise and fill in missing values. The pretreatment process is helpful to improve the performance and generalization ability of the

model. Next, we will use the unique intrusion detection model proposed in this paper to perform classification detection on the preprocessed data.

In order to improve the prediction accuracy of the model, we optimize the polynomial regression algorithm. During the experiment, we compare multiple models such as first-order polynomial regression, second-order polynomial regression, and third-order polynomial regression, and evaluate their performance according to their prediction accuracy and generalization ability. The experimental results show that the cubic polynomial regression model achieves satisfactory prediction results on the test set, and shows good generalization ability.

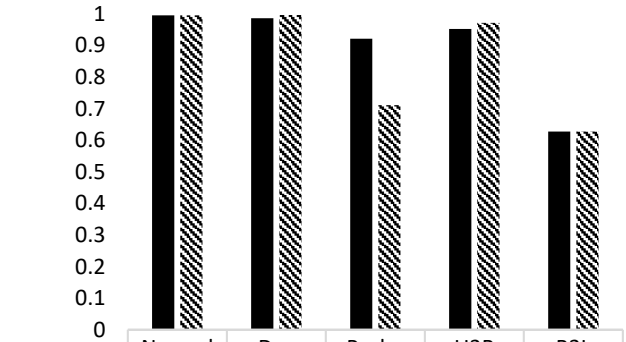
Based on this result, we choose the cubic polynomial regression model as the optimal model and apply it to the intrusion detection task. The experimental results show that this model has high accuracy and reliability in detecting intrusion behavior, and provides strong support for practical scenarios.




| | Normal | Dos | Probe | U2R | R2L |
|---|---|---|---|---|---|
| Recall | 0.955 | 0.898 | 0.899 | 0.74 | 0.008 |
| Precision | 0.995 | 0.933 | 0.265 | 0.937 | 0.308 |

Figure 2: Fusion algorithm of random forest and first order polynomial regression

As shown in Figure 2, the algorithm combined with random forest and a polynomial regression has poor fitting effect in the data obtained from multiple experiments. In terms of accuracy rate, the accuracy rate of U2R and R2L attack types is only about 70%, while the accuracy rate of Probe type is only 50%, and the recall rate of R2L is only 10%. This model works poorly.

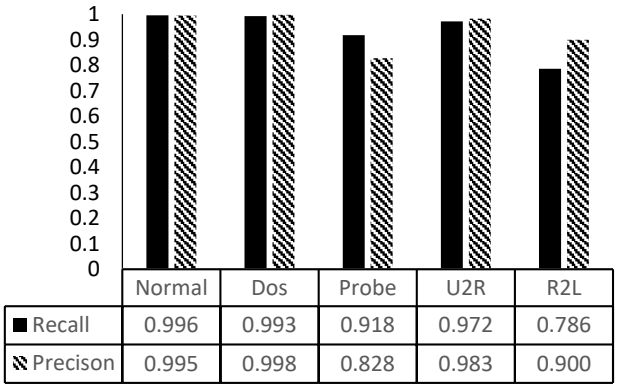



| | Normal | Dos | Probe | U2R | R2L |
|---|---|---|---|---|---|
| Recall | 0.995 | 0.986 | 0.921 | 0.952 | 0.627 |
| Precision | 0.995 | 0.996 | 0.711 | 0.971 | 0.627 |

Figure 3: Fusion algorithm of random forest and quadratic polynomial regression

As shown in Figure 3, after the polynomial regression algorithm is upgraded, the data is classified by the random forest algorithm, and the poor index of the combined algorithm model of random forest and primary polynomial regression mentioned in Figure 1 has been significantly improved. For example, the accuracy rate of Probe type increased from 50.70% to 71.08%, and the recall rate and accuracy rate of U2R type both increased from about 70% to more than 95%, with a significant increase. The recall rate of the R2L type has increased from 10% to 62.70%, and the accuracy rate

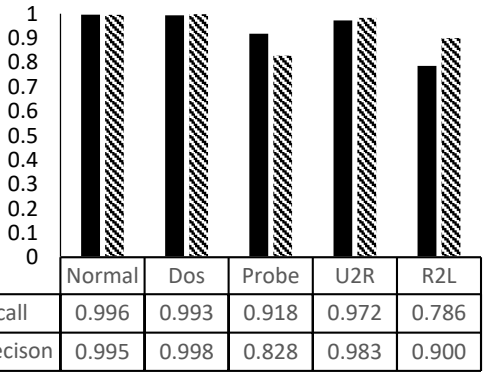
has increased from 70.27% to 87.5%. Although the overall model effect has improved, it is still not ideal.




| | Normal | Dos | Probe | U2R | R2L |
|---|---|---|---|---|---|
| Recall | 0.996 | 0.993 | 0.918 | 0.972 | 0.786 |
| Precison | 0.995 | 0.998 | 0.828 | 0.983 | 0.900 |

Figure 4: Fusion algorithm of random forest and cubic polynomial regression

As shown in Figure 4, after several experiments, we obtained a stable model whose performance improved significantly from the first-order polynomial regression model (multiple linear regression) to the cubic polynomial regression model. In the prediction of normal behavior and several types of attack behavior, the recall rate and accuracy have been significantly improved, especially for Nolmal type and Dos type, and several other types of indicators have reached a relatively good level. On the whole, the model is better than the random forest combined quadratic polynomial regression algorithm in performance, and the algorithm is relatively low complexity, easy to implement and good effect. For the selected optimal model, namely the random forest-cubic polynomial regression model, we then choose to evaluate its experimental data.




| | Normal | Dos | Probe | U2R | R2L |
|---|---|---|---|---|---|
| Recall | 0.996 | 0.993 | 0.918 | 0.972 | 0.786 |
| Precison | 0.995 | 0.998 | 0.828 | 0.983 | 0.900 |

Figure 5: Random forest-quartic polynomial regression fusion algorithm

As shown in Figure 5, after the polynomial regression algorithm is increased to four times, the impact on different types of data is more complex. While some types of accuracy have improved significantly, recall rates and accuracy rates have declined for some types of data. Possible causes include overfitting, data imbalance, noise interference, etc. These factors can cause models to perform poorly on certain types of data, reducing recall rates. In this case, it is obviously not reasonable to blindly insist on continuing to upgrade in order to improve the accuracy.

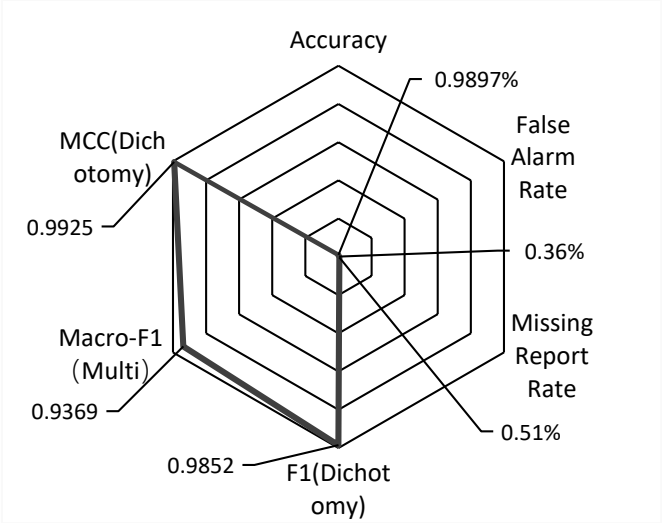
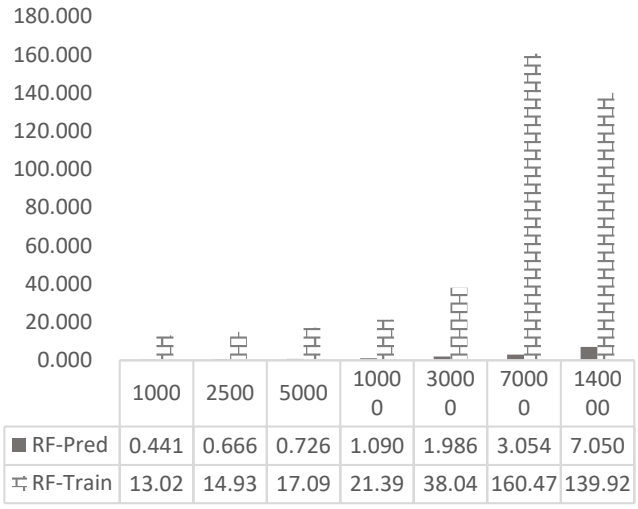



Figure 6: Random forest - cubic polynomial regression fusion model evaluation




| | 1000 | 2500 | 5000 | 10000 | 30000 | 70000 | 140000 |
|---|---|---|---|---|---|---|---|
| RF-Pred | 0.441 | 0.666 | 0.726 | 1.090 | 1.986 | 3.054 | 7.050 |
| RF-Train | 13.02 | 14.93 | 17.09 | 21.39 | 38.04 | 160.47 | 139.92 |

Figure 7: Training and prediction time of random forest algorithm (seconds / 10,000)

As shown in Figure 6, the accuracy rate of the stable model reaches 98.97%, the false positive rate is 0.36%, and the false negative rate is 0.51%. F1, Macro-F1 and MCC all have high indicators, reflecting the overall good effect of the model.

As shown in Figure 7, with the continuous increase in the amount of training data, the training time of the random forest model also presents a significant increasing trend. When the amount of data reached 10,000, the training time had reached about 21 seconds. As the amount of data grew, the training time increased significantly, and when the amount of data reached 140,000, the training time was as high as 160 seconds.

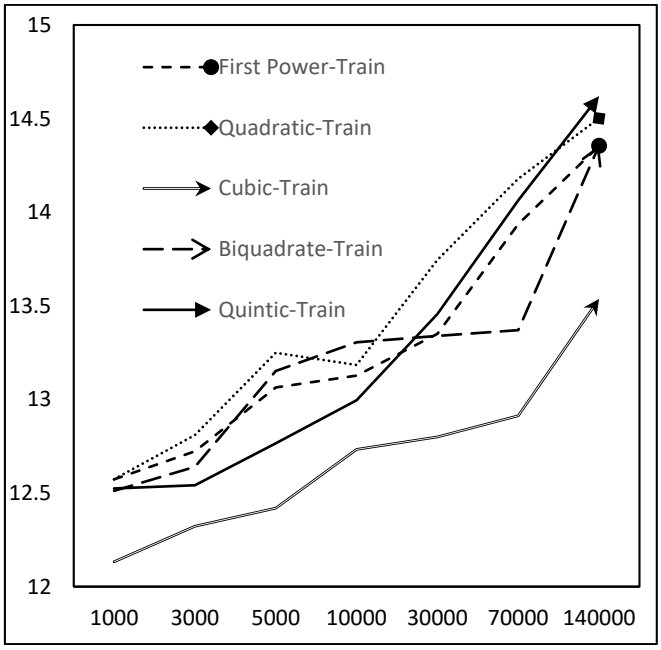



Figure 8: Training time of polynomial regression algorithm model (seconds / 10,000)

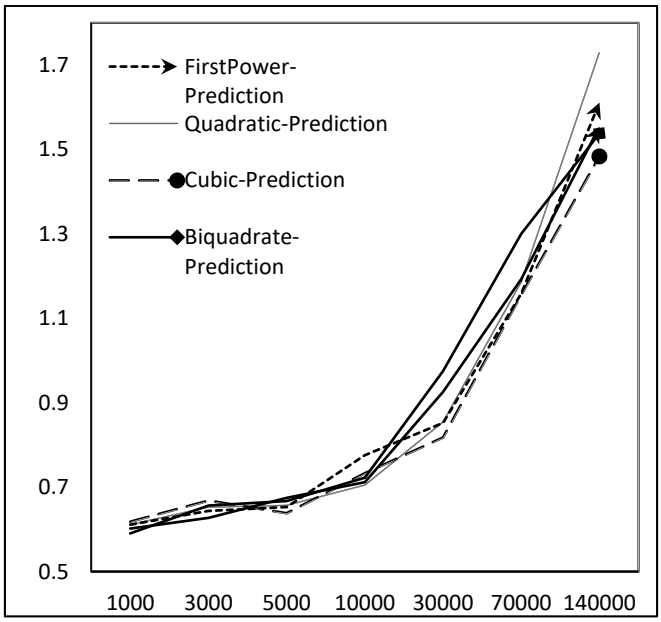



Figure 9: Prediction time of polynomial regression algorithm model (seconds / 10,000)

Therefore, it is difficult to train the random forest model in real time when the predictive model is trained with massive data. In order to improve the training efficiency of random forest model, some optimization measures should be taken to improve the defects of too long training time.

In addition, as shown in Figures 8 and 9, from first-order polynomial regression to fifth-degree polynomial regression, the increase in prediction and training time is relatively small compared to the increase in data volume. When training the model, the training time of cubic polynomial regression is shorter. When training 140,000 pieces of data, it takes about 13 seconds, or about 0.096 milliseconds, to process each piece. In the model prediction, the cubic polynomial regression predicted 140,000 data in about 1.48 seconds, about 0.01 milliseconds to process each, which is extremely fast.

To sum up, it is difficult to train random forest model in real time for forecasting model training with massive data [20]. In contrast, the training and prediction time of cubic polynomial regression is relatively short, showing better efficiency.

## 5. Conclusion

Based on the NSL-KDD dataset, this paper proposes a network intrusion detection model combining improved polynomial regression and random forest. Compared with the traditional model, this model significantly improves the prediction performance, reduces the risk of overfitting, and has obvious advantages in anti-noise ability, automatic feature selection, model interpretability and adaptability. This hybrid model can effectively deal with different types and complexity of data, and provides a powerful tool for solving practical problems. In this study, we explore prediction methods that combine random forest and polynomial regression to take full advantage of the advantages of both methods in capturing nonlinear relationships and improving prediction performance. However, in practical applications, selecting the appropriate model fusion strategy is crucial to achieve the best performance of the model, which may require a lot of experiments to adjust the model parameters.

In summary, this study provides an effective framework for the combination of random forest and polynomial regression. We hope that this approach will provide researchers and practitioners with a valuable tool for solving predictive tasks in a variety of practical problems. Future research can be expanded in several aspects, such as exploring more model fusion strategies to improve predictive performance and computational efficiency, and achieve a wider range of applications.

## References


*[1] Ghorbani A A, Lu W, Tavallaee M. Network intrusion detection and prevention: concepts and techniques[M]. Springer Science & Business Media, 2009.*

*[2] Ponomarev S, Atkison T. Industrial control system network intrusion detection by telemetry analysis[J]. IEEE Transactions on Dependable and Secure Computing, 2015, 13(2): 252-260.*

*[3] J.P. Anderson, Computer security threat monitoring and surveillance[R]. Technical report, James P.Anderson Company, Fort Washington, Pennsylvania, 1980.*

*[4] Liao H J, Lin C H R, Lin Y C, et al. Intrusion detection system: A comprehensive review[J]. Journal of Network and Computer Applications, 2013, 36(1): 16-24.*

*[5] Jieying Zhou, Pengfei He, Rongfa Qiu, Guo Chen, Weigang Wu. Intrusion Detection Based on Random Forest and Gradient Boosting Tree.Journal of Software, 2021, 32(10):3254-3265*

*[6] Ji Jun, Jun Li, Chen Chen, et al. Network Intrusion Detection Method Based on Random Forest[J]. Computer Engineering and Applications, 2020, 56(2):7. DOI:10.3778/j.issn.1002-8331.*

*[7] Hu Zhipeng, Yan Bingyong, Peng Yigong. Cost-sensitive random forest algorithm for hierarchical sampling and its application [J].Computer Engineering and Design, 2019, 40(12):6. DOI:CNKI:SUN:SJSJ.0.2019-12-001.*

*[8] XIA Jingming, LI Chong, TAN Ling, et al. Improved Network Intrusion Detection Method for Random Forest Classifier [J]. Computer Engineering and Design, 2019, 40(8): 2146-2150.*

*[9] Shi T, Horvath S. Unsupervised learning with random forest predictors [J]. Journal of Computational and Graphical Statistics, 2006, 15(1): 118-138.*

*[10] Prasad A M, Iverson L R, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction [J]. Ecosystems, 2006, 9: 181-199.*

*[11] Kwok S W, Carter C. Multiple decision trees[M]//Machine intelligence and pattern recognition. North-Holland, 1990, 9: 327-335.*

*[12] Ali J, Khan R, Ahmad N, et al. Random forests and decision trees[J]. International Journal of Computer Science Issues (IJCSI), 2012, 9(5): 272.*

*[13] Kursa M B. Robustness of Random Forest-based gene selection methods[J]. BMC bioinformatics, 2014, 15: 1-8.*

*[14] M. Tavallaee, E. Bagheri, W. Lu, et al. A detailed analysis of the KDD CUP 99 data set[C]. IEEE International Conference on Computational Intelligence for Security & Defense Applications (CISDA), 2009: 1-6.*

*[15] Zhang Hongzhuo, Li Zhihua, Wu Pengwei. Research on Intrusion detection technology for high-dimensional unbalanced data [J]. Network Security Technology and Application, 2023(06):61-63.*

*[16] van der Gaag M, Hoffman T, Remijsen M, et al. The five-factor model of the Positive and Negative Syndrome Scale*

*II: a ten-fold cross-validation of a revised model [J]. Schizophrenia research, 2006, 85(1-3): 280-287.*
*[17] Parsaei M R, Rostami S M, Javidan R. A hybrid data mining approach for intrusion detection on imbalanced NSL-KDD dataset [J]. International Journal of Advanced Computer Science and Applications, 2016, 7(6): 20-25.*
*[18] Yu P S, Yang T C, Chen S Y, et al. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting [J]. Journal of hydrology, 2017, 552: 92-104.*
*[19] Faruk D Ö. A hybrid neural network and ARIMA model for water quality time series prediction [J]. Engineering applications of artificial intelligence, 2010, 23(4): 586-594.*
*[20] Shi M, Hu W, Li M, et al. Ensemble regression based on polynomial regression-based decision tree and its application in the in-situ data of tunnel boring machine[J]. Mechanical Systems and Signal Processing, 2023, 188: 110022.*