The research of Random Forest Intrusion Detection Model based on Optimization in Internet of Vehicles

Yalin Li , Fei Li*, Jiaqi Song
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China

*lifei@cuit.edu.cn

Abstract. Security model is the main means to protect the network information security of vehicle. Due to the rapid development of artificial intelligence in recent years, machine learning technology is also emerging in the field of Internet of vehicles security. The random forest model is a strong classifier and can prevent overfitting better than the decision tree model. However, only using the traditional random forest invasion detection model has some problems, such as: the model detection time is long, the false alarm rate is high, the ability of using platform transplantation is poor, etc. In this paper, it is optimized in a lightweight way to reduce the time consumption and improve the accuracy of intrusion detection in the vehicle networking intrusion detection model.

Keywords: Random forest, Intrusion detection, Artificial intelligence, Car networking, Automotive safety

1. Introduction

1.1. Background
With the rapid development of the Internet, the security situation of cyberspace is becoming increasingly complex, and the means of network intrusion are becoming more diversified. The intrusion behavior brings great threats to the network ecological environment, and such threats also exist in the field of Internet of vehicles. How to detect abnormal samples more accurately and quickly is the research focus of intrusion detection at present. Thanks to the rapid development of artificial intelligence in recent years, machine learning technology is also emerging in the field of Internet of vehicles security.
In addition to the basic electronic control system. At present, there are other auxiliary driving systems, such as automobile tire pressure detection, brake auxiliary equipment and intelligent parking system, etc. Most of the data information of these systems is obtained from CAN (vehicle controller LAN) bus. At the beginning of the design of CAN bus [1], connecting to external networks was not considered. Therefore, in the process of realizing the Internet of vehicles, many remaining security loopholes need to be filled up.

In 2015, Chrysler was forced to recall 1.4 million vehicles after Jeep's on-board entertainment system was found to have a bug that allowed the brakes and steering to be controlled remotely. In 2016, Tencent Cohen Lab realized a remote attack on Tesla. They changed the main screen of Tesla into the logo of Cohen Lab, and the car owner could not do any operation. Later, they realized remote unlock of the car and control some functions of the car while moving, such as rearview mirror and trunk, etc. In February 2017, the safety performance evaluation standard of the Internet of Vehicles points to the data security and privacy security of customers. In June of the same year, the database of an American business group was attacked, and the distribution data of up to 10 million cars of several brands were exposed [2].

The intrusion detection model based on random forest algorithm is an important detection method of machine learning in the field of vehicle networking intrusion detection. However, only using the traditional RANDOM forest intrusion detection model has some problems, such as long detection time, high false alarm rate, poor transplantation ability of using the platform and poor generalization. Based on the random forest algorithm, network traffic feature selection algorithm and unbalanced data classification technology in machine learning, this paper aims to reduce the time consumption of intrusion detection model and improve the accuracy of intrusion detection, so as to improve the application of intrusion detection model in the Internet of vehicles.

1.2. Research Actuality
As for the exploration of automobile safety model, many scholars in China have put forward their own feasible schemes or theoretical models. In Yu He [3] 'Research on Networked Automobile Information Security And CAN Bus Anomaly Detection Technology', he proposed CAN bus anomaly detection based on information entropy and CAN bus anomaly detection based on decision tree respectively. Both of these two theoretical models show good effects in their simulation environment and can detect abnormal messages well. However, for attacks with short time intervals, CAN bus anomaly detection based on information entropy cannot effectively feedback the results. However, CAN bus anomaly detection based on decision tree may overfit the results due to the characteristics of the decision tree itself, resulting in inaccurate results. In The study of VANET Information Security Issues and Anomaly Detection Technology by Huang Yue [4], the author proposed VANET anomaly detection based on random forest. Due to the nature of random forest, this model can detect anomalies well, but its resistance to witch attacks is too low. In Yang Zhe [5], he proposed a variety of Sybil attack algorithm models for different stages and carried out relevant experimental analysis, but they were only for this kind of attack alone, and there was no further explanation for the model of flood attack.

LIN designs a security policy IDT&C that USES CAN message counter and ID table to generate CAN message code MAC [6]. Schweppe et al. designed a 32-bit MAC security architecture based on the EVITAHSN security system, which CAN be applied to THE CAN bus [7]. Groza et al., focusing
on CAN bus security, proposed several lightweight authentication protocols such as EPSB and libr-can, and verified their usability in small-scale ECU networks [9,10]. By analyzing the signal characteristics of CAN bus, he proposed a method to identify the sender of information [11].

WOO et al. designed a lightweight encryption communication method using 32-bit AES algorithm, and compared the bus load rate and transmission response time with IDT&C and EPSB algorithm. In the bus environment with less than 20 ECU, the bus load rate of their method was about 50%[12]. LU applied Markov decision Process algorithm (MDP) to build a model for attacking the car process, and did relevant research on the method of encrypting the STORAGE system of ECU [13]. Lu proposed a method to detect attack CAN bus, and used security rules to detect the behavior of malicious ECU, and gave an example of relevant security rules [14].

Based on the above research contents, it can be seen that although all the scholars have built security models and optimized and improved the efficiency, accuracy and complexity, substantive defense measures or simulated anomaly detection have not achieved good results. When using information entropy correlation model, the detection result of information entropy for illegal messages with a very short existence time is not ideal. When using decision tree correlation model, it is easy to overfit. The stochastic forest model can solve the above problems well while maintaining higher accuracy. Therefore, this paper will propose a set of optimized security threat detection model based on random forest, so that it CAN be applied in CAN bus.

2. Anomaly detection method and optimization

The security research of vehicle-mounted CAN bus message is mainly to design a series of measures such as encryption and decryption at the protocol layer to ensure the security of CAN bus message transmission, but this method will lead to a large amount of computing overhead and there are problems such as backward compatibility [15]. Message detection focuses on checking whether the frame format and number of digits of CAN bus messages are correct, while it is usually impossible to determine whether the message data itself is abnormal or tampered, especially the abnormal detection of data fields carrying a lot of information and commands, which has an important impact on each ECU of the vehicle.

When the message of vehicle-mounted CAN bus is tested for anomalies, it is necessary to check whether the message data of CAN bus is normal. In this paper, the random forest classification method of machine learning is used for detection. The random forest algorithm performs well in the processing of data sets, and can handle multi-attribute (multi-feature) data sets well. Compared with other classification algorithms, the training speed is faster and the realization difficulty is lower. In this paper, CAN bus messages are used to construct the random forest model, and then the new data are put into the classification model as test data sets for classification.

2.1. Construction of random forest model based on CAN bus message

For the original data set \( \{s_1, s_2, s_3, \ldots, s_n\} \). In this paper, N training sets were extracted with Bootstrap method. The amount of data in each training set is consistent, and the amount of data is consistent with the original data set. Since each data collection is a random extraction with put back, the same data elements may appear in each newly constructed training set, and some data elements may not appear
at all (being selected). This effectively avoids falling into the problem of local optimal solution in the decision tree model.

For the selected N training subsets, \( \{ h_1, h_2, h_3, \ldots, h_n \} \) decision tree. When constructing each decision tree, M attributes are randomly selected from M features. In this paper, the index order of Gini coefficient is adopted to select M optimal features, and then the decision tree is generated according to CART node splitting algorithm. Finally, the test set packet data is input, classified by the N decision trees, and then voted and counted to obtain the final prediction result. This is the overall process of stochastic forest model prediction.

Whether CAN bus messages are correct is essentially a binary classification problem. Gini coefficient of CAN bus messages is:

\[
Gini_{CAN}(D) = 1 - \sum_{i=1}^{2} p_i^2 = 2p(1-p)
\]  

(1)

Where, D represents the training set message of a certain type of CAN bus. The specific result of classification only needs to be clear whether it is normal or not. Therefore, I has only two values of 1 or 2. If I =1, it means that the CAN bus message is normal; if I =2, it is otherwise. P represents the probability of occurrence of normal samples of CAN bus message.

The Gini coefficient was used to construct the decision tree, and the Gini coefficient after splitting was compared and analyzed for each of the randomly selected M features. Since the data fields of the CAN bus total 8 bytes, 64-bit binaries. Therefore, it is divided into 16 attributes, and the Gini coefficient after splitting is:

\[
Gini(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)
\]  

(2)

Where, \( D_1 \) and \( D_2 \) mean that training set D is divided into two parts and D is non-null. \( |D_1| \) and \( |D_2| \) represent the sample sizes of the two subsets respectively. \( |D| \) represents the total number of training set D samples. A feature may have multiple Gini coefficients. The smaller the Gini coefficient, the better the feature division effect. In the process of constructing the random forest, it is necessary to select the optimal points of each feature to split.

### 2.2. Optimization of random forest algorithm

At present, more commonly used feature selection algorithms can be divided into filtering and wrapping. Filtering algorithm takes advantage of the basic characteristics of data and adopts a method independent of the classifier to evaluate the correlation between features. It has a small time cost, but because it is independent of the classification model, the selected features may be redundant features [16], or even not conducive to the features of data classification.

The encapsulation algorithm is combined with a specific classification algorithm to select the optimal feature subset while possessing excellent classification performance. However, this method is related to a specific classifier and generally has a high computational cost, resulting in a long response time of the intrusion detection model [17].

In order to overcome the disadvantages of low accuracy and high time cost caused by using only two methods, a hybrid algorithm based on chi-square test and random forest is used to predict the intrusion detection of Internet of vehicles. This method can be divided into three stages: in the first stage, chi-square statistics are used as the evaluation index to sort all the features contained in the
training set and eliminate the quantitative features. In the second stage, the inclusion method is used to further improve the feature subset, and the stochastic forest algorithm is mainly used to determine an optimal feature set. In the third stage, this method is applied to the intrusion detection of Internet of vehicles, that is, the intrusion detection model based on this principle is evaluated.

2.2.1. Chi-square test
Chi-square test, a common feature selection method, evaluates the importance of each feature by measuring the chi-square statistics of sample categories. The chi-square is used to describe how far the actual observation deviates from the expected value. The larger the chi-square value is, the greater the deviation between the actual observation value and the expected value is, which can also be understood as the weaker the independence between the two events [18]. The chi-square value calculation method of a feature is as follows:

$$x^2 = \sum_{i=1}^{t} \sum_{j=1}^{l} \frac{(A_{ij}-E_{ij})^2}{E_{ij}}$$  \hspace{1cm} (3)

Among them, $E_{ij} = \frac{R_i \cdot L_j}{S}$, $T$ represents the number of values of the feature, $L$ represents the number of tags of the class, $A_{ij}$ represents the number of samples of the $i$th eigenvalue and the $j$ class, $R_i$ represents the number of samples of the $i$th eigenvalue, $L_j$ represents the number of samples of the $j$ class, $S$ represents the total number of samples, and $E_{ij}$ represents the expected frequency of $A_{ij}$.

After calculating the chi-square of all the eigenvalues considered in each sample, they are arranged in the order of the chi-square value from the largest to the smallest, and then the required features are filtered in turn by the required number of eigenvalues set.

2.2.2. The package method is used for optimization
The second stage is essentially a further feature optimization of the feature subset selected in the first stage. Random forest is a combinatorial classifier that USES multiple class classification regression tree to train and predict samples. In the second stage, an optimal feature set is determined by evaluating the feature candidate set by wrapping the random forest algorithm. When the optimal classification performance is reached, the process stops and the optimal feature subset is output.

2.2.3. An optimized random forest model based on the algorithm
After two steps of 3.2.1 and 3.2.2, the basic framework of the intrusion detection model can be built, as shown in Figure 1. This method can identify the target of attack category, and has the characteristics of light weight, but also can maintain high detection performance.
2.3. Optimization of random forest model in data processing

Unbalanced network traffic data is a relatively common problem existing in data sets of random samples [19]. Using SMOTE and KNN, this paper made abnormal detection of the sample. Based on this method, unbalanced data processing method is introduced into the intrusion detection model in the second chapter to build a new detection model and improve the fine granularity detection performance of intrusion categories, which is helpful to realize a more lightweight intrusion detection system. The specific steps are as follows:

1. For the small sample set, use SMOTE to generate the new sample point and get the new training set \( T_1 \).
2. For each sample point \( S_i \) with a small number of samples, KNN algorithm is used to calculate its nearest K neighbors, and the category of these K neighbors is recorded, which is denoted as \( N_{\text{max}} \) and \( N_{\text{min}} \) according to the majority class and minority class.
3. Take an \( R = \frac{N_{\text{max}}}{N_{\text{min}}} \). When R is greater than or less than a set parameter value, remove this sample from the newly generated training set \( T_1 \), and get data set \( T_2 \); Thus, the framework of the random forest model after simultaneous optimization of data processing and algorithm can be obtained, as shown in Figure 2.

![Fig 1: Basic framework of intrusion detection model](image-url)
3. Experimental results and analysis

Veins of the experimental results based on Omnet++ component simulation environment: in the Omnet++ module output window and monitor, can be view the status of the simulation. The model classification result was compared with the original test set label, and the inconsistency between the classification result and the original label was counted. Use two-dimensional matrix to describe statistical data:

\[
\begin{array}{ccc}
0 & 0 & \beta_{0-1} \\
1 & \alpha_{1-0} & 0 \\
2 & \alpha_{2-0} & \beta_{2-1}
\end{array}
\]

The statistics of each row of the matrix to a certain type of label. Line I represents the misclassification and quantity statistics of label I data.

Where, \(\alpha_{i\rightarrow j}\) represents the number of j type samples misjudged by the tag data of I. Of which \(i,j \in [0,1,2]\). Since CAN bus messages CAN be subdivided into 52 types, only 3 representative messages are selected as the research objects in the following experiment. "00000133H", "00000244H" and "000003D9H" are respectively represented by 0,1,2. These three kinds of messages all have the characteristics of large amount of data transmitted and more times of transmission. The three messages are set to 20,000 in the simulation environment.

The classification results of the unoptimized random forest model are as follows:

\[
\begin{array}{ccc}
0 & 0 & 1232 \\
1 & 680 & 0 \\
2 & 1030 & 1330
\end{array}
\]

The first row of the two-dimensional matrix indicates that \(\alpha_{0\rightarrow 1}=1232\), that is, the \(\alpha_{0\rightarrow 2}=750\) is counted as a 0 data tag, and the total number of no 2 tags is 20,000, with an error rate of 9.91%. Similarly, \(\alpha_{1\rightarrow 0}=680\), that is, the data of 680 tags with no 0 is counted as no 1 tag, and the data of no 1 tag is 20,000, with an error rate of 4.4%. The number 2 tag had an 11.8% error rate.

The classification results of the random forest model optimized only in terms of algorithm are as follows:
In the same calculation method as above, the error rate is 7.3%, 3.7% and 8.7% respectively. It can be seen that the stochastic forest model after simultaneous optimization in terms of algorithm has better optimization effect.

After data processing and algorithm optimization, the classification results of the random forest model are as follows:

\[
\begin{array}{ccc}
0 & 0 & 785 \\ 1 & 610 & 0 \\ 2 & 788 & 955 \\
\end{array}
\]

In the same calculation method, the error rate is 5.8%, 2.6% and 5.6%, respectively. It can be seen that the random forest model optimized by both data processing and algorithm has a lower error rate than that optimized by algorithm alone. This is only a sample in multiple simulations, and the results in the subsequent dozens of experiments all show the above trend. The simulation results show that the optimized stochastic forest model CAN be used to better predict intrusion detection based on CAN bus.

4. Conclusion
The intrusion detection design in this paper aims to optimize the traditional random forest model, starting with algorithm optimization and non-equilibrium data processing respectively, but there are still more optimization ideas for readers' reference, and the author hopes that more efficient and accurate optimization models can appear.

In terms of the random forest model to resist witch attacks, the paper "A Study of A Random Forest Improved Model in The Internet of Vehicles" proposed by the author has been elaborated in more detail, which is not listed here. The author hopes that a random forest model with identity authentication mechanism with better preprocessing capability can be developed. It will bring better protection to the Internet of vehicles in terms of safety and promote the wider application of intelligent unmanned vehicles.

References
[1] Congcong L 2019 Research on the Security mechanism of Internet of Vehicles Information Security Beijing Jiaotong University
[2] Liang C 2018 Security Threats and Research Status of Internet of Vehicles. China Information Security
[3] He Y 2016 Research on information security of Networked Automobiles and CAN Bus Anomaly Detection Technology. Jilin University
[4] Yue H 2017 Research on VANET Information Security and Anomaly Detection Technology. Jilin University
[5] Zhe Y 2019 Research on the Security Mechanism and Key Technologies for Internet of Vehicles. Beijing University of Posts and Telecommunications
[6] Chung-Wei L and Sangiovanni-Vincentelli A 2012 Cyber-Security for the Controller Area
Network (CAN) Communication Protocol.CyberSecurity, 2012 International Conference

[7] Schewppe H and Roudier Y 2012 Security and privacy for in-vehicle networks. 2012 IEEE 1st International Workshop

[8] Schewppe H and Roudier Y, Weyl B, et al 2011 Car2x communication: securing the last meter-a cost-effective approach for ensuring trust in car2x applications using in-vehicle symmetric cryptography. Vehicular Technology Conference (VTC Fall), 2011 IEEE

[9] Groza B and Murvay P-S 2012 Broadcast Authentication in a low Speed Controller Area Network. E-Business and Telecommunications: Springer

[10] Groza B and Murvay S 2013 Efficient protocols for secure broadcast in controller area networks. Industrial Informatics, IEEE Transactions

[11] Groza B, Murvay S and Van Herreweghe A, et al 2012 Libra-can: a lightweight broadcast authentication protocol for controller area networks. Cryptology and Network Security: Springer

[12] Murvay P-S and Groza B 2014 Source identification using signal characteristics in controller area networks. Signal Processing Letters, IEEE 21(4) pp 395-399

[13] Woo S, Jo H J and Lee D H 2015 A practical wireless attack on the connected car and security protocol for in-vehicle can. Intelligent Transportation Systems, IEEE Transactions. 16(2) pp 993-1006

[14] Yu L, Deng J and Brooks R R, et al 2015 Automobile ECU Design to Avoid Data Tampering. Proceedings of the 10th Annual Cyber and Information Security Research Conference

[15] Lingyun W, Guihe Q and He Y 2016 Journal of Jilin University (Science Edition) 56(03) pp 663-668

[16] Yonghong P, Zhiqing W and Jianmin J 2010 A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics

[17] Jinjie H, Yunze C and Xiaoming X 2007 A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognition Letters

[18] Singh Deepak, Sisodia Dilip Singh and Singh Pradeep 2020 Multi-objective Evolutionary Approach for the Performance Improvement of Learners using Ensembling Feature Selection and Discretization Technique on Medical Data. Current medical imaging

[19] Paria Soltanzadeh and Mahdi Hashemzadeh 2021 RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. Information Sciences

[20] Peng X and Sen L 2009 Flow classification method based on C4.5 Decision tree. Chinese Journal of Software Engineering 20(10) pp 2692-2704

[21] Ronghua H, Xiaomei D and Daling W 2015 Research on node replication attack and witch attack defense mechanism in wireless sensor networks. Acta electronica sinica.

[22] Zhengfeng C 2014 Stochastic Forest Algorithm Optimization Research. Capital University of Economics and Business

[23] Zhou T 2017 Research on an On-board Network Security Protection Mechanism. Chengdu University of Information Engineering

[24] Yalin L, Fei L and Jiayan Z. Study on A Random Forest Improvement Model in Internet of
Vehicles. *IOP Conference Series Earth and Environmental Conference*