Study on A Random Forest Improvement Model in Internet of Vehicles

Yalin Li, Fei Li and Jiayan Zhang
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China
598528873@qq.com

Abstract. Security model is the main means to protect the information security of automobile network. Among many relevant security models, stochastic forest model is a strong classifier model and can better prevent overfitting than decision tree model. It has good characteristics in resisting flood attacks and other aspects, but it has poor ability to resist Sybil attacks. Therefore, an identity authentication system is added on the basis of the original random forest model, which can resist Sybil attacks while supporting the work in the Internet of vehicles environment. Experimental results show that the model can achieve the described effect.

1. Introduction

1.1. Background
With the advent of the Internet era, the cooperation between traditional automobile industry and internet-type enterprises is increasingly close, and more and more electronic information products are applied to automobiles. At the same time, scholars also put forward a number of automobile network security models.

Despite a number of theoretical security models proposed by scholars, car safety problems continue to occur: in 2013, hackers at DEFCON implemented a software flaw in the ford wing tiger and Toyota prius that enabled them to connect to a network of vehicles inside the vehicle and control critical systems such as throttle and brake. In 2014, 360 company took advantage of the process design vulnerability of tesla's automobile application to realize a series of operations such as remote unlocking and switching lights [1]. In 2015, a bug in Jeep grand Cherokee's vehicle-mounted entertainment system allowed the brakes and steering to be remotely operated, forcing Chrysler to recall 1.4 million vehicles and causing considerable property damage. In 2016, tencent Cohen lab implemented a remote attack on tesla. They replaced the main screen of tesla with the logo of Cohen lab, and the owner could not perform any operation. Then, they realized remote unlocking of the car and controlling part of the vehicle in progress, such as braking, rearview mirror, trunk and so on. At the beginning of 2017, the safety performance evaluation standard of Internet of vehicles pointed to the data security and privacy security of customers. In June of the same year, the database of a business group in the United States was attacked, involving the distribution data of up to 10 million cars of several brands. In December of the same year, nissan officially announced that the database data of its financial company were stolen by hackers, including the personal information of customers and loan information [1].

Intelligent car is the general trend and is the security threat of intelligent car is an imminent problem. We are in urgent need of an efficient, accurate and practical safety model to protect the safety of cars and even our people.
1.2. Research Actuality

For the exploration of automobile safety model, many scholars in China have put forward their own feasible schemes or theoretical models. In his paper "research on networked automobile information security and CAN bus anomaly detection technology" [2], he proposed CAN bus anomaly detection based on information entropy and CAN bus anomaly detection based on decision tree. These two theoretical models show good special effects in the simulation environment and can detect abnormal messages well. But for the attack with short time interval, CAN bus anomaly detection based on information entropy cannot effectively feedback the result. However, CAN bus anomaly detection based on decision tree may lead to over-fitting and inaccurate results due to the characteristics of decision tree itself. In Huang Yue's [3] research on VANET information security and anomaly detection technology, the author proposed VANET anomaly detection based on random forest. Due to the characteristics of random forest, the model can detect anomalies well, but its resistance to witch attack is too low. In Yang Zhe [4], he proposed a variety of Sybil attack algorithm models targeted at different stages and carried out relevant experimental analysis. However, they were only aimed at such attacks alone, and the model was not further explained for flood attacks.

Lin designed a security policy IDT&C [5] that uses CAN message counter and ID table to generate CAN message message code MAC. Schewepe et al. designed a 32-bit MAC security architecture based on the EVITA/HSM security system, which can be applied to CAN bus [6]. Focusing on CAN bus security, Groza et al. proposed a number of lightweight authentication protocols such as EPSB and libra-can, and verified their usability in small-scale ECU networks [7, 8]. 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 encrypted 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 the method was about 50% [12]. Lu used Markov decision process algorithm (MDP) to establish a model for attacking the vehicle process, and made relevant research on the method of encrypting the ECU storage system [13]. Lu proposed a method to detect the attack on CAN bus and used security rules to detect the behavior of malicious ECU, and gave an example of relevant security rules [14].

It is not difficult to see from these literatures that although scholars have built security models and optimized and improved efficiency, accuracy and complexity, substantial defense means or simulated anomaly detection have not achieved good results. When the information entropy correlation model is used, the detection result of the information entropy for the illegal message with very short existence time is not ideal. When using decision tree correlation model, it is easy to overfit. The random forest model can better solve the above problems while maintaining higher accuracy, but its defense against witch attack is not satisfactory. This paper proposes a set of random forest-based security threat detection model that can resist witch attacks under the criteria of maintaining efficiency, accuracy and computational complexity that are basically similar to the parameters of the previous excellent models.

2. Existing Random Forest Security Threat Detection Model

2.1. Random Forests

The process of security threat detection is nothing more than a classification process. Decision tree model and stochastic forest model are good classifiers. However, the decision tree model also has some problems -- overfitting is easy to occur. If the number of samples of a certain category is large, the result of information gain is more inclined to such characteristics. The random forest contains multiple decision trees, and the detection results are averaged through multiple decision trees, which can significantly reduce the over-fitting phenomenon [2].

Since the random forest model has such excellent performance, the research of experts and scholars on it is also a hundred flowers blooming, only under certain conditions to compare the relevant indicators can we say whose security threat detection model is superior or inferior.
2.2. The Existing Random Forest Security Threat Detection Model is More Concerned

2.2.1. Vanet Information Security Problems and Abnormal Detection Technology Research. Huang yue’s study [3] pointed out that random forest has high detection accuracy for flood attack and a large detection error for witch attack. Although the grid search method, including coarse- and fine-grained search, can greatly ensure the accuracy of the characterization parameters, the experiment also proves its effectiveness. In addition, the author also makes a stochastic forest optimization scheme of algorithm and carries out relevant experiments. But none of this is a good test for witch attacks.

2.2.2. Other Improved Detection Models Based on Random Forest. Wang hao pointed out in his research that the algorithm based on random forest can improve the identification efficiency, effectively solves the problem caused by data imbalance, and has a good classification effect [22]. He also optimized the pretreatment of the data by using the improved SMOTE. Through experimental comparison, it can be concluded that the stochastic forest algorithm can reach a relatively balanced state in terms of accuracy and comprehensiveness of detection. However, the author did not simulate the defense of the means of attack, let alone know the security of the results.

Hu miao et al. optimized the decision tree construction stage of the stochastic forest model: in the process of constructing the classification decision tree of the stochastic forest, the fuzzy method was introduced into the nodes of the binary decision tree, in which the fuzzy region about classification division was designed, and the normal and abnormal membership functions were designed on the fuzzy region. When a sample passes through the fuzzy region of the decision tree node, if the abnormal membership degree of the sample is larger than the normal membership degree, the sample is judged as an abnormal class. Otherwise, the sample enters the sub-tree node of the decision tree, and if there is no sub-tree node, it will be judged as a normal class. The final category of this sample is determined by the voting steps in the random forest algorithm [23]. Through experiments, it is proved that the scheme has better comprehensive performance and more stable algorithm than RFV, SVDD and RFP. However, there is no relevant research on the defense of witch attack and other means of attack.

In a summary paper of hangqi, the author mentioned that there are three directions for the optimization of stochastic forest: optimization of stochastic forest algorithm combined with data preprocessing; Optimization of stochastic forest algorithm construction process and introduction of new algorithm for optimization of stochastic forest [24]. A large number of related models are also presented in this paper, but these models do not solve the problem of witch attack.

2.3. Sybil Attack and Identification

Sybil attack is a malicious node to create a number of false nodes to deceive the detection, so as to achieve their own illegal purposes of attack. As long as the nodes are authenticated, the fake nodes can be defeated and the witch attack can be successfully resisted.

Random forest has higher detection accuracy for flood attack and replay attack, and a higher detection error for witch attack. The witch attack can be well resisted by the identity authentication mechanism [27], and the idea of random forest model with identity authentication mechanism was born.

3. Improved Random Forest Security Threat Detection Model

3.1. Build Random Forest Model

3.1.1. Pretreatment. Through the optimization of grid search parameters, the optimal parameter selection is sought. The parameters include 3: extracted M eigenvalues, number of decision trees N, and number of layers L. If the original data is non-balanced, SMOTE can be used to strike a balance.

3.1.2. Sampling of Put Back. The sampling algorithm adopted in this paper is bagging algorithm: for a given sample set T, M subsets of samples of size N are selected evenly from N total samples. Then M
models can be obtained from M sample subsets. After voting, the results of the sub-algorithm are finally obtained. The essence of this algorithm is to vote on a model by multiple weak classifiers, which is a strong classifier and can obtain higher accuracy.

3.1.3. Node Split. Node splitting is the core of constructing random forest, and splitting produces a decision tree. When splitting, select a subset of the total attributes randomly and without replacement. Here, node splitting hybrid algorithm [28] is selected to construct the following function:

\[ \theta(\alpha) = \alpha_1 Gini - \alpha_2 GainRatio \]  

(1)

Where \( \alpha_1 \) and \( \alpha_2 \) cannot be both 0 and 1.

When nodes split, the CART algorithm split rule is Gini coefficient is the smallest, and the C4.5 algorithm split rule is information gain rate (calculated by GainRatio) is the largest. Therefore, when both of them reach the optimal value, the minimum value of \( \theta \) (protective) is the optimal rule for splitting nodes.

3.1.4. Generating Random Forest. Repeat steps 3.1.2-3.1.3 to obtain the optimal decision tree number \( N \) determined by 3.1.1, and stop.

![Random Forest Generation Flow Chart](image)

**Figure 1.** Random Forest Generation Flow Chart

3.2. Anomaly Detection Based On Random Forest

As shown in figure 1, the collected original data can be divided into test data and training data after pre-processing. The test data will be used to generate the decision tree, while the training data can verify the newly generated random forest model. Since the training data can be generated in real time, the adjustment of the random forest model can be more real-time and easily adjusted.
3.3. Introduce an Identity Authentication Mechanism

From the second chapter and the cited literatures, it can be known that the pure random forest model is nothing more than the selection and optimization of preprocessing, splitting algorithm and the number of attribute feature layers from the perspective of its own model optimization. No previous literature has solved the identification problem based on random forest security model. Therefore, when it is attacked by witches and other malicious attacks with pseudo-identity type, its defensive effectiveness will be low. Identity authentication, also equivalent to each "person" to set the corresponding legal "name", and this type of pseudo-identity attack is not a natural way to start.

3.3.1. Identity Authentication Model Design

(1) As shown in figure 3, in the car network certified by CA, TBOX transmits the request signal encrypted by TBOX private key to TSP.

(2) After the TSP decrypts it with the public key of TBOX, it sends the allowed signal encrypted by the private key of TSP.

(3) TBOX decrypts the allowed signal with the TSP public key and completes the authentication.
The identity authentication model needs a sufficiently superior encryption algorithm for two reasons: first, the vehicle system's own data storage and processing capacity is relatively low, and it cannot process too complex and large amounts of data; Second, if the encryption and decryption time is too long, it will seriously affect the customer experience. But too simple encryption algorithm is too easy to be broken by violence, which can also lead to security incidents. Therefore, to design a set of efficient and secure encryption communication algorithm mechanism is an important guarantee measure for identity authentication.

3.3.2. Encryption Communication Design Based on Key Matrix Algorithm. Since there are 100-200 ecus (electronic control units) in the car, a symmetric matrix can be constructed based on this. It can be transformed into matrix A by more complex power transformation so as to make it more resistant to cracking. Matrix A can contain various information about the car, which is determined by the manager.

A unary quadratic function can be constructed:

\[
F(x, y) = \sum_{0 \leq i, j \leq y} a_{ij}x^i y^j
\]  

(2)

Constructed matrix A:

\[
A = \begin{pmatrix}
    a_{00} & a_{01} & \ldots & a_{0y} \\
    a_{10} & a_{11} & \ldots & a_{1y} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{y0} & a_{y1} & \ldots & a_{yy}
\end{pmatrix}
\]

Since A is originally composed of symmetric matrix, A is still A symmetric matrix. Any two symmetric elements in A can be used as encryption and decryption keys. In this way, not only security is greatly improved, but the storage load on each unit is reduced.

3.4. Random Forest Based Identification Detection Model

Combined with the second and third sections of this chapter, the detection model with identity authentication based on random forest can be obtained. The identity authentication of this design scheme also needs the base station that can be trusted, which USES the car and the nearby base station in the limited range to carry on the identity authentication.

4. Experimental Results, Analysis and Exploration

The experimental results based on the Veins in Omnet++ component simulation environment. Omnet++ has module output window and monitor, which can view the state of simulation. The model classification result is compared with the original test set label, and the statistical classification result is inconsistent with the original label. Use two-dimensional matrix to describe statistical data:

\[
\begin{pmatrix}
    0 & 0 & \beta_{0 \rightarrow 1} & \gamma_{0 \rightarrow 2} \\
    1 & \alpha_{1 \rightarrow 0} & 0 & \gamma_{1 \rightarrow 2} \\
    2 & \alpha_{2 \rightarrow 0} & \beta_{2 \rightarrow 1} & 0
\end{pmatrix}
\]

Each row of the matrix corresponds to the statistics of a certain type of label. Line i represents misclassification and quantity statistics of label i data.

Where, \(\alpha_{i \rightarrow j}\) represents the number of samples of type j that the data of no. i tag misjudge. Of which \(i, j \in [0, 1, 2]\)

The results of the model against hongfang attack are as follows:
The first row of two-dimensional matrix represents the $\alpha_{1,0} = 6$, that is, the data of 7 no. 1 tags are counted as no. 0 data tags, the total number of no. 1 tag data is 1006, and the error rate is 0.6%. $\alpha_{2,1} = 13$, that is, 132 tag data are counted as 1 tag data, 2 tag data is 1700, and the error rate is 0.8%. From the analysis of the results, the random forest model has a good effect on flood attack detection.

The results of resisting witch attack of the random forest model before optimization are as follows:

$$
\begin{align*}
0 & \rightarrow 0 & 0 & 0 \\
1 & \rightarrow 6 & 0 & 0 \\
2 & \rightarrow 0 & 13 & 0 \\
\end{align*}
$$

With the above calculation method, the error rates are: 1%, 7.2% and 37.2% respectively. It can be seen that the random forest model has poor resistance to witch attacks.

The optimized random forest model can resist witch attack as follows:

$$
\begin{align*}
0 & \rightarrow 0 & 0 & 0 \\
1 & \rightarrow 6 & 0 & 0 \\
2 & \rightarrow 190 & 13 & 0 \\
\end{align*}
$$

With the above calculation method, the error rates are 1.2%, 0.8% and 1.6% respectively. It can be seen that the optimized random forest model has a better resistance to witch attack. In terms of time efficiency, due to the addition of identity authentication, the time consumption of the algorithm is almost the same as the previous model. Therefore, it can be approximately considered that only the time required for three times of transmission of identity authentication is increased, which is at the level of microseconds, while the time for the base station to cover the vehicle is above the level of seconds. Therefore, it can be considered feasible theoretically.

5. Conclusion
From the above experimental results, it can be concluded that the optimized random forest model can greatly reduce the error rate caused by witch attack. In the application of Internet of vehicles, some limitations of random forest are removed and it will get better development. However, how to push this application to a broader platform, which party can escort the normal and effective operation of this project, which party is willing to pay for this infrastructure and the optimization of receiving and receiving information of the base station are all the problems we will face now.

6. Reference
[1] Chen liang. Security threats and research status of Internet of vehicles [J]. China information security, 2018(06):103-105.
[2] Yu he. Research on networked automobile information security and CAN bus anomaly detection technology [D]. Jilin university, 2016.
[3] Huang yue. VANET information security problems and abnormal detection technology research [D]. Jilin university, 2017.
[4] Yang zhe. Research on security mechanism and key technologies for Internet of vehicles [D]. Beijing university of posts and telecommunications, 2019.
[5] Chung-Wei L, Sangiovanni-Vincentelli A. Cyber-Security for the Controller Area Network (CAN) Communication Protocol[C]. CyberSecurity, 2012 International Conference on, 2012: 1-7.
[6] Schweppe H, Roudier Y. Security and privacy for in-vehicle networks[C]. Vehicular Communications, Sensing, and Computing (VCSC), 2012 IEEE 1st International Workshop on, 2012: 12-17.

[7] Schweppe H, Roudier Y, Weyl B, et al. Car2x communication: securing the last meter-a cost-effective approach for ensuring trust in car2x applications using in-vehicle symmetric cryptography[C]. Vehicular Technology Conference (VTC Fall), 2011 IEEE, 2011:1-5.

[8] Groza B, Murvay P-S: Broadcast Authentication in a low Speed Controller Area Network[J]. E-Business and Telecommunications: Springer,2012: 330-344.

[9] Groza B, Murvay S. Efficient protocols for secure broadcast in controller area networks[J]. Industrial Informatics, IEEE Transactions on, 2013, 9(4):2034-2042.

[10] Groza B, Murvay S, Van Herreweghe A, et al.: Libra-can: a lightweight broadcast authentication protocol for controller area networks[J]. Cryptology and Network Security: Springer, 2012:185-200.

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

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

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

[14] Larson U E, Nilsson D K, Jonsson E. An approach to specification-based attack detection for in-vehicle networks[C]. Intelligent Vehicles Symposium, 2008 IEEE, 2008: 220-225.

[15] Muter M, Groll A, Freiling F C. A structured approach to anomaly detection for in-vehicle networks[C]. Information Assurance and Security (IAS), 2010 Sixth International Conference on, 2010: 92-98.

[16] Wolf M, Weimierskirch A, Paar C. Security in automotive bus systems[J]. Proceedings of the Workshop on Embedded Security in Cars, 2004.

[17] Chen Minxiou, Lin Bingyang. Design remote monitor system based on OSGi platform in vehicle gateway for vehicle network[C]. Proceeding 2nd International Conference on Next Generation Information Technology. USA: IEEE,2011:12-17.

[18] Li Fei, Zhang H, Wang J, et al. Toward Intelligent Vehicle Intrusion Detection Using the Neural Knowledge DNA. Cybernetics & Systems, 2018:412-419.

[19] Qin zhou. Research on vehicle network security protection mechanism [D]. Chengdu university of information engineering,2017.

[20] Wang yamen. Design and implementation of openssl-car gateway authentication system [D]. Jilin university,2010.

[21] Malerba D, Esposito F, Semeraro G. A further comparison of simplification methods for decision-tree induction[M]/ Learning from data. Springer New York, 1996: 365-374.

[22] Wang hao. Network attack detection method based on random forest [J]. Software,2016,37(11):60-63.

[23] Hu miao, wang kaijun, li haichao, Chen lifei. Random forest and anomaly detection of fuzzy tree nodes [J]. Journal of nanjing university (natural science),2018,54(06):1141-1151.

[24] Hang qi, Yang jinghui. Application status of stochastic forest algorithm for machine learning [J]. Electronic technology and software engineering,2018(24):125-127.

[25] Liu yanqi, fu feng, zhu jianqi, et al. DoS attack detection model based on active entropy [J]. Journal of jilin university: engineering science, 2011, 41(4): 1059-1064.

[26] Xu peng, Lin sen. Traffic classification based on C4.5 decision tree [J]. Journal of software, 2009, 20(10):2692-2704.

[27] Hu ronghua, dong xiaomei, wang daling. Research on defense mechanism of node replication attack and witch attack in wireless sensor networks [J]. Acta electronica sinica,2015,43(04):743-752.

[28] Cao zhengfeng. Optimization of stochastic forest algorithm [D]. Capital university of economics and business,2014.