Vehicle Network Node Behavior Classification based on Optimized Bayesian Classifier

Jijin Wang[1, a], Xiaoqiang Xiao[1, b] and Peng Lu[1, c]
[1]College of Computer, National University of Defense Technology, Changsha 470001, China

Keywords: Bayesian Classification; Optimization; Trust Value; VANET; Behavior Classification.

Abstract: Vehicle Ad-hoc Network (VANET) is a special type of mobile Ad hoc network. It has a high-speed and dynamic topological structure, which is limited by single path planning and buildings, random density distribution of network nodes, intermittent interruption of wireless transmission, shadow effect and other obstacles. In VANET, trust value of vehicle nodes has always been a problem of special concern to researchers, and also a key problem in research field of VANET. Accurate calculation of trust value of vehicle nodes can reduce the damage of traffic order by malicious nodes and avoid traffic accidents due to the transmission of wrong information. Trust value of Nodes computation depends on the own behavior and experiences of vehicles, and also depends on detection of neighbor nodes. Suggested that the behavior of the vehicle in VANET can be accurately classified by Bayesian classifier, provides the accurate basis for the node trust value computation, promote to develop in the direction of more security and stability of VANET field.

1 INTRODUCTION

Vehicle Ad-hoc Network (VANET) is a special network architecture composed of vehicle-mounted unit (OBU) and roadside facility (RSU). It connects with cellular network, WI-FI and terrestrial wireless devices through wireless data transmission, and then sends them to mobile management center through public network. VANET is a special type of mobile Ad hoc network, which has a high-speed and dynamic topological structure and is limited by single path planning and buildings, random density distribution of network nodes, intermittent interruption of wireless transmission, shadow effect and other obstacles. In recent years, the proposal of smart city has drawn more attention to VANET. For users, the most important function of VANET is by improving the traffic flow to reduce the probability of accident, and to improve the traffic flow is through the provide the right information to the driver or vehicle, but for real-time information changes or delay may lead to the vehicle-mounted system error, finally brings to the driver and passenger safety hidden trouble. In VANET, we use wireless communication, and it is very difficult to protect the safety of this type of communication. However, the information spread in VANET involves the life safety of drivers and passengers, which may bring immeasurable losses due to the tampering of a location message or accident message. In the present study phase, by studying the vehicle on-board network related scholars node trust value computation mechanism, according to the vehicle's own experience, neighbor node information feedback, TA (Trusted Authority) node's assessment of the security properties parameters, such as classifying the behavior of the node, and then based on the behavior of nodes in the number of honest behavior and honest behavior node trust value computation, judging by the node trust value of node sends out the truth of news. At present, the existing classification method is based on naive Bayes classification. In this paper, through the optimization of naive Bayes, node behavior classification is more efficient.

2 RELATED WORK

Based on the current attack mode of the vehicle-mounted network, we dig out a series of security attribute parameters (K1, K2, K3,...), for example:

Wang, J., Xiao, X. and Lu, P.
Vehicle Network Node Behavior Classification based on Optimized Bayesian Classifier.
DOI: 10.5220/0008386200050009
In *Proceedings of 5th International Conference on Vehicle, Mechanical and Electrical Engineering (ICVMEE 2019)*, pages 5-9
ISBN: 978-989-758-412-1
Copyright © 2020 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

the garbage rate of the information sent by the node, the integrity of the information sent by the node, the ratio of the node speed to the speed of the current section, etc. Through the statistics of existing attack modes and attack characteristics, we can get a priori data set, each node Vi (K1, K2, K3,...) corresponds to one or more behavior types (A, B, C, D...,$\bar{A}$,$\bar{B}$,...), where A is honest behavior and $\bar{A}$ is non-honest behavior (aggressive behavior), α is the count of honest behavior, and β is the count of non-honest behavior. Beta distribution is used to describe the Vi trust distribution of the current node, and the trust distribution of node Vi can Be expressed as Ri-Be (αi,βi).Beta distribution refers to a group of continuous distributions defined in the interval (0,1), which is a commonly used fitting distribution model in Bayesian theory (B. Coppin, 2004).Where, parameters α and β are both greater than zero, and the probability fitting of uncertain events determined by the two variables can be carried out by adjusting the values of alpha and beta. In trust management, the trust space is defined in the interval [0, 1]. α and β are defined as variables associated with the normal behavior and abnormal behavior of the evaluation object. Honest behavior and non-honest behavior of vehicle nodes meet the basic conditions of Beta distribution and can be fitted. The trust distribution Ri follows the Beta distribution of parameters α and β, denoted as Ri ~ Be (α, β).The probability density function of Beta distribution is:

$$f(R_i;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}r_i^{\alpha-1}(1-r_i)^{\beta-1} \quad (1)$$

Type, Ri (0, 1),α > 0, β > 0.

When α <1, Ri≠0; β <1, Ri ≠ 1.The expected value of its probability density function can be expressed as:

$$\mu=E(R_i)=\frac{\alpha}{\alpha+\beta} \quad (2)$$

That is, the trust evaluation value of node Oi is $\frac{\alpha}{\alpha+\beta}$

3 THE OPTIMIZATION OF NAIVE BAYESIAN CLASSIFICATION

Classification is a common method of machine learning and data mining. Depending on the number of target classifications used to categorize data sets, different methods can be selected to perform the classification. For binary classification, decision tree and support vector machine are usually adopted, but they are limited by the number of target classification not exceeding two. This strict constraint makes them difficult to generalize to fit the broad meaning of the actual classification work, where the number of target classifications is usually more than two. From the point of view of obtaining general toolbox, naive Bayes classifier is more suitable for general classification expectation. A variety of successful practical applications based on naive Bayesian classifier, such as weather forecasting service, customer credit assessment, health status classification, etc. Simply preprocess the format of the data sets in the problem domain to tabular format. This mathematical classifier can continue to calculate the validity of fitting a new piece of data into each possible classification. In this way, the classification with the highest fitness value can be selected as the most suitable classification for this data. Naive Bayes classifier is a probability classification mechanism based on Bayes' theorem, which is the last word theory of Thomas Bayes (T. Bayes, 1763; J. Tabak, 2004). From a classification perspective, the main goal is to find the best mapping between a new piece of data and a set of classifications in a particular problem domain. In order to make this mapping probabilistic, some mathematical operations are carried out to convert the joint probability into the product of prior probability and conditional probability. As a method of machine learning and data mining, this mathematical transformation is a bit unnatural and may be difficult for beginners to understand, because it turns a simple division into a long list of molecules divided by a long list of denominators. However, these unnatural transformations are necessary because prior probabilities and conditional probabilities can be easily summarized from a given data set by simply calculating the number of instances with or without given conditions (F. Yang, 2016, Hasrouny H , Samhat A E , Bassil C , et al, 2017).The classifier considers the given problem domain composed of n attributes and m attributes, and calculates the attribute vector (A1,

A2,...,An)And a set of categories (C1, C2,...,Cn).Given the validity of attribute vector fitting, (A1 = k1, A2 = k2..., An = kn), and the posterior probability is calculated by Bayesian reasoning to obtain classification Ci:

$$\begin{aligned} &P(C_i \mid (\mathrm{A}_1 = \mathrm{k}_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)) \\ &= \frac{P(C_i \bigcap (\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n))}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \mathrm{A}_n = k_n)} \\ &= \frac{P((\mathrm{A}_1 = \mathrm{k}_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n) \bigcap C_i)}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)} \\ &= \frac{P((\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n) \mid C_i) \times P(C_i)}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)} \end{aligned} \quad (3)$$

Type: 1 < i < m

After this series of probability calculation, the fitness value between the given attribute vector and the possible classification can be expressed quantitatively. When the naive Bayesian classification is applied to the behavior classification of vehicle nodes, it is necessary to convert the value of safety attribute parameters into 0/1 (for example, when the speed attribute is overspeed, it is 1 and not 0 when it is not overspeed) or regional value. In this way, it is more effective in attribute vector fitting (Wu qi-wu, liu qing-zi; 2015; R.S. Raw, M. Kumar, 2013; Li Q, Malip A, Marin K, et a1, 2012).

In equation (1), the calculation of $P((\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n) \mid C_i)$ is very complicated and greatly increases the computational complexity of such classification. In real life, a hypothesis is usually made to reduce the computational complexity, that is, assume that each attribute parameter is independent of each other, and in the vehicle-mounted network, each safety attribute parameter of the vehicle node is independent of each other, and then the following transformation can be carried out (F. Yang, 2017):

$$\begin{aligned} &P((\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n) \mid C_i) \\ &= P(\mathrm{A}_1 = k_1 \mid C_i) \times P\ (\mathrm{A}_2 = k_2 \mid C_i) \times ... \times P(\mathrm{A}_n = k_n \mid C_i) \times P(C_i) \end{aligned} \quad (4)$$

Type: 1 < i < m

Therefore, the posterior probability can be simplified to:

$$\begin{aligned} &P(C_i \mid (\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)) \\ &= \frac{P(C_i \bigcap (\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n))}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \mathrm{A}_n = k_n)} \\ &= \frac{P((\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n) \bigcap C_i)}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)} \\ &= \frac{P((\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n) \mid C_i) \times P(C_i)}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)} \\ &= \frac{P(\mathrm{A}_1 = k_1 \mid C_i) \times P\ (\mathrm{A}_2 = k_2 \mid C_i) \times ... \times P(\mathrm{A}_n = k_n \mid C_i) \times P(C_i)}{P(\mathrm{A}_1 = k_1 \bigcap \mathrm{A}_2 = k_2 \bigcap ... \bigcap \mathrm{A}_n = k_n)} \end{aligned} \quad (5)$$

Type: 1 < i < m

Since all the posteriori probabilities have the same denominator in their calculations, for comparative purposes, you can remove the denominator from the fraction just by calculating and comparing the numerator. So mapping a given property vector (A1 = k1; A2 = k2,...,An = kn)which has its validity to a classification Ci can be further simplified as:

$$P(\mathrm{A}_1 = k_1 \mid C_i) \times P\ (\mathrm{A}_2 = k_2 \mid C_i) \times ... \times P(\mathrm{A}_n = k_n \mid C_i) \times P(C_i) \quad (6)$$

Type: 1 < i < m

After calculating all the validity, the highest validity shows the best mapping between the attribute vector and its best fitting classification. When the vehicle node behavior classification is applied, threshold Cij is set for different attack types. If the threshold value is exceeded, the behavior belongs to this kind of attack behavior, adding a dishonest behavior; on the contrary, if the threshold value is not exceeded, adding an honest behavior (Wu qi-wu, liu qing-zi; 2015).

4 TEST RESULTS OF VEHICLE NODE BEHAVIOR CLASSIFICATION

In order to test the improved classifier, we selected an example from an artificial intelligence book to verify and ensure the mathematical correctness of the improved classifier. The selected data set is shown in table 1, in which each section of data is composed of attributes k1, k2, k3, k4, and k5, where the value of k1-k5 is an integer within the range of 0 to 4, and the available categories are C1, C2,C3, $\overline{C1}$, $\overline{C2}$, $\overline{C3}$.

Table 1. Prior data sets.

k1	K2	K3	K4	K5	Classification
0	1	3	0	1	C1, C2, $\overline{C3}$
1	2	1	0	1	$\overline{C1}$, $\overline{C2}$, $\overline{C3}$
0	0	0	4	1	C1, C2, C3
1	3	1	0	2	$\overline{C1}$, C2, $\overline{C3}$
1	0	2	0	4	$\overline{C1}$, $\overline{C2}$, $\overline{C3}$
0	1	2	4	4	C1, C2, $\overline{C3}$
2	0	1	4	3	C1, C2, C3
2	3	3	1	1	C1, $\overline{C2}$, C3
2	4	4	0	2	C1, C2, $\overline{C3}$
2	2	1	1	0	$\overline{C1}$, $\overline{C2}$, C3
0	1	3	4	1	$\overline{C1}$, C2, $\overline{C3}$
1	2	3	1	4	C1, C2, C3
4	1	0	1	2	$\overline{C1}$, C2, $\overline{C3}$
1	2	0	1	4	C1, C2, $\overline{C3}$
3	0	1	2	2	C1, $\overline{C2}$, $\overline{C3}$

To find the best classification for a given new data (k1=1, k2=0, k3=3, k4=1, k5=4), prior probabilities and conditional probabilities must be derived so that they can be applied to the above probability calculations in equations (1) and (4).Based on the data in table 1, the prior probabilities of classification C1, C2, C3, $\overline{C1}$, $\overline{C2}$, $\overline{C3}$ and are respectively:

Among the 45 sample data, 9 belong to C1 class. Since no specific attribute values are checked, the prior probability of C1 class is 9/45, that is, P (C1) = 9/45.Of the 45 sample data,

Among the 45 sample data, 10 belong to C2 class. Since no specific attribute values are checked, the prior probability of C2 class is 10/45, that is, P (C2) = 10/45.

Among the 45 sample data, 5 belong to C3 class. Since no specific attribute values are checked, the prior probability of C3 class is 5/45, that is, P (C3) = 5/45.

Among the 45 sample data, 6 belong to the class. Since no specific attribute values are checked, the prior probability of the class is 6/45, that is, P ($\overline{C1}$) = 6/45.

Among the 45 sample data, 5 belong to the class. Since no specific attribute values are checked, the prior probability of the class is 5/45, that is, P ($\overline{C2}$) = 5/45.

Among the 45 sample data, 10 belong to the class. Since no specific attribute values are checked, the prior probability of the class is 10/45, that is, P ($\overline{C3}$) = 10/45.

Table 2. Classify subset data for C1.

K1	K2	K3	K4	K5	Classification
0	1	3	0	1	C1
0	0	0	4	1	C1
0	1	2	4	4	C1
2	0	1	4	3	C1
2	3	3	1	1	C1
2	4	4	0	2	C1
1	2	3	1	4	C1
1	2	0	1	4	C1
3	0	1	2	2	C1

Among the 9 sample data belonging to C1 class, 2 of them have k1=1.In the case of C1 classification, the probability of k1=1 is 2/9., P (k1=1|C1) = 2/9;

Among the 9 sample data belonging to C1 class, 3 have k2 =0.K2 is equal to 0 with a probability of 3/9 in C1 classification.K2 = 0, |, C1 = 3/9;

Among the 9 sample data belonging to C1 class, 3 have k3=3.The probability that k3 is equal to 3, given the C1 classification, is 3/9. P (k3= 3|C1) = 3/9;

Among the 9 sample data belonging to C1 class, 3 have k4=1.The probability that k4 is equal to 1 is 3/9 in the case of C1 classification. P (k4= 1|C1) = 3/9;

Among the 9 sample data belonging to C1 class, 2 have k5=4.The probability that k5 is equal to 4 is 3/9 in the case of C1 classification, P (k5= 4|C1) = 3/9;

For classification C2, C3, $\overline{C1}$, $\overline{C2}$, $\overline{C3}$ and respectively according to the prior probability calculation method of C1 classification, the respective results can be obtained through python programming.

Through the comparison of the five validity degrees, it is concluded that the best fitting classification of k1=1, k2=0, k3=3, k4=1, k5=4 is (C1, $\overline{C2}$, C3).However, after removing the denominator from the posterior probability calculation, these three values no longer represent probabilities. They are simply to support the relative comparison, the extent to which a given block of data is suitable for every possible classification.

```
The Classification:
When (k1=1, k2=0, k3=3, k4=1, k5=2), the validity to be in class C1 is 0.0016
When (k1=1, k2=0, k3=3, k4=1, k5=2), the validity to be in class C2 is 0.0012
When (k1=1, k2=0, k3=3, k4=1, k5=2), the validity to be in class C3 is 0.0021
When (k1=1, k2=0, k3=3, k4=1, k5=2), the validity to be in class ~C1 is 0.0006
When (k1=1, k2=0, k3=3, k4=1, k5=2), the validity to be in class ~C2 is 0.0014
When (k1=1, k2=0, k3=3, k4=1, k5=2), the validity to be in class ~C3 is 0.0007
```

Fig 1. The result of the experiment.

Experimental results show that the optimized Bayesian classifier reduces the time complexity of probability calculation, and the calculation results are more accurate for node behavior judgment.

Through the behavior classification, we can find that this node has two honest behaviors and one dishonest behavior, which can provide basis for the next step to calculate the trust value of this node.

5 SUMMARY

Optimization of Bayesian classifier is greatly reduced the time on the probability calculation complexity, reduced the amount of writing test code, to be able to rely on a priori data set to determine the current node behavior, the behavior of the node classification more accurate and intuitive to car, for the node trust value computation provides effective basis, to solve the problem of on-board network traffic node trust value computation, the next step is to rely on the trust value is calculated in this paper, the behavior of the classification results, and the node trust value of high and low used in vehicle network in all kinds of agreement, to provide reliable basis for the next hop node selection.

REFERENCES

B. Coppin,"Artificial Intelligence Illuminated, MA: Jones and Bartlett Publishers Inc., USA, 2004, pp. 352-353.

F. Yang, "A General Purpose Probabilistic Inference Engine," IAENG Transactions on Engineering Sciences - Special Issue for the International Association of Engineers Conferences 2016, Singapore: World Scientific, 2017, pp. 33-44.

F. Yang," Crafting a Lightweight Bayesian Inference Engine," Proceedings of The World Congress on Engineering (WCE 2016), Vol. II, London, UK, 2016, pp. 612-617.

Hasrouny H, Samhat A E, Bassil C, et al. VAN et security challenges and solutions: A survey [J]. Vehicular Communications, 2017, 7:7-20.

J. Tabak,"Probability and Statistics: The Science of Uncertainty," NY: Facts on File, Inc., USA, 2004, pp. 46-50.

Li Q, Malip A, Marin K, et a1. A reputation - based announcement scheme for VANETs [J]. IEEE Transactions on Vehicular Technology, 2012, 61(9): 4095—4108.

R.S. Raw, M. Kumar, N. Singh, "Security Challenges, issues and their solutions for VANET", published in International Journal of Network Security & Its Applications (IJNSA), September 2013, Vol.5, No. 5.

T. Bayes,"An Essay Towards Solving a Problem in the Doctrine of Chances," Philosophical Transactions of the Royal Society, Volume 53, Issue 1, 1763, pp. 370-418.

Wu qi-wu, liu qing-zi. VANET secure routing trust model based on Bayesian theory [J]. Journal of Sichuan university (engineering science edition), 2015, 47(2).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS