PAPER Special Section on Cloud and Services Computing

Data Mining Intrusion Detection in Vehicular Ad Hoc Network

Xiaoyun LIU[†a)], **Gongjun YAN**[††b)], **Danda B. RAWAT**[†††c)], *Nonmembers*, *and* **Shugang DENG**[†d)], *Member*

SUMMARY The past decade has witnessed a growing interest in vehicular networking. Initially motivated by traffic safety, vehicles equipped with computing, communication and sensing capabilities will be organized into ubiquitous and pervasive networks with a significant Internet presence while on the move. Large amount of data can be generated, collected, and processed on the vehicular networks. Big data on vehicular networks include useful and sensitive information which could be exploited by malicious intruders. But intrusion detection in vehicular networks is challenging because of its unique features of vehicular networks: short range wireless communication, large amount of nodes, and high mobility of nodes. Traditional methods are hard to detect intrusion in such sophisticated environment, especially when the attack pattern is unknown, therefore, it can result unacceptable false negative error rates. As a novel attempt, the main goal of this research is to apply data mining methodology to recognize known attacks and uncover unknown attacks in vehicular networks. We are the first to attempt to adapt data mining method for intrusion detection in vehicular networks. The main contributions include: 1) specially design a decentralized vehicle networks that provide scalable communication and data availability about network status; 2) applying two data mining models to show feasibility of automated intrusion detection system in vehicular networks; 3) find the detection patterns of unknown intrusions.

key words: *vehicular networks, mobile networks, intrusion detection, vehicular ad hoc network, security*

1. Introduction

The past decade has witnessed a growing interest in vehicular networking, a subset of mobile computing [1], [2]. The initial vision vehicular networking originally sees that radio-equipped vehicles can somehow network together and, by exchanging and aggregating individual views, can build the drivers informed about potential traffic safety risks, and can heighten their awareness of road conditions and other traffic-related events. In support of vehicular networking and, more generally, of traffic-related communications, the US Federal Communications Commission (FCC) has allocated 75MHz of spectrum in the 5.850 to 5.925 GHz band for the exclusive use of Dedicated Short Range Communications (DSRC) [3]. As a result, in the near future, vehicles equipped with computing, communication and sensing capabilities will be organized into ubiquitous and pervasive networks with a significant Internet presence while on the move.

Most, if not all, applications on vehicular networks, communicate with *big data*, such as safety message which are very important and under a stringent time limit, finance information which are used online payment (e.g. ez-pass toll station and smart parking [4]), medical information such as emergency rescue [5], etc.

Malicious attackers, aiming at the big data, can launch intrusion detection to obtain access of the vehicular system. Vehicular networks are new systems, although many experts think vehicular networks will bring revolutionary to our daily driving [6], [7]. However, the new invented systems suggest that new system security holes and new intrusion methods are continuously being discovered. Additionally, newly invented intrusion attacks can go most likely undetected, resulting to unacceptable false negative error rates. There are some data mining based intrusion detection systems have been proposed in mobile ad hoc network [8], [9], sensor networks [10] and cloud computing [11]–[13]. However, in vehicular networks, this is the first attempt to adopt data mining in IDS, based on our best knowledge. Data mining in vehicular network is a process that extracts valuable information and knowledge from large, complex, noisy, and digital data sets which hide unknown but useful information. The purpose of the data mining process in this paper is to extract intrusion information from data package and transform intrusion information into an understandable structure for intrusion detection and prevention.

In addition, the most distinct feature of data mining on vehicular networks is the challenge of availability and scalability. Vehicles, the nodes of the networks, have high mobility and can only communicate with each other in short range wireless networks. The network topology will inherently and constantly change. Therefore, vehicular networks have difficulties to scalably collect data for data mining and to robustly communicate data for network status information availability. The large number cars require that intrusion detection method must be scalable. Therefore, in this paper, we propose to construct vehicular cloud [7], [14] to scalably collect and transmit network data for preparation of data mining as well as for updating new categories and new rules adopted in data mining. The whole vehicular network is partitioned into smaller subnetworks based on geographic traffic map.

Manuscript received October 29, 2013.
Manuscript revised January 15, 2014.

[†]The authors are with Anhui University, Hefei, Anhui 230601, China.
[††]The author is with University of Southern Indiana, Evansville, IN 47712 USA.
[†††]The author is with Georgia Southern University, Statesboro, Georgia 30460, USA.
a) E-mail: liujody@hotmail.com
b) E-mail: gyan@usi.edu
c) E-mail: drawat@georgiasouthern.edu
d) E-mail: 27793600@qq.com

DOI: 10.1587/transinf.E97.D.1719

2. Related Work

Although data mining application on intrusion detection has been proposed in 1998 by Lee, *et al.* [15]. But very little data mining methods has been extensively proposed to detect intrusions in vehicular ad hoc networks or vehicular networks, although there are some applications in mobile ad hoc network [8], [9], sensor networks [10] and cloud computing [11]–[13].

In vehicular networks, Grover, *et al.* proposed to apply machine learning to detect misbehavior, i.e. transmitting inaccurate messages to trigger unsafe situation [16]. But misbehavior may not be intrusion. Yan et al. [17] adopted box-counting algorithm in vehicular ad hoc network to filter the outliers of location information of vehicles to reach integrity of location information. The location coordinates are placed in a plane and are partitioned into grids. The population of the location points are counted and compared. The more location points at a Cell, the more accurate the location data is. The mean and variance values are computed after partitioning and filtering. Nilsson, *et al.* proposed forensic investigations to attacks in vehicle-to-vehicle communication [18]. The Brian Carrier's Digital Crime Scene Model is adopted as a template to investigate the attacks. But data mining as a investigating method is not applied.

In mobile ad hoc networks, Kalman filter [19] and particle filter (Monte Carlo filter) [20] are applied in the position estimation as well. To apply these filters, the input data must follow the same distribution (normal distribution). But the input data from malicious attackers does not follow a constant distribution. The compromised input from malicious attackers will cause big error in these filters. A hybrid data mining anomaly detection technique for node-based IDS has been proposed [21]. Two data mining techniques, i.e. association-rule mining and cross-feature mining, are used to characterise normal behaviours of mobile nodes and detect anomalies by finding deviance from the norm.

In sensor networks, Loo, *et al.* presented an intrusion detection algorithm for routing attacks in wireless sensor networks [10]. In cloud computing, Modi et al. [11] and Patel et al. [12] reviewed different intrusion detection in cloud computing. The author suggested that IDS and IPS should be combined in next generation of cloud computing. Sood et al. [13] proposed one combined method to improve security in cloud computing. Patel et al. [12] integrated four technologies: autonomic computing self-management, ontology, risk management, and fuzzy theory are leveraged to enable IDS.

3. The Proposed System

3.1 Available and Scalable Cloud Platform

The vehicular network is associated with a number of grids. A city or a traffic area is partitioned into grids (called Cells).

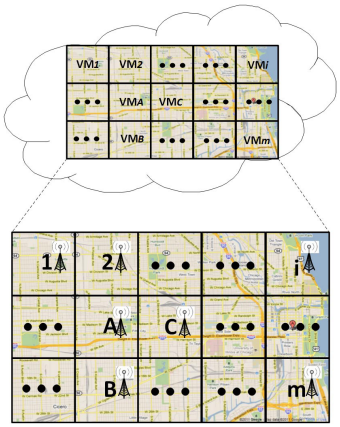


Fig. 1 A city is partitioned into Cells and the Cell is mapped with a virtual machine. A Cell is installed with a wireless tower.

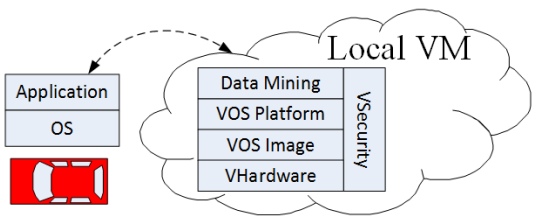


Fig. 2 A vehicle node in a cell can communicate with a virtual machine where IDS is locally built and computed.

The Cell size is predefined, e.g. 700 meters square as shown in Fig. 1. The size of the square is determined by the transmission range (i.e. 1 kilometers) of DSRC. Each Cell is associated with a virtual machine which executes intrusion detection algorithms. In each Cell, there is a transmission tower, i.e. roadside infrastructure. The tower acts as network backbone to interconnect various Local Area Networks (i.e. a Cell) and connects to Internet. A message will be encrypted and broadcasted inside the Cell. The tower will propagate the message through Internet to the tower of the destination Cell where the message is broadcasted and the receiver vehicle will pick up the message. This network platform provides a path for the exchange of intrusion information between different Cells in a *scalable* way and ensures the messages are *available*.

Each Cell executes its own local data mining models and rules, as shown in Fig. 2. New intrusion methods can be quickly detected and responded in each subnetwork. The detection in subnetwork is efficient as the size of data set is smaller than the whole network. New rules can be informed to each individual subnetwork through backbone network infrastructure.

3.2 Numerical Data Set Profile

According to current stage of intrusion detection technology [22], the following metrics shown in Table 1 and 2 are applied.

Table 1 Numerical Data Set 1.

% of establishment error/second
of establishment errors
of other errors
of to the same host
to the same service
average duration all services
average duration current service
bytes transferred all services
bytes transferred current service
% on same host to same service

Table 2 Numerical Data Set 2.

Timestamp
Source socket (IP + port)
Destination socket (IP + port)
protocol
Source bytes
Destination bytes
TCP Flags

3.3 Nature Language Data

Nature language is often harder than numerical number processing. Vehicular networks apply computational linguistics and adopt a parser-handler pattern for read text out of files and processing them. The pattern adopts the design of XML's SAX parser and content handlers, which are open available and built into Java in package org.xml.sax. A parser is created for the format in which the nature language text is represented. Then a handler for the text is created and attached to the parser. Therefore, the parser can be used to parse text from a file and all text found in the file will be passed off to the handler for processing. We also applied classes for suffix arrays of characters, of tokens, or of tokenized documents with links back to the documents from the suffix array. Suffix arrays support finding arbitrary length repeated strings in a large nature language text collection.

3.4 Data Mining Architecture

As in stated in Sect. 3.1, vehicular networks are partitioned into cells. Both a vehicle and the tower in a cell will collect network activities. The vehicle applies data mining rules to filter normal user behaviors and intrusion behaviors. Likewise, the tower can filter intrusion behavior by executing data mining models on intrusion data sets, referring intrusion detection rules which are generated by rules generator. All towers are connected by Internet networks therefore they can synchronize the intrusion detection rules along with each other. Because of dynamics of vehicular networks, the intrusion detection rules will need to be updated consistently. The advantages of the proposed system include, as shown in Fig. 3: 1) High scalability and availability. Local cells can easily receive new rules that have been developed by a remote cell no matter how far it is. The wired towers and local cell communication can ensure the scalability

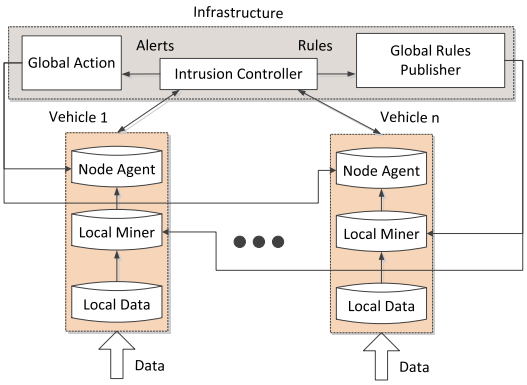


Fig. 3 The data mining architecture for IDS.

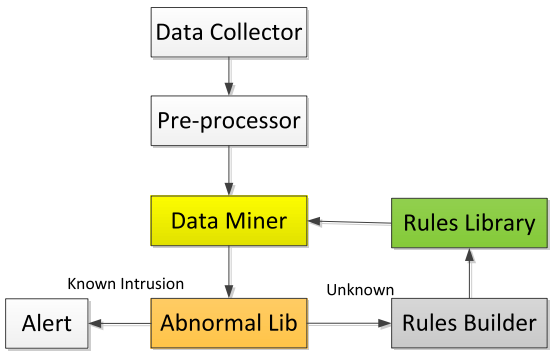


Fig. 4 The IDS data mining model on a vehicle.

and availability. 2) Dynamical updates. Each local cell contributes the system by communicating report new rules to infrastructure which thereafter broadcasts the new rules to all the other cells for references. 3) High adaptivity. Local cells know local traffic best and accumulate rules that fit the local cell best.

We propose an intrusion detection system framework by using data mining models, as shown in Fig. 4. Data collector model, as a network sniffer, captures vehicular network data which represents network health status of local area where the collector is located. Since vehicular network data is often complex big data, we pre-process the collected data and save information into specially designed data structures and then form data sets that data mining models require. The data sets are computed similarity to the known attacks and normal data in data miner module to filter outliers. If the outliers are known attacks, we send alerts. If the outliers are unknown, we put them in abnormal data sets for further analysis to generate detection rules. Therefore, it is important to dynamically update the rules library to reflect new intrusions.

The new rules will be enclosed in the rules library. The classic Apriori and extension algorithms [23] are applied to obtain the new rules. The basic steps as shown in Fig. 5 include:

- Pre-processing abnormal data. Only those related data fields are selected for generating new rules, such as

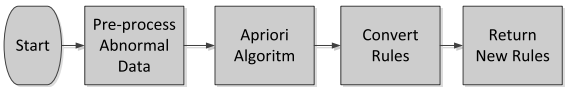


Fig. 5 The procedure of generate rules from unknown intrusion.

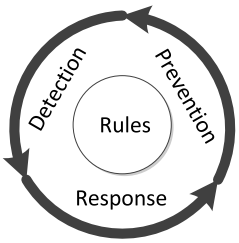


Fig. 6 The dynamical detection-response-prevention model.

source socket (IP + port), destination socket (IP + port), duration of session, etc.
- Applying Apriori algorithms to find associations between different sets of data.
- Enlisting the new associations as rules in the rules library.

Instead of static and passive detection and response to intrusion, a dynamical intrusion detection and prevention model is applied in this paper. There are sub-models including rules, detection, response, and prevention, as shown in Fig. 6. The core of the model is rules, such as new patterns of intrusions, prevention strategy and applications, etc. Detection sub-model dynamically monitors the status of vehicular network. Response sub-model alert the system administrator when intrusion is detected.

3.4.1 Optimized Data Mining Models

Two fundamental data mining models: Naïve Bayes [24] and Logistic Regression classifiers [25] are applied in this paper. The reasons that we adopt the two models are two: 1) conceptual verification. The main contribution of this paper is to show feasibility and novelty of IDS by data mining in the vehicular networks. The novelty of data mining is not our main focus. 2) The two models are well accepted and proved in many applications. Therefore, it avoids the errors caused by data mining models itself.

Let's define a document space $\mathbb{X} = \{x_1, \cdots, x_d\}$ where all documents are represented in this space. We use x for $\mathbb{X} = x_i$. A fixed set of classes is marked $\mathbb{C} = \{c_1, c_2, \cdots, c_m\}$. We use c for $\mathbb{C} = c_i$ and a training set $\mathbb{D} = \{(x_1, c_1), \ldots, (x_n, c_n)\}$ of classified documents. Each labeled document $< x_i, c_i > \in \mathbb{X} \times \mathbb{C}$. A classifier is to map documents to classes: $\gamma : \mathbb{X} \to \mathbb{C}$. In this section, we are interested two models: Naive Bayes Classifier and Logistic Regression Classifier.

3.4.2 Naïve Bayes Classifier

In Naïve Bayes model, we compute the probability of a document x being in a class c by using Bayes rules:

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)} \tag{1}$$

A document x is a list of words: $x = \{w_1, \cdots, w_n\}$. Referring previous work [24], we write

$$p(c|w_1, \cdots, w_n) = \frac{p(c)\prod_{i=1}^{n} p(w_i|c)}{\sum_{c'}\left[p(c')\prod_{i=1}^{n} p(w_i|c')\right]}. \tag{2}$$

We are interested to find the "*best*" class of Naïve Bayes classification. The best class is the most likely or maximum a posteriori (MAP) class

$$c_{max} = \arg\max_{c\in\mathbb{C}} p(c)\prod_{i=1}^{n} p(w_i|c) \tag{3}$$

3.4.3 Optimization

In this section, we are interested in optimization problem of Naïve Bayes Classifier. Let each document be represented by a word count vector $x = (t_1, \cdots, t_v)^T$. We assume for each class c, the probability distribution of a document follows the multinomial distribution with parameter θ_c:

$$p(x|c) \propto \prod_{w=1}^{v} \theta_{c_w}^{t_w} \tag{4}$$

The log likelihood is

$$\log p(x|c) = x^T \log \theta_c + const. \tag{5}$$

We also assume that the multinomial distribution assume conditional independence of feature dimensions $1, \cdots, v$ given the class c. Given a training set $\{(x_1, c_1), \ldots, (x_n, c_n)\}$. Our task, in this step, is to find the best parameters $\Theta = \{p(c = j), \theta_1, \ldots, \theta_v\}$. Therefore, we translate the model as $p(x|c = j) \propto \prod_{w=1}^{n} \theta_{j_w}^{x_w}$. According to the maximum-likelihood estimation (MLE), the maximum of the joint (*log*) likelihood of the training set:

$$c^*_{max} = \log p((x, y)_{1:n}|\Theta); \text{skip } \Theta \text{ below} \tag{6}$$

$$= \sum_{i=1}^{n} \log p(y_i) \log p(x_i|y_i) \tag{7}$$

It is easy to solve it using Lagrange multipliers [26] and arrive at

$$p(c = j) = \frac{\sum_{i=1}^{m}[c_i = j]}{m} \tag{8}$$

and

$$\theta_{j_w} = \frac{\sum_{i:c_i=j} x_{iw}}{\sum_{i:c_i=j}\sum_{u=1}^{v} x_{iu}} \tag{9}$$

The above results are intuitively explained as following: they are class frequency in the training data set, and the word frequency of each class.

3.4.4 Logistic Regression Classifier

Naïve Bayes is a generative model and models the joint $p(x, c)$, with the independence assumption on the words of x. The Logistic regression, a discriminative model, estimates $p(c|x)$ directly.

We are interested in learning classifiers, $c = f(x)$, for a set of training data sets $\{(x_1, c_1), \ldots, (x_n, c_n)\}$. For a document, the vector $x_i = (t_1, \cdots, t_d)^T$ consist of transformed word frequencies $t_i, i \in d$ from the training document.

The values $c_i \in [-1, +1]$ are class labels encoding membership (+1) or nonmembership (-1) of the vector in the category of c_i. To map x to real number, we compute the inner product between x and a parameter vector $\theta \in \mathbb{R}^v$:

$$\theta^T x \tag{10}$$

According [25], we minimizes log likelihood loss:

$$\min_{\theta} \lambda \|\theta\|^2 + \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)) \tag{11}$$

where θ is assumed as Gaussian distribution with covariance $\frac{1}{2\lambda} I$:

$$\theta \sim N\left(0, \frac{1}{2\lambda} I\right). \tag{12}$$

Recall that both Naïve Bayes and Logistic Regression are linear classifiers. They both divide the documents $\mathbb{X}$ with a hyperplane. But they differ from each other: Naïve Bayes optimizes a generative objective function, while Logistic Regression optimizes a discriminative objective function. In practice, logistic regression often has higher accuracy when training data set size is large and Naïve Bayes has an advantage when the training data set size is small.

4. Simulations

4.1 Vehicular Networks Simulation Setup

We were interested to investigate the network performance of our proposed method. We first applied SUMO [27], a mobility simulator, to generate a mobility trace file and then fed the trace file to NS-2 [28], a network simulator version 2, where wireless network is simulated. We choose to SUMO and NS-2.30 not only because they are publicly available, but also because they are both well-maintained and well-accepted in the research community. We assumed a 700m × 700m area of city streets to represent a cell, shown in Fig. 7. Vehicles entered the cell from the border streets and then randomly moved on streets in the cell. We initially placed 150 vehicles. Vehicles make random turning decisions at each intersection. The speed limits on the streets range from 5 to 20 m/s (11-45 mile per hour). Traffic lights are randomly simulated as well. SUMO generates a mobility trace file that can be imported into NS-2. Then, NS-2 is executed and nodes in NS-2 follow the nodes in SUMO respectively. Each traffic flow sends UDP packets (512 bytes per packet). We compared three scenarios with different transmission range (TR): 233m, 350m and 700m for each car.

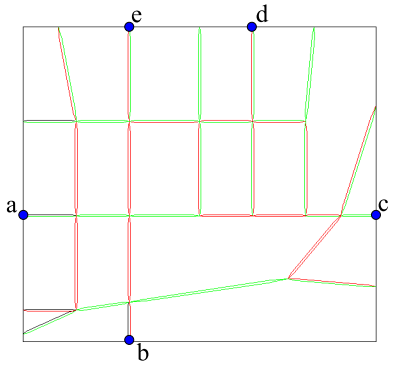


Fig. 7 Illustrating our assumed map topology which served as an example. The points (a, b, c, d, e) are assumed as the initial major entries of traffic. The green and red colored edges show the traffic lights at time t.

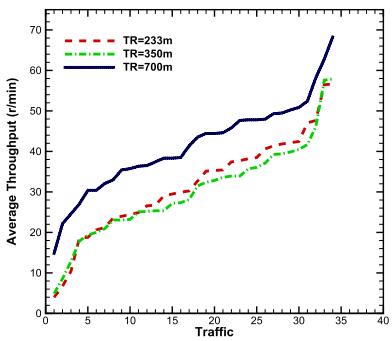


Fig. 8 Throughput of server.

The average throughput of each street requests is of interest in network simulations. We collected the throughput of each traffic flow and computed the average throughput of each traffic flow which stands for an individual street. The result is shown in Fig. 8. As expected, the throughput value of 700m TR is about 50% higher than for a TR of 350m. This is because cars in the scenario two can directly communicate with the pseudonym server but the cars in the scenario one will need to relay request to the pseudonym server when the direct connection is unavailable. It is interesting to notice that the throughput of 233m TR is similar to the one of 350m TR. Since cars are moving, the connection is not reliable and some relaying requests will be lost. Additionally, the relay cars may not be found all the time and the packet will be dropped when time is out. As long as relay is needed, two-hop communication does not make significant difference to three-hop communication.

We were also interested in the request response delays. The packet response delays of each flow were collected and computed. The result is shown in Fig. 9. As expected, the delay values of requests of the case $TR = 233m$ and the case $TR = 350m$ are slightly larger than the case

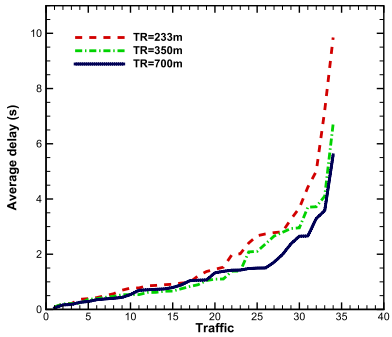


Fig. 9 Delay of requests.

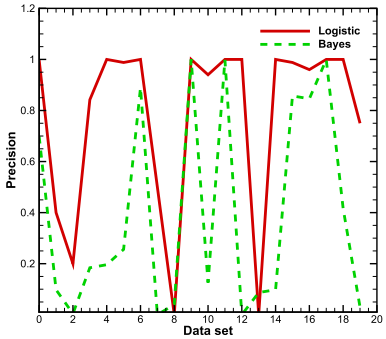
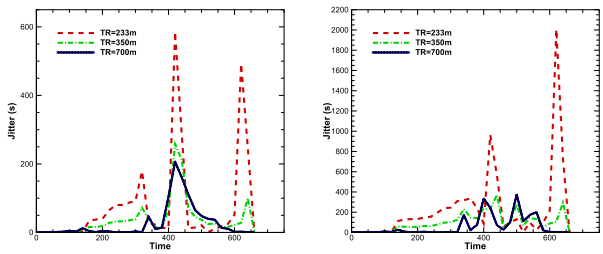


Fig. 11 Precision.



(a) Jitter 1. (b) Jitter 2.

Fig. 10 Jitters from a street.

$TR = 700m$. The reason is obviously because of the unreliable vehicular networks communication. The more hops in communication, the bigger delay values will be. In our proposed scheme, our assumption that the communication can directly reach to the cell shows both theoretical value and empirical meaning in this simulation.

We investigated the jitter of requests in the selected street. As expected, the jitter values (both Jitter 1 and Jitter 2 definitions) shows that the jitter in scenarios $TR = 233m$ and $TR = 350m$ has a significantly larger amplitude and fluctuation than the one in scenario $TR = 700m$. The reason, as mentioned earlier, lies in the mobility of vehicles and in the differences in transmission range. It is fairly interesting to note that the jitter values are higher at the middle of the day and the end of the day. For middle of day, more cars are on street and the wireless channels become more crowded. Wireless transmission is more likely to collide and the jitter values increase. Towards the end of the day, the population of vehicles is greatly decreased. Vehicles in scenario one could fail to connect the pseudonym server because no intermediate cars can be used as communication relay nodes. Comparing Fig. 10 (a) and Fig. 10 (b), we note that jitter values will be different if the jitter is defined differently.

4.2 Intrusion Classification Experiments

In this paper, simulation experiments are configured as follows. We loaded linux on five cars and each car runs several network applications such as email, stream video, online chatting, Internet browsing, etc. We will need to collect network status from these cars. The task is to train and learn a predictive model that can differentiate legitimate and illegitimate connections in a computer network.

All these data sets are obtained from TCPdump and are about 5,000,000 records recorded in about nine months. Every connection is marked as attack or normal. There are 4 attack categories with 39 attack types. The examples of the attack types are: buffer overflow, ftp write, guess passwd, loadmodule, multihop, portsweep, rootkit, etc,. We placed 22 attack types are in the training data sets. There are 17 attack types only exist in testing data sets. We notice that there are new intrusion attacks on application layer, such as script attacks, database SQL injection, crossing site scripts, etc. As we are more interested on network layer, we adopted the well-accepted benchmark data set. The four categories are

1. DOS, denial-of-service, e.g. ping-of-death, syn flood, smurf;
2. R2L, unauthorized access from a remote machine to a local machine, e.g. guessing password;
3. U2R, unauthorized access to local superuser privileges by a local unpivileged user, e.g. buffer overflow attacks;
4. PROBING, surveillance and probing, e.g. port-scan, ping-sweep;

Experiments were executed on a java simulator WEKA 3 [29]. The two adopted models (Naïve Bayes and logistic regression) were used in the experiments. We show the simulation results using widely used metrics of data mining. Recall is the ratio of the texts that are relevant to the classification and are successfully classified. Recall can also be defined as the probability that a relevant text is classified by the classification. As shown in Fig. 12, logistic regression model has higher recall than naïve bayes. Precision takes all the classified text into account and is the fraction of the texts classified that are relevant to the user's information need. As positive predictive value, precision results of logistic regression are higher than the ones of naïve bayes, shown in Fig. 11. We are interested in F_1 measure which evenly weights recall and precision. The F-measure is defined:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}.$$

Figure 13 shows the result of F-measure. As expected, the

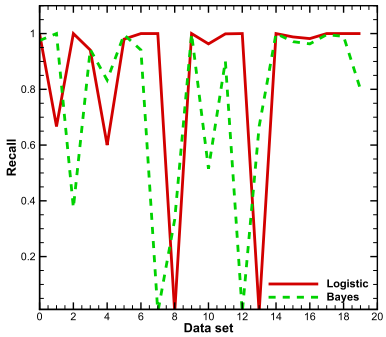


Fig. 12 Recall.

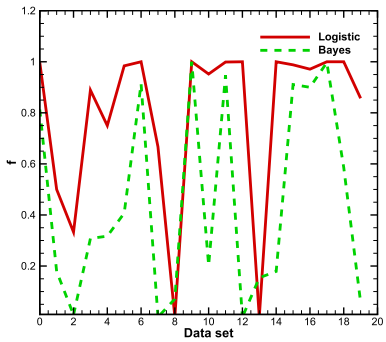


Fig. 13 F-measure.

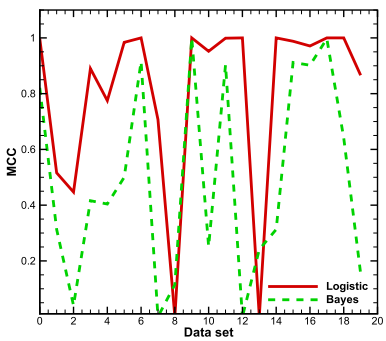


Fig. 14 MCC.

F-measure of is higher than the one of naïve bayes.

As F-measures do not take the true negative rate into account, we also showed another measure to assess the binary classifier. The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of binary (two-class) classifications. There are four types of true/false and positive/negative combinations. As expected, the MCC values of the logistic regression model are higher than the ones of naïve bayes model, shown in Fig. 14.

5. Conclusion and Future Work

In this paper, we adopt two data mining models into intrusion detection and to compare the performance of the two models in simulations. Experiments showed that logistic regression out-performed naïve bayes when the larger data sets present. As a novel attempt, this paper can improve security of vehicular networks.

As future work, we will explore other data mining models, especially ones that are suitable to vehicular networks. Multiple metrics, such as time cost, computing cost, and information retrieval metrics, will be further explored to compare the performance of different models. More importantly, new attack data sets will be searched and applied in the future. More applications will be studied to better promote the vehicular data cloud and benefit community.

References

[1] J. Lin, W. Yang, G. Yan, D.B. Rawat, and B. Wang, "Cooperative collision warning based on copula method," J. Application Research of Computers, vol.10, no.1, pp.1–11, 2012.
[2] G. Yan, S. Olariu, and D. Popescu, "Notice: An architecture for the notification of traffic incidents," IEEE Intelligent Transportation Systems Magazine, vol.4, no.4, pp.6–16, 2012.
[3] US Federal Communications Commission (FCC), "Standard specification for telecommunications and information exchange between roadside and vehicle systems - 5 ghz band dedicated short range communications (DSRC) medium access control (MAC) and physical layer (PHY) specifications," Sept. 2003.
[4] G. Yan, W. Yang, D.B. Rawat, and S. Olariu, "Smartparking: A secure and intelligent parking system," IEEE Intelligent Transportation Systems Magazine, vol.3, no.1, pp.18–30, 2011.
[5] G. Yan, Y. Wang, M.C. Weigle, S. Olariu, and K. Ibrahim, "WE-Health: A secure and privacy preserving eHealth using NOTICE," Proc. International Conference on Wireless Access in Vehicular Environments (WAVE), Dearborn, MI, USA, Dec. 2008.
[6] G. Yan, D. Wen, S. Olariu, and M.C. Weigle, "Security challenges in vehicular cloud computing," IEEE Trans. Intelligent Transportation Syst., vol.14, no.1, pp.284–294, 2012.
[7] S. Arif, S. Olariu, J. Wang, G. Yan, W. Yang, and I. Khalil, "Datacenter at the airport: Reasoning about time-dependent parking lot occupancy," IEEE Trans. Parallel Distrib. Syst., vol.99, no.PrePrints, 2012.
[8] P. Mitra and C. Poellabauer, "Efficient group communications in location aware mobile ad-hoc networks," Pervasive Mob. Comput., vol.8, no.2, pp.229–248, April 2012.
[9] N. Pham, R.K. Ganti, Y.S. Uddin, S. Nath, and T. Abdelzaher, "Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing," Proc. 7th European Conference on Wireless Sensor Networks, ser. EWSN'10, pp.114–130, Berlin, Heidelberg: Springer-Verlag, 2010.
[10] C.E. Loo, M.Y. Ng, C. Leckie, and M. Palaniswami, "Intrusion detection for routing attacks in sensor networks," International Journal of Distributed Sensor Networks, vol.2, no.4, pp.313–332, Nov. 2006.
[11] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "Review: A survey of intrusion detection techniques in cloud," J. Network and Computer Applications, vol.36, no.1, pp.42–57, Jan. 2013.
[12] A. Patel, M. Taghavi, K. Bakhtiyari, and J. Celestino JúNior, "Review: An intrusion detection and prevention system in cloud computing: A systematic review," J. Network and Computer Applications, vol.36, no.1, pp.25–41, Jan. 2013.
[13] S.K. Sood, "A combined approach to ensure data security in cloud computing," J. Network and Computer Applications, vol.35, no.6, pp.1831–1838, Nov. 2012.
[14] S. Olariu, T. Hristov, and G. Yan, "The next paradigm shift: From vehicular networks to vehicular clouds," in Mobile Ad hoc networking: the cutting edge directions, ed. S. Basagni, S.G. Marco Conti, and I. Stojmenovic, Wiley, USA, 2012.

[15] W. Lee, S.J. Stolfo, and K.W. Mok, "Mining audit data to build intrusion detection models," in KDD, pp.66–72, 1998.
[16] J. Grover, V. Laxmi, and M.S. Gaur, "Misbehavior detection based on ensemble learning in vanet," Proc. 2011 International Conference on Advanced Computing, Networking and Security, ser. ADCONS'11, pp.602–611, 2012.
[17] G. Yan, X. Chen, and S. Olariu, "Providing vanet position integrity through filtering," Proc. 12th International IEEE Conference on Intelligent Transportation Systems (ITSC2009), pp.569–574, St. Louis, MO, USA, Oct. 3-7 2009.
[18] D.K. Nilsson and U.E. Larson, "Conducting forensic investigations of cyber attacks on automobile in-vehicle networks," Proc. 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop, 2008, pp.8:1–8:6, 2008.
[19] S. il Ko, J. suk Choi, and B. hoon Kim, "Performance enhancement of indoor mobile localization system using unscented Kalman filter," SICE-ICASE, 2006. International Joint Conference, pp.1355–1360, Oct. 2006.
[20] Widyawan, M. Klepal, and D. Pesch, "Influence of predicted and measured fingerprint on the accuracy of rssi-based indoor location systems," 4th Workshop on Positioning, Navigation and Communication, 2007. WPNC '07, pp.145–151, March 2007.
[21] Y. Liu, Y. Li, H. Man, and W. Jiang, "A hybrid data mining anomaly detection technique in ad hoc networks," Int. J. Wire. Mob. Comput., vol.2, no.1, pp.37–46, May 2007.
[22] W. Lee and S.J. Stolfo, "A framework for constructing features and models for intrusion detection systems," ACM Trans. Information and System Security (TISSEC), vol.3, no.4, pp.227–261, Nov. 2000.
[23] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," Proc. 20th International Conference on Very Large Data Bases, ser. VLDB '94, pp.487–499, San Francisco, CA, USA: Morgan Kaufmann Publishers, 1994, [Online]. Available: http://dl.acm.org/citation.cfm?id=645920.672836
[24] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," Proc. 23rd International Conference on Machine Learning, ser. ICML '06, pp.161–168, New York, NY, USA, 2006, [Online]. Available: http://doi.acm.org/10.1145/1143844.1143865
[25] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York, Secaucus, NJ, USA, 2006.
[26] R.T. Rockafellar, "Lagrange multipliers and optimality," SIAM Rev., vol.35, no.2, pp.183–238, June 1993.
[27] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "SUMO (simulation of urban mobility) — an Open-source traffic simulation," MESM 2002, 4th Middle East Symposium on Simulation and Modelling, M. Al-Akaidi, Ed., pp.183–187, Erlangen, Germany, Oct. 2002.
[28] S. McCanne and S. Floyd, "NS network simulator," software available at http://www.isi.edu/nsnam/ns/
[29] Weka, "Weka 3: Data Mining Software in Java," http://www.cs.waikato.ac.nz/ml/weka/, 2012.

Xiaoyun Liu obtained M.S. in Management Information System in 1994, is currently working as Associate Professor for Anhui University, Hefei, Anhui, China. Her research interests include information system, information analysis, and e-commerce management.

Gongjun Yan received his Ph.D. in Computer Science from Old Dominion University in 2010. He is currently an Assistant Professor in University of Southern Indiana. His main research areas include intelligent vehicles, security, privacy, routing, and intelligent systems. He had more than 70 publications including journal/conference papers, book chapters, and patents. He also serves as associate editors in journals such as IEEE Transaction on Intelligent Transportation System, Ad Hoc & Sensor Wireless Networks, etc.

Danda B. Rawat received the Ph.D. in Electrical and Computer Engineering (Wireless Communications and Networking) from Old Dominion University. He is currently with the Department of Electrical Engineering at Georgia Southern University. His current research interests include design, analysis, and evaluation of cognitive radio networks, vehicular ad hoc networks, wireless sensor networks, network security, and cyber physical systems. He has published more than 60 scientific/technical papers on these topics. Dr. Rawat is a Senior Member of IEEE, and a member of ACM and ASEE.

Shugang Deng obtained B.S. in Management Information System in 2009, is currently working on his Master Degree in School of Business Anhui University, Hefei, Anhui, China. His research interests include information system and e-commerce management.