Behavior Classification based Self-learning Mobile Malware Detection

Dai-Fei Guo[1], Ai-Fen Sui[1], Yi-Jie Shi[2], Jian-Jun Hu[1], Guan-Zhou Lin[1] and Tao Guo[1]
1. Add-on IT Security, Corporate Technology, Siemens Ltd., China, Wangjing Zhonghuan Nanlu, Chao yang District, P.O.Box 8543, Beijing, 100102, China, Email: daifei.guo@siemens.com
2. State key laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China, Email: yijieshi2000@126.com

***Abstract*— More and more mobile malware appears on mobile internet and pose great threat to mobile users. It is difficult for traditional signature-based anti-malware system to detect the polymorphic and metamorphic mobile malware. A mobile malware behavior analysis method based on behavior classification and self-learning data mining is proposed to detect the malicious network behavior of the unknown or metamorphic mobile malware. A network behavior classification module is used to divide the network behavior data of mobile malware into different categories according to the behavior characteristic in the training and detection phase. Three types of network behavior data of mobile malware and normal network access are employed to train the different Naïve Bayesian classifier respectively. Those classifiers are used to analyze the corresponding type of network behavior to detect the new or metamorphic mobile malware. An incremental self-learning method is adopted to gradually optimize those Naïve Bayesian Classifiers for different behavior. The simulation results showed that those Naïve Bayesian Classifiers based on behavior classification have better accuracy rate of analysis on mobile malware network behavior. Performance simulation results showed that the network behavior analysis system based on the proposed method can analyze the mobile malware on mobile internet in real time.**

***Index Terms*—mobile internet, mobile malware, data mining, behavior classification**

I. INTRODUCTION

With the rapid development of mobile internet, many attacks driven by economic interest have appeared and caused great damage to the mobile users and operators. Those attacks exploiting mobile malware on mobile internet are more dangerous and complicated since mobile terminals contain lots of private information and have diversified and powerful communication capability.

Mobile malware can spread through multiple methods among mobile terminals and the infected terminal may be employed to launch various attacks to the mobile internet and terminal. For example, "Lanpackage.A" Trojan malware can send beguiling message with malicious website information to other mobile subscribers via MMS and tempt the user to download the malware. The infected terminal can be forced to send many MMS that will cause involuntary fee consuming. The malware can also connect the remote server to update or receive attacking target information. Its mutation "Lanpackage.C" with nickname "short message pirate" can even upload short message of the infected terminal to the remote server. Some studies have been made to analyze the behavior of mobile malware. Becher et al. [1] investigate the attack channel employed by mobile malware which can spread via MMS, Bluetooth and any mobile internet file transmission protocol. Milligan et al. analyzed the security risk of mobile phone that includes data leakage, data theft, malware spreading, and network spoofing and network congestion by spamming [2].

The existing mobile malware detection technology includes signature-based mobile malware scan and network monitor, static sample analysis and dynamic behavior analysis, etc. The traditional signature-based match technique [3] can detect the known malware with high accuracy but is not flexible enough to analyze the network behavior of the new malware or metamorphic malware [4].

Some static sample analysis methods have been employed to the malware analysis in the terminal. Batyuk et al. [5] proposed a static analysis and reporting system for android application which worked at application-binary-level and can disable malicious features from an application. Rassameeroj et al. [19] proposed an Android application contextual analysis based on a permission security model with clustering algorithms which can explore the similarity of application by visualization techniques. Barrera et al. [20] employed the Self-Organizing Map (SOM) algorithm in the permission security analysis of Android application and provided the expressiveness of permission set with visualization of permission-based systems.

Some dynamic behavior analysis technologies in the terminal have been proposed to monitor the malicious application behavior in the terminal. Blasing et al. [6] proposed an Android application dynamic analysis system which can detect the suspect or malicious behavior of applications by executing them in a sandbox. The system can also perform static analysis by comparing the application file with mobile malware pattern. Burguera et al. [15] proposed a dynamic behavior collection and analysis framework which obtained the application behavior by monitoring the kernel system

doi:10.4304/jcp.9.4.851-858

calls with a lightweight client named Crowdroid. A central server was used to collect the application behavior data and built the behavior data subset of benign application and mobile malware with a k-means partitional clustering algorithm. A dynamic malware analysis system called CWSanbox [16] has been proposed to monitor all system call of application and analyze the malicious malware behavior by executing the application in a simulated environment. Portolakidis et al. [17] proposed a cloud-based malware analysis framework which collects the execution information of application and sent to the cloud. The mobile phone replicas were created in the cloud according to the collected information and the application behavior was analyzed by running those replicas and performing security check on the application behavior in a secure virtual environment. Berthomé et al. [18] proposed an application behavior monitoring mechanism which can log and alert the application sensitive behavior, e.g. accessing private data, by repackaging the compiled application and injecting a security reporter.

In this paper, a mobile malware behavior analysis method based on behavior classification and self-learning data mining is proposed to detect unknown or metamorphic mobile malware. Our contributions are summarized below.

- The network behavior of mobile malware is analyzed according to the behavior characteristic and divided into different categories. An improved Naïve Bayesian anomalous network behavior analysis method based on behavior classification is proposed to detect the different types of network behavior of mobile malware.
- An incremental self-learning method is used to adjust the proposed behavior-classification based Naïve Bayesian Classifiers to adapt the variable network behavior of mobile malware.
- A Naïve Bayesian behavior analysis system based on behavior classification is designed to detect the mobile malware behavior on mobile internet.
- The detection accuracy of the proposed behavior-classification based Naïve Bayesian Classifiers is compared with the traditional two-category Naïve Bayesian. And their execution performance is studied by simulation.

The rest of the paper is organized as follows. The related works is reviewed and the mobile malware behavior is analyzed in Section 2. Section 3 presents the behavior-classification based Naïve Bayesian analysis method and incremental self-learning algorithm. Section 4 describes the system architecture and design. Section 5 evaluates the detection accuracy and performance of the proposed behavior-classification based analysis method on mobile internet and Section 6 concludes the paper.

II. Related Work and Behavior Classification

A. Related Works

Many machine learning technologies have been proposed to analyze the static file and dynamic behavior of malware to detect the new malware. Schultz et al. [7] propose a data-mining based framework containing RIPPER, Naïve Bayesian and Multiple Naïve Bayesian classifiers to analyze new malicious executables which can effectively detect new malicious malware samples. An automatic dynamic malware clustering analysis method based on Self-Organizing Map (SOM) and simple K-means has been applied to analyze local malwares [8] in the mobile terminal. Bayer et al. [9] proposed a malware programs dynamic behavior analysis based scalable clustering approach to group malware samples that detect the similar malware behavior. Shamili et al. [10] proposed a distributed light-weight system deployed on a network of mobile terminal to monitors the dynamic behavior and detect the malware with a distributed Support Vector Machine (SVM) algorithm.

Chiang et al. [11] proposed a mobile malware detection method based on behavioral analysis to detect new and unknown mobile malware. They employ ontology, certainty factor theory and fuzzy Petri nets to describe the behavior and automatically detect mobile malware. On the other hand, malware behaviors categorization and observation technology have been proposed by Bose et al. to create a malicious behavior pattern database and train a Support Vector Machines (SVMs) with the known behavior which can detect new Symbian malware [12].

The Naïve Bayesian classifier is of high detection accuracy and high performance when attributes of the class are independent. It is also very robust even if the given class violates the independence assumption [13]. Chen et al. [14] adopted the Naïve Bayesian to analyze the malware in the virus reporting.

B. Mobile Malware Behavior Classification

There are three phases during the mobile malware infection: dissemination phase, phase of accessing malicious server and phase of attacking. In the dissemination phase, mobile malware can be sent to the other mobile terminals via MMS, HTTP, FTP and Email, etc. After mobile malware infects mobile terminal, it may connect the malicious server to download the updated file or control command. In the end, mobile malware can launch various attacks including stealing the privacy data, sending MMS to the other terminal and even access the toll value-added service, etc.

The malware behaviors in each infection phase have different features so the behavior classification in the mobile malware analysis can improve the detection accuracy. The behavior can be divided into three categories and analyzed with different classifiers to improve accuracy: dissemination, accessing malicious website and attack.

III. Behavior Classification Based Naïve Bayesian Analysis Method

A. Behavior Classification and Data Mining

There are two stages in the network behavior data mining: analyzer training and network behavior detection. The procedure of behavior classification data mining in

training phase is shown in Fig. 1. During the training phase, the network behavior data of known mobile malware and normal network access are chosen as training data to train the behavior classification based analyzer. Those malicious training behavior data will contain attack behavior, malicious access and dissemination behavior data of mobile malware and the normal network access data should also have the similar types of behavior data such as behavior of file downloading, accessing website, normal file upload and so on. The behavior classification module is used to divide the training data set into three subsets according to the behavior characteristic: dissemination behavior subset, malicious access behavior subset and attack behavior data subset. Then these three data subsets are used to train three Naïve Bayesian classifiers: attack behavior classifier F1, malicious access classifier F2 and dissemination behavior classifier F3 respectively.

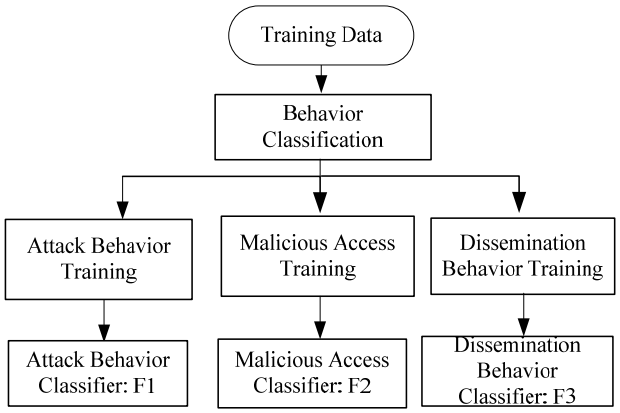

Figure 1. Behavior Classification Training

The procedure of mobile malware behavior detection with those above behavior classification based analyzer is shown in Fig. 2.

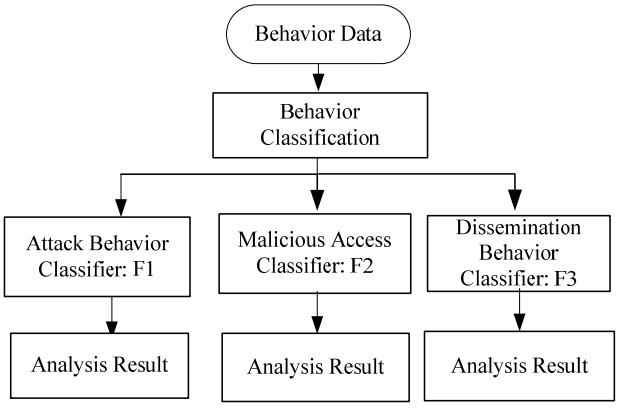

Figure 2. Behavior Classification Detection

During the detection phase, the behavior data from real mobile internet is input to the behavior classification based analysis system. The Behavior Classification module is used to divide the network behavior data into three types of subset according to the behavior characteristic. Then the corresponding malicious behavior classifier is chosen to analyze the behavior data to decide whether they are malicious behavior.

B. Naïve Bayesian Analysis

D is the training dataset of the network behavior data with data tuple X. Every data tuple is a k dimensions attribute vector $X=\{x_1,x_2,\ldots\ldots,x_k\}$. The given network behavior X can be divided into n categories: $C=\{C_1, C_2,\ldots\ldots,C_n\}$. In the Naïve Bayesian analysis on the three types of network behavior, the malicious behavior would be discriminated from the normal action according to category support probability. The number of the category can be set n=2. The category support probability of the network behavior data X belonging to category C_i can be computed as in (1):

$$P(C_i \mid X)=\frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{1}$$

Bayesian algorithm determines X is the category of C_l with the maximum probability:

$$P(C_l \mid X) > P(C_i \mid X) \quad 1 \le i \le n, i \ne l \tag{2}$$

In (1), $P(X)$ is the identical for all the categories, to find the maximum probability $P(C_l \mid X)$ is equivalent to computing the maximum of $P(X \mid C_i)\ P(C_i)$ as in (3) :

$$P(X \mid C_l)P(C_l) > P(X \mid C_i)P(C_i) \quad 1 \le i \le n, i \ne l \tag{3}$$

The category probability $P(C_i)$ can be calculated as in (4) based on the training dataset D:

$$P(C_i)=\frac{|C_{i,D}|}{|D|} \tag{4}$$

$|D|$ is the total of network behavior data X in the training dataset D, and $|C_{i,D}|$ is the number of the network behavior X of category C_i in the training dataset D. The $P(X \mid C_i)$ can be computed by calculating the attribute probabilities: $P(x_1 \mid C_i)$, $P(x_2 \mid C_i)\ldots$, $P(x_k \mid C_i)$ respectively since the feature attributes of network behavior are independent:

$$P(X \mid C_i)=\prod_{l=1}^{k} P(x_l \mid C_i) \tag{5}$$

C. Incremental Self-learning Method

The network behavior of mobile malware always changing on the mobile internet and the behavior-classification based Naïve Bayesian Classifiers is trained with the limited normal and malicious behavior training data which cannot adapt the changeful network behavior. The new malicious and normal behavior can be employed to train the behavior-classification based classifier to optimize the detection accuracy gradually. An incremental self-learning method can improve the training performance which allows training the classifiers only with the new behavior data. The self-learning method of the classifiers means to add the new dataset $D_T^{'}$ belonging

to category C^t into the training dataset D. Then the total dataset become $D^{'}$ as in (6):

$$D^{'} = D + D_T^{'} \tag{6}$$

The incremental self-learning method only adopt the new dataset $D_T^{'}$ containing normal and malicious behavior as in (7) to train the classifiers

$$D_T^{'} = \{< X^t, C^t >\} \tag{7}$$

The incremental self-learning will change the category probability $P(C_i)$ and conditional probability $P(X \mid C_i)$ which can be computed as in (8) and (9) respectively:

$$P^{'}(C_i) = \frac{1 + count\,(C_i) + count^{'}(C_i)}{|C| + |D| + |D_T^{'}|} \tag{8}$$

$$P^{'}(x_j | C_i) = \frac{1 + count\,(C_i \wedge x_j) + count^{'}(C_i \wedge x_j)}{|A_i| + count\,(C_i) + count^{'}(C_i)} \quad 0 \le i \le n, 1 \le l \le k, 1 \le j \le |A_l| \tag{9}$$

$count(C_i)$ is the number of the network behavior belonging to the category C_i in the training dataset D. $count^{'}(C_i)$ is the number of the network behavior belonging to the category C_i in the new incremental training dataset $D_T^{'}$. $|C|$ is the total number of the network behavior category in the training dataset D. $|D|$ is the total number of the network behavior in the training dataset D. $|D_T^{'}|$ is the total number of the network behavior in the new incremental training dataset $D_T^{'}$. $count(C_i \wedge x_j)$ is the number of the network behavior belonging to the category C_i and the values of its attribute A_i is x_j in the training dataset D. $count^{'}(C_i \wedge x_j)$ is the number of the network behavior belonging to the category C_i and the values of its attribute A_i is x_j in the training dataset $D_T^{'}$. $|A_l|$ is the number of value of attribute A_l.

IV. The System Design and Implement

A mobile malware network behavior analysis system based on the behavior classification Naïve Bayesian method is designed whose structure is shown as Fig. 3. The system is deployed on the mobile internet and analyzes the network behavior of mobile terminal. In the training phase, the training dataset captured from the real mobile internet contains three types of mobile malware behavior and known normal network action are chosen to train three types of Naïve Bayesian classifiers: attack behavior classifier F1, malicious access classifier F2 and dissemination behavior classifier F3 respectively. In the detection phase, the Data Capturing module will capture the network traffic from mobile internet and parse the behavior data. The Whitelist Detection engine and Signature-based Detection are employed to identify the normal action and known malicious behavior. The behavior of normal access and known mobile malware do not need to be analyzed by the behavior classification based malicious malware analyzer and can be removed. Then the remaining data is suspect behavior data which is sent to Behavior Classification module to determine the category of the behavior and then one of the three classifiers is chosen to analyze whether the network behavior is malicious behavior.

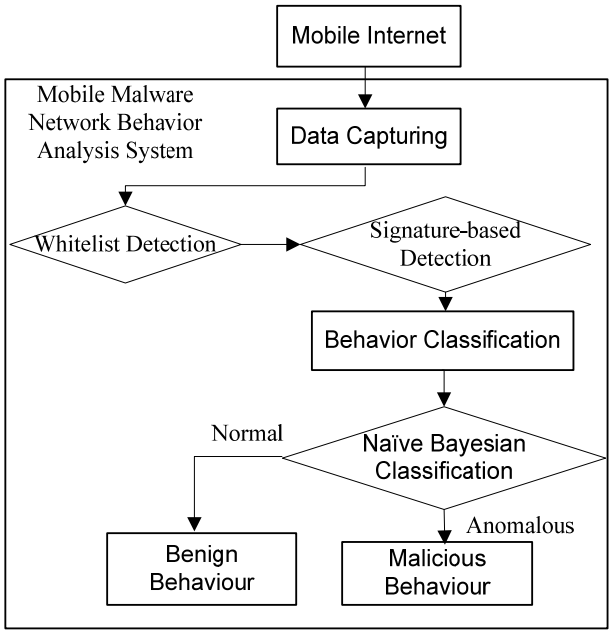

Figure 3. Behavior Classification Based Mobile Malware Network behavior Analysis System

The normal behavior output from the Whitelist Detection engine and malicious behavior from the Signature-based detection engine can be chosen as the training dataset of incremental self-learning for the behavior classification based Naïve Bayesian analyzer. The procedure of the incremental self-learning is shown as Fig.4.

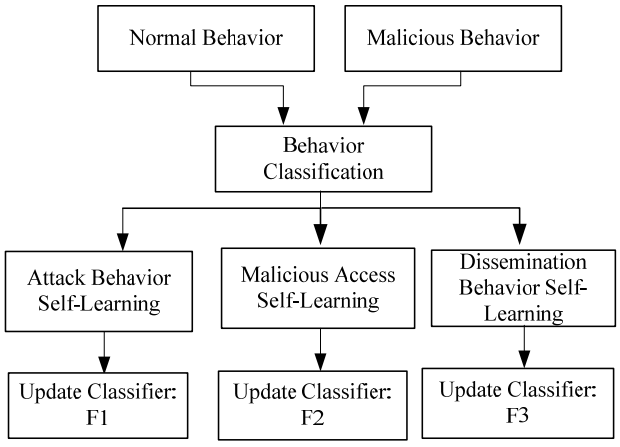

Figure 4. Behavior Classification based Self-learning

The new behavior data of normal access and mobile malware is divided into three types of training subset by Behavior Classification: attack behavior and normal action, malicious access and normal access and the data

subset of dissemination behavior and normal file downloading. The new three training data subset are used to train the attack behavior classifier F1, malicious access classifier F2 and dissemination behavior classifier F3 respectively. In the end, those three revised classifiers are updated into the network behavior analysis system of mobile malware according to the result of the incremental self-learning.

V. Experiment Result

A. Introduction of Experiment.

The simulation is made in the PC server with dual CPU Intel Xeon E5620 and 16GB RAM. The training data set *D* contains normal access behavior and three types of malicious behavior data subset: attack behavior subset D1, malicious access D2 and propagation behavior D3 that is used to train the three Naïve Bayesian classifiers: attack behavior classifier F1, malicious access classifier F2 and dissemination behavior classifier F3. The detection accuracy of the three classifiers and the average detection accuracy are compared with that of traditional two-category Naïve Bayesian classifier. The performance effect on the classifier of those training data subset with 10, 20, 30 kinds of malicious malwares behaviors are also studied subsequently.

Every training data subset contains 100,000 malicious network behaviors data of mobile malware and 200,000 network behaviors of normal access or file downloading. And each of the trained Naïve Bayesian classifiers is used to detect 300,000 real mobile internet test data subset containing corresponding behavior respectively. For example, dissemination behavior classifier F3 is used detect the test data subset containing network data of malicious dissemination behavior and normal file downloading. The test data subset contains 100,000 malicious behaviors of 40 types of malwares and 200,000 normal network action data.

The detection accuracy (*DA*) of mobile malware behavior analysis with the proposed classifiers is calculated as in (10):

$$DA = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Herein, TP (true positive) is the number of network behavior of mobile malware is correctly identified. TN (true negative) is the number of normal network behaviors is correctly detected. FP (false positive) is the number of network behaviors of mobile malware are classified as normal network behavior. FN (false negative) is the number of normal network behavior is classified as network behaviors of mobile malware.

B. Result of 10 Types of Malware Behavior Classification Analysis.

Detection accuracy comparison results of those behavior classification based analyzers trained by behavior data containing 10 types of mobile malware are shown as Fig. 5. S_10_1 is the detection accuracy of the classifier F1 trained by data containing 10 types of mobile malware attack behavior and similar normal action; S_10_2 is that of the classifier F2 trained by data containing 10 types of malicious access behavior and normal network access; And S_10_3 is that of the F3 trained by data containing 10 types of propagation behavior and normal file downloading action. S_10_A is the average detection accuracy of the above three kinds of classifiers, and S_10 is the traditional two-category Naïve Bayesian trained by data with the combination set of the above three group of data subsets which are used to detect the combination set of the three groups of real mobile internet behavior data.

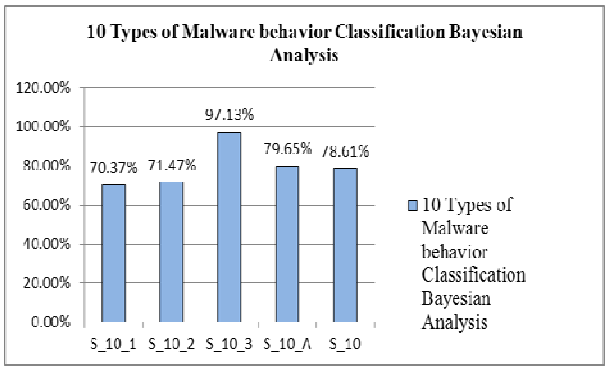

Figure 6. 10 types of malware behavior Classification analysis

In this scenario, the dissemination behavior classifier F3 can achieve the best detection accuracy 97.13% (S_10_3). And the attack behavior classifier F1 has the worst detection accuracy 70.37% (S_10_1). The result show that the proposed behavior classification based analyzer can improve the detection accuracy in the scenario of 10 types of malware. The average detection accuracy is improved to 79.65% (S_10_A) compared with 78.61% (S_10) of two-category Naïve Bayesian.

C. Result of 20 Types of Malware Behavior Classification Analysis.

Detection accuracy comparison results of those behavior classification based analyzers trained by behavior data containing 20 types of mobile malware are shown as Fig. 6.

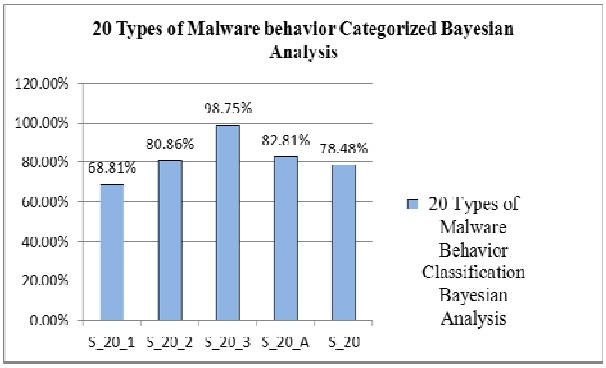

Figure 5. 20 types of malware behavior Classification analysis

Similarly, S_20_1 is the detection accuracy of the classifier F1 trained by data containing 20 types of mobile malware attack behavior and similar normal action; S_20_2 is that of the classifier F2 trained by data containing 20 types of malicious access behavior and

normal network access; And S_20_3 is that of the classifier F3 trained by data containing 10 types of propagation behavior and normal file downloading action. S_20_A is the average detection accuracy of the above three kinds of classifiers, and S_20 is the traditional two-category Naïve Bayesian trained by data with the combination set of the above three group of data subsets which are used to detect the combination set of the three groups of real mobile internet behavior data.

The detection accuracy of F1, F2 and F3 is 68.81% (S_20_1), 80.86% (S_20_2) and 98.75% (S_20_3) respectively. The average detection accuracy is improved to 82.81% (S_20_A) compared with 78.48% (S_20) of two-category Naïve Bayesian.

D. Result of 30 Types of Malware Behavior Classification Analysis.

Detection accuracy comparison results of those behavior classification based analyzers trained by behavior data containing 30 types of mobile malware are shown as Fig. 7.

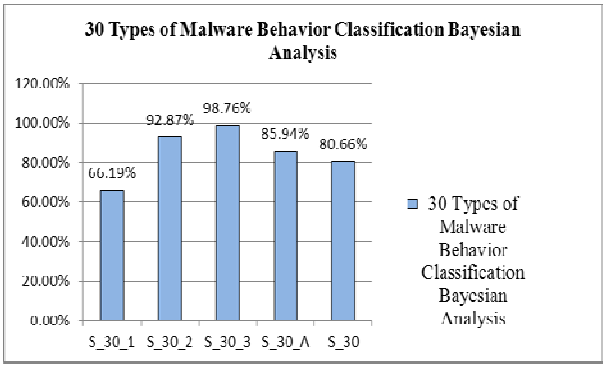

Figure 7. 30 types of malware behavior Classification analysis

The detection accuracy of F1, F2 and F3 is 66.19% (S_30_1), 92.87% (S_30_2) and 98.76% (S_30_3) respectively. The average detection accuracy is improved to 85.94% (S_30_A) compared with 80.66% (S_30) of two-category Naïve Bayesian

E. Comparison on Three Kinds of Malware Behavior Classification Analyzer

Fig. 8 is the comparison results of the detection accuracy of the above three scenarios.

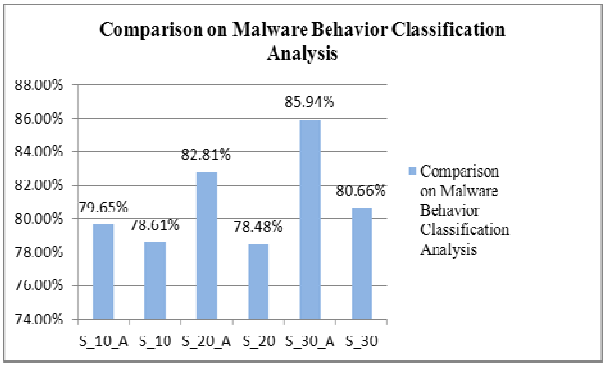

Figure 8. Comparison of different number of malware

The scenario of 30 types of mobile malware can achieve the best detection accuracy 85.94% S_30_A. And the detection accuracy of two-category Naïve Bayesian in the scenarios of training data containing 10 and 20 types of malwares has the worst detection accuracy: 78.61% (S_10) and 78.48% (S_20) respectively.

F. Performance Comparison.

The performance comparison of those behavior classification based analyzers trained by network behavior data containing 10 types of mobile malware is shown in table I. The training speed of the classifier F1, F2, F3 can reach 532,859.68, 190,114.07 and 174,723.35 TPS (Transactions per Second) respectively and they can detect 274,725.27, 311,526.48 and 299,401.20 pieces of test data per second respectively. The average training speed of the three classifiers can reach 233,281.49 TPS and the detection speed can reach 294,406.28TPS. On the other hand, the training speed of the traditional two-category Naïve Bayesian classifier is 209,302.33 TPS and detection speed is 315,015.75 TPS. On the typical mobile internet based on GPRS (General Packet Radio Service) with peak throughput 2Gbps (Giga Bit per Second), the actual network transactions speed is about 14,000 TPS. The proposed behavior classification based analyzers can meet the requirement of real-time process on the mobile internet.

TABLE I.
PERFORMANCE OF 10 TYPES OF MALWARE BEHAVIOR ANALYSIS

No.	Name	Train time[s]	Average Speed [TPS]	Detect time[s]	Average Speed [TPS]
1	S_10_1	0.56	532859.68	1.09	274725.27
2	S_10_2	1.58	190114.07	0.96	311526.48
3	S_10_3	1.72	174723.35	1.00	299401.20
4	S_10_A	3.86	233281.49	3.06	294406.28
5	S_10	4.30	209302.33	2.86	315015.75

The performance comparison of those behavior classification based analyzers trained by network behavior data containing 20 types of mobile malware is shown in table II.

TABLE II.
PERFORMANCE OF 20 TYPES OF MALWARE BEHAVIOR ANALYSIS

No.	Name	Train time[s]	Average Speed [TPS]	Detect time[s]	Average Speed [TPS]
1	S_20_1	1.60	187265.92	0.81	369458.13
2	S_20_2	1.64	183262.06	0.81	369913.69
3	S_20_3	1.61	186219.74	0.82	364520.05
4	S_20_A	4.85	185567.01	2.45	367947.67
5	S_20	4.16	216346.15	2.40	375312.76

The training speed of the classifier F1, F2, F3 can reach 187,265.92, 183,262.06 and 186,219.74 TPS

(Transactions per Second) respectively and they can detect 369,458.13, 369,913.69 and 364,520.05 pieces of test data per second respectively. The average training speed of the three classifiers can reach 185,567.01TPS and the detection speed can reach 367947.67 TPS. On the other hand, the training speed of the traditional two-category Naïve Bayesian classifier is 209,302.33 TPS and detection speed is 315,015.75 TPS. They can meet the real-time process requirement on the mobile internet.

The performance comparison of those behavior classification based analyzers trained by network behavior data containing 30 types of mobile malware is shown in table III. The average training speed of the three classifiers can reach 181,744.75TPS and the detection speed can reach 302,622.73TPS. On the other hand, the training speed of the traditional two-category Naïve Bayesian classifier is 213,118.64 TPS and detection speed is 286,898.31TPS. They can meet the real-time process requirement on the mobile internet.

TABLE III.
PERFORMANCE OF 30 TYPES OF MALWARE BEHAVIOR ANALYSIS

No.	Name	Train time[s]	Average Speed [TPS]	Detect time[s]	Average Speed [TPS]
1	S_30_1	1.68	178997.61	0.93	323275.86
2	S_30_2	1.64	183486.24	1.07	280373.83
3	S_30_3	1.64	182815.36	0.98	307377.05
4	S_30_A	4.95	181744.75	2.97	302622.73
5	S_30	4.22	213118.64	3.14	286898.31

The performance comparison of the three scenarios is shown in table IV.

TABLE IV.
PERFORMANCE COMPARISON OF THE THREE SCENARIOS

No.	Name	Train time[s]	Average Speed [TPS]	Detect time[s]	Average Speed [TPS]
1	S_10_A	3.86	233281.49	3.06	294406.28
2	S_10	4.30	209302.33	2.86	315015.75
3	S_20_A	4.85	185567.01	2.45	367947.67
4	S_20	4.16	216346.15	2.40	375312.76
5	S_30_A	4.95	181744.75	2.97	302622.73
6	S_30	4.22	213118.64	3.14	286898.31

The network behavior analyzer in scenario of 20 types of mobile malware can achieve the highest detection speed. The detection speed of the behavior classification based analyzers can reach 367,947.67 TPS and the detection speed of the traditional two-category Naïve Bayesian classifier in this scenario can reach 375,312.76 TPS. The Training speed of the behavior classification based analyzers trained by the data containing 10 types of mobile malware is the fastest which can reach 233,281.49 TPS. The behavior classification based analyzers trained by behavior data containing 30 types of mobile malware has the best detection accuracy. If it is chosen in the detection of mobile malware on mobile internet, it can also achieve very high speed of training and detection. They can conduct training and analyze the network behavior of mobile internet in real time.

VI. CONCLUSION AND FUTURE WORK.

A behavior classification based mobile malware detection method is proposed to analyze the network behavior of the new or metamorphic mobile malware which is improved gradually with an incremental self-learning method. The malicious network behavior data is divided into three categories: attack behavior, malicious access behavior and dissemination behavior based on the behavior characteristic of mobile malware. A behavior classification module is used to distinguish these three types of behaviors data in the training and detection phase. In the training phase, the network behavior data containing the three types of malicious behaviors and normal access action are used to train the attack behavior classifier F1, malicious access classifier F2 and Dissemination Behavior F3 respectively. Those classifiers are used to detect the corresponding type of network behavior data. The detection accuracy of the behavior classification based analyzers is compared with the traditional two-category Naïve Bayesian classifier. At the same time, detection accuracy of those proposed analyzers are compared by changing the number of malware in the training data set. The experiment result shows that detection accuracy can be improved to 85.94% with the behavior classification based analyzers compare with 80.66% of two-category classifier in the scenarios of the analyzers trained by data containing 20 and 30 types of malwares, the proposed behavior classification based analysis system can detect 367947.67 and 302,622.73 transactions behavior per second (TPS) respectively which can meet the real-time process requirement on mobile internet.

Future Work includes evaluating more abnormal behavior analysis algorithm, optimal behavior features selection and more elaborate behavior categorization method.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Dr. Li Ming Zhu for his support to our research work.

This paper is supported by Corporate Technology, Siemens Ltd. China.

REFERENCES

[1] M. Becher, F. C. Freiling, J. Hoffmann, T. Holz, S. Uellenbeck, and C. Wolf, “Mobile Security Catching Up? Revealing the Nuts and Bolts of the Security of Mobile Devices,” in Proceedings of IEEE Symposium on Security and Privacy, May 2011, pp.96-111.

[2] P. M. Milligan and D. Hutcheson, “Business risks and security assessment for mobile devices,” in MCBE’07:

Proceedings of the 8th Conference on 8th WSEAS Int. Conference on Mathematics and Computers in Business and Economics. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2007, pp. 189–193.
[3] D. Venugopal, G. N. Hu, Efficient signature based malware detection on mobile devices, Mobile Information Systems, Vol.4 No.1, 2008, pp.33-49.
[4] J. A. Morales, P. J. Clarke, Y. Deng, Testing and evaluating virus detectors for handheld devices, Journal in Computer Virology, vol. 2, no. 2,2006, pp. 135-147.
[5] L. Batyuk, M. Herpich, S.A. Camtepe, K. Raddatz, A.-D.Schmidt, S. Albayrak, "Using static analysis for automatic assessment and mitigation of unwanted and malicious activities within Android applications", Proceedings of 6th International Conference on Malicious and Unwanted Software (MALWARE), 2011, pp. 66 – 72.
[6] T. Blasing, L. Batyuk, A. Schmidt, S. Camtepe, and S. Albayrak, An android application sandbox system for suspicious software detection, Proceedings of 5th International Conference on Malicious and Unwanted Software (MALWARE),2010, pp. 55–62.
[7] M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo, Data mining methods for detection of new malicious executables. Proceedings of IEEE Symposium on Security and Privacy, 2001, pp.38–49.
[8] R. Christian, C. Lim, A. S. Nugroho, M. Kisworo, Integrating Dynamic Analysis Using Clustering Techniques for local Malware in Indonesia, Proceedings of 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies,2010, pp.167-169.
[9] U. Bayer, P. Milani Comparetti, C. Hlauscheck, C. Kruegel, and E. Kirda. Scalable, Behavior-Based Malware Clustering. In 16th Symposium on Network and Distributed System Security (NDSS), 2009.
[10] A.S. Shamili, C. Bauckhage, Alpcan, Tansu, Malware Detection on Mobile Devices Using Distributed Machine Learning, Proceedings of the 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 4348 – 4351.
[11] H. S. Chiang and W. J. Tsaur, Identifying Smartphone Malware Using Data Mining Technology, Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN), 2011,pp.1-6.
[12] A. Bose, X. Hu, K. G. Shin, and T. Park, "Behavioral detection of malware on mobile handsets," in MobiSys '08: Proceeding of the 6th international conference on Mobile systems, applications, and services. New York, NY, USA: ACM, 2008, pp. 225–238.
[13] P. Dominigos, M. PaZZani, on the optimality of the simple Bayesian classifier under Zero-one loss. Machine Learning Vol.29, 1997, pp.103-130.
[14] L. Chen, N. Zhen, Y. H. Guo, M. Xu, Y.T. Hu, Applying Naive Bayesian Incremental Learning In Virus Reporting and analyzing, Computer Applications and Software Vol.27,No.1 (2010),pp.92-95(in Chinese)
[15] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, "Crowdroid: behavior-based malware detection system for Android," in Proceedings of the 1st ACM workshop on security and privacy in smartphones and mobile devices, Chicago, USA, 2011, pp. 15-26.
[16] C. Willems, T. Holz, and F. Freiling, "Toward Automated Dynamic Malware Analysis Using CWSandbox," IEEE Security and Privacy, vol. 5, 2007, pp. 32-39.
[17] G. Portokalidis, P. Homburg, K. Anagnostakis and H. Bos, "Paranoid Android: versatile protection for smartphones," in Proceedings of the 26th Annual Computer Security Applications Conference, Austin, Texas, 2010, pp. 347-356.
[18] P. Berthomé, T. Fécherolle, N. Guilloteau and J.-F. Lalande, "Repackaging Android Applications for Auditing Access to Private Data", 2012 Seventh International Conference on Availability, Reliability and Security (ARES), 2012, pp. 388 – 396.
[19] I. Rassameeroj and Y. Tanahashi, "Various approaches in analyzing Android applications with its permission-based security models," in International Conference on Electro/Information Technology. Mankato, Minnesota, USA: IEEE Computer Society, May 2011, pp. 1–6.
[20] A. Barrera, David and Kayacik, H. Güne¸s and van Oorschot, Paul C. and Somayaji, "A methodology for empirical analysis of permission-based security models and its application to android," Proceedings of 17th ACM conference on computer and communications security. Chicago, Illinois, USA: ACM Press, Oct.2010, pp. 73–84.

Dai-Fei Guo is currently a technical manager of Siemens Corporate Technology, Beijing, China. He was born in China in 1975. He received his Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications in China in 2004. He has gotten his B.S. degree in electronics and information system and master degree in communications and information system from Shandong University, Jinan, China. His recent research interests include wireless network security, anti-malware and data mining technique, etc.

Ai-Fen Sui is currently a senior key expert of Siemens Corporate Technology, Beijing, China. She was born in 1974. She received his Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications in China. Her recent research interests include data mining technique and network security, etc.

Yi-Jie Shi is a PhD candidate in State key laboratory of Networking and Switching Technology at Beijing University of Posts and Telecommunications, Beijing, China. She was born in 1974. Currently, her PhD research is in information security under Prof. Qiaoyan Wen. Her recent research interests include mobile internet Security and cryptography.

Jian-Jun Hu is currently a technical manager of Siemens Corporate Technology, Beijing, China. He was born in 1980. He received his B.S. and M.S. degree in computer science from Peking University. His recent research interests include Application security and network security, etc.

Guan-Zhou Lin is a senior engineer of Siemens Corporate Technology, Beijing, China. He was born in 1985 in China. He has gotten his Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China. His recent research interests include data mining technique and network security.

Tao Guo is a senior engineer of Siemens Corporate Technology, Beijing, China. He was born in 1982. He has gotten his B.S. and M.S. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China. His recent research interests include Mobile internet security and network security, etc.