Situational Awareness Framework for Threat Intelligence Measurement of Android Malware

Mookyu Park[1], Junwoo Seo[1], Jaehyeok Han[1], Haengrok Oh[2], and Kyungho Lee[1*]
[1]*School of Information Security, Korea University*
145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea
{ctupmk, junuseo, one01h, kevinlee}@korea.ac.kr
[2]*Agency for Defense Development(ADD)*
460, Ogeum-ro, Songpa-gu, Seoul, Republic of Korea
haengrok@add.re.kr

Abstract

With the development of the Internet of Things (IoT) technology, various devices are connected to the network. The availability of mobile devices is increasing to remotely control these electronic products. As the importance of mobile devices increases, operating systems such as Android OS and iOS are targeted for cyber attacks. In addition, mobile devices are used to manage business data as well as private areas, including text messages and contacts, so the risk of attack is also increasing. This paper proposes threat intelligence evaluation for mobile malware from the viewpoint of situational awareness by extracting features that can detect Android malware using machine learning algorithms.

Keywords: Situational Awareness, Threat Intelligence, Android Malware, Threat Measurement

1 Introduction

Internet of Things (IoT) technology is being applied to various fields such as safety, transportation, industrial, healthcare and building. According to the Gartner report, the global IoT market is expected to grow at an annual average of 28.8% to $ 300 billion in 2015 to $ 1 trillion in 2020. In particular, smart home, smart city, and smart car were selected as-as future businesses that utilize IoT technology. With the expansion of the IoT technology area, the importance of mobile devices such as smartphones and remotely control them is increasing. Large companies such as Apple and Samsung, which are taking over the smartphone market, are increasing their investments to construct the IoT ecosystem. Apple has begun to embed W1 chips in its smartphones and IoT devices that can connect its products to a single ecosystem. Samsung is working on its own operating system, "Tizen," for smartphones and smart cars[1].

However, since IoT equipment makes the boundary between cyberspace and real world disappear, the damage caused by cyber attacks can be extended to the damage of real world. In the past cyber attack, if software or hardware of the state or enterprise was selected as an attack target, it is now causing secondary damage that seized personal information. A typical threat is mobile malware, such as spyware. Thirty-five percent of attackers are exploiting this attack method and are focused on dealing with users' personal information. In addition, according to RSA's Current State of Cybercrime report published in 2016, mobile fraud in 2013 and 2015 has increased by about 173%[2].

Global companies are talking about the possibility of a cyber attack on mobile users. In particular, they focus on the diversified North Korean cyber attacks. Palo Alto Networks has discovered a malicious

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 9:3 (Sept. 2018), pp. 25-38
*Corresponding author: Kyungho Lee, School of Information Security, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea, Tel: +82-(0)2-3290-4885, Email: kevinlee@korea.ac.kr

app that disguised as a specific app on Google Play, and that the app is targeting Korean users with a Samsung smartphone. McAfee pointed to the North Korean hacker organization "Lazarus" as the backbone of the malicious apps of Android backdoor malware disguised as Bible apps. In addition, security companies in South Korea installed spy apps on 10 smartphone users in the national defense, diplomacy, and security fields in 2016. They are behind the spy apps that steal phone calls, text messages, and contact information. Of hacker organizations. The reason for the increase of attacks on mobile devices is that they acquire Permissions on major personal information and functions through malicious apps, and have a lot of information assets (pictures, travel routes, etc.)[3].

Increased usability of IoT devices increases the usability of mobile devices (ex: smartphones). Because of this phenomenon, an attacker is more likely to attack the mobile device, which could cause damage to the real space, such as personal information leakage. To cope with this situation, it is necessary to not only detect the threats to mobile malicious apps but also to evaluate them. This paper suggests a method for evaluating threats based on feature detection through android malware detection using machine learning from a perspective of situational awareness.

2 Backgrounds

Initially, cyber attacks targeted things that are limited to cyberspace, such as hardware and software. However, recent cyber attacks such as ransomware, cryptocurrency exchange hacking, and information leak on mail or electronic equipment of government official officer are changing in the direction of causing damage to the real world by expanding in cyberspace. This section describes cyberspace layers and the methodology for measuring threats for situational awareness via Android malware.

2.1 Research of Cyberspace Layers

Cyberspace is a virtual space that enables communication of information environment by overcoming the temporal and spatial limits of reality through a virtual network environment composed of electronic devices and electronic spectrum. In this environment, cyberspaces are interdependent of the Internet, networks, embedded processes, people, society, and policy.

The concept of cyberspace is described in a hierarchical structure. Joint Publication (JP) 3-12R US Cyber Operation defines cyberspace as three layers: physical layer, logical layer, and persona layer. The physical layer is the layer where the physical network devices (router, switch etc.) are located geographically or physically. A logical layer is a layer in which network or communication devices located at the physical layer connect each node through a protocol or program for the logical connection. A representative example of this layer is OSI (Open Systems Interconnection) 7 layers. The persona layer is a layer where the electronic service provided by the physical layer and the logical layer influences human beings and is a layer where human beings such as IP address, e-mail, and ID can express themselves in cyberspace[4].

Many studies on cyberspace layer have been done and it is applied as an essential concept to construct cybersecurity policy. D.Clark extended the cyberspace to the information layer and the top layer-people layer in addition to the physical and logical layers mentioned in JP 3-12R. This concept of cyberspace layers describes the information flow in cyberspace by adding an information layer. This provided a basis for the dependence of cyberspace services on the hierarchy. Through this information layer, the Top Layer-people layer is described so that human decisions can be projected into cyberspace[5]. The US Department of Homeland Security defined a cyber ecosystem that extended cyberspace in a 2011 report. The cyber ecosystem is a concept that not only human individuals but also social organizations such as private companies, nonprofit organizations, governments, and electronic devices such as hardware and

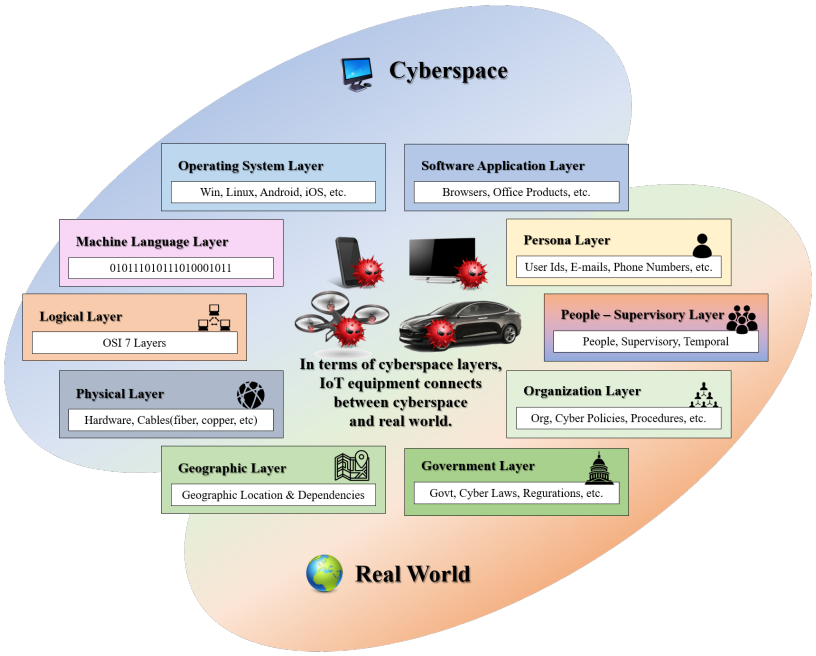


Figure 1: Cyberspace is connected with the real world around human being, and the threat generated in cyberspace can act as a real threat to the real world[6].

software interact to form a virtual ecosystem. This concept reflects the situation that the development of IoT and AI (Artificial Intelligence) technology will rapidly integrate cyberspace and reality. The cyber ecosystem consists of fifteen layers of complex interaction for each element. In particular, these layers show that the actions in cyberspace may ultimately affect social or political systems. This is shown in Figure 1[6].

These cyberspace layers are used in the key terrain of military strategy. N.T.Pantin conceptually analyzed the applicability of key terrain in the cyber domain. This study examined the three layers (physical layer, logical layer, and persona layer) of cyberspace and applied the concept of key terrain. As a result, we confirmed that cyber key terrain requires a certain revaluation. In addition, this research have further confirmed that cyberspace is a unique area, but does not require a specific key terrain definition[7]. The cyberspace layer is also important in policy making. N.Choucri and D. Clark paper used cyberspace layer to conceptualize international relations. This study analyzed the interdependence of cyberspace and international relations using DSM (Domain Structure Matrix) for conceptualization. This research have explained how cyber international relations form influence in the real world[8]. This concept of cyberspace layer is an essential concept in situational awareness that reflects recent cyber attacks.

2.2 FAIR(Factors Analysis of Information Risk)

The FAIR model is a risk management framework developed by J.A. Jones. This method is intended to understand, analyze and measure information risks. In order to measure the risk, this model is structured so that it can measure the assets and threats by elements differently from the existing risk measurement models. The FAIR model is divided into two components: *LEF(Loss Event Frequency)*, which represents the threat, and *LM(Loss Magnitude)*, which represents the loss of the asset. *LEF* is a variable indicating the frequency of loss and is composed of *TEF(Threat Event Frequency)* and *VUL(Vulnerabilities)*. *VUL* consists of *TCap(Threat Capability)* and *CS(Control Strength)*. *LM* consists of *PL(Primary Loss)* and *SL(Secondary Loss)*. *PL* consists of *PLM(Primary Loss Magnitude)* and *PLEF(Primary Loss Event*

Frequency). *SL* is *SLM(Secondary Loss Magnitude)* and *SLEF(Secondary Loss Event Frequency)*. This can be shown in Figure 2.

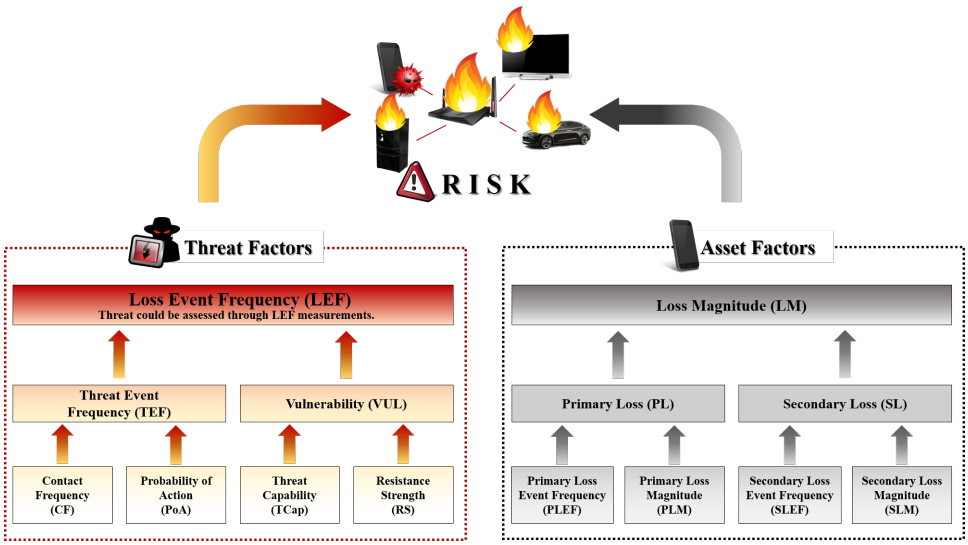


Figure 2: The FAIR model can measure risk based on the frequency of threats to the threat agent and the primary and secondary losses on the assets it owns. This study uses the LEF of the FAIR model for threat assessment[9].

Along with these factors, FAIR needs to set up a *TCom(Threat Community)* to profile the attacker(threat agent) that is a threat to the asset. TCom consists of motive, primary intent, sponsorship, preferred general target characteristics, preferred targets, capability, personal risk tolerance, and concern for collateral damage. These factors are used to profile malicious attackers or attacking activities. Through this process, it is possible to store the attack model and its characteristics against the threat agent[9].

Cyber threats no longer affect cyberspace, but affect the real world as well. The risk measurement through the FAIR model is a suitable method for appropriate situational awareness of these changes. The reason for this is that it is possible to measure not only the primary damage but also the additional possible damage. This paper proposes a situational awareness framework based on the detection results for Android malware, which reflects the results of comprehensive calculations of threat.

3 Related Works

As the importance of cyberspace has been emphasized by the development of IoT environment and AI technology, the concept of situational awareness which is concentrated in existing physical space has been extended to cyberspace. In addition, research on malware detection is actively under way in order to utilize this frame technically. This section describes studies on situation awareness and malware detection.

3.1 Situational Awareness

Situational awareness means recognizing environmental factors in the time and space where a situation occurs and responding to future threats. This method is widely used as a framework for recognizing and responding to threats in the private sector such as terrorism, security, disaster and cyber security. From a military point of view, SA recognizes threats, including all the environmental factors that affect its

assets and enemy forces. Through this analysis, the military-oriented SA provides the current status of the cyberspace operation environment and provides forecast data for the future.

A representative framework for the situational awareness is based on Endsley's model. The Endsley's model consists of a recognition of the elements of the environment within the volume of time and space, an understanding of the meaning, and a process of projecting the state in the near future. Endsley's model consists of three levels: perception, comprehension, and projection. The perception is a level that recognizes the status and attributes of related elements in the environment. The comprehension level is the step of synthesizing the elements of the perception level by analyzing and evaluating the situation. Projection level predicts how information analyzed at comprehension level will affect the state of the future operating environment over time[10].

This Endsley's model has been developed for various environments and goals. J.R.Beyod developed an OODA (Observe, Orient, Decide, Act) model that focuses on cognitive decision making as in Endsley's model. OODA is a process that supports decision making for dynamic environments[11]. A.N.Steinberg et al. studied JDL DFM (JDA Data Fusion Model) combining processing, data fusion, and situational awareness. The JDL DFM consists of a structure that predicts and evaluates the environment under observation based on the collected data under certain circumstances. This model has the advantage of handling large amounts of data, such as network traffic[12].

J.Okolica et al developed a CSAM (Cyber Situational Awareness Model) that reflects business continuity planning. This model aims to build an automation engine that updates the environment in real time and predicts future threats through sense, evaluates and assess[13]. G.P.Tadda and J.S.Salerno developed SARM (Situational Awareness Reference Model) that combines Endsley's model with JDL DFM to improve understanding of the data. The SARM has the advantage of being able to respond flexibly to changing threats in real time[14]. N.Evancich et al studied ECSA (Effective Cyber Situational Awareness), a situational awareness through network monitoring. ECSA is divided into three stages: Network Awareness, Threat Awareness, and Operational Awareness. Network awareness is the step of recognizing the charity and security characteristics of the network. Threat awareness is the step of detecting the attack vectors for the network. Operational awareness is a measure of the impact of an attack on network operational capability. ECSA is an improved situational awareness model than CSAM in decision making, collaboration, and resource management[15]. Table 1 summarizes these results.

Table 1: Situational awareness models have been developed to provide quantitative indicators in decision making[16].

Year	Model	Focus	Proponent
1995	SAM (Situational Awareness Model)	Cognitive decision making	M.R.Endsley
1996	OODA Loop (Observe, Orient, Decide, Act)	Cognitive decision making	J.R.Boyd
1998	JDL DFM (JDL Data Fusion Model)	Processing and fusion of data and SA	A.N.Steinberg et al
2009	CSAM (Cyber Situational Awareness Model)	Business continuity planning and CSA	J.Okolica et al
2010	SARM (Situation Awareness Reference Model)	Situational awareness	G.P.Tadda & J.S.Salerno
2014	ECSA (Effective Cyber Situational Awareness)	CSA in computer networks	N.Evancich et al

In addition, the Cyberware project consists of the cyber situational awareness framework with the asset, configuration, impact, threat, and visualization. In this framework, the recognition of assets means tools or techniques that can automatically extract the latest information. The configuration means to automatically extract dependencies between mission modeling and asset and mission. The impact is a framework for automatically assessing the impact of an attack through a mission and analyzing the correlation through it. The threat is a way of predicting future behavior by recognizing the type and purpose of the enemy. The visualization generates the results for the mission in semantic perspective.

Each of these elements is assigned a task within the framework, which is then measured, and the resulting values are combined to provide context awareness. In addition, by predicting what threats to the system, decision makers can provide effective countermeasures.

3.2 Research of Malware

Most mobile malware detection has been studied through various machine learning applications. These studies are focused on improving training data configuration and detection algorithms. Zarni Aung and Win Zaw performed clustering using K-means clustering, and performed classification using J48 Decision Tree Algorithm, Random Forests (RF), and Classification and Regression Tree (CART). They select the k best feature among the various permission features drawn from the Android apk file. The k best features are selected using the feature selection method: Information Gain. This method depends on the entropy of the features and chooses the highest value of gain as the best feature. For 500 sample android applications, RF showed the highest accuracy of 0.918 and CART showed the lowest accuracy of 0.849[17].

Wu et al. proposes a static feature-based mechanism, DroidMat, to provide a static analysis paradigm for detecting android malware. DroidMat clustering with K-means algorithm by setting information such as permission, intent, etc. in the manifest file as feature. They use the kNN algorithm to classify android malware as benign or malware. To find optimal efficiency, they experimented with two series. First, they conducted experiments using several feature sets. In the paper, they experimented with four feature sets: intent + API, without intent + API, permission + API, and permission + intent + API. When viewed in precision's view, the dataset including all of the permissions, intents, and APIs showed the highest precision. Second, several different clustering and classification algorithms are combined. In the paper, EM + kNN, K-means + kNN, EM + Naïve Bayes and K-means + Naïve Bayes were tested with four sets of algorithms. The K-means + kNN algorithm set showed a best result in terms of both recall and precision's view. The detection accuracy reaches 97.87%[18].

Schultz et al. compared four machine learning algorithms (signature method, RIPPER, Naïve Bayes, Multi-Naïve Bayes) trained with three features: a program-generated DLL and a system call, strings found in a program binary, and a raw hexadecimal representation of the binary. The Multi-Naïve Bayes method showed a higher detection rate of 97.76% than any algorithm, which is twice the detection rate of signature-based methods[19]. Justin Sahs and Latifur Khan attempted to detect Android malware using One-Class Support Vector Machine algorithm. They used permissions and CFGS of the input applications as features. For 2081 benign and 91 malicious Android applications, true negative was 0.5 and true positive was 0.9[20].

These researches contribute to the detection effectiveness or efficiency of malware, but they have limitations in determining the criteria for threats. Therefore, this study reflects the threat assessment to the situational awareness so that it could be expanded from the threat detection to become the standard of decision making.

4 Process of Threat Assessment in Situational Awareness

Most situational awareness models use qualitative methods to classify threats and judge the current situation. However, these qualitative methods tend to depend on heuristic techniques that utilize human perception and experience. This paper proposes a threat assessment scheme that minimizes these qualitative judgments. This paper measures threat based on malware detection and the FAIR model of risk measurement. In addition, this paper applies the Gaussian Mixture Model (GMM) to the proposed model to supplement existing qualitative methods.

The basic situational awareness model follows the four steps of Endsley's model: State of the Environment, Situational Awareness, Decision and Performance of Action. This study excludes the first step, "State of the Environment", as research is organized assuming that we have collected Android malware. This study focuses on the second step "Situational Awareness" and consists of *Malware Detection* stage using semi-supervised learning and *Threat Assessment* stage through FAIR model applying K-Means algorithm. In addition, for the semantic visualization, *Decision-Making Awareness*, which is the threat grade optimization step through GMM (Gaussian Mixture Model), was constructed. This is shown in Figure 3.

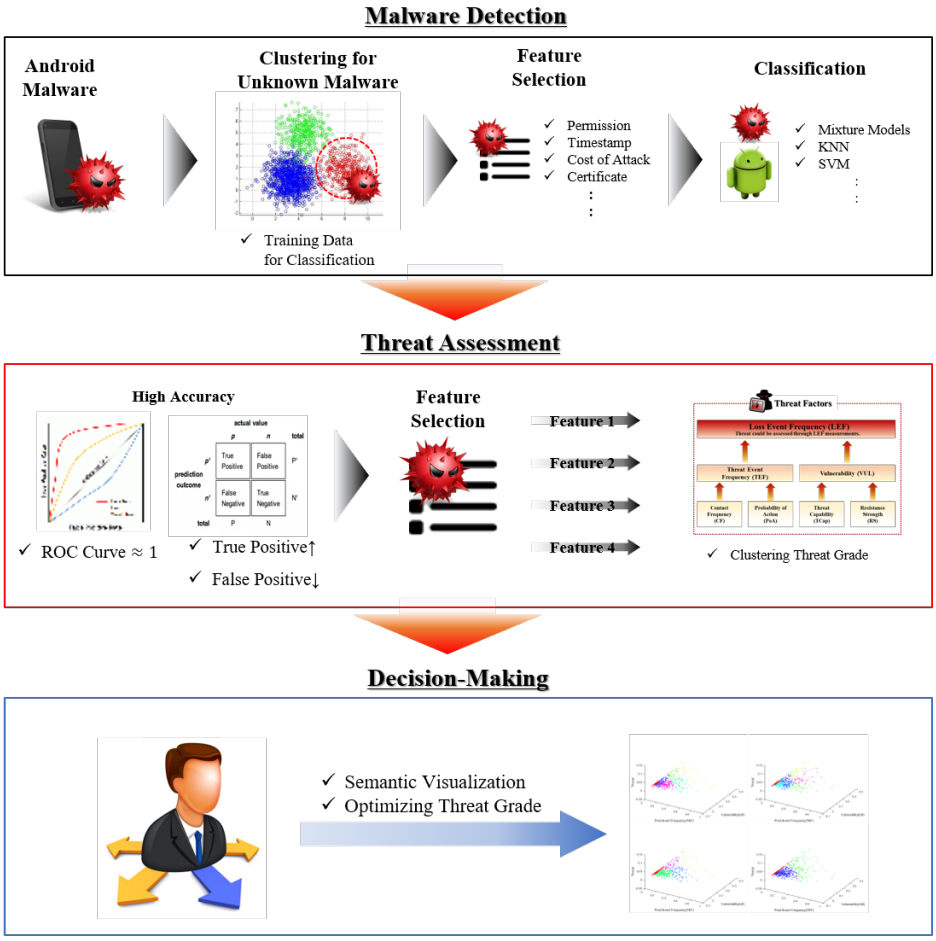


Figure 3: In order to make decisions about an unknown attack, the situational awareness proposed in this study is composed of three steps of *Malware Detection*, *Threat Assessment*, and *Decision-Making* applying the machine learning method.

4.1 Malware Detection

As the cyber attack becomes intelligent, the unknown attack is the biggest threat than the known attack in the real Android environment. However, existing rule-based algorithms and classification algorithms that require learning data have limitations in distinguishing new threats. In order to solve this problem, it is necessary to update training data through clustering in order to judge whether it is a malicious code when given data. Based on the updated training data, feature selection and classification are used to

extract features showing high detection results (ROC, True Positive). These features are used as data for each element of the FAIR model for threat measurement[21].

4.2 Threat Assessment

The features extracted from the Android Malware Awareness phase are applied to the *LEF(Loss Event Frequency)*, an indicator of the threat among the elements of the FAIR model. *LEF* is a factor that indicates the frequency with which a threat is a loss to an asset. This element consists of *TEF(Threat Event Frequency)* and *VUL(Vulnerabilities)*. The *VUL* consists of *TCap(Threat Capability)* and *CS(Control Strength)* which indicates difficulty in attack success. The existing FAIR model measures the risk by measuring each factor from "Very High" to "Very Low" in five grades. However, in this case, numerical values or variables that are not reflected in the risk may arise. This study clusters threats with the K-Means algorithm to overcome these limitations. The data used for clustering are f_N, features matching *TEF* and *VUL*. We cluster these N data with K which is the number of threat classes with D dimensions. If there are five grades from "Very High" to "Very Low", K becomes 5. if the n th f_n corresponds to $k(k = 1, \cdots, K)$ clusters, $r_{nk} = 1$, and $r_{nk} = 0$ otherwise.

$$r_{nk} = \begin{cases} 1 & \text{if } k = argmin_g ||x_n - \mu_g||^2 \\ 0 & \text{Otherwise} \end{cases}$$

Thus, the distortion measure function, which is an objective function for the threat grade, is given by the following equation(1).

$$G = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||f_n - \mu_k||^2 \tag{1}$$

At this time, an EM(Expectation Maximization) algorithm, which is an iterative procedure, is used to minimize the function G value. The iterative procedure obtains the convergence value by repeatedly fixing μ_k and r_{nk}. In this case, the G value is the minimum value for μ_k, so if using the derivative, μ_k can cluster the threat grade as shown in equation(2)[22][23].

$$\mu_k = \frac{\sum_n r_{nk} f_n}{\sum_n r_{nk}} \tag{2}$$

4.2.1 Decision-Making

If semantic visualization is difficult even in a cluster through the K-Means algorithm, there may be a limit in decision-making. To reduce this uncertainty, this paper optimizes clustering class through GMM. Assuming that any complex probability distributions are combined into the number of k Gaussian distributions, the GMM can statistically deduce the characteristics of the subgroups. The performance measure of the GMM is the log-likelihood function. In this study, GMM aims at parameter estimation that maximizes the probability of G for threats classified in the *Threat Awareness* stage. The log-likelihood function is shown in equation(3). In the log-likelihood function, the parameter θ represents the mean μ_k, covariance Σ_k, and the probability of the distribution of π_k in the Gaussian distributions.

$$ln\ p(G|\pi, \mu, \Sigma) = \sum_n ln \sum_k N(g_n|\mu_k, \Sigma_k) \tag{3}$$

Since GMM is difficult to estimate jointly update, this method uses EM algorithm. In other words, this study is an alternative update that fixes π to minimize the log-likelihood function, calculates mu, Σ, fixes

μ, Σ and computes π. The equations for threat grade are (4), (5), and (6). At this time, z is a latent variable that makes it easy to calculate the optimization[24][25].

$$\mu_k = \frac{1}{\sum_n pz_k = 1|g} \sum p(z_{nk} = 1|g) g_n \tag{4}$$

$$\Sigma_k = \frac{1}{\sum_n P(z_k = 1|g)} \sum_n p(z_{nk} = 1|g)(g_n - \mu_k)(g_n - \mu_k)^T \tag{5}$$

$$\pi_k = \frac{\sum_n P(z_k = 1|g)}{n} \tag{6}$$

This paper proposed the situational awareness model for Andorid Malware as three steps of *Malware Detection*, *Threat Assessment*, and *Decision-Making Awareness*. The next section explains the results through the analysis of Android malware.

5 Result

Attacks on most mobile devices are aimed at information leakage, illegal surveillance, and eavesdropping on users. However, as the connectivity of mobile devices grows around IoT technologies, various threats could occur[26]. This study assumes that information leakage occurs when a malicious app is installed on a mobile device from the perspective of a socio-technical system. This section applies a process of situational awareness as described above to Android malware to rank threats to decision making.

5.1 Result of Malware Detection

This study utilized malicious app data extracted from a static analysis in research of J.Jang et al. This data is based on the working principle of Andro-AutoPsy and utilizes extracted Android app data related to malicious behavior. This data includes APIs related to malicious behavior, a list of system commands, and so on. The data consists of 2000 app files, 1500 normal apps, and 500 malicious apps[27]. Based on this research, this study was analyzed based on KNN (K-Nearest Neighbors), SVM (Support-Vector Machine) and Logistic Regression.

In this paper, KNN analysis detects malicious apps by distance measurement and distance weight change. The four KNNs used in this study are basic KNN using Euclidean distance measurement, Cosine KNN using Cosine distance measurement, Cubic KNN using Minkowski space distance measurement, and Weighted KNN with distance weighting as the inverse square. In the case of SVM, the kernel function is modified from the first to the third, and the kernel function is measured by applying the Gaussian function. The confusion matrix was used for the measured results.

5.1.1 Confusion Matrix

The confusion matrix is expressed as an error matrix and is mainly used as an index for evaluating algorithm performance. The confusion matrix consists of TP(True Positive), FN(False Negative), FP(False Positive) and TN(True Negative). TP is a case of predicting the normal behavior of the app as a normal behavior. FN is a case of predicting normal behavior as malicious behavior. FP is a case of predicting malicious behavior as normal behavior, and TN is a case of predicting malicious behavior as malicious behavior. Through this, it is possible to measure TPR(True Positive Rate), PPV(Positive Predictive Value), FPR(False Positive Rate), FDR(False Discovery Rate), Accuracy and F1 Score. The equations for each index are as follows, and the results are shown in Table 2[28].

$$
\begin{aligned}
TPR &= \frac{TP}{TP+FN} \\
PPV &= \frac{TP}{TP+FP} \\
FPR &= \frac{FP}{FP+TN} \\
FDR &= \frac{FP}{FP+TP} \\
Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\
F_1 &= \frac{2TP}{2TP+FP+FN}
\end{aligned}
$$

Table 2: Detection of Android Malware showed the highest accuracy of SVM series and logistic regression, and features of permission and intent were extracted.

Indicators	KNN	Cosine KNN	Cubic KNN	Weighted KNN	Linear SVM	Quadratic SVM	Cubic SVM	Gauss SVM	Logstic Regression
TP	0.990	0.980	0.990	0.990	0.990	0.990	0.990	0.990	0.990
FN	0.140	0.100	0.190	0.090	0.010	0.020	0.020	0.170	0.010
FP	0.010	0.020	0.010	0.010	0.010	0.010	0.010	0.010	0.010
TN	0.860	0.900	0.810	0.910	0.990	0.980	0.980	0.830	0.990
TPR	0.876	0.907	0.839	0.917	0.990	0.980	0.980	0.853	0.990
PPV	0.990	0.980	0.990	0.990	0.990	0.990	0.990	0.990	0.990
FPR	0.011	0.022	0.012	0.011	0.010	0.010	0.010	0.012	0.010
FDR	0.010	0.020	0.010	0.010	0.010	0.010	0.010	0.010	0.010
Accuracy	0.925	0.940	0.900	0.950	0.990	0.985	0.985	0.910	0.990
F1 Score	0.930	0.942	0.908	0.952	0.990	0.985	0.985	0.917	0.990

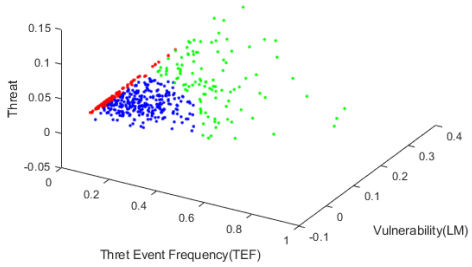
Based on these measurement results, algorithms with high accuracy are Weighted KNN, SVM series (Linear SVM, Quadratic SVM, Cubic SVM) and Logistic Regression. In addition, 136 features corresponding to permission and intent were extracted. This paper grades threat through a probability distribution of these features.

5.2 Result of Threat Assessment & Decision-Making

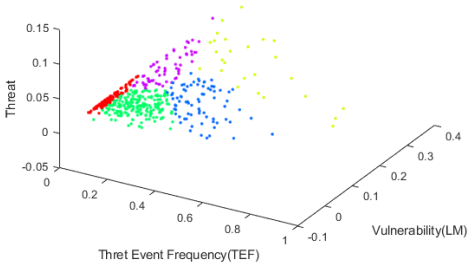
Detection of malicious apps can determine whether an app is a threat, but there is a limit to the degree of threat. This process has a negative effect on decision-makers in preparing countermeasures. This paper uses factors of *LEF(Loss Event Frequency)* of FAIR model to classify the threat of mobile malware in order to compensate for the weakness of this threat detection. *LEF* is composed of *TEF(Threat Event Frequency)* and *VUL(Vulnerability)*.

5.2.1 TEF(Threat Event Frequency)

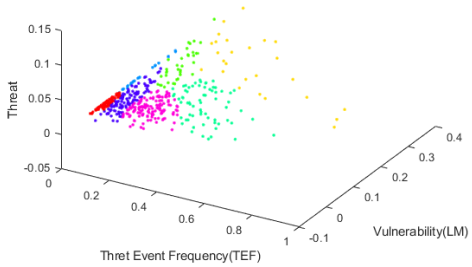
Factors included in the *TEF* are *CF(Contact Frequency)* and *PoA(Probability of Action)*. These factors indicate the frequency of actual threats and the success of malicious apps in the Android environment due to threats. The results of this actual action are integrated through the data used in this study. In particular, it can be assumed that the features of permission represent the probability of an attack by malicious apps. This leads to the finding that the related works, which were presented above, also function as a feature of malicious apps. Based on this evidence, this study defined the *TEF* as a probability distribution of the permissive features.



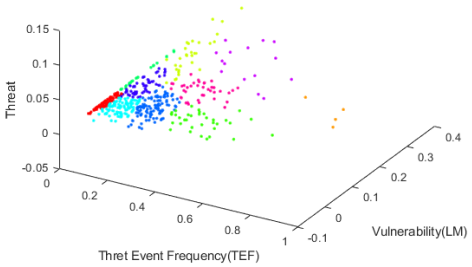
(a) 3 Cluster of Threat Grade



(b) 5 Cluster of Threat Grade



(c) 7 Cluster of Threat Grade



(d) 10 Cluster of Threat Grade

Figure 4: Based on threat clustering through machine learning, decision makers can effectively provide semantic visualization for situational awareness and construct delicate countermeasures.

5.2.2 VUL(Vulnerability)

The elements included in VUL consist of *TCap(Threat Capability)* and *CS(Control Strength)*. *TCap* is defined as the probability of malicious features included in Android malicious apps. The more features that malicious apps have, the higher their ability to respond to threats. In other words, Android malicious apps that have various malicious features can judge that they have various attack methods. *CS* is defined based on the price of Android malware. The data is expressed as the probability of price distribution of Android OS and Any OS among ”Zerodium Payouts for Mobiles”[29][30].

Based on the above definition, *TEF* and *VUL* are expressed as probability distributions. In this case, the probability distribution of *VUL* is calculated as the joint probability of TCap and CS. Finally, threats to Android malware can be evaluated with TEF and VUL combinations. However, in the case of such a combination, it is necessary to classify and optimize it because it has complexity. This study classifies threats based on **K-Means algorithms** and uses **GMM** as a method to optimize them. The results are shown in Figure 4.

There are limitations to classify existing threats into 3 classes (low, moderate, high) or 5 classes (very low, low, moderate, high, very high). However, this study can provide a semantic visualization in that it allows the decision maker to evaluate the threat in detail through the proposed process and methodology.

6 Conclusion

As high-value information such as personal information and financial information is stored and managed on a mobile device, it is subject to cyber attacks. To counter these threats, many studies have focused on effectively detecting malware. However, evaluation is essential to make it intelligence. This study proposed a situational awareness framework based on detection of Android malware. The proposed situational awareness model consists of three steps: *Malware Detection* aimed at recognizing the malicious behavior of the app, *Threat Assessment* to rank it, and *Decision-Making* to optimize threat rating. Existing situational awareness is limited in that it confirms the threat from the cognitive and empirical perspectives of the analyst or decision maker. However, applying the machine learning to Situational Awareness will improve the objectivity of the decision support system. This study visualizes machine learning in each procedure to enable objective and semantic situational awareness. Through this process, this paper could contribute to decision support by evaluating threat intelligence for detecting attacks.

Acknowledgments

This work was supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract. (UD060048AD)

References

[1] J. Rivera, "Gartner identifies the top 10 strategic technology trends for 2014," *Gartner, Inc. Retrieved at March*, vol. 10, p. 2016, 2013.

[2] RSA, "2016: Current state of cybercrime," *White Paper*, May 2016.

[3] D. A. Pinkston, "Inter-korean rivalry in the cyber domain: The north korean cyber threat in the sŏn'gun era," *Georgetown Journal of International Affairs*, vol. 17, no. 3, pp. 60–76, 2016.

[4] C. Operations, "Joint publication 3-12 (r)," *Joint Chief of Staffs*, February 2013.

[5] D. Clark, "Characterizing cyberspace: past, present and future," *MIT CSAIL*, vol. 1, pp. 2016–2028, March 2010.

[6] R. Philip *et al.*, "Enabling distributed security in cyberspace," *Departament of Homeland Security*, March 2011.

[7] N. T. Pantin, "Key terrain: application to the layers of cyberspace," Ph.D. dissertation, Naval Postgraduate School, March 2017.

[8] N. Choucri and D. D. Clark, "Integrating cyberspace and international relations: The co-evolution dilemma," *MIT Political Science Department Research Paper*, vol. 2012, no. 29, November 2012.

[9] J. Jones, "An introduction to factor analysis of information risk (fair)," *Norwich University journal of information assurance*, vol. 2, no. 1, November 2006.

[10] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, March 1995.

[11] J. R. Boyd, "The essence of winning and losing," *Unpublished lecture notes*, vol. 12, no. 23, pp. 123–125, January 1996.

[12] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the jdl data fusion model," in *Proc. of the Sensor Fusion: Architectures, Algorithms, and Applications III (SPIE3719'99), Orlando, FL, United States*, vol. 3719. SPIE, March 1999, pp. 430–442.

[13] J. Okolica, J. T. McDonald, G. L. Peterson, R. F. Mills, and M. W. Haas, "Developing systems for cyber situational awareness," in *Proc. of the 2nd Cyberspace Research Workshop (CRW'09), Shreveport, Louisiana*, June 2009, p. 46.

[14] G. P. Tadda and J. S. Salerno, "Overview of cyber situation awareness," in *Cyber situational awareness*, S. Jajodia, P. Liu, V. Swarup, and C. Wang, Eds. Springer, Boston, MA, 2010, pp. 15–35.

[15] N. Evancich, Z. Lu, J. Li, Y. Cheng, J. Tuttle, and P. Xie, "Network-wide awareness," in *Cyber Defense and Situational Awareness*, A. Kott, C. Wang, and R. F. Erbacher, Eds. Springer, Cham, 2014, pp. 63–91.

[16] T. Pahi, M. Leitner, and F. Skopik, "Analysis and assessment of situational awareness models for national cyber security centers," in *Proc. of the 3rd International Conference on Information Systems Security and Privacy (ICISSP'17), Porto, Portugal*, vol. 1. SCITEPRESS, February 2017, pp. 334–345.

[17] Z. Aung and W. Zaw, "Permission-based android malware detection," *International Journal of Scientific & Technology Research*, vol. 2, no. 3, pp. 228–234, March 2013.

[18] D.-J. Wu, C.-H. Mao, T.-E. Wei, H.-M. Lee, and K.-P. Wu, "Droidmat: Android malware detection through manifest and api calls tracing," in *Proc of the 7th Asia Joint Conference on Information Security, Tokyo, Japan*. IEEE, August 2012, pp. 62–69.

[19] M. G. Schultz, E. Eskin, F. Zadok, and S. J. Stolfo, "Data mining methods for detection of new malicious executables," in *Proc. of the 2001 IEEE Symposium on Security and Privacy (S&P'01), Oakland, California, USA*. IEEE, May 2000, pp. 38–49.

[20] J. Sahs and L. Khan, "A machine learning approach to android malware detection," in *Proc. of the 2012 European Intelligence and Security Informatics Conference (EISIC'12), Odense, Denmark*. IEEE, August 2012, pp. 141–147.

[21] K. Wipawayangkool and E. Villafranca, "Exploring millennials' malware awareness and intention to comply with information security policy," *Review of Integrative Business and Economics Research*, vol. 4, no. 3, pp. 153–161, January 2015.

[22] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, June 2010.

[23] G. Hirschberger, T. Pyszczynski, and T. Ein-Dor, "Vulnerability and vigilance: Threat awareness and perceived adversary intent moderate the impact of mortality salience on intergroup violence," *Personality and Social Psychology Bulletin*, vol. 35, no. 5, pp. 597–607, May 2009.

[24] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 381–396, March 2002.

[25] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, July 2015.

[26] F. Arnold, H. Hermanns, R. Pulungan, and M. Stoelinga, "Time-dependent analysis of attacks," in *Proc. of the 3rd International Conference on Principles of Security and Trust (POST'14), Grenoble, France*, ser. Lecture Notes in Computer Science, vol. 8414. Springer, Berlin, Heidelberg, April 2014, pp. 285–305.

[27] J.-w. Jang, H. Kang, J. Woo, A. Mohaisen, and H. K. Kim, "Andro-autopsy: Anti-malware system based on similarity matching of malware and malware creator-centric information," *Digital Investigation*, vol. 14, pp. 17–35, September 2015.

[28] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, January 2011.

[29] A. Delgado, "Zero-day exploits pricing in white and gray markets: a case-study," https://www.academia.edu/22674761/Zero-Day_Exploits_Pricing_in_White_and_Gray_Markets_a_case-study [Online; accessed on August 20, 2018].

[30] K. Sean, Michael, "Zerodium paying up to $500k for mobile messaging app vulnerabilities," August 2017, http://www.eweek.com/security/zerodium-paying-up-to-500k-for-mobile-messaging-app-vulnerabilities [Online; accessed on August 20, 2018].

Author Biography

Mookyu Park received the B.S. degrees from Sejong University in 2014, Currently he is an Ph.D student in the Center for Information Security Technologies(CIST) of Korea University since 2014. His research interests include Situational Awarenee and Risk management.

Junwoo Seo is a B.S student in Korea University since 2015. His research interests include Battle Damage Assessment and Situational Awareness.

Jaehyeok Han received the B.S. degrees from University of Seoul in 2011, M.S. degrees in Information security from Korea University in 2016. Currently he is an Ph.D student in the Center for Information Security Technologies(CIST) of Korea University. His research interests include Digital forensics, Data minig and Risk management.

Haengrok Oh received the Ph.D degrees from Cheongju University in 2008. Currently he work for Agency for Defense Development. His research interests include Cyber defense and Cyber Situational Awareness.

Kyungho Lee received the B.S. degrees from Sogang University of in 1989 , M.S. degrees from Sogang University in 1997, Ph.D degrees from Korea University in 2009. Currently he is Associate Professor in the Department of Cyber Defense and the Center for Information Security Technologies(CIST) of Korea University. His research interests include Risk management, Information security consulting, Privacy policy and Privacy impact assessment.