Multiple Profiles based Ensemble Model for Analytical Classification of Cyber Incident

Hyung-Jin Mun1 and Byoung Yup Lee2*

1Division of Information and Communication Engineering, Baekseok University, Korea; jinmun@gmail.com
2Department of Cyber Security, Pai Chai University, Korea; bylee@pcu.ac.kr

Abstract

Background/Objectives: Cyber incidents collected from security information & event management system are growing rapidly due to expanding malicious code and companies got to collect more data and to use a variety of information with the advent of big data. Methods/Statistical Analysis: It is difficult for cyber incident analysts to extract and classify similar features due to Cyber Attacks. To solve these problems, the analytical classification of cyber incidents formerly generated for one of the profiles from the features of cyber incidents and cyber observable, and by evaluating the degree of similarity based on this profile, similar cyber incident is identified. Findings: Analytical classification from big data of cyber incident requires various features of cyber observables that compose the cyber incident. Therefore, it is necessary to improve classification accuracy of the similarity by using multi-profile which is classified as the same features of cyber observables. When utilizing an ensemble algorithm and grouping similar features, we calculate the similarity, it shows higher accuracy of the classification than it is calculated based on the same criteria. Improvements/Applications: We propose a multi-profile ensemble model performed similarity analysis on cyber incident based on both attack type and cyber observable that can enhance the accuracy of the classification.

Keywords: Classification, Cyber Incident, Cyber Observable, Ensemble Model, Intrusion, Profiles

1. Introduction

An attacker attempts to attack a system in different ways. Even while an Internet user is web surfing, the attack can be made. Due to malicious codes with worms or viruses, various cyber incidents occur. Data of accumulated cyber incidents is increasing exponentially. From the data, information can be obtained by diversified technologies like big data. The information is accumulated by intrusion analyses1–5. As classifying accumulated information, we obtain attack patterns and technology to prevent attacks.

Analytical classification of cyber incidents is a concept which divides each cyber incident with different features. Therefore, it is necessary to derive similar features that can identify cyber incident1–9.

A variety of studies are in progress regarding ways to classify the type of cyber incidents10–13.

Clustering ensemble model offers the ability to classify similar cyber incidents and improves the accuracy and reliability of analytical classification of cyber incidents by combining the strengths of clustering results14–16.

Also, this method demonstrates that combining the clustering results of clustering algorithms with various features yields a better clustering solution than selecting from a single clustering process alone9.

Figure 1 shows the structure of the ensemble algorithm and finds the final solution by combining solutions.

Figure 1. Structure of ensemble algorithm.

Decision making method of the combination algorithm is as shown in Figure 2; the figure shows majority voting and weighted majority voting.
When the number of classes is $M=3$ and the number of classifiers is $T=5$, the class label vector printed out from five classifiers corresponding to an unknown pattern, $x$, is shown. In the majority voting combination algorithm, the second group scores the highest and is classified as $w_2$.

However, in the case of the weighted majority voting combination algorithm considering the weighted value at the third group gets the highest value and is classified as $w_3$. The result comes out because the third classifier, $c_3$, showing high credibility (0.4) chooses $w_3$.

Figure 2. Decision making method of ensemble algorithm.

Clustering ensemble model which analyzes a variety of data sources is more effective than using a single data source by cross-validating cyber incidents with the actual observations.17

2. The Composition of Cyber Incident and Profile

Dataset which used to analyze the cyber incident is composed of heterogeneous columns vertically partitioned. Two data sets with different features are connected by a common unique identifier called the cyber incident. Therefore they can be analyzed by Interconnection

Tables 1 and 2 show the two data sets that are composed of cyber incidents according to the attack type and cyber observable.

| Table 1. Cyber incident based attack types |
|-------------------------------------------|
| Incident | Probe attack | DoS attack | Unauthorized access attack |
|----------|--------------|------------|---------------------------|
| #no_01   | Satan        | Land       | Rootkit                   |
| #no_02   | Saint        | Smurf      | Worm                      |

| Table 2. Cyber incident based cyber observables |
|-----------------------------------------------|
| Incident | IP           | URL         | Hash                      |
|----------|--------------|-------------|---------------------------|
| #no_01   | 1.1.1.1      | www.abc.com | ab2c7defg                 |
| #no_02   | 2.2.2.2      | www.zyx.com | z1yxw9vut                 |

We derive two profiles based on cyber incident information such as attack type and cyber observable.

There is a wide variety of similarity evaluation method to choose from cyber incident analysis system. Therefore, the suitable evaluation method is to be used according to the type of the feature.18,19

This algorithm is necessary for generating a cyber incident profile that is the basis for the similarity evaluation.

Figure 3 shows a profiling algorithm that derives cyber incident which can represent many cyber incidents by the clustering. The profile is composed of the median cyber incidents.

Because it is impossible to quantify the cyber observables, designing the median value of the cyber observables as the profile is difficult. Considering the features of cyber observables and comparing them one another, the most repetitive cyber observable is designated as the profile. Although the representative value of cyber incidents, the profile, is designated as single one, multiple profiles to different cyber incidents can be generated, considering the accuracy of the classification.

Figure 3. A profiling algorithm.

3. Similarity Evaluation Model of Cyber Incident

The similarity between cyber incidents is calculated based on the profiles and then neighbors of cyber incident are found from the results. Figure 4 shows a similarity evaluation algorithm.

By evaluating the distance of cyber observables of cyber incidents, which has already been classified to a new
cyber incident, the similarity is calculated. By calculating the similarity after weighting based on an attack type and cyber observables, it is possible to figure out the degree of danger and to evaluate the similarity, considering features of the cyber observables.

The algorithm we suggest does not calculate the similarity after averaging outputs from the single distance evaluation algorithm to the multiple profiles but evaluate the similarity utilizing the majority voting algorithm to the multiple similarity from the multiple distance algorithm.

4. Multiple Profiles based Ensemble Model

The ensemble algorithm uses majority voting algorithm to the process combining the outputs of multiple classifiers to make a classification result.

We propose this novel two ensemble models that performed similarity analysis on the hierarchically classified attack type and cyber observable based on cyber incident.

Figure 4. Similarity evaluation algorithm.

Figure 5 shows an individual classification model based of a combined profile with attack type and cyber observable.

Figure 6 shows an ensemble model that uses the average similarity of each separated profile with attack type and cyber observable. Figure 6 does not process feature information uniformly but group information by the attribute of the feature information to calculate the similarity of each group in a combined form.

Figure 7 shows a similarity evaluation method that uses an ensemble model of Figure 5.

Our suggested similarity evaluation method in Figure 5 integrates features of each profile into a combined profile used by the individual classification algorithms and calculates the degree of similarity.

Figure 8 shows a similarity evaluation method that uses an ensemble classification model based of separated profile of Figure 6.

Our suggested ensemble classification model in Figure 6 creates the separated profile used by the ensemble algorithms and calculates the degree of similarity based on each profile.

Figure 5. An individual classification model based on a combined profile.
Multiple Profiles based Ensemble Model for Analytical Classification of Cyber Incident

The composition of dataset for verification is as shown in Table 3 and data of 653 cases of cyber incidents were used. The data consist of 6 attributes; 295 data were labeled as Positive and 358 data were labeled as Negative. The ratio of training to test was 7:3.

Table 3. Composition of DataSet

| Incident | probe attack | DoS attack | access attack | IP      | URL         | Hash    |
|----------|--------------|------------|---------------|---------|-------------|---------|
| #no_03   | Satan        | land       | rootkit       | 1.1.1.2 | test.abc.com| ab2c7defg|
| #1       | Satan        | land       | rootkit       | 1.1.1.1 | www.abc.com | ab2c7defg|
| #2       | saint        | smurf      | worm          | 2.2.2.2 | www.zyx.com | z1yxw9vut|

It can be confirmed that with multiple profiles-based ensemble model proposed in this study, the accuracy of the classification gets higher as the number of multiple profiles get higher according to the test result as shown in Figure 9.

Overall, the separated profile shows higher accuracy of the classification than the combined profile.

In Figure 9, combined profile represents the evaluated value of the individual classification model in Figure 5 and separated profile represents the outcome evaluated based on the ensemble classification model in Figure 6. As separated profile is applied to the combination algorithm in Figure 6, it can be found out that weighted majority
voting displays higher accuracy of the classification than Majority Voting.

Figure 9. Accuracy of classification by training repetition of classifier.

5. Conclusion

This paper has the significance in suggesting what kind of information we could use to create cyber incident profiles in the environment of big data and how we could combine and utilize them effectively according to multi-profile ensemble model. Our suggested ensemble model offers the ability to classify similar cyber incidents and improves the accuracy and reliability of analytical classification of cyber incidents. However, our model should be further studied to compare the differences in accuracy of classification by applying the proposed method to different algorithms.

6. Acknowledgement

This work was supported by the research grant of Pai Chai University in 2016.

7. References

1. Ten C, Manimaran G, Liu C. Cybersecurity for Critical Infrastructures: Attack and defense modeling, IEEE Transactions on Systems. 2000; 40(4):853–65.
2. Faysel MA, Syed S, Haque. Towards cyber defense: Research in intrusion detection and intrusion prevention systems. International Journal of Computer Science and Network Security. 2010; 10(7):316–25.
3. Singh S, Agrawal S, Murtaza A, Rizvi, Thakur RS. Improved support vector machine for cyber-attack detection. Proceeding of WCECS IEEE; 2011.
4. Bapuji V, Kumar RN, Govardhan A, Sarma SSVN. Soft computing and artificial intelligence techniques for intrusion detection system. Network and Complex Systems. 2012; 2(4):24–33.
5. Nguyen HD, Cheng Q. An efficient feature selection method for distributed cyber attack detection and classification. 2011 45th Annual Conference on Information Sciences and Systems (CISS); 2011. p. 1–6.
6. Kumar B, Mishra, Saini H. Cyber attack classification using game theoretic weighted metrics approach. World Applied Sciences Journal. 2009;7(Special Issue of Computer & IT):206–15.
7. Du H, Murphy C, Bean J, Yang SJ. Toward unsupervised classification of non-uniform cyber attack tracks. International Conference on Information Fusion; 2009.p. 1919–25.
8. Jain A, Singh AS. Distributed Denial Of Service (DDOS) Attacks - classification and implications. Journal of Information and Operations Management. 2012; 3(1):136–40.
9. Dharamkar B, Singh R. Cyber-attack classification using improved ensemble technique based on support vector machine and neural network. International Journal of Computer Application. 2014; 103(11):1–7.
10. Amudha P, Karthik S, Sivakumari S. An experimental analysis of hybrid classification approach for intrusion detection. Indian Journal of Science and Technology. 2016 Apr; 9(13):1–8.
11. Sharma M, Singh SK, Agrawal P, Madaan V. Classification of clinical dataset of cervical cancer using KNN. Indian Journal of Science and Technology. 2016 Jul; 9(28):1–5.
12. Suganthi RLS, Hanumanthappa M. Classification of event image set using mining techniques. Indian Journal of Science and Technology. 2016 Jun; 9(22):1–6.
13. Verma A, Kaur I, Kaur A. Algorithmic approach to data mining and classification techniques. Indian Journal of Science and Technology. 2016 Jul; 9(28):1–22.
14. Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence. 2011;25(3):337–72.
15. Singh S, Silakari S. An ensemble approach for cyber attack detection system: A generic framework. 2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing(SNPD); 2013. p. 79–84.
16. Rathore D, Jain. Design hybrid method for intrusion detection using Ensemble cluster classification and SOM network. International Journal of Advanced Computer Research. 2012; 2(5):181–6.
17. Jouve PE, Nicoloyannis N. A new method for combining partitions, applications for distributed clustering. Proceeding of the International Workshop on Parallel and Distributed Machine Learning and Data Mining:2003.
18. Kim J, Ahn B-H, Jeong D. A recommender system using mixed filtering for health products. The Journal of Internet Electronic Commerce Research. 2012; 12(2):109–24.
19. Marin J, Ragsdale D, Sirdu J. A hybrid approach to the profile creation and intrusion detection. Proceedings of DARPA Information Survivability Conference and Exposition II, (DISCEX ‘01). 2001; 1:69–76.
20. Kim YS, Mun H-J, Cho H, Kim B, Lee J, Lee JW, Lee BY. The composition and analytical classification of cyber incident based Hierarchical Cyber Observables. Journal of the Korea Content Association.2016;16(11):139–53.