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 analyses. 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 incident. A variety of studies are in progress regarding ways to classify the type of cyber incidents.

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 results.

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 alone.

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

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
<table>
<thead>
<tr>
<th>Incident</th>
<th>Probe attack</th>
<th>DoS attack</th>
<th>Unauthorized access attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>#no_01</td>
<td>Satan</td>
<td>Land</td>
<td>Rootkit</td>
</tr>
<tr>
<td>#no_02</td>
<td>Saint</td>
<td>Smurf</td>
<td>Worm</td>
</tr>
</tbody>
</table>

Table 2. Cyber incident based cyber observables
<table>
<thead>
<tr>
<th>Incident</th>
<th>IP</th>
<th>URL</th>
<th>Hash</th>
</tr>
</thead>
<tbody>
<tr>
<td>#no_01</td>
<td>1.1.1.1</td>
<td><a href="http://www.abc.com">www.abc.com</a></td>
<td>ab2c7defg</td>
</tr>
<tr>
<td>#no_02</td>
<td>2.2.2.2</td>
<td><a href="http://www.zyx.com">www.zyx.com</a></td>
<td>z1yxw9vut</td>
</tr>
</tbody>
</table>

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.

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, designating 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 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 4.** Similarity evaluation algorithm.

**Figure 5.** An individual classification model based on a combined profile.
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

<table>
<thead>
<tr>
<th></th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>244</td>
<td>102</td>
</tr>
<tr>
<td>Negative</td>
<td>213</td>
<td>94</td>
</tr>
<tr>
<td>Total</td>
<td>457</td>
<td>196</td>
</tr>
</tbody>
</table>

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