FEATURE ANALYSIS ON THE CONTAINMENT TIME FOR CYBER SECURITY INCIDENTS

GULSUM AKKUZU[1], BENJAMIN AZIZ[1], HAN LIU[2]

[1]School of Computing, University of Portsmouth, UK
[2]School of Computer Science and Informatics, Cardiff University, UK
E-MAIL:gulsum.akkuzu@port.ac.uk, benjamin.aziz@port.ac.uk, liuh48@cardiff.ac.uk

Abstract:

Data mining techniques have been widely used as a common goal to discover hidden patterns from big data sets, so researchers have been motivated to make use of data in discovering useful information. The main contribution of this paper lies in its identifying relevant features from an open data set to predict the containment time of Cyber incidents. In particular, 13 relevant features were identified and selected to come up with a predictive model. Our results are discussed in the context of the organization's' information security.

Keywords:

Data-driven security management; Feature selection; Organizational data set; WEKA tool; Information security; Machine learning

1. Introduction

In the recent years, companies have been pushed to transform themselves into data-driven organizations because of technological developments in the big data infrastructure [9]. This has brought an increment inaccessible organizational data for researchers or users. There is even a huge amount of data available in organizations. However, these data cannot be used directly, which means that the raw data needs to be pre-processed towards extracting the hidden patterns, semantic information, and useful features [10]. This operation is referred to as feature extraction from data. Data driven security management has been commonly adopted to uncover hidden patterns in data and respond with countermeasures. In the business world, data is used to identify and analyze the situation of an organization [15]. In order to achieve data-driven security management in the big data area, undertaking feature extraction to mine useful information is necessary.

Prediction, forecasting, and extraction of associated features have become an inevitable approach in the business area. In analyzing social cases [18], extraction of useful information from a big data set plays a key role for organizations [13]. There are various approaches to achieve feature extraction from big data. In our case, data-driven security management is adopted through feature evaluation, and relevant features are then selected to predict the containment time of an incident. The containment time is defined as a period of time from the moment of incident discovery until the moment of containment. The importance of containment is explained in [3], which is extremely important in stopping the attacker from making more damage to the system. It is stated that the main part of containment is decision making [3]. Thus, if there are predefined strategies to contain the incident, decision making can be much easier. Hence, acceptable risks need to be defined by organizations; and on top of that, some containment strategies need to be developed to deal with incidents.

We claim that if organizations know the relevant features on the containment time, it can be much easier to predict the incident's containment time, adopt specific strategies for it, and decide which features (risk) are acceptable or unacceptable. For this reason, this paper focuses on identifying relevant features on the containment time of an incident. In particular, we used an open organizational data set to understand organization's situation better in terms of information security. The main contributions of this paper can be summarized as follows: 1) it identifies the relevant features on the containment time of a cyber security incident, and discusses them with regards to organizations' security issues (feature selection), 2) it interprets key findings from experimental results in terms of information security, and 3) it develops and interprets a model by using selected features, which is a transformation from data through information to knowledge.

2. Related Works

Different detection approaches can be utilized to find attack patterns from data. One of these approaches is referred to as feature selection. Feature selection can be explained essentially as identifying features relevant for predicting a class value by using machine learning approaches [7]. In other words, it is utilized to identify the relationships between the features and class values in a data set. In the context of security management, it is aimed at identifying which features have an effect on the containment time of incident in order to minimize the exorbitance and maximize the relevance of the features (a subset of the selected features) [12]. Once a set of relevant features is selected, meaningful points can then be deduced.

According to [12], a large number of features can be irrelevant to the class value in a security data set. In this case, it is highly necessary to adopt machine learning algorithms for the automatic evaluation and selection of features. The feature selection can be achieved by using three main approaches; namely, filter, wrapper, and embedded ones [12]. Features are graded by looking at their statistical importance in the setting of filter based methods [5]. In the setting of wrapper methods, features are selected heuristically based on the classification accuracy of the classifier trained on the subset of the features [8]. Embedded methods predicate feature selection on classifier [14]. The performance of these methods is considered reputable because of the determined class value [21]. Based on this, we utilized the approach proposed in [12].

When an incident is detected by a system or organization, it needs to be undertaken for the information security of a system. As explained by [1], to mitigate an incident's damaging impacts on a system, an effective reaction needs to be given. There are some significant points of incident response such as containment, discovery, eradication, and so on [2]. In this paper, containment time will be the main point of investigation. The goal of incident containment is preventing the incident from spreading, stopping the incident, and regaining control of the compromised system [20]. In [16], the containment time is explained thoroughly and defined as the discovery and contain of incident.

The previous studies produced different methods and techniques to tackle the issues of organizational information security. However, it has never been tried to deduce which features could determine containment time of a Cyber incident. It has thus motivated us to detect relevant features on the containment time of a Cyber incident. Concurrently, we aim to discover new information by using raw data.

3. Proposed Solution

In this section, we present steps that need to be taken to find a solution to the problem. The most important points are choosing suitable data sets and tools to obtain efficient results. We chose VERIS‘s data set VCDB to achieve our aim for the reason that it has enough records and each record contains excessive features. As it is known, the more features you have on a data set, the more questions you get. The VCDB has over six thousand and eight hundred records and over two thousand features. The VERIS data set aims to help organizations collect useful incident and sharing qualified information with others [16]. VERIS Community Database offers a solution to researchers interested in security incident problems, and to risk managers who lack reliable, unrestricted, and comprehensive raw data sets of security incidents available for download [4]. The second step is data cleaning and pre-processing, which depend on the research questions with a view and focus of the data set. Firstly, we need to analyze which features of the original data set are irrelevant for the study. Secondly, whether an incident has distinctive value or not needs to be judged, e.g., if an incident only involves a binary value (true or false), it does not make sense to include that value in our case. Finally, it is necessary to identify feature redundancy cases, e.g., there can be different incidents which have intersections and these incidents can be gathered under an incident.

As previously mentioned, it is crucial to decide on appropriate tools. In this paper, we use the WEKA tool, since it has a powerful feature selection function with many different options in line with our purpose. In particular, we chose the best-first and ranker options. The best-first option brings the relevant features based on their importance for a class value (for this paper, class value is containment time). The ranker part arranges the importance of the values by attributing numbers to them. We focused on feature selection by identifying irrelevant features on the VCDB data-set. It allowed us to select relevant features to specific class value.

4. Results

This section presents our experimental setup and results reported using WEKA [19]. Effective features of the containment time of a Cyber incident were found by applying feature selection algorithms on the tool.

As mentioned earlier, the main aim of this paper is to explore features relevant to containment time. To achieve this goal, two parameters need to be set for feature selection; namely, attribute evaluator and search method. Each attribute is evaluated within the context of class value. The experiment's results are presented in tables. More details on [19] are provided by the University of Waikato wikipage [19]. We created our own data set after editing the VCDB original data set. Figure 1 indicates how many relevant features we obtained from VCDB.280 data set by using feature selection filters.

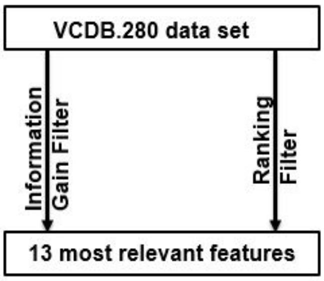

Figure 1. Directions of the experiments

4.1. Experimental Results

Through feature selection, relevant attributes are selected, i.e. redundant and/or irrelevant attributes are removed from the feature set. As described in Section 2, there are two types of feature evaluation methods; namely, filter and wrapper. The former is considered to be an independent evaluation and the latter is essentially based on learning algorithms. In our experiment, filter approach is adapted to select a subset of attributes, which is based on the information gain and gain ratio filters because they are capable of capturing high correlation between features and the class attribute [19].

The results are presented in Table 1, which includes the results of running information on WEKA. In [19], different attribute evaluators are used to compare their performance. The most important parts are underlined and the first underlined one is the instances and attribute values. As it was mentioned in previous sections, 224 attributes were taken from the VCDB data set after editing and shrinking/cleaning. Another point is that the data set contains

6861 records, and none of the original records were removed. The second underlined point is the class value, which, in our case, is containment time. The section shows the relevant features, which affect the containment time. The results show that the aim of this paper is achieved from the experiments. The meanings of these features are explained in the data analysis part of this paper.

Table 1. Results from the WEKA tool

Selected Attributes
victim.revenue.iso.currencycodeUSD
attribute.availability.variety.Interruption
attribute.availability.duration.unit.Hours
attribute.availability.duration.unit.Days
asset.governance.Personally owned
timeline.discovery.unit
discovery.method.Ext.customer
action.physical.variety.Theft
timeline.compromise.unit
timeline.incident.year
victim.industry.name
victim.state
plus.timeline.notification.year
Relevant Attributes:: 13

Table 2. Selected features' Value

Selected attributes	**Provided Information**
Attribute	It shows which security attributes were compromised during an incident
Asset	Asset describes compromised information gains during an incident
Action	it identifies what is the cause and contribution of the threat to an incident
Response	it answers questions are first how the incident was covered, what was the remediation process of it, and how the discovery of incident is done?
Victim	It shows importance of victim' details

Table 2 presents a brief explanation what information selected features provide. Features are summarized in five main types that are namely action, asset, response, attribute and victim. We used the ranking filter to see features' effectiveness on the containment time. The Correlation Ranking Filter gave us the result "Selected attributes: 11,5,4,10,3,2,9,6,1,8,13,12,7". In Table 1, the features could be put in order from top to bottom. Based on the ranking filter, it is evident that the most effective feature on the containment time is action incident. It helps us to decide the sizes of our model, which is given in Section 5. As can be seen from the experiment, our aim is achieved by performing experiments on [19] and applying filters on it. 13 specific attributes out of 224 are selected to predict the containment time (to clear a possible confusion; 13 features excluding the class value of data set).

4.2. Interpretation of Results and VCDB Analysis

We analyzed our work with security issues in mind. The first analysis was conducted through feature prediction analysis, which was done by using the WEKA tool. The second analysis was carried out by using extracted features from the WEKA on the VCDB data set, which was the original data set used in our work. Figure 2 shows our technique, which provides information about our directions for doing each analysis. The first analysis is from containment time to features, and it indicates that if a new Cyber incident's containment time needs to be predicted, then the relevant features need to be known. For the second part, we are basing our analysis on our experiments' result with regards to the containment time of incidents. This means that if we know the effective features on the containment time of incident then we can identify how long it will take to contain it. The features related to containment time of an incident were presented in the previous Section 4 in Table 1 that provides related features' names. Based on the results of those experiments, five main classes can be determined; namely, action, asset, attribute, response, and victim. The benefit of classification features is that it helps one to understand these features in terms of security in organizations.

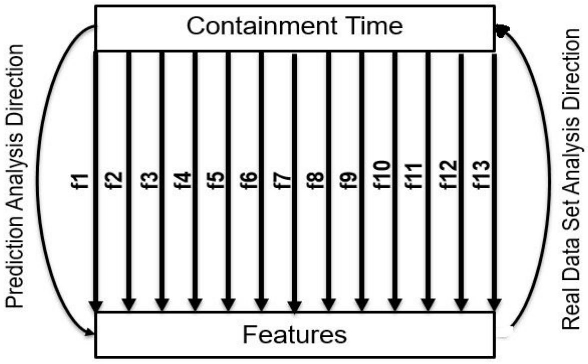
Containment Time
Prediction Analysis Direction
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
f11
f12
f13
Real Data Set Analysis Direction
Features

Figure 2. Analysis technique

Action as it is given in the requirement and analysis part, the description of the action in the VERIS is that it identifies the cause and contribution of the threat to an incident. It can be either malware cause, hacking cause, misuse, error, social, environmental cause, or physical cause. **Action.physical.variety.Theft** was the result of the experiments in our case. The primary question is how that feature can be interpreted in the sense of security. The statement of "Action.Physical.variety.Theft" is that if the action is deliberate physical theft threat, then it can be claimed that the prediction of its containment time could be much quicker than other possibilities. Other possibilities are given in [16]. There are six possible primary categories apart from physical action. By looking at the results of the experiments, it can be argued that although there are three other possibilities under the physical category, predicting the containment time of theft could be the shortest time for organizations. To make it clear, in our world, employers have given computers, laptops, or mobile devices to their employees. Physical theft is completely related to physical devices. This means that in case the physical devices are stolen, the company can lose crucial data that is on the device. As it has been understood from the experiments' results; if it is physical theft, then its containment time prediction may be quicker then others because of the sensitiveness.

The second feature is attribute value. It shows which security attributes were compromised during an incident. The security attributes are confidentiality, possession, integrity, authenticity, availability, or utility. When the results of experiments are evaluated, it is clearly seen that availability is one of the most relevant features on the containment time. All the results received, other possibilities, and their meaning in terms of security are clarified. **Attribute.availability** is one of the results in our case. This means that if the availability of data is compromised, prediction of its containment time can take less time than the containment of either confidentiality or integrity. Availability of data is one of the main points of security issues and organizations since it provides the retrieval and update of data to support relying party systems [6]. Data needs to be available whenever and wherever needed. Therefore, it can be clearly deduced that the result of the experiments was true since the results indicate that the most relevant security attribute is availability rather than confidentiality and integrity.

The asset value was the last relevant feature from the incident details features of the VERIS on containment time. Asset describes compromised information gains during an incident. Before making the results of experiment clear, a table is given to show subtitles of asset value in the VERIS web site [16]. According to our results, the most relevant one is ownership. The result is **asset.governance.Personally owned**, it means that if the compromised information is owned personally, its prediction of containment time will be shorter than other assets. The meaning of ownership in the VERIS is who owns (the asset) effects on an incident. It can be explained with an example: if one person's data confidentiality is affected by an attack, the sufferer takes care of her data confidentiality more than others such as her colleagues, her employees, and so on. From an information security perspective, it makes the result of experiments meaningful and true.

The following result is related to the response feature of an incident from the experiments. The response focuses on the timeline of incidents and aims to answer these questions: how was the incident covered, what was the remediation process of it, and how is the discovery of the incident done? The result of experiments: **Timeline.Notification,Timeline. compromise.unit, Timeline.incident.year, DiscoverymethodExt.customer**

Before discussing the results of the experiments, the Timeline and Discovery method need to be explained based on the VERIS community database explanations. In the VERIS, four different types of time line events are explained: compromise, exfiltration, discovery, and containment. In our case, the discovery method variety was **the external customer**. The time line concept of an incident depends on the discovery method. It can be made clear with examples: if an incident is reported by a user, which is internally reported, the containment time for it can be predicted. If the actor disclosed the information externally, then the containment time of it can be predicted as well. Another discovery method might be from customer: if it is discovered by customers, containment time of incident could again be predicted since it is a relevant feature. After discussing the time line concepts based on the VERIS explanations, we can go back to the experiments' results. The prediction of containment time may be longer or shorter.

The last result is related to victim information. As it is shown in the result of experiments, **victim.state** and "victim.industry.name" are effective features on the containment time based on our experiment results. If the victim industry name is known, the containment time of the incident can be shorter than when it is unknown. The most relevant industries are **the healthcare and public industries**. In healthcare and public industries incidents, the prediction of containment time could be longer than others. The main idea here is that if the industry name is known, the containment time can be predicted. Another feature is the victim state. A clear deduction can be made that an incident's containment time can be predicted if the victim state (country region) is known. Having discussed relevant features on the containment time of an incident, it could be seen our work brings a new view to information security area of organisations.

VCDB Analysis:
In this part, we are going back to our original data set by taking our features to analyze each feature's containment time. By using its containment timeline concepts, we created tables that show percentages of the containment time concepts.

We claim that if an incoming incident has any of relevant features given to us, then we may be able to tell its specific time percentage of containment period. Figure 3 gives percentages of the containment time in each time period from our original data set. For example, if the feature is action physical theft, then we can calculate its containment time in the hour period. We used the formula percentage(p)=specific time period's containment time(st)/total containment time that belongs to specific feature(N) p=st/N. Let us take hour period of time percentage on physical theft feature containment time, if an incoming incident has physical theft features, then its percentage of hourly containment time is 11% (p=118/1038). We have excluded all cases of unknown times because these cases are undisclosed. All the following Figures 3–7 represent the percentage of time units of the containment time corresponding to each features we predicted in our analysis.

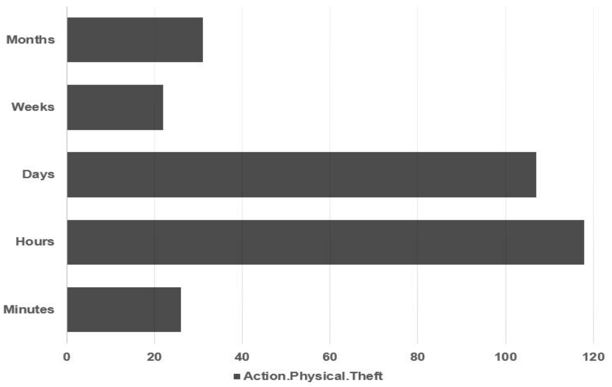

Figure 3. Percentage of action physical theft

We created all figures above by using our original data set (VCDB). Features' time concepts are different: for example, while attribute availability feature has months time concept in our data set, discovery method does not have it. We calculated all percentages by using the formula (p=st/N).

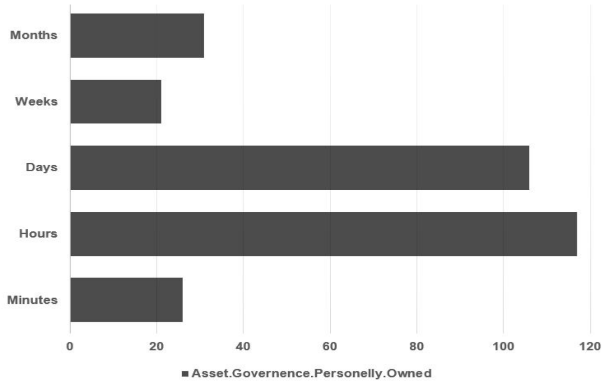

Figure 4. Asset governance personally owned

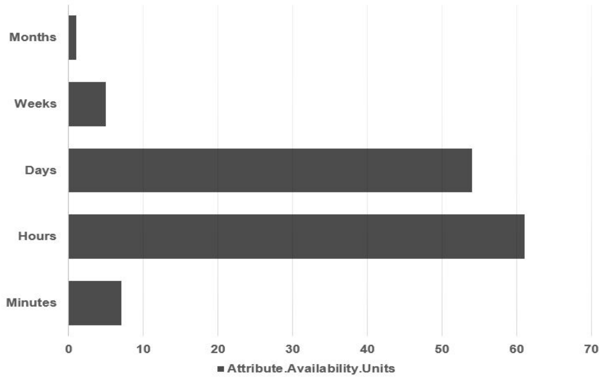

Figure 5. Percentages of specific time period

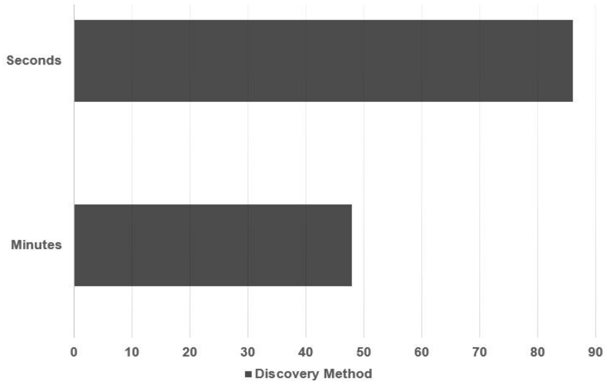

Figure 6. Discovery method time percentages

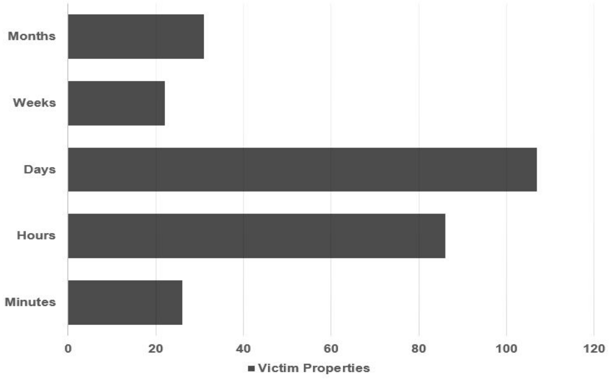

Figure 7. Victim Properties time percentage

When an incoming incident has any of relevant features, we can tell its specific containment period of time percentage.

5. Model Development

Based on the experiment findings, relevant features have contributed to developing an effective containment time feature model, Figure 8. It is an abstract module that highlights relevant features, removing all the schema detail and impact rating figures. We think that the model could be useful for people who have no expertise in either VERIS schema or [19] and its feature selection options. The model given shows the ratio of features on the containment time of an incident. Based on Figure 8 we claim that if the time line concepts of an incident (such as discovery time, compromise, and exfiltration) are known, then a prediction about the containment time of that incident can be made so that organizations make a decision based on their circumstances. For example, they may opt for a short time or a longer time containment strategy. The second feature is the attribute details. We need to expand on the attribute part as the availability of attribute was one of the certain areas in our experiment results. Thus, if the organizations decide to come up with a containment strategy for an incident, then they need to think about service availabilities such as sources, network connection etc. Based on Figure 8, it can be clearly seen that if the security attribute detail affected is either availability, confidentiality, or integrity, then prediction of its containment time may take less time than the others. The next feature is victim properties, it makes sense that if the victim property is known, then the containment time of incident can be predicted easily. It helps organizations think about their weakest points. The arrows show that the direction of the effect is from feature side to class value, which, in our case, is containment time.

6. Conclusion and Future Work

In this paper, we analyzed the most relevant features related to the containment time of Cyber security incidents as reported in the VERIS open data set of Cyber incidents. Over 200 features from 6800 records were selected, and these were then utilized as input through the WEKA filtering algorithms. Thirteen features out of the 200 were deemed relevant for predicting containment times of incidents. These features could form part of a model of organizations' information security and could help companies increase their information security levels by focusing on the most important features when considering the containment times of incidents.

Future work could focus on widening the scope of the classes considered to be relevant to express a more general notion of Cyber response. For example, other times could be considered, such as incident discovery and data exfiltration times. Additionally, other response aspects could also be included alongside containment time, which would be triggered by Cyber incidents, e.g., response action types and the impact of the incidents on the organization.

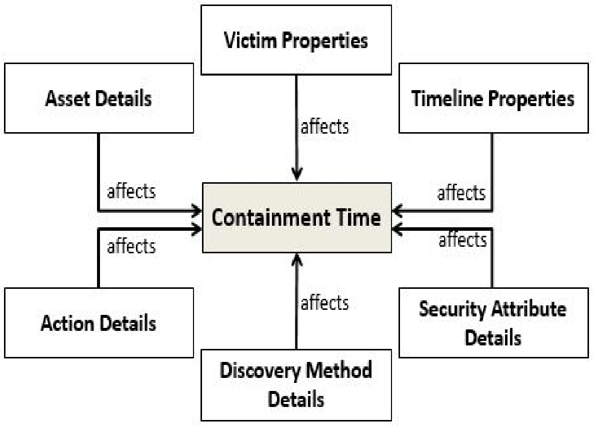

Figure 8. The model of features relevant to containment time

References

[1] Nurul Hidayah. A.R., Kim-Kwang Raymond. C. A survey of information security incident handling in the cloud. Computers and Security, 49, 45-69, 2015.

[2] Chris. A., Audrey. D.,Georgia. K., Robin. R., Mark. Z. Defining incident management processes for csirts: A work in progress.2004.

[3] Cichonski, Paul; Millar, Tom; Grance, Tim; Scarfone, Karen. International Journal of Computer Research; Huttington Vol. 20, Iss. 4, (2013): 459-530.

[4] GitHub. 2017.

[5] Guyon, I., Elisseeff, A. . An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.2003.

[6] Vincent C. Hu., David F.,Rick K., Adam S., Kenneth. S., Robert M., Karen S. Attribute-based access control. Computer, 48(2), 85-88.2015.

[7] Gunes K., A.Nur Z.H., Malcolm I.H.. Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection data sets. In Proceedings of the third annual conference on privacy, security, and trust.October, 2015.

[8] P. Ganesh K., T. Aruldoss A.V., Ponnusamy R., Devaraj, D. Design of fuzzy expert system for micro array data classification using a novel genetic swarm algorithm. Expert Systems with Applications, 39(2), 1811-1821.2012.

[9] In L. Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303.2017.

[10] David M., Michal B., Pavel Z. Speeding up the multimedia feature extraction: a comparative study on the big data approach. Multimedia Tools and Applications, 76(5), 7497-7517. 2017.

[11] S. Murugarasan M., Sellapan P. Security metrics maturity model for operational security. In Computer Applications Industrial Electronics (ISCAIE), 2016 IEEE Symposium on (pp. 101-106). IEEE. May, 2016.

[12] Yamuna P., Dinesh K., K.K. B. Max-Margin Feature Selection. Pattern Recognition Letters.2017.

[13] Michael S., Homayoon D.,George A., Chester E., Sergio G., Donovan M., Ali M. Probabilistic risk assessment procedures guide for NASA managers and practitioners. 2011.

[14] Mingkui T., Li W., Ivor W. T. Learning sparse svm for feature selection on very high dimensional datasets. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 1047-1054).2010.

[15] Bhavani T., Murat K., Kevin H., Latifur K., Tim F., Anupam J., Tim O., Elisa B. A Data Driven Approach for the Science of Cyber Security: Challenges and Directions. In Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on (pp. 1-10). IEEE. July, 2016.

[16] VERIS. 2017.

[17] Wang, S., Wei, J. (2017). Feature selection based on measurement of ability to classify subproblems. Neurocomputing, 224, 155-165

[18] Zhigang W., Kamran N. Statistical Characterization, Pattern Identification, and Analysis of Big Data. SAE International Journal of Materials and Manufacturing, 10. Jan 2017.

[19] WEKA. 2017.

[20] Wilcox, S., and Brown, D. 2005. Responding to security incidents – sooner or later your systems will be compromised. Journal of Health Care Compliance. 7, 2 (April. 2005), 41- 48.

[21] Yiteng Z., Mingkui T.,Ivor T., Yew S. O. Discovering support and affiliated features from very high dimensions. arXiv preprint arXiv:1206.6477.2012.