Using Data Mining and Machine Learning (DM-ML) for the Classification and Prediction of Significant Cyber Incidents (SCI)

[1]Atigadda Kavyasree, [2]Vuttunoori Ashritha, [3]Patil Manikrao

[12]Under Graduate, Department of IT-GNITC-Hyderabad
[3]Assistant Professor, Department of IT-GNITC-Hyderabad

ABSTRACT:

The rapid growth in technology and several IoT devices make cyberspace unsecure and eventually lead to Significant Cyber Incidents (SCI). Cyber Security is a technique that protects systems over the internet from SCI. Data Mining and Machine Learning (DM-ML) play an important role in Cyber Security in the prediction, prevention, and detection of SCI. The dataset (SCI as per the report of the Center for Strategic and International Studies (CSIS)) is divided into two subsets (pre-pandemic SCI and post-pandemic SCI). Data Mining (DM) techniques are used for feature extraction and well know ML classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) for classification. A centralized classifier approach is used to maintain a single centralized dataset by taking inputs from six continents of the world. The results of the pre-pandemic and
post-pandemic datasets are compared and finally conclude this paper with better accuracy and the prediction of which type of SCI can occur in which part of the world. It is concluded that SVM and RF are much better classifiers than others and Asia is predicted to be the most affected continent by SCI.

INTRODUCTION

The speedy advancement in technology and the boom in the IoT industry increase the possibility of cyber incidents. Especially, after the pandemic COVID-19, this ratio is in-creased . It is expected that the number of IoT devices count will reach around 75 billion by 2025. As per the handbook 'Cybersecurity Almanac' released by 'Cybersecurity Ventures, the global cybercrime cost is expected to reach USD 10.5 trillion in 2025, from USD 6 trillion in 2021. In 2021, an organization suffered from a ransomware attack after every 11 seconds, and it is expected to suffer after every 2 seconds in 2031 [4]. Table 1 depicts up-to-date statistics about the internet and social media users from January 2020 to October 2022. There is an alarming increase in the percentage of 24.5 active social media users. Cyber security is a technique to protect systems over the internet from cyber incidents. A cyber incident means an activity or event which occurred through the internet and jeopardizes the Confidentiality, Integrity, and Availability (CIA Triad) of the communication system through any means. The term Significant Cyber Incident (SCI) means a cyber incident that results in

manifest damage to the national security and economy. Cyber security is used by individuals as well as organizations to protect their information and systems over the internet from unauthorized access. With the increase in SCI, cyber security measures also improved to tackle these incidents. Data Mining and Machine Learning (DM-ML) play an important role in cyber incidents prediction, prevention, and Detection by using different approaches. In this paper, the outfall of SCI has been predicted based on the datasets, collected from the report of the Center for Strategic and International Studies (CSIS). The datasets consist of textual data comprising of SCI type and the continent where it occurred. First, it is divided into two parts (prepandemic SCI and post-pandemic SCI) and then analyzed ten types of SCI that occurred in six continents of the world. Pre-pandemic (before COVID-19) dataset includes those SCI which happened during the period from 2003 to December 2019. Similarly, the post-pandemic (after COVID-19) dataset includes those SCI which happened during the period from January 2020 to till date. As there are no countries in the seventh continent 'Antarctica', so the only six continents in our study are considered. Further, it is also investigated how the data can be used for classification accuracy and eventually the better classifier for distinguishing different SCI. The results achieved by focusing on which type of SCI occurred at which continent of the world. The main objective of this study is to explore the benefits of centralized classifier for treating future SCI. Data Mining features like n-grams and Bag of Words (BoW) are more useful now for the feature extraction from the collected data. ML algorithms like Naïve Bayes (NB), Support Vector Machine (SVM) Logistic Regression (LR) and Random Forest (RF) are used for data classification. Finally, the results of pre- and post-pandemic datasets are compared which concludes with the best results of SVM, and RF classifiers and Asia (the most affected continent by SCI) is predicted.

Related Work

I. Secure Training of Decision Trees With Continuous Attributes:

We apply multiparty computation (MPC) techniques to show, given a database that is secretshared among multiple mutually distrustful parties, how the parties may obliviously construct a decision tree based on the secret data. We consider data with continuous attributes (i.e., coming from a large domain), and develop a secure version of a learning algorithm similar to the C4.5 or CART algorithms. Previous MPC-based work only focused on decision tree learning with discrete attributes (De Hoogh et al. 2014). Our starting point is to apply an existing generic MPC protocol to a standard decision tree learning algorithm, which we then optimize in several ways. We exploit the fact that even if we allow the data to have continuous values, which a priori might require fixed or floating point representations, the output of the tree learning algorithm only depends on the relative ordering of the data. By obliviously sorting the data we reduce the number of comparisons needed per node to O(N log2 N) from the naive O(N2), where N is the number of training records in the dataset, thus making the algorithm feasible for larger datasets. This does however introduce a problem when duplicate values occur in the dataset, but we manage to overcome this problem with a relatively cheap subprotocol. We show a procedure to convert a sorting network into a permutation network of smaller complexity, resulting in a round complexity of O(log N) per layer in the tree. We implement our algorithm in the MP-SPDZ framework and benchmark our

implementation for both passive and active three-party computation using arithmetic modulo 2 64. We apply our implementation to a large scale medical dataset of ≈ 290 000 rows using random forests, and thus demonstrate practical feasibility of using MPC for privacy-preserving machine learning based on decision trees for large datasets.

II.Privately Evaluating Decision Trees and Random Forests.

Decision trees and random forests are common classifiers with widespread use. In this paper, we develop two protocols for privately evaluating decision trees and random forests. We operate in the standard two-party setting where the server holds a model (either a tree or a forest), and the client holds an input (a feature vector). At the conclusion of the protocol, the client learns only the model's output on its input and a few generic parameters concerning the model; the server learns nothing. The first protocol we develop provides security against semi-honest adversaries. We then give an extension of the semi-honest protocol that is robust against malicious adversaries. We implement both protocols and show that both variants are able to process trees with several hundred decision nodes in just a few seconds and a modest amount of bandwidth. Compared to previous semi-honest protocols for private decision tree evaluation, we demonstrate a tenfold improvement in computation and bandwidth.

III.Web page multiclass classification.

As the internet age evolves, the volume of content hosted on the Web is rapidly expanding. With this ever-expanding content, the capability to accurately categorize web pages is a current challenge to serve many use cases. This paper proposes a variation in the approach to text preprocessing pipeline whereby noun phrase extraction is performed first followed by lemmatization, contraction expansion, removing special characters, removing extra white space, lower casing, and removal of stop words. The first step of noun phrase extraction is aimed at reducing the set of terms to those that best describe what the web pages are about to improve the categorization capabilities of the model. Separately, a text preprocessing using keyword extraction is evaluated. In addition to the text preprocessing techniques mentioned, feature reduction techniques are applied to optimize model performance. Several modeling techniques are examined using these two approaches and are compared to a baseline model. The baseline model is a Support Vector Machine with linear kernel and is based on text preprocessing and feature reduction techniques that do not include noun phrase extraction or keyword extraction and uses stemming rather than lemmatization. The recommended SVM One-Versus-One model based on noun phrase extraction and lemmatization during text preprocessing shows an accuracy improvement over the baseline model of nearly 1% and a 5- fold reduction in misclassification of web pages as undesirable categories.

IMPLEMENTATION

Naive Bayes(NB) algorithm

Naive Bayes is a probabilistic machine learning algorithm that can be utilized in a wide assortment of grouping tasks. The name naive is utilized on the grounds that it accepts the provisions that go into the model are free of one another. Numerically Given the Bayesian calculation is addressing a class variable and the arrangement of qualities are, Conditional probability of A given B can be registered as:

$$P(A \mid B) = P(A \cap B) / P(B) \quad (1)$$

Logistic Regression

Logistic Regression is a classification algorithm for categorical variables like Yes/No, True/False, 0/1, etc,. Logistic regression transforms its product using the logistic sigmoid

function to return a chance value. The definition of the logistic function is given in equation (2)

$$\sigma(t) = 1/1+et \quad (2)$$

Equation (3) function is used to transform the typical linear regression formula

$$f(x) = \beta 0 + \beta 1x \quad (3)$$

The resulting equation is shown in Equation 4. In this formula, p(x) represents the probability that an input sample belongs to the target 1. That is, the probability that an application is malicious given that it is making the observed system calls.

$$p(x) = 1/1+ e-\beta 0 - \beta 1x \quad (4)$$

Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way. This is the equation used in Logistic Regression. Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorize, logistic regression may be able to help.

Results

The proposed algorithm Naive Bayes and existing algorithm Logistic Regression (LR) algorithm were run at a time in an Anaconda-Jupyter. As the sample sets are executed for a number of iterations the accuracy values of Naive Bayes(NB) and Logistic Regression(LR) Algorithm classifiers vary for the classification of accuracy shown in Table 1. The observed values for the metrics of Group Statistics, the mean accuracy, and the standard deviation for the Naive Bayes(NB) Algorithm are 62.2 and 0.37014. The Logistic Regression(LR) Algorithm's mean accuracy is 49.92 and the standard deviation is 0.66106. The Naive Bayes(NB) Algorithm also obtained a standard error mean rate of 0.16553 whereas the Logistic Regression(LR) Algorithm obtained an error mean rate of 0.27563 as shown in table 2. Analysis of the overall classification of Detection of Malware in Cloud storage Data by Naive Bayes and Logistic Regression Algorithm models shows the classification of the detecting malware. Naive Bayes (62.7%) shows better accuracy than Logistic Regression (50%). Statistical Analysis of Mean, Standard deviation and Standard Error and Accuracy of Naive Bayes and Logistic Regression Algorithm is done. Then an independent sample test of 5 samples was performed, Naive Bayes Algorithm obtained a mean difference of 12.01 and a standard error difference of 0.33882. When compared to other algorithm performance, the Naive Bayes Algorithm performed better than the Logistic Regression Algorithm and the significance value of 0.053.

TABLES AND FIGURES

Table 1. Comparing accuracy values with the different sample sizes. It represents Detection of Novel Malware Attacks Analysis, the accuracy of Naive Bayes (62%), and

the Logistic Regression algorithm (50%).

Iteration	Naive Bayes	Logistic Regression
1	62%	50.0%
2	62.5%	49.5%
3	61.5%	50.5%
4	62.3%	49.9%
5	61.9%	50.3%

CONCLUSION

This paper focuses on the research based on Significant cyber incidents (SCI) from September 2003 to October 2022 as per the report of the Center for strategic and international studies (CSIS). The datasets are analyzed and classified using data mining and machine learning algorithms. Four different classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) are used and predicted the output (name of the continent based on the type of SCI). It is also predicted which continent is more affected by SCI during the period. Finally, it is concluded that SVM and RF are both better than other classifiers against our models, in both cases (pre-pandemic and post-pandemic) and Asia is the most affected continent by SCI.

REFERENCES:

[1] Y. Li and Q. Liu, ‘‘A comprehensive review study of cyber-attacks and cyber security; emerging trends and recent developments,’’ Energy Rep., vol. 7, pp. 8176–8186, Nov. 2021, doi: 10.1016/j.egyr.2021. 08.126.

[2] J. Kaur and K. R. Ramkumar, ‘‘The recent trends in cyber security: A review,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 34, no. 8, pp. 5766–5781, Sep. 2022, doi: 10.1016/j.jksuci.2021. 01.018.

[3] H. Hejase, H. Kazan, A. Hejase, and I. Moukadem, ‘‘Cyber security paper,’’ Comput. Inf. Sci., vol. 14, pp. 10–25, Mar. 2021, doi: 10.5539/cis.v14n2p10.

[4] Cybersecurity Almanac by Cyber Security Ventures. [Online]. Available:

[6] P. S. Seemma, S. Nandhini, and M. Sowmiya, ‘‘Overview of cyber security,’’ Int. J. Adv. Res. Comput. Commun. Eng., vol. 7, no. 11, pp. 125–128, Nov. 2018, doi: 10.17148/IJARCCE.2018.71127.

[7] Q. E. Hodgson, A. Clark-Ginsberg, Z. Haldeman, A. Lauland, and I. Mitch, Managing Response to Significant Cyber Incidents: Comparing Event Life Cycles and Incident Response Across Cyber and Non-Cyber Events. anta Monica, CA, USA: RAND Corp., 2022, doi: 10.7249/RRA1265-4.

[8] A. Handa, A. Sharma, and S. K. Shukla, ‘‘Machine learning in cybersecurity: A review,’’ WIREs Data Mining Knowl. Discovery, vol. 9, no. 4, p. e1306, Jul. 2019, doi: 10.1002/widm.1306.

[9] A. E. Ibor, F. A. Oladeji, and O. B. Okunoye, ‘‘A survey of cyber security approaches for attack detection, prediction, and prevention,’’ Int. J. Secur. Appl., vol. 12,

[10] K. Shaukat Dar, S. Luo, S. Chen, and D. Liu, ‘‘Cyber threat detection using machine learning techniques: A performance evaluation perspective,’’ in Proc. Int. Conf. Cyber Warfare Secur. (ICCWS), Oct. 2020,

pp. 1–6, doi: 10.1109/ICCWS48432.2020.9292388.

[11] Significant Cyber Incidents (SCIs).

[12] A. Pektaş, M. Eris, and T. Acarman, ‘‘Proposal of n-gram based algorithm for malware classification,’’ in Proc. 5th Int. Conf. Emerg. Secur. Inf., Syst. Technol., Jan. 2011, pp. 14–18.

[13] C. Wressnegger, G. Schwenk, D. Arp, and K. Rieck, ‘‘A close look on n-grams in intrusion detection: Anomaly detection vs. classification,’’ in Proc. ACM workshop Artif. Intell. Secur., Nov. 2013, pp. 14–18, doi: 10.1145/2517312.2517316.

[14] S. Soni and B. Bhushan, ‘‘Use of machine learning algorithms for designing efficient cyber security solutions,’’ in Proc. 2nd Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT), vol. 1, Jul. 2019, pp. 1496–1501, doi: 10.1109/ICICICT46008.2019. 8993253.

[15] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. M. Hossain, S. Ikhlaq, and S. Hossain, ‘‘Cyber intrusion detection using machine learning classification techniques,’’ in Proc. Int. Conf. Comput. Sci., Commun. Secur., Singapore, 2020, pp. 121–131.

[16] A. Terai, S. Abe, S. Kojima, Y. Takano, and I. Koshijima, ‘‘Cyberattack detection for industrial control system monitoring with support vector machine based on communication profile,’’ in Proc. IEEE Eur. Symp. Secur. Privacy Workshops, Apr. 2017, pp. 132–138, doi: 10.1109/EuroSPW.2017.62.

[17] K. Ghanem, F. J. Aparicio-Navarro, K. G. Kyriakopoulos, S. Lambotharan, and J. A. Chambers, ‘‘Support vector machine for network intrusion and cyber-attack detection,’’ in Proc. Sensor Signal Process. Defense Conf. (SSPD), Dec. 2017, pp. 1–5, doi: 10.1109/SSPD.2017. 8233268.

[18] N. Bhusal, M. Gautam, and M. Benidris, ‘‘Detection of cyber attacks on voltage regulation in distribution systems using machine learning,’’ IEEE Access, vol. 9, pp. 40402–40416, 2021, doi: 10.1109/ACCESS.2021.3064689.

[19] R. Bapat, ‘‘Identifying malicious botnet traffic using logistic regression,’’ in Proc. Syst. Inf. Eng. Design Symp. (SIEDS), Apr. 2018, pp. 266–271, doi: 10.1109/SIEDS.2018.8374749.

[20] A. Kajal and G. Sardana, ‘‘Protection from cyber attacks using IDS security mechanism with random forest classifier: A review,’’ J. Crit. Rev., vol. 7, no. 19, p. 8516, 2020.

[21] S. Ustebay, Z. Turgut, and M. A. Aydin, ‘‘Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier,’’ in Proc. Congr. Big Data, Deep Learn. Fighting Cyber Terrorism, Dec. 2018, pp. 71–76, doi: 10.1109/IBIGDELFT.2018. 8625318.

[22] M. Malik, M. W. Iqbal, S. K. Shahzad, M. T. Mushtaq, M. R. Naqvi, M. Kamran, B. A. Khan, and M. Usman Tahir, ‘‘Determination of COVID19 patients using machine learning algorithms,’’ Intell. Autom. Soft Comput., vol. 31, no. 1, pp. 207–222, 2022, doi: 10.32604/iasc.2022. 018753.

[23] N. M. Chayal and N. P. Patel, ‘‘Review of machine learning and data mining methods to predict different cyberattacks,’’ in Data Science and Intelligent Applications, Singapore, 2021, pp. 43–51.

[24] S. Ali, A. Rauf, N. Islam, H. Farman, and S. Khan, ‘‘User profiling: A privacy issue in online public network,’’ SINDH Univ. Res. J., vol. 49, pp. 125–128, Mar. 2017.

[25] A. Sapienza, A. Bessi, S. Damodaran, P. Shakarian, K. Lerman, and E. Ferrara, ‘‘Early warnings of cyber threats in online discussions,’’ in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Nov. 2017, pp. 667–674, doi: 10.1109/ICDMW. 2017.94.

[26] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, ‘‘Darknet and deepnet mining for proactive cybersecurity threat intelligence,’’ in Proc. IEEE Conf. Intell. Secur. Informat. (ISI), Sep. 2016, pp. 7–12, doi: 10.1109/ISI.2016.7745435.