IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 4, Ser. I (Jul. –Aug. 2021), PP 55-60
www.iosrjournals.org

Investigation and Classification of Cyber Crime using Deep Learning

Prof. Ms. A. B. Bavane[1], Aishwarya Chavan[2], Mayuri Gaikwad[3], Varsha Khillare[4], Sonali Kolekar[5]

[1]*Asstt. Prof. Department of Information Technology, DVVP COE Ahmednagar, Maharashtra, India*
[2,3,4,5]*Department of information Technology, DVVP COE Ahmednagar, Maharashtra, India*

Abstract: *An intrusion detection system is software that monitors a single or a network of computers for malicious activities that are aimed at stealing or censoring information or corrupting network protocols. Most technique used in today's intrusion detection system are not able to deal with the dynamic and complex nature of cyber-attacks on computer networks. Even though efficient adaptive methods like various techniques of Deep learning can result in higher detection rates, lower false alarm rates and reasonable computation and communication cost. With the use of data mining can result in frequent pattern mining, classification, clustering and mini data stream. An advanced method for intrusion detection system based on Data mining and Deep Learning is proposed in this proposal. Intrusion Detection system is divided into two types Host based IDS and Network based IDS. In this proposal, Network based IDS is used to protect computer network and its resources from malicious attacks. Based on the number of citations or the relevance of an emerging method, papers representing each method were identified, read, and summarized. Because data are so important in Deep learning and data mining approaches, well-known cyber data sets are used in Deep learning and data mining.*

Key Word: *Cyber Crime, Deep Learning, Data mining, Intrusion Detection System.*

Date of Submission: 10-07-2021 Date of Acceptance: 26-07-2021

I. Introduction

Cyber security systems are composed of network security systems and computer security systems. Each of these has, at a minimum, a firewall, antivirus software, and an intrusion detection system .Intrusion detection system s help discover, determine, and identify unauthorized use, duplication, alteration, and destruction of information systems. The security breaches include external intrusions attacks from outside the organization and internal intrusions. Recommendation The Deep learning, Data Mining methods are described, as well as several applications of each method to cyber intrusion detection problems. The complexity of different Deep learning and data mining algorithms is discussed, and the proposal provides a set of comparison criteria for Deep learning and data mining methods and a set of recommendations on the best methods to use depending on the characteristics of the cyber Problem to solve Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction. There are three main types of cyber analytics in support of intrusion detection systems: misuse-based, anomaly-based, and hybrid. Misuse-based techniques are designed to detect known attacks by using signatures of those attacks. They are effective for detecting known type of attacks without generating an overwhelming number of false alarms. They require frequent manual updates of the database with rules and signatures. Misuse-based techniques cannot detect novel attacks. Anomaly-based techniques model the normal network and system behavior, and identify anomalies as deviations from normal behavior. They are appealing because of their ability to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, thereby making it difficult for attackers to know which activities they can carry out undetected. Additionally, the data on which anomaly-based techniques alert can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviors may be categorized as anomalies.

Use of Internet services are increasing day by day the threats from the internet to computer systems, data are also increasing. Attackers can easily get access to the important data resources in our systems. It is very important to protect the data from such attackers as they can use this data for their personal needs and can sell the data for their personal needs or it can end up in wrong hands. Large amount of data is stored in the servers and computers of companies. So it is very important to make sure the valuable data is safe and secure. This can be done with the help of real time Intrusion detection system which detects any kind of suspicious activity and

alerts to the administrator to prevent attacks. This can be achieved by using various methods of Deep Leaning and Data mining.

II. Literature Survey

Songnian Li et.al.in their paper "Geospatial big data handling theory and methods: A review and research challenges", mentioned the proposed system, they made review on various geospatial theory and methods used to handle geospatial big data. Given some special attributes. Authors considered that customary data taking controlling methodologies and techniques are lacking and the following domains were recognized as in requirement for further advancement and examination in the control. This incorporates the advancements in calculations to manage real-time analytics and to support ongoing flooding data, as well as improving new spatial indexing techniques. The improvement of theoretical and methodological ways to deal with transfer of big data from illustrative and parallel research and applications to ones that investigates easygoing and illustrative connections.

Yang C, Goodchild M et.al. in the paper entitled "Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?", International Journal of Digital Earth, pp. 305-329,Vol. 4, No. 4, July 2011 have proposed another technique for overseeing gigantic remote sensing image data by utilizing HBase and MapReduce framework. At first they have divided the actual image into various tiny pieces, and store the blocks in HBase, which is dispersed in a gathering of hubs. They have used Map Reduce programming model on handling the stored blocks, which can be simultaneously executed in a group of hubs. Besides, as a result of the high versatility of Hadoop, it is anything but difficult to add new hubs to the cluster, which was normally exceptionally troublesome in general ways. Finally they notice that the speeds of data commerce and processing increase because the cluster of HBase grows. The outcomes demonstrate that HBase is extremely reasonable for large image information stockpiling and handling.

Duffy DQ, Schnase et.al. in their paper entitled "Preliminary Evaluation of Map Reduce for High-Performance Climate Data Analysis", NASA new technology report white paper, 2012 has given that Map Reduce is an approach to high-performance analytics that may be useful to data intensive problems in climate research. It offers an analysis paradigm that uses clusters of computers and combines distributed storage of large data sets with parallel computation. We are particularly interested in the potential of Map Reduce to speed up basic operations common to a wide range of analyses. In order to evaluate this potential, we are prototyping a series of canonical MapReduce operations over a test suite of observational and climate simulation datasets. Our initial focus has been on averaging operations over arbitrary spatial and temporal extents within Modern Era Retrospective-Analysis for Research and Applications (MERRA) data. Preliminary results suggest this approach can improve efficiencies within data intensive analytic workflows.

Gema Bello Orgaza et.al. in their paper titled "Social big data: Recent achievements and new challenges", Journal of Information Fusion, Science Direct, pp. 45– 59,Volume 28, March 2016 gave that Big data has become an important issue for a large number of research areas such as data mining, Deep learning, computational intelligence, information fusion, the semantic Web, and social networks. The rise of different big data frameworks such as Apache Hadoop and, more recently, Spark, for massive data processing based on the Map Reduce paradigm has allowed for the efficient utilization of data mining methods and Deep learning algorithms in different domains. A number of libraries such as Mahout and Spark MLib have been designed to develop new efficient applications based on Deep learning algorithms. The combination of big data technologies and traditional Deep learning algorithms has generated new and interesting challenges in other areas as social media and social networks. These new challenges are focused mainly on problems such as data processing, data storage, data representation, and how data can be used for pattern mining, analyzing user behaviors, and visualizing and tracking data, among others. In this paper, we present a revision of the new methodologies that is designed to allow for efficient data mining and information fusion from social media and the new applications and frameworks that are currently appearing under the "umbrella" of the social networks, social media and big data paradigms.

III. Need for Research

As the use of Internet services are increasing day by day the threats from the internet to computer systems, data are also increasing. Attackers can easily get access to the important data resources in our systems. It is very important to protect the data from such attackers as they can use this data for their personal needs and can sell the data for their personal needs or it can end up in wrong hands. Large amount of data is stored in the servers and computers of companies. So it is very important to make sure the valuable data is safe and secure. This can be done with the help of real time Intrusion detection system which detects any kind of suspicious activity and alerts to the administrator to prevent attacks. This can be achieved by using various methods of Deep Leaning and Data mining. These cyber-attacks may steal your data and corrupt the system. The cyber attackers can use the internet after getting access of specific PC by cyber-attack for the purpose of data mining

or crypto currency mining or some other illegal work. This will affect the complete network and at last the organization will face a big loss in the form of data or money.

IV. Proposed System

Python 3.6 was used to create the application files. Before running the files, it must be ensured that Python 3.6 and the following libraries are installed. Sklearn :- Machine Learning Library Numpy :- Mathematical Operations Pandas :- Data Analysis Tools Mat-plotlib :- Graphics and Visuality The implementation phase consists of 5 steps, which are: 1- Pre-processing 2- Statistics 3- Attack Filtering 4- Feature Selection 5-machine Learning Implementation. Each of these steps contains one or more Python files. The same file was saved with both ”py” and ”ipynb” extensions. The code they contain is exactly the same. The file with the ipynb extension has the advantage of saving the state of the last run of that file and the screen output.

Python

Python is an interpreter, high-level, general-purpose programming language. Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Van Rossum led the language community until July 2018. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python features a comprehensive standard library, and is referred to as ”batteries included”. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open-source software and has a community based development model. Python and CPython are managed by the non-profit Python Software Foundation.

Support Vector Machine

Support Vector Machine or SVM algorithm is a simple yet powerful Supervised Machine Learning algorithm that can be used for building both regression and classification models. SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets. Even with a limited amount of data, the support vector machine algorithm does not fail to show its magic.

Case 1: Consider the case in figure that to find the best hyperplane which can separate the Two classes. In SVM we try to maximize the distanced between hyperplane And Nearest data point

Case 2: In this case all decision boundaries are separated classes.

Case 3: In this case data is not evenly distributed on left and right.

Case 4: While selecting hyperplane, SVM will automatically ignores the data point and Selects the best performing hyperplane.

Case 5: In this case linear classifiers are highlighted and Data can be separated by any Straight line

Procedure methodology

1.Data Set:- The process of preparing data include several step .there are variations in the step listed by different data preparation.

2. Preprocess and visualize:-Data preprocessing is a data mining technique that involves transforming raw data into an understandable format Data preprocessing includes cleaning instance, selection ,normalization ,transformation features extraction & selection.

3.Training Model Using SVM:-Machine learning involves predicating & classifying data and to down employee various machine learning algorithms according to the data set.

4.User Input:- Input is the raw data that is processed output.

5.Save Model :- The data set to save and load user machine learning model in python using sklearn This allows you to save user model to file and load later in order to make predation.

6.Predict :- On attack is found is check.

V. Results and Discussion

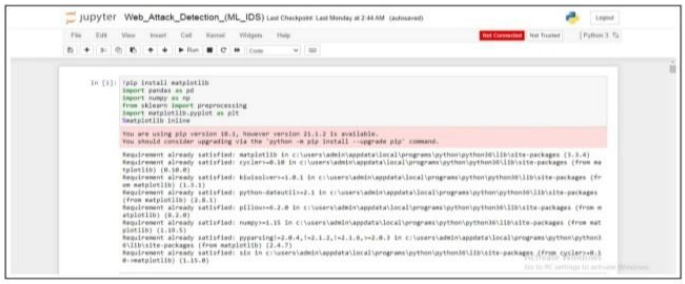
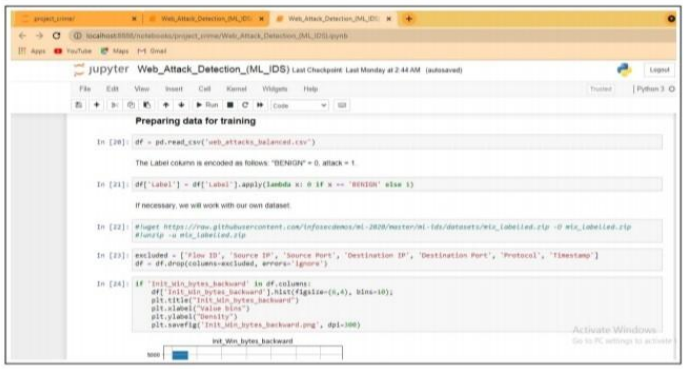


Figure 6.1: Library Import



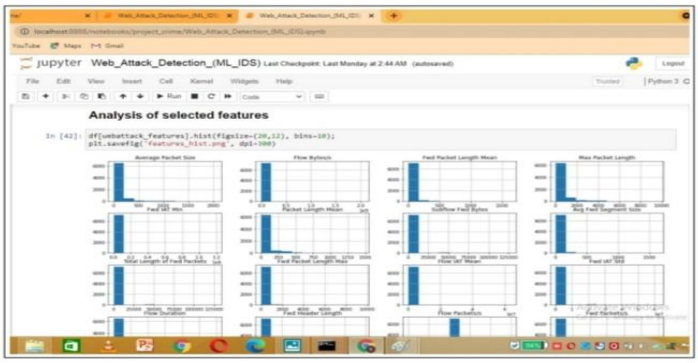


Figure 6.3: Analysis of Features

Figure 6.4: Removal of Correlated Feature

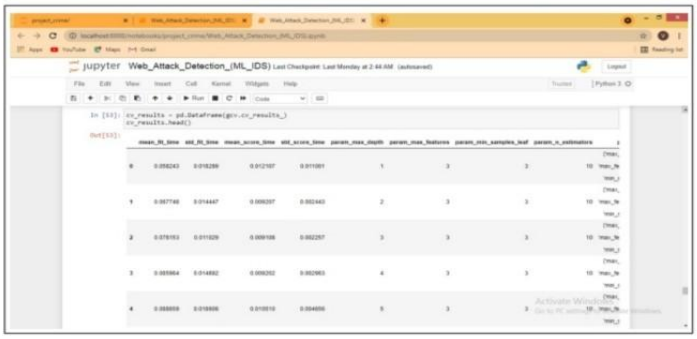


Figure 6.5: Cv*results*

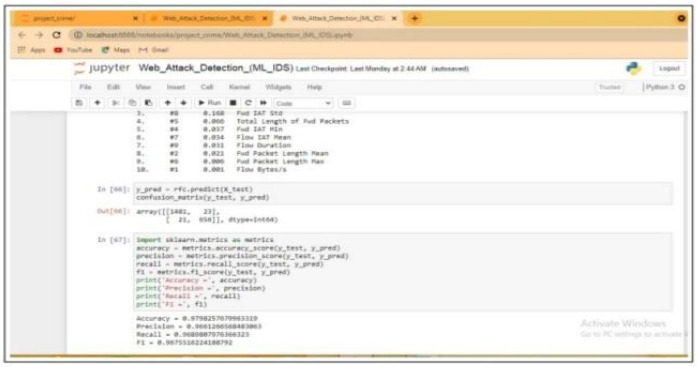


Figure 6.6: Accuracy of System

VI. Conclusion

In proposed system we presented a system were the dataset containing information about cyber attacks in the company's was been used. Then, preprocessing and normalization will be performed on these dataset using some techniques of Machine Learning. The proposed system can give best accuracy as compared to other techniques. The functionalities of this project can be scaled up in the future. These functionalities could be: Real-time data analysis of crime data: This could help us obtain crime patterns and forecasts of the future instantly using real-time datasets. Data mining of social media to generate datasets, and then preprocess and analyze them to spot trends of the current crime situation in a particular place or region. Compare and display the results of all available and applicable forecasting, predicting and classification models side by side, such that the user can select any of those methods.

References

[1] Songnian Li, Suzana Dragicevic ,Francesc Anton Castro , Monika ester, Stephan ´ Winter, Arzu Coltekin, Christopher Pettit, "Geospatial big data handling theory and methods: A review and research challenges".

[2] Deepak A Vidhate, Parag Kulkarni, 2019, "Performance comparison of multiagent cooperative reinforcement learning algorithms for dynamic decision making in retail shop application", International Journal of Computational Systems Engineering, Inderscience Publishers (IEL), Volume 5, Issue 3, pp 169-178.

[3] Yang C, Goodchild M, Huang Q, Nebert D, Raskin R, "Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?", International Journal of Digital Earth, pp. 305-329,Vol. 4, No. 4, July 2011.

[4] Deepak A Vidhate, Parag Kulkarni, 2019, "A Framework for Dynamic Decision Making by Multi-agent Cooperative Fault Pair Algorithm (MCFPA) in Retail Shop Application", Information and Communication Technology for Intelligent Systems, Springer, pp 693-703.

[5] Duffy DQ, Schnase JL, Thompson JH, Freeman SM, Clune TL, "Preliminary Evaluation of Map Reduce for High- Performance Climate Data Analysis", NASA new technology report white paper, 2012.

[6] Deepak A Vidhate, Parag Kulkarni, 2018," A Novel Approach by Cooperative Multiagent Fault Pair Learning (CMFPL)", Communications in Computer and Information Science, Springer, Singapore, Volume 905, pp 352-361.

[7] Gema Bello-Orgaza, Jason J. Jungb, David Camacho, "Social big data: Recent achievements and new challenges", Journal of Information Fusion, Science Direct, pp. 45– 59,Volume 28, March 2016.

[8] Deepak A Vidhate, Parag Kulkarni, 2018, Exploring Cooperative Multi-agent Reinforcement Learning Algorithm (CMRLA) for Intelligent Traffic Signal Control, Smart Trends in Information Technology and Computer Communications. SmartCom 2017, Volume 876, pp 71-81.

[9] Kleiman. D., Wright. C., Varsalone. J., Clinton. T., Gregg. M., 2007, "The Official CHFI Study Guide (Exam 312-49): for Computer Hacking Forensic Investigator", Published Book by Syngress.

[10] Deepak A. Vidhate and Parag Kulkarni, 2017, "A Framework for Improved Cooperative Learning Algorithms with Expertness (ICLAE)", International Conference on Advanced Computing and Communication Technologies Advances in Intelligent Systems and Computing, Springer Singapore, volume 562, pp. 149-160.
[11] Deepak A. Vidhate and Parag Kulkarni, 2017, "Expertise Based Cooperative Reinforcement Learning Methods (ECRLM)", International Conference on Information & Communication Technology for Intelligent System, Springer book series Smart Innovation, Systems and Technologies (SIST, volume 84), Springer Cham, pp 350-360.
[12] Amol Borkar, Akshay Donode, Anjali Kumari "A Survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and Protection System (IIDPS)" Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.
[13] Deepak A. Vidhate and Parag Kulkarni, 2016, "Innovative Approach Towards Cooperation Models for Multi-agent Reinforcement Learning (CMMARL) "International Conference on Smart Trends for Information Technology and Computer Communications Springer, Singapore, pp. 468-478.
[14] Deepak A. Vidhate, Parag Kulkarni, "New Approach for Advanced Cooperative Learning Algorithms using RL methods (ACLA)" VisionNet'16 Proceedings of the Third International Symposium on Computer Vision and the Internet, ACM DL pp 12-20, 2016.
[15] S. Zegeye, B. De Schutter, J. Hellendoorn, E. A. Breunesse, and A.Hegyi, "A predictive traffic controller for sustainable mobility using parameterized control policies," IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 3, pp. 1420–1429, 2012.
[16] Deepak A. Vidhate, Parag Kulkarni, "Enhanced Cooperative Multi-agent Learning Algorithms (ECMLA) using Reinforcement Learning", International Conference on Computing, Analytics and Security Trends (CAST), IEEE Xplorer, pp 556-561, 2017.

Prof. Ms. A. B. Bavane, et. al. "Investigation and Classification of Cyber Crime using Deep Learning." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(4), 2021, pp. 55-60.