Cyber Safety System by Implementing Integration of Expert Learning and Data Mining Concepts

Mrinal Paliwal

Abstract— ***the continuous growth in the rate of cyber-attacks in recent years uplifts the worry for the cyber security of industrial control systems. The current efforts of the cyber security system are depended on firewalls, data diodes and other basic methods for prevention of infringement. A cyber threat, intrusion or infringement detection system detects malicious or noxious activities by scanning a system and investigate digitally by employing "machine learning" and "data digging" techniques for handling dynamic and complex functioning of malicious assaults in computer systems and extracting essential information from an input data. In this research paper, the techniques we have used to complete this research may bring advancement in recognition rates, decrease the fault rate which also led to a decrease in the cost factor..***

Rundown phrases— ***threat, infringement, cyber, machine learning, data digging.***

I. INTRODUCTION

Cyber safety is fundamentally related to the framework, method, and techniques, practices introduced to safeguard the cyber networks including network devices, protocol and data from a malicious activity like unauthorized access, stealing of data, damage of data or any kind of manipulation of protected data. Safety towards computer devices and the network is essential because the government, corporate, medical and other sectors in an area or country generate data, a process that generated data and load that data on computer storage devices. A particular part of the stored data may be confidential data which cannot be shared, it may be intellectual data, transaction data, account details which can be manipulated by unlawful access leading to undesirable consequences. Many confidential, security firms transfer secured, confidential or sensitive data by communication networks, cyber safety field disclose the protocols related to the protection of that transmitted information and the information system that is used in the above-mentioned process. With the increasing growth of cyber-attacks rates in organizations and worldwide led to the invention of cyber safety systems. The machine teaching and data mining approaches presented in this article are entirely relevant to problems of cable and distant devices interfering and detecting crime. Various types of digital investigation techniques aid for recognition of intrusion. Some of them are exploitation based, irregularity based and hybrid technique. Exploitation based approaches are proposed to recognize acknowledged attacks by processing the marks of those malicious users deprived of generating an attention blogging amount of incorrect cautions. But they lack in detecting new attacks because they don't have any stored parameters and marks. Irregularity based method compares the original system and the directed system and detects the irregularity as additional deviations from the original system.

Present-day situation of the intrusion detection system detects the inside and outside mischievous activities happening to the system. The protection of devices against disruptions or attacks is becoming more challenging each day as cyber-attacks in the networks are very technologically advanced in nature and increasing rapidly. More internet consumers have increased the probability of information harm, fraud and interference.

II. RELATED LITERATURE

A published paper [1] conducted an inquiry into various geospatial hypotheses and methods for dealing with enormous geospatial data. Due to a number of unusual characteristics, researchers felt that the normal data about monitoring concepts. Another researcher in the paper [2] suggested an additional method for the regulation of huge distant image detection by HBase and MapReduce. In the early stages, they divided the true picture into several little fragments and kept the HBase squares dispersed on a social event. They employed the MapReduce programming model to handle the removed fragments that could be carried out in a collection of midpoints around the same time. The middle positions in the Hadoop cluster have no inferior and to get an economic output as a goal of their researcher. Furthermore, it is certainly not difficult to incorporate fresh facilities for this community, due to the elevated adaptability of Hadoop. Finally, the steps of the trade and management of data have increased because of the development of HBase. The findings show that HBase is sensitive to the collection and handling of significant image information.

Another developer presented a paper [3] expected the replacement of huge NetCDF logical information that is maintained pursuant to Hadoop in parallel to capability and entry technology. On Map Reduce the recovery scheme is implemented. In order to demonstrate the suggested approach, the Argo data is used. The execution is done by identifiable data and different task digits in a space given the personal computer. The results of the assessments show the simultaneous technique which can be used for the profitability of the huge NetCDF scale. Huge data became a major feature of the general picture and logically reveals the

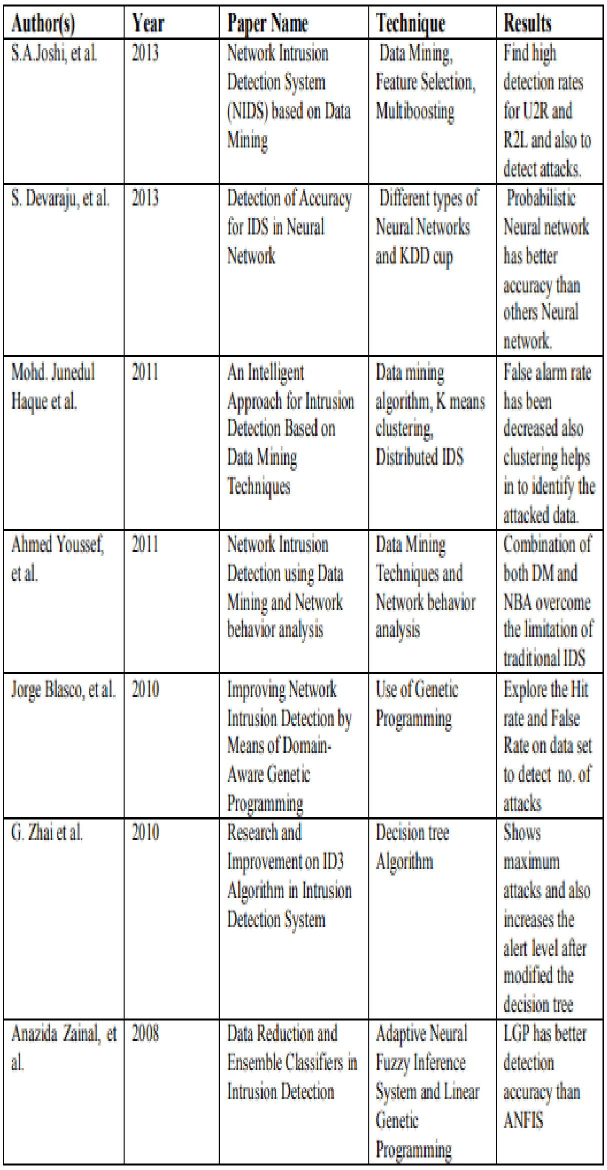
Revised Manuscript Received on 14 September, 2019.

Mrinal Paliwal, Department of Computer Science and Engineering, Sanskriti University, Uttar Pradesh,India. (E-mail: sanpubip@gmail.com)

organization, sector, state and other membership that is notified.

Another researcher presented a paper in 2013 [4] which disclose about the enhanced growth in technology, cyberspace security is one of the greatest stimulating issues. And also talked about increasing the efficiency and accuracy of the system.

Various other research papers surveyed disclosed about “anomaly-based network intrusion detection, techniques, systems and challenges” but some drawbacks were there like it does not present a full set of state of the art machine learning methods, no proper explanation of the technical details of methods disclosed. Table 1 shows a comparative study on such existing techniques for intrusion detection.



Author(s)	Year	Paper Name	Technique	Results
S.A.Joshi, et al.	2013	Network Intrusion Detection System (NIDS) based on Data Mining	Data Mining, Feature Selection, Multiboosting	Find high detection rates for U2R and R2L and also to detect attacks.
S. Devaraju, et al.	2013	Detection of Accuracy for IDS in Neural Network	Different types of Neural Networks and KDD cup	Probabilistic Neural network has better accuracy than others Neural network.
Mohd. Junedul Haque et al.	2011	An Intelligent Approach for Intrusion Detection Based on Data Mining Techniques	Data mining algorithm, K means clustering, Distributed IDS	False alarm rate has been decreased also clustering helps in to identify the attacked data.
Ahmed Youssef, et al.	2011	Network Intrusion Detection using Data Mining and Network behavior analysis	Data Mining Techniques and Network behavior analysis	Combination of both DM and NBA overcome the limitation of traditional IDS
Jorge Blasco, et al.	2010	Improving Network Intrusion Detection by Means of Domain-Aware Genetic Programming	Use of Genetic Programming	Explore the Hit rate and False Rate on data set to detect no. of attacks
G. Zhai et al.	2010	Research and Improvement on ID3 Algorithm in Intrusion Detection System	Decision tree Algorithm	Shows maximum attacks and also increases the alert level after modified the decision tree
Anazida Zainal, et al.	2008	Data Reduction and Ensemble Classifiers in Intrusion Detection	Adaptive Neural Fuzzy Inference System and Linear Genetic Programming	LGP has better detection accuracy than ANFIS

III. PROPOSED WORK

Proposed new methods, such as CFS subgroup algorithm and the mobile network with WEKA instruments, in order to solve the current issue. The CFS subgroup selects the most significant and common features of the technology. The preference of features is intended to recognize and remove useless and inadequate features. The attribute and trait assessment are highly coefficient.

Completely fair scheduler algorithm

The attribute selection is a method which enables selection of true sub-set as the appropriate function. The choice of the feature in the domain of information preprocessing in information mining is the most common and significant method. The attribute selection is for the identification and removal of useless and unsuitable attributes. Supervised and unsupervised learning are the two most common type of learning process and this function can be implemented with both techniques. The sub-set attributes of the optimality are evaluated by assessment requirements. The domain dimension expands in N number of features. Detecting a subset of ideal feature is usually intransigent and the NP-hard has been shown many other problems appropriate to the choice of features. Some steps are been carried out in the selection method i.e. firstly the production of subset, secondly valuation of subset and discontinuing standard and lastly checking the outcome.

A neural network is considered as another method, where multilayer perceptron, logistic regression and learning features are used. The multilayer perceptron is employed specifically for training the neural network. Exploration of regression that is employed for the estimating output. It is also used in the evaluation of linking between the independent variable and dependent variables which may be multinomial. The linear correlation coefficient for measuring characteristic is given below:

$$\mathrm{Correlation(r)} = \frac{N\sum XY - \sum X - \sum Y}{\sqrt{N\sum X^2 - \sum X^2\, N\sum Y^2 - \sum Y^2}}$$

$$H(Y) = -\sum_{yR_y} p(y)\,\log(p(y))$$

$$H(Y/X) = -\sum_{yR_y} p(x) \sum_{yR_y} p(y/x)\,\log(p(x/y))$$

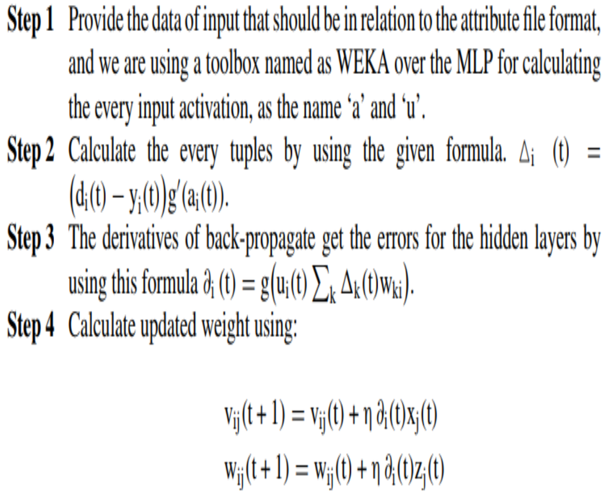
$$C(Y/X) = \frac{H(Y) - H(Y/X)}{H(Y)}$$

In the above equation:

X, Y belongs to two attributes.

For precise result, multilayer perceptron is employed in which neural network learns by a set of weight for labeling of class, wherein the label here used is an attack on the network. And the time consumed in training is reduced.

Algorithm of the proposed method is



Step 1 Provide the data of input that should be in relation to the attribute file format, and we are using a toolbox named as WEKA over the MLP for calculating the every input activation, as the name 'a' and 'u'.

Step 2 Calculate the every tuples by using the given formula. $\Delta_i\ (t) = \left(d_i(t) - y_i(t)\right)g'(a_i(t))$.

Step 3 The derivatives of back-propagate get the errors for the hidden layers by using this formula $\partial_i\ (t) = g\left(u_i(t) \sum_k \Delta_k(t) w_{ki}\right)$.

Step 4 Calculate updated weight using:

$$v_{ij}(t+1) = v_{ij}(t) + \eta\, \partial_i(t) x_j(t)$$

$$w_{ij}(t+1) = w_{ij}(t) + \eta\, \partial_i(t) z_j(t)$$

Figure 1 illustrates the proposed framework architecture comprising of massive datasets with discrete volume status to increase the rate of performance of intrusion uncovering system that presents the malicious activities detected with the implementation of Waikato environment for knowledge analysis.

Waikato environment tool is a java programming language based program considered as a machine learning software. It is accessible on any computer platform and it is open-source software. It also supports data mining techniques like gathering, data preprocessing and regression. In the proposed method the aforementioned tool is used to enable the detection of concealed information or data chunks from the file present on the system and even in the database. This tool is controlled by a user interface.

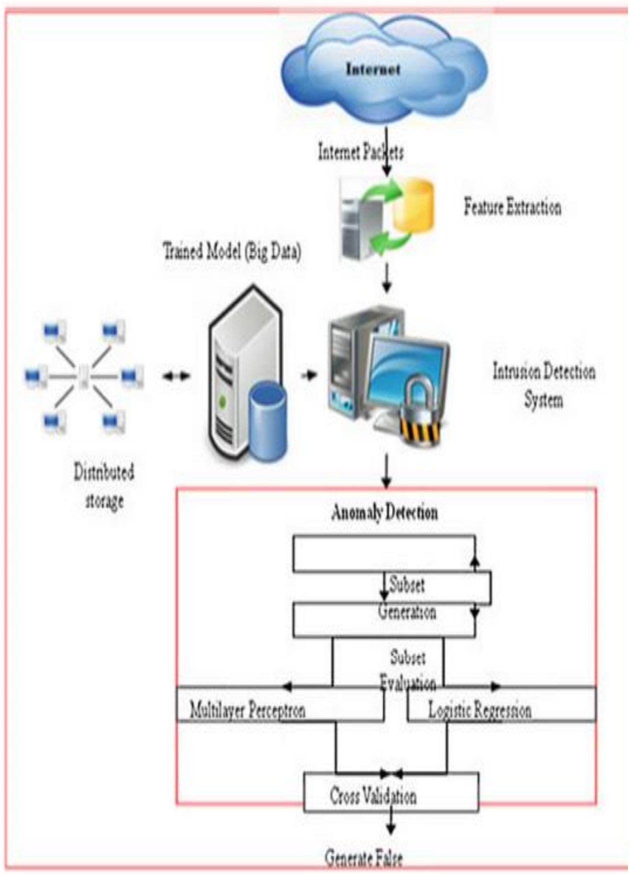


Figure 1 Proposed framework

IV. RESULT

A dataset is selected and various experiments are performed on that selected database to compute the performance of the proposed framework. The arrangement on which the trials are been carried out: window7, Intel processor with 2.00 GHz speed respectively.

The selected dataset comprises of train data belonging to 2000 linking accounts and 5000 test data. Dataset also includes a group of 41 selected attributes from every link also a collection of tag for detecting the link status, where the status may be normal or attacked labeled.

Figure 2 explains the summary of different implementation instants with different dataset sizes. The suggested intruder tracking scheme requires less moment on all levels rather than traditional computer training methods. The reason behind this is less training dataset.

Figure 3 illustrates the communications network anomaly identification frequency. Almost all types of assaults like samples, DoS, U2R and R2L are identified in the suggested intrusion uncovering system. The identification frequency for anomalies relies on the test information. The presumed data settings will function as an intrusion dataset.

Figure 4 expresses the comparison between proposed and other methods on the basis of the size of the dataset, approaches followed.

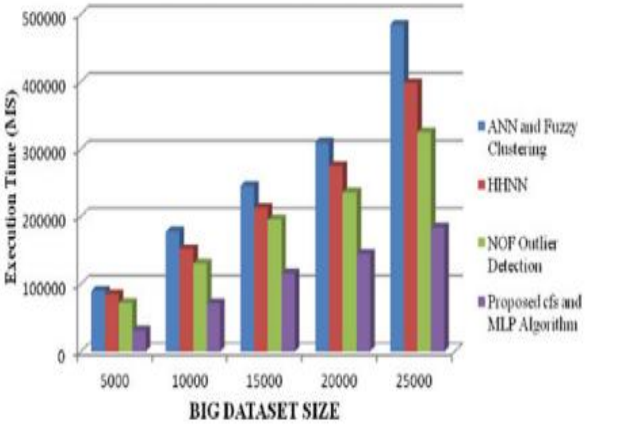


Figure 2 Vast dataset dimension versus performance time

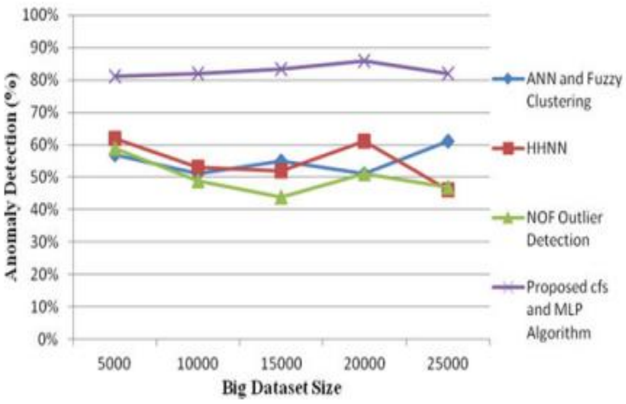


Figure 3 Communications network irregularity identification frequency

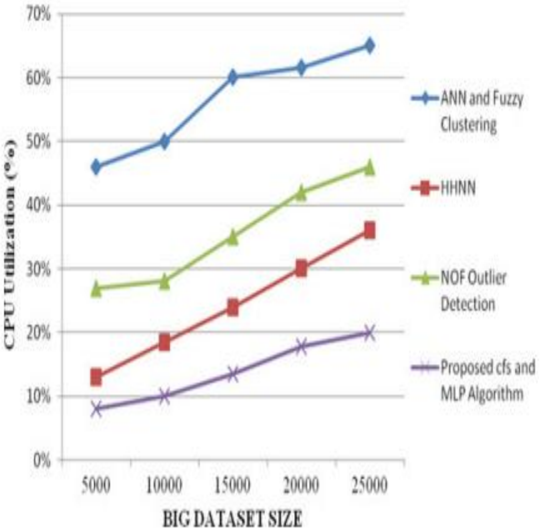


Figure 4 Comparison between proposed and other methods

V. CONCLUSION

This work presented a new technique for spotting interruption present in the network by a completely fair scheduler algorithm and neural network where the training model comprises two vast amount dataset with the scattered base that enhances the process of intrusion detection. On comparing with the existing infringement detection methods proposed method takes less time for implementation and loading the dataset. Here in this paper, the performance rate of the proposed framework is higher than other existing frameworks. In the future, the proposed work might be employed in numerous computation purposes.

REFERENCES

1. Songnian Li, SuzanaDragicevic, FrancesAnton Castro, Monika Sester, Stephan Winter, ArzuColtekin, Christopher Pettit, "Geospatial big data handling theory and methods: A review and research challenges", Volume2 | Issue2 || March-April-2017 | www.ijsrcseit.com 97 ISPRS Journal of Photogrammetry and Remote Sensing, pp. 119-133, Volume 115, May 2016.
2. Yu Zheng, "Methodologies for Cross-Domain Data Fusion: An Overview", IEEE Transactions on big Data, pp. 16-34, Volume:1, Issue:1, TBD-2015-05-0037, March 2015.
3. Yu Zheng, "Crowdsourcing geospatial data", ISPRS Journal of Photogrammetry and Remote Sensing, ScienceDirect, pp.550-557, Volume 65, Issue 6, November 2010.
4. S.A.Joshi, Varsha S.Pimprale "Network Intrusion Detection System (NIDS) based on Data Mining" International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013.
5. Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein. "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model." Journal of Computational Science 25 (2018): 152-160.
6. Anazida Zainal, Mohd Aizaini Maarof and SitiMariyam Shamsuddin "Data Reduction and Ensemble Classifiers in Intrusion Detection" in 2008 IEEE.
7. Ndonda, Gorby Kabasele, and Ramin Sadre. "A Two-level Intrusion Detection System for Industrial Control System Networks using P4." Proceedings of the 5th International Symposium for ICS & SCADA Cyber Security Research. 2018.
8. Choi, Wonsuk, et al. "VoltageIDS: Low-level communication characteristics for automotive intrusion detection system." IEEE Transactions on Information Forensics and Security 13.8 (2018): 2114-2129.
9. Liu, Chih-Hsiung, Shaw-Ben Shi, and Yu Chen Zhou. "Multi-sensor intrusion detection system." U.S. Patent Application No. 10/026,283.
10. Guangqun Zhai, Chunyan Liu "Research and Improvement on ID3 Algorithm in Intrusion Detection System" in 2010 IEEE.
11. Jorge Blasco, Agustin Orfila, Arturo Ribagorda "Improving Network Intrusion Detection by Means of Domain-Aware Genetic Programming" DOI 10.1109/ARES.2010.53 in IEEE 2010.
12. Ahmed Youssef and Ahmed Emam "Network Intrusion Detection using Data Mining and Network ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume No. 2, Issue No. 6, June 2013 2194 www.ijarcet.org Behavior Analysis" International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011
13. Mohd. Junedul Haque, Khalid.W. Magld, Nisar Hundewale "An Intelligent Approach for Intrusion Detection Based on Data Mining Techniques" in 2012 IEEE.
14. N.S.Chandolikar, V.D.Nandavadekar "Comparative analysis of two algorithm for Intrusion attack classification using dataset" in International Journal of Computer Science and Engineering (IJCSE) in 2012.
15. Devendra kailashiya, Dr. R.C. Jain "Improve Intrusion Detection Using Decision Tree with Sampling" in IJCTA | MAY-JUNE 2012.
16. S.A.Joshi, Varsha S.Pimprale "Network Intrusion Detection System (NIDS) based on Data Mining" International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013.
17. S. Devaraju, S .Ramakrishnan "Detection of Accuracy for Intrusion Detection System using Neural Network Classifier" International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013).
18. Yacine Bouzida, Frederic Cuppens "Neural networks vs. decision trees for intrusion detection" in 2011.