An Enhanced Intelligent Intrusion Detection System using Machine Learning

Dhikhi T, M.S. Saravanan

Abstract: ***Numerous Intrusion detection techniques are used to find the anomalies that depends on the accuracy, detection rate etc. The purpose of the system is to detect the anomalies based on the given dataset thereby improving the accuracy. A CWS IDS is proposed to find the anomalies in the network, that combines machine learning techniques autoencoder and support vector machine for feature extraction and classification. This is evaluated on the training and testing datasets of NSL KDD dataset that accomplishes well in terms of reduction rate and precision. By combining autoencoder and support vector machine for finding the anomalies, the performance metrics of the system is improved.The system is related with single SVM and Random forest classifier. The performance measures such as precision, recall, accuracy and F-measure is equated with the SVM, random forest, and CWS IDS for training data and test data. Thereby the recognition rate is enhanced and both false positives, false negatives are lesser.***

Keywords: Contractive Encoder, Intrusion detection, NSL-KDD Dataset, Support Vector Machin,.

I. INTRODUCTION

Due to the extensive use of network information, Intrusion Detection has turned out to be a network security challenge. The primary assignment of Intrusion Detection is to perceive the unseen attacks in the network or a system. Intrusion Detection System can exploit anomaly detection technique or signature detection technique which finds novel attacks and known attacks respectively. Signature intrusion detection techniques identify attacks on the source of rules that are already defined in the network so it is practical to distinguish only recognized attacks in the network. In complex intrusion detection system standard behavior of the traffic is studied, if any traffic which deviates the normal pattern is defined as intrusion. As novel attacks can be found using anomaly detection techniques it is highly advantageous than signature based intrusion detection techniques. Intrusion Detection algorithms can be applied for both network and a system. According to anomaly detection technique the network stream of traffic that violates from regular activities pattern is categorized as intrusion. The process of anomaly detection includes classifying the features of abnormal traffic using any of the modern wide techniques. Many methodologies are used to categorize the attack by feature selection using machine learning and fuzzy logic.

Machine learning approach for the classification of attacks can be supervised learning, unsupervised or semi supervised approach. The different machine learning classification algorithms are linear classifiers, support vector machines, decision trees, and random forest, nearest neighbors, logistic regression, naïve Bayes, auto encoders, and deep belief networks. In supervised learning, all the input data is labeled and the output learns to predict from the input data while in unsupervised information, all the info information is unlabeledand learns to inherent output from input data. Semi supervised algorithms are mixture of supervised and unsupervised learning.

Intrusion detection procedures are approved on an average dataset, KDD. The NSL– learning detection and information mining (Knowledge Discovery in Databases - KDD) [17] dataset, it is an enhanced type of KDD which is viewed as a standard in the assessment of interruption detection strategies to prepare all models on preparing dataset while never presenting test dataset to the model amid preparing and after that assessed the models on testing datasets. The attacks that are classified in KDDCUP99 dataset are different type of Denial of Service, User to Root, Remote to Local and Probing.

NSLKDD dataset comprise of 41 input in addition to class names. Input 1 to 9 characterize to the essential highlights made from TCP/IP association devoid of payload review. Features from 10 to 22 involved substance highlights, created since the payload of TCP fragments of packets. Input from 23 to 31 be removed from instance sensitive traffic properties while highlights 32 to 41 enclose relevance based traffic types that were intended to gauge intruder inside interims longer than 2 seconds.

II. RELATED WORK

Nathan Shone et al [2] offers novel deep learning methodology for interruption detection that compromises deep learning classification demonstrate built utilizing stacked NDAEs. This has been employed in graphics processing unit (GPU)-enabled Tensor Flow assessed spending the standard KDD Cup ’99 and NSL-KDD datasets. They also calculated the training time necessary for stacked NDAE model, moreover a DBN model to investigate the KDD ’99 dataset which provides great stages of accuracy.

I. Ahmad et al[1] analyzed the well-known machine learning techniques viz. support vector mechanism and extreme learning machine. The dataset NSL and data mining datasets are taken for the assessment of interruption detection mechanism. In their analysis result it’s concluded that ELM is more precise than RF, SVM on complete data samples and SVM more precise on partial samples, besides in quarter dataset SVM is better.

Revised Manuscript Received on July 05, 2019

Dhikhi, Research scholar in Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. She is currently Assistant Professor with SRM Institute of Science and Technology, Ramapuram, Chennai India

Dr. M.S. Saravanan, Associate Professor in the Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

M. Al-Qatfet al.[4] proposed a IDS technique using self-taught learning(STL) which is an active deep learning technology for feature learning and dimensionality. This is made using the scant auto encoder device that is a good learning method for restructuring a novel feature illustration in an unsupervised manner. The paper presently enhances SVM categorization accurateness and faster training and testing times. Moreover it reveals upright calculations in two-category and five-category classification. A higher precision rate in five-category classification is achieved in this approach when compared to other shallow classification means like J48, Naive Bayesian, RF, and SVM.

C. Xu et al[5] introduced a deep learning speculation for IDS that uses feature extraction that develop a deep learning model. He proposed intrusion detection that comprises of a discontinuous neural system with gated recurrent units (GRU), multilayer perceptron (MLP), softmaxmodule. The research was prepared on both KDD dataset and NSL-KDD data sets. This paper concluded that the outcome of BGRU and MLP together for the KDD 99 and the NSL-KDD datasets is better.

Naseer et al[6] investigated suitable approach for anomaly based IDS created on various deep neural networks such as convolutional neural systems, auto encoders, and periodic neural systems. These were skilled on NSLKDD dataset and estimated on NSLKDDTest+ and NSLKDDTest21 and performed on a GPU-based test bed using keras with theano back end. In this, evaluations were done using organization metrics viz. recipient working attribute, area under curve, precision-recall curve, mean average precision and accuracy of classification for deep as well as conventional machine learning techniques.

M.H. Ali et al [7] introduced an established knowledge model for fast learning network (FLN) supported particle swarm optimization (PSO) is planned. This is functional to the matter of detecting an intruder and valid supported the eminent dataset KDD99. The established system is correlated against a good vary of meta-heuristic systems to tutor extreme learning system as well as FLN classifier. PSO-FLN has beaten different learning approaches within the testing accuracy of the training. Much differentiation has been accomplished with a special variety of neurons within the unseen layer of FLN, and therefore the unique ELM that improve the FLN guidelines to boost the IDS accurateness in the work projected many rules like Genetic Algorithm, Harmony Search improvement (HSO)[15]. P. Tao et al [8] proposes a new genetic procedure based on the features of SVM and GA algorithm known as FWP-SVM-genetic algorithm. This method lessens the SVM error rate by using a feature selection strategy of genetic algorithm with amending the fitness algorithm. The characteristic weights and constraints of SVM are optimized simultaneously, allowing to optimal feature subset. The outcome of this paper describes escalation in correct positive rate and decline the error speed.

Q. Zhang et al [9] used kernel-based fuzzy – rough set for validating IDS and assessed using KDD 99 dataset. These fuzzy classifiers can work with the inaccuracy and vagueness of discrete, noise data thereby it performs well concerningreduction impact and precision. The feature selection methods have been typically used laterally with classifiers for network interruption recognition, AL-JARRAH et al [14] presented a multiple randomized meta-learning technique called T-IDS that rely on data partitioned learning model. Due to the precision and less training time on botnet dataset, this technique is more advantageous that other machine learning techniques such as random tree, C4.5 and sequential minimal optimization. Moreover different techniques are used to detect botnet intrusions namely Voronoi clustering-based data segregating techniques and innovative feature ranking.

H. Peng et al [10] used better feature selection, FACO procedure merging an algorithm named ant colony optimization algorithm with feature collection. FACO is implemented for betterment cataloging of different classifiers. This optimization algorithm is a simulation optimization algorithm that builds a comprehensive directed graph over n features, mimics the scavenging behavior of ants. Additionally the redundant features are designated thereby reducing the time hurdles of classification algorithms and enlarge the precision of traffic allotment. Development in the path transfer possibility mode of ant colony is done. In the interim, the two phase pheromone stimulating guideline was applied to add pheromones to preserve the calculation from falling into a neighborhood ideal prematurely.

Z. Wang et al [12] complete paper evaluates various algorithms of intrusion detection domain that use deep learning strategies and identified diverse component use designs for the assault algorithms. The study suggests that the most usually utilized highlights demonstrate its more contributed to the exposure of the intense knowledge established intrusion detection and accordingly they justify added consideration. Moreover it provides better security in the identification along with barrier endeavors.

NISIOTI et al [11] gives a complete review of unsupervised and hybrid strategies for interruption identification, examining their potential in the space. This present and feature the significance of highlight building methods and also confer current IDSs ought to advance from basic location to relationship and attribution. Advanced data analytics techniques can be used to recreate and associate attacks to recognize attackers. This paper also proposed three new modules concerning the outbound network communication. PCA method, allow to convert a large dataset into a novel, minor and uncorrelated for feature selection and dim

III. PROPOSED WORK

The proposed system designed to combine the supervised and unsupervised learning algorithms for improving the accuracy and performance system. This model consists of different phases such as dataset, preprocessing, feature extraction, classification, and detection evaluation. The system is portrayed in the fig.1 with diverse stages and the flow of each point to the other is also illustrated.

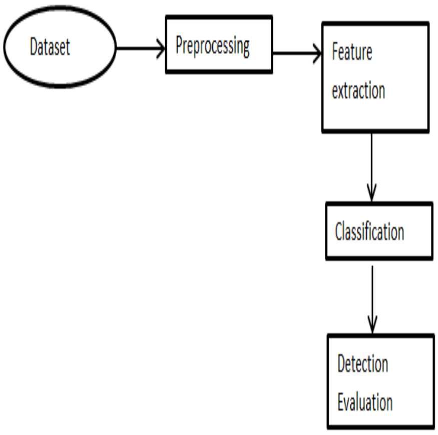

Fig.1 Intrusion Detection Model

A. Dataset

The accomplishment of the organization is directly proportional to the rightness of the dataset such that the data selection is an important task. KDD 99 [13] is used for assessment of anomaly detection, holds a set of data to inspect which includes a wide variety of intrusions simulated in it. KDD training dataset comprehends 41 features. It is characterized as either an attack or normal which states assessment of the attacks detected. NSL-KDD[3] is a data set projected to unravel approximately of the key difficulties of the KDD99 data set is sensible. All the records in NSL-KDD train and test sets are sensible. This leads cheap for executing the experimentations on complete set of requirement for arbitrarily selecting a minor portion.

B. Preprocessing

Preprocessing is done to eliminate the non-numeric, symbolic features that are not involved in the detection process. The classifier is not able to process these types of symbolic data improving the performance of detection progression.

C. Feature Extraction

Feature extraction is the procedure of gathering explicit information from a set of samples. Further feature is selected from the subset of those extracted for a particular domain. Feature extraction can be done using auto encoders, an unsupervised learning technique. Contractive Auto encoder (ContAE) is an approach to avoid uninteresting solutions to add an explicit term in the loss that penalizes the solution. It is used to learn the features by studying depictions which are forceful to trivial variations in training data. This is accomplished by commanding a fine span established on Frobenius norm of the Jacobian matrix for the encoder initiations according to the input sample. This is calculated using the formula (1). According to [16], a fine term is supplementary to the cost function which is sensitive to training input. This penalty term help in learning depictions equivalent to non-linear feature space but balanced to the maximum of guidelines equals to the characteristic break.

$\|J_h(x)\|^2_F=\sum_{ij}(\partial h_j(x)/\,\partial x_i)(1)$

Loss function formula for ContAE is calculated as:

$TCAE(\theta)=\sum_{x\,\in D n}(L(x, g(f(x)))+\lambda\,\|J_f\,(x)\|F^2$ (2)

Where ||Jf (x)||F2 is the Frobenius norm of the jacobian matrix, ie. the addition of squares over all elements inside the matrix.

In ContAE the learned encoding will be related for very alike inputs. It is possible to train the model by demanding that the derivative of the hidden layer is little in accordance to the input. ie, if there is small change in the input, analogous encoded state should be conserved. Fig. 2.shows that similar inputs are contracted to a constant output within a neighborhood, based on what the model observed during training [19].

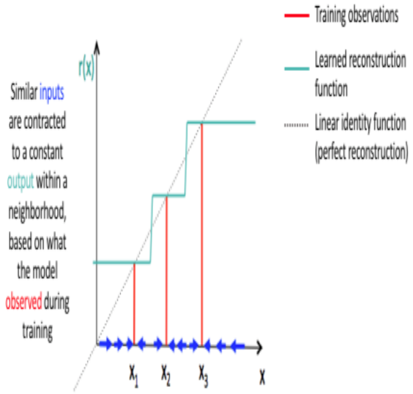

Fig. 2 Slope of ContAE

D. Classification

Support Vector Machine (SVM) is for classifying according to binary or multi classification to separate from group of optimistic instance from a group of harmful instances. This is done by a hyperplane that isolates its training data hence the distance between the hyperplane and the neighboring point from each session is exploited. SVM determine which class the datapoint fits to. Structural risk minimization rules are followed which solves regression and classification task in this manner enlightening correctness and performance. The proportion of erroneous classification will be high if the training set has uneven quantity of negative and positive set where the report of datain different classes are unstable. The basic plan of weighted support vector machine (WSVM) is to apportion each information a dissimilar weight agreeing to its comparative significance within category such altered information has completely different role to the learning of the result evident. Weighted SVM is (3)

$\min F(\omega,b,\mathcal{E})=½\,\|\,\omega\|^2+C\lambda_{vi}\sum_{I=1}\mathcal{E}_i$

$s.t y_i\,(\omega^r\,\Phi(x_i)+b)\geq 1-\mathcal{E}_{i,}$

$\mathcal{E}_i\geq 0, i=1,2,\ldots.l$ (3)

The decision function is (4)

$f\,(x)=\operatorname{sign}((\omega.\emptyset\,(x))+b\,)$ (4)

E. Detection Evaluation

The system CWS IDS is assessed on the dataset NSL-KDD that consists of full, half and one-fourth data set with 65535, 32767, 18383 samples respectively. The evaluation metrics are considered and compared and s can be classified as follows:

True positive (TP): irregularity cases properly categorized as an anomaly.

False positive (FP): ordinary cases imperfectly categorized as an anomaly.

True negative (TN): regular cases appropriately categorized as normal.

False negative (FN): abnormality cases erroneously categorized as normal.

Retrieval Number :H6932068819/19©BEIESP
DOI: 10.35940/ijitee.H6932.078919

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The following metrics are considered
Accuracy: tells the fraction of accurate classification of the entire records in the testing set, as shown in (6).
A = (TP+TN) / (TP+TN +FP+FN) (6)
Precision: tells fraction of right estimate of intrusion with overall of predictable intrusions as in (7).
P = TP/ (TP+FP) (7)
Recall: tells the fraction of approved estimate of intrusions separated by the full amount of legitimate intrusion possibilities in the testing set, as in (8).
R=TP/(TP+FN) (8)
F-measure: is measured excessive crucial metric of system ID that bank on prediction and recall, as in (9).
F = (2∗P∗R) / (P+R) (9)

IV. PERFORMANCE EVALUATION

The performance of model is compared with single SVM and Random forest classifiers. All the performance metrics are higher than the existing. The training and testing periods of the CWS IDS are minor than single SVM. Thus the model is proficient compared to individual SVM. The performance metrics that are evaluated in the detection evaluation is equated with the SVM, random forest, and CWS IDS for training data and test data. Figure 3 and 4 represents the performance metrics after the calculation on training dataset and test dataset respectively.

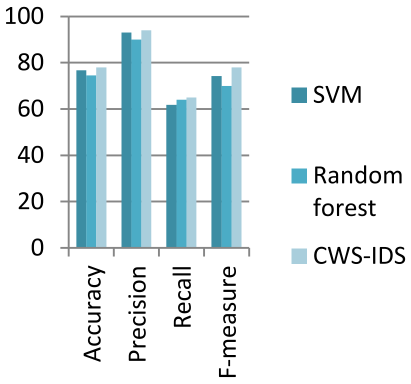

Fig. 3 Comparison of performance metrics on training dataset

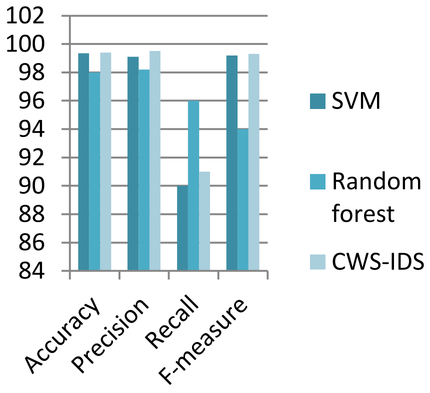

Fig. 4 Comparison ofperformance metrics on test dataset

V. CONCLUSION

The proposed system CWS IDS is an improved intrusion technique that uses machine learning techniques for feature selection and classification. This method is a promise for reducing the false positives and wrong negatives. The model has compared with the existing SVM and RF techniques for IDS and outperformed the current learning approaches in testing and training accuracy. Further step can be done by applying this to the real network for implementing it more efficiently. This can be applied to all class categories classification for an enhanced performance.

REFERENCES

1. Iftikhar Ahmad , Mohammad Basheri, Muhammad JavedIqbal, And Aneel Rahim" Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection" IEEE Transactions on SPECIAL SECTION ON SURVIVABILITY Strategies For Emerging Wireless Networks, Volume 6 May 2018 pp. 33789-33795.
2. Nathan Shone, Tran Nguyen Ngoc, Vu DinhPhai, Qi Shi, "A Deep Learning Approach to Network Intrusion Detection", IEEE Transactions on Emerging Topics In Computational Intelligence, Vol. 2, No. 1, February 2018,pp. 41-50.
3. MajdLatah ,LeventToker, "Towards an efficient anomaly-based intrusion detection for software-defined networks" IET Netw., 2018, Vol. 7 Iss. 6, pp. 453-459.
4. Majjed Al-Qatf , Yu Lasheng, Mohammed Al-Habib, And Kamal Al-Sabahi "Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection" IEEE. Translations and content mining, VOLUME 6, 2018, pp. 52843-52856.
5. CONGYUAN XU, JIZHONG SHEN , XIN DU, AND FAN ZHANG "An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units" IEEE Access,Volume: 6, 2018, Page(s)**:** 48697 – 48707.
6. Sheraz Naseer1,2, Yasir Saleem1, Shehzad Khalid3, Muhammad Khawar Bashir1,4, Jihun Han5, Muhammad MunwarIqbal, And Kijun Han" Enhanced Network Anomaly Detection Based on Deep Neural Networks" IEEE Transactions on Special Section On Cyber-Threats And Countermeasures In The Healthcare Sector Volume 6, 2018 pp.48231-48246.
7. Mohammed HasanAli ,Bahaa Abbas Dawood Al Mohammed , Alyani Ismail, And MohamadFadliZolkipli " A New Intrusion Detection System Based on Fast Learning Network and Particle Swarm Optimization" IEEE Transactions, Volume 6, 2018,pp. 20255-20261.
8. PeiyingTao ,Zhe Sun, And Zhixin Sun "An Improved Intrusion Detection Algorithm Based on GA and SVM" IEEE Transactions on Special Section On Human-Centered Smart Systems And Technologies, Volume 6,2018 pp. 13624-13631.
9. Qiangyi Zhang, YanpengQu, Ansheng Deng "Network Intrusion Detection Using Kernel-based Fuzzy-rough Feature Selection", IEEE International Conference on Fuzzy Systems,2018.
10. HuijunPeng , Chun Ying, Shuhua Tan , Bing Hu , And ZhixinSun,"An Improved Feature Selection Algorithm Based on Ant Colony Optimization", IEEE Transactions Volume 6, 2018, pp. 69203-69209.
11. Antonia Nisioti, Student Member, AlexiosMylonas , Paul D. Yoo,Senior Member, and VasiliosKatos "From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 20, NO. 4,2018, pp. 3369-3388.
12. Zheng Wang, "Deep Learning-Based Intrusion Detection With Adversaries", IEEE Transactions on Special Section On Challenges And Opportunities Of Big Data Against Cyber Crime, Volume 6, 2018,pp.38367-38384.
13. "GauravMeena, Ravi Raj Choudhary ", A review paper on IDS classification using KDD 99 and NSL KDD dataset in WEKA, IEEE International Conference on Computer, Communications and Electronics (Comptelix),2017, Page s: 553 - 558
14. Omar Y. Al-Jarrah, Omar Alhussein, Paul D. Yoo, Senior Member, IEEE, Sami Muhaidat, Senior Member, IEEE, Kamal Taha, Senior Member, IEEE, and Kwangjo Kim, Member, IEEE "Data Randomization and Cluster-Based Partitioning for Botnet Intrusion Detection" IEEE TRANSACTIONS ON CYBERNETICS, VOL. 46, NO. 8, AUGUST 2016 pp.1796 -1806.
15. G. Li, P. Niu, W. Zhang, and Y. Liu, ''Model NOx emissions by least squares support vector machine with tuning based on ameliorated teachingŰlearning-based optimization,'' ChemometricsIntell. Lab. Syst., vol. 126, pp. 11–20, Jul. 2013.

16. S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, ''Contractive auto-encoders: Explicit invariance during feature extraction,'' in Proc. 28th Int. Conf. Int. Conf. Mach. Learn. (ICML), 2011, pp. 833–840. [Online]. Available: http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Rifai_455.pdf
17. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, ''A detailed analysis of the KDD CUP 99 data set,'' in Proc. IEEE Symp. Comput.Intell.Secur. Defense Appl. (CISDA). Piscataway, NJ, USA: IEEE Press, 2009, pp. 53–58. [Online]. Available: http://dl.acm.org/citation.cfm?id=1736481.1736489
18. C.-F. Tsai, Y.-F.Hsu, C.-Y.Lin, and W.-Y. Lin, ''Intrusion detection by machine learning: A review,'' Expert Syst. Appl., vol. 36, no. 10, pp. 11994–12000, 2009.
19. Introduction to autoencoders. Jeremy Jordan, Data Science, 19 March 2018. https://www.jeremyjordan.me/autoencoders/

AUTORS PROFILE

Dhikhi T, is Research scholar in Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. She is currently Assistant Professor with SRM Institute of Science and Technology, Ramapuram, Chennai with total of 10 years' teaching experience. She received B.Tech degree in Computer Science from Cochin University of Science and Technology (CUSAT) and M.E in Computer Science and Engineering from Anna University. Her current research includes Network Security, Machine learning, Cryptography. She has published around 25 research papers in reputed journals and presented paper in 9 Conferences including national and international. She is a fellow of ISTE and IET.

Prof. M.S.Saravanan, Received B.Sc degree in computer science from Madras University in 1996, the MCA degree from Bharathidasan University in 2001, M.Tech degree from IASE University in 2005 and PhD degree in the Bharathiar University in March 2013. His current research interests include Data and Process Mining, Cloud Computing and Big Data Analytics. He is working as a professor in the Department of Computer Science and Engineering in Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India. He is a member of IEEE, IAENG, ICSA and reviewed more than twenty five journal publication including, IEEE Knowledge Engineering. He has received best researcher cash award of worth INR 1,08,000 from the VelTech University. He has published one hundred and two international publications and presented twenty four research papers in international and national conferences, having more than twenty years of teaching experience in various institutions in India and rest of the World.

Retrieval Number :H6932068819/19©BEIESP
DOI: 10.35940/ijitee.H6932.078919

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication