IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

A Semi-supervised Approach for Abnormal User Behaviour Detection in Network

Nandit Malviya[1*], Mukta S. Takalikar[2]

[1]Computer Engineering, Pune Institute of Computer Technology, Pune University, Pune, India
[2]Computer Engineering, Pune Institute of Computer Technology, Pune University, Pune, India

[*]*Corresponding Author: malviyanandit@gmail.com, Tel.: +91 8962293560*

Available online at: www.ijcseonline.org

Accepted: 15/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are generic. Detecting abnormal user behavior is of great significance for a secured network. The traditional detection method, which is based on machine learning, usually needs to accumulate a large amount of abnormal behaviour data from different times or even different network environments for training, so the data gathered is not in line with practical data and thus affects. There are many systems being developed which analyzes big data logs and recognizes patterns in it with already predefined classes using machine learning algorithm. The current research in this area implements algorithm like SVM (support vector machines), PCA (principal component analysis) mostly to classify data. Apart from this many are working to find different classes to classify anomalous activities. In this project, analysis of various machine learning algorithms will be carried out irrespective of user behaviour.

***Keywords*—** Anomaly Detection, Learning process, Machine Learning, Security.

I. INTRODUCTION

User Behaviour Analytics (UBA) technology analyzes historical data logarithm including network and authentication, logs are collected and stored in log management and SIEM system is used to identify normal records of traffic caused by user behaviors, both normal and malicious. UBA systems are primarily intended to provide cyber security teams with actionable insights. While UBA systems does not take action based on their findings, they can be configured to automatically adjust the difficulty of authenticating users who show anomalous behaviour.

Static perimeter defenses are no longer adequate in a world where data falling out increasingly are carried out using stolen user Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun credential. A new glide path called UBA, can eliminate this guesswork using big data and auto learning algorithms to assess the risk, in near-real time, of user activity.[1] UBA employs modeling to establish what normal behavior looks like. This user modeling incorporates information about: network user activities, request send, data received from network. This data is correlated and analyzed based on past and on-going activities. Network user behaviour analysis develop normal versus abnormal behavior profiles by collecting information on users' activities across IP addresses, accounts and devices. Unlike signature-based threat technologies, user behavior analytics creates a baseline for each individual user and then uses categorical, numerical and contextual information to identify anomalies and flag risky behavior.[2]

A. Intrusion Detection System

An IDS screens organization's action for suspicious activity. It may be comprised of gear, computer program, or a combination of the two. IDSes are comparable to firewalls, but are arranged to screen action that has entered or orchestrated, rather than foreseeing to get to an organization totally. This grants IDSes to recognize attacks that start from interior in an organization.

An intruder detection systems can be arranged for either an organization or a particular gadget. An arranged interruption location framework (NIDS) screens inbound and outbound activity, as well as information exchanged between frameworks inside a range. NIDSes are regularly spread out over a few distinctive focuses in an organization to create records that are not being monitored but detected as irregular activities. beyond any doubt there a no escape clauses where activity may be unmonitored.

B. Malware Detection

Malware has threatened computers, networks, and infrastructures since the eighties. There are two major

technologies to defend against this, but most organizations rely almost exclusively on just one approach, the decade's old signature-based methodology. The more advanced method of detecting malware via behavior analysis is gaining rapid traction, but is still largely unfamiliar.

Rest of the paper is organized as follows, Section I contains the introduction of user behavior based anomaly detection and types of network attacks , Section II contain the related work of abnormal user behavior detection, Section III contain the some measures of the process proposed to solve the challenges, Section IV contain the sytem overview, Section V contains architecture and the results obtained, section VI is all about concluding and future work that can be done.

II. RELATED WORK

Anomaly detection in network is the prime task for the monitoring tools available over internet. Each tools have there own method to solve the problem. One of the aspects of detection is User Behaviour(UB), which monitors over netflow data using features describing UB. When user behaviour comes into picture duration, srcbytes, dstbytes, logins attempted. Discovery is an dynamic security innovation for the irregular client conduct interruption which gives real-time security to caught and react to inner assaults as well as outside assaults when the organize framework is being jeopardized. Interruption is characterized as a collection of pernicious behaviors that endeavor to weaken the keenness, secrecy or accessibility of assets.

Meng Bi, Jian Xu, Mo Wang, Fucai Zhou [17] have analysed user behaviour with the principal component analysis algorithm. With the ability of PCA to extract features from whole dataset, PCA has an advantage over other since it has more efficient feature extraction methodology which increases the overall detection of the system.

You Lu*, Xutfeng Xi, Ze Hua, Hongjie Wu, NI Zhang [1] proposed a method by actualizing collaborative learning with semi-supervised learning where they supplanted cross approval with integration of part classifiers to diminish the overhead of naming. By this approach they have attempted to fathom two challenges: Require of labelled information, overhead of naming information.

Khurum Nazir Junejo, Jonathan Goh [2] proposed nine state-of-the-art ML classifiers that are quick and scalable, in spite of the fact that to some degree delicate to noise. Three of classifiers speak to discriminative classifiers, to be specific support vector machine (SVM), neural Network (NN), and instance-based learning (IBL). Three other classifiers are based on choice trees, to be specific irregular timberland (RF), J48, and best-first tree. The remaining three are measurable classifiers, specifically Naive Bayes (NB), Bayesian network (BayesNet), and polynomial logistic regression (LR). This all work for protecting framework as of now breached additionally classify which assault it was in spite of the fact that numerous interruption location framework are show which work over organize layer.

K. Hanumantha Rao, G. Srinivas, Ankam Damodhar and M. Vikas Krishna[4] have portrayed sorts of interruption Discovery Frameworks specifically network intrusion detection system(NIDS), Host-based Intrusion Detection Framework (HIDS), Protocol-based Intrusion Detection Framework (PIDS). Their proposed strategy comprises of two calculation working together i.e K-means and ID3 Choice trees. This strategy was basically planned for two challenges- abuse of location and peculiarity location.

Hamed Haddad Pajouh, GholamHossein Dastghaibyfard, Sattar Hashemi [5] have given a strategy where they have utilized naive bayes for to begin with arrange classification and for way better division between normal and anomalous activities KNN-classifier is utilized. They have utilized direct discriminant analysis (LDA) for highlight diminishment. Numerous assaults (like DoS, R2L, U2R) have been identified and produced wrong caution rate which was extra thing from past inquire about.

N. Pandeswari, Ganesh Kumar [6] have deployed their anomaly detection process over cloud. Their cloud environment have cloudsim 3.0 installation and anomaly detection process have naive bayes and ANN-classifiers for detection.the attack types are categories such as Denial of service, Probe, R2L, U2R.

Laskov et al. [8] put forward one-class SVM method for intrusion detection , which performed well in respect of false alarm rate; Tsang et al. [9] held up core vector machine CVM, which can finish fast training based on large data set; Khan et al. [10] combined SVM and hierarchical clustering.
Robert Mitchell and Ing-Ray Chen[11] proposed demonstrate to overcome interior aggressors that abuse the judgment of the medical cyber physical system (MCPS) with the objective to cripple the MCPS usefulness and whereas constraining the wrong caution likelihood to ensure the welfare of patients is of most extreme significance.

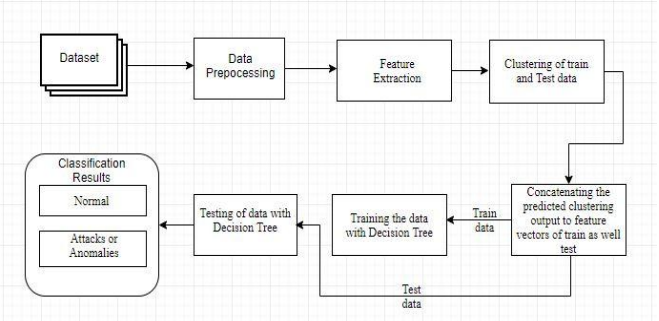

Figure 1: Methodology

Jaime Devesa, Igor Santos, Xabier Cantero, Yoseba K. Penya, Pablo G. Bringas [12] gave a detection method where they extracted features by defining their own set of rules in form of regular expressions and these expressions includes

behaviour as well family of malwares. For detection, training and testing ML algorithms naive Bayes, Rnadom forest (with the forest of 100), J48 (confidence of 0.25), SVM (Weka model tained SVM).

Shaohua Teng, Naiqi Wu, Haibin Zhu, Luyuo Teng [13] proposed and interative model of SVM model which classifies the data set into two categories the flow is normal and suspicious, on next iteration suspicious data into DOS/probe and R2L/U2R next classifier classifies DOS and probe, R2L and U2R. In the following the scheduling of classifier is one of the challenges for which they have given the scheduling policy.

III. METHODOLOGY

Figure 1 describes the proposed anomalous detection method. For detecting the anomalies in the network in form of attacks happening in network, we are using an unsupervised learning approach for determining the pseudo labels for partially labelled data and a supervised learning approach for decision making of normal or anomaly detected. The system will be analysed based on three scenarios described in section IV.

There are many modules in system and are described below.

A. MODULES

1. **Pre-processing**: Takes input one or more data log, performs pre-processing functions on them. Pre-processing functions are: - tokenizing, null values removal or initiating them etc.
2. **Clustering**: Form cluster from the feature vectors.
3. **Concating**: Concate the cluster output to feature vector.
4. **Classification**: Classifies the feature vector based on training.
5. **Calculate Accuracy**: Accuracy calculation is done by equating the predicted label with true labels.

The Pre-processing module is the most important one as feature extraction, feature encoding is being performed. For feature extraction we have calculate feature importance with the use of Random Forest with the importance of features we have extracted 14 features out of 42 which actually contribute in detecting the attacks.

B. ALGORITHM APPLIED

1. **K-means**: K-means perform the clustering over the features provided. It divides a set of N samples X into K disjoint clusters C, each described by the mean of the samples in the cluster. The K-means algorithm aims to choose clusters that minimize the inertia, or within-cluster sum of squared criterion:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_j - \mu_i||^2)$$

2. **Decision Tree Classifier**: It classifies the data after being trained. If a target is a classification outcome taking on values 0,1,…,K-1, for node m, representing a region Rm with Nm observations, let be the proportion of class k observations in node m.

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i = k)$$

C. DATA SET

The data set KDD 99' [19] will be used to train and test the detection system. As user behaviour is recognized by various attacks include in the data set. The KDD data set comprises of total 40 type of attacks where 21 attacks are common in both training and testing data set. In the data set among the features, duration of requests, source port, destination port, bytes send, bytes received determines the behaviour of user over network. The table 1,2 shows the insight of testing and training data set.

Table 1: Testing Data Set

	Original records	Distinct records
Attacks	250,436	29,378
Normal	60,591	47,911
Total	311027	77,289

Table 2: Training Data Set

	Original records	Distinct records
Attacks	3,925,650	262,178
Normal	972,781	812,814
Total	4,898,431	1,074,992

IV. SYSTEM OVERVIEW

The proposed system is shown in figure 1 which is designed in such a way that all challenges in work can be resolved. The main challenge is handling the unlabelled data for that we have used semi-supervised approach so that unlabelled data can be made of use. For the system conda environment is set up in linux operating system with required packages like numPy, sciPy, Sklearn, pandas.

A. DESIGN

There are two data set in KDD 99' which will be used for training and testing of the system. The pre-processing module helps in performing various transformation to get the input in desired form. Since it is the main module as transformation, encoding, feature extraction takes place in this phase which is utmost important for further process.

Then to the unlabelled data features clustering is performed which is described in section **III.** The results of clustering are appended to feature vector as new vector. These new transformed data is used for training and testing of system and evaluating accuracy.

B. REQUIREMENT

The abnormal user behaviour detection system which checks the anomalies based on user behaviour in network is built using python 2.7 with IDE Atom and conda environment. The minimum CPU requirements will be i5 processor, 500 GB ROM, 8 GB RAM.

V. RESULTS AND DISCUSSION

For our system we have created three scenarios, firstly normal and unique attacks label records from the train and test data, second all attacks labels and in last the labels other than 'normal' are encoded as 'attack'.

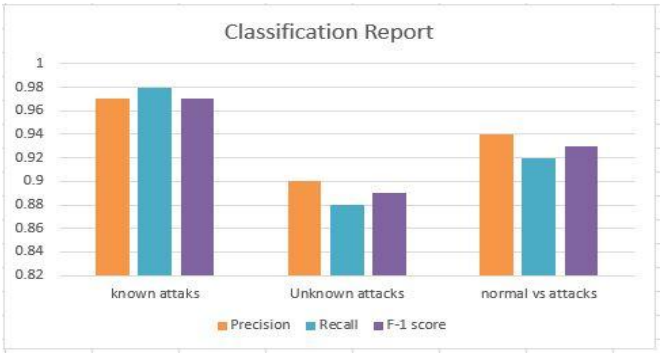

Figure 2: Performance Evaluation

Table 3: Classification Results

Scenarios	Known Attacks	Unknown Attacks	Normal vs Attacks
Accuracy	0.9768	0.8978	0.9246

As table 3 shows the classification results based on scenarios created.

A. RESULT ANALYSIS

Here the comparison of existing system with our proposed system is given:

Refernce	Methods	Result
[12]	KNN,CART,Random Forest	95.4%
[13]	ANN, Spark	94%
[14]	LSSVM, IDS	96.75%
	Our method	97.68%

Table 4: Comparison of existing work.

B. PERFORMANCE EVALUATION

The evaluation of the proposed system is done to check the current performance and improve it as per the metric suggest:

- **Precision**: It is described as correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = \frac{tp}{tp + fp}$$

- **Recall**: It is the ratio of correctly predicted positive observations to the all observations in actual class.

$$\text{Recall} = \frac{tp}{tp + fn}$$

- **F-1 Score** : The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

In figure 2 performance evaluation graph is depicted which gives an idea of area to be improved.

VI. CONCLUSION AND FUTURE SCOPE

Detecting abnormal user behaviour over a collected data from a different time and locations involves overhead as data is not structured. By traditional method it is difficult to detect anomaly efficiently. Hence with the proposed solution we can figure out the ways of detecting abnormal user behaviour some best suited some normal. Being able to detect the abnormal behaviour over network will enhance the security of system over the network. By use of a semi-supervised method we have tried to solve the problem of run-time detection and the efficiency of model is checked in different scenarios so we can say if any unsuitable circumstances occur our model will perform efficiently.

Current system is detecting the anomalies from the standard testing data set further with time batch over network monitoring can make the system work in real time which can be a positive aspect in securing any cyber security application (CSA).

ACKNOWLEDGMENT

I also sincerely convey my gratitude to my guide Prof. M.S. Takalikar, Department of Computer Engineering for her constant support, providing all the help, motivation and encouragement for the work. Above all I would like to thank my parents for their wonderful support and blessings, without which I would not have been able to accomplish my goal.

REFERENCES

[1] You Lu, Xuefeng Xi, Ze Hua, Hongjie Wu, Ni Zhang "An abnormal user behavior detection method based on partially labelled data" Computer Modelling New Technologies, pp.132-141,March 2014.

[2] Bi M, Xu J, Wang M, Zhou F. "Anomaly detection model of user behavior based on principal component analysis". Journal of Ambient Intelligence and Humanized Computing, pp.547-554, August 2016.

[3] Khurum Nazir Junejo, Jonathan Goh, "Behaviour-Based Attack Detection and Classification in Cyber Physical Systems Using Machine Learning", CPSS,ACM, 2016.
[4] Hanumantha Rao, G. Srinivas, Ankam Damodhar and M. Vikas Krishna "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms", International Journal of Computer Science and Telecommunications, Volume 2, Issue 3, June 2011.
[5] Pajouh HH, Dastghaibyfard G, Hashemi S. "Two-tier network anomaly detection model: a machine learning approach". Journal of Intelligent Information Systems, pp.61-74, Feb 2017.
[6] Pandeeswari N, Kumar G. "Anomaly detection system in cloud environment using fuzzy clustering based ANN". Mobile Networks and Applications, pp.494-505, Jun 2016 .
[7] Deepaa A J, Kavitha V "A Comprehensive Survey on Approaches to Intrusion Detection System", Procedia Engineering, pp.2063-9, 2012.
[8] Kloft M, Brefeld U, Duessel P, Gehl C, Laskov P. "Automatic feature selection for anomaly detection", Proceedings of the 1st ACM workshop on Workshop on AISec. ACM, 2008.
[9] Tsang IW, Kwok JT, Cheung PM. "Core vector machines: Fast SVM training on very large data sets". Journal of Machine Learning Research, pp.363-9, 2005.
[10] Khan L, Award M, Thuraisingham B "A new intrusion detection system using support vector machines and hierarchical clustering". VLDB Journal, pp.507-21 2007.
[11] Mitchell, R. and Chen, R., "Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems", IEEE Transactions on Dependable and Secure Computing, pp.16-30, 2015.
[12] Jaime Devesa, Igor Santos, Xabier Cantero, Yoseba K. Penya and Pablo G. Bringas "Automatic behaviour-based Analysis and Classification System for Malware Detection",Deusto Technological Foundation, Bilbao, Spain 2010.
[13] Teng, Shaohua, Naiqi Wu, Haibin Zhu, Luyao Teng, and Wei Zhang. "SVM-DT-based adaptive and collaborative intrusion detection", IEEE/CAA Journal of Automatica Sinica, pp.108-118, 2018.
[14] Yao, H., Y. Liu, and C. Fang, "An Abnormal Network Traffic Detection Algorithm Based on Big Data Analysis".International Journal of Computers,Communications Control, 2016.
[15] Hsieh, C.-J. and T. Y. Chan. "Detection DDoS attacks based on neural network using Apache Spark,International Conference in Applied System Innovation, 2016.
[16] Ambusaidi MA, He X, Nanda P, Tan Z., "Building an intrusion detection system using a filter-based feature selection algorithm". IEEE transactions on computers, pp.2986-98, 2016.
[17] Meng Jiang and Peng Cui, Christos Faloutsos, "Suspicious Behavior Detection: Current Trends and Future Directions", IEEE Computer Society, January/February 2016.
[18] Thomas Dietterich, Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns, "Adaptive Computation and Machine Learning" MIT press, 2011.
[19] Stephen D. Bay and Dennis F. Kibler and Michael J. Pazzani and Padhraic Smyth, "The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation", SIGKDD Explorations, 2000.

