IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

A survey on Detecting Network Intrusions Using Machine Learning

K. Haritha[1*], CH. Mallikarjuna Rao[2]

Department of Computer Science and Engineering,
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

DOI: https://doi.org/10.26438/ijcse/v7i5.11011105 | Available online at: www.ijcseonline.org

Accepted: 19/May/2019, Published: 31/May/2019

Abstract- Intrusion Detection (ID) is a basic part of security, for example, versatile security machines. Earlier different ID procedures are utilized; however their execution is an issue. ID execution relies upon precision, which needs to enhance to diminish false alarms and to expand the detection rate. To determine concerns on execution, multi-layer network, SVM, Naïve Bayes and different procedures have been utilized in later. Such procedures demonstrate restrictions and are not proficient for use in huge data, for example, complex and system data. The ID framework is utilized in breaking down immense traffic data; thus, a proficient classification method is important to beat the issue. This issue is considered in this paper. Popular data mining and machine learning methods are used. They are SVM, and Random forest and KNN, Decision Tree, Extreme Learning Machine (ELM). These methods are outstanding a direct result of their capacity in classification. NSL_KDD dataset is used.

Keywords: ID, Anomaly Detection, False Alarms, NSL_KDD dataset, Ensemble Approaches.

I. Introduction

Intrusion is a serious issue in security and a prime issue of security rupture, on the grounds that a solitary case of intrusion can take or erase data from Network systems and PC in almost no time. Intrusion can likewise harm system hardware. Furthermore, intrusion can cause immense harms monetarily and consensus the IT basic foundation, accordingly prompting data mediocrity in digital war. In this manner, ID is crucial and its prevention is essential. Different ID methods are accessible, yet their efficiency remains an issue; it relies upon rate of detection and false rate. The issue on exactness should be routed to decrease the false alarms rate and to increase the detection rate. This idea was the stimulus forward is explore work. Consequently, SVM, and Random Forest, KNN, Decision Tree, ELM are observed in this work; these strategies have been demonstrated successful in their capacity to address the classification issue.

This work utilized the NSL-KDD dataset and KDD dataset which is an enhanced type of the KDD and is viewed as a benchmark in the assessment of ID methods [12]. Securing PC and system data is essential for associations and people in light of the fact that bargained data can cause impressive harm. To maintain a strategic distance from such conditions, Intrusion Detection systems are vital.

II. Survey on detecting Network Intrusions

Wang *et al.* [1] suggested a framework for ID system using Support Vector Machine and verified their approach with the dataset i.e., NSL_KDD. They declared that the approach which they used has an effectiveness of 99.92 percentage, which was more than the other approaches. The efficiency of Support Vector Machine declines when it involves huge data, and it's not good to prefer, in evaluating enormous network traffic for detecting the intrusions.

S.Duque and Omar *et al.* [2] applied K-Means algorithm on NSL-KDD and tested on various five-clusters. The best outcomes are acquired when 22-clusters were used. Also K-Mean is utilized in hybrid approaches.

B.Sharma and H.Gupta *et al.*[3] used K-Mean in hybrid methods that utilizes two strategies, association rule and clustering. Apriori and K-Mean algorithms are utilized to identify the intrusions. The execution measures are execution time-120ms, CPU usage-74% and also memory use-54%.

[4] and [5] applied hybrid method for clustering and classification. Ravale & Nilesh *et al.* [4] applied a hybrid method of RBF kernel function of SVM and K-Means algorithm. The efficient outcome of the hybrid approach is 93 percent and the rate of detection is 95 percent. Chao and Wen *et al.* [5] applied a hybrid method of KNN and K-Means. Then the accuracy outcome is improved that is, 99 percent. KDD 99 dataset is used in both the hybrid methods.

Magld & Hundewale *et al*. [6] applied K-Means approach and some other data mining techniques and concluded that K-Mean is better than any other approach. They didn't mention the dataset name .
Nannan and Liang *et al*. [7] used an approach, which is collection of Fuzzy C Mean (FCM) and K-means techniques to remove or reduce the false positives from DARPA 2000 dataset. Conclusion of this task is, the effect of FCM algorithm is more than K-Means algorithm.
Zhengjie and Yongzhong *et al.* [8] used hybrid technique of K-Means & particle Swarm Optimization method. The rate of detection of unknown attacks is 60.8 percent and the known attacks is 75 percent.

In order to increase the efficiency of Support Vector Machine, Horng & Yang *et al*. [9] used the hierarchical clustering along with the hybrid Support Vector Machine. For Feature selection procedure to remove unnecessary attributes from the dataset, the BRICH hierarchical clustering technique is used. Then the efficiency in classification of SVM is increased. The rate of efficiency of proposed framework is 95.7 percentage and rate of false positive is 0.7 percentage. Shivshankar E *et al.* [11] suggested a framework for detecting the SQL injection attacks using K-Nearest Neighbour algorithm.

Chan and Yang *et al.* [13] used SVM and Random Forest for identifying the intrusion in the network. They suggested that only 14 attributes out of 41 are most important to detect the attacks correctly and thus rate of detection increases with this.
Huang and Zhu *et al.* [16] suggested that ELM performs best generalization and classification for complex data.

Table I: Different types of existing mechanisms

Classifier	Method	Parameters	Advantages	Disadvantages
K-NN	Assigned to the class of its nearest neighbor.	The number k of nearest neighbor and the feature space transformation.	1. Scientifically tractable. 2. Simple usage. 3. Itself easily to parallel usage.	1. More storage required. 3.Slow in classify the test tuples.
ANN	It changes its format based on in and out data.	It is from a low optionability.	1.Difficult non-linear links with related variables	1. computational Burden. 2.More time required.
Bayesian technique	Based probabilities of sample observations and classes.	Parameters approximated the input set.	1.Bayesian streamline classifier 2.Good accuracy for large data-set.	1 The suppositions made in class restrictive autonomy

III. Existing Data Processing Models of IDS

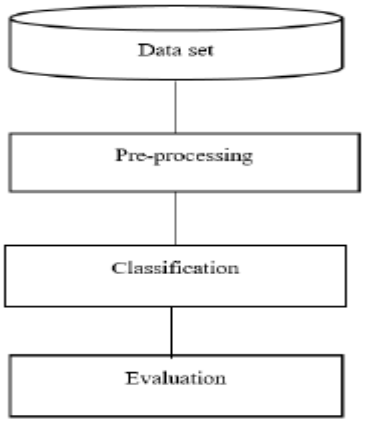

Fig1: Model flow of ID system

Case Study:
Writing audit speaks to that numerous analysts completed investigate on methodologies of DM to recognize the intrusions and each methodology has distinctive precision, false alarm rate and detection-rate.

Preprocessing:
Because of the contrast between the organizations data, it is important to pre-process it like to change the char data into number data.

K-Means Clustering
K-Mean [2,3,4], algorithm is a strategy that groups all the identical data , depending on their conduct. K-Mean is an unsupervised errand that is, the data does not determine what we attempt to learn. Numerous specialists use K-Mean in the half and half ways to deal with the peculiar data.

Algorithmic Procedure

Step1: Select quantity of centroids of dataset as the beginning centroids.

Step2: After that, determine Euclidean separation between every datum point and the centroids.

Step3: If the datum point is nearest to the centroid, at that point abandon it and don't do any modifications in its position. However, on the off chance that the information point isn't nearest to the centroid, at that point moves it to its nearest one.

Step4: Regenerate the centroid of both changed clusters.

Step5: Repeat stage 3 if not.

$$M = \sum_{a=1}^{k} \quad \sum_{b=1}^{n} \quad d_{ab}\ (x_b, y_a)$$

Where, $d_{ab}(x_b,y_b)$ is an Euclidean measure between the data of x_b and y_a the centroid

$$d(x_b, y_a) = \| x_b - y_a \|$$

IV. Types of Attacks in IDS

1. Dos Attacks

A Dos (Denial of Service) attack is a kind of an attack, in which an intruder makes a memory resource excessively occupied or too full [15]. For example smurf, Neptune, so forth are on the whole DoS attack.

2. R2L Attacks

Remote to Local attack is a kind of an attack in which a client sends bundles to a machine over the network, which he/she doesn't have permission to reveal computer susceptibilities and makes use of those benefits which a nearby client would have on the PC [15]. For example xlock, visitor, xnsnoop, phf, sendmail word reference and so on.

3. U2R Attacks

These attacks are misuses in which the programmer begins off on the system as an account of an ordinary client and endeavors to mishandle things in the system so as to increase super client benefits .For example perl, xterm.

4. Probing Attacks

This is an attack in which the attacker checks a machine or a system's administration gadget so as to decide shortcomings that may later be misused in order to bargain the system. This strategy is usually utilized in data mining for example portsweep, saint, mscan, nmapetc.

V. Dataset

Statistical examinations on KDD CUP 99, demonstrated that this dataset has shortcomings that impact on systems' execution. Subsequent to researching and investigating this data, it was realized that 75 percent of testing data and 78 percent of the training data are frequently appearing. So NSL_KDD dataset is used [10]. There are total of 41 features, which incorporate formal, numeric, binary attributes.

Table II: Represents different types , their features and numbers

Type	Features with their Numbers
Nominal	protocol_type(2),service(3),flag(4)
Binary	land(7),logged_in(12),root_shell(14),su_attempted(15),is_host_login(21),is_guest_login(222)
Numeric	duration(1),src_bytes(5),dst_bytes(6),wrong_fragment_urgent(9),hot(10),num_failed_logins(11),num_compromised(13),num_root(16),num_file_creations(17),num_shells(18),num_access_files(19),num_outbound_cmds(20),count(23),srv_count(24),serror_rate(25),srv_serror_rate(26),rerror_rate(27),srv_rerror_rate(28),same_srv_rate(29),diff_srv_rate(30),srv_diff_host_rate(31),dst_host_count(32),dst_host_srv_count(33),dst_host_same_srv_rate(34),dst_host_diff_srv_rate(35),dst_host_same_src_port_rate(36),dst_host_srv_diff_host_rate(37),dst_host_serror_rate(38),dst_host_srv_serror_rate(39),dst_host_rerror_rate(40),gst_host_srv_rerror_rate(41).

VI. Survey Evaluation parameters:

This examination utilizes some evaluation measures, for example, precision, detection rate, and false alarm rate as assessment variables, and these are calculated based on the confusion matrix in table III [14].

Table III: Confusion matrix

Predicted value→ Actual value↓	Normal	Attacks
Normal	TN	FP
Attacks	FN	TP

Performance measures are:

Accuracy, Detection Rate, False Alarm Rate

True Positives(TP): Correctly identified count.

True Negatives(TN): Safe application Correctly identified as safe.

False Postive(FP): Safe applications falsely identified.

False Negative(FN): Count of Falsely identified as usual.

Survey results analysis of Existing models:

The trial outcomes are assessed from the suggested structure in figure1 on NSL_KDD dataset. Each of the record has 41 features. Among all the clusters, 22nd cluster has shown the highest accuracy. Table 1V shows that the exactness of the proposed framework test information is shown underneath.

Table IV: KDD cup-99 Dataset

Protocol	Flag	Dst_bytes	count	Srv_count	Dst_host_count	Serror_rate	Attacks
Udp	SF	146	1	1	255	0	R2l
Udp	SF	146	2	2	255	0	Dos
Udp	S3	146	12	4	187	0	Dos
Tcp	S2	146	22	12	196	0	U2r
Tcp	SF	0	5	21	71	0	Normal
Tcp	S0	185	2	13	3	0	Normal
Icmp	REJ	185	3	20	54	0	Prob
Icmp	SF	260	21	11	174	0	Prob
Udp	SF	146	15	15	255	0	R2l
Tcp	S3	329	2	23	255	0	R2l
Udp	S2	923	22	1	177	0	Dos
Icmp	S0	137	13	4	196	0	U2r
Icmp	RSTU	735	2	12	54	0	Normal
Udp	RSTU	260	1	2	255	0	Normal
Tcp	SF	185	3	13	255	0	Normal

VII. Conclusion

Because of our dependence on online, thus developing number of intrusions cases are increasing, so building successful IDS are fundamental for securing Internet assets but then it is an incredible test. In written works, numerous analysts used supervised learning techniques SVM, K-NN. But for the large data these techniques could not perform well. So we propose ELM technique for ID effectively. Here, based on training data with the label dataset, ELM maps the system traffic into pre-defined class's i.e. normal or explicit attack type.

References

[1] H.Wang,J.Gu,andS.Wang,''An effective intrusion detection framework based on SVM with feature augmentation,'' Knowl.-Based Syst., vol. 136, pp. 130–139, Nov. 2017, doi: 10.1016/j.knosys.2017.09.014.

[2] S. Duque, N.B Omar, "Using data Mining Algorithms for Developing a Model for ID System (IDS)", Proceedings of

Science direct: Procedia Computer Science 61, pp. (46-51), 2015.
[3] B. Sharma and H. Gupta, "A design and Implementation of ID System by using Data Mining", IEEE Fourth International Conference on Communication Systems and Network Technologies, pp.700-704, 2015.
[4] U. Ravale, M. marathe, P. Padiya, "Feature Selection based Hybrid Anomoly ID System using K Means and RBF Kernal Function", Proceedings of Science Direct: International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 428-435, 2015.
[5] W. C. Lin, S. W. Ke, C. F. Tsai, "CANN: An ID system based on combining cluster centers and nearest neighbors", Proceedings of Science direct: Knowledge-Based Systems, pp. 13-21, 2015.
[6] J. Haque, K.W. Magld, N. Hundewale, "An Intelligent Approach for ID based on Data Mining Techniques", Proceedings of IEEE, 2012.
[7] Liang Hu, Taihui Li, NannanXie, Jiejunhu, "False Positive Elemination in ID based on Clustering", IEEE International Conference on Funny System and Knowledge Discovery (FSKD), pp. 519-523, 2015.
[8] Zhengjie Li, Yongzhong Li, Lei Xu, "Anomoly ID Method based on K-Means Clustering Algorithm with Particle Swarm Optimization", IEEE International Conference of Information Technology, Computer Engineering and Management Sciences, pp. 157- 161, 2011.
[9] S. J. Horng, M.Y. Su, Y. H. Chen, T. W. Kao, R. J. Chen, J. L. Lai, C. D. Perkasa, "A novel ID system based on hierarchical clustering and support vector machines", Proceedings of Science direct: Expert Systems with Applications, pp. 306-313, 2011.
[10] http://nsl.cs.unb.ca/NSL-KDD/
[11] K. Shivshankar E., "Combination of Data Mining Techniques for ID System", IEEE International Conference on Computer, Communication and Control (IC4-2015).
[12] L. Dhanabal, S.P. Shantharajah, "A study of NSL-KDD Dataset for ID System based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue 6, pp. (446-452), June 2015.
[13] Chang, Y., Li, W., & Yang, Z. (2017). Network Intrusion Detection Based on Random Forest and Support Vector Machine. 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). doi:10.1109/cse-euc.2017.118
[14] Basic Evaluation Measures From the Confusion Matrix. Accessed: May 20, 2018. [Online]. Available: http://WordPress.com and https:// classeval.wordpress.com
[15] Swati Paliwal and Ravindra Gupta. Article: Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm. International Journalof Computer Applications 60(19):57-62, December 2012.
[16] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, ''Extreme learning machine: A new learning scheme of feedforward neural networks,'' in Proc. IEEE Int. Joint Conf. Neural Netw., vol. 2, Jul. 2004, pp. 985–990, doi: 10.1109/IJCNN.2004.1380068.

