IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Smart Intrusion Detection Using Machine Learning Techniques

Ashish Puri[1]*, Md Tabrez Nafis[2]

[1,2]Dept. of Computer Science and Technology, Jamia Hamdard (Hamdard University), N.Delhi, India

*Corresponding Author: ashishjsn9@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.483488 | Available online at: www.ijcseonline.org

Accepted: 14/Apr/2019, Published: 30/Apr/2019

Abstract **-** Intrusion Detection is one of the most effective and widely used implementation against the attacks and threats. Further more attackers keeps on varying their attacking techniques and tools .In this paper we have tried to perform a simulation study to evaluate the performance of varied machine learning classifiers to detect intrusion detection based on KDD 99 cups data set [1] focusing on enhancing the proficiency of Intrusion Detection system (IDS).

Keywords— DoS- Denial Of Service; U2R-User to root; R2L: Root to local; CIA-Confidentilaity,Integrity,availability ; CM-Confusion Matrix; MLP-Multi Layer perceptron ; NEA-Nearest Cluster Algorithm.; GAU:- Gaussian; K-M: K means algorithm; CPE: - Cost Per example; Pd:Probabilty Detection; FaR:False alrm Rate;

I. INTRODUCTION

Internet has become the driving force of our day to day life with tremendous progression of information technology in the last 2 decennary. We are using the internet and computers networks in every sphere of life from social networking, business, and entertainment to education, health etc. –which leaves us more vulnerable to different types of attack, thus Network security has become major challenge for researchers. To ensure the security and prevention from major classes' attacks [2] like–DOS, U2R, R2L, Probe – to handle these threats we need a powerful & intelligent Intrusion detection system (IDS).

There are many type of attacks threatening the CIA triad[3] i.e. Confidentiality, Integrity and Availability of information and cyber space. The Dos –Denial of Service is one of the most commonly encountered attacks. The Denial of services (DoS) momentarily denies or blocks the end user services. DoS consume and overload the network resources in general. In present scenarios web applications and social networking website are the prime focus of DoS . R2L –Remote to local is another class of attack which provides local right permission of network resources which should be exclusive for local users. The examples for R2L are SPY and PHP – whose main aim is to gain unauthorized access to network resources.

U2R: User-to-root attack is related to network and computer resources- it switches the assailant access authorization from normal user to the root user having full access rights to the computer and network resources. As attackers keeps on adapting new attacking methodologies in exploiting different kind of vulnerabilities. Hence it is very difficult to detect all types of attacks via single solutions. For this intrusion detection system becomes the essential part of the network security. It is implemented to monitor network traffic and generates alerts in case of any attack. IDS can be used to monitor specific device (host IDS) also and or monitors all the network traffic (Network IDS).

In general, two main classes of Intrusion detection system [4] – (1) Anomaly based Intrusion (2) Misuse based intrusion detection systems. Anomaly based IDS are enforced to detect attacks on the basis of recorded normal behavior. It is stands on statistical evaluation. It spots attacks based on abnormalities in the pattern w.r.t to regular pattern, Advantages-Anomaly IDS is capable of detecting new unknown threats from communicational assets. Disadvantages- Time needed in training dataset is more and increased problem of false positive alerts persists.

Misused or Signature based Intrusion detection system illustrates the attacks in the form of signatures. A data base of these patterns and signatures is maintained for comparing the data received on the network to identify the intrusion occurred. Advantage- Signature based IDS produces very low false positives. They are easy to develop and requires less communicational resources ,Disadvantages- Detects only those threats that are present in the database and updating of database is a time consuming process Literature survey depicts that, for IDs majority of researchers applied only one algorithm to detect diverse attack categories. The set of machine algorithm applied in the literatures comprises of small subset which potentially capable for intrusion detection. On the basis of comprehensive analysis conducted

shows that there is considerable divergence from one attack category to another. In this paper we have tried & tested different algorithms and to identify the most suitable algorithm capable & effective in detecting particular attack category is our main area of concern for this smart intrusion detection handler have been proposed which incorporates 3 main algorithms: Multilayer Perceptron algorithm , K-means algorithm and Gaussian algorithm, the KDD 99 data-set is passed through the SIDH (tool) to test and obtain the most accurate & promising result of intrusion occurred. The main objective of the research is to specify whether the packet is normal/regular packet or irregular attack packet. The Smart Intrusion Detection handling tool is the proposed solution for the above said challenge.

The structure of the paper is as follows: Section II Related Work, III Methodology, Section IV Results & Discussion section V Conclusion and Future scope.

II. RELATED WORK

.The Kernel Minor tool was used by Levin and the KDD data-set [5] was utilized to create a set of locally alleviated decision trees to select optimal subset of trees for forecasting up to the minute cases, in the study only 15% of KDD training dataset was used by Levin from entire training dataset. To detect different attack categories, multiclass detection was used on the KDD data set. A substantial intrusion detection rate for varied classes was achieved. The detail of the detection observed is as follows:

Table 1

Sno	Attack Type	Detection Rate in %
01	Probing	84.5 %
02	Denial of Service	97.2 %
03	U2R	12.1 %
04	R2L	7.25

And False Alarm rate (FAR) observed in different attack classes is as follows:

Table 2

Sno	Attack Type	Detection Rate in %
01	Probing	21.1 %
02	Denial of Service	72.3 %
03	U2R	35.6 %
04	R2L	1.8%

Steinback and Ertoz utilized Shared nearest neighbor technique which was suitable for locating data clusters of varied sizes, shapes and density. Here data mainly contains substantial amount of data outlier and noise. All the attack caseloads were picked from KDD testing and training datasets with capping of 10000 records for each attack category. There were almost 36 attack categories, at random 10000 records were selected from testing data sets and training datasets.95000 records were selected from KDD data sets. After duplicate KDD records were removed the size of the data was downsized to 44000 records. This set was then employed in order to train 2 clustering algorithm viz.

- K-means algorithm [8]
- C4.5

K-means Algorithm – with 300 clusters able to detect different intrusions as mentioned below: -

Table 3

SNo.	Attack Type	Detection Rate in %
01	Probing	91.25 %
02	Denial of Service	96.25 %
03	U2R	5.21 %
04	R2L	77.04 %

While the SNN technique performance in detecting the attack categories is as follows:

Table 4

SNo.	Attack Type	Detection Rate in %
01	Probing	74.25 %
02	Denial of Service	78.25 %
03	U2R	37.8 %
04	R2L	69.15 %

The above records show that SNN technique scores fairly well in detecting U2R, while nothing was reported by the author. A point to be noted that this model determines whether a record is intrusive or not if the record pertains to a particular class.

Similarly Chow and Yeung used non parametric density estimation approach in order to construct Id system using regular data only. It is based on parzen window estimator. This methodology was utilized for detecting attack category in KDD data set, in this they have tested randomly regular records from KDD training data set in order to determine the density of the said model. This technique could spot various intrusions and the results obtained are tabulated below:

Table 5

SNo.	Attack Type	Detection Rate in %
01	Probing	99.11 %
02	Denial of Service	96.25 %
03	U2R	93.8 %
04	R2L	31.21 %

According to the literature survey, it is correct to state that most of the researcher have employed single algorithm in order to detect multiple attack category and on the basis of the comprehensive study carried out it can be stated that there is substantial variations from one attack category to another.

In this paper we have tested different algorithms to obtain the most effective algorithm as per the attack category or

intrusion. For this a tool have been proposed in which three main algorithm were housed viz.

- Multi perceptron Algorithm.[10]
- Nearest Cluster (NEA) Algorithm.[11]
- Gaussian Algorithm.[12]

This tool is named as SIDH-Smart Intrusion Detection handling tool .The KDD 99 datasets after filtration and proper classification of Network data were passed through this tool to test and obtain the most accurate result of intrusion occurrence.

III. METHODOLOGY

A. Data Classification & pre-processing of KDD 99 Data

Relevant The KDD 99 data set consists of varied forms of discrete, continuous, and symbolic attributes with variable resolution and range and is very tedious to process such data hence pre-processing is required before pattern classification .The Pre-processing of data set requires two steps:

1. Mapping of symbolic value attribute to numeric value attribute.
2. Scaling of Data-set.

In mapping of scaling attribute the attack names like buffer-overflow, guess Password were mapped to different class-types as depicted in the table below:

Table 6

SNo.	Attack Class	Classes Assigned
01	Normal Data	0
02	Probing	1
03	Denial of Service (Dos)	2
04	U2R	3
05	R2L	4

Symbolic feature were described as follows:

SNo.	Symbolic Feature	Number Symbols Assigned
01	Protocol Type	03
02	services	70
03	Flags	11

The above mentioned symbolic features were plotted to integer values in the range from 0 to N-1, were N is the number of symbols. After this these features were linearly scaled as mentioned below:

Table 7

SNo.	Feature Name	Range assigned
01	Duration	[0,60000]
02	Wrong Fragment	[0,3]
03	Urgent	[0,1]
04	Number of Failed Logins	[0,5]
05	Non Compromised	[0,9]
06	Su-attempted	[0,2]
07	Num-root	[0,7500]
08	Num_file_creation	[0,100]

Large number of duplicate data set are present in KDD 99 data and needed to be removed for training purposes via different classifier models.

The number of datasets present in the original data set is as follows:

SNo.	Attack Class	Quantity
01	Normal Data	972,780
02	Probing	41,102
03	Denial of Service (Dos)	388370
04	U2R	52
05	R2L	1126

After filtration of duplicates records there are total of 812,813 records. The details of records are as follows:

SNo.	Attack Class	Quantity
01	Normal Data	812,813
02	Probing	13,860
03	Denial of Service (Dos)	247,267
04	U2R	52
05	R2L	1126

For pattern recognition linkNet is the tool which is freely available for simulation purposes for recognition of patterns & machine learning models. All the simulation was performed on Ubuntu platform Intel core i5-3230 2.6 Ghz, 8th Generation processor and 8 Gb RAM.

IV. RESULTS AND DISCUSSION

A. Performance comparison.

In order to obtain the optimal settings for the topology and parameters through the empirical means multi instances were built and tested on KDD data sets. For this purpose few models were developed by utilising multi-layer perceptron algorithm alone .A comparative study was carried out for shortlisting the elite model for the provided classifier algorithm. One of the measures is cost per example that requires two quantities:

- Confusion Matrix (CM) [13]
- Cost Matrix (C)[14]

Cost Matrix (C): Here the associated classes are labelled in rows and the current context of KDD datasets instances are placed in columns, there are 5 categories: {Normal, Probe, DoS, U2R, R2L}. Thus matrix of 5x5 dimension was developed .An Entries at column i and column j represent the non-negative cost of misclassifying paradigm. These figures were assigned for assessing result of KDD 99 dataset. The Magnitude of values thus obtained is directly proportional to

the impact of the computing platform or resources under the attack if test record lands into the category.

Confusion Matrix (CM): Similar to cost matrix, in confusion matrix 5x 5 matrixes was taken into account. The entries in row i and column j represent the misclassified patterns & attack categories respectively. By obtaining the values of cost matrix (C) and the Confusion matrix –to calculate the cost per example following formula was used:

$$CPE = \frac{1}{N}\sum_{i=1}^{5}\sum_{j=1}^{5} CM(i,j) * C(i,j)$$

CM- Confusion Matrix;

C- Cost Matrix;

N- Is the number of pattern tested

Lesser the value of cost per example superior the classifier model .The performance comparison was done by using detection of probability in conjunction with False Alarm Rate (FAR) and is considered to be widely accepted measure.

B. Deployment of Pattern Recognition and machine algorithm.

As mentioned earlier since Pattern recognition and machine algorithm techniques were applied for intrusion detection. Testing of six distinct pattern recognition and machine learning algorithm was carried out on KDD 99 dataset .These algorithm represents varied fields :- neural networks ,decision trees, and statistical model. A brief outline of how particular instance of the above mentioned algorithms in order to define their performance of intrusion detection is described as follows

1) Multilayer perceptron (MLP)

MLP is the most commonly used neural network classification algorithm .For simulation purpose the architecture of MLP with KDD dataset is set into 3 layered format structures:

- Input layer
- Hidden layer
- Output layer

In this simulation on each neuron an unipolar sigmoid transfer function was used in hidden layer as well as output layer with slope value of 1.0. Here the contingent gradient function along with mean square error function was utilized learning algorithm. In the output layer there are 41 neurons and in the input layer there are 5 neurons Several simulation were conducted with number of concealed layer nodes ranging from 45 to 85 with an increment of 10.A constant learning rate was maintained for each simulation viz. 0.1,0.2,0.3,0.4 ,these 4 values have 0.6 as the weight change momentum value. Learning rates varies with different simulations. The final model was able to attain 0.2344 as the cost per example value.

2) Gaussian classifier (GAU)

In Gaussian classifier inputs were assumed to be uncorrelated and distribution of classes varies in their mean values only. The Gaussian classifier is predicted on Bayes Decision theorem. There are 4 distinct models which were developed using Gaussian classifier were as follows:

- Diagonal covariance matrix which is Quadratic Classifier type.
- Tilted covariance matrix which is Quadratic Classifier type.
- Diagonal covariance matrix which is a type of Linear Classifier.
- Tilted covariance matrix is also class of linear classifier.

In our case study the cost per example value obtained is 0.3622.

3) K means Clustering Algorithm (K-M)

In K-means clustering algorithm generation of diverse clusters were obtained which were specified for every single output class. The simulations were run various clusters .IN every single simulation there were equivalent numbers of clusters for every single attack class. Let us assume the number of cluster is k which is non-integer power of 2, after generation of P clusters where P is greater than k the cluster centre have minimal variance with in its pattern. This variance was eliminated one by one till the value of clusters was shortened to k. An Epoch representing total training pattern in a resulting class was generated. Training of clusters were performed till the average square error difference was reduced to 1% or less .To form the new clusters the centres disturbed +/-1 % of standard deviation in each clusters. A randomized offset of plus/minus 1% was appended during every single split .This model obtained the lowest cost per example value of 0.2379.

4) Nearest Cluster Algorithm (NEA)

Nearest Cluster Algorithm is considered as the abridged version of K nearest neighbour clustering algorithm .A group of cluster centre was created from the training dataset which

is provided as input to nearest clustering algorithm which can be used like K-means ,leader algorithm .Multiple simulation were carried out by using primary clusters on each resulting class. .The value of cost per example for KDDs testing data came near around 0.2466.

5) Decision tree (C4.5)

This algorithm was devised by Quinlan which is based on information theoretic methodology. The main purpose of this algorithm is to build a decision tree with minimal number of nodes that leads to trivial number of misclassification on training data set. The C 4.5 algorithms adopted divide and rule scheme. The primary phase was set to 20% of records present in KDD data set and this 20% of record was appended after every single iteration and the retraining of tree was done. In all the test conducted the last two branches should contains at least 2 records. The cost per example for the elite decision tree classifier model came to be 0.2396.

C. Performance rating of Algorithm on KDD testing data set

Out of five classifiers developed through KDD training data set. Best performance instances were evaluated on KDD training data set. The Probability Detection(Pd) and False alarm Rate (FaR) performance was recorded for each classifier. The simulation result so obtained is presented in Table 1. For each classifier algorithm each attack category with Probability of Detection (PD) and FAR values are stated. The Table 1 indicates that the single algorithm is incapable to spot all attack categories with high contingency Detection rate and low False alarm Rate. The result so obtained also reflects that certain algorithm proves to have better performance in comparison to others .The K-M,GAU,NEA identify more than 87 % of attack records for probing .The result of detection of different algorithm is as follows:

Attack Class	Classifier	Detection rate(in%)
Probing	K-M,GAU,NEA	85%
Denial of Service (Dos)	MLP,NEA,K-M	97%
U2R	GAU,K-M	22%
R2L	GAU	10%

After identifying the algorithm as per the attack category, the best algorithm can be selected to obtain the most promising end result and in these FaR values is also being taken into account. As per the study conducted the obtained result is tabulated below:

The Pd and FaR values for each algorithm is tabulated below:

TABLE-1

Classifier Algorithm		PROBE	DoS	U2R	R2L
GAU	Pd	0.90	0.831	0.230	0.098
	FaR	0.110	0.008	0.006	0.002
K-M	Pd	0.880	0.970	0.291	0.06
	FaR	0.27	0.035	0.035	0.002
NEA	Pd	0.777	0.969	0.021	0.030
	FaR	0.45	0.025	0.065	0.016
MLP	Pd	0.880	0.952	0.505	0.051
	FaR	0.005	0.025	0.291	0.014
C4.5	Pd	0.888	0.980	0.017	0.04
	FaR	0.006	0.003	0.00002	0.00005

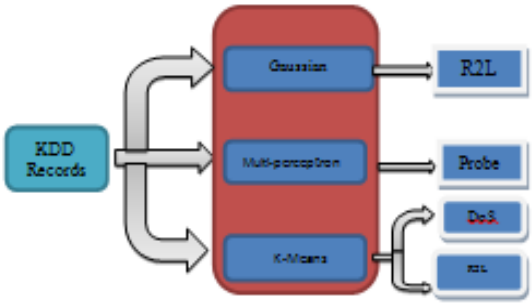
It is logical to mention that machine learning algorithm and pattern recognition algorithm verified on KDD data sets. Showed an fair degree of exploitation of performance in detecting only 2 attack classes specifically Probing and DoS .On the contrary, all five algorithm failed to provide an acceptable level of detection performance for remaining two categories ie U2R and R2L. Subsequent to the observation as per the given attack categories few subset of classifier algorithm provided the most promising result.

Proposed Study

SMART INTRUSION DETECTION HANDLER

The Result in the previous section suggest that the performance can further be improved if an integrated tool encompassing different intrusion detection algorithm is proposed to handle the variation in attacks and intrusions ,the proposed SMART intrusion detection handling tool can be the solution to this ,providing quite impressive result .This tool is equipped with GAUSSIAN (GAU), MULTI-PERCEPTRON(MLP) and K-MEANS(K-M) algorithm as they proved to be the best in detection of Intrusion .i.e. GAU for R2L,MLP for Probing and K-means for DoS, as shown in **Figure -1**

SMART INTRUSION DETECTION HANDLER (**SIDH**)



Further testing the performance of the tool so developed was tested on allied datasets and the result of performance comparison is tabulated in Table -2:

TABLE-2

Data Set		PROBE	DoS	U2R	R2L
KDD CUP Winner	Pd	0.81	0.831	0.230	0.098
	FaR	0.110	0.008	0.006	0.002
KDD CUP Runner-up	Pd	0.880	0.970	0.291	0.06
	FaR	0.27	0.035	0.035	0.002
Aggarwal and Joshi	Pd	0.777	0.969	0.021	0.030
	FaR	0.45	0.025	0.000006	0.0001
SIDH	Pd	0.880	0.952	0.0005	0.051
	FaR	0.005	0.025	0.291	0.0001

V. CONCLUSION AND FUTURE SCOPE

By the above stated comparative statement in TABLE-2 it is reasonable to state that SMART Intrusion Detection Handler is capable of handling better result in detection of different attacks and intrusions more effectively, apart from this the cost per test example can also be achieved with this tool.

The IDS uses Eccentric Classifier which is adaptive to the network traffic characteristics since the features selected to focus on the inclined nature of the network protocol. In addition, pattern matching operations are now integral. They are activated after performing light validations and advantage from an affable domain pursuit of signatures. The system had been validated in web atmosphere and the end results are provided. Results demonstrate enhanced performance in view of the detection rate and the time needed to detect intrusion.

At a future date, there is potential to render evolution or alteration to the proposed clustering and classification algorithms using artificial intelligence to achieve further improved performance. Finally, the intrusion detection system can be expanded as an intrusion prevention system [15] to strengthen the functioning of the system.

REFERENCES

[1] A. M., C. and K., R. (2012). Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set. International Journal of Information and Network Security (IJINS), 1(4).

[2] K. Park and Y. Cheong, "Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm," 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), Bamberg, Germany, 2018, pp. 282-286

[3] Durcikova and M. E. Jennex, "Introduction to Confidentiality, Integrity, and Availability of Knowledge, Innovation, and Entrepreneurial Systems Minitrack," 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA, 2016, pp. 4010

[4] K. Alrawashdeh and C. Purdy, "Toward an Online Anomaly Intrusion Detection System Based on Deep Learning," IEEE International Conference 2016 on Machine Learning Applications , Anaheim, California, USA, 2017, pp. 195-200

[5] Levin,KDD-99 classifier learning Contest Llsofts Result Overview",SIGKDD ,January 2000,Vol.I (2) pp 67-75.

[6] L.Ertoz,M.Steinbock and V.Kumar in finding cluster of different sizes and shapes,and densities in Noisy and high dimensional data" Technical Report.

[7] D.Y Yeung and Chow „ Parzen Window network Intrsion Detection".Sixteenth international conferance on platform recognition ,Quback city canada P.P 11-15.

[8] C4.5 Simulator downloaded from: https://github.com/scottjulian/C4.5/tree/master/src/main/java/myc45

[9] Zaghian, A. & Noorbehbahani, F. Pattern Anal Applic (2017) 20: 701. https://doi.org/10.1007/s10044-015-0527-6.

[10] Pérez-Miñana E. (1998) A Generative Learning Algorithm that uses Structural Knowledge of the Input Domain yields a better Multi-layer Perceptron. In: Bullinaria J.A., Glasspool D.W., Houghton G. (eds) 4th Neural Computation and Psychology Workshop, London, 9–11 April 1997. Perspectives in Neural Computing. Springer, London

[11] Parvin H., Mohamadi M., Parvin S., Rezaei Z., Minaei B. (2012) Nearest Cluster Classifier. In: Corchado E., Snášel V., Abraham A., Woźniak M., Graña M., Cho SB. (eds) Hybrid Artificial Intelligent Systems. HAIS 2012. Lecture Notes in Computer Science, vol 7208. Springer, Berlin, Heidelberg

[12] A. K. Amoura, J. König and E. Bampis, "Scheduling Algorithms for Parallel Gaussian Elimination With Communication Costs," in IEEE Transactions on Parallel & Distributed Systems, vol. 9, no. , pp. 679-686, 1998. doi:10.1109/71.707547

[13] M. Ohsaki, et al.,"Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification" in IEEE Transactions on Knowledge & Data Engineering, vol. 29, no. 09, pp. 1806-1819, 2017.

[14] V. Lesser, E. Durfee and D. Corkill, "Trends in Cooperative Distributed Problem Solving" in IEEE Transactions on Knowledge & Data Engineering, vol. 18, no. 01, pp. 63-83, 1989.

[15] G. Smith, et al., "Anatomy of a Real-Time Intrusion Prevention System," in Autonomic Computing, International Conference on, null, 2008 pp. 151-160.

[16] P. Anitha, D. Rajesh, K. Venkata Ratnam, "Machine Learning in Intrusion Detection – A Survey", International Journal of Computer Sciences and Engineering, Vol.7, Issue.3, pp.112-119, 2019.

[17] Madhavi Dhingra, "Survey on Intrusion Detection System Based on Feature Classification and Selection", International Journal of Computer Sciences and Engineering, Vol.7, Issue.3, pp.399-403, 2019.

[18] P. Patil, T. Bagwan, S. Kulkarni, C. Lobo, S.R. Khonde, "*Multi-Attacks Detection in Distributed System using Machine Learning*", International Journal of Computer Sciences and Engineering, Vol.7, Issue.1, pp.601-605, 2019.

