International Journal of Engineering & Technology, *7 (4.36) (2018) 261-263*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Performance Analysis of Misuse Attack Data using Data Mining Classifiers

Dr. Anitha Patil[1]*, M. Srikanth Yadav[2]

[1]Proferssor, Department of CSE, Pillai HOC College of Engineering and Technology, Maharashtra, India
[2]Associate Professor, Department of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P, India

Abstract

Data mining can be characterized as the extraction of certain, already un-known, and conceivably valuable data from information. Various analysts have been creating security innovation and investigating new techniques to recognize digital assaults with the DARPA 1998 dataset for Intrusion Detection and adjusted renditions of this dataset KDDCup99 and NSL-KDD, yet as of not long ago nobody have inspected the execution of Top information mining calculations chose by specialists in information mining. The execution of these calculations are contrasted and precision, blunder rate and normal cost on changed renditions of NSL-KDD prepare and test dataset where the occasions are ordered into typical and four digital assault classes: DoS, Probing, R2L and U2R. Furthermore, the most vital highlights to identify digital assaults in all classifications and in every classification are assessed with Weka's Attribute Evaluator and positioned by Information Gain. The goal of this paper is to estimate the performance of classification models like logistic regression, artificial neural networks and support vector machines for predicting intrusions and these techniques are examined to improve the accuracy and performance of these models on KDDCUP dataset. The predictive models are developed using 42 input variable and 23 output variables from the attack set. We examined these data mining models in terms of their accuracy, sensitivity, specificity and FAR. The regression model achieved an accuracy of 99.62%, sensitivity is 99.01%, and specificity is 92.18% with a FAR of 7.82. The Multilayer perceptron (ANN) model achieved an accuracy of 99.62%, sensitivity is 99.01%, and specificity is 91.03% with a FAR of 8.97. The last model Support vector machine model achieved an accuracy of 99.62%, sensitivity is 99.01%, and specificity is 88.00% with a FAR of 12.00. The logical regression model had the better false alarm, sensitivity and specificity, followed by the Multilayer perceptron model and the support vector machine model. The most imperative highlights to distinguish digital assaults are essential highlights, for example, the quantity of seconds of a system association, the convention utilized for the association, the system benefit utilized, ordinary or mistake status of the association and the quantity of information bytes sent. The most vital highlights to distinguish DoS, Probing and R2L assaults are essential highlights and the minimum critical highlights are content highlights. Dissimilar to U2R assaults, where the substance highlights are the most imperative highlights to identify assaults.

Keywords*:* Intrusions, KDD Cup, Misuse, Neural Networks, Regression, Support vector Machines

1. Introduction

The customary events of the world are in need of speedy access and processing of information. In this scenario, demand would increase and correspondingly larger amount of information and resources need to be stored in different computers with a necessary correlation between them. Due to the proliferation of systems and increased network connection, the illicit access and interfering of data would be aggravated. Consequentially, a virtual access path would be created to unauthorized users in the networks. Normally, intruders have an ability to determine the flaw in systems or networks and take advantage of them for misusing them.

Access controls and protection procedures are not adequate for the compromised and inside threats. To recognize the intruders and intrusions is the absolute elucidation to shield systems and networks. So, the intrusion detection systems should not only identify threats and also to monitor the attempts made by intruders.

A trustworthy structure should secure its resources and data from unauthorized access, tampering, and a denial of use. The function of any computer network system should have some expected level of trust and confidence. For each and every system, the protection policy is to be formulated based on the predictable performance. Normally, the computer security is based on the realization of the following factors in a computer machine.

- Sys_Confidentiality – It is the measure, to check whether the information is going to be accessed only by authoritative people.
- Sys_Integrity – The status of information should not alter in any malicious manner.
- Sys_Availability – Computer systems should function without the degradation of admittance and to allocate resources to genuine users when they need it.

In general, an intrusion is described as a sequence of events that tries to negotiation the confidentiality, integrity violation and denial of resources. Anderson (1980) defined an intrusion as the impending opportunity of an intentional unauthorized attempt to right to use information, influence information, or make a system untrustworthy. So, an intrusion is an attempt to break or violate the security policy of the intrusion detection system.

An Intrusion Detection System (IDS) was commercially introduced in the year 1990. It behaves like a burglar alarm which detects any kind of invasion and triggers alarms like audible, visual or messages like e-mail. The IDS is mainly used to protect the machine from the intruders, which may cause to generate an attack or abuse the system, in order to detect new attacks and to deal with known attacks, the attack database has to be updated periodically and it should be documented. But the mechanism should have low false alarms while ensuring the detection of invasion. Intrusion detection systems are appropriate everywhere to defense current networks and no complete and systematic methodology is available to test the effectiveness of these systems. Though there are various approaches, they are relatively ineffective in the classification and alarm rate dimensions. The Data mining based misuse detection methods have been effectively used in the network intrusion detection systems. Because of their extensive capabilities of discovering new attacks.

2. Classifier Model with Logistic Regression

The Misuse attack dataset has been supplied to the logistic regression model. The dataset consisting of 5857 instances with 41 attributes. The pseudo code for logistic regression is shown in figure 3.1 and it is used to find edge based estimation. Let there are k classes for m instances and n attributes, the attribute matrix M is going to be calculated by using n*(k-1) matrix. The elementary equation of a generalized model can be represented as

$$lk\ (E(x)) = \alpha + \beta y1 + \gamma y2 \text{ --- } (3.1)$$

Here, lk () is the linkage function, E(x) is anticipation of target variable and the linear predictor can be generated by using α + βy1 + γy2. The role of linkage function is to connect the expected values of x with other values by using the linear predictors α, β and γ.

The linear relationship between dependent and independent variables can be represented as, here 'Attack' is dependent variable.

$$lk(x) = \beta(\text{Attack}) + \beta_i \text{ --- } (3.2)$$

The linkage function, 'lk()' is established using primarily two things, First one Probability of Success (p) and the second one probability of Failure (1-p), and the criteria for p is either p>=0 or p<=1. To get the logistic regression results, we have to satisfy both the conditions. The probability of having an attack can be predicted by using the following equation

$$p = e\text{^}(\beta i + \beta(\text{Attack})) \text{ --- } (3.3)$$

To get the probable value less than 1, we must split the probability value p by a number higher than p, and it is represented as,

$$p = e\text{^ }(\beta i + \beta(\text{Attack})) / e\text{^}(\beta i + \beta(\text{Attack})) + 1\text{--- } (3.4)$$

Redefine the probability using equations (3.2), (3.3) and (3.4) as:

$$p = (e \text{ ^ } x/ 1 + e \text{ ^ } x) \text{ --- } (3.5),$$

Here p is the probability of success and probability of failure can be represented as

$$q = 1 - p = 1 - (e \text{ ^ } x/ 1 + e \text{ ^ } x) \text{ --- } (3.6),$$

Where q is the probability of Failure. The linkage function can be derived by applying logarithmic function and the derived equation of (d) and (e) as

$$\text{Log } (p/1\text{-}p) = x = \beta(\text{Attack}) + \beta_i \text{ --- } (3.7)$$

The equation (f) is used in logistic regression.

3. Classifier Model with Artificial Neural Networks

A Multilayer perceptron is a classifier meant for linear activity. It classifies the given input into two groups with a straight line. Input value is characteristically a feature vector indicated as "x" multiplied with its weights values say "w" and the resultant is added to a bias value "b" as b:y=w*x + b.

A Multilayer perceptron generates a distinct output depends on a number of real-valued inputs by forming a linear arrangement by means of its input weights, and the process can be represented mathematically as follows in equation 4.1, where the value "w" represents the weight of a vectors, the vector of inputs is i, and the bias value b.

$$y = \emptyset\left(\sum_{i=1}^{n} wixi + b\right) = \emptyset(wTx + b) \qquad (4.1)$$

Neural network techniques have been largely used for intrusion detection in view of the fact that these techniques are does not have need of more parameters to get optimized result. In the neural networks method, initially N data samples were given to input layer and it tries to predict the behavior of next sample using first N samples, and it is considered as output. This paper essentially deals with the Multilayer Perceptron Neural Network model for host based intrusion systems using log files, which are generated by a personal computer.

MLP method is categorized into Feed Forward Neural Network and Back Propagation Neural Network. This method contains a total of three layers namely, the hidden layer, an input layer and an output layer. In this training dataset, each data point represents either normal or abnormal class. The abnormal data points are unspecified as intruder data.

The following formulas are going to be used to measure the performance of the recommended system.

Detection Accuracy = (TN + TP) (TN + TP + FN + FP) (4.2)

Precision = (TP) (TP + FP) --- (4.3)

Recall = (TP) (TP + FN) --- (4.4)

In this experimental study, the dataset has been examined to detect whether there is an attack or not. The results are depicted in Figure 4.1. The attributes duration, protocol and service is taken into consideration to detect network data packet is either normal or attack. In order to detect the type of attack, attribute protocol is taken into account for detecting protocol related attack. The processed result is shown in figure 4.2.

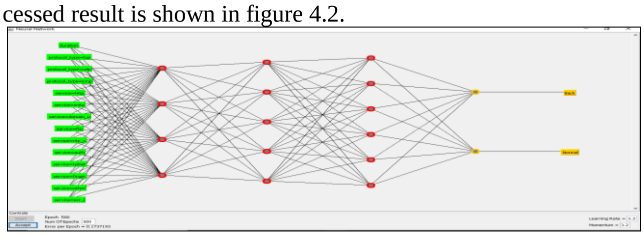


Fig 4.1: To check whether the captured data packet is a normal or malicious

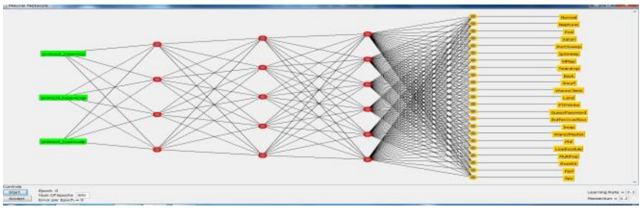


Fig 4.2: To Detect the type of Attack, using protocol

An attempt has been made to use neural network based data mining models on network dataset. We have taken a few earlier days of training and testing data from log files, which are stored in comma separated values design for investigational examination. The dataset contains 5857 records, which are described by using a total of 42 attributes.

4. Classifier Model with Support Vector Machines

Support vector Machine method is going to replaces all missing values and transforms the nominal attributes into numerical or binary values. Support vector machines algorithm also tries to normalize all attributes by default. Multi-class problems are solved using pair-wise classification.

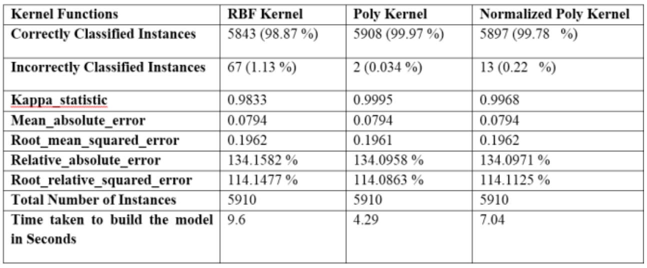
The Following Kernel functions have used for experimental study. The statistics like time taken to build the classifier model, accuracy have been compared. The kernel functions are represented mathematically in the following equations. The results obtained for all three types of kernel functions used in Support Vector Machines are shown in Table 4.1.

(i) *The RBF kernel:* Kernel (a, b) = e^-(<a-b, a-b>^2 * gamma)

(ii) *The polynomial kernel:* Kernel (a, b) = <a, b>^p or Kernel (a, b) = (<a, b>+1) ^ p

(iii) *The normalized polynomial kernel:* Kernel (a, b) = <a, b>/ sqrt (<a, a><b, b>)
Where <a, b> = Poly_kernel (a, b)

Table 4.1: Classifier Results based on kernel function



Kernel Functions	RBF Kernel	Poly Kernel	Normalized Poly Kernel
Correctly Classified Instances	5843 (98.87 %)	5908 (99.97 %)	5897 (99.78 %)
Incorrectly Classified Instances	67 (1.13 %)	2 (0.034 %)	13 (0.22 %)
Kappa_statistic	0.9833	0.9995	0.9968
Mean_absolute_error	0.0794	0.0794	0.0794
Root_mean_squared_error	0.1962	0.1961	0.1962
Relative_absolute_error	134.1582 %	134.0958 %	134.0971 %
Root_relative_squared_error	114.1477 %	114.0863 %	114.1125 %
Total Number of Instances	5910	5910	5910
Time taken to build the model in Seconds	9.6	4.29	7.04

5. Results and Observations

The consolidated results obtained from the above mentioned three classification's techniques is shown in table 5.1,

Table 5.1: Experimental results of comparison of all the three classifiers

Parameter	Logistic Regression	Multilayer Perceptron	Support Vector Classifier
No. of instances classified Correctly	2250 (77.75 %)	2250 (77.75 %)	2242 (77.47 %)
No. of instances classified Incorrectly	644 (22.25 %)	644 (22.25 %)	652 (22.53 %)
Kappa statistic	0.1	0.1	0.0835
Mean_absolute_ error	0.3435	0.3318	0.2253
Root_ mean_squared_error	0.4144	0.4156	0.4747
Relative_ absolute_error	94.47%	91.25%	61.96%
Root_ relative_squared_error	97.20%	97.49%	111.33%
Total Number of Instances	2894	2894	2894
Time taken to build model	0.24 seconds	23.13 seconds	2.33 seconds

6. Conclusion

The objective of this paper is to classify the network dataset using the classifiers like Logistic Regression, Multilayer Perceptron and the Support Vector Machines. The proposed model has produced better results with limited resources.

References

[1] J.Quinlan, "Programs for machine learning" .Morgan Kaufmann, (1993)
[2] G. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pp 338-345 (1995)
[3] C. Chang and C. Lin.,"LIBSVM : a Library for Support Vector Machines", (2001)
[4] Software at http://www.csie.ntu.edu.tw/ cjlin/libsvm
[5] NSL-KDD: available on
[6] http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html, (2009)
[7] KDD CUP 1999, available on:
[8] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[9] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln Laboratory",ACM Transations on Information and system security, vol. 3, no. 4, pp. 262–294 (2000)
[10] C. E. Rasmussen, fmincg minimization function.
[11] http://learning.eng.cam.ac.uk/carl/code/minimize/
[12] "Waikato environment for Knowledge analysis (Weka) and Using Weka in Matlab"
[13] http://www.mathworks.com/matlabcentral/fileexchange/50120-using-weka-in-matlab
[14] M. A. Hall and L. A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach", University of Waikato, (1997)
[15] Denning DE, Edwards DL, Jagannathan R, Lunt TF, Neumann PG, "A prototype IDES: A real-time intrusion detection expert system". Technical report, Computer Science Laboratory, SRI International, Menlo Park
[16] M. Roesch, "Snort — lightweight intrusion detection for networks". In Proceedings of the 13th Systems Administration Conference, pp 229 – 238, Seattle, WA, USA, Usenix Association, , (1999)
[17] Jagannathan R, Lunt TF, Anderson D, Dodd C, Gilham F, Jalali C, Javitz HS, Neumann PG, Tamaru A, Valdes A, "System Design Document: Next-generation intrusion-detection expert system (NIDES)". Technical report, Computer Science Laboratory, SRI International, Menlo Park, (1993)
[18] Darren Anstee, Denial of service attack data, Arbor Networks Inc.,2015.
[19] Andy Meek, DDoS attacks are getting much more powerful and the Pentagon is scrambling for solutions,2015.
[20] Joseph Steinberg, Denial of Service Attacks Are Growing Increasingly Problematic: Here's What You Need To Know, 2015.
[21] Carl G, Kesidis G, Brooks RR, Rai S. Denial-of-service attack-detection techniques. Internet Computing, IEEE. 2006 Jan;10(1):82-9.
[22] Gavaskar S, Surendiran R, Ramaraj DE. Three Counter Defense Mechanism for TCP SYN Flooding Attacks. International Journal of Computer Applications. 2010 Sep;6(6):0975-8887.
[23] Kavisankar L, Chellappan C. A Mitigation model for TCP SYN flooding with IP Spoofing. Proceeding of IEEE International Conference on Recent Trends in Information Technology (ICRTIT), 2011, pp. 251-256.
[24] Ng J, Joshi D, Banik SM. Applying Data Mining Techniques to Intrusion Detection. Proceeding of IEEE 12th International Conference Information Technology-New Generations, 2015, pp. 800-801.