Regression Based Network Security Scenario Prediction Model

Ahmed Mutar Awad[1*], *Yousif* Hamad Efan[1] *and Qays* Neamah Ibrahim[1]

[1]General directorate of education, Anbar, Iraq

Abstract. Analysing issues in which one or more independent variables predict an outcome may be done using logistic regression. A binary or dichotomous dependent variable is used to quantify the result, which only comprises data coded as 1 (True, Success, etc.) or 0 (False, Failure, etc). Logistic regression is used to identify the best model to represent the connection between a dependent variable (outcome or response variable) and a collection of independent (predictor or explanatory) variables. Biomedical applications of LR (Linear Regression) include cancer detection, survival prediction, and more. This is a popular and well-established data analysis approach in statistics and biomedicine. It is also recommended to compare data mining approaches with logistic regression when mining clinical data. In this research, various ideas about network security assessment are described and compared,, and a technique for evaluating the current state of network security using a virtual Honeynet is proposed. A binary linear regression model is developed based on the correlations between Honeynet active, host active of computer networks, and IP active when network intrusion occurs. Prototype systems for data gathering and regression fitting are built, proving the validity of regression prediction models.

I. Introduction

The proliferation of computer networks has resulted in a high number of economic losses as a result of network security incidents. It is necessary to conduct an analysis of network security from the perspective of all networks and to forecast and evaluate trends in network security conditions. Because of this, network security scenario assessment has been a study focus in recent years. Network security situational awareness is described as the ability to predict network security trends by collecting and synthesising all relevant information that represents the current state of security. In the past several years, situation assessment has been used at computer network security research, which was originally used in airports. This paper proposes and evaluates computer network security using data fusion and data mining methods to combine (Distributed Intrusion Detection System)DIDS data. There is a degree of difficulty in application because of the low performance of real-time calculating. NVisionIp can dynamically show network connection status and network flow as well as data filtering capabilities created by the SIFT project team. Despite this, the only security assessment element is a high level of expertise required by its users. A B-class network connection status may be shown by NVisionIP in three different hierarchies. An (Intrusion Detection System) IDS alarm and performance indicators-based quantitative hierarchical network security threat assessment technique has been developed by Chen Xiuzhen et al. Security threat indexes may be assessed in a hierarchical manner with the use of information about the holes in the host system.

Even so, it will take a while to assess. An attack-based network analysis model was suggested by Zhang Haixia et al. It is more appropriate to use the attack ability's increment to explain the aim than the attack ability itself. After taking these factors into account, it evaluated network security based on cost of attack route. A solid basis for network security assessment models and algorithms has been laid by both domestic and international academics. There are, however, a few issues. The evaluation method will be sluggish and difficult to achieve assessment prediction requirements if network security elements are not taken into account comprehensively. On the other hand, if network security factors are taken into account comprehensively, the evaluation will be rapid but incorrect.

* Corresponding author: amaacs2@gmail.com

This study presents a technique for network security scenario evaluation using Honeynet in response to the aforesaid issues. Honeynet is a network of active decoys made up of several nodes. Nodes in Honeynet are designed to be simple to attack and to impersonate a wide range of genuine services or applications. By using Honeynet, researchers may learn about and analyse the habits of hackers, and they can use this information to entice an invading attempt on their own systems. Normal network traffic will not be seen in Honeynet because of its structure. To assess the present state of network security, we may use Honeynet active to presume that all network connection activity in Honeynet are malicious.

II. Literature survey

As computer technology has progressed in terms of speed, affordability, as well as access to enormous quantities of processing power in a short period of time, there has been a rise in the use of data mining techniques. The majority of these data mining applications have made use of machine learning. A learning system's knowledge comes mostly from what it learns from the data it receives as input. The usefulness of several feature selection strategies for pre-processing input data has been evaluated in very few research on pre-processing data used as input into these data mining systems. These solutions are evaluated using real-world financial credit-risk data.

Financial institutions such as banks and home loan businesses face various challenges in decision-making, which can be addressed by using a range of processing techniques and tools. It is widely accepted that neural networks are one of the most promising methods. Credit loan application categorization has been studied in terms of the ideal parameters of three models (MLP, Ensemble Averaging, and Boosting by Filtering) that have been compared in terms of efficiency and accuracy. In this case, the objective was to determine which of the three neural network models would be the most useful in this particular decision situation. Committee Machine Models (CMMs) outperformed a single Multi-Layer Perception model, while Ensemble Averaging outperformed Boosting by Filtering (Meliha Handzic et al., 2003).

The study of corporate credit ratings has sparked a lot of debate. Artificial intelligence (AI) technologies outperform conventional statistical methods, according to recent research. Support Vector Machines (SVMs) are introduced in this thesis to answer the issue in an effort to give a model with more explanatory power. As a benchmark, the Back Propagation Neural Network (BNN) was utilised, and a prediction accuracy of roughly 80% was achieved for both the United States as well as Taiwan markets. Only a little increase in SVM performance was seen. Another goal of the study was to make AI-based models easier to understand. Relative relevance of input financial variables from neural network models was determined by using recent research findings. It was determined that there were significant disparities between the American and Taiwanese markets when it came to the deciding elements (Zan Huanng et al., 2004).

For the logistic regression, an efficient feature assessment issue was examined. Features are ranked in order of their predicted impact on the model's performance in this thesis’s forward feature selection methodology. The coefficient of each additional feature in the logistic regression model may be estimated quickly and accurately using an approximation optimization based on back fitting. We can also swiftly analyze billions of possible features even for extremely large datasets because of the parallelization of the algorithm across features and records, which is very scalable (Singh et al., 2009).

For the logistic regression, an efficient feature assessment issue was examined. Features are ranked in order of their predicted impact on the model's performance in this thesis’s forward feature selection methodology. The coefficient of each additional feature in the logistic regression model may be estimated quickly and accurately using an approximation optimization based on back fitting. We can also swiftly analyze billions of possible features even for extremely large datasets because of the parallelization of the algorithm across features and records, which is very scalable (Singh et al., 2009).

In remote sensing data processing, feature selection is critical, especially when it comes to classification from hyper spectral pictures. Based on input characteristics, an LR model may be used to forecast the class probabilities based on their relative relevance. Remotely sensed pictures are classified using the LR model, which generates inherently more informative soft classifications as well as features for feature selection. There was no significant drop in classification accuracy for both soft and hard classes with less restrictive assumptions in the LR model, according to the data (Qi Cheng et al., 2006). Diagnostics of power distribution faults rely on logistic regression, whereas studies into power system dependability rely heavily on neural networks. Criteria for evaluating the classifier's performance include: accuracy, true-positive rate, true-negative rate, and geometric mean (mean square error) (Le Xu et al., 2005).

Logistic regression and Artificial Neural Networks (ANN) have been compared and an experiment on kidney transplant outcomes prediction has been carried out. For kidney transplant outcomes, the findings show that ANN and bagging are useful data mining methods. This also supports the possibility of combining several strategies in order to improve forecast accuracy. Results show that ANN beats logistic regression in the majority of scenarios.

III. Methodology and metrics

Building of Regression Model: As the number of Honeynet nodes on a computer network grows, so does the importance of Honeynet Active in terms of network security. Increasing infiltration activity will lead to an increase in data traffic, which might lead to a severe network security scenario. As a result, Honeynet Active's predictions may be utilised to forecast the network's security. Hence,

$$y = \text{HoneynetActive} --(1)$$

As the number of Honeynet nodes on a computer network grows, so does the importance of Honeynet Active in terms of network security. Increasing infiltration activity will lead to an increase in data traffic, which might lead to a severe network security scenario. As a result, Honeynet Active's predictions may be utilised to forecast the network's security. Hence.

This type of predictive model can be used when the target variable is a categorical variable with two categories - for example, active or inactive, healthy or unhealthy, win or lose, purchase product or does not purchase product etc. Logistic regression is also known as logistic model and logit model. The chance of an event occurring may be predicted by fitting the data into a logistic curve using logistic regression. classification error, squared error, and negative log-likelihood are three methods to characterise the prediction error that may be taken into account. It uses predictor variables, which may be numerical or categorical, much like other regression analysis methods. For example, knowing a person's age, sex, and Body Mass Index (BMI) might help estimate the likelihood that he or she would have a heart attack within a certain period of time. Medical and social science research, as well as commercial applications like predicting whether a consumer will buy or cancel their subscription, employ logistic regression. When an illness is present, the subject's reaction, Y, is 1; when it is absent, it is 0 (for example, Y=1 in the presence of a disease). A vector of explanatory variables, X, may be defined as x1 (or x2,... xn). Explanatory factors are explained in binary form using the logistic regression model.

$$\text{Logit}\{\Pr(Y = 1 \mid x)\} = \log\left\{\frac{\Pr(Y = 1 \mid x)}{1 - \Pr(Y = 1 \mid x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k - (1)$$

Where the variables are:

*Y: This represents the binary outcome variable. In logistic regression, Y typically represents the binary outcome, such as 0 or 1, success or failure, yes or no, etc.

*x1, x2, ..., xk: These represent the predictor variables or features. In logistic regression, these are the independent variables that are used to predict the probability of the binary outcome (Y).

*β0, β1, β2, ..., βk: These are the coefficients associated with each of the predictor variables (x1, x2, ..., xk). These coefficients are estimated during the model training process and represent the strength and direction of the relationship between the predictor variables and the log-odds of the binary outcome.

“Regression coefficients” of x1, x2, x3 are termed "intercepts," while 0 is called the “intercept.” Every one of the regression coefficients is a measure of how much the risk factor contributes. An increase in the probability of an outcome is indicated by a positive regression coefficient, while a decrease in the probability is indicated by a negative regression coefficient. A large regression coefficient indicates that the risk factor has a strong influence on the probability of that outcome, while a non-zero regression coefficient indicates that the risk factor has little influence on the probability of an outcome.

The logistic function is given by

$$P = \frac{1}{1 + e^{-\log it(p)}} ----(2)$$

Where the variables are:

P: This represents the probability of the event occurring, where P is between 0 and e: This denotes the mathematical constant approximately equal to 2.71828.

logit(p): The logit function transforms the probability (P) into the log-odds of the event occurring, calculated as the natural logarithm of the odds ratio.

Figure 1 depicts a logistic regression function graph. For example, the logistic function may accept any number from -infinity to -infinity as an input, while the result is limited to values between 0 and 1.

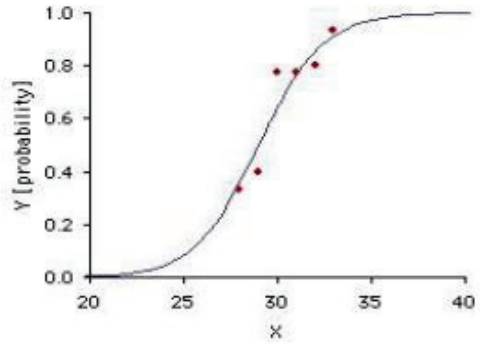

Fig. 1. A graph of logistic regression function

Linear models come in a plethora of varieties. Although the response variable has been assumed to be normally distributed in all the models analysed, this is not the case here. Only two potential results may be achieved when the response variable is a categorical random variable, as discussed in this module. We see much of this type of data. There are many examples of responses, such as whether or not an individual gets better, whether or not an object passes quality control in a manufacturing process, and so on and so forth. Accordingly, statistical techniques cannot be used in this study due to the fact that the response variables are dichotomous (i.e., they have only two potential outcomes). Logistic regression is the most often used approach for data analysis including dichotomous answer variables.

In our network penetration tests, we show that the host's behaviour changes when a large number of worm infections appear or when hackers launch scanning assaults. For example, viruses may compel some hosts to disconnect from the network.. Also, IP Active is going to change. Here, the link between Host Active, IP Active, and Honeynet Active is examined. While 1PActive is labelled as X2, Hostvar is marked as Xl. In this country, computer networks and honeynets have developed into a complicated system. Create the variables Hostvar and IPActive in the Xl and X2 formats, respectively. Both Hostvar and IPActive affect Honeynet Active, however the results of our experiments show that there is no straight line between the two. There is a power-law connection between Honeynet, IP, and Host actives if this is the case.

$$y = b_0 x_1^{b_1} x_2^{b_2} - - - - - - - - (3)$$

Where the variables are:

(y = b_0 \times x_1^{b_1} \times x_2^{b_2}), the variables are defined as:

(y): This represents the response variable or the dependent variable. It is the quantity that is being modeled or predicted based on the values of the predictor variables.

(x_1): This is the first predictor variable or independent variable. It is one of the factors that is believed to have an impact on the value of the response variable (y).

(x_2): This is the second predictor variable or independent variable. Similar to (x_1), it is another factor that is believed to have an impact on the value of the response variable (y).

(b_0): This is the intercept coefficient. It represents the value of the response variable when all predictor variables are equal to zero.

(b_1): This is the coefficient for (x_1). It represents the effect of (x_1) on the response variable (y).

(b_2): This is the coefficient for (x_2). It represents the effect of (x_2) on the response variable (y).

Once this is done, transform the model to a linear one by changing the variables. Following this procedure, both sides of relation (2) were subjected to the logarithm operation.

$$\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 --(4)$$

Where the variables are:

(y): This represents the dependent variable, and the natural logarithm transformation is applied to it. The natural logarithm of the dependent variable is used to transform it into a linear form to meet the assumptions of linear regression.

(b_0): This is the intercept term, also referred to as the constant term, which represents the value of the dependent variable when all independent variables are equal to 1.

(x1) and (x_2): These are the independent variables, and the natural logarithm transformation is applied to each of them.

(b_1): This is the coefficient for (ln(x_1)) , representing the effect of the natural logarithm of (x_1) on the natural logarithm of the dependent variable.

(b_2): This is the coefficient for (ln(x_2)), representing the effect of the natural logarithm of (x_2) on the natural logarithm of the dependent variable.

Commands $Y = \ln y, B_0 = \ln b_0, X_I = \ln x_I$ and $X_2 = \ln x_2$, then

$$Y = B_0 + B_1X_1 + B_2X_2 ------(5)$$

The extended linear model in equation (5). Xli, X2i, and Yi are the observation data for the I time of n tests on a two-dimensional surface. When Xli = lnXLi, X2i= Inx2i' Yi = In yi, the logarithm procedure was completed.

$$\hat{Y}_i = B_0 + B_1X_1 + B_2X_2 ----- (6)$$

According to least square method, makes $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ minimum, and gets B_0, B_I and B_2 using the method of extremum seeking.

IV. Experiments and Analysis

30 machines with internal network IP addresses were chosen for the experiment, one of which serves as the virtual Honeynet's master node. Twenty virtual honeypots with the same domain as the host computer may be simulated by utilising a physical computer. The rest of the machines, including 18 Windows PCs, seven Linux systems, and three FreeBSD computers, are operating regularly. Tcpdump is used to gather IP data packets in computer networks, and Perl is used to interpret the Tcpdump log. As a result, the experiment's host and IP become active at time ti. Honeynet Active is difficult to get in real time from virtual node logs, however it is possible to do so via statistical analysis. For the purpose of conducting experiments in which several assaults are made to the experiment network segment using current intrusion software or self-created worm programmes and data is extracted for regression analysis, the regular working period is split into three-time quanta.

The next experiment is to see whether the binary regression model can forecast the future. To begin, check the model's false positive rate. True positive rate is the percentage of false alarms in which an assault really occurs. Honeynet Active's anticipated and actual values were compared using a 10-fold dilution of the typical working day's IP Active and Host Active data. Figure 2 shows the comparison results. The red broken line depicts the expected values based on the model, whereas the blue broken line depicts the actual values that were actually observed. The projected values of Honeynet Active are much higher than the actual values when the network is operating normally. As a result, the regression prediction model has a false positive rate that is nearly nil. There is no presumption of a power function connection between the Host Active and IP Active in a normal network as shown by this result.

Table 1. Network intrusion detection table

Prediction	Actual	Accuracy difference
0.001251	0.563585	0.279326
0.193304	0.808740	0.391283
0.585009	0.479873	0.370964
0.350291	0.895962	0.448847
0.822840	0.746605	0.498549
0.174108	0.858943	0.402505
0.710501	0.513535	0.406161
0.303995	0.014985	0.175294
0.303995	0.014985	0.175294
0.091403	0.364452	0.237616

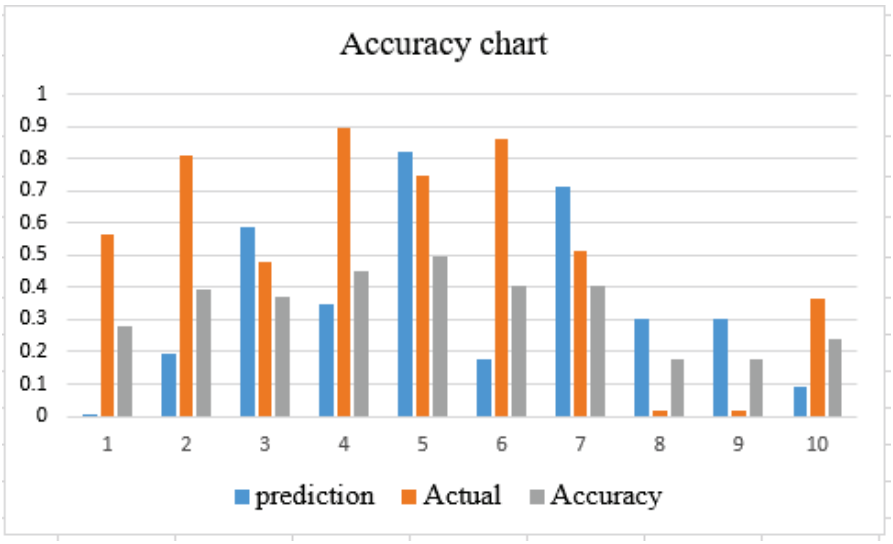

Fig. 2. Network intrusion detection between prediction and actual measures

Simulated network intrusion was performed. Host Active and IP Active have a direct connection to Honeynet Active when a network intrusion occurs. The simulated assault on computer networks was used to carry out ten attacks, and the real Honeynet activity and regression predicted values were computed. Figure 2 depicts the results of tests in which satisfactory match was shown.

V. Conclusion

In this study, a computer network virtual honeynet is developed and its regression prediction model is experimentally proven. Using a regression model, it is possible to anticipate the security condition of a network in a particular range of threshold values and to do so quickly. 619 prediction threshold values can only be selected by artificial intelligence and expert knowledge nowadays. We'll use our fuzzy math expertise to increase threshold value selection accuracy and tackle challenges of specific deviation.

References

1. Brijesh Kumar, Baradwaj and Saurabh Pal. (2011). “Mining Educational Data to Analyze Students Performance”. International Journal of Advanced Computer Science and Applications, 2(6): 64-39.

2. Mahendra Tiwari, Randhir Singh and Neeraj Vimal. (2013). “An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education”. International Journal of Computer Science and Mobile Computing., 2(2):53 – 57.

3. Sherine Dominick and Abdul Razak, T. (2014). “Analyzing the Student Performance using Classification Techniques to find the better Suited Classifier”. International Journal of Computer Applications., 104(4):1-3.

4. WEI Yong LIAN and Yi-Feng, “A Network Security Situational Awareness Model Based on Log Audit and Performance Correction [J]”, Chinese Journal of Computers, vol. 4, pp. 763-764, 2009.

5. Abdullah AL-Malaise, Areej Malibari and Alkhozae. (2014). “Students’ Performance Prediction System Using Multi Agent Data Mining Technique”. International Journal of Data Mining and Knowledge Management Process., 4(5):1-20.

6. Tribhuvan, A.P., Tribhuvan, P.P. and Gade, J.G. (2015). “Applying Naive Bayesian Classifier for Predicting Performance of a Student Using Weka”. Advances in Computational Research., 7(1): 239-242.

7. Humera Shaziya, Raniah Zaheer and Kavitha, G. (2015). “Prediction of Students Performance in Semester Exams using a Naïve bayes Classifier”. International Journal of Innovative Research in Science, Engineering and Technology., 4(10): 9823-9829.

8. Bass T. “Intrusion detection systems & multisensor data fusion: Creating Cyberspace Situational Awareness [J]”. Communications of the ACM, 2000, 43 (4): 992105.

9. Yin Xiaoxin, Yurcik W, Treaster M, et al. VisFlowConnect: “Net Flow visualizations of link relationships for security situational awareness [C] PPProc of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security”. New York: ACM, 2004: 26234.

10. Lakkaraju K, Yurcik W, Lee A J. NVisionlP: “Net Flow visualizations of system state for security situational awareness [C] PPProc of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security”. New York: ACM, 2004: 65272.

11. Jyoti Bansode Shah. (2016). “Mining Educational Data to Predict Student ‘s Academic Performance”. International Journal on Recent and Innovation Trends in Computing and Communication., 4(1): 01-05.

12. Tismy Devasia, Vinushree, T.P. and Vinayak Hegde, “Prediction of Students Performance using Educational Data Mining”. Proc. International Conference on Data Mining and Advanced Computing, Ernakulam, India, pp. 91-95, 2016. ISBN: 9781467385954

13. Mashael A. Al-Barrak and Muna Al-Razgan. (2016). “Predicting Students Final GPA Using Decision Trees”: A Case Study. International Journal of Information and Education Technology., 6(7):528-533.

14. Chen Xiuzhen, Zheng Qinghua, Guan Xiaohong, et al. “Quantitative hierarchical threat evaluation model for network security [J]”. Journal of Software, 2006, 17 (4): 8852897.

15. Zhang Haixia, Su Purui, Feng Dengguo. “A Network Security Analysis Model Based on the Increase in Attack Ability [J] “. Journal of Computer Research and Development, 2007, 44 (12): 201222019.

16. Abu Zohair, L.M. (2019) “Prediction of Student’s performance by modelling small dataset size”. International Journal of Educational Technology in Higher Education.,16(27):1-18.

17. Micheline Apolinar Gotardo. (2019). “Using Decision Tree Algorithm to Predict Student Performance”. Indian Journal of Science and Technology., 12(5): 1-8.

18] Ramanathan. L, Angelina Geetha, Khalid and Swarnalatha. (2016). “Student Performance Prediction Model Based on Lion-Wolf Neural Network”. International Journal of Intelligent Engineering and System., 10(1): 114-123.

19. Oyerinde, O. D and Chia, P. A. (2017). “Predicting Students’ Academic Performances – A Learning Analytics Approach using Multiple Linear Regression”. International Journal of Computer Applications., 157(4):37- 44.

20. Fumiya Okubo., Yamashita, T., Shimada, A. and Ogata, H. “A Neural Network Approach for Students' Performance Prediction”. Proc. Seventh International Learning Analytics and Knowledge Conference, Vancouver, British Columbia, Canada, 2017, pp.598-599.