RESEARCH ARTICLE OPEN

Performance Analysis of Machine Learning Techniques Used in Intrusion Detection Systems

Ufuk Murtaza* Prof. Dr. Zafer ASLAN**
**(Department of Computer Engineering, Istanbul Aydin University,*
*** (Department of Computer Engineering, Istanbul Aydin University,*
Corresponding Author: Ufuk Murtaza

ABSTRACT
In today's age, new developments are constantly occurring in the internet world. These developments, such as the number of internet users and the increase in web applications, have brought some risks in a matter called data security. Intrusion Detection Systems (IDS), a tool used for data and network security, prevent attacks on secure internal networks by developing specific simulations. In addition, it detects unexpected login and access requests and successfully removes threats. In recent history, many researchers have been working on safer IDS to prevent these threats. However, there is a limited number of performance comparisons of IDS machine learning techniques. Different techniques have been studied in the applications. Machine learning techniques such as decision trees, neural networks, random forest, AdaBoost, logistic regression, Naive Bayes, K-nearest neighbor algorithms on a data set of their performance and success rates were evaluated. F-measure, precision, specifity, accuracy and sensitivity analyses were performed and their classification was observed. NSLKDD was used as a data set on the studies. To solve the problem, the data set was analyzed in the Waikato Environment Knowledge Analysis (WEKA) environment. Although many algorithms applied on the data set gave close values, it was determined that K-nearest neighbor (KNN) applications showed the highest performance with a classification rate of %98,56.

Keywords–Intrusion Detection Systems, Machine Learning, K-nearest neighbor, Data Mining

Date of Submission: 18-06-2020 Date of Acceptance: 06-07-2020

I. INTRODUCTION

In this technology age, the internet, which is a place of access and storage of information, has very important place. The fact that the Internet is so important raises security concerns at certain points as well as benefits. It is critically important, especially in sectors such as banking, health, education, finance and electronic government services. The addition of increasing service transactions to the internet environment has attracted the attention of those who want to benefit from this issue and increased the attack rates on the network environment. Although information seems to have come forward, in fact, information; While it is the keys of yesterday and today, information always has key roles in shaping the future [1].

The intrusion detection system is tasked with detecting normal or abnormal behavior by controlling traffic on the network. These are the mechanisms that inform the system specialist of this abnormality. It is the last defense force in security breaches or vulnerabilities of the system. Therefore, the intrusion detection system is of critical importance. Performance and effectiveness of these systems shows that these systems need to be much more powerful. Machine learning techniques are successful in order to improve the effectiveness and performance of the intrusion detection system.These cyberattacks on individuals and institutions affect negatively. These attacks disrupt data and services entered over the internet. Intrusion detection systems have been developed to reduce and detect this negative impact in IT networks. Intrusion detection systems are monitoring and supervising all network traffic. It perceives outgoing or incoming attacks as a threat. It issues a warning by identifying threat as a suspect [2].

II.PURPOSE

In this study, NSLKDD data set is used in different IDS comparison analysis. It is aimed to obtain information about the attack detection system success by using machine learning techniques on the dataset. In order to evaluate the attack types, 42 attributes are used in the data set. A total of 7 machine learning methods were applied to these qualifications. The aim of this study is to compare the results of machine learning techniques applied on the intrusion detection system within the WEKA software and to determine the algorithm that gives the best performance and success rate.

III. SCOPE

Nowadays, it is used in many areas due to the big data problem. In complex problems, machine learning techniques are applied if there is no formula or equation. Machine learning algorithms support making good decisions and making predictions by finding appropriate patterns for the input data [3].

In this study, the success rates of Naïve Bayes, K nearest neighbor, logistic regression, artificial neural networks, decision tree, random forest, adaboost techniques were analyzed according to their attack detection systems and their results were evaluated in terms of performance [4].

IV. METHOD

In this study, data set selection is important for the performance of the data to be tested and the developed algorithms. According to the results of the studies conducted in this field, it is necessary to use the current data set which is also applied worldwide. After the data mining inference competition organized by the MIT Lincoln Laboratory, the data set we use was created by extracting residual information from the KDDCup99 data set. Performance evaluation of machine learning techniques used in intrusion detection systems has been performed. NaiveBayes, K nearest neighbor, decision trees, artificial neural networks, random forest, adaboost, logistic regression algorithm has been used. For the results, a 10-fold cross-passing method was used on the weka software.

Logistic Regression

Logistic regression is often used as a classification. It is used to calculate the relationship between a variable that is independent with logistic regression and the variables that are the result output [5].

Naive Bayes

The Naive Bayes algorithm is called classifiers, which calculate probability by collecting frequentuses of all the values in the data set and combinations that may occur with the data in that set [6].

Random Forest

The random forest algorithm is used for regression analysis and classification. In the random forest algorithm, test data and training data are applied to the decision tree model to obtain results [7].

Artificial Neural Networks

Artificial Neural Networks (ANN) is a powerful classification tool that models biological nerve cells (neuron) as a scheme in its appearance [8].

Adaptive Boosting (AdaBoost)

The Adaptive Boosting (ADABoost) algorithm can be used in conjunction with many learning algorithms for performance improvement. AdaBoost also combines different learning outcomes in one center with classification representing performance improvement and empowerment [9].

K-nearest neighbor (KNN)

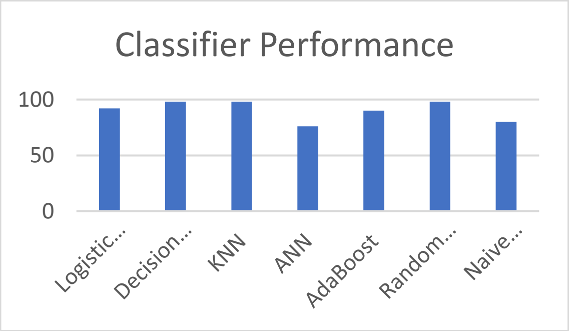
K is the closest neighbor algorithm is generally used for classification and regression analysis [10].

DecisionTree

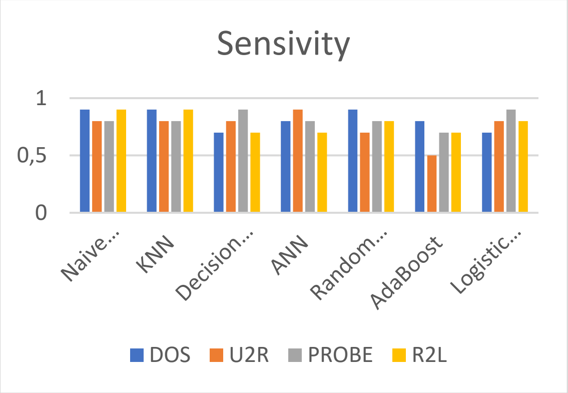
The decision tree algorithm is often used on data mining as its field of use. The goal of this method is to perform the process of classifying the data with in the dataset [11].

V. RESULTS

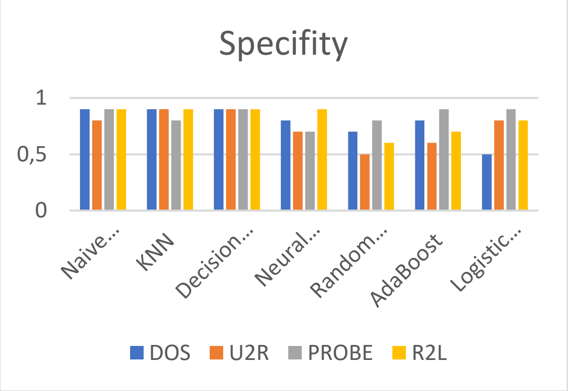
A total of 7 machine learning techniques were used on the data set.



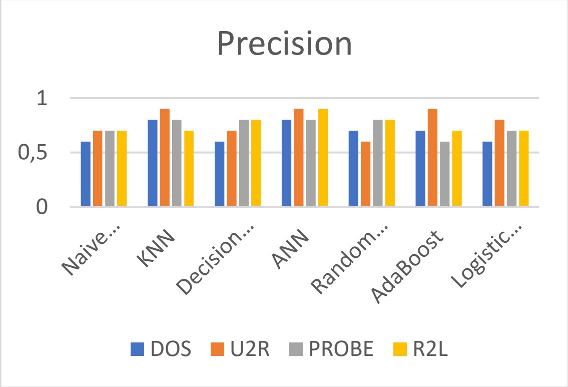
Classifiers, we evaluate success based on its classification, to distinguish normal behavior, KNN are more successful than other classifiers. The detection of DOS attacks KNN, decision tree and random forest has reached a near %99 success. KNN,random forest and decision tree give better results in accurate detection of PROBE attacks. In the R2L and U2R attacks than other classifiers,



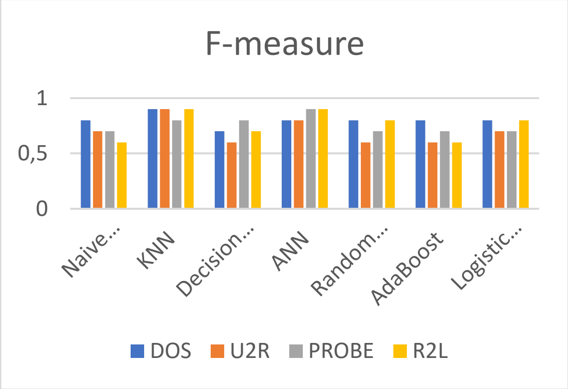
When looking at sensitivity values, Naive Bayes, KNN, ANN, and decision trees are a better classifier in perceiving normal behavior. When we compare classifiers based on sensivity values, decision trees are better classifiers in determining normal behavior correctly, Naive Bayes and decision trees in DOS attacks, KNN, ANN and decision trees in R2L attacks, AdaBoost, KNN, ANN and decision trees in PROBE and U2R attacks.



When comparing classifiers based on their specifity values, decision trees to detect the normal behavior, Naive Bayes and decision trees in DOS attacks, KNN in R2L attacks, ANN and decision trees, in PROBE and U2R attacks AdaBoost, KNN, neural network and decision trees are better classifiers in terms of specifity.



According to the criteria of precision, decision trees in normal behavior, all classifiers are examined in DOS attacks, in PROBE and R2L attacks KNN, whilethe U2R attacks ANN classifier is better.



Since the F-measure includes false positive and false negative values, it gives more reliable results as a measure of accuracy compared to the accuracy measure and is used in the comparison analysis of algorithms. When we look at the F-measure values, decision trees are the best classifiers in correct perception of normal behavior, ANN and decision trees in DOS attacks, decision trees in PROBE and R2L attacks, and KNN in U2R attacks.

V. CONCLUSION

Nowadays, for the reasons like: aspecially globolized world, increasing number of Technologies and networks connecting to each other intrusion detection systems are one of the areas that need to be developed by conducting research on them [12]. Machine learning techniques are important in terms of more effective use of intrusion detection systems [13].

In this study, NSLKDD dataset is used in attack detection systems in terms of classification success, processing time, sensitivity, selectivity, precision and F-measure using Naive Bayes, random forest, K closest neighbor algorithm, logistic regression, artificial neural networks, adaboost, and decision trees. Performance was examined.

NSLKDD has often been the preferred public data set. Of course, researchers can collect their own data. However, this will raise some problems both in terms of costand in terms of comparing the work put forward. The training data allocated for system training is as important as the data set used. Sufficient training data should be used during the system training phase. [14]

Sufficient test data is needed to measure the success of the generated Intrusion Detection System (IDS). A sufficient amount of test data plays an important role in accurately measuring system success [15]. A sufficient amount of test data plays an important role in accurately measuring system success.

REFERENCES

[1] Canbek, G., & SAĞIROĞLU, Ş. Bilgi, Bilgi Güvenliği ve Süreçleri Üzerine Bir İnceleme, *Politeknik Journal,* 2015,165-174.

[2] Buczak L., &Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE, Vol. 18. 2016.

[3] Bhavsar, Y., & Waghmare, K., Intrusion Detection System Using Data Mining Technique: Support Vector Machine, *International Journal of Emerging Technology and Advanced Engineering*, 2013, 581-586.

[4] U. Murtaza, *Saldırı Tespit Sistemlerinde Kullanılan Makine Öğrenmesi Tekniklerinin Performans Analizi* master diss., İstanbul Aydın University, İstanbul, 2020.

[5] Press, S., & Wilson, S., Choosing Between Logistic Regression and Discriminant Analysis, *Journal of the American Statistical Association*, 1978, 699-705

[6] Sharma, N., & Mukherjee S., Intrusion Detection using Naïve Bayes Classifier With Feature Reduction, *Procedia Technology 4*, 2012, 119-128.

[7] Hasan, Mehedi., & Nasser, M., Feature Selection for Intrusion Detection Using Random Forest, *Journal of Information Security*, 2016, 129-140.

[8] Ibrahim, L., & Basheer, D., A Comparison Study for Intrusion Database (KDD99, NSLKDD) Based On Self Organization Map (SOM) Artificial Neural Network, *Journal of Engineering Science and Technology*, 2013, 107-119.

[9] Hu, W., & Maybank, S., AdaBoost Based Algorithm for Network Intrusion Detection, *IEEE Transactions on Systems*, 2008, 577-583.

[10] Li, W., Yi, P., A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network, *Hindawi Publishing Corporation Journal of Electrical and Computer Engineering*, 2014, 01-08.

[11] Farid, D., & Harbi N., Combining Naïve Bayes and Decision Tree for Adaptive Intrusion Detection, *International Journal of Network Security*, 2010, 12-25.

[12] Hoque, M., & Bikas N., An Implementation of Intrusion Detection System Using Genetic Algorithm, *International Journal of Network Security*, 2012, 109-120.

[13] Livingstone, F., Implementation of Breiman's Random Forest Machine Learning Algorithm, *Machine Learning Journal Paper*, 2005, 01-13.

[14] Revathi, S., &Malathi A., A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection, *International Journal of Engineering Research & Technology*, 2013, 1848-1853.

[15] Zainal, A., Shamsuddin, M., Ensemble Classifiers for Network Intrusion Detection System *Journal of Information Assurance and Security 4*, 2009, 217-225.