Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220087

Predict Attacker Behaviour on IDS with High Accuracy Using K-Nearest Neighbor Algorithm

Bhavana M[a] and RajendranT[b, 1]
[a]*Research Scholar, Dept. of CSE, Saveetha School of Engineering,*
[b]*Asso. Prof., Dept. of CSE, Saveetha School of Engineering,*
[a,b] *SIMATS, Chennai, Tamilnadu, India*

Abstract. This work's main objective is to predict the attacker's behavior pattern with high accuracy by using machine learning methods. According to the experimental and statical analysis, the proposed model has improved accuracy. The study was performed with two machine learning algorithms, K-Nearest Neighbor (KNN) and Decision Tree (DTA). On a dataset of 19,864 items, the algorithms were implemented, trained, and assessed. Two iterations have extracted the trained and tested dataset on the sample size. Each algorithm has undergone ten iterations with different test sizes to get different result sets. This study's G-Power test for machine learning algorithms is roughly 80%. The result sets of the programming experiment have been further analyzed with statistical tools and observed that the accuracy of KNN is 99.76, while DTA is 98.84, according to the testing data. By conducting the independent samples t-tests, the statistical difference is $p<0.05$. This research aims to create an innovative intruder behavior prediction that uses machine learning techniques to identify data as usual or invasive. While comparing the decision tree algorithm with the K-Nearest Neighbor algorithm, the final extracted results demonstrate that the KNN was outperformed.

Keywords: Innovative Intruder Behaviour, K-Nearest Neighbor, Decision Tree Algorithm, IDS, Statistical Analysis, Network security.

1. Introduction

The research aims to develop an innovative intruder behavior prediction model using KNN and DTA algorithms to actively scan the attacker behaviors and then efficiently classify the regular or anomaly activity. Through this model, we could defend the network against attacks and enrich safe work zone. The development of the network user behavior model should be more concerned with minimizing the false rate by detecting the positive activity [1]. There is a rapid increase in cyber security attacks, particularly over confidential data work environments [2]. The current scenario demands innovative and efficient Intrusion Detection systems (IDS) developments. The earlier mechanisms, such as encryption techniques and firewalls, are insufficient to

[1]Rajendran T, Corresponding Author, Department of Computer Science and Engineering Saveetha School of Engineering Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India, Pincode:602105.Email:rajendrant.sse@saveetha.com.

handle the current pressure on attacks [3]. IDS is not only operated on the network perimeter but also scans all layers' traffic. IDS is trained with machine learning models to intensively predict the behavior of the event by classifying it into normal or anomaly [4]. Detecting unknown attacks and misuse detection are innovative intrusion detection systems applications using machine learning algorithms [5]. Nearly 30,500 research articles were published and indexed related to IDS using machine learning models. In the study [6], Text categorization techniques were used for intrusion detection systems relying on the K-Nearest Neighbor Algorithm. In the study [7], the false alarm filtering using an IDS depends on the K-Neighbor classifier. In the study [8], the Feature selection approach for network intrusion detection systems based on the Decision tree and K-Neighbor Algorithm was cited four times. Study [9], Improving the Accuracy of Intrusion Detection System using a combination of machine learning algorithms, cited six times. The best study is a dependable network intrusion detection system using the K-Neighbor algorithm [10].

An intrusion detection system is most widely used for network security. Our literature survey shows that it has given fewer accurate results in detecting unknown attacks, and it will take more time for testing and training. We have proposed this work under the guidance of our department team, which has helped in this prediction algorithm to get an accurate result in work. The main advantage of machine learning algorithms is that they can detect unknown attacks. The goal is to develop a viable solution, an innovative intruder behavior prediction system capable of delivering precise results. The experiment results demonstrate that the proposed model will have a high degree of accuracy. Our theory holds on to reality since the significance is less than 0.05.

2. Materials and Methods

The research has been conducted in the CISCO Lab of SSE, SIMATS. This project will use two machine learning algorithms in two groups, Group1 is the KNN algorithm, and Group2 is the DTA algorithm. The sample size will be of two iterations used in this research. Iteration-1 for the Train Set and Iteration2 for the Test set. The method evaluated ten iterations of every algorithm with a sample size of 10 to check whether the data was an anomaly or normal. The G-Power test will be 80% used for machine learning algorithms [11].

The flow chart shows how the training process has been done while fitting the model and improving the performance of the model. The flowchart defines the steps of the training set for innovative intrusion detection systems. It consists of Feature extraction, Reduces Features, Data Preprocessing, Learning algorithms, and Model Evaluation, as given in Figure 1.

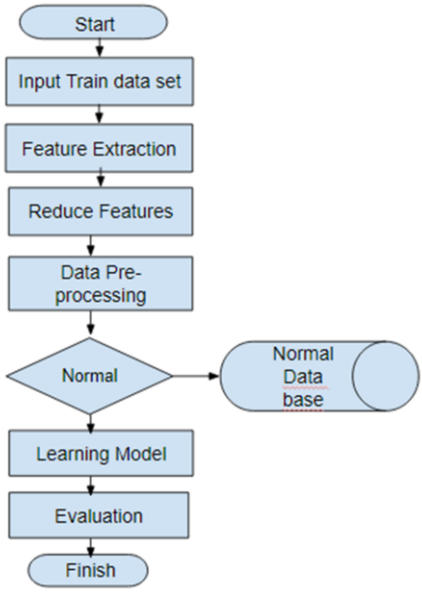

Figure 1. Flowchart for Training the Data Set.

2.1. Dataset Description

This study is based on the 'intrusion detection dataset,' made available and accessible through an open forum domain Kaggle. The dataset contains the Train dataset and the test dataset. The training dataset contains 25193 rows and 42 columns. The test dataset contains 19,864 rows and 41 columns. TCP, UDP, FTP, HTTP, and ICMP protocols are used in the dataset. The training dataset is used to develop the machine learning technique. The validation data has been used to assess the machine learning model's performance.

2.2. K-NearestNeighbor Algorithm

The K-Nearest Neighbor is a machine learning technique that can be used to overcome clustering issues. It determines the distance between the testing set and the source and makes a prediction based on that distance. K-Neighbor algorithm calculates the distance using equation (1) Euclidean Distance formula and also to find the k nearest neighbors using the calculated Euclidean distance.

$$d(p,q) = d(q,p) = \sqrt{(x-a)^2 + (y-b)^2 36} \qquad (1)$$

The KNN algorithm will be used for classification and regression, but it is more commonly used for clustering problems. KNN is a non-parametric approach, which means it makes no assumptions about the data it uses. The normal and anomaly behavior is predicted by finding the score for each process N over the test dataset T. If the average score is within the threshold behavior, then it is classified as normal, or else it is classified as an anomaly. The same fashion of prediction is carried out over all

available test datasets T. Finally, the accumulated score has been summarized to produce the accuracy score.

2.3. Decision Tree Algorithm

The Decision Tree Algorithm is a supervised machine learning model which can be used to solve problems in classification and regression. The decisions or tests are made based on the characteristics of the given dataset. The DTA model evaluates each child from the tree based on the test_attributes. The dtree model will level up the score if the test_attributes match. The iterations over the child will continue till its reaches the leaf node. Finally, the accuracy score of the DTA model will be calculated.

2.4. Experiment Setup

The popular Python IDE, Jupyter Notebook, is used to develop and implement machine learning models. The Intrusion detection dataset is collected. The dataset was partitioned into two different classes after loading the data. The two classes after the data partition are normal and attack. The data preprocessing has been done. The data can be divided into train data and test data. Machine learning algorithms can be implemented. Through the training, process algorithms can be built. Then the test dataset was fed into both the machine learning algorithms to get the accuracy. Finally, the statistical analysis packages have been used to analyze and visualize the various metrics in the study.

2.5. Statistical Analysis

The statistical software tool used for machine learning algorithms is the SPSS tool for statistical analysis. The statistical computations are performed using the SPSS tool, and the findings acquired for different testing sizes of significance are less than 0.05. The mean, standard deviation, and standard error statistical significance between the groups were determined using an independent sample t-test. A comparison of the two algorithms, KNN and DTA, was performed. The variables used in the next part of the trained data set are independent variables and dependent variables in the test set. The trained data is an independent variable, and the test data is the dependent variable.

3. Results

Compared to their accuracy rate, the KNN algorithm provided us with an accuracy of 99%, and the Decision Tree algorithm offered us an accuracy of 98%. The accuracy for every test size is different for both algorithms. The K-Neighbor algorithm has better accuracy than the decision tree algorithm. The algorithm's accuracy changes due to whatever is inside the test size, as given in Table 1.

Table 1. Accuracy of KNN and DTA based in different test sizes.

Test size	0.12	0.16	0.22	0.26
K-Neighbor	99.87	99.76	99.73	99.45
Decision Tree	98.93	98.84	98.76	98.65

From the obtained results, the accuracy is different for every test size. Comparing the accuracy of both the KNN and DTA gives the result of high accuracy for the proposed model. The mean value for K-Neighbor and Decision Tree algorithm is 99.51 and 95.42. Standard Deviation for K-Neighbor and Decision Tree algorithm is 0.294 and 0.235. The standard error mean of algorithms is 0.093 and 0.074as given in Table 2.

Table 2. Group Statistics the Mean and SD for KNN and DTA.

	KNN, DT	**N**	**Mean**	**Standard Deviation**	**Std.Error Mean**
Accuracy	KNN	10	99.51	.294	.093
	DT	10	95.42	.235	.074

The Mean Difference for both the algorithms is 4.093. The Standard Error Difference for both the algorithms is 0.119. The significant value is less than 0.05 in the independent sample test using the SPSS tool, as given in Table 3.

Table 3. Independent Sample T-test for KNN and DTA

		Sig.	**Mean Difference**	**Std.Error Difference**	**95% Confidence Lower**
Accuracy	Equal variances assumed	.000	4.093	.119	3.843
	Equal variances not assumed	.000	4.093	.119	3.842

The bar chart shows the accuracy of the K-Nearest method, which looks to generate the most consistent results with the slightest standard deviation. There is an optimal substantially different between the K-Neighbor algorithm and the Decision Tree algorithm was achieved, which is $p<0.05$ through the Independent sample test as given in Figure 2.

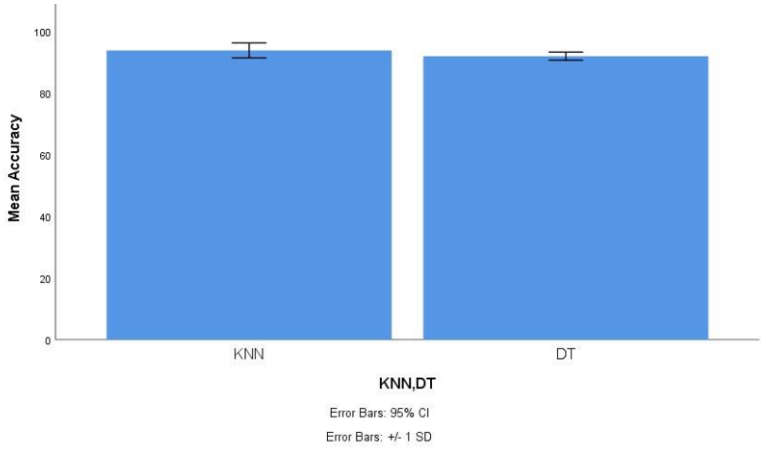

Figure 2. The bar chart represents the comparison of the mean accuracy of KNN and DTA.

4. Discussion

The accuracies of KNN and DTA are achieved at 99% and 98%. For every test size, the accuracy will be different for both algorithms. The K-Neighbor algorithm has better accuracy than the decision tree algorithm. The data is gathered through several iterations of the experiment to determine different accuracy rates for various test sizes. An independent sample t-test is performed with the SPSS application for the analytical findings. Comparing the accurateness of both the algorithms KNN and DTA gives the result of high accuracy for the proposed model. The mean value for K-Neighbor and Decision Tree algorithm is 99.51 and 95.42. Standard Deviation for K-Neighbor and Decision Tree algorithm is 0.294 and 0.235. The study [12] shows that the naive Bayes has better performance detecting accuracy, but according to [6], KNN algorithms have better performance. Based on a literature review, it has been established that the K-Neighbor method is more accurate than other algorithms.

The Mean Difference for both the algorithms is 4.093. The Standard Error Difference for both the algorithms is 0.119. The significant value is less than 0.05 for two algorithms through independent sample tests with the SPSS application. In previous findings, the mean error value of the Random Forest algorithm and KNN is 0.0682. The standard error means the difference is high in the K-Neighbor algorithm [11]. In this study [13], logistic regression has better accuracy than decision trees. In the study [14] SVM algorithm has better accuracy according to the KNN algorithm [15].

5. Conclusion

The results of our experiments encourage further research into the effects of different classifiers and exploration of other developments in the innovative intrusion detection system domain for providing network security. The accuracy of intrusion detection systems using the K-Neighbor algorithm and Decision Tree algorithm is proved that the K-Neighbor algorithm has higher accuracy than the Decision Tree. The limitation of

our experiment is that calculation gets very complex, mainly if many values are unknown. A small change in data leads to a significant change in structure. Future work and experiments would enable the model to be more robust as the current model in this research uses binary class classification. The comparison of algorithms proves the proposed model will get its accuracy.

References

[1] Li W, Yi P, Wu Y, et al. A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network. J Electr Comput Eng; 2014. Epub ahead of print 30 June 2014. DOI: 10.1155/2014/240217.

[2] Liu Z, Zhang X. Prediction and Analysis for Students' Marks Based on Decision Tree Algorithm. 2010 Third International Conference on Intelligent Networks and Intelligent Systems. Epub ahead of print 2010. DOI: 10.1109/icinis.2010.59.

[3] Pathan A-SK. The State of the Art in Intrusion Prevention and Detection. CRC Press, 2014.

[4] Vigna G, Jonsson E, Kruegel C. Recent Advances in Intrusion Detection: 6th International Symposium, RAID 2003, Pittsburgh, PA, USA, September 8-10, 2003, Proceedings. Springer, 2003.

[5] Bhattacharyya DK, Kalita JK. DDoS Attacks: Evolution, Detection, Prevention, Reaction, and Tolerance. CRC Press, 2016.

[6] Binbusayyis A, Vaiyapuri T. Comprehensive analysis and recommendation of feature evaluation measures for intrusion detection. Heliyon 2020; 6: e04262.

[7] Aburomman AA, Reaz MBI. A novel SVM-kNN-PSO ensemble method for intrusion detection system. Applied Soft Computing 2016; 38: 360–372.

[8] Zhao Z, Ge L, Zhang G. A novel DBN-LSSVM ensemble method for intrusion detection system. 2021 9th International Conference on Communications and Broadband Networking. Epub ahead of print 2021. DOI: 10.1145/3456415.3456431.

[9] Liang D, Liu Q, Zhao B, et al. A Clustering-SVM Ensemble Method for Intrusion Detection System. 2019 8th International Symposium on Next Generation Electronics (ISNE). Epub ahead of print 2019. DOI: 10.1109/isne.2019.8896514.

[10] Moukhafi M, El Yassini K, Bri S. A novel hybrid GA and SVM with PSO feature selection for intrusion detection system. International Journal of Advances in Scientific Research and Engineering 2018; 4: 129–134.

[11] Pathak A, Pathak S, SGSITS, et al. Study on Decision Tree and KNN Algorithm for Intrusion Detection System. International Journal of Engineering Research and; V9. Epub ahead of print 2020. DOI: 10.17577/ijertv9is050303.

[12] Bhattacharyya DK, Kalita JK. Network Anomaly Detection: A Machine Learning Perspective. CRC Press, 2013.

[13] Lagman AC. Predictive Decision Support System using Logistic Regression and Decision Tree Model Combination for Student Graduation Success Determination. Proceedings Journal of Interdisciplinary Research 2015; 2: 144–153.

[14] Khan L, Awad M, Thuraisingham B. A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB Journal 2007; 16: 507–521.

[15] Kambourakis G, Shabtai A, Kolias C, et al. Intrusion Detection and Prevention for Mobile Ecosystems. CRC Press, 2017.