International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-8, Issue-2 , February 2021 Regular Edition E-ISSN: 2349-7084

An Optimized KNN Model for Signature-Based Malware Detection

Tsehay Admassu Assegie[1*]

Department of Computer Science, Aksum Institute of Technology, Aksum University, Axum, Ethiopia
E-mail: tsehayadmassu2006@gmail.com

Available online at http://www.ijcert.org

Received: 05/02/2021, Revised: 12/03/2021, Accepted: 18/03/2021, Published 26/03/2021

Abstract: Malware is a computer program developed with the intent of disrupting, stealing, and compromising a computer system. In recent advances in technology and internet use, malware has become the major problem in computer society. In this research, an optimal K-nearest Neighbor (KNN) based malware detection and classification model is proposed. The proposed malware detection model is based on application programming interface (API) call sequence analysis and classification. The dataset is collected from an online Kaggle data repository which consists of 42,797 malicious application programming interface (API) call sequences and 1,079 non-malicious application programming interface (API) call sequences. The Nearest Neighbor (KNN) algorithm is applied to the dataset to create a model that detects malware. Finally, the accuracy of the proposed KNN based malware detection model is evaluated, and the result shows that the accuracy of 98.17% is achieved in the detection of malware using the model. The proposed model is significantly essential for detecting real-time intrusion on computer systems.

Keywords: Computer security, Intrusion detection, KNN, Malware detection, network security

1. Introduction

Malware is a computer program developed to disrupt the normal function of a system or a network. In recent years, due to widespread malware in the network, network security incidents are increasing day by day [1]. To mitigate the damage caused by the malware, different software programs such as antimalware are developed. The existing malware detection approaches are signature-based malware detection systems that cannot detect unknown malware [2]. The rules of application, programming interface (API) call sequence of malware, and their variants machine learning models are developed to overcome the current malware detection systems' problems to learn the new features based on the previous observation given during the training phase to the learning algorithm.

Classification refers to the mapping of observation in the data repository to the predefined class or group. Based on the particular observation features in the data repository, a classifier can unknown observation to a given class or group. Before training, the dataset is split into training and test set; then, the machine learning algorithms are trained on the training set and test on the test set. By employing different machine learning algorithms, a model that detects malware based on the previous association rule and characteristics of the algorithm can effectively detect unknown malware. This study is aimed at answering the following research questions:

1) How to develop a malware detection model with a KNN algorithm with an acceptable level of accuracy?
2) What is the accuracy of KNN on malware detection?
3) How can we find the optimum K value in KNN to optimize the efficiency of KNN on malware detection?

In upcoming sections, the study discusses the related works and summarizes the previous work. Section 3 discusses research methodology, and in section 4, the results and research findings are discussed. Finally,

 https://doi.org/10.22362/ijcert/2021/v8/i2/v8i206

section 5 discusses the conclusion and claims reached by the study.

2. Related Work

Machine learning algorithms and machine learning applications have become essential to signature-based malware detection [4-13]. In signature-based malware detection, supervised machine learning algorithms are trained using malware behaviors so that the algorithm can detect unknown activity as malicious or non-malicious. In this section, some of the research conducted by the researcher on solving and dealing with signature-based malware detection is discussed.

In [3], data mining-based API call sequence identification is proposed to identify an API call as malware or legitimate activity. The API provides the interface through which different programs interact with each other. The operating system APIs provide the interface via which a program interacts with the system. When a program performs some activity, such as writing or opening a file, the operating system API is used to perform the required operation to permit the process. In the same way, malware programs use system APIs to perform illegitimate functions or actions on the system. The authors trained KNN, Naïve Bayes, and Decision tree algorithm to API dataset, and each algorithm's performance is tested against malware detection accuracy. The result shows that the Naïve Bayes algorithm better understands malware detection compared to the KNN and decision tree algorithm.

In [4], KNN and Naïve Bayes-based hybrid approach is proposed for malware detection. The malware detection accuracy of the unique algorithm and hybrid model is compared, and the result shows that the hybrid system is better on malware detection. Based on the literature survey, we have a thorough understanding of state of the art in malware detection and noted that malware detection requires a highly accurate and better perfuming model for the malware to be detected with higher precision the problem to be minimized. Thus, this study focused on improving the existing work by designing and implementing an optimized KNN neighbor-based model for malware detection.

3. Methodology

The data repository used in this research is collected from an online kaggle malware dataset. Malicious activities and their critical API call sequence pattern are used to classify an API call sequence as negative behavior on-malicious behavior. The KNN algorithm is applied to the collected data repository to create the model for malware detection by classifying the API call sequences into malicious and non-malicious classes based on the class labels given to the algorithm as input during the training. The data repository consists of 43,876 observations. Each in the data repository has 3 attributes, namely, hash description, API call type, and class label. Each of the attributes of the observations is summarized in table 1.

3.1.K-nearest neighbor algorithm

The K-nearest neighbor model is a distance-based learning algorithm widely implemented for classification tasks with a lower dataset [5-6].

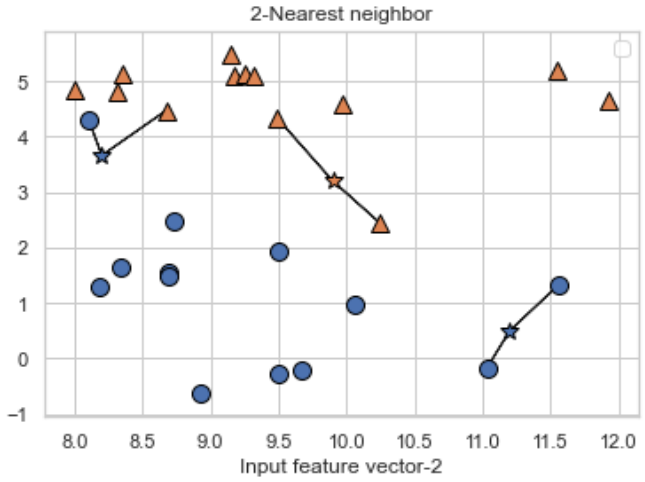

Fig. 1. 2-Nearest neighbor classifier

K-nearest neighbor works by calculating the distance between a data point and classes. A given data point is classified as malware or not malware class based on the distance between the data point and the classes; the closer distance is selected. As shown in figure 1, the stars indicate data points. The lines are the Euclidean distance between the data point and the classes, namely the red triangles indicating malware class and the blue circles show not malware class. The Euclidean distance measure used to define the distance between a given data point and class is defined as follows:

$$\text{Distance}\{(x0, y0), (x1, y1)\} = \sqrt{(x1 - x0)^2 + (y1 - y0)^2} \quad (1)$$

Where x0,x1,yo,y1, are data points.

Table 1: API dataset features

Attribute	Description
Hash description	An MD5 hash of the example Type: 32 bytes string
API call type	Integer (0-306)
Class label	Integer: 0 (non-malware) or 1 (Malware)

4. Results and Discussion

In this section, the performance of the proposed malware detection model is analyzed. The analysis of the model's performance is performed using the accuracy and classification errors for different values of K. the accuracy and classification error for the proposed model on malware

 https://doi.org/10.22362/ijcert/2021/v8/i2/v8i206

detection are discussed in section 4.1 and section 4.2, respectively.

4.1. Accuracy vs. k-value

The accuracy of the KNN algorithm depends on K. the malware detection error by the proposed model is shown in figure 1. As shown in figure 2, the error rate is high when higher K values are used. The highest error rate is at a K value of 39, and the malware detection error is lowest when the K value is 3.

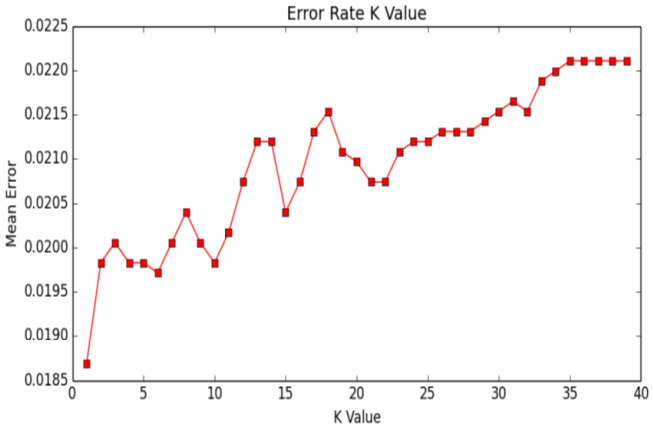

Fig. 2. Errors by the model for K values between 1 and 40

The error rate for the proposed KNN algorithm for different values of k between 1 and 40 is shown in figure 2. As shown in figure 2, the error rate is slightly increased when larger K values are used for training the proposed KNN model. Thus, with an increase in K value, the performance of the proposed model tends to decrease. The lower error rate indicates better malware detection accuracy. The maximum possible malware detection accuracy is achieved when a smaller value is used for training the model. The highest malware detection accuracy is achieved when K=3 is used in preparing the model. We have employed a grid search method for finding the optimal K value with for loop, testing the cross-validation accuracy against each value of K from K=1 to K=40, as shown in figure 2.

4.2. Classification error

We have experimented on the proposed model for testing overfitting. The model overfitting is tested with a learning curve—overfitting results when a model fits nicely on the training set and cannot generalize on the test set. The learning curve for the proposed KNN based malware detection model is demonstrated in figure 3. As shown in figure 3, the test error is above the training error, indicating that the training error is lower than the test error. Thus, the model performs with lower error in training or performs higher than the test, indicating that the model is not suffering from overfitting. Moreover, we observe from figure 3 that the test error is below 0.0235, which shows that the model performs with 99.97% accuracy on testing.

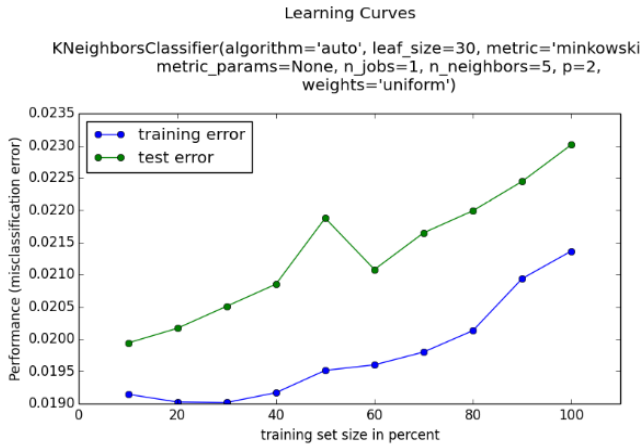

Fig. 3. The learning curve of the proposed model

4.3. Discussions

The training and test accuracy of the proposed model on neighbors' different values is shown in figure 4. As shown in figure43, the test and training accuracy of the model on malware detection has the lowest performance of detection accuracy when K=2, and the highest efficiency on malware detection is achieved when the value of K=3. As demonstrated in figure 4, the training and test accuracy of the malware detection model gets lower for the higher K values. There is a slight decrease in both the training and test accuracy on malware detection for higher values of K. in contrast to the learning curve, which shows error rate for both training and validation; the accuracy-test shows the performance of classification accuracy for the proposed model. We observe from figure 4 that the classification accuracy drops with an increase in the values of K, particularly when the k value increases beyond 3. Initially, the accuracy looks highest due to overfitting, but the latter gets higher at K=3.

Fig. 4. K value and the training and test accuracy

 https://doi.org/10.22362/ijcert/2021/v8/i2/v8i206

4.4. Comparative study

This section compares the proposed model with exiting work. The results obtained in previous work is compared with the result achieved in this work. For comparison, we have employed accuracy as a performance measure or comparison criteria. Table 2 demonstrates the result achieved in this study and the development obtained by other researchers in the existing work of literature.

Table 2: Comparative study with existing work

Existing and proposed work	Algorithm employed	Accuracy in %
[10]	KNN	97
[12]	SVM	92
[13]	SVM	87
proposed work	KNN	98.17

5. Conclusion and Future Scope

This study proposed a signature-based malware detection model based on application programming interface (API) call sequence behavior classification and analysis by employing K-Nearest Neighbour (KNN) algorithm on a dataset collected from the online Kaggle data repository. The accuracy of the proposed malware detection model is tested, and the result shows that the model has 98.17% performance on malware detection. Moreover, the grid search method is employed to find the optimal k value for achieving higher classification accuracy. The K value that provides the highest maximum possible accuracy on malware detection for KNN is analyzed. Overall, the proposed model is efficient when the K value of 3 is used when the KNN model is trained on the malware dataset.

References

[1] Yu-Lun Wan, Jen-Chun Chang, Rong-Jaye Chen, Shiuh-Jeng Wang, Feature-Selection-Based Ransomware Detection with Machine Learning of Data Analysis, IEEE, International Conference on Computer and Communication Systems, 2018.

[2] Aziz Mohaisen, Omar Alrawi, Jeman Park, Network-based Analysis and Classification of Malware using Behavioral Artifacts Ordering, Association for Computing Machinery, 2019.

[3] Om Prakash Samantray, Satya Narayan Tripathy, Susanta Kumar Das, A Data Mining Based Malware Detection Model using Distinct API Call Sequences, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7, May 2019.

[4] Niranjan A, Akshobhya KM, P Deepa Shenoy, Venugopal K R, Ensemble of kNN, Naïve Bayes Kernel and ID3 for Efficient Botnet Classification using Stacking, IEEE, 2018.

[5] Assegie, T.A, Nair, P.S, Comparative Study On Methods Used In Prevention And Detection Against Adress Resolution Protocol Spoofing Attack, Journal of Theoretical and Applied Information Technology 31st August 2019.

[6] Assegie, T.A, A Predictive Model For Improving Employee Attrition Rate With K-Nearest Neighbor Classifier, International Journal of Research and Reviews in Applied Sciences,, Jan-Mar. 2021.

[7] Assegie, T.A, An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.

[8] Maryam Nisa, Jamal Hussain Shah, Shansa Kanwal, Mudassar Raza, Muhammad Attique Khan, Robertas Damaševi˘cius, Tomas Blažauskas, Hybrid Malware Classification Method Using Segmentation-Based Fractal Texture Analysis and Deep Convolution Neural Network Features, Applied Sciences, 2020.

[9] Assegie, T.A, Nair, P.S, Comparative Study On Methods Used In Prevention And Detection Against Adress Resolution Protocol Spoofing Attack, Journal of Theoretical and Applied Information Technology 31st August 2019.

[10] Sunoh Choi, Combined KNN Classification and Hierarchical Similarity Hash for Fast Malware Detection, Applied science, 2020.

[11] P HarshaLatha, R Mohanasundaram, Classification of Malware Detection Using Machine Learning Algorithms: A Survey, International Journal of Scientific & Technology Research Volume 9, Issue 02, February 2020.

[12] Usha Narra, Clustering versus SVM for Malware Detection, A Project Presented to The Faculty of the Department of Computer Science San Jose State University In Partial Fulfilment of the Requirements for the Degree Master of Science, 2015.

