An ensemble based stacking approach for Network Intrusion Detection System

Vikas Kumar[1], Ritika[2]

[1]PG Student, Sharda University ,Greater Noida, India

[2]PG Student, Guru Jambheshwar University Hisar, India

Abstract

To detect malicious activities from the network various classification techniques are designed. The existing technique is SVM which is applied to classify data into malicious and normal classes. In this research work, stacking approach is designed for the classification of network traffic. The base classifier which is used in this research work is SVM classifier and meta-classifier is KNN classifier. The performance of existing and proposed approach is tested in terms of accuracy and execution time. It is analyzed that proposed technique performs well as compared to existing SVM classifier for the network traffic classification SVM classifier given 77% accuracy with in 0.02 second and KNN classifier given 79% accuracy within 0.05 second. We can analyze that the accuracy of KNN is better than SVM but in case of execution time the SVM is better. Thus we applied an stacking approach (SVM+KNN) and we get 87.50% accuracy with in 0.04 second. **Keywords:**

SVM, KNN, Stacking, KDD

Introduction

The interconnectivity between the computers that provide a single network with lots of advantages is termed as the network. They are connected with other in order to provide the communication by which information can be exchange easily. Hence, the communication is facilitated by these connected computers in order to perform assigned task. The scenario in which numerous computers are gathered and connected with each other to exchange information and provide facilities to other resources is called a network [1]. The information such as data communication is provided with the help of networking technology. There are software and hardware types of resources present within the sharing devices. A communication platform through which the interaction and transmission of information is provided amongst users is known as a social network. Today, social networks are practically included in each domain with respect to one way or another. The services involved in education, business, excitement etc. all these applications include social networks. For example, there are several such business as well that advertise their brands and products on the social networking platforms such that the products that they wish to sell can gain popularity [2]. The popularity has increased due to the increase in positive feedbacks given online and the flexibility of utilizing these products available. The patterns are carried on and obeyed by the malicious users in very unique manner such that they can become the part of secure networks. For example, as per the association rule, a genuine user sends the messages to normal user on regular basis. A noise within the data is a random error or variance that is caused in a variable and the examining of data does not affect it much. For example, the individual's purchasing activities can be taken as criteria to detect the credit card faults using the behavior [3]. This study utilizes the NSL-KDD data set that includes 42 attributes. By eliminating duplicate instances such that the biased classification results 6-9 can be removed from data set, the KDD'99 data set 4, 5 is enhanced. There are several versions of data set available amongst which the utilization of only 20% of training data is done. In the form of KDDTrain+_20Percent, this data is represented and there are 25192 instances included in it. With the name of KDDTest+, the test data set is recognized and there are around 22544 instances included within it. With variation of number of instances, there are several configurations of this data set available. However, 42 numbers of attributes are available here [4]. The 'class' attribute is the attribute that is labeled 42 in the data set through which it can be distinguished whether there is a normal connection instance or an attack present in the given instance. Table below provides the description of

KDD data set attributes along with the labeled classes. In K-Nearest Neighbor, a patter x is classified by assigning class label to it that is most frequently represented among its k nearest patterns. The class with minimum average distance is used to assign a test pattern that shows that this method is sensitive to distance function. The Euclidean distance metric is employed for getting minimum average distance. All features are normalized into same range this is the main requirement of this metric approach. A nonparametric classifier that generates better performance for k number of optimal values is known as k-nearest neighbor classifier [5]. Another feed forward artificial neural network based classifier is known as multi-layer perceptron (MLP). In order to improve the performance of classification, initially the simplification is done using single hidden layer. Further, two hidden layers are moved in this process. Numbers of hidden units across the number of trials are selected for every data set. Numbers of hidden neurons across the number of trails are also identified by conducting certain experiments. The rule of thumb is applied here through which rough total of number of weights is calculated to be n/10 for which n is the total number of training points [6]. For training neural networks back propagation algorithm is applied here. As per the multi-layer perceptron, Bayesian theory provides optimal discriminant function which is approximated by this algorithm. A predictive model is present within the SVM classification approach. In the form of input, the data is provided and output is achieved in the form of classified data is separate two categories. A model is implemented using SVM training algorithm for text that includes each training sample that belongs to any of the two classes.

Literature Review

Amrita, et.al (2018) proposed a novel approach for network intrusion detection system in this paper. The proposed system is named as Hybrid Feature Selection Approach {Heterogeneous Ensemble of Intelligent Classifiers (HyFSA-HEIC) and within the traffic that is coming as input, the anomaly is classifier here [7]. The HyFSA and HEIC systems are integrated hierarchically here. The optimal numbers of features are achieved here using HyFSA and with the help of generated optimal features, HEIC is generated further. As per the simulation results it is seen that in comparison to other existing ensemble and single classifier methods, the proposed system provides better results.

Bayu Adhi Tama et.al (2017) presented that to protect the computer systems, an intrusion detection system (IDS) is very important. In order to perform IDS task, there are several approaches introduced such as data mining, machine learning and so on. The performance of multiple classifiers is enhanced in comparison to the performance of single classifiers as per the analysis of approaches [8]. The five popularly known ensemble approaches which are bagging, stacking, boosting, rotation forest, and voting are studied comparatively in this paper. In comparison to bagging, rotation forest, and voting mechanism, the boosting and stacking approaches perform better as per the simulation results.

M. Paz Sesmero, et.al (2015) presented a study in which the stacking and its variants are reviewed along with various applications in which they are used. There are various contradictory results made here and stacking configuration is not optimally provided here. There is a need of prior information such that the parameters can be configured which is contradictory to the Wolpert's statement [9]. The domain independency is provided by stacking system. In comparison to other ensemble methods, the stacking approach is applied very less within the real-world applications. However, the ensembles are generated due to the huge diversity of these data within stacking.

Necati DEMIR, et.al (2018) presented that the model generation is enhanced such that stacking method is improved. In the form of combiner method various classification algorithms are utilized. The subsets of dataset are utilized along with randomly chosen features for generating the model [10]. For the combiner however, not all the models can be used as input. Within model selection, several metrics are utilized and for the combiner method, however, only the chosen models are utilized. For achieving highly accurate results, several comparisons are made with pure machine learning approaches in the experiments conducted in this paper using stacking technique. Within experiments results, the highest detection rate is achieved in comparison to other studies for the user-to-root attacks.

Nanak Chand, et.al (2016) presented in this paper the popularly known research area which is anomaly detection that uses machine learning approaches. The researchers working in different regions have been highly interested in the machine learning algorithms due to their adaptability and learning power [11]. Here, the intrusions are detected in the network by applying machine learning algorithm using SVM classifier. 9 different machine learning algorithms were used to evaluate the performance of SVM and its stacking. The performances of various classifiers were analyzed using NSL-KDD99 data set. As per the simulation results it is seen that the usage of random forest and SVM has resulted in increasing the efficiency of proposed approach in comparison to other existing approaches.

Mazhar Rathore, et.al, (2016), presented that detection of the KDD calls has been done by the telecommunication authorities and Internet service providers. This detection is performed to block the illegal commercial KDD or to give importance to the users using paid KDD calls. This technique has the complex security and mechanism of tunneling due to which most of the KDD detection techniques are not providing effective and efficient results [12]. As per performed experiments, it is demonstrated that proposed technique has better performance as compared to other. 97.54% TP and .00015% FP was provided by this technique. Hence, detection of the KDD calls in the fast growing environment it is considered as the optimal choice for the authorities.

Research Methodology

In this methodology, three steps have been utilized for the classification of network traffic. K-mean clustering was applied in the first step, which clustered the similar and dissimilar type of data. The removal of the redundancy and missing values, lead to refining of the dataset which is considered for input. In the second step, they implemented the technique for the calculation of arithmetic mean of the whole dataset. It is considered as the central point of the dataset. The central point is calculated using the Euclidian distance by which the points are differentiated. The clustering of the similar points is done in one cluster and others in second. They applied the SVM classifier in order to classify the data into two classes which is the last step of classification technique. They implemented the technique of KNN classifier in order to improve the performance of the previously developed methods. This technique clusters the uncluttered points so that classification accuracy can be increased. The Euclidian distance is calculated using the KNN classifier in which they classify nearest neighbor. In this clusters having same distance are clustered in one class and rest in other.

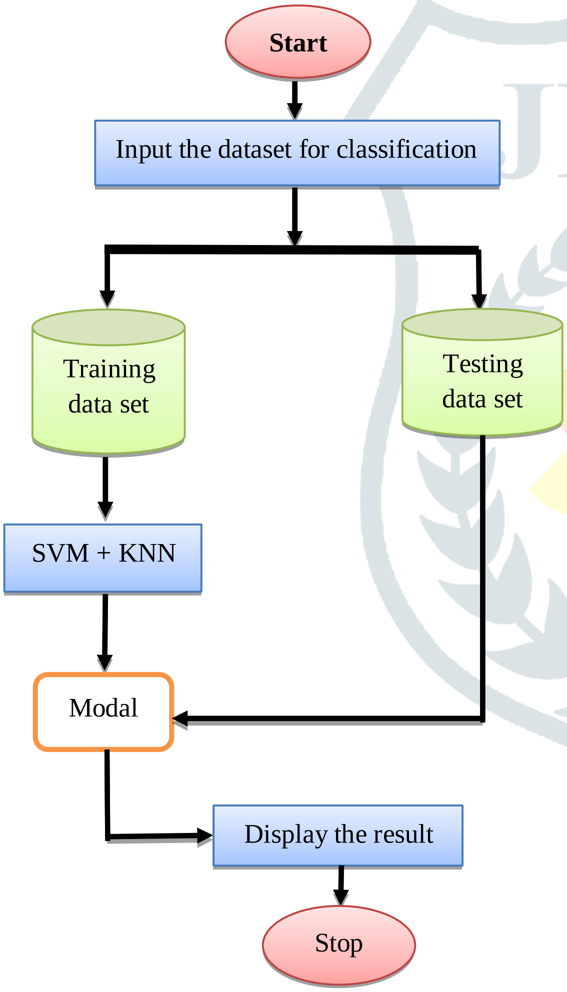

Figure 1: Proposed Flowchart

Following are the various steps which are following in the flowchart:

1. Input Dataset: - In the first step, the KDD dataset is given as input which is used to classify network traffic into certain classes

2. Selection of base classifier: - In the second step, the SVM classifier is applied as the base classifier to classify network traffic

3. Selection of Meta classifier: - In the third step, the KNN classifier is applied as the Meta classifier for the classification of network traffic

4. Evaluate: - In the last step, the models are evaluated in terms of accuracy and execution time

Experimental Results

The proposed results are implemented in Python and the results are evaluated with respect to accuracy and execution time to make comparisons amongst proposed and existing approaches.

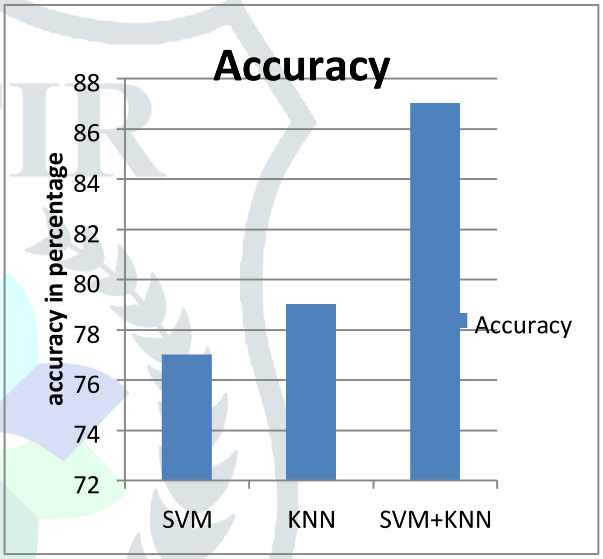

Fig 2: Accuracy Comparison

As shown in figure 2, comparisons are made amongst the algorithm that used SVM classifier, the algorithm that used KNN classifier and the proposed algorithm in terms of accuracy. As per the results it is seen that there is enhancement in accuracy of SVM+KNN classifier.

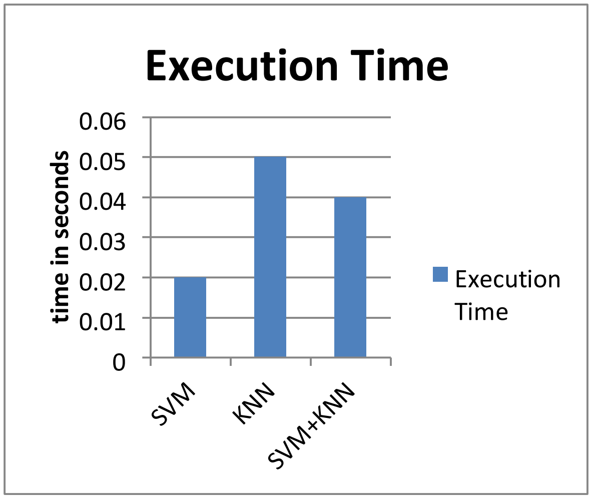

Fig 3: Execution Time

As shown in figure 3, the execution times of proposed and existing approaches are compared which show that less time is consumed by SVM+KNN classifiers.

Conclusion

In the machine learning we carried out a comparative study of well know classifier ensemble technique such as a staked generalization. We use SVM as a base classifier and KNN as a meta classifier. The aim of ensemble stacking approach is to combine these two classifier in an intelligent way that the accuracy of ensemble modal is better than base classifier.

We have measured performance of SVM, KNN and stacked of both classifier using NSL_KDD99 dataset. We analyze that the result of stacked modal is better than base classifier that is approximate 88% but there is time complexity is greater than the base classifier.

In future, we can do more better work on that proposed modal for reduce the time complexity and also can be determine the different types of attack ratio in network.

References

[1] Elsayed M. Elshamy, Aziza I. Hussein, Hesham F. A. Hamed, Hamdy M. Kelash, M. A. Abdelghany, “Secure KDD System Based on Biometric Voice Authentication and Nested Digital Cryptosystem using Chaotic Baker’s map and Arnold's Cat Map Encryption”, IEEE, 2017

[2] Sarwar Khan, NoumanSadiq, “Design and Configuration of KDD based PBX using Asterisk server and OPNET platform”, IEEE, 2017

[3] Sudipta Dey1, Tamal Chakraborty2 , ItiSahaMisra, “Sub-band based CAC Scheme using Adaptive Codec Switching for improved Capacity and GoS of Cognitive KDD Users”, 2017 4th International Conference on Signal Processing, Communications and Networking (ICSCN - 2017), March 16 – 18, 2017

[4] AboagelaDogman, Reza Saatchi, “Multimedia traffic quality of service management using statistical and artificial intelligence techniques”, The Institution of Engineering and Technology 2014, vol. 8, pp. 367–377, 2014.

[5] Eko Ramadhan, Ahmad Firdausi,3SetiyoBudiyanto, “Design and Analysis QoSKDD using Routing Border Gateway Protocol (BGP)”, IEEE, 2017

[6] Wonjung Kim, Taewon Song, Taeyoon Kim, Hyunhee Park, and Sangheon Pack, “KDD Capacity Analysis in Full Duplex WLANs”, IEEE, 2015

[7] Amrita, Kiran Kumar Ravulakollu, “A Hybrid Intrusion Detection System: Integrating Hybrid Feature Selection Approach with Heterogeneous Ensemble of Intelligent Classifiers”, International Journal of Network Security, Vol.20, No.1, PP.41-55, Jan. 2018

[8] Bayu Adhi Tama and Kyung-Hyune Rhee, “Performance evaluation of intrusion detection system using classifier ensembles”, Int. J. Internet Protocol Technology, Vol. 10, No. 1, 2017

[9] M. Paz Sesmero, Agapito I. Ledezma and Araceli Sanchis, “Generating ensembles of heterogeneous classifiers using Stacked Generalization”, WIREs Data Mining Knowl Discov 2015, 5:21–34

[10] Necati DEMIR, Gokhan DALKILIC, “Modified stacking ensemble approach to detect network intrusion”, 2018, Turkish Journal of Electrical Engineering & Computer Sciences, 26: 418-433

[11] Nanak Chand, Preeti Mishra, C. Rama Krishna, Emmanuel Shubhakar Pilli and Mahesh Chandra Govil, “A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection”, 2016, IEEE

[12] M. Mazhar, U. Rathore, “Threshold-based generic scheme for encrypted and tunneled Voice Flows Detection over IP Networks”, Journal of King Saud University Computer and Information Sciences, vol. 27, pp. 305–314, 2015.