NETWORK ANOMALIES CLASSIFICATION USING BIASED CLASSIFIER WITH KERNEL-BASED STRATEGIES.

Prachi D. Waghmare[1], Prof. S. G. Shikalpure[2]
[1,2]Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, (India)
(E-mail: prachiwaghmare08gmail.com, shikalpure@gmail.com)

***Abstract*—** Protecting the structure of network, system and other supplementary internet devices from attacks of unauthorised users or any malicious activity appears to be a significant task to be achieved in today's world. The society is all regarding the internet, all our day to day activity is through the internet from bank transactions to e-commerce purchase. The intrusion detection system (IDS) interprets an essential function in determining attacks (or intrusions). IDS generate false-alarms that are non flexible or rigid in the environment whenever anomalies detected within the network. In this paper, we use the kernel methods or kernel-based strategies, feature selection metrics and biased classifier by comparing it with trained threshold values to obtain high precision and diminish the degree of false-alarms generated.

***Keywords*—** *Intrusion detection system (IDS), datasets, Kernel methods or Kernel-based strategy, feature selection metrics, biased classifier.*

I. INTRODUCTION

In today's world with the advent of the internet, access to areas such as finance, medical, entertainment and many more have become a much easier task to be achieved. Protecting the system from any malicious activity over the internet becomes decisive. Over the long run, Cyber-security has been affected and are facing different challenges from various activities or attacks from the intruders. Hence to protect our data over the internet, prior knowledge about such activities is must so that legal action is to be taken against any unwanted threat. Firewall and Anti-virus software which secures a system are trivial and could not handle the uninterrupted network under threat of malignant activity or breaching of data. The need for the intelligent and efficient system comes into picture wherein the system learns and detects such intrusions without affecting the day to day activities. An intrusion detection system (IDS) is one such that helps predict malicious activity and generates false-alarms to any such intrusions [1].

The pioneer of the intrusion detection system (IDS) traced back in 1980 acquainted by Anderson for monitoring network activity and protecting a system from various types of attacks internal or external [2]. The IDS system based on two detection methods misuse or signature-based intrusion detection system and anomaly-based intrusion detection system. An anomaly-based system is a paradigm in which the rules are pre-defined during the systems network activity. If the rules match the pre-defined activities it is classified as a normal activity and others classified as an abnormal activity. The occurrence of abnormal activity generates several false-alarms. A signature-based system consists of a database with already known types of attacks. These known types are collected from different honey spot areas over the period and stored in the form of a database. A signature-based detection system cannot detect new types of intrusion activity occurring in a system [3].

Different datasets are to consider while developing an intrusion detection system. The commonly used datasets which are available freely are KDDCup'99[1], NSL-KDD[2], Kyoto2006+[3] and ISCX[4]. These datasets are useful for implementing a better IDS system to diminish the rate of false alarms and are collected from different timeline which cover many new attack types [4].

Kernel-based strategies in addition to the kernel classifiers possess a greater inclination towards statistics and mathematical concepts adopted in implementing machine-learning approaches to procure efficient outputs. These Kernel-based strategies are used in intrusion detection (IDS) to help estimate the significance concerning a particular feature amidst datasets. A key component in the kernel method is the weight vector which estimates the feature-space selection using a positive definite kernel function. With the significance of weight vector or feature-space selection, the Kernel-based strategies focus on the feature mapping using the equation (1) [5].

$$K\ (Xi,\ Xj) = \phi(x).\ \phi(y)\ldots\ldots\ (1)$$

To increase the performance of the IDS there are many techniques one such is the feature selection metrics. The feature selection metrics are two types filter-based and wrapper-based. The most popular are filter based; these are easy to understand and use the ranking search method. The wrapper-based feature selection method uses different learning algorithms which are more complex. We are using mutual

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[2] https://www.unb.ca/cic/datasets/nsl.html
[3] http://www.takakura.com/Kyoto_data/
[4] https://www.unb.ca/cic/datasets/ids.html

information, information gain and chi-square for feature-selection metrics. These feature selection metrics techniques use the score obtained by the kernel methods to obtain which feature or attribute will be useful for better results to obtain [6].

A threshold value obtained during the training process is a deciding factor for the classifier whether the newly arrived network packets are normal-type or anomaly-type. Threshold values estimated for every dataset with different values which are adaptive in nature [7]. An IDS system should distinctly classify whether any activity over the network is normal-type or anomaly-type. Different classifiers are used in machine learning, one such is the biased classifier which is used in kernel clustering to address the problem of data inhomogeneity. In IDS, the classifier divides the set of incoming packets or network contents as normal-type or anomaly-type [8].

The paper incorporates section 2 as related work, section 3 of proposed architecture, section 4 is implementation, section 5 focus on experimental analysis and obtained results and section 6 is the conclusion and future scope.

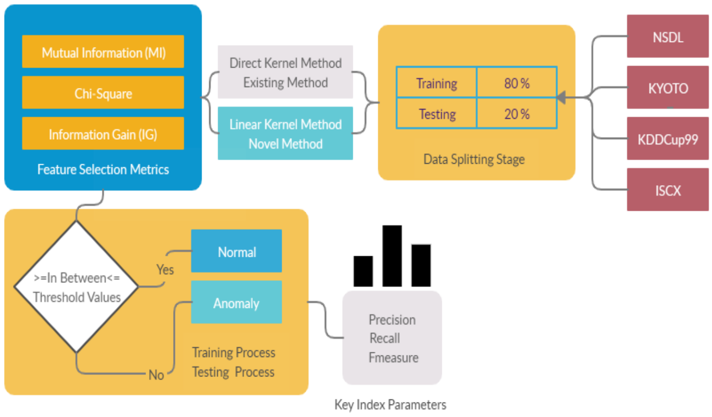
II. Related work

Following are a few of the works discussed on intrusion detection systems based on different classification techniques and feature extraction.

Aumreesh Ku. Saxena, Dr. Sitesh Sinha, Dr. Piyush Shukla, in 2017 have made a general study of intrusion detection system with survey of agent-based intrusion detection system. This review paper guides through the types of intrusion detection system with advantages and disadvantages [9]. In 2019, Kazi, Billal and Md Mahbubur have developed a novel supervised learning system. The idea is to combine the most prominent features to be selected using wrapper-based feature selection along with support vector machine classifiers. The results are combined with artificial neural network (ANN) methodology to give output as classifying the network traffic as normal-type or attack-type [10]. Back then in 2015, Xiaoyan, Yanbin and Yanxia had studied the relevance and performance of kernel functions in a grid search of support vector regression. Support vector machines are considered to be a core part of kernel functions. They tried combining support vector and regression classification to give better outcomes for the energy efficiency dataset [11]. In 2014, Ambusaidi, Xiangjian, Zhiyuan, Priyadarsi, Liang and Upasana proposed a method of feature-selection for IDS. In thi they have implemented a hybrid feature selection method combining both the filter and wrapper-based techniques to be applied on the IDS [12]. An IDS technique with a genetic algorithm, Kohonen map (SOM) to identify network anomaly by Anil S. and Remya R. in 2013. The implementation uses genetic algorithm and SOM as feature and information enhancement on the given datasets, where genetic algorithm is used for anomaly detection and SOM is used for feature extraction [13]. To diminish the false-positive rate in addition to the given dataset Kathleen Goeschel in 2016 using data-mining approaches in combination with Decision tree and Naïve Bayes algorithms had put forward the idea of a new IDS [14]. A study on SVM based linear and non-linear pattern classification with different kernels by Sourish Ghosh, Anasuya Dasgupta, Aleena Swetapadma in 2019. This survey indicates the various concepts of support vector machines, some of its real-life applications and future aspects of SVM [15].

III. Proposed System

There have been many intrusion detection systems which prove to be a helping hand in finding the intruders. Many of them guarantee greater accuracy to protect the systems but, with the advent of the new attack type, the system fails to guarantee the same accuracy. For our proposed system we thus tried to cover the newly available set of the dataset along with the previous commonly used dataset. In the proposed system architecture, principal sections are the Kernel-based strategies, feature selection metrics and the anomaly detection using threshold values. Kernel-based strategies focus on calculating the weights of each attribute mentioned in the dataset. The task of the kernel is mainly to focus on ranking weight values. Further, each score or weight values of features or attributes estimated using the feature-selection metrics. The feature-selection metrics calculate based on mutual information, chi-square and information gain. These metrics are the deciding factors to which features must further approve for anomaly detection. The scores compared with a threshold value in the classification module. The threshold value is regarded as a deciding factor so that the system detects whether the occurred packet is a normal packet or anomalous packet. In addition to the classification, the point that needs more focus is the utilisation of memory. Thus, classifying the dataset in such a manner becomes quite convenient and to some extent decreases the memory constraints which are generally faced by intrusion detection systems.



IV. Implementation

The implementation part includes the details of the proposed system. This section puts forward the relevance of every module and its specification.

A) **Dataset Description-** While studying different intrusion detection system (IDS) and machine-learning techniques applied on them, the commonly observed datasets were KDDCup'99, NSL-KDD, Kyoto 2006+ and ISCX dataset each having varying features or attributes and numerous records. These datasets are available openly and are endured within various researches while coping network-based detection. KDDCup’99 dataset is considered to be a landmark dataset whenever any IDS system gets trained. ISCX a new dataset created mainly to deal with the emerging new types of attack. The feature-selection method is determined using the score value or the threshold-value of each feature present within the dataset is analysed using the kernels and feature-selection metrics.

B) Data Splitting- In data splitting, to avoid overfitting we divide our dataset as a training data and testing data. Splitting proportion in the dataset decided on assuming the size and category of data present. In a generalised form if a dataset incorporates more than 10,00,000 then, it is proportioned as 90:10, if the data is less then it is either 60:40 or 50:50. We have estimated here 80:20 data splitting proportion assuming the given dataset size. Training-set should not imply inconsistency or the model will be biased to other values present apart from those tested set. Scientifically it is proven that when data trained in greater quantity as compared to the testing quantity. It implies that the system proposed gives better output results.

C) System Learning and feature-selection metrics- For system learning we are using kernel-based strategies, feature selection metrics and threshold values. Kernel-based strategies are a statistical expression that helps us find scores or weight of the features which is the similarity measure between the two record set. The scores are used to obtain the best threshold values to decrease the error rate values of the attributes. Every time the system generates different threshold values till, we compare all the attributes within the dataset. After training whenever we obtain a new packet on the network firstly the kernel calculates the weight or score of the occurred packet. Then the score value goes through the relevant feature or attributes from the datasets for further classification and evaluation. For improving the performance evaluation of the classification feature selection metrics are applied that ensures a better outcome of the classifier.

D) Classification- In classification part, we segregate the records in two distinct class type one is normal and another anomaly (i.e. Benign, DOS, u2r, r2l, Prob). If the value of the threshold-value obtained from the kernel ranges from >= threshold <= and is classified as either normal-type or attack-type. The boundary or threshold-values obtained from the weights or scores of the features of the dataset. Value reckoned from the kernel using the weights and scores of them from the dataset. These values mapped into the higher dimensional space later results are in the form of the matrix. From the matrix, we obtain key indexing parameters of the system.

V. Experimental Results

Our idea intends to perform experiments, examine particular execution regarding the developed method and divide the given input collection from a dataset that whenever a new packet occurs over the network it is classified as 'Normal' or 'Anomaly'. We have applied a biased classifier which divides the classes. During our experimentation, we have considered here four different datasets from KDDCup'99 dataset comprising the oldest amidst any to ISCX a newly available one. Our results are assuming correctness and each low false-positive rate. These factors help us analyse whether a system implies attainability through each incoming data or not. All individual training and testing have been done considering the splitting about records as 80%-20%. Here the training part is 80% whereas the testing part is 20% respectively.

Whenever we talk about intrusion detection systems, the extensively related dataset is KDDCup'99. That denotes some subset from this DARPA'98 application which comprises an unusual simulation about a specific virtual network. It contains around 4,898,431 records, including 41 distinct characteristics. The dataset incorporates five classes, one normal-class and four attack-classes which imply DoS (Denial of Service attack), U2R (User to root attack), Probe (Probing attack) and R2L (Remote to local). Down about these total records, we examine 55000 as training reports including 22000 as testing reports. Individual deficiencies amidst this KDDCup'99 is it fails to reflect real-time traffic also it cannot represent new-flanged network-traffic since formed or fabricated for simulation in extension through the redundancy of data. To work out amidst particular problems faced by KDDCup'99 revamped variants named NSL-KDD were created with features intact identical as that of KDDCup'99 but with additional numeric-value till some end. The numeric-value denotes essentially some characteristic that classifies a dataset as either a normal-type or anomaly-type. But concerning this system, there exist no constraining class-labels to occur a numeric-value hence discarded it. In NSL-KDD we took around 40000 records as training and 12000 as testing.

The next in line denotes this Kyoto2006+ dataset built on three years (from 2006 to 2009) of real-time traffic data consolidated through Kyoto University from different honeypots. Kyoto2006+ contains 24 peculiarities of which 14 are the same as KDDCup'99 whereas additional features like an enhanced characteristic investigation of network-traffic including traversing new irregularities present inside convolution. This dataset comprises 784,000 21-dimensional records, where 388,632 attack-type credentials and 395,368 normal-type records. Concerning a particular system, we took 65000 as training-records and 54555 as testing-records.

One novel and realistic data generated during 2011, especially for identifying intrusion exposure and cyber-security testing scheme comprises ISCX (Information Security Centre of Excellence). It denotes some synthesised dataset considered to be one of the benchmarks against specific prevention of malware and security scrutiny. Ideally, a dataset not compelled to indicate unplanned properties, both network-wise and traffic-wise it includes both normal-type and anomalous-type traffic. The total feature count is 11 with the last feature depicting the attack-type. We have considered training-record of 55000 and testing-record of 44500.

For classification we have considered different threshold values which are estimated to be the deciding factor between the obtained score or weights from the kernel methods applied. Following table 1 denotes the threshold values obtained during the training process of our system. These threshold values are compared with every feature or attribute within the dataset and are compared to further get classified as normal or anomaly.

TABLE 1

Dataset	Threshold for rate	Threshold for count	Threshold for Positive to Negative values
KDDCup’99	0.99	0.99	0.50
NSL-KDD	0.90	0.90	0.60
Kyoto2006+	0.78	0.78	0.23
ISCX	0.22	0.89	0.27

With the obtained threshold from the training-set whenever we test the dataset with our trained model, we obtain different precision and f-measure values. The system evaluation is based on these values. An ideal intrusion detection system believed to have 100 percent of accuracy and 0 per cent of the False-positive rate. This situation symbolises the system will detect all the possible attacks without any mis sorting or error. Thus, for a system evaluation we will consider accuracy and FRP. Accuracy consists of the correctly indexed testing instances from the total number of records trained from each dataset. Accuracy also implies the success rate of the system developed. The formula for accuracy is-

$$\text{Accuracy}= (TN+TP)/(TP+FP+TN+FN)$$

The accuracy of KDDCup’99 is maximum in contrast with all the dataset. Figure 2. shows the comparison values in the accuracy of each dataset used to develop the system. For ISCX dataset the accuracy is comparatively low, but nearer to the highest accuracy procured.

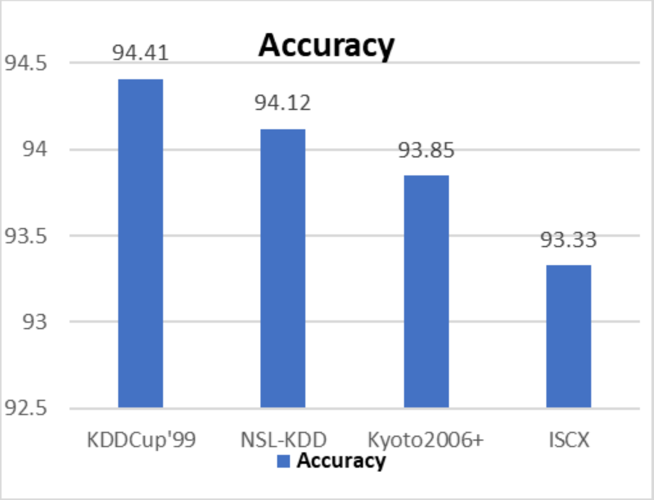


Figure 2. Comparing accuracy of datasets

False-positive rate represents all misclassified testing instances of the trained dataset. That is the instance which should be normal-type but misclassified as anomaly-type. The formula for False-positive rate is-

$$\text{False positive rate (FRP)} = FP/(FP+TN)$$

Where FP is false positive, TP true positive, TN true negative and FN is false negative all these values obtained from the confusion matrix created during the testing process. False-positive rate of KDDCup’99 is minimal compared to all the dataset shown in Figure 3.

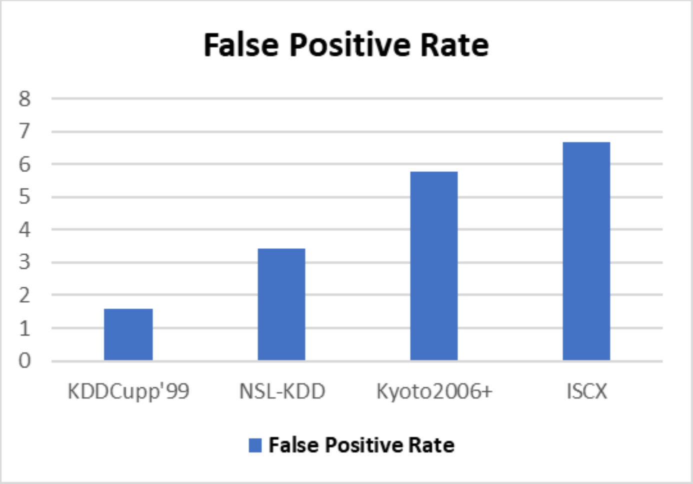


Figure 3. Comparing False-positive rates of dataset

VI CONCLUSION

In this paper, we have implemented an intrusion detection system using kernel-based strategies or kernel methods with feature selection metrics and biased classifiers using threshold hold values as the deciding factor for the developed system. The idea to use a kernel method is to obtain better results for intrusion detection as these methods use mathematical calculations. For feature optimization the feature extraction techniques have proven to be great help. This feature selection metrics diminish irrelevant use of features. A system with more accuracy and minimum false-positive rates are considered to be better for the intrusion detection system. We obtain the highest accuracy of 94.41% and lower false-positive value of 1.6 in our implemented system. For future research work, we may consider different kernel methods with feature selection metrics and other classifiers for getting better results outcomes.

REFERENCES

[1]. Md Zahangir Alom, Tarek M. Taha, “Network Intrusion Detection for Cyber Security using Unsupervised Deep Learning Approaches” in National Aerospace Electronics Conference (NAECON) IEEE, 2017.

[2]. J.P. Anderson, "Computer security threat monitoring and surveillance", Technical Report, James P. Anderson Co., Fort Washington, PA, 1980.

[3]. Bashir, U., & Chachoo, M., “Intrusion detection and prevention system: Challenges & opportunities”, in International Conference on Computing for Sustainable Global Development, 2014.

[4]. Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, Ali A. Ghorbani, “Toward developing a systematic approach to generate benchmark datasets for intrusion detection”, Computers & Security, Volume 31, Issue 3, May 2012, Pages 357-374.

[5]. Akash G. Gedam, “Direct Kernel Method for Machine Learning with Support Vector Machine”, in International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017.

[6]. Azar Abid Salih, Maiwan Bahjat Abdulrazaq, “Combining Best Features Selection Using Three Classifiers in Intrusion Detection System”, in International Conference on Advanced Science and Engineering (ICOASE), 2019.

[7]. Zeya Zhang, Zhiheng Zhou, & Dongkai Shen., “Sample selection method in supervised learning based on adaptive estimated threshold”, in the International Conference on Machine Learning and Cybernetics, in 2013.

[8]. Dmitrii Marin, Meng Tang, Ismail Ben Ayed, Yuri Boykov, “Kernel clustering: density biases and solutions” in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[9]. Saxena, A. K., Sinha, S., & Shukla, P., “General study of intrusion detection system and survey of agent-based intrusion detection system”, in International Conference on Computing, Communication and Automation (ICCCA), 2017.

[10]. Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahman Department, “Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection”, in International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2019.

[11]. Xiaoyan Ma, Yanbin Zhang, Yanxia Wang, “Performance evaluation of kernel functions based on grid search for Support Vector Regression”, in *IEEE 7th International Conference on CIS & RAM, 2015.*

[12]. Mohammed A. Ambusaidi, Xiangjian He, Zhiyuan Tan, Priyadarsi Nanda, Liang Fu Lu and Upasana T. Nagar, "A novel feature selection approach for intrusion detection data classification" in IEEE 13th ICTSPCC,2014.

[13]. Anil S, Remya R, "A hybrid method based on Genetic Algorithm, Self-Organised Feature Map, and Support Vector Machine for better Network Anomaly Detection", in 4th ICCCNT, Tiruchengode, India, 2013.

[14]. Kathleen Goeschel, "Reducing False Positive in Intrusion Detection Systems Using Data-Mining Techniques Utilizing Support Vector Machines, Decision Trees, And Naive Bayes For Off-Line Analysis", in IEEE, 2016.

[15]. Sourish Ghosh, Anasuya Dasgupta, Aleena Swetapadma, “A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification” in International Conference on Intelligent Sustainable Systems (ICISS), 2019.