PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR IDS

K.Priyanka, Dr. R. Jegadeesan, V.Neelima, Divya, Sairamachari, M.Ganesh

Assoc.prof., B.Tech Final Year Students-Computer Science and Engineering

Jyothishmathi Institute of Technology and Science, Karimnagar, India

Abstract: Any abnormal activity will be assumed to be anomalies intrusion. Intrusion could be a major issue that is to be addressed and should be detected. Intrusion detection could be a central part of security tools, such as reconciling security appliances, intrusion detection systems, intrusion interference systems, and firewalls. Different techniques have been used for intrusion detection however their attainment is a problem. The performance of intrusion detection system primarily depends on accuracy that must produce less warning and high detection rate. To resolve the problems related to performance the techniques like support vector machine (SVM), Multi layer perceptron and different techniques are utilized in recent work. These techniques impose bound limitations and conjointly not applicable for big knowledge sets, like system and network knowledge. This downside is addressed during this paper by introducing well-known economical machine learning algorithms like SVM, random forest, and extreme learning machine (ELM). These techniques are well-known due to their ability in classification.

Index Terms - Detection rate, extreme learning machine, SVM, Warning.

I. INTRODUCTION

Intrusion may be a severe issue in security and a chief drawback of security breach, as a result of one instance of intrusion will steal or delete information from laptop and network systems in a few seconds. Intrusion can even harm system hardware. Furthermore, intrusion will cause large losses financially and compromise the IT vital infrastructure, thereby leading to data inferiority in cyber war. Therefore, intrusion detection is very important. Different intrusion detection techniques are obtainable, but their accuracy remains associate issue; accuracy depends on detection and warning rate, the matter on accuracy desires to be self-addressed to cut back the false alarms rate and to increase the detection rate. This notion was the impetus for this analysis work. Thus, support vector machine (SVM), random forest (RF), and extreme learning machine (ELM) applied during this work; these ways are proved effective in their capability to deal with the classification. Intrusion detection mechanisms are valid on a standard dataset, KDD. This work used the NSL-knowledge discovery and data processing (KDD) dataset, that is associate degree improved ty of the KDD and is taken into account a benchmark within the analysis of intrusion detection strategies. The remainder of the paper is organized as careful below. The connected work is given in Section II. The projected model of intrusion detection to that totally different machine learning techniques are applied is delineate in Section III. The implementation and results ar mentioned in Section IV. The paper is finished in Section V, that provides a summary and directions for future work.

II. RELATED WORK

Verifying pc and system data is imperative for associations and individuals because of traded off data will cause sizeable damage. To maintain a strategic distance from such conditions, interruption recognition frameworks zone unit essential. As of late, unique AI approaches are anticipated to improve the execution of interruption location frameworks. Wanget al arranged partner degree interruption location system dependent on SVM and substantial their strategy on the NSL-KDD dataset. They guaranteed that their procedure, that has 99.92% adequacy rate, was better than various methodologies; in any case, they neglected to make reference to utilized dataset measurements, scope of preparing, and testing tests. Yet in addition, the SVM execution diminishes once huge data zone unit concerned, partner degree and it is anything but a perfect option for dissecting huge system traffic for interruption identification.
Cortes C, Vapnik VN [1], exhibited SVM which utilizes the NSL-KDD informational index is a refined adaptation of its forerunner KDD”99 informational collection with a sizable measure of value information which emulates the ongoing can just train and test an interruption location framework. The NSL-KDD dataset is utilized to assess the proposed strategy, and the exact outcomes demonstrate that it accomplishes a superior and more rigorous execution than existing strategies as far as exactness, location rate, false caution rate and preparing speed.

An epic help vector machine (SVM) show consolidating part essential segment investigation (KPCA) with hereditary calculation (GA) is proposed by J.H. Lee, J.H. Lee, S.G. Sohn, et al [2], for interruption recognition. In the proposed model, a multi-layer SVM classifier is received to appraise whether the activity is an assault, KPCA is utilized as a preprocessor of SVM to lessen the component of highlight vectors and abbreviate preparing time. So as to decrease the commotion brought about by highlight contrasts and improve the execution of SVM, an improved bit work (N-RBF) is proposed.

In AI, a blend of classifiers, known as an ensemble classifier, which is a model created by Nelsileno Virgúfio de Souza Araújo [3], frequently beats singular ones. While numerous gathering approaches exist, it remains, be that as it may, a troublesome errand to locate a reasonable ensemble design for a specific dataset. This paper proposes a novel gathering development strategy that utilizes PSO produced loads to make troupe of classifiers with better exactness for interruption discovery.

Random forest (RF) is a group classifier this model was presented by Ujwala Ravale, Nilesh Marathe, Puja Padiya [4], used to improve the precision. Random forest comprises of numerous choice trees. Random forest has low characterization error compared with other customary grouping calculations. Number of trees, least hub size and number of highlights utilized for part every hub. Points of interest of RF are recorded beneath. 1) Generated forests can be put something aside for future reference. 2) Random forest defeats the issue over fitting. 3) In RF exactness and variable significance is naturally produced. Highlight choice (FSS) is a preparing step usually utilized in information mining. It is viable in dimensionality decrease and evacuates irrelevant features along these lines builds exactness.

Intrusion Detection Systems provide police assistance, prevention and resistance to unauthorized access. Thus, Aburomman Reaz Reaz [5] planned a classification of the group, which could be a combination of PSO and SVM; this rating surpassed 92.90% on various roads. The data discovery and processing dataset was used in 1999 (KDD99), which has previously reported defects. What's more, SVM is not a good choice for massive analytics, as its performance deteriorates as the volume of information grows. Raman et al. Fellow planned in the detection mechanism of nursing penetration supported the base graph Algorithm Genetic Hyper (HG-GA) to prepare the parameter and have an option in SVM. They claimed that their methodology outperformed current approaches with a 97.14% detection rate on associates in the NSL-KDD nursing dataset. It was used to test and verify intrusion detection systems.

We consider the problem of intrusion detection in the computer network, investigate the use of maximum learning machines (ELMs) to classify and detect interventions. With increased communication between networks, the risk of information systems on external attacks or interventions has increased significantly. Chi cheng [6] Automated learning methods such as transmission support devices (SVMs) and neural networks have been used extensively to detect intrusion. These methods generally suffer from long training periods, requiring control of transactions, or not performing well in classifying multiple layers. We suggest a basic ELM method based on random features, the ELM-based method of the nucleus of the classification. We compare our methods with commonly used SVM techniques in both binary and multi-layered classifications. The simulation results show that the proposed ELM core proposal outperforms SVM in the speed of training and testing, while the kernel-matching ELM achieves higher detection accuracy than SVM in a multi-layer classification.

Farnaz and Jabbar [7] developed a model for the RF-assisted intrusion detection system. They tested the effectiveness of their model on associates in the NSL-KDD nursing dataset, and their results were non-tabular at 99.67% detection compared to J48. The most fundamental limitation of the RF algorithm is that many trees can create a slow mathematical base for time prediction. Elbasiony et al. Planned for the RF-supported intrusion detection model and k-weighted; it corrects its model on the KDD99 data set. The results of the system are indisputable with 98.3% accuracy. RF is not suitable for predicting real traffic due to its slowness, thanks to the formation of a large range of trees. To boot, the KDD99 data set refers to some restrictions as has been said.

III. PROPOSED MODEL

The key phases of the planned model embrace the dataset, preprocessing, classification, and evaluation analysis. Every part of the planned system is very important and adds valuable influence on its performance. The core focus of this work is to research the performance of various classifiers, namely, SVM,
RF, and ELM in intrusion detection. Figure one demonstrates the model of intrusion detection system planned during this work. The proposed model involves certain steps which are represented in the architecture of proposed model. The loaded data set is preprocessed and then it is subjected to classification and then evaluation is performed.

![Proposed model for intrusion detection system](image)

**A. Dataset**

Dataset choice for experimentation may be a vital task, as a result of the performance of the system is predicated on the correctness of a dataset. The more accurate the information, the bigger the effectiveness of the system. The dataset will be collected by various suggests that, like 1) change dataset, 2) simulated dataset, 3) testbed dataset, and 4) normal dataset. However, complications occur within the application of the primary three methodologies. A true traffic methodology is dear, whereas the change methodology is unsafe. The event of a simulation system is additionally complicated and difficult to boot. Differing kinds of traffic square measure required to model numerous network attacks, that is complicated and expensive. To beat these difficulties, the NSL–KDD dataset is employed to validate the projected system for intrusion detection.

**B. Preprocessing**

The classifier is unable to method the raw dataset as a result of a number of its symbolic options. Thus, preprocessing is essential, during which non-numeric or symbolic options are eliminated or replaced, as a result of they are doing not indicate important participation in intrusion detection. However, this method generates overhead as well as a lot of coaching time; the classifier’s design becomes advanced and wastes memory and computing resources. Therefore, the non-numeric options are excluded from the raw dataset for improved performance of intrusion detection systems.

**C. Classification**

Placing an activity into traditional and intrusive classes is the core activity of intrusion detection system, which is known as the intrusive analysis engine. Thus, different classifiers are applied as intrusive analysis engines in intrusion detection within the literature, like multilayer perceptron, SVM, naive Thomas Bayes, self-organizing map, and DT. However, during this study, the 3 completely different classifiers of
SVM, RF, and ELM area unit applied supported their proved ability in classification issues. Details of every classification approach area unit provided

1) Support Vector Machine
SVMs were at the start projected by Vapnik (1995) for determination problems of classification and multivariate analysis. SVM is a supervised learning technique that’s trained to classify different classes of information from numerous disciplines. These have been used for two-class classification issues and area unit applicable on each linear and non-linear information classification tasks. SVM creates a hyperplane or multiple hyperplanes in a high-dimensional house, and also the best hyperplane in them is that the one that optimally divides information into completely different classes with the biggest separation between the categories. A non-linear classifier uses numerous kernel functions to estimate the margins. The most objective of those kernel functions (i.e., linear, polynomial, radial basis, and sigmoid) is to maximize margins between hyperplanes. Recently, several extremely promising applications are developed by researchers because of the increasing interest in SVMs.

![Architecture of the SVM for intrusion detection](image)

Figure 2. Architecture of the SVM for intrusion detection

The figure 2 illustrates the design of the SVM classification model within the projected intrusion detection system. We have used the radial basis perform (RBF) kernel for the implementation of the SVM model within the projected system. The kernel perform uses square geometrician distance between 2 numeric vectors and maps information(input , file) to a high dimensional area to optimally separate the given data into their various attack categories. Therefore, kernel RBF is especially effective in separating sets of information that share complicated boundaries. In our study, all the simulations are conducted victimization the freely accessible LibSVM package.

Given that the chosen drawback may be a multiclass classification drawback, it uses the notion of 1 vs all for attack classification. during this notion, the multiclass drawback is split into a two-class drawback. The radial basis operate (RBF) kernel is employed during this study, that is diagrammatical as follows:

$$K(x,y) = e^{-\gamma|x-y|^2}, \quad \gamma > 0$$

For given coaching samples ($x_i, y_i$), $i = 1, 2, \ldots, n$, where $i$ is the maximum range of samples within the coaching knowledge, $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, wherever one shows samples from a positive class and $-1$ represents sequences from the negative category. When exploitation SVM, the answer of the subsequent downside is provided.

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

subject to $y_i(\phi(x_i)^T w + b) \geq 1 - \xi_i$

Here, $\phi$ transforms the coaching vector $x_i$ to the upper dimensional area. Following this, the SVM shows a hyper-plane having most margin to separate totally different categories of knowledge.

The determined results via the SVM model aren't considerably convincing compared with those from the opposite classifiers. The advantage of SVM is that stripped-down parameter adjustment is needed. The disadvantages of it embody the necessities of a mathematician perform for every instance of the coaching set, thereby increasing coaching time and performance degradation on terribly giant datasets with thousands
of instances, as within the case classification. Just in case most margin classifier fails to search out any separating hyperplane, soft margin is employed to beat this downside. Soft margin uses positive slack variables $\xi_i$, $i = 1, 2, \ldots, N$ within the constraints as follows:

$$(w.x_i-b) \geq +1-\xi_i \text{ for } y_i = +1$$

$$(w.x_i-b) \geq -1+\xi_i \text{ for } y_i = -1$$

$\xi_i \geq 0$

When a slip happens, $\xi_i$ need to exceed unity. Then, $i \xi_i$ is associate edge on the coaching error. The Lagrange during this situation is as follows

$$LP=\frac{1}{2}\|W\| + C\sum_{i=1}^{n} \xi_i - \sum_{j} \alpha_j (y_j(x_i.w - b) - 1 + \xi_i) - \sum \mu_i \xi_i$$

where, $\mu_i$ represents Lagrange multipliers wont to acquire the positive worth of $\xi_i$.

2) Random Forest

RFs are ensemble classifiers, that are used for classification and regression analysis on the intrusion detection knowledge. RF works by making numerous call trees within the coaching section and output category labels those have the majority vote. RF attains high classification accuracy and might handle outliers and noise within the knowledge. RF is employed during this work as a result of it's less prone to over-fitting and it's antecedently shown smart classification results.

Figure 3 shows the implementation of the random forest classification model within the information classification within the projected system. A preprocessed sample of n samples is fed to the random forest classifier. RF creates $n$ totally different trees by employing a range of feature subsets. Every tree produces a classification result, and also the results of the classification model depends on the bulk ballot. The sample is appointed to the category that obtains highest ballot scores. The antecedently earned classification results indicate that RF is fairly suitable within the classification of such information as a result of in some cases, it's obtained higher results than produce other classifiers. Alternative advantages of the RF embody its higher accuracy than Adaboost and fewer probabilities of overfitting. The architecture of the random forest for intrusion detection is given in Figure 3.

3) Extreme Learning Machine

ELM is sometimes referred as single or multiple hidden layer feedforward neural networks. ELM will be wont to solve numerous classification, clustering, regression, and have engineering issues. This learning rule involves input layer, one or multiple hidden layers and therefore the output layer. Within the ancient neural networks, the tasks of adjustment of the input and hidden layer weights square measure terribly computationally high-ticket and long as a result of it needs multiple rounds to converge.
To overcome this drawback, Huang et al. planned an SLFN by indiscriminately choosing input weights and hidden layer biases to attenuate the coaching time. The excellent detail of ELM is accessible in Huang et al. and Qayyum et al. The authors claim that these models learn quicker and attain higher generalization capability as compared with different feedforward network models. ELM performance is comparable SVM or different progressive machine learning classifiers. ELM has the best ability to perform higher in extremely complicated datasets. The design of the planned system is shown in Figure 4.

N input samples \((z_i, y_i)\) are gift, wherever \(z_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T\) is that the ith sample with n completely different options and \(y_i = [y_{i1}, y_{i2}, \ldots, y_{im}]^T\) describes the particular labels of \(x_i\) with ancient SLFN with K hidden neurons that is outlined as follows:

\[
\sum_{n=1}^{K} \beta_j h(w_n, a_j + c_n) = \alpha_j \quad j=1,2,\ldots,N
\]

Where \(w = [w_{n1}, w_{n2}, \ldots, w_{nk}]^T\) is the chosen weight vector associate degree indicates an jth hidden somatic cell reference to the input nodes. \(\beta_j=[\beta_{j1}, \beta_{j2}, \ldots, \beta_{jn}]^T\) shows the burden vector with affiliation of jth hidden neuron and also the output nodes and cm is that the threshold of the jth hidden neuron \(\alpha_k = [\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kn}]^T\) is the kth output neuron. \(h(.)\) represents the activation operate and SLFN used for Mhidden neurons and activation operate will approach these N coaching samples with zero error. Numerous alternative techniques are applied to find and classify intrusion of wired and wireless atmosphere.

**D. Evaluation**

The designed system is evaluated supported the quality dataset NSL–KDD, that is irregular and divided into 3 elements, namely, the total dataset, and therefore the 1/4 dataset. the total dataset consists of 65,535 samples, the includes 32,767 samples, and therefore the 1/4 th dataset consists of 18,383 samples. Accuracy, precision, and recall are used as analysis metrics.

**Accuracy**: Accuracy is computed as “the total variety of correct prediction, True Positive (TP) + True Negative (TN) divided by the whole variety of a dataset Positive (P) + Negative (N)”.

\[
\text{Accuracy} = \frac{(TP + TN)}{(P + N)}
\]
Figure 5. Accuracy of the SVM, RF and ELM (80% training and 20% testing)

**Precision**: Preciseness is computed as “the variety of correct positive predictions (TP) divided by the whole variety of positive predictions (TP + FP)”. Preciseness is additionally called a positive prognosticative worth. 

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

Figure 6. Precision of the SVM, RF and ELM (80% training and 20% testing)

**Recall**: Recall is computed as “the variety of correct positive predictions (TP) divided by the whole variety of positives (P)”. Recall is additionally called verity positive rate or sensitivity.

\[
\text{Recall} = \frac{TP}{P}
\]
Figure 7. Recall of the SVM, RF and ELM (80% training and 20% testing)

IV. RESULTS

The exactness/accuracy of SVM (Linear), SVM (RBF), RF, and ELM on 20% testing and 80% training information tests is appeared in Figure 5. ELM performs better than SVM (Linear), SVM (RBF) and RF on full information tests, while SVM (RBF) shows improved exactness over RF and ELM on half information tests. SVM (Linear) beats different procedures on 1/4 information tests, as delineated in Figure 5.

The precision of SVM (Linear), SVM (RBF), RF, and ELM on 20% testing and 80% preparing information tests is appeared in Figure 6. The precision of ELM is superior to that of SVM Linear and RBF on the full information tests, and it moreover outflanks that of RF. On half information tests, the precision of SVM (Linear) is higher than that of SVM (RBF), ELM, and RF. On 1/4 information tests, the precision of SVM (Linear) is equivalent to that of SVM (RBF). Besides, the SVM performs superior to ELM and RF in the 1/4 dataset.

The recall of SVM (Linear), SVM (RBF), RF, and ELM on 20% testing and 80% preparing information tests is appeared in Figure 7. On full information tests, the recall of ELM performs superior to those of SVM (Linear), SVM (RBF), and RF. The recall of SVM (Linear) is more prominent than those of SVM (RBF), ELM, and RF. The positioning of review on 1/4 of information tests is as per the following: first for SVM (RBF), second for SVM (Straight), third for RF, and fourth for ELM. The above mentioned dialog demonstrates that SVM performs better on a little dataset, though EML beats others approaches on huge datasets.

The exactness/accuracy of SVM (Linear), SVM (RBF), RF, and ELM on 10% testing and 90% preparing information tests is appeared in Figure 8. On the full information tests, the exactness of ELM is superior to that of SVM (direct), SVM (RBF), and RF. The SVM (RBF) beats SVM (Linear), ELM, and RF on the half information tests. The SVM (straight) demonstrates better execution on 1/fourth information tests as contrasted and SVM (RBF), RF, and ELM.
The precision of SVM (Linear), SVM (RBF), RF, and ELM on 10% testing and 90% preparing information tests is appeared in Figure 9. The outcomes demonstrate that the ELM shows better precision than RF, SVM (RBF), and SVM (Linear) on full information tests, while SVM (Linear) demonstrates better precision on the half information tests. Besides, SVM (Linear) performs superior to ELM and RF on 1/4 dataset.

The recall of SVM (Linear), SVM (RBF), RF, and RLM on 10% testing and 90% preparing information tests is appeared in Figure 10. On full information tests, the recall of ELM out per frames those of SVM (direct), SVM (RBF), and RF, while the recall of SVM (straight) is superior to those of SVM (RBF), ELM, and RF on half information tests. On the 1/4 information tests, SVM (RBF) is practically equivalent to SVM (Linear), while it demonstrates better outcomes over RF and ELM, as appeared in Figure 10.
V. CONCLUSION

Intrusion detection and bar area unit essential to current and future networks and data systems, as a result of our daily activities area unit heavily addicted to them. what is more, future challenges can become a lot of intimidating thanks to the net of Things. During this respect, intrusion detection systems are vital within the previous few decades. many techniques are utilized in intrusion detection systems, however machine learning techniques area unit common in recent literature. in addition, totally different machine learning techniques are used, however some techniques area unit a lot of appropriate for analyzing vast information for intrusion detection of network and data systems. to handle this drawback, totally different machine learning techniques, namely, SVM, RF, and ELM area unit investigated and compared during this work. ELM outperforms alternative approaches in accuracy, precision, and recall on the complete information samples that comprise sixty five,535 records of activities containing traditional and intrusive activities. what is more, the SVM indicated higher results than alternative datasets in half the info samples and in 1/4 of the info samples. Therefore, ELM could be a appropriate technique for intrusion detection systems that area unit designed to investigate a large quantity of information. In future, ELM are going to be explored more to research its performance in feature choice and have transformation techniques.

REFERENCES

[6] chi cheng; ELM for intrusion detection 30 july 2012,10.1109/IJCNN.2012.6252449


