Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM

Anish Halimaa A
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
anishhalimaatce@gmail.com

Dr. K.Sundarakantham
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India,
kskcse@tce.edu

***Abstract*— In order to examine malicious activity that occurs in a network or a system, intrusion detection system is used. Intrusion Detection is software or a device that scans a system or a network for a distrustful activity. Due to the growing connectivity between computers, intrusion detection becomes vital to perform network security. Various machine learning techniques and statistical methodologies have been used to build different types of Intrusion Detection Systems to protect the networks. Performance of an Intrusion Detection is mainly depends on accuracy. Accuracy for Intrusion detection must be enhanced to reduce false alarms and to increase the detection rate. In order to improve the performance, different techniques have been used in recent works. Analyzing huge network traffic data is the main work of intrusion detection system. A well-organized classification methodology is required to overcome this issue. This issue is taken in proposed approach. Machine learning techniques like Support Vector Machine (SVM) and Naïve Bayes are applied. These techniques are well-known to solve the classification problems. For evaluation of intrusion detection system, NSL– KDD knowledge discovery Dataset is taken. The outcomes show that SVM works better than Naïve Bayes. To perform comparative analysis, effective classification methods like Support Vector Machine and Naive Bayes are taken, their accuracy and misclassification rate get calculated.**

Keywords*— *Intrusion Detection, Support Vector Machine Naive Bayes, Machine Learning.

I.INTRODUCTION

In order to recognize abnormal behaviour that occurs in a computer or network, Intrusion detection system (IDS) is used. IDSs can be characterized in several ways, among them misuse-based and anomaly based IDSs are the most common. To detect known attack like snort, Misuse-based IDS can perform proficiently. This type of IDSs has less false alarm rate. It incapable to recognizes new attacks which does not personalize any instruction in database. In Anomaly-based IDS, it develops a model of regular behaviour after that; it separates any essential deviations from this model and consider that deviation as intrusion. This type of IDS has the ability to detect both known and unknown attacks, but encounters a high false alarm rate. Various machine learning techniques are incorporated to decrease false alarm rate.

A. Intrusion Detection System

A distinct existence of intrusion can steal or eliminate information from computer or network systems in limited duration. Hence intrusion is one of the major issues in network security. System hardware also gets harm due to intrusion. Various techniques of intrusion detection are performed; however accuracy is one of the major problems. Detection rate and false alarm rate plays an essential role for the analysation of accuracy. Intrusion detection must be enriched to reduce false alarms and to increase the detection rate. Thus, Support Vector Machine (SVM) and Naïve Bayes are applied. Classification can be addressed by these algorithms. Apart from that, Normalization and Feature Reduction are also applied to make a comparative analysis.

B. Machine Learning

Machine Learning is used to automate analytical model building. It is a technique of data analysis. It is one of the branches of Artificial Intelligence which works on the concept that a system gets trained, make decisions and learn to identify patterns with fewer interventions of humans. Supervised and Unsupervised learning are the two most extensively used machine learning techniques. Labeled examples like an input with preferred output are taken for training algorithms. Instances without historical labels get trained using unsupervised learning. To discover some structure within the data and to explore the data are the two main objective of unsupervised learning. Apart from these methods, approaches like Semisupervised learning and Reinforcement learning are used.

For training purpose, semisupervised learning uses fewer amounts of labeled data and huge amounts of unlabeled data. Trial and error method is used in Reinforcement Learning in which the actions yield the best rewards. Classification, regression and prediction are used. Agent, environment and actions are the three primary component used in this type of learning. The goal is that, the agent has to select those actions, which exploit the predictable reward. By applying good policy, the agent able to reach the goal much faster.

C. Support Vector Machine

Support Vector Machine (SVM) comes under supervised learning method, in which various types of data from different subjects get trained. In a high-dimensional space, SVM creates hyperplane or multiple hyperplanes. The hyperplane which optimally separates the given data into various classes with the major partition, consider as a best hyperplane. For evaluate the margins between hyperplanes, a non-linear classifier applies various kernel functions. Maximizing margins between hyperplanes is the main aim of these kernel functions like linear, polynomial,

radial basis, and sigmoid. Due to the growing attention in SVMs, the eminent applications have been established by the developers and researchers. SVM deals a main role in image processing and pattern recognition applications.

Usually a classification task mainly involves dividing data into two sets namely, training datasets and testing datasets. In that class label will be defined as "target variables" and attributes will be defined as features or "observed variables.

D. Naive Bayes

Bayesian classifiers are statistical classifiers. They are capable to forecast the probability that whether the given model fits to a particular class. It is based on Bayes' theorem. It constructed on the hypothesis that, for a given class, the attribute value is independent to the values of the attributes. This theory is called class conditional independence.

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)\ P(H)}{P(\mathbf{X})}$$

II. LITERATURE SURVEY

Protecting computer and network information of an organizations and individuals become an important task, because compromised information can cause huge loss. Hence, intrusion detection system is used to prevent this damage. To enrich the function of IDS, different machine learning approaches get developed. The main objective [2] is to address the problem of adaptability of Intrusion Detection System (IDS).The proposed IDS has the proficiency to recognize the well-known attacks as well as unknown attacks. The proposed IDS consist of three major mechanisms: Clustering Manager (CM), Decision Maker (DM), Update Manager (UM). NSL-KDD dataset is applied to estimate the working of the proposed IDS. Both supervised and unsupervised techniques were accompanied. The information received to the system is grounded on the education of an agent who disregards the correction proposals presented by IDS. This technique is applied on supervised mode. Both known and unknown traffics can be detected by the system, when they work under unsupervised mode. After updating recently arrived data from both supervised and unsupervised modes, the function of the system has been improved. Performance of the system gets improved, when it runs in unsupervised mode.

By incorporating machine learning techniques like, [3] SVM and Extreme Learning Machine (ELM), a hybrid model get developed. Modified K-means is used to construct high quality dataset. It builds small dataset that denote overall original training datasets. By this step, the training time of the classifier gets reduced. KDDCUP 1999 is used for implementation. It shows accuracy of about 95.75 percentages.

Various machine learning techniques like SVM, Random Forest (RF) and ELM are examined to report this problem. ELM shows better result when compared to other techniques in accuracy. Datasets get divided into one-fourth of the data samples, half of the dataset and full datasets. However, SVM produces better results in half of the data samples and one-fourth samples of data. ELM is the best method to handle the huge amount of data of about two lakh instances and more.

A new hybrid classification algorithm on Artificial Bee Colony (ABC) and Artificial Fish Swam (AFS) is proposed [6]. Nowadays computer system is prone to different information thefts due to the widespread usage of internet, which leads to the emergence of IDS. Fuzzy C-Means Clustering (FCM) and Correlation-based Feature Selection (CFS) is applied [6] for separating training datasets and to eliminate irrelevant features. If-then rues are generated by using CART technique, which is applied to differentiate normal and anomaly records according to the selected features.

Correlation-based feature selection method which is a simple filter-based model is used in the proposed system. Datasets containing the features, highly correlated with the class, yet uncorrelated with the others are applied. By using NSL-KDD and UNSW-NB15 datasets this approach get achieved 99 percentages of detection rate of anomalies and 0.01 percentages of false positive rate. A hybrid method for A-NIDS using AdaBoost algorithms and Artificial Bee Colony to obtain low false positive rate (FPR) and high detection rate (DR).

III.EXISTING METHODOLOGIES

The performance of the proposed model is evaluated by the KDD Cup dataset. In order to train classifiers like SVM and ELM, 10 percent KDD training dataset is taken which contains large number of instances. 10 percent KDD dataset is taken rather than entire dataset, because applying entire dataset will cause several problems. Symbolic attributes like protocol, service and flag get changed or removed. Finally, the instances get labeled under four categories: Normal, DoS, Probe, and R2L. They have trained SVM and ELM with the Dataset. For testing process, they have used multi-level model with corrected KDD dataset. Accuracy of the proposed model has attain up to 95.75 percentages and false alarm rate of 1.87 percentages by using KDD Cup 1999 dataset.

IV.PROBLEM STATEMENT

Due to the excessive volume of data, false alarm report of intrusion to network gets increased and detection accuracy gets reduced. This is one of the major issues when the system encounters unknown attacks. The main objective is to increase the accuracy rate and to lessen the false alarm rate. To meet the above challenges machine

learning algorithm like SVM and Naïve Bayes has been used.

V.PROPOSED APPROACH

Dataset pre-processing, classification and result evaluation are the vital phases in the proposed model. In proposed system each phase is essential and enhances important influence on its performance. To examine the function of SVM and Naïve Bayes classifiers are the essential steps of this work.

A. Pre-Processing

Dataset contains symbolic features; these features are unable to process by the classifier. Hence, pre-processing takes place. In this phase all non-numeric or symbolic features get removed or exchanged. Elimination or replacement of non-numeric or symbolic features is done in pre-processing phase.

The overall process of pre-processing is essential, in which non-numeric or symbolic features are eliminated or replaced, as they do not perform any important participation in intrusion detection. Symbolic attributes like protocol, service and flag get changed or removed. Finally, the instances get labeled under four categories: Normal, DoS, Probe, and R2L.

B. Methodology

- Comparative analysis done between SVM and Naïve Bayes for classification of dataset, to analyze their accuracy and Misclassification Rate. At first raw dataset is taken and the class attribute contains 24 different types of attack which get labeled under 4 categories. They are normal, Dos, Probe, r2l.

- After, labeling Pre-processing is done to convert nominal attribute to binary attribute. In order to obtain improved performance of intrusion detection system, non-numeric features get removed.

- For randomization, the dataset is allowed to get processed in WEKA tool by incorporating the filter Randomize. Randomize filter randomly shuffles the order of instances passed through it by setting a random number generator, in which the seed value get reset. Collecting the first 19,000 instances for comparative analysis.

- In order to get different result and to improve the performance of the dataset, methodologies like CfsSubsetEval is done for feature reduction. The given dataset after preprocessing under goes feature reduction and normalization.

- CfsSubsetEval is one of the methods of attribute selection. It calculates the value of attributes by considering the individual predicting estimation of all features along with the degree of redundancy between them.

- About classification under SVM, it comes under supervised learning method, in which various types of data from different subjects get trained. In a given high dimensional space, Support Vector Machine creates hyperplane or multiple hyperplanes in a high-dimensional space. SVM creates hyperplane or multiple hyperplanes.

- The hyperplane which optimally separates the given data into various classes with the major partition, consider as a best hyperplane. For evaluate the margins between hyperplanes, a non-linear classifier applies various kernel functions. Maximizing margins between hyperplanes is the main aim of these kernel functions like linear, polynomial, radial basis, and sigmoid.

- For the first 19,000 instances, classification of raw dataset using SVM, SVM under different Normalization techniques and SVM along with Feature Reduction is done for comparative analysis. Accuracy and Misclassification rate also noted.

- Same, process is done using Naïve Bayes. Bayesian classifiers are statistical classifiers. They are capable to forecast the probability that whether the given model fits to a particular class. It is based on Bayes' theorem. It works on the hypothesis that, for a given class, the attribute value is independent to the values of the attributes. This theory is called class conditional independence.

- Naive Bayes classifier works as follows: Training set of samples get denoted by T, each with their class labels. There are k classes, X1,X2,X3,..,X(k-1),Xk. A = {a1,a2,...,an}, depicting n measured values of the n attributes, m1,m2,...,mn, whereas A depicting n-dimensional vector,. b) For a given sample A, the classifier will calculate A, which fits to the class having the maximum posteriori probability, conditioned.

The block diagram of this approach is given in "Fig. 1". Accuracy has been calculated and a graph has been plotted based on the obtained results. From the graph, we have can analyze, SVM outperforms Naïve Bayes.

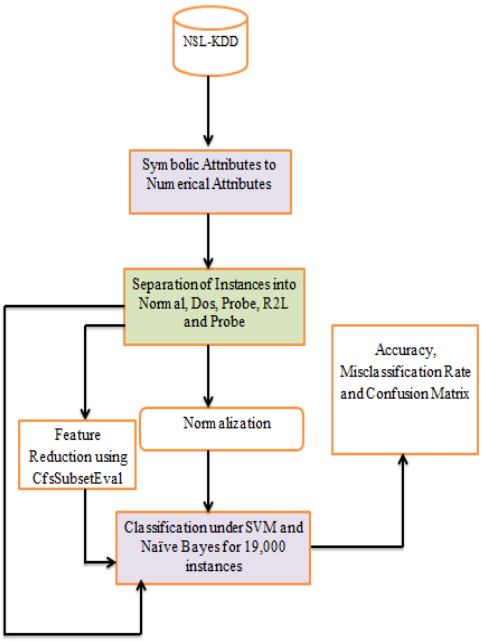

Fig.1. Block Diagram

NAIVE BAYES ALGORITHM

INPUT: Training Set T
Predictor Variable P
OUTPUT: A group of dataset for testing.
STEPS:

1. Read the Training set T.
2.Calculate the conditional probability P for every class
H ←> dependent class
X ←> class variable

$$P(X|H) = \frac{P(H|X)*P(X)}{P(H)}$$

3.Find the class with maximum probability.
4.Generate confusion matrix
5.Find Accuracy and Misclassification rate

SUPPORT VECTOR MACHINE

INPUT: Preprocessed Data
OUTPUT: Output Classes
STEPS:

1 Calculate Objective Function T
2.Objective function= $\min w\lambda \| w \|2 + \Sigma(1-y_i(x_i,w))$
Where xi is the input sample, y_i is the output label,
W is weight vector, λ is reguralaization parameter
3. Apply gradient descent learning w.r.t weight
4. Update rule for weight for misclassified output
$w=w+\eta(y_i x_i - 2\lambda w)$.
5. Update rule for weight for correctly classified output
$w=w+\eta(i - 2\lambda w)$, where η is the learning vector
6. Return T
7.end function

VI. PERFORMANCE ANALYSIS

By analyzing accuracy rate and misclassification rate, the performance of SVM and Naïve Bayes algorithm has been evaluated for 19,000 instances. The performance metrics of these algorithms is evaluated by the information from confusion matrix

Methodology	Accuracy Rate	Misclassification Rate
SVM	97.29	2.705
Naïve Bayes	67.26	32.73
SVM-CfsSubsetEval	93.95	6.04
Naïve Bayes-CfsSubsetEval	56.54	43.45
SVM-Normalization	93.95	2.705
Naïve Bayes-Normalization	71.001	28.998

TABLE I. Accuracy and Misclassification rate of algorithms

A. Evaluation

The model is evaluated based on NSL-KDD dataset, after applying methodologies like pre-processing and randomization. The dataset consists of 19,000 samples. Accuracy rate and Misclassification rate are taken as evaluation metrics.

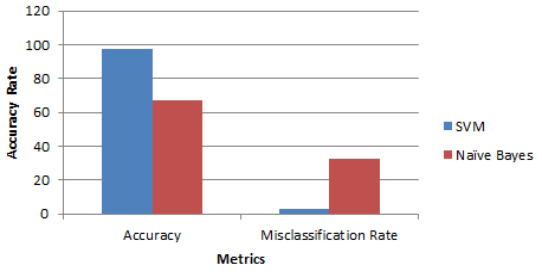

Fig 2.Accuracy and misclasssification rate of SVM and Naive Bayes for 19,000 instances

The above graph describes the comparison of classification accuracy and Misclassification rate of the original dataset after preprocessing. From the graph it can be infer that SVM attains accuracy of 97.29 percentages and Naive Bayes attains accuracy rate of 67.26 percentages for 19000 instances. Naive Bayes has high Misclassifcation rate than SVM and for 19000 instance.

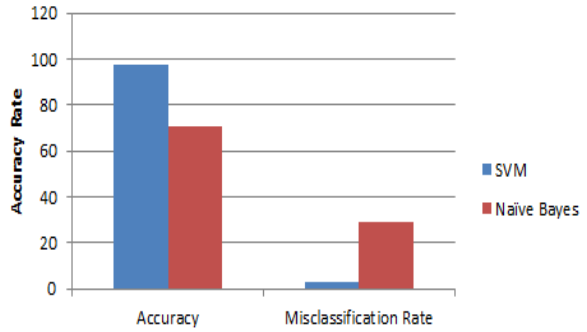

Fig 3. Accuracy and misclassification rate of SVM and Naive Bayes for 19,000 instances after Normalization

The above graph describes the comparison of classification accuracy and Misclassification rate of the dataset after Normalization. From the graph it can be infer that SVM attains accuracy of 93.85 percentages and Naive Bayes attains accuracy rate of 71.001 for 19000 instances. Naive Bayes has high Misclassification rate than SVM and for 19000 instances. the accuracy rate has been decreased for Naive Bayes for 19000 instances.

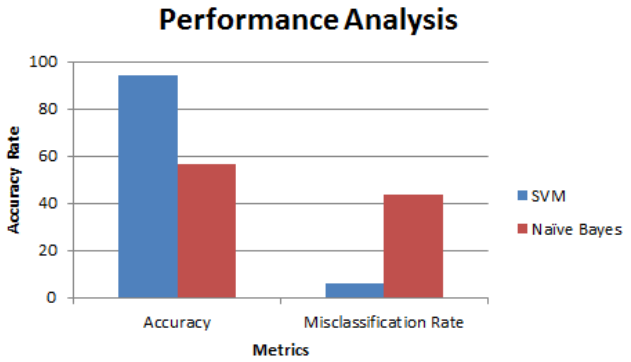

Fig 4. Accuracy and Misclassification rate after feature reduction

The above graph describes the comparison of classification accuracy and misclassification rate of the dataset after Feature reduction. From the graph it can be infer that SVM attains accuracy of 93.95 percentages and Naive Bayes attains accuracy rate of 56.54 for 19000 instances. Naive Bayes has high Misclassification rate than SVM for 19000 instances.

VII. CONCLUSION

Intrusion detection and Intrusion prevention are needed in current trends. As our regular events are mainly dependent on networks and information systems, intrusion detection and intrusion prevention are very vital. Many approaches have been applied in intrusion detection systems. Among them machine learning plays a vital role. This analysis deals with machine learning algorithms like SVM and Naïve Bayes. It proposes while dealing with 19,000 instances SVM outperforms Naïve Bayes.

VIII. FUTURE WORK

Future work deals with large volume of data, a hybrid multi-level model will be constructed to improve the accuracy. It deals with building an more effective model based on well-organised classifiers which are capable to categorise new attacks with better performance.

REFERENCES

[1]H.Wang,J.Gu,andS.Wang,‘‘An effective intrusion detection framework based on SVM with feature augmentation,’’ Knowl.-Based Syst., vol. 136, pp. 130–139, Nov. 2017.

[2]Setareh Roshan, Yoan Miche, Anton Akusok, Amaury Lendasse; “Adaptive and Online Network Intrusion Detection System using Clustering and Extreme Learning Machines”, ELSEVIER, Journal of the Franklin Institute, Volume.355, Issue 4,March 2018,pp.1752-1779.

[3]Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman , Mohd Zakree Ahmad Nazri; “Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, ELSEVIER, Expert System with Applications, Volume.66,Jan 2017,pp.296-303.

[4]Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, Aneel Raheem; “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection”, IEEE ACCESS, Survivability Strategies for Emerging Wireless Networks, Volume.6,May 2018,pp.33789-33795.

[5]BuseGulAtli1, YoanMiche,AapoKalliola, IanOliver, SilkeHoltmanns, AmauryLendasse; “Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space” SPRINGER, Cognitive Computation, June 2018,pp. 1-16

[6]Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu; “A Feature Reduced Intrusion Detection System Using ANN Classifier”, ELSEVIER, Expert Systems with Applications,Vol.88,December 2017 pp.249-247

[7]Vajiheh Hajisalem, Shahram Babaie; “A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection”, ELSEVIER, Department of Computer Engineering, Vol. 136, pp. 37-50, May 2018.

[8]Karen A. Garcia, Raul Monroy , Luis A. Trejo, Carlos Mex-Perera and Eduardo Aguirre,“Analyzing Log Files for Postmortem Intrusion Detection”,IEEE Transactions on Systems,Man, and Cybernetics, part C(Application and Reviews)42.6(2012),pp.1690-1704.

[9]R.M.Elbasiony,E.A.Sallam,T.E.Eltobely,andM.M.Fahmy,‘‘A hybrid network intrusion detection framework based on random forests and weighted k-means,’’ Ain Shams Eng. J.,vol. 4,no. 4,pp. 753–762, 2013.

[10]Hudan Studiawan, Christian Payne, Ferdous Sohel; “Graph Clustering and Anomaly Detection of Access Control log for Forensic Purposes”, ELSEVIER, Digital Investigation, Vol. 21, pp.76-87, June 2017 .

[11]Mazini, Mehrnaz, Babak Shirazi and Iraj Mahdavi; “Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms”, Journal of King Saud University- Computer and Information Sciences, 2018.

[12] Huang, G.-B., Zhou, H., Ding, X., & Zhang, R.“Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics” 42(2), 513–529, 2012.