AN EFFICIENT INTRUSION DETECTION SYSTEM USING SVM AND NAIVE BAYES ALGORITHM

[1] Bhuvaneshwari.S, Aishwarya.B[*2], Dr.Veeralakshmi.P

[1,2]Student, Department of IT, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, Tamilnadu, India,
[3]Associate Professor, Department of IT, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, Tamilnadu, India.

Abstract : ***Machine learning techniques are being widely used to develop an intrusion detection system (IDS) for detecting and volumes requiring a scalable solution. Networks play important roles in modern life, and cyber security has become a vital research area. An intrusion detection system (IDS) which is an important cyber security technique, monitors the state of software and hardware running in the network. The detection system is mainly based on feature selection and classifiers. Traditional algorithm doesn't work well with large dataset so to enhance the performance accuracy a combined method of Support Vector Machine and Naïve bayes algorithm are processed to analyse the Intrusion detection dataset. Finally, the accuracy, precision, recall, and predictions are calculated using the confusion matrix.***

Index Terms - **Support Vector Machine, Naïve bayes, Intrusion detection, attacks,cyber security.**

I. INTRODUCTION

Networks have increasing influences on modern life, making cyber security an important field of research. Network environments change quickly, attack variants and novel attacks emerge constantly.It is necessary to develop IDSs that can detect unknown attacks. To address the above problems, researchers have begun to focus on constructing IDSs using Cyber security techniques mainly include anti-virus software, firewalls and intrusion detection systems (IDSs).The IDS can be distinguished on the basis of where the detection is taking place and how or by which technique it is being detected. The IDS is classified into two segment one being Network Intrusion Detection System (NIDS) and the other being Host Intrusion Detection System(HIDS).

The first system mentioned helps in the analysis the incoming networking traffic whereas the HIDS functioning is based on the activity of the operating system. The IDS that is based on the method of detection being applied which can be classified as the Signature based (misuse) detection, which acknowledges inadequate model - anomaly detection method and analyzes the deviations of the network from a "good" model using Machine Learning techniques. There are essentially two main challenges that arise while generating an effective IDS for new attacks. First, the feature selection from the dataset is very difficult as it will tell us how important a feature can be. The feature selection changes with the change in attack type. Secondly, there does not exist a labelled traffic real-time networking. [1] Data Mining is an analysis technique that is used to analyze Big Data. Data Mining techniques were first applied to the IDS.

The main aspects of data mining on IDS that were dealt with originally were termed as clustering and classification. Since there exist no label for the initial data set for clustering issue, the object created for the clustering algorithm was allocated the same class with similar data records. The behavior of the packet was termed as a normal class or abnormal class according to the features and characteristics of already existing data.

The Classification system works on mining from the already clustered data. This implies that the data is labeled. As time has passed by, numerous techniques have been used for data mining but the recent researches are using the concept of Deep Learning and Neural Networks to implement an effective IDS. [2] Clustering is the process of creating the partition on data such that each partition or group has the same characteristic. A similar pattern is found out between the data and then on the basis of it, the data is segregated. Clustering has a significant benefit in the intrusion detection system that it can learn from the record or the audit data itself. Mini-batch K-means clustering is also an upcoming concept in the data mining field where the concept of K- means is used over IDS. Minibatch K-means algorithm's principal idea is using different random groups of distinct memory size so that it can be easy to store. Each group of data is computed under the algorithm and the output is again fed into the process. [3]

II. Related works:

The SVM is particularly attractive to analysis large dataset due to its ability to handle noise, large dataset and large input spaces and mapping of non-linear input data into a high dimensional feature space with minimum error on training set and the other algorithm is naïve bayes classifiers that include probabilistic machine learning model that's used for classification task. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

The work uses NSL-KDD data set and some common algorithms such as decision tree, random forest and deep neural network to train the model. By adjusting the proportion of training data and setting up multiple decision trees, a MultiTree and adaptive voting algorithm are proposed which obviously improve the effect of intrusion detection. It includes decision tree, random forest, KNN, DNN, and design an ensemble adaptive voting algorithm [1]

The proposed system uses CNN to select traffic features from raw data set automatically and set the cost function weight coefficient of each class based on its numbers to solve the imbalanced data set problem. The model not only reduces the false alarm rate (FAR) but also improves the accuracy of the class with small numbers. The use of standard NSL-KDD data set to evaluate the performance of the proposed CNN model show that the accuracy, FAR, and calculation cost of the proposed model perform better than traditional standard algorithms. The experimental results shows the accuracy and calculation cost of the proposed model perform better than traditional standard algorithms. [2]

This system proposes a 5-level hybrid classification system based on flow statistics in order to attain an improvement in the overall accuracy of the system. For the first level, the k-Nearest Neighbor approach (KNN); for the second level, Extreme Learning Machine (ELM); and for the remaining levels, the Hierarchical Extreme Learning approach. In comparison with conventional supervised machine learning algorithms based on the NSL- KDD benchmark dataset, the experimental study showed that this system achieves the highest level of accuracy (84.29%). Therefore, it presents an efficient approach for intrusion detection in SDNs. [3]

A novel approach called SCDNN, which combines spectral clustering (SC) and deep neural network (DNN) algorithms. To adopt SC to capture the features of complex network datasets, for attack types similar to normal access using clustering to find features that divide the dataset into subsets with different cluster centres. Six KDD-Cup99 and NSL-KDD datasets and a sensor network dataset were employed to test the performance of the model. Finally, the experimental results indicate that the accuracy, detection and false alarm rates of SCDNN are better than those of conventional methods. This technique is adapted in the proposed system.[4]

KDD'99 Dataset is used to find out which one is the best intrusion detector for the dataset. Statistical analysis on KDD'99 dataset found important issues which highly affect the performance of evaluated systems and results in a very poor evaluation of anomaly detection approaches. The most important deficiency in the KDD'99 dataset is the huge number of redundant records. It includes any redundant records in the train set as well as in the test set, more frequent records. The performances of these two approaches have been observed on the basis of their accuracy, false negative rate and precision. The results indicate that the ability of the SVM classification produces more accurate results than Random Forest and RF takes less time to train the classifier than SVM. [5]

III. Problem definition:

In recent years, advanced threat attacks are increasing. IDS has become an important and integral part of over-all security architecture. Networking has become ubiquitous in people's lives and work. Hence, network security bears increasing importance for network users and operators. The traditional network intrusion detection system is based on feature filtering which has some drawbacks that makes it difficult to find new attacks. The detection performance is tightly related to selected features and classifiers, but traditional feature selection and classification algorithms cannot perform well in massive data environment. Also, the raw traffic data are imbalanced, which has a serious impact on the classification results. In the modern network, IDS has become an important and integral part of over-all security architecture. An event or action causes breach of integrity if it allows to change the states of resources, residing in a computer in an unauthorized manner. Similarly, an event or action causes breach of availability if it prohibits legitimate users to access resources or services, residing in a computer. With the rapid growth of attacks, several intrusion detection systems have been proposed in the literature

IV. Problem description:

Support Vector Machine:

The theory of SVM is from statistics and the basic principle of SVM is finding the optimal linear hyperplane in the feature space that maximally separates the two target classes [17,18]. Geometrically, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes, which requires to solve constraint problem

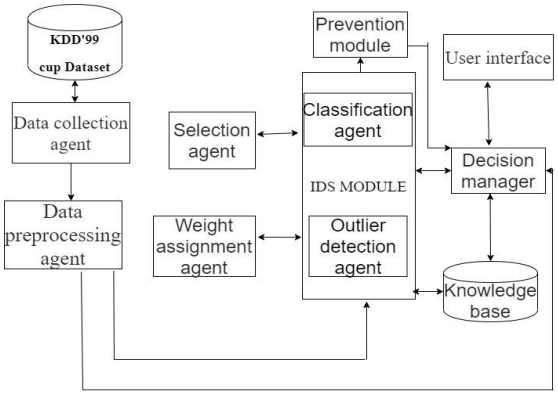

Fig1.System Architecture

Module description:

- Dataset Loading module:

Fig1.The dataset is collected from kaggle repository. It contains separately training and testing dataset fig1.1. The dataset will be in .arff format.[15] An .arff file is an ASCII text **fi**l**e** that describes a list of instances sharing a set of attributes. The KDD training dataset consists of approximately 4,900,000 single connection vectors. The feature or attribute 42 include target class as labeled [20]

In real world data are generally dirty such as it contains errors, missing value, duplicate, outlier, incomplete, irrelevant and inconsistent data. The purpose of data pre processing is to clean the noise data, extract features, and transforms the preliminary data into a format that will be more easily and effectively processed for the purpose of the user. This technique helps to improve the efficiency of the algorithm to classify the data correctly.[3]
The data preprocessing involves three stages

1- *Data cleaning*: this stage is responsible for removing any records contains a missing value an inconsistent value. It also removes duplicate records in the data set.
2- *Data transformation*: after data cleaning the next stage of pre-processing is to transform or convert the features that have text forms to numeric form to be suitable input for classification algorithms.
3- *Data normalization*: as a step of data preprocessing when datasets are too large, attribute normalization is important to detection performance. The ranges of the feature are normalized by scaling it is value so that they fall within the small specified range 0 to 1. The method min- max normalization is also applied.

DESCRIPTION

S.No.	Name of the file	Description
1	KDDTrain+.ARFF	The full NSL-KDD train set with binary labels in ARFF format
2	KDDTrain+.TXT	The full NSL-KDD train set including attack-type labels and difficulty level in CSV format
3	KDDTrain+_20Percent.ARFF	A 20% subset of the KDDTrain+.arff file
4	KDDTrain+_20Percent.TXT	A 20% subset of the KDDTrain+.txt file
5	KDDTest+.ARFF	The full NSL-KDD test set with binary labels in ARFF format
6	KDDTest+.TXT	The full NSL-KDD test set including attack-type labels and difficulty level in CSV format
7	KDDTest-21.ARFF	A subset of the KDDTest+.arff file which does not include records with difficulty level of 21 out of 21
8	KDDTest-21.TXT	A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21

Fig1.1 KDD'99 cup dataset

Feature Selection: is a process to select the most relevant feature set by removing irrelevant or redundant features. Only subsets of original features are selected.[11]

The Objectives of feature selection techniques are[14]:
-Reduce the dimensionality of feature space to avoid in curse of dimensionality.
-Speed of learning algorithms.
-Reduce large amount of data
-Less memory storage.
-Increasing the training (learning model)performance.
- Enhance Predictive accuracy by decreasing overfitting.
-Sampling the data more efficiently and improving the data quality.

- **SVM module:**

The support vector machines (SVM) are universal binary classifiers based on statistical and optimizing theories. The SVM is particularly attractive to analysis large dataset due to its ability to handle noise, large dataset and large input spaces and mapping of non-linear input data into a high dimensional feature space with minimum error on training set fig1.1

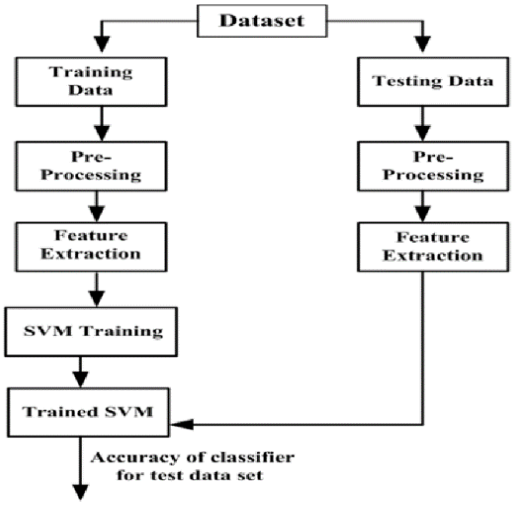

Fig1.1 Support Vector Machine

Algorithm of Support vector machine:

Algorithm: SVM classifier
Input: Dataset
Output: Accuracy and Validity

1. Start
2. Input the dataset
3. Classify the dataset
4. Apply the SVM machine learning with four kernel functions(linear, polynomial, Sigmoid and Radial Based Function(RBF))
5. Specify the Hyper-plane
6. If obtain Accuracy and Validity is NOT acceptable the goto step 4
7. End

- **Naïve Bayes module:**

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning. Bayesian classifiers a simple approach based on the inferences of probabilistic graphical models. It is very easy to construct, not needing any complicated iterative parameter estimation schemes and also predicts the class label in the fastest time[3] The Naive Bayes classifier is a supervised learning algorithm based applying Bayes' theorem. This method assumes that all the features are independent values of each other predictors.
Fig1.2. This assumption is called class conditional independence.
Using Bayesian theory:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Algorithm for naïve Bayes:

Input:
Training dataset T,
F=(f1,f2,f3,….fn) // value of the predictor variable in the testing dataset.

Output:
A class of testing dataset.
Steps:

1. Read the training dataset T.
2. Calculate the mean and standard deviation of the predictor variable in each class;
3. Repeat

Calculate the probability of f1 using the gauss density variable in each class;

Until the probability of all predictor variables (f1,f2,f3,…fn) has been calculated.

4. Calculate the likelihood for each class;
5. Get the greatest likelihood

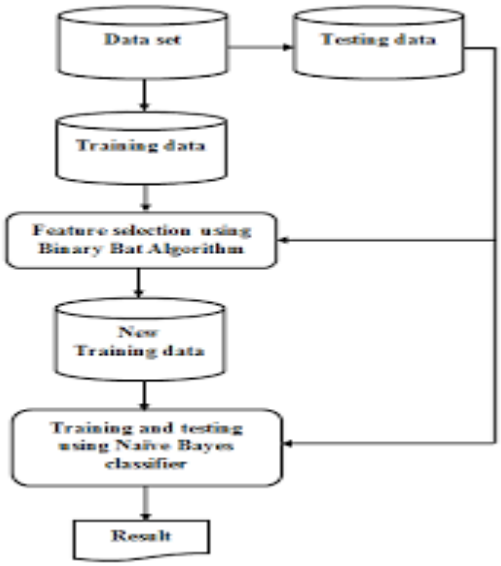

Fig 1.2 Naïve Bayes model

- **Confusion matrix module:**

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class

Classification Rate/Accuracy:
1) True Positive (TP) - Attack data that is correctly classified as an attack.
2) False Positive (FP) - Normal data that is incorrectly classified as an attack.
3) True Negative (TN) - Normal data that is correctly classified as normal. 4) False Negative (FN) - Attack data that is incorrectly classified as normal.
The following measures are used to evaluate the performance of our proposed solution:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy measures the proportion of the total number of correct classifications.

$$\text{Precision} = \frac{TP}{TP + FP}$$

The precision measures the number of correct classifications penalized by the number of incorrect classifications.

$$\text{Recall} = \frac{TP}{TP+FN}$$

The recall value can be calculated using the above formula.

Block diagram:

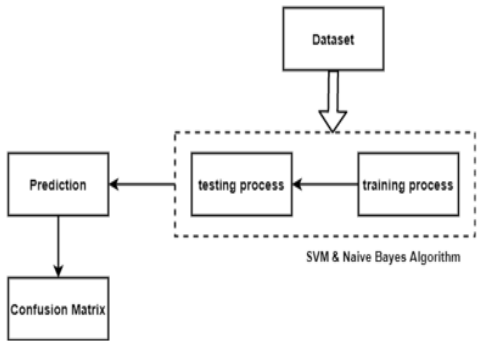

Fig2. Block diagram

V. Results and discussion:

The proposed methodology is a robust classification method and an efficient feature selection algorithm. It is analysed using the confusion matrix. The results have demonstrated that the developed approach offers high levels of accuracy, precision and recall together with reduced training time. In the existing system making use of decision tree algorithm is not good at handling large dataset. So in this proposed system support vector machine and naïve bayes algorithm has been used which is good at handling large dataset.it also gives performance accuracy in less time compared to the exisiting system. A comparison study between the Support Vector Machine, Naïve bayes, and combined model is given below:

Algorithm	Accuracy
Support Vector Machine Algorithm	79.0
Naïve bayes Algorithm	76.0
Combined	82.0

The individual accuracy of algorithm is given by Support vector Machine is 79.0 and Naïve bayes algorithm is given by 76.0 and the combined value is given by confusion matrix of 82.0 which is higher than the existing system that is 79.2 using decision tree.

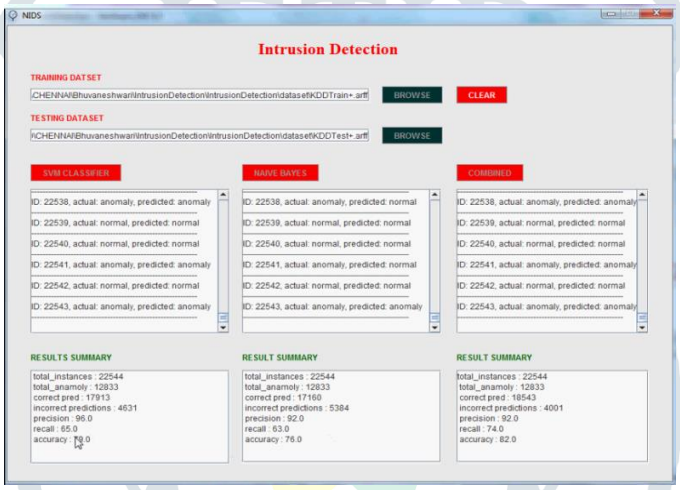

Figure 3. Experimental result analysed in java language

VI. Conclusion and Future enhancement:

This paper introduced a learning algorithm for detecting network intrusions using SVM and Naive Bayesian classifier with data mining. The algorithm is suitable for analyzing large number of network logs or audit data. The main purpose of this paper is to improve the performance of naïve Bayesian classifier for intrusion detection. The proposed system algorithm has been tested on KDD99 dataset that shows it maximized the balance detection rates for 2 classes in KDD99 dataset and minimized false positives at acceptable level. The future work focus on applying this algorithm in real time network and ensemble with other data mining algorithms for improving the detection rates in intrusion detection and also the performance accuracy of intrusion detection can be increased further.

VII. REFERENCE:

[1] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, and R. Atkinson, ``Shallow and deep networks intrusion detection system:A taxonomy and survey," 2017, *arXiv:1701.02145*. [Online]. Available: https://arxiv.org/abs/1701.02145.

[2] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, ``A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 4150, Feb. 2018.

[3] M. Tao,W. Fen, and C. Jianjun, Y. Yang, and C. Xiaoyun, ``A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks," *Sensors*, vol. 16, no. 10, p. 1701, 2016.

[4] K. Wu, Z. Chen, and W. Li, ``A novel intrusion detection model for a massive network using convolutional neural networks," *IEEE Access*, vol. 6, pp. 5085050859, 2018.

[5] H. Nkiama, S. Z. M. Said, and M. Saidu, ``A subset feature elimination mechanism for intrusion detection system," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 3, pp. 148157, 2016.

[6] "Types of Intrusion Detection System." [Online]. Available: https://en. wikipedia.org/wiki/Intrusion_detection_system
[7] K. S. Desale, C. N. Kumathekar, and A. P. Chavan, "Efficient Intrusion Detection System using Stream Data Mining Classification Technique,," in International Conference on Computing Communication Control and Automation,, 2015.
[8] K. A. I. PENG, V. C. M. LEUNG, and Q. HUANG, "Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System Over Big Data," SPECIAL SECTION ON CYBERPHYSICAL- SOCIAL COMPUTING AND NETWORKING, , 2018. [Online]. Available: 0.1109/ACCESS.2018.2810267
[9] AHMAD, M. BASHERI, M. J. IQBAL, and A. RAHIM, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection." [Online]. Available: 0.1109/ACCESS.2018.2841987
[10] Q. Niyaz, M. Alam, W. Sun, and A. Y. Javaid, "A Deep Learning Approach for Network Intrusion Detection System,," in Conference Paper in Security and Safety, 2015.
[11] S. Revathi and D. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion De-tection," International Journal of Engineering Research & Technology (IJERT), vol. 2, no. 12, 2013.
[12] "Mrutyunjaya Panda and Manas Ranjan Patra, "Network Intrusion Detection using Naive Bayes"," International Journal of Computer.
[13] "Sparsity-driven weighted ensemble classifier." [Online]. Available: https://arxiv.org/abs/1610.00270
[14] Z.WANG, "Deep Learning Based Intrusion Detection with Adversaries," SPECIAL SECTION ON CHALLENGES AND OPPORTUNITIES OF BIG DATA AGAINST CYBER CRIME, 2018. [Online]. Available: 10.1109/ACCESS.2018.2854599
[15] H. su Chae and S. H. Choi, "Feature Selection for efficient Intrusion Detection using Attribute Ratio," International Journal of Computers and Communications, vol. Volume 8, 2014.
[16] Prof.S.S. Manivannan and Dr. E. Sathiyamoorthy, "An Efficient and Ac-curate Intrusion Detection System to detect the Network Attack Groups using the Layer wise Individual Feature Set ," International Journal of Engineering and Technology (IJET).
[17] H. Nkiama, S. Z. M. Said, and M. Saidu, "A Subset Feature Elimination Mechanism for Intrusion Detection System," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. Vol. 7, no. No. 4, 2016.
[18] "Artificial Neural Networks Defination" [Online]. Available: https: //www.digitaltrends.com/cool-tech/what-is-an-artificial-neuralnetwork/
[19] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying con-volutional neural network for network intrusion detection." In ICACCI 2017, pp. 1222–1228.

[20] An Adaptive Ensemble Machine Learning Model for Intrusion Detection"; XIANWEI GAO, CHUN SHAN, CHANGZHEN HU, ZEQUN NIU, ZHEN LIU; 2019, IEEE Access