A Scalable Feature Selection Approach for IDS

[1]Bharti Harode, [2]Anurag Jain, [3]Chetan Agrawal

[1]M.Tech. Scholar, [2]Professor, [3]Professor
Computer Science and Engineering,
Radharaman Institute of Science and Technology, Bhopal, India

***Abstract*: Communication plays a vital role in everyone's life, as it is the only way for ideas and data to be exchanged amongst each other. There are various ways by which we can communicate with one another such as Telephones, Radio, Television and most recently is the Internet. With internet communication effortlessly communication is possible at anytime from anywhere to anyplace on the planet which increases the productivity of work. Computer network was developed to make the communication easier amidst individuals. The tremendous growth in network and accessibility of internet increased the security issues in the field of networking. Over the last few years as the usage of internet has increased, the no of attacks over the network has increased, its performance along with security has also been affected to multiple folds. Intrusion is set of actions that attempt to compromise the integrity, confidentiality, or availability of a network and its resource and an intrusion detection system (IDS) is a system for the detection of such intrusions. Intrusion Detection System consists of three components Data collection, normalization and classification. In this paper InfoGain, a feature selection technique is used for the reducing the size of dataset. Different classifiers viz Naïve Bayes, Random Forest and J48 were implemented using WEKA to find the algorithm with best accuracy.**

***Index Terms*: Intrusion Detection, NSL-KDD, Classification Techniques, Feature Selection.**

I. INTRODUCTION

Intrusion Detection Systems

Now a day, internet play a wide role in the technological development of society, but intrusion is a big challenge to maintain security of facts above the vertical lines. The way of identifying and tracking down-break-ins in the network security is called "intrusion detection. "Intrusion detection systems (IDS), which have long been a subject for academic research and development, are achieving mainstream demand as companies move more of their business communications to the internet. An intrusion detection system [01] can help us in giving some relevant knowledge of attacks or intrusion endeavors by identifying an intruder's movements. In this respect, intrusion detection systems are powerful mechanism in the organization's war to keep the information processing system secure.

An intrusion detection system (IDS) is made of hardware and software equipment that work together to know the unforeseen matter which may point out an assault can occur, is occurring, or has occurred[02]. Some notify at the time of attacking, some inform before the war take place, and some after the effects of attack. Simply IDS can be explains as a way to supervise phenomenon occurring in a computer system or network and examine them to point intrusion. IDS may be segment of constructed software or an outer device that monitors traffic in order to avoid unpredictable venture like malicious and illegitimate traffic, traffic violating, and certainty norms, and barriers that breach the uses of policies.

Dataset

NSL-KDD is a dataset proposed by Tavallaee. NSL-KDD dataset is a reduced form of the first KDD 99 dataset[06]. NSL-KDD comprises of indistinguishable highlights from KDD 99. The NSL-KDD data set[07] has the accompanying focal points over the first KDD data set: Redundant records are absent in the train set, in this manner the classifiers won't deliver any one-sided result. There is no copy records in the test sets; which results in better decrease rates. The quantity of records being chosen from every trouble level gathering is contrarily corresponding to the level of records present in the first KDD data set. Subsequently, the characterization rates of AI strategies shift in an immense range, which makes it progressively powerful to have an exact assessment of various learning systems. The quantity of records in the train and test sets are sensible, which makes it productive to run the examinations on the total set without the need to randomly pick a little bit which makes assessment aftereffects of various research attempts to be predictable and equivalent[09].

Feature Selection and Classification

Feature Selection [03] is where you consequently or physically select those features which contribute most to your forecast variable or yield in which you are keen on. Feature determination strategies help you in your main goal to make an exact prescient model. They help you by picking features that will give you as great or better precision while requiring less information. Feature choice techniques can be utilized to recognize and expel unneeded, superfluous and repetitive properties from information that don't add to the precision of a prescient model or may in actuality decline the exactness of the model. Data Gain evaluates the value of a property by estimating the data gain as for the class. Gain Ratio Evaluates the value of a property by estimating the gain ratio regarding the class. Relationship evaluates the value of a property by estimating the connection (Pearson's) among it and the class.

Classification [03] is technique of forecasting the class from the provided points of data. Classes are often termed as labels or targets. The forecasting modelling of classification [06][10] is the operation of evaluating a function of map (f) from input variables (X) to discrete output variables (y).

The Data Classification process includes two steps −

- Building the Classifier or Model
- Using Classifier for Classification

Credulous Bayes Classifier [01][04] Particularly dependent on Bayesian hypothesis, the Naïve Bayes Classifiers are an accumulation of arrangement calculations, where a gathering of calculations share a typical standard, for example each pair of features being ordered is autonomous of one another and the likelihood of one quality does not influence the other. Random Forest[01] It utilizes Ensemble calculations, which consolidates at least two calculations of same or diverse kind for arrangement of articles. Random forest classifier makes a lot of choice trees from randomly picked subset of preparing set. At that point total of the votes from various choice trees is done to choose the last class of the test object. J48[01] It utilizes separate and-vanquish strategy. A choice tree is made recursively dependent on the eager calculation. Choice tree so shaped is comprises of the root hub, branches, parent hubs, tyke hubs and leaf hubs. A hub in a tree indicates characteristics present in dataset.

II. LITERATURE SURVEY

A lot of work has been done in the area of intrusion detection so far. Some of the related previous researches are discussed in this paper. In [01] author proposed a filter method feature selection algorithm as a preprocessing step to improve accuracy and system performance. They developed three different class structures on NSL-KDD dataset and classification models were created and evaluated on three different class structures namely all attack types, main attack types and two attacks type. In [02] creator proposed a novel multi-classifier layered methodology, by consolidating naive bayes classifier with NBTree to improve recognition rate and exactness of minority class without harming the exhibition of greater part class. In [03] creator lead an examination and assessment on different information mining calculations to consider the presentation of every classifier against clamor free dataset and uproarious (10 percent and 20 percent) dataset. In [06] author performed their analysis on the NSL-KDD dataset with the help of figures and tables and found that NSL-KDD dataset is a best candidate data set to simulate and test the performance of IDS, as mostly the attacks are launched through the inherent drawbacks of the TCP protocol. In [12], author proposed a model "Intelligent Intrusion Detection System" which employed AI approach for detection of intrusion. The technique employed neural network, fuzzy logic, and network, with simple data mining techniques to process the data. In [13] author shows that by selecting right features and good training parameters for designing Artificial Neural Network, the performance and preciseness of an IDS can be improved. In [14] Presented a technique for intrusion detection that applies GA to detect intrusion in networks through effective feature selection. Their methodology utilizes data hypothesis to remove applicable features and decrease the multifaceted nature. At that point, they framed a direct structure rule from the chose features so as to arrange organize practices into typical and peculiar conduct. Be that as it may, their methodology thinks about just discrete features. In their work, each of the attack types includes five relevant features for classification. The rules generated are simple and the performance of the rule was evaluated using KDD CUP 1999 dataset. The generated rules specify specific attack types and hence the system is suitable to detect only specific types of attacks.

III. PERFORMANCE MEASURES

In our work we used the following performance measures:

	Predicted-1	Predicted-0
Actual-1	True Positive (TP)	False Negative (FN)
Actual-0	False Positive (FP)	True Negative (TN)

Fig. I Confusion Matrix Predictive Table

TP = Detects condition when it is present
FP = Detects condition when it is absent
FN = Does not detect condition when it is present
TN = Does not detect condition when it is absent

1. **Accuracy**: It is the most instinctive presentation measure and it is essentially a ratio of effectively anticipated perception to the all out perceptions. One may feel that, on the off chance that we have high precision, at that point our model is ideal. Truly, precision is an incredible measure however just when you have symmetric datasets where estimations of false positive and false negatives are practically same.

Accuracy = (TP + TN) / (TP + TN +FP + FN)

2. **Precision:** Precision (likewise called positive prescient worth) is the part of pertinent occasions among the recovered occurrences, What extent of positive IDs was really right.

Precision = TP / TP + FP

3. **Recall**: Recall (otherwise called affectability) is the part of significant occasions that have been recovered over the aggregate sum of pertinent examples. What extent of real positives was recognized accurately.

Recall = TP / TP + FN

4. Performance: Performance is calculated as the total time taken by the algorithm to build model and then time taken by it to test the test data with that model.

Where,

TP = no. of true positives

FP = no. of false positives

FN = no. of false negatives

TN= no. of true negatives

IV. EXPERIMENTS

All experiments were performed on a Windows platform having configuration Intel® core™ i3-3110M CPU 2.50 GHZ, 4 GB RAM. We have used Weka as our data mining tool. Weka contains a gathering of AI calculations which are helpful for information mining assignments like pre-handling, moreover with grouping, including idea of relapse, just as bunching, in addition to the affiliation guidelines, and representation. In Weka, feature selection is divided two parts : (1) attribute evaluator and (2) search method. In our work we have used InfoGain as attribute evaluator with Ranker's search as search method.

We have used Information Gain Method for reducing the size of dataset. 42 attributes were present in the actual dataset, which were reduced after applying Information Gain calculation. We have done scaling of attributes on the basis of Infogain of each attribute viz. we have divided the dataset into four parts namely dataset with attributes having InfoGain more than 10% (20 attributes), InfoGain more than 30% (16 attributes), InfoGain more than 50%(7 attributes) and complete dataset with all attributes. Scaling of attributes is shown in Table I.

Table I. Feature Selection Results

Information Gain	No. Of Attribute	Selected Attributes
Above 10%	20	3,4,5,6,12,23,25,26,29,30,31,32,33,34,35,36,37,38,39,42
Above 30%	16	3,4,5,6,12,23,25,26,29,30,33,34,35,38,39,42
Above 50%	7	3,4,5,6,29,30,42
Full Data	42	1.,2,3,…..,42
Base Paper Attributes	5	6,12,23,31,32

After feature selection step, classification of attacks to different classes of attacks is done by applying classification algorithms to all above mentioned four dataset groups. In our work we have employed and compared results of three single different classifiers viz. **Naïve Bayes, Random Forest and J48.** Evaluation results of classification models are shown in Table II, Table III, Table IV.

Table II. Classification Results

Information Gain	No. Of Attribute	Precision(%)		
		Random Forest	***J48***	***Naïve Bayes***
Above 10%	20	0.841	0.844	0.798
Above 30%	16	0.825	0.852	0.797
Above 50%	7	0.837	0.880	0.804
Full Data	42	0.852	0.858	0.809
Base Paper Attributes	5	0.794	0.795	0.706

Table III. Classification Results

Information Gain	No. Of Attribute	Recall(Sensitivity)(%)		
		Random Forest	***J48***	***Naïve Bayes***
Above 10%	20	0.793	0.791	0.734
Above 30%	16	0.782	0.805	0.732
Above 50%	7	0.819	0.858	0.698
Full Data	42	0.805	0.815	0.761
Base Paper Attributes	5	0.749	0.749	0.572

Table IV shows the accuracy of each classifier for each subset of NSL-KDD dataset. From the table we can observe that the highest accuracy is achieved with J48 algorithm in the case where we have taken attributes with InfoGain more 50%. We can also see that varying no. of attributes is also varying the accuracy directly, lesser attributes, higher will be the accuracy.

Table IV. Classification Results

Information Gain	No. Of Attribute	Accuracy (%)		
		Random Forest	***J48***	***Naïve Bayes***
Above 10%	20	79.2628%	79.0720%	73.3632%
Above 30%	16	78.2381%	80.5358%	73.1813%
Above 50%	7	81.8666%	85.7878%	69.8456%
Full Data	42	80.4516%	81.5339%	76.1178%
Base Paper Attributes	5	74.8581%	74.8625%	57.217%

Table V. Performance Results

Information Gain	No. Of Attribute	Performance Results(Seconds)		
		Random Forest	***J48***	***Naïve Bayes***
Above 10%	20	65.52/1	13.69/0.45	0.67/0.67
Above 30%	16	52.89/1.14	10.3/0.31	0.58/0.83
Above 50%	7	28.22/1	3.08/0.06	0.17/0.22
Full Data	42	142.85/1.44	40.35/0.5	1.17/1.09
Base Paper Attributes	5	49.77/1.26	4.27/0.23	0.33/0.45

Training and test time results are shown in Table V. According to the results we have achieved the higher value of accuracy, sensitivity and precision values of our tested data. The Accuracy, sensitivity and precision values that are produced using J48 classification algorithm are best with feature selection. It also gives the best result in terms of performance. Therefore, we can state that the performance can be improved by doing scaling of data by using feature selection algorithms.

IV. CONCLUSION

In this paper, the presentation of different characterization calculations has been thought about and assessed based on NSL-KDD dataset. Different characterization calculations like NB, J48, RF from various grouping calculation families were tried and looked at. Feature determination is a significant procedure that includes limiting the quantity of important features for augmenting the prescient intensity of the model. This decreases the dimensionality of feature space, expels repetitive, unimportant, or loud data. It brings the quick impacts for application: accelerating a data mining calculation, improving the data quality and thereof the exhibition of data mining, and expanding the intelligibility of the mining results. It has been seen that choice of features improves the exhibition of the model. The consequences of our examinations and assessment urge us to continue further research on different half breed IDS procedures. Future work may include different feature selection method to increase the accuracy and performance of classification algorithms needs assistance to decrease the time taken by models to make them faster with a precise result. Exploration of different revealed characterization calculations against genuine system traffic alongside the impact of different feature choice method will be the focal point of our future works.

REFERENCES

[01] Ayse Gul, Esref Adali" A feature selection algorithm for IDS" 2nd International Conference on Computer Science and engineering 978-1-5386-0930-9/17/$31.00 © 2017 IEEE

[02] Neelam Sharma, Saurabh Mukherjeeb "A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in IDS " 2nd International Conference on Communication, Computing & Security (ICCCS-2012)

[03] Jamal Hussain and Samuel Lalmuanawma "Feature analysis, evaluation and comparisons of classification algorithms based on noisy intrusion dataset" 2nd International Conference on Intelligent Computing, Communication & Convergence(ICCC-2016)

[04] Dr. Saurabh Mukherjeea, Neelam Sharmaa "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" © 2011 Published by Elsevier

[05] Yasmen Wahba1, Ehab ElSalamouny 2 and Ghada ElTaweel "Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction" IJCSI International Journal of Computer Science Issues, Volume 12, Issue 3, May 2015

[06] L.Dhanabal1, Dr. S.P. Shantharajah" A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015

[07] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009) IEEE2009

[08] Debar, H., Dacier, M., and Wespi, A., A Revised taxonomy for intrusion detection systems, Annales des Telecommunications, Vol. 55, No. 7–8, 361–378, 2000.

[09] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, October 2007.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and a. Alonso- Betanzos, "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," Expert Syst. Appl., vol. 38, no. 5, pp. 5947–5957, 2011.

[12] Norbik Bashah, Idris Bharanidharan Shanmugam, and Abdul Manan Ahmed," Hybrid Intelligent Intrusion Detection System" World Academy of Science, Engineering and Technology, 2005

[13] Saman M. Abdulla, Najla B. Al-Dabagh, Omar Zakaria, Identify Features and Parameters to Devise an Accurate Intrusion Detection System Using Artificial Neural Network, World Academy of Science, Engineering and Technology 2010.

[14] A. H. Sung, S. Mukkamala. (2004) The Feature Selection and Intrusion Detection Problems. In Proceedings of the 9th Asian Computing Science Conference, Lecture Notes in Computer Science 3029 Springer 2004, pp.

[15]S Zaman, F Karray Features selection for intrusion detection systems based on support vector machinesCCNC'09 Proceedings of the 6th IEEE Conference on Consumer Communications and Networking Conference 2009