A Novel Approach for Feature Selection Technique to Identify Intrusion Related Data-A Survey

[1]YUVRAJ D PAWAR, [2]SATHYA PRAVEEN D

[1,2]Dept of Computer Science & Engineering, BAMU, Shreeyash College of Engineering, Aurangabad, India.

***Abstract*—The development of intrusion detection systems (IDS) depends on pre-processing and selection of important data properties. Another important factor is the design of an effective learning algorithm that is organized into normal and abnormal patterns. The purpose of this research was to propose a new and better Naive Bayes assortment to improve intrusion detection accuracy in IDS. The proposed classification should take less time compared to existing classifiers. In order to obtain accurate and fast network data processing, this study uses three standard features. This study tested the effectiveness of the proposed new classification algorithms. Bayes Naïve, J48 and REPTree, which measure different performance parameters using 10-fold cross-fold detection, are evaluated using this classifier. The empirical results show that the improved version of the Naive Bayes break gives better results in terms of Intrusion Detection and Error Rate.**

Keywords— Machine learning; Intrusion Detection System (IDS); Naïve Bayes algorithm; Feature selection; NSL KDD data set.

I. INTRODUCTION

The chances of data loss, hacking and intrusion have increased as the use of internet and popularity. Continued Internet attacks are a serious challenge in developing flexible and adaptable security approaches. Intrusion can be defined as a set of actions that compromises the integrity, confidentiality, or availability of computer resources. [1, 2] Intrusion Detection System (IDS) is the most important element used in the detection. Attacks Internet attacks, which may be either hosted or network-based. [3,4] Intrusion detection is the process of monitoring and analyzing activity occurring in a computer system. [5] In literature, various techniques are used to develop effective IDS. But such techniques often have some drawbacks. Traditional intrusion prevention techniques such as firewalls, access control, or password-based encryption can not fully protect networks and systems from severe attacks and malware. Research for intrusion detection is based on machine learning. This technique has the ability to detect independent packets with high detection rates and low false positives, while the system can quickly adapt itself in a dynamic environment. One of the major problems in network intrusion detection systems is the amount of data generated and collected by network users. As the number of Internet users grows, the data generated is increasing day by day in the computer network and decreasing the capacity of the IDS. [7] The appropriate feature set must be identified by the feature separation. This reduces the processing time and improves detection accuracy in the IDS [8, 9]. The present document offers an improved Naive Bayes (NB) rating. Tung, which can overcome the shortcomings of existing Naïve Bayes algorithm and provides more accurate and more precise in detecting intrusion. For the selection of features, this study has applied the technique of selecting features such as the selection of attributes based on relationships, evaluation of attributes, perception of information, evaluation of attributes, profit Using a feature selection technique removes irrelevant or useless features that do not contribute much to intrusion detection. The proposed version of the Naive Bayes breaks is tested using the NSL KDD datagram to detect attacks under four major categories: Probe, DoS, U2R, And R2L (remote to local). The proposed version will be compared to the existing classifier.

Intrusion Detection (IDS) is a powerful security technology that detects, blocks and responds to malicious activity on a computer. [10, 11] ID examines and analyzes statistics for network activity. [12] IDS can be used to detect types of malicious network communications and computer systems use. Because assembly techniques such as firewalls are vulnerable to attacks, they tend to be prone to errors in the case of faulty configurations or vague security policies. [13, 14] , 15]. So, to overcome the problem of traditional intrusion detection methods, Machine Learning (MLM) has introduced machine learning into the field of artificial intelligence. [6, 17]. The main motivation in learning to machine learning is to learn automatically in a computer-based learning environment. Pattern recognition and complex rules to make informed decisions based on historical data and past experience. In the context of intrusion detection, detection patterns learn from the previously recorded attack patterns, (Called a signature) and

 DOI : 10.18231/2454-9150.2018.0859

detect similar objects in the incoming traffic that have never been seen before.

We can consider several factors that affect the success rate of IDS, which is based on the machine learning classifier in a given environment. One such factor is the representation and quality of information that we will use for intrusion detection. In theory, having a lot of information with additional features and features will result in more discrimination and more accuracy. But in practice, many machine learning mechanisms have shown that this is not always true. Given a set of multiple properties, the learning algorithm produces a biased estimate of the probability of the class label. [20] If the database contains irrelevant and excessive data, learning during the training is more difficult redundant data directly leads to over fitting problems and overall system performance decreases. Naive Bayes may be affected by redundancy due to its assumption that the classes that specify the attributes are independent. Decision tree algorithms such as C4.5 can send training data to a large tree size. It has often been seen that the removal of irrelevant and redundant data on the production of trees is small by C4.5 algorithm. [19] [20] Therefore, all the problems mentioned above can be solved by the attribute selection technique or the attribute selection technique. Feature selection is used for intrusion detection to eliminate redundant and irrelevant information. This means the process of selecting a subset of related attributes that fully describes the problem with minimal degradation. [14] In the algorithm selection phase, the algorithm finds the best subset of the attributes in the set. Your information automatically a subset of the data set is provided with an algorithm for creating a subset.

Loop is used until enough attributes are selected from the dataset without affecting system performance. The evolutionary function of the subset is used to track the algorithm of this activity. The whole process of selecting features is shown in Figure 1.

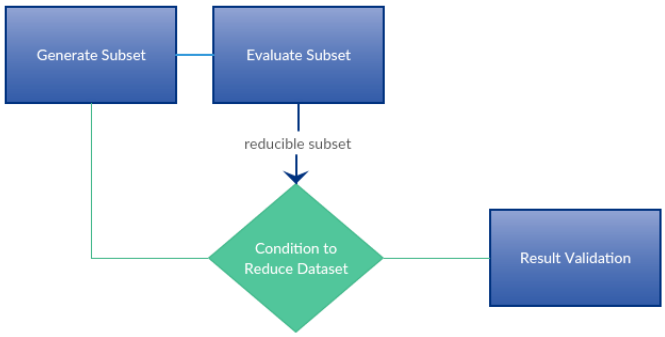

Figure 1 Feature Selection

Feature selection

In this study, we used three standardized qualitative methods in the Naïve Bayes re-classification process, which we proposed. They will select the related attributes. The Gain Information, Gain Ratio, using the first search. And strategic alignment.

Evaluate the value of a feature group by referring to the individual predictability of each feature, along with the possibility of repetition between attributes. CFS evaluates attribute that are highly correlated with the class. but it's not about each other.

Information Gain Attribute Evaluates the value of an attribute by measuring the information received from the class. Data reception is based on the concept of entropy, which is widely used in the domain of information theory. Give a set of instances S that contain positive and negative examples of certain concepts.

The profit margin is a complement to the information given above. Trying to overcome data reception and want to choose a lot of valuable features. So we can say that the Gain Ratio feature evaluator is more accurate under some problems where the data is well organized and there is no overlap. [18, 19]

II. RELATED WORK

In 1987, Denning presented models for the development of IDS, based on the Markov Series, Time Series, etc., in the Denning IDS format, identifying normal users and malicious users based on behavior such as if user behavior deviated from normal Behavior is abnormal. [14]

The first IDS to achieve this goal in real time was developed in early 2007. Prof. Chou et al. Proposed a dynamic "intrusion detection system" based on a specific artificial inventive method, such as neural and fuzzy systems for invasive detection of Chou et al. [16] A number of hybrid techniques have been used in the field of machine learning to solve the problem of selecting features in a network. Intrusion Detection Hybrid methods of classifying and classifying genes, fusions, or fuzzy genes together to increase the effectiveness of IDS. [16] [17]

Al-Dabagh et al. [18]Demonstrated that improving the accuracy and efficiency of IDS can be improved by selecting effective neural network models (ANNs) and training parameters.

K Franke et al. (CFS) [19]provides a methodology for the selection of relational and cognitive features that work automatically and efficiently using continuous and sequential features.

Abraham A. et al.[20] Bayesian Integration and Tee Regression and Regression, and proposed a hybrid model for feature selection algorithms that provide better results in identifying unknown attacks.

Panda [21]and the team propose an intelligent hybrid approach, using a combination of filtering with the classifier to make intelligent decisions to maximize overall IDS performance.

 DOI : 10.18231/2454-9150.2018.0859

Saurabh et al.[22] Demonstrate the importance of selecting features for efficient and effective intrusion detection systems. They propose a powerful attribute reduction (FVBRM) method to identify the weakest set of key input features using the NSL-KDD dataset.

Alhaddad Mohammed J et al[23]. conducted experiments to study the application of different classification methods and the effects of using classification machines in classification and accuracy.

Axellson proposed [24] implicit and error rates for intrusion detection systems that worked on the principles of Bayesian rule of probability conditions.

E.Nutu Lutu offers Naive Bayes (NB) classification [25] for classification. Stream mining is a data mining, sequential and sequential data mining process in real time. The performance of the naive classifier classification has been improved by eliminating unrelated properties from the modeling process.

III. PROPOSED ARCHITECTURE

Although the Naïve Bayes algorithm yields satisfactory results. It has some flaws, such as Bayes Naive. It is assumed to be very independent, for example, assuming that all features are independent of each other. Based on a literature survey, we have learned that a great deal of effort has been made to improve the Naïve Bayes identifier using two methods: subset selection and systematic independence hypothesis. In this study, the author presents a new algorithm that works both ways. In the current work, the Naive Bayes classification algorithms are updated without the assumption that the independence of the terms of the attributes is different. The improved algorithm relies on the Corr (Xi, C) formula defined in Equation (1.5). In the next step, we will change the order, such as the set X. The X series is sorted from | Corr (Xi, C) | from the given set X *. The arc from the first property set is merged with the second set of properties. Finally, for all remaining features, we calculate the conditional probability of each feature with the help of the previous feature, using the class of the order X *. The maximum probability value between all the calculated features, Used to distinguish the parent of each feature from the ball. Correlation coefficients between random variables X i and X j.

- To select feature from initial input that maximizes the input I(C;fi) and minimizes the average of redundancy MRs simultaneously.

where I(C;fi) is the amount of information that feature fi carries about the class C.

$$G_{MI} = \arg\max_{f_i \in F}(I(C; f_i) - \frac{1}{|S|}\sum_{f_s \in S} MR),$$

MR, is the relative minimum redundancy of feature fi against feature fs

$$MR = \frac{I(f_i; f_s)}{I(C; f_i)},$$

where fi belongs to F and fs belongs to S.

- In case I (c;fs) = 0 then the feature can be discarded without computing.
- In case fi and fs are high, feature will contribute to redundancy to reduce the number of features that need to be examined, a numerical threshold (Th = 0) value is applied to GMI.

A. GMI should have following properties :

1. If (GMI = 0), then the current feature fi is irrelevant or unimportant to the output C because it cannot provide any additional information to the classification after selecting the subset S of features. Thus, the current candidate fi is removed from S.
2. If (GMI > 0), then the current feature fi is relevant or important to the output C because it can provide some additional information to the classification after selecting the subset S of the feature. Thus, the current candidate fi is added into S.
3. If (GMI < 0), then the current feature fi is redundant to the output C because it can cause reduction in the amount of MI between the selected subset S and the output C. Thus, feature fi is removed from S.

Proposed Algorithm
Input : Set of features F = { fi, i =1,.....,n}
Output : S - Selected Feature
Start
Initialize set s = EMPTY
Calculate information carried by class, denoted as I (C) for every feature
for(each feature i=1 ..... to n)
n_f = n ;
Select the feature f_i such that arg max(I(C,f_i)),i = 1,.....,n
Set Feature Selected as **F**,
end for

Calculate GMI and find fi such that

while F **is not empty do**
if(GMI >0) then
Add Feature f to set S
s <- S **union** {f_i}
end if

end while

Based on features classified we will develop a set S and process further.

We will use Weka SVM classifier to classify data from NSDL and KDD dataset.

DOI : 10.18231/2454-9150.2018.0859

Step 1: Allocate data that is defined in a class of approximate size. Since the data set has a very unequal class, the proposed study used a k-fold cross

Step 2: Create the primary structure of the Naive Bayes identifier X = {X1, X2, ...Xn}.

Step 3: Calculate the relationship and type between each attribute Xi, i = 1 ..... n of all classes using Corr Correlation Coefficient (X i Y i).

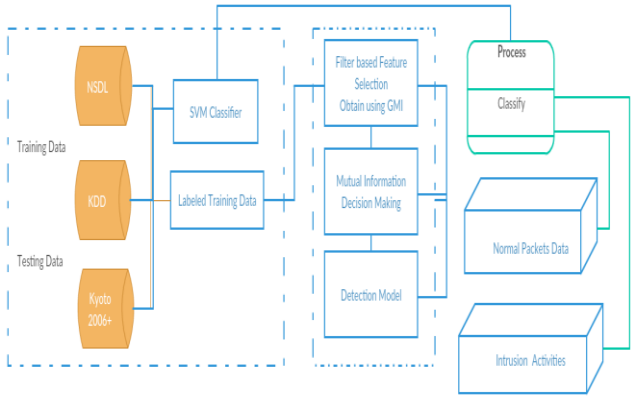

Figure 2 Proposed Architecture

Steps involved

1. Select dataset
2. Load Training set
3. Train SVM using training set
4. Load Master Dataset NSDL, KDD, Kyoto
5. Obtain Feature set F using f_i
6. Build Selection Set S
7. Using detection model to filter data records
8. Classify test set using selected features.

- **Experimental Setup**

For experiments, we have used the updated NSDKDD data set, which consists of a complete set of KDD datasets. The training kit used for experimental purposes has a broken attack. 21 out of a total of 37 sets in the test suite. The NSL KDD data set consists of 41 attributes and five classes, which are normal and the other four are attack types.

Dataset used : KDD99 is preliminary dataset and NSL-KDD is a new revised version of the KDD Cup 99.

Like KDD Cup 99 dataset, each record in the NSL-KDD dataset is composed of 41 different quantitative and qualitative features.

KDD Cup 99 and NSL-Kdd includes three different sets:

training (the "10 percent KDD Cup 99" data and "KDDTrainþ" respectively),

test ("kddcup testdata" and "KDDTest" respectively)

The NSD KDD dataset we used for our study intrusion detection was 125973 and 42 attributes. Select Features Although our focus algorithm is better than the existing classifier and its performance is consistent with the number of relevant attributes that are contained in the dataset, Table 1 shows the results based on Of binary classes such as attackers or normal users.

Classifiers	Accuracy	Precision	F-measure
Naïve Bayes	45.60	83.13	81.44
J-48	91.67	87.70	91.12
REPTree	89.12	88.31	72.21
Proposed	91.20	96.23	97.66
Algorithm			

Table 1 Classification Results

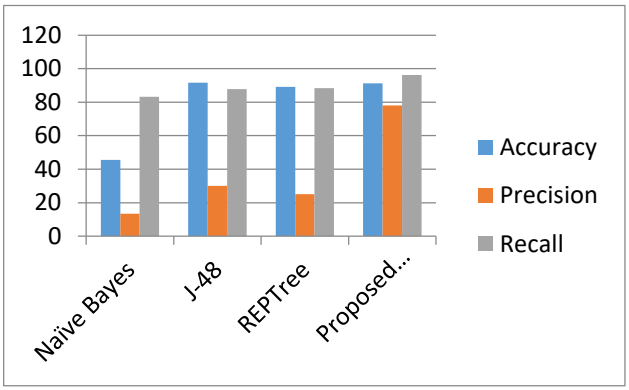

FIGURE3 GRAPHICAL RESULTS

IV. KEY INDEX PARAMETERS FOR RESULT CLASSIFICATION

In information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, whereas recall (also called sensitivity) is the fraction of the relevant instances which are recovered. Accuracy and recall are therefore based on an understanding and measurement of relevance.

In simple terms, high accuracy means that an algorithm returned results significantly more relevant than irrelevant, while high recall means that an algorithm returned most relevant results.

True positive (TP): Classifying an intrusion as an intrusion. The true positive rate is synonymous with detection rate.

False positive (FP): Incorrectly classifying normal data as an intrusion also known as a false alarm rate.

True negative (TN): Correctly classifying normal data as normal, it true negative rate is also referred to as specificity.

False negative (FN): Incorrectly classifying an intrusion as normal.

The most important category measurements for binary categories are:

Precision	$P = TP/(TP + FP)$
Recall	$R = TP/(TP + FN)$

V. CONCLUSION

We have worked with the Naive Bayesian learning paradigm because it assumes a very independent feature between attributes that offer a new algorithm that

 DOI : 10.18231/2454-9150.2018.0859

approximates interactions between features using probability. Condition Performance comparison between different classifiers with classifiers is made to understand performance in terms of performance measures. Based on the findings, it was found that all the features in the data set were not equally important because we could ignore certain attributes above anything else that was not related to intrusion detection. Therefore, this study uses the technique of selecting features and found better results than before. The results show that a subset of features identified by the gain + Ranker have improved our Naïve Bayes classification. In the future, we will try to use feature selection using soft-calculation techniques to identify invasions.

REFERENCES

[1] Chih-Fong Tsai a, Yu-Feng Hsu b, Chia-Ying Lin c, Wei-Yang Lin d "Intrusion detection by machine learning A review" Expert Systems with Applications Elsevier 2009.

[2] Tanya Garg and Surinder Singh Khurana IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, 2014, Jaipur, India

[3] Jian Pei Shambhu J. Upadhyaya Faisal Farooq Venugopal Govindaraju. Proceedings of the 20thInternational Conference on Data Engineering published In IEEE 2004.

[4] 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.

[5] Gulshan Kumar, Krishan Kumar & Monika Sachdeva (2010) "The use of artificial intelligence based techniques for intrusion detection: a review" Published online: 4 September 2010 © Springer Science+Business Media.

[6] Biesecker, Keith, Elizabeth Foreman, Kevin Jones and Barbara Staples (2008) "Intelligent Transportation Systems (ITS) Information Security Analysis." United States Department of Transportation Technical Report FHWA-JPO-98-009, 16 November 2008.

[7] Siva S. Sivatha Sindhu, Geetha , A. Kannan " Decision tree based light weight intrusion detection using a wrapper approach ".Expert Systems with Applications 39 (2012) 129–141 published in Elsevier

[8] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsin "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment". Procedia Computer Science 3 (2011) 1237–1242

[9] F. Maggi, M. Matteucci and S. Zanero, "Reducing false positives in anomaly detectors through fuzzy alert aggregation". Information Fusion, 10, 300–311. 2009

[10] C-C. Lin and M-S. Wang, "Genetic-clustering algorithm for intrusion detection system. International Journal of Information and Computer Security", 2, 218–234. 2008

[11] Dr. Saurabh Mukherjee, Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" Published by Elsevier 2012.

[12] O. Y. Al-Jarrah1, A. Siddiqui1, M. Elsalamouny, P. D. Yoo1, S. Muhaidat1, K. Kim "Machine- Learning-Based Feature Selection Techniques for Large- Scale Network Intrusion Detection" 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops.

[13] Heberlein, L. Todd, Dias, Gihan V, Levitt, Karl N, Mukherjee, Biswanath, Wood, Jeff, and Wolber, David, "A Network Security Monitor," 1990 Symposium on Research in Security and Privacy, Oakland, CA, pages 296-304

[14] Paxson, Vern, Bro, "A System for Detecting Network Intruders in Real-Time," Proceedings of The 7th USENIX Security Symposium, San Antonio TX, 1998.

[15] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey ," Intrusion Detection Using Data Mining Techniques" 978-1-4244-5651-2/10/$26.00 ©2010 IEEE

[16] PAT LANGLEY, STEPHANIE SAGE," Induction of Selective Bayesian Classifiers" Institute for the Study of Learning and Expertise 2451 High Street, Palo Alto, CA 94301 [17] ENGEN, "Machine learning for network based intrusion detection," Doctoral dissertation, Bournemouth University, 2010.

[17] 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.

[18] H Nguyen, K Franke, S Petrovic "Improving Effectiveness of Intrusion Detection by Correlation Feature Selection" , 2010 International Conference on Availability, Reliability and Security, IEEE Pages-17-24

[19] A Abraham, S Chebrolu, J P. Thomas "Feature deduction and ensemble design of intrusion detection systems" Computers & Security,Volume 24, Issue 4, June 2005, Pages 295-307

[20] Panda, Mrutyunjaya, Ajith Abraham, and Manas Ranjan Patra, "A Hybrid Intelligent Approach for Network Intrusion Detection," International Conference on Communication Technology and System Design 2011,Procedia Engineering 30 (2012), 1-9. [42] Saurabh, Mukherjee, and Neelam Sharma, "Intrusion detection using naive Bayes classifier with feature reduction," Procedia Technology 4 (2012), 119-128 doi: 10.1016/j.protcy.2012.05.017

[21] Alhaddad, Mohammed, Amir Ahmed, Sami M. Halawani "A study of the modified KDD 99 dataset by using classifier ensembles approach," IOSR Journal of Engineering, May. 2012, Vol. 2(5) pp: 961-965.

[22] S. Axelsson, "The base rate fallacy and its implications for the difficulty of Intrusion detection" Proc. Of 6th. ACM conference on computer and communication security 1999.

[23] Patricia E.N. Lutu, "Fast Feature Selection for Naive Bayes Classification in Data Stream Mining," Proceedings of the World Congress on engineering, Vol III, WCE 2013.

[24] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification," 2007.

[25] S Chebrolu, A Abraham, J P. Thomas Feature deduction and ensemble design of intrusion detection systems, Computers & Security,Volume 24, Issue 4, June 2005.

 DOI : 10.18231/2454-9150.2018.0859