Comparison of NBTree and VFI Machine Learning Algorithms for Network Intrusion Detection using Feature Selection

Rupali Malviya
M.Tech Student
Dept. of Computer Science and Engineering
UIT, Allahabad, India

Brajesh K. Umrao
Assistant Professor
Dept. of Computer Science and Engineering
UIT, Allahabad, India

ABSTRACT
The security of computer networks is of great importance. But, with the proliferation of electronic devices and the internet, there has been an exponential rise in malicious activities. The security perpetrators take the advantage of the intricacy of the internet and carry out intrusions. There have been certain researches to find out solutions for detecting intrusions. In this paper, the research has been the application of machine learning techniques to the field of network intrusion detection. Machine learning techniques can learn normal and anomalous patterns from training data and generate classifiers which can be used to detect intrusions in a network. The machine learning techniques used are Naïve Bayes Tree algorithm and the Voting Feature Intervals algorithm. Also, Feature Selection Methods to improve the performance of these algorithms were used because the input to classifiers is in a high dimension feature space, but all features available are not relevant for classification. Two approaches were taken into consideration for feature selection, Chi Square and Gain Ratio. Using these feature selection approaches a comparative study of the two algorithms NBTree and VFI as classifiers has been done. The NSL-KDD data set has been used to train and test the classifiers.

Keywords
Machine learning, NBTree, VFI, Feature selection, Chi Square, Gain Ratio.

1. INTRODUCTION
With the decreasing cost of information processing and Internet accessibility, there has been an increase in more and more organizations becoming vulnerable to a wider variety of cyber threats. New and complicated methods of attacks are being developed by the attackers owing to the rapid expansion of the internet, the intricacy of communication protocols and anonymity on the internet. According to a survey done by CERT/CC i.e. Computer Emergency Response Team/Coordination Center [9], the rate of cyber attacks is more than doubling every year in recent times. An organization could suffer immense loss if its systems and networks are attacked. Therefore, it has undoubtedly become important to make the systems robust, especially those that are used for critical functions like in the military of a nation, or commercial sectors which keep highly classified information on their systems. Attackers, most of the time, harm the systems and the networks by intruding into them. Intrusion detection is a much concerned problem in the current day scenario. It includes identifying those actions that could be an act of intrusion. The activity of intrusion compromises the three goals of security, i.e. the integrity, confidentiality, and availability. The tralatitious methods for intrusion detection rely on the panoptic knowledge of signatures of the attacks that are known. Events taking place on a system are monitored and logged. These logged events are compared with the signatures of already known attacks to detect intrusions. If the signatures of the two are matched an intrusion is said to be detected. Features are extracted from several audit streams, and intrusions are detected by comparing the feature values to a set of attack signatures provided by human experts working in the information security domain in the organization. The signature database used is to be revised manually for each and every new kind of intrusion that the information security personnel come across. Another limitation of signature-based methods is that they are unable to detect emerging cyber threats. This is because of their inability to detect novel attacks. Also, after a new attack is known, its signature is developed and most of the time there is a considerable lag in updation of signature database and deployment across the network. These limitations of signature-based intrusion detection methods have attracted the researcher's interest in intrusion detection techniques which are based upon machine learning [10, 13].

This paper describes the use of two of the machine learning techniques which can be employed for detection of intrusion in the computer networks. But before employing these machine learning algorithms feature selection was performed on the data set using two different methods. These methods have helped reduce unwanted features and select the most significant ones which determine the most accurate categorization of the data. Feature selection, is a preprocessing step for choosing a subset of relevant features for building robust learning models [1, 7, 11]. It is the process of choosing a subset of original features in such a way that the feature space is reduced in an optimal manner to evaluation criterion. The crude data collected is generally very large, so it is required to select a subset of data by creating vectors of features which represent almost all of the information from the data. The current existing methods of feature selection for machine learning fall into two categories. First, there are methods which evaluate the worth of features using the learning algorithm which is to be ultimately applied to the data. These methods are called wrappers. Second, those which evaluate the worth of features by using heuristics based on general characteristics of the data. These are called filters [8, 12].

2. CLASSIFICATION
The classification of large data sets is an important problem in machine learning. For a database with millions of records and large no. of classes, such that each record belongs to one of the given classes, the problem of classification aims to determine the class to which a given record belongs. In supervised type of classification, there is a set of records

called the training data and for each of the record of this set, the respective class which it belongs to is also given. Using this training set, the classification process attempts to generate the descriptions of the classes, and these descriptions help to classify the unknown records [19]. There are several approaches to supervised classifications. Naïve Bayes Tree and Voting Feature Intervals are two of them.

2.1 Naive Bayes Tree Algorithm

The NBTree algorithm is a hybrid of the Naïve Bayes [2] and the Decision Tree algorithm [3]. The learned knowledge is represented in the form of a tree. This tree is constructed recursively. But, the leaf nodes are Naive Bayes categorizers [4]. In order to limit the entropy measure a threshold is chosen for continuous attributes. For finding a node utility, data is discretized and 5-fold cross validation accuracy estimation is computed using Naive Bayes at the node. The utility of a split is the weighted sum of utility of the nodes. Also, this is dependent upon the number of instances going through that node. The NBTree algorithm strives to approximate whether the generalization accuracy of Naive Bayes at each leaf is higher than a single Naive Bayes classifier at that node. A split is considered to be significant if relative reduction in the error is greater than 5% and there are a minimum of 30 instances in the node [4]. For discrete valued attributes, the Naive Bayes method performs quite well. With the increase in data size, the performance also improves. But in case of continuous valued attributes, Naive Bayes method does not take into account the attribute interactions. Whereas, the decision trees do not give good performance when the data size is very large. These shortcomings are overcome by the NBTree algorithm [16].

2.2 Voting Feature Intervals

The VFI algorithm [6] is a classification algorithm which is based on the concept of voting frequency intervals (therefore the name given VFI). In this algorithm, every training instance is represented as a vector of features. This also has a label which represents the class of that instance. Then for each feature, feature intervals are constructed. A set of values for a given feature where the same subset of class values is observed, is represented by an interval. Therefore, two adjacent intervals represent different classes. Two phases are there, training phase and the classification phase. In the training phase, the feature intervals are to be found. These are calculated by calculating the lowest and highest feature value for each linear feature for each class. The observed feature values are taken into consideration for nominal features. For every linear feature with k classes, 2k values are found. These are then sorted and every pair of consecutive points forms a feature interval and point intervals are formed for nominal values. Every interval is represented in the form of a vector as (lower, $count_1$, $count_2$....$count_k$), where lower denotes the lowest feature value and count i denotes the number of training instances of class i which fall into that interval. Next, in the classification phase, the interval i of a new instance e is found out. Then, for every class a feature vote is calculated. These votes are normalized and the class that has the highest feature vote is the class predicted for the new instance.

3. FEATURE SELECTION

Feature Selection is also known as subset selection, attribute selection or variable selection. It is an important preprocessing step used in machine learning, where a subset of the features that are available in the original data is selected for subsequent application of a learning algorithm [5]. Feature Selection is necessary because it is computationally infeasible to use all the features in hand. There is a curse of dimensionality, which refers to the fact that the number of data samples required to estimate some arbitrary multivariate probability distribution increases exponentially and the number of dimensions in the data increases linearly. Selection is a very important step in classification with a search procedure to find the optimal feature set [19]. Also, feature selection methods need to use some sort of evaluation function together. The evaluation functions can be divided into two main groups: Filters and Wrappers. Filters measure the relevance of feature subsets independently of any classifier, whereas wrappers use the classifier's performance as the evaluation measure. In this paper, two different approaches for feature selection have been considered, Chi Square and Gain Ratio which are based on filter approach.

4. EXPERIMENTS AND RESULTS

For performing intrusion detection, NSL-KDD data set was selected. This data set is an improvement over the commonly used data set KDDcup 1999 data set. The KDD dataset had certain inherent problems which are mentioned in [17]. The NSL-KDD data set is an improvement over the KDDcup'99 data set. It has the following advantages over the original KDD data set:

1. The redundant records have been removed from the training set. This will prevent the classifiers from being biased towards records that are more frequent.
2. Duplicate records are not there in the test sets that have been proposed. This prevents the learner's performance from being biased by the methods that have better detection rates on the records that are more frequent.
3. From each difficulty level group, the number of records that have been selected is inversely proportional to the percentage of records in the KDDcup data set present originally. This enables it to be more efficient in having an accurate evaluation of different machine learning techniques as the classification rates of different machine learning methods vary.
4. The instances or records given in the training and test sets are reasonable in number. Evaluation of results of different research works will be consistent and comparable [17].

A subset of the NSL-KDD 1999 data set was used. This subset had 10000 instances of the original data set. Each of the instances belonged to either of the two classes i.e. normal or anomaly. To test the performance of NBTree and VFI algorithms for intrusion detection the algorithms were trained and then these performed binary classification, i.e. they determined whether a particular instance belonged to the normal (benign) class or anomaly (intrusion) class. But first, on this data set feature selection was performed. The data set, thus, obtained with fewer number of features was used for evaluation.

For experiments an open source workbench for machine learning called Weka version 3.6.11 was used. Weka is a framework of machine learning algorithms used for tasks of data mining. It contains tools for preprocessing of data, as well as tools for classification, regression, clustering, association rules and visualization.

The Chi Square and Gain Ratio implementation provided by Weka was performed one by one on the data set and the top 5, 10, 15 and 20 features were selected. Doing so, gave the most relevant and necessary features to work with. Now on these reduced data sets classification was performed by using NBTree algorithm and VFI algorithm adopting 10-fold cross validation to verify the feature selection results. Performances of each were recorded. The comparison of the accuracy of the two algorithms based on feature selection is shown in the Table1 below.

Table 1. Performance of VFI and NBT with Chi Square and Gain Ratio (Accuracy in %age)

No. of Attributes	VFI		NBTree	
	Chi Square	Gain Ratio	Chi Square	Gain Ratio
5	96.38	88.64	99.38	93.53
10	96.20	88.04	99.41	98.97
15	93.15	93.18	99.44	99.41
20	93.15	93.16	99.35	99.47

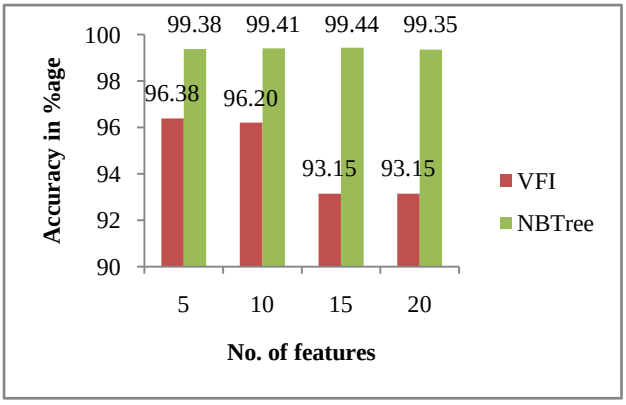

Fig 1: Performance of VFI and NBT algorithms based on the features selected by Chi Square feature selection method.

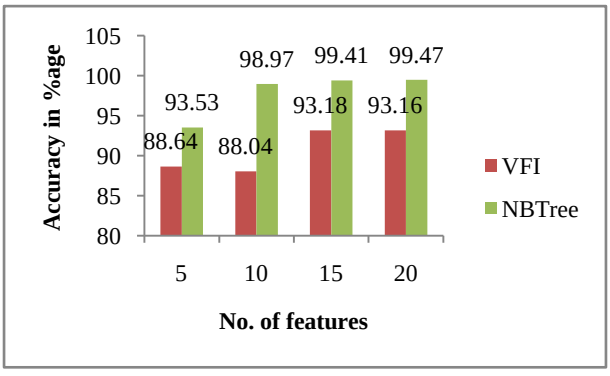

Fig 2: Performance of VFI and NBT algorithms based on the features selected by Gain Ratio feature selection method.

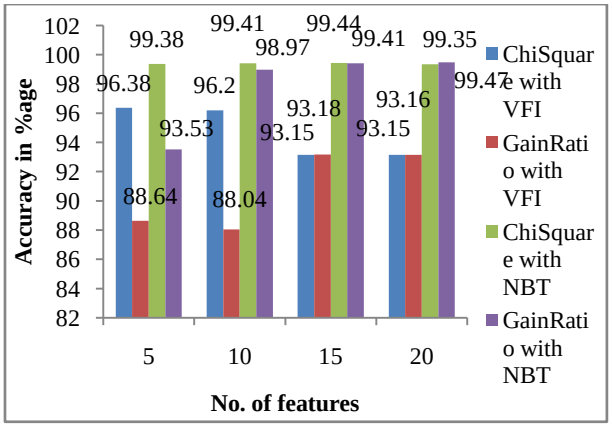

Fig 3: For fewer no. of features Chi Square performs better, and for more no. of features accuracy with Gain Ratio becomes comparable to that with Chi Square.

In the results of the experiments, it was noticed that the NBTree algorithm gives better accuracy than VFI as shown in Table 1. The Figure 1 shows the performance of the VFI and NBTree algorithms based on their accuracy, with the features selected by the Chi Square feature selection method. While, The Figure 2 shows the performance of the VFI and NBTree algorithms with the features selected by the Gain Ratio feature selection method. In Figure 3, we can see that for fewer number of attributes i.e. 5 and 10 Chi Square, with both the algorithms, gives better accuracy than Gain Ratio. Whereas, with more number of attributes i.e. 15 and 20, the Gain Ratio method gives accuracy comparable to the Chi Square method.

5. FUTURE SCOPE

In this paper, only two learning algorithms were applied and compared. The Weka machine learning workbench provides a collection of many learning schemes, which can be tested and evaluated. Also, the default parameters of the machine learning schemes were used. Further improvement may be seen if in intrusion detection by optimizing these parameters. In addition to this, a reduced subset of the actual data set was used. These experiments can be performed by using the entire data set which may lead to further improvement in the performance of the learner. The feature selection methods used in this work are Chi Square and Gain Ratio. Experiments can be done using different feature selection methods.

6. CONCLUSIONS

From the experiments performed it was found that the NBTree classifier performs better than the VFI classifier. NBTree gave better accuracy results with both the type of feature selection methods as compared to the VFI algorithm. Also, it can be said that, for fewer number of attributes Chi Square feature selection method would be better choice than Gain Ratio with both the algorithms. Whereas, with more number of attributes the Gain Ratio method gives accuracy comparable to the Chi Square method. Therefore, it can be concluded that for performing intrusion detection NBTree is a better choice than VFI. In addition to this, for scenarios where fewer features are to be used for classification Chi Square performs better than Gain Ratio but, for scenarios where more number of features are to be used Gain Ratio can also be used as an option.

7. REFERENCES

[1] Doak. 1992. An evaluation of feature selection methods and their application to computer security, Technical report, DavisCA: University of California, Department of Computer Science.

[2] Pat Langley, Wayne Iba, Kevin Thompson.1992. An analysis of bayesian classifiers. National Conference on Artificial Intelligence, 223–228.

[3] Ross R. Quinlan.1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.

[4] Ron Kohavi.1996. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid, In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 202–207.

[5] Blum and P Langley.1997. Selection of relevant features and examples in machine learning, Artificial Intelligence, 97(1-2):245–271.

[6] Gulsen Demiroz, H. Altay Guvenir. 1997. Classification by voting feature intervals, In European Conference on Machine Learning, 85–92.

[7] M Dash, H Liu. 1997. Feature Selection for Classification, Intelligent Data Analysis: An International Journal, vol. 1, no. 3,131-156.

[8] R Kohavi and G H John. 1997. Wrapper for Feature Subset Selection, Artificial Intelligence, vol. 97, 273-324.

[9] A. K Jones, R S Sielken. 2000. Computer system intrusion detection: A survey.

[10] Eric Bloedorn et al. 2001. Data Mining for Network Intrusion Detection: How to get started.

[11] Isabelle Guyon, Andr´e Elisseeff. 2003. An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, Vol. 3, 1157-1182.

[12] D. Włodzisław, W. Tomasz, B. Jacek, K. Adam. 2003. Feature Selection and Ranking Filters.

[13] Gary Stein, Bing Chen, Annie S Wu, Kein A Hua. 2005. Decision tree classifier for network intrusion detection with GA based feature selection, Proceedings of the 43rd ACM Annual Southeast Conference, Kennesaw, Georgia, Vol 2.

[14] Ian H Witten, Eibe Frank. 2005. Data Mining Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann.

[15] Marco Barreno,Blaine Nelson, Russell Sears, Anthony D. Joseph, J. D. Tygar. 2006. Can machine learning be secure? In asiaccs ’06: 46 proceedings of the 2006 acm symposium on information, computer And communications security, ACM Press, 16–25.

[16] Sandeep V. Sabnani. 2008. Computer Security: A Machine Learning Approach, Technical Report, MSc in Information Security at Royal Holloway, University of London.

[17] M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani. 2009. A Detailed Analysis of the KDD CUP 99 Data Set, Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA).

[18] Xiaofeng Liao, Liping Ding,Yongji Wang. 2011. Secure Machine Learning, A Brief Overview, Fifth International Conference on Secure Software Integration and Reliability Improvement – Companion.

[19] Shina Sheen, R. Rajesh. 2008. Network intrusion detection using feature selection and decision tree classifier. TENCON 2008 IEEE conference.

[20] Jiawei Han, Micheline Kamber. 2006. Data Mining: Concepts and Techniques, Harcourt India Pvt Ltd.

[21] Weka Machine Learning Project.

[22] Nsl-kdd data set for network-based intrusion detection systems.Available on: http://nsl.cs.unb.ca/NSL-KDD.

[23] K. Stroeh, E.R.M. Madeira, S.K. Goldenstein. 2013.An approach to the correlation of security events based on machine learning techniques. Journal of Internet services and Applications. doi 10.1186 /1869-0238-4-7