A Naïve Gain Approach to Intrusion Detection Systems

SonalPorwal
M.E Computer Student
Vidyalankar Institute of Technology
Mumbai, India

DeepaliVora
Asst. Professor and Head of the Information and Technology Department
Vidyalankar Institute of Technology
Mumbai, India

ABSTRACT

Today the world is dependent upon so many advanced technologies and network systems, that their protection from those which intent to break the system with malicious attacks, or trying some unauthorized access with an intention of financial gain or simply trying to intrude the system has become essential. This leads to the need of Intrusion Detection Systems.

Many algorithms have been suggested to implement this system, which requires building of a training model by using a training data set. In this paper,NSL KDD data set will be used to train the system using Naïve Bayes approach and then there is an attempt to improve its accuracy by proposing an algorithm based on feature selection. A concept of threshold is also introduced which works on the principle of C4.5 algorithm.

The proposed algorithm is applied on another data set that is supplied by the user which is also a part of NSL KDD.

This paper discusses the proposed algorithm which is used to improve the performance of the classification system of the Naïve Bayes Classifier and reduce the number of false alarm rate to some extent.

General Terms

Naïve Bayesian Classifier, Feature Selection, Decision Trees,

Keywords

NAVGAIN,IDS. *Thresholding*

1. INTRODUCTION

Information Systems are subjected to electronic attacks; attempts to breach the information security are risingevery day with an intension of monetary gains and obtaining secretive information by gainingunauthorized access to the system. Techniques like security passwords, firewalls, Intrusion Prevention Systems andIntrusion Detection Systems help information systems to avoid and deal with these attacks. They do this by collecting information from a variety of network systems and then analyzing the information (data) for possible attacks or intrusions. HoweverIDS cannot conduct investigations of attacks without human interventions, hencean administrator is needed monitor the traffic and review the log record.

The IDS is classified as follows:

a) Active IDS: it is configured in such a manner that it automatically blocks the suspected attacks. It does not need any intervention of the operator. It provides the advantage of real-time correction in response to attacks. The response could be in the form of logging off of a user who performed the suspicious activity or blocking the network traffic from the suspicious source.

b) Passive IDS: it is configured for monitoring and analyzing network traffic activity. An operator is alerted about the vulnerabilities and attacks occurring in the system. Unlike Active IDS it does not carry the ability of protective and corrective actions. It only detects and maintains the log of the security breach and signals an alert to the operator.

c) Network Intrusion Detection System (NIDS): it is classified on the basis of the data sources; it is placed along a network boundary and monitors all the incoming traffic (network packets) on that boundary.

d) Host Intrusion Detection System (HIDS): it is classified on the basis of the data sources, unlikeNIDS; it does not monitor the entire network but examines the activity on each individual computer level or host level and sends an alert to the operator.HIDS is usually used to monitor intrusions in mission critical servers to ensure that the performance is not degraded. It however suffers from the disadvantage of compromisation.

e) Knowledge-based IDS: also known as Signature based IDS [1] [2] stores patterns of well-known attacks (signatures) to identify the intrusions, any match with the signatures is flagged as an attack. It however suffers from inability to detect the novel attacks. Hence regular update of the signature database is required.

f) Behavior-based IDS: also known as Anomaly based IDS builds model for normal network behavior and uses it to detect new patterns that deviate from them.Any significant change from the expected behavior is flagged as an attack. However it suffers from a disadvantage called as *false alarms (*false positives*) i*.e.;any new behavior could be misunderstood as an attack while it is not. i.e.; any behavior which does not match with the regular ones in network model is labeled as an attack.

Applying Data Mining techniques suchas clustering, classification and association rules, etc., on network data improves the performance of IDS. Algorithms like decision trees, naive Bayesian classifiers, neural network, K nearest neighbour (KNN) have been widely used to analyse network logs to gain intrusion related knowledge and to improve the performance of IDS. Few techniques that have been used in past will be discussed in the further section.

2. LITERATURE SURVEY:

Data mining commonly includes the following methods [3]:

a) Classification analysis:Classification is aprocess of obtaining a model which describes the data classes, which is then used to predict the class of objects with unknown class labels. Themodel obtained is based on the analysis of a training data set (i.e., data objects whose class label is known).e.g. Decisiontrees [4] like C4.5, ID3, CARTetc., Naive Bayes Classification [5] [6]

b) Clustering analysis: opposite to classification, clustering is an unsupervised learning where no information is available on the labels of the training data.Clusters of objects are formed on the basis of distance measurements so that objects with similarities form clusters, while each individual cluster is dissimilar as compared to another. Each cluster forms a classof objects, from which rules can be derived e.g.:K-Means algorithm[6][7]

c) Association analysis: The main objective of association rules analysis is to discover association relationships between the specific values of features in large datasets e.g.: Apriorialgorithm.

The classification Techniques:

2.1 Decision Trees:

A decision tree performs classification of a given data sample through various levels of decisions to help reach a final decision. Its construction process is top down, divide and rule. The classification starts from the root node to a suitable end leaf node, which represents a classification category. When the leaf nodes are symbolic in nature then the tree is a classification tree and regression trees are the one with continuous. One disadvantage of the decision trees is when there are too many categories; the classification accuracy is significantly reduced.

2.2 Support Vector Machines:

An SVM classifier is designed for binary classification. The generalization in this approach usually depends on the geometrical characteristics of the given training data, and not on the specifications of the input space. This procedure transforms the training data into a feature space of a huge dimension. That is, to separate a set of training vectors that belongs to two different classes [8].

2.3 Fuzzy Logic:

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicatelogic [9]. The data can be classified based on statistical metrics and fuzzy logic rules can be applied to these portions of data to classify them as normal or attack.

A fuzzy system comprises of a group of linguistic statements based on expert knowledge which are used in the form of if-then rules;which helps to distinguish data using a set of fuzzy logic rules based on the attribute's linguistic value.

The Fuzzy Intrusion Recognition Engine (FIRE) [10] is an anomaly based intrusion detection system that uses fuzzy logic to assess whether malicious activity is taking place on a network. The metrics are evaluated as fuzzy sets.FIRE uses a fuzzy analysis engine to evaluate the fuzzy inputs and trigger alert levels for the security administrator.

2.4 Naïve Bayes:

NB classifier uses the concept of probability for performing supervised learning. It predicts the class of an unknown example optimally and hence is very popular in Data Mining. In NB classifier class conditional probabilities for each attribute value are calculated from the given dataset which are used to classify the known or unknown examples. Several researchers have adapted ideas from NB classifier to create models for anomaly detection [11]

The following sections will contain further discussion on this technique.

3. CLASSIFICATION APPROACH TO IDS:

The classification approach can be used in intrusion detection systems with the goal to assign the data set an appropriate class label on the basis of the model generated by using the values of the attribute features of the dataset.

They can be used for anomaly and misuse detection system as well. In the former system a normal behaviour model is built from the training data set that are known to be "normal" using various learning algorithms, in the later systemnetwork traffic data are collected and assigned the class label as "normal" or "attack" and this labelled data set is used as the training data set to learn classifiers of different types e.g.:Naïve Bayes, Decision Trees etc. which can be used to detect the known intrusions (attacks).

3.1 Naïve Bayes Classifier

As per [3], Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes (conditional independence).

There are classes, say $C_k$for the data to be classified into. Each class has a probability$P(C_k)$ that represents the prior probability of classifying an attribute intoC_k; the values of $P(C_k)$ can be estimated from the training dataset.

Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class C_i if and only if

$\mathrm{P(Ci/X)} > P(Cj/X)$ for $1 \leq j \leq m, j \neq i$.

Thus we find the class that maximizes$\mathrm{P(Ci/X)}$. The class $C_i$for which P(Ci/X)is maximized is called the maximumposteriori hypothesis.

By Bayes' theorem [3]

$$\mathrm{P(Ci/X)} = \frac{\mathrm{P(X/Ci)P(Ci)}}{\mathrm{P(X)}}$$

As P(X) is the same for all classes, only $P(X/Ci)P(Ci)$needs to be maximized. In order to reduce computation in evaluating $P(X/Ci)$.The naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means that

$$\mathrm{P(X/Ci)} = \sum_{k=1}^{n} \mathrm{P}(X_k/C_i)$$

The probabilities $\mathrm{P(X1/Ci)}, \mathrm{P(X2/Ci)} \dots . \mathrm{P(Xn/Ci)}$ can easily be estimated from the training set. $X_k$refers to the value of attribute$A_k$for sample X.

In order to predict the class label of X$P(X/Ci)P(Ci)$is evaluated for each class. The classifier predicts that the class label of X is C_i if and only if it is the class that maximizes.

$$P(X/Ci)P(Ci)$$

Although research shows that irrelevant features in data set should not theoretically affect the accuracy of Naïve Bayes, they do degrade the performance in practice.

3.2 Proposed Naïve Gain Classifier

3.2.1 Purpose:

The filtration of redundant and irrelevant attributes can improve the performance of Naïve Bayes Classifier. This can be done by using the concept of entropy of the attribute featuresi.e.;by using the concept of C4.5 decision trees [12] The tree obtained by C4.5 is used to choose the attributes that are most contributing (nearer to the root node) in the classification task.

3.2.2 Feature Selection(relevance analysis)

The features that are selected by C4.5 to construct the decision trees are most descriptive in nature with respect to the idea of classification. The algorithm selects the best attribute from training set to build the training model that yields most information for classification. As the number of training examples increase, it is observed that the attributes become less correlated. This is because C4.5 will use only one of a set of correlated features for making good splits in training set. The C4.5 algorithm ensures that the attribute nearer to the root node has highest information content. Thus this method of feature selection will help the Naïve Bayes algorithm to perform better and achieve high accuracy rates.

3.2.3 Information Gain

The information gain of each attribute is computed using the concept of entropy. Entropy is a measure of the purity in an arbitrary collection of samples. Lesser the entropy more is the information content in that attribute. Suppose the class label attribute has m distinct values defining m distinct classes,C_kLet S be a set consisting of s data samples and $s_i$be the number of samples of S in classC_k.The entropy of attribute A with *v* distinct values in sample set S having s data samples is given by

$$\mathrm{E(A)} = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{\mathrm{s}} \mathrm{I}(s_{1j}, \dots, s_{mj})$$

Where,

$$\mathrm{I}(s_{1j}, s_{2j}, \dots, s_{mj}) = -\sum_{k=1}^{m} p_{kj} \log_2(p_{kj})$$

Here $\mathrm{I}(s_{1j}, s_{2j}, \dots, s_{mj})$ is the expected information needed to classify a given sample S and p_{kj} is the probability that a sample inS_j belongs to class C_k and is calculated by $\frac{s_{kj}}{S_j}$

Where S_j contains those samples in S that have values a_j of A,s_{kj} is the number of samples of class C_k in a subset S_j

Thus Information Gain of attribute A is calculated as

$$Information\ Gain(A) = \mathrm{I}(s_1, s_2, \dots, s_m) - E(A)$$

The attributes with the highest information gain are taken into consideration for feature selection. The methodology used for feature selection will be discussed in the following sections.

4. SYSTEM DESIGN

4.1 System Architecture

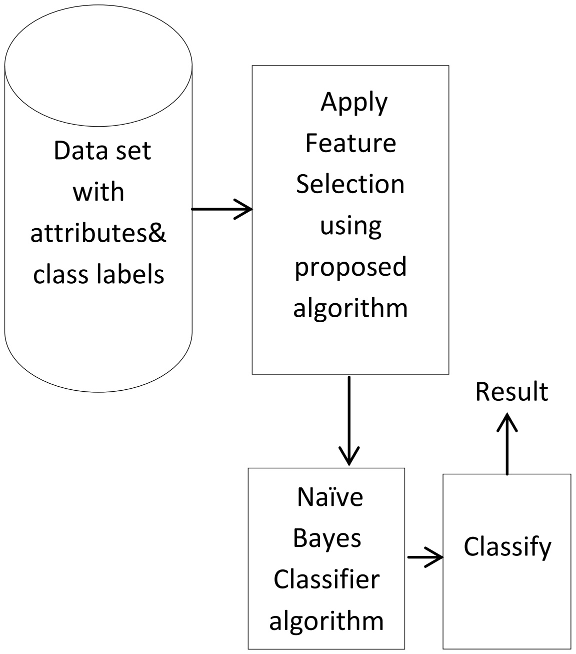
Data set with attributes& class labels
Apply Feature Selection using proposed algorithm
Result
Naïve Bayes Classifier algorithm
Classify

Fig 1: Architecture of Naïve Gain Classifier

4.2 Feature Selection Methodology

As discussed in the previous session the attributes with the higher information gain will mean the attributes that contribute more to the classification of the data as compared to the other attributes. Thus the usage of the *thresholding* technique to select the attributes serves the purpose. The information gain of each attribute is calculated from the data and then a specific threshold is decided which is of course numeric in nature. Theninformation gain of each attribute is compared with the defined threshold and whichever attribute has lesser value than the threshold value, is ignored and removed from the attribute list. The whole process defines thatthe attributes with higher information gain should be considered or one can say that the attributes which are nearer to the root node in a C4.5 decision tree should be considered as they have maximum information content, and hence contributing the most in classification.

4.3 Data set

For the purpose of experiments, NSL KDD data set having 41 fields as attributes (features) and 42nd field as the class label is used. The 42nd field can be generalized as normal or an attack. The attack being of various types such as neptune, satan, smurf etc. All attacks can be categorized into four classes of attacks as follows:

Denial of Service (DOS): Attacker tries to prevent legitimate users from using a service.

Remote to Local (r2l): Attacker does not have an account on the victim machine, hence tries to gain access.

User to Root (u2r): Attacker has local access to the victim machine and tries to gain super user privileges.

Probe: Attacker tries to gain information about the target host.

NSL KDD data set is applied into an information gain concept, all the 41 attribute's information gain is calculated and compared to the defined threshold ,when the thresholdvalue is less than the information gain of the attribute, then that attribute is selected and considered for further classification thus yielding to higher detection rates for the data set given by.

5. EXPERIMENTAL EVALUATION

5.1 Experimental data set design

For training the system, a training set is used with 12190 records of the network connections out of which 6503 records are of class normal non-malicious category,0 connections of land, 4041 connections of neptune,81 connections of warezclient,321 connections of ipsweep,87 connections of teardrop,273 connections of portsweep,30 connections of pod,12 connections of guess_passwd,145 connections of nmap,333 connections of satan,258 connections of smurf,5 connections of multihop,83 connections of back,2 connections of ftp_write,4 connections of buffer_overflow,2 connections of imap,2 connections of phf,3 connections of rootkit,5 connections of warezmaster.

A user supplied test data is used with 5001 record connections out of which 2569 records are of the class normal, and others being attack of various categories as that of training set.

5.2 Feature Selection Results

When the proposed algorithm is applied on the data set, it gives varying results depending on the threshold applied.

The algorithm is applied on the training data set initially and then for testing purpose a user supplied data set is used.

The threshold which is used on the basis of the concept of information gain is also chosen in different ranges to cover the maximum range of possibilities. As different set of thresholds are selected, the number of features selected also differs and hence their accuracy of detection. This can be seen in the graph as follows.

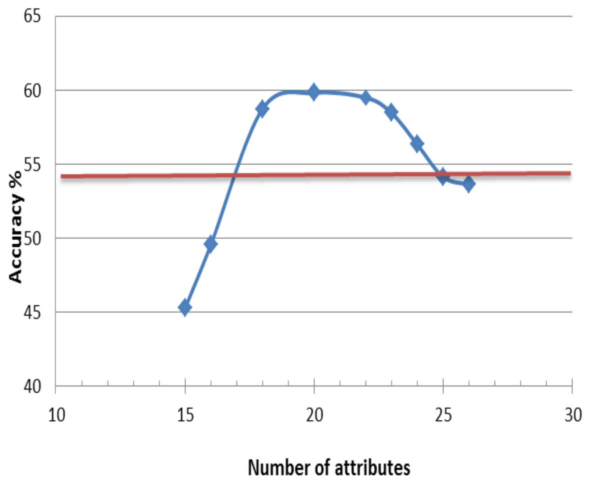

Fig 2: Result of applying feature selection and its corresponding trend in accuracy.

The above results show that, when different sets of thresholdsare applied on the data set, the accuracy ranges from approximately 49% to 59%, both being on the negative and positive side of the improvements respectively.

It has also been observed that the behavior of the algorithm has maximum variation around the threshold valueof average information gain.

The brown line in the graph above, (fig 2), refers to the accuracy obtained *without feature selection*, i.e.;the accuracy obtained with 41 attributes into consideration, and 42nd being the class classified as attack or normal.

It has been observed that out of 41 attributes if only most contributing attributes are chosen or rather filtered for feature selection i.e.;attributes with highest information gain, then the accuracy decreases as it is not able to classify well for a new data set acting as test data.And on the other hand, iftoo many attributes are selected on the basis of threshold i.e.; attributes not contributing much to the classification, the performance of the algorithm decreases as it adds noise to the data. The algorithm wrongly classifies instances due to lack of full knowledge and presence of partial information which is moreover not much contributing,hence creating confusion.

Thus the algorithm behaves best and serves optimum result at around the average threshold value.

5.3 Experimentalanalysis

The following are the experimental results obtained, when applied on the testing data set. We observe that the accuracy of detection changes when proposed algorithm is applied on the same dataset as compared to the traditional Naïve Bayes algorithm.

Table 1. Comparative analysis of Naïve Bayes algorithm and proposed algorithm: Naive Gain Approach

Accuracy obtained with Naïve Bayes Approach	Accuracy obtained after applying Naïve Gain Approach	Number Of Attributes Filtered after applying Threshold	Increase/ Decrease In Accuracy After applying (proposed)

			Naïve Gain Approach
54.929	45.2509	15	-9.6781
54.929	49.5501	16	-5.3789
54.929	58.7283	18	3.7993
54.929	59.808	20	4.879
54.929	59.4881	22	4.5591
54.929	58.4483	23	3.5193
54.929	56.3087	24	1.3797
54.929	54.1292	25	-0.7998
54.929	53.6093	26	-1.3197

From the above results, it is observed that the accuracy of Naïve Gain approach highly depends on the number of attributes selected, which in turn depends on the value of threshold. The threshold value is the information gain of attributes, which gives best result at average.

Once the system has been trained on the given dataset and a training model has been built, it can be tested for its performance on different data sets. A user supplied test set is used to detect its accuracy rate and the number of instances classified as correct and incorrect. A confusion matrix is also built to know the number of instances correctly classified as normal, incorrectly classified as an attack, number of instancescorrectly classified as variouscategories of attack, and number of instances misunderstood as others respectively.

6. CONCLUSION

Intrusion detection using Naive Bayesian classifier with the proposed algorithm is suitable for analysing large number of network logs or audit data. It improves the performance of detection rates for different types of intrusions. The main propose of this paper is to improve the performance of Naïve Bayesian classifier for intrusion detection.

In this paper, the proposed algorithm is tested on NSL KDD dataset which shows that it maximizes the accuracy of detection of intrusions near the average value of threshold. The future work focuses on applying this algorithm in real time network traffic and the improvement of classification system using another modified approach in decision trees and compare and achieve the better of the two.

7. ACKNOWLEDGMENTS

Mysincere thanks to the members who have guided me throughout the research.

8. REFERENCES

[1] Ahmed Youssef and Ahmed Emam,”Network Intrusion Detection Using Data Mining and Network Behaviour Analysis”, International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011Network Behaviour Analysis”, International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011

[2] Stefan Axelsson,”Intrusion Detection Systems:A Survey and Taxonomy”, Department of Computer Engineering,Chalmers University of Technology,March2000

[3] Jiawei Han and MichelineKamber,”Data Mining: Concepts and Techniques”, Simon Fraser Universitty,3rdedition, July 2011.

[4] Devi Prasad Bhukyaand S. Ramachandram,” Decision Tree Induction: An Approach for Data Classification Using AVL-Tree”,International Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010.

[5] Mrutyunjaya Panda1 and ManasRanjan Patra2,”Network Intrusion Detection UsingNaïve Bayes”,IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007

[6] Panda, M.; Patra, M.R., "A Comparative Study of Data Mining Algorithms for Network Intrusion Detection," Emerging Trends in Engineering and Technology, 2008. ICETET '08.IEEE, vol., no., pp.504,507, 16-18 July 2008

[7] K KBharti,”Intrusion detection using clustering”,IJCCT,Vol.1 for International Conference [ACCTA-2010], August 2010.

[8] S.VijayaPeterraj,”Study of Data Mining Techniques in Intrusion Detection”,Fast Processing Peer Reviewed International Journals,2012.

[9] George J. Klir and Bo Yuan, ”Fuzzy Sets and Fuzzy Logic:Theory and Applications”.Prentice Hall,*1995.*

[10] John E. Dickerson,” Fuzzy Network Profiling for Intrusion Detection”, NAFIPS,pp.301-306,July 2000,IEEE Explore.

[11] Kruegel, C.; Mutz, D.; Robertson, W.; Valeur, F., "Bayesian event classification for intrusion detection," Computer Security Applications Conference, 2003. Proceedings. 19th Annual, vol., no., pp.14, 23, 8-12 Dec. 2003, IEEEExplore.

[12] Juan Wang, Qiren Yang,” An intrusion detection algorithm based on decision tree technology”, Asia-Pacific Conference on Information Processing, 2009, IEEE