IJCAT International Journal of Computing and Technology, Volume 1, Issue 3, April 2014
ISSN : 2348 - 6090
www.IJCAT.org

Combining Naive Bayesian and Support Vector Machine for Intrusion Detection System

[1] Amit D. Sagale, [2] Swati G. Kale

[1] Department of Information Technology,Yeshwantrao Chavan College of Engineering Nagpur University
Nagpur,Maharashtra 441 110, INDIA.

[2] Department of Information Technology,Yeshwantrao Chavan College of Engineering Nagpur University
Nagpur,Maharashtra 441 110, INDIA.

Abstract - Computer networks are nowadays subject to an increase number of attacks. Intrusion Detection Systems (IDS) are designed to protect them by identifying anomaly behaviors or improper uses. Since the scope is different in every case (register already-known menace to later recognize them or model legal uses to trigger when a variation is detected), so far to respond against both kind of attacks IDS have failed. System developed with the single algorithms like classification, neural networks, clustering etc. gives better detection rate and less false alarm rate. Recent papers show that the cascading of two algorithm yields much better performance than the system developed with the single algorithm. Intrusion detection systems that uses solo algorithm, the accuracy and detection rate were not up to mark. Increase in the false alarm rate was also encountered. Cascading of algorithm (Hybrid model) is performed to solve this problem. This paper represents two algorithms for developing the intrusion detection system. Naïve Bayesian (NB) and Support Vector Machine (SVM) are combined to maximize the accuracy, which is the advantage of NB and diminish the wrong alarm rate which is the advantage of SVM. In this paper we proposed hybrid model which give higher detection rate and low false positive rate for IDS.

***Keywords* - IDS, NaiveBayesian, SVM, KDD dataset, Hybrid.**

1. Introduction

Nowadays, many organizations use Internet services as their communication and marketplace to do business. Together with the growth of computer networks uses, the growing rate of network attacks has been advancing, impacting to the availability, confidentiality, and integrity of critical information data. Therefore a network administrator must use one or more security tools such as firewall, antivirus, IDS and Honey Pot to prevent important data from criminal enterprises. A network system using a firewall only is not sufficient to prevent networks from all attack types. The firewall cannot protect the network against intrusion attempts during the opening port. Hence a Real-Time Intrusion Detection System is a prevention tool that gives an alarm signal to the computer user or network administrator for opposed activity on the opening session, by inspecting hazardous network activities. There are two general categories of attacks which intrusion detection technologies attempt to introduce anomaly detection and misuse detection. Anomaly detection identifies all activities that different from established patterns for users, or groups of users. Anomaly detection usually involves the creation of knowledge bases that contain the profiles of the monitored activities. The second general categories to intrusion detection are misuse detection.

These techniques involve the comparison of a user's activities with the known behaviors of attackers attempting to penetrate a system. While anomaly detection typically utilizes threshold monitoring to indicate when a certain well-known metric has been reached, misuse detection techniques frequently use rule-based approaches. When applied to misuse detection, the rules become follows for network attacks. The intrusion detection mechanism identifies a potential attack if a user's activities are found to be consistent with the established rules. The use of full rules is critical in the application of expert systems for intrusion detection. We present hybrid approaches for modelling IDS. Naive Bayesian (NB) and Support Vector machines (SVM) are combined as hierarchical hybrid intelligent system model (NB-SVM) and an ensemble approach combining the base classifiers. The hybrid intrusion detection model combined the individual base classifiers and other hybrid machine learning paradigms to maximize detection accuracy and minimize computational complexity.

2. Intrusion Detection Systems

Intrusion detection systems are used to identify, classify and possibly, to respond to benign activities. Also, Intrusion Detection System (IDS) is used to monitor all or traffic, detect malicious behavior activities, and respond to the activities. Network intrusion detection system was establish for the purpose of malicious activities detection to strengthen the security, confidentiality, and integrity of critical information systems. These systems can be network-based or host-based. HIDS(Hybrid intrusion detection System) is used to analyze the internal event such as process identifier while NIDS is to analyze the external event such as traffic volume, IP address, service port and others. The challenge of the study is how we can have an IDS with higher detection and low false positive rate? [4]

Intrusion detection has two main techniques which are misuse-based intrusion detection and anomaly based intrusion detection. Misuse-based intrusion detection IDSs that employ misuse detection approach detect attacks by comparing the predefine signatures against the network traffics captured by the IDSs. When a match is found, the IDSs will take action as the traffics are considered unsafe to computer systems or computer networks. Actions taken by the IDSs will normally include sending alerts to network administrator. IDSs that implement misuse detection approach are, however, incapable of detecting novel attacks. The network administrator will need to update the stored signatures frequently to make sure that the IDSs perform well in detecting intrusions. [5]Anomaly based intrusion detection IDSs that employ anomaly detection are capable of identifying new attacks, that contain activities deviate from the normal. Such IDSs utilize the build profiles that are learned based on normal activities in computer networks. This system has two stapes.

1) Learning: It works on profiles. The profiles represent the normal activities of the users, systems, or network connections, applications. Great care should be taken while defining profiles because currently there is no effective way to define normal profiles that can achieve high detection rate and low false positives at the same time.

2) Detection: The profile is used to detect any deviance in user normal behavior. [7]

Different Types of Attacks

- Denial of Service (DOS): Making some machine resources too busy to answer to legitimate users requests.
- User to Root (U2R): Exploiting vulnerability on a system to obtain a root access.
- Remote To Local (R2L): Using vulnerability in order to obtain a local access like a machine user.
- Probing: Collecting useful information or known vulnerabilities about a network or a system. [8]

3. Machine Learning Algorithm

Machine learning studies how to automatically discover to make accurate predictions based on past observations. This type of algorithm we provided Knowledge with result for learning purpose and then we providing knowledge algorithm give result on past observation. At the time of learning algorithm makes some rules or threshold value for each classifier. There are many algorithm are used as machine like Decision tree, Naïve Bayesian, Support Vector Machine etc.

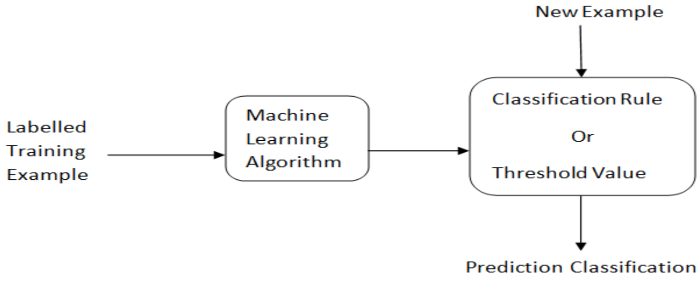
New Example
Labelled Training Example
Machine Learning Algorithm
Classification Rule
Or
Threshold Value
Prediction Classification

Figure 1: Flow of Machine Algorithm

3.1. Naive Bayesian

Naive Bayesian classifier is a simple classification scheme, which estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label c. The conditional independence assumption can be formally stated as follows:

$$P(A \mid C = c) = \prod_{i=1}^{n} P(A_i \mid C = c) \qquad (1)$$

Where each attribute set A = {A1,A2,….,An}consists of n attribute values. With the conditional independence assumption, instead of computing the class conditional probability for every grouping of A, only estimate the conditional probability of each Ai, given C. The latter approach is more practical because it does not require a very large training set to obtain a good estimate of the probability. To classify a test example, the naïve Bayesian classifier computes the posterior probability for each class C.

$$P(C \mid A) = \frac{P(C)\prod_{i=1}^{n} P(A_i \mid C)}{P(A)} \qquad (2)$$

Since P(A) is fixed for every A, it is sufficient to choose the class that maximizes the numerator term,

$$P(C)\prod_{i=1}^{n} P(A_i \mid C) \qquad (3)$$

The naïve Bayesian classifier has several advantages. It is easy to use, and unlike other classification approaches, only one time scan of the training data is required. The naïve Bayesian classifier can easily handle missing attribute values by simply omitting the probability when calculating the likelihoods of membership in each class.

3.2. Support Vector Machines

Support Vector Machines have been proposed as a novel technique for intrusion detection. A Support Vector Machine (SVM) maps input (real-valued) feature vectors into a higher dimensional feature space through some nonlinear mapping. SVMs are powerful tools for providing solutions to classification, regression and density estimation types of problems. These are developing on the principle of structural risk minimization. Structural risk minimization seeks to find a hypothesis for which one can find lowest probability of error. The structural risk minimization can be achieved by finding the hyper plane with maximum separable margin for the data.

Computing the hyper plane to separate the data points i.e. training a SVM leads to quadratic optimization problems. SVM uses a feature called kernel to solve this problems. Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; some of them are Polynomial, radial basis function, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training classifier, which selects support vectors along the surface of this function. SVMs classify data by via these support vectors, which are member of the set of training inputs that outline a hyper plane in feature space. The implementation of SVM intrusion detection system has two phases: training and testing. The main advantage of this method is speed of the SVMs, as the capability of detecting intrusions in real-time is very important. SVMs can learn a larger set of patterns and be able to better scale, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification. The main disadvantage is SVM can only handle binary-class classification whereas intrusion detection requires multi-class classification.

4. Proposed Algorithm

Given a training data *D* contains the following attributes {*A1, A2,…,An*} and each attribute *Ai* contains the following attribute values {*Ai1, Ai2,…,Aih*}. The attribute values can be discrete or continuous. Also the training data *D* contains a set of classes *C* = {*C1, C2,…,Cm*}. Following are steps for implementation hybrid model.

1) In proposed algorithm using naïve Bayesian the prior *P*(*Cj*) and conditional *P*(*Aij*|*Cj*) probabilities in the training data *D*. The prior probability *P*(*Cj*) for each class is estimated by counting how often each class occurs in the training data *D*.
2) For each attribute *Ai* the number of occurrences of each attribute value *Aij* can be counted to determine *P*(*Ai*). Similarly, the conditional probability *P*(*Aij*|*Cj*) for each attribute values *Aij* can be estimated by counting how often each attribute value occurs in the class in the training data *D*.
3) Then the algorithm classifies all the examples in e_i with some targeted class the training data *D* with these prior *P*(*Cj*) and conditional *P*(*Aij*|*Cj*) probabilities. For classifying the examples, the prior and conditional probabilities are used to make the prediction.
4) To classify the example, the algorithm estimates the likelihood that *ei* is in each class. The probability that *ei* is in a class is the product of the conditional probabilities for each attribute value with prior probability for that class. The posterior probability *P*(*Cj* | *ei*) is then found for each class and the example classifies with the highest posterior probability for that example. After classifying all the training examples, the class value for each example in training data *D* updates with Maximum Likelihood (ML) of posterior probability *P*(*Cj*|*ei*).

$$C_j = C_i \longrightarrow P_{ML}(C_j | e_i).$$

5) Then again the algorithm calculates the prior $P(Cj)$ and conditional $P(Aij|Cj)$ probabilities using updated class values in the training data D, and again classifies all the examples of training data using these probabilities. If any of the training example is misclassified then the algorithm calculates the information gain for each attributes $\{A1, A2,\ldots,An\}$ in the training data D.
6) And after the full training providing using naïve Bayesian Classify the data misclassified data and classified data. classified data will provided to SVM also provided targeted class and unlabeled data using multiclass SVM classified in two DOS, R2l, U2R and PROB.

5. Experiment Analysis

In order to evaluate the performance of proposed algorithm for network intrusion detection, we performed 5-class classification using KDD99 intrusion detection benchmark dataset. we compare with Naïve Bayesian(NB) single algorithm and Support Vector Machine(SVM) single algorithm with Hybrid model(HM) in term of Detection Rate(DR) and False Positive rate (FP).

Table 1: Experimental Result

Method	Normal	Probe	DOS	U2R	R2L
HM DR	99.72	99.25	99.75	99.20	99.26
HM FP	0.06	0.39	0.04	0.11	6.81
NB DR	99.27	99.11	99.69	64.00	99.11
NB FP	0.05	0.32	0.04	0.12	6.87
SVM DR	99.71	98.22	99.63	86.11	97.79
SVM FP	0.06	0.51	0.04	0.12	7.34

6. Conclusions and Future Works

This paper introduced a new hybrid algorithm for network intrusion detection using naive Bayesian classifier and SVM algorithm, which analyzes the large volume of network data and considers the complex properties of attack behaviors to improve the performance of detection speed and detection accuracy. In this paper we have concentrated on the development of the performance of naïve Bayesian classifier. It has been successfully tested that this hybrid algorithm minimized false positives, as well as maximize balance detection rates on the 5 classes of KDD99 benchmark dataset. The attacks of KDD99 dataset detected with 99% accuracy using Hybrid algorithm. The future work focus on improving the false positives of remote to user (R2L) attack and apply this detection model into real world IDS.

References

[1] Network Intrusion Detection Using Improved Decision Tree Algorithm K.V.R. Swamy, K.S. Vijaya Lakshmi Department Of Computer Science and Engineering V.R.Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India IJCSIT-2012.

[2] M. Moorthy , Dr. S. Sathiyabama A Study of Intrusion Detection using Data Mining IEEE-2012.

[3] Network Intrusion Detection Using Tree Augmented Naive-Bayes R. Najafi Mohsen Afsharchi IEEE-2012.

[4] Attacks Classification in Adaptive Intrusion Detection using Decision Tree Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman WASET-2010.

[5] Modeling Intrusion Detection Systems Using Hybrid Intelligent Systems Sandya Peddabachigari, IEEE-2005.

[6] Intrusion Detection System using Support Vector Machine and Decision Tree IJCA(2010) Snehal A. Mulay P.R. Devale, G.V. Garje.

[7] Combining Naivie Bayes And Decision Tree For Adaptive Intrusion Detection IJNSA-2010. Dewan Md. Farid, Nouria Harbi, and Mohammad Zahidur Rahman.

[8] Effective Network Intrusion Detection using Classifiers Decision Trees and Decision rules G.MeeraGandhi , Kumaravel Appavoo , S.K. Srivatsa IJANA-2010

[9] Intrusion Detection System using Support Vector Machine Jayshree Jha , Leena Ragha IJAIS-2013.

[10] Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System Jashan Koshal, Monark Bag IJCNIS-2012

[11] Adaptive Intrusion Detection based on Boosting and Naïve Bayesian Classifier Dewan Md. Farid , Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman IJCA-2011

[12] A Detailed Analysis of the KDD CUP 99 Data Set Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani CISDA-2009

[13] Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets. H. Güneş Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood

[14] Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features Adetunmbi A.Olusola., Adeola S.Oladele. and Daramola O.Abosede PWCECS-2010

[15] Evaluation Effectiveness of Hybrid IDS Using Snort with Naïve Bayes to Detect Attacks Safwan Mawlood Hussein , Fakariah Hani Mohd Ali, Zolidah Kasiran.

[16] Feature selection and design of Intrusion Detection System based on k-means and triangle area support vector machine , Pingj ie Tang , Rang-an Jiang , Mingwei Zhao. IEEE-2010.

Authors

Amit D. Sagale received the B.E. degree in Information Technology from PCEA Higna, Nagpur, INDIA in 2009 and perusing MTech. degree in Information Technology from YCCE, Nagpur in 2013. He worked as a Teaching Assistant at 2-WEEK ISTE DBMS-WORKSHOP conducted by IIT Bombay (21st may to 31st may 2013) in YCCE Nagpur.

Swati G. Kale She now currently working as a Asst. Prof. in Information Technology Department at YCCE Nagpur.