IRJEAS
NURTURING RESEARCH

INTERNATIONAL RESEARCH JOURNAL
OF
ENGINEERING & APPLIED SCIENCES

ISSN: 2322-0821(O) VOLUME 5 ISSUE 4
ISSN: 2394-9910(P) Oct 2017 - Dec 2017

www.irjeas.org

NEW NAÏVE BAYES CLASSIFIER FOR IMPROVE INTRUSION DETECTION SYSTEM ACCURACY

Jitendra Malviya[1], Vaibhav Patel[2], Anurag Srivastava[3]

[1] *M.Tech Scholar, Department of Computer Science& Engineering ,NRI Institute of research &Technology,Bhopal*
[2] *Asst. Professor, Department of Computer Science& Engineering ,NRI Institute of research &Technology,Bhopal*
[3] *Professor &Head, Department of Computer Science& Engineering ,NRI Institute of research &Technology,Bhopal*

ABSTRACT- Information security is becoming a more important issue for modern computer system, progressively. Intrusion Detection System (IDS) as the main security defensive technique is widely used against many category of attacks. IDS are used to detect various kinds of attacks in very large data. Data mining and machine learning technology have applied in intrusion detection systems. Many machine learning methods have also been introduced by researcher recently to obtain high correctness and uncovering rate. Unfortunately a potential drawback of all those methods is that how to distinguish abnormal association activities effectively. Other shortcomings are low detection rate, long training times in large data set. This paper proposes new naïve bayes classification algorithms based on log probability. With the help of this naïve bayes classifier, IDS improve the performance. This paper talks about classification of abnormal association, high accuracy and detection rate with low false alarm.

Keywords- Naïve Bayes, New naïve bayes, machine learning technique, IDS, Classification.

1. INTRODUCTION

We securing information either in private or government sector has become an essential requirement. System vulnerabilities and valuable information magnetize most attackers' attention. Traditional intrusion detection approaches such as firewalls or encryption are not sufficient to prevent system from all attack types. The number of attacks through network and other medium has increased dramatically in recent years. Efficient intrusion detection is needed as a security layer against these malicious or suspicious and abnormal activities. Thus, intrusion detection system (IDS) has been introduced as a security technique to detect various attacks. IDS can be identified by two techniques, namely misuse detection and anomaly detection. Misuse detection techniques can detect known attacks by examining attack patterns, much like virus detection by an antivirus application. However they cannot detect unknown attacks and need to update their attack pattern signature whenever there is new attacks .On the other hand, anomaly detection identifies any unusual activity pattern which deviates from the normal usage as intrusion. Although anomaly detection has the capability to detect unknown attacks which cannot be addressed by misuse detection, it suffers from high false alarm rate .In recent years, and interest was given into machine learning techniques to overcome the constraint of traditional intrusion techniques by increasing accuracy and detection rates. New techniques based genetic algorithm used in our detection approach. The advantage of IDS (Intrusion Detection system) can greatly reduce the time for system administrators/users to analyze large data and protect the system from illicit attacks. In this research, we will only be looking at Machine Learning technique based on genetic algorithms.

A. Machine Learning

When a computer needs to perform a certain task, a programmer's solution is to write a computer program that performs the task. A computer program is a piece of code that instructs the computer which actions to take in order to perform the task. The field of machine learning is concerned with the higher-level question of how to construct computer programs that automatically learn with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E Thus, machine learning algorithms automatically extract knowledge from machine readable information. In machine learning, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts applied to the research at hand, this means that a computer would learn to classify alerts into incidents and non-incidents (task T). A possible performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly. The training experiences (E) could be labeled instances.

B. Naïve Bayes Algorithms

Naive Bayes classifier is the simplest among Bayesian Network classifiers. It has shown to be very efficient on a variety of data classification problems. However, the strong assumption that all features are conditionally independent given the class is often violated on many real world applications. Therefore, improvement of the Naive Bayes classifier by alleviating the feature independence assumption has attracted much attention.

C. Dataset

KDD Cup 1999 Data (KDD99) is the dataset used in the evaluate machine learning technique. In practice, we recognize that this dataset is decade old and has many criticisms for Current research. But we believe that it is still sufficient for our experiment which aims to reflect the performance of distinct machine learning approaches in a general way and find out relevant issues. In addition, the full KDD99 dataset Contain 4,898,431 records and each record contain 41 features. Due to the computing power, we do not use the full dataset of KDD99 in the experiment but a 10% portion use of it. This 10% KDD99 dataset contains 494,069 records (each with 41 features) and 4 categories of attacks. The details of attack categories and specific types are shown in Table1.

According to Table1, there are four attack categories in 10% KDD99 dataset:

(1) Probing: Scan networks to gather deeper information
(2) DoS: Denial of service
(3) U2R: Illegal access to gain super user privileges
(4) R2L: Illegal access from a remote machine

Every attack categories contain some specific attack types. For example, DoS has 6 specific attack types (e.g. back, land, neptune), R2L has 8 specific attack types (e.g ftp write, guess passwd, imap). There are totally 22 specific attack types within the 10% KDD99 dataset, while the full KDD99 dataset has 39 specific attack types. Although the number of specific attack types is different between 10% KDD99 dataset and full KDD99 dataset, we believe that there are no negative effects on our evaluation purpose.

D. Feature selection

The attribute (also called feature) in the dataset is a key element that can affect the performance results of machine learning schemes. There are 41 features of each record in the dataset which belong to 4 main categories: TCP connection content basic characteristic, time-based network traffic and host-based network traffic. The full features are shown in Table2. Intuitively, some features are insignificant to the training of machine learning algorithms as well the improvement of detection rate.

E. Log Probability

A log probability is simply the logarithm of a probability. The use of log probabilities means representing probabilities in logarithmic space, instead of the standard [0, 1] interval. In most machine learning tasks we actually formulate some probability p which should be maximized, here we would optimize the log probability log(p) instead of the probability for class θ. The use of log probabilities determines better numerical stability, when the probabilities are close to each other and very small.

ex = y

loge (y) = x

Where x is probability. To get back the values of probability take log of y on base e.

2. RELATED WORK

There are many nonparametric approaches for intrusion Detection in large traffic, such as neural networks, SVM and decision trees in a uniform environment with the purpose of exploring the practice and issues of using these approaches in detecting abnormal Behaviors. With the analysis of experimental results, we claim that the real performance of machine learning algorithms depends heavily on practical context. Therefore, the machine learning approaches are supposed to be applied in an appropriate way in terms of the actual settings.[1] the use of ELM methods to classify binary and multi-class network traffic for intrusion detection. The performance of ELM in both binary-class and multi-class scenarios are investigated, and compared to SVM based classifiers. The advantages of ELM include its scalability and significant reductions in training time, as compared to SVM. Simulation results show that the proposed method can detect intrusions even in large datasets with short training and testing times. [2] This paper presents a neural-network-based active learning procedure for computer network intrusion detection. Applying data mining and machine learning techniques to network intrusion detection often faces the problem of very large training dataset size. The practical problems associated with such a large dataset include very long model training times, redundant information, and increased complexity in understanding the domain-specific data. We demonstrate that a simple active learning procedure can dramatically reduce the size of the training data, without significantly sacrificing the classification accuracy of the intrusion detection model. The network traffic instances are classified into one of two categories – normal and attack. A comparison of the actively trained neural network model with a C4.5 decision tree indicated that the actively learned model had better generalization accuracy. [3]. In this paper the performance of a Machine Learning algorithm called Decision Tree is evaluated and compared with two other Machine Learning algorithms namely Neural Network and Support Vector Machines. The algorithms were tested based on accuracy, detection rate, false alarm rate and accuracy of four categories of attacks. From the experiments conducted, it was found that the Decision tree algorithm outperformed the other two algorithms. In this research, we intend to compare the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset. The dataset is very large, and it is almost impossible to test the data using ordinary computer. The large dataset requires high performance machines. Thus in this research, we need to resemble the data set into smaller dataset, and run it using an ordinary computer. [4] Authors develop a new version of the Naive Bayes classifier without assuming independence of features. The proposed algorithm approximates the interactions between features by using conditional probabilities. We present results of numerical experiments on several real world data sets, where continuous features are discretized by applying two different methods. These results demonstrate that the proposed algorithm significantly improve the performance of the Naive Bayes classifier, yet at the same time maintains its robustness.[11] Author [3] Presented a comprehensive analysis on Probe attacks, by applying various popular machine learning techniques such as Naïve Bayes, SVM, Decision Trees etc. Author used KDDcup99 data set to build the model. Author proposed three layers architecture for detection of probe attacks. Principal Component Analysis is used for dimensionality reduction. Author removed duplicate samples from the training data set. Here author compared the performance of each classifier with the help of a line chart.[16]

3. Proposed Work

Some research in machine learning community has addressed the strategy for improve the performance of intrusion detection system. To classify network activities as normal or abnormal while minimizing misclassification propose a classification framework based on new naïve bayes algorithm. The Proposed naïve bayes algorithm is using the concept of log probability. Detail about the log probability discuss in Introduction.

Proposed new naïve bayes Algorithm:

```
General Naive Bayes Algorithm
Start To get Class of particular Instance
Declare Array probs of size = n
Loop For j=0 to n-1  //where n is no. of classes in dataset
For each class get value of probability and save in probs[j]
End For
Get no. of attributes
Loop, While (there exists no. of attributes)
Declare variable temp and max=0;
Loop For j=0 to n-1
Get probability estimates of each attribute and product over of these with each class probabilities.
Get max of these probability obtained in previous step and store in array of probabilities.
Now get / Take log of probabilities and update in array of probabilities.
Take max value from array of log of probabilities
End For
End while
```

This is the proposed new naïve bayes algorithm which is used for improving the IDS performance. Better accuracy of naïve bayes classifier. We used KDD Cup99 dataset. The first step is pre-processing and select appropriate attribute from dataset. In the next step, we applied classification on training dataset in order to classify normal and abnormal data. Now apply new naïve bayes classier and measure performance. This same process also applied for general naïve bayes classifier algorithms and compare result. Architecture of the proposed work are shown in figure 2. For experiment purpose weka 3.8 tool is used.

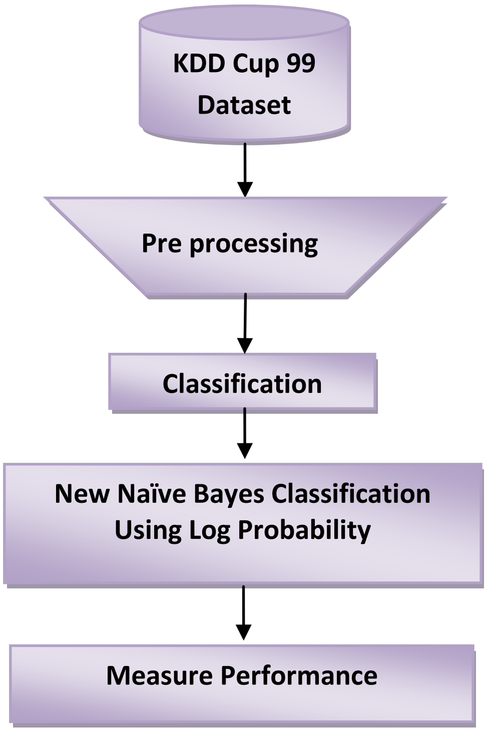

Figure 2. Architecture of the system

4. CONCLUSION

In this paper, new naïve bayes algorithms based on log probability have been proposed. This algorithms show the better performance in terms of accuracy and covering and uncovering rate. The purpose of this proposed method efficiently classify abnormal and normal data by using very large data set and detect intrusions even in large datasets with short training and testing times. With proposed method we get high accuracy for many categories of attacks and detection rate with low false alarm. The proposed method results compare with general naïve bayes algorithms using KDD Cup 99 dataset to improve the performance of intrusion detection system.

5. REFERENCES

[1] YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE

[2] Chi Cheng, Wee Peng Tay and Guang-Bin Huang "Extreme Learning Machines for Intrusion Detection" - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[3] Naeem Seliya , Taghi M. Khoshgoftaar "Active Learning with Neural Networks for Intrusion Detection" IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10/$26.00 ©2010 IEEE

[4] Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O International Conference on Networking and Information Technology 978-1-4244-7578-0/$26.00 © 2010 IEEE

[5] Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, Member, IEEE "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming" IEEE, JANUARY 2011

[6] Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" - 2010 Second International Conference on MultiMedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[7] Jingbo Yuan , Haixiao Li, Shunli Ding , Limin Cao "Intrusion Detection Model based on Improved Support Vector Machine" Third International Symposium on Intelligent Information Technology and Security Informatics 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[8] Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze "A Novel Intrusion Detection Method Based on Support Vector Machines" IEEE 2010.

[9] W. Yassin, Z. Muda, M.N. Sulaiman, N.I.Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification" IEEE 2011.

[10] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques" IEEE 2010.

[11] SONA TAHERI1 MUSA MAMMADOV1,2 ADIL M. BAGIROV1, Improving Naive Bayes Classifier Using Conditional Probabilities, Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia.

[12] Rajesh Wankhede, Vikrant Chole " Intrusion Detection System using Classification Technique" International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[13] S. Revathi, Dr. A. Malathi " A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December – 2013

[14] Koushal Kumar, Jaspreet Singh Batth Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016

[15] Purushottam R. Patil, Yogesh Sharma, Manali Kshirasagar," Performance Analysis of Intrusion Detection Systems Implemented using Hybrid Machine Learning Techniques" International Journal of Computer Applications (0975 – 8887) Volume 133 – No.8, January 2016

Ch.Ambedkar, V. Kishore Babu Detection of Probe Attacks Using Machine Learning Techniques International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 2, Issue 3, March 2015, PP 25-29 ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online) www.arcjournals.org