INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O) ISSN: 2394-9910(P)
VOLUME 5 ISSUE 2
Apr 2017 - Jun 2017
www.irjeas.org

"AN IMPLEMENTATION OF INTRUSION DETECTION SYSTEM USING MACHINE LEARNING CLASSIFICATION TECHNIQUE"

Anurag Shrivastava[1], Jyoti Sondhi[2], Sadaf Khan[3]

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

***Abstract*—Use of internet is increasing progressively, so that large amount of data and it security is also an issue. IDS (Intrusion detection system) are main defending technique. Sampling technique is one the solution of large dataset. This work proposes a sampling technique for obtaining the sampled data. Sampled dataset represent the whole dataset with proper class balancing. The purpose of this paper is to propose classification framework based on different model. This model also based on machine learning and sampling to improve the IDS classification performance. The Proposed work is tested on basis of Accuracy, Error rate, Detection rate and False Alarm rate. In this paper compare results of all models. NSL-KDD dataset used for this approach.**

***Keywords*— IDS, Sampling, Classification, Machine learning technique, NSL KDD Dataset**

I. 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 machine learning based IDS with sampling is 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. Improve the performance of IDS and the low false alarm rate.

A. Data Mining

Data Mining is defined as the technique of extracting information or knowledge from huge amount of data. In other words, we can say that data mining is mining knowledge from large data.

B. Machine Learning Technique :

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

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 5 Issue 2, Apr 2017- Jun 2017, Page 14-17

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.

II. Related work:

The authors [1] have proposed to use data mining technique including classification tree and support vector machines for intrusion detection. Utilize data mining for solving the problem of intrusion because of following reasons: It can process large amount of data. User's subjective evolution is not necessary, and it is more suitable to discover the ignored and unknown information. Machine learning based ID3 and C4.5 two common classification tree algorithms used in data mining. Author said C4.5 algorithm is better than SVM in detecting network intrusions and false alarm rate in KDD CUP 99 dataset.

In [2], the author said 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 which has been conducted by A. 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. Compare the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset.

Author [3] says common problem of IDS that are high false positives and low detection rate. An unsupervised machine learning using k-means was used to propose a model for Intrusion Detection System (IDS) with higher efficiency rate and low false positives and false negatives. The NSL-KD data set was used which consisted of 25,192 entries with 22 different types of data.

In [5] authors said Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. It is the process of selecting representatives which indicates the complete data set by examining a fraction. This paper focuses on different types of sampling strategies applied on neural network. Here sampling technique has been applied on two real, integers and categorical dataset such as yeast and hepatitis data set prior to classification. Authors give the comparison of different sampling strategies for classification which gives more accuracy.

In [6] authors investigate the effect of sampling methods on the performance of quantitative bankruptcy prediction models on real highly imbalanced dataset. Seven sampling methods and five quantitative models are tested on two real highly imbalanced datasets. A comparison of model performance tested on random paired sample set and real imbalanced sample set is also conducted. The commonly used re-sampling strategies include oversampling and under-sampling. Two widely used oversampling methods: Random Oversampling with Replication (ROWR) and Synthetic Minority Over-sampling Technique (SMOTE) are employed in this paper. Two Under-sampling sampling method: Random under sampling (RU) and Under-sampling Based on Clustering from Gaussian Mixture Distribution (UBOCFGMD). Under-sampling method is better than oversampling method because there is no significant difference on performance but oversampling method consumes more computational time.

The work [7] discusses imbalanced dataset. A dataset is imbalanced if the classification categories are not approximately equally represented. Authors discuss some of the sampling techniques used for balancing the datasets, and the performance measures more appropriate for mining imbalanced datasets. Over and under-sampling methodologies have received significant attention to counter the effect of imbalanced data sets. Sampling methods are very popular in balancing the class distribution before learning a classifier.

III. NSL KDD data set and sampling:

A. NSL KDD Dataset :

In Earlier days the researcher focused on DARPA dataset for analyzing intrusion detection [13]. It consist of seven weeks of training and also two weeks of testing raw tcpdump data. The main drawback is its packet loss. The refined version of DARPA dataset which contains only network data (i.e. Tcpdump data) is termed as KDD dataset [11]. Which consist on 5 million single connection for training records and 2 million connection for testing. Due to its huge size the researcher used on 10%% of dataset to analysis intrusion accuracy which affects the performance of the system, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [9] is proposed, which consists of selected records of the complete KDD data set. The advantage of NSL KDD dataset is 1. No redundant records in the train set, so the classifier will not produce any biased result 2. No duplicate record in the test set which have better reduction rates. 3. The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set. The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test dataset i.e. not available in the training datasets.

B. Sampling:

Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. The researcher's initial task is to formulate a rational justification for the use of sampling in his research. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. It is the process of selecting representatives

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 5 Issue 2, Apr 2017- Jun 2017, Page 14-17

which indicates the complete data set by examining a fraction. Due to sampling we overcome the problems like; i) in research it is not possible to collect and test each and every element from the data base individually; and ii) study of sample rather than the entire dataset is also sometimes likely to produce more reliable results.

C. Feature selection

Due to the large amount of data flowing over the network real time intrusion detection is almost impossible. Feature selection can reduce the computation time and model complexity. Research on feature selection started in early 60s [9]. Basically feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features [10] from the data for building an effective and efficient learning model [11].
A number of feature selection algorithms are proposed by various authors.[] Attribute evaluator is basically used for ranking all the features according to some metric.

IV. Proposed Work

Some research in machine learning community has addressed the strategy of re-sampling the original dataset to deal with the issue of class imbalance []. The commonly used re-sampling strategies include oversampling and under-sampling. Oversampling is to sample the minority class over and over to achieve the balanced distribution of the two classes, while under-sampling is to select a portion of the majority class to achieve the distribution balance of the two classes. In the original imbalanced training dataset, let the original sample set of minority class and majority class denoted by Cmin and Cmax separately, the size of minority class Cmin is much less than the size of majority class Cmax.

Under sampling:

Under sampling is to select a portion of the majority class to achieve the distribution balance of the two classes. In Random under sampling the majority class is under-sampled by randomly removing samples from the majority class Population until the majority class becomes minority class.

V. ARCHITECTURE OF THE PROPOSED MODEL

In Architecture of the proposed model shows that in NSL KDD dataset Firstly we are applying sampling technique and get sampled dataset now we are using preprocessing technique in sampled dataset and applying feature selection method.

Now going to classification part and determine the training and testing data in very short period after that applying classification technique in trained data and evaluate the result. Same procedure is applying in different machine learning classifier and measure result. Also measure the classifier performance with un-sampled dataset.

Parameter of the performance measures in the terms of high detection rate, low false alarm rate, less training and testing time, and high accuracy.

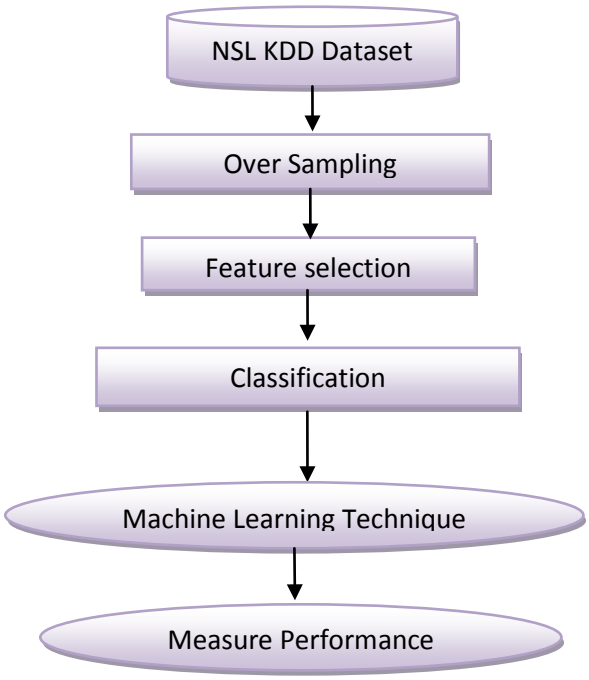

Figure 1. Architecture of the system

VI. CONCLUSION

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. Most importantly when using this method redundant information, complexity with abnormal behaviors are reduced. 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 other machine learning technique using intrusion detection to improve the performance of intrusion detection system. Experimental results and analysis shows that the proposed system gives better performance in terms of high detection rate, low false alarm rate, less training and testing time, and high accuracy.

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] Solane Duquea*, Dr.Mohd. Nizam bin Omarb Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS) www.sciencedirect.com, Elsevier 2015.

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

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

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

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

[12] S. SobinSoniya, S. Maria Celestin Vigila, Intrusion Detection System: Classification and Techniques 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[13] Anand Motwani, Vaibhav Patel, Anita Yadav Optimal Sampling for Class Balancing with Machine Learning Techniquefor Intrusion Detection System ISSN No. (Online): 2277-2626 (2): 47-51(2015)

[14] Mr. Sachin S. Patil, Prof. Deepak Kapgate, Prof. P.S. Prasad
A Review on Detection of Web Based Attacks using Data Mining Techniques December 2013 ISSN: 2277 128X

[15] [1] Solane Duquea*, Dr.Mohd. Nizam bin Omarb Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS) www.sciencedirect.com, Elsevier 2015.

[16] [2] 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

[17] [3] 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

[18] [4] 1Premansu sekhara rath, 2Manisha mohanty, 3Silva acharya, 4Monica aich Optimization of ids algorithms using data mining technique International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-4, Issue-3, Mar.-2016

[19] [5] S. Revathi 1 Dr. A. Malathi2 Network Intrusion Detection Based On Fuzzy Logic International Journal of Computer Application Issue 4, Volume 1 (February 2014) Available online on http://www.rspublication.com/ijca/ijca_index.htm ISSN: 2250-1797

[20] [6] M.Saraswathi1, N.Kowsalya2 Anomaly Detection via Online Oversampling Principal Component Analysis International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 8, August 2015.

[21] [7] Mr. Kamlesh Patel1, Mr. Prabhakar Sharma2 An Implementation of Intrusion Detection System Based on Genetic Algorithm International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 11, November 2016