Analysis of Intrusion Detection Technique In Data Mining: A Survey

Deepa Hindoliya[1] ,Prof. Avinash Sharma[2]

[1]PG Scholar, CSE ,[2] Associate Professor, CSE

MITS, Bhopal, India

Abstract:- Cyber terrorism has recently been said to be the biggest threat to our modern society. Every day a new cyber-scare story makes the headlines. The national government recognizes the importance of cyber security, as several officials have made clear in the past few years "Cyber security is among the most serious economic and national security challenges we will face in the 21st Century, we face a long-term challenge in cyberspace from foreign intelligence agencies and militaries, criminals, and others, and, this struggle will wreak serious damage on the economic health and national security. For the prevention and detection of cyber terrorism used intrusion detection system, intrusion detection system detects illegal behavior of network over data. In current research trend performance of intrusion detection system is important issue. Now various authors' used machine learning and feature optimization technique for intrusion detection system. Machine learning technique is collection of all learning algorithm such as classification, clustering and regression. For the improvement of machine learning technique used feature optimization technique. In this paper presents review of intrusion detection technique using machine learning and feature optimization process.

Key Terms:- IDS, Machine Learning, Feature Optimization, KDDCUP, Traffics features, PCA.

1. INTRODUCTION

The advancement of information communication network contributes the improvement of the quality of daily life of people, and now considered as fundamental social and economic infrastructure. However, the increase of incidents and threats against this infrastructure has turned out to be a serious problem. These days it is very significant to maintain a high level security to ensure protected and trust information message among different groups. But integrity of data over internet and any other network is always under threat of intrusions and misuses. So Intrusion Detection Systems (IDS) have become crucial components in computer and network security [6]. Improvement of intrusion detection technique in major concern of financial sector and social networking site use of common user. The computer security community has developed a variety of intrusion detection systems to prevent attacks on computer systems. Feature optimization and feature reduction is major challenges in current researcher trend in intrusion detection technique. Irrelevant and redundant attributes of intrusion detection dataset may lead to complex intrusion detection model as well as reduce detection accuracy. The network based intrusion detection is called as mysterious attacks and this attack is analyzed on the basis of normal attack scenario [20]. Despite all the applied mechanism for intrusion detection system not provide the complete secured data. Therefore, intrusion detection is becoming an increasingly important technique that monitors network traffic and identifies network intrusions attacks to computer systems. A number of machine learning based approaches have been used for detecting abnormal threats. Machine learning refers to a group of techniques that develop the easiness for ambiguity, improbability, incomplete fact, and estimate to achieve toughness and low solution price. The principle constituents of machine learning are Fuzzy Logic (FL) [18, 15], Artificial Neural Networks (ANNs), Probabilistic Reasoning (PR), and Genetic Algorithms (GAs). The Genetic Algorithm is used to detect the intrusions in networks [3, 19]. It considers both temporal and spatial information of network connections during the encoding of the problem using Genetic Algorithm. Data mining is an efficient method for intrusion detection, which can dig out the unknown knowledge and rules from a large number of network data or audit data from host. The Genetic Algorithm is more helpful for identification of network anomalous behaviors. The Rough Set Neural Network Algorithm is used to reduce a number of computer resources required to detect an attack. The KDDCup'99 dataset is used to test the data and gives the better and robust result. The various feature reduction techniques such as Independent Component Analysis, Linear Discriminate Analysis and Principal Component Analysis used to reduce the computational intensity. KDD cup 99 dataset is used to reduce computation time and improves the accuracy of the systems. Section-I gives the introduction of the intrusion detection. Section-II gives the traffic feature of intrusion detection system. Problem formulations in intrusion detection have been reviewed in section-III. Section IV discusses comparative result evaluation. Finally, section-V is the conclusion and future scope.

II.TRAFFIC FEATURE OF NETWORK

The generation of network traffic is very large amount, processing of this traffic data is very difficult for firewall, intrusion detection system and other security analysis of tools. The generated traffic is not formatted, due to this reason the classification of traffic categories is very difficult. For the analysis of traffic data used KDD mining tools and converted into connection and sequence data [12]. These sequence and connection data have 42 features on different categories such as basic feature data, content feature, time based feature and host traffic based feature.

1. Basic Features: - the basic feature of these categories gets information from packet header without information of payload. The content of these categories is 1 to 8.

2. Content Features: - In this group new TCP packets analyzed with help of area information. An example of this category is number of "hot" indicator.

3. Time-based Traffic Features: - for gathering these types of features a window of 2 second interval is defined. In this interval, some properties of packets are measured. For example number of connections to the same service as the current connection in the past two seconds.

4. Host-based Traffic Features: - In this category instead of a time based window, a number of connections are used for building the window. This category is designed so that attacks longer than 2 second can be detected.
The processing of feature and description of feature discuss in table 1, 2 and 3 according to their description and data type.

Table 1: Basic features of individual TCP connections

Feature name	Description	Type
hot	number of ``hot" indicators	continuous
num_failed_logins	number of failed login attempts	continuous
logged_in	1 if successfully logged in; 0 otherwise	discrete
num_compromised	number of ``compromised'' conditions	continuous
root_shell	1 if root shell is obtained; 0 otherwise	discrete
su_attempted	1 if ``su root" command attempted; 0 otherwise	discrete
num_root	number of ``root" accesses	continuous
num_file_creations	number of file creation operations	continuous
num_shells	number of shell prompts	continuous
num_access_files	number of operations on access control files	continuous
num_outbound_cmds	number of outbound commands in an ftp session	continuous
is_hot_login	1 if the login belongs to the ``hot" list; 0 otherwise	discrete
is_guest_login	1 if the login is a ``guest"login; 0 otherwise	discrete

Table 2: Content features within a connection suggested by domain knowledge.

Feature name	Description	Type
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror_rate	% of connections that have ``SYN" errors	continuous
rerror_rate	% of connections that have ``REJ" errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
srv_serror_rate	% of connections that have ``SYN" errors	continuous
srv_rerror_rate	% of connections that have ``REJ" errors	continuous
srv_diff_host_rate	% of connections to	continuous

Table 3: Traffic features computed using a two- second time window.

Feature name	Description	Type
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror_rate	% of connections that have ``SYN" errors	continuous
rerror_rate	% of connections that have ``REJ" errors	continuous
same_srv_rate	% of connections to the same service	Continuous
diff_srv_rate	% of connections to different services	Continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	Continuous
srv_serror_rate	% of connections that have ``SYN" errors	Continuous
srv_rerror_rate	% of connections that have ``REJ" errors	Continuous
srv_diff_host_rate	% of connections to different hosts	continuous

Feature selection and feature reduction is an important data processing step prior to perform of intrusion detection technique [17]. Feature optimization and feature reduction process used some heuristic function such as genetic algorithm, particle of swarm optimization and neural network. In the all categories of feature some features play ideal role in connection stream in case of normal connection and abnormal connection. if reduces these feature improve the performance of intrusion detection technique.

III PROBLEM FORMULATION

The environment in which the feature extraction and feature reduction is done it is crucial section for intrusion detection. This means that the network traffic contains user confidential information. In general, only the header fields of the packets can be checked but not the user data in the payload. Scalability is an issue with IDS. Because of the huge amount of data flowing through the mobile operator's network, it is not an easy task to find out the right information needed for IDS. The problem is to find an answer to the question: "What features need to be taken into account when calculating or analyzing whether the activity is malicious or not?"Based on prior research on IDS it is clear that either one of the techniques alone cannot detect everything but the combination of the both is the most promising approach. For example misuse detection can be used to filter known threats from the traffic to make it easier for the anomaly detection system to focus on the unknown. Even though IDS have been researched over 20 years, we still do not have an answer to the question of what features should be monitored. So far different kinds of methods and algorithms have been developed for anomaly detection but the focus has been on making them more efficient. Almost all of them are lacking the same information; what features are important for IDS, especially in telecommunications networks? For some reason information on the used features is not easily found from IDS research publications. No matter what the reason is the result is the same; every researcher has to figure out by themselves which features should be used for the monitoring.

1. The pre-processing of KDDCUP99 takes more time.
2. The rate of false alarm generation is high.
3. Some data mining classifier are ambiguous situation for selection of base classifier
4. Entropy based intrusion detection system suffered by high false rate
5. The detection of dynamic feature evaluation.

IV COMPARATIVE STUDY OF DETECTION TECHNIQUE

In this section discuss the comparative study of intrusion detection based on machine learning and feature optimization. The study used method detection rate and finally demerits of method.

Table 4: Comparative study of different intrusion detection techniques.

Sr No.	Method/Approach used	Rate of deduction	Demerits
1.	Fuzzy Genetic algorithm [1]	Average detection rate 97.00%	Used only limited number of data attribute
2.	Neural network classifier [2]	Approximate efficiency 97.50%	Efficiency can be improve further with using feature reduction
3.	Learning Lamstar Neural Network [4]	Improved efficiency 98.00%	SOM which shows poor result for PROBE class.
4.	Decision tree classifier with feature based GA techniques [5]	Improved classification rate	Decision tree not included the forest tree.
5.	Feature reduction with KNN and Bayes classifier [7]	Eliminate non-useful information based feature	Sometimes Computation cost may not be supported
6.	Random forest classifier with SMOTE	Build a model in less time	Not useful for Real time adaptive ID System

7.	Genetic Approach [9]	Improved model for misuse detection system	Method support only some feature in KDDCUP
8.	Data mining and neural network approach [10]	Detection rate for RBF is 98.50%	Results not compared with any PCA Approach
9.	Attribute selection and classification method [11]	Gives Accuracy is 99.00%	Model does not apply with Real world
10.	Unsupervised Neural network [13]	Accuracy 97.00% with ART and 95.00% SOM nets	ART-2 offered a little lower detection rate performance than ART-1
11.	Genetic Algorithm [14]	Accuracy for training data set is 97.25% and testing set is 95.00%	For the Large training rate of datasets, it is neither efficient nor feasible
12.	Fuzzy Based Divide Conquer Algorithm [16]	Classification rate is 99.96%	Works only for selected features
13.	Feature selection and Ensemble method [16]	Accuracy for Probe is 100% and for R2L is 99.97%	U2R gives not better result, it's suffered
14	FNN-SVM [21]	Average accuracy performance is 97%	Very complex model used more time for execution

V CONCLUSION AND FUTURE SCOPE

In this paper study of intrusion detection technique using machine learning and feature optimization technique. In the study we seen that features of network data is very complex due to mixed categories. For the classification task feature selection and feature optimization is important technique. For the optimization of feature and reduction of feature used neural network technique, genetic algorithm and particle of swarm optimization technique. We also saw that merging a different classification technique also improved the detection of intrusion detection. All the method used in this survey gives average 98% detection rate with false alarm generation. In future increase the detection rate approximate 100 % and reduces the false alarm generation.

REFERENCES

[1] P. Jongsuebsuk , N. Wattanapongsakorn, C. Charnsripinyo "Network Intrusion Detection With Fuzzy Genetic Algorithm For Unknown Attacks" IEEE, (Icoin) 2013. Pp 1-5.

[2] S.Devaraju , Dr. S.Ramakrishnan "Performance Analysis Of Intrusion Detection System is Using Various Neural Network Classifiers" In IEEE International Conference On Recent Trends In Information Technology, Icrtit 2011. Pp 1033-1038.

[3] Li Liu, Pengyuan Wan, Yingmei Wang, Songtao Liu "Clustering And Hybrid Genetic Algorithm Based Intrusion Detection Strategy " Telkomnika Indonesian Journal Of Electrical Engineering Vol.12, 2014. Pp 762-770.

[4] V.Venkatachalam and S.Selvan "Intrusion Detection Using An Improved Competitive Learning Lamstar Neural Network" In Ijcsns International Journal Of Computer Science And Network Security, Vol.7, 2007. Pp 255-264.

[5] Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua "Decision Tree Classifier For Network Intrusion Detection With Ga-Based Feature Selection" 2011. Pp 1-6.

[6] Mohammad A. Faysel , Syed S. Haque "Towards Cyber Defense: Research In Intrusion Detection And Intrusion Prevention Systems" Ijcsns International Journal Of Computer Science And Network Security, Vol.10, 2010. Pp 316-325.

[7] Shafigh Parsazad , Ehsan Saboori , Amin Allahyar "Fast Feature Reduction In Intrusion Detection Datasets" Mipro 2012, Pp 1023-1029.

[8] Abebe Tesfahun, D. Lalitha Bhaskari "Intrusion Detection Using Random Forests Classifier With Smote And Feature Reduction" International Conference On Cloud & Ubiquitous Computing & Emerging Technologies, 2013. Pp 127-133.

[9] Zorana Bankovic , Dusan Stepanovic, Slobodan Bojanic, Octavio Nieto-Taladriz "Improving Network Security Using Genetic Algorithm Approach" Computers And Electrical Engineering, 2007. Pp 1-14.

[10] Hachmi Fatma , Limam Mohamed "A Two-Stage Technique To Improve Intrusion Detection Systems Based On Data Mining Algorithms" IEEE, 2013. Pp 1-6.

[11] Tansel Ozyer , Reda Alhajj , Ken Barker "Intrusion Detection By Integrating Boosting Genetic Fuzzy Classifier And Data Mining Criteria For Rule Pre-Screening" Journal Of Network and Computer Applications 2007. Pp 99-113.

[12] Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nguyen Huu Hoa, Mohammad Zahidur Rahman."Adaptive Network Intrusion Detection Learning: Attribute Selection And Classification" 2008. Pp 1-5.

[13] Morteza Amini, Rasool Jalili, Hamid Reza Shahriari "Rt-Unnid: A Practical Solution To Real-Time Network-Based Intrusion Detection Using Unsupervised Neural Networks" Computers & Security, 2006. Pp 459-468.

[14] Ren Hui Gong, Mohammad Zulkernine, Purang Abolmaesumi "A Software Implementation Of A

ISSN: 2395-2946

Genetic Algorithm Based Approach To Network Intrusion Detection" IEEE, 2005. Pp 250-258.

[15] Ajith Abraham, Ravi Jain, Johnson Thomas, Sang Yong Han "D-Scids: Distributed Soft Computing Intrusion Detection System" Journal Of Network And Computer Applications, 2005. Pp 1-19.

[16] Anish Das And S. Siva Sathya "A Fuzzy Based Divide And Conquer Algorithm For Feature Selection In Kdd Intrusion Detection Dataset" International Journal Of Computer Science And Informatics, Vol-2, 2012. Pp 46-50.

[17] Srilatha Chebrolu , Ajith Abraham , Johnson P. Thomas "Feature Deduction And Ensemble Design Of Intrusion Detection Systems" Computers & Security, 2005. Pp 295-307.

[18] V. Bapuji , R. Naveen Kuma ,2 Dr. A. Govardhan , S.S.V.N. Sarma "Soft Computing And Artificial Intelligence Techniques For Intrusion Detection System" Network And Complex Systems, Vol-2, 2012. Pp 24-33.

[19] Ren Hui Gong, Mohammad Zulkernine And Purang Abolmaesumi "A Software Implementation Of A Genetic Algorithm Based Approach To Network Intrusion Detection" In IEEE 2005.

[20] Jonatan Gomez and Dipankar Dasgupta "Evolving Fuzzy Classifiers for Intrusion Detection" In IEEE 2002.

[21] A.M. Chadrashekhar and K. Raghuveer " Intrusion detection technique by using K-means, Fuzzy Neural network and SVM classifier" (ICCCI), 2013. Pp 450-456.