IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EFFECTIVE INTRUSION DETECTION USING DATA MINING TECHNIQUES

[1]Kasturi Dhabale, [2]Prof. A. A. Chinchamalatpure

[1]ME Student, [2]Assistant Professor

[1]Department of Computer Science and Engineering

[1]Dr. Sau. Kamaltai Gawai Institute of Engineering & Technology, Darapur, Maharastra, India.

Abstract: Intrusion-Detection Systems (IDSs) is an evolving technology for protecting computer networks. For instance, in earlier day's Denial-of-Service (DoS) attack cannot cause serious disasters, but today, successful DoS attacks can cause great financial loss to organizations. The goal of intrusion-detection systems is to detect anomalous or misuse behavior of system and notify to network administrators about the activities. Many intrusions-detection tools have security weaknesses such as failing to encrypt the log files, ignoring access control, and failing to perform integrity checks, etc., An IDS is more secure than other security tools, such as firewalls [2]. Earlier research system based on two major concepts known as anomaly detection and signature detection based on abnormal behavior of the system. Initially IDS consists of collection of audit data from the observed system. Then this data is either preprocessed or directly applied to the detector to generate an alarm. The main aim of IDS is to increase detection rate and to reduce false alarm rate in detecting attacks [5]. Recently, the researcher mainly focused on anomaly detection based on proposed methodologies such as data mining, SVM, Fuzzy Genetic in detecting intrusive behavior of the system.

Index Terms - Data mining, SVM, and Fuzzy Genetic in detecting intrusive behavior.

I. INTRODUCTION

The rapid growth in technology leads to a possibility where computers and networks are under threat from worms, viruses and attacks. The use of devices connected to Internet is increasing every year very rapidly. The increase in the number of network devices has led to an increase in unauthorized activity, not only from external attackers, but also from internal attackers, through various types of intrusions. An intrusion is defined as a set of actions that compromise the integrity, confidentiality or availability of a resource, and is a type of attack that attempts to bypass the security mechanism of a computer system. Intrusion detection [1] is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems.

Computer networks have developed rapidly over the years contributing significantly to social and economic development. International trade, healthcare systems and military capabilities are examples of human activity that increasingly rely on networks. This has led to an increasing interest in the security of networks by industry and researchers. The importance of Intrusion Detection Systems (IDS) is critical as networks can become vulnerable to attacks from both internal and external intruders.

Currently, use of smart IDS is viewed as an effective solution for network security and protection against external threats. However, the existing IDS often has a lower detection rate under new attacks and has a high overhead when working with audit data, and thus machine learning methods have been widely applied in intrusion detection. SVM, one of the machine learning technologies, is a new algorithm based on statistical learning theory higher performance than the traditional learning methods in solving the classification problem GA shows excellent global optimization ability via population search strategies and information exchange between individuals. Different from the traditional multi-point search algorithm, GA can easily avoid local optima. In this paper, GA and SVM are used to select the optimal feature subset and optimize the SVM parameters and feature weights to improve the performance of the network intrusion detection system.

Also several researchers focused on fuzzy rule learning for effective intrusion detection using data mining techniques. By taking into consideration these motivational thoughts, we will develop a fuzzy rule based system in detecting the attacks. Proposed system, anomaly-based intrusion detection will make use of effective rules identified in accordance with the designed strategy, which will be obtained by mining the data effectively. The fuzzy rules generated from the proposed strategy will be able to provide better classification rate in detecting the intrusion behavior.

II LITERATURE SURVEY

In recent years plenty of Intrusion detection systems have been developed commercially and noncommercial that identify intrusions in the system. Latest methods are being utilized to get better success rate of such kind of schemes. Data mining techniques could handle huge dataset and permits automation of IDS. Local anomaly detection models have been extended that could identify an intrusion with an immense degree of accurateness. According to the reviewed research, two types of profiling are made.

Some IDS systems sustain a database of probable intrusion action patterns and activate alarm when such action is identified. These systems result in less fake alarms because of a difference in node usage prototypes, however, intrusion behavior with new prototypes are likely to be underreported. Another class of IDS schemes maintains a usual operational profile formed by a learning process. Anything that falls outside such a profile of behavior is categorized as a possible intrusion. These schemes have a superior fake alarm rate, but are more probable to determine unknown intrusions.

Xiao et al [8] presented a method for detection of intrusion that applies GA to identify intrusion in networks during valuable feature selection methods. Their technique utilizes information theory to mine related features and decrease the difficulty. After that, they created a linear structure rule from the selected features in order to categorize network activities into normal and anomalous activities. However, their technique considers only discrete features. Ciaulkns et al [9] stated a dynamic data mining system for identifying anomalies utilizing
decision tree in networks. Gudadhe et al [10] presented a model that utilizes boosted decision tree i.e. hoeffding tree categorization method to amplify the efficiency of the intrusion detection system. Boosting technique improves ensemble performance by utilizing adaptive window and range hoeffding tree like base learner. The primary idea of boosting is to merge simple rules to form an ensemble such that the efficiency of the single ensemble element is improved. The boosting algorithm begins by giving all data training tuples the similar weight w0. Later than a classifier is built the load of every tuple is modified according to the categorization given by that classifier. Then, a second classifier is constructing the reloaded training tuple. The concluding classification of intrusion detection is a loaded average of the individual classifications of overall classifiers.

Pan et al [11] stated a misuse detection scheme utilizing the grouping of neural network and C4.5 algorithm. Gaddam et al [12] stated supervised anomaly detection scheme by cascading KMeans clustering and ID3 Decision Tree learning algorithms. Yasami and Mozaffari [13] stated host based IDS using a grouping of K-Means clustering and ID3 Decision Tree learning algorithms used for unsupervised classification of abnormal and normal behavior in existing networks. In their proposed work, the K-Means clustering algorithm was primarily applied to the normal training data and it was division into K clusters utilizing the Euclidean distance
measure. Decision Tree was created on every cluster using ID3 algorithm. The anomaly score value of the K-Means clustering algorithm and decision rules from ID3 were mined. Resultant anomaly score value was acquired using a special algorithm which merges the output of the two algorithms. The threshold rule was applied for making the decision on the test instance normality. The efficiency of the merged approach was evaluated with individual K-Means clustering, ID3 categorization algorithm and the other approaches based on Markov chains and stochastic learning automata. Improvement in correctness had been monitored in the merged approach when evaluated with other approaches. In almost all study work, SVM has been utilized for categorization of network traffic patterns. The disadvantage with this method is that it obtains a long time for training the scheme. So, it is significant to optimize that difficulty utilizing clustering, fuzzy logic genetic algorithm and neural networks. Platt [14] stated an express training technique for SVM utilizing sequential minimal optimization. Lin and Wang [15] & Tang and Qu [16] stated Fuzzy Support Vector Machines in which a fuzzy membership to every input point was applied to reformulate the SVMs such that different input points can create different contributions to the learning of decision surface. Kim et al [17] stated a GA based approach to get better the capability of the SVM based intrusion detection models utilized in network intrusion detection systems. The rules produced in their research work were more capable in categorization of recognized and unidentified prototypes since the proposed neuro tree detection pattern incorporates neural network to preprocess the data in to amplify the generalization capability.

Khan et al [18] presented a method for optimizing the training time of SVM, mainly when handling huge datasets, utilizing hierarchical clustering analysis. A dynamically rising self-organizing tree algorithm for clustering was utilized by them because it has verified to conquer the problems of existing hierarchical clustering algorithms. Clustering analysis helps in discovering the edge points, which are mainly competent data prototype to train SVM, among two classes, abnormal and normal. Their algorithm added considerably in improving the training stage of SVM with superior generalization precision. A novel algorithm for multiclass SVM was proposed by Guo et al [19]. The tree build in their algorithm consists of a sequence of two class SVMs. Considering both reparability and balance, in every iteration multiclass prototypes are separated into two sets according to the distances among pair wise classes and the number of prototypes in every class. This algorithm could well treat with the irregularly distributed difficulties. Lei and Zhao [20] projected a model utilized on IDS; this model is based on Rough Set theory and Fuzzy Support Vector Machines (RS-FSVM). Experimental results were shown that the RS-FSVM achieves the most excellent recognition capacity.

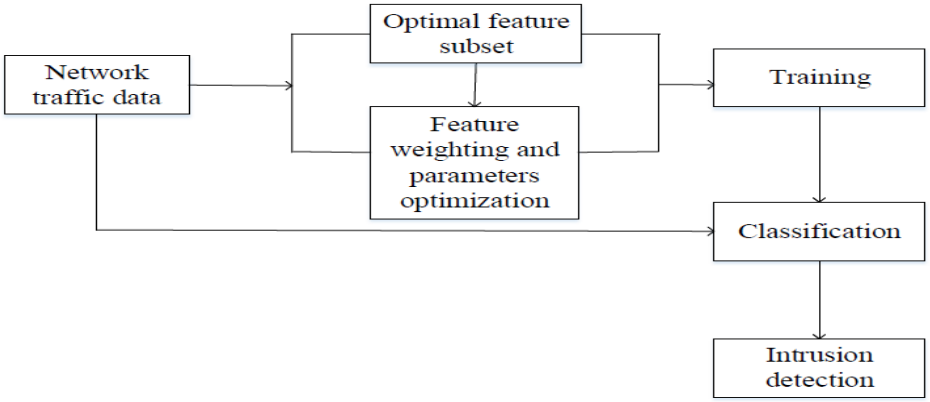
Mulay et al [21] projected the IDS based on SVM utilizing decision tree. Chen et al [22] applied Support Vector Machine to multiclass categorization difficulties and solved the multiclass error diagnosis tasks. They measured the restrictions of conventional techniques, and hence projected the Decision Tree based SVM (DTSVM) that utilizes genetic algorithm (GA) which preserves the higher generalization ability. In their work, decision tree was created by utilizing the GA with maximum distance to create the two subclasses as divisible as possible and hence offers relatively improved generalization capability in the majority of cases.

Yi et al [23] projected a modified Radial Basis kernel Function (U-RBF), through the mean and mean square diversity values of feature attributes inserted in Radial Basis kernel Function (RBF). They recommended an enhanced incremental kernel function U-RBF, which is based on Gauss kernel function. This method decreases the noise between attributes, so the recognition rate of the U - RBF is elevated than RBF. U-RBF plays significant role in saving the training and testing time. This technique is unsuccessful to discover user to root (U2R) and remote to local (R2L) attacks.

III Methodology

3.1 SVM-GA

We propose a combination of the genetic algorithm (GA) with support vector machine (SVM). First, we optimize the crossover probability and mutation probability of GA, generate the population to speed up the search in the early evolution of the population and accelerate the convergence of the algorithm in the later evolution of the population. In the stage of optimal feature set selection, a new fitness function is proposed to decrease the error rate while increasing the true positive rate. Finally, the feature weights and parameters of SVM are optimized simultaneously, and the robustness of SVM is improved.



3.2 System Using Fuzzy Logic

Recently, several researchers focused on fuzzy rule learning for effective intrusion detection using data mining techniques. By taking into consideration these motivational thoughts, we will develop a fuzzy rule based system in detecting the attacks. Proposed system, anomaly-based intrusion detection will make use of effective rules identified in accordance with the designed strategy, which will be obtained by mining the data effectively. The fuzzy rules generated from the proposed strategy will be able to provide better classification rate in detecting the intrusion behavior. The different steps involved in the proposed system for anomaly-based intrusion detection (shown in figure 1) are described as follows:

(1) Classification of training data
(2) Strategy for generation of fuzzy rules
(3) Fuzzy decision module
(4) Finding an appropriate classification for a test input

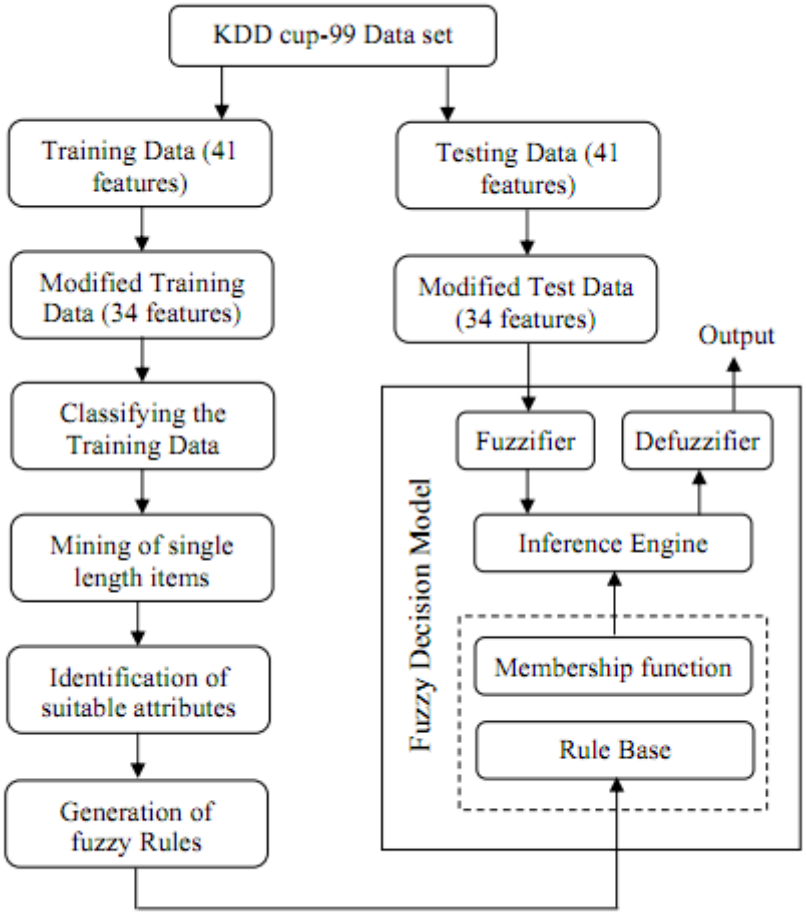


Fig.1. The overall steps of the proposed intrusion detection system

The dataset we will be taken for analyzing the intrusion detection behavior using the proposed system is KDD-Cup 1999 data. The first component of the proposed system is of classifying the input data into multiple classes by taking in mind the different attacks involved in the intrusion detection dataset. Based on the analysis, the KDD-Cup 1999 data contains four types of attacks and normal behavior data with 41 attributes that have both continuous and symbolic attributes. The proposed system will be designed only for the continuous attributes because the major attributes in KDD-Cup 1999 data are continuous in nature. Therefore, we will take only the continuous attributes for instance, 34 attributes from the input dataset by removing discrete attributes. The class label describes several attacks, which comes under four major attacks (Denial of Service, Remote to Local, U2R and Probe) along with normal data. The five subsets of data will be then used for generating a better set of fuzzy rules automatically so that the fuzzy system will learn the rules effectively.

CONCLUSION AND FUTURE WORK

In this paper, a method of applying genetic algorithms with fuzzy logic is presented for network intrusion detection system to efficiently detect various types of network intrusions. To implement and measure the performance of the system I carried out a number of experiments using the standard KDD Cup 99 benchmark dataset and obtained reasonable detection rate. To measure the fitness of a chromosome I used the fuzzy confusion matrix where the fuzzy membership value and fuzzy membership function for the complement of a fuzzy set are two different concepts because the surface value is not always counted from the ground level. The proposed detection system can upload and update new rules to the systems as the new intrusions become known. Therefore, it is cost effective and adaptive. The method suffers from two aspects. Firstly, it generates false alarms which are very serious problem for IDS. Secondly, for high dimensional data, it is hard to generate rules that cover up all the attributes.

REFERENCES

[1] Sourcefire. Snort open source network intrusion prevention and detection system (ids/ips).
[2] Liao, Hung-Jen, Chun-Hung Richard Lin, Ying-Chi Lin, and Kuang-Yuan Tung. "Intrusion detection system: A comprehensive review."_Journal of Network and Computer Applications_36, no. 1 (2013): 16-24.
[3] Debar, Herve. "An introduction to intrusion-detection systems."_Proceedings of Connect_2000(2000).
[4] Liao, Hung-Jen, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. "Intrusion detection system: A comprehensive review."_Journal of Network and Computer Applications_36, no. 1 (2013): 16-24.
[5] Jyothsna, V., VV Rama Prasad, and K. Munivara Prasad. "A review of anomaly based intrusion detection systems."_International Journal of Computer Applications_28, no. 7 (2011): 26-35.
[6] Bashah, Norbik, Idris Bharanidharan Shanmugam, and Abdul Manan Ahmed. "Hybrid intelligent intrusion detection system."_World Academy of Science, Engineering and Technology_11 (2005): 23-26.

[7] Ghosh, Anup K., Aaron Schwartzbard, and Michael Schatz. "Learning Program Behavior Profiles for Intrusion Detection." In_Workshop on Intrusion Detection and Network Monitoring, vol. 51462. 1999.
[8] Xia, Tao, Guangzhi Qu, Salim Hariri, and Mazin Yousi. "An efficient network intrusion detection method based on information theory and genetic algorithm." In_Performance, Computing, and Communications Conference, 2005. IPCCC 2005. 24th IEEE International, pp. 11-17. IEEE, 2005.
[9] Caulkins, Bruce D., Joohan Lee, and Morgan Wang. "A dynamic data mining technique for intrusion detection systems." In_Proceedings of the 43rd annual Southeast regional conference- Volume 2, pp. 148-153. ACM, 2005.
[10] Gudadhe, Mrudula, Prakash Prasad, and Kapil Wankhade. "A new data mining based network intrusion detection model." In_Computer and Communication Technology (ICCCT), 2010 International Conference on, pp. 731-735. IEEE, 2010.
[11] Pan, Zhi-Song, Song-Can Chen, Gen-Bao Hu, and Dao-Qiang Zhang. "Hybrid neural network and C4. 5 for misuse detection." In_Machine Learning and Cybernetics, 2003 International Conference on, vol. 4, pp. 2463-2467. IEEE, 2003.
[12] Gaddam, Shekhar R., Vir V. Phoha, and Kiran S. Balagani. "K-means+ id3: A novel method for supervised anomaly detection by cascading k-means clustering and id3 decision tree learning methods."_Knowledge and Data Engineering, IEEE Transactions on_19, no. 3 (2007): 345-354.
[13] Yasami, Yasser, and Saadat Pour Mozaffari. "A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods."_The Journal of Supercomputing_53, no. 1 (2010): 231-245.
[14] Platt, J.C. “Fast Training of Support Vector Machines using Sequential Minimal Optimization", Advances in Kernel Methods: Support Vector Learning, pp. 185-208, 1998.
[15] Lin, Chun-Fu, and Sheng-De Wang. "Fuzzy support vector machines."Neural Networks, IEEE Transactions on_13, no. 2 (2002): 464-471.
[16] Tang, Hao, and Liang-sheng Qu. "Fuzzy support vector machine with a new fuzzy membership function for pattern classification." In_Machine Learning and Cybernetics, 2008 International Conference on, vol. 2, pp. 768-773. IEEE, 2008.
[17] Kim, Dong Seong, Ha-Nam Nguyen, and Jong Sou Park. "Genetic algorithm to improve SVM based network intrusion detection system." In_Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on, vol. 2, pp. 155-158. IEEE, 2005.
[18] Khan, Latifur, Mamoun Awad, and Bhavani Thuraisingham. "A new intrusion detection system using support vector machines and hierarchical clustering."The VLDB Journal—The International Journal on Very Large Data Bases_16, no. 4 (2007): 507-521.
[19] Guo, Jun, Norikazu Takahashi, and Wenxin Hu. "An efficient algorithm for multi-class support vector machines." In_Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on, pp. 327-331. IEEE, 2008.
[20] Li, Lei, and Ke-nan Zhao. "A new intrusion detection system based on rough set theory and fuzzy support vector machine." In_Intelligent Systems and Applications (ISA), 2011 3rd International Workshop on, pp. 1-5. IEEE, 2011.
[21] Mulay, Snehal, P. R. Devale, and G. V. Garje. "Decision tree based support vector machine for intrusion detection." In_Networking and Information Technology (ICNIT), 2010 International Conference on, pp. 59-63. IEEE, 2010.
[22] Chen, Huanhuan, Qiang Wang, and Yi Shen. "Decision tree support vector machine based on genetic algorithm for multi-class classification."_Systems Engineering and Electronics, Journal of_22, no. 2 (2011): 322-326.
[23] Yi, Yang, Jiansheng Wu, and Wei Xu. "Incremental SVM based on reserved set for network intrusion detection."_Expert Systems with Applications_38, no. 6 (2011): 7698-7707.
[24] Lu, Wei, and Issa Traore. "Detecting new forms of network intrusion using genetic programming."_Computational Intelligence_20, no. 3 (2004): 475-494.
[25] Mabu, S., Chen, C., Lu, N., Shimada, K. and Hirasawa, K. “An Intrusion-Detection Model Basedon Fuzzy Class-Association-Rule Mining Using Genetic Network Programming”, IEEE Trans. On Systems, Man, and Cybernetics, Part C: Applications and Reviews,Vol. 40, No 99, pp. 1-10, 2010
[26] Khayam, Syed Ali, Ayesha Binte Ashfaq, and Hayder Radha. "Joint network-host based malware detection using information-theoretic tools."_Journal in computer virology_7, no. 2 (2011): 159- 172.
[27] Roesch, Martin. "Snort: Lightweight Intrusion Detection for Networks." InLISA, vol. 99, no. 1, pp. 229-238. 1999.
[28] Cisco, I. O. S. "NetFlow." (2008).
[29] Rothberg, Michael S. "Disk drive for receiving setup data in a self monitoring analysis and reporting technology (SMART) command." U.S. Patent 6,895,500, issued May 17, 2005.
[30] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update."_ACM SIGKDD explorations newsletter_11, no. 1 (2009): 10-18.
[31] kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[32] Wu, Su-Yun, and Ester Yen. "Data mining-based intrusion detectors."_Expert Systems with Applications_36, no. 3 (2009): 5605-5612