International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-10)
Research Article | **October 2014**

A Efficient Method for Intrusion Detection using GNP based on Fuzzy Class Association Rule Mining

Ms. Smita D. Shinde*
Department of Computer Engg. & Pune University
Pune, India

Prof. G. M. Bhandari
Department of Computer Engg. & Pune University
Pune, India

Abstract—

Due to increasing internet services, many kinds and a large number of security threats are increasing. Therefore, intrusion detection systems are effectively used for detecting intrusion accesses. This paper describes a novel fuzzy class association rule mining method based on genetic network programming (GNP) for detecting network intrusions.GNP is an evolutionary optimization technique, which uses directed graph structures instead of strings in genetic algorithm or trees in genetic programming, which enhance the representation ability with compact programs derived from the reusability of nodes in a graph structure. By combining fuzzy set theory with GNP, the proposed method can deal with the mixed database that contains both discrete and continuous attributes and also extract many important class association rules that contribute to enhancing detection ability. Therefore, the proposed method can be flexibly applied to both misuse and anomaly detection in network-intrusion-detection problems. There are number of features of proposed methods like sub attribute utilization, accuracy in mining rules and high detection rates. Experimental results with DARPA98 and DARPA99 databases from MIT Lincoln Laboratory shows that the proposed method provides competitively high detection rates(DR) compared with other machine-learning techniques and GNP with crisp data mining.

Keywords— Class-association rule mining, Fuzzy membership function, Genetic Network Programming, Intrusion detection, Sub attribute Utilization

I. INTRODUCTION

Intrusion is defined as the attempts to bypass the security mechanisms of a computer or network. The basic goals of computer security are integrity, confidentiality, and availability. As integrity involves no duplicity in data, confidentiality means privacy of the data and availability involves the presence of the data in the accurate manner when it is to be required. So, Intrusion is a set of unwanted actions aimed to compromise these security goals. To prevent these actions, intrusion prevention (authentication, encryption, etc.) alone is not sufficient. So before Intrusion prevention, Intrusion detection is needed. Rate of detection and strength in detection are two major parameters to evaluate IDS. Techniques like categorization, clustering, artificial neural network, rule based systems and expert system are used in intrusion detection system. These methods can be used in combination to provide better performance

Large number of data mining techniques has been introduced to improve accuracy in detection and constancy in detection but they suffer from large computational difficulty for rule extraction from dense database. In order to detect remarkable rules from a dense database, genetic algorithm (GA) [18] and genetic programming (GP) are together used to form association rule and mining of rules properly. Using this idea a FCRM method based on Genetic Network Programming (GNP) is proposed. By combining fuzzy set theory with GNP, the proposed method deals with the mixed database that contains both isolated and constant attributes. GNP can withdraw rules that include both isolated and constant attributes continuously. Fuzzy sets can help us to defeat sharp boundary problem by allowing different degrees of memberships.

The concept of GNP-based FCRM is introduced in detail. The fuzzy membership values are used for fuzzy rule extraction, and use of sub attributes method is proposed to avoid the information loss. In the meantime, a novel GNP structure for association-rule mining is built up so as to carry out the rule extraction step. In addition, a new fitness function that provides the flexibility of mining more new rules and mining rules with higher accuracy is given in order to detect different kinds of detection.

After the extraction of class -association rules, these rules are used for categorization. Two kinds of classifiers are built up for misuse detection and anomaly detection, respectively, in order to categorize fresh data correctly. For misuse detection, the normal-pattern rules and intrusion-pattern rules are retrieve from the training dataset. Classifiers are built up according to these retrieved rules and for anomaly detection; focus is on retrieving as many normal-pattern rules as possible. Retrieved normal-pattern rules are used to detect novel or unidentified intrusions by evaluating the divergence from the normal behaviour [22].

II. LITERATURE SURVEY

Intrusion detection is categorized into misuse detection and anomaly detection. Misuse detection mainly searches for particular patterns or sequence of programs and user behaviours that match well-known intrusion scenarios and anomaly detection develops models of normal network behaviour, and new intrusions are detected by evaluating considerable divergence from the normal activities.

In order to detect the intrusion, various approaches have been developed and proposed over number of years. One of this approaches, Genetic Network Programming (GNP) procedure and data mining are extensively used. W. Lu and I. Traore has proposed a rule evolution approach based on Genetic Programming (GP) called "Detecting New Forms of Network Intrusion Using Genetic Programming" for detecting novel attacks on network is presented and four genetic operators that is reproduction, mutation, crossover and dropping condition operators are used to develop new rules. New rules are used to detect novel or well-known network attack. But the detection result is not good for some runs because the selection of crossover and mutation points in equivalent operations is arbitrary [1]. Therefore rules extracted using this approach is less which directly affects on accuracy in detection.

A framework for constructing features and models for intrusion detection system (WENKE LEE, SALVATORE J. STOLFO) [5], this article describes a novel framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection. This framework uses data mining algorithms to compute activity patterns from system audit data and extracts predictive features from the patterns. It then applies machine learning algorithms to the audit records that are processed according to the feature definitions to generate intrusion detection rules.

Z. Bankovi´c, D. Stepanovi´c, S. Bojani´c, and O. Nieto - Taladriz have proposed a system for Improving network security using genetic algorithm. This approach is used for evolving and testing new rules for intrusion detection . In this method, the KDD99Cup training and testing dataset are used [4]. But this approach is not works good with mixed database which contains binary and constant attributes.

Chuanhuan Yin, Shengfeng Tian, Houkuan Huang, and Jun He have proposed "Applying Genetic Programming to Evolve Learned Rules for Network Anomaly Detection". In this approach GP is used to evolve new rules from the initial learned rules through genetic operations. GP-based rule learning approach outperforms the original rule learning algorithm. The limitation of this method is that the algorithm needs two passes during training, resulting in the inefficiency of detector [23].

Daniel Barbarra, Julia Couto, Sushil Jajodia, Leonar Popyack, and Ningning Wu have proposed "ADAM: Detecting Intrusions by Data Mining". This method mainly detects probe attacks and DoS attacks but the accuracy in detection is low as compared to other techniques [12]

After analyzing existing approaches, a system is proposed called GNP and Fuzzy Rule Based Intrusion Detection System to solve some drawbacks of the existing system i.e. low detection precision, use of only one kind of database and extraction of minimum number of rules.

III. PROPOSED SYSTEM

In this section, proposed approach is elaborated. First architecture of proposed system is presented. Then modules of this system are described**.**

A. System Architecture

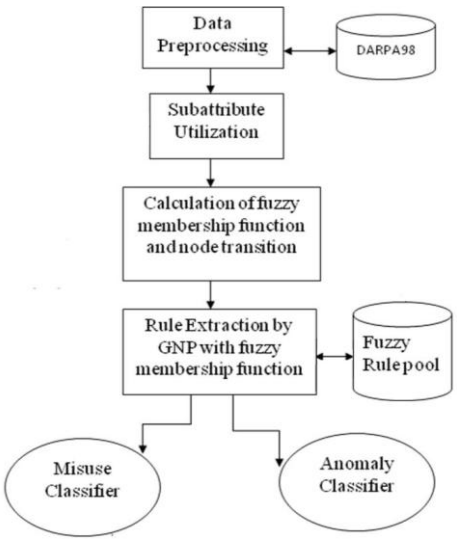

Fig. 1.System Architecture

Data Preprocessing is done on input DARPA98 or DARPA99 dataset. The fuzzy membership values are used for fuzzy rule extraction, and Sub attribute-utilization mechanism is proposed to avoid the information loss. Meanwhile, a new GNP structure for association-rule mining is built up so as to conduct the rule extraction step. In addition, a new fitness function that provides the flexibility of mining more new rules and mining rules with higher accuracy is given in order to adapt to different kinds of detection. After the extraction of class association rules, these rules are used for classification. Two kinds of classifiers are built up for misuse detection and anomaly detection, respectively, in order to classify new data correctly. For misuse detection, the normal-pattern rules and intrusion-pattern rules are extracted from the training dataset. Classifiers are built up according to these extracted rules. While, for anomaly detection, we focus on extracting as many normal-pattern rules as possible. Extracted normal-pattern rules are used to detect novel or unknown intrusions by evaluating the deviation from the normal behavior.

a. Data Pre-processing

First data pre-processing is done. In this the DARPA98 training data includes "list file", which identifies each network connection's time stamps, service type, source IP address, source port, destination IP address, destination port and the type of each attack [3]. Tcptrace utility software [5] is used to extract information on packets to construct new intrinsic features such as data bytes, SYN and FIN packets flowing from the source to the destination as well as from the destination to the source. After the data pre-processing for each network connection, 30 attributes including the list file features, intrinsic features and time based features are obtained.

b. Sub attribute Utilization

Network connection data have their own characteristics, such as discrete and continuous attributes, and these attribute values are important information that cannot be lost. Therefore for avoiding information los s Subattribute utilization method is used. It concern about binary, symbolic, and continuous attributes to keep the completeness of data information. Binary attributes are divided into two subattributes corresponding to judgment functions. For example, binary attribute *A*1 (=land) was divided into *A*_11 (representing land= 1) and *A*_12 (representing land= 0). The symbolic attribute was divided into several subattributes, while the continuous attribute was also divided into three subattributes concerning the values represented by linguistic terms (low, middle, and high) of fuzzy membership functions predefined for each continuous attribute

c. Fuzzy Membership Function for Continuous Attributes and Node Transition

Each continuous attribute is divided into three subattributes with linguistic terms. A predefined membership function is assigned to each continuous attribute and the linguistic terms can be expressed by the membership function
The parameters α, β, and γ in a fuzzy membership function for attribute *Ai* is set as follows:
β = average value of attribute *Ai* in the database
γ = the largest value of attribute *Ai* in the database
$\alpha + \gamma = 2\beta$.

d. Rule Extraction by GNP with Fuzzy Membership Functions

GNP examines the attributes of tuples at judgment nodes and calculates the measurements of association rules at processing nodes [2]. Judgment nodes judge the values of the assigned subattributes, e.g., Land= 1, Protocol=tcp, etc. The GNP-based fuzzy class -association rule mining with sub attribute utilization successfully combines discrete and continuous values in a single rule.
The extracted fuzzy class -association rules are stored in a rule pool through generations. When an important rule is extracted by GNP, it is stored in the pool with its support, confidence, $\chi 2$ value, and the parameters of the fuzzy membership function.
Calculation of $\chi 2$ value of rule $X \rightarrow Y$ is shown as follows. Assume support(X) = x, support(Y) = y, support($X \rightarrow Y$) = z, and the total number of tuples is N.
$\chi 2$ is calculated using following formula

$$\chi 2 = \frac{N(z-xy)^2}{xy(1-x)(1-y)} \qquad (1)$$

If required, a fuzzy rule already stored in the pool would be extracted again. In that case, the membership function and $\chi 2$ value might be changed. If the fuzzy rule has higher $\chi 2$ value, it will replace the same old fuzzy rule in the pool along with its fuzzy parameters. Therefore, the pool is updated every generation and only important fuzzy rules with higher $\chi 2$ values and better-adapted fuzzy parameters are stored.

e. Classifiers for misuse and anomaly detection

After rule Extraction by GNP, classifiers for anomaly detection and misuse detection are built up for classifying new data correctly.

IV. RESULTS

a. Data Sets

The Defence Advanced Research Projects Agency (DARPA) DARPA98 and DARPA99 datasets provided by MIT Lincoln Laboratory are used as training datasets and real-time testing dataset is used for evaluation of system

b. Results

First classification of input training dataset is done as shown in Table 1. The training dataset are DARPA98 and DARPA99. From these datasets connections are classified as normal or intrusion.

Table 1: Classification of input training dataset as DARPA98 and DARPA99

Dataset	**Total number of Normal Connections**	**Total number of Intrusion Connections**
1998	114	432
1999	153	544

After classification of input training dataset, real time testing dataset is classified as normal connection or intrusion connection as shown in Table 2.

Table 2: Classification of input real time testing dataset

Dataset	Total number of Normal Connections	Total number of Intrusion Connections
1998	8	6
1999	10	1

V. CONCLUSION

In this paper, a GNP-based fuzzy class-association-rule mining with Sub attribute utilization and the classifiers based on the extracted rules have been proposed, which can consistently use and combine discrete and continuous attributes in a rule and efficiently extract many good rules for classification. The important function of the proposed method is to efficiently extract many rules that are statistically significant and they can be used for several purposes. When we use them for misuse detection, the matching of a new connection with the normal rules and the intrusion rules are calculated, respectively, and the connection is classified into the normal class or intrusion class. When we use the rules for anomaly detection, only the rules of the normal connections are used to calculate the deviation of a new connection from the normal area. Therefore, many rules extracted by GNP cover the spaces of the classes widely.

VI. FUTURE ENHANCEMENT

By using probability density function based on Fuzzy GNP, enhancement can be done in developed system. In this technique data can be classified as normal or intrusion. In addition a new data can be labeled as normal or intrusion with approximate probability
e.g.95% reliability. The technique will improve detection rate and reduce positive false ratio.

REFERENCES

[1] W. Lu and I. Traore, "Detecting new forms of network intrusion using genetic programming," *Comput. Intell.*, vol. 20, no. 3, pp. 474–494, 2004.
[2] K. Shimada, K. Hirasawa, and J. Hu, "Genetic network programming with acquisition methods of association rules," *J. Adv. Comput. Intell. Intell. Inf.*, vol. 10, no. 1, pp. 102–111, 2006.0
[3] W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 227–261, 2000.
[4] Z. Bankovi´c, D. Stepanovi´c, S. Bojani´c, and O. Niet o-Taladriz, "Improving network security using genetic algorithm approach," *Comput. Elect. Eng.*, vol. 33, pp. 438–451, 2007.
[5] T cptrace Software T ool. [Online]. Available: www.tcptrace.org.
[6] KDDCUP 1999 DAT A [Online]Available: www.11mit .edu
[7] J. Zhang, M. Zulkernine, and A. Haque, "Random -forests based network intrusion detection systems," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 5, pp. 649 –659, Sep. 2008.
[8] T amas Abraham, "IDDM: Intrusion Detection Using Data Mining Techniques", DSTO Electronics and Surveillance Research Laboratory, Salisbury, Australia, May 2001.
[9] R. P. Lippmann, D. J. Fried, I. Graf, J. Haines, K. P. Kendall, D. McClung, D.Weber, S.Webster,D.Wyschogrod, R. K.Cunningham, andM.A. Zissman, "Evaluating intrusion detection systems: The 1998 DARPA offline intrusion detection evaluation," in *Proc. DARPA Inf. Survivability Conf. Expo.*. vol. 2, Los Alamitos, CA: IEEE Comput . Soc. Press, 2000.
[10] K. Hirasawa, T . Eguchi, J. Zhou, L. Yu, and S. Markon, "A doubledeck elevator group supervisory control system using genetic network programming," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 4, pp. 535–550, Jul. 2008.
[11] J. Luo, "Integrating fuzzy logic with data mining methods for intrusion detection," Master's thesis, Dept. Comput . Sci., Mississippi State Univ., Starkville, MS, 1999.
[12] Daniel Barbarra, Julia Couto, Sushil Jajodia, Leonar Popyack, and Ningning Wu, "ADAM: Detecting Intrusions by Data Mining", *Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security T1A3 1100 United States Military Academy*, West Point, NY, June 2001
[13] Jianxiong Luo and Susan M. Bridges, "Mining Fuzzy Association Rules and Fuzzy Frequency Episodes for Intrusion Detection", *International Journal of Intelligent Systems*, Vol. 15, No. 8, pp.687-704, 2000.
[14] W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 227–261, 2000.
[15] D. E. Denning, "An intrusion detection model," *IEEE Trans. Softw. Eng.*, vol. SE-13, no. 2, pp. 222–232, Feb. 1987.
[16] W. Hu, W. Hu, and S. Maybank, "Adaboost-based algorithm for network intrusion detection," *IEEE Trans. Syst.,Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 577–583, Apr. 2008.
[17] Z. Yu, J. J. P. T sai, and T . Weigert, "An automatically tuning intrusion detection system," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 2, pp. 373–384, Apr. 2006.

[18] D. E. Goldberg, *Genetic Algorithm in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[19] K. Shimada, K. Hirasawa, and J. Hu, "Genetic network programming wit h class association rule acquisition methods from incomplete databases," in *Proc. SICE Annu. Conf.*, Kagawa, Japan, 2007, pp. 2708 – 2714.

[20] K. Hirasawa, M. Okubo, H. Katagiri, J. Hu, and J. Murata, "Comparison between genetic network programming (GNP) and genetic programming (GP)," in *Proc. Congr. Evol. Comput.*, 2001, pp. 1276–1282.

[21] T . Eguchi, K. Hirasawa, J. Hu, and N. Ota, "A study of evolutionary multi agent models based on symbiosis," *IEEE Trans. Syst.,Man, Cybern. B, Cybern.*, vol. 36, no. 1, pp. 179–193, Feb. 2006.

[22] C. C. Aggarwal and P. Yu, "Outliers detection for high dimensional data," in *Proc. ACM SIGMOD Conf.*, 2001, pp. 37–46.

[23] Chuanhuan Yin, Shengfeng T ian, Houkuan Huang, and Jun He, " Applying Genetic Programming to Evolve Learned Rules for Network Anomaly Detection", ICNC 2005, LNCS 3612, pp. 323 – 331, 2005

