GRAVITATIONAL SEARCH ALGORITHM FOR FEATURE SELECTION IN INTRUSION DETECTION SYSTEM

VAHID KAVIANI JABALI

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JANUARY 2013

I would like to thanks almighty Allah swt because of the guidance and helping me to complete this project. This project is dedicated to my family for their endless support and encouragement.

ACKNOWLEDGEMENT

First and foremost, all praise be to Allah, for blessings and guidance and for giving me the inspiration to embark on this project. I would like to express our sincerest gratitude to Dr. Anazida Zainal for her supervision and assistance throughout running of this project and for providing invaluable insight and directions whenever needed. Her guidance in this composition course helped me a lot and I'm indebted to her in this regard as well as in bringing this project to a completion. I'm also thankful to my families for their cooperation and support throughout the semester for carrying out and realizing this project.

Besides, I would like to thank the authority of Universiti Teknologi Malaysia (UTM) for providing me with a good environment and facilities such as Computer laboratory to complete this project with software which I need during process.

ABSTRACT

This project was carried out to use the Gravitational Search Algorithm for feature selection in IDS to selectively choose significant features which represents categories of network such as DoS, Probe, U2R and R2L and to improve the accuracy and effectiveness of feature selection and to have better detection. This project aimed to study trends of feature selection in IDS and to implement BGSA for selectively choose features for IDS and to test and validate the performance and feedback of BGSA. The significance of feature selection can be viewed in two aspects. First is to filter out noise and remove redundant and irrelevant features and over load of features which causes significant loss of accuracy and time consumption in detection. In this project, it validates and evaluates the BGSA algorithm and focuses on the feature selection by implementing of BGSA. The results of BGSA program proves that the selected features which proposed by BGSA in terms of accuracy and efficiency are quite acceptable. The comparison of classification rates for all the five classes with other approaches which are using the same dataset shows that the BGSA is more accurate than others.

ABSTRAK

Projek ini dijalankan untuk menggunakan Algoritma Carian Graviti bagi pemilihan ciri dalam IDS supaya ia menjadi selektif dalam memilih ciri-ciri penting yang mewakili kategori rangkaian seperti DoS, Probe, U2R dan R2L dan untuk meningkatkan ketepatan dan keberkesanan pemilihan ciri serta mempunyai pengesanan yang lebih baik. Projek ini bertujuan untuk mengkaji trend pemilihan ciri dalam IDS dan melaksanakan BGSA supaya ia menjadi selektif dalam memilih ciri-ciri untuk IDS dan untuk menguji dan mengesahkan prestasi serta maklum balas daripada BGSA. Kepentingan pemilihan ciri boleh dilihat dari dua aspek. Pertama adalah untuk menapis bunyi dan menghapuskan ciri-ciri yang berlebihan iaitu tidak relevan dan beban lebih ciri-ciri yang menyebabkan kerugian ketara ketepatan dan penggunaan masa dalam pengesanan. Dalam projek ini, ia mengesahkan dan menilai algoritma BGSA dan memberi tumpuan kepada pemilihan ciri dengan melaksanakan BGSA. Keputusan program BGSA membuktikan bahawa ciri-ciri yang dipilih yang dicadangkan oleh BGSA dari segi ketepatan dan kecekapan boleh diterima. Perbandingan kadar klasifikasi untuk kesemua lima kelas dengan pendekatan lain yang menggunakan dataset yang sama menunjukkan bahawa BGSA adalah lebih tepat berbanding yang lain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1

INTRODUTION

1.1 Background Information

By developing of networks and computers, in the same time, keeping data safe and secure in computers becomes one of most interesting and challenging area in Network and security. In spite of the fact that attackers try to achieve the sensitive and critical data to take advantage of them. Due to many motivations, there are plenty number of news about misusing information and attacking computers across the globe which have done by intruders. However, many studies and investigations have been conducted to increase the safety and security of networks and computers; there is various attack and most of them still new and opened scope for research. Today after passing a half of century from emerging computer to the world and growing a vast varieties of countermeasures and mitigation approaches against hackers but the necessity of developing new method for reducing exposure and penetration is undeniable due to arriving more novel attacks day by day.

The progress of computer technology has affected communication technology. From 1980s, many devices have been invented and developed. The progress in the network technology changes the way of communication and data distribution in the world because many businesses and companies use this

technology for trading and marketing their products and contacting their partner and customers properly. Due to the completion and surviving in this generation among all organizations, the importance of safeguard and other countermeasures to stop penetration of intruders to their sensitive or critical information has been raising significantly. To begin with definition in terms of attack, intruder is somebody who can maliciously interrupt, captures, modify, steal or delete important information in the computers and applications by network access or by direct access like run executable code in PC. Attackers use different resources of victim to do the attack. Specifically, they misuse hardware vulnerabilities or software weakness to penetrate the system.

Nowadays security countermeasures such as access control [2] and authentication [3] have been developed to achieve Confidentiality, Integrity and Availability and to block unauthorized intruders from accessing and modifying information. These prevention methods are developed as a front line of defense system. The advantages of the Internet, namely the availability and amount of information, also it is apparent exposure method and the largest threat to the sensitive and critical security. [4] stated that the Intrusion Detection System is second line of defense or detection method against any kind of external threats. The aim of IDS is to identify and preserve computer system from penetrations of intrusions. In fact there are two techniques for detection in IDS systems which are anomaly detection and misuse detection. Different approaches purpose own different technique.

Some examples about intrusion concerns are [1]:

i. Unauthorized modifications in system files or user information.
ii. Illegal access or modification of user files or information.
iii. Unauthorized modifications of system information in network components

For instance: modifications of router tables in an Internet to deny use of the network. Some of the necessary features an intrusion detection system should possess include [1]:

i. Be able to protect them self or be a fault tolerant and run continually with minimum human control. The IDS must recover themselves from system crashes, either accidental or caused by malicious activity.
ii. Be able to work automatically which is preventing an attacker to manipulate the IDS easily. Moreover, the IDS must be able to track any modifications.
iii. Enforce IDS with the optimized overhead on the system to avoid interfering with the normal operation of the system.
iv. The IDS have to be adaptable and configurable in order to changes in system and user behavior over time. In terms of accuracy easy to implement the security policies and user behavior of the systems that are being monitored.
v. Able to detect different types of attacks accurately and must not track any legitimate activity as an intrusion or false positives and conversely at the same time, the IDS must not fail to recognize any real attacks (false negatives).

1.2 Statement of Problem

Based on the other researches done in this area, it is clear that the effectiveness of an IDS model relies on retraining of the reference models and enhancing the recognition of classifiers. One of most important issue in IDS, in order to have better detection, is Feature selection. Feature selection is where a feature subset is selected to represent the data. The significance of feature selection can be viewed in two aspects. First is to filter out noise and remove redundant and irrelevant features and over load of features which causes significant loss of accuracy and time consumption in detection. In this project, it will validate and evaluate the BGSA algorithm and focus on the feature selection by implementing of BGSA will apply GSA algorithm in the feature selection instead of current approaches at end of the project, we'll

determine by the result of this study that this algorithm can be optimized and be more accurate for detection system or not.

1.3 Purpose of the Study

The purpose of this project is to deploy and use the Gravitational Search Algorithm in feature selection of IDS to selectively choose significant features which represents categories of network such as DoS, Probe, U2R and R2L and enable the Intrusion Detection System (IDS) to learn the pattern in network traffic.

1.4 Objectives of the Study

Following are the objectives of this project:

i. To study trends of feature selection in IDS
ii. To implement B-GSA for selectively choose features for IDS
iii. To test and validate the performance and feedback of B-GSA.

The performance of BGSA to do feature selection will be evaluated based on detection accuracy of classifier using the selected features proposed by BGSA.

1.5 Scope of Study

The study is limited to the following:

1. The domain problem is feature selection technique to choose well features by BGSA which is represent sort of traffic classes and these classes will be determined by SVM which allows doing classification and detection accuracy % on classes.
2. Classification of attacks are based on four established dominant categories which are Denial of Service (DoS), Probe, User to Root (U2R) and Remote to Local (R2L) as widely used in other studies in the field of IDS (Abraham et al., 2007; Shafi and Abbas, 2009; Tajbakhsh *et al*., 2009; Farid *et al*., 2010; Teng *et al*., 2010).
3. The data used in this research is KDD Cup 1999 Intrusion Detection data set as widely used by other researchers in the field (Abraham *et al*., 2007; Jemili *et al*., 2007; Shafi an Abbas, 2009; Tajbakhsh *et al*., 2009; Farid *et al*., 2010).

1.6 Significance of the Study

It is important to minimize the recognition time or the time for classification an attack. As an instance, when an attack may already has destroyed the sensitive and critical data which caused disruption of services or denial of service in networks. This project will examine the means of minimizing the time for recognition process using the concept of selective recognition and minimal feature set. Another significance of this project is that efficiency and accuracy classification of an attack has long been researched and continued to be pursued due to the dynamic nature of the network traffic itself. The attacks become more complex and more frequent (higher intensity) which lead to more vulnerable computer network.

1.7 Organization of Report

Chapter 1 or introduction is an overview from the project. Chapter 2 provides background information and a review of related literature that leads to the statement problem. Chapter 3 provides project methodology. Chapter 4 discusses the design and solution approach in improving effectiveness. Chapter 5 is about implementation and results of program. Chapter 6 is about discussion and conclusion which is about the validation of algorithm and performance of program.

REFERENCES

[1] Bishop Matt. Computer security e art and science: Addison Wesley; 2003.

[2] Russel, D. & Gangemi, G.T. 1992. Computer security basics.CA: O=Reilly & Associates Inc. 448p.

[3] Caelli,W., Dennis, L. & Shain, M. 1994. Information Security Handbook. First edition.Wilthire:Macmillan Press Ltd. 833p.

[4] Anderson, J.P. 1980. Computer Threat Monitoring and Surveillance. (In Anderson, J.P. Technical report, Fort Washington, Pennsylvania.) Bishop Matt. Computer security e art and science: Addison Wesley; 2003.

[5] Lee W, Stolfo S, Mok K. A data mining framework for building intrusion detection models. proceedings of the IEEE symposium on security and privacy; 1999a.

[6] Sung AH, Mukkamala S. Identifying important features for intrusion detection using support vector machines and neural networks. In: Proceedings of International Symposium on Applications and the Internet (SAINT 2003); 2003. p.209e17.

[7] Forrest S, Perelson AS, Allen L, Cherukuri R. Self-nonself discrimination in a computer. In: Proceedings of the 1994 IEEE symposium on research in security and privacy. Los Alamitos, CA: IEEE Computer Society Press; 1994.

[8] I. Corona, G. Giacinto, F. Roli ,Intrusion detection in computer systems using multiple classifier systems O. Okun, G. Valentini (Eds.), Supervised and Unsupervised Ensemble Methods and Their Applications, Springer-Verlag, Berling/Heidelber (2008), pp. 91–104

[9] Technical Report James P Anderson Co Fort Washington (1980) : Technical report, James P. Anderson Company, Fort Washington, Pennsylvania, Pages: 56

[10] T. Crothers, Implementing Intrusion Detection Systems, A Hands-On Guide for Securing the Network, Wiley Publishing, Inc., 2003.

[11] Mykerjee, B., et al, 1994, "Network Intrusion Detection", IEEE Network, Vol.8, No.3, pp.26-41.

[12] Forrest, S. et al, (1994) “Self-Nonself Discrimination in aComputer”, Proceeding of 1994 IEEE Symposium onResearch in Security and Privacy, Los Alamos, CA: IEEEComputer Society Press.

[13] Kim and Spafford, 1995 Gene H. Kim, Eugene H. Spafford Experiences with tripwire: using integrity checkers for intrusion detection <http://citeseer.ist.psu.edu/kim95experiences.html> (1995)

[14] Estévez-Tapiadoret al., 2004, J.M. Estévez-Tapiador, P. García-Teodoro, J.E. Díaz-Verdejo Anomalydetection methods in wired networks: a survey and taxonomy Computer Networks, 27 (16) (2004), pp. 1569–1584

[15] Mukkamala et al., 2004b Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham, Vitorino Ramos Intrusion detection systems using adaptive regression splines ,in: I. Seruca, J. Filipe, S. Hammoudi, J. Cordeiro (Eds.), Sixth international conference on enterprise information systems, ICEIS'04, Portugal, vol. 3 (2004), pp. 26–33 ISBN 972-8865-00-7

[16] Giacintoetal., 2003 G. Giacinto, F. Roli, L. Didaci Fusion of multiple classifiers for intrusion detection in computer networks Pattern Recognition Lett., 24 (12) (2003), p. 1795

[17] Krugeletal., 2002 Krugel Christopher, Toth Thomas, Kirda Engin. Service specific anomalydetection for network intrusion detection. In: Proceedings of Symposium on Applied Computing; 2002.

[20] JoIo B. D. Cabrera, et al., "Statistical Traffic Modelin g For Network Intrusion Detection", Proc of the 8th international Symposium on Modeling, Analysis and Sirnulation of Computer and Telecommunication Systems, San Fra ncisco, CA, 2000

[22] Anderson, J. P. (1980). Computer security threat monitoring and surveillance, (Technical Report), Washington, PA, James P. Anderson Co.

[23] [Denning, 1987] Denning,D. E. (1987).An intrusion-detectionmodel. IEEE Transactions on Soft-ware Engineering, 13, 222-232.

[24] Smaha, S. E. (1988). Haystack: An intrusion detection system. Proceedings of the Fourth Aerospace Computer Security Applications Conference (pp. 37-44).

[25] M. Dekker, "Security of the Internet," The Froehlich/Kent Encyclopedia of Telecommunications, Volume 15, pp. 231-255, New York, 1997.

[26] Jolliffe IT. Principal component analysis: Springer-Verlag; 1986. KDD cup 99 intrusion detection data set, !http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gzO.

[27] T.F. Lunt, A. Tamaru, F. Gilham, R. Jagannathm, C. Jalali, P.G. Neumann, H.S. Javitz, A. Valdes, T.D. Garvey, A Real-time Intrusion Detection Expert System (IDES), Computer Science Laboratory, SRI International, Menlo Park, CA, USA, Final Technical Report, February 1992.

[28] D. Anderson, T. Frivold, A. Tamaru, A. Valdes, Next- generation intrusion detection expert system (NIDES), Software Users Manual, Beta-Update release, Computer Science Laboratory, SRI International, Menlo Park, CA, USA, Technical Report SRI-CSL-95-0, May 1994

[29] P.A. Porras, P.G. Neumann, EMERALD: event monitor- ing enabling responses to anomalous live disturbances, in: Proceedings of the 20th NIST-NCSC National Information Systems Security Conference, Baltimore, MD, USA, 1997, pp. 353–365.

[30] E. Tombini, H. Debar, L. Me ´, M. Ducasse ´, A serial combination of anomaly and misuse IDSes applied to HTTP traffic, in: Proceedings of the 20th Annual Computer Security Applications Conference, Tucson, AZ, USA, 2004.

[31] Te-Shun Chou, Kang K. Yen, Liwei An, Niki Pissinou, and Kia Makki. (October, 2007). Fuzzy Belief Pattern Classification of Incomplete Data. IEEE International Conference on Systems, Man and Cybernetics, pp. 535-540, Montreal, Quebec, Canada.

[32] Haykin, S. (1999). "Neural Networks: A Comprehensive Foundation." 2nd. Ed. Upper Saddle River, N.J: Prentice Hall.

[33] Bishop, C. M. (1995). Neural networks for pattern recognition. England: Oxford University

[34] Manocha, S., & Girolami, M. A. (2007). An empirical analysis of the probabilistic K- nearest neighbour classifier. Pattern Recognition Letters, 28, 1818–1824.

[35] Vapnik, V. (1998). Statistical learning theory. New York: John Wiley. Wang, W., & Battiti, R. (2006). Identifying intrusions in computer networks with principal component analysis. In Paper presented at the proceedings of the first international conference on availability, reliability and security (ARES'06).

[37] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59–69.

[38] Mitchell, T. (1997). Machine learning. New york: McGraw Hill. Moradi, M., & Zulkernine, M. (2004). A neural network based system for intrusion detection and classification of attacks. In Paper presented at the proceeding of the 2004 IEEE international conference on advances in intelligent systems – Theory and applications. Luxembourg.

[39] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, P. J. (1984). Classification and regressing trees. California: Wadsworth International Group.

[40] Pearl, J. (1988). Probabilistic reasoning in intelligent systems. Morgan Kaufmann. Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. Journal of Network and Computer Applications, 30, 114–132.

[41] Koza, J. R. (1992). Genetic programming: On the programming of computers by means of natural selection. Massachusetts: MIT.

[42] Zimmermann, H. (2001). Fuzzy set theory and its applications. Kluwer Academic Publishers.

[43] Jang, J.-S., Sun, C.-T., & Mizutani, E. (1996). Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. New Jersey: Prentice Hall.

[44] Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226–239.

[45] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, 1996, pp. 123–140.

[46] L. Breiman,"Random Forests,"Machine Learn- ing, vol. 45, no. 1, 2001, pp. 5–32.

[47] Chebrolu Srilatha, Abraham Ajith, Thomas Johnson. Hybrid feature selection for modeling intrusion detection systems. In: Pal NR, et al, editor. 11th International conference on neural information processing, ICONIP'04. Lecture Notes in Computer Science. vol. 3316. Germany: Springer Verlag; 2004. p. 1020e5. ISBN 3-540-23931-6.

[48] KDD'99 archive: The Fifth International Conference on Knowledge Discovery and Data Mining. URL: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (Last browsed in December 2007)

[49] DARPA Intrusion Detection Evaluation, MIT Lincoln Laboratory. URL: http://www.ll.mit.edu/IST/ideval/ (Last browsed in December 2007)

[50] W. Lee, S.J. Stolfo, K.W. Mok, Adaptive intrusion detection: a data mining approach, Artif. Intell. Rev. 14 (6) (2000) 533–567.

[51] Shelly Xiaonan Wu, , Wolfgang Banzhaf The use of computational intelligence in intrusion detection systems: A review Memorial University of Newfoundland, St John's, NL A1B 3X5, Canada

[52] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

[53] A. Zainal, M. A. Maarof, and S. M. Shamsuddin, "Using Rough-PSO in Anomaly Intrusion Detection."

[54] Srilatha Chebrolua, Ajith Abrahama, Johnson P. Thomas (2004) Feature deduction and ensemble design of intrusion detection systems , Oklahoma State University

[55] Shailendra K. Preeti J. ,Effective Anomaly based Intrusion Detection using Rough Set Theory and Support Vector Machine, International Journal of Computer Applications (0975 – 8887), Volume 18– No.3, March 2011

[56] Esmat Rashedi, Hossein Nezamabadi-pour, and Saeid Saryazdi, "Gsa: A gravitational search algorithm," Information Sciences, vol. 179, no. 13, pp. 2232–2248, 2009.