INTRUSION DETECTION SYSTEMS USING K-MEANS CLUSTERING SYSTEM

NOR DZUHAIRAH HANI BINTI JAMALUDIN

Bachelor of Computer Science (Software Engineering) with Honors

UNIVERSITI MALAYSIA PAHANG

INTRUSION DETECTION SYSTEM USING K-MEANS CLUSTERING SYSTEMS

NOR DZUHAIRAH HANI BINTI JAMALUDIN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science (Software Engineering)

Faculty of Computer Systems and Software Engineering

UNIVERSITI MALAYSIA PAHANG

2019

ACKNOWLEDGEMENTS

I would like to express my gratitude and thanks to Encik Mohd Hafiz bin Mohd Hassin, my supervisor, who gave me a thorough guidance in this research from the beginning until the end. He provided me various advices and insights for me to understand the flow of this research more.

Next, I would like to thank my friends, classmates, and lecturers for additional supports and help during the progress to finish this research. They gave additional comments and opinions to enhance this research.

Last but not least, I am thankful for the endless support I received from my parents, Tuan Haji Jamaludin bin Abdullah and Puan Hajah Siti Nuriah binti Parlan, also my family members. It is a bless to have their encouragement to push me further to complete this research.

ABSTRACT

Internet is the biggest platform for people all over the world to connect with each other, and to search for important information and such. Along with the raising of internet usage, the number of cases of intrusion attacks also increases. Because of this, intrusion detection is important, especially for large companies which held huge and confidential data and information. This system works to detect the abnormal connection of a network so that a stronger protection could be build. Since the attack is not restricted into only one type, a data mining technique is applied to classify all types of attack from a huge amount of data entering into a network. By using this technique, the intrusion detection system could work better. In this research, data mining technique of K-means clustering system are used to detect the intrusion and attack. 1999 KDD Cup Dataset is used for training, testing, and validation of the system. The dataset is famous among intrusion detection system researcher for its data which resembles real attacks at real times.

ABSTRAK

Internet ialah platform terbesar yang digunakan oleh segenap dunia untuk berhubung dengan satu sama lain, dan mencari maklumat penting dan terkini. Dengan peningkatan jumlah penggunaan internet, jumlah kes-kes yang melibatkan serangan pencerobohan juga turut meningkat. Disebabkan oleh hal ini, pengesanan pencerobohan adalah penting, terutama sekali untuk syarikat-syarikat besar yang mempunyai data dan maklumat yang banyak dan sulit. Sistem ini berfungsi untuk mengesan sambungan rangkaian yang abnormal supaya satu sistem perlindungan yang lebih kuat boleh dibina. Oleh kerana serangan pencerobohan tidak terhad kepada satu jenis sahaja, teknik perlombongan data diaplikasikan untuk mengklasifikasikan semua jenis serangan daripada jumlah data yang banyak yang memasuki sesebuah rangkaian. Dengan menggunakan teknik ini, sistem pengesanan pencerobohan boleh berfungsi dengan lebih baik. Dalam kajian ini, teknik perlombongan data yang digunakan ialah sistem pengelompokan K-means digunakan untuk mengesan pencerobohan dan serangan. Set data yang digunakan untuk latihan, ujian, and pengesahan sistem adalah 1999 KDD Cup Dataset. Set data ini adalah terkenal dikalangan pengkaji pengesanan pencerobohan kerana data nya adalah menyerupai serangan yang sebenar di masa yang sebenar.

TABLE OF CONTENT

CHAPTER 5 CONCLUSION

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CFS	Correlation-based feature selector
DARPA	Defence Advanced Research Projects Agency
DOS	Denial of Service
EM	Expectation Maximum
FN	False Negative
FP	False Positive
IDS	Intrusion Detection System
KDD	Knowledge Discovery in Database
KNN	K-Nearest Neighbor
MAP	Maximum a Posteriori
R2L	Remote to Local
TN	True Negative
TP	True Positive
U2R	User to Remote

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Nowadays, internet is the main means of connection between people around the world. Almost everywhere in the world have internet connection, whether it is as wired or wireless. Without proper use of the internet, one could be exposed to the risk of having cyber intruders attacking them. This would let the perpetrators to gain personal data or company confidential illegally. There are a lot of ways for the intruders to intrude such as unauthorized access, denial of service and sending of viruses. In Malaysia, there was a total of 5,078 cybercrimes incidences reported in between January and August 2018 based on Cybercrime Malaysia statistics. Fraud and intrusion each has the total of 2,907 and 699 cases, thus making the two incidents top the charts of MyCERT's statistic.

Intrusion Detection System (IDS) is a system for the users to detect the anomalies of their internet network traffic. There are two categories of IDS which are anomaly detection and misuse detection. Misuse detection will match the computer activity with the stored signatures of the attacks the systems have known. Anomaly detection is a method to detect intrusions by learning the attribute of normal activity, and it will detect anything that is different from the normal activity.

1.2 PROBLEM STATEMENT

Cybercrimes incidents reached thousands of reports every year. This shows that most of the internet users are still not aware about the problems of the cybercrime and cyber intruder around them. This lack of awareness might be because the users do not know how to check their network traffic activity, and thus, making them vulnerable to the attack of the intruder. Some users would also think that they would never have any intruder because of their position of not being someone important or somebody with high rank. In reality, all internet users should be able to detect their own network traffic activity.

In order to help internet users for this reason, we proposed a research of Intrusion Detection System which is based on misuse and anomaly detection systems. The system would be easy to use, thus making anyone can learn it. This would help them to know whether they are being attacked or not, and be able to find a solution to tighten their cyber security.

1.3 OBJECTIVE

The main objective of this research is to find the right systems to detect the abnormal network traffic activity using IDS. The objectives are:

i. To design an effective system for cyber intrusion detection to detect abnormal network traffic activity.
ii. To develop Intrusion Detection System using K-Means clustering algorithm.
iii. To test the findings and performance of the proposed Intrusion Detection System.

1.4 SCOPE

The scope of this research is divided into several categories, which are;

i. Dataset
   a. The dataset that will be used in this research is the 1999 KDD Dataset Intrusion Detection Evaluation Data.

ii. Language
   a. The programming language to implement the algorithm used is Java.

1.5 THESIS ORGANIZATION

This thesis consists of five chapters. Chapter 1 will discussed about the introduction of the research, with its problem statement, objectives, and scope. In this chapter, the definition of intrusion detection systems is explained.

Chapter 2 discusses about the literature review. This consists of the review of related works, and to choose the appropriate algorithm to be used in the system.

Chapter 3 is about the methodology to be used in this work. A proposal to solve the existing problem and the flow of conducting the research is being discussed in this chapter.

Chapter 4 shall discuss the implementation on the proposed algorithm towards the dataset. The results of the implementation will be analyzed and discussed.

Chapter 5 is the summary of this research. A summarized version about the experiments and the constraints of the experiment are concluded in this chapter.

REFERENCES

Adetunmbi, A. O., Falaki, S. O., Adewale, O. S., & Alese, B. K. (2008). Network Intrusion Detection Based on Rough Set and K-Nearest Neighbour. *International Journal of Computing and ICT Research*, *2*(1), 60–66.

Ahmad, H., Uppal, M., Javed, M., & Arshad, M. J. (2014). An Overview of Intrusion Detection System (IDS) along with its Commonly Used Techniques and Classifications. *International Journal of Computer Science and Telecommunications*, *5*(2), 20–24.

Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017). Evaluation of Machine Learning Algorithms for Intrusion Detection System, (Iv), 277–282.

Amudha, P., Karthik, S., & Sivakumari, S. (2013). Classification Techniques for Intrusion Detection – An Overview. *International Journal of Computer Applications*, *76*(16), 975–8887. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.2060&rep=rep1&type=pdf

Ashoor, A. S., & Gore, S. P. (2011). Importance of Intrusion Detection System (IDS). *International Journal of Scientific and Engineering Research*, *2*(1), 1–6. http://doi.org/10.1016/j.jnca.2012.08.007

Balogun, A. O., Balogun, A. M., Adeyemo, V. E., & Sadiku, P. O. (2015). A Network Intrusion Detection System : Enhanced Classification via Clustering, *6*(4), 53–58.

Bharti, K. K., Shukla, S., & Jain, S. (2010). Intrusion detection using clustering. *PROCEEDING OF ACCTA International Conference*, *1*(2), 158–165.

Chae, H., Jo, B., Choi, S., & Park, T. (2013). Feature Selection for Intrusion Detection using NSL-KDD. *Recent Advances in Computer Science 20132*, 184–187.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (n.d.). The Knowledge Discovery Process. *Data Mining*, 9–24. http://doi.org/10.1007/978-0-387-36795-8_2

Darra, E., & Katsikas, S. K. (2017). A survey of intrusion detection systems in wireless sensor networks. *Intrusion Detection and Prevention for Mobile Ecosystems*, 393–458. http://doi.org/10.1201/b21885

Denning, D. E. (1987). An Intrusion-Detection Model. *IEEE Transactions on Software Engineering*, *SE-13*(2), 222–232. http://doi.org/10.1109/TSE.1987.232894

Depren, O., Topallar, M., Anarim, E., & Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. *Expert Systems with Applications*, *29*(4), 713–722. http://doi.org/10.1016/j.eswa.2005.05.002

Dewa, Z., & A., L. (2016). Data Mining and Intrusion Detection Systems. *International Journal of Advanced Computer Science and Applications*, *7*(1). http://doi.org/10.14569/IJACSA.2016.070109

Dias, L. P., Cerqueira, J. J. F., Assis, K. D. R., & Almeida, R. C. (2017). Using artificial neural network in intrusion detection systems to computer networks. *2017 9th Computer Science and Electronic Engineering Conference, CEEC 2017 - Proceedings*, 145–150. http://doi.org/10.1109/CEEC.2017.8101615

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems*, *13*(2–3), 99–115. http://doi.org/10.1016/S0167-739X(97)00015-0

García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers and Security*, *28*(1–2), 18–28. http://doi.org/10.1016/j.cose.2008.08.003

Gecchele, G., Rossi, R., Gastaldi, M., & Caprini, A. (2011). Data Mining methods for Traffic monitoring data analysis: A case study. *Procedia - Social and Behavioral Sciences*, *20*(December), 455–464. http://doi.org/10.1016/j.sbspro.2011.08.052

Gendreau, A. A., & Moorman, M. (2016). Survey of intrusion detection systems towards an end to end secure internet of things. *Proceedings - 2016 IEEE 4th International Conference on Future Internet of Things and Cloud, FiCloud 2016*, 84–90. http://doi.org/10.1109/FiCloud.2016.20

Goeschel, K. (2016). Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis. *Conference Proceedings - IEEE SOUTHEASTCON*, *2016–July*. http://doi.org/10.1109/SECON.2016.7506774

Guide, S. (n.d.). Today’s enterprise-security reality: Devices, data, and applications are outside of your physical control—and all of it must be managed and secured.

Gupta, M. (2015). Hybrid Intrusion Detection System: Technology and Development. *International Journal of Computer Applications*, *115*(9), 975–8887.

Hofmeyr, S. A. (1998). Intusion Detection Using Sequences of System Calls.pdf.

Jianliang, M., Haikun, S., & Ling, B. (2009). The application on intrusion detection based on K-means cluster algorithm. *Proceedings - 2009 International Forum on Information Technology and Applications, IFITA 2009*, *1*, 150–152. http://doi.org/10.1109/IFITA.2009.34

Juvonen, A. (2014). *Intrusion detection applications using knowledge discovery and data mining*.

Katayama, N., Nohara, O., Moriyama, H., & Fujimaki, H. (1994). Attempt to isolate mast-cell precursors based on the differential sensitivity to UV-B and X-irradiation. *Toxic Substances Journal*, *13*(2), 85–95. http://doi.org/10.1609/AIMAG.V17I3.1230

Kemmerer, R., & Vigna, G. (2002). Intrusion detection: A brief history and overview (supplement to computer magazine). *Computer*, 27–30. http://doi.org/http://doi.ieeecomputersociety.org/10.1109/MC.2002.10036

Khanbabapour, H., & Mirvaziri, H. (2014). An Intelligent Intrusion Detection System Based On Expectation Maximization Algorithm in Wireless Sensor Networks. *International Journal of Information*, *4*(1), 1–10. Retrieved from http://esjournals.org/journaloftechnology/archive/vol4no1/vol4no1_1.pdf

Kumar, S., & Chandel. (2017). Intrusion Detection System using K-Means Data Mining and Outlier Detection Approach.

Lakshmi, S. V., & Prabakaran, T. E. (2014). Application of k-Nearest Neighbour Classification Method for Intrusion Detection in Network Data. *International Journal of Computer Applications*, *97*(7), 975–8887. Retrieved from https://pdfs.semanticscholar.org/b364/bdcead6644066805f37dc755fa43d3b466b2.pdf

Laskov, P., Patrick, D., & Sch, C. (2011). Image Analysis and Processing – ICIAP 2011, *6978*(September 2005). http://doi.org/10.1007/978-3-642-24085-0

Laskov, P., Patrick, D., Sch, C., & Rieck, K. (2005). Learning intrusion detection : supervised or unsupervised ? *Distribution*, 50–57.

Leung, K., & Leckie, C. (2005). Unsupervised anomaly detection in network intrusion detection using clusters. *Conferences in Research and Practice in Information Technology Series*, *38*(January), 333–342. http://doi.org/10.1080/15374416.2013.769171

Li, W., Yi, P., Wu, Y., Pan, L., & Li, J. (2014). A new intrusion detection system based on KNN classification algorithm in wireless sensor network. *Journal of Electrical and Computer Engineering*, *2014*(January 2016). http://doi.org/10.1155/2014/240217

Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. *Computers and Security*, *21*(5), 439–448. http://doi.org/10.1016/S0167-4048(02)00514-X

Lin, W. C., Ke, S. W., & Tsai, C. F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-Based Systems*, *78*(1), 13–21. http://doi.org/10.1016/j.knosys.2015.01.009

Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., & Das, K. (2003). Recent Advances in Intrusion Detection, *2820*(October 2000). http://doi.org/10.1007/b13476

Lu, W. (2014). A detailed analysis of the KDD CUP 99 data set NRC Publications Archive (NPArC), (July 2009), 1–6. http://doi.org/10.1109/CISDA.2009.5356528

Magnano, C., & Lekas, C. (n.d.). CPSC097 Project Proposal : Network Intrusion Detection Using Random Forests And Expectation Maximization Preprocessing.

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011). Intrusion detection based on K-means clustering and OneR classification. *Proceedings of the 2011 7th International Conference on Information Assurance and Security, IAS 2011*, 192–197. http://doi.org/10.1109/ISIAS.2011.6122818

Nalavade, K., & Mehsram, B. B. (2014). Evaluation of K-Means Clustering for Effective Intrusion Detection and Prevention in Massive Network Traffic Data. *International Journal of Computer Applications*, *96*(7), 9–14.

Naphade, M. R. A., Raut, M. P. D., Year, B. E. F., & Chikhli, A. E. C. (2016). A Review of Intrusion Detection System Basic Concepts, *5*(3), 482–485.

Nath, S. V. (2007). Crime pattern detection using data mining. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops Proceedings)*, *1*(954), 41–44. http://doi.org/10.1109/WI-IATW.2006.55

Patil, P. R. (2016). Performance Analysis of Intrusion Detection Systems Implemented using Hybrid Machine Learning Techniques, *133*(8), 35–38. http://doi.org/10.5120/ijca2016907997

Rajan, S. S., & Cherukuri, V. K. (2010). An overview of intrusion detection systems. *Retrieved May*, *12*(3), 559–563. http://doi.org/10.1109/surv.2010.032210.00054

Rajeshkumar, G., Mangathayaru, N., & Narsimha, G. (2016). Intrusion Detection – A Text Mining Based Approach, *14*(February), 76–88.

Reguianski, T. L. (1962). The Air Force Institute of Technology. *IRE Transactions on Education*, *E-5*(2), 117–118. http://doi.org/10.1109/TE.1962.4322266

Rutman, R. S. (1994). On the paper by R. R. Nigmatullin "fractional integral and its physical interpretation." *Theoretical and Mathematical Physics*, *100*(3), 1154–1156. http://doi.org/10.1007/BF01018580

Sabhnani, M. (n.d.). Meysam - 2013 - Do Debt Markets Price Ṣukūk and Conventional Bonds Differently.pdf.

Sahasrabuddhe, A., Naikade, S., Ramaswamy, A., Sadliwala, B., & Futane, P. P. (2017). Survey on Intrusion Detection System using Data Mining Techniques, 1780–1784. http://doi.org/10.1111/cea.13091

Sammany, M., Sharawi, M., El-beltagy, M., & Saroit, I. (2007). For Intrusion Detection Systems and Classification of Attacks, (January).

Sharma, S., & Gupta, R. K. (2015). Intrusion detection system: A review. *International Journal of Security and Its Applications*, *9*(5), 69–76. http://doi.org/10.14257/ijsia.2015.9.5.07

Shenfield, A., Day, D., & Ayesh, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *ICT Express*, *4*(2), 95–99. http://doi.org/10.1016/j.icte.2018.04.003

Siddiqui, M. K., & Naahid, S. (2013). Analysis of KDD CUP 99 Dataset using Clustering based Data Mining. *International Journal of Database Theory and Application*, *6*(5), 23–34. http://doi.org/10.14257/ijdta.2013.6.5.03

Singh, A., Banafar, H., & Pippal, R. S. (2015). Intrusion Detection on KDD99cup Dataset using K-means , PSO and GA : A Review.

Solanki, M., & Dhamdhere, V. (2014). Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier. *International Journal of Emerging Trends & Technology in Computer Science*, *3*(6). Retrieved from http://www.ijettcs.org/Volume3Issue6/IJETTCS-2014-11-02-10.pdf

SylviaaS, M., & Professor, A. (2015). Intrusion Detection System-a Study. *International Journal of Security, Privacy and Trust Management (IJSPTM)*, *4*(1), 31–44. http://doi.org/10.5121/ijsptm.2015.4104

Ted Holland. (2004). Interested in learning SANS Institute InfoSec Reading Room tu , A ho ll r igh ts. *Information Security*, 18. http://doi.org/10.9780/22307850

Thomas, C., Sharma, V., & Balakrishnan, N. (2008). Usefulness of DARPA dataset for intrusion detection system evaluation, (September 2014), 69730G. http://doi.org/10.1117/12.777341

Urvashi, M., & Jain, M. A. (2015). A survey of IDS classification using KDD CUP 99 dataset in WEKA. *International Journal of Scientific & Engineering Research*, *6*(11), 947–954. Retrieved from http://www.ijser.org

Wang, G., Hao, J., Mab, J., & Huang, L. (2010). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. *Expert Systems with Applications*, *37*(9), 6225–6232. http://doi.org/10.1016/j.eswa.2010.02.102

Wikipedia. (2012). Intrusion detection system. *Wikipedia*, (Wikipedia), 1–5. http://doi.org/10.1016/S0009-2509(02)00111-2

Zakaria, W. Z. A. (2015). Application of Case Based Reasoning in IT Security Incident Response, (September).

Zanero, S., & Savaresi, S. M. (2004). Unsupervised learning techniques for an intrusion detection system. *Proceedings of the 2004 ACM Symposium on Applied Computing - SAC '04*, (December), 412. http://doi.org/10.1145/967900.967988

Zanero, S., & Serazzi, G. (2008). C3-239.pdf, 1043–1048. http://doi.org/10.1109/NOMS.2008.4575276

Academy, C. (2017). Feature Selection in the Corrected KDD-dataset Shahrzad Zargari, (May).

Zhao, Z., & Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. *Proceedings of the 24th International Conference on Machine Learning - ICML '07*, 1151–1157. http://doi.org/10.1145/1273496.1273641