Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 46 – 51

Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)

Solane Duque[a]*, Dr.Mohd. Nizam bin Omar[b]

[a]*College of Arts and Sciences,University Utara Malaysia*
[b]*College of Arts and Sciences,University Utara Malaysia*

Abstract

A common problem shared by current IDS is the high false positives and low detection rate. An unsupervised machine learning using k-means was used to propose a model for Intrusion Detection System (IDS) with higher efficiency rate and low false positives and false negatives. The NSL-KD data set was used which consisted of 25,192 entries with 22 different types of data. Results of the study using 11, 22, 44, 66 and 88 clusters, showed an efficiency rate of 70.75%, 81.61%, 65.40%, 61.30% and 55.43% respectively; false positive rates of 0.74%, 4.03%, 15.55%, 21.47% and 31.91% respectively; and false negative rates of 99.82%, 98.14%, 97.76%, 96.32% and 95.70%, respectively. Interestingly, the best results were generated when the number of clusters matches the number of data types in the data set. In the light of the findings, it is recommended that other data mining techniques be explored; a study using k-means data mining algorithm followed by signature-based approach is proposed in order to lessen the false negative rate; and a system for automatically identifying the number of clusters may be developed.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: data mining; clustering; machine learning; unsupervised learning; k-means

1. Introduction

The latest developments in computer systems and the internet have revolutionized the way people think and do things. A process like sending traditional mail that normally takes hours or even days can now be completed in a click of a mouse or a touch of a finger through electronic mail or e-mail. People communicate with each other from

* Corresponding author. Tel.: +968-9609-7575;
E-mail address: solane@unizwa.edu.om

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2015.09.145

different places through integrated relay chat, or video conferencing as a much convenient mode of communication.

However, along with the many advances in computer systems and IT infrastructures are the risks associated with the use of these technologies. Over the last two decades, computer threats and cybercrimes have proliferated at the disadvantage of the general public, and newer threats are introduced each day that compromise the integrity, validity and confidentiality of data. Companies, nations, and individual persons can be victims of malicious activities in the internet. As a consequence of cybercrimes, millions of dollars have been spent on mitigation strategies.

People who exploit the vulnerabilities of the information systems are usually adept at using sophisticated programming techniques and take advantage of the interconnectivity of the systems so much so that they do not even need local access to the network because they can launch the attacks remotely.

Malicious activities in the internet are also known as intrusion. An intrusion is defined as any activity that violates security policy of the network [1]. Intrusion detection system (IDS) is software and hardware deployed to carry out the process of detecting unauthorized use of, or attack upon, a computer or a telecommunications network – which is supposed to bridge the gaps in firewall and anti-viruses. An IDS provides monitoring and analysis of user and system activity, can audit system configuration and vulnerabilities, assess the integrity of critical system and data files, provide statistical analysis of activity patterns based on the matching with known attacks, analyze abnormal activity, and operate system audit [2]. One advantage of the IDS is its ability to document the intrusion or threat to an organization, thereby providing bases for informing the public regarding the latest attack patterns through system logs.

The types of computer attacks detected by IDS are categorized into three, namely: (i) scanning attacks, (ii) denial of service (DOS) attacks, and (iii) penetration attacks [3]. Each of these three categories of computer attacks has distinct signatures and behaviours - to which IDS is designed to analyze, detect and triggers an alarm when encountered. Once an alarm is set, network administrators will have to analyze the logs to decide whether the suspected activity is indeed anomalous.

In most IDS however, there is a high instances of false positives and false negatives which can be cumbersome to deal with for the network administrators. A false positive is an instance where an IDS incorrectly identifies a benign activity to be malicious while a false negative occurs when the IDS fails to detect a malicious activity [4]. During normal operation, an IDS can generate thousands of false alarms per day [5]. Network intrusion detection systems - no matter if they are anomaly-based or signature-based - share a common problem: the high number of false alerts or false positives. The number of alerts collected by an IDS can be up to 15,000 per day per sensor, and the number of false positives (FP) can be thousands per day. These problems usually cause the final user, the security manager to lose confidence in the alerts, lower the defence levels in order to reduce the number of false positives, or to have an overload of work to recognize true attacks due to IDS mistakes [6].

This paper proposes using machine learning and the k-means data mining algorithm to develop an IDS model with higher efficiency rate and lower false alarms.

1.1 Problem Statement

The study proposes machine learning and the k-means data mining algorithm to develop an IDS model with higher efficiency and lower false using the NSL-KDD data set.

1.2 Research Questions

Consequently, it will answer the following research questions:

1.2.1 To what extent can the k-means detect (i.e. detection rate) attack and normal data?

1.2.2 What are the factors affecting the implementation of an IDS model using k-means data mining algorithm?

1.3 Research Objectives

1.3.1 To be able to detect normal vis-a-vis attack data within the data set.

1.3.2 To be able to identify the false positive rate generated using the k-means algorithm.

1.3.3 To be able to identify the false negative rate generated using the k-means algorithm.
1.3.4 To be able to identify the efficiency rate generated using the k-means algorithm.

2. Review of Related Literature

There has been a lot of research papers conducted using the KDD CUP '99 Dataset for developing models for instruction detection system. Although, there is much debate as to whether the dataset is in fact a good or valid record to be used as basis for proposing models for intrusion detection system, the fact that there is no other substitute dataset available for such purpose, makes it still the widely-used and accepted dataset for experimentation. The KDD CUP '99 or KDD'99 is 10% of the original DARPA 98 dataset which was used in the MIT Lincoln Laboratory. The DARPA'98 has around 5 million records of activity from different users and connections, and the large volume of data makes it difficult for processing by ordinary machines – hence the KDD'99. As more researches are conducted, a new dataset is proposed, the NSL-KDD.

Data mining is defined as searching for knowledge (interesting patterns) in data [7]. Many people associated data mining to Knowledge Discovery in Database (KDD), but data mining can also be viewed as a single step towards knowledge discovery. In [31], a new version of the KDD data set is proposed known as the NSL-KDD due to supposed inherent problems of the KDD CUP'99. In [25], The NSL-KDD is a reduced version of the KDD'99 dataset. The NSL-KDD has the same features as the KDD'99 but it does not include the redundant records of the KDD'99, and there are also no duplicate records which make it unbiased to frequent and redundant entries.

3. Research Methodology

3.1 NSL-KDD Data Set

The NSL-KDD data set has 25,192 entries and 43 attributes – where the 41 attributes are the same as the KDD'99; the 42nd attribute is the data label, and the 43rd attribute is the level of difficulty. There are 22 different types of data: (1) normal, (2) back, (3) buffer_overflow, (4) guess_passwd, (5) imap, (6) ipsweep, (7) multihop, (8) neptune, (9) nmap, (10) phf, (11) pod, (12) portsweep, (13) rootkit, (14) satan, (15) smurf, (16) teardrop, (17) warezclient, (18) warezmaster, (19) ftp_write, (20) load_module, (21) land and (22) spy.

3.2 Pre-processing

Pre-processing involves cleaning the data of inconsistencies and/or noise, and combining or removing redundant entries. Pre-processing also involves converting the attributes of the dataset into numeric data and saving in a format readable because k-means works only on numerical data. Alphanumeric data were converted to numeric values starting from 0.001, 0.002, and so on. Smaller values (instead of 1, 2, etc.) were used to make sure that it will not affect the computations.

3.3 K-means Clustering

K-means is a centroid-based technique, and is the simplest and most fundamental clustering by partitioning is the k-means, wherein the objects are organized into *k* partitions ($k \leq n$). The k-means is particularly used to identify outliers because when there is a value that is far away from the majority of the data, the mean value of the cluster will be significantly distorted. This study will use k-means clustering as a method of outlier detection. In this outlier detection model, it is assumed that normal behaviour pattern are far more frequent than the outliers or anomalous behaviours.

3.3.1 The K-Means Clustering Formula [8]

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, C_i)^2 \quad (1)$$

Where:
E – is the sum of the squared error for all objects in the data set

p – is the point in space representing a given object

3.3.2 The K-Means Clustering Algorithm

The k-means clustering algorithm for partitioning, where each cluster's centre is represented by the mean value of the objects in the cluster:

Input:
k – the number of clusters,
D- a dataset containing *n* objects
Output: A set of *k* clusters.

Method:

(1) Arbitrarily choose *k* objects from *D* as the initial cluster centres;
(2) repeat
(3) (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster means that is, calculate the mean value of the objects for each cluster;
(5) Until no change.

3.4 Performance Measures

The following formula will be used to measure the performance using 4 different clusters (22, 44, 66, and 88).

$$DR_{Normal} = \frac{Number\ of\ True\ Normal\ Data\ Detected}{Number\ of\ Normal\ Data\ Detected}\ x100\% \quad (2)$$

$$DR_{Attack} = \frac{Number\ of\ True\ Attack\ Data\ Detected}{Number\ of\ Attack\ Data\ Detected}\ x100\% \quad (3)$$

$$FPR = \frac{Number\ of\ False\ Positives}{Number\ of\ Normal\ Data} x100\% \quad (4)$$

$$FNR = \frac{Number\ of\ False\ Negatives}{Number\ of\ Attack\ Data}\ x100\% \quad (5)$$

$$Efficiency\ Rate = (DR_{Normal}) + (DR_{Attacks}) \quad (6)$$

where:
DR – Detection Rate
FPR – False Positive Rate (i.e. normal data classified as attacks)
FNR – False Negative Rate (i.e. attacks classified as normal)

4. Results and Findings

The results showed an efficiency of 81.61%; 65.40%; 61.30%; and 55.43% depending on the number of clusters used (11, 22, 44, 66, or 88). Further, it can be noted that as the number of cluster increases above the number of data types, the detection rate, false negative rate, and efficiency rate, decreases; but the false positive rate increases.

It is interesting to know that the best results were generated when 22 clusters were used – corresponding to the number of data types. This shows that the performance of the k-means is dependent on the number of clusters, and therefore number of clusters should be determined beforehand.

It is also notable that the false positive rate is significantly lower than the false negative rate for clusters all clusters. Although the issue of the IDS not being able to detect malicious data is still a problem and is something

that still needs further investigation, lessening the false alarms (ie.e false positives) generated is at a low 4.03% for 22 clusters, and 0.74% for 11 clusters.

Table 1: k-means clustering results

	Number of Clusters				
	11	22	44	66	88
True Normal Data Detected	13350	12907	11358	10562	9157
Total Normal Data Detected	25072	24431	22720	21873	20395
True Attacks Detected	21	219	381	432	505
Total Attacks Detected	120	761	2472	3319	4797
False Positives	99	542	2091	2887	4292
False Negatives	11722	11524	11362	11311	11238
Normal Data Detection Rate	53.25%	52.83%	49.99%	48.29%	44.90%
Attack Data Detection Rate	17.50%	28.78%	15.41%	13.02%	10.53%
Efficiency Rate	70.75%	81.61%	65.40%	61.30%	55.43%
False Positive Rate	0.74%	4.03%	15.55%	21.47%	31.91%
False Negative Rate	99.82%	98.14%	96.76%	96.32%	95.70%

5. Conclusion and Recommendation

Results of k-means clustering showed that a higher efficiency rate is achieved when the correct number of clusters is applied, and increasing or decreasing the cluster beyond the number of data types only lessens the efficiency of the model.

Identifying the number of clusters therefore significantly changes the results. One has to know at the onset how many clusters are expected in order to get good results. In this model 22 clusters were used based on the different types of data. However, in a dynamic network, the challenge of identifying the number of clusters will be difficult since there is no “ground data” to serve as basis for deciding the number of clusters.

In the light of the findings, the following are the recommendations:

1. Other data mining techniques (like Bayesian, hierarchical, etc) may explored to compare results.
2. A study using k-means data mining algorithm followed by signature-based approach is proposed in order to lessen the false negative rate.
3. A system for automatically identifying the number of clusters may be developed.

References

1. Bischof, H., Leonardis, A., and Selb, A. MDL principle for robust vector quantisation. Pattern Analysis and applications. 2:59-72,1999.
2. SANS Institute. Understanding Intrusion Detection System. 2001.
3. Bace, R., and Mell, P. Intrusion Detection System, NIST Special Publications SP800. November. 2001.
4. Scarfone, K. and Mell, P. Guide to Intrusion Detection and Prevention System. National Institute of Standards and Technology. Special Publication 800-94. February 2007.
5. J. Ioannidis, A. Keromytis, and M.Yung (Eds.): “IDS False Alarm Reduction using Continuous and Discontinuous Patterns”, ACNS 2005, LNCS 3531, pp. 192–205, 2005. c Springer-Verlag Berlin Heidelberg 2005.
6. Owen, D., “What is a False Positive and Why are False Positives a problem?”, available online at http://www.sans.org/security-resources/idfaq/false_positive.php , last accessed in June 2015.
7. Han, K., Kamber, M., Pei, J. Data Mining Concepts and Techniques. Third Edition. Morgan Kaufmann, Elsevier Inc. 2012. ISBN 978-0-12-381479-1.

8. Jiawei, H. and Micheline, K. Data Mining Concepts and techniques, second edition, China Machine Press, pp. 296-303. 2006.
9. Chapman, S.J. Matlab Programming for Engineers. International Student Edition. Fourth Edition. Thomson Learning, part of Thomson Corporation. ISBN-10:0-495-24451-1. ISBN-13:978-0-495-24451-6. 2008.
10. Gilat, A. Matlab An Introduction with Applications. Fourth Edition. SI Version. John Wiley and Sons, Inc. 2011.ISBN: 978-0-470-87373-1. Printed in Asia.2011.
11. Gilat, A. Matlab An Introduction with Applications. Third Edition. SI Version. John Wiley and Sons, Inc. 2011.ISBN: 978-0-470-10877-2. Printed in the United States of America. 2007.
12. Kayacik, H.G., Zincir-Heywood, A.N. and Heywood, M.L. Selecting Features for Intrusion Detection: A Feature Analysis of KDD 99 Intrusion Detection Datasets.2006.
13. KDD Cup 1999 Data available at: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html October 2007, last accessed in August 2014.
14. Lane, T., and Brodley, C.E. Sequence matching and learning in anomaly detection for computer security. In AAAI Workshop: AI Approaches to Fraud Detection and Risk Management pp. 43-49.AAAI Press. July 1997.
15. Lee,W., Stolfo, S.J. and Mok,K.W. Data mining approaches for intrusion detection. N Proceedings of the 7th USENIX Security Symposium, March 1999.
16. Lee, W., Stolfo, S.J. and Mok, K. Data Mining in work flow environments: Experiments in intrusion detection. In Proceedings of the 1999 Conference on Knowledge Discovery and Data Mining.1999.
17. Lee, W.and Stolfo, S.J. A Framework for Constructing Features and Models for Intrusion Detection Systems. 1999.
18. Lippmann, R.P., et.al. MIT Lincoln Laboratory Offline Component of DARPA 1998 Intrusion Detection Evaluation. MIT Lincoln Laboratory. PI Meeting. 1998.
19. Mannila, H., Toivonen, H., and Verkamo, A.I. Discovering frequent episodes in sequences. In Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining. Montreal, Canada. August 1995.
20. MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available at: http://www.II.mit.edu/mission/communications/ist/corpora/ideval/index.html last accessed on February 2014.
21. Mukkama, S., Janoski, G., Sung, A. Intrusion detection using neural networks and support vector machines. Proceedings of IEEE International Joint Conference on Neural Networks, pp. 1702-1707.2002.
22. Nguyen, H.A., and Choi, D. Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. APNOMS 2008, LNCS 5297, Springer-Verlag Berlin Heidelberg 2008. pp.399-408. 2008.
23. NSL-KDD Data Set for network-based intrusion detection systems available at: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.
24. Olusola, A.A., Oladele, A.S., and Abosede, D.O. Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features. Proceedings of the World Congress on Engineering and Computer Science Vol I WCECS 2010. October 20-22, 2010, San Francisco, USA. 2010.
25. Patel A.,Sammarvar,S., and Naik, A. DataMining Vs.Statistical Techniques for Classification of NSL-KDD Intrusion Data. International Journal of Computer Science and Information Technologies, Vol 5(4), 2014.ISSN:075-9646.
26. Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993).
27. Sabhani, M., and Serpen, G. Why Machine Learning Algorithms Fail in Misuse Detection on KDD Intrusion Detection Dataset. Intelligent Data Analysis, vol 6. (Jne 2004).
28. Stanford-Chen, S. Common intrusion detection framework. Available at: http://seclab.cs.ucdavis.edu/cidf.
29. Siddiqui, M.K., and Naahid, S.Analysis of KDD CUP 99 Dataset using Clustering based Data Mining. International Journal of Database Theory and Application Vol.6, No. 5. pp.23-24. 2013.
30. Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Application (CISDA 2009).
31. The DARPA Intrusion Detection Data Sets. Lincoln Laboratory Massachusetts Institute of Technology. Available at: www.II.mit.edu
32. Waikato Environment for Knowledge Analysis (WEKA) version3.5.7. Available at: http://www.cs.waikato.ac.nz/ml/weka/, June 2008.
33. Witten, I. H., Franck, E. Data Mining Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco. 2005.
34. Xu, X. Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction, International Journal of Web Services Practices 2(1-2), 49-58. 2006.