Identifying False Alarm Rates for Intrusion Detection System with Data Mining

Fatin Norsyafawati Mohd Sabri[1], Norita Md.Norwawi[2], and Kamaruzzaman Seman[3]
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai, Negeri Sembilan, Malaysia

Abstract
Intrusion Detection Systems (IDS) are very important in determining how secure a system is, and to discover several types of attack such as Denial of Service (DOS), Probes and User to Root (U2R) attacks. However, recently false alarm rates and accuracy of detection are happens to be the most important issues and challenges in designing effective IDS. Therefore, this study is aimed at detecting denial of services attack and normal traffic using Knowledge Discovery and Data Mining Cup 99(KDD CUP 99) dataset to reduce the false alarm rates. Data mining is used to extract the useful information from large databases. The results have shown that the data mining technique reduces the false alarm rates and increase the accuracy of the system.

Keywords
Intrusion detection system; accuracy; false alarm rate; data mining; denial of service

1. Introduction

During the past few years, the number of intrusions in computer networks has grown extensively, and many new hacking tools and intrusive methods have appeared. Using IDS is one way of dealing with suspicious activities within a network[1].The intrusion behaviours cause the great damage of systems. So enterprises search for intrusion detection systems to protect their systems. The traditional technology such as firewall is used to defense attacks. Thus, the IDS usually used to enhance the network security of enterprises. The major difference between firewall and IDS system is that firewall is a manual passive defense system.

Comparatively, IDS could collect packets online from the network. After collecting them, IDS will monitor and analyze these packets. So, IDS system acts as the "second line of defense". Finally, it will provide the detecting results for managers. The detecting results could be either attack or normal behavior. An ideal IDS system has a 100% attack detection rate along with a 0% false positive rate, but it is hard to achieve. Detecting illegal behaviours of the host or network is the major object of IDS. The IDS is actually such a system to detect some illegal behavior. One of the ability of IDS is it could monitor various activities on the network. IDS will send a warning message to the managers if it detects an attack.

Briefly, the aim of IDS is to detect intruders or attacks effectively[2]. There are two main methods of IDS, misuse or signature detection and anomaly detection. Signature or Misuse Detection is known intrusions are detected by looking at the computer system behavior some characteristic pattern of such intrusions. This approach uses some collected information about the system behavior under normal conditions and under some known intrusions to determine the current state of them system. In this case, the intrusion detection problem is a classification problem[3].

Misuse detection refers to techniques that characterize known methods to penetrate a system. These penetrations are characterized as a 'pattern' or a 'signature' that the IDS look for. The pattern or signature might be a static string or a set sequence of actions. System responses are based on identified penetrations. The idea of misuse detection is to establish a pattern or a signature form so that the same attack can be detected. The other idea here is to establish a normal activity profile for system. False alarm rate refers to the proportion that normal data is falsely detected as attack behavior, namely a false positive (FP) situation. Accuracy is defined as the proportion of data correctly classified that is a true positive (TP) and true negative (TN)[4].

The intrusion detection technology is the process of identifying network activity that can lead to a compromise of security policy[5]. It is a system for detecting intrusions and reporting them accurately to the proper authority[6]. Intrusion Detection Systems are usually specific to the operating system that they operate in and are an important tool in the overall implementation an organization's information security policy.

An intrusion detection system can be compared with a house burglar alarm: if somebody tries to enter illegally in the house, one of the sensors will detect it and will trigger the alarm bell and alert the house owner and the police. Similarly, if somebody tries to compromise the confidentiality, the integrity or the availability of a

Manuscript received April 5, 2011
Manuscript revised April 20, 2011

computer system or network, or tries to break the security protections, an intrusion detection system will alert the system owner and the security team[7].

Data mining is used most by statisticians, database researchers and more recently by the MIS and business communities. Here, we used the term "KDD" to refer to the overall process of discovering useful knowledge from data [8].It is defined as the process of extracting useful and previously unnoticed models or patterns from large data stores[9].There are two techniques used to extract the data known as classification and clustering [10].

A classification based IDS attempts to classify all traffic as either normal or malicious in some manner. The primary difficulty in this approach is how accurately the system can learn what these patterns are. This ultimately affects the accuracy of the system both in terms of whether non-hostile activity is flagged; false positive and whether malicious activity will be missed; false negative. For classification, there are lot of algorithms can be used for this purpose such as decision tree, fuzzy logic, genetic algorithms, support vector machine, neural network and Bayesian. Meanwhile for clustering, K-Nearest Neighbour (KNN), K-mean or other techniques can be used for it [10].

The advantages of data mining are in the decision support and application development. Decision support is the mining process exchanging information to facts, rules and graphic presentation where all the data have entity relationship among those elements.

Denial of Service (DOS) attack is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine. For example: ping of death and SYN flood[11]. The aim of DOS is to disrupt, destroy, or render services unavailable. A typical DOS exhausts the target's resources. The server is rendered unavailable for honest clients, who then proceed to request similar services from competitors. To prevent malicious requests, a server needs to filter out bogus connection requests and honour those from legitimate clients [12].Therefore, data mining is used in this study to increase the accuracy of the system and reduce the false alarm rates.

2. Method

We used data mining software tools known as iDA analyzer. The software is a GUI based software and easy to use. IDA analyzer software is capable of classifying large amount of data within seconds depending on the speed and condition of computer processors. It provides data mining rule-based algorithm. Figure 1 shows three phases in data mining process. In phase 1, data preparation involves collection and assessment towards the data. After collecting and confirm the format of the data, data will be selected based on the requirement needed for this study. This is due to some of the mining functions that only accept data in a certain format. Fortunately, KDD CUP 99 dataset is already in binary format. KDD CUP 99 is a dataset of computer network intrusion detection downloaded from http://kdd.ics.uci.edu website.

Meanwhile in phase 2, a knowledge model is obtained representing behavior patterns in relevant problems with variables for relation between them. In the next step, we analyzed the results and the rules generated by iDA analyzer software.

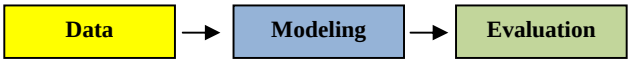

Fig. 1: Data Mining Process

The steps involve in the experiment process include identify which data of KDD CUP 99 dataset will be chosen, then selecting appropriate category of attacks such as DOS attacks data and normal data, identify the availability of software classification to process large amount of data and finally the data is divided into five partitions and start the mining sessions.

We conducted the experiment onto 7000 of KDD CUP 99 dataset. The amount of data was used due to the limitation software used to perform this classification. iDA classification software is selected to classify all the data into six classes of DOS attacks and one class of normal data as shown in Table 1.

Table 1: Data Classification

Class of DOS Attacks	Total
Back	1434
Land	21
Neptune	1434
Normal	1434
Pod	264
Smurf	1434
Teardrop	979

3. Results and Discussions

3.1 Data Partitioning

The volume of data is divided into two parts, testing data and training data. Training data is a data instances used to create supervised learning models while testing data is a data instances used to test models built with supervised learning. For this study we already created five groups of data partition which are 6300 training data and 700 testing data, 5600 training data and 1400 testing data, 4900 training data and 2100 testing data, 4200 training data and 2800 testing data and lastly is 3500 for both training data and testing data. The partition that produces the best results is the third partition as tabulated in Table 2.

Table 2: Data partitioning and percentage of correctness

Partition	Training	Testing	Accuracy%
1	6300	700	99.86
2	5600	1400	99.93
3	4900	2100	99.90
4	4200	2800	95.96
5	3500	3500	95.67

3.2 Classification

In order to determine the number of false positive and false negative, confusion matrix analysis has been used. The result is presented in Table 3.

Table 3: Number of False Positive and False Negative Using Rule-based

Class of Attack	Number of records	False Positive	False Negative
Land.	0	0	0
Teardrop.	110	0	0
Back.	476	0	0
Neptune.	476	0	0
Normal.	476	0	2
Smurf.	476	0	0
Pod.	86	0	0
	4900	0	2

Table 4: Confusion Matrix Classification

Computed Result \ Actual Result	Land.	Teardrop	Back.	Neptune	Normal.	Smurf.	Pod.	Total	PSP
Land.	0	0	0	0	0	0	0	0	0%
Teardrop	0	110	0	0	0	0	0	110	100%
Back.	0	0	475	0	1	0	0	476	99.79%
Neptune	0	0	0	476	0	0	0	476	100.0%
Normal.	0	0	0	0	476	0	0	476	100.0%
Smurf.	0	0	0	0	0	476	0	476	100.0%
Pod.	0	0	0	0	0	0	85	85	100.0%
Total	0	110	476	476	476	476	86	2100	
Psp	0%	100%	99.79%	100.0%	100.0%	100.0%	99.79%		

Meanwhile, accuracy of classification can be calculated based on percentage of successful prediction (PSP).Accuracy of Intrusion Detection System is very important in order to ensure the ability of the system to detect intruders. Table 4 shows the results for this calculation.

$$PSP = \frac{\text{Number of incidents that have been successfully classified}}{\text{Total of incidents}} \times 100\%$$

Percentage of detection rate (DR) also can be calculated based on confusion matrix table by using the following formula:-

$$DR = TP/(TP+TN)\ X100\%$$

TP = amount of attack when it actually attack
TN = amount of normal detect when it actually normal

TP=1622
TN=476

DR = (1622/2098) X100%
= 77.31%

3.3 Rules Extraction Generates by iDA Data Mining Software

iDA data mining software takes about less than one minute to classify the data based on their pattern and characteristics. Figure 4 shows the example of rules generated by iDA software. Meanwhile Figure 5 shows the example of rules that have been paraphrased into *IF THEN RULES*.

```
1480.00 <= src_bytes <= 1480.00
  :rule accuracy 100.00%
  :rule coverage 98.31%

1.00 <= wrong_fragment <= 1.00
  :rule accuracy 95.63%
  :rule coverage 98.31%

1.00 <= dst_host_count <= 5.00
  :rule accuracy 53.39%
  :rule coverage 66.29%
  :rule accuracy 100.00%
  :rule coverage 98.31%

  **Total Percent Coverage = 100.00%
```

Fig. 4: Example of Rules Generated by iDa Software

IF **src_bytes** IS EQUAL TO **1480**

THEN

IF **wrong_fragment** IS EQUAL TO **1** THEN

IF **dst_host_count** GREATER THAN **1** OR LESS

THAN **5** THEN **LAND**

Figure 5: If -Then Rules

4. Conclusion

In conclusion, this technique has reduced the numbers of false alarm rates and increases the accuracy of the systems in the 7000 samples of randomly selected KDD CUP dataset. In future, we hope that we can get better results of accuracy and increase the detection rate for the IDS system.

Acknowledgements
The authors wish to thank to her university, Universiti Sains Islam Malaysia, Faculty of Science and Technology Universiti Sains Islam Malaysia and colleagues for their willingness to review the paper and contributed to its technical content.

REFERENCES

[1] Dong S, Malcom I& Zinchir-Heywood A.N. 2005."Training Genetic Programming on Half a Million Patterns: An Example from Anomaly Detection", IEEE Transactions on Evolutionary Computation.Vol.9 No.3, pp.225-239,.
[2] Vaibhav G,Csill F & Valtorta M, “Paid: A probabilistic agent-based intrusion detection system”. Journal of Computers and Security. Vol.24.No.7. October, pp.529-545. 2005.
[3] Jonatan Gomez & Dipankar Dasgupta,Evolving Fuzzy Classifier for Intrusion Detection,Proceedings of the 2002 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY, 2001.
[4] Osareh Alireza & Shadgar Bita, “Intrusion Detection in Computer Networks based on Machine Learning Algorithms”. International Journal of Computer Science and Network Security, Vol.8.No.11. November, pp.15-23, .2008.
[5] Benattou M & K. Tamine, “Intelligent Agents for Distributed Intrusion Detection System”. World Academy of Science, Engineering and Technology 6.June,pp.190,200.5
[6] Jones, Anita. K. & Robert. S. Sielken, “Computer System Intrusion Detection: A Survey”, Technical Report. Department of Computer Science, University of Virginia, Charlottesville, Virginia, 2000.
[7] Bace R. & Mell P, “Intrusion Detection Systems”. NIST Special Publication,pp. 800-31,2001.
[8] Usama Fayyad, Gregory Piatetsky-Shapiro, dan Padhraic Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine American Association for Artificial Intelligence.

[9] Bass T, "Intrusion detection systems and multisensor data fusion", Communications of the ACM 43 (4),April,pp. 99–105,2000.
[10] Terry Brugger S. 2004. "Data Mining Methods for Network Intrusion Detection". University California, Irvin.
[11] Shilpa L, Sini J & Bhupendra V, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD",International Journal of Engineering Science and Technology. Vol. 2.No.6.June.pp.1790-1799,2010.
[12] Douglas S & Berkant U, "Towards Denial- of -Service – Resilient Key Agreement Protocols", Proceeding 14th Australasian Conf. on Information Security and Privacy (ACISP), LNCS,Tokyo, Japan, pp.389 – 406.2009.
[13] http://www.aaai.org/AITopics/assets/PDF/AIMag1703-2-article.pdf

Fatin Norsyafawati Mohd Sabri received the B.S degrees in Computer Science from Universiti Sains Islam Malaysia in 2009 respectively. Currently, she is doing her master in computer science focusing on Intrusion Detection System field.

Norita Md Norwawi is an Associate Professor at Universiti Sains Islam Malaysia. She obtained her Bachelor in Computer Science in 1987 from the University of New South Wales, Australia. She received her Masters degree in Computer Science from National University of Malaysia in 1994. In 2004, she obtained her PhD specializing in Temporal Data Mining and Multiagent System from University Utara Malaysia. As an academician, her research interests include artificial intelligence, multi-agent system, temporal data mining, text mining, knowledge mining, information security and digital Islamic application and content. Her works have been published in international conferences, journals and won awards on research and innovation competition in national and international level.

Kamaruzzaman Seman is a professor at Universiti Sains Islam Malaysia. He is also one of the senior members IEEE USA since 1998.He obtained his Bachelor from Univertiti Teknologi Malaysia, Malaysia in 1985. Then in 1986, he got his M.Sc in Telematics from Essex University. In1994, he obtained his PhD in Broadband ISDN from Strathclyde University UK. He experienced as being a chairman of Sub-Working Group 3 (Information Superhighway) of MNIC (now known as MNIIF),chairman of Next Generation Network (NGN), Telekom R&D S/B Task Force. (June 2003 – July 2005) and many more.