WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

Anomaly Detection Using Supervised learning Techniques in Social

Prof .Dr. Chezalina Binti Zulkifli*

Faculty of Computing and Creative Industries, UniversitiPendidikan Sultan IDris, Malaysia

*Corresponding Author: Prof .Dr.Chezalina Binti Zulkifli

DOI: https://doi.org/10.31185/wjcm.58

Received: June 2022; Accepted: August 2022; Available online: September 2022

ABSTRACT: Intrusion detection corresponds to a suite of techniques that are used to identify attacks against computers and network infrastructures. As the cost of the information processing and Internet accessibility falls, more and more organizations are becoming vulnerable to a wide variety of cyber threats. Web mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labelled as 'normal' or 'intrusive' and a learning algorithm is trained over the labelled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labelled appropriately. Evaluation results show that the proposed approach can reduce the number of alerts by 94.32%, effectively improving alert management process. Because of the use of ensemble approach and optimal algorithms in the proposed approach, it can inform network security specialist the state of the monitored network in an online manner.

Keywords: Web Data Mining Techniques, Anomaly Detection, Fraud DetectionMethod

1. INTRODUCTION

Intrusion detection corresponds to a suite of techniques that are used to identify attacks against computers and network infrastructures. As the cost of the information processing and Internet accessibility falls, more and more organizations are becoming vulnerable to a wide variety of cyber threats. According to a recent survey by CERT/CC [1], the rate of cyber-attacks has been more than doubling every year in recent times. Therefore, it has become increasingly important to make our information systems, especially those used for critical functions in the military and commercial sectors, resistant to and tolerant of such attacks. The most widely deployed methods for detecting cyber terrorist attacks and protecting against cyber terrorism employ signature-based detection techniques. Such methods can only detect previously known attacks that have a corresponding signature, since the signature database has to be manually revised for each new type of attack that is discovered. These limitations have led to an increasing interest in intrusion detection techniques based on data mining [2–6].

Web mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labelled as 'normal' or 'intrusive' and a learning algorithm is trained over the labelled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labelled appropriately. Research in misuse detection has focused mainly on classification of network intrusions using various standard data mining algorithms [2, 4–6], rare class predictive models, association rules [7, 8] and cost sensitive modelling. Unlike signature based intrusion detection systems, models of misuse are created automatically, and can be more sophisticated and precise than manually created signatures. A key advantage of misuse detection techniques is their high degree of accuracy in detecting known

attacks and their variations. Their obvious drawback is the inability to detect attacks whose instances have not yet been observed [9].

The term fraud here refers to the abuse of a profit organisation's system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.

2. LEARNING FROM RARE CLASSES

In misuse detection related problems, standard data mining techniques are not applicable due to several specific details that include dealing with skewed class distribution, learning from data streams and labelling network connections. The problem of skewed class distribution in the network intrusion detection is very apparent since intrusion as a class of interest is much smaller i.e. rarer than the class representing normal network behaviour [10, 11]. In such scenarios when the normal behaviour may typically represent 98-99% of the entire population a trivial classifier that labels everything with the majority class can achieve 98-99% accuracy. It is apparent that in this case classification accuracy is not sufficient as a standard performance measure.

3. RELATED WORKS OF ANOMALY DETECTION TECHNIQUES

Most research in supervised anomaly detection can be considered as performing generative modelling. These approaches attempt to build some kind of a model over the normal data and then check to see how well new data fits into that model. An approach for modelling normal sequences using look ahead pairs and contiguous sequences is presented in [12]. A statistical method for ranking each sequence by comparing how often the sequence is known to occur in normal traces with how often it is expected to occur in intrusions is presented in [13]. One approach uses a prediction model obtained by training decision trees over normal data [14], while others use neural networks to obtain the model [15] or non-stationary models [16] to detect novel attacks. Lane and Brodley [17] performed anomaly detection on unlabelled data by looking at user profiles and comparing the activity during an intrusion to the activity during normal use. Similar approach of creating user profiles using semi-incremental techniques was also used in [18]. Barbara used pseudo-Bayes estimators to enhance detection of novel attacks while reducing the false alarm rate as much as possible [19]. A technique developed at SRI in the EMERALD system [20] uses historical records as its normal training data. It then compares distributions of new data to the distributions obtained from those historical records and differences between the distributions indicate an intrusion. Recent works such as [21] and [22] estimate parameters of a probabilistic model over the normal data and compute how well new data fits into the model.

The proposed approach is able to process alerts produced by heterogeneous IDS systems. The approach is evaluated using DARPA 1999 dataset and Standard data set for a Port Complex dataset. Evaluation results show that the proposed approach can reduce the number of alerts by 94.32%, effectively improving alert management process. Because of the use of ensemble approach and optimal algorithms in the proposed approach, it can inform network security specialist the state of the monitored network in an online manner.

4. PROPOSED WORK OF ANOMALY DETECTION TECHNIQUES

This research has been applied the proposed detection schemes to 1999 DARPA Intrusion Detection Evaluation Data [23] as well as to the real network data from the Web. The DARPA'99 data contain two types: training data and test data. The training data consists of 7 weeks of network- based attacks inserted in the normal background data.

Attacks in training data are labelled. The test data contained 2 weeks of network-based attacks and normal background data. 7 weeks of data resulted in about 5 million connection records. Although DARPA'99 evaluation represents a significant advance in the field of intrusion detection, there are many unresolved issues associated with its design and execution. In his critique of DARPA evaluation, starting from usage of synthetic simulated data for the background (normal data) and using attacks implemented via scripts and programs collected from a variety of sources. In addition, it is known that the background data contains none of the background noise (packet storms, strange fragments . . .) that characterize real data. However, in the lack of better benchmarks, vast amount of the research is based on the experiments performed on this data. The evaluation of any intrusion detection algorithm on real network data is extremely difficult mainly due to the high cost of obtaining proper labelling of network connections. Figure 1 shows that Architecture proposed approach

for managing Alerts. The main reason for this procedure is to associate new constructed features with the connection records from "list files" and to create more informative data set for learning.

Since the amount of available data is huge (e.g. some days have several million connection records), this work has sampled sequences of normal connection records in order to create the normal data set that had the same distribution as the original data set of normal connections. This Research has used this normal data set for training our anomaly detection schemes, and then examined how well the attacks may be detected using the proposed schemes.

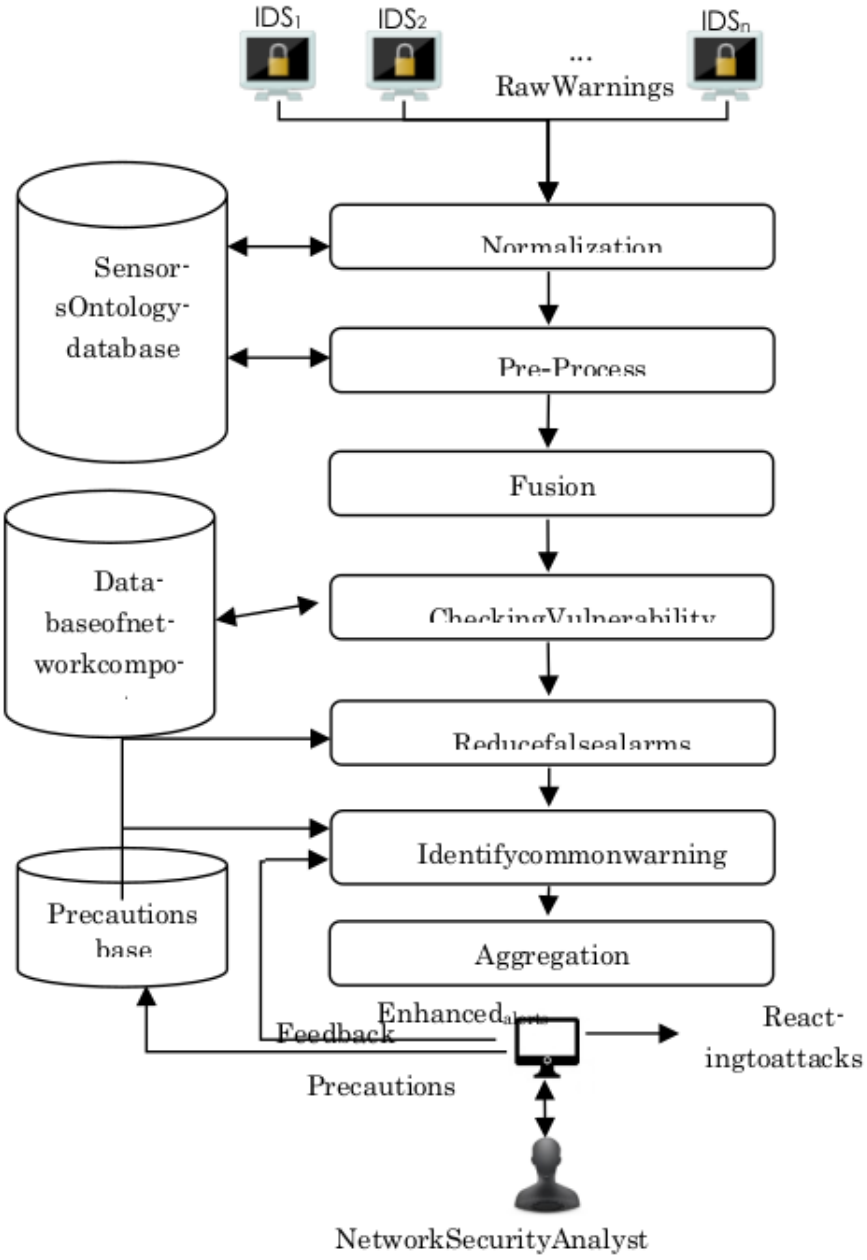

FIGURE 1. Architecture proposed approach for managing Alerts

```
Normalize(rawAlert)
{
alert = new Alert();
```

```
  alert.AID=AttackNamesDB.GetAttackName(rawAlert.Name,raw
Alert.Sensor.Type);
  foreach(attribute in Alert.Attributes) {
  alert.attribute.Value=NormalizationDB.GetStandardAt-
tributeValue(attribute, rawAlert);
  }
  send alert to next component;
  }
```

FIGURE 2. Sub-component code Normalization

Table 1. Evaluating the results of the proposed approach in data collection DARPA 99

The fifth week	The fourth week	The third week	The second week	The first week	
6090	260	82	7512	104	# alerts()
4385	219	82	464	104	# FP alerts()
12	10	9	12	18	# output alerts()
2	5	9	6	18	# output FP alerts()
99.80%	96.15%	89.02%	99.84%	82.69%	RR (%)
99.95%	97.72%	89.02%	98.71%	82.69%	FPRR (%)

Table 1 report on additional metrics for evaluating the results of the proposed approach in data collection DARPA 1999. Furthermore, shows that results of the proposed approach at the dataset DARPA 1999 on bursty attacks.

5. CONCLUSION & OPEN ISSUES

The Web and its Services are growing rapidly, so is the complexity and the number of cyber-attacks. Thus it is essential to use different security tools in order to protect computer systems and networks. Among these tools, Intrusion Detection Systems (IDSs) are one of the components of Defences-in-depth. One major drawback of IDSs is the generation of a huge number of alerts, most of which are false, redundant, or unimportant. Among different remedy approaches, many researchers proposed the use of data mining. Most of the research done in this area could not address the problems completely. Also, most of them suffer from human dependency and offline functionality.

To aid applicability involving anomaly diagnosis techniques, an activity intended for extracting practical statistical written content based along with temporal attributes is additionally applied. Experimental results performed on DARPA 98 data set indicate that the most successful anomaly detection techniques were able to achieve the detection rate of 74% for attacks involving multiple connections and detection rate of 56% for more complex single connection attacks, while keeping the false alarm rate at 2%. When the false alarm rate is increased to 4%, the achieved detection rate reaches 89% for bursty attacks and perfect 100% for single-connection attacks. Computed ROC curves indicate that the most promising technique for detecting intrusions in DARPA'99 data is the LOF approach. In addition, when performing experiments or real network data, the LOF approach was very successful in picking several very interesting novel attacks.

Considering the DARPA'99 data, performed experiments also demonstrate that for different types of attacks, different anomaly detection schemes were more successful than others. For example, the unsupervised SVMs were very promising in detecting new intrusions since they had very high detection rate but very high false alarm rate too. Therefore, future work is needed in order to keep high detection rate while lowering the false alarm rate. In this research, an online approach is proposed in order to manage alerts issued by IDSs. The proposed approach is able to process alerts produced by heterogeneous IDS systems. The approach is evaluated using DARPA 1999 dataset and Standard data set for a Port Complex dataset.

Our long-term goal is to develop an overall framework for defending against attacks and threats to computer systems. Although our developed techniques are promising in detecting various types of intrusions they are still preliminary in nature. Data generated from network traffic monitoring tends to have very high volume, dimensionality and heterogeneity, making the performance of serial data mining algorithms unacceptable for on-line analysis. Therefore, development of new anomaly detection algorithms that can take advantage of high performance computers is a key component of this project. According to our preliminary results on real network data, there is a significant non-overlap of our anomaly detection algorithms with the SNORT intrusion detection system, which implies that they could be combined in order to increase coverage. The approach is looked at utilizing DARPA 1999 dataset and Standard data set for a port intricate dataset. Evaluation outcomes display that the suggested approach may reduce the amount of alerts by simply 94. 32%,

properly increasing notify management procedure. Because of the by using attire approach and optimum algorithms inside suggested approach, it might enlighten circle safety measures practitioner their state on the monitored circle in an on the internet approach.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] Y. Xu, C. Yan, J. Shi, Z. Lu, X. Niu, Y. Jiang, and F. Zhu, "An anomaly detection and dynamic energy performance evaluation method for HVAC systems based on data mining," *Sustainable Energy Technologies and Assessments*, vol. 44, pp. 101092–101092, 2021.
[2] A. Rajesh and S. &kiran, "Anomaly Detection Using Data Mining Techniques in Social Networking," *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, pp. 1268–1272.
[3] S. Jose, D. Malathi, B. Reddy, and D. &jayaseeli, "A survey on anomaly based host intrusion detection system," *Journal of Physics: Conference Series*, vol. 1000, pp. 12049–12049, 2018.
[4] M. H. Oh and G. &iyengar, "Sequential anomaly detection using inverse reinforcement learning," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & data mining*, pp. 1480–1490, 2019.
[5] J. Gao, X. Song, Q. Wen, P. Wang, L. Sun, and H. &xu, "Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks," 2020.
[6] C. Fan, F. Xiao, Y. Zhao, and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data," *Applied energy*, vol. 211, pp. 1123–1135, 2018.
[7] V. Hajisalem and S. &babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," *Computer Networks*, vol. 136, pp. 37–50, 2018.
[8] S. Thudumu, P. Branch, J. Jin, and J. J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data*, vol. 7, no. 1, pp. 1–30, 2020.
[9] T. Wen and R. Keyes, "Time series anomaly detection using convolutional neural networks and transfer learning," 2019.
[10] M. Riveiro, G. Pallotta, and M. &vespe, "Maritime anomaly detection: A review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 5, pp. 1266–1266, 2018.
[11] A. Guezzaz, Y. Asimi, M. Azrour, and A. &asimi, "Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection," *Big Data Mining and Analytics*, vol. 4, no. 1, pp. 18–24, 2021.
[12] L. Basora, X. Olive, and T. &dubot, "Recent advances in anomaly detection methods applied to aviation," *Aerospace*, vol. 6, no. 11, pp. 117–117, 2019.
[13] A. Dogan and D. &birant, "Machine learning and data mining in manufacturing," *Expert Systems with Applications*, vol. 166, pp. 114060–114060, 2021.
[14] T. Nolle, A. Seeliger, and M. &mühlhäuser, "BINet: multivariate business process anomaly detection using deep learning," in *International Conference on Business Process Management*, pp. 271–287, Springer, 2018.
[15] R. K. Pandit and D. Infield, "SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes," *IET Renewable Power Generation*, vol. 12, no. 11, pp. 1249–1255, 2018.
[16] L. Erhan, M. Ndubuaku, M. D. Mauro, W. Song, M. Chen, G. Fortino, . . &liotta, and A, "Smart anomaly detection in sensor systems: A multi-perspective review," *Information Fusion*, vol. 67, pp. 64–79, 2021.
[17] A. Capozzoli, M. S. Piscitelli, S. Brandi, D. Grassi, and G. &chicco, "Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings," *Energy*, vol. 157, pp. 336–352, 2018.
[18] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series," *Ieee Access*, vol. 7, 1991.
[19] L. Feremans, V. Vercruyssen, B. Cule, W. Meert, and B. Goethals, "Pattern-based anomaly detection in mixed-type time series," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 240–256, Springer, 2019.
[20] P. J. Rousseeuw and M. Hubert, "Anomaly detection by robust statistics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 2, pp. 1236–1236, 2018.
[21] M. Idhammad, K. Afdel, and M. &belouch, "Distributed intrusion detection system for cloud environments based on data mining techniques," *Procedia Computer Science*, vol. 127, pp. 35–41, 2018.
[22] E. Stripling, B. Baesens, B. Chizi, and S. &vandenbroucke *Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud. Decision Support Systems*, vol. 111, pp. 13–26, 2018.
[23] A. Chaudhary, H. Mittal, and A. Arora, "Anomaly detection using graph neural networks," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 346–350, 2019.