An Anomaly Intrusion Detection System Using C5 Decision Tree Classifier

Ansam Khraisat(✉), Iqbal Gondal, and Peter Vamplew

Internet Commerce Security Laboratory (ICSL), Federation University Australia, Ballarat, Australia
{a.khraisa,iqbal.gondal,p.vamplew}@federation.edu.au

Abstract. Due to increase in intrusion activities over internet, many intrusion detection systems are proposed to detect abnormal activities, but most of these detection systems suffer a common problem which is producing a high number of alerts and a huge number of false positives. As a result, normal activities could be classified as intrusion activities. This paper examines different data mining techniques that could minimize both the number of false negatives and false positives. C5 classifier's effectiveness is examined and compared with other classifiers. Results should that false negatives are reduced and intrusion detection has been improved significantly. A consequence of minimizing the false positives has resulted in reduction in the amount of the false alerts as well. In this study, multiple classifiers have been compared with C5 decision tree classifier using NSL_KDD dataset and results have shown that C5 has achieved high accuracy and low false alarms as an intrusion detection system.

Keywords: Malware · Intrusion detection system · NSL_KDD
Anomaly detection

1 Introduction

Anomaly Intrusion Detection Systems (AIDS) [3] have attracted the interest of many researchers due to their potential to detect a zero-day attack. AIDS recognizes abnormal user behavior on a computer system. The assumption for this technique is that attacker activity differs from normal user activity. AIDS [4] creates a behavior profile of normal user's activity by using selected features and machine learning approaches. It then examines the behaviors of new data with the predefined normal behavior profile and tries to identify abnormalities. Those behaviors of users which are unusual are identified as potential attacks.

In this research work, a range of data mining techniques including SVM, Naive Bayes, C4.5 implemented in the WEKA package (developed by the University of Waikato, New Zealand) as well as the C5 algorithm [10] were applied on the NSL-KDD dataset.

M. Ganji et al. (Eds.): PAKDD 2018, LNAI 11154, pp. 149–155, 2018.
https://doi.org/10.1007/978-3-030-04503-6_14

The rest of the paper is organized as follows. Related worked is discussed in Sect. 2. The IDS model with the dataset details is discussed in Sect. 3. Conceptual framework of our IDS model is proposed in Sect. 4. In Sect. 5, the experiment details are given and evaluation results are presented and discussed. Finally, we conclude the paper in Sect. 5.

2 Related Works

Some prior research has examined the use of different techniques to build AIDSs. Chebrolu et al. examined the performance of two feature selection algorithms involving Bayesian networks (BN) and Classification Regression Trees (CRC), and combined methods [2]. Karan et al. proposed a technique for feature selection using a combination of feature selection algorithms such as Information Gain (IG) and Correlation Attribute evaluation then they tested the performance of the selected feature by applying different classification algorithms such as C4.5, Naive Bayes, NB-Tree and Multi-Layer Perceptron [1]. Subramanian et al. propose classifying NSL-KDD dataset using decision tree algorithms to construct a model with respect to their metric data and studying the performance of decision tree algorithms [11].

C5 algorithm's performance is explored very well in a different domain such as modelling landslide susceptibility. Miner et al. used data mining techniques in the topic of landslide susceptibility mapping. They used C5 classifier to handle the complete dataset and address some limitations of WEKA, one of the best results were obtained from C5 applications [9].

3 IDS Model

A prediction model has two main components which are training phase and testing phase. In the training phase the normal profile is created, and in the testing phase the user actions are verified against the corresponding profile. We classify each of the collected data records obtained from the feature phase as normal or an anomaly. In the testing stage, we examine each model.

3.1 Classification

A classification technique is a systematic approach for building classification models from an input data set. Classification is the task of mapping a data item into one of a number of predefined classes [7]. Figure 1 shows a general approach for applying classification techniques.

Decision Trees. are considered one of the most popular classification techniques. Quinlan (1993) has advocated for the decision tree approach and the latest implementation of Quinlan's model is C5 [10]. In this paper we will apply C5 classifier, the algorithm has many advantages like:

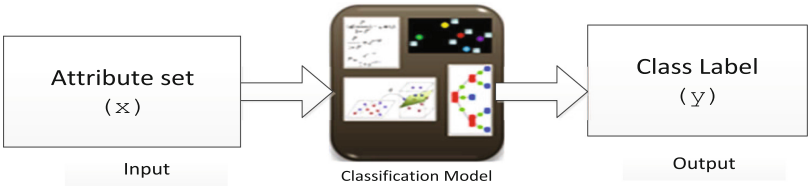

Fig. 1. Classification techniques

- Easy to understand the tree, as the large decision tree can be viewed as a set of rules. C5 can provide the knowledge for the noisy or missing data.
- Addresses over fitting and error pruning issues. Winnowing technique in C5 classifier can predict which attributes are relevant and which are not in the classification. It is useful while dealing with big datasets.

In machine learning, **Naive Bayes** classifiers are a family of least complex probabilistic classifiers based on using Bayes' theorem with robust (naive) independence assumptions between the attributes [8]. It is simple to build, with no complex iterative parameter estimation which makes it suitable for very large datasets. **SVM Model** is a demonstration of the examples as points in space, mapped so that the examples of the separate categories are split by a clear space that is as varied as possible. New examples are then matched into that similar space and predicted to belong to a group based on which side of the gap they belong to [6].

3.2 Framework of Intrusion Detection System

Our purpose is to examine different machine learning techniques that could minimize both the number of false negatives and false positives and to understand which techniques might provide the best accuracy for each category of attack patterns. Different classification algorithms have been applied and evaluated. Figure 2 shows a conceptual framework of our IDS.

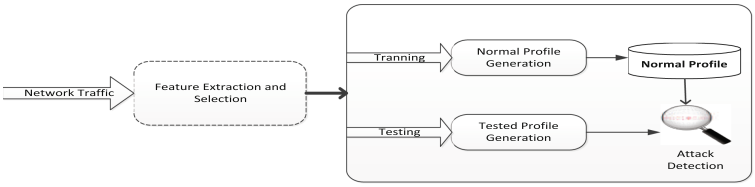

Fig. 2. Overall approach

Collected data is a network traffic, which is used to do feature extraction and selection. In the training phase, a normal profile is developed and in this

stage, the classifier is trained to detect the attacks. In the detection phase, data mining techniques are used to generate rule sets that are considered as abnormal activities and used by the classification algorithm already learned to classify the item set as an attack. After testing stage, we compute the accuracy rate, and other performance statistics to distinguish which classifier has predicted successfully.

4 Experimental Analysis

WEKA platform is used [5] to study J48, Naive Bayes and SVM. A commercial system from RuleQuest Research is used for C5 algorithm's [10]. NSL-KDD dataset is used [12]. We compared four different classifiers: C4.5, SVM, Naive Bayes and C5 to evaluate the performance of classification techniques.

4.1 Dataset Description

NSL-KDD data set has been used to overcome KDD cup99 dataset problem. A statistical analysis have been done on KDD cup99 dataset and found issues which have affected the ability to evaluate anomaly detection approaches. It is revealed the main issue is that KDD cup99 dataset has a huge number of redundant records [17]. NSL-KDD is considered as benchmark dataset in evaluating the performance of intrusion detection techniques [12].

The amount of training and testing records in NSL-KDD dataset are significant so the performance of classifiers can be evaluated reliably. The dataset has 125,973 records, where 67,343 are normal cases and 58,630 are anomalies. The dataset contains 22 types of attack, and 41 features.

4.2 Model Evaluation and Results

Our model will be evaluated based on the following standard performance measures:

- True positive (TP): Number of cases correctly predicted as anomaly. True negative (TN): Number of cases correctly predicted as normal.
- False positive (FP): Number of wrongly predicted as anomalies, when the classifier labels normal user activity as an anomaly. False negative (FN): Number of wrongly predicted as normal cases, when a detector fails to identify the anomaly.

Table 1 shows the confusion matrix for a two-class classifier. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

In the paper, we have used k-fold cross validation technique for performance evaluation. In this technique, dataset is randomly divided into k different parts.

In the evaluation, we measured the effectiveness and efficiency of different classification algorithms that wrongly identify the percentage of the False Negative alarm Rate (FN rate) and False Positive (FP rate). Table 2 provide the overall results of our experiments, which indicate that C5 classifiers are best at classifying the intrusions; it has successfully distinguished between normal and anomalous activity with minimum number of false alarm.

Table 1. Confusion matrix for an anomaly detection system

	Predicted class		
Actual class		Normal	Attack
	Normal	True Negative(TN)	False Positive(FP)
	Attack	False Negative(FN)	True Positive(TP)

Table 2. Confusion matrix for different classification algorithms

Classification algorithm	C5		C4.5		SVM		Naive Bayes	
Classified as	a	b	a	b	a	b	a	b
a = normal	67249	94	67200	143	66370	973	63060	4283
b = anomaly	121	58509	132	58498	2296	56334	7832	50798

Table 3 showed the accuracy for all the classifiers and shows that C5 classifiers have outperformed other classifiers in the study. C5 classifier has the highest accuracy of 99.82% which is followed by C4.5, SVM and Naive Bayes respectively. The number of false alarms, accuracy and time of building IDSs should be considered for IDS evaluation. Although C5 decision tree classifier wasn't faster classifier as shown in Table 4 C5 is the best in term of the accuracy and low false alarm. Naive Bayes is the fastest, but has the lowest accuracy by a substantial margin. The time that takes for generating the ruleset in C5 is 2.06, while the time that takes for generating the ruleset in c4.5 is 29.98, which is slower than C5. The reasons for this, in C5 the rules are generated separately.

Table 3. Accuracy in detection by using different algorithms

Classification algorithm	Accuracy
C5	99.82%
C4.5	99.78%
SVM	97.40%
Naive Bayes	90.38%

Table 4. Time Consuming for each classifier in seconds

Classification algorithm	Time
C5	70.6
C4.5	27.35
SVM	1423.92
Naive Bayes	1.02

5 Discussion and Conclusion

In this paper, an AIDS is proposed with the use of C5 classifier to detect both the normal and anomalous activities accurately. The aim of this approach is to identify attacks with enhanced detection accuracy and decreased false-alarm rates. We have established the robustness of our proposed techniques for intrusion detection by testing them on a NSL-KDD dataset that contains various types of intrusions. Our proposed method is evaluated on NSL-KDD dataset. Our experimental results indicate that our approach can detect malware traffic with a high detection rate of 99.82%. This demonstrates the significance of using C5 classifier in AIDS and makes the detection more effective. C5 are more powerful than C4.5, SVM and Naive Bayes because the memory usage is minimum, good speed and it also has excellent accuracy. In other words, C5 classifier provides high computational efficiency for classifier training and testing.

References

1. Bajaj, K., Arora, A.: Dimension reduction in intrusion detection features using discriminative machine learning approach. IJCSI Int. J. Comput. Sci. Issues **10**(4), 324–328 (2013)
2. Chebrolu, S., Abraham, A., Thomas, J.P.: Feature deduction and ensemble design of intrusion detection systems. Comput. Secur. **24**(4), 295–307 (2005)
3. Denning, D.E.: An intrusion-detection model. IEEE Trans. Softw. Eng. **2**, 222–232 (1987)
4. Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: techniques, systems and challenges. Comput. Secur. **28**(1–2), 18–28 (2009)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
6. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4), 18–28 (1998)
7. Lee, W., Stolfo, S.J., Mok, K.W.: A data mining framework for building intrusion detection models. In: Proceedings of the 1999 IEEE Symposium on Security and Privacy, pp. 120–132. IEEE (1999)
8. McCallum, A., Nigam, K., et al.: A comparison of event models for Naive Bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48. Citeseer (1998)

9. Miner, A., Vamplew, P., Windle, D., Flentje, P., Warner, P.: A comparative study of various data mining techniques as applied to the modeling of landslide susceptibility on the Bellarine Peninsula, Victoria, Australia (2010)
10. Quinlan, R.: Data mining tools See5 and C5. 0 (2004)
11. Subramanian, S., Srinivasan, V.B., Ramasa, C.: Study on classification algorithms for network intrusion systems. J. Commun. Comput. **9**(11), 1242–1246 (2012)
12. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009, pp. 1–6. IEEE (2009)