

A model for anomaly classification in intrusion detection systems

V O Ferreira, V V Galhardi, L B L Gonçalves, R C Silva, A M Cansian

Department of Computer Science and Statistics (DCCE), São Paulo State University (UNESP), São José do Rio Preto, Brazil.

E-mail: viniciusoliveira@acmesecurity.org

Abstract. Intrusion Detection Systems (IDS) are traditionally divided into two types according to the detection methods they employ, namely (i) misuse detection and (ii) anomaly detection. Anomaly detection has been widely used and its main advantage is the ability to detect new attacks. However, the analysis of anomalies generated can become expensive, since they often have no clear information about the malicious events they represent. In this context, this paper presents a model for automated classification of alerts generated by an anomaly based IDS. The main goal is either the classification of the detected anomalies in well-defined taxonomies of attacks or to identify whether it is a false positive misclassified by the IDS. Some common attacks to computer networks were considered and we achieved important results that can equip security analysts with best resources for their analyses.

1. Introduction

Intrusion Detection Systems (IDS) are aimed at monitoring the computer network traffic in some environment. They can be divided in two groups according to its detection method: anomaly detection and misuse detection [1]. In abuse detection, signatures representing previously known attacks are confronted with the current traffic in the monitored network to detect intrusions. In the other hand, anomaly detection tries to learn the normal pattern of the monitored traffic and detect any unusual behavior or anomaly. Although anomaly based methodologies appears quite efficient because they do not need signatures and have the capability to detect unknown attacks, they have some downsides. Anomaly based IDS suffer with the generation of high amounts of false positives and, also, the detected anomalies often have no clear information about the malicious events they represent, causing the analysis process difficult to take place.

In this context, this paper presents a model for automatic classification of anomalies detected by an anomaly based IDS in well-known taxonomies of attacks. Our main goal is to help security analysts in their analysis so that they can take the adequate counter measures in timely fashion. Besides, our methodology helps to identify the false positives, by classifying them into different classes of those that represent real attacks. We built a model based on the algorithm Autoclass and showed its capability to anomaly classification as well as the identification of false positives.

This paper is organized as follows: in section 2 we provide a brief discussion about anomaly classification field. In section 3 are detailed the Autoclass, the features we use for classification and other details about our model. In section 4 some discussion and results are presented. Finally, conclusions are shown in the section 5.



2. Anomaly Classification

Machine learning algorithms are widely used to classify data. Among machine learning algorithms, there are supervised and unsupervised methods. The main difference between these methods is that the supervised methods require previous labeled data while unsupervised methods work with unlabeled data. Nguyen and Armitage, in [2], do an extensive review about the use of machine learning algorithm for network traffic classification. They showed the efficiency of these algorithms compared to more traditional methods (port-based and payload based). Autoclass is an unsupervised algorithm which has shown good results in network classification as can be observed in [3], [4] and [5].

Although Autoclass has been used to classify legitimate network traffic, as far as we know, it had not been used yet to classify anomalous network traffic. Anomalous traffic classification is quite unexplored [6], with just a few works related in the literature, e.g. [6] and [7]. Therefore, our work presents a model which uses the unsupervised method Autoclass to classify common anomalies in computer networks.

3. Materials and Methods

The main component of our model is the machine learning algorithm we used for the classification. Autoclass is a partitional clustering algorithm and a fuzzy classifier. It uses an unsupervised approach which allows automatic discovery of the clusters inherent in a training dataset. This dataset contains unlabelled instances which are grouped according to its similar features, then, the clusters generated can be used in later steps to classify new unseen instances (i.e. anomalies, in the context of this work). To accomplish this, Autoclass combines Expectation Maximization algorithm with Bayesian theory to build a probabilistic model with distinct probability distributions that governs the classification in each cluster. As a fuzzy methodology, Autoclass allows the instances to be partly classified in more than one cluster. For further details about Autoclass we strongly encourage the reader to read [8].

Table 1. Autoclass classification features.

Name	Description
total_con	Number of flows/connections.
total_src_pkts	Number of packets.
total_src_bytes	Number of bytes.
synrate	Rate of flows with the TCP SYN flag set.
nullrate	Rate of flows without answer.
rstrate	Rate of flows with the TCP RST flag set.
distinct_dstports	Rate of distinct destination ports.
distinct_dstwkps	Rate of distinct well know destination ports.
distinct_srcwkps	Rate of distinct well know source ports.
distinct_dstaddrs	Rate of distinct destinations IPs accessed.

In our model, the clusters represent the possible attack classes into which the anomalies can be classified, in the case an anomaly is classified in more than one cluster we choose the classification with the highest probability as the anomaly's attack class. Autoclass classify the anomalies according to some pre-selected features of them, which must be grouped by each source IP address that access the monitored network. In this work we selected the features shown in Table 1, they are able to represent the traffic behaviour and can be easily extracted from protocols for traffic summarization such as Netflow, a widely deployed protocol in operational environments [6].

We designed our model to serve as generic approach for anomaly classification for different anomaly based IDS. In order to provide this interoperability, our model implements the IDMEF (Intrusion Detection Message Exchange Format) protocol, described by RFC 4765. Therefore, any IDS that exports its alerts in IDMEF standard, including the information from Table 1, can be served by the anomaly classification model proposed in this paper.

4. Tests and Results

In order to evaluate our model, we established a data collection environment from both: 1) a simulated network, with virtual machines, where real attacks were performed against target machines and 2) the internal network from ACME! Cybersecurity Research, the laboratory where this research was conducted. In the former network we collected anomalous traffic from the five attack classes our model is able to classify: Denial of Service (DoS), Dictionary Attack, Port Scan, Network Scan and Web Vulnerability Scan (Webscan). From the latter network we collected the legitimate traffic data generated from the laboratory users using the network in business days. This legitimate traffic includes HTTP traffic, chatting, remote access and other common operations performed by legitimate users.

The IDS we choose for the tests is described in [9]. We configured it to analyse the traffic generated in the data collection environment for two days. The IDS generated 7411 alerts, in the IDMEF standard, of which 5333 are anomalies caused by the attacks and 2078 are false positives caused by the legitimate traffic.

In order to use our model for the first time, one has to train the Autoclass so that it can learn about the different attack classes and the normal traffic. We trained the Autoclass with one day data from the data collection environment. At the end, Autoclass found 35 clusters of which 30 were mapped to legitimate traffic and the other five were mapped one for each attack. This mapping was possible due the attacks have been performed with specific source IP address.

After the training step we submitted the alerts to our model, a correct classification is when an alert of an unknown anomaly is correctly classified in the cluster mapped to the attack that generates such anomaly. Besides classify the anomalies, we also innovate in classifying the false positives detected by the IDS. False positives represent the misclassification of legitimate traffic as anomaly. Once we can classify the false positives, from the IDS, as legitimate traffic it is possible to save the time security analysts spend analysing these false alerts. Our model considers all clusters mapped to legitimate traffic as one comprehensive cluster able to classify the false positives detected by the IDS. In order to measure the classification performance we used three standard metrics [6] showed in equations 1, 2 and 3. Respectively, the first one measures the global performance and the last ones evaluate the performance for each cluster.

$$overall_accuracy = \frac{TruePositives}{TruePositives + FalsePositives} \quad (1)$$

$$recall(x) = \frac{TruePositives(x)}{TruePositives(x) + FalseNegatives(x)} \quad (2)$$

$$precision(x) = \frac{TruePositives(x)}{TruePositives(x) + FalsePositives(x)} \quad (3)$$

The recall and precision for each cluster as well as the overall classification accuracy is show in Figure 1. Our model had an overall accuracy of 87.17% which is in conformity with other works, presented in [3], [4] and [5], that used Autoclass for legitimate traffic classification. We achieved moderate results with some classes, in the Dictionary Attack cluster we achieved a recall of 71.83% because some instances were misclassified as false positives, the traffic generated for this attack was not so disparate from the legitimate traffic generated by the users. In the Webscan cluster we achieved a precision of 78.89% because some instances of DoS alerts were misclassified in this class, we believed it happened because these attacks are similar, since both perform several requests to a certain target.

Despite that, we achieved a recall of 99.9% in the false positives classification which means that almost all false positives detected by the IDS were correctly identified.

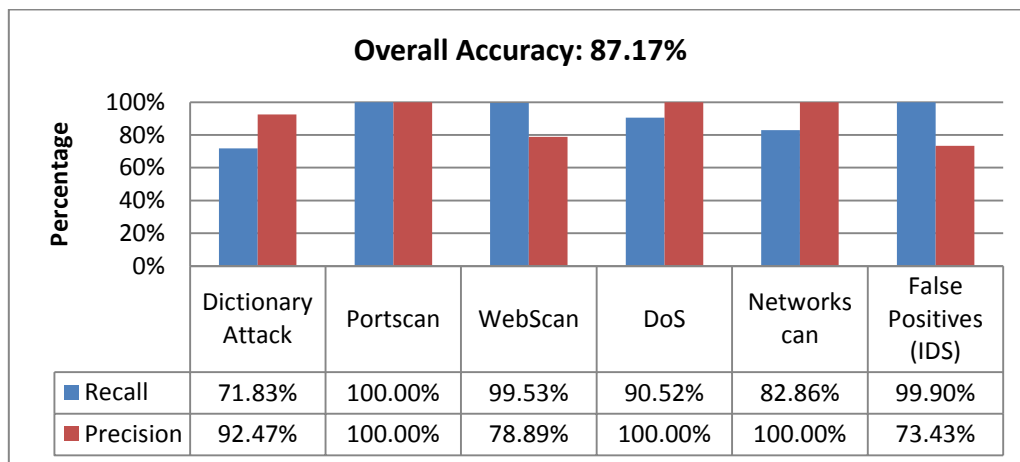


Figure 1. Autoclass classification accuracy.

5. Conclusion

In this work we present a new model for anomaly traffic classification using the algorithm Autoclass. We achieved an overall accuracy of 87.17%, a value in conformity with related work that used Autoclass for legitimate traffic classification. Therefore, we achieved our goal to show the Autoclass potentiality to also classify previously unknown anomalies in well-defined taxonomies of attacks. Moreover, we showed another benefit of anomaly classification, since our model could identify 99.9% of false positives detected by the IDS, it could be used as an additional feature to reduce false alerts in Intrusion Detection Systems.

6. Acknowledgments

The authors thank FAPESP, CAPES and FAPERP (proc. 68/2015) for supporting this research.

References

- [1] Wu S X, Banzhaf W 2010 The use of computational intelligence in intrusion detection systems: A review *Appl Soft Comput.* **10** 1–35
- [2] Nguyen T, Armitage G 2008 A survey of techniques for internet traffic classification using machine learning *IEEE Commun Surv Tutor.* **10** 56–76
- [3] Zander S, Nguyen T, Armitage G 2005 Automated traffic classification and application identification using machine learning. *IEEE Conf Local Comput Networks 30th Anniv.* 250–257
- [4] Erman J, Arlitt M 2006 Traffic classification using clustering algorithms *SIGCOMM work on min net data* 281–286
- [5] Erman J, Mahanti A, Arlitt M 2006 QRP05-4: Internet Traffic Identification using Machine Learning *IEEE Globecom* 1–6.
- [6] Paredes-Oliva I, Castell-Uroz I, Barlet-Ros P, Dimitropoulos X, Sole-Pareta J 2012 Practical anomaly detection based on classifying frequent traffic patterns *Comput Commun Work (INFOCOM WKSHPs) IEEE Conf* 49–54
- [7] Tellenbach B, Burkhart M, Schatzmann D, Gugelmann D, Sornette D 2011 Accurate network anomaly classification with generalized entropy metrics *Comput Networks.* **55** 3485-3502.
- [8] Cheeseman P and Stutz J 1996 Bayesian classification (AutoClass): theory and results *In Advances in knowledge discovery and data mining.* American Association for Artificial Intelligence, Menlo Park, CA, USA 153-180.
- [9] Batista M L and Cansian A M 2011 Detecção de eventos em redes de computadores utilizando detecção de novidade *IADIS Ibero Americana WWW/Internet.* Rio de Janeiro, RJ, Brazil.