Adaptive Alarm Filtering by Causal Correlation Consideration in Intrusion Detection*

Heng-Sheng Lin, Hsing-Kuo Pao, Ching-Hao Mao, Hahn-Ming Lee, Tsuhan Chen, and Yuh-Jye Lee

Abstract. One of the main difficulties in most modern Intrusion Detection Systems is the problem of massive alarms generated by the systems. The alarms may either be false alarms which are wrongly classified by a sensitive model, or duplicated alarms which may be issued by various intrusion detectors or be issued at different time for the same attack. We focus on learning-based alarm filtering system. The system takes alarms as the input which may include the alarms from several intrusion detectors, or the alarms issued in different time such as for multi-step attacks. The goal is to filter those alarms with high accuracy and enough representative capability so that the number of false alarms and duplicated alarms can be reduced and the efforts from alarm analysts can be significantly saved. To achieve that, we consider the causal correlation between relevant alarms in the temporal domain to re-label the alarm either to be a false alarm, a duplicated alarm, or a representative true alarm. To be more specific, recognizing the importance of causal correlation can also help us to find novel attacks. As another feature of our system, our system can deal with the frequent changes of network environment. The framework gives the judgment of attacks adaptively. An ensemble of classifiers is

* This work was partially supported by the iCAST project sponsored by the National Science Council, Taiwan, under the Grant No. NSC97-2745-P-001-001.

Heng-Sheng Lin
Trade-Van Information Services Co., Taipei, Taiwan
e-mail: ed.lin@tradevan.com.tw

Hsing-Kuo Pao, Ching-Hao Mao, Hahn-Ming Lee, and Yuh-Jye Lee
National Taiwan University of Science and Technology, Taipei, Taiwan
e-mail: pao@mail.ntust.edu.tw, d9415004@mail.ntust.edu.tw, yuh-jye@mail.ntust.edu.tw

Hahn-Ming Lee
Academia Sinica, Taipei, Taiwan
e-mail: hmlee@mail.ntust.edu.tw

Tsuhan Chen
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: tsuhan@cmu.edu

K. Nakamatsu et al. (Eds.): New Advan. in Intel. Decision Techno., SCI 199, pp. 437–447.
springerlink.com

adopted for the purpose. Accordingly, we propose a system mainly consisting of two components: one is for alarm filtering to reduce the number of false alarms and duplicated alarms; and one is the ensemble-based adaptive learner which is capable of adapting to environment changes through automatic tuning given the expertise feedback. Two datasets are evaluated.

Keywords: Intrusion detection, alarm filtering, false alarm, adaptive learning, ensemble.

1 Introduction

Over the last couple of decades, the intrusion methods are getting sophisticated and diversified. Variety of rootkits and exploit codes are easily obtained for the hackers to attack the systems. Therefore, individual data can be illegally read or overwritten by intruders. Many different Intrusion Detection Systems (IDSs) have been provided to detect those malicious attacks. However, one of the weakest points with those IDSs is the problem of massive false alarms (or false positives). As revealed in several reports, IDS usually generates nearly 99% of false alarms in the detection [1, 5]. On the other hand, various alarms, which could be issued by naïve decision rules may come from the same unique attack. For instance, several minor alarms may suggest a multi-step attack. A clever detection system should be able to single out the reason for further automatic or non-automatic analysis. Overall, false alarms or duplicated alarms can waste significant time from human analyzers. Without considering those issues, an IDS can be virtually useless. The problem is more serious when no enough human analyzers can be assigned for further analysis of the generated alarms. E.g., a personal IDS will not afford such overhead. In this work, we propose an alarm filtering (AF) framework which can significantly reduce these two kinds of alarms: false alarms and duplicated alarms. The framework considers causal correlation between alarms and alarms will then be issued with high accuracy and no redundancy. As another important feature, to apply our system to real network, we would like to make the final decision of alarm classification adaptively to different periods and to different environment. An ensemble of classifiers called ensemble-based adaptive learner (EAL) will be adopted to adjust the prediction precision and sensitivity for the system according to network conditions. The feedback from alarm analysts will be used to tune the setting periodically.

To reduce the false alarms and duplicated alarms, our AF system considers causal correlation between several alarms when they are either temporally correlated or associated with a single attack. Alarm correlation [12, 13] has been used in discovering the intentions or root cause of the attackers [3] and how they achieve their goals [8], i.e. the attack methods. Based on our observations, single minor alarm in small scale may be confusing, but minor alarms collected as a whole may indicate a serious attack. When lacking of considerations in large scale, some alarms may be mislabeled, so called the *false alarm problem*. On the other hand, multi-step attacks often trigger a bunch of alarms in a sensitive system to downgrade the performance of the system, so called the *duplicated alarm*

problem. To deal with these two problems, one has to consider causal correlated alarms instead of a single alarm. Note that reduction of false alarms may lower the detection sensitivity (also known as recall). A trusted system must still be sensitive enough to detect serious attack and at the same time only a small number of false alarms are generated. As a challenging but an important extreme, our AF system will have the ability to identify novel alarms or alarms related to *novel attacks*. Causal correlation is considered for those alarms which may be associated with some anomaly behaviors and the final judgment can then be given with high confidence. We need to emphasize that the single alarm is usually issued from some naïve decision rules or signature alignment, to deal with various special cases or to solve some particular problems. Such rules may be created by a simple-minded consideration without too much rigorous efficiency analysis of the whole system. On the other hand, some rules may be created with a global view and lack of ability to fit into special environment, e.g., the period when new attacks just being released, or the "normal" period with low number of "background alarms". Our system offers a solution for that. To consider the causal correlation for a set of alarms, we are possible to relate the alarms to true attack or the attack of high risk, including novel attacks.

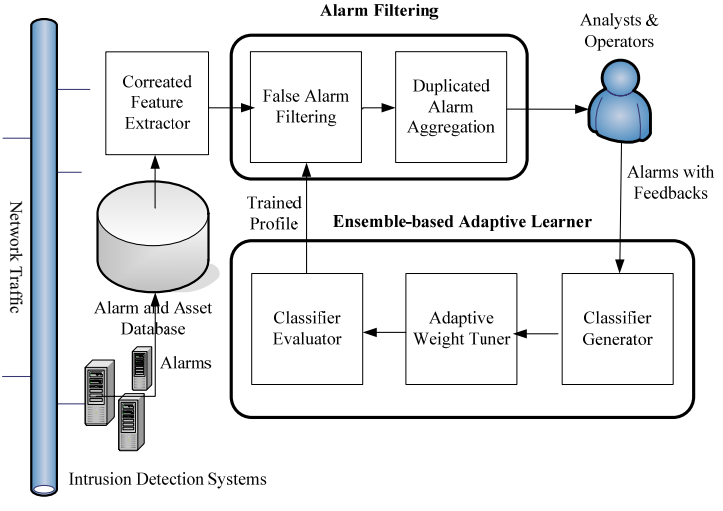


Fig. 1 Architecture of Adaptive Alarm Filtering System

The consideration of causal correlation can reduce both of the false alarms and duplicated alarms, and at the same time can deal with novel attacks. However, the network environment is usually not stable. As a result, operators have to tune and confirm the setting of IDSs frequently for the changes. That creates a burden for the operators. Due to the changes of network environment including devices, services or attack approaches, the pre-trained classifier will be getting to lose its accuracy on prediction after a certain period. This phenomenon is called *concept drift* [11] and happens very often in the real world. To address the problem, we

consider automatically tuning the system so that the intrusion detection can adapt to the environment. An *adaptive* learning system filters false and duplicated alarms after IDSs and adaptively learns from responses of experts recurrently [6, 9]. That helps system operators relieved from laboring works on parameter tuning. Different from previous research, our proposed framework gets more focused on practical issues of changing network environment over different periods or even different sites. Using the proposed *ensemble*-based adaptive learner (called EAL) is eligible to give robust performance on prediction as time goes by or for different network environment. It is understood that the network data are large-scale stream data, and usually highly unbalanced between attacks and normal data. Besides, most attacks happen in a continuous fashion in a very short period of time. To deal with those concerns, our EAL is proposed based on entropy computation, also inspired by AdaBoost [10]. Also, some aging effect is added to the system. By our approach, the rare attacks are not to be overlooked and can contribute some to our system. Other than adapting to time, we can also make our system adaptive to different commercial organizations. We will take the risk of assets as the input for further system improvement. Overall, we aim at designing a system which can be applied to real environment.

2 Feature Extraction, Alarm Filtering and Adaptation

In this section, we discuss our proposed system in full details, as illustrated in Fig. 1. Different from most IDSs, we focus on the reduction of false alarms and duplicated alarms after IDSs issue alarms. To deal with real network data, we also consider an adaptive system where the attack call may depend on the time information. Our system can take alarms from several sources, e.g., distributed IDSs, as the input. Basically, the system can be separated into three parts: Feature Extraction Unit, Alarm Filtering Unit, and Ensemble-based Adaptive Learning Unit. We proceed to give details for those different units.

2.1 Feature Extraction

Many factors combined together to decide which alarm comes from a true attack. They include causal correlated alarms, unusual changes of frequency in alarm issuing, and asset information, etc. That is, the feature extraction set is beyond the common attributes of intrusion alarms, such as packet size, signature names, IP addresses, port numbers and so on. We need to know that, in some cases, the IP address may limit the generalization ability of the model [3]. Opposed to that, the properties of hosts are more strongly relevant to attacks. Therefore, an IP address is replaced by its corresponding asset information. Below, we illustrate those features one by one.

Causal Alarm Features. Our idea is to correlate the alarm in the previous step and the alarm at this moment as causal correlation features. As we can imagine, the features will help us to detect a multi-step attack. More than that, the features

also provide more information on recognizing alarms which may be from novel attacks. For instance, if an alarm never happened in historical data, it is still possible to be classified correctly according to the conditional probability of the alarm after observing the pre-step alarms. To be more specific, those features consist of the combination of present alarm tag and the alarm tag in the previous stage (or pre-step alarm), under three different conditions, pre-step alarm with the same source, or same target as the current alarm, or both being the same.

Abnormal Frequency Value. Several triggered alarms are like "background noises" which happen all the time and may not suggest any meaningful information. The true alarm usually has an instant change that is out of the range of its normal frequency. We can compute mean μ, and variance σ^2, of each individual alarm a, on daily basis from historical data. The attribute, being able to real-time measures the degree of anomaly on individual alarm, is formulated as follows:

$$\mathrm{AFV}(a,x)=\begin{cases}\ln\left(\dfrac{(x-\mu_a)^2}{2\sigma_a^{\,2}}\right) & x>\mu_a \\ 0 & \text{others}\end{cases} \tag{1}$$

where x is an accumulated amount of corresponding alarms of the present day, a refers to the alarm identity of x, μ_a is the average number of alarms correspondent to alarm a, and σ_a^2 is the correspondent variance.

2.2 Alarm Filtering

Alarm Filtering (AF) is to reduce the amount of alarms. To provide succinct alarm report to users, we first classify alarms to a class of relevant or irrelevant to an attack, and then aggregate duplicated alarms as an alarm group on behalf of a high level event for users. We consider causal correlation of alarms to reduce the false alarms, as described previously. An ensemble-based classifier, combined with basic learners is adopted as false alarm filtering. In order to adapt to changing network condition, the filter employs ensemble-based adaptive learner to keep updating (discussed later) as time goes by.

2.3 Ensemble-Based Adaptive Learner

Inspired by AdaBoost, the proposed EAL algorithm is a meta-learning method aimed to combine multi-classifiers when data are incrementally grown with time. Different from many related works [2, 4], the proposed algorithm is specifically focused on the characteristics of computer network and practical requirement of SOC.

In our ensemble of classifiers, there are two types of weights should be optimized. They are example weight for each instance and classifier weight, so called voting weight for each weak classifier. To learn from feedbacks, alarm log is separated by day as $D_J = D_{J-1} \bigcup d_J$ and $D_0 = \phi$ where $d_j = \{(\mathbf{x}_{ji}, w_{ji}, z_{ji})_{i=1}^{n(j)}\}$, $n(j)$

denotes the amounts of alarms in the *j*-th day, $\boldsymbol{x}_{ji}$ is an alarm feature vector of the *i*-th example, $w_{ji} \in [0, 1]$ denotes the corresponding example weight, and $z_{ji} \in Z = \{-1,1\}$ indicates the corresponding real class in the *j*-th day. We set different example weights according to its class in our experiments. Borrowing the concept of entropy from information theory, we can deal with the problem of unbalanced data. Moreover, sample re-weighting, like AdaBoost, and complementary learning from previous wrongly predicted examples, strongly enhance the robustness of our system. The function is listed as follows:

$$h_{final}(x) = \arg\max_{z \in Z} \sum_{k=1}^{m} \varphi(a_k, \lambda, \tau) \cdot v_k \cdot h_k(x, z) \tag{2}$$

Where $h_{final}(\circ)$ is the finial hypothesis of committee decision with *m* member classifiers, $h_k(\circ)$ represents the hypothesis of the *k*-th day, $v_k = (1 + entropy(P(d_k))) \cdot \log((1-\varepsilon_k)/\varepsilon_k)$ is the corresponding voting weight decided by its error rate ε, and the entropy defined by, $entropy(P) = -P\ln(P) - (1-P)\ln(1-P)$, indicating the information of the *k*-th training data point for the distribution. The $P(d_k)$ is the portion of true alarms in a training set. Finally, in order to being adaptive to changes, the Memory Decline Ratio (MDR) listed as follows will be used:

$$\varphi_k(a_k, \lambda, \tau) = \frac{\exp(-\lambda(a_k - \tau))}{1 + \exp(-\lambda(a_k - \tau))}, \tag{3}$$

which is inspired by aging-forgetting mechanism and modified by sigmoid function. It is employed to tune the voting weight through the time in each individual classifier. The forgetting slope λ, is set for how fast to drop out a useless classifier with tolerating time τ. Setting the pair of parameters will be discussed in experiment and, actually, depends on the degree of concept drift or change in each dataset.

3 Experiments

We have built a prototype system to demonstrate the proposed approach that is able to filter the alarms of both kinds, the false alarms and the duplicated alarms. Below, we discuss different measurement on the system of alarms filtering such as False Positive (FP), which analysts have to pay extra effort with, and True Negative (TN), which means that filtered alarms are indeed not correlated to attacks. Of course, filtering out the true alarms associated to attacks is more serious than anything else, e.g. achieving high TP rate. Moreover, the ability to identify novel alarms is also taken into account. There are two experiments designed for demonstration. One is to evaluate that the proposed feature set including the causal correlation features is able to enhance the performance of intrusion detection. The second experiment is to compare with different learning schemes to support that our approach is able to filter out false or duplicated alarms and effectively adapt to network change especially when novel alarms happen.

Novel alarms make operators tune the setting of IDSs from time to time. The novel alarms also make pre-trained model useless after a while. Hence, we especially discuss the ability of our framework to identify novel alarms, meaning that our AF is able to give the correct predicted class on an unseen alarm under an acceptable level of false positive rate. By means of Receiver Operating Characteristic (ROC) curve, the False Positive (FP) rate referring to cost and the True Positive (TP) rate referring to detection ability allow analysts to know the trade off between detection rate and cost.

3.1 Datasets

The system was validated by two datasets. The first one is made by a popular benchmark, DARPA 1999 [7]; the other is a real world private alarm dataset, which is provided from an SOC operated in Taiwan, called A-SOC 2007. The center offers a service of security surveillance to their clients, including many organizations, government departments and companies. Both of the alarms of DARPA and A-SOC are manually labeled for evaluation according to its official report and warning tickets to monitored client, respectively.

Data Distribution. Their data distributions are shown in Table 1.

Table 1 Distribution of each dataset for Experiments. Novel alarms represent that an alarm is never seen before the day

Dataset	Duration	Distinct IP	Distribution of Alarm Label		
			Total (Novel)	True (Novel)	False (Novel)
DARPA	1999.3.1 ~ 1999.4.10	546	55,473 (3,693)	19,109 (2,191)	36,364 (1,502)
A-SOC	2007.8.30 ~ 2007.9.7	5,368	308,063 (56,984)	6,743 (2,978)	301,320 (54,006)

3.2 Performance Measurement

Receiver Operating Characteristic (ROC) curve is adopted as the main performance measurement. The unbalanced problem makes the performance hard to be evaluated. Because the number difference between true alarms and false alarms is large, the enhancement of performance on identifying rare true alarms is easy to be overlooked if using Accuracy as a measurement. ROC curve, which consists of TP rate and FP rate, is suitable to be a performance measurement for this. In the viewpoint of system security operators, they want to know how much cost (FP rate) they have to pay if keeping a level of sensitivity of recognizing rare attacks. ROC curve can serve this purpose because false positive rate is like a cost we have to pay if we want to reach a level of true positive rate (alarm detection rate).

Table 2 Performance Test with Different Feature Combinations. The performance comparison is TP rate (detection rate) vs. FP rate (cost). The detection rate of novel alarms is specifically demonstrated for revealing the ability to detect novel alarms with different feature combinations. Basic feature set is the original attributes generating from Snort IDS but excluding source and destination IP addresses

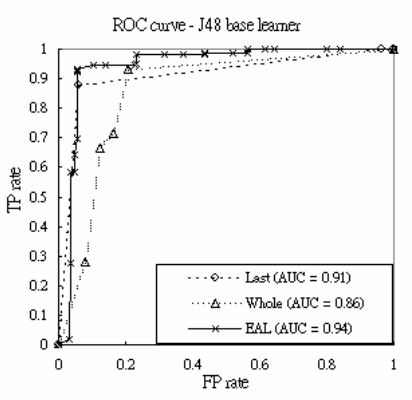
Dataset	Feature Combination	Cost	All Alarms	Novel Alarms	Correctly Filtered
		FP rate	TP rate	TP rate	
DARPA	Basic	18.83%	82.45%	85.76%	89.8%
	Causal	5.26%	95.08%	87.36%	97.34%
A-SOC	Basic	37.19%	64.01%	87.58%	98.73%
	Causal	28.03%	62.36%	88.68%	98.84%

Note: Our experiment takes all alarms as the input. Correctly Filtered Alarms represent the re-labeled false alarms that are indeed not associated to attacks and can be filtered out appropriately

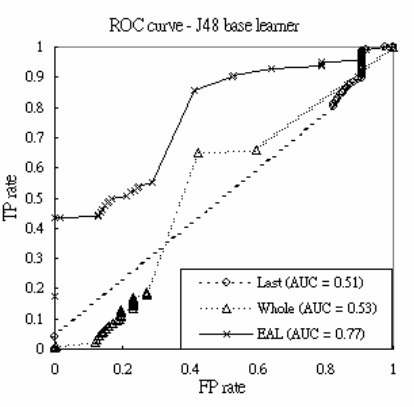
3.3 Results and Discussion

Feature Set Evaluation. The first experiment is to demonstrate that the causal correlation is helpful on alarms filtering without sacrificing the detection rate. To achieve that, we test different feature combinations with DARPA and A-SOC datasets and analyze the performance result. As shown in Table 2, the proposed causal feature set including basic alarm information, asset, causal correlation and variance frequency has the best performance, higher detection rate and lower FP rate, especially on detecting novel alarms. Moreover, adopting the causal correlation features to classify SOC dataset causes that the false positive rate is greatly reduced about 10%. The reason is that the SOC dataset gathered in 2007 has more sophisticated multi-step attacks than DARPA 1999. Therefore, the causal correlation feature set has greater enhancement on SOC dataset than on DARPA 1999.

Cost and Detection Rate. To evaluate our approach on the tradeoff of cost and detection rate, our proposed EAL are compared with two other generic schemes in the second experiment. The first controlled scheme is that the decision model only keeps the last classifier in alarm ensemble classifiers for prediction. The second one is to keep all previous trained classifiers and combine them with the same weight in alarm ensemble classifiers for prediction. The comparison of ROC curve is shown in Fig. 2, which demonstrates that our approach receives the largest area-under-curve (AUC) values in both of the DARPA and A-SOC datasets. The ROC curve provides analysts the view of how much the cost has to pay if the model can identify alarms including novel alarms, as shown in Fig 2(a) and 2(b). The cost implicitly means that analysts have to spend their time to pick out the false alarms. Obviously, the high cost is unpractical when the system is operated on real environment. All comparison results are illustrated in Table 3. Our proposed EAL also performs very well on correctly filtering alarms without



(a) ROC curve of DARPA dataset



(b) ROC curve of A-SOC dataset

Fig. 2 Capability of detecting false alarms - ROC curves of different incremental classifier schemes using a base learner, J48 from Weka

Table 3 Performance comparison, TP rate (detection rate) vs. FP rate (cost), with different learning schemes. The scheme of only last classifier is to predict alarms by the classifier trained from the previous day. And, whole classifier means the alarm ensemble classifiers combining all of previous trained classifiers for prediction. The last scheme is our proposed approach EAL, for prediction

Dataset	Learning scheme	Cost FP rate	All Alarms TP rate	Novel Alarms TP rate	Correctly Filtered Alarms TN / (TN + FN)
DARPA	Only Last Classifier	5.92%	87.96%	81.79%	93.7% (34,210 / 36,511)
	Whole Classifier	16.46%	71.52%	20.58%	84.81% (30,378 / 35,820)
	EAL	5.7%	93.22%	87.17%	96.36% (34,292 / 35,588)
A-SOC	Only Last Classifier	6.43%	48.63%	56.28%	98.79% (281,938 / 285,402)
	Whole Classifier	30.08%	49.35%	80.86%	98.4% (210,690 / 214,105)
	EAL	28.03%	62.36%	88.68%	98.84% (216,854 / 219,392)

sacrificing much to deal with novel or rare true alarms. In traditional methods on false alarm reduction, assessing risk through an asset table or setting a correlation rule to recognize true alarms is possible to ignore novel alarms. Therefore, they can only aware of known attack and lack of ability to defense from a new threat. However, in our experiment, it actually shows the ability to find out novel alarms.

Aggregation Duplicated Alarm. There are 55,473 individual alarms in DARPA dataset grouped into 11,197 groups, including only two impure groups, which include true and false alarms in the same group. The reduced amount significantly relieves about 75% of overall alarms. On the other hand, the 308,063 individual alarms in A-SOC dataset are grouped into 25,662 groups with 27 impure groups but still helpful for analysts. As a result, analysts just need to confirm the succinct alarm groups with predicted true class of alarms instead of a large number of indi-

vidual alarms. About 92% labor is saved in A-SOC dataset by means of our approach. Actually, our grouping approach does not only save work for analysts but also make analysts easily giving feedback for enhancing the ability of the alarm ensemble classifiers on further prediction.

4 Conclusions

We proposed a system for adaptive alarm filtering. Our goal is to enhance the performance of IDS through reducing the number of false alarms and duplicated alarms. Other than that, our system can be operated in an adaptive fashion. The proposed learning-based alarm filtering system does not only classify alarms with high confidence but also adaptively change with time goes by according to feedback from experts. Moreover, through our feature set including the causal correlation features, the system also makes identifying novel alarms possible. After evaluation of experiments on DARPA and A-SOC dataset, all individual alarms are aggregated as groups, which reduces size to about 25% and 8% from original alarms respectively. In the mean while, with at least 87% novel alarm detection rate, about 96% to 98% of the false alarms have been correctly filtered out in DARPA and A-SOC dataset. After false alarms significantly identified by proposed approaches, analysts can actually pay more attentions on the main courses such as intrusion analysis and related responses.

References

1. Alharbt, A., Imai, H.: IDS False Alarm Reduction Using Continuous and Discontinuous Patterns. In: Proc. of the 3th International conf. on Applied Cryptography and Network Security (ACNS 2005), pp. 192–205 (2005)
2. Fern, A., Givan, R.: Online ensemble learning: An empirical study. Machine Learning 53(1), 71–109 (2003)
3. Julisch, K.: Clustering Intrusion Detection Alarms to Support Root Cause Analysis. ACM Trans. on Information and System Security (TISSEC) 6(4), 443–471 (2003)
4. Kidera, T., Ozawa, S., Abe, S.: An Incremental Learning Algorithm of Ensemble Classifier Systems. In: Proc. of the International Joint Conf. on Neural Networks (IJCNN 2006), BC, Canada, pp. 3421–3427 (2006)
5. Law, K.H., Kwok, L.F.: IDS False Alarm Filtering Using KNN Classifier. In: Lim, C.H., Yung, M. (eds.) WISA 2004. LNCS, vol. 3325, pp. 114–121. Springer, Heidelberg (2005)
6. Liaw, K.-K., Wu, Y.-L.: False Alarm Filtering Using SVM and Sliding Window. In: The 3rd Joint Workshop on Information Security (JWIS 2008), Seoul, Korea (July 2008)
7. Mahoney, M.V., Chan, P.K.: An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection. In: Vigna, G., Krügel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 220–237. Springer, Heidelberg (2003)

8. Ning, P., Cui, Y., Reeves, D.S., Xu, D.: Techniques and tools for analyzing intrusion alerts. ACM Trans. on Information and System Security (TISSEC) 7(2), 274–318 (2004)
9. Pietraszek, T.: Using adaptive alert classification to reduce false positives in intrusion detection. In: Jonsson, E., Valdes, A., Almgren, M. (eds.) RAID 2004. LNCS, vol. 3224, pp. 102–124. Springer, Heidelberg (2004)
10. Schapire, R., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margins: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26(5), 1651–1686 (1998)
11. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning 23(1), 69–101 (1996)
12. Valeur, F., Vigna, G., Kruegel, C., Kemmerer, R.A.: Comprehensive approach to intrusion detection alert correlation. IEEE Trans. on Dependable and Secure Computing 1(3), 146–169 (2004)
13. Zhu, B., Ghorbani, A.A.: Alert Correlation for Extracting Attack Strategies. International Journal of Network Security 3(3), 224–258 (2006)