Improvement in Intrusion Detection with Advances in Sensor Fusion

Ciza Thomas and N. Balakrishnan

***Abstract*—Various Intrusion Detection Systems reported in literature have shown distinct preferences for detecting a certain class of attack with improved accuracy, while performing moderately on the other classes. In view of the enormous computing power available in the present-day processors, deploying multiple Intrusion Detection Systems in the same network to obtain best-of-breed solutions has been attempted earlier. The paper presented here addresses the problem of optimizing the performance of Intrusion Detection Systems using sensor fusion with multiple sensors. The trade-off between the detection rate and false alarms with multiple sensors is highlighted. It is illustrated that the performance of the detector is better when the fusion threshold is determined according to the Chebyshev inequality. In the proposed Data-dependent Decision fusion method, the performance optimization of individual Intrusion Detection Systems is first addressed. A neural network supervised learner has been designed to determine the weights of individual Intrusion Detection Systems depending on their reliability in detecting a certain attack. The final stage of this Data-dependent Decision fusion architecture is a sensor fusion unit which does the weighted aggregation in order to make an appropriate decision. This paper theoretically models the fusion of Intrusion Detection Systems for the purpose of demonstrating the improvement in performance, supplemented with the empirical evaluation.**

***Index Terms*—Intrusion Detection Systems; Sensor Fusion; Neural Network; Data-dependent Decision Fusion; Chebyshev Inequality.**

I. Introduction

An Intrusion Detection System (IDS) gathers information from a computer or a network, and analyzes this information to identify possible security breaches against the system or the network. An observation of various IDSs available in literature shows distinct preferences for detecting a certain class of attack with improved accuracy, while performing moderately on the other classes. The availability of enormous computing power has made it possible for developing and implementing IDSs of different types on the same network. The integration of the decisions coming from different IDSs has emerged as a technique that could strengthen the final decision. Sensor fusion can be defined as the process of collecting information from multiple and possibly heterogeneous sources and combining them to obtain a more descriptive, intuitive and meaningful result [1].

Most of the related works [1], [2], [3], [4], [5], in the field of sensor fusion have been carried out mainly with one of the methods like probability theory, evidence theory, voting fusion theory, fuzzy logic theory or neural network, to aggregate information. It is clear from all the previous works in sensor fusion that there occur more effective means of analyzing the information provided by existing IDSs, thereby causing an effective data refinement for knowledge recovery. In spite of all such attempts, the state-of-the-art IDS performance leaves room for further improvement.

An analysis of the poorly detected attacks reveals the fact that the attacks are characterized by features that do not discriminate them much. An attempt to show the improved performance of multiple IDSs using rule-based fusion has been attempted in the work of Thomas and Balakrishnan [6]. The rule-based fusion systems work only with small input data and there is the need for machine learning algorithms to handle the type of data appearing on the network traffic. The rule-based fusion also has the limitation of being dependent on the individual IDSs that are used in sensor fusion. It is necessary to incorporate an architecture that considers a method of improving the detection rate by gathering an in-depth understanding about the input traffic and also the behavior of the individual IDSs. This helps in automatically learning the individual weights for the combination when the IDSs are heterogeneous and show difference in performance. The architecture should thus be data-dependent and hence the rule set has to be developed dynamically. A new Data-dependent Decision(DD) fusion architecture underpinning sensor fusion to significantly enhance the IDS performance is proposed in the work of Thomas and Balakrishnan [7]. The improved performance of the data-dependent decision fusion architecture can be shown both theoretically as well as experimentally with an approach adopted for optimizing both the local sensors and the fusion unit with respect to the error rate. The optimal performance along with the complexity of the task bring to the fore the need for theoretically sound basis for the sensor fusion techniques in IDSs.

The motivation of the present work was the fact that the empirical evaluation as seen in [7] was extremely promising with the data-dependent decision fusion. The modeling can be extremely useful with a complete addressing of the problem with sound mathematical and logical concepts. Thus the present work employs modeling to augment the effective mathematical analysis of the improved performance of sensor fusion and to develop a rational basis which is free from the various techniques used.

The remaining part of the paper is organized as follows. Section II discusses the related work of sensor fusion in IDS. In section III, the model of the proposed data-dependent decision fusion architecture is presented by modeling the constituent parts and also by stating the problem. Algorithms

Ciza Thomas and N. Balakrishnan are with the Department of Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore, India, e-mail: ciza_thomas@yahoo.com

Manuscript received June 24, 2008; revised December 29, 2008; accepted March 23, 2009.

for optimizing the local detectors along with a data-dependent decision fusion architecture for optimizing the fusion criterion are also presented in section III. Section IV contains the experimental results along with the discussions regarding the higher performance of the proposed architecture within the threshold bounds and the trade-off between various network parameters. Finally, the concluding comments are presented in section V.

II. Related Work

Tim Bass [2] presents a framework to improve the performance of intrusion detection systems based on data fusion. A few first steps towards developing the engineering requirements using the art and science of multi-sensor data fusion as an underlying model is provided in [2]. Giacinto et al. [3] propose an approach to intrusion detection based on fusion of multiple classifiers. Didaci et al. [4] attempt the formulation of the intrusion detection problem as a pattern recognition task using data fusion approach based on multiple classifiers. Wang et al. [5] present the superiority of data fusion technology applied to intrusion detection systems. The use of data fusion in the field of DoS anomaly detection is presented by Siaterlis and Maglaris [1]. The detection engine is evaluated using the real network traffic. Another work incorporating the Dempster-Shafer theory of evidence is by Hu et al. [8].

Siraj et al. [9] discuss a Decision Engine for an Intelligent Intrusion Detection System (IIDS) that fuses information from different intrusion detection sensors using an artificial intelligence technique. Thomopolous in one of his work [10], concludes that with the individual sensors being independent, the optimal decision scheme that maximizes the probability of detection at the fusion for fixed false alarm probability consists of a Neyman-Pearson test at the fusion unit and the likelihood ratio test at the sensors. The threshold based fusion of combining multiple IDSs by fixing a certain number of false alarms is discussed in the work of Thomas and Balakrishnan [6]. This is a case of combining the top ranking outputs of each IDS after removing the duplicate alerts and setting the maximum acceptable false alarm rate.

The other somewhat related works albeit distantly are the alarm clustering method by Perdisci et al. [11], aggregation of alerts by Valdes et al. [12], combination of alerts into scenarios by Dain et al. [13], the alert correlation by Cuppens et al. [14], the correlation of Intrusion Symptoms with an application of chronicles by Morin et al. [15], and aggregation and correlation of intrusion-detection alerts by Debar et al. [16].

In the paper by Thomas and Balakrishnan [7], a sensor fusion architecture, which is data-dependent and different from the conventional fusion architecture is attempted. The focus of the present work is modeling the data-dependent decision fusion in an attempt to optimize both the fusion rule as well as the sensor rules.

III. Modeling the Data-dependent Decision Fusion System

This work is an extension of the data-dependent decision fusion approach proposed in our earlier work [7]. The paper presented here includes modeling with optimization done at every single stage, thereby arriving at an optimum architecture showing improved performance than what has been reported so far in literature. This architecture consists of three stages, optimizing the individual IDSs as the first stage, determining the weights of the individual IDSs with a neural network learner as the second stage, and performing the weighted aggregation with a fusion unit as the final stage.

A. Modeling the Intrusion Detection Systems

Consider an IDS that either monitors the network traffic connection on the network or the audit trails on the host. The network traffic connection or the audit trails monitored are given as $x \in X$, where X is the entire domain of network traffic features or the audit trails respectively. The model is based on the hypothesis that the security violations can be detected by monitoring the network for traffic connections of malicious intent in the case of network-based IDS and a system's audit records for abnormal patterns of system usage in the case of host-based IDS. The model is independent of any particular operating system, application, system vulnerability or type of intrusion, thereby providing a framework for a general-purpose IDS.

When making an attack detection, a connection pattern is given by $x_j \in \Re^{jk}$ where j is the number of features from k consecutive samples used as input to an IDS. As seen in the DARPA dataset, for many of the features the distributions are difficult to describe parametrically as they may be multi-modal or very heavy-tailed. These highly non-Gaussian distributions has led to investigate non-parametric statistical tests as a method of intrusion detection in the initial phase of IDS development. The detection of an attack in the event x is observed as an alert. In the case of network-based IDS, the elements of x can be the fields of the network traffic like the raw IP packets or the pre-processed basic attributes like the duration of a connection, the protocol type, service etc. or specific attributes selected with domain knowledge such as the number of failed logins or whether a superuser command was attempted. In host-based IDS, x can be the sequence of system calls, sequence of user commands, connection attempts to local host, proportion of accesses in terms of TCP or UDP packets to a given port of a machine over a fixed period of time etc. Thus IDS can be defined as a function that maps the data input into a normal or an attack event either by means of absence of an alert (0) or by the presence of an alert (1) respectively and is given by:

$$IDS : X \rightarrow \{0, 1\}.$$

To detect attacks in the incoming traffic, the IDSs are typically parameterized by a threshold T. The IDS uses a theoretical basis for deciding the thresholds for analyzing the network traffic to detect intrusions. Changing this threshold allows the change in performance of the IDS. If the threshold is very low, then the IDS tends to be very aggressive in detecting the traffic for intrusions. However, there is a potentially greater chance for the detections to be irrelevant which result in large false alarms. A large value of threshold on the other hand will have an opposite effect; being a bit conservative in detecting attacks. However, some potential attacks may get missed this way. Using a 3σ based statistical analysis, the higher threshold (T_h) is set at $+3\sigma$ and the lower threshold (T_l) is set at -3σ. This is with the assumption that the traffic signals are normally distributed. In general the traffic detection with s being the sensor output is given by:

$$Sensor\quad Detection = \begin{cases} attack, & T_l < s < T_h \\ normal, & s \leq T_l, \quad s \geq T_h \end{cases}$$

The signature-based IDS functions by looking at the event feature x and checking whether it matches with any of the records in the signature database D.

$$\begin{aligned} Signature\text{-}based\, IDS &: X \rightarrow \{1\} \qquad \forall x \in D, \\ &: X \rightarrow \{0\} \qquad \forall x \notin D. \end{aligned}$$

Anomaly-based IDS generates alarm when the input traffic deviates from the established models or profiles P.

$$\begin{aligned} Anomaly\text{-}based\, IDS &: X \rightarrow \{1\} \qquad \forall x \notin P, \\ &: X \rightarrow \{0\} \qquad \forall x \in P. \end{aligned}$$

B. Modeling the fusion IDS

Consider the case where n IDSs monitor a network for attack detection and each IDS makes a local decision s_i and these decisions are aggregated in the fusion unit f. This architecture is often referred to as the parallel decision fusion network and is shown in Figure 1. The fusion unit makes a global decision, y, about the true state of the hypothesis based on the collection of the local decisions gathered from all the sensors. The problem is casted as a binary detection problem with the hypothesis "$Attack$" or "$Normal$". Every IDS participating in the fusion has its own detection rate D_i, and false positive rate F_i, due to the preferred heterogeneity of the sensors in the fusion process. Each IDS indexed i gives an alert or no-alert indicated by s_i taking a value one or zero respectively, depending on the observation x.

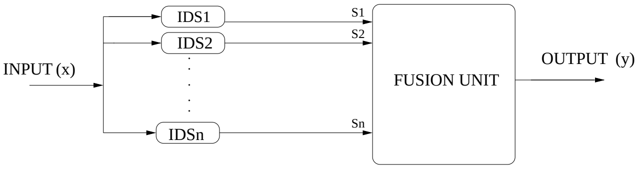

Fig. 1. Parallel Decision Fusion Network

$$s_i = \begin{cases} 0, & normal\; is\; declared\; to\; have\; been\; detected \\ 1, & attack\; is\; declared\; to\; have\; been\; detected \end{cases}$$

The fusion center collects these local decisions s_i and forms a binomial distribution y as given by y=$s = \sum_{i=1}^{n} s_i$, where n is the total number of IDSs taking part in fusion.

Theorem 1

The output of a binary fusion unit is decided by a function f given by:

$f : s_1$ x s_2..... x s_n x $x \Rightarrow \{0,1\}$, where the decisions of the individual detectors given by s_i are deterministic and the data x is a random parameter.

Lemma 1

The decision rule used by each of the individual detectors is deterministic and can be expressed as a function f_i given by:

$f_i : s_i \Rightarrow \{0,1\}$ defined as:

$$f_i(x_j) = \begin{cases} 0, & if\; p(s_i^j = 0|x_j) = 1 \\ 1, & otherwise \end{cases}$$

where j corresponds to the class of the network traffic on which the fusion rule as well as the respective sensor outputs depend. Since fusion center makes the final decision, the assumption is made that the output of the fusion rule is binary, i.e., either $Normal$ or $Attack$. It is the same case with all the individual IDSs: each IDS classifies the incoming traffic as $Normal$ or $Attack$.

C. Statement of the problem

The problem statement is defined in the following steps:

- The random variable x represents the observation to be made. This observation belongs to either of the two groups of the hypothesis: $Normal$ or $Attack$ with probabilities p or $q = 1 - p$, respectively.
- A set of n IDSs monitors the random variable x and detects the presence of attack in the traffic. The set of detections by the n sensors is given by $\{s_1, s_2.....s_n\}$, where s_i is the output of the IDS indexed i. Each s_i is a function of the input x, i.e., $s_i = f_i(x)$.
- The problem of optimum detection with n IDSs selecting either of the two possible hypotheses is considered from the decision theory point of view. The loss function ℓ is defined in terms of the decisions made by each IDS along with the observation and is given by:

  $\ell : \{0,1\}$ x $\{0,1\}$ x .....$\{0,1\}$ x $\{Normal, Attack\} \Rightarrow R$. Then the average of this loss is minimized. The objective of the decision strategy is to minimize the expected penalty (loss function) incurred as:

  $$min\, E[\ell(s_1, s_2.....s_n, H)], \tag{1}$$

  where H is the hypothesis and the minimization is over the decision rules of each detector.

- With $\ell(s_1, s_2.....s_n, H) = k$ being the cost incurred for the IDS_1 deciding s_1, IDS_2 deciding s_2, and so on. The minimum value of this cost function ℓ occurs when all the sensors make the correct decisions as $\ell(0,0,...0, Normal) = \ell(1,1,...1, Attack) = 0$ and increases to “1” if any one IDS only is incorrect and so on. Thus the cost function ℓ takes the maximum value of “n” when all the IDSs are unable to make the correct decision. This is a trivial case where all cases of n errors is penalized by the same amount and the function ℓ can be reduced by using affine transformations. From the cost matrix of the KDD IDS evaluations [20], it is clear that $\ell(0, s_1, s_2.....s_n, Attack) > \ell(1, s_1, s_2.....s_n, Normal)$, or it is more costly for any detector to miss an attack compared to a false alarm, regardless of the detection of other detectors. The minimization of the loss leads to sets of coupled inequalities in terms of the likelihood ratio of each IDS and the decisions made at the other sensors.

- As k decreases from 2 to 1, the thresholds would change in a way which increases the probability of error, as double errors are discounted to single ones. As k increases from 2, double errors become prohibitively expensive, so it is to be expected that some mechanism will emerge to reduce their likelihood. Thus, for k varying from 1 to n, there are n solutions to minimize equation 1, one of which being the global minimum and thus the optimal threshold pair.

D. The effect of setting threshold

To detect the attack in the incoming traffic, the IDSs are typically parameterized with a threshold, T. Changing this threshold allows the change in performance of the IDS. If the threshold is very large, some potentially dangerous attacks get missed. A small threshold on the other hand results in more detections, with a potentially greater chance that they are not relevant.

The final step in the approach towards solving of the fusion problem is taken by noticing that the decision function $f_i(.)$ is characterized by the threshold T_i and the likelihood ratio (if independence is assumed). Thus the necessary condition for optimal fusion decision occurs if the thresholds $(T_1, T_2, ...T_n)$ are chosen optimally. However, this does not satisfy the sufficient condition. These refer to the many local minima, each need to be checked to assure the global minimum.

The counterintuitive results at the individual sensors with the proper choice of thresholds will be advantageous in getting an optimum value for the fusion result. They are excellent paradigms for studying distributed decision architectures, to understand the impact of the limitations, and even suggest empirical experiments for IDS decisions.

The structure of the fusion rule plays a crucial role regarding the overall performance of the IDS since the fusion unit makes the final decision about the state of the environment. While a few inferior IDSs might not greatly impact the overall performance, a badly designed fusion rule can lead to a poor performance even if the local IDSs are well designed. The fusion IDS can be optimized by searching the space of fusion rules and optimizing the local thresholds for each candidate rule. Other than for some simple cases, the complexity of such an approach is prohibitive due to exponential growth of the set of possible fusion rules with respect to the number of IDSs. Searching for the fusion rule that leads to the minimum probability of error is the main bottleneck due to discrete nature of this optimization process and the exponentially large number of fusion rules. In our experiment we are trying to maximize the true positive rate by fixing the false positive rate at α_0. α_0 determines the threshold T by trial and error. We have noticed that within two or three trials in our case. This is done with the training data and hence it is done off line.

The computation of thresholds couples the choice of the local decision rules so that the system-wide performance is optimized, rather than the performance of the individual detector. This requirement is taken care of by the data-dependent decision fusion architecture in [7] and shown in Figure 2. This architecture has three-stages; the IDSs that produce the alerts as the first stage, the neural network supervised learner determining the weights to the IDSs' decisions depending on the input as the second stage, and then the fusion unit doing the weighted aggregation. The neural network learner can be considered as a pre-processing stage to the fusion unit. The neural network is most appropriate for weight determination, since it becomes difficult to define the rules clearly, mainly as more number of IDSs are added to the fusion unit. When a record is correctly classified by one or more detectors, the neural network will accumulate this knowledge as a weight and with more number of iterations, the weight gets stabilized. This information is used to fine-tune the fusion unit, since the fusion depends on the input feature vector. The fusion output is represented as:

$y = F_j(w_i^j(x_j, s_i^j), s_i^j)$,

where the weights w_i^j are dependent on both the input x_j as well as individual IDS's output s_i^j, where the prefix j refers to the class label and the suffix i refers to the IDS index. The fusion unit used gives a value of “1” or “0” depending on the set threshold being higher or lower than the weighted aggregation of the IDS's decisions. The fusion unit is optimized using this set up with the proper weighting to each one of the input to the fusion unit. The individual IDS are optimized by the proper choice of the threshold which is decided by the detection-false alarm trade-off. ROC curves are used to evaluate IDS performance over a range of trade-offs between detection rate and the false positive rate. Each IDS will have an operating point in the ROC curve and the optimum operating point is located at the relatively top-left point. The optimal decision fusion detection rule is obtained by forming the output of the fusion unit as: $y = s = \sum_{i=1}^{n} w_i^j s_i^j$. The architecture is independent of the data set and the structures employed, and

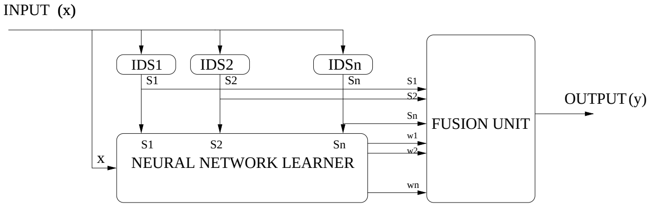

Fig. 2. Data-dependent Decision Fusion architecture

can be used with any real valued data set.

E. Modeling the Neural Network learner unit

The neural network unit in the data-dependent architecture is a supervised learning system which learns from a training data set. The training of the neural network unit by back propagation involves three stages: the feed forward of the output of all the IDSs along with the input training pattern, which collectively form the training pattern for the neural network learner unit, the calculation and the back propagation of the associated error, and the adjustments of the weights. After the training, the neural network is used for the computations of the feedforward phase. Learning can be defined over an input space X, an output space Y and a loss function ℓ. The training data can be specified as $\{(x_i, y_i)\}$, where $x_i \in X$ and $y_i \in Y$. The output is a hypothesis function $f_w : X \rightarrow Y$. f_w is chosen from a hypothesis space $\mathcal{F}$ to minimize the prediction error given by the loss function. The hypothesis function is that of the neural network and it represents the non-linear function from the input space X to the output space Y.

It is simple to assume stationarity by assuming the distribution of data points encountered in the future to be the same as the distribution of the training set. For simplicity, the DARPA data set is assumed to represent the real time traffic pattern distribution. Stationarity allows us to reduce the predictive learning problem to a minimization of the sum of the loss over the training set.

$$\begin{aligned} f_w^* &= argmin \sum \ell(f_w(x_i), y_i) \\ s.t \quad & f_w \in \mathcal{F} \ \& \ \ (x_i, y_i) \in s. \end{aligned} \tag{2}$$

Loss functions are typically defined to be non-negative over all inputs and zero when $f_w(x_i) = y_i$.

F. Dependence on the data and the individual IDSs

Often, the data in the databases is only an approximation of the true data. When the information about the goodness of the approximation is recorded, the results obtained from the database can be interpreted more reliably. Any database is associated with a degree of accuracy, which is denoted with a probability density function, whose mean is the value itself.

In order to maximize the detection rate it is necessary to fix the false alarm rate to an acceptable value, taking into account the trade-off between the detection rate and the false alarm rate. The threshold (T) that maximizes the TP_{rate} and thus minimizes the FN_{rate} is given as:

$$\begin{aligned} & FP_{rate} = P[alert|normal] \\ & = P\left[\sum_{i=1}^{n} w_i s_i \geq T \,|normal \right] = \alpha_0, \end{aligned} \tag{3}$$

$$\begin{aligned} & TP_{rate} = P[alert|attack] \\ & = P\left[\sum_{i=1}^{n} w_i s_i \geq T \,|attack \right]. \end{aligned} \tag{4}$$

The fusion of IDSs becomes meaningful only when $FP \leq FP_i \quad \forall\ i$ and $TP \geq TP_i \quad \forall\ i$; where FP and TP correspond to the false positives and the true positives of the fused IDS and FP_i and TP_i correspond to the false positives and the true positives of the individual IDS indexed i. It is required to provide low value of weight to any individual sensor that is unreliable, hence meeting the constraint on false alarm as given in equation 3. Similarly, the fusion improves the TP_{rate} as the detectors get weighted according to their performance.

G. Threshold Optimization

Tenney and Sandell in their work [19] establish the optimum strategy that minimizes a global cost in the case where the *a priori* probabilities of the hypotheses, the distribution functions of the local observations, the cost functions, and the fusion rule are given. They concluded that each local detector is optimally a likelihood ratio detector but that the computation of the optimum thresholds for these local detectors is complicated due to cross coupling.

The global optimization criterion for a distributed detection system would encompass local decision statistics, local decision thresholds, the fusion center decision statistic, and the fusion center decision threshold.

For each input traffic observation x, the set of n local thresholds should be optimized with respect to the probability of error. With a fusion rule given by a function f, the average probability of error at the fusion unit is given by the weighted sum of false positive and false negative errors.

$$P_e(T,\, f) = p{*}P(s = 1|Normal) + \ q{*}P(s = 0|Attack), \tag{5}$$

where p and q are the respective weights of false positive and false negative errors.

Assuming independence between the local detectors, the likelihood ratio is given by:

$$\frac{P(s|Attack)}{P(s|Normal)} = \frac{P(s_1, s_2, ..., s_N|Attack)}{P(s_1, s_2, ..., s_N|Normal)}$$
$$= \prod_{i=1}^{n} \frac{P(s_i|Attack)}{P(s_i|Normal)}.$$

The optimum decision rule for the fusion unit follows:

$$f(s) = log \frac{P(s|Attack)}{P(s|Normal)}$$

Depending on the value of $f(s)$ being greater than or equal to the decision threshold, T, or less than the decision threshold, T, the decision is made for the hypothesis as "*Attack*" or "*Normal*" respectively. Thus the decisions from the n detectors are coupled through a cost function. It is shown that the optimal decision is characterized by thresholds as in the decoupled case. As far as the optimum criterion is concerned, the first step is to minimize the loss function of equation 1. This leads to sets of simultaneous inequalities in terms of the generalized likelihood ratios at each detector, the solutions of which determine the regions of optimum detection.

H. Fixing threshold bounds at the fusion unit

Instead of using a fixed threshold to determine the intrusion, the problem can be solved by using more flexible bounds and adjusting the bounds for optimum detection-false alarm trade-off taking into account the stochastic nature of network traffic.

Let D and F denote the unanimous detection rate and the false positive rate respectively. The mean and the variance of the traffic patterns can be used to compute the probability of anomaly assuming the normal distribution. This refers to the amount of anomalous traffic quantitatively. The mean and variance of s in case of attack and no-attack, are given by the following equations:

$$E[s|alert] \; = \; \sum_{i=1}^{n} D_i, \qquad Var[s|alert] \; = \; \sum_{i=1}^{n} D_i(1-D_i);$$

in case of attack,

$$E[s|alert] \; = \; \sum_{i=1}^{n} F_i, \qquad Var[s|alert] \; = \; \sum_{i=1}^{n} F_i(1-F_i);$$

in case of no-attack.

The fusion IDS is required to give a high detection rate and a low false positive rate. Hence the threshold T needs to be chosen well above the mean of the false alerts and well below the mean of the true alerts. Consequently, the threshold bounds are given as:

$$\sum_{i=1}^{n} F_i < T < \sum_{i=1}^{n} D_i.$$

The detection rate and the false positive rate of the fusion IDS is desired to surpass the corresponding weighted averages and hence:

$$D > \frac{\sum_{i=1}^{n} D_i^2}{\sum_{i=1}^{n} D_i} \tag{6}$$

and

$$F < \frac{\sum_{i=1}^{n} (1-F_i)F_i}{\sum_{i=1}^{n} (1-F_i)}. \tag{7}$$

Now, using simple range comparison,

$$D = P\{s \geq T|attack\}$$

$$= P\left\{ |s - \sum_{i=1}^{n} D_i| \; \leq \; \left(\sum_{i=1}^{n} D_i - T \right) |attack \right\}.$$

With the mean and variance known, a new observation is defined to be abnormal if it falls outside a confidence interval that is dependent on the standard deviation. This modeling requires no prior knowledge about normal activity in order to set the limits; instead, it learns what constitutes normal activity from its observations, and the confidence intervals automatically reflect this increased knowledge. Also, the confidence interval depends on the observed data. Using the Chebyshev inequality on the random variable s with

$mean = E[s] = \sum_{i=1}^{n} D_i$ and

$variance = Var[s] = \sum_{i=1}^{n} D_i(1-D_i),$

$P\left\{|s - E(s)| \geq k\right\} \; \leq \; \frac{Var(s)}{k^2}$

With the assumption that the threshold T is greater than the mean of normal activity,

$$P\left\{ |s - \sum_{i=1}^{n} D_i| \; \leq \; \left(\sum_{i=1}^{n} D_i - T \right) |attack \right\} \quad \geq$$

$$1 \; - \; \frac{\sum_{i=1}^{n} D_i(1-D_i)}{\left(\sum_{i=1}^{n} D_{i-T} \right)^2}$$

From (6) it follows that :

$$1 \; - \; \frac{\sum_{i=1}^{n} D_i(1-D_i)}{\left(\sum_{i=1}^{n} D_{i-T} \right)^2} \; \geq \; \frac{\sum_{i=1}^{n} D_i^2}{\sum_{i=1}^{n} D_i}$$

The upper bound of T is derived from the above equation as:

$$T \quad \leq \quad \sum_{i=1}^{n} D_i - \sqrt{\sum_{i=1}^{n} D_i}$$

Similarly, for the false positive rate, $F \; = \; P\{s \; \geq \; T \,|\, no\text{-}attack\}$, in order to derive the lower bound of T,

From (7) it follows that :

$$\frac{\sum_{i=1}^{n} F_i(1-F_i)}{\left(T - \sum_{i=1}^{n} F_i \right)^2} \; \leq \; \frac{\sum_{i=1}^{n} F_i(1-F_i)}{\sum_{i=1}^{n} (1-F_i)}$$

The lower bound of T is derived from the above equation as:

$$T \quad \geq \quad \sum_{i=1}^{n} F_i + \sqrt{\sum_{i=1}^{n} (1-F_i)}$$

The threshold bounds for the fusion IDS is:

$$\left[\; \begin{matrix} n \\ \sum F_i \\ i=1 \end{matrix} \; + \sqrt{\sum_{i=1}^{n} (1-F_i)} \; , \quad \sum_{i=1}^{n} D_i - \sqrt{\sum_{i=1}^{n} D_i} \; \right].$$

It is necessary to have the minimum required number of IDSs in the fusion architecture so that the bound exists, i.e., the upper bound is ensured to be larger than the lower bound as:

$$\sum_{i=1}^{n} D_i - \sqrt{\sum_{i=1}^{n} D_i} \geq \sum_{i=1}^{n} F_i + \sqrt{\sum_{i=1}^{n} (1 - F_i)}$$

This condition poses a restriction on the individual detection rates, individual false alarm rates and the number of individual IDSs used in the fusion process.

Since the threshold T is assumed to be greater than the mean of normal activity, the upper bound of false positive rate F can be obtained from the Chebyshev inequality as:

$$F \leq \frac{Var[s]}{(T - E[s])^2} \tag{8}$$

In order to reduce the false positive rate, it is important to reduce the variance of the normal traffic. In the ideal case with normal traffic the variance is zero. From equation 8 it can be seen that as the variance of the normal traffic approaches zero, the false positive rate also approaches zero. Also, since the threshold T is assumed to be less than the mean of the intrusive activity, the lower bound of the detection rate D can be obtained from the Chebyshev inequality as:

$$D \geq 1 - \frac{Var[s]}{(E[s] - T)^2} \tag{9}$$

For an intrusive traffic, the factor $D_i(1 - D_i)$ remains almost steady and hence the value of Variance $=\sum_{i=1}^{n} D_i(1 - D_i)$ is an appreciable value. Since the variance of the attack traffic is above a certain detectable minimum, from equation 9 it is seen that the correct detection rate can approach an appreciably high value. Similarly the true negatives will also approach a high value since the false positive rate is reduced with IDS fusion.

IV. Results and Discussion

A. Test set up

The test set up for the experimental evaluation consisted of three Pentium machines with Linux operating system. For a good protection, a combination of shallow and deep sensors is necessary in many systems. Hence, for the purpose of fusion we have incorporated two sensors, one that monitors the header of the traffic packet and the other that monitors the packet content. The experiments were conducted with the simulated IDSs PHAD and ALAD [21]. This choice of heterogeneous sensors in terms of their functionality is to exploit the advantages of fusion IDS [2]. In addition, complementary IDSs provide versatility and similar IDSs ensure reliability. The PHAD being packet-header based and detecting one packet at a time, is totally unable to detect the slow scans. However, PHAD detects the stealthy scans much more effectively. The ALAD being content-based will complement the PHAD by detecting R2L(Remote to Local) and U2R(User to Root) attacks with appreciable efficiency.

The weight analysis of the IDS data coming from PHAD and ALAD was carried out by the neural network supervised learner before it was fed to the fusion element. The detectors PHAD and ALAD produces the IP address along with the anomaly score of the alert. The alerts produced by these IDSs are converted to a standard binary form. The neural network learner inputs these decisions along with the particular traffic input which was monitored by the IDSs. The Internet Engineering Task Force Intrusion Detection working group's Intrusion Detection Message Exchange Format (IDMEF), which enable different types of IDSs to generate the events by using unified language can be used instead.

The neural network learner was designed as a feed forward back propagation algorithm with a single hidden layer and 25 sigmoidal hidden units in the hidden layer. Experimental proof is available for the best performance of the Neural Network with the number of hidden units being $log(T)$, where T is the number of training samples in the data set [22]. In order to train the neural network, it is necessary to expose them to both normal and anomalous data. Hence, during the training, the network was exposed to weeks 1, 2, and 3 of the training data and the weights were adjusted using the back propagation algorithm. An epoch of training consisted of one pass over the training data. The training proceeded until the total error made during each epoch stopped decreasing or 1000 epochs had been reached.

The fusion element analyzes the IDS data coming from PHAD and ALAD distributed across the single subnet and observing the same domain. The fusion unit performed the weighted aggregation of the IDS outputs for the purpose of identifying the attacks in the test data set. It used binary fusion by giving an output value of one or zero depending on the value of the weighted aggregation of the various IDS decisions. The packets were identified by their timestamp on aggregation. A value of one at the output of the fusion unit indicated the record to be under attack and a zero indicated the absence of an attack.

B. Data set

The MIT Lincoln Laboratory under DARPA and AFRL sponsorship, has collected and distributed the first standard corpora for evaluation of computer network intrusion detection systems. The network traffic including the entire payload of each packet was recorded in tcpdump format and provided for evaluation. This MIT-DARPA data set (IDEVAL 1999) [23] was used to train and test the performance of Intrusion Detection Systems. The data for the weeks one and three were used for the training of the anomaly detectors and the weeks four and five were used as the test data. The training of the neural network learner was performed on the training data for weeks one, two and three, after the individual IDSs were trained. Each of the IDS was trained on distinct portions of the training data (ALAD on week one and PHAD on week three), which is expected to provide independence among the IDSs and also to develop diversity while being trained.

It is important to mention at this point that the proposed architecture can be generalized beyond the data set or the IDSs that are used in fusion. Even with the criticisms by McHugh [24] and Mahoney and Chan [25] against the DARPA dataset, the dataset was extremely useful in the IDS evaluation undertaken in this work. Since none of the IDSs perform exceptionally well on the DARPA dataset, the aim is to show that the performance improves with the proposed method. If a system is evaluated on the DARPA dataset, then it cannot claim anything more in terms of its performance on the real network traffic. Hence this dataset can be considered as the base line of any research [26]. Also, even after nine years of its generation, there are still a lot of relevant attacks in the data set for which signatures are not available in database of even the frequently updated signature based IDSs.

C. Evaluation metrics

Let TP be the number of attacks that are correctly detected, FN be the number of attacks that are not detected, TN be the number of normal traffic packet/connections that are correctly classified, and FP be the number of normal traffic packet/connections that are incorrectly detected as attack. In the case of an IDS, there are both the security requirements and the usability requirements. The security requirement is determined by the $TPrate$ and the usability requirement is decided by the number of FPs because of the low base rate in the case of a network traffic.

The commonly used IDS evaluation metrics on a test data are the overall accuracy and F-score.

$$Overall\ Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Overall Accuracy is not a good metric for comparison in the case of network traffic data since the true negatives abound.

Precision is a measure of what fraction of test data detected as attack are actually from the attack classes.

$$Precision = \frac{TP}{TP+FP}$$

Recall is a measure of what fraction of attack class was correctly detected.

$$Recall = \frac{TP}{TP+FN}$$

There is a trade-off between the two metrics precision and recall. As the number of detections increase by lowering of the threshold, the recall will increase, while precision is expected to decrease. The recall-precision characterization of a particular IDS is normally used to analyze the relative and absolute performance of an IDS over a range of operating conditions. F-score, which is the harmonic mean of recall(R) and precision(P), scores the balance between precision and recall. The F-score is given by:

$$F\text{-}score = \frac{2*P*R}{P+R}$$

The standard measures, namely precision, recall, and F-score are grounded on a probabilistic framework and hence allows one to take into account the intrinsic variability of performance estimation. The comparison of IDSs with the metric F-score has the limitation in directly applying tests of significance to it in order to determine the confidence level of the comparison. The primary goal was to achieve improvement in recall as well as precision, and hence P-test [27] is used which account for the improvement in both precision and recall.

To compare two IDS X and Y, let (R_X, P_X) and (R_Y, P_Y) be the values of recall and precision w.r.t attack respectively. Let IDS X and Y predict N_X^{Pos} and N_Y^{Pos} positives respectively and N^{Pos} be the total number of positives in the test sample. Then the P-test is applied as follows:

$$Z_R = \frac{R_X - R_Y}{\sqrt{2R(1-R)/N^{Pos}}}$$

$$Z_P = \frac{P_X - P_Y}{\sqrt{2P(1-P)(1/N_X^{Pos} + 1/N_Y^{Pos})}}$$

where $R = \frac{R_X + R_Y}{2}$ and $P = \frac{N_X^{Pos} P_X + N_Y^{Pos} P_Y}{N_X^{Pos} + N_Y^{Pos}}$

If $Z_R \geq 1.96$, then R_X can be regarded as being significantly better than R_Y at the 95% confidence level.

If $Z_R \leq -1.96$, then R_X can be regarded as being significantly poorer than R_Y at the 95% confidence level.

If $|Z_R| \leq 1.96$, then R_X can be regarded as being comparable to R_Y.

Similar tests are applied to compare P_X and P_Y.

Now, to compare the two IDSs X and Y;
IDS X is better than IDS Y if either of the following criteria is satisfied:

$R_X \gg R_Y$ and $P_X \sim P_Y$
$R_X \gg R_Y$ and $P_X \gg P_Y$
$R_X \sim R_Y$ and $P_X \gg P_Y$

$R_X \sim R_Y$ and $P_X \sim P_Y$, then $X \sim Y$.

It may so happen that one metric is significantly better and the other metric is significantly worse. In such cases of conflict, the non-probabilistic metric F-score can be used instead of applying the significance test.

D. Experimental evaluation

All the IDSs that form part of the fusion IDS were first evaluated separately with the same data set, and then the combined fusion IDS was evaluated. The experimental results of Table I and Table II clearly show that none of the individual IDS was able to provide an acceptable value for

• TABLE I
ATTACKS OF EACH TYPE DETECTED BY PHAD AT 0.00002 FP RATE

Attack type	Total attacks	Attacks detected	% detection
Probe	37	26	70%
DoS	63	27	43%
R2L	53	6	11%
U2R/Data	37	4	11%
Total	190	63	33%

TABLE II
ATTACKS OF EACH TYPE DETECTED BY ALAD AT 0.00002 FP RATE

Attack type	Total attacks	Attacks detected	% detection
Probe	37	9	24%
DoS	63	23	37%
R2L	53	31	59%
U2R/Data	37	15	41%
Total	190	78	41%

TABLE V
F-SCORE AND OVERALL ACCURACY FOR DIFFERENT CHOICE OF FALSE POSITIVES FOR ALAD

FP	TP	Precision	Recall	Overall Accuracy	F-score
50	52	0.29	0.27	0.99	0.28
100	78	0.44	0.41	0.99	0.42
200	86	0.30	0.45	0.99	0.36
500	90	0.15	0.47	0.99	0.23

TABLE VI
F-SCORE AND OVERALL ACCURACY FOR DIFFERENT CHOICE OF FALSE POSITIVES FOR FUSION IDS

FP	TP	Precision	Recall	Overall Accuracy	F-score
50	58	0.54	0.31	0.99	0.39
100	93	0.48	0.49	0.99	0.49
200	122	0.38	0.64	0.99	0.48
500	135	0.21	0.71	0.99	0.32

all the performance measures.

The analysis of PHAD and ALAD have resulted in a clear understanding of the individual IDSs expected to succeed or fail under a particular attack. The combination of the two sensor alerts provide an improved rate of detection as shown in Table III.

In each of the individual IDSs, the number of detections were observed at false positives of 50, 100, 200 and 500, when trained on inside week 1 and week 3 along with additional normal traffic collected from our University traffic, and tested on weeks 4 and 5.

The performance in terms of F-score of PHAD, ALAD and the combination of PHAD and ALAD is shown in the Table IV, V and VI for various values of false positive by setting the threshold appropriately. The improved performance of the combination of the alarms from each system can be observed in Table VI, corresponding to the false positives between 100 and 200, by fixing the threshold bounds appropriately. Thus the combination works best above a false positive of 100 and much below a false positive of 200. The metric F-score reaches a comparably better value of 0.49 for the fusion IDS from the scores of 0.36 and 0.42 for PHAD and ALAD respectively.

The ROC curves of Figure 3 show the improved performance of the Data-dependent Decision fusion IDS compared to the two IDSs PHAD and ALAD. The Table VII shows the result of P-test. The DD fusion method performs significantly better than PHAD and comparable to ALAD according to the significance test.

TABLE III
ATTACKS OF EACH TYPE DETECTED BY FUSION IDS AT 0.00004 FP RATE

Attack type	Total attacks	Attacks detected	% detection
Probe	37	30	81%
DOS	63	42	67%
R2L	53	32	60%
U2R/Data	37	18	49%
Total	190	122	64%

TABLE IV
F-SCORE AND OVERALL ACCURACY FOR DIFFERENT CHOICE OF FALSE POSITIVES FOR PHAD

FP	TP	Precision	Recall	Overall Accuracy	F-score
50	40	0.44	0.21	0.99	0.28
100	63	0.39	0.33	0.99	0.36
200	63	0.24	0.33	0.99	0.28
500	63	0.11	0.33	0.99	0.17

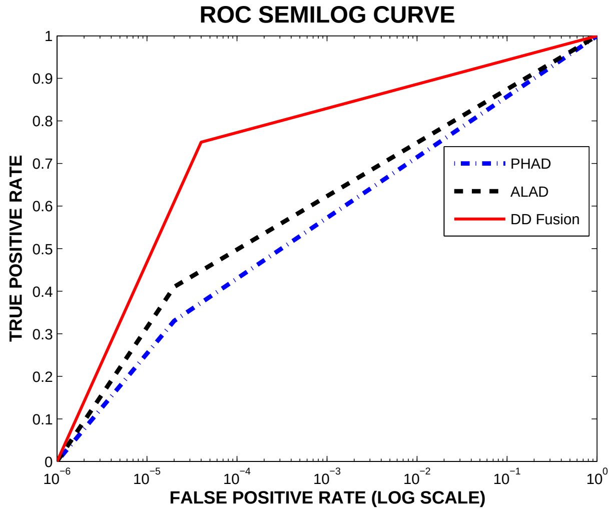

Fig. 3. ROC curve of PHAD, ALAD and fusion IDS

TABLE VII
PERFORMANCE COMPARISON OF INDIVIDUAL IDSS AND THE DATA-DEPENDENT FUSION METHOD

Detector pairs/ Z-number	DD Fusion and PHAD	DD Fusion and ALAD
Z_R	3.17	1.6
Z_P	1.33	0.61

V. CONCLUSION

The sensor fusion techniques works effectively by gathering complementary information that can improve the overall detection rate without adversely affecting the false alarm rate. Simple theoretical model is initially illustrated in this paper for the purpose of showing the improved performance of IDS using sensor fusion. The detection rate and the false positive rate quantify the performance benefit obtained through the fixing of threshold bounds. The theoretical proof was supplemented with an experimental evaluation, and the detection rates, false positive rates, and F-score were measured. Also the significance test (P test) was carried out for showing the improved performance of sensor fusion in IDS. In order to understand the importance of setting a threshold, the anomaly-based IDSs, PHAD and ALAD have been individually analyzed. The experimental results obtained prove the correctness of the theoretical analysis made in this work.

ACKNOWLEDGMENT

The authors would like to thank the unknown reviewers for their constructive suggestions, which in turn improved the quality of the paper.

REFERENCES

[1] C. Siaterlis and B. Maglaris, Towards Multisensor Data Fusion for DoS detection, ACM Symposium on Applied Computing, 2004

[2] T. Bass, Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems, IRIS National Symposium, 1999

[3] G. Giacinto, F. Roli, and L. Didaci, Fusion of multiple classifiers for intrusion detection in computer networks, Pattern recognition letters, 2003

[4] L. Didaci, G. Giacinto, and F. Roli, Intrusion detection in computer networks by multiple classifiers systems, International Conference on Pattern recognition, 2002

[5] Y. Wang, H. Yang, X. Wang, and R. Zhang, Distributed intrusion detection system based on data fusion method, Intelligent control and automation, WCICA 2004

[6] C. Thomas and N. Balakrishnan, Selection of Intrusion Detection Threshold bounds for effective sensor fusion, Proceedings of SPIE International Symposium on Defense ans Security, vol.6570, Apr 2007

[7] C. Thomas and N. Balakrishnan, Advanced Sensor Fusion Technique for Enhanced Intrusion Detection, IEEE International Conference on Intelligence and Security Informatics, Jun. 2008

[8] W. Hu, J. Li, and Q. Gao, Intrusion Detection Engine on Dempster-Shafer's Theory of Evidence, Proceedings of International Conference on Communications , Circuits and Systems, vol.3, pp. 1627-1631, Jun 2006

[9] A. Siraj, R.B. Vaughn, and S.M. Bridges, Intrusion Sensor Data Fusion in an Intelligent Intrusion Detection System Architecture, Proceedings of the 37th Hawaii international Conference on System Sciences, 2004

[10] S.C.A. Thomopolous, R. Viswanathan, and D.K. Bougoulias, Optimal distributed decision fusion, IEEE Transactions on aerospace and electronic systems, vol. 25, No. 5, Sep. 1989

[11] R. Perdisci, G. Giacinto, and F. Roli, Alarm clustering for intrusion detection systems in computer networks, Engg. applications of Artificial intelligence, Elsevier publications, March 2006

[12] A. Valdes and K. Skinner, Probabilistic alert correlation, Springer Verlag Lecture notes in Computer Science, 2001

[13] O.M. Dain and R.K. Cunningham, Building Scenarios from a Heterogeneous Alert Stream, IEEE Workshop on Information Assurance and Security, 2001

[14] F. Cuppens and A. Miege, Alert correlation in a cooperative intrusion detection framework, Proceedings of the 2002 IEEE symposium on security and privacy, 2002

[15] B. Morin and H. Debar, Correlation of Intrusion Symptoms : an Application of Chronicles, RAID 2003

[16] H. Debar and A. Wespi, Aggregation and Correlation of Intrusion-Detection Alerts, RAID 2001

[17] M.V. Mahoney and P.K. Chan, Detecting Novel attacks by identifying anomalous Network Packet Headers, Florida Institute of Technology Technical Report CS-2001-2

[18] M.V. Mahoney and P.K. Chan, Learning non stationary models of normal network traffic for detecting novel attacks, SIGKDD, 2002

[19] R.R. Tenney and N.R. Sandell, Detection with distributed sensors, IEEE Transactions on Aerospace and Electronic Systems, vol. 17, No.4, Jul 1981

[20] C. Elkan, Results of the KDD'99 classifier learning, SIGKDD Explorations, Vol. 1, Issue 2, pp.62-63, Jan 2000

[21] M.V. Mahoney, A Machine Learning approach to detecting attacks by identifying anomalies in network traffic, PhD Dissertation, Florida Institute of Technology, 2003

[22] R.P. Lippmann, An introduction to computing with Neural Nets, IEEE ASSP Magazine, Vol.4, pp. 4-22, April 1987

[23] DARPA Intrusion Detection Evaluation Data Set, http://www.ll.mit.edu/IST/ideval/data/data_index.html

[24] J. McHugh, Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA IDS evaluations as performed by Lincoln Laboratory, ACM Transactions on Information and System Security, vol.3, No.4, Nov. 2000

[25] M.V. Mahoney and P.K. Chan, An analysis of the 1999 DARPA /Lincoln Laboratory evaluation data for network anomaly detection, Technical Report CS-2003-02

[26] C. Thomas and N. Balakrishnan, Usefulness of DARPA data set in Intrusion Detection System evaluation, Proceedings of SPIE International Defense and Security Symposium, 2008

[27] M.V. Joshi, On evaluating performance of classifiers for rare classes, Proceedings of the 2002 IEEE International Conference on data mining, pp. 641-644, 2002

Ciza Thomas took her B.Tech and M.Tech in Electronics and Communication Engineering from College of Engineering, Trivandrum, India. She is currently pursuing for Doctoral Degree in the Supercomputer Education and Research Centre of the Indian Institute of Science, Bangalore, India.

N. Balakrishnan took his B.Tech in Electronics Engineering from the Madras University, India and PhD from the Indian Institute of Science, Bangalore, India. His topics of interest include Signal Processing, Numerical Electromagnetics, and Network Security. He is the Associate Director of the Indian Institute of Science, Bangalore, India.