10

Mathematical Basis of Sensor Fusion in Intrusion Detection Systems

Ciza Thomas
Assistant Professor, College of Engineering, Trivandrum
India

Balakrishnan Narayanaswamy
Associate Director, Indian Institute of Science, Bangalore
India

1. Introduction

Intrusion Detection Systems (IDS) gather information from a computer or a network, and analyze this information to identify possible security breaches against the system or the network. The network traffic (with embedded attacks) is often complex because of multiple communication modes with deformable nature of user traits, evasion of attack detection and network monitoring tools, changes in users' and attackers' behavior with time, and sophistication of the attacker's attempts in order to avoid detection. This affects the accuracy and the reliability of any IDS. An observation of various IDSs available in the literature shows distinct preferences for detecting a certain class of attack with improved accuracy, while performing moderately on other classes. The availability of enormous computing power has made it possible for developing and implementing IDSs of different types on the same network. With the advances in sensor fusion, it has become possible to obtain a more reliable and accurate decision for a wider class of attacks, by combining the decisions of multiple IDSs.

Clearly, sensor fusion for performance enhancement of IDSs requires very complex observations, combinations of decisions and inferences via scenarios and models. Although, fusion in the context of enhancing the intrusion detection performance has been discussed earlier in literature, there is still a lack of theoretical analysis and understanding, particularly with respect to correlation of detector decisions. The theoretical study to justify why and how the sensor fusion algorithms work, when one combines the decisions from multiple detectors has been undertaken in this chapter. With a precise understanding as to why, when, and how particular sensor fusion methods can be applied successfully, progress can be made towards a powerful new tool for intrusion detection: the ability to automatically exploit the strengths and weaknesses of different IDSs. The issue of performance enhancement using sensor fusion is therefore a topic of great draw and depth, offering wide-ranging implications and a fascinating community of researchers to work within.

The mathematical basis for sensor fusion that provides enough support for the acceptability of sensor fusion in performance enhancement of IDSs is introduced in this chapter. This chapter justifies the novelties and the supporting proof for the Data-dependent Decision (DD) fusion architecture using sensor fusion. The neural network learner unit of the Data-dependent Decision fusion architecture aids in improved intrusion detection sensitivity and false alarm reduction. The theoretical model is undertaken, initially without any knowledge of the available detectors or the monitoring data. The empirical evaluation to augment the mathematical analysis is illustrated using the DARPA data set as well as the real-world network taffic. The experimental results confirm the analytical findings in this chapter.

2. Related Work

Krogh & Vedelsby (1995) prove that at a single data point the quadratic error of the ensemble estimator is guaranteed to be less than or equal to the average quadratic error of the component estimators. Hall & McMullen (2000) state that if the tactical rules of detection require that a particular certainty threshold must be exceeded for attack detection, then the fused decision result provides an added detection up to 25% greater than the detection at which any individual IDS alone exceeds the threshold. This added detection equates to increased tactical options and to an improved probability of true negatives Hall & McMullen (2000). Another attempt to illustrate the quantitative benefit of sensor fusion is provided by Nahin & Pokoski (1980). Their work demonstrates the benefits of multisensor fusion and their results also provide some conceptual rules of thumb.

Chair & Varshney (1986) present an optimal data fusion structure for distributed sensor network, which minimizes the cumulative average risk. The structure weights the individual decision depending on the reliability of the sensor. The weights are functions of probability of false alarm and the probability of detection. The maximum *a posteriori* (MAP) test or the Likelihood Ratio (L-R) test requires either exact knowledge of the *a priori* probabilities of the tested hypotheses or the assumption that all the hypotheses are equally likely. This limitation is overcome in the work of Thomopoulos et al. (1987). Thomopoulos et al. (1987) use the Neyman-Pearson test to derive an optimal decision fusion. Baek & Bommareddy (1995) present optimal decision rules for problems involving n distributed sensors and m target classes.

Aalo & Viswanathan (1995) perform numerical simulations of the correlation problems to study the effect of error correlation on the performance of a distributed detection systems. The system performance is shown to deteriorate when the correlation between the sensor errors is positive and increasing, while the performance improves considerably when the correlation is negative and increasing. Drakopoulos & Lee (1995) derive an optimum fusion rule for the Neyman-Pearson criterion, and uses simulation to study its performance for a specific type of correlation matrix. Kam et al. (1995) considers the case in which the class-conditioned sensor-to-sensor correlation coefficient are known, and expresses the result in compact form. Their approach is a generalization of the method adopted by Chair & Varshney (1986) for solving the data fusion problem for fixed binary local detectors with statistically independent decisions. Kam et al. (1995) uses Bahadur-Lazarsfeld expansion of the probability density functions. Blum et al. (1995) study the problem of locally most powerful detection for correlated local decisions.

The next section attempts a theoretical modeling of sensor fusion applied to intrusion detection, with little or no knowledge regarding the detectors or the network traffic.

3. Theoretical Analysis

The choice of when to perform the fusion depends on the types of sensor data available and the types of preprocessing performed by the sensors. The fusion can occur at the various levels like, 1) input data level prior to feature extraction, 2) feature vector level prior to identity declaration, and 3) decision level after each sensor has made an independent declaration of identity.

Sensor fusion is expected to result in both qualitative and quantitative benefits for the intrusion detection application. The primary aim of sensor fusion is to detect the intrusion and to make reliable inferences, which may not be possible from a single sensor alone. The particular quantitative improvement in estimation that results from using multiple IDSs depends on the performance of the specific IDSs involved, namely the observational accuracy. Thus the fused estimate takes advantage of the relative strengths of each IDS, resulting in an improved estimate of the intrusion detection. The error analysis techniques also provide a means for determining the specific quantitative benefits of sensor fusion in the case of intrusion detection. The quantitative benefits discover the phenomena that are likely rather than merely chance of occurrences.

3.1 Mathematical Model

A system of n sensors $IDS_1, IDS_2, ..., IDS_n$ is considered; corresponding to an observation with parameter x; $x \in \Re^m$. Consider the sensor IDS_i to yield an output s^i; $s^i \in \Re^m$ according to an unknown probability distribution p_i. The decision of the individual IDSs that take part in fusion is expected to be dependent on the input and hence the output of IDS_i in response to the input x_j can be written more specifically as s^i_j. A successful operation of a multiple sensor system critically depends on the methods that combine the outputs of the sensors, where the errors introduced by various individual sensors are unknown and not controllable. With such a fusion system available, the fusion rule for the system has to be obtained. The problem is to estimate a fusion rule $f : \Re^{nm} \rightarrow \Re^m$, independent of the sample or the individual detectors that take part in fusion, such that the expected square error is minimized over a family of fusion rules.

To perform the theoretical analysis, it is necessary to model the process under consideration. Consider a simple fusion architecture as given in Fig. 1 with n individual IDSs combined by means of a fusion unit. To start with, consider a two dimensional problem with the detectors responding in a binary manner. Each of the local detector collects an observation $x_j \in \Re^m$ and transforms it to a local decision $s^i_j \in \{0, 1\}, i = 1, 2, ..., n$, where the decision is 0 when the traffic is detected normal or else 1. Thus s^i_j is the response of the ith detector to the network connection belonging to class $j = \{0, 1\}$, where the classes correspond to normal traffic and the attack traffic respectively. These local decisions s^i_j are fed to the fusion unit to produce an unanimous decision $y = s_j$, which is supposed to minimize the overall cost of misclassification and improve the overall detection rate.

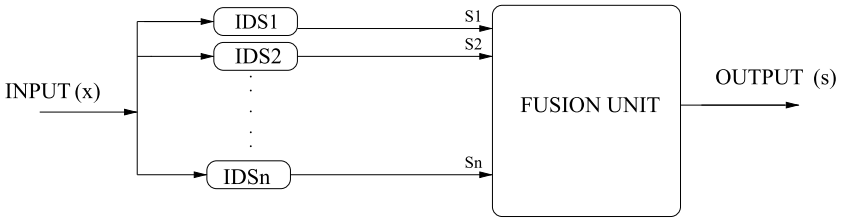


Fig. 1. Fusion architecture with decisions from n IDSs

The fundamental problem of network intrusion detection can be viewed as a detection task to decide whether network connection x is a normal one or an attack. Assume a set of unknown features $e = \{e_1, e_2, ..., e_m\}$ that are used to characterize the network traffic. The feature extractor is given by $e_e(x) \subset e$. It is assumed that this observed variable has a deterministic component and a random component and that their relation is additive. The deterministic component is due to the fact that the class is discrete in nature, i.e., during detection, it is known that the connection in either normal or an attack. The imprecise component is due to some random processes which in turn affects the quality of extracted features. Indeed, it has a distribution governed by the extracted feature set often in a nonlinear way. By ignoring the source of distortion in extracted network features $e_e(x)$, it is assumed that the noise component is random (while in fact it may not be the case when all possible variations can be systematically incorporated into the base-expert model).

In a statistical framework, the probability that x is identified as normal or as attack after a detector s_θ observes the network connection can be written as:

$$s^i = s_{\theta_i}(e_e(x)) \tag{1}$$

where x is the sniffed network traffic, e_e is a feature extractor, and θ_i is a set of parameters associated to the detector indexed i. There exists several types of intrusion detectors, all of which can be represented by the above equation.

Sensor fusion results in the combination of data from sensors competent on partially overlapping frames. The output of a fusion system is characterized by a variable s, which is a function of uncertain variables $s^1, ..., s^n$, being the output of the individual IDSs and given as:

$$s = f(s^1, ..., s^n) \tag{2}$$

where $f(.)$ corresponds to the fusion function. The independent variables (i.e., information about any group of variables does not change the belief about the others) $s^1, ..., s^n$, are imprecise and dependent on the class of observation and hence given as:

$$s_j = f(s_j^1, ..., s_j^n) \tag{3}$$

where j refers to the class of the observation.

Variance of the IDSs determines how good their average quality is when each IDS acts individually. Lower variance corresponds to a better performance. Covariance among detectors measures the dependence of the detectors. The more the dependence, the lesser the gain benefited out of fusion.

Let us consider two cases here. In the first case, n responses are available for each access and these n responses are used independent of each other. The average of variance of s_j over all $i = 1, 2, ..., n$, denoted as $(\sigma^j_{av})^2$ is given as:

$$(\sigma^j_{av})^2 = \frac{1}{n}\sum_{i=1}^{n}(\sigma^j_i)^2 \tag{4}$$

In the second case, all n responses are combined using the mean operator. The variance over many accesses is denoted by $(\sigma^j_{fusion})^2$ and is called the variance of average given by:

$$(\sigma^j_{fusion})^2 = \frac{1}{n^2}\sum_{i=1}^{n}(\sigma^j_i)^2 + \frac{1}{n^2}\sum_{i=1,i\neq k}^{n}\sum_{k=1,i\neq k}^{n}\rho^j_{i,k}\sigma^j_i\sigma^j_k \tag{5}$$

where $\rho^j_{i,k}$ is the correlation coefficient between the ith and kth detectors and for j taking the different class values. The first term is the average variance of the base-experts while the second term is the covariance between ith and kth detectors for $i \neq k$. This is because the term $\rho^j_{i,k}\sigma^j_i\sigma^j_k$ is by definition equivalent to correlation. On analysis, it is seen that:

$$(\sigma^j_{fusion})^2 \leq (\sigma^j_{av})^2 \tag{6}$$

When two detector scores are merged by a simple mean operator, the resultant variance of the final score will be reduced with respect to the average variance of the two original scores. Since $0 \leq \rho^j_{m,n} \leq 1$,

$$\frac{1}{n}(\sigma^j_{av})^2 \leq (\sigma^j_{fusion})^2 \tag{7}$$

Equation 6 and equation 7 give the lower and upper bound of $(\sigma^j_{fusion})^2$, attained with correlation and uncorrelation respectively. Any positive correlation results in a variance between these bounds. Hence, by combining responses using the mean operator, the resultant variance is assured to be smaller than the average (not the minimum) variance.

Fusion of the scores reduces variance, which in turn results in reduction of error (with respect to the case where scores are used separately). To measure explicitly the factor of reduction in variance,

$$\frac{1}{n}(\sigma^j_{av})^2 \leq (\sigma^j_{fusion})^2 \leq (\sigma^j_{av})^2 \tag{8}$$

Factor of reduction in variance, $(v_r) = \frac{(\sigma^j_{av})^2}{(\sigma^k_{fusion})^2}$

$1 \leq v_r \leq n$

This clearly indicates that the reduction in variance is more when more detectors are used, i.e., increasing n, the better will be the combined system, even if the hypotheses of underlying IDSs are correlated. This comes at a cost of increased computation, proportional to the value of n. The reduction in variance of the individual classes results in lesser overlap between the class distributions. Thus the chances of error reduces, which in turn results in an improved detection. This forms the argument in this chapter for why fusion using multiple detectors works for intrusion detection application.

Following common possibilities encountered on combining two detectors are analyzed:

1. combining two uncorrelated experts with very different performances;
2. combining two highly correlated experts with very different performances;
3. combining two uncorrelated experts with very similar performances;
4. combining two highly correlated experts with very similar performances.

Fusing IDSs of similar and different performances are encountered in almost all practical fusion problems. Considering the first case, without loss of generality it can be assumed that system 1 is better than system 2, i.e., $\sigma_1 < \sigma_2$ and $\rho = 0$. Hence, for the combination to be better than the best system, i.e., system 1, it is required that

$$(\sigma^j_{fusion})^2 < (\sigma^j_1)^2; \quad \frac{(\sigma^j_1)^2 + (\sigma^j_2)^2 + 2\rho\sigma^j_1\sigma^j_2}{4} < (\sigma^j_1)^2; \quad (\sigma^j_2)^2 < 3(\sigma^j_1)^2 - 2\rho\sigma^j_1\sigma^j_2 \tag{9}$$

The covariance is zero in general for cases 1 and 3. Hence, the combined system will benefit from the fusion when the variance of one ($(\sigma^j_2)^2$) is at most less than 3 times of the variance of the other $(\sigma^j_1)^2$ since $\rho = 0$. Furthermore, correlation [or equivalently covariance; one is proportional to the other] between the two systems penalizes this margin of $3(\sigma^j_1)^2$. This is particularly true for the second case since $\rho > 0$. Also, it should be noted that $\rho < 0$ (which implies negative correlation) could allow for larger $(\sigma^j_2)^2$. As a result, adding another system that is negatively correlated, but with large variance (hence large error) will improve fusion ($(\sigma^j_{fusion})^2 < \frac{1}{n}(\sigma^j_{av})^2$). Unfortunately, with IDSs, two systems are either positively correlated or not correlated, unless these systems are jointly trained together by algorithms such as negative correlation learning Brown (2004). For a given detector i, s_i for $i = 1, ..., n$, will tend to agree with each other (hence positively correlated) most often than to disagree with each other (hence negatively correlated). By fusing scores obtained from IDSs that are trained independently, one can almost be certain that $0 \leq \rho_{m,n} \leq 1$.

For the third and fourth cases, we have $(\sigma^j_1)^2 \approx (\sigma^j_2)^2$. Hence, $\rho(\sigma^j_2)^2 < (\sigma^j_1)^2$. Note that for the third case with $\rho \approx 0$, the above constraint gets satisfied. Hence, fusion will definitely lead to better performance. On the other hand, for the fourth case where $\rho \approx 1$, fusion may not necessarily lead to better performance.

From the above analysis using a mean operator for fusion, the conclusion drawn are the following:

The analysis explains and shows that fusing two systems of different performances is not always beneficial. The theoretical analysis shows that if the weaker IDS has (class-dependent)

variance three times larger that of the best IDS, the gain due to fusion breaks down. This is even more true for correlated base-experts as correlation penalizes this limit further. It is also seen that fusing two uncorrelated IDSs of similar performance always result in improved performance. Finally, fusing two correlated IDSs of similar performance will be beneficial only when the covariance of the two IDSs are less than the variance of the IDSs.

It is necessary to show that a lower bound of accuracy results in the case of sensor fusion. This can be proved as follows:

Given the fused output as $s = \sum_i w_i s_i$, the quadratic error of a sensor indexed i, (e_i), and also the fused sensor, (e_{fusion}) are given by:

$$e_i = (s_i - c)^2 \tag{10}$$

and

$$e_{fusion} = (s_{fusion} - c)^2 \tag{11}$$

respectively, where w_i is the weighting on the ith detector, and c is the target. The ambiguity of the sensor is defined as:

$$a_i = (s_i - s)^2 \tag{12}$$

The squared error of the fused sensor is seen to be equal to the weighted average squared error of the individuals, minus a term which measures average correlation. This allows for non-uniform weights (with the constraint $\sum_i w_i = 1$). Hence, the general form of the ensemble output is $s = \sum_i w_i s_i$.

The ambiguity of the fused sensor is given by:

$$a_{fusion} = \sum_i w_i a_i = \sum_i w_i (s_i - s)^2 \tag{13}$$

On solving equation 13, the error due to the combination of several detectors is obtained as the difference between the weighted average error of individual detectors and the ambiguity among the fusion member decisions.

$$e_{fusion} = \sum_i w_i (s_i - c)^2 - \sum_i w_i (s_i - s)^2 \tag{14}$$

The ambiguity among the fusion member decisions is always positive and hence the combination of several detectors is expected to be better than the average over several detectors. This result turns out to be very important for the focus of this chapter.

4. Solution Approaches

In the case of fusion problem, the solution approaches depend on whether there is any knowledge regarding the traffic and the intrusion detectors. This section initially considers no knowledge of the IDSs and the intrusion detection data and later with a knowledge of available IDSs and evaluation dataset. There is an arsenal of different theories of uncertainty and methods based on these theories for making decisions under uncertainty. There is no consensus as to which method is most suitable for problems with epistemic uncertainty, when information is scarce and imprecise. The choice of heterogeneous detectors is expected to result in decisions that conflict or be in consensus, completely or partially. The detectors can be categorized by their output s_i, i.e., probability (within the range [0, 1]), Basic Probability Assignment (BPA) m (within the range [0, 1]), membership function (within the range [0, 1]), distance metric (more than or equal to zero), or log-likelihood ratio (a real number).

Consider a body of evidence $(F; m)$, where F represents the set of all focal elements and m their corresponding basic probability assignments. This analysis without any knowledge of the system or the data, attempts to prove the acceptance of sensor fusion in improving the intrusion detection performance and hence is unlimited in scope. In this analysis the Dempster-Shafer fusion operator is used since it is more acceptable for intrusion detection application as explained below.

Dempster-Shafer theory considers two types of uncertainty; 1) due to the imprecision in the evidence, and 2) due to the conflict. Non specificity and strife measure the uncertainty due to imprecision and conflict, respectively. The larger the focal elements of a body of evidence, the more imprecise is the evidence and, consequently, the higher is non specificity. When the evidence is precise (all the focal elements consist of a single member), non specificity is zero. The importance of Dempster-Shafer theory in intrusion detection is that in order to track statistics, it is necessary to model the distribution of decisions. If these decisions are probabilistic assignments over the set of labels, then the distribution function will be too complicated to retain precisely. The Dempster-Shafer theory of evidence solves this problem by simplifying the opinions to Boolean decisions, so that each detector decision lies in a space having 2^{Θ} elements, where Θ defines the working space. In this way, the full set of statistics can be specified using 2^{Θ} values.

4.1 Dempster-Shafer Combination Method

Dempster-Shafer (DS) theory is required to model the situation in which a classification algorithm cannot classify a target or cannot exhaustively list all of the classes to which it could belong. This is most acceptable in the case of unknown attacks or novel attacks or the case of zero *a priori* knowledge of data distribution. DS theory does not attempt to formalize the emergence of novelties, but it is a suitable framework for reconstructing the formation of beliefs when novelties appear. An application of decision making in the field of intrusion detection illustrates the potentialities of DS theory, as well as its shortcomings.

The DS rule corresponds to conjunction operator since it builds the belief induced by accepting two pieces of evidence, i.e., by accepting their conjunction. Shafer developed the DS theory of evidence based on the model that all the hypotheses in the FoD are exclusive and the frame is exhaustive. The purpose is to combine/aggregate several independent and equi-reliable sources of evidence expressing their belief on the set. The aim of using the DS theory of

fusion is that with any set of decisions from heterogeneous detectors, sensor fusion can be modeled as utility maximization. DS theory of combination conceives novel categories that classify empirical evidence in a novel way and, possibly, are better able to discriminate the relevant aspects of emergent phenomena. Novel categories detect novel empirical evidence, that may be fragmentary, irrelevant, contradictory or supportive of particular hypotheses. The DS theory approach for quantifying the uncertainty in the performance of a detector and assessing the improvement in system performance, consists of three steps:

1. Model uncertainty by considering each variable separately. Then a model that considers all variables together is derived.
2. Propagate uncertainty through the system, which results in a model of uncertainty in the performance of the system.
3. Assess the system performance enhancement.

In the case of Dempster-Shafer theory, Θ is the Frame of Discernment (FoD), which defines the working space for the desired application. FoD is expected to contain all propositions of which the information sources (IDSs) can provide evidence. When a proposition corresponds to a subset of a frame of discernment, it is said that the frame discerns that proposition. It is expected that the elements of the frame of discernment, Θ are assumed to be exclusive propositions. This is a constraint, which always gets satisfied in intrusion detection application because of the discrete nature of the detector decision. The belief of likelihood of the traffic to be in an anomalous state is detected by various IDSs by means of a mass to the subsets of the FoD.

The DS theory is a generalization of the classical probability theory with its additivity axiom excluded or modified. The probability mass function (p) is a mapping which indicates how the probability mass is assigned to the elements. The Basic Probability Assignment (BPA) function (m) on the other hand is the set mapping, and the two can be related $\forall\ A \subseteq \Theta$ as $m(A) = \sum_{B \epsilon A} p(B)$ and hence obviously $m(A)$ relates to a belief structure. The mass m is very near to the probabilistic mass p, except that it is shared not only by the single hypothesis but also to the union of the hypotheses.

In DS theory, rather than knowing exactly how the probability is distributed to each element $B \in \Theta$, we just know by the BPA function m that a certain quantity of a probability mass is somehow divided among the focal elements. Because of this less specific knowledge about the allocation of the probability mass, it is difficult to assign exactly the probability associated with the subsets of the FoD, but instead we assign two measures: the (1) belief (Bel) and (2) plausibility (Pl), which correspond to the lower and upper bounds on the probability,

i.e., $Bel(A) \leq p(A) \leq Pl(A)$

where the belief function, $Bel(A)$, measures the minimum uncertainty value about proposition A, and the Plausibility, $Pl(A)$, reflects the maximum uncertainty value about proposition A.

The following are the key assumptions made with the fusion of intrusion detectors:

- If some of the detectors are imprecise, the uncertainty can be quantified about an event by the maximum and minimum probabilities of that event. Maximum (minimum) probability of an event is the maximum (minimum) of all probabilities that are consistent with the available evidence.
- The process of asking an IDS about an uncertain variable is a random experiment whose outcome can be precise or imprecise. There is randomness because every time a different IDS observes the variable, a different decision can be expected. The IDS can be precise and provide a single value or imprecise and provide an interval. Therefore, if the information about uncertainty consists of intervals from multiple IDSs, then there is uncertainty due to both imprecision and randomness.

If all IDSs are precise, then the pieces of evidence from these IDSs point precisely to specific values. In this case, a probability distribution of the variable can be build. However, if the IDSs provide intervals, such a probability distribution cannot be build because it is not known as to what specific values of the random variables each piece of evidence supports.

Also the additivity axiom of probability theory $p(A) + p(\bar{A}) = 1$ is modified as $m(A) + m(\bar{A}) + m(\Theta) = 1$, in the case of evidence theory, with uncertainty introduced by the term $m(\Theta)$. $m(A)$ is the mass assigned to A, $m(\bar{A})$ is the mass assigned to all other propositions that are not A in FoD and $m(\Theta)$ is the mass assigned to the union of all hypotheses when the detector is ignorant. This clearly explains the advantages of evidence theory in handling an uncertainty where the detector's joint probability distribution is not required.

The equation $Bel(A) + Bel(\bar{A}) = 1$, which is equivalent to $Bel(A) = Pl(A)$, holds for all subsets A of the FoD if and only if $Bel's$ focal points are all singletons. In this case, Bel is an additive probability distribution. Whether normalized or not, the DS method satisfies the two

axioms of combination: $0 \leq m(A) \leq 1$ and $\sum_{A \subseteq \Theta} m(A) = 1$. The third axiom $\sum m(\phi) = 0$ is not satisfied by the unnormalized DS method. Also, independence of evidence is yet another requirement for the DS combination method.

The problem is formalized as follows: Considering the network traffic, assume a traffic space Θ, which is the union of the different classes, namely, the attack and the normal. The attack class have different types of attacks and the classes are assumed to be mutually exclusive. Each IDS assigns to the traffic, the detection of any of the traffic sample $x \in \Theta$, that denotes the traffic sample to come from a class which is an element of the FoD, Θ. With n IDSs used for the combination, the decision of each one of the IDSs is considered for the final decision of the fusion IDS.

This chapter presents a method to detect the unknown traffic attacks with an increased degree of confidence by making use of a fusion system composed of detectors. Each detector observes the same traffic on the network and detects the attack traffic with an uncertainty index. The frame of discernment consists of singletons that are exclusive ($A_i \cap A_j = \phi$, $\forall i \neq j$) and are exhaustive since the FoD consists of all the expected attacks which the individual IDS detects or else the detector fails to detect by recognizing it as a normal traffic. All the constituent IDSs that take part in fusion is assumed to have a global point of view about the system rather than

separate detectors being introduced to give specialized opinion about a single hypothesis.

The DS combination rule gives the combined mass of the two evidence m_1 and m_2 on any subset A of the FoD as $m(A)$ given by:

$$m(A) = \frac{\sum_{X \cap Y = A} m_1(X) m_2(Y)}{1 - \sum_{X \cap Y = \phi} m_1(X) m_2(Y)} \tag{15}$$

The numerator of Dempster-Shafer combination equation 15 represents the influence of aspects of the second evidence that confirm the first one. The denominator represents the influence of aspects of the second evidence that contradict the first one. The denominator of equation 15 is $1 - k$, where k is the conflict between the two evidence. This denominator is for normalization, which spreads the resultant uncertainty of any evidence with a weight factor, over all focal elements and results in an intuitive decision. i.e., the effect of normalization consists of eliminating the conflicting pieces of information between the two sources to combine, consistently with the intersection operator. Dempster-Shafer rule does not apply if the two evidence are completely contradictory. It only makes sense if $k < 1$. If the two evidence are completely contradictory, they can be handled as one single evidence over alternative possibilities whose BPA must be re-scaled in order to comply with equation 15. The meaning of Dempster-Shafer rule 15 can be illustrated in the simple case of two evidence on an observation A. Suppose that one evidence is $m_1(A) = p$, $m_1(\Theta) = 1 - p$ and that another evidence is $m_2(A) = q$, $m(\Theta) = 1 - q$. The total evidence in favor of A = The denominator of equation 15 = $1 - (1 - p)(1 - q)$. The fraction supported by both the bodies of evidence = $\frac{pq}{(1-p)(1-q)}$

Specifically, if a particular detector indexed i taking part in fusion has probability of detection $m_i(A)$ for a particular class A, it is expected that fusion results in the probability of that class as $m(A)$, which is expected to be more that $m_i(A)\ \forall\ i$ and A. Thus the confidence in detecting a particular class is improved, which is the key aim of sensor fusion. The above analysis is simple since it considers only one class at a time. The variance of the two classes can be merged and the resultant variance is the sum of the normalized variances of the individual classes. Hence, the class label can be dropped.

4.2 Analysis of Detection Error Assuming Traffic Distribution

The previous sections analyzed the system without any knowledge about the underlying traffic or detectors. The Gaussian distribution is assumed for both the normal and the attack traffic in this section due to its acceptability in practice. Often, the data available in 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. Formally, each database value is indeed a random variable; the mean of this variable becomes the stored value, and is interpreted as an approximation of the true value; the standard deviation of this variable is a measure of the level of accuracy of the stored value.

Assuming the attack connection and normal connection scores to have the mean values $y^i_{j=I} = \mu_I$ and $y^i_{j=NI} = \mu_{NI}$ respectively, $\mu_I > \mu_{NI}$ without loss of generality. Let σ_I and σ_{NI} be the standard deviation of the attack connection and normal connection scores. The two types of errors committed by IDSs are often measured by False Positive Rate (FP_{rate}) and False Negative Rate (FN_{rate}). FP_{rate} is calculated by integrating the attack score distribution from a given threshold T in the score space to ∞, while FN_{rate} is calculated by integrating the normal distribution from $-\infty$ to the given threshold T. The threshold T is a unique point where the error is minimized, i.e., the difference between FP_{rate} and FN_{rate} is minimized by the following criterion:

$$T = argmin(|FP_{rate_T} - FN_{rate_T}|) \tag{16}$$

At this threshold value, the resultant error due to FP_{rate} and FN_{rate} is a minimum. This is because the FN_{rate} is an increasing function (a cumulative density function, cdf) and FP_{rate} is a decreasing function ($1 - cdf$). T is the point where these two functions intersect. Decreasing the error introduced by the FP_{rate} and the FN_{rate} implies an improvement in the performance of the system.

$$FP_{rate} = \int_T^{\infty} (p^{k=NI}) dy \tag{17}$$

$$FN_{rate} = \int_{-\infty}^{T} (p^{k=I}) dy \tag{18}$$

The fusion algorithm accepts decisions from many IDSs, where a minority of the decisions are false positives or false negatives. A good sensor fusion system is expected to give a result that accurately represents the decision from the correctly performing individual sensors, while minimizing the decisions from erroneous IDSs. Approximate agreement emphasizes precision, even when this conflicts with system accuracy. However, sensor fusion is concerned solely with the accuracy of the readings, which is appropriate for sensor applications. This is true despite the fact that increased precision within known accuracy bounds would be beneficial in most of the cases. Hence the following strategy is being adopted:

- The false alarm rate FP_{rate} can be fixed at an acceptable value α_0 and then the detection rate can be maximized. Based on the above criteria a lower bound on accuracy can be derived.
- The detection rate is always higher than the false alarm rate for every IDS, an assumption that is trivially satisfied by any reasonably functional sensor.
- Determine whether the accuracy of the IDS after fusion is indeed better than the accuracy of the individual IDSs in order to support the performance enhancement of fusion IDS.
- To discover the weights on the individual IDSs that gives the best fusion.

Given the desired false alarm rate which is acceptable, $FP_{rate} = \alpha_0$, the threshold (T) that maximizes the TP_{rate} and thus minimizes the FN_{rate};

$$TP_{rate} = Pr[\sum_{i=1}^{n} w_i s_i \geq T \,|attack] \tag{19}$$

$$FP_{rate} = Pr[\sum_{i=1}^{n} w_i s_i \geq T | normal] = \alpha_0 \tag{20}$$

The fusion of IDSs becomes meaningful only when $FP \leq FP_i \quad \forall i$ and $TP \geq TP_i \quad \forall i$. In order to satisfy these conditions, an adaptive or dynamic weighting of IDSs is the only possible alternative. Model of the fusion output is given as:

$$s = \sum_{i=1}^{n} w_i s_i \quad and \quad TP_i = Pr[s_i = 1 | attack], \quad FP_i = Pr[s_i = 1 | normal] \tag{21}$$

where TP_i is the detection rate and FP_i is the false positive rate of any individual IDS indexed i. It is required to provide a low value of weight to any individual IDS that is unreliable, hence meeting the constraint on false alarm as given in equation 20. Similarly, the fusion improves the TP_{rate}, since the detectors get appropriately weighted according to their performance.

Fusion of the decisions from various IDSs is expected to produce a single decision that is more informative and accurate than any of the decisions from the individual IDSs. Then the question arises as to whether it is optimal. Towards that end, a lower bound on variance for the fusion problem of independent sensors, or an upper bound on the false positive rate or a lower bound on the detection rate for the fusion problem of dependent sensors is presented in this chapter.

4.2.1 Fusion of Independent Sensors

The decisions from various IDSs are assumed to be statistically independent for the sake of simplicity so that the combination of IDSs will not diffuse the detection. In sensor fusion, improvements in performances are related to the degree of error diversity among the individual IDSs.

Variance and Mean Square Error of the estimate of fused output

The successful operation of a multiple sensor system critically depends on the methods that combine the outputs of the sensors. A suitable rule can be inferred using the training examples, where the errors introduced by various individual sensors are unknown and not controllable. The choice of the sensors has been made and the system is available, and the fusion rule for the system has to be obtained. A system of n sensors $IDS_1, IDS_2, ..., IDS_n$ is considered; corresponding to an observation with parameter x, $x \in \Re^m$, sensor IDS_i yields output s^i, $s^i \in \Re^m$ according to an unknown probability distribution p_i. A training $l-$sample (x_1, y_1), (x_2, y_2), ..., (x_l, y_l) is given where $y_i = (s_i^1, s_i^2, ..., s_i^n)$ and s_j^i is the output of IDS_i in response to the input x_j. The problem is to estimate a fusion rule $f : \Re^{nm} \rightarrow \Re^m$, based on the sample, such that the expected square error is minimized over a family of fusion rules based on the given $l-$sample.

Consider n independent IDSs with the decisions of each being a random variable with Gaussian distribution of zero mean vector and covariance matrix diagonal $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. Assume s to be the expected fusion output, which is the unknown deterministic scalar quantity to be

estimated and $\hat{s}$ to be the estimate of the fusion output. In most cases the estimate is a deterministic function of the data. Then the mean square error (MSE) associated with the estimate $\hat{s}$ for a particular test data set is given as $E[(s-\hat{s})^2]$. For a given value of s, there are two basic kinds of errors:

- Random error, which is also called precision or estimation variance.
- Systematic error, which is also called accuracy or estimation bias.

Both kinds of errors can be quantified by the conditional distribution of the estimates $pr(\hat{s}-s)$. The MSE of a detector is the expected value of the error and is due to the randomness or due to the estimator not taking into account the information that could produce a more accurate result.

$$MSE = E[(s-\hat{s})^2] = Var(\hat{s}) + (Bias(\hat{s},s))^2 \tag{22}$$

The MSE is the absolute error used to assess the quality of the sensor in terms of its variation and unbiasedness. For an unbiased sensor, the MSE is the variance of the estimator, or the root mean squared error $(RMSE)$ is the standard deviation. The standard deviation measures the accuracy of a set of probability assessments. The lower the value of $RMSE$, the better it is as an estimator in terms of both the precision as well as the accuracy. Thus, reduced variance can be considered as an index of improved accuracy and precision of any detector. Hence, the reduction in variance of the fusion IDS to show its improved performance is proved in this chapter. The Cramer-Rao inequality can be used for deriving the lower bound on the variance of an estimator.

Cramer-Rao Bound (CRB) for fused output

The Cramer-Rao lower bound is used to get the best achievable estimation performance. Any sensor fusion approach which achieves this performance is optimum in this regard. CR inequality states that the reciprocal of the Fisher information is an asymptotic lower bound on the variance of any unbiased estimator $\hat{s}$. Fisher information is a method for summarizing the influence of the parameters of a generative model on a collection of samples from that model. In this case, the parameters we consider are the means of the Gaussians. Fisher information is the variance, (σ^2) of the score (partial derivative of the logarithm of the likelihood function of the network traffic with respect to σ^2).

$$score = \frac{\partial}{\partial\sigma^2} ln(L(\sigma^2;s)) \tag{23}$$

Basically, the score tells us how sensitive the log-likelihood is to changes in parameters. This is a function of variance, σ^2 and the detection s and this score is a sufficient statistic for variance. The expected value of this score is zero, and hence the Fisher information is given by:

$$E\left\{[\frac{\partial}{\partial\sigma^2} ln(L(\sigma^2;s))]^2|\sigma^2\right\} \tag{24}$$

Fisher information is thus the expectation of the squared score. A random variable carrying high Fisher information implies that the absolute value of the score is often high.

Cramer-Rao inequality expresses a lower bound on the variance of an unbiased statistical estimator, based on the Fisher information.

$$\sigma^2 \geq \frac{1}{Fisher\ information} = \frac{1}{E\left\{[\frac{\partial}{\partial\sigma^2} ln(L(\sigma^2;X))]^2|\sigma^2\right\}} \tag{25}$$

If the prior probability of detection of the various IDSs are known, the weights $w_{i|i=1,---n}$ can be assigned to the individual IDSs. The idea is to estimate the local accuracy of the IDSs. The decision of the IDS with the highest local accuracy estimate will have the highest weighting on aggregation. The best fusion algorithm is supposed to choose the correct class if any of the individual IDS did so. This is a theoretical upper bound for all fusion algorithms. Of course, the best individual IDS is a lower bound for any meaningful fusion algorithm. Depending on the data, the fusion may sometimes be no better than Bayes. In such cases, the upper and lower performance bounds are identical and there is no point in using a fusion algorithm. A further insight into CRB can be gained by understanding how each IDS affects it. With the architecture shown in Fig. 1, the model is given by $\hat{s} = \sum_{i=1}^{n} w_i s_i$. The bound is calculated from the effective variance of each one of the IDSs as $\hat{\sigma_i^2} = \frac{\sigma_i^2}{w_i^2}$ and then combining them to have the CRB as $\frac{1}{\sum_{i=1}^{n} \frac{1}{\hat{\sigma_i^2}}}$.

The weight assigned to the IDSs is inversely proportional to the variance. This is due to the fact that, if the variance is small, the IDS is expected to be more dependable. The bound on the smallest variance of an estimation $\hat{s}$ is given as:

$$\hat{\sigma^2} = E[(\hat{s}-s)^2] \geq \frac{1}{\sum_{i=1}^{n} \frac{w_i^2}{\sigma_i^2}} \tag{26}$$

It can be observed from equation 26 that any IDS decision that is not reliable will have a very limited impact on the bound. This is because the non-reliable IDS will have a much larger variance than other IDSs in the group; $\hat{\sigma_n^2} \gg \hat{\sigma_1^2}, - - - , \hat{\sigma_{n-1}^2}$ and hence $\frac{1}{\hat{\sigma_n^2}} \ll \frac{1}{\hat{\sigma_1^2}}, --- , \frac{1}{\hat{\sigma_{n-1}^2}}$. The bound can then be approximated as $\frac{1}{\sum_{i=1}^{n-1} \frac{1}{\hat{\sigma_i^2}}}$.

Also, it can be observed from equation 26 that the bound shows asymptotically optimum behavior of minimum variance. Then, $\hat{\sigma_i^2} > 0$ and $\hat{\sigma_{min}^2} = min[\hat{\sigma_i^2}, - - -, \hat{\sigma_n^2}]$, then

$$CRB = \frac{1}{\sum_{i=1}^{n} \frac{1}{\hat{\sigma_i^2}}} < \hat{\sigma_{min}^2} \leq \hat{\sigma_i^2} \tag{27}$$

From equation 27 it can also be shown that perfect performance is apparently possible with enough IDSs. The bound tends to zero as more and more individual IDSs are added to the fusion unit.

$$CRB_{n\to\infty} = Lt_{n\to\infty} \frac{1}{\frac{1}{\hat{\sigma_1^2}} + - - - + \frac{1}{\hat{\sigma_n^2}}} \tag{28}$$

For simplicity assume homogeneous IDSs with variance $\hat{\sigma^2}$;

$$CRB_{n\to\infty} = Lt_{n\to\infty}\frac{1}{\frac{n}{\hat{\sigma^2}}} = Lt_{n\to\infty}\frac{\hat{\sigma^2}}{n} = 0 \tag{29}$$

From equation 28 and equation 29 it can be easily interpreted that increasing the number of IDSs to a sufficiently large number will lead to the performance bounds towards perfect estimates. Also, due to monotone decreasing nature of the bound, the IDSs can be chosen to make the performance as close to perfect.

4.2.2 Fusion of Dependent Sensors

In most of the sensor fusion problems, individual sensor errors are assumed to be uncorrelated so that the sensor decisions are independent. While independence of sensors is a good assumption, it is often unrealistic in the normal case.

Setting bounds on false positives and true positives

As an illustration, let us consider a system with three individual IDSs, with a joint density at the IDSs having a covariance matrix of the form:

$$\bigwedge = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix} \tag{30}$$

The false alarm rate (α) at the fusion center, where the individual decisions are aggregated can be written as:

$$\alpha_{max} = 1 - Pr(s_1 = 0, s_2 = 0, s_3 = 0|normal) = 1 - \int_{-\infty}^{t}\int_{-\infty}^{t}\int_{-\infty}^{t} P_s(s|normal)ds \tag{31}$$

where $P_s(s|normal)$ is the density of the sensor observations under the hypothesis *normal* and is a function of the correlation coefficient, ρ. Assuming a single threshold, T, for all the sensors, and with the same correlation coefficient, ρ between different sensors, a function $F_n(T|\rho) = Pr(s_1 = 0, s_2 = 0, s_3 = 0)$ can be defined.

$$F_n(T|\rho) = \int_{-\infty}^{-\infty} F^n(\frac{T - \sqrt{\rho}y}{\sqrt{1-\rho}})f(y)dy \tag{32}$$

where $f(y)$ and $F(X)$ are the standard normal density and cumulative distribution function respectively.

$F^n(X) = [F(X)]^n$

Equation 31 can be written depending on whether $\rho > \frac{-1}{n-1}$ or not, as:

$$\alpha_{max} = 1 - \int_{-\infty}^{\infty} F^3(\frac{T - \sqrt{\rho}y}{\sqrt{1-\rho}})f(y)dy \quad for \quad 0 \leq \rho < 1 \tag{33}$$

and

$$\alpha_{max} = 1 - F^3(T\,|\rho) \quad for \quad -0.5 \leq \rho < 1 \tag{34}$$

With this threshold T, the probability of detection at the fusion unit can be computed as:

$$TP_{min} = 1 - \int_{-\infty}^{\infty} F^3\left(\frac{T - S - \sqrt{\rho}y}{\sqrt{1-\rho}}\right) f(y)dy \quad for \quad 0 \leq \rho < 1 \tag{35}$$

and

$$TP_{min} = 1 - F^3(T - S\,|\rho) \quad for \quad -0.5 \leq \rho < 1 \tag{36}$$

The above equations 33, 34, 35, and 36, clearly showed the performance improvement of sensor fusion where the upper bound on false positive rate and lower bound on detection rate were fixed. The system performance was shown to deteriorate when the correlation between the sensor errors was positive and increasing, while the performance improved considerably when the correlation was negative and increasing.

The above analysis were made with the assumption that the prior detection probability of the individual IDSs were known and hence the case of bounded variance. However, in case the IDS performance was not known *a priori*, it was a case of unbounded variance and hence given the trivial model it was difficult to accuracy estimate the underlying decision. This clearly emphasized the difficulty of sensor fusion problem, where it becomes a necessity to understand the individual IDS behavior. Hence the architecture was modified as proposed in the work of Thomas & Balakrishnan (2008) and shown in Fig. 2 with the model remaining the same. With this improved architecture using a neural network learner, a clear understanding of each one of the individual IDSs was obtained. Most other approaches treat the training data as a monolithic whole when determining the sensor accuracy. However, the accuracy was expected to vary with data. This architecture attempts to predict the IDSs that are reliable for a given sample data. This architecture is demonstrated to be practically successful and is also the true situation where the weights are neither completely known nor totally unknown.

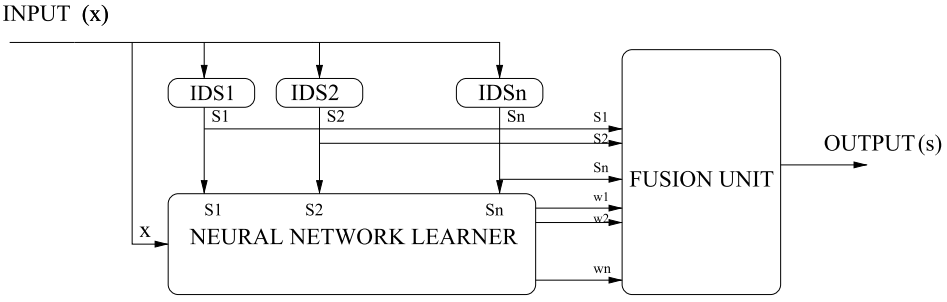


Fig. 2. Data-Dependent Decision Fusion architecture

4.3 Data-Dependent Decision Fusion Scheme

It is necessary to incorporate an architecture that considers a method for improving the detection rate by gathering an in-depth understanding on the input traffic and also on the behavior of the individual IDSs. This helps in automatically learning the individual weights for the

combination when the IDSs are heterogeneous and shows difference in performance. The architecture should be independent of the dataset and the structures employed, and has to be used with any real valued data set.

A new data-dependent architecture underpinning sensor fusion to significantly enhance the IDS performance is attempted in the work of Thomas & Balakrishnan (2008; 2009). A better architecture by explicitly introducing the data-dependence in the fusion technique is the key idea behind this architecture. The disadvantage of the commonly used fusion techniques which are either implicitly data-dependent or data-independent, is due to the unrealistic confidence of certain IDSs. The idea in this architecture is to properly analyze the data and understand when the individual IDSs fail. The fusion unit should incorporate this learning from input as well as from the output of detectors to make an appropriate decision. The fusion should thus be data-dependent and hence the rule set has to be developed dynamically. This architecture is different from conventional fusion architectures and guarantees improved performance in terms of detection rate and the false alarm rate. It works well even for large datasets and is capable of identifying novel attacks since the rules are dynamically updated. It also has the advantage of improved scalability.

The Data-dependent Decision fusion 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 as the final stage. 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. The architecture is independent of the dataset and the structures employed, and can be used with any real valued dataset. Thus it is reasonable to make use of a neural network learner unit to understand the performance and assign weights to various individual IDSs in the case of a large dataset.

The weight assigned to any IDS not only depends on the output of that IDS as in the case of the probability theory or the Dempster-Shafer theory, but also on the input traffic which causes this output. A neural network unit is fed with the output of the IDSs along with the respective input for an in-depth understanding of the reliability estimation of the IDSs. The alarms produced by the different IDSs when they are presented with a certain attack clearly tell which sensor generated more precise result and what attacks are actually occurring on the network traffic. The output of the neural network unit corresponds to the weights which are assigned to each one of the individual IDSs. The IDSs can be fused with the weight factor to produce an improved resultant output.

This architecture refers to a collection of diverse IDSs that respond to an input traffic and the weighted combination of their predictions. The weights are learned by looking at the response of the individual sensors for every input traffic connection. The fusion output is represented as:

$$s = F_j(w_j^i(x_j, s_j^i), s_j^i), \tag{37}$$

where the weights w_j^i are dependent on both the input x_j as well as individual IDS's output s_j^i, where the suffix j refers to the class label and the prefix i refers to the IDS index. The fusion unit used gives a value of one or zero depending on the set threshold being higher or lower than the weighted aggregation of the IDS's decisions.

The training of the neural network unit by back propagation involves three stages: 1) 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, 2) the calculation and the back propagation of the associated error, and 3) the adjustments of the weights. After the training, the neural network is used for the computations of the feedforward phase. A multilayer network with a single hidden layer is sufficient in our application to learn the reliability of the IDSs to an arbitrary accuracy according to the proof available in Fausett (2007).

Consider the problem formulation where the weights $w_1, ..., w_n$, take on constrained values to satisfy the condition $\sum_{i=1}^{n} w_i = 1$. Even without any knowledge about the IDS selectivity factors, the constraint on the weights assures the possibility to accuracy estimate the underlying decision. With the weights learnt for any data, it becomes a useful generalization of the trivial model which was initially discussed. The improved efficient model with good learning algorithm can be used to find the optimum fusion algorithms for any performance measure.

5. Results and Discussion

This section includes the empirical evaluation to support the theoretical analysis on the acceptability of sensor fusion in intrusion detection.

5.1 Data Set

The proposed fusion IDS was evaluated on two data, one being the real-world network traffic embedded with attacks and the second being the DARPA-1999 (1999). The real traffic within a protected University campus network was collected during the working hours of a day. This traffic of around two million packets was divided into two halves, one for training the anomaly IDSs, and the other for testing. The test data was injected with 45 HTTP attack packets using the HTTP attack traffic generator tool called libwhisker Libwhisker (n.d.). The test data set was introduced with a base rate of 0.0000225, which is relatively realistic. The MIT Lincoln Laboratory under DARPA and AFRL sponsorship, has collected and distributed the first standard corpora for evaluation of computer network IDSs. This MIT- DARPA-1999 (1999) was used to train and test the performance of IDSs. 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.

The classification of the various attacks found in the network traffic is explained in detail in the thesis work of Kendall (1999) with respect to DARPA intrusion detection evaluation dataset and is explained here in brief. The attacks fall into four main classes namely, Probe, Denial of Service(DoS), Remote to Local(R2L) and the User to Root (U2R). The Probe or Scan attacks

automatically scan a network of computers or a DNS server to find valid IP addresses, active ports, host operating system types and known vulnerabilities. The DoS attacks are designed to disrupt a host or network service. In R2L attacks, an attacker who does not have an account on a victim machine gains local access to the machine, exfiltrates files from the machine or modifies data in transit to the machine. In U2R attacks, a local user on a machine is able to obtain privileges normally reserved for the unix super user or the windows administrator.

Even with the criticisms by McHugh (2000) and Mahoney & Chan (2003) 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 Thomas & Balakrishnan (2007). Also, even after ten years of its generation, even now there are lot of attacks in the dataset for which signatures are not available in database of even the frequently updated signature based IDSs like Snort (1999). The real data traffic is difficult to work with; the main reason being the lack of the information regarding the status of the traffic. Even with intense analysis, the prediction can never be 100 percent accurate because of the stealthiness and sophistication of the attacks and the unpredictability of the non-malicious user as well as the intricacies of the users in general.

5.2 Test Setup

The test set up for experimental evaluation consisted of three Pentium machines with Linux Operating System. The experiments were conducted with IDSs, PHAD (2001), ALAD (2002), and Snort (1999), distributed across the single subnet observing the same domain. PHAD, is based on attack detection by extracting the packet header information, whereas ALAD is application payload-based, and Snort detects by collecting information from both the header and the payload part of every packet on time-based as well as on connection-based manner. This choice of heterogeneous sensors in terms of their functionality was to exploit the advantages of fusion IDS Bass (1999). The PHAD being packet-header based and detecting one packet at a time, was totally unable to detect the slow scans. However, PHAD detected the stealthy scans much more effectively. The ALAD being content-based has complemented the PHAD by detecting the Remote to Local (R2L) and the User to Root (U2R) with appreciable efficiency. Snort was efficient in detecting the Probes as well as the DoS attacks.

The weight analysis of the IDS data coming from PHAD, ALAD, and Snort 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 whereas the Snort produces the IP address along with severity 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 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 dataset Lippmann (1987). The values chosen for the initial weights lie in the range of -0.5 to 0.5 and the final weights after

training may also be of either sign. The learning rate is chosen to be 0.02. 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. If the neural network stops learning before reaching an acceptable solution, a change in the number of hidden nodes or in the learning parameters will often fix the problem. The other possibility is to start over again with a different set of initial weights.

The fusion unit performed the weighted aggregation of the IDS outputs for the purpose of identifying the attacks in the test dataset. It used binary fusion by giving an output value of one or zero depending 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.

5.3 Metrics for Performance Evaluation

The detection accuracy is calculated as the proportion of correct detections. This traditional evaluation metric of detection accuracy was not adequate while dealing with classes like U2R and R2L which are very rare. The cost matrix published in KDD'99 Elkan (2000) to measure the damage of misclassification, highlights the importance of these two rare classes. Majority of the existing IDSs have ignored these rare classes, since it will not affect the detection accuracy appreciably. The importance of these rare classes is overlooked by most of the IDSs with the metrics commonly used for evaluation namely the false positive rate and the detection rate.

5.3.1 ROC and AUC

ROC curves are used to evaluate IDS performance over a range of trade-offs between detection rate and the false positive rate. The Area Under ROC Curve (AUC) is a convenient way of comparing IDSs. AUC is the performance metric for the ROC curve.

5.3.2 Precision, Recall and F-score

Precision (P) is a measure of what fraction of the test data detected as attack are actually from the attack class. Recall (R) on the other hand is a measure of what fraction of attack class is correctly detected. There is a natural trade-off between the metrics precision and recall. It is required to evaluate any IDS based on how it performs on both recall and precision. The metric used for this purpose is F-score, which ranges from [0,1]. The F-score can be considered as the harmonic mean of recall and precision, given by:

$$F\text{-}score = \frac{2 * P * R}{P + R} \tag{38}$$

Higher value of F-score indicates that the IDS is performing better on recall as well as precision.

Attack type	Total attacks	Attacks detected	% detection
Probe	37	22	59%
DoS	63	24	38%
R2L	53	6	11%
U2R/Data	37	2	5%
Total	190	54	28%

Table 1. Attacks of each type detected by PHAD at a false positive of 0.002%

Attack type	Total attacks	Attacks detected	% detection
Probe	37	6	16%
DoS	63	19	30%
R2L	53	25	47%
U2R/Data	37	10	27%
Total	190	60	32%

Table 2. Attacks of each type detected by ALAD at a false positive of 0.002%

5.4 Experimental Evaluation

All the IDSs that form part of the fusion IDS were separately evaluated with the same two data sets; 1) real-world traffic and 2) the DARPA 1999 data set. Then the empirical evaluation of the data-dependent decision fusion method was also observed. The results support the validity of the data-dependent approach compared to the various existing fusion methods of IDS. It can be observed from tables 1, 2 and 3 that the attacks detected by different IDS were not necessarily the same and also that no individual IDS was able to provide acceptable values of all performance measures. It may be noted that the false alarm rates differ in the case of snort as it was extremely difficult to try for a fair comparison with equal false alarm rates for all the IDSs because of the unacceptable ranges for the detection rate under such circumstances.

Table 4 and Fig. 3 show the improvement in performance of the Data-dependent Decision fusion method over each of the three individual IDSs. The detection rate is acceptably high for all types of attacks without affecting the false alarm rate.

The real traffic within a protected University campus network was collected during the working hours of a day. This traffic of around two million packets was divided into two halves, one for training the anomaly IDSs, and the other for testing. The test data was injected with 45 HTTP attack packets using the HTTP attack traffic generator tool called libwhisker Libwhisker (n.d.). The test data set was introduced with a base rate of 0.0000225, which is relatively realistic. The comparison of the evaluated IDS with various other fusion techniques is illustrated in table 5 with the real-world network traffic.
The results evaluated in Table 6 show that the accuracy (Acc.) and AUC are not good metrics with the imbalanced data where the attack class is rare compared to the normal class. Accuracy was heavily biased to favor majority class. Accuracy when used as a performance measure assumed target class distribution to be known and unchanging, and the costs of FP and FN to be equal. These assumptions are unrealistic. If metrics like accuracy and AUC are to be used, then the data has to be more balanced in terms of the various classes. If AUC was to be used as an evaluation metric a possible solution was to consider only the area under

Attack type	Total attacks	Attacks detected	% detection
Probe	37	10	27%
DoS	63	30	48%
R2L	53	26	49%
U2R/Data	37	30	81%
Total	190	96	51%

Table 3. Attacks of each type detected by Snort at a false positive of 0.02%

Attack type	Total attacks	Attacks detected	% detection
Probe	37	28	76%
DoS	63	40	64%
R2L	53	29	55%
U2R/Data	37	32	87%
Total	190	129	68%

Table 4. Attacks of each type detected by Data-dependent Decision Fusion architecture at a false positive of 0.002%

the ROC curve until the FP-rate reaches the prior probability. The results presented in Table 5 indicate that the Data-dependent Decision fusion method performs significantly better for attack class with high recall as well as high precision as against achieving the high accuracy alone.

The ROC Semilog curves of the individual IDSs and the DD fusion IDS are given in Fig. 4, which clearly show the better performance of the DD fusion method in comparison to the three individual IDSs, PHAD, ALAD and Snort. The log-scale was used for the x-axis to identify the points which would otherwise be crowded on the x-axis.

Detector/ Fusion Type	Total Attacks	TP	FP	Precision	Recall	F-score
PHAD	45	10	45	0.18	0.22	0.20
ALAD	45	18	45	0.29	0.4	0.34
Snort	45	11	400	0.03	0.24	0.05
OR	45	28	470	0.06	0.62	0.11
AND	45	8	29	0.22	0.18	0.20
SVM	45	23	94	0.2	0.51	0.29
ANN	45	25	131	0.16	0.56	0.25
Data-dependent Decision Fusion	45	27	42	0.39	0.6	0.47

Table 5. Comparison of the evaluated IDSs with various evaluation metrics using the real-world data set

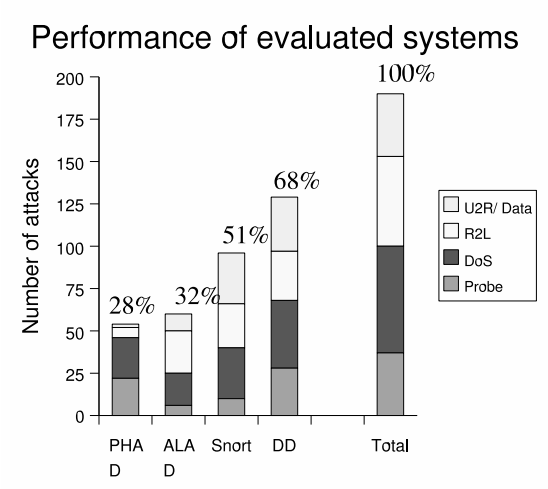


Fig. 3. Performance of Evaluated Systems

Detection/ Fusion	P	R	Acc.	AUC	F-Score
PHAD	0.35	0.28	0.99	0.64	0.31
ALAD	0.38	0.32	0.99	0.66	0.35
Snort	0.09	0.51	0.99	0.75	0.15
Data-Dependent fusion	0.39	0.68	0.99	0.84	0.50

Table 6. Performance Comparison of individual IDSs and the Data-Dependent Fusion method

6. Conclusion

A discussion on the mathematical basis for sensor fusion in IDS is included in this chapter. This study contributes to fusion field in several aspects. Firstly, considering zero knowledge about the detection systems and the traffic data, an attempt is made to show the improved performance of sensor fusion for intrusion detection application. The later half of the chapter takes into account the analysis of the sensor fusion system with a knowledge of data and sensors that are seen in practice. Independent as well as dependent detectors were considered and the study clarifies the intuition that independence of detectors is crucial in determining the success of fusion operation. If the individual sensors were complementary and looked at different regions of the attack domain, then the data-dependent decision fusion enriches the analysis on the incoming traffic to detect attack with appreciably low false alarms. The approach is tested with the standard DARPA IDS traces, and offers better performance than any of the individual IDSs. The individual IDSs that are components of this architecture in this particular work were PHAD, ALAD and Snort with detection rates 0.28, 0.32 and 0.51 respectively. Although the research discussed in this chapter has thus far focused on the three

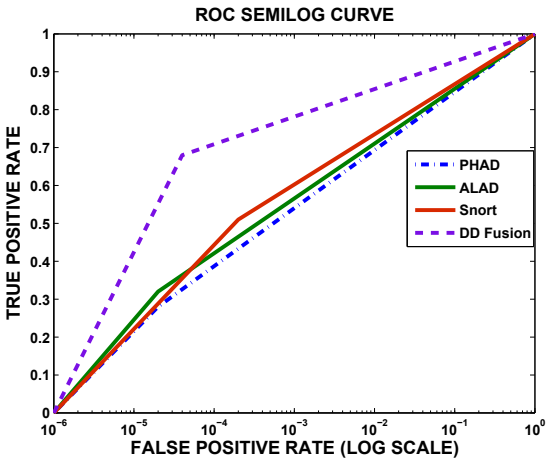


Fig. 4. ROC Semilog curve of individual and combined IDSs

IDSs, namely, PHAD, ALAD and Snort, the algorithm works well with any IDS. The result of the Data-dependent Decision fusion method is better than what has been predicted by the Lincoln Laboratory after the DARPA IDS evaluation. An intrusion detection of 68% with a false positive of as low as 0.002% is achieved using the DARPA data set and detection of 60% with a false positive of as low as 0.002% is achieved using the real-world network traffic. The figure of merit, F-score of the data-dependent decision fusion method has improved to 0.50 for the DARPA data set and to 0.47 for the real-world network traffic.

7. References

Aalo, V. & Viswanathan, R. (1995). On distributed detection with correlated sensors: Two examples, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. 25**(No. 3): 414–421.

ALAD (2002). Learning non stationary models of normal network traffic for detecting novel attacks, *SIGKDD*.

Baek, W. & Bommareddy, S. (1995). Optimal m-ary data fusion with distributed sensors, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. 31**(No. 3): 1150–1152.

Bass, T. (1999). Multisensor data fusion for next generation distributed intrusion detection systems, *IRIS National Symposium*.

Blum, R., Kassam, S. & Poor, H. (1995). Distributed detection with multiple sensors - part ii: Advanced topics, *Proceedings of IEEE* pp. 64–79.

Brown, G. (2004). Diversity in neural network ensembles, *PhD thesis* .

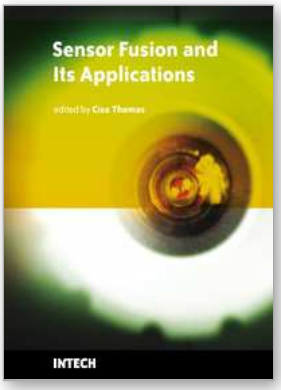
Chair, Z. & Varshney, P. (1986). Optimal data fusion in multiple sensor detection systems, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. 22**(No. 1): 98–101.

DARPA-1999 (1999). `http://www.ll.mit.edu/IST/ideval/data/data_index.html`.

Drakopoulos, E. & Lee, C. (1995). Optimum multisensor fusion of correlated local, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. 27**: 593–606.

Elkan, C. (2000). Results of the kdd'99 classifier learning, *SIGKDD Explorations*, pp. 63–64.

Fausett, L. (2007). *My Life*, Pearson Education.
Hall, D. H. & McMullen, S. A. H. (2000). *Mathematical Techniques in Multi-Sensor Data Fusion*, Artech House.
Kam, M., Zhu, Q. & Gray, W. (1995). Optimal data fusion of correlated local decisions in multiple sensor detection systems, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. 28**: 916–920.
Kendall, K. (1999). *A database of computer attacks for the evaluation of intrusion detection sytsems*, Thesis.
Krogh, A. & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning, *NIPS* (No.7): 231–238.
Libwhisker (n.d.). rfp@wiretrip.net/libwhisker.
Lippmann, R. (1987). An introduction to computing with neural nets, *IEEE ASSP Magazine*, pp. 4–22.
Mahoney, M. & Chan, P. (2003). An analysis of the 1999 darpa /lincoln laboratory evaluation data for network anomaly detection, *Technical Report CS-2003-02*, Publisher.
McHugh, J. (2000). Improvement in intrusion detection with advances in sensor fusion, *ACM Transactions on Information and System Security* **Vol. 3**(4): 543–552.
Nahin, P. & Pokoski, J. (1980). Nctr plus sensor fusion equals iffn or can two plus two equal five?, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. AES-16**(No. 3): 320–337.
PHAD (2001). Detecting novel attacks by identifying anomalous network packet headers, *Technical Report CS-2001-2*.
Snort (1999). www.snort.org/docs/snort_htmanuals/htmanual_260.
Thomas, C. & Balakrishnan, N. (2007). Usefulness of darpa data set in intrusion detection system evaluation, *Proceedings of SPIE International Defense and Security Symposium*.
Thomas, C. & Balakrishnan, N. (2008). Advanced sensor fusion technique for enhanced intrusion detection, *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, IEEE, Taiwan.
Thomas, C. & Balakrishnan, N. (2009). Improvement in intrusion detection with advances in sensor fusion, *IEEE Transactions on Information Forensics and Security* **Vol. 4**(3): 543–552.
Thomopoulos, S., Vishwanathan, R. & Bougoulias, D. (1987). Optimal decision fusion in multiple sensor systems, *IEEE Transactions on Aerospace and Electronic Systems* **Vol. 23**(No. 5): 644–651.



Sensor Fusion and its Applications

Edited by Ciza Thomas

ISBN 978-953-307-101-5
Hard cover, 488 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

This book aims to explore the latest practices and research works in the area of sensor fusion. The book intends to provide a collection of novel ideas, theories, and solutions related to the research areas in the field of sensor fusion. This book is a unique, comprehensive, and up-to-date resource for sensor fusion systems designers. This book is appropriate for use as an upper division undergraduate or graduate level text book. It should also be of interest to researchers, who need to process and interpret the sensor data in most scientific and engineering fields. The initial chapters in this book provide a general overview of sensor fusion. The later chapters focus mostly on the applications of sensor fusion. Much of this work has been published in refereed journals and conference proceedings and these papers have been modified and edited for content and style. With contributions from the world's leading fusion researchers and academicians, this book has 22 chapters covering the fundamental theory and cutting-edge developments that are driving this field.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ciza Thomas (2010). Mathematical Basis of Sensor Fusion in Intrusion Detection Systems, Sensor Fusion and its Applications, Ciza Thomas (Ed.), ISBN: 978-953-307-101-5, InTech, Available from: http://www.intechopen.com/books/sensor-fusion-and-its-applications/mathematical-basis-of-sensor-fusion-in-intrusion-detection-systems

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821