I.J. Information Engineering and Electronic Business, **2009, 1, 32-40**
Published Online October 2009 in MECS (http://www.mecs-press.org/)

A Data-Fusion-Based Method for Intrusion Detection System in Networks

Xiaofeng Zhao
Department of Information Management, Hebei University of Engineering, HanDan, China
Qboy_best@163.com

Hua Jiang, LiYan Jiao
Department of Information Management, Hebei University of Engineering, HanDan, China
Department of Medical Science , Hebei University of Engineering, Handan, China
hd_jianghua@126.com

***Abstract*—Hackers' attacks are more and more intelligent, which makes it hard for single intrusion detection methods to attain favorable detection result. Therefore, many researches have carried out how to combine multiple security measures to provide the network system more effective protection. However, so far none of those methods can achieve the requirement of the practical application. A new computer information security protection system based on data fusion theory is proposed in this paper. Multiple detection measures are "fused" in this system, so that it has lower false negatives rate and false positive rate as well as better scalabilities and robust.**

***Index Terms*—intrusion detection system; data fusion; D-S theory**

I. INTRODUCTION

After being put into research in the 1980s，intrusion detection has become one of the hot pots in the field of computer security. Over the past several decades, research has been focused on the searching for an efficient detection method. And, many algorithms and methods are put forward or introduced to this field. However, so far none of those methods can achieve the requirement of the practical application. The main reason is that network intrusions are launched by "real person", and are more complex than other destructive actions (such as computer virus). So, all the single detection methods can only be effective to certain intrusions, but not ideal for others, which makes the high false negatives rate and false positive rate.

In view of the problem of single detecting method at present, we hope to find a way to get diverse intrusion detection methods combined, so that they can cooperate together to process the detection.

After being put forward in the last century and decades of rapid development, data fusion technology has been widely used in the military, geological and chemical industry. And that multi-sensor data fusion has become an important method to analyze the large scale of heterogeneous data in the complex systems. In the multi-sensor data fusion system, the multiple sensors can gain more targets' information, and using these information and data appropriately can improve the system's measurement accuracy, enhance the fault-tolerance, improve its stability and reliability, and ultimately improve the system's overall performance.

In recent years, some scholars try to apply data fusion technology to IDS to cover the shortage of single methods and to improve the detection accuracy of IDS [1,2]. A successful fusion system consists of two important issues: the selection of the base detectors[3,4], and the fusion mechanism[5]. The latter plays the key role, and many fusion mechanisms were put forward[6,7], which are usually put into two categories: Winner-take-all type and Weighted sum. The Winner-take-all type includes Majority vote, Weighted Majority vote, Behavior Knowledge Space, Naïve-Bayes combination and Dempster-Shafer combination, since they all have a decision for each base detector and the final decision is the one with the highest measurement value. For Weighted sum type such as average and neural network, each base classifier is given a weight which depends on the ability of individual base detector. A weight of each base detector is computed, and then final decision is given by summing up their outputs with the weights. AKI P.F.CHAN[8] has compared the methods mentioned above, and the results are shown in Table 1,2.

It is shown in table 1 and 2 that the accuracies (false alarm rates) of neural network and Dempster-Shaffer are much higher(lower) than other methods. But the scalability of neural network is not good, and the decision-making is too complicated. For Dempster-Shaffer theory, the definition of frame of discernment is too difficult. Therefore, both the methods are seldom used in intrusion detection area.

Manuscript received March 11, 2009; revised June 12, 2009; accepted August 12, 2009.

Table I. AVERAGE TESTING ACCURACY USING DIFFERENT FUSION METHODS ON KDDCUP'99 DATASET[8].

Fusion Methods	Average Testing Accuracy
Neural Network	99.59%
Dempster-Shaffer	99.08%
Weighted Majority Vote	80.66%
Majority Vote	80.09%
Average	80.01%
Navie-Bayes	79.86%

Table II. AVERAGE FALSE ALARM RATE USING DIFFERENT FUSION METHODS ON KDDCUP'99 DATASET[8]

Fusion Methods	Average False Alarm Rate
Neural Network	0.63%
Dempster-Shaffer	0.71%
Weighted Majority Vote	1.91%
Majority Vote	2.27%
Average	2.29%
Navie-Bayes	4.99%

In this paper, a data fusion based intrusion detection model is put forward, in which the detection process is divided into 3 levels: basic detection level, information level and knowledge level. And then an input matrix is put forward to introduce the D-S theory to the information level of the whole detection, so that the results of different detecting methods and heterogeneous data in the system can be "fused" together. Further more, the intrusion scenario and the system's security situation can be extracted at a higher level.

II. HIERARCHICAL MODEL OF THE DATA FUSION BASED INTRUSION DETECTION SYSTEM

A 3-level structure which is generally accepted by scholars [9] is adopted in this paper. The whole system contains three layers: the basic detection layer, information layer and knowledge layer (as shown in the Fig. 1).

A. The basic detection layer

Various IDS agents are arranged in the basic detection layer, different agents adopt different detecting methods to give detection to the collected system information. Each detecting agent can be an independence intrusion detection system (For example: Snort), one kind of detection method (For example: SVM), or other computer security system (For example: Firewall). And, all the detecting results are transmitted to the data fuse module of upper layer to carry out the fusing. Though, each agent can only make a partial judgment to the system's security situation, combining the detecting information of all basic agents can provides sufficient and all-round safety information for the upper layer's fusing module, which is because the focus points of each kind of basic detecting agents are different (For example: some agents adopt misuses detection but others adopt anomaly detection, or, some agents adopt host intrusion detection but others adopt network-based intrusion detection).

B. The calibration layer

Because the system uses a variety of detection methods, and different methods may generate different output formats, for example: the misuse detection gives a probability of whether an intrusion happens, but the results of anomaly detection is the probability of whether there exists some kind of anomaly state in the system. So the output of these different agents must be integrated into a unified format for upper layers to fuse. In addition, this layer should also include a certain amount of filters to remove the information which is incomplete or can not be processed by the fusing agents.

C. The information layer

Different basic detection agents may make different judgment from different angle to a certain intrusion, for

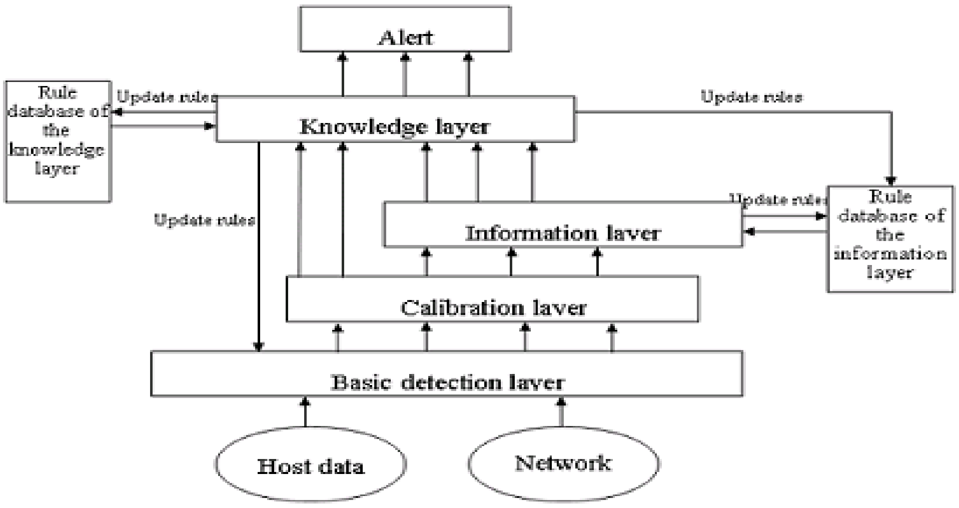


Figure 1. Hierarchical model of the data-fusion based intrusion detection system

example: for an U2R attack, the Anomaly Detection agent using state transition technology may generate alarms according to the illegal state transition of a certain user, and it just knows that the some anomalies exist in the system , but doesn't know these anomalies are occurred from which intrusion behavior; A rule-based misuse agent may detect that a U2R intrusion happens in the system correctly; A SVM based classification detection engine may mistaken it for a R2L intrusion because there is not enough obvious features; And network-based intrusion agent don’t generate alarms because it cannot detect the intrusion behavior. The most important task of the information layer is to give a reasonable and effective evaluation to the judgments of different kinds of agents to the same event, and finally give a correct decision-making.

D. The knowledge layer

Although the information layer fusion consumedly reduced the amount of original alarms, and the correctness of the decision has been greatly improved than the original decision-making, but the number of alarms is still too much for the system’s administrator, and the decision is still on a relatively low level. So, the result of the information layer still need to be further extracted, so that the administrator can not only acquire system’s security situation from a higher level, but also make the intrusion scenario clear. Therefore, a knowledge layer is intercalated into the system to process the information layer’s output, and to get a more precise and comprehensive understanding to the system’s security situation.

III. OUR FUSING METHOD IN THE INFORMATION LAYER

The most important section of the whole system is how the output of the basic layer be fused in the information layer. Whether the fusing method of the information layer is good enough will directly influence the effect on the whole system’s examination.

A successful fusion method for intrusion detection should meet the requirement as follows:

(1) Should use the information of basic detection layer as comprehensive as possible.

(2) Should have high fusion accuracy.

(3) Can process different type of input data.

(4) Should have good scalability.

(5) Fusion algorithm should have high executive efficiency.

As mentioned in section 1, so far, among all the fusion methods only neural net and D-S theory have high fusion accuracy, but the scalability and executive efficiency of neural net is not good.

The D-S evidence [10] theory has been applied to many fields of data fusion widely. And it can meet all the requirement of intrusion detection. However, it is still lack of researches to apply D-S evidence theory to the intrusion detection system so far. The main reason is that the frame of discernment is difficult to define. To solve the problem, we propose a new method, so that D-S evidence theory is introduced into the fuse method of information layer in this paper. And we wish that it would be brought to the attention by researchers working in the field of intrusion detection.

A. Dempster-Shafer evidence theory

The D-S framework is based on the view that proposition can be regarded as subsets of a given set of hypotheses. For example, in the intrusion detection system, we can regard the set of hypotheses as the set of categories of intrusion. Each anomalous event, then, is a subset of Θ. Thus, the propositions of interest are in a one-to-one correspondence with the subsets of Θ, and the set of all propositions corresponds to the set of all subset of Θ, which is denoted 2^{Θ}. Θ is named a frame of discernment, and the proposition are said to be discerned by the frame.

Definition 1: Basic Probability Assignment

Beliefs can be assigned to propositions to express their uncertainty. The beliefs are usually computed based on a density function m：$2^{\Theta}\rightarrow[0,1]$ called a basic probability assignment(bpa) or mass function:

$$\sum\{m(A) \mid A \subseteq \Theta\} = 1 \quad m(\phi) = 0 \tag{1}$$

m(A) represents the belief exactly committed to A.

Definition 2: Belief Function

Given a body of evidence with bpa m, we can compute the total belief provided by the body of evidence for a proposition. This can be done by a belief function Bel: $2^{\Theta}\rightarrow [0, 1]$ defined upon m as follows:

$$Bel(A) = \sum_{B\subseteq A} m(B) \tag{2}$$

Bel(A) is the total belief committed to A, that is, the mass of A itself plus the mass attached to all subsets of A. Bel(A) can he interpreted as follows:

i) Bel(A)=0 means that we have no knowledge about A or that A is false,

ii) Bel(A)=l means that A is true,

iii) 0<Bel(A)<l means that the evidence provides partial support for A.

Definition 3: Plausibility Function

Let Bel: $2^{\Theta}\rightarrow [0, 1]$ be the belief function, the plausibility function is defined as

$$\text{PI(A)=1-Bel(A),} \quad \text{for all } A\subseteq\Theta$$

PI(A) is called the plausibility of A, which quantifies the strength how we don not doubt A or A is reliable.

Theorem 1: Set Θ be the frame of discernment, then the Bel: $2^{\Theta}\rightarrow[0,1]$ is a Belief Function if and only if:

(1) Bel(ϕ) = 0,

(2) Bel(Θ) = 1,

(3) To any natural number n， A1， A2， …， An$\subseteq\Theta$.

$$Bel(A_1\cup A_2\cup\cdots\cup A_n)\geq\sum_i Bel(A_i)-\sum_{i>j} Bel(A_i\cap A_j)+\cdots+(-1)^n Bel(A_1\cup A_2\cup\cdots\cup A_n)$$

Theorem 2: Dempster’s rule of combination.

Dempster’s rule of combination represents the conjunctive operation of the evidence. Given several belief functions on the same frame of discernment based

	A_1	A_2	...	A_j	...	A_s	$\bar{A}_j$	$Normal$
$m1$	$m1(A_1)$	$m1(A_2)$	...	$m1(A_j)$	...	$m1(A_s)$		$m1(Normal)$
$m2$			...	$m2(A_j)$	...		$m2(\bar{A}_j)$	
...	...	...	...	...	...	...	...	...
mi		$mi(A_2)$	...		...		$mi(\bar{A}_j)$	
$m(i+1)$	$m(i+1)(A_1)$	$m(i+1)(A_2)$	...	$m(i+1)(A_j)$	...	$m(i+1)(A_s)$		$m(i+1)(Normal)$
...	...	...	...	...	...	...	...	...
mn	$mn(A_1)$	$mn(A_2)$	...	$mn(A_j)$	...	$mn(A_s)$		$mn(Normal)$

Figure 2. The input matrix of information layer

In which, i＝n2+1， n=n2+n3+1, and the value of m2 to mi depends on the specific

on the different evidences, if they are not entirely conflict, we can calculate a belief function using Dempster's rule of combination. It is called the orthogonal sum of the several belief functions.

The orthogonal sum Bel1 ⊕ Bel2of two belief functions is a function whose focal elements are all the possible intersections between the combining focal elements and whose bpa is given by

$$m(A)=\frac{\sum_{A_i\cap B_j=A} m_1(A_i)m_2(B_j)}{1-\sum_{A_i\cap B_j=\phi} m_1(A_i)m_2(B_j)}=\frac{1}{N}\sum_{A_i\cap B_j=A} m_1(A_i)m_2(B_j),\ A\neq\phi$$

$$m(A)=0,\ A=\phi$$

$$N=1-\sum_{A_i\cap B_j=A} m_1(A_i)m_2(B_j)>0 \tag{3}$$

B. Input of the information layer

Corresponding with different basic detection agents, the information layer may get the following input:

(1) 0-1 input.

In this case, the basic detection agents can only tell whether a certain event is an intrusion or not, and usually the basic detection agents using rule-based method give this output. To the data fusion, information will be lost with the largest number by using the method of 0-1 Input. We name it the first kind of input.

(2)Single probability input.

The basic detection agent regards a event as a certain kind of intrusion (or an anomalous thing), and gives the probability of its decision. Mostly we can acquire the results like that when we use the state transition as our checking method. To the data fusion, some information of the basic detection will be lost by using this kind of method, but more information is reserved than use the first kind of input. We name it the second kind of input.

(3)Multi-probability input.

The basic detection agent doesn't make the final decision of one certain event, but it lists the probable classifications with the probability. For example, all the possible classifications of one certain event given by a basic detector are listed below: Neptune, Smurf, Pod, Self ping, Normal and Unknown. And the relative probabilities are: 35%,45%,5%,5%,2%,8%.This kind of input remains the most comprehensive information for making a decision. We name it the third kind of input.

In the information layer, we use the D-S evidence theory to fuse the basic detection agents' outputs.

Suppose in a detection period, there are M basic detection agents detecting a certain IP. And, the numbers of the first, second, third kind of input corresponding with the M basic detection agents are m1，m2，m3. In the M basic detection agents, N agents produce outputs, which means that there are N inputs for this certain IP to the information layer. And, the numbers of the first, second, third inputs are n1, n2, n3(n1＋n2＋n3=N). The intrusion events detected by the basic detection agents are A1,…,As, and As is the unknown events.

We defines the frame of discernment Θ as (A1, A2, …, As, Normal).

For the first kind of input, there are n1i $(\sum_{1\le i\le s} n_{1i}=n_1)$ basic detection agents classify a event as category Ai, then the basic probability assignment of Ai is $m1(A_i)=\frac{n_{1i}}{m_1}$, and the basic probability assignment of As is m1(Normal)= $\frac{m_1-n_1}{m_1}$. Obviously, the Bel of the first kind is a belief function.

For the second kind of input, suppose the basic detection agent j classify a event as category Ai with the probability Pji, and we set $\bar{A_i}=\Theta-A_i$, then the basic probability assignment is mj(Ai)=Pji, mj($\bar{A_i}$)=1－Pji, $2\le j\le n_2+1$. Obviously, the Bel of the second kind is a belief function.

For the third kind of input, suppose the probability distribution of the basic detection agent k is

A1, A2, …, As, Normal

P_{k1}, P_{k2},…, P_{ks}, $P_{kNormal}$

The basic probability assignment mk(Ai)=P_{ki}, $n_2+2\le k\le n_2+1+n_3$. Obviously, the Bel of the third kind is a belief function.

Combining the Basic Probability Assignment of the three kinds of inputs, we can get the input matrix of the information layer as shown in Fig 2:

Using the Dempster's rule of combination, we can fuse the input matrix into new evidence. And the final decision is made by the new evidence. The uncertainty of new evidence is given by the belief interval.

C. Frame of discernment of the system

Shafter pointed out[10] that all the hypothesis should describe all the possibility of problem completely. We demand that when we classify the intrusion of the system, we should classify them as detailed as possible by the method of intrusion. This could lead to the number of intrusions defined by whole system much larger. But in our module we defined the frame of discernment as: the alarm collection produced by all the different detection method in the basic detection layer during a detection period. Compared with all kinds of intrusions defined by system, the kinds and numbers of intrusion in a detection period (such as 30 seconds) are much less. So the hypothesis number of the Frame of discernment will not be very large, which ensures the fuse efficiency.

IV. Reappearance of intrusion scenario

A complete intrusion process is made up of a series of relevant intrusion behaviors, when a illegal(intrusion) event is find, we believe that it is not isolate, and should have its "cause and effect". Generally, we divide intrusions into two categories: destructive attack and invasive attack. Destructive attack intends to destruct the network of victims, and they make the victim network go wrong by malicious methods (DOS is a typical destructive attack). Invasive attack aims at "intrusion", and their ultimate goal is to control the victims' computers to steal victims' secret or conduct other sabotage.

Correspondingly, a complete intrusion process can be divided into 3 stages: Probe, Intrusion conduction and Clearance, and there are many attack methods in each stage. A complete intrusion process is shown in fig 3.

Because hackers' attacks are not isolate events, when a single intrusion behavior that belongs to certain intrusion stage is detected, we can be certain that the intrusion must have passed previous stages. For instance, when a illegal root access is detected, we can confirm that the attack must have passed probe stage and illegal promotion stage.

In order to find out the real intrusion attempt and process, we should further fuse isolate alarms from information layer to series of intrusion scenarios. The definition of intrusion scenario is given below.

Definition 4 intrusion scenario: Set S = {A1,..., An} be alarms of a complete intrusion process, the intrusion scenario is a alarm sequence generated by time-series, which is expressed by A1→A2→… →An.

The fusion process of intrusion scenario is as follows:

All the known intrusions are classified as probe, intrusion conduction and clearance types, and intrusion conduction is further classified as DOS，R2L，U2R，Other types. Correspondingly, set 3 alarm pools: probe，intrusion_conduction，trace_clearance, and set DOS，R2L ， U2R and Other alarm pools in intrusion_conduction pool. For unknown intrusion types, set Unidentified alarm pool, and put all the unknown alarms into it.

For the convenience of system's extensions, first assign all the alarm pools corresponding codes. Set the codes of probe, intrusion_conduction, trace_clearance, unidentified pool be P1, P2, P3, P4, and set DOS, R2L, U2R, Other subpools of intrusion_conduction be P_{21}， P_{22}， P_{23}， P_{24}. Suppose there are n incomplete intrusion scenarios been found. When a new alarm A comes, first determine whether A belongs to those n incomplete scenarios, if it belongs to Si,, and Si becomes a complete intrusion scenarios with A, then report the scenarios to administrator and extract the intrusion pattern to intrusion pattern database, and if Si still be incomplete with A then wait for other new alarms. If A don't belong to any incomplete intrusion scenarios, then look for the most relevant single alarm that don't belong to any intrusion scenarios in each alarm pool, and if the correlation degrees are higher than the defined threshold then generate a new incomplete intrusion scenario, and if all the correlation degrees are lower than the threshold then put it into corresponding alarm pool, and wait for following alarms. When all the alarm pools are full, put the alarms in the pools into the alarm database, and find the potential intrusion scenarios by offline data mining. The process is shown in Algorithm 1.

The key part of this algorithm is the compute of alarms correlation degree. Currently, related algorithm are being widely studied[11,12,13], and it is not the important point of this paper, so we just choose the real time method of literature[14].

A typical complete intrusion scenario is shown in Fig 4.

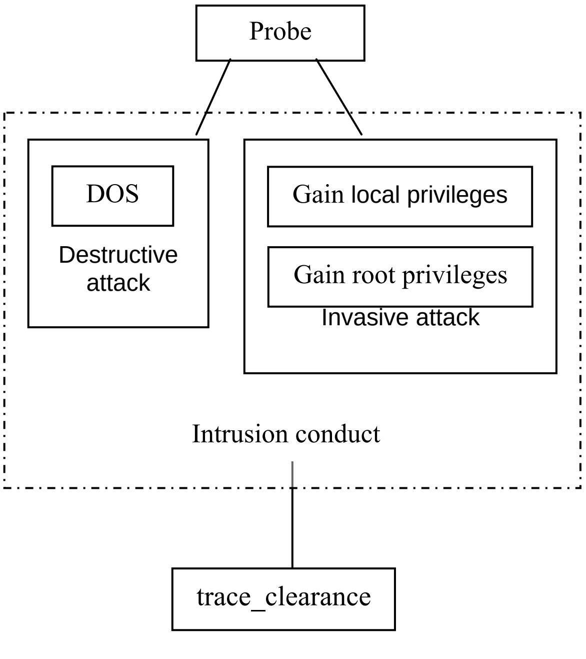


Figure 3. Intrusion stage of hacker

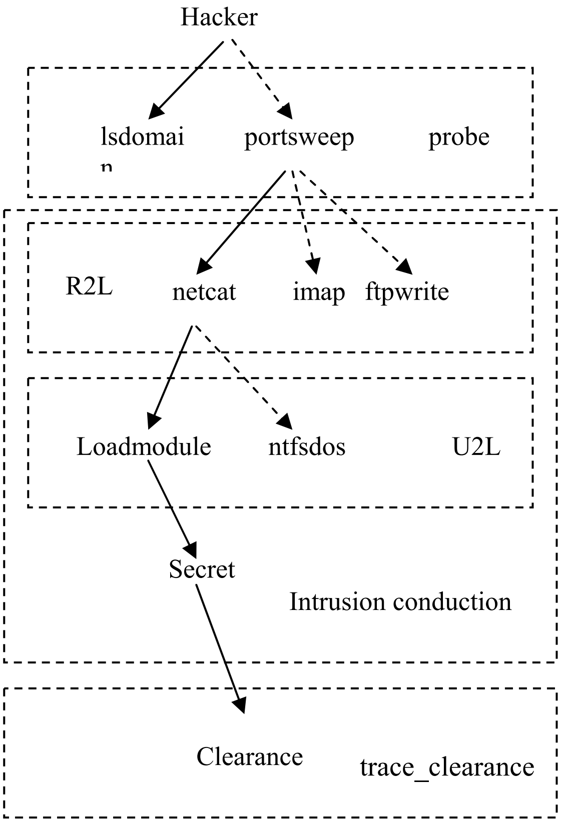


Figure 4. A complete intrusion scenarios

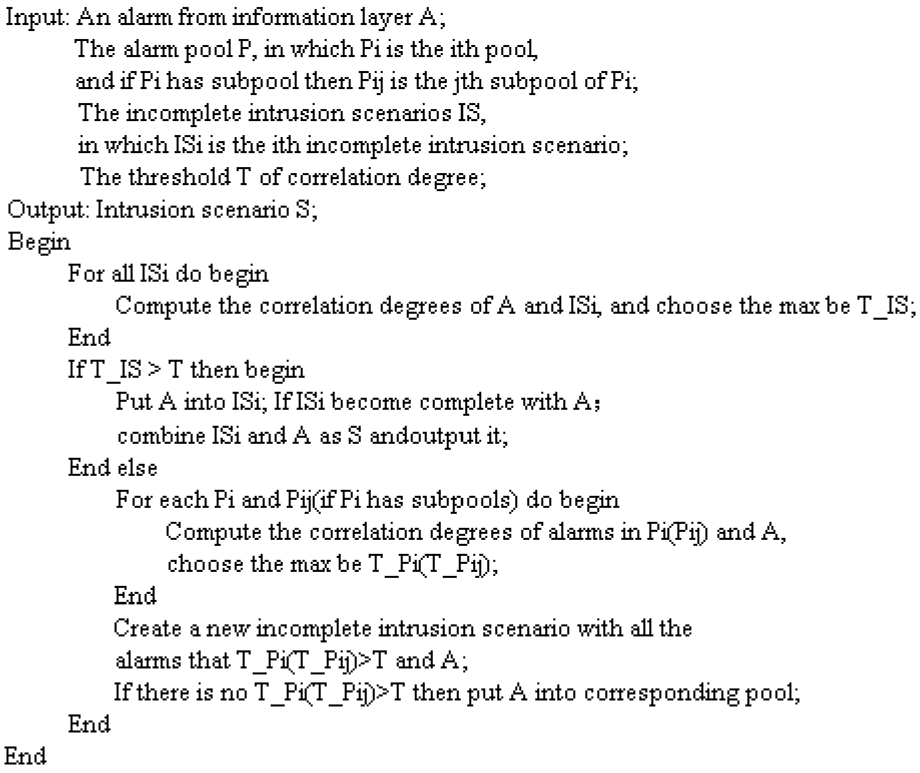
V. Experiments and Analysis

The experiments use the dataset provided by KDD CUP'99[15]. The data set contains 5 million connection records, and each record contains 41 properties. There are four kinds of intrusions: Dos, Probing, R2L, U2R, and a total of 22 attack types. The whole dataset is too large, so generally, only 10% subset is used. It is composed of all the low frequency attack records, the 10% of normal records and high frequency attack records, such as smurf, neptune, portsweep and satan. These four types of the attack records occupy 99.51% of the whole KDDCUP99 dataset and 98.45% of the 10% subset[15].

The basic detecting methods are: C4.5, Naïve Bayes, Neural Network, MDT[16], KNN(K=10), KNN(K=5), SVM.

A. Experiments of correct rate

In this experiment we choose murphy's method [17] as our fusion algorithm. In order to test the reliability and stability, three groups of experiments were carried out. To ensure the independent of basic detectors, for the three experiments, 30000, 20000, 10000 records are chosen randomly from the training data set to train the basic detectors. And a 30000 records set are chosen randomly from the test data set for test. In each experiment, we first use the trained basic detectors to process the test records, and then fuse the outputs of the basic detectors using the method in this paper. The results of the three experiments are shown in the following three tables, and the data in the table is the detection correctness of each detector and our method to different intrusions (Dos, Probe, U2R, R2L) and normal records (Normal).



```
Input: An alarm from information layer A;
       The alarm pool P, in which Pi is the ith pool,
       and if Pi has subpool then Pij is the jth subpool of Pi;
       The incomplete intrusion scenarios IS,
       in which ISi is the ith incomplete intrusion scenario;
       The threshold T of correlation degree;
Output: Intrusion scenario S;
Begin
    For all ISi do begin
        Compute the correlation degrees of A and ISi, and choose the max be T_IS;
    End
    If T_IS > T then begin
        Put A into ISi; If ISi become complete with A;
        combine ISi and A as S andoutput it;
    End else
        For each Pi and Pij(if Pi has subpools) do begin
            Compute the correlation degrees of alarms in Pi(Pij) and A,
            choose the max be T_Pi(T_Pij);
        End
        Create a new incomplete intrusion scenario with all the
        alarms that T_Pi(T_Pij)>T and A;
        If there is no T_Pi(T_Pij)>T then put A into corresponding pool;
    End
End
```

Algorithm 1: intrusion scenario reappearance algorithm

Table III. COMPARISON TABLE 1 OF THE D-S METHOD TO THE BASIC

Table IV. DETECTION METHODS

	Normal	Dos	Probe	U2R	R2L
C4.5	98.9%	93.5%	74.7%	40.0%	7.0%
Bayes	96.7%	73.9%	89.1%	40.0%	24.1%
Neural	99.2%	91.5%	85.6%	20.0%	7.0%
MDT	86.3%	89.4%	42.8%	20.0%	39.0%
D-S	99.1%	93.7%	90.0%	20.0%	10.3%

Table V.COMPARISON TABLE 2 OF THE D-S METHOD TO THE BASIC DETECTION METHODS

	Normal	Dos	Probe	U2R	R2L
C4.5	97.8%	91.6%	55.0%	37.4%	12.5%
Bayes	96.7%	73.8%	81.2%	20.5%	24.4%
Neural	99.5%	74.1%	59.0%	20.0%	8.0%
MDT	85.7%	89.8%	43.7%	20.0%	27.6%
D-S	98.9%	90.7%	69.7%	20.0%	10.3%

Table VI. COMPARISON TABLE 3 OF THE D-S METHOD TO THE BASIC DETECTION METHODS

	Normal	Dos	Probe	U2R	R2L
C4.5	90.8%	87.5%	36.7%	21.0%	14.6%
Bayes	89.5%	66.6%	9.6%	15.0%	13.1%
Neural	88.1%	92.6%	0.0%	15.6%	21.9%
MDT	75.9%	73.2%	43.7%	29.7%	8.5%
D-S	93.2%	88.7%	39.2%	21.0%	11.0%

All kinds of basic detectors in the first experiment are trained well, in which C4.5 has the best result to detect DOS intrusion, however, the result of Naïve Bayes, Hybrid Neural Network and MDT are not so good. The fused result is more close to the result of C4.5. To Probe, all of the four detector's detecting results cannot exceed 90%, and the result of D-S exceeds 90%, but the last two kinds of intrusions merely appear in the training data set, for example, the R2L intrusion only appears hundreds of time in the training data set, however we must select training record randomly to ensure the result be general. That is the reason why the last two kinds of intrusions' frequency of occurrence is so low, which made the detecting result of all detectors on them are not perfect.

We can see from the tables, each basic detector has higher detected ratio to certain kinds of intrusions, but is not perfect to others, for example, C4.5 algorithm has better correct rate to Normal and DOS, but not to Probe. However , the correct rates of fused results to every kind of intrusions are all close to, or even higher than, the highest correct rates of all basic detectors, which makes the system has higher correct rate to all intrusions , so this makes up for the flaw that single detecting method cannot have good detecting result to all intrusions. Moreover, the correct rates of the Normal in the three experiments are all close to, or higher than, the highest correct rates of all basic detectors, especially in the third experiment, the results of basic detectors to the category Normal are all not ideal, but the fused correct rate is 93.2%, that means the fusion makes a lower false positive rate.

When a small number of the results of basic detectors is not ideal, it will not affect the result of the fusion greatly(For example, in the first experiment, for the Bayes method, the correct rate of Probe is relatively low, but the fusion result surpasses the best results of all basic detectors). That makes the whole system more stable, and the damage or false alarm of a few basic detectors may not influence the detecting results of the system. But, if the majority results of the basic detectors are not ideal, it is hard to get better fusion results (For example, in the third experiment, the result of Probe). In short, the D-S fusion algorithm in this paper can mutually make up respective deficiency of the basic detectors in the system, and the results of detection are more stable. However, the results of the fusion still depends on the quality of basic detectors, if there exist too many low performance detectors, even if using the fusion it is also very difficult to get satisfactory results.

B. The influence of the basic detector's number to the results

In order to test the influence of the basic detectors' number to the results of the fusion, 6 detection algorithm are adopted: C4.5, Naive Bayes, Hybrid Neural Network, MDT, KNN (K = 10), KNN (K = 5), SVM.

A total of five experiments are carried out. In the first experiment we choose the first two methods as the basic detectors. And then, in each following experiment a new detecting method is added to the system as the basic detector. Training and test are same as the first experiment in the section 5.1. The results of the experiment are shown in the figure 5.

As can be seen from the figure 5, when only two basic detectors are used, the fusion result is not very good, but with the increasing of the basic detectors' number the fusing correct rate increases. For the third experiment, the fusion result of Probe intrusion doesn't increase, that is because the detecting result of the added MDT algorithm to Probe intrusion is not ideal, which improves the dependence of the fusion to the basic detectors.

The average fusion time for 10000 test records is: 0.02s for 2 basic detectors, 0.03s for 3 detectors, 0.04s for 4 detectors, 0.06s for 5 detectors and 0.09s for 6 detectors. But in the experiment, there are only 4 intrusions (and Normal), so the test for time performance does not have practical significance. And in the practical application, the time performance is still needed to be further proven.

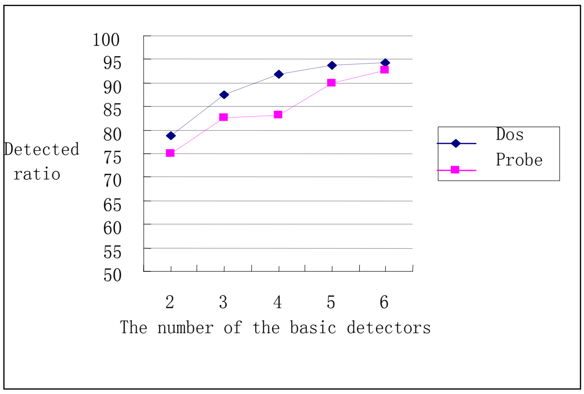


Figure 5. The influence of the basic detector's number to the results

VI.The architecture for practical application

A muti-agents based architecture of tree structure is designed for practical application as shown in Fig 6. The main reason of the chosen of tree structure is that it can match the organizational structure of enterprises or other institutions well. We can see form Fig 6 that the whole architecture is divided into 3 layers corresponding to the 3 layers of hierarchical model, and in each layer certain agent is placed. The function of each detection and fusion agent is the same as the function of corresponding layer in the hierarchical model. We name fusion agent of knowledge layer FKA, fusion agent of information layer FIA, basic detection agent BDA and registration agent RGA.

The concept of "domain" is introduced into the model as definition 5.

Definition 5 Domain: In the tree structure of the architecture, a domain is a subtree with a certain FKA as the root. The whole system is called the global domain.

A domain can have several BDAs and FIAs, but only one FKA, and protects a department of certain corporation or institution. For security considerations, components of the same domain can trust each other and exchange information directly, but components of different domains don't trust each other and the information exchange must through their root nodes.

There is a RGA in each domain. The RGA collect all the information of agents in the domain. When an agent need some information, it asks the RGA which agents have the information, and contact them directly to get the information.

The detecting procedure is as follows:

When initializing, the RGA broadcasts to the whole domain, and ask all the agents report their information, including: the name of the agent, the information it has, the information it need and the detecting method etc. When an agent need some information, it asks the RGA which agents have the information, and contact them directly to get the information. The RGA of the global domain has the information of the whole system, and if a agent cannot get the information from its own domain, it seek the information from the RGA of the global domain.

BDA sends its detecting result to corresponding FIAs. FIAs fuse all the information and give the final decision of whether there is certain intrusion in the system, and send their decision to the FKA. FKA analyses all the decision and gives the intrusion scenario and security situation from a higher level. The FKA of the global domain analyses the intrusion scenario and security situation of the whole system. The detection agents and fusion agents should have sel-flearning functions, so the upper nodes should feedback the detecting results and new intrusions to lower nodes, so that they can update their rule database.

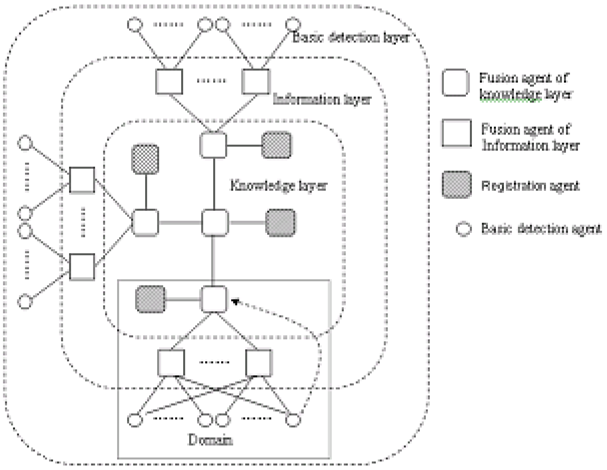


Figure 6. The architecture of the system

VII. Conclusion

In this paper, a data fusion-based intrusion detection model is introduced, and the intrusion detection problem is converted into a abstracting process that abstracts the system's information from the low-level to high-level, further more points out that different basic detectors' output should be fully fused in order to get the detection more accurate. And the D-S evidence theory is introduced to the information layer of the model.

The experiments shows: by fusing, the final detected rate of the entire system to every intrusions will be higher than or close to the best detected rate of all the basic detectors to them, which makes the entire system has a relative high detected rate to all intrusions, and makes up for the flaw that single detecting method cannot has good detecting result to all intrusions. However, the experiments also show that the validity of existing D-S Fusion algorithms also partly depends on the validity of the basic detectors, and the searching for a more stable D-S fusion algorithm is one of our major research directions.

References

[1] Tim Bass, "Intrusion detection systems and multisensor data fusion", Communications of the ACM, Vol.43, April 2000, pp.99-105.

[2] Tim Bass, Silk Road, "Multisensor data fusion for next generation distributed intrusion detection systems", IRIS National Symposium Draft, 1999, pp.24-27.

[3] L.I.Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment", IEEE Trans.on Systems, Man and Cybernetics, 2002, Vol.32, pp.146-156.

[4] K.Tumer and J.Ghosh, "Error correlation and errorreduction in ensemble classifiers", Connection Science, 1996Vol.8, pp.385-404.

[5] F.Roli, G.Giacinto and G.Vernazza, "Methods for designing multiple classifier systems", MCS, LNCS2096, 2001, pp.78-87.

[6] Klein, L.A.. "A Boolean algebra approach to multiple sensor voting fusion", IEEE Trans Acrosp. Electron Syst, 2004, pp.317-327.

[7] A. P. F. Chan, W. W. Y. Ng, D. S. Yeung and E. C. C. Tsang, "Multiple classifier system with feature

grouping for intrusion detection: mutual information approach", To appear in the 9th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, 2005, pp.215-221.

[8] AKI P.F.CHAN, WING W.Y.NG, DANIEL S.YEUNG, "Comparison of different fusion approaches for network Intrusion Detection Using Ensemble of RBFNN", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 2005, pp.18-21.

[9] You He, Guohong Wang, "The applications of multi-sensors data fusion", Beijing: Publishing House of Electronics Industry, 2000, 11.

[10] Shafer, G. "A mathematical theory of evidence", Princeton U.P., Princeton, N .J., 1976.

[11] VALDES A, SKINNER K. "Probabilistic alert correlation", Fourth International Symposium on Recent Advance in Intrusion Detection, 2001. pp.54-69.

[12] Chengpo Mu , Houkuan Huang , Shengfeng Tian, "Intrusion detection alert verification based on multi level fuzzy comprehensive evaluation", In : Proc. 2005 International Conference on Computational Intelligence and Security , Lecture Notes in Artificial Intelligence 3801 , Berlin , 2005, pp, 9-16.

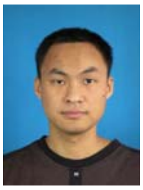
[13] Mu Chengpo, Huang Houkuan , Tian Shengfeng. "Intrusion detection alerts processing based on fuzzy comprehensive evaluation", Journal of Computer Research and Development , 2005 , 42(10),pp.1679-1685.

[14] Soojin Lee, Byungchun Chung. "Real-time analysis of intrusion detection alerts via correlation". Computers & Security. (2006)25, pp.169-183.

[15] KDD CUP 1999 Data . http://www.ics.uci.edu/~kdd/databases/kddcup99/kddcup99.html

[16] Xiaofeng Zhao, Zhen Ye. "Research on weighted multi Random decision tree and its application to intrusion detection". Computer Engineering and Applications. 2007, 27(5), pp.1041-1043.

[17] Grundel D, Murphey R, Paralos P, eds. "Theory and algorithms for cooperative systems", Singapore: World Scientific, 2005, pp.239-310.



Xiaofeng Zhao was born in Han Dan, China, in 1978. He finished the Bachelor degree in computer scienence and technology from the Hebei University of Technology of Tianjin, China in 2000, and the Master degree in computer software and theory from the Hefei University of Technology of Hefei, China in 2007. His field of interest is computer security and data mining.

He joined in Information Management Department, Economics and Management School, Hebei University of Engineering in 2007, and is a Lecturer now. He was a Computer Engineer of Handan Iron and Steel Group from 2000 to 2004.

Hua Jiang, born in 1977-1-9, Handan, Hebei Province, China. In March, 2006, graduated from Hebei University of Engineering and obtained postgraduate qualifications. Main research fields: network security, information management, supply chain management.

Now works in Information Management Department, Economics and Management School, Hebei University of Engineering, Lecturer. Mainly published articles:

Jiang, Hua, "Study on mobile E-commerce security payment system", Proceedings of the International Symposium on Electronic Commerce and Security, ISECS 2008, Aug 3-5 2008, pp.754-757;

Jiang Hua, Ruan Junhu, "Analysis of Influencing Factors on Performance Measurement of the Supply Chain Based on SCOR-model and AHM", IEEE/SOLI'2008; Beijing, China October 12-15, 2008, pp.2141-2146;

Xiaofeng Zhao, Hua Jiang, Liyan Jiao, "A Data Fusion Based Intrusion Detection Model", International Symposium on Education and Computer Science (ECS2009), 7-8 March, 2009 Wuhan, Hubei, China (in press).

Current research interests: management optimization and scientific decision-making.

Liyan Jiao was born in Shijiazhuang, China, in 1978, She is a graduate student in Medical College of Hebei University of Engineering. Her field of interest is medical informatics.