Combine a Population-based Incremental Learning with Artificial Immune System for Intrusion Detection System

Jheng-Long Wu, Pei-Chann Chang and Hsuan-Ming Chen

Abstract—This research focus on the intrusion detection system (IDS) development which using artificial immune system (AIS) with population based incremental learning (PBIL). AIS have powerful distinguished capability to extirpate antigen when the antigen intrude into human body. The PBIL is based on past learning experience to adjust new learning. Therefore we propose an intrusion detection system call PBIL-AIS which combine two approaches of PBIL and AIS to evolution computing. In AIS part we design three mechanisms such as clonal selection, negative selection and antibody level to intensify AIS performance. In experimental result, our PBIL-AIS IDS can capture high accuracy when an intrusion connection attacks.

Keywords—Artificial immune system, intrusion detection, population-based incremental learning, evolution computing.

I. INTRODUCTION

This research is focus on network intrusion problem to build an intrusion detection system. Intrusion detection based on statistical pattern recognition approaches has attracted a wide range of interest in response to the growing demand of reliable and intelligent intrusion detection systems (IDS), which are required to detect sophisticated and polymorphous intrusion attacks. In general, IDS are calling the anomaly detection in the literature. Most detection system can reliably identify intrusion attacks in relation to the known signatures of discovered vulnerabilities.

In traditional detection system, many researcher use statistics method to analysis intrusion problem which is calculating the frequency phenomenon and analyst intrusion attacks probability. The detection approaches such as fuzzy control, artificial neural network, decision tree, SVM and so on which these approaches to solve the anomaly statements has good performance in a lot of anomaly problems [1]–[4]. Recently years, another computationally intelligent approach that Evolutionary Computing (EC) has attracted a wide range of interest in the computer science or artificial intelligence. EC approaches such as artificial immune system (AIS), Genetic Algorithm (GA), Evolutionary Strategy (ES), Genetic Programming (GP), Ant Swarm Optimization (ACO) and Particle Swarm Optimization (PSO) which they impersonate the natural evolution mechanism to solve complex problems[5]–[8]. AIS are a special approach among these EC approaches that it has powerful evolution mechanism according to immune response.

Many researches use AIS to solve the parameter combination optimization [9], [10]. In the AIS evolution which find effective antibody, create new antibody, clonal selection, negative selection and antibody level are very important mechanism [11].

In this research, we want to apply PBIL to induce AIS for improve the AIS effectiveness. According to immune rules of human immune system, we want to use the antibody (detection rules) to extirpate the antigen (intrusion connections). AIS work on detection process which need a learning approach for create new antibody. We want to use PBIL to incremental learning the new antibody which affinity is better than old antibodies at is time.

The main purposes of this research are:
1) we want to develop an intrusion detection system by combines PBIL and AIS.
2) our propose system can provide a network security service when the intrusion connections attacks.

II. LITERATURE REVIEW

A. Intrusion Detection System

The intrusion detection system is a security protection problem when appear sophisticated and polymorphous intrusion attacks. In general, there are two main types of IDS such as network intrusion detection system (NIDS) and host-based intrusion detection system (HIDS). The anomaly detection approach usually uses statistical analysis and pattern recognition to solve. It is able to detect anomaly intrusion without prior knowledge. Therefore the model has the generalization capability to extract intrusion rule during training. The statistical anomaly-based IDS determines what bandwidth is generally used, what ports and devices generally connect to each other and alert the manager when the connection is detected which is anomaly or normal [12].

Most approach to build IDS such as Abadeh et al. [13] refer to uses the genetic fuzzy systems (GFSs) hybrid model to solve intrusion attacks problem. They presented three kinds of genetic fuzzy systems based on Michigan, Pittsburgh and iterative rule learning approach to deal with intrusion detection. Alteaijary and Algarny [14] presented a Bayesian based intrusion detection system which based on Bayesian probability theory to filter the intrusion attacks. Horng et al. [15] presented a hierarchical clustering and support vector machines hybrid model to build an IDS. These three researches presented three approaches to solve the same KDD Cup 1999 dataset. However the intrusion detection dataset of KDD Cup 1999 is very popular. In this paper we use the dataset of KDD Cup 1999 to develop our IDS.
B. Artificial Immune Algorithms

In computer science, artificial immune systems are based on principles and processes that are inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving [16]. AIS approach has high performance compared to artificial neural networks, GAs, fuzzy systems and so on, also it has been successfully applied to many fields such as clustering, classification, pattern recognition, computer defense, optimization, and so on [16]–[22].

C. Population-based Incremental Learning

The concept of incremental learning is similar step learning which can increase toleration when learning process. Many researches refer a lot of incremental learning methods which in this paper is using population based incremental learning (PBIL) approach induct into AIS. The PBIL algorithm, first proposed by Baluja [23], is an evolutionary optimization algorithm, and estimation of distribution algorithm. This is type of genetic algorithm (GA) where explicitly maintains the statistics contained in a GA’s population and the genotype of an entire population is evolved rather than individual members [24]. The PBIL algorithm is as follows: (1) a population is generated from the probability vector. (2) The fitness of each member is evaluated and ranked. (3) Update population probability vector based on fittest individual. (4) Mutate. (5) Repeat steps 1–4. The PBIL has been very successful when compared to standard GAs on a lot of benchmark and real time problems [23], [25], [26].

III. PBIL-AIS INTRUSION DETECTION SYSTEM

This study develops an intrusion detection system to solve network intrusion problem which is called PBIL-AIS Intrusion Detection System (PBIL-AIS-IDS). Our proposed model has three phases: first phase is data collection and preprocessing; second phase is Evolution by PBIL-AIS which there have cone selection mechanism and Negative selection mechanism; third phase is detection the intrusion occurrence from detection rules of Antibody of our model in Fig. 1. The detail process of propose model as follows:

A. Data Collection and Preprocessing

In the step we want to collect the connection records from internet user to form the training data for learning the IDS. Each connection record is a user connect to service system for provide services according to user requirement. According to classification problem each connection will tag the answer label for supervise learning.

B. PBIL-AIS Evolution Flow

We propose a PBIL-AIS IDS for network environment in Fig. 2. There have 6 main steps to learning anomaly intrusion knowledge which are including initial antibody pools, calculate affinity between antibodies and antigen, create new antibody, Clonal Selection mechanism, antibody level mechanism and Negative Selection mechanism.

Step 1: Initial Antibody Pools

In this step we will initial the antibody pool which normal antibody pool and anomaly antibody pool. In each pool initial N populations of antibody are randomly random selection. We want to generate good information in each pool for the AIS evolution.
Step 2: Calculate Affinity between Antibodies and Antigen

In this paper, we call each connection an antigen in our system. The affinity value calculation is from antibody with antigen. There are two data type such as nominal data and continuous data which the nominal data is used the Hamming Distance to get the difference degree; the continuous data is used the Euclidean Distance. Before the affinity calculation we need to transform data scale which only continuous data needed. The normalization transformation as follows:

\[ x_{j,g} = \frac{x_{j,g} - \min(x_j)}{\max(x_j) - \min(x_j)}, \]

where \( x_{j,g} \) denotes normalization value in \( j \)th antibody of \( g \)th feature. \( \min(x_j) \) denotes value in original in \( j \)th antibody of \( g \)th feature. \( \max(x_j) \) denotes minimal value in \( g \)th feature. \( \max(x_j) \) denotes maximal value in \( g \)th feature.

The affinity value between antibody and antigen as follows:

\[ Affinity_j = \frac{1}{h_j + e_j} \quad j = 1,2,3,...,v, \]

where \( Affinity_j \) denotes affinity value in \( j \)th antibody with antigen currently. \( h_j \) denotes the total value of distance as nominal data. \( e_j \) denotes the total value of distance as continuous data. In the nominal data we used this Eq. (3) as follow:

\[ h_j = \sum_{f=1}^{H} w_f I_{j,f}, \quad I_{j,f} = \begin{cases} 0, & \text{if } x_{j,g} = x_{j,g}^{Ag} \\ 1, & \text{if } x_{j,g} \neq x_{j,g}^{Ag} \end{cases}, \]

where \( w_f \) denotes weight value in \( f \)th feature of \( j \)th antibody. \( I \) denotes difference degree value as nominal data. \( H \) denotes a set of nominal data. \( x_{j,g}^{Ag} \) denotes value of \( f \)th feature of antigen.

Continuous data as follows:

\[ e_j = \sqrt{\sum_{g=1}^{E} w_g (x_{j,g} - x_{j,g}^{Ag})^2}, \]

where \( w_g \) denotes weight value in \( g \)th feature of \( j \)th antibody. \( x_{j,g}^{Ag} \) denotes value of \( g \)th feature of antigen.

From the affinity formula, we can know if the value is trend to less that is meaning the antibody with antigen has good affinity. In classification problem we need to know which antibody can extirpate antigen. According to Eq. (2) if class of antibody and antigen are the same then the antigen is extirpated by this antibody which the affinity value of the antibody is highest than other antibodies. Therefore we can define the best antibody \( Ab_{bext} \) is highest affinity currently, the Eq. (5) is the determine condition.

\[ Ab_{bext} = \arg \max_{Class} Affinity^g, \]

Class denotes the classes of anomaly and normal. Sometime the highest affinity antibody \( Ab_{bext} \) is not a correct class so we need to create a new antibody to fit the characteristic of antigen.

Step 3: Create New Antibody by evolution Approach

In this step, we want to learn a new antibody for extirpate antigen which the design concept is from PBIL approach. First we select the highest affinity antibody with antigen from individual antibody pool.

Let the antibody \( Ab^{right}_{bext} \) is the same class of the highest affinity antibody with antigen, on the contrary the antibody \( Ab^{wrong}_{bext} \) is not the same class. How to detect intrusion attacks that we use match approach to know which antibody can matching the antigen. The antibody and antigen match step as follow: first using Memory Cell antibodies to match antigen, if it can match success so the new connection is anomaly intrusion. According to affinity formula we can know the affinity relation between two classes antibody pool is \( Affinity^{Ab_{bext}} > Affinity^{Ab^{wrong}} \).

If Memory Cell antibody cannot match success for antigen then go to next Mature Cell to find can match antigen if else go to Immature Cell. The feature of new antibody \( Ab^{new} \) learning has two methods to calculation. If the feature belongs to nominal data set then used this Eq. (6) to calculation.

\[ x_{j}^{Ab_{bext}} = x_{j}^{Ag} \]

If the feature belongs to continuous data set then use the evolving new feature value. Let \( LR \) is the learning rate; \( N \) is the maximal times of evolution; according to the gradient descent algorithm to set \( LR \) which in front period the learning rate is high and in rear period the learning rate change to low. We suppose the number of \( d \) continuous features, and then the \( g \)th features \( x_{g}^{new} \) in new antibody evolution from Eq. (7) as follows:

\[ x_{g}^{new} = x_{g}^{Ab_{bext}} + LR \times (x_{g}^{Ag} - x_{g}^{Ab_{bext}}), \quad 0 < LR < 0.5 \]

As in Fig. 3, for flowchart, the new antibody evolution is flowing (1) to make out the antibody \( Ab^{right}_{bext} \) and \( Ab^{wrong}_{bext} \). (2) To check out which antibody feature has high information for learning, the mean is scanning which feature difference value between antibody with antigen equal to 0 and maximal. (3) Other antibody features are following PBIL to learn. (4) Recalculate the affinity of the new antibody. (5) If the new antibody is the highest affinity then go to the next step, otherwise go back to step (3). (6) The class of new antibody is defining from antigen class. (7) The new antibody inject to antibody pool according to the antibody pool class.

Step 4: Clonal Selection Mechanism

According to Clonal Selection approch, we need to find the effective feature of antibody to elevate the important weight. This process calls Clonal Expansion. In this paper we design a clonal expansion approach different traditional approach which deep into the antibody feature. After this process the antibody feature can capture high effect on intrusion detection. The clonal selection step as follow:

1. Antibody feature selection: selecting high affinity difference degree feature between antibody and antigen.
2. Clonal expansion: these selected features will expand feature weight which the expansion is fixed value to adjustment.
3. Genetic mutation: using random expansion rate to adjust the clonal expansion.
(4) Recombine antibody feature set: combine these expanded feature and not has selected feature to form weight set. The clonal expansion rate can use Eq. (7) to adjust antibody feature weights. Form Eq. (3) and (4) that the $w_f$ and $w_g$ are thought Eq. (7) to give. Let weight $w_t$ and $w_{new}$ is a weight set with a feature set $WS = \{w_{f1}, w_{f2}, \ldots, w_{fn}\}$ which the feature set including nominal data and continuous data.

\[ w_{new}^t = (1 - \delta) \times w_{old}^t + \Delta \tau_i. \]  

(8)

\[ \Delta \tau_i = \begin{cases} 1/r & \text{if } dw_{ab^{right}} - dw_{ab^{wrong}}, 0 < r < R \quad \text{,} \\ 0 & \text{otherwise} \end{cases} \]

(9)

\[ dw_i = |x_t^Ag - x_t^{Ab}| \]  

(10)

where $w_{new}^t$ denotes new weight of $t$th feature in feature set.
\( \delta \) denotes coefficient. $\Delta \tau_i$ denotes expand rate in $t$th feature. $dw$ denotes difference value of $t$th feature between antibody and antigen. $r$ denotes a random value. $R$ denotes a maximal value of random by user definition. According to formula (8), (9) and (10) we can adjust weights of feature for learning high useful weight set by clonal selection mechanism. For an example see Fig. 4, there have two antibodies as $Ab^{right}$ and $Ab^{wrong}$ which the class of $Ab^{right}$ is same to $Ag$ and the class of $Ab^{wrong}$ is not same to $Ag$. The original affinity of $Ab^{right}$ is 0.39 and $Ab^{wrong}$ is 0.35. The weight set $WS$ of all feature are 1, the $Ab^{right}$ difference set $Ab^{right}_{dw}$ of all feature between $Ab^{right}$ and $Ag$ are $dw^{Ab^{right}} = \{1.0, 0.9, 0.8, 0\}$, the $Ab^{wrong}$ difference set of all feature between $Ab^{wrong}$ and $Ag$ are $dw^{Ab^{wrong}} = \{1.0, 1.2, 0.6, 0\}$. From the Fig. 4 we can see the second feature is the maximal difference that this feature weight thought the clonal expansion process is change from 1 to 1.5. The new weight set is $WS = \{1, 1, 1.5, 1\}$, use the new weight $WS$ to recalculate affinity of two antibodies of $Ab^{right}$ and $Ab^{wrong}$ which new affinity value are 0.32 and 0.29. The affinity difference with two antibodies changes from 0.01 to 0.03. However, if the affinity relation is $\text{Affinity}^{Ab^{right}} < \text{Affinity}^{Ab^{wrong}}$ then go thought clonal selection to adjust the feature weights, we can gave the relation change to $\text{Affinity}^{Ab^{right}} > \text{Affinity}^{Ab^{wrong}}$.

Step 5: Antibody Level Mechanism

In this step we want to design a multiple level antibody pools which are including Memory Cell, Mature Cell and Immature Cell.

If antibody belongs to Memory Cell that denotes it have high-distinguish capability which the usage rate of these antibodies are highest more than Mature Cell antibody. If antibody belongs to Mature Cell that denotes it have mid-distinguish capability which the usage rate least than Memory Cell but more than Immature Cell. If antibody belongs to Immature Cell that denotes it have low-distinguish capability which the usage rate is lower least than Memory Cell and Mature Cell. In antibody level part we also design three cycles for which detection times need to adjust antibody level, the cycle parameters including $\lambda$, $\beta$ and $\alpha$. The relation of detection times is $\alpha < \beta < \lambda$ that meaning is Memory Cell antibody has long existence term among evolution. Mature Cell antibody has middle existence term among evolution. Immature Cell antibody has short existence term among evolution. The detail antibody level rules in TABLE I. There have 8 statements for determine upgrade, demotion and nothing which each Cell have the nothing statement. About the adjustment process of each Cell as follows:

Fig. 3 The flowchart of the proposed new antibody evolution

Fig. 4 The example for clonal selection
usage rate but affinity is high. If the AIS model not considers the antibody. The negative selection in our model is wants find low antibody level which if antibody has high usage rate it is positive process in AIS. The positive selection like the clonal section and immature pool goes to mature pool. Let selection design to antibody upgrade when antibody from negative antibody affect degree; if the cycle are up to mature antibody pool. Negative selection start threshold then the new antibodies in this cycle.

\begin{equation}
\text{Negative}_{\text{imm}} = \text{affinity}_{\text{old}} - \text{affinity}_{\text{new}},
\end{equation}

\begin{equation}
\text{affinity}_{\text{new}} = \sum_{j=1}^{n} \text{affinity}_j, \text{iff } \forall \text{affinity}_j \in \text{immature Ab}
\end{equation}

\begin{equation}
\text{affinity}_{\text{old}} = \sum_{j=1}^{n} \text{affinity}_j, \text{iff } \forall \text{affinity}_j \in \text{immature Ab}
\end{equation}

where Negative\textsubscript{imm} denotes the negative selection start threshold. Affinity\textsubscript{new} denotes the total affinity value of all immature new antibodies in this cycle. Affinity\textsubscript{old} denotes the total affinity value of all immature old antibodies in least cycle.

C. Detect the intrusion for testing

In this part, we want to detect the intrusion attacks which if true intrusion attacks then put forth the alert for the network manager just in time to solve anomaly. In this paper we only use the rules of Memory Cell antibody to find which connections is the intrusion. The Memory Cell has high-distinguish capability for most intrusion connections because the three mechanisms can capture perfect antibodies which they high usage rate and high affinity.

**IV. EXPERIMENTAL RESULTS**

The KDD-Cup99 data set from UCI repository [27] is widely used as the benchmark data for IDS evaluation. In our experiments, we random select 49252 records from its 10% training data consisting of 494021 connection records for training. There have 31124 records for testing as in Fig. 5. Each connection record represents a sequence of packet transmission starting and ending at a time period, and can be classified as normal class and 4 different classes of attacks.

(1) Denial-of-service (DOS): Denial of the service that are accessed by legitimate users, e.g., SYN flooding.
(2) Remote-to-local (R2L): Unauthorized access from a remote machine, e.g., password guessing.
(3) User-to-root (U2R): Unauthorized access to gain local super-user (root) privileges, e.g., buffer overflow attack.
(4) Probing (Probe): Surveillance and probing for information gathering, e.g., port scanning.

In model evaluation approach which we use accuracy to measure the classification effectiveness from different classification techniques, as in Equation 14.

\begin{equation}
\text{Accuracy} = \frac{TP}{TP + TN + FN + FP},
\end{equation}

where TP denotes ture positive (antibody is normal and antigen is positive); FP denotes false positive (antibody is normal but antigen is anomaly); FN denotes false negative (antibody is anomaly but antigen is normal); TN denotes ture negative (antibody is normal but antigen is anomaly).

From the experimental result in TABLE II, our PBIL-AIS IDS has 91.92% accuracy in samll case. Our IDS has good performance compare to other machine learning approach.
In this paper we proposed a PBIL with AIS model for intrusion detection problems in network environment. The PBIL has a good learning process that can improve AIS evolution effect for create new antibody step. In this paper we propose three mechanisms such as clonal selection, negative selection and antibody level to intensify AIS performance is very well. Therefore, a robust intrusion detection system need complete evolution conditions. Experimental results of the PBIL-AIS intrusion detection system demonstrate high performance.

REFERENCES


