The Designation of Bio-Inspired Intrusion Detection System Model in Cloud Computing Based on Machine Learning

Yufei Ge, Hong Liang, Lin Chen, Qian Zhang

College of Computer and Communication Engineering, China University of Petroleum, Qingdao, 266580, China

Keywords: Intrusion Detection System, Bio-Inspired, Dendritic Cell Algorithm, Cloud Computing, Model Designation.

Abstract. The paper proposed a new model by applying bio-inspired theory into intrusion detection system in cloud computing. Dendritic Cell algorithm (DCA) based on Danger Theory get involved in the model designation in allusion to the existing problem, high false negative rate, low detection efficiency, low real-time process, and lack of adaptive of the unknown aggressive behavior for nowadays detection mechanisms. Combing the pertinent knowledge of machine learning theory, then to design one improved bio-inspired intrusion detection system model in cloud computing based on machine learning. Finally we introduce our system module as well as the detailed work processes.

Introduction

The human immune system (HIS) has successfully protected our bodies against attacks from various harmful pathogens, and this provides a rich source of inspiration for computer security systems, especially intrusion detection systems. Jerne proposed the first model of the immune system in 1974[1], features gleaned from HIS satisfied the requirements of designing a competent intrusion detection system [2-4]. Applying the theoretical immunology and observed immune functions, models of an intrusion detection system (IDS) has gradually developed into a new research field, called artificial immune system (AIS). But from the current situation, most of the research are still in the exploratory stage, and traditional immunity cannot explain transplants, tumours, and autoimmunity, in which some non-self antigens are not eliminated. Therefore, Aickelin started to work on a Danger Project which intended to apply a new theory, called Danger Theory. Danger theory specially focuses on building more biologically realistic algorithms which consider not only adaptive, but also innate immune reactions [5] [6]. Hence the Dendritic Cell Algorithm (DCA) was proposed by Greensmith et al. [7] [8].

In this paper, based on the existing approach and techniques, we propose a bio-inspired intrusion detection system model, which attempts to design a hybrid intrusion detection model that based on a combination of bio-inspired theory and machine learning technology in cloud computing.

The paper is organized as follows. In the next section, we give the design background of our problem that we will research in this paper. In section 3, we propose the model that we research in this paper, and some definitions and assumptions are given, as well as the bio-inspired intrusion detection engine architecture. Section 4 presents the model work progress. In Section 5, the experiment is made to illustrate the efficiency of the DCA algorithm. Finally, we conclude our paper in section 6.

Design Background

Traditional IDS has the following problems: (1) single detection method; (2) lack of self-learning and adaptive ability; (3) inaccurate control of both false positive rate and false negative rate. In response to these problems, we should trace back to the composition of the data from the network, i.e. consists of three parts: (1) the normal behavior; (2) the aggressive behavior; (3) the unknown behavior. Any single detection method will be invalid when conducting pattern matching in any type of behavior.

Cloud computing obtains elastic computing capabilities and storages file system, intrusion detection in such background will be an acceleration when deploying IDS in clouds. Nevertheless, with other obstacles, such as solution of detection methods, network acceleration speed uncertainty, audit data source, locus of detection especially, those characters will lead to the final architecture of the model.

Our model selects the correct targets in laying solution foundation around the based three problems of traditional IDS:

The most innovative breakthrough is that we introduce the DCA which is a bio-inspired approach break the traditional self non-self limitations, and claim that immune responses are triggered by unusual death of normal tissues, not by non-self antigens, i.e. combines the misuse detection method and anomaly detection method.

Then consider preprocessing network data under certain machine learning approaches, including using the Cluster Analysis Algorithm to the data before any pattern matching, and reducing the dimensions of data attributes in accordance with principal components analysis (PCA) approach. Therefore, machine learning techniques can be utilized to extend the adaptive ability in IDS.

An improved Apriori algorithm plays a significant role in matching links between existing large amounts of data and existing rules to build relevant rule base for detection. We can use this way to achieve the self-learning extension mechanism of the rule set.

Bio-Inspired Intrusion Detection Model

We proposed the system model base on above ideas, as shown in Figure 1:

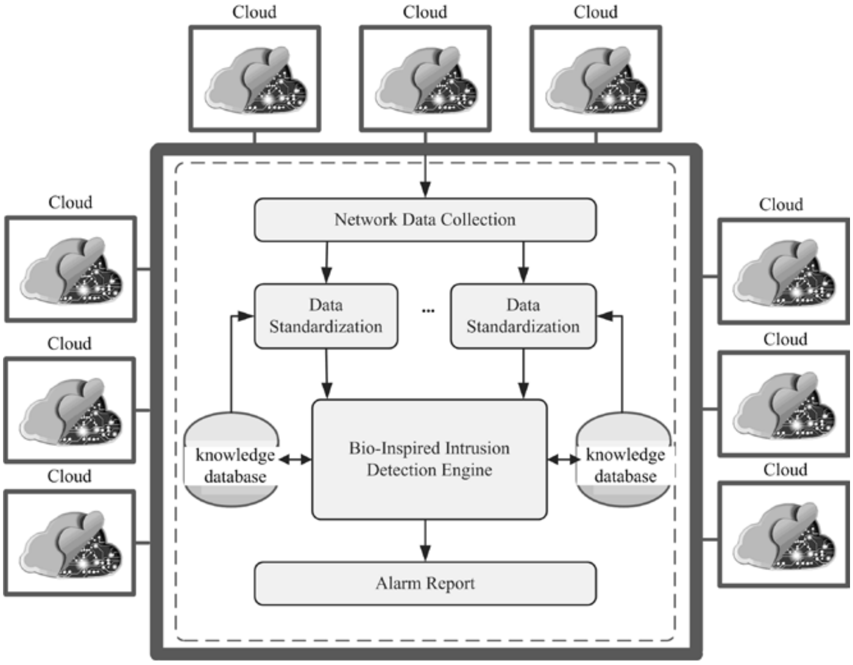

Fig.1: Bio-Inspired Intrusion Detection System in Cloud Computing

In Fig.1, the model is mainly applied in SaaS of Cloud. The basic functions of the modules are as follows:

Network Data Collection

The module is mainly responsible for collecting network data such as TCP events, UDP events, and ICMP events etc. Then the collection will be delivered to the next module.

Data Standardization

The goal of this module is to make the data standardized. While in our model, the bio-inspired approach will use small training sets but some classified and standardized data, i.e. raw network data contains various attributes need some classified demands and dimension reduction to standardize. In this module, we proposed Cluster Analysis algorithm to classify the raw data as well as principal components analysis (PCA) in dimension reduction. Then the raw data will be automatically processed step by step through above procedures.

Bio-Inspired Intrusion Detection Engine

This module plays a core role in the whole model which integrates some major function modules, as is shown is Figure 2:

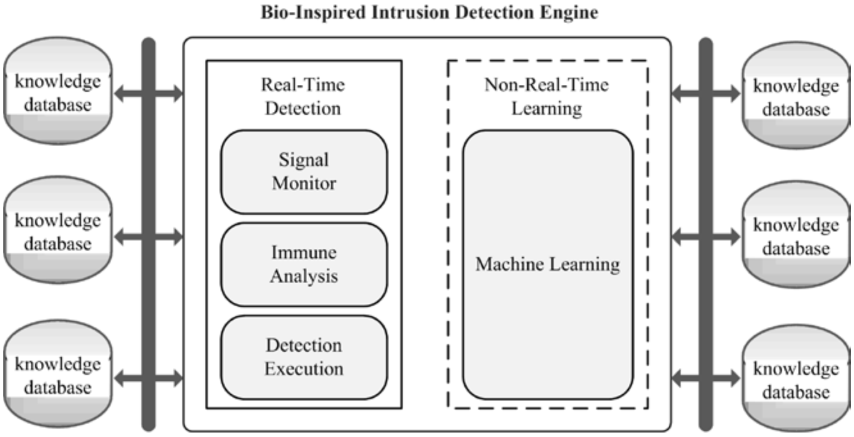

Fig.2: Internal Components of Bio-Inspired Intrusion Detection Engine

The engine will provide two types of function module. The solid line means the real-time detection modules, and the dotted line represents the non-real-time learning modules.

A. Signal Monitor Module

This module mainly obtains the preprocessed data and extracts the danger signal inside.

B. Immune Analysis Module

In this module, the DCA algorithm will be a major and long-term factor in data analysis. Just like DCs in immune system, which have the capacity to combine a multitude of molecular information and to interpret this information for the T-cells of the adaptive immune system. When the input signal arrives, the module will combine the signal in a linear signal model to simulate the DCs signal biological processes.

C. Detection Execution Module

This module will help the Immune Analysis Module evaluate the anomaly state of the antigen data, then a final detection result will be generated.

D. Machine Learning Module

- Feature Extraction
- Rule/Feature Self-Learning
- Rule/Feature Update

This module will finish three main functions above. First, the module will help extract the feature for unknown attack data signal storing in the knowledge database. Second, the module will establish new danger attack rule, and automatically learn the new rules in a non-real-time style. At last, the new rule and feature will update in the knowledge database. This paper proposes an improved Apriori algorithm intends to build relevant rule base for detection, and the reference data set is KDD CUP 99.

Alarm Report

After the detection analysis working in the engine, this module will give some warning with some advices to remind the cloud client checking their service running state etc.

The principles of design our robust models consists of two major phases: Requirements Definition answers the broad questions about what IDS function is and what it should do, and Framework Definition answers questions about how a system behaves and how it is structured to meet user goals.

Model Work Process

The mainly process of the whole system model includes the following steps:

Training the machine learning knowledge database

There will be two methods to train such machine learning algorithm, one is collecting raw network data from a mature IDS which distinguish endanger (normal) behavior data from the danger (abnormal) behavior data, the other is in support of official network data sets such as KDD. The Cluster Analysis algorithm will process the above data into a few categories as an initial standardized endanger classified database.

Bio-inspired intrusion detection.

The technique of Sniffer Discovery is a hard problem in network security. We deposit sniffer application over the cloud, with a verbose packet sniffer that displays a great amount of detail on each packet it reads, including application layer information.

Through decoding the data packet, and PCA will make some dimension reduction work to help read the structure of the raw data.

The DCA core engine will receive the standardized data input stream, if it belongs to antigen, then update the antigen profile; if it belongs to signal, then transform the signal which required more detection until a condition that will cause a termination of a loop, then the danger signal will be sent to alarm module; besides, the endanger signal will look forward to the machine learning pattern matching to do the further analysis.

The improved Apriori algorithm will do the alternative pattern matching work to make the final decisions, i.e. the danger data will send to the alarm module and the endanger rule or feature will store in knowledge database at the same time, which establish a approach in increasing robustness of the detection preference.

Experimental setup and results

In order to illustrate the feasibility and effectiveness of the DCA algorithm in our model, feature selection improves classification by searching for the subset of features is used, which best classifies the training data [9]. The data for our experiments was prepared by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Labs MIT [10]. The data set has 41 attributes for each connection record plus one class label as given in Table 1.

Table 1: Variables for DCA algorithm data set

Variable No.	Variable Name	Variable Type	Variable Label
1	duration	Continuous	A
2	protocol_type	Discrete	B
3	service	Discrete	C
4	flag	Discrete	D
5	src_bytes	Continuous	E
6	dst_bytes	Continuous	F
7	land	Discrete	G
8	wrong_fragment	Continuous	H
9	urgent	Continuous	I
10	hot	Continuous	J
11	num_failed_logins	Continuous	K
12	logged_in	Discrete	L
13	num_compromised	Continuous	M
14	root_shell	Continuous	N
15	su_attempted	Continuous	O
16	num_root	Continuous	P
17	num_file_creations	Continuous	Q
18	num_shells	Continuous	R
19	num_access_files	Continuous	S
20	num_outbound_cmd	Continuous	T

	s		
21	is_host_login	Discrete	U
22	is_guest_login	Discrete	V
23	count	Continuous	W
24	srv_count	Continuous	X
25	serror_rate	Continuous	Y
26	srv_rerror_rate	Continuous	Z
27	rerror_rate	Continuous	AA
28	srv_rerror_rate	Continuous	AB
29	same_srv_rate	Continuous	AC
30	diff_srv_rate	Continuous	AD
31	srv_diff_host_rate	Continuous	AE
32	dst_host_count	Continuous	AF
33	dst_host_srv_count	Continuous	AG
34	dst_host_same_srv_rate	Continuous	AH

Continued from Table 1: Variables for DCA algorithm data set

Variable No.	Variable Name	Variable Type	Variable Label
35	dst_host_diff_srv_rate	Continuous	AI
36	dst_host_same_src_port_rate	Continuous	AJ
37	dst_host_srv_diff_host_rate	Continuous	AK
38	dst_host_serror_rate	Continuous	AL
39	dst_host_srv_serror_rate	Continuous	AM
40	dst_host_rerror_rate	Continuous	AN
41	dst_host_srv_rerror_rate	Continuous	AO

Table 2: Variables for DCA algorithm data set

Attack Type.	Classification accuracy on test data set (%)
	DCA
Normal	99.98
Probe	99.90
DoS	99.89
U2R	90.24
R2L	99.98

As can be seen from Table 1, the first four classes except U2R will result in a high accuracy, maybe the U2R class will not be suitable in our algorithm, and we need some more hard work to do.

Conclusions

This paper designs a new bio-inspired intrusion detection system model in cloud computing which based on machine learning. The core detection module use DCA which does not require large training sets and the knowledge of normality and anomaly is acquired through a machine learning approach, which is the second improvement to achieve a more accurate detection rate and a low false negative rate, that enhance the whole performance of the entire detection system. Overall, our proposed model architecture gave a high accuracy for most of the attacks, but there should be more research job to deal with in the future.

Acknowledgements

The research was supported by the Fundamental Research Funds for the Central Universities.

References

[1] United States general accounting office. Information security: computer attacks at department of defence pose increasing risks. GAO/AIMD-96-84, USA, 1996.

[2] Kim J W. Integrating artificial immune algorithms for intrusion detection. University College London (University of London), 2002.

[3] Somayaji A, Hofmeyr S, Forrest S. Principles of a computer immune system. Proceedings of the 1997 workshop on new security paradigms. ACM, 1998: 75-82.

[4] Hofmeyr S A, Forrest S. An immunological model of distributed detection and its application to computer security. PhD thesis, University of New Mexico, 1999.

[5] Aickelin U, Bentley P, Cayzer S, et al. Danger theory: The link between AIS and IDS? . Artificial Immune Systems. Springer Berlin Heidelberg, 2003: 147-155.

[6] Aickelin U, Cayzer S. The danger theory and its application to artificial immune systems. ArXiv preprint arXiv: 0801.3549, 2008.

[7] Greensmith J, Aickelin U. Dendritic cells for real-time anomaly detection. ArXiv preprint arXiv: 1001.2405, 2010.

[8] Greensmith J, Aickelin U. Dendritic cells for SYN scan detection. Proceedings of the 9th annual conference on Genetic and evolutionary computation. ACM, 2007: 49-56.

[9] Chebrolu S, Abraham A, Thomas J. Feature deduction and ensemble design of intrusion detection systems. Computers and Security. Elsevier, 2005: 295–307.

[10] MIT Lincoln Laboratory. http://www.ll.mit.edu/IST/ideval/