DB-OLS: An Approach for IDS[1]

Vikas Pareek, Aditi Mishra, Arpana Sharma,
Rashmi, Shruti Bansal

Apaji Institute Of Mathematics and Applied Computer Technology
Banasthali Vidyapith, India-304022

Email :{ er_pareekvikas@yahoo.co.in,aditimishra29@gmail.com,
shrutib23@gmail.com, chauhan.rashmiarya@gmail.com}

ABSTRACT: *An intrusion detection system plays a major role in network security. We propose a model "DB-OLS: An Approach for IDS" which is a Deviation Based-Outlier approach for Intrusion detection using Self Organizing Maps. In this model "Self Organizing Map" approach is to be used for behavior learning and "Outlier mining" approach, for detecting an intruder by calculating deviation from known user profile. This model aims to improve the capability of detecting intruders.*

KEYWORDS: *IDS, SOM, Outlier mining, Network security.*

1 INTRODUCTION

The use and importance of computers along with internet is increasing rapidly. With the growth of the Internet and its potential there has been subsequent change in business model of organizations across the world. More and more people are getting connected to the Internet to take advantage of the new business model popularly known as e-Business.

There can be two aspects of business on the Internet. On one aspect, the Internet brings in tremendous potential to business in terms of reaching the users. At the same time it also brings in lot of risk to the business. A significant security problem for networked systems is hostile, or at least unwanted, trespass by users or software. One of the two most publicized threats to security is the intruder (the other is viruses), generally referred to as a hacker or cracker.

[1] Initial draft of this paper appeared in the proceedings of IWTMP2PS as Springer CCIS series, 2010, Volume 89, Part 2, 395-401.

DOI : 10.5121/ijp2p.2010.1103

1.1 Intruder: Intruder is the person who tries to gain unauthorized access to the network or a host based system. Anderson identified three classes of intruders, viz. Masqueraders, Misfeasors and Clandestine**.** [1]:

1.2 Intrusion: Intrusion is caused by attackers accessing the systems from the Internet, authorized users of the systems who attempt to gain additional privileges for which they are not authorized, and authorized users who misuse the privileges given to them.

1.3 Intrusion Detection: An Intrusion detection System assumes that an attacker has outsmarted the security guards and gained an authorized access. It tries to identify attackers by scanning the behavior of active users.[1] Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions, defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer or network [2].

Intrusion can be detected using the signature and behavior based knowledge of intruder (can be either masquerader or misfeasor) and intended user for a network based intrusion detection system.

2 IDS CLASSIFICATION

An Intrusion Detection System (IDS) is categorized either as network-based or host based. In the former, header fields of the various network protocols are use to detect intrusions. In the later approach (host-based IDS), the focus shifts to the operating system level.

There are two possible approaches to detect the intrusion. The first is misuse *(*signature*)* detection: this technique is similar to pattern matching. Initially the system has been designed on the basis of known attack patterns and the test data is checked for the occurrence of these patterns.

These systems have a high degree of accuracy, very effective at detecting attacks without generating an overwhelming number of false alarms. But they are unable to detect new attacks, can only detect those attacks they know about—therefore they must be constantly updated with signatures of new attacks. To overcome the drawback of misuse detection it is required to have the knowledge base system in which regular updates need to be made in order to add new intrusion scenarios.

The second technique is *anomaly detection:* this technique works by analyzing the deviation from normal activities and usually at the user level or system level. They function on the assumption that attacks are different from “normal” (legitimate) activity and can therefore be detected by systems that identify these differences. They can detect unusual behavior and thus have the ability to detect symptoms of attacks without specific knowledge of details and can produce information that in turn be used to define signatures for misuse detectors but usually produce a large number of false alarms due to the unpredictable behaviors of users and networks [2]. Anomaly detection is further categorized as statistical, predictive patterns or neural network based anomaly detection.

3 CONTEMPORARY RESEARCH

There are numerous data mining techniques that have been used for detecting intruders. These include association rule mining, frequent pattern, classification, clustering, nearest neighbor (KNN), outlier mining etc. Association rule mining was one of the popular techniques but its place was taken by some other techniques like clustering or classification because it was quite slow.

SVM is Support Vector Machine, comparatively a new classification technique with a higher performance than other previous learning methods. It

has been used in many applications such as bioinformatics and pattern recognitions. In the field of security, lots of the work has been done by using SVM and it has also been used in IDS. Though SVM based IDS enhanced the performance of IDS in terms of detection rates and speed of processing, but still there are lots of work to do.

Latifur Khan, et al. [3] proposed - "A new intrusion detection system using support vector machines and hierarchical clustering" a method which has scalable solutions for detecting network based anomalies by support vector machine for classification. It has high generalization accuracy but takes long time for training.

Clustering became the next choice because generally we have large amount of data in multidimensional form. Some of the models for IDS have been designed on the basis of k-means or k-monoids clustering technique of data mining but it suffers from various shortcomings like, this algorithm is very sensitive to outliers, and generally terminates at a local optimum. Secondly, it is necessary for K-means algorithm to determine the number K of clusters in advance. Therefore, the quality of the result is not satisfactory. The dependency on clusters effects critically on the clustering results.

A brief summary of some clustering methods based on competitive learning is shown in the table below:

S.No.	Approach	Description
1)	Self Organizing Maps (SOM)	It is well-liked for cluster analysis, feature extraction and data visualization. Topology-preserving clustering method. Especially powerful for the visualization of high-dimensional data. Can be used for complex processing systems, making it into a simple subsystem.
2)	Learning vector quantization (LVQ)	Takes much space for storage. More time is needed for computation
3)	*C*-means/ K-means clustering	The number of clusters must be defined early. Requires much time. This algorithm is very sensitive to outliers, and generally terminates at a local optimum. The quality of the result is not satisfactory.

4)	Adaptive resonance theory (ART)	It is complex and sensitive.
5)	Mountain and subtractive clustering	Complexity increases when it grows exponentially and dimensionally. It takes less time for clustering when training set is small but much more time when training set is very large.
6)	Neural gas	Needs high computational effort.

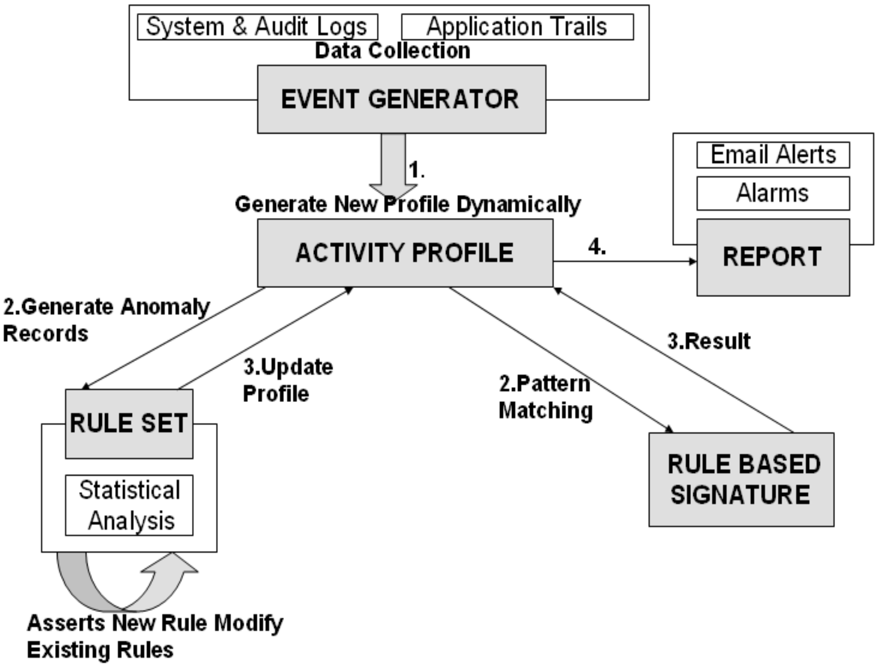
The models are not continuously learning the behavior thereby leading to the possibility that a new intrusion will not be detected and a false alarm may be generated [4].The IDS model that we propose has the ability of continuous updation in behavior learning, which guarantees the system administrator that false alarms generated will be low. New intrusion will be detected because continuous updation in behavior learning is being made. We are considering some specific audit data for the confirmation of audit patterns, so it requires only a specific portion of audit records to be reviewed.

For learning the behavior of the system Artificial Neural Network approach "Self Organizing Approach (SOM)" is to be used. SOM is a type of neural network that is trained using unsupervised learning. SOM uses neighborhood functions which makes it different from other artificial neural networks.

4 PROPOSED APPROACH

4.1 Approach:

We propose a model for Intrusion Detection, with the help of deviation based outlier and self organizing map approach, using signature and behavior based knowledge of intruder (masquerader & misfeasor) and intended user for a network based intrusion detection system. When the intrusion detection system is learning the behavior, at the same time the network may experience an attack. In this interval of learning the system is open for attacks. Our system provides security with the help of network packet monitoring, audit logs and signature matching. This will also reduce false alarm rate.



4.2 Model:

-

Figure: 1. Model of DB-OLS

4.2.1 Working of Model:

Whenever the user logs into the system, some activities will be performed and there corresponding events will get traced in audit log files. Data will be collected for the event generated from audit log files. As soon as the data is collected the next step is generation of the activity profile of the current user on the basis of certain parameters, which has been selected from audit log files.

Initially some rules are defined on the basis of known attacks, which has been stored in the rule set. Once activity profile has been generated, two threads will be executed simultaneously. On one side, with the help of the existing rule pattern matching will be done which defines whether the behavior is of intruder or of legitimate user. The result will be reported back to activity profile and report will be generated in the form of alarms.

Now it may be possible that the behavior of legitimate user may change or there may be the possibility of new attacks, so it is necessary to update the existing rule set. Thus it requires continuous updation in the existing rule set. Therefore on the other side the second thread will executed which will update the existing rules in the rule based signature.

4.2.2 Description of Model:

Event Generator: It is a data collector and the events may include audit records, system logs and application trails.

Activity Profile: It contains the variables that are used to calculate the behavior of the system based on some predefined statistical measure. The variables are associated with certain pattern specifications, which come into, play when filtering the event records.

Rule Based Signature: Involves an attempt to define a set of rules that can be used to decide that a given behaviour is that of an intruder. Rules are developed to detect deviation from previous usage patterns that searches for suspicious behaviour.

Rule Set: The rule set represents inferencing mechanisms, such as a rule-based system. It uses event records, anomaly records, and other data to Control the activity of the other components of the IDS and to update their state.

Report: Intrusions detection must be brought to attention. Reporting means Printed reports, email alerts, audible alerts, graphical displays etc [5].

4.2.3 Flow Diagram of DB-OLS:

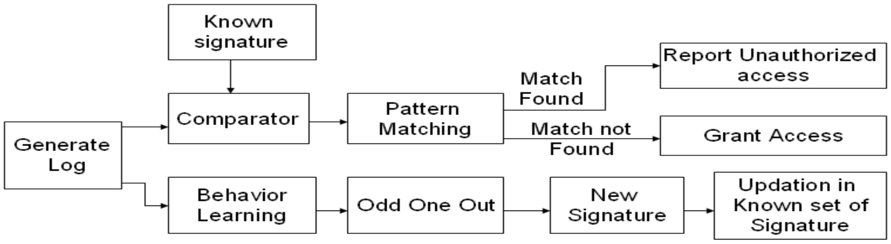


Figure:2 Flow Diagram of DB-OLS

The system monitors the user’s activity and collects audit logs corresponding to each event. When the logs are generated, system can have signature of current user who has just logged into it. This signature is then fed into two units.

In one of the two units, the current signature and stored known signature (of attacker) will be compared. Here pattern matching will take place. If the patterns are matched that means signature of current user has matched with intruder’s signature and report will be generated for unauthorized access.

In the other unit, the system will make use of signatures to learn the behavior of user’s current activity. This behavior learning is necessary to have an updated system in order to have false alarm rate. For learning the behavior of the system Artificial Neural Network approach "Self Organizing Approach (SOM)" has been used.

They provide a way of representing multidimensional data in much lower dimensional spaces - usually one or two dimensions. This process, of reducing the dimensionality of vectors, is essentially a data compression technique known as *vector quantization*. SOM operates in two modes: training and mapping. We are using SOM during the behavior learning because it is unsupervised learning. In this we have taken certain parameters from the audit logs, and some weightage has been assigned to each parameter selected. Weightage is applied because on the basis of one or two parameters we cannot decide whether a person is an intruder or a legitimate user. Different parameters are holding different weightage, that is, a value is assigned to each parameter. This map is feed forward map which helps us to know about the future learning because under this each time the dataset will be updated.

In SOM we will be using U matrix. The U matrix value of a particular parameter is the average distance between the parameter and its neighborhood parameters. Euclidian distance is being used to calculate the value of U matrix. Whenever a new data set is being inputted in the network, that is whenever behavior learning is in process, and then the Euclidian distance to all the weight factors is computed. The parameter with weight factor most similar to the input is called Best Matching Unit (BMU). The magnitude of change that is the magnitude of learning new behavior decreases with time. This is so because more and more new cases have been already set as input as a case of learning. Hence change of magnitude decreases with time and with distance from BMU. Figure shows the best matching unit. [6]

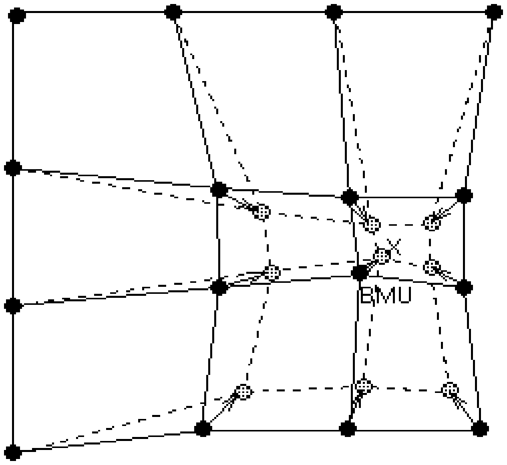


Figure: 3: Updating the best matching unit (BMU) and its neighbors towards the input sample marked with x. The solid and dashed lines correspond to situation before and after updating, respectively.

The formula for updation of data set is given as follows:

$$Wv(t+1) = Wv(t)+\Theta(v, t)\ \alpha(t)\ (D(t)-W\ v(t)) \qquad (1)$$

where:
$\alpha(t)$ is monotonically decreasing learning coefficient

A function is said to be monotonically decreasing whenever $x \leq y$ then $f(x) \geq f(y)$.
t is current iteration
α is limit on time iteration
Wv is current weight vector
$\Theta(t)$ is restraint due to distance from BMU usually called neighborhood function
$\alpha(t)$ is learning restraint due to time
D(t) is the input vector
$\Theta(v, t)$ is neighborhood function. It depends on the distance between the BMU and parameter p.
The U matrix is formed with the help of Euclidian distance. Euclidian distance between point's p and q is the length of the line segment pq

$$d(p,q) = \sqrt{}\ (p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots\ldots\ldots\ldots\ldots + (p_n - q_n)^2 \qquad (2)$$

Here each parameter has been assigned a weight factor. The size of the neighborhood function decreases with time. This process is repeated for each input vector, for a certain number of cycles in order to decrease the false alarm rate. [8, 9, 10]

Algorithm:

```
1.  Select the parameters.
2.  Randomize the parameters with weight factors.
3.  Learn the behavior.  It can be new from the existing behavior.
4.  Calculate the Euclidian distance for each parameter.
5.  Track the parameter with the smallest distance. This gives the BMU of the parameter.
6.  Update the parameter learning:
    $Wv(t+1) = Wv(t) + \Theta(v,t)\ \alpha(t)\ (D(t) - W\ v(t))$
7.  Increment t and repeat from step 2 while $t < \alpha$
```

We will use Outlier Mining. It is mainly used for fraud detection and for several other applications. In this, we are given a set of 'n' data points or objects and 'k' the expected number of outliers, and we have to find the top 'k' objects that are considerably dissimilar, exceptional or inconsistent with respect to the remaining data.

"Deviation Based Approach" identifies outliers by examining the main characteristics of objects in the group. Objects that deviate from this description are considered as outliers. We will sequentially compare the values in a given set. A set S of 'n' objects, a subsequence $\{S_1, S_2 \ldots S_m\}$ of these objects is made with $2 \leq m \leq n$ such that S_{j-1} is subset of S_j. In a given set of values it will return notification if the objects are similar to one another. Set a threshold value, if the difference is above the threshold value then alarm will be generated [4, 7].

5 CONCLUSION

Our model is compliant with the IETF-CIDF (common intrusion detection architecture) standard of DARPA. The event generator, Rule set; Activity profile, signature and report correspond to E-box (event generator), A-box (Analyzer), D-box (database) and R-box (response unit) respectively of CIDF. Therefore the design will be standard and cost effective.

One drawback of our system is that the issue of fail-closure has not been addressed. That is, in case of sudden failure, it is not clear how the system will close its operation, so as to leave the network protected.

The use of two approaches for the intrusion detection system can raise the amount of resources of the system, able of finding attacks which may occur for a long period of time, a number of user sessions, or by many attackers working in concert. Our system can provides security with the help of network packet monitoring, audit logs and signature matching. This will also reduce false alarm rate.

6 FUTURE WORK

As the use of SOM results in slow operation, for better efficiency, other techniques like adaptive resonance theory may be used. Besides that, a thorough evaluation of the scheme and the claims made here will pave the way for better understanding of the strengths and weaknesses of this system. For this we intend to build an IDS based on this approach.
DB-OLS: An Approach for IDS overcomes the drawbacks of Rule based & statistical anomaly based approaches. We put forward a new approach that contains advantages of both the Rule based and statistical anomaly based approaches by using self organizing maps and deviation based outlier approach. This combination ensures the detection of intrusion and updation of signature base.

ACKNOWLEDGEMENTS:

We give our special thanks to our colleagues in Computer Science, Ms. Kanchan Sharma and Ms. Sonal Dixit for their comments and encouragement during this work. We owe a great amount of thanks to our parents for their endless support and love.

6 REFERENCES:

[1]. Pieprzyk, J., Hardjono, T., Seberry, J.: Fundamentals of Computer Security, Springer International Edition (2003)

[2]. Bace, R., Mell, P.: NIST Special Publication on Intrusion Detection Systems (2000)

[3] Khan, L., Awad, M., Thuraisingham, B.: A new intrusion detection system using support vector machines and hierarchical clustering, 2006, http://www.springerlink.com/content/v77p215n66071087/

[4]. Cannady, J., Harrell, J.: A Comparative Analysis of Current Intrusion Detection Technologies, Georgia Tech Research Institute, http://www.neurosecurity.com/articles/IDS/TISC96.pdf

[5]. Hen, J., Kamber, M.: Data Mining- Concepts and Techniques, Morgan Kaufmann (2000)

[6] http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml

[7]. Seleznyov, A., Puuronen, S.: Anomaly Intrusion Detection Systems- Handling Temporal Relations between Events, University of Jyväskylä, Finland, http://www.raid-symposium.org/raid99/PAPERS/Seleznyov.pdf

[8]. Jones*,* AK*,* Sielken*,* RS*:* Computer system intrusion detection: A survey. Technical report*,* University of Virginia Computer Science Department *(1999)*

[9]. Stallings, W.: Cryptography and Network Security, Third Edition, Prentice Hall (2003)

[10]. Anderson, J.P.: Computer Security- Threat Monitoring and Surveillance. Technical Report, J.P. Anderson Company, Fort Washington, Pennsylvania (1980)

[11]. Deepa, S.N., Sivanandam, S.N.: Principles of Soft Computing, Wiley India (2007)

[12]. Kim, D.S., Nguyen, H.N., Park, J.S.: Genetic Algorithm to Improve SVM Based Network Intrusion Detection System, Hankuk Aviation University, Seoul, KOREA(2005)
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1423667&userType=inst

[13]. Lichodzijewski, P., Heywood, A.N.Z.,Heywood, M.I.: Dynamic Intrusion Detection Using Self-Organizing Maps, Dalhousie University, Halifax, NS (2003)

[14]. Zanero, S., Savaresi, S.M.: Unsupervised learning techniques for an intrusion
detection system, Milan, Italy (2004)

[15]. BRUGGER, S.T.: Data Mining Methods for Network Intrusion Detection, University of California, Davis (2004)

[16]. Lazarevic, A., Ozgur, A., Ertoz, L., Srivastava, J., Kumar, V.: A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, University of Minnesota, USA (2003)

[17]. Pareek, V., Mishra, A., Sharma, A., Chauhan, R., Bansal, S. : A Deviation Based Outlier Intrusion Detection System, Recent Trends in Network Security and Applications, Springer Communications in Computer and Information Science, 2010, Volume 89, Part 2, 395-401.

[18] CIDF working group. The common detection framework. Version 06 available at http://seclab.cs.ucdavis.edu/cidf,1999.

ABOUT AUTHORS

Vikas Pareek is Assistant Professor (Computer Science) at Banasthali Vidyapith. He has done B.E. (CSE) and is pursuing Ph.D. on “Integer Factorization and breaking RSA algorithm”. His research interests include information security, algorithmics and mobile computing.

Aditi Mishra has done B.E. (I.T.) from RGTU, Bhopal (M.P.) in June 2009. She is pursuing M.Tech in Information Technology from department of Computer Science at Banasthali University, India. Her research interests include Information Security, Data Mining and Wireless communication. She loves to play Tennis on hard court, Reading biographies, poems, Writing poems and listening to music.

Arpana Sharma has done her M.Sc. (I.T.) from Rajasthan University in June 2009. She is pursuing M.Tech in Information Technology from department of Computer Science at Banasthali University, India. Her research interests include Natural Language Processing, Artificial Intelligence and Machine Translation. She loves to surf open source code.

Rashmi has done her M.Sc. (Mathematics) from C.C.S. from University Meerut in June 2008. She is pursuing M.Tech in Information Technology from department of Computer Science at Banasthali University, India. Her research interests include Natural Language Processing, Theory of Computation and Information Security. She loves Painting.

Shruti Bansal has done her B.Sc. (Mathematical Science) from Delhi University in June 2006 and M.C.A. from Nice Management College, UPTU in June 2009. She is pursuing M.Tech in Information Technology from department of Computer Science at Banasthali University, India. Her hobbies are playing badminton and cooking.