Incremental Hybrid Intrusion Detection Using Ensemble of Weak Classifiers

Amin Rasoulifard, Abbas Ghaemi Bafghi, Mohsen Kahani

Ferdowsi University of Mashhad, Department of computer science, Faculty of Engineering
Mashhad, Iran
rasoulikarbar@yahoo.com, ghaemib@um.ac.ir, kahani@um.ac.ir

Abstract

It is important to increase the detection rate for known intrusions and detect unknown intrusions. It is also important to incrementally learn new unknown intrusions. Most current intrusion detection systems employ either misuse detection or anomaly detection. In order to employ these techniques, we propose incremental hybrid intrusion detection system. This framework combines incremental misuse detection and incremental anomaly detection. Our framework can learn new class of intrusions that not exist in previous data which used for training incremental misuse detection. The framework has lower computational complexity so it is suitable for real-time or on-line learning. Experimental evaluation of KddData also presented.

Keywords

Incremental learning, hybrid IDS, weak classifier, ensemble of classifier.

1. Introduction

With the fast growing of network-based services and sensitive information on the networks, the number and the severity of network-based computer attacks have significantly increased. Although a wide range of security technologies such as information encryption, access control, and intrusion prevention can protect network-based systems, there are still many undetected intrusions.

An intrusion can be defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource". IDS can detect and identify intrusion behavior or intrusion attempts in a computer system by monitoring and analyzing network packets or system audit logs, and then sends intrusion alerts to system administrators in real time. Intrusion detection techniques can be categorized into misuse detection and anomaly detection. Misuse detection systems use patterns of well-known attacks or weak spots of the system to identify intrusions. The main shortcoming of such systems are: known intrusion patterns have to be hand-coded into the system; they are unable to detect any future(unknown) intrusions that have no matched patterns stored in the system [1,2,3]. Anomaly detection systems firstly establish normal user behavior patterns (profiles) and then try to determine whether deviation from the established normal profiles can be flagged as intrusions. The main advantage of anomaly detection systems is that they can detect new types of unknown intrusions [4,5,6].

In recent years, the continual emergence of new attacking methods has caused great loss to the whole society. So, the advantage of detecting future attacks has specially led to an increasing interest in incremental learning techniques. The traditional methods commonly build a static intrusion detection model on the prior training dataset, and then utilize this model to predict on new network behavior data. However, the network behavior model will not change continually along with detecting and analyzing process. Thus the initially learnt intrusion detection model can not adapt to the new network behavior pattern, which causes that the false positive rate of detection model increases and the detection precision of detection model declines.

In order to improve intrusion detection with high detection rate, with the ability of detection new unknown attacks, and continually adapt model to cope with new network behaviors, we propose hybrid intrusion detection system which combine the incremental misuse intrusion detection and incremental anomaly detection. In addition, when intrusion detection dataset is so large that whole dataset can't be load into main memory, the original dataset can be partitioned into several subsets, and then the detection model is dynamically modified according to other training subsets after the detection model built on one subset.

Several hybrid intrusion detection systems have been proposed for combining misuse detection and anomaly detection [7,8,9]. We proposed hybrid intrusion detection system based on incremental learning. We use ensemble

of weak classifiers for implementing incremental misuse intrusion detection system. Intrusion detection systems using ensemble of weak classifiers generally possesses lower computational complexity than other frameworks which using strong classifier, because of using weak classifier with suitable parameter to satisfy weak hypothesis. We use on-line k-mean algorithm for incremental anomaly detection to detect unknown intrusions.

The rest of the paper is organized as follows: related work presented in section 2, hybrid system architecture presented in section 3, the proposed architecture presented in section 4, experimental evaluation presented in section 5 and conclude the paper in conclusion section.

2. Related Work

Hybrid intrusion detection systems composed of misuse detection and anomaly detection system. It can detect both known intrusions and unknown intrusions. Various hybrid methods have been proposed for improve the productivity misuse intrusion detection and anomaly detection systems.

ADAM (Audit data analysis and mining) is a hybrid on-line intrusion detection system which uses association rules for detecting intrusions [7]. This framework consists of two phases: training phase, on-line phase. In training phase, the dataset without any class of intrusions applied to the model and constitute a profile of normal activities as a set of association rules pattern. In on-line phase, ADAM use sliding window, on-line algorithm that find frequent pattern in the last D connections and compare them with those stored in normal profile, discard those that similar to the pattern of normal profile. With the rest, ADAM uses a classifier which has been previously trained to classify the suspicious data as a known type of attack, unknown types and a false alarm.

The Next Generate Intrusion Expert System (NIDES) is a hybrid system [8]. It is consists of rule-based misuse detection and anomaly detection that use statistical approaches. This framework employs misuse detection and anomaly detection in parallel for detecting intrusions. The random forest algorithm used for hybrid intrusion detection system in [9]. It use ensemble of classification tree for misuse detection and use proximities to find anomaly intrusions. Such as ADAM it has two phases: on line phase, off-line phase. In on-line phase the classification trees are used to generate the pattern of known intrusions and in the off-line of algorithm, system can detect unknown intrusions and build patterns of unknown intrusions then add its to the database of known intrusion patterns.

FLIPS is the framework which uses hybrid approach for intrusion prevention systems [10]. The core of this framework is an anomaly-based classifier that incorporates feedback form environment to both tune its model and automatically generate signatures of known malicious activities. It uses PayL as an anomaly detection component. The misuse detection component of this framework is signature-based intrusion detection system which use pattern of intrusions for detection.

In proposed incremental hybrid system, we use Learn++ algorithm for incremental misuse detection component and on-line k-mean algorithm is used for incremental anomaly detection component.

3. Hybrid System Architecture

To improve the productivity of misuse and anomaly detection, several hybrid intrusion detection systems have been proposed. These frameworks combine misuse and anomaly detection to achieve the detection rate of misuse detection and to detect unknown intrusions. There are three ways to combine misuse and anomaly detection: use anomaly detection at first then misuse detection, use misuse and anomaly detection in parallel and use misuse detection and then anomaly detection afterwards.

Some hybrid intrusion detection systems use anomaly detection at first to detect suspicious activities and then use misuse detection to detect attacks from suspicious activities [7,10]. Suspicious activities are those that deviate from the profile of normal activities. The framework of this approach is shown in Fig.1. Observed activities applied to the anomaly detection to produce suspicious activities and then misuse detection is used to detect attacks. Connections that match to the pattern of attacks are labeled as attack, those that match to false alarm patterns are labeled as normal and the others are labeled as unknown attacks.

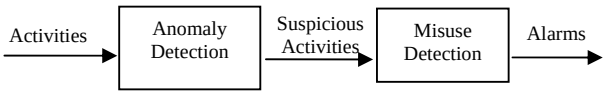

Fig.1: Anomaly detection at first then misuse detection

In hybrid intrusion detection systems that use anomaly detection at first, to reduce false positive rate of framework, this component must have high detection rate and misuse detection component must have an ability to detect false positive rate. So these frameworks will not suitable for hybrid intrusion detection systems. Fig.2 has shown the framework of these hybrid systems.

Some hybrid intrusion detection systems use misuse detection and anomaly detection component in parallel [8]. In these frameworks both components generate suspicious activities individually. Then, the correlation component used to elicit intrusions from suspicious activities. Fig.3 has shown the framework of these hybrid systems.

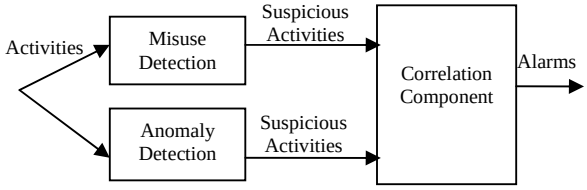

Fig.2: Parallel approach

Other hybrid intrusion detection systems employ misuse detection at first and use anomaly detection afterwards [9]. These frameworks can detect known attacks in real time and generate suspicious activities from the observed activities. Anomaly detection is used to detect unknown intrusions from the suspicious activities.

Our proposed incremental hybrid intrusion detection system uses this type of hybrid intrusion detection systems. Our hybrid system is suitable for detect known intrusions in real time because of using weak classifiers with lower complexity and has an ability to learn new intrusions incrementally.

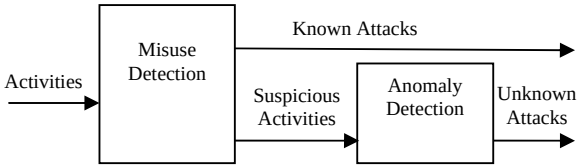

Fig.3: Misuse detection at first then anomaly detection

4. The Proposed Architecture

It is important to increase the detection rate for known intrusions and detect unknown intrusions. It is also important to incrementally learn new intrusions. Due to the fast growing of new intrusions in recent years, Detecting and learning future intrusions will be the main interest in intrusion detection systems. We propose incremental hybrid intrusion detection system which works based on ensemble of weak classifiers, for incrementally learning new intruions that, to the best of our knowledge, is not considered in intrusion detection systems. Intrusion detection systems using ensemble of weak classifiers generally possesses lower computational complexity than other frameworks which using strong classifier, because of using weak classifiers with suitable parameter to satisfy weak hypotheses. This property is very attractive and promising in intrusion detection systems, because the classifiers should be retained in short periods in practice

4.1. Ensemble of Weak Classifiers

Ensemble of Classifier developed to improve the classification performance of weak classifiers. In essence, an *ensemble of weak classifiers* are trained using different distributions of training samples, whose outputs are then combined using one method for combining classifiers [11] to obtain final classification rule. This approach exploits the so-called *instability* of the weak classifiers, which allows the classifiers to construct sufficiently different decision boundaries for minor modifications in their training datasets, causing each classifier to make different errors on any given instance. A strategic combination of these classifiers then eliminates individual errors, generating a strong classifier.

4.2. Proposed Hybrid Architecture

Misuse detection has high detection rate for known intrusions and cannot detect unknown intrusions. Anomaly detection can detect unknown intrusions but having a low detection rate and high computational complexity. It is also important to incrementally learn new intrusions. In order to obtain intrusion detection with aforementioned techniques, we propose an incremental hybrid system which combines the incremental misuse intrusion detection and incremental anomaly detection. The framework for proposed hybrid intrusion detection system is shown in Fig.4.

The proposed framework divided into two phases: on-line phase and off-line phase. Misuse intrusion detection component is used in the on-line phase. It can learn new class of intrusions that not exist in previous data which used for training the existing classifiers. In other words, it can learn new class of intrusions in supervised mode. It is also suitable for learning known intrusions in on-line mode because of using ensemble of weak classifiers with lower computational complexity.

In the off-line section of our framework, we use on-line k-mean algorithm. It can identify new unknown intrusions and can incrementally learn new instance of data. The new identified intrusions by anomaly detection component must be applied to the misuse intrusion detection component in the next learning phase. Therefore, we must determine the class type of new intrusions. For this reason, another component is used for determining the class type of new intrusions. This component is optional and can be done by the administrator of systems. Any supervised or unsupervised clustering algorithms can be used for this component. In our experiment, for the sake of simplicity, we manually determine the class type of intrusions that detected by anomaly detection component.

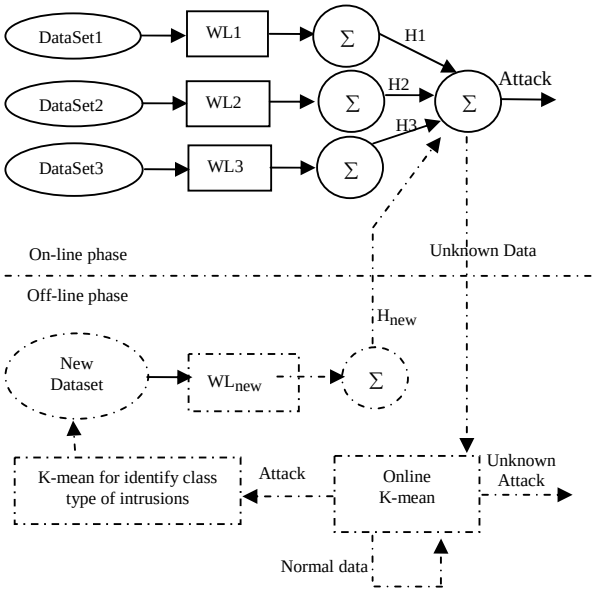

Fig.4: incremental hybrid intrusion detection system

4.3. Misuse Intrusion Detection

Misuse detection systems use patterns of well-known attacks or weak spots of the system to identify intrusions. The main shortcomings of such systems are that it cannot detect new unknown intrusions.

The fast growing of new intrusions in computer systems led to an increasing interest in incremental learning algorithms for intrusion detection systems. Learn++ algorithm is an incremental learning algorithm that used ensemble of weak classifiers for incrementally learn new information [12].

We use Learn++ algorithm for incremental misuse detection component of proposed hybrid intrusion detection system. It has an ability to incrementally learn new class of intrusions that not trained as output for existing classifiers. In other words, this algorithm can learn new class of intrusions in supervised mode. Intrusion detection systems using ensemble of weak classifiers generally possesses lower computational complexity than the other frameworks which using strong classifiers, because of using weak classifiers with suitable parameter to satisfy weak hypotheses. The framework of incremental misuse intrusion detection system is shown in Fig.5 which has the following components:

WL: Learn++ algorithm requires a group of **weak Learner** (classifiers) designed before hand. Weak Classifier can obtain a 50% correct classification performance on its own training dataset. We use multi layer perceptron for implementing a weak classifier.

$\sum$: **Weighted Majority voting** which used for calculating the final classification accuracy based on the classification accuracy of the weak classifiers.

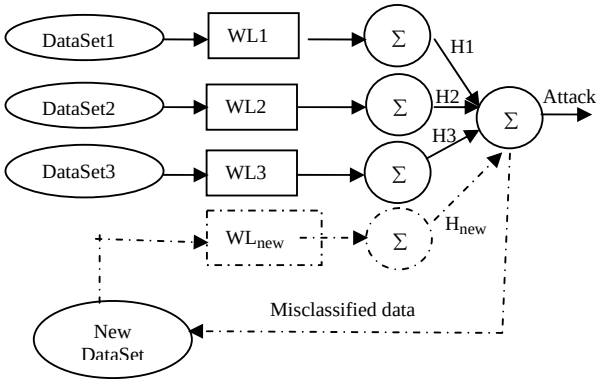

Fig 5: incremental misuse intrusion detection

4.4. Anomaly Detection Component

Anomaly detection amounts to training models for normal behavior and then classifying as intrusions any network behavior that significantly deviates from the known normal patterns. Clustering algorithms have recently gained attention in intrusion detection systems because of having advantages to find new attacks not seen before. With the fast growing of new attacks in recent years, incremental learning gained interest attention for detecting future intrusions. We use incremental clustering algorithm in proposed hybrid system for incrementally learn new unseen intrusions. We use on-line k-mean algorithm [13]. It has low time complexity, fast convergence and it is suitable for incremental learning. The pseudo-code for on-line k-mean algorithms is shown if Fig.6.

Algorithm: online k-means (kmo)

Input: A set of *N* data vectors $X = \{x_1,...,x_N\}$ in $\Re^d$ and number of clusters *K*.

Output: A partition of the data vectors given by the cluster identity vector

$$Y = \{y_1,...,y_N\}, y_n \in \{1,...,k\}$$

Steps:

1. Initialization: initialize the cluster centroid Vectors $\{\mu_1,...,\mu_K\}$;
2. Loop for M iterations
   For each data vector x_n, set

$$y_n = \arg\min_k \|x_n - \mu_k\|^2,$$

   Update the centroid μ_{y_n} as

$$y_{y_n}^{(new)} = \mu_{y_n} - \frac{\partial E}{\partial \mu_{y_n}} = \mu_{y_n} + \xi(x_n - \mu_{y_n}),$$

Where ξ is a learning rate usually set to be a small positive number (e.g., 0.05). the number can also gradually decrease in the learning process.

Fig.6: on-line k-mean algorithm

5. Experimental Evaluation

For simulations, the weak learner used to generate individual hypotheses was a single hidden layer MLP with 41 hidden layer nodes and 4 nodes in output layer. The 4 nodes in output layer correspond to the four class type of intrusions. The mean square error goals of all MLPs were preset to values of 0.02 to prevent over-fitting and to ensure sufficiently weak learning. We note that any neural network can be turned into weak learning algorithms by selecting its number of hidden layers and the number of hidden layer nodes small, and the error goal high, with respect to the complexity of problem.

5.1. Intrusion Dataset

To simulate the presented ideas, we used the 1998 DARPA Intrusion Detection Evaluation program data provided by MIT Lincoln Labs [10]. The TCP dump raw data was processed into connection records, which are about five million connection records. The data set contains 24 attack types. The attacks fall into four main categories as follows:

Denial of Service (DOS): Attacker makes some computing or memory resources too busy or too full to handle legitimate requests, or denies legitimate users access to a machine

Remote to User (R2L): Attacker who does not have an account on a remote machine sends packets to that machine over a network and exploits some vulnerability to gain local access as a user of that machine.

User to Root (U2R): Attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system.

Probing: Attacker scans a network of computers to gather information or find known vulnerabilities. An attacker with a map of machines and services that available on a network can use this information to look for exploits.

5.2. Results and Analysis

For validating the effectiveness of Incremental intrusion detection using ensemble of weak classifiers, the following experiments are done. In all experiment, we assume that the available data not used in previous stage of learning and this data is new unknown data that previous model classify its incorrectly

A 10% sample consisting of about 500,000 records obtained from the UCI machine learning repository was used in our study as training set and entire labeled test set is used for testing set. The labeled test dataset includes 311029 records with different distribution from the training set. We use intrusions of training set to generate the model of incremental misuse component and use normal instances of training set to construct profile of normal activities.

The intrusions dataset divided into five sections. Each section used to generate a respective classifier H. Fig.7 is shown that adding the additional training sample to generate classifier (H) caused an increasing manner in detection rate of misuse detection on test dataset. This means that misuse detection component can learn new information incrementally.

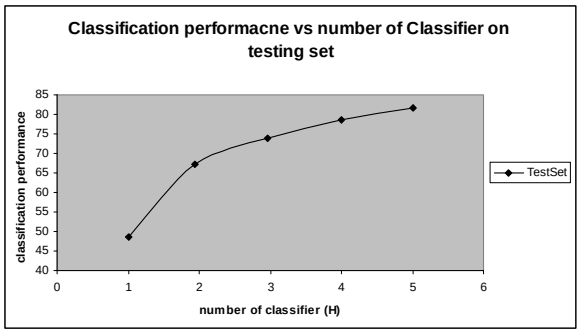

Fig.7: Classification performance vs number of classifier

The effectiveness of anomaly detection components on performance classification of hybrid intrusion detection system investigated in following scenario. We remove the instances of data that correctly classified by misuse detection component from the testing dataset and apply anomaly detection component on remaining instances of testing dataset. These remaining instances are those that misuse detection component clarify them as unknown data. Then incremental anomaly detection applied to remaining dataset to detect anomaly intruions. Many instances of remaining dataset detected as intrusions that misuse detection component could not clarify the classification output of its.

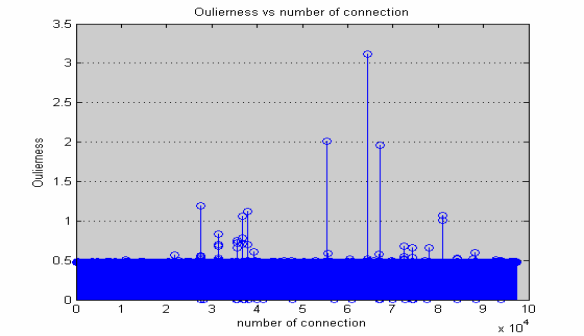

Fig.8: Intrusions detected by anomaly detection

As indicated in Fig.8 there are instances of data in remaining dataset which detected as intrusions by anomaly detection. These instances were not predicted by misuse detection component. This means that combining misuse detection and anomaly detection can detect more intrusions than each of them individually. These intrusions are those that used to generate new classifier based on Learn++ algorithm and will be detected in on-line phase of our framework in the next time.

6. Computational Complexity

After analysis pseudo-code of LEARN++ algorithm [12], we calculate that in training phase of our framework, the computational complexity of on-line phase is $O(nKT_k\alpha)$, where n is the number of instances, K is the number of available dataset for generating classifiers (H), T_k is the number of weak hypotheses that must be generated, α is the complexity of BaseClassifier which in our framework is simple multi layer perceptron. For testing phase, computational complexity of our framework is $O(n\alpha')$, where n is the number of test instances, α' is the complexity of BaseClassifier in testing phase.

Clustering algorithms can be divided in two categories [13]: similarity based and centroid based. Similarity algorithms have a complexity at least $O(N^2)$, where N is the number of instances. In contrast centroid-based algorithms have a complexity of O(NKM), where K is the number of clusters, M is the number of batch iteration and N is the number of instances. The on-line k-mean algorithm is a centroid-based which can be a desirable choice for on-line learning. Because it has high clustering quality, relatively lower complexity and fast convergence.

Our framework use simple multi layer perceptron in order to generate weak hypotheses. The complexities of these hypotheses are very lower than the strong classifier that can be constructed with neural network, so the framework has lower complexity than strong classifier.

In general, the frameworks that use ensemble of weak classifiers(using neural network) possesses lower computational complexity than other framework which using strong neural network, because of using neural network with suitable parameter to satisfy weak hypothesis. This property is very attractive and promising in intrusion detection, because classifiers should be retained in short periods in practice.

Any other classification algorithms can be used for generating weak hypotheses. This is a research interest in intrusion detection that we want to work on in future.

7. Conclusion

The fast growing of new intrusions in recent years led to an increasing interest in incrementally new intrusions. In this paper, we propose an incremental hybrid intrusion detection system to combine incremental misuse detection and incremental anomaly detection. For adaptively improving intrusion detection model to network behavior, we use an incremental intrusion detection system based on ensemble of weak classifiers for misuse component of our hybrid system. It has an ability to detect instances of data that belong to new class of intrusions in supervised mode. The new class of intrusion is not used in the previous training data.

There are some intrusions that could not detected by misuse intrusion detection. In order to detect these intrusions we use incremental anomaly detection along with misuse detection for implementing of hybrid intrusion detection to detect known attacks and unknown attacks.

The results of simulation show that combining misuse detection and anomaly detection can improve the productivity of intrusion detection systems

References

[1] Mounji, A., Charlier, B.L., Zampuniéris, D., & Habra, N. (1995). Distributed audit trail analysis. In D. Balenson & R. Shirey (Eds.), *Proceedings of the ISOC'95 symposium on network and distributed system security* (pp. 102--112), IEEE Computer Society, Los Alamitos, CA.

[2] Lindqvist, U., & Porras, P.A. (1999). Detecting computer and network misuse through the production-based expert system toolset (PBEST). In L. Gong & M. Reiter (Eds.), *Proceedings of the 1999 IEEE symposium on security and privacy* (pp. 146--161), IEEE Computer Socitey, Los Alamitos, CA.

[3] Ilgun, K., Kemmerer, R.A., & Porras, P.A. (1995). State transition analysis: A rule-based intrusion detection approach. *IEEE* Transactions *on Software Engineering*, *21* (3), 181--199.

[4] F. Neri, "Comparing local search with respect to genetic evolution to detect intrusions in computer networks," in Proceedings of the 2000 Congress on Evolutionary Computation, vol. 1, pp. 238–243, Mar-seille, France, July 2000. IEEE, IEEE. Source: IEEE Xplore

[5] J. Guan, D. X. Liu, and B. G. Cui, "An induction learning approach for building intrusion detection models using genetic algorithms," in Proceedings of Fifth World Congress on Intelligent Control and Automation WCICA, vol. 5, pp. 4339–4342. IEEE, June 2004.

[6] C. Kruegel, T. Toth, and E. Kirda, "Service specific anomaly detection for network intrusion detection," in Proceedings of the 2002 ACM symposium on Applied computing, pp. 201–208. ACM, Symposium on Applied Computing, ACM Press New York, NY, USA, Mar. 2002.

[7] Daniel Barbarra, Julia Couto, Sushil Jajodia, Leonard Popyack, and Ningning Wu, "ADAM: Detecting Intrusion by Data Mining", *Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security T1A3 1100 United States Military Academy*, West Point, NY, June 2001.

[8] Debra Anderson, Thane Frivold, And Alfonso Valdes, *Next-Generation Intrusion Detection Expert System (NIDES)-A Summary*, Technical Report SRICLS-95-07, SRI, May 1995.

[9] J. Zhang and M. Zulkernine, "A Hybrid Network Intrusion Detection Technique Using Random Forests," Proc. of the International Conference on Availability, Reliability and Security (AReS), IEEE CS Press, pp. 262-269, Vienna, Austria, April 2006.

[10] M. Locasto, K. Wang, A. Keromytis, and S. Stolfo. Flips: Hybrid adaptive intrusion prevention. In Proceedings of the 8th International Symposium on Recent Advances in Intrusion Detection (RAID), September 2005.

[11] Lei Xu, Adam Krzyzak, Ching Y, Suen, Methods of Combining Multiple Classifier and Their Application to Handwriting Recognition, IEEE TRANSACTION ON SYSTEMS, MAN AND CYBERNETICS, VOL. 22, NO. 3, MAY/JUNE 1992.

[12] Polikar R., Udpa L., Udpa, S., Honavar, V., "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Transactions on System, Man and Cybernetics (C), Special Issue on Knowledge Management*, vol. 31, no. 4, pp. 497-508, 2001.

[13] Shi Zhong, Taghi Khoshgoftaar, and Naeem Seliya Clustering-Based Network Intrusion Detection, International Journal of Reliability, Quality and Safety Engineering.