I J C T A, 9(28) 2016, pp. 93-98

INTEGRATED APPROACH FOR INTRUSION DETECTION USING CONDITIONAL RANDOM FIELDS WITH LAYERED APPROACH

M. Azhagiri* A. Rajesh and Divya Meena. S*****

Abstract: Intrusion Detection, one of the perplexing tasks for security professionals, is defined as the process of detecting actions that tries to compromise the CIA i.e., Confidentiality, Integrity and Availability of the system. The main goal of the intrusion detection system is to recognize the entities that attempt to destabilize the security controls in situ. Intrusion Detection System (IDS) can be a device or software application that is able to monitor the network activities for any malicious activities or any violations in the policy and must be able to cope with the increasing network traffic. As the network has become pervasive and the number of intrusion events increasing, every organization is implementing different system to monitor any cracks in IT Security [1]. Yet, intrusion detection is still a challenging issue. The accuracy and efficiency of the IDS is addressed using Conditional Random Fields (CRF) and Layered Approach respectively, in this paper. The result of CRF is promising than the data mining approaches like Decision tree and Naïve Bayes. The method is able to improve the attack detection accuracy for U2R and R2L attacks. To end with, the system has a benefit of increasing or decreasing the number of layers, based on the environment where the system is deployed and this gives flexibility to the network admins.

Keywords: Network security, Intrusion Detection, Conditional Random Fields, Layered Approach, Decision Tree, Naïve Bayes.

1. INTRODUCTION

Intrusion detection is defined as the process of detecting actions that tries to compromise the CIA i.e., Confidentiality, Integrity and Availability of the system. Its main responsibility is to inspect all the inbound and outbound network activities and to identify any apprehensive pattern that may be a network attack from someone who is trying to break compromise our system. According to SANS Institute, it is the art of detecting unsuitable, imprecise, or abnormal action. It is one of the most perplexing tasks for security professionals and network admins because, the possible technologies for attacking have become even more sophisticated and at the same time, the technical ability required for a novice attacker is very less because of the easy availability of the proven methods in the web. This increases the need to protect the network and the system from different types of susceptibilities. Also, steps have to be taken to detect the novel and concealed exploitations on the system by developing a more consistent and competent intrusion detection system that can produce only a minimal number of false alarms. But, all IDS are not efficient. Those that can handle large amount of traffic in the network, quickly in decision making and fewer false alarms can only become efficient IDS [2]. Started off in the year of 1980, IDS is classified into Network based IDS (NIDS) and Host based IDS (HIDS), depending on the deployment mode and data usage. Based on the attack detection method, IDS is classified into Signature based method and Anomaly based method. The main difference between the two is that, Signature based method extracts patterns from previously known attacks but anomaly based system learns from the normal data, where the anomalous activities are not recorded. A combination of Signature based method and Anomalous based method, commonly called

* Dept of CSE St.Peter's University, Avadi,Chennai-600054 **Email:** azhagiri1687@gmail.com
** Dept of CSE C Abdul Hakeem College of Engineering and Technology, Tamil Nadu 632509 **Email;** amrajesh73@gmail.com
*** Dept of CSE Kingston Engineering College Tamil Nadu 632059 **Email:** dhivya.mina18@gmail.com

as Hybrid method is more efficient and can provide better detection [3]. One of the concerns for Hybrid method is the availability of completely anomalous and attack-free data, but this is difficult to achieve.

Conditional Random Field (CRF) is a new probabilistic graph model that has the advantage of expressing the elements of long-distance dependent and overlapping features; normalizing all the features and solving the label bias problem of HMM. It shows a good performance when dealing with natural language activities such as English shallow parsing and English name reorganization of entity [4]. The accuracy and efficiency of IDS has been improving every other day. The main idea of CRF lies in the random process theory that connects all kinds of conjunction information and its relativity within the information's data sequence that includes the relations among feature sets itself [16]. After ascertaining the most probable classification of logged behaviours, it can move on to detecting the attacks and the normal discovery. On comparing with most other methods, CRF has been found to produce promising results.

2. CONDITIONAL RANDOM FIELDS FOR INTRUSION DETECTION

Conditional random field (CRF) was initially proposed by Lafferty and his colleagues in 2001 and is mostly based on MEMM (Maximum Entropy Markov Model) [17]. CRF can be combined with almost all variety of features, because of their sturdy inference power based index value style. It is a new probabilistic graph model that has the advantage of expressing the elements of long-distance dependent and overlapping features; normalizing all the features and solving the label bias problem of HMM[15]. They are helpful in modelling the conditional distribution for a set of random variables. As stated earlier, they are widely used in natural language processing activities. They do not make any unnecessary assumption on any observation and so they are said to be a better framework. Some of the Conditional models include CRF, MEM (Maximum Entropy Markov) and Maxent. One of the main features of CRF is that they are free from Label bias and Observation bias, as they are undirected in nature [5]. The following gives a brief description of the CRF.

Let X is a random variable over the data sequence that is to be labelled and let Y is a random variable over the corresponding label sequence. Let us assume that all the parts consisting Y as Y*i* are included in the fixed symbol sets of y. Let G be a graph connecting nodes and edges and is represented as G = (V, E) such that Y = (Y*v*), *v*εV and that Y is indexed by the vertex of G. (X, Y) becomes a CRF in case when (X, Y) is conditioned on X and the random variable Y*v* obeys the Markov property with respect to the graph G; [7]

$$p\,(Yv \mid X, Yw\,, w \neq v) = P\,(Yv \mid X, Yw, w \sim v) \quad (1)$$

where $w \sim v$ means that w and v are neighbours in graph G. Y is a tree in Graph G and its cliques are the edges and the vertices. By the joint distribution over the label sequence Y given X is in the form;

$$p\Theta\,(y \mid x)\ \alpha\ \exp\,(\Sigma\ \lambda kfk\,(e, y \mid e, x) + \Sigma\ \mu kgk\,(v, y \mid v, x\,)) \quad (2)$$

where X is a data sequence, Y is a label sequence, Y|*e* is a set that consist of parts of Y as defined by edge *e*. *y*|*v* is a set that consist parts of Y as defined by vertices V. Assuming that the feature fk and gk are given as a fixed parameter estimation is to train $\Theta = (\lambda 1\ ,\ \lambda 2\ ,..\mu 1\ ,\ \mu 2\ ..)$ out of the training data, i.e., the parameters in CRF model are determined by the distribution knowledge of the training data sets. The main goal is to improve the malicious attack detection accuracy. On comparing with other methods, CRF is found to be better in detecting the attacks, especially in case of "Unauthorized access to Root" (U2R), "Remote to Local" (R2L) and "Denial of Service" (DOS) attacks [8]. Though CRF is expensive for training and testing, the long-time benefit is high. The complexity for training simple linear structure CRF is O (TL2NI), where T is the length of sequence, L is the number of labels, and N is the number of

iterations. Intrusion detection has only two labels namely “Normal” and “Attack”[18]. The efficiency of the system can be improved with Layered approach, which can reduce the length of the sequence, T.

The intrusion detection system normally has to classify different features that are highly correlated and there exist a complex relationship between them. As a basic classification of “Normal and Attack” [9], the system has to take into account several features such as if the “system is logged in”, “how many files are created” and many more. Analysing this information individually will not provide any useful knowledge. Only on analysing them together, they will provide meaningful knowledge that can help in making the classification easier. The better performance of CRF when compared to others is mainly because they don’t analyse features individually. The features are represented in the form of sequence and the labels are assigned to every feature in the sequence. Though this increases the complexity, it also increases the intrusion detection accuracy.

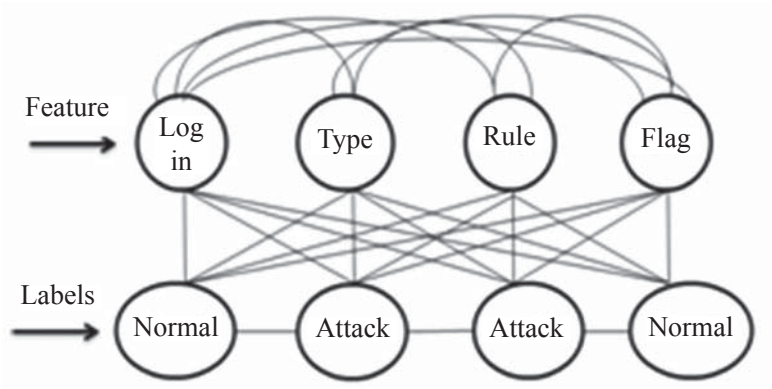

Figure 1. Labelling based on Dependency among Features

Every label is connected to each of the input features, indicating that only the combination of features can make an appropriate label for the feature and so CRF models using dependency among the features. No other model makes such dependency among features. One main advantage of such dependency is that, even if some data is missing, the feature can well be labelled with minimal number of features.

3. LAYERED APPROACH FOR INTRUSION DETECTION

In Layer-based Intrusion Detection System (LIDS), a number of confidence authorizations are performed in a sequence one after other. It signifies a sequential layered approach to ensure CIA of data over the network [7]. The first main goal is to reduce the computation required and the overall time required to detect the anomalous events in the network. Time to detect is a significant one and can be reduced only be eliminating the communication overhead among the different layers. To achieve this, the layers are made to act autonomous and self- sufficient enough to block an attack without requiring a central decision-maker. Each layer in the framework is trained individually and then they are deployed sequentially.

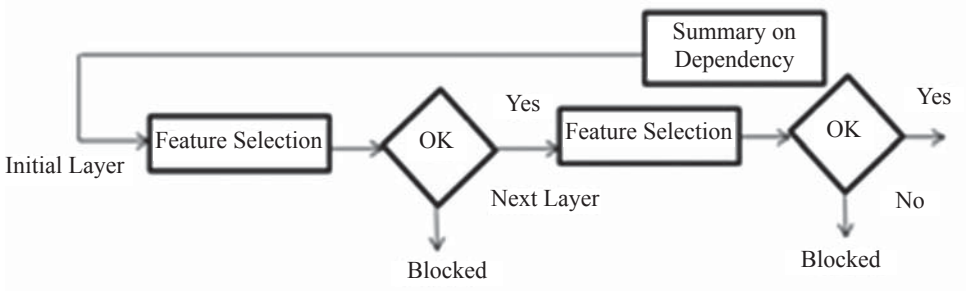

Figure 2. Layered Approach

We use Probe layer Dos Layer, R2L Layer and U2R Layer to correspond to the four different attack groups [8]. Each layer has small set of features relevant to it. This process of feature selection is most important. The layers have to act independently, so contradicting features are placed in more than one layer. Each layer will filter out and block anomalous connection and so the next layers need not perform any further processing and this makes a quicker response to the intrusion. The second most important goal is to improve the speed of the operation of the system. To improve the speed, the minimal set of features relevant to the layers ischosen. This type of feature selection improves both the training and testing speed. It is hard to achieve both accuracy and efficiency in intrusion detection, but the combination of CRF and LIDS can achieve both [9]. CRF and LIDS together called as Layered CRF has better performance than any other models.

4. INTEGRATING LAYERED APPROACH WITH CONDITIONAL RANDOM FIELD

While integrating the Layered approach with CRF, the first step is to connect the network by connecting each of the neighbouring nodes and deploying them independently in the network area. Port number for each node is authorized in the node itself [10]. IDS will appropriately detect the inappropriate, inaccurate and anomalous activities in the network.

1. ***Randomized Field Detection:*** The source or target file is chosen and the data selected is converted into fixed size of packets and is sent to the detector.
2. ***Probe layer:*** Probe layer is susceptible to probe attacks, which acquires information of the target network from a source that is external to the network [11]. This layer takes “duration of connection” and “source bytes” as essential features. Some other features like “number of files created an number of files accessed” does not provide any useful information for detecting probes.
3. ***DOS layer:*** This layer considers “section of connections having same host and destination”, “section of connections having same service”, “source bytes” and “section of packets with errors” as significant features. Other features like “login and log out” does not provide any useful information for detecting DOS attacks.
4. ***R2L Layer:*** This kind of attacks is the most complicate one to detect, because they involve feature from network and host. So, the significant features should include both network-level features such as “connection duration and service requested” and host level feature such as “number of failed attempts to log in”.
5. ***U2R layer:*** This type of attack is also difficult because they involve semantic level details that are difficult to detect at an early stage. It mostly involves content based attacks and typically targets an application. So, the significant features include “number of files created and number of shell prompts appealed”.

The following chart describes the attack detection accuracy of four techniques namely Layered CRF, KDD, Decision Tree and Support Vector Machine [12]. Each of the technique is assessed for the accuracy in detection with respect to Probe layer, QoS layer, R2L Layer and U2R Layer.

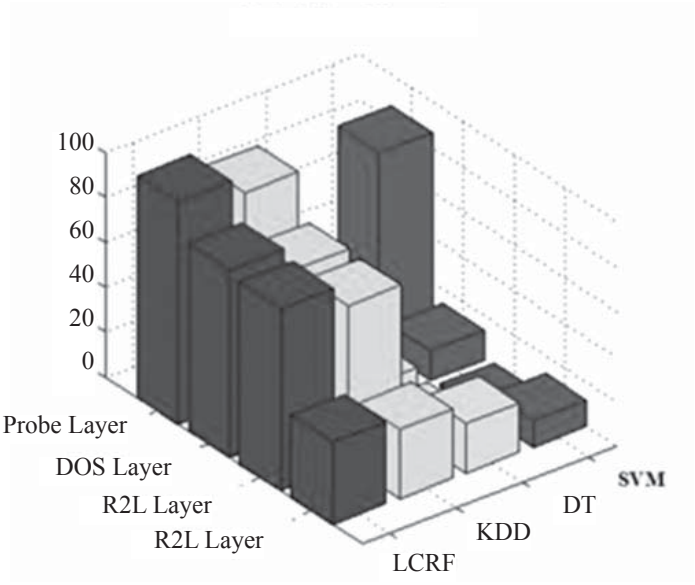

Figure 3. Probability of Detection

It is clear from the figure that LCRF outperforms all other techniques in all layers.

5. CONCLUSION AND FUTURE WORK

It is commonly known that decision trees and naïve Bayes algorithm are highly known for their performance in detection but Layered CRF performs much better than those methods and this is mainly attributed to the fact that, CRF does not consider features individually [13]. CRF provides higher accuracy and efficiency in detection and it is strong enough and can handle noisy data and still provide high performance. CRF acquires feature sets without pre-processing the data and can find out abnormal behaviour in the network accurately. This paper has focused on improving the accuracy and efficiency of the intrusion detection system. The model requires only a little time to train and test the data [14]. Once the attack is detected in a particular layer, the system accelerates the intrusion response and minimizes the effect of the attack. The number of layers required can be increased or decreased based on the environment in which the system is deployed and this gives flexibility to the network administrators. In future, the model will focus on developing the signatures for signature-based system that can be deployed at the end of the network to filter out the attacks that are frequently occurring and previously known. In this case, the new unknown attacks are left out for the anomaly and hybrid based system to detect. Also, the CRF model can be enhanced to compute and predict the network security situations.

Reference:

1. T. Abraham, IDDM: Intrusion Detection Using Data MiningTechniques,http://www.dsto.defence./gov.au/publications/2345/DSTO-GD-0286.pdf, 2008.
2. R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD, vol. 22, no. 2, pp. 207-216, 1993.
3. N.B. Amor, S. Benferhat, and Z. Elouedi, “Naive Bayes vs. Decision Trees in Intrusion Detection Systems,” Proc. ACM Symp. Applied Computing (SAC ’04), pp. 420-424, 2004.
4. J.P. Anderson, Computer Security Threat Monitoring and Surveillance, pdf, 2010. http://csrc.nist.gov/publications/history/ande80.
5. R. Bace and P. Mell, Intrusion Detection Systems, Computer Security Division, Information Technology Laboratory, Nat’l Inst. of Standards and Technology, 2001.

6. D. Boughaci, H. Drias, A. Bendib, Y. Bouznit, and B. Benhamou, "Distributed Intrusion Detection Framework Based on Mobile Agents," Proc. Int'l Conf. Dependability of Computer Systems (DepCoS-RELCOMEX '06), pp. 248-255, 2006.

7. Y. Bouzida and S. Gombault, "Eigenconnections to Intrusion Detection," Security and Protection in Information Processing Systems, pp. 241-258, 2004.

8. H. Debar, M. Becke, and D. Siboni, "A Neural Network Component for an Intrusion Detection System," Proc. IEEE Symp. Research in Security and Privacy (RSP '92), pp. 240-250, 1992.

9. T.G. Dietterich, "Machine Learning for Sequential Data: A Review," Proc. Joint IAPR Int'l Workshop Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR '02), LNCS 2396, pp. 15-30, 2002.

10. P. Dokas, L. Ertoz, A. Lazarevic, J. Srivastava, and P.-N. Tan, "Data Mining for Network Intrusion Detection," Proc. NSF Workshop Next Generation Data Mining (NGDM '02), pp. 21-30, 2002.

11. Y. Du, H. Wang, and Y. Pang, "A Hidden Markov Models-Based Anomaly Intrusion Detection Method," Proc. Fifth World Congress on Intelligent Control and Automation (WCICA '04), vol. 5, pp. 4348-4351, 2004.

12. S. Dzeroski and B. Zenko, "Is Combining Classifiers Better than Selecting the Best One," Proc. 19th Int'l Conf. Machine Learning (ICML '02), pp. 123-129, 2002.

13. L. Ertoz, A. Lazarevic, E. Eilertson, P.-N. Tan, P. Dokas, V. Kumar, and J. Srivastava, "Protecting against Cyber Threats in Networked Information Systems," Proc. SPIE Battlespace Digitization and Network Centric Systems III, pp. 51-56, 2003.

14. S. Forrest, S.A. Hofmeyr, A. Somayaji, and T.A. Longstaff, "A Sense of Self for Unix Processes," Proc. IEEE Symp. Research in Security and Privacy (RSP '96), pp. 120-128, 1996.

15. Y. Gu, A. McCallum, and D. Towsley, "Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation," Proc. Internet Measurement Conf. (IMC '05), pp. 345-350, USENIX Assoc., 2005.

16. K.K. Gupta, B. Nath, and R. Kotagiri, "Network Security Framework," Int'l J. Computer Science and Network Security, vol. 6, no. 7B, pp. 151-157, 2006.

17. K.K. Gupta, B. Nath, and R. Kotagiri, "Conditional Random Fields for Intrusion Detection," Proc. 21st Int'l Conf. Advanced Information Networking and Applications Workshops (AINAW '07), pp. 203-208, 2007.

18. K.K. Gupta, B. Nath, R. Kotagiri, and A. Kazi, "Attacking Confidentiality: An Agent Based Approach," Proc. IEEE Int'l Conf. Intelligence and Security Informatics (ISI '06), vol. 3975, pp. 285-296, 2006.