Network Security and Situational Awareness Data Pre-processing Method Based on Conditional Random Fields

Rajesh.P [1], Krishnamoorthy.P [2], Gopi.S[3], Sivasankari.S [4]
Assistant Professor CSE[*1,2,]
Assistant Professor IT [#3 #4,]
Kingston Engineering College, Vellore, India
Email - [1]rajeshpcse@kingston.ac.in, [2]krishnancse0206@gmail.com, [3]gopi.scse@gmail.com,
[4]sivasankari_cse@yahoo.co.in

Abstract: The examination of Network security situational Awareness (NSSA) is imperative since it can progress the system checking capacities, Emergency reaction limit and anticipate the advancement pattern of system security. In light of the substantial measure of Intrusion Detection System (IDS), We propose another strategy for information pre-processing for NSSA in light of Conditional Random fields (CRFs). It takes points of interest of the CRFs models which can line to arrangement information stamping and add irregular ascribes to manage the measure of information from IDS, and give the information to NSSA. It utilizes KDD Cup 1999 information sets as exploratory information and arrives at a conclusion that our proposed strategy is practicable, solid and productive. , This paper explains the situational familiarity with the three fundamental explore content: This paper expounds on the situational awareness of the three main research content: extraction the factors of NSSA, situation understanding and situation prediction.

Key Words: Network security situational Awareness (NSSA), Information, Detection, CRFs

INTRODUCTION:

With the broad use of system innovation, its scale is constantly growing and opening up, the system is influenced by different security dangers, for example, the intrusion of outside aggressors, Trojans, DDoS, worms, infections, inner assaults, and new sorts of assaults keep on emerging, for example, Web code infusion, Botnet and so forth. Some of customary measures are received to guarantee the system framework security, for example, firewall, IDS, infection identification, fixing vulnerabilities and so forth, yet these techniques are a piece of careful steps for assault conduct, arrange overseers cannot set up the system's status overall to locate the potential threats and take compelling measures.

In 1999, Tim Bass [1] proposed the idea of the internet circumstance mindfulness and built up a utilitarian structure for it, which developed a hypothetical establishment for resulting research on NSSA. Stephen G. Batsell,[2], Jason Shifflet[3] likewise made a comparable model which incorporated the current system security framework to understand the framework system, adapted to the substantial scale organize security occurrences. Yet, these strategies were just restricted recognition of assaults, which couldn't really execute the system security situational mindfulness. The system circumstance alludes to the present state and the adjustments in patterns of system which incorporates the operation of an assortment of system types of gear, system acts, and client practices and so on. It is significant that the circumstance is an express, a pattern all in all and the general idea. Organize security situational mindfulness is characterized to gain, comprehend, and show the security components which can change the system security state, and to anticipate the future improvement incline among the huge scale arrange environment. This requires to coordinated information of system security status which has a place with various levels and sorts, to measure organize security circumstance, to draw a guide of the present security circumstance state, and to give a premise basic leadership to director.

Xi'an Jiao tong [4] University executed an incorporated system security checking stage in view of IDS firewall, and assessed the system circumstance, and they proposed a strategy for quantitative progressive danger assessment show for system security in light of factual examination. In proposed a technique for utilizing nonstop shrouded Markov models (HMM) to quantitatively ascertain the danger of system security circumstance. The

strategy makes them deficiency of the presumption of yield autonomy, which brings about its failure to consider the elements of the specific circumstance and pick the right components.

Data mining is the non-insignificant procedure to separate the helpful data from a lot of information, which is avoided the expansive, deficient, boisterous, fluffy, irregular and down to earth information to locate the concealed, consistent and individuals obscure ahead of time, however possibly helpful and eventually reasonable data and learning. The extricated information can be communicated as an idea, rules, regularities, design and different structures. Information mining is the center connection of Knowledge Discovery in Database (KDD).

CONDITIONAL RANDOM FIELDS:

Conditional random fields (CRFs) are a class of statistical modeling method often applied in pattern recognition and machine learning, where they are used for structured prediction. CRFs are a type of discriminative undirected probabilistic graphical model [5]. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text or biological sequences and in computer vision. CRF on observations X and random variables Y as follows:

Let G=(V,E) be a graph such that Y= (Yv)v €V, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field when the random variables Yv, conditioned on X, obey the Markov property with respect to the graph: p(Yv|X,

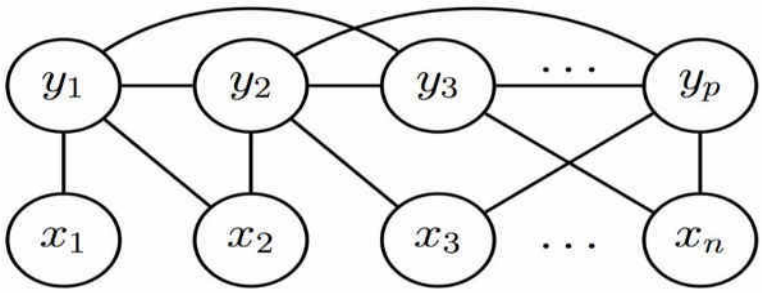

Fig 1: CRF

Yw,w≠v)=p(Yv|X, Yw,w≈v) Where w≈v means that w and v are neighbors in graph 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 Θ = (λ1 , λ2 ,..μ1 , μ2 ..) 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 [7]. 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"[8]. 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” , 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. Analyzing this information individually will not provide any useful knowledge. Only on analyzing 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 analyze 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.

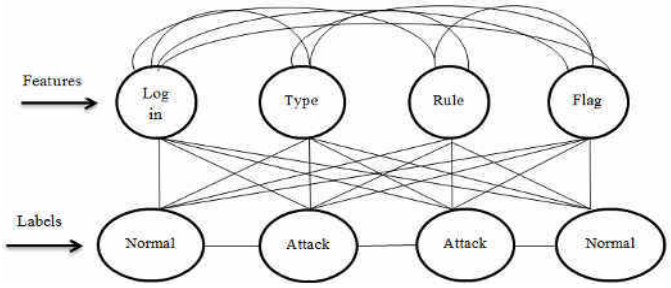

Fig 2: Labeling based on Dependency among Features

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 labeled with minimal number of features.

DESCRIPTIONS OF FEATURE SETS:

Experimental data used in CRFs models detection are KDD Cup 1999 data sets from standard database. Among them there are large numbers of normal network flow and various attack and have strong representative factors. Totally four attacks:

- DoS: denial-of-service, e.g. SYN flood, land attack;
- R2L: unauthorized access from a remote machine, e.g. guessing password;
- U2R: unauthorized access to local super user (root) privileges, e.g. various buffer overflow attacks;
- Probe: surveillance and other probing, e.g. port scanning.

A complete TCP connected talking is considered as a connection record, such as each UDP and ICMP packet. Each conjunction record is independent from other records. The basic property is the coherent property of each conjunction information such as area property, flow property and main processor flow property which are abstracted property relative to intrusion detection by Wenke Lee through data mining and comparing between normal style and intrusion style, and it has 41 different features which can be classified as 4 feature sets: Basic feature sets, Content feature sets, Flow feature sets, Traffic of hosts feature sets.

DESIGN OF A MULTI-LEVEL ANALYSIS FRAMEWORK OF NSSA:

A multi-level investigation system of NSSA, which roll out a little improvement from Ensley s three level model of circumstance mindfulness. To begin with it suggests that each sort of information ought to have a comparing procedure motor for distinguishing the information has a place with a specific element. Second, it partitions the recognition into two sections; calculate distinguishing proof and connection rules, on the grounds that the reason for observation is to get information of who will partake in the exercises and how they act. Last, it clears up that the center procedure of NSSA is circumstance assessment, and this procedure will produce the information of current circumstance and after that conjecture the circumstance in two days or a week time.

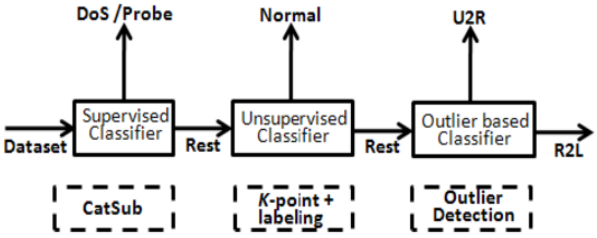

Fig:3 multi-level analysis framework of NSSA

EVALUATION INDEX:

To evaluate the capability of proposed model, we adopt following statistics measures as the Test standard:

Accuracy

Accuracy is the nearness of measurement results to the true value.

Accuracy = number of correct judged classified sample / number of total sample

$$Acc = \frac{Cl.S1}{S}$$

CONCLUSION:

This paper focuses on the challenges of Network Security Situation Awareness and tries to resolve it using CRF. We point out that the relationship between the situation evaluation and the situation awareness, and then propose a method for situation evaluation. According to the proposed model, we implemented a situation awareness system. The evaluation of a simulated network indicates that the approach is suitable for network environment, and the evaluation results are precise and efficient.

REFERENCES:

1. Bass, T., "Intrusion Detection Systems and Multisensory Data Fusion," Communications of the ACM, Vol. 43, No. 4, April 2000.
2. Bat sell S G, etc. "Distributed Intrusion Detection and Attack Containment for Organizational Cyber Security". http://www. ioc. Ornl. gov/projects/documents/containment.pdf, 2005.
3. Shifflet J. "A Technique Independent Fusion Model for Network Intrusion Detection". Proceedings of the Midstates Conference on Undergraduate Research in Computer Science and Mathematics, vol.3, 2005, pp.13-19.
4. Huimin Zhang, etc. "Study and implementation of integrated network security monitoring system". Journal on Communications, vol.24, 2003, pp.155-163.
5. J. Lafferty, A. McCallum, and F. Pereira. "Conditional random fields: probabilistic models for segmenting and labeling sequence data". In International Conference on Machine Learning, 2001. Pp.282–289.
6. Kapil Kumar Gupta, Baikunth Nath, Kotagiri Ramamohanarao, "Conditional Random Fields for Intrusion Detection". Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops. Melbourne, Australia, AINAW.2007, pp. 203-208
7. Jianping Li, Huiqiang Wang, Jianguang Yu. "Research on the Application of CRFs Based on Feature Sets in Network Intrusion Detection". Proceedings - 2008 International Conference on Security Technology, SecTech 2008, pp.194-197.