2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.2.31

A Method of Data Preprocessing for Network Security Situational Awareness Based on Conditional Random Fields

Aiping Lu[+] and Jianping Li

School of Computer and Information Technology, Northeast Petroleum University

Daqing, China, 163318

Abstract—Network Security Situational Awareness(NSSA) has been a hot research in the network security domain. Because of the large amount of Intrusion Detection System (IDS), We propose a new method of data preprocessing for NSSA based on conditional random fields(CRFs). It takes advantages of the CRFs models which can stitch to sequence data marking and add random attributes to deal with the amount of data from IDS, and provide the data for NSSA. It uses KDD Cup 1999 data sets as experimental data and comes to a conclusion that our proposed method is practicable, reliable and efficient.

Keywords- network security; situational awareness;conditional Random Fields

1. Introduction

With the extensive application of network technology, its scale is continually expanding and opening up, the network is affected by various security threats, such as the invasion of external attackers, Trojans, DDoS, worms, viruses, internal attacks, and new types of attacks continue to emerge, such as Web code injection, Botnet etc. Some of traditional measures are adopted to ensure the network system security, such as firewall, IDS, virus detection, patching vulnerabilities etc, but these methods are part of precautionary measures for attack behavior, network administrators can not establish the network's status as a whole to find the potential dangers and take effective measures.

In 1999, Tim Bass[1] first proposed the concept of cyberspace situation awareness and established a functional framework for it, which constructed a theoretical foundation for subsequent research on NSSA. Stephen G. Batsell[2], Jason Shifflet[3] also made a similar model which integrated the existing network security system to realize the system framework, coped with the large-scale network security incidents. But these methods were only limited detection of attacks, which could not truly implement the network security situational awareness. The network situation refers to the current state and the changes in trends of network which includes the operation of a variety of network equipments, network acts, and user behaviors etc. It is worth noting that the situation is a state, a trend as a whole and the overall concept. Network security situational awareness[4] is defined to acquire, understand, and display the security elements which can change the network security state, and to predict the future development trend among the large-scale network environment. This requires to integrated data of network security status which belong to different levels and types, to quantify network security situation, to draw a map of the current security situation state, and to provide a basis decision-making for administrator.

Xi'an Jiaotong University implemented an integrated network security monitoring platform based on IDS and firewall[5], and evaluated the network situation, as well as they proposed a method of quantitative

[+] Corresponding author.
E-mail address: leejp@126.com

hierarchical threat evaluation model for network security based on statistical analysis[6]. Paper [7] proposed a method of network security situational awareness based on Rough Set theory. Haslum[8] proposed a method of using continuous hidden Markov models (HMM) to quantitatively calculate the threat of network security situation. The method has one shortcoming of the assumption of output independence, which results in its inability to consider the features of the context and choose the right features. Conditional Random Fields(CRFs)[9]is a new model of the probabilistic graph, it has the advantage of expressing the elements of long-distance dependency and overlap features, normalizes all of features and solves the *label bias* problem of HEMM. CRFs Model has showed good performance in dealing with natural language tasks such as English shallow parsing and English name reorganization of entity. Recently, Kapil Kumar Gupta[10], Jianping Li[11] have used CRFs Model in IDS, and improved the detection accuracy and efficiency, but the application of data preprocessing for network security situational awareness based on CRFs has not been reported.

In this paper we put forward a CRFs model based on feature sets to data preprocessing for NSSA. No need to prepare knowledge and deal with the training data and data assumption, this model obtains CRFs' features, and is used to study abnormal structure data sets and establish CRFs detection model to mark irregular data. The essence of CRFs is based on random process theory to connect all kinds of conjunction information and its relativity within the information data sequence which includes relations among feature sets. After ascertain the most possible classification of recorded behaviors, it can move to attack detection and normal discovery. According to the results in the experiments, we know that after comparing with former technologies, CRFs can be more suitable to intrusion detection.

2. Conditional Random Fields

Conditional random fields(CRFs) was firstly proposed by Lafferty and his colleagues in 2001[9], whose model ideal mainly came from MEMM(Maximum Entropy Markov Model). Just like the MEMM, CRFs models are also index value style which have strong inference power and can be mixed with all kinds of features. CRFs calculates the probability distribution of the whole sequence, when the observing sequences waiting for marking are given, but not to define the next state distribution under current state condition. This distributed conditional feature of label sequence makes CRFs well fit to the real world's data. In these data, condition probability of label sequence relies on the dependent, mutual effect features in observing sequence, and by giving these features different weight values to show the variety importance of them.

Define *X* is a random variable over data sequences to be labeled, *Y* is a random variable over corresponding label sequences. Assuming all the consisting parts of *Y* as *Yi* included in fixed symbol sets of *y*. For example, *X* may include connective record of data sequence, while *Y* includes the sequence of record type label. *Y* refers to a set recording type labeled types set.

Definition: Let $G=(V,E)$ be a graph such that $Y=(Y_v)_{v\in V}$, so that *Y* is indexed by the vertices of *G*. Then (X,Y) is a conditional random field in case when conditioned on *X*, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v \mid X, Y_w, w\neq v)=p(Y_v \mid X, Y_w, w\sim v)$, ($w\sim v$ means that *w* and *v* are neighbors in *G*). In the graph $G=(V,E)$ of *Y* is a tree, its cliques are the edges and vertices. Therefore, by the fundamental theorem of random fields, the joint distribution over the label sequence *Y* given *X* has the form:

$$p_\theta(y \mid x) \propto \exp\Big(\sum_{e\in E,k}\lambda_k f_k(e, y|_e, x)+\sum_{v\in V,k}\mu_k g_k(v, y|_v, x)\Big) \tag{1}$$

X-data sequence, *Y*-label sequence. $Y|_e$ is a set of consist parts of *Y* defined by edge e. $y|_v$ is a set of consist parts of *Y* defined by vertices *V*. Assuming featured f_k and g_k are given and fixed, parameter estimation is mainly train $\theta=(\lambda_1,\lambda_2,\ldots;\mu_1,\mu_2,\ldots)$ out of training data, that is to say, parameters in CRFs models are ascertained by the distribution knowledge of training data sets.

In the experiments, for the conjunction record sequence X and record type sequence y, we can define a linear CRFs model as follows:

$$p(y|x)=\frac{1}{Z(x)}\exp\Big(\sum_{i,k}\lambda_k f_k(y_{i-1}, y_i, x)+\sum_{i,k}\mu_k g_k(y_i, x)\Big) \tag{2}$$

Z is normalization factor. Each $f_k(y_{i-1}, y_i, x)$ means features of input nodes and output nodes located between i and $i-1$ in observing sequence x. While $g_k(y_i, x)$ is features of in put nodes and output nodes located in i, λ and μ mean the weights of featured function. Weights parameters are made out by studying training data. Due to the differences among properties and among property sets, as well as differences among different condition information, it comes to a conclusion that the weights parameters of CRFs would be different. During the test, we work out the y of maximum the P(y|x) by taking the advantages of parameters and features which are turn out during the training, which is finding the best marking property from all possible outcomes.

3. 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.

The conjunction characteristic of each attack is incompletely the same, but they have a lot of features in common. By Data Mining method, Wenke Lee and his colleagues discovered these shared features and discovered that it is more efficient to detect different attacks with different property sets. The attacks of DoS and Probe need mainly detect which based on basic features and flow feature group. However the attacks of R2L and U2R need mainly detect which based on basic features and content feature group.

4. Data Preprocessing Model for NSSA Based on CRFs

4.1. Data Preparation

Experiments in the thesis only adopt the 10_percent data concentrated and separately provided by KDD Cup 1999 data, totally 494021 records.

According to the four features talked above, we cut each record in original data sets into four sub-records, each of which has a symbol standing feature set, respectively is A, B, C, D and 41 properties, and in the last row of the sub-record is the great style's name namely DoS, Probe, R2L, U2R or Normal. In sub-records, except one property set relative data in original record, other properties are all considered 0. Four sub-record make up to a new observing sequence.

The 10_percent data sets are divided into four teams, respectively marking 0, 1, 2, and 3.

4.2. Evaluation Index

To evaluate the capability of CRFs detection model, we adopt following seven statistics measures as the test standard:

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

Precision = TP/(TP+FP), TP means amount of the correct judged samples of the positive, FP means amount of the correct judged samples of the negative.

Recall rate= TP/(TP+FN), FN means amount of the incorrect judged samples of the positive.

F-Value = (2*precision* recall rate) / (precision + recall rate);

Detection rate = number of correctly detected intrusion / number of total intrusion in test sets

False alarm rate = normal sample mistaken for abnormal sample / number of total normal sample

Missing alarm rate = abnormal sample mistaken for normal sample / number of total abnormal sample.

4.3. Evaluation of Credibility Based on Restricted Forward-backward Algorithm

Forward-backward algorithm[12] can be used to calculate the probability for all possible state sequence under the condition of the given sequence of observation. In CRFs Model, we need to define a modified "forward variable" $\alpha_t(s_i)$, the recursion definition as follow:

$$\alpha_{t+1}(s) = \sum_{s} \alpha_t(s') \exp\left(\sum_{k=1}^{K} l_k f_k(s', s, o, t) \right) \tag{3}$$

In order to evaluate the credibility of the host's state, it must restrict the forward-backward algorithm as: each path must be through the sub-path which satisfied with constraint $C = \langle s_t, s_{t+1}, ... \rangle$, in which, $s_t \in C$ may be a positive constraint(the sequence should be through s_t) or a negative constraint t(the sequence should not be through s_t). In identifying the host's state, constraint *C* is corresponding to the identified host's state, the positive constraint in *C* is the internal state of host, and the negative constraint is the border of the host's state. In restricted forward-backward algorithm, the forward variable value must be satisfied with constraint *C*. The model based on CRFs, for all $s_q \in C$, is corresponding to define the restricted forward variable as follow:

$$\alpha_q(s_i) = \begin{cases} \max\left[\alpha_{q-1} \exp\left[\sum_k l_k f_k(s', s_i, o, t) \right] \right], if s_i \cong s_q \\ 0, \quad otherwise \end{cases} \tag{4}$$

In the above formula, the symbol $s_i \cong s_q$ denotes that s_i is satisfied with the constraint $s_i \cong s_q$. If $\alpha'_{t+1}(s_i)$ is a restricted forward variable, $Z'_0 = \sum_i \alpha_T'(s_i)$ is the restricted lattice value, therefore, the credibility of the identified host's state can be denoted as the normalization of Z'_0, that is, Z'_0 / Z_0, $Z_0 = \sum_i \alpha_T(s_i)$.

5. Experiment and Result Analysis

5.1. Training Method of Feature Parameters

The goal of training the CRFs model is to maximize the log-likelihood for the training data set under the condition of the given training data set (9).

$$D = \{ < o, l >^{(1)}, ..., < o, l >^{(j)}, ..., < o, l >^{(N)} \tag{5}$$

$$L_{\wedge} = \sum_{j=1}^{N} \log(P_{\wedge}(l^{(j)} | o^{(j)})) - \sum_{k=1}^{K} \frac{l_k^2}{2\sigma^2} \tag{6}$$

In the above formula, the second item is the Gauss transcendental value which can smooth the feature parameters. σ^2 is denoted as the transcendental variance. In this paper, we use the L-BFGS[12] algorithm to solve the optimization objective function.

5.2. Analysis of Experimental Results

The four teams of data sets in experiment are all including four attack types and normal types of data and the rate of the same type in each team are almost same. Our aim of experiment is that efficiently and correctly separating all kinds of data when different types of normal and abnormal data are mixed together.

Team 1, 2 and 3 of data sets are used as training data sets separately and get three kinds of CRFs detection models, and use data set 0 as test data set. To eliminate the unbalance of the data sets, we will take the average value of three experimental results. When the CRFs model detects the sequence, it will judge the connected record property sets to mark the great type of each sub-record. To get the best classification, we take the mutual marking results of the four sub-records as the original conjunction record types being recognized. The detection average results of the three teams are showed as following table I and table II.

TABLE I. EXPERIMENT RESULTS OF STATISTICS BY EVALUATION INDEX

	Accu-racy	Detection rate	Missing alarm rate	False alarm rate
Average of the three experiments	99.97%	99.97%	0.03%	0.03%

TABLE II. EXPERIMENT RESULTS OF STATISTICS BY ATTACK TYPES

	Average Value	
0 test dataset	Precision/%	Recall rate/%
Normal	99.91	99.97
DoS	100.00	100.00
U2R	94.44	60
R2L	98.16	96.41
Probe	99.55	97.15

From the average value of the three experimental statistics results, the detection efficiencies can meet the requests. From the situation of detection, in the four attacks, DoS and Probe are both with high accuracy while that of R2L and U2R are relatively lower. That's because of the large sample quantity and various categories of training data of DoS and Probe. But the sample quantity and categories of R2L and U2R are relatively less. Especially the U2R has the least sample quantity, with 18 records in each team at most, that is the reason for U2R's low recall rate.

From the detection of U2R and the comparison with other attack's test results, we know that the recall rate and accuracy will increase along with the quantity and categories of sample. Moreover, in each data group of training, R2L has 288 records at most, leading to a high level of accuracy and recall rate.

In addition, we have used training data two times larger than test data, and compared required test model with the former quantity's test set model. The results show that all evaluation index of U2R type detection are obviously improved, with evaluation standards results of other types are almost same. All of these show that the CRFs need high quantity and categories of sample. By using small quantity samples to train it can create a more powerful detection model. For the detection model through training data of abundant sample varieties, whose detection performances enhance are not so obvious comparing with that of more abundant ones.

In the following part, we compare the results of detection precision with the results from SVM and multi-class SVM. The detailed as data is show in table III.

TABLE III. DETECTION PRECISION OF VARIOUS DETECTION MODELS

	CRFsFS /%	SVM /%	Multi-class SVM /%
DoS	100.00	76.86	97.00
U2R	94.44	66.7	78.51
R2L	98.16	31.58	24.91
Probe	99.55	93.24	73.55

Dissimilar with other detection models, when we adopt CRFs detection model there is no need to preprocess the data and it can fully used various featured information of conjunction data. From the comparison in table 3, we can see the precision of CRFs detection model of detecting attack type is higher than that of other methods. That means the CRFs detection model has higher anti-interference, and higher power than other methods.

To sum up, advantages of using CRFs based on features sets in network intrusion are as following:

(1) CRFs can consider the dependent relations between features and feature sets well, which is also advantage of CRFs itself.

(2) While in training if the sample category is enough, the amount of sample it needs is little.

(3) In test process, it can recognize data with many types and with a higher accuracy, detective rate, precision, detective speed while lower false alarm rate and missing alarm rate.

(4) It is very efficient to detect the huge data set.

6. Conclusions

In this paper, we present in detail a data preprocessing method for network security situational awareness based on the CRFs and show the encouraging results on theKDD Cup 1999 data sets.

This thesis makes use of the characteristics of the CRFs marking and slice cutting sequence data process, and various feature information of network conjunction information data, to establish CRFs detection model based on feature sets. CRFs has strong learning power toward detection samples. It can acquire detection model based on feature sets without preprocess with data, and can find out abnormal behavior accurately. This kind of detection method is not only theoretically valid, but also can be applied in actual system.

7. Acknowledgment

The authors should thank the other cooperators in this project for their contributions in this paper. We are also grateful to the anonymous referees for their insightful comments and suggestions, which clarified the presentation. This work is supported by the National High Technology Research and Development Foundation of China(863). (No.2007AA01Z401).

8. References

[1] Tim Bass. “Intrusion Detection System and Multi-sensor Data Fusion”.Communications of the ACM, 43, 4 (2000), pp.99-105.

[2] Bat sell S G,etc. “Distributed Intrusion Detection and At tack 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 Mat hematics , vol.3,2005,pp.13-19.

[4] Huiqiang Wang,etc. “Survey of Network Situation Awareness System”.Computer Science,vol.33,2006,pp.5-10.

[5] Huimin Zhang, etc. “Study and implementation of integrated network security monitoring system”.Journal on Communications,vol.24,2003, pp.155-163.

[6] Xiuzhen Chen,etc. “Quantitative Hierarchical Threat Evaluation Model for Network Security”.Journal of Software, vol.17,2006 ,pp.885-897.

[7] Ying Liang, Huiqiang Wang, Jibao Lai. “A Method of Network Security Situation Awareness Based on Rough Set Theory”. Computer Science,vol.34, 2007,pp.95-147.

[8] Haslum Kjetil , Arnes André. “Multisensor real-time risk assessment using continuous time hidden Markov models”.Proceedings of the International Conference on ComputationalIntelligence and Security (CIS) . Guangzhou , China , 2006 :6942703.

[9] 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.

[10] 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

[11] 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.

[12] Junsheng Zhou,etc. “Automatic Recognition of Chinese Organization Name Based on Cascaded Conditional Random Fields”. Chinese Journal of Electronics,vol.34,2006.pp.804-809.