Scholars Journal of Engineering and Technology (SJET)
Sch. J. Eng. Tech., 2015; 3(3B):327-333
©Scholars Academic and Scientific Publisher
(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

ISSN 2321-435X (Online)
ISSN 2347-9523 (Print)

Research Article

A New Algorithm of Network Intrusion Detection base on the Application of Conditional Random Fields

Jianping Li*, Siyuan Zhao
School of Computer and Information Technology, Northeast Petroleum University, Dqing 163318, China

***Corresponding author**
Jianping Li
Email: leejp@126.com

Abstract: While the network brings convenience to people, its own fragility offers intrusion opportunities for hackers and malicious attackers. Along with the diversity and complexity of intrusion attack, high performance intrusion detection techniques are required, and so the study of on-line detection, adaptive detection and multiclass detection techniques becomes current hotspot. To improve the performance of multiclass intrusion detection system(IDS), this thesis puts forward a method of CRFs (Conditional Random Fields) based on attribute sets in IDS. This Algorithm uses varied connection information and its relativity in network connection information data sequence as well as the feature sets relativity in data sequence to attack detection and discovery of abnormal phenomenon. In this thesis, after the discussion of the work process of the models and the comparison between KDD cup'99 data sets' detective conclusion and other test methods. The simulation results show that the proposed algorithm is practicable, reliable and efficient.

Keywords: Network Security, intrusion detection system, Conditional Random Fields, KDD cup'99

INTRODUCTION

With the constant development of computer technology, new network attacking means also continuously renewed, such as violent attack, network tapping, original code analysis, IP masquerade, service attack decline, network scan, dispersed attack, attack applying known loopholes and defects of agreements and etc. Because of its opening attributes, computer information is characterized by character of sharing and easy spreading. And varied types of attacking are constantly emerging.

In the circumstance that more and more attacks discovered or undiscovered keep on emerging, current IDS (Intrusion Detection System) can't detect all the insisting attack under the condition of its acceptable reliability, and have many defects such as low accuracy, high rate of wrong warning and leaked warning. Former researchers brought many kinds of detecting methods into IDS, such as Data Mining technology[1][2], Decision Tree[3][4],SVM[5][6],HMM[7] and so on. These classified models are based on known knowledge and data hypothesis to establish more precise classified models, to correctly differentiate normal and abnormal behaviors, and to enhance the IDS detection rate.

Models in the past were mostly established under the premise of known knowledge or presumed data, intending to form ideal classified models. However, the data acquired form intrusion detection area always couldn't meet the needs of the system study, but presenting variability, small samples[8], wouldn't surely accord with those model training samples. In addition, in the process of establishing a model, we should pre-process the training data before picking out some of the data or property and be sure about the special treatment for the few types of abnormal behavior samples. This is necessary for a more accurate reliable detection system.

Recently, Kapil Kumar Gupta and his colleagues applied CRFs in the intrusion detection[9]. They establish a model according to CRFs' characteristics[10], properties and features between properties. But this method regardless of the connection between feature sets, did not present the capability of CRFs.

CRFs can be regarded as a undirected diagram model, which is a data statistics frame model to mark and separate sequence data. The model can use the relations between the properties to mark the sequence. CRFs has showed good performance in dealing with natural language tasks such as English shallow parsing and English name reorganization of entity[11][12][13]. The characteristics and study achievements show it is capable to deal with many sequence marking research tasks.

We put forward in this paper that this CRFs based on feature sets to network intrusion detection. No need to prepare knowledge and deal with the training data and data assumption, this model obtained CRFs' features, is used to study abnormal structure data sets before 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 result in the experiments, we know that after comparing with former technologies, CRFs can be more suitable in detecting intrusion.

This article is organized as follows: We simply introduce CRFs theory in the second part, describe the network conjunction feature sets in the third part, discussion about not only CRFs based on feature sets' detection model and its applications, but also experiment results in the forth part, and the conclusions in fifth part.

Conditional Random Fields(CRFs)

CRFs was firstly proposed by Lafferty and his colleagues in 2001[10], whose model ideal mainly came from MEMM(Maximum Entropy Markov Model). CRFs is a undirected diagram model which calculate and output the conditional rates of nodes when the input nodes condition is given. It is a differentiation models maintaining the advantage of conditional rate frame in MEMM, and overcoming the shortcomings of generation model of HMM and also overcoming the strictly independent assumption condition without any additional features[14]. While CRFs also solved the problem of label bias of MEMM and other disadvantage of non-generation models[10]. Just like the MEMM, CRFs models are also index value style which have strong inference power and can use complex, over trapped and dependent features to train and deduct. In addition, the differentiation models insist the features of observing data determine the states, and can be mixed with all kinds of features.

CRFs calculate the probability distribution of the whole sequence, when the observing sequence waiting for marking are given, but not to define the next state distribution under current state condition. This distribution condition property of label sequence makes CRFs well while appropriates the real world's data. In these data, condition probability of label sequence is rely on the dependent, mutual effect features in observing sequence, and by giving these features different Weight values to show the variety importance of them.

Visit and operation between main processors in network will produce a series of conjunction information, so in the information records external users' visits to the local computers. Each piece of conjunction information can be regarded as a sequence. The judgment to the visit behavior can also be a judgment to the conjunction information. Therefore, according to its features, the thesis classifies the connective record information by CRFs models. The Differences from the methods in Kapil Kumar Gupta's thesis are as followings, we not only use the property and information between properties but also fully used the training focused on information between property sets as features when marking the record data, to provide more features information for data marking.

3. Descriptions of Feature Sets

Experimental data used in CRFs models detection are KDD cup 1999 data sets[15] from standard database. The data sets are a gathered network conjunction record sets whose original data is resumed conjunction information based on data required by Wenke Lee and his colleagues, in 1998, in DARPA's IDS estimation[1]. Five million conjunction records used as training data and about two million used as test data. 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. synflood;
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, so are each UDP and ICMP packet. Each conjunction record is independent from other records. And the basic property is the coherent property of each conjunction information. While area property, flow property and main processor flow property are abstracted property relative to invasion 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 following 4 feature sets[1]:

(1)Basic feature sets, such as connective continuous time, agreement, service, the number of send out bytes and the number of received bytes etc.

(2)Content feature sets, which use area knowledge to acquire property relative with information packets, such as numbers of hot mark in conjunctions, times of failure debarkation, whether successfully debark or not, etc.

(3)Flow feature sets, which are the sets based on time and network flows. It can be divided in to two sets; one is Sam Host feature sets which include some relative negotiation, service statistics information in the conjunction during last 2 seconds and current connective main processor with same aim; the other one is Same Service feature set which can make out some statistics information during last 2 seconds and current same connective service.

(4)Traffic of hosts feature sets, namely features related with network traffic which is based on hosts. This kind of features are to discover slow scan, the way of obtaining is dealing with statistics of past 100 covariance features in conjunction, such as the number of connective host purposes between the same old ones host purpose and currents ones , as well as the ratio of the same service conjunction.

The conjunction characteristic of each attack is also incompletely the same, but they have a lot of features in common. By data mining, Wenke Lee and his colleagues discovered these shared features and discovered that it is more efficient to detect different attacks with different property sets[1]. The attacks of DoS and Probe need mainly detect which based on basic features and flow features group. However the attacks of R2L and U2R need mainly detects which based on basic features and content features group.

Application of CRFs Based on Feature Sets Model

Definition of CRFs and its Detection Model

Define X as random variable in data sequence to label, Y is relevant label sequence random variable. Presume all the consisting parts of Y as Y_i 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: Given an undirected diagram $G = \left(V,\ E\right)$, V as the top points set, E as the edges set. Then make top points as the index of labeled Y, that is $Y = (Y_v)_{v\in V}$,Y of every top point can be a random label in labels set. When the appearance of Y rely on X and the random variable sequence of Y_V according to diagram structure, which is the same as Markov's, namely $p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$, ($w \sim v$ refers to the connecting edge between two top points), we name (X, Y) a conditional random area. In $G = (V, E)$, Y is a tree, and the son-diagrams are edges and top points, So according to the basic theory of random area model, we can describe the conjunction distribution of given *Y-label* sequence of X , as follows:

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

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 top point *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, it is to say, parameters in CRFs models, are ascertained by the distribution 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:

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

where 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 and among property groups, as well as differences among different condition information, it comes to a conclusion that the weights parameter of CRFs would be different. During the test, we work out the y of maximum the *P(y|x)* by taking the advantages of parameter and features which are turn out during the training , which is finding the best marking property from all possible outcomes.

Data Preparation

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

Among the *10_percent* data sets, DoS attack record holds the most part, reaching to 390 thousand recodes. And the one next to it is Normal type which holds 97 thousand records. While the attacks of Probe and R2L hold above one thousand records. And U2R hold the fewest, only tens of records.

According to the four features talked above, we cut each record in original data set into four son-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 son-record is the great style's name namely DoS, Probe, R2L, U2R or Normal. In son-records, except one property set relative data in original record, other properties are all considered 0. Four son-record make up to a new observing sequence.

Each conjunction record information converts like this, when CRFs model established, it can melt into different attack types property characteristics and also can melt into the relative information of properties or feature sets.

Experiments in the thesis employ different structure data sets including discrete data and constant data, without preprocessing to data such as cutting out noises. And the data are all original without any assumption. The 10_percent datasets are divided into four teams, respectively marking 0, 1, 2, and 3. Through covariance, we discover that the distributions of each team are basically the same. In order to compare with Kapil Kumar Gupta's method, 10_percent data can be divided into different teams according to his thesis[9].

Evaluation Index

To evaluate the capability of CRFs detection model, we adopt following eight 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 = TP/(TP+FN), FN means amount of the incorrect judged samples of the positive

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

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

Average detection time = total detection time / total sample

Experiment results and analysis

The four teams of data sets in experiment are all including four attack types and normal types of data and the ratios of the same type in each team are almost the 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 trainings 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 experiment results. When the CRFs model detect the sequence, it will get all kinds of feature information by training among which feature information between property sets to mark each son-record. And it is also equal to judge the connected record property sets to mark the great type of each son-record, namely to mark out the four great attack types and normal behavior types. To get the best classification, we take the mutual marking results of the four son-records as the original conjunction record types being recognized. The detection results of each team are showed as following table-1and table-2.

Table-1: Experiment results of each team statistics(1)

	Accuracy/%	Average time of detection/ record/ms	Detection rate	Missing alarm rate	False alarm rate
1 training dataset	99.97	0.62ms	99.97%	0.03%	0.05%
2 training dataset	99.98	0.62ms	99.98%	0.02%	0.03%
3 training dataset	99.97	0.62ms	99.97%	0.03%	0.02%
Average of the three experiments	99.97	0.62ms	99.97%	0.03%	0.03%

Table-2: Experiment results of each team statistics(2)

	1 training dataset		2 training dataset		3 training dataset		Average Value	
0 test dataset	Precision/ %	Recall/%	Precision/ %	Recall/%	Precision/ %	Recall/%	Precision /%	Recall/%
Normal	99.92	99.95	99.92	99.97	99.89	99.98	99.91	99.97
DoS	100.00	100.00	100.00	100.00	100.00	99.99	100.00	100.00
U2R	83.33	50.00	100.00	70.00	100.00	60.00	94.44	60
R2L	97.79	96.38	98.18	96.46	98.52	96.38	98.16	96.41
Probe	99.22	93.32	99.71	99.02	99.71	99.12	99.55	97.15

The number of 100.00 in table II is a value through accepting and rejecting, so are all the data in all the tables in this paper.

In the experiment, we discover that the results of the marking appears of son-recode which been output, wrong or right, come to the came conclusion. That is because the CRFs based on feature sets model will compare the info which been marked in detected sequences while under testing, there is some relativity between the two neighboring labels in one sequence, which come to the results as above.

From the average value of the three experiment 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, and this is the reason for U2R’s low recall ratio.

From the detection of U2R in each team and the comparison with other attack’s test results, we know that the recall 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.

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 the same. Further more, we have used CRFs detection model to recognize the concrete attack style of every conjunction record. But the results are a little worse. 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 with the results from multi-class SVM[16] and UVSVM[5]. Because of different data quantity, here we use test time of each record, the detailed as data is show in table-3.

Table-3: Detection speed and false alarm rate row forms of various calculate ways

	CRFsFS	UVSVM	Multi-class SVM
average detection time / record/ms	0.62	0.89	1.74
false alarm rate/%	0.03	0.78	1.89

CRFsFS represents the detective model of CRFs based on feature sets in each tables in this paper. From the table III, we could see the speed of CRFs detection is higher than multi-class SVM and UVSVM while the false alarm rate is lower. Let’s take a look at another table, showing the detection precision of different attacks using different detection methods.

Table-4: 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 methods, 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 4, we can see the precision rate 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 in table IV.

In order To compare with Kapil Kumar Gupta’s detection method[9], we use the same datasets as in Kapil Kumar Gupta’s thesis, and the same method of dividing the data sets. Then we take experiments respectively according to Kapil Kumar Gupta method and the CRFs based on feature sets’ detection method. The comparations of results are listed as the following table-5.

Table-5: The comparation between detection method based on feature set and Kapil Kumar Gupta method

			Accuracy /%	Precision Rate/%	Recall /%	F-Value /%	False Alarm Rate /%	Missing Alarm Rate /%
DoS		Kapil's Method	99.98	99.99	99.97	99.98	0.02	0.03
		CRFsFS	99.99	100.00	99.99	99.99	0.01	0.01
Probe		Kapil's Method	99.90	99.17	98.35	98.75	0.04	1.66
		CRFsFS	99.97	99.95	99.27	99.61	0.01	0.73
R2L		Kapil's Method	99.92	97.99	95.20	96.58	0.02	4.80
		CRFsFS	99.97	98.93	98.05	98.48	0.01	1.95
U2r	Large size	Kapil's Method	99.97	93.75	57.69	71.43	0.00	42.31
		CRFsFS	99.98	100.00	65.39	79.07	0.00	34.62
	Small size	Kapil's Method	99.08	96.00	92.31	94.12	0.33	7.69
		CRFsFS	98.47	100.00	80.77	89.36	0.00	19.23

In the experiment, results from Kapil Kumar Gupta detection method are not quite the same as the results in its thesis, and in Kapil Kumar Gupta thesis, they think the U2R detection experiments employing large or small scale of normal data mixed with U2R samples training detection model will receive the familiar results. But the experiment results we get have some kind of different. This may be because though we use the same data sets, but the final divided concrete training sets and testing sets may be different, what's more, the parameter used in order may be different in the process of training. In the comparison experiment, parameter C in training orders is 1.5. From table V, we can see that results by our detection method are better than those by Kapil Kumar Gupta method.

To distinguish the difference of results of CRFs detection model in these two experiment method, come out because of error or not, we carry out a test of the significance of difference. In the test, P-Value of DoS, Probe and R2L are far less than 0.01 by using sign test method. The data shows the different test results of the three attacks by two methods are not happened occasionally, and the difference is obvious. The test of the two methods results of U2R are respectively 0.625 and 0.581, which are far large than 0.05. This means the two methods in test U2R have no obvious difference. That is to say, when the abnormal behavior is small quantity sample, the sample quantity is very large, the difference will be obvious. And the capability of detection model based on features sets is much better than that of CRFs detection model in Kapil Kumar Gupta thesis.

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

- CRFs can consider the dependent relations between features and feature sets well, which is also advantage of CRFs itself.
- The detective data can be different structure of data set, and does not need to prepare for the processing, also does not need to put forward any assumption conditions.
- Though the training time of the model turn an index growth along with the increment of the sample quantity, once the training complete, the CRFs detection model acquired will swiftly and efficiently recognize normal and abnormal conjunction record.
- While in training if the sample category is enough, the amount of sample it needs is little.
- In test, 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.
- It is very efficiency to detect the huge data set.
- So we declare the superior detection function of CRFs detection model based on features sets and it is very suitable to apply in network intrusion detection.

CONCLUSIONS

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.

Our next work aim is to apply the CRFs detection model in opening experiment environment, and study how to better use CRFs in the identify work of unknown attack types.

REFERENCES

1. Wenke Lee, S J Stolfo, and K W Mok; A data mining framework for building intrusion detection

models. The 1999 IEEE Symposium on Security and Privacy, CA, Oakland, 1999; .120-132.

2. Yu Feng, Ma Xiaochun, Gao Xiang; Application of Frequent Episodes Mining on Intrusion Detection", Application Research of Computers, 2005; (7):153-155
3. Gary Stein, Bing Chen, Annie S Wu; Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection. Proceedings of the 43rd annual Southeast regional conference - Volume 2, ACM, Georgia, 2005; 136 -141
4. Xiangyang Li, Nong Ye; Decision tree classifiers for computer intrusion detection", Real-time system security, 2003; 77-93.
5. Luo Min, Yin Xiaoguang, and Zhang Huanguo; A Research on Intrusion Detection Based on Unsupervised Clustering and Support Vector Machines", Computer Engineering and Applications, 2006; (18):4-7.
6. Rao Xian, Dong Chunxi, and Yang Shaoquan; An Intrusion Detection System Based on Support Vector Machine. Journal of Software, 2003;14(4):798-803.
7. Jha S, Tan K, Maxion RA; Markov chains, classifiers and intrusion detection", The 14th IEEE Computer Security Foundations Workshop, IEEE Computer Society, Canada, Washington, 2001; 206.
8. Li Hui, Guan Xiaohong, and Zan Xin; Network Intrusion Detection Based on Support Vector Machine", Journal of Computer Research and Development, 2013; 40(6): 799-807.
9. Kapil Kumar Gupta, Baikunth Nath, and Kotagiri Ramamohanarao; Conditional Random Fields for Intrusion Detection. Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops, IEEE Computer Society, Melbourne, Australia, 2012; 203-208.
10. Lafferty J, McCallum A, Pereira F; Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, 2001; 282-289.
11. Burr Settles; Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, 2010; 104-107.
12. Fei Sha, Fernando Pereira, "Shallow Parsing with Conditional Random Fields", Proceedings of HLT-NAACL 2003, Association for Computational Linguistics, Edmonton, Canada, 2013, pp. 134-141.
13. Andrew McCallum, Wei Li; Early Results for Named Entity Recognition with Conditional Random Fields Feature Induction and Web-Enhanced Lexicons. Proceedings of the 7th Conference on Natural Language Learning, Association for Computational Linguistics, Edmonton, Canada, 2003; 188-191.
14. Hong MingCai, Zhang Kuo, and Tang Jie; A Chinese Part-of-speech Tagging Approach Using Conditional Random Fields", Computer Science, 2012; 133(10):148-158.
15. Kdd cup 1999 intrusion detection data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm
16. Xiao Yun, Han Chongzhao, and Zheng qinghua; Network Intrusion Detection Method Based on Multi-Class Support Vector Machine", Journal of Xi'an Jiaotong University, 2013;39(6): 562-565.