SVM-based Dynamic Risk Recognition and Complex Risk Assessment

Xue Liu[1, a], Wenjing Qi[2,b*] and Weihua Yuan[2,c]

[1] Dept. of Computer Science, Shandong College of Electronic Technology

[2] School of Computer Science and Technology, Shandong Jianzhu University

[a] liuxue71@126.com, [b] qiwj@sdjzu.edu.cn, [c] huahua_qingdao@126.com

Keywords: Dynamic risk recognition; SVM; Risk factors; Complex risk; Quantified assessment

Abstract. Risk assessment is a critical step for the robust operation of an information system. We incorporate machine learning and statistical theory together in risk recognition and evaluation to accommodate the dynamic and complex characters of information systems. first, SVM classifier is employed to recognize dynamic risk; then risk factor is defined for very single risk based on historical experiences; further, a complex risk assessment model is proposed to quantify risk to capital loss, which provide an intuitive way for user to understand the severity of risks . Experiments show that our method is feasible and effective in practical application environments.

Introduction

Nowadays, information systems (IS) are increasingly becoming basic infrastructures of our society and economy. But there exist many potential risks during the operation of information systems due to their intrinsic flaws and outside threats. Risk assessment is the process that makes use risk management theories and scientific method to analyze vulnerabilities and confronted threats of IS systematically, evaluates the hazard degree and provides countermeasures to solve potential attacks or to control risk to an acceptable level.

It is believed that properly used of risk quantification tools will help to reduce the cost on maintenance of system security and the chances of being intruded , the benefits there out will be far more than investments[1,2]. The existing risk assessment approaches can be grouped into two major categories: qualitative approaches [3-5] and quantitative approaches [8-10]. OCTAVE approach [3], PARA and Facilitated Risk Analysis Process[4] are qualitative methods. VaR-based risk assessment[6], Markov Model based security risks evaluation[7] and risk evaluation model based on correlation rules[8], modular attack trees[10] are representatives of quantitative approaches.

Most methods for IS security risk assessment are relatively static or one-off method, which means it can’t adapt to highly dynamic security risk, and lead to the lagging assessment results with little use for current situation. In this paper, we proposed a method to recognize and evaluate IS security risk dynamically based on online SVM classifier and historical statistic. Firstly, SVM classifier is used to recognize present risk; risk factors are obtained from historical statistic data; then, a complex risk assessment model (CRAM) is proposed to quantitatively evaluated capital loss arose from various risks. Finally, we apply this method in a real application environment to test its feasibility and performance.

Recognition of Dynamic Risk Based on SVM

Information security threats change greatly with time, this means data we will deal with are complex and in high volume; at same time, few samples of attacks can be obtained if we want to quickly response to prevent further loss. These problems bring great difficulties in recognizing threats effectively. SVM is a classifier specially suit for small sample data set, and is not sensitive to data dimension, it becomes a widely used method in network abnormal events detection [11] and other pattern recognition problems.

Theorem of SVM. SVM is developed from the optimal classification hyper-plane of linear separable condition. For a linear separable sample set $(x_i, y_i), i=1,...,n, x_i \in R^d, y_i \in \{1,-1\}$, it satisfies:

$$y_i(w \cdot x_i + b) - 1 \geq 0,\ i=1,...,n \tag{1}$$

The plane that satisfies Eq.1 and minimize class interval is called optimal classification hyper-plane. This is a quadratic programming problem; its optimal solution is the saddle point of Lagrange function:

$$L(w,b,a)=\tfrac{1}{2}\|w\|^2-\sum_{i=1}^{l}a_l\{y_i[(w\cdot x_i)+b]-1\} \tag{2}$$

Where α is non-negative Lagrange multiplier. Optimal classification plane problem is then converted to the following optimizing problem:

$$\min\ L(w,b,a)=\tfrac{1}{2}\|w\|^2-\sum_{i=1}^{l}a_l\{y_i[(w\cdot x_i)+b]-1\} \tag{3}$$

$$subject\ to\ \ y_i[(w\cdot x_i)+b]-1\geq 0\ \ \forall i$$

Eq.3 can be further converted to dual problem with Lagrange optimization method, that is, to optimize a_i in order to maximize function Eq.4 under constraint $\sum_{i=1}^{l}y_i a_l=0,\ ai\geq 0,\ i=1,...,n$

$$Q(a)=\sum_{i=1}^{n}a_i-\tfrac{1}{2}\sum_{i,j=1}^{n}a_i a_j y_i y_j(x_i\cdot x_j) \tag{4}$$

There exists a unique solution to this quadratic function optimizing problem, and only a small part of the solution with nonzero a_i^*, these samples that corresponding to a_i^* are support vectors, then we get optimal classification function $f(x)$:

$$f(x)=\mathrm{sgn}\{(w\cdot x+b)\}=\mathrm{sgn}\{\sum_{i=1}^{l}a_i^* y_i\ (x_i\cdot x)+b^*\} \tag{5}$$

Index for Evaluating Performance. Several indices are used to evaluate performance of classifier. Let *tp* denote number of normal samples that correctly classified, *tn* is number of abnormal samples that correctly classified, *fp* is number of normal samples that wrongly classified, and *fn* is number of abnormal samples that wrongly classified. Then accuracy of classification is defined as:

$$Accuracy=\frac{tp+tn}{tp+tn+fp+fn} \tag{6}$$

In security risk recognition, we concern more about false negative rate (FNR) and false positive rate (FPR), FPR and FNR represent wrongly classified rate in normal sample set and abnormal sample set:

$$FPR=\frac{fp}{tp+fp}\quad ,FNR=\frac{fn}{fn+tn} \tag{7}$$

In intrusion detection, FNR is more critical index, since we don't want to leave out any suspicious events.

Experiments. We performed an intrusion detection experiment in a company LAN. We collected original data flows of three weeks, randomly select a part of these data, analyzed it manually and annotated as two class: *normal* and *abnormal*. Then one-third of it is taken as training data, and the rest is testing data. Data is normalized to the format that can be used by SVM classifier[12] as follows: (*0,tcp,http,SF,224,1658,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,18,18,0.00,0.00,0.00,0.00,1.00,0.00,0.00,91, 255,1.00,0.00,0.01,0.04,0.00,0.00,0.00,0.00,normal*). Compared to the ground truth, the recognition accuracy rate is 94.6%, FPR is 5.46% and FNR is 4.96%, which are in an acceptable range, and we find that most of FNR are caused by unknown attacks. Also, we can train the classifier to recognize specific class of threats.

Complex Risk Assessment Model

In any conditions, risk is the total effects of some factors, such as threats, vulnerabilities, asset value and non-asset value (e.g. influence on productivity or social benefits). In an information system, security risks that arousing from diverse threats and exploiting various system flaws are complex, we proposed a complex risk assessment model (CRAM) in this section to evaluate total hazard degree.

Description of Threat. We defined some symbols as follows to describe the risk factors and model.

Types of threats: Risk scale of a threat is related to its type, let *E* denote the set of all possible threats, $E=\{e_i|i=1,2,....n\}$, and threats that an organization really confronted is a subset of *E*.

Possibility of occurrence of a threat: let $p(e_i)$ dente the occurrence possibility of threat e_i under certain confidence degree. We evaluate the possibility of an organization using historical data and prediction technologies.

Number of observed threats: How many threats are recorded and recognized in a given time interval, denote by A.

Number of threats of type e_i: denoted by N_i , $N_i=p(e_i)*A$.

Representation of Vunerability. Vulnerabilities are channels that used to perform attacks. Vulnerability set is denoted as: $V=\{v_j| j=1,2,...,m\}$. We defined a variable f_{ji}, which reflects the probability that vulnerability v_j can be employed by threat e_i, value of f_{ji} is obtained from historical data by experts. A relation set is defined with respect to the relations between vulnerabilities and threats, $R=\{R_{ij}| i=1,2,....., j=1,2,...n\}$, where $R_{ij}=(v_j,e_i, f_{ji})$, it means that threat of type e_i attack an object by employing the vulnerability v_j at a possibility f_{ji} .

Quantified Assessment of Complex Risks. Consequence of risk is loss of capital or assets, that is, if potential threats turned into actual attacks, they will affect information resources, productivity and cooperation image, and this can be valued by capital. So it is a comprehensible way to represent risk with capital lo**ss.** Let A denotes total events per day, $C(e_i)$ is the value loss that threat of type e_i will induce, it is the function of threat influence factor(w_i) and capital value factor(l_i), $C(e_i)=l_i*w_i$, l_i and w_i are experiential value represents average loss value and its influence weight of event e_i. Then we can evaluate the daily capital loss L with:

$$L=A\times\sum_{R_{ij}\in R}P(e_i)\times f_{ji}\times C(e_i) \qquad (8)$$

From Eq. 8 we can see that L reflects complex risk generated by various threats. What we expect is to control the average daily capital loss under an acceptable level, this requires proper security investment to compensate system weakness. Value of L is a great support for managers to decide the proper investment amounts.

Practical Results and Evaluation

Training and Classification. We apply our risk recognition and assessment method in a government's portal website, which provides online services for the public, such as news broadcasting, online consulting and online approving etc. We recorded data traffics of 3 months, there are 7545228 access records. 10% of data for a week are taken as training samples and analyzed manually; there are 51045 records of normal events and 2530 records of abnormal events. From analysis, four types of abnormal events are recognized; the occurrence distribution of abnormal events is shown in Table 1.

Table 1. Distribution of Threats in Training data

	Volume	$p(e_i)$	Description
1	1091	43.12%	Maintain error
2	453	17.90%	Unauthorized Access
3	557	22.02%	Worms
4	429	16.96%	Trojan horse

Table 2. Recognition of Abnormal Events

Tpye	Threats Volume	$p(e_i)$	Threats per day
e_1	143629	41.01%	1561
e_2	80251	22.91%	872
e_3	63518	18.13%	690
e_4	53654	15.31%	583
e_{un}	9204	2.63%	100

SVM classifier is used to distinguish abnormal and normal events in the rest data, and recognized 350256 abnormal events. Then we recognize these events further based on classes of training data. The results are shown in Table.2, there are 2.63% events that can't be classified to any type of e_1,e_2,e_3,e_4, we marked them as unknown events, e_{un}.

Assessments and Quantification of Risk. According to the definition of capital loss in Eq. 8, A is total events per day; $p(e_i)*$ f_{ji} reflects possibility that actually happened, in our experiments, we calculate the happened threats, so f_{ji} is assigned value 1. Threat influence factor(w_i) and capital value factor(l_i) are given by experts, without loss of generality, the risk level of unknown events is considered to be media, then threat influence factor is set to 0.5，we summarize evaluation parameters in Table 3.

With parameters in Table 3, the daily capital loss is calculated with Eq. 8, Fig.1 shows daily loss caused by security events in 3 months. The average capital loss in a single day is $\widetilde{L}=(9.9\pm3.437)\times10^4$.

Table 3. Parameters of Threats and Influence

Type	$p(e_i)$	w_i	l_i	$C(e_i)$
e_1	41.01%	0.3	10000	3000
e_2	22.91%	0.7	100000	70000
e_3	18.13%	0.5	30000	15000
e_4	15.31%	0.7	50000	35000
e_{un}	2.63%	0.5	50000	25000

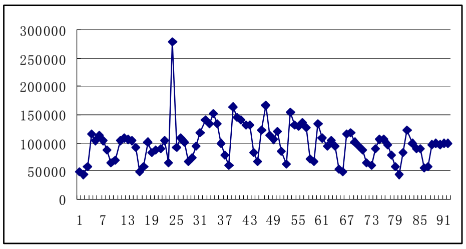

Figure 1. Daily Loss

Conclusions

We proposed a systematic method to recognize and quantify complex risk dynamically. SVM classifier, which always uses latest threat samples to train classifier, is good at recognizing dynamic risk in information system. Complex risk assessment model turns risks into capital loss, user or decision maker can get an intuitive and objective view of the risk level that organization currently confronted without disturbed by complicated techniques and managerial details. In future, we'll further our work in improving accuracy and optimizing quantification parameters.

Acknowledgement

The author would like to thank supports from Shandong International Cooperation Training program for Young Teachers and Shandong NSF Grant (ZR2013FL024).

References

[1] Mercuri, R. T.. Analyzing security costs[J]. Comm. ACM 2003,46(6):15–18.

[2] Longstaff, T. A., C. Chittister, R. Pethia, Y. Y. Haimes. Are we forgetting the risks of information technology?[J]. IEEE Compute, 2000,33(12) 43-51.

[3] C. Alberts, A. Dorofee, Managing Information Security Risks: The OCTAVE Approach, Pearson Education, Inc., Upper Saddle River, New Jersey, 2002.

[4] T. Peltier, Information Security Risk Analysis, second ed., Auerbach Publications, Boca Raton, FL, 2007.

[5] Peide Liu, Zhengwei Du. Application of E-commerce Risk Assessment Research with Weight Unknown TOPSIS Method[C]. International Symposiums on Information Processing, 2008,345-349

[6] Wenjing Qi, Xue Liu, Jian Zhang, Weihua Yuan.Dynamic Assessment and VaR-based Quantification of Information Security Risk[C], EBISS2010，

[7] Weiming L, Zhengbiao G． Hidden Markov Model Based Real Time Netw ork Security Quantification Method [C]． 2009 International Conference on Networks Security， Wireless Communications and Trusted Computing，2009

[8] Xiaorong Cheng, Yan Wei, Xin Geng, Network Security Risk Assessment Based on Association Rules ， Proceedings of 2009 4th International Conference on Computer Science & Education,1142-1145

[9] L.A. Gordon,M.P. Loeb, The economics of information security investment, ACMTransactions on Information and SystemSecurity, 5 (4) (2002) :438–457.

[10] L. Grunske, D. Joyce, Quantitative risk-based security prediction for component-based systems with explicitly modeled attack profiles, Journal of Systems and Software 81 (8) (2008) 1327–1345.

[11] Snehal A. Mulay, P.R. Devale, G.V. Garje. Intrusion Detection System using Support Vector Machine and Decision Tree. International Journal of Computer Applications .Volume 3 – No.3, June 2010

[12] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html