International Journal of Applied Operational Research
Vol. 1, No. 2, pp. 31-35, Autumn 2011

Journal homepage: www.ijorlu.ir

Network Risk Evaluation by Data Mining

N. Najafli*, B. Farnoushfar, M. Nasiri, B. Minaei

Received: April 8, 2011 ; **Accepted:** August 20, 2011

Abstract Risk management is one of the most prominent concepts which has recently been brought into sharp focus regarding security issues in computer networks. Scientifically speaking, risk in the field of network security is a generalized matter leading the organization to the provision of resolutions which target resources and profits of the organization. This paper has discussed what methods are to be adopted with respect to identification and classification of risk assessment and evaluation of computer networks using data mining. The proposed method uses support vector machine to generate and analyze results of risk evaluation. Considering available resources in the organization, the tasks performed by means of this method are of high accuracy; therefore, results are hereby in accordance with theories of management and security and based on resources and organizational profits depending on adjustable parameters as inputs. Outcomes of this study can be implemented in a dynamically real-time defensive system.

Keywords Risk Management, Risk Evaluation Model, Network Vulnerability, Data Mining, Support Vector machine.

1 Introduction

When it comes to network risk evaluation by data mining, regardless of any specific infrastructure, not only must it be independent of the related executive environment, but also it includes vital information of the organization, prioritizes security cases and considers management and financial costs. To achieve this goal, the so-called “risk” takes the place of vulnerability-related issues, by which it means we have come up with features converting it to a discernable and measurable data structure in computer programs. Also, we have used Support Vector Machine to verify accuracy and evaluate the model, so results will appear as a model of highest accuracy, precision, and efficiency.

2 Related works

Lee Kong and his colleagues [1] in a paper on *application of RBF-SVM in network security risk* have compared SVM superiority to other methods and in another paper on *combined kernel SVM in network security risks,* they have compared combined kernel performance with the others. In this paper, we have used SVM as a tool to assess and evaluate a risk data structure in network security. However, in our research, polynomial kernel is better than RBF;

* **Corresponding author. (✉)**
E-mail: n.najafli@yahoo.co.uk (N. Najafli)

N. Najafli, B. Farnoushfar, M. Nasiri, B. Minaei
Faculty of Computer Engineering, Iran University of Science and Technology.

due to the risk data structure which was used. It is a better approach in terms of performance and efficiency rather than RBF and combined kernel.

2.1 Risk and vulnerability

Vulnerability is a weakness of a system. Any vulnerability is considered a risk which can be exploited. Risk existence conditions are: 1- The value of organizational assets leading to identifying critical sections. 2- Threats that target these assets. 3- Sorts of Vulnerability influencing these assets. The figure1 indicates the risk triangle with its components.

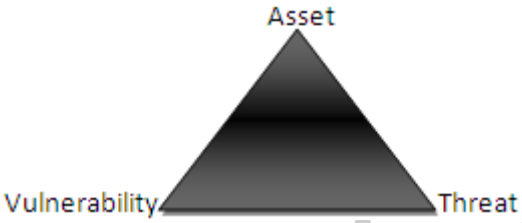

Fig. 1 Risk triangle

2.2 Risk features

The risk features which were investigated include: Window of vulnerability, CIA components, Stolen organizational assets by unauthorized persons, Loss of organizational assets, Manipulation of organizational assets, minimizing or eliminating risks, Risks related to operating system, Network complexity increase, Risk quality, Risk amount based on annual loss, Risk similarity to known risks, Vulnerability weight, approaches to challenging or minimizing risk, Risk name and Host name in which the risk is detected.

2.3 Proposed method

Data mining has many tools for analyzing data to discover patterns in mass data. Also, a lot of data records exist, so we applied data mining to evaluate network risks and discover any new relation.

To come up with a model and prepare an experimental model, we have used a combination of the above risk attributes to classify records. Then, we have examined different kernels to find a kernel with the highest accuracy and the best performance because it must be useable in real-time systems too.

2.4 Preprocess

In this section, we have chosen the Feature-Selection algorithm. The feature-Selection algorithm can optimize the data models which we want to use in SVM; by using this algorithm, some attributes of data structure having low impact are detected and ignored, and thus consequently, the curse of dimensionality problem is reduced extremely and the model accuracy increases impressively. Table1 is showing the settings we adjust for this algorithm.

Table 1 Feature selection parameters

Acceptable miss value	70%
Maximum records in a category	95%
Importance	x<0.8
Marginal	0.5<x<=0.8
Ignore	X<0.5
Coefficient of variance	0.1

Among the sixteen selected fields, according to collected data, three fields have been ignored by the algorithm and others with importance above 0.95 have been chosen.

3 Support vector machine

Support Vector Machine is a kind of statically learning machine which is used to label data points. This machine makes use of the classification and regression methods. Outcomes will be classified in two classes as positive and negative. This machine uses hyper planes for classification and if this couldn't classify, it would use kernel functions to map data points to a higher dimensional space. Stopping condition are determined by a parameter which is called "C" that is an upper bound for Lagrange coefficients in solving quadratic problems for distance of two margins of the hyper plane. This machine uses an upper bound to limit and improve leaning capacity of the learning machine. This inequality is:

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h(\log(\frac{2l}{h}) + 1 - \log(\frac{n}{4}))}{l}} \quad (1)$$

In inequality 1, we call R(α) as calculated risk and $R_{emp}(\alpha)$ as empirical risk that came from training sets and the second part of the right side of this inequality, is called the VC confidence. By this inequality for error measurement, we don't have to verify the training set anymore and it will generalize the error measurement. Also, the number of configuration parameters are lesser than the other algorithms and solving non-linear problems as linear and this, and this makes this method better to be used for generalization.
In order to map to a higher dimensional space, it uses these kernels:

- Polynomial $K(x, y)=(x.y+\delta)^p$ (2)
- Radial basis(Gaussian) $AFc = K(x, y)=e^{\frac{-\|x-y\|^2}{2\delta^2}}$ (3)
- Sigmoid $\tan(kx.y-\delta)$ (4)

The polynomial kernel with degree of zero becomes a linear kernel. Also, if gamma becomes zero, it will be inhomogeneous.

4 Results

The mentioned attributes have checked on hosts in a video game company. The hosts in this company were an online game server; several support hosts and a web server. They have been examined by a program which had been created by us. We have gathered 3092 records included mentioned attributes from them. Then, we examined them on four kernels of SVM. Table 2, is showing the kernels configurations.

Table 2 Kernels configuration

RBF Gamma	Bias	Stopping Condition	Regression Precision	Degree	Gamma	C	Kernel
-	0	0.001	0.1	-	1	10	Sigmoid
-	-	0.001	0.1	-	-	10	Linear
0.1	-	0.001	0.1	-	-	10	RBF
-	0	0.001	0.1	3	1	10	Polynomial

The worst result has been made by the sigmoid kernel but the other kernels have better outcomes. The reason of sigmoid weakness is that it is applied to mercer conditions for specified values for δ and κ. The results of each kernel are shown in table 3:

Table 3 The results for each kernel

F-measure	Training time per ms	accuracy	Kernel
0.09	3000	90.9	Sigmoid
0.953	2000	99.61	RBF
0.953	1000	99.61	Linear
0.979	1000	99.84	Polynomial

Because our risk data structure uses many flag fields, it was possible to predict that a polynomial or linear kernel would be appropriate. As shown in the results, that's why the polynomial kernel is much better for our data mining model. Now, we have a kernel which has the highest accuracy with the most correct result and the lowest training time.

The F-measure parameter is used to evaluate the kernel exact accuracy. It is a measure that combines precision and recall. Greater precision decreases recall and greater recall leads to decreased precision. The F-measure is the harmonic-mean of *P* and *R* and takes account of both measures.

$$F = \frac{2t_p}{t_p + f_n + f_p} \tag{5}$$

This equation shows how F-measure can be calculated. Each parameter is: t_p is true positive, f_n is false negative and f_p is false positive. They have calculated from table 4 which is called coincidence matrix:

Table 4 Coincidence matrix for polynomial

	Low	Moderate	High
Low	121		5
Moderate	0	155	0
High	0	0	2811

5 Conclusion

In this paper, we have studied a method to evaluate and manage the risk concept in computer networks by data mining. Also, we have introduced appropriate tools to evaluate risk with a suitable kernel. In our research, polynomial kernel according to this kind of risk data structure is more efficient than RBF kernel. Because of the method that we have used to describe risk, not only has it shown its capacity to consider organizational requirements, but also it is a generalized method that is independent of platform and environment. This method benefits from polynomial kernel which has low time and computational complexity that is useful and more efficient for real-time processing. Also, the described method uses data mining techniques to improve accuracy and efficiency of the computer networks security risk evaluations. This method can be used for designing a dynamically real-time defense in depth system to create intelligent and efficient security systems.

References

1. Guo, A. L., Li, C. C., (2008). International Symposium on Intelligent Information Technology Application Workshops. Shanghai 21-22 Dec, 40-43, china.
2. CMU/SEI-2006-HB-003, (2006). Defense-in-Depth: Foundations for Secure Resilient IT Enterprises. Carnegie mellon university and software engineering institute. Pittsburgh,Pennsylvania, USA.
3. Chen, W. H., Hsu, S. H., (2005). Application of SVM and ANN for intrusion detection. Computers & Operations Research, 32, 2617–2634.
4. Li, C. C., Guo, A. L., Li, D., (2008). International Symposium on Intelligent Information Technology Application Workshops, Shanghai 21-22 Dec, 36 – 39, china.
5. Wuling, R., Xianjie, W., (2008). International Conference on Advanced Computer Theory and Engineering, Phuket, 20-22 Dec, 378 – 382, china.
6. Hui-sheng, G., (2008). Wireless Communications,Nwtworking and Mobile Computing, Dalian, 12-14 Oct, 1-4 , Liaoning china.
7. Li, Z., (2009). Research of Information Security Risk. Management Based on Statistical Learning Theory, 3,436-438.
8. Jiawei, H., (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann publishers, USA.
9. Feature Selection, http://en.wikipedia.org/wiki/Feature_selection,10 june 2010
10. Data mining and artificial intelligence, http://www. irandataminer.ir, 17 June 2010.