Network security analysis of weighted neural network with association rules mining

WANG Ziqiao, FU Weinan

School of Software and Microelectronics, Northwestern Polytechnical University, Shaan Xi'an, 710072, China

Email: awe1233430@hotmail.com

Key words: Intrusion classification; mark; security; network; neural network

Abstract. This article applies Co-S3OM semi-supervised learning algorithm to intrusion detection field and proposes specific semi-supervised network intrusion classification scheme. In accordance with different type of attack, different mark samples are selected as training set to complete initialization of three S3OM classifiers; marked sample data is expanded with coordinative vote by three classifiers. Test structure process is given in detail to use KDD Cup 99 data set to perform semi-supervised classification. It shows in test that intrusion classification model based on Co-S3OM is of high data sample marking rate and high intrusion classification rate.

Introduction

As a dynamic network security technology, intrusion detection is able to comprehensively supervise computer network or various application programs in host computer, audit and analyze massive data on the basis of intelligent security strategy, actively identify and respond massive intrusion actions in system, effectively guarantee security of system and have become a hot research point in network security field[1]. Earlier intrusion detection algorithm is based on supervised learning, namely, requiring all training data samples to be marked by classification. This supervised learning method is of high detection rate but unable to detect unknown attack effectively; and data collected in training is required to be correctly marked to be normal or abnormal. However, in real network setting, there is missive unmarked data. There is great difficulty and high price in obtaining marked data. It is nearly impossible to mark all data.

Network intrusion classification

Network intrusion detection is a matter of classification in fact, namely, classifying detection data into normal and abnormal data. However, data required to be classified in intrusion detection is complex and generally appears to be high-dimensional and unclassified. Intrusion detection classification issue can be solved by applying SOM to intrusion detection field.

This article will coordinate application of semi-supervised classification algorithm Co-S3OM to network intrusion detection field and solve lesser network intrusion detection issues in marked sample. The main work of solving intrusion detection issues with Co-S3OM includes: The first one, selecting valid marked sample to complete initialization of three SSOM classifiers training set and ensuring otherness among three classifiers; the second one, fulfilling marking of the final classifier by coordination among three SSOM classifiers; the third one, expanding training set of three classifiers and adding newly marked sample to three classifiers with repeated iteration to form the

final classifier for classification.

(1) Data normalization

KDD Cup 99 data set is selected to perform intrusion detection classification test; tested data is classified into training set and testing set. Training set is divided into marked set and unmarked set. To make it convenient for statistical classification accuracy rate, marked sample is used for all testing sets. In samples of unmarked set, each data has 41 different attributes (32 continuous attributes and 9 discrete attributes); each data in marked sample has a total of 41 different attributes and 1 attack-type label.

Input singular values and corresponding type label of each data are withdrawn respectively from training set and testing set; discrete attributes are numbered in test to be numerical. For example, attack type is divided into normal data and attack data, which is numbered with 1 and 2; TCP, UDP, ICMP and other attributes of protocols are numbered with 1, 2, 3 and etc; service types, aol, auth, …, and whois, are numbered with 1, 2,…,68; name of types in Flag, OTH, REJ, …, SH, are numbered with 1, 2, …, 11.

To eliminate difference of order of magnitudes among each dimensional data, large error in network prediction is avoided to occur due to large difference between input and output. Data normalization function shown in format (1) is used to normalize input singular values to [0,1]. At this point, to ensure higher testing rate, training set and training set are put together as an overall data set for normalization.

$$x_k = (x_k - x_{\min}) / (x_k - x_{\max}) \tag{1}$$

Where $x_{\min}$ represents minimum value of data and $x_{\max}$ represents maximum value of data

(2) Initialization

During initialization, the most important work is how to select valid marked sample set from existing intrusion data set. Selected marked samples are divided into three levels as initialization training set of three SSOM classifiers respectively. To ensure three SSOM classifiers vote unmarked sample with high accuracy rate, samples of three SSOM classifiers training set shall be of high otherness, diversity and dissimilarity.

In network intrusion detection, due to numerous attack types, one type of attack data is selected from numerous types of attack data to fulfill initialization for one SSOM classifier. For example, DOS attack data fulfills initialization for SSOM classifier 1; R2L attack for SSOM classifier 2; probe for SSOM classifier 3. In this way, three SSOM classifiers training set is inconsistent totally with large otherness and network model generated training SSOM neural network is also inconsistent. Through voting by three classifiers, marking of unmarked sample is fulfilled with a higher marking accuracy rate.

(3) Data marking

Semi-supervised classification algorithm Co-S3OM is mainly used to fulfill marking of unmarked sample, explore implicit information about unmarked sample and expand number of marked sample in this process.

Through initialization, three SSOM classifiers have respective marked sample training sets T1, T2 and T3. Three training sets are used to train classifiers W1、W2 and W3 respectively. During coordination among three SSOM classifiers, when three SSOM classifiers predict certain unmarked sample uniformly, such predicted mark is used to mark predicted sample, add it to marked sample set and continuously expand number of marked sample to form new training set.

During testing in data set KDD Cup 99, since each sample has information about 41

characteristic attributes and the 42nd attribute is type of sample, if three SSOM predict certain unmarked sample to be normal, the 42nd attribute is marked as 1 directly; if predicting to be abnormal, the 42nd attribute is marked as 2 directly; if predication is inconsistent, the next sample is selected from unmarked set for predication and marking.

Newly added marked samples are added to one of T1, T2 and T3 successively each time to form new marked sample training set and ensure inconsistent training sample of three classifiers. After marked sample training set is renewed, T1, T2 and T3 is reused to train and generate new classifiers W1、W2 and W3; new classifiers are used for repeated iteration until unmarked sample set is empty.

(4) Intrusion classification

Three expanded T1, T2 and T3 are combined as the final intrusion detection training set to train SSOM neural network, generate final classifier and use such classifier to fulfill testing for data to be tested.

Test

PC of Intel Core2 Duo CPU 2.0GHz and 2.0GB internal storage is used and operation system Windows XP and programming setting MATLAB 7.8.0 (R2009.0a) fitted in test platform.

Data set KDD Cup 99 is used for test data, including massive normal network flows and various attacks, which can be divided into 4 types, namely DOS、R2L、U2R and PROBE.

(1) DOS: Refuse service attack, such as SYN Flood and land;

(2) R2L: Unauthorized remote access, such as password guess;

(3) U2R: Various privilege escalations, such as various local and remote Buffer Overflow attacks;

(4) PROBE: Various port scanning and vulnerability scanning

In test, 500 normal data and 500 attack type data are drawn randomly from data set “10% KDD” of KDD Cup 99 as unmarked training data set; 36 data is drawn randomly from data set “10% KDD” of KDD Cup 99 as marked training data set; 500 normal data and 500 attack type data are drawn randomly from data set “Corrected KDD” of KDD Cup 99 as tested data set.

Marked sample training sets are divided into three parts as initialization training set for three SSOM classifiers respectively. Three tests are performed as per selected number from less to more. In each test, when each classifier performs initialization, marked sample training set is composed as shown in Table 1. For example, in the 1st test, 5 marked samples are selected as training set to be tested for three classifiers. Training set of classifier 1 is composed of 2 normal samples and 3 DOS attack samples; training set of classifier 2 is composed of 2 normal samples, 2 R2L attack samples and 1 U2R attack sample; classifier 3 is composed of 2 normal samples and 3 Probe attack samples. Training set of each classifier comes from different types of attack. There is an obvious difference among three classifications.

Table 1 Tested Data Set (Unit: Nos)

		Normal	DOS	R2L	U2R	Probe
Test 1	Classifier 1	2	3			
	Classifier 2	2		2	1	
	Classifier 3	2				3
Test 2	Classifier 1	3	6			
	Classifier 2	3		3	3	
	Classifier 3	3				6
Test 3	Classifier 1	6	6			
	Classifier 2	6		3	3	
	Classifier 3	6				6

Table 2 Results of Test

	SSOM	Co-SSOM	
	rate2	rate1	rate2
Test 1	54.50%	67.83%	72.63%
Test 2	57.03%	77.45%	79.03%
Test 3	66.90%	82.50%	89.00%

In each test, accurate marking rate and accurate classification rate are calculated respectively in accordance with formats (2) and (3). Rate 1 means accurate marking rate of unmarked sample; rate 2 means accurate marking rate of testing set sample. Each test is performed three times respectively; average is valued as final results. Results of test are as shown in Table 2. Figure 1 is changing curves of rate 1 and rate 2 with continuous growth of training set data.

It can be seen from Table 2 and Figure 1 that: firstly, since Co-SSOM has fulfilled expansion of marked sample, compared with using SSOM of initialization marked sample training set only, its classification rate is greatly enhanced; secondly, with initialization training concentration and continuous increase of number of marked samples, both accurate marking rate and intrusion detection rate are increasingly enhanced. As seen, semi-supervised intrusion detection method based on Co-SSOM is able to enhance intrusion detection classification rate of network effectively.

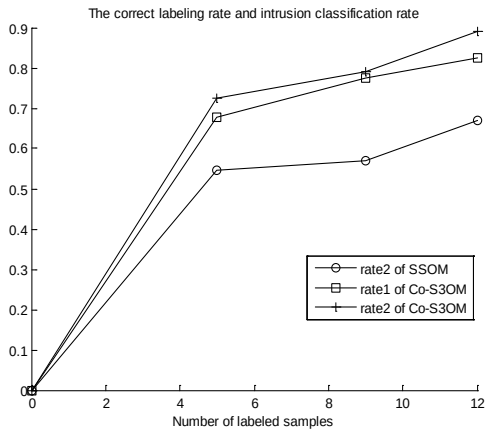

Figure 1 Comparison among Results of Test

Conclusions

To solve the issue of a high price in obtaining marked intrusion data in network setting, semi-supervised learning is introduced to network intrusion classification field. In accordance with different types of network types, slight marked intrusion data is divided into three parts as initial training set training classifiers respectively so as to form three initialized classifiers of large difference. Through coordinative learning of three classifiers, marking of unmarked intrusion data is fulfilled. A detailed introduction is given about the process of using KDD Cup 99 data set to construct semi-supervised tested data set. It shows in results of test that semi-supervised learning is able to explore unmarked intrusion data information effectively and it is of higher intrusion classification rate.

Reference

[1] Liu Y, Yang J, Meng Q, et al. Stereoscopic image quality assessment method based on binocular combination saliency model[J]. Signal Processing, 2016, 125: 237-248.

[2] Song, X., and Geng, Y.Distributed Community Detection Optimization Algorithm for Complex Networks. Journal of Networks, 9(10), 2758-2765.

[3] Pahlavan, K., Krishnamurthy, P., and Geng, Y. Localization Challenges for the Emergence of the Smart World, Access, IEEE, 3(1), 1-11.

[4] Ying Liang*, Xiukun Wang. (2013). Developing a new perspective to study the health of survivors of Sichuan earthquakes in China: a study on the effect of post-earthquake rescue policies on survivors' health-related quality of life, Health Research Policy and Systems, 11:41,1-12.