Protocol-Based Classification for Intrusion Detection

Kun-Ming Yu[1], Ming-Feng Wu[2], and Wai-Tak Wong[3]
[1] *Department of Computer Science and Information Engineering, Chung Hua University*
[2, 3]*Department of Information Management, Chung Hua University*
Chung Hua University, No. 707, Sec. 2,
WuFu Road, HsinChu, Taiwan.,
yu@pdlab.csie.chu.edu.tw, insperation410@gmail.com, wtwong316@gmail.com

Abstract: A lightweight network intrusion detection system is more efficient and effective for real world requirements. Higher performance may result if insignificant and/or useless features are eliminated. Logistic Regression is one feature selection method. In this study, protocol type and Logistic Regression were used to pick up the feature sets which can get nearly the same performance as the full feature using a Support Vector Machine. Evaluation was done over a benchmark dataset used KDD CUP'99. In terms of time efficiency, the proposed method performs more than seven times better than other feature selection methods.

Key-Words: Intrusion detection, Logistic Regression, Protocol, Support Vector Machine

1. Introduction

An Intrusion Detection System can be used to detect threatening breaches in information security. With the rapid growth in Internet business, malicious usage, attacks, stealing of sensitive information and sabotage, and information security have become prime concerns for many governments as well as corporations. Timely detection of perpetration from millions of connection records and getting an increased accuracy rate are important issues in information technology.

An Intrusion Detection System (IDS) is an important component of the defense-in-depth security mechanisms in computer network systems. At present, IDS analyzing packages and log files are used to prevent attacks. In general, two methods are used: network-based IDS and host-based IDS. Network-based IDS examines the content as well as format of network traffic. Therefore, a NIDS detects probes, scans, malicious and anomalous activity across the whole network. The primary advantage of a NIDS is that it can observe the whole network or any subsets of the network from one location. The other primary advantage is the low deployment cost. A NIDS is the only system that can monitor the entire domain, thus eliminating the need to install each host, which would lead to intrusion detection cost. However, a NIDS has several inherent weaknesses. These weaknesses are susceptibility to generating false alarms, as well as an inability to detect certain attacks called false negatives. A NIDS cannot detect host specific processes or protect from unauthorized physical access. On the other hand, host-based IDS places its reference monitor in the kernel/user layer and watches for anomalies in the system call and command sequences. Host-based IDS can analyze all activities belonging to the host. Unfortunately, the complicated log files decrease reaction time. However, if the log files are too sketchy, the Host-based IDS cannot effectively detect the invasion of normal activities. Generally speaking, most of these Host-based IDSs have common architectures, meaning that most host systems work as host agents reporting to a central console. Thus, the prime cost is considerable.

In order to make IDS more efficient, reducing the dimensions and data complexity have been used as simplifying features. In this study communication protocol was used as one of the primary conditions for making intrusion detection models. This was then combined with logistic regression for selection.

The remainder of this paper is organized as follows: In Section 2, the selection methods are discussed. The protocol-based intrusion detection

model is discussed in detail in section 3 and, in the final section; directions for future research are discussed.

2. Related Work

Feature selection can reduce both the data and the computational complexity. It can also get more efficient and find out the useful feature subsets. The feature selection methods used in our research are compared in this paper.

2.1 Principal Component Analysis

K. Person proposed Principal Component Analysis (PCA) [3], depending on the field of application. This is also known as the discrete Karhunen-Loève transform. PCA is based on transforming a large number of variables into a smaller number of uncorrelated variables by finding a few orthogonal linear combinations of the original variables with the largest variance. In [4], 14 features were chosen to predict the accuracy which was 99.8734％ of the KDD CUP'99 full data (kddcup.data.gz).

2.2 Discriminant Analysis

Discriminant analysis (DA) is used to determine discriminating variables between two or more naturally occurring groups. DA works by creating discriminant functions (DFs) which predict to which group each case belongs. DFs are interpreted by standardized coefficients and the structure matrix. DFs create the boundary between groups. Wong [5] used DA as the feature selection method and the false alarm rate was 0.37% in 9 selected features.

2.3 Logistic Regression Analysis

The Logistic Regression (LR) Model was proposed by the 19th century Belgian mathematician Verhulst. The main contribution of LR is that it solves the traditional linear regression models, the strain cannot be dealt with by the number of variables in the two categories of dependent variable error. There is a critical point increment (threshold) S-function through maximum probability estimates (Maximum Likelihood Estimation; MLE) predictors of the best estimate of parameters, which can deal with two kinds of nominal variables making forecast more accurate. In [11], LR was used as the IDS model feature selection method and the testing data were the KDDCUP'99 full data. The correct rate was 99.95%.

2.4 Support Vector Machine

The Support Vector Machine (SVM) is a relatively new statistical learning algorithm that provides a powerful tool for generalized classification with inadequate candidate ability. It was proposed by the Russian statistician and mathematician Vladimir Naumovich Vapnik. Ambwani [10] used SVM as classification tool. In the Ambwani [10] experiment, compared with the KDDCUP'99 test dataset, the accuracy rate was 92.46%. In [12] it was pointed out that KDDCUP'99 test dataset do not facilitate in forecasting results. In [4] KDDCUP'99 full dataset was used as test based on the proposed Ambwani theory, the prediction accuracy rate was 99.9382%.

Discriminant Analysis and Principal Component Analysis are clustering features as well as simplifying feature subsets to reduce the number of features replacing the original feature set, therefore the discrete eigenvalue can easily be removed or neglected. Logistic Regression reduces the data dimensional and calculating complexity by distinguishing protocol as a conditional for reducing the possibility of the discrete eigenvalue being ignored and enhancing the efficiency of Logistic Regression. Therefore, for this study, LR was chosen as the main method, and it was then compared with the methods for PCA [4], DA [5], LR [11].

3. Protocol-based Logistic Regression feature selection method

Protocol Anomaly Detection (PAD) works by analyzing application-level traffic, commands and behaviors and then blocking and denying

undesirable or otherwise inappropriate commands. Application protocols have been published in RFCs and vendor documents [1]. The application protocols can be used to check for proper or expected behavior, even in the absence of identification; new attacks can be effectively intercepted.

In [2], 90% of the attacks are protocol usage anomalies. The reason for that is most of the attacks exploit breaches in badly defined areas of protocols both in the protocol standard itself as well as its implementations. For example, CodeRed used buffer overflow to determine attacks. Thus, using communication protocol makes intrusion detection models more efficient.

In this paper, different communication protocols were used to sort different protocol-based IDS. There were five stages in our experiment. See the flow chart in Figure 1.

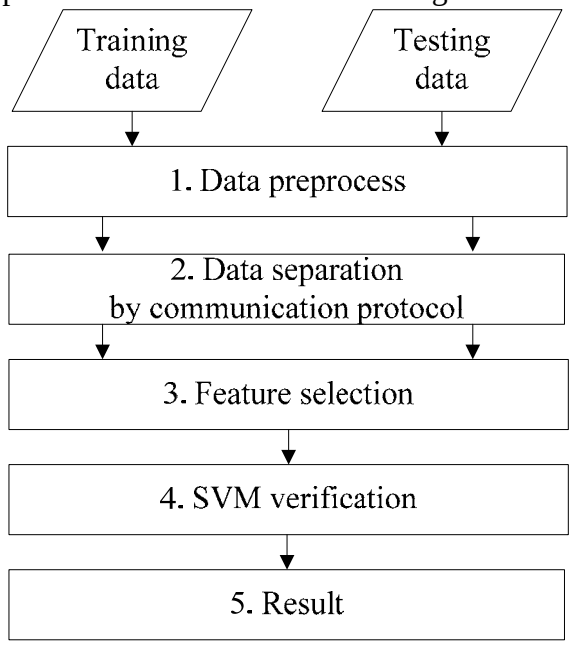

Figure 1. Algorithm flowchart

In the data preprocessing step, there were forty-one features in the KDDCUP'99 dataset in which Protocol type, Service, Flag and label were non-metric. Before SVM training and testing, non-metric data must be converted to metric data.

In Figure 1, separating data in the communication protocol step, the data were divided according to three protocols (ICMP, TCP, and UDP). Secondly, because of the likelihood of the KDDCUP'99 data numerical difference being misleading [13]. In order to increase the prediction accuracy rate and decrease the difference between data, data were normalized before the next step.

The feature selection step in figure.1, LR procedure of SPSS on the preprocessed data to obtain the feature subsets; details of the feature selection are given in Figure 2.

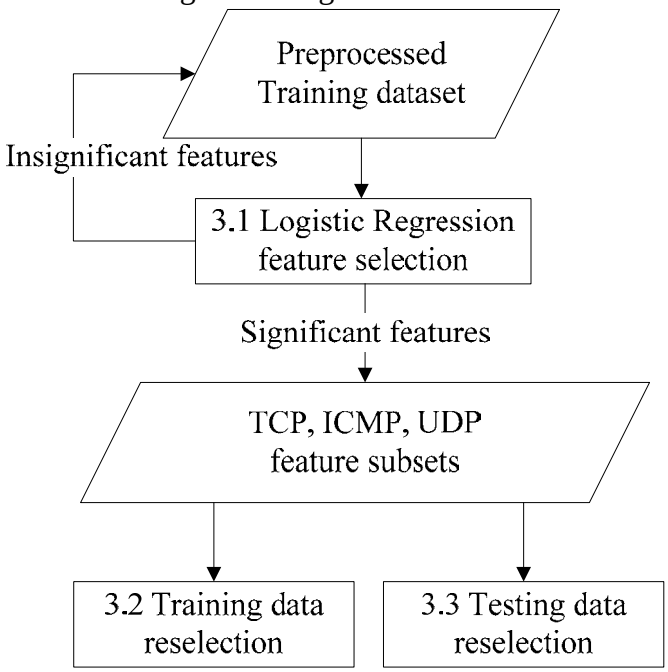

Figure 2. Feature selection step

For the fourth step in figure.1about SVM verification, we used LIBSM which is a kind of supervise SVM. SVM has well efficiency in classification, for instance, financial analysis, image distinguish, biological information analysis. SVM is the technique use to solve the classification problem which is how to classify the data of unknown data class into the correct data categories. If there are some data have been classified into specific categories, but don't know anything about the original rules used for classification, when new data comes, SVM can predict which category it should belong to by some statistical learning theories.

Due to the fact that problems in the real

world can be divided into more than two classes, the above may not be feasible because of overlapping distribution. Therefore, Corinna Cortes and Vapnik [6] propose the slack value ζ to handle misclassification and outlier data. Even though the prediction rate of SVM classification is good, there are two key subjects that influence prediction. One is the "kernel function selection"; the other is the "hyper-parameters search". Choosing a suitable kernel function and the best hyper-parameters are critical issues for SVMs. Unfortunately, up to now; the common way for solving these problems has been trial and error.

There are four kernel functions: linear, polynomial, radial basis function (RBF), and sigmoid. There is not one single standard for choosing a suitable kernel function, but according to Smola's research [7], the RBF function is a reasonable general first choice. Thus, RBF was used as raining reigning kernel function in this study.

Two parameters are necessary in the RBF kernel; C and γ must be searched. The goal is to identify good parameters (C, γ) so that the classifier can accurately predict unknown data. Presently, experts and scholars in an effort to solve the Support Vector Machine parameter selection some solutions [8] have been proposed. Chang, J. Lin [9] developed the LIBSVM which uses cross-validation parameters to achieve the best approach. In addition to the above methods can be selected SVM best parameters, academics Ambwani proposes other solutions [10].

LIBSVM RBF kernel function only provides two parameters, C and γ. The method selected for the first as a numerical value of the fixed value of γ (γ LIBSVM default value of 1 / k, k values for the input attributes number [11]), and t the parameters for C numerical interval were set as volatile Support Vector Machine module training and forecast information. The best C and γ from the test results were chosen.

4. Experiment

All experiments were performed on a Microsoft XP machine with a Pentium IV CPU 3.00GHz processor and 1 GB RAM. Kddcup.data_10_percent.gz with 494,021 records was used as the training dataset; KDDCUP'99 full dataset (kddcup.daata.gz) with 4,898,431 records for testing data. KDDCUP'99 originated from a study at the MIT Lincoln Lab and was post-processed by Columbia University. Involving four categories of attack: Dos（Denial of Service）, the feature list is shown in Table 1. In the past, KDDCUP'99 competition used the corrected.gz as test dataset, but according to [12] the huge data difference will lead to poor detection accuracy.

Table 1. KDDCUP '99 feature List

No	Feature name	No	Feature name
1	Duration	22	Is_guest_login
2	Protocol_type	23	Count
3	Service	24	Srv_count
4	Flag	25	Serror_rate
5	Src_bytes	26	Srv_serror_rate
6	Dst_bytes	27	Rerror_rate
7	Land	28	Srv_rerrot_rate
8	Wrong_flagment	29	Same_srv_rate
9	Urgent	30	Diff_srv_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_rate
14	Root_shell	35	Dst_host_diff_srv_rate
15	Su_attempted	36	Dst_host_same_src_port _rate
16	Num_root	37	Dst_host_srv_diff_host_R ate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_files	40	Dst_host_rerror_rate
20	Num_outbound_ cmds	41	Dst_host_srv_rerror_rate
21	Is_host_login		

This study characteristic according to the different communication protocols for data classification, using the logistic regression theory stepwise to simplify the eigenvalues. The protocols were divided into the following five parts: the pre-processing of data, distinguishing

data with communication protocols, feature selection using SPSS13.0 statistical software, and validating the SVM classification test and t According to the above, 20 features (deduct duplicate features in ICMP_LR, TCP_LR, UDP_LR) were chosen to set up models with different Communication protocols. The best feature subset shows in Table 2.

Table 2. Extraction of feature sets

Model name	Features used	Feature set
Full	41	1-41
DA	9	2, 12, 23, 24, 29, 31, 32, 36, 39
PCA	14	2, 3, 4, 23, 24, 25, 26, 29, 30, 33, 34, 36, 38, 39
Full_LR	15	2,4,6,8,10,12,22,23, 29,30,32,33,36,37,38
ICMP_LR	6	5, 24, 31, 32, 33, 37
TCP_LR	12	4, 13, 22, 24, 27, 28, 30, 32, 33, 34, 35, 36
UDP_LR	10	3, 5, 6, 8, 27, 29, 34, 35, 36, 40

The extraction feature sets of training data were fed to the SVM for training. The testing dataset was examined before the training process was finished. The two parameters of the Gaussian Radial basis function (RBF), C and γ, must be determined. The 10-fold Cross validation (CV) technique was used to train the dataset to find the parameters yielding the best results. The parameters tried in the 10-fold CV process wereγ ={2, 1, 0.5, 0.1, 0.01, 0.001} and C={1000, 750, 500, 250, 100, 50, 10, 2, 1}. The optimal parameters are shown in Table 3. The results were compared with the singular models of [4] and [11], and the results are shown in Figure 4. In Table 4 PL*R* is the total amount of ICMP_LR, TCP_LR and UDP_LR.

Table 3. The best parameters of LR method

Model name	C	γ
Full_LR	1000	0.5
ICMP_LR	750	2
TCP_LR	1000	1
UDP_LR	2	0.1

Table 4. Accuracy compared with other methods

Method Used	Features Used	Accuracy (%)
DA[5]	9	99.7305%
PCA[4]	14	99.8734%
Full	41	99.9382%
LR[11]	15	99.9587%
PLR	20	99.9634%

In Table 4, the performance of full features is 99.9382%, but using feature selection methods can get similar or better results than when using full features. This means that some features in KDDCUP '99 would have negative impact on accuracy. Besides, the performance of FP, FN, TP and TN were also compared in Table 5. Figure 5, giving the correct classification and misclassification rate, shows that our feature model performed better.

Table 5. Performance of different methods

%	TN	FP	FN	TP
DA[5]	99.94	0.06	0.32	99.68
PCA[4]	99.68	0.32	0.11	99.89
FULL	99.86	0.14	0.04	99.96
LR[11]	99.88	0.12	0.02	99.98
PLR	99.97	0.03	0.04	99.96

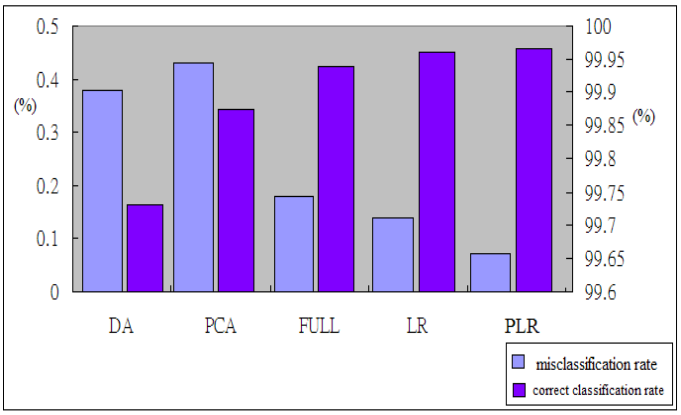

Figure 5. Performance of different methods

In [10], an accuracy of 99% was also achieved with a prediction time of 7.35 seconds, but the number sampled was only 6890. In this study, 4,898,431 (KDD CUP'99 full data) were used, which is 710 times larger than [10]. The experiment took 11 minutes 10 seconds. Multiplying this 710 times with the prediction time of [10] gives 89 minutes 4seconds. That means that our research saved 10 times on prediction time, thus being more efficient. The prediction time is shown in Table 6.

Table 6. Time efficiency of the different methods

Method Used	SVM predict time (hr：min：sec)
DA[5]	3:37:02
PCA[4]	1:53:22
FULL	1:35:05
LR[11]	0:12:51
PLR	0:11:10

5. Conclusion

We have presented a statistical method, Logistic Regression with separate protocol, has been presented for the selection of important features with different protocols for anomaly-based network intrusion detection systems using the KDDCUP'99 dataset. This approach is a theoretical method for finding features; it is fast and precise. Using the Support Vector Machine, the separate protocol model provided positive results.

Although no great improvement in detection accuracy was obtained, the elimination of features in this approach leads to a simplification of the problem. Faster and more accurate detection can be expected.

Future work will focus on adding new features to extract more suitable feature subsets or a lightweight Intrusion Detection System.

References:

[1] Freemont Avenue Software, 2004, " White paper Protocol Anomaly Detection," September 9.

[2] Erwan Lemonnier – Defcom., 2001, "Protocol Anomaly Detection in Network-based IDSs", June 28.

[3] Jolliffe, I. T., 2002, "Principle Component Analysis" Springer, 2nd ed., New York, USA.

[4] Lai, C. Y., 2007, "A Novel Approach to Multi-classifier Based on Multiple feature Sets with SVM for Network Intrusion Detection," *Chung Hua University* thesis, July 24.

[5] Wong, W. T., and Lai, Y. C., 2006, "Identifying Important Features For Intrusion Detection Using, Discriminant Analysis And Support Vector Machine," International Conference of Machine Learning and Cybernetics, Vol. 6, pp. 3563-3567.

[6] Cortes, C., and Vapnik, V., 1995, "Support-Vector networks," Machine learning, Vol. 20, pp. 273-297.

[7] Smola, A. J., 1998, " Learning with Kernels PhD Thesis", GMD, Birlinghoven, Germany.

[8] Grandvalet, Y., and Canu, S., 2002, "Adaptive scaling for feature selection in SVMs," Neural Information Processing System, Vol. 15, pp. 553-560.

[9] Hsu, C. W., Chang, C. C., and Lin, J. C., 2003, "LIBSVM:a library for support vector machines," Available http://ww.csie.ntu.edu.tw/~cjlin/libsv

[10] Ambwani, T., 2003, "Multi class support vector machine implementation to intrusion detection," Proceedings of the International Joint Conference of Neural Networks, vol. 3, pp. 2300-2305, January 27.

[11] Huang, W. C., 2007, "Using Regression Theiry fir Feature Selection in Intrusion Detection System," *Chung Hua University* thesis, August 10.

[12] Pavel, L., Patrick, D., Christin, S. and R. Konrad, 2005, "Learning intrusion detection: supervised or unsupervised?" 13th International Conference on Image Analysis and Processing, pp. 50-57.

[13] Stolfo, S. J., Wei, F., Wenke, L., Prodromidis, A., and Chan, P. K., 2000, "Cost-based modeling for fraud and intrusion detection: results from the JAM project,"

Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2, pp. 130-144.