Anomaly Detection Using LibSVM Training Tools

Jung-Chun Liu [1], Chu-Hsing Lin [1], Jui-Ling Yu [2], Wei-Shen Lai [3], Chia-Han Ho [1]
[1]*Department of Computer Science and Information Engineering, Tunghai University, Taiwan*
[2]*Department of Applied Mathematics, Providence University, Taiwan*
[3]*Department of Information Management, Chienkuo Technology University, Taiwan*
[1]*{ jcliu, chlin, g95280077 }@thu.edu.tw*
[2]*jlyu@pu.edu.tw*
[3]*WeiShenLai@gmail.com*

Abstract

Intrusion detection is the means to identify the intrusive behaviors and provide useful information to intruded systems to respond fast and to avoid or reduce damages. In recent years, learning machine technology is often used as a detection method in anomaly detection. In this research, we use support vector machine as a learning method for anomaly detection, and use LibSVM as the support vector machine tool. By using this tool, we get rid of numerous and complex operations and do not have to use external tools for finding parameters as needed by using other algorithms such as the genetic algorithm. Experimental results show that high average detection rates and low average false positive rates in anomaly detection are achieved by our proposed approach.

1. Introduction

Intrusion detection system (IDS) forms the second line of defense, and the intrusion detection technology has become critical to protect systems and users in the Internet age [1]. Intrusion detection is the means to identify and indicate the intrusive behaviors. Information of users is monitored and collected, and is analyzed to find the users' patterns of behavior. The gathered information is compared with known data to detect invasions, attacks, and abnormal activities. Upon detection of intrusions, intruded systems respond to avoid or reduce further damages.

There are mainly two types of intrusion detection techniques: anomaly detection and misuse detection. We will focus on learning-based anomaly detection in this paper.

Anomaly detection uses statistical analysis methods to analyze normal users' behaviors on the Internet plus internal information flow statistics and records to build a profile. Then, this profile is used as a benchmark to classify activities of system operations. Abnormal activities are detected when events occur outside the scope of normal activities. The advantage of anomaly detection is that one needs not to worry about various possible attacks until the first occurrence of abnormal behaviors is recorded.

For anomaly detection, we train data by support vector machine (SVM) [2]. There are many researches with good results about learning-based IDS with SVM [3, 4] and anomaly detection [5, 6]. To improve the efficiency for anomaly detection, some researchers propose to combine SVM with other technologies, for example, neural networks [7, 8, 9], and genetic algorithm [10, 11, 12]. We study the feasibility of using LibSVM for anomaly detection in this research.

SVM is a statistical learning theory based on machine learning methods. A special property of SVM is that it simultaneously minimizes empirical classification errors and maximizes geometric margins. By training with lots of data, SVM learns to find the best compromise and gives the best projection with limited information.

We use KDDCUP 1999 dataset as training and testing data [13]. Two forms of SVM: C-SVM and one-class SVM are used as classification technologies and LibSVM [14] is chosen as the SVM tool.

We find that we can get good results by using this tool without evolved procedures such as the selection of parameters, which is hard to decide when using SVM. There are many ways to try out best parameters, such as genetic algorithm (GA) which needs lots of computations and consumes much time.

The suitable SVM is found by observation of the experimental outcomes of anomaly detection by different types of SVM. The high average detection rates and low average false positive rates in anomaly detection show that our proposed approach is feasible.

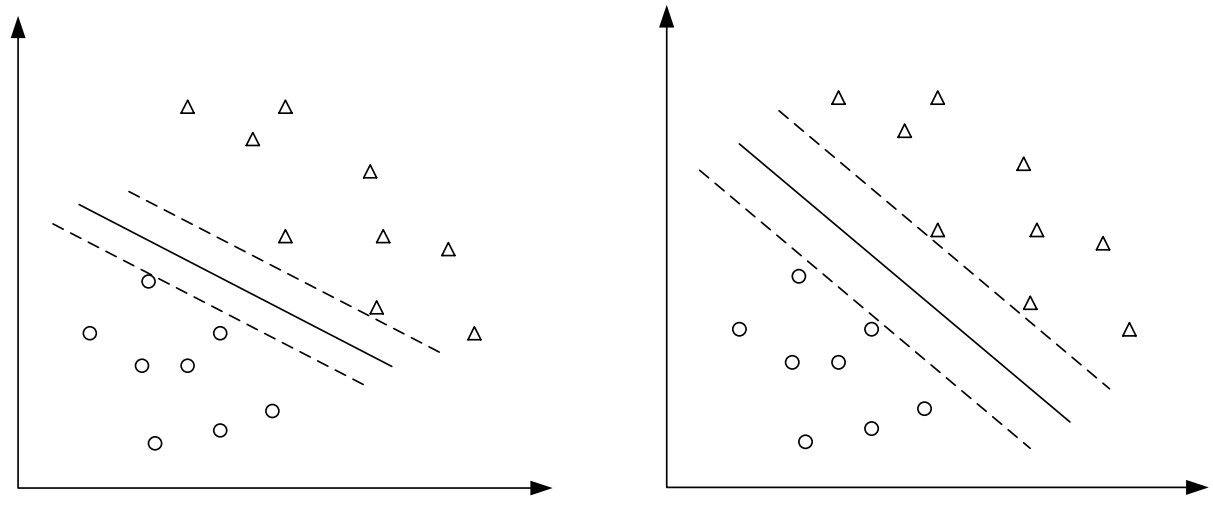
The rest of this paper is organized as follows. In Section 2, we will introduce briefly the two forms of SVM. In Section 3, we will present our experiment methods and in Section 4, the results. Conclusions will be given in Section 5.

2. SVM

Sometimes we want to catachrestically classify data into two groups. There exist a few good technologies for classification such as the naïve Bayes and neural networks. When applied correctly, these technologies give acceptable results. Most important advantages of SVM are simple to use and high precision.

SVM is a statistical learning theory based on machine learning methods. SVM is widely used in the respect of bioinformatics, data mining, image recognition, text categorization, hand-written digit recognition. The earlier SVM was designed to solve binary classification problems. It is important for SVM to solve multi-class classification in efficient ways. Some scholars propose related researches about multi-class SVM [15, 16].

The basic concept of SVM is to classify separable data in space R^d. We want to find a hyper-plane that separates these data into two groups, group A and group B in the R^d space. As shown in Figure 1, the data of group A are in the right and upper side of the hyper-plane, and the data of group B are in the other side of the hyper-plane. The margin between the two parallel hyper-planes in Figure 1 (a) is narrower than the gap between the two parallel hyper-planes in Figure 1 (b). Since hyper-planes with wider margin are preferred and so the hyper-plane in Figure 1 (b) is better.



1 (a) 1 (b)

Figure1. Concept of SVM

In some non-linear cases, by transforming source data with a kernel function into high dimension space, one can solve non-linear data in original dimensions by separating into two parts with linear method in high dimensions to reduce the error [17]. The concept is shown in Figure 2.

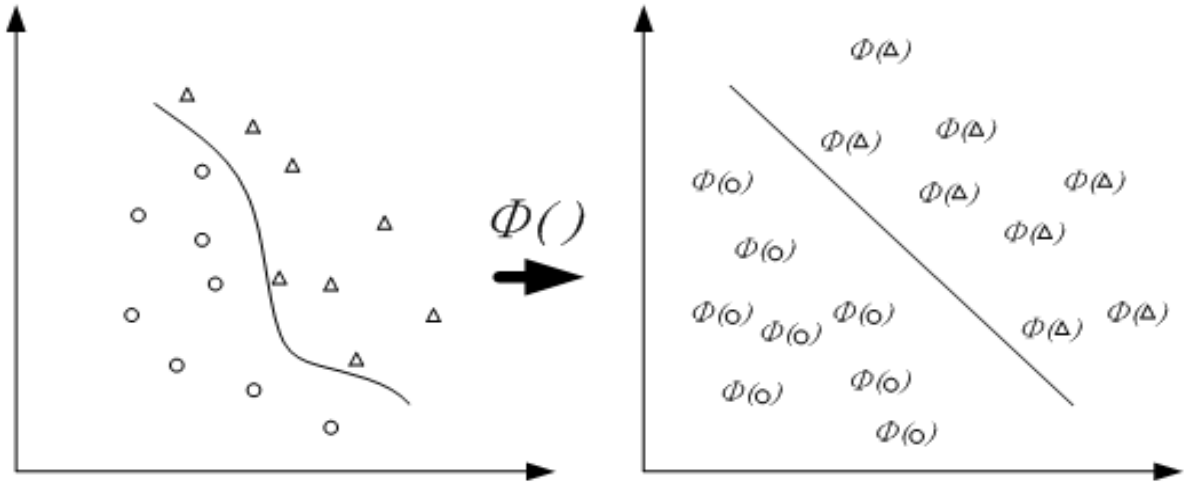


Figure2. Concept of non-linear data classifying

2.1. C-SVM

The C-SVM is proposed by Cortes and Vapnik in 1995 [18] and Vapnik in 1998 [15].
The primal form is:

$$\min_{w,b,\xi,\rho} \quad \frac{1}{2} w^t w - v\rho + \frac{1}{l}\sum_{i=1}^{l} \xi_i$$
$$s.t. \quad y_i\left(w^T \cdot \phi(x_i) + b\right) \geq \rho - \xi_i$$
$$,\xi_i \geq 0, i = 1 \ldots l, \rho \geq 0$$

Where vectors $X_i \in R^n$, $i = 1,...,l$ in two classes, and the vector $y_i \in R^l$ such that $y_i \in \{+1, -1\}$;
The decision function is:

$$\text{sgn}(\sum_{i=1}^{l} y_i a_i K(x_i, x) + b)$$

The dual is:

$$\min \quad \frac{1}{2}\alpha^T Q \alpha - e^T \alpha$$
$$Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$$
$$s.t. \quad y^T \alpha = 0$$
$$0 \leq \alpha_i \leq C, \quad i = 1,..., l$$

The geometry interpretation of C-SVM is shown in Figure 3.

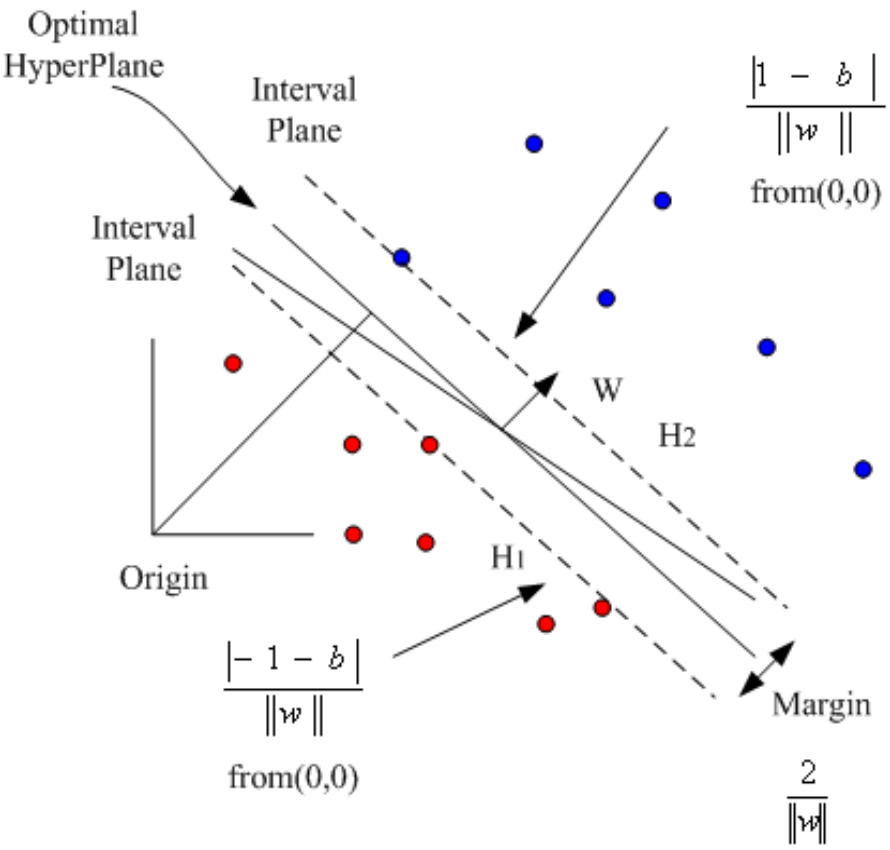


Figure 3. Geometry interpretation of C-SVM

As shown in Figure 3 the solid lines are the found hyper-planes. We call H_1 and H_2 the supporting hyper-planes. We want to find the best classification hyper-planes that have widest margin between the two supporting hyper-planes.

Definition of classification hyper-plane is:

$$w^T x = -b(w^T x + b = 0)$$

Therefore we can present supporting hyper-planes H_1 and H_2 as:

$$H_1 : w^T x + b + \delta$$

$$H_2 : w^T x + b - \delta$$

We scale H_1 and H_2 with constants w, b, and δ:

$$H_1 : w^T x + b = 1$$

$$H_2 : w^T x + b = -1$$

The distance from H_1 to the origin is $\frac{|1-b|}{\|w\|}$. The distance from H_2 to the origin is $\frac{|-1-b|}{\|w\|}$. The distance between H_1 and H_2 is $\frac{|2|}{\|w\|}$.

By above equation the data points should satisfy the following equations in R^d:

$$w^T x_i + b \geq 1 \quad for\ y_i = 1$$

$$w^T x_i + b \le -1 \quad for\ y_i = -1$$

We can combine the two above inequality as:

$$y_i\left(w^T x_i + b \right) \ge 1$$

And we get the widest margin between two Support Hyper-planes by:

$$\max\left(\frac{2}{\| w \|}\right), \text{ or } \min\left(\frac{\| w \|}{2}\right)$$

2.2. One-class SVM

One-class SVM was proposed by Schölkopf et al. in 2001 for estimating the support of a high-dimensional distribution [19]. The base idea of one-class SVM is to separate data from the origin. Schölkopf et al. proposed a method to adapt the SVM one-class classification problem. After transforming the feature by the kernel function, the origin is treated as the only member of the second class. Then the image of the one class is separated from the origin.

The algorithm can be summarized as mapping the data into a feature space H using a fit kernel function, and then trying to separate the mapped vectors from the origin with maximum margin:

Given training vectors $X_i \in R^n, \quad i = 1,..., l$, without any class information, the primal form is:

$$\min_{w,b,\xi,\rho} \quad \frac{1}{2} w^t w - \rho + \frac{1}{vl} \sum_{i=1}^{l} \xi_i$$

$$s.t. \quad \left(w \cdot \phi(x_i)\right) \ge \rho - \xi_i, \xi_i \ge 0$$

The decision function is:

$$\text{sgn}\left(\sum_{i=1}^{l} a_i K(x_i, x) - \rho \right)$$

The dual is:

$$\min \quad \frac{1}{2} \alpha^T Q \alpha \qquad s.t. \quad 0 \le \alpha_i \le 1, \quad i = 1,..., l$$

$$Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$$

$$e^t \alpha = vl$$

The geometry interpretation of one-class SVM is shown in Figure 4.

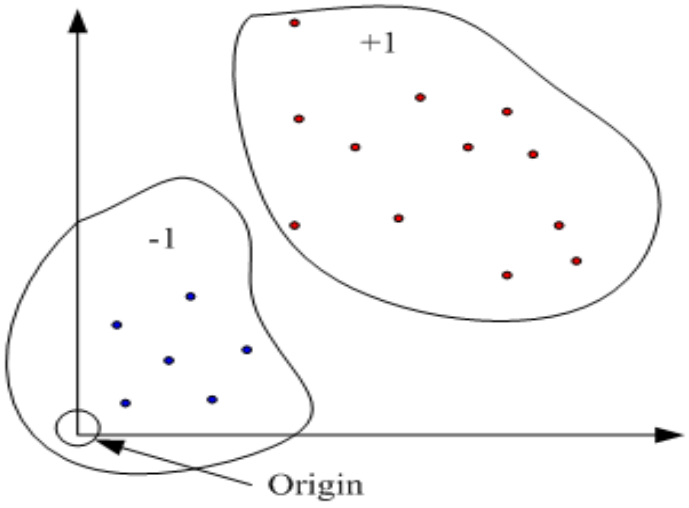


Figure 4. Geometry interpretation of one-class SVM

2.3. LibSVM

LibSVM is a library for support vector machines. Its goal is to promote SVM as a convenient tool. It integrates C-SVM classification, nu-SVM classification, one-class-SVM, epsilon-SVM regression, and nu-SVM regression. It also provides an automatic model selection tool for C-SVM classification. The latest version 2.85 is released on November 6, 2007.

3. Experiment Setup

3.1. Data Source

In our experiment, we used 1999 KDD Cup data set. These data are prepared and managed by MIT Lincoln Labs. Lincoln Labs set up an environment to acquire nine weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks. KDD Cup [20] is the leading Data Mining and Knowledge Discovery competition in the world, organized by ACM SIGKDD - Special Interest Group on Knowledge Discovery and Data Mining, the leading professional organization of data miners. In recent years, this data set has been widely used as a benchmark for evaluation of the intrusion detection technology.

We separated the source data into normal part and abnormal parts. There were twenty four known types of attacks in the source data. A connection was established when a sequence of TCP packets start and end at some well defined time span, in which data flowed between a source IP address and a target IP address under some well defined protocol. Each connection was labeled as either normal for normal users, or abnormal for attacks with exact one specific attacking type. Each connection record consisted of about 100 bytes.

3.2. Experimental Environment

The experiment was performed in the following experimental environment:
CPU: Pentium Core 2 Duel E6750
RAM: DDR II 667 2GB
OS: Microsoft Windows XP SP2
SVM tool: LibSVM 2.84 (released on April 1, 2007).

3.3. Data Processing

Non-numerical data items were changed into numerical data formats to make the data trainable by LibSVM. The characteristic of this data set was that 80% of it belonged to abnormal behaviors.

3.4. Data Feature

The KDD cup 1999 dataset contains all 41 features in Table 1, include basic features of individual TCP connections, content features within a connection suggested by domain knowledge, and traffic features computed using a two-second time window. The exactly feature descriptions are in [13]. We train SVM by using a large number of data contains in this feature.

Table1. Data features of KDD cup 1999 dataset

duration	is_guest_login
Protocol_type	Count
Service	serror_rate
src_byte	rerror_rate
dst_byte	same_srv_rate
flag	diff_srv_rate
land	srv_count
wrong_fragment	srv_serror_rate
urgent	srv_rerror_rate
hot	srv_diff_host_rate
num_failed_logins	dst_host_count
logged_in	dst_host_srv_count
num_compromise	dst_host_same_srv_rate
root_shell	dst_host_diff_srv_rate
su_attempted	dst_host_same_src_port_rate
num_root	dst_host_srv_diff_host_rate
num_file_creations	dst_host_serror_rate
num_shells	dst_host_srv_serror_rate
num_access_files	dst_host_rerror_rate
num_outbound_cmds	dst_host_srv_rerror_rate
is_hot_login	

4. Experiment Processes and Results

We performed our experiment in two parts:

A. Use C-SVM for classifying technology

We processed the source data to make them fit with the format of SVM. First we took 20,000 of them as a unit, and tested 5 times, labeled as A1~A5, respectively. Second we took 50,000 of them as a unit, and tested 5 times, labeled as B1~B5, respectively. Third, we took 100,000 of them as a unit, and tested 5 times, labeled as C1~C5, respectively.

We drew randomly 60% of each test as training data. The left 40 % of data were validation data. The results are listed in Table 2. It shows good test results with very high average detection rates above 97%.

Table2. Detection rates by using C-SVM for classifying technology

	Avg. Detection Rate	Avg. False Positive rate
A	98.7%	2.0%
B	97.6%	1.5%
C	97.3%	1.7%

B. Use one-class SVM for classifying technology

To compare the effectiveness of different detection algorithms, we duplicated the experiment environments in [21]. The p-kernel is a new kind of kernel described in [22] together with other techniques of detection. The normal data were divided into three parts: the training data, the test data and

the validation data. Then, we drew 10,000 samples from source data, among them 6000 samples as training data, 2000 samples as test data, and 2000 samples as validation data. The results are listed in Table 3. A high average detection rate of 95% is obtained by using LibSVM.

Table3. Detection rates by using one-class SVM for classifying technology.

Algorithm	Avg. Detection rate	Avg. False Positive rate
K-NN	91%	10%
Naïve Bayes	89%	8%
SVM(light)	91%	10%
P-kernel SVM	98%	6%
SVM (LibSVM)	**95%**	**7%**

5. Conclusions and future works

We use SVM on the learning-based anomaly detection system, whereas in the choice of tools, we use LibSVM as a SVM tool. We compare the effectiveness of this SVM tool with other algorithms of unsupervised SVM based on p-kernels for anomaly detection. In this paper we make use of p-kernel with SVM-light and get nearly perfect results. We can easily obtain good result with average detection rate up to 95% by using LibSVM alone, with its default parameters and kernel (RBF), and without the need of other external kernels.

As for C-SVM (standard SVM), by using the default parameters, the result of detection rate and false positive are very good. This proves that our method is simple and effective and that it can achieve high detection rates.

By comparing experimental results in A and B, the standard SVM is better than one-class SVM in this case.

We have obtained nice results by using LibSVM with the KDD Cup 1999 dataset and two forms of SVM. But since both the attacking technology and the detection technology keep updating very fast, our future work is to design new methods and train them with latest data in extreme cases, and to classify data efficiently with new forms of SVM.

6. Acknowledgement

This work was supported in part by Taiwan Information Security Center (TWISC), National Science Council under the grants NSC 95-2218-E-001-001, NSC95-2218-E-011-015, NSC 97-2115-M-126-003, and NSC 97-2221-E-029 -023.

7. References

[1] G.Giacinto, F.Roli, L.Didaci, “Fusion of multiple classifiers for intrusion detection in computer networks,” Pattern Recognition Letters. Vol.24, pp. 1795–1803, 2003

[2] N. Cristianini, and J.S. Taylor, “An introduction to support vector machine,” Cambridge University Press, Cambridge, UK, 2000.

[3] B.J. Kim, “Kernel Based Intrusion Detection System,” International Conference on Information Systems archive Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05) , pp.13-18, 2005

[4] H.Li, XH.Guan, X.Zan, et al, “ Network intrusion detection based on support vector machine,” Journal of Computer Research and Development, vol.40, no.6, pp.799-807, 2003

[5] M. Fugate and J.R. Gattiker, “Anomaly Detection Enhanced Classification in Computer Intrusion Detection,” In Pattern Recognition with Support Vector Machines, First International Workshop, Niagara Falls, Canada, August 10, 2002, Lecture Notes in Computer Science 2388, pp. 186-197.

[6] W. Hu and Y. Liao and V. R. Vemuri, “Robust Support Vector Machines for Anomaly Detection in Computer Security,” In Proceedings of 2003 International Conference on Machine Learning and Applications, Los Angeles, CA, June 23-24, 2003.

[7] S. Mukkamala, G. Janoski, A H. Sung, “Intrusion Detection Using Neural Networks and Support Vector Machines,” In Proceedings of IEEE International Joint Conference on Neural Networks, IEEE Computer Society Press, 2002, pp.1702-1707.

[8] S Mukkamala, A H. Sung, “Feature Selection for Intrusion Detection Using Neural Networks and Support Vector Machines,” Proceedings of the 82nd Annual Meeting of the Transportation Research Board, National Academics.

[9] L. Yang, G. Li, “An Efficient Network Anomaly Detection Scheme Based on TCM-KNN Algorithm and Data Reduction Mechanism” Information Assurance and Security Workshop, 2007. IAW '07. IEEE SMC20-22 Page(s):221 – 227, June 2007

[10] A. T. Quang, Q. Zhang, X. Li, “Attack recall control in anomaly detection,” In Proceedings of ICCT 2003, International Conference on Communication Technology, vol. 1, Page(s): 382 - 384, 9-11 April 2003

[11] S.-Y. Ohn, H.-N. Nguyen, D. S. Kim, J. S. Park, “Determining Optimal Decision Model for Support Vector Machine by Genetic Algorithm,” In International Symposium on Computational and Information Sciences, Shanghai, China, December 16-18, 2004, Lecture Notes in Computer Science, pp. 895-902.

[12] D. S. Kim, H.-N. Nguyen, J. S. Park, “Genetic algorithm to improve SVM based network intrusion detection system,” Advanced Information Networking and Applications, 2005. AINA 2005, 19th International Conference on Volume 2, 28-30 March 2005 Page(s):155 - 158 vol.2

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

[14] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[15] V. Vapnik, “Statistical learning theory,” John Wiley and Sons, New York, 1998.

[16] J. Weston and C. Watkins, “Multi-class support vector machines,” In Proceedings of ESANN99, Brussels, 1999.

[17] B. E. Boser, I. M. Guyon, V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," Fifth Annual Workshop on Computational Learning Theory, ACM, (1992).

[18] C.Cortes., and V. Vapnik, “Support vector networks,” Machine Learning, vol.20, no2, pp.273-297, 1995\

[19] B.Schölkopf et al. “Estimating the support of a High-Dimensional Distribution,” Technical Report, Department of Computer Science, University of Haifa, Haifa, 2001.

[20] http://www.sigkdd.org/kddcup/index.php

[21] K. Li and G. Teng, “Unsupervised SVM Based on p-kernels for Anomaly Detection,” ICICIC’06, 2006 International Conference on Innovative Computing, Information and Control.

[22] J.P. VERT, “Support Vector Machine Prediction of signal peptide cleavage site using a new class of kernels for strings,” Pacific Symposium on Biocomputing, vol 7:649-660, 2002.

Authors

Jung-Chun Liu received his B.S. degree in electrical engineering from National Taiwan University in 1990. He received M.S. and Ph.D. degrees from the Electrical and Computer Science Engineering Department at University of Texas at Austin, in 1996 and 2004, respectively. He is an assistant professor in the Computer Science Department at the Tunghai University, Taiwan. His research interests include digital signal processing, VLSI design, RF and microwave engineering, watermarking, embedded systems, and computer networks.

Chu-Hsing Lin received both of his B.S. and M.S. degrees in applied mathematics from National Tsing Hua University and National Chung

Hsing University, respectively. In 1991, he received his Ph.D. degree in computer sciences from National Tsing Hua University, Taiwan. Since then he has been a faculty of the Department of Computer Science and Information Engineering, Tunghai University. Dr. Lin is currently a professor and the chair of the CSIE department of Tunghai University. From 1995 to 1999, he has ever been the Director of the Computer Center of Tunghai. He has also been one of the Board Directors of the Chinese Information Security Association (CCISA) from 2001 till now. Dr. Lin has published over 50 papers in academic journals and international conferences. He has received over twenty project grants from government departments and private companies in recent years. In 2006, he was awarded the Outstanding Instructor Award of Master & Ph.D. Thesis by the IICM (Institute of Information & Computing Machinery). He was the winner of the 1991 Acer Long-Term Award for Ph.D. Dissertation. His current research interests include multimedia information security, wireless ad hoc networks, embedded systems applications.

Jui-Ling Yu received her Ph.D. degree in Department of Mathematics from Michigan State University, USA, in 2005. She currently is an assistant professor in the Department of Applied Mathematics at the Providence University, Taichung, Taiwan, since 2005. Her current research interests include numerical methods, industrial mathematics, and mathematical biology.

Wei-Shen Lai received his B.S. and M.S. degrees in computer science and information engineering from Feng Chia University and National Chiao Tung University, respectively. In 2002, he received his Ph.D. degree in computer science and information engineering from National Chiao Tung University, Taiwan. In 2004, he has been a faculty of the Department of Information Management, Chienkuo Technology University. His current research interests include network security and cryptography.

Chia-Han Ho received both of his B.S. and M.S. degrees in computer science and information engineering from Tunghai University in 2006 and 2008, respectively. Under the instruction of his adviser Professor Chu-Hsing Lin, he has published two international conference papers. The topics of his research interests include intrusion detection system, network security, and support vector machine. His current research focus on the analysis and improvement of digital forensic.