ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Parallelization of Intrusion Detection System

Diwakaran M

Assistant Professor, Dept. of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, India

ABSTRACT: Currently the interconnections between computer systems have been increasing rapidly. As a result of which network security has become a challenge to the security professionals. A hybrid Intrusion Detection System (IDS) integrating both misuse detection and anomaly detection is being proposed to enhance the security of the network. The misuse detection module and the anomaly detection module are connected in parallel and the results are integrated using a decision support system. The misuse detection module is built based on c5.0 decision tree algorithm and anomaly detection module using multi class SVM (Support Vector Machine) algorithm. The proposed intrusion detection system is to be evaluated by conducting experiments with KDD Cup99 dataset. Integration of misuse detection and anomaly detection module is more efficient than conventional intrusion detection systems as it reduces the false error rates and increases the efficiency of the detection system.

GENERAL TERMS: Intrusion Detection System: Signature Detection and Anomaly Detection.

KEYWORDS: C5.0 Decision Tree Algorithm, Multi-class Support Vector Machine.

I. INTRODUCTION

An intrusion is a set of actions performed to compromise a system's security, intrusion detection is process of identifying these malicious activities. Intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station. Intrusion Detection Systems are c classified based on various categories. Based on the scope of protection, they are classified into network based (NIDS) and host based (HIDS) intrusion detection systems. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, and reporting attempts. In addition, organizations use IDPS for other purposes, such as identifying problems with security policies, documenting existing threats and deterring individuals from violating security policies. IDPS has become a necessary addition to the security infrastructure of nearly every organization. IDPS typically records the information related to observed events, notify security administrators of important observed events and produce reports. Many IDPSs can also respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g. reconfiguring a firewall) or changing the attack's content.

II. DETECTION METHODOLOGIES

Intrusion detection methodologies are classified into three major categories: Signature based Detection (SD), Anomaly based Detection (AD) and Stateful protocol analysis (SPA).

A.Signature Based Detection

A signature is a pattern or string that corresponds to a known attack or threat. Signature detection is the process to compare patterns against captured events for recognizing possible intrusions. Because of using the knowledge accumulated by specific attacks and system vulnerabilities. Signature detection is also known as knowledge based detection or misuse detection. But in these methods new attacks cannot be detected, constant updation of new signatures is required.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

B.Anomaly Based Detection

An anomaly is a deviation to a known behavior, and profiles represent the normal or expected behaviors derived from monitoring regular activities, network connections, hosts or users over a period of time. Anomaly detection compares normal profiles with observed events to recognize significant attacks. Hence it is also called as behavior based detection. In case of anomaly detection, the rate of false positive is higher as it detects everything that deviates from normal activity.

C.Stateful Protocol Analysis

The word stateful indicates that IDS could know and trace the protocol states (eg., pairing requests with replies). Though stateful protocol analysis process looks like anomaly detection process, they are essentially different. Anomaly detection adopts preloaded network or host specific or host specific profiles, whereas stateful protocol analysis depends on vendor developed generic profiles to specific protocols. Generally, the network protocol models in stateful protocol analysis are based originally on protocol standards from international standard organizations, e.g., IETF. Stateful protocol analysis is also called as specification based detection.

Most IDSs use multiple methodologies to provide more extensive and accurate detection. For example, signature detection and anomaly detection are complementary methods, because the former concerns certain attacks/threats and the later focuses on unknown attacks.

III. RELATED WORKS

Many research have been done in the field of intrusion detection and prevention in networks. Gesung Kim, Seungmin Lee, Sehun Kim [1] have developed a new approach towards intrusion detection. These works are done in order to provide maximum security to the network by integrating intrusion detection and anomaly detection methods. A new hybrid intrusion detection method that hierarchically integrates a misuse detection model and an anomaly detection model in a decomposition structure is proposed. First, a misuse detection model is built based on the C4.5 decision tree algorithm and then the normal training data is decomposed into smaller subsets using the model. Next, multiple one-class SVM models are created for the decomposed subsets. As a result, the detection model uses the known attack information and also builds the profiles of normal behaviour very precisely.

In other works Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely [2] has proposed a hybrid detection framework that depends on data mining classification and clustering techniques is proposed. In misuse detection, random forests classification algorithm is used to build intrusion patterns automatically from a training dataset, and then matches network connections to these intrusion patterns to detect network intrusions. In anomaly detection, the k-means clustering algorithm is employed to detect novel intrusions by clustering the network connections' data to collect the most of intrusions together in one or more clusters.

Amuthan Prabakar Muniyandi, R. Rjeswari, R. Rajaram [3] designed a anomaly detection systems (ADS) monitor the behaviour of a system and flag significant deviations from the normal activity as anomalies an anomaly detection method using "K-Means + C4.5", a method to cascade k-Means clustering and the C4.5 decision tree methods for classifying anomalous and normal activities in a computer network. The k-Means clustering method is first used to partition the training instances into k clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances, decision trees were build using C4.5 decision tree algorithm. The decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster. To obtain a final conclusion, the results derived from the decision tree on each cluster are exploited.

Dr. Saurabh Mukherjee, Neelam Sharma have also proposed a paper called "Intrusion Detection using Naïve Bayes Classifier with Feature Detection" where the IDS was tested using performance of three standard feature selection methods using Correlation-based Feature Selection, Information Gain and Gain Ratio. This paper proposes method Feature Vitality Based Reduction Method, to identify important reduced input features. We apply one of the efficient classifier naive bayes on reduced datasets for intrusion detection. Empirical results show that selected reduced attributes give better performance to design IDS that is efficient and effective for network intrusion detection. For

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

building efficient and effective IDS we investigate the performance of three standard feature selection algorithms involving Correlation-based Feature Selection (CFS), Information Gain (IG) and Gain Ratio (GR) to identify important reduced input features. The reduced data sets are further classified by using common Naïve Bayes classifier on discretized values. Since results using discretized features are usually more compact, shorter and accurate than using continuous values.

IV. SYSTEM DESIGN

In the proposed system, the misuse detection module and anomaly detection modules are connected in parallel and the result of both are integrated using a decision support module. The basic block diagram of a parallel hybrid intrusion detection system is given below:

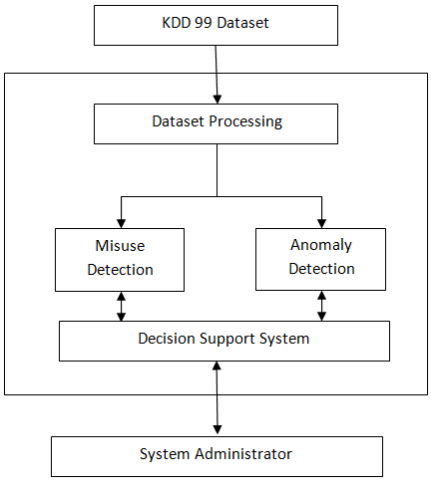

Fig 1: Hybrid IDS

The implementation process consists of three steps: dataset processing, modeling of C5.0 Decision Tree algorithm and design of Multiclass SVM.

A.DATASET PROCESSING

The KDD training dataset consist of 10% of original dataset that is approximately sixty four thousand single connection vectors each of which contains 41 features and is labelled with exact one specific attack type. . Each vector is labelled as either normal or an attack, with exactly one specific attack type. Deviations from normal behaviour, everything that is not normal, are considered attacks. Each connection record consists of about 100 bytes. Attacks fall into four main categories:

DoS: In this category the attacker makes some computing or memory resources too busy or too full to handle legitimate request, or deny legitimate users access to machine. DoS contain the attacks: neptune, back, smurf, pod, land, and teardrop.

R2L: It provides unauthorized access from a remote machine. In this category the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to obtain root access to the system. U2R contains the attacks: buffer_overflow, loadmodule, rootkit and perl.

U2R: It provides unauthorized access to local superuser (root) privileges. In this category the attacker sends packets to machine over a network but who does not have an account on that machine and exploits some vulnerability to gain

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

local access as a user of that machine. R2L contain the attacks: warezclient, multihop, ftp_write, imap, guess_passwd, warezmaster, spy and phf.

Probing: It provides surveillance and other probing. In this category the attacker attempt to gather information about network of computers for the apparent purpose of circumventing its security. Probe contains the attacks: portsweep, satan, nmap, and ipsweep.

The major objectives performed by detecting network intrusion are stated as recognizing rare attack types such as U2R and R2L, increasing the accuracy detection rate for suspicious activity, and improving the efficiency of real-time intrusion detection models.

There are several categories of derived features.

- The same host features examine only the connections in the past two seconds that have the same destination host as the current connection, and calculate statistics related to protocol behavior, service, etc.
- The similar same service features examine only the connections in the past two seconds that have the same service as the current connection.
- Same host and same service features are together called time-based traffic features of the connection records.

Some probing attacks scan the hosts (or ports) within a minute. Therefore, connection records were also sorted by destination host, and features were constructed using a window of 100 connections to the same host instead of a time window. This yields a set of so-called host-based traffic features. Unlike most of the DOS and probing attacks, there appear to be no sequential patterns that are frequent in records of R2L and U2R attacks. This is because the DOS and probing attacks involve many connections to some host(s) in a very short period of time, but the R2L and U2R attacks are embedded in the data portions of packets, and normally involve only a single connection.

Stolfo et al. used domain knowledge to add features that look for suspicious behavior in the data portions, such as the number of failed login attempts. These features are called content features. A complete listing of the set of features defined for the connection records is given in the three tables below. The data schema of the contest dataset is available in machine-readable form.

B.C5.0 DECISION TREE ALGORITHM

Quinlan popularized the decision tree approach (Quinlan, 1996). The latest public domain implementation of Quinlan's model is C5.0 which is a advanced version of the popular C4.5 decision tree algorithm. The primary issue of the decision tree algorithms is to locate the attribute that best divides the data into their corresponding classes. C5.0 builds decision trees from training data sets using the concept of information entropy. That is, it is based on the highest gain of each attribute. The gain is calculated using the following formula:

$$\mathbf{Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} f_s(A_i) \times Entropy(S_{A_i})}$$

Where Gain(S, A) is the gain of set S after a split over the A attribute; Entropy(S) is the information entropy of set S; n is the number of different values of attribute A in S; A is the proportion of items possessing Ai as the value for A in S; Ai is the ith possible value of A; and SAi is a subset of S containing all items where the value of A is Ai. Here, the entropy is obtained as follows:

$$\mathbf{Entropy(S) = -\sum_{j=1}^{m} f_s(j) \times log f_s(j)}$$

Where m is the number of different values of the attribute in S (entropy is computed for one chosen attribute) and $f_S(j)$ is the proportion of the value j in the set S. After the tree is created by maximizing the gain, the C5.0 model decomposes the data space such that certain decomposed regions become homogeneous. Then, C5.0 performs the final pruning step. This step reduces the classification errors caused by specializations in the training set; thus, it makes the tree more general. Here, the C5.0 is used to train the misuse detection model in the hybrid intrusion detection system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Both normal and attack data are used to train the model: C5.0 divides the data into decomposed regions and labels the regions as the classes of major data belonging to each decomposed region.
C5.0 algorithm was chosen over C4.5 algorithm because of the fact that C5.0 has more advanced features like:

- Variable misclassification cost
- Case weight attribute
- Provides option to view large decision tree as a set of rules which is easy to understand
- More memory efficient
- Reduce error pruning technique
- Cross reference window
- It is faster than C4.5 and supports boosting
- Consumes less memory
- Lower error rates on unseen dataset

C.MULTI CLASS SVM

SVM (Support Vector Machine) is a classifier derived from statistical learning theory by Vapnik and Chervonenkis. Currently, SVM is closely related to: Kernel methods, large margin classifiers, reproducing kernel Hilbert space, Gaussian process.

SVM separates between these two classes via hyperplane that is optimally positioned to maximize the margin between the positive samples and the negative ones, then 'plot' the test data at the high dimensional space, distinguishing whether it belongs to positive or negative side according to the optimal hyperplane. Although SVMs were originally designed as binary classifiers, approaches that address a multiclass problem as single all together optimization problems exist, but are computationally much more expensive than solving several binary problems. A variety of techniques for decomposition of the multi class problem into several binary problems using support vector machines as binary classifiers have been proposed, and several widely used are:

One against all: for the N class problem (N>2), N 2-class SVM classifiers are constructed. The i^{th} SVM is trained while labeling all samples in the i^{th} class as positive examples and the rest as negative examples. In the recognition phase, a test example is presented to all N SVMs and is labeled according to the output among N classifiers. The disadvantage of the method is that the number of training samples is too large, so it is difficult to train.

One against one: this algorithm constructs {N(N-1)/2}, 2-classifiers, using all the binary pair wise combinations of the N classes. Each classifier is trained using the samples of the first class as positive examples and the samples of second class as negative examples. To combine these classifiers, it naturally adopts Max Wins algorithm that finds the resultant class by first voting the classes according to the results of each classifier and then choosing the class that is voted most. The disadvantage of this method is that every last test sample has to be presented to large number of classifiers (N(N-1)/2). This results in faster training but slower testing, especially when the number of classes in the problem is big.

Directed acyclic graph SVM (DAGSVM): Introduced by Platt the algorithm for training a N(N-1)/2 classifiers is the same as in one against one. In the recognition phase, DAGSVM depends on a rooted binary directed acyclic graph to make a decision. When a test sample reaches the leaf node, the final decision is made. A test sample is only presented only to the N-1SVMs in the nodes on decision path. This results in significantly faster testing while keeping a very similar recognition rate as one against one.

The approach used here to classify the attacks is one against all classifier, because of its minimum error rate. For a binary classification problem with input space X and binary class labels Y: $Y \in \{-1, 1\}$. Giving training samples (y_1, x1), ..., (y_i, xi). $Yi \in \{-1, 1\}$. The goal of SVM is to search for the optimal hyperplane.

$$\mathbf{W}.\mathbf{X} + \mathbf{B} = \mathbf{0}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

with variables w and b that satisfy the following inequality yi(w·xi+b)≥1, i= 1, . . . ,i, defining the minimum distance between two class groups in the new projection.

$$\mathbf{d(w,b)} = \min_{x/y} \frac{\mathbf{x^t.w}}{\|\mathbf{w}\|} - \max_{x/y} \frac{\mathbf{x^t.w}}{\|\mathbf{w}\|}$$

For a given training set w,b that maximizes d(w0,b0) solve the following quadratic optimization problem: $\mathbf{min}_{w} \frac{1}{2} \boldsymbol{w}.\boldsymbol{w}$, satisfying yi(w·xi+b)≥1,i= 1, . . . ,i.

If the given training sample is linearly seperable, the optimization problem has feasible solutions. The optimal solution w and b forms the best hyperplane that maximizes the margin between two different classes in the new projection. Because the SVM search is best for the hyperplane separation instead of highest training sample accuracy. If the parameters are properly selected, the SVM typically produces both excellent classification results and good generalization. Not every problem is guaranteed to be linearly separable, so soft margin hyperplane SVM was developed to separate the training dataset with minimal number of errors. A number of candidate kernel training functions have been used in SVM, including polynomial.

$$\mathbf{k(x,y)} = \mathbf{(1 + x.y)^d}$$

Exponential RBF:

$$\mathbf{k(x,y)} = \mathbf{exp}\ (-\frac{\|\mathbf{x-y}\|}{\mathbf{2\sigma^2}})$$

And Gaussian RBF:

$$\mathbf{k(x,y)} = \mathbf{exp}\ (-\frac{\|\mathbf{x-y}\|^2}{\mathbf{2\sigma^2}})$$

For the new data point x, the classification is then performed as,

$$\mathbf{y = sign(f(x))}, \qquad \mathbf{f(x)} = \sum_{i=1}^{N_{sv}} \alpha_i y_i \mathbf{K(x, x_i)} + \mathbf{b}$$

Where N_{SV} is the number of support vectors

SVMs are also used for Pattern recognition, Signal processing, Data mining, Abnormality detection, Interpretation procedure in different areas of medicine, Intrusion detection (internet, networking, wireless), Security issues, identifying different type of Attacks, Information retrieval systems, Enhancement of search engines results, and mail filtering.

V. RESULT ANALYSIS

The total number of attack signatures in the dataset is 65536. After processing the dataset into machine readable format, a decision tree is constructed with the same. To construct the tree, entropy gain values of each attribute in the dataset is calculated. The attribute with maximum gain value is used for classification process and the attribute with maximum gain is the “flag”. Thus, flag acts as the root node of the tree which classifies all other attributes.

- The number of classes with “normal” attack = 65532
- The number of classes with “buffer overflow” attack = 2
- The number of classes with “load module” attack = 1
- The number of classes with “perl” attack = 1
- The number of attacks with icmp protocol = 248
- The number of attacks with tcp protocol = 64080
- The number of attacks with udp protocol = 1208

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Table.1-Comparison of Execution Time of C4.5 & C5.0 Algorithms

Features	C4.5 Decision Algorithm	C5.0 Decision Algorithm
Training Time	56.58 seconds	40 seconds
Testing Time	11.2 seconds	10 seconds

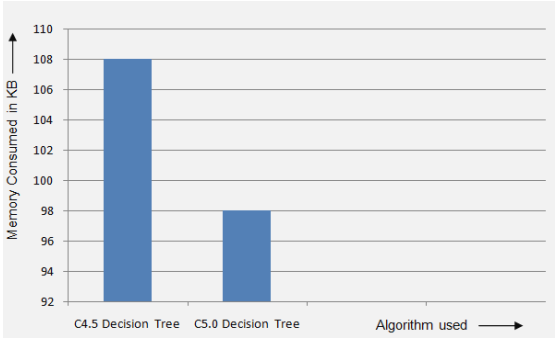

Fig 2: Comparison of C4.5 & C5.0 Algorithms

Then info gain of each attribute in the dataset is calculated with these entropy values. The maximum info gain is obtained for "flag" attribute, the gain value is 726817.4999 The flag is classified with the different types of attack classes in the dataset.

SVM classification is done with "protocol type" and "class name" as main attributes. The maximum number of attacks in the given dataset has "tcp" connection with "normal" class. Since most of attacks are done using tcp protocol, the classifications for tcp with different error classes are done.

For example, tcp connection with buffer overflow attack is taken as separate calss and tcp connection with load module attack is classified as separate class. Similarly all connection are classified.

The time taken for the C5.0 algorithm is comparatively lesser than that of previously used C4.5 algorithm, thus fastening the detection rate of the system. Also using multiclass SVM gives more accuracy to the system. Now a days multiclass SVM is the best method used for outliner classification which is because of its high accuracy character.

Features	1-Class SVM	Multiclass SVM
Training Time	41.42 seconds	30 seconds
Testing Time	29.1 seconds	19.5 seconds

Table.2- Comparison of Execution Time of 1-class & Multiclass SVM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

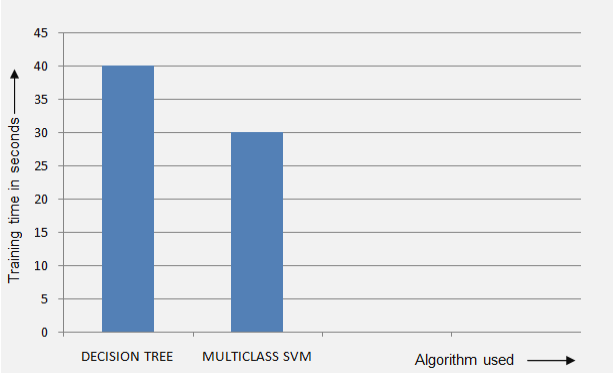

Fig 3: Graphical Representation of time taken for detection

Hence, a Hybrid Intrusion Detection System with more accuracy in terms of anomaly detection and faster execution in terms of Signature Detection is designed.

VI. CONCLUSION

Thus, a new hybrid intrusion detection method that integrates a misuse detection model and an anomaly detection model in parallel is designed. First, the C5.0 decision tree (DT) was used to create the misuse detection model that is used to decompose the normal training data into smaller subsets. Then, the multi-class support vector machine (multi-class SVM) was used to create an anomaly detection model. The results show that the proposed hybrid model of intrusion detection system performs faster than the previous conventional models. This increases the rate of detection of attacks and also reduces the number of false positives on the detection system. Hence, the security of a network can be enhanced. The performance of new parallelized intrusion detection system was tested with KDD Cup99 Dataset.

REFERENCES

1. Gisung Kim, Seungmin Lee, Sehun Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", published in Expert Systems with Applications, Volume 41, March 2014.
2. Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, Mahmoud M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means", published in Ain Shams Engineering Journal, Volume 4, Issue 4, December 2013.
3. Amuthan Prabakar Muniyandi , R. Rajeswari , R. Rajaram "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm", Procedia Engineering, Volume 30, 2012.
4. G.V. Nadiammai, M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal, 2014.
5. Mrutyunjaya Panda, Ajith Abraham, Manas Rajan Patra, "A Hybrid Intelligent Approach for Network Intrusion Detection", Procedia Engineering, Volume 30, 2012.
6. Dr. Saurabh Mukherjee, Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", Procedia Technology, Volume 4, 2012.
7. Basant Agarwal, Namita Mittal, "Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques", 2nd International Conference on Communication, Computing & Security, Porcedia Technology, 2012
8. Nagaraju Devarakonda, Srinivasulu Pamidi, Valli Kumari V, Govardhan A, "Intrusion Detection System using Bayesian Network and Hidden Markov Model", Procedia Technology, Volume 4, 2012
9. Zhaoyang Qu, Xiaobo, "Improving Pattern matchinf algorithm of intrusion detection", Procedia Engineering, Volume 15, 2011.
10. Weijun li, Zhenyu Liu, "A method for SVM with normalization in intrusion detection", Procedia Environmental Sciences, Volume 11, 2011.