# Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (10): 1865-1870, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.1865.1870


# An Novel Intrusion Detection System Based on Naive Bayesian Algorithm

Hui Wang, Hongyu Chen and Shanshan Yang
College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

**Abstract:** With increasing Internet connectivity and traffic volume, recent intrusion incidents have reemphasized the importance of network Intrusion Detection System (IDS). According to the deficiency of the Naive Bayesian (NB) algorithm, this paper presents an improved NB algorithm, which is addition of an attribute-added method to the traditional NB. This algorithm which is based on the original model and combined with a controlling parameter to enhance the accuracy of classification, the best parameter obtained by experiments can not only simplify both the time complexity and space complexity of the intrusion detection but also optimize the classification performance. The experimental results prove that this proposed approach applied into the intrusion detection framework can drastically reduce the false alarm rate of IDS so as to improve the detection efficiency and decrease economic damage brought by the cyber attack.

**Key words:** Naive bayesian, IDS, attribute-added method, controlling parameter, false alarm rate

## INTRODUCTION

In recent years, network attacks have increased in number and severity along with the rapid growth of online users. The 2013 Internet Security Threat Report of Symantec uncovered that the quantity of network intrusion targeted at credit card numbers, passwords and other financial information have increased from approximately 9 million in 2004 to 60 million in 2012 which has risen by more than 750% ,and the global economic losses reached about $1140 billion per year. Therefore, the research and development of intrusion detection technology is a great challenge and intrusion Detection System (IDS) has become a essential security infrastructure of most organizations.

Naïve Bayesian (NB) algorithm is a bayes method based on a simple hypothesis that assumes the different feature attributes of the samples are independent with each other. Naïve Bayesian classifier (NBC) is an application based on the Bayesian theory, which is one of the most widely used classifier now. The method of implementing intrusion detection is actually to design a network events classifier that will distinguish the normal and abnormal data from the data flow, so that it can realize the alarm function of network attacks. Thus, the contribution of this paper is to make the IDS has a better detection efficiency and robustness by using the improved NBC.

## RELATED RESEARCH

**Research of intrusion detection:** As demonstrated, Data Mining based intrusion detection falls into two categories: misuse detection and anomaly detection (Mohammad *et al*., 2011). Misuse detection attempts to match the features of the known attacks in the sample set with the unknown actions in the network traffic. Its advantages are the lower false positive rate and the faster detection speed. However, the disadvantage of misuse detection is that it cannot detect novel attacks that were not included in the sample set, but it relies on a learning algorithm to make up for these deficiencies. This learning algorithm is trained by a dataset in which each instance is labeled as either a normal event or an intrusion. Although the algorithm cannot detect novel attacks that were not included in the training set, it can be automatically retrained with the new attack instances through a new training dataset. In contrast, Anomaly detection can identify the new or unseen attacks because it builds models of normal network events and detects the events that deviate from these models. Despite its advantages, the anomaly detection method suffers from a high false alarm rate due to previously unnoticed normal events.

Therefore, the key question for the development of the effective IDS is how to reduce the false alarm rate of the system and improve the classification accuracy. Currently, a great deal of technologies have been used in IDS and machine learning approach such as: Decision tree learning (Sivatha Sindhu *et al*., 2012), Support vector machine (Yi *et al*., 2011) and Artificial neural network (Wang *et al*., 2010).

**Research of Naïve Bayesian:** Until now, NBC has been widely used in the fields of today's computer information such as: Spam filtering (Yang *et al*., 2011) and intrusion detection (Panda and Patra, 2007). Compared with other

**Corresponding Author:** Hui Wang, College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

classifiers, the advantages of NBC are that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification and it is not very sensitive to the missing data.

In the early time, Panda and Patra (2007) have effectively used NBC into IDS and established a primary intrusion detection framework based on NBC. Farid *et al.* (2010) presented a new hybrid learning algorithm for adaptive network intrusion detection using NBC and decision tree, which analyzes the large volume of network data and considers the complex properties of attack behaviors to improve the performance of detection speed and detection accuracy. A new algorithm Averaged Tree Augmented Naive Bayes (ATAN) is proposed by Jiang and Cai to improve the class probability estimation performance of NB (Jiang *et al.*, 2012). The algorithm can weaken the conditional independence assumption of Naive Bayes effectively to enhance the detection rates of NBC.

In this study, based on the above analysis and practical experience, we proposed an improved Naïve Bayesian model. This model extends the existing intrusion detection framework and introduces the new preprocessing of feature selection and discretization into the framework to decrease the complicacy of network data to a certain extent and improve the speed and the accuracy of classification greatly. Distinct from the ordinary Naive Bayesian classification model, the improved Naive Bayesian algorithm adds the attribute-added method and controls the accuracy and misclassification rate of the classification of network events by selecting proper parameters. In addition, it combined with machine learning and automatically retraining approach to find the best classification of NBC.

## INTRUSION DETECTION FRAMEWORK

**Proposed model for naïve bayesian classifier:** Naive Bayes is the simplest form of Bayesian network classifiers. Compared with the other classifiers, Naive Bayesian algorithm has more widely been used in the area of events classification for the simply method can classify events correctly and more quickly. The mainly core of this algorithm is as follows.

Let $A_i$ be a sample of network events represented as a vector of n feature values, $A_i = \{a_1, a_2, ..., a_n\}$, where i is the number of the sample. Given a set of class labels $C_j \in \{f: A_i \rightarrow C_j | C_1, C_2, ..., C_m\}$, where the function mapping relation f is letting $A_i$ be the true class label of $C_j$. Now, given a training set of attack instances $X = \{X_1, X_2, ..., X_n\}$ and the true class label C of it, where $X_i$ is represented by an attribute vector $\{x_1, x_2, ..., x_n\}$ and the set $C = \{C_1, C_2, ..., C_m\}$. Then, we used the NBC to estimate the probability that $X_i$ belongs to C. The calculation steps are as follows:

To estimate the prior probability $P(C = c_j)$ using Eq. 1.

$$P(C = c_j) = \frac{\sum_{j=1}^{n} L(c_i) + 1}{Total\_N + Total\_C} \quad (1)$$

where, $\sum_{j=1}^{n} L(c_i)$ is the frequency of occurrence of the class label C in the training set, $c_j$ $(1 \le j \le n)$ is the jth class sample, Total_N is the total number of the training samples, Total_C is the quantity of classes.

Using Eq. (2) to estimate the conditional probability $P(A = a_i | C = c_j)$.

$$P(A = a_j | C = c_j) = \frac{\sum_{i=1,\ j=1}^{n} L(a_i, c_i) + 1}{\sum_{j=1}^{n} L(c_i) + Total\_A} \quad (2)$$

where, Total_A is the total values of the feature $a_i (1 \le i \le n)$ in the training samples:

$$\sum_{i=1,\ j=1}^{n} L(a_i, c_i)$$

is the frequency of occurrence of the feature $a_i$ belongs to the class label $c_j$ in the training set.

According to the above two equations, the total probability $P(A = a_i)$ can be calculated by using Eq. 3.

$$P(a_i)\sum_{j=1}^{n} P(a_i | c_j)P(c_j) = \sum_{j=1}^{n} \frac{\sum_{i=1,\ j=1}^{n} L(a_i, c_i) + 1}{\sum_{j=1}^{n} L(c_i) + Total\_A} \times P(c_j) \quad (3)$$

Based on the derived prior probability $P(X = c_j)$, the law of total probability and Bayes formula, the posterior probability $P(c_j|a)$ of each training sample will be computed in Eq. 4:

$$P(C = c_j | A = a_i) = \frac{P(C = cj)\prod_{i=1}^{n} P(a_i | c_j)}{P(A = a_i)} \quad (4)$$

The above equation is used to figure out the probability $P = \{f: Y \rightarrow C | P_1, P_2, ..., P_n\}$ of $y_i$ belongs to each of the classes, where function mapping relation f is letting Y be the true class label of C, Y is the sample set of

$\{y_1, y_2,..., y_n\}$ and C is the class set of $\{c_1, c_2, ..., c_n\}$. Then, these several probabilities are sorted and normalized for getting the class similarity of the sample $y_i$ belongs to each of the classes. Finally, Getting the Maximum A Posteriori Probability (MAP) $P_{MAP}$ for selecting the final class of the sample in Eq. 5:

$$\begin{aligned} P_{MAP} &= \arg\max_{\exists m \in M} P(c_j \mid y_i) \\ &= \arg\max_{\exists m \in M} \frac{\prod_{i=1}^{n} P(y_i \mid c_j) \times P(c_j)}{P(y_i)} \\ &= \arg\max_{\exists m \in M} \prod_{i=1}^{n} P(y_i \mid c_j) \times P(c_j) \end{aligned} \quad (5)$$

Given the above derivation process, the corresponding classifier NBC is the function of classification defined in Eq. 6.

$$c(y_1, y_2,..., y_n) = \arg\max_{\exists c_j \in (1 \le j \le n)} \prod_{i=1}^{n} P(y_i \mid c_j) \times P(c_j) \quad (6)$$

In this thesis, a multistage structure of NBM is established based on the above mathematical model. There is only one father node C and child nodes $A_i$ $(1 \le i \le n)$ in the structure and every child node is represented by a number of mutually independent attribute nodes $a_i$ $(1 \le i \le n)$, where the relationship of these defined nodes is expressed as the mapping function f = $\{C_1:< A_1, (a_1, a_2, ..., a_n)>, C_2:< A_2, (a_1, a_2, ..., a_n)>, ..., C_n:< A_n, (a_1, a_2, ..., a_n)>\}$. All of the sample nodes $A_i$ belong to a common class node C and they are the pairwise independent events, as shown in Fig. 1.

Figure 1 not only illustrates the independent relationship and structural meanings of each node, but it also reflects both the core idea of obtaining the posteriori probability from the priori probability and the learning process of updating the training set by the posteriori probability. And then, it will find out the maximum possible classification. Despite the fact that the far-reaching independence assumptions are often inaccurate, the multistage NBC has several properties that make it surprisingly useful in practice.

**Attribute-added method:** Theoretically, NB algorithm finally determines the optimum class node $c_j$ $(1 \le j \le n)$ by using the calculated MAP of the true class label that the sample node belongs to . But in fact, according to the different influence of various factors, such as choosing diverse feature attributes and training sets would definitely lead to attenuation to varying degrees on the classification accuracy of NBC.

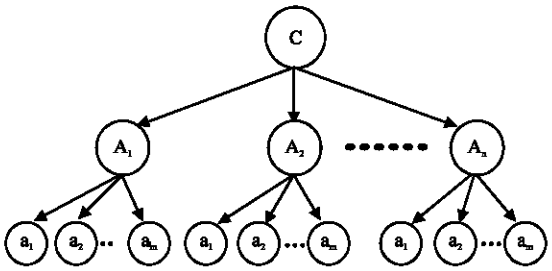


Fig. 1: A multistage Naïve Bayes structure

According to the intrusion detection technology, network events $A_i$ $(1 \le i \le n)$ can be divided into normal events $C_n$ and abnormal events $C_j$ $(1 \le j \le n)$, where N is a constant and their probabilistic relation is represented as $Uc = \{\exists C_N, C_j \mid P(C_N)+P(C_j) = 1\}$. Generally, using Bayes formula to work out the posterior probabilities $P(C_N|A_i)$, $P(C_j|A_i)$ and relying only on MAP to roughly judge that if one probability value is larger than another to classify a event into one type is not precise.

So in this article, on the above algorithm's foundation, the attribute-added method is added into the NB algorithm to improve the classification accuracy by regulation parameter $\theta$. Finally, making the classification efficiency of NBC to attain the best optimization by correctly choosing the regulation parameter $\theta$. The next calculation of above algorithm is as follows: when the inequality $P(C_j|A_i) > P(C_N|A_i)$ is established, converting it to another inequality:

$$\frac{P(Cj \mid Ai)}{P(CN \mid Ai)} > 1$$

Then introducing the controlling parameter $\theta$ into the inequality to work out a new discriminate influenced by the value of $\theta$. The above process of calculation and the final equation are represented in Eq. 7:

$$\begin{aligned} &\frac{P(C_j \mid A_i)}{P(C_N \mid A_i)} > 1 \Rightarrow \ln \frac{P(C_j \mid A_i)}{P(C_N \mid A_i)} > \theta \Rightarrow \frac{P(C_j \mid A_i)}{P(C_N \mid A_i)} > e^{\theta} \\ &\Rightarrow \frac{P(C_j \mid A_i)}{1 - P(C_i \mid A_i)} > e^{\theta} \Rightarrow P(C_j \mid A_i) > \frac{e^{\theta}}{1 + e^{\theta}} = \psi \end{aligned} \quad (7)$$

The maximum value $\Psi$ can be calculated by using the relationship of these two parameters which is expressed as the mapping relation f: $\{MAX [e^{\theta}/(1+e^{\theta})] \rightarrow \Psi\}$. Then, according to the Squeeze Theorem, given a limiting equation of the posterior probability $\lim P(C_j|A_i) = \Psi$ and for any arbitrary integer $\alpha$, when the condition $\exists n > N$ holds up, the inequality $\alpha - \Psi \le P(C_j|A_i) = \Psi$ will be proved, where the two premises $\alpha - \Psi > 0$ and $\alpha + \Psi < 1$should be

satisfied simultaneously. Finally, the best value $\Psi$ can be estimated based on the calculated limiting value $f(x_i)$ so as to more accurately determine which class $C_j$ is the type that the event $A_i$ belongs to. The best selected value $\theta$ is obtained from the experimental data.

**Intrusion detection framework based on NBC:** Detecting network intrusion events of IDS is a uncertainty behavior and Naïve Bayesian is suitable for solving the probability event. So, it is scientific and reasonable using the intrusion detection technology based on NBC to design IDS.

Figure 2 illustrates both training and testing processes of intrusion detection flow. In the first phase, firstly matching the known network traffic data $D_k$ with the known class labels $C_k$ in the sample and labeling them into the training set for training by the relationship $U_D = \{(D_k, C_k)|f: D_k \neg C_k\}$, where k is a variable from 1 to n. Then extracting the simplified and effective data and performing statistical analysis on the data by preprocessing the complex data labeled in training set with data discretization and feature selection. Finally, building Naïve Bayesian classifier according to the estimated priori probability $P(D_k|C_k)$, $(1 \leq k \leq n)$. The whole detection flow during this phase is a supervised learning process which is to gradually supplement the labeled data in the training set that will make the classifier has a good predicting effect.

In the second stage, firstly, by discretizing the new network traffic data $D_u$ $(1 \leq u \leq n)$ in the test set and using NBC generated by the first phase to test classification accuracy and there is a mapping function:

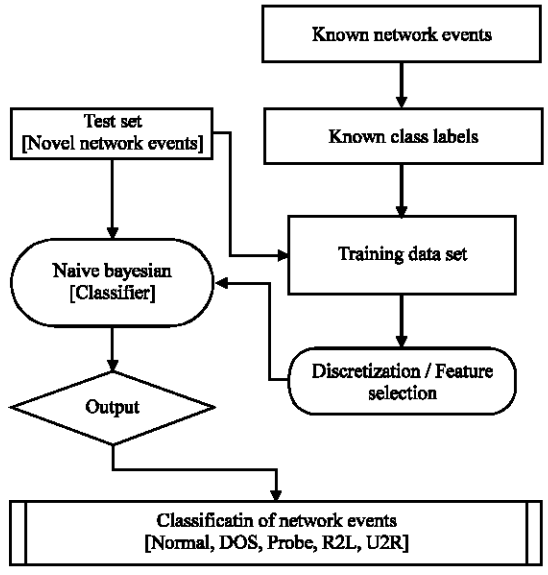


Fig. 2: Intrusion detection framework based on NBC

$$f_{NBC} = \{(U_n:<D_u, C_k>, U_k:<D_k, C_k>)|U_u \neg U_k\}$$

where, $U_u$ and $U_k$ in the function respectively denote the training set and test set. Then, to obtain the correct classes C which the unknown network events X belong to by using the approach of matching unknown network events with the labeled samples in the set, where C belongs to the set of five event types $U = \{C_j \neq \varnothing, 1 \leq j \leq 5\}$.

DOS, Pr obe, R2L, U2R and X is a set contains unknown events $\{X_1, X_2, ..., X_n\}$. And the data of the training set will be simultaneously updated again to perfect the detection performance of the classification model. The adoptive training set and test set in the framework all from the classic KDD Cup 1999(KDD'99) intrusion detection data set which can check the generalization ability and predictive ability of the classifier with its design mode.

## EXPERIMENTAL RESULTS AND ANALYSIS

**KDD'99 intrusion detection dataset:** This experiment adopts the KDD'99 intrusion detection dataset to test our IDS which consists of two parts: training data of approximately 5 million network connection records during 7 weeks and testing data of about 2 million network connection records during the last 2 weeks. The descriptions and examples of all the four attack types are DOS, R2L, U2R and Probe. Table 1 shows the data of 5 classes of network events records and their distribution in the training set and test set respectively of the 10% KDD data set.

**Experiment and results:** The experiment is operated in a PC of 2.3GHz CPU with 4GB RAM and 500GB HDD in MATLAB 8.0 Environment. In order to conduct simulated test with the above 10% KDD'99 dataset and the different simulation results are respectively shown in Table 2 and Table 3.

It is observed from Fig. 3 that the accuracy of the proposed algorithm increased nonlinearly with the rate rising to a peak at H(V), then fell gradually, but the error rate can decrease low point at L(V) along with a proper parameter . So, the best value $\theta_{bset}$ will be obtained that

Table 1: Characteristics of the 10% KDD'99 dataset

| Class | Training samples | Testing samples | 10% KDD training data distributions (%) | 10% KDD test data distributions (%) |
|---|---|---|---|---|
| Normal | 97277 | 60592 | 19.69 | 19.48 |
| DOS | 391458 | 237594 | 79.24 | 73.91 |
| R2L | 1126 | 8606 | 0.23 | 5.20 |
| U2R | 52 | 70 | 0.01 | 0.07 |
| Probe | 4107 | 4166 | 0.83 | 1.34 |
| Total | 494020 | 311028 | 100.00 | 100.00 |

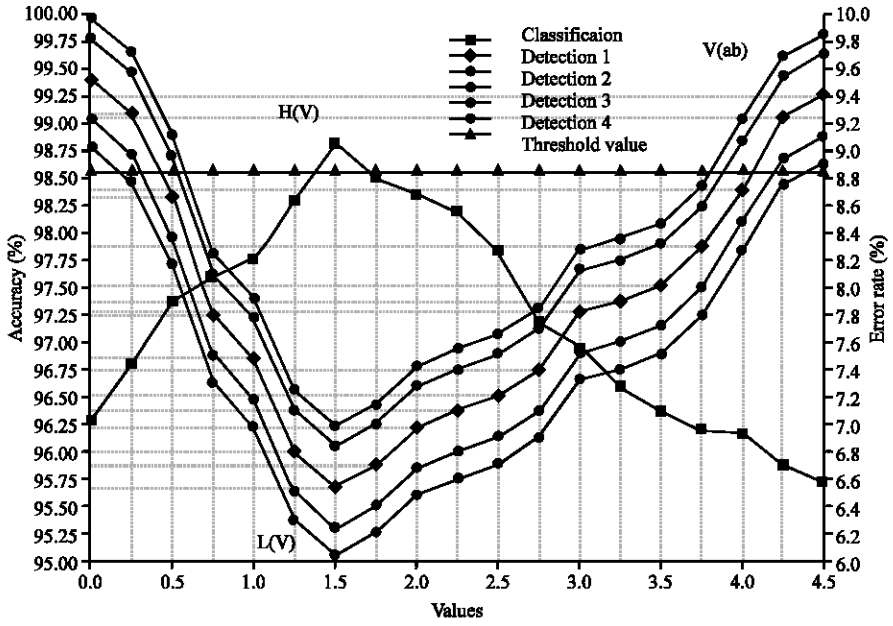


Fig. 3: Probability distribution of classification according to different values

Table 2: Comparison of detection rates in training set and test set

| | Method | | | |
|---|---|---|---|---|
| | Naïve bayesian | | Proposed Naïve bayesian | |
| Class | Training set (%) | Test set (%) | Training set (%) | Test set (%) |
| Normal | 98.89 | 96.99 | 99.54 | 99.58 |
| DOS | 99.51 | 97.76 | 99.86 | 99.84 |
| R2L | 99.46 | 99.62 | 99.59 | 99.13 |
| U2R | 92.18 | 92.33 | 95.62 | 96.12 |
| Probe | 91.27 | 92.38 | 98.32 | 98.36 |

Table 3: Comparison of the overall classifier performance

| Model | Normal | DOS | R2L | U2R | Probe |
|---|---|---|---|---|---|
| Proposed Naïve Bayesian | 99.56 | 99.85 | 99.36 | 95.91 | 98.34 |
| Decision tree | 98.75 | 98.68 | 97.73 | 94.83 | 96.89 |
| Support vector machine | 99.23 | 98.49 | 96.88 | 95.57 | 93.71 |
| Naïve Bayes | 97.94 | 98.63 | 99.54 | 92.26 | 91.82 |
| Artificial neural network | 97.74 | 98.66 | 98.41 | 93.37 | 92.97 |
| Genetic algorithm | 97.69 | 98.92 | 98.72 | 93.18 | 90.99 |

simultaneously corresponds to both H(V) and L(V). Firstly, setting the threshold value to limit the value range which is from 0.75 to 3.5 and the part of values exceeding the set point are considered as loss values V(ab), which will not occur in the allowable range of the chosen $\theta$. Then the best $\theta$ value which is 1.5 can be calculated with the built set of parameters $U_V = \{(\theta \epsilon [0.75, 3.5]) | f_u\}$, where the relationship of values is represented as the mapping function $\theta_{bset}$. Finally, injecting the selected $\theta_{bset}$ into the intrusion detection model based on the multistage NBC to improve the detection efficiency, where the accuracy and error rate can be reached 98.82 and 6.54%, respectively.

## CONCLUSION

For the complex and volume network events on the Internet, the intrusion detection technology is actually a set consist of a train of indeterminate actions. For the reason that Naïve Bayesian is fit for solving the probabilistic problem, so applying the NBC into the IDS is totally feasible. Therefore, the improved NB based on an attribute-added method is presented by referring to the experience and idea of the predecessors and making up for deficiencies of the traditional Naïve Bayesian.

On the basis of the above analysis, the best controlling parameter $\theta_{bset}$ is calculated by experiments, then using $\theta_{bset}$ to produce the best performance of NBC. Finally, practice proves that the false alarm rate of IDS based on the proposed algorithm can be decreased greatly. And comparing to other methods, the classification rate of the improved NB is also reflecting a better advantage. However, the performance in many aspects of the proposed method also has comparatively huge rise space. To further enhance the predictive accuracy of the classifier for complicated and changeable network attacks is also the following research projects.

## ACKNOWLEDGMENT

This project is supported by the National Natural Science Foundation of China (No. 51174263), supported by Research Fund for the Doctoral Program of Higher

Education of China (No. 20124116120004), supported by the Doctor Funds of Henan Polytechnic University (No. B2010-62) and supported by Educational Commission of Henan Province of China (No. 13A510325).

## REFERENCES

Farid, D.M., N. Harbi and M.Z. Rahman, 2010. Combining naive bayes and decision tree for adaptive intrusion detection. Int. J. Network Secur. Appl., 2: 12-25.

Jiang, L., Z. Cai, D. Wang and H. Zhang, 2012. Improving tree augmented naive bayes for class probability estimation. Knowl. Based Syst., 26: 239-245.

Mohammad, M.N., N. Sulaiman and O.A. Muhsin, 2011. A novel intrusion detection system by using intelligent data mining in weka environment. Procedia Comput. Sci., 3: 1237-1242.

Panda, M. and M.R. Patra, 2007. Network intrusion detection using naive bayes. Int. J. Comput. Sci. Network Secur., 7: 258-263.

Sivatha Sindhu, S.S., S. Geetha and A. Kannan, 2012. Decision tree based light weight intrusion detection using a wrapper approach. Expert Syst. Appl., 39: 129-141.

Wang, G., J. Hao, J. Ma and L. Huang, 2010. A new approach to intrusion detection using artificial neural networks and fuzzy clustering. Expert Syst. Appl., 37: 6225-6232.

Yang, J., Y. Liu, Z. Liu, X. Zhu and X. Zhang, 2011. A new feature selection algorithm based on binomial hypothesis testing for spam filtering. Knowl. Based Syst., 24: 904-914.

Yi, Y., J. Wu and W. Xu, 2011. Incremental SVM based on reserved set for network intrusion detection. Expert Syst. Appl., 38: 7698-7707.