International Journal of Computer Science
and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 4, July 2014, 99-112

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

RESULTS AND COMPARISONS OF DATA MINING TECHNIQUES TO IMPROVE SOFTWARE RELIABILITY

NADHEM SULTAN ALI EBRAHIM & V. P PAWAR

SRT Marathwada University, Nanded, Maharashtra, India

ABSTRACT

The aim of our study is to improve the reliability of the software, in this study we have implemented some of the data mining techniques to improve the software reliability they techniques used are the machine learning intelligence those techniques are the neural network and enhanced support vector machine, as the reliability of the software is main important in the system to avoid attacking the systems, we have taken some real data for training the system from KDD then implemented it to train the system and define all the attacks that may occur from any existing software that running in the systems or coming through the network or from the internet, by the use of the windows logs that exist pre-defined in the windows we can read all event occurring in the system as the system has to report the events occurring in the machine, by monitoring the logs and analyzing with the help of the machine learning algorithms of the NW and ESVM we get the results of what are the normal process and the attacks process and which type of attacks. we have been using both of the data mining techniques to define the attack's type usually the results we got was almost same in neural network and ESVM slightly littlemore accurate and efficient while using the ESVM, we have implemented and experimented both of the techniques in dot net software using C# language.

KEYWORDS: Data Mining Techniques, SE Include Generalization, Characterization, Classification, Clustering, Associative Tree, Support Vector Machines

INTRODUCTION

Data mining is the process of extracting useful data or knowledge from a scattered data and it employs various analytic tools to extract patterns and information from large datasets. Today, large numbers of datasets are collected and stored. Human are much better at storing rather extracting knowledge from it, especially the accurate and valuable information needed to create good software. Large datasets are hard to understand, and traditional techniques are infeasible for finding information from those raw data. Data mining helps scientists in hypothesis formation in physics, biology, chemistry, medicine, and engineering. There are few steps of data mining, data integration, data cleaning, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining techniques that can be applied in improving SE include generalization, characterization, classification, clustering, associative tree, decision tree or rule induction, frequent pattern mining.

Software Reliability

Reliability of the software is the ability of a computer program to perform its intended functions and operations in a system's environment, without experiencing failure (system crash).

IEEE 610.12-1990 defines reliability as "The ability of a system or component to perform its required functions under stated conditions for aspecified period of time."

Using these definitions, software reliability is comprised of three activities:-

- Error prevention
- Fault detection and analysing.
- Measurements to maximize reliability.

When it comes to reliability we have to mention some aspects

- **Scalability**

It is defined as how the application or the software scales with increasing of the workload and ability of a system, a network, or a process to continue to function well, when it is changed in size or volume in order to meet a growing need.

- **Durability**

Durability is the time duration of the software to meet its performance requirements.

- **Sustainability**

Capable of being continued with minimal long-term effect on the environment.

- **Stability**

The software being capable to be stable as long as it is still there.

TECHNIQUES USED

We have used two popular strategies for supervised learning and classification, most the research going in those data mining techniques, the first techniques Neural Network and the second is Support vector machine

Neural Network

NN is an information processing paradigm that is inspired by biology nervous system, it is composed of a large number of highly interconnected processing elements called neuron.

We have selected neural network because of its ability to derive meaning from complicated or imprecise data.

In this techniques we have used four modules

Activation Functions

In this module there are four functions

- **Bipolar Sigmoid Function**

It is an activation function the range of its output is [1, -1]

F(x) =2/ (1+exp (-alpha * x))-1

- **Activation of Function Interface**

Used for all the activation functions, which use with neurons to calculate their output as function of weighted sum of their inputs.

- **Sigmoid Activation Function**

  Represented by the expression

  F(x) =1/ (1+alpha (-alpha * x)).

  Output range is [1, 0].

- **Threshold Activation Function**

  This class represented by the expression

  F(x) = 1, if x >= 0, otherwise 0

  Output range [0, 1].

Layers

This class is scattered into three modules

- **Activation Layer**

  The purpose of this layer is to activate the neurons it is used in multi-layer neural network.

- **Distance Layer**

  This layer is for distance neurons, it is single layer of such network

  Elastic net, Kohonen self-organizing map.

- Layer

- This base neural network layer it represents collection of neurons.

Learning

Learning is the main module of the neural network technique

Here are seven learning of neural network we used

- **Back Propagation Learning Algorithm**

  It is used widely for multi-layer neural network training.

- **Delta Rule Learning Algorithm**

Here when the activation Neurons in the neural network is being this method used to train one layer of the neural network.

- **Elastic Network Learning**

  It concern to train the data in the distance network when couples of computers are connected through the network.

- **Supervised Learning**

When the desired output is already known in the learning stages, the system should adopt its internals to produce the correct answer or close to the correct.

- **Unsupervised Learning**

It used when the output still not known in the learning stage, so the system will calculate the output based on the existence samples which was provided in dataset.

- **Perceptron Learning**

The learning here used to train one layer on neural network of the activation

Neurons with the existing of the threshold.

- **Kohonen Self Organizing Map**

This class purpose is allowing to train the distance netowrks.

Networks

In the networks module we have three modules which they considered the activation of the network or the distance network, based on multi-layer neural network with activation function and activation layers.

Neurons

Neurons modules concern about the activation neurons weighted sum of inputs and it adds the threshold values and then applies activation function. And another module is about distance neuron, computes its output as distance between its weighted and inputs.

Support Vector Machine

SVM has been developed in the reverse order to the development of neural networks, it evolved from the sound theory to the experiments and implementation, while the NNs followed more heuristic path, from applications and extensive experimentation to the theory."Wang (2005) The SVM adjusts the degree of nonlinearity automatically during training, it is popular strategy method for supervised machine learning and classification.

Separating maximum margin hyperplane whose position is determined by maximizing its distance from the support vectors is the fundamental feature of SVM.

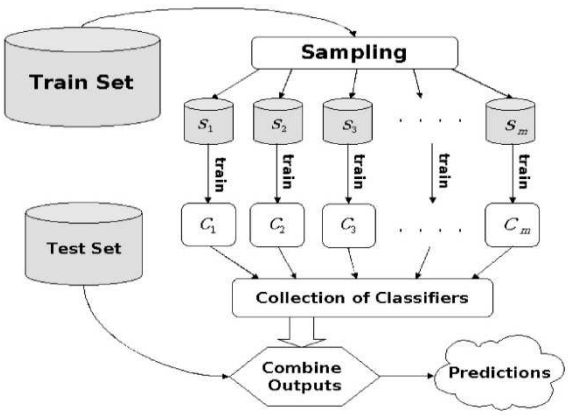

Figure 1

It is useful tool for classification, in versus to neural network each input of training data set has only one classifier output as shown in the figure above, then it combines the output together and test it and take a prediction decision.

In ESVM module we have

- **Learning Module Which Contains**
  - Support vector machine the common interface that used for Support Machine Vector learning algorithms.
  - Sequential Minimal Optimization (SMO) Algorithm.
  - Sequential Minimal Optimization (SMO) Algorithm for Regression.
- Sparse Kernel Support Vector Machine (kSVM).
- Sparse Linear Support Vector Machine (SVM).

We have used two type of testing phase to do the experimental work

The first one is with training data available in the dataset which contain 1000 of records available already in the Microsoft access database which we collected from KDD CUP1999

And the second one is based on live data which we monitor through the network packets with the help of WinPcapto capture the packets coming and going to the network.

Using both techniques we got nice results for both, ESVM is slightly more accuracy and efficiency compare to neural network.

Sequence Diagram

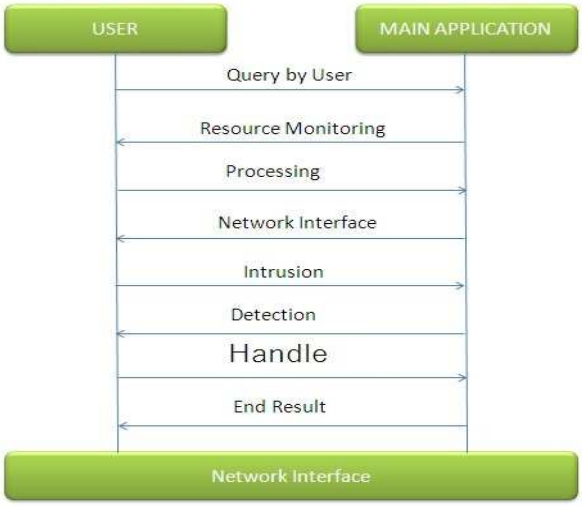

Figure 2

In the above sequence diagram it is explained briefly how the scenario of capturing errors of the software and intrusion that exist in the system.

Starting from the query by the user in front of the monitor who give order by processing application then to the network interface there the detection can be happened and handle

In the experimental work three protocols were used TCP, ICMP and UDP.

Types of Attacks

Normal, Multihop, Snmpgetattack, Snmpguess, Teardrop, Udpstorm, Apache2, Buffer_overflow

ftp_write, guess_passwd, httptunnel, imap, land, loadmodule, mailbomb, mscan, named, Neptune

Xlock, Xsnoop, Xterm, Nmap, Perl, Phf, Portsweep, Processtable, Ps, Rootkit, Saint, Satan, Sendmail

Sqlattack, Warezmaster, Worm, Ipsweep, Pod, Smurf.

Block Diagram

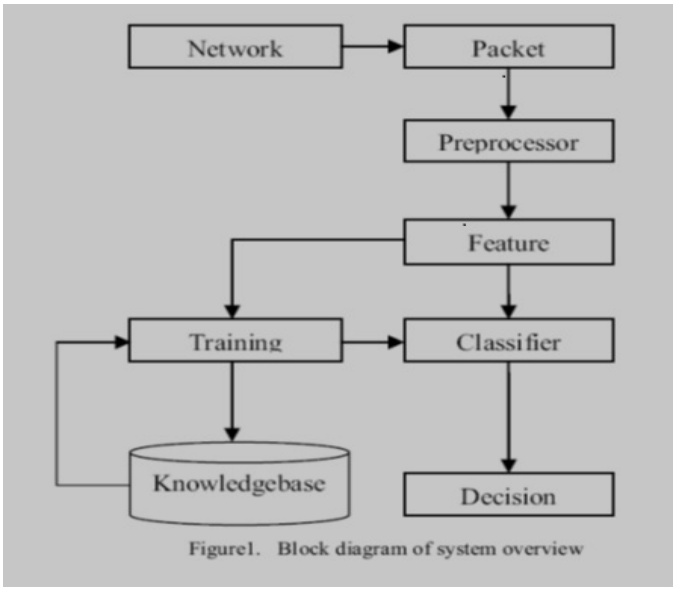

Figure 3

Briefly in the block diagram explained how the process of the proposed system works

In the network layer many packets goes in and out in the system, we preprocess the packets using the technique proposed system and identify its feature and is it normal packet or intrusion packet using the training and the knowledge base database and finally take decision to allow or to block it.

We have used significant techniques for improving the software reliability and both the techniques are usable nowadays and many of the research is done on them.

The entire experimental work was done using Microsoft visual studio 2010 in C#.

Two Techniques

The Process of Experimental Work

- Load the training dataset.

In the training dataset we got the data already from KDD1999 we organized it to be used in the project for the experimental work.

- Set the rules for each selected dataset.

- Identify the classifier.

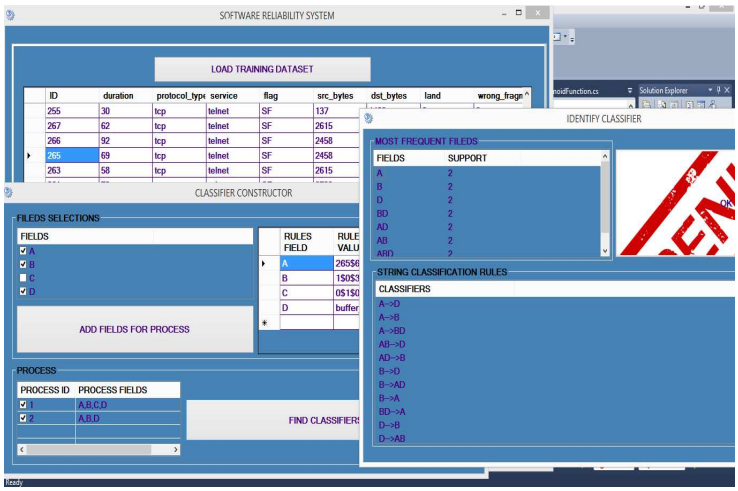

Figure 4

- Find the classifiers after completing the uploading the training dataset we can see all the classifiers we have set rules for.
- Testing phase in this step there are two choices to generate test data.
  - Ready data which already exists in the database.
  - Live data where to monitor the data coming through the network.

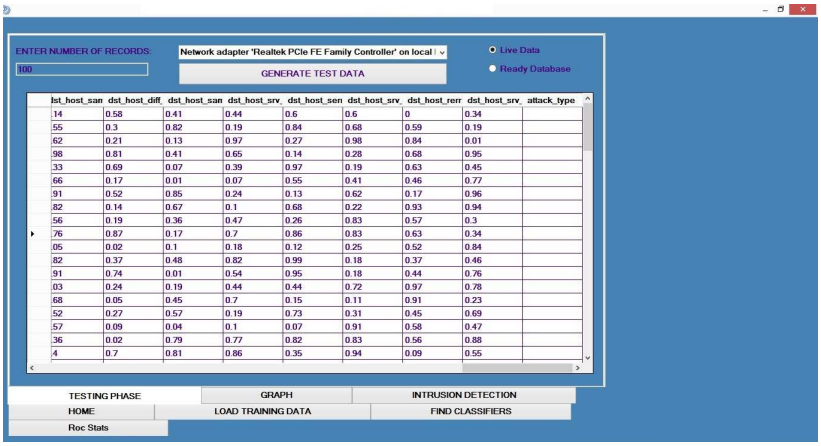

Figure 5

- Attack detection, in this phase there are the two techniques we used

Neural Network

Support Vector Machine

- Find the ROC (Receiver Operating Characteristics), after the attack detection in either technique we can check the ROC of the technique.
- Draw the graph to represent the attack types.

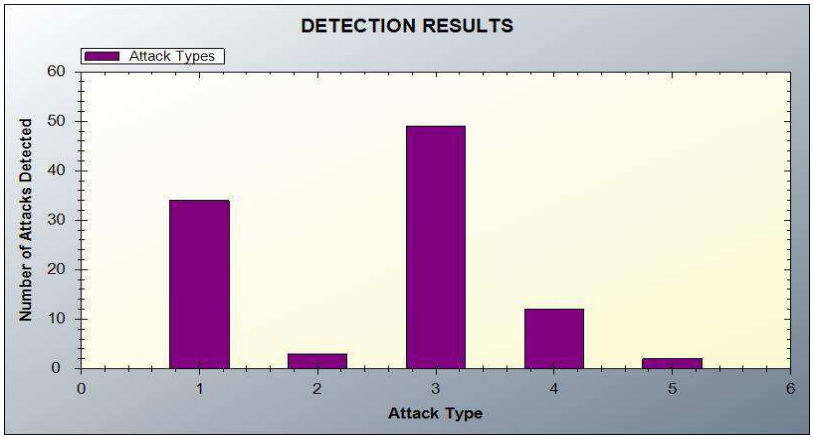

Figure 6

Detectin the attack comes positives when we use ready data but only one plot we get when we use live data as the attack type is normal and it depends on the system whether it is free of intrusion or no.

Detecting the attack

First step is to train the system by uploading training dataset.

Test Cases

Table 1

Test number	Test Tittle	Expected Results	Results
Start the application	**Login Form**	Authentication Successful	**Pass**
Load the KDD data set	**Load Training Data**	Load successfully all data set	**Pass**
Testing Phase	Generate Test Data	Loading data should be successful	**Pass**
	Select Live data		
	Ready data set		
Intrusion Detection	Load the data set & Generate test data	Intrusion detection Examine	**Pass**
Implementation of Algorithm	Neural Network	Implementation occurs	**Pass**
	Enhanced Support Vector Machine		

Before explaining about the techniques that were implemented would like to use those shortcut:

TP → True Positive
FP → False Positive
TN → True Negative
FN → False Negative
ROC → Receiver Operating Characteristics

- Neural Network experimental results :-

The first test stage was on training data we analyzed from KDD

Table 2

100 Records Results Using Neural Network	
TP=35	FP=4
FN=8	TN=53

Table 3

Roc Calculation Using Neural Network of 100 Records	
Sensitivity= 0.8242	Specificity= 0.922
Efficiency= 0.873	Accuracy= 0.8799

The second stage was on live data in cyber system

Table 4

100 Live Records Results Using Neural network	
TP=38	FP=4
FN=8	TN=50

Table 5

Roc Calculation Using Neural Network of 100 LIVE Records	
Sensitivity= 0.8202	Specificity= 0.921
Efficiency= 0.871	Accuracy= 0.875

Enhanced Support Vector Machine

In the ESVM technique the Experimental results we got using ready data is:-

Table 6

100 LIVE Records Results Using Support Vector Machine(ESVM)	
TP=47	FP=1
FN=3	TN=49

Table 7

Roc calculation Using ESVM of 100 LIVE Records	
Sensitivity= 0.9418	Specificity= 0.9706
Efficiency= 0.9562	Accuracy= 0.9562

And Using the Live data

Table 8

100 LIVE Records Results Using Neural Network	
TP=48	FP=2
FN=3	TN=47

Table 9

Roc Calculation Using Neural Network of 100 LIVE Records	
Sensitivity= 0.9471	Specificity= 0.9679
Efficiency= 0.9572	Accuracy= 0.957

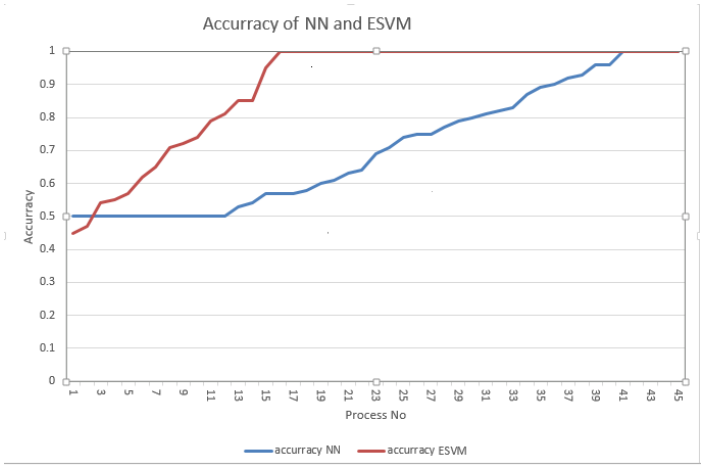

Figure 7

Using ESVM the accuracy better than the Neural network

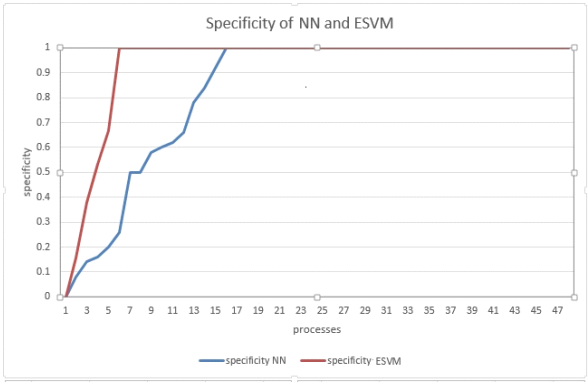

Figure 8

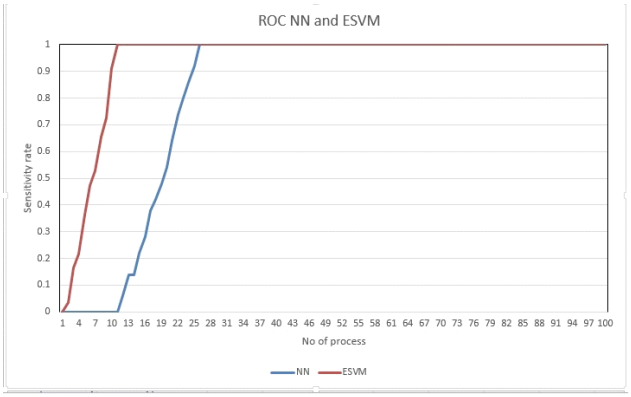

Figure 9

Here in the above screen shot shows the ROC of both neural network results and ESVM results

The ESVM ROC is higher than the ROC of the Neural network.

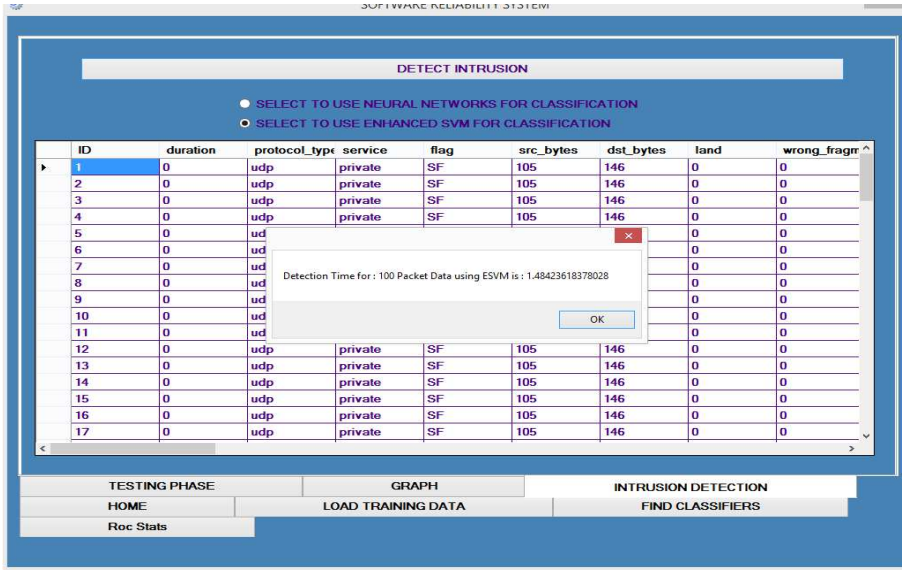

Figure 10

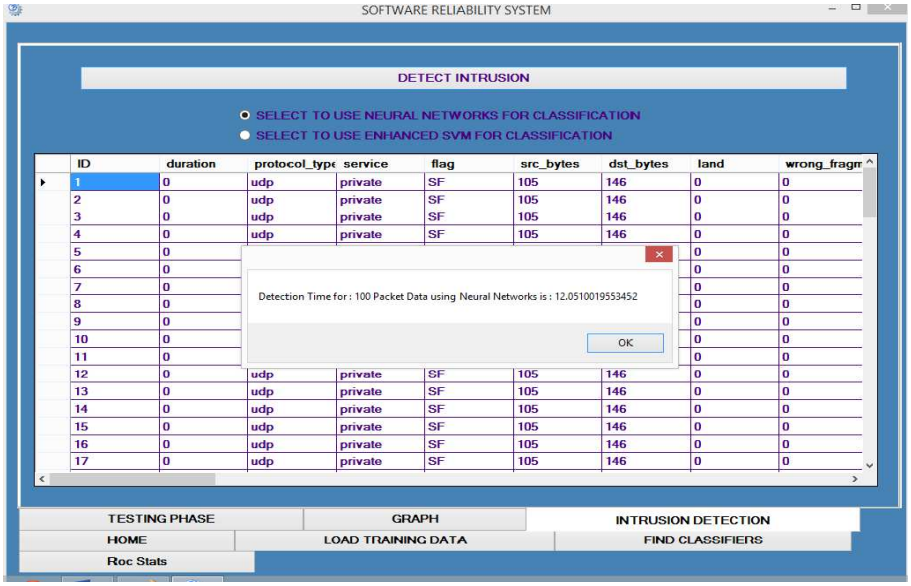

Figure 11

Table 10

Detecting Time in Seconds Using ESVM and NN	
ESVM	1.484
NN	12.051

CONCLUSIONS

Attacks are generated to servers and log files are created for every 30 minutes, dataset are created with different types of attacks and log files are exists in the system form both we can identify types of the attack from the behavior of the software and monitoring the ports and the duration of each connection,

We have used supervised learning algorithms of the neural network and ESVM to detect the error after training the system on the existence of the dataset derived from KDD or generate the live dataset on a live connections in the

network. Using the created dataset the derived form KDD or live will evaluate the performance of the machine learning algorithms.

REFERENCES

1. N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.
2. N.L. Johnson, S. Kotz, Distributions in Statistics: Continuous Univariate Distributions, vol. 1–2, Wiley, New York, 1970.
3. C. Saunders, A. Gammerman. V. Vovk, Ridge regression learning algorithm in dual variables, Proceedings of the 15th International Conference on Machine Learning CML'98), MorganKaufmann, 1998, pp. 515–521.
4. Joong-Hee Leet, Jong-Hyouk Leet, Seon-Gyoung Sohn, Jong-Ho Ryu, "Effective Value of decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System", ICACT 2008, pp:1170-1175.
5. Jie Yu and Zhoujun Li, "A Detection and Offense Mechanism to Defend against Application Layer DDoS Attacks" *IEEE Third International Conference on Networking and Services*, pp.54 – 54, 2007.
6. Yoohwan Kim, Wing Cheong Lau, MooiChooChuah and Jonatan Chao"Packet Score: A Statistics-Based Packet Filtering Scheme againstDistributed Denial-of-Service Attacks", IEEE Trans. On dependable and. secure computing, Vol. 3, No. 2, PP. 2594-2604, 2006.
7. Yi Xie, and Shun-Zheng YU, "A Large-Scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors",IEEE/*ACM Trans. on networking*, vol. 17, No.1, Pp. 54-65, 2009.
8. J.A.K. Suykens, L. Lukas, J. Vandewalle, Sparse least squares support vector machine classiEers, European Symposium on ArtiEcial Neural Networks (ESANN 2000), Bruges Belgium, April 2000, pp. 37–42.
9. T. Van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, J. Vandewalle, Benchmarking least squares support vector machine classiEers, Internal Report00-37, ESAT-SISTA, K. U. Leuven.
10. Giuseppe Ateniese, Chris Riley,ChristianScheideler, Computer Society,1242 IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 5, NO. 9, SEPTEMBER 2006 Survivable Monitoring in Dynamic Networks.
11. Robert Mitchell and Ing-Ray Chen, IEEE TRANSACTIONS ON RELIABILITY, VOL. 62, NO. 1, MARCH 2013 Effect of Intrusion Detection and Response on Reliability of Cyber Physical Systems.
12. M. Anand, E. Cronin, M. Sherr, M. Blaze, Z. Ives, and I. Lee, "Security challenges in next generation cyber physical systems," in *Beyond SCADA: Networked Embedded Control for Cyber Physical Systems*. Pittsburgh, PA, USA: NSF TRUST Science and Technology Center, Nov. 2006.
13. F. B. Bastani, I. R. Chen, and T. W. Tsao, "Reliability of systems with fuzzy-failure criterion," in *Proc. Annu. Rel. Maintainability Symp.*,Anaheim, California, USA, January 1994, pp. 442–448.
14. A. Cárdenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Proc. 1st Workshop Cyber-Phys. Syst.Security*, Virginia, USA, Jul. 2009.

15. I.R. Chen, F. B. Bastani, and T. W. Tsao, "On the reliability of AI planning software in real-time applications," *IEEE Trans. Knowledge Data Eng.*, vol. 7, no. 1, pp. 4–13, 1995.

16. P. Oman and M. Phillips, "Intrusion detection and event monitoring in SCADA networks," in *Critical Infrastructure Protection*, E.Goetz and S. Shenoi, Eds. Boston: Springer, 2007, vol. 253, International Federation for Information Processing, pp. 161–173.

17. C.-H. Tsang and S. Kwong, "Multi-agent intrusion detection system in industrial network using ant colony clustering approach and unsupervised feature extraction," in *Proc. IEEE Int. Conf. Ind. Technol.*, Hong Kong, Dec. 2005, pp. 51–56.

18. J. H. Cho, I. R. Chen, and P. G. Feng, "Effect of intrusion detection on reliability of mission-oriented mobile group systems in mobile ad hoc networks," *IEEE Trans. Reliability*, vol. 59, no. 1, pp. 231–241, Mar.2010.

19. Usha Banerjee, Gaurav Batra, K. V. Arya Feedback Reliability Ratio of an Intrusion detection System, *Journal of Information Security*, 2012, 3, 238-244 doi:10.4236/jis.2012.33030 Published Online July 2012 (http://www.SciRP.org/journal/jis)

20. *Journal of Information Security*, 2012, 3, 238-244 doi:10.4236/jis.2012.33030 Published Online July 2012 (http://www.SciRP.org/journal/jis).

21. R. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. H. Webster, D. Wyschograd, R. K. Cunningham and M. A. Zissman, "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation," IEEE Computer Society Press, vol. 2, 2000, pp.12-26.