Proposed Network Intrusion Detection System In Cloud Environment Based on Back Propagation Neural Network

Shawq Malik Mehibs

Computer science department, University of Technology. Baghdad-Iraq.

shouq90@ymail.com

Soukaena Hassan Hashim

Computer science department, University of Technology. Baghdad-Iraq.

Soukaena.hassan@yahoo.com

Abstract

Cloud computing is distributed architecture, providing computing facilities and storage resource as a service over the internet. This low-cost service fulfills the basic requirements of users. Because of the open nature and services introduced by cloud computing intruders impersonate legitimate users and misuse cloud resource and services. To detect intruders and suspicious activities in and around the cloud computing environment, intrusion detection system used to discover the illegitimate users and suspicious action by monitors different user activities on the network .this work proposed based back propagation artificial neural network to construct t network intrusion detection in the cloud environment. The proposed module evaluated with kdd99 dataset the experimental results shows promising approach to detect attack with high detection rate and low false alarm rate.

Keywords: cloud computing environment, intrusion detection, back propagation algorithm, machine learning.

الخلاصة

الحوسبة السحابية هي هيكيلة موزعة توفر قدرات حسابية, موارد تحزين كخدمة عبر الانترنت للأيفاء بمتطلبات المستخدم بسعر منخفض .بسبب طبيعة الحوسبة السحابية المفتوحة والخدمة المقدمة المتسللين ينتحلون المستخدمين المخولين وبعد ذلك يسيئون استخدام موارد وخدمات الحوسبة السحابية . لكشف المتسللين والانشطة المشبوة في بيئة الحوسبة السحابية ،نظام كشف التطفل يستخدم لكشف المستخدمين الغير مخولين والانشطة المشبوهة بواسطة فحص نشاطات المستخدم على الشبكة .في هذا البحث استخدمت خوارزمية الشبكات العصبية الاصطناعية (BP) لبناء نظام كشف تطفل في بيئة السحابية . النظام المقترح اختبر باستخدام بيانات KDD99. اظهرت النتائج ان النظام المقترح يشكل طريقة واعدة تتميز بدقة عالية مع نسبة انذار كاذبة منخفضة.

الكلمات المفتاحية :بيئة الحوسبة السحابية ،كشف التطفل،خوارزمية الانتشار العكسي،تعلم الآلة.

1. Introduction

In the recent years cloud computing used in most organizations, due to its Characteristics like high scalability, high flexibility and low operational cost. Cloud computing supplies the customer with essential requirements as a service, where Service providers of could environment provided Infrastructure, Platform, Software as a service (Muthukumar and Rajendran, 2015). The customers can access and use the service provided by cloud computing via internet, as a result, leads to internet intrusion. Cloud computing revolve around user, as such the customers must be ensure about their data and applications storage in cloud server. Intrusion can be defined as unauthorized access to the system and tries to adversely affect the confidentiality, availability or integrity of a resource. To detect intruders' intrusion detection systems (IDSs) are used, they compare the observed events with suspicious patterns. According to the technique used for detection, there are two types of IDSs signature-based and anomaly-based. Anomaly intrusion detection can be classified into host-based and network-based, depending on the target which monitoring (Fung and Boutaba, 2014). Network intrusions detection system (NIDS) monitoring traffic of the network and system activities for malicious activity or policy violations. With increment the service provided over network and sensitive information transmitted via the Internet, this leads to the increase in number and type of attack, for that network intrusion Detection System (NIDS) had become important part in network security, which represent protection layer that is used to pre-defined malicious activity or suspicious pattern by monitored the network traffic, as result for that the administrative capacity of the system administrator's security will be reinforced ,and operational performance of the system will be optimized .Intrusion detection system represent a protection layer after the firewall. In recent years data mining techniques are used for intrusion detection in wide range because the automation of intrusion detection .the pattern of intrusion and normal behavior can be computed using data mining . Intrusion Detection mechanisms (IDS) based on data mining technique are not only automated but are also provided for a significantly elevated accuracy and efficiency. It can help in revealing of new intrusions and policy violations, promoting decision support for intrusion management, and discovering behavior patterns of attackers that unknown previously. One of the data mining technique that has successful in solving complex practical problems neural network. Artificial neural networks have the ability to solve several problems confronted by the other present technique used in intrusion detection. There are three advantage of intrusion detection based on neural network (Yassin *et.al., 2012*; Wu and Banzhaf, 2010).

- Neural network provides elasticity in intrusion detection process, where the neural network has the ability to analyze and ensure that data right or partially right. Likewise, neural network is capable of performing analysis on data in nonlinear fashion
- Neural network has the ability to process data from a number of sources in a non-linear fashion .this is very important especially when coordinated attack by multiple attackers is conducted against the network.
- neural network is characterized by high speed in processing data

In This work proposes a network intrusion which depends on back propagation (BP) neural network. Back propagation is multilayer supervised learning neural network algorithm. Back propagation as machine learning algorithm has a learning process is composed the two stages ,the learning stage , where the network is learned by modification of the weights ,in the testing stage the weight from learning stage is used to forecast the class label of the new input patterns. This work proposes an algorithm for intrusion detection capable to classifying of the normal behavior and the four types of attack (DOS, Probe, U2R, and R2L). The proposed algorithm is evaluated with KDD99 dataset. The rest of paper is organized as the following: in section 2 we present the previous related work, in section 3 we explain the proposed system, in section 4 the KDD99 dataset is included and described, in section 5 the preprocessing of datasets is presented, in section 6 neural network and back propagation algorithm are illustrated in details, section 7 evaluation performance is illustrated, in section 8 the experiments and results and finally the conclusion.

2. Related work

Tian and Liu (2010) developed new approach for intrusion detection where the neural network is used to train the module, and to enhance the ANN particle swarm optimization algorithm used to optimize the parameters of ANN, and to increase the performance of proposed algorithm . This approach used rough set as a feature selection approach to reduce the dimension of dataset.

Pandeeswari and Kumar (2015) proposed anomaly intrusion detection system in cloud environment at virtual machine monitor (VMM) layer called hypervisor detector. The Proposed system is designed by integration of fuzzy c mean (FCM) and artificial neural network(ANN) .This system involves three stage, in the first stage the fuzzy c mean algorithm is used to generate clusters of the same class to improve the leaning ability of ANN, in the second phase the generated clusters are used as input to train various ANN algorithm, in the third stage the error of various ANN is reduced by using fuzzy aggregation module which integrates the result of different ANNs.

Xiangmei and Zhi (2011) depended on advantage of improved GA of global search ability and fast local searching of the levenberg-marquard back propagation algorithm. The author proposed Hybrid neural network which consist of sub classifiers .every sub classifiers is neural network classifiers used to classify one type of attack where the dataset for the training and testing is divided into four subsets, namely, DOS training sample sub-set, Probe training sample subset, R2L training sample subset and U2R training sample subset. The fusion rule is used to combine the result from sub classifiers

Guangjun *et.al*., 2008 developed intrusion detection system using back propagation neural network .To improve the detection efficiency the learning rate change dynamics with the error of network according to specific formula. Then the BP neural network combines with simulated annealing algorithm to increase the detection accuracy.

Chen and Qian ,2009 developed hybrid approach for intrusion detection system depending on particle swarm optimization algorithm (PSO) and radial basis function (RBF) network. In the proposed PSO-RBF neural networks the Particle swarm optimization algorithm (PSO) is used to optimize the parameter of Radial basis function (RBF) neural network ,then the author compared the result of proposed approach with conventional RBF neural network .The experiment revealed that the new approach is better than the conventional RBF neural network.

3. The proposed system

In the proposed system the KDD99 dataset is used to generate the training and testing samples .Then after back propagation parameters determine the algorithm used to train network to be used later in intrusion detection .The proposed system consisted of following steps:

1. Selecting subsets of samples for training and test phases from KDD99 dataset.
2. Preprocessing of the selected subsets of samples.

3. Training back propagation neural network algorithm using the samples of training set.
4. Test the BPNN using the testing samples.
5. Calculating the performance of trained module.

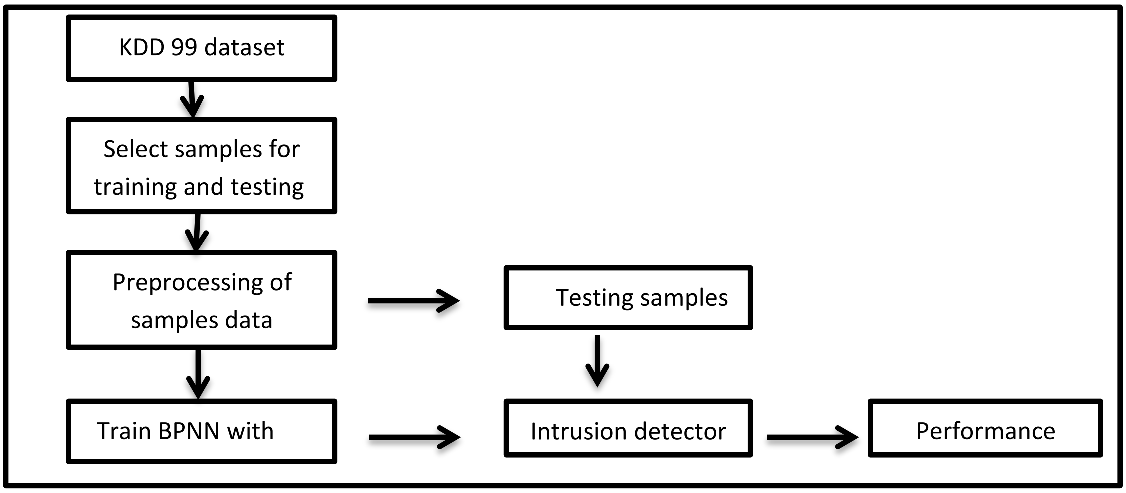

Figure 1.the Block diagram of the proposed system.

4. KDD Cup 99 Dataset

The KDD cup 99 dataset was benchmark dataset used in data Mining tools competition and international knowledge discovery that is selected from DARPA98 network traffic dataset in 1999 by collecting single TCP dump into TCP connections. The KDD cup 99 dataset is considered the most a favorable and developed standard dataset used in research of evaluating algorithms used in intrusion detection .In KDD 99 dataset each TCP connection consists of 41 attribute with label which determine the status of connection either being normal or specific attack type. The attack types include four categorized:

- Denial of Service (DOS) attacks: In this type of attack, attackers attempt to prohibit authorized users from using service by consuming the resources of network.
- Probe attack: Attacker gathers information about the target host by scanning and polling activities for future attack
- Remote-to-Local (R2L) attack: Attacker attempts to access machine which is not authorized to access to that machine.
- User-to-Root (U2R) attack: Attacker has already local access to the system and attempt to get non-authorized access to gain super-user privileges.

In KDD 99 dataset consist of 41 feature divided into numeric features (38) and symbolic features (3) are classified into the following classes:

- Basic features: These taken from packet headers and they describe each single TCP connection and the number of them is equal to 9.

- Content features: These feature are related to domain knowledge and used to refer to suspicious behavior where network traffic has no sequential patterns. These feature are equal to13.
- Time-based traffic features: These features are calculated according to window interval by capturing properties of the connections in the past 2 second that have the same service as the current connection .These feature are equal to 9.
- Host-based traffic features: These features calculated using historical window of 100 connections that have the same destination host. These features used to assess attack take time longer 2 second .These feature equal to 10.

The KDD cup 99 datasets consist of training and test set .There are 4,940,000 data sample in the training set , these samples are distributed between normal behavior and 24 attacks. On the other hand, there are 311,029 data samples include normal network traffic and 38 types of attack,24 attack existed in training set besides 14 new attack . As the training set contain large number of data samples, other training set formed include 10% of data samples used in wide range (Wu and Banzhaf,2010).

Table (1): number of samples in KDD cup 99

dataset	normal	Dos	Probe	U2R	R2L	Total
corrected KDD 99"	60593	229853	4166	70	1126	311029
"10% KDD	97277	391458	4107	52	1126	494020

5. Data preprocessing

The KDD cup 99 contains numeric feature in addition to symbolic features for example protocol type (tcp,udp,icmp) ,back propagation algorithm accepted numerical data. Therefore, the symbolic feature is transformed to numeric value of sequential integer value. Data normalization improved the efficiency and the accuracy of mining algorithms including ANN's .Where these algorithms provide better result when the data to be analyzed fall between [0, 1]. Min-Max normalization method which is a linear transformation is used to scale data between [0, 1] .The following formula is used to find the new value:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

Table (2) numeric value of symbolic feature of KDD cup99 dataset

Protocol type	Feature value	Service	Feature value	Service	Feature value	Service	Feature value	Flag	Feature value
Tcp	1	Private	1	time	23	shell	45	SF	1
Udp	2	Smtp	2	mtp	24	Efs	46	SH	2
icmp	3	http	3	gopher	25	login	47	S0	3
		ftp_data	4	rje	26	printer	48	S1	4
		IRC	5	link	27	netbios_ssn	49	S2	5
		telnet	6	Ctf	28	csnet_ns	50	S3	6
		Domain	7	Hostnames	29	nntp	51	RSTR	7
		Finger	8	iso_tsap	30	supdup	52	REJ	8
		Other	9	pop_2	31	http_443	53	RSTO	9
		ftp	10	netbios_dgm	32	uucp_path	54	RSTOSO	10
		Imap4	11	netbios_ns	33	domain_u	55	OTH	11
		pop_3	12	sql_net	34	ntp_u	56		
		Sunrpc	13	bgp	35	ecr_i	57		
		pm_dump	14	vmnet	36	eco_i	58		
		Echo	15	Z39_50	37	tim_i	59		
		Discard	16	ldap	38	urp_i	60		
		Systat	17	nnsp	39	red_i	61		
		Daytime	18	kshell	40	Remote_job	62		
		Netstat	19	klogin	41	X11	63		
		Ssh	20	uucp	42	http_8001	64		
		Name	21	courier	43	urh_i	65		
		whois	22	exec	44				

6. Artificial Neural Network

An artificial neural network (ANN) is a computational model inspired from biological nervous systems. The artificial neural network is characterized by a set of node (neurons) and connection between nodes (weights). The neurons are arranged in layers and connected through weights, each neuron represents a computational unit that receives input to process it to get the output. The connections set the direction of information outflow, which can be unidirectional or bidirectional. The process of ANNs learning is based on examples. The ANNs have the ability to overcome incomplete, noisy and limited data. Therefore, it was successfully implemented in a wide range of data intensive applications. In the neural network, the method is used to adjust the weights and the way that nodes are connected determines the type if neural network (Wu and Banzhaf, 2010). In our work, the back propagation multilayer feed-forward neural networks algorithm is used to solve a multi-class problem of intrusion detection to predict intrusion. Multilayer layer neural network topology means that network consist of three type's layers each of them includes a collection of nodes. The first type is the input layer which receives value in accordance to samples of issue. The second type is the output layer which represents the outcome of the network. The third is the type hidden layer which used to learn the feature of the input.

6.1 Back Propagation Neural Network

Back propagation algorithm is a supervised learning technique which means the algorithm training with samples of input and output that network should calculate. Back Propagation (BP) is multilayer feed forward neural network contains one input layer, one or more hidden layer and one output layer, neurons are arranged in layers. The learning course of back propagation consists of two phases: the forward phase where the input is presented and propagation it towards the output layer: the backward phase where the error is computed and the weight is adjusted to reduce the error so that the ANN learn the data. In the forward each neurons in the input layer has input multiplied by weights between the input and the hidden layers. Each hidden neurons (j) in hidden layer receives value Zj(j) using equation:

$$z_j = \theta + \sum_{i=1}^{n} x_i \; w_{ij} \qquad (2)$$

The binary sigmoid function actviation function used to process the output of the hidden layer using equation:

$$f(x) = \frac{1}{1+\exp^{-x}} \qquad (3)$$

$$z_j = f(z_j)$$

The output of hidden layer broadcasts to the nodes in the output layer as equation:

$$y_k = \theta + \sum_{j=1}^{p} z_j \; w_{jk} \qquad (4)$$

$$y_k = f(y_k)$$

Where θ is the basis between layers.

the mean square error value E is considered the quantative measure of learning which reflects the degree of learning. Generally, an MSE under (0.1) indicates that the net learned its training set .the mean square error value E is calcualted according to the following equation :

$$E = \frac{1}{2} \sum_p \sum_k \left(T_{pk} - Y_{pk}\right)^2 \qquad (5)$$

p – Represent the number of samples.

k – Represent the number of output unit.

Y_{pk} – the value of actual output.

T_{pk} – the value of target output.

In the backward phase if the output of network is different from target output then the output error will be calculated, the error then propagated towards the input layer to update the weight between neurons in layers. The error between the output and hidden layer can be calculated using the following equation:

$$\delta_{2k} = y_k(1 - y_k)(T_k - y_k) \qquad (6)$$

The error between hidden and input layer can be computed using equation:

$$\delta_{1j} = z_j\left(1 - z_j\right)\sum_{k=1}^{m} \delta_{2k}\, w_{jk} \qquad (7)$$

Then the weights are updated to reduce the error according the following equations:

$$w_{jk}(\text{new}) = \eta * \delta_{2k} * z_j + \propto * w_{jk}(\text{old}) \qquad (8)$$
$$w_{ij}(\text{new}) = \eta * \delta_{1k} * x_i + \propto * w_{ij}(\text{old}) \qquad (9)$$

w_{jk} – the weights between the output and hidden layer.

w_{ij} – the weights between the hidden and input layer.

η – Learing rate.

$\propto$ – Momentum cofficient

The goal of backward phase is to find the global optimum of network weights and reduce the gradient error. The learning course is achieved by minimizing the mean absolute error valueE_m.

6.2 The proposed algorithm

Back propagation for intrusion detection (BP-ID) module consist of two stages, the training stage and the test stage . Algorithm (1) shows the training stage of the proposed module and algorithm (2) shows the testing stage of the proposed intrusion detection module.

```
Algorithm(1) : traing stage of proposed (BP-ID) module

Input: number of samples selected form KDD 99 ,Minerror.
Output: Vector of wights $w_{ij}$ between input layer and hidden layer,vector of weight
        $w_{jk}$ between hidden layer and output layer.
Steps:
Initialize the weights $w_{ij}$ and $w_{jk}$ between neurons in the layers (set to small random value).
Compute the desired ouput T from labeled training samples.
Each feature in training sample represents input unit ($X_i$) then broadcasts to all neurons in hidden layer.
Compute the output of each unit in the hidden layers ($z_j$) using equation (2) then the output of hidden layer is processed using equation (3).
Compute the output of each unit in the output layer (Yk) using equation (4) then the output is processed using equation (3).
Compute the error value for each unit in the output layer using the equation (6).
Compute the error value for each unit in the hidden layer using the equation(7)
update the weights $w_{ij}$ and $w_{jk}$ between layers according to equations (8)(9):
Calculate the absolute error value $E_m$ using equation (5).
Repeat until $E_m$ <=Minerror
```

```
Algorithm (2):testing stage of proposed (BP-ID) module.

Input: number of testing samples selected from KDD99, vector of wights $w_{ij}$
       between input layer and hidden layer,vector of weight $w_{jk}$ between hidden
       layer and output layer.
Output: Vector of ouput value match or very closer to target value.
Steps:
The features of training sample represent input unit ($X_i$) then they broadcast to all neurons in hidden layer.
Compute the output of each unit in the hidden layers ($z_j$) using equation (2) then the output of hidden layer is processed using equation (3).
Compute the output of each unit in the output layer (Yk) using equation (4) then the output is processed using equation (3).
```

7. Performance Evaluation

The measure efficiency of IDS depends on its ability to make the right detection depending on the nature of the given status compared with the result of intrusion detection system (IDS). Four possible results can be obtained and they are called confusion matrix described in table (3). The four outcomes are true negative (TN) which indicates the correct prediction of normal behavior, true positive (TP) which indicates the correct predication of attack behavior, false positive (FP) which indicates the wrong predication of normal behavior as attack, false negative (FN) which indicates mistake predication of attack behavior as normal. Both (TN) and (TP) are considered guide of the correct operation of the IDS. In addition, both (FP) and (FN) reduce the effectiveness of IDS. Therefore (FP) and (FN) should be minimized to increase the efficiency of IDS system. To evaluate the proposed model for intrusion detection and its ability to detect intrusion. The following measure is used to detect the performance of the IDSs.

Accuracy (ACC): it measures performance represent the rate of samples which are properly detected as normal or attack to the overall number of samples and is calculated using the equation:

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (10)$$

Detection Rate (DR) :It measures performance which indicates the ratio of the number of samples that are correctly classified as attack to the total number of attack samples and is calculated using equation :

$$\text{DR} = \frac{\text{TP}}{\text{TP+FN}} \qquad (12)$$

False Alarm Rate (FAR): It measures performance which represents the rate of samples which is improperly categorized as attack to the overall number of samples of normal behavior and is calculated using equation:

$$\text{FAR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \qquad (13)$$

Table (3): confusion matrix

Actual class	Predicted Class	
	Negative class(normal)	Positive class(attack)
normal	True negative (TN)	False positive (FP)
attack	False negative (FN)	True positive (TP)

8. Experiments and Results

The proposed algorithm is evaluated with KDD99 dataset. The proposed algorithm is trained with samples selected from KDD 99 dataset includes normal behavior samples besides the other four types of attack (Dos, Probe, U2R, R2l) to specify normal samples from attack samples and also to detect the type of attack. Three evaluation criteria used to assess the proposed module. To check the efficiency of the proposed module three experiments are conducted, in the first experiments the algorithm is tested with (500) records containing normal behavior in addition n to four attack types. The results obtained show high detection rate of the module up to (0.99) and low false alarm rate (0.03). The second and third experiments are conducted with (1000) (1500) records respectively and they also include four types of attack. The results of the experiments remain in same range as shown in table (4). The topology of network consist of three layer (input layer, hidden layer, output layer), the number of unit in input layer is equal to 41 which is the number of feature in KDD99 dataset. The number of neurons in hidden layer is equal to 20 which is the number of hidden unit determined after trial and error. In the output layer the number of neuron is equal to 5 which represents the normal behavior and the four types of attack. Back propagation parameters are shown in table (5).

Table (4): Experimental result of the proposed algorithm

Number of experimental	Number of samples	DR	Accuracy	FAR
Experimental 1	500	0.99	0.97	0.03
Experimental 2	1000	0.99	0.98	0.02
Experimental 3	1500	0.99	0.97	0.9

Table (5): Parameters of back propagation algorithm

Parameters name	Value parameters
basis	1
Learing rate	1
Momentum cofficient	1
No.of unit in input layer	41
No.of unit in hidden layer	20
No.of unit in output layer	5
Mean square error	0.001
Maximum number of iteration	20

9. Conclusion

In this work back propagation neural network is proposed for intrusion detection in cloud environment .The proposed work contain two stages; in the first stage we train the model with back propagation algorithm using KDD 99 dataset to classify normal behavior and the other four types; in the second stage the trained module is tested with three datasets to detect normal behavior and four types of attack. The purpose of this work is to implements intrusion detection system characterized by high detection rate and low false alarm in cloud environment. Where, the most important aspect of intrusion detection system in cloud environment is false positive alarm rate where the high positive false alarm lead to consuming time which affects the quality of the service provided to customer. The proposed module is evaluated using different evaluation criteria .The experimental result proves effectiveness of the proposed algorithm characterized by high detection rate and low false alarm.

Reference

Carol Fung, Raouf Boutaba, 2014,**"intrusion detection networks:A Key to Collaborative Security",**Taylor & Francis Group.

Li Xiangmei, Qin Zhi, 2011**, "The Application of Hybrid Neural Network Algorithms in Intrusion Detection System"**, IEEE, ISBN:978-1-4244-8691-5,pp.1-4.

Muthukumar B., Praveen Kumar Rajendran, 2015**," security in computing and communications"**, springer, volume 536, pp. 54–65.

Pandeeswari N., Ganesh Kumar, 2016**,"anomaly detection System in cloud environment using fuzzy clustering based ANN"**, Mobile Networks and Applications, Volume 21,issue 3, pp. 494-505.

Shelly Xiaonan Wu, Wolfgang Banzhaf, 2010**,"The use of computational intelligence in intrusion detection systems: A review"**, Applied Soft Computing, Vol. 10, pp.1-35.

Song Guangjun, Zhang Jialin, Sun Zhenlong,2008**," The Research of Dynamic Change Learning Rate Strategy in BP Neural Network and Application in Network Intrusion Detection"**, IEEE, ISBN:978-0-7695-3161-8,pp.513.

WenJie Tian, JiCheng Liu,2010**," A New Network Intrusion Detection Identification Model Research"**, IEEE, ISBN:978-1-4244-5192-0,pp.9-12**.**

Yassin W., Udzir N.I., Muda Z., Abdullah A. and Abdullah M.T., 2012,**"A Cloud-Based Intrusion Detection Service Framework ",** IEEE, **ISBN:**978-1-4673-1425-1,pp. 213 – 218**.**

Zhifeng Chen, Peide Qian, 2009, **"Application of PSO-RBF Neural Network in Network Intrusion Detection"**, IEEE, ISBN:978-0-7695-3859-4,pp. 362 – 364.