Wrapper based Intrusion Detection System with Duration and Local Area Network Denial Features

Amol A. Dhiwar
SSBT's College of Engineering and Technology,
Bambhori, Jalgaon
Maharashtra (India)

Girish K. Patnaik, PhD
SSBT's College of Engineering and Technology,
Bambhori, Jalgaon
Maharashtra (India)

ABSTRACT
The use of internet has become a popular way of getting connected with each other, as a result networking attacks get increased. The networking attacks are termed as intrusions that are based on the values of features. Features based Intrusion Detection Systems (IDS), mostly used for Denial of Service (DoS) attacks, have low response in terms of intrusion detection because of missing Local Area Network Denial (LAND) and duration features. Hence, precise security of a system is not assured without considering LAND and duration features. In order to minimize DoS attacks and to make the system more secured, it warrants additional features. All the features are having their certain values that indicate the presence or absence of an intrusion. An existing genetic algorithm has considered 16 features for intrusion detection but, still some DoS and Remote to Local (R2L) attacks are not covered in it. These attacks depend on duration and LAND features of dataset. In the proposed work these two features are focused and extracted using genetic algorithm so that detection response of IDS is improved.

General Terms
Intrusion Detection System, Denial of Service, Intrusion Preventions System

Keywords
KDD Cup Dataset, LAND, Naive Bayesian Classifier, DoS Attacks, Training Datasets

1. INTRODUCTION
Intrusion detection is an important method that monitors network traffic and finds network intrusions such as faulty network behaviors, illegal network access and hostile attacks to computer systems. An Intrusion Detection System (IDS) detects intrusions and reports it accurately to the proper authority [1]. Conventional intrusion prevention strategies, such as firewalls, access control schemes or encryption methods, have failed to prove themselves to effectively protect networks and systems from increasingly sophisticated attacks and malwares. Generally intrusion detection systems are divided into two variations, misuse detection and anomaly detection. Misuse detection depends on the prior representation of specific patterns for intrusions, allowing any matches to it in current activity to be reported. The anomaly based system builds a model of the normal behavior of the system and then looks for anomalous activity such as activities that do not confirm to the developed model. The anomaly detection systems are adaptive that is it deal with new attack but it unable to identify all types of attacks. Many researchers have proposed and implemented various models for IDS but it often generates too many false alerts due to their simplistic analysis.

1.1 Problem statement
It has been observed that many research intended to enhance the intrusion detection system. The researchers worked on limited features that are optimum. But there are still certain features like duration and LAND, have impact on intrusion detection system, which are missed by many researchers. Hence duration and LAND oriented records create problems in network. So to solve the problem of duration and LAND oriented records, it warrants to consider duration and LAND features.

1.2 Proposed solution
In the proposed work, the features like duration and LAND are considered. The features are extracted from the dataset using Genetic Algorithm (GA). The selected features along with duration and LAND are then given to the classifier. The combination of Genetic Algorithm and Classifier is termed as wrapper approach. For feature extraction, training dataset is taken on the basis of which intrusion or anomaly is identified. To identify the anomaly, it requires testing dataset from which anomalies are identified. Finally, mapping process is done on testing dataset that gives result containing various statistics values about an intrusion.

2. RELATED WORK
As bypassing the system security is one of the part of information stealing process that plays a role of theft [1]. But whenever such efforts are taken from outsider, it needs to have a system that easily and quickly identifies the behavior of system. Here the one who does so is called as an intruder. Now-a-days such cases are happening regularly and therefor it is important now to have a system that stops it. A system that cares of all such activities is commonly referred as intrusion detection system [2]. To develop such system, a concept of data mining is used where raw data is taken as an input and corresponding structured data is identified.

Data mining is a domain that gives corrected dataset which is used for experiment purpose. There are various domains available that are used for information retrieval purpose. Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the needs of web-based applications [4]. Usage data captures the identity or origin of Web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered. Web mining

process of extracting important data from big data is domain of data mining.

Information Retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources [5]. Searching is based on metadata or on full-text (or other content-based) indexing. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

Financial data consists of pieces or sets of information related to the financial health of a business. The pieces of data are used by internal management to analyze business performance and determine whether tactics and strategies must be altered. People and organizations outside a business will also use financial data reported by the business to judge its credit worthiness, decide whether to invest in the business and determine whether the business is complying with government regulations.

Automated information retrieval systems are used to reduce what has been called information overload. Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications. In intrusion detection system, features are most relative and important component. These features are need to be accessed and handled by proper channel. Generally, there are various ways to deal with features. The conditional random field is an old approach for developing an IDS.

Feature extraction is one of the most important part of any intrusion detection system that is done with many feature extraction algorithms. Genetic Algorithm (GA) is one of among them that do the feature extraction task effectively that the others [6]. The extracted features are need to be classified using some sort of classifier. The classification process gives related values required for intrusion detection system. Decision Tree is generally used classifier for intrusion detection system. First subsection of this section presents related work on Decision Tree. Support Vector Machine (SVM) is an approach that do the task of feature classification with more precise results than decision tree classifier. The Naive Bayesian classifier is an effective classification method that classifies the feature on probability basics with more precise results. Fuzzy Logic is feature extraction algorithm that do extraction on the basis of fuzzy sets. Fuzzy sets are collection of attributes of the relation.

3. PROPOSED APPROACH

3.1 System architecture

The architecture is a system that unifies its components or elements into a coherent and functional blocks. The architecture shows the structure of system. The architecture of proposed intrusion detection system is divided into two phases

1. Training Phase

2. Testing Phase

3.1.1 Training Phase

The architecture of proposed system is shown in figure 1

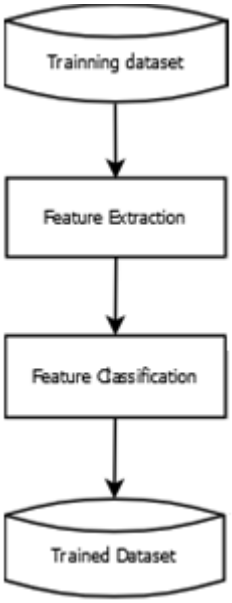

Figure 1: System Architecture for Training Phase

The training phase Architecture is composed from two parts:
An input to training phase is training dataset. The training dataset is dataset by which samples are made to analyze the system. KDD CUP 99 is a dataset that are used here as training dataset. It is an initial input to the proposed system on the basis of that result is calculated.

1. Feature Extraction

An intrusion is easily identified by observing the values of features. The values of features are of 41 different types [6]. As different researcher taken various combination of features for developing an intrusion detection system. The combination of features decides the accuracy of resultant IDS. The accuracy is bind to the most relevant features of dataset. The relevant features not only the way by which an intrusion is identified. The intrusion, sometimes, also detected by combining few features together. In an existing system, the duration and Local Area Network Denial (LAND) features are not considered. These features are also responsible to indicate some Denial of Service (DoS) attacks in combination with other features. Hence in proposed system, duration and LAND features are also extracted. The extraction process starts from an initial population. The population is generated for a number of generations while progressively improving the qualities of the individuals by increasing the fitness value as the measure of quality. During each generation selection, cross over, and mutation are one after the other applied to each individual with certain probabilities. First, the numbers of best fit individuals are selected based on a user defined fitness function. The remaining individuals are selected and paired with each other. Each individual pair produces one offspring by partially exchanging their genes around one or more randomly selected crossing points. At the end, a certain number of individuals are selected and the mutation operations are applied. Selection is the phase where population individuals with better fitness are selected, otherwise it gets damaged. Feature extraction is start with finalizing the gene length. Let, alpha be a gene length that is calculated by expression

$$\alpha = Gl = L \quad (3.1)$$

Where Gl : Gene Length

The gene length is easily calculated using Equation 3.1. But, it is required to set the gene value at initial level. The setting the gene value is required for the purpose of starting an execution of an proposed system. The gene value is also set using a simple formula but with setting the initial fitness value to zero.

$$\beta = Sg = vi \quad (3.2)$$

Where Sg : Setted Gene

Vi: ith value of particular gene

The gene value is set by using Equation 3.2. That gene value has to be accessed within predetermine methods. The methods that use the gene value must get access of it. To give access of gene value, the following expression is used.

$$Gg = vi$$

In this way basic work with gene is completed. The gene value has importance, since it start the system to run. After getting the gene value some portion of feature extraction gets finished. Here, once the gene is initialized and make available to others, the rest part of feature extraction is interested. Because only getting the gene value every time is not so important since gene value is required only to make the start. The major task afterword's is to analyze the gene value and decides whether to involve the current gene in feature extraction process or not. In other word it is termed as analyzing the fitness of a gene. As the gene fitness decides whether to participate the gene into feature extraction process or not. To get the fitness of any gene needs special method that is represented as

$$Gf = fn \quad (3.3)$$

Where fn : _tness function

Here fitness of gene is nothing but,

$$fn = \sum i<Gl;i<Sl, fitness + + \quad (3.4)$$

Where Sl: solution Length

Gl: Gene Length

The gene is a part of chromosome that comprises a population. The population is also referred as an individual. As the fitness value of a gene decides the role of gene, it is very important to collect all such participated genes to form an individual. An individual is composed of various combination.

Crossover

A crossover operator is used to recombine two strings to get a better string. In crossover operation, recombination process creates different individuals in the successive generations by combining material from two individuals of the previous generation. The two strings participating in the crossover operation are known as parent strings and the resulting strings are known as children strings. The crossover operator recombines good sub-strings from good strings together, hopefully to create a better sub string. In cross over process fitness calculation is one of the major task that depends upon candidate solution. The candidate solution is a solution that gives reference as a good one to complete cross over process. To obtain a candidate solution crossover step is required to execute. Whatever solutions are formed during feature extraction process, are examined and opted best solution for further process.

Mutation

Mutation adds new information in a random way to the genetic search process and ultimately helps to avoid getting trapped at local optima. It is an operator that introduces diversity in the population whenever the population tends to become homogeneous due to repeated use of reproduction and crossover operators. Mutation in a way is the process of randomly disturbing genetic information. They operate at the bit level, when the bits are being copied from the current string to the new string, there is probability that each bit become mutated. This probability is usually a quite small value, called as mutation probability. The mutation operator alters a string locally expecting a better string.

2. Feature classification

The classification of extracted features are required in order to get the appropriate information. An information extracted during extraction process has not enough understandable format. To make the information understandable, it require classifies it using some sort of classifier. For proposed system, Naive Bayesian classifier is used. The Bayes networks are one of the most widely used graphical model to represent and handle uncertain information. The Bayes networks are specified by two components that are graphical component and Numerical Component

The Naive Bayesian model is a heavily simplified Bayesian probability model. The model, consider the probability of an end result given several related evidence variables. The probability of end result is encoded in the model along with the probability of the evidence variables occurring given that the end result occurs. The probability of an evidence variable given that the end result occurs is assumed to be independent of the probability of other evidence variables given that end results occur. Assume that a set of examples that monitor some attributes such as whether it is raining, whether an earthquake has occurred etc are available. Let's assume that it is known, using the monitor, about the behavior of the alarm under these conditions. In addition, having knowledge of these attributes, it recorded whether or not a theft actually occurred. Consider the category of whether a theft occurred or not as the class for the Naive Bayesian classifier. The other attributes considered are as knowledge that gives evidence about the theft has occurred. The Naive Bayesian classifier operates on a strong independence assumption. It means that the probability of one attribute does not affect the probability of the other. Given a series of n attributes, the naive Bayes classifier makes 2n! Independent assumptions. Nevertheless, the results of the naive Bayes classifier are often correct. Training data noise is minimized by choosing good training data. The training data must be divided into various groups by the machine learning algorithm. Bias is the error due to groupings in the training data being very large. Variance is the error due to those groupings being too small.

3.1.2 Testing Phase

The architecture of proposed system for testing phase is shown in Figure 2. The testing phase architecture consist only one process that is testing As the trained dataset, received from training phase as their output, contains all related values since it is developed by extracting the features and classifying it. This trained dataset along with testing dataset is applied to testing process as an input. The testing dataset is a dataset that is going to be test according to the trained dataset. Testing dataset is chosen from KDD CUP 99 dataset. Testing dataset contains records of feature oriented information. This information is used and compared with trained dataset.

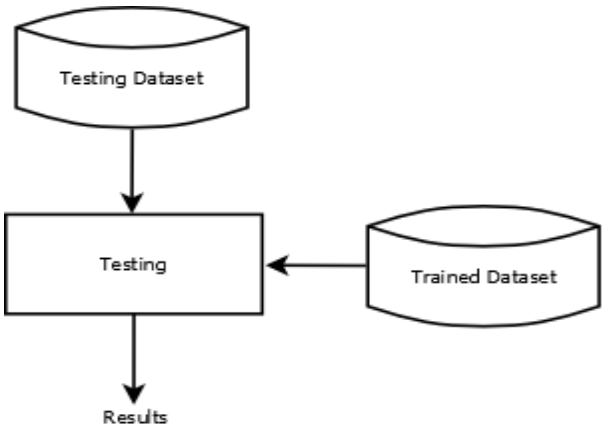

Figure 2: System Architecture for Testing Phase

Testing: As the features are extracted and classified into their equivalent known information, it need to check presence or absence of an intrusion. To check the status of an intrusion, it require to make the training dataset. The training dataset, on the basis comparison is made, is a collection of data that contains trained samples. The trained samples are collection of records that describes the values of all extracted features. The extracted features with corresponding values decides whether it is count as an anomaly or normal record. The anomaly indicates presence of intrusion whereas normal indicates absence of intrusion. The identification of an intrusion is totally based on training dataset. According to the values of anomaly oriented records, training dataset is designed. Trained dataset is compared with testing dataset. The testing dataset is collection of record in which intrusion findings are supposed to.

After checking the testing dataset for an intrusion, the results are obtained using the help of weka inbuilt functions. The weka inbuilt functions are provided by weka family that gives statistics about an intrusion by referring to the datasets. The statistics contains number of record for testing, percentage of anomaly, percentage of normal entries, kappa statistics etc.

4. RESULTS AND DISCUSSIONS

4.1 Implementation details

The proposed system is implemented using JAVA with NetBeans IDE 7.3.1 and MySQL-5.5. The variant of JAVA module jdk with version 7.2 is used here for developing an implementation of proposed solution. The implemented solution of proposed work is then opened in NetBeans IDE where it can also be modified. Initially an IDE is opened and code is typed by following the guideline of an algorithm that are described in previous chapter. Here training phase gets completed. Using an IDE, new form for testing phase is designed where testing dataset is taken as an input along with output of previous that is training phase. Finally, trained dataset is compared with testing dataset by referring weka vector library.

4.2 Experimental test cases

The test cases are generated for 50 records with considering both duration and LAND features. The values of duration and LAND features are responsible for anomaly in association with other features values too.

4.2.1 Test cases without duration and LAND features

Test case (a):

Training dataset: tcp, http, SF, 253, 11905, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 10, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.20, 87, 255, 1.00, 0.00, 0.01, 0.02, 0.00, 0.00, 0.00, 0.00

Testing Dataset: tcp, http, SF, 253, 11905, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 10, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.20, 87, 255, 1.00, 0.00, 0.01, 0.02, 0.00, 0.00, 0.00, 0.00

The Table 1 shows that if the parameter in the form of features, have the mentioned values then it is considered as LAND anomaly. But the duration and LAND features are not considered in testing dataset, therefore the test gets failed.

Table 1 Test case (a)

Attack Type	Parameters	Attack Present	Result
LAND	Protocol=TCP, Service type=HTTP, Flag=SF, src bytes=253, dst bytes=11905	Y	fail

Test Case (b):

Training Dataset: udp, private, SF, 105, 147, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 41, 5, 0.12, 0.05, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00

Testing Dataset: udp, private, SF, 105, 147, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 41, 5, 0.12, 0.05, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00

Table 2 Test case (b)

Attack Type	Parameters	Attack Present	Result
DoS	Protocol=UDP, Service Type=private, Flag=SF, src bytes=105 and dst bytes=147	Y	fail

The Table 2 shows that if the parameter in the form of features, have the mentioned values then it seems that it is normal record. Protocol name is UDP, service type is private, flag is Status flag, source bytes are 105 and destination bytes are 147, then it is not an anomalous record. Hence, on the basis of trained dataset, the test is failed as duration field is absent.

4.2.1 Test cases with duration and LAND features

Test Case (a):

Training dataset: 0, tcp, http, SF, 253, 11905, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 10, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.20, 87, 255, 1.00, 0.00, 0.01, 0.02, 0.00, 0.00, 0.00, 0.00

Testing Dataset: 0, tcp, http, SF, 253, 11905, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 10, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.20, 87, 255, 1.00 ,0.00 ,0.01 ,0.02, 0.00, 0.00, 0.00, 0.00

Table 3 Test case (a)

Attack Type	Parameters	Attack Present	Result
LAND	Duration=0, Protocol = TCP, Service type = HTTP, Flag=SF, src bytes =253, dst bytes=11905, land=1	Y	Passed

The Table 3 shows that if the parameter in the form of features, have the mentioned values then it is considered as LAND anomaly. The duration features value is 0, protocol name is TCP, service type is HTTP, flag is Status flag, source bytes are 253, destination bytes are 11905 and land value is 1, then it is a LAND attack (anomaly). Hence, on the basis of trained dataset, the test is passed.

Test Case (b):
Training Dataset: 1000, udp, private, SF, 105, 147, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 41, 5, 0.12, 0.05, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00
Testing Dataset: 1000, udp, private, SF, 105, 147, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 41, 5, 0.12, 0.05, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00

Table 4 Test case (b)

Attack Type	Parameters	Attack Present	Result
DoS	Duration=1000, Protocol = UDP, Service Type =private, Flag=SF, src bytes=105, dst bytes = 147, land=0	Y	Pass

The Table 4 shows that if the parameter in the form of features, have the mentioned values then it is not normal record. The duration features value is 1000, protocol name is UDP, service type is private, flag is Status flag, source bytes are 105, destination bytes are 147 and land value is 0, then it is an anomalous record. Hence, on the basis of trained dataset and presence of duration value, the test is passed.

20 test cases are taken and their results are mentioned in the following table with considering True Positive, False Positive etc values.

Table 5. Intrusion Detection with Overall Observation

Sr No	No of Records	LAND oriented entries in records	Duration oriented entries in records	True Positive	False Positive	Intrusion Detection rate in percentage
1	20	2	2	20	0	100
2	20	0	2	18	2	90
3	20	2	0	18	2	90
4	20	1	0	17	3	85
5	20	0	0	16	4	80

In Table 5 True positive parameter shows the successive detection of an intrusion and False positive shows unsuccessful or wrong analysis of record. By analyzing all intrusion detection rate of true positive records, it is clear that detection rate is much greater than existing systems that exclude LAND and duration features.

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., “Table 1” or “Figure 2”), please note that the word for Table and Figure are spelled out. Figure’s captions should be centered beneath the image or picture, and Table captions should be centered above the table body.

4.3 Discussion

Wrapper based intrusion detection system is one of the best way to design an Intrusion Detection System (IDS). The used wrapper approach for proposed system takes genetic algorithm (GA) and Naive Bayesian classifier. In the proposed system GA is used to extract the features from the dataset where as Naive Bayesian is used to classify the extracted features. The detection of an intrusion using wrapper approach is most efficient one. The response rate gets maximized for the DoS as well as LAND attacks using proposed system. The DoS attacks are having different way to come into the system, so it is important to consider all responsible features for DoS type of attacks. The features, duration and LAND, are focused in proposed system that reflect the result of an intrusion detection system. From the experimental results, it is clear that DoS attacks are reduced. The reduction of DoS attacks is achieved by involving duration and LAND features. From the experimental results it is also sure now that response rate of an IDS's is improved. As the said features play an important role to achieve the predefined target in terms of an objectives. Hence the objectives of proposed system are achieved with wrapper approach. Now-a-days, many researcher is working on IDS's but, they generally avoid to consider these features due to less chances of these attacks to enter into the system.

5. CONCLUSION AND FUTURE WORK

The wrapper based intrusion detection system with considering Duration and LAND features gives good results. The implemented system detects many DoS class intrusions that are considered as a normal record previously. The proposed system not only gives detection of an intrusions, but it also shows that while developing an intrusion detection system, duration feature and LAND feature must be considered. The proposed system shows that, leaving the duration and LAND features is problematic for resultant intrusion detection system. For results, various test cases are taken and analyzed the output. To analyze the proposed system it is required to have the training dataset and testing dataset. The results of test cases are observed and recorded. The observations are defined on the basis of number of records found fault with considering duration and LAND features. According to observations, it is clear that the said features are responsible for DoS attacks. The proposed system mostly focuses on duration and LAND features of dataset.

In the future work, Web mining along with semantic web, known as semantic web mining, is to be concentrated that is evolving and that helps to overcome the cons of web mining

6. ACKNOWLEDGMENTS

First of all I would like to extend our deep gratitude to almighty God, who has enlightened us with power of knowledge. The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. My deepest thanks to the guide of the work and the Head of Computer Engineering Department, Prof. Dr. Girish K. Patnaik, for guiding and correcting various documents of mine with attention and care. His guidance and encouragement contributed greatly to the completion of this work.

7. REFERENCES

[1] S.S.Sindhu, S. Geetha and A. Kannan, "Decision Tree based light weight intrusion detection using a wrapper approach", Expert System with Applications, 2012, vol.39, pp.129-141.

[2] K.K.Gupta, B.Nath and R. Kotagiri, "Layered approach using conditional random field for intrusion detection", IEEE Transaction on Dependable and Secured Computing, 2010, vol.7, pp.35-47.

[3] D.S.Mukharjee and N. Sharma, "Intrusion detection using naïve bayes classifier with feature reduction", Elsevier Publications, 2012, vol.4, pp.119-128.

[4] M.S.Hoque, M.A.Mukit and M.A.N.Bikas, "An implementation of intrusion detection system using genetic algorithm" International Journal of Network Security and It's Applications, 2012, vol.4, pp.109-120.

[5] M.M.M.Hassan, "Network intrusion detection system using genetic algorithm and fuzzy logic", International Journal of Innovative Research in Computer and Communication Engineering, 2013, vol.1, pp.1435-1445.

[6] A.A.Olusola, A.S.Oladele, D.O.Abosede, "Analysis of kdd 99 intrusion detection dataset for selection of relevance features", Proceeding of World Congress on Engineering and Computer Science, 2011, vol.1, pp. 978-988.

[7] D.Jayalutchumy and D.P.Thambiduri, "Web mining research issues and feature directions a survey", International Organization of Scientific Research Journal of Computer Engineering, 2013, vol.4, pp.20-27.

[8] P.Mehataa, B. Parekh, K.Modi and P. Solanki, "Web personalization using web mining: concept and research issue", International Journal of Information and Education Technology, 2012, vol.2, pp.510-512

[9] M. Yadav and P.Mittal. "Web mining: an introduction", Intrnational Journal of Advanced Research in Computer Science and Software Engineering, 2013, vol.3 pp. 683-687.