International Journal of Electrical, Electronics and Computer Engineering **4**(1): 40-46(2015)

ISSN No. (Online): 2277-2626

Intrusion Detection on KDD99cup Dataset using K-means, PSO and GA: A Review

Akhilesh Singh*, Harsha Banafar** and Ravi Singh Pippal**

**M. Tech. Scholar, Department of Computer Science Engineering, REC Bhopal (M.P.), INDIA*

***Professor, Department of Computer Science and Engineering, REC Bhopal (M.P.), INDIA*

(Corresponding author: Akhilesh Singh)

(Received 04 January, 2015 Accepted 14 February, 2015)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Nowadays use of computer with internet becomes very popular technology but the use of this is most susceptible to severe attack. It is key challenge for the user to protect their personal information and resources from the attack. An intrusion detection system is the process for identifying attacks on network. Intrusion detection system is categorized into two types: Anomaly based and misuse based detection. The data mining techniques make feasible to observe the network to thwart it from the intruders such as K-means and PSO etc. Different researcher works for the detection of intrusion on network. In this KDDCUP1999 dataset is used for the evolution for the performance of network. This paper presents literature of the different approach developed by the author to prevent the network from serious threats such as: Denial of services, remote to local (R2L), user to root (U2R) etc together with its merits and demerits.

Keywords: Anomaly, IDS, K-means, KDD dataset, Misuse, PSO, Threats

I. INTRODUCTION

The use of computer with internet rapidly increasing in present days but the use of computer network now becomes more susceptible to various intruders. The principle of the Intrusion detection system (IDS) is to thwart the computer system from attack. The IDS is the most critical part of the security infrastructure for the networks connected to the internet because diverse ways to compromise the stability and security of network. IDS can be classified into two types: Anomaly and Misuse detection. Anomaly detection system generates a database of normal activities and any deviations from the normal behavior are occurred watchful is triggered regarding the occurrence of intrusions. Misuse Detection system stores the predefined attack patterns in the database if an analogous data and if analogous circumstances occur it is classified as attack. Based on the source of data the intrusion detection system are classified to Host based IDS and Network based IDS. In network based IDS the individual packet flowing throughout the network are examined. The host based IDS analyzes the behavior on the single computer or host. The major shortcoming of the misuse detection (signature detection) method is that it cannot sense novel attacks and discrepancy of known attacks. To evade these disadvantages we go for anomaly based detection methods. With this approach, known and fresh attacks can be detected. The difficulty is that it will produce more false alarms [1]. The intrusion detection technique based on unsupervised learning has a high detection rate but also a high False positive rate.

In Section II discuss related work for decreasing the energy consumption. The Section III discusses about the different routing techniques. Section IV describes the proposed methodology and last section presents conclusion the paper.

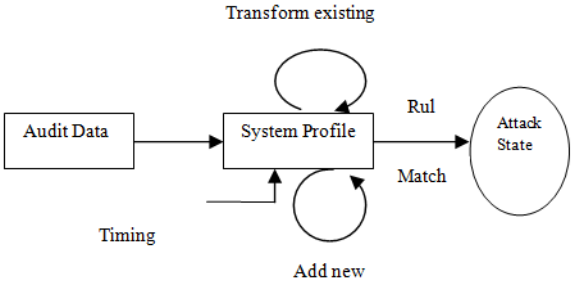

Fig. 1. Misuse based detection systems.

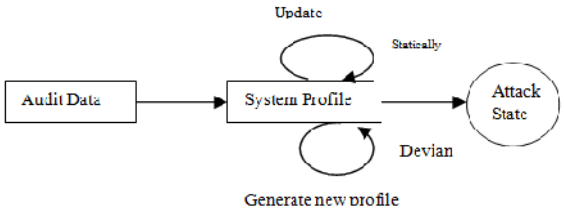

Fig. 2. Anomaly based detection system.

A. KDDCUP'99 Dataset

The KDD Cup 1999 dataset has been used for the evaluation of intrusion detection methods. The KDD Cup 1999 training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type [2].

In KDD Cup 1999 dataset has the different types of attacks: back, buffer_overflow, ftp_write, guess_passwd, imap, ip_sweep, land, load module, multi-hop, neptune, nmap, normal, perl, phf, pod, port sweep, root kit, satan, smurf, spy, teardrop, warez client, warez master. The datasets contain a total number of 24 training attack types, with an additional 14 types in the test data only. These attacks can be divided into 4 groups [3]. The Table.1 shows the list of attacks in group wise:

Table 1: List of attacks - group wise.

Testing Dataset	**Types of attack**
DoS	back, land, Neptune, pod, smurf, teardrop
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy warezclient, Warezmaster
U2R	buffer_overflow, load module Perl, rootkit
Probe	IP_sweep, n_map, port_sweep, satan

Denial of Service (DoS) attacks: Deny justifiable appeals to a system, e.g. flood, User-to-Root (U2R) attacks: unauthorized access to local super user(root) privileges, e.g. diverse buffer overflow attacks, Remote-to-Local (R2L) attacks: unauthorized access from a remote machine, e.g. guessing password, and Probing: surveillance and other probing, e.g. port scanning [4].

The sets are named as A, B, C, D, and E correspondingly. The set 'A' attains data from DoS class. The set 'B' attains data from U2R class. The set 'C' acquires data from R2L class. The set 'D' attains data from Probe Class. The set 'E' attains data from Normal class. The subsequent sets of data can be used for training and testing the data from KDD Cup 1999 dataset.

Table 2: Training and Testing Data Set.

Dataset	**Training set**	**Testing set**
DoS	300	300
R2L	20	19
U2R	300	300
Probe	300	300
Normal	300	300
Total	1220	1219

The 41 featured dataset and concentrated featured dataset for each class is used to sense the attacks in KDD Cup 1999 dataset. The 41 features are listed in the website [2]. For converting symbols into numerical form, an integer code is assigned to each symbol. For instance, in the case of protocol_type feature, 0 is assigned to tcp, 1 to UDP, and 2 to the ICMP symbol and so on.

Attack names are first mapped to one of the five classes, 'A' for DoS, 'B' for U2R, 'C' for R2L, 'D' for Probe and 'E' for Normal. Two features spanned over a extremely large integer range, namely src_bytes [0, 1.3 billion] and dst_bytes [0, 1.3 billion]. Logarithmic scaling (with base 10) is applied to these features to diminish the range to [0.0, 9.14]. All other features are Boolean, in the range [0.0, 1.0].

Consequently scaling is not indispensable for these attributes. 300 signals from DoS, R2L, Probe and Normal class each and 20 signals from U2R class are selected for training the network. Four dissimilar neural networks are used for training the KDD Cup 1999 data. The networks are typically trained to execute tasks such as pattern recognition and decision-making. Table 2 corresponds to the training set. 300 signals from DoS, R2L, Probe and Normal class each and 19 signals from U2R class are selected for testing the network. Four dissimilar neural networks are used for testing the KDD Cup 1999 data. By testing the KDD Cup 1999 data, the exactness of the each neural networks are measured. The Table.2 corresponds to the testing set [5].

II. RELATED WORK

A new technique of combining swarm intelligence (Simplified Swarm Optimization) and data mining algorithm (Random Forest) for feature selection and reduction. SSO is used to find more appropriate set of attributes for classifying network intrusions, and Random Forest is used as a classifier. In the preprocessing step, we optimize the dimension of the dataset by the proposed SSO-RF approach and find an optimal set of features. SSO is an optimization method that has a strong global search capability and is used here for dimension optimization [6].

Two techniques C5.0 and artificial neural network (ANN) are utilized with feature selection [7]. Feature selection techniques will discard some irrelevant features while C5.0 and ANN acts as a classifier to classify the data in either normal type or one of the five types of attack. KDD99 data set is used to train and test the models; C5.0 model with numbers of features is producing better results with all most 100% accuracy. Performances were also verified in terms of data partition size.

A technique which is divided into four steps: initial step, k-means clustering is used to generate different training subset then based on the obtained subset, various neuro-fuzzy data model are trained [8]. Consequently, a vector for SVM classification is obtained and in last, classification using radial SVM is applied to detect the intrusion occurred or not. To demonstrate the applicability and ability of the new method, the result of KDD dataset is confirmed in which it shows that the proposed methods produce better result than the BP, multi-class SVM and other approach such as decision tree etc.

Bharat *et al.*, proposed a method for intrusion detection using Particle Swarm Optimization with Genetic Algorithm based feature selection and using Adaptive Mutation for sluggish convergence of optimization algorithm [9].

The results thus obtained are around 92% that proves the proposed method to be reasonably effective in intrusion detection.

Vahid developed a hybrid method of C5.0 and SVM and investigate and evaluate the performance of our proposed method with DARPA dataset. The motivation for using the hybrid approach is to improve the accuracy of the intrusion detection system when compared to using individual SVM and individual SVM [10].

Hu introduced a new host-based anomaly intrusion detection methodology using discontinuous system call patterns, in an endeavor to increase detection rates whilst plummeting false alarm rates [8]. The main idea is to apply a semantic structure to kernel level system calls in order to replicate inherent activities hidden in high-level programming languages which can help comprehend program anomaly behavior. Outstanding results were demonstrated using a multiplicity of decision engines evaluating the KDD98 and UNM data sets and a new, modern data set. The ADFA Linux data set was created as part of this research using a recent operating system and contemporary hacking methods and is now openly available. Additionally, the new semantic method possesses an inherent flexibility to mimicry attacks and demonstrated a high level of portability between dissimilar operating system versions.

The clustering method by using hybrid method based on Principal Component Analysis (PCA) and Fuzzy Adaptive Resonance Theory (FART) for classifying diverse attacks. The PCA is concerned to random selects the best provenance and reduction the feature space. The FART is implementing which is used to classifying dissimilarity in collection of data, regular and irregular. The proposed method can improves the high performance of the detection rate and to reduce the false alarm rate and this is computed approach on the benchmark data from KDD Cup 99 data set [19].

The method which is based on MAHALANOBIS Distance characteristic ranking and an improved comprehensive search to choose an improved combination of features. They evaluated the approach on the KDD CUP 1999 datasets using SVM classifier and KNN classifier. The results showed that classification is done with high classification rate and low misclassification rate with the reduced feature subsets [20].

III. OVERVIEW OF INTRUSION DETECTION TECHNIQUES

There are different data mining techniques used for Intrusion detection. In this section we are describing general idea of these techniques:

A. K-means Clustering

K-means clustering [12] is one of the simplest unsupervised clustering algorithms. The algorithm takes input parameter „k and partition the „n dataset into k cluster so that the intra-cluster similarity is high and inter-cluster similarity is low. "K is a positive integer number given in advance. K means clustering takes less time as compared to the hierarchical clustering and yields better results.

With the help of clustering training dataset is clustered into 5 dataset wherein 4 dataset will be a type of intrusion called attack dataset and one with normal data type called normal dataset. Here are the four steps of the clustering algorithms:

1. Define the number of clusters K.
2. Initialize the K-cluster centroids. This can be done by randomly dividing all objects into K clusters, computing their centroids, and authenticating that all centroids are diverse from each other. Otherwise, the centroids can be initialized to K arbitrarily preferred, diverse objects.
3. Iterate over all objects and calculate the distances to the centroids of all clusters. Allot each object to the cluster with the adjoining centroid.
4. Recalculate the centroids of both customized clusters.
5. Reiterate step 3 until the centroids do not change any more.

A distance function is obligatory in order to calculate the distance (i.e. similarity) among two objects. The regularly used distance function is the Euclidean one which is defined as:

$$d(x,y) = \sqrt{\sum (x_i - y_i)^2}$$

Where x = (x1 . . . xm) and y = (y1…ym) are two input vectors with m quantitative features. In the Euclidean distance function, all features contribute equally to the function value. However, since different features are usually measured with different metrics or at different scales, they must be normalized before applying the distance function.

B. Genetic Algorithm

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics. The aim of development of GAs is developing a system as robust and as adaptable to the environment as the natural systems [13].

Genetic algorithms are search procedures often used for optimization problems. In this algorithm an initial population of chromosomes is generated randomly where each chromosome represents a possible solution to the problem (a set of parameters). From each chromosome different positions are encoded as, characters, bits or no.s. These positions could be known as genes. Goodness of each chromosome calculated by evaluation function, according to the desired solution; this function is known as "Fitness Function" [14]. It holds three phases after calculating fitness function i.e. selection, crossover, and mutation. In selection it selects most optimal solution of a problem calculated by using fitness function. The selected chromosomes are called parents. After selection phase crossover phase comes in which characteristics of different parent chromosomes exchange and they produce offspring, there are various methods for crossover, for example N point crossover, uniform crossover etc. Mutation involves flipping of one or more bits of chromosomes and then evaluated using some fitness criteria. After termination chromosomes having the highest fitness function called the best solution of the problem. Mutation maintains diversity in the population. Genetic algorithm from other algorithm because it implemented at machine code level, it is fast to detect in real time. Some of its good properties, e.g. robust to noise, no gradient information is needed to find global optimal or sub-optimal solution, self-learning capabilities made it best approach.

C. Principle Components Analysis

It is well known that principal component analysis (PCA) is an essential technique in data compression and feature extraction [15], and it has been also applied to the field of ID. It is well known that PCA has been widely used in data compression and feature selection. Feature selection refers to a process whereby a data space is transformed into a feature space, which has a reduced dimension. Some basic knowledge of PCA is briefly described in the next. Assume that {xt} where t = 1, 2 . . . , N are stochastic n dimensional input data records with mean (μ). It is defined by the following Equation:

$$\mu = \frac{1}{N}\sum_{t-1}^{N} x_t \qquad (1)$$

The covariance matrix of xt is defined by

$$C = \frac{1}{N}\sum_{t=1}^{N}(x_t - \mu).(x_t - \mu)^T \qquad (2)$$

PCA solves the following Eigen value problem of covariance matrix C:

$$Cv_i = \lambda_i v_i \qquad (3)$$

Where λ_i (i = 1, 2... n) are the Eigen values and vi (i = 1, 2, ..., n) are the corresponding eigenvectors. To represent data records with low dimensional vectors, we only need to compute the m eigenvectors (called principal directions) corresponding to those m largest Eigen values (m < n). It is well known that the variance of the projections of the input data onto the principal direction is greater than that of any other directions. Let

$$\Phi = [v_1, v_2 \dots\dots v_m], \Lambda = diag[\lambda_1, \lambda_2 \dots\dots \lambda_m] \quad (4)$$

Then

$$C\Phi = \Phi\Lambda \qquad (5)$$

The parameter v denotes to the approximation precision of the m largest eigenvectors so that the following relation holds.

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq v \qquad (6)$$

Based on (5) and (6) the number of eigenvectors can be selected and given a precision parameter v, the low dimensional feature vector of a new input data x is determined by

$$x_f = \Phi^T x \qquad (7)$$

D. Support Vector Machine

The Support Vector Machine is one of the most successful classification algorithms in the data mining area.SVM uses a high dimension space to find a hyper-plane to perform binary classification.SVM approach is a classification technique based on Statistical Learning Theory (SLT). It is based on the idea of hyper plane classifier. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized. The SVM uses a portion of the data to train the system. It finds numerous support vectors that correspond to the training data. These support vectors will form a SVM model. According to this model, the SVM will categorize a given unknown dataset into target classes [16].

E. Particle Swarm Optimization

Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Kennedy and Eberhart [17], inspired by social behavior such as bird flocking or fish schooling. A high-level view of PSO is a collaborative population based search model. Individuals in the population are called particles, representing potential solutions. The performance of the particles is evaluated by a problem-dependent fitness. These particles move around in a multidimensional searching space. They move toward the best solution (global optimum) by adjusting their position and velocity according to their own experience (local search) or the experience of their neighbors (global search), as shown in Equation 7. In a sense, PSO combines local search and global search to balance exploitation and exploration.

$$v_i(t) = w \times v_i(t-1) + c_1 \times r_1(p_i^l - x_i(t-1) + c_2 \times r_2(p_i^g - x_i(t-1) \qquad (7a)$$

$$x_i(t) = x_i(t-1) + v_i(t) \qquad (7b)$$

where i = 1; 2; : : : ;N, population size N; vi(t) represents the velocity of particle i, which implies a distance traveled by i in generation t; xi(t) represents the position of i in generation t; p_i^l represents the previous best position of i; p_i^g represents the previous best position of the whole swarm; w is the inertia weight which balances the local and global searching pressure;c1 and c2 are positive constant acceleration coefficients which control the maximum step size of the particle; r1 and r2 are random number in the interval [0,1], and introduce randomness for exploitation.

PSO has shown good performance in solving numeric problems. In the perspective of intrusion detection, PSO algorithms have been used to learn classification rules. Chen et al: [18] demonstrated a “divide-and-conquer” approach to incrementally learning a categorization rule set using a ordinary PSO algorithm. This algorithm starts with a full training set. One run of the PSO is expected to fabricate the best classifier which is added to the rule set. In the meantime, data covered by this classifier are deleted from the training dataset. This development is continued until the training dataset is empty. Table 3 shows the merits and demerits of the discussed method for the detection of intrusion is illustrated below.

Table 3: Merits and demerits of different IDS methods.

Methods	Merits	Demerits
K-Means Clustering	1. Very effective at detecting known threats 2. Fast processing time	1. It does not work with global cluster 2. Low false positive rate
Genetic algorithm (GA)	1. High detection rate 2. Reduces the complexity	1. Training process needs more time 2. It is not able to detect group features
Principle component Analysis	1. High detection rate 2. Low false alarm rate	1. It requires extensive training time 2. Low performance
Support Vector Machine	1. It gives better visibility of behavior of each 2. Highly accurate	1. Difficult to train system for dynamic 2. It takes more training time
Particle Swarm Optimization (PSO)	1. Detection rate is higher 2. Speed of convergence is higher	1. Degeneracy 2. Cluster dependence

Table 4: KDD Cup 99 Dataset 41 Features.

S. No.	Attributes Name	S. No.	Attributes Name
1	Duration	22	Count
2	protocol_type	23	srv_count
3	service	24	serror_rate
4	flag	25	srv_serror_rate
5	src_bytes	26	rerror_rate
6	dst_bytes	27	srv_rerror_rate
7	land	28	same_srv_rate
8	wrong_fragment	29	diff_srv_rate
9	urgent	30	srv_diff_host_rate
10	hot	31	dst_host_count
11	num_failed_logins	32	dst_host_srv_count
12	logged_in	33	dst_host_same_srv_ rate
13	num_compromised	34	dst_host_diff_srv_rate
14	root_shell	35	dst_host_same_src_ port_rate
15	su_attempted	36	dst_host_srv_diff_host_rate
16	num_root	37	dst_host_serror_rate
17	num_file_creations	38	dst_host_srv_serror_rate
18	num_shells	39	dst_host_rerror_rate
19	num_access_files	40	dst_host_srv_rerror_ rate
20	num_outbound_ cmds	41	Count
2 1	is_guest_login		

CONCLUSION

The transmission of personal information or sharing of resources over the network becomes very critical task for the users because the information or resources are trapped by different intruders. For the detection of intruders there are many system and methods has been developed and proposed.

In this paper we present the survey of literature of different approaches implemented and different methods with their merits and demerits is described. In this some methods improves the performance and able to detect the intrusion. In future work, develop an algorithm which requires less training time and efficient to detect the intruders.

REFERNCES

[1]. Feng Guorui, Zou Xinguo, Wu Jian, (2012). "Intrusion detection based on the semi supervised Fuzzy C- Means clustering algorithm", Department of Information Science Technology, Shandong University, China, pp. 2667-2670, 2012.

[2]. The UCI KDD Archive, "KDD Cup 1999 Data", Information and Computer Science, 1999.

[3]. S. Devaraju and S. Ramakrishnan, "Performance Analysis of Intrusion Detection System Using Various Neural Network Classifiers", *International Conference on International Conference on Recent Trends in Information Technology*, pp. 1033-1038, 2011.

[4]. V. Venkatachalam and S. Selvan, (2007). "Intrusion detection using an improved competitive learning lamstar neural network", *International Journal of Computer Science and Network Security*, Vol. **7**, No. 2, pp. 255-259.

[5]. S. Devaraju and S. Ramakrishnan, "Performance Comparison of Intrusion Detection System using Various Techniques – A Review", *ICTACT Journal on Communication Technology*, Vol. **4**, No. 3, pp. 802-812, 2013.

[6]. S. Revathi, A. Malathi "Network Intrusion Detection Using Hybrid Simplified Swarm Optimization and Random Forest Algorithm on NSL-KDD Dataset", *International Journal of Engineering And Computer Science*, ISSN: 2319-7242 Volume **3** Issue 2 February, 2014 Page No. 3873-3876.

[7]. Pratibha Soni, Prabhakar Sharma, "An Intrusion Detection System Based on KDD-99 Data using Data Mining Techniques and Feature Selection", *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Vol. **4** Issue-3, July 2014.

[8]. A. M. Chandrasekhar, K. Raghuveer " Intrusion Detection Techniques by using K-means, Fuzzy Neural network and SVM classifier", ICCCI- 2013, Jan. 04-06, 2013, Coimbatore, INDIA.

[9]. Bharat Rathi, Dattatray V. Jadhav "Network Intrusion Detection Using PSO Based on Adaptive Mutation and Genetic Algorithm", *International Journal of Scientific & Engineering Research*, Volume **5**, Issue 8, August -2014, ISSN 2229 - 5518

[10]. Vahid Golmah "An Efficient Hybrid Intrusion Detection System based onC5.0 and SVM", *International Journal of Database Theory and Application*, Vol.**7**(2): (2014), pp.59-70 http://dx.doi.org/10.14257/ijdta.2014.7.2.06

[11]. Jiankun Hu, "A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns", *IEEE Transactions on Computers*, vol. **63**, no. 4, pp. 807-819, April 2014.

[12]. Pankaj Pandey , Sanjay Kumar Sharma, Mahendra Singh Sisodiya and Susheel Kumar Tiwari "An Improved Intrusion Detection Based on K-means Clustering via Naïve Bayes Classification", *Proceedings of ICAESM*, pp. 417-422,2012.

[13]. Swati Sharma, Santosh Kumar, Mandeep Kaur "Recent trend in Intrusion detection using Fuzzy-Genetic algorithm" *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. **3**, Issue 5, May 2014, ISSN (Online) : 2278-1021.

[14]. Mostaque Md and M. Hassan, "Current studies on intrusion detection system, genetic and fuzzy logic," *International Journal of Distributed and Parallel Systems* (IJDPS), vol. **4**, (2013).

[15]. Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham "Principle Components Analysis and Support Vector Machine based Intrusion Detection" 2010. *10th International Conference on Intelligent Systems Design and Applications. In proceeding of IEEE*.

[16]. Yogita B. Bhavsar, Kalyani C. Waghmare "Intrusion Detection System Using Data Mining Technique: Support Vector Machine", *International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459, Volume **3**, Issue 3, March 2013.

[17]. J. Kennedy and R. Eberhart. Particle swarm optimization. *In Proceedings of IEEE International Conference on Neural Networks*, Volume **4**, pages 1942–1948. IEEE Press, Nov/Dec 1995.

[18]. G. Chen, Q. Chen, and W. Guo. A PSO-based approach to rule learning in network intrusion detection. In B.-Y. Cao, editor, *Fuzzy Information and Engineering, of Advances in Soft Computing*, Volume **40** pages 666–673. Springer Berlin / Heidelberg, 2007.

[19]. Preecha Somwang, Woraphon Lilakiatsakun: "intrusion detection technique by using fuzzy ART on computer network security" 2011, 978-1-4577-2119, in IEEE.

[20]. Zhao Y ongli, Zhang Yungui, Tong Weiming, Chen Hongzhi, "an improved feature selection algorithm based on MAHALANOBIS distance for network intrusion detection", international conference on sensor network security technology and privacy communication system (SNS & PCS), 2013.

[21]. Devendra kailashiya, R.C. "Improve Intrusion Detection Using Decision Tree with Sampling", *Int. J. Computer Technology & Applications*, Vol. **3** (3), 1209-1216, ISSN:2229-6093.