IJCTA, 9(27), 2016, pp. 503-511 ISSN: 0974-5572
© International Science Press

A Hybrid Classification Model for IP Spoofing in Network Forensic

N. Venkataramanan[1], Dr. T.N. Ravi[2]

ABSTRACT

The attackers can hide their real identity by using spoofed IP addresses. To track the spoofers, a various number of different mechanisms are suggested. However, due to the difficulties of implementation, there has been no commonly adopted IP trace back mechanism, at least at the Internet-level. Consequently, the mist on the places of spoofers has never been dissipated until now. In this paper, a novel hybrid methodology is proposed to detect the spoofed IP address. Another computational system was proposed to perform the classification task and separating the elements from the given dataset KDDCUP 99 datasets utilizing a hybrid feature selection as a part of the pre-processing and Naïve Bayes is utilized to group that ideal dataset for Spoofed IP address.

Keywords: IP Spoofing, Information Gain, Quick Reduct, Naïve Bayes Classification, Network Forensic.

1. INTRODUCTION

IP Spoofing, which is network utilized by assailants for starting attacks utilizing fake source IP locations, is considered as a genuine security issue on the web. Attackers use delivers that are apportioned to others or unassigned locations, to counteract revealing their real areas, or enhance the effect of assault, or to dispatch reflection based assaults. Some understood assaults that rely on upon IP spoofing are DNS amplification, SYN Flooding and SMURF and so forth. A DNS amplification assault which genuinely crumbled the working of a Top Level Domain (TLD) name server is accounted for in [1]. A report of ARBOR on NANOG 50th conference reveals spoofing is still critical in watched DoS strikes [2]. Recognizing the sources of IP spoofing traffic is of extraordinary significance. For whatever length of time that their areas are not uncovered, they can't be debilitated from dispatching further assaults. Indeed, even simply nearing the spoofers, for instance, deciding the ASes (Autonomous Systems) or networks they live in, attackers can be situated in a minimal measured spot, and filtration mechanism can be set nearer to the assailant, before the satirizing activity gets packaged. Moreover, this can build up a reputation framework for ASes, which would be helpful to constrain the relating ISPs to check IP addresses. A key strategy for passive perception of spoofing exercises is the utilization of Network Telescopes [3].Network telescope gets non-requested messages, which are chiefly created by casualty frameworks hit by traffic with source prefix set in the scope of the telescope. At present, the greatest reach telescope is the CAIDA UCSD telescope, which holds 1/256 of all the IP addresses and is essentially used to screen DDoS exercises. Moore el at. [4] gave a procedure known as backscatter examination which induces components of DoS strikes in light of records assembled by the network telescope. The MIT Spoofer Project [5] tries to uncover which networks can discharge spoofing focused strikes. Volunteer members set up a customer that survey the spoofing capacity of their networks and hosts. The outcomes uncovers 6700 ASes out of 30205 don't filter spoofing.

2. PROPOSED HYBRID CLASSIFICATION MODEL FOR IP SPOOFING

The proposed Hybrid Classification model for IP spoofing is exhibited in the figure 1. From the figure 1, the dataset KDDCUP 99 [10] is given as the info dataset. Information preprocessing is an imperative stride in the machine learning processing that dispenses with out of missing value, range value, impossible data combinations and so on. For the most

[1] Research Scholar, P.G & Research Department of Computer Science, Periyar E.V.R College (Autonomous), Trichy, Tamilnadu, India

[2] Assistant Professor, P.G & Research Department of Computer Science, Periyar E.V.R College (Autonomous), Trichy, Tamilnadu, India

part information preprocessing incorporates feature selection and extraction, transformation, normalization and learning. The yield of the information preprocessing is the last preparing set that concentrates learning for the testing stage. The accompanying strides utilized for information preprocessing:

- Identifying components (features) and its related qualities.
- Converting unique features information esteem into numerical information esteem.
- Applying information standardization in view of min-max standardization.
- Perform remove null values and similarity check.

In the pre-processing step, the two feature selection strategies are hybridized to lessen the attribute size of the dataset. In this structure, information gain and quick reduct rough set is utilized for hybridization. Naive Bayes is utilized to group the ideal dataset as Spoofed IP address and Non-Spoofed IP Address classification.

A. Information Gain Feature Selection Method

In this technique, the perceptibility capacity is utilized [6]. The perceptibility capacity is given as takes after: For a data framework (H,I), s detectability capacity DC is a boolean capacity of m Boolean variables f1,f2… … ..fm comparing to the traits f1,f2..… fm individually, and characterized as takes after: DC(f1,f2… .fm)=Q1 ‘” Q2 ‘”… ..Qn where fk • Q. The proposed calculation for the data pick up characteristic subset assessment is characterized as underneath:

Step 1:Compute perceptibility lattice for the chose dataset. By utilizing S[K,J]={ f • F, where

A[K] ≠ A[J] and B[K] ≠ B[J]} K,J=1,2,… .n   Eq1

Where A are condition attribute and B is a decision attribute. This detectability matrix S is symmetric. Where S[c,d]=P[d,c] and P[c.c]=0. Along these lines, it is adequate to consider just the lower triangle or the upper triangle of the matrix.

Step 2:Compute the discernibiliy capacity for the perceptibility matrix S[c,d] by utilizing

DC(c) = ‘” {‘” S[c,d]/c,d • H; S[c,d] `”0}   Eq2

Step 3: Select the property, which has a place with the extensive number of conjunctive sets, numbering no less than two, and apply the law of expansion.

Step 4: Repeat steps 1 to 3 until the law of expansion can’t be connected for every part.

Step 5:Substitute all unequivocally proportional classes for their relating properties.

Step 6:Calculate the Information gain for the improved perceptibility capacity contained qualities by utilizing

$$\text{Gain(Ij)} = \text{F(Pj)} - \text{F(Ij)} \qquad \text{Eq.3}$$

where

$$F(1) = \Sigma_{k=1}^{n} P_K log_2 P_K \qquad \text{Eq.4}$$

$$\Sigma_{k=1}^{n} P_K log_2 P_K = -\frac{p1}{p} log_2 \frac{p1}{p} - \frac{p2}{p} log_2 \frac{p2}{p} - \frac{pn}{p} log_2 \frac{pn}{p} \qquad \text{Eq.5}$$

Where Pk is the proportion of contingent quality P in dataset. At the point when Ij has | Ij | sorts of property estimations and condition characteristic Pk segments set P utilizing trait Ij, the estimation of data F(Gi) is characterized as

$$F(I) = \Sigma_{j=1}^{Ij} Wj * F(Ij) \qquad \text{Eq.6}$$

Step 7: Choose the most elevated Gain esteem and add it to the lessening set, and expel the trait from the perceptibility capacity. Goto step 6 until the perceptibility capacity achieves invalid set.

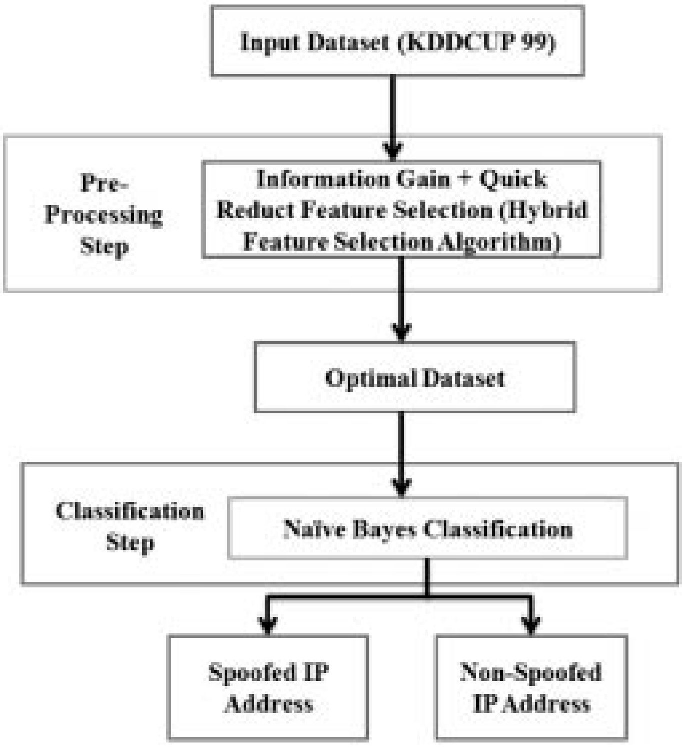
Input Dataset (KDDCUP 99)
Pre-Processing Step
Information Gain + Quick Reduct Feature Selection (Hybrid Feature Selection Algorithm)
Optimal Dataset
Classification Step
Naïve Bayes Classification
Spoofed IP Address
Non-Spoofed IP Address

Figure 1: Proposed Hybrid Classification Model for IP Spoofing

B. Quick Reduct Rough Set

Quick Reduct Algorithm is the most understood calculation for feature selection utilizing Rough sets [7][8]. This is an incremental methodology; where it begins with a void set and in every stride a feature is added to the Reduct, in such way that dependency quantifies increments. The methodology stops when the dependency measure of the arrangement of elements being considered is equivalent to the dependency measure utilizing all the conditional features. The calculation endeavors to figure a reduct without comprehensively producing every single conceivable subset [8]. Its pseudo code calculation is given underneath:

Input: Original Dataset,

D the set of all conditional features; S- the set of decision features, a reduct is defined as Q subset.

Input: Original Dataset

```
    Begin
    Initialize Q as Empty set (is represented by {})
    $R \leftarrow Q$
    $\forall \chi \in_{(D-Q)}$
    if $\gamma_{Q \cup \{\chi\}}(S) > \gamma_q(S)$
    $R \leftarrow Q \cup \{\chi\}$
    $Q \leftarrow R$
    Until $\gamma_{q(S)} = \gamma_{d(S)}$
    Return Q
End
```

Output: A Reduct Dataset

The QUICKREDUCT algorithm endeavors to ascertain a reduct without completely creating every single possible subset. It begins off with a vacant set and includes turn, each one in turn, those features that outcome in the best increment in the rough set dependency metric, until this delivers its most extreme conceivable quality for the dataset.

C. Naïve Bayes Classification

Naive Bayes is a strategy for assessing probabilities of individual variable qualities, given a class, from preparing information and to then permit the utilization of these probabilities to order new elements, which is a term in Bayesian insights managing a straightforward probabilistic classifier taking into account applying Bayes' hypothesis (from Bayesian measurements) with strong (guileless) autonomy assumptions. In basic terms, a strong Bayes classifier expect that the nearness (or nonappearance) of a specific feature of a class is disconnected to the nearness (or nonattendance) of some other element. The Naive Bayesian classifier, fills in as taking after inference [9]-[18]:

Step 1: Let T be a training set of tuples and their related class names. Each tuple is spoken to by a m-dimensional attribute vector, A = (a1, a2, … .. , am), m estimations made on the tuple from m properties, individually, X1, X2, … , Xm.

Step 2: Suppose that there are n classes D1, D2, … . , and Dn. Given a tuple, A, the classifier will anticipate that A has a place with the class having the most noteworthy back likelihood, adapted on A. That is, the guileless Bayesian classifier predicts that tuple A has a place with the class Tj if and just if

$$P\Big((D_j|A \;>\; P\ (D_k|\mathrm{A}\ \ \text{for}\ 1\ \le\ \mathrm{k}\ \le\ \mathrm{n},\ \mathrm{k}\ \ne\ \mathrm{j}$$

The boost P(Dj|A). The class Dj for which P(Dk|A) is amplified is known as the most extreme posterior hypothesis. By Bayes' hypothesis (Next condition)

$$P(D_j|\,A) = \frac{P(A|D_j)\,P(D_j)}{P(A)}$$

Step 3: Since P(A) is consistent for all classes, just (P(Dj|A) = P(A |Dj)P(Dj)) should be amplified.

Step 4: Based on the supposition is that properties are restrictively free (i.e., no reliance connection between attributes), the registering of P(A|Dj) utilizing the accompanying condition:

$$P(A|D_j) = \prod_{i=1}^{m} P(a_i|D_j)$$

Diminishes the calculation cost by Equation (P(Dj|A) = P(A |Dj)P(Dj), just numbers the class appropriation. On the off chance that Xi is unmitigated, P(Ai|Dj) is the no. of tuples in Dj having esteem Ai for Xi separated by |Dj, T| no. of tuples of Dj in T. Also, if Xi is persistent esteemed, P(Ai|Dj) is typically processed in view of Gaussian circulation with a mean ì and standard deviation ó and P(Ai|Dj) is:

$$\mathrm{g}\ (\mathrm{x},\ \mu,\ \sigma) = \frac{1}{\sqrt{2\pi\sigma}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(D_j|\,A) = g(a_i,\ \mu_{Dj},\ \sigma_{Dj})$$

Where ì is the mean and ó is the difference. On the off chance that a property estimation doesn't happen with each class esteem, the likelihood will be zero, and a posteriori likelihood will likewise be zero.

D. Pseudo Code for Proposed Hybrid Classification Model for IP Spoofing

```
    Input: S ($A_0$, $A_1$,... $A_{m-1}$) // A training dataset with M features
        $R_0$                  // a subset from which to start the search
    Algorithm:
        Begin
        Initialize: $R_{best}$ = $R_0$;
        $d_0$ = card ($R_0$) // Calculate the cardinality of $R_0$
                $\theta_{best} = eval\ (R_0, S, I)$ ; // evaluate $S_0$ by an independent measure I
                $\gamma_{best} = eval\ (R_0, S, Q)$; // Evaluate $S_0$ by a Quick Reduct algorithm Q
        for d = $d_0$+1 to M begin
        for j =0 to M-d begin
        R= $R_{best}$ *” {Ai};//Generate a subset with cardinality c for   evaluation
        $\theta$ = eval (R, S, I); //evaluate the current subset R by Independent measure I
        if ($\theta$ is better than $\theta_{best}$)
        $\theta_{best}$ = $\theta$;
        $R'_{best}$ = R
        end;
        $\gamma$= eval ($R'_{best}$, S, F (G)) //evaluate by Information Gain Algorithm
        if ($\gamma$ is better than $\gamma_{best}$)
        $R_{best}$ = $R'_{best}$;
        $\gamma_{best}$ = $\gamma$;
        Else;
        Break and return $R_{best}$;
        End;
        Return $R_{best}$
    Output: Reduct Dataset
```

In the pseudo code, S represents the Original dataset with list of attributes; R_0 is used to start the search by using sequential forward search method. d_0 holds the result of the cardinality of R_0. is the evaluation of the cardinality of the dataset by means of independent measure. is the evaluation of cardinality in the dataset by using Quick Reduct algorithm. Up to M-d attributes, the cardinality is calculated by R from R_{best}. is the evaluation of the by Information Gain algorithm.

4. SIMULATION RESULTS AND DISCUSSIONS

The proposed computational hybrid classification model for IP Spoofing was implemented in MATLAB. Amid the assessment, 10 percent named information of KDDCUP 99 was utilized for preparing the proposed hybrid classification model for IP spoofing. This dataset contains three sorts of traffics and six sorts of DoS assault around four gigabytes and every traffic record has 41 features names whose values facilitate to recognize the sort classification either as ordinary or assault. It contains a sum of 24 assault sorts that fall into four noteworthy classifications, for example, Test, R2L (Remote to User), U2R (User to Root) and DoS (Denial of Service). DoS assaults are hard to manage on the grounds that they are anything but difficult to dispatch, hard to track furthermore it is difficult to reject the solicitations of the attacker. PoD (Ping of Death), Teardrop, Neptune (Syn Flood), Land, Back and smurf are the six sorts of DoS assaults in KDDCUP 99 [10]. Back sort of denial of service assaults against the Apache web server, an attacker submits demands with URL's containing numerous front cuts. As the server tries to prepare these solicitations it will back off and gets unable to proceed other request. Back assault needs to realize that solicitations for archives with more than some number of front cuts in the URL ought to be viewed as an assault. In the "smurf " assault, attackers use ICMP echo demand packets coordinated to IP broadcast addresses from remote areas to make a denial of service assault. The Land assault happens when an aggressor sends a spoofed SYN packet in which the source location is the same as the destination address. Teardrop happens because of IP fragmentation re-assembly code which does not legitimately handle covering IP fragments. This assault by searching for two extraordinarily fragmented IP datagram.

The principal datagram is a 0 counterbalance section with a payload of size N, with the MF bit on (the information substance of the parcel is unimportant). The second datagram is the last section (MF = 0), with a positive balance more prominent than N and with a payload of size not as much as N. Neptune assault depicts that every half-open TCP connection made to a machine causes the "tcpd" server to add a record to the information structure that stores data about every single pending connection. This information structure is of limited size, and it can be made to flood by purposefully making excessively numerous incompletely open communication. Neptune assault can be recognized from ordinary network traffic by searching for various synchronous SYN packets bound for a specific machine that are originating from an inaccessible host. A host-based intrusion detection framework can screen the extent of the tcpd connection information structure and alarm a client on the off chance that this information structure nears its size point of confinement. Ping of Death assault has been accounted for when the framework respond in a flighty manner while getting larger than average IP packets. Conceivable responses incorporate smashing, solidifying and rebooting. Ping of Death can be distinguished by noticing the span of all ICMP packets and flagging those that are longer than 64000 bytes.

In this proposed model, the concealed related data from the features was watched. Learners talked about among others, about conceivable potential varieties in movement records which understand the earlier information of abnormal practices ahead of time. This proposed highlight determination system encourages brief recognition and refinement of conceivable individual movement records from group.

Table 1: Simulation Result of Hybrid Feature Selection Method

Attack Type	*Original Dataset*	*Information Gain Technique*	*Quick Reduct Rough Set Technique*	*Proposed Feature Selection (IG + Quick Reduct*
Normal	97, 277	97,277	97,277	97,277
Tear Drop	979	805	762	698
Back	2203	1985	1754	1685
Smurf	2,80,790	60451	10789	1024
Pod	264	189	157	123
Neptune	107201	50421	15478	845
Land	21	19	17	15

Table 2: Performance Comparison of Quick Reduct, Information Gain and Hybrid Feature Selection Method

Performance Metrics	*Quick Reduct*	*Information Gain*	*Proposed Hybrid Feature Selection Method(RR+IG)*
Deduction (in %)	86%	85%	92%
Execution Time	3.54 mins	2.96 mins	0.98 mins
Error Rate	88%	85%	70%

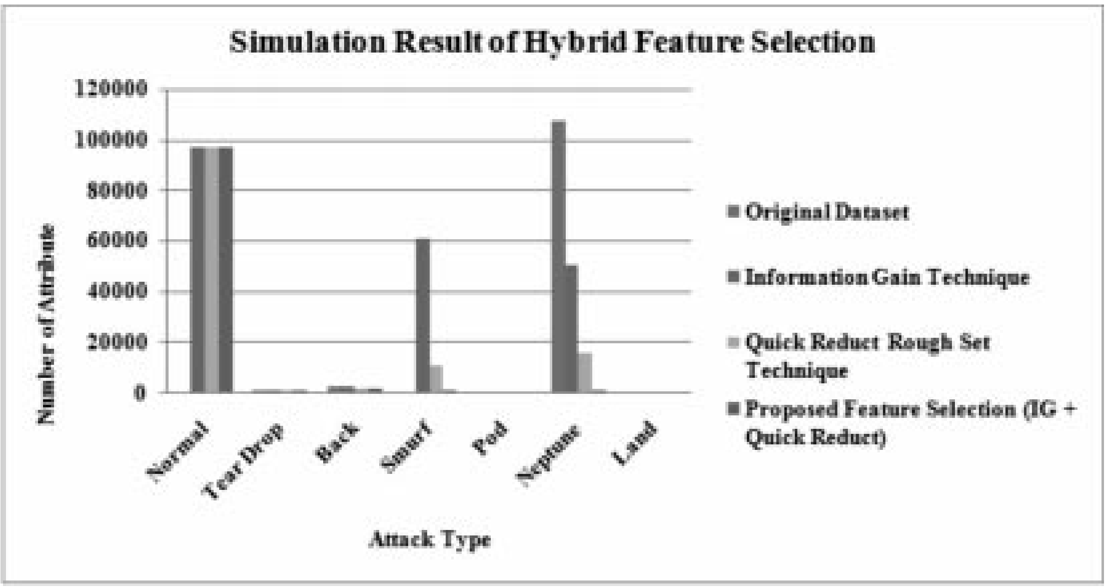

Figure 2: Feature Selection using Proposed Hybrid Feature Selection Algorithm

Table 3: Performance Comparison of Original Dataset, Information Gain, Quick Reduct and Proposed Hybrid Feature Selection Algorithm

	Original Dataset	*Information Gain*	*Quick Reduct*	*Proposed Hybrid Algorithm*
Deduction Accuracy	79.74	80.47	81.01	82.61
Kappa Statistic	0.45	0.47	0.51	0.66
Mean Absolute Error	0.26	0.25	0.23	0.22
Root Mean Squared Error	0.44	0.41	0.37	0.35
Relative Absolute Error	52.57	51.75	48.52	46.25
Root Relative Absolute Error	87.67	85.57	81.23	78.64
True Positive Rate	0.69	0.71	0.75	0.81
False Positive Rate	0.32	0.30	0.25	0.21
Precision	0.69	0.70	0.75	0.81
Recall	0.69	0.70	0.75	0.81
Receiver Operating Characteristic Curve	0.73	0.75	0.77	0.84

To approve the outcomes got from the mixture calculation, the accompanying parameters Kappa Statistic, Performance, False Positive Rate (FPR), True Positive Rate (TPR), Relative Absolute Error (RAE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Relative Absolute Error (RRAE), Mean Squared Error (MSE), Receiver Operating Characteristic Curve (ROC), Confusion Matrix, Precision, Recall are considered. The average square difference between the outputs and targets is Mean Squared Error. If zero it means no error whereas lower values are better. The correlation between targets and output is measured by value called Regression R. The random relationship is indicated by 0 whereas close relationship is given by 1.

The TPR against FPR is considered to plot the ROC curve. The obtained result is considered as good only when the ROC value areas are nearer to the value of 0.80 to 0.90. The predictive model performance is estimated by using Cross Validation technique. In the 10 fold cross validation, the data sets are divided into 10 sets in which 9 data sets are used for training and 1 is used for testing. Table 3 represents the performance of the proposed hybrid feature selection with original dataset, Information Gain and Quick Reduct feature selection methods. The table 4 gives the performance comparison of the proposed hybrid Classification model for IP Spoofing attacks with Fuzzy Hybrid model. From the table 4, it is clear that the proposed hybrid model performs better than the existing Fuzzy Hybrid model.

Table 4: Performance Comparison of Proposed Hybrid Classification model with Existing Fuzzy Hybrid Model for IP Spoofing attacks

	Fuzzy Hybrid Model	*Proposed Hybrid Classification Model*
Deduction Accuracy	81.45	82.61
Kappa Statistic	0.67	0.66
Mean Absolute Error	0.24	0.22
Root Mean Squared Error	0.38	0.35
Relative Absolute Error	47.25	46.25
Root Relative Absolute Error	80.78	78.64
True Positive Rate	0.74	0.81
False Positive Rate	0.25	0.21
Precision	0.74	0.81
Recall	0.74	0.81
Receiver Operating Characteristic Curve	0.81	0.84

6. CONCLUSIONS

In this paper, another calculation hybrid classification model was proposed by hybridizing the quick reduct rough set and information gain for removing the features from the duplications. At that point Naïve Bayes order is utilized to arrange the given set of attributes as Spoofed IP and Non-Spoofed IP. This proposed Hybrid Classification Model are adequate interoperability and high reliability and practically identical with a few surely understood calculations like Artificial Neural Network. Results on the given dataset demonstrate that the proposed hybrid classification model would be fit for arranging the IP address by different sorts of assaults with higher exactness. The consequences of hybrid classification model are better than Artificial Neural Network Classification.

REFERENCES

[1] ICANN Security and Stability Advisory Committee, “Distributed denial of service (DDOS) attacks”, *SSAC, Tech. Rep. SSAC Advisory SAC008*, Mar. 2006.

[2] C. Labovitz, “Bots, DDoS and ground truth”, presented at the *50th NANOG*, Oct. 2010.

[3] The UCSD Network Telescope. [Online]. Available: http://www.caida.org/projects/network_telescope/.

[4] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, “Inferring internet denial-of-service activity” *ACM Trans. Comput. Syst.*, vol. 24, no. 2, 115–139, May 2006. [Online]. Available: http://doi.acm.org/10.1145/1132026.1132027.

[5] R. Beverly, A. Berger, Y. Hyun, and K. Claffy, “Understanding the efficacy of deployed internet source address validation filtering” in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf. (IMC)*, 356–369, 2009.

[6] Danny Roobaert, Grigoris Karakoulas and Nitesh V. Chawla, “Information Gain, Correlation and Support Vector Machines”, *Springer*, 463-470, 2006.

[7] Xiuyi Jia, Lin Shang, Bing Zhou, Yiyu Yao, “Generalized Attribute Reduct in Rough Set Theory*”, Three-way Decisions and Granular Computing, Knowledge Based System- Elsevier,* **91**, January, 204-218, 2016.

[8] Jun Wang, Jiaxu Peng, Ou Liu, “A Classification Approach for Less Popular Webpages based on Latent Semantic Analysis and Rough Set Model”, *Expert Systems with Applications - Elsevier*, **42(1)**, 642-648, January 2015.

[9] Dr. Saurabh Mukherjee, Neelam Sharma, “Intrusion Detection Using Naive Bayes Classifier with Feature Reduction”, *Procedia Technology*, 119-128, 2012.

[10] Shashikant Upadhyay, Rajini Ranjan Singh, “Comparative Analysis based Classification of KDD’99 Intrusion Dataset”, *International Journal of Computer Science and Information Security*, **13(3)**, 14-20, March 2015.