Accepted Manuscript

Firefly algorithm based Feature Selection for Network Intrusion Detection

Selvakumar B , Muneeswaran K

PII: S0167-4048(18)30393-6
DOI: https://doi.org/10.1016/j.cose.2018.11.005
Reference: COSE 1432

To appear in: *Computers & Security*

Received date: 20 April 2018
Revised date: 22 October 2018
Accepted date: 26 November 2018

Please cite this article as: Selvakumar B , Muneeswaran K , Firefly algorithm based Feature Selection for Network Intrusion Detection, *Computers & Security* (2018), doi: https://doi.org/10.1016/j.cose.2018.11.005

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Firefly algorithm based Feature Selection for Network Intrusion Detection

Selvakumar B[1], Muneeswaran K[2]
[1]*selvakumar.b@mepcoeng.ac.in,*[2]*kmuni@mepcoeng.ac.in*
[1,2]*Mepco Schlenk Engineering College, Sivakasi*

Abstract - Network intrusion detection is the process of identifying malicious activity in a network by analyzing the network traffic behavior. Data mining techniques are widely used in Intrusion Detection System (IDS) to detect anomalies. Dimensionality reduction plays a vital role in IDS, since detecting anomalies from high dimensional network traffic feature is time-consuming process. Feature selection influences the speed of the analysis and the proposed work, deploys filter and wrapper based method with firefly algorithm in the wrapper for selecting the features. The resulting features are subjected to C4.5 and Bayesian Networks (BN) based classifier with KDD CUP 99 dataset. The experimental results show that 10 features are sufficient to detect the intrusion showing improved accuracy. The proposed work is compared with the existing work showing promising improvements.

Keywords - Network security, Network Intrusion Detection System (NIDS), Feature Selection, Firefly Algorithm, Mutual Information.

I. INTRODUCTION

Intrusion detection system (IDS) is an important component of secure information systems. Intruders in the network are trying to access to the unauthorized resources in the network. It is highly required to monitor and analyse the activities of the user and the system behaviours. Simply by modifying the configuration of the system parameters, the behaviour of the system could be erratic. Hence the system has to be provided with the features for the periodic monitoring and its behavioural patterns both for normal and abnormal activities.There are two types of IDS [13], which are based on deployment in real time and detection mechanism. The IDS based on deployment is categorized into Host based IDS (HIDS) and Network based IDS (NIDS). HIDS monitors the internal activities of a computing system. NIDS dynamically monitors the logs of network traffic in real time to identify the potential intrusions in a network using appropriate detection algorithms. The IDS based on detection mechanism is categorized into misuse detection, anomaly detection and hybrid IDS. Misuse detection uses the predefined set of rules or signatures to detect known attacks. Anomaly detection builds a normal activity profile to detect unknown attacks by checking whether the system state varies from the established normal activity profile. Hybrid IDS detects known and unknown attacks. Nowadays, all kinds of IDS use the data mining techniques for detecting intrusions. Most of the existing NIDS detect attacks by using all attributes constructed from network traffic. But, not all the attributes are needed for detecting attacks. Reduced number of attributes or features can reduce the detection time and increases the detection rate also. In this work, we combined filter and wrapper based approach to select appropriate features for detecting Network Intrusion. The motivation of the work is in reducing the number of features with improved performance for an uncompromised detection rate. The proposed work focuses on NIDS. Though various techniques exist in the literature for NIDS in terms of selection of features, classifiers, the proposed work concentrates on the Meta heuristic approach called firefly technique for feature selection and C4.5 classifier and compared with Bayesian network classifier.

The remainder of this paper is organized as follows. Section II outlines the related work in the literature. The dataset description is given in the section III and the section IV presents proposed work for intrusion detection. The results and discussions are made in the Section V followed by the concluding remarks in section VI.

II. RELATED WORK

NIDS monitors the network activity based on payload information and statistical features of network traffic. A detailed survey of the existing methods outlining the methods, and their applicability associated with the tools in NIDS is done by Monowar et al. [13]. Also they listed the complete attacks related to the HIDS and NIDS. In addition, they emphasized the need for the extraction of the effective features which play major role in the detection of intruders. The detection methods along with the metrics used to assess the performance of the NIDS was discussed. A neural network based NIDS was proposed by Gowrison et al. [7] along with boosting algorithm with less computational complexity. Also they demonstrated the relation between the combination of features and attacks in the form of grammar [29]. The experiments were conducted on KDDCUP'99. A similar work is also carried out by Weiming et al. [22] with online Adaboost-based parameterized methods.

Unsupervised anomaly detection system for detecting intruders was carried out Jungsuk et al. [9] with the unlabelled data. Despite the advantages, it is still hard to deploy them into a real network environment. To overcome the disadvantages of the clustering based work, Deepak et al. [6] proposed a hybrid approach which is the combination of K-Medoids clustering and Naïve-Bayes classification. In their work, first they applied clustering on all data to form a group and after that applied a classifier for classification purpose to identify intrusion in the network. Data mining techniques were deployed by Vaishali et al. [18] and Uday et al. [3] to detect both known and unknown patterns of attacks.

Due to the various combinations of the features in the network traffic records, the optimization techniques were introduced by the researchers. Revathi et al. [16] carried their work using swarm intelligence technique to solve complex optimization problem and for data preprocessing. Genetic based algorithms were used by Skalak et al. [4] where the random mutation was deployed to select or deselect the features with hill climbing heuristic approach for the IDS. More than one weak classifiers are used by Akhilesh et al. [1] by the ensemble of Artificial Neural Network (ANN) and Bayesian Net with Gain Ratio (GR) feature selection technique for intrusion detection. Principal Component Analysis (PCA) was one of the instrumental methods to select the features. One such approach was used by Keerthi et al. [10] for dimensionality reduction. They carried out experiments with PCA using Random forest and C4.5 classifier algorithms with KDD CUP and UNB ISCX dataset. In their work, classification accuracy obtained by 10 principal components was compared with 41 features using C4.5 classifier.

Filter and wrapper based feature selection method was proposed by Wei et al. [20] and the experiments were conducted on KDD'99 data. In their work, instead of constructing a large number of features from massive network traffic, the authors aim to select the most prominent features and use them to detect intrusions in a fast and effective manner. They first employed feature selection based on filter method and wrapper method. Filter based feature selection uses the Information Gain to select important features based on relevance between an attribute and class and important attributes are selected based on rank. Wrapper based feature selection used some searching methods to select subset of the features and selected subset is evaluated using C4.5 and Bayesian network. However Siva et al. [26] used Genetic search as a searching strategy for wrapper based feature selection to select the optimal subset. But Lei Yu et al [12] created a filter based correlation model to select the features in faster manner without giving up the efficiency. The classifiers based on Support Vector Machines and neural network was used by Sung et al. [2] with selected 13 features. A parallel computing model and a nature inspired feature selection technique was attempted by Natesan et al. [27] proposing an efficient feature selection and classification in order to obtain optimized detection rate. Also Map Reduce programming model is used for selecting optimal subset with low computational complexity. IDS using Rough Set Theory (RST) along with SVM was constructed by Chen et al. where RST was used for selecting the important features [15]. The NSL-KDD dataset which is variation of the KDDCUP'99 was used by Dhanabal et al. [11] whose work differs from the others in terms of the data set usage.

III. DATASET DESCRIPTION

In this work, we used KDD CUP 99 data set [14], which consists of normal and attack types (22 different types).Each record of data is constructed from group of packets measured over 2 second window of a connection established to the same destination. Each data record has 41 features (34-numeric, 4-binary, and 3-nominal). First nine features represent the basic statistical information of the packets over a connection, next thirteen features represent the content of the packets, and another nine features represent the traffic information. The last nine features represent the host based features. There are different types of attack which are entering into the network over a period of time and the attacks are classified into the following four main classes. They are briefly described as follows:

- Denial of Service (Dos): Attacker tries to prevent legitimate users from using a service.
- Remote to Local (R2L): Attacker does not have an account on the victim machine, but tries to gain access.
- User to Root (U2R): Attacker has local access to the victim machine and tries to gain super user privileges.
- Probe: Attacker tries to gain information about the target host.

IV. PROPOSED WORK

Over the past few years, a growing number of research works have applied data mining techniques to various problems. In the proposed work, we have adapted them in intrusion detection system. Fig. 1 illustrates the architecture of the proposed work. Selection of important features is the first step for intrusion detection. Feature selection is the process of selecting a subset of original features according to certain criteria, and is important for high dimension data. Let F be the feature set having 'n' number of features. The different subset of features is coined with the complexity 2^n-1. The collection of feature subsets is denoted as S and given by:

$S = \{S_1, S_2, ...S_{2^n-1}\}$ Where n=41 in KDD cup data set.

The number of subsets is very large and exhaustive. Working with all such subsets and getting an enumerative solution is beyond the practical solution and hence different strategies have to be adapted. The algorithm for feature selection can be grouped into two categories: Filter based feature selection and Wrapper based feature selection [17]. The Meta heuristic firefly algorithm which was originally developed by Xin-She Yang [25] and is included in the wrapper approach, which has not been considered in any of the existing work in NIDS so far. Constructing fewer features also improve the efficiency of network intrusion detection. Though every work concentrated on the IDS with benchmark dataset, Wei Wang et al. [19] constructed the attributes from the real time environment and weighted the attributes using KNN and Principle Component Analysis.

A. Filter based feature selection

Features are evaluated based on the general characteristics of the training data without relying on any mining algorithms. It evaluates subset by their information content either with mutual information or with information gain. We have selected the feature with largest Mutual Information (MI).

The mutual information of two random variables is measured by entropy, which is able to quantify the uncertainty of random variables and scale the amount of information shared by them effectively [24].

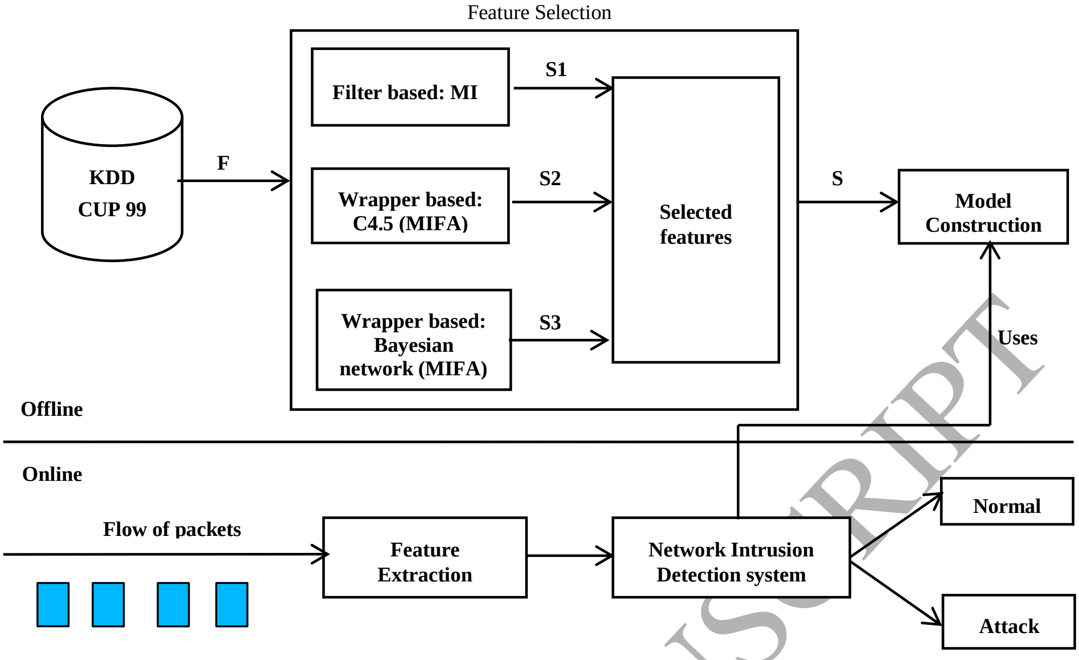

Fig.1 Architecture of proposed work

Let X is a discrete random variable, and its uncertainty can be measured by entropy H(X), which is calculated as follows:

$$H(X) = \sum_i p(x_i)\log_2(p(x_i)) \quad (1)$$

Where the Shannon entropy with probability distribution p(x) for each possible event x ∈ Ω (all possible events). Let Y be the class label of X, and we have the joint entropy H(X, Y):

$$H(X,Y) = -\sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2(p(x,y)) \quad (2)$$

Where p(x, y) is the joint probability distribution function of X and Y. The mutual information I(X; Y) between the variable representing the dataset X and the class labels Y is defined as

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \quad (3)$$

Let S_s be a subset of features on F, and C be the class labels. If the contributed information about the class C provided by the feature F_i∈F having largest mutual information among all the selected features in the subset S_s then the feature F_i is selected. Fig. 2 illustrates the filter based feature selection method.

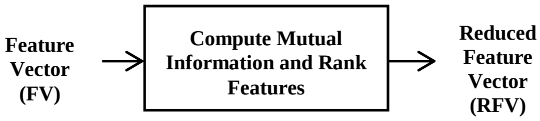
Feature Vector (FV)
Compute Mutual Information and Rank Features
Reduced Feature Vector (RFV)

Fig.2 Filter based feature selection

B. Wrapper based feature selection

It uses a classifier to evaluate subset of features by their predictive accuracy (on test data). The survey paper by Monowar et al [13] discusses many methods for searching the best subset, in which one of the methods is wrapper based feature selection. In our proposed work, Mutual Information Firefly algorithm (MIFA) is used as a feature selection strategy in wrapper based feature selection with C4.5 [8] and Bayesian Network [5] as a classifier. Fig.3 illustrates the wrapper based feature selection method.

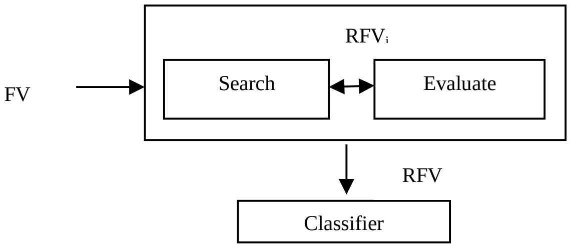

Fig.3Wrapper based feature selection

Nature Inspired meta-heuristic

In general there are two types of stochastic algorithms: heuristic and meta-heuristic. Heuristic means “to find” or “to discover by trial and error” and meta heuristic is the improved version of heuristic and firefly algorithm is one such approach originally developed by Xin-She Yang [25], where it is assumed that the fireflies are attracted to each other by their brightness. The objective function of any optimization problem can be mapped into the brightness of a firefly.

In the firefly algorithm, there are two important issues: the variation of brightness and formulation of the attractiveness. Thus the attractiveness between two fireflies i and j varies with respect to the distance, and brightness which decreases with the distance from its source. One more factor is the absorption coefficient due to the media which influences the attractiveness. Hence the brightness of a firefly at a radius (r) from another firefly with a source brightness B is:

$$B(r) = B_0 e^{-\gamma r} \tag{4}$$

Where B_0 is the original brightness; r is the distance between any two fireflies and γ is a light absorption coefficient which controls the decrease in light intensity. As a firefly's attractiveness is proportional to the brightness seen by another firefly, the attractiveness A of a firefly is given as:

$$A(r) \quad = A_0 e^{-\gamma r} \tag{5}$$

Where A_0 is the attractiveness at r = 0.Then the i^th firefly is attracted by j^th firefly, and the movement is formulated by

$$v_i^{t+1} = A_0 e^{-\gamma r_{ij}^2}(v_j^t - v_i^t) + \beta(R - 0.5) \tag{6}$$

Where β the randomization parameter, and R is a random number generator uniformly distributed between 0 and 1. The term 't' indicates the iteration number. The number of dimensions is D (d=1...D) and r_{ij} is the distance between the i^th firefly and the j^th firefly, which is defined by the equation (7).

$$r_{ij} = \| v_i - v_j \| = \sqrt{\sum_{d=1}^{D}\left(v_{id} - v_{jd}\right)^2} \tag{7}$$

In the proposed work, the number of dimensions (D) is 41 indicating the total number of features related to network intrusion detection. The complexity of the feature selections is 2^D, which is a kind of Non-deterministic Polynomial problem. Hence there is a need for the selection of effective features for reducing the computation complexity and storage for real time deployment. The pseudo code for the selection of features based on both mutual information and firefly algorithm mapped to the NIDS is given in the algorithm**.**

In this work for the evaluation of the feature selection at the intermediate stages, we have used C4.5 and Bayesian Network. Each firefly is represented as a binary vector with D number of features and is denoted by v_i= (v_{i1},v_{i2},v_{i3}, v_{id}…,v_{iD}), i=1…n where 'n' is the number of fireflies. Each element in v_i is limited to 0 or 1 indicating whether that traffic feature is selected or not. In other words, each firefly v_i is positioned as a point in D-dimensional vector space. The subset of feature set (41 features) is represented by different combinations of the presence of 0 or 1 in the feature set.

Each firefly is moving in a direction in the searching space to find the optimal feature subset based on the accuracy of the classifier model with the selected subset of features. The accuracy of the evaluator (classifier model) involving the selected feature is considered as an objective function or brightness of the firefly. The firefly having less accuracy/Brightness will move towards the higher accuracy/brightness using the Eq.6 and the distance between the two fireflies is calculated using the Eq.7. The resulting number of features due to the firefly algorithm is varying. In order to have the fixed number of features for effective implementation in the Mutual Information based Firefly Algorithm (MIFA), adaptive strategy is proposed in the current work, which is the novelty of the proposed work in NIDS. It is controlling the process of the adding or removing the resulting features by Firefly algorithm making fixed number of features. Long Zhang et.al [28] worked on the selection of features using firefly algorithm for various benchmark data set without fixing the number of features needs to be selected. In the proposed work, the number of features to be selected is fixed at k. If the resulting number of features |v_d==1| say 'm' is less than k, then (k-m) number of remaining features are added to it based on the mutual information (MI) from the unselected features. It m is greater than k, the MI for the resulting features is calculated and (m-k) number of features with lowest MI are removed. This strategy we mean as Mutual Information based Adaptive Strategy (MIAS).

The position of the i^th firefly in MIFA uses update rule as given in equation (8).

$$v_{id}^{t+1} = \begin{cases} 1 & if \;\; p_{id} > rand \\ 0 & otherwise \end{cases} \tag{8}$$

Where,

$$p_{id} = \frac{1}{1 + e^{v_{id}^t}}$$

v_{id}^{t+1} - represents the current value of d^th feature in the i^th firefly, v_{id}^t - represents the previous value of d^th feature in the i^th firefly and rand is a uniformly distributed random number between 0 and 1.

The Algorithm for Mutual information based firefly algorithm (MIFA) is given as follows, where the algorithm is iterative and runs for T_{max} times.

Algorithm MIFA(n,L,C,k,T_{max})

*//Input:*n - number of fireflies,
//Input: L - Attack class labels, C - Classifier,
// Input: k- number of features to be selected
//Output v - modified position of the firefly

```
// set of selected network flow features
// $T_{max}$ – Maximum number of iterations
    {
    $v_{i,i=1..n} = InitRand\binom{D}{k}$
    $\theta_{i,i=1..n} = evaluateclassifier(v_i)$
    $sort(\theta)$
    $t = 1$
    $while(t < T_{\max})$
    {
      $for\ \ i = 1\ to\ n$
      {
        $for\ \ j = 1\ to\ n\ \ \ // i \neq j$
        {
        $if\,(\theta_i < \theta_j)$
        {
          $move(i, j)\ \ // use\ equation 6$
          $updateAttractivenss(i, j)\ // use eqn.5$
          $v_{id} = update(v_{id})\ \ // use\ equation 8$
          $v_i = MIAS(v_i, k)$   //returns modified v
          //as discussed in section IV
          $\theta_i = evaluateclassifier(v_i)$
        }
        }
      }
     $sort(\theta)$
     $t = t + 1$
    }
    $return(v)$
    }
```

In the proposed work three different types of feature selection strategies are adapted as follows:

- Feature set obtained based on Mutual information (S1)
- Feature set obtained by wrapper method MIFA with C4.5 as evaluator (S2)
- Feature set obtained by wrapper method MIFA with Bayesian network as evaluator (S3)

Voting based selection of features is used from these feature sets (one feature is selected from these three sets iff it is available in minimum two sets) as in equation (9)

$$S = \{ f : f \in ((S1 \cap S2) \cup (S1 \cap S3) \cup (S2 \cap S3)) \} \qquad (9)$$

The final resulting feature set is used as input to the C4.5 classifier.

V. Results and Discussions

A. *Experiments based on KDD CUP 99*

In this paper KDD CUP 99 data set is used for experimental setup, which is one of the popular dataset for intrusion detection. As mentioned, records are well labelled as either normal, or as an exact type of attack in NSL-KDD. Table I describes the distribution of the attack samples which is used in our experiment.

TABLE I. Data Set description

Attack Category	Types	Training Size	Testing Size
DoS	Normal	40,000	40000
	smurf	10000	10000
	neptune	5000	5000
	back	1000	1203
	land	10	11
	teardrop	100	579
	pod	400	164
	Subtotal	**56510**	**56957**
Probe	Normal	40000	40000
	satan	800	789
	portsweep	500	540
	nmap	110	121
	ipsweep	600	647
	Subtotal	**42010**	**42097**
R2L	Normal	40000	40000
	ftp_write	4	4
	guess_passwd	23	30
	multihop	7	5
	imap	23	4
	warezclient	520	500
	warezmaster	10	10
	phf	4	0
	spy	2	0
	Subtotal	**40573**	**40553**
U2R	Normal	40000	40000
	buffer_overflow	15	15
	rootkit	4	6
	loadmodule	4	5
	Perl	0	3
	Subtotal	**40023**	**40029**

A. *Results of proposed work*

In KDD CUP 99 dataset all 22 types of attack are not equally distributed. This may degrade the performance of intrusion detection. To avoid impact on unbalanced data distribution we form the training data and test data, which are described in Table I. Table II shows an important features selected by filter and wrapper based MIFA methods. Table III shows a set of features selected by the proposed voting method for various types of attacks. It represents only 10 features are sufficient for detecting intrusion. It is observed from the table II, there are some overlapping features between the proposed method and the existing methods and are highlighted. In most of the cases, the proposed work has unique features compared to the existing methods. In mutual information based firefly algorithm α=0.1 (randomization parameter), B_0=1 (Base attraction), γ=1 (Absorption coefficient), n=10 (number of fireflies), T_{max}=100 (maximum number of iteration) are the initial parameters to the algorithm.

TABLE II. Important features for detecting all type of attacks using different methods

Type	Methods	Important features selected
DOS	MI	$\mathbf{f_{41}},\mathbf{f_{40}},\mathbf{f_{13}},\mathbf{f_{10}},\mathbf{f_5},\mathbf{f_6},\mathbf{f_{23}},f_{28},\mathbf{f_{24}},\mathbf{f_{27}}$
	Wrapper(C4.5)	$f_2,f_3,\mathbf{f_5},\mathbf{f_6},f_{11},\mathbf{f_{12}},\mathbf{f_{23}},\mathbf{f_{24}},\mathbf{f_{27}},\mathbf{f_{41}}$
	Wrapper(BN)	$f_1,\mathbf{f_5},\mathbf{f_{12}},f_{22},\mathbf{f_{23}},f_{25},\mathbf{f_{27}},f_{31},f_{34},\mathbf{f_{40}}$
PROBE	MI	$\mathbf{f_{41}},\mathbf{f_{28}},\mathbf{f_{27}},f_{40},\mathbf{f_5},\mathbf{f_6},\mathbf{f_{33}},f_4,f_{35},\mathbf{f_3}$

	Wrapper(C4.5)	$f_{1}, \mathbf{f_{2}, f_{3}, f_{5}}, f_{10}, f_{16}, f_{31}, f_{39}, \mathbf{f_{40}, f_{41}}$
	Wrapper(BN)	$\mathbf{f_{2}}, \mathbf{f_{5}, f_{6}}, f_{19}, f_{22}, f_{26}, \mathbf{f_{27}}, f_{29}, \mathbf{f_{31}}, f_{38}$
R2L	MI	$\mathbf{f_{41}}, \mathbf{f_{40}}, f_{27}, \mathbf{f_{28}}, f_{3}, f_{33}, \mathbf{f_{5}, f_{6}}, f_{11}, \mathbf{f_{24}}$
	Wrapper(C4.5)	$\mathbf{f_{5}, f_{6}, f_{7}, f_{13}}, f_{14}, f_{18}, f_{21}, \mathbf{f_{22}}, \mathbf{f_{25}, f_{28}}$
	Wrapper(BN)	$\mathbf{f_{5}, f_{6}, f_{7}, f_{13}}, f_{15}, \mathbf{f_{22}, f_{24}, f_{25}}, f_{32}, f_{36}$
U2R	MI	$\mathbf{f_{41}}, \mathbf{f_{27}, f_{28}, f_{40}, f_{33}}, \mathbf{f_{3}, f_{5}, f_{6}}, f_{24}, f_{23}$
	Wrapper(C4.5)	$\mathbf{f_{3}, f_{5}}, f_{8}, f_{13}, f_{14}, \mathbf{f_{15}}, f_{16}, \mathbf{f_{25}}, f_{35}, \mathbf{f_{40}}$
	Wrapper(BN)	$\mathbf{f_{5}}, f_{10}, f_{11}, \mathbf{f_{15}}, f_{20}, \mathbf{f_{25}}, f_{26}, f_{29}, f_{32}, f_{39}$

TABLE III. IMPORTANT FEATURE SELECTED BY OUR PROPOSED METHOD

Attack Type	**Important features selected**
DOS	$f_{5}, f_{6}, f_{10}, f_{12}, f_{13}, f_{23}, f_{24}, f_{27}, f_{40}, f_{41}$
PROBE	$f_{2}, f_{3}, f_{5}, f_{6}, f_{27}, f_{28}, f_{31}, f_{33}, f_{40}, f_{41}$
R2L	$f_{5}, f_{6}, f_{7}, f_{13}, f_{22}, f_{24}, f_{25}, f_{28}, f_{40}, f_{41}$
U2R	$f_{3}, f_{5}, f_{6}, f_{15}, f_{25}, f_{27}, f_{28}, f_{33}, f_{40}, f_{41}$

All the experiments are conducted on a computer with 3.00 GHZ i5 CPU and 8.00GB RAM memory. The experiments are conducted with the selected features (10 numbers) and all 41 features. Many classifiers such C4.5, Naive Bayes, Bayesian Network and Random Forest and the promising results are obtained for C4.5 and Bayesian Network. The accuracy of the classifier on the resulting features set using C4.5 and Bayesian network is compared and shown in Table IV. It is observed from the table that, the proposed method shows improved performance compared to the classification with all 41 features. Also the false alarm rate, F-measure of the proposed work are compared and shown in Table V and VI exhibiting improved results.

TABLE IV. COMPARISON OF ATTACKS DETECTION RATE BY C4.5 AND BN CLASSIFIER WITH IMPORTANT 10 FEATURES AND 41 FEATURES

AttackType	Methods	DR	
		With 41 features	With 10 features
DOS	BN	99.78	**99.95**
	C4.5	99.95	**99.98**
PROBE	BN	87.74	**93.42**
	C4.5	63.04	**63.85**
R2L	BN	99.90	97.83
	C4.5	92.95	**98.73**
U2R	BN	75.86	68.97
	C4.5	31.03	17.24

TABLE V. COMPARISON OF ATTACKS FALSE POSITIVE RATE BY C4.5 AND BN CLASSIFIER WITH IMPORTANT 10 FEATURES AND 41 FEATURES

Attack Type	Methods	FPR	
		With 41 features	With 10 features
DOS	BN	0.06	**0.01**
	C4.5	0.02	0.03
PROBE	BN	0.05	**0.01**
	C4.5	0.04	**0.00**
R2L	BN	0.018	**0.01**
	C4.5	0.00	**0.00**
U2R	BN	0.29	**0.00**
	C4.5	0.00	0.00

Comparison of attacks detection performance by c4.5 and Bayesian network classifier with 10 features and 41 features are also shown in Table VI, VII, and VIII. The values with bold font in the table mean that attack detection performance with 10 features gives better performance than that of 41 features.

TABLE VI. COMPARISON OF ATTACKS F-MEASURE BY C4.5 AND BN CLASSIFIER WITH IMPORTANT 10 FEATURES AND 41 FEATURES

Attack Type	Methods	F-Measure	
		With 41 features	With 10 features
DOS	BN	0.93	**0.99**
	C4.5	0.97	0.97
PROBE	BN	0.63	**0.92**
	C4.5	0.52	**0.76**
R2L	BN	0.61	**0.74**
	C4.5	0.96	**0.99**
U2R	BN	0.26	**0.44**
	C4.5	0.47	0.29

Comparison of attacks training time by c4.5 and Bayesian network classifier for the selected 10 features with 41 features is shown in Fig.4 and Fig.5. It shows that time taken to build a model with 10 features take less time than building model with 41 features. Comparison of detection time by c4.5 and Bayesian network classifier for the selected 10 features with 41 features is shown in Fig.6 and Fig.7. Time taken to detect the intrusion with 10 features takes less time than that with 41 features. As a result of feature selection the computing time during both training and testing is saved. A recent work by Chuanlong Yin et.al [30] on deep learning approach for intrusion detection using recurrent neural network worked with the same KDD cup data set and the accuracy for the attacks DoS, Probe, R2L and U2R is shown to be 83.5%, 24.7%, 11.5% and 83.4% respectively. The false positive rate is shown to be 2.1, 0.8, 0.1, and 2.2 for these attacks. In the proposed work both the accuracy and the false alarms are improved. The improved results for the accuracy are: 99.98%, 93.42%, 98.73%, 68.97% and the improved false positive rates are: 0.01, 0.01, 0, and 0 respectively for the attacks DoS, Probe, R2L, and U2R.

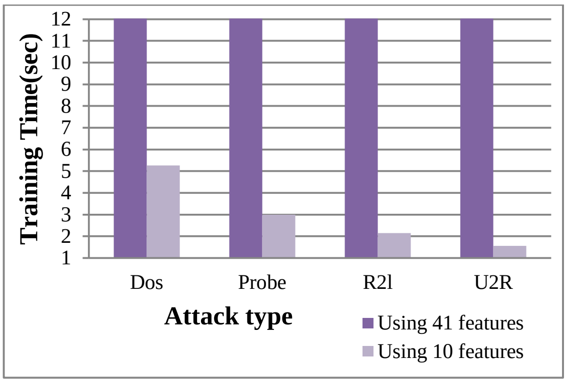

Fig. 4 Comparison of attacks Training time by C4.5 classifier for the selected 10 features and 41 features

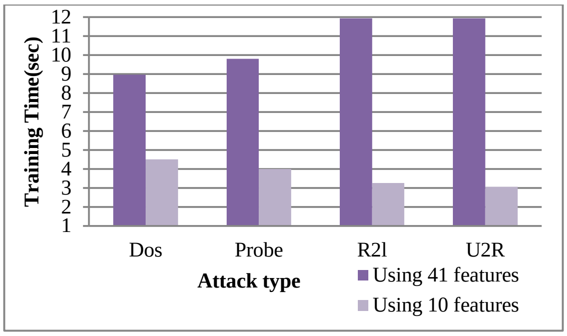

Fig. 5 Comparison of attacks Training time by BN classifier for the selected 10 features and 41 features

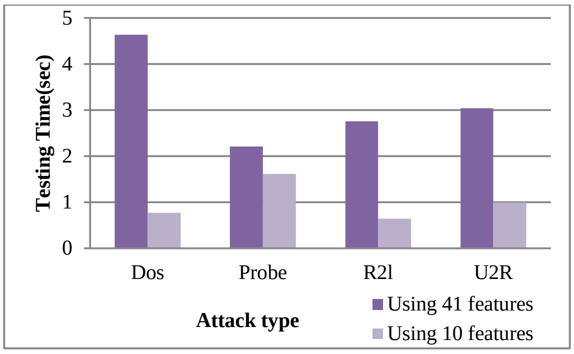

Fig. 6 Comparison of attacks testing time by C4.5 classifier for the selected 10 features and 41 features

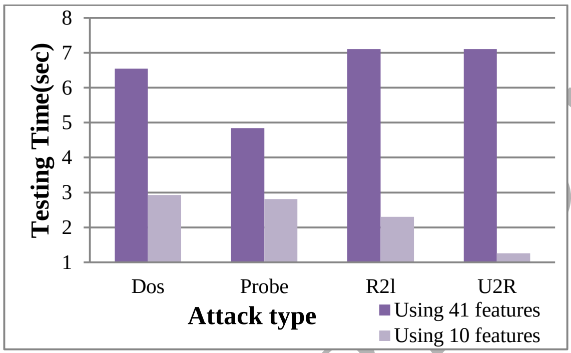

Fig. 7 Comparison of attacks testing time by BN classifier for the selected 10 features and 41 features

VI. Conclusion and Future work

One of the most challenges of network intrusion detection is to handle massive data for intrusion detection. Detection rate of the NIDS is based on number of samples as well as number of features. Reducing dimensionality, increasing the detection accuracy and reducing the false positive rate is the crucial task of Data Mining techniques for intrusion detection. Most existing method fails and used all or most of 41 features to identify intrusion in the network and based on KDD CUP 99 and NSL-KDD dataset. In this work we proposed a new feature selection algorithm for feature selection using KDD CUP 99 dataset. We selected the appropriate features from total number of features (41) for detecting intrusion in the network. Several feature selection methods, based on Mutual Information (MI) and wrapper with Bayesian network, C4.5 are used for feature selection. With only the most appropriate 10 features, the detection performance is better than with 41 features and reducing the computational cost for the classifier. The detection efficiency is improved with appropriate features. Our proposed technique for feature selection is producing better result rather than existing method for feature selection. The extended work is in progress using GPU facilities to decrease the time taken for the computation and improved results.

References

[1] Akhilesh Kumar Shrivas and Amit Kumar Dewangan,"An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set," International Journal of Computer Applications (0975 – 8887) Vol. 99 – No.15,August 2014.

[2] Andrew H. Sung, SrinivasMukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," Sympon Applications and the Internet, 2003.

[3] B.UdayBabu,C.G.Priya and Vishakh,"Survey on intrusion detection techniques using data-mining domain," IJERT, 2014. Vol. 3.

[4] David B.Skalak, "Protopype and feature selection by sampling and Random Mutation Hill Climbing algorithms".

[5] David Heckerman,"A Tutorial on Learning with Bayesian Networks," Microsoft Research, Technical Report MSRTR-95-06, March 1995.

[6] Deepak Upadhyaya and Shubha Jain, "Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification," IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, pp 231-236, May 2013.

[7] G.Gowrison,K.Ramar,K.Muneeswaran and K.Revathi, "Minimal complexity attack classification intrusion detection system," Appl. Soft Comput., 2013, 13, (2), pp. 921–927 .

[8] J.Ross Quinlan."C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, 1993.

[9] Jungsuk Song, HirokiTakakurb, yasuo Okabe, and Koji Nakao,"Toward a more practical unsupervised anomaly detection system," Inf. Sci., 2013, 231, (10), pp. 4–14.

[10] K. KeerthiVasan and B.Surendiran, "Dimensionality reduction using Principal Component analysis for network intrusion detection," Elsevier, 2016.

[11] L.Dhanabal and Dr.S.P. Shantharajah,"A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," IJARCCE,Vol. 5,6,June 2015.

[12] Lei Yu, Huan Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," ICML, 2003, pp. 856–863.

[13] Monowar H. Bhuyan,D.K.Bhattacharyya and J.K. Kalita "Network anomaly detection: Methods,systems and tools," IEEE Commun. Surv. Tutor. 2014, 16, (1), pp. 303–336.

[14] NSL-KDD data set,"https://github.com/defcom17/NSL_KDD".

[15] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set And Support Vector Machine For Network Intrusion Detection," International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009.

[16] S. Revathi and A. Malathi, "Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques," International Journal of Computer Applications , Volume 75– No.6, August 2013.

[17] SwathiV.Jadhav, VishwakamaPinki, "A survey on feature selection methods for High dimensional data," IJRITCC, 2016, pp. 83-86.

[18] Vaishali B Kosamkar and Sangita S Chaudhari, "Data Mining Algorithms for Intrusion Detection System: An Overview," International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS), 2012.

[19] Wei Wang, Xiangliang Zhang and Sylvain Gombault "Constructing attribute weights from computer audit data for effective intrusion detection," J. Syst. Softw., 2009, 82, (12), pp. 1974–1981.

[20] Wei Wang, Yongzhong He, Jiqiang Liu and Sylvain Gombault,"Constructing important features from massive network traffic for lightweight intrusion detection," IET, 2015, pp. 374-379.

[21] Wei Wang,SylvainGombault,”Efficient detection of DDoS attacks with important attributes,” CRiSIS, 2008, pp. 61–67.
[22] Weiming Hu, Jun Gao, Yanguo Wang, Ou Wu, and Stephen Maybank, “Online adaboost-based parameterized methods for dynamic distributed network intrusion detection,” IEEE Trans. Cybern., 2014, 44, (1), pp. 66–82.
[23] Wenke Lee and Salvatore J. Stolfo, “A framework for constructing features and models for intrusion detection systems,” ACM Trans. Inf. Syst. Sec., 2000, 3, (4), pp. 227–261.
[24] Cover TM, Thomas JA (2006) Elements of information theory (Wiley series in telecommunications and signal processing). Wiley-Interscience, London.
[25] X.-S. Yang,“Firefly algorithm, Levy flights and global optimization”, in: Research and Development in Intelligent Systems XXVI (Eds M. Bramer, R. Ellis, M. Petridis), Springer London, pp. 209-218 (2010)
[26] Siva S., Sivatha Sindhu,Geetha S.,Kannan a.,“Decision tree based light weight intrusion detection using a wrapper approach”, Elsevier, Expert Systems with Applications,pp. 129-141,2012.
[27] Natesan P., Rajalaxmi R.R., and Gowrison G., “Hadoop based parallel Binary Bat Algorithm for Network Intrusion Detection”, Springer,Int J Parallel Prog,PP. 1-20,2016.
[28] Long Zhang, LinlinshanandJianhuaWang,“Optimal feature selection using distance–based firefly algorithm with mutual Information criterion”,Springer,2016.
[29] G. Gowrison., et.al “Efficient context-free grammar intrusion detection system” International Journal of Innovative Computing, , Volume 7, Number 8, pp.1-20, August 2011
[30] Chuanlong Yin , Yuefei Zhu, Jinlong Fei, And Xinzheng He, A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks, DOI: 1 0.1109/ACCESS.2017.2762418, 2017

B. Selvakumar was born in the year 1985 at Sivakasi, Tamilnadu, India. He completed B.Tech Degree in Information Technology from Kamaraj College of Engineering and Technology, Virudhunagar, India in the year 2006 and M. E Degree in Computer Science and Engineering from Mepco Schlenk Engineering College, Sivakasi, India in the year 2011. His area of interest includes Network Security and Cryptography. He has got eleven years of teaching experience. He has authored or co-authored about 5 publications in journal/conference level. He is currently working as Assistant Professor in the Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India and he is a Life Member of Indian Society for Technical Education (ISTE)

Muneeswaran Karuppiah received the bachelor of engineering degree in Electronics and Communication engineering from Madurai Kamarajar University, Tamilnadu, India in 1984 and the master of engineering in computer science and engineering from Bharathiyar University, Tamilnadu, India, in 1990. In 2006, he received the Ph.D. degree in computer science engineering from M.S. University, Tamilnadu, India. He is in teaching and research for the past 34 years and 17 years respectively and currently, he is working as senior professor in Computer Science and Engineering Department at Mepco Schlenk Engineering College, Tamilnadu. His research interests are image analysis, machine learning techniques, and Data Analytics. He has authored or co-authored about 125 publications in journal/conference level and one book on compiler design with Oxford University. He is currently working as Senior Professor in the Department of Computer Science and Engineering at Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India and he is a life member of Computer Society of India (CSI), Indian Society for Technical Education (ISTE) and Institute of

Electronics and Telecommunication Engineers (IETE)