Hybrid Feature Selection for Network Intrusion Detection Using Data Mining

V.Manikandan [1], S.Karthikeyan [2], Ms.T.Bhuvaneswari [3]

Department of computer science and Engineering, National Engineering College, Kovilpatti

DOI: 10.29322/IJSRP.10.05.2020.p101110
http://dx.doi.org/10.29322/IJSRP.10.05.2020.p101110

Abstract - Network intrusion detection is a dynamic and important research area. It is the process of identifying malicious activity in a network by analyzing the network traffic. 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 is time consuming process. Feature selection is one of the prominent dimensionality reduction technique widely used in network traffic analysis. It focuses on selecting the important features from network traffic to identify intrusion. In our proposed work, appropriate features are selected for detecting intrusion in a network. We carried out our experiments on KDD CUP 99 dataset with filter and wrapper based feature selection methods using the classifiers C4.5 and Bayesian Networks (BN). Experiments show that 10 features are sufficient to detect the intrusion which produces promising accuracy with reduced training time. The proposed work compares the IDS with Existing method as feature selection.

Index Terms - Network security, Network Intrusion Detection System (NIDS), Feature Selection, Random Mutation Hill Climbing Algorithm, Information Gain.

I.INTRODUCTION

Intrusion detection system (IDS) is an important component of secure information systems. Intrusions are the violation of information security policy. Intrusion detection functions include: monitoring and analyzing both user and system activities, analyzing system configuration and vulnerabilities, assessing system and file integrity, ability to recognize patterns, types of attacks, analysis of abnormal patterns and tracking user policy violations. There are two types of IDS [13], which are based on deployment in real time and detection mechanism. The IDS based on deployment in real time 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 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 for an uncompromised detection rate.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 presents dataset description. Section 4 presents proposed work for intrusion detection. Section 5 presents proposed feature selection strategy. Section 6 shows Experiment and results. Finally, Section 7 shows conclusion and future work of this paper.

II.RELATED WORK

NIDS monitors the network activity based on payload information and statistical features of network traffic. Monowar et al. [13] described the existing network anomaly detection methods, systems and tools. Gowrison et al. [7] suggested an intrusion detection method using Neural Network and Boosting algorithm with less computational complexity. Jungsuk et al. [9] constructed a more unsupervised anomaly detection system for detecting intrusion. Unsupervised anomaly detection system constructed the intrusion detection system without labelled training data. Despite the advantages, it is still hard to deploy them into a real network environment. Keerthi et al. [10] used Principal Component Analysis (PCA) 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. Weiming et al. [22] proposed online Adaboost-Based parameterized methods for network intrusion detection. 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. Vaishali et al. [18] used various data mining Algorithms to detect both known and unknown patterns of attacks. Revathi et al. [16] introduced a new swarm intelligence technique to solve complex optimization problem and for data preprocessing. Akhilesh et al. [1] proposed ensemble of Artificial Neural Network (ANN) and Bayesian Net with Gain Ratio (GR) feature selection technique for intrusion detection. Wei et al.[20]proposed filter and wrapper based feature selection method based 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 important features and use them to detect intrusions in a fast and effective manner.

The authors 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 method to select subset of the features and selected subset is evaluated using C4.5 and Bayesian network.

Dhanabal et al. [11] analyzed the NSL-KDD dataset for Intrusion detection system based on classification algorithm. Uday et al. [3] survey the intrusion detection technique using Data Mining concepts. Siva et al. [26] used Genetic search as a searching strategy for wrapper based feature selection to select the optimal subset. Natesan et al. [27] proposed an efficient feature selection and classification in order to obtain optimized detection rate. They were used a parallel computing model and a nature inspired feature selection technique. Also Map Reduce programming model is used for selecting optimal subset with low computational complexity.

III.Dataset description

In this work, we used KDD CUP 99 data set [14] and used Information Gain for filter based approach and Wrapper based feature selection with an alternative way to identify the important features with C4.5 and Bayesian network as classifier. KDD CUP 99 data set consist of normal and 22 different type of attacks. Out of 41 features (34-numeric, 4-binary, 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, another nine features represent the traffic information of a group of packets measured over past 2 second window of a current connection established to the same destination. The last nine features represent the host based features.

A. *Attack Types inKDD CUP 99*

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.

- 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, hence 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. Figure 1 illustrates the architecture of our proposed work. Selection of important features is the first step for intrusion detection. In this work, filter and two wrapper based feature selection methods are used for feature selection. Let 's' be the total number of features in the final feature subset S and each method generates different subset of features namely S1, S2, S3. The final subset S contains features that are selected by at least two methods. If the number of features in the final subset S is less than the 's' (Number of features need to be selected) then the feature with high Information Gain is added into final subset S. Otherwise, feature with low information gain is removed from the subset S. Finally, model for intrusion detection is constructed with optimal feature subset S. NIDS uses learned model to detect the nature of the traffic which may be normal or specific type of attack. The experiment results demonstrate that with only the most important 10 features selected from all the original 41 features are sufficient to detect intrusion in a network. Constructing fewer features can also improve the efficiency of network intrusion detection.

V.Feature selection

Feature selection is the process of selecting a subset of original features according to certain criteria, and is important for high dimension data reduction. Feature selection reduces the number of features, removes irrelevant, redundant, or noisy features. The algorithm for feature selection can be grouped into two categories: Filter based feature selection and Wrapper based feature selection [17].

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 (e.g.: Information theoretic measures). Choose the feature with more information gain. Fig. 2 illustrates the filter based feature selection method.

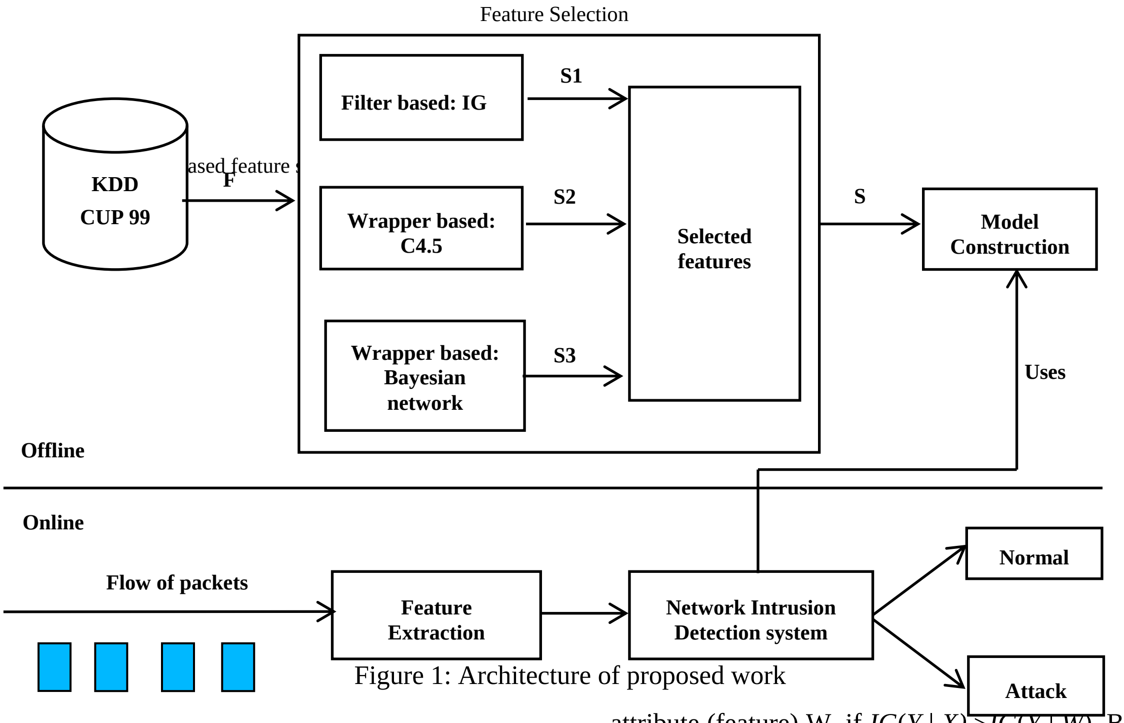


Figure 1: Architecture of proposed work

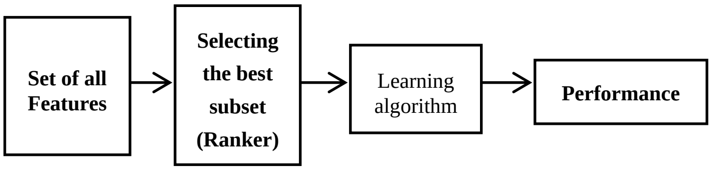


Figure 2: Filter based feature selection

Information Gain (IG):*The information gain of a given* attribute X with respect to the class attribute Y is the reduction in uncertainty about the value of Y, after observing values of X. It is denoted as IG(Y | X). The uncertainty about the value of Y is measured by its entropy defined as,

$$H(Y)=-\sum p(y_i)\log_2(p(y_i)) \qquad (1)$$

Where $P(y_i)$ is the prior probabilities for all values of *Y*.The uncertainty about the value of *Y* after observing values of *X* is given by the conditional entropy of *Y* given *X* defined as,

$$H(Y|X)=-\sum p(x_j)\sum p(y_i|x_j)\log_2(p(y_i|x_j)) \qquad (2)$$

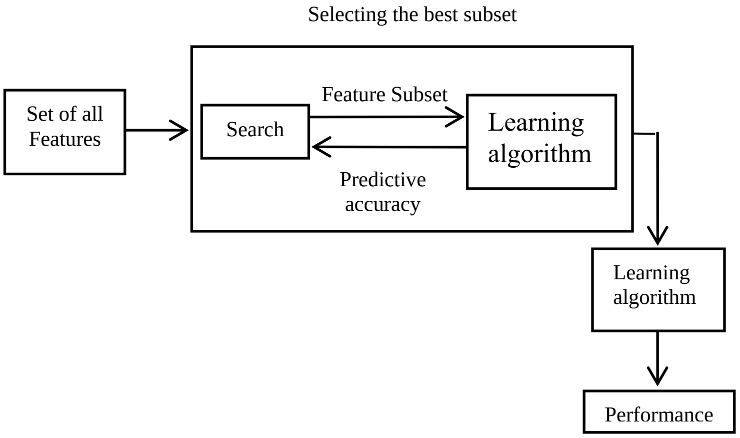
Where $P(y_i|x_i)$ is the posterior probabilities of *Y* given the values of *X*. The information gain is thus defined as

$$IG(Y|X)=H(Y)-H(Y|X) \qquad (3)$$

According to this measure, an attribute (feature) *X* is regarded more correlated to class *Y (attack category)* than attribute (feature) W, if $IG(Y \mid X) > IG(Y \mid W)$. By calculating information gain, we can rank the correlations of each attribute to the class and select key attributes based on this ranking[21].:

B. Wrapperbased feature selection

It uses a classifier to evaluate subsets by their predictive accuracy (on test data). The survey paper by Monowar et al. discusses many methods for searching the best subset as per the survey paper [13]. In the proposed work, Information Gain criterion based Random mutation hill cimbing algorithm (IGCBRHA) is used for wrapper based feature selection with c4.5 and Bayesian Network. Figure 3 illustrates the wrapper based feature selection method.



http://dx.doi.org/10.29322/IJSRP.10.05.2020.p101110

Figure 3: Wrapper based feature selection

Random Mutation Hill Climbing

Random mutation hill climbing is a local search method that has a stochastic component[Papadimitriou and Steiglitz, 1982]. The basic random mutation hill climbing algorithm(RMHC) is as described by Mitchell and Holland[1993].

Random Mutation Hill Climbing algorithm

```
Begin
Objective function f(x), x=(x_1,x_2,x_3,….x_d)T;
Step 1:Choose a binary string at random call this string best evaluated.
Step 2:Mutate a bit chosen at random in best evaluated.
Step 3:Compute the fitness at the mutated string. If the fitness is strictly greater than the fitness of best evaluated. Then set best evaluated to the mutated string.
Step 4:If the maximum number of iteration have been performed return best-evaluated. Otherwise go to step 2.
End;
```

C4.5 based feature selection:The decision tree models are found to be very useful in the domain of data mining since they obtain reasonable accuracy and they are relatively inexpensive to compute. Decision tree classifiers are based on the "divide and conquer" strategy to construct an appropriate tree from a given learning set S containing a set of labeled instances. As a well-known and widely used algorithm, C4.5 algorithm developed by Ross [8] generates accurate decision trees that can be used for effective classification. C4.5 builds decision trees from a set of training data also with the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets.C4.5 examines the information gain ratio (can be regarded as normalized Information Gain) that results from choosing an attribute for splitting the data. The attribute with the highest information gain ratio is the one used to make the decision. Given a learning set S and a non-class attribute X, the Information Gain Ratio is defined as:

$$IGR(S \mid X) = \frac{IG(S \mid X)}{-\sum_{i} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}} \qquad (9)$$

Where S_i is the subset of S for which attribute X have a value and | S | is the number of samples in S. The decision trees are constructed as a set of rules during learning phase. Finally, it is used to predict the classes of new samples based on the rules [21].

Bayesian network based feature selection:A Bayesian network is used to model a domain containing uncertainty in some manner [5]. It is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. In intrusion detection, for example, a Bayesian network could represent the probabilistic relationships between feature sets and types of intrusions. Given a feature vector of a sample, the Bayesian network is used to compute its probabilities of the presence of various classes (normal or individual type of attacks). Formally, Bayesian networks are Directed Acyclic Graphs (DAG) whose nodes represent variables, and whose arcs represent conditional dependencies between the variables. Each node contains the states of the random variable that it represents and a Conditional Probability Table (CPT). The CPT of a node contains probabilities of the node being in a specific state given the states of its parents. The variable represented by the child node is dependent on the ones represented by its parents. There are some efficient algorithms that can be used to perform inference and learning in Bayesian networks [5].Suppose there is an arc from node A to another node B, A is called a parent of B, and thus B is a child of A. The set of parent nodes of a node Xi is denoted by parents (Xi). A directed acyclic graph is a Bayesian Network relative to a set of variables if the joint distribution of the node values can be written as the product of the local distributions of each node and its parents:

$$P(X_1, X_2, \ldots\ldots X_n) = \prod_{i}^{n} P(X_i \mid parents(X_i)) \qquad (10)$$

Given a training set *S*, learning a Bayesian network is to find a network that best matches *S*. The learned network represents an approximation to the probability distribution governing the training set. For classification, given a test data vector represented with attributes, we use this network to compute its probability based on which for classification [21].

The algorithm for proposed feature selection method is shown as follows,

A. Proposed feature selection Algorithm

```
Input:   //Features in Training set $F = \{f_{i=1,2,3.........d}\}$
         n //Number of features in the Training set
         n1//Number of features need to be selected by S1
         n2//Number of features need to be selected by S2
         n3//Number of features need to be selected by S3
Output:  S =best subset with 's' features where s<D
Begin
Step 1: Initialization: Set S=Ø, S1=Ø, S2=Ø, S3=Ø
Step 2://Filter based feature selection
           S1 = find IG(C, F_i) where, i=1.2….d
           // using equation (3)
           S1=top (S1, n1)
Step 3: Generate S2 //Result of Wrapper based feature selection with C4.5.
Step 4: Generate S3//Result of Wrapper based feature selection with BN.
Step 5: Generate final subset S
```

```
    $S = \{ f : f \in ((S1 \cap S2) \cup (S1 \cap S3) \cup (S2 \cap S3)) \}$
Step 6:
        if (||S||<s) then
          Get features from set S1 (feature with Higher IG)
        else if  (||S|| >s)
          Remove features from set S (feature with low IG)

      Endif
   Return S
End
```

VI. Experiments and Results

A. Experiments based on KDD CUP 99

In this paper KDD CUP 99 data set is used for experimental setup. As mentioned, records are well labelled as either normal, or as an exact type of attack in KDD CUP 99.

Table I: Data described in the experiments

Table I describes the distribution of the attack samples which is used in our experiment.

B. Results of proposed work

In KDD CUP 99 dataset all 22 types of attack are not equally distributed. This may degrade the performance of our proposed work. 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 methods. Table III shows an important features selected by our proposed method to identify various type of attacks. It represents only 10 features are sufficient for detecting intrusion.

Attacks type	**Training set**	**Test set**
Dos	Normal 40,000,smurf 10,000,neptune 5000,back 1000,land 10,pod 100, teardrop 400 . (Total:56510)	Normal 40,000,smurf 10,000,neptune 5000,back 1203,land 11, pod 164, teardrop 579 .(Total:56957)
Probe	Normal 40,000,satan 800,portsweep 500,nmap 110,ipsweep 600. (Total:42010)	Normal 40,000,satan 789,portsweep 540,nmap 121,ipsweep 647. (Total:42097)
R2L	Normal 40,000,ftp_write 4,guess_passwd 23,imap 7,multihop 23,warezclient 520, waremaster 10, phf4,spy 2. (Total:40573)	Normal 40,000,ftp_write 4,guess_passwd 30,imap 5,multihop 4,warezclient 500, waremaster10. (Total:40553)
U2R	Normal 40,000,buffer_overflow 15,rootkit 4,loadmodule 4 . (Total:40023)	Normal 40,000,buffer_overflow 15,rootkit 6,loadmodule 5, perl3.(Total:40029)

Table II: Important features for detecting all type of attacks using different methods

Type	**Methods**	**Important features selected**
DOS	IG	$f_3,f_4,f_5,f_{13},f_{23},f_{29},f_{30},f_{33},f_{34},f_{35}$
	Wrapper(C4.5)	$f_2,f_5,f_9,f_{10},f_{24},f_{25},f_{31},f_{36},f_{37},f_{40}$
	Wrapper(BN)	$f_5,f_6,f_8,f_4,f_{22},f_{40},f_{17},f_{26},f_{11},f_{31}$
PROBE	IG	$f_3,f_4,f_5,f_6,f_{12},f_{27},f_{33},f_{34},f_{35},f_{40}$
	Wrapper(C4.5)	$f_5,f_6,f_{13},f_{14},f_{15},f_9,f_{19},f_{23},f_{36},f_{41}$
	Wrapper(BN)	$f_1,f_{32},f_2,f_8,f_{13},f_{36},f_{20},f_{15},f_3,f_5$
R2L	IG	$f_1,f_3,f_5,f_6,f_{10},f_{23},f_{24},f_{33},f_{36},f_{37}$
	Wrapper(C4.5)	$f_2,f_3,f_9,f_{10},f_6,f_{13},f_{16},f_{33},f_{39},f_{41}$
	Wrapper(BN)	$f_{29},f_{17},f_{11},f_{40},f_{39},f_3,f_{13},f_{35},f_{21},f_{30}$
U2R	IG	$f_1,f_3,f_5,f_{10},f_{13},f_{14},f_{17},f_{33},f_{36},f_{37}$
	Wrapper(C4.5)	$f_5,f_6,f_7,f_9,f_{17},f_{19},f_{20},f_{16},f_{37},f_{39}$
	Wrapper(BN)	$f_{21},f_{12},f_{25},f_{41},f_5,f_{39},f_{20},f_{32},f_1,f_{15}$

Table III: Important feature selected by our proposed method

Type	**Important features selected**
DOS	$f_5,f_{31},f_{40},f_4,f_3,f_{13},f_{23},f_{29},f_{30},f_{33}$
PROBE	$f_5,f_6,f_{13},f_{15},f_{36},f_3,f_4,f_{12},f_{27},f_{33}$
R2L	$f_3,f_6,f_{10},f_{33},f_{13},f_{39},f_1,f_5,f_{23},f_{34}$
U2R	$f_5,f_{17},f_{37},f_{20},f_{39},f_1,f_3,f_{10},f_{13},f_{14}$

All the experiments are conducted on a computer with 1.99GHZ Quad core CPU and 2.00G RAM memory.Comparison of attacks detection rate by c4.5 classifier and BN with proposed method and existing method for feature selection is shown in Table IV. It is observed from the Table IV that with proposed method for feature selection is giving higher Detection rate than Existing method for feature selection.Attacks false positive rate for the proposed method using c4.5 classifier and BN classifier is compared with existing method for feature selection in Table V. It is observed from the Table V that with proposed method for feature selection is giving minimum false positive rate than Existing method for feature selection.Comparison of attacks F-measure by c4.5 classifier and BN with proposed method and existing method for feature selection is shown in Table VI. It shows that our proposed work for feature selection gives better results than existing method[29] to detect different type of attacks.

Table IV: Comparison of attacks Detection rate by C4.5 and bn Classifier

Attack Type	**Methods**	**DR/TPR/Recall**	
		Existing Method	**Proposed Method**
DOS	BN	99.95282	99.94692
	C4.5	99.97641	**99.98231**
PROBE	BN	93.41917	**99.71388**

	C4.5	63.85312	**99.71388**
R2L	BN	97.83002	94.39421
	C4.5	98.73418	**99.63834**
U2R	BN	68.96552	13.7931
	C4.5	17.24138	**62.06897**

Table V: Comparison of attacks false positive rate by C4.5 and bn Classifier

Attack Type	**Methods**	**FPR**	
		Existing Method	**Proposed Method**
DOS	BN	0.006925	**0.005825**
	C4.5	0.02535	**0.00245**
PROBE	BN	0.005075	0.01515
	C4.5	0.002325	0.0061
R2L	BN	0.00925	**0.00225**
	C4.5	0.000025	**0**
U2R	BN	0.001025	**1e-04**
	C4.5	0.000025	**2.5e-05**

Table VI: Comparison of attacks F-Measure by C4.5 and bn classifier

Attack Type	**Methods**	**F-Measure**	
		Existing Method	**Proposed Method**
DOS	BN	0.9916625	**0.99291111**
	C4.5	0.970851	**0.9970302**
PROBE	BN	0.9199343	0.8723404
	C4.5	0.7588552	**0.9435921**
R2L	BN	0.739071	**0.8961373**
	C4.5	0.9927273	**0.9981884**
U2R	BN	0.4444444	0.2162162
	C4.5	0.2857143	**0.75**

Table VII: Comparison of attacks Detection rate by C4.5 and bn Classifier with important 10 features and 41 features

Attack Type	**Methods**	**Accuracy**	
		With 41 features	**With 10 features**
DOS	BN	95.64584	**99.57512**
	C4.5	98.33383	**99.82267**
PROBE	BN	94.76447	**98.54621**
	C4.5	94.22049	**99.40613**
R2L	BN	98.26647	**99.70163**
	C4.5	99.90136	**99.99507**
U2R	BN	99.70163	**99.92755**
	C4.5	99.95004	**99.97002**

Table VIII: 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	0.061075	**0.005825**
	C4.5	0.023525	**0.00245**
PROBE	BN	0.048675	**0.01515**
	C4.5	0.04145	**0.0061**
R2L	BN	0.017575	**0.00225**
	C4.5	2.5e-05	**0**
U2R	BN	0.00225	**1e-04**
	C4.5	0	2.5e-05

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

Table IX: 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	0.9317181	**0.99291111**
	C4.5	0.9727667	**0.9970302**
PROBE	BN	0.6254249	**0.8723404**
	C4.5	0.52078	**0.9435921**
R2L	BN	0.6113875	**0.8961373**
	C4.5	0.09625468	**0.9981884**
U2R	BN	0.8961373	0.2162162
	C4.5	0.4736842	**0.75**

Comparison of attacks accuracy by c4.5 and Bayesian network classifier for the selected 10 features with 41 features is shown in Figure 4 and Figure5.It is observed from Figure4 and Figure5 that with a reduced number of features attacks accuracy is better than that with 41 features.

Comparison of attacks accuracy by c4.5 and Bayesian network classifier for the selected 10 features with 41 features is shown in Figure 4 and Figure5.It is observed from Figure4 and Figure5 that with reduced number of features attacks accuracy is better than that with 41 features.

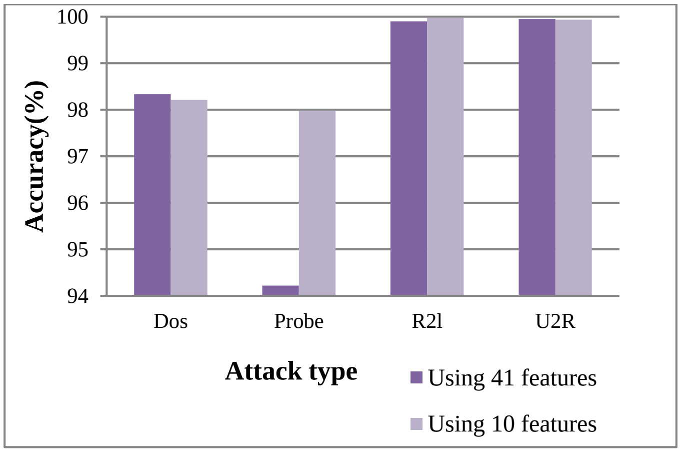


Figure 4: Comparison of attacks accuracy by C4.5 classifier for the selected 10 features and 41 features

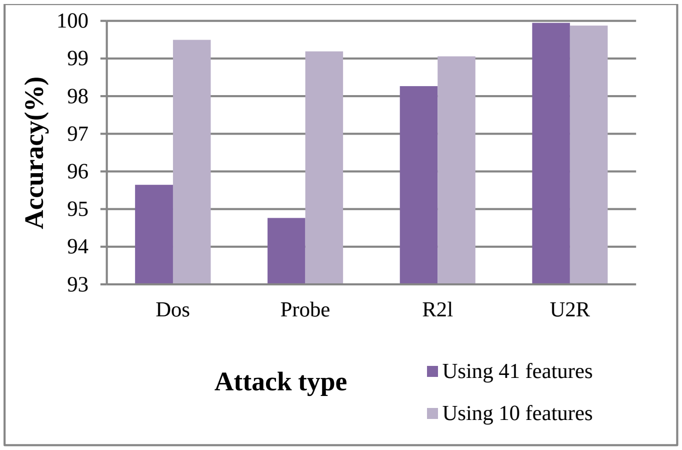


Figure 5: Comparison of attacks accuracy by BN classifier for the selected 10 features and 41 features

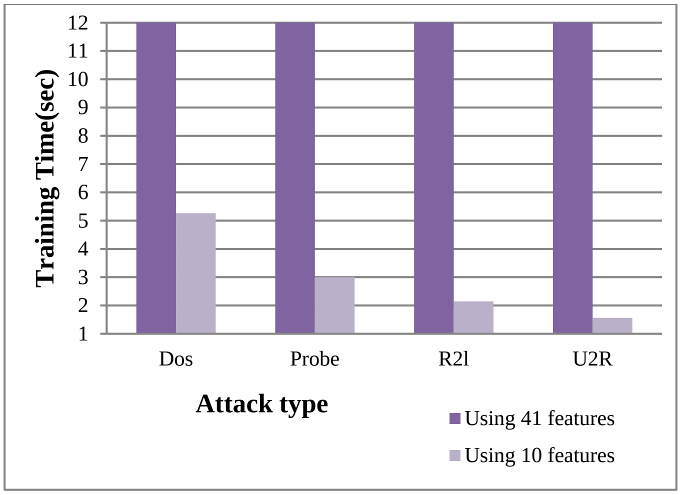


Figure 6: Comparison of attacks Training time by C4.5 classifier for the selected 10 features and 41 features

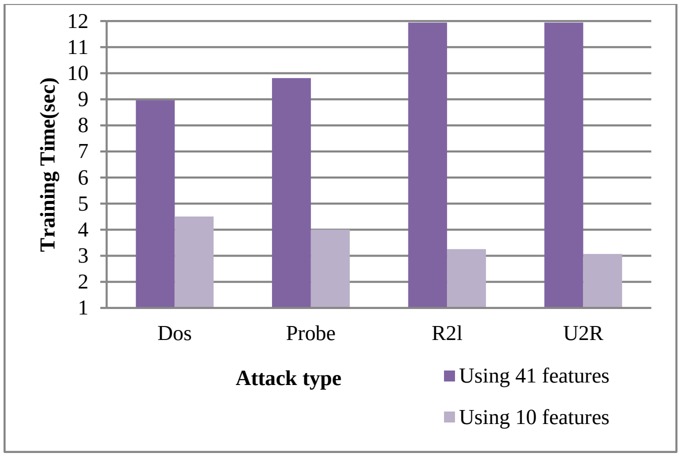


Figure 7: Comparison of attacks Training time by BN classifier for the selected 10 features and 41 features

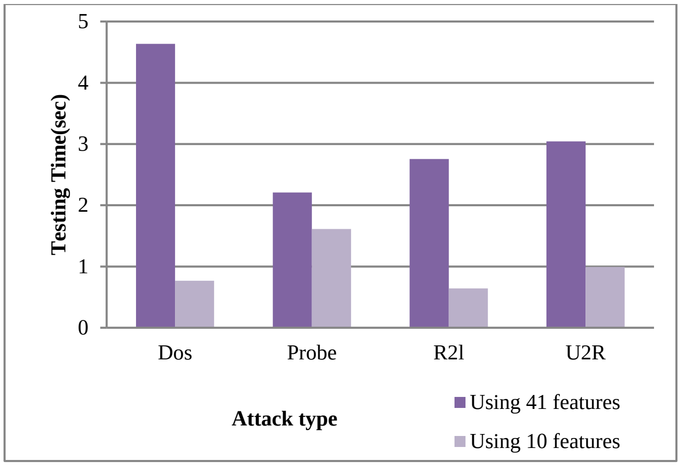


Figure 8: Comparison of attacks Testing time by C4.5 classifier for the selected 10 features and 41 features

Comparison of attacks training time by c4.5 and Bayesian network classifier for the selected 10 features with 41 features is shown in Figure 6 and Figure7.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 Figure 8 and Figure9. Time taken to detect the intrusion with 10 features take less time than that with 41 features. As a result of feature selection the computing time during training and testing is saved.

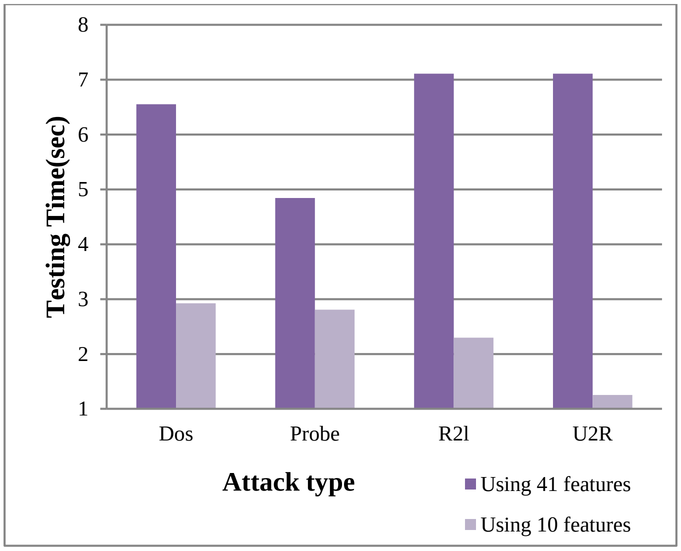


Figure 9: Comparison of attacks Testing time by BN classifier for the selected 10 features and 41 features

VII.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 Information Gain(IG) 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[29]. The extended work is in progress using the bat algorithm for selection of still appropriate features 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] 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

[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] Xin-She Yang,"Nature-Inspired Metaheuristic Algorithms Second Edition",Luniver press,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] Selva Kumar B,Muneeswarn K,Firefly algorithm based feature selection for Network Intrusion Detection,Computers & Security (2018).

Authors

First Author- V.Manikandan,BE., Final Year,National Engineering College, Kovilpatti, manikandanvkvp@gmail.com.

Second Author - S.Karthikeyan,BE., Final Year, National Engineering College,Kovilpatti,karthimamsai@gmail.com.

Third Author - Ms.T.Bhuvaneswari,ME.,Assistant Professor, National Engineering College,Kovilpatti,bhuvaneswari_cse@nec.edu.in.