Journal of Theoretical and Applied Information Technology
15th December 2016. Vol.94. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE SELECTION IN INTRUSION DETECTION, STATE OF THE ART: A REVIEW

[1]HELMI B MD RAIS, [2]TAHIR MEHMOOD
[1,2]Department of Computer and Information Sciences
Universiti Teknologi PETRONAS, Perak, Malaysia
E-mail: [1]helmim@petronas.com.my, [2]tahir_g02756@utp.edu.my

ABSTRACT

With the increase of internet usage the need of security for organizations network also increased. Network anomaly intrusion detection systems are designed to monitor abnormal activity in the network. These systems find the behavior that is deviated from the normal behavior. Network anomaly detection methods are implemented using different approaches including machine learning, data mining, and many more. However, intrusion detection systems highly depend on the features of the input data. These input features give information to the learning algorithms which used in intrusion detection system in the form of the detection method. With irrelevant and redundant features learning algorithm builds detection model with less accuracy rate. Also, ambiguous features increase the time complexity and consume other computational resources as well. By removing these irrelevant and redundant features accuracy of the learning algorithms can be increased. In this paper implementation of different feature selection techniques have been reviewed. Novel feature selection techniques have been developed due to its importance in network intrusion domain. We have discussed some of it in a technical aspect. These techniques are being discussed in detail. Moreover, features from these methods are also given and their results are being. We categorized these techniques according to their implementation. Different comparison of these techniques have been given and been discussed. Moreover, the benchmark dataset that is KDD99 widely used for anomaly detection is also discussed in this paper.

Keywords: *Intrusion Detection System, Intrusion Detection Methods, Feature Selection, KDD99*

1. INTRODUCTION

Today advancement in the internet brings advantages to the organization but at the same time it benevolent the hackers to undermine the security of the internet. There are usually two defense walls on the network; the first one is the firewall that scans and stops the traffic at the port level by using some rules for ports and source, destination IP addresses. The second wall is the network intrusion detection system (NIDS), which analyzes the whole packet including a packet header and payload for any suspicious activity. NIDS is the second line of defense, which let the network administrator to know that some attacks or malicious activities have bypassed the firewall. There are many vendors that provide NIDS products but not a single product can claim 100% protection against all attacks.

Intrusion Detection System (IDS) is designed to detect intrusions in the data. It monitors the data for any suspicious activity. Network anomaly based intrusion detection systems are used to observe the activities that deviated from normal behavior of the network. Network anomaly detection method makes a baseline for the normal activity, any activity that deviates from that baseline is considered as a possible intrusion.

Many algorithms have been applied to develop the anomaly detection model for intrusion detection system. These algorithms, however, highly depend on input features. High dimensional data causes learning algorithm to generate detection model with high time complexity and low accuracy rate. Poor features selection can affect the accuracy of these algorithms badly which leads to errors in the form of false negatives and false positives, which are needed to be minimized. This is the challenging task, and it is the main reason to prevent deployment of intrusion detection systems. The purpose of feature selection is to decrease the computational time for the learning algorithms and enhance the performance of these algorithms through removing redundant and irrelevant features [1]. Good features can lead to higher accuracy of the detection. Features selection improves the learning process, good generalization model, and decreases computational complexity [2].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In this paper different feature selection techniques from previous work have been discussed from network anomaly detection methods perspective.

method. The feature selection methods are discussed in detail. Also the selected features from these methods are also given and their results are being discussed in detail. Section 1 gives a brief introduction to types of intrusion detection system, the methods used in the intrusion detection system, and why feature selection is necessary for intrusion detection method. Section 2 gives previous work related to feature selection method, followed by dataset and evaluation in section 3. The dataset discussed in section 3 is KDD99 dataset which is a benchmark dataset for network anomaly detection. Also feature selection methods that are being discussed in section 2 all used KDD99 dataset. In section 4 Discussion section, discusses the results given by the papers given in section 2. Some issues and challenges are given in section 5. In section 6 paper is concluded in the conclusion section.

1.1. Types of Intrusion Detection System

Intrusion detection systems are software applications that analyze the unauthorized activities in network or system data and generate alerts about intrusions [3]. There are two types of intrusion detection system (IDS) named, Network-based IDS and Host-based IDS [4]. These systems are mainly based on data source [5]. Types of intrusion detection systems are based on where the intrusion detection system is implemented i.e. on network or host.

1.1.1. Host-based Systems

Host-based intrusion detection system (HIDS) deals with the information of a certain single host or computer. The HIDS collects data about the ongoing events in the monitoring system. Host- based intrusion detection system acquires data from system logs or audit trails and other logging information that are generated by the operating system. These log files consist of a sequence of system calls and give useful information about host system [6].

Host-based intrusion detection system should have the capability to collect information in sufficient detail to identify abnormalities in the host. But as the HIDS collects a finer level of detail, it requires significant storage. It also results in increased load on a host system. Thus, Host-based IDS has high cost and results in a tradeoff between time and storage complexity to its reliable detection rate. If the setup of the organization is large, then host- based approach could be economically infeasible. Its cost cannot be reduced as HIDS are not portable because it depends on the configuration of the specific host system as well. Furthermore, to access some information from the operating system needs user privileges to access the information and it is the one of the limitation of such intrusion detection systems.

1.1.2. Network-based Systems

A Network-based intrusion detection system (NIDS) deals with detecting intrusions in network data [7]. These systems gather information from the network itself, instead of each separate host. Information is gathered from the network traffic, as data travels on the network segments. Network- based intrusion detection system detects an attack by examining the content and header information of all packets that are moving across the network. Network intrusion detection system has been considered to be one of the most prominent methods for detection of complex and dynamic intrusion behaviors [8].

As NIDS is implemented on the network, it is transparent to the user of the system, and this is also important for intrusion detection system itself. Transparent monitoring reduces the possibility that the intruder will be able to detect and cancel security defense capabilities of the monitoring system without significant effort. Moreover, NIDS monitors traffic over specific network segments and is not dependent on the operating system this results in enhanced portability. Deployed network- based intrusion detection system detects all attacks, regardless of the implemented operating system. Such system is useful in situations where the network topology changes. However, NIDS does not work well on high-speed networks. It starts dropping packets on high-speed networks. In such condition, NIDS has less enough time to monitor network traffic efficiently and detect the intrusions.

1.2. Techniques Used in Intrusion Detection System

There are two main intrusion detection methods, signature based detection method and anomaly based detection method [9]. These are the techniques that are used by the intrusion detection system to detect intrusions.

Unlike we mainly focused the feature selection

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1.2.1. Signature Based Detection

Signature-based intrusion detection schemes seek defined patterns, or signatures, within the analyzing data [10]. The main advantage of this technique is that the signatures are very easy to develop if the behavior of the network is known.

This approach has been proved to be very effective at detecting known threats but ineffective to detect unknown threats [11]. Since signature based detection method can detect attacks whose signatures are previously stored in the database. Therefore, the signature must be created for every attack, due to this reason signature based detection method cannot detect novel attacks. Fixed behavior pattern as the bases of signature based technique fails to detect those worms, having self-modifying behavior like encrypting themselves. Furthermore, as signature-based detection method has to create new signatures for every variation, the performance of the system degrades significantly. By adding new signature will increase the amount of pattern to be detected and with the passage of time it will result in performance degradation of the detection system.

1.2.2. Anomaly Based Detection

The anomaly based detection method is based on defining the network behavior. If the behavior of the network is according to pre-defined behavior, then it is accepted. Otherwise, it triggers the event for the detection of the abnormal behavior [4]. Anomaly based technique requires detail knowledge of the network or host to build a baseline for normal behavior [12]. The baseline is considered as a threshold for the normal behavior. To detect anomalies correctly, detailed knowledge of accepted network or system behavior is required to be developed carefully. Conversely, the malicious behavior goes unnoticed if the behavior of malicious user falls below the baseline. The main advantage of anomaly based detection method over signature based detection method is the detection of novel attacks for which no signatures exist [13]. This can be seen when the system automatically detects new worms.

1.3. Feature Selection

Features are the trait of the system or object that predict the behavior or state of the system [14]. In other words, features are the distinctive characteristics of the system that predict the behavior, nature or state of the system. Abstractly all data points are considered as features, and can be utilized to predict the state or behavior of the system. Researchers generally extract the dominant features from all the data points that are invariant in nature to achieve robustness of the system. Researchers are aiming to select the features that can enhance the accuracy of the learning algorithm, this process is known as feature selection [15].

Feature selection must have the capability of detection and removal of noisy and misleading features among features.

Feature selection process selects a subset of features that represents the whole feature set [16]. Which features should be included or excluded is being decided in this process. Feature selection techniques hinges on two pillars, relevancy and redundancy of the features [17]. Relevant features are those that predict the desired system response, on the other hand, redundant features have a high degree of correlation among themselves. Thus, removal of the redundant features is desired. The predictive accuracy of the machine learning algorithms can be increased by the feature selection. Robust feature set also reduces the training time of the classifier as robust features are invariant in nature and reduces the dimension in high-dimensional data [18]. Reduced dataset also decreases dataset which acquire less storage space. In network intrusion detection, features are extracted from protocols header at different layers of network architecture and contents of data packets. Due to this reason noise in channels propagates to extracted features, this leads to the false intrusion alarm. There are two types of feature selection methods: Filter and wrapper. Filter method selects the subset of features without involving learning algorithm in evaluation phase and is mainly based on ranking of features, which represents the relevancy of the features [19]. In contrast, wrapper method evaluates a subset of features using learning algorithm [20]. This evaluating algorithm is called iteratively unless a robust subset of the feature is selected. Filter based approach is computationally fast compared to wrapper based as it doesn't involve any learning algorithm during ranking of features. However, wrapper based feature subset accomplishes good accuracy rate as it involves learning algorithm in the subset evaluation phase [21].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. FEATURE SELECTION TECHNIQUE AND INTRUSION DETECTION

Many works have been done for feature selection in intrusion detection domain. This is due to the high dependency of intrusion detection method on input features in the learning phase. Inadequate and noisy features affects the learning algorithm model. The detection model came from the learning of the ambiguous features cannot efficiently detect the attack, if the learning algorithm doesn't get adequate and concise information from the features. In this paper we discussed feature selection technique for network

anomaly detection using machine learning based, rough set theory, and evolution computing. The detection techniques and feature selection techniques are the aspect for the discussion in the following sections.

2.1. Machine learning based

Mukkamala et al. [22] Used SVM for intrusion detection on the kdd99 dataset. For feature selection, SVM was used to identify the most significant features for detecting attacks. Feature reduction was done by deleting one feature at a time, and train SVM with that reduced data set. Those features, whose deletion resulted in more accurate performance (as compared to the original SVM trained with 41 features) was deemed insignificant. It does feature selection by deleting single feature at a time and comparing the result with the whole features. It increases the time for feature selection as it has to test the reduced data for each feature and then comparing the result with the whole feature datasets classification result. Thus, the method consumed more computing memory and required more processing cycle.

Lin et al. [23] Used Simulated Annealing (SA) with Support Vector Machine (SVM) and Decision Tree (DT) together for detection of anomaly and feature selection. Simulated Annealing found the best features whereas Decision Tree was used to get rules from the dataset. The data was then classified using Support Vector Machine. Simulated Annealing was also used to adjust automatically the parameter setting for the DT and SVM. Intrusion detection method was trained and tested on the kdd99 dataset. The method had a very good result for Normal and DOS classes but for the other three attack classes Probe, Remote-to-Local (R2L), and User-to-Remote (U2R), this technique had not good results.

George [24] used Principle Component Analysis (PCA) along with SVM for intrusion detection. PCA was used for dimension reduction (feature selection) while SVM used for classification. Evaluation, based on the SVM with PCA approach, gave less misclassification compared to SVM method. The result was good as compared to using SVM alone, but the result still needs to improve also false negative rate (FNR) was so high.

Authors in [25] used two-phase technique for the classification of KDD99 dataset into normal and anomalous class. The approach used three feature ranking techniques, gain ratio, information gain, and global method for data handling (GMDH) for feature selection. The proposed method used GMDH for the classification as well in the second phase. The abductive network of GMDH was used for the validation for feature subset. The abductive network included monolithic abductive model and ensemble abductive model. Proposed method converted the nominal features to binary. Thus, the total number of features after transformation were

123. After transformation above mentioned three features selection techniques were carried out on the set to get a subset of features. The top common features, among above mentioned feature ranking approaches, were selected for the construction of the GMDH classifier model. The paper claimed to have less training time for feature subset. Less training time was due to the removal of the instances of the not selected features.

In [26] new method called, cluster center and nearest neighbor (CANN) was used to transform the data dimension into a single dimension, which was based on the Euclidean distance. CANN

calculated two different distances for the data point.

One was the distance between data point to the centroid of each cluster. Second distance was calculated by the sum of the distances between that data point to its neighbor data points. Both distances were summed to single new dimension for the data point. The data points with new dimension were validated using K-NN classifier. Before using K-NN, two subset of features were used. One subset consisted of 6 features while the second one consisted of 19 features. These features were taken directly from the work of [27] [28]. The author claimed to have a good detection rate of K- NN with new data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

formation compared to the K- NN result with the original set, but work didn't put the same k value for both cases of data formation. Also, the result was not validated for the whole feature set that makes it difficult to decide whether the reduced feature set as well as new data formation resulted better than whole feature set and original data formation.

2.2. Rough Set Theory

The rough set theory deals with the analysis of uncertain, imprecise and incomplete information [29]. It gives rational about the object represented by features. The rough set theory assumes that every object has some information associated with it, which helps to find the indiscernible objects according to the available information [30]. Many authors use this principle in intrusion detection feature selection and found it facilitating. Few of them are given below,

Ghali [31] used the rough set theory along with the artificial neural network. The aim of the author was to reduce the dataset for intrusion detection which resulted in less consumption of computer resources.

RSNNA (Rough Set Neural Network Algorithm) was used for feature reduction, which found dependencies among the features while feed forward neural network was used for the classification of the data. The result was good as compared to NN result for the full feature set and one hidden layer with four neurons. Moreover, from the results it can depicts that the aim of the methodology was just the reduction of data as the author didn't consider the detection rate for intrusion detection method.

Rassam and Maarof [32] used unsupervised learning method, artificial immune network (aiNet).

The artificial immune network was used for the clustering of the data into two clusters, normal class

and attack classes. Before the clustering features, a selection was done using rough set theory. The motivation of using clustering approach was that as

it is difficult to get the labelled data. Also, the

accurate labelling of the data is a hard task and is a time consuming job. This research used two phases. In phase one, the rough set theory was used, with the help of ROSSETA tool, to find feature subset. This input feature subset was then used in the second phase for aiNet to cluster the data into its respective classes. Feature subset was selected by rough set theory. With the reduced feature set aiNet was used for clustering. But before using aiNet the normalization process was done to convert nominal attributes into discrete values. The result was compared with K-means clustering and was outperformed. The author also used a rough set classifier to see the performance of the classifier with the whole feature set and with the reduced feature set. The classifier trained with the reduced feature set had a better result compared to the result of the classifier trained with the whole feature set.

In [33] enhanced SVM has been proposed. The idea behind is that SVM kernel function treats all features equally, which means that redundant and irrelevant features are treated the same way as other features. This work introduced to give weights to the features according to their importance. The rough set theory was used for this purpose. Important features were given high weights while least features got very low weight and thus deleted. Before applying rough set theory, redundant records were removed and nominal features were converted into the numeric using feature-value format. After getting the feature subset, SVM was trained with the reduced feature set to get detection model. In the experiment along with KDD99 dataset another dataset named, University of New Mexico (UNM), was used. This dataset consists of system calls. Result for the enhanced SVM was compared with the SVM trained with the whole feature set. For KDD99 dataset, the result was comparable but not significant. For the UNM dataset result was substantially good but the same result hold for the SVM trained and tested with the whole feature set.

Shirivastava and Jain [34] used research rough set theory to find the feature subset. The dataset was normalized first into numeric values. Then ROSETTA, rough set theory tool, was used to find feature reduct. Authors used Johnson's algorithm and genetic algorithm two different approaches in ROSETTA to find feature reduct. From Johnson's one feature reduct was obtained while genetic gave

39 feature reduct set from which 4 were selected. All five reduct sets, Johnsons' and genetic, contains six features. This feature subset was evaluated using SVM accuracy rate as well as false positive rate were measured for the original feature set and reduced feature set. The result was good for reduct feature set as compared to the original feature set.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3. Evolution Computation

Evolutionary computation is based on biological inspiration and involves survival of the fittest. Evolution computation involves many techniques like genetic algorithm (GA), ant colony optimization, particle swarm optimization, genetic programming etc. These algorithms are population based and are optimization algorithm, therefore it has been widely adopted for feature selection problem which is an NP-problem. The following literature gives the application of evolutionary computation for feature selection in intrusion detection method.

In [35] Tsang et al. proposed, multi-objective genetic fuzzy intrusion detection system (MOGFIDS) that was applied to improve the accuracy of classification algorithm by making effective interpretability of fuzzy knowledge base rules for classification algorithm. The proposed MOGFIDS applied an agent-based evolutionary computation framework to develop anomaly rule-based intrusion detection. This research addressed the problem that rule-based intrusion detection should have interpretable knowledge that can assist security experts for intrusion analysis. The multi- agent learning system, MOGFIDS, comprised of arbitrator agent (AA) and fuzzy set agent (FSA). Each FSA was an autonomous and intra-behaviors of FSA was done by interpretability-based regulation strategy. This included merging of similar fuzzy sets and removal of fuzzy sets identical to singleton set. The fuzzy sets distribution is comprised of steps included initialization of rule based population, Crossover and mutation, evaluation criteria, and selection mechanism of fuzzy rule set candidate.

Rufai et al. [36] combined membrane computing (MC) and bee algorithm (BA) for their work. Motivated by membrane structure and operations of living cells MC, gives the solution for BA to find the best feature subset. Thus, it improved BA for feature selection. BA was run on different membranes in the main membrane to get the initial solution. The best solution from the individual membrane was collected, mixed and passed to output membrane. At the output membrane, BA was again run for some specific number of iterations. After which the best solution was collected which served as a final solution. Feature subsets were selected for a different run with different fitness accuracy. The feature subset that gave best fitness accuracy was finally selected from a different run. The selected feature subset was validated using SVM classification. The attack detection rate was not outperformed as compared to other research methods, but false alarm rate was remarkably low compared to other feature selection methods.

Zainal et al. [37] used a 2-tier approach, which included rough set and particle swarm optimization (PSO), Rough-PSO. SVM was used for classification while fitness function was used to find out the fitness of the proposed feature subset. The 2-tier structure phase used rough set in coarse phase and particle swarm optimization in granular phase. Coarse phase removed the redundant and irrelevant feature, which was followed by PSO to refine and select most prominent features out of it. The purpose of using rough set before PSO was to reduce the complexity in PSO. The methods result was compared with other research work and proposed method has outperformed among them. Hasani et al. [38] used two evolutionary algorithms, bees algorithm (BA) and linear genetic algorithm (LGP), to find feature subset in their research. The proposed model called, LGP_BA, used in a wrapper approach. First candidate chromosome generated by LGP was used the input to BA, which performed modification using neighborhood search. New generations were performed by LGP that randomly selected features from the KDD99 dataset. Crossover and mutation were applied on each generation. It was performed to categories it with highest fitness values. After which fitness function was applied, and chromosomes that fulfill the condition by the fitness function were passed to BA. BA modified the solution from LGP. It evaluated the fitness solution using queen-bee evaluation. It also helped LGP to converge slowly. After evaluating BA applied crossover and mutation until it found proper fitness value. The best candidate solution was then evaluated by SVM.

Hua et al. [39] used foraging behavior inspired algorithm ant system [40], for feature selection of normal and intrusive data in the kdd99 dataset. Detection was done by classification in which SVM was used to classify the data into their respective classes. Features were represented by graphs and ants select the next node using probability function, which calculated pheromone value of the edge and heuristic information. The Fisher discrimination rate was adopted as the heuristic information in the probability function. A number of ants used was equal to the total number of features i.e. 41 and was positioned on the nodes randomly. Stopping criteria of the ant system was selected using the maximum predefined generations or the global optimal

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

solution. Global pheromone update was used to update the pheromone level on those edges which belonged to the optimal solution i.e. having the less squared error. Features subset was validated using SVM.

In [41], the authors proposed a novel feature selection algorithm based on cuttlefish and is called, cuttlefish algorithm (CFA). CFA used to search prominent features among KDD99 features. This algorithm is based on reflection and visibility same as cuttlefish do when light strikes it. CFA algorithm was modified to select features. Six cases were introduced in the feature selection process using CFA. These cases were carried out in steps. It followed the initialization process in which population for some initial randomly generated solutions were generated. Each population was associated with two subsets named, selected features and unselected features. After which case 1 and case 2 were carried out in parallel to selected best feature subset from each population. In case 3 and case 4 single feature extraction was used to produce new feature subset from its above cases. In case 5 new subset was generated and case 6 remaining solutions for populations were carried out. Each subset was validated using decision tree in terms of the fitness function. The paper claimed not to consider false positive rate as KDD99 test dataset contains attacks, which are not included in the training set. But the aim of anomaly detection The task to detect the unseen attacks with less FPR. So we cannot ignore FPR in anomaly detection case.

2.4. Dataset And Evaluation Analysis

2.4.1. KDD CUP 99 Data Set

KDD99 dataset is the network intrusion dataset that is utmost extensively used for the evaluation of anomaly detection methods in network intrusions [42]. The dataset came from DAPRA'98 IDS evaluation program [43, 44]. Training data was collected from seven weeks of data in which few week's data are attack free while other weeks of data consist of attacks.

Testing data consists of two weeks of data which consists of normal and attacks data. Kdd99 has huge records and that's why its subset is widely used and is called kddcup.data_10_percent (kdd99_10%). There are 22 attacks are in the training set and 16 additional attacks are in the testing set. The training set contains 492020 instances while 311029 instances are for testing dataset [45].

KDD99 contains four attack classes, Denial of Service (DoS), Probe, Root-to-Local (R2L), User-to-Root (U2R), and one legitimate data class called, Normal [46]. Sub attacks of these attack classes are given in Table 1.

Table 1: Attack Classes and sub attacks

Attack Class	Attack name	Additional attack in test dataset
DoS	smurf, neptune, back, teardrop, pod, land.	Processtable, apache2, mailbomb.
PROBE	satan, ipsweep, portsweep, nmap.	saint, mscan.
U2R	buffer_overflow, rootkit, load_module, perl.	sqlattack, xterm, httptunnel, ps.
R2L	warezclient, guess_password, warezmaster, imap, ftp_write, multihop, phf, spy.	sendmail, xsnoop, named, snmpguess, xlock, snmpgetattack, worm.

Journal of Theoretical and Applied Information Technology
15th December 2016. Vol.94. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: KDD99 features with labels

S.No	Label	Feature	S.No	Label	Feature	S.No	Label	Feature
1	A	duration	15	O	su_attempted	29	AC	same_srv_rate
2	B	protocol_type	16	P	num_root	30	AD	diff_srv_rate
3	C	service	17	Q	num_file_creations	31	AE	srv_diff_host_rate
4	D	flag	18	R	num_shells	32	AF	dst_host_count
5	E	src_byte	19	S	num_access_files	33	AG	dst_host_srv_count
6	F	dst_byte	20	T	num_outbound_cm	34	AH	dst_host_same_srv_ rate
7	G	land	21	U	is_host_login	35	AI	dst_host_diff_srv_rate
8	H	wrong_	22	V	is_guest_login	36	AJ	dst_host_same_src_port_rate
9	I	urgent	23	W	count	37	AK	dst_host_srv_diff_host_rate
10	J	hot	24	X	srv_count	38	AL	dst_host_serror_rate
11	K	num_failed_	25	Y	serror_rate	39	AM	dst_host_srv_serror_rate
12	L	logged_in	26	Z	srv_serror_rate	40	AN	dst_host_rerror_rate
13	M	um_promised	27	AA	rerror_rate	41	AO	dst_host_srv_rerror_rate
14	N	root_shelL	28	AB	srv_rerror_rate			

Table 3: KDD99_10% data distribution

Category	Instances	Distinctive Instances
NORMAL	97277	87831
DoS	391458	54572
PROBE	4107	2130
R2L	1126	999
U2R	52	52
TOTAL	**494020**	**145584**

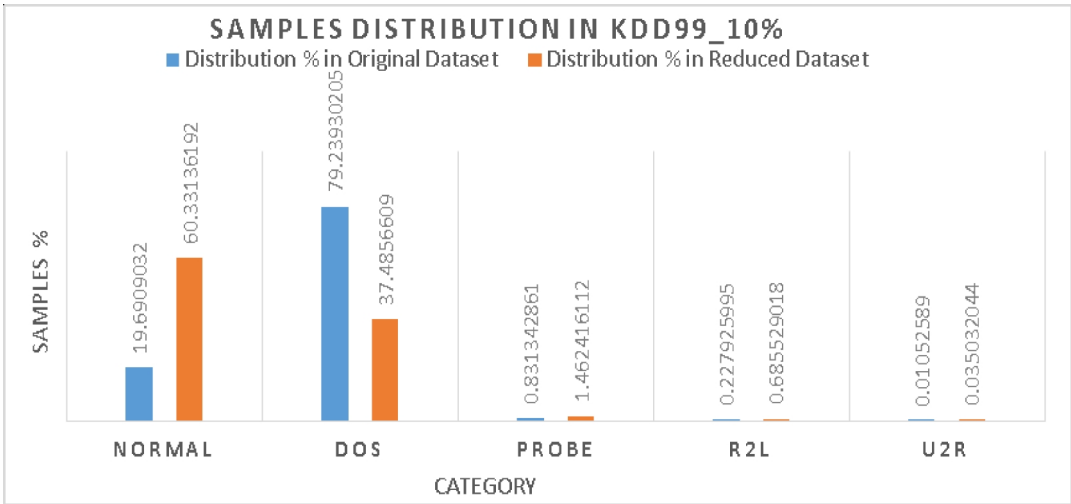

Figure 1: % Distribution Among Original KDD99_10% And Reduced Dataset

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The instances of each class are described by 41 features as shown in Table 2.
KDD99 features can be divided into three main categories [47] as follows,
1) Basic features: These features are extracted from single TCP/IP connections.
2) Traffic features: Also called time-based features and is extracted from the connections window interval.
3) Content feature: These features deal with the data portion of the TCP/IP packet.

KDD99 has many redundant data both in training and testing dataset. The distribution of different classes in KDD99 has been shown in Table 3.
It can be seen that in original dataset DoS and PROBE attack classes have a huge number of instances compared to U2R and R2L, which have very few instances. One reason for that is DoS and PROBE attacks occur more frequently in a short period. Many researches, therefore, claim having high detection rate for DoS and PROBE classes and low detection rate for R2L and U2R classes. Furthermore, as this dataset is the subset of DARPA 1998 which has many flaws, detailed discussion can be found in [48].
The percentage distribution in KDD99 (10%) of different classes has been shown in Figure 1. It also shows the distribution in KDD99 (10%) of different classes when the redundant data from KDD99 are removed.

2.4.2. Evaluation

The performance of the classifier is analyzed by measuring the error rate of the detection system. To estimate the error rates and accuracy of the system, the confusion matrix is used, given in Figure 2.

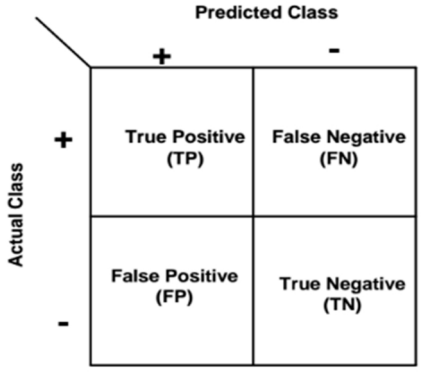

Figure 2: Confusion Matrix

NORMAL, DoS, and PROBE class but for U2R and R2L it has a very poor result.
While [32] has high TPR for the R2L class in which the proposed feature selection method

Selected eight features. In [39] detection method has high TPR, 99.4% among listed detection methods for probe class. This work used SVM classification for their detection model and SVM was trained with 32 features. It also combined theinstances of R2L and U2R classes and treated them as a single class. The TPR is high for that combined class, i.e. 98.7%. Table 4 also depicts the type of feature selection method that has been adopted in different methodology. For understanding features

-

of KDD99 has been labeled which is shown in Table 2. From the table, it can be seen that both*Figure 2: Confusion Matrix*

All other performance parameters are derived from TP, FN, FP, and TN.

- True Positive (TP): when the data of a positive class is correctly classified in its positive class.
- False Positive (FP): when the data of a negative class is incorrectly classified into positive class.
- False Negative (FN): when the data of a positive class is incorrectly classified in negative class.
- True Negative (TN): when the data of a negative class is correctly classified in its negative class.

The aim of the detection algorithm is to reduce the FPs and FNs, as these represents error and are misclassified data, meanwhile increase TPs and TNs. For some attacks, detection algorithm works perfectly but for others it is unable to detect. To detect that types of attack, the algorithm needs to be modified accordingly so for that particular type of attacks its performance gets improve.

3. DISCUSSION

In this section results of different listed methods in literature is discussed. Summary of different literature has been given in Table 4. It also includes a number of features selected, using feature selection method, and learning methods, used for the detection of an intrusion. Not a single detection method has a good result for all the classes of KDD99.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In [35] genetic and fuzzy method selected 25 features and detection method, fuzzy knowledge base rule, was used to develop a detection method using these 25 features. It has good TPR for types, wrapper based and filter based, of feature selection has been adopted.

Moreover, some author used nominal to binary conversion in data preprocessing process of KDD99. Some of them treated all different nominal values with different numeric values while some used same numeric value for different nominal values. The overall accuracy rate for detection methods used in the above literature is given in Figure 3. This includes the total number of correctly detected instances into its respective classes. It can be seen that the rough set classifier, which was trained by eight features has very low accuracy rate. Fuzzy Association Rule [35] which was trained by 25 features has better result compared to [32]. There are four detection methods that used SVM classifier. Proposed method in [39] is slightly better than [37] SVM detection model but [35] SVM used 25 feature while [37] SVM used just 6 features. Similarly, although SVM [36] has less accuracy rate compared to SVM [39] having 99.4% accuracy rate but [36] SVM has used ten features while SVM [36] collectively used 32 features. In [36] author found separate features for each class and also combined R2L and U2R classes into one class. Here it is worth noticing that KDD99 consists of a normal class and four attack classes, some methodologies just used two classes for their work, by combining all four attack classes into, one attack or abnormal class and the second one is a normal class. Whereas some of them treated all the five classes separately. If all the attack classes are combined it will give good TPR as DoS class has a huge number of examples in KDD99 dataset. Enormous examples of DoS dominant other attack classes. As most of other attack classes have less TPR due to an insufficient amount of examples. So by combining DoS class with other attack classes, commutatively high TPR ignoring the FPR for other attack classes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Results Of Different Detection Methods For Feature Selection Methods

Feature Selection Method	No. of Feature	Feature Selection Type	Selected Features	Detection Algorithm	Authors
SVM	13	Wrapper based	A, B, C, E, F, I, W, X, AC, AF, AG, AH, AJ	SVM	Mukkammal et al. [22]
SA and DT	23	Wrapper based	----------	SVM	Lin et al. [23]
SVM	28	Filter based	----------	SVM	George [24]
Information Gain, Gain	20	Filter based	----------	GMDH	Baig et al. [25]
CANN	19	Filter based	B, D, F, L, W, X, Y, Z, AC, AD, AE, AF, AG, AH, AI, AJ, AK,	K-NN	Lin et al. [26]
Genetic Fuzzy	25	Filter based	E, F, W, Y, AF, AG, AI	Fuzzy Knowledge Base Rules	Tsang et al. [35]
Rough Set	7	Filter based	----------	NN	Ghali [31]
Rough Set Theory	8	Filter based	----------	Rough Set Classifier	Rassam and Maroof [32]
Rough SET	16	Filter based	A, C, D, E, F, J, K, M, N, P, Q, R, S, V, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AI,	SVM	Yao et al. [33]
Rough set	6	Wrapper based	----------	SVM	Shrivastava and Jain [34]
MC and BA	10	Filter based	B, C, H, M, T, X, AF,	SVM	Rufai et al. [36]
Rough Set And PSO	6	Wrapper based	B, D, X, AA, AH, AI	SVM	Zainal et al. [37]
LGP and BA	6	Wrapper based	C, L, W, X, AA, AI	SVM	Hasani et al. [38]
ACO	32	Wrapper based	----------	SVM	Hua et al. [39]
Cuttlefish Algorithm	10	Wrapper based	----------	DT	Eesa et al. [41]

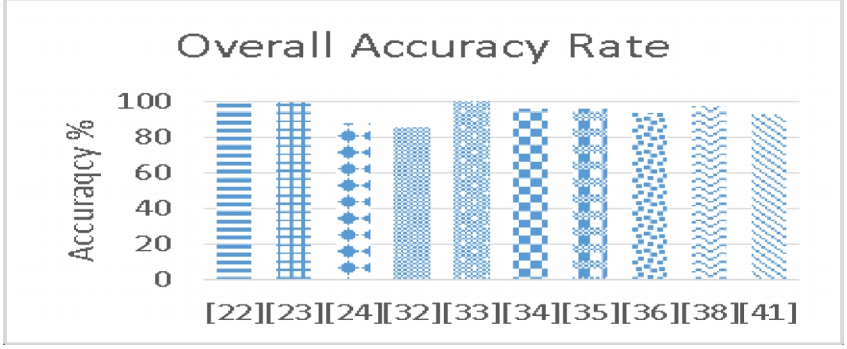

Figure 3: Accuracy Rate For Intrusion Detection

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. OPEN ISSUES AND CHALLENGES

Network speed is moving at high speed like Gbps for which real time anomaly detection methods are demanding. Therefore, only the relevant and correlated features can decrease time complexity for classification algorithm to develop efficient detection model. Mostly detection models are being evaluated using KDD99, which is a very old dataset. Although, many analyst believe that attacks in nowadays are the variants of the attacks from KDD99. If a detection model cannot achieve 100% TPR and 0% of FPR for all classes of KDD99, so how it will work for other latest attack classes. The data distribution in KDD99 is not uniform that has an impact on the detection model. The attacks in KDD99 are generated by the synthetic way, so the flaws in those systems are also included in KDD99. Many work has done by considering the data reduction to achieve less training and detection time for IDS while ignoring detection rate.

KDD99 test dataset contains some novel attacks, but many authors ignored to find out whether their model was able to detect those novel attacks or not. They only give the cumulative result for all attack classes. Many IDS used hybrid detection approach, by combining signature based and anomaly based. This can help to reduce FPR but it can increase time complexity for IDS operation.

5. CONCLUSION

Intrusion detection plays a vital role in the security of communication. IDS must be intelligent enough to have high detection rate with the less false positive rate. The high amount of data and irrelevant, redundant features make it difficult to build the prediction model for anomaly detection method. Furthermore, fast and efficient detection method can be achieved using robust features. This leads to real time detection of the intrusions in networks. This paper reviews feature selection techniques in the field of network anomaly detection. Feature selection techniques have been discussed in machine learning, rough set theory, and evolution computing. Comparative results for these techniques have been given and discussed critically. Open issues and challenges are being discussed. It can be concluded that feature selection for intrusion detection method is very critical because the detection method highly depends on the features of the input data.

REFERENCES

[1] Shah, B. And B.H. Trivedi. "Reducing Features Of KDD CUP 1999 Dataset For Anomaly Detection Using Back Propagation Neural Network". In Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference On. IEEE, 2015.

[2] Zheng, D. And C. Zhang, "Selecting Feature Subset For Large-Scale Data Via Fuzzy Rough Approach". Journal Of Convergence Information Technology, 8(9): P. 109, 2013.

[3] Kenkre, P.S., A. Pai, And L. Colaco. "Real Time Intrusion Detection And Prevention System". In Proceedings Of The 3rd International Conference On Frontiers Of Intelligent Computing: Theory And Applications (FICTA) 2014, Springer International Publishing, 2015.

[4] Chandola, V., A. Banerjee, And V. Kumar, "Anomaly Detection: A Survey". ACM Computing Surveys (CSUR), 41(3): P. 15, 2009.

[5] Xiaohui, Z. "Research And Design Of Intrusion Detection System In Computer Network". In 2015 International Conference On Social Science And Technology Education, Atlantis Press, 2015.

[6] Caselli, M., E. Zambon, And F. Kargl. "Sequence- Aware Intrusion Detection In Industrial Control Systems". In Proceedings Of The 1st ACM Workshop On Cyber-Physical System Security, ACM, 2015.

[7] Bhuyan, M.H., D. Bhattacharyya, And J.K. Kalita, "Network Anomaly Detection: Methods, Systems And Tools". Communications Surveys & Tutorials, IEEE, 16(1): P. 303-336, 2014.

[8] Nguyen, H.A. And D. Choi, "Application Of Data Mining To Network Intrusion Detection: Classifier Selection Model", In Challenges For Next Generation Network Operations And Service Management, Springer. P. 399-408, 2008.

[9] Othman, Z.A., Et Al., "Record To Record Feature Selection Algorithm For Network Intrusion Detection". International Journal Of Advancements In Computing Technology, 6(2): P. 163, 2014.

[10] Garcia-Teodoro, P., Et Al., "Anomaly-Based Network Intrusion Detection: Techniques, Systems And Challenges". Computers & Security, 28(1): P. 18-28, 2009.

[11] Farhoud Hosseinpour, P.V.A., Fahimeh Farahnakian, Juha Plosila, Timo Hämäläinen, "Artificial Immune System Based Intrusion

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Detection: Innate Immunity Using An Unsupervised Learning Approach". JDCTA (International Journal Of Digital Content Technology And Its Applications), Volume 8(5): P. 01-12, 2014.

[12] Friedberg, I., F. Skopik, And R. Fiedler, "Cyber Situational Awareness Through Network Anomaly Detection: State Of The Art And New Approaches". E & I Elektrotechnik Und Informationstechnik 132(2): P. 101-105, 2015.

[13] Chen, L.-S. And J.-S. Syu. "Feature Extraction Based Approaches For Improving The Performance Of Intrusion Detection Systems". In Proceedings Of The International Multiconference Of Engineers And Computer Scientists. 2015.

[14] Stańczyk, U. And L.C. Jain, "Feature Selection For Data And Pattern Recognition: An Introduction, In Feature Selection For Data And Pattern Recognition", Springer. P. 1-7, 2015.

[15] Pramokchon, P. And P. Piamsa-Nga, "Content-Adaptive Feature Selection For Classifying Class- Imbalanced Data". International Journal Of Advancements In Computing Technology, 6(5): P. 66, 2014.

[16] García, S., J. Luengo, And F. Herrera, "Feature Selection, In Data Preprocessing" In Data Mining, Springer. P. 163-193, 2015.

[17] Gediga, G. And I. Düntsch, "Rough Set Data Analysis—A Road To Non-Invasive Knowledge Discovery", Methodos Publishers, UK, 2000.

[18] Wang, S., Et Al., "Subspace Learning For Unsupervised Feature Selection Via Matrix Factorization". Pattern Recognition, 48(1): P. 10-19, 2015

[19] Zhang, F., Et Al., "Adversarial Feature Selection Against Evasion Attacks", 2015.

[20] Pitt, E. And R. Nayak. "The Use Of Various Data Mining And Feature Selection Methods In The Analysis Of A Population Survey Dataset". In Proceedings Of The 2nd International Workshop On Integrating Artificial Intelligence And Data Mining-Volume 84, Australian Computer Society, Inc, 2007.

[21] Wang, A., Et Al., "Accelerating Wrapper-Based Feature Selection With K-Nearest-Neighbor". Knowledge-Based Systems, 83: P. 81-91, 2015.

[22] Mukkamala, S., G. Janoski, And A. Sung. "Intrusion Detection: Support Vector Machines And Neural Networks". In Proceedings Of The IEEE International Joint Conference On Neural Networks (ANNIE). 2002.

[23] Lin, S.-W., Et Al., "An Intelligent Algorithm With Feature Selection And Decision Rules Applied To Anomaly Intrusion Detection". Applied Soft Computing, 12(10): P. 3285-3290, 2012.

[24] George, A., "Anomaly Detection Based On Machine Learning: Dimensionality Reduction Using PCA And Classification Using SVM" '. International Journal Of Computer Applications (0975–8887) Volume, 2012.

[25] Baig, Z.A., S.M. Sait, And A. Shaheen, "GMDH- Based Networks For Intelligent Intrusion Detection. Engineering Applications Of Artificial Intelligence", 26(7): P. 1731-1740, 2013.

[26] Lin, W.-C., S.-W. Ke, And C.-F. Tsai, CANN: "An Intrusion Detection System Based On Combining Cluster Centers And Nearest Neighbors". Knowledge-Based Systems, 2015.

[27] Tsai, C.-F. And C.-Y. Lin, "A Triangle Area Based Nearest Neighbors Approach To Intrusion Detection". Pattern Recognition, 43(1): P. 222-229, 2010.

[28] Xue-Qin, Z., G. Chun-Hua, And L. Jia-Jin. "Intrusion Detection System Based On Feature Selection And Support Vector Machine". In Communications And Networking In China, 2006. Chinacom'06. First International Conference On, IEEE, 2006.

[29] Rissino, S. And G. Lambert-Torres, "Rough Set Theory–Fundamental Concepts, Principals, Data Extraction, And Applications". Data Mining And Knowledge Discovery In Real Life Applications, P. 438, 2009.

[30] Pawlak, Z., "Rough Sets And Intelligent Data Analysis". Information Sciences, 147(1–4): P. 1-12, 2002.

[31] Ghali, N.I., "Feature Selection For Effective Anomaly- Based Intrusion Detection". International Journal Of Computer Science And Network Security, 9(3): P. 285-289, 2009.

[32] Rassam, M.A. And M.A. Maarof, "Artificial Immune Network Clustering Approach For Anomaly Intrusion Detection". Journal Of Advances In Information Technology, 3(3): P. 147-154, 2012.

[33] Yao, J., S. Zhao, And L. Fan, "An Enhanced Support Vector Machine Model For Intrusion Detection, In Rough Sets And Knowledge Technology", Springer. P. 538-543, 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[34] Shrivastava, S.K. And P. Jain, "Effective Anomaly Based Intrusion Detection Using Rough Set Theory And Support Vector Machine". International Journal Of Computer Applications, 18(3): P. 35-41, 2011.

[35] Tsang, C.-H., S. Kwong, And H. Wang, "Genetic- Fuzzy Rule Mining Approach And Evaluation Of Feature Selection Techniques For Anomaly Intrusion Detection". Pattern Recognition, 40(9): P. 2373- 2391, 2007.

[36] Rufai, K.I., R.C. Muniyandi, And Z.A. Othman, "Improving Bee Algorithm Based Feature Selection In Intrusion Detection System Using Membrane Computing". Journal Of Networks, 9(3): P. 523-529, 2014.

[37] Zainal, A., M.A. Maarof, And S.M. Shamsuddin, "Feature Selection Using Rough-DPSO In Anomaly Intrusion Detection", In Computational Science And Its Applications–ICCSA 2007, Springer. P. 512-524, 2007.

[38] Hasani, S.R., Z.A. Othman, And S.M.M. Kahaki, "Hybrid Feature Selection Algorithm For Intrusion Detection System". Journal Of Computer Science, 10(6): P. 1015-1025, 2014.

[39] Gao, H.-H., H.-H. Yang, And X.-Y. Wang. "Ant Colony Optimization Based Network Intrusion Feature Selection And Detection". In Machine Learning And Cybernetics, 2005. Proceedings Of 2005 International Conference On, IEEE, 2005.

[40] Rais, H.M., Z.A. Othman, And A.R. Hamdan. "Improved Dynamic Ant Colony System (DACS) On Symmetric Traveling Salesman Problem (TSP)". In Intelligent And Advanced Systems, 2007. ICIAS 2007. International Conference On, IEEE, 2007.

[41] Eesa, A.S., Z. Orman, And A.M.A. Brifcani, "A Novel Feature-Selection Approach Based On The Cuttlefish Optimization Algorithm For Intrusion Detection Systems". Expert Systems With Applications, 42(5): P. 2670-2679, 2015.

[42] Tavallaee, Mahbod, Et Al. "A Detailed Analysis Of The KDD CUP 99 Data Set." Proceedings Of The Second IEEE Symposium On Computational Intelligence For Security And Defence Applications 2009, 2009.

[43] Kayacik, H. Günes, A. Nur Zincir-Heywood, And Malcolm I. Heywood. "Selecting Features For Intrusion Detection: A Feature Relevance Analysis On KDD 99 Intrusion Detection Datasets." Proceedings Of The Third Annual Conference On Privacy, Security And Trust, 2005.

[44] Lincoln Laboratory, Massachusetts Institute Of Technology. [Cited 2015 January 15, 2015]; Available From: Http://Www.Ll.Mit.Edu/Mission/Communications/Cybe R/Cstcorpora/Ideval/Data/.

[45] Faisal, M.A., Et Al., "Data-Stream-Based Intrusion Detection System For Advanced Metering Infrastructure In Smart Grid: A Feasibility Study". Systems Journal, IEEE 9(1): P. 31-44, 2015.

[46] KDD Cup 1999 Data. Available From: Http://Kdd.Ics.Uci.Edu/Databases/Kddcup99/Kddcup99.Html.

[47] Lei, L., T"He Network Abnormal Intrusion Detection Based On The Improved Artificial Fish Swarm Algorithm Features Selection". Journal Of Convergence Information Technology, 8(6), 2013.

[48] Mchugh, J., "Testing Intrusion Detection Systems: A Critique Of The 1998 And 1999 DARPA Intrusion Detection System Evaluations As Performed By Lincoln Laboratory". ACM Transactions On Information And System Security, 3(4): P. 262-294, 2000.