Jurnal Teknologi

Full paper

Identifying Generic Features of KDD Cup 1999 for Intrusion Detection

Hamid H. Jebur[*], Mohd Aizaini Maarof, Anazida Zainal

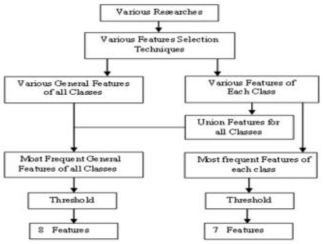
Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding authors: hamedhh59@yahoo.com

Article history

Received :4 February 2013
Received in revised form :
13 December 2014
Accepted :15 March 2015

Graphical abstract


Various Researches
Various Features Selection Techniques
Various General Features of all Classes
Various Features of Each Class
Union Features for all Classes
Most Frequent General Features of all Classes
Most frequent Features of each class
Threshold
Threshold
8 Features
7 Features

Abstract

Detection accuracy of Intrusion Detection System (IDS) depends on classifying network traffic based on data features. Using all features for classification consumes more computation time and computer resources. Some of these features may be redundant and irrelevant therefore, they affect the detection of traffic anomalies and the overall performance of the IDS. The literature proposed different algorithms and techniques to define the most relevant sets of features of KDD cup 1999 that can achieve high detection accuracy and maintain the same performance as the total data features. However, all these algorithms and techniques did not produce optimal solutions even when they utilized same datasets. In this paper, a new approach is proposed to analyze the researches that have been conducted on KDD cup 1999 for features selection to define the possibility of determining effective generic features of the common dataset KDD cup 1999 for constructing an efficient classification model. The approach does not rely on algorithms, which shortens the computational cost and reduces the computer resources. The essence of the approach is based on selecting the most frequent features of each class and all classes in all researches, then a threshold is used to define the most significant generic features. The results revealed two sets of features containing 7 and 8 features. The classification accuracy by using eight features is almost the same as using all dataset features.

Keywords: Intrusion detection; accuracy; feature selection; classification

Abstrak

Ketepatan pengesanan Sistem Pengesanan Pencerobohan (IDS) bergantung kepada engklasifikasikan trafik rangkaian berasaskan kepada ciri-ciri data. Menggunakan semua ciri-ciri untuk pengelasan mengambil lebih masa pengiraan dan sumber komputer. Sebahagian daripada ciri-ciri ini mungkin tidak relevan dan tidak diperlukan. Oleh itu, ia mempengaruhi pengesanan anomali trafik dan prestasi keseluruhan IDS. Kajian literatur mencadangkan algoritma dan teknik yang berbeza untuk menentukan set ciri-ciri KDD Cup 1999 yang paling relevan yang boleh mencapai ketepatan pengesanan yang tinggi dan mengekalkan prestasi yang sama dengan jumlah ciri-ciri data yang sama. Walau bagaimanapun, semua algoritma dan teknik ini tidak menghasilkan penyelesaian yang optimum walaupun mereka menggunakan set data yang sama. Dalam kertas kerja ini, satu pendekatan baru dicadangkan untuk menganalisis kajian yang telah dijalankan ke atas KDD Cup 1999 bagi ciri pemilihan untuk menentukan ciri-ciri generik berkesan dataset biasa KDD Cup 1999 untuk membina model klasifikasi yang cekap. Pendekatan ini tidak bergantung kepada algoritma, yang mana akan memendekkan kos komputasi dan mengurangkan sumber komputer. Intipati pendekatan berasaskan memilih ciri-ciri yang paling kerap bagi setiap kelas dan semua kelas dalam set data bagi dari semua kajian, kemudian batas digunakan untuk menentukan ciri-ciri generik yang paling penting. Keputusan menunjukkan dua set ciri mengandungi 7 dan 8 ciri. Ketepatan pengelasan dengan menggunakan lapan ciri adalah hampir sama dengan menggunakan kesemua ciri-ciri dataset.

Kata kunci: Pencerobohan pengesanan; ketepatan; pemilihan ciri; klasifikasi

■1.0 INTRODUCTION

Intrusion detection is the process of detecting any suspicious activities that compromise the security [1]. It becomes a key technology in achieving computer and network security and one of the effective means of early warning and strong defense in achieving information security, and avoiding the losses caused by various attacks [2] . The key function of intrusion detection systems is to monitor and analyze network traffic to find any malicious or abnormal activities using different detection techniques. Network traffic is very large, and it contains many features. Analyzing these data with all features is difficult, consuming more computation time, and can affect IDS accuracy because exotic features can increase the difficulty of detecting suspicious behavior [3], and increase system overload [4]. Some of these features may be redundant, irrelevant or noisy, which lead to reduce the detection accuracy [4,5]. Therefore, these features must be removed and a sub of relevant features should be extracted [6]. Feature selection is a promise approach for minimizing the extracted features to a small significant subset efficient to build optimal classification model and improve detection accuracy rate [7] with taking into account that using a few features may lead to lose some significant information.

The algorithms of feature selection are classified into the filter and wrapper models. The filter model selects some features depending on training data characteristics without using a learning algorithm, while the wrapper model needs a learning algorithm in feature selection [4, 5]. Most of the researches [8, 9, 10 and 11] have been done on total features of DARPA as a dataset for detecting intrusions [4], while less research have been done on feature selection. On the other hand, few researches have been done on real data [12, 13]. However, all algorithms and techniques for feature selection did not produce optimal solutions even when they utilize same datasets, and they did not agree on same features even for same attack class.

In this paper, several researches that have been done on DARPA dataset (KDD cup 1999) for features selection are investigated to define a subset of generic features relevant for classification, on the condition, that this process does not affect the detection accuracy and IDS performance. The features are selected based on a new approach that selects the most frequent features of each class and of all classes, and then assign a threshold to choose the significant features. This paper is organized as follows. Section 2 highlights DARPA dataset features. Section 3 reviews researches that have been done on feature selection demonstrating the techniques that have been used and the selected features. Section 4 shows the proposed approach. The results and discussion are shown in section 5, while the conclusion and future work are in sections 6.

2.0 DARPA DATASET

DARPA dataset is a sample of network traffic and audit logs of a simulated military network used for evaluating intrusion detection systems. These data were collected in 1998 by the Information Systems Technology Group of MIT Lincoln Laboratory, under Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL/SNHS) sponsorship. Two datasets were used for IDS evaluation; 1998 and 1999 DARPA which contains more attacks including Windows NT audit logs [14]. In 1999, the Third International Knowledge Discovery and Data Mining Tools Competition (KDD cup 1999) selected the DARPA 1998 datasets as their target dataset [15]. DARPA dataset totaled 4 gigabytes, and was processed into five million connection records [16] as KDD cup 1999 dataset. It contains 22 different attacks types classified into four groups: denial of services (DOS), probing, user to root attacks (U2R) and remote to user attacks (R2L) [17]. Each network connection record in this dataset has 41 (features) features listed in Table 1 and can be classified as follows [18]:

13 Content features, such as flag, number of failed logins, hot indicators, etc.

10 Host-based features depend on the past 100 connections similar to the one under consideration.

9 Time-based features, such as SYN error rates, Rejection rates, etc.

9 Basic features, such as connection duration, service requested, bytes transferred between source and destination machine, etc.

KDD cup 1999 dataset is still used for testing most of the IDSs [19, 20 and 21], although the traffic characteristics have been changed slightly such as the rising use of encrypted protocols, which cannot be detected by the IDSs [22].

Table 1 Network connection features

No	Network Features	No	Network Features	No	Network Features	No	Network Features
1	Duration	12	logged_in	23	Count	34	dst_host_same_srv_rate
2	Protocol_type	13	num_compromised	24	srv_count	35	dst_host_diff_srv_rate
3	Service	14	Root shell	25	serror_rate	36	dst_host_same_src_port_rate
4	Flag	15	su_ attempt	26	srv_serror_rate	37	dst_host_srv_diff_host_rate
5	Src_bytes	16	Num_root	27	rerror_rate	38	dst_host_serror_rate
6	Dst_bytes	17	Num_file_creations	28	srv_rerror_rate	39	dst_host_srv_serror_rate
7	Land	18	Num_shells	29	same_srv_rate	40	dst_host_rerror_rate
8	Wrong fragment	19	num_access_files	30	diff_srv_rate	41	dst_host_srv_rerror_rate
9	Urgent	20	num_outbound_cmds	31	srv_diff_host_rate		
10	Hot	21	is_host_login	32	dst_host_count		
11	Num_failed_logins	22	is_guest_login	33	dst_host_srv_count		

Source: [23]

3.0 RELATED WORK

Feature selection is the process of removing redundant and irrelevant features of network traffic data, and extracting the relevant ones that achieve high classification rate and enhance the overall performance of IDSs. Several researches have been held for features selection of KDD cup 1999 dataset to reduce computer resources and computation time when detecting traffic anomalies. Several techniques have been applied to extract the effective and relevant features to improve the detection accuracy or at least keep it unaffected.

All previous studies conducted on selected features from KDD cup 1999 dataset showed that the extraction of relevant features leads to an improvement in IDS classification and has no negative impact on the detection accuracy. It leads often to improve the accuracy, and overall IDS performance, in addition to its role in reducing the computer resources and computation time.

Mukkamala *et al*. [23] used support vector machines to extract 13 features as the most significant ones as shown in Table 2.

Mukkamala and Sung [24] addressed the importance of feature selection and ranking in intrusion detection since the reduction of irrelevant features leads to enhance the detection accuracy and IDS performance. They used two methods to rank the features of KDD cup 1999, the first one is performance-based ranking method (PBRM) and the second is SVM-specific feature ranking method (SVDFRM). The important features for each attack class are shown in Table 3. The union features are derived from the features of all classes and added to the table to be used in the analysis too.

Table 2 Extracted features

Author	**Approach**	**Attack Class**	**Features**
Mukkamala *et al*.[23]	SVMs	For all class	1,2,3,5,6,9,23,24,29,32,33,34,36

Table 3 Important features

Author	**Approach**	**Attack Class**	**Features**	**Union Features** Added by the researcher
Mukkamala and Sung [24]	PBRM	Normal	1,3,5,6,8-10,14,15,17,20-23, 25-29,33,35,36,38,39,41	1,3,5,6,8-10,14-29, 32,33,35,36,38,39,41
		Prope	3,5,6,23,24,32,33	
		Denial of Service	1,3,5,6,8,19,23-28, 32,33,35,36,38-41	
		U2R	5,6,15,16,18,32,33	
		R2L	3,5,6,24,32,33	
	SVDFRM	Normal	1-6,10,12,17,23,24,27,28,29, 31-34,36,39	1,2,3,4,5,6,10,12,17,23-29, 31,32,33,34,36,38,39
		Prope	1-6,23,24,29,32,33	
		Denial of Service	1,5,6,23-26,32,36,38,39	
		U2R	1-6,12,23,24,32,33	
		R2L	1,3,5,6,32,33	

Chebrolu *et al*. [25] indicated that selecting effective features is important to build IDSs. They used two feature selection techniques to select the important features of KDD cup 1999 dataset such as Bayesian networks (BN) and Classification and Regression Trees (CART). The features selected by their techniques are shown in Table 4.

Table 4 Selected features

Author	**Approach**	**Features**
Chebrolu *et al*.[25]	Bayesian Network (BN)	1,2,3,5,7,8,11,12,14,17,22-26,30,32
	Classification and Regression Trees (CART)	3,5,6,12,23,24,25,28,31,32,33,35

Zainal *et al*. [26] showed that rough set achieved a good performance in selecting effective features of KDD cup 99 data similar to other techniques such as linear genetic programming (LGP), multivariate adaptive regression splines (MARS), and support vector decision function (SVDF). The selected features are shown in Table 5.

Wang *et al*. [4] used different feature selection techniques such as wrapper with Bayesian networks (BN), wrapper with Decision trees (C4.5) and information gain (IG) to select effective features of KDD cup 1999 dataset. They considered the most 10 features, commonly selected by all techniques at the same time, as the most significant ones for each attack class. The features are shown in Table 6. The union features also are derived from the features of all classes to be used in the analysis.

Table 5 Most significant features

Author	**Approach**	**Attack Class**	**Features**
Zainal *et al*.[26]	Rough set	For all class	3,4,5,24,32,41

Table 6 Most significant features with union set

Author	Attack class	Approach	Features	Selected Features	Union Features Added by the researcher
Wang *et al.*[4]	DoS	IG	5, 23, 3, 24, 6, 2, 36	3,4,5,6,8,10,13,23, 24,37	1-6,8,10,12,13,14,22-24,29,30-33, 35-37,39,40
		Wrapper (BN)	4, 5, 8, 10, 13, 23, 37		
		Wrapper (C4.5)	3, 5, 6, 13, 23		
	Probe	IG	5, 3, 4, 6, 23,27, 33-35	3,4,5,6,29,30,32,35, 39,40	
		Wrapper (BN)	3, 4, 5, 29, 32, 35		
		Wrapper (C4.5)	5, 29, 30, 35, 39, 40		
	R2L	IG	5, 3, 6, 33, 36, 10,37, 24,1	1,3,5,6,12, 22, 23,31, 32,33	
		Wrapper (BN)	1, 5, 6, 22, 23, 32		
		Wrapper (C4.5)	1, 3, 5, 6, 12, 31		
	U2R	IG	3,33,13,14,1,10,5,17,32,36	1,2,3,5,10,13,14,32, 33,36	
		Wrapper (BN)	1, 2, 5, 14, 36		
		Wrapper (C4.5)	1, 13, 14, 32		

Zainal *et al.* [26] used a 2-tier feature selection process based on rough set and discrete particle swarm (DPSO) to reduce the features of KDD cup 1999 dataset of each attack class. The selected features are shown in Table 7 beside the union features, which added to be used in the analysis

Farid *et al.* [17] selected 19 features from KDD cup 1999 dataset as the important features by proposing new learning approach using naïve Bayesian classifier and ID3 algorithm. Table 8 shows the selected features.

Table 7 Selected features

Author	Approach	Attack Class	Features	Union Features Added by the researcher
Zainal *et al.* [26]	Rough Set and Discrete Particle Swarm (DPSO)	Normal	12,31-33,35,37,41	2,3,4,5,6,10,12,14,17, 22,23,24,29,31-38, 40,41
		Probe	2,3,23,34,36,40	
		DoS	5,10,24,29,33,34,38,40	
		U2R	3,4,6,14,17,22	
		R2L	3,4,10,23,33,36	

Table 8 Important features

Author	Approach	Attack Class	Features
Farid *et al.* [17]	New algorithm	For all class	1,3-6,8-11,13,15-19,23,24,32,33

Wa'el *et al.* [28] used rough set classification based parallel genetic algorithm (RSC-PGA) to reduce the selected features of KDD cup1999 to 22 features. Then, they selected the top rank five features as the important features. Unfortunately the 22 features are not mentioned in the research, and only the important ones are addressed. The selected features are shown in Table 9. Ghali [3] selected seven features as the most significant ones by using a new hybrid algorithm based on rough set and neural network (RSNNA). Table 10 shows the significant features.

Table 9 Extracted features

Author	Approach	Attack Class	Features
Wa'el *et al*.[28]	RSC-PGA	For all class	3,4,5,24,41

Table 10 Significant features

Author	Approach	Attack Class	Features
Ghali [3]	RSNNA	For all class	5,6,23,24,32,33,36

Othman *et al.* [29] used genetic algorithm GA to select features from customized features that resulted from preprocessing KDD cup 99 data. Three different sizes data sets are chosen for features selection, which yield three subsets of features 7,11and 15 as shown in Table 11.

Olusola *et al*. [30] identified the most relevant features of each attack class of KDD cup 99 dataset by using rough sets based on determining the dependency degree and the dependency ratio of each attack class. Table 12 shows these features. The union features also are derived from the features of all classes to be used in the analysis.

Table 11 Selected features

Author	Approach	Attack Class	Dataset	Features
Othman et al.[29]	GA	For all class	1	5,6,23,24,31,36,37
			2	5,6,13,23-26,33,36,37,38
			3	12,24-30,32,34,35,38-41

Table 12 Relevant features

Author	Approach	Attack Class	Features	Union Features Added by the researcher
Olusola *et al.* [30]	Rough Set	Normal	29	3,5,6,7,8,11,14,23,24,28,30,36,39
		Probe	5,28,30,36	
		DoS	5,7,8	
		U2R	3,14,24,36	
		R2L	3,6,11,23,39	

Revathi and Ramesh [31] minimized the 41 features of KDD cup 1999 dataset into two subsets of 7 and 14 features by using best first search method. Table 13 shows the two sets of features.

Table 13 Two sets of selected features

Author	Approach	Attack Class	Features	
			First set	Second set
Revathi and Ramesh [31]	Best First Search	For all class	1,3,4,5,6,23,24,25,27,33,34,35,36,40	2,3,5,6,23,30,33

Srinivasulu *et al*. [32] used genetic algorithm to select 10 relevant features out of the 41 features of KDD cup 1999 dataset. Table 14 shows the selected features.

Chen *et al*. [33] used rough set theory to select 29 features as the important features of KDD cup 1999 dataset. Table 15 shows the selected features.

Table 14 Selected features

Author	Approach	Attack Class	Features
Srinivasulu *et al*. [32]	Genetic Algorithm	For all class	2,3,5,6,23,36-40

Table 15 Selected features

Author	Approach	Attack Class	Features
Chen *et al*.[33]	Rough Set	For all classes	1,2,5,6,8,11-14,16-19,23,25,27,29-41

Hlaing [34] utilized feature selection based on the continuous features of the KDD cup 1999 dataset, which equal to 34 features. He extracted 10 features as the optimal features using mutual correlation. Table 16 shows the selected features.

Table 16 Selected features

Author	Approach	Attack Class	Features
Hlaing [34]	Mutual Correlation	For all classes	1,5,6,8,9,10,16,18,28,31

Alomari and Othman [35] used Bees Algorithm for feature selection and selected 6 features as the significant features of KDD cup 1999 dataset. Table 17 shows the selected features.

Chung and Wahid [36] utilized intelligent dynamic swarm based rough set (IDS-RS) for feature selection and extracted six features as the most significant features. The selected features are shown in Table 18.

Table 17 Selected features

Author	Approach	Attack Class	Features
Alomari & Othman [35]	Bee algorithm	For all classes	3,12,24,25,32,37

Table 18 Selected features

Author	Approach	Attack Class	Features
Chung and Wahid [36]	IDS-RS	For all classes	3,5,6,27,33,35

Pundir and Amrita [37] selected 15 features as the optimal features subset by using random forest. The selected features are shown in Table 19.

Devaraju and Ramakrishnan [38] used various feature selection techniques such as independent component analysis, linear discriminant analysis and principal component analysis to reduce the KDD cup 1999 dataset dimensionality and select the relevant features of the four attacks classes of the dataset. Table 20 shows these features and the union features.

Madbouly *et al*. [39] developed a relevant feature selection model to select the important features set of KDD cup 1999. They defined 11 features as the best features of the dataset. The features are shown in Table 21.

Table 19 Selected features

Author	Approach	Attack Class	Features
Pundir and Amrita [37]	Random forest	For all classes	2,3, 5,8,9, 12,14-17,27,28,32,33,38

Table 20 Selected features

Author	Approach	Attack Class	Features	Union Features Added by the researcher
Devaraju and Ramakrishnan [38]	Different Techniques	Probe	1,2,3,4,5	1-5,10-14,16-19,21-23,34,38-40
		DoS	1,2,4,5,23,34,38,39,40	
		U2R	10,13,14,16,17,18,19,21	
		R2L	1,2-5,10-13,17-19,21,22	

Table 21 Best features

Author	Approach	Attack Class	Features
Madbouly *et al*. [39]	New model	For all classes	1, 3, 4, 5, 6, 10, 14, 23, 25, 30, 35

4.0 THE PROPOSED APPROACH

The literature includes a wide range of researches conducted on KDD cup 1999 dataset for the purpose of extracting the significant features that can lead to build a high performance classifier for detecting anomalies in the dataset. Different techniques were used in these researches and they achieved satisfactory successes, but they did not achieve the desired ambition. Most of these researches used the same 10% of KDD cup 1999 but they did not give the same results, and gave different sets of features, and this is a big problem because the real network traffic data is not similar and the behavior of attacks is constantly evolving. The idea of this paper is simple and can be summarized as follows. These researches selected different features, but they must agree on key features of KDD cup 1999 that can build a high performance classifier. Therefore, the most frequent features in all researches are considered as the key features. This approach is used in many aspects such as ensemble technique and some rough set theory algorithms [40, 41]. This approach represents the essence of all researches that have been conducted previously regardless the type of the feature selection methods. It somehow converges, balances and ensembles the behaviors and functions of all the feature selection methods mentioned in the related work. The merit of this approach can be manifested in the non-use of any algorithm, which reduces the computational time and computer resources. The approach is shown in Figure 1, and the pseudo code is described below.

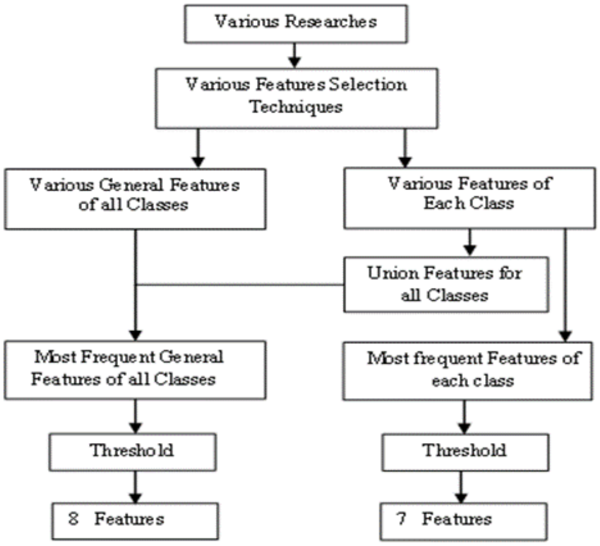

Various Researches
Various Features Selection Techniques
Various General Features of all Classes
Various Features of Each Class
Union Features for all Classes
Most Frequent General Features of all Classes
Most frequent Features of each class
Threshold
Threshold
8 Features
7 Features

Figure 1 The proposed approach

The Pseudo Code

```
-   Various researches on feature selection
-   Define the selected features of each research
-   For each research, define features of each class and
    general features of all classes if found
-   For each research, find the union features of all
    classes
-   For all researches, find the most frequent features of
    each class
-   Define a threshold
-   Select the final set of features
-   For all researches, find the most frequent features of
    union and general features
-   Define a threshold
-   Select the final set of features
```

■5.0 RESULTS AND DISCUSSION

KDD cup 1999 dataset has not changed and is still used to evaluate the performance of intrusion detection systems. Although many techniques have been used for feature selection, but they did not agree on specific common features for all attack classes of KDD cup 1999 dataset, or for each class. However, even they provided good results, but they did not attain optimal solutions. Actually, this somehow may be related to many reasons such as: 1) attack patterns and behaviors, 2) algorithms behaviors and structure, 3) inefficient algorithms and techniques, 4) improper preprocessing of KDD cup 1999 dataset, and 5) using different sizes of dataset.

The researches using NSL-KDD dataset are excluded because this dataset differs from KDD cup 1999 dataset, where it contains less redundant data. Using this data in this research can cause disparity in the performance of the algorithms and techniques of feature selection. On the other hand, some researches work were sketchily described. Khalaf *et al.* [42] used self-organizing map (SOM) and Principle component analysis (PCA) for feature selection, but they did not mention the selected feature. Raut and Singh [43] extracted general feature subsets (5, 21, and 33) from three rough set based feature selection techniques: Entropy-based, Open loop and Closed loop. However, there is a remarkable difference between the selected subsets in terms of feature numbers. In addition, they did not demonstrate which subset achieved better results. Revathi and Malathi [44] used hybrid simplified swarm optimization technique for both feature selection and classification, but the selected features were not mentioned. They only compared their classification results with other algorithms that utilized different feature selection techniques.

For this research, Microsoft excel is used to determine the most frequent features of each class of all researches. It also determines the most frequent features of the union features and the general features (same features for all class) for all researches. Then, the frequency and percentage of the extracted features are determined. A threshold 45% is defined to determine the degree of importance of each attribute, where each one exceeds this threshold is considered as one of the most important features. This consideration depends on the fact that these features are the most important features that have been extracted by all previous researches that used different techniques. The most frequent one in all researches, in spite of different selection techniques used, actually gains more importance. The final results of Excel, after taking the threshold into consideration, are shown in Tables (22, 23, 24, 25, 26, and 27).

Table 22 Significant normal class features

No. of Research	Attribute	Count	Percentage
23	3	15	65.21%
	5	19	82.60%
	6	15	65.21%
	23	14	60.86%
	24	13	56.52%
	32	12	52.17%
	33	13	56.52%

Table 23 Significant probe class features

No. of Research	Attribute	Count	Percentage
25	3	18	72.00%
	5	22	88.00%
	6	16	64.00%
	23	15	60.00%
	24	14	56.00%
	32	13	52.00%
	33	12	48.00%

Table 24 Significant DoS class features

No. of Research	Attribute	Count	Percentage
25	3	15	60.00%
	5	23	92.00%
	6	16	64.00%
	23	16	64.00%
	24	16	64.00%
	32	12	48.00%
	33	12	48.00%

Table 25 Significant U2R class features

No. of Research	Attribute	Count	Percentage
25	3	17	68.00%
	5	20	80.00%
	6	16	64.00%
	23	13	52.00%
	24	14	56.00%
	32	13	52.00%
	33	13	52.00%

Table 26 Significant R2L class features

No. of Research	Attribute	Count	Percentage
25	3	19	76.00%
	5	21	84.00%
	6	17	68.00%
	23	15	60.00%
	24	13	52.00%
	32	13	52.00%
	33	14	56.00%

Table 27 Significant union features

No. of Research	Attribute	Count	Percentage
25	3	19	76.00%
	5	23	92.00%
	6	18	72.00%
	23	18	72.00%
	24	17	68.00%
	32	14	56.00%
	33	14	56.00%
	36	13	52.00%

The proposed approach led to extracted two sets of features; the first one contains the significant 7 features extracted from the different features of each class. These features are 3, 5, 6, 23, 24, 32, and 33. The second set contains the significant 8 features extracted from the general features of all classes. These features are 3, 5, 6, 23, 24, 32, 33, and 36. The features are generic for all classes and are the same except the extra feature number 36 in the second feature set.

A subset of 10% of KDD cup 1999 dataset is used in this study, which includes five pairs of datasets processed by[25] and used by [25, 26]. Each pair represents training and testing data for one type of five classes of network attacks namely, Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probing, besides the normal class. The training data comprises of 5,092 records, and the testing data comprises of 6,890 records, where these data samples preserve the real KDD cup1999 distribution. The final training and testing subsets are determined by considering the selected features.

Lib SVM 3.11 is used as a classification tool to determine the detection accuracy of the two features sets derived by this study, in addition to the original set of the traffic that contains 41 features. The classification results are shown in Table 28. It is shown from the table that the performance of the 8 features is generally better than of the 7 features. The results illustrate an accuracy reduction in DOS and R2L classes, while the accuracy is almost the same in other classes. Also, the results indicate that feature number 36 significantly enhances the accuracy of the probe class in the case of seven features, which means it is essential for this class.

■6.0 CONCLUSION AND FUTURE WORK

The primary motivation for this work is the different results of many techniques used in features selection, although they often use the same datasets. The second motivation is that even these techniques differ; they must be involved in selection of key features able to a large extent to detect the malicious activities in network traffic. The results of this study confirm that, where the two sets of selected features involve mostly the same features. This approach also confirms, in a simple way, that the feature selection is NP-hard problem, where different algorithms and techniques failed to find the optimal subset of features of the KDD cup 1999 dataset in spite of their impressive results. In fact, feature selection is the outcome of several issues, which must be addressed and studied carefully. These issues include attack behavior and pattern, algorithms and techniques efficiency, preprocessing of KDD cup 1999 dataset and the traffic data size.

The two sets of features contain 7 and 8 features. The classification accuracy by using eight features is almost the same as using all dataset features. Future work aims to use different thresholds, different sizes of training and testing data sets of KDD cup 1999 dataset, in addition to try other classification techniques rather than the support vector machine to measure the effectiveness of the study results.

References

[1] Srinivasan, N. & Vaidehi, V. 2007. Reduction of False Alarm Rate in Detecting Network Anomaly Using Mahalanobis Distance and Similarity Measure. *Signal Processing, Communications and Networking, 2007. ICSCN'07. International Conference on*. 2007: IEEE. 366–371.

[2] Zaman, S. & Karray, F. 2009. Collaborative Architecture for Distributed Intrusion Detection System. *Computational Intelligence for Security and Defense Applications. CISDA 2009. IEEE Symposium on*. 2009: IEEE. 1–7.

[3] Ghali, N. I. 2009. Feature Selection for Effective Anomaly-based Intrusion Detection. *IJCSNS International Journal of Computer Science and Network Security*. 9: 285–289.

[4] Wang, W., Gombault, S. & Guyet, T. 2008. Towards Fast Detecting Intrusions: Using Key Attributes of Network Traffic. *Internet Monitoring and Protection, 2008. ICIMP'08. The Third International Conference on*. 2008: IEEE. 86–91.

[5] Hall, M. A. & Holmes, G. 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *Knowledge and Data Engineering, IEEE Transactions on*. 15: 1437–1447.

[6] Münz, G., Li, S. & Carle, G. 2007. Traffic Anomaly Detection Using K-Means Clustering. *Proc. of Leistungs-, Zuverlässigkeits-und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemen*. 4.

[7] Pernkopf, F. 2005. Bayesian Network Classifiers Versus Selective k-NN Classifier. *Pattern Recognition*. 38: 1–10.

[8] Mahoney, M. V. & Chan, P. K. 2003. Learning Rules for Anomaly Detection of Hostile Network Traffic. *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003: IEEE, 601–604.

[9] Li, Y., Wang, R., Xu, J., Yang, G. & Zhao, B. 2009. Intrusion Detection Method Based on Fuzzy Hidden Markov Model. *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*. 2009: IEEE. 470–474.

[10] Zhang, Y., Han, Z. & Ren, J. 2009. A Network Anomaly Detection Method Based on Relative Entropy Theory. *Electronic Commerce and Security. ISECS'09. Second International Symposium on*. 2009: IEEE. 231–235.

[11] Han, L. 2010. Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis. *Intelligence Information Processing and Trusted Computing (IPTC), 2010 International Symposium on*. 2010: IEEE. 458–462.

[12] Kabir, M. M., Shahjahan, M. & Murase, K. 2007. A Backward Feature Selection by Creating Compact Neural Network Using Coherence Learning and Pruning. *Journal ref: Journal of Advanced Computational Intelligence and Intelligent Informatics.* 11: 570–581.

[13] Kent, A., Fisk, M. & Gavrilov, E. 2010. Network Host Classification Using Statistical Analysis of Flow Data.

[14] Brugger, S. T. & Chow, J. 2007. An Assessment of the DARPA IDS Evaluation Dataset using Snort. *UCDAVIS Department of Computer Science.* 1: 2007.

[15] Brown, C., Cowperthwaite, A., Hijazi, A. & Somayaji, A. 2009. Analysis of the 1999 Darpa/Lincoln Laboratory Ids Evaluation Data with Netadhict. *Computational Intelligence for Security and Defense Applications, CISDA 2009. IEEE Symposium on*. 2009: IEEE, 1–7.

[16] Miller, M. 1999. Learning Cost-sensitive Classification Rules for Network Intrusion Detection Using Ripper.

[17] Farid, D. M., Darmont, J., Harbi, N., Nguyen, H. H. & Rahman, M. Z. 2009. Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification. *Proceedings of the International Conference on Computer Systems Engineering (ICCSE 2009)*.

[18] Xuren, W. & Famei, H. 2006. Improving Intrusion Detection Performance Using Rough Set Theory and Association Rule Mining. *Hybrid Information Technology, 2006. ICHIT'06. International Conference on*. 2006: IEEE. 114–119.

[19] Porto-Díaz, I., Martínez-Rego, D., Alonso-Betanzos, A., & Fontenla-Romero, O. 2009. Combining Feature Selection and Local Modelling in the KDD cup 99 dataset. *In Artificial Neural Networks–ICANN 2009.* Springer Berlin Heidelberg. 824–833.

[20] Shrivastava, A., Baghel, M., Gupta, H. 2013. A Novel Hybrid Feature Selection and Intrusion Detection Based On PCNN and Support Vector Machine. *Int.J.Computer Technology & Applications*. 4(6).

[21] Aggarwal, M., Amrita. 2013. Performance Analysis of Different Feature Selection Methods in Intrusion Detection. *International Journal of Scientific & Technology Research*. 2(6).

[22] Mahoney, M. V. 2003. Network Traffic Anomaly Detection Based on Packet Bytes. *Proceedings of the 2003 ACM symposium on Applied computing*, 2003: ACM, 346–350.

[23] Mukkamala, S., Janoski, G., & Sung, A. 2002. Intrusion Detection: Support Vector Machines and Neural Networks. In *Proceedings of the IEEE international joint conference on neural networks (ANNIE)*. 1702–1707.

[24] Mukkamala, S. & Sung, A. H. 2003. Feature Selection for Intrusion Detection Using Neural Networks and Support Vector Machines. *Transportation Research Record: Journal of the Transportation Research Board.* 1822: 33–39.

[25] Chebrolu, S., Abraham, A. & Thomas, J. P. 2005. Feature Deduction and Ensemble Design of Intrusion Detection Systems. *Computers & Security.* 24: 295–307.

[26] Zainal, A., Maarof, M. A. & Shamsuddin, S. M. 2006. Feature Selection Using Rough Set in Intrusion Detection. *TENCON 2006. 2006 IEEE Region 10 Conference*. 2006: IEEE. 1–4.

[27] Zainal, A., Maarof, M. A. & Shamsuddin, S. M. 2009. Ensemble Classifiers for Network Intrusion Detection System. *Journal of Information Assurance and Security*. 4: 217–225.

[28] Wa'el, M., Agiza, H. N. & Radwan, E. 2009. Intrusion Detection Using Rough Sets Based Parallel Genetic Algorithm Hybrid Model. *Proceedings of the World Congress on Engineering and Computer Science*.

[29] Othman, Z. A., Bakar, A. A. & Etubal, I. 2010. Improving Signature Detection Classification Model Using Features Selection Based on Customized Features. *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*. 2010: IEEE. 1026–1031.

[30] Olusola, A. A., Oladele, A. S. & Abosede, D. O. 2010. Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features. *Proceedings of the World Congress on Engineering and Computer Science.* 20–22.

[31] Revathi, M. & Ramesh, T. 2011. Network Intrusion Detection System Using Reduced Dimensionality. *Indian Journal of Computer Science and Engineering.* 2: 61–67.

[32] Srinivasulu, P., Prasad, R. S. & Babu, I. R. 2010. Intelligent Network Intrusion Detection Using DT and BN Classification Techniques. *Int. J. Advance. Soft Comput. Appl.* 2: 124–141.

[33] Chen, R. C., Cheng, K. F., Chen, Y. H. & Hsieh, C. F. 2009. Using Rough Set and Support Vector Machine for Network Intrusion Detection System. *Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on*. 2009: IEEE. 465–470.

[34] Hlaing, T. 2012. Feature Selection and Fuzzy Decision Tree for Network Intrusion Detection. *International Journal of Informatics and Communication Technology (IJ-ICT).* 1(2): 109–118.

[35] Alomari, O., & Othman, Z. A. 2012. Bees Algorithm for Feature Selection in Network Anomaly Detection. *Journal of Applied Sciences Research.* 8(3): 1748–1756.

[36] Chung, Y. Y., & Wahid, N. 2012. A Hybrid Network Intrusion Detection System Using Simplified Swarm Optimization (SSO). *Applied Soft Computing.* 12(9): 3014–3022.

[37] Pundir, S. L., Amrita. 2013. Feature Selection Using Random Forest In Intrusion Detection System. *International Journal of Advances in Engineering & Technology*. 6(3): 1319–1324

[38] Devaraju, S., & Ramakrishnan, S. (2014). Performance Comparison For Intrusion Detection System Using Neural Network With Kdd Dataset. *ICTACT Journal on Soft Computing*. l: 4(3).

[39] Madbouly, A. I., Gody, A. M., & Barakat, T. M. 2014. Relevant Feature Selection Model Using Data Mining for Intrusion Detection System. *arXiv preprint arXiv:*1403.7726

[40] Parvin, H., Minaei, B., Beigi, A., & Helmi, H. 2011. Classification Ensemble by Genetic Algorithms. *In Adaptive and Natural Computing Algorithms.* Springer Berlin Heidelberg. 391–399.

[41] Li, X. 2014. Attribute Selection Methods in Rough Set Theory. Master thesis, San Josй State University.

[42] Khalaf, I. A., Abualkishik, A. M., Aburomman, A. A., Reaz, I., & Bin, M. 2013. Two Features Selection Algorithms Based on Ensemble of SVM Classifier for Intrusion Detection. *Australian Journal of Basic & Applied Sciences*. 7(7).

[43] Raut, A. S., Singh K. R. 2014. Feature Selection for Anomaly-Based Intrusion Detection using Rough Set Theory. *In International Conference on Industrial Automation and Computing* (ICIAC) April, 2014.

[44] Revathi, S. Malathi, A. 2013. Network Intrusion Detection Using Hybrid Simplified Swarm Optimization Technique. *International Journal of P2P Network Trends and Technology (IJPTT)*. 3(8).