International Journal of Research in Advent Technology, Vol.7, No.4, April 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Genetic Algorithm based Feature Selection on Malware System Call Datasets

Devaki Trivedi, Bhavesh Borisaniya
Dept of Information Technology Engineering, Shantilal Shah Engineering College, Bhavnagar
Email:devki.trivedi95@gmail.com,borisaniyabhavesh@gmail.com

Abstract—Feature selection is dominant data mining phase. It is important to diminish dimensionality of dataset by neglecting duplicate and unrelated features. Because higher the size of feature vector requires high processing power to analyze the process. In this paper, we have evaluated the genetic algorithm as a feature selection technique with the aim of reducing the size of feature vector by selecting only important features from the standard malware system call dataset. To check efficiency of the selected feature subset, the reduced feature vector tested using J48 classifier.

Index Terms-Feature Selection, MVSR, Genetic Algorithm, J48 classifier.

1. INTRODUCTION

The objective of feature selection is to display large number of features with reduced subset of features. Feature selection method is used to select related features and discard irrelevant features [1]. *Embedded*, *Wrapper* and *Filter* are standard approaches for Feature selection [12]. In the filter method, selection of a feature is classifier independent. While in the wrapper method classifier involved for feature selection process. Wrapper method use machine learning algorithm for selection of subset features. Wrapper approach is expensive but gives more accurate results than filter. Embedded method has similar working as wrapper method. Difference in embedded method and wrapper method is that embedded method uses intrinsic model for metric building during learning [12]. L1 (Lasso) regularization and Genetic algorithm are example of Embedded method.

Ensemble Attribute Selection Method [1] with discriminatory and characteristic premises for malware detection. Depending on categorized and uncategorized data, the most related features with respect to the class is extracted [1]. Every feature in the classification increases the computational cost and time. So, it is necessary to work with fewer subset of features. Unrelated and duplicate attributes lower the rate of detection. As a necessary data mining phase, feature selection can discard unrelated and duplicate attributes from the large feature subset. Therefore, it can enhance detection rate and decrease false positive rate for malware detection. System calls are operating system functions used by program to make request of service to the kernel. Sequence of system calls can describe behavior of the process. Order of system call sequence is more important to modeling malware behavior. As length of system call (Term-size) increase, size of feature vector also increases [13]. Higher no of feature requires high processing power to analyze the process [13]. So, there is a need for applying feature selection on malware system call dataset that can select important features while retaining the classification accuracy.

The nature inspired algorithm used for feature selection can provides better results. John Holland in 1975 had proposed genetic algorithm (GA) and for the evolution of population in a specific environment used as a prototype [3]. Many optimization problems can be solved using genetic algorithm. Kaya et al. [4]-[5] for classifying ECG arrhythmias in several research used GA as a feature selection method for choosing important features. In further research, GA used to extract features from the airborne visible imaging spectrometer data [6]. Oh et al. [7] proposed a hybrid GA to choose attributes and proved that it has greater outcome than the classical GA method. They used a GA and Random-Forest depends on proposal for achieved result in esophageal cancer dataset. They applied GA to select features. Aniconic and Subasi [8] prefer an approach for cancer diagnosis by selection of features using GA to discard inefficient attributes.

The remaining section of this paper organized as follows. Section 2 describes genetic algorithm. Section 3 includes related work found in literature. Section 4 includes details about Evaluation Methodology and datasets used for this experiment. Section 5 covers experiment results and discussion about results. Section 6 concludes this research work and future direction of work.

2. GENETIC ALGORITHM

Genetic algorithm (GA) is a fluctuating heuristic search method [9]. It is a randomized search technique influenced by advancement theory of Darwin's – "Endurance of the fittest" [9]. Genetic algorithm starts with a collection of chromosomes known as population. Collection of genes could be bits, numbers or characters restrain by individual chromosome. Based on the reproduction, fitness value chromosome is chosen. As the fitness value is higher there is high opportunity of a chromosome being stipulate [10]. New population can be produced by crossover and mutation. Development of population go faster by crossover. Lost data of population can be recovered by global or local search is

possible in mutation. Iteration will be repeated until stopping condition is satisfied or optimal solution is achieved [11].

3. RELATED WORK

Feature selection is the process of minimize dimensionality of dataset by selecting features from the original feature subset [12]. There are number of approaches found in literature that uses genetic algorithm for feature selection. However, here we discuss few notable approaches that uses GA for feature selection in their work.

Ferriyan et al. [14] addressed feature selection based on genetic algorithm for intrusion detection systems. They used one-point crossover instead of two-point crossover applied on previous work. Because it is faster to use one-point crossover than two-point crossover. They used NSL-KDD Cup 99 dataset for their experiment. They applied five classifiers on datasets in the absence of feature selection. They explored random forest gave the better results in terms of the training time and classification rate. Then they used feature selections, in that they applied one crossover instead of two crossovers using classifier Random Forest and concluded that their parameters gave good results.

Suman Khatwani and Arti Arya [15] improved institution performance by predicting learners' performances to find the weak areas to guide their students. They work on genetic algorithm and decision tree to investigate learner's performance. Multiple decision tree is produced using ID3 Algorithm, each tree based on the distinct feature set to forecast the achievement of a student. To achieve better results in terms of accuracy, genetic algorithm was also included. On the n -ary trees Genetic algorithm used. GA designing fitness of each tree and crossover operations to achieve multiple generations. As the generation grows trees produced with a better fitness and at the end with the best accuracy decision tree is produced.

Senthilnayaki et al. [16] present algorithm to detect the network attacks whether it is anomaly or normal by using classification and pre-processing. They perform feature selection with the help of genetic algorithm, which helps in pre-processing and for classification modified J48 is used. The algorithm presented by author can detect features that are necessary for normal and anomaly records classification. Results of their experiment shows that the GA and modified J48 gives better accuracy of detection than the methods present in previous research in terms of reduced false rate and for detection rate.

Fei He et al. [17] described framework combining Information Gain and Genetic Algorithm for feature selection. For generation of feature subset cross propagation was used. In the case of large number of attributes on UCI data sets the hybrid algorithm perform better than the other methods.

Bidi and Elberrichi [18] produced genetic algorithm as a feature selection for distinct text representation methods. In the first select subset of features that provides better performance in classifier. In another way searches a subset of attributes that reduce dimensionality and provides best accuracy result in classification. They concluded that, feature selection using genetic algorithm provide good performance result in text classification with reduce dimensionality using F-measure. They performed evaluation on two dataset Reuter-21578 and 20Newsgroups.

Ketan Desai and Roshni Ade [19] proposed an approach using genetic algorithm to select features from NSL KDD data set. Selection of features is done using techniques like CFS, IG and CAE and their performance is tested using Naïve Bayes and J48. From the experiment result, it can be observed that their method reduces attributes from the dataset results into better accuracy for classification.

4. EVALUATION

4.1. Datasets

The Experiments carried out on two sets of standard malware system call datasets namely Anubis [25]and ADFA datasets. Anubis dataset defined by Davide Canali et al. [20] consists system calls of benign process labelled as goodware and malware. Table 1 provides the details of sub datasets found in Anubis dataset. The malware dataset consists the traces of different types of malware collected from Anubis. The goodware dataset includes the execution traces extracted from 10 distinct real-world machines. Anubis-good consist 36 benign application traces executed under Anubis. The malware-test includes the malware sample traces collected from machines other than Anubis. The dataset is collected in 1-gram format.

Table 1. Number of traces and system calls in Anubis dataset

Dataset	**Number of Traces**	**Number of System Calls**
Malware	5,855	3,28,99.160
Goodware	612	65,55,20,685
malware-test	1,133	13,18,8452
Anubis-good	36	44,127
Total	**7,636**	**70,16,52,424**

Table 2. Traces of system call in distinct category of ADFA-LD and ADFA-WD dataset

Data Type	**ADFA-LD**		**ADFA-WD**	
	Traces	System Calls	Traces	System Calls
Training Data	833	3,08,077	355	**1**,35,04,419
Validation Data	1,372	21,22,085	1,827	11,79,18,735
Attack Data	746	3,17,388	5,542	7,42,02,804
Total	5,951	27,47,550	7,724	20,56,25,958

The ADFA dataset consists of two datasets namely ADFA-LD (Linux Dataset) and ADFA-WD (Windows Dataset). This dataset is built by G. Creech et al. [21, 22].

Table 2 provides description about system call traces extracted from [21] for ADFA-LD and ADFA-WD dataset for each category. Traces of system call for distinct kind of attacks included in ADFA-LD. Windows dataset (ADFA-WD) presents collection of system calls for a various attacks and DLL access request.

Table 3 and table 4 describes details of each attack class in ADFA-LD and ADFA-WD dataset respectively [21, 22].

Table 3. Attack class details of ADFA-LD dataset

Attack	Payload/Effect	Trace Count
Hydra-FTP	FTP by Hydra – Password bruteforce	162
Hydra-SSH	SSH by Hydra – Password bruteforce	176
Adduser	Client-side poisoned executable – Add new superuser	91
Java-Meterpreter	TikiWiki vulnerability exploit – Java based meterpreter	124
Meterpreter	Client side poisoned executable	75
Webshell	PHP remote file inclusion vulnerability	118

Table 4. Attack class details of ADFA-WD dataset

ID	Vulnerability Exploited and Exploit Mechanism	Trace Count
V1	CVE:2006-2961 - Reverse Ordinal Payload Injection	454
V2	EDB-ID: 18367 - Upload and execute malicious payload using Xampp_webdav	470
V3	CVE: 2004-1561 - Metasploit exploit	382
V4	CVE: 2009-3843 - Metasploit exploit	418
V5	CVE: 2008-4250 - Metasploit exploit	355
V6	CVE: 2010-2729 - Metasploit exploit	454
V7	CVE: 2011-4453 - Metasploit exploit	430
V8	CVE: 2012-0003 - DNS Spoofing using Pineapple	487
V9	CVE:2010-2883 - Reverse Shell spawn through malicious PDF	440
V10	Backdoor - Reverse Inline Shell spawned	536
V11	CVE: 2010-0806 - Metasploit exploit	495
V12	Infectious Media - Blind Shell spawned	621

4.2. Experiment Methodology

In the first step we have extracted the features from the system call traces dataset using MVSR [13]. For each dataset term size 1, 2 and 3 used for carried out these experiments. Then for classification we have applied J48 classifier. We have evaluated Genetic algorithm as a feature selection method to select relevant and important features from the system call dataset. Size of original feature vector is reduced by applying genetic algorithm on subset of feature. To check efficiency of the selected features by genetic algorithm the reduced dataset tested using J48 classifier.

5. RESULTS AND DISCUSSION

For experiment, features extracted from system call trace dataset using MVSR. Term-size 1, 2 and 3 selected for these experiments. Classifier J48 applied on the MVSR. We have selected features from dataset using GA. Classifier used to classify selected features. We have compared results obtained by MVSR without applying GA and results with GA.

Table 5 describes experiment result on Anubis dataset. From the results, we can observe that around 50 % features are reduced while accuracy is almost retained. For MVSR, FP-Rate is low and accuracy is almost same for term-size 2 and 3. lowest FP-Rate for term-size 3 in GA.

Table 6 shows experiment result on ADFA-LD dataset. It describes no of selected features, obtained accuracy and FP-Rate result for MVSR and GA. With the help of this results we can analyze that approximately 47%, 49% and 50% and features are reduced for term-size 1,2 and 3 respectively, while there is no major difference in accuracy. In MVSR term-size1 provides lowest FP-Rate result and highest accuracy. For GA, term-size 2 provides lowest FP-Rate and highest accuracy result.

Table 7 shows result on ADFA-WD dataset. From the results, we can deduce that, alike results of Anubis dataset here also only half of the features are selected while accuracy is almost the same as we get with the full set of features. Term-size 3 provides highest accuracy and Term-size 1 provides lowest FP-Rate for MVSR. For GA Term-size 1 provides best accuracy result and Term-size3 provides lowest FP-Rate result.

Table 8 and Table 9 shows multiclass result of MVSR and GA for term-size1, term-size2 and term-size3 on ADFA-LD and ADFA-WD dataset respectively.

Figure 1 shows ROC curves of MVSR and GA results for term-size 1 (a), term-size 2 (b) and term-size 3 (c) on Anubis dataset. From the AUC we can observed that although no of feature reduces, there is no major difference in accuracy. Same can be concluded for experiment results of term-size 1 to 3 on ADFA-LD and ADFA-WD from ROC curves as shown in figure 2 and figure 3 respectively.

Figure 4 and figure 5 shows ROC curves for MVSR and GA multiclass results of term-size 3 on ADFA-LD and ADFA-WD dataset respectively. Due to space constraint we have not shown the multiclass ROC for term-size 1 and term-size 2 of ADFA-LD and ADFA-WD datasets. From the results it is evident that classifier is not able to perform well on multiclass dataset with both MVSR and GA. However, compare to MVSR reduction in number of features didn't affect the average accuracy and FP rate results.

Table 5. Anubis dataset experiment results

	Selected Features			Accuracy			FP-Rate		
	Term 1	**Term 2**	**Term 3**	**Term 1**	**Term 2**	**Term 3**	**Term 1**	**Term 2**	**Term 3**
MVSR	65	2188	27508	99.22	99.39	99.39	0.068	0.052	0.052
GA	32	1230	15896	99.05	99.22	99.14	0.085	0.071	0.064

Table 6. ADFA-LD dataset experiment results for binary class classification

	Selected Features			Accuracy			FP-Rate		
	Term1	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**
MVSR	176	3793	24818	95.98	95.84	95.21	0.154	0.176	0.219
GA	94	1922	12404	95.31	95.61	95.46	0.193	0.189	0.206

Table 7. ADFA-WD dataset experiment results for binary class classification

	Selected Features			Accuracy			FP-Rate		
	Term1	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**
MVSR	1310	4802	13473	92.85	92.81	93.00	0.129	0.132	0.133
GA	685	2398	6777	91.64	91.59	92.46	0.164	0.178	0.140

Table 8. ADFA-LD dataset experiment results for multiclass classification

	Selected Features			Accuracy			FP-Rate		
	Term1	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**
MVSR	176	3793	24819	92.43	92.08	91.08	0.204	0.205	0.240
GA	90	1916	12496	91.74	91.64	91.49	0.268	0.252	0.240

Table 9. ADFA-WD dataset experiment results for multiclass classification

	Selected Feature			Accuracy			FP-Rate		
	Term1	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**	**Term1**	**Term2**	**Term3**
MVSR	1310	4802	13473	49.58	49.55	50.03	0.060	0.061	0.062
GA	677	2376	6773	48.32	46.65	49.90	0.062	0.103	0.071

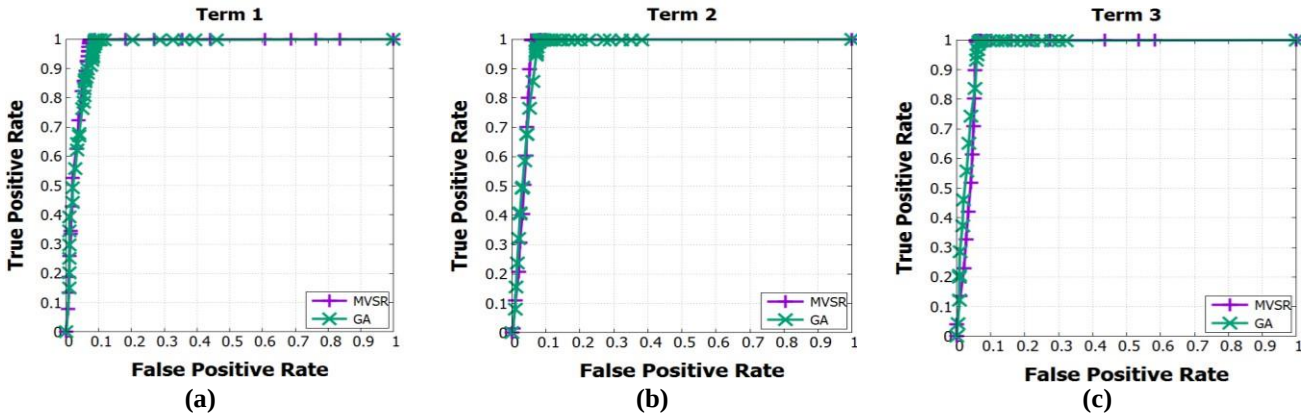

Fig 1. ROC curves of MVSR and GA on Anubis dataset for (a) term-size 1, (b) term-size 2, and (c) term-size 3

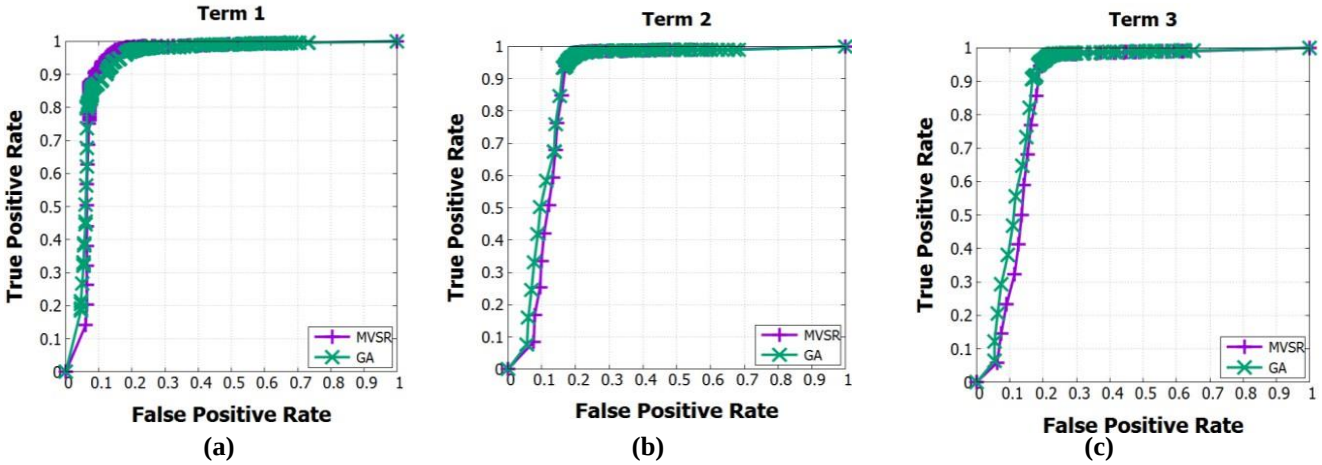

Fig 2. ROC curves of MVSR and GA on ADFA-LD dataset for (a) term-size 1, (b) term-size 2, and (c) term-size 3

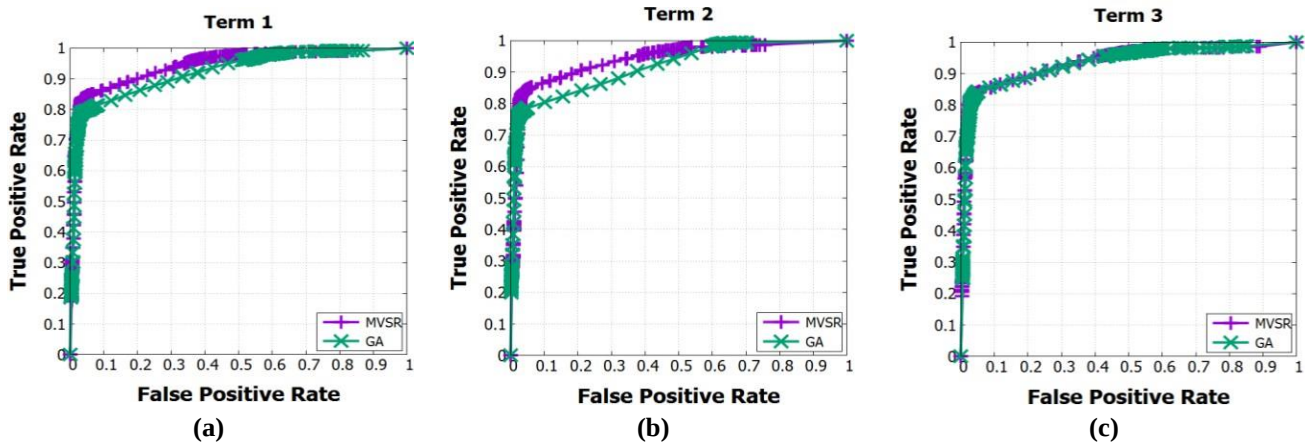

Fig 3. ROC curves of MVSR and GA on ADFA-WD dataset for (a) term-size 1, (b) term-size 2, and (c) term-size 3

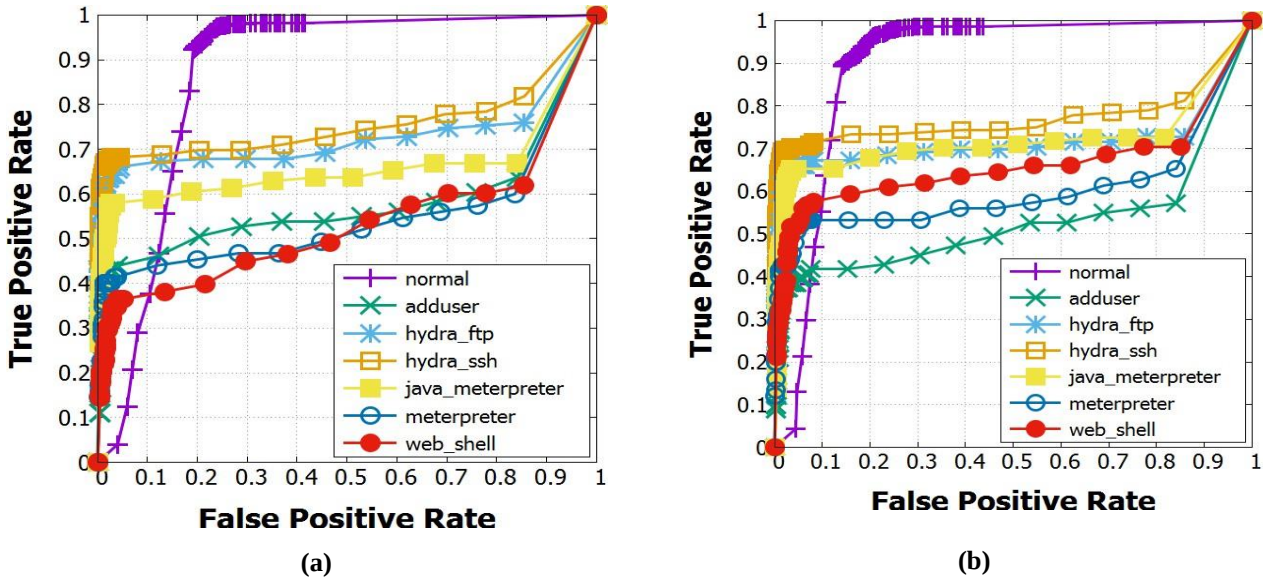

Fig 4. ROC curves of (a) MVSR and (b) GA multiclass classification result on ADFA-LD dataset for term-size 3

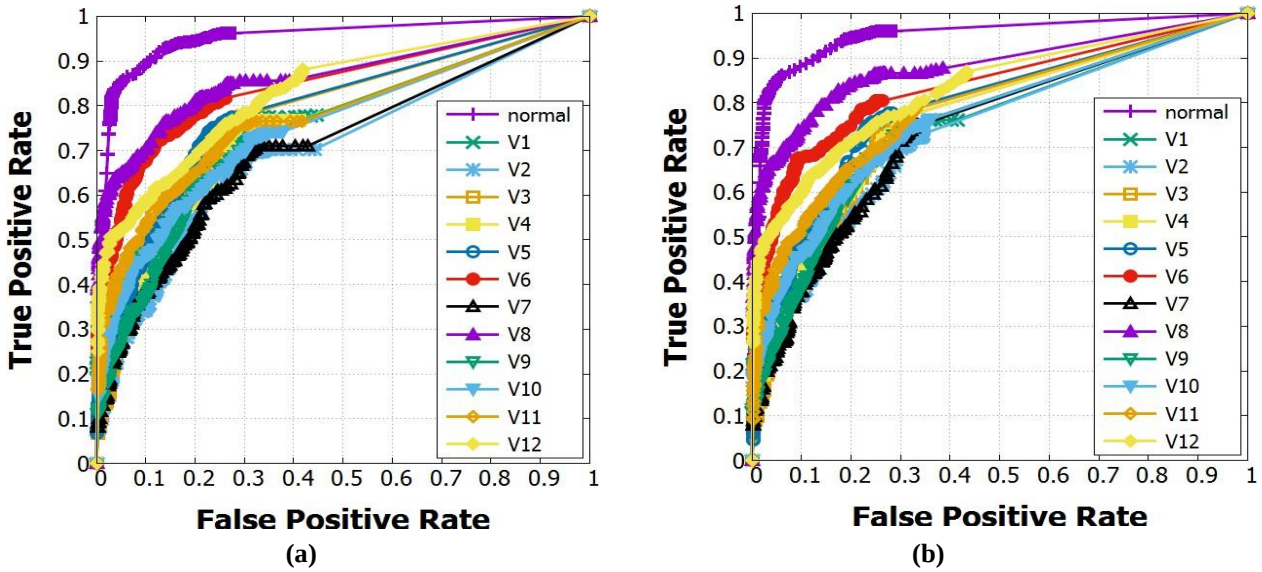

Fig 5. ROC curves of (a) MVSR and (b) GA multiclass classification result on ADFA-LD dataset for term-size 3

6. CONCLUSION AND FUTURE DIRECTIONS

Feature Selection is a data mining technique, which is used to select important feature by removing redundant and irrelevant features from the dataset. We used genetic algorithm as a feature selection method to select relevant features from feature set of malware system call traces. The features of malware system call traces are extracted using modified vector space representation (MVSR) method. We have applied J48 decision tree classifier on the reduced subset of features. Decision tree is significantly affected by the added attribute set because it helps the tree generation in more efficient manner. We have compared the results achieved by MVSR and Genetic Algorithm based feature selection method. From the obtained experiment result we can conclude that by using genetic algorithm number of features reduced by approximately 50%. while there is no major difference in accuracy result. This research work can be further evaluated towards applying genetic algorithm-based feature selection for different classifiers to enhance the performance on different malware datasets.

REFERENCES

[1] Totorkkola K. (2003): Feature Extraction by Non-Parametric Mutual Information Maximization, J. Mach.learn. Res, Vol.3, PP.1415-1438.

[2] iang Zhao and Huang. (2011): A Feature Selection Method for Malware Detection, International Conference on Information and Automation Shenzhen, china IEEE.

[3] Holland, J. (1975): Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, vol. 1, no. 1.

[4] Kaya, Y.; Pehlivan, H.; Tenekeci, H. (2017): Effective ECG beat classification using higher order statistic features and genetic feature selection, Biomed. Res., vol. 28, no. 17.

[5] Kaya, Y.; Pehlivan, H. (2015): Feature Selection Using Genetic Algorithms for Premature Ventricular Contraction Classification, in 9th International Conference on Electrical and Electronics Engineering, ELECO.

[6] Yu, S.; De Backer, S.; Scheunders, P. (2002): Genetic feature selection combined with composite fuzzy nearest neighbour classifiers for hyperspectral satellite imagery, Pattern Recognise, 23(1–3), pp. 183–190.

[7] Seok, Oh.; Jin-Seon, Lee.; Byung-Ro, Moon. (2004): Hybrid genetic algorithms for feature selection, IEEE Trans. Pattern Anal. Mach. Intell, 26(11), pp. 1424–1437.

[8] Alickovic, E.; Subasi, A. (2017): Breast cancer diagnosis using GA feature selection and Rotation Forest, Neural Computer. Appl., vol. 28, no. 4, pp. 753–763.

[9] Wei Li (2004): Using Genetic Algorithm for Network Intrusion Detection, Proceedings of the United States Department of Energy Cyber Security Grou, Training Conference, Vol. 8, pp. 24-27.

[10] Anup Goyal.; Chetan Kumar. (2008): GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System.

[11] Bharat, S.; Dhak.; Shrikant Lade. (2012): An Evolutionary Approach to Intrusion Detection System using Genetic Algorithm, International Journal of

Emerging Technology and Advanced Engineering, Volume 2, Issue 12.

[12] Aksoy, S. (2008): Feature Reduction and Selection, Department of Computer Engineering, Bilkent University, CS 551.

[13] Borisaniya, B.; Patel. D. (2015): Evaluation of modified vector space representation using adfa-ld and adfa-wd datasets, journal of information security.

[14] Ferriyan, Andrey et.al. (2017): Feature Selection Using Genetic Algorithm to Improve Classification in Network Intrusion Detection System, International Electronics Symposium on Knowledge Creation and Intelligent computing.

[15] Suman Khatwani.; Arti Arya, D. (2013): A Novel Framework for Envisaging a Learner's Performance using decision trees and genetic algorithm, International Conference on computer communication and informatics.

[16] Senthilnayaki et.al (2013): An Intelligent Intrusion Detection System Using Genetic Based Feature Selection and Modified j48 Decision Tree Classifier "Fifth international conference on advance computing, IEEE.

[17] Fei He et al (2016): A Hybrid feature Selection Method Based on Genetic Algorithm and Information Gain"5th international conference on Computer science and Network Technology, IEEE.

[18] Bidi.; Elberrichi (2016): Feature Selection for Text Classification Using Genetic Algorithms,8th international conference on modelling, identification and control, IEEE.

[19] Desale.; Ade. (2015): Genetic Algorithm based Feature Selection Approach for effective intrusion detection system, International conference on computer communication and informatics, IEEE.

[20] Canali, D.; Lanzi, A.; Balzarotti, D.; Kruegel, C.; Christodorescu M.; Kirda E. (2012): A quantitative study of accuracy in system call-based malware detection, in Proceedings of the International Symposium on Software Testing and Analysis, ser. ISSTA.New York, NY, USA: ACM, pp. 122–132.

[21] Creech, G.; Hu, J. (2013): Generation of a New IDS Test Dataset: Time to Retire the KDD Collection. Wireless Communications and Networking Conference, Shanghai, 7-10.

[22] Creech, G.; Hu, J. (2014): A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontinuous System Call Patterns. IEEE Transactions on Computers, 63, 807-819.

[23] Creech, G. (2014): Developing a High-Accuracy Cross Platform Host-Based Intrusion Detection System Capable of Reliably Detecting Zero-Day Attacks. Ph.D. Dissertation, University of New South Wales, Sydney.

[24] Anubis - malware analysis for unknown binaries. [Online]. Available: https://anubis.iseclab.org/

[25] The ADFA Intrusion Detection Datasets. http://www.cybersecurity.unsw.adfa.edu.au/ADFA IDS Datasets/