ISSN 2278-3091

Volume 9, No.1.1, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse7791.12020.pdf

https://doi.org/10.30534/ijatcse/2020/7791.12020

Classification of Metamorphic Virus using N-grams Signatures

Isredza Rahmi A Hamid[1]*, Nur Sakinah Md Sani [1], Zubaile Abdullah[1], Cik Feresa Mohd Foozy[1], Kuryati Kipli[2]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor Malaysia

[2]Faculty of Engineering, Universiti Malaysia Sarawak, Kota Samarahan 94300, Malaysia

*Corresponding author E-mail: rahmi@ uthm.edu.my

ABSTRACT

Metamorphic Virus has a capability to change, translate, and rewrite its own code once infected the system. The computer system can be seriously damage by undetected metamorphic virus. Due to this, it is very vital to design a metamorphic virus classification model. This paper focused on detection of metamorphic virus using Term Frequency Inverse Document Frequency (TF-IDF) technique. This research was conducted using Second Generation virus dataset. The first step is the classification model to cluster the metamorphic virus using TF-IDF technique. Then, the virus cluster is evaluated using Naïve Bayes algorithm in terms of accuracy. The dataset have different types of class and features used extracted from bi-gram assembly language. The result shows that the proposed model was able to classify metamorphic virus using TF-IDF with optimal number of virus group.

Key words: Metamorphic virus; classification; Term Frequency Inverse Document Frequency (TF-IDF

1. INTRODUCTION

Currently, security threat has become vicious and countermeasure must be taken seriously. The number of security threat towards the user is increasing each year. The virus inventor becomes more creative in order to penetrate the system. Once the virus was in the system, it will either corrupting the system or remains dormant until it gets to attack the target. Thus, the system must become more alert towards the virus intrusion in order to protect it from the virus attack.

Metamorphic virus has capabilities to change, translate, and rewrite its own code when it infects a system. It is the most viral and if it is not detected earlier the system can be seriously damage. The difference between Polymorphic and Metamorphic virus is that the Polymorphic Virus keeps the original code and only encrypt the code. The Metamorphic Virus is much more complex and requires programming expert to create this virus [1]. This research has three main objectives, which are:

- To design a virus detection model on metamorphic virus using static detection.
- To classify metamorphic virus using Term Frequency Inverse Document Frequency.
- To evaluate the proposed model using Naïve Bayes algorithm in terms of accuracy and efficiency.

The rest of the paper is organized as follows: Section 2 describes the related work on metamorphic virus detection techniques. Section 3 presents the proposed classification model for metamorphic virus detection based on Term Frequency Inverse Document Frequency (TF-IDF). Section 4 shows experimental setup. Section 5 will discuss about the result from the experiment. Finally, Section 6 concludes the work and highlights a direction for future research.

2. RELATED WORK

During the former phases of virus creation, virus programmers tried to infect a large number of victims. Virus was created similar in type of infection, but the malicious actions performed were different. However, the methods employed to infect a host machine and spread to other machines were similar to all virus. Most of the early stage of virus detection was discovered based on its signature files and activities performed by the virus. As virus detection systems managed to detect and stop the infections with increasing strength, virus programmers started to implement new methods to spread the virus infections [2]. The evolution of virus becomes more advanced that it produce virus that used encryption technique to obfuscate their presence. This makes the virus existence unclear to confuse the virus detector.

Metamorphic virus changed its code while propagate. Thus, it can avoid detection by static signature-based virus scanners. This leads to possibility of undetectable breed of malicious programs. Moreover, static analysis metamorphic virus also uses code obfuscation techniques which could beat dynamic analyzers, such as emulators. Hence, the metamorphic virus managed to alter its behavior when discovered executing under a manipulated environment. The metamorphic virus used several metamorphic transformations to differ the visual aspect, such as register usage exchange, code permutation, code expansion, code shrinking, and garbage code insertion [4]. Metamorphic virus also capable to create a new generation that looks different to their parents.

Table 1 shows the comparison between three virus detection approaches. Signature based is the most efficient approach as compared to anomaly based and code emulation in term of detection strength, accuracy and low at cost. However, it is only limited to new malware variant.

Table 1: Comparison of virus detection approaches [5]

Methods / Parameter	Signature based	Anomaly Based	Code Emulation
Strength	Efficient	New malware	Encrypted virus
Limitation	New malware	Unproven	Complex
Cost	Low	High	High
Accuracy	More database if updated	Less	More

Qu used behaviour-based features consists of 602 malware from the VCL family to create the signature of the virus family [6]. The algorithm used is backward regression model and regression model. The regression model was used to determine the identification of VCL malware and act as indicator to show the influential of VCL malware. The backward regression model achieved 90.3% accuracy in identifying VCL malware.

Kuriakose [7] used feature selection method to detect the presence of metamorphic virus. This research used 3344 malware sample and 1218 benign of 32 win XP. A significant bi - gram of variable lengths is used for constructing learning models using AdaboostM1 (using J48 as base classifier) and Random forest with default settings in WEKA [8]. They managed to achieve 99.8% and 92.6% for benign and virus detection.

Shabani [8] used the Bayesian Network features to detect metamorphic virus tested on 600 samples. Bayesian Network learning is known as a NP-hard problem because it's utilizing exploratory research proved that was helpful in many learning approach although it does not guarantee optimistic result. They used Hill climbing algorithm because it is a popular algorithm just because of exchanging between computational demands and the quality of the model [10]. Their approach managed to achieve above 90% accuracy.

Our work differ than other researchers [6][7][8] as shown in Table 2 in such a way that we used Second Generation virus dataset. Then, the dataset was classified using Term Frequency Inverse Document Frequency (TF-IDF) algorithm and tested with Naïve Bayes classification algorithm. We used bi-gram features to identify that metamorphic virus classes based on its own features can distinct themselves from each other.

Table 2: Virus classification approach

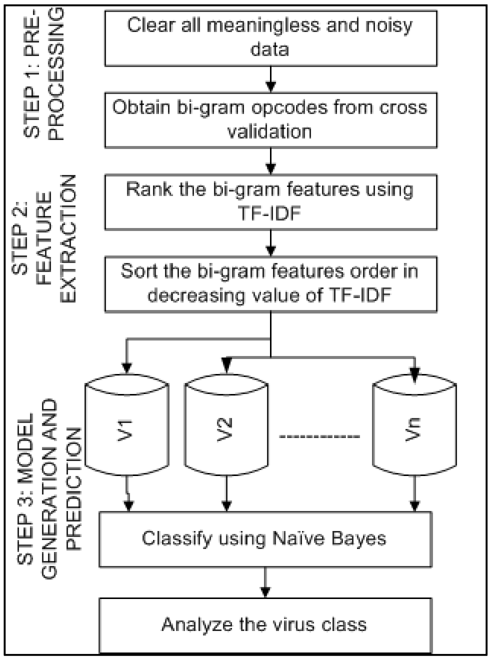
Work By	Features Used	Sample Used	Machine Learning Algorithm	Results (Acc)
Qu et al [6]	Behaviour-Based feature.	602	Backward Logistic Regression Model or Logistic Regression Model	90.3%
Kuriakose et al [7]	Feature Selection	3344 malware and 1218 benign	Adaboost and Random Forest	Benign - 99.8% Virus - 92.6%
Shabani et al [8]	Bayesian Network	600	Hill Climbing	Above 90%

3. METAMORPHIC VIRUS CLASSIFICATION MODEL

This section explains about the metamorphic virus classification model. The proposed model consists of three important phases including the pre-processing, feature extraction, and model generation with prediction. Figure 1 shows the flow of the proposed virus classification model.

3.1. Step 1: Pre-processing

The pre-processing phase involves raw data cleansing. The dataset consists of 152 viruses which 96 viruses from Second Generation. Data describing allows the distribution of data values, while data transformation help in performing calculation on existing columns. The pre-processing involves data mining techniques to train the classification models with set of rule based about the virus function. Then, the model will be trained and used to classify the testing data. Normally, data mining techniques is used for large of datasets for pattern detection [8]. The final process in data preparation is data sampling, which help the creation of training and to validate the datasets.


STEP 1: PRE-PROCESSING
Clear all meaningless and noisy data
Obtain bi-gram opcodes from cross validation
STEP 2: FEATURE EXTRACTION
Rank the bi-gram features using TF-IDF
Sort the bi-gram features order in decreasing value of TF-IDF
STEP 3: MODEL GENERATION AND PREDICTION
V1
V2

Vn
Classify using Naïve Bayes
Analyze the virus class

Figure 1: Virus Classification Model

3.2. Step 2: Feature Extraction

The feature extraction process is to create set of new features that can be used for classification. First, we obtained the bi-gram feature value. Then, the bi-gram feature was further classified using Term Frequency Inverse Frequency Document (TF-IDF). The TF-IDF algorithm gives weight value for each word in the whole document as shown in Figure 2. TF-IDF method allows each word to be considered as important and is inversely proportional on how often it occurs in whole document.

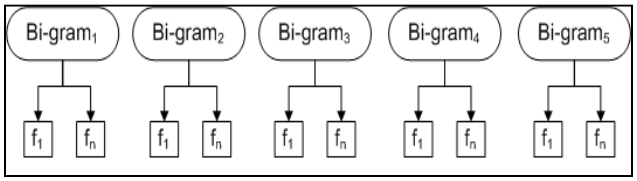


Figure 2: Classification of dataset

The classification using Term Frequency Inverse Document Frequency has three main steps:

1. The frequency of bi-gram features was calculated for the whole dataset.
2. The frequency value was normalized to avoid biased result by calculating the bi-gram feature value (Tf) divide by maximum bi-gram value in the dataset as shown in Equation 1.

$$Tf / Max(Tf \dots n) \quad (1)$$

3. Then, calculate the Inverse Document Frequency(IDF) using Equation 2, where, N is total number of documents in the corpus, $\{d \in D: \in d\}$ is number of documents where the term appears and $(t, d) \neq 0$ if the term is not in the corpus. This will lead to a division-by-zero.

$$\mathrm{idf}(t, d) = \log N / (|\{d \in D : t \in d\}|) \tag{2}$$

3.3. Step 3: Model Generation and Prediction

Term Frequency is used to categorize the text document. This method does not involve any binary values. Generally, Term Frequency means number of times the word or term$t$, exists in a document, d. To get a better result, the Term Frequency will be divided with maximum number of raw Term Frequency according to the length of the document. This can be simplified by Equation 3.

$$tf, \mathrm{Normalized} = \frac{tf}{\max/(tf, dl1 \ldots tf, dln)} \tag{3}$$

The second step will be the calculation the IDF using Equation 4.

$$\log\,(n, d/\mathrm{countif}(tf, d1 \ldots tf, dn)) \tag{4}$$

The last step is to calculate the Term Frequency Inverse Document Frequency by multiplying the values of tf, normalized with idf value. Finally, the TF-IDF value will be obtained through Equation 5.

$$tf - idf = (tf, \mathrm{Normalized} * idf) \tag{5}$$

4. EXPERIMENTAL SETUP

This section discuss the experimental setup for metamorphic virus classification

4.1. Metamorphic Virus Detection System

In our study, the virus classification was performed using WEKA. The experiment will show the accuracy and effect of classification algorithm used on the dataset. This experiment uses the Second Generation Virus Kits which has 52 features. The dataset contained long string of uni-gram assembly language. However, the uni-gram feature does not have any significant meaning for virus classification. So, the bi-gram feature is selected for this experiment. The bi-gram features are calculated using Term Frequency Inverse Frequency Document (TF-IDF) to classify the virus class. We only used the first five bi-gram features from the whole string to show that the TF-IDF algorithm is the effective method to classify virus.

4.2. Performance Metric

In order to measure the effectiveness of the classification approach, we refer to three possible outcomes as: True Positive (TP) and False Positive (FP) and Receiver Operating Curve (ROC).
True positive is where the number of correctly identified as metamorphic virus.

$$TP = \frac{TP}{TP + FN} \tag{6}$$

While False positive is where the number of wrongly identified as metamorphic virus.

$$FP = \frac{FP}{TN+FP} \tag{7}$$

ROC value is where the classification made and algorithm used can be determine it certainty. The best result of ROC value should be close to one. This shows that the datasets has equal sensitivity and specificity.

Table 3: TP and FP rate of Second Generation Virus Kits

bi-gram	Features	Number of class	TP Rate	FP Rate
1	movint	20	0.854	0.01
	callpop	13	0.958	0.003
	jmpmov	13	0.906	0.05
	jmpcall	3	0.99	0.24
2	intmov	6	0.969	0.018
	subpush	6	0.948	0.017
	movsub	6	0.979	0.016
	movadd	6	0.969	0.007
	movxor	7	0.927	0.067
	addadd	6	0.948	0.012
	popsub	5	0.948	0.051
	addsub	5	0.958	0.358
	addinc	4	0.979	0.101
	xoradd	4	0.979	0.201
	xorinc	3	0.958	0.958
3	Pushpush	10	0.979	0.001
	Subpush	7	0.948	0.004
	Submov	7	0.938	0.005
	Incloop	4	0.958	0.002
	loopcall	5	1	0
	Sublea	4	0.854	0.071
	subloop	4	0.958	0.313
	loopmov	5	0.979	0.146
	pushmov	10	0.938	0.004
	Incinc	3	0.969	0.001
	addloop	3	0.979	0
4	movlea	18	0.854	0.027
	movint	9	0.958	0.052
	Intmov	18	0.906	0.015
	callpop	9	0.885	0.018
	intcmp	10	0.927	0.006
	pushpush	17	0.844	0.009
	intpush	11	0.885	0.008
	poppush	10	0.917	0.007
	popsub	6	0.969	0.003
	loopcall	6	0.979	0.057
	intjz	2	0.99	0.99
	pushmov	15	0.844	0.017
5	leamov	22	0.813	0.024
	intmov	22	0.833	0.028
	popmov	16	0.896	0.012
	intpush	16	0.865	0.094
	pushpush	18	0.844	0.015

	jmov	11	0.917	0.017
	subpush	12	0.927	0.014
	movsub	12	0.875	0.01
	cmpjz	12	0.875	0.054
	pushlea	10	0.906	0.02
	poppush	15	0.854	0.077
	sublea	4	0.958	0.076
	popsub	7	0.948	0.181
	pushmov	18	0.823	0.017

5. RESULT AND DISCUSSION

This section discusses number of class and performance for each bi-gram in each datasets.

5.1. TF and FP rate

Table 3 shows the bi-gram features and number of classes for five bi-gram intended for Second Generation virus kit dataset. The dataset has different types of bi-gram features and total number of classes. We selected the first five bi-gram features from the whole string to show that the TF-IDF algorithm is the effective method to classify virus.

For the first bi-gram, jmpcall has the highest value of True Positive and False Positive rate with 0.99 and 0.24 respectively. However, callpop has better result compared to jmpcall because it has only slightly lower True Positive rate at 0.958 plus they have lower False Positive rate at 0.003.

As for second bi-gram, movsub, addinnc, and xoradd features have highest True Positive value with 0.979. Moreover, movadd has the lowest value for FP rate and xorinc has the highest value for False Positive rate with 0.958.

The loopcall features in third bi-gram achieved the highest True Positive rate and lowest False Positive rate that are 1 and 0 respectively. Other features also have high value of True Positive which shows that the virus classification is more diverse and correctly classified.

In fourth bi-gram, intjz feature has highest True Positive and False Positive value with 0.99 and 0.99 correspondingly. Thus, posub feature produce better result as it has high True Positive rate at 0.969 and lower False Positive rate at 0.003.

The sublea feature in fifth bi-gram has the highest True Positive value with 0.958. However, popsub has better result as though it has slightly lower True Positive rate at 0.927 compared to subpush, but it has much lower False Positive rate at 0.014.

5.2. ROC Value

The ROC value should be close to one to be considered as good. Figure 3 to 7 shows the ROC value for all bi-grams for Second Generation virus kit dataset. All features in fifth bi-grams shows the highest ROC value as compared to other bi-grams. This demonstrates that the dataset has equal sensitivity and specificity when classify using fifth bi-grams.

6. CONCLUSION

The classification of metamorphic viruses shows that the viruses can be reduced into small group. However, the technique used to cluster the metamorphic viruses is depends on the types of dataset used. The metamorphic virus classification model used Term Frequency Inverse Document Frequency (TF-IDF) to cluster the virus. This technique was widely implement in many research field that had terms or words as their dataset. In addition, this technique gives weight to important terms that need to be highlighted in a document. The proposed model managed to get high True Positive and low False Positive value when classifying the virus based on bi-gram features.

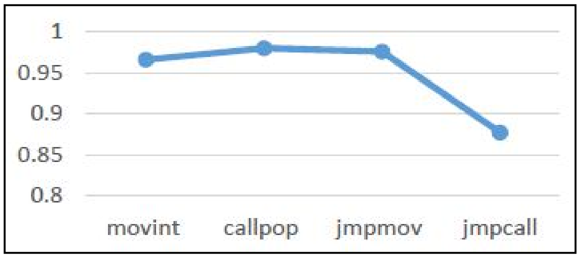


Figure 3. ROC of first bi-gram features

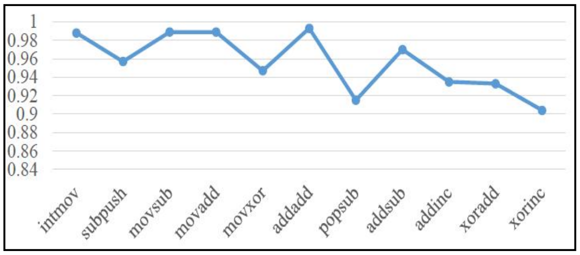


FIGURE 4: ROC OF SECOND BI-GRAM FEATURES

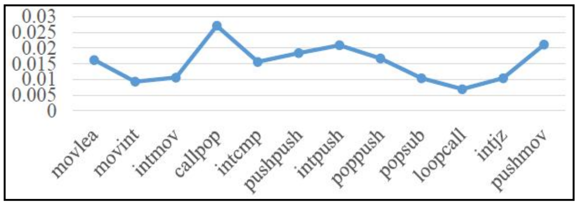


FIGURE 5:. ROC VALUE OF THIRD BI-GRAM FEATURES

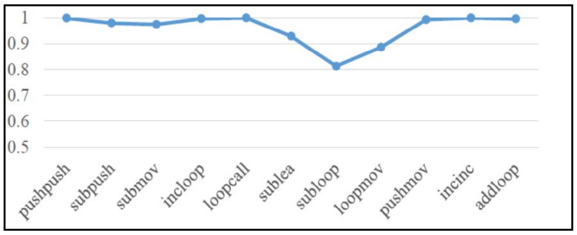


Figure 6: ROC value of fourth bi-gram features

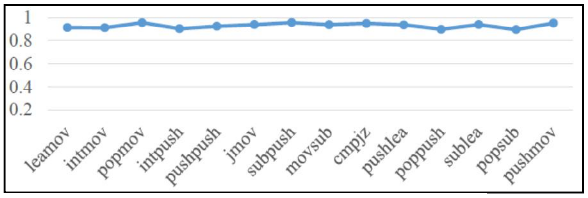


Figure 7: ROC value of fifth bi-gram features

ACKNOWLEDGEMENT

This research is sponsored by Ministry of Higher Education (MOHE), Universiti Tun Hussein Onn Malaysia [Tier 1

Vot: U892]. The authors would like to thank Universiti Tun Hussein Onn Malaysia and Ministry of Higher Education Malaysia for the facilities and financially supporting this research..

REFERENCES

[1] D. Kumar, N. Kumar, and A. Kumar, "Computer Viruses and Challenges for Anti-virus Industry," vol. 3, no. 2, 2014.

[2] S. Venkatachalam and M. Stamp, "Detecting Undetectable Metamorphic Viruses," in Proceedings of the 2011 International Conference 1on Security &amp; Management (SAM 2011), pp. 340, 2011.

[3] A. Venkatesan, "Code Obfuscation and Virus Detection," Master's Proj., pp. 1–47, 2008.

[4] E. Konstantinou, "Metamorphic Virus: Analysis and Detection," no. January, 2008.

[5] A. R. Kakad, S. G. Kamble, S. S. Bhuvad, and V. N. Malavade, "Study and Comparison of Virus Detection Techniques," vol. 4, no. 3, pp. 251–253, 2014.

[6] Y. Qu and K. Hughes, "Detecting metamorphic malware by using behavior-based aggregated signature," Internet Secur. (WorldCIS), 2013 World …, pp. 13–18, 2013.

[7] J. Kuriakose and P. Vinod, "Metamorphic virus detection using feature selection techniques BT - 5th IEEE International Conference on Computer and Communication Technology, ICCCT 2014, September 26, 2014 - September 28, 2014," 2014, pp. 141–146. https://doi.org/10.1109/ICCCT.2014.7001482

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, p. 10, 2009. https://doi.org/10.1145/1656274.1656278

[9] N. Shabani and M. V Jahan, "Metamorphic virus detection based on Bayesian network," in 2014 International Congress on Technology, Communication and Knowledge (ICTCK), pp. 8, 2014. https://doi.org/10.1109/ICTCK.2014.7033515

[10] J. L. M. Jose A. Gamez, Jose M. Puerta, "Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood," Data Min. Knowl. Discov., vol. 22, no. 1–2, pp. 106–148, 2011 https://doi.org/10.1007/s10618-010-0178-6