IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLASSIFICATION OF METAMORPHIC VIRUS USING HMM APPROACH

Ankur Singh Bist*, Dr. Anuj Sharma
* PHD Scholar SVU Gajraula
SVU Gajraula

DOI: 10.5281/zenodo.253984

ABSTRACT

Recent research work shows that HMM (Hidden Markov Model) is widely used in metamorphic virus detection. Virus generated from kits like NGVCK are detected effectively by HMM approach. Our purpose is to examine various flavours of HMM approach in virus detection.

KEYWORDS: Hidden Markov Model, Metamorphism, Observation Sequence.

INTRODUCTION

Internet has become target of malicious codes due to its increasing use. Malicious codes are executable code and have the capability to replicate. It makes their survival strong. Viruses design and evolution attached with the area of programming. Similar to other computer programs viruses carry functions that are intelligent for providing protection in such a manner that detection remains not easy for virus scanner [1].

Viruses have to take various procedures of intellect for continued existence. That is why they may have complex encrypting and decrypting engines. These are the most frequent methods used by computer viruses in current scenario. They make use of these techniques to mask the antivirus and to adopt the certain environment for their expansion [2].

Polymorphic viruses try to hide the decrypting module. More complex methods were developed enabling the virus designers to change the code of one virus file and make multiple morphed copies while maintaining its functionalities. These are the type of viruses which have the ability to mutate itself with the code changed but without changing its functionalities. Metamorphic virus can become a serious threat considering the fact that there can be thousands of variants of one virus file with their signature being totally different.

Metamorphic viruses transform its code in a specific manner very frequently and require to be prohibited. Their analysis will lead to evolve a framework where the overall process of detection will be bounded in specific outcomes of continuing evolving results. It is essential to make a distinction between replicating programs and its similar forms. Reproducing programs will not necessarily damage your system [3] [8]. There is big fight between designers of virus and antivirus. The enhanced knowledge about the certain patterns, specifications can be designed. Various malicious codes can be evolved and incremented in well precise and efficient manner. For perfect identification of a metamorphic virus, identification routines must be written that can generate the essential instruction set of the virus code from the actual occurrence of the infection [9] [10] [11] [12] [13].

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

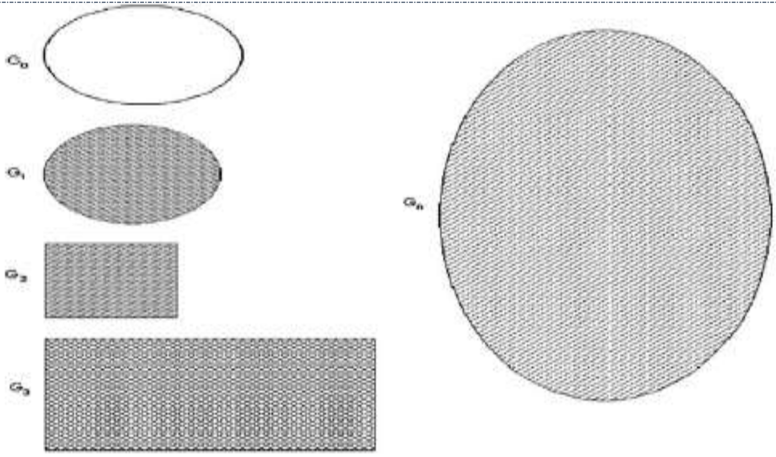

Figure1: Analogy of Metamorphic Viruses

MALWARE CLASSIFICATION APPROACH USING HMM

LET

T= Length of Observation Sequence
N= Number of States in the Model
M= Number of Observation Symbols
Q={ $q_0, q_1, \ldots \ldots \ldots q_{n-1}$ }=Distinct States of Markov Process
V={0, 1,…….M-1}= Set of Possible observations
A= State Transition Probabilities
B= Observation Probability Matrix
Ψ= Initial state Distribution
€=($€_0, €_1, \ldots \ldots \ldots €_{T-1}$)=Observation Sequence (O)

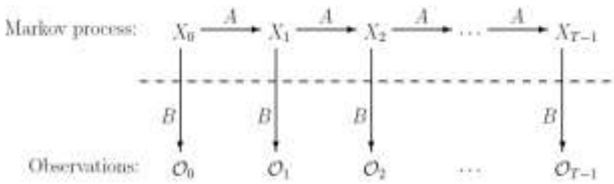

Figure 2: Hidden Markov Model

Hidden markov model are widely used for protein sequence analysis, speech recognition, software piracy detection and malware detection.

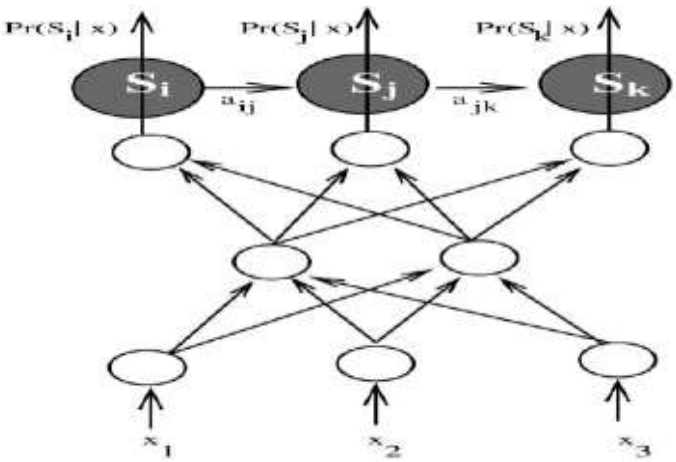

Figure 3: Basic Hybrid Architecture where a Two Layer Feed forward ANN estimates the posterior probabilities of states Si, Sj, Sk, of a left to right HMM given an hypothetic acoustic observation

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

A markov process or model has set of states and fixed probabilities for the state transition. In the hidden markov model the states are not directly visible to the observer. HMM is a machine learning technique that extracts the information during training phase. Score is generated by HMM that can be used for classification.

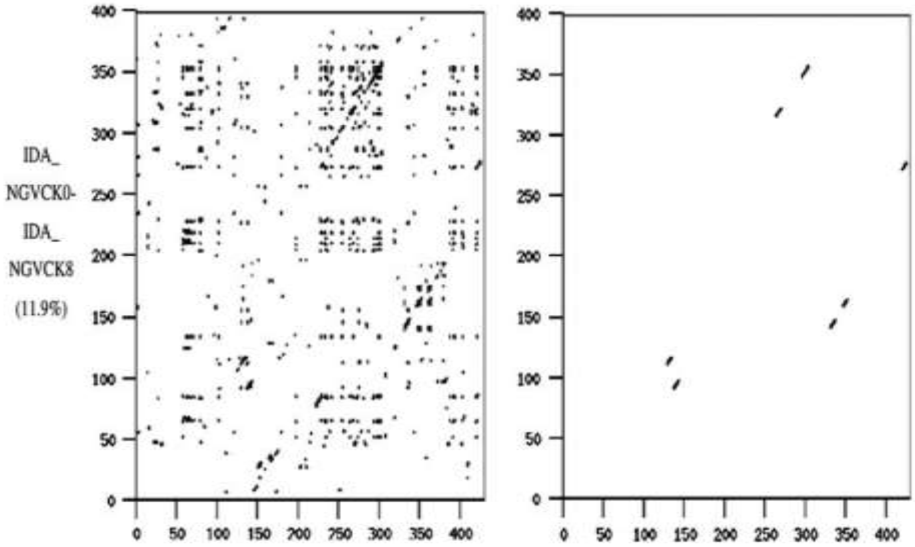

Figure 4: NGVCK Similarity Graph

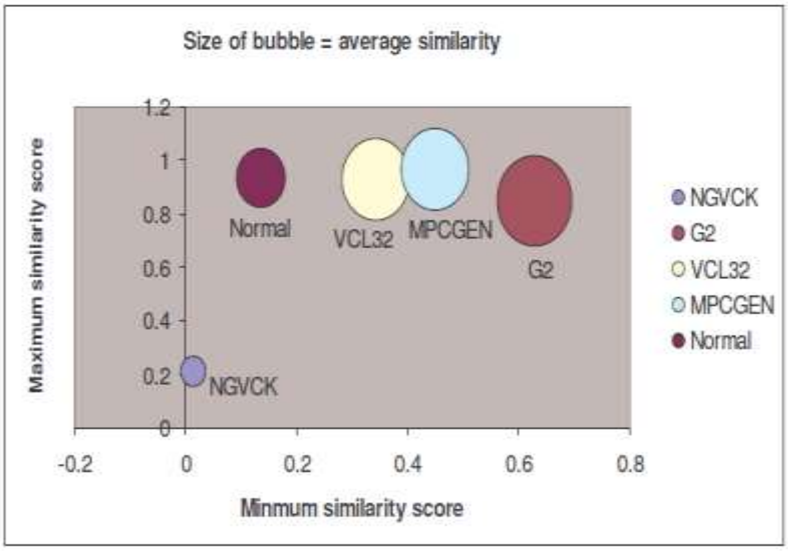

Figure 5: Diagram justifies the impact of NGVCK kit as compare with other kits.

METAMORPHIC VIRUS DETECTION USING HMM

String scanning is the easiest technique used by anti-virus software to identify computer viruses. It searches for sequence of strings that are part of a specific virus. This sequence of bytes is often called the signature of the virus, which is extracted for each different virus and organized in a database. Antivirus Engine will then use this database to search files and system areas for presence of the virus.

Wong and stamp presented detector based on Hidden Markov Models in 2006. They determined that how well the HMM can separate viruses from normal files. NGVCK (New Generation Virus Creation Kit) is used for analysis as the challenging dataset. HMM is found to be very efficient in the domain of malware detection. Similarity scores and threshold specifications found experimentally are given in table 1.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Table 1 Similarity Scores

Comparing IDA_N146 to:					
family viruses	scores	normal files	scores	non-family viruses	scores
IDA_N0	0.0728	IDA_R0	0	IDA_V0	0
IDA_N1	0.1133	IDA_R1	0	IDA_V1	0
IDA_N2	0.0925	IDA_R2	0	IDA_V2	0
IDA_N3	0.0684	IDA_R3	0	IDA_V3	0
IDA_N4	0.0791	IDA_R4	0	IDA_V4	0
IDA_N5	0.1162	IDA_R5	0	IDA_V5	0
IDA_N6	0.0970	IDA_R6	0	IDA_V6	0
IDA_N7	0.1376	IDA_R7	0	IDA_V7	0
IDA_N8	0.0403	IDA_R8	0	IDA_V8	0
IDA_N9	0.1764	IDA_R9	0	IDA_V9	0
IDA_N10	0.1886	IDA_R10	0	IDA_V10	0
IDA_N11	0.1390	IDA_R11	0	IDA_V11	0
IDA_N12	0.1364	IDA_R12	0	IDA_V12	0
IDA_N13	0.1462	IDA_R13	0	IDA_V13	0
IDA_N14	0.1257	IDA_R14	0	IDA_V14	0
IDA_N15	0.1066	IDA_R15	0	IDA_V15	0.0188
IDA_N16	0.1238	IDA_R16	0	IDA_V16	0.0215
IDA_N17	0.1044	IDA_R17	0	IDA_V17	0.0153
IDA_N18	0.0781	IDA_R18	0	IDA_V18	0.0163
IDA_N19	0.1172	IDA_R19	0	IDA_V19	0.0235
IDA_N20	0.1052	IDA_R20	0	IDA_V20	0.0146
IDA_N21	0.1456	IDA_R21	0	IDA_V21	0.0184
IDA_N22	0.1379	IDA_R22	0	IDA_V22	0.0188
IDA_N23	0.0967	IDA_R23	0	IDA_V23	0.0192
IDA_N24	0.0871	IDA_R24	0	IDA_V24	0.0190
IDA_N25	0.1041	IDA_R25	0		
IDA_N26	0.1327	IDA_R26	0		
IDA_N27	0.0597	IDA_R27	0		
IDA_N28	0.1667	IDA_R28	0		
IDA_N29	0.0813	IDA_R29	0		
IDA_N30	0.0383	IDA_R30	0		
IDA_N31	0.1386	IDA_R31	0		
IDA_N32	0.0999	IDA_R32	0		
IDA_N33	0.0661	IDA_R33	0		
IDA_N34	0.1243	IDA_R34	0.0175		
IDA_N35	0.1021	IDA_R35	0		
IDA_N36	0.1010	IDA_R36	0		
IDA_N37	0.0845	IDA_R37	0		
IDA_N38	0.0549	IDA_R38	0		
IDA_N39	0.1292	IDA_R39	0		

Threshold determination: Comparing IDA_N146 to 40 NGVCK viruses	
min score	**0.0349**
max score	0.1894

PHMMs explicitly accounts for positional information. Following Notations are used in PHMM.
X={ $x_1, x_2, \dots \dots \dots x_i$ }is the sequence of emitted symbols/ Observation sequence
N is the total number of states
α is the alphabet for the model/ possible observation symbol
M represents the match states, $M_1, M_2, \dots \dots \dots M_N$
I represent the insert states, $I_1, I_2, \dots \dots \dots I_N$
D represents the insert states, $D_1, D_2, \dots \dots \dots D_N$
π represent initial state probability distribution
A is the state transition probability matrix

A_{kl} is the transition frequency from state k to state l as determined from the given MSA
$a_{M1\ M2}$ is the transition probability from match state M_1 to match state M_2.
E is the emission probability matrix
E_{M1}(k) is the emission frequency of symbol k at state M_1
e_{M1}(k) is the emission probability of symbol k at state M_1
λ=(A, E, π) represents the PHMM model.

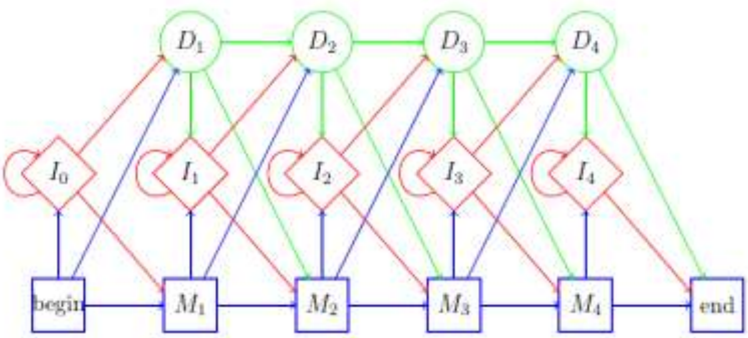

Figure 6: Profile Hidden Markov Model

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Srilatha Attaluri, Scott Mcghee and mark stamp explained about Profile Hidden Markov Models for metamorphic virus detection. Profile hidden markov model explicitly accounts for positional information. This information can be very useful for analysing computer viruses especially metamorphic viruses. It is widely used in bioinformatics especially for finding the related sequences of DNA and proteins. The authors observed that PHMM is well suitable for certain type of metamorphic viruses but not for others. Following are some important results observed experimentally.

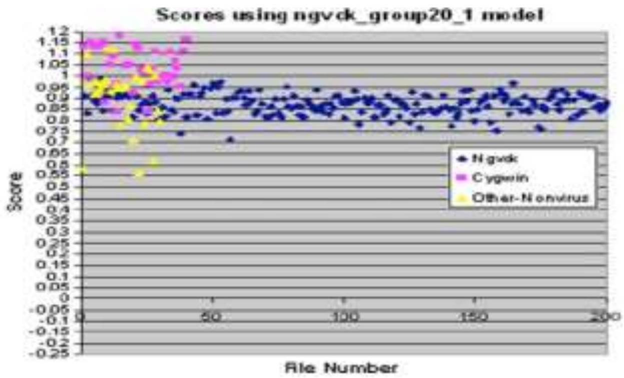

Figure 7: Scores using ngvck_group20_1model

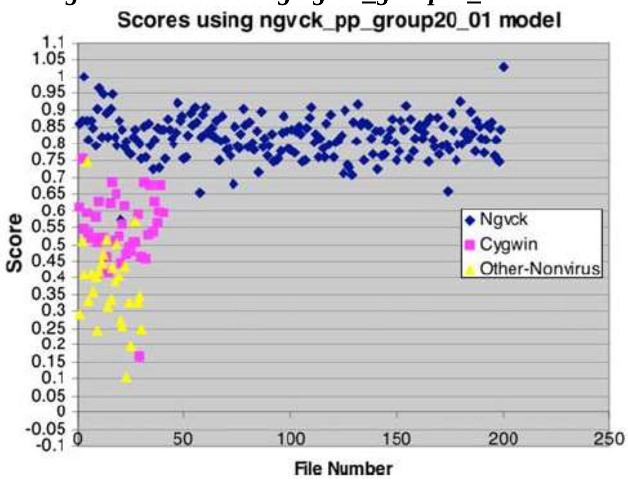

Figure 8: Scores using ngvck_pp_group20_01model

Mangesh Musale explained about hunting for metamorphic Java script malware. A recent trend in attack is observed through web pages where malicious codes inserted in Java Script. Author analysed metamorphic Java Script malware. To detect metamorphic Java Script malware Hidden markov model, opcodes graph similarity, singular value decomposition are used for finding out the similarity between morphed files and random benign files.

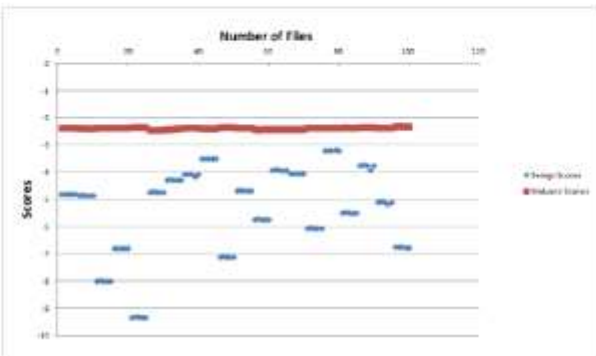

Figure 9: HMM score analysis N=2

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

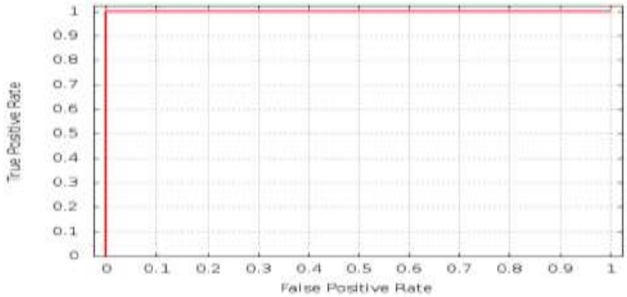

Figure 10: ROC curve for HMM

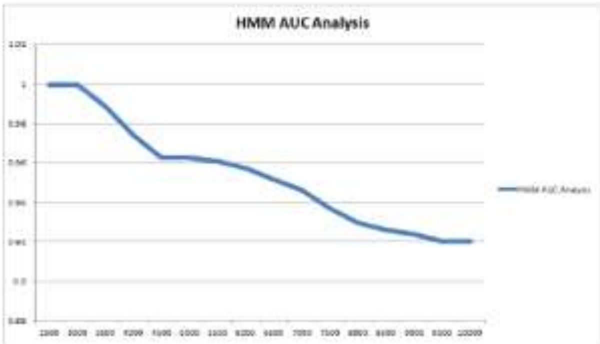

Figure 11: HMM AUC Analysis

Annachhatre, Austin and stamp explored malware classification based on Hidden markov models. More than 8000 malware samples are then scored against these models and clusters are created based on these scores. Some important experimental observations made by authors are depicted in following graphs. Authors obtained quite interesting results and leave remarks for future work like the suggestion to explore variations of k-means algorithm.

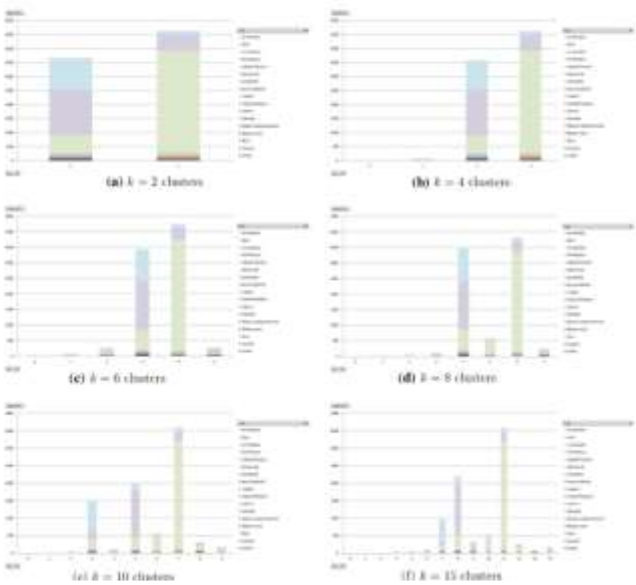

Figure 12: Stacked column chart group by cluster

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

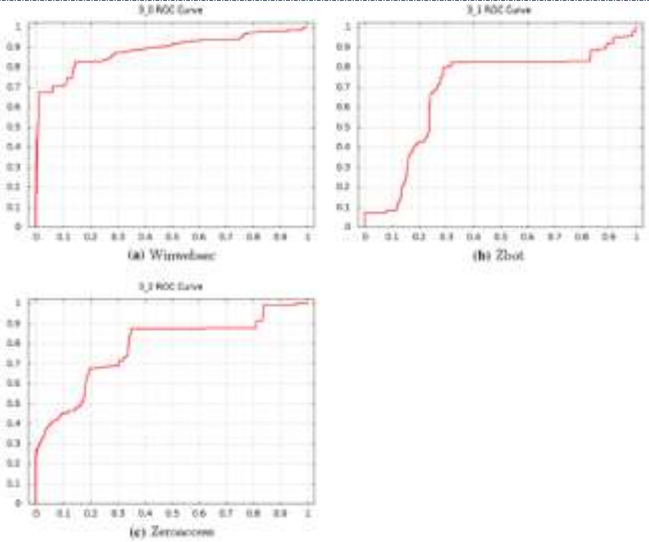

Figure 13: ROC curves for k=3

Ashwin Kalbhor, Thomas H. Austin, Eric Filiol, and Stamp developed the duelling HMM Strategy concept for more accurate classification. Meaning of Hidden states are analysed in order to reveal the in depth issues underlying in it and finally results are tested on four different compilers, hand written assembly code, three virus construction kits and two metamorphic malware families.

Virus family	Threshold				Dueling				Tiered			
	False positives	False negatives	Accuracy	Total time (ms)	False positives	False negatives	Accuracy	Total time (ms)	False positives	False negatives	Accuracy	Total time (ms)
G2	62/370	4/50	84.29	348.22	0/370	0/58	100.00	1,030.57	0/370	0/50	100.00	613.35
MPCGEN	80/370	0/50	78.81	405.47	0/370	0/58	100.00	981.13	0/370	0/50	100.00	685.36
NGVCK	166/370	16/200	68.07	590.07	2/370	75/200	86.49	1,195.64	2/370	75/200	86.49	737.00
MetaPHOR	108/370	0/60	74.88	538.66	4/370	0/60	98.98	1,133.47	4/370	0/60	98.98	840.57
MWOR (PR 1)	0/370	0/100	100.00	866.27	0/370	0/100	100.00	2,029.15	0/370	0/100	100.00	1,312.65
MWOR (PR 2)	1/370	0/100	99.79	964.16	0/370	0/100	100.00	2,415.05	0/370	0/100	100.00	1,968.83
MWOR (PR 3)	4/370	0/100	99.15	1,152.90	0/370	0/100	100.00	2,887.53	0/370	0/100	100.00	2,531.72
MWOR (PR 4)	4/370	0/100	99.15	1,341.37	0/370	0/100	100.00	3,433.05	0/370	0/100	100.00	3,187.04

Figure 14: Comparison of detection methods

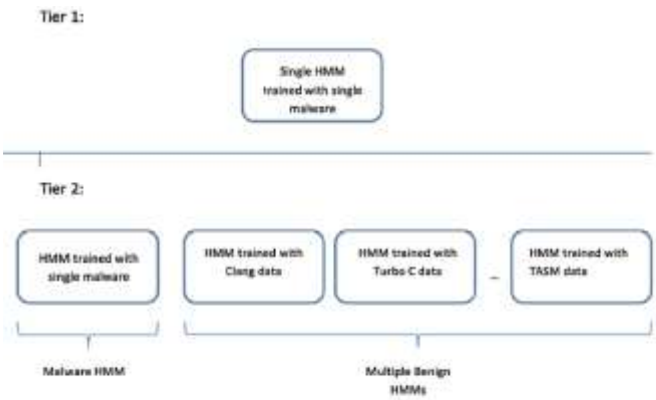

Figure 15: Design of the tiered HMM approach

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

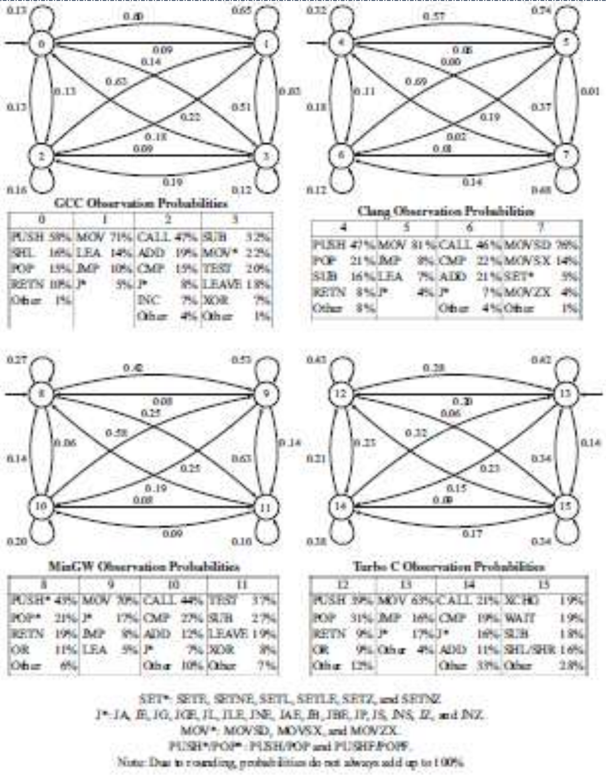

Figure 16: HMMs for GCC, clang, MinGW and Turbo C compilers from disassembled code

Threshold based approach and duelling HMM approach are combined together in tiered fashion in order to improve the performance thus HMM model show promising behaviour towards malware detection especially towards metamorphic malware detection.

CONCLUSIONS

Hidden Markov Model is a powerful statistical tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence. In this paper a detailed study is made to understand the impact of Hidden Markov Models in malware detection especially in metamorphic virus detection. Literature study depicts the various dimensions of HMMs that are being explored by researchers in order to enhance its utility in malware detection.

ACKNOWLEDGMENT

I would like to give my special thanks to all who directly or indirectly supported in my work.

REFERENCES

[1] J. Aycock, Computer Viruses and Malware, Vol 22, New York, NY, USA: Springer, pp. 5-32. 2006.
[2] H. Bidgoli, Handbook of information security, Wiley. 2006.
[3] F. Cohen, Computer Viruses. PhD thesis, University of Southern California. 1986.
[4] M. Mangesh, Hunting for Metamorphic Java script Malware, Master's Project, pp 359, 2014.
[5] M. Stamp, A Revealing Introduction to Hidden Markov Models, San Jose University, 2012.
[6] S. Attaluri, S. McGhee, M. Stamp, Profile Hidden Markov Models and Metamorphic Virus Detection, Springer, pp. 151-170, 2008.
[7] C. Annachhatre, T. Austin, M. Stamp, Hidden Markov Models for malware classification, Journal of Computer Virology, Springer, 2015.
[8] A. Kalbhor, T. Austin, E. Filiol, S. Josse, M. Stamp, Dueling Hidden Markov Models for Virus Analysis, Journal of Computer Virology, Springer, 2015.
[9] Bist, Ankur Singh, and Sunita Jalal. "Identification of metamorphic viruses."Advance Computing Conference (IACC), 2014 IEEE International. IEEE, 2014.

[10] Bist, Ankur Singh. "Detection of metamorphic viruses: A survey." Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE, 2014.
[11] Bist, Ankur Singh. "Classification and identification of Malicious codes." IJCSE. 2012.
[12] Bist, Ankur Singh. "Fuzzy Logic for Computer Virus Detection." IJESRT, ISSN: 2277-9655.
[13] Bist, Ankur Singh. "Hybrid model for Computer Viruses: an Approach towards Ideal Behavior." International Journal of Computer Applications 45 (2012).