International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 1370 – 1374

Use of Metaheuristic Algorithms in Malware Detection

Wasiur Rhmann
Babasaheb Bhimrao Ambedkar University
Lucknow, India
wasiurrhmann786@gmail.com

Gufran Ahmad Ansari
Assistant Professor, Department of Information Technology,
College of Computer, Qassim University, Al-Qassim,
Kingdom of Saudi Arabia (KSA)
gufranansari@qu.edu.sa

***Abstract*—**Metaheuristic algorithms are the general framework for optimization problems. They are not problem dependent and are heavily deployed in different domains. Due to rise in number of malware, malware detection techniques are updated very often. In the present work different metaheuritics algorithm used in malware detection and are available in the literature are discussed. Metaheuristics algorithm like harmony search, clonal selection, genetic algorithm and Negative selection algorithms are discussed.

Keywords*-Metaheuristics, Malware, Signature Based Technique, Polymorphic.*

*****

I. INTRODUCTION

Any code added, removed or changed with the intent to harm system functionality refers to malware. Increase in the number of malware propagation posed the economical loses to persons and organizations. So malware detection is a hot research area among security experts. Earlier signature based algorithms were used to identify threat by matching the suspicious byte code from malware database [1]. But that approach cannot detect unknown malware. New malware variants can be easily created by metamorphic and polymorphic techniques. Signature based techniques cannot detect polymorphic viruses; polymorphic viruses change themselves with each iteration of executions. Malware spread across the network through internet, USB/ DVD, emails and social networking sites.

Due to deficiency is signature based approach behavior based techniques are prevalent for detection of malware. Some techniques are based on static analysis and others are based on dynamic analysis. In the static analysis file is not executed while in dynamic analysis suspected file is executed to gather information about malware. Recently data mining techniques like associative mining, clustering are used in malware detection [2][3].

Metaheuristic algorithms are self learning algorithm to solve hard optimization algorithm up to nearest optimal solution [2]. A number of real life optimization problems cannot be solved by an exact optimization method. Approximate algorithms are designed to solve such types of problems and are classified as heuristics and metaheuristics. Heuristics algorithms are problem dependent and are based on experience while metaheuristics algorithms are like framework for optimization and they guide the design of heuristics. Metaheuristcs often find good solution in less computational cost. They are classified into two categories namely: single solution based or population based algorithms.

Single solution based Metaheuristics
These types of metaheuristics give single solution at a time like Tabu search, guided local search and iterated local search.
Population based Metaheuristics
In these types of algorithms solutions are updated iteratively to get the optimal solution.
Types of Malicious programs
Different types of malicious programs are classified into following categories:
Virus: It recursively replicates itself and infect host file or system.
Worm: They replicate on the networks.
Logic Bomb: Programmer writes malicious code for a legitimate program, which is executed at a specific time after the execution of application.
Torjan Horses: they are simplest malicious program and entice the users to some useful functionality to execute the malicious code.
Spammer Program:
They send unsolicited messages to messaging groups by mails or messages.
Different types of metaheuristics algorithms can be classified into following categories:

1.1 Evolution Based Approach
In evolution based optimization approach candidate solutions are improved using different operators**.**

1.1.1 Genetic Algorithm
1.1.2 Differential Evolution
1.1.3 Genetic Programming

1.2 Swarm Intelligence Based Approach

In swarm intelligence based approach updated candidate solutions are obtained using differential position update rule.

1.2.1 **Firefly Algorithm**

Firefly algorithm is an optimization algorithm which is inspired from the flashing behavior of fireflies.

1.2.2 **Cuckoo Search**

This optimization algorithm is inspired from parasitic behavior of cuckoo birds which places their child to some others nest.

1.2.3 **Bat Algorithm**

Bat algorithm is used for global optimization and inspired from movement of micro bat with varying speed and voices.

1.3 Artificial Immune System

Artificial Immune System is developed for optimization problems and it is based on biological immune system's ability to fight antigens. The immune system protects body from pathogens, diseases.

1.3.1 **Negative Selection Algorithm**

Negative selection algorithm is based on negative selection of T-cells in thymus.

1.3.2 **Clonal Selection Algorithm**

It is developed on the idea of the clonal selection theory of immune system that how B and T lymphocyte cells to improve their affinity.

1.3.3 Artificial Immune Network

It is inspired from immune network theory of immune system.

1.3.4 Danger Theory

Danger model is based on the idea that immune system distinguishes between cell which may be dangerous to system and that may not.

1.4 Particle Swarm Optimization

In this algorithm particles which are candidate solutions of the problem are moved in search space with a velocity.

1.5 Harmony Search

Harmony search algorithm is inspired from musicians who play a good harmony. Each musician plays a decision variable and note produced by them is a value for that variable.

1.6 Ant Colony Optimization

This algorithm can be applied where the optimal path is required in form of a reduced graph.

1.8 Science Based Approach

These metaheuristics are based on different laws related to the natural phenomenon, physics.

1.8.1 Newton's law based Serach

1.8.2 Electromagnetic law based Serach

1.8.3 Thermal energy principle based Serach

1.9 Memetic Algorithm

This algorithm is population based with procedure of local improvement.

2.0Tabu Search and Scatter Search

Tabu search employs local search to give an optimal solution of the mathematical optimization problem.

Metaheuristics algorithms are applied in different field like wireless sensor network [3], reactive power planning [4], Software testing [5, 6]. The rest of the paper is presented as follows: section 2 is described different metaheuristics based malware detection techniques and in section 3 authors concluded the work.

.

II. DIFFERENT METAHEURISTICS USED IN MALWARE DETECTIONS

1. Malware detection with the Harmonic search algorithm

Shen et al. [7] applied Harmony search algorithm for malware detection. Harmonic search algorithm is music inspired algorithm and population based metaheuristics. There are four steps in harmonic search algorithm:

First step is Initialization of harmonic memory, then improvisation of a harmony and inclusion of newly generated harmony, repeat the steps until termination criteria met. Control parameters are defined as HMS(Harmonic Memory Size), HMCR(Harmonic Memory Control Rate), PAR(Pitch Adjusting Rate), NI(Maximum number of iterations).

Random vectors ($x^{1,}$……,x^{hms}) are generated as the harmonic memory size and their fitness value is stored

Harmonic memory is given by the table:

Table 1. Different candidate solutions and there fitness

x_1^1	x_n^1	$f(x^1)$
.	.	.
.	.	.
.	.	.
x_1^{hms}	x_n^{hms}	$f(x^{hms})$

For detection of malware structural feature like byte n-grams, Portable Executable (PE) features, Strings and API call sequences are extracted from executable file. The optimal set of classifiers is derived using Harmony Search (HS). Authors pruned ensemble for malware detection. Ensemble is considered by using multiple heterogeneous classifiers in parallel fashion and pruned set is obtained by selection of best set of classifiers from ensemble. After pruning the ensemble final ensemble are combined. Results of each classifiers are combined with different techniques. The majority vote method is used to combine the result of different classifiers. In some situation more than one classifier is required for final output and in such situation more than one classifier is selected as an ensemble and final output is decided based on majority decision given by selected set of classifiers. A class value is given by each classifiers and ensemble proposed the class with the highest votes. The presented algorithm showed higher detection accuracy and outperformed the existing ensemble algorithm. Malware detection accuracy is calculated using the formula given

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

True Positive (TP)=Correctly predicted samples as malware
False Positive (FP)=Incorrectly predicted samples as malware
True Negative (TN)= Correctly predicted samples as benign
False Negative (FN)=Incorrectly predicted samples as benign

2. Malware detection with the Clonal selection and Genetic Algorithm

Afaneh [8] used artificial immune system (AIS) based algorithm called Virus Detection Clonal Algorithm. Virus detection clonal algorithm is based on the concept of the clonal selection algorithm and main steps of the presented algorithms are: cloning, hyper-mutation and stochastic reselection. In the proposed model viruses are antigens and antibodies are signatures. Signatures with high fitness values are selected for cloning then are mutated to provide a different signature to detect unknown malware. Two parameters signature per clone (fat) and hyper mutation probability (hp) are determined using a genetic algorithm,

VDC is based on validation and have two phases: learning and testing.

In clonal selection, parents with good affinity value are selected for cloning to produce new offspring. These generated offspring goes for maturation of affinity by mutation and offspring with improved fitness value replace the parents.

Generate initial population
While (Given criteria is not satisfied)
{Calculate the **affinity** of each antibody;
Generate clones of candidate solutions;
Hyper mutate each clone;
Reselect candidate solutions from the pool of parents and clones;
Replace candidate solutions of poor affinity;
} **End while**

Fig. 1 Clonal selection algorithm

In Genetic algorithm initial solutions are generated and they are optimized by iterating to het desired level of fitness. The fitness function is problem dependent. Recombination and mutation are used to improve the diversity of candidate solutions and generate new candidate solutions whose fitness are evaluated based on defined fitness function. The sample Genetic algorithm is defined below:

Start
Initialization: Generation of candidate solutions
Evaluation: Calculation of fitness function of each candidate solution
While (termination condition is satisfied)
{**Selection**: select the individual for next generation;
Recombination: Combine the individuals to generate new candidates;
Mutation: New generated candidates are change randomly;
Evaluation: Calculate the fitness of newly generated candidates;
}
Stop
Fig. 2 Genetic Algorithm

3. Malware detection using Negative selection algorithm with penalty factor

Negative selection algorithm is an Artificial Immune System (AIS) algorithm. It is inspired from disease protection system of human body. The immune system protects human body from pathogens and antigens. These pathogens and antigens are foreign substance like viruses, fungi, bacteria. Similarly generated antibodies are destroyed in defecnce of body. There are two types of cells in white blood cells:

B Lymphocytes (B cells) and T lymphocytes (T cells). Both types of cells produce in bone marrow and can be cloned, mature and may encounter antigens. B cells mature in spleen while T cells mature in thymus. The process of negative selection of imune system is performed in thymus and T cells that mathes with itself are excluded. Malware instruction library (MIL) is extracted from deep analysis of instruction frequency and file frequency [9]. A malware candidate signature library (MCSL) and a benign program malware like signature library (BPMSL) is created by splitting programs into various short bit strings. Negative selection algorithm divides the MCSL library into MDSL1 and MDSL2 based on signature matches to self. Model classifies the program as malicious and benign based on MDSL1 and MDSL2. Negative selection algorithm has two steps which are given below:

1. **Generation stage(Training stage)**
1. **Start**
2. Random candidate are generated
3. **If** generated candidate match self samples
4. **Then** new candidate is generated
5. **Else** generated candidate will be in new detector set.
6. **If** Enough detectors are generated
7. **Then** Stop
8. **Else**
9. **Generate** new detector

A number of detectors are generated in Negative selection algorithm. If new generated data matched any available detector then it will be discarded else it will be a member of detector set.

2. **Detection stage(Testing stage)**
1. **Start**
2. **New input data is generated**
3. **If** generated input matches any detector
4. **Then** Input is categorized **as non-self**

5. Else self

Detection stage is used to classify the input as self and non self. If input data matches input then input will be as non self.

4. Malware detection based on Metric based and Genetic algorithm

Malware detection algorithm based on Metric method and Genetic algorithm is presented [10]. Characteristics of programs are represented as vectors by metric method. The hybrid genetic algorithm is used to detect malicious part of the program. Malware scripts are detected by counting token frequency. There are four parts of the presented model: Decision algorithm determines the nature of program, whether a program is malicious or not. Malicious part of code is extracted with GA by Malicious code finder. Metric calculator converts the program to numerical vectors containing various metric. Distance calculator measures the distance between vectors.

GA removed the junk part of the code of the program.

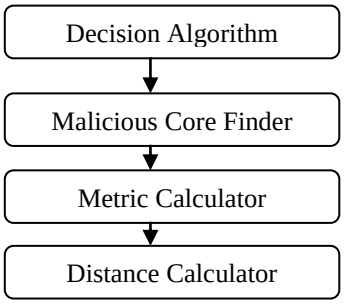

Fig. 3 Malware detection based on Genetic Algorithm and Metric Based Method

5. MALWARE DETCTION USING GENETIC PROGRAMMING

Le et al. [11] have presented a model for malware detection based on genetic programming. Proposed model is based on different features of file which are extracted as feature extraction.

5.1 Feature Extraction

Different features are extracted from file to distinguish between malicious codes legitimate files. Method based on n-grams are used to extract features. N-contiguous items, extracted in a sequence from a text are called a n-gram. After conversion of test file into hex format different factors are calculated;

$$TextFrequency = \frac{TermFrequency}{Max(TermFrequencyinmalcode)}$$

$$TFIDF = TF * log\frac{N}{DF}$$

N is No. Of documents in entire file collection

DF is No. of files in which term appears

5.2 Genetic Programming based model

In genetic program candidate solutions are a set of programs which are evolved to produce better programs using evolutionary algorithm.

There are two stages in solving malware detection problems namely: training and testing.

In training stage model is evolving to be able to classify the files as malicious or legitimate based on its features and in testing phase unseen data is predicted as malicious or benign based on the model of training stage.

III. CONCLUSIONS

This paper presented different metaheuristics techniques available in the literature which are used for malware detection. As polymorphic and metamorphic malware cannot be detected by signature based malware detection techniques. Signature based techniques are based on malware database and cannot detect unknown malware. Metaheuristics based techniques are effective against unknown malware as they require not to match from malware database. The main aim of these metaheuristics techniques is to reduce the false positive and false negative numbers to improve the efficiency of malware detection techniques.

REFERENCES

[1] Peter Szor, The Art of Computer Virus Research and Defense, Addison Wesley, 2005.

[2] Yanfang Ye, Dingding Wang, Tao Li, Dongyi Ye, Qingshan Jiang, An Intelligent PE malware detection system based on association mining, Journal of Computer Virology and Hacking Techniques, 2008, Vol. 4, No. 4, pp. 323-334.

[3] Younghee Park, Douglas S. Reeves, Mark Stamp, Deriving common malware behavior through graph clustering, Computer and Security, 2013, Vo. 39, pp. 419-430.

[4] Ke-Lin Du and M. N. S. Swamy, Search and Optimization by Metaheuristics, Springer, Switzerland, 2016.

[5] Palvinder Singh Mann, Satvir Singh, Energy Efficient Clustering Protocal Based on improved Metagheuristic in wireless sensor network, Journal of Network and Computer Applications, Vol. 83, 2017, pp. 40-52.

[6] Abdulla M. Saheen, Shimaa R. Spea, Sobhy M. Farrang, Mohammed A. Abido, A review of meta-heuristic algorithms for reactive power planning problem, Ain Shams Engineering Journal, In press, 2015.

[7] Wasiur Rhamnn, Taskeen Zaidi, Vipin Saxena, Use of Genetic Approach for test Cases Prioritization from UML Activity Diagram, International Journal of Computer Application, Vol. 115, No. 4, 2015, pp. 8-12.

[8] Wasiur Rhmann, Vipin Saxena, Optimized and Prioritized Test Paths Generation from UML Activity Diagram using Firefly Algorithm, International Journal of Computer Application, Vol. 145, No. 6, 2016.

[9] Shina Sheen, R. Anitha, P. Sirisha, Malware detection by pruning ensemble using harmony search, Pattern Recognition Letters, Article in Press, 2013, pp. 1-8.

[10] Suha Afaneh, Raed Abu Zitar, Alaa Al-Hamami, Virus detection using clonal selection algorithm and Genetic Algorithm, Applied Softcomputing, Vol. 13, No. 1, 2013.

[11] Zgang Peng Tao, Wang Wei and Tan Ying, A malware detection model based on a negative selection algorihm with penalty factor, Science China Information Science, Vol. 53, No. 12, 2010, pp. 2461-2471.

[12] Jingyun Kim and Byung-Ro Moon, New Malware Detection System using Metric Based Method and Hybrid Genetic Algorithm, GECCO'12 Companion, USA, July 7-11, 2012, pp. 1527-1528.

[13] Thi Anh Le, Le Quy Quang Uy Nguyen,Xuan Hoai Nguyen, Malware detection using Genetic programming, IEEE, 2014.