



Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University

Intelligent Technique for Computer Virus Detection

THESIS

Submitted in partial fulfillment of the requirements for the degree
of Master of Science in Computer and Information sciences

To the department of Computer Science, Faculty of Computer and
Information Sciences, Ain Shams University.

BY

MOHAMED MABROUK MAWED MORSEY
B.Sc., Faculty of Computer and Information Sciences,
Ain Shams University.

SUPERVISED BY

Prof. Dr. **MOSTAFA MAHMOUD SYIAM**
Former Vice Dean for Student Affairs, Faculty of Computer and
Information Sciences, Ain Shams University.

Prof. Dr. **MOHAMED HASHEM**
Chairman of the Technical Research and Development Center,
Air Defense Forces.

Dr. **SHYMAA ARAFAT**
Lecturer in Computer Science Department, Faculty of Computer
and Information Sciences, Ain Shams University.

Cairo – 2006

Acknowledgments

Thanks are due to ALLAH for getting this work done.

I am very grateful for the encouragement of Professor M. Essam Khalifa our faculty dean. Many thanks are given to my Advisors, Professor Mostafa Syiam, Doctor Mohamed Hashem, and Doctor Syhmaa Arafat for their continuous support, and help in all fields of my life and research work. Many work of thesis weren't being done without their valuable comments that helped me to cut many huge technical problems. I want to specially thank Doctor Mohamed Hashem, and Doctor Syhmaa Arafat for their constant support, interest, and encouragement of my work. I learned too much from them and I hope to continue the co-operation with them in following research. In this acknowledgment I should highlight the continuous encouragement and support of Professor Said Ghoniemy. Many thanks are given to everyone who helped me to achieve this work.

List of Figures

2.1 Functional diagram of the virus.....	9
2.2 The process of loading a .COM program into memory...	12
2.3 The structure of a .EXE program.....	14
2.4 The process of loading a .EXE program into memory	14
2.5 The boot record is located in track 0, head 0, and sector 1	15
2.6 The virus places itself in place of boot record, and saves the original boot record to the last record.....	17
2.7 A flowchart indicating the infection process of a boot record virus.....	18
2.8a An overwriting virus, it overwrites the code of the original program rendering it useless.....	20
2.8b A prepending virus, it places a copy of itself at the beginning of the program file.....	21
2.8c An appending virus, it places a copy of itself at the end of the file and inserts a jump instruction at the beginning of the file.....	21
2.9 The structure of a .EXE file after infection by an appending virus.....	22
2.10 A flowchart indicating how the stealth read virus works	25
3.1 Typical model of a neuron.....	32
3.2 The effect produced by the presence of a bias.....	33
3.3 The graph of the threshold function.....	34

3.4 The graph of the piecewise-linear function.....	34
3.5 The graph of the sigmoid function for a varying slope parameter a	35
3.6 A multilayer perceptron network with one hidden layer..	36
3.7 A signal flow graph indicating the details of output neuron j.....	38
3.8 A signal flow graph indicating the details of output neuron k connected to hidden neuron j.....	41
3.9 A signal flow graph indicating the back-propagation of error signal.....	43
3.10 A set of data items that are grouped into several disjoint groups.....	55
3.11 The agent interacts with environment through sensors and actuators.....	58
3.12 A static agent sends its mobile agents to their new host servers.....	59
3.13 The agent is sent to data source and only important data is sent through the network.....	61
3.14 The stages of the lifecycle of a mobile agent.....	61
4.1 An illustration of a typical Turing machine.....	68
5.1 An imaginary example of a viral input vector.....	82
5.2 An imaginary example of a non-viral input vector....	82
5.3 An imaginary example of a 2-dimensional array of feature values for viral and non-viral files.....	83

5.4 The architecture of the neural network classifier.....	85
5.5 A flowchart describing the entire process of the neural network detector.....	87
5.6 The effect of changing the number of training samples on the false positive and false negative errors...	88
5.7 The effect of changing the number of training samples on the success ratio.....	89
5.8 The effect of the threshold value on the number of errors.....	90
5.9 The effect of the threshold value on the success ratio.....	90
5.10 The effect of changing the number of training samples on the false positive and false negative errors...	91
5.11 The effect of changing the number of training samples on the success ratio.....	92
5.12 The effect of changing the threshold value on the number of errors (program file viruses).....	93
5.13 The effect of changing the threshold value on the success ratio (program file viruses).....	93
5.14 Changing the number of training samples affects the number of errors.....	96
5.15 Changing the number of training samples affects the success ratio.....	97
5.16 Changing the number of training samples affects the	98

number of errors.....	
5.17 Changing the number of training samples affects the success ratio.....	98
5.18 Changing the number of training samples affects the number of errors.....	101
5.19 Changing the number of training samples affects the success ratio.....	102
5.20 Changing the number of training samples affects the number of errors.....	103
5.21 Changing the number of training samples affects the success ratio.....	103
5.22 Changing the number of training samples affects the number of errors.....	106
5.23 Changing the number of training samples affects the success ratio.....	106
5.24 Changing the number of training samples affects the number of errors.....	107
5.25 Changing the number of training samples affects the success ratio.....	108
5.26 The structure of the virus detection system.....	109
5.27A flowchart describing the process of training, moving and testing the agents.....	112
5.28 The effect of changing the number of training samples on the number of errors.....	113

5.29 The effect of changing the number of training samples on the success ratio.....	114
5.30 Indicates the effect of changing the number of training samples on the number of errors for the method..	115
5.31 Indicates the effect of changing the number of training samples on the success ratio of the method.....	115

List of Tables

2.1 A simple comparison between viruses, Trojan horses, and worms.....	8
2.2 A simple comparison between .COM, and .EXE files..	11
2.3 A simple comparison between the antivirus techniques.....	30
3.1 Fictional data describing the weather conditions for playing some game.....	50
3.2 Frequencies of each weather feature value.....	50
3.3 Probabilities of each weather feature value.....	50
3.4 A list of some of the available agent platforms.....	64
5.1 The effect of changing the number of training on the number of false positives and false negatives.....	88
5.2 The effect of changing the threshold value T on the NN classifier when the number of training samples=15....	89
5.3 The effect of changing the number of training on the number of false positives and false negatives.....	91
5.4 The effect of changing the threshold value T on the NN classifier when applied on program viruses (the number of training samples is 50).....	92
5.5 The effect of changing the number of training samples on the performance of the naïve Bayes method when applied on boot record viruses.....	96

5.6 The effect of changing the number of training samples on the performance of the naïve Bayes method when applied on program viruses.....	97
5.7 The effect of changing the number of training samples on the performance of the multi-naïve Bayes method when applied on boot record viruses.....	101
5.8 The effect of changing the number of training samples on the performance of the multi-naïve Bayes method when applied on program viruses.....	102
5.9 The effect of changing the number of training samples on the performance of the K-means method when applied on boot record viruses.....	106
5.10 The effect of changing the number of training samples on the performance of the K-means method when applied on program viruses.....	107
5.11 The effect of changing the number of training samples on the performance of the mobile agent method when applied on boot record viruses.....	113
5.12 The effect of changing the number of training samples on the number of false positives and false negatives.....	114
6.1 A simple comparison among the intelligent antivirus techniques.....	119

Contents

LIST OF FIGURES.....	xiii
LIST OF TABLE.....	xviii
1 INTRODUCTION.....	1
1.1 Motivation.....	1
1.2 Research Goals.....	3
1.3 Thesis Outlines.....	3
2 AN OVERVIEW ON COMPUTER VIRUSES.....	5
2.1 What is a Computer Virus?.....	5
2.2 Viruses, Trojan Horses, and Worms.....	6
2.2.1 Trojan Horses.....	6
2.2.2 Worms.....	7
2.3 Computer Virus Components.....	8
2.4 Types of Computer Viruses.....	10
2.4.1 Classification of Viruses According to The Type of Host.....	10
2.4.1.1 Executable File Viruses.....	11
2.4.1.2 Boot Record Viruses.....	15
2.4.1.3 Macro Viruses.....	19
2.4.2 Classification of Viruses According to The Structure	20

2.4.3 Classification of Viruses According to The Method of Infection.....	22
2.5 Antivirus Techniques.....	25
2.5.1 Overly Broad Detection.....	26
2.5.1.1 Activity Monitors.....	26
2.5.1.2 Integrity Management Systems.....	27
2.5.2 Overly Specific Detection.....	27
2.5.2.1 Virus Scanners.....	28
2.5.2.2 The Determination of The Signature of a Virus	29
 3 AN OVERVIEW ON THE APPLIED INTELLIGENT TECHNIQUES	 31
3.1 An Overview On Neural Networks.....	31
3.1.1 The Model of a Neuron.....	31
3.1.2 Types of Activation Function.....	33
3.1.3 Multilayer Perceptron Neural Network	35
3.1.4 The Back-Propagation Training Algorithm.....	36
3.1.5 The Applications of Neural Networks.....	44
3.2 An Overview On Data Mining.....	45
3.2.1 Major Tasks of Data Mining.....	45
3.2.2 Applications of Data Mining.....	47
3.2.3 Naïve Bayes Method.....	49
3.2.4 Multi-Naïve Bayes Method.....	52
3.2.5 K-means Clustering Method.....	55

3.3 An Overview On Mobile Agents.....	58
3.3.1 Mobile Agents.....	59
3.3.2 Why Mobile Agents?.....	60
3.3.3 The Agent Lifecycle.....	61
3.3.4 Mobility.....	62
3.3.5 Judging An Agent.....	62
3.3.6 The Agent Platforms.....	63
3.3.6.1 Aglets.....	64
3.3.6.2 Grasshopper.....	64
3.3.7 Applications of Mobile Agents.....	65
 4 MODELING COMPUTER VIRUSES.....	 68
4.1 Overview on Turing Machines.....	68
4.2 The Proposed Mathematical Model.....	69
4.2.1 The Notation of The Model.....	69
4.2.2 Trojan Horses.....	71
4.2.3 The Virus Method.....	72
4.3 Applying The Framework.....	73
4.3.1 Viral Resistance.....	73
4.3.2 Applying the framework on each type of virus..	74
4.4 The Proposed Validity of The Model.....	75
4.5 Detectability of Viruses.....	77
4.5.1 Cohen's Proof of Undecidability of Virus Detectability...	77
4.5.2 Applying Cohen's Proof on The New Framework	79

4.6 Conclusions.....	79
 5 APPLYING THE INTELLIGENT COMPUTER VIRUS DETECTION TECHNIQUES.....	 81
5.1 Feature Extraction.....	81
5.2 The Proposed Neural Network-Based Virus Detector.	83
5.2.1 Building and Training The Network.....	84
5.2.2 Results Of Applying The Neural Network- Based Method.....	87
5.2.2.1 Applying The Neural Network method on Boot Record Viruses.....	87
5.2.2.1.1 The Effect of Changing The Number of Training Samples.....	88
5.2.2.1.2 The Effect of Changing The Output Threshold	89
5.2.2.2 Applying The Neural Network method on Program File Viruses.....	90
5.2.2.2.1 The Effect of Changing The Number of Training Samples.....	91
5.2.2.2.2 The Effect of Changing The Output Threshold..	92
5.3 The Proposed Data Mining-Based Virus Detector...	94
5.3.1 Naïve Bayes Method.....	94
5.3.1.1 The Algorithm of Naïve Bayes Method.....	94
5.3.1.2 The Results of Applying The Naïve Bayes Method	95
5.3.1.2.1 Applying The Naïve Bayes Method on Boot Record Viruses.....	96

5.3.1.2.2 Applying The Naïve Bayes Method on Program Viruses.....	97
5.3.2 Multi-Naïve Bayes Method.....	98
5.3.2.1 The Algorithm of The Method.....	99
5.3.2.2 The Results of Applying The Multi-Naïve Bayes Method.....	100
5.3.2.2.1 Applying The Multi-Naïve Bayes Method on Boot Record Viruses.....	101
5.3.2.2.2 Applying The Multi-Naïve Bayes Method on Program Viruses.....	102
5.3.3 K-means Clustering Method.....	103
5.3.3.1 The Algorithm of The Method.....	104
5.3.3.2 The Results of Applying The K-Means Method	105
5.3.3.2.1 Applying The K-Means Method on Boot Record Viruses.....	105
5.3.3.2.2 Applying The K-Means Method on Program Viruses.....	107
5.4 The Proposed Mobile Agent-Based Viruses Detector.	108
5.4.1 The Structure of Mobile Agent-Based Virus Detector	108
5.4.2 Results of Applying Mobile Agents.....	112
5.4.2.1 Applying The Mobile-Agent Method on Boot Record Viruses.....	112
5.4.2.2 Applying The Mobile-Agent Method on Program File Viruses.....	114

5.5 The Conclusions.....	116
6 CONCLUSIONS AND FUTURE WORK.....	118
REFERENCES.....	121