ANALYSIS OF MACHINE LEARNING TECHNIQUES USED IN BEHAVIOR-BASED MALWARE DETECTION

By

Ivan Firdausi

A Bachelor's Thesis
Submitted to the Faculty of

INFORMATION TECHNOLOGY

in partial fulfillment of the
requirements for the Degree of

BACHELOR OF SCIENCES
WITH A MAJOR IN INFORMATION TECHNOLOGY

SWISS GERMAN UNIVERSITY
EduTown BSD City
Tangerang – 15339
Island of Java, Indonesia
www.sgu.ac.id

July 2010

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, not material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Ivan Firdausi Date

Approved by:

Charles Lim, Msc., ECSA, ECSP, ECIH, CEH, CEI Date

Anto Satriyo Nugroho, Dr.Eng Date

Chairman of the Examination Steering Committee Date

ABSTRACT

ANALYSIS OF MACHINE LEARNING TECHNIQUES USED IN
BEHAVIOR-BASED MALWARE DETECTION

By

Ivan Firdausi

SWISS GERMAN UNIVERSITY
Bumi Serpong Damai

Charles Lim, Msc., ECSA, ECSP, ECIH, CEH, CEI, Major Lecturer

The increasing of malware that are exploiting the Internet daily has become a serious threat. The manual heuristic inspection of malware analysis is no longer considered effective and efficient compared against the high spreading rate of malware. Conventional signature matching-based antivirus systems fail to detect polymorphic, obfuscated, and new, previously unseen malicious executables. Hence, automated behavior-based malware detection using machine learning techniques is considered a profound solution. The behavior of each malware on an emulated (sandbox) environment will be automatically analyzed and will generate behavior reports. These reports will be preprocessed into sparse vector models for further machine learning (classification). The classifiers used in this research are *k*-Nearest Neighbors (*k*NN), Naïve Bayes, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN). According to the analysis of the test and experiment results of all the 5 classifiers, the overall best performance goes to J48 with a recall (true positive rate) of 95.9%, a false positive rate of 2.4%, a precision (positive predictive value) of 97.3%, and an accuracy of 96.8%. In summary, it can be concluded that a proof-of-concept based on automatic behavior-based malware analysis and the use of machine learning techniques could detect malware quite effectively and efficiently.

Keywords—malware analysis, dynamic analysis, behavior analysis, data mining, machine learning, classification, malware detection

DEDICATION

I dedicate this thesis to my family, my relatives, and my friends.

ACKNOWLEDGMENTS

The author wishes to give the utmost gratitude to Allah SWT. His blessing and guidance had helped me to give me strength, great health, great ideas, and finally led me to complete this thesis.

The author would also like to give special thanks to Charles Lim, Msc., Anto Satriyo Nugroho, Dr.Eng, and Alva Erwin, Msc., MTI for acting as my thesis advisors. They have had given me support, suggestions, and guidance in completing this thesis.

In addition, the author would also like to give thanks to Attur S. Widjaja (Aat) for giving us (SGU Malware Analysis Team) Indonesian malware samples for our data sets and to Anubis for giving us the permission to use their XML behavior report files of the malware and benign software that we had submitted for our research.

Last but not least, the author would like to give thanks to James Purnama, Msc. and Dipl.-Inf. Kho I Eng for supporting and helping us (SGU Malware Analysis Team) in the establishment of our new malware laboratory in SGU and to IDSIRTII Malware Analysis Team for sharing ideas, discussions, and their cooperation with us.

TABLE OF CONTENTS