SUPERVISED MACHINE LEARNING APPROACH FOR DETECTION OF MALICIOUS EXECUTABLES

YAHYE ABUKAR AHMED

A project submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JANUARY 2013

This project is dedicated to my beloved brother for his endless support and encouragement, and to my parents.

ACKNOWLEDGEMENT

First and foremost, I would like to express heartfelt gratitude and my sincere appreciation to my supervisor **Professor Dr. Mohd Aizaini Maarof** and my co-supervisor **Dr. Anazida Zainal** for their constant support, encouragement, guidance and friendship. They inspired me greatly to work in this project. Their willingness to motivate me contributed tremendously to our project. I have learned a lot from them and I am fortunate to have their as my mentor and supervisor

Besides, I would like to thank my beloved brother who supported advices and financially, and also the authority of Universiti Teknologi Malaysia (UTM) for providing me with a good environment and facilities such as Computer laboratory to complete this project and software which I need during process.

ABSTRACT

Malware can be described as any type of malicious code that has the potential harm to the computer or network. these threats came from various sources like the internet, local networks and portable drives. Virus which replicates itself is growing faster every year and poses a serious global security threat. The purpose of this research is to classify portable executable new malicious files from benign files. In recent years, data mining methods are investigated for detecting unknown malicious executables, and the result show high and acceptable detection rate. Therefore, this project applied machine learning to detect malicious executable files through Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms. These algorithms are compared together and selected the best accuracy model. The result of this research indicated that the accuracy of the SVM and ANN rely on the settings of the parameters used; ANN showed higher accuracy of 98.76 than SVM in terms of data set used while SVM performed a speed three times less than ANN and low computational power. The main conclusions drawn from this research were that current detection approaches of the antivirus are deficient because they fail to detect new unseen malicious files and they have higher false negative rates.

ABSTRAK

Malware boleh didefinisikan sebagai sebarang jenis kod pelbagai yang mempunyai potensi sebagai ancaman kepada komputer atau rangkaian dimana ancaman ini berpunca daripada pelbagai sumber seperti internet, rangkain setempat atau peranti luaran. Virus yang menyerupai diri sendiri berkembang dengan pesat setiap hari dan menyebabkan ancaman keselamatan di peringkat global. Tujuan kajian ini adalah untuk mengklasifikan pelaksanaan mudah alih pelbagai fail baru . Beberapa tahun kebelakangan ini, kaedah data mining telah disiasat untuk mengesan perlaksanaan kod pelbagai yang tidak diketahui puncanya dan keputusan menunjukkan kadar pengesanan yang tinggi dan diterimapakai. Dengan itu projek ini akan menggunapakai pembelajaran mesin untuk mengesan pelaksanaan pelbagai fail melalui algoritma Sokongan Mesin Vektor (SVM) dan Rangkaian Neural Kebijaksanaan (ANN) . Algoritma ini akan dibandingkan dan model yang paling tepat akan dipilih, Keputusan kajian akan menunjukkan ketepatan SVM dan ANN bergantung ke atas konfigurasi parameter yang digunakan dan kajian menunjukkan ANN mempunyai tahap ketepatan 98.76 berbanding SVM daripada segi konfigurasi parameter yang digunakan dengan kelajuan prestasi SVM adalah kurang tiga kali berbanding ANN dan kuasa pengkomputeran juga adalah rendah. Kesimpulan yang dapat dibuat daripada kajian ini adalah pendekatan pengesanan antivirus sedia ada masih banyak kekurangan kerana gagak untuk mengesan fail mengandungi kod pelbagai yang tidak dapat dilihat dan mereka mempunyai kadar negatif penipuan yang lebih tinggi.