SPAM DETECTION USING HYBRID OF ARTIFICIAL NEURAL NETWORK AND GENETIC ALGORITHM

ANAS W.A. ARRAM

UNIVERSITI TEKNOLOGI MALAYSIA

SPAM DETECTION USING HYBRID OF ARTIFICIAL NEURAL NETWORK AND GENETIG ALGORITM

ANAS W.A. ARRAM

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

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2013

To my beloved parents Wasif Arram and Tamam Madi, who have much faith in me.

To all my brothers and sisters who have stood by me.

To my respected supervisor, Dr.Anazida Zainal.

To my beloved country, Palestine.

ACKNOWLEDGEMENT

First and foremost, I give thanks and praise to Allah for his direction and blessings and for granting me knowledge, fortitude, and determination in the successful achievement of this research work and project.

I would like to express my gratitude to my supervisor, Doctor Anazida Zainal for hir guidance, trust, and support. I thank her for her insightful conversations and comments on the work.

I would like to thank my guide through life, my mother, who nourished the love of science in me, and who showed patience in raising me to become who I am today. She always acted as an encouraging educational model in my life. I thank her for her continued support and prayers for me. A tremendous amount of thanks goes to my father; I will always remember his encouragement and support to me since I began my postgraduate work. I will not forget his unlimited help through many difficulties as I pursued my degree.

A word of gratitude is also extended to my brothers and sisters for their support, encouragement, and patience.

ABSTACT

Spam detection is a significant problem which is considered by many researchers by various developed strategies. In this study, the popular performance measure is a classification accuracy which deals with false positive, false negative and accuracy. These metrics were evaluated under applying two supervised learning algorithms: hybrid of Artificial Neural Network (ANN) and Genetic Algorithm (GA), Support Vector Machine (SVM) based on classification of Email spam contents were evaluated and compared. In this study, a hybrid machine learning approach inspired by Artificial Neural Network (ANN) and Genetic Algorithm (GA) for effectively detect the spams. Comparisons have been done between classical ANN and Improved ANN-GA and between ANN-GA and SVM to show which algorithm has the best performance in spam detection. These algorithms were trained and tested on a 3 set of 4061 E-mail in which 1813 were spam and 2788 were non-spam. Results showed that the proposed ANN-GA technique gave better result compare to classical ANN and SVM techniques. The results from proposed ANN-GA gave 93.71% accuracy, while classical ANN gave 92.08% accuracy and SVM technique gave the worst accuracy which was 79.82. The experimental result suggest that the effectiveness of proposed ANN-GA model is promising and this study provided a new method to efficiently train ANN for spam detection.

ABSTRAK

Pengesanan spam adalah masalah yang besar dimana ia dianggap oleh ramai penyelidik dengan pelbagai strategi-strategi yang telah dibangunkan. Dalam kajian ini, pengukuran prestasi yang selalu digunakan adalah ketepatan pengelasan yang berurusan dengan positif palsu, negatif palsu dan ketepatan. Metrik-metrik ini telah dinilai dengan menggunakan dua algoritma pembelajaran yang dikawal: Hybrid of Artificial Neural Network (ANN) dan Genetic Algorithm (GA), Support Vector Machine (SVM) yang berdasarkan klasifikasi kandungan spam di dalam email telah dinilai dan dibandingkan. Dalam Kajian ini, pendekatan pembelajaran mesin hibrid yang mendapat inspirasi daripada Artificial Neural Network (ANN) dan Genetic Algorithm (GA) untuk mengesan spam-spam dengan berkesan. Perbandingan antara ANN biasa dan ANN-GA yang telah dipertingkatkan dengan ANN-GA dan SVM telah dibuat untuk menunjukkan algoritma yang mempunyai prestasi yang terbaik dalam mengesan spam. Algoritma-algoritma ini telah dilatih dan diuji dalam set 3 4061 E-mail dimana 1813 adalah spam dan selebihnya iaitu 2788 adalah tidak. Keputusan menunjukkan teknik ANN-GA yang dicadangkan memberi hasil yang lebih baik berbanding dengan teknik-teknik ANN yang biasa dan SVM. Keputusan dari ANN-GA yang dicadangkan memberi hasil ketepatan 93.71%, sementara ANN biasa hanya mendapat ketepatan 92.08% dan teknik SVM mendapat hasil ketepatan yang paling teruk iaitu 93.71%. Keputusan eksperimen ini mencadangkan bahawa keberkesanan model ANN-GA yang dicadangkan adalah cerah dan kajian ini memberi kaedah baru untuk melatih ANN dengan berkesan untuk mengesan spam.