A COMPARATIVE EVALUATION OF MACHINE LEARNING APPROACHES IN SMS SPAM DETECTION

SABER SALEHI

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

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JULY 2011

To my beloved parents, thank you for always being there for me, supporting me and encouraging me to be the best that I can be.

ACKNOWLEDGMENT

Praises to God for giving me the patience, strength and determination to go through and complete my study. I would like to express my appreciation to my supervisor, Assoc. Prof. Dr. Ali Selamat, for her support and guidance during the course of this study and the writing of the thesis. Without his continued support and interest, this thesis would not have been the same as presented here. I would like to dedicate this thesis to my family. Without their love and support I would have never come this far. Finally, I would also like to extend my thanks to my friends who have given me the encouragement and support when I needed it.

ABSTRACT

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 three supervised learning algorithm (Hybrid of Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO), Naive Bayes Classifier (NBC), Enhanced Genetic Algorithm (EGA)) based on classification of SMS contents were evaluated and compared. In this research, SAIS was hybridized by particle swarm optimization (PSO) for optimizing the performance of SAIS for spam filtering. PSO was used with mutation to reinforce the immune system's searches to find the best class in exemplar for classification. Results were improved using Hybrid SAIS and PSO. The proposed EGA was to achieve the best chromosomes which were grouped by the keywords. Then, the best chromosome with highest fitness value was selected as classifier. Simulated annealing (SA) was used with classical mutation and crossover to reinforce the efficiency of genetic searches. Achieved results represent the enhanced GA is markedly superior to that of a classical GA. These algorithms were trained and tested on a set of 4601 SMS messages in which 1813 were spams and 2788 were non-spams. Results showed that the proposed EGA technique gave better result compare to the hybrid SAIS and PSO and NBC techniques. Results also showed that the proposed EGA technique gave 99.87% accuracy, and the proposed NBC, hybrid of SAIS and PSO techniques gave 97.457% and 88.33% accuracy, respectively.

ABSTRAK

Pengesanan spam adalah masalah besar yang dianggap oleh ramai penyelidik menerusi pelbagai strategi yang dibangunkan. Dalam kajian ini, pengukur prestasi yang popular adalah ketepatan pengelasan yang memyokong keadaan positif palsu, negatif palsu dan ketepatan. Metrik ini telah dinilai melalui pengaplikasian tiga algoritma pembelajaran diselia iaitu *(Hybrid of Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO), Naive Bayes Classifier (NBC), Enhanced Genetic Algorthm (EGA))* berdasarkan penilaian dan perbandingan klasifikasi kandungan SMS. Dalam kajian ini, SAIS dihaibritkan oleh *particle swarm optimization (PSO)* dalam mengoptimumkan prestasi SAIS untuk penapisan spam. PSO beserta kaedah mutasi telah digunakan bagi mengukuhkan carian sistem imun dalam mencari kelas yang terbaik dalam contoh untuk pengelasan. Dengan menggunakan Hibrid SAIS dan PSO keputusan telah bertambah baik. EGA yang dicadangkan adalah untuk mencapai kromosom terbaik yang dikumpulkan menurut kata kunci. Kemudian, kromosom yang terbaik dengan nilai kecergasan tertinggi dipilih sebagai pengelas. *Simulated annealing (SA)* telah digunakan dengan mutasi klasik dan *crossover* untuk mengukuhkan kecekapan carian genetik. Keputusan yang mewakili GA yang telah dipertingkatkan menunjukkan keputusan yang ketara lebih tinggi daripada GA klasik. Algoritma ini telah dilatih dan diuji ke atas set yang mengandungi 4601 SMS yang mana 1813 daripadanya adalah spam dan 2788 bukan spam. Keputusan menunjukkan bahawa teknik EGA yang dicadangkan memberikan hasil yang lebih baik berbanding dengan SAIS Hibrid, PSO dan teknik NBC. Keputusan juga menunjukkan bahawa teknik EGA yang dicadangkan memberi 99.87% ketepatan, dan NBC yang dicadangkan iaitu Hibrid SAIS dan teknik PSO masing-masing memberikan ketepatan 97.457% dan 88.33%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

All wireless networks are comprised of independent computing nodes, have security problems, potentially. In wireless networks that are sorted by traditional mainframe computers without checking by central control, users capable to disturb the network with unlike intentions such as decreasing the wireless networks' confidentiality or integrity for accessing the data, using the capabilities of your network components for illegal or unsavory purposes or interfering the wireless network users with the reasonable activities. According to these unlike intentions, with the coming of internet and its protocols to control wide area connectivity, big and enterprise companies decided to protect their wireless networks, data, and applications against intrusive attacks and other threats by establishing perimeter defenses such as firewalls, VPNs, antivirus systems, intrusion prevention and intrusion detection systems (John*,* 2006).

Wireless sensor networks deploy rapidly. This kind of network is flexible and self-organizing system. This subject incurs to decrease the maintenance and deployment cost. Therefore, many application scopes such as health, home, military or environmental encourage using this kind of network. One of the significant challenges in many of these application areas is security. Sensor nodes must be far from the enemy because whenever the sensor network is reachable; sensor nodes could be destroyed and disturbed by the attackers. Cellular phone is one of the main areas in wireless network.

SMS service is the significant and main needs of people after voice among the fixed wireless users and cellular phone. For example, 80 percent of the U.S. work force used some sort of wireless devices such as mobile computing devices, cell phones and pagers. This news has positive and negative points for employee and companies, respectively. Because, this situation improves the employee's productivity. On the other hand, that is bad news for companies. Because, companies not ready for detecting the wireless network against breaches and debilitating viruses (Gohring, 2001).

On the other hand, by increasing the SMS users, two problems is created for operator. Firstly, the number of spam via SMS is increased. In addition, it causes the growth of SMS traffic. Spam occurs some problems in the wireless network area. Some of these problems are as the delay and lost non-spam SMS.

Totally, there has been a shortage of developed organization and functionality. So, that is exclusive the wireless viruses aren't more harmful. But, it makes change in the future. Industry analysts predict the spread of using wireless handheld and it causes increasing the abilities of new mobile. According to expected report which has been published, 5.9 billion personal digital assistants (PDAs), handsets and internet equipment will be provided with wireless ability at the end of 2008 (Gohring, 2001).

1.2 Problem Background

Over the years, many researchers proposed numerous techniques for detecting and handling the spam to mitigate the impact of spam on several scopes such as wireless network and e-mail and internet users. Most of these researches attend to increase and develop the accuracy of spam detection techniques. Several classifiers such as naïve Bayes, text compression and artificial neural network have been proposed to detecting and handling the spam. These classifiers are based on probabilistic techniques and machine learning. In the following, the existing problems which are solved by other researchers or in this dissertation are discussed. In this study most of the effort was done to remove these problems.

1.2.1 Self Learning and Self Adaptability of Naïve Bayes

Luo et al., (2010) proposed a new spam filtering based on naïve Bayes and AIS, and analyses the key problems of these algorithms. They assume that naïve Bayes is the most popular statistical-based anti spam method for its strong categorization and high precision. But it is weak in self-learning and self-adaptability. Artificial immune system has impressive performance on recognition, learning and memorizing. They attempt to combine the mechanism of naïve Bayes and artificial immune system, propose a hybrid spam filtering algorithm based on the two algorithms, and then solve the key problems of the algorithm. Their findings demonstrate that the hybrid algorithm not only achieves high classification accuracy at first, but also has self-learning, self-adaptability and robustness.

1.2.2 Mature Convergence in Genetic Algorithm

Sanpakdee et al., (2006) proposed a mechanism for filtering incoming spam mails by generating spam mail prototypes using genetic algorithm (GA). The keywords are categorized by their relating meaning into 7 groups as: adult, financial, commercial, beauty and diet, traveling, home-based business and gambling. Then, the string of chromosome which has 7 genes are defined. Each gen represents each word in chromosome by binary value. The accuracy of this mechanism is about 85% in average. But Sanpakdee et al. proposed technique has some weaknesses as below:

- For large words, it spends time to categorize the words however it was better to use feature selection methods to reduce dimensionality.
- Sanpakdee et al., technique is not flexible to create the chromosomes. Its method actually limited the chromosomes genes creation from limited keywords.
- The other limitation in Sanpakdee et al.'s technique is in GA operations. The cross over and mutation was for words of gene with the same group only.

The proposed technique by using GA in this dissertation solved the mentioned problems. In addition, in order to reinforce the efficiency of genetic searches and provide mature convergence, enhanced GA (EGA) using simulated annealing (SA) was proposed.

1.2.3 Deficiency in Optimization Methods

In the other research, Oda et al. (2003) used the artificial immune system for detecting spam. In particular, it tests the spam immune system against the publicly available spam assassin corpus of spam and non-spam, and extends the original system by looking at several methods of classifying email messages with the detectors produced by the immune system. The resulting system classifies the messages with similar accuracy compare to other spam filters but uses fewer detectors to do so which makes it an attractive solution for circumstances where processing time is at a premium. There is a deficiency in number of optimization methods in SAIS. This problem can be solved and eliminated using the other optimization methods besides mutation. In this dissertation, SAIS was hybridized by particle swarm optimization (PSO) for optimizing the performance of SAIS.

1.2.4 Reduction the Dimensionality

Feature selection is an important method for improving the efficiency and accuracy of text categorization algorithms by removing redundant and irrelevant terms from the corpus. Extensive researches have been done to improve the performance of spam detection by reduction to the dimensionality of features space. Almedia et al. (2009) compared the performance of most popular methods used as term selection techniques, such as document frequency (DF), information gain (IG), mutual information (MI), X^2 statistic, and odds ratio (OR) used for reducing the dimensionality of the term space with four well-known different versions of Naive Bayes spam filter. Regarding term selection techniques, Almedia et al. have found IG and $\chi 2$ statistic, most effective in aggressive term removal without losing categorization accuracy. On one hand, DF and

OR also usually provided an improvement on the filter's performance. On the other hand, the employment of MI offered poor results which frequently worsened the classifier's performance. But, in this dissertation the feature space is not that large for reduction so feature selection methods were not used in this dissertation.

1.3 Problem Statement

In this research, several existing classification methods was analyzed and simulated to investigate the efficiency of each technique. Accuracy of spam detection in SMS exchanging was evaluated using the naïve Bayes, genetic algorithm, simple artificial immune system based on classification of SMS contents. GA is well known and famous as optimization and searching method and in this research was used as classification. On the other hand, in naïve Bayes, estimation of probability of each email which occurs in spam category ($p(X|C)$) is different among the various data model. So, the main questions of this research are:

i. How we can find the optimization algorithms and their functionalities to understand the main concepts in spam detection?

ii. How we can adopt the naïve Bayes classifier, GA, enhanced GA, SAIS and hybrid SAIS and PSO for spam detection in SMS exchanging?

iii. How we can find the efficiency of naïve Bayes classifier, GA, enhanced GA, SAIS and hybrid SAIS and PSO in spam detection?

1.4 Dissertation Aim

The aim of this dissertation is to propose the hybrid SAIS and PSO, enhanced GA and naïve Bayes classifier to detect spam and evaluate the accuracy of spam detection in SMS exchanging.

1.5 Objectives

This research follows three objectives:

i. To study the existing methods in artificial immune system (AIS) and optimization methods to solve the convergence problem in GA for classification in spam detection.

ii. To develop and apply classical GA, enhanced GA by simulated annealing (SA), hybrid SAIS and PSO, SAIS and naïve Bayes classifier algorithms.

iii. To analyze effectiveness of enhanced GA, hybrid SAIS and PSO and naïve Bayes classifier algorithms in detecting spam.

1.6 Dissertation Scopes

This research is focusing on increasing accuracy and decreasing false positive and false negative based on classification of SMS content. The scopes of this research are as follow:

a. Initially, spam detection is applied in enhanced genetic algorithm (EGA), hybrid simple artificial immune system (SAIS) and particle swarm optimization (PSO) and Naive Bayes classifier (NBC).
b. The result of enhanced GA and hybrid SAIS and PSO as false positive, false negative and accuracy are compared to classical GA and SAIS, respectively.
c. The result of EGA, hybrid SAIS and PSO and NBC as false positive, false negative and accuracy are compared to each other.
d. The data sets used in this research has 4601 instances in which 39.4% are spam and each instance has 57 attributes. These data sets are prepared from UC Irvine. This website contains the data sets which are related to machine learning and intelligent systems (http://archive.ics.uci.edu/ml/datasets/Spambase).
e. Implementations of algorithms are done by java programming language.

1.8 Thesis Contribution

The contributions of this research are as follow:

- To develop Enhanced GA and hybrid of SAIS and PSO.
- To apply Enhanced GA, hybrid of SAIS and PSO, Naïve Bayes, GA and SAIS for spam detection.

- Analytical comparison of Enhanced GA, hybrid of SAIS and PSO, Naïve Bayes, GA and SAIS and the impact of PSO and simulated annealing (SA) on SAIS and GA.

1.9 Thesis Overview

Today, there is a need to detect spam in SMS because spam has fatal effect on SMS. A few works has been reported on spam detection in SMS. In our research, spam detection is modeled by hybrid SAIS and PSO, enhanced GA and naïve Bayes and the result is implemented by Java. These techniques are compared based on accuracy, false positive and false negative.

In the first chapter, wireless networks and the various technologies that are growing especially in SMS, has been described. Then, in the problem background the numerous techniques that are applied in spam detection by researchers are described shortly. Then, the main question of research, dissertation aim, dissertation objectives and dissertation scope were mentioned in detail. Finally, thesis overview and expected result has been described briefly.

Then, in the second chapter, spam and phishing and their fatal effects are described and then power of spam filtering is mentioned. Then, white List, black list and grey List which are as one of the current enterprise techniques to detect the client-side and server-side against spam and phishing are mentioned. The next subject is pre-acceptance and post-acceptance. There are two responses to spam at the server. The false positive and false negative is described after that and then machine learning and its methods as named inductive and deductive are described. Then some of the popular machine learning algorithms which are used in spam detection are described.

Then, in the research methodology chapter the related research methodologies that are used in spam detection is explained. Then, the dissertation objectives and main phases of research (contains 4 phases) which are used are described. In chapter 4, the data sets information is described. Then, the popular performance measure is described which is used as classification accuracy. Then, the proposed techniques by using Naive Bayes classifier (NBC), enhanced genetic algorithm (EGA), hybrid simple artificial immune system (SAIS) and particle swarm optimization (PSO) are presented, respectively. Finally, the results of mentioned techniques are represented and discussed. Finally in chapter 5 denotes the conclusion of this research.

1.10 Summary

In this chapter, wireless networks and the various technologies that are growing especially in SMS, has been described. Then, in the problem background the numerous techniques that are applied in spam detection by researchers are described shortly. Then, the main question of research, dissertation aim, dissertation objectives and dissertation scope were mentioned in detail. Finally, thesis overview and expected result has been described briefly.

REFRENCES

Abdel-Galil, T. K., Sharkawy, R. M., Salama, M. M. A., and Bartnikas, R. (2005). Partial discharge pulse pattern recognition using an inductive inference algorithm. IEEE Transactions on Dielectrics and Electrical Insulation, 12(2), 320-327.

Aksoy, M., Çagıl, G., and Türker, A. (2000). Number-plate recognition using inductive learning. Robotics and Autonomous Systems, 33(2-3), 149-153.

Alpaydin, E. (2004). Introduction To Machine Learning (Adaptive Computation And Machine Learning): The MIT Press.

Bhaduri, A. (2009). Credit scoring using Artificial Immune System algorithms: A comparative study. 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), 1540-1543.

Blanzieri, E., and Bryl, A. (2008). A Survey Of Learning-Based Techniques Of Email Spam Filtering. Department of Information and Communication Technology, University of Trento, Trento, Italy, 149-153.

Bonabeau, E., Dorigo, M., and Theraulaz, G. (2002). Inspiration For Optimization From Social Insect Behavior. Nature, 406(6).

Chhabra, S. (2005). Fighting Spam, Phishing And Email Fraud: University Of California Riverside, 540-543.

Cormack, G. V., Hidalgo, J.M., and Sánz, E.P. (2007). Feature engineering for mobile (SMS) spam filtering. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07, 871, 149-153.

Dasgupta, D., Ji, Z., and Gonzalez, F. (2003). Artificial immune system (AIS) research in the last five years. The 2003 Congress on Evolutionary Computation, 2003. CEC '03., 123-130.

De Castro, L., and Von Zuben, F. (2002). Learning and optimization using the clonal selection principle. IEEE Transactions on Evolutionary Computation, 6(3), 239-251.

Dorigo, M., Maniezzo, V., and Colorni,A. (1996). Ant system: optimization by a colony of cooperating agents. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 26(1), 29-41.

Dorigo, M., and Gambardella, L. (1997). Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation, 1(1), 53-66.

Dorigo, M. (2009). Swarm-bots and Swarmanoid : Two experiments in embodied swarm intelligence. In Proceedings of Web Intelligence, 123-130.

Eberhart, R. C. (1995). Particle swarm optimization: developments, applications and resources. Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), 81-86.

Eberhart, R. C., Shi, Y. (2001). Particle Swarm Optimization: Developments, Applications And Resources: Proceedings Of The Congress On Evolutionary Computation, Seoul, South Korea, 123-130.

Epstein, J. M., and Axtell, P. (1998). Growing Artifcial Societies: MIT Press.

Graham, j. (2003). The Spammers Compendium: MIT Spam Conference.

Guo, P., Wang, X., and Han, Y. (2010). The Enhanced Genetic Algorithms for the Optimization Design. Architectural Engineering, 2990-2994, 163-160.

Guzella, T. S., and Caminhas, W.M. (2009). A review of machine learning approaches to Spam filtering. Expert Systems with Applications, 36(7), 10206-10222.

Huang, H., and Hsu, C. (2002). Bayesian classification for data from the same unknown class. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 32(2), 137-145.

Jerome, H., Carter, M. (2000). The Immune System as a Model for Pattern Recognition and Classification. Section of Medical Informatics, Division of General Internal Medicine, University of Alabama, 123-130.

Kennedy, J., and Eberhart, R. (1995). Particle Swarm Optimization. Proc.Proceedings Of The IEEE International Conference On Neural Networks, Perth, Australia, IEEE Service Center, 183-189

Kubotani, H., and Yoshimura, K. (2003). Performance evaluation of acceptance probability functions for multi-objective SA. Computers and Operations Research, 30(3), 427-442.

Kumar, P. (1990). Convergence of adaptive control schemes using least-squares parameter estimates. IEEE Transactions on Automatic Control, 35(4), 416-424.

Leung, K., Cheong, F., and Cheong, C. (2007). Consumer credit scoring using an artificial immune system algorithm. IEEE Congress on Evolutionary Computation, 3377-3384.

Lingling L., Z., C., and Gao, Z. (2003). Multi-Strategy Combined Learning Mechanism Suitable For Designing Expert System. Proceedings Of The Second International Conference On Machine Learnkg And Cybernetics, 123-130.

Luo, Q., Liu, B., Yan, J., and He, Z. (2010). Research of a Spam Filtering Algorithm Based on Naïve Bayes and AIS. International Conference on Computational and Information Sciences, 152-155.

Merton, R. C. (1974). On The Pricing Of Corporate Debt: The Risk Structure Of Interest Rates. Journal Of Finance, 29(2), 449–470.

Mitchell, T. M. (2006). The Discipline of Machine Learning. Machine Learning.

Nature, T., Theory, L., and Vapnik, V.N. (2011). Book reviews. Applied physiology, nutrition, and metabolism Physiologie appliquée, nutrition et métabolisme, 36(3).

Oda, T., and White, T. (2003). Increasing the accuracy of a spam-detecting artificial immune system. The 2003 Congress on Evolutionary Computation, 2003. CEC '03., 390-396.

Pan, Q., Wang, L., Tasgetiren, M., and Zhao, B. (2007). A hybrid discrete particle swarm optimization algorithm for the no-wait flow shop scheduling problem with makespan criterion. The International Journal of Advanced Manufacturing Technology, 38(3-4), 337-347.

Preisach, B., Schmidt, H. , Decker, L., and Reinhold. (2008). Data Analysis, Machine Learning and Applications: Proceedings Of The 31st Annual Conference Of The Gesellschaft Klassifikation E.V, 123-130.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian Approach To Filtering Junk Email. Technical Report WS-98-05, AAAI. AAAI Workshop On Learning For Text Categorization, Madison, Wisconsin, 223-230.

Sanpakdee, U., Walairacht, A., and Walairacht, S. (2006). Adaptive Spai Mail Filtering Using Genetic Algorithm. 2006 8th International Conference Advanced Communication Technology, 441-445.

Secker, A., Freitas, A., and Timmis, J. (2003). AISEC: An Artificial Immune System for E-mail Classification. The 2003 Congress on Evolutionary Computation, 2003. CEC '03., 131-138.

Serafni, P., and Jaszkiewicz, A. (1992). Simulated Annealing For Multi Objective Optimization Problems. Proceedings Of The 10th International Conference On Multiple Criteria Decision Making, 123-130.

Sharvani, .G.S, Rangaswamy, T.M. (2009). Different Types of Swarm Intelligence Algorithm for Routing. International Conference on Advances in Recent Technologies in Communication and Computing, 604-609.

Shi-yong, L. (2003). Progresses in Ant Colony Optimization Algorithm with Applications. Computer Automated Measurement and Control, 623-630.

Tran, M., Armitage, G. (2004). Evaluating The Use Of Spam-Triggered TCP/IP Rate Control To Protect SMTP Servers. Proceedings Of The Australian Telecommunications Networks And Applications Conference (ATNAC 2004), Sydney, Australia, 123-130.

Twining, R. d., Williamson, M., Mowbray, M., Rahmouni, M., and Sergeant, M. (2003). Internet Level Spam Detection And Spamassassin 2.50. In Proceedings Of The 2003 Spam Conference, Cambridge MA, 153-160.

Twining, R. D., Williamson, M.M., Mowbray, M., and Rahmouni, M. (2004). Email Prioritization : reducing delays on legitimate mail caused by junk mail increased dramatically . These unwanted messages clutter up users ' Email Prioritization : reducing delays on legitimate mail caused by junk mail, 143-150.

Vapnik, V. N. (1995). The Nature Of Statistical Learning Theory. New York: Springer.

Wang, H., Ma, C., and Zhou, L. (2009). A Brief Review of Machine Learning and Its Application. International Conference on Information Engineering and Computer Science, 1-4.

Wang, Z., and Cui, D. (2009). A Hybrid Algorithm Based on Genetic Algorithm and Simulated Annealing for Solving Portfolio Problem. 2009 International Conference on Business Intelligence and Financial Engineering, 106-109.

Xin Jin, R. B., and Gao, X.Z. (2006). Notice of Violation of IEEE Publication Principles An Artificial Immune Recognition System-based Approach to Software Engineering Management: with Software Metrics Selection. Intelligent Systems Design and Applications. Sixth International Conference, 123-130.

Yang, W., Wan, W., Lin, G., and Zhang, L. (2007). An Efficient Intrusion Detection Model Based on Fast Inductive Learning. 2007 International Conference on Machine Learning and Cybernetics, 3249-3254.

Zhang, j., Liu, k., Tan, Y., and He, x. (2008). Random Black Hole Particle Swarm Optimization And Its Application. Signal Processing, 359-365.

Zhu, Z., and Sun, Y. (2009). Transmission Line Fault Classification Based on Wavelet Singular Entropy and Artificial Immune Recognition System Algorithm. 2009 2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS), 154-157.