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.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
SVM	-	Support Vector Machine
BP	-	Back Propagation
MLP	-	Multilayer Perceptron
PSO	-	Particle Swarm Optimization
PCA	-	Principal Component Analysis
FFNN	-	Feed Forward Neural Network
TP	-	True Positive
TN	-	True Negative
FN	-	False Negative
FP	-	False Positive
A	-	Accuracy
SR	-	Spam Recall

CHAPTER 1

INTRODUCTION

1.1 Introduction

A new channels of communication has offered by an internet to enable sending an e-mail to thousand of kilometers away. A lot of gates have opened by this middle of communication for actually free prevent e-mailing, reaching out to hundreds of thousands clients during seconds. Anyway, this independence of communication can be abused. A phenomenon of spam has increased in the last year and become a serious problem that will prevent the capability of communication via e-mail (Kågström, 2005).

There is no a specific and suitable sentence that says what exactly the spam, even if every spam message will be rapidly recognized by email user. Merriam-Webster Online Dictionary describes spam as “unsolicited usually commercial e-mail sent to a large number of addresses”. Many kind of commercial spam aim to express religious or political views, trick the aim audience with undertakings of fortune, spread senseless chain letters and infect the receivers’ computer with viruses. Most users agree that unsolicited mail is pubic frustration, even if one can talk over that what is unsolicited mail for one user can be an exciting email message for another.

Spam is usually economically beneficial to the sender; for this reason spam has become a critical problem. The low budget of e-mail as a communication medium virtually guaranties returns. Even though a little percentage of people reply to the spam advertising message by purchasing the product, this can be value the money and the time spent for sending unsolicited e-mails.

Spam has growth over the last years. Up to this time, 86.6% of all e-mail messages are unsolicited mail, according to the Spam-o-meter website. And In 2005 the total worldwide financial losses caused by spam were expected at $50 billion, in 2009 the same value was expected already at $130 billion. The main problem concerning unsolicited mail is which it is the victim who is spends money for the spam in the field of their bandwidth, time and disk space. This will be very expensive even for a non-large corporation with only 20 workers who each receive 21 unsolicited mail e-mails per day. If the classify and delete spam will takes 5 seconds, then the corporation will takes around 30 minutes daily to detect spam from non-spam. 20 spam messages every day as the statistics shows is a very small number for a corporation that is critical to unsolicited mail.

Increasing the flood of spam steadily requires a significant need to be managed. A many of on-going studies are attempt to solve the problem. However, there is an increasing necessity for speedily current anti-spam solutions to protect them because the most of e-mail users are restless.

1.2 Problem background

Over the years, a numerous techniques have proposed by many researchers for handling and detecting the unsolicited mail to mitigate the influence of spam on different scopes such as wireless network and internet and e-mail users. Most of these researches aimed to extend and develop the accuracy of spam detection techniques. Different classifiers such as naïve Bayes, artificial neural network and

text compression have been proposed to handle and detect the spam. These classifiers are depending on probabilistic techniques and machine learning. In the following, the existing problems which are partially solved by other researchers are discussed.

Many researches are founded in applying Artificial Neural Network (ANN). The major cons of ANN are that it needs considerable time to select the parameter and network training. On the other hand, ANN can detect very accurate results based on previous researches have shown that, which are from time to time more accurate than those of the symbolic classifiers. Özgür *et al.* (2004) also applied the tests that a full of 750 emails (410 spams and 340 non-spam emails) were implemented and a success rate of about 90% was done. Unfortunately, the existing ANN with poor parameter produces low detection rate. These researches displayed that ANN can be effectively used for automatic email filing into mailboxes and unsolicited mail filtering.

Özgür *et al.* (2004) proposed means of antispam dynamic checking for agglutinative languages in common and for Turkish in specific, depend on Bayesian filters and (ANN). There are two components for algorithms that adaptive. The first one categorizes the e-mails by using the roots and the second one deal with the morphology. The ANN has two structures, which are multi-layer perceptron and single layer perceptron, and they are dedicated and the binary and probabilistic are using to determine the inputon the network. Three approaches are dedicated for Bayesian classification, which are: probabilistic, binary, and advance probabilistic models. In the experiments, a total of 750 e-mails (410 spam and 340 non-spam) were used and a success rate of about 90% was achieved.

A performance evaluation of different types of ANN presented by Silva *et al.* (2004) used to mechanically filter and classify real samples of web spam depend on their contents. Some of evaluated approaches which indicated by the result have a big potential while they are proper to deal with the problem and clearly outperform the state-of-the-art techniques.

Support Vector Machine (SVM) considered as a classification technique that doesn't require a statistical data model to reduce the classification errors immediately. This method will be popular when developed to most of real world classification cases because of it has permanently high classification accuracy and uncomplicated implementation. Drucker *et al.* (1999) implemented the methods to spam filtering, testing it against three other text classification algorithms: Ripper, Rocchio and Boosting Decision Trees (BDT). Both BDT and SVM provided "acceptable" performances, with SVM given lower training requirements.

Christina *et al.* (2010) generated spam and non-spam message corpus from the latest mails and employed machine learning techniques to create the model. The efficiency of the model is estimated using 10-fold cross validation and observed that Multilayer Perceptron classifier out performs other classifiers and the false positive rate also very low compared to other algorithms. Email spam filters using this approach can be adopted either at mail server or at mail client side to minimize the quantity of unsolicited mail messages and to minimize the risk of productivity loss, bandwidth and storage usage.

1.3 Problem Statement

There are many elements to be considered in Artificial Neural Network (ANN), such as the number of input, hidden and output nodes, bias, minimum error and the type of activation/transfer function. All these elements will influence the convergence of ANN learning. There are some algorithms such as Genetic Algorithm that have been used to determine some parameters and supply the best pattern of weight in order to enhance the ANN learning. There is another problem with ANN, which existing algorithm without parameter tuning which gives poor detection on spam email and requires considerable time to select training rate parameter.

1.4 Purpose of Study

The purpose of this study is to increase accuracy and reduce false positive and false negative to find optimum parameter for ANN with the use of optimization algorithm (GA) to efficiently train ANN for spam detection.

1.5 Project Objectives:

This study follows three objectives:

i. Collect the dataset and select the most significant features.
ii. Implement classical ANN and develop improved ANN by incorporating Genetic Algorithm.
iii. Analyze effectiveness of improved ANN by comparing with classical ANN and SVM.

1.6 Project Scope:

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

i. Initially, spam detection is applied in improved ANN, support vector machine (SVM) based on email content.
ii. The result of classical ANN and SVM will be measured in terms of false positive, false negative and accuracy.
iii. Comparative will be done with support vector machine.

iv. The dataset will be collected from UCI Irvine. This website includes the datasets that are regarding to machine learning and intelligent systems (http://archive.ics.uci.edu/ml/datasets/Spambase).

1.7 The Significant of Study

Today, there is a need to detect spam in E-Mail because spam has fatal effect on E-Mail. Many works have been reported on spam detection in E-Mail. In this study, improving of neural network algorithm has been done to save time for parameter selection and to optimize network training in order to increase the performance of spam detection, and to evaluate the performance of ANN and SVM based on accuracy, false positive and false negative to show which one is suitable to be used in spam detection; the finding of this study are important to increase the performance of spam detection.

1.8 Organization of Report

The study consists of 6 chapters. Chapter one describes the introduction, background of the study, study objectives and problem statement, the scope of the study and its significant. The second chapter reviews available and related literature on Spam Detection, Artificial Neural Network (ANN), and Support Vector Machine (SVM). Chapter three describes the methodology of the study along with the appropriate framework. The fourth chapter provides the analysis of the preliminary findings of the study and conclusion of initial work. Chapter four discusses the process of implementation. Chapter five covers the statistical analyses of results. Finally, chapter six summarizes the whole study.

REFERENCES

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

Agrawal, B., Kumar, N. & Molle, M. Controlling spam emails at the routers. Communications, 2005. ICC 2005. 2005 IEEE International Conference on, 2005. IEEE, 1588-1592.

Aksoy, M. S., Çağıl, G. & Türker, A. K. (2000). Number-plate recognition using inductive learning. *Robotics and Autonomous Systems,* 33**,** 149-153.

Alsmadi, M. K. S., Omar, K. B. & Noah, S. A. (2009). Back propagation algorithm: the best algorithm among the multi-layer perceptron algorithm. *IJCSNS International Journal of Computer Science and Network Security,* 9**,** 378-383.

Chhabra, S. 2005. *Fighting spam, phishing and email fraud.* UNIVERSITY OF CALIFORNIA.

Christina, V., Karpagavalli, S. & Suganya, G. (2010). Email Spam Filtering using Supervised Machine Learning Techniques. *International Journal on Computer Science and Engineering (IJCSE) Vol,* 2**,** 3126-3129.

Cui, B., Mondal, A., Shen, J., Cong, G. & Tan, K. L. On effective e-mail classification via neural networks. Database and Expert Systems Applications, 2005. Springer, 85-94.

Drake, C. E., Oliver, J. J. & Koontz, E. J. Anatomy of a phishing email. Conference on Email and Anti-Spam, 2004.

Drucker, H., Wu, D. & Vapnik, V. N. (1999). Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on,* 10**,** 1048-1054.

Freeman, J. A. 1994. *Simulating neural networks with Mathematica,* Addison-Wesley Reading (Mass.) etc.

Graham-Cumming, J. The spammers' compendium. Proceedings of the spam conference, 2003.

Gulyás, C. 2006. *Creation of a Bayesian network-based meta spam filter, using the analysis of different spam filters.* Citeseer.

Hershkop, S. 2006. *Behavior-based email analysis with application to spam detection.* Columbia University.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98***,** 137-142.

John, G. H., Kohavi, R. & Pfleger, K. Irrelevant features and the subset selection problem. Proceedings of the eleventh international conference on machine learning, 1994. San Francisco, 121-129.

K, P. 2009. *Neural Network Classification Based on Qualification of Uncertainty.* Doctor of Philosophy Thesis, Murdoch University.

Kågström, J. 2005. *Improving naive bayesian spam filtering.* Master thesis, Mid Sweden University.

Kohavi, R. & Sommerfield, D. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 1995. 192-197.

Lai, C. C. & Tsai, M. C. An empirical performance comparison of machine learning methods for spam e-mail categorization. Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on, 2004. IEEE, 44-48.

Lai, G. H., Chen, C. M., Laih, C. S. & Chen, T. (2009). A collaborative anti-spam system. *Expert Systems with Applications,* 36**,** 6645-6653.

Lee, T. L. (2006). Neural network prediction of a storm surge. *Ocean Engineering,* 33**,** 483-494.

Levine, J. R., Young, M. L. & Everett-Church, R. 2004. *Fighting spam for dummies*, For Dummies.

Li, L. L., Li, Z. G., Zhou, C. L. & Gao, Z. H. Multi-strategy combined learning mechanism suitable for designing expert system. Machine Learning and Cybernetics, 2003 International Conference on, 2003. IEEE, 2291-2295.

Marzi, H. & Turnbull, M. Use of neural networks in forecasting financial market. Granular Computing, 2007. GRC 2007. IEEE International Conference on, 2007. IEEE, 516-516.

Montana, D. J. & Davis, L. Training feedforward neural networks using genetic algorithms. Proceedings of the eleventh international joint conference on artificial Intelligence, 1989. San Mateo, CA, 762-767.

Özgür, L., Güngör, T. & Gürgen, F. (2004). Spam mail detection using artificial neural network and Bayesian filter. *Intelligent Data Engineering and Automated Learning–IDEAL 2004***,** 505-510.

Romano, M., Liong, S. Y., Vu, M. T., Zemskyy, P., Doan, C. D., Dao, M. H. & Tkalich, P. (2009). Artificial neural network for tsunami forecasting. *Journal of Asian Earth Sciences,* 36**,** 29-37.

Sahami, M., Dumais, S., Heckerman, D. & Horvitz, E. A Bayesian approach to filtering junk e-mail. Learning for Text Categorization: Papers from the 1998 workshop, 1998. Madison, Wisconsin: AAAI Technical Report WS-98-05, 98-105.

Sergeant, M. Internet level spam detection and SpamAssassin 2.50. Proceedings of the 2003 Spam Conference, Cambridge MA, 2003.

Sexton, R. S. & Sikander, N. A. (2001). Data mining using a genetic algorithm-trained neural network. *Intelligent Systems in Accounting, Finance and Management,* 10**,** 201-210.

Silva, R. M., Almeida, T. A. & Yamakami, A. (2004). Artificial Neural Networks for Content-based Web Spam Detection.

St Sauver, J. (2005). Spam zombies and inbound flows to compromised customer systems.

Taylor, J. W. & Buizza, R. (2002). Neural network load forecasting with weather ensemble predictions. *Power Systems, IEEE Transactions on,* 17**,** 626-632.

Twining, R. D., Williamson, M. M., Mowbray, M. & Rahmouni, M. Email prioritization: Reducing delays on legitimate mail caused by junk mail. USENIX Annual Technical Conference, 2004.

Umamaheswari, K., Sumathi, S., Sivanandam, S. N. & Anburajan, K. K. N. Efficient Finger Print Image Classification and Recognition using Neural Network Data Mining. Signal Processing, Communications and Networking, 2007. ICSCN '07. International Conference on, 22-24 Feb. 2007 2007. 426-432.

Vainio, M. 2001. *Artificial neural network based prosody models for Finnish text-to-speech synthesis*, Citeseer.

Vapnik, V. 1999. *The nature of statistical learning theory*, springer.

Wang, H., Ma, C. & Zhou, L. A Brief Review of Machine Learning and Its Application. Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on, 19-20 Dec. 2009 2009. 1-4.

Woitaszek, M., Shaaban, M. & Czernikowski, R. Identifying junk electronic mail in microsoft outlook with a support vector machine. Applications and the Internet, 2003. Proceedings. 2003 Symposium on, 2003. IEEE, 166-169.

Yang, W., Wan, W., Guo, L. & Zhang, L. J. An Efficient Intrusion Detection Model Based on Fast Inductive Learning. Machine Learning and Cybernetics, 2007 International Conference on, 2007. IEEE, 3249-3254.

Yao, X. (1993). A review of evolutionary artificial neural networks. *International journal of intelligent systems*, 8**,** 539-567.

Zdziarski, J. 2005. *Ending spam: Bayesian content filtering and the art of statistical language classification*, No Starch Press.

Zhu, Y. & Tan, Y. (2011). A Local-Concentration-Based Feature Extraction Approach for Spam Filtering. *Information Forensics and Security, IEEE Transactions on*, 6**,** 486-497.

Weiyu Yi (2005). Artificial Neural Networks.339229

C.J. Lin, Foundations of support vector machines: A note from an optimization point view, Neural Computation 13 (2) (2001) 307–317.

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

B. Sullivan, "Who profits from spam? Surprise," MSNBC, 2003 http://www.msnbc.msn.com/id/3078642/.

Haza Nuzly (2006). Particle swarm optimization for neural network learning enhancement. M.Sc. Thesis, University Technology of Malaysia.

(http://www.securelist.com/en/analysis/204792249/Spam_in_September_2012)