**UNIVERSITI PUTRA MALAYSIA**

# *AN ENSEMBLE LEARNING METHOD FOR SPAM EMAIL DETECTION SYSTEM BASED ON METAHEURISTIC ALGORITHMS*

**AMIR RAJABI BEHJAT**

**FSKTM 2015 49**

# AN ENSEMBLE LEARNING METHOD FOR SPAM EMAIL DETECTION SYSTEM BASED ON METAHEURISTIC ALGORITHMS

By

AMIR RAJABI BEHJAT

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

June 2015

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

**AN ENSEMBLE LEARNING METHOD FOR SPAM EMAIL DETECTION SYSTEM BASED ON METAHEURISTIC ALGORITHMS**

By

**AMIR RAJABI BEHJAT**

**June 2015**

**Chairman: Aida Mustapha, PhD**

**Faculty: Computer Science and Information Technology**

In email spam detection, not only different parts and content of emails are important, but also the structural and special features of these emails have effective rule in dimensionality reduction and classifier accuracy. For example, the spammer changes patterns of message for making spam such as writing the message by JavaScript, using different advertising images and words to form features or attributes. Even the smart people are unable to report an email as a spam when the spammer tries to defraud them.

The aim of data mining is to search and find undetermined patterns in huge databases. A well known task is classification that predicts the class of new instances using known features or attributes automatically. Major problems in classification task are large amount of training data, large number of features and different behavior of data streams that reduce accuracy and increase computational cost in classifier training phase. Feature subset selection and classifier ensemble learning are familiar techniques with high ability to optimize above problems. Recently, various techniques based on different algorithms have been developed. However, the classification accuracy and computational cost are not satisfied.

In order to address the challenges that mentioned above in this study, in the first phase, a novel architecture based on ensemble feature selection techniques include Modified Binary Bat Algorithm (NBBA), Binary Quantum Particle Swarm Optimization (QBPSO) Algorithm and Binary Quantum Gravitational Search Algorithm (QBGSA) is hybridized with the Multi-layer Perceptron (MLP) classifier in order to select relevant feature subsets and improve classification accuracy. In the second phase, a classifier ensemble learning model is proposed consisting of separate outputs: (i) To select a relevant subset of original features based on Binary Quantum Gravitational Search Algorithm (QBGSA), (ii) To mine data streams using various data chunks and overcome a failure of single classifiers based on SVM, MLP and K-NN algorithms.

An experimental analysis is conducted by several experiments to evaluate the performance of the proposed ensemble methods which has been tested on the 4 benchmark datasets, namely LingSpam, SpamAssassin, Spambase and CS-DMC2010. In comparison to different single algorithms for feature selection, experimental results show that the proposed ensemble method is able to reduce dimensionality, the number of irrelevant features and produce reasonable classifier accuracy. Experiments demonstrate that ensemble classifier learning method produces better accuracy mining data streams and selecting subset of relevant features comparing other single classifiers.

In addition, experiments prove that the ensemble algorithms select highly relevant features to feed the MLP comparing individual techniques in terms of classifier performance through lower false positive, higher accuracy, and better CPU time.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**KAEDAH PEMBELAJARAN ENSEMBEL BAGI SISTEM PENGESANAN EMEL SPAM BERASASKAN ALGORITMA METAHEURISTIK**

Oleh

**AMIR RAJABI BEHJAT**

**Jun 2015**

**Pengerusi: Aida Mustapha, PhD**

**Fakulti: Sains Komputer dan Teknologi Maklumat**

Dalam pengesanan spam e-mel, bukan sahaja bahagian-bahagian yang berbeza dan kandungan e-mel adalah penting, tetapi juga ciri-ciri struktur dan istimewa e-mel ini mempunyai peraturan yang berkesan dalam pengurangan dimensi dan ketepatan pengelas. Sebagai contoh, penceroboh itu menukar corak mesej untuk membuat spam seperti menulis mesej dengan JavaScript, menggunakan imej pengiklanan yang berbeza dan kata-kata untuk membentuk ciri-ciri atau sifat-sifat. Malah orang pintar tidak mampu melaporkan e-mel sebagai spam apabila pengiklan itu cuba untuk menipu mereka.

Tujuan perlombongan data adalah untuk mencari dan mendapati corak yang belum ditentukan di dalam pangkalan data yang besar. Satu tugas yang terkenal adalah pengelasan yang meramalkan kelas contoh baru menggunakan ciri-ciri yang diketahui atau atribut secara automatik. Masalah utama dalam tugas pengelasan adalah jumlah besar data latihan, bilangan besar ciri-ciri dan tingkah laku yang berbeza aliran data yang mengurangkan ketepatan dan meningkatkan kos pengkomputeran dalam fasa latihan pengelas. Ciri-ciri pemilihan subset dan pembelajaran pengelas gabungan adalah teknik biasa dengan keupayaan yang tinggi untuk mengoptimumkan masalah di atas. Baru-baru ini, pelbagai teknik berdasarkan algoritma yang berbeza telah dibangunkan. Walau bagaimanapun, ketepatan klasifikasi dan kos pengkomputeran tidak kepuasan.

Dalam usaha untuk menangani cabaran-cabaran yang dinyatakan di atas dalam kajian ini, dalam fasa pertama, seni bina novel berdasarkan teknik-teknik pemilihan ciri gabungan termasuk Modified Binary Bat Algoritma (NBBA), Binary Kuantum Zarah Swarm Optimization (QBPSO) Algoritma dan Binary Graviti Kuantum Carian algoritma (QBGSA) adalah hibrid dengan Multi-lapisan Perceptron (MLP) pengelas untuk memilih subset ciri yang berkaitan dan meningkatkan ketepatan pengelasan. Dalam fasa kedua,model gabungan pembelajaran pengelas adalah dicadangkan terdiri daripada dua peringkat: (i) Untuk memilih subset yang berkaitan dengan ciri-ciri asal berdasarkan Binary Kuantum Graviti Cari Algoritma (QBGSA), (ii) untuk melombong data menggunakan pelbagai ketulan data dan mengatasi kegagalan penjodoh tunggal berdasarkan SVM, MLP dan algoritma K-NN.

Analisis eksperimen dijalankan oleh beberapa eksperimen untuk menilai prestasi kaedah gabungan yang dicadangkan yang telah diuji pada 4 dataset penanda aras, iaitu LingSpam, SpamAssassin, Spambase dan CSDMC2010. Berbanding dengan algoritma tunggal yang berbeza untuk pilihan ciri, keputusan eksperimen menunjukkan bahawa kaedah gabungan yang dicadangkan mampu mengurangkan kematraan, bilangan ciri-ciri yang tidak relevan dan menghasilkan ketepatan pengelas berpatutan. Eksperimen menunjukkan bahawa kaedah pembelajaran gabungan pengelas menghasilkan yang lebih baik perlombongan ketepatan aliran data dan memilih subset ciri-ciri yang berkaitan membandingkan penjodoh tunggal lain.

Di samping itu, eksperimen membuktikan bahawa algoritma gabungan pilih ciri-sangat relevan untuk memberi makan MLP membandingkan teknik individu dari segi prestasi pengelas melalui positif palsu yang lebih rendah, ketepatan yang lebih tinggi, dan masa CPU yang lebih baik.

**ACKNOWLEDGEMENTS**

First of all I would like to thank my Allah Almighty Who gave me the courage, health, and energy to accomplish this dissertation in due time and without Whose help this study would have not been possible to complete within the time limits.

I would like to express my sincere gratitude to my supervisor Dr.Aida Mostapha for giving me an opportunity to start this study. Her comments and suggestions for further development as well as her assistance during writing this thesis are invaluable to me. Her background of study on data mining and research style has provided for me an exceptional opportunity to learn more.

I would like to express my sincere thanks and appreciation to the supervisory committee members Professor Dr. Md Nasir B Sulaiman, Associate Professor Dr. Norwati mustapha and Professor Dr. Hossein Nezamabadi-pour for their guidance, valuable suggestions and advice throughout this work in making this success.

My deepest appreciation to my father and mother who has been supportive and patiently waiting for me to complete my study. At last not at least, i owe my sincere thanks to my wife Ms. Fariba Dolati and my son Saman for their encouragement and affirmation, which made it possible for me to achieve this work.

For the others who have directly or indirectly helped me in the completion of my work, I thank you all.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

**Aida Mustapha, PhD**
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

**Md. Nasir Sulaimane, PhD**
Associate professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

**Norwati Mustapha, PhD**
Associate professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

**Hossein Nezamabadi-pour, PhD**
Professor
Faculty of Engineering
Shahid Bahonar University of Kerman, Kerman, Iran
(Member)

**BUJANG BIN KIM HUAT, PhD**
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

**Declaration by graduate student**

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ______________________ Date: ____________________

Name and Matric No.: ________________________________________

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| BBA | Binary Bat Algorithm |
| BGA | Binary Genetic Algorithm |
| BGSA | Binary Gravitational Search Algorithm |
| BQGSA | Binary Quantum Gravitational Search Algorithm |
| BQPSO | Binary Quantum Particle Swarm Optimization |
| KNN | $K$-Nearest Neighbor |
| MLP | Multi-Layer Perceptron |
| PSO | Particle Swarm Optimization |
| QC | Quantum Computing |
| QEIA | Quantum Evolutionary-Inspired Algorithm |
| SVM | Support Vector Machine |

# CHAPTER 1

# INTRODUCTION

## 1.1 Motivation

Electronic mails (e-mails) are the most efficient and effective communication in the world. Recently, this technology has posed a serious spamming problem via spam or junk emails over the Internet (Wu et al., 2010). Large number of spam and junk emails consumes high bandwidth resources in a network environment. They are also able to quickly block a server by occupying storage space, which is highly risky for large sites that have thousands of users (Jindal and Liu, 2007; Lee et al., 2010). In a more recent development, spam emails have started to alter the content of emails. As the patterns of spam emails change over time, existing detection models that are built on old data has become unsuitable for classifying new incoming emails (Aggarwal, 2012). This scenario motivates a continuous effort in building better spam detection systems with higher accuracy.

In general, a spam detection system is related to a classification problem with two classes; spam or non-spam. The aim of spam detection is to separate spam and non-spam emails accurately (Batista, 2000; Islam et al., 2005; Mohammad and Zitar, 2011) with the lowest error rate and the highest accuracy (Michalak and Kwasnicka, 2006; Chang and Poon, 2009). Although there are a number of studies that have attempted various classification techniques to classify emails into spam and non-spam, the researches are constantly challenged by the large number of features in email content (Androutsopoulos, Koutsias, Chandrinos, Paliouras and Spyropoulos, 2000; Chuan et al., 2005), high computational cost for feature extraction and classification (Mohammad and Zitar, 2011; Androutsopoulos, Paliouras, Karkaletsis, Sakkis, Spyropoulos and Stamatopoulos, 2000), gap in the size of training data, unstable error rate (Fawcett, 2003; Fan, 2004), changes in spam email content over time, and finally imbalance between False Positive (FP) rate and False Negative (FN) rate (Blanzieri and Bryl, 2008).

The challenge in classification of emails is mainly attributed to their content; the large number of features, which are mostly words. A high number of features increases the number of examples during the training phase, therefore simultaneously increases the complexity and computational cost (Aha et al., 1991). When the number of words in emails, whether spam or non-spam, is large, the amount of undesired features increases the speed of training. On the other hand, a small number of features is not equally desireable because it may be insufficient for the training phase and to mask messages. In effort to decrease dimensionality of header and content features in spam detection systems, feature selection is highly critical to train only a subset of features from the entire set, hence removing all irrelevant features (Gomez et al., 2012).

This research formulates the feature selection problem in spam email detection as an optimization problem, for which it is to find the best solution from all feasible solutions. In spam detection system, this is essentially the task of finding the best set of features that accurately represents the spam emails from all features available in the email content. One obvious avenue to tackle an optimization problem is by exploring metaheuristic algorithms, which are widely recognized as a practical approach for optimization.

Metaheuristic algorithms follow the biological behavior in the nature (Yang, 2011), for example, the Particle Swarm Optimization (PSO) is based on fish schooling and birds flocking behaviors. These algorithms are applied as individual feature selection algorithm in most spam detection systems. They suffer from the risk of choosing a wrong feature as a solution among many equally optimal features in the feature space. Most algorithms are also prone to get trapped in local optimum and maybe not be truly functioning to select the exact relevant features (Saeys et al., 2008; Yang, 2010*a*; Faritha Banu and Chandrasekar, 2013). Because, the algorithm uses the sigmoid function as fitness function such as in updating particle position ($x$) in the Binary Particle Swarm Optimization (BPSO), where it decreases the performance of this algorithm and trap it into local optima. These problems are the same in other binary heuristic algorithms such as Bat Algorithm (BA) that follows the principles of BPSO.

To push the performance of metaheuristic algorithms, this research explores ensemble approach in feature selection and classification. In ensemble approach, instead of executing individual feature selection algorithms, we combine various metaheuristic algorithms to improve robustness of feature selection model and classification performance. Our main hypothesis is that ensemble approach will overcome the disadvantages in an individual metaheuristic algorithm by balancing the number of features, decreasing the feature set dimensionality, and finally enhancing the classification performance. The literature has also shown several detection and filtering models that applied ensemble classifiers to detect spam emails such as by Wang et al. (2003), but to the best of our knowledge there is no work on ensemble feature subset selection for spam detection.

In effort to enhance the global search ability in the proposed metaheuristic algorithms as well as to increase the speed of evolutionary algorithms, this research also explores into merging the evolutionary computation and quantum computing. These algorithms are based on the principles in quantum mechanics such as qubit representation that have ability of processing huge numbers of quantum states. Unlike Quantum Computing (QC), Quantum Evolutionary Algorithm (QEA) does not work with a quantum machine. For example, Binary Quantum Particle Swarm Optimization (BQPSO) and Binary Quantum Gravitational Search Algorithm (BQGSA) are algorithms for solving optimization problems based on quantum computing rules and riding on BPSO and BGSA algorithms.

## 1.2 Problem Statement

Although data mining techniques have improved classification accuracy, because spammer constantly changes the pattern in emails. Feature selection techniques select a subset of relevant features within original features in order to improve classification performance. Mentioned that one of the familiar techniques that is able to decrease dimensionality is feature selection, where the subset of features is selected from the whole set of features in order to remove irrelevant features. A useful subset of features for a classifier may be useful for other classifiers in the same time. As the result, an individual technique selects a relevant subset of features, but possibly out of a set of irrelevant features (Gomez et al., 2012).

Nonetheless, although ensemble approach has provided an environment to overcome shortcomings of individual algorithms (Saeys et al., 2008; Valentini and Masulli, 2002; Attik, 2006), the performance of such approach needs to be improved by changing a number of parameters in classifiers or feature selection algorithms to increase classification accuracy. One example is Binary Particle Swarm Optimization (BPSO) algorithm that has been previously applied in solving optimization problems such as feature subset selection. New algorithms including the Bat Algorithm (BA) use the advantages of BPSO to improve optimization process. However, there are two main problems in BBA whereby the algorithm is often trapped the search into local optimum, hence causing overfitting.

The first problem in the BBA algorithm concerns the sigmoid function. Conceptually, the high value of bat speed towards a negative or positive value shows that the bat position should change for a more specific dimension. In the binary algorithm, the speed steer the bat position towards 0 or 1. Additionally, the velocity ($v$) near to 0 shows that the position of bat ($x$) is satisfied and the sigmoid function demonstrates an equal probability of 0 or 1 for bat position. The second problem in BBA concerns on means to update the bat position ($x$). In the average of initial iterations, all bats come up the optimal solution; however after several iterations these bats keep out the optimal solution. While the optimal solution is near to 0, but the probability of 0 or 1 decrease to 50% within this time (Yang, 2010$a$; Izui et al., 2008). Since accurate models use thousands of features, most of the detection model overfit the feature dataset.

The ensemble learning approach needs to consider streaming data in email spam detection system. Most of data mining techniques mine stream data from large amount of data with limited memory. These techniques scan training data severally, so their performance (accuracy) is unsuitable in the higher rate data environment (Wang et al., 2003; Fan, 2004). Other mining methods are incremental or online data stream methods that refine and modify new arrived data. These methods update the model trained costly (Hulten et al., 2001; Katakis et al., 2006). Many studies have mined data stream based on single model that show the whole data stream. Their techniques consumes time and space with low efficiency such as decision trees (Domingos and Hulten, 2000). Ensemble learning is a famous method for mining data stream and concept drift by using statistical-based weighted voting technique. However, discarding old data based on the time creates the problem of conflicting and overfitting concepts (Fan, 2004).

## 1.3 Research Objectives

The main objective of this research is to propose novel ensemble learning methods that consist of ensemble feature selection and ensemble classification based on metaheuristic algorithms to improve classification accuracy. To achieve the objective, the following tasks are to be undertaken:

- To propose a novel wrapper-based ensemble feature selection method based on three metaheuristic algorithms, which are Binary Gravitational Search Algorithm (BGSA), Quantum Binary Particle Swarm Optimization (QBPSO), and Modified Binary Bat Algortihm (MBBA). This method selects a set of relevant features to decrease dimensionality obtaining a better classification accuracy comparing individual feature selection methods.

- To propose MBBA and NBPSO algorithms to prevent overfitting and trapping algorithm in local optimum during feature selection process.

- To propose an ensemble feature selection approach based on New Binary Particle Swarm Optimization (NBPSO)using three parts of email (header, subject, body) in order to select relevant features. This technique proves a set of relevant features may be not suitable for different classifiers in the same time. Due to this, this research proposes an ensemble feature selection method to identify a relevance of features in various parts of email based on different partition of training data.

- To propose a new ensemble learning classifiers using Quantum Binary Gravitational Search Algorithm (QBGSA)using Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) to avoid conflicting and overfitting data problem in classification problem instead of discarding data based on arrival time. In mining data streams in order to detect concept drifts, decrease computational cost, the increase of accuracy and efficiency of learning algorithms, QBGSA selects relevant features after desired iteration instead of discard training data according arrival time.

## 1.4 Research Scope

The feature selection problem studied in this research is scoped to email spam detection system and covers both structural and content-based features from email such as the header, subject and body. This research focuses on feature selection in spam detection system based on ensemble feature selection methods using metaheuristic algorithms in order to decrease dimensionality and training data while at the same time improving the classifier accuracy and computational cost. -

## 1.5 Research Significance

The main contributions of this research is the spam detection framework for ensemble learning in feature selection and classification. Ensemble feature selection method concerns on selection of a set of relevant features in spam emails using metaheuristic algorithms such as BGSA, NBPSO, QBPSO and MBBA algorithms. Ensemble classification concerns on high prediction accuracy using combination of Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) based on Quantum Binary Gravitational Search Algorithm (QBGSA). The detailed contributions in ensemble learning are as follows:

- The ensemble feature selection method is based on three metaheuristic algorithms, which are Binary Gravitational Search Algorithm (BGSA), Quantum Binary Particle Swarm Optimization (QBPSO), and Modified Binary Bat Algorithm (MBBA). In this technique, the metaheuristic algorithms are improved to overcome the defects of individual basic algorithms. This proposed technique is able to achieve a subset of features based on three feature selection algorithms by aggregate their results for better classifier accuracy.

- Relevant and robust features are obtained by the proposed ensemble feature selection technique that they may not be achieved using individual feature selection methods. In the large feature space, there are many relevant feature subsets with equal efficiency. Some individual feature selection algorithms are trapped in local optimum such as BPSO and BBA finding the best solution or feature, thus useful feature subsets unable to reach in feature selection process. As the result, the ensemble feature selection technique is able to decrease the risk of choosing an irrelevant feature subset by aggregating the results from various feature selection algorithms.

- The NBPSO and the MBBA algorithms are proposed to select a set of relevant features in the ensemble feature selection technique based on new fitness function. We also update the new position of particles and bats in NBPSO and MBBA respectively. These modified algorithms prevent trapping algorithm in local optimum and over fitting for increase of ensemble feature selection efficiency and better classifier accuracy.

- A novel ensemble feature selection method in spam detection system based on three parts of spam email (header, subject and body) using NBPSO algorithm as a feature selector, to identify a relevancy of features in three different parts of email. The relevant selected features increase the classifier accuracy and improve computational cost.

- A weighted ensemble classifiers based on QBGSA algorithm is able to mine data stream and concept drifts instead of single model application. This algorithm is trained by different data chunks. One of the important points in streaming data is keeping the balance data in order to avoid conflicting and over fitting training data. Thus, classifier ensemble method in this research applies QBGSA to decrease or delete old data by decrease irrelevant features instead of data arrival time in spam detection system.

## 1.6 Thesis Organization

This thesis is organized in accordance to the standard structure of thesis dissertations for Universiti Putra Malaysia. The thesis is divided into seven chapters.

Chapter 1 – Introduction. This chapter introduces the background of the research. It defines the problem area and explains the objectives of the research.

Chapter 2 – Literature Review. This chapter reviews the related field of study and similar researches. It introduces spam emails and traditional filters as well as novel methods based on machine learning techniques that are available in detecting spam. Then it explains the feature selection methods and current methods that have been applied in spam detection. In addition, this chapter presents few studies that focus on ensemble methods in spam detection including data stream mining. This chapter also discusses the efficiency of feature selection algorithms using different classifier and ensemble methods.

Chapter 3 – Framework for Ensemble Learning. This chapter presents the methodology adopted for the current research and how it is conducted. The methodology is clarified by flowcharts and figures that give detailed information of the research process.

Chapter 4 – Ensemble Feature Subset Selection Method. This chapter proposes a new feature subset selection method based on metaheuristic algorithms as an ensemble feature selection method in spam detection. It also explains how to select a subset of relevant features from different parts of email such as header, subject, body and attached files are explained using MLP classifier.

Chapter 5 – Ensemble Classification. This fifth chapter discusses the ensemble classifiers combined with QBGSA as a hybrid system. This chapter explains ensemble classifiers how prevent overfitting and conflicting in data stream classification and detect concept drifts using QBGSA feature selector to improve classifier accuracy.

Chapter 6 – Results and Discussions. In this chapter, a comprehensive experimental study is presented based on various experiments based on metaheuristic algorithms as feature selection methods and three methods of ensemble feature selection. At first, the experiments show the performance of metaheuristic algorithms and the second phase, the role of ensemble feature selection methods in the spam detection system in terms of classifier accuracy and computational cost is discussed. All the experimental results are obtained by charts and graphs.

Chapter 7 – Conclusion and Recommendations. This chapter concludes the research findings and introduces some suggestions for future work.

## REFERENCES


Abu-Nimeh, S., Nappa, D., Wang, X. and Nair, S. (2008), Bayesian additive regression trees-based spam detection for enhanced email privacy, *in* 'Availability, Reliability and Security, 2008. ARES 08. Third International Conference on', IEEE, pp. 1044–1051.

Aggarwal, C. C. (2012), Mining text streams, *in* 'Mining Text Data', Springer, pp. 297–321.

Aggarwal, C. C., Hinneburg, A. and Keim, D. A. (2001), *On the surprising behavior of distance metrics in high dimensional space*, Springer.

Aggarwal, C. C. and Zhai, C. (2012), A survey of text classification algorithms, *in* 'Mining text data', Springer, pp. 163–222.

Aha, D. W., Kibler, D. and Albert, M. K. (1991), 'Instance-based learning algorithms', *Machine learning* **6**(1), 37–66.

Alshawabkeh, M. (2013), 'Hypothesis margin based weighting for feature selection using boosting: theory, algorithms and applications'.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G. and Spyropoulos, C. D. (2000), 'An evaluation of naive bayesian anti-spam filtering', *arXiv preprint cs/0006013* .

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V. and Spyropoulos, C. D. (2000), An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages, *in* 'Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 160–167.

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D. and Stamatopoulos, P. (2000), 'Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach', *arXiv preprint cs/0009009* .

Attik, M. (2006), Using ensemble feature selection approach in selecting subset with relevant features, *in* 'Advances in Neural Networks-ISNN 2006', Springer, pp. 1359–1366.

Barber, M. J. (2012), Classification ensemble methods for mitigating concept drift within online data streams, PhD thesis, Colorado State University.

Batista, E. (2000), 'A fight to ban cellphone spam', *WiredNews, July* **6**.

Biggio, B., Fumera, G., Pillai, I. and Roli, F. (2007), Image spam filtering by content obscuring detection., *in* 'CEAS'.

Blanzieri, E. and Bryl, A. (2008), 'A survey of learning-based techniques of email spam filtering', *Artificial Intelligence Review* **29**(1), 63–92.

Blum, C. and Roli, A. (2003), 'Metaheuristics in combinatorial optimization: Overview and conceptual comparison', *ACM Computing Surveys (CSUR)* **35**(3), 268–308.

Byun, B., Lee, C.-H., Webb, S. and Pu, C. (2007), A discriminative classifier learning approach to image modeling and spam image identification., *in* 'CEAS', Citeseer.

Carpinteiro, O. A., Lima, I., Assis, J. M., de Souza, A. C. Z., Moreira, E. M. and Pinheiro, C. A. (2006), A neural model in anti-spam systems, *in* 'Artificial Neural Networks–ICANN 2006', Springer, pp. 847–855.

Carpinter, J. and Hunt, R. (2006), 'Tightening the net: A review of current and next generation spam filtering tools', *Computers & security* **25**(8), 566–578.

Carpinter, J. M. (2005), Evaluating ensemble classifiers for spam filtering, Technical report, Technical Report, University of Canterbury.

Carreras, X. and Marquez, L. (2001), 'Boosting trees for anti-spam email filtering', *arXiv preprint cs/0109015* .

Chang, M. and Poon, C. K. (2009), 'Using phrases as features in email classification', *Journal of Systems and Software* **82**(6), 1036–1045.

Chen, J. and Guo, C. (2006), Online detection and prevention of phishing attacks, *in* 'Communications and Networking in China, 2006. ChinaCom'06. First International Conference on', IEEE, pp. 1–7.

Chharia, A. and Gupta, R. (2013), Email classifier: An ensemble using probability and rules, *in* 'Contemporary Computing (IC3), 2013 Sixth International Conference on', IEEE, pp. 130–136.

Chuan, Z., Xianliang, L., Mengshu, H. and Xu, Z. (2005), 'A lvq-based neural network anti-spam email approach', *ACM SIGOPS Operating Systems Review* **39**(1), 34–39.

Cohen, W. W. (1996), Learning rules that classify e-mail, *in* 'AAAI Spring Symposium on Machine Learning in Information Access', Vol. 18, California, p. 25.

Cormack, G. V. and Lynam, T. R. (2005), Spam corpus creation for trec., *in* 'CEAS'.

Črepinšek, M., Liu, S.-H. and Mernik, M. (2013), 'Exploration and exploitation in evolutionary algorithms: a survey', *ACM Computing Surveys (CSUR)* **45**(3), 35.

DeBarr, D. and Wechsler, H. (2012), 'Spam detection using random boost', *Pattern Recognition Letters* **33**(10), 1237–1244.

Del Valle, Y., Venayagamoorthy, G. K., Mohagheghi, S., Hernandez, J.-C. and Harley, R. G. (2008), 'Particle swarm optimization: basic concepts, variants and applications in power systems', *Evolutionary Computation, IEEE Transactions on* **12**(2), 171–195.

Delany, S. J., Cunningham, P., Tsymbal, A. and Coyle, L. (2005), 'A case-based technique for tracking concept drift in spam filtering', *Knowledge-Based Systems* **18**(4), 187–195.

Domingos, P. and Hulten, G. (2000), Mining high-speed data streams, *in* 'Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 71–80.

Duda, R. O., Hart, P. E. and Stork, D. G. (1999), *Pattern classification*, John Wiley & Sons,.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000), *Pattern classification*, John Wiley & Sons,.

Dunne, K., Cunningham, P. and Azuaje, F. (2002), 'Solutions to instability problems with sequential wrapper-based approaches to feature selection', *Journal of Machine Learning Research* .

El-Alfy, E.-S. (2009), Discovering classification rules for email spam filtering with an ant colony optimization algorithm, *in* 'Evolutionary Computation, 2009. CEC'09. IEEE Congress on', IEEE, pp. 1778–1783.

El-Alfy, E.-S. M. and Abdel-Aal, R. E. (2011), 'Using gmdh-based networks for improved spam detection and email feature analysis', *Applied Soft Computing* **11**(1), 477–488.

Engelbrecht, A. P. (2006), *Fundamentals of computational swarm intelligence*, John Wiley & Sons.

Fan, W. (2004), Streamminer: A classifier ensemble-based engine to mine concept-drifting data streams, *in* 'Proceedings of the Thirtieth international conference on Very large data bases-Volume 30', VLDB Endowment, pp. 1257–1260.

Farid, D. M., Zhang, L., Hossain, A., Rahman, C. M., Strachan, R., Sexton, G. and Dahal, K. (2013), 'An adaptive ensemble classifier for mining concept drifting data streams', *Expert Systems with Applications* **40**(15), 5895–5906.

Faritha Banu, A. and Chandrasekar, C. (2013), 'An optimized approach of modified bat algorithm to record deduplication.', *International Journal of Computer Applications* **62**.

Fawcett, T. (2003), 'In vivo spam filtering: a challenge problem for kdd', *ACM SIGKDD Explorations Newsletter* **5**(2), 140–148.

Fawcett, T. (2006), 'An introduction to roc analysis', *Pattern recognition letters* **27**(8), 861–874.

Fern, A. and Givan, R. (2003), 'Online ensemble learning: An empirical study', *Machine Learning* **53**(1-2), 71–109.

Ferreira, A. J. and Figueiredo, M. A. (2012), 'An unsupervised approach to feature discretization and selection', *Pattern Recognition* **45**(9), 3048–3060.

Fukunaga, K. (1990), *Introduction to statistical pattern recognition*, Academic press.

Fumera, G., Pillai, I. and Roli, F. (2006), 'Spam filtering based on the analysis of text information embedded into images', *The Journal of Machine Learning Research* **7**, 2699–2720.

Gaber, M. M., Zaslavsky, A. and Krishnaswamy, S. (2007), A survey of classification methods in data streams, *in* 'Data Streams', Springer, pp. 39–59.

Gao, H. and Diao, M. (2009), Quantum particle swarm optimization for mc-cdma multiuser detection, *in* 'Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on', Vol. 2, IEEE, pp. 132–136.

Gazi, M. A. (2010), 'Increase reliability for skin detector using backprobgation neural network and heuristic rules based on ycbcr', *Scientific Research and Essays* **5**(19), 2931–2946.

Gomez, J. C., Boiy, E. and Moens, M.-F. (2012), 'Highly discriminative statistical features for email classification', *Knowledge and information systems* **31**(1), 23–53.

Gütlein, M., Frank, E., Hall, M. and Karwath, A. (2009), Large-scale attribute selection using wrappers, *in* 'Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on', IEEE, pp. 332–339.

Guyon, I. and Elisseeff, A. (2003), 'An introduction to variable and feature selection', *The Journal of Machine Learning Research* **3**, 1157–1182.

Han, J. and Kamber, M. (2006), *Data Mining, Southeast Asia Edition: Concepts and Techniques*, Morgan kaufmann.

Han, K.-H. and Kim, J.-H. (2000), Genetic quantum algorithm and its application to combinatorial optimization problem, *in* 'Evolutionary Computation, 2000. Proceedings of the 2000 Congress on', Vol. 2, IEEE, pp. 1354–1360.

Han, K.-H. and Kim, J.-H. (2002), 'Quantum-inspired evolutionary algorithm for a class of combinatorial optimization', *Evolutionary Computation, IEEE Transactions on* **6**(6), 580–593.

Han, X., Quan, L., Xiong, X. and Wu, B. (2013), 'Facing the classification of binary problems with a hybrid system based on quantum-inspired binary gravitational search algorithm and k-nn method', *Engineering Applications of Artificial Intelligence* **26**(10), 2424–2430.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J. and Tibshirani, R. (2009), *The elements of statistical learning*, Vol. 2, Springer.

Hua Li, C. and Xiangji Huang, J. (2012), 'Spam filtering using semantic similarity approach and adaptive bpnn', *Neurocomputing* **92**, 88–97.

Huang, C.-L. and Dun, J.-F. (2008), 'A distributed pso–svm hybrid system with feature selection and parameter optimization', *Applied Soft Computing* **8**(4), 1381–1391.

Huang, C.-L. and Wang, C.-J. (2006), 'A ga-based feature selection and parameters optimizationfor support vector machines', *Expert Systems with applications* **31**(2), 231–240.

Hulten, G., Spencer, L. and Domingos, P. (2001), Mining time-changing data streams, *in* 'Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 97–106.

Ibrahim, A. A., Mohamed, A. and Shareef, H. (2012), Application of quantum-inspired binary gravitational search algorithm for optimal power quality monitor placement, *in* 'Proceedings of the 11th WSEAS international conference on artificial intelligence, knowledge engineering and data bases (AIKED âĂŸ12)', pp. 22–24.

Islam, M. R., Chowdhury, M. U. and Zhou, W. (2005), An innovative spam filtering model based on support vector machine, *in* 'Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on', Vol. 2, IEEE, pp. 348–353.

Islam, R. and Abawajy, J. (2013), 'A multi-tier phishing detection and filtering approach', *Journal of Network and Computer Applications* **36**(1), 324–335.

Izui, K., Nishiwaki, S., Yoshimura, M., Nakamura, M. and Renaud, J. E. (2008), 'Enhanced multiobjective particle swarm optimization in combination with adaptive weighted gradient-based searching', *Engineering Optimization* **40**(9), 789–804.

Jeong, Y.-W., Park, J.-B., Jang, S.-H. and Lee, K. Y. (2010), 'A new quantum-inspired binary pso: application to unit commitment problems for power systems', *Power Systems, IEEE Transactions on* **25**(3), 1486–1495.

Jindal, N. and Liu, B. (2007), Analyzing and detecting review spam, *in* 'Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on', IEEE, pp. 547–552.

Kanaris, I., Kanaris, K. and Stamatatos, E. (2006), Spam detection using character n-grams, *in* 'Advances in Artificial Intelligence', Springer, pp. 95–104.

Katakis, I., Tsoumakas, G. and Vlahavas, I. (2006), 'Dynamic feature space and incremental feature selection for the classification of textual data streams', *Knowledge Discovery from Data Streams* pp. 107–116.

Katakis, I., Tsoumakas, G. and Vlahavas, I. (2010), 'Tracking recurring contexts using ensemble classifiers: an application to email filtering', *Knowledge and Information Systems* **22**(3), 371–391.

Kennedy, J. (2010), Particle swarm optimization, *in* 'Encyclopedia of Machine Learning', Springer, pp. 760–766.

Klinkenberg, R. (2004), 'Learning drifting concepts: Example selection vs. example weighting', *Intelligent Data Analysis* **8**(3), 281–300.

Koprinska, I., Poon, J., Clark, J. and Chan, J. (2007), 'Learning to classify e-mail', *Information Sciences* **177**(10), 2167–2187.

Lai, C.-C. and Wu, C.-H. (2007), 'Particle swarm optimization-aided feature selection for spam email classification', *IEEE, Kumamoto* p. 165.

Laorden, C., Sanz, B., Santos, I., Galán-García, P. and Bringas, P. G. (2013), 'Collective classification for spam filtering', *Logic Journal of IGPL* **21**(4), 540–548.

Lee, S. M., Kim, D. S., Kim, J. H. and Park, J. S. (2010), Spam detection using feature selection and parameters optimization, *in* 'Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on', IEEE, pp. 883–888.

Lee, S. M., Kim, D. S. and Park, J. S. (2011), 'Cost-sensitive spam detection using parameters optimization and feature selection.', *J. UCS* **17**(6), 944–960.

Liu, H. and Motoda, H. (1998), *Feature extraction, construction and selection: A data mining perspective*, Springer.

Lopes, C., Cortez, P., Sousa, P., Rocha, M. and Rio, M. (2011), 'Symbiotic filtering for spam email detection', *Expert Systems with Applications* **38**(8), 9365–9372.

Luke, S. (2009), *Essentials of metaheuristics*, Vol. 3, Lulu Raleigh.

Luo, X. and Zincir-Heywood, N. (2005), Comparison of a som based sequence analysis system and naive bayesian classifier for spam filtering, *in* 'Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on', Vol. 4, IEEE, pp. 2571–2576.

Luo, Y. and Xiong, S. (2009), Ensemble method for unsupervised feature selection, *in* 'Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on', Vol. 4, IEEE, pp. 513–516.

Maciá-Pérez, F., Mora-Gimeno, F., Marcos-Jorquera, D., Gil-Martínez-Abarca, J. A., Ramos-Morillo, H. and Lorenzo-Fonseca, I. (2011), 'Network intrusion detection system embedded on a smart sensor', *Industrial Electronics, IEEE Transactions on* **58**(3), 722–732.

Manjusha, K., Rahul, B. and Shahabas, S. (2013), 'An efficient method of spam classification by multiclass support vector machine classifier'.

Méndez, J. R., Fdez-Riverola, F., Iglesias, E. L., Díaz, F. and Corchado, J. M. (2006), Tracking concept drift at feature selection stage in spamhunting: An anti-spam instance-based reasoning system, *in* 'Advances in Case-Based Reasoning', Springer, pp. 504–518.

Michalak, K. and Kwasnicka, H. (2006), Correlation-based feature selection strategy in neural classification, *in* 'Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on', Vol. 1, IEEE, pp. 741–746.

Minku, L. L. and Yao, X. (2012), 'Ddd: A new ensemble approach for dealing with concept drift', *Knowledge and Data Engineering, IEEE Transactions on* **24**(4), 619–633.

Mohammad, A. H. and Zitar, R. A. (2011), 'Application of genetic optimized artificial immune system and neural networks in spam detection', *applied soft computing* **11**(4), 3827–3845.

Mukkamala, S. and Sung, A. H. (2005), Significant feature selection using computational intelligent techniques for intrusion detection, *in* 'Advanced Methods for Knowledge Discovery from Complex Data', Springer, pp. 285–306.

Nakamura, R. Y., Pereira, L., Costa, K., Rodrigues, D., Papa, J. P. and Yang, X.-S. (2012), Bba: A binary bat algorithm for feature selection, *in* 'Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on', IEEE, pp. 291–297.

Narayanan, A. and Moore, M. (1996), Quantum-inspired genetic algorithms, *in* 'Evolutionary Computation, 1996., Proceedings of IEEE International Conference on', IEEE, pp. 61–66.

Okun, O. and Global, I. (2011), *Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations*, Medical Information Science Reference.

Oliveira, A. L., Braga, P. L., Lima, R. M. and Cornélio, M. L. (2010), 'Ga-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation', *information and Software Technology* **52**(11), 1155–1166.

Olivo, C. K., Santin, A. O. and Oliveira, L. S. (2011), 'Obtaining the threat model for e-mail phishing', *Applied Soft Computing* .

Patidar, V., Singh, D. and Singh, A. (2013), 'Article: A novel technique of email classification for spam detection', *International Journal of Applied Information Systems* **5**(10), 15–19. Published by Foundation of Computer Science, New York, USA.

Payne, T. R. and Edwards, P. (1997), 'Interface agents that learn an investigation of learning issues in a mail agent interface', *Applied artificial intelligence* **11**(1), 1–32.

Pazoki, A. and Pazoki, Z. (2013), 'Classification system for rain fed wheat grain cultivars using artificial neural network', *African Journal of Biotechnology* **10**(41), 8031–8038.

Peng, H., Long, F. and Ding, C. (2005), 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(8), 1226–1238.

Puniškis, D., Laurutis, R. and Dirmeikis, R. (2006), 'An artificial neural nets for spam e–mail recognition', *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)* **5**(69), 73–76.

Ramadan, R. M. and Abdel-Kader, R. F. (2009), 'Face recognition using particle swarm optimization-based selected features', *International Journal of Signal Processing, Image Processing and Pattern Recognition* **2**(2), 51–65.

Rashedi, E., Nezamabadi-Pour, H. and Saryazdi, S. (2009), 'Gsa: a gravitational search algorithm', *Information sciences* **179**(13), 2232–2248.

Rashedi, E., Nezamabadi-Pour, H. and Saryazdi, S. (2010), 'Bgsa: binary gravitational search algorithm', *Natural Computing* **9**(3), 727–745.

Ripley, B. D. (1996), *Pattern recognition and neural networks*, Cambridge university press.

Ruan, G. and Tan, Y. (2007), Intelligent detection approaches for spam, *in* 'Natural Computation, 2007. ICNC 2007. Third International Conference on', Vol. 3, IEEE, pp. 672–676.

Ruan, G. and Tan, Y. (2010), 'A three-layer back-propagation neural network for spam detection using artificial immune concentration', *Soft Computing* **14**(2), 139–150.

Saeys, Y., Abeel, T. and Van de Peer, Y. (2008), Robust feature selection using ensemble feature selection techniques, *in* 'Machine learning and knowledge discovery in databases', Springer, pp. 313–325.

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. (1998), A bayesian approach to filtering junk e-mail, *in* 'Learning for Text Categorization: Papers from the 1998 workshop', Vol. 62, pp. 98–105.

Sarafrazi, S. and Nezamabadi-pour, H. (2013), 'Facing the classification of binary problems with a gsa-svm hybrid system', *Mathematical and Computer Modelling* **57**(1), 270–278.

Sebastiani, F. (2002), 'Machine learning in automated text categorization', *ACM computing surveys (CSUR)* **34**(1), 1–47.

Shen, Q., Diao, R. and Su, P. (2012), 'Feature selection ensemble.', *Turing-100* **10**, 289–306.

Sirisanyalak, B. and Sornil, O. (2007), Artificial immunity-based feature extraction for spam detection, *in* 'Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on', Vol. 3, IEEE, pp. 359–364.

Soranamageswari, M. and Meena, C. (2010), An efficient feature extraction method for classification of image spam using artificial neural networks, *in* 'Data Storage and Data Engineering (DSDE), 2010 International Conference on', IEEE, pp. 169–172.

Steinbach, M. and Tan, P.-N. (2009), 'knn: k-nearest neighbors', *The Top Ten Algorithms in Data Mining* pp. 151–162.

Stuart, I., Cha, S.-H. and Tappert, C. (2004), A neural network classifier for junk e-mail, *in* 'Document Analysis Systems VI', Springer, pp. 442–450.

Tak, G. K. and Tapaswi, S. (2010), Knowledge base compound approach towards spam detection, *in* 'Recent Trends in Network Security and Applications', Springer, pp. 490–499.

Tan, Y. and Ruan, G. (2014), 'Uninterrupted approaches for spam detection based on svm and ais', *International Journal of Computational Intelligence* **1**(1), 1–26.

Tretyakov, K. (2004), Machine learning techniques in spam filtering, *in* 'Data Mining Problem-oriented Seminar, MTAT', Vol. 3, pp. 60–79.

Tsymbal, A., Puuronen, S. and Patterson, D. W. (2003), 'Ensemble feature selection with the simple bayesian classification', *Information Fusion* **4**(2), 87–100.

Tu, C.-J., Chuang, L.-Y., Chang, J.-Y., Yang, C.-H. et al. (2007), 'Feature selection using pso-svm', *IAENG International journal of computer science* **33**(1), 111–116.

Unler, A. and Murat, A. (2010), 'A discrete particle swarm optimization method for feature selection in binary classification problems', *European Journal of Operational Research* **206**(3), 528–539.

Unler, A., Murat, A. and Chinnam, R. B. (2011), 'Pso: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification', *Information Sciences* **181**(20), 4625–4641.

Vafaie, H. and De Jong, K. (1992), Genetic algorithms as a tool for feature selection in machine learning, *in* 'Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., Fourth International Conference on', IEEE, pp. 200–203.

Valentini, G. and Masulli, F. (2002), Ensembles of learning machines, *in* 'Neural Nets', Springer, pp. 3–20.

Vieira, S. M., Sousa, J. and Kaymak, U. (2012), 'Fuzzy criteria for feature selection', *Fuzzy Sets and Systems* **189**(1), 1–18.

Wang, H.-b., Yu, Y. and Liu, Z. (2005), Svm classifier incorporating feature selection using ga for spam detection, *in* 'Embedded and Ubiquitous Computing–EUC 2005', Springer, pp. 1147–1154.

Wang, H., Fan, W., Yu, P. S. and Han, J. (2003), Mining concept-drifting data streams using ensemble classifiers, *in* 'Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 226–235.

Wang, H., Khoshgoftaar, T. M. and Napolitano, A. (2010), A comparative study of ensemble feature selection techniques for software defect prediction, *in* 'Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on', IEEE, pp. 135–140.

Wang, H., Philip, S. Y. and Han, J. (2010), Mining concept-drifting data streams, *in* 'Data Mining and Knowledge Discovery Handbook', Springer, pp. 789–802.

Wang, R., Youssef, A. M. and Elhakeem, A. K. (2006), On some feature selection strategies for spam filter design, *in* 'Electrical and Computer Engineering, 2006. CCECE'06. Canadian Conference on', IEEE, pp. 2186–2189.

Wang, W. (2008), Some fundamental issues in ensemble methods, *in* 'Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on', IEEE, pp. 2243–2250.

Wang, W. (2010), Heterogeneous bayesian ensembles for classifying spam emails, *in* 'Neural Networks (IJCNN), The 2010 International Joint Conference on', IEEE, pp. 1–8.

Wang, W., Zeng, G. and Tang, D. (2010), 'Using evidence based content trust model for spam detection', *Expert Systems with Applications* **37**(8), 5599–5606.

Wu, C.-H. and Tsai, C.-H. (2009), 'Robust classification for spam filtering by back-propagation neural networks using behavior-based features', *Applied Intelligence* **31**(2), 107–121.

Wu, C.-T., Cheng, K.-T., Zhu, Q. and Wu, Y.-L. (2005), Using visual features for anti-spam filtering, *in* 'Image Processing, 2005. ICIP 2005. IEEE International Conference on', Vol. 3, IEEE, pp. III–509.

Wu, Q., Wu, S. and Liu, J. (2010), 'Hybrid model based on svm with gaussian loss function and adaptive gaussian pso', *Engineering Applications of Artificial Intelligence* **23**(4), 487–494.

Xue, B. (2014), 'Particle swarm optimisation for feature selection in classification'.

Yan, K., Hongwei, Y. and Jiayin, F. (2010), Design of optimal stack filters using qpso, *in* 'Information Science and Engineering (ICISE), 2010 2nd International Conference on', IEEE, pp. 3639–3643.

Yang, J.-a., Li, B. and Zhuang, Z. (2003), Multi-universe parallel quantum genetic algorithm its application to blind-source separation, *in* 'Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on', Vol. 1, IEEE, pp. 393–398.

Yang, P., Liu, W., Zhou, B. B., Chawla, S. and Zomaya, A. Y. (2013), Ensemble-based wrapper methods for feature selection and class imbalance learning, *in* 'Advances in Knowledge Discovery and Data Mining', Springer, pp. 544–555.

Yang, X.-S. (2010*a*), *Nature-inspired metaheuristic algorithms*, Luniver press.

Yang, X.-S. (2010*b*), A new metaheuristic bat-inspired algorithm, *in* 'Nature inspired cooperative strategies for optimization (NICSO 2010)', Springer, pp. 65–74.

Yang, X.-S. (2011), Metaheuristic optimization: algorithm analysis and open problems, *in* 'Experimental Algorithms', Springer, pp. 21–32.

Yang, Y. and Pedersen, J. O. (1997), A comparative study on feature selection in text categorization, *in* 'ICML', Vol. 97, pp. 412–420.

Yang, Z., Nie, X., Xu, W. and Guo, J. (2006), An approach to spam detection by naive bayes ensemble based on decision induction, *in* 'Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on', Vol. 2, IEEE, pp. 861–866.

Yeh, C.-Y., Wu, C.-H. and Doong, S.-H. (2005), Effective spam classification based on meta-heuristics, *in* 'Systems, Man and Cybernetics, 2005 IEEE International Conference on', Vol. 4, IEEE, pp. 3872–3877.

Ying, K.-C., Lin, S.-W., Lee, Z.-J. and Lin, Y.-T. (2010), 'An ensemble approach applied to classify spam e-mails', *Expert Systems with Applications* **37**(3), 2197–2201.

Yourui, H., Chaoli, T. and Shuang, W. (2007), Quantum-inspired swarm evolution algorithm, *in* '2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)', pp. 208–211.

Yu, L. and Liu, H. (2003), Feature selection for high-dimensional data: A fast correlation-based filter solution, *in* 'ICML', Vol. 3, pp. 856–863.

Zahran, B. M. and Kanaan, G. (2009), 'Text feature selection using particle swarm optimization algorithm 1'.

Zhang, G. (2011), 'Quantum-inspired evolutionary algorithms: a survey and empirical study', *Journal of Heuristics* **17**(3), 303–351.

Zhang, L., Zhu, J. and Yao, T. (2004), 'An evaluation of statistical spam filtering techniques', *ACM Transactions on Asian Language Information Processing (TALIP)* **3**(4), 243–269.

Zhang, Y., Li, H., Niranjan, M. and Rockett, P. (2008), Applying cost-sensitive multiobjective genetic programming to feature extraction for spam e-mail filtering, *in* 'Genetic Programming', Springer, pp. 325–336.

Zhang, Y., Wang, S., Phillips, P. and Ji, G. (2014), 'Binary pso with mutation operator for feature selection using decision tree applied to spam detection', *Knowledge-Based Systems* **64**, 22–31.

Zhao, M., Fu, C., Ji, L., Tang, K. and Zhou, M. (2011), 'Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes', *Expert Systems with Applications* **38**(5), 5197–5204.

Zhao, W. and Zhu, Y. (2006), Classifying email using variable precision rough set approach, *in* 'Rough Sets and Knowledge Technology', Springer, pp. 766–771.

Zhou, S. and Sun, Z. (2005), A new approach belonging to edas: quantum-inspired genetic algorithm with only one chromosome, *in* 'Advances in Natural Computation', Springer, pp. 141–150.

Zhou, Y., Mulekar, M. S. and Nerellapalli, P. (2007), 'Adaptive spam filtering using dynamic feature spaces', *International Journal on Artificial Intelligence Tools* **16**(04), 627–646.

Zhuang, X., Yang, G. and Zhu, H. (2008), A model of image feature extraction inspired by ant swarm system, *in* 'Natural Computation, 2008. ICNC'08. Fourth International Conference on', Vol. 7, IEEE, pp. 553–557.

Zhuo, L., Zheng, J., Li, X., Wang, F., Ai, B. and Qian, J. (2008), A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine, *in* 'Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images', International Society for Optics and Photonics, pp. 71471J–71471J.

Zmyślony, M., Krawczyk, B. and Woźniak, M. (2013), Combined classifiers with neural fuser for spam detection, *in* 'International Joint Conference CISISâĂŹ12-ICEUTE´ 12-SOCO´ 12 Special Sessions', Springer, pp. 245–252.

Zorkadis, V., Panayotou, M. and Karras, D. A. (2005), Improved spam e-mail filtering based on committee machines and information theoretic feature extraction, *in* 'Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on', Vol. 1, IEEE, pp. 179–184.