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
COPYRIGHT

All material contained within the thesis, including without limitation text, logos, icons, photographs and all other artwork, is copyright material of Universiti Putra Malaysia unless otherwise stated. Use may be made of any material contained within the thesis for non-commercial purposes from the copyright holder. Commercial uses of material may only be made with the express, prior, written permission of Universiti Putra Malaysia.

Copyright © Universiti Putra Malaysia
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 Gravita-
tional 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 dikenalkan 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


Tujuan perlombongan data adalah untuk mencari dan mendapatkan corak yang belum ditentukan di dalam pangkalan data yang besar. Satu tugas yang terkenal adalah pengelasan yang meramal klas 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-pemilihan ciri gabungan termasuk Modified Binary Bat Algoritma (NBBA), Binary Kuantum Zarah Swarm Optimization (QBPSO) Algoritma dan Binary Graviti Kuantum Cari algoritma (QBGSA) adalah hibrid dengan Multi-lapisan Perceptron (MLP) pengelas untuk memilih subset ciri yang berkaitan dan meningkatkan ketepatan pengelas. 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 kemutangkusan, 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.
I certify that a Thesis Examination Committee has met on 5 June 2015 to conduct the final examination of Amir Rajabi Behjat on his thesis entitled "An Ensemble Learning Method for Spam Email Detection System Based on Metaheuristic Algorithms" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Ali bin Mamat, PhD  
associate professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Chairman)

Razali bin Yaakob, PhD  
Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Internal Examiner)

Masrah Azrifah binti Azmi Murad, PhD  
Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Internal Examiner)

Xin-She Yang, PhD  
Professor  
Middlesex University  
United Kingdom  
(External Examiner)

ZULKARNAIN ZAINAL, PhD  
Professor and Deputy Dean  
School of Graduate Studies  
Universiti Putra Malaysia

Date: 17 June 2015
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.: ________________________________
Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature:
Name of Chairman of Supervisory Committee:

Signature:
Name of Member of Supervisory Committee:

Signature:
Name of Member of Supervisory Committee:

Signature:
Name of Member of Supervisory Committee:
TABLE OF CONTENTS

ABSTRACT i
ABSTRAK iii
ACKNOWLEDGEMENTS v
APPROVAL vi
DECLARATION viii
LIST OF TABLES xiii
LIST OF FIGURES xiv
LIST OF ABBREVIATIONS xvi

CHAPTER

1 INTRODUCTION 1
1.1 Motivation 1
1.2 Problem Statement 3
1.3 Research Objectives 4
1.4 Research Scope 5
1.5 Research Significance 5
1.6 Thesis Organization 6

2 LITERATURE REVIEW 8
2.1 Introduction 8
2.2 The History of Spam Email 8
2.3 Spam Email Detection Systems 9
2.4 Data Classification 11
2.5 Feature Subset Selection 12
  2.5.1 Problems in Feature Subset Selection 14
  2.5.2 Feature Subset Selection Using Metaheuristic Algorithm in Spam Detection System 16
2.6 Ensemble learning techniques and Ensemble Feature Selection 21
  2.6.1 Ensemble Feature Selection 21
  2.6.2 Classifier Ensemble Learning in Spam Detection System 24
2.7 Summary 26
3 RESEARCH METHODOLOGY

3.1 Introduction 27
3.2 An Overview of the Problem 27
3.3 Research Steps 29
3.4 Experimental Design 34
  3.4.1 Benchmark Datasets 34
  3.4.2 Evaluation Metrics 35
3.5 Summary 37

4 MODIFIED BINARY BAT ALGORITHM AND ENSEMBLE FEATURE SUBSET SELECTION METHOD BASED ON METAHEURISTIC ALGORITHMS 38

4.1 Introduction 38
4.2 Metaheuristic Algorithms for Feature Subset Selection 38
  4.2.1 Binary Genetic Algorithm (BGA) 39
  4.2.2 Gravitational Search Algorithm (GSA) 40
  4.2.3 Binary Particle Swarm Optimization (BPSO) 42
  4.2.4 New Binary Particle Swarm Optimization (NBPSO) 44
  4.2.5 Binary Bat Algorithm (BBA) 45
  4.2.6 Modified Binary Bat Algorithm (MBBA) 49
  4.2.7 Quantum-Inspired Evolutionary Algorithm (QIEA) 50
4.3 Applied Classifiers in a Spam Email Classification System 56
  4.3.1 K-Nearest Neighbor Classifier 56
  4.3.2 Multi-Layer Perceptron (MLP) 58
  4.3.3 Support Vector Machines 59
4.4 Summary 60

5 ENSEMBLE FEATURE SELECTION 61

5.1 Introduction 61
5.2 Feature Subset Selection based on Metaheuristic Algorithms in Spam Detection System 61
5.3 Ensemble Learning Methods based on Feature Subset Selection in Spam Detection System 62
  5.3.1 Ensemble Learning Method based on Three Parts of Email 63
  5.3.2 Ensemble Learning Method based on Metaheuristic Algorithms 64
  5.3.3 Classifier Ensemble Learning using QBGSA Algorithm 65
5.4 Summary 69
6 RESULTS AND DISCUSSION

6.1 Introduction 71
6.2 Experimental Results of Feature Subset Selection based on Meta Heuristic Algorithms 71
  6.2.1 Experimental Environment 71
  6.2.2 Analysis on Classifier Performance based on Accuracy, Recall and Precision Measurement Methods 72
  6.2.3 Analysis on a Set of Selected Features, CPU Time and Elapsed time 77
6.3 Comparison of Feature Selection Algorithms 79
6.4 Experiment Results of Ensemble Methods in Spam Detection System 80
  6.4.1 Performance Evaluation for Ensemble Feature Selection 81
  6.4.2 Performance Evaluation for Ensemble Feature Selection based on Data Partitioning 83
  6.4.3 Performance Evaluation for Data Stream Classification based on Ensemble Learning Classifier using QBGSA 85
6.5 Summary 87

7 CONCLUSION

7.1 Introduction 89
7.2 Recommendations for Future Work 90

REFERENCES 92
BIODATA OF STUDENT 104
LIST OF PUBLICATIONS 106
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Various Datasets used in the Experimental Evaluation</td>
<td>34</td>
</tr>
<tr>
<td>6.1 Accuracy, Precision, Recall for LingSpam</td>
<td>72</td>
</tr>
<tr>
<td>6.2 Accuracy, Precision, Recall for SpamAssassin</td>
<td>72</td>
</tr>
<tr>
<td>6.3 Selected Features by Metaheuristic Algorithms based on LingSpam, SpamAssassin, Spambase and CSDMC2010</td>
<td>75</td>
</tr>
<tr>
<td>6.4 CPU Time Consumed by Metaheuristic Algorithms based on the Selected Features</td>
<td>77</td>
</tr>
<tr>
<td>6.5 Elapsed Time Consumed by Metaheuristic Algorithms based on the Selected Features</td>
<td>78</td>
</tr>
<tr>
<td>6.6 Comparison of Feature Selection Algorithms</td>
<td>80</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>29</td>
</tr>
<tr>
<td>3.2</td>
<td>32</td>
</tr>
<tr>
<td>3.3</td>
<td>36</td>
</tr>
<tr>
<td>4.1</td>
<td>47</td>
</tr>
<tr>
<td>4.2</td>
<td>50</td>
</tr>
<tr>
<td>4.3</td>
<td>51</td>
</tr>
<tr>
<td>5.1</td>
<td>64</td>
</tr>
<tr>
<td>5.2</td>
<td>66</td>
</tr>
<tr>
<td>5.3</td>
<td>70</td>
</tr>
<tr>
<td>6.1</td>
<td>76</td>
</tr>
<tr>
<td>6.2</td>
<td>76</td>
</tr>
<tr>
<td>6.3</td>
<td>77</td>
</tr>
<tr>
<td>6.4</td>
<td>78</td>
</tr>
<tr>
<td>6.5</td>
<td>82</td>
</tr>
<tr>
<td>6.6</td>
<td>82</td>
</tr>
<tr>
<td>6.7</td>
<td>83</td>
</tr>
<tr>
<td>6.8</td>
<td>83</td>
</tr>
<tr>
<td>6.9</td>
<td>84</td>
</tr>
</tbody>
</table>
6.10 ROC Curve of Partitioning Ensemble Feature Selection for SpamAssassin Dataset 84

6.11 CPU Time of Partitioning Ensemble Feature Selection for SpamAssassin Dataset 85

6.12 AUC Under ROC Curve of Ensemble Classifier using QBGSA based on Different Size of Training Data 86

6.13 CPU Time of Ensemble Classifier using QBGSA based on Different Size of Training Data 86

6.14 ROC Curve of Ensemble Classifier using QBGSA for SpamAssassin Dataset 87
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBA</td>
<td>Binary Bat Algorithm</td>
</tr>
<tr>
<td>BGA</td>
<td>Binary Genetic Algorithm</td>
</tr>
<tr>
<td>BGSA</td>
<td>Binary Gravitational Search Algorithm</td>
</tr>
<tr>
<td>BQGSA</td>
<td>Binary Quantum Gravitational Search Algorithm</td>
</tr>
<tr>
<td>BQPSO</td>
<td>Binary Quantum Particle Swarm Optimization</td>
</tr>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbor</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
</tr>
<tr>
<td>PSO</td>
<td>Particle Swarm Optimization</td>
</tr>
<tr>
<td>QC</td>
<td>Quantum Computing</td>
</tr>
<tr>
<td>QEIA</td>
<td>Quantum Evolutionary-Inspired Algorithm</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
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 Stamatospolous, 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 desirable 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, 2010a; 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, 2010a; 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 Algorithm (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


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


97


98


