# IMPROVED SCHEME OF E-MAIL SPAM CLASSIFICATION USING META-HEURISTICS FEATURE SELECTION AND SUPPORT VECTOR MACHINE

NADIR OMER FADL ELSSIED HAMED

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

FEBRUARY 2015

To my beloved parents, lovely wife Rania, My daughters (Gina, Taleen), brothers,sisters, and to whole Muslim Umma.

# ACKNOWLEDGEMENT

"Praise be to Allah, the cherisher and the sustainer of the world", "praise be to him he who taught by the pen, taught man, that which he did not know"

First and foremost, my sincere thanks to God, who endowed me to complete this PhD thesis. I would like to thank my supervisor, Associate Professor Dr. Othman Ibrahim, for all your guidance, support, brilliant ideas, numerous hours of discussions, patience, and the opportunities you have presented me. Your managerial skills and uncompromising quest for excellence always motivated me to present the best of what I can. I wish to thank Associate Professor Dr. Ahmad Kamil bin Mahood and Professor Dr. Siti Mariyam binti Shamsuddin for accepting to review and examine my PhD thesis.

The International Doctoral Fellowship (IDF) has played an important role in this research, accordingly, many thanks from the depths of my heart to UTM for granting me the IDF award. I also owe my great deal of thank to the Faculty of Computing for providing the best environment to carry out the research, and to our professors and staff for their friendly support and help throughout the study and especial thanks to a Sudanese research group for supporting me in discussion. And most importantly, I would like to thank to my family for their love and support, especially my mother, my wife and my daughters for their continuous support and supplication. Finally, I have made many friends during my time at UTM and I would like to thank all my friends and colleagues for their support and help.

# ABSTRACT

With the technological revolution in the 21st century, time and distance of communication are decreased by using electronic mail (e-mail). Furthermore, the growing use of e-mail has led to the emergence and further growth problems caused by unsolicited bulk e-mails, commonly referred to as spam e-mail. Many of the existing supervised algorithms like the Support Vector Machine (SVM) were developed to stop the spam e-mail. However, the problem of dealing with large data and high dimensionality of feature space can lead to high execution-time and low accuracy of spam e-mail classification. Nowadays, removing the irrelevant and redundant features beside finding the optimal (or near-optimal) subset of features significantly influences the performance of spam e-mail classification; this has become one of the important challenges. Therefore, in order to optimize spam e-mail classification accuracy, dimensional reduction issues need to be solved. Feature selection schemes become very important in order to reduce the dimensionality through selecting a proper subset feature to facilitate the classification process. The objective of this study is to investigate and improve schemes to reduce the execution time and increase the accuracy of spam e-mail classification. The methodology of this study comprises of four schemes: one-way ANOVA f-test, Binary Differential Evolution (BDE), Opposition Differential Evolution (ODE) and Opposition Particle Swarm Optimization (OPSO), and combination of Differential Evolution (DE) and Particle Swarm Optimization (PSO). The four schemes were used to improve the spam e-mail classification accuracy. The classification accuracy of the proposed schemes were 95.05% with population size of 50 and 1000 number of iterations in 20 runs and 41 features. The experiment of the proposed schemes were carried out using *spambase* and *spamassassin* benchmark dataset to evaluate the feasibility of proposed schemes. The experimental findings demonstrate that the improved schemes were able to efficiently reduce the number of features as well as improving the e-mail classification accuracy.

# ABSTRAK

Dengan revolusi teknologi pada abad ke-21, masa dan jurang komunikasi menurun dengan penggunaan mel elektronik (e-mel). Tambahan pula, penggunaan e-mel yang semakin meningkat telah mengakibatkan kebangkitan pertumbuhan masalah yang disebabkan oleh e-mel pukal yang tidak dipesan, biasanya dirujuk sebagai e-mel spam. Kebanyakan algoritma pengelasan e-mel spam sedia ada seperti Mesin Vektor Sokongan (MVS) dibangunkan untuk menghentikan e-mel spam. Walau bagaimanapun, masalah berurusan dengan data yang besar dan ruang ciri berdimensi tinggi boleh membawa kepada ketepatan pengelasan yang rendah dan kerumitan komputasi yang tinggi. Pada masa kini, mencari subset ciri-ciri yang optimum (hampir optimum) mempengaruhi prestasi pengelasan e-mel spam; ini telah menjadi salah satu cabaran yang penting. Namun, untuk mengoptimumkan keupayaan pengelasan e-mel spam, isu-isu pengurangan dimensi perlu diselesaikan. Skema pemilihan ciri subset menjadi sangat penting untuk mengurangkan dimensi dengan memilih ciri subset yang sesuai untuk memudahkan proses pengelasan. Objektif kajian ini adalah untuk mengkaji dan membangunkan skema untuk meningkatkan dan mengekalkan ketepatan, dan mengurangkan kerumitan komputasi bagi pengelasan e-mel spam. Metodologi kajian ini terdiri daripada empat skema: ANOVA ujian-f satu hala, Evolusi Perbezaan Binari (BDE), Evolusi Perbezaan Bertentangan (ODE) dan Pengoptimuman Kelompok Zarah Bertentangan (OPSO), dan gabungan antara Evolusi Perbezaan (DE) dan Pengoptimuman Kelompok Zarah (PSO). Empat skema tersebut telah digunakan untuk meningkatkan ketepatan pengelasan e-mel spam. Ketepatan pengelasan pendekatan yang dicadangkan adalah 95.05% dengan jumlah populasi 50 dan 1000 lelaran dalam 20 turutan dan 41 ciri-ciri. Percubaan pendekatan yang dicadangkan dilaksanakan dengan menggunakan penanda aras set data *spambase* dan *spamassassin* untuk menilai ketersavran pendekatan yang dicadangkan. Penemuan eksperimen menunjukkan bahawa, pendekatan baru dapat mengurangkan dengan cekap bilangan ciri-ciri serta meningkatkan ketepatan pengelasan e-mel.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| ACO | - | Ant Colony Optimization |
| ACO | - | Ant Colony Optimization |
| AI | - | Artificial Intelligence |
| BDE | - | Binary Differential Evolution |
| CA | - | Collaborative Approach |
| CAI | - | Computational Artificial Intelligence |
| CBA | - | Content-Based Approach |
| CBR | - | Case-Based Reasoning |
| CC | - | Correlation Coefficient |
| CV | - | Cross Validation |
| DE | - | Differential Evolution |
| DEPSO | - | DE and PSO |
| DF | - | Document Frequency |
| DR | - | Dimension Reduction |
| EAs | - | Evolutionary Algorithms |
| ECs | - | Evolutionary Computations |
| E-mail | - | Electronic Mail |
| EP | - | Evolutionary Programming |
| ES | - | Evolutionary Strategies |
| FN | - | False Negatives |
| FP | - | False Positive |
| FS | - | Feature Selection |
| FT | - | Feature Transformation |
| GA | - | Genetic Algorithm |
| IG | - | Information Gain |
| ISPs | - | Internet Service Providers |

| | | |
|---|---|---|
| LDA | - | Linear Discriminant Analysis |
| MI | - | Mutual Information |
| ML | - | Machine Learning |
| MLP | - | Multi-Layer Perceptron |
| NB | - | Naïve Bayes |
| OBL | - | Opposition-Based Learning |
| ODE | - | Opposition Differential Evolution |
| OPSO | - | Opposition Particle Swarm Optimization |
| PCA | - | principle Component Analysis |
| PSO | - | Particle Swarm Optimization |
| RBF | - | Radial Basis Function |
| SA | - | Simulating Annealing |
| SC | - | Soft Computing |
| SRDA | - | Spectral Regression Discriminant Analysis |
| SVD | - | Singular Value Decomposition |
| SVM | - | Support Vector Machine |
| TF | - | Term Frequency |
| TN | - | True Negative |
| TP | - | True Positive |
| TS | - | Term Strength |
| UBE | - | Unsolicited Bulk E-mail |
| UCE | - | Unwanted Commercial E-mail |
| UCI | - | University of California |

# LIST OF APPENDIX

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

In recent years, with the rapid development of Internet technologies, the number of people using electronic mail (e-mail) is continuously increasing (Allias et al., 2014; Trivedi and Dey, 2013). The usage of e-mail in everyday communication is mainly due to its time saving and cost reduction as well as being the fastest and easiest means of the delivery of messages. It has gained extremely wide popularity among internet users in information exchange (Chhabra *et al.*, 2010; Naksomboon *et al.*, 2010; Youn and McLeod, 2007a; Zhang *et al.*, 2011b). However, the increase in e-mail users leads to the appearance of unwanted and harmful e-mails known as "spam e-mail" (Kesharwani and Lade, 2013; Kumar et al., 2012). According to many researchers, spam e-mail forms a threat to the Internet user community and service providers (Almeida *et al.*, 2010). It has negative impacts on the usability of e-mail and IT infrastructure by occupying important resources such as wasting network bandwidth, producing unnecessary network congestion (Lai *et al.*, 2009b). As well as consuming computing resources and time, spam e-mail reduces the effectiveness of legitimate advertising, filling mailboxes, storage space (Pour *et al.*, 2012). As the usage of e-mail continues to increase, the ratio of spam e-mail is also increasing and thereby becomes more difficult, time-consuming and costly to be classified manually (Oda and White, 2003; Soranamageswari and Meena, 2010). In recent years, researchers efficiently solved the above issue (classifying e-mails). E-mail spam classification methods were investigated, and many studies have been proposed to improve their performance (Almeida *et al.*, 2011; Yevseyeva *et al.*, 2013). In addition, e-mail spam identification is a difficult task because spammers use tricks in order to avoid spam classifiers to ensure their delivery (Méndez *et al.*, 2008; Sousa *et al.*, 2013).

E-mail spam classification schemes are affected negatively when dealing with large data and a high dimensionality of the feature space (Almeida *et al.*, 2011; Islam and Yang, 2010). The high dimensionality of feature space may contain a large number of irrelevant and redundant features that can result in low accuracy and high execution time for the classifier (Behjat *et al.*, 2012a; Fagbola Temitayo, 2012). With the proliferation of high-dimensional feature space, feature selection has become an essential task of learning process. Generally, feature selection is widely used in e-mail spam classification to reduce the high dimensionality of feature space without sacrificing the performance of the classification (Khatri and Emmanuel, 2013). Reducing the dimensionality of the feature space allows the algorithm to work faster and more efficiently (Behjat *et al.*, 2013; He *et al.*, 2009b). The large number of features affects the execution time and led to reduced performance of e-mail classification (Lai *et al.*, 2009a). As a part of any feature selection algorithm, there are numerous factors that need to be considered. The existing evaluation measure utilized in feature selection techniques are divided into three categories namely filter, wrapper and embedded approaches (Cortez *et al.*, 2012; Khoshgoftaar *et al.*, 2013b; Unler *et al.*, 2011b). The main aim of the feature selection in e-mail spam classification is to overcome the high dimensionality of the feature space through removing the irrelevant and redundant features (Behjat et al., 2013; Xue et al., 2012). The irrelevant and redundant features increase the amount of the search space and make e-mail spam classification more difficult (Gomez et al., 2012). To overcome these challenges, reduction of high dimensionality is proposed, which decreases the number of features in order to achieve higher classification accuracy (Wu et al., 2011a). In recent work Sousa et al (2013) reported that the correct selection of subset features is a key issue in the task of discriminating between spam e-mail and non-spam e-mail. Another survey by (Guzella and Caminhas, 2009) stated that the biggest challenge in e-mail spam classification is to provide a classification scheme to reduce the execution time and improve the classification accuracy. The key of this study is to evaluate how different feature subset selection schemes can affect the performance of learning algorithm such as SVM-based e-mail classification system via reducing the dimensionality of the feature space.

## 1.2 Problem Background

The increasing risk size of spam e-mail has become a serious issue and uncontrollable not only to the Internet, but also for users and for Internet Service Providers (ISPs) (Idris *et al.*, 2014; Sathawane, 2013). Spam e-mail is an intrusion of privacy, with problematic content such as online fraud, phishing attacks or viruses (Méndez *et al.*, 2008). Spam e-mail creates a serious threat to the security of networked systems and computer users everywhere (Islam and Yang, 2010). Furthermore, a large amount of space is occupied in user's mailbox, and there is no relation between spam message content and receivers (Alguliev *et al.*, 2011; Pour *et al.*, 2012). Further, users of e-mails are affected by spam e-mail due to time spent to distinguish between spam e-mail and non-spam e-mail (Guzella and Caminhas, 2009). The importance of safeguarding the Inbox mail against spam e-mail is an essential issue and e-mail spam classification plays an important role in ensuring a non-spam e-mail. Recently, the high increase of spam e-mail has become a challenge. The e-mail spam classification problem is growing because spammers will always find new ways to attack e-mail spam classifier due to the economic benefits of sending spam e-mail (Sathawane, 2013). Therefore, there is a need to develop spam e-mail classifiers that can effectively eliminate the increasing of spam e-mail automatically before the spam enters the user's mailbox (Chhabra *et al.*, 2010). Among the approaches developed to combat spam e-mail, classification is an important and popular one (Zhang *et al.*, 2014).

E-mail spam classifiers have become obsolete in a short period of time and need to be updated on a regular basis due to the continuous changing of techniques used to send spam e-mail (Yevseyeva *et al.*, 2013). Currently, there are two major approaches for e-mail spam classification: Collaborative Approach (CA) and

Content-Based Approach, (CBA) and some researchers have combined features from both of them to develop new approach (He *et al.*, 2009b; Kuang *et al.*, 2014; Sathawane, 2013; Sousa *et al.*, 2010). The CA is based on sharing information about spam e-mails, while CBA uses a data mining classifier to analyze content (.e.g. word frequencies) (Cortez *et al.*, 2012). The two approaches have their own drawbacks; CA often suffers from the sparsity of data although many techniques have been developed to improve this drawback. On the other hand, CBA behavior is dependent not only on the classifier learning capabilities, but also the type of FS method adopted (Sousa *et al.*, 2013). In regards to e-mail spam classification, current research on CBA relies mainly on improving individual classifier performance via selecting the optimum subset features. The effectiveness of CBA e-mail spam classifier relies on the appropriate choice of the features (Méndez *et al.*, 2006). The increasing importance of e-mail spam classification motivates various aspects of classification-related study that provide a new solution, which may not be achievable by conventional e-mail classification approaches. The main goal of e-mail spam classification is to pre-sort messages into two categories of spam e-mail and non-spam e-mail with a high accuracy rate and low execution time (Terri Oda, 2005). Although, there are many algorithms such as SVM that have been developed for e-mail spam classification problems, but the problem is still not being solved completely (Ashok and Shrivastav, 2014; Kumar *et al.*, 2012).

The big challenge is to develop better schemes that automatically classifies spam e-mail from non-spam e-mail (Lai *et al.*, 2009a). The survey by Allias et al (2014) suggests that the e-mail spam classification has low classification accuracy due to a high dimensionality of feature space that contains redundant and irrelevant features (Allias *et al.*, 2014; Suebsing and Hiransakolwong, 2012; Wu *et al.*, 2011a). Unfortunately a high dimensionality of feature space after preprocessing became a significant challenge for the classifier (Allias *et al.*, 2014). In addition to the large number of data issues, the excessive number of features can also degrade the e-mail spam classification accuracy (Parimala and Nallaswamy, 2012). This is because the large number of features leads to the problem of high dimensional feature space (Behjat *et al.*, 2012c). The irrelevant and redundant features lead to a high execution time. Yet, not all of these features have the same importance for the e-mail spam classification, and some of them may be unimportant due to redundancy or may even be irrelevant. One of the fundamental motivations for feature selection is to overcome the problem of high dimensionality (Mendez *et al.*, 2006). To overcome this problem, dimensionality reduction schemes have been proposed, which can reduce the number of irrelevant features in order to achieve higher classification accuracy (Wu *et al.*, 2011b). Therefore, a wide variety of methods have been proposed in the literature in order to determine the most important features for classification (Aggarwal and Zhai, 2012). The correct Feature Selection (FS) approaches are a key challenge to improve the e-mail spam classification (Cortez *et al.*, 2012). Also, the result of eliminating irrelevant and redundant features leads to an optimized classification process that is efficient and accurate (Yevseyeva *et al.*, 2013). Nowadays the use of a finite subset of features to classify the e-mail as spam e-mail or non-spam e-mail is an important research topic (Kumar *et al.*, 2012; Lai *et al.*, 2009a; Yevseyeva *et al.*, 2013).

Recently, most of the researchers in the area of e-mail spam classification have concentrated more on obtaining the optimal classification accuracy via decreasing of the dimensionality feature space (Parimala and Nallaswamy, 2012; Yevseyeva *et al.*, 2013). There are many solution methods available for e-mail spam classification. Most of these methods are based on Machine Learning (ML) algorithms such as classification techniques (Fagbola Temitayo, 2012; Guzella and Caminhas, 2009). The literature review presents various ML methods that have been proposed for e-mail spam classification, such as SVM algorithm. SVM algorithm was introduced in mid 1990s, and it is one of the robust methods for binary classification (Rakse and Shukla, 2010). It is a popular algorithm applied in the binary classification. Furthermore, it is one of the top ten influential algorithms for data mining as well as the most accurate method among all well-known algorithms (Maali and Al-Jumaily, 2013; Wu *et al.*, 2008). Although many learning algorithms such as SVM have been widely used in e-mail spam classification, yet the problem of dealing with huge data and high dimensional feature space that leads to low accuracy and high execution time (Chhabra *et al.*, 2010; Wang, 2008). Figure1.1 illustrates the scenario leading to the problem addressed by this research.

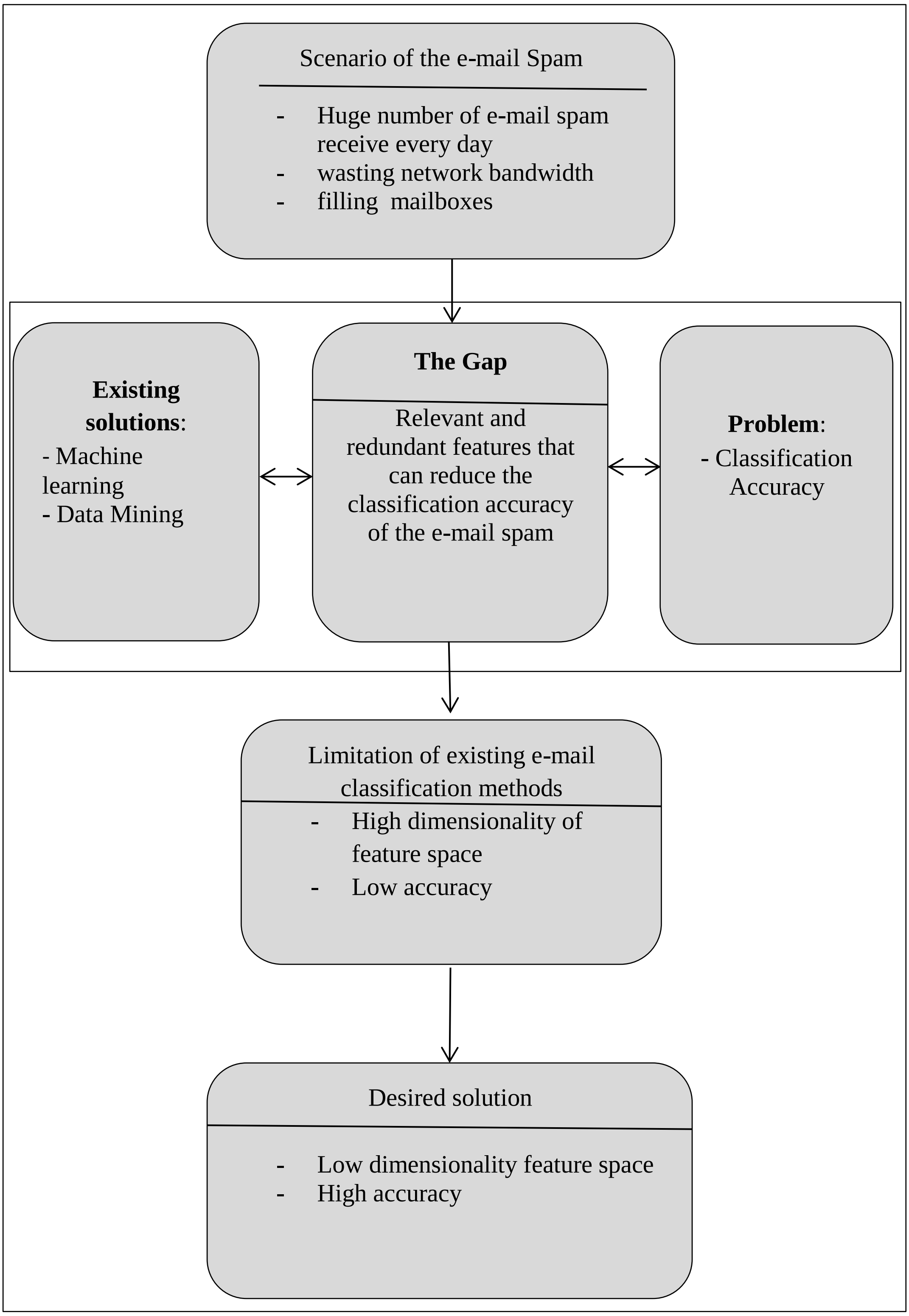


**Figure 1.1** Scenario leading to the problem

The problem of SVM with a high dimensional feature space and the large dataset classification still remains a challenge (Liu *et al.*, 2013). Yet the problem of optimizing the SVM in terms of improving the classification accuracy is the subject of ongoing research (Zhang *et al.*, 2012). The accuracy of SVM as a classifier was affected by the high dimensionality of features (number of variable) (Wang *et al.*, 2005). For example Genetic Algorithm (GA) algorithm was adopted in the work of Wang et al (2005) to select the optimal subset feature. SVM algorithm has continuously received increasing attention from researchers in spam e-mail classification. Some research has been done in SVM with e-mail spam classification (Lai *et al.*, 2009a; Lai and Wu, 2007). The conventional SVM algorithm is insufficient because it considers all e-mail features as equal in importance. All the e-mail features are used even if they have irrelevant or redundant features. There are different processes and methods used in order to enhance the accuracy and computational complexity of the learning algorithms such as SVM. Many researchers approach this problem of computational complexity and the classification accuracy via performance feature selection (Fagbola Temitayo, 2012; Maldonado and L'Huillier, 2013).

This thesis focuses on reducing the number of features (high dimensionality) via selecting the optimal (or near-optimal) subset features based on ECs algorithm. Additionally, reducing the number of features increases the e-mail spam classification accuracy. The literature presents that the researchers have tried to enclose feature selection schemes to select the optimal subset feature in classification problems (Parimala and Nallaswamy, 2012). Empirically, the FS approaches lead to a higher accuracy rate of e-mail classification. In the area of data mining, many researchers have mentioned that the maximum performance is not achieved by using all available features but by using a subset of all features. Moreover, the correlation between the features influences the classification accuracy. There is little research that provides a study to choose the optimal (or near-optimal) subset features in e-mail spam classification (Behjat *et al.*, 2012b). Generally, the problems in e-mail spam classification can be classified into three groups: high dimensional feature space, execution time, and low accuracy. and various researchers put in considerations (Suebsing and Hiransakolwong, 2012). From the above there are further needs to design and build better approaches by using ECs algorithm to select the optimal (near-optimal) subset features. Additionally, to differentiate between low and high features in terms of importance, researchers in this thesis consider the features that have emerged in order to obtain higher e-mail spam classification accuracy. Also, this study focuses on using feature subset selection schemes that help to select the subset features related to the performance of the e-mail classification system. The selection of the features is based upon some accuracy criteria, without significantly reducing the performance from the classifier system.

## 1.3 Problem Statement

Although several supervised algorithms have been widely used in e-mail spam classification, the problem of dealing with huge data and high dimensionality (many feature) is low accuracy as many researches are being carried out (Chhabra *et al.*, 2010; Fagbola Temitayo, 2012; Morariu *et al.*, 2006).

*"A high dimensionality of the feature space based on a large number of irrelevant and redundant features can affect the classification accuracy of various supervised algorithms during run for e-mail spam classification"*.

## 1.4 Research Questions

This research is intended to deal with the problems related to e-mail spam classification. This research seeks to answer the following main question:

*How can the removal of irrelevant and redundant features besides the selection of the optimal (or near-optimal) subset features reduce the high* *dimensionality of the feature space and increase the classification accuracy of e-mail spam classification?*

In order to answer the main question raised above, the following sub-questions need to be addressed:

i. How can effectively eliminate the irrelevant and redundant features to reduce the high dimensionality of e-mail spam classification?

ii. What are the optimum features that could significantly improve the execution time for e-mail spam classification?

iii. How can OBL approach improve the DE and PSO algorithms in terms of enhancing the e-mail spam classification accuracy?

iv. How can the hybridization of PSO and DE as feature selections enhance the e-mail spam classification accuracy?

## 1.5 The Aim of the Study

This study aims to propose schemes of selecting the optimal (or near-optimal) subset features and obtaining the optimum (or near-optimum) overall classification accuracy of e-mail spam classification regardless of the number of the optimal subset features selected. This study aims to remove the irrelevant and redundant features. The proposed schemes must ensure that reducing of the high execution time as well as improving the accuracy of e-mail spam classification.

## 1.6 Objectives of the Study

The main goal of this study is to proposed schemes to select an optimal feature and to improve the e-mail spam classification accuracy. The objectives of this study are:

i. To reduce the high dimensionality for e-mail spam classification based on one-way ANOVA F-test.
ii. To improve the execution time of e-mail spam classification based on BDE scheme as feature selection.
iii. To improve the accuracy of e-mail spam classification based on ODE and OPSO scheme.
iv. To improve hybrid scheme based on combination of PSO and DE for feature selection to enhance the accuracy of e-mail spam classification.

## 1.7 Scope of the Study

The scope of this study is to answer the research questions stated above in order to draw conclusions. Furthermore, the preceding section stated the objectives of this research which focus on how to produce an optimal (or near-optimal) e-mail spam classification schemes. The following aspects are the scope of the study for the stated objectives:

i. Binary e-mail spam classification (spam detection) due to the nature of e-mail datasets.
ii. Support Vector Machine (SVM) algorithm as classification algorithm due to popular algorithm that are used as classifiers for e-mail spam classification area.
iii. Correlation coefficient as fitness functions

iv. Binary Differential Evolution (BDE) algorithm to reduce the execution time for e-mail spam classification.
v. Improved Differential Evolution (DE) algorithm based on Opposition Based Learning (OBL) and Particle Swarm Optimization (PSO) algorithm to improve the accuracy of e-mail spam classification.
vi. Improved Particle Swarm Optimization (PSO) algorithm based on Opposition Based Learning (OBL) to improve the accuracy of e-mail spam classification.
vii. Evaluation of the performance of the proposed schemes using two standards dataset: "spambase" and "spamassassin." These are obtained from the machine learning repository of the Center for Machine Learning and Intelligent.
viii. The accuracy, precision, F-measure, false positive, computational complexity (execution time) and recall were selected to measure and evaluate the systems generated.
ix. The statistical significant test (Pearson correlation coefficient) was used to measure the agreement level between the proposed and the other methods such as e-mail spam classification with SVM using all features.

## 1.8 Significance of the Study

Since the start of research in e-mail spam classification, all proposed schemes aim to increase the performance accuracy of classification results without putting other issues into consideration such as high dimensionality feature space. Thus, our experiments show that the accuracy increased by selecting the optimal (near-optimal) subset features before classification via designing a classification scheme or by combining methods of other techniques. This research desires to make a significant contribution by presenting a new evolutionary feature subset selection scheme with SVM for e-mail spam classification. The improve schemes are to increase the accuracy and at the same time to reduce the execution time for e-mail spam classification. According to Symantec Intelligence Report in September 2012 the percentage of spam in e-mail traffic was increased by 2.7 percentage points from

August and averaged 75%, in addition to the Kaspersky Lab annual report the in which the total amount of spam in mail traffic was 78.5% (Bulletin, 2012; Wood, 2012).

## 1.9 Research Contributions

The contributions of this study are as follows:

i. An improved e-mail spam classification scheme using feature selection scheme based on one-way ANOVA and SVM as a classifier to remove the irrelevant and redundant features.
ii. An improved e-mail spam classification scheme using a significant feature selection based on binary DE and SVM as a classifier to reduce the high execution time.
iii. An improved e-mail spam classification scheme using OPSO and ODE to select the optimal (or near-optimal) features and SVM classifier to improve the accuracy of classification.
iv. An improved e-mail spam classification scheme using a feature selection based on hybrid (PSO and DE) and SVM classifier to improve the accuracy of classification.

## 1.10 Thesis Organization

This thesis is organized into eight chapters as:

**Chapter 1: Introduction**

The introductory chapter of this thesis provides a brief overview of some of the issues that are of concern to those working in the field of e-mail spam classification. This chapter will also look at the overview of the whole studies, the problem statements, research questions, aims, and the scope of this study as well as examining the contributions this research can make to the field of e-mail spam classification.

**Chapter 2: Literature Review**

This chapter provides background information and reviews of the related work in the area of e-mail spam classification. The chapter reviews recent surveys presented in the area. Since this study proposes ECs algorithm based solutions, the chapter also reviews, most especially, the feature selection schemes and classification research based on similar or other EAs. It reviews ML approaches that have been presented to improve the search performance of EAs. The chapter covers available datasets utilized in the methodology.

**Chapter 3: Research Methodology**

This chapter defines the methodology followed in this research to achieve the study's objectives. The main experiments of this study are to be conducted through four main approaches: a new feature subset selection based on one-way ANOVA F-test for e-mail spam classification, feature subset selection based Binary Differential Evolution Algorithm for e-mail spam classification, opposition DE and opposition PSO Algorithm based feature selection, and a hybrid of DE and PSO as feature subset selection in E-mail Spam Classification.

**Chapter 4: A New Feature Selection Based on One-Way ANOVA F-test**

This chapter provides a new feature subset selection based on one-way ANOVA f-test to remove the irrelevant and redundant features so these researchers are left with the most relevant features that increase or maintain accuracy of e-mail spam classification.

**Chapter 5: Feature Selection Based Binary Differential Evolution (BDE) Algorithm**

This chapter demonstrates feature subset selection based on Binary Differential Evolution (BDE) scheme and uses correlation coefficient as fitness function to determine the most optimal subset features that could contribute to reduce the execution time and improve the e-mail spam classification accuracy.

**Chapter 6: Feature Subset Selection Based ODE and OPSO**

The main goal of this chapter is to avoid the problem of generating solutions based on random estimates. The problem of the application based on random estimates (guesses) is that it may give different solutions each time which are far from the optimal solutions. This chapter presents feature subset selection based on ODE and also based on OPSO for efficient feature subset selection and evaluated by e-mail spam classification rate using SVM as classifier to present a better accuracy. The DE and PSO algorithms are computationally expensive due to the slow nature of the evolutionary process. In this chapter researchers use the correlation coefficient as a fitness function for OPSO while simultaneously using the correlation coefficient as a fitness function for ODE. Additionally, they use OBL to improve the convergence rate of classical DE and PSO and to improve the speed of DE and PSO.

### Chapter 7: Hybridization of PSO and DE as Feature Selection

These chapters present evolutionary feature selection based on the hybridization of DE and PSO. In this chapter, PSO operator such pbest and gbest were used instead of three random generations in mutation phase of DE to accelerate the convergence rate of DE algorithm, which indicates that researchers use the PSO before the mutation phase in DE. Then, the feature selection based on hybrid of DEPSO approach was applied to determine the most important features contributing to e-mail spam classification accuracy.

### Chapter 8: Conclusion and Future Work

Chapter 8 will review the conclusion of the research discussed throughout this study. This section will also put forward recommendations for future studies.

# REFERENCES


Abraham, A., Das, S., and Konar, A. (2006, 0-0 0). Document Clustering Using Differential Evolution. Paper presented at the IEEE Congress on Evolutionary Computation, 2006. CEC 2006. , 1784-1791.

Aggarwal, C. C., and Zhai, C. (2012). A survey of text classification algorithms. In Mining Text Data (pp. 163-222): Springer.

Ahandani, M. A., and Alavi-Rad, H. (2012). Opposition-based learning in the shuffled differential evolution algorithm. Soft computing, 16(8), 1303-1337.

Al-Qunaieer, F. S., Tizhoosh, H. R., and Rahnamayan, S. (2010). Opposition based computing—A survey. Paper presented at the The 2010 International Joint Conference on Neural Networks (IJCNN), , 1-7.

Alguliev, R. M., Aliguliyev, R. M., and Nazirova, S. A. (2011). Classification of textual E-mail spam using data mining techniques. Applied Computational Intelligence and Soft Computing, 2011, 10.

Alguliyev, R., and Nazirova, S. (2012). Two Approaches on Implementation of CBR and CRM Technologies to the Spam Filtering Problem. Journal of Information, 3.

Ali, M., Pant, M., and Abraham, A. (2009). Simplex differential evolution. Acta Polytechnica Hungarica, 6(5), 95-115.

Allias, N., Megat, M. N., Noor, M., and Ismail, M. N. (2014). A Hybrid Gini PSO-SVM Feature Selection Based on Taguchi Method: An Evaluation on Email Filtering.

Almeida, T. A., Almeida, J., and Yamakami, A. (2011). Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of Internet Services and Applications, 1(3), 183-200.

Almeida, T. A., Yamakami, A., and Almeida, J. (2010). Probabilistic anti-spam filtering with dimensionality reduction. Paper presented at the Proceedings of the 2010 ACM Symposium on Applied Computing, 1802-1806.

Alper Kursat Uysal, S. G. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems.

Arani, S. H. S., and Mozaffari, S. (2013a). Genetic-based feature selection for spam detection. Paper presented at the Electrical Engineering (ICEE), 2013 21st Iranian Conference on, 1-6.

Arani, S. H. S., and Mozaffari, S. (2013b). Genetic-based feature selection for spam detection. Paper presented at the 21st Iranian Conference on Electrical Engineering (ICEE), 2013 1-6.

Arun Rajput , D. T. (2009). Adaptive Spam Filtering based on Bayesian Algorithm. International Journal on Advanced Computer Engineering and Communication Technology, Vol-1(1 : ISSN 2278 - 5140).

Ashok, A. R., and Shrivastav, G. (2014). A Spam Filtering Technique using SVM Light Tool.

Babaoglu, İ., Findik, O., and Ülker, E. (2010). A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. Expert Systems with Applications, 37(4), 3177-3183.

Behjat, A. R., Mustapha, A., Nezamabadi-pour, H., Sulaiman, M., and Mustapha, N. (2012a). GA-based feature subset selection in a spam/non-spam detection system. Paper presented at the International Conference on Computer and Communication Engineering (ICCCE), 2012, 675-679.

Behjat, A. R., Mustapha, A., Nezamabadi-pour, H., Sulaiman, M. N., and Mustapha, N. (2013). A PSO-Based Feature Subset Selection for Application of Spam/Non-spam Detection. In Soft Computing Applications and Intelligent Systems (pp. 183-193): Springer.

Blum, C., and Roli, A. (2008). Hybrid metaheuristics: an introduction. In Hybrid Metaheuristics (pp. 1-30): Springer.

Bonissone, P. P. (1997). Soft computing: the convergence of emerging reasoning technologies. Soft computing, 1(1), 6-18.

Borse, A. A., and Kamlapur, S. M. (2012). Selection Of Feature Regions Set For Digital Image Using Optimization Algorithm. International Journal of Communication.

Cervante, L., Xue, B., Zhang, M., and Shang, L. (2012). Binary particle swarm optimisation for feature selection: A filter based approach. Paper presented at the IEEE Congress on Evolutionary Computation (CEC), 2012 1-8.

Chakravarty, S. (2010). A survey on text classification techniques for e-mail filtering. Paper presented at the Second International Conference on Machine Learning and Computing (ICMLC), 2010 32-36.

Chen, J., Huang, H., Tian, S., and Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36(3, Part 1), 5432-5435.

Chhabra, P., Wadhvani, R., and Shukla, S. (2010). Spam Filtering using Support Vector Machine. International Conference [ACCTA-2010], Vol.1( 2).

Chuang, L.-Y., Hsiao, C.-J., and Yang, C.-H. (2009). An Improved Binary Particle Swarm Optimization with Complementary Distribution Strategy for Feature Selection. Paper presented at the International Conference on Machine Learning and Computing, 244-248.

Çıltık, A., and Güngör, T. (2008). Time-efficient spam e-mail filtering using n-gram models. Pattern Recognition Letters, 29(1), 19-33.

Cormack, G. V. (2007). Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval, 1(4), 335-455.

Cortez, P., Vaz, R. F. M., Rocha, M., Rio, M., and Sousa, P. (2012). Evolutionary symbiotic feature selection for email spam detection.

Cosme, R. d. C., and Krohling, R. A. (2011). Support Vector Machines applied to noisy data classification using differential evolution with local search.

Dancey, C., and Reidy, J. (2008). Statistics without Maths for Psychology, 408-413. Harlow: United States: Prentice Hall

Das, S., Abraham, A., and Konar, A. (2008). Automatic clustering using an improved differential evolution algorithm. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 38(1), 218-237.

Das, S., and Suganthan, P. N. (2011). Differential evolution: A survey of the state-of-the-art. Evolutionary Computation, IEEE Transactions on, 15(1), 4-31.

Drucker, H., Wu, D., and Vapnik, V. N. (1999). Support vector machines for spam categorization. Neural Networks, IEEE Transactions on, 10(5), 1048-1054.

Eiben, A. E., and Smith, J. E. (2010). Introduction to evolutionary computing (Vol. 2): Springer Berlin.

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.

Elloumi, M., Hayati, P., Iliopoulos, C., Mirza, J. A., Pissis, S. P., and Shah, A. (2013). Comparison for the detection of Virus and spam using pattern matching tools. Paper presented at the International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 304-311.

Engelbrecht, A. P., and Pampara, G. (2007). Binary differential evolution strategies. Paper presented at the IEEE Congress on Evolutionary Computation, 2007. CEC 2007. , 1942-1947.

Fagbola Temitayo, O. S., digun Abimbola (2012). Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification Computer Engineering and Intelligent Systems Vol 3, No.3, 2012(No.3,), 17-28.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research, 3, 1289-1305.

Forman, G. (2004). A pitfall and solution in multi-class feature selection for text classification. Paper presented at the Proceedings of the twenty-first international conference on Machine learning, 38.

Fu, W., Johnston, M., and Zhang, M. (2011a). Hybrid particle swarm optimisation algorithms based on differential evolution and local search. In AI 2010: Advances in Artificial Intelligence (pp. 313-322): Springer.

Fu, W., Johnston, M., and Zhang, M. (2011b). A hybrid particle swarm optimisation with differential evolution approach to image segmentation. In Applications of Evolutionary Computation (pp. 173-182): Springer.

GA, D. K., and Okdem, S. (2004). A simple and global optimization algorithm for engineering problems: differential evolution algorithm. Turk J Elec Engin, 12(1).

Garcia-Nieto, J., Alba, E., and Apolloni, J. (2009). Hybrid de-svm approach for feature selection: Application to gene expression datasets. Paper presented at the 2nd International Logistics and Industrial Informatics, 2009. LINDI 2009. , 1-6.

Gawande, P. Y. P. a. S. H. (2012). A Comparative Study on Different Types of Approaches to Text Categorization. International Journal of Machine Learning and Computing, vol.2,no.4,pp.423-426.

Ghosh, A., Das, S., Chowdhury, A., and Giri, R. (2011). An improved differential evolution algorithm with fitness-based adaptation of the control parameters. Information Sciences, 181(18), 3749-3765.

Ghosh, A., Datta, A., and Ghosh, S. (2012). Self-adaptive differential evolution for feature selection in hyperspectral image data. Applied Soft Computing.

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.

Günal, S., Ergin, S., Gülmezoğlu, M., and Gerek, Ö. (2006). On feature extraction for spam e-mail detection. Multimedia Content Representation, Classification and Security, 635-642.

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

Han, J., Kamber, M., and Pei, J. (2006). Data mining: concepts and techniques: Morgan kaufmann.

Han, L., and He, X. (2007). A novel opposition-based particle swarm optimization for noisy problems. Paper presented at the Third International Conference on Natural Computation, 2007. ICNC 2007. , 624-629.

He, X., Zhang, Q., Sun, N., and Dong, Y. (2009). Feature selection with discrete binary differential evolution. Paper presented at the Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on, 327-330.

Hong, C. (2011). Improving classification in Bayesian networks using structural learning. World Academy of Science, Engineering and Technology, 75, 1407-1411.

Huaitie, X., Guoyu, F., Zhiyong, S., and Jianjun, C. (2010). Hybrid optimization method for parameter selection of support vector machine. Paper presented at the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), 2010 613-616.

Huang, J., Cai, Y., and Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognition Letters, 28(13), 1825-1844.

Idris, I. (2012). Model and Algorithm in Artificial Immune System for Spam Detection. International Journal, 3.

Idris, I., and Selamat, A. (2014). Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing, 22, 11-27.

Idris, I., Selamat, A., and Omatu, S. (2014). Hybrid email spam detection model with negative selection algorithm and differential evolution. Engineering Applications of Artificial Intelligence, 28, 97-110.

Imran, M., Hashim, R., and Khalid, N. E. A. (2012). Opposition based Particle Swarm Optimization with student T mutation (OSTPSO). Paper presented at the 4th Conference on Data Mining and Optimization (DMO), 2012 80-85.

Islam, R., and Yang, X. (2010, 25-27 Aug. 2010). Email classification using data reduction method. Paper presented at the 5th International ICST Conference on Communications and Networking in China (CHINACOM), 2010 1-5.

Ismaila Idris, A. S. (2012). Optimized spam classification approach with negative selection algorithm Journal of Theoretical and Applied Information Technology Vol. 39 (No.1 ), 22-31.

Jacob, E. (2013). A soft computing model for data clustering and application to gene grouping.

Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University.

Ji, Z., and Dasgupata, D. (2004). Augmented negative selection algorithm with variable-coverage detectors, 1081-1088 Vol. 1081.

Jia, D., Duan, X., and Khan, M. K. (2013). An Efficient Binary Differential Evolution with Parameter Adaptation. International Journal of Computational Intelligence Systems, 6(2), 328-336.

Jin, Q., and Ming, M. (2011). A method to construct self set for IDS based on negative selection algorithm, 1051-1053.

Jindal, N., and Liu, B. (2007). Review spam detection, 1189-1190.

Jun, Z., Zhi-Hui, Z., Ying, L., Ni, C., Yue-Jiao, G., Jing-hui, Z., et al. (2011). Evolutionary Computation Meets Machine Learning: A Survey. Computational Intelligence Magazine, IEEE, 6(4), 68-75.

Kamiya, A., Ovaska, S. J., Roy, R., and Kobayashi, S. (2005). Fusion of soft computing and hard computing for large-scale plants: a general model. Applied Soft Computing, 5(3), 265-279.

Kennedy, J. (2010). Particle swarm optimization. In Encyclopedia of Machine Learning (pp. 760-766): Springer.

Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. Paper presented at the IEEE International Conference on Neural Networks, 1995. Proceedings.,, 1942-1948.

Kennedy, J., and Eberhart, R. C. (1997). A discrete binary version of the particle swarm algorithm. Paper presented at the IEEE International Conference on Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 4104-4108.

Kennedy, J. F., Kennedy, J., and Eberhart, R. C. (2001). Swarm intelligence: Morgan Kaufmann.

Kesharwani, Y., and Lade, S. (2013). Spam Mail Filtering Through Data Mining Approach–A Comparative Performance Analysis. International Journal of Engineering, 2(9).

Khare, A., and Rangnekar, S. (2012). Particle Swarm Optimization: A Review. Applied Soft Computing.

Khatri, S., and Emmanuel, M. (2013). Review on Classification Algorithms in Email Domain.

Khoshgoftaar, T. M., Fazelpour, A., Wang, H., and Wald, R. (2013a). A survey of stability analysis of feature subset selection techniques. Paper presented at the IEEE 14th International Conference on Information Reuse and Integration (IRI), 2013 424-431.

Khoshgoftaar, T. M., Fazelpour, A., Wang, H., and Wald, R. (2013b). A survey of stability analysis of feature subset selection techniques. Paper presented at the IEEE 14th International Conference on Information Reuse and Integration (IRI), 2013, 424-431.

Khushaba Rami, N., and Ahmed, A.-A. (2009). Feature Subset Selection Using Differential Evolution.

Khushaba, R. N., Al-Ani, A., and Al-Jumaily, A. (2008). Differential evolution based feature subset selection. Paper presented at the 19th International Conference on Pattern Recognition, 2008. ICPR 2008. , 1-4.

Khushaba, R. N., Al-Ani, A., and Al-Jumaily, A. (2011). Feature subset selection using differential evolution and a statistical repair mechanism. Expert Systems with Applications, 38(9), 11515-11526.

Kim, H., and Chen, S.-S. (2009). Associative Naïve Bayes classifier: Automated linking of gene ontology to medline documents. Pattern Recognition, 42(9), 1777-1785.

Kokash, N. (2005). An introduction to heuristic algorithms. Department of Informatics and Telecommunications.

Kuang, B. Q., Lin, P. Y., Huang, P. J., Zhang, J. F., and Liang, G. Q. (2014). Spam Filter Based on Multiple Classifiers Combinational Model. Paper presented at the Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 1, 689-698.

Kumar, R. K., Poonkuzhali, G., and Sudhakar, P. (2012). Comparative Study on Email Spam Classifier using Data Mining Techniques. Paper presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists.

Lai, C.-C., Wu, C.-H., and Tsai, M.-C. (2009a). Feature selection using particle swarm optimization with application in spam filtering. Int J Innov Comput Inf Control, 5, 423-432.

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

Lai, G.-H., Chen, C.-M., Laih, C.-S., and Chen, T. (2009b). A collaborative anti-spam system. Expert Systems with Applications, 36(3), 6645-6653.

Langin, C., and Rahimi, S. (2010). Soft computing in intrusion detection: the state of the art. Journal of Ambient Intelligence and Humanized Computing, 1(2), 133-145.

Lee, C.-H., and Yang, H.-C. (2009). Construction of supervised and unsupervised learning systems for multilingual text categorization. Expert Systems with Applications, 36(2, Part 1), 2400-2410.

Liang, T., and Yu, Q. (2012, 2-4 Nov. 2012). Spam Feature Selection Based on the Improved Mutual Information Algorithm. Paper presented at the Fourth

International Conference on Multimedia Information Networking and Security (MINES), 2012, 67-70.

Lim, K. S., Buyamin, S., and Ibrahim, Z. (2011). Convergence and Diversity Measurement for Vector Evaluated Particle Swarm Optimization Based on ZDT Test Problems. Paper presented at the Modelling Symposium (AMS), 2011 Fifth Asia, 32-36.

Lin, S.-W., Ying, K.-C., Chen, S.-C., and Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications, 35(4), 1817-1824.

Liu, D., Qian, H., Dai, G., and Zhang, Z. (2013). An iterative SVM approach to feature selection and classification in high-dimensional datasets. Pattern Recognition, 46(9), 2531-2537.

Long, X., Cleveland, W. L., and Yao, Y. L. (2011). Methods and systems for identifying and localizing objects based on features of the objects that are mapped to a vector: Google Patents.

Luo, L., and Li, L. (2014). Defining and Evaluating Classification Algorithm for High-Dimensional Data Based on Latent Topics. PloS one, 9(1), e82119.

Maali, Y., and Al-Jumaily, A. (2013). Self-advising support vector machine. Knowledge-Based Systems, 52(0), 214-222.

Madkour, A., Hefni, T., Hefny, A., and Refaat, K. S. (2008). Using semantic features to detect spamming in social bookmarking systems. Paper presented at the Proc. of ECML PKDD Discovery Challenge Workshop, 55-62.

Magdalena, L. (2010). What is soft computing? revisiting possible answers. International Journal of Computational Intelligence Systems, 3(2), 148-159.

Maldonado, S., and L'Huillier, G. (2013). SVM-Based Feature Selection and Classification for Email Filtering. In Pattern Recognition-Applications and Methods (pp. 135-148): Springer.

Manjusha, K., and Kumar, R. (2010). Spam Mail Classification Using Combined Approach of Bayesian and Neural Network, 145-149.

Mark-Hopkins, Erik-Reeber , George-Forman, and Jaap-Suermondt. (1999).SpambaseDataset.ftp://ftp.ics.uci.edu/pub/machinelearningdatabases/spambase/.

Marsono, M. N. (2007). Towards Improving E-mail Content Classification for Spam Control: Architecture, Abstraction, and Strategies. PhD Thesis -University of Victoria.

Méndez, J. R., Cid, I., Glez-Peña, D., Rocha, M., and Fdez-Riverola, F. (2008). A comparative impact study of attribute selection techniques on Naïve Bayes spam filters. In Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects (pp. 213-227): Springer.

Mendez, J. R., Fdez-Riverola, F., Diaz, F., Iglesias, E. L., and Corchado, J. M. (2006). A comparative performance study of feature selection methods for the anti-spam filtering domain. In Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining (pp. 106-120): Springer.

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.

Mohammed M Mazid, A. B. M. S. A., Kevin S Tickle (2010). Improved C4.5 Algorithm for Rule Based Classification RECENT ADVANCES in ARTIFICIAL INTELLIGENCE, KNOWLEDGE ENGINEERING and DATA BASES.

Morariu, D., Vintan, L., and Tresp, V. (2006). Evolutionary feature selection for text documents using the SVM, 215-221.

Naksomboon, S., Charnsripinyo, C., and Wattanapongsakorn, N. (2010, 11-13 May 2010). Considering behavior of sender in spam mail detection. Paper presented at the 6th International Conference on Networked Computing (INC), 2010 1-5.

Natarajan, A., Subramanian, S., and Premalatha, K. (2012). Optimized Bin Bloom Filter for Spam Filtering using Particle Swarm Optimization. European Journal of Scientific Research, 68(2), 199-213.

Nazirova, S. (2012). Anti-Spam Software for Detecting Information Attacks. International Journal of Intelligent Systems and Applications (IJISA), 4(10), 25.

Noble, W. S. (2006). What is a support vector machine? Nature biotechnology, 24(12), 1565-1567.

Nwankwor, E., Nagar, A., and Reid, D. (2013). Hybrid differential evolution and particle swarm optimization for optimal well placement. Computational Geosciences, 1-20.

Oda, T., and White, T. (2003). Increasing the accuracy of a spam-detecting artificial immune system, 390-396 Vol. 391.

Omar, N., Jusoh, F., Ibrahim, R., and Othman, M. (2013). Review of Feature Selection for Solving Classification Problems. Journal of Information System Research and Innovation, 3.

Omran, M. G. (2009). Using opposition-based learning with particle swarm optimization and barebones differential evolution. Particle Swarm Optimization, InTech Education and Publishing.

Ovaska, S. J., Dote, Y., Furuhashi, T., Kamiya, A., and VanLandingham, H. F. (1999, 1999). Fusion of soft computing and hard computing techniques: a review of applications. Paper presented at the IEEE International Conference on Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999, 370-375 vol.371.

Parimala, R., and Nallaswamy, R. (2012). Feature Selection using a Novel Particle Swarm Optimization and It’s Variants. International Journal of Information Technology and Computer Science (IJITCS), 4(5), 16.

Pearson, K. (1920). Notes on the history of correlation. Biometrika, 13(1), 25-45.

Pour, A. N., Kholghi, R., and Roudsari, S. B. (2012). Minimizing the Time of Spam Mail Detection by Relocating Filtering System to the Sender Mail Server. International Journal, 4.

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Qin, A., and Li, X. (2013). Differential evolution on the CEC-2013 single-objective continuous optimization testbed. Paper presented at the IEEE Congress on Evolutionary Computation (CEC), 2013, 1099-1106.

Qin, A. K., and Suganthan, P. N. (2005). Self-adaptive differential evolution algorithm for numerical optimization. Paper presented at the The 2005 IEEE Congress on Evolutionary Computation, 2005. , 1785-1791.

R.Deepa Lakshmi, N. R. (2010). Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools (IJCSE) International Journal on Computer Science and Engineering Vol. 02, (No. 08), 2760-2766.

Rahnamayan, S., Tizhoosh, H. R., and Salama, M. M. (2006). Opposition-based differential evolution algorithms. Paper presented at the IEEE Congress on Evolutionary Computation, 2006. CEC 2006. , 2010-2017.

Rahnamayan, S., Tizhoosh, H. R., and Salama, M. M. (2007). Opposition-based differential evolution (ODE) with variable jumping rate. Paper presented at the IEEE Symposium on Foundations of Computational Intelligence, 2007. FOCI 2007. , 81-88.

Rahnamayan, S., Tizhoosh, H. R., and Salama, M. M. (2008). Opposition-based differential evolution. Evolutionary Computation, IEEE Transactions on, 12(1), 64-79.

Rakse, S. K., and Shukla, S. (2010). Spam Classification using new kernel function in Support Vector Machine. International Journal on Computer Science and Engineering, 2(5), 1819.

Saad, O., Darwish, A., and Faraj, R. (2012). A survey of machine learning techniques for Spam filtering. IJCSNS, 12(2), 66.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

Salcedo-Campos, F., Díaz-Verdejo, J., and García-Teodoro, P. (2012). Segmental parameterisation and statistical modelling of e-mail headers for spam detection. Information Sciences, 195(0), 45-61.

Salehi, S., and Selamat, A. (2011). Hybrid simple artificial immune system (SAIS) and particle swarm optimization (PSO) for spam detection, 124-129.

Sanasam Ranbir Singh, H. A. M., Timothy A. Gonsalves. (2010). Feature Selection for Text Classification Based on Gini Coeffcient of Inequality. Journal of Machine Learning Research, volume 10 , FSDM 2010 Proceedings(10), 76-85.

Sathawane, M. K. (2013). A Robust Spam Detection System using a collaborative approach with an E-Mail Abstraction Scheme and Spam Tree Data Structure. International Journal Of Computer Science And Applications, 6(2).

Schneider, K. M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering, 307-314.

Selamat, A., and Omatu, S. (2004). Web page feature selection and classification using neural networks. Information Sciences, 158, 69-88.

Sevkli, M., Mamedsaidov, R., and Camci, F. (2012). A Novel Discrete Particle Swarm Optimization for P-Median Problem. Journal of King Saud University-Engineering Sciences.

Shahzad, F., Masood, S., and Khan, N. K. (2013). Probabilistic opposition-based particle swarm optimization with velocity clamping. Knowledge and Information Systems, 1-35.

Shang, C., Li, M., Feng, S., Jiang, Q., and Fan, J. (2013). Feature selection via maximizing global information gain for text classification. Knowledge-Based Systems, 54, 298-309.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang, Z. (2007). A novel feature selection algorithm for text categorization. Expert Systems with Applications, 33(1), 1-5.

Sheehan, K. B. (2001). E- mail survey response rates: A review. Journal of Computer-  Mediated Communication, 6(2), 0-0.

Shi, W.-X., and Huang, Y.-L. (2010). A spam filtering technique based on feature fingerprint. Paper presented at the International Conference on Machine Learning and Cybernetics (ICMLC), 2010 1065-1070.

Soranamageswari, M., and Meena, C. (2010, 9-11 Feb. 2010). Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks. Paper presented at the Second International Conference on Machine Learning and Computing (ICMLC), 2010 101-105.

Sousa, P., Cortez, P., Vaz, R., Rocha, M., and Rio, M. (2013). Email spam detection: a symbiotic feature selection approach fostered by evolutionary computation. International Journal of Information Technology & Decision Making, 12(04), 863-884.

Sousa, P., Machado, A., Rocha, M., Cortez, P., and Rio, M. (2010). A Collaborative Approach for Spam Detection. Paper presented at the Second International Conference on Evolving Internet (INTERNET), 2010 92-97.

Storn, R., and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4), 341-359.

Suebsing, A., and Hiransakolwong, N. (2012). A Novel Technique for Feature Subset Selection Based on Cosine Similarity. Applied Mathematical Sciences, 6(133), 6627-6655.

Sun, J., Zheng, C., Li, X., and Zhou, Y. (2010). Analysis of the distance between two classes for tuning SVM hyperparameters. Neural Networks, IEEE Transactions on, 21(2), 305-318.

Talbi, H., and Batouche, M. (2004). Hybrid particle swarm with differential evolution for multimodal image registration. Paper presented at the IEEE International Conference on Industrial Technology, 2004. IEEE ICIT'04. 2004, 1567-1572.

Tang, G., Pei, J., and Luk, W.-S. (2013). Email mining: tasks, common techniques, and tools. Knowledge and Information Systems, 1-31.

Terri Oda, T. W. (2005). A Spam-Detecting Artificial Immune System. Master Thesis.

Tizhoosh, H. R. (2005). Opposition-based learning: a new scheme for machine intelligence. Paper presented at the international conference on Computational intelligence for modelling, control and automation, 2005 and international conference on intelligent agents, web technologies and internet commerce, , 695-701.

Tizhoosh, H. R. (2006). Opposition-Based Reinforcement Learning. JACIII, 10(4), 578-585.

Trivedi, S. K., and Dey, S. (2013). Effect of feature selection methods on machine learning classifiers for detecting email spams. Paper presented at the Proceedings of the 2013 Research in Adaptive and Convergent Systems, 35-40.

Tušar, T., and Filipič, B. (2007). Differential evolution versus genetic algorithms in multiobjective optimization. Paper presented at the Evolutionary Multi-Criterion Optimization, 257-271.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024-1032.

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. (2011b). mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Information Sciences, 181(20), 4625-4641.

Uysal, A. K., and Gunal, S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems.

Wang, G. (2008). A survey on training algorithms for support vector machine classifiers. Paper presented at the Fourth International Conference on Networked Computing and Advanced Information Management, 2008. NCM'08. , 123-128.

Wang, H., Li, H., Liu, Y., Li, C., and Zeng, S. (2007a). Opposition-based particle swarm algorithm with Cauchy mutation. Paper presented at the IEEE Congress on Evolutionary Computation, 2007. CEC 2007., 4750-4756.

Wang, H., Wu, Z., Rahnamayan, S., Liu, Y., and Ventresca, M. (2011). Enhancing particle swarm optimization using generalized opposition-based learning. Information Sciences, 181(20), 4699-4714.

Wang, J., Wu, X., and Zhang, C. (2005). Support vector machines based on K-means clustering for real-time business intelligence systems. International Journal of Business Intelligence and Data Mining, 1(1), 54-64.

Wang, X., and Cloete, I. (2005). Learning to classify email: a survey, 5716-5719 Vol. 5719.

Wang, X., Yang, J., Teng, X., Xia, W., and Jensen, R. (2007b). Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters, 28(4), 459-471.

Wood, P. (2012). Symantec Intelligence Report: September 2012. (September 2012).

Wu, H., Li, H.-z., Wang, G., Chen, H.-l., and Li, X.-k. (2011a). A Novel Spam Filtering Framework Based on Fuzzy Adaptive Particle Swarm Optimization. Paper presented at the International Conference on Intelligent Computation Technology and Automation (ICICTA), 2011 38-41.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.

Xia, H., Fu, Y., Zhou, J., and Xia, Q. (2013). Intelligent spam filtering for massive short message stream. COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, 32(2), 586-596.

Xingshi, H., Qingqing, Z., Na, S., and Yan, D. (2009, 7-8 Nov. 2009). Feature Selection with Discrete Binary Differential Evolution. Paper presented at the International Conference on Artificial Intelligence and Computational Intelligence, 2009. AICI '09. , 327-330.

Xu, C., Su, B., Cheng, Y., and Pan, W. (2013). An Adaptive Fusion Algorithm for Spam Detection.

Xu, Q., Wang, L., He, B., and Wang, N. (2011). Modified opposition-based differential evolution for function optimization. Journal of Computational Information Systems, 7(5), 1582-1591.

Xue, B., Zhang, M., and Browne, W. N. (2012). Multi-objective particle swarm optimisation (PSO) for feature selection. Paper presented at the Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference, 81-88.

Yevseyeva, I., Basto-Fernandes, V., Ruano-Ordás, D., and Méndez, J. R. (2013). Optimising anti-spam filters with evolutionary algorithms. Expert Systems with Applications.

Youn, S., and McLeod, D. (2007a). A comparative study for email classification. Advances and Innovations in Systems, Computing Sciences and Software Engineering, 387-391.

Youn, S., and McLeod, D. (2007b). A comparative study for email classification. In Advances and Innovations in Systems, Computing Sciences and Software Engineering (pp. 387-391): Springer.

Yu, B., and Xu, Z.-b. (2008). A comparative study for content-based dynamic spam classification using four machine learning algorithms. Knowledge-Based Systems, 21(4), 355-362.

Yu, H., and Kim, S. (2009). SVM tutorial: Classification, regression, and ranking. Handbook of Natural Computing.

Zhang, H., and Li, D. (2007). Naïve Bayes Text Classifier, 708-708.

Zhang, J., Zhan, Z.-h., Lin, Y., Chen, N., Gong, Y.-j., Zhong, J.-h., et al. (2011a). Evolutionary computation meets machine learning: A survey. Computational Intelligence Magazine, IEEE, 6(4), 68-75.

Zhang, Q., YANG, H., WANG, P., and MA, W. (2011b). Fuzzy Clustering based on Semantic Body and its Application in Chinese Spam Filtering. JDCTA: International Journal of Digital Content Technology and its Applications, 5(4), 1-11.

Zhang, W.-J., and Xie, X.-F. (2003). DEPSO: hybrid particle swarm with differential evolution operator. Paper presented at the IEEE International Conference on Systems, Man and Cybernetics, 2003. , 3816-3821.

Zhang, X., Zhou, J., Wang, C., Li, C., and Song, L. (2012). Multi-class support vector machine optimized by inter-cluster distance and self-adaptive deferential evolution. Applied Mathematics and Computation, 218(9), 4973-4987.

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