A COMPARATIVE EVALUATION OF MACHINE LEARNING APPROACHES IN SMS SPAM DETECTION

SABER SALEHI

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

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JULY 2011
To my beloved parents, thank you for always being there for me, supporting me and encouraging me to be the best that I can be.
ACKNOWLEDGMENT

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

Spam detection is a significant problem which is considered by many researchers by various developed strategies. In this study, the popular performance measure is a classification accuracy which deals with false positive, false negative and accuracy. These metrics were evaluated under applying three supervised learning algorithm (Hybrid of Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO), Naive Bayes Classifier (NBC), Enhanced Genetic Algorithm (EGA)) based on classification of SMS contents were evaluated and compared. In this research, SAIS was hybridized by particle swarm optimization (PSO) for optimizing the performance of SAIS for spam filtering. PSO was used with mutation to reinforce the immune system’s searches to find the best class in exemplar for classification. Results were improved using Hybrid SAIS and PSO. The proposed EGA was to achieve the best chromosomes which were grouped by the keywords. Then, the best chromosome with highest fitness value was selected as classifier. Simulated annealing (SA) was used with classical mutation and crossover to reinforce the efficiency of genetic searches. Achieved results represent the enhanced GA is markedly superior to that of a classical GA. These algorithms were trained and tested on a set of 4601 SMS messages in which 1813 were spams and 2788 were non-spams. Results showed that the proposed EGA technique gave better result compare to the hybrid SAIS and PSO and NBC techniques. Results also showed that the proposed EGA technique gave 99.87% accuracy, and the proposed NBC, hybrid of SAIS and PSO techniques gave 97.457% and 88.33% accuracy, respectively.
Pengesanan spam adalah masalah besar yang dianggap oleh ramai peneliti menerusi pelbagai strategi yang dibangunkan. Dalam kajian ini, pengukur prestasi yang popular adalah ketepatan pengelasan yang memyokong keadaan positif palsu, negatif palsu dan ketepatan. Metrik ini telah dinilai melalui pengaplikasian tiga algoritma pembelajaran diselia iaitu (Hybrid of Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO), Naive Bayes Classifier (NBC), Enhanced Genetic Algorithm (EGA)) berdasarkan penilaian dan perbandingan klasifikasi kandungan SMS. Dalam kajian ini, SAIS dihaibritkan oleh particle swarm optimization (PSO) dalam mengoptimumkan prestasi SAIS untuk penapisan spam. PSO beserta kaedah mutasi telah digunakan bagi mengukuhkan carian sistem imun dalam mencari kelas yang terbaik dalam contoh untuk pengelasan. Dengan menggunakan Hibrid SAIS dan PSO keputusan telah bertambah baik. EGA yang dicadangkan adalah untuk mencapai kromosom terbaik yang dikumpulkan menurut kata kunci. Kemudian, kromosom yang terbaik dengan nilai kecegasan tertinggi dipilih sebagai pengelas. Simulated annealing (SA) telah digunakan dengan mutasi klasik dan crossover untuk mengukuhkan kecekapan carian genetik. Keputusan yang mewakili GA yang telah dipertingkatkan menunjukkan keputusan yang ketara lebih tinggi daripada GA klasik. Algoritma ini telah dilatih dan diuji ke atas set yang mengandungi 4601 SMS yang mana 1813 daripadanya adalah spam dan 2788 bukan spam. Keputusan menunjukkan bahawa teknik EGA yang dicadangkan memberikan hasil yang lebih baik berbanding dengan SAIS Hibrid, PSO dan teknik NBC. Keputusan juga menunjukkan bahawa teknik EGA yang dicadangkan memberi 99.87% ketepatan, dan NBC yang dicadangkan iaitu Hibrid SAIS dan teknik PSO masing-masing memberikan ketepatan 97.457% dan 88.33%.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
<td></td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
<td></td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xi</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xii</td>
<td></td>
</tr>
</tbody>
</table>

1 INTRODUCTION 1

1.1 Introduction 1
1.2 Problem Background 3
  1.2.1 Self Learning and Self Adptability of Naive Bayes 3
  1.2.2 Mature Convergence in Genetic Algorithm 4
  1.2.3 Deficiency in Optimization Methods 5
  1.2.4 Reduction the Dimentionality 5
1.3 Problem Statement 6
1.4 Project Aim 7
1.5 Objectives 7
1.6 Dissertation Scope 8
1.7 Thesis Contribution 8
1.7 Thesis Overview 9
1.8 Summary 10
2 LITERATURE REVIEW

2.1 Introduction 11
2.2 Spam 12
2.3 Phishing 15
2.4 SMS Spam Filtering 17
2.5 WhiteList, BlackList, GreyList 17
  2.5.1 WhiteList 18
  2.5.2 Blacklist 19
  2.5.3 Greylist 19
2.6 Pre-acceptance and Post-acceptance Responses 20
  2.6.1 Pre-acceptance Response 21
  2.6.2 Post-acceptance Response 22
2.7 Evaluation Criteria 22
  2.7.1 False positive and False negative 23
  2.7.2 Spam Recall 24
  2.7.3 Spam Precision 24
  2.7.4 Accuracy 25
  2.7.5 F_B Measure 26
2.8 Machine Learning Methods 26
  2.8.1 Inductive Learning 28
  2.8.2 Deductive Learning 29
2.9 Supervised and Unsupervised Learning 30
2.10 Swarm Intelligence 31
  2.10.1 Ant Colony Optimization (ACO) 33
  2.10.2 Particle Swarm Optimization (PSO) 35
    2.10.2.1 Standard PSO 36
    2.10.2.2 Clonal PSO 37
    2.10.2.3 Random Black Hole Particle Swarm Optimization (RBH-PSO) 38
2.11 Background Immunology 38
2.12 Artificial Immune System (AIS) 40
  2.12.1 Artificial Immune Recognition System (AIRS) 41
  2.12.2 CLONALG 44
  2.12.3 Immunos Algorithm 46
  2.12.4 Simple Artificial Immune System (SAIS) 47
2.13 Naive Bayes Classifier (NBC) 47
2.14 Genetic Algorithm (GA)  
2.15 Simulated Annealing (SA)  
2.16 Algorithms Functionalities  
2.17 Summary  

3  
RESEARCH METHODOLOGY  
3.1 Introduction  
3.2 Phases of Research  
3.2.1 Phase 1: To study the Existing Methods in Artificial Immune System (AIS) and Optimization Methods to Solve the Convergence Problem in GA for Classification in Spam Detection  
3.2.2 Phase 2: To Formulate and Apply Classical GA by Simulated Annealing (SA), Simple Artificial Immune System (SAIS), Hybrid SAIS and PSO and Naive Bayes Classifier Algorithm for Spam Detection  
3.2.2.1 Apply Simple Artificial Immune System (SAIS)  
3.2.2.2 Develop Hybrid Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO)  
3.2.2.3 Apply Genetic Algorithm (GA)  
3.2.2.4 Develop Enhanced Genetic Algorithm (EGA)  
3.2.2.5 Naive Bayes Classifier (NBC)  
3.2.3 Phase 3: Calculating the Accuracy, False Positive, False Negative for each Algorithm  
3.2.4 Phase 4: Analyze Effectiveness of Static Optimization Algorithms (Enhanced GA, GA, Hybrid SAIS and PSO, SAIS and Naive Bayes) in Detecting Spam  
3.2.5 Phase 5: Compare the Performance of Hybrid SAIS and PSO, Enhanced GA and Naive Bayes According to Achieved Accuracy  
3.3 Summary  

4  
SIMULATION RESULT  
4.1 Introduction  
4.2 Data Set Information
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>Enhanced Genetic Algorithm (EGA) Techniques</td>
<td>91</td>
</tr>
<tr>
<td>4.4</td>
<td>Naive Bayes Technique</td>
<td>96</td>
</tr>
<tr>
<td>4.5</td>
<td>Hybrid Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO)</td>
<td>97</td>
</tr>
<tr>
<td>4.6</td>
<td>Experimental Results</td>
<td>101</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion on the Result</td>
<td>102</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary</td>
<td>104</td>
</tr>
</tbody>
</table>

**Conclusion**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Conclusion</td>
<td>105</td>
</tr>
<tr>
<td>5.2</td>
<td>Research Contributions</td>
<td>106</td>
</tr>
<tr>
<td>5.3</td>
<td>Future Works</td>
<td>107</td>
</tr>
</tbody>
</table>

**REFERENCES**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td>108</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Types of offers made via spam in a random sample of 1000 spam messages. Source: “False Claims in Spam,” FTC Division of Marketing Practices (April 2003)</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Genetic algorithm, naïve Bayes and simple artificial immune systems’ functionalities</td>
<td>53</td>
</tr>
<tr>
<td>4.1</td>
<td>Data set information</td>
<td>90</td>
</tr>
<tr>
<td>4.2</td>
<td>Calculating group weight from spam corpus</td>
<td>91</td>
</tr>
<tr>
<td>4.3</td>
<td>Testing result for 1000 generation using classical GA</td>
<td>93</td>
</tr>
<tr>
<td>4.4</td>
<td>Testing result for 1000 generation using enhanced GA</td>
<td>94</td>
</tr>
<tr>
<td>4.5</td>
<td>Testing result for SAIS</td>
<td>99</td>
</tr>
<tr>
<td>4.6</td>
<td>Testing result for hybrid SAIS and PSO</td>
<td>99</td>
</tr>
<tr>
<td>4.7</td>
<td>Testing result for 1000 generation using enhanced GA</td>
<td>101</td>
</tr>
<tr>
<td>4.8</td>
<td>Testing result in Naive Bayes</td>
<td>102</td>
</tr>
<tr>
<td>4.9</td>
<td>Testing result for hybrid SAIS and PSO</td>
<td>102</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Different types of spam in the media</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Types of offers made via spam in a random sample of 1000 spam messages. Source: “False Claims in Spam,” FTC Division of Marketing Practices (2003)</td>
<td>15</td>
</tr>
<tr>
<td>2.3</td>
<td>A phishing website that present as eBay</td>
<td>16</td>
</tr>
<tr>
<td>2.4</td>
<td>Current technical initiatives for fighting spam and phishing (Shalendra Chhabra, 2005)</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>False positive and false negative categorization</td>
<td>23</td>
</tr>
<tr>
<td>2.6</td>
<td>Lifecycle overview of AIRS algorithm</td>
<td>42</td>
</tr>
<tr>
<td>2.7</td>
<td>Overview of CLONALG Algorithm</td>
<td>45</td>
</tr>
<tr>
<td>2.8</td>
<td>Overview of Immunos algorithm</td>
<td>46</td>
</tr>
<tr>
<td>2.9</td>
<td>Naive Bayes Classifier, where the predictive features are conditionally independent given the class attribute (C)</td>
<td>49</td>
</tr>
<tr>
<td>2.10</td>
<td>A simplified flow chart of GA process</td>
<td>50</td>
</tr>
<tr>
<td>3.1</td>
<td>The main phases of research</td>
<td>58</td>
</tr>
<tr>
<td>3.2</td>
<td>Hybrid SAIS and PSO, Naïve Bayes and enhanced GA procedure for classification in spam detection</td>
<td>61</td>
</tr>
<tr>
<td>3.3</td>
<td>Mathematical model of normalization</td>
<td>62</td>
</tr>
<tr>
<td>3.4</td>
<td>SAIS flowchart</td>
<td>64</td>
</tr>
<tr>
<td>3.5</td>
<td>Mathematical model of Euclidean function</td>
<td>65</td>
</tr>
<tr>
<td>3.6</td>
<td>Hybrid SAIS and PSO framework</td>
<td>67</td>
</tr>
<tr>
<td>3.7</td>
<td>pi lists initialization mathematical model</td>
<td>68</td>
</tr>
<tr>
<td>3.8</td>
<td>Update pi and pg mathematical model</td>
<td>69</td>
</tr>
<tr>
<td>3.9</td>
<td>Update velocity and position mathematical model</td>
<td>70</td>
</tr>
<tr>
<td>3.10</td>
<td>Mathematical model of chromosomes generation</td>
<td>72</td>
</tr>
</tbody>
</table>
3.11 Mathematical model of fitness function 74
3.12 Mathematical model of calculating weight of groups 75
3.13 Classical GA flowchart 76
3.14 Enhanced GA framework 78
3.15 Simulated annealing mathematical model 80
3.16 Naïve Bayes classifier framework 82
3.17 Mathematical model of average frequency of each attributes on total spam or legitimate list 83
3.18 Mathematical model of average probability of each attributes on total spam or legitimate list 84
3.19 Mathematical model of average probability of each attributes which occurs in spam category 85
3.20 Performance Measure 87
4.1 Accuracy comparison between EGA and GA 94
4.2 False positive comparison between EGA and GA 95
4.3 False negative comparison between EGA and GA 95
4.4 Accuracy comparison between hybrid SAIS and PSO and SAIS 100
4.5 FP comparison between hybrid SAIS and PSO and SAIS 100
4.6 Accuracy comparisons between EGA, NBC, hybrid SAIS and PSO 103
4.7 False positive (FP) comparisons between EGA, NBC, hybrid SAIS and PSO 103
4.8 False negative comparisons between EGA, NBC, hybrid SAIS and PSO 104
CHAPTER 1

INTRODUCTION

1.1 Introduction

All wireless networks are comprised of independent computing nodes, have security problems, potentially. In wireless networks that are sorted by traditional mainframe computers without checking by central control, users capable to disturb the network with unlike intentions such as decreasing the wireless networks’ confidentiality or integrity for accessing the data, using the capabilities of your network components for illegal or unsavory purposes or interfering the wireless network users with the reasonable activities. According to these unlike intentions, with the coming of internet and its protocols to control wide area connectivity, big and enterprise companies decided to protect their wireless networks, data, and applications against intrusive attacks and other threats by establishing perimeter defenses such as firewalls, VPNs, antivirus systems, intrusion prevention and intrusion detection systems (John, 2006).
Wireless sensor networks deploy rapidly. This kind of network is flexible and self-organizing system. This subject incurs to decrease the maintenance and deployment cost. Therefore, many application scopes such as health, home, military or environmental encourage using this kind of network. One of the significant challenges in many of these application areas is security. Sensor nodes must be far from the enemy because whenever the sensor network is reachable; sensor nodes could be destroyed and disturbed by the attackers. Cellular phone is one of the main areas in wireless network.

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

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

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

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

1.2.1 Self Learning and Self Adaptability of Naïve Bayes

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

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

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

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

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

1.2.4 Reduction the Dimensionality

Feature selection is an important method for improving the efficiency and accuracy of text categorization algorithms by removing redundant and irrelevant terms from the corpus. Extensive researches have been done to improve the performance of spam detection by reduction to the dimensionality of features space. Almedia et al. (2009) compared the performance of most popular methods used as term selection techniques, such as document frequency (DF), information gain (IG), mutual information (MI), $\chi^2$ statistic, and odds ratio (OR) used for reducing the dimensionality of the term space with four well-known different versions of Naive Bayes spam filter. Regarding term selection techniques, Almedia et al. have found IG and $\chi^2$ statistic, most effective in aggressive term removal without losing categorization accuracy. On one hand, DF and
OR also usually provided an improvement on the filter’s performance. On the other hand, the employment of MI offered poor results which frequently worsened the classifier’s performance. But, in this dissertation the feature space is not that large for reduction so feature selection methods were not used in this dissertation.

1.3 Problem Statement

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

- i. How we can find the optimization algorithms and their functionalities to understand the main concepts in spam detection?
- ii. How we can adopt the naïve Bayes classifier, GA, enhanced GA, SAIS and hybrid SAIS and PSO for spam detection in SMS exchanging?
- iii. How we can find the efficiency of naïve Bayes classifier, GA, enhanced GA, SAIS and hybrid SAIS and PSO in spam detection?
1.4 Dissertation Aim

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

1.5 Objectives

This research follows three objectives:

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

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

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

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

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

1.8 Thesis Contribution

The contributions of this research are as follow:

- To develop Enhanced GA and hybrid of SAIS and PSO.
- To apply Enhanced GA, hybrid of SAIS and PSO, Naïve Bayes, GA and SAIS for spam detection.
• Analytical comparison of Enhanced GA, hybrid of SAIS and PSO, Naïve Bayes, GA and SAIS and the impact of PSO and simulated annealing (SA) on SAIS and GA.

1.9 Thesis Overview

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

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

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

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


De Castro, L., and Von Zuben, F. (2002). Learning and optimization using the clonal
Jerome, H., Carter, M. (2000). The Immune System as a Model for Pattern Recognition and Classification. Section of Medical Informatics, Division of General Internal Medicine, University of Alabama, 123-130.


in Communication and Computing, 604-609.


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


