ARTIFICIAL NEURAL NETWORK FOR SPAM FILTERING: CLASSIFICATION AND ANALYSIS OF SPAM AND NON-SPAM USING MULTI-LAYER PERCEPTRON THROUGH BACK-PROPAGATION LEARNING ALGORITHM AND RADIAL BASIS FUNCTION NETWORK

GAN SWAT NGO

This project is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science with Honours (Cognitive Science)

Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak
2004
ACKNOWLEDGEMENT

I would like to express my appreciation to my supervisor, Mr. Syafiq Fikri Abdullah @ Mr. Lee Nung Kion for his supervision, encouragement and patient during the time completing this thesis. Thank you for giving me a chance to learn from this project and thus updated my knowledge.

I also like to express my gratitude to all lectures, who are involved directly or indirectly in this thesis. They are Dr. Tay Yong Haur, Dr. Lim Chee Peng, Mr. Bong Chih How, Mr. Then Hang Hui and Mr. Lee Beng Yong. Their valuable idea, suggestion and invaluable knowledge about neural network and text classification has helped me a lot throughout this project, especially when faced some tricky problems.

Other than that, I would like to thank Mr. Loo Wai Chong, Mr. Hong Yiu Chuan, Mr. Tang Weng Chin and Mr. Nelson Aloysius for their ideas in implementing source code as well as provided neural network information.

Last but not least, I would like to express my thanks to my family, and fellow friends, especially Ms. Lim Yar Fen, Ms. Tan Siang Na and Mr. Lau Guan Sit for their help, understanding and support until this thesis submitted.

Thank you also for my coursemates who giving their helping hand to me for this thesis and other problems among these three years. Thank you very much to all.
TABLE OF CONTENTS

Acknowledgement iii
Table of Contents iv
List of Figures ix
List of Tables xii
Abstract xiv
Abstrak xv

1. Introduction 1
  1.0 Background 1
  1.1 Problem Statements 2
  1.2 Objectives and Scope 4
  1.3 Significance of Research 5
  1.4 Synopsis 5

2. Literature Review 7
  2.0 Preliminaries 7
  2.1 Background of Artificial Neural Network 7
  2.2 What is Spam? 9
  2.3 Spam Filtering 10
    2.3.1 Content Filtering 12
    2.3.2 Image Filtering 13
  2.4 Cased-Based Spam Filtering 14
2.5 Current Technology 16
2.6 Existing Applications of Spam Filtering 17
2.7 Back-propagation Neural Net (BPNN) 18
  2.7.1 The Learning Algorithm of Back-propagation 20
  2.7.2 The Advantages and Drawbacks of MLP 23
2.8 Naïve Bayesian Network 24
  2.8.1 Bayes’ Rule 24
  2.8.2 Classification of E-mail Messages 28
    2.8.2.1 Corpus preprocessing 28
    2.8.2.2 Naïve Bayesian classification 29
2.9 Radial Basis Function Network (RBF network) 32
  2.9.1 Radial Basis Function Neuron Model 34
  2.9.2 Types of Radial Basis Functions (RBF) 35
    2.9.2.1 Gaussian 35
  2.9.3 Radial Basis Function Networks Training Algorithms 36
2.10 The Advantages and Drawbacks of RBF 39
2.11 Summary 39

3. Methodology and System Design 41
  3.0 Introduction 41
  3.1 Data Collection 41
  3.2 Data Pre-processing 42
    3.2.1 Feature Extraction / Selection 43
    3.2.2 Preprocessing 44
3.2.3 Transformation

3.2.3.1 Vector Encoding

3.2.3.2 Scaling

3.3. Creation of Neural Network Models

3.3.1 Back-propagation Network

3.3.1.1 NEWFF Model

3.3.1.2 NEWFF Algorithm

3.3.1.3 Back-propagation Network Operation

3.3.1.4 Back-propagation Training Phase

3.3.1.5 Back-propagation Validation and Testing Phase

3.3.2 Radial Basis Function Network

3.3.2.1 NEWRB Model

3.3.2.2 NEWRB Algorithm

3.4 Simulation of Neural Network

3.4.1 Number of Hidden Nodes

3.4.2 Number of Epochs

3.4.3 Training Function

3.4.4 Learning Rate

3.4.5 Spread Constant

3.5 Performance Analysis

3.6 System Design

3.6.1 Determining the Frequency Selection for Each Message

3.6.2 Program Implementation of MLP

3.6.3 Program Implementation of RBF
### 4. Implementation and Performance Analysis

4.0 Introduction

4.1 Data Collection and Data Pre-processing

4.2 Creating and Training a Neural Network Model
   - 4.2.1 Multi-layer Perceptron
   - 4.2.2 Radial Basis Function Network

4.3 Simulation and Performance Analysis
   - 4.3.1 Experiment I – Classification using MLP
     - 4.3.1.1 Number of Hidden Nodes
       - 4.3.1.1.1 One-hidden layer
       - 4.3.2.2 Two-hidden layers
     - 4.3.1.2 Number of Epochs
   - 4.3.1.3 Training Function
   - 4.3.1.4 Learning rate
   - 4.3.1.5 MLP without Validation

4.3.2 Experiment II – Classification using RBF
   - 4.3.2.1 RBF Results

4.5 Summary

### 5. Conclusion and Recommendation

5.1 Introduction

5.2 Achievement
5.3 Conclusion

5.4 Future Work and Recommendation

5.4.1 Increasing the Training Set

5.4.2 Enhancing Pre-processing

5.4.3 Using Other Learning Algorithms

5.4.4 Unsupervised Training Method

5.4.5 Filter Based on Pair Words

5.4.6 Content Abstraction

5.4.7 Multi-Layer Spam Detection

5.4.8 On-line Spam Filtering

5.5 Summary

References

Appendix A

Appendix B
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>8</td>
</tr>
<tr>
<td>A biological neuron structure</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>9</td>
</tr>
<tr>
<td>An idealized model of neuron and network</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>11</td>
</tr>
<tr>
<td>Alternatives for spam filtering in Internet e-mail</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>15</td>
</tr>
<tr>
<td>The components of a Case-Based spam filtering system</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>19</td>
</tr>
<tr>
<td>Typical three-layer back-propagation networks</td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td>23</td>
</tr>
<tr>
<td>The idea of spam filtering using BP algorithm</td>
<td></td>
</tr>
<tr>
<td>2.7</td>
<td>32</td>
</tr>
<tr>
<td>Bayesian networks corresponding to Naïve Bayesian classifier</td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>34</td>
</tr>
<tr>
<td>One-dimensional radial basis functions</td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>34</td>
</tr>
<tr>
<td>Gaussian with three different standard deviations</td>
<td></td>
</tr>
<tr>
<td>2.10</td>
<td>34</td>
</tr>
<tr>
<td>Radial basis function network with $R$ inputs</td>
<td></td>
</tr>
<tr>
<td>2.11</td>
<td>35</td>
</tr>
<tr>
<td>Radial basis transfer function</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.12
RBF network in pattern classification

Figure 3.1
The sequence processes of spam classification

Figure 3.2
Interface of messages' pre-processing using Visual Basic program

Figure 3.3
An example of transformation: vector encoding and scaling before fed into
the ANN models

Figure 3.4
NEWFF architecture - the two-layer tansig / purelin network

Figure 3.5
General operation of back-propagation network algorithm

Figure 3.6
A hidden radial basis layer of $S^1$ neurons and an output linear layer of $S^2$ neurons

Figure 3.7
The desired output from the two networks after testing process

Figure 3.8
The pseudo-code for determination of the words' frequencies into scaled values

Figure 3.9
The MLP and RBF networks processes using MATLAB neural network toolbox (GUI)

Figure 3.10
The pseudo-code for implementation of the MLP network
Figure 3.11  
The pseudo-code for implementation of RBF network

Figure 4.1  
An example of graph target versus test set (testing phase) when applied to one-hidden layer

Figure 4.2  
An example of graph target versus test set (109 mails) for testing phase when applied to two-hidden layers

Figure 4.3  
An example of graph target versus test set (109 mails) for testing phase when using TRAINRP as training function in MLP

Figure 4.4  
An example of graph target versus test set (109 mails) for testing phase in RBF
LIST OF TABLES

Table 3.1
Different training functions

Table 4.1
Number of inputs (messages) for MLP network

Table 4.2
Number of inputs (messages) for RBF network

Table 4.3
The results of number of hidden nodes (1 hidden layer) versus average accuracy in MLP

Table 4.4
The results of number of hidden nodes (2 hidden layers) versus average accuracy in MLP

Table 4.5
The results of number of epochs versus accuracy in MLP

Table 4.6
The results of training function versus average accuracy in MLP

Table 4.7
The results of learning rate in MLP

Table 4.8
$SPREAD$ value (1.0 –10.0) versus accuracy in RBF

Table 4.9
$SPREAD$ value (0.1 – 0.9) versus accuracy in RBF

Table 4.10
$SPREAD$ value (0.010 – 0.200) versus accuracy in RBF
| Table 4.11 |  
| SPREAD value (0.051 – 0.59) versus accuracy in RBF | 81 |
ARTIFICIAL NEURAL NETWORK FOR SPAM FILTERING: CLASSIFICATION AND ANALYSIS OF SPAM AND NON-SPAM USING MULTI-LAYER PERCEPTRON THROUGH BACK-PROPAGATION LEARNING ALGORITHM AND RADIAL BASIS FUNCTION NETWORK

Gan Swat Ngo

This study is about spam filtering using Artificial Neural Networks (ANNs). The purpose of this study is to do comparisons and analysis for both multi-layer perceptron (MLP) and radial basis function network (RBF) in order to determine the better network for classifying e-mails correctly. Pre-processing of electronic mails or e-mails is done to extract a series of keywords as input vectors for the networks. The pre-processing is implemented by using Visual Basic language. Dataset that used in this project is downloaded from personal e-mail account. Experiments are carried out for the networks using MATLAB Neural Network Toolbox. In the experiments, few parameters are used to evaluate the networks performance. The study findings show that the most suitable network to classify e-mails is MLP network using two-hidden layer. The performance in term of average accuracy for MLP network is 72.48%. While performance of RBF network is slightly less than MLP network is 67.89%. Even though the MLP network performed better than RBF network but the RBF network still consider good in term of its stability.
ABSTRAK

RANGKAIAN NEURAL BUATAN UNTUK PENAPISAN E-MEL: KLASIFIKASI DAN ANALISIS SPAM DAN BUKAN-SPAM MENGGUNAKAN MULTI-LAPISAN PERCEPTRON DENGAN ALGORITMA PEMBELAJARAN ROMBATAN BALIK DAN RANGKAIAN RADIAL BASIS FUNCTION

Gan Swat Ngo

CHAPTER 1

INTRODUCTION

1.0 Background

Artificial neural networks (ANNs) or neurocomputing is an information-processing system, which applies the idea of biological neural network in its certain performance characteristics, such as parallel execution (Fausett, 1994). Besides, he also said that the applications of ANNs are widely used in signal processing (recovery of telecommunications from faulty software), control system (emulator and controller), pattern recognition (handwritten word recognition), medical field (diagnosis of hepatitis), speed production (NETtalk), speech recognition (recognition of a speaker in communication) as well as business (loan application).

According to Stergiou (n.d.), he suggested that neural network is suitable for prediction or forecasting needs, such as sales forecasting, industrial process control, customer research, data validation, risk management, and target marketing due to its ability in identifying patterns. Moreover, neural networks can apply into textual recognition, such as spam recognition.

Based on the Help menu of MSN Hotmail, spam mail refers to junk mail. It is the Internet's version of unsolicited direct mail. Furthermore, spam is any kind of message or posting that is send to multiple recipients who do not request the e-mail. Spam mail can be multiple postings of the same message from newsgroups. Spammer usually uses designed software to collect e-mail addresses from newsgroups, mailing lists, and e-mail programs. The spammer often uses unprotected server of other companies to distribute e-mail in order to avoid the cost of paying distributed e-mail.
Actually, user can complain to the junk mailer. However, it is not a good idea because as user’s mail is being read, he also activates his e-mail account. Thus, it attracts more and more junk mail in the future. Generally, spammers are regarded as the marketing low life of the Internet, whose business model is based on abusing the open structure of the Internet in order to use other companies’ resources (Ryals & Payne, 2001).

1.1 Problem Statements

Nowadays, Malaysia is moving into technological age and emphasizes knowledge acquisition. The Internet is a powerful tool to achieve this goal. One of the facilities that are provided by the Internet is e-mail account. Through registering an e-mail account, such as http://hotmail.msn.com, http://mail.yahoo.com, http://www.microsoft.com, and so forth, user gains more information rapidly. These services are provided almost free of charge. By using the e-mail account, user can exchange a variety of information from friends, lecturers, employers, employees, leaders, customers, retailers, etc. Moreover, users can communicate with their friends without geographical obstacle. By adding the filtering, user can also chat with friends through using the http://www.microsoft.com. Unfortunately, these facilities have misused for distributing unsolicited / inappropriate messages and documents or known as junk mails.

The spam can be sent with almost no cost to the sender. In fact, others are paid the costs associated with the spam, such as the Internet Service Provider (ISP) and the receiver. Besides, it is difficult to have a legal action against spammers for preventing the receipt of spam within that jurisdiction (Cunningham et al., n.d.). When this situation occurs, user will
face a lot of troubles in receiving mail from others because the size of mail account is limited as well as user cannot send his mail out due to mail traffic.

Moreover, user will waste much time to clean out the mailbox if he does not fix any device or software, which can detect whether the mail is junk mail or real mail. Therefore, the spam filter is needed in order to let the system to check the e-mails before downloading them. In other words, spam is harmful because it utilizes resources for other tasks, such as bandwidth, screen area, disk space, and user’s time. In addition, spam can be disreputable or entire illegal. For instance, various frauds, illegal products, and other inappropriate materials are advertised via spam (Martin, 2002).

Furthermore, user will feel difficult to search his desired e-mails if someone broadcasts unsolicited mass e-mail or newsgroup postings simply because he wants to spread messages. This is referring to the "signal-to-noise ratio". The purpose of spam filter is to help user to keep the Internet useful information readily available and keeps "junk mail" to a minimum level. The main problem of the existing filter’s software is that they cannot be trained and learn instead of fixing a set of filter’s rules. It is tedious and difficult to construct robust rules to detect the naturally changeable junk mail too.

This project applied artificial neural network techniques for spam filtering by using back-propagation neural net (BPNN) and radial basis function (RBF) network to solve the problem of spam recognition. There are few questions need to be investigated in order to know whether both networks can classify e-mails correctly. For instance, “Is any fixed rule use to classify e-mails in both networks?”, “Which one is better in classification based on the accuracy rate counted?” “Is the feature selection an ideal idea to find out the indicative terms
for spam mails after filter out the stop words?”, and “Can neural networks learn the pattern of the e-mail, such as header of e-mails that consist in both spam and ham?”.

1.2 Objectives and Scope

The objective of this project is to classify and make analysis of spam and non-spam through using ANN models, such as back-propagation learning algorithm and radial basis function network. This research focuses pattern classification of e-mail content in order to determine whether it is a spam or a non-spam. When a set of data samples is given, the network will carries out training to learn the pattern of e-mails. The trained network (filter) has to decide on which type of dataset categories (spam versus. non-spam) could be matched most closely when testing with the test set. The test set, which indicates advertisement, business’s information, pornographic issues, etc., will be classified as spam (filtered out). The rest of the mails are classified as ham mails. There are two specific objectives in this project.

a) To implement the ideas of multi-layer perceptron network (back-propagation learning algorithm) and the radial basis function network (clustering and least square learning algorithm) for spam filtering.

b) To evaluate the performance of back-propagation learning algorithm and RBF network models using keywords selection method as well as to quantify their results by statistical measures.

The scope of this study focuses on two specified ANN models as mentioned previously, which are the feed-forward back-propagation network model (BP) and the radial basis function network (RBF). The architectures and learning algorithms of both ANN models in classification mails problem will be investigated. The trained network that obtained
from training phase will be used in testing phase. Then the comparison of both models will be analyzed. Besides, the project also concerns about Naïve Bayesian classifier, which famous applied to spam detection application. However, it is mentioned in theoretically.

1.3 Significance of Research

This project is to implement the classification of spam and non-spam through two neural network paradigms, which are back-propagation network and radial basis function network. Comparison between both paradigms can help designer to recognize the potential of the model and network. Besides, this research is vital to improve existing spam filtering functions in order to protect e-mail’s user from spammers. Thus, it make user facilities e-mail account easier because he does not face any spam mails' problem that stated previously.

1.4 Synopsis

The content of thesis can be summarized as follows. Chapter 1 consists of a brief introduction of this project - neural network for spam filtering, which includes information about artificial neural network and definition of spam. Besides that, there are problem statements, objective of the study and scope of this project. Chapter 2 contains literature review about the spam filtering using variety of methods, such as case-based reasoning (CBR) and description of neural network paradigms in theoretically. Through reading other materials, knowledge about existing systems can be understood. Thus, the prior knowledge is used for this project as well. Sequentially, Chapter 3 discusses about the methodology and system design. Via the description of the methodology, user can understand the flow of the system and use suitable methods for training the spam and non-spam in order to classify the legitimate e-mail and junk
mail. This chapter also concerns about training and simulation processes using MATrix LABoratory (MATLAB) software for different learning paradigms. Chapter 4 discusses about the implementation processes and its performance analysis. Finally, the conclusion and recommendation for current and future research is in the chapter 5.
CHAPTER 2
LITERATURE REVIEW

2.0 Preliminaries

Kulkarni (1993) noted that Artificial Neural Networks (ANNs) model mimic the human brain. Their computing architectures are very different from the traditional computers. They are massively parallel system. ANN models are also known as connectionist models of parallel-distributed processing (PDP) models. According to the DARPA Neural Network Study (1988, AFCEA International Press, p. 60), neural network can define as follow:

"... a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes" (cited in Sarle, 2002).

2.1 Background of Artificial Neural Network

ANN models imitate the idea of human nervous system, which consists of cells called neurons. There are one trillion neurons \((10^{12})\) and \(10^{15}\) synapses interconnections. The functions' of neurons are receiving, processing, and transferring electrochemical signals over the neural pathways that include the brain's communication system (Kulkarni, 1993). Anyway, the brain is the most complex and powerful biological machinery on the earth. Therefore, researchers put more efforts in popularize the ANN models recently.
Based on Fausett (1994), there are three basic parts of the neurons, such as the cell body (soma); axon, and dendrites (see Figure 2.1). The soma supports the functions in the neuron through receiving input signals from the dendrites and synapses. The axon carries the fired signal out of the neuron (output of the signals). The dendrites consist of a very dense fiber-type of structure. Thus, they are able to receive incoming signals from input signals. While the synapses are the primary gateways for neurons' communicate with each other. The learning process plays an important role as the chemical reactions at the synapses are changed.

Generally, ANN is build corresponding to biological neuron. For instance, the inputs resemble to the dendrites, the weights resemble to the synapses, the activation function (neuron $j$) resembles to the soma, and the outputs resemble to the axon (see Figure 2.2).

![Figure 2.1 A biological neuron structure](image)
(Adapted from University of Victoria CENG 420 Artificial Intelligence, n.d.)
2.2 What is Spam?

Cunningham et al., (n.d.) pointed out that the term of “Spam” derives from a Monty Python sketch. It represents a group of Vikings who wish to eat in a restaurant but the menu contains too much of spam (the food) that it is hard to determine what else is available. Spam is unsolicited and unwanted email that is sent in bulk or large mailing lists by a stranger. Usually, it has commercial motive, such as promoting his products or services. However, spam should not include emails looking for employment or positions. He also stated that the definition of spam is based on the receiver or user. Different user defines mails as spam mails differently. Thus, uses individual mails as dataset will strengthen the case for personalized spam filtering. In this study, the chain letter sent by friends, such as legitimate articles, and forwarded messages are classified as non-spam category.

Spam can be considered as a junk email, junk postal mail and junk faxes. They can bring problems to Internet’s user. In comparison between junk email and other junk advertising, junk mail caused a lot of problem. Therefore, filtering mechanisms have been
developed to detect spam. Examples of the spam are forged header details to beat blacklisting, disguised words (e.g. Adult with ‘1’ replaced as a number) and random text to beat signatures based on text hashing.

According to Spring (2001), deleting the spam is a daunting task that is time consuming. Besides, unsolicited commercial e-mail seems to be getting worse. In average user receives three spam messages (business e-mail) per day. This number of spam will swell to 40 in 2003. Based on Ferris Research, user will spend 15 hours for deleting e-mail in year 2003 compared to 2.2 hours in the year 2000. In addition, the cost of the average business in the future will be increase to $400 per in-box contrast $55 today. The worst is spam mails can threaten privacy and bring viruses to user’s system.

Unfortunately, it is difficult to avoid the spam out of mailbox. Spammers are very clever to develop software to fight against anti-spam software makers who are Internet Service Providers, and worthy anti-spam groups, such as Coalition Against Unsolicited Commercial E-mail. The anti-spam groups is a nonprofit group, which is working to get legislation enacted ‘to stop’ the massive flow of spam over the Internet. On the other hands, by sending spam to others, one can make money. It is a ‘dream job’ for some people. "Make Money While You Sleep." said by Spring (2001).

2.3 Spam Filtering

There are two levels of operation the spam filtering for e-mail messages, which are individual user level and enterprise level (see Figure 2.3). Commonly, the individual user refers to a person who working at home, sending and receiving e-mail via an ISP. If the user wants to