

FEATURE SELECTION USING NEURAL NETWORK IN WRITER
IDENTIFICATION DOMAIN

GOH HUI LI



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: FEATURE SELECTION USING NEURAL NETWORK IN
WRITER IDENTIFICATION DOMAIN

SESI PENGAJIAN: 2010/2011

Saya _____ GOH HUI LI _____
(HURUF BESAR)

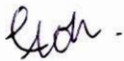
mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila tandakan(/)

_____ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

_____ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

/ _____ TIDAK TERHAD



(TANDATANGAN PENULIS)

Alamat tetap: 43, Taman Bahagia, Bt 36 1/4,
Jalan Johor, 82000 Pontian, Johor.



(TANDATANGAN PENYELIA)

Dr Azah Kamilah Bt Draman@
Muda

Nama Penyelia

Tarikh: 12 JULY 2011

Tarikh: 15/7/11

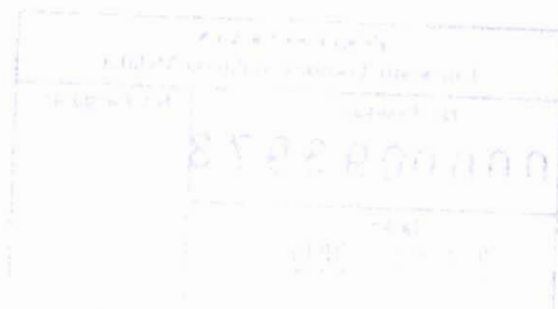
CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

FEATURE SELECTION USING NEURAL NETWORK IN WRITER IDENTIFICATION DOMAIN

GOH HUI LI

The report is submitted in partial fulfillment of the requirements for the
Bachelor of Computer Science (Artificial Intelligence)


FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2011

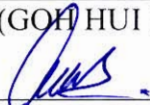


DECLARATION

I hereby declare that this project report entitled
**FEATURE SELECTION USING NEURAL NETWORK IN WRITER
IDENTIFICATION DOMAIN**

is written by me and is my own effort and that no part has been plagiarized
without citation.

STUDENT :  Date: 12 JULY 2011
(GOH HUI LI)

SUPERVISOR :  Date: 15/7/11
(DR AZAH KAMILAH BINTI DRAMAN @ MUDA)

DEDICATION

To my beloved parents, Mr. Goh Hup Chye and Mrs. Lee Bee Giok and my family,
for their expression of love and fully support...

To my supervisor, Dr. Azah Kamilah Binti Draman @Muda, for making it all
worthwhile...

ACKNOWLEDGEMENT

The special thanks go to my helpful supervisor, Dr Azah Kamilah Binti Draman @ Muda. The supervision and support that she gave truly help the smoothness of the project. Her assistances and advices was indeed appreciated and acknowledged.

My grateful thanks also go to both Dr Choo Yun Huoy and Pn. Norashikin. Thanks for helping me when I seek to you whenever I encountered with some problems and point out the negligence so that I will keep on the right track.

Lastly, I would like to thanks my family for giving me their support and encouragements throughout the whole project. Their advices and encouragements had pulled me through all the obstacles and finally completed the project.

ABSTRACT

Handwriting is different for different individuals. However, every single person has their own writing styles which are the unique features that are underlying in their writing. Yet, not every feature is important in identifying the writer. In order to identify these relevant features, feature selection technique is used. Today, feature selection has been used in handwriting aspect using different approaches but not neural networks. In this study, neural network will be used for feature selection in identifying writers in order to classify the handwritten into an appropriate class label by using Neural Network as classifier. Throughout the study, a three layered MLP architecture will be used as the neural net structure and Visual C++ will be used to develop the program for the research. The research is basically referred to the technical paper which entitled "Neural Network Feature Selector", but it has been modified to adapt to this problem situation. As a conclusion, the proposed research in this study has fulfilled the objectives. Feature selection using neural network has performed well in writer identification domain in this study. However, there is still room for improvement to make it better and computation friendly.

TABLE OF CONTENTS

CHAPTER	SUBJECT	PAGE
	DECLARATION	i
	DEDICATION	ii
	ACKNOWLEDGEMENTS	iii
	ABSTRACT	iv
	TABLE OF CONTENTS	v
	LIST OF TABLES	ix
	LIST OF FIGURES	x
	LIST OF ABBREVIATIONS	xii
CHAPTER 1	INTRODUCTION	
	1.1 Project Background	1
	1.2 Problem Statement(s)	2
	1.3 Objectives	2
	1.4 Scope	3
	1.5 Project Significance	3
	1.6 Expected Output	3
	1.7 Conclusion	4

CHAPTER II

LITERATURE REVIEW AND
RESEARCH METHODOLOGY

2.1	Introduction	5
2.2	Facts and findings	
2.2.1	Writer Identification Domain	6
2.2.2	Neural Network	7
2.2.3	Multilayer Perceptron	8
2.2.4	Feature Selection	10
2.2.5	Activation Function	11
2.2.6	Learning Algorithm	12
2.2.7	Weka	14
2.2.8	Existing Techniques	15
2.3	Experimental Settings	
2.3.1	Neural Net Setting	17
2.3.2	Software Used	18
2.4	Research Methodology	
2.4.1	Literature Review	20
2.4.2	Further studies on NN as Feature Selection method	20
2.4.3	Design Feature Selection Algorithm	21
2.4.4	Coding	21
2.4.5	Analysis	21
2.5	Project Schedule and Milestone	21
2.6	Conclusion	22

CHAPTER III	NEURAL NETWORK FEATURE SELECTOR	
3.1	Introduction	23
3.2	Feature Selection using Neural Network	23
3.3	Neural Network Training	24
3.4	Neural Network Feature Selection Algorithm	26
3.5	Flowcharts	28
3.6	Conclusions	47
CHAPTER IV	ANALYSIS	
4.1	Introduction	48
4.2	Problem Analysis	48
4.3	Model Used	49
4.4	Experimental Result	
4.4.1	Result Analysis	50
	4.4.1.1 Network Accuracy	50
	4.4.1.2 Feature Accuracy	51
	4.4.1.3 Feature Selection (Selected Attribute)	52
4.4.2	Program Screenshots	54
4.5	Discussion	
4.5.1	Feature Selection using Neural Networks	55
4.5.2	Feature Selection Algorithm	56
	4.5.1.1 Error Function	56
	4.5.1.2 Feature Selection Process	57
4.6	Conclusions	58

CHAPTER V	PROJECT CONCLUSION	
5.1	Observation on Weaknesses and Strengths	59
5.2	Proposition for Improvement	60
5.3	Contribution	60
5.4	Conclusion	60
REFERENCES		61
BIBLIOGRAPHY		63
APPENDIX		64

LIST OF TABLES

DIAGRAM

TABLE	TITLE	PAGE
2.1	Functions of Layers	8
2.2	Determining number of hidden layers	9
4.1	Network Accuracy	50
4.2	Feature Accuracy (on Training Set)	51
4.3	Feature Accuracy (on Testing Set)	52
4.4	Comparisons of the Results	52
4.5	Comparison of Important Attributes	53

LIST OF FIGURES

DIAGRAM	TITLE	PAGE
Figure 2.1	McCulloch Pits	7
Figure 2.2	Multilayer Perceptron	9
Figure 2.3	Architecture of Backpropagation Network	13
Figure 2.4	Modules of Project Studies	19
Figure 3.1	Neural Net Model of the experiment	24
Figure 3.5.1	MLP structure initialization and data read into arrays	29
Figure 3.5.2	Weight and Bias Initialization	31
Figure 3.5.3	Weight Training	33
Figure 3.5.4	Accuracy Checking	34
Figure 3.5.5	Propagation	36
Figure 3.5.6	Output Rounding	38
Figure 3.5.7	Mean Squared Error Calculation	39
Figure 3.5.8	Error Gradient Calculation	41
Figure 3.5.9	Weight Update	43
Figure 3.5.10	Feature Accuracy Calculation	44
Figure 3.5.11	Feature Selection	46

LIST OF ABBREVIATIONS

MLP	-	Multilayer Perceptron
FS	-	Feature Selection
NN	-	Neural Network
MSE	-	Mean Squared Error
ANN	-	Artificial Neural Network

CHAPTER I

INTRODUCTION

1.1 Project Background

As the technology is getting advanced, everything can be duplicated easily. However, Handwriting is the only thing that cannot be totally imitate by anyone in the real world. It is because every individual has their own unique writing styles which are the unique features for their handwriting. These are the information needed for writer identification. Even though these features are useful to identify the author, but not every features are really contribute much to the identification process. Therefore, feature selection is used to select the relevant attributes while discard those irrelevant attributes in order to make the learning faster. Feature selection is commonly done by using other methods such as ID3's other than neural networks. Most of the methods that used for feature selection that used in the past researches select one attribute at a time. However, it will be more sensible to make use of combination of attributes in some cases.

The proposed approach is neural networks used for feature selection as the process will start with whole set of attributes and remove the irrelevant one by one which is different from the common method like ID3. In fact, neural networks have been used in writer identification for the past researches. Yet, feature selection using neural networks in writer identification is still a new try. Throughout this project studies, neural network feature selection algorithm will be developed and the result will be analyzed to investigate the effectiveness use of neural networks in feature selection for writer identification domain.

1.2 Problem Statement(s)

Handwriting is different for different individuals. However, every single person has their own writing styles which are the unique features that are underlying in their writing. Yet, not every feature is important in identifying the writer. In order to identify these relevant features, feature selection technique is used. In this project studies, feature selection will be done by using neural network model which proposed by the paper written by Rudy Setiano and Huan Liu. (Rudy Setiano, H.L., 1997) Unlike the general feature selection, select one attribute at a time, it is start with whole set of attributes and remove the irrelevant one by one.

1.3 Objectives

- To investigate feature selection using neural network as method in writer identification domain
- To perform neural network as feature selection algorithm in identifying the author
- To design feature selection algorithm using neural network
- To classify the handwritten into an appropriate class label by using Neural Network as classifier (mostly use multilayer perceptron as classifier)

1.4 Scope

The scopes of the project are listed as below:

- The project is just mainly focus on writer identification domain
- The data for the experiment has resampled from the full UMI dataset
- Feature selection will be done by using neural network technique.
- Neural Network will be used as classifier to classify into appropriate class

1.5 Project Significance

This project is significant as it can research on whether the neural network model can used for feature selection in writer identification domain. There are some researchers had research in feature selection by using other approaches. So, an analysis of neural network feature selector is made in order to investigate the effectiveness of the result from the experiment.

1.6 Expected Output

This project is a research study which will produce the result of analysis for the feature selection by applying with neural network model in writer identification domain. There are two main modules for this analysis, which are Classification and Feature Selection module. In classification module, weight training will be carried out to obtain the optimized weight connections and minimize the mean squared error in order to yield a better training accuracy. The optimized weights will be used in the feature selection module. For feature selection module, the network will be retrained to get the accuracy rates for each feature respectively for feature ranking. Feature

ranking is done to select the important features out of all features by comparing with the allowable maximum decrease accuracy rates having one less attributes. At last, the important features will be selected.

1.7 Conclusion

As a conclusion, neural network will be used for feature selection in writer identification domain in this project. The result will be produced with the best features among total based on the accuracy rates.

For the following chapter, there will be a literature review of this topic which emphasis on the knowledge of the approaches. Project methodology will be included and discussed in Chapter II.

CHAPTER II

LITERATURE REVIEW AND RESEARCH METHODOLOGY

2.1 Introduction

In this chapter, literature of the research study will be discussed. It includes the study of Neural Network and writer identification domain. Besides, the components that are included in this project studies will be discussed to have better understanding before go into research stage. Moreover, the research methodology will also be determined and justified on the selection. Then project requirements will be stated clearly as the minimum requirements for the project studies. Lastly, the project schedule and milestones are prepared in order to ensure the studies can be complete within the specific time

2.2 Facts and finding

At the beginning of the project, there are some facts and findings which have to be made so that the research is on the right track.

2.2.1 Writer Identification Domain

Writing is a personal act: each writer is characterized by his writing, by the reproduction of details and unconscious practices ^[1]. Nowadays, there are many problem arises in the conflicts of originality of the documents such as a will, signature verification of banks, and analyze the text to identify the authors or the genetics of texts. In order to deliberate the authenticity of documents, writer identification is important for forensic analysis to help the experts in identifying process. Writer Identification generally requires two steps, there are:

- a) Image processing steps, which using feature extraction on images
- b) Classification, the document will categorized to the most similar document in the dataset according to the differences between the features

However, there are wide scopes for writer identification because of various types of different languages are existed other than English such as Chinese, Arabic. Therefore, there are a number of researchers worked on writer identification for different languages of text. The researchers researched the handwriting identification by using their proposed methods. Here are some proposed methods by the researchers.

- a) Complex wavelet transform for Chinese writer identification. (Da-Yuan Xu, Z.-W. S.-Y, 2008)
- b) Gradient feature and Neural Networks for Persian writer identification (Moghaddam, S.S. , 2009)

2.2.2 Neural Network

Neural Network is a set of processing elements and connections with adjustable weights. It has a large appeal to many researchers due to their great closeness to the structure of the brain. In an analogy of brain, an entity made up of interconnected neurons, neural networks are made up of interconnected processing elements called units which respond in parallel to a set of input given to each. It resembles the brain in the way:

- A neural network acquires knowledge through learning.
- A neural network's knowledge is stored within inter-neuron connection strengths known as synaptic weights.

Neural network was firstly introduced by physiologists Warren S. McCulloch and Walter Pitts in 1943. The model they created with two inputs and a single output.

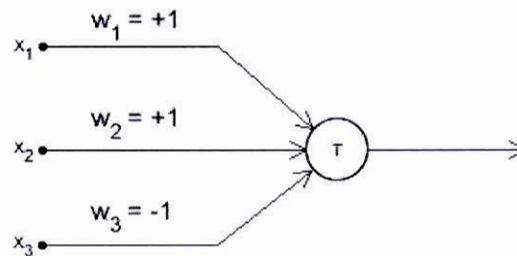


Figure 2.1 McCulloch-Pitts Model

However, the technology available at that time did not allow them to do too much. Neural network has been overlook for a period of frustration and disrepute time because of the limitations of neural network. Currently, this field enjoys a resurgence of interests. It has been applied into many applications, for example target recognition, voice recognition and medical diagnosis and as extra. There are three main types of neural network, which are:

- i. Single Layer Perceptron
- ii. Multilayer Perceptron
- iii. Radial Basis Function

However, multilayer perceptron are the most commonly used since it can solve many types of problems.

2.2.3 Multilayer Perceptron

Multilayer perceptron (MLP) is introduced to overcome the weakness of single layer perceptron (SLP) which can only deal with linearly separable sets of patterns. It can be used to It is known as the most widespread neural network architecture. It made useful until 1980s, because of lack efficient training algorithms (McClelland and Rumelhart 1986). MLPs are feedforward neural networks trained with the standard backpropagation algorithm. They are supervised networks so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input-output map. Most of the neural networks applications involve MLPs. In a MLP, there are at least three layers: 1 input layer, 1 or infinite hidden layers and 1 output layer.

Table 2.1 Function of layers

Input Layer	Receives the data about something
Hidden Layer	Used for calculations
Output Layer	Results of what the network has calculated, within 0.0 and 1.0

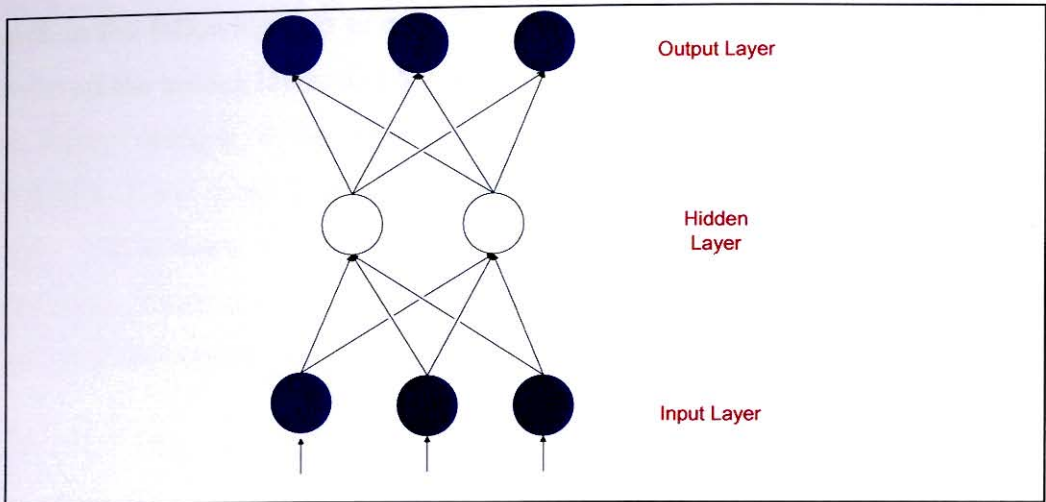


Figure 2.2 Multilayer Perceptron

However, the problem arises when the number of hidden layers and the number of hidden neurons are required to determine, it is normally determined by experts and experience. Generally, there is only required one hidden layer since it can be used to solve nearly all problems. Problems that require two layers are rarely encountered even though they can be used to represent the problem with any kind of shape. To determine the number of hidden layers (seen in Table 2.2):

Table 2.2 Determining hidden layers

Number of Hidden Layers	Result
None	Represent only linear separable functions or decisions
1	Can approximate any function contains continuous mapping from one finite space to another
2	Can represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy

After determining the number of hidden layers, the number of hidden units has to decide in the following step as it has tremendous influence on the final output. Both number of the hidden layers and hidden units need to revise properly. If there are too few hidden neurons, it can lead to underfitting, and if vice versa, it can lead to overfitting. If too many neurons in hidden layer, it will cause the training time to be longer. The amount of time increases when it is impossible to train the data sufficiently. There are few rules of thumbs need to take into account in order to correctly determine the number of hidden neurons:

- The number of hidden neurons should be between the size of input units and output units
- The number of hidden neurons can be determine by using the formula below

$$\text{no. of hidden neurons} = \frac{2}{3}(\text{input neurons} + \text{output neurons}) \text{ ---- (1)}$$

For the number of output units, it can be determined based on the class of the data used.

2.2.4 Feature Selection

Feature selection is known as attribute selection, variable selection, feature reduction or variable subset selection. It is the technique that used to select the subset of the attributes that are relevant for classification. The importance of feature selection is well-known. Besides, a classifier with better predictive accuracy may be found from the classification process by excluding redundant attributes.

The process of feature selection is often incorporated into classification. Generally, classification algorithm that builds decision trees such as ID3 is commonly used for feature selection as classifier. From a theoretical perspective, if large numbers of features are available, it is impractical for learning process as it is exhaustive search. Feature selection algorithm typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric