COMPARING FEATURE SELECTION METHOD FOR NEURAL NETWORK CLASSIFICATION

SITI KHALIJAH BINTI MOHD YASIN

This report is submitted in partial fulfillment of the requirements for the award of Bachelor of Electronic Engineering (Computer Engineering) with Honors

Faculty of Electronic and Computer Engineering

Universiti Teknikal Malaysia Melaka

JUN 2012

UNIVERSTI TEKNIKAL MALAYSIA MELAKA
FAKULTI KEJURUTERAAN ELEKTRONIK DAN KEJURUTERAAN KOMPUTER

BORANG PENGESAHAN STATUS LAPORAN
PROJEK SARJANA MUDA II

Tajuk Projek	:	COMPARING FEATURE SELECTION METHOD FOR NEURAL NETWORK CLASSIFICATION
Sesi Pengajian	:	1 1 / 1 2

Saya SITI KHALIJAH BINTI MOHD YASIN mengaku membenarkan Laporan Projek Sarjana Muda ini disimpan di Perpustakaan dengan syarat-syarat kegunaan seperti berikut:

1. Laporan adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan laporan 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**	

Disahkan oleh:

(TANDATANGAN PENULIS)

(COP DAN TANDATANGAN PENYELIA)

Tarikh:15 JUN 2012

Tarikh: 15 JUN 2012

"Saya akui laporan ini adalah hasil kerja saya sendiri kecuali ringkasan dan petikan yang tiap-tiap satunya telah saya jelaskan sumbernya."

Tandatangan :

Nama Penulis : Siti Khalijah Binti Mohd Yasin

Tarikh : 15 Jun 2012

"Saya/kami akui bahawa saya telah membaca karya ini pada pandangan saya/kami karya ini adalah memadai dari skop dan kualiti untuk tujuan penganugerahan Ijazah Sarjana Muda Kejuruteraan Elektronik (Kejuruteraan Komputer)."

Tandatangan :

Nama Penyelia : En Muhamad Noorazlan Shah Bin Zainuddin

Tarikh : 15 Jun 2012

ACKNOWLEDGEMENT

First of all, I would like to thank all lectures especially the one in Faculty of Electronic and Computer Engineering (FKEKK) for their guidance, encouragement, advice and ideas that have been poured out to me during the progress of my final year project. Indeed, all the knowledge given to me, I will try to practice and apply for my future working experience to express my appreciation.

I also wish to express our appreciation to thank Mr Muhamad Noorazlan Shah bin Zainudin for the guidance during my direction to ensure that I could complete my final year project. Not forgetting the 4 BENC classmates for their willingness instruction and guidance to me in carrying out the task and assisting me in solving the entire problem that I have been encountered during the progress of my project.

Finally, I would like to thank my parent as well as students from other educational institutes, who were also guided me and often injecting enthusiasm as well as offering me brilliant ideas.

ABSTRACT

Feature selection plays an important part in classifying systems in Neural Networks. A set of attributes which are relevant, irrelevant or redundant is desirable. The purpose of this project is to compare two methods of feature selection (Principle Component Analysis and Support Vector Machine) in order to gain the best result. Sample of images that has been converted into numeric data, it will be extracted and select by any feature selection technique. Back propagation is used to recognize the dataset and classified into the correct group. The extracted image or output from the two techniques has been used as an input for classification purpose. Back propagation is an iterative process that can often take a great deal of time to complete since the input data is feed forward network. Finally, the output or result of the classification will be analyzed in term of classification accuracy.

ABSTRAK

Pemilihan ciri memainkan peranan penting dalam mengklasifikasikan sistem dalam Rangkaian Neural. Satu set ciri-ciri yang relevan, tidak relevan atau berlebihan adalah wajar dalam pemilihan ciri. Tujuan projek ini adalah untuk membandingkan dua kaedah pemilihan ciri iaitu Prinsip Analisis Komponen dan Sokongan Mesin Vektor untuk mendapat hasil yang terbaik. Sampel imej yang telah ditukar ataupun diekstrak kepada data bernombor akan pilih oleh mana-mana teknik pemilihan ciri. Kaedah Rambatan balik digunakan untuk mengkelaskan dataset kepada kumpulan yang betul. Imej yang telah diekstrak atau hasil daripada dua teknik pemilihan ciri akan digunakan sebagai input untuk tujuan pengelasan. Proses lelaran bagi teknik Rambatan balik sering kali mengambil masa yang lama bagi tujuan pemprosesan bermula daripada data input memasuki rangkaian. Akhirnya, hasil daripada klasifikasi akan dianalisis dari segi ketepatan pengelasan.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATIONS

MI	-	Myocordial Infraction
PCA	-	Principal Component Analysis
SVM	-	Support Vector Machine
NN	-	Neural Network
BP	-	Back propagation
SPECT	-	Single Proton Emission Computed Tomography

LIST OF APPENDIX

CHAPTER I

INTRODUCTION

This chapter explains the background of the project which is giving the introduction about the project. This chapter also states the project objectives, problems statement and scopes of work.

1.1 Background

There were a lot of health products in today's market. Regular medicine, health supplements, nutrition supplements, weight loss supplements, placebo pills and alternative cure are just some of the option a person gets to choose. No matter which product that were chosen, the basic system of restoring the body back to health is the same. Understanding the background of the disease creation is the first step to understanding how health products work.

Sometimes it was hard to believe whether the person actually suffering the disease or not such as heart attack. A heart attack occurs if the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked. If blood flow isn't restored quickly, the section of heart muscle begins to die [1].

The classic symptom of myocardial infarction (MI) is an intense, sometimes squeezing, pressure or pain in or around the chest, often radiating to the jaw or left arm, and sometimes accompanied by profuse sweating, or a nearly overwhelming sense of fear or impending doom. Sometimes the discomfort may be relatively mild, and may be felt in the back, abdomen, shoulders, either or both arms. Unexplained sudden shortness of breath, nausea and vomiting, or merely a feeling of heartburn, may be the only symptoms [2]. Unfortunately, we can't count on having this classic pattern.

Although a lot of technologies to support this disease, but there were certain people that did not believe with that. Human error in taking the result also sometime will cause the danger to another. Let say, person A did not suffer any disease of heart attack but he was a little bit tension and cause hard to breath. Person B is the one who suffers the heart attack. Both of them was the diagnosis, the result of heart x-ray will help the doctor to determine whether they got a problem or not.

This research will purpose the best technique of feature selection to determine whether the patient has a normal or abnormal heart. The picture of x-ray that has been converted to numeric value will go through the feature selection method. Then it will analyze again using back propagation method of neural network classification to ensure the correctness of the picture.

1.2 Project Objective

The focus of this research is to come out with strong evidence to classify whether the patient has a normal or abnormal heart. Numeric data of certain cases will go through the three methods of feature selection that have been chosen. The selected data will be classified using back propagation method of neural network.

The objectives of this project are:

1. To study Back propagation Neural Network classifier to classify the sample dataset.
2. To study the two methods in feature selection which are Principle Component Analysis (PCA) and Support Vector Machine (SVM).
3. To analyze the result of percentage correctness classification.

1.3 Problem Statement

As mentioned above, medical crew will need evidence to prove whether their patient really suffering the heart disease or not. Although there were a lot of technologies and experimentation provided before to classify the disease, there are still make doubt of certain people. This research might be helped medical crew to come out with strong evidence to classify whether the patient has a normal or abnormal heart. People might be believed what that they have seen. With this kind of aids, probably they believe and ready to go further treatment.

This research also might be an aid for a medical trainer to improve their knowledge and skill before going through the real life situation. This research will come out with simple algorithms, easy to understand and might be suitable for them.

To come out with that solution, firstly the data need to filter by using a feature selection method. Recently, solutions of feature selection method problem have been dealt intensively. One of the most important contributions have been made was decision tree method. This method, uncovers relevant attributes one by one iteratively. The decision tree algorithm is by excluding the input feature of the neural network (NN) one by one and retraining the network rapidly. A lot of attribute require the process of retraining for almost every combination of the input feature.

To overcome this problem, a fast training algorithm such as Back propagation (BP) is used. But, the consideration of amount of time is needed to analyze the input. BP will take a lot of time to retrain the entire attribute. To improve the BP performance, feature selection method is used. Feature selection can reduce the number of input or predictable attribute. In other word, feature selection method will improve analysis and reduce the processing load. The data will be first to go through the feature selection technique before to retrain in the back propagation neural network classifier.

1.4 Scope of Project

This project will focus on the comparative study between two feature selection methods that have been chosen which are Principal Component Analysis (PCA) and Support Vector Machine (SVM). These methods were chosen because it was available for discrete attributes (numeric data). All attribute values in the database have been entered as numeric values corresponding to their index in the list of attribute values for that attribute domain.

In the classification of neural network system, feature selection plays a major role for the purpose of classification problem. A set of attribute that relevant, irrelevant or redundant can be huge. Reducing the number of attributes by selecting only the relevant ones is desirable. In doing so, higher performance with lower computational effort is expected.

The dataset is describing diagnosis of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories which is normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As the result, 44 continuous feature patterns were created for each patient [3].

The purpose of this project is to choose the best method to gain the best result of the feature pattern of neural network classification.

CHAPTER II

LITERATURE REVIEW

This chapter contains the literature review on the theoretical concept that applied in this research. It contains the information about the previous research and other information that are related. The main sources are journals and article obtained from the internet and other reference books.

2.1 Dataset

2.1.1 Cardiac Single Proton Emission Computed Tomography (SPECT)

SPECT is a nuclear medicine tomographic imaging technique using gamma rays. It is very similar to conventional nuclear medicine planar imaging using a gamma camera [4]. This image can provide information about the function of tissues. SPECT is a technique whereby cross sectional images of tissue function can be produced allowing the removal of the effect of overlying and underlying radioactivity [5].

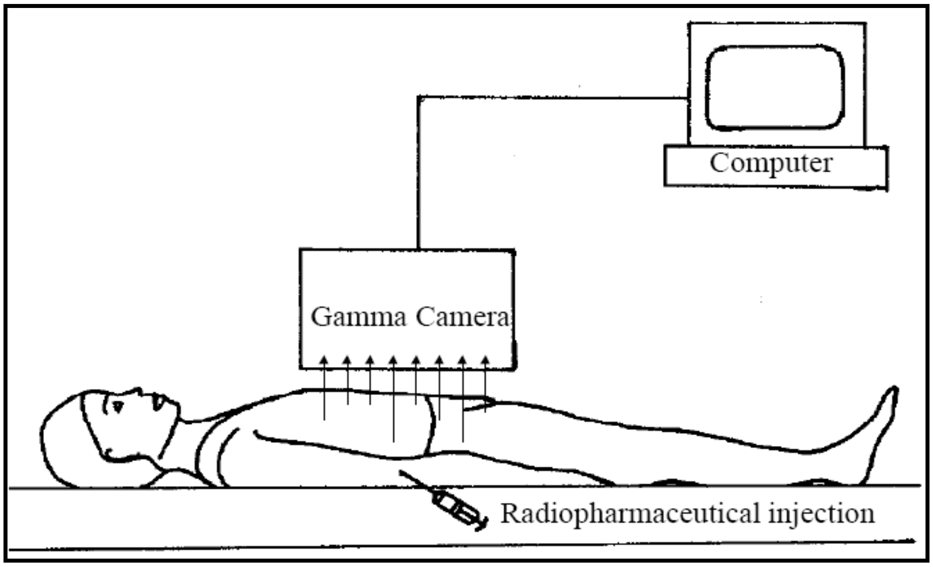
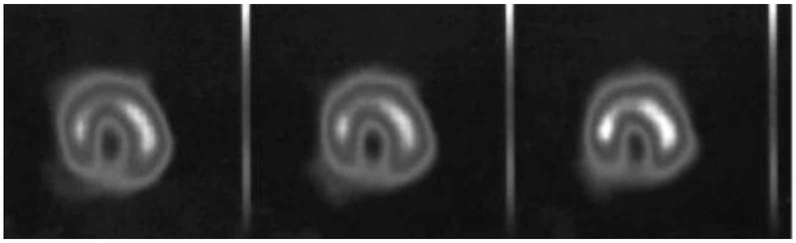


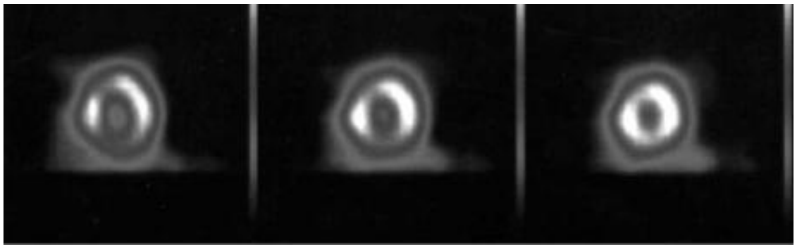
Figure 2.1.1.1: Schematic diagram of the production of radionuclide images using a gamma camera [5].

In cardiology, SPECT using either 201TlCl or 99mTc sestamibi is used to assess the viability of the heart muscle to help differentiate between ischemia and infarction. Both these tracers mimic the perfusion of the left ventricle and the use of SPECT highlights regions of poor flow during stress. These regions of the heart will reperfuse post stress if the tissue is ischemia rather than infected [5].

There were many other applications of SPECT example likes liver disease, tumor, lung and so on. It will help the diagnosis of disease and in some cases it can be an aid to therapy.



a) Normal heart



b) Abnormal heart (effect of poor perfusion in the posterior wall of the ventricle)

Figure 2.1.1.2: Comparison between normal and abnormal heart.

This information is necessary as aiding the treatment of patients with cardiovascular disease especially to determine which part of the heart is abnormal. The Cardio vascular disease also called as coronary artery disease. This disease is because of plaque (plak) that are built up inside the coronary arteries. An arteries role is to supply blood to our heart muscle. A plaque was made up of fat, cholesterol, calcium, and other substances found in the blood. The existence of plaque in the blood causes limits the flow of oxygen. It will become worse when plaque completely block blood flow through a coronary artery. Then, a heart attack may occur.

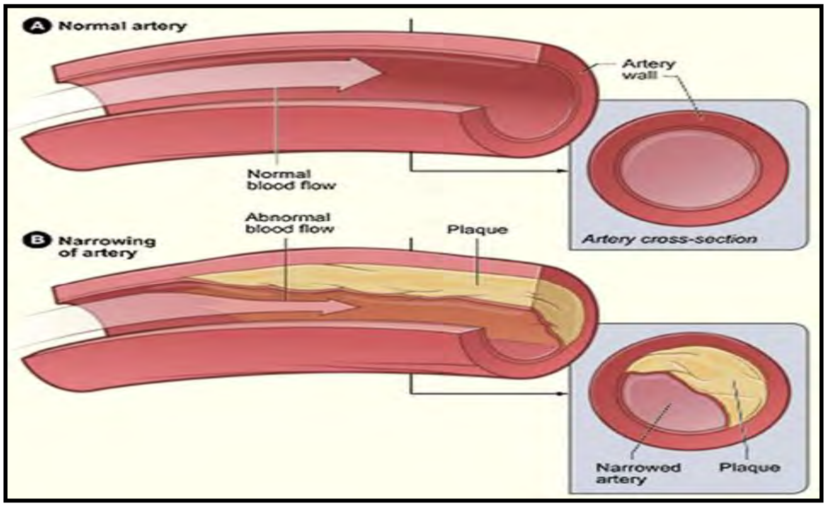


Figure 2.1.1.3: (A) shows a normal artery with normal blood flow. (B) Shows an artery with plaque buildup.

2.2 Feature Selection

Feature selection is a technique of selecting a subset of relevant features for building robust learning models [6]. By removing the irrelevant and redundant features from data, feature selection will help in improving the performance by alleviating the effect of the curse of dimensionality. Curse of dimensionality phenomena occurs when the dimensionality increase which is the volume of the space increase and the available data become sparse. It will affect the feature selection methods that require statistical significance. It also helps in enhancing generalization capability, speeding up the learning process and to improve model interpretability.

By improving analysis and reducing the processing load, feature selection is considered as a good step to go through before proceeding with the attribute classification. The method used for feature selection in neural network depends on the data type attribute.

Table 2.2.1: Feature selection methods that have been used in neural network and logistic regression models.

Algorithm	Method of Analysis	Comments
Neural Network	Interestingness scores Principal Component Analysis Support Vector Machine	The Neural Networks algorithm can use both entropy-based and Bayesian scoring methods, as long as the data contains continuous columns. Default.

Algorithm	Method of Analysis	Comments
Logistic Regression	Interestingness scores Shannon's Entropy Bayesian	Because you cannot pass a parameter to this algorithm to control feature election behavior, the defaults are used. Therefore, if all attributes are discrete or discretized, the default is BDEU.

Feature selection for neural network model can be controlled by using this parameter which is MAXIMUM_INPUT_ATTIBUTES, MAXIMUM_OUTPUT_ATTRIBUTES and MAXIMUM_STATES. To overcome the complexity of the network in which when the number of input attributes is larger than MAXIMUM_INPUT_ATTIBUTES or the number of predictable attributes is greater than the value of the MAXIMUM_OUTPUT_ATTRIBUTES, feature selection method was used. The number of hidden layers also can be controlled by setting the HIDDEN_NODE_RATIO parameter.

2.2.1 Principal Component Analysis (PCA)

The analysis of the paper on ‘Feature Selection using Principal Component Principal Analysis’ state that PCA has the potential to perform feature selection and is able to select a number of important feature from all the feature component [7]. Wake has a PCA attribute transformer on the attribute selection package and also a filter by the same name. PCA transforms the original data into a new space comprised of direction that are linear combinations of the original attributes. Component Analysis is a dimension reduction method that creates variable from the datasets for both supervised and unsupervised classification purposes. This is based on the consideration that a large