COMPARATIVE STUDY OF FEATURE SELECTION METHOD OF
MICROARRAY DATA FOR GENE CLASSIFICATION

NURULHUDA BINTI GHAZALI

UNIVERSITI TEKNOLOGI MALAYSIA
COMPARATIVE STUDY OF FEATURE SELECTION METHOD OF MICROARRAY DATA FOR GENE CLASSIFICATION

NURULHUDA BINTI GHAZALI

A project report 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

OCTOBER 2009
To my beloved Mummy and Abah…
Hazijun bt. Abdullah and Ghazali bin Sulong

My beloved sisters..
Nurhanani and Nur Hafizah

My beloved brother..
Ikmal Hakim

My brother-in-laws..
Saiful Azril and Faridun Naim

My beloved nieces..
Sarah Afrina and Sofea Alisya

My supervisor..
Assoc. Prof. Dr. Puteh Saad

and last but not least to all my supportive friends especially
Syara, Radhiah, Zalikha, Umi and Hidzir..

“Thank you for all the support and love given”
ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful.

All praise and thanks be to Allah for His guidance that had lead me in completing this research. His blessings had given me strength and courage throughout this past year and had helped me overcome difficulties during this research period.

First and foremost, I would like to take this opportunity to express my sincere gratitude to those who had assisted me in finishing this research. To my dear supervisor, Assoc. Prof. Dr. Putheh Saad, thank you for all your supports and guidance in showing me the right path towards completing this research. I really appreciated your advices and motivations that you had given me within the period of this research.

My infinite thank you are dedicated to my loving and caring family, who had cherish me and give me full support in any kind. I am deeply appreciated for all the motivations and inspirations. Without them, it is impossible for me to finish my research.

And last but not least, an endless appreciation to all my fellow friends and classmates for all the supports and encouragements. Their friendships never fail to amaze me.

May Allah S.W.T bless them all and repay all of their kindness and sacrifices.
Recent advances in biotechnology such as microarray, offer the ability to measure the levels of expression of thousands of genes in parallel. Analysis of microarray data can provide understanding and insight into gene function and regulatory mechanisms. This analysis is crucial to identify and classify cancer diseases. Recent technology in cancer classification is based on gene expression profile rather than on morphological appearance of the tumor. However, this task is made more difficult due to the noisy nature of microarray data and the overwhelming number of genes. Thus, it is an important issue to select a small subset of genes to represent thousands of genes in microarray data which is referred as informative genes. These informative genes will then be classified according to its appropriate classes. To achieve the best solution to the classification issue, we proposed an approach of minimum Redundancy-Maximum Relevance feature selection method together with Probabilistic Neural Network classifier. The minimum Redundancy-Maximum Relevance feature selection method is used to select the informative genes while the Probabilistic Neural Network classifier acts as the classifier. This approach has been tested on a well-known cancer dataset which is Leukemia. The results achieved shows that the gene selected had given high classification accuracy. This reduction of genes helps take out some burdens from biologist and better classification accuracy can be used widely to detect cancer in early stage.
ABSTRAK

# 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>x</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
<td></td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xiv</td>
<td></td>
</tr>
</tbody>
</table>

1 INTRODUCTION 1

1.1 Introduction 1

1.2 Background of the Problem 3

1.3 Problem Statement 5

1.4 Objectives of Research 5

1.5 Scope of Research 6

1.6 Importance of the Study 6
2 LITERATURE REVIEW

2.1 Introduction 8
2.2 Genes and Genes Expression 9
2.3 Microarray Technology 11
2.4 Feature Selection 12
  2.4.1 ReliefF Algorithm 13
  2.4.2 Information Gain 15
  2.4.3 Chi Square 16
  2.4.5 Minimum Redundancy-Maximum Relevance 16

2.5 Classification 18
  2.5.1 Random Forest 18
  2.5.2 Naïve Bayes 19
  2.5.3 Probabilistic Neural Network 20

2.6 Challenges in Genetic Expression Classification 22
2.7 Summary 23

3 METHODOLOGY

3.1 Introduction 24
3.2 Research Framework 25
  3.2.1 Problem Definition 27
  3.2.2 Related Studies 27
  3.2.3 Study on Proposed Method 28
  3.2.4 Data Preparation 29
  3.2.5 Feature Selection 31
  3.2.6 Classification 32
  3.2.7 Evaluation and Validation 34
  3.2.8 Result Analysis 34

3.3 Leukemia 35
3.4 Software Requirement 36
3.5 Summary 37
### 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>Schemes in mRMR Optimization Condition</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison of $k$-NN and PNN using 4 Datasets</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>Leukemia Dataset</td>
<td>56</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>DNA Structure</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>Process of Producing Microarray</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Sample of Microarray</td>
<td>12</td>
</tr>
<tr>
<td>2.4</td>
<td>Comparison of 3 Methods of Feature Selection</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Architecture of PNN</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Research Framework</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Sample of Dataset</td>
<td>30</td>
</tr>
<tr>
<td>3.3</td>
<td>Sample of Dataset</td>
<td>30</td>
</tr>
<tr>
<td>3.4</td>
<td>Process of Feature Selection</td>
<td>31</td>
</tr>
<tr>
<td>3.5</td>
<td>Process of Classification</td>
<td>32</td>
</tr>
<tr>
<td>3.6</td>
<td>Overall Process of Feature Selection and Classification</td>
<td>33</td>
</tr>
<tr>
<td>3.7</td>
<td>Abnormal Proliferation of Cells in Bone Marrow Compared To Normal Bone Marrow</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>Original Dataset in ARFF Format Showing Genes Values</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Original Dataset in ARFF Format Showing Class Names</td>
<td>40</td>
</tr>
<tr>
<td>4.3</td>
<td>Dataset in IOS GeneLinker Software before Discretization</td>
<td>41</td>
</tr>
<tr>
<td>4.4</td>
<td>Dataset in IOS GeneLinker Software after Discretization</td>
<td>42</td>
</tr>
<tr>
<td>4.5</td>
<td>Discretized Data in CSV Format</td>
<td>43</td>
</tr>
<tr>
<td>4.6</td>
<td>Continuous Data in CSV Format</td>
<td>44</td>
</tr>
<tr>
<td>4.7</td>
<td>ReliefF Algorithm</td>
<td>48</td>
</tr>
<tr>
<td>4.8</td>
<td>Chi Square Algorithm</td>
<td>51</td>
</tr>
</tbody>
</table>
5.1 Classification using PNN for Different Types of Data
5.2 Classification Accuracy using PNN for Different Scheme in Feature Selection using mRMR
5.3 Classification using PNN by Different Number of Selected Features
5.4 Comparison of Classification Accuracy by Different Feature Selection Method using PNN
5.5 Comparison of Classification Accuracy using Different Classifier
5.6 Classification Accuracy using 10-fold Cross Validation
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Acute Lymphoblastic Leukaemia</td>
</tr>
<tr>
<td>AML</td>
<td>Acute Myeloid Leukaemia</td>
</tr>
<tr>
<td>ARFF</td>
<td>Attribute-Relation File Format</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Values</td>
</tr>
<tr>
<td>mRMR</td>
<td>Minimum Redundancy Maximum Relevance</td>
</tr>
<tr>
<td>PNN</td>
<td>Probabilistic Neural Network</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>$k$-NN</td>
<td>$k$-Nearest Neighbor</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger Ribonucleic Acid</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Project 1 Gantt Chart</td>
<td>77</td>
</tr>
<tr>
<td>B</td>
<td>Project 2 Gantt Chart</td>
<td>82</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

Every living organism has discrete hereditary units known as genes. Each gene provides some function or mechanism either by itself or it will combine with other genes that will eventually producing some property of its organism. Genome is a complete set of genes for an organism and is said as the “library” of genetic instruction that an organism inherits (Campbell and Reese, 2002). Each gene is made of deoxyribonucleic acid (DNA) molecule which consists of two long strands that tightly wound together in a spiral structure known as double helix (Amaratunga and Cabrera, 2004). Along each of these strands located various form of genes that differs by its sequences for each organism. This makes each organism unique and different from each other. The DNA molecule of an organism is located in a cell. A cell is the fundamental units of all living organism and it contains many substructure such as nucleus, cytoplasm and plasma membrane. The nucleus is where DNA is embedded. Genes in DNA is expressed by transferring its coded information into proteins that dwell in the cytoplasm. This process is called as gene expression (Russell, 2003). There are several experimental techniques to measure gene
expression such as expression vector, reporter gene, northern blot, fluorescent hybridization, and DNA microarray.

DNA microarray technology allows the simultaneous measurement of the expression level of a great number of genes in tissue samples (Paul and Iba, 2005). It yields a set of floating point and absolute values. Many explored on classification methods to recognize cancerous and normal tissues by analyzing microarray data. The microarray technology typically produces large datasets with expression values for thousands of genes (2000-20000) in a cell mixture, but only few samples are available (20-80) (Huerta et al.).

This study is focused on gene selection and classification of DNA microarray data in order to identify tumor samples from normal samples. Gene selection is a process where a set of informative genes is selected from the gene expression data in a form of microarray dataset. This process helps improve the performance of the classifier. On the other hand, classification is a process to classify microarray data in several classes that have its own characteristics. There are several techniques that have been used in gene selection such as ReliefF Algorithm, Information Gain, minimum Redundancy Maximum Relevance (mRMR) and Chi Square. For classification of microarray data, a few techniques have been applied in the bioinformatics field to classify the highly dimensional data. These techniques include Random Forest, Naïve Bayes and Probabilistic Neural Network (PNN).

The proposed method involved two stages where the first stage is the gene selection stage and the second one would be the classification stage. In gene selection method, the technique chosen is a technique called minimum Redundancy-Maximum Relevance (mRMR) feature selection and will be compared to three other method namely ReliefF, Information Gain and Chi Square. mRMR is a feature selection framework that was introduced by Ding and Peng in 2005. They supplement the maximum relevance criteria along with minimum redundancy criteria to choose additional features that are maximally dissimilar to already identified ones.
This can expand the representative power of the feature subset and help improves their generalization properties. The classification problem will be handled by Probabilistic Neural Network (PNN) technique. PNN has been widely used in solving classification problems. This is because it can categorize data accurately (Nur Safawati Mahshos, 2008). Both techniques will be assessed on a benchmark cancer dataset which is Leukemia (Golub et al, 1999).

### 1.2 Background of the Problem

Cancer is a killer disease to everyone worldwide. There are at least 100 different types of cancer that has been identified. Traditionally cancer is diagnosed based on the microscopic examination of patients’ tissue. This kind of diagnosis may fail when dealing with unusual or atypical tumors. Currently, cancer diagnosis is based on clinical evaluation and also referring to medical history and physical examination. This diagnosis takes a long time and might however limit the finding of tumor cell especially in early tumor detection (Xu and Wunsch, 2003). If tumor cell is found in its critical stage, then it might be too late to cure the patient.

Thus, classification for cancer diseases has been widely carried out for the past 30 years. Unfortunately, there has been no general or perfect approach to identify new classes or assigning tumors to known classes. This happens because there are various ways that can cause cancer and too many types of cancer that sometimes difficult to distinguish. By depending on morphological appearance of tumors, it is hard to discriminate between two similar types of cancer (Golub et al, 1999).
In order to overcome the above issues, a new technique based on cancer classification has been introduced. The technique employs an advanced microarray technology that measures simultaneously the expression level of a great number of genes in tissue samples. Nevertheless, this technique contributes to a new problem whereby there exist a numerous number of irrelevant genes or overlapping of genes. Hence, selection and classification must be done in order to select the most significant genes from a pool of irrelevant genes and noises.

Nowadays, there are a lot of selection and classification techniques that has already been studied and developed to help in better classification of microarray data. Among these techniques, there are a few that gives promising result such as mRMR, ReliefF, Information Gain and Chi Square for gene selection and PNN classification. mRMR is chosen as the primary technique for gene selection since this technique are proposed originally for gene selection (Ding and Peng, 2003). The advantage of this technique is it focuses on redundancy of genes together with the relevance of genes. Unlike other techniques; ReliefF (Kononenko, 1994), Information Gain (Cover and Thomas, 1991) and Chi Square (Zheng et al, 2003), they were firstly proposed only for general feature selection, rather than genes. For comparison, these four techniques are used to select genes in order to measure the performance.

As for classification, the technique chosen in this research is Probabilistic Neural Network (PNN) classifier. PNN has been use in many studies of feature classification (Pastell and Kujala, 2007; Shan et al, 2002). These studies have proved that PNN yield better result in classification accuracy compared to other existing classifiers. Thus, this research combines a few feature selection methods together with PNN classifier to classify microarray data according to its classes.
1.3 Problem Statement

The challenging issue in gene expression classification is the enormous number of genes relative to the number of training samples in gene expression dataset. Not all genes are relevant to distinguish between different tissue types (classes) and introduced noise (Liu and Iba, 2002) in the classification process and thus it drowns out the contribution of the relevant genes (Shen et al, 2007). On top of that, a major goal of diagnostic research is to develop diagnostic procedures based on inexpensive microarrays that have adequate number of genes to detect diseases. Hence, it is crucial to recognize whether a small number of genes will be sufficient enough for gene expression classification.

1.4 Objectives of Research

The aim of this research is to select a set of meaningful genes using a minimum Redundancy-Maximum Relevance feature selection technique and to classify them using Probabilistic Neural Network. In order to achieve aim, the following objectives must be fulfilled:

1. To select a set of meaningful genes using Minimum Redundancy-Maximum Relevance (mRMR), Information Gain, ReliefF and Chi Square.
2. To evaluate the effectiveness of feature selection method using Probabilistic Neural Network (PNN) classifier.
3. To compare the performance of mRMR as feature selection method using PNN, Random Forest, and Naïve Bayes classifiers.
1.5 Scope of Research

The scope of study is stated as below:

- mRMR, ReliefF, Information Gain and Chi Square is utilized for gene selection.
- PNN technique is used for gene expression classification.
- Leukemia microarray dataset is used for testing (Data source: Weka Software Package, http://www.cs.waikato.ac.nz/ml/weka/)
- 10-fold cross validation is utilized to perform the validation.
- The tools used are Matlab, Knime, Weka and IOS GeneLinker

1.6 Importance of the Study

This study is carried out to aid in classification of cancer diseases. Cancer diseases are lethal to human. Several methods have been conducted to detect this deadly disease. Unfortunately, the time taken is too long to confirm that someone has the disease. This is due to the symptoms that can only be seen after a very long time and by the time, cancer level has reached a critical stage.

Common examination of patients require weekly checkup to precisely identify the presence of the disease. Due to the long term of examination, the disease might get more critical without exact cure or treatment. The advanced technology of microarray lessens the burden among medical staffs. The microarray of human genes can be used to detect cancer diseases earlier.
Despite the fact that microarray technology is said has the capability to solve the problems, but unfortunately this technology requires an excellent technique to select only the best subset of all genes to give enough information about a particular cancer disease. This is due to the overwhelming number of genes produce by microarray in a few sample sizes.

Thus, by doing this research, the best approach can be achieved to solve the problems in gene selection and classification. The idea was to apply the minimum Redundancy-Maximum Relevance feature selection technique (compared with other feature selection techniques) together with Probabilistic Neural Network to give a tremendous result in a short time. This research provides knowledge in the field of bioinformatics and it gives benefit in medical area. Apart from that, it helps saving human life by detecting cancer disease in early stage.
Ahlers, F.J., Carlo, W.D., Fleiner, C., Godwin, L., Mick, Nath, R.D., Neumaier, A.,
http://www.icsi.berkeley.edu/~storn/code.html

expression revealed by clustering analysis of tumor and normal colon tissues
probed by oligonucleotide arrays. PNAS. Vol 96: 6745-6750

Microarray and Protein Array Data. New Jersey, USA: Wiley Inter-Science.
8-10

Babu, B.V. and Chaturvedi, G. Evolutionary Computation Strategy for Optimization
of an Alkylation Reaction. Birla Institute of Technology and Science.

Trickle-bed Reactor using Differential Evolution and Orthogonal Collocation.
Elsevier Science.

Balasundaram Karthikeyan, Srinivasan Gopal, Srinivasan Venkatesh and
Approach to PD Pattern Classification Compared with BPA Network. Journal

Bi, C., Saunders, M. C. and McPheron, B. A. (2007). Wing Pattern-Based
Classification of the Rhagoletis pomonella Species Complex Using Genetic
Vol 4: 1-14


http://www.encyclopedia.com/doc/1O11-datapreparation.html


Principal Component Analysis, [Wikipedia](http://en.wikipedia.org/wiki/Principal_components_analysis), (accessed May 14, 2009)


