A HYBRID GENETIC ALGORITHM AND SUPPORT VECTOR MACHINE CLASSIFIER FOR FEATURE SELECTION AND CLASSIFICATION OF GENE EXPRESSION

MOHD SABERI BIN TAN AH CHIK @ MOHAMAD

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA

BORANG PENGESAHAN STATUS TESIS♦

JUDUL: **A HYBRID GENETIC ALGORITHM AND SUPPORT VECTOR MACHINE CLASSIFIER FOR FEATURE SELECTION AND CLASSIFICATION OF GENE EXPRESSION**

SESI PENGAJIAN: 2004/2005

Saya **MOHD SABERI BIN TAN AH CHIK @ MOHAMAD**
(HURUF BESAR)

mengaku membenarkan tesis (~~PSM~~/Sarjana/~~Doktor Falsafah~~)* ini disimpan di Perpustakaan Universiti Teknologi Malaysia dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Teknologi Malaysia.
2. Perpustakaan Universiti Teknologi Malaysia dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan 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)

Disahkan oleh

(TANDATANGAN PENYELIA)

Alamat Tetap:

LOT 772, KAMPUNG PADANG PAK AMIN, 16370 JELAWAT, BACHOK KELANTAN, MALAYSIA

PROFESOR DR. SAFAAI DERIS
Nama Penyelia

Tarikh: **1hb APRIL 2005**

Tarikh: **1hb APRIL 2005**

CATATAN:
* Potong yang tidak berkenaan.
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT atau TERHAD.
♦ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda (PSM).

"I/~~We~~* hereby declare that I/~~We~~* have read this thesis and that in my/~~our~~* opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Master of Science (*Computer Science*)".

Signature : …….....

Name of Supervisor : Professor Dr. Safaai Deris

Date : 1st April 2005

* Delete as necessary

BAHAGIAN A – Pengesahan Kerjasama*

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui kerjasama antara _____ dengan _____

Disahkan oleh:

Tandatangan : ……. Tarikh : …….

Nama : …….

Jawatan : …….

(Cop rasmi)

** Jika penyediaan tesis/projek melibatkan kerjasama*.

BAHAGIAN B – Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar : **Prof. Madya Dr. Ramlan Bin Mahmod**
Fakulti Sains Komputer & Teknologi Maklumat
Universiti Putra Malaysia
43400 UPM Serdang
Selangor

Nama dan Alamat Pemeriksa Dalam I : **Prof. Madya Dr. Naomie Binti Salim**
Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia
81310 UTM Skudai
Johor

Pemeriksa Dalam II :

Nama Penyelia Lain (jika ada) :

Disahkan oleh Penolong Pendaftar di SPS:

Tandatangan : ……..…….…. Tarikh : …….

Nama : **GANESAN A/L ANDIMUTHU**

A HYBRID GENETIC ALGORITHM AND SUPPORT VECTOR MACHINE CLASSIFIER FOR FEATURE SELECTION AND CLASSIFICATION OF GENE EXPRESSION

MOHD SABERI BIN TAN AH CHIK @ MOHAMAD

A thesis submitted in fulfilment 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

APRIL 2005

I declare that the thesis entitled "*A Hybrid Genetic Algorithm and Support Vector Machine Classifier for Feature Selection and Classification of Gene Expression*, is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : …….....……

Name : Mohd Saberi Tan Ah Chik @ Mohamad

Date : 1st April 2005

To my beloved parents and grandmother

ACKNOWLEDGMENT

It was my good fortune to have Professor Dr. Safaai Deris as my advisor. He was always available to support my research effort during my years at UTM. Despite of his busy schedule, he was constantly able to offer intelligent advice, comments, criticisms and suggestions whenever I had consulted him. I have also particularly appreciated his immediate and very thoughtful responses to my problems.

I am also indebted to Malaysian Ministry of Science, Technology and Environments for funding my Master program. My sincere appreciation also goes to all the people that have answered my email queries so promptly, the open source community, the maintainers of Google, CiteSeer and DBLP, and to all those authors who have generously made their publications and lecture notes available online.

My friends in the house rent and Artificial Intelligence and Bioinformatics Laboratory are the best people I know. Without from their help getting things done and letting off steam, I would never have make it.

My deepest thanks go to my parents and my family. Their influence made me realize the importance of education from a very early age. I cannot give enough thanks to them for the great love and support that they gave me throughout their lives.

ABSTRACT

Advancement in gene expression technology offers the ability to measure the expression levels of thousand of genes in parallel. Gene expression microarray data is expected to significantly aid in the development of efficient cancer diagnosis and classification platforms. Key issues that need to be addressed under such circumstances are the efficient selection of a small subset of genes that might profoundly contribute to disease identification from the thousand of genes measured on microarrays that are inherently noisy. This research deals with finding a small subset of informative genes from gene expression data which maximizes the classification accuracy. This research proposed a hybrid between Genetic Algorithm and Support Vector Machine classifier for selecting an optimal small subset of informative genes and classifying the optimal subset. Two benchmark data sets were used to evaluate the usefulness of the approach for small and high dimension data. Although, the experimental results showed that the hybrid method performed better than some of the best previous methods on small dimensional data, its performance deteriorated significantly on the higher dimensional data. An improved version of the hybrid method was designed by introducing a new algorithm for features selection based on improved chromosome representation to replace the original algorithm on the hybrid method which appeared to perform poorly on high dimensional data. The results of the gene expression microarray classification demonstrated that the proposed method performed better than the original and the previous methods. The informative genes from the experiment results proved to be biologically plausible when compared with the biological results produced from biologist and computer scientist researches.

ABSTRAK

Peningkatan teknologi pengekspresan gen yang berterusan membolehkan ribuan tahap pengekspresan bagi gen-gen diukur secara serentak. Data pengekspresan gen dijangka dapat memberikan faedah yang besar dalam pembangunan diagnosis kanser dan platfom pengelasan yang efisyen. Isu utama yang perlu diatasi dalam hal ini adalah pemilihan subset kecil bagi gen-gen secara efisyen daripada ribuan gen yang diukur oleh microarray dan dapat menyumbang kepada pengenalpastian penyakit. Tetapi, data gene yang dihasilkan oleh microarray mempunyai kebisingan. Kajian ini melibatkan pencarian gen-gen yang berinformatif dalam jumlah yang kecil daripada data pengekspresan gen microarray untuk memaksimumkan ketepatan proses pengelasan. Kajian ini telah mencadangkan pendekatan hibrid di antara Algoritma Genetik dan pengelas Mesin Sokongan Vektor untuk memilih subset kecil yang optimum bagi gen-gen berinformatif dan mengelaskan subset tersebut. Dua set data perbandingan yang berdimensi kecil dan besar telah digunakan untuk menilai kebolehgunaan pendekatan tersebut. Sungguhpun hasil-hasil eksperimen telah menunjukkan kaedah hibrid tersebut mengatasi kaedah-kaedah terbaik yang terdahulu pada data berdimensi kecil, namun prestasinya jatuh mendadak pada data yang lebih berdimensi besar. Kaedah hibrid yang lebih baik telah dibangunkan dengan memperkenal satu algoritma baru berasaskan perwakilan kromosom yang ditingkatkan penggunaanya untuk pemilihan ciri-ciri bagi menggantikan algoritma asal dalam kaedah hybrid terbabit yang didapati tidak sesuai bagi data yang berdimensi besar. Hasil-hasil daripada pengelasan pengekspresan gen microarray telah menunjukkan bahawa prestasi kaedah yang telah dicadangkan menandingi kaedah asal dan keadah-kaedah lain yang terdahulu. Gen-gen yang berinformatif daripada hasil eksperimen itu telah dibuktikan kepentingan biologinya melalui perbandingan dengan hasil eksperimen yang telah dikeluarkan oleh kajian ahli biologi dan saintis komputer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

λ	–	Weight decay factor / eigen value
ξ	–	Slack variable
π	–	Initial state probability
σ	–	Kernel scaling parameter or standard deviation
α	–	Lagrange multiplier
η	–	Learning rate
ϕ	–	Mapping
Err_{emp}	–	Empirical error
Err_{st}	–	Structureal risk
$k(x, x^i)$	–	Kernel function
μ	–	Means
$\sum$	–	Finite alphabet
a	–	Random value
A	-	Total of samples
b	–	Bias
c	–	Class
C	–	Soft-margin parameter
C_k	–	set of samples
Cy	–	Cyanine dyes
D	–	Probability distribution or data set
e	–	Expression level
E	–	Matrix
f	–	Classification function

f_i	–	Real index *i* in chromosome
F_i	–	i^{th} feature in data set
g	–	Gene
h	–	VC dimension
H	–	Relative entropy
k	–	Kernel
l	–	Lower bound
L	–	Lagrangian
Mar	–	Margin
n	–	Number of training samples, examples, features or instances
n_c	–	Number of feature subsets
N	–	Number of dimension space
r	–	Radius
s	–	Features selected
t	–	Threshold
T	–	Number of correctly samples
u	–	Upper bound
v	–	Vector
w	–	Weight vector
x	–	Instance or features subset
y	–	Class label

LIST OF ABBREVIATIONS

AB	–	AdaBoost
ABR	–	Regularized AdaBoost
ART-NN	–	Adaptive Resonance Theory Neural Network
ALL	–	Acute Myeloid Leukemia
AML	–	Acute Lymphoblastic Leukemia
DBNN	–	Denoeux Belief Neural Network
DNA	–	Deoxybonucleic Acid
ERM	–	Empirical Risk Minimization
GA	–	Genetic Algorithm
GASVM	–	Hybrid of GA with SVM classifier
GAWV	–	Hybrid of GA with WV classifier
JCFO	–	Joint Classifier and Feature Optimization
KFD	–	Kernel Fisher Discriminant
KM	–	Kernel Method
K-NN	–	k-Nearest Neighbour
LD	–	Logistic Discriminant
LOOCV	–	Leave One Out Cross Validation
MLP	–	Multilayer Perceptron
MN	–	Modular Neural Network
mRNA	–	messenger RNA
New-GASVM	–	New Hybrid of GA and SVM classifier
PCA	–	Principal Components Analysis
PGA	–	Parallel Genetic Algorithm
PLS	–	Partial Least Squares
RBF	–	Radial Basis Function
RNA	–	Ribonucleic Acid

SASOM	–	Structure Adaptive Self-Organizing Map
SRM	–	Structural Risk Minimization
SVM	–	Support Vector Machine
TNoM	–	Threshold Number of Misclassification
VC	–	Vapnik-Chervonenkis
WV	–	Weight Voting

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The development of microarray technology has produced a large amount of gene expression microarray data. To precisely classify gene expression microarray, needs to be selected informative genes because the gene expression microarray data have much noise and variety of such problems. Most previous works concerning gene (feature) selection task were based on filtering approaches and applied before the classification process. However, these filter approaches are unsuitable to be applied to the gene expression microarray data. Therefore this research is mainly concerned with the selection of informative genes from the data using wrapper (hybrid) approach. The purpose of this research is to investigate the properties of a hybrid Genetic Algorithm and Support Vector Machine classifier and its applicability in gene expression microarray classification. A review of gene expression is given in Appendix A. For quick reference, a glossary of structural genomic terms is also provided in Appendix B. This chapter begins by discussing the challenges involved in the classification of gene expression microarray. It also discusses the motivations for gene expression microarray classification and the aim well as the research objectives. The scope for classifying gene expression is also described. An overview of the thesis is presented in the final section of this chapter.

1.2 Challenges of Gene Expression Microarray Classification

Machine learning method has been used for the last 6 years to classify gene expression microarray and consistently demonstrated to be the best approach. One of the major challenges is the overwhelming number of genes relative to the number of training samples in the data set. Many of the genes are not relevant to differentiate between different tissue types (classes) and have introduced noise in the classification process, and thus potentially drowning out the contribution of the relevant ones (Ben-Dor et al., 2000). Moreover, a major goal of diagnostic research is to develop diagnostic procedures based on inexpensive microarrays that have enough probes to detect diseases (Liu et al., 2001; Ben-Dor et al., 2000). Thus, it is crucial to recognize whether a small number of genes will be sufficient enough for good classification task (Krishnapuram et al., 2004).

Key issues that need to be addressed under such circumstances are the efficient selection of small subset of informative genes which contributes to a disease from the thousands of genes measured on microarrays that are inherently noisy. The gene expression data sets are problematic due to the large number of genes. Consequently, a method that search over subsets of features can be prohibitively expensive. Moreover, these data sets contain only a small number of samples, so that the detection of irrelevant genes can suffer from statistical instabilities (Ben-Dor et al., 2000). Therefore, most previous methods for cancer classification that are based on gene expression data started with feature selection methods (Cho and Won, 2003). In gene expression classification, there is a practical need to reduce the number of measurements without significantly degrading the performance of the system. In the application, i.e., gene expression microarray that has involves a very large number of features, the performance of the classifier often degrades if the number used increases beyond a certain value (Ferri et al., 1993).

The gene expression microarray data of a sample is a vector that contains the gene expression levels of each sample measured simultaneously by microarray. From

the point of view of pattern recognition, the task of cancer classification based on gene expression data is a pattern classification problem and the feature vector for the classification is the gene expression vector. However, this problem is an extremely difficult one for many methods, since the feature dimension is usually very high (several thousands) and the training samples are usually very scarce, around 100 known samples or less (Mukherjee, 2001). If work is done in this high dimensional space with limited samples, most conventional pattern recognition algorithms may have not worked well (Nguyen and Rocke, 2002). Some algorithms that involve matrix inversion operation may not be able to arrive at a solution when the number of samples is less than the dimensions specified. For others that can achieve a solution, it may not be able to work properly on samples other than that used for training. This is called the generalization problem in pattern recognition and machine learning (Vapnik, 1995).

1.3 Research Motivations

Bioinformatics is a study of biological systems using computational techniques. It represents a relatively new area of computer science to handle and manage large amounts of data generated by advance technologies which are designed for measuring biological systems. The use of machine learning techniques in analyzing the biological data is currently at the forefront of the field and represents a major opportunity for the machine learning community. It has become biologically feasible to record large amounts of data from biological systems only recently, and this therefore explains the relatively recent emergence of the field of gene expression analysis.

Although cancer classification has improved over the past 30 years, there has been no general and perfect approach for identifying new cancer classes or assigning tumors to known classes (Golub et al., 1999; Ryu and Cho, 2002). It is because there

can be so many pathways causing cancer and many varieties. Traditional classification methods are mostly dependent on morphological appearance of tumors and their applications are limited by existing uncertainties (Golub et al., 1999). This approach has various limitations especially in discriminating between two similar types of cancer.

Recent technological researches have made some advancement in molecular genetics such as microarray (Lockhart et al., 1996; DeRisi et al., 1996). The microarray makes it possible to measure and generate gene expression levels of thousands of genes simultaneously under different cancerous or normal samples. Gene expression data itself consists of the activation levels of a number of genes from a cell, tissue or organism (Liu et al., 2001). The microarray experiments are used to gather information from tissue and cell samples about gene expression differences that will be useful in diagnosing diseases (Furey et al., 2000). Therefore, it provides a new way for people to understand molecular behaviours in abnormal tissues and make more accurate classifications in cancer diagnosis and treatment. Another important purpose of gene expression analysis is to improve understanding of cellular responses to drug treatment. Cancer genes classification has been central to advances in cancer treatments. Correct classification is crucial to cancer diagnosis and treatment. Moreover, for diagnostic purposes it is important to find small sets of genes that are sufficiently informative to distinguish between cells of different types (Liu et al., 2001). If the small number of genes has succeeded in the distinguishing, then researchers might be able to easily understand the biological significance of these genes.

Microarray technologies provide possibilities to investigate gene activities from whole genome. At the same time, they lead to many issues for computational biologists with large amount of data generated (Xu et al., 2002). The analysis of several thousands of genes at once and relating them to biologically or clinically relevant labels have required molecular biologists and oncologists to collaborate with statisticians and computer scientists who have some experience in producing models of given data (Mukherjee, 2001). For the chemist this might mean determining which

of these compounds might possibly be used as a drug. The molecular biologist may be concerned with which genes are important for certain cell functions and how this genetic pathway works. The development of statistical and computational procedures to address the scientific questions inquired by these experimenters is developing rapidly. Also it is becoming evident that statistical and computational issues such as methods or technologies raise what scientific questions can be answered and what breakthroughs will be made. Hence, computational techniques may provide assistance to computer scientists to improve the identification accuracy system (Su et al., 2002).

Currently there are two types of analysis of gene expression microarray data. The types of analysis are called clustering (unsupervised) and classification (supervised). Most approaches in the early era of gene expression microarray classification were based on clustering methods. The clustering methods are aimed at partitioning the set of genes into subsets that are expressed similarly across different conditions. A variety of approaches to sort and cluster gene expression data have been proposed (Alon et al., 1999; Cho et al., 1998; Eisen et al., 1998; Heyer et al., 1999; Tamayo et al., 1999). The clustering methods have been demonstrated to identify functionally related families of genes (Ben-Dor et al., 1999; DeRisi et al., 1997; Chu et al., 1998; Eisen et al., 1998; Iyer et al., 1999; Wen et al., 1998). Similarly, the clustering methods can be used to divide a set of cell samples into clusters based on their expression profile.

Clustering method, however, does not use any tissue annotation (e.g. tumor vs. normal) in the partitioning step (Ben-Dor et al., 2000). This information is only used to assess the success of the method. Moreover, regardless of the method used for class discovery, the challenge faced is in validation of the clusters (Slonim et al., 2000). Any clustering algorithm will find clusters of samples in gene expression data. However, given relatively few samples and thousands of gene expression vectors, one needs to show that the class distinction discovered is biologically interesting rather than coincidental artifacts of the data. Although this method provides a very informative visualization of the clustered data, it impact lack

robustness and does not have favorable scalability properties (Tamayo et al., 1999). This is mainly because of its huge memory demands in the case of very large data sets, which is typical of genome expression data clustering problems (Xu et al., 2002).

In contrast, supervised method attempts to classify the new tissues based on their gene expression profiles after training on examples that have been classified by classifier. This process is called classification in machine learning communities. The supervised methods have been shown to be able to distinguish various biological classes with a very low error rate without using any prior biological knowledge or expert interpretation (Golub et al., 1999; Brown et al., 1999; Furey et al., 2000; Mukherjee, 2001; Liu et al., 2001). For the gene expression microarray classification, classifier that discriminates between classes is constructed using samples in training set and the constructed classifier will be evaluated using the samples in testing set. Gene expression classification model typically consists of two steps: selecting informative genes (features) and doing the classification from the expression patterns of these genes. Figure 1.1 illustrates a general model for classifying gene expression microarray into the defined biological classes.

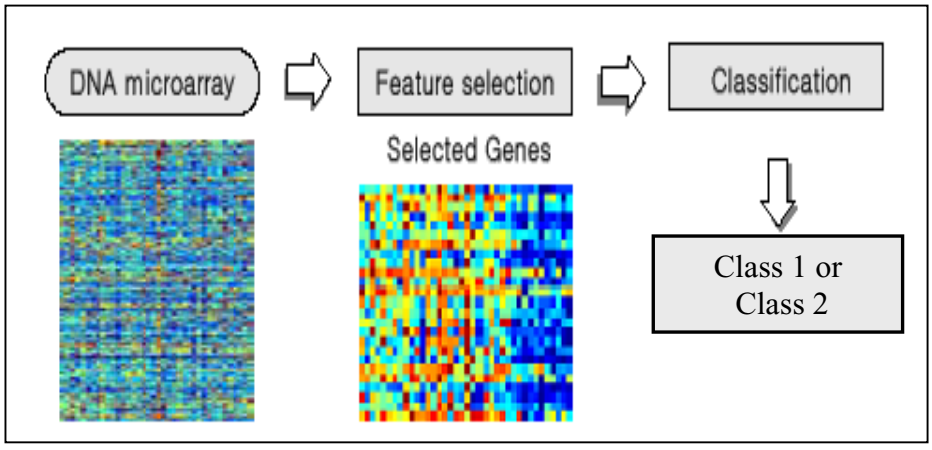

Figure 1.1: A general model for gene expression microarray classification.

The model falls in the learning from examples paradigm of supervised learning. In this paradigm, a mapping is learned from training data (gene expression

patterns) to a label that can be a biological class or a continuous value. One then tests the accuracy of this mapping on testing data. In this type of analysis, the computational task is to correctly classify and predict the state of the individual based on the given genetic profile, i.e., diseased or not diseased. The researches that apply machine learning techniques to perform the classification on real microarray data can be found in papers such by Golub et al. (1999), Slonim et al. (2000) and Furey et al. (2000).

From the favorable results of Support Vector Machine classifier (SVM) in previous experimental studies, the classification of gene expression microarray may potentially benefit from the classifier's performance as well as (Mukherjee, 2001), support the specific challenges imposed by the gene expression microarray classification. Moreover, SVM classifier has many advantages such as flexibility in choosing a similarity function, sparseness of solution when dealing with large data set, the ability to handle large feature space, and the ability to identify outlier (Brown et al., 1999).

Following good results obtained from the application of hybrid Neural Network classifier and Genetic Algorithm (GA) (Yang and Hanovar, 1998), hybrid SVM classifier and GA (Spulveda-Sanchis et al., 2002; Eads et al., 2002; Li et al., 2005), and hybrid Weight Voting (WV) classifier via GA (Liu et al., 2001) experimental studies, have evidently showed the GA offers a particularly attractive approach for optimization of feature subset selection. The results reported in the Yang and Hanovar (1999) experiments used a wide range of real world data sets such as document and artificial data sets from machine learning data repository at the University of California. Sepulveda-Sanchis et al. (2002) have predicted the Unstable Angina data set, while Eads et al. (2002) have classified the Time Series data set. Liu et al. (2001) have used gene expression microarray data sets for their research. The data sets are Leukemia Cancer and Colon Cancer. A hybrid between GA and SVM classifier was first used to classify gene expression microarray and proposed by Li et al. (2005). They have applied this method to the gene expression microarray data, namely diffuse large B cell lymphoma.

More importantly, in spite of excellent performance of hybrid GA and classifier in the experiments conducted by the previous works, currently, there has been little of further work reported pertaining to its applications in other classification problems or any improvement to the original hybrid scheme. Taking these factors into consideration, the gene expression microarray classification, which poses several challenges, will present a good platform to further investigate and possibly improve the hybrid method. The great performances and ability of GA and SVM classifier for the applications also may guarantee their application successfully in gene microarray classification. Hence, this work will incorporate the GA with SVM classifier for classification of gene expression microarray. The hybrid will be known as GASVM.

1.4 Objectives of Research

The goal of this research is to develop and improve a hybrid of Genetic Algorithm and Support Vector Machine classifier (GASVM) for feature selection and classification of gene expression microarray data. In order to reach the goal, several objectives need to be achieved:

- To determine the attributes of the GASVM on small and high dimension data in order to know the limitations.
- To design and develop an improved GASVM (New-GASVM) by using new algorithm based on the improved chromosome representation in order to overcome the challenges posed by gene expression microarray classification.
- To classify the gene expression microarray data using the New-GASVM and to see how it is comparable with existing methods.

1.5 Scope of Research

Since the goal of this research is to evaluate the applicability of the hybrid GA via SVM classifier for gene expression microarray classification, the scope of study is stated below:

- This research focuses on hybrid GA and SVM classifier for feature selection and classification process.
- Chromosome representation in GASVM will be improved and replaced in order to improve the GASVM performances.
- The classification can be confined to two classes only because the results of the two classes would be sufficient enough to gauge the performance of the GASVM when making comparison with existing methods.
- This research uses two data sets for the experiment, namely Leukemia Cancer and Colon Cancer data sets.

1.6 Overview of the Thesis

A general description of the contents of subsequent chapters in this thesis is given as follows:

- Chapter II describes and compares related researches on gene expression microarray classification, feature selection and hybrid of feature selection method via classifier.
- Chapter III describes the overall methodology adopted. The data sets used and the evaluation measures of the method performance are also discussed.
- Chapter IV provides a description of GASVM to select subset of features and make classification process. This chapter also empirically examines

the attributes of the GASVM. Lastly, this chapter also lists several possible limitations.

- Chapter V introduces the efforts in analyzing the possible limitations of GASVM, proposes potential solution and then evaluates them.
- Chapter VI presents the gene expression microarray classification method employed in this research. The proposed method is tested, analyzed and evaluated in the gene expression microarray benchmark data sets.
- Chapter VI concludes the thesis and provides suggestions for future research.