Studies on Intelligent Approaches to Select Informative Genes from Gene Expression Data for Cancer Classification

メタデータ	言語: eng 出版者: 公開日: 2010-07-26 キーワード (Ja): キーワード (En): 作成者: Mohd, Saberi Bin Mohamad メールアドレス: 所属:
URL	https://doi.org/10.24729/00000030

Studies on Intelligent Approaches to Select Informative Genes from Gene Expression Data for Cancer Classification

Mohd Saberi Bin Mohamad

January 2010

Doctoral Thesis at Osaka Prefecture University

Abstract

Gene expression technology namely microarrays, offers the ability to measure the expression levels of thousands of genes simultaneously in biological organisms. Gene expression data produced by the microarrays are expected to be of significant help in the developments of efficient cancer diagnoses and classification platforms. Many researchers have analyzed gene expression data to select a small subset of informative genes for cancer classification using various intelligent approaches. As a result, the selection of the small subset has improved classification accuracy. However, due to the small number of samples compared to the huge number of genes (high-dimension), irrelevant genes, and noisy genes, the most approaches face difficulties to select the small subset. Therefore, the ultimate goal of this research is to propose intelligent approaches for selecting a small (near-optimal) subset of informative genes from gene expression data for cancer classification. Support vector machine classifiers (SVMs) were used to measure classification accuracies on the gene subsets that produced by all the proposed approaches. The first six proposed approaches were produced based on genetic algorithms (GAs), whereas the remaining approaches were extensions of particle swarm optimization (PSO).

First, a multi-objective strategy in a hybrid of GAs and SVMs (GASVM) was proposed to improve the performance of GASVM that uses a single-objective approach. It is called MOGASVM. The strategy has been developed based on maximization of classification accuracy and gene subset size minimization. In this strategy, multi-objective problems have been accommodated by using specialized fitness functions in GAs. The ultimate goal of the strategy is to search and select a nondominated gene subset Pareto front. It was tried on four benchmark gene expression data sets and obtained encouraging results on those data sets as compared with an approach that used a single-objective strategy in GASVM.

Second, an approach using two hybrid methods was then introduced. This approach includes MOGASVM and an improved GASVM (GASVM-II). It was proposed to overcome the limitations of MOGASVM and GASVM-II that developed separately before. In the first phase, GASVM-II is applied to manually select genes from overall gene expression data in order to produce a subset of genes. It is used to reduce the dimensionality of the data, and therefore the complexity of the search or solution spaces can also be decreased. In the second phase, MOGASVM is used to select and optimize a small subset of informative genes from the subset that is produced by the first phase. The approach was assessed and evaluated on four well-known gene expression data sets, showing competitive results.

Third, a cyclic hybrid method based on GASVM-II has been proposed. It differs from other GASVM-based methods in one major part, namely it involves a cyclic approach, whereas the GASVM-based methods did not use any cyclic approach. Basically, the cyclic hybrid method repeats the process of GASVM-II to iteratively reduce the dimensionality of data and produce potential gene subsets. Five real gene expression data sets were used to test the effectiveness of the method. Experimental results show that the performance of the proposed method is superior to other experimental methods and previous related works in terms of classification accuracy and the number of selected genes. In addition, a scatter gene graph and the list of informative genes in the best gene subsets are also presented for biological usage.

Fourth, an iterative approach based on MOGASVM is then developed. Generally, it is almost completely the same with the proposed cyclic hybrid, but it uses MOGASVM to replace GASVM-II in the process to yield potential gene subsets and reduce the dimensionality of data iteratively. To demonstrate its effectiveness, four gene expression data sets are used. Experimental results show that the approach is efficient in finding genes for classifying cancer classes.

Fifth, a two-stage method was proposed to surmount the drawbacks of GASVM-based methods in previous related works. In the first stage, a filter method such as gain ratio (GR) or information gain (IG) is applied on overall gene expression data to preselect genes and finally produce a subset of genes. The dimensionality of data is also can be decreased. The second stage applies MOGASVM to automatically optimize the gene subset that is produced by the first stage. As a result, it yields a small (near-optimal) subset of informative genes. The two-stage method was evaluated on four publicly

available gene expression data sets. The results show that the proposed method outperforms existing methods and other experimental methods.

Sixth, since a two-stage method does not perform well as expected, a three-stage method that includes frequency analysis in the third stage was proposed. The frequency analysis is implemented to identify the most frequently selected genes in near-optimal gene subsets. The most frequently selected genes are presumed to be the most relevant for the cancer classification. The three-stage method differs from methods in previous works in one major part. The major difference is that it involves three stages (using a filter method, a hybrid method, and frequency analysis), whereas the previous works usually had only one stage (using a filter method or a hybrid method) or two stages (using a filter method and a hybrid method). The proposed method has been tested and evaluated for gene selection on five gene expression data sets that contain binary classes and multi-classes of tumor samples. Based on the experimental results, the performance of proposed method is better than that of other methods in previous related works. The list of informative genes in the final gene subset is also presented for biological usage.

Seventh, a modification of binary PSO was proposed to overcome the limitations of the conventional version of binary PSO and previous PSO-based methods. A scalar quantity called particle’s speed and a novel rule for updating particle’s positions are introduced in this modified binary PSO. This particle’s speed and rule are proposed in order to reduce the probability of genes to be selected for the cancer classification. By performing experiments on 12 different gene expression data sets, the modified binary PSO outperforms other previous related works, including the conventional version of binary PSO in terms of classification accuracy, the number of selected genes, and running times.

Eighth, an enhancement of binary PSO with the constraint of particle’s velocities was proposed. The constraint is introduced in the enhanced binary PSO to increase the probability of genes to be unselected for the classification. Experimental results on five actual gene expression data sets show that the performance of the proposed approach is superior to other previous related works, as well as to conventional binary PSO tried in this work.

Ninth, a modified sigmoid function and the particle's speed were introduced and implemented in binary PSO. This modified sigmoid function and particle's speed decrease the probability of genes to be selected for the cancer classification. The proposed method was experimentally assessed on five well-known gene expression data sets. In this sense, comparisons with the existing of binary PSO and several PSO-based methods show competitive results.

As a conclusion, 12 benchmark gene expression data sets have been used in this research to test the effectiveness of the proposed intelligent approaches. Overall, experimental results show that the performances of the proposed approaches are superior to previous related works as well as methods experimented in this work in terms of classification accuracy, the number of selected genes, and running times.

Acknowledgments

This thesis is an outcome of the contributions of many people. I would like to express my sincere gratitude and appreciation to my supervisor, namely Professor Dr. Sigeru Omatu for his intelligent guidance, encouragement, patience, motivation, and invaluable suggestion that made this research possible to be successfully completed in short time. I also appreciate his immediate and very thoughtful responses to my studies and life problems. Sincere appreciations and thanks are extended to Professor Dr. Safaai Deris, who proposed me to pursue my studies under supervision of Professor Dr. Sigeru Omatu. Furthermore, kindly thanks are extended to Professor Dr. Hidetomo Ichihashi and Professor Dr. Keinosuke Matsumoto those who have reviewed and commented this thesis in order to improve it. It is also a great pleasure to express my appreciation to Associate Professor Dr. Michifumi Yoshioka and all other members of the Intelligent Signal Processing Laboratory for their kind help and friendship in many aspects of this research and my life in Japan.

In addition, I would like to take this opportunity to thank Universiti Teknologi Malaysia for permission to fully allow my studies in Osaka Prefecture University, Japan. I also owe a great debt to Malaysian Ministry of Higher Education and Universiti Teknologi Malaysia for the scholarships of my studies. My sincere appreciation also goes to all the people that have answered my email queries so promptly, open source communities, Google, and to all those authors who have generously made their publications and lecture notes available online. Special thanks also to Osaka Prefecture University and Universiti Teknologi Malaysia for subscribing the databases of journals, conferences, and lecture notes where I can freely download the articles by online.

On a personal note, I would also like to give my best thanks to my family especially for my wife (Nur Yumni Binti Yahya), parents, grandmother, parents in law, and my supervisor’s wife for their love, help, support, encouragement, and understanding throughout this endeavor. I am greatly indebted to all of them. Above all, the thanks and gratefulness go to my God (ALLAH) for all the good award.

Contents

List of Figures

List of Tables

List of Abbreviations

3-SGS	A three-stage method
ADCA	Adenocarcinoma
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
BPSO	The conventional version of binary particle swarm optimization
C-GASVM	A cyclic hybrid method
CPSO	An enhancement of binary particle swarm optimization with the proposed constraint of particle’s velocities
DLBCL	Diffuse large B-cell lymphomas
Filter+MOGASVM	A two-stage method
GA	Genetic algorithm
GASVM	A hybrid of genetic algorithms and support vector machines
GASVM-II	GASVM version II
GASVM-II+MOGASVM	A combination between GASVM-II and MOGASVM
GPSO	Geometric particle swarm optimization
GR	Gain ratio
IG	Information gain
I-MOGASVM	An interactive approach based on MOGASVM
IBPSO	An improved binary particle swarm optimization
IPSO	An improved (modified) binary particle swarm optimization
LOOCV	Leave-one-out-cross-validation
MOGASVM	Multi-objective GASVM
MLL	Mixed-lineage leukemia
MOO	Multi-objective optimization
MPM	Malignant pleural mesothelioma
New-GASVM	An improved GASVM

PSO	Particle swarm optimization
PSOGA	A hybrid of PSO and GAs
PSOTS	A hybrid of PSO and tabu search
SPSO	An improvement of binary particle swarm optimization based on a modified sigmoid function
SRBCT	Small round blue cell tumors
SVM	A support vector machine classifier

List of Appendixes

Chapter 1

Introduction

Bioinformatics is defined as an application of computation tools to capture, analyze, and interpret biological data. It is an interdisciplinary field, which harnesses computer science, mathematic, engineering, and biology. It represents a relatively new area of computer science and engineering to handle and manage large amounts of data generated by advance technologies which are designed for measuring biological systems. The use of computational intelligence methods in analyzing the biological data is currently at the forefront of its field and represents a major opportunity in the research of computational intelligence communities.

The problem of cancers in this world is a growing one. A traditional cancer diagnosis relies on a complex and inexact combination of clinical and histopathological data. This classic approach often fails when dealing with atypical tumors or morphologically indistinguishable tumor subtypes. Recent advances in microarrays technology have led to a promising future of cancer diagnosis using new molecular-based approaches. This technology allows scientists to measure the expression levels of thousands of genes simultaneously in biological organisms. The most important application of gene expression data that are produced by the microarrays technology is the selection of informative genes for cancer classification. For quick reference, a glossary of structural genomic terms is also provided in Appendix B.

1.1 Microarrays

Microarrays technology is a machine that can be used to measure the expression levels of thousands of genes simultaneously under different cancerous or normal samples [22]. Microarrays experiments are used to gather information from tissue and cell samples about gene expression differences that are useful in diagnosing diseases [9]. It produces gene expression data as the final product. Therefore, it provides a new way for people to understand molecular behaviors in abnormal tissues and improve classification performances for accurate in cancer diagnosis and treatment. At the same time, the microarrays lead many issues for biologists with the large amount of data generated [37]. These issues have required molecular biologists to collaborate with computer scientists who have some experience in the development of intelligent approaches for processing and analyzing the huge amount of data [28]. This research uses gene expression data to select information genes for cancer classification.

Usually, the matrix of gene expression data, $G_{N\times(M+1)}$ contains different values of gene expression levels on a large scale. This matrix is organized as shown in Fig. 1.1, where
M = the total number of genes in each sample of $G_{N\times(M+1)}$,
N = the total number of samples in $G_{N\times(M+1)}$,
$g_{i,j}$ = a numeric value of the gene expression level of the *j*th gene in the *i*th sample,
$i = 1, 2, .., N;\ \ j = 1, 2, .., M$,
l_i = a class label for the *i*th sample, $l_i \in \{-1,+1\}$ for binary classes and $l_i \in \{+1,+2,...,+C\}$ for multi-classes where $+C =$ the total number of classes in $G_{N\times(M+1)}$.

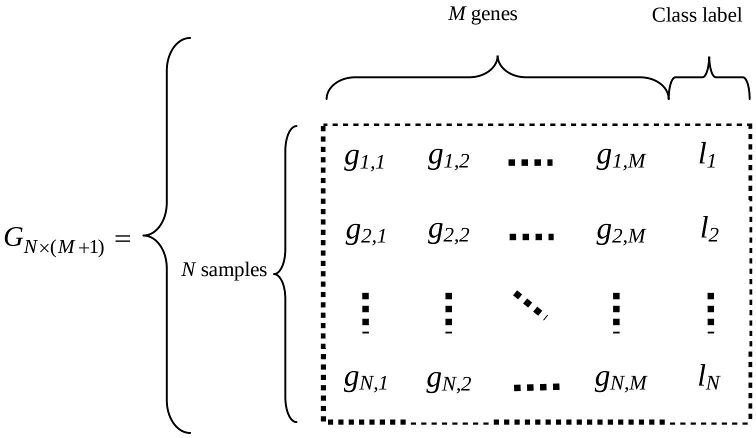


Fig. 1.1. The matrix of gene expression data.

1.2 Cancer Classification Based on Gene Expression Data

Recently, there are many classifiers such as support vector machines (SVMs), neural networks, etc. have been used for cancer classification of gene expression data. Based on the favorable results of SVMs from previous works [6],[9],[28], this research uses SVMs to classify cancer classes. Moreover, SVMs have many advantages such as flexibility in choosing a similarity function, sparseness of solution when dealing with large data, the ability to handle large feature space, and the ability to identify outliers [6]. The detail of SVMs can be found on in Mukherjee's thesis [28].

A cancer classification model has two phases: 1) Gene selection; and 2) classification [10]. The first phase uses a gene selection method to select genes, while in the phase stage; a classifier is implemented to perform classification process. Figure 1.2 shows the model of cancer classification.

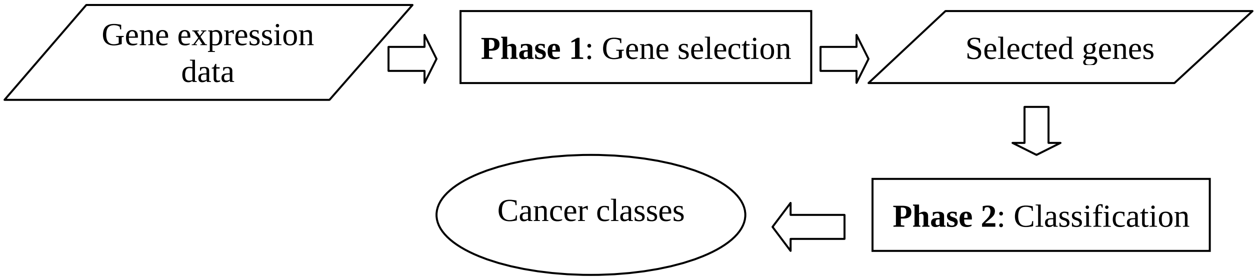


Fig. 1.2. The model of cancer classification.

Most previous works widely used two manners for measuring the accuracy of cancer classification. The manners are the leave-one-out-cross-validation (LOOCV) procedure to obtain LOOCV accuracy, and test accuracy measurement to produce test accuracy. For the LOOCV procedure, one sample from the training set is excluded, and the rest of training samples, *N*-1 are used to build a classifier. Then, the built classifier is used to predict the class that has been left out, and this process is repeated for each sample in the training set. The LOOCV accuracy is obtained by the overall number of correct classifications, divided by the number of samples in the training set, *N*. For the test accuracy measurement, the final classifier is built using all the training samples, and the classes of test samples from the testing set are classified one by one using the built classifier. The test accuracy is estimated by the number of the correctly classified test samples, divided by the number of samples in the testing set. In this research, the LOOCV procedure is used for measuring classification accuracy on the training set due to the small number of samples in gene expression data, and most previous works also used it; whereas for the calculation of classification accuracy on the testing set, the test accuracy measurement is used.

1.3 Gene Selection from Gene Expression Data

The selection of a small subset of informative genes from thousands of genes is a critical step for accurate cancer classification. Usually, a gene selection method is used to select a subset of informative genes that maximizes the classifier's ability to classify samples more accurately [24]. In pattern recognition domain, gene selection is called feature selection. The gene selection has several advantages:

- Maintain or improve classification accuracy.
- Reduce the dimensionality of data.
- Yield a small subset of genes.
- Remove irrelevant and noisy genes.
- Decrease computational times.
- Reduce the cost in a clinical setting.

In the context of cancer classification, gene selection methods can be classified into two categories [24]. Figure 1.3 shows the difference between the categories. If a gene selection method is carried out independently from a classification procedure, it belongs to the filter method. Otherwise, it is said to follow a hybrid (wrapper) method. Signal to noise ratio [9],[10], threshold number of misclassification scores [4], cosine coefficient, information gain, and euclidean distance [7] are some of the widely known the filter method. The hybrid method is performed dependently on classifiers and conducted in search space for selecting and evaluating subsets of genes. In this approach, a classifier is included as a part of its evaluation function. Furthermore, it performs some form of state space searches to select genes in order to maximize the evaluation function. This evaluation process is repeated until a condition has been satisfied.

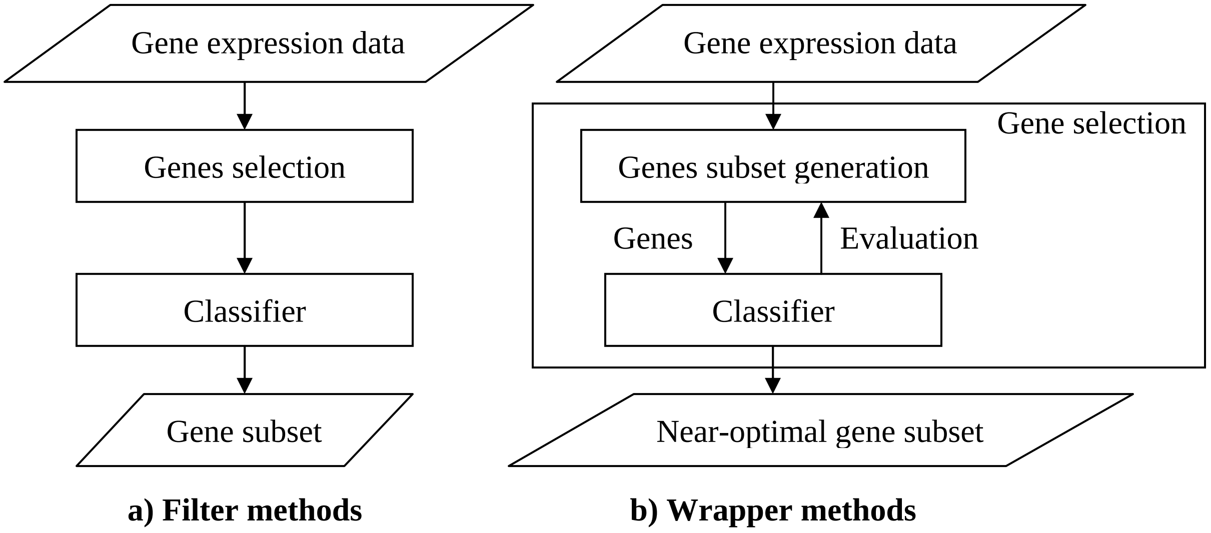


Fig. 1.3. The categories of gene selection methods.

In the early era of microarrays analyses, most previous works have used the filter method to select genes since it is computationally more efficient than the hybrid method [3],[11],[19],[36]. Many filter methods are normally mentioned as individual gene-ranking methods. Usually, they evaluate a gene based on its discriminative power for the target classes without considering its correlations with other genes. This mechanism may result in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification. These genes increase the dimensionality of the gene subset, and in turn affect the classification performance. A few years ago, several hybrid methods, especially a hybrid of genetic algorithms (GAs) and classifiers, have been implemented to select informative genes [8],[23],[26],[31]. Recently, several gene selection methods based on particle swarm optimization (PSO) have been proposed to replace GA in the hybrid method. PSO is a new population based stochastic optimization technique proposed by Kennedy and Eberhart [14]. It is motivated from the simulation of social behaviors of organisms such as bird flocking and fish schooling. The hybrid methods usually provide greater accuracy than the filter methods since genes are selected by considering and optimizing correlations among genes.

1.4 Problem Statements

Although the mechanism of cancer classifications 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 [10]. It is because there can be so many pathways causing cancer. The traditional methods of cancer classifications are mostly dependent on the morphological appearance of tumors and their applications are limited by existing uncertainties [10]. Moreover, the methods also have various limitations especially in discriminating between two similar types of cancers. Therefore, microarrays technology has been introduced to solve the limitations of the methods by offering ability to measure the gene expression levels of thousands of genes simultaneously in biological organisms such as human and animal. Due to the large amount of gene expression data generated by the microarrays technology, computational intelligence approaches are needed to analyze and process the data.

Almost all the computational intelligence approaches for the cancer classification of gene expression data started with gene selection methods [7]. Thus, there is a need to firstly select informative genes that contribute to a cancerous state by using gene selection methods [26] in order to maximize the classifier's ability to classify samples more accurately. An informative gene is a gene that is useful for cancer classification. However, the gene selection process poses a major challenge because of the following characteristics of gene expression data: The huge number of genes compared to the small number of samples (high-dimensional data), irrelevant genes, and noisy data.

1.5 Goal and Objectives of the Research

The ultimate goal of this research is to propose intelligent approaches based on GAs and PSO for selecting a small (near-optimal) subset of informative genes from gene expression data for cancer classification. In order to reach the goal, several objectives need to be achieved:

- To propose a multi-objective strategy in GASVM for improving the performance of GASVM that uses a single-objective approach.
- To propose an approach using two hybrid methods in order to reduce the complexity of data and optimize genes subsets.
- To propose a cyclic hybrid method based on GASVM-II for repeatedly reducing the dimensionality of data and producing potential gene subsets.
- To propose an iterative approach based on MOGASVM in order to decrease the complexity of data.
- To propose a two-stage method using a filter method and a hybrid method for preselecting and optimizing gene subsets, respectively.
- To propose a three-stage method that includes frequency analysis in its third stage in order to identify the most frequently selected genes in near-optimal gene subsets.
- To propose a modification of binary particle swarm optimization based on introduced particle's speed and a novel rule in order to overcome the limitations of the conventional version of binary PSO.
- To propose an enhancement of binary PSO with the constraint of particle’s velocities for increasing the probability of genes to be unselected for cancer classification.
- To propose an improvement of binary particle swarm optimization based on a modified sigmoid function and introduced particle's speed in order to solve the weaknesses of the conventional version of binary PSO and previous PSO-based methods.

1.6 Scopes of the Research

Since the goal of this research is to introduce and propose intelligent approaches in selecting a small subset of informative genes for cancer classification, the scopes of this research are stated as follows:

- Focusing on the modifications and enhancements of GASVM-based methods and PSO-based methods for selecting genes from gene expression data.
- Using and applying SVMs to classify samples from genes subsets that produced by GASVM-based methods and PSO-based methods.
- Conducting experiments on 12 public and benchmark gene expression data sets that contain binary classes and multi-classes: Colon, leukemia (Leukemia1), lung, MLL (Leukemia2), SRBCT, 11_Tumors, 9_Tumors, Brain_Tumor1, Brain_Tumor2, Lung_Cancer, Prostate_Tumor, and DLBCL. These data sets can be freely accessed by online and will be explained in the next chapters.

1.7 Organization of the Thesis

This thesis is organized into 11 chapters. The general information of each chapter is given as follows:

- Chapter 1 introduces main keywords that used in this research such as GAs, PSO, gene selection, microarrays, gene expression data, and cancer classification. It also describes about problem statements, goal, and scopes of this research.
- Chapter 2 describes a multi-objective strategy in GASVM (MOGASVM) for gene selection. It is proposed to improve the performance of GASVM that uses a single-objective approach. In this strategy, multi-objective problems have been accommodated by using a specialized fitness function in GAs.
- Chapter 3 discusses an approach using two hybrid methods. These hybrid methods are MOGASVM and GASVM-II. It was developed to overcome the limitations of MOGASVM and GASVM-II that developed separately before.

- Chapter 4 describes a cyclic hybrid method based on GASVM-II. Generally, this method repeats the process of GASVM-II to produce potential gene subsets and reduce the dimensionality of data repeatedly.
- Chapter 5 concerns on the discussion of an iterative approach based on MOGASVM. Basically, it is almost completely the same with the proposed cyclic hybrid, but it uses MOGASVM to replace GASVM-II.
- Chapter 6 discusses a two-stage method. It is proposed by using a filter method and a hybrid method to surmount the drawbacks of GASVM-based methods in previous related works.
- Chapter 7 describes a three-stage method that includes frequency analysis as an extra process in the last stage. This frequency analysis is implemented to identify the most frequently selected genes in near-optimal gene subsets.
- Chapter 8 introduces a modification of binary PSO. A scalar quantity called particle's speed and a novel rule for updating particle's positions are introduced in this modified binary PSO.
- Chapter 9 discusses an enhancement of binary PSO. The constraint of particle's velocities is introduced in the enhanced binary PSO to increase the probability of genes to be unselected for the classification.
- Chapter 10 describes a modification of binary PSO. A modified sigmoid function and the particle's speed are implemented in this modified binary PSO. Both the implementations decrease the probability of genes to be selected for the cancer classification.
- Chapter 11 gives the conclusion remarks of obtained results of all the proposed intelligent approaches, and suggests interesting ideas for the future research.

Chapter 2

A Multi-Objective Genetic Algorithm

2.1 Introduction

Multi-objective optimization (MOO) is an optimization problem that involves multiple objectives or goals. Generally, the objectives estimate different aspects of solutions. It is necessary to be aware that gene selection is a MOO problem in the sense of classification accuracy maximization and gene subset size minimization. Therefore, Chapter 2 describes and proposes a multi-objective strategy in GASVM for gene selection and the classification of gene expression data. This is known as MOGASVM.

2.2 A Multi-Objective Strategy in GAs

MOGASVM was developed to improve the performance of GASVM that uses a single-objective [23]. All information about GASVM such as flowcharts, algorithms, chromosome representations, fitness functions, and parameter values are available in Mohamad *et al*. [23].

In the sense of classification accuracy maximization and gene subset size minimization, gene selection can be viewed as an MOO problem. Formally, each gene subset (a solution) is represented by *x* (an *n*-dimensional decision vector). It is associated with a vector objective function $f(x)$:

$$f(x) = (f_1(x), f_2(x), ..., f_m(x)) \tag{2.1}$$

with $x = (x_1, x_2, ..., x_n) \in X$ where X is the decision space, i.e., the set of all expressible solutions. The vector objective function $f(x)$ maps X into $\Re^m$, where $\Re$ is the objective space and $m \geq 2$ is a number of objectives. f_i is the ith objective. The vector $z = f(x)$ is an objective vector. The image of X in the objective space is the set of all attainable points z (Fig. 2.1). If all objective functions are for maximization, a subset x is said to dominate another x (x^*) if and only if:

$x > x^*$ if

$$\forall i \in \{1,...,m\}, f_i(x) \geq f_i(x^*) \wedge \exists j \in \{1,...,m\}, f_j(x) > f_j(x^*)$$

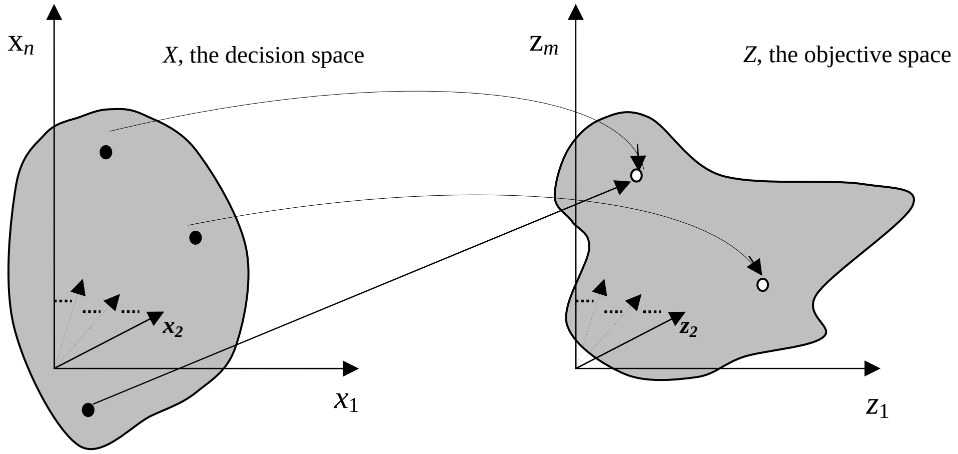


Fig. 2.1. The n-dimensional decision space maps to the m-dimensional objective space.

A solution (gene subset) is said to be Pareto optimal if it is not dominated by any other solutions in the decision space. A Pareto optimal solution cannot be improved with respect to any objective without worsening at least one other objective. The set of all feasible nondominated solutions in X is referred to as the Pareto optimal set, and for a given Pareto optimal set, the corresponding objective function values in the objective space are called the Pareto front [12].

The Pareto front in this research is defined as the set of nondominated gene subsets. MOGASVM is one promising approach to find or approximate the Pareto front. The roles of this approach are guided by the search towards the Pareto front while keeping the nondominated solutions as diverse as possible. Therefore, the original GASVM is customized to accommodate multi-objective problems by using specialized fitness functions. The ultimate goal of MOGASVM is to identify a nondominated gene subset Pareto front. This subset (individual) is evaluated by its accuracy on the training data and the number of genes selected in it. These criteria are denoted as f_1 and f_2 separately, and are used in a fitness function. Therefore, the fitness of individuals is calculated by Eq.(2.4) as follows:

$$f_1 = w_1 \times A(x) \tag{2.2}$$

$$f_2 = w_2 \times ((M - R(x)) / M) \tag{2.3}$$

$$fitness(x) = f_1 + f_2 \tag{2.4}$$

where $A(x) \in [0,1]$ is the leave-one-out-cross-validation (LOOCV) accuracy of the training data using only the expression values of the selected genes in a subset x, where $R(x)$ is the number of selected genes in x. M is the total number of genes, w_1 and w_2 are two priority weights corresponding to the importance of the accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$, and f_2 is calculated as above in order to support the maximization function of the minimization of gene subset size. In this research, the accuracy is more important than the number of selected genes (gene subset size).

Ambroise and McLachlan [2] have indicated that because of “selection bias”, the test results could be over-optimistic if the test samples were not excluded from the classifier building process in a hybrid approach. Therefore, the proposed MOGASVM totally excludes the test samples from the classifier building process in order to avoid the influence of the bias.

2.3 Experimental Results

2.3.1 Data sets

Two benchmark gene expression data sets are used to evaluate the proposed approach; leukemia cancer, colon, lung cancer, and mixed-lineage leukemia (MLL) cancer. Table 2.1 summarizes the data sets.

Table 2.1. The summary of gene expression data sets.

Data set	No. classes	No. samples in the training set	No. samples in the test set	No. genes	Source
Leukemia	2 (ALL and AML)	38 (27 ALL and 11 AML)	34 (20 ALL and 14 AML)	7,129	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Lung	2 (MPM and ADCA)	32 (16 MPM and 16 ADCA)	149 (15 MPM and 134 ADCA)	12,533	http://chestsurg.org/publications/2002-microarray.aspx.
MLL	3 (ALL, MLL, and AML)	57 (20 ALL, 17 MLL, and 20 AML)	15 (4 ALL, 3 MLL, and 8 AML)	12,582	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Colon	2 (Normal and tumor)	62 (22 normal and 40 tumor)	Not available	2,000	http://microarray.princeton.edu/oncology/affydata/index.html

Note:

MPM = malignant pleural mesothelioma.
ADCA = adenocarcinoma.
ALL = acute lymphoblastic leukemia.
MLL = mixed-lineage leukemia.
AML = acute myeloid leukemia.

For the leukemia, Lung, and MLL cancer data set, the LOOCV procedure is applied on the training set to obtain LOOCV accuracies, and the accuracy test measurement is applied on the testing set to measure test accuracy. However, for the colon cancer data set, only the LOOCV procedure is used because this data set only has the training set.

2.3.2 Experimental setup

Three important criteria are used to evaluate the MOGASVM performances; test accuracy, LOOCV accuracy, and the number of selected genes.

The experimental results presented in this section pursue two objectives. The first objective is to show that gene selection using MOGASVM is needed in order to reduce the number of genes and achieve better classification of the gene expression data. The second objective is to show that MOGASVM is better than the original version of GASVM that used a single-objective approach. To achieve these objectives, several experiments were conducted, 10 times each, for both data sets using different values of w_1 and w_2 ($w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$). The subset that produces the highest LOOCV accuracy with the lowest number of selected genes is chosen as the best subset. SVM, GASVM (single-objective), and GASVM version 2 (GASVM-II) were also used in this research as a comparison with MOGASVM. GASVM-II has been proposed by Mohamad *et al*. [23].

2.3.3 Result analysis and discussion

Table 2.2 and Table 2.3 show the results of the experiments for all the data sets using different values of w_1 and w_2. A value of the form $x \pm y$ represents an average value x with a standard deviation y. Overall, the classification accuracy and the number of selected genes for both data sets fluctuated because of the diversity of the solutions based on adjusted weights (w_1 and w_2). Moreover, multiple objectives search simultaneously in a run, and consequently populations tend to converge to the solutions which are superior in one objective, but poor at others. The highest averages of LOOCV and test accuracies for classifying leukemia samples were 95.53% and 84.41%, respectively, using $w_1 = 0.8$ and $w_2 = 0.2$, while 93.23% LOOCV accuracy was obtained for the colon data set using $w_1 = 0.7$ and $w_2 = 0.3$. The highest averages of LOOCV accuracy and test accuracy for classifying the lung data set were 73.31% and 85.84%, respectively, while 94.74% and 90%, respectively of the MLL data set. The highest averages of the accuracies of both the data sets were obtained by using $w_1 = 0.7$ and $w_2 = 0.3$.

Table 2.2. Classification accuracies for different gene subsets using MOGASVM on the leukemia and colon data sets (10 runs on average).

Weight		Average for the leukemia data set			Average for the colon data set	
w_1	w_2	Accuracy (%)		No. selected genes	LOOCV accuracy (%)	No. selected genes
		LOOCV	Test			
0.1	0.9	94.74 ± 0.00	84.12 ± 1.52	2,196.5 ± 10.88	90.65 ± 1.27	398.8 ± 6.36
0.2	0.8	95.26 ± 1.11	83.24 ± 2.79	2,205.1 ± 15.19	91.45 ± 1.08	419.5 ± 7.95
0.3	0.7	95.00 ± 0.83	83.24 ± 3.12	2,199.1 ± 25.83	92.58 ± 0.83	429.2 ± 12.22
0.4	0.6	95.53 ± 1.27	83.53 ± 2.48	2,220.8 ± 31.60	92.74 ± 0.85	430.1 ± 10.50
0.5	0.5	95.26 ± 1.11	82.65 ± 3.24	2,231.2 ± 26.84	92.90 ± 0.83	443.0 ± 9.19
0.6	0.4	95.26 ± 1.11	82.65 ± 2.93	2,210.9 ± 25.09	92.26 ± 0.68	429.0 ± 10.37
0.7	0.3	95.00 ± 0.83	83.24 ± 2.79	2,201.4 ± 15.87	93.23 ± 1.02	446.3 ± 18.90
0.8	0.2	95.53 ± 1.27	84.41 ± 2.42	2,212.6 ± 26.63	92.90 ± 1.13	445.9 ± 27.92
0.9	0.1	95.53 ± 1.27	83.82 ± 2.50	2,218.3 ± 28.29	92.26 ± 0.68	435.3 ± 12.89

Note: The best results are shown in the shaded cells. The colon data set only has LOOCV accuracy since it only has the training set.

A total of 2212.6 average genes in a subset were finally selected to obtain the highest accuracies (LOOCV and test) of the leukemia data set, whereas 446.3 average genes were selected of the colon data set. The averages genes of the lung and MLL data sets were 4418.5 and 4465.2 genes, respectively. Hence, the subsets were chosen as the best subsets. The best subsets are called the best-known Pareto front because it is close to the true Pareto front.

MOGASVM has found the best subsets since it distributed successfully diverse gene subsets over a solution space.

Table 2.3. Classification accuracies for different gene subsets using MOGASVM on the lung and MLL data sets (10 runs on average).

Weight		Average for the lung data set			Average for MLL the data set		
w_1	w_2	Accuracy (%)		No. selected genes	Accuracy (%)		No. selected genes
		LOOCV	Test		LOOCV	Test	
0.1	0.9	75 ± 0	84.43 ± 4.16	4,416.5 ± 17.90	94.74 ± 0	88.67 ± 5.49	4,472.1 ± 29.40
0.2	0.8	75 ± 0	85.24 ± 4.68	4,421.3 ± 21.53	94.74 ± 0	89.33 ± 4.66	4,470.6 ± 16.54
0.3	0.7	75 ± 0	84.16 ± 3.79	4,416.6 ± 13.59	94.74 ± 0	88.67 ± 7.06	4,466.9 ± 21.25
0.4	0.6	75 ± 0	81.75 ± 4.30	4,410.3 ± 26.30	94.74 ± 0	89.33 ± 4.66	4,471.4 ± 19.50
0.5	0.5	75 ± 0	84.10 ± 4.78	4,415.7 ± 25.40	94.74 ± 0	89.33 ± 5.62	4,465.3 ± 24.60
0.6	0.4	75 ± 0	84.90 ± 4.04	4,423.2 ± 19.62	94.74 ± 0	88.67 ± 3.22	4,479.2 ± 21.73
0.7	0.3	75.31 ± 0.99	85.84 ± 3.97	4,418.5 ± 50.19	94.74 ± 0	90.00 ± 3.51	4,465.2 ± 18.34
0.8	0.2	75 ± 0	83.22 ± 4.86	4,419 ± 15.25	94.74 ± 0	88.00 ± 6.13	4,479.3 ± 22.24
0.9	0.1	75 ± 0	83.83 ± 4.30	4,423.3 ± 19.66	94.74 ± 0	88.00 ± 6.13	4,468.4 ± 16.03

Note: The best results are shown in the shaded cells.

Table 2.4. The results of the best subsets in 10 runs ($w_1 = 0.8$ and $w_2 = 0.2$ of the leukemia data set, $w_1 = 0.7$ and $w_2 = 0.3$ of the colon, lung, and MLL data sets).

Data set	LOOCV (%)	Test (%)	Experiment no.	No. selected genes
Leukemia	97.37	88.24	4	2,252
Colon	95.16	-	7	446
Lung	78.13	93.29	7	4,433
MLL	94.74	93.33	7	4,437

Table 2.5. The benchmark of MOGASVM with GASVM (single-objective) and SVM on the leukemia and colon data sets.

Method	Leukemia data set (Average; the best)			Colon data set (Average; the best)	
	No. selected genes	Accuracy (%)		No. selected genes	LOOCV accuracy (%)
		LOOCV	Test		
MOGASVM	(2,212.6 ± 26.63; 2,252)	(95.53 ± 1.27; 97.37)	(84.41 ± 2.42; 88.24)	(446.3 ± 18.90; 446)	(93.23 ± 1.02; 95.16)
GASVM (Single-objective)	(3,574.9 ± 40.05; 3,531)	(94.74 ± 0; 94.74)	(83.53 ± 2.48; 88.24)	(979.8 ± 35.80; 940)	(91.77 ± 0.51; 91.94)
SVM	(7,129 ± 0; 7,129)	(94.74 ± 0; 94.74)	(85.29 ± 0; 85.29)	(2,000 ± 0; 2,000)	(85.48 ± 0; 85.48)

Note: The best results are shown in the shaded cells.

Table 2.6. The benchmark of MOGASVM with GASVM (single-objective) and SVM on the lung and MLL data sets.

Method	Lung data set (Average; the best)			MLL data set (Average; the best)		
	No. selected genes	Accuracy (%)		No. selected genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
MOGASVM	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)	(4,465.2 ± 18.34; 4,437)	(94.74 ± 0; 94.74)	(90.00 ± 3.51; 93.33)
GASVM (single-objective)	(6,267.8 ± 56.34; 6,342)	(75.00 ± 0; 75.00)	(84.77 ± 2.53; 87.92)	(6,298.8 ± 51.51; 6,224)	(94.74 ± 0; 94.74)	(87.33 ± 2.11; 86.67)
SVM	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)	(12,582 ± 0; 12,582)	(92.98 ± 0; 92.98)	(86.67 ± 0; 86.67)

Note: The best results are shown in the shaded cells.

All LOOCV results of the leukemia data set were much higher than the test results due to the problem of over-fitting. The data set properties, i.e., thousands of genes with less than a hundred samples in the training sets, probably cause the over-fitting, where a decision surface of the classifier performs well on the training set, but poorly on the test set.

Table 2.4 shows that the best performances (LOOCV and test accuracies) were 97.37% and 88.24%, respectively, for the leukemia data set using 2252 genes. For the colon data set, the highest LOOCV accuracy was 93.55% using 446 genes. The best performances for the leukemia and colon data sets were found in the fourth and seventh experiments, respectively, while for the lung and MLL data sets, the best performances have been found in the seventh experiments.

In Tables 2.5 and 2.6, the LOOCV accuracy, the test accuracy, and the number of selected genes are given in parentheses. The average results are given in the parentheses and the best results are highlighted in the shaded cells. This table shows that the performances of MOGASVM were better than that of GASVM and SVM in terms of the LOOCV accuracy, the test accuracy, and the number of selected genes on average and for the best results. In general, MOGASVM reduces the number of genes to about a quarter of the total, whereas GASVM reduces the number to about a half of the total. This is due to the ability of MOGASVM to search different regions of a solution space simultaneously, and therefore, it is possible to find a diverse set of solutions in a high-dimensional space. Moreover, it may also exploit the structures of good solutions with respect to different objectives to create new nondominated solutions in unexplored parts of the Pareto optimal set. This suggests that gene selection using the multi-objective approach is needed for disease classification of gene expression data.

2.4 Summary

In this chapter, MOGASVM has been designed, developed, and analyzed to solve gene selection problems. By performing experiments on MOGASVM, the present work found that the classification accuracy and the number of selected genes for both data sets fluctuated and were not equal when using different values of w_1 and w_2. This result shows that there are many irrelevant genes in gene expression data, and some of them act negatively on the accuracy acquired by the relevant genes. Generally, MOGASVM achieved significant LOOCV accuracy, test accuracy, and the number of selected genes, and was better than GASVM (single-objective) and SVM because its multi-objective strategy could find a diverse solution in a Pareto optimal set. MOGASVM can also be extended to other applications such as pattern recognitions, computer visions, and cognitive sciences. However, MOGASVM did not achieve the greatest accuracy, and the number of selected genes was still high. Therefore, the next chapter (Chapter 3) will propose GASVM-II+MOGASVM to reduce the number of selected genes and increase classification accuracy.

Chapter 3

An Approach Using Two Hybrid Methods

3.1 Introduction

Mohamad *et al*. [23] have reported that a hybrid of GAs and SVMs (GASVM), and also an improved GASVM called GASVM-II which has both advantages and disadvantages. This present work proposes a new approach called GASVM-II+MOGASVM which utilizes the advantages of MOGASVM and GASVM-II. The advantage of MOGASVM is that it can automatically select and optimize a number of genes to produce a gene subset. However, it performs poorly with high-dimensional data. In contrast, GASVM-II performs well with high-dimensional data. It can also reduce the complexity of search spaces, and may be able to evaluate all possible subsets of genes. Nevertheless, the drawback of GASVM-II is that it selects a number of genes manually to yield a gene subset. Therefore, this chapter proposes and describes an approach using two hybrid methods for selecting informative genes. It is developed to improve the performances of MOGASVM and GASVM-II.

3.2 The Proposed Approach

Figure 3.1 shows that the flowchart of GASVM-II+MOGASVM involves two phases. In the first phase, GASVM-II is applied to manually select genes from the overall gene expression data in order to produce a subset of genes. It is used to reduce the dimensionality of the data, and therefore, the complexity of the search or solution spaces can also be decreased.

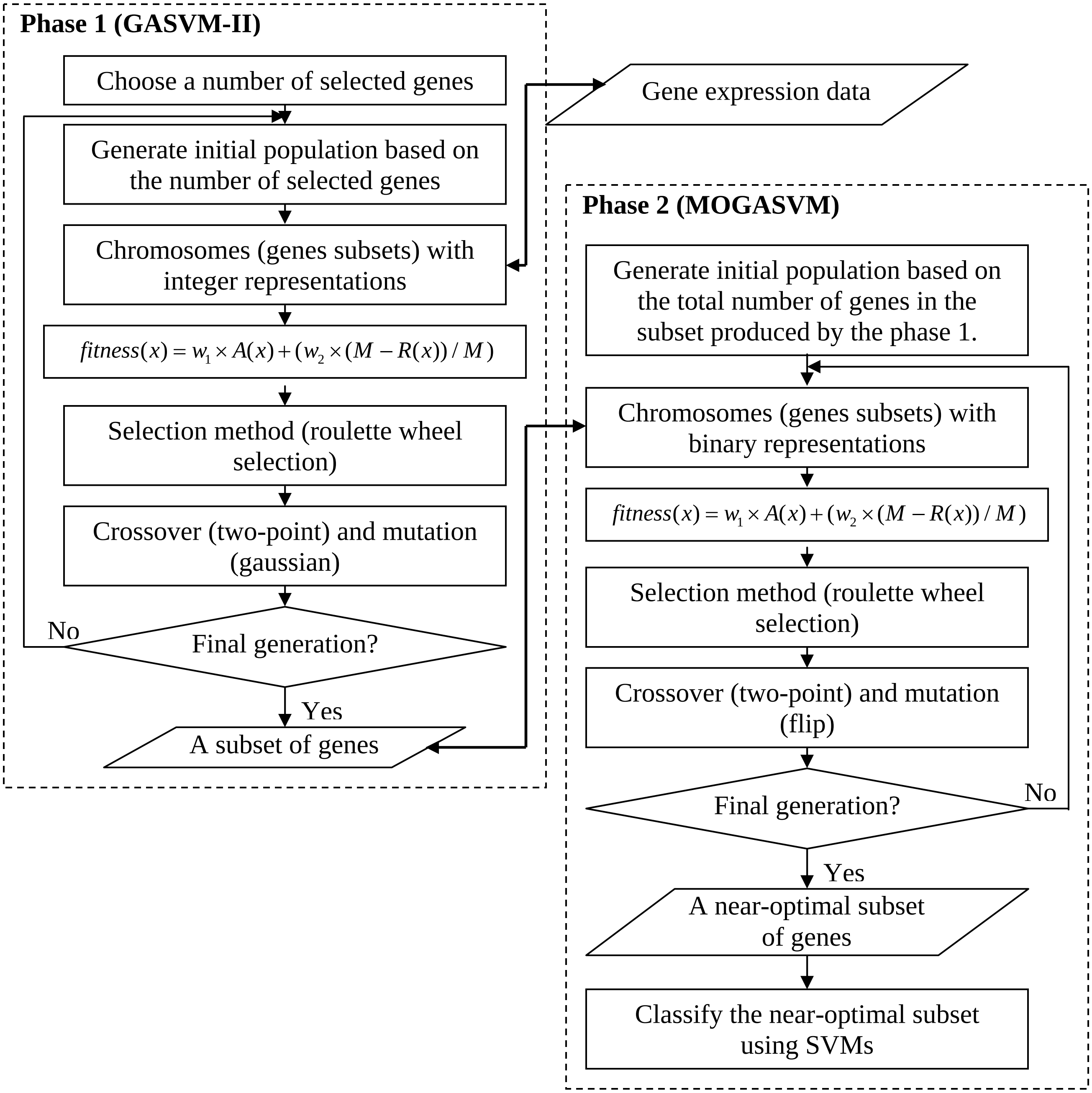

Phase 1 (GASVM-II)
Choose a number of selected genes
Generate initial population based on the number of selected genes
Chromosomes (genes subsets) with integer representations
$fitness(x) = w_1 \times A(x) + (w_2 \times (M - R(x)) / M)$
Selection method (roulette wheel selection)
Crossover (two-point) and mutation (gaussian)
No
Final generation?
Yes
A subset of genes
Gene expression data
Phase 2 (MOGASVM)
Generate initial population based on the total number of genes in the subset produced by the phase 1.
Chromosomes (genes subsets) with binary representations
$fitness(x) = w_1 \times A(x) + (w_2 \times (M - R(x)) / M)$
Selection method (roulette wheel selection)
Crossover (two-point) and mutation (flip)
Final generation?
No
Yes
A near-optimal subset of genes
Classify the near-optimal subset using SVMs

Fig.3.1. The flowchart of GASVM-II+MOGASVM.

In the second phase, MOGASVM is used to select and optimize a small subset of informative genes from the subset that is produced by the first phase. If the size of the subset is small and the combination of genes is not complex, MOGASVM can easily find and optimize the subset. MOGASVM is applied because it can automatically select a number of genes and finally produce an optimized gene subset. This second phase can also remove noisy genes because the first phase has reduced the size and complexity of the search spaces. The fitness of individuals is calculated as follows:

$$fitness(x) = w_1 \times A(x) + (w_2 \times (M - R(x)) / M) \tag{3.1}$$

in which $A(x) \in [0,1]$ is the LOOCV accuracy on the training data using only the expression values of the selected genes in a subset x, and $R(x)$ is the number of selected genes in x. M is the total number of genes, and w_1 and w_2 are two priority weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$. In this chapter, the accuracy is more important than the number of selected genes.

3.3 Experimental Results

3.3.1 Data set

Four benchmark gene expression data sets are used to evaluate the proposed approach, i.e., those for leukemia cancer, colon cancer, lung, and MLL. The summary of the data sets has been shown on Table 2.1 in Chapter 2. For the leukemia cancer, lung cancer, and MLL cancer data sets, LOOCV is applied on the training set, and an accuracy test measurement is carried out on the test set to measure the classification accuracy. However, for the colon cancer data set, only the LOOCV procedure is used because this data set only has the training set.

3.3.2 Experimental Setup

Table 3.1 contains the parameter values for GASVM-II+MOGASVM. These values are chosen based on the results of preliminary runs. In order of importance, three criteria are considered to evaluate the performances of the proposed approach; test accuracy, LOOCV accuracy, and the number of selected genes.

Table 3.1. Parameters of the proposed approach (GASVM-II+MOGASVM).

Data Set Parameters	Leukemia	Colon	Lung	MLL
No. populations	100	100	100	100
No. generations	1000	1000	1000	1000
Replacement rate (Roulette wheel selection)	0.8	0.8	0.8	0.8
Crossover rate (Two-point)	0.7	0.7	0.7	0.7
Mutation rate (Flip & gaussian)	0.01	0.01	0.01	0.01
w_1	0.8	0.7	0.7	0.7
w_2	0.2	0.3	0.3	0.3
Cost for generalization of SVMs	100	100	0.7	100

The experimental results presented in this section pursue two objectives. The first objective is to show that gene selection using GASVM-II+MOGASVM is needed for better classification of the gene expression data. The second objective is to show that GASVM-II+MOGASVM is better than GASVMs (single-objective and multi-objective) and GASVM-II. To achieve these objectives, several experiments are conducted on the proposed approach, 10 times on each data set. In the first stage, different numbers of preselected genes are chosen (10, 20, 30,..., 600). Furthermore, in the second stage, GASVM chooses a number of the final selected genes automatically. Lastly, it produces an optimized gene subset that contains the final selected genes. The subset that produces the highest LOOCV accuracy with the least possible number of selected genes is chosen as the best subset. SVM, GASVMs, and GASVM-II were also experimented for comparison with GASVM-II+MOGASVM.

Table 3.2. Classification accuracies for different gene subsets using GASVM-II+MOGASVM (10 runs on average).

No. preselected genes	Average for the leukemia data set			Average for the colon data set	
	Accuracy (%)		No. final selected genes	LOOCV accuracy (%)	No. final selected genes
	LOOCV	Test			
600	100 ± 0	77.94 ± 8.00	34.5 ± 3.92	98.23 ± 0.51	58.5 ± 4.81
400	100 ± 0	76.77 ± 10.78	11.9 ± 1.66	99.52 ± 0.78	26.9 ± 4.70
200	100 ± 0	76.47 ± 8.20	3.2 ± 0.79	99.52 ± 0.78	10.7 ± 2.00
100	100 ± 0	69.71 ± 11.52	3.1 ± 0.74	98.39 ± 3.48	9.9 ± 1.20
90	100 ± 0	77.35 ± 5.20	4.3 ± 1.16	99.03 ± 1.13	11.6 ± 2.01
80	100 ± 0	74.41 ± 9.51	4.0 ± 1.49	99.36 ± 0.83	11.8 ± 3.12
70	100 ± 0	69.71 ± 11.01	4.4 ± 1.35	99.52 ± 0.78	11.6 ± 2.17
60	100 ± 0	75.59 ± 7.08	19.4 ± 15.59	98.87 ± 1.09	9.8 ± 1.75
50	100 ± 0	73.24 ± 6.42	4.8 ± 1.14	96.94 ± 4.20	9.8 ± 1.48
40	100 ± 0	67.06 ± 10.26	4.8 ± 2.15	97.74 ± 3.42	10.7 ± 1.89
30	100 ± 0	71.77 ± 8.34	4.8 ± 1.69	96.94 ± 4.84	8.6 ± 1.84
20	100 ± 0	74.71 ± 12.10	4.5 ± 1.72	94.03 ± 7.37	7.0 ± 3.09
10	100 ± 0	85.88 ± 8.86	4.4 ± 1.35	92.74 ± 6.90	6.0 ± 3.23

Note: The results of the best subsets are shown in the shaded cells.

3.3.3 Result analysis and discussion

Table 3.2 shows that the highest averages of LOOCV and test accuracies for classifying leukemia cancer samples are 100% and 85.88%, respectively, while 99.52% LOOCV accuracy is obtained for the colon data set. Figure 3.2 shows that GASVM-II+MOGASVM has produced 100% LOOCV accuracy and 94.16% test accuracy for the lung data set, while Fig. 3.3 displays the highest averages of LOOCV accuracy and test accuracy for the MLL data set are 100% and 92%, respectively. In Table 3.2, Fig. 3.2, and Fig. 3.3, the values of the form $x \pm y$ represent an average value x with a standard deviation y.

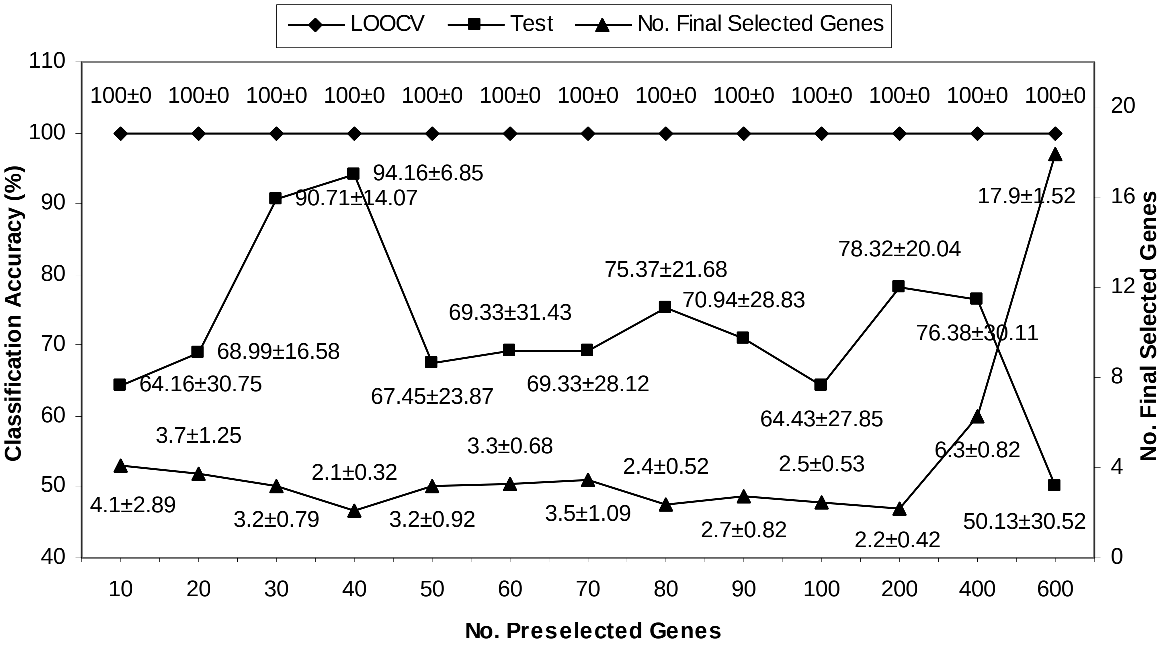


Fig. 3.2. A relation between the classification accuracies and the numbers of selected genes on the lung data set (10 runs on average).

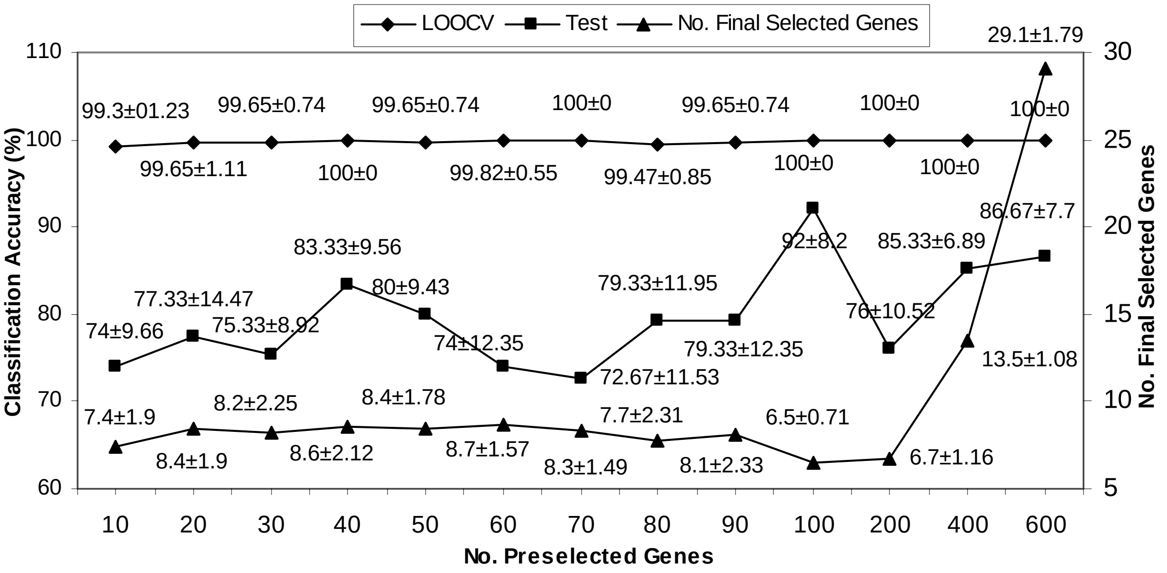


Fig. 3.3. A relation between the classification accuracies and the numbers of selected genes on the MLL data set (10 runs on average).

Table 3.3. The result of the best gene subsets in 10 runs.

Data set	No. preselected genes	LOOCV (%)	Test (%)	Experiment no.	No. final selected genes
Leukemia	10	100	97.06	2;4;5	2
Colon	70	100	-	1	9
Lung	40	100	98.66	2;6;7;8;9;10	2
MLL	100	100	100	1;2;6;9	6

Table 3.4. The benchmark of GASVM-II+MOGASVM with GASVMs and SVM on the leukemia and colon data sets.

Method	Leukemia data set (Average; the best)			Colon data set (Average; the best)	
	No. final selected genes	Accuracy (%)		No. final selected genes	LOOCV Accuracy (%)
		LOOCV	Test		
GASVM-II+MOGASVM	(3.4 ± 1.35; 2)	(100 ± 0; 100)	(85.88 ± 8.86; 97.06)	(11.6 ± 2.17; 9)	(99.52 ± 0.78; 100)
GASVM-II	(10 ± 0; 10)	(100 ± 0; 100)	(81.18 ± 10.21; 94.12)	(30 ± 0; 30)	(99.03 ± 0.83; 100)
MOGASVM	(2,212.6 ± 26.63; 2,252)	(95.53 ± 1.27; 97.37)	(84.41 ± 2.42; 88.24)	(446.3 ± 18.90; 446)	(93.23 ± 1.02; 95.16)
GASVM (single-objective)	(3,574.9 ± 40.05; 3,531)	(94.74 ± 0; 94.74)	(83.53 ± 2.48; 88.24)	(979.8 ± 35.80; 940)	(91.77 ± 0.51; 91.94)
SVM	(7,129 ± 0; 7,129)	(94.74 ± 0; 94.74)	(85.29 ± 0; 85.29)	(2,000 ± 0; 2,000)	(85.48 ± 0; 85.48)

Note: The best results are shown in the shaded cells.

Only 4.4 and 11 genes (the values of averages) were finally selected to obtain the highest averages of the accuracies of the leukemia and colon data sets, whereas 2.1 and 6.5 average genes were selected in the lung and MLL data sets. Almost all the different numbers of preselected genes and the final selected genes have obtained 100% LOOCV accuracy on the leukemia, lung, and MLL data sets. This result has proven that the proposed approach searches and selects the near-optimal solution (the best gene subset) in the solution space successfully. However, the LOOCV accuracies on the three data sets were much higher than the test accuracy due to over-fitting of these data sets. Over-fitting is a major problem in the classification of gene expression data when the LOOCV accuracy is much higher than the test accuracy. This problem happens because the number of training samples is smaller than the number of test samples, and many expression values of the test samples may be different from those of the training samples. Table 3.3 shows that the best performance of the best subsets of all the data sets.

Table 3.5. The benchmark of GASVM-II+MOGASVM with GASVMs and SVM on the lung and MLL data sets.

Method	Lung data set (Average; the best)			MLL data set (Average; the best)		
	No. final selected genes	Accuracy (%)		No. final selected genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
GASVM-II+MOGASVM	(2.1 ± 0.32; 2)	(100 ± 0; 100)	(94.16 ± 6.85; 98.66)	(6.5 ± 0.71; 6)	(100 ± 0; 100)	(92 ± 8.20; 100)
GASVM-II	(10 ± 0; 10)	(100 ± 0; 100)	(59.33 ± 29.32; 97.32)	(30 ± 0; 30)	(100 ± 0; 100)	(84.67 ± 6.33; 93.33)
MOGASVM	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)	(4,465.2 ± 18.34; 4,437)	(94.74 ± 0; 94.74)	(90 ± 3.51; 93.33)
GASVM (single-objective)	(6,267.8 ± 56.34; 6,342)	(75 ± 0; 75)	(84.77 ± 2.53; 87.92)	(6,298.8 ± 51.51; 6,224)	(94.74 ± 0; 94.74)	(87.33 ± 2.11; 86.67)
SVMs	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)	(12,582 ± 0; 12,582)	(92.98 ± 0; 92.98)	(86.67 ± 0; 86.67)

Note: The best results are shown in the shaded cells.

The benchmark of the proposed approach comparing GASVM-II, GASVMs (single-objective and multi-objective), and SVM is summarized in Table 3.4 and Table 3.5. The LOOCV accuracy, the test accuracy, and the number of selected genes are written in parentheses; the first and second parts are the averages and the best results, respectively. GASVM-II+MOGASVM outperformed GASVM-II, GASVMs, and SVM in terms of LOOCV accuracy, test accuracy, and the number of selected genes on average and the best results. Generally, GASVM-II was better than GASVMs and SVM on all the data sets. A small gene subset that is produced by GASVM-II+MOGASVM results in the high classification accuracy. This suggests that gene selection using the proposed approach is useful for cancer classification of gene expression data.

3.4 Summary

In this chapter, an approach using two hybrid methods (GASVM-II+MOGASVM) has been proposed, developed, and analyzed for gene selection and cancer classification. This research found that many combinations of gene subsets that did not contain equal numbers of genes produced different classification accuracy. This finding suggests that there are many irrelevant and noisy genes in gene expression data. In addition, the performance of GASVM-II+MOGASVM were superior to those of GASVM-II, GASVMs, and SVM. Focusing attention on a small subset of genes is useful not only because it produces good classification accuracy, but also because informative genes in this subset may provide insights into the mechanisms responsible for the cancer itself. The proposed approach can also be applied in other applications such as robotics, computer intrusion detections, and computer graphics. Even though the approach has classified tumors with high accuracy, it still cannot avoid the over-fitting problem. A cyclic hybrid method in Chapter 4 will study on how to find a good way to reduce the risk of the over-fitting problem and select a smaller subset of genes for cancer classification.

Chapter 4

A Cyclic Hybrid Method

4.1 Introduction

At the moment, several hybrid methods, especially combinations between GAs and SVMs (GASVM), have been implemented to select informative genes [13],[18],[20],[23],[25],[29]. The drawbacks of the hybrid methods (GASVM-based methods) in the previous works are intractable to efficiently produce a small (near-optimal) subset of informative genes when the total number of genes is too large (high-dimensional data); and the high risk of over-fitting problems. In order to overcome the limitations of the previous works and solve the problems derived from gene expression data, this chapter introduces and proposes a cyclic GASVM-based method (C-GASVM). This proposed method is optimal in the sense that it minimizes the number of selected genes and maximizes the classification accuracy.

4.2 Previous Works

Several hybrid methods, i.e., GASVM-based methods have been proposed for genes selection of gene expression data [13],[18],[20],[23],[25],[29]. Generally, the previous GASVM-based methods performed well in high-dimensional data, e.g., gene expression data since a modified chromosome representation and a multi-objective approach has been proposed [23],[25]. However, the methods yielded inconsistent results when they were run independently.

Li *et al*. [18] have proposed a GASVM-based method for the same purpose. Next, the work of Huang and Chang [13] can simultaneously optimize genes and SVM parameter settings by using a GASVM-based method. An improved GASVM-based method has been recently introduced in Li *et al*., [20] to produce a small subset of genes. Peng *et al*. [29] have introduced a recursive feature elimination post-processing step after the step of a GASVM-based method in order to reduce the number of selected genes again.

Nevertheless, the GASVM-based methods of the previous works are still intractable to produce a near-optimal subset of genes from high-dimensional data due to their binary chromosome representation drawback [13],[18],[20],[23],[25],[29]. The total number of gene subsets produced by GASVM-based methods is calculated by $M_c = 2^M - 1$, where M_c is the total number of gene subsets, and M is the total number of genes. Based on this equation, the GASVM-based methods are almost impossible to evaluate all possible subsets of selected genes if M is too many (high-dimensional data). Although the works of Peng *et al*. [25] and Li *et al*. [20] have implemented a preprocessing step to decrease the dimensionality of data, but it can only reduce a small number of genes, and many genes are still available in the data. The GASVM-based methods also face with the high risk of over-fitting problems. The over-fitting problem that occurred on hybrid methods (e.g., GASVM-based methods) was also reported in a review paper written by Saeys *et al*. [30].

4.3 The Proposed Cyclic Hybrid Method

This chapter proposes C-GASVM for gene selection from gene expression data. C-GASVM is a GASVM-based method. C-GASVM in the present work differs from the GASVM-based methods in the previous works [13],[18],[20],[23],[25],[29] in one major part. The major difference is that the proposed method involves a cyclic approach, whereas the previous works did not use any cyclic approach for gene selection. The flowchart of C-GASVM is shown in Fig. 4.1. The algorithm of C-GASVM is shown in Fig. 4.2. Basically, C-GASVM repeats the process of GASVM-II to produce potential subsets and reduce the dimensionality of data repeatedly.

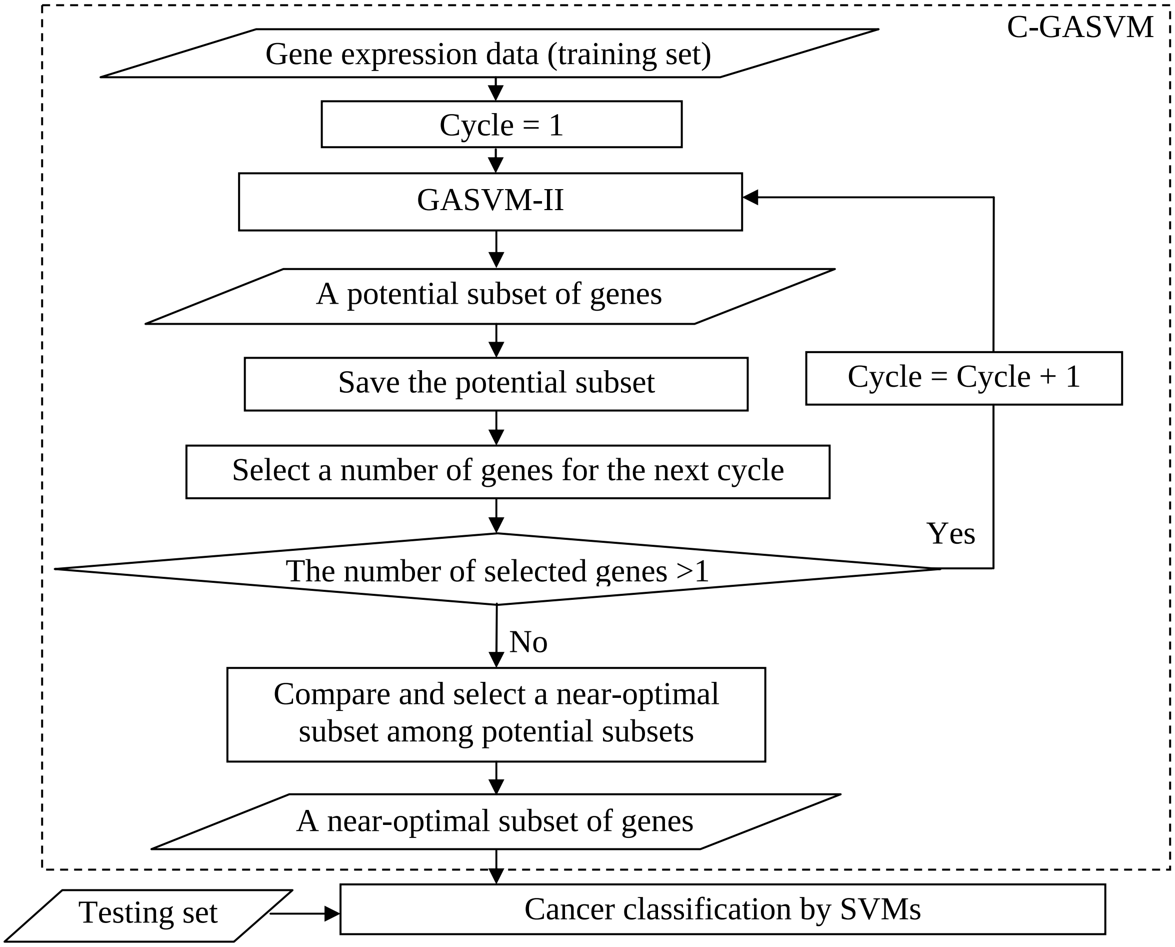

C-GASVM
Gene expression data (training set)
Cycle = 1
GASVM-II
A potential subset of genes
Save the potential subset
Cycle = Cycle + 1
Select a number of genes for the next cycle
Yes
The number of selected genes >1
No
Compare and select a near-optimal subset among potential subsets
A near-optimal subset of genes
Testing set
Cancer classification by SVMs

Fig.4.1. The flowchart of C-GASVM.

```
VARIABLE:
  $c$ : the cth cycle.    $n_s$ : the number of selected genes.    $x_a$ : ath chromosome.
  $x_a.fitness$ : the fitness of ath chromosome.    $x_a.\#gene$ : the number of genes in ath chromosome.
  $S_c$ : a potential subset of genes of cycle c.    $S_c.fitness$ : the fitness value of $S_c$.
  $S_c.\#gene$ : the number of genes in $S_c$.    $gen$ : generation.    $N$ : the total number of samples
  $M$ : the total number of genes.    $div\_gene$ : the divider for the number of selected genes.
INPUT:
  $G_{N\times(M+1)}$ : gene expression data (training set).    $pop\_num$ : the number of population.
  $gen\_num$ : the number of generation.    $cross\_rate$ : the rate of crossover operator.    $mut\_rate$ : the rate of mutation operator.
OUTPUT:
  $S_{opt}$ : a near-optimal subset of genes.    $S_{opt}.fitness$ : the fitness value of $S_{opt}$.    $S_{opt}.\#gene$ : the number of genes in $S_{opt}$.
Begin
  $gen := 0;$    $c := 1;$    $n_s := M / div\_gene;$
  $S_c := G_{N\times(M+1)};$    $S_c.fitness := 0;$    $S_c.\#gene := M;$
  $S_{opt} := 0;$    $S_{opt}.fitness := 0;$    $S_{opt}.\#gene := 0;$
  while ($S_c.\#gene > 1$) do                  // Step 1: Starting a cyclic process
        for ($a = 1; a \le pop\_num; a++$)
              $x_a := initialise(int, n_s, S_c);$
        end_for
        while ($gen < gen\_num$) do           // Step 2: Starting GASVM-II to produce a potential gene subset
            for ($a = 1; a \le pop\_num; a++$)
              $SVM(x_a);$
              $x_a.fitness := w_1 \times A(x_a) + (w_2 \times (M - R(x_a)) / M);$
            end_for
            $selection\_method(roulette\_wheel, gen);$
            $crossover(two\_point, cross\_rate);$
            $mutation(gaussian, mut\_rate);$
            $gen := gen + 1;$
        end_while                              // Step 3: Ending GASVM-II
        return ($S_c$);                 // Step 4: Producing and saving the potential subset for the cycle c
        if ($S_c\#gene > 100$) then     // Step 5: Selecting  a number of genes for the next cycle (cycle c+1)
              $n_s := S_c\#gene / div\_gene;$
              if ($n_s < 100$) then
                 $n_s := 100;$
              end_if
        end_if
        else if ($10 < S_c\#gene \le 100$) then
              $n_s := S_c\#gene - 10;$
        end_else_if
        else if ($1 < S_c\#gene \le 10$) then
              $n_s := S_c\#gene - 1;$
        end_else_if                     // Step 6: Ending the selection process
       $c := c + 1;$   $gen := 0;$
  end_while                             // Step 7: Ending the cyclic process
  for ($i = 0; i < c; i++$)             // Step 8: Compare and select an optimal subset among potential subsets
      if ($S_i.fitness > S_{opt}.fitness$) then
        $S_{opt} := S_i;$    $S_{opt}.fitness = S_i.fitness;$    $S_{opt}\#gene = S_i.\#gene;$
      end_if
  end_for
  return ($S_{opt}$);                   // Step 9: Producing a near-optimal subset of selected genes
End
```

Fig.4.2. The algorithm of C-GASVM.

4.3.1 Chromosome representation for C-GASVM

The present work uses integer chromosome representation in C-GASVM in order to overcome the limitation of the binary chromosome representation in previous related works [13],[18],[20],[25],[29]. The present work modifies the mechanism of gene selection of C-GASVM based on the representation to efficiently select gene subsets from high-dimensional data. The modification idea is based on Eq.(4.1) to reduce the number of gene subsets by fixing the number of selected genes. The fixing process is automatically done by a cyclic process in C-GASVM for each cycle.

$$y = {}_{M}C_{x} = \frac{M!}{x!(M-x)!} \tag{4.1}$$

where ${}_{M}C_{x}$ is the total number of subsets of selected genes x from the total number of genes M.

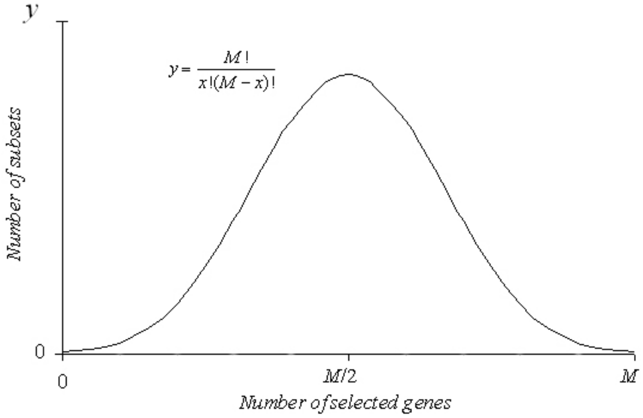


Fig.4.3. A relation between the number of subsets *y* and the number of selected genes *x* from the total number of genes *M*.

Figure 4.3 shows a graph based on Eq.(4.1). A maximum number of subsets are reached when the number of selected genes is chosen at *M*/2. Hence, the selection number at *M*/2 or about *M*/2 should be avoided. If the selection uses the number, C-GASVM is impossible to evaluate all subsets due to the huge number of subsets. Conversely, all subsets of genes are possible evaluated if a small or large number of the selected genes are chosen. In this research, C-GASVM only chooses the large number of selected genes in each cycle in order to avoid an over-fitting problem. If the selection chooses the small number, C-GASVM faces with the problem. This is reported and proved by the subsection of experimental results in this chapter.

g^1	g^2	$\dots$	$g^{n_s - 1}$	g^{n_s}

Note:
n_s = a number of selected genes from an input set (S_{c-1}), $1 \le n_s \le M$.
$M =$ the total number of genes in an input set (S_{c-1}).
g^j = an integer value in a chromosome, $1 \le g^j \le M$.
$j =$ the *j*th gene in a chromosome, $1 \le j \le n_s$.

Fig.4.4. Integer chromosome representation in C-GASVM.

Therefore, in C-GASVM, the chromosome representation is modified as shown in Fig. 4.4 which has integer representation. It includes values of integers g^j that indicate which genes are needed to be selected among the total genes in a data set. For example, if $g^j = 10$, then C-GASVM selects the 10th gene from the data set, and groups it into a subset of genes. The number of selected genes is represented by n_s. The number of g^j in a chromosome is equal to n_s. The binary chromosome representation of GASVM-based methods in the related previous works [13],[18],[20],[25],[29] is encoded with all genes and its size depends entirely on the total number of genes, M. In contrast, the integer chromosome representation in C-GASVM is only encoded with a number of selected genes that is automatically fixed by the cyclic process. Hence, the total number of genes, M does not really affect the size (length) of the chromosome so as to keep its size relatively small. Its size can vary according to M and n_s. The size of chromosomes and the number of selected genes are also the same for a similar cycle, but they are different for dissimilar cycles. Finally, a chromosome (a gene subset) is represented as $x = (g^1, g^2, ..., g^{n_{s-1}}, g^{n_s})$. For example, the *a*th chromosome is represented by $x_a = (g_a^1, g_a^2, ..., g_a^{n_{s-1}}, g_a^{n_s})$.

4.3.2 A fitness function for C-GASVM

A fitness value of individuals (gene subsets) is calculated as follows:

$$fitness(x) = w_1 \times A(x) + (w_2 \times (M - R(x)) / M) \tag{4.2}$$

where $A(x) \in [0,1]$ is the LOOCV accuracy on the training set using the only expression values of the selected genes in a gene subset, x. This accuracy is provided by SVM. $R(x)$ is the number of selected genes in x. M is the total number of genes for each sample in the training set. w_1 and w_2 are two priority weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$.

4.4 Experiments

4.4.1 Data sets

Five real gene expression data sets that contain binary classes and multi-classes are used to evaluate the performance of C-GASVM; leukemia, colon, lung, and mixed-lineage leukemia (MLL), and small round blue cell tumors (SRBCT) data sets. The summary of the first four data sets has been shown on Table 2.1 in Chapter 2. The SRBCT data set is a multi-classes data set. It has four classes; ewing family of tumors (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and burkitt lymphomas (BL). The training set contains 63 samples (22 EWS, 20 RMS, 12 NB, and 8 BL), whereas the test set contains 20 samples (6 EWS, 5 RMS, 6 NB, and 3 BL). There are 2,308 genes in each sample. It can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/.

4.4.2 Experimental setup

Since the number of training samples in gene expression data is small, the accuracy on the training set is calculated through the LOOCV procedure. For the test accuracy, SVM is built using all the training samples, and the classes of test samples from the test set are predicted one by one using the SVM. The test accuracy is estimated by the number of the correctly classified test samples, divided by the number of samples in the test set.

Table 4.1 contains parameter values for C-GASVM. These values are chosen based on the results of preliminary runs. Three criteria following their importance are considered to evaluate the performances of C-GASVM and other experimental methods; test accuracy, LOOCV accuracy, and the number of selected genes. Higher accuracy and a smaller number of selected genes are needed to obtain an excellent performance.

Table 4.1. Parameter settings for C-GASVM.

Data set Parameters	Leukemia	Colon	SRBCT	Lung	MLL
No. populations	50	50	50	50	50
No. generations	100	100	100	100	100
Replacement rate (Roulette wheel selection)	0.8	0.8	0.8	0.8	0.8
Crossover rate (Two-point)	0.7	0.7	0.7	0.7	0.7
Mutation rate (Gaussian)	0.01	0.01	0.01	0.01	0.01
w_1	0.8	0.7	0.8	0.7	0.7
w_2	0.2	0.3	0.2	0.3	0.3
div_gene	1.33	1.25	1.33	1.33	1.33
Cost for generalization of SVMs	100	100	100	0.7	100

Experimental results presented in this chapter pursue four objectives. The first objective is to show that a gene selection using C-GASVM is needed to produce a small (near-optimal) subset of informative genes for better classification accuracy. The second objective is to display a scatter gene graph and a list of informative genes in the best subsets produced by C-GASVM for biological usage. The third objective is to show that C-GASVM is better than other experimental methods such as GASVM (single-objective), MOGASVM, GASVM version 2 (GASVM-II), and SVM. The last objective is to compare C-GASVM with other previous works that only used GASVM-based methods. To achieve the four objectives, several experiments are conducted 10 times on each data set using C-GASVM and other experimental methods. Next, an average result of the 10 independent runs is obtained. A near-optimal subset that produces the highest classification accuracies with the possible least number of genes is selected as the best subset.

4.4.3 LOOCV and test accuracies of selected genes with C-GASVM

Table 4.2 shows the classification accuracy for each run using C-GASVM on all data sets. Interestingly, almost all runs have achieved 100% LOOCV accuracy on all data sets. This has proven that C-GASVM has efficiently selected and produced the near-optimal solution in a solution space. This is due to the fact of its ability to automatically reduce the dimensionality and complexity of the solution space on a cycle by cycle basis. C-GASVM also removes irrelevant and noisy genes in order to yield the high accuracy. The small gene subsets that are produced by the proposed C-GASVM result in the high classification accuracy.

Generally, near-optimal subsets that obtained from almost all run on the data sets contain less than 10 genes. This is inline with the diagnostic goal of developed medical procedures that needs the least number of possible informative genes to detect diseases. The conservativeness of the results in Table 4.2 is controlled and maintained by the cyclic approach and the fitness function of C-GASVM that maximizes the classification accuracy and meanwhile, minimizes the number of selected genes.

Table 4.2. Classification accuracies for each run using C-GASVM.

Data set	Run# Evaluation	1	2	3	4	5	6	7	8	9	10	Average ± S.D.
Leukemia	LOOCV (%)	100	100	100	100	100	100	100	100	100	100	100 ± 0
	Test (%)	88.24	88.24	88.24	88.24	88.24	91.18	91.18	94.12	82.35	88.24	88.82 ± 3.04
	#Genes	5	2	2	2	2	2	3	2	7	2	2.9 ± 1.73
SRBCT	LOOCV (%)	100	100	100	100	100	100	100	100	100	100	100 ± 0
	Test (%)	90	85	80	85	80	85	85	85	85	85	84.5 ± 2.84
	#Genes	20	7	6	7	7	7	7	7	7	8	8.3 ± 4.14
Colon	LOOCV (%)	100	98.39	100	100	100	98.39	98.39	96.77	100	98.39	99.03 ± 1.13
	#Genes	30	20	30	20	10	20	30	40	20	20	24 ± 8.43
Lung	LOOCV (%)	100	100	100	100	100	100	100	100	100	100	100 ± 0
	Test (%)	94.63	93.96	94.63	90.60	93.96	98.66	94.63	94.63	90.60	90.60	93.69 ± 2.52
	#Genes	2	5	2	2	5	4	2	2	2	2	2.80 ± 1.32
MLL	LOOCV (%)	100	100	100	100	100	100	100	100	100	100	100 ± 0
	Test (%)	100	93.33	93.33	86.67	80.00	86.67	93.33	93.33	93.33	93.33	91.33 ± 5.49
	#Genes	20	8	20	20	10	10	8	8	8	8	12.0 ± 5.58

Note: The results of the best subsets of each data set are shown in the shaded cells. S.D. denotes the standard deviation, whereas Run# and #Genes represent a run number and a number of selected genes, respectively. The colon data set only has LOOCV accuracy since it only has the training set.

Practically, the best subset of a data set is firstly chosen and the genes in it are then listed for biological usage. These informative genes among the thousand of genes may be the excellent candidates for clinical and medical investigations. Biologists can save much time since they can directly refer to the genes that have high possibility to be useful for cancer diagnosis and drug target in the future. The best subset is chosen based on the highest classification accuracy with the smallest number of selected genes. The highest accuracy gives confidence to us for the most accurate classification of cancer types. Moreover, the smallest number of selected genes for cancer classification can reduce the cost in clinical settings.

4.4.4 A list of informative genes for biological usage

Informative genes in the best gene subsets as produced by the proposed C-GASVM and reported in Table 4.2 are listed in Table 4.3. A gene accession number or probe-set name is used for searching the biological information of genes in the public database of genes. Some of these genes are already identified to be highly possible clinical markers for cancer diagnosis by biological research. For example, the patent of the United States entitled "Methods and Compositions for the Identification, Assessment, and Therapy of Human Cancers" (Patent number: 7338758; Publication date: March 4, 2008) has found the gene D50930 as highly expressed and sensitive genes. Furthermore, the gene Y00638 was identified by another patent in the United States (Patent number: 7011947; Publication date: March 14, 2006) as an over-expressed gene in MLL compared to ALL. Some of the remaining genes may be the excellent candidates for further clinical investigation.

Since only two genes were found in the best subset of the leukemia data set, it became possible to visualize the gene expression profiles with respect to the distinct leukemia subtypes (ALL and AML). Only samples in the training set of the best subset were picked to generate a graph because they have achieved 100% LOOCV accuracy with the smallest number of selected genes. Figure 4.5 shows the scatter graph of combinations of two-genes (Gene X59417 versus Gene X95735). In this graph, the clusters of ALL and AML are clear and the boundary can be easily drawn. Finally, the following simple prediction rules are obtained, which helps biologists to make an accurate diagnosis of the two subtypes of leukemia samples:

- A sample has ALL if and only if the expression level of the Gene X95735 is less than -0.53.
- A sample has AML if and only if the expression level of the Gene X95735 is higher than -0.50.

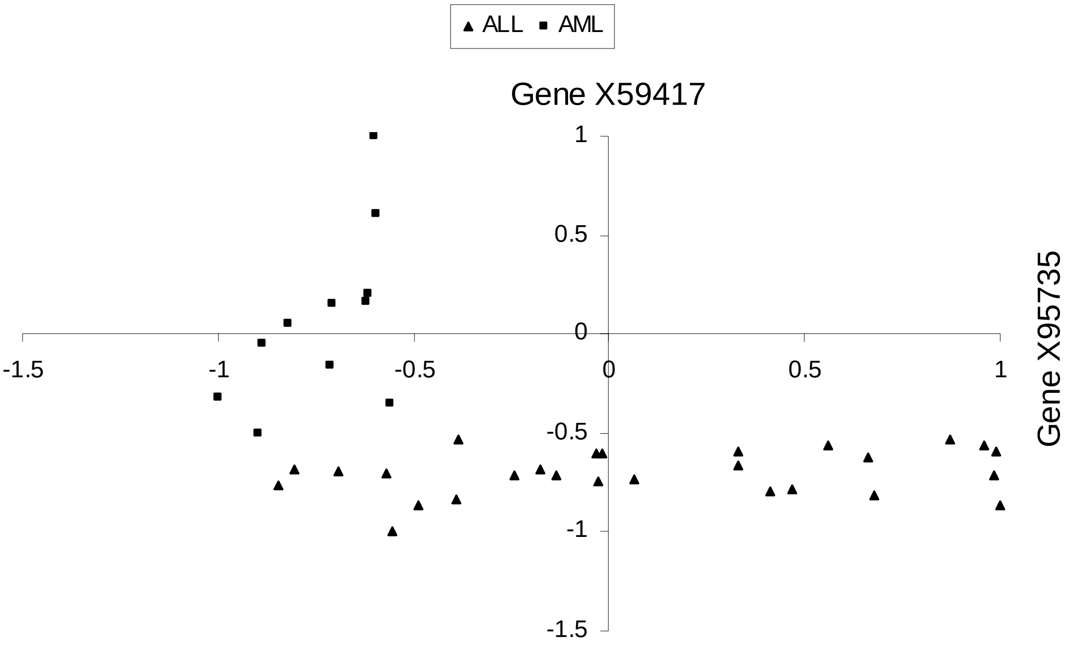


Fig.4.5. The scatter graph of the two-genes in the best subset of the leukemia data set. The vertical axis represents the value of expression levels of Gene the X95735, whereas the horizontal denotes the value of expression levels of the Gene X59417.

Table 4.3. The list of informative genes in the best gene subsets.

Data set	Run no.	Probe-set name / Image ID	Gene accession number / Gene card identifier
Leukemia	8	36122_at	X59417
		36958_at	X95735
SRBCT	1	884867	GC14P102870
		868304	GC10M090684
		1323448	GC14P105024
		450152	GC09P000263
		298963	GC14P104957
		725188	GC02P063727
		823696	GC10P091142
		295985	GC07M092072
		139705	GC13M072181
		244652	GC09P130485
Colon	5	NA	T95318
		NA	M64445
		NA	H69872
		1291_s_at	L03840
		NA	J03210
		32135_at	U00968
		NA	T96873
		38119_at	X12496
		NA	R81170
		NA	K03474
Lung	6	32551_at	U03877
		33634_at	AF038007
		35708_at	W27414
		36938_at	U70063
MLL	1	33636_at	U87459
		34129_at	AB023223
		34583_at	U02687
		34441_at	AF052090
		37832_at	AL080062
		38601_at	AF073500
		39212_at	AF038179
		35659_at	U00672
		40143_at	D50930
		40520_g_at	Y00638
		41750_at	D49489
		33863_at	U65785
		35804_at	AB022785
		41594_at	M64174
		326_i_at	HG1800

4.4.5 C-GASVM versus other experimental methods

The benchmark of C-GASVM in comparison with other experimental methods that have been experimented in this work is summarized in Table 4.4. GASVM (single-objective) is developed to implement a single-objective approach in its fitness function, while MOGASVM is developed for multi-objective approach. Binary chromosome representation has been used in these hybrid methods. GASVM-II and C-GASVM are almost the same in terms of chromosome representations, algorithms, etc. The difference is that GASVM-II not implement the cyclic process in its mechanism. It is developed to prove that an over-fitting problem is happen when the selection using a small number of selected genes, and compare its experimental results with C-GASVM.

Overall, the LOOCV and test accuracies of C-GASVM for all the data sets were higher than MOGASVM, GASVM (single-objective), GASVM-II, and SVM. Moreover, the number of selected genes by using C-GASVM was also lower. Based on the standard deviations of LOOCV accuracy, test accuracy, and the number of selected genes, C-GASVM was also more consistent than the other experimental methods except for SVM. SVM achieved 0 for the standard deviations in all experiments since it did not implement any gene selection approach. The gap between LOOCV accuracy and test accuracy that resulted by C-GASVM was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. On the other hand, the results of LOOCV accuracy of the others were much higher than their test accuracy because they were unable to avoid or reduce the risk of over-fitting problems. Over-fitting is a major problem of hybrid methods in classification of gene expression data when the classification accuracy on training samples, e.g., LOOCV accuracy is much higher than the test accuracy. This problem occurred in gene expression data because the number of genes greatly exceeds the number of samples, and many patterns of the test samples may be different from those of the training samples.

Table 4.4. The benchmark of C-GASVM with other experimental methods.

Data set	Experiment Evaluation	C-GASVM	GASVM-II [23]	MOGASVM [25]	GASVM [23]	SVM [25]
Leukemia (Average ± S.D; The best)	LOOCV (%)	(100 ± 0; 100)	(100 ± 0; 100)	(95.53 ± 1.27; 97.37)	(94.74 ± 0; 94.74)	(94.74 ± 0; 94.74)
	Test (%)	(88.82 ± 3.04; 94.12)	(81.18 ± 10.21; 94.12)	(84.41 ± 2.42; 88.24)	(83.53 ± 2.48; 88.24)	(85.29 ± 0; 85.29)
	No. selected genes	(2.9 ± 1.73; 2)	(10 ± 0; 10)	(2,212.6 ± 26.63; 2,252)	(3,574.9 ± 40.05; 3,531)	(7,129 ± 0; 7,129)
Colon (Average ± S.D; The best)	LOOCV (%)	(99.03 ± 1.13; 100)	(99.03 ± 0.83; 100)	(93.23 ± 1.02; 95.16)	(91.77 ± 0.51; 91.94)	(85.48 ± 0; 85.48)
	No. selected genes	(24 ± 8.43; 10)	(30 ± 0; 30)	(446.3 ± 18.90; 446)	(979.8 ± 35.80; 940)	(2,000 ± 0; 2,000)
SRBCT (Average ± S.D; The best)	LOOCV (%)	(100 ± 0; 100)	(99.84 ± 0.50; 100)	(100 ± 0; 100)	(98.41 ± 0; 98.41)	(98.41 ± 0; 98.41)
	Test (%)	(84.5 ± 2.84; 90)	(68 ± 9.49; 85)	(81.5 ± 7.47; 85)	(78.5 ± 3.38; 85)	(80 ± 0; 80)
	No. selected genes	(8.3 ± 4.14; 20)	(10 ± 0; 10)	(444.7 ± 19.09; 429)	(1146 ± 10.33; 1134)	(2,308 ± 0; 2,308)
Lung (Average ± S.D; The best)	LOOCV (%)	(100 ± 0; 100)	(100 ± 0; 100)	(75.31 ± 0.99; 78.13)	(75 ± 0; 75)	(65.63 ± 0; 65.63)
	Test (%)	(93.69 ± 2.52; 98.66)	(59.33 ± 29.32; 97.32)	(85.84 ± 3.97; 93.29)	(84.77 ± 2.53; 87.92)	(85.91 ± 0; 85.91)
	No. selected genes	(2.80 ± 1.32; 4)	(10 ± 0; 10)	(4,418.5 ± 50.19; 4,433)	(6,267.8 ± 56.34; 6,342)	(12,533 ± 0; 12,533)
MLL (Average ± S.D; The best)	LOOCV (%)	(100 ± 0; 100)	(100 ± 0; 100)	(94.74 ± 0; 94.74)	(94.74 ± 0; 94.74)	(92.98 ± 0; 92.98)
	Test (%)	(91.33 ± 5.49; 100)	(84.67 ± 6.33; 93.33)	(90 ± 3.51; 93.33)	(87.33 ± 2.11; 86.67)	(86.67 ± 0; 86.67)
	No. selected genes	(12.0 ± 5.58; 20)	(30 ± 0; 30)	(4,465.2 ± 18.34; 4,437)	(6,298.8 ± 51.51; 6,224)	(12,582 ± 0; 12,582)

Note: The best results of each data set are shown in the shaded cells. S.D. denotes the standard deviation. The colon data set only has LOOCV accuracy since it only has the training set.

GASVM (single-objective) and MOGASVM cannot produce a near-optimal subset of informative genes because they perform poorly in high-dimensional data due to their chromosome representation drawback. The LOOCV accuracy of GASVM-II is much higher than its test accuracy. These findings prove that GASVM-II causes the over-fitting problem even if it uses a small numbers of selected genes. This problem happens since the small selections not involve many relations among genes. This method would also be difficult for the usage because it needs to manually select the number of genes.

On the contrary, C-GASVM selects a large number of genes automatically in each cycle of the cyclic process to finally produce a small (near-optimal) subset of informative genes. The gap between LOOCV accuracy and test accuracy was also lower. Therefore, C-GASVM is more efficient than other experimental methods since it has produced the higher classification accuracies, smaller numbers of selected genes, smaller standard deviations, and smaller gap between LOOCV accuracy and test accuracy. However, due to the cyclic process, C-GASVM is computationally more extensive than other methods.

4.4.6 C-GASVM versus previous related works

For an objective comparison, the present work only compares C-GASVM with related previous works that used GASVM-based methods in their works [13],[20],[23],[25],[29]. The previous works also produced the average of classification accuracy results since they used hybrid approaches. The present work makes the comparison using the averages of LOOCV accuracy and the number of selected genes. This is due to the most previous works only evaluated the performance of their approaches using the LOOCV procedure or *k*-fold cross-validation and the number of selected genes on averages. At the moment, they used high-dimensional data such as the leukemia, SRBCT, and colon data sets for experimental usage. Additionally, the present work has used very high-dimensional data (more than 12,000 genes) such as the lung and MLL data sets to test the effectiveness of C-GASVM. The experimental result of the very high-dimensional data is only shown in Tables 4.2, 4.3, and 4.4.

Table 4.5 displays the benchmark of this work and previous related works. For the leukemia data set, the averages of LOOCV accuracy and the number of selected genes of the present work were 100% and 2.9 genes, respectively. The latest previous works such as Huang and Chang [13], Li *et al*. [28], and Peng *et al*. [29] also came up with the similar LOOCV result to the present results, but they used slightly more genes to obtain the same result. The LOOCV accuracy and test accuracy that produced in Mohamad *et al*. [23] and Mohamad *et al*. [25] were also less than the present work.

In the present work, there was increase in the average of LOOCV accuracy (100%) for the SRBCT data set as compared to Huang and Chang [13]. However, the average of the number of selected genes (8.3 genes) was slightly higher than the previous work. The work of Huang and Chang [13] only achieved 98.75% LOOCV accuracy on average using 6.2 average genes. For the colon data set, it was noted that the best result (99.03% LOOCV accuracy on average) of this work was higher than the best result from the latest previous work [28]. This work needed 24 genes on average to achieve the best result, whereas only 15 average genes have been used in the work of Li *et al*. [28]. The work of Peng *et al*. [29] achieved 93.55% LOOCV accuracy on average, but they used 12 average genes. The LOOCV accuracy and test accuracy genes of the SRBCT and colon data sets that produced in Mohamad *et al*. [23] and Mohamad *et al*. [23] were also less than the present work. Overall, this work has outperformed the related previous works on the data sets in terms of LOOCV accuracy and the number of selected genes. The previous works are intractable to efficiently produce a near-optimal subset of genes in high-dimensional data due to their binary chromosome representation drawback [13],[28],[25],[29]. Although the works of Li *et al*. [28] and Peng *et al*. [29] have implemented a preprocessing step to decrease the dimensionality of data, but they can only reduce a small number of genes, and many genes are still available in the data.

Table 4.5. The comparison between C-GASVM and other previous GASVM-based methods.

Data set	Experiment Evaluation	The present work	[13]	[28]	[23]	[25]	[29]
Leukemia (Average ± S.D; The best)	LOOCV (%)	(100 ± 0; 100)	(100 using 10-CV ± NA; NA)	(100 ± NA; NA)	(100 ± 0; 100)	(95.53 ± 1.27; 97.37)	(100 ± NA; NA)
	Test (%)	(88.82 ± 3.04; 94.12)	NA	NA	(81.18 ± 10.21; 94.12)	(84.41 ± 2.42; 88.24)	NA
	No. selected genes	(2.9 ± 1.73; 2)	(3.4 ± NA; NA)	(4 ± NA; NA)	(10 ± 0; 10)	(2,212.6 ± 26.63; 2,252)	(6 ± NA; NA)
Colon (Average ± S.D; The best)	LOOCV (%)	(99.03 ± 1.13; 100)	NA	(93.55 ± NA; NA)	(99.03 ± 0.83; 100)	(93.23 ± 1.02; 95.16)	(93.55 ± NA; NA)
	No. selected genes	(24 ± 8.43; 10)	NA	15 ± NA; NA	(30 ± 0; 30)	(446.3 ± 18.90; 446)	(12 ± NA; NA)
SRBCT (Average ± S.D; The best)	LOOCV (%)	(100 ± 0; 100)	(98.75 using 10-CV ± NA; NA)	NA	(99.84 ± 0.50; 100)	NA	NA
	Test (%)	(84.5 ± 2.84; 90)	NA	NA	(68 ± 9.49; 85)	NA	NA
	No. selected genes	(8.3 ± 4.14; 20)	(6.2 ± NA; NA)	NA	(10 ± 0; 10)	NA	NA

Note: The best results of each data set are shown in the shaded cells. ‘NA‘ means that the result is not reported in the related previous works. S.D. denotes the standard deviation, whereas 10-CV means 10-fold-cross-validation. The colon data set only has LOOCV accuracy since it only has the training set.

4.5 Summary

In this chapter, a cyclic GASVM-based method (C-GASVM) has been proposed and tested for gene selection on five real gene expression data that contain binary classes and multi-classes of tumors samples. Based on the experimental results, the performance of C-GASVM was superior to the other experimental methods and previous related works. This is due to the fact that C-GASVM can automatically reduce the dimensionality of the data on a cycle by cycle basis. When the dimensionality was reduced, the combination of genes and the complexity of solution spaces can also be automatically decreased repeatedly. This cyclic process is done to produce potential gene subsets from high-dimensional data (gene expression data), and finally generate a near-optimal subset of informative genes. Hence, the gene selection using C-GASVM is needed to produce a small subset of informative genes for high cancer classification. Moreover, focusing the attention on the informative genes in the best subset may provide insights into the mechanisms responsible for the cancer itself. Even though C-GASVM has reduced the risk of the over-fitting problem, it is still not able to completely avoid the over-fitting problem. Thus, Chapter 5 will propose I-MOGASVM to reduce the risk of the over-fitting problem again and select a smaller subset of genes for cancer classification

Chapter 5

An Iterative Approach

5.1 Introduction

The drawbacks of the hybrid methods (GASVM-based methods) in previous work [13],[20],[23],[25],[29] are an inability to efficiently produce a near-optimal subset of informative genes when the total number of genes is too large (high-dimensional data) due to the drawback of binary chromosome representation; and the high risk of over-fitting problems. The over-fitting problem that occurred in hybrid methods (e.g., GASVM-based methods) was also reported in a review paper by Saeys *et al*. [30]. In order to overcome the limitations of the previous work and solve the problems derived from gene expression data, the present work proposes an iterative approach based on MOGASVM.

5.2 The Proposed Iterative Approach Based on GASVM

This chapter proposes an interactive approach based on MOGASVM (I-MOGASVM) for gene selection. Details of MOGASVM can be found in Mohamad *et al*. [25]. I-MOGASVM in the present work differs from the methods in previous work in one major way [13],[20],[23],[25],[29]. This difference is that the present work proposed method involves an iterative approach, whereas the previous work did not use any iterative process for gene selection. The general procedure of I-MOGASVM is shown in Fig. 5.1.

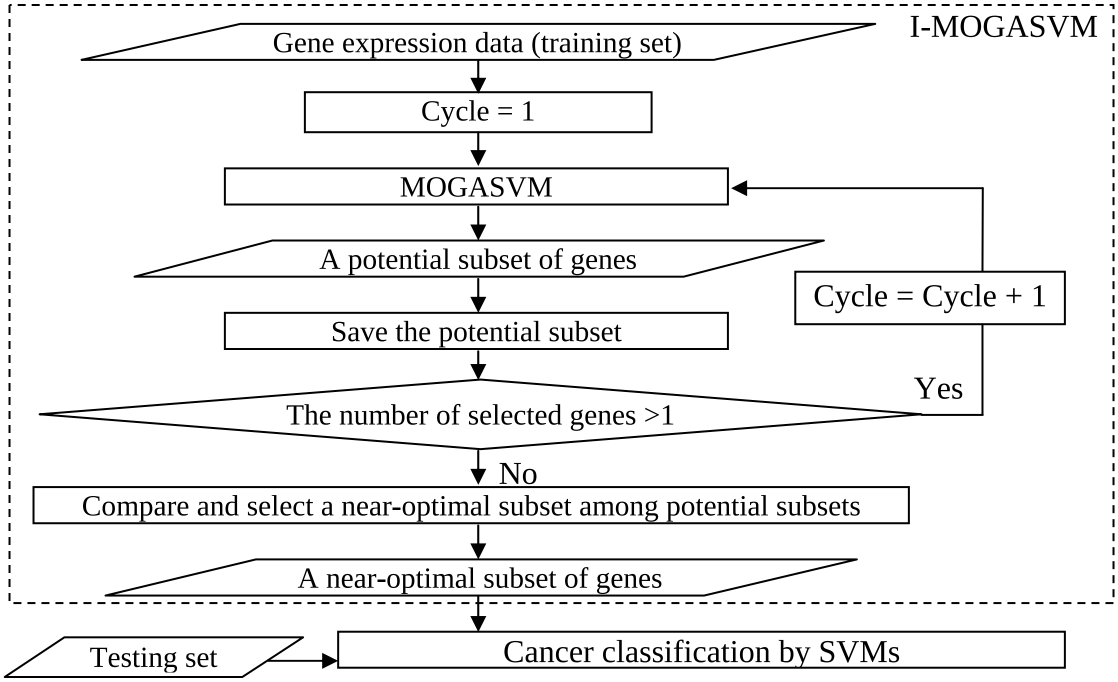

I-MOGASVM
Gene expression data (training set)
Cycle = 1
MOGASVM
A potential subset of genes
Save the potential subset
Cycle = Cycle + 1
Yes
The number of selected genes >1
No
Compare and select a near-optimal subset among potential subsets
A near-optimal subset of genes
Testing set
Cancer classification by SVMs

Fig.5.1. The general flowchart of I-MOGASVM.

Basically, I-MOGASVM repeats the process of MOGASVM to reduce the dimensionality of data iteratively. A description of each step is given below.

Step 1: Starting the iterative process. This is repeated until the number of selected genes in the potential subset of the current cycle *c* is equal to or less than 1. The number of cycles is based on the satisfied condition of genes numbers. In each cycle of I-MOGASVM, a number of selected genes are automatically selected by MOGASVM and the dimensionality is iteratively reduced.

Step 2: Starting MOGASVM to find and produce a potential subset of genes.

Step 3: Producing and saving the potential subset of selected genes. This potential subset is used for the next cycle (cycle *c*+1) as an input set. The selection of genes in the next cycle (cycle *c*+1) only uses genes in the potential subset that are the result of the previous cycle (cycle *c*). Therefore, the dimensionality and complexity of solution spaces can be decreased on a cycle-by-cycle-basis.

Step 4: A near-optimal subset is selected among the potential subsets based on the highest fitness value (the highest LOOCV accuracy with the smallest number of selected genes).

Step 5: An iterative process (Steps 1-4) results a near-optimal subset of genes. This subset can be found due to the dimensionality of data has been iteratively reduced. The near-optimal subset is then used to construct SVM, and the constructed SVM are tested by using the test set.

5.3 Experiments

5.3.1 Data sets

Four real gene expression data sets are used to evaluate I-MOGASVM; leukemia cancer, colon cancer, lung cancer, and mixed-lineage leukemia (MLL) cancer data sets. Table 2.1 in Chapter 2 shows the summary of the four data sets.

5.3.2 Experimental setup

Three criteria following their importance were considered to evaluate the performances of I-MOGASVM and other experimental methods; test accuracy, LOOCV accuracy, and the number of selected genes. Several experiments were conducted 10 times on each data set using I-MOGASVM and other experimental methods such as GASVM, MOGASVM, GASVM-II, and SVM. Next, the average result of the 10 independent runs was obtained. A near-optimal subset that produces the highest classification accuracies with the least possible number of genes is selected as the best subset.

5.3.3 Experimental results

Table 5.1 and Table 5.2 show the classification accuracy for each run using I-MOGASVM on all data sets. Interestingly, all runs have achieved 100% LOOCV accuracy on the data sets. This has proven that I-MOGASVM has efficiently selected and produced a near-optimal solution in a solution space. This is due to its ability to automatically reduce the dimensionality and complexity of the solution space on a cycle-by-cycle basis. Therefore, I-MOGASVM yields the near-optimal gene subset (a small subset of informative genes with high classification accuracy) successfully.

Table 5.1. Results for each run using I-MOGASVM on the leukemia and lung data sets.

Run no.	Leukemia data set			Lung data set		
	LOOCV (%)	Test (%)	No. selected genes	LOOCV (%)	Test (%)	No. selected genes
1	100	85.35	5	100	90.60	2
2	100	91.18	5	100	95.30	2
3	100	91.18	3	100	93.29	3
4	100	85.29	5	100	95.30	4
5	100	85.29	5	100	85.24	2
6	100	82.35	5	100	83.22	3
7	100	82.35	4	100	92.62	2
8	100	100	5	100	97.32	2
9	100	88.24	5	100	96.64	2
10	100	85.29	4	100	95.30	3
Average ± S.D	100 ± 0	87.65 ± 5.33	4.60 ± 0.70	100 ± 0	92.48 ± 4.80	2.5 ± 0.71

Note: The results of the best subsets are shown in the shaded cells. S.D. denotes the standard deviation.

Table 5.2. Results for each run using I-MOGASVM on the MLL and colon data sets.

Run no.	MLL data set			Colon data set	
	LOOCV (%)	Test (%)	No. selected genes	LOOCV (%)	No. selected genes
1	100	86.67	8	100	13
2	100	100	6	100	13
3	100	80	9	100	14
4	100	73.33	9	95.16	5
5	100	86.67	8	96.77	6
6	100	80	6	100	7
7	100	86.67	7	100	10
8	100	93.33	8	98.39	9
9	100	93.33	7	100	10
10	100	80	6	100	10
Average ± S.D	100 ± 0	86 ± 7.98	7.4 ± 1.17	99.03 ± 1.73	9.70 ± 3.06

Note: The results of the best subsets are shown in the shaded cells. S.D. denotes the standard deviation.

Table 5.3. The list of informative genes in the best gene subsets.

Data set	Run no.	Probe-set name
Leukemia	8	L15388_at M95678_at X15357_at X55668_at S76473_s_at
Lung	8	33328_at 609_f_at
MLL	2	35083_at 36436_at 36873_at 40518_at 35794_at 41827_f_at
Colon	6	H80240 T62220 H22688 T88902 U00968 T84082 T62947

Generally, near-optimal subsets that obtained from almost all run on the data sets contain less than 10 genes. This is inline with the diagnostic goal of developed medical procedures that needs the least number of possible informative genes to detect diseases. The conservativeness of the results in Tables 5.1 and 5.2 is controlled and maintained by the iterative approach and the fitness function of I-MOGASVM that maximizes the classification accuracy and meanwhile, minimizes the number of selected genes.

Practically, the best subset of a data set is firstly chosen and the genes in it are then listed for biological usage. The best subset is chosen based on the highest classification accuracy with the smallest number of selected genes. The highest accuracy gives confidence to us for the most accurate classification of cancer types. Moreover, the smallest number of selected genes for cancer classification can reduce the cost in clinical settings.

Table 5.4. The benchmark of the proposed I-MOGASVM with the other experimental methods and previous related works on the leukemia and lung cancer data sets.

Method	Leukemia data set (Average ± S.D; The best)			Lung data set (Average ± S.D; The best)		
	No. selected genes	Accuracy (%)		No. selected genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
I-MOGASVM	(4.60 ± 0.70; 5)	(100 ± 0; 100)	(87.65 ± 5.33; 100)	(2.5 ± 0.71; 2)	(100 ± 0; 100)	(92.48 ± 4.80; 97.32)
GASVM-II [23]	(10 ± 0; 10)	(100 ± 0; 100)	(81.18 ± 10.21; 94.12)	(10 ± 0; 10)	(100 ± 0; 100)	(59.33 ± 29.32; 97.32)
MOGASVM [25]	(2,212.6 ± 26.63; 2,189)	(95.53 ± 1.27; 97.37)	(84.41 ± 2.42; 88.24)	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)
GASVM [23]	(3,574.9 ± 40.05; 3,531)	(94.74 ± 0; 94.74)	(83.53 ± 2.48; 88.24)	(6,267.8 ± 56.34; 6,342)	(75 ± 0; 75)	(84.77 ± 2.53; 87.92)
SVM [23]	(7,129 ± 0; 7,129)	(94.74 ± 0; 94.74)	(85.29 ± 0; 85.29)	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)
Li *et al*. [20]	(4 ± NA; NA)	(100 ± NA; NA)	NA	NA	NA	NA
Peng *et al*. [29]	(6 ± NA; NA)	(100 ± NA; NA)	NA	NA	NA	NA
Huang and Chang [13]	(3.4 ± NA; NA)	(100 using 10-CV ± NA; NA)	NA	NA	NA	NA

Note: The best results are shown in the shaded cells. S.D. denotes the standard deviation, whereas 10-CV represents 10-fold-cross-validation. ‘NA’ means that a result is not reported in the related previous works. Methods in *italic* style are experimented in this research.

Informative genes in the best gene subsets, as produced by the proposed I-MOGASVM and reported in Tables 5.1 and 5.2, are listed in Table 5.3. These informative genes among thousands of other genes may be excellent candidates for clinical and medical investigations. Biologists can save much time, since they can refer directly to the genes that have the greatest possibility of being useful for cancer diagnosis and drug targeting in the future. A probe-set name is used for searching the biological information of genes in the public database of genes.

Table 5.5. The benchmark of the proposed I-MOGASVM with the other experimental methods and previous related works on the MLL and colon cancer data sets.

Method	MLL data set (Average ± S.D; The best)			Colon data set (Average ± S.D; The best)	
	No. selected genes	Accuracy (%)		No. selected genes	LOOCV Accuracy (%)
		LOOCV	Test		
I-MOGASVM	(7.4 ± 1.17; 6)	(100 ± 0; 100)	(86 ± 7.98; 100)	(9.7 ± 3.06; 7)	(99.03 ± 1.73; 100)
GASVM-II [23]	(30 ± 0; 30)	(100 ± 0; 100)	(84.67 ± 6.33; 93.33)	(30 ± 0; 30)	(99.03 ± 0.83; 100)
MOGASVM [25]	(4,465.2 ± 18.34; 437)	(94.74 ± 0; 94.74)	(90 ± 3.51; 93.33)	(446.3 ± 8.90; 446)	(93.23 ± 1.02; 95.16)
GASVM [23]	(6,298.8 ± 51.51; 224)	(94.74 ± 0; 94.74)	(87.33 ± 2.11; 86.67)	(979.8 ± 5.80; 940)	(91.77 ± 0.51; 91.94)
SVM [23]	(12,582 ± 0; 12,582)	(92.98 ± 0; 92.98)	(86.67 ± 0; 86.67)	(2,000 ± 0; 2,000)	(85.48 ± 0; 85.48)
Li *et al*. [20]	NA	NA	NA	15 ± NA; NA	(93.55 ± NA; NA)
Peng *et al*. [29]	NA	NA	NA	(12 ± NA; NA)	(93.55 ± NA; NA)

Note: The best results are shown in the shaded cells. S.D. denotes the standard deviation. ‘NA’ means that a result is not reported in the related previous works. Methods in *italic* style are experimented in this research.

For an objective comparison, the present work only compares I-MOGASVM with related previous works that used GASVM-based methods in their work [13],[20],[23],[25],[29]. Moreover, the previous works also produced the average of classification accuracy results since they used hybrid approaches. The present work makes the comparison using the averages of LOOCV accuracy and the number of selected genes. This is due to the most previous works only evaluated the performance of their approaches using the LOOCV procedure or *k*-fold-cross-validation and the number of selected genes on averages.

According to Tables 5.4 and 5.5, I-MOGASVM has outperformed the other experimental methods and previous work in terms of LOOCV accuracy, test accuracy, and the number of selected genes. The gap between LOOCV accuracy and test accuracy that resulted from using I-MOGASVM was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. Therefore, I-MOGASVM is more efficient than other experimental methods since it has produced higher classification accuracies, smaller numbers of selected genes, smaller standard deviations, and smaller gaps between LOOCV accuracy and test accuracy.

5.4 Summary

In this chapter, I-MOGASVM has been proposed and tested for gene selection on four sets of real gene expression data. Based on the experimental results, the performance of I-MOGASVM was superior to the other experimental methods and previous related work. This is due to the fact that I-MOGASVM can automatically reduce the dimensionality of the data on a cycle-by-cycle basis. When the dimensionality was reduced, the combination of genes and the complexity of solution spaces can also be automatically decreased iteratively. This iterative process is done to generate potential gene subsets from high-dimensional data (gene expression data), and finally produce a near-optimal subset of informative genes. Hence, gene selection using I-MOGASVM is needed to produce a near-optimal (small) subset of informative genes for cancer classification. Moreover, focusing attention on the informative genes in the best subset may provide insights into the mechanisms responsible for the cancer itself. Even though I-MOGASVM has achieved excellent performances, it is still not able to completely solve the over-fitting problem. Therefore, Chapter 6 will propose a two-stage method to solve the over-fitting problem.

Chapter 6

A Two-Stage Method

6.1 Introduction

This chapter introduces and proposes a two-stage gene selection method. The proposed method is to perform well in high-dimensional data and reduce a risk of over-fitting problems since it has two stages as follows: Stage 1 to decrease the dimensionality of data; stage 2 to produce a small (near-optimal) genes subset. The diagnostic goal is to develop a medical procedure based on the least number of possible genes that needed to detect diseases. Thus, the ultimate goal of this chapter is to select a small subset of informative genes (minimize the number of selected genes) for yielding high cancer classification accuracy (maximize the classification accuracy). To achieve the goal, the present work adopts the proposed two-stage method.

6.2 The Proposed Two-Stage Method (Filter+MOGASVM)

The proposed two-stage method is called Filter+MOGASVM because it uses a filter and MOGASVM in its stages. Filter+MOGASVM in the present work differs from the methods in the previous works in one major part [13],[23],[25],[26],[27],[29]. The major difference is that the proposed method involves two stages (using a filter method and a hybrid method), whereas the previous works usually used only one stage (using a hybrid method) for gene selection. The difference is necessary in order to produce a small (near-optimal) gene subset from high-dimensional data and reduce the high risk of over-fitting problems. For more understanding, the general flowcharts of the present work and the previous works are shown in Fig. 6.1 (a) and Fig. 6.1 (b), respectively. The detailed stages of Filter+MOGASVM are described as follows in the next two subsections.

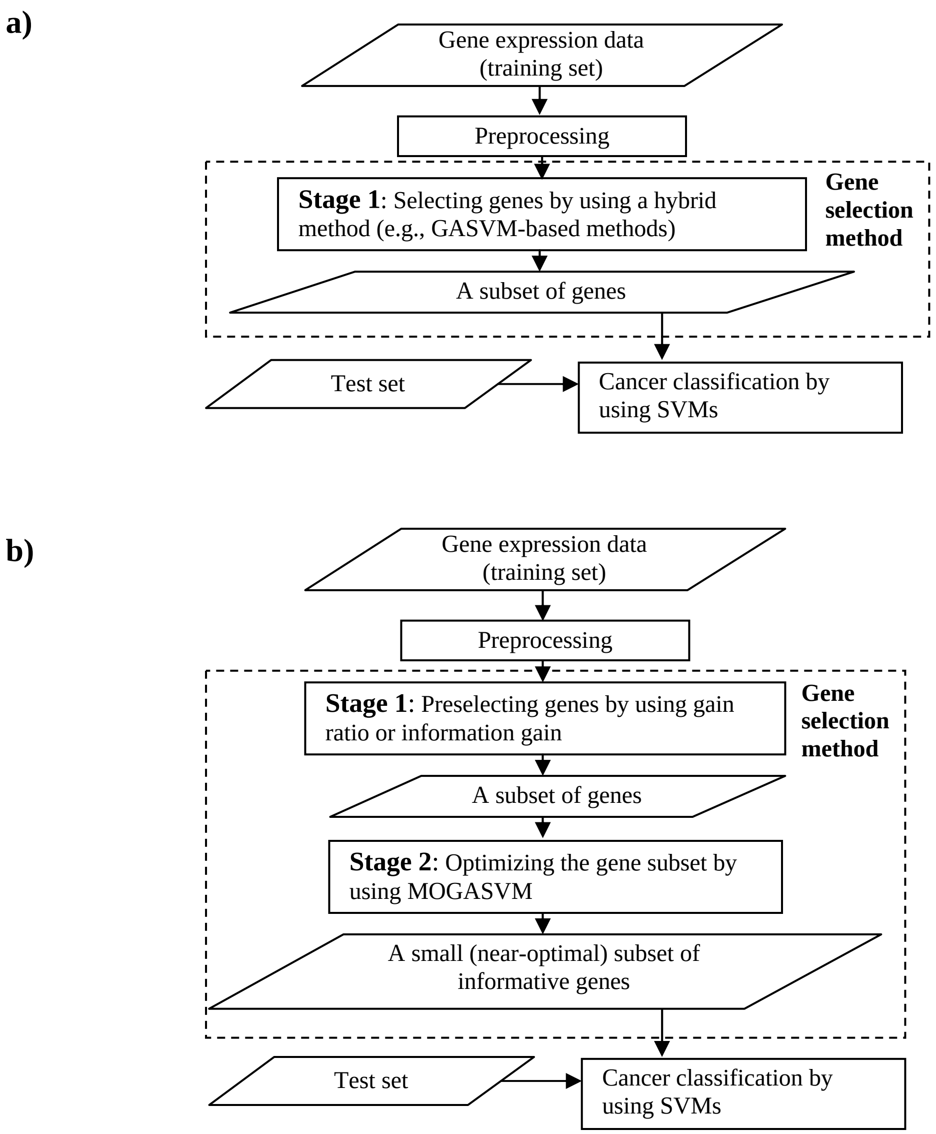

a)
Gene expression data (training set)
Preprocessing
Stage 1: Selecting genes by using a hybrid method (e.g., GASVM-based methods)
Gene selection method
A subset of genes
Test set
Cancer classification by using SVMs
b)
Gene expression data (training set)
Preprocessing
Stage 1: Preselecting genes by using gain ratio or information gain
Gene selection method
A subset of genes
Stage 2: Optimizing the gene subset by using MOGASVM
A small (near-optimal) subset of informative genes
Test set
Cancer classification by using SVMs

Fig.6.1. General flowcharts of (a) previous works (GASVM-based methods); (b) the present work (Filter+MOGASVM).

6.2.1 Stage 1: Preselecting genes using a filter method

In the first stage, the present work applies a filter method such as gain ratio (GR) or information gain (IG) on the training set to preselect genes and finally produce a subset of genes. After the preselection process, the dimensionality of data is also decreased. The filter method calculates and ranks a score for each gene. Genes with the highest scores are selected and put into the gene subset. This subset is used as an input to the second stage.

Since GASVM-based methods in previous works performs poorly in high-dimensional data, and meanwhile, the present work uses a GASVM-based method (MOGASVM) in the second stage of Filter+MOGASVM, a filter method (GR or IG) in this first stage is used to reduce the high-dimensional in order to overcome the drawback of GASVM-based methods. If the subset that is produced by the filter method is small-dimension, the combination of genes is not complex, and then MOGASVM in the next stage can possible to produce a small (near-optimal) subset of informative genes.

6.2.2 Stage 2: Optimizing a gene subset using MOGASVM

In this stage, the present work develops and uses MOGASVM to automatically optimize the gene subset that is produced by the first stage, and finally yield a small (near-optimal) subset of informative genes. This small subset is identified by an evaluation function in MOGASVM that uses two criteria; maximization of the leave-one-out-cross-validation (LOOCV) accuracy and minimizations of the number of selected genes. MOGASVM selects and optimizes genes by considering relations among them in order to remove irrelevant and noisy genes. The small subset is possible to be found due to the dimensionality and complexity of data has been firstly reduced by the first stage. The high risk of over-fitting problems can be also decreased because of the reduction. The detail of MOGASVM can be found in Mohamad *et al*. [25].

Finally, the small subset of the training set is used to construct SVM for cancer classification, and the constructed SVM is then tested by using the test set (independent set). This chapter has produced two methods of Filter+MOGASVM that are obtained from combinations of two different filter methods (GR and IG) and MOGASVM. These methods are GR+MOGASVM and IG+MOGASVM.

6.3 Experiments

6.3.1 Data sets

Three benchmark gene expression data sets that contain binary classes and multi-classes are used to evaluate Filter+MOGASVM. These data sets are the lung cancer, mixed-lineage leukemia (MLL) cancer, and leukemia cancer data sets. They are summarized in Table 2.1 in Chapter 2.

6.3.2 Experimental setup

Since the number of training samples in gene expression data is small, the cross-validation (CV) accuracy on the training set is calculated through the LOOCV procedure. For the test accuracy, SVM is built using all the training samples, and the classes of test samples from the test set are predicted one by one using the built SVM. The test accuracy is estimated by the number of the correctly classified samples, divided by the number of samples in the test set.

Table 6.1 contains parameter values for Filter+MOGASVM. These values are chosen based on the results of preliminary runs. Three criteria following their importance are considered to evaluate and compare the performance of Filter+MOGASVM with existing methods [13],[23],[25],[26],[27],[29] from viewpoints of the test accuracy, CV accuracy, and the number of selected genes. High accuracies and a small number of selected genes are needed to obtain an excellent performance. The top 200 genes are preselected by using GR and IG in the first stage of the proposed method, and are then used for the second stage. Several experiments are conducted 10 times on each data set using Filter+MOGASVM and other experimental methods such as GASVM (single-objective), MOGASVM, GASVM-II, and SVM. Filter+GASVM methods (IG+GASVM and GR+GASVM) are also experimented for the comparison. Next, an average result of the 10 independent runs is obtained.

Table 6.1. Parameter settings for Filter+MOGASVM.

Parameters	Lung data set	MLL data set	Leukemia data set
No. populations	100	100	100
No. generations	300	300	300
Crossover rate (Two-point)	0.7	0.7	0.7
Replacement rate (Roulette wheel selection)	0.8	0.8	0.8
Mutation rate (Flip)	0.01	0.01	0.01
w_1	0.7	0.7	0.8
w_2	0.3	0.3	0.2
Cost for SVMs	0.7	100	100

6.3.3 LOOCV and test accuracies of selected genes with Filter+MOGASVM

Tables 6.2, 6.3, and 6.4 show the results for each run on the lung, MLL, and leukemia data sets, respectively. The results of the best subsets are shown in the shaded cells, whereas the results in boldface display the best results of averages. S.D. denotes the standard deviation. Almost all runs have achieved 100% LOOCV accuracy on all the data sets. This has proved that Filter+MOGASVM has efficiently selected and produced a near-optimal gene subset from a solution space.

Table 6.2. Classification accuracies using Filter+MOGASVM on the lung data set.

Run no.	GR+MOGASVM (Filter+MOGASVM)			IG+MOGASVM (Filter+MOGASVM)		
	LOOCV (%)	Test (%)	No. selected genes	LOOCV (%)	Test (%)	No. selected genes
1	100	98.66	2	100	97.99	2
2	100	94.63	2	100	96.64	2
3	100	95.30	2	100	97.32	2
4	100	97.32	2	100	97.32	2
5	100	95.97	2	100	94.63	2
6	100	97.99	2	100	95.30	2
7	100	95.97	2	100	95.30	2
8	100	95.97	2	100	95.97	2
9	100	95.97	2	100	99.33	2
10	100	93.96	2	100	93.29	2
Average ± S.D.	100 ± 0	96.18 ± 1.45	2 ± 0	**100** **± 0**	**96.31** **± 1.77**	**2** **± 0**

Table 6.3. Classification accuracies using Filter+MOGASVM on the MLL data set.

Run no.	GR+MOGASVM (Filter+MOGASVM)			IG+MOGASVM (Filter+MOGASVM)		
	LOOCV (%)	Test (%)	No. selected genes	LOOCV (%)	Test (%)	No. selected genes
1	100	93.33	6	100	93.33	7
2	100	93.33	6	100	93.33	6
3	100	100	5	100	100	7
4	100	93.33	7	98.25	100	6
5	100	100	5	100	93.33	7
6	100	93.33	6	100	93.33	5
7	100	100	5	100	100	7
8	100	100	7	100	100	6
9	100	100	5	100	100	5
10	100	93.33	4	100	86.67	7
Average ± S.D.	**100** **± 0**	**96.67** **± 3.51**	**5.60** **± 0.97**	99.83 ± 0.56	96.00 ± 4.66	6.30 ± 0.82

Table 6.4. Classification accuracies using Filter+MOGASVM on the leukemia data set.

Run no.	GR+MOGASVM (Filter+MOGASVM)			IG+MOGASVM (Filter+MOGASVM)		
	LOOCV (%)	Test (%)	No. selected genes	LOOCV (%)	Test (%)	No. selected genes
1	100	91.18	3	100	91.18	3
2	100	88.24	3	100	91.18	3
3	100	94.12	3	100	94.12	3
4	100	91.18	2	100	91.18	2
5	100	91.18	3	100	91.18	3
6	100	94.12	3	100	88.24	2
7	100	91.18	2	100	94.12	2
8	100	91.18	3	100	88.24	3
9	100	94.12	3	100	85.30	3
10	100	91.18	3	100	91.18	3
Average ± S.D.	**100 ± 0**	**91.77 ± 1.86**	**2.70 ± 0.48**	100 ± 0	90.59 ± 2.70	2.70 ± 0.48

6.3.4 Filter+MOGASVM versus other experimental methods

The benchmark of Filter+MOGASVM in comparison with other experimental methods that have been experimented in this work is summarized in Tables 6.5 and 6.6. Overall, the LOOCV and test accuracies of Filter+MOGASVM for all the data sets were higher than Filter+GASVM, MOGASVM, GASVM-II, GASVM, and SVM. Moreover, the number of selected genes by using Filter+MOGASVM was also lower.

Based on the standard deviations of LOOCV accuracy, test accuracy, and the number of selected genes, Filter+MOGASVM was also more consistent than the other experimental methods except for SVM. This SVM achieved 0 for the standard deviations in all experiments since it did not implement any gene selection approach. The gap between LOOCV accuracy and test accuracy that are obtained by Filter+MOGASVM was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. On the other hand, the results of LOOCV accuracy of the others were much higher than their test accuracy because they were unable to avoid or reduce the risk of over-fitting problems. Over-fitting is a major problem of hybrid methods in gene selection and classification of gene expression data when the classification accuracy on training samples, e.g., LOOCV accuracy is much higher than the test accuracy.

Table 6.5. The benchmark of Filter+MOGASVM with Filter+GASVM and the previous methods on the leukemia data set.

Method	Leukemia data set (Average ± S.D.; The best)		
	No. selected genes	Accuracy (%)	
		LOOCV	Test
GR+MOGASVM (Filter+MOGASVM)	2.70 ± 0.48; 3	100 ± 0; 100	91.77 ± 1.86; 94.12
IG+MOGASVM (Filter+MOGASVM)	2.70 ± 0.48; 2	100 ± 0; 100	90.59 ± 2.70; 94.12
GR+GASVM (Filter+GASVM)	97.40 ± 4.43; 91	100 ± 0; 100	86.18 ± 1.99; 88.24
IG+GASVM (Filter+GASVM)	99.30 ± 6.29; 96	100 ± 0; 100	88.53 ± 2.93; 91.18
A cyclic hybrid method [27]	2.9 ± 1.73; 2	100 ± 0; 100	88.82 ± 3.04; 94.12
GASVM-II+GASVM [26]	3.4 ± 1.35; 2	100 ± 0; 100	85.88 ± 8.86; 97.06
GASVM-II [23]	10 ± 0; 10	100 ± 0; 100	81.18 ± 0.21; 94.12
MOGASVM [25]	2,212.6 ± 26.63; 2,189	95.53 ± 1.27; 97.37	84.41 ± 2.42; 88.24
GASVM [23]	3,574.9 ± 40.05; 3,531	94.74 ± 0; 94.74	83.53 ± 2.48; 88.24
SVM [25]	7,129 ± 0; 7,129	94.74 ± 0; 94.74	85.29 ± 0; 85.29

Note: The best results of each data set are shown in the shaded cells. S.D. denotes the standard deviation.

Table 6.6. The benchmark of Filter+MOGASVM with Filter+GASVM and the previous methods on the lung and MLL data sets.

Method	Lung data set (Average ± S.D.; The best)			MLL data set (Average ± S.D.; The best)		
	No. selected genes	Accuracy (%)		No. selected genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
GR+MOGASVM (Filter+MOGASVM)	2 ± 0; 2	100 ± 0; 100	96.18 ± 1.45; 98.66	5.60 ± 0.97; 5	100 ± 0; 100	96.67 ± 3.51; 100
IG+MOGASVM (Filter+MOGASVM)	2 ± 0; 2	100 ± 0; 100	96.31 ± 1.77; 99.33	6.30 ± 0.82; 5	99.83 ± 0.56; 100	96.00 ± 4.66; 100
GR+GASVM (Filter+GASVM)	101 ± 8.50; 105	100 ± 0; 100	86.04 ± 3.66; 90.60	100.40 ± 6.42; 98	100 ± 0; 100	90.67 ± 5.62; 100
IG+GASVM (Filter+GASVM)	100.3 ± 8.02; 87	100 ± 0; 100	84.30 ± 7.86; 88.59	100.20 ± 7.63; 99	100 ± 0; 100	88.67 ± 3.22; 93.33
A cyclic hybrid method [6]	2.80 ± 1.32; 4	100 ± 0; 100	93.69 ± 2.52; 98.66	12.0 ± 5.58; 20	100 ± 0; 100	91.33 ± 5.49; 100
GASVM-II+GASVM [26]	2.1 ± 0.32; 2	100 ± 0; 100	94.16 ± 6.85; 98.66	6.5 ± 0.71; 6	100 ± 0; 100	92 ± 8.20; 100
GASVM-II [23]	10 ± 0; 10	100 ± 0; 100	59.33 ± 29.32; 97.32	30 ± 0; 30	100 ± 0; 100	84.67 ± 6.33; 93.33
MOGASVM [25]	4,418.5 ± 50.19; 4,433	75.31 ± 0.99; 78.13	85.84 ± 3.97; 93.29	4,465.2 ± 18.34; 4,437	94.74 ± 0; 94.74	90 ± 3.51; 93.33
GASVM [23]	6,267.8 ± 56.34; 6,342	75 ± 0; 75	84.77 ± 2.53; 87.92	6,298.8 ± 51.51; 6,224	94.74 ± 0; 94.74	87.33 ± 2.11; 86.67
SVM [25]	12,533 ± 0; 12,533	65.63 ± 0; 65.63	85.91 ± 0; 85.91	12,582 ± 0; 12,582	92.98 ± 0; 92.98	86.67 ± 0; 86.67

Note: The best results of each data set are shown in the shaded cells. S.D. denotes the standard deviation.

GASVM and MOGASVM cannot produce a near-optimal subset of informative genes because they perform poorly in high-dimensional data due to their chromosome representation drawback. GASVM-II method is impractical to be used in real applications because a variety number of selected genes should be tested in order to obtain the near-optimal one. On the contrary, the proposed Filter+MOGASVM that preselects a number of genes in the first stage can automatically optimize the selected genes by the second stage in order to remove irrelevant genes and produce a small (near-optimal) subset of informative genes.

6.3.5 Filter+MOGASVM versus previous related works

For an objective comparison, the present work only compares Filter+MOGASVM with related previous works that used GASVM-based methods in their works [13],[29]. This is due to the present work also uses a GASVM-based method (MOGASVM) in Filter+MOGASVM. The previous works also produced the averages of LOOCV accuracy or *k*-fold-cross-validation, and the number of selected genes since they used hybrid approaches in their works.

Table 6.7. The comparison between the proposed method (Filter+MOGASVM) and other previous GASVM-based methods.

Data	Experiment Evaluation	The present work (Filter+MOGASVM)	Huang and Chang, [13]	Peng *et al*., [29]
Leukemia (Average ± S.D; The best)	CV Accuracy (%)	100 ± 0; 100 (using LOOCV)	100 ± NA; NA (using 10-CV)	100 ± NA; NA (using LOOCV)
	Test Accuracy (%)	91.77 ± 1.86; 94.12	NA	NA
	No. selected genes	2.70 ± 0.48; 3	3.4 ± NA; NA	6 ± NA; NA

Note: The best result is shown in the shaded cells. ‘NA‘ means that results are not reported in the related previous works. S.D. denotes the standard deviation, whereas 10-CV means 10-fold-cross-validation.

At the moment, only the leukemia data set is experimented as a comparison between the present work and other previous works [13],[29]. This is due to the previous works have used the leukemia in their experiments. Table 6.7 displays the benchmark of this work and previous related works on the leukemia data set. The averages of LOOCV accuracy and the number of selected genes of the present work were 100% and 2.7 genes, respectively. The latest previous work [13] also came up with the similar LOOCV result to the present ones, but the number of selected genes is slightly higher in order to obtain the same result. The work of Peng *et al.*, [29] analyzed this data set and finally yielded 100% average LOOCV accuracy with six average selected genes. Overall, this work has outperformed the related previous works on the data set in terms of classification accuracy and the number of selected genes. Filter+MOGASVM in the present work has produced a near-optimal (small) gene subset from high-dimensional data and reduced the high risk of over-fitting problems. This is due to the fact that a filter method in the first stage of Filter+MOGASVM reduces the dimensionality of the solution space in order to produce a gene subset. Next, MOGASVM in the second stage of Filter+MOGASVM optimizes the subset automatically to yield a small subset of informative genes with high classification accuracy. This small subset is obtained since Filter+MOGASVM considers and optimizes a relation among genes.

Unfortunately, the previous works [13],[29] did not provide any test accuracy result on the test set (independent data set) and did not show any standard deviation result for comparative comparison with the present work. GASVM-based methods in the previous works may almost possible face with a high risk of over-fitting problems and the difficulty to obtain a near-optimal solution in high-dimensional data since they used binary chromosome representation for gene selection mechanisms. This was also supported by a review paper in [30] which reported that hybrid methods (e.g., GASVM-based methods) confront with the risk of over-fitting problems because of the high-dimensional data.

6.4 Summary

In this chapter, Filter+MOGASVM has been proposed and tested for gene selection on three real gene expression data sets that contain binary classes and multi-classes of tumor samples. The performance of Filter+MOGASVM was superior to the other experimental methods and related previous works. This is due to the fact that the filter method in the first stage of the proposed method can preselect genes and reduce dimensionality of data in order to produce a subset of genes. When the dimensionality was reduced, the combination of genes and complexity of solution spaces were automatically decreased. The second stage of Filter+MOGASVM can automatically optimize the subset that is yielded by the first stage. This optimization process is done to remove irrelevant and noisy genes, and finally produce a small (near-optimal) subset of informative genes. Hence, the gene selection using Filter+MOGASVM is needed to produce a small subset of informative genes for excellent cancer classification of gene expression data. Based on the experimental results, a two-stage method (Filter+MOGASVM) could solve the over-fitting problem. Filter+MOGASVM has achieved excellent performance in terms of classification accuracy and the number of selected genes, but it is still not very high accuracy. Therefore, Chapter 7 will propose a three-stage method to highly increase classification accuracy.

Chapter 7
A Three-Stage Method

7.1 Introduction

There are two main drawbacks of the hybrid methods (GASVM-based methods) in the previous works [13],[23],[25],[26],[29],[31] are intractable to efficiently produce a small subset of informative genes when the total number of genes is too large (high-dimensional data); 2) the high risk of over-fitting problems. In order to solve the problems derived from gene expression data and overcome the limitations of the hybrid methods in the previous works, this chapter describes and proposes a three-stage method (3-SGS) for gene selection.

7.2 The Proposed Three-Stage Method (3-SGS)

A three-stage method (3-SGS) contains three stages for gene selection. 3-SGS in the present work differs from the methods in the previous works in one major part. The major difference is that the proposed method involves three stages (using a filter method, a hybrid method, and frequency analysis), whereas the previous works usually used only one stage (using a hybrid method) [13],[23],[25],[26],[31] or two stages (using a filter method and a hybrid method) [29]. The difference is necessary in order to produce near-optimal gene subsets from high-dimensional data, reduce the high risk of over-fitting problems, and finally yield a small subset of informative genes. The computational flow of 3-SGS for gene selection is shown in Fig. 7.1.

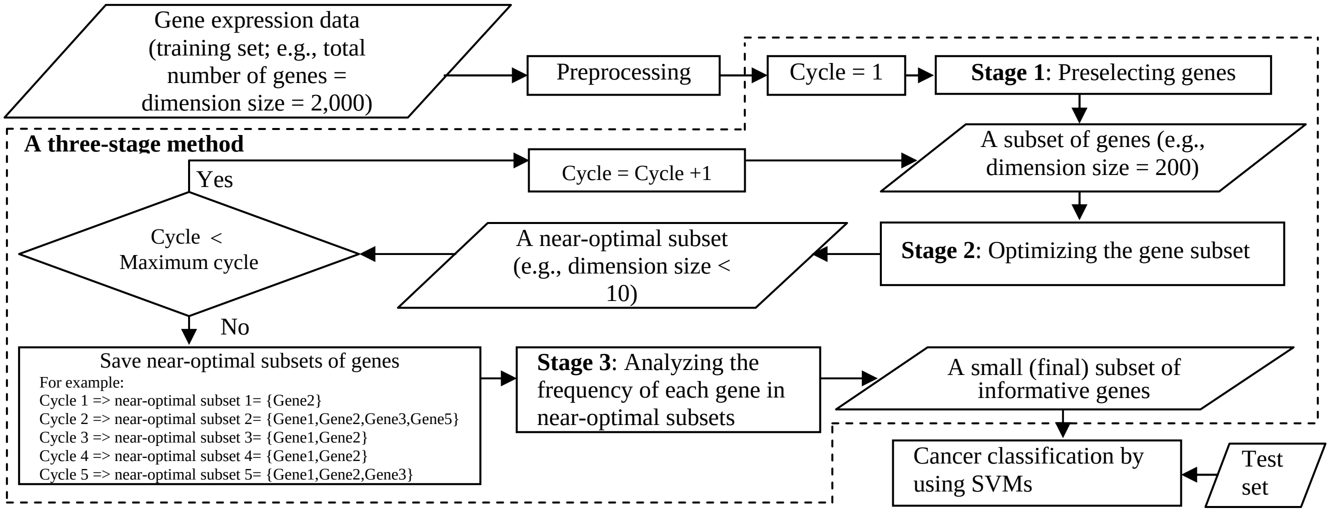

Gene expression data (training set; e.g., total number of genes = dimension size = 2,000)
Preprocessing
Cycle = 1
Stage 1: Preselecting genes
A subset of genes (e.g., dimension size = 200)
A three-stage method
Yes
Cycle = Cycle +1
Cycle < Maximum cycle
A near-optimal subset (e.g., dimension size < 10)
Stage 2: Optimizing the gene subset
No
Save near-optimal subsets of genes
For example:
Cycle 1 => near-optimal subset 1= {Gene2}
Cycle 2 => near-optimal subset 2= {Gene1,Gene2,Gene3,Gene5}
Cycle 3 => near-optimal subset 3= {Gene1,Gene2}
Cycle 4 => near-optimal subset 4= {Gene1,Gene2}
Cycle 5 => near-optimal subset 5= {Gene1,Gene2,Gene3}
Stage 3: Analyzing the frequency of each gene in near-optimal subsets
A small (final) subset of informative genes
Cancer classification by using SVMs
Test set

Fig.7.1. The proposed three-stage method (3-SGS).

7.2.1 Stage 1: Preselecting genes using a filter method

A filter method such as gain ratio (GR) or information gain (IG) is used in this stage (stage 1) to preselect genes and produce a subset of genes. After the preselect process, the dimensionality of data is also decreased. The filter method calculates and ranks a score for each gene. Genes with the highest scores are selected and put into a gene subset. This subset is then used as an input to the second stage. A GASVM-based method, i.e., MOGASVM that performs poorly in high-dimensional data is implemented in the second stage of 3-SGS. Therefore, the filter method (GR or IG) is firstly used to reduce the high-dimension in order to overcome the drawback of the GASVM-based method. If the subset that produced by the filter method is in small-dimension, the combination of genes is not complex, and then MOGASVM is possible to produce near-optimal genes subsets.

7.2.2 Stage 2: Optimizing a gene subset using MOGASVM

In this stage, MOGASVM optimizes gene subsets that are produced by the first stage, and finally yields near-optimal subsets of genes. This stage is cycled until the maximum number of cycles is satisfied. The near-optimal subsets are identified by an evaluation function in MOGASVM that uses two criteria: maximization of LOOCV accuracy and minimization of the number of selected genes. MOGASVM selects and optimizes genes by considering relations among them in order to remove irrelevant and noisy genes. The near-optimal subsets can be obtained since the dimensionality and complexity of data has been firstly reduced by the first stage. The high risk of over-fitting problems can be also decreased because of the reduction in the first stage. The detail of MOGASVM can be found in the previous work [25].

7.2.3 Stage 3: Analyzing the frequency of each gene in near-optimal subsets

In this stage, frequency analysis is implemented to identify the most frequently selected genes in near-optimal gene subsets. The frequency of appearance of each gene in each near-optimal gene subset is examined and analyzed to assess the relative importance of genes for cancer classification. The most frequently selected genes in near-optimal gene subsets are presumed to be the most relevant for the classification. Finally, a small (final) subset of informative genes (*K* genes, *K* is a number of genes) is produced and used to construct SVM. This subset contains a small number of informative genes with high classification accuracy. Table 7.1 shows an example on how to obtain the frequency of each gene and the final subset of informative genes. This chapter has produced two methods of 3-SGS obtained from combinations of two different filter methods (GR and IG) and MOGASVM. These methods are 3-SGS-GR and 3-SGS-IG.

Table 7.1. An example to obtain the frequency of each genes (assume that the maximum number of cycles is five).

Cycle	Near-optimal gene subset			
	Gene 1	Gene 2	Gene 3	Gene 4
1	N	Y	N	N
2	Y	Y	Y	N
3	Y	Y	N	N
4	Y	Y	N	N
5	Y	Y	Y	N
Frequency	4	5	2	0
A final subset of informative genes (following the most frequently selected genes)	Gene 2;	Gene 1;	Gene 3;	

Note: 'Y' means that the corresponding gene is included in a near-optimal gene subset. Otherwise, 'N' means that the corresponding gene is not included.

7.3 Experiments

7.3.1 Data sets and experimental setup

Five benchmark gene expression data sets that contain binary classes and multi-classes of cancer samples are used to evaluate 3-SGS. These data sets are the leukemia cancer, colon cancer, lung cancer, and mixed-lineage leukemia (MLL) cancer, and small round blue cell tumors (SRBCT) data sets. The first four data sets have been summarized on Table 2.1 in Chapter 2. The SRBCT data set is a multi-classes data set. It has four classes; ewing family of tumors (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and burkitt lymphomas (BL). The training set contains 63 samples (22 EWS, 20 RMS, 12 NB, and 8 BL), whereas the test set contains 20 samples (6 EWS, 5 RMS, 6 NB, and 3 BL). There are 2,308 genes in each sample. It can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/.

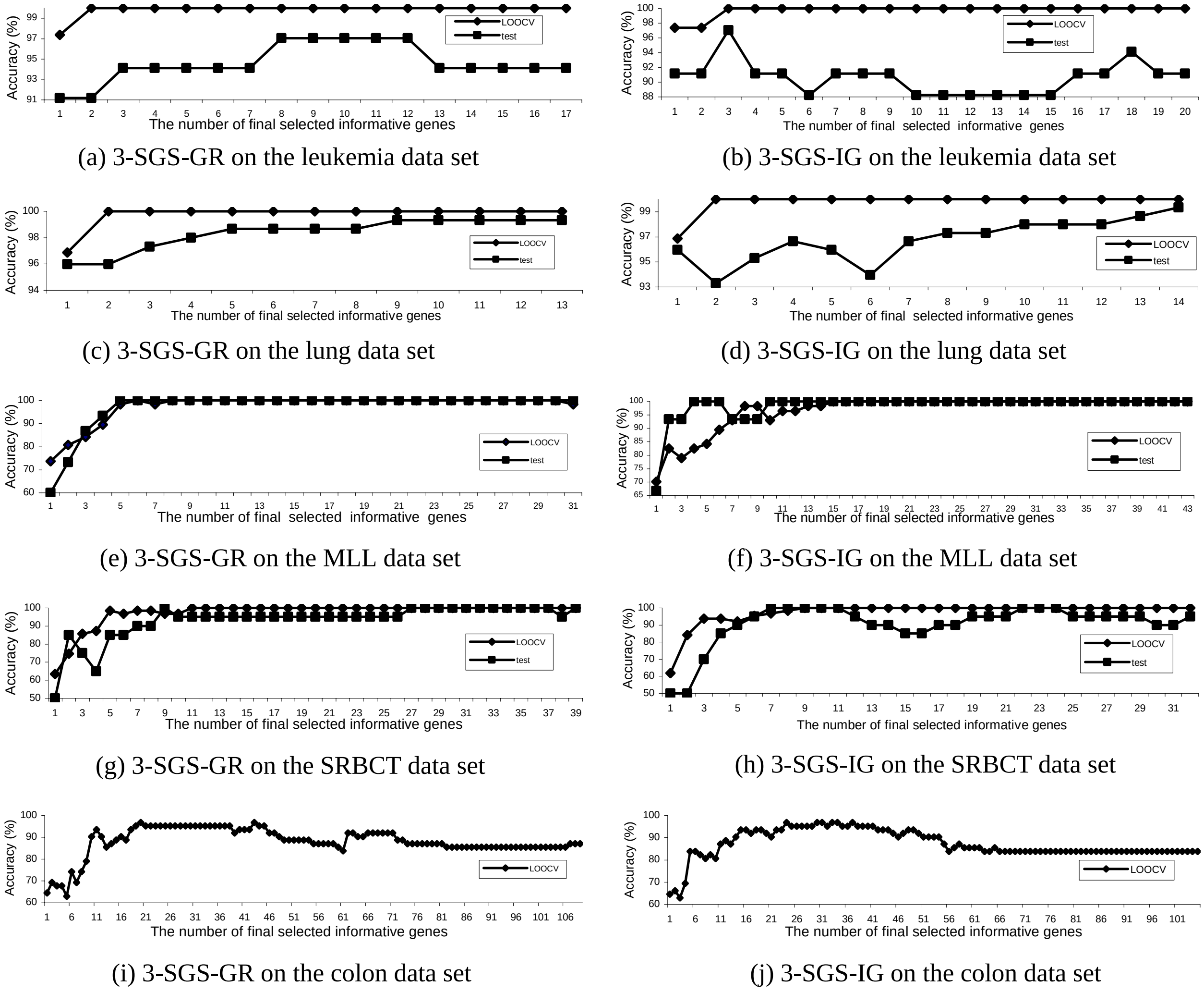
Table 7.2 contains parameter values for 3-SGS. These values are chosen based on the results of preliminary runs. Three criteria following their importance are considered to evaluate the performance of 3-SGS; test accuracy on the test set, LOOCV accuracy on the training set, and the number of selected genes. High accuracies and a small number of selected genes are needed to obtain an excellent performance. The top 200 genes are preselected by using GR and IG in the first stage of the 3-SGS, and are then used for the second stage.

Table 7.2. Parameter settings for 3-SGS.

Data set Parameters	Leukemia	Lung	MLL	SRBCT	Colon
No. populations	100	100	100	100	100
No. generations	300	300	300	300	300
Crossover rate (Two-point)	0.7	0.7	0.7	0.7	0.7
Mutation rate (Flip)	0.01	0.01	0.01	0.01	0.01
The maximum number of cycles	10	10	10	10	10
Cost for SVMs	100	0.7	100	100	100

7.3.2 Classification accuracies of the final subset of informative genes

As shown in Fig. 7.2, the best results of the leukemia (100% LOOCV and 97.06% test accuracies), the lung (100% LOOCV and 99.33% test accuracies), the MLL (100% LOOCV and 100% test accuracies), the SRBCT (100% LOOCV and 100 % test accuracies), and the Colon data sets (96.77% LOOCV accuracy) are obtained by using the only three (using 3-SGS-IG), nine (using 3-SGS-GR), six (using 3-SGS-GR), nine (using 3-SGS-IG), and 20 (using 3-SGS-GR) final selected informative genes (*K* genes), respectively. Many runs have achieved 100% LOOCV accuracy on all the data sets. These results have proved that 3-SGS has efficiently selected and produced a small subset of informative genes from high-dimensional data.



(a) 3-SGS-GR on the leukemia data set

(b) 3-SGS-IG on the leukemia data set

(c) 3-SGS-GR on the lung data set

(d) 3-SGS-IG on the lung data set

(e) 3-SGS-GR on the MLL data set

(f) 3-SGS-IG on the MLL data set

(g) 3-SGS-GR on the SRBCT data set

(h) 3-SGS-IG on the SRBCT data set

(i) 3-SGS-GR on the colon data set

(j) 3-SGS-IG on the colon data set

Fig.7.2. A relation between classification accuracies and the number of final selected informative genes (*K* genes) using 3-SGS.

7.3.3 A list of informative genes for biological usage

The informative genes and their rank scores (frequency) of the final subsets as produced by the proposed 3-SGS and reported in Fig. 7.2 are listed in Table 7.3. These informative genes among the thousand of genes may be the excellent candidates for clinical and medical investigations. Biologists can save much time since they can directly refer to the genes that have high possibility to be useful for cancer diagnosis and drug target in the future. A gene

ID or gene card ID is used for searching the biological information of genes in the public database of genes.

Table 7.3. The list of informative genes in the final gene subsets.

Data set	Rank score	Gene ID / Gene card ID
Leukemia	5	M23197
	2	X95735
	2	X85116
Lung	5	W28612
	3	AL050224
	2	AB020647
	1	X05323
	1	AI201108
	1	AL049381
	1	S71043
	1	AJ012737
	1	Y00318
MLL	9	M11722
	7	M13143
	3	U41843
	3	Z83844
	2	L08895
	2	U59878
SRBCT	5	GC16M088332
	4	GC01P149298
	4	GC02M091189
	3	GC02P191818
	3	GC08M042151
	3	GC13M046243
	2	GC07P115952
	2	GC18M023784
	2	GC11M002110

7.3.4 3-SGS versus other previous methods

Tables 7.4, 7.5, 7.6, and 7.7 show the benchmark of this work and previous related works. Both 3-SGS and the two-stage method had 100% LOOCV accuracy on the leukemia data set, but the increase for 3-SGS was 3.22% in the colon data set. Comparing 3-SGS with the best results of one-stage categories on all the data sets except the Colon data set, 3-SGS showed improvements between 0.67% and 18.5% on test accuracy. 3-SGS not obtain excellent results on the colon data set because this data set has smaller total number of genes (only 2,000 genes) compared to other data sets. This prove that 3-SGS is suitable only for very high-dimensional data (more than 2,000 genes).

Table 7.4. The benchmark of 3-SGS with previous methods on the leukemia and colon data sets.

Category	Gene selection method	Leukemia data set				Colon data set		
		No. selected genes	Accuracy (%)		Time taken (Hour)	No. selected genes	CV accuracy (%)	Time taken (Hour)
			CV	Test				
Three-stage	3-SGS	3	100	97.07	(0.17)	20	96.77	(3.23)
Two-stage	GASVM [29]	6	100	-	-	12	93.55	-
One-stage	GASVM [13]	(3.4)	(100)	-	-	-	-	-
	Signal-to-noise-ratio [10]	50	94.74	85.29	-	-	-	-
	t-test [11]	-	-	-	-	-	-	-
	GASVM-II+GASVM [26]	(4.5)	(100)	(85.88)	(2.22)	(11.6)	(99.52)	(11.87)
	GASVM-II [23]	(10)	(100)	(81.18)	(1.37)	(30)	(99.03)	(10.24)
	MOGASVM [25]	(2,212.6)	(95.53)	(84.41)	(94.65)	(446.3)	(93.23)	(76.46)
	GASVM [23]	(3,574.9)	(94.74)	(83.53)	(101.23)	(979.8)	(91.77)	(98.23)

Note: The results of the best subsets are shown in the shaded cells. '-' means that the results are not reported in the related previous work. A result in '( )' denotes an average result. CV represents cross-validation. Methods in *italics* style are experimented in this work.

Overall, the proposed 3-SGS has outperformed the previous works (one-stage and two-stage methods) on all the data sets except the colon data set in terms of test accuracy, LOOCV accuracy, and the number of selected genes. This is due to the fact that a filter method in the first stage of 3-SGS reduces the dimensionality of the solution space in order to produce a gene subset. Next, MOGASVM in the second stage of 3-SGS optimizes the subset automatically to yield near-optimal subsets of genes. These subsets are obtained since MOGASVM in 3-SGS considers and optimizes a relation among genes. Finally, the first *K* genes appearing most frequently are selected as the final selected informative genes for cancer classification.

Table 7.5. The benchmark of 3-SGS with previous methods on the MLL data set.

Category	Gene selection method	MLL data set			
		No. selected genes	Accuracy (%)		Time taken (Hour)
			CV	Test	
Three-stage	3-SGS	6	100	100	(9.23)
One-stage	GASVM [13]	(3.5)	(100)	-	-
	Principal component analysis [3]	100	95	-	-
	GASVM-II+GASVM [26]	(6.5)	(100)	(92)	(46.48)
	GASVM-II [23]	(30)	(100)	(84.67)	(22.64)
	MOGASVM [25]	(4,465.2)	(94.74)	(90)	(260.54)
	GASVM [23]	(6,298.8)	(94.74)	(87.33)	(534.08)

Note: The results of the best subsets are shown in the shaded cells. '-' means that the results are not reported in the related previous work. A result in '( )' denotes an average result. CV represents cross-validation. Methods in *italics* style are experimented in this work.

Table 7.6. The benchmark of 3-SGS with previous methods on the SRBCT data set.

Category	Gene selection method	SRBCT data set			
		No. selected genes	Accuracy (%)		Time taken (Hour)
			CV	Test	
Three-stage	3-SGS	9	100	100	(3.51)
One-stage	GASVM [13]	(6.2)	(98.75)	-	-
	Principal component analysis [16]	78	100	-	-
	GASVM-II [23]	(10)	(99.84)	(68)	(12.86)
	MOGASVM [25]	(444.7)	(100)	(81.5)	(86.56)
	GASVM [23]	(1146)	(98.41)	(78.5)	(157.69)

Note: The results of the best subsets are shown in the shaded cells. '-' means that the results are not reported in the related previous work. A result in '( )' denotes an average result. CV represents cross-validation. Methods in *italics* style are experimented in this work.

Table 7.7. The benchmark of 3-SGS with previous methods on the lung data set.

Category	Gene selection method	Lung data set			
		No. selected genes	Accuracy (%)		Time taken (Hour)
			CV	Test	
Three-stage	3-SGS	9	100	99.33	(1.24)
One-stage	GASVM [31]	8	100	98.66	-
	t-test [11]	4	-	97.32	-
	GASVM-II+GASVM [26]	(2.1)	(100)	(94.16)	(7.57)
	GASVM-II [23]	(10)	(100)	(59.33)	(7.10)
	MOGASVM [25]	(4,418.5)	(75.31)	(85.84)	(110.23)
	GASVM [23]	(6,267.8)	(75)	(84.77)	(113.57)

Note: The results of the best subsets are shown in the shaded cells. '-' means that the results are not reported in the related previous work. A result in '( )' denotes an average result. CV represents cross-validation. Methods in *italics* style are experimented in this work.

Generally, filter methods in previous works [3],[10],[11],[16] achieved poor performances since they may result in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification. These bad performances occurred because the methods evaluated a gene based on its discriminative power for the target classes without considering its relations with other genes.

GASVM-based methods [13],[23],[25],[26],[29],[31] may be unable to produce a small subset of informative genes because they perform poorly in high-dimensional data due to their chromosome representation drawback. GASVM-II [23] method is impractical to be used in real applications because a variety number of selected genes should be tested in order to obtain the near-optimal one. On the contrary, the proposed 3-SGS that preselects a number of genes at the first stage can reduce the data dimensionality and produce a gene subset. This subset is then optimized by MOGASVM in the second stage of 3-SGS to yield near-optimal subsets. Finally, the first *K* genes appearing most frequently are selected as the final selected informative genes (a small subset) for cancer classification.

The gap between LOOCV accuracy and test accuracy that resulted by 3-SGS was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. On the other hand, the results of LOOCV accuracy of the related previous works [23],[25],[26],[31] were much higher than their test accuracy because they were unable to avoid or reduce the risk of over-fitting problems. Other previous works by GASVM-based methods [13],[29] did not provide any test accuracy results and thus, the over-fitting problem could not be investigated in their works. Over-fitting is a major problem on hybrid methods in gene selection and classification of gene expression data when the classification accuracy on training samples, e.g., LOOCV accuracy is much higher than the test accuracy. This was also supported by a review paper in Saeys *et al*. [31] which reported that hybrid methods (e.g., GASVM-based methods) confront with the high risk of over-fitting problems because of the high-dimensional data.

7.4 Summary

In this chapter, 3-SGS has been proposed and tested for gene selection on five gene expression data sets that contain binary classes and multi-classes of tumor samples. Based on the experimental results, the performance of 3-SGS was superior to other methods in related previous works. This is due to the fact that the filter method in the first stage of the 3-SGS can preselect genes and reduce dimensionality of data in order to produce a subset of genes. When the dimensionality was reduced, the combination of genes and complexity of solution spaces were automatically decreased. The second stage of 3-SGS can automatically optimize the subset that is yielded by the first stage in order to produce near-optimal gene subsets. Finally, the first *K* genes appearing most frequently are selected as the final selected informative genes (a small subset) for cancer classification. Generally, 3-SGS in this chapter also obtains short running time because of the large number of genes are removed by a filter technique in the first stage. As a conclusion, 3-SGS has obtained high classification accuracy with a few numbers of selected genes. However, due to the application of a filter method in the first stage of 3-SGS, the number of preselected genes is difficult since it is manually done. Moreover, 3-SGS is difficult to select the best number of *K* genes in implementation because there are many *K* genes in the third stage. Therefore, Chapter 8 will introduce particle swarm optimization because it is easy to implement, has few parameters to adjust, and has been successfully applied in many area.

Chapter 8

Modified Binary Particle Swarm Optimization Based on Introduced Particle's Speed and a Novel Rule

8.1 Introduction

Recently, several gene selection methods based on particle swarm optimization (PSO) have been proposed to select informative genes from gene expression data [1],[8],[21],[33]. PSO is a new population based stochastic optimization technique proposed by Kennedy and Eberhart [14]. It is motivated from the simulation of social behavior of organisms such as bird flocking and fish schooling. Alba *et al*. [1] have firstly evaluated a new version of PSO, called geometric PSO for gene selection. Unfortunately, the experimental results are less significant because the geometric PSO is more about generalizing optimizers based on a notion of distance where different distance metrics give a rise to different operators with regards to the predefined geometric operators. Shen *et al*. [33] have proposed a hybrid of PSO and tabu search approaches for gene selection. However, the results obtained by using the hybrid method are less meaningful since the application of tabu approaches in PSO is unable to search a near-optimal solution in search spaces. Next, an improved binary PSO have been proposed by Chuang *et al*. [8]. This approach produced 100% classification accuracy in many data sets, but it used a high number of selected genes (large gene subset) to achieve the high accuracy. It uses the high number because of the global best particle is reset to zero position when its fitness values do not change after three consecutive iterations. After that, Li *et al*. [21] have introduced a hybrid of PSO and GAs for the same purpose. Unfortunately, the accuracy result is still not high and many genes are selected for cancer classification since there are no direct probability relations between PSO and GAs.

Generally, the PSO-based methods [1],[8],[21],[33] are intractable to efficiently produce a small (near-optimal) subset of informative genes for high classification accuracy. This is mainly because the total number of genes in gene expression data is too large (high-dimensional data). Thus, in order to solve the problem, the present works proposes an improved (modified) binary PSO (IPSO) to select a small (near-optimal) subset of informative genes that is most relevant for the cancer classification. The small subset means that it contains the small number of selected genes.

8.2 The Conventional Version of Binary PSO (BPSO)

BPSO is initialized with a population of particles. At each iteration, all particles move in a problem space to find the optimal solution. A particle represents a potential solution in an *n*-dimensional space. Each particle has position and velocity vectors for directing its movement. The position vector and velocity vector of the *i*th particle in the *n*-dimension can be represented as $X_i = (x_i^1, x_i^2, ..., x_i^n)$ and $V_i = (v_i^1, v_i^2, ..., v_i^n)$, respectively, where $x_i^d \in \{0,1\}$; $i = 1, 2, \ldots, m$ (*m* is the total number of particles); and $d = 1, 2, \ldots, n$ (*n* is the dimension of data) [15]. v_i^d is a real number for the *d*-th dimension of the particle *i*, where the maximum v_i^d , $V_{\max} = (1/3) \times n.$

In gene selection, the vector of particle positions is represented by a binary bit string of length *n*, where *n* is the total number of genes. Each position vector (X_i) denotes a gene subset. If the value of the bit is 1, it means that the corresponding gene is selected. Otherwise, the value of 0 means that the corresponding gene is not selected. Each particle in the *t*-th iteration updates its own position and velocity according to the following equations:

$$v_i^d(t+1) = w(t) \times v_i^d(t) + c_1 r_1^d(t) \times (pbest_i^d(t) - x_i^d(t)) + c_2 r_2^d(t) \times (gbest^d(t) - x_i^d(t)) \qquad (8.1)$$

$$Sig(v_i^d(t+1)) = \frac{1}{1 + e^{-v_i^d(t+1)}} \qquad (8.2)$$

$$\text{if } Sig(v_i^d(t+1)) > r_3^d(t), \text{ then } x_i^d(t+1) = 1; \text{ else } x_i^d(t+1) = 0 \qquad (8.3)$$

where c_1 and c_2 are the acceleration constants in the interval [0,2]. $r_1^d(t), r_2^d(t), r_3^d(t) \sim U(0,1)$ are random values in the range [0,1] that sampled from a uniform distribution. $Pbest_i(t) = (pbest_i^1(t), pbest_i^2(t), ..., pbest_i^n(t))$ represents the best previous position of the ith particle, whereas $Gbest(t) = (gbest^1(t), gbest^2(t), ..., gbest^n(t))$ denotes the global best position of the swarm (all particles), respectively. They are assessed base on a fitness function. $Sig(v_i^d(t+1))$ is a sigmoid function where $Sig(v_i^d(t+1)) \in [0,1]$. $w(t)$ is an inertia weight and initialized with 1.4. It is updated as follows:

$$w(t+1) = \frac{(w(t) - 0.4) \times (MAXITER - Iter(t))}{(MAXITER + 0.4)} \tag{8.4}$$

where $MAXITER$ is the maximum iteration (generation) and $Iter(t)$ is the current iteration. Figure 8.1 shows the flowchart of BPSO.

8.2.1 Investigating the drawbacks of BPSO and previous PSO-based methods

Before attempting to propose IPSO, it would be prudent to find the limitations of BPSO and previous PSO-based methods [1],[8],[21],[33]. This subsection investigates theoretically the limitations by analyzing Eq.(8.2) and Eq.(8.3). These equations are analyzed because they are most important equations for genes selection in binary spaces. Both the equations are also implemented in BPSO and the PSO-based methods.

The sigmoid function (Eq.(8.2)) represents a probability for $x_i^d(t)$ to be 0 or 1 ($P(x_i^d(t) = 0)$ or $P(x_i^d(t) = 1)$). For example,

if $v_i^d(t) = 0,$ then $Sig(v_i^d(t) = 0) = 0.5$ and $P(x_i^d(t) = 0) = 0.5.$
if $v_i^d(t) < 0,$ then $Sig(v_i^d(t) < 0) < 0.5$ and $P(x_i^d(t) = 0) > 0.5.$
if $v_i^d(t) > 0,$ then $Sig(v_i^d(t) > 0) > 0.5$ and $P(x_i^d(t) = 0) < 0.5.$

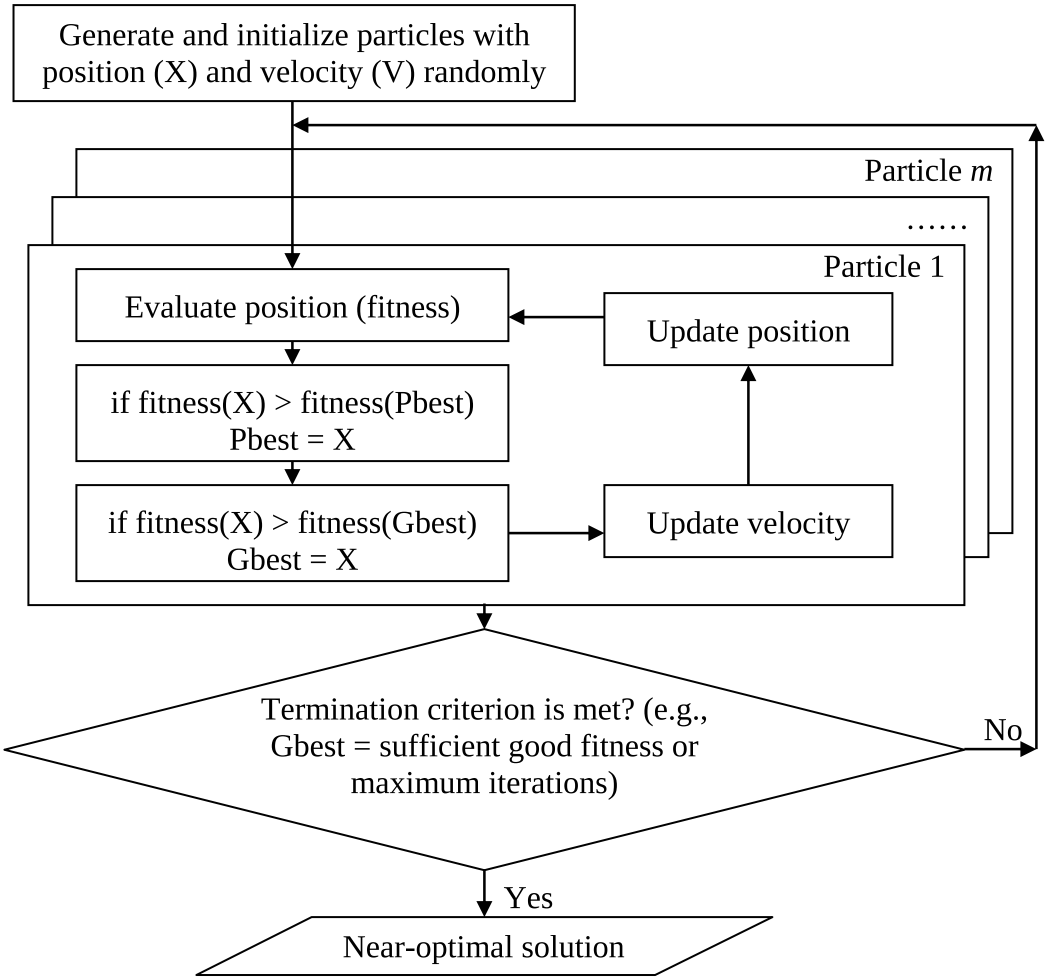

Generate and initialize particles with position (X) and velocity (V) randomly
Particle *m*
……
Particle 1
Evaluate position (fitness)
Update position
if fitness(X) > fitness(Pbest)
Pbest = X
if fitness(X) > fitness(Gbest)
Gbest = X
Update velocity
Termination criterion is met? (e.g., Gbest = sufficient good fitness or maximum iterations)
No
Yes
Near-optimal solution

Fig. 8.1. The flowchart of BPSO.

Also note that $P(x_i^d(t)=0)=1-P(x_i^d(t)=1)$. From the analysis, it concludes that $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$ because Eq.(8.2) is a standard sigmoid function without any constraint and no modification. Hence, by using this standard sigmoid function in high-dimensional spaces (gene expression data), it only reduces the number of genes to about half of the total number of genes. This is reported and proved in the section of experimental results. Therefore, Eq.(8.2) and Eq.(8.3) are potentially being the drawbacks of BPSO and the previous PSO-based methods in selecting a small number of genes for producing a near-optimal (small) subset of genes from gene expression data.

8.3 A Modification of Binary PSO with Introduced Particle's Speed and a Novel Rule (IPSO)

Almost all previous works of gene expression data researches have selected a subset of genes to obtain excellent cancer classification. Therefore, this chapter proposes IPSO for selecting a near-optimal (small) subset of genes. It is proposed to overcome the limitations of BPSO and previous PSO-based methods [1],[8],[21],[33]. IPSO in the present work differs from BPSO and the PSO-based methods on two parts; 1) introduce a scalar quantity that called particle’s speed (s); and 2) propose a rule for updating $x_i^d(t+1)$, whereas BPSO and the PSO-based methods have used the original rule (Eq.(8.3)) and no particle’s speed implementation. The particle’s speed and rule are introduced in order to

- increase the probability of $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$ and
- reduce the probability of $x_i^d(t+1)=1$ $(P(x_i^d(t+1)=1))$.

The increased and decreased probability values cause a small number of genes are selected and grouped into a gene subset. $x_i^d(t+1)=1$ means that the corresponding gene is selected. Otherwise, $x_i^d(t+1)=0$ represents that the corresponding gene is not selected.

Definition 1. s_i is a speed or length or magnitude of V_i for the particle *i*. In a real Euclidean space $\Re^n$, where $\Re$ denotes the field of real numbers, and *n* is the dimension of $\Re$, s_i can be derived by the Euclidean norm as follows:

$$s_i = \|V_i\| = \sqrt{\left(v_i^1\right)^2 + \left(v_i^2\right)^2 + \ldots + \left(v_i^n\right)^2} \, . \tag{8.5}$$

Therefore, the following properties of s_i are crucial;

- non-negativity: $s_i \geq 0$,
- definiteness: $s_i = 0$ if and only if $V_i = 0$,

- homogeneity: $\|\alpha V_i\| = \alpha \|V_i\| = \alpha s_i$ where $\alpha \geq 0$ and
- the triangle inequality: $\|V_i + V_{i+1}\| \leq \|V_i\| + \|V_{i+1}\|$ where $\|V_i\| = s_i$ and $\|V_{i+1}\| = s_{i+1}$.

The particle's speed and the rule for IPSO in binary spaces are proposed as follows:

$$s_i(t+1) = w(t) \times s_i(t) + c_1 r_1(t) \times dist(Pbest_i(t) - X_i(t)) + c_2 r_2(t) \times dist(Gbest(t) - X_i(t)) \tag{8.6}$$

$$Sig(s_i(t+1)) = \frac{1}{1+e^{-s_i(t+1)}} \quad \text{subject to } s_i(t+1) \geq 0 \tag{8.7}$$

$$\text{if } Sig(s_i(t+1)) > r_3^d(t), \text{ then } x_i^d(t+1) = 0; \text{ else } x_i^d(t+1) = 1 \tag{8.8}$$

where $s_i(t+1)$ represents the speed of the particle *i* for the *t*+1 iteration, whereas in BPSO and previous PSO-based methods (Eqs.(8.1), (8.2), and (8.3)), $v_i^d(t+1)$ represents a single element of a particle velocity vector for the particle *i*. In IPSO, Eqs.(8.6), (8.7), and (8.8) are used to replace Eqs.(8.1), (8.2), and (8.3), respectively. $s_i(t+1)$ is the rate at which the particle *i* changes its position. Based on Definition 1, the most important property of $s_i(t+1)$ is $s_i(t+1) \geq 0.$ Hence, $s_i(t+1)$ is used instead of $v_i^d(t+1)$ so that its positive value can increase $P(x_i^d(t+1) = 0).$

In Eq.(6), the calculation for updating $s_i(t+1)$ is mainly based on the distance between $Pbest_i(t)$ and $X_i(t)$ $(dist(Pbest_i(t) - X_i(t))),$ and the distance between $Gbest(t)$ and $X_i(t)$ $(dist(Gbest(t) - X_i(t))),$ whereas the original formula (Eq.(8.1)) is used to calculate $v_i^d(t+1)$ and it is essentially based on the difference between $Pbest_i^d(t)$ and $x_i^d(t),$ and the difference between $Gbest^d(t)$ and $x_i^d(t).$ The distances are used in the calculation for updating $s_i(t+1)$ in order to always satisfy the property of $s_i(t+1),$ namely $(s_i(t+1) \geq 0)$ and finally increase $P(x_i^d(t+1) = 0).$ Subsection 8.3.1 explains how to calculate the distance between two positions of two particles, e.g., $dist(Gbest(t) - X_i(t)).$

Equations (8.6), (8.7), and (8.8) and $s_i(t) \geq 0$ increase $P(x_i^d(t)=0)$ because the minimum value for $P(x_i^d(t)=0)$ is 0.5 when $s_i(t)=0$ $(\min P(x_i^d(t)=0) \geq 0.5).$ Meanwhile, they decrease the maximum value for $P(x_i^d(t)=1)$ to 0.5 $(\max P(x_i^d(t)=1) \leq 0.5).$ Therefore, if $s_i(t) > 0,$ then $P(x_i^d(t)=0) >> 0.5$ and $P(x_i^d(t)=1) << 0.5.$

Figure 8.2(a) shows that Eqs.(8.6), (8.7), and (8.8) and $s_i(t) \geq 0$ in IPSO increase $P(x_i^d(t)=0)$; whereas Fig. 8.2(b) denotes that Eqs.(8.1), (8.2), and (8.3) in BPSO yield $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5.$ For example, the calculations for $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ in Fig. 8.2(a) are shown as follows:

if $s_i(t)=1,$ then $P(x_i^d(t)=0)=0.73$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.27,$

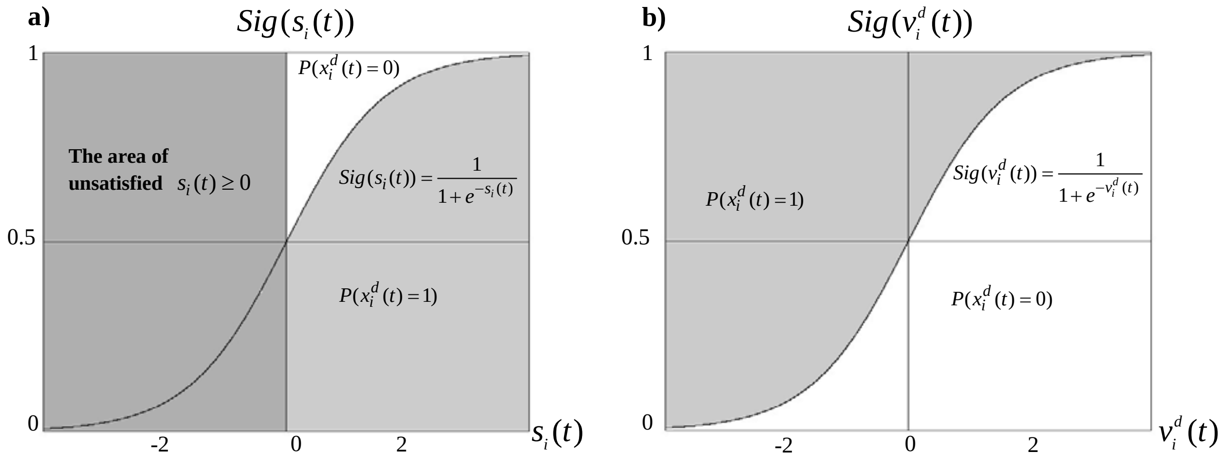
and

if $s_i(t)=2,$ then $P(x_i^d(t)=0)=0.88$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.12.$

This high probability of $x_i^d(t)=0$ $(P(x_i^d(t)=0))$ causes a small number of genes are selected in order to produce a near-optimal (small) gene subset from high-dimensional data (gene expression data). Hence, IPSO is proposed to overcome the limitations of BPSO and the previous PSO-based methods, and finally produce a small subset of informative genes.

8.3.1 The calculation of the distance of two particles' positions

The number of different bits between two particles relates to the difference between their positions. For example, $Gbest(t)=[0011101000]$ and $X_i(t)=[1110110100].$ The difference between $Gbest(t)$ and $X_i(t)$ is $[-1-1010-11-100].$ The value of 1 indicates that compared with the best position, this bit (gene) should be selected, but it is not selected, which may decrease classification quality and lead to a lower fitness value. In contrast, a value of -1 indicates that, compared with the best position, this bit should not be selected, but it is selected. The selection of irrelevant genes makes the length of the subset longer and leads to a lower fitness value. Assume that the number of 1 is $a,$ whereas the number of -1 is $b.$ The present work uses the absolute value of $a-b$ ($|a-b|$) to express the distance between two positions. In this example, the distance between $Gbest(t)$ and $X_i(t)$ is $dist(Gbest(t)-X_i(t))=|a-b|=|2-4|=2.$



Legends:

▩	The area of unsatisfied $s_i(t)\geq 0.$
□	The area of $P(x_i^d(t)=0).$
░	The area of $P(x_i^d(t)=1).$

Fig. 8.2. The areas of $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ based on sigmoid functions in (a) IPSO; (b) BPSO.

8.3.2 Fitness functions

The fitness value of particles (gene subsets) is calculated as follows:

$$fitness(X_i)=w_1\times A(X_i)+(w_2\times(n-R(X_i))/n) \tag{8.9}$$

where $A(X_i)\in[0,1]$ is LOOCV classification accuracy that uses the only genes in a gene subset (X_i). This accuracy is provided by SVM. $R(X_i)$ is the number of selected genes in X_i. n is the total number of genes for each sample. w_1 and w_2 are two priority weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$.

8.4 Experiments

8.4.1 Data sets and experimental setup

There are 12 gene expression data used in this chapter for evaluating IPSO and BPSO. The 10 data sets are summarized in Table 8.1. They included binary-classes and multi-classes data sets and were downloaded from http://www.gems-system.org. The other two data sets such as the colon and lung data sets have been described on Table 2.1 in Chapter 2.

Table 8.1. The description of 10 gene expression data sets.

Data sets	No. samples	No. genes	No. classes
11_Tumors	174	12,533	11
9_Tumors	60	5,726	9
Brain_Tumor1	90	5,920	5
Brain_Tumor2	50	10,367	4
Leukemia1	72	5,327	3
Leukemia2	72	11,225	3
Lung_Cancer	203	12,600	5
SRBCT	83	2,308	4
Prostate_Tumor	102	10,509	2
DLBCL	77	5,469	2

Note:
SRBCT = small round blue cell tumor.
DLBCL = diffuse large B-cell lymphomas.

All experimental results reported in this chapter are experimented in Rocks Linux version 4.2.1 (Cydonia) on the IBM xSeries 335 (cluster node) that contains 13 compute-nodes. Each compute-node has four high performances 3.0GHz Intel Xeon CPUs with 512MB of memories. Thus, the total number of CPUs for the 13 compute-nodes is 52. In order to make sure the running time of every run using the same capacity of CPUs usages, each run has been independently experimented on only one CPU. This situation is important because the comparison of running times between IPSO and BPSO is conducted for evaluation of their performances.

Table 8.2. Parameter settings for IPSO and BPSO.

Parameters	Values
No. particles	100
No. iterations	300
w_1	0.8
w_2	0.2
c_1	2
c_2	2

Experimental results that produced by IPSO are compared with an experimental method (BPSO) and other previous PSO-based methods for objective comparisons [1],[8],[21],[33]. Firstly, the present work applied the gain ratio technique for preprocessing in order to preselect 500-top-ranked genes. These genes are then used by IPSO and BPSO. Next, SVM is used to measure LOOCV accuracy on gene subsets that produced by IPSO and BPSO. In order to avoid selection bias, the implementation of LOOCV is in exactly the same way as Chuang *et al*. [8] where the only one cross-validation cycle (outer loop), namely LOOCV is used. Moreover, the present work uses LOOCV accuracy for comparisons because the previous related works also applied it to measure classification accuracy on the same data sets. Several experiments are independently conducted 10 times on each data set using IPSO and BPSO. Next, an average result of the 10 independent runs is obtained. Two criteria following their importance are considered to evaluate the performances of IPSO and BPSO such as

LOOCV accuracy and the number of selected genes. Additionally, running times are also measured for the comparison between IPSO and BPSO. High accuracy, the small number of selected genes, and low running time are needed to obtain an excellent performance. Table 8.2 contains parameter values for IPSO and BPSO. These values are chosen based on the results of preliminary runs.

Table 8.3. Experimental results for each run using IPSO on the 11_Tumors, 9_Tumors, Brain_Tumor1, and Brain_Tumor2 data sets.

Run no.	11_Tumors		9_Tumors		Brain_Tumor1		Brain_Tumor2	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	95.40	239	73.33	226	92.22	9	92	11
2	94.83	254	73.33	238	92.22	21	92	27
3	94.83	240	75	237	93.33	8	92	5
4	95.40	245	75	240	92.22	6	92	5
5	94.83	230	76.67	255	92.22	27	92	4
6	94.83	232	78.33	248	92.22	10	92	16
7	95.40	251	75	235	92.22	6	94	4
8	94.83	237	76.67	247	93.33	5	90	5
9	95.40	228	76.67	240	93.33	11	92	6
10	94.83	253	75	240	92.33	9	92	7
Average ± S.D.	95.06 ± 0.30	240.90 ± 9.55	75.50 ± 1.58	240.60 ± 7.95	92.56 ± 0.54	11.20 ± 7.15	91.00 ± .05	6.40 ± 1.90

Note: The results of the best subsets are shown in the shaded cells. A near-optimal subset that produces the highest classification accuracy with the smallest number of genes is selected as the best subset. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively.

8.4.2 Experimental results and discussion

Based on the standard deviations in Tables 8.3, 8.4, and 8.5, results that produced by IPSO were almost consistent on all data sets. Interestingly, all runs have achieved 100% LOOCV accuracy with less than 30 selected genes on the Leukemia1, Leukemia2, SRBCT, DLBCL, and lung data sets. Moreover, over 92% classification accuracies have been obtained on other data sets, except for the 9_Tumors data set. This means that IPSO has efficiently selected and produced a near-optimal gene subset from high-dimensional data (gene expression data).

Figure 8.3 shows that the averages of fitness values of IPSO increase dramatically after a few generations on all the data sets. A high fitness value is obtained by a combination between a high classification rate and a small number (subset) of selected genes. The condition of the proposed particle's speed that should always be positive real numbers started in the initialization method, and the new rule for updating particle's positions provoke the early convergence of IPSO. In contrast, the averages of fitness values of BPSO was no improvement until the last generation due to $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$.

Table 8.4. Experimental results for each run using IPSO on the Leukemia1, Leukemia2, Lung_Cancer, and SRBCT data sets.

Run no.	Leukemia1		Leukemia2		Lung_Cancer		SRBCT	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	100	4	100	7	95.07	9	100	10
2	100	2	100	6	95.57	27	100	22
3	100	4	100	7	96.55	10	100	25
4	100	4	100	6	95.57	5	100	8
5	100	3	100	8	95.57	12	100	28
6	100	4	100	4	96.06	6	100	12
7	100	4	100	5	95.57	7	100	14
8	100	3	100	7	96.55	28	100	26
9	100	4	100	8	96.55	34	100	24
10	100	3	100	9	95.57	11	100	6
Average ± S.D.	100 ± 0	3.50 ± 0.71	100 ± 0	6.70 ± 1.50	95.86 ± 0.53	14.90 ± 10.57	100 ± 0	17.50 ± 8.32

Table 8.5. Experimental results for each run using IPSO on the Prostate_Tumor, DLBCL, colon, and lung data sets.

Run no.	Prostate_Tumor		DLBCL		Colon		Lung	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	98.04	18	100	7	93.55	5	100	9
2	97.06	6	100	8	93.55	5	100	6
3	98.04	17	100	4	96.77	4	100	6
4	98.04	26	100	6	93.55	5	100	5
5	98.04	9	100	5	93.55	4	100	6
6	98.04	11	100	5	95.16	5	100	8
7	98.04	8	100	6	93.55	4	100	4
8	98.04	7	100	7	95.16	4	100	5
9	98.04	26	100	5	93.55	5	100	7
10	98.04	8	100	7	93.55	4	100	6
Average ± S.D.	97.94 ± 0.31	13.60 ± 7.68	100 ± 0	6.00 ± 1.25	94.19 ± 1.13	4.5 ± 0.53	100 ± 0	6.20 ± 1.48

According to the Table 8.6, overall, it is worthwhile to mention that the classification accuracy and the number of selected genes of IPSO are superior to BPSO in terms of the best, average, and standard deviation results on all the data sets except for the Lung_Cancer and Prostate_Tumor data sets. The classification accuracy of BPSO on both the data sets were slightly higher than IPSO. This is probably because both the data sets need many genes for more accurate classification of cancer classes. Meanwhile, IPSO produces a smaller number of genes compared to BPSO. The running times of IPSO are also lower than BPSO in all the data sets. IPSO can reduce its running times because of the following reasons:

- IPSO selects the smaller number of genes compared to BPSO,
- The computation of SVMs is fast because it uses the small number of features (genes) that selected by IPSO for classification process, and
- IPSO only uses the speed of particles for comparing with $r_3^d(t)$, whereas BPSO practices all elements of a particle’s velocity vectors for the comparison.

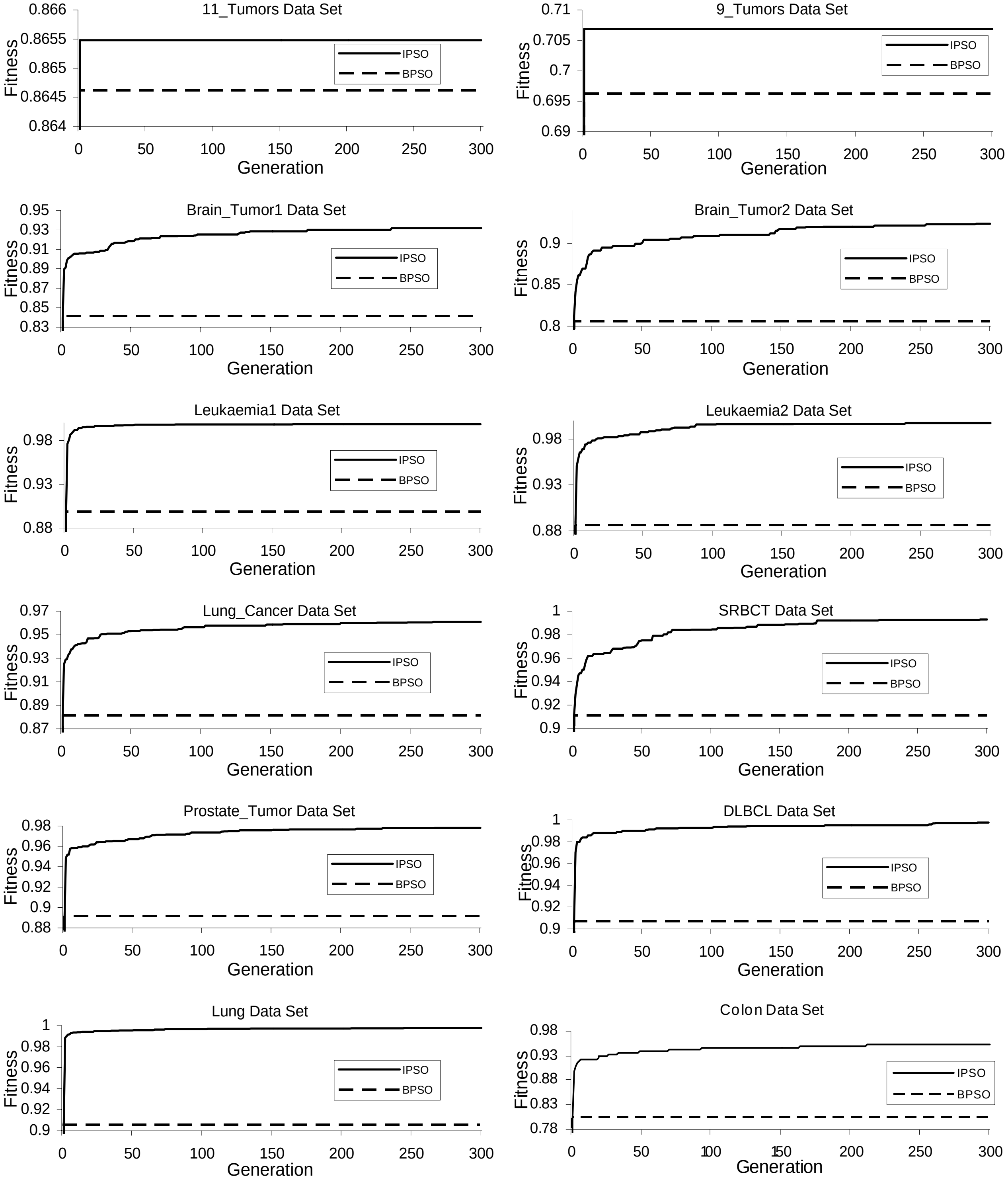


Fig. 8.3. A relation between the average of fitness values (10 runs on average) and the number of generations for IPSO and BPSO.

Table 8.6. Comparative experimental results of IPSO and BPSO.

Data set	Method / Evaluation	IPSO			BPSO		
		Best	#Ave	S.D	Best	#Ave	S.D
11_Tumors	#Acc (%)	95.40	95.06	0.30	95.98	94.94	0.85
	#Genes	228	240.9	9.55	245	241.10	12.80
	#Time	56.40	57.00	0.37	409.71	409.93	0.23
9_Tumors	#Acc (%)	78.33	75.50	1.58	78.33	73.33	1.92
	#Genes	248	240.6	7.95	244	236.00	12.38
	#Time	3.02	3.34	0.17	31.36	31.57	0.12
Brain_Tumor1	#Acc (%)	93.33	92.56	0.54	92.22	92.00	0.47
	#Genes	5	11.20	7.15	220	236.30	11.94
	#Time	10.63	12.08	0.88	46.65	46.77	0.10
Brain_Tumor2	#Acc (%)	94.00	92.00	0.94	90	88.20	0.63
	#Genes	4	9.10	7.34	251	245.30	11.30
	#Time	0.62	0.66	0.03	10.58	10.60	0.02
Leukemia1	#Acc (%)	100	100	0	98.61	98.61	0
	#Genes	2	3.50	0.71	216	224.70	5.23
	#Time	2.28	2.31	0.02	13.86	13.94	0.03
Leukemia2	#Acc (%)	100	100	0	97.22	97.22	0
	#Genes	4	6.70	1.50	218	228.11	4.86
	#Time	2.24	2.72	0.25	19.37	19.90	0.35
Lung_Cancer	#Acc (%)	96.55	95.86	0.53	97.54	96.60	0.63
	#Genes	10	14.90	10.57	245	228.70	9.70
	#Time	90.34	96.24	6.64	282.75	285.33	1.34
SRBCT	#Acc (%)	100	100	0	100	100	0
	#Genes	6	17.50	8.32	206	221.30	7.35
	#Time	5.52	5.96	0.39	44.86	44.88	0.01
Prostate_Tumor	#Acc (%)	98.04	97.94	0.31	98.04	98.04	0
	#Genes	7	13.60	7.68	217	231.50	8.40
	#Time	3.59	3.64	0.03	48.11	48.61	0.26
DLBCL	#Acc (%)	100	100	0	100	100	0
	#Genes	4	6	1.25	215	230.10	10.09
	#Time	1.60	1.62	0.02	11.21	12.49	1.11
Colon	#Acc (%)	96.77	94.19	1.13	87.10	86.94	0.51
	#Genes	4	4.50	0.53	214	231	10.19
	#Time	4.22	4.33	0.06	30.58	30.65	0.27
Lung	#Acc (%)	100	100	0	99.45	99.39	0.18
	#Genes	4	6.20	1.48	219	223.33	4.24
	#Time	8.22	8.31	0.05	110.71	111.07	0.23

Note: The best results of each data set are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively, whereas #Genes and #Ave represent the number of selected genes and an average, respectively. #Time stands for running time in the hour unit.

Table 8.7. A comparison between IPSO and previous PSO-based methods.

Data set	Method Evaluation	IPSO	IBPSO [8]	PSOTS [31]	PSOGA [21]	GPSO [1]
11_Tumors	#Acc (%)	(95.06)	93.10	-	-	-
	#Genes	(240.9)	2948	-	-	-
9_Tumors	#Acc (%)	(75.50)	78.33	-	-	-
	#Genes	(240.6)	1280	-	-	-
Brain_Tumor1	#Acc (%)	(92.56)	94.44	-	-	-
	#Genes	(11.20)	754	-	-	-
Brain_Tumor2	#Acc (%)	(92.00)	94.00	-	-	-
	#Genes	(9.10)	1197	-	-	-
Leukemia1	#Acc (%)	(100)	100	(98.61)	(95.10)	-
	#Genes	(3.50)	1034	(7)	(21)	-
Leukemia2	#Acc (%)	(100)	100	-	-	-
	#Genes	(6.70)	1292	-	-	-
Lung_Cancer	#Acc (%)	(95.86)	96.55	-	-	-
	#Genes	(14.90)	1897	-	-	-
SRBCT	#Acc (%)	(100)	100	-	-	-
	#Genes	(17.50)	431	-	-	-
Prostate_Tumor	#Acc (%)	(97.94)	92.16	-	-	-
	#Genes	(13.60)	1294	-	-	-
DLBCL	#Acc (%)	(100)	100	-	-	-
	#Genes	(6)	1042	-	-	-
Colon	#Acc (%)	(94.19)	-	(93.55)	(88.7)	-
	#Genes	(4.50)	-	(8)	(16.8)	-
Lung	#Acc (%)	(100)	-	-	-	(99)
	#Genes	(6.20)	-	-	-	(4)

Note: The results of the best subsets are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. '-' means that a result is not reported in the related previous work. A result in '( )' denotes an average result. #Genes and #Acc represent the number of selected genes and the classification accuracy, respectively.

IBPSO = Improved binary PSO.
PSOGA = A hybrid of PSO and GAs.
PSOTS = A hybrid of PSO and tabu search.
GPSO = Geometric PSO.

For an objective comparison, the present work compares IPSO with previous related works that used PSO-based methods in their proposed methods [1],[8],[21],[33]. It is shown in Table 8.7. For all the data sets, the averages of the number of selected genes of the present work were smaller than the previous works [1],[8],[21],[33]. The present work also have resulted the higher averages of classification accuracies on seven data sets (11_Tumors, Leukemia1, Leukemia2, SRBCT, Prostate_Tumor, DLBCL, and colon) compared to the previous works. However, the classification accuracies of Chuang *et al*. [8] were slightly higher than the present work on four data sets (9_Tumors, Brain_Tumor1, Brain_Tumor2, and Lung_Cancer).

Even though the previous work [8] achieved better classification accuracies on the four data sets, but they used high numbers of selected genes (at least 750 selected genes) to obtain the results. Moreover, they could not have statistically meaningful conclusions because their experimental results were obtained by only one independent run on each data set, and not based on average results. The average results are important since their proposed method is a stochastic approach. Additionally, in their approach, the global best particle's position is reset to zero position when its fitness values do not change after three successive iterations. Theoretically, their approach is almost impossible to result a near-optimal gene subset from high-dimensional spaces (high-dimension data) because the global best particle's position should make a new exploration and exploitation for searching the near-optimal solution after its position reset to zero. Overall, the present work has outperformed the previous related works in terms of LOOCV accuracy and the number of selected genes. Running times between IPSO and the previous works cannot be compared because they were not reported in their articles.

According to Fig. 8.3 and Tables 8.3, 8.4, 8.5, 8.6, and 8.7, IPSO is reliable for gene selection since it has produced the near-optimal solution from gene expression data. This is due to the proposed particle's speed and the introduced rule increase the probability $x_i^d(t+1)=0 \ \ (P(x_i^d(t+1)=0))$. The particle's speed is introduced to provide the rate at which a particle changes its position, whereas the rule is proposed to update particle's positions. The increased probability value for $x_i^d(t+1)=0$ causes the selection of a small number of informative genes and finally produces a near-optimal subset (a small subset of informative genes with high classification accuracy) for cancer classification.

8.5 Summary

In this chapter, IPSO has been proposed for gene selection on 12 gene expression data sets. Overall, based on the experimental results, the performance of IPSO was superior to BPSO and PSO-based methods that proposed by previous related works in terms of classification accuracy and the number of selected genes. IPSO was excellent because the probability $x_i^d(t+1)=0$ has been increased by the proposed particle's speed and the introduced rule. The particle's speed and the introduced rule have been proposed in order to yield a near-optimal subset of genes for better cancer classification. IPSO also obtains lower running times because it selects the small number of genes compared to BPSO. The next chapter (Chapter 9) will propose a constraint approach in BPSO in order to increase the probability $x_i^d(t+1)=0$.

Chapter 9

Enhanced Binary Particle Swarm Optimization with the Constraint of Particle's Velocities

9.1 Introduction

Chapter 9 also focuses on how to solve the weaknesses of BPSO and the previous related works [1],[21],[33] as described in Subsection 8.2.1 in Chapter 8. Thus, this chapter proposes and discusses an enhancement of binary PSO with the proposed constraint of particle's velocities (CPSO). The constraint is introduced in CPSO to increase the probability of genes to be unselected for the classification. It is evaluated by using five gene expression data sets.

9.2 An Enhancement of Binary Particle Swarm Optimization Based on the Constraint of Particle's Velocities (CPSO)

This chapter introduces an enhancement of binary PSO with the proposed constraint of particle's velocities (CPSO) to select a small (near-optimal) subset of informative genes that is most relevant for the cancer classification. It is also proposed to overcome the limitations of BPSO and the previous PSO-based methods [1], [21], [33]. CPSO in the present work differs from BPSO and the PSO-based methods on one part; it proposes the constraint of elements of particle velocity vectors; whereas BPSO and the previous PSO-based methods have not implemented any constraint for the elements of particle velocity vectors. CPSO also applies the proposed rule as described in Chapter 8 for updating particle's velocities. The constraint and rule are sequentially implemented to

- increase the probability of $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$ and
- reduce the probability of $x_i^d(t+1)=1$ $(P(x_i^d(t+1)=1))$.

The increased and decreased probability values cause a small number of genes are selected and grouped into a gene subset. $x_i^d(t+1)=1$ means that the corresponding gene is selected. Otherwise, $x_i^d(t+1)=0$ represents that the corresponding gene is not selected. The constraint of elements of particle velocity vectors is proposed as follows:

$$Sig(v_i^d(t+1))=\frac{1}{1+e^{-v_i^d(t+1)}} \qquad (9.1)$$
$$\text{subject to } v_i^d(t+1)\geq 0$$

where v_i^d is a real number for the *d*-th dimension of the particle *i* in the *t*+1 iteration with the maximum v_i^d, $V_{\max}=(1/3)\times n.$ $Sig(v_i^d(t+1))$ is a sigmoid function where $Sig(v_i^d(t+1))\in[0,1].$

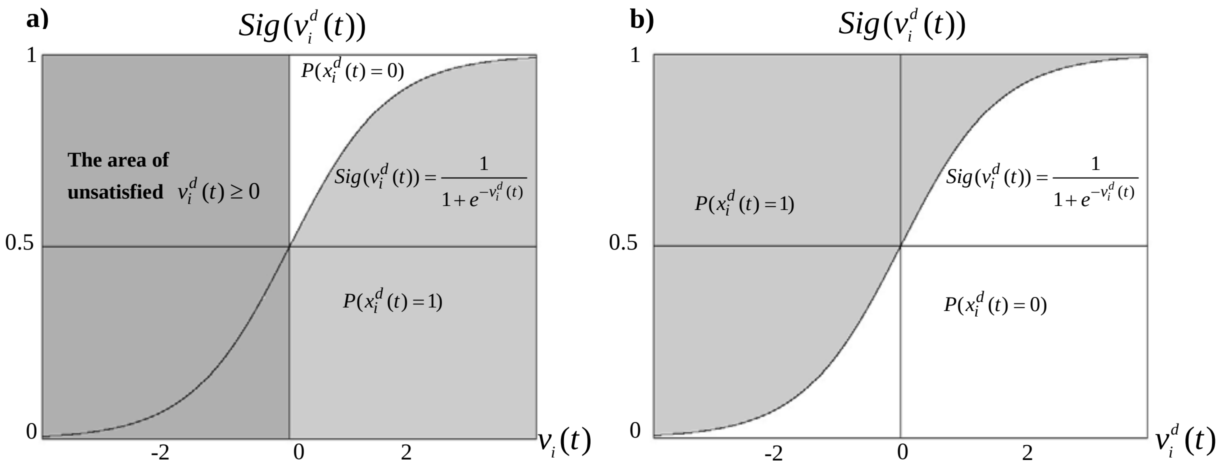
The constraint of elements of particle velocity vectors and the rule increase $P(x_i^d(t)=0)$ because the minimum value for $P(x_i^d(t)=0)$ is 0.5 when $v_i^d(t)=0$ $(\min P(x_i^d(t)=0)\geq 0.5).$ Meanwhile, they decrease the maximum value for $P(x_i^d(t)=1)$ to 0.5 $(\max P(x_i^d(t)=1)\leq 0.5).$ Therefore, if $v_i^d(t)>0,$ then $P(x_i^d(t)=0)>>0.5$ and $P(x_i^d(t)=1)<<0.5.$

Figure 9.1(a) shows that the constraint of elements of particle velocity vectors and the rule in CPSO increase $P(x_i^d(t)=0)$; whereas Fig.9.1(b) displays that Eqs.(8.1), (8.2), and (8.3) in BPSO as stated in Chapter 8 yield $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5.$ For example, the calculations for $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ in Fig. 9.1(a) are shown as follows;

if $v_i^d(t)=1,$ then $P(x_i^d(t)=0)=0.73$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.27,$

and

if $v_i^d(t)=2,$ then $P(x_i^d(t)=0)=0.88$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.12.$



Legends:

▩	The area of unsatisfied $v_i^d(t) \geq 0$.
□	The area of $P(x_i^d(t)=0)$.
▒	The area of $P(x_i^d(t)=1)$.

Fig. 9.1. The areas of $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ based on sigmoid functions in (a) CPSO; (b) BPSO.

The fitness value of particles (gene subsets) is calculated as follows:

$$fitness(X_i) = w_1 \times A(X_i) + (w_2 \times (n - R(X_i))/n) \tag{9.2}$$

where $A(X_i) \in [0,1]$ is the LOOCV classification accuracy that uses the only genes in a gene subset (X_i). This accuracy is provided by SVM. $R(X_i)$ is the number of selected genes in X_i. n is the total number of genes for each sample. w_1 and w_2 are two priority weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$.

9.3 Experimental Results

9.3.1 Data sets and experimental setup

Five gene expression data sets are used in this chapter to test the effectiveness of CPSO compared to BPSO and PSO-based methods from the previous related works [1],[21],[33]. These data sets are the leukemia, colon, lung, and mixed-lineage leukemia (MLL), and small round blue cell tumors (SRBCT) data sets. The first four data sets are summarized on Table 2.1 in Chapter 2. The SRBCT data set is a multi-classes data set. It has four classes; ewing family of tumors (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and burkitt lymphomas (BL). The training set contains 63 samples (22 EWS, 20 RMS, 12 NB, and 8 BL), whereas the test set contains 20 samples (6 EWS, 5 RMS, 6 NB, and 3 BL). There are 2,308 genes in each sample. It can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/.

Table 9.1. Parameter settings for CPSO and BPSO.

Parameters	Values
No. particles	100
No. iterations	500
w_1	0.8
w_2	0.2
c_1	2
c_2	2

Firstly, the present work applied the gain ratio technique for pre-processing in order to pre-select 500-top-ranked genes. These genes are then used by CPSO and BPSO. Next, SVM is used to measure LOOCV accuracy on gene subsets produced by CPSO and BPSO. Several experiments are independently conducted 10 times on each data set using CPSO and BPSO. Next, an average result of the 10 independent runs is obtained. High LOOCV accuracy, the small number of selected genes, and low running time are needed to obtain an excellent performance. Table 9.1 contains parameter values for CPSO and BPSO. These values are chosen based on the results of preliminary runs.

9.3.2 Result analysis and discussion

Based on the standard deviation of classification accuracies in Tables 9.2 and 9.3, results that produced by CPSO were almost consistent on all data sets. Interestingly, all runs have achieved 100% LOOCV accuracy with less than 50 selected genes on the SRBCT data set. Moreover, over 97% classification accuracies have been obtained on other data sets, except for the colon data set. This means that CPSO has efficiently selected and produced a near-optimal gene subset from high-dimensional data (gene expression data).

Table 9.2. Experimental results for each run using CPSO on the leukemia, colon, and lung data sets.

Run no.	Leukemia data set		Colon data set		Lung data set	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	100	10	90.32	4	99.45	9
2	100	5	90.32	6	99.45	9
3	100	3	88.71	28	99.45	7
4	98.61	9	91.94	10	99.45	30
5	98.61	9	88.71	8	99.45	8
6	100	31	88.71	8	99.45	9
7	98.61	11	88.71	7	98.90	8
8	98.61	10	88.71	7	99.45	5
9	98.61	8	88.71	5	99.45	15
10	98.61	9	88.71	130	99.45	13
Average	99.17	10.50	89.36	21.30	99.39	11.30
± S.D.	± 0.72	± 7.61	± 1.13	± 38.80	± 0.15	± 7.17

Note: The results of the best subsets are shown in the shaded cells. A near-optimal subset that produces the highest classification accuracy with the smallest number of genes is selected as the best subset. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively.

Table 9.3. Experimental results for each run using CPSO on the MLL and SRBCT data sets.

Run no.	MLL data set		SRBCT data set	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	97.22	32	100	20
2	98.61	113	100	48
3	97.22	38	100	42
4	97.22	28	100	50
5	97.22	6	100	21
6	95.83	6	100	37
7	97.22	11	100	32
8	97.22	37	100	27
9	97.22	88	100	21
10	97.22	33	100	50
Average ± S.D.	97.22 ± 0.66	39.20 ± 35.04	100 ± 0	34.80 ± 12.30

Note: The result of the best subsets is shown in the shaded cells. A near-optimal subset that produces the highest classification accuracy with the smallest number of genes is selected as the best subset. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively.

Figure 9.2 shows that the averages of fitness values of CPSO increase dramatically after a few generations on all the data sets. A high fitness value is obtained by a combination between a high classification rate and a small number (subset) of selected genes. The condition of the proposed constraint of elements of particle velocity vectors that should always be positive real numbers started in the initialization method and the new rule for updating particle's positions provoke the early convergence of CPSO. In contrast, the averages of fitness values of BPSO was no improvement until the last generation due to $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$.

For an objective comparison, CPSO is compared with BPSO. According to the Table 9.4, overall, it is worthwhile to mention that the classification accuracy and the number of selected genes of CPSO are superior to BPSO in terms of the best, average, and standard deviation results on all the data sets. The classification accuracies of BPSO and CPSO were the same on the lung and SRBCT data sets. However, the number of selected genes of BPSO was higher than CPSO to achieve the same accuracy. CPSO also produces smaller numbers

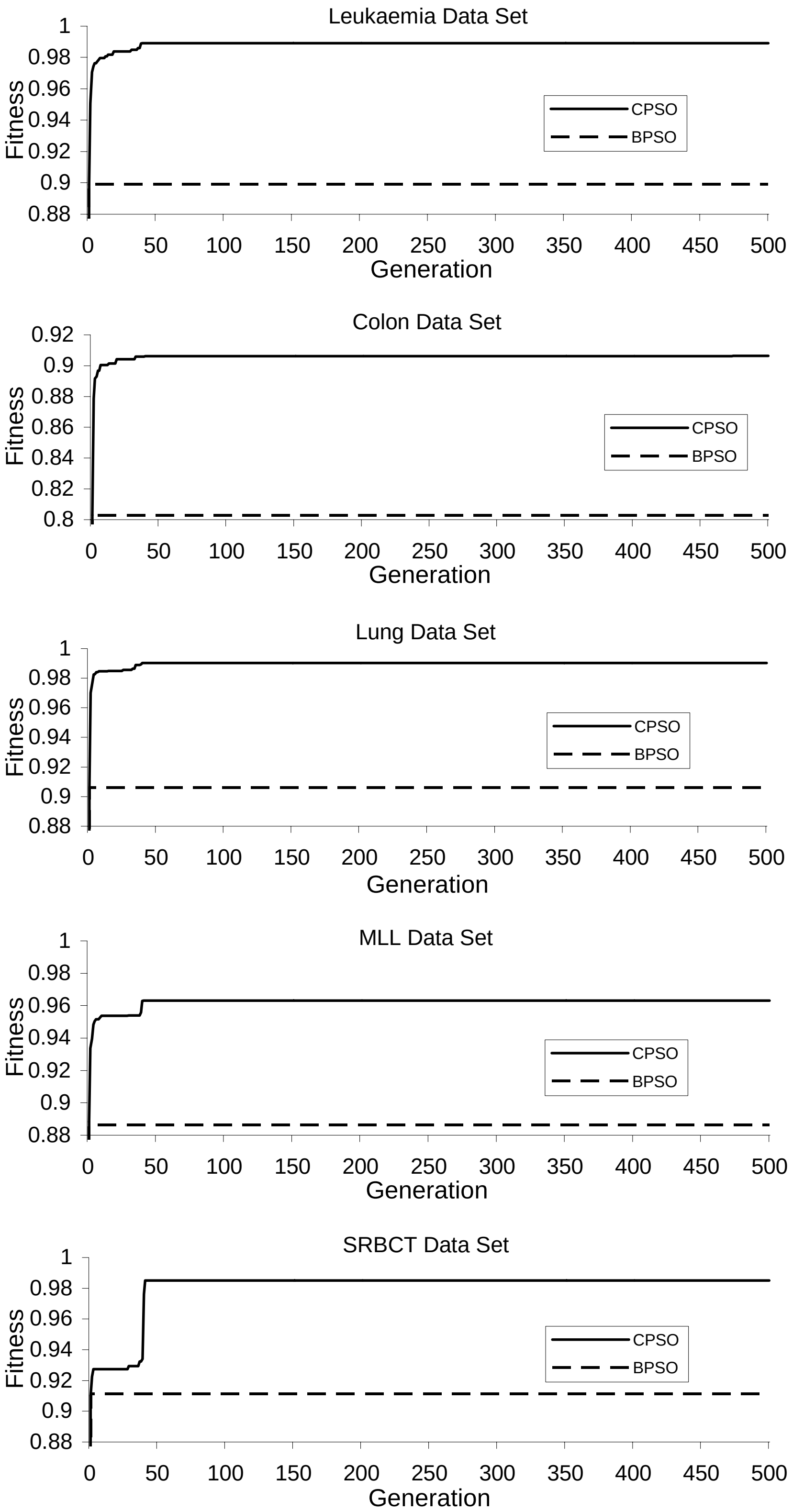


Fig. 9.2. A relation between the average of fitness values (10 runs on average) and the number of generations for CPSO and BPSO.

Table 9.4. Comparative experimental results of CPSO and BPSO.

Data set	Method / Evaluation	CPSO			BPSO		
		Best	#Ave	S.D	Best	#Ave	S.D
Leukemia	#Acc (%)	100	99.17	0.72	98.61	98.61	0
	No. selected genes	3	10.50	7.16	216	224.70	5.23
	Running time (hour)	5.26	6.13	1.44	13.86	13.94	0.03
Colon	#Acc (%)	91.94	89.36	1.13	87.10	86.94	0.51
	No. selected genes	10	21.30	38.80	214	231	10.19
	Running time (hour)	8.78	9.26	0.70	30.58	30.63	0.27
Lung	#Acc (%)	99.45	99.39	0.18	99.45	99.39	0.18
	No. selected genes	5	11.30	7.17	219	223.33	4.24
	Running time (hour)	63.53	64.40	0.87	110.71	111.07	0.23
MLL	#Acc (%)	98.61	97.22	0.66	97.22	97.22	0
	No. selected genes	113	39.20	35.04	218	228.11	4.86
	Running time (hour)	9.51	11.64	4.98	19.37	19.90	0.35
SRBCT	#Acc (%)	100	100	0	100	100	0
	No. selected genes	20	34.80	12.30	206	221.30	7.35
	Running time (hour)	21.67	21.76	1.32	44.86	44.88	0.01

Note: The best results of each data set are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively, whereas #Ave represents an average.

of genes and lower running times compared to BPSO on all the data sets. CPSO can reduce its running times because of the following reasons;

- CPSO selects the smaller number of genes compared to BPSO and
- The computation of SVMs is fast because it uses the small number of features (genes) that selected by CPSO for classification process.

The present work also compares CPSO with previous related works that used PSO-based methods in their proposed methods [1],[21],[33]. It is shown in Table 9.5. For all the data sets, the averages of the number of selected genes of the present work were smaller than the previous works. The present work also have resulted the higher averages of classification accuracies on the leukemia data set compared to the previous works. However, experimental results produced by Shen *et al*. [31] were better than the present work on the colon data sets.

Running time between CPSO and the previous works cannot be compared because it was not reported in their articles.

According to Fig. 9.2 and Tables 9.3, 9.4, and 9.5, CPSO is reliable for gene selection since it has produced the near-optimal solution from gene expression data. This is due to the proposed constraint of elements of particle velocity vectors and the introduced rule increase the probability $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$. The increased probability value for $x_i^d(t+1)=0$ causes the selection of a small number of informative genes and finally produces a near-optimal subset (a small subset of informative genes with high classification accuracy) for cancer classification.

Table 9.5. A comparison between CPSO and previous PSO-based methods.

Data set	Method Evaluation	CPSO	PSOTS [31]	PSOGA [21]	GPSO [1]
Leukemia	#Acc (%)	(99.17)	(98.61)	(95.10)	-
	No. selected genes	(10.50)	(7)	(21)	-
Colon	#Acc (%)	(89.36)	(93.55)	(88.7)	-
	No. selected genes	(21.30)	(8)	(16.8)	-
Lung	#Acc (%)	(99.39)	-	-	(99)
	No. selected genes	(11.30)	-	-	(4)
MLL	#Acc (%)	(97.22)	-	-	-
	No. selected genes	(39.20)	-	-	-
SRBCT	#Acc (%)	(100)	-	-	-
	No. selected genes	(34.80)	-	-	-

Note: The results of the best subsets are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. '-' means that a result is not reported in the related previous work. A result in '( )' denotes an average result. #Acc represents the classification accuracy.

PSOGA = A hybrid of PSO and GAs.

PSOTS = A hybrid of PSO and tabu search.

GPSO = Geometric PSO.

9.4 Summary

Overall, based on the experimental results, the performance of CPSO was superior to BPSO and previous PSO-based methods in terms of classification accuracy and the number of selected genes. CPSO was excellent because the probability $x_i^d(t+1)=0$ has been increased by the proposed constraint of elements of particle velocity vectors and the introduced rule. The constraint and rule have been proposed in order to yield a near-optimal subset of genes for better cancer classification. CPSO also obtains lower running times because it selects the small number of genes compared to BPSO. Chapter 10 will propose a modified sigmoid function to more increase the probability $x_i^d(t+1)=0$.

Chapter 10

Improved Binary Particle Swarm Optimization Based on a Modified Sigmoid Function

10.1 Introduction

This chapter proposes an improvement of binary particle swarm optimization. It is introduced to surmount the limitations of BPSO and the previous related works [1],[8],[21],[33]. The limitations have been described in Subsection 8.2.1 in Chapter 8. A simple modified sigmoid function is proposed in the improved BPSO. In order to test the effectiveness of the proposed method, the present work applies it to five gene expression data sets, including binary-classes and multi-classes data sets.

10.2 An Improvement of Binary Particle Swarm Optimization with a Modified Sigmoid Function (SPSO)

In order to overcome the limitations of BPSO and previous PSO-based [1],[8],[21],[33] for selecting a small subset of genes, this chapter proposes an improvement of BPSO based on a modified sigmoid function (SPSO). SPSO in the present work differs from BPSO and the PSO-based methods on one major part; the present work modifies the existing sigmoid function, whereas BPSO and the PSO-based methods have used the standard sigmoid function as shown on Eq.(8.2) in Chapter 8. Moreover, SPSO also implements the proposed rule and particle’s speed as introduced in Chapter 8. The modified sigmoid function, rule, and particle’s speed are consecutive applied to;

- increase the probability of $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$ and
- reduce the probability of $x_i^d(t+1)=1$ $(P(x_i^d(t+1)=1))$.

The increased and decreased probability values cause a small number of genes are selected and grouped into a gene subset. $x_i^d(t+1)=1$ means that the corresponding gene is selected. Otherwise, $x_i^d(t+1)=0$ represents that the corresponding gene is not selected. The modified sigmoid function is proposed as follows:

$$Sig(s_i(t+1))=\frac{1}{1+e^{-5s_i(t+1)}} \qquad \text{subject to } s_i(t+1)\geq 0 \tag{10.1}$$

where $s_i(t+1)$ represents the speed of the particle *i* for the *t*+1 iteration, whereas in BPSO and previous PSO-based methods, $v_i^d(t+1)$ represents a single element of a particle velocity vector for the particle *i*. $s_i(t+1)$ is the rate at which the particle *i* changes its position.

Equations (10.1) and $s_i(t)\geq 0$ increase $P(x_i^d(t)=0)$ because the minimum value for $P(x_i^d(t)=0)$ is 0.5 when $s_i(t)=0$ $(\min P(x_i^d(t)=0)\geq 0.5)$. Meanwhile, they decrease the maximum value for $P(x_i^d(t)=1)$ to 0.5 $(\max P(x_i^d(t)=1)\leq 0.5)$. Therefore, if $s_i(t)>0$, then $P(x_i^d(t)=0)>>0.5$ and $P(x_i^d(t)=1)<<0.5$.

Figure 10.1(a) shows that Eq.(10.1) and $s_i(t)\geq 0$ in SPSO increase $P(x_i^d(t)=0)$; whereas Fig.10.1(b) denotes that Eqs. (8.1), (8.2), and (8.3) in BPSO as stated in Chapter 8 yield $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$. For example, the calculations for $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ in Fig. 10.1(a) are shown as follows:

if $s_i(t)=1$, then $P(x_i^d(t)=0)=0.993307$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.006693$,

and

if $s_i(t)=2$, then $P(x_i^d(t)=0)=0.999955$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.000045$.

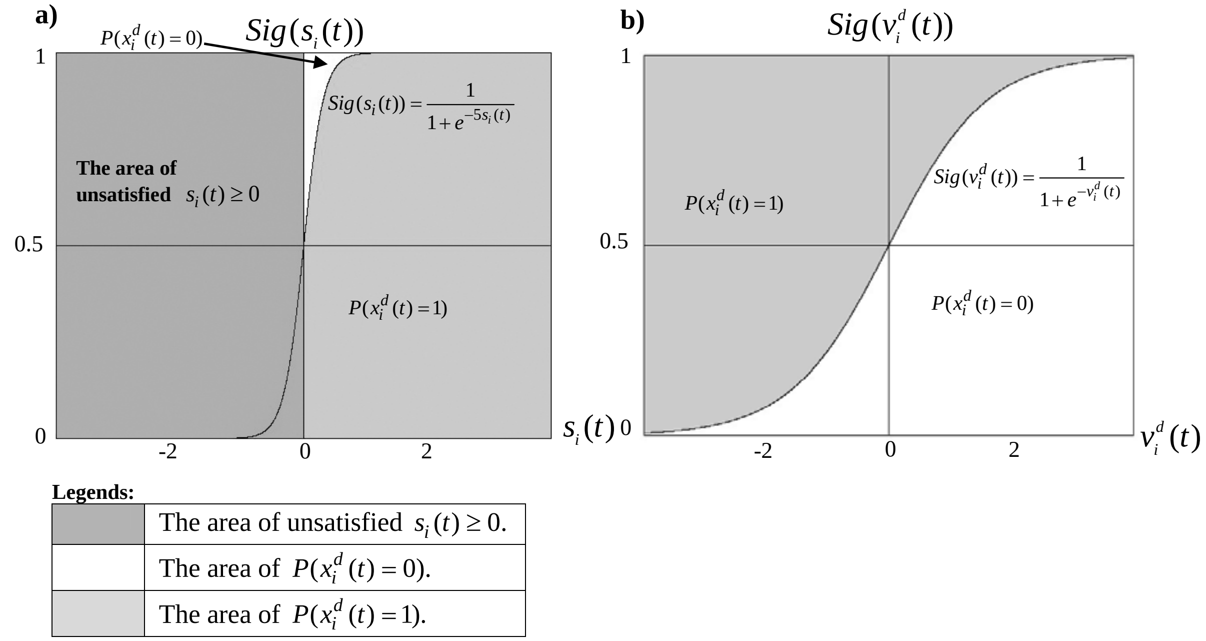


Fig. 10.1. The areas of unsatisfied $s_i(t)\geq 0$, $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ in (a) SPSO; (b) BPSO.

The high probability of $x_i^d(t)=0$ ($P(x_i^d(t)=0)$) causes a small number of genes are selected in order to produce a near-optimal (small) gene subset from high-dimensional data (gene expression data). Hence, SPSO is proposed to overcome the limitations of BPSO and the previous PSO-based methods, and finally produce a small subset of informative genes. The fitness value of particles (gene subsets) is calculated as follows:

$$fitness(X_i)=w_1\times A(X_i)+(w_2\times(n-R(X_i))/n) \tag{10.2}$$

where $A(X_i)\in[0,1]$ is the LOOCV classification accuracy that uses the only genes in a gene subset (X_i). This accuracy is provided by SVM. $R(X_i)$ is the number of selected genes in X_i. n is the total number of genes for each sample. w_1 and w_2 are two priority weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1\in[0.1,0.9]$ and $w_2=1-w_1$.

10.3 Experimental Results

10.3.1 Data sets and experimental setup

Five benchmark gene expression data sets used in this chapter. They included binary-classes and multi-classes data sets. These data sets are the leukemia, colon, lung, and mixed-lineage leukemia (MLL), and small round blue cell tumors (SRBCT) data sets. The first four data sets are summarized on Table 2.1 in Chapter 2. The SRBCT data set is a multi-classes data set. It has four classes; ewing family of tumors (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and burkitt lymphomas (BL). The training set contains 63 samples (22 EWS, 20 RMS, 12 NB, and 8 BL), whereas the test set contains 20 samples (6 EWS, 5 RMS, 6 NB, and 3 BL). There are 2,308 genes in each sample. It can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/.

Table 10.1. Parameter settings for SPSO and BPSO.

Parameters	Values
No. particles	100
No. iterations (generation)	500
w_1	0.8
w_2	0.2
c_1	2
c_2	2

Experimental results produced by SPSO are compared with an experimental method (BPSO) and other previous PSO-based methods for objective comparisons [1],[8],[21],[33]. SVM is used to measure LOOCV accuracy on gene subsets that produced by SPSO and BPSO. In order to avoid selection bias, the implementation of LOOCV is in exactly the same way as did by Chuang *et al.* [8] where the only one cross-validation cycle (outer loop), namely LOOCV is used. Several experiments are independently conducted 10 times on each data set using SPSO and BPSO. Next, an average result of the 10 independent runs is obtained. Two criteria following their importance are considered to evaluate the performances of SPSO and BPSO; LOOCV accuracy and the number of selected genes. Additionally, running times are also measured for the comparison between SPSO and BPSO. High accuracy and the small number of selected genes are needed to obtain an excellent performance. Table 10.1 contains parameter values for SPSO and BPSO. These values are chosen based on the results of preliminary runs.

10.3.2 Result analysis and discussion

Based on the standard deviation of classification accuracy in Table 10.2 and Table 10.3, results produced by SPSO were consistent on all data sets. Interestingly, all runs have achieved 100% LOOCV accuracy with less than 131 selected genes on the leukemia, SRBCT, and MLL the data sets. Moreover, over 91% classification accuracies have been obtained on the lung and colon data sets. This means that SPSO has efficiently selected and produced a near-optimal gene subset from high-dimensional data (gene expression data).

Figure 10.2 shows that the averages of fitness values of SPSO increase dramatically after a few generations on all the data sets. A high fitness value is obtained by a combination between a high classification rate and a small number (subset) of selected genes. The condition of the proposed particle's speed that should always be positive real numbers started in the initialization method, the new rule for updating particle's positions, and the modified sigmoid function provokes the early convergence of SPSO. In contrast, the averages of fitness values of BPSO was no improvement until the last generation due to $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$.

According to the Table 10.4, overall, it is worthwhile to mentioning that the classification accuracy of SPSO are superior to BPSO in terms of the best, average, and standard deviation results on all the data sets. Moreover, SPSO also produces a smaller number of genes compared to BPSO. The running times of SPSO are lower than BPSO in all the data sets. SPSO can reduce its running times because SPSO selects the smaller number of genes compared to BPSO.

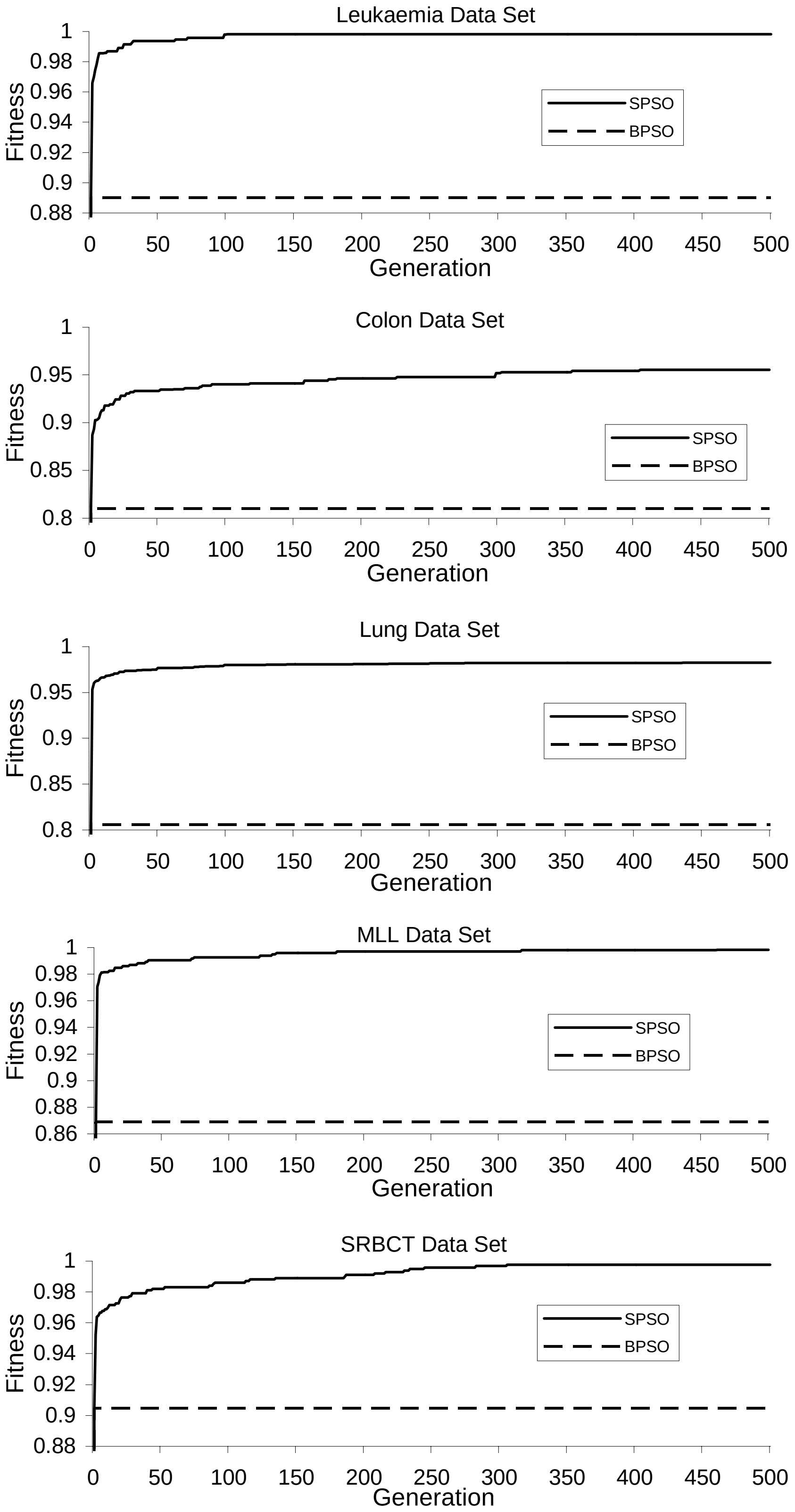


Fig 10.2. A relation between the average of fitness values (10 runs on average) and the number of generations for SPSO and BPSO.

Table 10.2. Experimental results for each run using SPSO on the leukemia, colon, and lung data sets.

Run no.	Leukemia data set		Colon data set		Lung data set	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	100	55	93.55	17	98.34	119
2	100	65	93.55	11	97.79	115
3	100	65	95.16	22	97.79	107
4	100	70	96.77	22	98.34	125
5	100	51	98.39	23	97.79	128
6	100	62	95.16	15	98.34	130
7	100	58	93.55	27	97.79	111
8	100	61	95.16	29	98.34	106
9	100	63	93.55	20	97.79	127
10	100	67	91.94	16	98.34	129
Average ± S.D.	100 ± 0	61.70 ± 5.72	94.68 ± 1.87	20.20 ± 5.55	98.07 ± 0.29	119.70 ± 9.37

Note: The results of the best subsets are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively.

Table 10.3. Experimental results for each run using SPSO on the MLL and SRBCT data sets.

Run no.	MLL data set		SRBCT data set	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	100	131	100	33
2	100	123	100	26
3	100	117	100	25
4	100	113	100	26
5	100	116	100	31
6	100	109	100	22
7	100	116	100	26
8	100	114	100	21
9	100	111	100	29
10	100	111	100	22
Average ± S.D.	100 ± 0	116.10 ± 6.56	100 ± 0	26.10 ± 3.96

Table 10.4. Comparative experimental results of SPSO and BPSO.

Data set	Method / Evaluation	SPSO			BPSO		
		Best	#Ave	S.D	Best	#Ave	S.D
Leukemia	#Acc (%)	100	100	0	98.61	98.61	0
	No. selected genes	51	61.70	5.72	3488	3528.75	26.83
	Running time (hour)	7.52	7.46	0.67	261.34	261.41	0.18
Colon	#Acc (%)	98.39	94.68	1.87	90.32	88.55	0.92
	No. selected genes	23	20.20	5.55	982	985.00	25.22
	Running time (hour)	5.06	5.02	0.07	64.45	64.63	0.18
Lung	#Acc (%)	98.34	98.07	0.29	88.40	88.12	0.32
	No. selected genes	106	119.70	9.37	6177	6193.25	26.99
	Running time (hour)	94.80	94.79	0.12	1040.55	1040.50	0.10
MLL	#Acc (%)	100	100	0	95.83	95.83	0
	No. selected genes	109	116.10	6.56	6101	6153.1	31.62
	Running time (hour)	13.51	13.83	0.18	236.759	239.00	1.34
SRBCT	#Acc (%)	100	100	0	100	100	0
	No. selected genes	21	26.10	3.96	1076	1098.33	12.46
	Running time (hour)	10.32	9.63	1.24	136.81	136.87	0.06

Note: The best results of each data set are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively, whereas #Ave represents an average.

For an objective comparison, the present work compares SPSO with the previous related works that used PSO-based methods in their proposed methods [1],[8],[21],[33]. It is shown in Table 10.5. For leukemia, lung, MLL, and SRBCT the data sets, the averages of classification accuracies of the present work were higher than the previous works. The present work also has resulted the smaller averages of the number of selected genes on the data sets compared to the previous works. The latest previous work also came up with the similar LOOCV results (100%) to the present work on the leukemia and SRBCT data sets, but they used many genes (more than 400 genes) to obtain the same results [8]. Moreover, they could not have statistically meaningful conclusions because their experimental results were obtained by only one independent run on each data set, and not based on average results. The average results are important since their proposed method is a stochastic approach.

Additionally, in their approach, the global best particle's position is reset to zero position when its fitness values do not change after three successive iterations. Theoretically, their approach is almost impossible to result a near-optimal gene subset from high-dimensional spaces (high-dimension data) because the global best particle's position should make a new exploration and exploitation for searching the near-optimal solution after its position reset to zero. Overall, the present work has outperformed the previous related works in terms of LOOCV accuracy and the number of selected genes.

Table 10.5. A comparison between SPSO and previous PSO-based methods.

Data set	Method Evaluation	SPSO	IBPSO [8]	PSOTS [33]	PSOGA [21]	GPSO [1]
Leukemia	#Acc (%)	(100)	100	(98.61)	(95.10)	-
	#Genes	(61.70)	1034	(7)	(21)	-
Colon	#Acc (%)	(94.68)	-	(93.55)	(88.7)	-
	#Genes	(20.20)	-	(8)	(16.8)	-
Lung	#Acc (%)	(98.07)	-	-	-	(99)
	#Genes	(119.70)	-	-	-	(4)
MLL	#Acc (%)	(100)	100	-	-	-
	#Genes	(116.10)	1292	-	-	-
SRBCT	#Acc (%)	(100)	100	-	-	-
	#Genes	(26.10)	431	-	-	-

Note: The results of the best subsets are shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. '-' means that a result is not reported in the related previous work. A result in '( )' denotes an average result. #Genes and #Acc represent the number of selected genes and the classification accuracy, respectively.

IBPSO = Improved binary PSO.
PSOGA = A hybrid of PSO and GAs.
PSOTS = A hybrid of PSO and tabu search.
GPSO = Geometric PSO.

According to Fig. 10.2 and Tables 10.2, 10.3, 10.4, and 10.5, SPSO is reliable for gene selection since it has produced the near-optimal solution from gene expression data. This is due to the fact that the proposed modified sigmoid function increases the probability $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$. This high probability causes the selection of a small number of informative genes and finally produces a near-optimal subset (a small subset of informative genes with high classification accuracy) for cancer classification. The sigmoid function is modified for increasing the probability of bits in particle's positions to be zero.

10.4 Summary

In this chapter, SPSO has been proposed for gene selection on five gene expression data sets. Overall, based on the experimental results, the performance of SPSO was superior to BPSO and PSO-based methods that proposed by the previous related works in terms of classification accuracy and the number of selected genes. SPSO was excellent because the probability $x_i^d(t+1)=0$ has been increased by the modified sigmoid function. The modified function has been proposed in order to yield a near-optimal subset of genes for better cancer classification. SPSO also obtains lower running times because it selects the small number of genes compared to BPSO. The next chapter will conclude this thesis.

Chapter 11
Conclusions

11.1 Introduction

This research has proposed intelligent approaches to select informative genes from gene expression data for cancer classification. The proposed approaches have been introduced based on GAs and PSO. Twelve gene expression data sets were used to test the effectiveness of the approaches in terms of the number of selected genes and classification accuracy. This chapter draws general conclusions about the achieved results, and offers several potential ideas for future works.

11.2 Conclusion Remarks

There were three main problems encountered when investigating and analyzing the applicability of intelligent approaches to select a small subset of informative genes from gene expression data for cancer classification, namely, the small number of samples compared to the huge number of genes (high-dimension), irrelevant genes, and noisy genes. The work in this research has addressed all the three challenges with the promising approaches.

Six intelligent approaches based on GAs have been proposed to select a small subset of informative genes when dealing with the data. These approaches are a multi-objective strategy in GASVM, a combination of two hybrid methods, a cyclic hybrid method, an iterative approach, a two-stage method, and a three-stage method. The approaches use some ideas such as dimensionality reduction, filter out irrelevant genes, and remove noisy genes, to produce near-optimal gene subsets for cancer classification. The ideas work in stochastic environments in which GAs have been implemented to search and find the near-optimal subsets. More importantly, by performing experiments on five gene expression data sets, the present work has found that the performances of the proposed approaches were superior to the other previous related works, as well as to several methods experimented in this research. The performances include classification accuracy and the number of selected genes.

The three remaining approaches were the extensions of binary PSO. These approaches are modified binary PSO, enhanced binary PSO, and improved binary PSO. Ideally, the three approaches were introduced to reduce the probability of genes to be selected. To decrease the probability, the following mechanisms have been proposed: particle’s speed, a new rule for updating particle’s velocities, the constraint of particle’s velocities, and a modified sigmoid function. The proposed approaches were evaluated on twelve benchmark gene expression data sets and obtained excellent results on those data sets as compared with other previous related works, including BPSO in terms of classification accuracy and the number of selected genes. The proposed approaches also produced lower running times compared to BPSO.

11.3 Direction for Future Works

Based on the findings in this research, several areas deserve further study. The proposed approaches can be extended for applications on other biological data such as protein structures, protein-protein interactions, etc. This can be implemented by combining the data into gene expression data for the same purposes such as gene selection and cancer classification. The data combination is important to improve the classification accuracy and provide useful biological information as the final product for biologists.

Since the last phase for analyzing gene expression data is classification process, the modification of classifiers are also needed to increase classification accuracy. The classifiers can be improved by providing functions for gene selection in their structures and modules which would make them applicable to a wider range of solutions. According to the findings of extended approaches based on binary PSO in this research (Chapters 8, 9, and 10), the proposed approaches could reduce the probability of genes to be selected and finally yield high classification accuracy. Therefore, any idea or formula to decrease the probability can be expected to produce good results on gene selection and classification processes.

APPENDIX A

The list of publications referred in this thesis

No.	Titles	Authors	Journals / Conferences	Related Chapters
1	A Multi-Objective Strategy in Genetic Algorithm for Gene Selection of Gene Expression Data	M.S. Mohamad S. Omatu S. Deris M.F. Misman M. Yoshioka	Int. J. Artif. Life Rob., Vol. 13, No. 2, pp. 410-413 (2009).	Chapter 2
2	Multi-Objective Optimization Using Genetic Algorithm for Gene Selection from Microarray Data	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Proc. of the Int. Conf. on Comput. and Commun. Eng. 2008, pp. 1331-1334 (Kuala Lumpur, Malaysia, 2008).	Chapter 2
3	A Multi-Objective Strategy in Genetic Algorithm for Gene Selection of Gene Expression Data	M.S. Mohamad S. Omatu S. Deris M.F. Misman	Proc. of the 13th Int. Symp. on Artif. Life and Rob., CD-ROM pp. 324-327 (Beppu, Japan, 2008).	Chapter 2
4	Selecting Informative Genes from Microarray Data by Using Hybrid Methods for Cancer Classification	M.S. Mohamad S. Omatu S. Deris M.F. Misman M. Yoshioka	Int. J. Artif. Life Rob., Vo. 13, No. 2, pp. 414-417 (2009).	Chapter 3
5	An Approach Using Hybrid Methods to Select Informative Genes from Microarray Data for Cancer Classification	M.S. Mohamad S. Omatu M. Yoshioka S. Deris	Proc. of the 2nd Asia Int. Conf. on Model. & Simul. 2008, pp. 603-608 (Kuala Lumpur, Malaysia, 2008).	Chapter 3
6	Selecting Informative Genes from Microarray Data by Using a Hybrid Algorithm for Cancer Classification	M.S. Mohamad S. Omatu S. Deris M.F. Misman	Proc. of the 13th Int. Symp. on Artif. Life and Rob., CD-ROM pp. 328-331 (Beppu, Japan, 2008).	Chapter 3
7	A Cyclic Hybrid Method to Select a Smaller Subset of Informative Genes for Cancer Classification	M.S. Mohamad S. Omatu M. Yoshioka S. Deris	Int. J. Innovative Comput., Inf. Control, Vol. 5, No. 10, pp. 2189-2202 (2009).	Chapter 4
8	A Recursive Genetic Algorithm to Automatically Select Genes for Cancer Classification	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Advances in Soft Computing, Vol. 49, pp. 166-174. Springer (Berlin / Heidelberg, 2009).	Chapter 4
9	Selecting Informative Genes from Microarray Data by Using a Cyclic GA-Based Method	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	The 1st Int. Conf. on Intell. Syst., Model. and Simul., Accepted for publication (Liverpool, United Kingdom, 2010).	Chapter 4

10	Gene Subset Selection Using an Iterative Approach Based on Genetic Algorithms	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Int. J. Artif. Life Rob., Vol. 14, No.1, pp. 12-15 (2009).	Chapter 5
11	Gene Subset Selection Using an Iterative Approach Based on Genetic Algorithms	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Proc. of the 14th Int. Symp. on Artif. Life and Rob., CD-ROM pp. 758-761 (Beppu, Japan, 2009).	Chapter 5
12	An Iterative GASVM-Based Method: Gene Selection and Classification of Microarray Data	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Lecture Notes in Computer Science, Vol. 5518, pp. 187-194. Springer (Berlin / Heidelberg, 2009).	Chapter 5
13	A Two-Stage Method to Select a Smaller Subset of Informative Genes for Cancer Classification	M.S. Mohamad S. Omatu M. Yoshioka S. Deris	Int. J. Innovative Comput., Inf. and Control, Vol. 5, No. 12, pp. 2959-2968 (2009).	Chapter 6
14	A Model for Gene Selection and Classification of Gene Expression Data	M.S. Mohamad S. Omatu S. Deris S.Z.M. Hashim	Int. J. Artif. Life Rob., Vol. 11, No. 2, pp. 219-222 (2007).	Chapter 6
15	A Model for Gene Selection and Classification of Gene Expression Data	M.S. Mohamad S. Omatu S. Deris S.Z.M. Hashim	Proc. of the 13th Int. Symp. on Artif. Life and Rob., pp. 320-323 (Beppu, Japan, 2008).	Chapter 6
16	Three-Stage Method for Selecting Informative Genes for Cancer Classification	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	IEEJ Trans. Electr. Electron. Eng., Vol. 4, No. 6, pp. 725-730 (2009).	Chapter 7
17	A Three-Stage Method to Select Informative Genes for Cancer Classification	M.S. Mohamad S. Omatu M. Yoshioka S. Deris	Int. J. Innovative Comput., Inf. and Control, in press.	Chapter 7
18	A Three-Stage Method to Select Informative Genes for Cancer Classification	M.S. Mohamad S. Omatu M. Yoshioka S. Deris	The 40th ISCIE Int. Symp. on Stochastic Syst. Theory and Its Appl., pp. 63-64 (Kyoto, Japan, 2008).	Chapter 7
19	A Three-Stage Method to Select Informative Genes from Gene Expression Data in Classifying Cancer Classes	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	The 1st Int. Conf. on Intell. Syst., Model. and Simul., Accepted for publication (Liverpool, United Kingdom, 2010).	Chapter 7
20	Particle Swarm Optimization for Gene Selection in Classifying Cancer Classes	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Int. J. Artif. Life Rob., Vol. 14, No. 1, pp. 16-19 (2009).	Chapter 8
21	Particle Swarm Optimization for Gene Selection in Classifying Cancer Classes	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Proc. of the 14th Int. Symp. on Artif. Life and Rob., CD-ROM pp. 762-765 (Beppu, Japan, 2009).	Chapter 8

22	An Improved Binary Particle Swarm Optimisation for Gene Selection in Classifying Cancer Classes	M.S. Mohamad S. Omatu S. Deris M. Yoshioka A. Zainal	Lecture Notes in Computer Science, Vol. 5518, pp. 495-502. Springer (Berlin / Heidelberg, 2009).	Chapter 8
23	Particle Swarm Optimization for Gene Selection Using Microarray Data	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Proc. of the 2009 Int. Conf. on Multimedia, Inf. Technol. and its Appl., pp. 37-40 (Sakai, Japan, 2009).	Chapter 8
24	An Enhancement of Binary Particle Swarm Optimization Based on the Proposed Constraint and Rule for Selecting a Small Subset of Informative Genes	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	Far East J. Math. Sci., (in press).	Chapter 9
25	Selecting Genes from Gene Expression Data by Using an Enhancement of Binary Particle Swarm Optimization for Cancer Classification	M.S. Mohamad S. Omatu S. Deris M. Yoshioka	The 2nd Int. Conf. on Agents and Artif. Intell., Accepted for publication (Valencia, Spain, 2010).	Chapter 10

APPENDIX B

The glossary of structural genomic terms

Word	Meaning
ab initio	From the beginning
amino acid	One of 20 naturally occurring amino carboxylic acid molecules
amino acid sequence	Arrangement of the residues in a protein
biopsy	Examination of severed tissue for diagnostic
C-terminal	The residue in a peptide that has a free carboxyl group
epidemiology	Scientific discipline studying the incidence, distribution, and control of disease in a population.
gene	A specific location of genetic coding that possesses part of the building blocks of an organism. Each gene is tailor designed to have the information required for a particular function, and can be switched on and off on demand. Genetics a segment of DNA that is involved in producing a polypeptide chain; it can include regions preceding and following the coding DNA as well as introns between the exons; it is considered a unit of heredity; "genes were formerly called factors".
gene expression	Pertaining to a gene that is active in nature (not dormant). Conversion of the information encoded in a gene first into messenger RNA and then to a protein.
gene expression level	Approximate number of copies of RNA in a cell.
gene induction	Activation of an inactive gene.
genome	the genetic material of an organism; the complete DNA component of an organism
in vivo	in the living body of a plant or animal
motif	commonly observed structural components of proteins formed by simple combinations of adjacent secondary structures
native state/structure	the final conformation for proteins in the intact cell
peptide	derived from two or more amino carboxylic acid molecules by formation of a covalent bond from the carbonyl carbon of one to the nitrogen atom of another with formal loss of water
polypeptide	a peptide containing ten or more amino acids
primary structure	amino acid sequence; the order of amino acids as they occur in a polypeptide chain
protein	a naturally occurring and extremely complex substance that consists of amino acid residues joined by peptide bonds
protein folding	a rapid biochemical reaction involved in the formation of proteins; it begins before a protein has been completely synthesized and proceeds through discrete intermediates (primary, secondary, and tertiary structures) before the final structure (quaternary structure) is developed
protein sequence	amino acid sequence of a protein
residue	an amino acid unit in the polypeptide chain: when two or more amino acids combine to form a peptide, the elements of water are removed, and what remains of each amino acid is called an amino acid residue
sequence homology	the degree of similarity between sequences of amino acids
target sequence	amino acid sequence which is considered for prediction

Bibliography

[1] E. Alba, J. Garcia-Nieto, L. Jourdan, and E. Talbi, "Gene Selection in Cancer Classification Using PSO/SVM and GA/SVM Hybrid Algorithms," *Proc. of the 2007 IEEE Congress on Evolutionary Comput.*, pp. 284–290, 2008.

[2] C. Ambroise and G. J. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-expression Data," *Proc. of the National Academic of Sciences*, Vol. 99, No. 10, pp. 6562–6566, 2002.

[3] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a Unique Leukemia," *Nature Genetics*, Vol. 30, pp. 41–47, 2002.

[4] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns," *J. of Comput. And Biol.*, Vol. 6, Nos. 3-4, pp. 281–297, 2000.

[5] F. Bergh and A. P. Engelbrecht, "A Study of Particle Swarm Optimization Particle Trajectories," *Inf. Sciences*, Vol. 176, No. 8, pp. 937–971, 2006.

[6] M. P. S. Brown, W. N. Grundy, D. Lin, N. Christianini, C. Sugnet, T. S. Furey, M. J. Ares, and D. Haussler, "Knowledge-Based Analysis of Microarray Gene Expression Data Using Support Vector Machines," *Proc. of the National Academic of Sciences*, Vol. 97, No. 1, pp. 262–267, 2000.

[7] S. B. Cho and H. H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification," *Proc. of the 1st Asia-Pacific Bioinformatics Conf. on Bioinformatics*, Vol. 19, pp. 189–198, 2003.

[8] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, “Improved Binary PSO for Feature Selection Using Gene Expression Data,” *Comput. Biol. and Chem.*, Vol. 32, No. 1, pp. 29–38, 2009.

[9] T. S. Furey, N. Cristianini, N. Duffy, M. Schummer, D. W. Bednarski, and D. Haussler, “Support Vector Machine Classification and Validation of Cancer Tissue Sample Using Microarray Expression Data,” *Bioinformatics*, Vol. 16, No. 10, pp. 906–914, 2000

[10] T. R. Golub, D. K. Slonim, P. Tomayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” *Science*, Vol. 286, No. 5439, pp. 531–537, 1999.

[11] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, “Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma,” *Cancer Research*, Vol. 62, No. 17, pp. 4963–4967, 2002.

[12] J. Handl, D. B. Kell, and J Knowles, “Multi-Objective Optimization in Bioinformatics and Computational Biology,” *IEEE/ACM Trans. on Comput. Biol. and Bioinformatics*, Vol. 4, No. 2, pp. 279–292, 2007.

[13] H. L. Huang and F. L. Chang, ESVM, “Evolutionary Support Vector Machine for Automatic Feature Selection and Classification of Microarray Data,” *BioSystems*, Vol. 90, No. 2, pp. 516–528, 2007.

[14] J. Kennedy and R. Eberhart, “Particle Swarm Optimization,” *Proc. of the 1995 IEEE Int. Conf. on Neural Networks*, Vol. 4, pp. 1942–1948, 1995.

[15] J. Kennedy and R. Eberhart, “A Discrete Binary Version of the Particle Swarm Algorithm,” *Proc. of the 1997 IEEE Int. Conf. on Syst., Man, and Cybern.*, Vol. 5, pp. 4104–4108, 1997.

[16] J. Khan , J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, “Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks,” *Nature Medicine*, Vol. 7, No. 6, pp. 673–679, 2001.

[17] S. Knudsen, *A Biologist’s Guide to Analysis of DNA Microarray Data*, John Wiley & Sons, 2002.

[18] L. Li, W. Jiang, X. Li, K. L. Moser, Z. Guo, L. Du, Q. Wang, E. J. Topol, Q. Wang, and S. Rao, “A Robust Hybrid Between Genetic Algorithm and Support Vector Machine for Extracting an Optimal Feature Gene Subset,” *Genomics*, Vol. 85, No. 1, pp. 16–23, 2005.

[19] J. Li, H. Liu, S. K. Ng, and L. Wong, “Discovery of Significant Rules for Classifying Cancer Diagnosis Data,” *Bioinformatics*, Vol. 19, No. 2, pp. 93–102, 2003.

[20] S. Li, X. Wu, and X. Hu, “Gene Selection Using Genetic Algorithm and Support Vectors Machines,” *Soft Computing*, Vol. 12, No. 7, pp. 693–698, 2008.

[21] S. Li, X. Wu, and M. Tan, “Gene Selection Using Hybrid Particle Swarm Optimization and Genetic Algorithm,” *Soft Computing*, Vol. 12, No. 11, pp. 1039–1048, 2008.

[22] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittman, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown, “Expression Monitoring by Hybridization to High-density Oligonucleotide Arrays,” *Nature Biotechnology*, Vol. 14, No. 13, pp. 1675–1680, 1996.

[23] M. S. Mohamad, S. Deris, and R. M. Illias, “A Hybrid of Genetic Algorithm and Support Vector Machine for Features Selection and Classification of Gene Expression Microarray,” *Int. J. of Comput. Intell. and Appl.*, Vol. 5, No. 1, pp. 91–107, 2005.

[24] M. S. Mohamad, S. Omatu, S. Deris, and S. Z. M. Hashim, “A Model for Gene Selection and Classification of Gene Expression Data,” *Int. J. of Artif. Life and Robotics*, Vol. 11, No. 2, pp. 219–222, 2007.

[25] M. S. Mohamad, S. Omatu, S. Deris, M. F. Misman, and M. Yoshioka, “A Multi-Objective Strategy in Genetic Algorithm for Gene Selection of Gene Expression Data,” *Int. J. of Artif. Life and Robotics,* Vol. 13, No. 2, pp. 410–413, 2009.

[26] M. S. Mohamad, S. Omatu, S. Deris, M. F. Misman, and M. Yoshioka, “Selecting Informative Genes from Microarray Data by Using Hybrid Methods for Cancer Classification,” *Int. J. of Artif. Life and Robotics*, Vol. 13, No. 2, pp. 414–417, 2009.

[27] M. S. Mohamad, S. Omatu, M. Yosioka, and S. Deris, “A Cyclic Hybrid Method to Select a Smaller Subset of Informative Genes for Cancer Classification,” *Int. J. of Innovative Comput., Inf. and Control*, Vol. 5, No. 8, pp. 2189–2202, 2009.

[28] S. Mukherjee, *Application of Statistical Learning Theory to DNA Microarray Analysis*, Ph.D. Thesis, Massachusetts Institute of Technology, 2001. Doi: 10.1.1.9.1695.

[29] S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, and L. Chen, “Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machines,” *FEBS Letters*, Vol. 555, pp. 358–362, 2003.

[30] Y. Saeys, I. Inza, and P. Larranaga, “A Review of Feature Selection Techniques in Bioinformatics,” *Bioinformatics,* Vol. 23, No. 19, pp. 2507–2517, 2007.

[31] S. Shah and A. Kusiak, “Cancer Gene Search with Data-mining and Genetic Algorithms,” *Computers in Biol. and Medicine,* Vol. 37, No. 2, pp. 251–261, 2007.

[32] Q. She, H. Su, L. Dong, and J. Chu, “Support Vector Machine with Adaptive Parameters in Image Coding,” *Int. J. of Innovative Comput., Inf. and Control*, Vol. 4, No. 2, pp. 359–367, 2008.

[33] Q. Shen, W. M. Shi, and W. Kong, “Hybrid Particle Swarm Optimization and Tabu Search Approach for Selecting Genes for Tumor Classification Using Gene Expression Data,” *Comput. Biol. and Chem.*, Vol. 32, No. 1, pp. 53–60, 2009.

[34] Ö. Uncu and I. B. Türksen, “A Novel Feature Selection Approach: Combining Feature Wrappers and Filters,” *Inf. Sciences*, Vol. 177, No. 2, pp. 449–466, 2007.

[35] L. Wang, F. Chu, and W. Xie, “Accurate Cancer Classification Using Expressions of Very Few Genes,” *IEEE/ACM Trans on Comput. Biol. and Bioinformatics*, Vol. 4, No. 1, pp. 40–53, 2007.

[36] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, “HykGene: A Hybrid Approach for Selecting Marker Genes for Phenotype Classification Using Microarray Gene Expression Data,” *Bioinformatics*, Vol. 21, No. 8, pp. 1530–1537, 2005.

[37] R. Xu, G. C. Anagnostopoulos, and D. C. Wunsch II, “Tissue Classification Through Analysis of Gene Expression Data Using a New Family of ART Architectures,” *Proc. of the 2002 Int. Joint Conf. on Neural Networks*, Vol. 1, pp. 300–304, 2002.

[38] K. Yang, Z. Cai, J. Li, and G. Lin, “A Stable Gene Selection in Microarray Data Analysis,” *BMC Bioinformatics*, Vol. 7, pp. 228–246, 2006.