FILTER-WRAPPER METHODS FOR GENE SELECTION IN CANCER CLASSIFICATION

OSAMA AHMAD SULEIMAN ALOMARI

UNIVERSITI SAINS MALAYSIA

2018

FILTER-WRAPPER METHODS FOR GENE SELECTION IN CANCER CLASSIFICATION

by

OSAMA AHMAD SULEIMAN ALOMARI

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

September 2018

ACKNOWLEDGEMENTS

In the Name of Allah, the Beneficent, the Merciful. First praise is to Allah, the Almighty, on whom ultimately we depend for sustenance and guidance. My sincere thanks goes to my supervisor Prof. Ahamad Tajudin Khader, from the School of Computer Sciences, Universiti Sains Malaysia (USM), for his wise counseling, valuable advice, continuous support, help and guidance throughout the duration of my research. Indeed, without his support and cooperation, I could not have completed this study. I owe my deepest gratitude to my co-supervisor, Assoc. Prof. Dr. Mohammed Azmi Al-Betar from Department of Information Technology, Al-Huson University College, Al-Balqa Applied University, Jordan, for his patience, motivation, enthusiasm, and immense knowledge. I am grateful to Universiti Sains Malaysia for financial support under the Fellowship Scheme throughout my Ph.D. candidature.

Most of all, I would like to express my deepest appreciation to my beloved parents who always encouraged me to go ahead with my study. All thanks to my father and my dearest mother for your prayers and support.

Finally, I would like to express my sincere and heartfelt thanks to all my dear friends who supported me during my research study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC Artificial Bee Colony

BA Bat-inspired Algorithm

GA Genetic Algorithm

HSA Harmony Search Algorithm

IG Information Gain

MB Markov Blanket

NB Number of Bats

mRNA Messenger Ribonucleic Acid

MRMR Minimum Redundancy Maximum Relevancy

MRMR-BA Hybrid Minimum Redundancy Maximum Relevancy and Adapted Bat Algorithm

rMRMR-BA Hybrid Robust Minimum Redundancy Maximum Relevancy and Adapted Bat Algorithm

rMRMR-MBA Hybrid Robust Minimum Redundancy Maximum Relevancy and Modified Bat Algorithm

rMRMR-HBA Hybrid Robust Minimum Redundancy Maximum Relevancy and Hybrid Bat Algorithm

PSO Particle Swarm Optimisation

PCA principal component analysis

RBF Radial basis kernel function

SBS Sequential Backward Strategy

SFS Sequential Forward Strategy

SSO Simplified swarm optimisation

SVM support vector machine

PENDEKATAN PENAPIS-PEMBALUT UNTUK PILIHAN GEN DALAM KLASIFIKASI BARAH

ABSTRAK

Dalam kajian ungkapan gen 'microarray,' menemukan subset terkecil gen bermaklumat daripada set data 'microarray' bagi tujuan klinikal dan pengkelasan kanser yang tepat adalah salah satu daripada cabaran paling sukar dalam tugas pembelajaran mesin. Ramai pengkaji telah cuba untuk menangani masalah ini dengan menggunakan kaedah penapis, kaedah pembalut ataupun gabungan kedua-dua pendekatan. Kaedah hibrid adalah merupakan kaedah penghibridan di antara kaedah penapis dan kaedah pembalut. Ia mendapat manfaat daripada kelajuan pendekatan penapis and ketepatan pendekatan pembalut. Beberapa kaedah penapis-pembalut hibrid telah dicadangkan bagi memilih gen bermaklumat. Namun, kaedah-kaedah hibrid berhadapan dengan beberapa halangan yang dikaitkan dengan pendekatan-pendekatan penapis dan pembalut. Subset gen yang dihasilkan daripada pendekatan-pendekatan penapis memiliki kekurangan dari segi ramalan dan kekukuhan. Kaedah pembalut berhadapan dengan masalah-masalah interaksi yang kompleks di kalangan gen-gen dan genangan dalam optima setempat. Bagi menangani kelemahan-kelemahan ini, kajian ini menyiasat kaedah-kaedah penapis dan pembalut bagi membentuk kaedah-kaedah hibrid yang berkesan bagi tujuan pemilihan gen. Kajian ini mencadangkan kaedah-kaedah penapis-pembalut hibrid yang baru berdasarkan Maximum Relevancy Minimum Redundancy (MRMR) sebagai suatu pendekatan penapis dan mengadaptasi algoritma

diinspirasikan dari kelawar (bat-inspired/BA) sebagai suatu kaedah pembalut. Pertamanya, penghibridan MRMR dan pengadaptasian BA disiasat bagi menyelesaikan masalah pemilihan gen. Kaedah yang dicadangkan dinamakan sebagai MRMR-BA. Kedua, pengubahsuaian kaedah penapis (iaitu MRMR) telah diperiksa. Satu himpunan pendekatan-pendekatan penapis (iaitu ReliefF, Chi-Square dan Kullback-Liebler) dihibridkan dengan mekanisma penapis MRMR bagi meningkatkan kekukuhannya, dan kaedah ini dirujuk sebagai rMRMR-BA. Ketiga, pengubahsuaian kaedah pembalut (iaitu BA) disiasat. Operator optimisasi tambahan yang berdasarkan penyelesaian inventif TRIZ selanjutnya meneroka interaksi di antara gen. Kaedah ini dirujuk sebagai rMRMR-MBA. Akhir sekali, kajian ini menyiasat penghibridan BA dengan algoritma carian setempat (iaitu Mendaki Bukit/β Hill Climbing). Keputusan-keputusan yang dicapai dalam kajian ini dibandingkan dengan keputusan-keputusan 10 kaedah yang lain menggunakan 14 set data-set data penanda aras 'microarray' yang mempunyai saiz dah kerumitan berbeza. rMRMR-HBA yang dicadangkan mencapai keputusan-keputusan terbaik bagi 8 daripada 14 set data. Lebih-lebih lagi, kaedah yang dicadangkan menghasilkan keputusan-keputusan yang kompetitif ke atas set data yang selebihnya.

FILTER-WRAPPER METHODS FOR GENE SELECTION IN CANCER CLASSIFICATION

ABSTRACT

In microarray gene expression studies, finding the smallest subset of informative genes from microarray datasets for clinical diagnosis and accurate cancer classification is one of the most difficult challenges in machine learning task. Many researchers have devoted their efforts to address this problem by using a filter method, a wrapper method or a combination of both approaches. A hybrid method is a hybridisation approach between filter and wrapper methods. It benefits from the speed of the filter approach and the accuracy of the wrapper approach. Several hybrid filter-wrapper methods have been proposed to select informative genes. However, hybrid methods encounter a number of limitations, which are associated with filter and wrapper approaches. The gene subset that is produced by filter approaches lacks predictiveness and robustness. The wrapper approach encounters problems of complex interactions among genes and stagnation in local optima. To address these drawbacks, this study investigates filter and wrapper methods to develop effective hybrid methods for gene selection. This study proposes new hybrid filter-wrapper methods based on Maximum Relevancy Minimum Redundancy (MRMR) as a filter approach and adapted bat-inspired algorithm (BA) as a wrapper approach. First, MRMR hybridisation and BA adaptation are investigated to resolve the gene selection problem. The proposed method is called MRMR-BA. Second, the modification of the filter approach (i.e., MRMR) is examined. An ensem-

ble of filter approaches (i.e., ReliefF, Chi-Square and Kullback-Liebler) is hybridised with the filtering mechanism of MRMR to increase its robustness, and this method is referred to as rMRMR-BA. Third, the modification of the wrapper approach (i.e., BA) is investigated. Additional optimization operators, which are based on TRIZ inventive solution, further explored the interaction between genes. This method is referred to as rMRMR-MBA. Finally, this study investigates BA hybridisation with local search algorithm (i.e., β Hill Climbing) to enhance local exploitation capability. This method is referred to as rMRMR-HBA. The obtained results of this study are compared with those of 10 other methods by using 14 benchmark microarray datasets of different sizes and complexity. The proposed rMRMR-HBA achieved the best results on 8 out of the 14 datasets. Moreover, the proposed method yielded competitive results on the remaining datasets.

CHAPTER 1

INTRODUCTION

1.1 Background

Precise diagnosis and effective treatment of diseases are key issues in scientific research, which bear positive meanings and implications for human health. Nowadays, cancer has become a very serious problem as it can infect anyone regardless of color, creed or status. The current cancer classification comprises more than 100 types (Berndt et al., 2017). For the cancer patient to receive an appropriate therapy, the clinician must classify as precisely as possible the cancer type.

Analyzing the morphologic characteristics of biopsy specimens has still been considered as a standard diagnostic method. However, it has its drawbacks. There exist very limited information in this regard and obviously important missing tumor aspects such as the capacity for the invasion and metastases, proliferation rate and evolution of resistance mechanisms to certain treatment agents (Perez-Diez et al., 2007).

Molecular diagnostic methods are essentially required to classify tumor subtypes in an appropriate manner. The classical molecular methods look for the DNA, RNA or protein of a defined marker, which is correlated with a specific type of tumor, and may or may not give biological information about cancer generation or progression. During the last two decades, the advent of microarray technology has enabled molecular biologists to extract a massive amount of molecular information, which can be used to

discover common patterns within a group of samples from a specific disease (Bolón-Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez and Herrera, 2014). There are many types of microarray, which have been generated such as DNA microarrays, protein microarrays, chemical compound microarrays, cell microarrays, tissue microarrays and antibody microarrays (Perez-Diez et al., 2007).

DNA microarray provides new insights into the mechanisms of the living systems through the possibility of analyzing thousands of genes simultaneously and getting significant information about the function of the cell. This particular information can be utilized for diagnosing many diseases such as Alzheimer (Panigrahi and Singh, 2013), diabetes (Yoo et al., 2009) and cancer (Chen et al., 2014). Gene expression data, which are extracted from the DNA microarray, have been widely employed to recognize cancer biomarkers or gene signature. This can complement the conventional histopathologic assessment, which can be done computationally through constructing machine learning algorithm on microarray dataset to generate a prediction model (i.e., cancer diagnostic tool). This tool is capable of classifying cancer tissues from normal tissues accurately (Alshamlan et al., 2015). Moreover, the tool can refine our understanding of the causes of cancers to discover a new therapy (Alba et al., 2007). Mullainathan and Spiess (2017) define machine learning as "" *an application of computer science that is related to artificial intelligence and allows algorithms to automatically recognize, classify, and extract data. The process of machine learning also optimizes the efficiency and accuracy of the information that it processes.*""

Gene expression data is considered as a high-dimensionality dataset, which typically consists of thousands of genes, but with only few numbers of patient samples

available for analysis. However, most of the genes are irrelevant, noisy and redundant. Many machine learning algorithms suffer from the curse of dimensionality. Therefore, data reduction is particularly required. Data reduction is a preprocessing technique, which is employed to overcome the dimensionality curse in the analysis of data. Data reduction boosts machine learning performance in terms of accuracy and simplicity, speed, and data interpretation. Feature selection (which is commonly known in the context of microarray dataset as gene selection) is a common data reduction technique, which is widely used to tackle the "curse of dimensionality" in microarray data analysis.

Gene selection is a process, which is carried out to find the most informative genes with respect to the improved predictive accuracy of diseases. Methods of gene selection are divided into three categories (Dash and Liu, 1997). The first is the wrapper approach. The second is the filter approach and the third is the hybrid approach of gene selection. The wrapper approach consists of two main components: subset generation (i.e., search techniques) and evaluation (i.e., machine learning algorithm). Many search techniques were used for subset generation, for example, Sequential Forward Strategy (SFS) (Whitney, 1971), Sequential Backward Strategy (SBS) (Kittler, 1986), genetic algorithm (Li and Yin, 2013), etc. The machine learning algorithms were used to perform the evaluation process, for example, the support vector machine (Vapnik, 1999), Naive bayes (Kelemen et al., 2003), and k nearest neighbor (Guo et al., 2004). The wrapper approach used machine learning algorithm to evaluate the reliability of genes or genes subsets. The filter approach does not include the machine learning algorithm for removing irrelevant and redundant features. Instead, it uses the principal characteristics of the training data to evaluate the significance of the genes subset or

genes (Kohavi and John, 1997). Filter approaches can be broadly classified into two sub-methods: 1) univariate and 2) multivariate. The univariate methods assess each individual gene independently of other genes according to various characteristics (distance, information, dependency, etc.), for example, Chi-Square (Su and Hsu, 2005), Kullback-Liebler (Kullback and Leibler, 1951), Fisher score (Gu et al., 2012), and Information Gain (Quinlan, 1986). The multivariate method considers the relevancy between the gene and the target class, as well as the redundancy between the genes. There are many existing multivariate filter methods based on the literature, for example, minimum redundancy maximum relevancy (MRMR) (Peng et al., 2005a) and ReliefF (Kononenko, 1994). The hybrid approach is hybridization between both filter and wrapper methods (Guyon and Elisseeff, 2003).

Metaheuristic techniques have been widely used as a role of subset generation in the wrapper approach to address gene selection problems. Their performance has been proven to be one of the best performing techniques, which have been used for solving gene selection problems (Jain et al., 2018, 2017; Salem et al., 2017; Mahajan et al., 2016; Shreem et al., 2014; Alshamlan et al., 2015). Osman and Laporte (1996) defined the metaheuristic as follows:

"*The metaheuristic-based method is an iterative improvement process that uses its operators and combines the problem specific knowledge for exploration and exploitation of the search space of the problem in order to reach acceptable solutions.*"

The search space is a bounded domain, which involves all possible solutions for the targeted problem. Any successful metaheuristic method should be able to make a

balance between exploration and exploitation during the search. Notably, the exploration process is the ability to explore new regions of the search space, which have not been visited before. On the other hand, the exploitation process requires an intensive search for the regions that have been already visited. Metaheuristic methods are classified into two categories, local search-based methods (or trajectory methods) and population-based methods (or evolutionary methods). The local search-based methods consider one solution at a time and attempts to enhance it using the neighborhood structures. The main advantages of these methods are the speed of search. However, the main drawback is that it is easy to be stuck in local optima by focusing on exploitation rather than exploration. Examples of local search-based methods are simulated annealing (Brooks and Morgan, 1995) and tabu search (Glover, 1989). In contrast, the population-based methods, which consider a population of solutions at a time, recombine the current solutions to generate one or more new solutions at iterations. The population-based methods are more concerned with exploration rather than exploitation. These include genetic algorithm (Holland, 1975), scatter search (Glover et al., 2000), ant colony optimization (Dorigo et al., 1996), and harmony search algorithm (Geem et al., 2001).

1.2 Motivations

The microarray technology facilitates biologist in monitoring the activity of thousands of genes in one experiment. This technology generates gene expression data, which are significantly applicable for cancer classification. However, due to the nature of gene expression dataset, as it is high-dimensional, biological heterogeneity, and innately noise that causes of generating irrelevant, redundant, noise genes (Bolón-

Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez and Herrera, 2014; Jain et al., 2018; Nguyen et al., 2015). These characteristics poses a challenges to the data interpretation and analysis, and for computational learning algorithms (i.e., Machine learning algorithms) to produce an accurate cancer diagnostic tool. From a computational point of view, Finding informative genes and isolating irrelevant and redundant genes are challenging tasks. However, they help enhance the predictive accuracy of a classifier technique and interpret the pattern of selected genes (Dash and Liu, 1997). Nevertheless, the existence of a large number of genes is a challenging issue in the development of an efficient classifier called machine learning algorithm (Lai et al., 2016). To address this challenge and to improve the predictive accuracy of diseases, researchers can apply gene selection, also known as feature selection, which is a data preprocessing step in data mining, to find the subset of most informative genes which can provide enhanced classification accuracy (Jain and Zongker, 1997),resulting in producing an accurate cancer diagnostic tool. Figure 1.1 illustrate the gene selection procedure on microarray gene expression data.

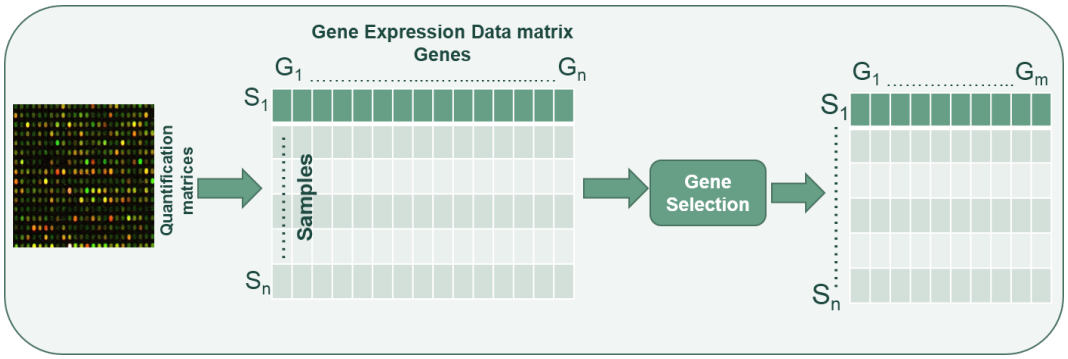

Figure 1.1: Gene selection procedure on microarray gene expression data.

1.3 Problem statement

Conventionally, gene selection methods are divided into three categories, namely, the wrapper approach, the filter approach, and the hybrid approach (Jain and Zongker, 1997). The hybrid approach, which is the last category in gene selection methods, is a hybrid of filter and wrapper approaches. The integration of the features of both approaches helps detect informative genes with high classification accuracy (Guyon and Elisseeff, 2003). Many hybrid filter-wrapper methods have been proposed to select informative genes. However, hybrid methods encounter a number of limitations associated with filter and wrapper approaches. These limitations are identified as follows.

In filtering-based approaches, each filter relies on a different metric of various characteristics, such as distance, probability distribution, information theory. Then, for specific dataset, each filter produces a varying subset of genes from another filter on the same dataset. Accordingly, performance results obtained by applying machine learning algorithm on each subset of genes are also varied (Bolón-Canedo et al., 2012; Seijo-Pardo, Porto-Díaz, Bolón-Canedo and Alonso-Betanzos, 2017; Ebrahimpour and Eftekhari, 2017; Nguyen et al., 2015; Liu et al., 2010). The filter approach that yields best results for a specific dataset, may not do so for another. Indicating that classification performance results are highly variable. In other words, the selected gene subset lacks robustness and predictively.

In wrapper-based approaches, the classification of the gene subset is accomplished through two stages: searching and evaluation. In the searching stage, search-based methods are utilised to generate a discriminative gene subset based on an efficient classifier (Jain and Zongker, 1997). Finding the optimal subset of genes has been shown

to be NP-hard problem (Jain et al., 2017; Narendra and Fukunaga, 1977). Therefore, metaheuristic-based approaches have been implemented as a searching method in wrapper-based approaches such as naturally inspired algorithms (Alshamlan et al., 2015). Several metaheuristic-based approaches have been been applied to solve gene selection problems, such as Correlation-based Feature Selection with improved-Binary Particle Swarm Optimisation (CFC-iBPSO) (Jain et al., 2018), Harmony search with a Markov blanket (HSA-MB) (Shreem et al., 2014), Information Gain and Standard Genetic Algorithm (IG-SGA) Salem et al. (2017), Binary Particle Swarm Optimisation and a Combat Genetic Algorithm (BPSO-CGA) (Chuang et al., 2012) and Genetic Algorithm with Artificial Bee Colony (Alshamlan et al., 2015). However, most of these approaches experience stagnation in local optima caused by complex interactions among genes and large gene search space (Jain et al., 2018; Alshamlan et al., 2015; Xue et al., 2013; Li et al., 2008; Shreem et al., 2014).

1.4 Research Objective

This research mainly aims to propose effective hybrid filter-wrapper methods for gene selection to detect cancer biomarker in certain diseases. To achieve the main goal, this research conducts to achieve the following objectives, which can be seen as follows.

- To propose a hybrid filter-wrapper method by using a proper filter-based approach and a suitable population-based algorithm for the subset generation in the wrapper approach and to solve the gene selection problem.
- To modify the selected filter-based approach by hybridising it with an ensemble of filters approaches, to produce a robust and predictive subset of genes and to

improve gene selection outcomes;

- To improve the performance of the wrapper approach in a way that allows the method to navigate the gene search space effectively. The performance can be improved by:
  - modifying the proposed population-based algorithm by adding extra optimization search operators. This is to further explore the interaction among genes to promote a wide coverage of the gene search space.
  - hybridizing the proposed population-based algorithm with the local search algorithm in order to enhance its local exploitation capability.

1.5 Research Scope

The scope of the study is stated as below:

- MRMR and BA are used as a hybridization method to solve the gene selection problem. Furthermore, an enhancement process is applied to both MRMR and BA to improve the gene selection outcomes.
- SVM classifier is used for gene expression classification.
- Microarray cancer benchmark datasets are used for testing.
- Classification accuracy, the number of the selected genes, the fitness value, sensitivity, specificity, and F1_score are used for evaluation. Moreover, statistical tests are carried out to determine any significant differences in the obtained results.

1.6 Contributions

The research has advanced a number of contributions as follows:

1. Hybrid filter-wrapper method based on Maximum Relevancy Minimum Redundancy (MRMR) as the filter approach and adapted Bat-Inspired Algorithm (BA) as the wrapper approach. The adaptation of the BA was carried out based on the genes, which were selected by the filter approach (i.e., MRMR). The adaptation process involves i) formulating the gene selection problem, ii) adapting the operators of the BA and iii) identifying suitable values for the parameters of the BA. This method is referred to as "Hybrid Minimum Redundancy Maximum Relevancy and Adapted Bat Algorithm" (MRMR-BA).

2. *Modified Maximum Relevancy Minimum Redundancy (MRMR)*: Hybridization of ensemble of filter methods (i.e., ReliefF, Chi-square, and Kullback-Liebler) with MRMR filtering process to improve its robustness and predictively. This method is referred to as "Hybrid Robust Minimum Redundancy Maximum Relevancy and Adapted Bat Algorithm" (rMRMR-BA).

3. *Modified Bat-Inspired Algorithm (MBA)*: The BA was modified by adding extra operators, which were inspired by TRIZ inventive solution to conduct further optimization search into the basic BA. This to further explore the interaction between genes that allow the most promising gene search space regions to be reached and refined. This, in turn, produces better gene selection outcomes. This method is called "Hybrid Robust Minimum Redundancy Maximum Relevancy and Modified Bat Algorithm" (i.e. rMRMR-MBA).

4. *Hybrid Bat-Inspired Algorithm (HBA)*: BA was hybridized with local search algorithm (i.e., β-Hill Climbing) to enhance the local exploitation capability. This method is called "Hybrid Robust Minimum Redundancy Maximum Relevancy and Hybrid Bat Algorithm" (i.e. rMRMR-HBA).

1.7 Structure of thesis

This thesis is organized into nine chapters as follows:

Chapter 2 provide a brief description of the microarray technology, gene selection process and approaches, and a survey of the previous approaches, which tackled gene selection problem. The chapter also discusses the basics of the BA, followed by description of the biological background of BA. The procedural steps of BA are also presented and discussed in this chapter. The chapter concluded with the BA applications and variants, which are provided and discussed.

Chapter 3 introduces the research design or methodology, which is adopted in this research. The chapter consists of five phases; namely, initial phase, preprocessing phase, construction phase, improvement phase, and finally the evaluation phase.

Chapters 4, 5, 6, 7 present the MRMR-BA, rMRMR-BA, rMRMR-MBA, rMRMR-HBA respectively. Each chapter describes a particular proposed method, which solves the gene selection problem. It also presents and discusses the experiments, as well as the obtained results.

In Chapter 8, a comparison between the results of the proposed filter-wrapper methods is made and discussed. The best results, which were obtained from the proposed

methods, are compared with those obtained by other comparative methods in the literature.

Chapter 9 provides and discusses the findings of this research. It also forwards a number of recommendation for further research work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter covers the literature that forms the theoretical background and motivation of the thesis. This chapter also introduces essential background and fundamental of microarray, gene selection process, and gene selection approaches (filter, wrapper, and hybrid). It reviews related work in gene selection using metaheuristic approaches and other approaches. Finally, a comprehensive study related BA algorithm include biological background, procedural steps, applications, and variants.

2.2 Biological and medical background

2.2.1 Introduction

Nucleic acids are the most important molecules in cells. They allow the process of building proteins. They also control the cell life cycle (Watson, 2008). There are two types of nucleic acids: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The functional parts of DNA are responsible for protein synthesis; these small parts are called genes. For instance, proteins play a significant role in cells such as catalysis, defense, movement, protection, regulation, signaling, structural support, transport, transcription, and translation. In order to start the protein synthesis process, gene transcriptions occur and produce mRNA, which is later translated to become a protein (Watson, 2008).

The mRNA amount is a key marker of the cell, where it can be possible to demonstrate facts about the current state of the cells and their activities. Observably, the gene expression patterns of cancer cells are effectively different from that of the intact cells. Microarrays can be used to study how these patterns change (Ochs and Godwin, 2003).

2.2.2 Cancer

Normally, the cell life cycle goes through stages including growth, maturity, division, and death. However, cancer cells are immortal cells and proliferate uncontrollably due to genetic mutations (Schulz, 2005). All cancer cells are characterized by the imbalance of expression between oncogenes and suppressor genes (Ochs and Godwin, 2003; Simon and Dobbin, 2003). These characteristics can be used to identify cancer types. Earlier, only clinical parameters were examined to identify cancers. Later on, however, microarray analysis technique was implemented to study the changes of the molecular characteristics. As a result, the gene expression can be measured by microarray analysis and used to identify cancer subtypes (Schulz, 2005).

2.2.3 DNA

DNA is defined as a double-stranded helix, which is constructed from consecutive nucleotides. Each nucleotide is composed of one of four nitrogen base (A: adenine, C: cytosine, G: guanine, T: thymine), a sugar called deoxyribose, and phosphate group. The double strands are joined together according to base pairing rules (A with T, and C with G) and they store the same biological information. In replication, the two strands separate and run in opposite direction to each other to create mRNA molecules (Berg et al., 2002; Nelson et al., 2008). Figure 2.1 contains a schematic view of the double

DNA helix.

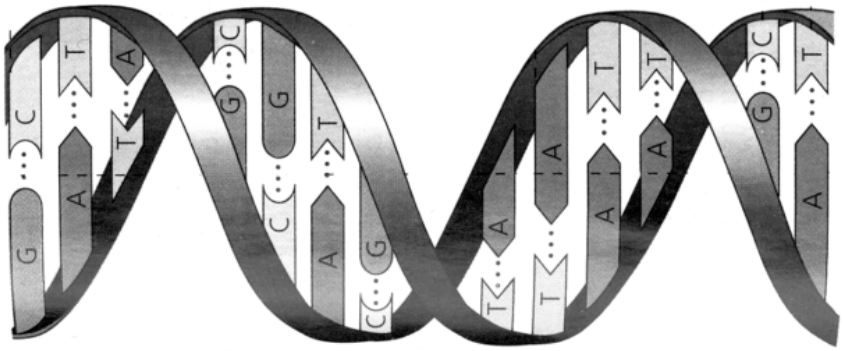

Figure 2.1: The double DNA helix

2.2.4 mRNA

The messenger ribonucleic acid (mRNA) is a single-stranded molecule. It contains a sequence of nucleotides. Unlike DNA, each nucleotide consists of a nitrogen base (A: adenine, C: cytosine, G: guanine, U: uracil), and ribose sugar (Berg et al., 2002; Nelson et al., 2008). mRNA is created through the process of transcription of DNA. For instance, the double-stranded DNA is separated and a protein called RNA polymerase binds to one of DNA strands, and uses it as a template. The messenger RNA is created and separated from the DNA-spiral. Then, DNA-strands bind together again. The amount of mRNA transcription reflects the activation of that gene. Microarray technique is used to probe into target mRNA in order to produce quantitative or qualitative analysis of the current state of the cell (Ochs and Godwin, 2003; OŠNeill et al., 2003). Figure 2.2 below illustrates the process of transcription.

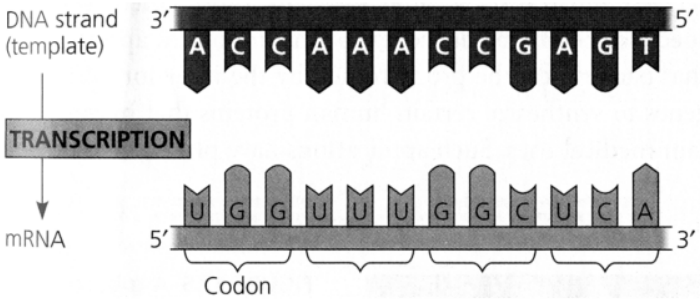

Figure 2.2: Transcription

2.3 Microarray

According to Usmani et al. (2016), Microarray is defined as follows:

"*A microarray is a multiplex lab-on-a-chip. It is a 2D array on a solid substrate such as a glass slide that assays large amounts of biological material using throughput screening, processing and detection methods.*"

DNA microarray (Scena et al., 1995) is the most popular type of microarrays, it is a high-throughput and large-scale technology. It has greatly fascinated the scientific and industry communities. As snapshots of the expression level of thousands of genes are given, DNA microarrays promise new insights into the world of fundamental biology.

DNA microarrays allow the measurement of activity and interactions of thousands of genes simultaneously. This qualifies and enables the technology to perform various scientific tasks, including the identification of co-expression genes, and the discovery of array or gene groups with similar expression pattern. Moreover, the identification of genes with evidently varying expression in term of a set of discerned biological entities (i.e., cancerous tissues), mapping of expression data to metabolic pathways, simulation of regulatory gene networks, etc. On the other hand, they are fast as they produce a

great amount of experiment data that have yet to be discovered by scientists.

2.3.1 DNA Microarray Technology

DNA microarray is a glass slide also called gene chip or DNA chip, which consists of many spots. There are single-stranded cDNA molecules corresponding to one of the mRNA strings that are attached to each spot. The microarray has been devised to measure the level of gene expression. Thousands of genes can be measured simultaneously, where the human genome is believed to have 20000-25000 genes (Alba et al., 2007; Pennisi, 2007). The microarrays are normally prefabricated for specific organisms. For further details regarding how microarrays are made, Berrar et al. (2003) is helpful study in this regard.

First of all, mRNA is extracted from single-type cells. Due to the fact that it is impossible to test absolute values of quantity for a certain mRNA string, the difference between two different samples is examined. In most cases, one cancer test and one reference sample of healthy cells of the same type are adopted. From the mRNA, cDNA is made via reverse transcription and two different fluorescent dyes Cy5 ("Red", for the test sample) , and Cy3 ("Green", for the reference) are attached to cDNA strands. The cDNA strings will be attached through base pairing to the spot at which the complementary probes are fixed. The strings, which are not binding the array, are cleaned. Next, the spot will fluoresce to a certain degree when the microarray is scanned at two wavelengths (red and green) (Simon and Dobbin, 2003; Holloway et al., 2002). Figure 2.3 shows a graphical representation of this process.

The amount of the various spots fluoresce can well indicate the presence of the

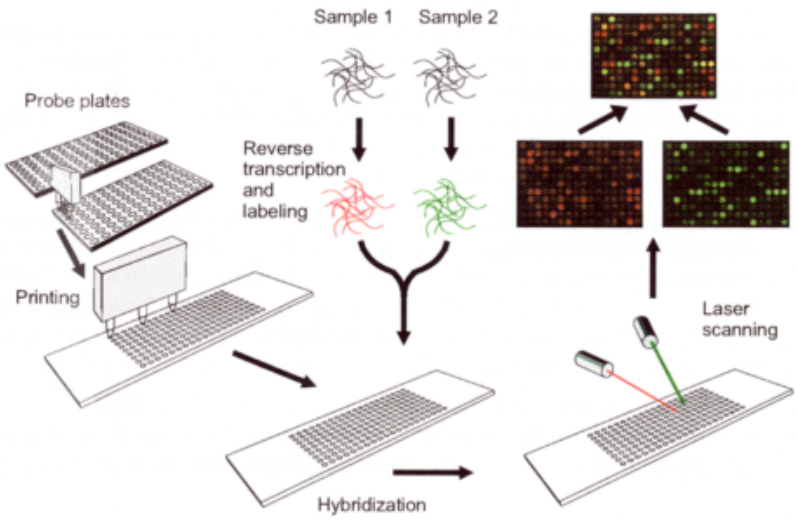

Figure 2.3: Overview of Microarray Technology.

mRNA in the sample. Since RNA has the tendency to degrade soon after transcription, this measurement is a snapshot of the quantity of the mRNA strings in the cancer cell sample in compared with a healthy cell. Therefore, the microarray experiment creates a profile of which genes in the genome are active in a particular cell type under a certain condition, as compared with a reference sample (Simon and Dobbin, 2003; Churchill, 2002).

The two intensities are calculated for each spot in the array. These intensities are proportional to the amount of mRNA in the sample. For the purpose of measuring the relative abundance of a particular mRNA the Cy5/Cy3 ratio is computed for each spot on the array ((Duggan et al., 1999).

$$Ratio = \frac{IntensityofCy5 - BackgroundIntensityofCy5}{IntensityofCy3 - BackgroundIntensityofCy3}$$

The calculation of the background intensity is carried outside the spot and it is an estimation of noise, which is caused by various external factors (like light and reflection) or any strands that are stuck (Rydén et al., 2006). The ratio is a value from 0 to ∞ where values from 0 to1 signify a decreasing expression and values, and values from 1 to ∞ signify an increasing expression rather than the reference. The values are often log2-transformed since it this increases and decreases the values and brings them to similar scales (Duggan et al., 1999; Midelfart et al., 2002). The processing of microarray gene expression data is shown in Figure 2.4.

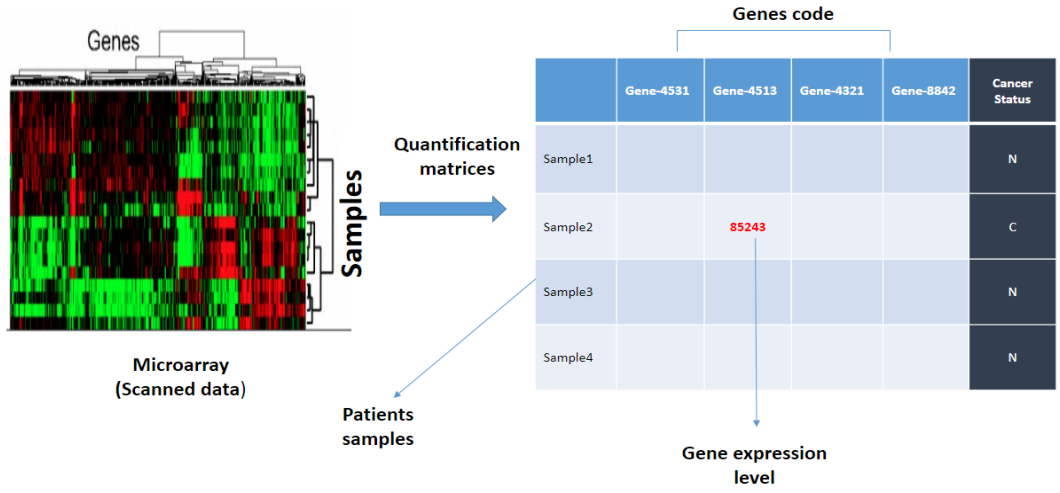

Figure 2.4: Formation of Microarray Gene Expression Data.

2.3.2 Challenges of Analysing Microarray Data

Several machine learning methods serve to analyse the microarray data. A significant amount of new discoveries have been reported. Nonetheless the microarray data have posed a great challenge to computer experts. Specifically, some main difficulties rest in the trait of the microarray data (Zexuan, 2007; Bolón-Canedo, Sánchez-Maroño,

Alonso-Betanzos, Benítez and Herrera, 2014; Jain et al., 2018; Nguyen et al., 2015):

1. Microarray data are extremely dimensional with thousands of genes, but with only tens or hundreds of samples (arrays). This makes it difficult to learn from the data under the curse of dimensionality, where complexity, time of computation and the resource of memory required growing with the dimension exponentially.

2. Microarray data are innately noisy. The natural fluctuations have the tendency to import the measurement variations and implicate the microarray analysis. Additionally, the experiment on microarray exhibits a complex scientific process, where there will be an introduction of errors due to flawed instruments, materials' impurity and scientists' own negligence.

3. The biological heterogeneity is another factor, which serves as a deterrent to the successful data analysis. The gene functional classes show wonderful intra-heterogeneity because of their difference, striking in the derivation organisms and the complex regulation systems.

4. Microarray data normally have genes, which are irrelevant and redundant. They are automatically affecting the learning algorithms' speed and accuracy.

Feature selection, which is also known as gene selection in the context of the microarray data analysis has been introduced extensively to address the issues mentioned earlier.

2.4 Gene Selection Problem

Gene selection refer to the issue of choosing a minimal subset of M genes from the original set of N genes (M<=N) so that the gene space can be reduced in the best way possible, and the learning algorithm's performance is better and is not decreased greatly (Liu and Yu, 2005; Dash and Liu, 1997; Liu and Motoda, 2012). Gene selection can also be defined in the following definitions:

"*Gene selection is defined as a process of identifying certain-disease related genes and finding a gene subset that contains the most discriminative information by removing noisy and irrelevant genes* " (Gheyas and Smith, 2010).

"*A necessary preprocess step to analyze these data, as this method can reduce the dimensionality of the datasets and often conducts to better analyses*" (Talbi et al., 2008).

"*A method for choosing the important subset of genes with high classification accuracy is needed to overcome this challenge. Such method would not only save computational costs, but will also enable doctors to identify a small subset of biologically relevant genes with certain cancers and target only a small number of genes in designing less expensive experiments*" (Li et al., 2008).

Besides reducing the dimensionality of the original gene space, gene selection offers a multitude of advantages (Bolón-Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez and Herrera, 2014; GHAZALI, 2008):

1. Help biologists identify the underlying biological mechanisms, which relates

gene expression to diseases.

2. Reduce cost in clinical settings.

3. Enhance the generalization ability of classifiers.

4. Reduce the training time.

2.5 Gene Selection Process

Gene selection is categorised into four major components namely: subset generation, evaluation function, stopping criterion, and validation procedure. Whole gene selection can be summarised as follows. Subset generation is based on searching techniques to produce a candidate of gene subsets, and each candidate subset is evaluated on the basis of some independent (i.e. without involving any machine learning algorithm) or dependent (the performance of machine learning algorithm) criteria and is continuously carried out until the stopping criteria are fulfilled. The chosen subset is validated (Zexuan, 2007). The general process of feature selection is shown in Figure 2.5. The Gene selection components are thoroughly discussed in the following subsections.

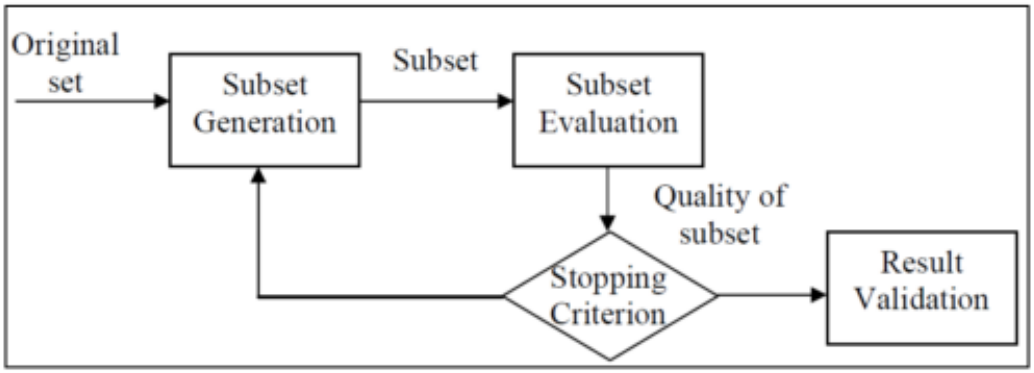

Figure 2.5: Gene Selection Process.

2.5.1 Subset generation

Subset generation is a search process conducted involving the starting point and search strategies to generate a subset of genes for evaluation. In term of the starting point, this search process may start with many strategies. For example, Sequential Forward Strategy (SFS) (Whitney, 1971), starts the search with a new empty set and successfully adds the most relevant genes from the original set into new set sequentially. In contrast to SFS, Sequential Backward Strategy (SBS) (Kittler, 1986) begins with a full set and successfully deletes the most irrelevant genes from the set. Another strategy is bidirectional selection (Caruana and Freitag, 1994), based on SFS and SBS, in which the starting point is located at both ends. In this strategy, genes are simultaneously added and deleted. In addition, a fourth choice of strategy starts the search with a chaotic selected subset based on SFS, SBS, or bidirectional strategy.

In the search strategy, the search space of potential subsets of genes expands exponentially as the number of genes increases. For example, exactly eight subsets (states) exist in the case of three genes (Figure 2.6) (Liu and Motoda, 2012).

Three categories of strategy, namely, complete, random, and heuristic, can be employed to achieve a search task as follows (Dash and Liu, 1997):

In Figure 2.6, the first state (full set) stands for the full subset in which three genes are selected, while the other state (empty set) stands for the empty subset in which no genes are chosen. Conventionally, the generation procedure in selecting a subset of genes from the whole set of genes is classified into three search strategies: a complete search, a heuristic search or a random search (Dash and Liu, 1997), which are discussed

in details as follows:

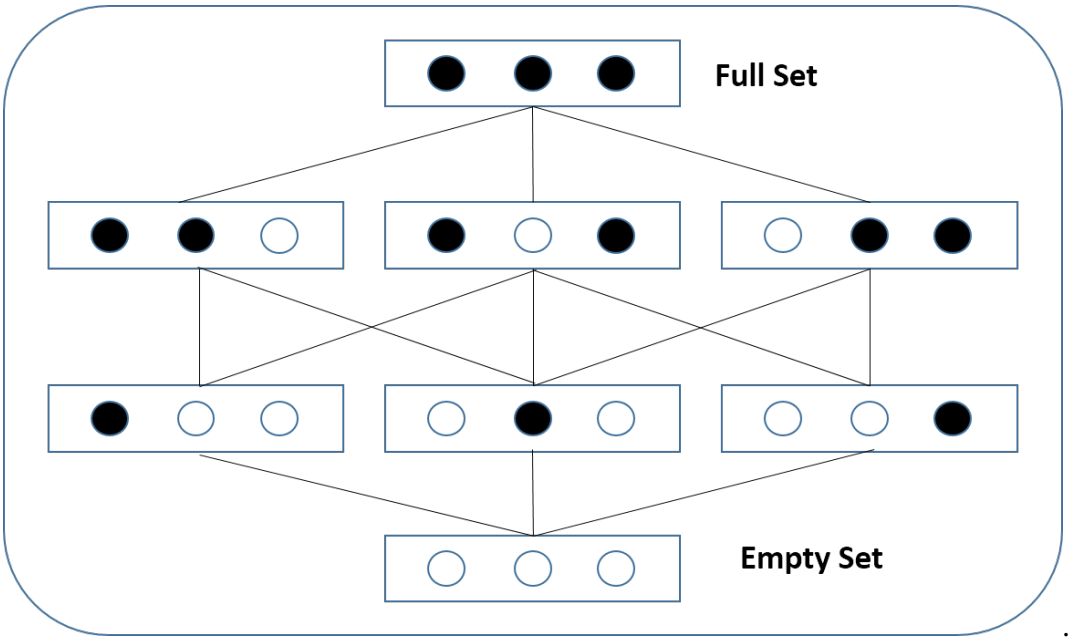

Figure 2.6: Gene selection as a search problem.

2.5.1(a) Complete Search

Complete or exhaustive search is performed to generate the entire possible candidate solutions and thus obtain the best possible subset. In other words, before the final selection is carried out, all 2^N possible subsets in the space must be generated and evaluated. This search method ensures that the optimal subset of genes is produced from the data given. Nonetheless, a complete search is feasible for a small dataset. The work regarded as seminal in complete search is known as the branch and bound method introduced by (Yu and Yuan, 1993). This method performs efficiently a complete search, and terminates the search along a certain branch if a certain limit is surpassed or if a solution is not promising.

2.5.1(b) Random Search

Algorithms involving this approach randomly produce a new subset at each iteration. Despite the fact that the search space stays of 2^N the exact number of subsets considered by the algorithm is controlled by the number of iterations. The reason behind the development of these algorithms is to avoid being stuck in the local minima as in the heuristic search.

2.5.1(c) Heuristic Search

A heuristic search relies on a heuristic approach to navigate a given search space and can be illustrated as a 'depth-first' search guided by heuristics. The cost of a heuristic search may be estimated via a path connecting two ends (Figure 2.6), which may take a maximum length of N. The cost of this process correspond to a path connecting the two ends, which may cover a maximum length of N. The space complexity of this process takes $O(N)$, where N is the number of subsets to be generated. A heuristic search is faster than a complete search because the former searches a particular path only. However, it prone to losing optimal solutions.

2.5.2 Subset evaluation

Identifying the final subset of genes involves selecting the best subset in terms of some evaluation measures. In this evaluation method, a value is fixed to every subset in consideration of the ability to differentiate varying target classes. Various evaluation methods have functioned well in gene selection. These methods can be categorised into five (Dash and Liu, 1997):