Gene Expression Data and Cancer Correlation Analysis by Emerging Pattern Based Projected Clustering

The Hong Kong Polytechnic University

Department of Computing

Yu, Tsz Him

A thesis submitted in partial fulfillment of the requirements for the Degree of Master of Philosophy

Jul 2004

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written nor material which has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_______________________________(Signed)

Yu, Tsz Him____________(Name of student)

Abstract

Cancer studies are one of the hot topics in medical and bioinformatics domains. Scientists are using microarray technologies and data mining techniques to study cancer at molecular levels. Two data mining techniques, namely, pattern mining and clustering, are heavily used in the field of bioinformatics to analyze gene expression data.

In this thesis, the basic problem in organizing the information from the gene expression data in an easy understandable way for the domain experts in the further knowledge discovery process are investigated. We have introduced the Emerging Pattern Based Projected Clustering (EPPC) approach to organize the gene expression data into meaningful clusters. We apply the ideas of the emerging patterns and projected clustering together to form emerging pattern based projected clusters for the biologists. The resulting clusters can be used in the cancer detection problem and the experiment results show that its classification performance is comparable with ORCLUS, the state-of-the-art clustering approach. With its strength in readability, we believed that the resulting clusters are useful for the domain experts in conducting further experiments and studies.

Acknowledgements

I would like to take this opportunity to express my thanks to my supervisors, Dr. Korris Chung and Dr. Stephen Chan, for their fully supports during these years. Korris provided a very stable and reinless environment for me and let me concentrate on my research. With his guidance, I started my research in a right direction and developed the ideas of my works steps by steps. He commented on my works and helped me to dissect the ways that I did not have a clear picture. He is nice, indulgent and supportive all the times. Stephen encouraged me and provided critical comments on my works. He is kind and knowledgeable.

I would like to thanks Dr. Daniel Lee in Department of Applied Biology and Chemical Technology, who taught me molecular biology and biotechnology. I would like to thanks Dr. Vincent Ng in Department of Computing, who commented on my works and provided additional financial support to me. And I want to express my gratitude to all my teachers, colleagues, classmates and all my friends. They provided different supports and cares for me in these years.

Finally, I wish to say sorry to my family. They gave me so much, but I gave so little in return.

Table of Contents

List of Figures

List of Tables

1 Introduction

Bioinformatics has recently become a hot research topic in computer science and molecular biology societies and it is a multi-disciplinary subject involving molecular biology, computer science, mathematics and statistics [1-3]. In this bioinformatics era, there exists huge volume of data but only limited useful information. Molecular biologists and computer scientists are working together as alliances since they are complementary to each other and can speed up the growth of bioinformatics for improving our understanding of nature. However, they are making troubles to each other that fast developments in one field will introduce a lot of works for the other one. The major difference between these two relationships is that alliances communicate and help each other but trouble makers do not. Therefore, a major goal of data miner with computing/informatics backgrounds is how to improve the communication with molecular biologists in the context of bioinformatics. A solution to this problem is to generate understandable data mining results so as to improve the communications with and minimize the workloads of molecular biologists in their bioinformatics knowledge discovery process.

In this thesis, we introduce a new clustering approach, called emerging pattern based projected clustering (EPPC), to assist the knowledge discovery process. EPPC is an integrated approach of emerging pattern mining [4] and projected clustering [5] to obtain easy-to-understand analysis results of high dimensional data. In order to show the effectiveness of EPPC, we apply the ideas of EPPC to the problem of cancer detection and classification.

1.1 Gene expression data and cancer correlation analysis

The studies of cancers have been an important research topic in medical areas for many years. It is important because cancers are still top killer diseases and many people lost their life because of cancers. There is a prediction from National Cancer Institute (NCI) that 500,000 U.S. people will die of cancer yearly [6].

While the causes of cancers are still mysterious, statistical based analysis has been suggested to solve the problem. It is insufficient and time-consuming. Scientists have identified some of the risk factors, such as smoking increases the chance of developing lung cancer, based on statistical based studies. These studies suggested the ways to prevent cancers but there are cases that cancer may still develop even all risk factors are absent and hypothesis from those observations do not hold all times. On the other hand, the lag time of such kinds of statistical based studies may be very long. An example given by NCI about smoking and lung cancer relationship has a lag time of 20 years [7]. In other words, we may not find out those side effects about smoking at very beginning. Therefore, new methods for cancers related studies are needed.

Although cancers cannot be cured completely at this stage, treatments to earlier cancers are always more effective. An example in [7] shows that the five-year survival rates for the stage I and stage III melanoma are around 90% and 20% respectively. It clearly shows that cancers are easier to be cured if they are detected and treated at the earlier stage. However, earlier cancers has no symptoms [7] and most of the patients only have medical check up until they feel pains or significant changes are found in their bodies. Most of the time when cancers are found, it is late.

Screening methods [6], such as cancer imaging and pap tests, were introduced to detect earlier cancers without symptoms. However, those existing screening methods are disease specific and making uncomfortable feeling to patients. Moreover, they have their own limitations.

Cancers are cells losing growth control that is originally controlled by genes. Those abnormal growth tissues invade surrounding tissues and cause functional damages of organs. Microarrays are now used to measure those cellular activities during protein synthesis and gene expression data from these experiments provides the opportunities to detect earlier cancer at molecular levels. Scientists found that the context of expression levels of cancerous and normal tissues are not the same. Therefore, large scale comparative studies of gene expression profiles have been conducted. Using gene expression data becomes a new direction in cancer detection and it requires the use of modern bioinformatics techniques, such as data mining, for efficient analysis and assistant of any further discovery.

1.2 Problems and objectives

Cancer detection and classification using gene expression data is one of the new research directions in bioinformatics and data mining originally designed to discover knowledge and information from data is now heavily used in this problem. In general, cancer detection using gene expression data and data mining techniques can be divided into two sub-problems:

~ How to classify the tissue samples correctly?

~ How to provide easy understandable result for molecular biologists to conduct further investigation?

In order to classify the tissue samples correctly, we need to tackle the specificity of gene expression data that has created certain challenges to existing data mining algorithms. The major problems in handling gene expression data include:

- The high dimensional gene expression data is not easy to manipulate and understand.
- There exist limited records of gene expression data and they are typically not sufficient to approximate the real world.

In data mining, there are two major approaches employed for classification problems. They are pattern based approach and clustering based approach. However, they are different when coping with gene expression data and are not easy to generate understandable results for molecular biologists to carry out further investigations. Specifically, we have to address the following questions:

- How can we help the user to understand large amount of patterns found by pattern based approach in an easier way?
- How can we improve the understandability of clusters by grouping samples according to biological meaningful information instead of using distance measure without reason behind?

In this research, our aim is to introduce a new data mining approach for effective knowledge discovery in bioinformatics databases. To demonstrate its effectiveness, we try to apply the proposed approach to the cancer detection and classification problems. To achieve this aim, the following issues are addressed:

~ Tackle the curse of dimensionality problem of high dimensional gene expression data using the idea of projected clustering

~ Make use of the easy-to-understand patterns (domain knowledge) extracted from gene expression data to organize the gene expression data into manageable number of clusters in order to minimize the effort in analyzing the huge amounts of patterns and enhance the biological meaningfulness of clusters at the same time

~ Investigate how the pattern mining approach and clustering approach, i.e., emerging pattern mining and projected clustering, can be integrated to combine their strengths and compensate their weaknesses.

~

1.3 Organization

This thesis consists of six chapters. In chapter 2, we look into the bioinformatics researches from the data mining perspectives. The cancer detection problem is highlighted. In chapter 3, we review the two data mining techniques that designed for high dimensional data. They are the emerging patterns and projected clustering techniques (ORCLUS). In chapter 4, we introduce the concept of emerging pattern based projected clustering (EPPC), the problems in mining EPPC and the framework for mining EPPC is introduce. We evaluate the performance of EPPC in cancer classification problem in chapter 5. The final chapter concludes the thesis and outlines future works.

2 Literature Reviews

In this chapter, we are going to "look into bioinformatics from a data mining perspective". Bioinformatics is a newly established research discipline and data mining techniques are becoming one of its most popular ingredients. The major reason for this merge is that existing problems in the field of bioinformatics are closely matched with the basic ideas/assumptions of data mining. In this chapter, we will go through the backgrounds of bioinformatics and briefly review the existing data mining techniques for bioinformatics and cancer detection applications.

2.1 A brief look into bioinformatics

The field of bioinformatics has experienced an explosive growth in recent years and there are many novel methods available now. It is very difficult to predict the growth of this young field. But we all sure that bioinformatics has already undergone its dark age, the embryonic stage, and it will grow rapidly in the near future. In this section, we will look at its development history, motivations, definition, aims, possible research directions and one of its applications – cancer detection.

2.1.1 Its development history

Molecular biology and computer science researches have been carried out separately at the beginning, but some of the most fundamental problems in the molecular biology seem to be appropriately addressed by computer science techniques. In late-1960s, researchers started to combine the computational information from computing techniques and the experimental information from molecular biology laboratory together to provide better understanding and new insight about macromolecules, genes and proteins. This era can be considered as the birth of computational biology. In 1970s, the computational requirements for the field became solid and computation methods are being used in the problem of sequence alignment, evolutionary tree analysis and construction, prediction of protein structure, protein folding problem and so on. In 1980s, the field of computational biology was dominated by sequence analysis, molecular databases, protein structure prediction and molecular evolution that were almost hopeless to be solved without the aids of computer. In 1990s, breakthroughs in hardware, database, internet and various computational technologies support the rapid development of the field and today it has already become an independent discipline with its own problems and achievements and called bioinformatics. Details about the development of bioinformatics can be found in [8].

At the time now, the coverage of bioinformatics may have unpredictable changes, but its skeleton has become solid and there are two promising trends we can observed. First, the diversity, volume and complexity of information available are increasing. Second, its dependency on the informatics techniques is also increasing. Since existing informatics techniques are not designed for bioinformatics originally, it is indispensable to design new techniques or tailor existing informatics techniques to meet the needs arisen from bioinformatics problems.

2.1.2 Its motivations

The motivations of bioinformatics come from both molecular biology and computer science domains. In the molecular biology area, there is a flood of new data and new problems. In the computer science area, improved computation power, disk storage capacity, database, internet technologies and various types of algorithms and analysis methods provide opportunities to solve those complicated problems.

Advancements in biotechnology, such as the invention of microarray technologies, have created the floods of new data and problems in field of bioinformatics. Scientists started to open the molecular black box within cells. They are not limited to study those large cell components by using electronic microscopes and it is possible to discover the secret of life at molecular level. Various types of new and complex biological data, such as raw DNA and protein sequence, macromolecular structure, genomes, gene expression, literature and metabolic pathways data [1], are now available and they are being produced at a phenomenal rate. For example, the GeneBank repository and SWISS-PROT database are doubling their size in every 15 months [1]. The flood of data not only provides opportunities in different research topics and also promotes the needs of using computer technologies.

Fortunately, the improvements in CPU, disk storage, database and internet technologies were also rapid in this decade. Research works requiring high computational power and large disk storage, such as matching between DNA sequences, become possible. The high connection speed of internet and advanced databases technologies also help scientists to access valuable data and summit new entries in anywhere around the world. Breakthroughs in computing technologies, such as new algorithms in data mining and machine learning areas, are essential to support and speed up the developments of molecular biology since those biological data are often too large and too complex to be manipulated by human. It is also important in supporting the rapid growth of bioinformatics communities.

2.1.3 Its definition & aims

Since bioinformatics has become an independent discipline, pioneers are trying to draw a clear picture for it by proposing a formal definition in recent years. Most of them [1-3] define bioinformatics as the application of science of informatics, including mathematics, statistics and computer science, to molecular biology. Luscombe [1] has proposed a very detail definition for bioinformatics, highlighting some of the most important aspects in this field. His proposed definition is quoted below.

> *"Bioinformatics is conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying **"informatics techniques"** (derived from disciplines such as applied maths, computer science and statistics) to **understand** and **organize** the **information** associated with these molecules, on a **large scale**. In short, bioinformatics is a management information system for molecular biology and has many **practical applications**."*

According to Luscombe's definition [1], the aims of bioinformatics can be summarized into four aspects. They are organizing data, developing intelligent tools using informatics techniques, analyzing data and interpreting results in biological meaningful manner and using the newly discovered knowledge in practical applications.

The first aim is perhaps the simplest one focusing on the organization of data for the researchers to access available information and to submit their new data efficiently. All breakthroughs in disk storage, database technology, internet technology have make contributions for this aim. For example, large sequence database called GeneBank which contains more than 22 millions sequence records is now freely accessible for the researchers around the world through the Internet. Scientists can search for the DNA sequences in it and submit new DNA sequences that they found. On the other hand, some researchers are focusing on the challenges come from the legacy data, and the heterogeneous, complex and geographically dispersed nature [9] of different data sources.

The second aim of bioinformatics is to develop more intelligent tools by using informatics techniques for scientists to analyze available data. Tools are needed to be efficient in front of the large amount of data. We also need intelligent tools like sequence searching tools called FASTA [10] and PSI-BLAST [11]. They are not just a simple text-based search tool using heuristics to solve the string matching problems efficiently, but also consider matches with biologically significance. Therefore, researchers who carefully model and develop appropriate algorithmic techniques for developing biologically intelligent and scalable tools are always expertise in computational theory, as well as having thorough understanding in biology [12].

The third aim is to use those intelligent tools to analyze the data and interpret results in biological meaningful manner. Biological studies are not limited to each individual biological system by using traditional laboratory experiments. Bioinformatics allows the global analysis for the available data to uncover new principles that may apply across many biological systems using informatics techniques. In recent years, scientists are not limited to study only few genes in their life time, but they can use microarray data to study relation between thousands of genes simultaneously. Tools like laboratory management information system (LIMS) [9] are always needed and they are typically integrated with intelligent analysis and visualization tools to form a platform for bioinformatics researches. Scientists may use such a research platform to incorporate biological models and domain knowledge to perform further studies and simulations in order to discover those common principles between different biological systems.

The forth aim is to use those newly discovered knowledge in practical applications. For example, the specific gene expression patterns found in cancerous tissues through bioinformatics studies can be used to improve the cancer diagnosis and improve the treatment plans [13]. Moreover, those newly obtained patterns may also provide valuable directions for scientists to conduct any further studies. For example, genes that are contained in a specific pattern discovered in cancerous tissues are likely to be correlated and they open up further research opportunities in understanding the cancer developments.

The ultimate goal of bioinformatics is likely to be the understanding of nature [12]. Integrative genomics [14], translational genomics [14], system biology [15] are being promoted by the pioneers as the coming future in bioinformatics. These researches focus on understanding the relationships between different types of data, different species and different biological system models and they are trying to discover the secret of nature from different levels and perspectives. In order to reach this ultimate goal, it is very important to understand the knowledge embedded in those available data thoroughly and there is a definite need in developing intelligent tools that is capable to provide easy understandable and usable form of biological meaningful information for further knowledge discovery at this time.

2.1.4 Its research directions

There are some common practices among scientists. It is because most of their ultimate goals are closely related to improve the understanding of nature even researchers are proceeding their own researches in various levels by using different types of data that are diverse in size and complexity. For examples, there are studies in separating coding and non-coding regions using raw DNA sequence data, studies in correlating expression patterns using gene expression data and so on [1]. Similar to the research methodologies in other disciplines, studying bioinformatics cannot leave out steps like choosing target data for a specific topic, understanding and organizing the data, applying suitable informatics techniques, analyze the results and formulate biological hypothesis for biologists to undergo further studies.

Although the steps of researches are almost the same in the field of bioinformatics, pioneers are having different views on the research directions for bioinformatics. For example, Molidor [16] said that bioinformatics should start from sequence analysis to gene expression data analysis, then it comes to the age of integrative and translational genomics and finally it opens the opportunities in personalized medicine; Durand [14] stressed the necessity of an integrated platform that provides fully integrations on heterogeneous and distributed data sources for easy comparisons between different analysis approaches; Yao [15] forecasted the future of system biology by illustrating the migration of the field from data-driven approach to model-driven approach; Wiemer [3] and Martin-Sanchez [17] illustrated the integration possibilities with medical informatics.

Luscombe has summarized his point of views into a two dimensional bioinformatics spectrum [1] that is easy to understand. The vertical axis of the spectrum demonstrates how the depth direction of bioinformatics can aid rational drug design to minimize the effort in biology laboratory. It starts with simulations by using a single sequence with the gene findings, structure prediction and miscellaneous algorithms to find the corresponding protein structure, force field at the protein surface and finally paving the ways for drugs design. For the horizontal axis of the spectrum, the breadth direction highlights how the biological data and informatics techniques have broadened the scope of biology studies. In this dimension, comparative analysis is always used to discover those hidden principles or novel patterns between data in large scales.

As mentioned previously, there is a definite need to develop informatics tools that can provide biological meaningful results and those results are needed to be easy-to-understand. In addition, it is not limited to use data driven approach in bioinformatics researches. Model driven approach that incorporates available biological knowledge or knowledge from other disciplines to discover the new biological knowledge may have potential advantages. Developing tools that can make use of the domain knowledge in knowledge discovery and generate biological meaningful and easy-to-understand results is one of the promising trends in bioinformatics.

On the other hand, the bioinformatics spectrum [1] has generalized those major research topics in the field of bioinformatics into two directions, i.e. the depth and the breadth directions. The depth direction focuses on the inference process of new biological knowledge which is always facilitated by the informatics techniques like case-based reasoning. On the other hand, the breadth direction focuses on the comparative analysis of biological data and techniques like data mining are deemed appropriate. Therefore, employing data mining and incorporating useful biological knowledge for large scale comparative studies is one of the promising trends in the coming future.

2.1.5 Its application: Cancer detection

Cancer is referred to a group of diseases related to cell growth through which the disease can spread throughout our body [7, 18]. It arises from the loss of growth controls in cells and those cells keep dividing even new cells are not needed. These extra cells form a mass of tissue and they are classified as cancer if they can invade and damage nearby tissues or organs. Scientists are trying to determine the possible causes of cancer and they have already identify some risk factors [7, 18], such as smoking tobacco, but there are still many people getting cancer with the absence of all known risk factors. It shows that our knowledge about cancer is still limited and the development of cancer is interpreted as a result of complex mix of factors related to lifestyle, heredity and environment [18].

Cancer detection is important because it is one of the top killer diseases. Our knowledge in cancer development is still limited and the number of new patients is increasing. In 2004, an estimation of 0.5 million patients in United States will be die of cancer [6]. Cancer detection is important because cancer is easier to be cured if it is detected and treated in its early stage. For example, the five-year survival rate for patients with Melanoma in stage III is just around 20% but it is around 90% for stage I cases [7]. However, cancer detection is not easy since early cancer may not have any symptoms. Most of the patients only visit their doctors when they noticed changes occurred in their body or felt pain. When such changes are noticeable, the development of cancer is no longer at its early stage. For these reasons, improving the cancer detection methods should be at high priority for cancer researchers.

In order to detect the earlier stage cancer, different screening techniques have been introduced but they have their own limitations. Screening techniques are now available to check for some of cancers in a person who does not have any symptoms and the estimates of deaths that can be avoided through screening vary from 3% to 35% [6]. It is because treatments for cancers are much easier in earlier stage. Cancer imaging, such as X-ray imaging and CT scans, is used to check any suspicious areas or abnormalities of a person that might be cancerous. However, some of the imaging procedures may be uncomfortable or require patients to stay in a small space for some time [19]. Exposure to more X-rays or radioactive substances are required. Some of the methods may require the injection of contrast agents, steroids or histamine blockers and they may cause discomfort or allergy [19]. Laboratory tests, such as blood and urine tests, pap test, are medical procedures that samples of blood, urine, or other tissues or substances in the body are checked for cancer and test values are usually matched with some normal ranges. However, many factors will affect the results of those tests. For example, sex, age, race, medical history, general health, specific foods, drugs and even how closely the patient follows the pre-test instructions will vary the testing results [20]. All laboratory test results must be interpreted in the context of the overall health of the patient and are generally used along with other exams or tests. So, family doctor who is familiar with the patient's medical history and current health conditions is critical for the accuracy of laboratory tests [20]. Sometimes, interpretations may be a bit subjective. Last but not least, most of the existing screening methods are specific to certain types of cancers and there are no effective screening methods available for some cancers [6].

Cancer genetics is rapidly expanding and DNA-based testing can be used to confirm a specific mutation as the cause of the inherited risk in developing cancer [7, 21]. Mutations related in cancer development can be detected by using DNA-based testing related data, such as microarray profiles of normal and cancerous tissues, in comparative analysis studies. The great differences in gene expression profiles obtained by microarray between the normal and tumor tissues are not limited to be used as a screening procedure and may also be employed by the physicians for cancer diagnosis [22]. Moreover, this method can be applicable to all cancers and therefore it is especially useful for those cancers that have no effective screening method currently.

Although cancer detection can benefit from the comparative studies in gene expression data, the gene expression data itself is complex to understand. In recent years, researchers have paid great efforts in this topic. Details about gene expression data and some of the recent works in bioinformatics data mining are reviewed in section 2.2.2 and 2.2.3 respectively. In general, bioinformatics techniques are important for us to develop some effective methods that can incorporate the biological knowledge and provide easy-to-understand and biologically meaningful information for cancer detection problem using gene expression data.

2.2 Use of data mining in bioinformatics

The field of data mining is evolving in this decade along with the rapid growth in our data generation, data collection, data warehousing and distribution abilities. Huge scientific databases and different types of data are waiting for data miners to discover new knowledge. Bioinformatics provides new types of data and applications for data miners and computer scientists are using different data mining techniques to solve the problems. In this section, motivations and aims of using data mining in bioinformatics, nature and challenges of gene expression data, data mining techniques used for the problem of cancer detection with gene expression data are reviewed.

2.2.1 Motivations and aims

The popularity of data mining is mainly due to the fact that the growth of our capabilities in discovery useful knowledge cannot catch up the advancements in data generation, collection, warehousing and distributions abilities. Huge amounts of available data has far exceeded our ability for comprehensions and it is considered as a "data rich but information poor" situation [23]. Therefore, data mining which refers to a class of methods that are used under some computational efficiency limitations in the knowledge discovery process [24] becomes one of the essential components in discovering new knowledge.

In the mean time, the advancements in biotechnologies, such as DNA sequencing, gene expression profiling techniques, help molecular biologists to generate huge amounts of biological data and the volume of data is challenging biologists for conducting further investigations. For example, sequence databases, GenBank and SWISS-PROT, are doubling their size almost yearly [1]. New types of data provide either different scale or granularity of researches or totally new research opportunities. For example, gene expression data provides opportunities for biologists to study large number of genes at one time instead of just focusing on few genes in the past. It comes to a "data rich and information poor" situation in the field of bioinformatics now and it demands on new methodologies for in-depth analysis. Therefore, data mining techniques that are motivated by similar challenges and have already successfully implemented in other domains, such as discovering consumers' behavior, are naturally be one of most appreciated informatics techniques for bioinformatics studies.

The aims of employing data mining techniques in bioinformatics are similar to other types of data, such as sales patterns analysis, that assists domain experts for efficient in-depth analysis. In general, it is used to transform biological data and observations into structured and meaningful information that scientists can access, visualize and understand easily [25]. Moreover, it should be capable to provide various degrees of details and different viewpoints of knowledge [25] and improve the drug target discovery, diagnostics, design of treatment plans and so on [26]. Finally, it should be capable to help scientists to discover the connections between different biological systems [25] and ultimately help us to understand the nature [12]. In summary, data mining is useful to extract information from biological data for various applications and further knowledge discovery in bioinformatics and the fundamental scope of using data mining in bioinformatics is to improve the readability and understandability of data.

2.2.2 New data, new challenges: Gene expression data

In this bioinformatics age, different types of new biological data and research topics are evolving. A precise list of them can be found in [1]. Among them, gene expression data is one of the popular types of data in both bioinformatics and data mining communities. It is because the gene expression data can provide different research opportunities in cancers, such as cancer detection and diagnosis mentioned in section 2.1.5. In this section, the backgrounds of gene expression data are first introduced, its features and its challenges for existing data mining techniques are also discussed.

“Gene expression” is the term used to describe the transcription process of the information contained within DNA into mRNA during the protein synthesis process. Genes are said to be “expressed” if they are turned on during protein synthesis and the amounts of mRNA produced are measured as the gene expression level [22]. By using the microarray containing many DNA samples, scientists can determine the expression levels of thousands genes in a single experiment by measuring the amount of mRNA bound to each spot on the array [27]. The details of microarray technologies can be found in [22, 27].

Every cell of our body contains a full set of chromosomes and identical genes and only a fraction of genes are turned on [22]. Scientists found that the contexts of gene expression are different among different types of tissues. For example, brain and muscles tissues are different in context [27] and they are also different under various conditions, such as under the influence of drugs [27]. Therefore, gene expression data is widely used in cancer detection studies since the expression profiles of normal and cancerous tissues are having significant differences.

Gene expression data is high dimensional in nature. Its high dimensionality is inherited by the physical properties of microarray technologies. For example, a single microarray experiment can examine 40,000 genes from 10 different samples under 20 different conditions and produces at least 8,000,000 pieces of information [27]. In addition, gene expression data contain relatively limited records when compared with the number of attributes. If genes are interpreted as attributes in the gene expression data and the attribute to samples ratio will be very large in the above example. The gene expression level is measured by the amount of mRNA in each spot of microarray labeled by fluorescent dyes [27] and the measured intensity of the dyes is numeric in nature. Since microarray experiments are still expensive in terms of time and cost, most of such experiments have been conducted to investigate preferred biological or medical significant properties, such as for cancer diagnosis purposes. Therefore, most of the gene expression data contain predefined class labels with important biological meaning.

The major challenge in studying gene expression data comes from its high dimensionality. Most of the existing data mining algorithms are computationally infeasible for data with high dimensionality. For example, the Apriori algorithm is almost computationally infeasible to mine long gene expression patterns since the number of candidate itemsets is too large in high dimensional gene expression data while the k-means algorithm may not form meaningful clusters because of the sparseness of data [5, 28, 29]. On the other hand, gene expressions always have large number of attributes but limited number of records. It is difficult for data mining algorithms to obtain good approximation for the real world and the models learned may not be robust for the new data. The readability and understandability of the mining results are also very important for gene expression data and some of the data mining techniques are weak in this issue. In the case of cancer diagnosis, it is very difficult to get the trusts from users, i.e. the medical officers, if class label is assigned with a numerical score only.

2.2.3 Techniques for cancer detection using gene expression data

Data mining is one of the appreciated techniques for bioinformatics with tones of published literatures. In section 2.1.5, we have discussed the cancer detection problem briefly and we have illustrated the potential of using of gene expression data in achieving high accuracy cancer detection. Then, we have studied the properties of gene expression data and the challenges for existing data mining algorithm in section 2.2.2. In this section, some popular data mining techniques for gene expression data in cancer detection problem are discussed below. In most of the literatures, cancer classification is used instead of cancer detection. They are more or less the same in technology basis but cancer classification is just a broader topic that its scopes also include the identification of cancer types and cancer subtypes. In this section, we review the techniques used for gene expression data and this two terms are considered as equivalent.

2.2.3.1   Naïve Bayes method

Naïve Bayes (NB) method uses probabilistic induction to assign the class labels to testing samples. It assumes the attributes in samples are conditionally independent given the class label and models each class as a set of Gaussian distributions. Each gene in the training data form Gaussian distributions for every class and the class label is assigned to the data instances with maximum probability.

NB method is simple to use but it has two major limitations. First, it assumes genes in samples are orthogonal to each other but this assumption seems not close to the truth [30]. Scientists are interesting to study the genes interaction and most of them believe there should be correlation among genes. The above assumption may provide inaccurate classification and it is incapable to discover biological information, like genes interaction. Second, it assumes data are in Gaussian distribution [30]. It is quite restrictive to use and the limited size of available gene expression dataset is very difficult to determine their distributions followed. Thus, its performance is limited because of these two fundamental assumptions.

2.2.3.2 Artificial neural networks

Artificial neural network (ANN) consists of basic units, called neurons, that simulating those biological neurons in our brain. Each neuron has multiple inputs and single output. They are connected to each other with a weight and form a network. The topology of ANN is problem dependent, but in general it has number of input layer neurons equivalent to the number of available attributes in data, number of output layer neurons equivalent to the number of available classes and neurons in hidden layer. In the training phase, the ANN is first trained with training samples by using a learning algorithm. The learning algorithm adjusts the weights in those connections between neurons and the learned information is existed as the patterns of connection weights in ANN. In the testing phase, a testing instance is input through the input layer and started to evaluate layer by layer. The output neuron with the maximum value in ANN indicated the corresponding class label for that testing sample.

ANN is able to handle many interacting variables and non-linear behavior of the tissue samples and class labels [31]. It also provided comparable results with other methods [30] that shown in the literatures. However, it performs classification in a black box manner [30]. Information learned during the training phase may not be extracted into rules or any easy understandable formats easily [31], so its benefits to bioinformatics become limited. Moreover, the risk of premature convergence and the problem dependency of choosing suitable topology and learning algorithm make ANN is very difficult to obtain a optimal network [31].

2.2.3.3 Decision trees

Decision trees also called as classification trees. It consists of a set of internal nodes and leaf nodes. The internal nodes are splitting criterions that comprise of a splitting attribute and predicates defined on this attribute. The leaf nodes are single class labels. In the tree construction process, the entropy based measure is always used to determine the best attribute with maximum information gain for splitting. This process is a top-down recursive divide-and-conquer process [23] and it stop if all samples are belongs to some classes or no remaining attributes available for further separation. Sometimes, tree is pruned by using heuristics to avoid overfitting problem.

Decision tree is an attractive approach in bioinformatics is mainly due to its readability. The result of decision tree are very interpretable [30] and understandable knowledge in forms of hierarchical trees or sets of rules extracted [31]. Those trees and rules are provided valuable information for the scientists for further studies. On the other hand, it do not need to provide any parameters and its construction is relative fast [30]. However, the resulting tree is always error-prone when the number of training examples per class is small [31], the robustness of the decision tree in front of new data is questionable for small gene expression dataset.

2.2.3.4 k-Nearest Neighbour

k-nearest neighbour (k-NN)is one of the similarity based methods. It tried to find the most similar set of training samples for the testing sample and use majority of the voting to determine the class for the testing sample. The distance metric that it used can be any similarity measure based on attributes' values. The most common similarity measures are Euclidean distance and Pearson correlation.

k-NN is relatively less prone to noise and bias in the data [30, 32] since the testing samples is evaluated by set of instances and it does not have the problem of repeat the training when new data is available. However, it is not scalable because the computation of similarity is very expensive if the dataset is large [30]. On the other hand, the similarity measure in high dimensional data, such as gene expression data, may suffer from the curse of dimensionality problem and affect the accuracy.

2.2.3.5 Association rule based classifier

CBA (Classification Based on Association) classifier is one of the successful application methods in using association rule in the classification problem. The basic idea of CBA is to extract a special type of association rules, class association rules (CARs), and use them in classifying testing samples. The major difference between class association rules and general association rules is that its consequence is class label, but there is no such restriction in a typical association rule. In the classification process, set of CARs that satisfied the minimum requirements of support and confidence are selected. Then, the best rule with highest confidence and support is used to classify the testing instances. The class label of the best CAR whose antecedent is contained in the testing instances is then assigned.

CBA classifier that incorporating the idea of association rules into the classification problem achieved high classification accuracy and the major advantage for gene expression studies is that the classification results are easy to understand. However, the large set of discovered rules is always a problem and it is more serious for the gene expression data that is high dimensional in nature. In general, discovering such a huge set of rules is computational expensive and it is difficult for the scientists to find a good target for further studies. Increasing the support threshold, or confidence threshold, or both may reduce the number of CARs found but it will suffer the risk of dramatically degrading in classification accuracy and it is not an ultimate solution.

2.2.3.6 Cluster based methods

Grouping similar objects is one of the most basic abilities of human being [33]. We always found that many individual objects may have properties in common and the basic idea of cluster based method is to group the training samples into different clusters based on different similarity or distance measures. For example, early people can classify plants into groups, such as poisonous or edible by using their appearance, their color or their tastes. Those cluster representatives, such as the centroids of clusters, are used to represent large number of instances that originally difficult for us to manage and understand. To form clusters from data, it is needed to define some effective measures, called objective functions, to evaluate the similarity or distance between objects. In the classification process, the similarities or distances between the testing sample and clusters representatives are measured and then the testing sample is assigned to its closest cluster.

There are many well known clustering algorithms, such as k-means and fuzzy c-means, are shown to be useful in organizing gene expression data into clusters for the classification of cancer. They are generally less prone to noise and bias in the data and information organized into manageable number of clusters are always easier for scientists to choose the appropriate target for further investigations. However, most of the clustering algorithms form clusters by using the tightness of data point and it is often lack of practical meaningful support, not easy to understand and not easy to be applicable in biology related domain. On the other hand, most of the clustering algorithms suffer from the curse of dimensionality problem for high dimensional gene expression data and meaningful clusters are not easy to obtain. Feature selection is often used to prune useless features before forming reliable clusters, but it is still not possible to prune off too many feature without information loss [5, 28].

2.3 Summary

In this chapter, we started from the history of bioinformatics and followed by its definitions, motivations and aims. We also discussed some of research directions for the field in general and stressed on the hot topic, cancer classification using gene expression data. From the data mining perspectives, the problem of cancer detection by gene expression data comparative analysis motivated the development of new data mining algorithms. The new data mining algorithms needed to solve the challenges come from the gene expression data, such as high dimensionality and limited records. In reviewing and appreciating those existing data mining techniques applied to the gene expression data cancer detection problem, understandability is always the most important requirement in bioinformatics communities.

3 Projected clustering and emerging patterns

Clustering and pattern mining approaches are two main streams in data mining domains for years and there are tones of algorithms proved to be applicable for different type of data effectively. However, most of them do not work well for high dimensional data, such as gene expression data mentioned in the previous chapter. Those traditional algorithms are challenged not only by the volume of available of data and also challenged by their complexity as mentioned in Section 2.2.2. In recent years, new clustering and pattern mining algorithms are developed to tackle the problems raised by those high dimensional data.

In this chapter, we will review an outstanding clustering and a new pattern mining algorithm, projected clustering and mining emerging patterns, in details. First of all, the motivation, assumptions, objectives and definitions of these two algorithms will be highlighted. Second, their own problem statement will be stated. Finally, we will comment on their strengths and weaknesses.

3.1 Projected clustering

One of the most important breakthroughs of the traditional clustering approach is the concept of the projected clustering introduced by Aggarwal in 1999 [5]. Projected clusters are defined as subset of data points whose distance between themselves within the cluster is minimal in the corresponding subspace of dimensions. Projected clusters can capture sets of closely related data points in different subspaces. It is believed that set of data points are closely related to each others in their own subspace and the projected clusters can offer experts useful and realistic new insights for the knowledge discovery purpose with additional subspace information.

3.1.1 Introduction

Advancements in data collection and storage technologies provide new opportunities to have data not only in large volume but also in high dimensionality. Such as the microarray technologies introduced in Section 2.2.2, it generates data with thousands of dimensions. It is promising that our ability in data generation will continue to improve in the coming future and there is a need to focus on development of new data mining techniques specified for high dimensional data.

The major reason leads us to focus on the concept of clustering using feature subspace is that almost all well known clustering algorithms, such as k-means, trends to break down in high dimensional feature space [5]. The quality of those clusters obtained by traditional clustering algorithm using full feature space degrade too fast while the dimensionality increase because of the inherent sparsity of the data [28, 29] and it is referred as curse of dimensionality problem in the literature [5, 28, 29]. Theoretical results have shown that the distance between every pair of points in high dimensional space are nearly the same [34]. Therefore, the meaningfulness of those resulting clusters are now being questioned [28, 29].

Feature selection is well adopted approach to reduce the dimensionality of data but it may not be sufficient for all situations. Example from Aggarwal [5] illustrated that feature selection is not good enough to solve the problem in forming clusters in three dimensional space effectively by discarding any features if different set of points are correlated with respect to different set of features. In such case, applying feature selection to form clusters may suffer from unpredictable information loss. Aggarwal [28, 29] commented that feature selection may not always be feasible to prune off too many dimension without information loss.

Therefore, the idea of projected clustering was introduced. It is a redefinition of the clustering problem with special consideration in the relationship between resulting clusters and their corresponding feature spaces that targeted to minimize the information loss. In short, its objective is to group data points with high dimensionality into clusters under different feature subspaces.

In the projected clustering problem, there are some basic assumptions [5, 28, 29]. It is assumed that not all the available dimensions are relevant to a cluster, some of the dimensions are irrelevant to a cluster and it can be considered as noise. Moreover, the dimensions that relevant to the clusters are data locality specific. It means that data points in different clusters are correlated with respect to a different and specific set of features. The projected clusters may have different numbers of relevant dimensions and the dimension projection may exist in arbitrarily oriented subspaces of lower dimensionality.

The formal definition of generalized projected cluster was given by Aggarwal in [29]. It was defined as a partition of a set C of data points with a set ξ of vectors such that points in C are closely clustered in subspace ξ and the dimensionality of ξ is much lower than the dimensionality of full feature space. Examples of projected clusters are available in [28, 29].

3.1.2 Projected clustering algorithm

Aggarwal has defined the problem of finding projected clusters as a two-fold problem [5]. The first problem is to locate a set of clusters' center and the second problem is to find the appropriate set of dimensions for each cluster. The problem of finding clusters' center in full dimensional environment has been studied for many years and there are many state-of-the-art approaches available, such as k-means methods. Therefore, the focus of the projected clustering algorithm, such as PROCLUS and ORCLUS, are stressed to the problem of finding set of projected dimensions for each cluster.

Both PROCLUS and ORCLUS are projected clustering algorithms introduced by Aggarwal [5, 29]. PROCLUS is the first algorithm of projected clustering. It is a simplified model that dimension projection can only be made on the axis-parallel manner. However, Aggarwal considered the distribution of data points may not be necessary parallel to the dimension axis. It is because there may be inter-attribute correlations existed in the real data and the projections in arbitrary directions may be more appropriated to capture the skews in data distributions. ORCLUS is a generalized projected clustering algorithm that allows the dimension projection in arbitrary directions, so we chose ORCLUS to illustrate the concept of projected clustering in this thesis.

3.1.2.1 Overview

In the model of ORCLUS, data points may get aligned along arbitrarily skewed and elongated shape in lower dimensional space because of the inter-attribute correlations exists [28, 29]. Each orthogonal set of vectors, called projected dimensions, defined a subspace for a projected cluster is then used to capture the nature of skews and correlations in the attributes. The objective of ORCLUS is to discover tightest projected clusters, in terms of projected energy [5, 28, 29], with unique subspace of dimensions. The subspace of dimensions in original feature space is represented by a set of projected dimensions.

There are four user inputs in ORCLUS. The number of resulting projected clusters k and their cardinality of dimensions l are specified by user. The rate of reduction in number of clusters (α) and dimensions (β) are two important user input parameters to control the quality and the efficiency of ORCLUS algorithm. The output of the ORCLUS are ($k+1$) projected clusters and each of them having a subset ξ of dimensions with cardinality equal to l.

3.1.2.2 ORCLUS: A three phase algorithm

ORCLUS employed the variant of hierarchical merging approach to maintain the feasibility in handling large dataset. Hierarchical approach is prohibitively expensive for large dataset. Therefore, cluster is used as a generic unit during the merging operation of ORCLUS instead of considering a single data instances as the merging unit. During each merging operation, the number of clusters and the dimensionality of clusters are decreased gradually with the rate equal to α and β respectively. It stops when k projected clusters with dimensionality equal to l are obtained.

The ORCLUS algorithm consists of three phases. They are the initialization, iterative and refinement phase. In this section, we will discuss the aims of each phase and how they work briefly.

The aim of the initialization phase is to find out a superset of piercing set of medoids for the iteration phase. In order to obtain a set of reliable clusters, we try to pick at least one seed in each natural cluster in this phases. Due to the computational complexity, it is not feasible to start up with a large number of seeds. So, greedy algorithm is applied to find out a small enough superset of piercing set of mediods that is few times larger than the user specified number of final clusters (*k*) in this initialization phase. The technique of the greedy algorithm is widely used in the partitioning approach. The idea is that every new initial cluster seed for the superset of piercing set of medoids is selected with maximum distances with the set of selected seeds.

The aim of iterative phase is to improve the quality of the set of cluster centers by hill climbing approach. In most of the cases, even domain experts do not know the number of natural clusters in advance. Therefore, it started with a superset of piercing set of cluster centers from previous phase and the projected dimensions for each projected clusters are initialized with full dimensionality. By merging closest clusters and evaluate a new set of projected dimensions iteratively, the best set of cluster centers are found.

There are three operations in this phase and they are the assignment, dimension projection and merging operation. Firstly, data points are assigned to the closest cluster seed among all available clusters and their corresponding set of projected dimensions. Secondly, the new projected dimensions for different clusters are evaluated from their own data points. Finally, the closest clusters are merged to form a new one.

The aim of the refinement phase is simple that targeted to ensure the quality of the clustering result by one pass over data with the best set of cluster seeds found in the iterative phase. In this phase, the data points are assigned to the closest cluster seeds found in iterative phase and then the projected dimensions of each cluster are deduced from this final set of clusters like the iterative phase.

3.1.2.3 Dimension projection process in ORCLUS

The major difference between projected clustering and traditional clustering is that it consists of the dimension projection process that does not exist in traditional clustering approach. In Figure 3.1, dataset with 3 available dimensions are used as an example to illustrate the difference between traditional clustering approach and projected clustering approach. In this example, two clusters (Cluster A & B) are formed under the 3 dimensional spaces by using traditional clustering approach. However, projected clustering approach form Cluster A and Cluster B under different sets of dimensional space by using the dimensional projection process. In most of the cases, the dimensionalities of the projected clusters are much smaller than the full dimensionality of dataset and tighter clusters can be formed.

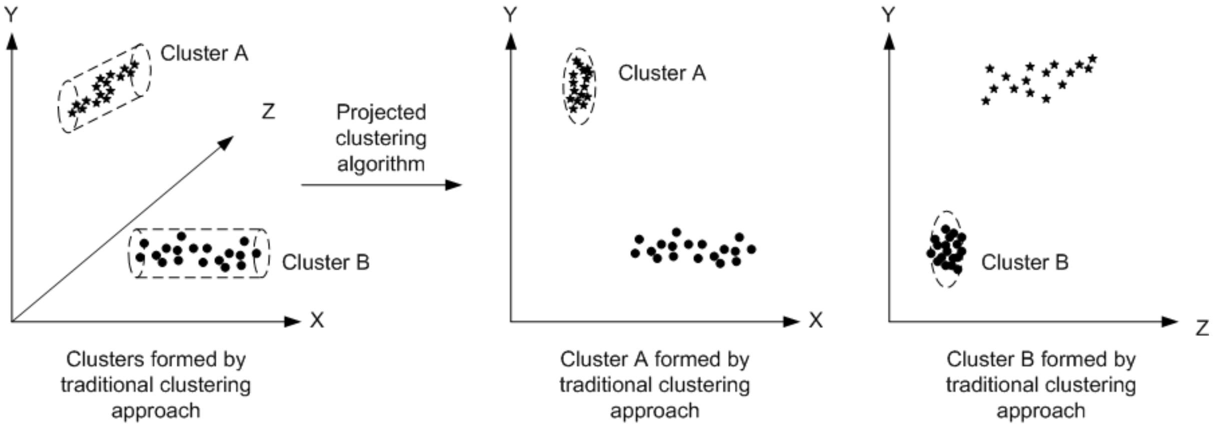
Y
Cluster A
Z
Cluster B
X
Clusters formed by traditional clustering approach
Projected clustering algorithm
Y
Cluster A
X
Cluster A formed by traditional clustering approach
Y
Cluster B
Z
Cluster B formed by traditional clustering approach

Figure 3.1 Example of projected clusters

The dimension projection mechanism of ORCLUS is employed the idea of singular value decomposition (SVD) that used in the feature reduction for years. The original idea of SVD is to transform the original data space into a new coordinate system in which the (second order) correlations in the data are minimized. In the transformed orthonormal system, the dimensions of transformed data space are defined by the eigenvectors and their eigenvalues denote the spread (or variance) of data points along each such newly defined dimensions.

The problem of choosing the projected dimension in ORCLUS is just opposite to the problem of feature reduction. In the feature reduction problem, dimensions preferred are those captured the world with least amount of information lost. So, those dimensions (eigenvectors) with maximum spread (largest in eigenvalue) are used to retain most of the information. However, in the projected clustering problem, we want to find the dimensions that capture the greatest amount of similarity among those points in the same cluster. Therefore, those dimensions (eigenvectors) with least spread (smallest in eigenvalue) that retain the information about the similarity of the points with least spread in each cluster are being used. In general, the dimension projection process of ORCLUS is a variant of the SVD techniques to obtain the projected dimensions that data points within a projected cluster are closest to each other.

3.1.3 Strength and weakness of ORCLUS

The strength of ORCLUS is that it can form reliable clusters for high dimensional data. It is originally designed for the high dimensional data and it provided a new view point in tackle the curse of dimensionality problem raised from the high dimensional data. Because of the sparsity of data points in the high dimensional space, most traditional algorithms trends to fail in providing meaningful clusters. With the dimension projection process in ORCLUS data points are clustered with different transformed feature spaces in lower dimensionality instead of full feature space. The distances between data points in the projected dimension space is more significant to be distinguished, thus data points can be partitioned into meaningful clusters. On the other hand, ORCLUS allows different sets of projected dimensions for different clusters. It overcomes the limitations in the traditional features selection approach that all clusters are partitioned under the same reduced feature space. Since feature selection techniques can not reduce too many features without any information loss [28, 29], ORCLUS always gives clusters with lower dimensionality and more descriptive sets of projected dimensions. In terms of the intra cluster similarity and inter cluster dissimilarity, ORCLUS with lesser information lost always give better results when compared with those traditional clustering approaches in reduced feature space.

There are different sets of projected dimensions are associated with those resulting projected clusters which provide additional information for the domain experts for further knowledge discovery. In those traditional clustering approaches, we only have distance related information about data instances in different clusters. ORCLUS projected clusters provide not only distance information but also the information on different sets of related features. This additional feature related information makes the clustering results more descriptive and it is potentially useful for the domain expert to analysis available data in more detail manner. It is absolutely the advantages of ORCLUS in further knowledge discovery.

ORCLUS is very powerful to provide reliable projected clusters. Its dimension projection ability can form projection clusters in arbitrarily oriented subspace that capture all possible data skews and feature correlation in data sucessfully. However, the resulting clusters are not very easy to understand. The sets of projected dimensions of ORCLUS clusters are eigenvectors that each of them is linear combinations of those original features. It is never easy for users to interpret, especially for biologists. Therefore, the idea of ORCLUS is potentially useful for molecular biologists in further knowledge discovery but make it applicable is still not an easy task.

3.2 Emerging pattern mining

Emerging patterns (EPs) are one of novel patterns introduced by Dong in 1999 [4]. EPs are defined as the patterns whose support values are having great differences between different data partitions. It captures sets of features that are important in differentiating different samples into correct data partitions. EPs have proved to be applicable in the classification problem in the literatures and it is also believed that it can offer domain experts useful and new insight in difference description for knowledge discovery process.

3.2.1 Introduction

In the past, the problem of mining interest patterns from raw data was focused on those patterns having high support values in the dataset which called frequent patterns. There existing quite a lot of state-of-the-art representations and mining algorithms for such frequent patterns. However, patterns that can be used in the different description between datasets are also important in the knowledge discovery process, such as helping us in making decision in classification problems.

On the other hand, not only patterns have high occurrence are important. In many problems, patterns with low to medium support are also important if they show great differences on different data partitions. For example, the patterns exists in tumor tissues are always having lower occurrence when compare with those present in normal tissues, but they are very important in both cancer classification application and further knowledge discovery.

The objective of emerging patterns is to define a new type of pattern that can be used to describe the difference between datasets. It is because only the occurrence of patterns may not be sufficient for the decision making. By introducing the difference in patterns occurrence between datasets, more complicated problems can be solved.

Another objective of the emerging patterns is that its discriminatory power is being employed and targeted for creating the new generation of classifier to solve the classification problem in high dimensional data related applications, such as cancer tissue classification with gene expression data.

In the problem of mining and using EPs, there are two basic assumptions. First, the importance of the support differences between two data partitions is assumed to be higher than the support value alone in a pattern. Therefore, the usefulness of EPs is no longer considered by the support value in the first place. The growth rate of the EPs, defined in [4], is the major consideration of EPs' usefulness. Second, low to medium support EPs are also important. It is because there are many applications in our real world are targeting for the low support patterns. One of the typical examples is the use of tumor patterns in cancer detection and classification problem. In this example, tumor patterns are most likely to have lower support values in the dataset when compare those patterns exists in normal tissues.

The formal definition of the growth rate and emerging patterns (EPs) can be found in [4]. In general, the growth rate is defined as the ratio of the support value between two datasets or data partitions. Any patterns with growth rates that are larger than the user specified threshold, are EPs. Examples of emerging pattern [35], {gene(K03001) ≥ 89.2} and {gene(R76254) ≥ 127.16} and {gene(D31767) ≥ 63.03}, discovered from colon tumor dataset [36]later, changes it occurrence of 0% in normal tissue to 75% in cancerous tissue.

3.2.2 Border based mining algorithms

The problem of mining EPs with growth rate larger than the user specified threshold is simply divided into the two problems. The first one is to find the set of patterns with high support values in dataset D_1, and the second one is to find the set of patterns with low support values in dataset D_2. The first problem is being one of the famous studies among different data mining parties and there are many innovative algorithms to solve it efficiently. However, mining patterns with low support value are still a very challenging task nowadays and most of the case it will become computational infeasible if the dataset is large in size.

Therefore, Dong [4] defines the problem of mining EPs in another way. The first problem is to find the patterns with high support values in both dataset. The second problem is to subtract the results obtained from the first problem in order to obtain the set of EPs that having high support value in dataset D_1 but not in D_2. The focus in the EPs mining is the subtraction process between two large itemsets.

3.2.2.1 Overview

The basic assumption in mining EPs is that the importance of the difference in occurrence between two datasets is higher than the occurrence alone. Therefore, the nice property used by those apriori-based algorithm, called subset-closedness [4], does not hold for EPs. In order to extract EPs from the data, another nice properties [4] called interval-closedness are introduced. The idea of border representation is used with respects to the interval-closedness property. Patterns are then represented by a border with the most general set of patterns as its left boundary and most specific set of patterns as its right boundary. For any patterns that is the superset of a pattern in the left boundary and is the subset of a pattern in the right boundary, it is said to be contained in the border. For example [4], there are total 12 itemsets ({1,2}, {1,2,3}, {1,2,4}, {1,2,5}, {1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,2,3,4,5}, {1,2,6}, {1,2,4,6}, {1,2,5,6}, {1,2,4,5,6}) contained in {{1,2}, {{1,2,3,4,5}, {1,2,4,5,6}}. By using the interval-closedness properties together with the border based mining algorithm, enumeration of patterns during the EPs mining process which is very time consuming is minimized.

The objective of those border based algorithms is to discover all EPs by using border representations. As mentioned in previous section, the mining for EPs is transformed to a subtraction process between two large itemsets in different datasets. Therefore, there are two objectives of those border based mining algorithms, they are generating the border representing the frequent itemsets of each dataset and performing subtraction between borders.

In order to extract EPs in border representation from the dataset, there is an important input from user is needed. It is the growth rate of EPs that is defined by the ratio between the background dataset and target dataset support values [4].

3.2.2.2 Border-Diff and MBD-LLBorder

The border based algorithm will take borders that representing collections of large itemsets as inputs. By using the set operation, difference between borders can be obtained and the output border is representing the EPs between two dataset with the growth rate larger than the user requirement. In [4], the border based algorithms called Border-Diff and MBD-LLBorder are described in detail and we will introduce these two algorithms in brief here.

The purpose of Border-Diff is to derive the differential between a pair of borders with special form: [L_2, {U}] = [{∅},{U}] – [{∅}, R_1]. [L_2, {U}] is the border that contains the small itemset of a dataset with left boundary equal to L_2 and right boundary that is the universal set {U}. [{∅},{U}] is the universe of the space and [{∅}, R_1] is the collection of large itemsets in dataset. By using the Border-diff algorithm, we can found out the collection of small itemsets in border representation by limited the calculation on the itemsets appeared in the borders and it is very important that for large dataset or high dimensional data. It is because the candidates of small itemsets are large in volume and it is not feasible to extract them easily.

MBD-LLBorder is the main algorithm that aims to discover EPs by manipulating only two input borders. The two input borders are the large borders that representing those large itemsets in two data datasets respectively. In each iteration of the algorithm, it uses one itemset in the right boundary of large border of dataset D_2 and the whole right boundary of large border of dataset D_1 to find out the itemsets that should exist in the left boundary of resulting border.

3.2.2.3 The EPs selection problems

The use of mined EPs always depends on the problem we want to solve. Although we can mine all the EPs by using the border based algorithms theoretically, it is not common to use all of them in solving problems. It is because there always exists in a large number of EPs, selection of EPs become one of the important issues in real applications. There is a specific types of EPs, called Jumping EPs [37], is use extensively in different applications. It is because JEPs has the highest discriminatory power with infinity growth rate. The infinity growth rate means that these EPs only found in one dataset but never exists in the others and their high discriminatory power is most important in classification related applications.

On the other hand, the occurrence of EPs is another consideration in selection of EPs. If EPs are having the same growth rate, the one with higher the occurrence in the dataset always interpreted as more important. The basic assumption of this interpretation is just what frequent pattern used historically. In general, itemsets in the left boundary of EPs border are more general one when compared with those existed in right boundary and they always have higher occurrence. Therefore, they are always employed in applications.

3.2.3 Strengths and weaknesses

The first strength of EPs is that it makes a great change in the view points on importance of patterns. It is because we only consider the occurrence of patterns is not sufficient in many applications. For example, patterns exist in tumor tissue are very important in cancer detection problem, but it always have low occurrence when compared with those patterns come from normal tissue. In many cases, the difference in occurrence between dataset is more important than occurrence alone and EPs are targeted to capture such difference.

The second strength is high discriminatory power of EPs and it can be specified by the user defined growth rate. By definition, EPs with larger growth rate have higher discriminatory power and those EPs with large growth rate has been proven to be useful in the classification problem in the literatures [38-41]. For example, JEP-Classifier and DeEPs which adopts the eager learning and lazy learning approaches are EPs based classifiers that give high performance in the classification problem.

The third advantage is that EPs are also easy to understand because they are just the collection of attributes with specified range of value in raw dataset. Readability is one of the important requirements in the knowledge discovery of biological domain because any hypothesis we mined should undergo the validation by domain experts in laboratories. If the hypothesis obtained is not easy to understand, it may not be helpful in the further knowledge discovery.

Finally, EPs with low support values can be obtained efficiently by using the border based mining algorithms. Like the example mention previously, EPs with low support value are potentially useful for cancer detection problem. Patterns with low support are difficult to find by most of the well known pattern mining algorithms and border based algorithms make EPs with low to medium support to be available in many real situations.

However, there are also limitations in using EPs. The number of EPs that obtained is large in volume especially for high dimension data. The large volume of patterns found may not be easy for us to apply in the application and it also makes difficulties for the domain experts in the further knowledge discovery process. Top EPs that have highest value in occurrence are always used as solutions instead of complete set of EPs are used in order to maintain the efficiency for the system. However, too less EPs are used will actually lower the accuracy and the robustness for unseen data but too many EPs will lower the system efficiency. It is very difficult to get such a magic number of EPs to be used in every application.

3.3 Summary

The major problem for pattern mining approach, such as using emerging patterns, is that there are too many patterns extracted from the data. The large amount of mined patterns are difficult for the domain experts to discovery any further knowledge and they are not easy to be applicable. Clustering are more efficient approaches when compared with pattern mining approach because of relative limited numbers of result clusters are obtained. However, most of the clustering approaches group data instances by geometric distances. The reasons behind the resulting clusters may not be biologically meaningful and easy to understand for biologists. We found that these two approaches of data mining techniques are rarely work together, but they are strong in different areas and there is an opportunities to integrate them to get both their strengths and compromise their weakness. This leading to our approach introduced in the rest of the thesis, the emerging pattern based projected clustering approach.

4 Emerging Pattern Based Projected Clusters

A new kind of knowledge representation, called emerging pattern based projected clusters (EPPCs), are introduced in this chapter. EPPCs are defined as those projected clusters whose projected dimensions are the collection of attributes in the emerging patterns (EPs) mined from the dataset using data mining techniques. It is believed that by using the EPPCs to group the data instances into different clusters, the correlation between those instances are much easier for the domain expert to analyze. In this chapter, the definition and problem of forming EPPCs is presented together with the EPPC framework.

4.1 Introduction

In previous chapter, we have introduced two powerful data mining techniques that targeted for high dimensional data. Emerging patterns are one of the outstanding pattern mining approaches and it has proved to be applicable in bioinformatics problems [35]. Since the number of available data record is still limited when compared with the number of available dimensions, the emerging patterns extracted may not be sufficient to approximate the real world. The large number mined emerging patterns are also not easy for the domain experts to use and conduct biological experiments for further knowledge discovery efficiently. Projected clustering is one of the state-of-the-art clustering approaches that it tackled the curse of dimensionality problem and let clustering in high dimensional data being possible. However, the resulting projected clusters with complicated projected dimensions are not easy to understand, especially by biologists. These two effective data mining approaches have their own limitations and they are still insufficient for the bioinformatics domain that results are required to be easy for understanding and efficient in supporting further knowledge discovery in laboratories.

In this chapter, a new knowledge representation, called emerging pattern based projected clusters (EPPCs), are introduced. The rationale of EPPCs is intended to integrate the emerging patterns and projected clusters to get both their strengths. The resulting EPPCs are potentially useful, efficient and understandable for the biologists. In the following section, we begin by an example of EPPCs and followed by the definition of EPPCs given in Section 4.3. Then the problem in finding EPPCs is discussed and framework to mine EPPCs is also introduced.

4.2 Example of emerging pattern based projected clusters

Example 4.2.1 There are some EPPCs generated from the colon dataset [36]. The following are three typical EPPCs consisting of different sets of cluster dimensions and data instances:

Table 4.1 Examples of EPPCs from colon dataset

	No. of sample	**Tissue Type**	**Cluster Dimension**				
			Gene 1	**Gene 2**	**Gene 3**	**Gene 4**	**Gene 5**
Cluster 1	7	Normal	H51015	R10066	U32519	T47377	Z50753
Cluster 2	6	Cancer	M76378	T47377			
Cluster 3	10	Cancer	H08393	M76378			

In above table, three EPPCs are selected to demonstrate the two important characteristics of resulting EPPCs obtained from the colon dataset [36]. The first feature is that EPPCs are projected clusters with subset of attributes as its projected dimensions. Unlike the k-mean clustering approach, it uses corresponding subset of attributes in evaluating the distances between each instance and cluster representatives instead of using the full dimensionality. For example, Cluster 3 using 2 genes, H08393 and M76378, out of 2000 in raw data as its projected dimensions to group 10 instances. Another feature of EPPCs is that each of them may have their own sets of projected dimensions in different size.

Unlike those ORCLUS projected clusters which consist of same number of projected dimensions (principal components) among all projected clusters. In our example, Cluster 1 has 5 projected dimensions; Cluster 2 and 3 have 2 projected dimensions.

4.3 Definition of emerging pattern based projected clusters

Before proceeding to describe our EP-based projected clustering (EPPC) algorithm, we introduce some notations and definitions. Let N be the total number of data points and n_i be the number of data points in cluster C_i. Assume that the dimensionality of full data space D is equal to d and the dimensionality of projected space D_i of cluster C_i is equal to d_i, where $d_i \leq d$. Let $X_i = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ be the set of data points in cluster C_i, $\vec{p}_j = \{x_{j1}, x_{j2}, \ldots, x_{jd_i}\}$ be the projected point and $\vec{s}_i$ be the centroid of cluster C_i, i.e. $\vec{s}_i = \sum_{j=1}^{n_i} \vec{p}_j / n_i$. Finally, the (projected) distance of projected point $\vec{p}_j$ to the center of cluster C_i is written as $Pdist(\vec{p}_j, \vec{s}_i, D_i)$.

By means of the projected distance, the emerging pattern based projected clusters are defined as below:

Definition 4.3.1 Given a set of data points N, an emerging pattern based projected cluster (EPPC) C_i consists of set of data points $X_i = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ that their projected distance $Pdist(\vec{p}_j, \vec{s}_i, D_i)$ between its centroid $\vec{s}_i$ are minimal among all available EPPC (C_1 ,…, C_n) in the data space. The projected dimensions D_i are the collection of attributes of emerging patterns mined from the data.

Note that using emerging patterns to form projected clustering provides certain flexibility to users. Users are not limited to use only one EP in dimension projection process to form

1-EPPCs. More than one EP are also applicable to generate *m*-EPPCs and the resulting clusters always give better results in the classification problem when proper number of EPs are used. Detail experimental results can be found in Chapter 5. Therefore, *m*-EPPC in this thesis is referred as an EPPC with its projected dimensions D_i equivalent to the union of all attributes among *m* EPs.

4.4 The problem of finding emerging pattern based projected clusters

Finding EPPCs is a two-fold problem that is similar to ORCLUS projected clustering algorithm in [5, 28]. First, we have to locate the clusters' center. Second, we needed to find out the projected dimensions for each clusters.

The first problem is well defined in those partitioning methods of clustering approach. In most of the cases, a set of seeds are carefully selected using different heuristics at the beginning and improve the quality of those seeds iteratively. For example, the new seeds are selected by maximizing the distance between the set of previously selected seeds in order to obtain a piercing set of seeds at the beginning. And the quality of those initial cluster centers is improved iteratively by some well known approaches, such as k-means and k-medoids algorithms. Therefore, our EPPC algorithm is focused on the second problem in finding projected dimensions for the projected clusters.

The second problem was stated as the redefining clustering for high-dimensional data in the literature [28]. Under traditional clustering problem definition, this problem does not exist. Full dimensional space of the data instances are used in grouping objects into clusters. However, the available dimensions of objects can be considered as infinity in reality and those available dimensions that we have considered in the traditional clustering approach can be interpreted as a simplified or reduced feature space obtained by our limited data acquisition abilities. And we can interpret the traditional clustering problem as a model of simplified clustering problem with the dimension projection problem excluded. But actually, the feature selection process can be interpreted as dimension projection with same set of projected dimensions for all clusters and it is performed by the domain experts using their domain knowledge in the data acquisition and preprocessing process. Since our knowledge in life science is still very limited, the complexity of the data is very difficult to reduce during the data preparation or collection stage. Data in the domain of bioinformatics, such as gene expression data, are high dimensional in nature. Dimension projection process becomes core problem in handling high dimensional data and it is the major focus in our project.

4.5 Emerging Pattern Based Projected Clustering Framework

We have discussed the problem of finding emerging pattern based projected clusters (EPPCs) in previous section. However, the successfulness of finding EPPCs also depends on the availability of the emerging patterns (EPs) from the dataset. In this chapter, we introduce the framework in finding and using EPPCs which consists of three phases. The phase one of the framework is discovering the emerging patterns from the dataset. The EPs mined are used as the input in the phase two to form the EPPCs. Finally, the EPPCs obtained are using in the classification problem in phase three.

4.5.1 Overview of the EPPC framework

In the past, mining EPs and projected clustering techniques were used independently in solving different types of problem since they are strong in different areas. In this section, they are integrated to form a framework in solving the bioinformatics problems, such as classification of the tumor and normal tissues with gene expression data.

As shown in the Figure 4.1, the whole framework consists of three phases. The aim of the phase one is to extract the EPs from the raw data. The mined EPs are large in volume and representatives are selected as the input of the phase two. The purpose of phase two is to form the EPPCs by using the selected EPs and our EPPC algorithm. The resulting EPPCs are used in the phase three that in classifying the new testing data samples.

In the following subsections, we discuss each phase in details with the problem of classifying cancer tissues and normal tissues in colon dataset [36] as an example for illustration.

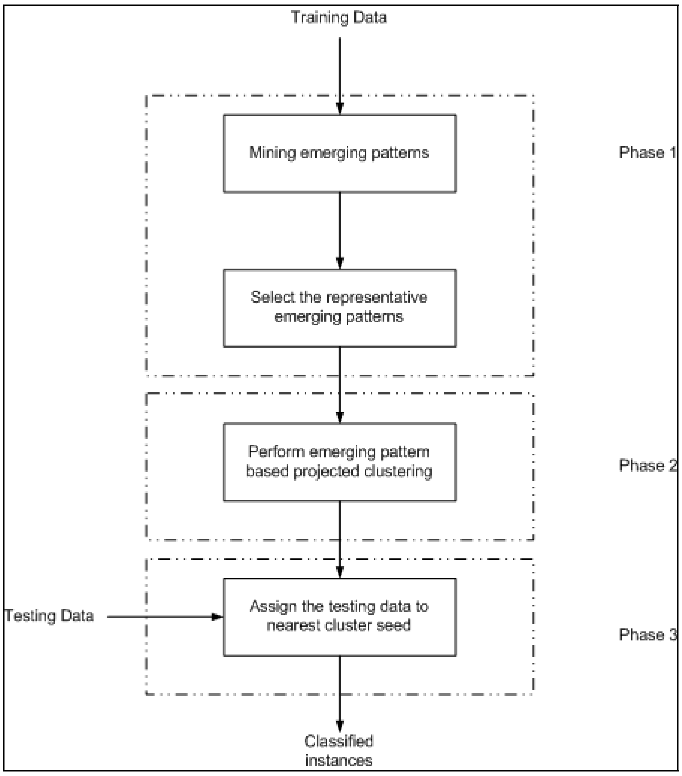
Training Data
Mining emerging patterns
Phase 1
Select the representative emerging patterns
Perform emerging pattern based projected clustering
Phase 2
Testing Data
Assign the testing data to nearest cluster seed
Phase 3
Classified instances

Figure 4.1 Flowchart of the EPPC framework

4.5.2 Phase 1: Mining emerging patterns

In this phase, our objective is to extract the domain knowledge in terms of EPs and the resulting EPs are used to generate the EPPCs. By using the border based mining algorithms introduced by Dong and Li [4, 37], those EPs are extracted for the next phase. The complete flow of mining EPs is shown in the Figure 4.2 below.

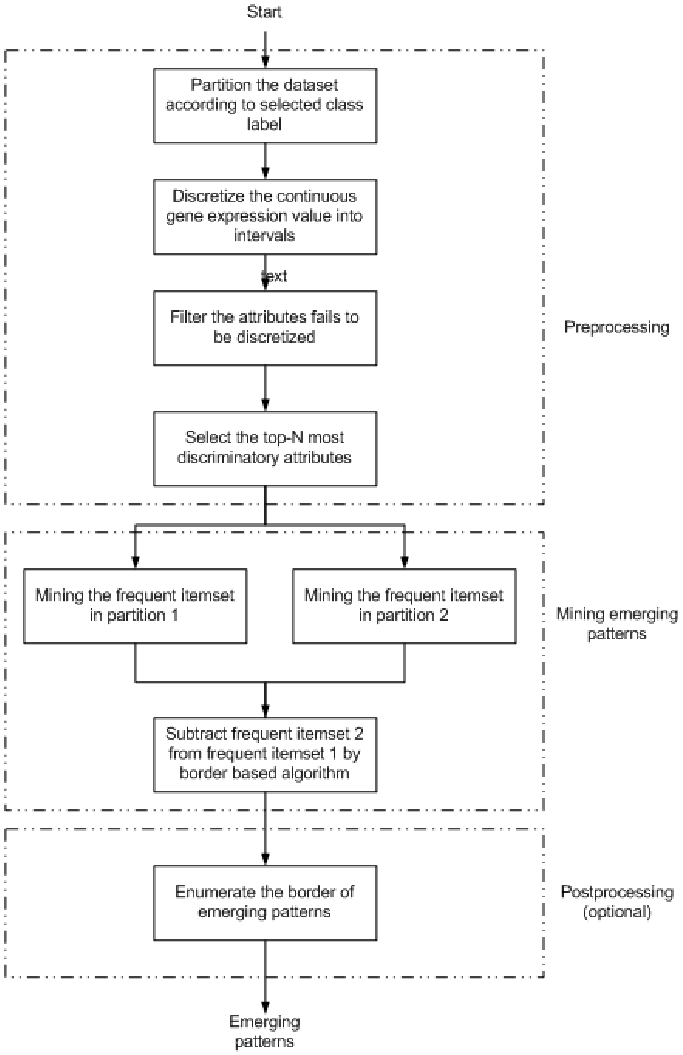
Start
Partition the dataset according to selected class label
Discretize the continuous gene expression value into intervals
text
Filter the attributes fails to be discretized
Select the top-N most discriminatory attributes
Preprocessing
Mining the frequent itemset in partition 1
Mining the frequent itemset in partition 2
Mining emerging patterns
Subtract frequent itemset 2 from frequent itemset 1 by border based algorithm
Enumerate the border of emerging patterns
Postprocessing (optional)
Emerging patterns

Figure 4.2 Flowchart of mining of EPs

4.5.2.1 Preprocessing

By definition [4], the EPs are patterns which having great differences in support values between different datasets. In mining EPs from the data for the classification problem of cancer and normal tissues, the first step is to divide the dataset into partitions by using available class labels, such as “cancerous” or “normal”. The next step is to discretize each attributes, gene expression values, that are continuous in nature into intervals with entropy discretization before proceed. Each interval of a gene expression values will be interpreted as an item in the data space in our pattern mining process later. After the discretization process, we found that not all the genes that their values can be discretized into intervals with respect to those available class labels. Those genes failed to be discretized are filtered out because they are considered as not discriminatory in distinguishing the cancerous and normal tissues. In most of the case, the raw data consists of large number of attributes and there may be still quite a large number of attributes left after discretization process. For example, there are 135 out of 2000 genes that can be discretized into 2 intervals and totally 270 items can be encoded from the colon dataset [36]. However, 270 items are not a small number in pattern mining problem. It always takes us a long time in the mining patterns and generates tons of resulting patterns. Too many patterns found may not be good in most of the cases that it leads difficulties in any further analysis and applications. Therefore, optional selection process is advised before the mining process actually started. The entropy value obtained during the discretization process is then used to select the top-n most discriminatory attributes from the data. In the colon dataset [36], 35 genes with smallest entropy value are selected because of they are the most discriminatory among 135 discretized genes and 35 attributes would be more than enough for our problem.

4.5.2.2 Mining emerging patterns by border based algorithms

As mentioned in Section 3.2, mining emerging patterns are not feasible by using apriori based algorithms. In this step, the border based algorithms introduced by Dong & Li [4, 37] are used. In general, BORDER-DIFF [4] and MDB-LLBORDER [4] is most suitable partners that they can find out all the EPs with growth rate large than the user supplied threshold. However, in terms of the clustering aspects, JEPs is always preferred since one of the fundamental objectives of forming clusters is trying to maximize the dissimilarity between clusters and JEPs has growth rate equal to infinity which is highest in discriminatory power among all EPs. In our framework, JEPs are used to form EPPCs because of its high discriminatory power and also obtaining JEPs by using border based algorithm HORIZON-MINER [37] and MDB-LLBORDER is more efficient. As shown in the Figure 4.2, HORIZON-MINER is used to mine the frequent itemsets from every partition of dataset. After that, MDB-LLBORDER use the output borders from HORIZON-MINER to perform the subtraction operation to obtain a JEP border. Since the mining process in extracting JEPs by using border representation do not needed to perform enumeration and support counting for pattern candidates. It is very efficient for the gene expression data that is high dimensional in nature.

4.5.2.3 Postprocessing (optional)

By using the border based algorithms, mining EPs or JEPs become feasible. A border that contains all the EPs that their growth rate larger than the threshold are obtained. There may be some cases to form emerging pattern based projected clustering under some constraints. In order to select the desirable set of EPs for the next phase, the border of EPs may be first enumerated. However, the number of EPs that we needed in next step is always small and we can use those EPs at the left boundary of border that has the highest occurrence and without undergo the enumeration process. Therefore, this step is optional and it exists for the completeness of the framework only.

4.5.3 Phase 2: Forming EP-based projected clusters by EPPC algorithm

In this phase, our goal is to form the EPPCs from the training data. By using EP-based projected clustering (EPPC) algorithm, EPPCs are obtained. Our proposed EPPC algorithm includes three phases that similar to [28], namely, initialization, iteration and refinement phase.

In the initialization phase, its goal is to pick up the initial cluster seeds for the iteration phase. In the iteration phase, data points are assigned to different clusters and projected dimension of those newly formed clusters are being evaluated. Its goal is to improve the quality the set of clusters continuously until the user specified numbers of clusters are obtained. Once a set of best cluster seeds is obtained after iterations, the refinement phase will start and all the data points will be reassigned to those cluster seeds obtained previously to form a set of final clusters. The goal of refinement phase is to ensure the quality of clusters with best clusters' center found in iterative phase. The overviews of the EPPC algorithms are shown in Figure 4.3 and Figure 4.4. The details of each phase are provided in the following sections.

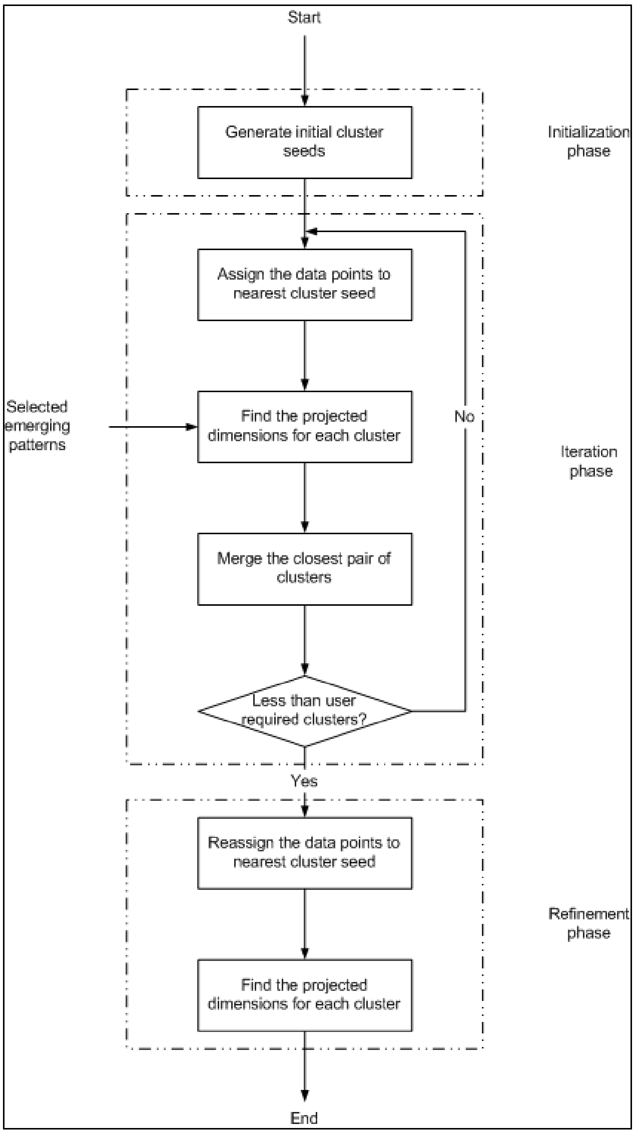
Start
Generate initial cluster seeds
Initialization phase
Assign the data points to nearest cluster seed
Selected emerging patterns
Find the projected dimensions for each cluster
No
Iteration phase
Merge the closest pair of clusters
Less than user required clusters?
Yes
Reassign the data points to nearest cluster seed
Refinement phase
Find the projected dimensions for each cluster
End

Figure 4.3 Flowchart of the EPPC algorithm

```
Algorithm EPPC ($k_0$, $k$, $E$){
  Initialization phase
  Pick $k_0 > k$ initial cluster seeds randomly from the
  dataset;
  Set no. of current cluster to no. of initial cluster;
  for each cluster {
    Set the cluster dimension to full dimensionality
  }
  Iterative phase
  While no. of current cluster > user requirement {
    Assign the data points to the nearest cluster seeds;
    Determine the cluster dimensions associated to each
    cluster;
    Merge the closest clusters and obtain the new seed for
    the newly merged cluster;
    Update the no. of current cluster;
  }
  Refinement phase
  Reassign the data points to the set of good seeds obtained
  from iteration phase;
  Determine the cluster dimensions associated to each
  cluster;
  Return the projected clusters with cluster seeds,
  corresponding dimensions and data points;
}
```

Figure 4.4 The EPPC algorithm

4.5.3.1 Initialization phase

In this phase, the number of final clusters is defined by the users. We randomly pick k_0 initial cluster seeds from the dataset, where k_0 should be several times larger than k, and the projected dimensions of all initial seeds are initialized to the full dimensions of the dataset initially.

4.5.3.2 Iterative phase

The goal of the iteration phase is to improve the quality of the cluster seeds iteratively in order to find the best clusters. There are three operations in this phase, namely, assignment, dimension projection and merging.

For the assignment operation, there should be k_c cluster seeds in the current iteration. In this operation, the data points in the dataset are assigned to their closest seed. We use the distance metric, such as City segmental distance or Euclidean distance, to measure the distances between the data points and cluster seeds under those projected dimensions, i.e. the projected distance, $Pdist(\vec{p}_j, \vec{s}_i, D_i)$. After the partitions are formed, the centroids of each partition are evaluated and they are used as the new seeds in the next iteration. This procedure is illustrated in Figure 4.5 below.

```
Algorithm Data_Point_Assignment {
  for each data point {
    for each cluster {
      Determine the projected distance between the data points and current seeds;
    }
    Add the data points to their nearest cluster;
  }
  Remove cluster from the set if it is empty;
  Set the centroids of those projected clusters as the new cluster seed;
  Return the cluster seed and data set in projected clusters;
}
```

Figure 4.5 Data point assignment algorithm

For the dimension projection operation, those partitions formed by the assignment operation consist of a set of data points. In this operation, the projected dimensions of each projected cluster are evaluated by their own data points. For each partition, we examine its data points and find those EPs embedded. The user specified numbers of EPs that are most frequently occurred are chosen and the union dimension comprised in this set of EPs act as the set of projected dimensions for that particular partition. This procedure is described in Figure 4.6.

```
Algorithm Dimension_Projection {
  for each cluster {
    Find the user specified number of EPs that having most frequent occurrence among the data points in the cluster;
    Find the corresponding attributes that make up the corresponding set of EPs;
    Set the projected dimensions to that collections of attributes;
  }
  Return the dimensions for the projected clusters;
}
```

Figure 4.6 Dimension projection algorithm

In the last operation, i.e. the merging operation, the closest pair of clusters is merged together to form a new cluster. The clusters undergo this operation is obtained by evaluating the average distance between the union of data points and new cluster seed of merged clusters. In short, the smaller the average distance, the closer the pair of clusters. The details of this operation can be found in Figure 4.7.

```
Algorithm Cluster_Merging {
  for each pair of cluster {
    Find the closest pair of clusters from the set of existing clusters;
    Merge the data points of the two clusters;
    Find the projected dimensions of the unified data points;
    Find the new seed of the unified data points;
    Evaluate the radius of the merged clusters;
  }
  Merge the closest pair of clusters such that the radius of the merged clusters is minimal;
  Return the set of new cluster seeds;
}
```

Figure 4.7 Cluster merging algorithm

4.5.3.3 Refinement phase

Finally in the refinement phase, the resulting cluster seeds obtained from the iteration phase are then used to form the final clusters by assigning all the data points to them once again. The goal of this phase is to ensure that all data points are assigned to the closest cluster seeds after the final cluster seeds are found. The assignment operation and dimension projection operation shown in Figure 4.5 and Figure 4.6 are process once more time in this refinement phase.

4.5.4 Phase 3: Using EP-based projected clusters

In this phase, our first objective is to assign the class labels to those EPPCs formed in previous phase and our second objective is using the resulting EPPCs to solve the classification problem for the new data.

In section 4.5.3.3, data instances are reassigned to the closest cluster seed to form EPPCs with best quality. However, EPPC may consists of data points with different class labels and we have to deduce its class label before use. The simplest way to assign the label to those resulting clusters is using class label of the majority of the data instances. Although it is rare, there may be still some cases that the number of data instances with different class labels are the same. In this case, we will compare the percentages of data instances for each class in the training dataset to make the final decision. If it is still a problem in giving the label to a cluster, it will be labeled as unclassified.

In the classification problem of tumor and normal tissues, we try to measure the projected distance between the new data instances and the centroids of our EPPCs. The new data instance is assigned to the cluster with minimal projected distances and also the corresponding class label is assigned to it.

5 Simulation Results

In this chapter, the performances of using EP-based projected clusters (EPPCs) are evaluated. First of all, the potential of using EPPCs on the gene expression data is illustrated by comparing K-means clusters and EPPCs for colon cancer data. Second, we compared our EPPCs with the projected clusters formed by state-of-the-art algorithm, ORCLUS, in classifying colon cancer data. In these experiments, the powers of EPPCs, high classification accuracy and readability, on the bioinformatics problems are demonstrated. Finally, we studied the performance of EPPCs with different cancer datasets and proved it can be applicable in cancer classification problem.

5.1 Forming generic EP-based projected clusters

In this experiment, we formed the generic EPPCs by using 1 EP in the dimension projection process. The clustering error of 1-EPPCs are compared with k-mean clusters and we proved that projected clusters are more applicable for gene expression data that it is high dimensional in nature.

5.1.1 Dataset and experimental settings

The colon tumor dataset was collected by Alon et al [36] at Princeton University. The dataset consists of 2000 gene expression values of 40 tumor and 22 normal colon tissues samples and it is publicly available at http://microarray.princeton.edu/oncology/affydata/index.html. Since the original dataset consists of 2000 gene expression values as attributes for total 62 samples and not all of those attributes are useful in separating samples into different classes, the dataset is first reduced its size by using the entropy discretization method provided by MLC++ [42]. The entropy method finds a total of 135 significant genes that are relevant to classify samples into different classes and we pick the 35 top-ranked genes as mentioned in [43] to generate a reduced dataset. In order to facilitate the study of the relationship between different gene expression values and the types of tissue, each gene expression value of the reduced dataset is normalized before they undergo the clustering process. The mean of the normalized gene expression values is equal to zero while the standard deviation is equal to one.

In this experiment, we report the clustering performance of our proposed EPPC algorithm. Different number of initial and final clusters combination are used and three most popular distance metrics, namely, Euclidean, City block and City segmental distance, are employed. All samples in the dataset are used in the experiments and the error of the resulting cluster is calculated as:

$$\text{Clustering Error} = \frac{\text{Number of samples in wrong clusters}}{\text{Total number of samples in dataset}}$$

In order to minimize the effect from different initializations, each combination of experiment settings was simulated 50 times and the average error rate is listed in Table 5.1-Table 5.3 in following section.

5.1.2 Clustering performance

5.1.2.1 Clustering error of EPPCs

According to our experimental result shown in Table 5.1 - Table 5.3, the City segmental distance gives the smallest cluster errors in general. It was observed that in every final cluster, the number of projected dimensions varies due to different EPs being used in projections. Thus, the smaller the clustering error means the better clustering subspaces were identified. Moreover, there is a most significant decrement in clustering error as a result of increasing the number of final clusters from 4 to 8. It is because there may have quite a number of sources causing the cancer to develop. In our clustering context, there may be more than two natural clusters exist in those gene expression data and using 8 final clusters gives a much better result than using 4. It suggested that the number of natural clusters may between 4 and 8. When the number of the final clusters is larger than the number of natural clusters available in the data, any further increase in the number of final clusters may not be so important and thus the improvement in the clustering error becomes limited. Although we only have two class labels obtained from the microarray experiments, we can guess the number of natural clusters from our clustering results and such kind of information is potentially useful in providing some directions for further biological studies.

Table 5.1 Clustering error based on Euclidean distance

	Final Cluster Num.								
Initial Cluster Num	.	4	8	12	16	20	24	28	32
	8	0.221	0.135						
	16	0.249	0.167	0.152	0.110				
	24	0.243	0.165	0.151	0.135	0.140	0.095		
	32	0.272	0.177	0.136	0.135	0.113	0.122	0.123	0.084
	40	0.257	0.174	0.137	0.116	0.103	0.092	0.102	0.096

Table 5.2 Clustering error based on City Block distance

	Final Cluster Num.								
Initial Cluster Num	.	4	8	12	16	20	24	28	32
	8	0.288	0.129						
	16	0.304	0.230	0.222	0.107				
	24	0.311	0.233	0.213	0.206	0.198	0.092		
	32	0.284	0.240	0.197	0.182	0.177	0.166	0.162	0.080
	40	0.318	0.237	0.199	0.170	0.154	0.145	0.138	0.114

Table 5.3 Clustering error based on City Segmental distance

	Final Cluster Num.								
Initial Cluster Num	.	4	8	12	16	20	24	28	32
	8	0.195	0.129						
	16	0.188	0.157	0.141	0.107				
	24	0.193	0.152	0.136	0.120	0.117	0.093		
	32	0.209	0.167	0.143	0.123	0.114	0.100	0.089	0.076
	40	0.197	0.169	0.147	0.119	0.096	0.085	0.076	0.074

5.1.2.2 Clustering Error of k-mean clusters

In order to demonstrate the effectiveness of the EPPC algorithm, the K-means algorithm implemented by a public domain package called NetLab3.2 was used to cluster the same set of reduced colon data that is used in previous experiment. Again, the simulation was repeated for 50 times for each individual setting. As shown in Table 5.4, the performance of the K-means algorithm is not as accurate as ours.

Table 5.4 Clustering error using k-means algorithm

Avg Error	**Cluster Num.**			
	8	**16**	**24**	**32**
	0.321	0.270	0.233	0.190

5.2 Classification using m-EPPCs

In this set of experiments, we divided the dataset into two partitions for training and testing. Projected clusters are formed with different combination of projected dimension numbers, initial and final cluster numbers. The classification accuracies of ORCLUS projected clusters and EPPCs on unseen testing data are examined and their performances are detailed in the following sections.

5.2.1 Dataset and experimental setting

We also employed the colon tumor dataset collected by Alon et al [36] at Princeton University, the details of the dataset and its preprocess steps can be found in Section 5.1.1. In the following experiments, we report and compare the classification performances of those projected clusters generated by our EPPC algorithm and ORCLUS [28]. Initially, a specified portion of tumor and normal samples are randomly selected and used as the training dataset to generate projected clusters. Then, the rest of the records act as testing samples and they are assigned to those resulting projected clusters accordingly. Measurement of the classification accuracy is defined below.

$$\text{Classification accuracy} = \frac{\text{No. of testing samples in correct clusters}}{\text{Total number of testing samples}}$$

In order to minimize the effect of initial points and ordering problems in clustering, we repeated every experiment for 50 times for different number of initial clusters and final clusters, with training and testing samples that selected and ordered randomly.

5.2.2 Classification performance

5.2.2.1 ORCLUS projected clusters

In this experiment, we studied the classification performance of ORCLUS projected clusters. The focus of this experiment is on the effects of ORCLUS projected clusters with different number of projected dimensions and their performance in classification for the colon dataset [36]. We implemented ORCLUS and used Euclidean distance as a distance metrics by following the methodology in [28]. Each projected dimension is equivalent to one principal component of the covariance matrix of the datasets. The details of the dimension projection mechanism are available in [28]. In this test, we use 70% of samples (43 samples with 28 tumor tissues and 15 normal tissues) from the dataset as training data and choose the number of principal components (*l*) that we are interested in to 1, 5, 10, 15, 16, 17, 18, 19, 20 and 35. The averaged classification accuracy of those 50 repetitions respect to different number of initial clusters and final clusters are studied.

From the experiment results, we found that the relationship between the classification accuracy and number of projected dimensions are similar among different of number of initial clusters. For example, Figure 5.1 and Figure 5.2 are the summaries of the experimental results in different number of initial clusters and they give the similar tendency in classification performances.

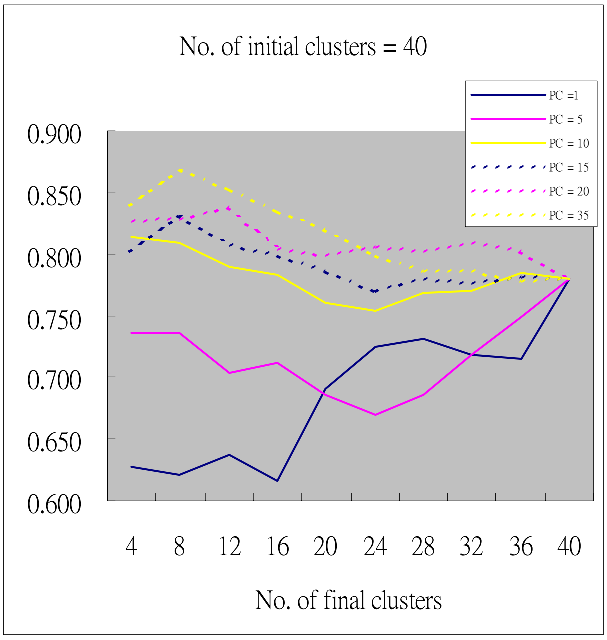

Figure 5.1 Classification accuracy of ORCLUS projected clusters (Initial cluster = 40)

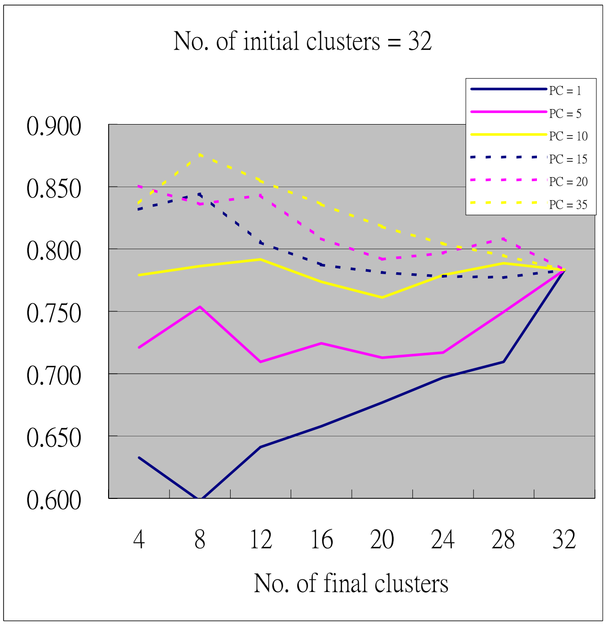

Figure 5.2 Classification accuracy of ORCLUS projected clusters (Initial cluster = 32)

Figure 5.1 summarized the experiment results with the number of initial clusters equal to 40 is employed to demonstrate the variation in classification performances of projected clusters with different number of projected dimensions in the followings. We found that if we used less than 15 principal components as projected dimensions to form ORCLUS projected clusters. The resulting clusters give relatively low classification rate for those testing samples in cancer detection. By increasing the number of the projected dimensions from 1 to 15, the classification accuracy of those projected clusters are increased significantly as shown in the Figure 5.1. In the case of more than 15 principal components are used, the classification rate is much higher in general but the improvements in terms of the classification accuracy by further increasing the number of projected dimensions are flattened. The reason is that the increase the number of projected dimensions will increase the volume of information extracted form the training data and form more reliable clusters, thus the higher classification rates are obtained in general when the numbers of projected dimensions are increased. But if the number of projected dimensions reaches the optimal, any further increment of projected dimensions will not introduce relevant information. Besides, it will introduce additional distance between the testing samples and the projected clusters' centers and it may cause the distance between the testing samples and those cluster centers become too close to distinguish from the most desirable cluster and affected the classification accuracy. It is referred as curse of dimensionality problem in literature [28].

We can deduce the optimal number of projected dimensions with respect to different number of final clusters from Figure 5.1. According to our experiment results, if the number of final clusters is smaller than 24, the projected clusters formed by 35 principal components give the highest classification rate. However, projected clusters generated by 20 principal components give the highest classification accuracy when the number of final clusters are larger than 24. It is explainable that the smaller in the number of the final clusters, cluster centers are far away to each other. The distance between those testing samples and different resulting projected cluster centers are always in greater differences. Therefore, the higher dimensionality may not degrade the classification accuracy very much and the additional dimensions may give more relevant information for the testing samples to distinguish the correct clusters from the rest of undesired clusters. That is the reason for those projected clusters with higher dimensionality classified those testing samples better if the number of final clusters is small. If the number of projected clusters increased, the distance between them decreased. The negative effects of increasing dimensionality of the projected clusters overwrite its benefits and then the drop in performance occurred. That is the reason why projected clusters with 20 dimensions give higher classification rate when compare with projected clusters with 35 dimensions if the number of clusters larger than 24 in Figure 5.1.

Since our reduced dataset consists of 35 dimensions, the projected clusters with 35 dimensions do not provide any advantages in dimensional reduction aspect and it may not very interesting for further studies. However, the classification accuracy of projected clusters with 20 and 15 projected dimensions is very close to those projected clusters with 35 projected dimensions as shown in Figure 5.1 and they are more applicable in real applications. In Figure 5.3, we provide a detail investigation to the classification performances of those projected clusters with 15 to 20 dimensions. In terms of classification accuracy, we cannot found an optimal number of dimensions easily from Figure 5.3 that can outperform the others in cancer detection. Since the differences in classification rate between those projected clusters as shown in Figure 5.3 is very small. We use the classification rate of projected clusters consists of 17 projected dimensions as representative, as shown in Table 5.5, for comparison purpose in the rest of the paper, because its performance is generally good in different number of final clusters.

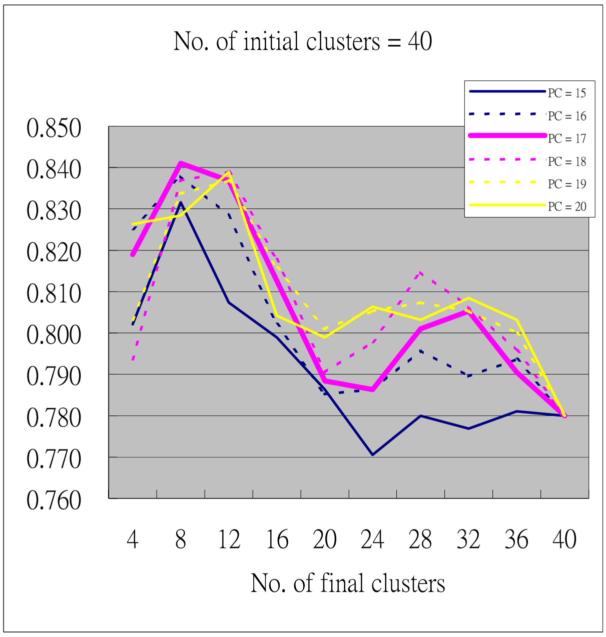

Figure 5.3 Classification accuracy of ORCLUS projected clusters (Initial cluster = 40)

Table 5.5 Classification accuracy of ORCLUS projected clusters

(No. of PC = 17; 70% training data)

	Final Cluster Num.										
Initial Cluster Num	**.**	**4**	**8**	**12**	**16**	**20**	**24**	**28**	**32**	**36**	**40**
	8	0.846	0.838								
	16	0.844	0.841	0.819	0.809						
	24	0.833	0.845	0.843	0.807	0.815	0.796				
	32	0.829	0.833	0.815	0.814	0.799	0.799	0.804	0.783		
	40	0.819	0.841	0.837	0.813	0.788	0.786	0.801	0.805	0.791	0.780

5.2.2.2 EPPC projected clusters

In this experiment, we study the effect of using different number of EPs to generate m-EPPCs by our proposed EPPC algorithm on the performance of the classification. We also used Euclidean distance as a metrics for the m-EPPC to obtain projected clusters in order to compare with ORCLUS. In this test, we employ the same set of training and testing samples that we used in previous experiment, i.e. 70% of samples from the dataset are selected and ordered randomly for training. Sets of EPs, consists of total 1, 3, 5, 6, 7, 8 and 10 EPs, are used to generate projection dimensions for the projected clusters in different sets of experiments and the average classification accuracy of those 50 repetitions are studied with different number of initial clusters and final clusters.

According to the experimental results, the relationship between the classification rate and the number of EPs used for dimension projection in EPPC is similar among different number of initial clusters and we used the Figure 5.4 with 40 initial clusters for illustration. Figure 5.4 shows the classification accuracy of emerging pattern based projected clusters (m-EPPCs) and we found that if only one EP is used to generate the projected clusters, the classification rate such 1-EPPCs are not good enough when compared with 3-EPPCs and other m-EPPCs ($1<m\leq10$) in any number of final clusters. The major reason is that a single EP is unlikely to be adequate to capture all the significant dimensions that should be included in desired projected clusters. In most of the cases, using set of EPs to generate projected clusters can include more relevant and significant attributes and thus they always obtain higher classification rate as shown in our results. In general, more EPs used in generating projected clusters, the higher the classification accuracy can be obtained. In Figure 5.4, we can observe this trend by considering the improvement of classification performance from 1-EPPCs to 5-EPPCs. However, the situation becomes messy when the number of EPs used is more than 5. 5-EPPCs, 7-EPPCs and 10-EPPCs give the optimal classification accuracy in different number of final clusters but when the number of final clusters is smaller than 24, 10-EPPCs give better results and 5-EPPCs give the best classification rate when the number of final cluster is larger than 24. These findings are very similar to the results obtained in the previous set of experiments about the ORCLUS projected clusters studies. That is the limited number of final clusters mean the distance between them are larger and clusters with higher dimensionality may give results that better than the lower dimensional clusters. In these studies, they proved that the dimension

projection by using EPs is reasonable since the experimental results obtained by m-EPPCs can be explained thoroughly by principles of clusters' dimensionality.

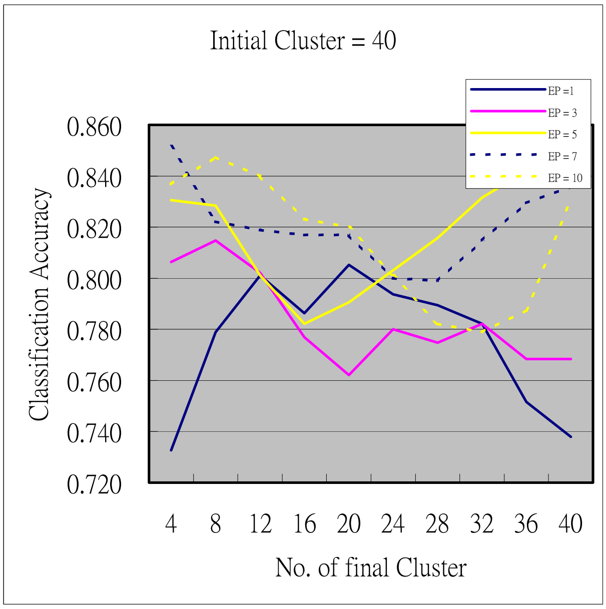

Figure 5.4 Classification accuracy of EPPC projected clusters (Initial cluster = 40)

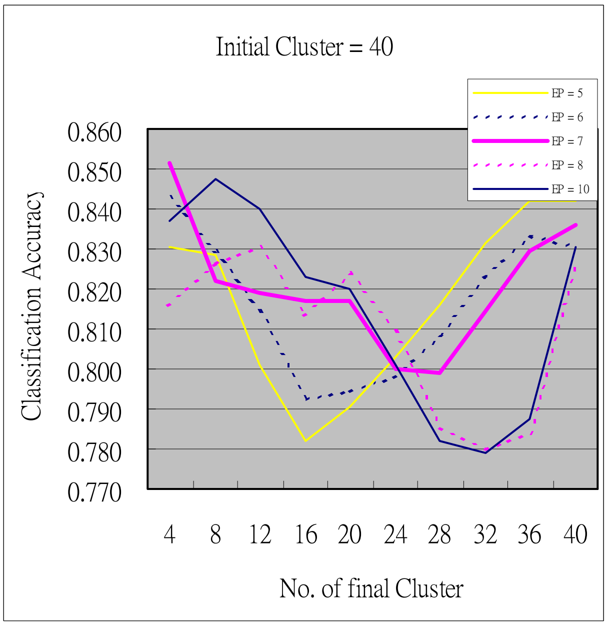

Figure 5.5 Classification accuracy of EPPC projected clusters (Initial cluster = 40)

It is interesting that the classification rates obtained by 1-EPPCs are especially low when the number of final clusters either very small or very large. Its shape in Figure 5.4 shows that it is totally different from the others m-EPPCs. The reason behind is that if the number of final clusters are limited, the set of projected clusters that generated by single EP are not enough to capture the real world since single EP can only provide limited information for the projected dimensions. On the other hand, if there are too many final clusters, some of clusters' projected dimensions will be very similar because similar EPs are likely to be employed in the generation of projected dimensions and those similar clusters with relatively smaller inter-cluster distance may not distinguish the samples accurately.

From Figure 5.4, we found that the projected clusters generated by set of EPs with a number of used EP larger than 5 give very impressive classification rates in different number of final cluster. In Figure 5.5, we take a closer look to the classification accuracy from 5-EPPCs to 10-EPPCs. 5-EPPCs give the best classification result when the number of final cluster is larger than 24 but it does not perform well if the number of final cluster is less than 24. 10-EP project clusters perform similarly in opposite manner. However, 7-EP projected clusters perform well in all combination of final clusters generally and it is used as the representative in the comparative studies between ORCLUS projected clusters and m-EPPCs. The performances of 7-EPPCs in different environments are summarized in Table 5.6.

Table 5.6 Classification accuracy of 7-EPPC projected clusters (70% training data)

	Final Cluster Num.										
Initial Cluster Num	**.**	**4**	**8**	**12**	**16**	**20**	**24**	**28**	**32**	**36**	**40**
	8	**0.852**	0.838								
	16	0.839	0.823	0.807	0.807						
	24	**0.857**	0.834	0.815	0.804	0.798	**0.807**				
	32	**0.844**	0.829	**0.822**	**0.826**	**0.816**	**0.802**	0.803	**0.823**		
	40	**0.852**	0.822	0.819	**0.817**	**0.817**	**0.800**	0.799	**0.815**	**0.829**	**0.836**

5.2.2.3 Comparative studies on ORCLUS and EPPC

In above experiments, we have examined the performances of ORCLUS projected clusters and m-EPPCs in the classification problem. Moreover, we have selected one representative from both technologies that have generally good performance in most of the examined conditions for comparison. They are the projected clusters with 17 principal components as projected dimensions obtained by ORCLUS and projected clusters with projected dimensions generated by set of 7 EPs with EPPC.

Their performances in classification respected to different combinations of initial and final clusters number are compared and shown in the Table 5.6. The bolded entries in Table 5.6 are the experimental results that 7-EPPCs give a better performance when compare with ORCLUS projected clusters with 17 projected dimensions in Table 5.5. We found that the performance of 7-EPPCs obtained from our proposed EPPC algorithm is slightly better than the representative ORCLUS projected clusters. We got 16 cases out of 30 that give better classification accuracy.

Table 5.7 Summary of ORCLUS and EPPC

	ORCLUS	**EPPC**
Use of class label (domain knowledge)	No	Yes
User inputs	No. of final cluster No. of principal component for dimension projection	No. of final cluster No. of EPs for dimension projection
Dimension projection	Using principal component	Using EPs
Individual set of dimensions for each cluster	Yes	Yes
Projected dimension	Linear combination of all existing attributes	Collection of attributes
No. of projected dimension for each cluster	Same in each cluster	Vary in different clusters
Classification accuracy	High	High
Readability	Bad	Good

In addition to the classification power, other differences between ORCLUS and EPPC are summarized in Table 5.7. The major difference of EPPC is that it utilized the information in predefined classes, domain knowledge, which always available in gene expression data but ORCLUS do not make use of it. In the readability aspect, the result clusters' projected dimension of EPPCs are more easy to interpret. Further discuss on readability are available in later section.

5.2.2.4 Readability – projected dimensions of resulting projected clusters

In traditional clustering algorithms, the resulting clusters only provide information on those data points that are said to be similar under a predefined distance metric. Most of the distance metrics, such as Euclidean distance, are geometrically meaningful with respect to the full dimensional data space. But it is questionable to group data points in meaningful clusters [5, 28] and it would be very difficult for users to interpret when the number of dimensions of data is large.

Projected clustering works on a step further and it provides flexibility to individual clusters in having their own set of dimensions that they are significant to group the data points. More meaningful clusters can be formed in different subspaces instead of using the full data space and it successfully tackled the curse of dimensionality problem. ORCLUS use the principal components generated from the covariance matrix of data as the projected dimensions of its projected clusters. In our previous experiment, we show that ORCLUS is powerful to form projected clusters and those resulting clusters are applicable in the cancer detection (classification) with high classification accuracy. However, those projected dimensions in terms of principal components are not easy to interpret by users, especially for the biologists. A sample of tumor cluster with 5 projected dimensions is shown in

Table 5.8. Each dimension of ORLCUS projected clusters is equivalent to a column in Table 5.8. Each projected dimension can be interpreted as a linear combination of 35 original dimensions in reduced dataset and they are never easy for user to interpret.

Table 5.8 Sample Tumor cluster - 5 dimensions projected cluster generated by ORCLUS

Gene No.	**D1**	**D2**	**D3**	**D4**	**D5**
M26383	0.097	0.239	-0.062	-0.030	0.051
M63391	-0.224	0.013	-0.277	-0.026	-0.074
R87126	0.163	-0.142	0.094	-0.017	0.300
M76378	-0.202	-0.053	0.551	0.064	-0.188
H08393	0.037	0.015	-0.040	0.002	0.025
X12671	0.077	0.034	0.011	-0.013	-0.029
R36977	-0.021	-0.051	-0.023	-0.017	-0.009
J02854	-0.011	0.271	0.065	-0.111	0.270
M22382	-0.012	-0.056	-0.114	0.010	-0.020
J05032	0.008	0.090	0.170	-0.013	-0.022
M76378	-0.157	0.040	0.029	0.068	-0.105
M76378	-0.055	0.052	-0.341	0.088	-0.090
M16937	-0.034	0.073	0.114	-0.018	-0.088
H40095	-0.012	-0.027	-0.309	0.003	-0.031
U30825	0.107	0.073	0.103	-0.080	0.090
H43887	-0.112	-0.084	-0.124	-0.046	-0.169
X63629	0.036	-0.030	0.162	0.029	-0.114
H23544	0.195	-0.084	-0.176	-0.027	-0.167
R10066	-0.060	-0.085	0.145	0.011	-0.077
T96873	0.045	-0.018	0.268	-0.033	0.035
T57619	0.001	0.253	-0.008	-0.034	0.015
R84411	0.147	0.110	0.003	-0.035	-0.254
U21090	-0.231	-0.336	-0.272	-0.083	0.425
U32519	-0.238	-0.396	0.086	-0.031	0.026
T71025	0.076	-0.120	0.167	-0.063	0.461
T92451	-0.169	-0.140	0.044	-0.052	-0.239
U09564	0.422	0.229	-0.048	0.024	0.067
H40560	0.177	-0.245	-0.155	-0.021	-0.055
T47377	-0.123	0.192	-0.007	-0.042	0.209
X53586	-0.064	0.013	0.047	-0.179	0.125
U25138	0.340	-0.070	-0.076	0.011	-0.192
T60155	0.178	0.119	0.009	-0.057	0.167
H55758	-0.024	-0.030	-0.012	-0.080	-0.092
Z50753	-0.474	0.493	-0.106	0.016	0.028
U09587	0.001	-0.011	0.000	0.946	0.145

Our EPPC algorithms formulate projected clusters by using the discrimination power of the emerging patterns. In previous experiments, we showed that EPPC can form reliable projected clusters and those resulting clusters are also applicable in cancer detection problem. The classification accuracy of EPPC is comparable or even better than ORCLUS in some situations. More important is that the projected clusters formed by EPPC are just collections of attributes. They are easy to interpret by the users. Three samples of EPPCs are extracted and shown in Table 4.1 previously. In that example, it is not difficult to understand that cluster 2 consists of 6 cancerous tissues that those tissues samples are similar in gene expression values with respect to two suspecting gene M76378 and T47377.

5.3 More classifications using n-EPPCs

In this section, we used different cancer gene expression datasets to evaluate the performance of m-EPPCs in cancer classification problem. Again, we tested projected clusters for classification with different combinations of projected dimensions, initial and final clusters. In addition to above three parameters, we also test m-EPPCs with different size of reduced datasets that having different number of selected attributes. The classification accuracies of state-of-art ORCLUS projected clusters are used to compare with m-EPPCs.

5.3.1 Datasets and experimental settings

In this set of experiments, we experiment three more gene expression data are employed together with the colon tumor [36] mentioned in previous section. They are ALL-AML leukemia [44], ovarian [45] and lung [46] cancer data and their details are list in Table 5.9. Initially, data sources that are not predefined into two partitions with be first divided with 70% of instances as training data and the rest of instances are identified as testing data. Then, entropy discretization method [42] provided by MLC++ is applied to the training dataset. The entropy method finds sets of significant genes that are most relevant to classify those training samples and we pick different number of top-ranked genes (attributes) with smallest entropy values to form reduced datasets with different size for our experiments. Value of reduced datasets is being normalized and the mean of normalized gene expression values is equal to zero and standard deviation is equal to one.

Finally, the testing dataset is also reduced its size with those selected top-ranked genes and normalized with values obtained in training samples.

Table 5.9 Cancer gene expression data details

Cancer type	No. of class	Predefined training and testing partition	No. of instance			No. of attributes
			Training	Testing	Total	
ALL-AML	2	Yes	27 (ALL) 11 (AML)	20 (ALL) 14 (AML)	38 (training) 34 (testing)	7129
Colon	2	No	28 (negative) 15 (positive)	12 (negative) 7 (positive)	43 (training) 19 (testing)	2000
Lung	2	Yes	16 (MPM) 16 (ADCA)	15 (MPM) 134 (ADCA)	32 (training) 149 (testing)	12533
Ovarian	2	No	64 (normal) 113 (cancer)	27 (normal) 49 (cancer)	177 (training) 76 (testing)	15154

In the following experiments, we report the classification performance of our m-EPPCs and ORCLUS projected clusters. Four cancer data are used to demonstrate the usefulness of above projected clustering techniques in solving cancer classification problem. Reduced dataset with different number of available attributes are used to generate projected clusters with different combinations of the number of initial, final clusters and number of projected dimensions. The measurement of classification error we use here is similar to Section 5.1.1.

$$\text{Classification error} = \frac{\text{No. of testing samples assigned to wrong clusters}}{\text{total number of testing samples}}$$

Again, 50 repeated experiments with training and testing samples that selected and ordered randomly are used to minimize the effect of initial points and ordering problems occurred in clustering.

5.3.2 Classification performance

In this set of experiments, we studied the classification performance of ORCLUS and EPPC projected clusters for different cancer gene expression data. We first focus on the classification performance of EPPC under different size of reduced datasets. Then, we also evaluated classification performance with the relationship between different size of reduced datasets and different number of EPs used in dimension projections selectively. In order to minimize the bias in comparison, the performances of different sizes of reduced datasets are being evaluated by using the average classification error obtained from experiments using different combination of three experimental parameters. They are the number of projected dimensions of projected clusters, the number of initial clusters and the number of resulting final clusters. The performance of EPPC is compared with the state-of-art projected clustering techniques ORCLUS.

5.3.2.1 Different size of reduced dataset

Before the classification take place, it is often to reduced size of dataset to improve the performance in terms of speed and accuracy. In previous classification experiment of colon tumor dataset [36], we use a reduced dataset with top-35 ranked genes (attributes) obtained by entropy discretization method [42]. The reason for us to select 35 top-ranked genes is influenced by the works of Li [35, 43]. In those literatures, they have shown that reduced dataset with 35 top-ranked genes are well performed to classify the colon data [36]. However, different datasets may require different size of the reduced datasets to obtain optimal solutions. In this section, we analyze the classification performance of ORCLUS and m-EPPCs with reduced datasets contain 20, 30, 35, 40 and 50 top-ranked genes.

In comparing the classification error with different number of attributes of the reduced dataset for ORCLUS and m-EPPCs, we take the mean of classification error with different possible combinations of the number of initial, final clusters and projected dimensions. The possible testing combinations of these three parameters are listed in Table 5.10 below.

Table 5.10 Possible combinations of testing parameters

	No. initial clusters	**No. of final clusters**	**No. of projected dimension / EP**
ORCLUS	8, 16, 24, 32	4, 8, 12, 16, 20, 24, 28, 32	5, 10, 15, 20, 25, 30, 40, 50
EPPC	8, 16, 24, 32	4, 8, 12, 16, 20, 24, 28, 32	5, 10, 15, 20, 25, 30, 40, 50

According to our experiment results, we found that the optimal numbers of attributes in reduced datasets for those cancer datasets are not the same. In Figure 5.6, the experimental results of cancer classification with ORCLUS and EPPC projected clusters are shown. We observed that there is an increase in classification error with increasing the number of attributes in reduced dataset of ALL-AML. It suggested that the optimal number of attributes for ALL-AML reduced dataset should be less than 20. It is because additional attributes that are not useful in classifying data instances will increase the additional distance between them and finally cause the curse of dimensionality problem and lower down the classification accuracy. Classification of colon and lung cancer datasets are performing the best with 35 top-ranked attributes used and the optimal number of attributes for ovarian cancer classification is found to be 40 according our experiment results. These 3 datasets show a drop in classification error when increasing the number of top ranked attributes from 20 to their own optimal value and finally classification error increased again with additional attributes are used.

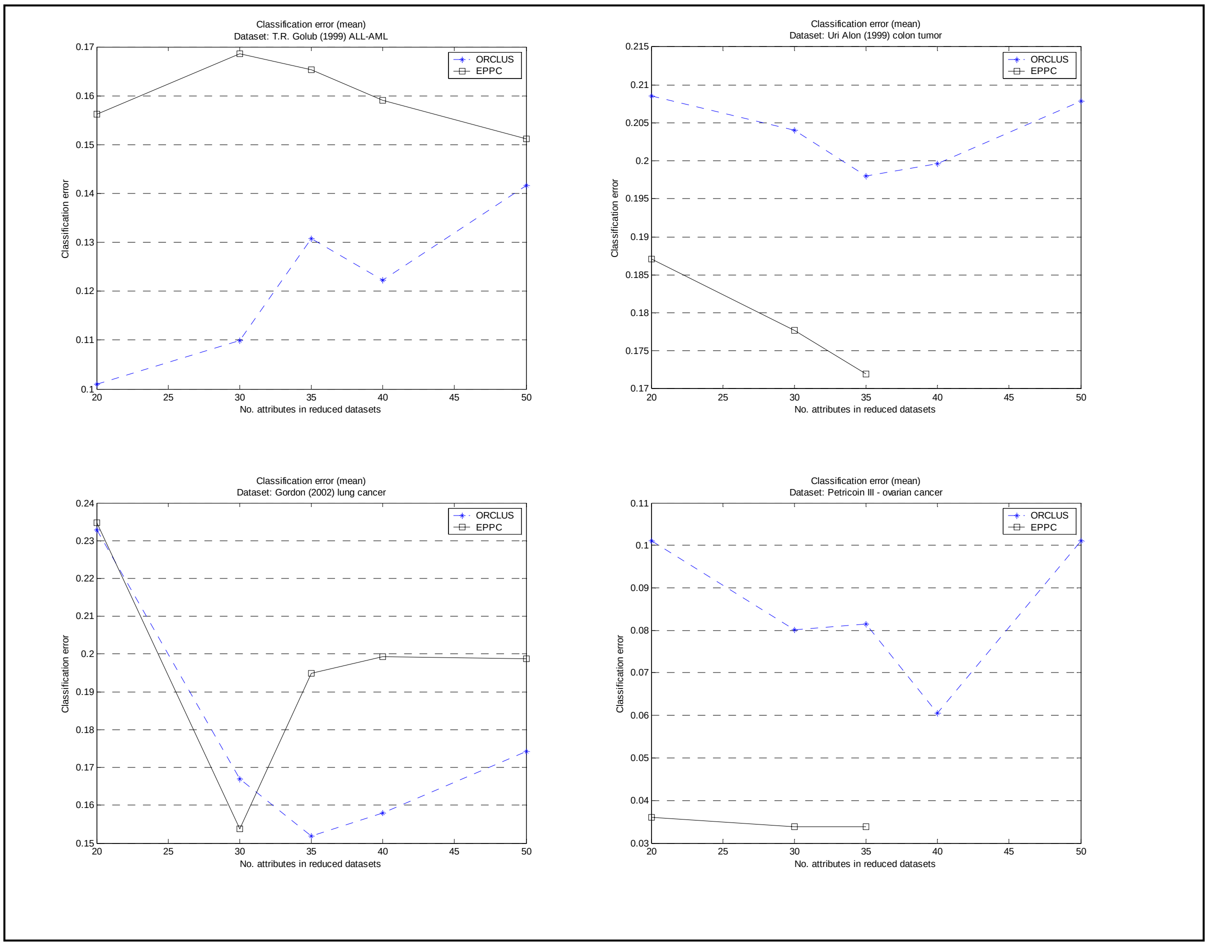

Figure 5.6 Classification error (mean) - different size of reduced dataset

In Figure 5.6, the experimental results of classification with EPPC are also shown. The shapes of the classification error curves that we have obtained are very similar to the line for ORCLUS. It proved that EPPC can also applicable to cancer classification like ORCLUS. It is interesting that the optimal number of attributes for the classification problem may not exactly the in those tested datasets. The major reason is that the dimension projection of EPPC and ORCLUS are not the same, they are using emerging patterns and principal components respectively, and therefore the number of attributes needed for forming reliable clusters may not be the same. These optimal numbers are datasets dependent and they are not predictable. Our experimental results have just show the approximation with limited data instances for problem space of these 4 cancer datasets.

In terms of the further knowledge discovery, the small the number of attributes needed in EPPC is likely to be beneficial. It is because the more the attributes we used, it will be more difficult for user to interpret for most of the cases. In Figure 5.6, the colon and ovarian dataset for EPPC are shown with the top-20, top-30 and top-35 reduced dataset only. The reason is that EPPC is now bottlenecked at the generation of EPs from data. Increase of attributes of dataset is still expensive in the process of EPs generation. In terms of classification error, these tested reduced datasets have already outperformed ORCLUS with all tested reduced dataset (for classification error, see Table 5.11 and Table 5.12 for details). Therefore, the further investigations with top-40 and top 50 reduced datasets are neglected here.

Table 5.11 ORCLUS Classification performance with different size of reduced dataset

Dataset	**Size of reduced dataset**	**Classification error (ORCLUS)**			
		Max	**Min**	**Average**	
ALL-AML	Top 20	0.2865	0.0294	0.1010	
	Top 30	0.2994	0.0359	0.1099	
	Top 35	0.3171	0.0465	0.1307	
	Top 40	0.3029	0.0371	0.1222	
	Top 50	0.3288	0.0535	0.1416	
Colon	Top 20	0.3126	0.1484	0.2085	
	Top 30	0.3137	0.1421	0.2040	
	Top 35	0.2716	0.1379	0.1980	
	Top 40	0.3032	0.1305	0.1996	
	Top 50	0.3095	0.1463	0.2079	
Lung	Top 20	0.4161	0.1544	0.2328	
	Top 30	0.3670	0.1007	0.1669	
	Top 35	0.2836	0.0847	0.1519	
	Top 40	0.3490	0.0871	0.1580	
	Top 50	0.3624	0.0819	0.1742	
Ovarian	Top 20	0.2729	0.0208	0.1010	
	Top 30	0.2571	0.0168	0.0802	
	Top 35	0.2600	0.0192	0.0815	
	Top 40	0.2661	0.0100	0.0605	
	Top 50	0.3008	0.0121	0.1010	

Table 5.12 EPPC Classification performance with different size of reduced dataset

Dataset	**Size of reduced dataset**	**Classification error (EPPC)**			**Orclus (Average) – EPPC (Average)**
		Max	**Min**	**Average**	
ALL-AML	Top 20	**0.1906**	0.1394	0.1562	-0.0552
	Top 30	**0.2141**	0.1465	0.1686	-0.0587
	Top 35	**0.2006**	0.1459	0.1653	-0.0346
	Top 40	**0.1882**	0.1400	0.1590	-0.0368
	Top 50	**0.2171**	0.1076	0.1511	-0.0095
Colon	Top 20	**0.2221**	**0.1389**	**0.1871**	**0.0214**
	Top 30	**0.2432**	**0.1389**	**0.1777**	**0.0263**
	Top 35	**0.2253**	**0.1368**	**0.1720**	**0.026**
	Top 40	NaN	NaN	NaN	NaN
	Top 50	NaN	NaN	NaN	NaN
Lung	Top 20	**0.3368**	0.1651	0.2348	-0.002
	Top 30	0.3672	**0.0340**	**0.1537**	**0.0132**
	Top 35	0.3934	**0.0695**	0.1949	-0.043
	Top 40	0.4217	**0.0443**	0.1994	-0.0414
	Top 50	0.4166	**0.0554**	0.1987	-0.0245
Ovarian	Top 20	**0.0737**	0.0213	**0.0360**	**0.065**
	Top 30	**0.0671**	0.0208	**0.0339**	**0.0463**
	Top 35	**0.0616**	0.0221	**0.0339**	**0.0476**
	Top 40	NaN	NaN	NaN	NaN
	Top 50	NaN	NaN	NaN	NaN

In general, we found that EPPC and ORCLUS both are very close in classification performance as shown in Table 5.11 and Table 5.12. The differences between the classification error is just around 5% and for those classification rate that EPPC outperform the ORCLUS are bolded in Table 5.12. We found that EPPC outperform the ORCLUS in colon and ovarian cancer data on average.

5.3.2.2 Different number of projected dimensions / EPs

In Section 5.2.2, we mentioned that the influences of initial clusters number for both ORCLUS and m-EPPCs are not that great if its value is larger than the number of nature clusters and we have also analyzed the relationship between the numbers of projected dimensions with the numbers of final clusters. In previous section, we found that m-EPPCs have comparable performance with ORCLUS or even outperform ORCLUS in some of the datasets. In this section, we focus on the relationship between the numbers of projected dimension and the sizes of reduced dataset with respect to the classification error.

According to our experiment results, we observed that there are some differences between ORCLUS and m-EPPCs under those situations with different combinations of Top-n reduced datasets and Top-n of projected dimensions. In Table 5.11 and Table 5.12, they show that the performance of ORCLUS for lung cancer dataset are slightly better than our m-EPPCs around 0.2%-4% with different sizes of reduced datasets and it shows that the ORCLUS under the larger size of reduced datasets, such as Top-35 and Top-40 cases, perform even better. In Figure 5.7, it shows the details in classification errors for both ORCLUS and EPPC with different sizes of reduced dataset against the number of projected dimensions separately. In general, we observed that with small number of projected dimensions, the performance of ORCLUS is better than EPPC, it is because the projected dimensions of ORCLUS are principal components that are the linear combinations of attributes in reduced datasets and normally each of them may likely to embed more information than an emerging pattern that is just a collection of few attributes. It is the major reason for ORCLUS gives better performance in the environment with small number of projected dimensions. While the number of projected dimension increased, the unions of EPs show their power in embedding discriminative information by providing much lower classification error in all tested reduced dataset of lung cancer data. Since the maximum principal components of ORCLUS are equivalent to the number of attributes in reduced dataset, we only compare performance of ORCLUS and EPPC with the same number of projected dimension and the majority of cases we compared are low in the number of projected dimensions. Therefore, the averaged classification error of m-EPPC for lung cancer dataset that shown in Table 5.11 and Table 5.12 is a little bit higher than ORCLUS. In Figure 5.7, we also found that the more the information for a clusters may not be a good news all the time in general. From those ORCLUS lines in Figure 5.7, they generally got a U-shape. It is because the classification accuracies are improved from not enough dimensions to an optimal point and they drop again when suffering the curse of dimensionality problem that caused by considering too many dimensions in ORCLUS. However, the EPs are collection of attributes in those reduced dataset, the union of EPs may not actually increase the number of dimension used all the time. Therefore, the classification performance of EPPC is much more stable in lung cancer that shown in Figure 5.7 with a L-shape.

Since the numbers of projected dimensions we choose to solve the classification problems are critical to the performance, it would be valuable if we can obtain a magic number in advance. However, it is data dependent and unpredictable in nature. We can only obtain it by some intelligent trial-and-error manner expensively [28]. Therefore, EPPC is more applicable to many problems since it is less sensitive to the available attributes in datasets and the number of projected dimensions that we choose.

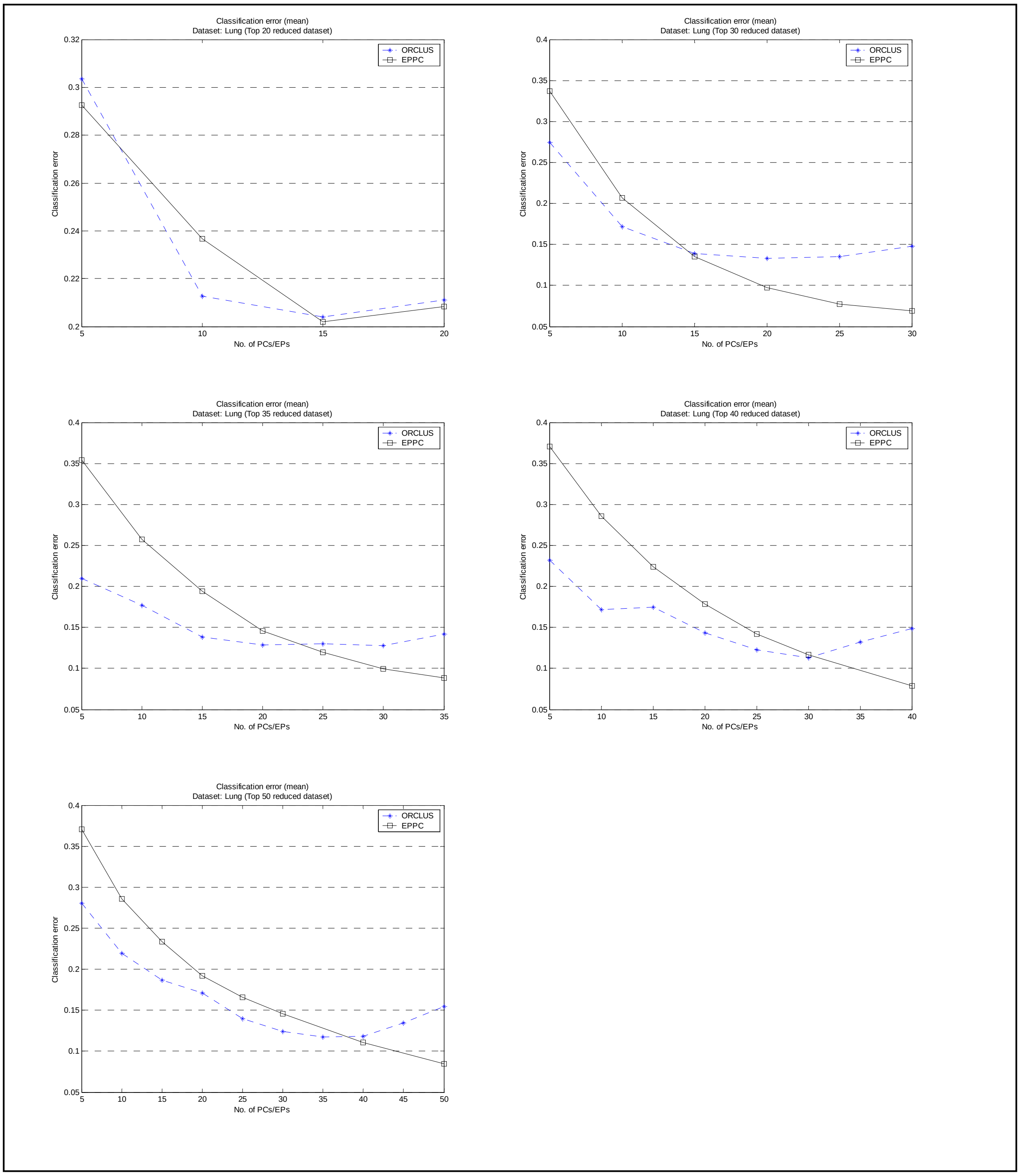

Figure 5.7 Classification error of Lung

(Top-n reduced dataset v.s. Top-n projected dimension)

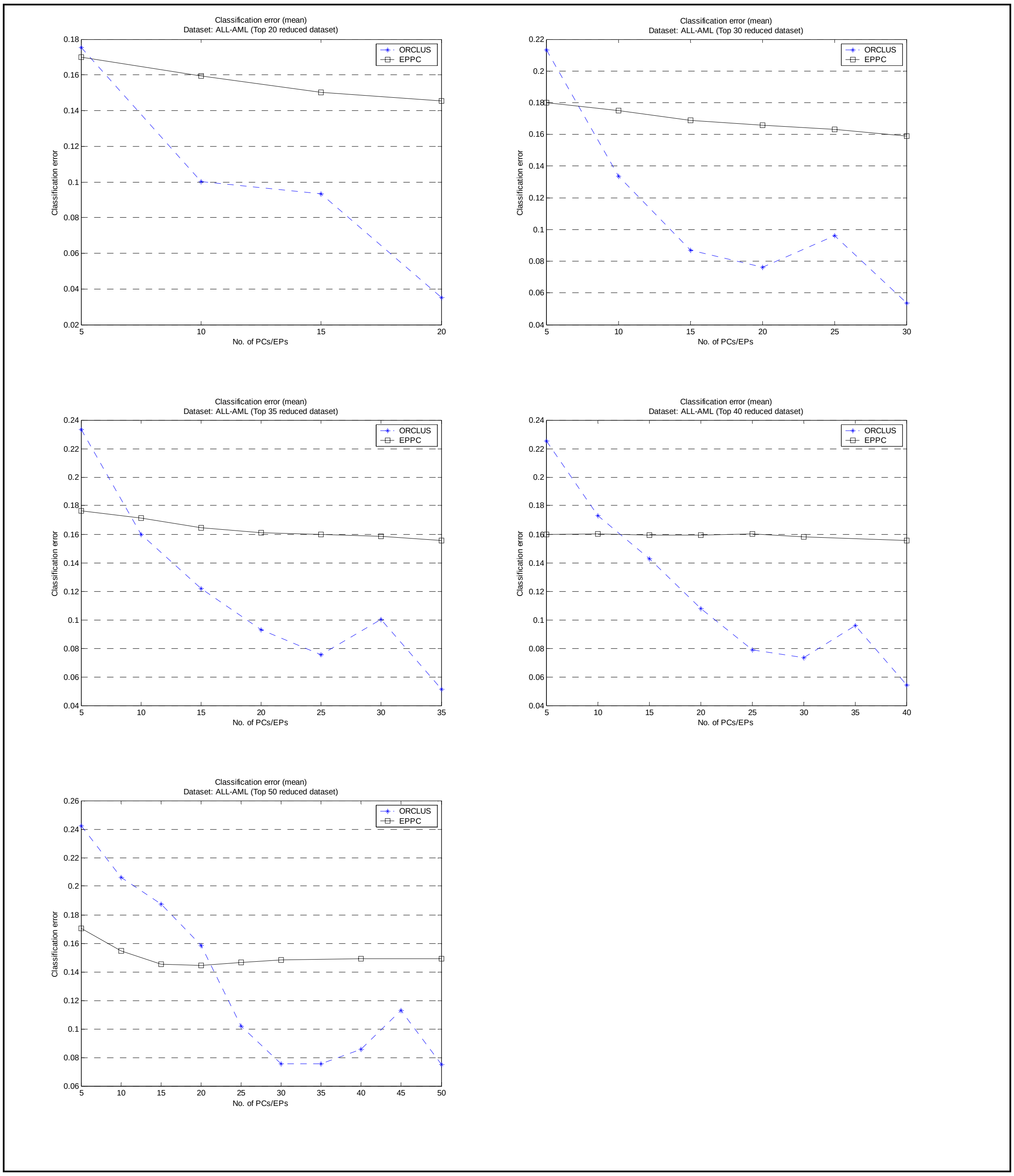

Figure 5.8 Classification error of ALL-AML

(Top-n reduced dataset v.s. Top-n projected dimension)

In Figure 5.8, we plot the classification error of the ALL-AML dataset. We found that m-EPPC is very stable in its classification performances and increases in number of projected dimension only improve the performance slightly. On the other hand, ORCLUS seems not worked as stable as EPPC, but it outperform EPPC algorithm in most of the cases with enough number of projected dimensions. The major reason of above findings is caused by the nature of those extracted EPs in ALL-AML dataset. We extracted large number of EPs with high occurrence in dataset but the portion of high occurrence EPs used in our experiments are not that large when comparing with experiments of lung cancer. Every EPs is a piece of knowledge extracted from the datasets and the top-m EPs that we selected are the most relevant knowledge for our problem. For ALL-AML dataset, it seems quite difficult to select out most significant, appropriate descriptions for the dataset without serious information lost and thus top-m EPs that we employed is limited in solving the problems. For lung cancer dataset, we got enough knowledge with top-m selected EPs with insignificant information lost. The usage details of mined EPs for these two datasets are shown in Table 5.13 for reference. In Figure 5.9 - Figure 5.11, we try to illustrate above hypothesis by introducing additional EPs in running EPPC for some of the ALL-AML reduced datasets and compared its performance with ORCLUS. We found that the extra EPs increased the portion of the use of high occurrence EPs in different amounts in Top-20, Top-30 and Top-35 reduced datasets respectively and the classification results show significant improvements. For example, in the case of Top-20 reduced dataset of ALL-AML, the EPPC gives comparable performance with ORCLUS.

On the other hand, the complicated principal components representations always do a better job in ALL-AML dataset. It is because the space described by principal components can be interpreted as the transformation of feature space and the information of the datasets are concentrated in those top ranked of components. Therefore, using small numbers of top ranked principal components are always good enough for the problem and improvements by further increasing the number of projected dimensions in ORCLUS are not as significant as EPPC.

One interesting points that we found from the results of ALL-AML experiments is that when very small number of projected dimensions are used. EPPC outperforms ORCLUS and it is very likely that EPPC can be identified most relevant knowledge from the dataset successfully.

Table 5.13 Percentage of high occurrence used in previous experiments

Dataset	**Size of reduced datset**	**Total no. of EPs**	**High Occurrence EPs**	**Max % of high occurrence EPs used in expt. shown in Figure 5.7, Figure 5.8**
ALLAML	Top20	81	81	24.69%
	Top30	228	218	13.76%
	Top35	438	310	11.29%
	Top40	717	468	8.55%
	Top50	1064	623	8.03%
Lung	Top20	39	39	51.28%
	Top30	95	95	31.58%
	Top35	144	144	24.31%
	Top40	193	193	20.73%
	Top50	344	322	15.53%

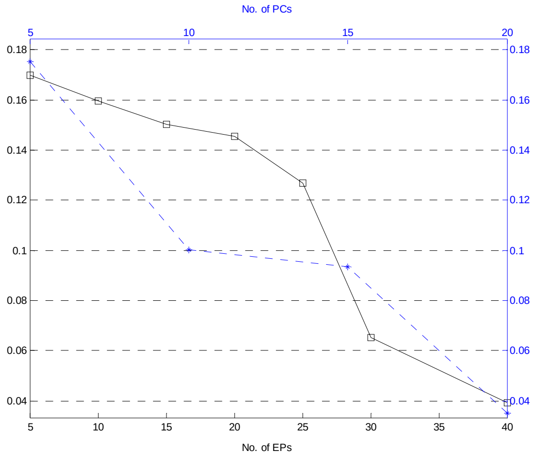

Figure 5.9 Classification Error (ALLAML - Top20 reduced dataset)

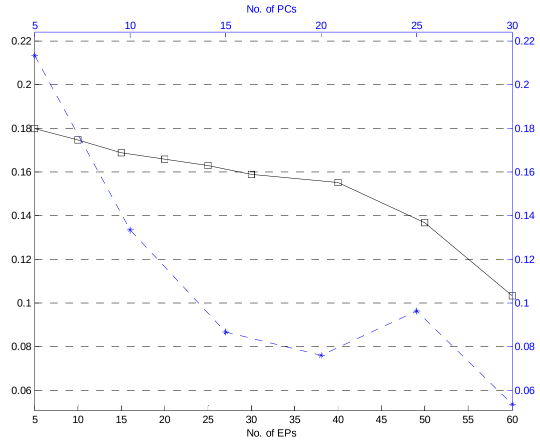

Figure 5.10 Classification Error (ALLAML –Top30 reduced dataset)

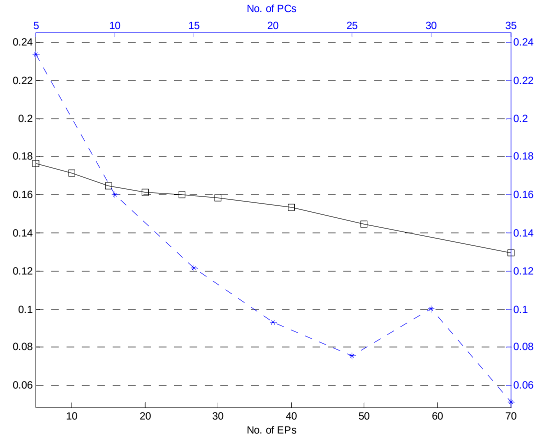

Figure 5.11 Classification Error (ALLAML –Top35 reduced dataset)

6 Conclusion and Future Works

In this thesis, we have studied one of the problems in the field of bioinformatics using data mining technique, it is the molecular classification of cancer by gene expression data. In this chapter, we summarize the results of our work in Section 6.1. In Section 6.2 - 6.4, we discuss the contributions of our works and some future research issues on Emerging Pattern-based Projected Clustering.

6.1 Summary of our works

We have investigated (in Chapter 2) the molecular classification of cancers which is a bioinformatics issue from the data mining point of view. We started from the motivation of the field bioinformatics, its definitions, its aims to the research trends and challenges of applying data mining in the problem of molecular classification of cancers. We identified the nature of this new type of data, the gene expression data, and the challenges that it caused for those well known data mining algorithms.

We have reviewed (in Chapter 3) two promising data mining techniques, emerging patterns and projected clustering. We have discussed the rationales of introducing these algorithms, their objectives, assumption they have made and their own framework. We have commented on their strength and weakness and before we show the opportunities of their integration.

We have proposed (in Chapter 4) the idea of Emerging Pattern-based Projected Clusters and its definitions. EPPC is introduced because of the insufficient of existing algorithm to provide easy understanding results of the problem of molecular cancer classification. Moreover, we have discussed the problem of finding EPPC and introduced a framework to find EPPC for the classification purpose.

We have evaluated (in Chapter 5) the performance of EPPC. We have used k-mean algorithm to compare with EPPC to show the dimension projection is essential for success. We has compared EPPC with state-of-art projected clustering algorithm ORCLUS in terms of classification accuracy, readability and stability.

6.2 Contributions

In this research, we have make contributions in the following ways:

1. Introduce the integration of two data mining techniques, they are the emerging patterns and projected clustering. In the past, these two techniques are used independently on different problems and we integrated them to get both their strengths.
2. Apply the emerging pattern based projected clustering techniques to molecular cancer classification by gene expression data. Gene expression data are high dimension in nature and it challenged most of the existing data mining algorithms. By using EPPC, we can classify cancer accurately by using gene expression data.
3. Improve the readability of projected clusters by using emerging patterns. Projected clusters are useful to handle the high dimensional data but it is not easy to understand. By using the emerging patterns in dimension projection process. The EPPCs are easy to understand and it facilitates the further knowledge discovery.

Publications:

1. Larry T. H. Yu, Fu-lai Chung and Stephen C. F. Chan, "Emerging Pattern Based Projected Clustering for Gene Expression Data," *Proceedings of European Workshop on Data Mining and Text Mining for Bioinformatics (with ECML/PKDD-2003)*, Dubrovnik, Croatia, pp. 71-75, 2003.

2. Larry T. H. Yu, Fu-lai Chung, Stephen C. F. Chan and Simon M. C. Yuen, "Using Emerging Pattern Based Projected Clustering and Gene Expression Data for Cancer," *Proceedings of 2nd conference on Asia-pacific Bioinformatics*, Dunedin, New Zealand, pp. 75-84, 2004.

6.3 Limitations

In this research, we have focused on the integration of emerging patterns and projected clustering techniques and our studies are limited in the following areas.

1. There are always huge amount of JEPs can be extracted for EPPC and it is possible to obtain better EPs by using constraints in mining EPs. However, mining EPs with constraints more close to the researches about mining EPs and it is not included in our studies at this moment.
2. Emerging patterns used in this research are JEPs with growth rate equal to infinity and it gave us promising results. Theoretically, not only JEPs are applicable to EPPC algorithm and the properties of using other types of EPs for EPPC not being studied at this stage. Moreover, using other data mining patterns, such as frequent patterns, together with projected clustering technique are not covered here.

6.4 Future works

We list below main problems of our interest for further research topics.

1. Theoretically, all EPs can be extracted by border based algorithms. In this research, we may be only interested in those top *n* (say 50) high occurrence JEPs in complex left boundary in the dimension projection process. Our approach is to select JEPs with maximum occurrence. If the size of candidate set is larger than *n*, we select *n* of them randomly. However, constraints are often existed in practical problems.

Instead of select JEPs with considering their occurrence, it is a problem for us to study in applying constraints in the problem of JEPs selection.

2. By definition, JEPs is the EPs with maximum discrimination power. However, EPs with high growth rate, or even different types of patterns, are also applicable in concept of EPPC potentially. It is an interesting topic to introduce other patterns in forming useful clusters for different type of data or problems.
3. Although n-EPPCs are readable in nature, it would be fruitful if we can visualize those resulting clusters interactively. Interactive visualization can enhance the users to discover further knowledge from inter-clusters relationships and intra-cluster relationships.

References

[1] M. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? A proposed definition and overview of the field," *Methods of Information in Medicine*, vol. 40, pp. 346-58, 2001.

[2] P. L. Elkin, "Primer on medical genomics part V: Bioinformatics," *Mayo Clinic Proceedings*, vol. 78, pp. 57-64, 2003.

[3] J. Wiemer, F. Schubert, M. Granzow, T. Ragg, J. Fieres, J. Mattes, and R. Eils, "Informatics united - Exemplary studies combining medical informatics, neuroinformatics, and bioinformatics," *Methods of Information in Medicine*, vol. 42, pp. 126-133, 2003.

[4] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," *Proceedings of Fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, California, United States, pp. 43-52, 1999.

[5] C. C. Aggarwal, C. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park, "Fast algorithms for projected clustering," *SIGMOD Record*, vol. 28, pp. 61-72, 1999.

[6] Cancer screening overview. National Cancer Institute, 2004, http://www.cancer.gov/cancerinfo/pdq/screening/overview

[7] Understanding cancer. National Cancer Institute, 2004, http://press2.nci.nih.gov/sciencebehind/cancer/cancer00.htm

[8] C. A. Ouzounis and A. Valencia, "Early bioinformatics: The birth of a discipline - a personal view," *Bioinformatics*, vol. 19, pp. 2176-2190, 2003.

[9] J. Li, S.-K. Ng, and L. Wong, "Bioinformatics adventures in database research," *Proceedings of 9th International Conference on Database Theory - ICDT 2003*, Siena, Italy, pp. 31-46, 2003.

[10] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, pp. 2444-2448, 1988.

[11] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389-3402, 1997.

[12] T. Lengauer, "Algorithmic research problems in molecular bioinformatics," *Proceedings of 2nd Israel Symposium on the Theory and Computing Systems*, pp. 177-192, 1993.

[13] J. Li and L. Wong, "Emerging patterns and gene expression data," *Proceedings of Workshop on Genome Informatics*, Tokyo, Japan, pp. 3-13, 2001.

[14] P. Durand, C. Medigue, A. Morgat, Y. Vandenbrouck, A. Viari, and F. Rechenmann, "Integration of data and methods for genome analysis," *Current Opinion in Drug Discovery & Development*, vol. 6, pp. 346-352, 2003.

[15] T. Yao, "Bioinformatics for the genomic sciences and towards systems biology. Japanese activities in the post-genome era," *Progress in Biophysics & Molecular Biology*, vol. 80, pp. 23-42, 2002.

[16] R. Molidor, A. Sturn, M. Maurer, and Z. Trajanoski, "New trends in bioinformatics: From genome sequence to personalized medicine," *Experimental Gerontology*, vol. 38, pp. 1031-1036, 2003.

[17] F. Martin-Sanchez, V. Maojo, and G. Lopez-Campos, "Integrating genomics into health information systems," *Methods of Information in Medicine*, vol. 41, pp. 25-30, 2002.

[18] What you need to know about cancer. National Cancer Institute, 2004, http://www.cancer.gov/cancerinfo/wyntk/

[19] Cancer imaging. National Cancer Institute, 2004, http://www.nci.nih.gov/clinicaltrials/learning/science-explained-imaging

[20] Interpreting laboratory test results. National Cancer Institute, 2004, http://cis.nci.nih.gov/fact/5_27.htm

[21] Cancer genetic overview. 2004, http://www.nci.nih.gov/cancerinfo/pdq/genetics/overview

[22] A science primer. National Center for Biotechnology Informfation, 2004, http://www.ncbi.nlm.nih.gov/About/primer/index.html

[23] J. Han and M. Kamber, *Data mining: Concepts and techniques*, 1st ed: Morgan Kaufmann, 2001.

[24] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds.: AAAI Press/The MIT Press, 1996, pp. 1-34.

[25] S. Kasif, "Datascope: Mining biological sequences," *IEEE Intelligent Systems*, vol. 14, pp. 38-43, 1999.

[26] L. Wong, "Datamining: Discovering information from bio-data," in *Current Topics in Computational Biology*, T. Jiang, Y. Xu, and M. Zhang, Eds.: MIT Press, 2003, pp. 317 - 342.

[27] A. Brazma, A. Robinson, G. Cameron, and M. Ashburner, "One-stop shop for microarray data - Is a universal, public DNA-microarray database a realistic goal?," *Nature*, vol. 403, pp. 699-700, 2000.

[28] C. C. Aggarwal and P. S. Yu, "Redefining clustering for high-dimensional applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, pp. 210-25, 2002.

[29] C. C. Aggarwal and P. S. Yu, "Finding generalized projected clusters in high dimensional spaces," *SIGMOD Record*, vol. 29, pp. 70-81, 2000.

[30] Y. Lu and J. W. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, pp. 243-268, 2003.

[31] W. Dubitzky, M. Granzow, and D. Berrar, "Data mining and machine learning methods for microarray analysis," *Methods of Microarray Data Analysis: Papers from CAMDA'00*, pp. 5-22, 2001.

[32] S.-B. Cho and H.-H. Won, "Data mining for gene expression profiles from DNA microarray," *International Journal of Software Engineering*, vol. 13, pp. 593-608, 2003.

[33] B. S. Everitt, S. Landua, and L. Morven, *Cluster Analysis*, 4th ed: Arnold, 2001.

[34] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," *Proceedings of International Conference on Database Theory*, Jerusalem, Israel, pp. 217-35, 1999.

[35] J. Li and L. Wong, "Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns," *Bioinformatics*, vol. 18, pp. 725-734, 2002.

[36] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, pp. 6745-6750, 1999.

[37] G. Dong, J. Li, and X. Zhang, "Discovering jumping emerging patterns and experiments on real datasets," *Proceedings of 9th International Database*

Conference on Heterogeneous and Internet Databases (IDC99), Hong Kong, pp. 155-168, 1999.

[38] J. Li, G. Dong, and K. Ramamohanarao, "Instance-based classification by emerging patterns," *Proceedings of 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2000)*, Lyon, France, pp. 191-200, 2000.

[39] J. Li, K. Ramamohanarao, and G. Dong, "Emerging patterns and classification," *Proceedings of 6th Asian Computing Science Conference on Advances in Computing Science - ASIAN 2000*, Penang, Malaysia, pp. 15-32, 2000.

[40] G. Dong, X. Zhang, L. Wong, and J. Li, "CAEP: Classification by aggregating emerging patterns," *Proceedings of Second International Conference on Discovery Science (DS'99)*, Tokyo, Japan, pp. 30-42, 1999.

[41] J. Li, G. Dong, K. Ramamohanarao, and L. Wong, "DeEPs: A new instance-based discovery and classification system," *Machine Learning*, vol. 54, pp. 99-124, 2004.

[42] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger, "MLC++: a machine learning library in C++," *Proceedings of Sixth International Conference on Tools with Artificial Intelligence. TAI 94*, New Orleans, LA, USA, pp. 740-3, 1994.

[43] J. Li and L. Wong, "Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns," *Bioinformatics*, vol. 18, pp. 1406-1407, 2002.

[44] T. R. Golub, D. K. Slonim, P. Tamayo, 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, pp. 531-537, 1999.

[45] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, pp. 572-577, 2002.

[46] 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, pp. 4963-4967, 2002.