CHAPTER 1

INTRODUCTION

1.1 DATA MINING

Data mining plays an important role in the information age. Data mining refers to extracting or mining knowledge from large amounts of data. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both “data” and “mining” became a popular choice. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. The discovered knowledge can be applied to decision making, process control, information management, and query processing. This chapter discusses about certain concepts of data mining, genetic information and issues in retrieving the gene data.

1.1.1 Data Mining Process

Data mining is an iterative process that involves the following phases:

**Problem definition:** A data mining project starts with the understanding of the problem. Data mining experts and domain experts work closely together to define the project objectives and the requirements from an application perspective. The project objective is then translated into a data mining problem definition.

**Data exploration:** Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the application experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.
**Data preparation:** Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

**Modeling:** Data mining experts select and apply various mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. When the final modeling phase is completed, a model of high quality has been built.

**Evaluation:** Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

**Deployment:** Data mining experts use the mining results by exporting the results into database tables or into other applications.

![Data Mining Process Model](image)
Data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in the information technology. Databases containing number of data in the order 109 and dimension in the order of 103 are becoming increasingly common. When the scale of data manipulation, exploration and inference goes beyond human capacities, people look to computing technologies for automating the process. These have prompted the need for intelligent data analysis methodologies, which could discover useful knowledge from data. The term KDD refers to the overall process of knowledge discovery in databases. Data mining is a step in the KDD process consisting of a particular enumeration of patterns over the data, subject to some computational limitations says Peyman et al (2013). It uses automated tools employing sophisticated algorithms to discover hidden patterns, associations, anomalies and/or structure from large amounts of data stored in data warehouses or other information repositories. Fields concerned with inferring models from data include statistical pattern recognition, applied statistics, machine learning and neural computing. Data mining is a form of knowledge discovery essential for solving problems in a specific domain.

1.1.2 Data Mining in Medicine

Data mining has been exploited to retrieve the valuable information in a wide spread fields, especially, in DNA microarray technology. Medical data, for example, often exists in vast quantities in an unstructured format. Medical data, however, requires a large amount of preprocessing in order to be useful. A robust preprocessing system is required in order to extract any kind of knowledge from even medium-sized medical data sets. The data must not only be cleaned of errors and redundancy, but organized in a fashion that makes sense to the problem.

1.1.3 Data Mining in Pattern Recognition

Data mining involves fitting models or determining patterns from observed data. The fitted models play the role of inferred knowledge. (Gaber 2010) explains
deciding whether the model reflects useful knowledge or not, is a part of the overall KDD process for which subjective human judgment is usually required. Two major reasons why a pattern is interesting from the subjective (user-oriented) point of view by (Laurent 2007) are as follows.

- Unexpectedness: when it is ‘surprising’ to the user.
- Actionability: when the user can act on it to her or his advantage.

Though both these concepts are important it has often been observed that actionability and unexpectedness are correlated. In literature, unexpectedness is often defined in terms of the dissimilarity of a discovered pattern from a vocabulary provided by the user. Data mining tasks can be descriptive, i.e., discovering interesting patterns describing the data, and predictive, i.e., predicting the behavior of the model based on available data.

1.1.4 Data Mining Components

Typically, a data mining algorithm constitutes some combination of the following three components.
- The model: The function of the model (e.g., classification, clustering) and its representational form (e.g., linear discriminant, neural networks). A model contains parameters that are to be determined from the data.
- The preference criterion: A basis for preference of one model or set of parameters over another, depending on the given data.
- The search algorithm: The specification of an algorithm for finding particular models and parameters, given the data, model(s), and a preference criterion. A particular data mining algorithm is usually an instantiation of the model or preference or search components. The more common model functions in current data mining practice include:
  1. Classification: classifies a data item into one of several predefined categorical classes.
  2. Regression: maps a data item to a real-valued prediction variable.
3. Clustering: maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models.
4. Rule generation: extracts classification rules from the data.
5. Discovering association rules: describes association relationship among different attributes.
6. Summarization: provides a compact description for a subset of data.
7. Dependency modeling: describes significant dependencies among variables.
8. Sequence analysis: models sequential patterns, like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time.

1.1.5 Challenges in Data Mining

The rapid growth of interest in data mining is due to the
1. Falling cost of large storage devices and increasing ease plenary field with a general goal of predicting outcomes and uncovering relationships in data. It uses automated tools employing sophisticated algorithms to discover hidden patterns, associations, anomalies and or structure from large amounts of data stored in data warehouses or other information repositories.
2. User interaction and prior knowledge. Data mining is inherently an interactive and iterative process says Ramamohanarao et al (2005). Users may interact at various stages, and domain knowledge may be used either in the form of a high level specification of the model, or at a more detailed level. Visualization of the extracted model is also desirable.
3. Overfitting and assessing the statistical significance. Data sets used for mining are usually huge and available from distributed sources.
4. Understandability of patterns. It is necessary to make the discoveries more understandable to humans. Possible solutions include rule structuring, natural language representation, and the visualization of data and knowledge.
5. Nonstandard and incomplete data. The data can be missing and or noisy.
6. Mixed media data. Learning from data that is represented by a combination of various media like, say numeric, symbolic, images and text.

7. Management of changing data and knowledge. Rapidly changing data, in a database that is modified or deleted or augmented, may make previously discovered patterns invalid. Possible solutions include incremental methods for updating the patterns.

8. Integration. Data mining tools are often only a part of the entire decision making system. It is desirable that they integrate smoothly, both with the database and the final decision making procedure.

1.2 PROBLEM STATEMENT

Microarray experiments normally produce a large amount of datasets with expression values for thousands of genes. Microarray data was found to be more useful when compared to other datasets. The sheer volume (in bytes), specialized formats (such as MIAME), and curation efforts associated with the datasets require specialized databases to store the data. Gene Expression Omnibus versatility is largely attributed that tabular data are treated as “blobs”, that is, compressed text tab-delimited tables that may contain any number of rows or columns. Data in selected columns are extracted to a secondary database and used in subsequent indexing and query applications. The number of data of this type required from the database can be huge, sometimes even exceeding 100 gigabytes, requiring several hours, if not days, to complete analysis, even using a supercomputer. For this reason, system requires a high-speed processing with accuracy and effective data retrieval. Berners-Lee et al (2009) performed reasoning on the metadata provided by the Semantic Web in order to extract knowledge from the Web more effectively.

The DNA micro array technology considerably expedite the procedure of discovering the utility of genes, the amount of data generated by this technology
also pretenses a challenge for the biologists to carry out the analysis. To design an intelligent system with existing search or retrieval technique to keep up with dynamic environments is a challenging one. For this, the solution is to integrate information retrieval algorithms, a user preference algorithm, an existing search or retrieval engine and an agent. A system that retrieves sequence data from multiple sources depending on their speed at the time of issuing the query thereby avoiding fetching data from the sites that are facing technical difficulty or delay. This research reviews how information retrieval techniques can be applied to microarray gene data. To improve the information retrieval performance of a search or retrieval engine based on specified measurable attributes like microarray gene data. Initially, the dimensionality diminution process is carried out in order to shrink the microarray gene data without losing information.

After the dimensionality diminution process, a robust classification method is indispensable to retrieve the gene information from the microarray experimental data. Jian et al (2006) compares different classification techniques and brings about the effectiveness among them. An effective classification technique is obtained to get back the gene information from the microarray experimental data. Information retrieval and data mining are powerful tools to extract information from the databases and/or information repositories. The integrative cluster analysis of both clinical and gene expression data has shown by Valarmathie et al (2012) to be an effective alternative to overcome the abovementioned problems. The focus is on how to improve the searching, retrieval and the clustering performance in genomic data from clustering techniques. These data can be used to the actual clinical application of gene expression data analysis and guide development of drugs and other research.
1.3 OBJECTIVES

Performance Issues in Genetic Information Retrieval is to provide optimal efficient and effective information about microarray gene data. The growing use of data or information brings with it all the performance issues that arise with high data volume and real-time delivery requirements. Noureddine & Mohammad (2004) say the size, heterogeneity and distributed architecture of the Web present tremendous performance problems for search engines, directories and other information resources must be addressed. It is an important and strategic operations management problem in information retrieval, and usually involves multiple and conflicting objectives such as time, accuracy level, error rate, specificity, sensitivity, etc.

- To create a frame work for dimensionality diminution process with the microarray data without losing information by techniques and utilized for information retrieval.
- To classify and identify a set of models that exhibit and distinguish data classes or concepts and predict the class of objects that has unknown class label.
- The dimensionality reduced gene data is applied with the SVM ensemble for the classification process. Here more than a single class is identified with the aid of this SVM ensemble.
- To measure the costs (efforts) and benefits of dimensionality diminution methods by gene data sets and obtain an efficient method.
- To improve the clustering performance in genomic data from commonly used clustering techniques.
1.4 MOTIVATION

Performance evaluation in Information Retrieval (IR) has a long tradition that has greatly boosted the development of information access systems (IASs). Evaluation of information retrieval (IR) system has a long tradition. There is a need for new evaluation methodologies, which are able to deal with various issues says Leung et al (2009). In research related to Genetic information retrieval system, evaluation is often related to the question of whether the eventual day-to-day users of a system will be more successful at (or simply prefer) using one access system over another. To attain the goals of the research, crucial issues in several IR evaluation methodologies were identified Mark et al (2008). Presently, the enhanced DNA microarray technology has resulted in expression levels of thousands of genes being recorded over just a few tens of different samples. In microarray data analysis, the process of information retrieval system includes diagnosis of disease, categorizing disease and getting information which is useful to give possible treatments. In order to address these issues, there is a need for the development of appropriate methodologies. There was a need for human expertise to make use of computational tools and development of better and safer drugs for treatment of all diseases of mankind. This research could bring about the recent developments in bioinformatics and the need for the development of molecular modeling.

1.5 SIGNIFICANCE

One of the developments of computer science in the last years tends to bring computer scientists closer to researchers in neuroscience and the genetic study. The GA is a probabilistic algorithm simulating the mechanism of natural selection of living organisms and is often used to solve problems having expensive solutions. Its traditional application is the optimization of a fitness function on a given set. It is chosen for its strength and can be applied without any special
knowledge of the domain. This research attempts to enhance the performance of information retrieval by using an adaptive genetic algorithm which can improve the quality of query and obtain more developed queries that fit the searcher’s needs. Investigate and evaluate different fitness functions. Reduce the search space which leads to saving time and reduction the number of iterations needed to generate the most optimized query. Finally obtain the best techniques for information retrieval system.

1.6 REQUIREMENTS

Hardware Requirement

Processor 32-bit, dual-core, 1.5 GHz minimum per core
RAM 1 GB
Hard disk 2 GB

Software Requirement

Operating system Windows 2000/XP, MacOSX or Linux
Visual requirements Office 2007 or later
Browser Internet Explorer 8 or later, Mozilla Firefox
Database Entrez, PubMed, GenBank and GEO
1.7 CONTRIBUTION

Microarray technology will help researchers to learn more about many different diseases, including heart disease, mental illness and infectious diseases, to name only a few. In the past, scientists have classified different types of cancers based on the organs in which the tumors develop. With the help of microarray technology, however, they will be able to further classify these types of cancers based on the patterns of gene activity in the tumor cells. This will then be able to design treatment strategies targeted directly to each specific type of cancer. Additionally, by examining the differences in gene activity between untreated and treated tumor cells - for example those that are radiated or oxygen-starved - scientists will understand exactly how different therapies affect tumors and be able to develop more effective treatments.

1.8 METHODOLOGY

In the process of information retrieval in DNA microarray technology, gene classification is quite tough task, because of the characteristics of the data, which contain high dimensionality and small sample size. A combination of the tactics is repeatedly used in practice for classification with gene expression data. Such classification measures normally contain the following steps: gene selection/dimension reduction, in which a small amount of gene components are constructed from a huge number of genes; classification, in which the samples are categorized into groups by applying standard statistical models on the gene components. For the purpose of retrieving information from a microarray gene expression dimensionality reduction technique based on LPP, PCA and PLS is developed. Especially the classification places vital role in mining information in the microarray gene data. The dimensionality reduced gene data is given to effective classification by SVM and ANN methods. The dimensionality reduced gene data is applied with the SVM ensemble for the classification process for more
than a single class. A comparative study is made with the dimensionality reduction and classification techniques for efficient retrieval.

1.9 LIMITATION AND DELIMITATION

Cancer classification is one major application of microarray data analysis. Due to the ultra high dimensionality nature of microarray data, data dimension reduction has drawn special attention for such type of data analysis. The available data dimension reduction methods like PCA and PLS are unsupervised. The system with ANN is trained with few samples of microarray data. PCA transfers features into linear combinations of original attributes but it does not consider data in terms of missing values and inconsistent relationship. Limitation of techniques like classification trees based algorithms is that they cannot generally handle more genes.

It is evident that application of supervised dimension reduction techniques could have a promising perspective in microarray data analysis. In order to exploit the class label information, one novel LPP method is proposed. It yields a map which is simple, linear, and defined everywhere (and therefore on novel test data points). The algorithm can be easily kernelized yielding a natural non-linear extension. Performance improvement of this method over Principal Component Analysis is demonstrated through several experiments. Though our method is a linear algorithm, it is capable of discovering the non-linear structure of the data manifold. Therefore LPP is chosen for linear, non-linear and supervised dimension reduction techniques. SVM is chosen for classification of more than one class for any number of sample genes. The support vectors identified by the SVM effectively define the boundary of the training set of genes. The effectiveness is evaluated by the classification accuracy of a SVM classifier. The SVM ensemble is the process of making a strong decision about the class of an input test gene.
1.10 INFORMATION RETRIEVAL

Information Retrieval (IR) is the science that addresses the efficient and effective gathering, indexing, organization and retrieval from billion-document collection. A system that retrieves sequence data from multiple sources depending on their speed at the time of issuing the query thereby avoiding fetching data from the sites that are facing technical difficulty or delay. Traditional IR systems focus on the retrieval of information from a collection based on a query provided to the system by the user. After processing the query, the system returns a list of data to the user, usually consisting of the title and a short extract. An IR system is used to retrieve data applicable to a user’s needs based on queries posed to the system by the user. Figure 1.2 illustrates the operation of a generic IR system. As shown, a user poses a query to the system. This query is parsed by the system, and is used to select data from the document collection indexed by the system. Results are then returned to the user.

![Figure 1.2 Basic Components of an IR System](image)

In many systems, the user is able to refine a query based on the results returned by the system, and can then resubmit the modified query. This process repeats itself until the user is satisfied. As seen in Figure 1.2, a generic IR system consists of a number of components:
1. A data store often referred to as the document collection, where data searchable by the system reside.
2. A query formulation subsystem, which allows users to pose queries to the system.
3. An indexing component, responsible for accepting raw document or data and converting them into a form usable (i.e. searchable) by the system. Information can only be added to the document collection by going through this component.
4. A processing component which operates on query and data representations, deciding what to return in response to a query, and in what order.
5. An output component, which displays the results of the user’s query. This component is usually closely linked with the query formulation system so as to enable the user to further operate on the results of a query.

1.10.1 Operation of IR

The process of information retrieval can operate on one of two broad types of information sources: structured and unstructured. Structured sources are exemplified by records within a database and standardized forms, where certain information must appear in specific locations was said by (Nir 2002). Unstructured sources are common in most fields, with examples including books, articles, and most other informal means of human communication. Performing retrieval on structured data sources is usually a simple task. A search is executed for the contents of a combination of predefined fields, for example, retrieving records from a database based on the ranges of a certain field. The focus of much research in information retrieval is the retrieval of information from unstructured data sources. Information Retrieval techniques studies information extraction and text mining, applied to genomic literature databases.

1.10.2 Aspects of IR

Two aspects of an IR system can be measured: efficiency, and effectiveness. Efficiency can be measured in terms of the resources required by the
system, including the storage space required to store the information collection, and the computing resources needed to perform operations on the collection, such as the addition and removal of information, and performing queries. While it is sometimes possible to compute the time and space complexity of the various aspects of the system, more concrete efficiency measures are often required when implementing a concrete system, but these measurements are difficult to perform in a machine independent manner. Effectiveness attempts to measure, as the name implies, the effectiveness of an IR system at satisfying a set of queries. Given a sufficiently general information and query collection, the effectiveness should provide a domain neutral measure of the system’s ability to satisfy user queries. For searching proper information that we need, it is necessary to construct efficient information retrieval agent systems helping many web clients’ requests.

1.10.3 IR Paradigm

There are three important paradigms in the area of information retrieval (IR): Probabilistic IR, Knowledge-based IR, and, Artificial Intelligence based techniques like neural networks and symbolic learning. At a broad level, research in IR can be categorized into three categories: 1. Probabilistic IR, 2. Knowledge based IR, and 3. IR based on machine learning techniques.

1. Probabilistic IR: Probabilistic retrieval is based on estimating a probability of relevance of a document to the user for the given user query. Typically relevance feedback from a few data is used to establish the probability of relevance for other data in the collection.

2. Knowledge based IR: This approach focuses on modeling two areas. First, it tries to model the knowledge of an expert retriever in terms of the expert's domain knowledge, that is, his or her search strategies and feedback heuristics. Although knowledge based approaches might be effective in certain domains, it may not be applicable in all domains.
3. Learning systems based IR: This approach is based on algorithmic extraction of knowledge or identifying patterns in the data. There are three broad areas within this approach: Symbolic learning, neural networks, and Evolution based algorithms. In the symbolic learning approach knowledge discovery is done typically by inductive learning by creating a hierarchical arrangement of concepts and producing IF-THEN type production rules. Neural networks are connectionist learning algorithms that typically simulate the way human brain learns and remembers knowledge says Praveen et al (2009).

Evolutionary algorithms are based on the Darwinian principles of natural selection. These algorithms can be further divided into: GA’s, evolutionary strategies, and evolutionary programming. While evolutionary programming utilizes changes at the level of species, the evolutionary strategies exploit changes at individual behavioral level. An approach for redescribing data descriptions and subsequently adopt a similar approach to data classification and clustering. The research work fits well in this paradigm.

1.10.4 Challenge of IR

Ensuring the timely retrieval of relevant items while not retrieving those items that are non-relevant to the searcher’s information need is the major challenge. Information retrieval remains a key research area within information science points Kalero & Wilson (2003). Over the past years, the study of information retrieval has expanded beyond efficiency and search mechanisms within the systems themselves to include human factors and searcher thought processes during IR sessions to better understand the mental processes user employ when performing information seeking tasks.
1.11 MICROARRAY GENE DATA

A DNA microarray is a multiplex technology used in molecular biology and in medicine. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10^{-12} moles) of a specific DNA sequence, known as probes (or reporters). This can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation. Each data point produced by a DNA microarray hybridization experiment represents the ratio of expression levels of a particular gene under two different experimental conditions. In standard microarrays, the probes are attached via surface engineering to a solid surface by a covalent bond to a chemical matrix. DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), to genotype or resequence mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data.

Microarray data was found to be more useful when compared to other similar datasets. The sheer volume (in bytes), specialized formats (such as MIAME), and curation efforts associated with the datasets require specialized databases to store the data. The goal of the molecular biology is to understand the regulatory mechanism that governs protein synthesis and activity. All the cells in an organism carries equal no of genes yet their protein synthesis can be different due to regulation. Protein synthesis is regulated by control mechanisms at different stages 1) Transcription, 2) RNA splicing, 3) Translation, and 4) Post-translational modifications cites Adi et al (2006).
Microarray techniques provide a platform where one can measure the expression levels of thousands of genes in hundreds of different conditions whereas using traditional methods in molecular biology can only report the expression levels of single genes. Many analysis tools help biologists to discover co-expressed genes. Microarray can also be used to determine which genes are expressed in which tissues and at which times during embryonic development. In bioinformatics, this process could refer to finding motifs in sequences to predict patterns, to discover genetic information underlying a database, to summarize clustering rules for multiple DNA (or) Protein sequences, and so on.

1.11.1 Microarray Experiment

With the substantial growth of biological data, Knowledge Discovery in Database (KDD) will play a significant role in analyzing the data and in solving emerging problems. The result, from an experiment with \( n \) genes on a single chip, is a series of \( n \) expression-level ratios. The data from a series of \( m \) such experiments may be represented as a gene expression matrix, in which each of the \( n \) rows consists of an \( m \)-element expression vector for a single gene. Microarray provides a medium for matching known and unknown DNA samples based on base-pairing rules and automating the process of identifying the unknowns. One major challenge with the use of microarray technology is the analysis of massive amounts of gene-expression data for various applications.
DNA microarrays are glass microscope slides onto which genes are attached at fixed and ordered locations. Each gene sequence is identified by a location of a spot in the array. Using a Microarray printer, the DNA is spotted directly onto the slide. With microarrays, it is possible to examine a gene expression within a single sample or to compare gene expressions within two tissue samples, such as in tumor and non-tumor tissues. The labeled DNA of an active gene will bind to the gene fragment, and produce a brighter color. The intensity of the spots is then measured by a scanning microscope and the results are recorded on a graph.

This addresses the key aspects of the microarray gene-expression data analysis for the two most common objectives: class comparison and class prediction. Class comparison mainly aims to select which genes are differentially expressed across experimental conditions. Gene selection is separated into two steps: gene ranking and assigning a significance level. Class prediction uses expression profiling analysis to develop a prediction model for patient selection,
diagnostic prediction or prognostic classification. DNA microarrays can simultaneously measure the expression level of thousands of genes within a particular mRNA sample. Such high-throughput expression profiling can be used to compare the level of gene transcription in clinical conditions in order to: 1) identify diagnostic or prognostic biomarkers 2) classify diseases (eg, tumors with different prognosis that are indistinguishable by microscopic examination) 3) monitor the response to therapy and 4) understand the mechanisms involved in the genesis of disease processes.

1.11.2 Types of Microarrays

Microarrays can be broadly classified according to at least three criteria: 1) length of the probes 2) manufacturing method and 3) number of samples that can be simultaneously profiled on one array. In general, the term “probe” is used to describe the nucleotide sequence that is attached to the microarray surface. The word “target” in microarray experiments refers to what is hybridized to the probes. There are three major types of applications of DNA. The first involves finding differences in expression levels between predefined groups of samples. This is called a “class comparison” experiment.

A second application, “class prediction,” involves identifying the class membership of a sample based on its gene expression profile. This requires the construction of a classifier (a mathematical model) able to analyze the gene expression profile of a sample and predict its class membership. The classifier is constructed based on a representative set of samples with known class membership. The third type of application involves analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features. This application is known as “class discovery.”
11.3 Data Pre-processing

Once the microarrays have been hybridized, the resulting images are used to generate a dataset. This dataset needs to be “preprocessed” prior to the analysis and interpretation of the results. Preprocessing is a step that extracts or enhances meaningful data characteristics and prepares the dataset for the application of data analysis methods.

11.4 Class Discovery Studies

Class discovery involves analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features. The example involved measuring the expression profiles of a large number of patients with diseases with the goal of classifying them into subgroups of patients having similar expression profiles. The medical and biological interest of this effort is aimed at understanding the mechanisms of disease underlying the syndrome. Classification technique is necessary to get the gene information from the microarray experimental data. High-throughput techniques have become a primary approach to gathering biological data. However, the deluge of data contains an overwhelming amount of unknown information about the organism under study. In DNA microarray technology, gene classification is quite tough task, because of the characteristics of the data, which contain high dimensionality and small sample size. After the dimensionality diminution process, a robust classification method is indispensable to retrieve the gene information from the microarray experimental data. Jian et al (2006) compares different classification techniques and brings about the effectiveness among them.

Most data mining applications have been assessed in this research. Although the activities are still in their early stages and should continue to develop as the Web evolves. Testing for elevated expression of certain genes can assist in predicting cancer. The difficulty in microarray analysis, however, is the ultra—high dimensionality of gene expression data (microarray image). The high
dimensionality of microarrays makes processing them a very difficult task with a high time and space complexity. Therefore, to make processing microarrays feasible, it is important to reduce the dimensionality of the microarray before further processing.

1.11.5 Microarray Clustering

Clustering microarray features helps investigate their biological significance. Thus, a major use of microarray data is to classify genes with similar expression profiles into groups. A wide variety of clustering algorithms have been employed in this field for research. Clustering aims at dividing the data points (genes or samples) into groups (clusters) using measures of similarity, such as correlation or Euclidean distance. In the past decade there have been advance in technologies, the amount of biological data such as DNA sequences and microarray data have been increased tremendously. To obtain knowledge from the data, explore relationships between genes, understanding severe diseases and development of drugs for patterns from the databases of large size and high dimensionality. Information retrieval and data mining are powerful tools to extract information from the databases and/or information repositories.

1.11.6 Dimensionality Reduction

To improve the information retrieval performance of a search or retrieval engine based on specified, measurable attributes and relative to the increased cost of adding the agent. Initially, the dimensionality diminution process is carried out in order to shrink the microarray gene data without losing information. This operation is applied to microarray gene data and thus information about gene can be retrieved. Data type, quality and dimensionality are some factors which affect performance of data mining task. Dimension reduction is a necessary part of multivariate analysis of high throughput assay data such as gene expression data. Dimensionality reduction is an effective solution to the problem of curse of
dimensionality. Dimension reduction is performed for the given microarray gene data and then class prediction or classification is done.

1.12 APPLICATIONS OF MICROARRAY

Molecular diagnostics is a rapidly advancing field in which insights into disease mechanisms are being elucidated by use of new gene-based biomarkers. Until recently, diagnostic and prognostic assessment of diseased tissues and tumors relied heavily on indirect indicators that permitted only general classifications into broad histologist or morphologic subtypes and did not take into account the alterations in individual gene expression. Microarray analysis may provide invaluable information on disease pathology, progression, resistance to treatment, and response to cellular microenvironments and ultimately may lead to improved early diagnosis and innovative therapeutic approaches for cancer says Li et al (2007). Microarray methods were initially developed to study differential gene expression using complex populations of RNA. Refinements of these methods now permit the analysis of copy number imbalances and gene amplification of DNA and have recently been applied to the systematic analysis of expression at the protein level.

1.12.1 Experimental Design and Choice of Reference

Careful design at the outset is crucial to the success of microarray experiments. In cancer research, case–control, blocked, and random profile designs predominate. In a case–control study, two samples from a single individual, e.g., tumor tissue and healthy tissues are compared directly. Because patient variability and genetic heterogeneity are key issues in microarray data analysis, the case–control design is an excellent solution when feasible. Blocked designs are typically used to study the effect of a treatment or growth condition on a sample such as a cell line. They have been successfully used to examine cell lines grown under
different conditions (e.g., cultured in the presence or absence of an anticancer
drug) or different related cell lines (e.g., wild type vs. mutant, non transected cells
vs. transected cells).

1.12.2 Image Acquisition and Quantification

Microarray image processing uses differential excitation and emission
wavelengths of the two flours to obtain a scan of the array for each emission
wavelength, typically as two 16-bit grayscale TIFF images. These images are then
analyzed to identify the spots, calculate their associated signal intensities, and
assess local background noise. Most image acquisition software packages also
contain basic filtering tools to flag spots such as extremely low-intensity spots,
ghosts spots (where background is higher than spot intensity), or damaged spots
(e.g., dust artifacts).

1.12.3 Databases and Normalization

The quantity of data generated in a microarray experiment typically requires
a dedicated database system to store and organize the microarray data and images.
The first role of a local microarray database is the storage and annotation
(description of experimental parameters) of microarray experiments by the
investigator who designed and carried out the microarray experiments.

1.12.4 Statistical Analysis and Data Mining

Analysis of large gene expression data sets is a new area of data analysis
with its own unique challenges. Data mining methods typically fall into one of two
classes: supervised and unsupervised. In unsupervised analysis, the data are
organized without the benefit of external classification information. Hierarchical
clustering, K-means clustering, or self-organizing maps are examples of
unsupervised clustering approaches that have been widely used in microarray
analysis. Supervised analysis uses some external information, such as the disease
status of the samples studied. Supervised analysis involves choosing from the entire data set a training set and a testing set and also involves construction of classifiers, which assign predefined classes to expression profiles.

### 1.12.5 Cancer Classification

Cancer classification is one major application of microarray data analysis. Due to the ultra high dimensionality nature of microarray data, data dimension reduction has drawn special attention for such type of data analysis. The currently available data dimension reduction methods are either supervised, where data need to be labeled or computational complex. DNA microarrays have enabled biology researchers to conduct large-scale quantitative experiments cites (Ahmad 2009). The response patterns have helped illuminate mechanisms of disease and identify disease sub phenotypes, predict disease progression, assign function to previously unannotated genes, group genes into functional pathways, and predict activities of new compounds.

Directed at the genome sequence itself, microarrays have been used to identify novel genes, binding sites of transcription factors, changes in DNA copy number, and variations from a baseline sequence, such as in emerging strains of pathogens or complex mutations in disease-causing human genes. Although all of the cells in the human body contain identical genetic material, the same genes are not active in every cell. Studying which genes are active and which are inactive in different cell types helps scientists to understand both how these cells function normally and how they are affected when various genes do not perform properly.

### 1.12.6 Gene Discovery

DNA Microarray technology helps in the identification of new genes, know about their functioning and expression levels under different conditions.
1.12.7 Disease Diagnosis

DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancer have been classified on the basis of the organs in which the tumors develop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer.

1.12.8 Drug Discovery

Microarray technology has extensive application in Pharmacogenomics. Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. The researchers can use this information to synthesize drugs which combat with these proteins and reduce their effect.

1.12.9 Toxicological Research

Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants.
1.13 PERFORMANCE IN GENE RETRIEVAL

The process of genomic information retrieval has allowed significant advances in biology and led to advancements in critical areas. Genomic information retrieval techniques are based on sequence alignment algorithms, literature search, genomic sequence searches and database searches. Research is to improve the information retrieval performance of a search or retrieval engine based on specified, measurable attributes and relative to the increased cost of adding the agent. The direction is to design intelligent system with existing search or retrieval technique to keep up with dynamic environments. From the microarray gene data, the process of extracting the required knowledge remains an open challenge for the biologists to carry out the analysis. There was a need for human expertise to make use of computational tools in drug discovery and the significance of computer aided drug discovery that could help in the design and development of better and safer drugs for treatment of all diseases of mankind. Also, the molecular biologists face the challenges in determining the required knowledge from this kind of enormous amount of data. In this kind of knowledge seeking applications, information retrieval is one of the primary and most important technologies to extract the entailed knowledge from the huge amount of data.

Performance Issues in Genetic Information Retrieval is to provide optimal efficient and effective information about microarray gene data. The performance par at new computational methods and a workable process must be implemented for effective and timely analysis and management of the genetic kind of data. To achieve the high classification accuracy in microarray gene expression data set, various classification methods using dimension reductions have been extensively studied. The growing use of multimedia in modern documents brings with it all the performance issues that arise with high data volume and real-time delivery requirements. To deal with multi-objective like enable the diagnosis of diseases and for its evaluation a greater number of alternative solutions, two different
approaches are implemented in the solution procedure. An experimental study using actual data from cancer disease is carried out into two stages. While the dimensionality reduction of the given data deals on the performance of solution in the first stage, the solution procedure and simulated annealing are compared with other classification techniques.

The objectives were feature tests that were used during diagnosis evaluation to rate how closely a candidate texture matched visual characteristics of a target texture image. The objectives included the minimization of total microarray part movement, and the minimization of within-cell load variation dealt by Adi et al (2006). The nature to learn boolean queries in information retrieval systems is another important problem. Two objectives are used: precision (the ratio between the relevant documents retrieved in response to a query and the total number of documents retrieved) and recall (the ratio between the relevant documents retrieved and the total number of documents relevant to the query in the database). The purpose of information retrieval techniques is to retrieve all the relevant information (high recall), while including as little of the non-relevant information as possible (high precision).

Classification problems are a very common target for machine learning strategies. These types of problems involve determining to which of a set of classes an input belongs. Most microarray platforms produce highly reproducible within-platform measurements when operating within their range of sensitivity. Microarrays are able to simultaneously monitor the expression levels of thousands of genes. Such gene expression information can be used in medicine for comparing clinically relevant groups (eg, healthy vs diseased), uncovering new subclasses of diseases, and predicting clinically important outcomes, such as the response to therapy and survival. The recent focus on accurate and fast access to biological information was triggered by the availability of a large volume of unstructured biological data. In this kind of knowledge seeking applications, information retrieval is one of the primary and most important technologies to extract the
entailed knowledge from the huge amount of data. There are several types of models which differ in respect to performance and accuracy and how certain obstacles are overcome.

1.14 ORGANIZATION OF THE THESIS

Chapter one dealt with the introduction to the performance in genetic information retrieval. The various aspects of information retrieval and its paradigm are discussed. It also gave brief description about microarray gene data and its applications. Issues on performance of genetic information retrieval were prescribed. The problem statement, objectives and motivation of the research were presented.

Chapter two reviews some of the related works available in the literature. It discusses about the advancement of genome computing techniques and microarray genes analysis technique. A review of the literature exploring the role of primary support in delivering genetic services is performed.

Chapter three investigates the dimensionality reduction and information retrieval. The efficacy of various dimensionality reduction techniques is investigated. A gene classification based on blending Locality Preserving Projection (LPP) and Support Vector Machine (SVM) is presented which improves the performance of the retrieval.

Chapter four details the Multiclass classification with SVM, the kind of dataset used in genetic information and also defines the proposed multiclass SVM ensemble. This chapter reports on the results obtained from running the experiments. This also interprets and analyses the results, attempting to understand what they mean in the context of the approach.
The results from the experiments of dimensionality reduction techniques are evaluated in Chapter five, which is concerned with the findings and the comparison of techniques in classification and clustering in detail. This chapter also discusses the context of the field of information retrieval.

Chapter 6 concludes the thesis and discusses the future works. It also suggests some ideas for the extension of the research.

1.15 SUMMARY

This chapter took a broader look at data mining. Some aspects of the processes involved in data mining were discussed. The discussion involved the Information retrieval systems responsible for retrieval of large amounts of data in an efficient manner. Microarray gene data, a widely used entity and which plays a prominent role and is well defined. In bioinformatics, the process of gene information retrieval and its performance were also considered. Next chapter deals about the literature survey of the previous work carried out in the area of microarray gene data and its efficient retrieval.