CHAPTER 5

EFFICIENT GENE RETRIEVAL TECHNIQUE

5.1 INTRODUCTION

Information retrieval is one of the most important technologies at present. Information retrieval is the task, given a set of documents and a user query, of finding the relevant documents. We can always get much information in the Internet or distributed computing systems using various information retrieval models. For searching proper information that we need, it is necessary to construct efficient information retrieval agent systems helping many web clients’ requests tells (Lee 2007). Information retrieval applications require speed, consistency, accuracy and ease of use in retrieving relevant texts to satisfy user queries.

Accuracy: The sheer volumes of information stored in electronic media magnify deficiencies in recall the percentage of relevant information retrieved. Giving the user reasonable-sized response with high precision can mean missing hundreds of relevant data.

Speed: As the quantity of data that must be searched increases, the speed of searching can become a reserve bottleneck. In practical terms, the need for fast search means that more computational-intensive processing such as NLP techniques must either apply very selectively or run as "batch" indexing tool prior to retrieval.

Consistency: Many information retrieval environments require indexing of the text by the groups of indexers or by the authors. This leads to a decrease in accuracy from the inevitable inconsistencies, which automatic processing could help to avoid.
Ease of use: The growth of personal computer has made obsolete the traditional model of information retrieval with a trained human intermediary, giving systems responsiveness a high priority.

5.1.1 Classification of Retrieval Techniques

Retrieval techniques can be further classified in terms of the characteristics of the retrieved set of data and the representation that are used. Figure 5.21 gives a diagrammatic view of the classification. The first distinction that make among retrieval techniques is whether the set of retrieved data contains only information whose representation as ream exact match with the query or a partial match with the query. Exact match techniques are currently at use in most conventional IR systems. Queries are usually formulated using Boolean expressions and the search pattern within the query has to exactly match the data representation inside of the document to be retrieved.

![Classification of Retrieval Techniques](image)

Figure 5.21 Classifications of Retrieval Techniques

Within there are many different variants. Individual techniques search single nodes without considering the collection as a whole. For the feature-based techniques, documents are represented by sets of features or index terms. The index can be either defined manually or be computed automatically. The most prominent representative of this category is the vector space model, which is based
on a formal model of retrieval and indexing. In the vector space model, each
document is represented by an index vector containing a set of weighted terms. For
each query, documents are ranked in decreasing order of similarity to the query
deals Ivica et el (2010). The probabilistic approach is similar to the vector space
model: the basic goal is to retrieve data in order of their probability of relevance to
the query.

In the probabilistic model, (as contrasted with a Boolean retrieval system) a
query typed by a user to retrieve information is taken as an unstructured form.
These terms are then matched to the data in the database. Some suggest a fuzzy set
approach for feature-based formal IR systems. Contrary to formal feature-based IR
systems, there have also been a number of ad hoc similarity measures. Most of the
feature-based approaches have the problem that small differences in weights can
lead to significant differences in results. For structure-based techniques, data are
represented in a more complicated structure than just a set of index terms as used
for the feature-based techniques. It is theoretically possible to represent the
contents collections in formal logic. Systems in this category could, e.g., use rules
to describe how relevant fragments are related to the query. Instead of using logic
to represent the contents, the contents also can be described as a graph where the
edges and nodes of the graph stand for ideas and relationships contained in the
document.

With network-based methods, the set of all data and their relationships are
used to find the most relevant information with respect to a query. The most
prominent method is clustering, where the most similar data are clustered together
and all data are grouped into a cluster hierarchy until a ranked list of lowest-level
clusters is produced. If the data are represented as a network of nodes, the user can
also browse through the network with system assistance. Through dialog with the
user, the system can use the network to build a model of the user. Based on the
user model, a model of the user's information needs can be constructed. Spreading
activation is similar to browsing in that, from the start node, other nodes connected
to that node are activated. Activated nodes then propagate or spread themselves through the network.

For a partial match, the set of retrieved data will include also those that are an exact match with the query. The next level of the classification distinguished between retrieval techniques that compare the query with individual representative and techniques that use a representation of data that emphasize connections to other data in a network. Much of the research in Information Retrieval has concerned improvements to similarity computations, statistics gathering, and term extraction, with the goal of improving effectiveness. This chapter provides information about efficient techniques on classification and clustering.

5.2 EFFICIENT CLASSIFICATION TECHNIQUE

Data and Knowledge Mining is learning from data. This learning from data comes in two flavors: supervised learning and unsupervised learning. In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set. The large amount of data that is usually present in Data Mining tasks allows splitting the data file in three groups: training cases, validation cases and test cases. Training cases are used to build a model and estimate the necessary parameters. The validation data helps to see whether the model obtained with one chosen sample may be generalizable to other data. Test data can be used to assess the various methods and to pick the one that does the best job on the long run.
Most data mining methods are supervised methods, however, meaning that (1) there is a particular pre specified target variable, and (2) the algorithm is given many examples where the value of the target variable is provided, so that the algorithm may learn which values of the target variable are associated with which values of the predictor variables. Supervised classification (here onwards to be referred as classification ) algorithms have been widely applied to speech, vision, robotics, diseases, and artificial intelligence applications etc where real time response with complex real world data is necessity. There have been wide ranges of machine learning and statistical methods for solving classification problems. Different parametric and non-parametric classification algorithms have been studied. Some of the algorithms are well suited for linearly separable problems. Non-linear separable problems have been solved by neural networks in Alireza & Bita (2009) and support vector machines etc.

However, in many cases it is desired to find a simple classifier with simple architecture and also to ensure the performance of the classifiers. To achieve optimal accuracy estimation, it is required to minimize the bias associated with the random sampling of the training and testing data samples. The classifier is trained and tested 10-times. It must be noted here that even we use the sample for testing of the algorithm and the evaluation based on thousands of testing set.

![Figure 5.22 Model for Performance Evaluation](image)
The robustness and domain-independent capabilities of algorithms attracts researchers to evolve a set of classification rules. There is a need for new evaluation methodologies, which are able to deal with various issues. From a user's point of view, evaluation of performance of IR systems should be in terms of how well they are supported with respect to whole retrieval sessions. Typically, users initiate a session with a specific goal (e.g. acquiring crucial information for making a decision). Thus, the overall quality of a system should be evaluated with respect to the user's goal. (Samuel 2006) say most supervised data mining methods apply the following methodology for building and evaluating a model.

- First, the algorithm is provided with a training set of data, which includes the pre classified values of the target variable in addition to the predictor variables. A provisional data mining model is then constructed using the training samples provided in the training data set.
- The next step in supervised data mining methodology is to examine how the provisional data mining model performs on a test set of data. In the test set, a holdout data set, the values of the target variable are hidden temporarily from the provisional model, which then performs classification according to the patterns and structure it learned from the training set. The efficacy of the classifications is then evaluated by comparing them against the true values of the target variable.
- The provisional data mining model is then adjusted to minimize the error rate on the test set.
- The adjusted data mining model is then applied to a validation data set, another hold out data set, where the values of the target variable are again hidden temporarily from the model. Estimates of model performance for future, unseen data can then be computed by observing various evaluative measures applied to the validation test.

These points are portrayed in the figure 5.23. The recent focus is on accurate and fast access to biological information was triggered by the availability of a large
volume of unstructured biological data. From the microarray gene data, the process of extracting the required knowledge remains an open challenge. In this, 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.

![Figure 5.23 Methodologies for Supervised Modeling](image)

To achieve the high classification accuracy in microarray gene expression data set, various classification methods using dimension reductions have been extensively discussed in the previous chapters. The growing use of medical data in modern world brings with it all the performance issues that arise with high data volume and real-time delivery requirements.

**5.2.1 Efficient Gene Classification**

In order to retrieve the required information, gene classification is vital; however, the task is complex because of the data characteristics, high
dimensionality and smaller sample size. As a first process in the proposed gene retrieval, the high dimensionality of the microarray gene data is reduced using dimensionality reduction technique. For the purpose of retrieving information from a microarray gene expression, an effective retrieval technique based on LPP and PCA is implemented. Since LPP and PCA were used in the previous chapters for efficient data reduction technique. Initially, the dimensionality diminution process is carried out in order to shrink the microarray data without losing information with the aid of LPP and PCA techniques and utilized for information retrieval. The technique like LPP and PCA is chosen for the dimensionality reduction for efficient retrieval of microarray gene data.

5.2.1.1 Dimensionality reduction by LPP

Firstly, dimensionality reduction, of the gene data is performed using LPP. The LPP is chosen for the dimensionality reduction because of its ability of preserving locality of neighborhood relationship. The LPP procedure for dimensionality reduction constitutes of three steps, namely, (1) generation of Distance matrix (2) determining adjacency matrix and (3) Calculating dimensionality reduced matrix. The steps for dimensionality reduction are as same as discussed in chapter 3. The \( \hat{Y} \) obtained from the above equation is the dimensionality reduced gene data with size \( n_s \times n_s \).

5.2.1.2 Dimensionality diminution through principal component analysis (PCA)

Let \( M_{xy} ; 0 \leq x \leq n, 0 \leq y \leq n \) here \( n \) indicates the number of genes in the sample in which it has been taken from and \( n_s \) indicates the number of samples in which has been taken for the process. These microarray gene data is of higher dimension and hence it must be reduced in order to do that the PCA mechanism is utilized. In this dimensionality diminution, data which is in high dimension is converted to low dimension. The dimensionality diminution is the process of
reducing the large dimensional data, in order to make comfort to the classification process. PCA is one of the dimensionality reduction techniques that are utilized. PCA is a great tool for the data analysis process and also it can be utilized for the dimensionality reduction without any loss of information. The steps for dimensionality reduction of microarray gene data using PCA is same as depicted in chapter 4.

5.2.1.3 Results and discussion

For evaluating the technique, the microarray gene samples of human acute leukemias were utilized. The data has been taken by two different classes of microarray gene data, namely, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Thus obtained microarray gene expression data is of dimension, \( n_c = 2, \ n_g = 7192 \) and \( n_s = 38 \).

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Number of Samples</th>
<th>Number of Genes</th>
<th>Dimensionality Reduced Data with the aid of LPP and PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>38</td>
<td>7192</td>
<td>27 X 38</td>
</tr>
<tr>
<td>AML</td>
<td>38</td>
<td>7192</td>
<td>11 X 38</td>
</tr>
</tbody>
</table>

A sample of microarray gene dataset of two classes that has been used for processing is given in the Table A 2.11 The process of the SVM is to identify the class of the gene. From the training gene data, the SVM learns well about the class under which the given gene dataset is present. Once the SVM is trained well, it attains the ability to classify any gene dataset in the similar fashion. The dimension-reduced matrix is given to the trained SVM and so the class of the given microarray gene data is obtained in an effective manner. Some six samples
for each cancer class are given in the Table A 2.12 for processing respectively. The efficacy of the techniques has been determined by comparing it with classification technique using Support Vector Machine (SVM).

Table 5.6 Performance comparison for the classified data with the aid of dimensionality reduced data (LPP)

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Specificity (In %)</th>
<th>Sensitivity (In %)</th>
<th>Accuracy (In %)</th>
<th>Error Rate (In %)</th>
<th>Scalability (In Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>92.62</td>
<td>86.26</td>
<td>97.30</td>
<td>2.703</td>
<td>80K</td>
</tr>
<tr>
<td>AML</td>
<td>72.50</td>
<td>97.63</td>
<td>92.22</td>
<td>2.778</td>
<td>120K</td>
</tr>
</tbody>
</table>

Table 5.7 Performance comparison for the classified data with the aid of dimensionality reduced data (PCA)

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Specificity (In %)</th>
<th>Sensitivity (In %)</th>
<th>Accuracy (In %)</th>
<th>Error Rate (In %)</th>
<th>Scalability (In Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>92.59</td>
<td>54.35</td>
<td>82.12</td>
<td>13.596</td>
<td>70K</td>
</tr>
<tr>
<td>AML</td>
<td>54.55</td>
<td>93.10</td>
<td>84.59</td>
<td>14.741</td>
<td>96K</td>
</tr>
</tbody>
</table>

The comparison of the technique with the LPP and PCA-based gene classification techniques with respect to the performance metrics, specificity, sensitivity, accuracy, error rate and scalability are given in the Table 5.6 and Table 5.7. Here SVMs is utilized for classifying ALL and AML genes and for this process, initially these diseases equivalent gene data was trained with the SVMs separately. In the testing process any of these genes given and they have identified in which the class they belongs to. The work has been obtained better results and they are compared with another dimensionality reduction algorithm i.e. PCA. As
the LPP have good positive features in their task of dimensionality reduction respectively. When technique like LPP is chosen for the dimensionality reduction of microarray gene data there is effective retrieval. The comparative results have shown that the technique LPP possesses better accuracy and lesser error rate than PCA techniques. Hence, this means of gene classification have paved the way for effective information retrieval in the microarray gene expression data. The results discussed here have shown the performance of the work.

5.3 EFFICIENT CLUSTERING TECHNIQUE

With the rapid increase in the availability of data for exploration and analysis it is important to develop techniques that efficiently perform data clustering and data classification. In the first case there is a need to get insight into the data and find out things about it in as objective a way as possible. This type of approach is usually classed as exploratory data analysis and includes clustering. The second concerns learning or identifying the extent to which the data conforms to known or hypothesized models. Data mining methods may be categorized as either supervised or unsupervised. In unsupervised methods, no target variable is identified as such. Instead, the data mining algorithm searches for patterns and structure among all the variables say Daxin et al (2004). The most common unsupervised data mining method is clustering. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given. For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases.

Clustering or cluster analysis involves the identification of subsets of the data that are similar. The subset usually intuitively corresponds to points that are more similar to each other than they are to points from another cluster. Points in
the same cluster have the same label. Clustering is carried out in an unsupervised way by trying to find subsets of points that are similar without having a predefined notion of the cluster. Clustering task usually involves the following steps:

1. Representation of the data;
2. Deciding on a similarity measure or distance metric that is most appropriate for the task in the domain;
3. Performing the clustering;
4. Cluster description;
5. Evaluation.

Data representation is the task of deciding the number of features available, their nature and scale, the size of the dataset and the number of clusters or classes. Feature selection is the process of selecting a set of features that are the most effective subset to use with the clustering algorithm. Feature extraction or feature combination is the use of transformations of the feature set to produce a set of features that are able to be effectively used by the clustering algorithm. In many problems of cluster analysis, there is little prior information available about the data, and as few assumptions about the data as possible can be made explains Chen et al (2009). The notion of clustering is relatively flexible as the aim is to identify and reveal clusters or groups in an exploratory data analysis sense. It is simply an object that summarizes and represents the objects in the cluster. It should be close to every object in the cluster in some average sense. The similarity of the objects to the objects is measured by a matching function or similarity function. Similarity is usually measured by some dissimilarity measure such as a metric defined on the dataset. However there are other ways of approaching the definition of similarity. This would then imply that similarity is a transitive notion which is stronger than what we usually expect in the notion of similarity. Most clustering algorithms use a number of empirically determined parameters such as:
• The number of clusters
• A minimum and maximum size of each cluster
• A threshold value on the matching function, below which an object will not be included in a cluster
• A control on overlap between clusters
• An objective function which is optimized

5.3.1 Clustering Microarray Gene Data

Microarray technique can be used to measure the expression levels of tens of thousands of genes in different conditions. These conditions may be time series during a biological process. With the substantial growth of biological data, it greatly increased the challenges of understanding the results of huge data. To meet the challenges, data mining techniques specially referred to clustering algorithms were used to reveal the structure of the data and identify the interesting patterns. Clustering is an unsupervised learning process which partitions the given data set into groups where patterns within the groups are similar to one another and dissimilar to the patterns in different groups. 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. Microarray can also be used to determine which genes are expressed in which tissues and at which times during embryonic development.

Two statistical operations commonly applied to microarray data are classification and clustering. Clustering is a form of learning by observation and don’t rely on predefined classes whereas classification is a learning by example which means they rely on class labeled training examples given by Anirban et al (2006). In bioinformatics, one important clustering task is to identify groups of co-
expressed genes which recognize coherent expression patterns. Another aspect of clustering is finding gene networks and gene interactions.

5.3.2 Clustering Techniques

In clustering, data or observations are usually represented as feature vectors and the individual scalar components of a feature vector \( x = (x_1, \ldots, x_n) \) are called features or attributes. The dimensionality of the data or the dataset is \( n \). In selecting or developing clustering techniques the method should exhibit some theoretical soundness. This may be assessed by certain criteria of adequacy. Some criteria are:

- The method produces a clustering which is unlikely to be altered drastically when additional objects are incorporated - stability under growth
- The method is stable in the sense that small perturbations in the description of the objects lead to small changes in the clustering
- The method is independent of the initial ordering of the objects.

Clustering techniques which support microarray expressed data in order to find the gene functions and regulations are

- Hierarchical clustering
- K-means clustering
- K-medoid
- Two way clustering
- Genetic K-means
- Fuzzy clustering

5.3.2.1 Hierarchical clustering

Hierarchical Clustering is a novel clustering technique that takes set of gene expression data and produce a hierarchy of cluster as a result. The result of hierarchical clustering is usually depicted by a binary tree (or) dendrogram. An example of dendrogram is portrayed in figure2.7. Hierarchical clustering starts by
calculating the distance matrix for all patterns in data set. The two closest patterns are merged and the distance matrix is calculated again but using the new cluster instead of the two merged patterns. This process is repeated until the complete dendrogram is built.

5.3.2.2 Partitional algorithms

A partitional clustering algorithm produces a single partition of the data with no hierarchical structure. This means that the algorithm usually requires the number of clusters to be specified. They usually optimize a criterion function defined on the dataset. Most algorithms have traditionally been run multiple times with different starting states and the best configuration produced is the one used as the clustering. The most frequently used objective function is the squared error criterion:

\[ f_k(x) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||a^{(j)}_i - x_j||^2 \]  

(5.1)

Where \( k \) is the number of clusters, \( n_j \) is the number of records in the cluster \( j, j = 1, \ldots, k \), \( a^{(j)}_i \) is the \( i \)-th element of the cluster \( j, i = 1, \ldots, n_j \) and \( x_j \) is the centroid of the \( j \)th cluster. This has been found to work well with isolated and compact clusters. The k-means algorithm is the most commonly used algorithm using this criterion. The k-means algorithm randomly chooses \( k \) cluster centers and iteratively reassigns data points to clusters based on the similarity between the pattern and the cluster centers until there is no further reassignment or the squared error no longer decreases significantly.

K-Means Clustering

K-means clustering is a typical partition based clustering method. It is an iterative process of assigning cluster memberships and re-estimating cluster parameters. Given a data set of \( n \) patterns and \( k \), the no. of clusters to form, the partition algorithm organizes the patterns into \( k \) partitions (\( k<n \)) where each
partition represents a cluster. The clusters are formed to optimize an objective portioning criterion, called a similarity function such as distance, so that objects within a cluster are similar whereas the objects of different clusters are dissimilar says (Vasanth 2009).

Algorithm
Input: The number of clusters k and a database containing n objects.
Output: A set of k clusters that minimizes the squared-error criterion.

1. Arbitrarily choose k objects as the initial cluster centers;
2. Repeat
3. Reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5. until no change;

The k-means always converge to a local minimum which means for different initial centroid, it leads to different results. For better clustering we have to find initial centroid that is consistent with the distribution of patterns. The advantages are it can be used for a wide verity of datasets; the results should be quite efficient, even though multiple runs are performed. The limitations are it has trouble for clustering data that contains outliers; empty clusters can be obtained if no points are allocated to the cluster during the assignment statement. Users must specify the no. of clusters in advance, when it converges to local optima.
5.3.2.3 K-Medoid

K-medoids also a partition based algorithm such as k-means. In contrast to the k-means algorithm k medoids chooses objects as centre (medoids). Medoid is the most centrally located object in a cluster.

Algorithm
1. Randomly select k objects as initial medoids.
2. Associate each data object to the cluster with the most similar medoids.
3. Randomly select non-medoid object O’.
4. Compute total cost, S of swapping initial medoid object to O’.
5. If S is less than zero, then swap initial medoid with the new one.
6. Repeat steps 2-5 until there is no change in the medoid.

The basic strategy of k-medoids algorithm is to find k clusters in n objects by first arbitrarily finding a representative object for each cluster.

5.3.2.4 Coupled two way clustering

In one way gene clustering the genes and samples are clustered independently, whereas in coupled two way clustering both the features are used and the data points are clustered. The main idea of coupled two way clustering is to identify subsets of genes and samples such that when one of these is used cluster the other. Coupled two way clustering begins with clustering the samples and the genes of the full data set and identify all stable clusters of either samples (or) genes. These stable clusters are candidates for the next iteration.

5.3.2.5 Genetic K-Means algorithm

Genetic algorithms are efficient randomized search algorithm in order to obtain optimum solutions using the principles of evolution and natural genetics,
having a large amount of implicit parallelism. As the amount of laboratory data in molecular biology grows exponentially, generic algorithms in Daxin et al (2004) are used to prove this growing amount of biological data. Genetic K-means clustering is a hybrid approach of combining genetic and K-means algorithm.

5.3.2.6 Fuzzy clustering

Fuzzy clustering provides a systematic, unbiased way to change precise values into several descriptors of cluster memberships. Here every object is assigned to each cluster with a membership degree. The membership degree should be in the range between 0 and 1 instead of using crisp assignments of the data.

Hard membership vector (0,1,0,1)  Fuzzy membership vector (0.2,0.8,0.3,0.4)

![Figure 5.24 Hard Clustering](image1)

![Figure 5.25 Fuzzy clustering](image2)

As single gene might involve multiple genetic functions, fuzzy clustering is more effective than hard clustering for analyzing gene expression profiles cites Valarmathie et al (2009). It can be used to find natural vague boundaries in data. Fuzzy logic can efficiently manage the uncertainty and the vagueness of the expression levels.
Fuzzy C-Means Clustering

Fuzzy c-means algorithm is the most widely used fuzzy clustering method. Let $X$ be the dataset with the samples $x_1, x_2, \ldots, x_n$, $V = \{v_1, v_2, \ldots, v_c\}$ is the center vector and an objective function is defined with the membership degree between each data $x_j$ and cluster center $v_i$

$$J_m(X, U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^m d^2(x_j, v_i)$$

(5.2)

Here, $\mu_{ij}$ is the membership degree of $x_j$ and the $i$th cluster, an element of the membership matrix $U = [\mu_{ij}]$, $d^2$ is the square of the Euclidean distance, and $m$ is the fuzziness parameter, which means the degree of fuzziness of the each data membership degree.

Algorithm
1. Set $c$, the number of clusters and $m$, fuzziness parameter
2. Initialize $\mu_{ij}$ as follows
   $$\sum_{j=1}^{c} \mu_{ij} = 1, \quad 1 \leq j \leq n$$
3. Compute $v_i$, each centre of all clusters.
4. Compute the membership matrix $U$
5. Repeat step 3 and 4 until step 4 is satisfied.

The advantages of Fuzzy clustering are it produces a clustering that provides an indication of the degree to which any point to any cluster. The limitations of fuzzy clustering are the users must specify the no of clusters in advance, and it is dependent on the choice of membership co-efficient.
5.3.3 Efficient Gene Clustering Technique

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. The integrative cluster analysis of both clinical and gene expression data has shown to be an effective alternative to overcome the abovementioned problems.

High-throughput techniques have become a primary approach to gathering biological data. These data can be used to the actual clinical application of gene expression data analysis and guide development of drugs and other research. However, the deluge of data contains an overwhelming amount of unknown information about the organism under study. Therefore, clustering is a common first step in the exploratory analysis of high-throughput biological data. Although a number of clustering methods have been proposed, they incur problems such as Quality and Efficiency points. In the aspect of efficiency, most clustering algorithms aim to produce the best clustering result based on the input parameters.

In the process of mining gene expressions under multi-conditions microarray experiments, gene clustering is relatively a tough task. A combination of the approaches is utilized continually in practice for clustering with microarray data. Such classification measures usually have the following stages: i) gene selection/dimension reduction, where a small amount of gene components are created from a vast number of genes; ii) clustering, where the samples are clustered into groups by applying standard models on the gene components. Normally, microarray experiments create a large number of datasets with expression values for thousands of genes but still not more than a few dozens of
samples, thus very accurate arrangement of tissue samples in such high dimensional problems is a complicated task says Jian et al (2006). Therefore, a robust clustering method is indispensable to retrieve the gene information from the microarray experimental data. Here is a work to group the microarray gene data with the aid of FCM. Initially, the dimensionality reduced data is applied with the FCM for clustering.

5.3.3.1 Clustering based on LPP and FCM

The challenging issue in microarray technique is to analyze and interpret the large volume of data. This can be achieved by clustering techniques in data mining. In hard clustering like hierarchical and k-means clustering techniques, data is divided into distinct clusters, where each data element belongs to exactly one cluster so that the outcome of the clustering may not be correct in many times. The problems addressed in hard clustering could be solved in fuzzy clustering technique. Among fuzzy based clustering, fuzzy c means (FCM) is the most suitable for microarray gene expression data. The problem associated with fuzzy c-means is the number of clusters to be generated for the given dataset needs to be specified in prior. This can be solved by combining this method with a popular probability related Expectation Maximization (EM) algorithm which provides the statistical frame work to model the cluster structure of gene expression data suggests (Vasanth 2009).

The main objective of this is hybrid fuzzy c-means method is to determine the precise number of clusters and interpret the same efficiently. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In some situations, the object may belong to more than one cluster, and associated with each element is a set membership levels. In this method, the fuzzy c-means combined with the EM (Expectation Maximization) algorithm which provides the statistical frame work to model the cluster structure of gene expression data. Dimensionality reduction, one of the two stages of the gene
clustering technique is performed using LPP. From the gene data of different classes \( Y_{ijk} \), a concatenated matrix is obtained. In the concatenated matrix, the gene data of all the classes are combined and it is given as a single matrix. The matrix \( Y_{con} \) is given as follows

\[
Y_{con,il} = \sum_{i=0}^{n_c-1} Y_{ijl}
\]

(5.3)

where,

\[
Y_{ijl} = \begin{cases} 
Y_{ijl} & \text{if } l \in (i, n_s(i+1) - 1) \\
0 & \text{otherwise}
\end{cases}
\]

(5.4)

The concatenated matrix \( Y_{con} \) of dimension \( n_g \times n_s' \); \( n_s' = n_s n_c \), \( n_s' < cn_g \), which is highly dimensional and so the dimensionality of the matrix is reduced using LPP. The LPP is a linear dimensionality reduction algorithm that shares most of the properties of data representation of nonlinear techniques, namely, locally linear Embedding or Laplacian Eigenmaps. The LPP procedure for dimensionality reduction constitutes of three steps, namely, (1) generation of Distance matrix (2) determining adjacency matrix and (3) Calculating dimensionality reduced matrix. The \( \hat{Y} \) obtained from the above equation is the dimensionality reduced gene data with size \( n_s' \times n_s' \). The \( \hat{Y} \) is utilized to cluster the input microarray gene data using FCM. Any clustering method aims to produce a \( K \times n \) partition matrix \( U(\hat{Y}) \) of the given data set \( \hat{Y} \), consisting of \( n \) objects, \( \hat{Y} = \{ y_1, y_2, \ldots, y_n \} \), where \( K \) is the number of clusters. The partition matrix may be represented as \( U = [uk_{kj}] \), \( k = 1, \ldots, K \) and \( j = 1, \ldots, n \), where \( uk_{kj} \) is the membership of pattern \( y_j \) to the \( k \)th cluster. Greater value of \( uk_{kj} \) implies that the probability of belongingness of point \( y_j \) to the \( k \)th cluster is more. The well known fuzzy C-means uses cluster centers encoded in chromosomes have been used in designing as shown in figure 5.26. FCM clustering method optimizes \( J_m \) and \( XB \) cluster validity indices respectively.
Here data set $\hat{Y}$ is partitioned into $K$ clusters with centers $Z = \{z1, z2, \ldots, zK\}$ and $m$ is the fuzzy exponent. $\|.\|$ represents the Euclidean norm and $D(z_k, y_j)$ denotes the distance of point $y_j$ from the center of the $k$th cluster. In this, the Euclidean norm is taken as a measure of the distance between two points. The $Jm$ index represents the global cluster variance, whereas $XB$ index is expressed as a function of the ratio of total intra-cluster variance to the minimum inter-cluster separation. Hence both the measures are to be minimized.

The technique for microarray gene clustering has been implemented in the working platform of MATLAB (version 7.11). For evaluating the technique, the microarray gene samples of human acute leukemia and colon cancer data are utilized. Thus LPP method is applied to identify informative genes and reduce gene dimensionality for clustering samples to detect their phenotypes.
Table 5.8 Microarray gene data dimension utilized for the evaluation process

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Number of Samples</th>
<th>Number of Genes</th>
<th>Dimensionality Reduced Data with the aid of LPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>38</td>
<td>7129</td>
<td>38 X 40</td>
</tr>
<tr>
<td>AML</td>
<td>34</td>
<td>7129</td>
<td>34 X 40</td>
</tr>
<tr>
<td>COLON</td>
<td>62</td>
<td>3000</td>
<td>62 X 42</td>
</tr>
</tbody>
</table>

Table 5.9 Sample of the microarray gene data to test the proposed technique

<table>
<thead>
<tr>
<th>Class</th>
<th>ALL</th>
<th>AML</th>
<th>COLON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Gene</td>
<td>ALL 16125 TA-Norel</td>
<td>ALL 23368 TA-Norel</td>
<td>AML SH 5</td>
</tr>
<tr>
<td>AFFX-CreX-5_at (endogenous control)</td>
<td>-172A</td>
<td>-93A</td>
<td>-271A</td>
</tr>
<tr>
<td>AFFX-CreX-3_at (endogenous control)</td>
<td>52A</td>
<td>10A</td>
<td>-12A</td>
</tr>
<tr>
<td>AFFX-BioB-5_st (endogenous control)</td>
<td>-134A</td>
<td>159A</td>
<td>-104A</td>
</tr>
</tbody>
</table>

While testing, when a gene dataset is given, the technique has to identify its belonging cluster. Clustering for microarray gene expression data whose amount is large can be fully calculated by determining the boundary of the clusters. Clustering algorithms, such as K-means and Fuzzy C-means approaches are applied both to group genes, to partition samples in the early stage and have proven to be useful. The performance of each clustering algorithm may vary greatly with different data sets. Complete-link clustering method uses the smallest similarity within a cluster as the cluster similarity, and every data object within the
cluster is related to every other with at least the similarity of the cluster. In order to test the performance of the data, N artificial m-dimensional feature vectors from a multivariate normal distribution having different parameters and densities were generated. Situations of large variability of cluster shapes, densities, and number of data points in each cluster were simulated.

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Accuracy FCM</th>
<th>Accuracy k-means</th>
<th>Correlation FCM</th>
<th>Correlation k-means</th>
<th>Distance FCM</th>
<th>Distance k-means</th>
<th>Error Rate FCM</th>
<th>Error Rate k-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>83.9</td>
<td>71.1</td>
<td>0.342</td>
<td>0.235</td>
<td>0.00379</td>
<td>0.00476</td>
<td>0.21</td>
<td>0.38</td>
</tr>
<tr>
<td>AML</td>
<td>80.6</td>
<td>70.6</td>
<td>0.024</td>
<td>0.013</td>
<td>0.00364</td>
<td>0.00472</td>
<td>0.30</td>
<td>0.31</td>
</tr>
<tr>
<td>COLON</td>
<td>79.0</td>
<td>67.6</td>
<td>0.119</td>
<td>0.062</td>
<td>0.02029</td>
<td>0.02653</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>

From the Table 5.10, it can be seen that the technique FCM has provided more accuracy, correlation and less distance and error rate rather than the k-means gene clustering techniques. More accuracy and less error rate leads to effective clustering of the given microarray gene data to the actual class of the gene.

### 5.4 SUMMARY

This chapter presents the efficient classification and clustering technique for retrieving genetic information. As the first process in genetic retrieval, dimensionality is reduced. Efficient dimensionality reduction technique is experimentally proved. Secondly, clustering techniques are described and efficient clustering technique for microarray gene data is also obtained from our
experimental calculation. This leads to the next chapter which concludes and explains about the future enhancement of the research.