BIOMEDICAL DATA CLASSIFICATION USING HIERARCHICAL CLUSTERING

By

Hu Yang

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
UNIVERSITY OF MANITOBA
WINNIPEG, MANITOBA
AUGUST 2005

0-494-08998-9

Your file Votre référence
ISBN:
Our file Notre reference
ISBN:

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a MSc thesis entitled:

Biomedical Data Classification using Hierarchical Clustering

submitted by: *Hu Yang*

in partial fulfillment of the requirements for the degree of: *MSc*

Dr. N. Pizzi, Advisor

Dr. Desmond Walton
Computer Science

Dr.W. Kinsner
Electrical and Computer Engineering

Date of Oral Examination: *August 22, 2005*

The student has satisfactorily completed and passed the MSc Oral Examination.

Dr. N. Pizzi, Advisor

Dr. Peter King
Chair of MSc Oral

Dr. Desmond Walton
Computer Science

Dr .W. Kinsner
Electrical and Computer Engineering

(The signature of the Chair does not necessarily signify that the Chair has read the complete thesis.)

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES
*****
COPYRIGHT PERMISSION PAGE

Biomedical Data Classification Using Hierarchical Clustering

BY

Hu Yang

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree

of

MASTER OF SCIENCE

Table of Contents

Abstract

Biomedical data such as those acquired from magnetic resonance (MR) spectrometers often have the characteristics of high dimensionality and small sample size. These two characteristics make the classification of such data difficult. Hierarchical clustering produces robust clustering results, especially when working on small size high-dimensional datasets. The goal of this research is to investigate the effectiveness of hierarchical clustering in the classification of high-dimensional biomedical data. In order to achieve the above goal, a new classification method and a new dimension reduction method, which use hierarchical clustering, were developed in this research. These methods were tested using MR spectra and the results were benchmarked against linear discriminant analysis.

Chapter 1

Introduction

1.1 Magnetic Resonance Spectroscopy

Magnetic resonance spectroscopy (MRS) [9] is concerned with the behaviour of atomic nuclei and their interaction with electromagnetic radiation. Certain nuclei, for example those of ^{1}H (hydrogen), ^{13}C (carbon) and ^{31}P (phosphorus) resonate when exposed to electromagnetic radiation at a particular frequency. This frequency is dependent on the type of nucleus and also on the intensity of the surrounding magnetic field. The use of MRS in medicine allows us to see what is going on inside the body without carrying out invasive surgery or inserting optical instruments. MRS is not unique in this; there are other techniques for imaging the body such as X-rays and ultrasound. However, unlike other methods, MRS makes it possible not only to visualize anatomical structure, but also to investigate physiological function. The extra dimensions of information offered by MRS and also the fact that the technique has no known harmful effects makes it a unique and powerful imaging technique for clinical medicine.

MRS [9], first developed in the 1940's, is based on the fact that a nucleus will resonate at a slightly different frequency depending on its molecular environment. Exploiting the interaction between an external homogeneous magnetic field and a nucleus that possesses spin, MRS is a spectroscopic modality that is reliable and versatile. Combined with robust classification strategies, MRS is specifically useful in the classification of biomedical spectra such as those acquired from the human brain.

Magnetic resonance (MR) spectra may contain unwanted artifacts as the tissue being examined will not be homogeneous. These artifacts together with a low signal-to-noise ratio may make identification and measurement of the metabolites that are present in the tissue difficult. The fact that MR spectra carry information on a large number of metabolites presents the non-trivial problem of how to extract and classify this information. Besides high dimensionality, MR spectra often suffer from small sample size. These two characteristics present serious challenges for the classification and interpretation of such spectra. In my thesis, I address this problem using hierarchical clustering methods.

1.2 Pattern Classification Approach

Pattern classification [14] is a discipline devoted to extracting context from data by identifying meaningful patterns. Pattern classification may be broken down into supervised and unsupervised methods. Unsupervised methods attempt to group similar patterns together typically using some type of distance measure. Supervised methods, which require that each pattern has an associated class (or group) label, attempt to predict the class to which a pattern belongs. A supervised pattern classification method, linear discriminant analysis (LDA) and an unsupervised pattern method, hierarchical clustering, were used in this thesis.

LDA [26], a computationally simple analysis technique, which assumes data normality and equal covariance matrices for the different classes, is often applied to MR spectra classification problems.

Hierarchical clustering is one of the most widely used clustering methods for separating individual items into groups. Ward's method is believed to produce the best clustering results among the various hierarchical clustering methods.

1.3 Scopira

Scopira [5], developed for analyzing high-dimensional biomedical data, is an open framework running in Linux for numerical algorithm module development, execution and interaction. Scopira allows new modules, data types and functions to integrate smoothly with existing systems. Developers can quickly build their algorithm modules with Scopira's powerful template library and simplified programming interface. Some Scopira modules were developed in this research.

1.4 Thesis Structure

My thesis analyzes the suitability of hierarchical clustering in biomedical spectra classification problems. I first present a supervised Ward's method and then introduce a new dimension reduction method using hierarchical clustering. These approaches were tested using MR spectra with LDA as a benchmark.

Following the introduction section, the subsequent sections of the thesis are organized as follows. Section 2 describes the concept and use of MR spectroscopy. Section 3 introduces the Scopira software developed by the National Research Council's Institute for Biodiagnostics. Section 4 briefly describes pattern classification, hierarchical clustering, LDA and dimensionality reduction. Section 5 presents the main research approaches in this thesis. Section 6 summarizes system development and the experiment results. Finally, Section 7 offers concluding remarks.

Chapter 2

Magnetic Resonance Spectroscopy

An MR signal is produced by inducing nuclei of interest to resonate by exposing them to a pulse of radiation at their resonance frequency, and then allowing the nuclei to relax when they will release radiation at this same frequency. Figure 2.1 shows the process of the target nuclei transfer from a lower energy state, where radiation can be absorbed, to a higher energy state, where radiation may be released.

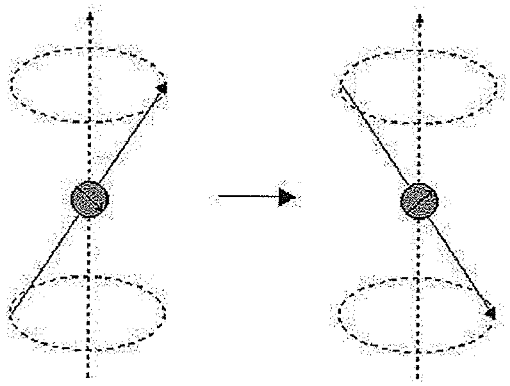


Figure 2.1: Nucleus transfer from lower to higher energy state

MRS is based on the theory that atomic nuclei are surrounded by a cloud of electrons, which slightly shield the nucleus from any external magnetic field. Figure 2.2 illustrates this effect, where B is the externally applied magnetic field and e is the electron cloud that produces a magnetic field and generates the small field shift. Because the strength of the resulting signal will depend on the number of nuclei present, it can be used to give a measure of the proportion of nuclei in a sample. As the structure of the electron cloud is specific to an individual molecule or compound, the magnitude of this screening effect is also a characteristic of the chemical environment of individual nuclei. The resonant frequency is proportional to the magnetic field that it experiences and will be determined not only by the externally applied magnetic field, but also by the small field shift generated by the electron cloud. This shift in frequency is called the chemical shift. It should be noted that chemical shift is a very small effect, usually expressed in parts per million of the main frequency.

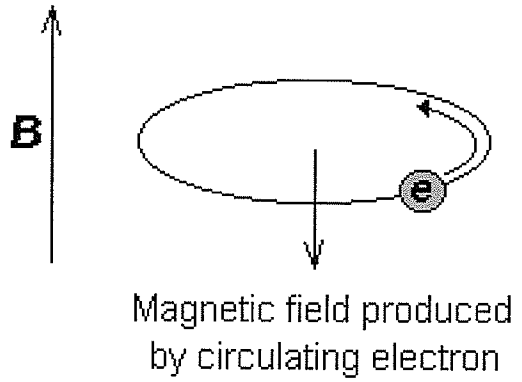


Figure 2.2: Shielding effect of electron's magnetic field against the external field, B.

MRS provides information on the types of metabolites present in tissue. It also provides a means of measuring these metabolites. In addition to giving information about the concentrations of specific metabolites, MRS provides information about the intracellular environment of the metabolites. MRS is widely employed as a research tool and a noninvasively diagnostic method. As MRS noninvasively monitors disease biochemistry, it can provide important new information for the clinician.

2.1 Clinical Uses of MR Spectroscopy

The potential of MRS for biology was understood initially during the time it was developed in the 1940's, but experiments were limited in scope by the relatively poor quality of the instrumentation that was then available. With the development of high-field superconducting magnets in the late 1960's together with the emergence of Fourier-based enhancements, it became possible to use MRS to study proteins and other biological molecules. This led to the realisation that MRS might have extensive applications in the study of the metabolism of living systems [11]. Figure 2.3 gives an example of a ^{1}H (hydrogen) MR spectrum acquired at 37°C on a 437 MHz MR spectrometer, which can be used for the detection and diagnosis of pathological tissue, e.g. brain tumors.

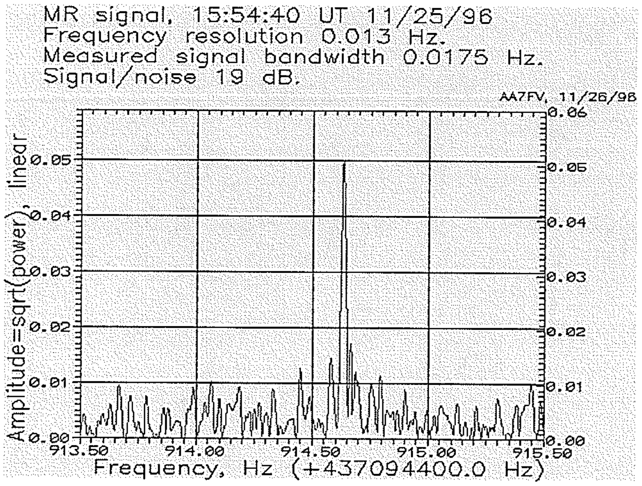


Figure 2.3: Example of a ^{1}H MR spectrum

One of the main areas in which MRS shows great potential as a clinical tool is in the diagnosis and treatment of cancer. Both ^{31}P and ^{1}H spectra show different metabolite patterns according to the type of tumours. It has been shown that MRS can be successfully used to discriminate between different types of human brain tumours and between normal tissue and tumours with 99% success compared with 77% pre-operative diagnosis for the same patients that was based on all the available clinical information including CT, MRI and angiography MRI [16]. MRS is also useful for the evaluation of metabolic myopathies. ^{31}P MRS is already used at a number of medical centres to determine the presence of metabolic myopathies from elevated inorganic phosphate peaks in resting muscle [33]. Figure 2.4, an MR image of a two-dimensional slice of a human brain, shows regions of normal (N) tissue and tumors (T). Figure 2.5(i) shows typical MR spectra of the tumours and Figure 2.5(ii) shows typical MR spectra of normal tissue.

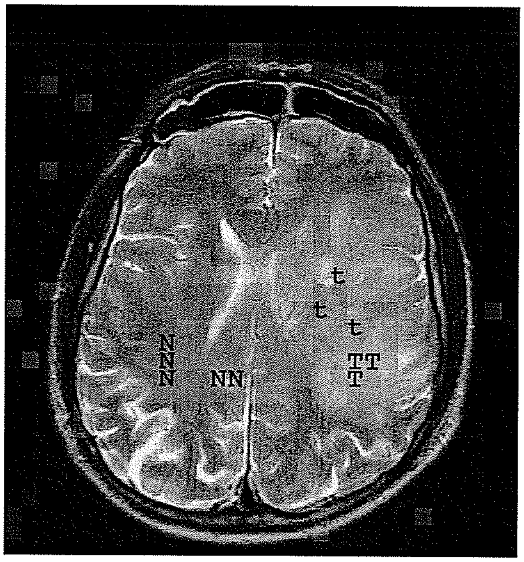


Figure 2.4: MR image of a human brain depicting normal tissue (N) and tumors (T).

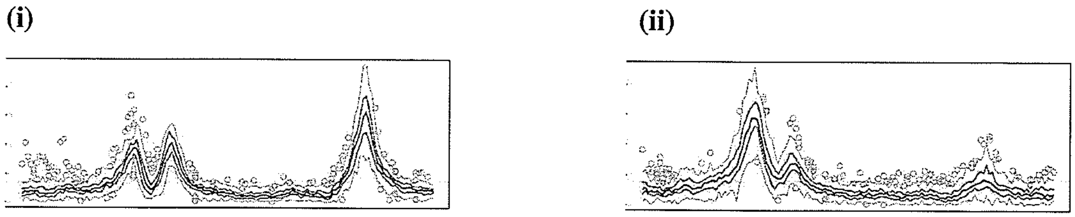


Figure 2.5: Typical human brain MR spectra of normal tissue (i) and tumors (ii) from Figure 2.4.

MRS has also been shown to be useful in the diagnosis and grading of prostate cancer. Investigators studying responses to therapy in animals have found that the ^{31}P spectrum changes in response to therapy often before there is a noticeable decrease in tumour size [21].

Another example of a potential application of MRS is in the treatment of epilepsy. MRS is already used in some medical centres to identify the focus of seizures before brain surgery. Currently EEG and other scanning methods do not provide accurate localization information in the majority of cases, but it has been shown that ^{31}P and ^{1}H MR spectra provide additional information that may avoid the use of invasive depth electrodes [23].

2.2 Problems with MR Spectral Analysis

One of the great advantages of MRS for medical applications is that it allows us to obtain information about the metabolic composition of living tissues *in situ*. On the other hand, the fact that MRS signals are obtained *in situ* presents considerable difficulties, both with acquiring the signals and extracting the context. It is impossible to control all the conditions of an investigation, which may cause the signal to contain unwanted artifacts. For example, artifacts are introduced by the movement of the patient, which effectively changes the sample that contributes to the MR signal. Another potential problem is that, while it may be possible to focus on a specific region, it is often not possible to focus on a specific tissue. Because the size of the smallest region that can at present be examined effectively by MRS is approximately 2 cm^3, it is likely that signal acquired from any region will include other signals in addition to those from the required tissue. [3].

Another problem concerns the low sensitivity of the MR signal [31]. The sensitivity, which can be expressed in terms of the signal-to-noise ratio of the spectrum, is dependent on several factors. These include the strength of the applied field, the design and performance of the MR instruments and the time taken to accumulate the data. One of the main factors that accounts for the low sensitivity of MR is that the interaction between the nuclei and the magnetic field is weak, that is, the amount of energy absorbed is low. This means that the amount of energy released is also low leading to a weak signal. Because of the limitations imposed on acquiring a signal from a living subject, the signal-to-noise ratio is generally lower for data acquired *in vivo*. The low signal-to-noise ratio cannot be improved using averaging techniques because of the short experimental times required for patient comfort.

Baseline distortion is another problem which may affect quantification of MRS data. One factor which can alter the shape of the baseline is the presence of metabolites with large peaks that have broad humps that spread the sides of these peaks. This distorts the signal by moving the contributions of other metabolites away from the baseline and towards the sides of the bumps. This problem particularly affects kidney, liver and tumour ^{31}P spectra, where a broad hump of signals from immobile phosphates underlies the spectrum [2].

Another factor which will affect spectral analysis methods is that while in principle, MRS spectra should have Lorentzian [17] peaks, the peaks observed from spectra obtained *in vivo* are often not of this ideal shape. This may be due to magnetic field inhomogeneity, magnetic susceptibility and other problems.

Chapter 3

Scopira

Scopira [5], developed for analyzing high-dimensional biomedical data, is an open framework for algorithm module development, execution and interaction. Used for biomedical data analysis, Scopira permits new modules, data types and functions to integrate smoothly with existing systems. Developers can quickly build their algorithm modules with Scopira's powerful template library and simplified programming interface. A Scopira map may contain input, output and algorithm modules. Figure 3.1 shows a typical Scopira map that contains 12 modules.

Scopira possesses a hierarchical, objected-oriented, data type tree, where each node represents one data type. Any two data types are considered to be compatible when these two data types are identical or when one is an ancestor of the other. A new data type can be added to the system by registering this data type in the Scopira data tree. After being assigned a base data type, Scopira can then operate on all descendants of this base data type automatically. Relying on the Scopira data tree to exchange data, developers can focus on designing the algorithm kernel of a new module. As an algorithm organizer, Scopira allows experienced data analysts to put multiple algorithm modules and their connections together to form an algorithm module network. Some Scopira modules were created and used in this research.

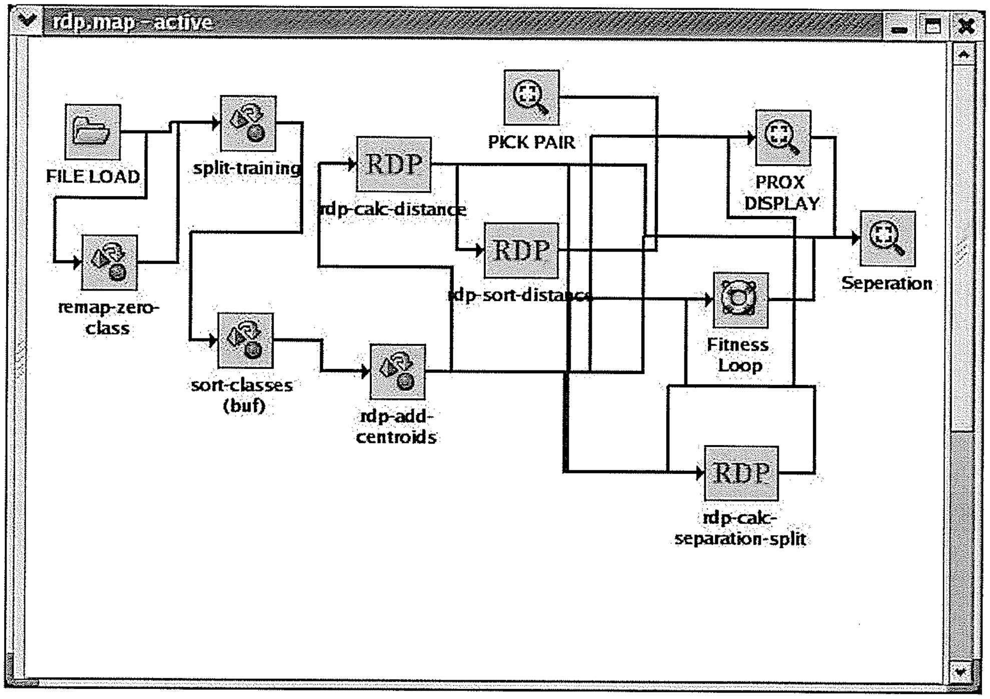


Figure 3.1: A typical Scopira map

3.1 Core Design

The Scopira architecture can be divided into three large software components: the engine core, the user interface (front ends), and the back end computation kits (See Figure 3.2).

The engine core is the central hub of Scopira. It is responsible for loading, maintaining and executing different modules within maps. The engine has three run-time selectable event schedulers used to determine module execution (possibly in parallel). The single-thread scheduler runs events sequentially using only one system thread per process. The multi-thread scheduler attempts to maximize a multi-processor machine by paralleling module execution. The network-aware scheduler manages a cluster of multiple machines, connected via a network, each with any number of processors. This scheduler may partition a map over these network nodes transparently, without requiring any special programming by the module developer.

The engine core manages module kits. A kit may contain any number of module kernels, micro functions, data types and graphical proponents. Kits are implemented as shared code libraries, dynamically loaded at run-time and selected by the user.

Finally, the front ends interact with the engine through its exposed, object-oriented interface. The interactive visual map editor and scripting systems both use this interface to manipulate and execute maps. Custom front ends may be built in a straight forward manner with no need to rebuild any part of Scopira.

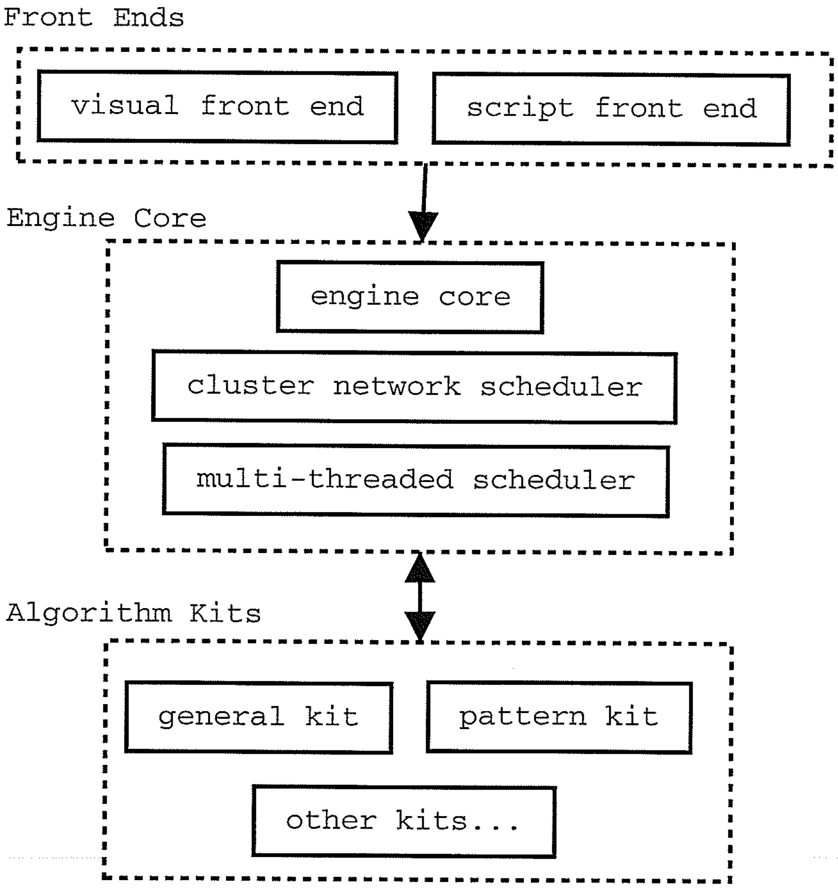


Figure 3.2: Scopira architecture layout

3.2 Portable

Scopira follows standard software engineering methods to maximize code portability. Platform dependent code like the thread and network communication systems are encapsulated within objects. Any changes required to these systems for new platforms or compilers need only be made to these objects.

The external routines used by the Scopira engine core and standard kits are those provided by the Standard C++ library and the system level C libraries. Because of this, it is quite straightforward to port the engine core and kits to other platforms.

The visual front end depends on the GTK+ [15] graphical user interface library. The script front end uses GUILE/Scheme [6]. Both GTK+ and GUILE are portable to all UNIX platforms. To maintain clean, portable code throughout the various development phases of the project, Scopira is routinely compiled and tested under various compilers. Currently, this list includes GNU C++ and Intel C++.

3.3 Generic Programming and Parallelism

Scopira follows the C++ use of generic templates to achieve performance gains. In-lining and specialized instantiations of constructs give the exact objects without the need for indirection, common base classes, or using the same simple data types. For instance, with generic classes, developers are not forced to work with double types when they want the smaller float types.

Scopira supports parallel execution at the inter- and intra-module levels. At the inter-module level, Scopira transparently schedules and executes modules simultaneously. Scopira does this by selecting a combination of modules to execute from the current run queue that would maximize the current state of free processors. All this is done transparently to the module developer.

At the intra-module level, Scopira provides an MPI-like interface that allows modules to request and use multiple processors within the context of their execution. This allows modules to be parallelized without having to be broken up into smaller units. Through both levels, Scopira constantly maintains and monitors the amount of processing resources, constantly attempting to maximize computational performance.

Chapter 4

Pattern Classification

Pattern classification [14] is a discipline devoted to extracting context from data by identifying meaningful patterns. A pattern (or sample) can be represented by an ordered set of n variables (or features) denoted by a vector $x = [x_1, x_2, ..., x_n] \in X \subseteq R^n$, where X is the input feature (pattern) space. Each pattern x belongs to one (and only one) of k classes, denoted as $y \in Y \subseteq \{1, 2, ..., k\}$, where Y is the output class space. Pattern classification may be regarded as a (mapping) function $f : X \rightarrow Y$, which maps (predicts) an output class label, y, for a sample, x.

With supervised pattern classification, we have a sample set V, which consists of N pairs of samples and class labels (x, y):

$$V = \{(x^1, y^1), \ldots, (x^N, y^N)\} \tag{4.0.1}$$

The sample set is separated into a training set and a validation set. We first build a classification function based on the information obtained from the training set and then apply this function to a validation set, which is used to validate its accuracy.

In unsupervised classification, class labels are not used during the training phase; rather we separate the input pattern space X into k groups (or clusters) using only the information contained in the input patterns. Only after the groups have been identified are the class labels subsequently used to assess the performance of the unsupervised method.

4.1 Cluster analysis

Cluster analysis [1], a pattern analysis method used for sorting a set of samples (patterns) into groups or clusters, may be used for classification. Each cluster may have a particular property. The degree of similarity is high between samples in the same cluster and low between samples from different clusters. The main objective of cluster analysis is to organize data into meaningful structures. To accomplish this objective, hierarchical clustering, a specific type of cluster analysis, classify samples into groups of nested classes.

Cluster analysis is a tool of discovery. It may reveal associations and structure in data which, though not previously evident, nevertheless are sensible and useful once found. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as a taxonomy for related animals, insects or plants; or suggest statistical models with which to describe populations; or indicate rules for assigning new cases to classes for identification and diagnostic purposes; or provide measures of definition, size and change in what previously were only broad concepts; or find exemplars to represent classes.

4.1.1 General Procedure

The two key steps within cluster analysis are the measurement of distances between objects and the grouping of the objects based upon the resultant distances (linkages). The distances provide for a measure of similarity between objects and may be measured in a variety of ways, such as Euclidean and Manhattan distances. The criteria used to then link (group) the features may also be undertaken in a variety of manners. Linkages are based upon how the association between groups is measured. For example, simple linkage or nearest neighbor distance measures the distance to the nearest object in a group. While furthest neighbor linkage or complete linkage measures the distance between furthest objects. Both linkages are based upon single data values within groups, whereas the average between-group linkage is based upon the distance from all objects in a group.

4.1.2 Various Cluster Analysis Techniques

Cluster analysis techniques may be hierarchical, i.e, the resultant classification has an increasing number of nested classes, resembling a phylogenetic classification. Others, such as k-means clustering [14] are non-hierarchical, where one must specify the number of clusters (k) into which the data are to be grouped. At the end of the analysis, the data will be split between k clusters. In this clustering procedure, only the final cluster membership for each case is presented.

Clustering techniques can also be grouped as divisive or agglomerative. A divisive method begins with all cases in one cluster. This cluster is gradually broken down into smaller and smaller clusters. Agglomerative techniques start with (usually) single member clusters. These are gradually fused until one large cluster is formed.

4.1.3 Hierarchical Clustering

In the hierarchical clustering [14] procedure, a series of partitions take place, which run from N clusters each containing a single pattern to a single cluster containing all N patterns.

Given a data set of N samples to be analyzed, hierarchical clustering first constructs a $N \times N$ similarity matrix. For example, we may construct such a similarity matrix based on the Euclidean distance [1] between these N samples:

$$\begin{pmatrix} d_{11} & d_{21} & \dots d_{N1} \\ d_{12} & d_{22} & \dots d_{N2} \\ \dots & \dots & \dots\dots\dots \\ d_{1N} & d_{2N} & \dots d_{NN} \end{pmatrix}$$

where d_{ij} is the Euclidean distance between samples x^i and x^j.

$$d_{ij} = \sqrt{(x_1^i - x_1^j)^2 + \cdots + (x_n^i - x_n^j)^2} \tag{4.1.1}$$

The subsequent hierarchical clustering steps are:

1. Assign N samples to N clusters. Each cluster contains one and only one sample.

2. Find the most similar pair of clusters and merge them into a single cluster. Here, two clusters are reduced and one new cluster is created.

   For example, given a similarity matrix based on Euclidean distance, if we have $d_{ij} \leq d_{pq}$, $for\ all\ p \neq q\ and\ p, q \leq N$, then cluster C^i and cluster C^j are the most similar pair of clusters. We may then delete C^i and C^j and add a new cluster C^{n+1}.

$$C^{n+1} = \frac{C^i + C^j}{2} \tag{4.1.2}$$

3. Compute similarities between the new cluster and each of the old clusters, and update the similarity matrix.

4. Repeat steps 2 and 3 until all samples are clustered into a single cluster of size N.

4.1.4 Ward's method

Ward's method [32] is a hierarchical method that enables clustering by assessing group variances. The group with the smallest increase in variance with the iterative inclusion of a sample will receive the sample. Ward's is a popular default linkage that produces compact groups of well distributed size. Ward's method is one of the best solutions in hierarchical clustering. In this research, our new classification methods will be created based on Ward's method. The key to Ward's method is the way it finds the most similar pair of clusters. In step 2, Ward's method calculates an error sum of squares (ESS) [13] between each pair of clusters and merges the pair of clusters that give the minimum ESS.

A similarity matrix based on ESS is:

$$\begin{pmatrix} ess_{11} & ess_{21} & \ldots ess_{N1} \\ ess_{12} & ess_{22} & \ldots ess_{N2} \\ \ldots & \ldots & \ldots\ldots\ldots \\ ess_{1N} & ess_{2N} & \ldots ess_{NN} \end{pmatrix}$$

where ess_{ij} is the error sum of squares between samples x^i and x^j:

$$ess_{ij} = (x_1^i - \frac{x_1^i + x_1^j}{2})^2 + \cdots + (x_n^i - \frac{x_n^i + x_n^j}{2})^2 + (x_1^j - \frac{x_1^i + x_1^j}{2})^2 + \cdots + (x_n^j - \frac{x_n^i + x_n^j}{2})^2 \tag{4.1.3}$$

4.1.5 Distance Measures

Hierarchical clustering uses the dissimilarities or distances between objects when forming the clusters. Three main distance measures [14] are typically used: Euclidean distance, Squared Euclidean distance and Manhattan distance (See Figure 4.1).

Euclidean distance: It is the geometric distance in the multidimensional space. Euclidean distance is computed as:

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \tag{4.1.4}$$

Squared Euclidean distance: By squaring the standard Euclidean distance, progressively greater weight is placed on objects that are further apart. This distance is computed as:

$$d_{xy} = (x_1 - y_1)^2 + \cdots + (x_n - y_n)^2 \tag{4.1.5}$$

Manhattan distance: This distance is simply the average difference across dimensions. In this measure, the effect of single large difference (outliers) is dampened. Manhattan distance is computed as:

$$d_{xy} = |x_1 - y_1| + \cdots + |x_n - y_n| \tag{4.1.6}$$

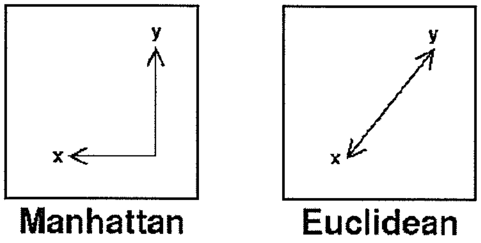


Figure 4.1: Typical distance measures

4.2 Bayesian Pattern Classification

The major role of discriminant analysis is to define rules for classifying samples into one of several classes. If the samples are represented by vectors in n-dimensional space, each object can be thought of as a point in this n-dimensional space.

In supervised classification, the formulation of a classification rule corresponds to an explicit or implicit construction of a boundary surface between the k classes in the training set so that the classes become as well separated as possible. Figure 4.2 illustrates two subspaces that may serve as class boundaries for a 3-class 2-dimensional dataset; clearly, the subspace on the right produces a more discriminatory boundary than the one on the left. We assume that each sample in the training set can be classified into one class with no measure of doubt. But many application problems cannot satisfy this assumption. So most statistical discriminant methods, such as Bayesian [29] methods, use probability theory to estimate the possibility of a sample belonging to a certain class. The Bayesian decision rule is:

$$d_i(x) = P(x|y_i)P(y_i), \;\; i = 1, \ldots, k \tag{4.2.1}$$

where $P(y_i)$ is the probability that sample, x, belongs to class y_i. $P(x|y_i)$ is the conditional density function defining the probability that x belongs to y_i, given that y_i is the case. If $d_i(x) > d_j(x)$ for all $j \neq i$, then x is assigned the class label, y_i.

The most important part in this decision function is to estimate the densities $P(x|y_i)$ from the training data. But this estimation is difficult when the dimension of the number of features is large.

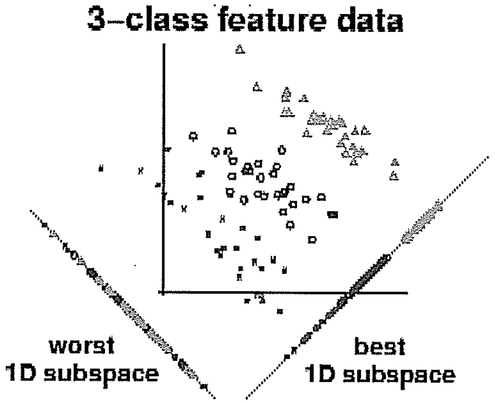


Figure 4.2: Three class two-dimensional dataset with two subspaces as possible discrimination boundaries.

4.3 Linear Discriminant Analysis

Some assumptions about the nature of the conditional probabilities can successfully simplify the estimation of $P(x|y_i)$. In one-dimensional data analysis, one often assumes a univariate normal distribution. The normal distribution's probability function is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.3.1}$$

where μ is the mean value or expected value and σ is the variance value or standard deviation.

Similarly, for n-dimensional data, it is assumed that the densities $P(x|y_i)$ are multivariate normal distributions. The multivariate normal density functions are of the form:

$$P(x|y_i) = \frac{\exp[-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i)]}{(2\pi)^{N/2}|C_i|^{1/2}}. \tag{4.3.2}$$

where μ_i is the mean vector or expected value vector and C_i is the n-dimensional covariance matrix. Figure 4.3 shows 2-dimensional (i) and 3-dimensional (ii) plots of the normal distribution function.

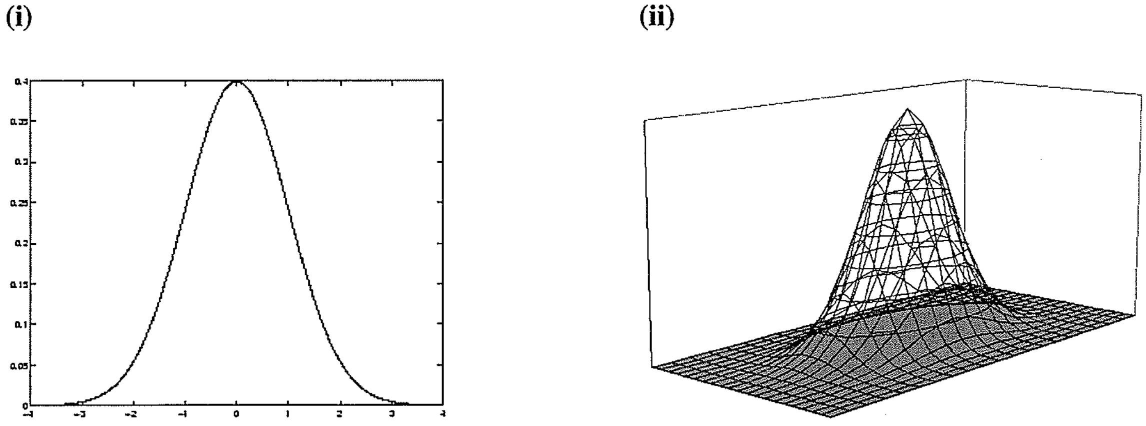


Figure 4.3: Two-dimensional (i) and three-dimensional (ii) plots of the normal distribution function.

Because of the exponential nature of the multivariate normal density function $P(x|y_i)$, it is more convenient to work with the natural logarithm of the Bayesian decision function. In other words, we may transfer the decision function $d_i(x) = P(x|y_i)P(y_i)$ to:

$$d_i(X) = \ln[P(x|y_i)P(y_i)] = \ln P(x|y_i) + \ln P(y_i). \tag{4.3.3}$$

Substituting equation 4.3.2 into equation 4.3.3 yields

$$d_i(X) = \ln P(y_i) - \frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|C_i| - \frac{1}{2}[(x-\mu_i)^T C_i^{-1}(x-\mu_i)]. \qquad (4.3.4)$$

Since the term $\frac{N}{2}\ln 2\pi$ does not depend on i, equation 4.3.4 can be simplified to:

$$d_i(X) = \ln P(y_i) - \frac{1}{2}\ln|C_i| - \frac{1}{2}[(x-\mu_i)^T C_i^{-1}(x-\mu_i)]. \qquad (4.3.5)$$

Equation 4.3.5 may be rewritten as:

$$d_i(X) = \ln P(y_i) - \frac{1}{2}\ln|C_i| - \frac{1}{2}x^T C_i^{-1}x + x^T C_i^{-1}\mu_i - \frac{1}{2}\mu_i^T C_i^{-1}\mu_i. \qquad (4.3.6)$$

Now, we may assume that all covariance matrices are equal, $C_i = C$ for $i = 1, ..., k$, then $-\frac{1}{2}\ln|C_i| - \frac{1}{2}x^T C_i^{-1}x$ is equal to $-\frac{1}{2}\ln|C| - \frac{1}{2}x^T C^{-1}x$, which does not depend on i. Finally, the decision function is simplified to:

$$d_i(X) = \ln P(y_i) + x^T C^{-1}\mu_i - \frac{1}{2}\mu_i^T C^{-1}\mu_i. \qquad (4.3.7)$$

This equation is called the linear discriminant function [26]. The Bayesian classification method based on this linear discriminant function is known as LDA.

LDA [26] is used as a benchmark for the proposed new classification method. The main idea of LDA is to find a transformation matrix that maximizes the ratio of overall variance to within class variance. The overall variance or total scatter measures the average diversity of all data, and the within class variance or the within scatter measures the average diversity of the data that belong to the same class. Figure 4.4 gives examples of good (i) and bad (ii) separation.

Assuming the conditional density $P(x|y_i)$ is a multivariate normal distribution and all class covariance matrices are equal, LDA produces optimal linear decision boundaries. Define the class discriminant function:

$$d_i(X) = \ln P(y_i) + x^T C^{-1}\mu_i - \frac{1}{2}\mu_i^T C^{-1}\mu_i. \qquad (4.3.8)$$

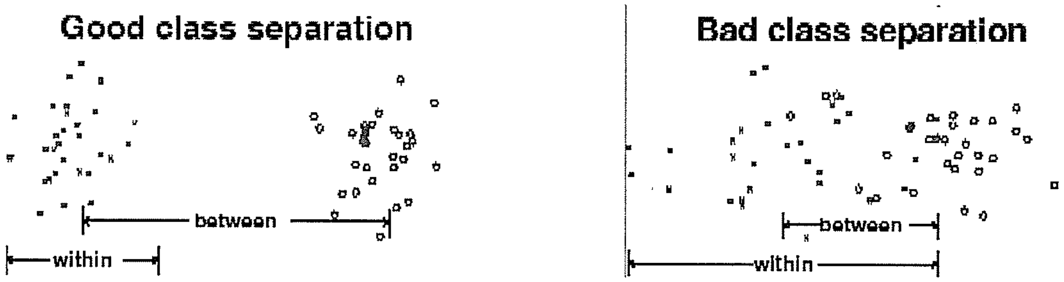


Figure 4.4: Good (i) and bad (ii) class separation

where C^{-1} is the inverse of the covariance matrix and μ_i is the mean for y_i. A sample x will be assigned to y_i, if $d_i(X)$ yields the smallest value for all discriminant functions.

Computing C^{-1} is unproblematic when the number of variables is small. However, biomedical data, such as MR spectra, often possess high dimensionality (many features), which can lead to an ill-conditioned matrix. So, before applying LDA to biomedical spectra, we must first reduce the dimensionality of the input (feature). In this dimension reduction procedure, the risk of information loss always exists.

4.4 Reduction of Dimensionality

Reduction of dimensionality is a key problem in biomedical data classification, particularly in cases where the number of variables is high compared with the number of samples. Most classification methods depend on a certain ratio of samples to variables; normally the number of variables should be no more than one third the number of samples [20]. Procedures that are sound in low-dimensional spaces can become completely impractical in a space of 100 or more dimensions [7].

Methods for reduction of dimensionality fall into two categories. In the first category are methods that aim to describe the data more succinctly, that is to express the data as concisely as possible with minimum loss of information. These methods do not rely on prior knowledge of class membership of subsets but attempt to remove irrelevant information by transforming the original data into a new set of variables. The second category of methods attempt to reduce the number of variables by selecting the best set of features for discrimination. In this case, knowledge of the class membership of the data are typically used.

4.5 Related Work

In order to give a complete description of research related to this thesis, this section approaches the discussion on related work from three directions. First, I give a description of dimension reduction methods used in MR spectra analysis. Second, I show some studies that use LDA in MR spectra classification. Finally, I describe research that use clustering methods in biomedical data analysis.

Principal component analysis (PCA) [14] is one of the most commonly used statistical method for reduction of dimensionality. PCA operates by transforming the original features into a new set of uncorrelated variables called principal components (PC's). These new variables are linear combinations of the original features derived in decreasing order of importance. The first PC accounts for the most variance in the original data. Howells et al. [18] used PCA to reduce the original 16,000 data points of a MR data set to 15 PC's. These 15 PC's, which accounted for 95% of the variance in the data set, were used as input to a neural network classification method with good results. However, a disadvantage of PCA is, in some cases, principal components may not be able to provide the best features for classification, because the PC's are ordered by variance and not by discriminatory power.

The averaging method is the most popular dimension reduction strategy. For instance, given an input feature space size of 1000, we can set an averaging window of 10 to average 1000 features to 100 features. Hagberg et al. [16] combined dimension reduction methods, such as the averaging method with LDA to classify brain tumours on the basis of metabolite measurements. The averaging method is relatively simple in computation. However, when 10 features are compressed into 1 feature, some useful information is unavoidably lost.

Feature selection is concerned with choosing the best features to use in the classification method. Fukunaga et al. [10] combined feature selection method with LDA in biomedical data classification. Finding the best subset of m features out of n may be carried out by evaluating a criterion of class separability for all possible combinations of the m features. However, the calculation of C_n^m combinations becomes prohibitively expensive for even fairly small values of n and m. On the other hand, most of the traditional methods [10] [28] of feature selection assume a small number of features and may not be of much help for data such as MR spectra which have a very large number of features.

Different from the above dimension reduction method, the proposed new dimension reduction method using clustering may cause less information loss than the averaging method and give a better computation performance than the common feature selection method.

LDA is often used in MR spectra classification. Preul et al. [25] used LDA to classify glial brain tumours on the basis of metabolite measurement. In this study, tumours were divided into three grades on the basis of biopsy data and clear separation was obtained between the three groups. The research group from the National Research Council's Institute for Biodiagnostics in Winnipeg has used LDA in a number of studies using MR spectra data. Results have shown that LDA can be used to successfully classify ^{1}H MR spectra of various diseases such as thyroid neoplasms [30] and human brain neoplasms [24].

All the above research combined LDA with some external feature reduction methods in biomedical data analysis. This research endeavors to find a new supervised pattern classification method with its own feature reduction procedure for high dimensional biomedical data classification.

Hierarchical clustering is the most widely used method for the analysis of patterns of gene expression. By grouping unknown genes with the similar structure, hierarchical clustering is very helpful in gene categorization. Eisen et al. [8] and Iyer et al. [19] have applied these techniques to the study of gene expression patterns. Gartland et al. [12] applied cluster analysis to MR spectra of samples from a variety of induced toxic states in rats. Hierarchical cluster analysis was used as an inductive method of analyzing the intensities of 16 metabolites obtained from these spectra. The cluster results showed that some of the different toxins formed a discrete cluster. Howells et al. [17] have used cluster analysis to categorize MR spectra obtained from perchloric acid extracts of normal and tumorous tissue in rats. The clustering results showed a partial separation of samples into groups representing the different tissue types.

However, in all of the above research, hierarchical clustering were used as an unsupervised classification method. Different from the above methods, our approach involves a supervised hierarchical clustering method.

Chapter 5

Research Methods and Material

5.1 Problem Statement

The success of using LDA in biomedical data classification is based on two assumptions. The conditional densities $P(x|y_i)$ are multivariate normal distributions and all covariance matrices are equal. Although many biomedical datasets can satisfy these two assumptions, there are still some exceptions. In equation 4.3.8, C^{-1} can always be found only when the number of samples is much larger than number of features. However, biomedical datasets, such as MR spectra, often have the characteristics of high dimensionality and small sample size. So, before applying LDA to biomedical datasets, we must first reduce the dimensionality of the feature space. In this dimension reduction procedure, the risk of information loss always exists.

This research endeavors to find a new supervised pattern classification method in high dimensional data classification area, especially for those cases which can not be successfully analyzed by LDA.

Although some common feature reduction methods, such as the averaging method, are simple in computation, the risk of information loss is high when using these kinds of dimension reduction methods. On the other hand, the feature selection method may provide a minimum loss of information, but this method is quite expensive in computation time.

My thesis introduces a new dimension reduction method reduces both information loss and computation complexity.

5.2 Proposed Solution

Based on finding the minimum error sum of squares, Ward's method can successfully be used in unsupervised high-dimensional data classification problems without any previous assumptions or restrictions. In this research, a new supervised pattern classification strategy and a new dimension reduction method using hierarchical clustering were build based on Ward's method.

5.2.1 Supervised Ward's Method

Suppose there are c classes of samples in the training set. The basic classification routine for our supervised Ward's method (SW) is as follows (See Figure 5.1):

1. Separate each class of samples in the training set into m clusters using Ward's method to obtain $m \times c$ clusters of patterns.

2. Calculate each cluster's centroid. Given a cluster C containing j samples: x^1, x^2 $\ldots x^j$. the cluster's centroid C_{cen} is:

$$C_{cen} = \frac{x^1 + x^2 + \cdots + x^j}{j} \tag{5.2.1}$$

3. For each sample in the validation set, calculate the Euclidean distance from this sample to the $m \times c$ cluster centroids. In other words, for each sample, we construct a distance vector of size $m \times c$.

4. Label each sample to a predicted class based on the minimal Euclidean distance between this sample and those $m \times c$ cluster centroid. A sample, x^i, will be labeled as class s if

$$d_{ip} \leq d_{iq}, \;\; p \in s, \;\; q = 1, \ldots, m \times c \tag{5.2.2}$$

where p and q are cluster centroids and p is derived from the cluster in class s.

5. After getting the predicted class labels for all the samples in the validation set, we can then compare the predicted class label with the actual class label for each sample in the validation set and calculate the classification accuracy.

6. Decrease the value of m down to 1 and repeat steps 1 to 5 to find the optimal classification accuracy.

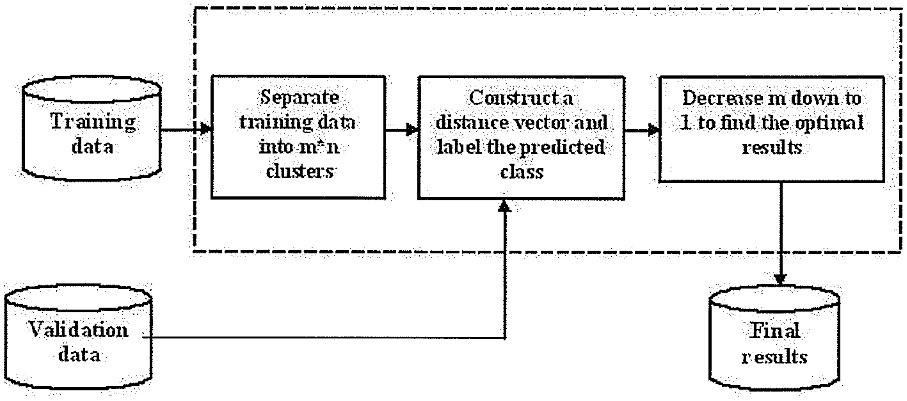

Training data
Separate training data into m*n clusters
Construct a distance vector and label the predicted class
Decrease m down to 1 to find the optimal results
Validation data
Final results

Figure 5.1: Supervised Ward's Method

5.2.2 Dimension Reduction Method Using Clustering

The intent of this method is to significantly reduce the dimensionality of the feature space by replacing the original features with distances between samples and a set of cluster centroids. This should, in general, simplify the classification task for the underlying classifier (for instance, LDA). Given a feature space size of g with e classes of samples in the training set, our dimension reduction method using hierarchical clustering based on Ward's method (DRW) is as follows:

1. Separate each class of samples in the training set into f clusters using Ward's method. Then we get $e \times f$ clusters of samples.

2. Calculate each cluster's centroid. Given a cluster C containing l samples: x^1, x^2 $\dots x^l$. the cluster's centroid C_{cen} is:

$$C_{cen} = \frac{x^1 + x^2 + \cdots + x^l}{l} \tag{5.2.3}$$

3. For each sample in the validation set, calculate the Euclidean distance from this sample to those $e \times f$ cluster centroids. For each sample, we construct a distance vector of size $e \times f$. In other words, we transform the feature space size of g into size of $e \times f$, where $e \times f \leq g$.

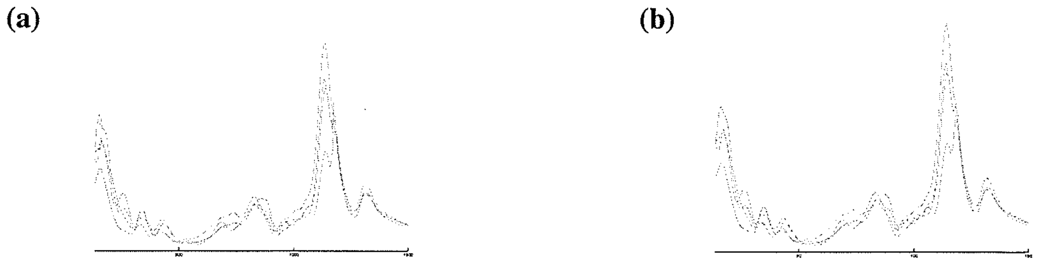
5.3 MR Spectra

In order to verify this new classification method's performance, we used two groups of biomedical datasets in our experiments: 206 MR spectra of human brain neoplasms [27] and 444 MR spectra of yeasts [22]. All data were ^{1}H MR spectra acquired at 37°C on a 360 MHz MR spectrometer.

5.3.1 Yeast

Yeasts are unicellar fungi that use the characteristics of the cell, ascospore and colony. They are found on the skin surfaces and in the intestinal tracts of warm-blooded aniamals, where they may live symbiotically or as parasties. The common yeast infection is typically Candidiasis and is caused by the yeast-like fungus Candida albicans. In addition to being the causative agent in vaginal yeast infections, Candida is also a cause of diaper rash and thrush of the mouth and throat [22]. The identification of closely related species or subspecies of yeasts is problematic.

The MR spectra of yeasts consisted of 5 different species of yeasts. A total of 444 spectra containing 1500 features were divided into five classes, according to the corresponding Candida species: 104 albicans, 91 glabrata, 81 krusei, 93 parapsilosis, and 75 tropicalis. A 250 sample training set was randomly selected from the spectra, containing 50 samples from each class. The validation set contained the remaining 194 samples. Figure 5.2 show the plots of the MR spectra of yeast with the original 1500 features (i) and 150 averaged features using a window size of 10 (ii).



1500 features

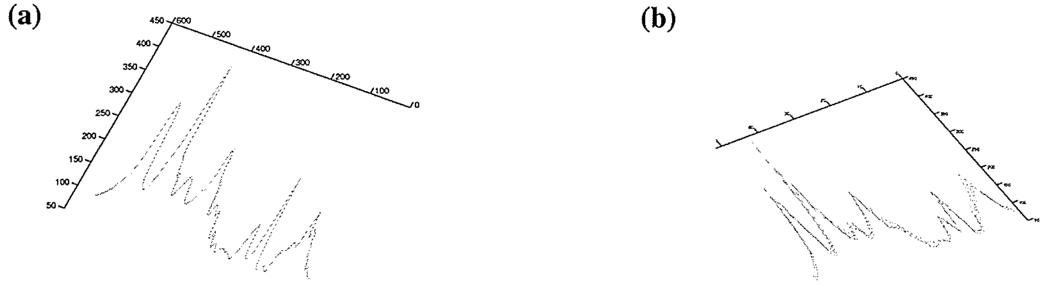
150 features

Figure 5.2: MR spectra of yeast species

5.3.2 Human Brain Neoplasms

A neoplasm or tumor, is a confined mass of abnormal tissue that proliferates rapidly and without cessation. Tumors involving the central nervous system are classified according to the type of cells and the location of the tumor. Tumors are considered primary brain tumors if they originate in the central nervous system, or secondary brain tumors if they originate elsewhere and cells of the tumor migrate to the central nervous system. A tumor is benign if it is slow growing, and malignant if it is rapidly growing and readily invades surrounding brain tissue. About 24,000 primary brain tumors and an even larger number of secondary brain tumors are diagnosed in United States each year. Meningiomas and astrocytomas are two common primary tumors.

The MR spectra of human brain neoplasms consisted of 2 different human brain neoplasms and one group of control samples of non-tumorous brain tissue from patients with epilepsy. A total of 206 spectra containing 550 features were divided into three classes: 95 meningiomas, 74 astrocytomas and 37 control samples (epilepsy). An 80 sample training set contained 29 meningioma, 31 astrocytoma and 20 epilepsy . The validation set contained the remaining 126 samples and was not used during the training phase. Figure 5.3 are the MR spectra plots of human brain neoplasms with the original 550 features (i) and 55 averaged features using a window size of 10 (ii). This MR spectra contains too much noise, so we rotate the plots to display it clearly.



550 features

55 features

Figure 5.3: MR spectra of human brain neoplasms with the original 550 features (i) and 55 averaged features (ii).

5.4 Evaluation

The first part of the evaluation is to compare the performance of DRW with the averaging method. For a given MR dataset, I first use LDA with the averaging method and then LDA with the DRW. The second part of the evaluation is to make a comparison between SW and LDA. Both of these methods were applied to the selected MR datasets.

We anticipate that SW will have greater classification accuracy than LDA in some cases and the DRW using clustering is superior to the averaging in most cases.

Chapter 6

Results and Discussion

6.1 System Development

1. **Algorithm design:** A hierarchical clustering algorithm was developed based on Ward's method [34]. This algorithm was first implemented in a separate C++ program using the GNU C++ compiler. **Algorithm 1** gives the pseudocode for the Ward's method used in this research.

Algorithm 1 Ward's clustering method

```
Form a cluster for each item
Until merge into one cluster do
    Pick arbitrary cluster as Current Cluster,
    Found = False;
    Until not Found do
        find the closest neighbor cluster to Current Cluster;
        if they are reciprocal nearest neighbors then
            Merge them;
            Found = True;
        else
            Change Current Cluster to its nearest neighbor;
        end if
    end-do
end-do
```

2. **Algorithm Test:**

   Some synthetic two dimensional data sets were used to test if this algorithm can successfully distinguish simple clusters. The 200 × 2 data set labelled as two_long_narrow_distribution_circles contains one thick circle with size of 100×2 and the other 100×2 long narrow circle. In the experiment, when the number of clusters is set to 5, Ward's method can clearly distinguish all 5 circles. Figure 6.1 shows the clustering results ranging from 2 to 5 clusters for this data set.

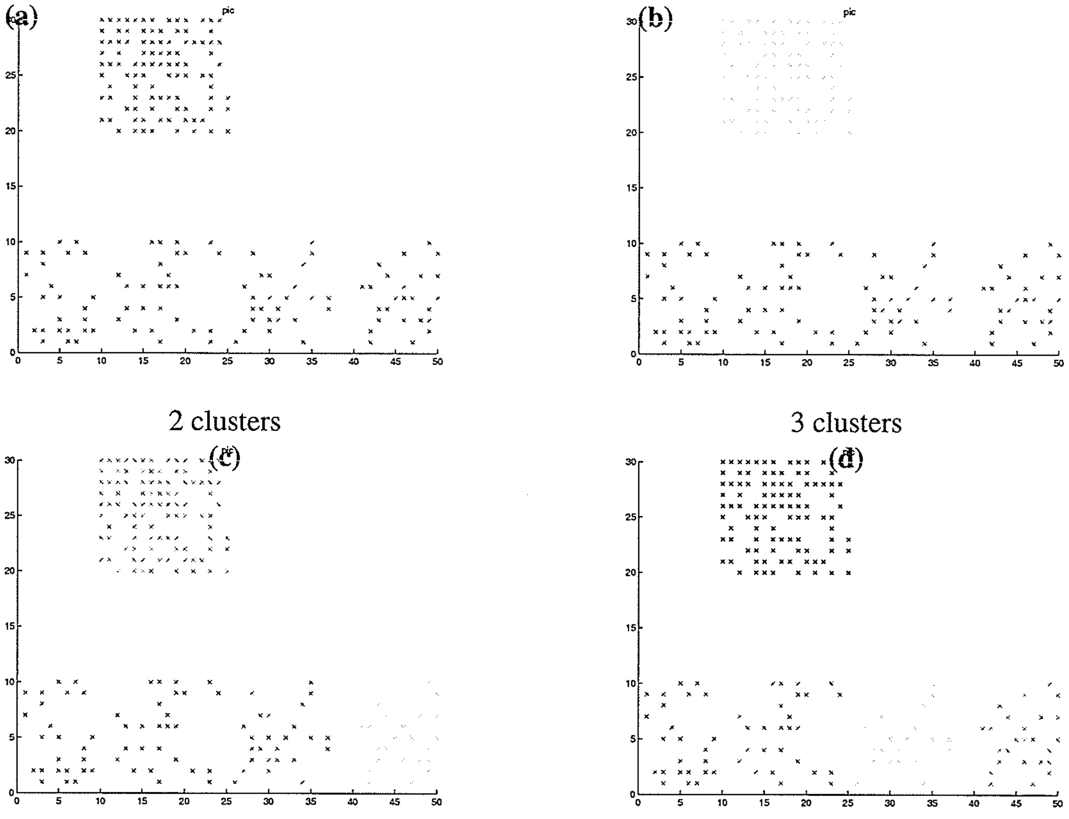
   The 150 × 2 data set labelled as variable_sized_circles contains one dense circle with size of 50 × 2 and several sparse circles that ranging from 10 × 2 to 30 × 2 in size. Starting with 4 clusters, our algorithm can easily determine the corresponding number of circles based on the distance measurement. Figure 6.2 gives the clustering results from 4 to 7 clusters for this data set.

   The 150×2 data set labelled as two_concentric_circles contains two 75×2 concentric circles. When the number of clusters is set to 2, our algorithm correctly predict the class labels. Figure 6.3 presents the clustering results ranged from 2 to 5 clusters for this data set.

   The test results show that this algorithm generates reasonable clustering results for all synthetic data sets. Some MATLAB programs were developed for creating these two dimensional data sets and analyzing their clustering results.

3. **Scopira module development:** The second phase was to implement the hierarchical clustering algorithm in Scopira [5]. Some Scopira modules were developed based on the above C++ program.

   In order to take advantage of Ward's method, three modules were created: the Ward's engine module used to generate the intermediate data for the data generating model



2 clusters

3 clusters

4 clusters

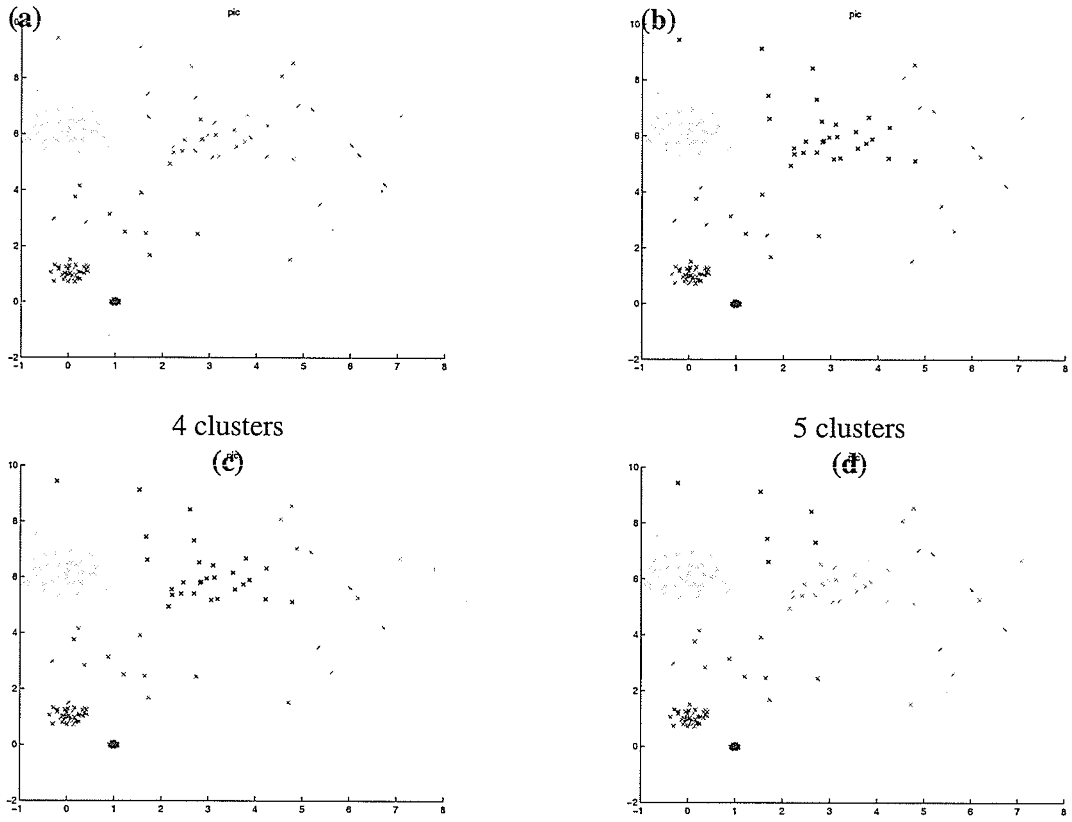
5 clusters

Figure 6.1: Clustering results for two long narrow distributions

and the tree map module; the data generating module used to generate clustering results based on the intermediate data; and the tree map module used to display a complete hierarchical cluster tree based on the intermediate data.

Figure 6.4 is a Scopira map of Ward's method. This map gives the clustering partition matrix and cluster centroids as outputs. In Figure 6.5, the ward_tree_map shows a hierarchical tree. Users can redraw this tree by changing the number of clusters.

4. **Methods implementation:** In the third phase of my research, the supervised Ward's method (SW) and the dimension reduction method using clustering (DRW) were



4 clusters

5 clusters

6 clusters

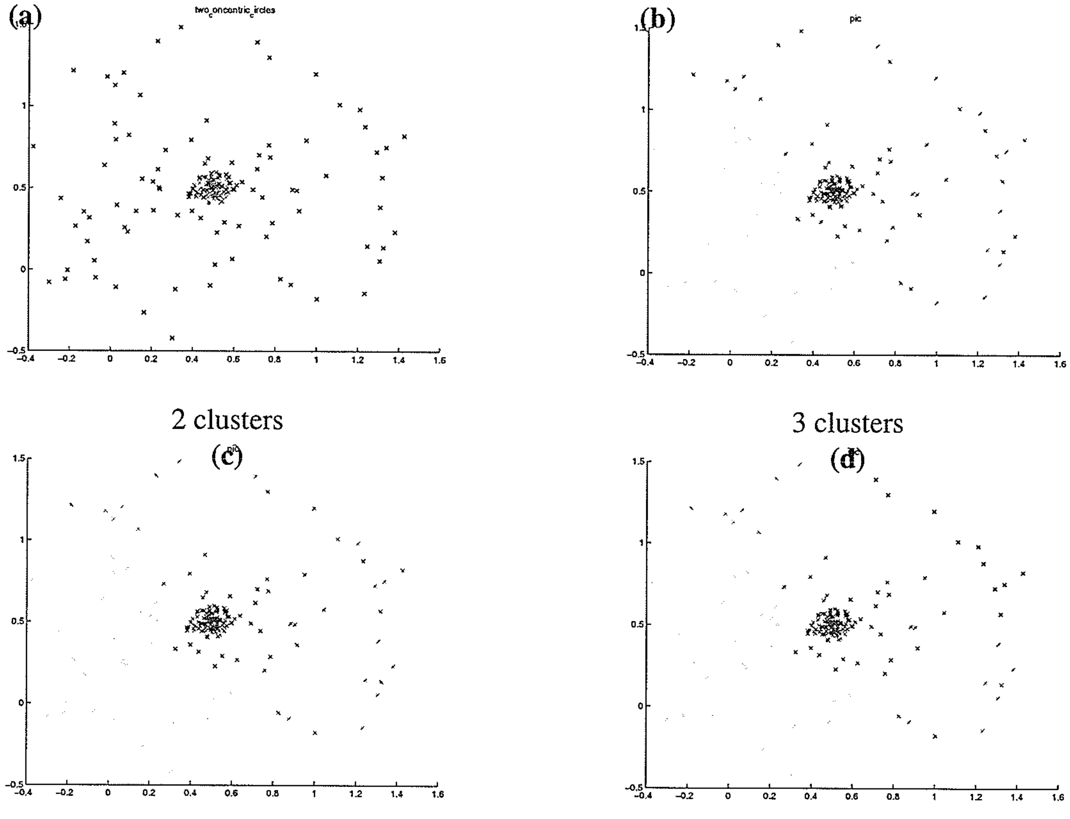
7 clusters

Figure 6.2: Clustering results for variable sized circles

implemented using MATLAB [4]. The MATLAB signal processing toolbox was used in this phase.

6.2 Results using Various Distance Metrics

In the experiment using the MR spectra of human brain neoplasms, SW with Euclidean distance produced a classification accuracy of 77.8% (see Table 6.1). SW with Squared Euclidean distance gave an accuracy of 78.6% (see Table 6.2). While Manhattan distance



2 clusters

3 clusters

4 clusters

5 clusters

Figure 6.3: Clustering results for two concentric circles

gave an accuracy of 76.2% (see Table 6.3). Figure 6.6 is a summary chart of these classification results.

In the experiment of yeast-candidiasis, SW with Euclidean distance produced a classification accuracy of 78.9% (see Table 6.4). SW with Squared Euclidean distance gave an accuracy of 79.9% (see Table 6.5). While Manhattan distance gave an accuracy of 77.3% (see Table 6.6). Figure 6.7 shows a chart that compares these classification results.

From the above results, we found the Squared Euclidean distance gives better accuracy compared to the other distance measures. So, in the rest of our experiments, we exclusively use this distance measure.

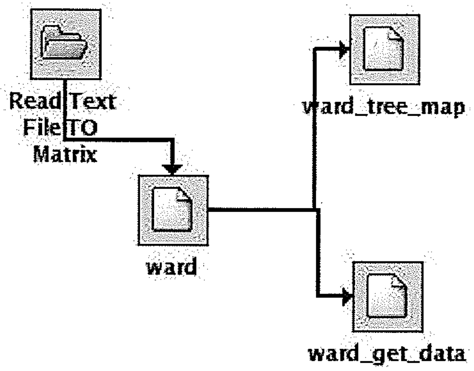
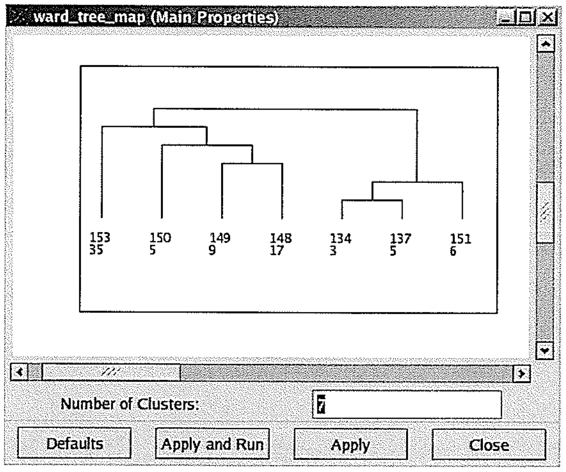


Figure 6.4: Ward map implemented in Scopira



Number of Clusters: 7

Figure 6.5: Hierarchical tree produced by Ward's method

6.3 Results using Original Features

We first apply LDA without any feature reduction method and SW to both datasets using all of the original features. In the experiment with human brain neoplasms, using all 550 features, LDA produced a classification accuracy of 57.9% (see Table 6.7). While SW had an accuracy of 78.6% (see Table 6.1). Figure 6.8 compares the classification results in a chart.

In the experiment of yeast-candidiasis, using all 1500 features, LDA could not produce

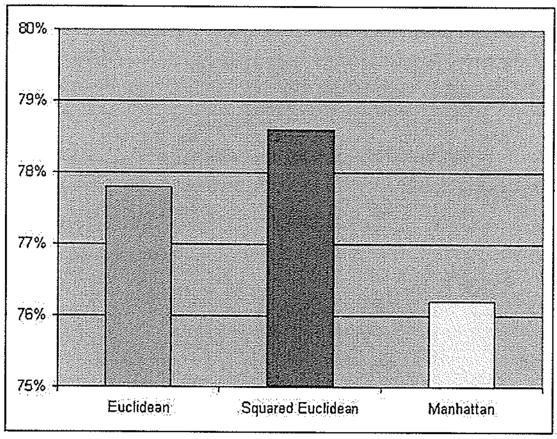


Figure 6.6: Human brain neoplasm classification results using SW with different distance measures.

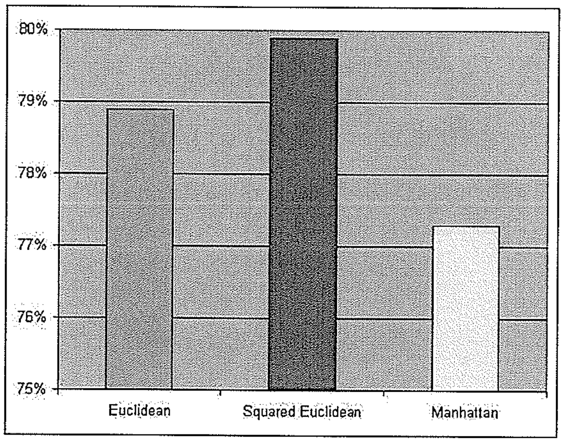


Figure 6.7: Yeast species classification results using SW with different distance measures.

a classification result due to the high dimensionality of the feature space. On the other hand, SW had an accuracy of 79.9% (see Table 6.8).

From the above results, we find that without feature reduction methods, LDA can not give proper classification results for datasets that have high dimensionality, while SW can produce reasonable classification results.

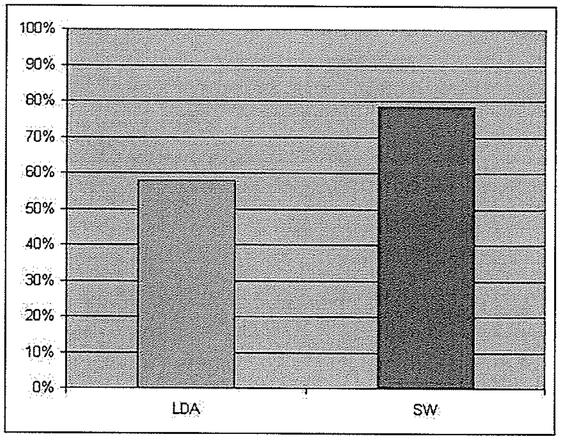


Figure 6.8: Human brain neoplasms results using original features

6.4 Feature Averaging versus Feature Clustering

In these experiments, DRW is compared against standard feature averaging. The confusion matrices generated for both feature reduction methods are listed in Table 6.9-6.12.

In the experiment with human brain neoplasms, averaging the 550 input features to 55 features, LDA produced a classification accuracy of 77.8% (see Table 6.9). On the other hand, DRW produced an accuracy of 83.3% (see Table 6.10). Figure 6.9 plots the classification results using both methods.

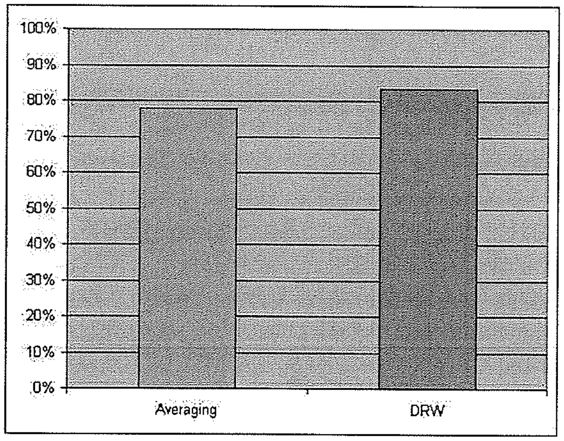


Figure 6.9: Human brain neoplasms results using feature averaging and DRW

In the experiment of yeast-candidiasis, averaging 1500 input features to 100 features, LDA produced a classification result with the accuracy of 89.7% (see Table 6.11). In this case, DRW produced the same classification results (see Table 6.12). Figure 6.10 compares the results of both methods.

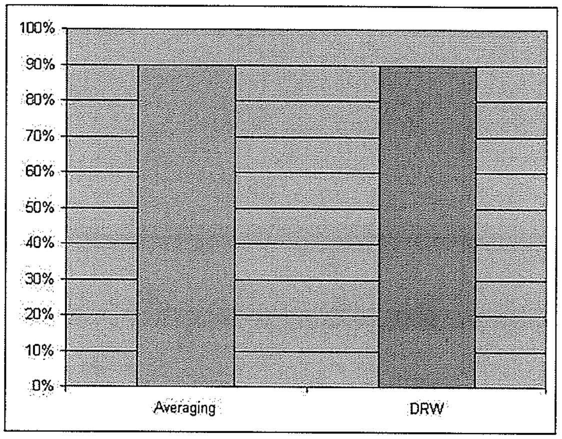


Figure 6.10: Yeast species results using feature averaging and DRW

6.5 DRW and Human Brain Neoplasms Misclassifications

Using DRW, we have seen that the overall classification accuracy increased by 7.1% compared to standard feature averaging (83.3% versus 77.8%). Moreover, with DRW, the classification errors were more conservative. DRW never misclassified ME or AS spectra (abnormal tissue) as EP (normal tissue) and 2 of the 17 EP samples were misclassified as AS. With standard feature averaging, on the other hand, 10 of the 109 abnormal tissue samples were misclassified as normal while 1 of the EP samples was misclassified as abnormal. Obviously, it is better to misclassify a type of tumor as another type of tumor rather than normal tissue. We may define the false normals and abnormals rate, F_{na}, as the percentage of misclassified normals and abnormals. In the case of standard averaging, F_{na} is 8.7%, while, with DRW, F_{na} was only 1.6%.

A conservative misclassification rate, that is, small F_{na}, is especially important in biomedical data analysis, which further justifies the use of DRW in this domain.

Table 6.1:
Human Brain Neoplasms: SW with Euclidean

77.8%	ME	EP	AS	Acc
ME (66)	52	2	12	78.8%
EP (17)	0	16	1	94.1%
AS (43)	5	8	30	69.8%

Table 6.2:
Human Brain Neoplasms: SW with Squared Euclidean

78.6%	ME	EP	AS	Acc
ME (66)	51	0	15	77.3%
EP (17)	0	15	2	88.2%
AS (43)	4	0	33	76.7%

Table 6.3:
Human Brain Neoplasms: SW with Manhattan

76.2%	ME	EP	AS	Acc
ME (66)	50	0	13	75.8%
EP (17)	0	17	0	100.0%
AS (43)	13	1	29	67.4%

Table 6.4:
Yeast Species: SW with Euclidean

78.9%	AL	GL	KR	PA	TR	Acc
AL (54)	36	8	0	0	10	66.7%
GL (41)	1	32	6	0	2	78.1%
KR (31)	0	3	25	0	3	80.7%
PA (43)	5	0	0	36	2	83.7%
TR (25)	0	1	0	0	24	96.0%

Table 6.5:
Yeast Species: SW with Squared Euclidean

79.9%	AL	GL	KR	PA	TR	Acc
AL (54)	36	8	0	0	10	66.7%
GL (41)	1	32	6	0	2	78.1%
KR (31)	0	1	27	0	3	87.1%
PA (43)	5	0	0	36	2	83.7%
TR (25)	0	1	0	0	24	96.0%

Table 6.6:
Yeast Species: SW with Manhattan

77.3%	AL	GL	KR	PA	TR	Acc
AL (54)	36	8	0	0	10	66.7%
GL (41)	3	30	6	0	2	73.2%
KR (31)	0	3	25	0	3	80.7%
PA (43)	5	0	0	36	2	83.7%
TR (25)	0	2	0	0	23	92.0%

Table 6.7:
Human Brain Neoplasms: LDA

57.9%	ME	EP	AS	Acc
ME (66)	38	8	20	57.6%
EP (17)	4	10	3	58.8%
AS (43)	11	7	25	58.1%

Table 6.8:
Yeast Species: SW

79.9%	AL	GL	KR	PA	TR	Acc
AL (54)	36	8	0	0	10	66.7%
GL (41)	1	32	6	0	2	78.1%
KR (31)	0	1	27	0	3	87.1%
PA (43)	5	0	0	36	2	83.7%
TR (25)	0	1	0	0	24	96.0%

Table 6.9:
Human Brain Neoplasms: LDA with Averaging

77.8%	ME	EP	AS	Acc
ME (66)	52	2	12	78.8%
EP (17)	0	16	1	94.1%
AS (43)	5	8	30	69.8%

Table 6.10:
Human Brain Neoplasms: LDA with DRW

83.3%	ME	EP	AS	Acc
ME (66)	51	0	15	77.3%
EP (17)	0	15	2	88.2%
AS (43)	4	0	39	90.7%

Table 6.11:
Yeast Species: LDA with Feature Averaging

89.7%	AL	GL	KR	PA	TR	Acc
AL (54)	46	0	1	1	6	85.2%
GL (41)	0	38	2	0	1	92.7%
KR (31)	0	5	23	2	1	74.2%
PA (43)	0	0	0	42	1	97.7%
TR (25)	0	0	0	0	25	100.0%

Table 6.12:
Yeast Species: LDA with DRW

89.7%	AL	GL	KR	PA	TR	Acc
AL (54)	46	0	1	1	6	85.2%
GL (41)	0	38	2	0	1	92.7%
KR (31)	0	5	23	2	1	74.2%
PA (43)	0	0	0	42	1	97.7%
TR (25)	0	0	0	0	25	100.0%

Chapter 7

Conclusion

This research yields a new classification method and a new dimension reduction method which constitute my thesis. Contributions to the study of biomedical data classification include the dimension reduction method using clustering (DRW) and the supervised Ward's method (SW).

1. Dimension Reduction Method Using Clustering (DRW)

    - reduces information loss compared to other dimension reduction methods, such as feature averaging.

    - combined with LDA, increases classification accuracy, especially when applied to human brain neoplasm MR spectra.

2. Supervised Ward's Method (SW)

    - apply hierarchical clustering method for supervised classification.

    - develop a new classification method.

Our experiment results show that SW can produce stable and acceptable classification results. In real-world classification problems, SW is more suitable for those cases that can not be successfully classified by LDA.

DRW generated results at least as good as the feature averaging method. In the case of the human brain neoplasm spectra, DRW produced significantly improved classification accuracy with more conservative misclassifications.

We also compared the performance of three main distance measures using MR spectra with Ward's method. The experiments showed that the Squared Euclidean distance is more suitable to be used with Ward's method in MR spectra analysis.

A research paper based on the proposed research has been presented at 2004 IEEE Canadian Conference on Electrical and Computer Engineering [35]. I expect that future results will be published in an international journal on biomedical informatics.

Bibliography

[1] Michael R. Anderberg, *Cluster analysis for applications*, ch. 6, Academic Press, New York, 1973.

[2] E. Raymond Andrew, Graeme Bydder, John Griffiths, and Peter Styles, *Clinical magnetic resonance imaging and spectroscopy*, ch. 4, JohnWiley and Sons, New York, 1990.

[3] J. L. Bock, *Nmr in clinical chemistry where do we stand?*, Clinical Chemistry **40** (1994), no. 7, 1215–1217.

[4] Eugene N. Bruce, *Biomedical signal processing and signal modeling*, ch. 3, John Wiley & Sons, Weinheim, 2001.

[5] Aleksander B. Demko, Nicolino J. Pizzi, and Ray L. Somorjai, *Scopira: A system for the analysis of biomedical data*, Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering (Winnipeg, Canada), 12 May-15 May 2002, pp. 1093–1098.

[6] David Drysdale, *Tutorial introduction to guile*, 2000.

[7] Richard O. Duda and Peter E. Hart, *Pattern classification and scene analysis*, ch. 2, Wiley, New York, 1973.

[8] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences of the United States of America **95** (1998), no. 25, 14863–14868.

[9] Horst Friebolin, *Basic one- and two-dimensional NMR spectroscopy*, ch. 4, Wiley-VCH, New York, 1998.

[10] Alvin K. Fukunaga, *Introduction to statistical pattern recognition*, ch. 2, Academic Press, Boston, 1990.

[11] David G. Gadian, *Nuclear magnetic resonance and its applications to living systems*, ch. 3, Oxford: Clarendon Press, second edition, 1995.

[12] K.P. Gartland, C.R. Beddell, J.C. Lindon, and J.K. Nicholson, *Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine*, Molecular Pharmacology **39** (1991), no. 5, 629–642.

[13] Karl Frederich Gauss, *Theory of the combination of observations least subject to errors, part one, part two, supplement*, ch. 1, Society for Industrial and Applied Mathematics, Philadelphia, 1995.

[14] Allan D. Gordon, *Classification: Methods of the exploratory analysis of multivariate data*, ch. 5, Chapman and Hall, New York, 1981.

[15] Arthur Griffith, *Gnome/gtk+ programming bible*, ch. 4, Hungry Minds, 2000.

[16] Gisela Hagberg, Allessandro P. Burlina, Irina Mader, Werner Roser, Ernst W. Radue, and Joachim Seelig, *In vivo proton MR spectroscopy of human gliomas: Definition*

of metabolic coordinates for multi-dimensional classification, Magnetic Resonance in Medicine **34** (1995), no. 2, 242–252.

[17] Sian L. Howells, Richard J. Maxwell, and John R. Griffiths, *Classification of tumour ^{1}H NMR spectra by pattern recognition*, NMR in Biomedicine **5** (1992), 59–64.

[18] Sian L. Howells, Richard J. Maxwell, A. C. Peet, and John R. Griffiths, *An investigation of tumor ^{1}H nuclear magnetic resonance spectra by the application of chemometric techniques*, Magnetic Resonance in Medicine **28** (1992), 214–236.

[19] Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, and Patrick O. Brown, *The transcriptional program in the response of human fibroblasts to serum*, Science **283** (1999), 83–87.

[20] Bruce R. Kowalski and Svante Wold, *Handbook of statistics*, ch. 5, North Holland Publishing Company, Amsterdam, 1982.

[21] M. O. Leach, *The physics of medical imaging*, ch. 4, Institute of Physics Publishing, Bristol, 3rd edition, 1992.

[22] Robert K. Mortimer, Rebecca Contopoulou, and John King, *Genetic and physical maps of saccharomyces cerevisiae*, Yeast **8** (1992), 817–902.

[23] William Negendank, *Studies of human tumours by MRS: A review*, NMR in Biomedicine (1992), no. 5, 303324.

[24] A. Nikulin, K. M. Briere, L. Friesen, I.C.P. Smith, and R. L. Somorjai, *Genetic algorithm-guided optimal attribute selection: A novel preprocessor for classifying*

MR spectra, Proceedings of Society for Magnetic Resonance in Medicine (Nice, France), 19 August-25 August 1995, pp. 1940–1948.

[25] M. C. Preul, Z. Caramanos, D.L. Collins, J-G. Villemure, W. Feindel, and D.L. Arnold, *Linear discriminant analysis based on proton MR spectroscopic imaging of human brain tumours improves pre-operative diagnosis*, Proceedings of the 2nd Meeting of the Society of Magnetic Resonance (San Francisco, United States), 10 August-15 August 1994, pp. 125–131.

[26] Alvin C. Rencher, *Methods of multivariate analysis*, ch. 5, Wiley, New York, 1995.

[27] Brian Ross and Stefan Bluml, *Magnetic resonance spectroscopy of the human brain*, The Anatomical Record **265** (2001), 54–84.

[28] David W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, ch. 4, Wiley, New York, 1992.

[29] G.A.F Seber, *Multivariate observations*, ch. 3, Wiley, New York, 1984.

[30] Ray L. Somorjai, Alexander E. Nikulin, Nicolino J. Pizzi, Dick Jackson, Gordon Scarth, Brion Dolenko, Heather Gordon, Peter Russell, Cynthia L. Lean, Leigh Delbridge, Carolyn E. Mountford, and Ian C. P. Smith, *Computerized consensus diagnosis: A classification strategy for the robust analysis of MR spectra. I. application to ^1H spectra of thyroid neoplasms*, Magnetic Resonance in Medicine **33** (1995), 257–263.

[31] Alberto Spisni, *Magnetic resonance spectroscopy in biology and medicine*, ch. 4, Pergamon Press, New York, 1992.

[32] Joe H. Ward, *Hierarchical grouping to optimize an objective function*, American Statistical Association Journal **58** (1963), no. 301, 236–244.

[33] Michael W. Weiner, *Clinical applications of mr spectroscopy and spectroscopic imaging*, Proceedings of the 2nd Annual Meeting of the SMR (1994), 185–190.

[34] David Wishart, *An algorithm for hierarchical classification*, Biometrics **18** (1969), no. 256, 165–170.

[35] Hu Yang and Nicolino J. Pizzi, *Biomedical data classification using hierarchical clustering*, Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (Niagara Falls, Canada), 02 May-05 May 2004, pp. 1861–1864.