Case-Base Retrieval of Childhood Leukaemia Patients Using Gene Expression Data

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Ali Anaissi

in

School of Software
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
JANUARY 2013

UNIVERSITY OF TECHNOLOGY, SYDNEY
SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled "**Case-Base Retrieval of Childhood Leukaemia Patients Using Gene Expression Data**" by **Ali Anaissi** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Dated: January 2013

Research Supervisor: _____
Dr Madhu Goyal

CERTIFICATE

Date: **January 2013**

Author: **Ali Anaissi**

Title: **Case-Base Retrieval of Childhood Leukaemia Patients Using Gene Expression Data**

Degree: **Ph.D.**

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

Signature of Author

Acknowledgements

My first acknowledgement is to God for being my source of spiritual strength. I would like to thank my supervisor Dr Madhu Goyal and Dr Paul Kennedy for the opportunity to work with them on this thesis, for all of the guidance they have given me along the way. I would like to thanks Prof Jie Lu for accepting me to be a member of QCIS lab. The University of Technology Sydney has provided me not only with a wonderful education but also the scholarship that made this work possible and for that I am grateful. The Children's Hospital at Westmead for providing me the gene expression datasets. I thank my friends and colleagues at the QCIS lab for the unforgettable memories. Thank you all.

To my father and mother who will always be an inspiration throughout my life. To my wife for being my partner in life. And, to my children Joelle and Jawad, who will always be my blessings from God. You two represent my greatest commitments in life. I pray for wisdom so I can accomplish my job as your father. All together, we have survived throughout this, in our situation, challenging journey. I love you all.

Table of Contents

List of Tables

List of Figures

Abstract

Acute Lymphoblastic Leukaemia (ALL) is the most common childhood malignancy. Nowadays, ALL is diagnosed by a full blood count and a bone marrow biopsy. With microarray technology, it is becoming more feasible to look at the problem from a genetic point of view and to perform assessment for each patient. This thesis proposes a case-base retrieval framework for ALL using a nearest neighbour classifier that can retrieve previously treated patients based on their gene expression data. However, the wealth of gene expression values being generated by high throughout microarray technologies leads to complex high dimensional datasets, and there is a critical need to apply data-mining and computational intelligence techniques to analyse these datasets efficiently. Gene expression datasets are typically noisy and have very high dimensionality. Moreover, gene expression microarray datasets often consist of a limited number of observations relative to the large number of gene expression values (thousands of genes). These characteristics adversely affect the analysis of microarray datasets and pose a challenge for building an efficient gene-based similarity model. Four problems are associated with calculating the similarity between cancer patients on the basis of their gene expression data: feature selection, dimensionality reduction, feature weighting and imbalanced classes.

The main contributions of this thesis are: (i) a case-base retrieval framework, (ii) a

Balanced Iterative Random Forest algorithm for feature selection, (iii) a Local Principal Component algorithm for dimensionality reduction and visualization and (iv) a Weight Learning Genetic algorithm for feature weighting.

This thesis introduces Balanced Iterative Random Forest (BIRF) algorithm for selecting the most relevant features to the disease and discarding the non-relevant genes. Balanced iterative random forest is applied on four cancer microarray datasets: Childhood Leukaemia dataset, Golub Leukaemia dataset, Colon dataset and Lung cancer dataset. Childhood Leukaemia dataset represents the main target of this project and it is collected from The Children's Hospital at Westmead. Patients are classified based on the cancer's risk type (Medium, Standard and High risk); Colon cancer (cancer vs. normal); Golub Leukaemia dataset (acute lymphoblastic leukaemia vs. acute myeloid leukaemia) and Lung cancer (malignant pleural mesothelioma or adenocarcinoma). The results obtained by BIRF are compared to those of Support Vector Machine-Recursive Feature Elimination (SVM-RFE) and Naive Bayes (NB) classifiers. The BIRF approach results are competitive with these state-of-art methods and better in some cases. The Local Principal Component (LPC) algorithm introduced in this thesis for visualization is validated on three datasets: Childhood Leukaemia, Swiss-roll and Iris datasets. Significant results are achieved with LPC algorithm in comparison to other methods including local linear embedding and principal component analysis. This thesis introduces a Weight Learning Genetic algorithm based on genetic algorithms for feature weighting in the nearest neighbour classifier. The results show that a weighted nearest neighbour classifier with weights generated from the Weight Learning Genetic algorithm produces better results than the un-weighted nearest neighbour algorithm.

This thesis also applies synthetic minority over sampling technique (SMOTE) to increase the number of points in the minority classes and reduce the effect of imbalanced classes. The results show that the minority class becomes recognised by the nearest neighbour classifier. SMOTE also reduces the effect of imbalanced classes in predicting the class of new queries especially if the query sample should be classified to the minority class.

Chapter 1

Introduction

Gene expression similarity measurement has received significant attention in the field of cancer diagnosis and treatment [52] [61]. However, a trustworthy and precise similarity measurement is essential for successful diagnosis and treatment of cancer. The ability to successfully distinguish between patients based on their gene expression data and to look in the neighbourhood space of patients to see how patients are similar or dissimilar to others requires numerous data mining and computational intelligence techniques to be incorporated into the similarity measurement process. The huge number of gene expression values generated by microarray technology leads to very complex datasets and raises numerous statistical and data mining questions. Gene expression datasets are typically noisy and have very high dimensionality [8]. Moreover, gene expression microarray datasets often consist of a limited number of observations (hundreds) relative to the large number of gene expression values (thousands of genes). These characteristics adversely affect the analysis of microarray datasets and pose a challenge for building effective similarity measurements. These characteristics also result in difficulties in working with standard machine learning techniques, which must be modified to deal with the complexities of gene expression

data.

The main objective of this thesis is to develop a case-base retrieval framework able to retrieve similar patients from a case base of ALL patients based on their gene expression data. Four problems are associated with calculating similarity between patients based on their gene expression data: feature selection, dimensionality reduction, feature weighting and imbalanced classes.

Feature selection involves identification of bio-markers that are strongly associated with the disease and that can be used to distinguish classes of patients. Too many features or genes in a dataset adversely affect similarity measurement as many of these genes are irrelevant to a specific trait of interest.

Dimensionality reduction is the transformation of a high-dimensional dataset into a lower dimensional one by extracting new features using a mapping (either linear or non-linear). This transformation of the dataset into a lower dimensional form is required due to the curse of dimensionality which adversely affects the distance measurement in such highly dimensional data [9].

Feature weighting is the estimation of the relative influence of individual features with respect to classification performance. The relative importance of feature plays a significant role in similarity measurement as the distance function can be very sensitive to the relative importance of features.

Imbalanced data is where the number of patients belonging to each class is not the same. Imbalanced classes is a common problem in the biomedical field [35] [3] [36]. For example, the main experiments of this thesis are performed on a real Childhood Leukaemia gene expression dataset collected from The Children's Hospital at Westmead and is composed of 110 patients. Each patient has more than twenty two

thousand gene expression values. Patients are classified into three categories based on the cancer's risk type: Medium, Standard and High risk. The majority of patients (78 patients) are classified as a Medium risk, 21 patients are classified as a Standard risk where the minority of patients (11 patients) are classified as High risk. This imbalanced classes problem adversely affects the case-base retrieval process and the classification performance in the feature selection process as it can results in a trivial classifier that classifies all patients as the majority class.

1.1 Motivation

Acute Lymphoblastic Leukaemia is the most common childhood malignancy [69]. It is a type of cancer that affects the blood and bone marrow. The causes of ALL are still unknown but are likely thought to result from mutations of genes [73].

Nowadays, ALL is diagnosed by a full blood count and a bone marrow biopsy. Based on these examinations, an ALL patient's risk of relapse and appropriate treatment are identified. Most children achieve an initial remission, yet approximately 20% of children with ALL suffer a relapse [29]. This relapse problem is considered as one of the major obstacles and difficulties where the cancer recurs facing clinicians in curing ALL patients. One reason for relapse is for incorrect therapy due to misclassification of risk factors of ALL patients [29]. Consequently, accurate risk assessment of patients is crucial for successful treatment.

The motivation of this thesis is to develop a point-of-care clinical software for better diagnosis and treatment of childhood leukaemia sufferers. This thesis develops a case-base retrieval framework with the eventual aim of supporting clinicians and biologists in predicting how paediatric cancer sufferers will react to treatment by comparing

current to previous patients on the basis of their gene expression data.

1.2 Objective and Aims

The objective of this thesis is to develop a case-base retrieval framework for childhood leukaemia patients. By observing neighbouring patient treatment and outcome, more reliable decisions about this new patient can be made. The assumption is made that patients with similar gene expression profiles will react similarly to therapy and should be treated similarly.

In order to achieve this objective, this thesis has four aims.

1. The first aim is to develop an algorithm for selecting features from gene expression datasets. Gene expression datasets have a huge number of genes but only a small set of these genes is related to the desired output and can be used in prediction and classification [38]. This thesis aims to select the most relevant features that play a major role in discriminating between patients. Many learning algorithms developed before the advent of microarray analysis do not effectively handle data with high dimensionality and small sample size. Microarray analysis raises the need for a special feature selection process. This aim is achieved by developing an algorithm for feature selection with the consideration of the difficulties that accompany these datasets such as the low number of observations, high amount of noise, correlation between features and imbalanced classes.

2. The second aim is to reduce the dimensionality of the dataset in order to enhance distance measurements which are affected by the curse of dimensionality. Feature selection on its own is not sufficient to reduce the dimensionality for accurate similarity measurement. Although feature selection reduces the dimensionality, the dataset may still exist in a high dimensional space. Therefore, further dimensionality reduction is needed.

3. The third aim is to develop a weight learning algorithm for feature weighting. Distance measurement aims to calculate the distance between two patients based on their associated gene expression. However, genes are not equally important as some genes are more significant and important than others [88] [84] and they should attract a higher weight. On the other hand, features with less influence should receive a low weight.

4. The fourth aim handles the problem of imbalanced classes and the small number of observations in the minority class, which affects classification performance and the case retrieval.

1.3 Contributions

This thesis makes the following contributions to knowledge:

1. A Case-Base Retrieval Framework

   The framework is a combination of different computational intelligence and data mining techniques for pre-processing the data before measuring the similarity between ALL patients. It is composed of two modules: module 1 pre-processes the training dataset in order to handle the complexities of the gene expression

datasets, module 2, on the other hand, processes the new query data using the outcomes of module 1.

2. The Balanced Iterative Random Forest algorithm

   This is an embedded feature selection algorithm with the capability of dealing with imbalanced microarray datasets. Performance comparisons show that BIRF out perform other gene selection methods tested particularly on class-imbalanced data.

3. The Local Principal Component algorithm

   This is a non-linear dimensionality reduction algorithm which maps high dimensional data to a lower dimensional space. The algorithm is based on the first principal component of the local neighbourhood of each data point. Performance comparisons show that LPC out perform Principal Component Analysis and competitive to local linear embedding algorithm.

4. A Weight Learning Genetic algorithm

   This algorithm is for feature weighting in the nearest neighbour classifier. Wrapper feature weighting method based on Genetic Algorithms is proposed to seek the genes encoding weights for the similarity measurement. Performance comparisons show that the weighted nearest neighbour classifier using the generated weights from the Weight Learning Genetic algorithm out performs the un-weighted nearest neighbour algorithm.

1.4 Organization of this Thesis

This thesis is presented as follows. Chapter 2 presents background concepts and literature review related to the topic of microarray data, case-based reasoning, feature selection, dimensionality reduction, feature weighting and imbalanced classes. Chapter 3 presents the proposed case-base retrieval framework. Chapter 4 introduces the Balanced Iterative Random Forest algorithm for gene selection from microarray data. Chapter 5 presents the dimensionality reduction and visualisation of gene expression datasets using Local Principal Component algorithm. Chapter 6 introduces the Weight Learning Genetic algorithm for feature weighting and the method to handle the problems of imbalanced classes and low number of observations. Chapter 7 concludes the thesis and outlines future research directions.

Chapter 2

Literature Review

This chapter provides the necessary background information for this thesis. Section 2.1 presents the description of the case-based reasoning systems and the difficulties of case base retrieval process in the microarray domain. Section 2.2 introduces the techniques for feature selection with the focus on the strengths and weakness of these techniques. Section 2.3 gives an overview of dimensionality reduction techniques and how they relate to this thesis. Section 2.4 presents the methodologies that can be used for feature weighting. Section 2.5 presents the approaches to solve the imbalanced classes problem. Section 2.6 introduces measures used in this thesis for evaluating classifier performance for imbalanced data.

2.1 Case-Based Reasoning

Case-based reasoning (CBR) [93] is a methodology for building intelligent systems for the storage and retrieval of past experiences to solve new problems. The main idea of CBR is that a new problem can be solved by adapting solutions that were used to solve previously encountered problems [49]. A case-based reasoning system is composed of four processes: retrieve, reuse, revise and retain (see Figure 2.1). The

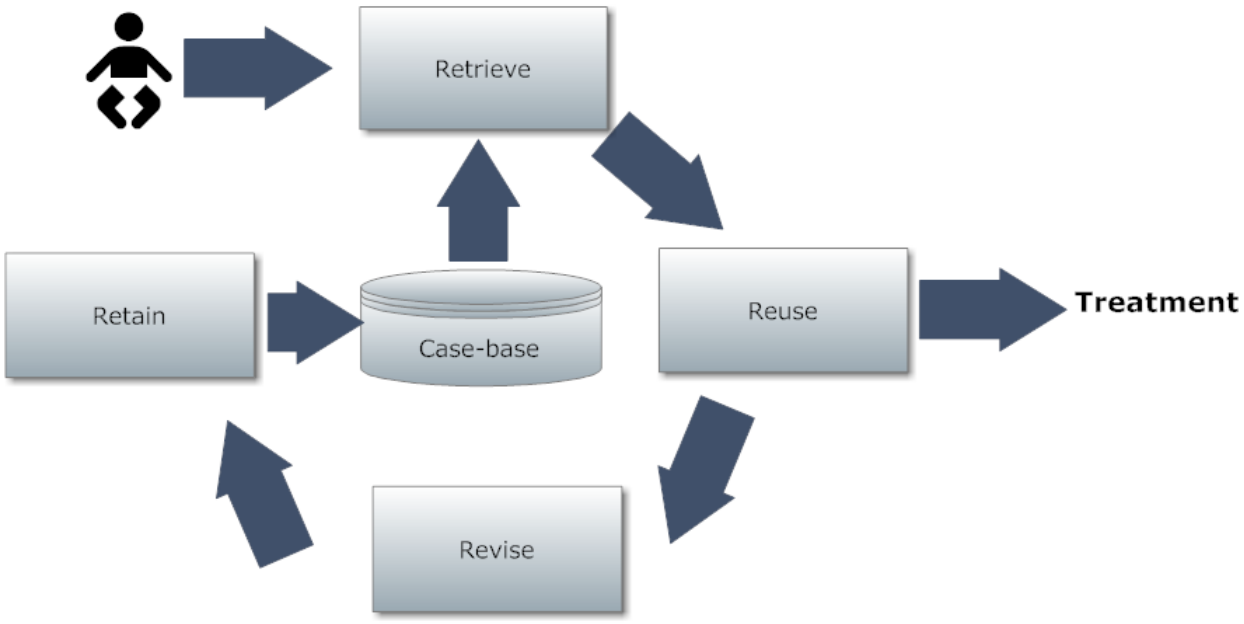

Figure 2.1: Case-base reasoning stages

retrieval process aims to retrieve the most similar previously experienced cases to the new problem. Each case is characterised by multiple features or attributes. When a new case is presented to the CBR system, the system measures the similarity between features for this query case with those of other cases in the case base in order to retrieve the most similar cases from the case base to the query. Usually, the CBR system is unlikely to retrieve an identical case, so it retrieves the most similar cases. Reusing the cases is the second process where copying or integrating the solutions from the retrieved cases is done. The revising process adapts the retrieved solution to solve the new problem. The last process is to retain the new solution once it has been confirmed or validated.

Case-based reasoning has been extensively used in the biomedical field and it is successfully applied in many applications. For instance, Fernando Diaz and et al [23] use a case-based reasoning system for cancer classification based on microarray data.

They employ a case-based reasoning model that incorporates a set of fuzzy prototypes, a growing cell structure network and a set of rules to provide an accurate diagnosis. This model is implemented and tested with microarray data belonging to bone marrow cases from 43 adult patients with cancer plus a group of six cases corresponding to healthy persons. The authors report that their case base reasoning system outperforms specific classification techniques such as Prediction Analysis of Microarrays (PAM).

Marling and Whitehouse [66] use case-based reasoning in the care of patients with Alzheimers disease. The authors report that case-based reasoning finds effective treatment by matching patients to treatments that were effective for similar patients in the past. Similarly, Lieber and Bresson [62] propose a CASIMIR/CBR system for breast cancer treatment and they claim that their system suggests appropriate solutions for new patients. Althoff et al [4] develop a knowledge-based medical decision support system using case-based reasoning for diagnosing intoxication by drugs. The authors claim that this system has potential use in the Russian Toxicology Information and Advisory Center in Moscow and in many Russian hospital ambulances.

2.1.1 Case-Based Retrieval For Gene Expression Data

The case base retrieval process is usually considered as the most important step within the case-based reasoning cycle and it gains a lot of interest from researchers [83] [65] [94]. Various techniques have been proposed to retrieve similar cases from the case base. The similarity measures used have a great influence on the performance of retrieval. The reason for this is that one of the most important assumptions in

case-based reasoning is that similar experiences can guide future reasoning, problem solving, and learning [82]. Most of today's case matching and retrieval algorithms are based on this assumption, and this leads to the development of various case clustering and classification techniques. Among these techniques, weighted feature-based similarity is the most common method used to compute relatedness among cases [55]. This technique is based on the computation of distance between cases, where the most similar case is determined by evaluation of a similarity measure.

However, simple feature-based similarity cannot be studied in an isolated manner, and there are many possible assumptions and constraints that could affect similarity measurements among cases. For example, similarity may be interpreted as relatedness among problem features (i.e., the problem space) which represents the main problem in this research project. Consequently, this thesis will deal with data mining and computational intelligence techniques to enhance the feature-based similarity for gene expression data, and to handle the complexities underlying these datasets.

Gene expression datasets are noisy with many irrelevant attributes and highly balanced classes. These characteristics raise difficulties in measuring the similarity between cases and in building an effective nearest neighbour classifier. Nearest-neighbour classifiers based on distance measures, tend to fare badly in situations where there are many features and where only a few are informative. Moreover, similarity measurement must also consider the relative importance of each feature and the effect of imbalanced classes on retrieving similar cases.

Accordingly, the main problem with a CBR system dealing with gene expression data, is the high number of attributes compared to the small number of cases. However, most of the reviewed papers deal with a large number of cases and the problem of

having a rapid case retrieval process [7]. Few papers deal with high dimensional data. In this sense, the work of Arshadi and Jurisica [7] propose a maintenance technique that integrates an ensemble of CBR classifiers with spectral clustering and logistic regression to improve the classification accuracy of CBR classifiers on (ultra) high-dimensional biological data sets. The authors apply spectral clustering followed by feature selection to pre-process the data. The authors evaluate the system on two publicly available microarray datasets that cover leukaemia and lung cancer samples. They report improvements in classification accuracy of approximately 20% from 65% to 79% for leukaemia and from 60% to 70% for lung cancer.

Stottler et al [83] propose three algorithms for rapid retrieval of cases from a case base. They claim that these algorithms can deal with up to a million cases but they didn't address the problem of the high dimensionality of the cases. Ma et al [65] propose a case retrieval algorithm based on ant colony clustering. They report improvements in the efficiency of case retrieval when there are many cases in the case base, but they didn't consider the problem of cases having many attributes.

This thesis explores problems of CBR where the case base has relatively few cases each having many features.

2.1.2 *k*-Nearest Neighbour Classifier

k nearest neighbour (kNN) [21] classifiers are based on similarity measurements and are considered as one of the most straightforward types of classifier in the machine learning techniques. This classifier is very useful and effective in situations where an explanation of the result is required as the process is transparent, easy to implement and to debug. kNN classifies query sample by identifying the k nearest neighbours

using a distance measure such as the Euclidean distance given in equation (2.1.1). Using those neighbours, the class of the query sample is determined based on the majority class among the nearest neighbours. If $\mathbf{c_p} = (f_{p1}, f_{p2}, \cdots, f_{pn})$ and $\mathbf{c_q} = (f_{q1}, f_{q2}, \cdots, f_{qn})$ are two cases in the case base, then the distance between $\mathbf{c_p}$ *and* $\mathbf{c_q}$ is defined as

$$d(\mathbf{c_p}, \mathbf{c_q}) = \sqrt{(f_{p1} - f_{q1})^2 + (f_{p2} - f_{q2})^2 + \cdots + (f_{pn} - f_{qn})^2} \tag{2.1.1}$$

Although the kNN algorithm is simple, it suffers from some disadvantages especially in gene expression datasets. These include problems dealing with irrelevant features, high dimensionality, importance of features and imbalanced classes.

k nearest neighbour (kNN) classifier is sensitive to irrelevant or redundant features because all features equally contribute to the similarity and thus to the classification. Therefore, feature selection is required in order to select the relevant features.

The second disadvantage is the effect of the high dimensionality on the distance measurement. This is known as the curse of dimensionality [56]. In order to handle this problem, dimensionality reduction algorithms are required to reduce the number of attributes of the dataset.

The third disadvantage is related to the relative importance of each feature and plays a significant role in instance-based learning algorithms such as kNN where the distance function is very sensitive to the relative importance of features. Consequently, un-weighted features may result in a weak similarity measurement and hinder the classification task. Accordingly, relevant features which play a major role in similarity measurement should be treated more importantly than other features. Based on the above scenario, weighted Euclidean distance is appropriate. Let $CB = \mathbf{c_1}, \mathbf{c_2}, \cdots, \mathbf{c_n}$

represent a case base having n cases. Each case in this memory has m features denoted as $\mathbf{c_p} = (f_{p1}, f_{p2}, \cdots, f_{pm})$ where $1 \leq j \leq m$ the number of features for each case. Each feature f_j, is assigned a weight w_j to indicate the importance of that feature. Then, for any pair of cases $\mathbf{c_p}$ and $\mathbf{c_q}$ in the case base, a weighted distance can be defined as

$$d_{pq} = d(\mathbf{c}_p - \mathbf{c}_q) = \left[\sum_{j=1}^{m} w_j (f_{pj} - f_{qj})^2\right]^{1/2} \tag{2.1.2}$$

The last disadvantage which also affects prediction is that class frequencies are imbalanced with a small number of samples in the case base [63]. Imbalanced classes affect prediction of the new query class especially if this query sample should be classified as the minority class. Another important factor is the parameter k, the number of similar patients to be retrieved. This parameter plays a major role in classifying query sample.

This thesis uses the k-nearest neighbour classifier for retrieving cases from the case base and explores the problem of the feature-based similarity in gene expression datasets.

2.2 Feature Selection

Feature selection is a critical technique in the field of bioinformatics [37] and it has been used in various domains for large and complex data such as gene expression datasets. Feature selection techniques are applied to select genes or proteins associated with a trait of interest [53] and to classify different types of samples in gene expression microarray data [2] or mass spectrometry data [1], to identify disease-associated genes and gene-gene interactions.

Feature selection provides a way for analysing high dimensional datasets such as gene expression microarray data. With this large number of genes, only a small number of them is strongly associated with the desired classification and relevant to the disease while other genes are irrelevant [78]. Consequently, feature selection becomes a big challenge and represents a real prerequisite for selecting the most informative genes from the original data.

Four benefits can be achieved in feature selection. The first one is removing the noise and irrelevant features which affect similarity measurement. Second, reducing the effect of the curse of dimensionality and enhancing the quality of dimensionality reduction algorithms. The third advantage is increasing the speed of learning algorithms such as classification, similarity measurement and prediction. The last benefit is in providing better understanding and interpretation for biologists and assisting them in identifying and monitoring the target disease or function types [25].

2.2.1 Feature Selection Techniques

Saeys et al [78] define a taxonomy of feature selection techniques which divides the feature selection into three approaches: filter methods, wrapper methods and embedded methods.

Filter based algorithms are independent of the classifiers. That is, the gene selection process and the classification process are separated. Examples are χ^2-statistics [92], t-statistic [43], RelieF [76], Information Gain [85]. With the filter approach, genes are selected based on an evaluation criterion regardless of the classifier as shown in Figure 2.2(a). The advantages of these algorithms are that they are fast, scalable and the selected genes generalize to unseen data [96]. However, the disadvantages are

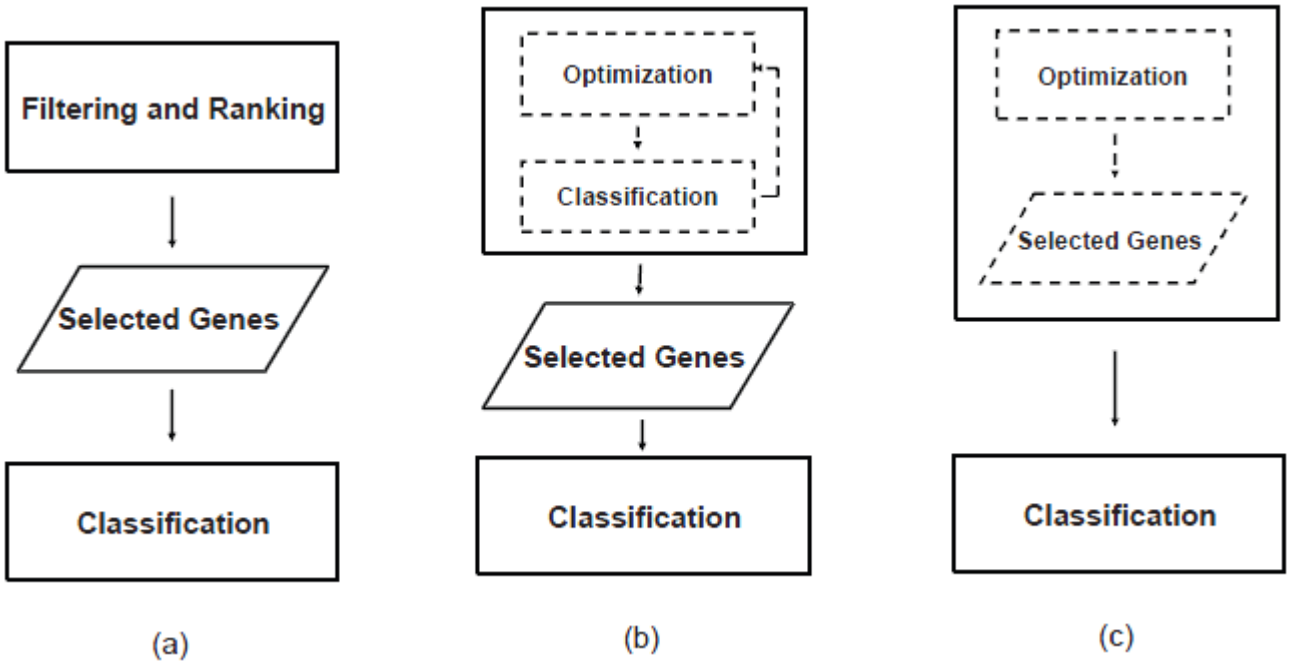

Figure 2.2: Selection methods. (a) Filter approach (b) Wrapper approach (c) Embedded approach [78]

that they ignore of feature interactions. The univariate filter method treats genes as independent of one another. This drawback of uni-variate filter methods can lead to poor classification performance. Thus, a multivariate filter methods were introduced to handle this problem but not fully. Another disadvantage of filter methods is that they ignore interaction to the classifier which also may lead to a loss of information for features that play a good role in discrimination [48].

The wrapper approach, on the other hand, selects a gene subset at each step and evaluates it based on the classification performance [48], as shown in Figure 2.2(b). This interaction between the features and the classifier is considered as an advantage of the wrapper approach in addition to consideration of correlation between features. However, the disadvantages of these approaches are that they have a high risk of over fitting and are computationally intensive. Examples includes the use of genetic algorithms, sequential forward selection and sequential backward elimination. Evolutionary based algorithms such as the Genetic Algorithm (GA) [40] have been introduced as wrapper algorithms for the analysis of microarray datasets [59] [71] [45]. Unlike classical wrappers which select genes incrementally [42], GA selects genes non-linearly by creating gene subsets randomly. Nevertheless, wrapper approaches like GA have long been criticized for suffering from over fitting [78]. Moreover, the use of a given inductive algorithm as the sole optimization guide often leads the system to pursue high classification accuracy on training data at the expense of poor generalization on unseen data.

Embedded methods, the third approach to feature selection, consider the interaction between the classifier and feature selection but with less computational cost and with

a lower risk of over-fitting. As illustrated in Figure 2.2(c), feature selection is a by-product of the classification process. Examples include classification trees such as ID3 [74] and C4.5 [75]. However, the drawback of embedded methods is that they are often greedy search-based algorithm [13], using only the top ranked features to perform sample classification in each step while an alternative split may perform better. Furthermore, additional steps are needed to extract the selected features from the embedded algorithms.

Feature selection in an essential approach in this thesis and it is important to identify a selection method suited to gene expression data and case base retrieval.

2.2.2 Adaptation of Ensemble Methods in Feature Selection

Ensemble methods [24] are effective learning techniques that have been introduced to improve overall prediction and feature selection accuracy. In this approach, several feature selection methods are combined to make the decision for elimination of features instead of choosing one particular classifier. Moreover, ensemble techniques have the advantage of handling the problems of small sample size and high-dimensionality associated with gene expression datasets. They also reduce the potential of over fitting the training data. With ensemble methods, the training dataset may be used more efficiently. Some ensemble methods such as random forests are particularly useful in this thesis as they deal with very high-dimensional datasets. Higher classification accuracy can be achieved by generating multiple prediction models each with a different feature subset. Moreover, ensemble methods are effective in dealing with data with small sample sizes such as gene expression data.

2.2.3 Random Forest For Feature Selection

The random forest algorithm was developed by Breiman [16], and is known as one of the most robust classification algorithms developed to date. It is an ensemble classifier consisting of many decision trees. Many classification trees are grown during training. A training set is created for each tree by random sampling with replacement from the original dataset. During the construction of each tree about one-third of the cases are left out of the selection and this becomes the out-of-bag cases which are used as a test set. The classification performance of the test set is evaluated based on the out-of-bag error rates.

Random forest has been used extensively in the biomedical domain [23] [6] because it is well suited for microarray data. Features will not be deleted based on one decision or one tree, but many trees will decide and confirm elimination of features. Another positive characteristic of random forest is that it is applicable to very high dimensional data with a low number of observations, a large amount of noise and high correlated variables. Moreover, random forest can handle the problem of imbalanced classes. All these characteristics make this algorithm an appropriate choice for gene expression datasets.

Diaz-Uriarte and Alvarez de Andres [23] applied random forests on a number of different microarray datasets and compare performance with many other algorithms such as k-nearest neighbours and support vector machine. They report that the classification accuracy of random forests is similar to those of the best algorithms that are already in use. Moreover, they explore the potential of random forest for attribute selection and propose a method for gene selection using the out-of-bag error rates.

The authors state that because of the algorithm's ability to perform well as a classifier while also allowing for excellent attribute selection, this algorithm should become an essential part of the tool-box for prediction and gene selection with microarray data. Archer and Kimes [6] perform a similar evaluation of the random forests algorithm and came to many of the same conclusions about its effectiveness of the algorithm. They apply the algorithm to Acute Lymphoblastic Leukaemia microarray data as a means of trying to discover the genes which are responsible for the difference between subtypes of the disease.

2.2.3.1 Algorithm of Random Forest

Random forest consists of four main steps

1. Choose t trees to grow.
2. Choose m potential variables as candidate to split each node, where $m << M$, and M is the number of input variables. The parameter m is hold constant while growing the forest.
3. Grow t trees. When growing each tree the following is done.
   (a) Construct a bootstrap sample of size n with replacement and grow a tree from this bootstrap sample.
   (b) When growing a tree at each node select m variables at random and use them to find the best split.
   (c) Grow the tree to the maximal extent.
4. To classify a point $\mathbf{x}$ collect votes from every tree in the forest and then use majority voting to decide on the class label.

2.3 Dimensionality Reduction

Dimensionality reduction is a significant technology in this thesis. In contrast to feature selection, dimensionality reduction aims to extract new features from the original set of features. The purpose of dimensionality reduction is to reduce, understand and visualize the structure of complex datasets with the high dimensionality and a low number of observations found in microarray data. Analysis and visualization is difficult in practice and becomes an obstacle for diagnosis and treatment of patients [47]. Lee and Verleysen [56] propose a taxonomy of dimensionality reduction. There are two main types of dimensionality reduction: linear dimension reduction such as Principal Component Analysis, and Non-Linear Dimensionality Reduction (NLDR) such as Sammon mapping [80], locally linear embedding [77], and ISOMAP [86]. There are two main approaches for NLDR: methods to preserve distances between the high and low dimensional spaces and methods to preserve the topology (or neighbourhoods). Methods can be further subdivided into those using Euclidian distance like Multi dimensional scaling [50] and those using geodesic distance (graph distance) such as ISOMAP. They identified two types of topology preservation methods, those using a predefined lattice(e.g Sammons mapping) and those using model-driven lattice (e.g locally linear embedding).

2.3.1 Linear Dimensionality Reduction

Linear dimensionality reduction approaches are the traditional methods for dimensionality reduction. However, their effectiveness is limited by its global linearity. Principal Component Analysis (PCA), introduced by Pearson [32], is one of the oldest and best known methods in data mining and analysis. This method is characterized

by its simplicity and it is a non-parametric method. It aims to extract the latent variables from a high dimensional dataset. It can act as a step to reduce complex datasets in a very dimensional space into a lower dimension in order to reveal the latent variables and simplified structures that often underlie it [56]. The goal of PCA is to find the directions or components where the data has maximum variance. This is achieved by finding the eigenvalues and eigenvectors of the covariance matrix of the data.

Several mathematical and statistical steps are followed in order to solve the eigenvector problem. The first step is to center the original data by subtracting the mean of each attribute. The next step is to calculate the covariance matrix which is the $d \times d$ matrix where d is the dimensionality of the data. Then the eigenvalues and eigenvectors of the covariance matrix are calculated by solving the following equation.

$$A - \lambda I = 0 \tag{2.3.1}$$

where A is the covariance matrix, I is an identity matrix and λ is the eigenvalues and eigenvectors of the covariance matrix.

After determining the eigenvalues and corresponding eigenvectors, the eigenvectors are sorted in decreasing order based on their eigenvalues. The highest eigenvalue represents the first principal component and so on. If the centred data is projected onto the first principal component, the dimensionality of the original data is reduced into just one dimension. With this simple dimensionality reduction algorithm, the question is how many principal components should be selected.

Principal Component Analysis has been successfully used in many applications in data

mining and pattern recognition. One paper has been published where PCA was used in graph mining in order to reduce the dimensionality and identify a smaller number of characteristic patterns [79]. Another paper used PCA for face recognition together with a radial basis function neural network. The role of PCA here was not only for dimensionality reduction of the image, but also to retain some of the variations in the image data [87]. Another paper used PCA and fractional-step linear discriminate analysis for face recognition. PCA successfully reduced the computational complexity of the algorithm and the dimensionality of the data [64]. Principal component analysis is an important tool in this thesis and is used to reduce the dimensionality of the dataset.

2.3.2 Non-Linear Dimensionality Reduction

This area has received increasing attention for researchers due to the increasing number of applications in the field of microarray data. Non-linear dimensionality reduction techniques have been regularly used for visualization of high-dimensional datasets [22]. Many non-linear techniques for dimensionality reduction have been proposed since 2000 including ISOMAP [86] and Locally Linear Embedding (LLE) [77]. These methods try to preserve the local neighbourhood of each point for the high dimensional space in the low dimensional space. They have been used for visualization of high dimensional data and perform well when the data is composed of one cluster and fail when the data has multi clusters [90]. Many papers have been published using non-linear dimensionality reduction. Cho and Park [20] use ISOMAP for NLDR with some additions to the base algorithm in order to take into consideration the class information when the data is clustered. Another paper proposed

a continuous ISOMAP that is able to compute the low dimension for out-of sample points [97]. That is it can project data points unseen by the original algorithm into the lower dimensional space. A neural network-based ISOMAP algorithm was proposed to efficiently achieve a robust and stable ISOMAP [17]. The benefits of the proposed method are that the time complexity is linear and the space complexity is constant. Another paper modifies LLE with the addition of a kernel function in bioinformatics [60]. It aims to reduce the dimensionality of a Lymphoma data set in order to classify samples as cancer or not through the analysis of microarray gene expression. The author claims an improved result with the kernel extension compared to the original LLE. Another paper uses LLE for text dimensionality reduction [39] and demonstrates the computational time of LLE is less compared to other methods. LLE has also been used for speech visualization where it enables deaf people to differentiate between different speech to improve and train their oral ability [95].

In contrast to traditional linear techniques, the non-linear techniques have the ability to deal with complex non-linear data such as microarray data. However and also in contrast to PCA, non-linear techniques have the problem of dealing with out-of-sample point [11]. That is, a new data point cannot be embedded into an existing low-dimensional data. In this thesis, non-linear dimensionality reduction algorithms will be used only for visualization purposes.

2.3.3 Trustworthiness

The Trustworthiness measurement is proposed by Kaski et al [46] for error estimation and quality measurement of non linear dimensionality reduction algorithms. One approach for evaluating dimensionality reduction methods is to compare neighbours in

input and output dimensional spaces. The Trustworthiness measurement compares the neighbourhood of the points in the input and output spaces. For example if point $\boldsymbol{x}$ is close to points $\boldsymbol{w}$ and $\boldsymbol{z}$ in the input space X, then point $\boldsymbol{x}$ should be also close to points $\boldsymbol{w}$ and $\boldsymbol{z}$ in the output reduced space Y otherwise an error arises after reduction. For example, in Figure 2.3, there is a point $\mathbf{x}_i$ in the input space (Figure 2.3a) with nearest neighbours points represented in red (point $\boldsymbol{w}$ and $\boldsymbol{z}$). These points are transformed and mapped into lower dimensional space. The point $\mathbf{x}_i$ transformed into another point $\mathbf{y}_i$. The red points are still the nearest neighbors for the point $\mathbf{y}_i$ except for point $\boldsymbol{z}$ which is lost from the nearest neighbors. On the other hand, a blue point becomes a new nearest point for the point $\mathbf{y}_i$ when it was not in the input space. In this case, we do not have complete trustworthiness because the neighborhood of the point $\mathbf{x}_i$ has changed between the input and output spaces.

Trustworthiness finds to which extent neighbors in the input space also have corre-

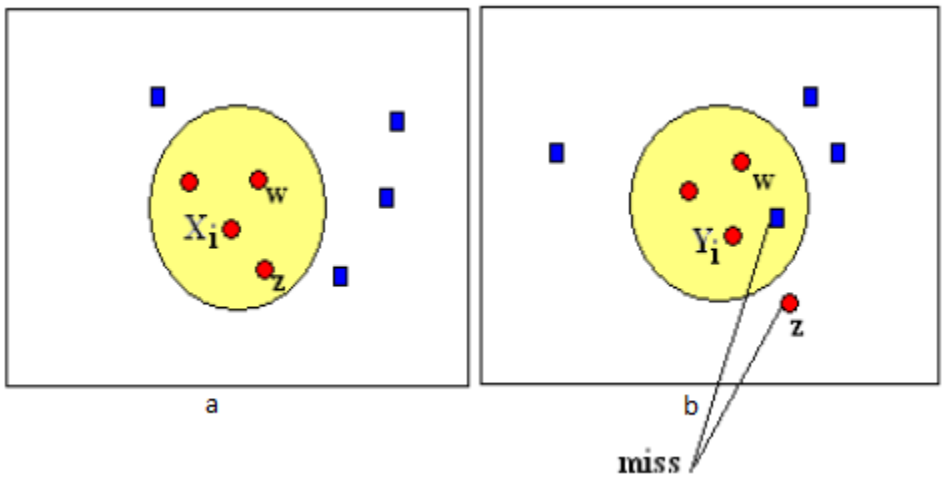

Figure 2.3: Neighbours problem in dimensionality reduction: a) high dimensional space and b) low dimensional space

sponding neighbors in the output space by ranking the neighborhoods of point sets

in the input and output spaces. Trustworthiness can be defined as follows: let N be the number of data samples and $r(i,j)$ be the rank of the data sample j in the ordering according to the distance from i in the original data space. Denote by $U_k(i)$ the set of those data samples that are in the neighborhoods of size k_{trust} of the sample i in the visualization display but not in the original data space [89]. The measure of trustworthiness $M_{trust}(k)$ of the dimensionality reduction is

$$M_{trust}(k_{trust}) = 1 - A(K_{trust}) \sum_{i=1}^{N} \sum_{j \epsilon U_{K_{trust}}(i)} (r(i,j) - K_{trust}) \tag{2.3.2}$$

Trustworthiness measure is important in this thesis because it provides error estimation and quality measurement for dimensionality reduction algorithms.

2.4 Genetic Algorithms

Genetic algorithm [40] is a global search technique used to solve optimization problems which do not already have a well defined efficient solution. Genetic algorithms (GA) have received a great deal of attention regarding their potential as optimization techniques for optimization problems and are often used to solve real-world problems [70]. Genetic algorithms have been applied in many applications with multi objectives, such as feature weighting [44], transportation problems [33], the minimum spanning tree problems [100] and the production process planning problems [99]. As a result, GA is an important tool in this thesis and can be used to weight the features. Four main steps are involved in GA: evaluation, selection, crossover and mutation. Figure 2.4 shows the different processes of GA.

1. The evaluation procedure measures the fitness of each individual solution in the

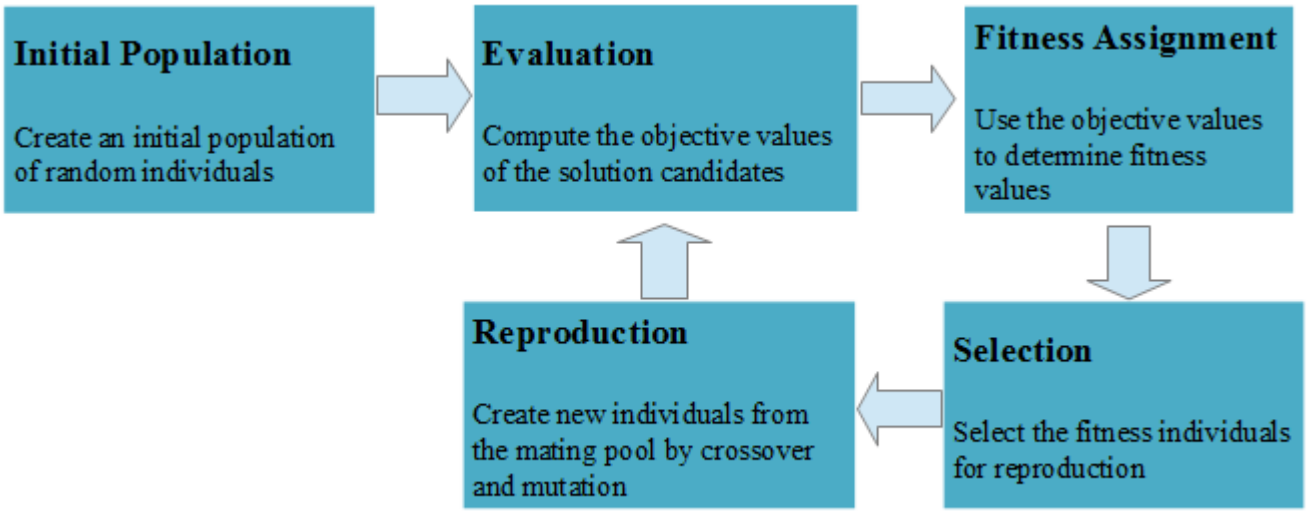

Figure 2.4: Genetic algorithm process based on [70]

population and assigns it a relative value based on the defining optimisation (or search) criteria. Typically in a non-linear programming scenario, this measure reflects the objective value of the given model.

2. The selection procedure randomly selects individuals of the current population for development of the next generation. Various methods have been proposed but all follow the idea that the fittest individuals have a greater chance of survival.

3. The crossover procedure takes two selected individuals and combines them about a crossover point thereby creating two new individuals. Simple (asexual) reproduction can also occur which replicates a single individual into the new population.

4. The mutation procedure randomly modifies the genes of an individual subject by a small mutation factor, introducing further randomness into the population.

2.5 Imbalanced Data

Many gene expression datasets have the imbalanced classes problem. That is, at least one of the classes constitutes only a very small minority of the data. For such problems, the effect is on the practical classification where the interest usually leans towards correct classification of the minor class. Generally, most of the classifiers used in selecting features suffer from the imbalanced classes and many have poor performance because they are biased to the large samples and pay less attention to the rare class. These algorithms do not work well for such problems because they aim to minimize the overall error rate. Consequently, unsatisfactory classification performance results and most of the rare class features are not recognized. Several researchers have tried to address the problem in many applications such as fraudulent telephone call detection [31], information retrieval and filtering [57], diagnosis of rare thyroid deceases [12] and detection of oil spills from satellite images [51].
The main dataset of this thesis is imbalanced data and has a small number of observations. Therefore, ignoring this critical characteristic of the data may result in very poor feature selection and case retrieval.

2.5.1 Approaches to Handle Imbalanced Classes

Two different approaches have been proposed to overcome this problem [27]. One is based on a sampling technique: either down-sampling the majority class or over-sampling the minority class, or both. The other approach is to use cost sensitive learning: assigning a high cost to misclassification of the minority class, and trying to minimize the overall cost [101].

2.5.2 Sampling Technique

This approach alters the distribution of the classes to be more balanced. This can be done either by over sampling or by down sampling and sometimes both [51].
Down sampling removes a number of observations from the majority class. It aims to attain the same number of samples in the majority class as in the minority class. SHRINC, proposed by [51], is used for down sampling by reducing the number of samples of the majority class. This method has some drawbacks because it can eliminate some useful information. Moreover, as was shown in the above discussion about the gene expression data set, a very small number of observations are available in the gene expression data. Consequently, elimination of observations is not justified with this kind of data.

On the other hand, over sampling increases the number of samples in the minority class by replicating minority samples so that they reach the same number as in the majority class. The synthetic minority over sampling technique (SMOTE) [18] is an approach used to form new minority class examples by interpolating between several minority classes examples that lie together. This method is based on the k-nearest neighbours and it is based on the distance measure. Consequently, it may not work well on a very high dimensional datasets like gene expression datasets due to the curse of dimensionality.
In this thesis, oversampling technique will be adopted in order to increase the number of samples in the minority class and reduce the effect of imbalanced classes.

2.5.3 Cost Sensitive Learning

The second approach to tackle the problem of imbalanced data is cost sensitive learning [26] [72] [30]. This approach assigns a high cost to misclassification of the minority class and low cost to misclassification of the majority class. Based on the above discussion about sampling technique, this method is more suitable to apply on the gene expression datasets for learning extremely imbalanced data. As random forest produces votes from each generated tree in the forest and these votes are then used to classify a new point, cost sensitive learning will be used in this thesis and applied on these votes by assigning a weight for each class. A higher misclassification error cost is given to the minority class by assigning a larger weight. These weights will play a role in calculating the votes at the terminal nodes. The weighted vote of a class is the weight for that class multiplied by the number of cases for that class.

2.6 Evaluation Measures of Imbalanced Classes

As gene expression datasets often have the problem of imbalanced classes, the minor class can have very little impact on the accuracy measurement. This leads to a fact that the traditional accuracy measure may be not adequate for extremely imbalanced classes. Other measures have been proposed for imbalanced class classification evaluation. Most of these measurement rules are based on the data obtained from the confusion matrix. Confusion matrix represents the outcome of the actual predicted classification obtained by the classifier. Table 2.1 illustrates a confusion matrix for a two-class classification. Conventionally, the majority class is represented as a negative class label and the minority class is represented as a positive class label. The actual

class label of the examples is presented in the first column, and their predicted class label presented in the first row. The numbers of positive and negative examples that are classified correctly are denoted by TP and TN, while the misclassified examples are denoted by FN and FP. In order to measure the performance of the classifier, the

Table 2.1: A confusion matrix for a two-class classification

	Predicted positive	Predicted negative
Actual positive	True Positive (TP)	False Negative (FN)
Actual negative	False Positive (FP)	True Negative (TN)

accuracy should be determined for each class separately. That is, the true negative rate and the true positive rate shown in equations (2.6.1) and (2.6.2) respectively. These two estimated values also known as Recall.

$$Recall^{+} = TrueNegativeRate = \frac{TN}{TN + FP} \tag{2.6.1}$$

$$Recall^{-} = FalsePositiveRate = \frac{TP}{TP + FN} \tag{2.6.2}$$

Another interesting measure is the precision of the predicted positive and negative cases which is the proportion of the correctly predicted positive cases and negative cases, respectively. This can be determined using equations (2.6.3) and (2.6.4).

$$Precision^{+} = \frac{TP}{TP + FP} \tag{2.6.3}$$

$$Precision^{-} = \frac{TN}{TN + FN} \tag{2.6.4}$$

Another evaluation measure that works well with imbalanced dataset is the Receiver Operating Characteristic (ROC) curve [14]. Receiver Operating Characteristic curve presents the trade-off between the true positive rate and the false positive rate. A ROC curve is considered to be good if it is closer to the top left corner, and the

straight line connecting (0,0) and (1,1) represents a random classifier with even odds. The area under the ROC curve (AUC) is often used to summarise a classifier performance into a single quantity, which represents the performance of a learner in general across different prediction costs, and the larger the AUC the better. These presented evaluation measures will be adopted in this thesis in order to evaluate the performance of the classification.

2.7 Summary

Many researchers have applied case base reasoning to biomedical applications. However, the case retrieval for gene expression data is still an active area of research. Moreover, feature-based similarity, in the microarray domain, still requires a lot of research studies to be done in order to have an accurate similarity measurement between patients due to the complexities that underlie gene expression datasets.

Feature selection has been extensively used in the microarray domain. However, the complexities of our dataset, obtained from The Children's Hospital at Westmead, requires development of a new algorithm for feature selection. As can be seen from this review, each feature selection method has advantages and disadvantages and no one algorithm is universally optimal. However, for feature selection, random forest has been shown to be well suited for microarray data. With respect to dimensionality reduction, although NLDR algorithms are designed to handle complex datasets, principal component analysis is preferable to be used in this thesis because the out-of-sample problem does not exist and new points can be accurately embedded in the existing low dimensional space. For feature weighting, genetic algorithms are effective

global search technique and have been extensively used to solve optimization problems. The imbalanced classes problem has been solved with two approaches which will be both used in this thesis in different situations.

Chapter 3

Case-Base Retrieval Framework For Gene Expression Datasets

This chapter presents a case-base retrieval framework for similarity measurement and case-base retrieval for gene expression datasets. The rest of this chapter is organised as follows. Section 3.1 presents the effects of irrelevant features, high dimensional space and low numbers of observations on the distance measurement and case retrieval process in gene expression data. Section 3.2 introduces a high level view of the case-based retrieval framework to show the different modules of the framework before going to discuss in detail about each module of the framework.

3.1 Case Base Retrieval in Gene Expression Datasets

The process of retrieving similar cases from a case base to a query case is regarded as a primary and fundamental step in case-based reasoning (CBR) [54] [65] [94], and the similarity measurement between cases plays a very important role in this process. Case-based reasoning systems are unable to function properly without employing effective case retrieval to retrieve the similar cases from the case base. As

a result, similarity measurement has a strong influence on the case retrieval process and has great value in providing better solutions for new problems. The most widely used methods for similarity measures are distance-based functions which calculate the distance between cases using some or all of the attributes constituting the cases. However, distance measurement is not applicable on cases with many attributes such as gene expression datasets. The reason for this is that distance measurement is very sensitive to irrelevant or redundant features because all features equally contribute to the similarity. As a result, similarity measurement is considered as a big challenge with gene expression datasets which contain a massive number of attributes for each case. With this kind of data, pre-processing steps such as feature selection, dimensionality reduction and feature weighting are required to prepare the data for a useful similarity measurement task.

Feature selection and dimensionality reduction are two essential techniques in the microarray domain. In feature selection, a subset of features that are strongly associated with a specific trait of interest are selected. It acts as an initial step for cleaning gene expression datasets by selecting the relevant features, enhancing the quality of machine learning, and providing a better understanding of gene expression datasets. In some cases, feature selection might not be enough on its own to reduce high dimensional datasets as the data may still exist in a high dimensional space. Accordingly, dimensionality reduction methods are required in the microarray domain and will also be involved in the case base retrieval process. In contrast to feature selection, dimensionality reduction aims to extract new features from the original set of features using some linear or non-linear mapping methods.

Feature weighting, which aims to find the most important genes by determining how

useful they are in distinguishing categories, is also required in similarity measurement. The reason for this is because feature-based similarity, which is used in this thesis, is very sensitive to the quality of features. Feature-based similarity suffers from the equal treatment of features where some features may be more significant than others and play a major role in discriminating categories. These features should be treated more importantly than other features and they should attract high weights. On the other hand, low weights are given to low impact features. Accordingly, feature weighting is required in order to enhance the calculation of the similarity between patients and to achieve more accurate results.

Furthermore, gene expression microarray datasets often have a low number of observations compared to the large number of genes. Oversampling techniques are used in this thesis in order to increase the number of samples of the minority classes. This leads to improvements in the classification and retrieval performance of test samples in the minority class. Moreover, oversampling techniques also reduce the effect of imbalanced classes.

Consequently, the acquisition of a huge number of genes with the complexities underlying gene expression datasets has given rise to the development of a framework for case base retrieval process. The motivation of building this framework is to process these kinds of datasets before measuring the similarity between patients.

3.2 Case-Base Retrieval Framework

Based on the characteristics and problems of gene expression datasets, this thesis presents a novel framework that involves several computational intelligence techniques. This framework is composed of two modules: module 1 for training processes

and module 2 for test processes (see Figure 3.1).

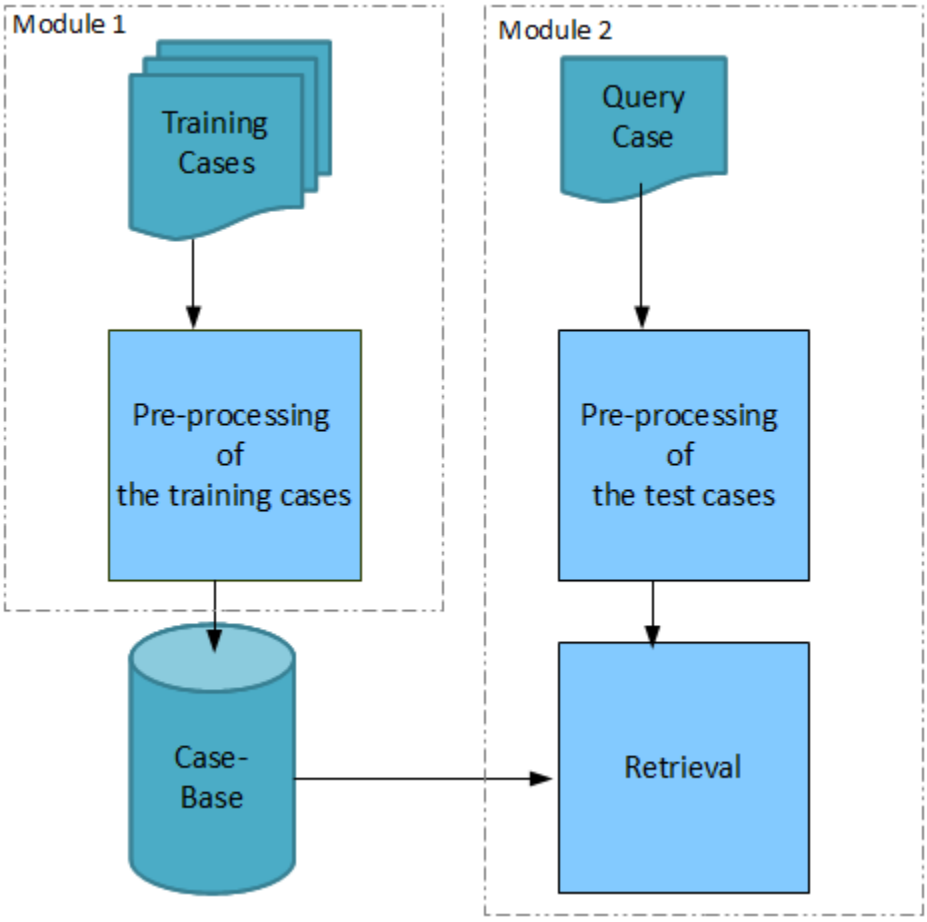

Figure 3.1: Case-base retrieval framework

Module 1 concerns the pre-processing steps of the training dataset. The purpose of this module is to pre-process the training dataset and handling the complexities of the gene expression datasets such as feature selection, dimensionality reduction and feature weighting.

Module 2, on the other hand, concerns a test process for a new query case. The purpose of this module is to use the outputs of the training process such as a list

of the selected features in order to pre-process the query before retrieving, from the case base, the similar previous cases. Detailed descriptions of each module in the case-base retrieval framework are given in the next sections.

3.2.1 Module 1: Pre-processing the Training Dataset

The first module of the case base retrieval framework pre-processes the training dataset and is composed of four steps: feature selection, dimensionality reduction and visualisation, feature weighting and oversampling (see figure 3.2). The first step aims to select a subset of genes that represent the most relevant features to a specific output. The second step applies dimensionality reduction algorithms on the dataset in order to reduce the impact of the curse of dimensionality on the similarity measure and to visualise the dataset. Once the dataset is processed by the dimensionality reduction algorithm and is transformed to a lower dimensional space, the dataset is presented to the feature weighting process to identify the relative importance of each feature for similarity and classification optimization. The last step applies an oversampling technique to increase the number of samples of the minority classes and to reduce the effect of the imbalanced classes. Figure 3.2 shows the steps of training process.

- Step 1.1: Feature Selection

  Feature selection algorithms are initially applied to the dataset. The training cases are presented to feature selection algorithms to select relevant features. The output of this step is the training dataset with a subset of genes that are strongly associated with the output of interest. The selected genes are stored

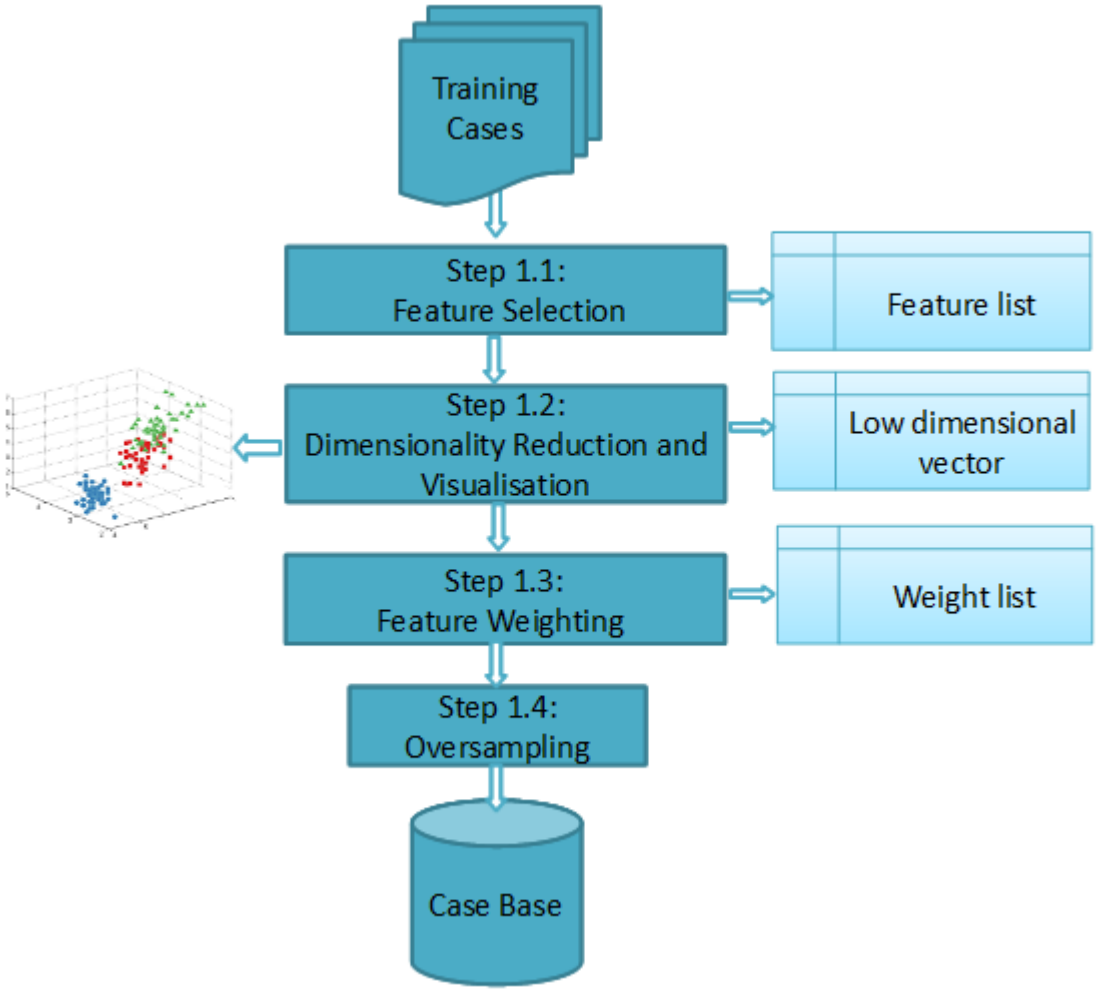

Figure 3.2: Pre-processing the training dataset in the framework

as a list of genes, labelled "Feature list" in Figure 3.2, to be used later for processing the test data samples.

Detailed description of this step such as the proposed algorithm, validation and number of selected attributes are given in chapter 4 "Feature Selection"

- Step 1.2: Dimensionality Reduction and Visualisation

  Linear dimensionality reduction algorithms are applied after feature selection in order to reduce the dimensionality of the dataset. Although the features of the

dataset are reduced after removing the irrelevant features, the dataset may still exist in a high dimensional space and similarity measurement may still suffer from the curse of dimensionality [56]. This transformation of the dataset into a lower dimensional form is required in order to enhance the distance measurement. Kernel Principal Component Analysis (KPCA) is applied to the training set samples. The output of this step is the training dataset in a low dimensional space and a vector, labelled "Low dimensional vector" in Figure 3.2, to be used later in processing the test data samples.

This step involves another procedure that aims to visualise the gene expression dataset. Biologists and clinicians may be able to better understand the structure of complex microarray datasets when visualised in two or three dimensions. Figure 3.3 shows the feature selection, dimensionality reduction and visualisation steps in the case base retrieval framework and how the dimensionality of the dataset is modified from one step to another. Further description of the number of selected components in KPCA and the validation process are given in Chapter 5 "Dimensionality Reduction and Visualisation".

- Step 1.3: Feature Weighting

  Once the training dataset is projected to a low dimensional space, the next step is to weight the features. The output of this step is a weight vector which is stored as a list of weights, labelled the "Weight list" in Figure 3.2, to be used in the distance measurement formula. Feature weighting is needed for instance-based learning algorithms such as nearest neighbourhood. Giving weights to the features based on their quality and usefulness has the potential to lead to more accurate distance measurement. Genetic algorithms are involved in

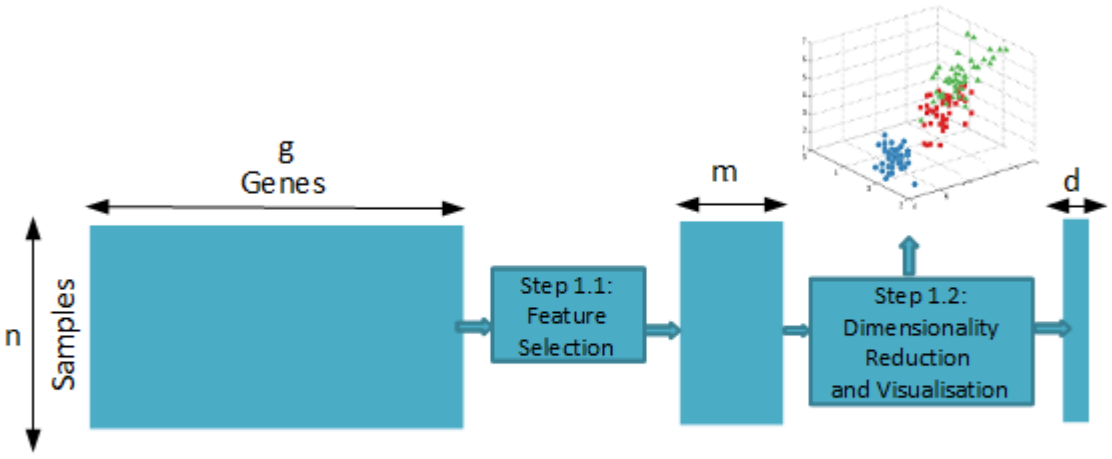

Figure 3.3: Modification of the data in the dimensionality reduction and visualisation section

this step in order to find the relative importance of each feature and assign it a corresponding weight. Detailed description of this step and the proposed technique are given in Chapter 6 "Similarity Measurement"

- Step 1.4: Oversampling

  Oversampling techniques are involved in this step to increase the number of samples in the gene expression dataset. Synthetic minority oversample technique (SMOTE) [19], which is considered as an intelligent oversampling method, is applied in this framework in order to add new artificial minority examples by interpolating between original minority class examples that lie together rather than duplicating the pre-existing examples. Figure 3.4 shows the feature weighting and oversampling steps in the case base retrieval framework. Further description about this step are given in Chapter 6 "Similarity Measurement".

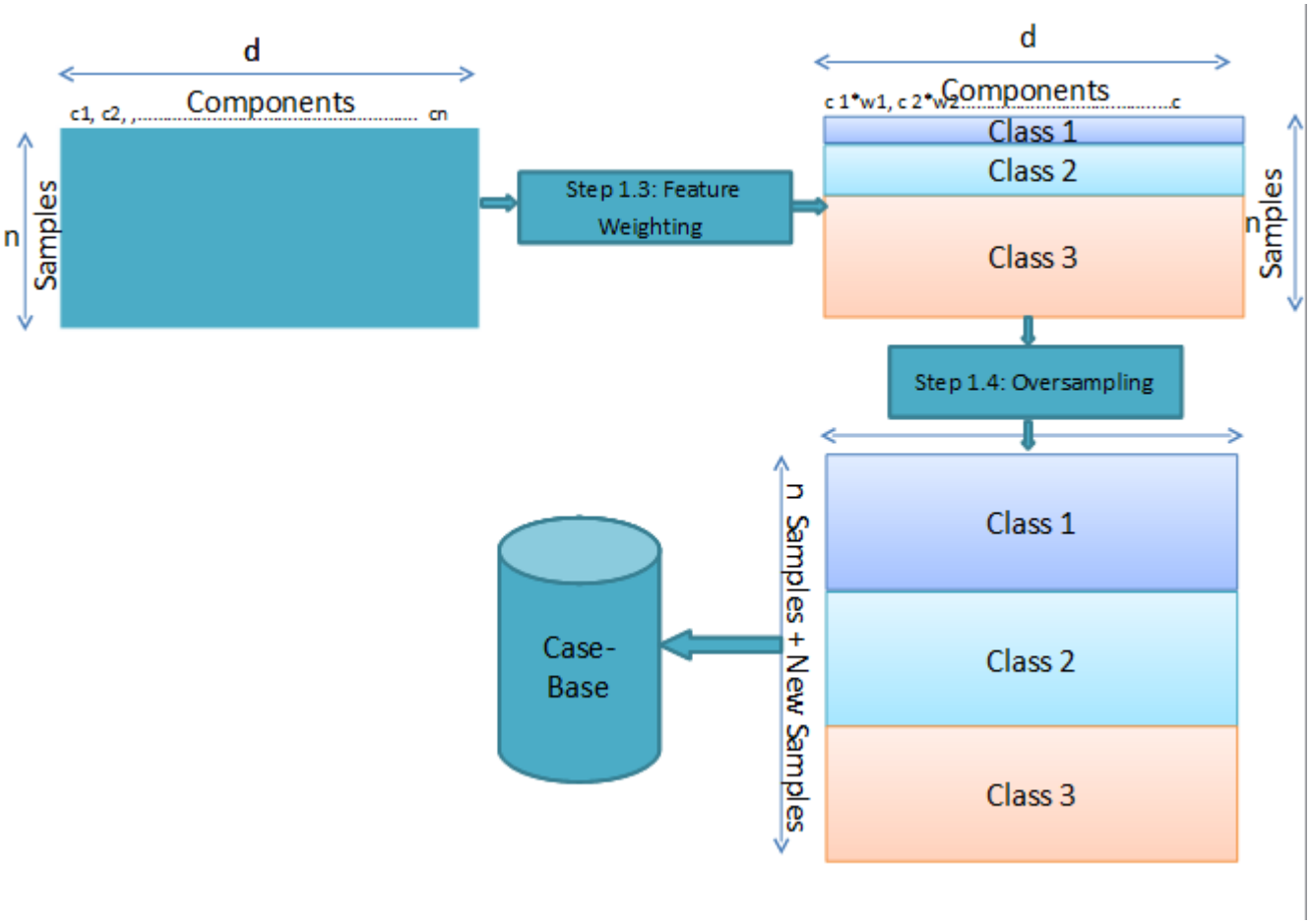

Figure 3.4: Modification of the data in the feature weighting and sampling techniques section

3.2.2 Module 2: Pre-processing the test dataset

The second module of the framework is related to the query samples (see Figure 3.5). A new query sample comes in the high dimensional space with irrelevant features. These features are filtered based on the determined relevant features obtained from the feature selection step 1.1 of module 1. The next step is to transform the new sample into a lower dimensional space by projecting the filtered features onto the dimensionality reduction vector obtained from the dimensionality reduction step 1.2

of module 1. Once the new sample is passed through the pre-processing steps of the

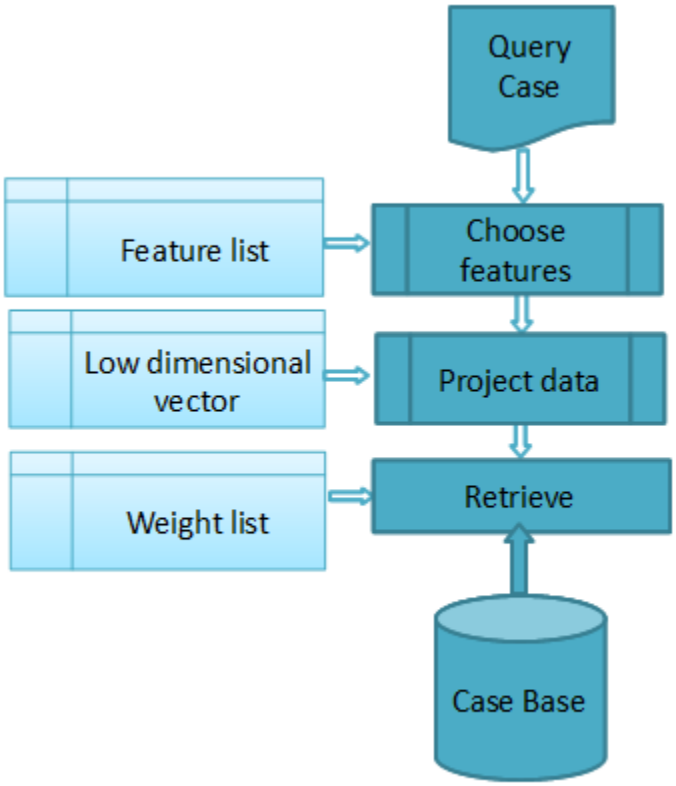

Figure 3.5: Pre-processing the test sample in the framework and retrieving the similar cases

test samples, the new sample is presented to the retrieving process in order to retrieve the old similar cases from the case base using the feature's weight obtained from the feature weighting step 1.3 of module 1.

3.3 Summary

A case base retrieval framework for gene expression data is presented in this chapter. This framework is contribution 1 of this thesis listed in section 1.3. It provides orientation to the important technologies of this research and the relationships among these technologies. The process of developing this conceptual framework brought out a number of significant realizations: that feature selection and linear dimensionality reduction have several positive effects on the gene expression datasets in terms of alleviating the curse of dimensionality and enhancing the similarity measurement; that the effectiveness of weighting the features is important in the distance measurement; and that one improvement in case base retrieval process is adopted by applying oversampling techniques on gene expression datasets.

The methods and steps forming part of the framework are described in more detail and validated in later chapters.

Chapter 4

Balanced Iterative Random Forest for Gene Selection from Microarray Data

This chapter proposes a feature selection method called Balanced Iterative Random Forest (BIRF) to select genes that are relevant to a specific trait of interest from gene expression datasets. Too many features or genes in a dataset adversely affect similarity measurement and classification performance as many of these genes are irrelevant to specific trait of interest [38]. Moreover, the huge number of features leads to the problem of high dimensionality which in turn affects the distance measurement. Hence, selection of a subset of genes that are relevant to a trait of interest is crucial and plays a vital role for building a successful gene expression similarity measurement model.

Based on the literature review of feature selection approaches presented in chapter 2 and the characteristics of gene expression datasets, there is no one algorithm or technique that is optimal for feature selection of gene expression datasets. However, random forest [16] has been found to handle most of the problems that accompany these kinds of datasets. Random forest is an ensemble classifier which can work with

high-dimensional data with a low number of observations, can handle the correlation between features and is robust on noise. The rest of this chapter is organised as follows.

Section 4.1 introduces the Balanced Iterative Random Forest algorithm with the pseudocode and the methods to handle the imbalanced classes and approach to validate. Section 4.2 describes datasets used for validations. Section 4.3 describes experiments validating the approach and results achieved on different gene expression datasets. Section 4.4 compares the proposed algorithm with other methods. Finally, section 4.5 gives a brief conclusion to summarize and bring together the main areas covered in this chapter.

4.1 Balanced Iterative Random Forest for Feature Selection

This thesis introduces a new method for feature selection based on random forest called Balanced Iterative Random Forest (BIRF). Balanced Iterative random forest is an embedded gene selector that follows a backward elimination approach. The base learning algorithm is random forest which is involved in the process of determining which features are removed at each step. The algorithm starts with the entire set of features in the dataset. At every iteration, the number of the attributes is reduced by removing those attributes which have zero importance value. After discarding those genes, a new random forest is built again with the selected set of genes that yields the smallest out-of-bag (OOB) error rate.

The R package `randomForest` is used in this thesis. The two main parameters of random forest are `mtry`, the number of input variables randomly chosen at each split and `ntree`, the number of trees in the forest. These two parameters are set to their default values (`ntree` $= 500$; ; `mtry` $= \sqrt{d}$). Two other parameters are very important in this thesis due to the problem of imbalanced classes and ignoring them may result in poor feature selection. The two parameters are `cutoff`, a vector weight for each class, and `sampsize`, the number of cases to be drawn to grow each tree. These two parameters are carefully tuned in order to achieve a successful feature selection process without ignoring the minority classes. Explanation about these parameters and how to handle the imbalanced classes are given in the next section before going to describe the BIRF algorithm.

4.1.1 Handling Imbalanced Classes Effect on Feature Selection

Balanced Iterative Random Forest algorithm also takes in consideration the problem of imbalanced classes and its effect on the classification process. Standard classifiers may have poor performance on datasets with imbalanced classes because they are biased to the class with a large number of samples and pay less attention to the rare class. This may result in poor accuracy for the minority class. Similar to standard classifiers, random forest also suffers from learning of extremely imbalanced classes dataset but random forest allows us to mitigate this problem. Two solutions are applied on the BIRF to alleviate this problem: balanced sample and cost sensitive learning. Balanced sample solution is based on the parameter `sampsize` which aims

to induce random forest to build trees from a balanced bootstrap sample, that is a bootstrap sample is drawn from the minority class with the same number of samples from the majority class. In the case of imbalanced data, there is a high probability that random forest will build a tree from a bootstrap sample that contains only a few samples from the minority class, resulting in poor performance for predicting the minority class.

The second solution aims to apply a cost sensitive learning technique through the parameter `cutoff` in order to make random forest more suitable to learn extremely imbalanced data. The idea of cost sensitive learning is to assign a high cost for misclassification of the minority class and trying to minimize the cost of the major class. As random forest generates votes to classify the input case, cost weights are applied on those votes in order to make the calculation of the votes as a proportion rather than whole. This solution aims to balance the distribution of classes without altering the semantics of the dataset neither by down-sampling nor over-sampling the dataset. This is because the down sampling technique aims to remove samples from the majority class in order to attain the same number of samples in the minority class. This method may result in losing some useful information from the dataset. Moreover, only a small number of observations are available in the gene expression datasets. Oversampling, on the other hand, is where the number of samples in the minority class is increased to attain the same number of samples in the majority class. However, increasing the number of samples using SMOTE [19] which was reviewed in section 2.5.2, cannot be applied on very high dimensional datasets due to the curse of dimensionality.

A Cost Weight Finder (CWF) algorithm is proposed to tune the parameter `cutoff` with the best weights required to balance the distribution of classes. This algorithm uses random forest with the default settings as a base classifier for classification measurement performance for different randomly generated weights.

Given an input data matrix $\mathbf{X}\epsilon\mathbb{R}^{n\times g}$ where n is the number of samples and g is the number of genes, iterative random forest takes a random sample $\mathbf{Y}\epsilon\mathbb{R}^{n\times f}$ where $f << g$ from the input data matrix and initializes a random cost weight for each class in the range [0,1]. In every iteration, classification performance of the classes is evaluated by measuring the area under the receiver operating characteristic curve (AUC). If this evaluation measure (i.e AUC) is less than one, new random values are generated again until the evaluation measure reaches the maximum value of one. The algorithm is run for a predefined number of iterations in order to avoid an infinite loop. When the algorithm reaches the predefined number of iterations, the algorithm then chooses the highest evaluation measure found with the corresponding cost weights. The proposed algorithm to find these weights is presented in algorithm 4.1.1.

Algorithm 4.1.1: CostWeightFinder( )

$itr \leftarrow numberofiterations$

$g \leftarrow numberofgenes$

$n \leftarrow numberofpatients$

$currentOOB.error \leftarrow 1$

$data \leftarrow \begin{cases} f\ features\ with\ n\ Patients\ uniformly\ randomly\ selected \\ from\ the\ dataset\ where\ f << g \end{cases}$

for $i \leftarrow 0$ **to** itr

do $\begin{cases} costWeight.vector \leftarrow Assign\ stratified\ randomly\ in\ [0,1]\ an\ error\ cost\ for\ each\ class \\ \textbf{REPEAT} \\ \quad \begin{cases} Run\ random\ forest\ on\ this\ sample \\ previousOOB.error \leftarrow currentOOB.error \\ currentOOB.error \leftarrow Read\ the\ outOfBag(OOB)\ error \\ Find\ the\ importance\ value\ for\ each\ feature \\ Delete\ features\ with\ importancevalue <= 0 \end{cases} \\ \textbf{UNTIL}\ \ currentOOB.error > previousOOB.error \\ \textbf{if}\ evaluationMeasure\ of\ AUC = 1 \\ \quad \textbf{then} \begin{cases} Save\ the\ costWeight.vector \\ \textbf{exit} \end{cases} \\ \quad \textbf{else} \begin{cases} array[] \leftarrow Store\ the\ assigned\ error\ cost\ values\ (costWeight) \\ continue\ with\ the\ For\ loop \end{cases} \end{cases}$

$costWeight.vector \leftarrow Find\ the\ best\ obtained\ error\ cost\ values\ in\ array$

$Save\ costWeight.vector$

return $(costWeight.vector)$

4.1.2 Algorithm of Balanced Iterative Random Forest

A balanced iterative random forest algorithm is proposed to select the most relevant genes to the disease and can be used in the classification and prediction process. The algorithm uses the cost weights obtained from the Cost Weight Finder (CWF) algorithm and the balanced sample solution to handle the imbalanced classes.

Due to the large size of gene expression datasets and the limitation of current hardware, it was not possible to run BIRF algorithm on all genes of the dataset. Consequently, the first stage in this algorithm is to divide the dataset, by the number of genes, randomly into different datasets. Then the BIRF algorithm is run on each dataset to select the informative genes. The selected genes from each dataset are then combined to form a new gene expression dataset with fewer attributes. With this obtained dataset, BIRF then starts with an iterative attribute elimination process. The developed algorithm is presented in algorithm 4.1.2

Algorithm 4.1.2: BIRF(*dataset*)

$currentOOB.error \leftarrow 1$

$randSamples \leftarrow \begin{cases} Divide\ the\ dataset\ randomly\ into\ different\ samples \\ by\ the\ number\ of\ features\ (without\ replacing) \end{cases}$

for $i \leftarrow 1$ **to** $number(randSamples)$

do $\begin{cases} Run\ random\ forest\ on\ randSamples[i]\ using\ the\ cost\ weights\ found\ by\ CWF \\ Find\ the\ importance\ value\ for\ each\ feature \\ tempRandSamples[i] \leftarrow Delete\ features\ with\ importance\ value <= 0 \\ Store\ randSamples[i]\ with\ the\ selected\ features \\ Accumulate\ the\ selected\ features\ in\ the\ dataset\ reducedDataset \\ reducedDataset \leftarrow reducedDataset + tempRandSamples[i] \end{cases}$

REPEAT

$\begin{cases} Run\ random\ forest\ on\ the\ reducedDataset \\ Find\ the\ importance\ value\ for\ each\ feature \\ reducedDataset \leftarrow Delete\ feature\ with\ importance\ value <= 0 \\ previousOOB.error \leftarrow currentOOB.error \\ currentOOB.error \leftarrow Calculate\ the\ OOB\ error \\ \textbf{UNTIL} \\ \textbf{if}\ currentOOB.error >= previousOOB.error \end{cases}$

return (*reducedDataSet*)

4.1.3 Validation of Over-Fitting

Over-fitting occurs in statistics and machine learning algorithms especially when these algorithms are dealing with complex datasets such as gene expression datasets (many attributes relative to small number of samples) [38]. It was also established in the

literature review that one of the characteristics of random forest is that it less prone to over-fitting. Nevertheless, to further support the process of feature selection, additional experiments are performed to ensure that there is no over fitting in the gene selection process. In each iteration, after removing the irrelevant genes from the training dataset, the same genes are eliminated from the test dataset and classification performance is evaluated. Once the classification error rate of the test dataset starts to increase after reaching a minimum value, it is assumed that the training dataset is over trained and that the algorithm should stop at this stage. Experiments about this technique are given in section 4.4.

4.1.4 Validation of the Selected Genes

The decision about how many attributes to select in feature selection process is critical and has two effects. Selecting too many attributes from the original dataset makes it difficult to analyse these genes in terms of their effect on the disease. On the other hand, in order to build a generalizable gene expression similarity measurement model, it is important to incorporate as much information as possible. Therefore, it is possible to make a principled decision by testing the effect of the selected number of attributes on the classification performance to know whether more genes provide new information or not.

Although the out-of-bag error rate of the training dataset with the selected genes may reach a minimum value and provides a good classification performance, these selected genes may still require further exploration to determine whether they are globally informative or they are just selected by chance and may be only predictive to that particular dataset. In order to support the gene selection process and to distinguish

between predictive attributes and those that only appear to be predictive, this thesis proposes a methodology to decide which genes best describe the original dataset. The methodology repeats experiments training the BIRF algorithm several times and reduces the training dataset into several subsets (resultant attribute lists). The resultant attributes in each subset are then compared to see which attributes are selected in multiple executions and which attributes are once selected. The assumption is that the attributes which appear in multiple subsets are more informative than attributes that appear in a single subset. The subset that contains the most common attributes with the minimum out-of-bag error is the one that best describes the original dataset.

4.2 Datasets

The experiments of this thesis are performed on a Childhood Leukaemia gene expression dataset which is collected from The Children's Hospital at Westmead. The dataset was normalized by the Distance Weighted Discrimination (DWD) algorithm [67]. The entire childhood leukaemia gene expression dataset is composed of 110 patients with expression values for 22,678 genes. However, stratified random sampling is applied on the gene expression dataset and it is divided into training and test datasets. The training dataset is composed of 70 patients who are classified as follows;

- Medium risk (53 patients)
- Standard risk (11 patients)
- High risk (6 patients)

The test dataset is composed of 40 patients and they are classified as follows;

- Medium risk (25 patients)
- Standard risk (10 patients)
- High risk (5 patients)

Other datasets have been used in this research project for evaluation of BIRF. Three publicly available microarray datasets: Golub Leukaemia cancer, Colon cancer and Lung cancer datasets were used in this chapter. All these datasets are characterized by a relatively small number of samples with a high dimensional space. For these three datasets, the same training and test data reported in the previous studies are used in these experiments, without changing the sample sizes, so that the obtained results can be objectively compared with earlier methods.

- Golub Leukaemia dataset is a well-studied publicly available microarray benchmark collected by Golub and colleagues and is produced with an Affymetrix oligonucleotide microarray [35]. This dataset consists of gene expression profiles for two acute types of leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloblastic leukaemia (AML). The dataset is normalized by z-score [68] and consists of 72 samples: 23 of AML, and 49 of ALL. Each sample is measured over 7,129 genes. These samples are broken into two datasets: a training dataset and a test dataset composed of 38 and 34 samples respectively (see Table 4.1).
- Colon dataset is another publicly available microarray dataset which was obtained with an Affymetrix oligonucleotide microarray [3]. The Colon dataset

Table 4.1: Microarray gene expression datasets

Datasets	Number of classes	Number of features	Number of training samples	Number of testing samples
Childhood Leukaemia	3	22,678	70	40
Golub Leukaemia	2	7,129	38	34
Colon cancer	2	2,000	38	34
Lung cancer	2	12,533	32	149

contains 62 samples, with each sample containing the expression values for 2000 genes. Each sample indicates whether it came from a tumour biopsy or not. This dataset is used in many different research papers on feature selection of gene expression datasets [10] [15] [34]. The dataset is quite noisy but the real challenge is the shape of the data matrix where the dimensionality of the feature space is very high compared to the number of cases. It is important to avoid over fitting in this dataset. Although the number of cases is very low, the dataset is split into two datasets: a training dataset and a test dataset composed of 38 and 34 samples respectively (see Table 4.1).

- Lung cancer dataset is also used in the experiments and it was generated with an Affymetrix oligonucleotide microarrays and normalized by z-score [36]. For each sample it is indicated whether it came from a malignant pleural mesothelioma (MPM) or adenocarcinoma (ADCA). There are 181 tissue samples (31 MPM and 150 ADCA) which already have been broken into training and testing samples. The training dataset contains 32 of samples, 16 MPM and 16 ADCA. The remaining 149 samples are used for test. Each sample is described by 12533 genes. Similar to the Colon dataset, the Lung cancer dataset is also noisy but with more samples and genes. These samples are broken into two datasets: a

training dataset and test dataset composed of 32 and 149 samples respectively (see Table 4.1).

4.3 Experiments

Several experiments are performed on the balanced iterative random forest algorithm in order to demonstrate the validity of the proposed algorithm, to evaluate the algorithm on different datasets and to compare our achieved results to other algorithms by using the same datasets.

4.3.1 Experiments on Childhood Leukaemia Dataset

Balanced Iterative Random Forest is validated with the Childhood Leukaemia gene expression dataset collected from The Children's Hospital at Westmead. It is important to note that the main purpose of these experiments is to find a subset of genes most closely correlated with the leukaemia risk type distinction. Also it is important to incorporate as much data as possible without including so much data that may result in losing interesting separations between patients. The set of informative genes to be used in the prediction of risk type was chosen to be the 107 genes selected at a lowest out-of-bag error rate 0.04. The confusion matrix on the training dataset for this 107 genes dataset is shown in Table 4.2. As can be seen from this matrix, patients with high risk are fully predicted. Forty nine patients with standard risk are predicted correctly. However, two patients are predicted as high risk and two patients as standard risk. With respect to the standard risk patients, ten patients are predicted correctly and only one patient is predicted as a medium risk. In order to

validate the results we have obtained from this experiment, test dataset is processed by random forest to see the classification performance of the selected genes. Table 4.3 shows the confusion matrix for classification of the test dataset and demonstrates that the classifier generalised reasonably well. Precision and accuracy are computed

Table 4.2: A confusion matrix for the childhood leukaemia training dataset

	Predicted High	Predicted Medium	Predicted Standard
Actual High	6	0	0
Actual Medium	2	49	2
Actual Standard	0	1	10

Table 4.3: A confusion matrix for the childhood leukaemia test dataset

	Predicted High	Predicted Medium	Predicted Standard
Actual High	3	1	1
Actual Medium	0	20	5
Actual Standard	0	3	7

on the test dataset in order to evaluate the classification performance of the selected genes. The precision of the high, medium and standard classes are 100%, 83.3% and 53.8% respectively. However, the accuracy of the high, medium and standard are 60%, 80% and 70% respectively.

4.3.1.1 Validation of Results in terms of Over-Fitting

It is important to be able to validate results and prove that they are not due to over-fitting the training data. A sophisticated technique is used in these experiments and it is described in section 4.2.3. The iterative random forest is run again on the training dataset for selecting the relevant bio-markers. Simultaneously, the test dataset is involved in this process to validate that the training process doesn't over train. The

training process of the iterative random forest on the Childhood Leukaemia dataset is shown in Figure 4.1.

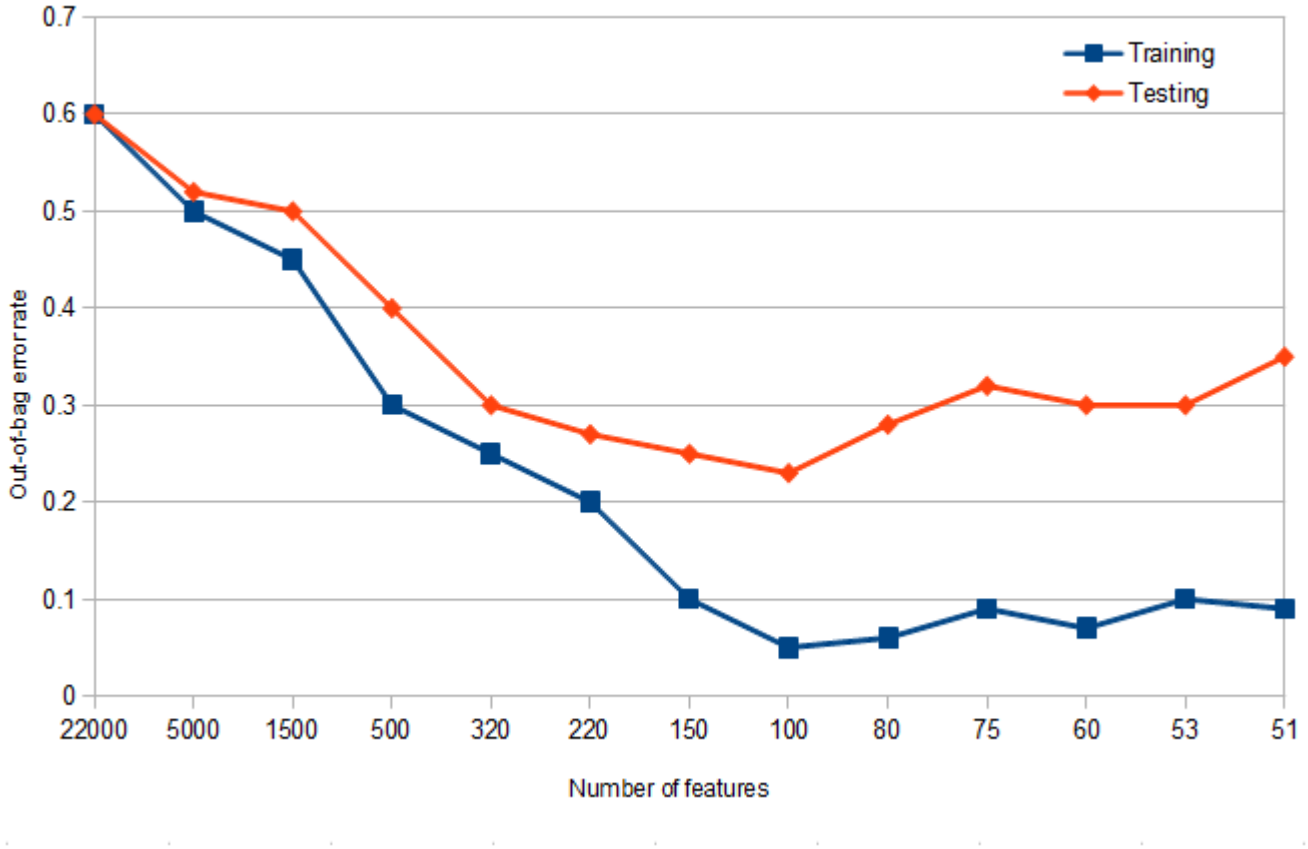

Figure 4.1: Variations of Out-of-bag error during selecting the features

As can be seen from the graph and based on the training dataset, the out-of-bag error decreases as the number of irrelevant features and noise is eliminated at each iteration. After several iterations, the out-of-bag error becomes stable in a range between 0.1 and 0.04. With respect to the test dataset, the classification performance is quite similar to the training dataset but with less accuracy. Moreover, it is important to note that there is no over-training of the dataset in the first eight

iterations and with the number of features is greater than 100. The out-of-bag error of the test dataset consistently decreases in the first eight iterations. After the eighth iteration, the out-of-bag error of the test dataset is increased again and becomes unstable for several iterations. The training stops at the eighth iteration with the lowest out-of-bag errors are achieved for the test and training datasets.

4.3.1.2 Analysis of Selected Genes

To further evaluate the attribute-selection process, experiments training the balanced iterative random forest algorithm are repeated three times. The resultant attribute lists from each repetition are then compared to the attributes obtained from the initial experiment where 107 genes have been selected. The goal of that is to see whether the 107 selected attributes appear in the three resultant attribute lists or not. It is interesting to note that 80% of the top 20 genes consistently appear in the three lists, and the top 20 genes remain near the top in the other three lists. 64% of the top 100 selected genes from each list are the same. This supports the fact that the top selected genes are globally predictive and have not been selected by chance. Moreover, it also indicates that the the feature selection process was not over-trained.

Classification performance of the three resultant attribute lists are also compared to see whether the list that contains the most common attributes provides good separation between the patients. Table 4.4, 4.5 and 4.6 show the classification performance of the first, second and third list respectively. The out-of-bag errors of the three lists are 17.27%, 14.55% and 16.73% respectively. It can be clearly noticed from this analysis that the selected 107 genes (obtained in section 4.3.1) contains the most common attributes, it provides the minimum out-of-bag error (4%) and it is the best

one describes the original dataset.

Table 4.4: A confusion matrix for the childhood leukaemia test dataset (first list)

	Predicted High	Predicted Medium	Predicted Standard
Actual High	3	1	1
Actual Medium	1	18	6
Actual Standard	1	3	6

Table 4.5: A confusion matrix for the childhood leukaemia test dataset (second list)

	Predicted High	Predicted Medium	Predicted Standard
Actual High	2	2	1
Actual Medium	0	20	5
Actual Standard	1	2	7

Table 4.6: A confusion matrix for the childhood leukaemia test dataset (third list)

	Predicted High	Predicted Medium	Predicted Standard
Actual High	3	1	1
Actual Medium	2	18	5
Actual Standard	1	3	6

4.3.2 Experiments on the Three Public Microarray Datasets

One of the most important aspects of any experiment is validating the algorithm. In this thesis, validation is done by applying the proposed algorithm on the three publicly available microarray datasets described in section 4.3. If the algorithm performs well then the feature selection process is done correctly.

Balanced Iterative Random Forest is initially validated on the Golub Leukaemia dataset [35]. The minimal out-of-bag error rate of zero is achieved at 50 features which are selected as the most important features in classification. This result is

Table 4.7: A confusion matrix for the Golub training dataset

	Predicted ALL	Predicted AML
Actual ALL	25	0
Actual AML	0	13

Table 4.8: A confusion matrix for the Golub test dataset

	Predicted ALL	Predicted AML
Actual ALL	24	0
Actual AML	0	10

also validated in order to make sure that the feature selection process has not overfitted to that training dataset. With the selected 50 features, 100% accuracy has been achieved on both the training and test datasets. Tables 4.7 and 4.8 show the obtained results.

The same procedure is applied to the Colon dataset [3]. Sixty nine features are selected as the most important features in classification with a minimal out-of-bag error rate of zero. For the test data, an accuracy of 96% has been achieved where only one patient is wrongly classified. Tables 4.9 and 4.10 show the obtained results.

With respect to the Lung cancer dataset [36], 57 features are selected from the training dataset with zero error rates. With the test data, however, 97% accuracy have been achieved with only one patient is wrongly classified. Tables 4.11 and 4.12 show the obtained results.

These results suggest that BIRF works well for several gene expression datasets.

Table 4.9: A confusion matrix for the Colon training dataset

	Predicted cancerous	Predicted normal
Actual cancerous	27	0
Actual normal	0	13

Table 4.10: A confusion matrix for the Colon test dataset

	Predicted cancerous	Predicted normal
Actual cancerous	12	1
Actual normal	0	9

4.4 Comparison With Other Algorithms

In the previous section, we performed experiments on three different public gene expression datasets that have been analysed by researchers using various gene selection methods. We report the results achieved by Support Vector Machine-Recursive Feature Elimination (SVM-RFE), Multiple SVM-RFE (MSVM-RFE) and Optimal Bayes (OB) as shown in Table 4.13. It can be seen that a very high accuracy is achieved on most of the datasets studied here. For instance, on the Golub Leukemia dataset, BIRF, SVM-RFE and MSVM-RFE [28] have a perfect prediction on the test dataset. The best obtained accuracy for NB was by Inza et al [41] at 95%. The best performance on the Colon dataset is achieved at 96% obtained by BIRF. On the other hand, the accuracy of SVM-RFE, MSVM-RFE and NB is 83.71% [28], 83.57% [28] and 87% [41] respectively. With respect to the Lung cancer dataset, 81% and 88% accuracies are reported (not shown in Table 4.13) using the bagging and boosting methods [58] where 97% is achieved by BIRF.

Table 4.11: A confusion matrix for the Lung cancer training dataset

	Predicted ADCA	Predicted MPM
Actual ADCA	16	0
Actual MPM	0	16

Table 4.12: A confusion matrix for the Lung cancer test dataset

	Predicted ADCA	Predicted MPM
Actual ADCA	134	0
Actual MPM	1	14

4.5 Summary

This chapter proposes a method called balanced iterative random forest for selecting features in imbalanced gene expression datasets. This algorithm represents contribution 2 listed in section 1.3. It is un-realistic to assume that the attribute-selection algorithm, in this case the balanced iterative random forest algorithm, will be able to pinpoint which attributes can describe the risk type of the patient and identify all of the biologically significant attributes with a such large complex dataset. Nevertheless, the attribute selection process is done carefully by validating the results, and eventually, it produces a small subset which contains the most informative genes. This result was validated and supported through two different experiments: over-fitting validation and analysis of the selected genes. The experiments showed that the classifier did not over-fit the training dataset. Also, the analysis of attributes to distinguish between predictive attributes and those that only appear to be predictive (over-fitted attributes) showed that most of these attributes appeared in multiple repeats of the algorithm runs. Balanced Iterative Random Forest is also applied to three other microarray datasets: Golub, Colon cancer and Lung cancer datasets. Overall, BIRF

Table 4.13: Accuracy results for Colon and Leukaemia datasets

Datasets	Measurements	BIRF	SVM-RFE from [28]	MSVM-RFE from [28]	NB from [41]
Golub	Number of genes	50	95	37	4
	Accuracy	1	1	1	0.95
Colon	Number of genes	69	7	3	2
	Accuracy	0.96	0.83	0.83	0.87
Lung	Number of genes	57	31	33	NA
	Accuracy	0.97	0.96	0.96	NA

resulted in classifiers comparable or superior in accuracy to SVM-RFE, MSVM-RFE and Naive Bayes on the Colon, Golub and Lung datasets.

Chapter 5

Dimensionality Reduction and Visualization

This chapter discusses the dimensionality reduction and visualization step in the case base retrieval framework. Section 5.1 gives a brief introduction about the importance of dimensionality reduction and data visualization in the microarray domain. Section 5.1.1 presents a methodology to reduce the dimensionality of the dataset by choosing a number of components in kernel principal components analysis (KPCA) to represent the data in fewer dimensions. Section 5.1.2 proposes a new algorithm for non-linear dimensionality reduction and visualization of gene expression datasets called Local Principal Component(LPC). Section 5.2 describes datasets used for validation before presenting experiments comparing results to other state-of-the-art algorithms. Section 5.3 gives a conclusion and summary of the contributions for this chapter.

5.1 Dimensionality Reduction of Gene Expression Datasets

Dimensionality reduction is the second stage in the case base retrieval framework illustrated in Figure 3.2. It aims to reduce the dimensionality of the data by transforming a high-dimensional dataset into a lower dimensional one which represents the most important variables that underlie the original data. Both linear dimensionality reduction and non-linear dimensionality reduction algorithms are involved in this thesis for two different purposes: dimensionality reduction and visualization. Linear dimensionality reduction (LDR) algorithm is applied on the Childhood Leukaemia gene expression dataset to reduce the number of attributes of the dataset and to enhance the quality of distance measurement. Non-linear dimensionality reduction, on the other hand, is applied for visualization of gene expression dataset once it is processed by LDR algorithm.

5.1.1 Linear Dimensionality Reduction Approach

Kernel Principal Components Analysis is involved in this work in order to reduce the dimensionality of the dataset. Kernel methods [81] are employed along with PCA to save considerable amounts of computation time in finding the effective principal components. This is because the number of attributes or features is very large in gene expression datasets and is much higher than the number of samples. In normal PCA, the size of covariance matrix is $m \times m$ where m is the number of attributes. However, by employing kernel methods, the size of the kernel matrix is $n \times n$ where n is the number of observations or samples. The idea behind KPCA is to find the

directions or components where the data has maximum variance. This is achieved by finding the eigenvalues with the corresponding eigenvectors for the kernel matrix of the dataset. Dimensionality reduction is then achieved by choosing the largest eigenvalues obtained by KPCA in order to represent the data in fewer dimensions. Dimensionality reduction based on KPCA takes as input $\mathbf{X}\epsilon\mathbb{R}^{n\times m}$ and produces output $\mathbf{Y}\epsilon\mathbb{R}^{n\times d}$ where $d < m$ is the dimensionality of the low dimensional space. The question in this process is: what is the minimum dimension that can be achieved? or which components of KPCA should be selected to represent the dataset in fewer dimensions?

This thesis proposes a wrapper method approach for choosing the best value of $d < m$ where m and d is the dimensionality of the input and output datasets respectively. The concept of this wrapper method is to use a nearest neighbourhood (NN) classifier to evaluate the classification performance on different low dimensionality representations of the data so as to choose the most appropriate value of d. Due to the small number of observations in the training dataset and in order to have a result that can generalise well, k-fold cross validation technique is used in order to determine the classification accuracy of the classifier. The accuracy is evaluated on different low dimensionality representation of the data in order to find the value of d that best describes the data. The training dataset is randomly partitioned into k subsamples. Each fold is taken from the training data for validation and the remaining data is processed by KPCA. For each low dimensional vector obtained by KPCA, the validatation fold is projected onto this vector and presented to the NN classifier to evaluate the classification accuracy. This process is repeated k times and the low dimensionality representation of the data that performs the best on each fold is then

selected. Experiments using this approach to choose d are given in section 5.2.1.

5.1.2 Non-Linear Dimensionality Reduction and Visualization of Gene Expression Datasets

This thesis proposes a non-linear dimensionality reduction algorithm called Local Principal Component (LPC) for dimensionality reduction and visualization of gene expression datasets. The algorithm is based on the first principal component of the local neighbourhoods of each data point. This algorithm overcomes a drawback of PCA when it is applied on non-linear data. This algorithm takes as input $\mathbf{X}\epsilon\mathbb{R}^{d\times n}$ and produces output $\mathbf{Y}\epsilon\mathbb{R}^{q\times n}$ where $q < d$ is the dimensionality of the embedding input vector $\mathbf{X}$ in the low dimensional space $\mathbf{Y}$ (see Figure 5.1). Note that the order

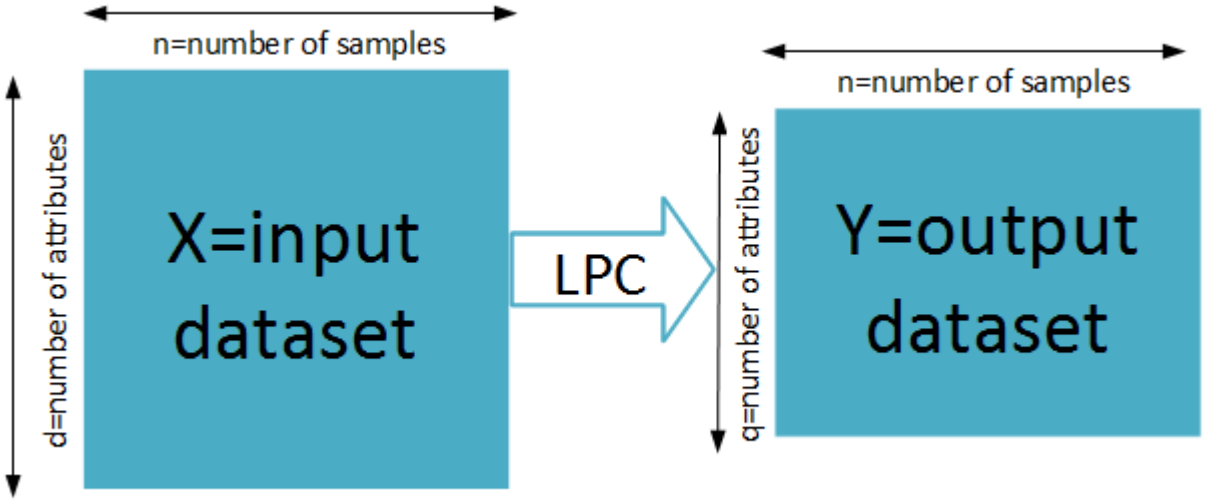

Figure 5.1: Structure of the input and output dataset of LPC

of coordinates differs from that presented in the last section. Four steps are involved in this algorithm. Step 1 computes the neighbors for each data point. For that, we

determine the k-nearest neighbours for each data point $\mathbf{x_i}$. After computing the k nearest neighbours for each data point, a matrix $\mathbf{C_i}$ is created of size $d \times k$. Column vectors of each $\mathbf{C_i}$ are the column vectors of $\mathbf{X}$ for the k nearest neighbours of each point $\mathbf{x_i}$ ordered by increasing distance with the point $\mathbf{x_i}$ itself excluded. Step 2 determines the first principal component for each matrix $\mathbf{C_i}$ by solving the eigenvalue problem for the kernel matrix. Step 3 calculates the orthogonal projection of the first eigenvector and stores it in a square matrix $\mathbf{M} \epsilon \mathbb{R}^{n \times n}$ based on the indexes of the neighbourhood indices obtained from the first step. Step 4 calculates the embedding coordinates $\mathbf{Y}$ using the $\mathbf{M}$ matrix by solving the global eigenvalues and eigenvectors of the squared matrix $\mathbf{M}$.

These steps are accomplished using the following algorithm:

Algorithm 5.1.1: LPC(*inputDataset*)

$d \leftarrow$ *dimensionality of the input dataset* $\mathbf{X}$
$q \leftarrow$ *dimensionality of the output dataset* $\mathbf{Y}$
$n \leftarrow$ *number of samples*
$k \leftarrow$ *number of nearest neighbours*
$\mathbf{C}[n] \leftarrow$ *Empty matrix of size* $d \times k$
$\mathbf{I} \leftarrow$ *Identity matrix of size* $k \times k$
$\mathbf{M} \leftarrow$ *Identity matrix of size* $n \times n$
$\mathbf{IND} \leftarrow$ *Empty matrix of size* $k \times 1$
for $i \leftarrow 1$ **to** n
do
- $\mathbf{C}[i] \leftarrow k$ *nearest neighbors to point* X_i
- $\mathbf{IND} \leftarrow$ *Indexes for the k nearest neighbors to point* X_i
- $\mathbf{P}$ (*size* $k \times 1$) $\leftarrow$ *First Principal component of* $\mathbf{C}[i]$
- $\mathbf{O}$ (*size* $k \times k$) $\leftarrow$ *Orthogonal project of* $\mathbf{P} \times \mathbf{t}(\mathbf{P}) - \mathbf{I}$
- $\mathbf{M}[ind[i], ind[i]] \leftarrow \mathbf{M}[ind[i], ind[i]] + \mathbf{O}$

eigenvalues $\leftarrow$ *Solve the eigenvalue problem for the matrix* $\mathbf{M}$
sortedEigenvalues $\leftarrow$ *Sort eigenvalues in decreasing order*
selectedEigenvalues $\leftarrow$ *Top q eigenvalues.*
outputDatset $\leftarrow$ *Corresponding eigenvectors for the selectedEigenvalues*
return (*outputDataset*)

5.2 Experiments

Three datasets have been used in this study for validation, error estimation and experiments. The Swiss-roll dataset, which was created to explore various dimensionality reduction algorithms, has been used in this study for validation of LPC. It is generated randomly by sampling a 3D Swiss-roll surface with no class label information.

The second one is the famous Iris dataset provided by Anderson [5]. The dataset has 4 features and 150 samples consisting of three species of Iris flower (Setosa, Versicolour and Virginica) with 50 samples of each species. A Childhood Leukaemia gene expression dataset provided by The Children's Hospital at Westmead and described in chapter 4 is visualized with LPC. This gene expression dataset has been preprocessed by applying a feature selection and linear dimensionality reduction algorithm in order to remove the noise and irrelevant features which affect the result of dimensionality reduction algorithm.

5.2.1 Choosing number of Components in KPCA

Kernel principal components analysis algorithm is applied on the Childhood Leukaemia gene expression dataset which has been pre-processed by the balanced iterative random forest to select the relevant bio-markers as described in section 4.4. The dataset now has 107 features and it is composed of 70 patients who are classified as follows:

- Medium risk (53 patients)
- Standard risk (11 patients)
- High risk (6 patients)

Cross validation is used in training in order to avoid over fitting and to achieve better generalized results. Many applications use ten-fold cross validation, but because there are not many samples in our dataset, and in order to have a reasonable number of training and test samples (especially for high risk patients), the value of d (the target dimensionality) is determined using 8-fold cross validation.

Eight-fold Cross validation is applied to evaluate classifications on the dataset. The main consideration of this process is to achieve the maximum possible data reduction that is compatible with accurate classification for training and test samples. As there is no definitive rule for choosing the number of eigenvectors to retain, the decision in this work is based on the classification performance plus the amount of the total variance accounted for by d, the number dimensions selected.

The training Childhood Leukaemia gene expression dataset, as described in Section 4.3, is randomly partitioned into eight sub-samples. Each sample is taken from the training data for validation and the remaining data is processed by KPCA and reduced into different values of d. After each reduction, the test sample is projected into the space of the reduced data and classification accuracy is evaluated. This process is repeated eight times and the eight results are averaged for each value of d so that we have a single estimate at each value of d for the eight validations. Classification accuracy of the obtained reduced dataset is evaluated for each value of d which is determined based on the classification performance. Figure 5.2 shows the classification results for different values of d. As can be seen, the best average value is realized at d=50. Moreover, the accuracy values of different folds are more stable at the value 50. The variance accounted for the chosen d represents 95% of the total variance.

An important reduction is achieved in the data size since the dataset will be in a low dimensional space and distance measurements can be applied on the dataset to compute the similarity between the patients in the case base and new patients. New samples can be projected into low dimensional space without re-computing the eigenvectors since the out-of-sample problem does not exist in KPCA.

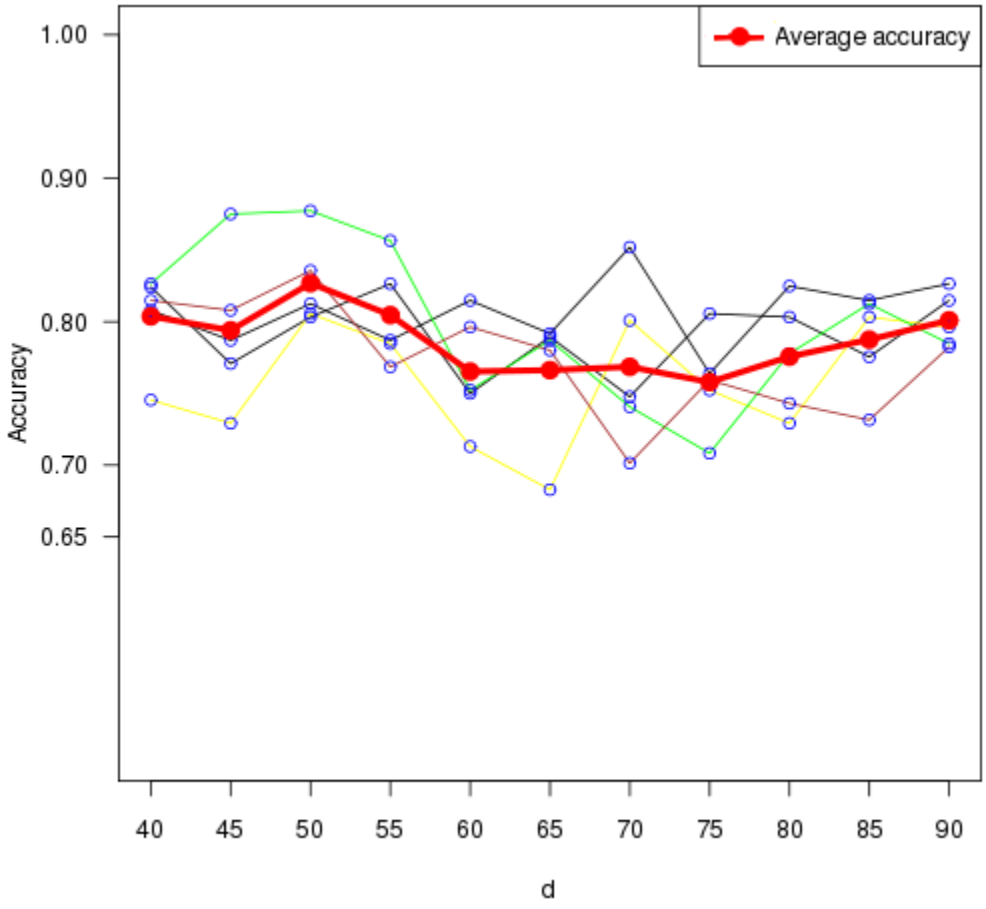

Figure 5.2: Accuracy according to the different dimensionality reduction process using 8-fold cross-validation

5.2.2 Performance Evaluation of LPC

In order to demonstrate the validity of the proposed algorithm, experiments are performed using the Iris dataset and the artificial Swiss roll data set for validation before testing the algorithm on the Childhood Leukaemia gene expression dataset.

5.2.2.1 Error Estimation of LPC

This thesis uses Trustworthiness measurement, which is proposed by Kaski et al [46] and described in Section 2.3.3, for error estimation and quality measurement of LPC algorithm. The aims of this error estimation is to find to which extent neighbors in the input space also have corresponding neighbors in the output space. Two data sets have been examined by trustworthiness; the first dataset is the Swiss-roll data set and the second is the Childhood Leukaemia dataset. The two data sets have been reduced several times with different values of the parameter k. Figure 5.3 and 5.4 show the obtained result of the trustworthiness of LPC applied on Swiss-roll data set and hospital gene expression dataset respectively. As can be seen from Figure 5.3, the trustworthiness is quite stable around the value of 0.98 for different values of parameter k. In Figure 5.4, the trustworthiness dramatically changes based on k, but it can be noticed that the trustworthiness has highest values for $k > 10$ and especially at $k = 14$.

5.2.2.2 Validation of LPC on the Iris Dataset

The algorithm is tested on the Iris dataset described in Section 5.2. Figure 5.5 represents the scattering of the original data set projected onto the 2D space of the first two principal components. Figure 5.6 shows the data reduced to two dimensions by LPC with k=14. The trustworthiness measurement of this reduction is 0.995 which means that the neighborhoods of the data set are preserved with a very small error in the low dimensional space.

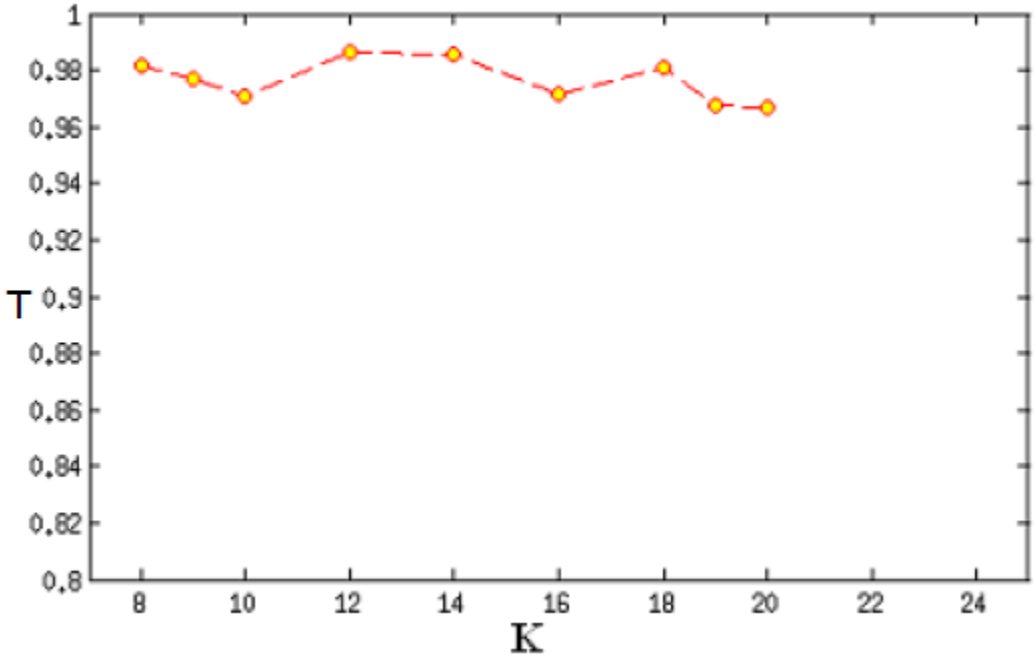

Figure 5.3: Trustworthiness of LPC on Swiss-roll data set

5.2.2.3 Validation of LPC on the Swiss Roll Dataset

The Swiss Roll data set was created to validate various dimensionality reduction algorithms. The LPC algorithm is tested on this data. In this experiment, we have generated 1000 points for the Swiss-roll distribution.

Figure 5.7 represents the original data set in 3D space. As can be seen, the data is folded to have the Swiss-roll form. Figure 5.8 shows that the data, reduced to two dimensions data by PCA, lacking the quality of visualization performance and dimensionality reduction as reported by Lee and Verleysen [56]. Results of projection of data using LPC for k=14 are given in Figure 5.9. In Figure 5.9, the data has better visualization than PCA and it shows that better embedding preserving the shape of the manifold can be achieved by LPC. In order to quantify the comparison

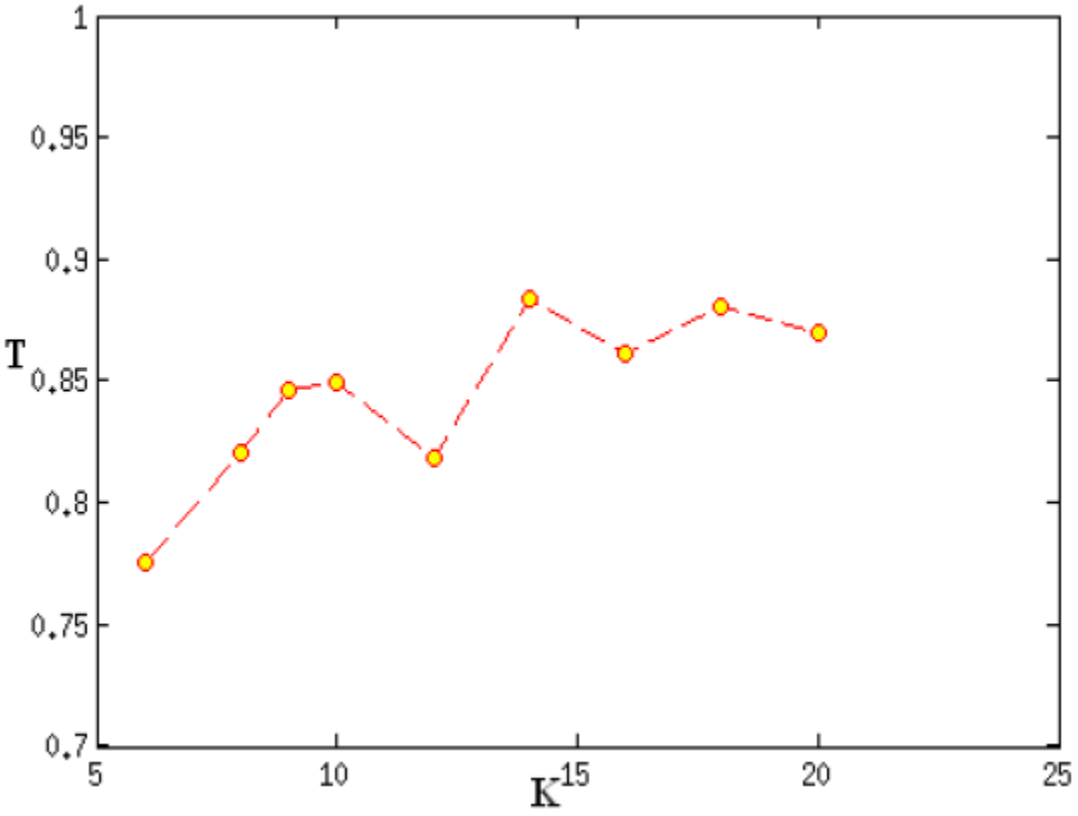

Figure 5.4: Trustworthiness of LPC on Childhood Leukaemia gene expression dataset

between the two outputs, we have measured the trustworthiness of dimensionality reduction performed by PCA and LPC. The trustworthiness of LPC for Swiss-roll is 0.997 compared to 0.848 for PCA which suggests that the LPC embedding is better than PCA for maintaining neighborhood relationships.

5.2.2.4 Visualization of Childhood Leukaemia Dataset By LPC

As our target from this algorithm is to visualize gene expression datasets, the Childhood Leukaemia gene expression dataset collected from The Children's Hospital at

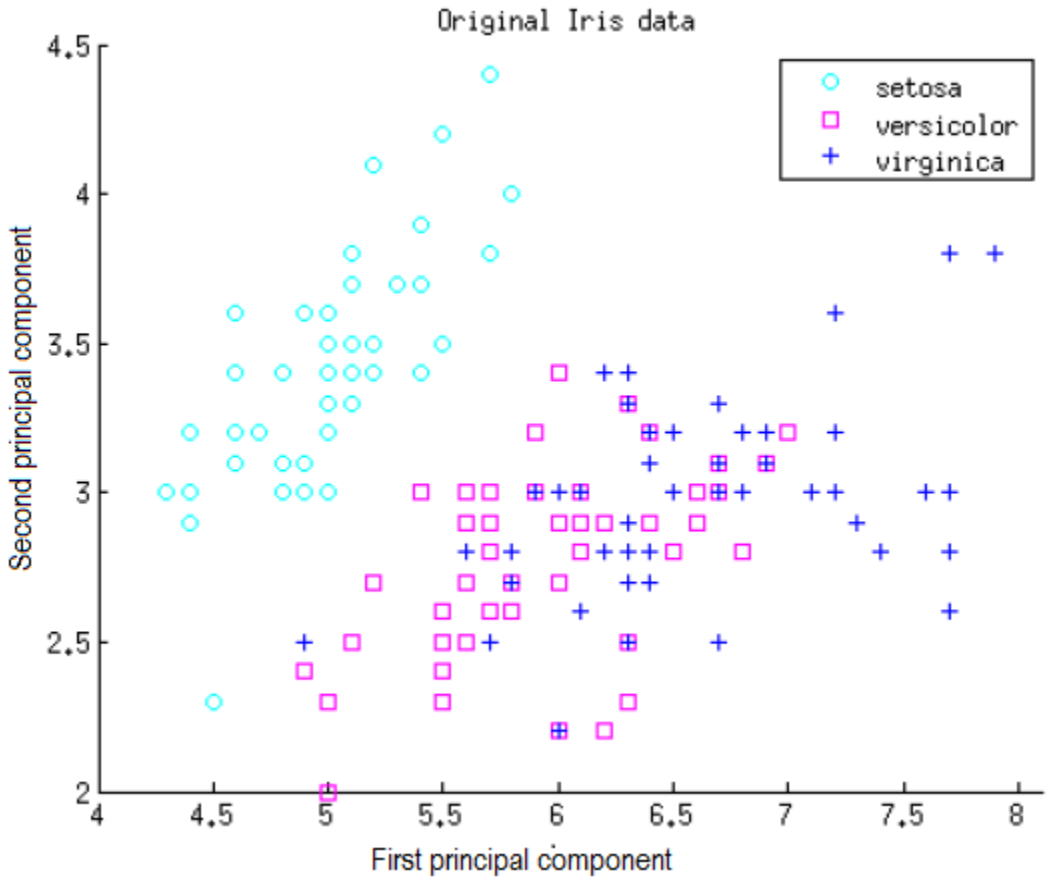

Figure 5.5: Original Iris data set

Westmead has been used to demonstrate this algorithm. This dataset represents the main target for this project. It has been pre-processed by the Balanced Iterative Random Forest to select the relevant bio-markers and kernel principal component analysis for reducing the dimensionality. The dataset has 50 features and is composed of 70 patients who are classified as follows:

- Medium risk (53 patients)
- Standard risk (11 patients)

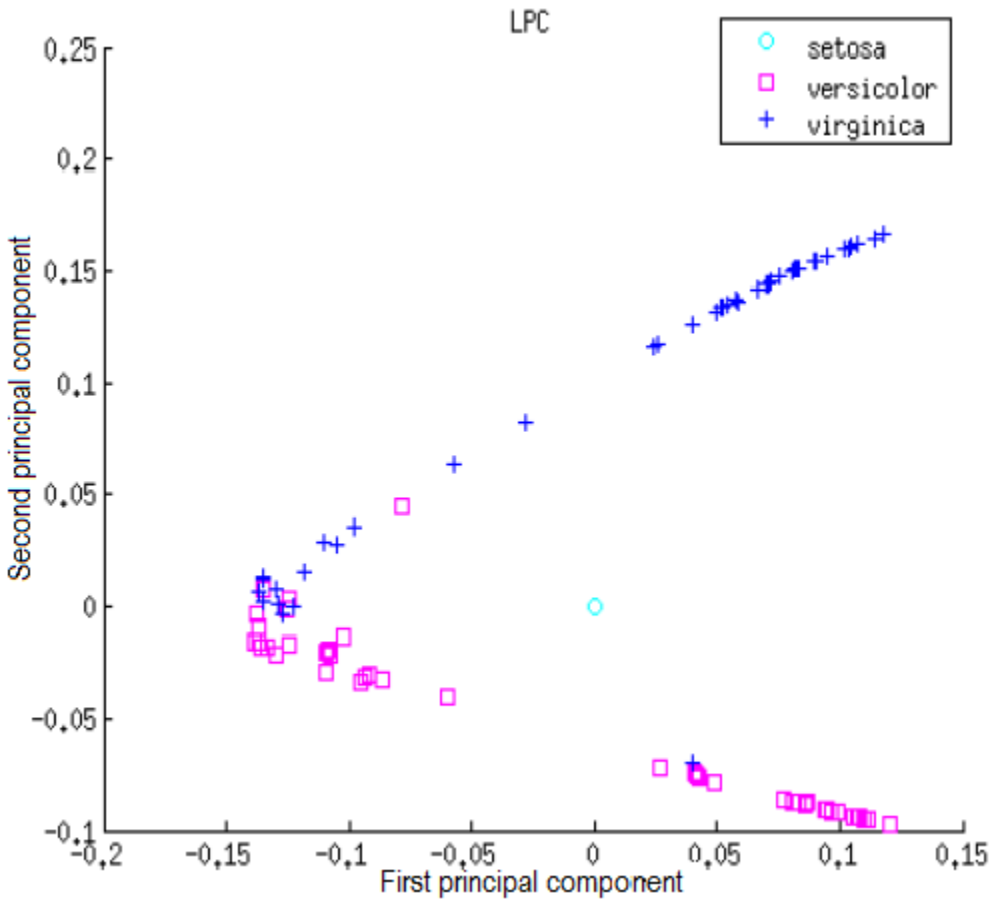

Figure 5.6: Iris data set processed by LPC

- High risk (6 patients)

Figure 5.10 and Figure 5.11 show the result of the obtained data set after applying PCA algorithm and LPC algorithm respectively with k again equal to 14. The trustworthiness of PCA is 0.80. On the other hand, LPC has a trustworthiness of 0.86 which suggests better preserving of neighbourhoods in the lower dimensional space.

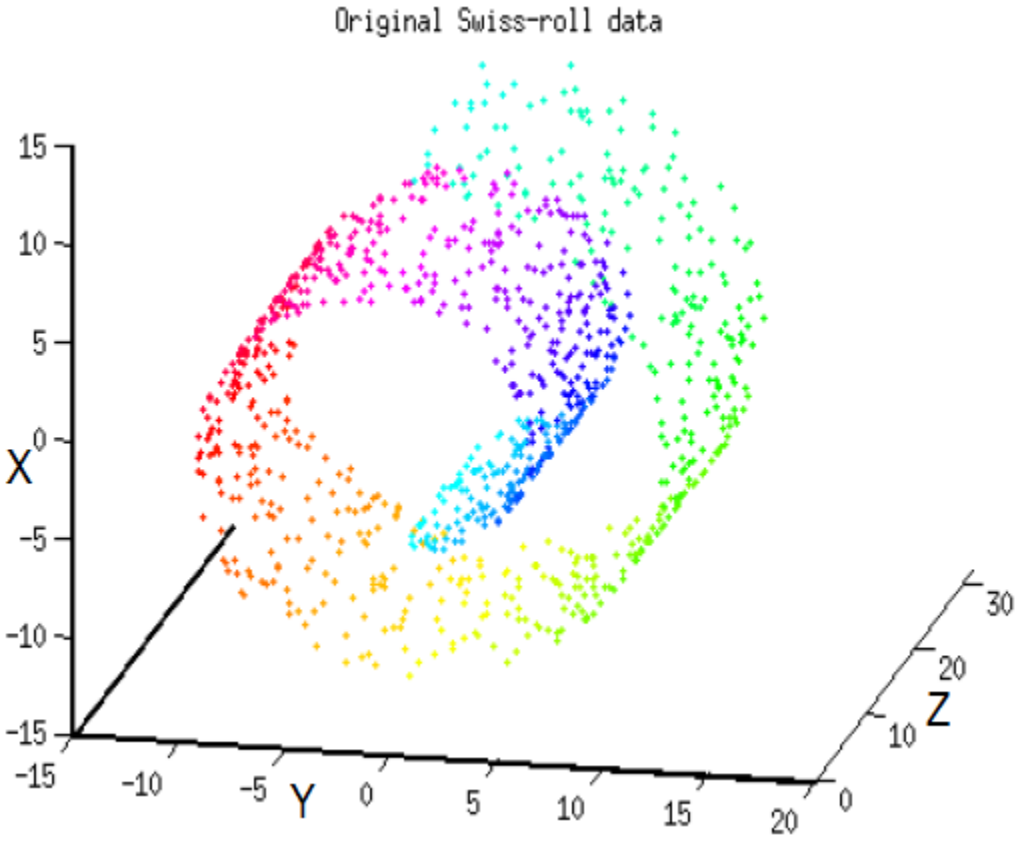

Figure 5.7: Swiss roll data set

5.2.2.5 Comparison With Other Methods

Local principal component is a technique that is similar to locally embedding algorithms (Local Tangent Space Alignment [98] and Local Linear Embedding [77]) in that it constructs a local linear embedding of the k nearest neighbors. The motivation behind LPC is to have good trustworthiness. The algorithm LPC aims to find local principal components around a data point x_i based on the k nearest neighbors of that point. This is followed by another step to extract the first principal component and then to construct the square matrix from these principal components.

Several experiments have been done to make sure that LPC has good dimensionality

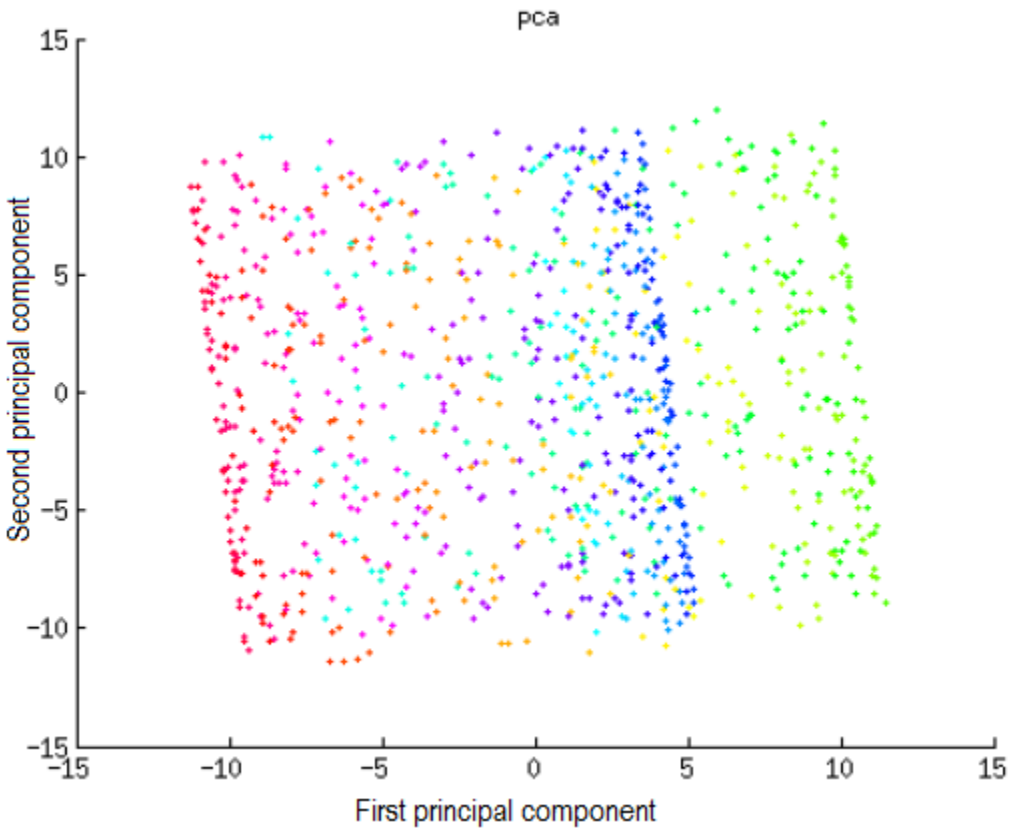

Figure 5.8: Swiss roll data reduced to 2D by PCA

reduction in terms of preserving the neighborhoods of the points in the low dimensional space as in the high dimensional space. As our algorithm is similar to LLE, we present some experiments comparing LPC to LLE. Consequently, we have compared the trustworthiness of LPC to LLE on the Swiss-roll and The Children's Hospital dataset (Childhood Leukaemia dataset). Figures 5.12 and 5.13 present the result of these measurements.

As can be seen from Figure 5.12, the trustworthiness of LPC applied on Swiss-roll dataset is quite stable around the value of 0.98 for different values of parameter k

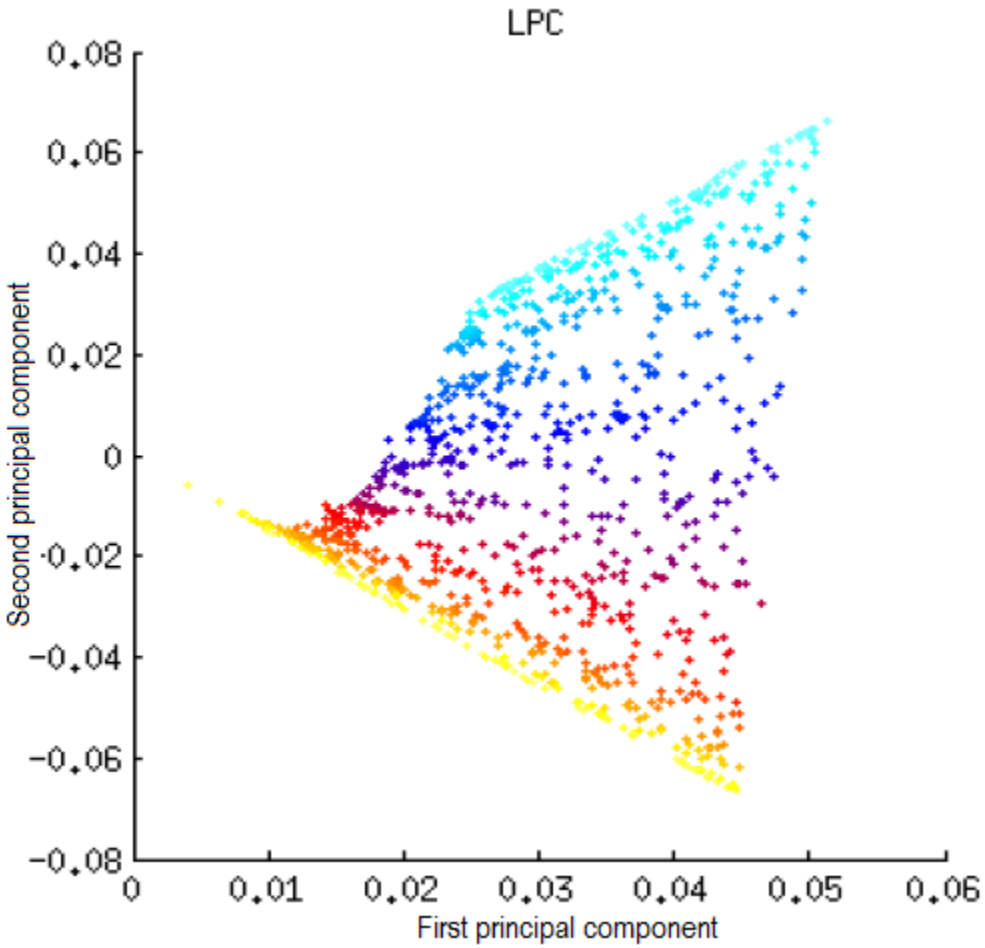

Figure 5.9: Swiss roll data reduced to 2D by LPC

where the trustworthiness of LLE dramatically changes with values of k.

With respect to the Childhood Leukaemia dataset, the trustworthiness of LPC is better than LLE at different values of k especially for $k > 10$. However, the trustworthiness of LPC is less than LLE at $k = 6$. For example at $k = 14$ the trustworthiness of LPC is 0.89 which represents the maximum value. On the other hand, the trustworthiness of LLE is 0.85 at $k = 9$ which represents the maximum value as well.

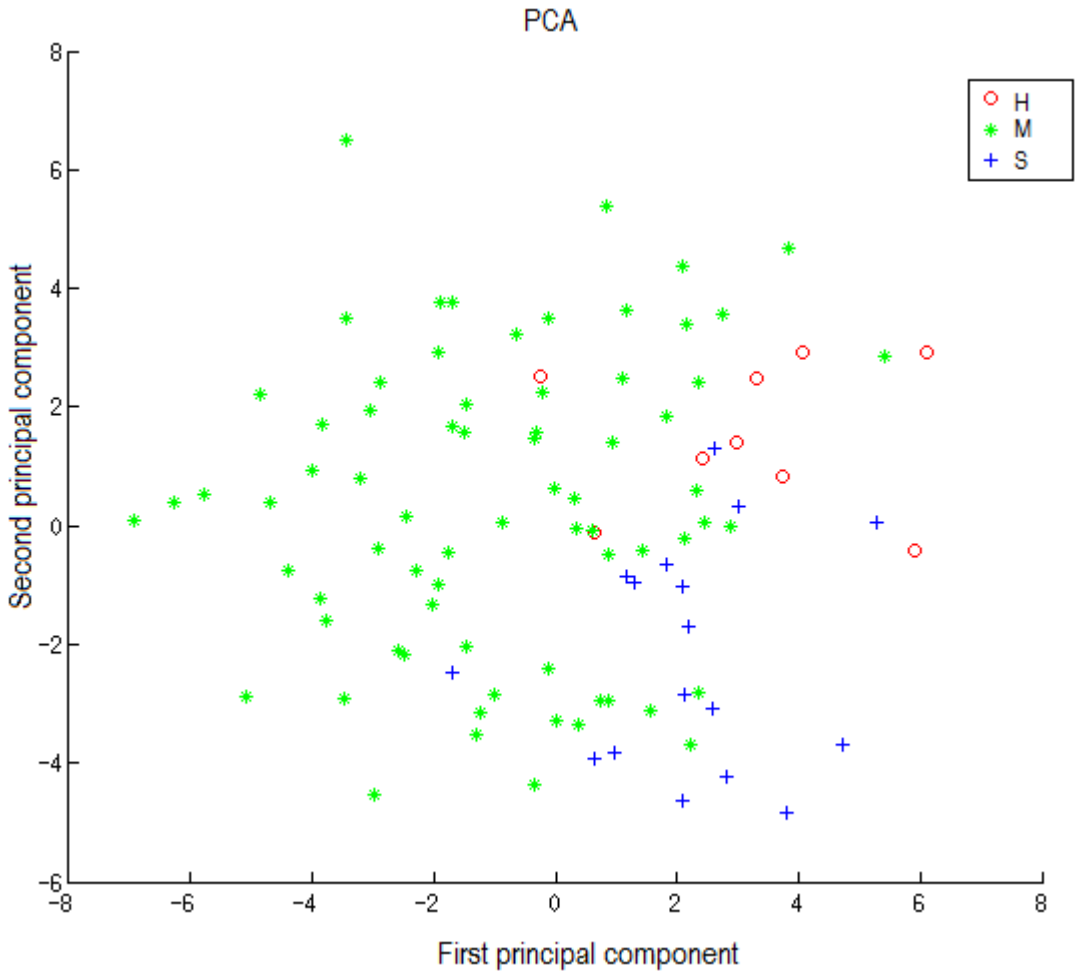

Figure 5.10: Leukaemia data reduced to 2D by PCA

5.3 Summary

This chapter proposed a methodology to reduce the dimensionality of dataset based on KPCA. KPCA reduces the dimensionality of the Childhood Leukaemia dataset by choosing fifty attributes from the dataset. The dataset is reduced into more easily low dimensional space for better quality distance measurement. This chapter also proposed LPC algorithm, contribution 3 in this thesis listed in section 1.3, for high dimensional data reduction and visualization. This algorithm is tested and validated

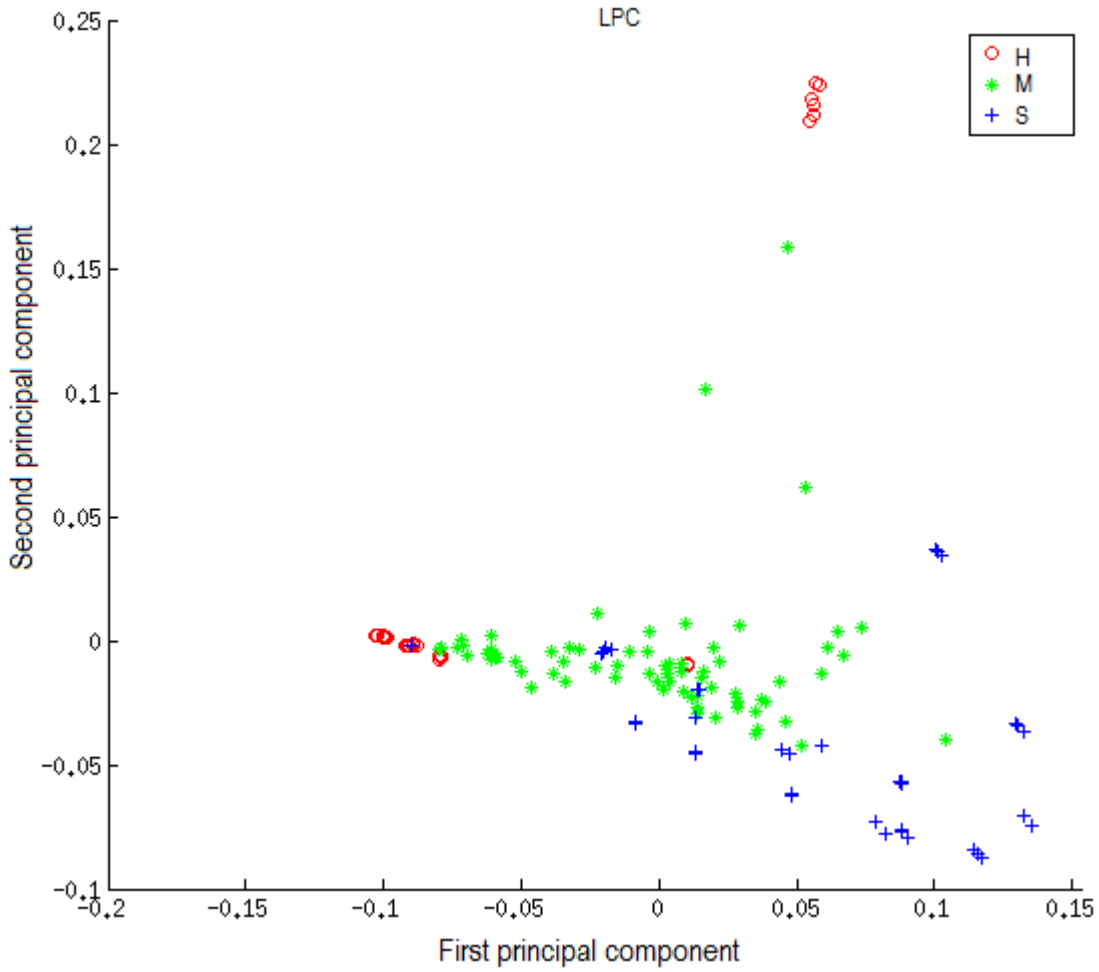

Figure 5.11: Leukaemia data reduced to 2D by LPC

on Iris and Swiss-roll datasets. Good trustworthiness values are achieved with LPC through these experiments. Moreover, LPC is applied on the Childhood Leukaemia gene expression dataset for visualization in two dimensional space. This algorithm provides a way to visualize the data in order to see the position of a patient with respect to other patients.

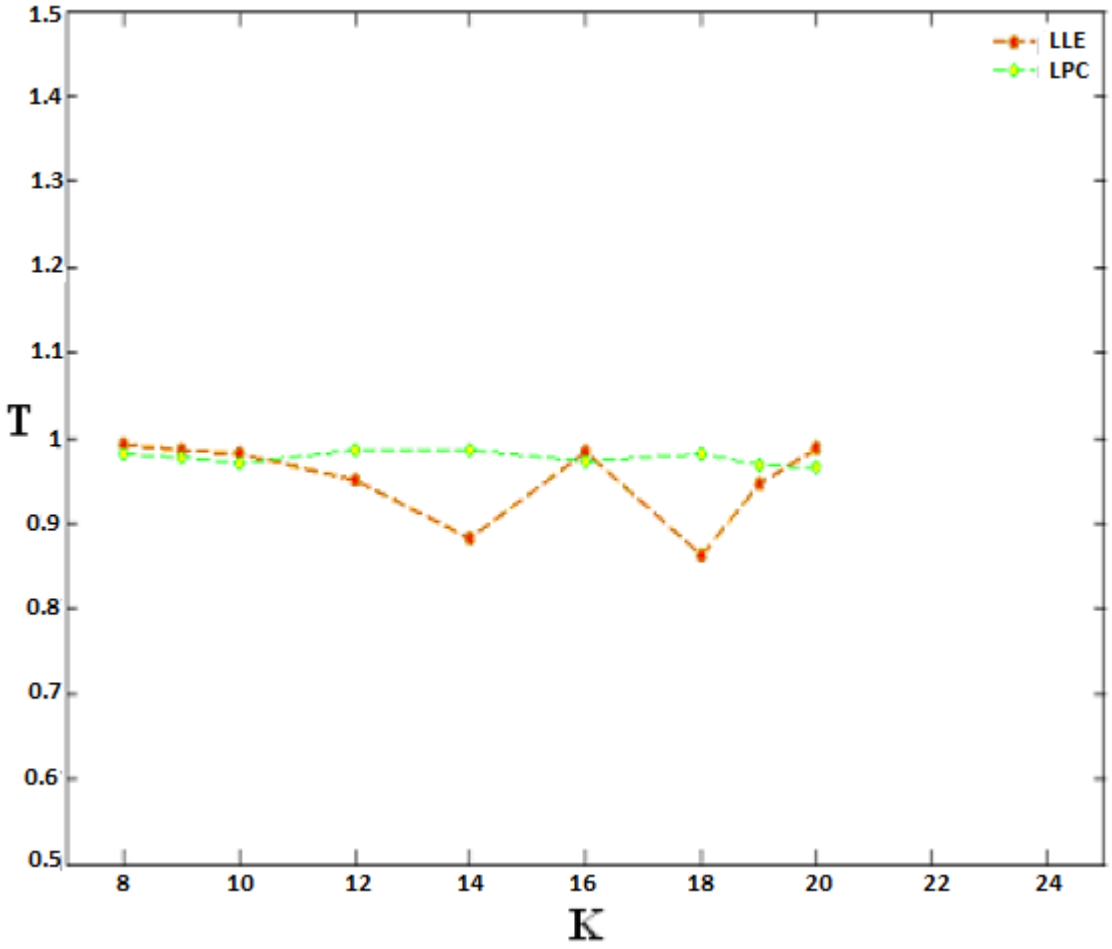

Figure 5.12: Trustworthiness of LPC Vs. LLE using Swiss-roll data set

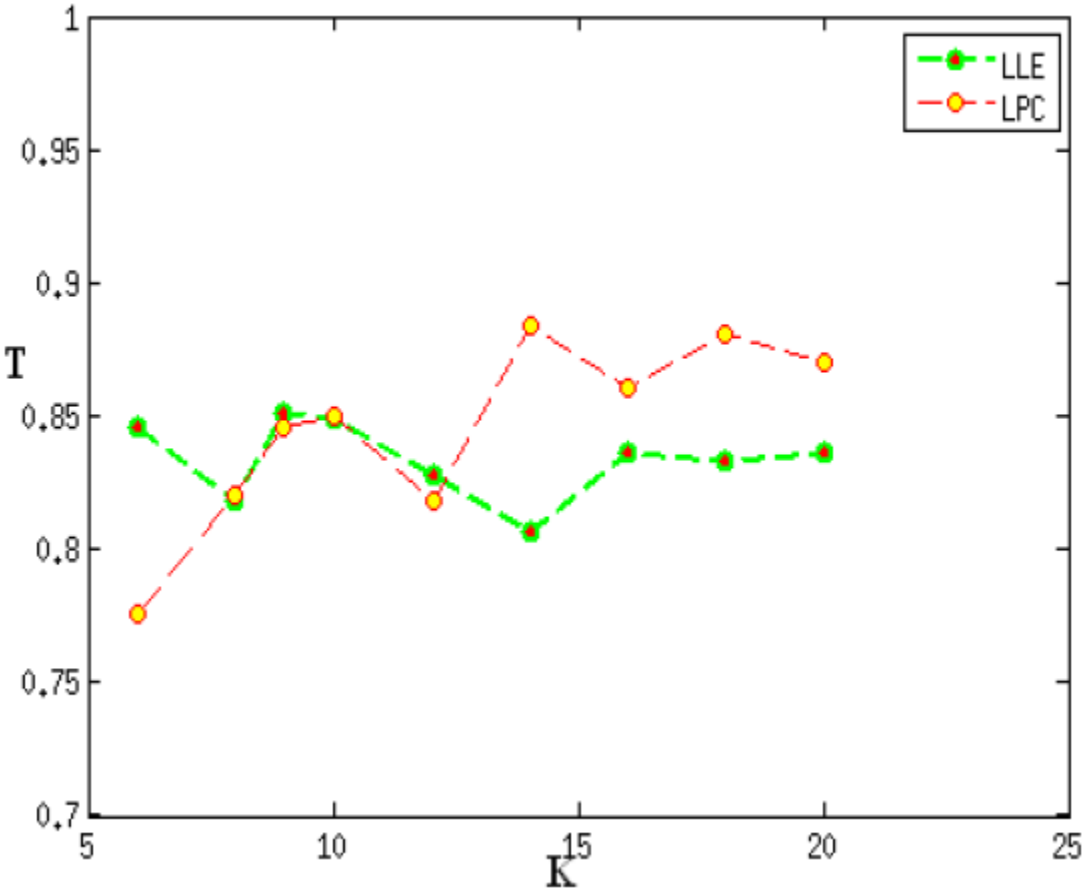

Figure 5.13: Trustworthiness of LPC Vs. LLE using Microarray data set

Chapter 6

Similarity Measurement

This chapter proposes a methodology for similarity measurement and classification in the case base retrieval process. Section 6.1 gives a brief introduction about the nearest neighbour classifier and the factors to be considered in the similarity measurement. Section 6.2 presents the experiments and proposes a Weight Learning Genetic algorithm for feature weighting and the methodology applied to handle the effect of imbalanced classes. Section 6.3 gives a brief conclusion to summarize the achievements and contributions of this chapter.

6.1 Similarity Measurement Approach

The similarity measurement task is one of the major challenges and the key process in the case-base retrieval framework. Accurate similarity measurement has a great value in providing better treatment and diagnosis in the field of patient classification. This thesis uses k Nearest Neighbor(kNN) classifier which is considered as one of the most popular classifier in the machine learning techniques.

According to the literature review, kNN has some disadvantages that already have

been addressed and solved in the previous chapters. kNN is very sensitive to irrelevant features because all features equally contribute to the similarity. This problem is handled in chapter 4 through selecting the relevant features by the Balanced Iterative Random Forest algorithm. Another disadvantage is the effect of the high dimensionality on the distance measurement which is known as the curse of dimensionality problem. This problem also has been addressed in chapter 5 by employing Kernel Principal Components Analysis (KPCA) to transform the data into a lower dimension.

Other two disadvantages of kNN relate to feature weighting and imbalanced classes will be addressed in this chapter. Although the problem of irrelevant features is handled, some features may be much more discriminative than other features. Accordingly, relevant features which play a major role in the similarity measurement should be treated more importantly than other features; i.e. high weight should be given. On the other hand, features with less significance and influence, low weights should be given. In practice, those important features should be weighted more than other features during the calculation of the distance between cases in order to achieve a high accurate retrieval process.

The other disadvantage which also may affect the result of retrieval process is the imbalanced classes with a small number of samples in the case base [63]. This problem is also addressed in this chapter since it affects the prediction of the new query especially if this query sample should be classified to a minority class.

6.2 Experiments

Four experiments are done in this chapter: distance metric selection, determination of parameter k, feature weighting and oversampling the data. Experiments are executed on the basis of the real gene expression dataset collected from The Children's Hospital at Westmead and described in chapter 4. To recall this dataset, initially it has more than 22000 gene expression values for 70 patients. This dataset is processed by the Balanced Iterative Random Forest in chapter 4 for selecting the relevant bio-markers and by KPCA in chapter 5 for dimensionality reduction. The training dataset of patients now has 50 attributes and they are classified as follow:

- Medium risk (53 patients)
- Standard risk (11 patients)
- High risk (6 patients)

However, the test dataset composed of 40 patients and they are classified as follow:

- Medium risk (25 patients)
- Standard risk (10 patients)
- High risk (5 patients)

6.2.1 Distance Metric Selection

The identification of the nearest neighbours in kNN is based on the computation of distances between the query case and cases in the case base, where the most similar cases are determined by evaluating a similarity measure (i.e metric). The performance

of the similarity function is sensitive to the choice of the similarity metric. Therefore, experiments are performed on four similarity metrics: Euclidean, Hamming, Jaccard and Cosine distance to choose the one that provides the better accuracy.

Due to the small number of samples and in order to generalize the results, experiments are performed using c-fold cross validation. It is usually called k-cross validation but c is used here to differentiate it from the parameter k of the nearest neighbour classifier. Most of the applications use the ten-fold cross validation, but because there are not many samples in our dataset, 8-fold cross validation is used to compare the performance of the four similarity metrics. The training dataset is randomly partitioned into 8 sub-samples. Each fold is taken from the training data and treated as a test fold for validation. Each similarity metric is used to calculate the similarity of each test sample with other training samples. This process is repeated for each fold and the eight results from the folds are averaged for each type of similarity metric so that we have a single estimate of accuracy for each similarity metric.

As the training dataset is imbalanced data, the minor class has a very little impact on the measurement of accuracy. This leads to a fact that the traditional accuracy measures cannot be adequate in a situation of extremely imbalanced classes. Therefore, another measurement of accuracy that is Area Under Curve (AUC), presented in section 2.6, is used to determine the accuracy of the classifier.

Based on these experiments, it is found that Euclidean distance consistently performed better and provides the best similarity measurement results than a number of other popular functions such as Hamming, Jaccard and Cosine distance (see Table 6.1).

Table 6.1: Performance comparison of the distance metrics for the nearest neighbour classifier.

	Average accuracy of the eight folds
Euclidean	**0.87**
Cosine	0.83
Hamming	0.77
Jaccard	0.75

6.2.2 Determination of Parameter k

Based on the literature review, it is readily perceived that the value of k affects the performance of the nearest neighbour classifier. This is very obvious in the case of gene expression dataset, since it has a small number of samples with imbalanced classes. The same cross validation procedure, used in the selection of the distance metric, is used here to determine the optimal value of k. The training dataset is randomly partitioned into 8 sub-samples. Each fold is taken from the training data and treated as a test fold for validation. Euclidean distance is used to calculate the similarity of each test sample with other training samples. This process is repeated for each fold at different values of k to find the one that gives the highest accuracy and can be used in classification of query samples. The eight results from the folds are averaged for each value of k so that we have a single estimate at each value of k for the eight validations.

As the dataset has three different classes and in order to avoid the equal voting of kNN, experiments are performed on different values of k where k is indivisible by them i.e (4, 5, 7, 8, 10 and 11). The highest value of k is evaluated at k=11 because the number of the training cases in the minority class is six. The best results were achieved at k=5. The accuracy does not improve as k further increases (see Table 6.2). This result can be justified by looking at the nature of the dataset and how the

three classes are distributed. It can be noticed that a new high risk patient is hard to classify correctly if k is large. All the results reported in the next experiments were obtained based on the k=5. The initial classification performance of the 5NN classifier is presented in the confusion matrix Table 6.3.

Table 6.2: Performance comparison of the kNN for different values of k.

k	Average accuracy of the eight folds
4	0.84
5	**0.87**
7	0.8
8	0.75
10	0.66
11	0.66

Table 6.3: Classification performance results of the test dataset

	Predicted High	Predicted Medium	Predicted Standard
Actual High	1	2	2
Actual Medium	0	23	2
Actual Standard	2	3	5

As can be seen from Table 6.3, the classification results were not good as most of the patients are predicted incorrectly, especially the high risk patients. As discussed before, the NN classifier is very sensitive to the relative importance of each feature. Therefore feature weighting may be required to enhance the classification task.

6.2.3 Feature Weighting

Feature weighting is a technique used to estimate the relative influence of individual feature with respect to the classification performance. When successfully weighted, high impact feature would receive a high value weight, whereas low weight should be

assigned to low impact features. Consequently, a weighted feature-based similarity (weighted Euclidean distance) is used in this thesis to compute the similarity between cases. However, the main issue is how to find the relative importance of each feature. Two hypotheses are proposed to address this issue. Details for these hypotheses will be discussed in the following sections.

6.2.3.1 Hypothesis 1: Eigenvalues can be used as weights for features

The first hypothesis for feature weighting is based on KPCA eigenvalues. As seen in the previous chapter, the dimensionality of the dataset is reduced by applying KPCA. Also has been observed that the dimensionality is achieved by discarding features with a low eigenvalue and keeping only those features with a high eigenvalue. One idea is to employ these eigenvalues in the 5NN as a vector weight and then use it in the Euclidean distance formula. Eight-fold cross validation is applied on the training dataset to compute the accuracies of the unwiehted k-NN classifier. Eight accuracies are obtained from this experiment for the eight test folds. The same procedure is applied on the weighted k-NN with the eigenvalues and eight accuracies are computed for the eight test folds. t-test was conducted to determine whether accuracy differences between the eigenvalues approach and un-weighted 5NN are significant or not. The p-values of the presented t-test is also used to judge the degree of the performance improvement. The paired t-test generates a p-value 0.0283, which indicates that eigenvalues-weighted 5NN and un-weighted 5NN doesn't have the same accuracy. The results of the one-tailed t-test indicates that eigenvalues approach reports accuracy improvements over the un-weightd 5NN. The accuracy increases from 0.74 to 0.82 which represents the highest value that achieved from the eight-fold cross

validation and from 0.53 to 0.61 which represents the lowest value that achieved from the cross validation. In spite of these improvements, the evaluation of the 5NN classifier using these weights was not efficient and the desired results are not achieved. Consequently, hypothesis 1 is not supported in this thesis.

6.2.3.2 Hypothesis 2: Genetic algorithm can be used to seek the weights for features

Genetic algorithm (GA) [40] is considered as a general purpose search process for optimization problems. The key feature of a genetic algorithm is the fitness function which contains the optimization search problem. Consequently, a fitness function is developed for optimization of classification performance by searching for the best genes encoding weights for the similarity measurement. The goal of the GA is to minimize the classification error of the training dataset. Wrapper feature weighting method based on Genetic Algorithms is used to propose a Weight Learning Genetic algorithm to seek the genes encoding weights for the similarity measurement. The concept of this wrapper method is to use a GA to seek the best weights of features with the 5NN classifier providing the GA's fitness function (see figure 6.1). The fitness measure is computed by subtracting accuracy from one (see Equation (6.2.1)). The accuracy is computed by averaging the accuracies of the eight-folds for the 5NN classifier using the generated weights in its Euclidean distance measure.

$$fitness = 1 - Accuracy \tag{6.2.1}$$

The training dataset is randomly partitioned into 8 sub-samples. Each fold is taken from the training data for validation and the remaining data is processed by GA. For

each generated feature weights by GA, the validation fold is presented to the 5NN classifier and the classification accuracy is evaluated by calculating the AUC of the obtained confusion matrix. This process is repeated eight times and the eight results from each fold are averaged for each set of weight so that we have a single estimate at each set of weight for the eight folds.

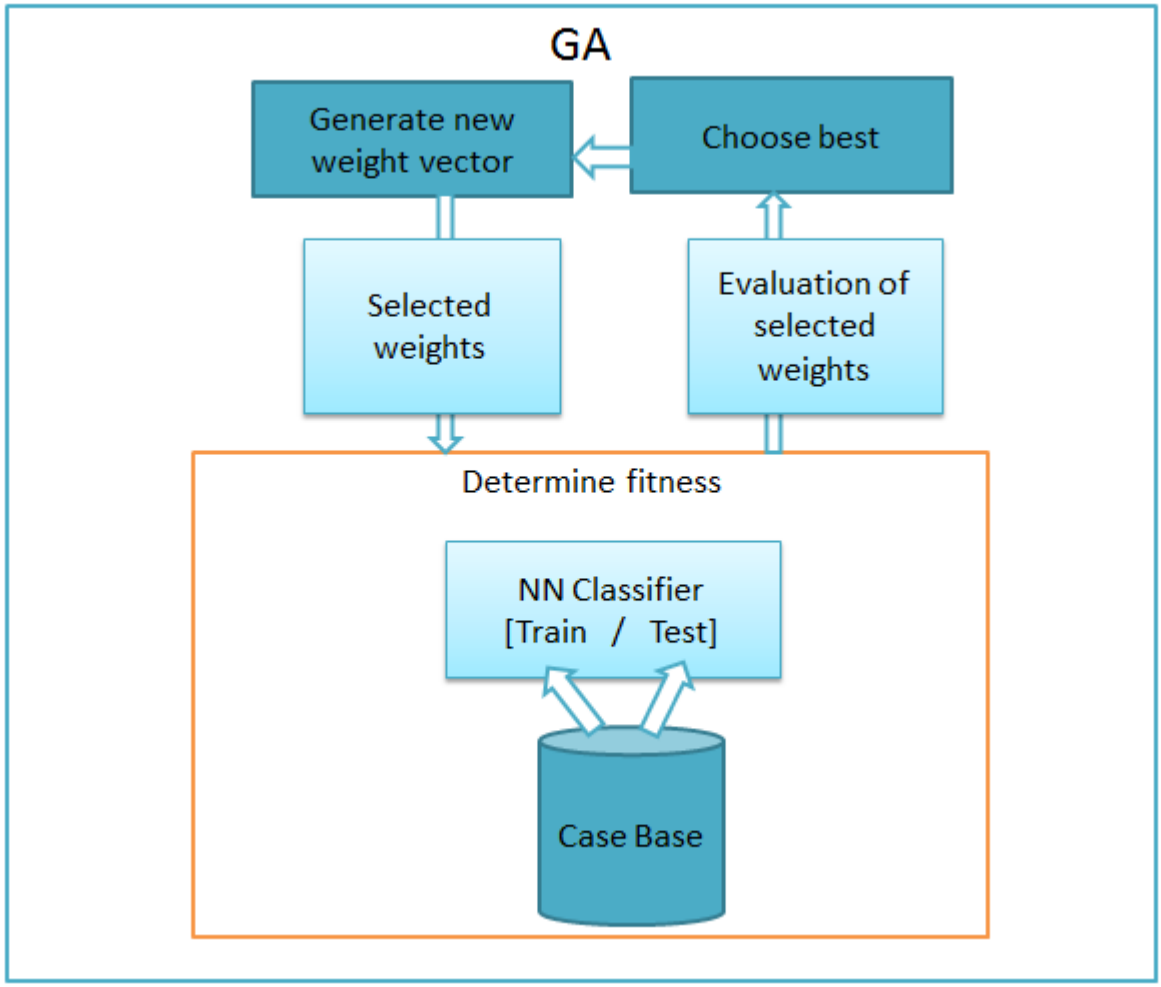

Figure 6.1: Weight Learning GA and 5NN

The methodology is implemented using the Matlab Global Optimization Toolbox.

The parameters for the Genetic Algorithm for this task are:

1. Number of Variables: 50
2. Aeq: ones(1,50)
3. Beq: 1
4. Lower: zeros(1,50)
5. Upper: ones(1,50)
6. Population Size: 100
7. Number of Generations: 50
8. Probability of Crossover: 0.8
9. Crossover Function: Scattered
10. Other Parameters: Default

The program runs for 50 generations with a population size of 100 individuals. Solutions from one population are taken and used to form a new population. In each iteration, if the selected population forms a solution with a better fitness value, population will be converged to be the relevant weights of feature. Figure 6.2 shows the running process of the GA by plotting the fitness value of each generation with the best weight for each individual. The program stops at the 50th iteration and the best individual is then selected as an encoding weight for the feature.

t-test also was conducted to determine whether accuracy differences between GA approach and the un-weighted 5NN are significant or not. The p-value ($3.53E-008$)

of the presented t-test indicates a significant performance improvement in the accuracy. The results of the one-tailed t-test indicates that GA approach outperform un-weighted 5NN. The average accuracy for the eight-fold cross validation increases from 0.7 to 0.96. Table 6.4 shows classification performance of the test dataset applied on the weighted 5-nearest neighbour classifier.

Obviously, this hypothesis is supported in this thesis as it leads to a significant enhancement in the classification process. The GA wrapper method provided good feature weights that substantially improve the performance of the nearest-neighbor classifier. These results outlined in Table 6.4 indicate that assigning different weights to features in the domain of gene expression datasets improves the classification accuracy of the nearest neighbour algorithm.

Table 6.4: Classification performance results of the test data applied on the weighted-5NN classifier

	Predicted High	Predicted Medium	Predicted Standard
Actual High	4	0	1
Actual Medium	0	25	0
Actual Standard	0	1	9

As the nearest neighbor is considered as a white box classifier that allows the user to look deeply in the classification outputs. The analysis of the obtained result reveals that the classification probability for some patients is not stable enough as some patients have two classifications with the same probabilities but they are classified correctly based on the score voting of its similarity to the neighbours samples. Classification probabilities for each patient are calculated and presented in Table 6.5. The probabilities are computed for each patient in the test dataset based on the five retrieved patients. For example, if the five retrieved patients for a such high risk patient in the test dataset are as follows: two medium, one standard and two high risk,

then the classification probabilities for this patient are 0.4 , 0.2 and 0.4 respectively. As can be seen from this table, some patients and especially patients in the minority class are hardly classified with the actual category. These results indicate that the imbalanced classes problem affects the classification performance of the minority class. Therefore oversampling may be required to enhance the classification performance of the minority class.

6.2.4 Oversampling

Most of the classification techniques assume that training samples are evenly distributed among different categories. However, in practical applications, datasets often exist in an unbalanced form. In addition to that, gene expression datasets also have low number of samples. With these kinds of datasets, a poor classification performance is achieved and it results trivial classifiers that completely ignore the minority class. Based on the literature review in chapter 2, there are two different methods to address this problem: sampling techniques and cost sensitive learning. With feature selection process, cost sensitive learning is employed to solve the imbalanced class problem because sampling techniques were not applicable during feature selection process due to the large number of irrelevant features and their effect on the distance measurement. However, at this stage, these problems have been resolved in chapter 4 and 5. Consequently, sampling techniques become very useful to be applied on the gene expression datasets in order to solve the problem of imbalanced classes and low number of samples.

Synthetic Minority Over-sampling Technique (SMOTE) [19] is used in this thesis to oversample the minority class by introducing synthetic samples. Generally, SMOTE is solely applied on the minority classes in order to increase their sample size to attain the same number of samples in the majority class. According to the results reported in the confusion matrix (Table 6.6), this technique of oversampling results intersection between classes boundaries so that some samples in the majority class become inside the decision region of the minority class. All patients in the minority class are predicted correctly but some patients from the majority class, which used to be correctly recognized by the classifier, fall in the wrong decision region (see Table 6.6). This might be because SMOTE is sensitive to complex datasets such as multi-classes and more than one minority class [91].

An approach is suggested to avoid the intersection between different classes. Minority classes are over-sampled at 30%, 50%, 100%, 200%, 300% and 400%. The best accuracy is achieved at 100%. The accuracy becomes stable if the minority classes are over-sampled at greater than 100% (Table 6.7). This methodology significantly increases the performance of the classification especially for the minority samples (see Table 6.8).

Oversampling the training dataset consistently provides an improvement in classification of test data. Moreover, it provides a more stable classifier in case of imbalanced classes. The confusion matrix and probabilities of the 5NN classifier are presented in Table 6.8 and Table 6.9 respectively.

For instance, SMOTE does not shows a major improvements in the classification performance. However, the results become more stable as the classification probability becomes higher than before when the weighed nearest neighbor classifier is used

without oversampling the training dataset (see Table 6.9).

6.3 Summary

A nearest neighbour classifier is proposed in this chapter for gene expression similarity measurements and case base retrieval. A Weight Learning Genetic Algorithm, contribution 4 of this thesis listed in section 1.3, is proposed for feature weighting in the nearest neighbour classifier. The weighted nearest neighbour classifier is successfully applied and enhances the classification performance. The results show that the a Weight Learning Genetic algorithm improved the un-weighted nearest neighbour algorithm. Introducing weights to the features in the nearest neighbour algorithm leads to improvement in the classification performance. SMOTE approach also provides an improvement in the classification of imbalanced class datasets. with 100% oversampled, minority and majority classes are fully recognized by the classifier and the intersections between the classes regions are avoided.

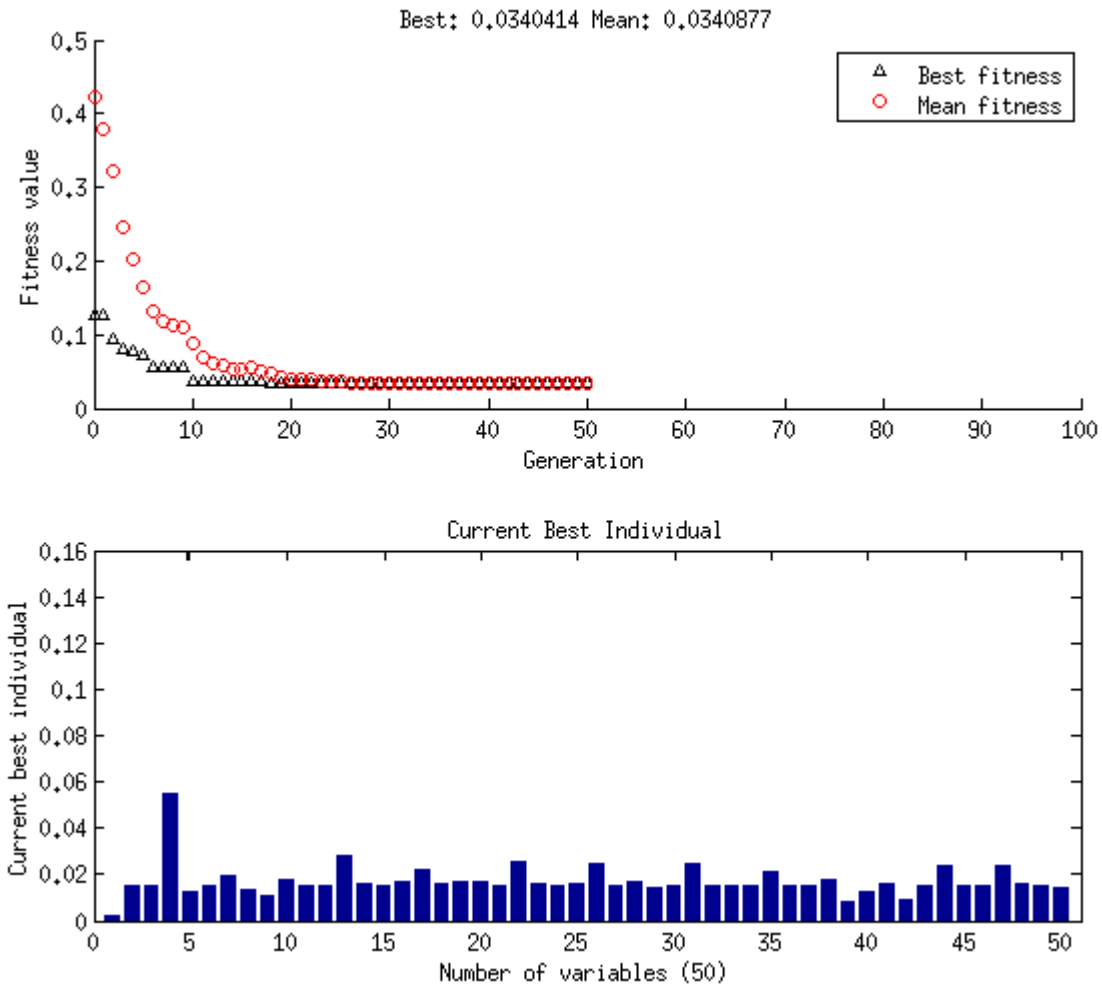

Figure 6.2: Fitness value for each generation

Table 6.5: Classification probability results of the test dataset

	Actual	Predicted High	Predicted Medium	Predicted Standard
1	High	0.4	0.2	0.4
2	High	0.4	0.4	0.2
3	High	0.4	0.4	0.2
4	High	0.8	0.2	0
5	High	0.2	0	0.8
6	Medium	0	1	0
7	Medium	0	1	0
8	Medium	0	0.6	0.4
9	Medium	0	0.8	0.2
10	Medium	0	1	0
11	Medium	0	0.6	0.4
12	Medium	0	1	0
13	Medium	0	0.6	0.4
14	Medium	0	0.6	0.4
15	Medium	0	0.6	0.4
16	Medium	0	0.8	0.2
17	Medium	0	1	0
18	Medium	0	0.8	0.2
19	Medium	0	1	0
20	Medium	0	0.8	0.2
21	Medium	0	1	0
22	Medium	0	0.6	0.4
23	Medium	0	1	0
24	Medium	0	1	0
25	Medium	0	1	0
26	Medium	0	0.6	0.4
27	Medium	0	0.6	0.4
28	Medium	0	0.8	0.2
29	Medium	0	0.8	0.2
30	Medium	0	0.8	0.2
31	Standard	0	0.2	0.8
32	Standard	0	0	1
33	Standard	0	0	1
34	Standard	0	0	1
35	Standard	0.2	0.2	0.6
36	Standard	0.2	0.2	0.6
37	Standard	0.2	0.2	0.6
38	Standard	0.2	0.2	0.6
39	Standard	0.2	0.4	0.4
40	Standard	0.2	0.4	0.4

Table 6.6: Classification performance of the one-side over-sampled

	Predicted High	Predicted Medium	Predicted Standard
Actual High	53	0	0
Actual Medium	7	33	13
Standard	0	0	53

Table 6.7: Classification performance of 100% over-sampled training dataset

	Predicted High	Predicted Medium	Predicted Standard
Actual High	18	0	0
Actual Medium	0	53	0
Actual Standard	0	0	33

Table 6.8: Classification performance results of the test dataset after 100% over-sampled the training dataset

	Predicted High	Predicted Medium	Predicted Standard
Actual High	4	0	1
Actual Medium	0	25	0
Actual Standard	0	0	10

Table 6.9: Classification probability results of the test dataset after SMOTE

	Actual	Predicted High	Predicted Medium	Predicted Standard
1	High	0.8	0.2	0
2	High	1	0	0
3	High	1	0	0
4	High	0.8	0.2	0
5	High	0.4	0	0.6
6	Medium	0	1	0
7	Medium	0	1	0
8	Medium	0	0.8	0.2
9	Medium	0	0.8	0.2
10	Medium	0	1	0
11	Medium	0	0.6	0.4
12	Medium	0	1	0
13	Medium	0	0.8	0.2
14	Medium	0	1	0
15	Medium	0	0.6	0.4
16	Medium	0	0.8	0.2
17	Medium	0	1	0
18	Medium	0	0.8	0.2
19	Medium	0	1	0
20	Medium	0	0.8	0.2
21	Medium	0	1	0
22	Medium	0	0.6	0.4
23	Medium	0	1	0
24	Medium	0	1	0
25	Medium	0	1	0
26	Medium	0	0.6	0.4
27	Medium	0	0.6	0.4
28	Medium	0	1	0
29	Medium	0	1	0
30	Medium	0	0.8	0.2
31	Standard	0	0.2	0.8
32	Standard	0	0	1
33	Standard	0	0	1
34	Standard	0	0	1
35	Standard	0.2	0	0.8
36	Standard	0	0.2	0.8
37	Standard	0.1	0.1	0.8
38	Standard	0.2	0.2	0.6
39	Standard	0.2	0.2	0.6
40	Standard	0.4	0	0.6

Chapter 7

Conclusion

The goal of this research was to develop a clinical tool that will help biologists in how they can predict the risk of relapse for childhood leukaemia sufferers by comparing them to previous patients based on the patient's gene expression data. In order to achieve this goal, this thesis proposed a case-base retrieval framework for retrieval process which represents the main challenge and the key process for a successful case-based reasoning system.

The main data used for this study was collected from The Children's Hospital at Westmead. The data was produced from microarray analysis of childhood leukaemia gene expression values (22678) for 110 patients and it is obtained on the Affymetrix U133 platform. These kinds of data are complex and highly dimensional, which raises many challenges in terms of the analysis, similarity measurement and classification tasks. Five main research questions have been addressed in this thesis: (i) how to build a framework for similarity measurement and case base retrieval for gene expression datasets; (ii) how to select the bio-markers that are strongly associated with the cancer disease; (iii) how to transform the data into lower dimensional space and visualize it; (iv) how to estimate the relative influence of each feature with respect

to the classification performance and (v) how to handle the problems of imbalanced classes and low number of observations.

7.1 Contribution 1: Case-Base Retrieval Framework

An innovative case-base retrieval framework for gene expression data, introduced in section 3.2, is presented to address the first research question. Two modules are presented in this framework: module 1 pre-processes the training dataset to select the relevant features, to reduce the dimensionality, to weight the features and to oversample the training dataset. Module 2 processes the new query data using the outputs of the training steps in module 1 such as feature list, low dimensional vector and weight list.

7.2 Contribution 2: Balanced Iterative Random Forest for Feature Selection

Data mining techniques are used to develop a Balanced Iterative Random Forest algorithm, introduced in section 4.2, to address the second research question that related to feature selection. Feature selection, as one of the most important processes, has been considered carefully in this study. It is unrealistic to assume that the feature selection algorithm, in this case the balanced iterative random forest algorithm, would be able to identify all of the biologically significant genes with such a large complex

and imbalanced dataset. This thesis shows that feature selection process is done in an intelligent way especially the way the imbalanced and over-fitting problems are handled and when the feature selection process is evaluated through reducing the dataset into several subsets of varying sizes to see whether the feature selection process was over-fitted or not. In this thesis, attributes which appeared in multiple subsets are considered as the most informative and predictive genes. Although feature selection process was unable to identify which genes are particularly responsible to define the patient's risk category, a small subset, containing the most informative genes, results from this process.

Balanced Iterative Random Forest is validated on four microarray datasets: childhood Leukaemia, Golub, Colon cancer and Lung cancer datasets. Overall, BIRF resulted in a classifier comparable or superior in accuracy to SVM-RFE, MSVM-RFE and Naive Bayes on the Colon, Golub and Lung datasets.

7.3 Contribution 3: Local Principal Component Analysis for Dimensionality Reduction and Visualisation

The third research question is addressed by introducing Local Principal Component algorithm, introduced in section 5.3.1, for visualization of gene expression datasets and by using kernel principal components analysis to transform the dataset into lower dimensional space. This thesis shows that the effect of highly dimensional data on the similarity measurement is avoided by transferring a high dimensional data into a

lower dimensional space through choosing components with high eigenvalues.

7.4 Contribution 4: Weight Learning Genetic Algorithm for Feature Weighting

This thesis also has used machine learning techniques to develop a weight learning genetic algorithm to address the fourth research question relates to feature weighting. A wrapper method based genetic algorithm, introduced in section 6.2.2.2, is proposed to tune feature weights for accurate nearest-neighbour retrieval. Basic settings of genetic algorithm with efficient fitness function achieved significant improvements with feature weighting.

Furthermore, SMOTE approach is used in this thesis to improve the accuracy of nearest-neighbour classifiers for minority class. SMOTE is used in this thesis in order to address the last research question of this thesis by solving the problems of imbalanced classes and low number of observations. SMOTE was tested on a variety of number of extra samples in the minority classes. Oversampling the minority classes at 100% provides improvement in the decision regions of the classes and performs better than oversampling the minority classes to attain the same number of samples in the majority class.

7.5 Future Work

The ultimate goal of this thesis is to be able to create a clinical tool which can be used to assist clinicians in how to predict the cancer sufferers will react to treatment by comparing them to other previous classified patients. Based on the retrieved similar patients, appropriate risk category is then assigned to the new patient so that successful treatment which applied on the previous patients will be targeted to a new patient. If the similar patients received a particular treatment and survived, then it would make sense to clinicians prescribe this treatment for the new patient otherwise it may be wise to explore another treatment for the new patient.

This study is a step in a new direction of using case-based reasoning system for diagnosis of ALL patient's based on their gene expression data. Data mining, machine learning and statistical techniques have proved to be useful and powerful in this research. Although case-base retrieval was able to identify the similarity between the previous patients and new patients, future work is required on this research project in order to create a domain knowledge for cancer patients. Firstly, this project needs to incorporate other genetic and clinical information in the case-based reasoning system so that more informed decisions about a new patient can be made. It is important to include other datasets such as single-nucleotide polymorphism data in the case base retrieval process. Thus we can have an ensemble datasets classifier for more trusted prediction in the cancer field. Also it is important to include clinical data in the case base reasoning and make comparison between the patients based on their clinical data along with their genetic data.

Secondly, it is also important to do gene selection based on another specific trait of output such as relapse and non-relapse patients. In this project genes are selected

based on the patient's risk of relapse. Moreover, it is good idea to complete the four processes of case-based reasoning system and develop a methodology for case base revising and maintenance.

This research project can be applied to many other types of diseases. As we have seen, feature selection process is applied on different microarray datasets with different diseases. It is important to take advantage of the wealth of knowledge hidden in these datasets and create a domain knowledge which will lead to more informed treatment decisions resulting in a higher percentage of individuals who survive cancer disease.

Author's publications:

- A. Anaissi, P.J. Kennedy, and M. Goyal. Feature selection of imbalanced gene expression microarray data. In Software Engineering, Artificial Intelligence, 12th ACIS International Conference on Networking and Parallel/Distributed Computing (SNPD), pages 73−78. IEEE, 2011.

- A. Anaissi, P.J. Kennedy, and M. Goyal, A framework for high dimensional data reduction in the microarray domain, Communications in Mathematical and in Computer Chemistry/MATCH (2011).

- A. Anaissi, P.J. Kennedy, and M. Goyal, Dimension reduction of microarray data based on local principal component. International Conference on Mathematical and Computational Biology (ICMCB). WASET, 2011.

- A. Anaissi, P.J. Kennedy, and M. Goyal. A framework for high dimensional data reduction in the microarray domain. International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth , pages 903−907. IEEE, 2010.

- A. Anaissi, P.J. Kennedy, and M. Goyal. Similarity-based retrieval for gene expression data, Bioinformatics and Biology Insights, 2012 (submitted).

- A. Anaissi, P.J. Kennedy, and M. Goyal. Balanced recursive random forest for genes selection, Bioinformatics and Biology Insights, 2012 (submitted).

Bibliography

[1] R. Aebersold, M. Mann, et al., *Mass spectrometry-based proteomics*, Nature **422** (2003), no. 6928, 198–207.

[2] D.B. Allison, X. Cui, G.P. Page, and M. Sabripour, *Microarray data analysis: from disarray to consolidation and consensus*, Nature Reviews Genetics **7** (2006), no. 1, 55–65.

[3] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proceedings of the National Academy of Sciences **96** (1999), no. 12, 6745.

[4] K.D. Althoff, R. Bergmann, S. Wess, M. Manago, E. Auriol, O.I. Larichev, A. Bolotov, Y.I. Zhuravlev, and S.I. Gurov, *Case-based reasoning for medical decision support tasks: the inreca approach*, Artificial Intelligence in Medicine **12** (1998), no. 1, 25–41.

[5] E. Andersen, *The irises of the gasp e peninsula*, Bulletin of the American Iris Society **59** (1935), 2–5.

[6] K.J. Archer and R.V. Kimes, *Empirical characterization of random forest variable importance measures*, Computational Statistics & Data Analysis **52** (2008), no. 4, 2249–2260.

[7] N. Arshadi and I. Jurisica, *Maintaining case-based reasoning systems: a machine learning approach*, Advances in Case-Based Reasoning (2004), 439–520.

[8] P. Baldi and G.W. Hatfield, *Dna microarrays and gene expression: from experiments to data analysis and modeling*, Cambridge University Press, 2002.

[9] R. Bellman, *Dynamic programming and lagrange multipliers*, Proceedings of the National Academy of Sciences of the United States of America **42** (1956), no. 10, 767.

[10] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, *Tissue classification with gene expression profiles*, Journal of Computational Biology **7** (2000), no. 3-4, 559–583.

[11] Y. Bengio, J.F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, *Out-of-sample extensions for LLE, ISOMAP, MDS, eigenmaps, and spectral clustering*, Advances in neural information processing systems **16** (2004), 177–184.

[12] C.L. Blake and C.J. Merz, *Uci repository of machine learning databases [http://www. ics. uci. edu/˜ mlearn/mlrepository. html]. irvine, ca: University of california*, Department of Information and Computer Science **55** (1998).

[13] A.L. Blum and P. Langley, *Selection of relevant features and examples in machine learning*, Artificial intelligence **97** (1997), no. 1-2, 245–271.

[14] A.P. Bradley, *The use of the area under the roc curve in the evaluation of machine learning algorithms*, Pattern recognition **30** (1997), no. 7, 1145–1159.

[15] A. Brazma and J. Vilo, *Gene expression data analysis*, FEBS letters **480** (2000), no. 1, 17–24.

[16] L. Breiman, *Random forests*, Machine learning **45** (2001), no. 1, 5–32.

[17] S.P. Chao, C.L. Yen, and C.C. Kuo, *Neural ISOMAP*, Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing, vol. 1, IEEE, 2007, pp. 333–336.

[18] N. Chawla, A. Lazarevic, L. Hall, and K. Bowyer, *Smoteboost: Improving prediction of the minority class in boosting*, Knowledge Discovery in Databases: PKDD 2003 (2003), 107–119.

[19] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, *Smote: synthetic minority over-sampling technique*, Arxiv preprint arXiv:1106.1813 (2011).

[20] M. Cho and H. Park, *Nonlinear dimension reduction using ISOMAP based on class information*, International Joint Conference on Neural Networks, IEEE, 2009, pp. 566–570.

[21] T. Cover and P. Hart, *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory **13** (1967), no. 1, 21–27.

[22] S. De Backer, A. Naud, and P. Scheunders, *Non-linear dimensionality reduction techniques for unsupervised feature extraction*, Pattern Recognition Letters **19** (1998), no. 8, 711–720.

[23] F. Díaz, F. Fdez-Riverola, and J.M. Corchado, *gene-CBR: A case-based reasoning tool for cancer dignosis using microarray data sets*, Computational Intelligence **22** (2006), no. 3-4, 254–268.

[24] T. Dietterich, *Ensemble methods in machine learning*, Multiple classifier systems (2000), 1–15.

[25] C. Ding and H. Peng, *Minimum redundancy feature selection from microarray gene expression data*, Bioinformatics Conference, IEEE, 2003, pp. 523–528.

[26] P. Domingos, *Metacost: A general method for making classifiers cost-sensitive*, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1999, pp. 155–164.

[27] C. Drummond, R.C. Holte, et al., *C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling*, Workshop on Learning from Imbalanced Datasets II, Citeseer, 2003.

[28] K.B. Duan, J.C. Rajapakse, H. Wang, and F. Azuaje, *Multiple SVM-RFE for gene selection in cancer classification with expression data*, IEEE Transactions on NanoBioscience **4** (2005), no. 3, 228–234.

[29] H.G. Einsiedel, A. von Stackelberg, R. Hartmann, R. Fengler, M. Schrappe, G. Janka-Schaub, G. Mann, K. Hählen, U. Göbel, T. Klingebiel, et al., *Long-term outcome in children with relapsed ALL by risk-stratified salvage therapy: results of trial acute lymphoblastic leukemia-relapse study of the berlin-frankfurt-münster group 87*, Journal of Clinical Oncology **23** (2005), no. 31, 7942–7950.

[30] C. Elkan, *The foundations of cost-sensitive learning*, International Joint Conference on Artificial Intelligence, vol. 17, Citeseer, 2001, pp. 973–978.

[31] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery in databases*, AI magazine **17** (1996), no. 3, 37.

[32] K.P. FRS, *Liii. on lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2** (1901), no. 11, 559–572.

[33] M. Gen and Y. Li, *Spanning tree-based genetic algorithm for the bicriteria fixed charge transportation problem*, Proceedings of the 1999 Congress on Evolutionary Computation, vol. 3, IEEE, 1999.

[34] G. Getz, E. Levine, and E. Domany, *Coupled two-way clustering analysis of gene microarray data*, Proceedings of the National Academy of Sciences **97** (2000), no. 22, 12079.

[35] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science **286** (1999), no. 5439, 531–537.

[36] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, and R. Bueno, *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma*, Cancer research **62** (2002), no. 17, 4963.

[37] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, The Journal of Machine Learning Research **3** (2003), 1157–1182.

[38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene selection for cancer classification using support vector machines*, Machine learning **46** (2002), no. 1, 389–422.

[39] C. He, Z. Dong, R. Li, and Y. Zhong, *Dimensionality reduction for text using LLE*, International Conference on Natural Language Processing and Knowledge Engineering, IEEE, 2008, pp. 1–7.

[40] J.H. Holland, *Genetic algorithms*, Scientific american **267** (1992), no. 1, 66–72.

[41] I. Inza, P. Larrañaga, R. Blanco, and A.J. Cerrolaza, *Filter versus wrapper gene selection approaches in DNA microarray domains*, Artificial intelligence in medicine **31** (2004), no. 2, 91–103.

[42] I. Inza, B. Sierra, R. Blanco, and P. Larrañaga, *Gene selection by sequential search wrapper approaches in microarray cancer class prediction*, Journal of Intelligent & Fuzzy Systems **12** (2002), no. 1, 25–33.

[43] P. Jafari and F. Azuaje, *An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors*, BMC Medical Informatics and Decision Making **6** (2006), no. 1, 27.

[44] J. Jarmulak and S. Craw, *Genetic algorithms for feature selection and weighting*, In Proceedings of the IJCAI, vol. 99, Citeseer, 1999, pp. 28–33.

[45] T. Jirapech-Umpai and S. Aitken, *Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes*, BMC bioinformatics **6** (2005), no. 1, 148.

[46] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén, *Trustworthiness and metrics in visualizing similarity of gene expression*, BMC bioinformatics **4** (2003), no. 1, 48.

[47] P. Kennedy, S. Simoff, D. Skillicorn, and D. Catchpoole, *Extracting and explaining biological knowledge in microarray data*, Advances in Knowledge Discovery and Data Mining (2004), 699–703.

[48] R. Kohavi and G.H. John, *Wrappers for feature subset selection*, Artificial intelligence **97** (1997), no. 1-2, 273–324.

[49] J.L. Kolodner, *An introduction to case-based reasoning*, Artificial Intelligence Review **6** (1992), no. 1, 3–34.

[50] J.B. Kruskal, *Nonmetric multidimensional scaling: a numerical method*, Psychometrika **29** (1964), no. 2, 115–129.

[51] M. Kubat and S. Matwin, *Addressing the curse of imbalanced training sets: one-sided selection*, Machine learning-international workshop, Morgan kaufmann publishers, 1997, pp. 179–186.

[52] L.J. Lancashire, C. Lemetre, and G.R. Ball, *An introduction to artificial neural networks in bioinformaticsapplication to complex microarray and mass spectrometry datasets in cancer studies*, Briefings in bioinformatics **10** (2009), no. 3, 315–329.

[53] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, et al., *Machine learning in bioinformatics*, Briefings in bioinformatics **7** (2006), no. 1, 86–112.

[54] D.B. Leake, *Case-based reasoning*, The knowledge engineering review **9** (1994), no. 01, 61–64.

[55] ———, *Case-based reasoning*, John Wiley and Sons Ltd., 2003.

[56] J.A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*, Springer Verlag, 2007.

[57] D.D. Lewis and J. Catlett, *Heterogeneous uncertainty sampling for supervised learning*, Proceedings of the eleventh international conference on machine learning, 1994, pp. 148–156.

[58] J. Li and L. Wong, *Using rules to analyse bio-medical data: A comparison between c4. 5 and pcl*, Advances in Web-Age Information Management (2003), 254–265.

[59] L. Li, C.R. Weinberg, T.A. Darden, and L.G. Pedersen, *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method*, Bioinformatics **17** (2001), no. 12, 1131–1142.

[60] X. Li and L. Shu, *Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis*, Expert Systems with Applications **36** (2009), no. 4, 7644–7650.

[61] JG Liao and K.V. Chin, *Logistic regression for disease classification using microarray data: model selection in a large p and small n case*, Bioinformatics **23** (2007), no. 15, 1945–1951.

[62] J. Lieber and B. Bresson, *Case-based reasoning for breast cancer treatment decision helping*, Advances in Case-Based Reasoning (2000), 1–10.

[63] W. Liu and S. Chawla, *Class confidence weighted kNN algorithms for imbalanced data sets*, Advances in Knowledge Discovery and Data Mining (2011), 345–356.

[64] X. Lu, Y. Wang, and A.K. Jain, *Combining classifiers for face recognition*, Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on, vol. 3, IEEE, 2003, pp. III–13.

[65] S. Ma, J. Li, and D. Liu, *The case retrieval strategy based on hierarchical clustering*, Second Pacific-Asia Conference on Web Mining and Web-based Application, IEEE, 2009, pp. 81–85.

[66] C. Marling and P. Whitehouse, *Case-based reasoning in the care of alzheimers disease patients*, Case-based reasoning research and development (2001), 702–715.

[67] JS Marron, M.J. Todd, and J. Ahn, *Distance-weighted discrimination*, Journal of the American Statistical Association **102** (2007), no. 480, 1267–1271.

[68] M.L. Marx and RJ LARSEN, *Introduction to mathematical statistics and its applications*, Pearson/Prentice Hall, 2006.

[69] N. Meydan, T. Grunberger, H. Dadi, M. Shahar, E. Arpaia, Z. Lapidot, J.S. Leeder, M. Freedman, A. Cohen, A. Gazit, et al., *Inhibition of acute lymphoblastic leukaemia by a jak-2 inhibitor*, (1996).

[70] Z. Michalewicz, *Genetic algorithms+ data structures= evolution programs*, springer, 1998.

[71] CH Ooi and P. Tan, *Genetic algorithms applied to multi-class prediction for the analysis of gene expression data*, Bioinformatics **19** (2003), no. 1, 37–44.

[72] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk, *Reducing misclassification costs*, Proceedings of the Eleventh International Conference on Machine Learning, 1994, pp. 217–225.

[73] C.H. Pui, *Acute lymphoblastic leukemia*, Childhood leukaemia. Second edition. Ed. Memphis Tennessee, USA: St. Jude Childrens Research Hospital (2006), 439–472.

[74] J.R. Quinlan, *Induction of decision trees*, Machine learning **1** (1986), no. 1, 81–106.

[75] ———, *C4. 5: programs for machine learning*, Morgan kaufmann, 1993.

[76] M. Robnik-Šikonja and I. Kononenko, *Theoretical and empirical analysis of ReliefF and RReliefF*, Machine learning **53** (2003), no. 1, 23–69.

[77] S.T. Roweis and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science **290** (2000), no. 5500, 2323–2326.

[78] Y. Saeys, I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics **23** (2007), no. 19, 2507–2517.

[79] H. Saigo and K. Tsuda, *Iterative subgraph mining for principal component analysis*, Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 1007–1012.

[80] J.W. Sammon Jr, *A nonlinear mapping for data structure analysis*, IEEE Transactions on Computers **100** (1969), no. 5, 401–409.

[81] B. Schölkopf, A. Smola, and K.R. Müller, *Kernel principal component analysis*, Artificial Neural NetworksICANN'97 (1997), 583–588.

[82] B. Smyth and M.T. Keane, *Adaptation-guided retrieval: questioning the similarity assumption in reasoning*, Artificial Intelligence **102** (1998), no. 2, 249–293.

[83] R.H. Stottler, A.L. Henke, and J.A. King, *Rapid retrieval algorithms for case-based reasoning*, International Joint Conference on Artificial Intelligence, vol. 11, Citeseer, 1989, pp. 233–237.

[84] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. de Grassi, C. Lee, et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes*, Science **315** (2007), no. 5813, 848–853.

[85] Y. Su, TM Murali, V. Pavlovic, M. Schaffer, and S. Kasif, *Rankgene: identification of diagnostic genes based on expression data*, Bioinformatics **19** (2003), no. 12, 1578–1579.

[86] J.B. Tenenbaum, V. De Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), no. 5500, 2319–2323.

[87] S. Thakur, JK Sing, DK Basu, M. Nasipuri, and M. Kundu, *Face recognition using principal component analysis and rbf neural networks*, First International

Conference on Emerging Trends in Engineering and Technology, IEEE, 2008, pp. 695–700.

[88] V.G. Tusher, R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*, Proceedings of the National Academy of Sciences **98** (2001), no. 9, 5116–5121.

[89] J. Venna and S. Kaski, *Visualizing gene interaction graphs with local multidimensional scaling*, (2006).

[90] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, *Non-linear dimensionality reduction techniques for classification and visualization*, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 645–651.

[91] S. Wang and X. Yao, *Diversity analysis on imbalanced data sets by using ensemble models*, Symposium on Computational Intelligence and Data Mining, IEEE, 2009, pp. 324–331.

[92] Y. Wang, F.S. Makedon, J.C. Ford, and J. Pearlman, *HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data*, Bioinformatics **21** (2005), no. 8, 1530–1537.

[93] I. Watson, *Case-based reasoning is a methodology not a technology*, Knowledge-Based Systems **12** (1999), no. 5, 303–308.

[94] S. Wess, K.D. Althoff, and G. Derwand, *Using k-d trees to improve the retrieval step in case-based reasoning*, Topics in Case-Based Reasoning (1994), 167–181.

[95] W. Xu, X. Lifang, Y. Dan, and H. Zhiyan, *Speech visualization based on locally linear embedding (LLE) for the hearing impaired*, International Conference on BioMedical Engineering and Informatics, vol. 2, IEEE, 2008, pp. 502–505.

[96] P. Yang, B.B. Zhou, Z. Zhang, and A.Y. Zomaya, *A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data*, BMC bioinformatics **11** (2010), no. Suppl 1, S5.

[97] H. Zha and Z. Zhang, *Continuum isomap for manifold learnings*, Computational Statistics & Data Analysis **52** (2007), no. 1, 184–200.

[98] Z. Zhang and H. Zha, *Principal manifolds and nonlinear dimensionality reduction via tangent space alignment*, Journal of Shanghai University (English Edition) **8** (2004), no. 4, 406–424.

[99] G. Zhou and M. Gen, *Evolutionary computation on multicriteria production process planning problem*, IEEE International Conference on Evolutionary Computation, IEEE, 1997, pp. 419–424.

[100] G. Zhou, M. Gen, and T. Wu, *A new approach to the degree-constrained minimum spanning tree problem using genetic algorithm*, IEEE International Conference on Systems, Man, and Cybernetics, vol. 4, IEEE, 1996, pp. 2683–2688.

[101] Z.H. Zhou, *Cost-sensitive learning*, Modeling Decision for Artificial Intelligence (2011), 17–18.