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.