Building a robust clinical diagnosis support system for childhood cancer using data mining methods

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Hamid Ghous

in

School of Software
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
JUNE 2016

CERTIFICATE

Date: **JUNE 2016**

Author: **Hamid Ghous**

Title: **Building a robust clinical diagnosis support system for childhood cancer using data mining methods**

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

I am greatly indebted to my supervisor, Dr. Paul Kennedy for his continuous encouragement, advice, help and invaluable suggestions. He is such a nice, generous, helpful and kindhearted person. I feel really happy, comfortable and unconstrained with him during my PhD study. I owe my research achievements to his experienced supervision. Many thanks are also due to my co-supervisor, Dr. Daniel Catchpoole for his valued suggestions and constant support, and for the numerous conversations with him. I appreciate the travel support for attending the international conferences which I received from the School of Software and the Vice-Chancellor's Conference Fund. I wish to thank my fellow research students and the staff of the school for providing various assistance for the completion of this research work.

I would like to thank my family for their understanding and assistance. I also like to thank them for the freedom to study for the long time necessary to complete postgraduate studies. This thesis could not have been completed without the support and encouragement of my family and friends. I would also like to thank Aedan Roberts from The Children's Hospital at Westmead for his support.

To My Family and friends

Table of Contents

List of Tables

List of Figures

Abstract

Progress in understanding core pathways and processes of cancer requires thorough analysis of many coding and noncoding regions of the genome. Data mining and knowledge discovery have been applied to datasets across many industries, including bioinformatics. However, data mining faces a major challenge in its application to bioinformatics: the diversity and dimensionality of biomedical data. The term 'big data' was applied to the clinical domain by Yoo et al. (2014), specifically referring to single nucleotide polymorphism (SNP) and gene expression data. This research thesis focuses on three different types of data: gene-annotations, gene expression and single nucleotide polymorphisms.

Genetic association studies have led to the discovery of single genetic variants associated with common diseases. However, complex diseases are not caused by a single gene acting alone but are the result of complex linear and non-linear interactions among different types of microarray data. In this scenario, a single gene can have a small effect on disease but cannot be the major cause of the disease. For this reason there is a critical need to implement new approaches which take into account linear and non-linear gene-gene and patient-patient interactions that can eventually help in diagnosis and prognosis of complex diseases. Several computational methods have been developed to deal with gene annotations, gene expressions and SNP data of

complex diseases. However, analysis of every gene expression and SNP profile, and finding gene-to-gene relationships, is computationally infeasible because of the high-dimensionality of data. In addition, many computational methods have problems with scaling to large datasets, and with overfitting. Therefore, there is growing interest in applying data mining and machine learning approaches to understand different types of microarray data.

Cancer is the disease that kills the most children in Australia (Torre et al., 2015). Within this thesis, the focus is on childhood Acute Lymphoblastic Leukaemia. Acute Lymphoblastic Leukaemia is the most common childhood malignancy with 24% of all new cancers occurring in children within Australia (Coates et al., 2001). According to the American Cancer Society (2016), a total of 6,590 cases of ALL have been diagnosed across all age groups in USA and the expected deaths are 1,430 in 2016.

The project uses different data mining and visualisation methods applied on different types of biological data: gene annotations, gene expression and SNPs.

This thesis focuses on three main issues in genomic and transcriptomic data studies:

(i) Proposing, implementing and evaluating a novel framework to find functional relationships between genes from gene-annotation data.

(ii) Identifying an optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression.

(iii) Proposing, implementing and evaluating a novel feature selection approach to identify related metabolic pathways in ALL

This thesis proposes, implements and validates an efficient framework to find functional relationships between genes based on gene-annotation data. The framework is built on a binary matrix and a proximity matrix, where the binary matrix contains information related to genes and their functionality, while the proximity matrix shows similarity between different features.The framework retrieves gene functionality information from Gene Ontology (GO), a publicly available database, and visualises the functional related genes using singular value decomposition (SVD). From a simple list of gene-annotations, this thesis retrieves features (i.e Gene Ontology terms) related to each gene and calculates a similarity measure based on the distance between terms in the GO hierarchy. The distance measures are based on hierarchical structure of Gene Ontology and these distance measures are called similarity measures. In this framework, two different similarity measures are applied:

(i) A hop-based similarity measure where the distance is calculated based on the number of links between two terms.

(ii) An information-content similarity measure where the similarity between terms is based on the probability of GO terms in the gene dataset.

This framework also identifies which method performs better among these two similarity measures at identifying functional relationships between genes. Singular value decomposition method is used for visualisation, having the advantage that multiple types of relationships can be visualised simultaneously (gene-to-gene, term-to-term and gene-to-term)

In this thesis a novel framework is developed for visualizing patient-to-patient relationships using gene expression values. The framework builds on the random forest

feature selection method to filter gene expression values and then applies different linear and non-linear machine learning methods to them. The methods used in this framework are Principal Component Analysis (PCA), Kernel Principal Component Analysis (kPCA), Local Linear Embedding (LLE), Stochastic Neighbour Embedding (SNE) and Diffusion Maps. The framework compares these different machine learning methods by tuning different parameters to find the optimal method among them. Area under the curve (AUC) is used to rank the results and SVM is used to classify between relapsed and non-relapsed patients.

The final section of the thesis proposes, implements and validates a framework to find active metabolic pathways in ALL using single nucleotide polymorphism (SNP) profiles. The framework is based on the random forest feature selection method. A collected dataset of ALL patient and healthy controls is constructed and later random forest is applied using different parameters to find highly-ranked SNPs. The credibility of the model is assessed based on the error rate of the confusion matrix and kappa values. Selected high ranked SNPs are used to retrieve metabolic pathways related to ALL from the KEGG metabolic pathways database.

The methodologies and approaches presented in this thesis emphasise the critical role that different types of microarray data play in understanding complex diseases like ALL. The availability of flexible frameworks for the task of disease diagnosis and prognosis, as proposed in this thesis, will play an important role in understanding the genetic basis to common complex diseases.

This thesis contributes to knowledge in two ways:

(i) Providing novel data mining and visualisation frameworks to handle biological data.

(ii) Providing novel visualisations for microarray data to increase understanding of disease.

Author's Publications

(i) H. Ghous, P. J. Kennedy, N. Ho, and D. R. Catchpoole. Comparing functional visualisations of lists of genes using singular value decomposition. Journal of Research and Practice in IT, 2013. Accepted, 02/10/13.

(ii) H. Ghous, N. Ho, D. R. Catchpoole, and P. J. Kennedy. Functional visualisation of genes using singular value decomposition. In Y. Zhao, J. Li, P. J. Kennedy, and P. Christen, editors, Proceedings of the Tenth Australasian Data Mining Conference (AusDM 12), volume 134 of Conferences in Research and Practice in IT (CRPIT), pages 53-59. Australian Computer Society, 2012.

(iii) H. Ghous, N. Ho, D. R. Catchpoole, and P. J. Kennedy. Comparing functional visualizations of genes. In J. Maria Pena and F. Famili, editors, 5th Workshop on Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions, European Conferencce on Machine Learning and Principals and Practice of Knowledge Discovery in Databases 2011, pages 12-21. 5-9 September 2011.

(iv) H. Ghous, P. J. Kennedy, D. R. Catchpoole, and S. J. Simoff. Kernel-based visualisation of genes with the gene ontology. In J. F. Roddick, J. Li, P. Christen, and P. J. Kennedy, editors, Data Mining and Analytics 2008: proceedings of the

Seventh Australasian Data Mining Conference (AusDM08), volume 87 of Conferences in Research and Practice in IT (CRPIT), pages 133-140. Australian Computer Society, Sydney, 2008.

Abbreviations and Glossary

Term	**Abbreviation**	**Definition**
Acute Lymphoblastic Leukaemia	ALL	
Biased data		Data where cases of one class is very high compared to other class
Classification		A supervised learning process of assigning cases to target class
Data transformation		A process to transform or consolidate data for effective use
Diffusion Maps	DM	
Dimensionality reduction		Reducing the number of random variables in a dataset
Gene Ontology	GO	
High dimensionality		Data whose dimension is higher than dimensions considered in traditional multivariate analysis
Kernel Principal Component Analysis	kPCA	
Kyoto Encyclopaedia of Genes and Genomes	KEGG	
Local Linear Embedding	LLE	
Nonlinear Dimensionality Reduction	NLDR	
Prediction		The process of predicting the outcome of certain case
Principal Component Analysis	PCA	
Prognosis		Predicting likely outcome of patient
Random Forest	RF	
Singular Value Decomposition	SVD	
Single Nucleotide Polymorphism	SNPs	
Stochastic Neighbour Embedding	SNE	