UNIVERSITI PUTRA MALAYSIA

INTEGRATED FRAMEWORK WITH ASSOCIATION ANALYSIS FOR GENE SELECTION IN MICROARRAY DATA CLASSIFICATION

ONG HUEY FANG

FSKTM 2011 29

INTEGRATED FRAMEWORK WITH ASSOCIATION ANALYSIS FOR GENE SELECTION IN MICROARRAY DATA CLASSIFICATION

By

ONG HUEY FANG

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

April 2011

© COPYRIGHT UPM

Dedicated to my beloved mother and father

© COPYRIGHT UPM

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

INTEGRATED FRAMEWORK WITH ASSOCIATION ANALYSIS FOR GENE SELECTION IN MICROARRAY DATA CLASSIFICATION

By

ONG HUEY FANG

April 2011

Chairman: **Norwati Mustapha, PhD**

Faculty: **Computer Science and Information Technology**

Microarray data classification is one of the major interests in health informatics that aims at discovering hidden patterns in gene expression profiles. The main challenge in building this classification system is the curse of dimensionality problem. Therefore, gene selection is an indispensable task in microarray data classification to identify smaller sets of relevant genes. However, most of the existing gene selection methods are statistical analyses and purely based on gene expression values in the identification of differentially expressed genes. As a result, the selected genes might be false positives and are not biologically meaningful.

The purpose of this study was to integrate microarray dataset with additional biological information for selecting genes that are not only differentially expressed but also informative for classifiers. To achieve that, an integrated framework with a new gene selection method was developed to improve classification performance in terms of accuracy and number of selected genes. The proposed gene selection method combined

© COPYRIGHT UPM

the strength of both filter method and association analysis to identify a set of discriminative and informative genes. Association analysis was employed to integrate more than one type of biological information in the same transaction database, and also to identify groups of genes that are frequently co-occurred in target samples. Modifications have been made on the existing association algorithm for mining frequent itemsets, where genes in each itemset were sorted according to their discriminative scores rather than according to lexicographic order. In addition to that, discriminative scores were used to compute interestingness of frequent itemsets before ranking them.

The proposed integrated framework has been tested on colon cancer, leukemia, breast cancer and lung cancer microarray datasets. Two types of biological information were incorporated in the selection process, namely the Gene Ontology (GO) and the KEGG Pathways (KEGG). The experimental results showed that the recommended GO based models, KEGG based models, and GO-KEGG based models outperformed the expression-only models by attaining better classification accuracies with less number of genes. In the experiments, leukemia and lung cancer datasets had achieved 100% accuracies in all the classifiers with number of selected genes as small as three. On the other hand, colon cancer and breast cancer datasets achieved better classification accuracies compared with the previous integrated method, which are 95.16% and 95.88% respectively. Moreover, the proposed integrated framework proved to build informative and interpretable microarray classification models. The selected genes can be traced back to their functional annotations and association groups for reasoning and creating new hypotheses for future investigation.

© COPYRIGHT UPM

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

RANGKA KERJA INTEGRATIF DENGAN ANALISIS SEKUTUAN UNTUK PEMILIHAN GEN DALAM PENGKELASAN DATA TATASUSUNAN-MIKRO

Oleh

ONG HUEY FANG

April 2011

Pengerusi: **Norwati Mustapha, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Pengkelasan data tatasusunan-mikro merupakan salah satu tumpuan utama dalam bidang informatik kesihatan yang bertujuan untuk mencari pola-pola tersembunyi di dalam profil ekspresi gen. Cabaran utama dalam pembinaan sistem pengkelasan ini ialah masalah *curse of dimensionality*. Oleh itu, pemilihan gen merupakan satu tugas penting dalam pengkelasan data tatsusunan-mikro untuk mengenalpasti set gen yang lebih kecil dan berkaitan. Namun demikian, sebahagian besar kaedah pemilihan gen yang sedia ada adalah analisis statistik dan berdasarkan sepenuhnya kepada nilai-nilai ekspresi dalam mengenalpasti gen-gen yang mempunyai ekspresi yang berbeza. Akibatnya, gen-gen yang terpilih mungkin *false positive* dan tidak bermakna dari segi biologi.

Tujuan kajian ini adalah untuk mengintegrasikan dataset tatsusunan-mikro dengan maklumat tambahan biologi bagi memilih gen-gen yang bukan sahaja berbeza ekspresi tetapi juga berinformatik untuk pengkelas-pengkelas. Bagi mencapai itu, sebuah rangka kerja integratif dengan satu kaedah pemilihan gen yang baru telah dibangunkan untuk

© COPYRIGHT UPM

meningkatkan prestasi pengkelasan dari segi ketepatan dan jumlah gen yang terpilih. Pendekatan yang dicadangkan menggabungkan kelebihan dari kedua-dua kaedah penapisan dan analisis sekutuan untuk mengenalpasti set-set gen yang diskriminasi dan berinformatik. Analisis sekutuan digunakan untuk mengintegrasikan lebih dari satu jenis maklumat biologi di dalam satu pangkalan data transaksi yang sama, dan juga untuk mengenalpasti kumpulan gen-gen yang sering muncul bersama dalam sampel target. Perubahan telah dilakukan pada algoritma sekutuan yang sedia ada bagi perlombongan *frequent itemsets*, di mana gen-gen di dalam setiap *itemset* akan disusun secara menurun berdasarkan nilai-nilai diskriminasi mereka daripada berdasarkan urutan *lexicographic*. Tambahan lagi, nilai-nilai diskriminasi digunakan untuk mengira nilai *interestingness* bagi setiap *frequent itemset* sebelum menentukan kedudukan mereka.

Rangka kerja integratif yang dicadangkan telah diujikan pada dataset tatasusunan-mikro kanser usus besar, leukemia, kanser payu dara dan kanser paru-paru. Dua jenis maklumat biologi telah digabungkan ke dalam process pemilihan iaitu *Gene Ontology* (GO) dan *KEGG Pathway* (KEGG). Keputusan kajian menunjukkan bahawa cadangan model-model berdasarkan GO, model-model berdasarkan KEGG, dan model-model berdasarkan GO-KEGG mengungguli model-model ekspresi-sahaja dengan mencapai ketepatan pengkelasan yang lebik baik dengan jumlah gen yang lebih kurang. Dalam eksperimen-eksperimen tersebut, dataset leukemia dan kanser paru-paru telah mencapai ketepatan 100% dalam kesemua pengkelas dengan jumlah gen terpilih sekecil tiga. Manakala, dataset kanser usus besar dan kanser payudara telah mencapai ketepatan pengkelasan yang lebik baik berbanding dengan kaedah integratif sebelum ini, dengan masing-masing 95.16% dan 95.88%. Selain itu, pendekatan integratif yang dicadangkan

© COPYRIGHT UPM

terbukti dapat membina model-model pengkelasan tatasusunan-mikro yang berinformatik dan boleh diinterpretasikan. Gen-gen terpilih boleh ditelusuri kembali pada anotasi-anotasi fungsi dan kumpulan-kumpulan sekutuan mereka untuk penalaran dan pembinaan hipotesis-hipotesis baru bagi tujuan penyelidikan pada masa depan.

© COPYRIGHT UPM

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Norwati Mustapha for her patient guidance, encouragement, and valuable advices. Through the course of my study, she was always available to support my research works. I am very fortunate to be supervised by her, as she taught me patiently on ways to improve my research and encourage me to explore more in the research field. Her knowledge in teaching and research has inspired me to become a better student and researcher.

Secondly, I would like to express my sincere thanks and appreciation to the supervisory committee member, Associate Professor Dr. Md. Nasir Sulaiman for his guidance and valuable suggestions in making this work a success. My sincere appreciation also goes to all my friends in UPM for their continuous help and sharing of knowledge.

Finally, I would like to thanks my family for their support and encouragement for me during my study. Without them it is not possible for me to achieve this work.

Ong Huey Fang
April 2011

© COPYRIGHT UPM

APPROVAL

I certify that an Examination Committee has met on 12th April 2011 to conduct the final examination of Ong Huey Fang on her Master of Science thesis entitled "Integrated Framework with Association Analysis for Gene Selection in Microarray Data Classification" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the Master of Science. Members of the Examination Committee are as follows:

Abu Bakar B. Md. Sultan, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Masrah Azrifah Azmi Murad, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Shyamala C. Doraisamy, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Siti Zaiton Mohd Hashim, PhD
Associate Professor
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
(External Examiner)

_____
NORITAH OMAR, PhD
Associate Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

© COPYRIGHT UPM

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Norwati Mustapha, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Md. Nasir B Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

_____

HASANAH MOHD. GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

© COPYRIGHT UPM

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at University Putra Malaysia or other institution.

_____

ONG HUEY FANG

Date: 12 April 2011

© COPYRIGHT UPM

TABLE OF CONTENTS

			Page
DEDICATION			ii
ABSTRACT			iii
ABSTRAK			v
ACKNOWLEDGEMENTS			viii
APPROVAL			ix
DECLARATION			xi
LIST OF TABLES			xv
LIST OF FIGURES			xvi
LIST OF ABBREVIATIONS			xviii
CHAPTER			
1	**INTRODUCTION**		1
	1.1	Motivation and Background	1
	1.2	Problem Statement	4
	1.3	Research Objectives	6
	1.4	Research Scope	7
	1.5	Research Contribution	8
	1.6	Organization of Thesis	8
2	**BACKGROUND**		10
	2.1	Introduction	10
	2.2	Gene Selection for Microarray Classification	10
		2.2.1 Filter Method	11
		2.2.2 Wrapper Method	12
		2.2.3 Embedded Method	13
		2.2.4 Hybrid Method	14
	2.3	Biological Information	15
		2.3.1 Gene Ontology	17
		2.3.2 KEGG Pathway	20
	2.4	Integrated Analysis in Biological Context	21
	2.5	Summary	25

© COPYRIGHT UPM

3	**LITERATURE REVIEW**		26
	3.1	Introduction	26
	3.2	General Scheme for Integrative Analyses	26
	3.3	Integrative Gene Selection for Microarray Classification	28
	3.4	Fundamental of Association Analysis	32
	3.5	Association Analysis Algorithms	35
		3.5.1 Apriori	35
		3.5.2 FP-Growth	36
	3.6	Association Analysis for Microarray Data	38
	3.7	Summary	41
4	**RESEARCH METHODOLOGY**		43
	4.1	Introduction	43
	4.2	Research Steps	43
	4.3	The Knowledge Discovery Process	46
	4.4	System Requirements	48
		4.4.1 Hardware Requirements	48
		4.4.2 Software Requirements	48
		4.4.3 Data Collection	49
		4.4.4 Evaluation Parameters	55
	4.5	Summary	55
5	**INTEGRATED FRAMEWORK WITH ASSOCIATION ANALYSIS**		56
	5.1	Introduction	56
	5.2	Integrated Framework for Microarray Classification	56
		5.2.1 Data Collection	58
		5.2.2 Data Preparation	58
		5.2.3 Data Integration	61
		5.2.4 Gene Selection	63
		5.2.5 Classification	63
	5.3	The Proposed Integrative Gene Selection Method	66
		5.3.1 Filter Method	67
		5.3.2 Association Analysis	68
		5.3.3 Gene Ranking	70
		5.3.4 Gene Selection Algorithm	71

© COPYRIGHT UPM

5.3.5 A Running Example 73
5.4 Summary 81

6 **RESULTS AND DISCUSSIONS** 82
6.1 Introduction 82
6.2 Experimental Remarks 82
6.3 Dataset and Preprocessing 83
6.3.1 Microarray Data 83
6.3.2 Additional Biological Data 85
6.4 Results 86
6.4.1 Results of Colon Cancer Dataset 87
6.4.2 Results of Leukemia Dataset 91
6.4.3 Results of Breast Cancer Dataset 95
6.4.4 Results of Lung Cancer Dataset 99
6.5 Discussion 105
6.6 Summary 106

7 **CONCLUSION AND FUTURE RESEARCH** 107
7.1 Introduction 107
7.2 Research Conclusion 107
7.3 Future Works 109

REFERENCES 112
APPENDICES 118
A Retrieved GO IDs for Genes in Colon Data 118
B Retrieved KEGG IDs for Genes in Colon Data 120
C Discriminative Scores for Genes in Colon Data 134
D KEGG Based Transaction Database for Colon Cancer 147
E GO Terms and KEGG Descriptions for Selected Genes in Colon Data 148
F GO Terms and KEGG Descriptions for Selected Genes in Leukemia Data 152
G GO Terms and KEGG Descriptions for Selected Genes in Breast Data 155
BIODATA OF STUDENT 160

© COPYRIGHT UPM