



UNIVERSITI PUTRA MALAYSIA

***PREDICTION OF BREAST CANCER RELAPSE TIME IN CONTINUOUS
SCALE BASED ON TYPE-2 TSK FUZZY MODEL***

SAYED HAMID MAHMOUDIAN

FK 2010 75

**PREDICTION OF BREAST CANCER RELAPSE TIME IN CONTINUOUS
SCALE BASED ON TYPE-2 TSK FUZZY MODEL**

By

SAYED HAMID MAHMOUDIAN

**Thesis Submitted to the School of Graduate Studies, University Putra Malaysia, in
Fulfilment of the Requirements for the Degree of Doctor of Philosophy**

November 2010

Abstract of thesis presented to the senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

PREDICTION OF BREAST CANCER RELAPSE TIME IN CONTINUOUS SCALE BASED ON TYPE-2 TSK FUZZY MODEL

By

SAYED HAMID MAHMOUDIAN

November 2010

Chairman : Associate Professor Mohammad Hamiruce Marhaban, PhD

Faculty : Engineering

Recently, microarray analysis and gene expression profiles have been widely applied in diagnosis and classification of different types of cancer such as liver, colon or breast cancer. As the number of breast cancer cases increased dramatically in many countries including Malaysia in recent decades, different types of studies have been done to control the disease or reduce the cost of their treatments. Gene expression profiles, which can screen the behavior of a large number of genes simultaneously, have been used in some studies to extract the significant genes related to breast cancer. Tumor classification, Estrogen Receptor status recognition or survival analysis has been usually considered as important objectives in these studies. Due to the fact that studies in survival analysis of breast cancer can reduce the cost of treatments and side effects of the adjuvant therapy,

different methods for predicting the outcome of the disease have been proposed by previous researcher.

The two major objectives of this research are to propose a fuzzy classifier to discriminate breast cancer tumors into two classes, which are high risk and low risk by some interpretable rules similar to linguistic words, and to predict the relapse time of breast cancer by TSK fuzzy models in continuous scale. For this reason, breast cancer dataset has been applied for training the models and two other independent samples have been used for validating the results. In addition, K-fold Cross Validation, B632 and B632+ methods have been used for error estimation.

In the first objective of the thesis, a lemma has been proven and a new hybrid algorithm based on Fuzzy Association Rule Mining has been proposed to gather some selected genes and generate fuzzy rules for classification.

In the second one, a method for generating the fuzzy rules to discriminate the samples of breast cancer into the different groups have been proposed and applied to predict the relapse time of samples in continuous scale while handling the uncertainties in linguistic terms of the rules.

The relapse time of two available independent samples of breast cancer have been predicted by the model and the results show the superiority of the proposed model with respect to the previous study. Finally 46 significant genes and 16 fuzzy rules have been introduced which can be used in a Type-2 TSK fuzzy model as a predictor.



ABSTRAK

Abstrak tesis dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

RAMALAN MASA RELAPS DALAM SKALA SELANJAR BAGI KANSER PAYUDARA BERDASARKAN MODEL KABUR TSK JENIS-2

Oleh

SAYED HAMID MAHMOUDIAN

November 2010

Pengerusi: Profesor Madya Mohammad Hamiruce Marhaban, PhD

Fakulti: Kejuruteraan

Kebelakangan ini, analisa jujukan mikro dan profil ekspresi gen telah digunakan dengan meluas dalam diagnosis dan pengelasan beberapa jenis kanser seperti kanser hati, kolon atau payudara. Disebabkan kes kanser payudara telah meningkat bilangannya dalam dekad ini di kebanyakan negara termasuk Malaysia, pelbagai kajian telah dijalankan untuk mengawal penyakit tersebut atau mengurangkan kos rawatannya. Profil ekspresi gen berupaya untuk mencerminkan tingkah laku gen berjumlah besar dengan serentak, telah digunakan dalam beberapa kajian mengekstrak gen-gen penting yang berkaitan dengan kanser payudara. Pengelasan tumor, pengekaman status penerima estrogen atau analisa peluang hidup biasanya dianggap sebagai objektif utama untuk kajian-kajian ini. Memandangkan kajian tentang peluang hidup daripada kanser payudara boleh

mengurangkan kos rawatan dan kesan sampingan daripada terapi adjuvan, pelbagai kaedah untuk meramal hasil penyakit tersebut telah dicadangkan oleh para penyelidik.

Dua objektif utama kajian ini ialah untuk mencadangkan satu pengelasan kabur dalam membezakan tumor kanser payudara kepada dua kelas iaitu kelas berisiko tinggi dan kelas berisiko rendah, mengikut peraturan yang boleh diinterpretasi, menggunakan kata-kata linguistik serta meramalkan masa relaps bagi kanser payudara menggunakan model kabur TSK dalam skala selanjut.

Untuk itu, set data kanser payudara van't Veer telah digunakan dalam melatih model-model yang dicadangkan serta 2 lagi sampel bebas yang diterbitkan oleh van't Veer dan van de Vijver telah digunakan untuk mengesahkan keputusan yang diperolehi. Kaedah keesahan silang, B632 dan B632+ telah digunakan untuk menganggarkan ralat.

Bagi objektif pertama, satu lemma telah dibuktikan dan satu algoritma hibrid berdasarkan Perlombongan Peraturan Berkaitan Kabur telah dicadangkan untuk mengumpul beberapa gen yang terpilih dan menjana peraturan kabur untuk pengelasan.

Dalam objektif kedua, satu kaedah untuk menjana peraturan kabur dalam membezakan sampel-sampel kanser payudara kepada kumpulan-kumpulan yang berbeza telah dicadangkan dan diaplikasikan untuk meramal masa relaps sampel-sampel tersebut dalam skala selanjur serta mengendali ketidakpastian dalam terma linguistik peraturan-peraturan.

Masa relaps bagi 2 sampel kanser payudara yang bebas telah diramalkan oleh model tersebut dan hasil menunjukkan model yang dicadangkan jauh lebih baik daripada model-model dari kajian sebelum ini. Subset gen-gen yang penting dan peraturan kabur yang dijana untuk ramalan masa relaps juga dikemukakan. Akhirnya, 46 gen penting dan 16 aturan kabur telah dicadangkan yang mana ia boleh digunakan oleh model kabur TSK Jenis-2 sebagai peramal.

ACKNOWLEDGEMENTS

First and foremost, my praise to Allah Jalla Jalaloh, who blessed me with patience, courage, consistency and good health during this study, I would like to express my thanks to the chairman of my supervisory committee Associate Professor Dr Mohammad Hamiruce Marhaban, who helped me to be able to continue this study and gave me great guidance, suggestion and encouragement. My gratitude goes to the member of supervisory committee, Professor Dr Raha Abdul Rahim, Associate Professor Dr. Rozita Rosli and Dr. M. Iqbal Saripan for his kind guidance.

Moreover, I am thankful to the staff in the department of Electrical and Electronic Engineering that helped me to continue this study.

Finally, it is needed to express the heartfelt thanks to my wife who support me in my research and patience, care and encouragement during the study. I am very grateful for my mother, for always being there when I needed her and also my kids Nima and Raha who encourage me to work hard by their beautiful smiles. I would also like to express my gratefulness to my friends for giving me a good time here.

I certify that a Thesis Examination Committee has met on (July 2010) to conduct the final examination of Hamid Mahmoodian on his thesis entitled “Prediction of Breast Cancer Relapse Time in Continuous Scale Based on Type-2 Fuzzy Model” in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Noorhisam b. Misron, PhD

Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Abdul Rahman b. Ramli, PhD

Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Internal Examiner)

Syed Abd. Rahman Al-Attas, PhD

Associate Professor
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
(External Examiner)

Golshah Naghdy, PhD

Associate Professor
Faculty of Electrical Computer and Telecommunications Engineering
University of Wollongong
Country Australia
(External Examiner)

BUJANG KIM HUAT, PhD

Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

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

Mohammad Hamiruce Marhaban, PhD

Associate Professor
Department of Electrical and Electronic Engineering
University Putra Malaysia
(Chairman)

Raha Abdul Rahim, PhD

Professor
Faculty of Biotechnology and Biomolecular Sciences
University Putra Malaysia
(Member)

Rozita Rosli, PhD

Associate Professor
Faculty of Medicine and Health Sciences
Universiti Putra Malaysia
(Member)

M. Iqbal b. Saripan, PhD

Lecturer
Department of Electrical and Electronic Engineering
Universiti Putra Malaysia
(Member)

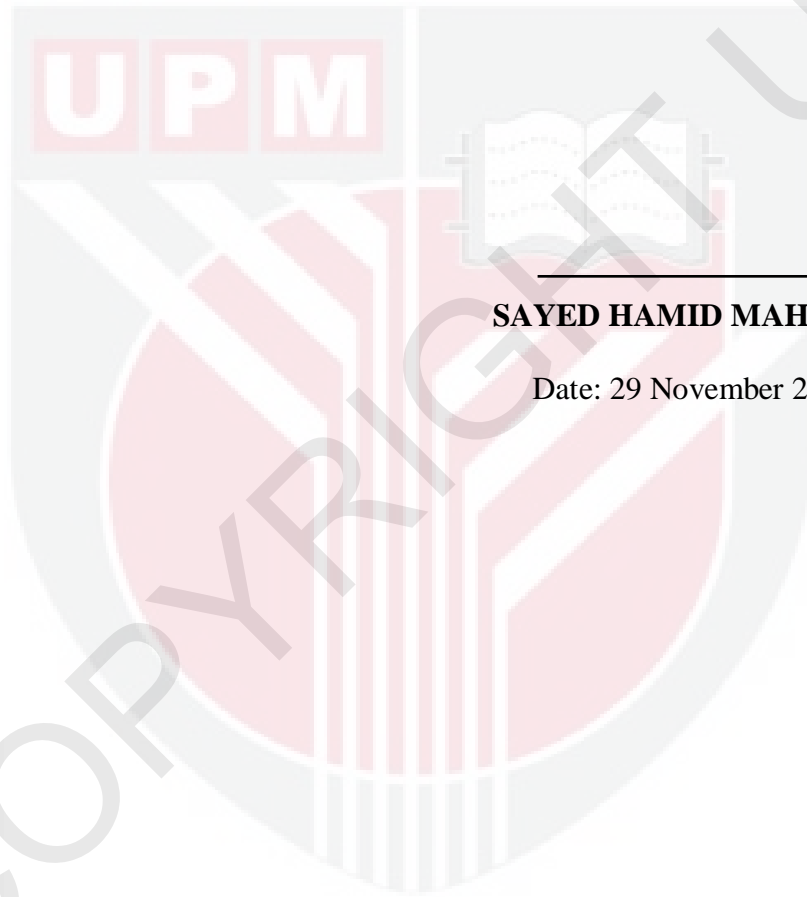
HASANAH MOHD GHAZALI, PhD

Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

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



SAYED HAMID MAHMOODIAN

Date: 29 November 2010



TABLE OF CONTENTS

	Page
ABSTRACT	ii
ABSTRAK	v
ACKNOWLEDGEMENTS	viii
APPROVAL	ix
DECLARATION	xi
LIST OF TABLES	xv
LIST OF FIGURES	xviii
LIST OF ABBREVIATIONS	xxii
CHAPTER	
1 INTRODUCTION	1
1.1 Preface	1
1.2 Motivation and Problem Statement	3
1.3 Aim and Objectives	5
1.4 Scope of the Work	6
1.5 Thesis Contributions	7
2 LITERATURE REVIEW	9
2.1 Microarray Technology and Breast Cancer Analysis	9
2.1.1 Spotted Microarray Technology	11
2.1.2 Oligonucleotide Microarray Technology	12
2.1.3 Microarray analysis	14
2.1.4 Gene Expression Profile in Breast Cancer	15
2.1.5 Previous Studies on Van't Veer dataset	26
2.2 Literature review on Gene Selection, classification and Validation	34
2.2.1 Formulation of Feature Subset Selection	37
2.2.2 Feature Extraction	38
2.2.3 Feature Selection	39
2.2.4 Fisher Gene Selection	41
2.2.5 Pearson Correlation Coefficient	43
2.2.6 Recursive Feature Elimination (RFE)	44
2.2.7 Penalized Logistic Regression (PLR)	47
2.2.8 Validation Procedure and Error Estimation	51
2.2.9 K-fold Cross Validation	51
2.2.10 B632 and B632+	52
2.3 Fuzzy Systems and Fuzzy Classifiers	56
2.3.1 Important Concepts of FLS	57
2.3.2 Mamdani Model	60
2.3.3 TSK model	61

2.3.4	Fuzzy classifier	63
2.4	Rule Mining in Fuzzy Classifier	65
2.4.1	Fuzzy Association Rule Mining (FARM)	67
2.4.2	Rule Mining Based on FARM	70
2.4.3	Rule Mining Based on SVM	72
2.5	Type-2 Fuzzy System	73
2.5.1	Interval Type-2 Fuzzy Logic Systems (IT2-FLS)	75
2.5.2	Fuzzification, Aggregation and Defuzzification of IT2-FLS	77
3	INTRODUCTION TO GENE SELECTION AND A NEW PROPOSED ALGORITHM	80
3.1	Introduction	80
3.2	Gene Selection in Breast Cancer Dataset	80
3.2.1	Selection of Penalty (Regularization) Parameter in PLR	82
3.2.2	Results for PLR classifier	82
3.2.3	Results for SVM Classifier	84
3.3	Proposed Algorithm: Conditional Gene Selection (CGS)	88
3.3.1	The Algorithm	89
3.3.2	Results	92
3.3.3	Colon dataset	93
3.3.4	Breast Cancer Dataset	95
4	BREAST CANCER SURVIVAL ANALYSIS BY FUZZY CLASSIFIER	99
4.1	Introduction	99
4.2	Using FARM in Fuzzy Classifier	100
4.3	Fuzzy Classification of Cancer Datasets	103
4.3.1	Using FARM in Breast Cancer Classification	106
4.3.2	Hybrid Classifier Algorithm	119
4.3.3	Breast Cancer Results with the Proposed Algorithm (Hybrid Algorithm)	123
4.3.4	Colon Cancer Results with Hybrid Algorithm	129
4.4	Discussion on Breast Cancer Results	133
4.4.1	Effect of Confidence Values	138
4.4.2	Genes with Linguistic Term ZE	139
4.5	Conclusions	142
5	RELAPSE TIME PREDICTION	144
5.1	Introduction	144
5.2	Mathematical Framework	146
5.2.1	Confidence Value and Uncertainty	146
5.2.2	Uncertainty Recognition	149
5.3	Generating the Fuzzy Predictor	151
5.3.1	Converting a Classifier Rulebase to a TSK Rulebase	153
5.3.2	Aggregation in FARM41 and SVM41	154

5.4	Rule Mining by Feature Partitioning Method (FPM)	155
5.4.1	Search Algorithm	158
5.4.2	FPM algorithm	163
5.4.3	Performance Criteria	166
5.5	Methodology of Using FPM Algorithm	167
5.5.1	Part 1- Model Selection	168
5.5.2	Part 2- Gene Subset Selection	171
5.5.3	Part 3-Optimization	173
5.5.4	Part 4- Final Tuning	176
5.5.5	Gene Selection in Part 1	176
5.5.6	Performance Estimation	178
5.6	Results of FARM41 and SVM41	180
5.7	Results of FPM	186
5.7.1	Results and Discussion of Model Selection in Part 1	187
5.7.2	Results and Discussion of Gene Selection in Part 2	191
5.7.3	Results and Discussion of Optimization in Part 3	199
5.7.4	List of Selected Genes	205
5.7.5	Results and Discussion of Tuning in Part 4	205
5.8	Conclusion	213
6	CONCLUSION AND FUTURE WORK	217
6.1	Future Work	220
	REFERENCES	221
	BIODATA OF STUDENT	237
	LIST OF PUBLICATIONS	238