AN INFORMATION-THEORETIC APPROCH TO PATTERN RECOGNITION AND ITS APPLICATION IN LIFE SCIENCE

MR.THEERA PIROONRATANA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING
DEPARTMENT OF ELECTRICAL ENGINEERING
GRADUATE COLLEGE
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK
ACADEMIC YEAR 2010

Thesis Certificate

The Graduate College, King Mongkut's University of Technology North Bangkok

Title An Information-Theoretic Approach to Pattern Recognition and Its Application in Life Science

By Mr.Theera Piroonratana

Accepted by the Graduate College, King Mongkut's University of Technology North Bangkok in Partial Fulfillment of the Requirements for the Doctor of Philosophy in Electrical Engineering

Mongkol Wangsathitwong Dean, Graduate College

(Dr.Mongkol Wangsathitwong)

18 March 2011

Thesis Examination Committee

Chairperson

(Assistant Professor Dr.Chatchawit Aporntewan)

Member

(Associate Professor Dr.Nachol Chaiyaratana)

Member

(Associate Professor Dr.Vara Varavithya)

Member

(Assistant Professor Dr.Marong Phadoongsidhi)

Member

(Assistant Professor Dr.Panrasee Ritthipravat)

AN INFORMATION-THEORETIC APPROCH TO PATTERN RECOGNITION AND ITS APPLICATION IN LIFE SCIENCE

MR.THEERA PIROONRATANA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

DEPARTMENT OF ELECTRICAL ENGINEERING

GRADUATE COLLEGE

KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK

ACADEMIC YEAR 2010

Name : Mr.Theera Piroonratana

Thesis Title : An Information-Theoretic Approach to Pattern Recognition and Its Application in Life Science

Major Field : Electrical Engineering

King Mongkut's University of Technology North Bangkok

Thesis Advisor : Associate Professor Dr.Nachol Chaiyaratana

Academic Year : 2010

Abstract

This thesis interests in two life science problems that can be tackled using information-theoretic approaches for pattern recognition. The first problem covers the identification of ancestry informative markers (AIMs) from genome-wide single nucleotide polymorphisms (SNPs). A protocol for AIM extraction is proposed. The protocol consists of three main steps: (a) identification of potential positive selection regions via F_{ST} extremity measurement, (b) SNP screening via two-stage attribute selection and (c) classification model construction using a naïve Bayes classifier. The two-stage attribute selection is composed of a newly developed round robin symmetrical uncertainty ranking technique and a wrapper embedded with a naïve Bayes classifier. The protocol has been applied to the HapMap Phase II data. Two AIM panels, which consist of 10 and 16 SNPs that lead to complete classification between CEU, CHB, JPT and YRI populations, are identified. Moreover, the panels are at least four times smaller than those reported in previous studies. The results suggest that the protocol could be useful in a scenario involving a larger number of populations. The second problem involves the application of a neural network and

decision trees in thalassaemia screening. The aim is to classify thirteen classes of thalassaemia abnormality and one control class by inspecting the distribution of multiple types of haemoglobin in blood specimens, which are identified via high performance liquid chromatography (HPLC). C4.5 and random forests are the chosen architecture for decision tree implementation. For comparison, multilayer perceptrons are explored in classification via a neural network. The stratified 10-fold cross-validation results indicate that the best classification performance is achieved when C4.5 is used in conjunction with samples which have been pre-processed with input attribute discretisation and redundant attribute removal. Subsequently, C4.5 is applied to an additional sample set in a clinical trial which results in acceptably high classification accuracy. These results suggest that a combination of C4.5 with haemoglobin typing analysis via HPLC may give rise to a guideline for further investigation of thalassaemia classification.

(Total 63 pages)

Keywords : Pattern Recognition, Thalassaemia, Ancestry Informative Marker

Nachol Chaiyaratana

Advisor

ชื่อ : นายธีระ พิรุณรัตน์
ชื่อวิทยานิพนธ์ : การรู้จำแบบเชิงทฤษฎีสารสนเทศและการประยุกต์ใน
วิทยาศาสตร์ชีวิต
สาขาวิชา : วิศวกรรมไฟฟ้า
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
ปีการศึกษา : 2553

บทคัดย่อ

วิทยานิพนธ์ฉบับนี้ให้ความสนใจปัญหาวิทยาศาสตร์ชีวิตซึ่งสามารถแก้โดยใช้การรู้จำแบบรูปเชิงทฤษฎีสารสนเทศทั้งหมดสองปัญหา ปัญหาที่หนึ่งครอบคลุมการคัดกรองเครื่องหมายพันธุกรรมจำเพาะบรรพบุรุษจากภาวะพหุสัณฐานนิวคลีโอไทด์เดี่ยว (สนิป) ระดับจีโนม เกณฑ์ที่ใช้ประกอบด้วยสามขั้นตอนคือ (ก) การระบุบริเวณซึ่งมีแนวโน้มการคัดเลือกบวกด้วยวิธีการหาค่าสุดขีด F_{ST} (ข) การคัดกรองสนิปด้วยการคัดเลือกลักษณะประจำสองระยะและ (ค) การสร้างแบบจำลองการจำแนกโดยใช้ตัวจำแนกแบบเบย์สามัญ การคัดเลือกลักษณะประจำสองระยะประกอบด้วยเทคนิควนรอบการจัดลำดับความไม่แน่นอนแบบสมมาตรที่ได้รับการพัฒนาขึ้นใหม่ และตัวห่อซึ่งฝังตัวด้วยตัวจำแนกแบบเบย์สามัญ เกณฑ์นี้ได้นำไปใช้กับข้อมูล HapMap Phase II ได้ผลลัพธ์เป็นแผงเครื่องหมายพันธุกรรมจำเพาะบรรพบุรุษสองแผงที่ประกอบด้วยสนิปจำนวน 10 และ 16 สนิป ซึ่งนำไปสู่การจำแนกระหว่างประชากร CEU, CHB, JPT และ YRI อย่างสมบูรณ์ ยิ่งกว่านั้นแผงดังกล่าวยังมีขนาดเล็กกว่าเป็นสัดส่วนอย่างน้อยสี่เท่าของขนาดแผงเครื่องหมายพันธุกรรมจำเพาะบรรพบุรุษที่มีการรายงานในการศึกษาก่อนหน้านี้ จากผลลัพธ์ที่ได้แสดงให้เห็นว่าเกณฑ์ที่นำเสนอสามารถใช้กับปัญหาซึ่งประกอบด้วยประชากรหลายประชากร

ปัญหาที่สองเกี่ยวพันกับการประยุกต์ข่ายงานระบบประสาทและต้นไม้ตัดสินใจในการคัดกรองโรคธาลัสซีเมีย จุดมุ่งหมายคือการจำแนกความผิดปกติของโรคธาลัสซีเมียสิบสามกลุ่มและกลุ่มควบคุมหนึ่งกลุ่มโดยการตรวจดูการแจกแจงของฮีโมโกลบินชนิดต่างๆ ในตัวอย่างเลือดซึ่งระบุได้ด้วยเครื่องโครมาโทกราฟีของเหลวสมรรถนะสูง C4.5 และเทคนิคป่าสุ่มคือสถาปัตยกรรมที่ถูกเลือกสำหรับการสร้างต้นไม้การตัดสินใจ ในขณะที่มัลติเลเยอร์เพอร์เซ็บตรอนคือสถาปัตยกรรมที่ถูกเลือกสำหรับการสร้างข่ายงานระบบประสาท ผลลัพธ์จากการตรวจสอบ

ความสมเหตุสมผลแบบไขว้เป็นชั้น 10 ชั้นชี้ให้เห็นว่าได้ประสิทธิภาพการจำแนกดีที่สุดเมื่อนำ C4.5 ไปใช้กับตัวอย่างที่ลักษณะประจำผ่านกระบวนการทำให้เป็นข้อมูลวิยุตและกำจัดความซ้ำซ้อนของลักษณะประจำ จากนั้น C4.5 ถูกประยุกต์กับตัวอย่างเพิ่มเติมจากการทดลองที่คลินิก ซึ่งผลลัพธ์ที่ได้มีความถูกต้องของการจำแนกในเกณฑ์ยอมรับได้อย่างสูง ผลลัพธ์ดังกล่าวยืนยันว่าการรวมกันของ C4.5 กับการวิเคราะห์ชนิดฮีโมโกลบินด้วยเครื่องโครมาโทกราฟีของเหลวสมรรถนะสูงสามารถให้แนวทางการคัดกรองโรคธาลัสซีเมีย

(วิทยานิพนธ์มีจำนวนทั้งสิ้น 63 หน้า)

คำสำคัญ: การรู้จำแบบ, ธาลัสซีเมีย, เครื่องหมายพันธุกรรมจำเพาะบรรพบุรุษ

ณชล ไชยรัตนะ

_____ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Associate Professor Dr.Nachol Chaiyaratana, for his assistance throughout the period of this research work.

Finally I would link to thank Thailand Research Fund for financial support through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/1.E.KN/50/A.1)

Theera Piroonratana

TABLE OF CONTENTS

		Page
Abstract (in English)		ii
Abstract (in Thai)		iv
Acknowledgements		vi
List of Tables		viii
List of Figures		x
Chapter 1	Prologue	1
Chapter 2	Identification of Ancestry Informative Markers	4
	2.1 Introduction	4
	2.2 Materials and Methods	7
	2.3 Results and Discussions	13
	2.4 Conclusions	24
Chapter 3	Thalassaemia Classification	26
	3.1 Introduction	26
	3.2 Materials and Methods	30
	3.3 Results and Discussions	39
	3.4 Conclusions	47
Chapter 4	Epilogue	51
Bibliography		52
Appendix A		61
Biography		63

LIST OF TABLES (CONTINUED)

Table		Page
3-9	Summary of classification errors from the clinical trial.	49

LIST OF FIGURES

Figure Page

1-1 Steps in pattern recognition. The pre-processing step may involve attribute selection and attribute discretisation. 2

1-2 An information theory influences various steps in pattern recognition. 2

1-3 Schematic diagram of the proposed AIM identification protocol. Details of each step are given in Chapter 2. 3

1-4 Schematic diagram of the procedure for solving the thalassaemia classification problem. Details of each step are given in Chapter 3. 3

2-1 Outline of the SU_2 ranking. In this example, the three-population problem consists of balanced 150 samples and 1,000 SNPs. The genotype distribution of SNP1 in all three populations is displayed. This leads to the SU_2 values of 0.016193, 0.009468 and 0.049025 for the population pairs (Pop_1, Pop_2), (Pop_1, Pop_3) and (Pop_2, Pop_3), respectively. After the calculation of SU_2 values for each SNP in every population pair is completed, SNPs are sorted according to their ranks. Three sets of top-ranked SNPs can be extracted from three population pairs. Only the top 50 SNPs are selected for each sorted set. The merging of three 50-SNP sets leads to the screened SNP set of size between 50 and 150. 11

2-2 Performance of CFS, NB-Wrap, the simple SU ranking and the SU_2 ranking in conjunction with a naïve Bayes classifier. 15

LIST OF FIGURES (CONTINUED)

Figure Page

2-3 Performance of NB-Wrap and the two-stage approach, consisting of the SU_2 ranking and NB-Wrap, in conjunction with a naïve Bayes classifier. 16

2-4 Empirical F_{ST} distribution for every population pair. 18

3-1 Schematic diagram for the methodology employed in the investigation. 29

3-2 Elution chromatograms of (a) a normal specimen and (b) a specimen from a person with Hb E trait that are obtained from an Hb Gold HPLC system. RT(s) represents the retention time in seconds for each fraction of elute. % of Hb represents the percentage of haemoglobin in the elution peak. 31

3-3 Schematic diagram of a multilayer perceptron: (a) computational model of a neuron and (b) feed-forward network structure that contains one hidden layer. 38

3-4 C4.5 decision tree which is constructed using discretised attributes. A set of screening rules can be extracted from the decision tree. For example, if the percentage of Hb E in a blood specimen is less than or equal to 6.60 while the combined percentage of Hb A1C and Hb F from the same specimen is between 24.35 and 67.75, then it is most likely that the specimen is taken from an HPFH patient. 38