CLASSIFICATION OF SNPs FOR OBESITY ANALYSIS USING FARNeM MODELLING

ONG PHAIK LING

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS

JUDUL: CLASSIFICATION OF SNPs FOR OBESITY ANALYSIS USING FARNeM MODELLING.

SESI PENGAJIAN: 2012/2013

Saya ONG PHAIK LING
(HURUF BESAR)

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
___/___	TIDAK TERHAD	

_____	_____
(TANDATANGAN PENULIS)	(TANDATANGAN PENYELIA)
Alamat tetap: A4-02A 162 Residency Km12 Jalan Ipoh Rawang 68100 Batu Caves, Selangor.	___DR. CHOO YUN HUOY_____
	Nama Penyelia
Tarikh : _____	Tarikh : _____

CATATAN: *Tesis dimaksudkan sebagai Laporan Projek Sarjana Muda(PSM)
**Jika Tesis ini SULIT atau TERHAD, sila Lampirkan surat daripada pihak berkuasa.

CLASSIFICATION OF SNPs FOR OBESITY ANALYSIS USING FARNeM MODELLING

ONG PHAIK LING

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Artificial Intelligence)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2013

DECLARATION

I hereby declare that this project report entitled

CLASSIFICATION OF SNPs FOR OBESITY ANALYSIS USING FARNeM MODELLING

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : _____ Date: _____

(ONG PHAIK LING)

SUPERVISOR : _____ Date: _____

(DR. CHOO YUN HUOY)

DEDICATION

To my beloved parents, Ong Beng San and Lim Sew Lean, your love and support are my greatest inspiration upon accomplish this project.

To my dear friends, especially Liew Siaw Hong and Ee Kim Hwe for your motivation and support throughout this project.

To my dearest lecturer, Dr Choo Yun Huoy for being responsible, helpful and always by my side to encourage and motivate me.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisor – Dr. Choo Yun Huoy and evaluator – Dr. Sharifah Sakinah for the useful comments, remarks, assistance, guidance and encouragement throughout this project. Without her, I could not have done this project successfully.

Furthermore, I would like to thank my beloved parents Ong Beng San and Lim Sew Lean who have been lovely and supportive to me.

Last but not least, an honourable mention goes to my lovely friends for their understanding and support throughout this project.

ABSTRACT

The current trend of obesity research is heading toward the field of Single Nucleotide Polymorphism (SNPs). It is because with a recognise SNPs through classification, a personalized medicine can be customized which in turn allow early diagnosis. However, it is costly and time consuming to deal with large size, redundant and noisy SNPs data. Therefore, feature selection has to precede classification task. This experiment is following a general methodology which consists of 6 phases- preliminary studies, data preparation, SNPs reduction, classification of SNPs, benchmarking and analysis, and lastly result validation. Forward attribute reduction based on neighbourhood rough set model (FARNeM) is used to select attribute that are disease related and to discard attribute that are not disease related because it can avoid information loss cause by discretization process in classical rough set. A common threshold, 0.1 and a common distance, Euclidean distance are implemented in FARNeM to perform feature selection. Then, the reduction result performance is compared among FARNeM, Correlation Feature Selection (CFS), ReliefF and with data without undergo feature selection. Both CFS and ReliefF were chosen based on their reduction properties that are subset reduction and ranking reduction respectively, which believe can produce a comparative result with FARNeM. It is at best to maximize positive predictive value and negative predictive value in diagnostic task. Thus, classification accuracy, sensitivity and specificity are used to further assess the flexibility of error rate. Experimental result shows that, it is encouraging to perform feature selection and FARNeM performs better than others technique in sensitivity and specificity measurements. However, the accuracy of FARNeM is affected badly by skewed data. Therefore, in future, improvement needs to be done when dealing with skewed data. Besides that, it is also suggested to tune the parameter of threshold as threshold is very important in determining the size of neighbourhood.

ABSTRAK

Bidang dalam Polymorphism Nukleotida Single (SNP) merupakan trend semasa pnyelidikan obesiti dan kegemukan. Ini adalah kerana melalui klasifikasi, SNPs dapat dikenalpasti dan seterusnya membatu dalam diagnosis awal.. Walau bagaimanapun, ia adalah mahal dan memakan masa untuk berurusan dengan data SNPs yang saiznya selalu besar, berlebihan dan bising. Oleh itu, pemilihan ciri adalah diperlukan sebelum klasifikasi. Eksperimen ini telah mengikut kaedah umum yang terdiri daripada 6 kajian frasa-awal iaitu, penyediaan data, pengurangan SNPs, klasifikasi SNP, penanda aras dan analisis, dan akhirnya pengesahan keputusan. Forward pengurangan sifat berdasarkan kejiranan model set kasar (FARNeM) digunakan untuk memilih sifat yang berkaitan dengan penyakit dan membuang sifat yang tidak berkaiatan kerana ia boleh mengelakkan kehilangan punca maklumat melalui proses pendiskretan dalam set kasar klasik. Ambang dan jarak yang biasa digunakan dengan 0.1 dan Euclidean jarak masing-masing digunakan dalam FARNeM untuk melaksanakan pemilihan ciri. Kemudian, prestasi hasil pengurangan telah dibandingkan di kalangan FARNeM, Pemilihan Ciri-ciri Korelasi (CFS), ReliefF dan dengan data tanpa menjalani pemilihan ciri. Kedua-dua CFS dan ReliefF dipilih berdasarkan ciri-ciri pengurangan mereka yang pengurangan subset dan pengurangan kedudukan masing-masing. Ia adalah yang terbaik untuk memaksimumkan nilai ramalan positif dan nilai negatif ramalan dalam tugas diagnostik. Oleh itu, selain ketepatan klasifikasi, sensitiviti dan spesifikasi juga digunakan untuk menilai fleksibiliti kadar kesilapan. Hasil eksperimen menunjukkan pemilihan cirri adalah penting dan FARNeM lebih baik daripada teknik yang lain dalam sensitiviti dan ukuran kekhususan. Walau bagaimanapun, ketepatan FARNeM terjejas oleh data pencong. Oleh itu, pada masa hadapan, penambahbaikan perlu dilakukan apabila berhadapan dengan data pencong. Selain itu, ia juga dicadangkan untuk mencuba pelbagai parameter ambang kerana ambang adalah sangat penting dalam menentukan saiz kejiranan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATONS

A	-	Adenine
Actr	-	Ancestral
ACE	-	Angiotensin Converting Enzyme
AD	-	Anderson Darling
AMD	-	Age-Related Macular Degeneration
AUC	-	Area under the curve
BDNF	-	Brain derived neurotrophic factor
BMI	-	Body Mass Index
BN	-	Bayesian Network
BVI	-	Body Volume Index
C	-	Cytosine
CNVs	-	Number variants
CSSP	-	Column Subset Selection Problem
D	-	Derived
DLDA	-	Diagonal Discriminant Analysis
DNA	-	Deoxyribonucleic acid
FARNeM	-	Forward Attribute Reduction based on neighbourhood rough set model
FN	-	False Negative
FP	-	False Positive
G	-	Guanine
GA	-	Genetic Algorithm
GWAS	-	Genome-wide association studies
htSNPs	-	Tagging SNPs

IBD5	-	Inflammatory Bowel Disease 5
ICA	-	Independent Component Analysis
KNN	-	K-Nearest Neighbour
LD	-	Linkage equilibrium
LEP	-	Leptin
LEPR	-	Leption receptor
LPL	-	Human Lipoprotein Lipase
MC4R	-	Melanocortin 4 receptor
NHMS	-	National Health and Morbidity Survey
NTRK2	-	Tyrosine kinase receptor type 2 gene
PCA	-	Principal Component Analysis
PCSK1	-	Prohormone convertase 1
POMC	-	Proopiomelanocortin
RS	-	Rough Set Theory
RST	-	Rough Set Theory
SIM1	-	Single-minded homolog 1
SNPs	-	Single Nucleotide Polymorphisms
T	-	Thymine
TN	-	True Negative
TP	-	True Positive
UTeM	-	Universiti Teknikal Malaysia Melaka
WEKA	-	Waikato Environment for Knowledge Analysis
WHO	-	World Health Organization
WHR	-	Waist to hip ratio

CHAPTER I

INTRODUCTION

1.1 Project Background

"Overweight and obesity are defined by the World Health Organization (WHO) as abnormal or excessive fat accumulation that presents a risk to an individual health." In Malaysia, the level of obesity has reach alarming proportions and is one of the serious public health problems that we are facing nowadays. The obesity-associated diseases such as diabetes type 2 are threatening and are now believe to be live shortening. According to scientist, the recent rapid rise in obesity might due to major changes in eating behaviour and lifestyle but who becomes obese at individual level is determined to a great extend by genetic susceptibility (Fontaine *et al*., 2003). According to Helene and David in their journal: Molecular Basis of Obesity, it was stated that they have evidence that obesity is strongly heritable disorder. Therefore, genetic plays an important role in affecting obesity.

The analysis of single nucleotide polymorphism (SNPs) data is the key component of disease-gene association studies. With the technological developments, high-throughput genotyping and sequencing techniques challenge researchers to

analyse genome-wide sequence datasets with hundreds of thousands of SNPs. Due to the large size of these datasets, educated reduction of the number of SNPs is required in order to meet the computational demands of association studies.

Genetic research using computational method in combating obesity issue in Malaysia has developed considerably. However, researchers nowadays are still focusing on biomedical approaches which are expensive and the result is not very satisfying. This is due to the rapid growth of genomic data volume has outranges human's ability to manage and deal with them. Gene selection is a typical example of application domain with high dimensional data. In gene selection problems, expression levels of many genes are mostly recorded by microarray data, but only a small number gene are critical for classification and diagnosis. In addition, for large size of features for example genes, usually only a smaller size of examples are available for training and testing purposes. This makes the process of learning even more difficult.

The main goal in gene selection is to improve the accuracy of classification besides reducing computational cost and noises. Therefore, the reduction of gene is an important issue. One of the important concepts in rough set reduction is dependency of attributes which can be used to find the degree of similarity between attribute and decision (Mei-Ling, 2010) which is useful in gene selection. In order to evaluate a better gene subset, there are two basic rules, which are relevance of the gene and interaction of gene. The characteristic of indicernability and optimization of rough set reduction are believed to be useful in dealing with bioinformatics data especially in gene selection as it has the ability to search objects in a multi-dimensional data space (Nordin *et al*., 2010).

The concept of rough set reduction is an effective method to select informative gene. However, implementing only the concept of rough set reduction in gene selection is not enough. Research shows that discretisation will cause the information loss as it will discard the outmost of redundancy and noise. This is because biological data normally contain missing values and by discretisation, it might discretise some relevant gene which will affect accuracy (Mei-Ling, 2010).

Therefore, the neighbourhood rough set model by Hu Qing-Hua is introduced (Qing-Hua *et al.*, 2008).

The concept of neighbourhood is that the neighbours lie homogeneously and is as close to sample data as possible. This neighbourhood of the data will be considered not only in terms of proximity but also in terms of their spatial distribution with respect to the dataset. Neighbourhood rough set has proved to be a powerful tool in attribute reduction, rule extraction, reasoning with uncertainty and feature selection. Therefore, neighbourhood and it's relation are a class of important concept in topology which can also be applied to biological data. The dependence function of neighbourhood rough set was used as the heuristic information. By implementing the concept of neighbourhood rough set reduction, it will help to select a minimal reduct, which decreases the likelihood of information loss as this technique combines the advantages of feature subset selection and neighbourhood-based classification. Besides that, the concept of neighbourhood rough set reduction is believed to be simple and straightforward to implement.

Thus, this research aims to reduce the number of SNPs through neighbourhood rough set reduction modelling in targeted obesity genes using soft computing techniques.

1.2 Problem Statements

The discovery of SNPs in the domain of obesity is on the trend now. With the high volume of genetic data, it is computational costly and time consuming to use the experimental through laboratory to determine various obesity related gene. According to the review of technique especially in the field of SNPs, FARNeM has achieved a very good performance in many domains such as cancer. However, up to my knowledge, it is still rare or no research showing FARNeM can perform in the domain of obesity.

1.3 Objectives

This project embarks on the following objectives:

1. To identify the suitable gene variants group for obesity diagnosis.
2. To design and implement FARNeM in SNPs obesity.
3. To test and analyse FARNeM.

1.4 Scopes

This project focuses mainly on tackling the issue of obesity and overweight using soft computing techniques such as feature selection in the area of SNPs. In this project, both threshold and distanced used is fixed at a commonly used parameter which is 0.1 and Euclidean distance respectively. The performance of FARNeM is then tested with a SNPs related dataset with 5 seed to avoid any bias learning. The performance measurement such as accuracy, specificity, sensitivity, positive predictive value and negative predictive value were used to measure the performance of the proposed technique. In order to validate the result, normality test was carried out to test the distribution of result before deciding on the type of validation test.

1.5 Project Significance

The project significance is to introduce the technique of Forward Attribute Reduction based on neighbourhood rough set model (FARNeM) in the domain of obesity and overweight particular in SNPs.

1.6 Expected Output

At the end of this project, I am expected to get a s*et al*gorithm and coding base on neighbourhood rough set to get better SNP selection and classification.

1.7 Conclusion

In conclusion, an AI technique based on the concept of neighbourhood rough set is proposed to get better SNP selection and classification. It is believed that a promising result in term of accuracy, specificity, sensitivity, positive predictive value and negative predictive value would be produced.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

In this chapter, a literature review of the biological related foundation and dimensionality reduction studies for SNPs selection are provided. The chapter contains four sections. The first section presents the importance of performing research in the field of obesity and overweight. The second section provides the essential concept of biology related studies to this research. The third section surveys on genetic data analysis literature for dimensionality reduction techniques in different categories. Lastly, the fourth section discusses on the evaluation of performance measurement for the experimental results analysis.

2.2 Importance of Performing Obesity and Overweight Research in Malaysia

Overweight and obesity has already reached an alarming health care level. According to the World Health Organization (WHO), the rate of obesity and overweight has been doubled since 1980 with 502 million adults classified as obese globally (Fontaine *et al*., 2003). Besides that, surprisingly, the rate of death caused by obesity and overweight far exceed the rate of death caused by underweight (The Star, 2012).

Back to our home front, the statistic of obesity and overweight cases in Malaysia has been increasing from time to time. Based on National Health and Morbidity Survey (NHMS) in 2006, 43% of Malaysian adults and 38% of Malaysian children were overweight or obese. In the year of 2010, WHO has ranked Malaysia as 6th in Asia-Pacific region with the highest adult obesity rate. Meanwhile, Malaysia is currently at top of obesity ranking and is rated as "the fattest country" in Southeast Asia with the statistic of 60% of Malaysians aged 18 and above is labelled as overweight and obesity.

These entire statistics have rung alarm bells at the Malaysia Ministry of Health. As according to statistics there are nearly 2 out of 3 adults in Malaysia are either obese or overweight. Figure 2.1 shows the rate of obesity in Malaysia according to state in year 2011. There is an increasing concern over this health issue from the Malaysia government as the consequences of obesity are harmful. According to research, there are many obesity-associated diseases that contribute to the percentage of death in Malaysia such as type 2 diabetes, cardiovascular disease, hypertension and certain cancers (WHO, 2001). These obesity-associated diseases have become the silent killer for the majority of Malaysian.

Besides that, obesity or overweight and obesity-associated diseases have a significant economic impact in terms of absence from work and depleting health resources in the country. Weight gain has also been linked with poor concentration levels, poor academic success and social exclusion in school. The quality of life of Malaysian was proven to be affected as overweight or obese people tend to have