

บทคัดย่อ

รหัสโครงการ: RMU5380004

ชื่อโครงการ: การตรวจหาอันตรายกิริยาแบบอภิสเตชิสปริสุทธิ์ระหว่างหลาย
ตำแหน่งพันธุกรรมโดยการทดสอบการเรียงสับเปลี่ยนเชิงสุ่มที่รวมผล
การวิเคราะห์อันตรายกิริยาระหว่างตำแหน่งพันธุกรรมสองตำแหน่ง

ชื่อนักวิจัย: รองศาสตราจารย์ ดร.ณชล ไชยรัตน์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

E-mail Address: nchl@kmutnb.ac.th

ระยะเวลาโครงการ: 2 ปี

ระยะเวลาสัญญา: 1 ปี

รายงานฉบับนี้สนใจโรคพันธุกรรมสองโรคที่สามารถประยุกต์ใช้การรู้จำแบบในการวิเคราะห์โรคแรกที่สนใจคือโรคเบาหวานชนิดที่สอง 2LOmb ซึ่งทำการคัดเลือกลักษณะประจำโดยการทดสอบการเรียงสับเปลี่ยนเชิงสุ่มที่รวมผลของการวิเคราะห์อันตรายกิริยาระหว่างตำแหน่งพันธุกรรมสองตำแหน่งเป็นขั้นตอนวิธีที่พัฒนาขึ้นสำหรับปัญหาแรก ขั้นตอนวิธีประกอบด้วยสี่ขั้นได้แก่การวิเคราะห์แบบสองตำแหน่งพันธุกรรม การทดสอบการเรียงสับเปลี่ยนเชิงสุ่ม การคำนวณค่าความน่าจะเป็นวงกว้าง และการค้นหาแบบก้าวหน้าเพื่อระบุการรวมการวิเคราะห์แบบสองตำแหน่งที่ดีที่สุด สมรรถนะของ 2LOmb ได้รับการเปรียบเทียบกับสมรรถนะของวิธีการหาความสัมพันธ์ทางพันธุกรรมแบบเซต การคัดเลือกลักษณะประจำแบบอิงสหสัมพันธ์ และเทคนิค ReliefF แบบปรับปรุง ผลการจำลองจากปัญหาอันตรายกิริยาหลายตำแหน่งแสดงให้เห็นว่า 2LOmb ให้ค่าผิดพลาดแบบบวกเท็จต่ำ นอกจากนั้น 2LOmb สามารถระบุภาวะพหุสัญญาณนิวคลีโอไทด์เดี่ยวที่เป็นสาเหตุโรคได้ทั้งหมด ซึ่งชี้ให้เห็นว่า 2LOmb มีกำลังการตรวจจับสูง 2LOmb ได้รับการนำไปประยุกต์ใช้กับข้อมูลโรคเบาหวานชนิดที่สองที่ประกอบด้วย 7,065 ภาวะพหุสัญญาณนิวคลีโอไทด์เดี่ยวจาก 370 ยีน 2LOmb ระบุ 11 ภาวะพหุสัญญาณนิวคลีโอไทด์เดี่ยวจากสี่ยีนที่สัมพันธ์กับโรคเบาหวานชนิดที่สอง โรคที่สองที่สนใจคือโรคธาลัสซีเมีย โดยปัญหาที่สนใจคือการจำแนกคุณสมบัติของเลือดโดยใช้ต้นไม้ตัดสินใจ C4.5 ตัวจำแนกแบบอย่างง่าย และมัลติเลเยอร์เพอร์เซ็ปตรอนสำหรับการคัดกรองโรคธาลัสซีเมีย เป้าหมายคือการจำแนกประเภทความผิดปกติธาลัสซีเมีย 18 กลุ่มซึ่งมีความชุกสูงในประเทศไทยและคนปกติหนึ่งกลุ่ม

โดยอาศัยข้อมูลจากการตรวจความสมบูรณ์ของเม็ดเลือดและการตรวจชนิดฮีโมโกลบิน ผลการทดลองแสดงให้เห็นว่าได้สมรรถนะการจำแนกสูงสุดซึ่งมีค่าความแม่นยำ 93.23% (ค่าเบี่ยงเบนมาตรฐาน 1.67%) และค่าความแม่นยำ 92.60% (ค่าเบี่ยงเบนมาตรฐาน 1.75%) จากตัวจำแนกเบี่ยงอย่างง่ายและมัลติเลเยอร์เพอร์เซ็ปตรอนตามลำดับเมื่อทำให้ลักษณะประจำมีค่าแบบกนทันะก่อน และผลจากการคัดเลือกลักษณะประจำแบบห่อแสดงให้เห็นว่าค่าความเข้มข้นฮีโมโกลบินซึ่งเป็นลักษณะประจำไม่มีผลต่อสมรรถนะการจำแนก นอกจากนี้สมรรถนะการจำแนกนี้ยังสูงกว่าสมรรถนะการจำแนกที่ได้เมื่อใช้ข้อมูลจากการตรวจชนิดฮีโมโกลบินเป็นข้อมูลเข้าสำหรับตัวจำแนกเพียงอย่างเดียว

คำหลัก: การคัดเลือกลักษณะประจำ การจำแนก การรู้จำแบบ ชาลส์ซีเมีย เบาหวานชนิดที่สอง

Abstract

Project Code: RMU5380004

Project Title: Detecting Purely Epistatic Multi-locus Interactions by an Omnibus Permutation Test on Ensembles of Two-Locus Analyses

Investigator: Associate Professor Dr. Nachol Chaiyaratana
King Mongkut's University of Technology North Bangkok

E-mail Address: nchl@kmutnb.ac.th

Project Period: 2 Years

Contract Period: 1 Year

This report interests in two genetic disease problems that can be tackled using pattern recognition. The first problem covers a genetic association study of type 2 diabetes mellitus (T2D). 2LOmb which performs attribute selection via an omnibus permutation test on ensembles of two-locus analyses is proposed. The algorithm consists of four main steps: two-locus analysis, a permutation test, global p -value determination and a progressive search for the best ensemble. 2LOmb is benchmarked against a set association approach, a correlation-based feature selection technique and a tuned ReliefF technique. The simulation results from multi-locus interaction problems indicate that 2LOmb has a low false-positive error. Moreover, 2LOmb has the best performance in terms of an ability to identify all causative single nucleotide polymorphisms (SNPs), which signifies a high detection power. 2LOmb is subsequently applied to a T2D data set, which contains 7,065 SNPs from 370 genes. The 2LOmb search reveals that 11 SNPs in four genes are associated with T2D. The second problem involves the classification of blood characteristics by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening.

The aim is to classify 18 classes of thalassaemia abnormality, which have a high prevalence in Thailand, and one control class by inspecting data characterised by a complete blood count (CBC) and haemoglobin typing. The stratified 10-fold cross-validation results indicate that the best classification performance with average accuracy of 93.23% (standard deviation = 1.67%) and 92.60% (standard deviation = 1.75%) is achieved when the naïve Bayes classifier and the multilayer perceptron are respectively applied to samples which have been pre-processed by attribute discretisation. The results from wrapper attribute selection also suggest that the haemoglobin concentration attribute is redundant. Moreover, the achieved classification performance is significantly higher than that obtained using only haemoglobin typing attributes as classifier inputs.

Keywords: Attribute Selection, Classification, Pattern Recognition, Thalassaemia, Type 2 Diabetes Mellitus