

ผลของค่าสถิติเฉพาะที่อิงการจำแนกต่อการวิเคราะห์บรรณนิทัศน์

นายเจษฎา วีระเดชกำพล

นายธิดิ รุ่งเรือง

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวិชากรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ปีการศึกษา 2561

Effects of Classification-Based Local Statistic on Annotation Analysis

Mr. Jessada Weeradetkumpon

Mr. Titi Rungruang

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF BACHELOR OF COMPUTER ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
FACULTY OF ENGINEERING
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK
ACADEMIC YEAR 2018

ปรินุญนัพนธ์รื่อง : ผลของค่าสถิติเฉพาะที่อิงการจำแนกต่อการวิเคราะห์บรรณนิทัศน์
 ชื่อ : นายเจษฎา วีระเดชกำพล
 นายชิตี รุ่งเรือง
 สาขาวิชา : วิศวกรรมคอมพิวเตอร์
 ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์
 คณะ : วิศวกรรมศาสตร์
 อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ
 ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี
 ผู้ช่วยศาสตราจารย์ ดร.ดำรงฤทธิ ศรีบุญศิริโชค
 ปีการศึกษา : 2561

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ อนุมัติให้
 ปรินุญนัพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต
 สาขาวิชาวิศวกรรมคอมพิวเตอร์

..... หัวหน้าภาควิชาวิศวกรรมไฟฟ้า
 (ผู้ช่วยศาสตราจารย์ ดร.นภค วิวัชร โกเศศ) และคอมพิวเตอร์

..... ประธานกรรมการ
 (รองศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ)

..... กรรมการ
 (รองศาสตราจารย์ ดร.วรา วราวิทย์)


..... กรรมการ
 (ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี)



..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ดำรงกฤต ศรีสุศรีโชค)


ลิขสิทธิ์ของภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

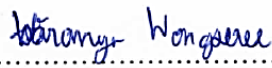
Project Report Title : Effects of Classification-Based Local Statistic on Annotation Analysis
Name : Mr. Jessada Weeradetkumpon
Mr. Titi Rungruang
Major Field : Computer Engineering
Department : Electrical and Computer Engineering
Faculty : Engineering
Project Advisors : Assoc. Prof. Dr. Nachol Chaiyaratana
Asst. Prof. Dr. Waranyu Wongseree
Asst. Prof. Dr. Damrongrit Setsirichok
Academic Year : 2018

Accepted by the Faculty of Engineering, King Mongkut's University of Technology North
Bangkok in Partial Fulfillment of the Requirements for the Degree of Bachelor of Computer
Engineering


.....
(Asst. Prof. Dr. Nophadon Wiwatcharagoses) Chairperson of Department of Electrical
and Computer Engineering


.....
(Assoc. Prof. Dr. Nachol Chaiyaratana) Chairperson


.....
(Assoc. Prof. Dr. Vara Varavithya) Member


.....
(Asst. Prof. Dr. Waranyu Wongseree) Member


.....

Member

(Asst. Prof. Dr. Damrongrit Setsirichok)

Copyright of the Department of Electrical and Computer Engineering, Faculty of Engineering

King Mongkut's University of Technology North Bangkok

บทคัดย่อ

ปริญญานิพนธ์นี้นำเสนอการบูรณาการระหว่างการวิเคราะห์บาทวิถีอิงภววิทยา และการจำแนกของสองคลาสข้อมูลไมโครอาเรย์การแสดงออกของยีน การวิเคราะห์นี้สำคัญของฟังก์ชันและการแสดงออกหรือ SAFE เป็นเครื่องมือที่ถูกใช้สำหรับการวิเคราะห์บาทวิถี ในขณะที่การทำให้เป็นปรกติของการถอดยโลจิสติกถูกใช้สำหรับการจำแนก สามบทลงโทษการทำให้เป็นปรกติที่ศึกษาได้แก่ การถอดยริดจ์ การหาค่าสัมบูรณ์น้อยสุดและตัวดำเนินการเลือกหรือแลชโซ และข่ายยืดหยุ่น ค่าสัมบูรณ์ของสัมประสิทธิ์ของยีนที่เป็นข้อมูลเข้าจากการทำให้เป็นปรกติของแบบจำลองการถอดยโลจิสติกถูกใช้ในส่วนของการปรับระดับยีนในกระบวนการ SAFE การปรับปรุง SAFE ที่ใช้ประโยชน์จากการถอดยริดจ์ แลชโซ และข่ายยืดหยุ่นในปริญญานิพนธ์นี้จะถูกเรียกว่า rSAFE lSAFE และ eSAFE ข้อมูลการแสดงออกของยีนจากห้าการศึกษาโรคมะเร็งที่สนใจได้แก่ ลูคีเมีย มะเร็งปอด, ตัวอ่อนเนื้องอกที่ระบบประสาทส่วนกลาง, มะเร็งต่อมลูกหมากและมะเร็งสมอง ผลลัพธ์จากการศึกษาทั้งหมดนี้ชี้ให้เห็นว่าบทลงโทษของการถอดยริดจ์เป็นตัวเลือกที่ดีที่สุดของการทำให้เป็นปรกติ ซึ่งจำนวนวิธีหรือหมวดหมู่ของยีนที่สำคัญจากการวิเคราะห์โดยใช้ rSAFE ให้ผลลัพธ์คล้ายกับการวิเคราะห์ด้วยวิธี SAFE อย่างไรก็ตามยังมีความแตกต่างระหว่างหมวดหมู่ของยีนที่สำคัญที่ถูกระบุโดย SAFE และ rSAFE จึงแนะนำว่าควรใช้เทคนิคทั้งสองควบคู่กันเมื่อวิเคราะห์ข้อมูลการแสดงออกของยีน

Abstract

This article presents an integration between gene ontology-based pathway analysis and classification of two-class gene expression microarray data. A significant analysis of function and expression or SAFE framework was chosen for the pathway analysis while regularized logistic regression was chosen for the classification. Three regularization penalties were explored: ridge regression, least absolute shrinkage and selection operator or lasso and elastic net penalties. Absolute values of coefficients of gene inputs from a regularized logistic regression model were then used as gene-level statistics in SAFE. The modified SAFE that exploited ridge regression, lasso and elastic net regularization was referred to as rSAFE, lSAFE and eSAFE, respectively. Gene expression data from five studies of cancer were of interest: leukemia, lung carcinoma, central nervous system embryonal tumor, prostate cancer and brain cancer studies. The overall results indicated that the ridge regression penalty was the best choice for regularization. Consequently, the number of significant pathways from the analysis using rSAFE was similar to that from the analysis using SAFE. However, there were differences between significant pathways identified by SAFE and rSAFE, suggesting that both techniques should be used when analyzing gene expression data.

กิตติกรรมประกาศ

ปริญญานิพนธ์เล่มนี้ไม่อาจเสร็จสมบูรณ์ได้หากปราศจากความช่วยเหลือจาก
รองศาสตราจารย์ ดร.ณชล ไชยรัตน์ ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี และผู้ช่วย
ศาสตราจารย์ ดร.ดำรงฤทธิ์ เศรษฐศิริโชค ที่คอยให้คำแนะนำและให้การสนับสนุน ตลอดทั้งการ
ให้ความช่วยเหลือในทุก ๆ ด้าน จนทำให้ปริญญานิพนธ์เล่มนี้เสร็จสมบูรณ์ออกมาครบถ้วน
ต้องขอขอบพระคุณอาจารย์ทุกท่านมา ณ โอกาสนี้

ข้าพเจ้าขอขอบคุณอาจารย์ท่านอื่น ๆ ในภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ ทุก ๆ ท่านที่คอยให้ความรู้ คำแนะนำ และคอยสั่งสอนข้าพเจ้าตลอด
ระยะเวลาที่ศึกษาอยู่ ณ ที่แห่งนี้ จนข้าพเจ้าสามารถนำความรู้ที่ได้มานำไปใช้ในการประกอบอาชีพ
ในอนาคต

สุดท้ายนี้ต้องขอขอบคุณเพื่อน ๆ ทุกคน รุ่นพี่ รุ่นน้อง และบุคลากรของสาขาวิชาวิศวกรรม
คอมพิวเตอร์ทุกท่าน ที่คอยให้ความช่วยเหลือในตลอดระยะเวลาที่ผ่านมา

เจษฎา วีระเดชกำพล

ธิดิ รุ่งเรือง

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	๗
บทคัดย่อภาษาอังกฤษ	๘
กิตติกรรมประกาศ	๙
สารบัญตาราง	๙
สารบัญภาพ	๙
บทที่ 1. บทนำ	1
บทที่ 2. ทฤษฎี	5
2.1 การถดถอยโลจิสติก	5
2.2 การถดถอยลงโทษ	7
2.3 การวิเคราะห์นัยสำคัญของฟังก์ชันและการแสดงออก	10
2.4 ขั้นตอนการสร้างแบบจำลอง	13
บทที่ 3. ผลการทดลอง	16
3.1 ผลลัพธ์จากการคัดเลือกลักษณะประจำ	16
3.2 ผลลัพธ์จำนวนหมวดหมู่ของยีนที่สำคัญจากการทำซ้ำ	18
3.3 ผลลัพธ์จำนวนหมวดหมู่ของยีนที่สำคัญของแต่ละชุดข้อมูล	23
3.4 เวลาที่ใช้ในแต่ละการทดลอง	32
บทที่ 4. สรุปผลการวิจัย	35
เอกสารอ้างอิง	36
ประวัติผู้แต่ง	40

สารบัญตาราง

ตารางที่	หน้า
2-1 เปรียบเทียบข้อดี - ข้อเสีย และความเหมาะสมในการใช้งานของแต่ละวิธี	9
2-2 แหล่งที่มาของชุดข้อมูลและวิธีการประมวลผลก่อน	14
2-3 การประมวลผลก่อน	15
2-4 จำนวนยีนที่เหลือหลังทำการประมวลผลก่อน	15
3-1 จำนวนยีนที่สำคัญที่ผ่านการคัดเลือกในแต่ละชุดข้อมูล โดยใช้ Elastic Net ที่ $\alpha = 0.1$	16
3-2 จำนวนของยีนที่สำคัญที่ค้นพบจากเทคนิค eSAFE โดยค่าที่อยู่ในวงเล็บคือ α ที่ใช้ในการคัดเลือกลักษณะประจำ	17
3-3 จำนวนหมวดหมู่ของยีนที่สำคัญที่พบในแต่ละเทคนิค โดยใช้ Elastic Net ที่ $\alpha = 0.1$	18
3-4 จำนวนหมวดหมู่ของยีนที่สำคัญที่ค้นพบจากเทคนิค eSAFE โดยค่าที่อยู่ในวงเล็บคือ α ที่ใช้ในการคัดเลือกลักษณะประจำ	19
3-5 ผลลัพธ์ที่ได้จากการทำ FDR ของ Central Nervous System Embryonal Tumor จากการทดลองที่เป็นตัวแทน	23
3-6 ผลลัพธ์ที่ได้จากการทำ FDR ของ Brain Cancer จากการทดลองที่เป็นตัวแทน	24
3-7 ผลลัพธ์ที่ได้จากการทำ FDR ของ Lung Carcinoma จากการทดลองที่เป็นตัวแทน	26
3-8 ผลลัพธ์ที่ได้จากการทำ FDR ของ Leukemia จากการทดลองที่เป็นตัวแทน	27
3-9 ผลลัพธ์ที่ได้จากการทำ FDR ของ Prostate Cancer จากการทดลองที่เป็นตัวแทน	29
3-10 ผลลัพธ์ที่ได้จากการทำ FDR ของ Prostate Cancer (Capsular penetration) จากการทดลองที่เป็นตัวแทน	30
3-11 ผลลัพธ์ที่ได้จากการทำผลรวมสถิติของวิลคอกซันระหว่าง FDR ที่ได้จาก SAFE กับ rSAFE	31
3-12 เวลาเฉลี่ยในการทำแต่ละเทคนิค แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือ ค่าเบี่ยงเบนมาตรฐาน	32
3-13 เวลาสูงสุดในการทำแต่ละเทคนิค แต่ละรอบในหน่วยวินาที	33
3-14 เวลาลดน้อยสุดในการทำแต่ละเทคนิค แต่ละรอบในหน่วยวินาที	33

สารบัญภาพ

ภาพที่	หน้า
2-1 ตัวอย่างการจำแนกสปีชีส์ของดอกไฮริส	6
2-2 แสดงกระบวนการทำงานของ 5 fold cross validation	7
2-3 การแสดงกระบวนการทำงานของ SAFE	12
2-4 การแสดงกระบวนการทำงานของ rSAFE	12
3-1 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Central Nervours System	20
3-2 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Brain Cancer	21
3-3 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Lung Carcinoma	21
3-4 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Leukemia	22
3-5 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer	22
3-6 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer (Capsular penetration)	23
3-7 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Central Nervous System Embryonal Tumor	24
3-8 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Brain Cancer	25
3-9 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Brain Cancer	25
3-10 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Lung Carcinoma	26

สารบัญภาพ (ต่อ)

ภาพที่	หน้า
3-11 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Lung Carcinoma	27
3-12 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Leukemia	28
3-13 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Leukemia	28
3-14 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer	29
3-15 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer	30
3-16 แผนภาพแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer (Capsular penetration)	31

บทที่ 1

บทนำ

การที่ยีน (Gene) จะสามารถกำหนดลักษณะทางพันธุกรรมของสิ่งมีชีวิตได้ ข้อมูลทางพันธุกรรมในยีนจะต้องถูกถ่ายทอดจาก DNA (Deoxyribonucleic Acid) สู่ RNA (Ribonucleic Acid) และ RNA สู่โปรตีน โดยที่ RNA และโปรตีนทำงานร่วมกันในการกำหนดลักษณะเฉพาะของสิ่งมีชีวิตทุก ๆ ชนิด ดังนั้นการถ่ายทอดข้อมูลทางพันธุกรรมจาก DNA สู่ RNA และโปรตีน จึงมีความสำคัญอย่างมากในการทำให้ยีนกำหนดลักษณะทางพันธุกรรมของสิ่งมีชีวิตได้ ขบวนการสังเคราะห์ RNA มี RNA ที่จำเป็นต่อการสังเคราะห์โปรตีนอยู่ 3 ชนิด ได้แก่ tRNA (Transfer RNA), rRNA (Ribosomal RNA) และ mRNA (Messenger RNA) ขบวนการสังเคราะห์ RNA และขบวนการสังเคราะห์โปรตีน เรียกรวมกันว่าการแสดงออกของยีน (Gene Expression)

การทำวิจัยทางการแพทย์ได้ถูกปฏิรูปด้วยการพัฒนาเทคโนโลยีไมโครอาร์เรย์ (Microarray) ที่สามารถวัดระดับการแสดงออกของยีนได้ครั้งละหลายพันยีนในการทดลองเพียงครั้งเดียว ซึ่งนำไปสู่การวินิจฉัยทางคลินิกที่ประสบความสำเร็จโดยเฉพาะอย่างยิ่งในโรคมะเร็ง การศึกษาไมโครอาร์เรย์มีปัญหาหลักอยู่สองปัญหาได้แก่ ปัญหาที่หนึ่งคือ ปัญหาการค้นหาคلاسซึ่งเป็นการระบุชนิดย่อยของเนื้องอกด้วยขั้นตอนวิธีจัดกลุ่มอิงความคล้ายของข้อมูลแสดงลักษณะเฉพาะ (Profile) การแสดงออกของยีน และปัญหาที่สองคือ ปัญหาการทำนายคลาสซึ่งใช้ข้อมูลแสดงลักษณะเฉพาะการแสดงออกของยีนทำนายชนิดของเนื้องอกที่กำหนดไว้ล่วงหน้าโดยขั้นตอนวิธีการจำแนก โดยเป้าหมายสูงสุดของปัญหาเหล่านี้คือการเพิ่มความเข้าใจในกลไกการเกิดมะเร็ง แต่การวิเคราะห์ข้อมูลไมโครอาร์เรย์นั้นมีข้อบกพร่องเรื่องจำนวนตัวแปรมีมากกว่าจำนวนตัวอย่างมาก ดังนั้นจึงจำเป็นต้องพัฒนาเครื่องมือที่เหมาะสมในการวิเคราะห์ข้อมูลที่ซับซ้อนเหล่านี้

การระบุยีนที่มีความแตกต่างของระดับการแสดงออกของยีนระหว่างสองกลุ่มตัวอย่างสามารถทำได้โดยการวิเคราะห์ทางสถิติแบบตัวแปรเดียว เช่น การทดสอบที (t-test) ซึ่งเป็นเทคนิคอิงพารามิเตอร์หรือ แมนน์-วิทนี (Mann-Whitney) ที่เป็นเทคนิคไม่อิงพารามิเตอร์ อย่างไรก็ตามการวิเคราะห์ความแตกต่างนี้ไม่ได้คำนึงถึงความสัมพันธ์ระหว่างยีน ดังนั้นการวิเคราะห์ทางสถิติแบบหลายตัวแปร เช่น การวิเคราะห์ความแปรปรวน การวิเคราะห์การถดถอย

และการวิเคราะห์การจำแนกจึงมีความเหมาะสมมากกว่า แต่การวิเคราะห์ข้อมูลที่มีจำนวนตัวแปรมากกว่าจำนวนตัวอย่างมากมีแนวโน้มที่จะมีความแปรปรวนสูง ดังนั้นการทำให้เป็นปรกติเป็นเทคนิคหนึ่งในการควบคุมการแปรปรวนของแบบจำลองที่ได้จากการวิเคราะห์ทางสถิติหลายตัวแปร จากการศึกษาเปรียบเทียบเทคนิคการวิเคราะห์ข้อมูลไมโครอาร์เรย์พบว่า การวิเคราะห์เชิงเส้นร่วมกับการทำให้เป็นปรกติมีประสิทธิภาพสูงไม่แตกต่างจากเทคนิคการวิเคราะห์ไม่เป็นเชิงเส้น

นอกจากการระบุยีนที่จำเป็นต่อการจำแนกตั้งแต่สองตัวอย่างขึ้นไปในข้อมูลการแสดงออกของยีน การวิเคราะห์บาทวิถี (Pathway Analysis) เป็นอีกวิธีการหนึ่งที่มีเป้าหมายในการได้ข้อมูลเชิงลึกที่แตกต่างกัน โดยที่การวิเคราะห์บาทวิถีไม่ได้พยายามที่จะระบุข้อมูลของยีน แต่ทำการจัดกลุ่มยีนให้เป็นชุดของยีนและระบุวิถีที่มีข้อมูลของยีนเหล่านั้น ถึงแม้ว่าการวิเคราะห์บาทวิถีจะครอบคลุมการวิเคราะห์ข้อมูลทางชีววิทยาอย่างกว้างขวางรวมไปถึงการวิเคราะห์ภววิทยายีน ความสัมพันธ์ระหว่างโปรตีนและความสมดุลของฟลักซ์ (Flux) แต่การวิเคราะห์บาทวิถีมุ่งเน้นความสนใจไปที่ชุดของยีนในบริบทของการตีความข้อมูลการแสดงออกของยีน

Khatri et al. (2012) ได้แบ่งการวิเคราะห์บาทวิถีออกเป็น 3 แบบ แบบแรกคือ การวิเคราะห์เหนือตัวแทน (The over – representation analysis) ซึ่งจะทำการแบ่งยีนออกเป็นสองชุด ชุดแรกคือชุดของยีนที่สนใจและชุดสองคือชุดของยีนที่ไม่สนใจ แต่ละชุดของยีนจะแบ่งต่อไปอีกสองชุด ชุดแรกคือชุดของยีนที่มีการแสดงออกแตกต่างกันและชุดสองคือชุดของยีนที่มีการแสดงออกไม่แตกต่างกัน ซึ่งต่อมามีการใช้วิธีตาราง 2x2 มาเพิ่มความสะดวกให้กับชุดของยีนทั้งสี่ชุดดังกล่าว โดยการทดสอบทางสถิติสามารถนำมาใช้ประเมินผลว่าความแตกต่างระหว่างการกระจายแบบไม่ต่อเนื่องของยีนที่เราสนใจและยีนที่เราไม่สนใจมีนัยสำคัญทางสถิติ (Goeman and Bühlmann, 2005) แบบที่สองคือ การให้คะแนนในระดับฟังก์ชัน (The functional class scoring) เป็นการรวบรวมผลที่เกิดขึ้นกับยีนที่เกิดจาก 3 ขั้นตอนดังต่อไปนี้ ในขั้นตอนแรกสถิติระดับยีน (A gene-level statistic) เป็นการคำนวณจากข้อมูลการแสดงออกของยีน ในขั้นตอนที่สองสถิติระดับยีนสำหรับทุก ๆ ยีนถูกรวบรวมจากสถิติระดับวิถี (A pathway-level statistic) และเนื่องจากการกระจายของสถิติระดับวิถีไม่สอดคล้องกับการแจกแจงมาตรฐานใด ๆ ดังนั้นในขั้นตอนที่สามการทดสอบการสับเปลี่ยน (A permutation test) จึงถูกนำมาใช้ในการประเมินนัยสำคัญทางสถิติของสถิติระดับวิถี ในทางเดียวกันกับแบบที่สามคือ ฐานโครงสร้างทางเดิน (The pathway topology - based) ที่มีการพิจารณาผลของยีนที่เกิดจากสามขั้นตอนเหมือนการให้คะแนนระดับฟังก์ชัน อย่างไรก็ตามสิ่งที่แตกต่างกันคือ ฐานโครงสร้างทางเดินต้องใช้ความรู้เกี่ยวกับโครงสร้างทางเดิน (Pathway topology)

มาร่วมด้วย ซึ่งรายละเอียดเกี่ยวกับฐานโครงสร้างทางเดินจะถูกถกเถียงเพิ่มเติมโดย Nguyen et al. (2018)

ถึงแม้ว่าการระบุข้อมูลของยีนผ่านการจำแนกและการวิเคราะห์บทบาทวิถีมีเป้าหมายที่แตกต่างกันและให้ผลลัพธ์ที่ไม่เหมือนกัน แต่เป็นไปได้ว่าการวิเคราะห์บทบาทวิถีจะยืมจุดเด่นของการระบุข้อมูลของยีนผ่านการจำแนกมาใช้ โดยเฉพาะอย่างยิ่งค่าสัมประสิทธิ์ของยีนที่เป็นข้อมูลเข้าสำหรับการจำแนกที่เป็นเชิงเส้นสามารถใช้สถิติระดับยีนในการให้คะแนนระดับฟังก์ชัน เมื่อมีการทำให้เป็นปรกติจะมียีนที่ไม่มีความสำคัญต่อการจำแนก ซึ่งได้แก่ ยีนที่ไม่มีความแตกต่างในการแสดงออก จะสามารถลดค่าสัมประสิทธิ์ให้เหลือศูนย์ได้ในทางตรงกันข้ามจะมียีนที่มีสหพันธ์กันซึ่งจำเป็นต่อการจำแนก ได้แก่ ยีนที่มีความแตกต่างกันในการแสดงออก ค่าสัมประสิทธิ์จะไม่เป็นศูนย์เมื่อผ่านการทำให้เป็นปรกติ คุณสมบัตินี้เองของการทำให้เป็นปรกติสำหรับค่าสัมประสิทธิ์การจำแนกที่ได้เปรียบมากกว่าการใช้สถิติแบบตัวแปรเดียว (Univariate statistics) ที่ซึ่งมีประสิทธิภาพมากกว่าหรือเท่าเทียมกับการใช้สถิติแบบหลายตัวแปร (Multivariate statistics) ในการวิเคราะห์การแสดงออกของยีน (Glazko and Emmert-Streib, 2009) ที่สถิติระดับยีน ดังนั้นจึงเหมาะสมอย่างมากในการรวมค่าสัมประสิทธิ์เข้าไปในสถิติระดับวิถี

ในปริณยานุพจน์นี้นำเสนอวิธีการใช้ค่าสัมประสิทธิ์ของยีนที่เป็นข้อมูลเข้าสำหรับการจำแนกที่เป็นเชิงเส้นที่สถิติระดับยีนในวิธีการให้คะแนนระดับฟังก์ชันดังที่ได้กล่าวไปแล้ว การจำแนกที่เป็นเชิงเส้นในการศึกษานี้ใช้แบบจำลองการถดถอยลอจิสติก (Logistics Regression Model) ค่าสัมประสิทธิ์ของยีนที่เป็นข้อมูลเข้าถูกทำให้เป็นปรกติโดยใช้การถดถอยริดจ์ (Ridge Regression) (Le Cessie and Van Houwelingen, 1992) การหาค่าสมบรูณ์น้อยสุดและตัวดำเนินการเลือกหรือแลชโซ (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) และข่ายยืดหยุ่น (Elastic Net) (Zou and Hastie, 2005) โดยการคัดเลือกข้อมูลของยีนสามารถดำเนินการได้ก็ต่อเมื่อใช้แลชโซหรือข่ายยืดหยุ่น ในขณะที่สหสัมพันธ์ระหว่างข้อมูลยีนจะถูกตรวจสอบได้ก็ต่อเมื่อใช้การถดถอยริดจ์หรือข่ายยืดหยุ่น และสุดท้ายการวิเคราะห์นัยสำคัญของฟังก์ชันและการแสดงออก (A significance analysis of function and expression) หรือ SAFE เป็นเครื่องมือที่ใช้สำหรับการให้คะแนนระดับฟังก์ชัน (Barry et al., 2005) SAFE ประสบความสำเร็จในการประยุกต์ใช้กับข้อมูลการแสดงออกของยีนจากการศึกษามะเร็งปอด (Barry et al., 2005) โรคซึมเศร้า (Jansen et al., 2016) เส้นทางการตอบสนองความเสียหายของดีเอ็นเอที่ควบคุมการเชื่อมต่อ mRNAs (Chenet et al., 2017) ชีสติก ไฟโบรซิส (Polineni et al., 2018) และแพร่หลายอย่างกว้างขวางโดย

Pounds et al. (2007) ด้วย 5 ชุดข้อมูลการแสดงผลของยีนที่ซึ่งมีสองคลาสต่อตัวอย่างถูกเลือกให้นำเสนอ ได้แก่ มะเร็งเม็ดเลือดขาว (Golub et al., 1999), มะเร็งปอด (Bhattacharjee et al., 2001), เนื้อเยื่อที่ระบบประสาทส่วนกลาง (Pomeroy et al., 2002), มะเร็งต่อมลูกหมาก (Singh et al., 2002) และมะเร็งสมอง (Nutt et al., 2003)

เนื้อหาในปริญญานิพนธ์นี้ประกอบไปด้วย 3 บทคือ บทที่ 2 อธิบายเกี่ยวกับทฤษฎีที่ใช้ในการสร้างแบบจำลองและเทคนิคต่าง ๆ ที่ใช้ในงานวิจัย บทที่ 3 แสดงผลการทดลองของแต่ละชุดข้อมูลผ่านการวิเคราะห์บรรณนิทัศน์และแสดงเวลาที่ใช้ในการประมวลผลของแต่ละชุดข้อมูล และแต่ละเทคนิคที่ใช้ในงานวิจัย บทที่ 4 สรุปผลการทดลองของปริญญานิพนธ์นี้

บทที่ 2

ทฤษฎี

2.1 การถดถอยโลจิสติก (Logistic Regression)

2.2.1 หลักการของการถดถอยโลจิสติก

การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) เป็นการวิเคราะห์ตัวแปรเชิงพหุที่มีวัตถุประสงค์เพื่อทำนายเหตุการณ์ที่สนใจหรือประมาณค่าว่าจะเกิดเหตุการณ์นั้นหรือไม่ ภายใต้อิทธิพลของตัวปัจจัย แบบจำลองการถดถอยโลจิสติก (Logistic Regression Model) เป็นตัวจำแนกเชิงเส้น (Linear Classifier) ที่ให้ข้อมูลออกเป็นค่าวิฤต (Discrete Value) หรือคลาส (Class) และรับข้อมูลเข้าหรือตัวแปรอิสระที่อาจมีตัวแปรเดียวหรือหลายตัวแปร โดยการถดถอยโลจิสติกที่ให้ข้อมูลออกเป็นสองคลาสเรียกว่าการถดถอยโลจิสติกทวินาม (Binomial Logistic Regression) และการถดถอยโลจิสติกที่ให้ข้อมูลออกเป็นหลายคลาสเรียกว่าการถดถอยโลจิสติกพหุนาม (Multinomial Logistic Regression) การถดถอยโลจิสติกเป็นเครื่องมือวิเคราะห์ข้อมูลที่มีวัตถุประสงค์เพื่อประเมินความเสี่ยงหรือทำนายเหตุการณ์ จึงมีการประยุกต์ใช้ในงานวิจัยหลากหลายสาขา อาทิ สาขาทางการแพทย์ วิศวกรรมศาสตร์ และเศรษฐศาสตร์ เป็นต้น

2.2.2 แบบจำลองการถดถอยโลจิสติก

การวิเคราะห์การถดถอยโลจิสติกเป็นการประมาณค่าความน่าจะเป็นของเหตุการณ์ Prob (Event) ที่จะเกิดขึ้น โดยมีต้นแบบมาจากฟังก์ชันโลจิสติก ซึ่งหากมีตัวแปรอิสระเพียงตัวเดียว ฟังก์ชันโลจิสติกจะมีสมการ ดังนี้

$$Prob(event) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad (2-1)$$

โดยที่ β_0 เป็นค่าคงที่

β_1 เป็นสัมประสิทธิ์ของตัวแปรอิสระ

X เป็นตัวแปรอิสระ

e เป็นลอการิธึมธรรมชาติ (ค่าประมาณ 2.71828...)

ในกรณีที่ตัวแปรอิสระหลายตัว ฟังก์ชันโลจิสติกจะมีสมการ ดังนี้

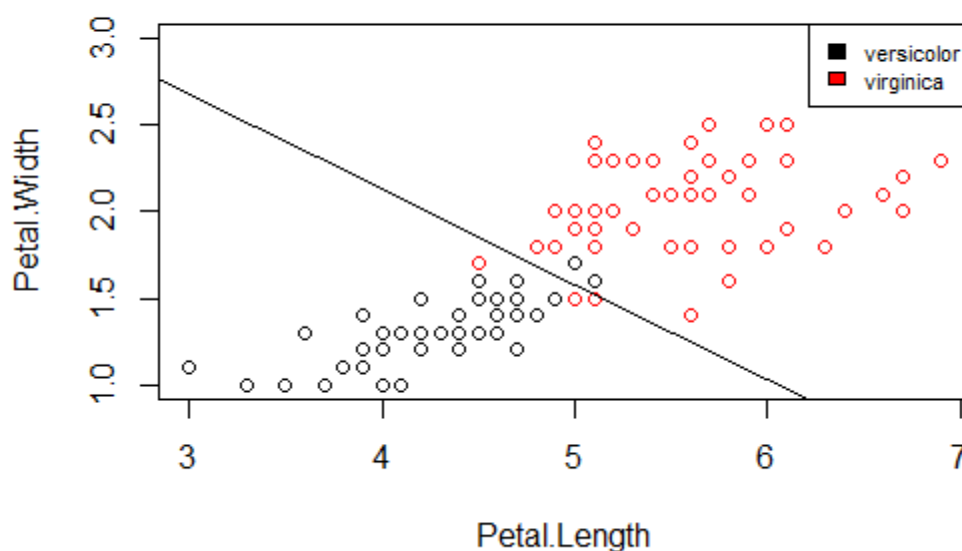
$$\text{Prob(event)} = \frac{1}{1 + e^{-z}} \quad (2-2)$$

โดยที่ Z อยู่ในรูปของ

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (2-3)$$

ถ้าแทน Prop(event) ด้วย y และแทนค่า Z ตามสมการที่ 2-3 เข้าไปในสมการที่ 2-2 จะได้สมการของแบบจำลอง ดังนี้

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (2-4)$$



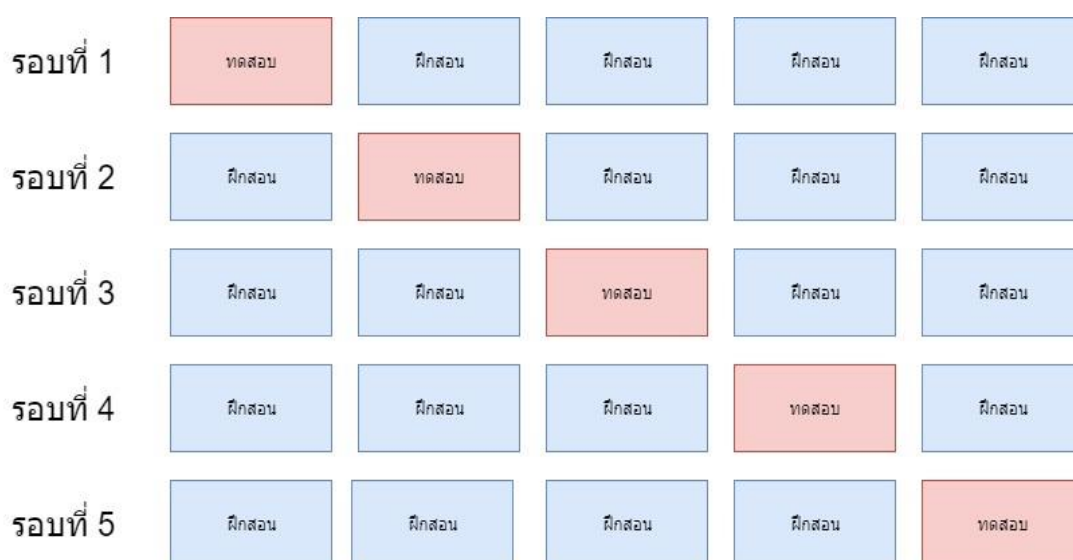
ภาพที่ 2-1 ตัวอย่างการจำแนกสปีชีส์ของดอกไอริส

2.2.3 การตรวจสอบความสมเหตุสมผลไขว้ (k-Fold Cross-Validation)

การตรวจสอบความสมเหตุสมผลไขว้เป็นวิธีประมาณค่าคลาดเคลื่อนนอกตัวอย่างที่ได้รับ ความนิยมมากที่สุดวิธีหนึ่ง ตัวอย่างเช่น ชุดข้อมูลจำนวน 15 ตัว หากกำหนด $k = 5$ (ส่วนทบ 5 ส่วน) หมายความว่า จะทำการแบ่งข้อมูลออกเป็น 5 ชุด โดยจะใช้ข้อมูล 4 ชุดใด ๆ เป็นชุดฝึกสอน ส่วน

อีกหนึ่งชุดที่เหลือเป็นชุดทดสอบ โดยจะสลับชุดทดสอบให้ครบทั้ง 5 ชุด ดังนั้นการตรวจสอบความสมเหตุสมผลไขว้จึงใช้ข้อมูลที่มีทั้งหมดในการประเมินสมรรถนะการจำแนก

ข้อดีของการตรวจสอบความสมเหตุสมผลไขว้คือ การประมาณค่าคลาดเคลื่อนนอกตัวอย่างมีความเอนเอียงและความแปรปรวนน้อย อีกทั้งยังใช้การคำนวณไม่มาก ซึ่งในการประมาณค่าคลาดเคลื่อนนอกตัวอย่างโดยปกตินิยมใช้ $k = 5, 10$ เนื่องจากมีความเอนเอียงและความแปรปรวนไม่สูง



ภาพที่ 2-2 แสดงกระบวนการทำงานของ 5 fold cross validation

2.2 การถดถอยลงโทษ (Penalized Regression)

การถดถอยลงโทษ เป็นวิธีที่นิยมใช้กันอย่างแพร่หลายในการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระ β เมื่อข้อมูลมีมิติสูง ซึ่งตัวประมาณดังกล่าวจะหาได้จากการหาค่า β ที่ทำให้ฟังก์ชันเป้าหมาย (Objective Function) มีค่าต่ำสุด ดังสมการ

$$\hat{\beta} = \arg \min_{\beta} \|Y - \sum_{j=1}^p X_j \beta_j\|^2 + P_{\lambda}(\beta) \quad (2-5)$$

จากสมการที่ 2-5 จะสังเกตได้ว่ามีฟังก์ชัน $P_{\lambda}(\beta)$ ซึ่งเรียกว่า ฟังก์ชันการลงโทษ (Penalty Function) เพื่อใช้ในการให้น้ำหนักของฟังก์ชันการลงโทษ ดังกล่าว

สำหรับฟังก์ชันการลงโทษนั้นมียู่อด้วยกันหลายรูปแบบ ซึ่งแต่ละวิธีให้แบบจำลองการถดถอยโลจิสติกที่แตกต่างกัน ในปริณญานิพนธ์นี้ใช้ 3 วิธีคือ

2.2.1 การถดถอยริดจ์ (Ridge Regression)

การถดถอยริดจ์ เป็นวิธีลดขนาดสัมประสิทธิ์ของตัวแปรอิสระ β ซึ่งจะทำให้สัมประสิทธิ์ของตัวแปรที่ไม่จำเป็นสำหรับแบบจำลองการถดถอยโลจิสติกมีค่าน้อยแต่ไม่เท่ากับศูนย์ โดยการถดถอยริดจ์มีฟังก์ชันการลงโทษ ดังสมการ

$$\lambda \sum_{j=1}^p \beta_j^2 \quad (2-6)$$

ข้อดีของการถดถอยริดจ์คือ เหมาะสมสำหรับปัญหาที่ตัวแปรมีสหสัมพันธ์สูง แต่มีข้อเสียที่ขาดคุณสมบัติในการคัดเลือกตัวแปร

2.2.2 การหาค่าสัมบูรณ์น้อยสุดและตัวดำเนินการเลือก (Least Absolute Shrinkage and Selection Operator) หรือ แลซโซ (Lasso)

แลซโซเป็นวิธีที่สามารถคัดเลือกตัวแปรเข้าแบบจำลอง (Model) โดยแบบจำลองการถดถอยโลจิสติกที่ได้จากแลซโซจะมีสัมประสิทธิ์ β ส่วนใหญ่เป็นศูนย์ และสัมประสิทธิ์ β บางส่วนไม่เท่ากับศูนย์ ซึ่งฟังก์ชันการลงโทษของแลซโซมีสมการดังนี้

$$\lambda \sum_{j=1}^p |\beta_j| \quad (2-7)$$

ข้อดีของแลซโซคือ สามารถคัดเลือกตัวแปรที่จำเป็นสำหรับแบบจำลองการถดถอยโลจิสติก แต่มีข้อเสียตรงที่ในกรณีที่ตัวแปรมีสหสัมพันธ์สูง แลซโซมีแนวโน้มที่จะคัดเลือกเพียงตัวแปรเดียวจากกลุ่มตัวแปรที่มีสหสัมพันธ์สูงโดยไม่สนใจว่าเป็นตัวแปรใด

2.2.3 ข่ายยืดหยุ่น (Elastic Net)

ข่ายยืดหยุ่นเป็นวิธีที่รวมข้อดีของการถดถอยริดจ์และแลซโซเข้าด้วยกัน กล่าวคือสามารถคัดเลือกตัวแปรได้เช่นเดียวกับแลซโซ และยังเหมาะสมสำหรับปัญหาที่ตัวแปรมีสหสัมพันธ์สูงเช่นเดียวกับการถดถอยริดจ์ โดยข่ายยืดหยุ่นมีฟังก์ชันการลงโทษดังสมการ

$$\lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j| \quad (2-8)$$

ตารางที่ 2-1 เปรียบเทียบข้อดี - ข้อเสีย และความเหมาะสมในการใช้งานของแต่ละวิธี

วิธี	ข้อดี	ข้อเสีย	ความเหมาะสมในการใช้งาน
การถดถอย ริดจ์	แก้ไขปัญหาครณีตัวแปรอิสระมีสหสัมพันธ์สูง	ขาดคุณสมบัติในการคัดเลือกตัวแปร	เหมาะสำหรับข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กที่ไม่เท่ากับศูนย์จำนวนมาก
แลชโซ	มีความสามารถในการคัดเลือกตัวแปร	กรณีที่ตัวแปรอิสระมีสหสัมพันธ์สูง วิธีแลชโซจะเลือกตัวแปรเข้าสู่แบบจำลองเพียงตัวแปรเดียวจากกลุ่มที่มีสหสัมพันธ์สูงโดยไม่สนใจว่าจะเป็นตัวแปรใด	เหมาะสำหรับข้อมูลที่มีสัมประสิทธิ์ขนาดกลางที่ไม่เท่ากับศูนย์จำนวนน้อยถึงปานกลาง
ข่ายยัดหยุ่น	มีความสามารถในการคัดเลือกตัวแปรและแก้ปัญหาคตัวแปรที่มีสหสัมพันธ์สูง อีกทั้งยังสามารถแก้ไขข้อเสียของวิธีแลชโซได้	ใช้เวลาทำงานนานกว่าวิธีการถดถอยริดจ์และแลชโซ เนื่องจากมีฟังก์ชันการลงโทษ 2 เทอม	เหมาะสำหรับข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กถึงปานกลางที่ไม่เท่ากับศูนย์จำนวนปานกลาง หรือข้อมูลที่ตัวแปรมีสหสัมพันธ์สูง

2.3 การวิเคราะห์นัยสำคัญของฟังก์ชันและการแสดงออก (Significance Analysis of Function and Expression) หรือ SAFE

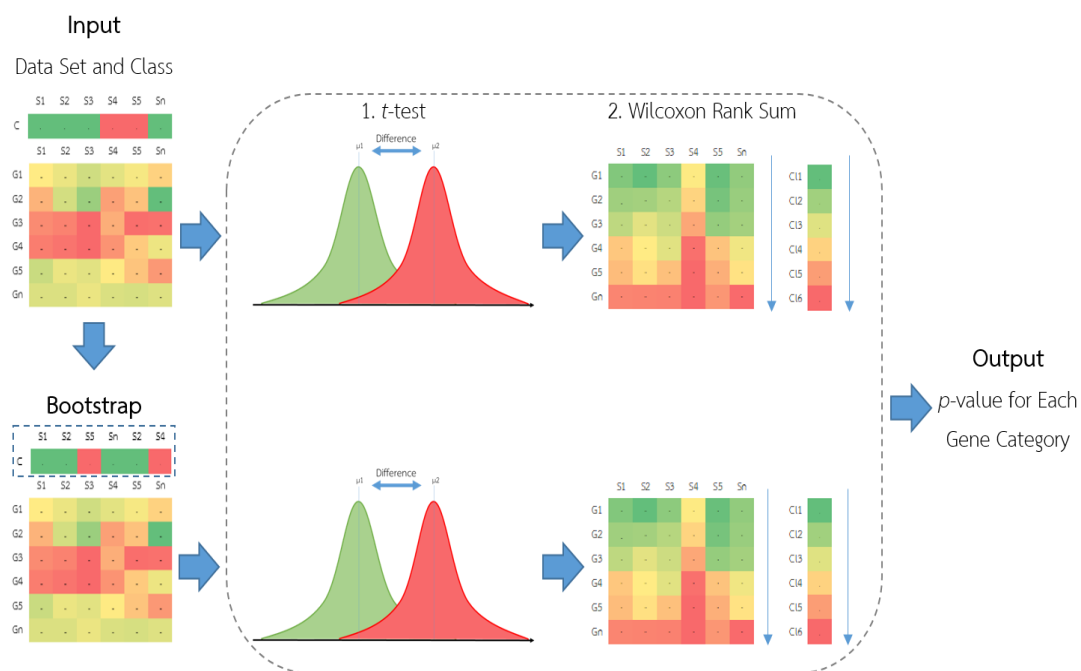
SAFE นั้นเป็นเครื่องมือในการให้คะแนนระดับฟังก์ชันของการวิเคราะห์บาทวิถี (Barry et al., 2015) โดยที่ SAFE ประกอบด้วยสามขั้นตอนดังนี้ การคำนวณสถิติระดับยีน การคำนวณสถิติระดับวิถี และการประเมินนัยสำคัญทางสถิติของสถิติระดับวิถีสำหรับชุดข้อมูลสองคลาส เช่น ชุดข้อมูลกรณีการควบคุม โดยที่ค่าสมบูรณ์ของสถิติ t สองตัวอย่าง เป็นการคำนวณจากทุก ๆ ตัวอย่างของแต่ละยีนและถูกใช้ในสถิติระดับยีน จากนั้นสถิติระดับวิถีจะคำนวณเพื่อประเมินความแตกต่างระหว่างการกระจายแบบวิเศษของสถิติระดับยีนสำหรับยีนในและนอกวิถีหรือหมวดหมู่ที่สนใจ สถิติผลรวมอันดับของวิลคอกซันและโคลโมโกรอฟ-สเมอ (Wilcoxon rank sum and Kolmogorov-Smirnov statistics) ได้รับการแนะนำให้ใช้สำหรับสถิติระดับวิถี ความสำคัญของสถิติระดับวิถีสำหรับแต่ละวิถีนั้นถูกกำหนดให้ผ่านการทดสอบการสับเปลี่ยนหรือการบูตสแตรป์ โดยการสับเปลี่ยนแต่ละครั้งจะถูกสร้างขึ้นจากชุดข้อมูลดั้งเดิมเพื่อให้ลักษณะคลาสของแต่ละตัวอย่างถูกสับเปลี่ยนอย่างสุ่มในขณะที่จำนวนของตัวอย่างในแต่ละคลาสยังคงเท่าเดิมในทุก ๆ ครั้งของการทดสอบการสับเปลี่ยน จากนั้นความสำคัญของสถิติระดับวิถีจะได้ออกจากการเปรียบเทียบสถิติระดับวิถีที่ได้จากชุดข้อมูลดั้งเดิมกับชุดข้อมูลที่ถูกสับเปลี่ยน ในทางตรงกันข้ามแต่ละการทำบูตสแตรป์ถูกสร้างจากการสุ่มตัวอย่างแบบแทนที่ด้วยชุดข้อมูลเดิมในข้อมูลบูตสแตรป์ หลังจากนั้นความสำคัญของสถิติระดับวิถีจะได้ออกจากการกระจายของสถิติระดับวิถีที่ได้จากชุดข้อมูลดั้งเดิมและการทำซ้ำแบบบูตสแตรป์ Barry et al. (2008) แสดงให้เห็นว่าการทำบูตสแตรป์ให้น้ำหนักทางสถิติที่สูงกว่าการทดสอบการสับเปลี่ยนเมื่อประเมินความสำคัญของสถิติผลรวมอันดับของวิลคอกซัน (Wilcoxon rank sum statistics) ที่ซึ่งใช้ในสถิติระดับวิถี นอกจากนี้หากมีการสันนิษฐานว่าการกระจายของสถิติ t ที่ถูกใช้ในสถิติระดับยีนและสถิติระดับวิถีที่ผ่านการทำบูตสแตรป์นั้นเป็นการประมาณปกติแล้วค่า p -values ของสถิติระดับยีนและสถิติระดับวิถีสามารถหาได้จากการกระจาย t ที่มีองศาอิสระเท่ากับจำนวนตัวอย่างของยีนที่มีการแสดงออกด้วยหนึ่งและได้รับจาก

$$p_{\text{gene}} = 1 - \Phi \left(\left| \sum_{i=1}^B v_i^* / B \right| / \sqrt{\frac{1}{B-1} \left(\sum_{i=1}^B v_i^{*2} - \left(\sum_{i=1}^B v_i^* \right)^2 / B \right)} \right) \quad (2-9)$$

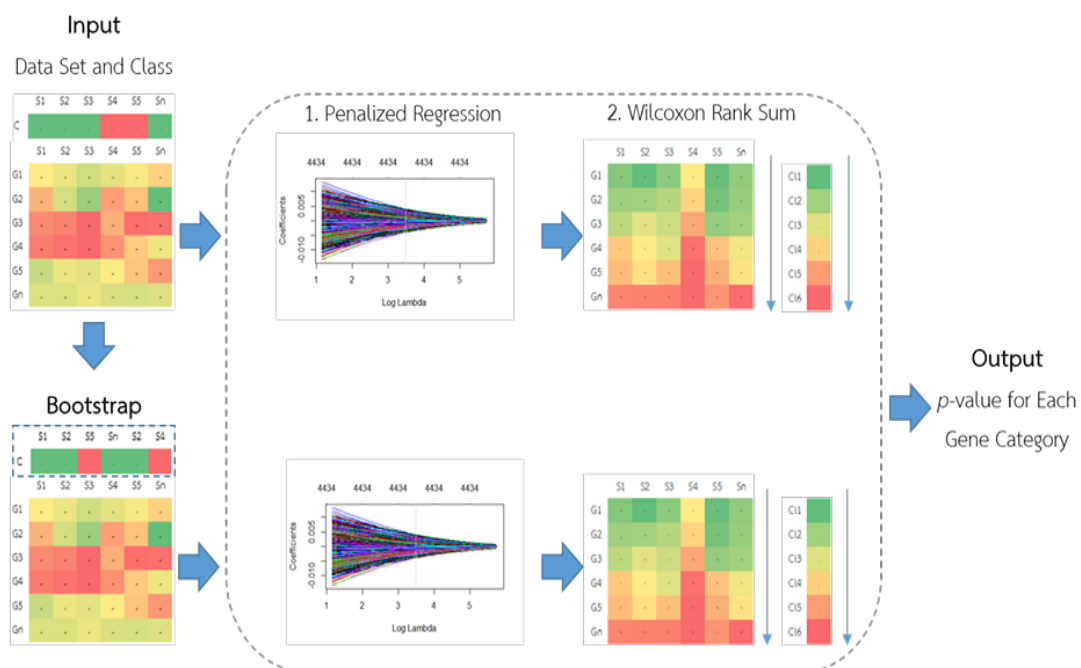
และ

$$P_{\text{pathway}} = 1 - \Phi \left(\left(\sum_{i=1}^B u_i^* / B - E_{H_0}(u) \right) / \sqrt{\frac{1}{B-1} \left(\sum_{i=1}^B u_i^{*2} - \left(\sum_{i=1}^B u_i^* \right)^2 / B \right)} \right) \quad (2-10)$$

ตามลำดับ Φ คือฟังก์ชันการกระจายแบบสะสม t และ B คือจำนวนการทำซ้ำแบบบูตสเตร็ปซึ่งถูกกำหนดที่ 200 ครั้งตามคำแนะนำของ Barry et al. (2008) v_i^* คือสถิติ t ที่ถูกใช้ในสถิติระดับยีนซึ่งได้มาจากการทำซ้ำครั้งใด ๆ ของบูตสเตร็ป u_i^* เป็นสถิติระดับวิถีที่ได้มาจากการทำซ้ำครั้งใด ๆ ของบูตสเตร็ป โดยการทำการซ้ำครั้งแรกของบูตสเตร็ปคือชุดข้อมูลดั้งเดิม $E_{H_0}(u)$ คือค่าคาดหวังของสถิติระดับวิถีภายใต้สมมติฐานว่างที่ซึ่งมีการกระจายแบบวิยคของสถิติระดับยีนเหมือนกันทั้งในและนอกวิถี $E_{H_0}(u)$ ถูกกำหนดโดย $m_c(m+1)/2$ ซึ่ง m คือจำนวนของยีนและ m_c คือจำนวนยีนในวิถีหรือหมวดหมู่ที่สนใจเมื่อ Wilcoxon rank sum ถูกใช้ในสถิติระดับวิถี และเนื่องจากมีหลายวิถีที่อยู่ภายใต้การพิจารณา จึงจำเป็นต้องมีการแก้ไขความถูกต้องสำหรับการทดสอบหลายสมมติฐาน โดยทั่วไปแล้วอัตราความผิดพลาดต่อวงศ์ (Family - Wise Error Rate) หรือ FWER (Westfall and Young, 1989) หรืออัตราการค้นพบเท็จ (False Discovery Rate) หรือ FDR (Yekutieli and Benjamini, 1999) สามารถประมาณค่า p -values ของสถิติระดับวิถีได้ Barry et al. (2005) แสดงให้เห็นว่า FWER สามารถควบคุมผ่านการแก้ไขความถูกต้องของ Bonferroni ได้ ในขณะที่ FDR สามารถควบคุมผ่านขั้นตอนที่อธิบายโดย Benjamini and Hochberg (1995) ได้ ซึ่งการประมาณค่าของ FDR ที่แนะนำโดย Storey and Tibshirani (2003) สำหรับการวิเคราะห์การแสดงออกของยีนจะได้รับการศึกษาในงานวิจัยนี้ ทั้งนี้วิธีการของ SAFE ดั้งเดิมในปฏิญานีพจน์นี้จะเรียกว่า SAFE ส่วนวิธีการ SAFE ที่ถูกพัฒนาด้วยการลดออลงโทษทั้ง 3 แบบ ได้แก่ การลดออลยริดจ์ แลชโซ และ ข่ายยัดหยุ่น จะเรียกว่า rSAFE, ISAFE และ eSAFE ตามลำดับ



ภาพที่ 2-3 การแสดงกระบวนการทำงานของ SAFE



ภาพที่ 2-4 การแสดงกระบวนการทำงานของ rSAFE

2.4 ขั้นตอนการสร้างแบบจำลอง

งานวิจัยนี้สร้างมาจากการเขียนโปรแกรมด้วยภาษา R ทั้งหมด โดยกระบวนการทำงานอิงจากภาพที่ 2-4 จะได้ว่าเริ่มต้นจากการนำข้อมูลดั้งเดิม (Original Data) เข้าสู่กระบวนการของการถดถอยรีดจ์เพื่อทำการลดค่าสัมประสิทธิ์ของยีนที่ไม่มีความสำคัญต่อการจำแนกให้เข้าใกล้ศูนย์ จากนั้นนำยีนที่ผ่านกระบวนการของการถดถอยรีดจ์มาจัดหมวดหมู่ของยีน เรียกว่า หมวดของยีน (Gene Category) แล้วนำหมวดของยีนที่ได้นั้นเข้ากระบวนการสถิติผลรวมอันดับของวิลคอกชันเพื่อจัดอันดับหมวดของยีน ในทำนองเดียวกันนั้นอีกด้านหนึ่งของกระบวนการจะนำข้อมูลดั้งเดิมมาทำการสับเปลี่ยนด้วยวิธีบูตสแตรป์ แล้วนำข้อมูลที่สับเปลี่ยนนั้นเข้าสู่กระบวนการของการถดถอยรีดจ์ จัดหมวดหมู่ของยีนและเข้ากระบวนการสถิติผลรวมอันดับของวิลคอกชัน ตามลำดับเช่นเดียวกัน สุดท้ายจึงนำหมวดหมู่ยีนที่ได้จากกระบวนการสถิติผลรวมอันดับของวิลคอกชันทั้งสองด้านของกระบวนการมาหาค่า p -value ด้วยการหาผลรวมของจำนวนครั้งที่อันดับของข้อมูลดั้งเดิมมีค่ามากกว่าหรือเท่ากับอันดับของข้อมูลที่ถูกสับเปลี่ยนด้วยวิธีบูตสแตรป์หารด้วยจำนวนครั้งที่ทำการสับเปลี่ยนด้วยวิธีบูตสแตรป์ ดังสมการ

$$p_{kl} = \frac{1}{K} \sum_{h=1}^K I\{v_{hl} \geq v_{kl}\} \quad (2-11)$$

โดยที่ p_{kl} คือค่า p -value

K คือจำนวนครั้งของการสับเปลี่ยนด้วยวิธีบูตสแตรป์

v_{hl} คืออันดับของข้อมูลดั้งเดิม

v_{kl} คืออันดับของข้อมูลที่ถูกสับเปลี่ยนด้วยวิธีบูตสแตรป์

งานวิจัยนี้ใช้ชุดข้อมูล 6 ชุด แต่ละชุดข้อมูลประกอบไปด้วย 2 คลาส โดยก่อนที่จะนำชุดข้อมูลเหล่านี้ไปใช้จะต้องทำการประมวลผลก่อน ซึ่งจะทำให้จำนวนยีนลดลงจากชุดข้อมูลดั้งเดิม เนื่องจากเราสนใจยีนที่อยู่ในช่วงที่เรากำหนดไว้เท่านั้น แหล่งที่มาของชุดข้อมูลและวิธีการประมวลผลก่อนแสดงดังตารางที่ 2-2

ตารางที่ 2-2 แหล่งที่มาของชุดข้อมูลและวิธีการประมวลผลก่อน

Dataset	Dataset Publication	Preprocess Publication
Central Nervous System Embryonal Tumor	Pomeroy et al., 2002	Pomeroy et al., 2002
Brain Cancer	Nutt et al., 2003	Chang et al., 2007
Lung Carcinoma	Bhattacharjee et al., 2001	-
Leukemia	Golub et al., 1999	Dudoit et al., 2002
Prostate Cancer	Singh et al., 2002	Singh et al., 2002
Prostate Cancer (Capsular penetration)	Singh et al., 2002	Singh et al., 2002

ข้อมูลในแต่ละชุดที่ได้มาจะถูกนำมาทำการประมวลผลก่อนโดยมีขั้นตอนดังนี้

1. ปรับค่าขีดเริ่มเปลี่ยน (Thresholding) ด้วยวิธีการจำกัดค่าที่เป็นไปได้สูงสุด (Ceiling) และต่ำสุด (Flooring)
2. ทำการกรอง (Filtering) ยีนที่มีค่ามากกว่าค่าที่กำหนด 2 ค่า นั่นคือค่าของผลต่างระหว่างค่ามากที่สุดกับค่าน้อยสุด (Max-Min) และผลหารระหว่างค่ามากที่สุดกับค่าน้อยสุด (Max/Min)
3. ทำการแปลงค่าของยีนด้วยลอการิทึมฐาน 10
4. ทำการทำให้เป็นมาตรฐาน (Standardize) ระหว่างยีน

โดยรายละเอียดของการปรับค่าขีดเริ่มเปลี่ยนและการกรองของแต่ละชุดข้อมูลแสดงดังตารางที่ 2-3

ตารางที่ 2-3 การประมวลผลก่อน

Dataset	Flooring	Ceiling	Max-Min	Max/Min
Central Nervous System Embryonal Tumor	20	16,000	500	5
Brain Cancer	20	16,000	100	3
Lung Carcinoma	20	16,000	-	-
Leukemia	100	16,000	500	5
Prostate Cancer	10	16,000	50	5
Prostate Cancer (Capsular penetration)	10	16,000	50	5

หลังจากผ่านการทำกรประมวลผลก่อนแล้วจำนวนยีนที่เหลือถูกแสดงดังตารางที่ 2-4

ตารางที่ 2-4 จำนวนยีนที่เหลือหลังทำการประมวลผลก่อน

Dataset	Original Gene	Preprocessed Gene	Sample
Central Nervous System Embryonal Tumor	7,129	4,739	60
Brain Cancer	12,625	4,434	50
Lung Carcinoma	12,600	3,312	203
Leukemia	7,129	3,571	72
Prostate Cancer	12,600	5,966	102
Prostate Cancer (Capsular penetration)	12,600	4,040	49

บทที่ 3

ผลการทดลอง

การวิเคราะห์ผลที่เกิดขึ้นต้องใช้ผลการทดลองจากการทำงานของเทคนิค SAFE ที่ใช้ t-test กับ rSAFE, eSAFE, lSAFE ในการคัดเลือกลักษณะประจำแทนการใช้ t-test เพื่อนำมาเปรียบเทียบผลลัพธ์ในการหาหมวดหมู่ของยีนที่สำคัญ

3.1 ผลลัพธ์จากการคัดเลือกลักษณะประจำ

จากวิธีการในบทที่ 2 หัวข้อที่ 2.4 จำนวนยีนที่สำคัญที่ผ่านการคัดเลือกจะได้จำนวน ดังตารางที่ 3-1

ตารางที่ 3-1 จำนวนยีนที่สำคัญที่ผ่านการคัดเลือกในแต่ละชุดข้อมูล โดยใช้ Elastic Net ที่ $\alpha = 0.1$

Dataset	Original	Ridge Regression	Elastic Net	Lasso
Central Nervous System Embryonal Tumor	4,739	4,739	0	0
Brain Cancer	4,434	4,434	85	8
Lung Carcinoma	3,312	3,312	111	7
Leukemia	3,571	3,571	117	6
Prostate Cancer	5,966	5,966	95	10
Prostate Cancer (Capsular penetration)	4,040	4,040	41	0

จากตารางที่ 3-1 จะเห็นว่าจำนวนยีนที่สำคัญที่ผ่านการคัดเลือกในแต่ละชุดข้อมูลนั้นมีแนวโน้มไปในทางเดียวกัน นั่นคือจำนวนยีนของวิธี Ridge Regression จะมีจำนวนยีนที่ผ่านการคัดเลือกมากที่สุดซึ่งมีจำนวนเท่ากับวิธี Original รองลงมาคือวิธี Elastic Net และน้อยที่สุดคือวิธี Lasso เนื่องจาก

จำนวนยีนที่สำคัญที่ผ่านการคัดเลือกเกิดจากการทำกระบวนการลดค่าสัมประสิทธิ์ของตัวแปรซึ่งวิธี Ridge Regression จะไม่มีตัวแปรใดที่ค่าสัมประสิทธิ์เป็นศูนย์จึงเหลือจำนวนยีนที่ผ่านการคัดเลือกมากที่สุด ในทางตรงกันข้ามกับวิธี Lasso ที่มีการลดค่าสัมประสิทธิ์จนเหลือศูนย์หลายตัวแปรจึงทำให้จำนวนยีนที่ผ่านการคัดเลือกมีน้อยที่สุด และวิธี Elastic Net เป็นวิธีที่ผสมผสานทั้งสองวิธีก่อนหน้าเข้าด้วยกันจึงทำให้จำนวนยีนที่ผ่านการคัดเลือกอยู่กึ่งกลางระหว่างวิธีทั้งสอง

โดยวิธีข้างยี้ดยุ่นจะแสดงจำนวนยีนที่สำคัญที่พบมากที่สุดเทคนิค eSAFE ซึ่งอ้างอิงจากการเลือกค่าการให้น้ำหนักของฟังก์ชันการลงโทษ (α) ที่ทำให้ค้นพบยีนที่สำคัญมากที่สุดจากตารางที่ 3-2

ตารางที่ 3-2 จำนวนของยีนที่สำคัญที่ค้นพบจากเทคนิค eSAFE โดยค่าที่อยู่ในวงเล็บคือ α ที่ใช้ในการคัดเลือกลักษณะประจำ

Dataset	Number of genes								
	(0.1)	(0.2)	(0.3)	(0.4)	(0.5)	(0.6)	(0.7)	(0.8)	(0.9)
Central Nervous System Embryonal Tumor	0	0	0	0	0	0	0	0	0
Brain Cancer	85	47	32	23	18	13	12	8	8
Lung Carcinoma	111	60	43	32	28	20	13	9	10
Leukemia	117	41	30	23	21	12	11	7	7
Prostate Cancer	95	51	33	18	26	25	11	12	11
Prostate Cancer (Capsular penetration)	41	24	17	6	0	0	0	0	0

จากตารางที่ 3-2 จะเห็นว่าแนวโน้มจำนวนของยีนที่สำคัญที่พบใน eSAFE ไม่แตกต่างจากตารางที่ 3-1 นั่นคือจำนวนของยีนที่สำคัญมีแนวโน้มที่จะลดลงตามการเพิ่มของค่า α ที่ส่งผลให้ค่าสัมประสิทธิ์มีค่าเป็น 0 เพิ่มขึ้น

3.2 ผลลัพธ์จำนวนหมวดหมู่ของยีนที่สำคัญจากการทำซ้ำ

การทำซ้ำการทดลองทั้งหมด 100 ครั้งของเทคนิค SAFE, rSAFE, eSAFE, ISAFE สิ่งที่น่าสนใจคือจำนวนหมวดหมู่ของยีนที่สำคัญที่ค้นพบจากแต่ละเทคนิคที่ให้ผลตรงกัน 80 ครั้งขึ้นไปมาเปรียบเทียบกัน ซึ่งจะแสดงให้เห็นถึงความแตกต่างของหมวดหมู่ของยีนที่สำคัญที่ค้นพบในแต่ละวิธีการ ดังตารางที่ 3-3

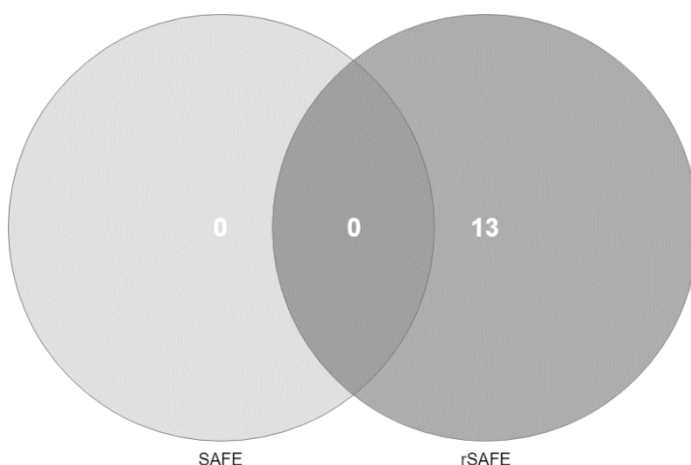
ตารางที่ 3-3 จำนวนหมวดหมู่ของยีนที่สำคัญที่พบในแต่ละเทคนิค โดยใช้ Elastic Net ที่ $\alpha = 0.1$

Dataset	Gene Categories	Total Number of Gene Categories			
		SAFE	rSAFE	eSAFE	ISAFE
Central Nervous System Embryonal Tumor	13,221	0	13	0	0
Brain Cancer	12,549	110	168	27	0
Lung Carcinoma	11,316	3,517	3,485	968	109
Leukemia	11,903	1,863	1,766	387	33
Prostate Cancer	13,812	539	671	85	57
Prostate Cancer (Capsular penetration)	11,925	0	0	0	0

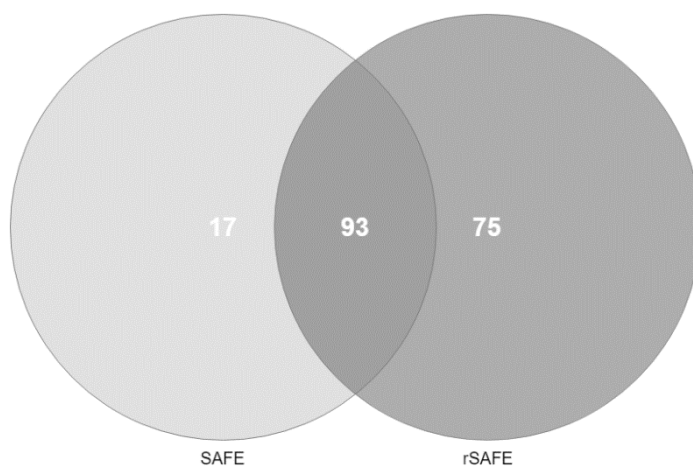
จากตารางที่ 3-3 จะเห็นว่าจำนวนหมวดหมู่ของยีนที่สำคัญที่พบในแต่ละเทคนิคนั้นมีแนวโน้มเหมือนจำนวนยีนที่สำคัญที่ผ่านการคัดเลือกในแต่ละชุดข้อมูล กล่าวคือจำนวนหมวดหมู่ยีนที่สำคัญที่พบใน rSAFE มีจำนวนมากที่สุด รองลงมาคือ eSAFE และพบน้อยที่สุดใน ISAFE เนื่องจากการหาหมวดหมู่ของยีนที่สำคัญเกิดจากการนำสัมประสิทธิ์ของตัวแปรมาจัดอันดับผลรวมของ

จากตารางที่ 3-4 จะเห็นว่าแนวโน้มหมวดหมู่ของยีนที่สำคัญที่พบใน eSAFE ไม่แตกต่างจาก ตารางที่ 3-3 นั่นคือหมวดหมู่ยีนที่สำคัญมีแนวโน้มที่จะพบลดลงตามการเพิ่มของค่า α ที่ส่งผลให้ค่าสัมประสิทธิ์มีค่าเป็น 0 เพิ่มขึ้น

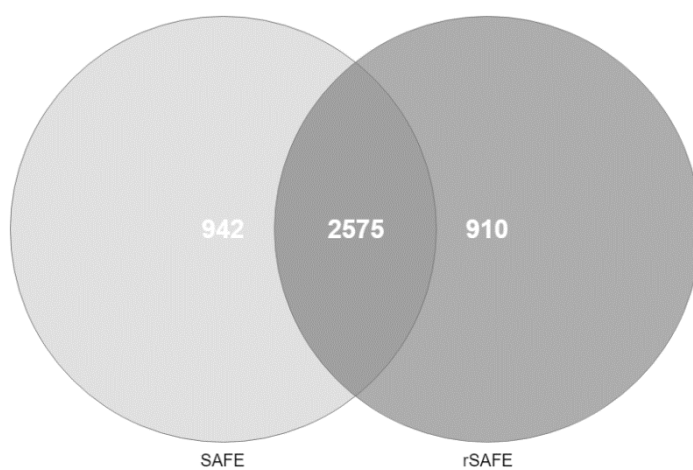
จากการหาหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ที่แสดงผลตรงกัน 80 ครั้งขึ้นไปสามารถนำมาสร้างแผนภาพเวนน์ (Venn Diagram) เพื่อแสดงให้เห็นถึงความเหมือนและแตกต่างกันของหมวดหมู่ของยีนที่สำคัญที่ค้นพบจาก 2 เทคนิคนี้ของแต่ละชุดข้อมูล ดังแสดงใน ภาพที่ 3-1 ถึง 3-6



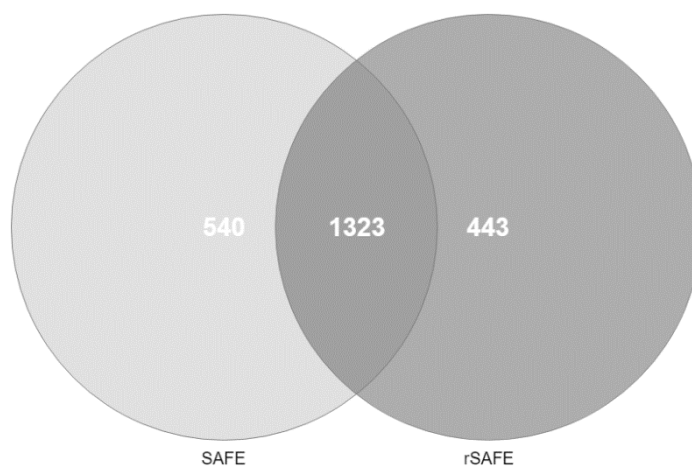
ภาพที่ 3-1 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Central Nervours System



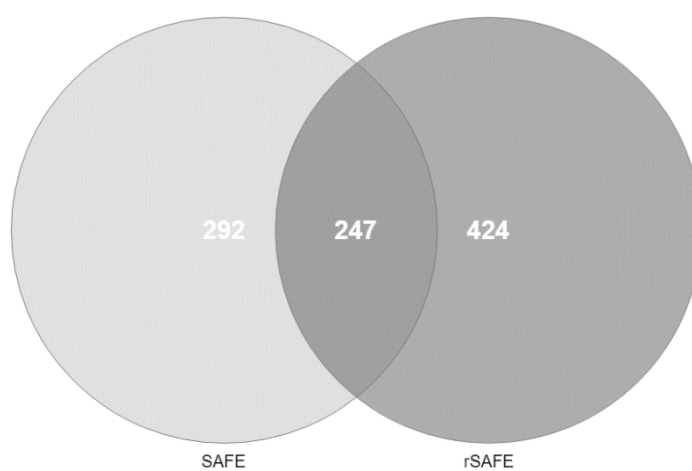
ภาพที่ 3-2 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Brain Cancer



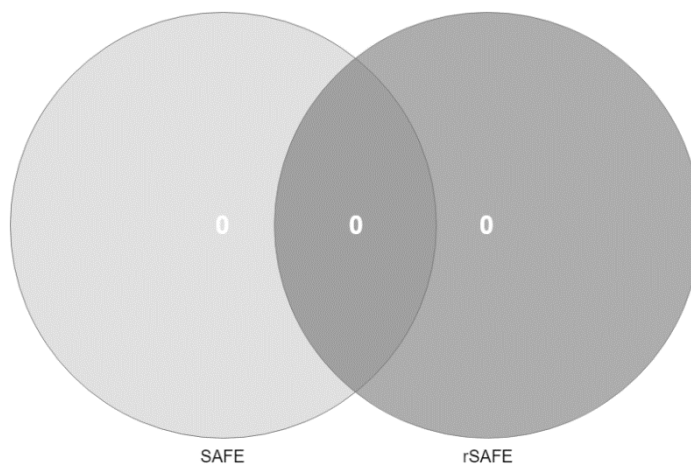
ภาพที่ 3-3 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Lung Carcinoma



ภาพที่ 3-4 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Leukemia



ภาพที่ 3-5 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer



ภาพที่ 3-6 แผนภาพเวนน์แสดงจำนวนหมวดหมู่ของยีนที่สำคัญจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer (Capsular penetration)

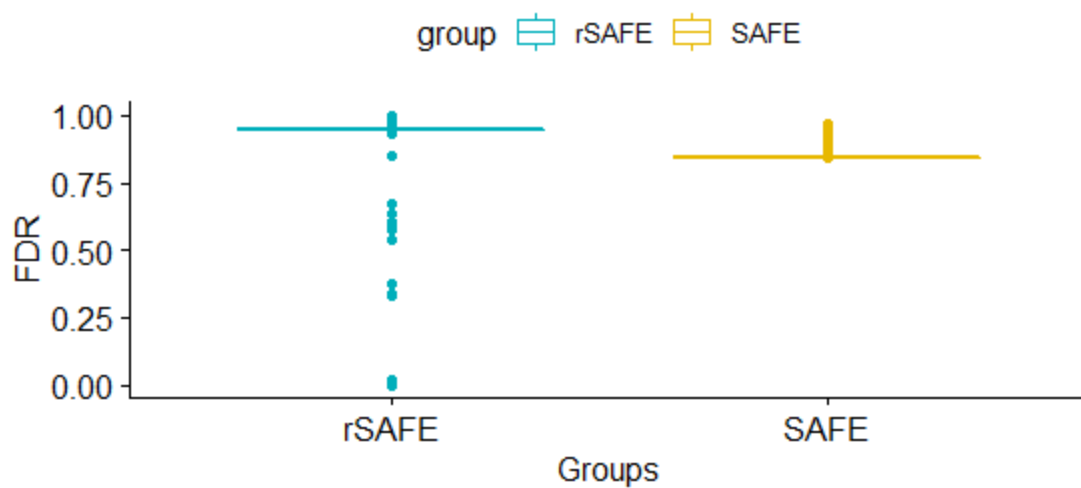
3.3 ผลลัพธ์จำนวนหมวดหมู่ของยีนที่สำคัญของแต่ละชุดข้อมูล

จากการเปรียบเทียบจำนวนหมวดหมู่ของยีนที่สำคัญจากการทำซ้ำในหัวข้อที่ 3.2 ทำให้สามารถเลือกการทดลองที่เป็นตัวแทนของการทดลองทั้ง 100 ครั้งได้ โดยเลือกจากหมวดหมู่ของยีนที่สำคัญของตัวแทนการทดลองที่ได้ค่าใกล้เคียงกับหมวดหมู่ของยีนที่สำคัญที่ให้ผลตรงกัน 80 ครั้งขึ้นไป

ตารางที่ 3-5 ผลลัพธ์ที่ได้จากการทำ FDR ของ Central Nervous System Embryonal Tumor จาก การทดลองที่เป็นตัวแทน

Algorithm	Total Number of Gene Categories	
	Total	Freq. ≥ 80
SAFE	0	0
rSAFE	13	13

ข้อมูลจากตารางที่ 3-5 สามารถนำมาเปรียบเทียบเทียบค่า FDR ในรูปแบบแผนภาพกล่องได้ตามภาพที่ 3-7

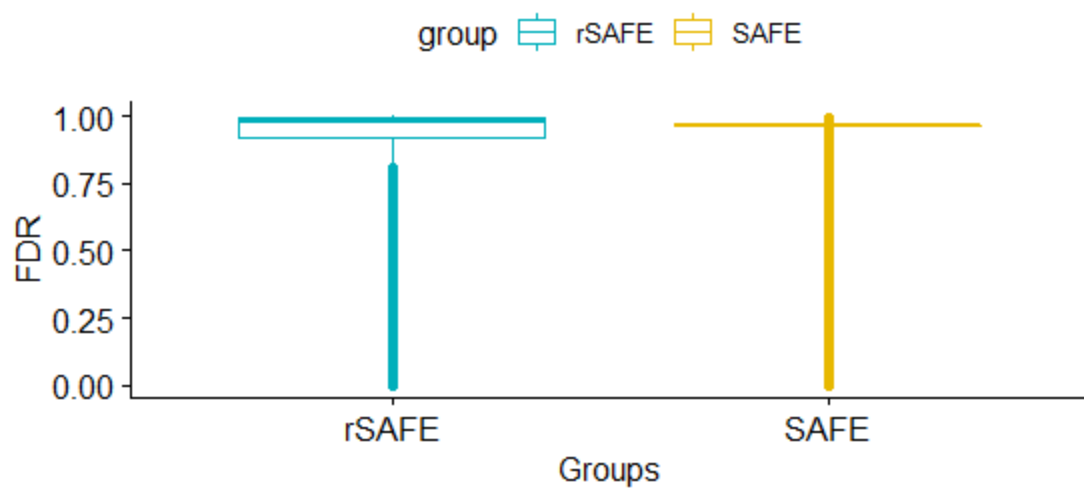


ภาพที่ 3-7 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Central Nervous System Embryonal Tumor

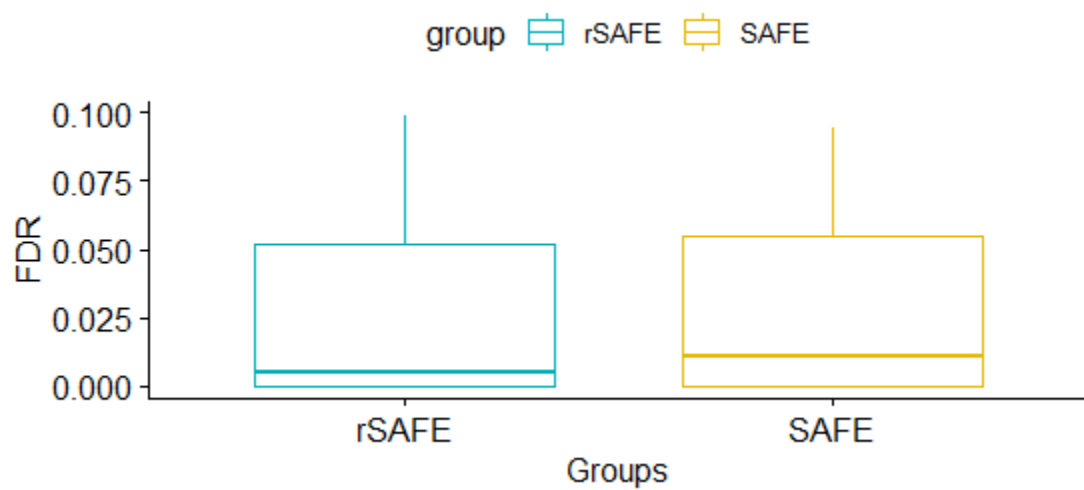
ตารางที่ 3-6 ผลลัพธ์ที่ได้จากการทำ FDR ของ Brain Cancer จากการทดลองที่เป็นตัวแทน

Algorithm	Total Number of Gene Categories	
	Total	Freq. ≥ 80
SAFE	125	110
rSAFE	192	168

ข้อมูลจากตารางที่ 3-6 สามารถนำมาเปรียบเทียบเทียบค่า FDR ในรูปแบบแผนภาพกล่องได้ตามภาพที่ 3-8 และ 3-9



ภาพที่ 3-8 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Brain Cancer

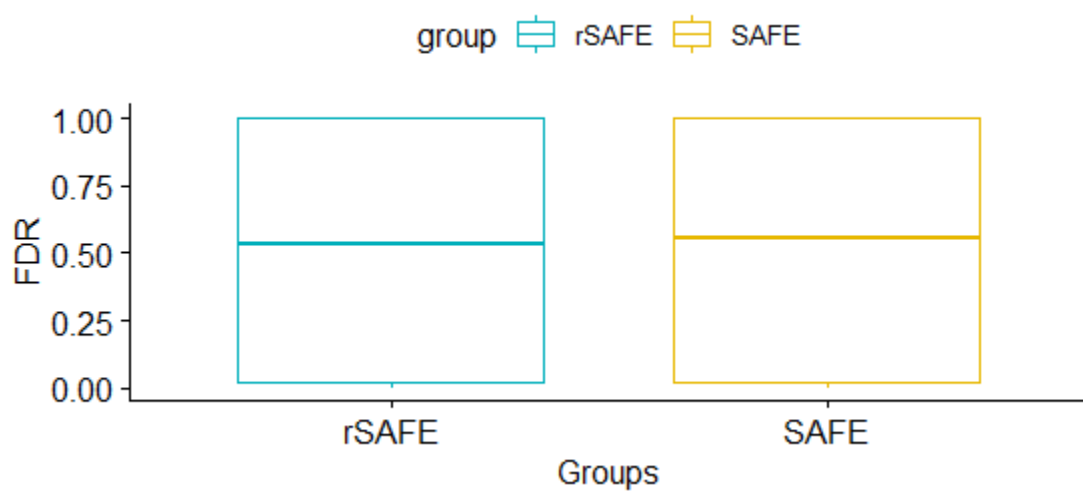


ภาพที่ 3-9 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Brain Cancer

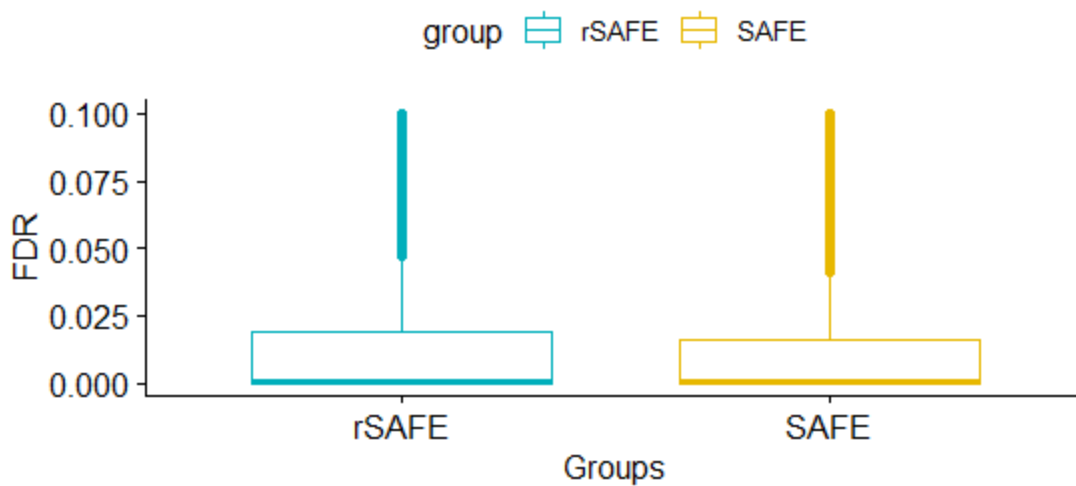
ตารางที่ 3-7 ผลลัพธ์ที่ได้จากการทำ FDR ของ Lung Carcinoma จากการทดลองที่เป็นตัวแทน

Algorithm	Total Number of Gene Categories	
	Total	Freq. ≥ 80
SAFE	3,714	3,517
rSAFE	3,753	3,484

ข้อมูลจากตารางที่ 3-7 สามารถนำมาเปรียบเทียบเทียบค่า FDR ในรูปแบบแผนภาพกล่องตามภาพที่ 3-10 และ 3-11



ภาพที่ 3-10 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Lung Carcinoma



ภาพที่ 3-11 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Lung Carcinoma

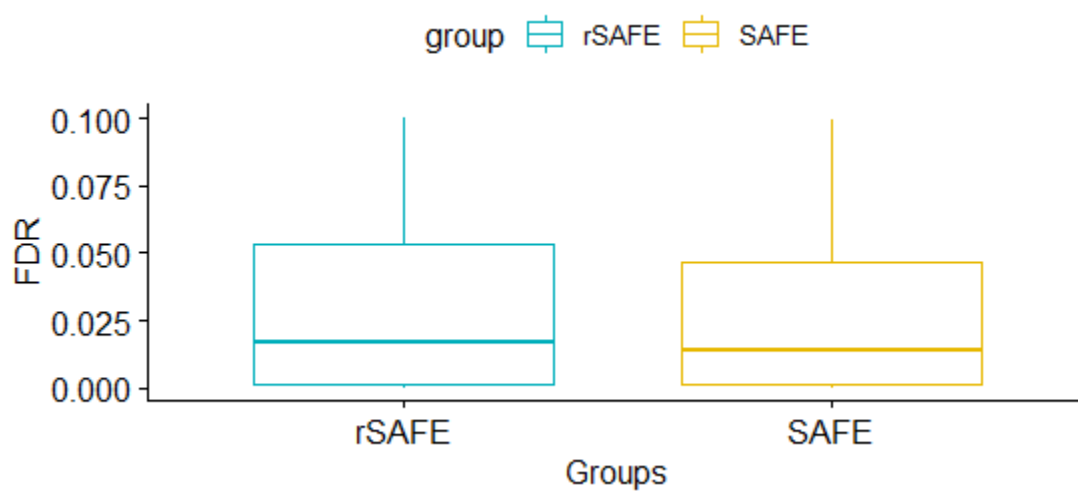
ตารางที่ 3-8 ผลลัพธ์ที่ได้จากการทำ FDR ของ Leukemia จากการทดลองที่เป็นตัวแทน

Algorithm	Total Number of Gene Categories	
	Total	Freq. ≥ 80
SAFE	2,044	1,860
rSAFE	1,967	1,762

ข้อมูลจากตารางที่ 3-8 สามารถนำมาเปรียบเทียบเทียบค่า FDR ในรูปแบบแผนภาพกล่องตามภาพที่ 3-12 และ 3-13



ภาพที่ 3-12 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Leukemia



ภาพที่ 3-13 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Leukemia

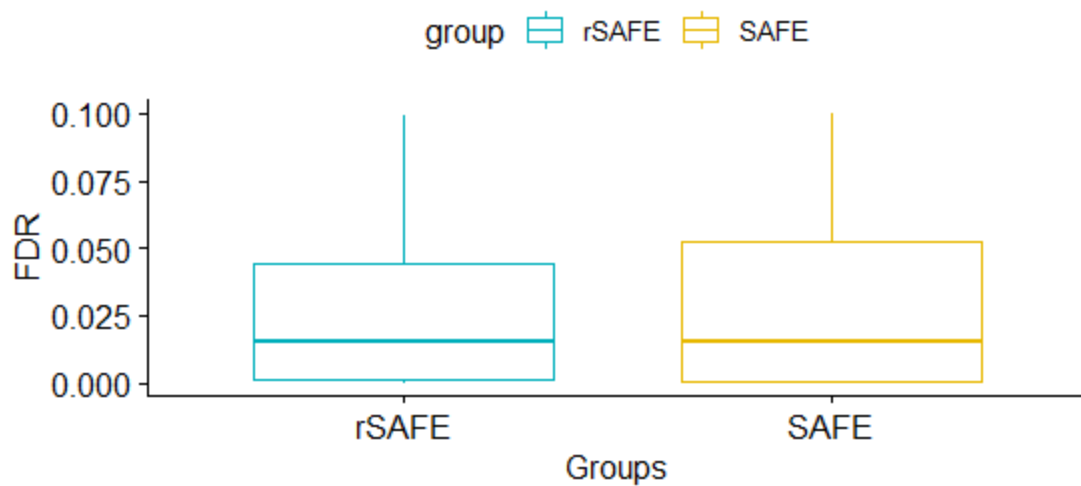
ตารางที่ 3-9 ผลลัพธ์ที่ได้จากการทำ FDR ของ Prostate Cancer จากการทดลองที่เป็นตัวแทน

Algorithm	Total Number of Gene Categories	
	Total	Freq. ≥ 80
SAFE	675	539
rSAFE	764	670

ข้อมูลจากตารางที่ 3-9 สามารถนำมาเปรียบเทียบเทียบค่า FDR ในรูปแบบแผนภาพกล่องตามภาพที่ 3-14 และ 3-15



ภาพที่ 3-14 แผนภาพกล่องแสดงค่า FDR ทั้งหมดจากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer



ภาพที่ 3-15 แผนภาพกล่องแสดงค่า FDR ≤ 0.1 จากเทคนิค SAFE และ rSAFE ของชุดข้อมูล Prostate Cancer

ตารางที่ 3-10 ผลลัพธ์ที่ได้จากการทำ FDR ของ Prostate Cancer (Capsular penetration) จากการทดลองที่เป็นตัวแทน

Algorithm	Total Number of Gene Categories	
	Total	Freq. ≥ 80
SAFE	0	0
rSAFE	0	0

ข้อมูลจากตารางที่ 3-10 สามารถนำมาเปรียบเทียบเทียบค่า FDR ในรูปแบบแผนภาพกล่องตามภาพที่ 3-16

3.4 เวลาที่ใช้ในแต่ละการทดลอง

ทำการทดลองโดยใช้ Computer Engine บน Google Compute Engine โดยมีรายละเอียดดังนี้

- Intel Xeon 2.3 GHz (family 6 model 63 stepping 0) processor 8 vCPUs
- RAM 32 GB memory
- OS Ubuntu 18.10 (Cosmic Cuttlefish)
- R version 3.5.1 programming language
- SAFE version 3.22.0

เวลาเฉลี่ยที่ใช้ในการทำเทคนิค SAFE, rSAFE, eSAFE, ISAFE ของแต่ละชุดข้อมูลทั้ง 100 ครั้ง เป็นไปตามตารางที่ 3-12

ตารางที่ 3-12 เวลาเฉลี่ยในการทำแต่ละเทคนิค แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือ ค่าเบี่ยงเบนมาตรฐาน

Dataset	SAFE	rSAFE	eSAFE	ISAFE
Central Nervous System Embryonal Tumor	8.71 (0.37)	1,283.62 (77.57)	254.02 (36.73)	163.35 (23.22)
Brain Cancer	7.71 (0.27)	1,204.98 (184.41)	204.86 (27.06)	147.24 (6.51)
Lung Carcinoma	9.05 (0.32)	4,385.15 (1,177.74)	528.73 (44.18)	292.73 (19.02)
Leukemia	7.48 (1.13)	1,262.93 (170.23)	185.86 (17.28)	114.60 (1.50)
Prostate Cancer	11.90 (0.66)	2,384.40 (598.52)	370.73 (50.87)	266.50 (39.45)
Prostate Cancer (Capsular penetration)	7.06 (0.38)	1,007.44 (129.54)	191.16 (21.98)	127.63 (17.20)

เวลามากที่สุดที่ใช้ในการทำเทคนิค SAFE, rSAFE, eSAFE, ISAFE ของแต่ละชุดข้อมูลทั้ง 100 ครั้ง เป็นไปตามตารางที่ 3-13

ตารางที่ 3-13 เวลามากสุดในการทำแต่ละเทคนิค แต่ละรอบในหน่วยวินาที

Dataset	SAFE	rSAFE	eSAFE	ISAFE
Central Nervous System Embryonal Tumor	10.04	1,705.90	289.66	183.06
Brain Cancer	8.51	1,429.89	227.47	156.98
Lung Carcinoma	9.98	7,720.03	560.97	357.03
Leukemia	9.65	1,674.01	201.48	118.12
Prostate Cancer	14.01	3,480.77	420.04	315.01
Prostate Cancer (Capsular penetration)	8.07	1,193.10	207.21	155.71

เวลาน้อยที่สุดที่ใช้ในการทำเทคนิค SAFE, rSAFE, eSAFE, ISAFE ของแต่ละชุดข้อมูลทั้ง 100 ครั้ง เป็นไปตามตารางที่ 3-14

ตารางที่ 3-14 เวลาน้อยสุดในการทำแต่ละเทคนิค แต่ละรอบในหน่วยวินาที

Dataset	SAFE	rSAFE	eSAFE	ISAFE
Central Nervous System Embryonal Tumor	7.91	1,226.29	185.02	125.84
Brain Cancer	7.25	916.69	149.72	105.30
Lung Carcinoma	8.54	3,466.61	414.18	275.77
Leukemia	6.26	1,125.78	134.61	111.21
Prostate Cancer	10.81	1,857.41	287.73	223.50
Prostate Cancer (Capsular penetration)	6.32	781.60	141.47	94.65

จากตารางที่ 3-12 ถึง 3-14 หากเราสนใจเวลาที่ใช้ในแต่ละเทคนิคจะได้ว่า rSAFE ใช้เวลาในการดำเนินการมากที่สุด รองลงมาคือ eSAFE, ISAFE และ SAFE ตามลำดับ เหตุผลที่เทคนิคต่างๆ ใช้เวลานานกว่า SAFE ก็เพราะว่าทั้งสามเทคนิคนั้นมีการสร้างแบบจำลองในการจำแนกแทนการใช้ t-test และที่ rSAFE ใช้เวลานานกว่า eSAFE และ ISAFE ก็เพราะว่า rSAFE นั้นไม่มีสัมประสิทธิ์ของตัวแปรใดที่ถูกลดค่าจนเหลือศูนย์ เมื่อมีสร้างแบบจำลองในการจำแนกจึงทำให้ต้องทำการปรับปรุงค่าสัมประสิทธิ์ทุกตัวแปร ในทางตรงกันข้าม ISAFE ทำการลดค่าสัมประสิทธิ์หลายตัว

แปรจนเหลือศูนย์ ทำให้การปรับปรุงค่าตัวแปรนั้นทำเพียงไม่กี่ตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับ ศูนย์เมื่อมีการสร้างแบบจำลองในการจำแนกเป็นเหตุผลให้ ISAFE ใช้เวลาน้อยกว่า rSAFE ส่วน eSAFE นั้นเป็นการรวมวิธีทั้งสองเข้าด้วยกันจึงทำให้เวลาที่ใช้อยู่กึ่งกลางระหว่างวิธีทั้งสอง

โดยเมื่อเราสนใจเวลาที่ใช้ในแต่ละชุดข้อมูลจะได้ว่า แต่ละชุดข้อมูลนั้นใช้เวลาใกล้เคียงกัน มีเพียงสองชุดข้อมูลที่ใช้เวลานานกว่าชุดข้อมูลอื่นอย่างเห็นได้ชัด นั่นก็คือชุดข้อมูลมะเร็งปอดและชุดข้อมูลมะเร็งต่อมลูกหมาก เหตุผลเพราะจำนวนตัวอย่างของยีนที่เหลือหลังจากทำการประมวลผลก่อนของทั้งสองชุดข้อมูลนั้นเยอะกว่าชุดข้อมูลอื่น ๆ อ้างอิงจากตารางที่ 2-4 เพราะการสร้างแบบจำลองในการจำแนกนั้นขึ้นอยู่กับจำนวนของตัวอย่างของยีนที่เหลือหลังทำการประมวลผลก่อน

บทที่ 4

สรุปผลการวิจัย

ปริญญานิพนธ์นี้นำเสนอผลของค่าสถิติเฉพาะที่อิงการจำแนกต่อการวิเคราะห์หับรณัทศน์เมื่อทำการใช้สัมประสิทธิ์ของตัวจำแนกจากการคัดเลือกขึ้นเป็นค่าสถิติเฉพาะเทียบกับค่าสถิติที่อธิบายความแตกต่างของการแสดงออกของยีน

ปริญญานิพนธ์นี้ใช้วิธีเทคนิคในการคัดเลือกลักษณะประจำคือ แลชโซ, ข่ายยึดหยุ่น และการถดถอยริดจ์ ซึ่งผลลัพธ์ที่พบคือค่าสัมประสิทธิ์กลายเป็นศูนย์เพิ่มขึ้นหรือจำนวนของยีนจะลดลงจากเดิมในแต่ละวิธี โดยที่จำนวนยีนที่พบเรียงจากน้อยไปมากคือ แลชโซ, ข่ายยึดหยุ่น และการถดถอยริดจ์ ตามลำดับ นอกจากนี้จำนวนหมวดหมู่ของยีนที่สำคัญที่พบเรียงจากน้อยไปมาก คือ ISAFE, eSAFE, rSAFE ตามลำดับ พบว่าปัจจัยสำคัญที่ทำให้จำนวนหมวดหมู่ของยีนที่สำคัญที่พบมีความแตกต่างกันคือค่าสัมประสิทธิ์ที่กลายเป็นศูนย์

จากการทดลองทั้งหมด 100 ครั้งของเทคนิค SAFE, rSAFE, eSAFE, ISAFE พบว่าเมื่อเทียบจำนวนหมวดหมู่ของยีนที่สำคัญที่พบในเทคนิค SAFE และ rSAFE มีผลลัพธ์ที่ใกล้เคียงกัน แต่มีชุดข้อมูล Central Nervous System Embryonal Tumor ที่เทคนิค SAFE ไม่พบหมวดหมู่ของยีนที่สำคัญ แต่เทคนิค rSAFE พบหมวดหมู่ของยีนที่สำคัญ ส่วนค่า FDR ของหมวดหมู่ยีนที่สำคัญที่ได้จากเทคนิค SAFE และ rSAFE ไม่มีความแตกต่างกัน แต่ค่า FDR ของหมวดหมู่ยีนที่พบทั้งหมดมีความแตกต่างกัน ซึ่งทำให้ผลลัพธ์หมวดหมู่ของยีนที่สำคัญที่พบจากทั้งเทคนิค SAFE และ rSAFE นี้มีทั้งส่วนที่เหมือนกันและส่วนที่แตกต่างกัน

เวลาที่ใช้ในการทดลองของแต่ละเทคนิคนั้นใช้เวลานานกว่าวิธีของ SAFE เนื่องจากในแต่ละเทคนิคนั้นมีการสร้างแบบจำลองการจำแนกแทนการใช้ t-test ส่วนเวลาที่ใช้ในแต่ละเทคนิคก็มีความแตกต่างกัน โดยที่ rSAFE ใช้เวลานานที่สุด รองลงมาคือ eSAFE และใช้น้อยที่สุดคือ ISAFE เป็นผลมาจากความแตกต่างกันของค่าสัมประสิทธิ์ที่กลายเป็นศูนย์

เอกสารอ้างอิง

1. Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2), e1002375.
2. Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, 23(8), 980–987.
3. Nguyen, T., Mitrea, C., & Draghici, S. (2018). Network-based approaches for pathway level analysis. *Current Protocols in Bioinformatics*, 61(1), 8.25.1–8.25.24.
4. Glazko, G. V., & Emmert-Streib, F. (2009). Unite and conquer: Univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, 25(18), 2348–2354.
5. Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1), 191–201.
6. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
7. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.
8. Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, 21(9), 1943–1949.
9. Jansen, R., Penninx, B. W. J. H., Madar, V., Xia, K., Milaneschi, Y., Hottenga, J. J., Hammerschlag, A. R., Beekman, A., van der Wee, N., Smit, J. H., Brooks, A. I., Tischfield, J., Posthuma, D., Schoevers, R., van Grootheest, G., Willemsen, G., de Geus, E. J., Boomsma, D. I., Wright, F. A., Zou, F., Sun, W., & Sullivan, P. F. (2016). Gene expression in major depressive disorder. *Molecular Psychiatry*, 21(3), 339–347.

เอกสารอ้างอิง (ต่อ)

10. Chen, J., Crutchley, J., Zhang, D., Owzar, K., & Kastan, M. B. (2017). Identification of a DNA damage-induced alternative splicing pathway that regulates p53 and cellular senescence markers. *Cancer Discovery*, 7(7), 766–781.
11. Polineni, D., Dang, H., Gallins, P. J., Jones, L. C., Pace, R. G., Stonebraker, J. R., Commander, L. A., Krenicky, J. E., Zhou, Y. H., Corvol, H., Cutting, G. R., Drumm, M. L., Strug, L. J., Boyle, M. P., Durie, P. R., Chmiel, J. F., Zou, F., Wright, F. A., O’Neal, W. K., & Knowles, M. R. (2018). Airway mucosal host defense is key to genomic regulation of cystic fibrosis lung disease severity. *American Journal of Respiratory and Critical Care Medicine*, 197(1), 79–93.
12. Pounds, S. B., Cheng, C., & Onar, A. (2007). Statistical inference for microarray studies. In D. J. Balding, M. Bishop, & C. Cannings (Eds.), *Handbook of statistical genetics* (3rd ed., pp. 231–266). Chichester, UK: John Wiley & Sons.
13. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531–537.
14. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., & Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24), 13790–13795.

เอกสารอ้างอิง (ต่อ)

15. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., & Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870), 436–442.
16. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D’Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2), 203–209.
17. Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R., & Louis, D. N. (2003). *Cancer Research*, 63(7), 1602–1607.
18. Barry, W. T., Nobel, A. B., & Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics*, 2(1), 286–315.
19. Westfall, P. H., & Young, S. S. (1989). P value adjustment for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84(407), 780–786.
20. Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1–2), 171–196.
21. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

เอกสารอ้างอิง (ต่อ)

22. Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), 9440–9445.

ประวัติผู้แต่ง

ปรินญาณินพนธ์เรื่อง : ผลของค่าสถิติเฉพาะที่อิงการจำแนกต่อการวิเคราะห์บรรณนิทัศน์
 สาขาวิชา : วิศวกรรมคอมพิวเตอร์
 ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์
 คณะ : วิศวกรรมศาสตร์
 ชื่อ : นายเจษฎา วีระเดชกำพล

ประวัติ

เกิดเมื่อวันที่ 21 กุมภาพันธ์ พ.ศ. 2540 อยู่บ้านเลขที่ 503 ถนนเพชรเกษม ตำบลห้วยจรเข้ม
 อำเภอเมือง จังหวัดนครปฐม สำเร็จการศึกษาระดับมัธยมศึกษาตอนปลาย จากโรงเรียนสิรินธรราช
 วิทยาลัย จังหวัดนครปฐม สาขาวิทยาศาสตร์-คณิตศาสตร์ ปีการศึกษา 2557 และสำเร็จการศึกษาใน
 ระดับปริญญาตรี สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์
 คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2561

ชื่อ : นายชิตี รุ่งเรือง

ประวัติ

เกิดเมื่อวันที่ 3 กุมภาพันธ์ พ.ศ. 2539 อยู่บ้านเลขที่ 190 หมู่บ้านชัยภัทรวิลเลจ ซอยลาดพร้าว
 วังหิน 83 ถนนลาดพร้าววังหิน แขวงลาดพร้าว เขตลาดพร้าว จังหวัดกรุงเทพมหานคร สำเร็จ
 การศึกษาระดับมัธยมศึกษาตอนปลาย จากโรงเรียนปิยะมหาราชาลัย จังหวัดนครพนม
 สาขาวิทยาศาสตร์-คณิตศาสตร์ ปีการศึกษา 2557 และสำเร็จการศึกษาในระดับปริญญาตรี สาขาวิชา
 วิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัย
 เทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2561