รายงานวิจัยฉบับสมบูรณ์

โครงการ

การพัฒนาการจำแนกผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์ร่วมกับข้อมูลไมโครอะเรย์

Development of Pathway-Based Multiclass Classification Of Complex Diseases Using Microarray Data

คณะผู้ร่วมทำวิจัย

หัวหน้าโครงการ: รศ.ดร.โจนาธาน ชาน
ผู้ร่วมวิจัย: ผศ.ดร.อัศวิน มีชัย

คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

งานวิจัยนี้ได้รับทุนอุดหนุนงบประมาณแผ่นดิน มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
ปีงบประมาณ 2557

รายงานวิจัยฉบับสมบูรณ์

ประจำปีงบประมาณ 2557

---

1. ชื่อโครงการวิจัย

ภาษาไทย การพัฒนาการจำแนกกลุ่มผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์ร่วม กับข้อมูลไมโครอะเรย์

ภาษาอังกฤษ Development of Pathway-Based Multiclass Classification of Complex Disease Using Microarray Data

2. หน่วยงานหลักที่รับผิดชอบงานวิจัย

ชื่อหน่วยงาน คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
เลขที่ 126 ถนนประชาอุทิศ
แขวงบางมด เขตทุ่งครุ กรุงเทพฯ 10140

3. คณะผู้วิจัย

3.1 หัวหน้าโครงการ : รศ.ดร.โจนาธาน ชาน
คุณวุฒิ : ปริญญาเอก (วิศวกรรมเคมี)
สาขาความเชี่ยวชาญ : Bioinformatics and Data Mining
สถานที่ทำงาน : คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
โทรศัพท์ : 0-2470-9819
โทรสาร : 0-2872-7145
ประสบการณ์ในงานวิจัย : Mathematical Modeling and Simulation, Bioinformatics, Knowledge Engineering, Parallel and Grid Computing

ความรับผิดชอบต่อโครงการที่เสนอ คิดเป็น 90% ของงานทั้งหมด

3.2 ผู้ร่วมงานวิจัย : ผศ.ดร.อัศวิน มีชัย
คุณวุฒิ : ปริญญาเอก (วิศวกรรมเคมี)
สาขาความเชี่ยวชาญ : วิศวกรรมเมตาโบลิซึมและชีววิทยาระบบ

สถานที่ทำงาน	:	ภาควิชาวิศวกรรมเคมี คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
โทรศัพท์	:	0-2470-9234 ต่อ 405
โทรสาร	:	0-2872-9118
ประสบการณ์ในงานวิจัย	:	Metabolic engineering and Systems biology

ความรับผิดชอบต่อโครงการที่เสนอ คิดเป็น 10% ของงานทั้งหมด

บทคัดย่อ

การศึกษาโรคทางพันธุกรรมที่มีความซับซ้อน เป็นงานวิจัยที่ท้าทายนักวิทยาศาสตร์เป็นอย่างมากเนื่องจากกลไกการเกิดโรคทางพันธุกรรมประเภทนี้ยังยากที่จะอธิบาย ทั้งนี้เป็นเพราะความซับซ้อนทางพันธุกรรมดังกล่าวเกิดขึ้นจากการมีการเปลี่ยนแปลงของพันธุกรรมในหลายส่วนร่วมกันแล้วจึงก่อให้เกิดโรคขึ้น ดังนั้นการ ศึกษาโรคทางพันธุกรรมแบบบูรณาการได้ถูกพัฒนาขึ้นเพื่อช่วยวิเคราะห์ข้อมูลทางพันธุกรรมในหลายส่วนพร้อมกันซึ่งจะช่วยให้การทำความเข้าใจโรคที่ซับซ้อนนั้นมีความเป็นไปได้มากยิ่งขึ้น งานวิจัยนี้มุ่งเน้นที่จะพัฒนาระเบียบวิธีการวิเคราะห์ข้อมูลการแสดงออกของยีนร่วมกับข้อมูลชุดยีนซึ่งระเบียบวิธีการที่มีอยู่ใน ขณะนั้นถูกจำกัดให้ใช้ได้กับข้อมูลสองกลุ่มตัวอย่างเท่านั้น โดยระเบียบวิธีการที่ได้พัฒนาและนำเสนอในงานวิจัยชิ้นนี้สามารถนำไปใช้กับชุดข้อมูลที่มีมากกว่าสองกลุ่มตัวอย่างได้ โดยระเบียบวิธีการที่พัฒนาขึ้นนั้น ได้ถูกทดสอบด้วยชุดข้อมูลจริง และแสดงให้เห็นว่ามีความเสถียรและความถูกต้องที่จะนำไปใช้ในการวินิจฉัยโรคได้ ทั้งนี้ ระเบียบวิธีการต่างๆได้ถูกพัฒนาให้อยู่ในรูปแบบของโปรแกรมออนไลน์ที่อนุญาตให้ผู้ที่สนใจใช้เพื่อวิเคราะห์หายีนที่เกี่ยวข้องกับการเกิดโรค หรือใช้ในการวินิจฉัยโรคเบื้องต้นได้

คำสำคัญ: โรคที่มีความซับซ้อน / การจำแนกโรคหลากหลายกลุ่ม / ข้อมูลพาธเวย์ / การวิเคราะห์ความแปรปรวนของกลุ่มข้อมูล / ไมโครอะเรย์เทคโนโลยี

Abstract

The study of complex diseases is one of the most challenging areas as the mechanisms driving the diseases are still unclear. The complexity of those diseases is generally agreed to be from the genetic alterations in multiple biological levels; so an integrative analysis was proposed in this work. By integrating multilayer of biological data, this would improve the understanding of the complex diseases, as the information obtained from one layer may provide the missing information that cannot be acquired from another layer. This dissertation aims to improve the state-of-art gene-set-based analyses that is currently limited to binary class and extend them to multiclass problems. In this work, the integration of pathway data in microarray data analysis for multiclass classification has been proposed. By evaluating the proposed method using actual cancer data, they have been shown to be accurate and robust in disease diagnosis. Finally, we also implemented the proposed method as an online tool that is expected to be able to facilitate the search for candidate disease-related genes and to be used to diagnose the patients as an early screening tool.

Keywords: Complex Disease / Multiclass Classification / Pathway-based / Analysis of Variance / Microarray Technology

สารบัญเรื่อง

สารบัญตาราง

สารบัญรูป

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของงานวิจัย

เทคโนโลยีดีเอ็นเอไมโครอะเรย์ได้ถูกนำมาใช้เพื่อวัดระดับการแสดงออกของยีนจำนวนมากพร้อมๆกัน Messenger RNA(mRNA) จะถูกสกัดออกมาจากเซลล์ที่สนใจและถูกนำไปไฮบริไดซ์ลงบนไมโครอะเรย์ชิพที่มีโพรพ (probe) จำนวนมากเพื่อวัดระดับการแสดงออกโดยใช้ระเบียบวิธีการทางด้าน Image processing (Lipshutz, 1995) ข้อมูลไมโครอะเรย์ได้ถูกใช้อย่างแพร่หลายในงานวิจัยทางด้านการแพทย์ เพื่อช่วยทำความ เข้าใจกลไกการเกิดโรค และเพื่อหาวิธีการป้องกันและรักษาต่อไป (Adi, 2006; Orihuela, 2004). จากการที่มีการใช้เทคโนโลยีนี้อย่างแพร่หลายทำให้เกิดการรวบรวมข้อมูลไมโครอะเรย์ในรูปแบบฐานข้อมูลสาธารณะ อาทิเช่น Gene Expression Omnibus (GEO), Gene Expression Databased (GXD), ArrayExpress ฯลฯ เพื่อที่จะใช้ข้อมูลเหล่านั้นซึ่งมีจำนวนมากให้เกิดประสิทธิผลสูงสุดจึงจำเป็นต้องมีการใช้ศาสตร์ใหม่ซึ่งผสมผสามความรู้ทางด้านการจัดการสารสนเทศและความรู้เชิงชีววิทยา หรือที่รู้จักกันในชื่อ ชีวสารสนเทศ

การวิเคราะห์ข้อมูลไมโครอะเรย์ทั่วๆไปมักจะมีการทำการวิเคราะห์ยีน (gene) ทีละยีน ซึ่งเป็นวิธีการวิเคราะห์ที่ให้ผลที่ดีสำหรับโรคที่มีความซับซ้อนไม่มาก แต่สำหรับโรคที่มีความซับซ้อนสูงโดยมีกลไกหลายอย่างเกิดขึ้นร่วมกันเพื่อก่อให้เกิดโรคนั้น จึงจำเป็นต้องวิเคราะห์ข้อมูลการแสดงออกของยีนร่วมกับข้อมูลทางด้านชีวภาพอื่นๆด้วย อาทิเช่น ข้อมูลกลุ่มยีน, ข้อมูลความสัมพันธ์กันของโปรยีน, ฯลฯ การวิเคราะห์ ข้อมูลไมโครอะเรย์ร่วมกับข้อมูลพาธเวย์เป็นหนึ่งแนวทางที่ถูกพัฒนาอย่างแพร่หลาย (Hosgood, 2008; Stelios, 2011; Wang, 2011) ข้อมูลพาธเวย์เป็นข้อมูลที่บอกถึงการทำงานร่วมกันของยีนต่างๆ ซึ่งก่อให้เกิดเป็นฟังกชั่นการทำงานต่างๆภายในเซลล์ หลายกลุ่มวิจัยได้พยายามพัฒนาระเบียบวิธีการปรับเปลี่ยนข้อมูลไมโครอะเรย์ให้อยู่ในรูปแบบของระดับการทำงานของพาธเวย์ ระเบียบวิธีการ Negatively Correlated Feature Set (NCFS-i) ถูกพัฒนาขึ้นโดย Sootanan et al., 2012 ซึ่งมีการใช้วิธีการค้นหาแบบ Greedy Algorithm และมีการคำนวนระดับการทำงานของพาธเวย์โดย อ้างอิงตามระดับการแสดงออกของกลุ่มยีนเด่นๆในแต่ละพาธเวย์ ผลการทดลองของงานวิจัยดังกล่าวแสดงให้เห็นว่า แบบจำลองที่มีการใช้ข้อมูลการทำงานของพาธเวย์ที่ได้จาก NCFS-i นั้น มีความสามารถในการแยกแยะกลุ่มผู้ป่วยได้ดีกว่าการใช้ข้อมูลไมโครอะเรย์เพียงอย่างเดียว

อย่างไรก็ตามระเบียบวิธีการดังกล่าวถูกออกแบบมาเพื่อใช้กับการศึกษาที่มีกลุ่มตัวอย่างเพียงสองกลุ่มเท่านั้น ดังนั้นในงานที่มีกลุ่มตัวอย่างมากกว่าสองกลุ่มขึ้นไปจะไม่สามารถปรับใช้ระเบียบวิธีการนี้ได้ ซึ่ง

โดยทั่วไปแล้วงานวิจัยทางด้านการแพทย์โดยใช้ข้อมูลไมโครอะเรย์ก็ถูกปรับใช้กับปัญหาที่มีกลุ่มตัวอย่างหลายกลุ่ม อาทิ เช่น การแยกแยะผู้ป่วยโรคมะเร็ง การแบ่งระยะของมะเร็ง ฯลฯ (Sridhar, 2001; Jane, 2004; Lukas, 2004; Yang, 2014). ดังนั้นงานวิจัยชิ้นนี้จึงมีจุดประสงค์เพื่อจะปรับปรุง และพัฒนาระเบียบวิธีการจำแนกกลุ่มผู้ป่วยหลากหลาย ด้วยวิธีการอ้างอิงระดับการทำงานของพาธเวย์ร่วมกับข้อมูลไมโครอะเรย์ เพื่อให้สามารถนำมาใช้ในการวินิจฉัยโรคได้อย่างมีประสิทธิภาพมากขึ้น

1.2 วัตถุประสงค์ของโครงการวิจัย

1. เพื่อพัฒนาระเบียบวิธีการอ้างอิงระดับการทำงานของพาธเวย์เพื่อใช้ในการวิเคราะห์และจำแนกกลุ่มตัวอย่างหลายกลุ่ม
2. เพื่อจะประยุกต์ใช้ระเบียบวิธีการที่พัฒนาขึ้นมากับข้อมูลไมโครอะเรย์จริง

1.3 ขอบเขตโครงงานวิจัย

1. งานวิจัยชิ้นนี้จะปรับปรุงและพัฒนาวิธีการอ้างอิงระดับการทำงานของพาธเวย์เพื่อใช้ข้อมูลการทำงานของพาธเวย์ในการวิเคราะห์และจำแนกกลุ่มตัวอย่างหลายกลุ่ม
2. ความถูกต้อง ความน่าเชื่อถือของระเบียบวิธีการใหม่ที่พัฒนาขึ้นมานั้นจะถูกวัดโดยการนำไปประยุกต์ใช้กับข้อมูลไมโครอะเรย์จริง และเปรียบเทียบผลลัพธ์ที่ได้กับระเบียบวิธีการอื่นๆ ที่เคยถูกพัฒนาขึ้นมาก่อนหน้านี้

1.4 ผลที่คาดว่าจะได้รับ

1. ได้วิธีการจำแนกผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์การเกิดโรคร่วมกับข้อมูลไมโครอาร์เรย์
   <u>กลุ่มเป้าหมาย</u> กลุ่มวิจัยด้านชีวสารสนเทศทางการแพทย์ของประเทศไทย
   ผลกระทบของงานวิจัยต่อกลุ่มเป้าหมาย สามารถนำวิธีการจำแนกผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์การเกิดโรคร่วมกับข้อมูลไมโครอาร์เรย์ไปใช้ประโยชน์ได้
2. การสั่งสมเทคโนโลยี และองค์ความรู้ทางด้านชีวสารสนเทศทางการแพทย์ จะช่วยเพิ่มศักยภาพในการเผยแพร่เทคโนโลยี องค์ความรู้ และบริการต่อสาธารณะ
   <u>กลุ่มเป้าหมาย</u> กลุ่มวิจัยด้านชีวสารสนเทศทางการแพทย์ของประเทศไทย
   ผลกระทบของงานวิจัยต่อกลุ่มเป้าหมาย เพิ่มศักยภาพในการวิจัยทางด้านการวิเคราะห์ข้อมูลไมโครอาร์เรย์ ชีววิทยาระบบและชีวสารสนเทศของประเทศไทย
3. เผยแพร่องค์ความรู้และบทความวิจัยในรูปแบบสิ่งตีพิมพ์และงานประชุมวิชาการในระดับชาติและ/หรือนานาชาติ
   <u>กลุ่มเป้าหมาย</u> กลุ่มวิจัยด้านชีวสารสนเทศทางการแพทย์ระดับชาติและนานาชาติ

ผลกระทบของงานวิจัยต่อกลุ่มเป้าหมาย มีวิธีการจำแนกผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์ร่วมกับข้อมูลไมโครอาร์เรย์

1.5 แผนการถ่ายทอดเทคโนโลยีหรือผลการวิจัยสู่กลุ่มเป้าหมาย

1. แผนการถ่ายทอดเทคโนโลยีและวิธีการจำแนกผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์ร่วมกับข้อมูลไมโครอาร์เรย์สู่กลุ่มวิจัยด้านชีวสารสนเทศทางการแพทย์ของประเทศไทย
   กลุ่มเป้าหมาย กลุ่มวิจัยด้านชีวสารสนเทศทางการแพทย์ของประเทศไทยและนานาชาติ
   วิธีการถ่ายทอด นำเสนอวิธีการจำแนกผู้ป่วยหลายกลุ่มด้วยการวิเคราะห์พาธเวย์ร่วมกับข้อมูลไมโครอาร์เรย์ในการประชุมวิชาการในประเทศไทยและ/หรือนานาชาติ

1.6 แผนการดำเนินงานตลอดโครงการ

ตารางที่ 1.1 แผนการดำเนินงาน

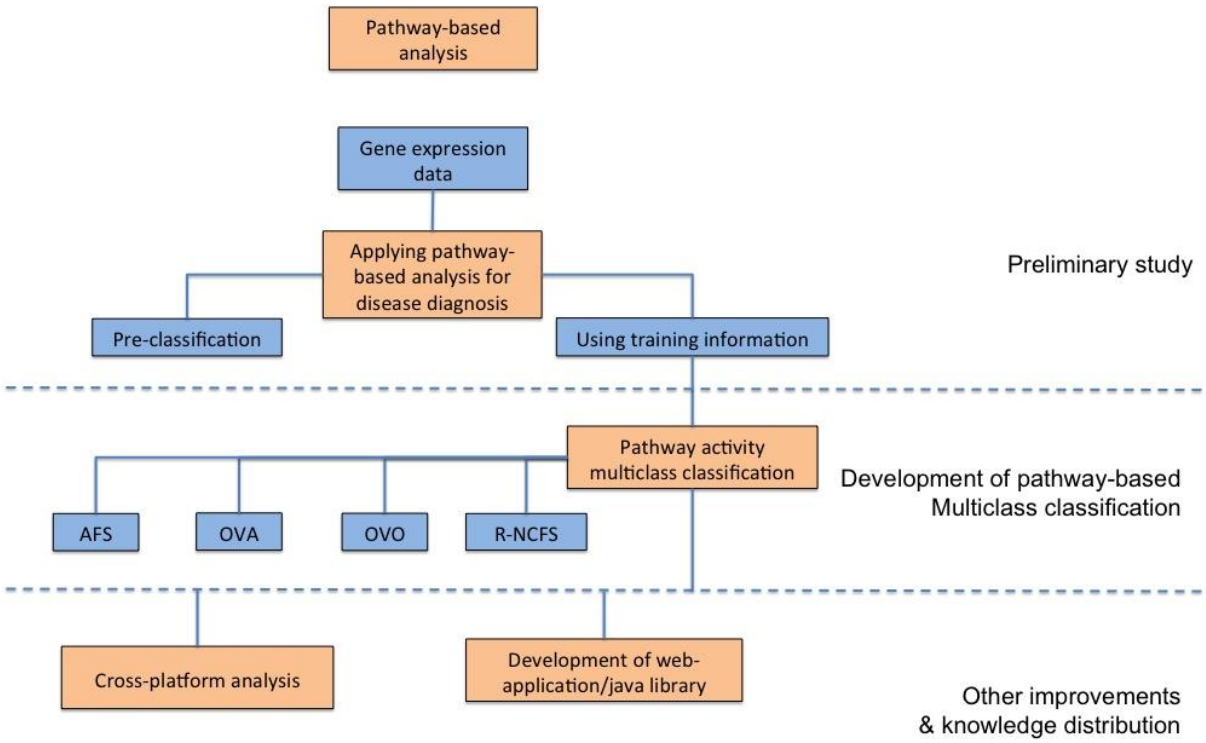
แผนการดำเนินงาน	เดือนที่ 1-3	เดือนที่ 4-6	เดือนที่ 7-9	เดือนที่ 10-12
1. ค้นหาข้อมูลไมโครอะเรย์เพื่อนำมาใช้ในการศึกษาและทดสอบระเบียบวิธีการที่พัฒนาขึ้น	←	→		
2. ทำการจัดระเบียบข้อมูลเบื้องต้นก่อนนำไปใช้ในการศึกษาจริง	←	→		
3. ค้นหาระเบียบวิธีการเพื่อนำมาช่วยในการหาข้อมูลการเป็นโรคและไม่เป็นโรคของกลุ่มตัวอย่างเบื้องต้นก่อนจะนำไปใช้อ้างอิงระดับการทำงานของพาธเวย์		←	→	
4. ทดสอบประสิทธิภาพของการจำแนกกลุ่มผู้ป่วยโดยอาศัยระเบียบวิธีการใหม่ที่มีการหาข้อมูลการเป็นโรคและไม่เป็นโรคเบื้องต้นเทียบกับระเบียบวิธีการเดิม		←	→	
5. ค้นหาระเบียบวิธีการเพื่อเข้ามาช่วยในการปรับปรุงให้การอ้างอิงระดับการทำงานของพาธเวย์แบบ NCFS สามารถปรับใช้กับกลุ่มตัวอย่างหลายกลุ่มได้			←	→
6. ทดสอบประสิทธิภาพของการจำแนกกลุ่มตัวอย่างหลายกลุ่มโดยอาศัยระเบียบวิธีการใหม่ เปรียบเทียบกับระเบียบวิธีการจำแนกกลุ่มตัวอย่างหลายกลุ่มอื่นๆ				← →

บทที่ 2

ระเบียบวิธีวิจัย

2.1 ภาพงานโดยรวม

งานวิจัยชิ้นนี้มุ่งหวังที่จะพัฒนาการวิเคราะห์ข้อมูลไมโครอะเรย์ร่วมกับข้อมูลพาธเวย์เพื่อใช้ในการจำแนกกลุ่มผู้ป่วยหลายกลุ่ม โดยงานวิจัยนี้ได้แบ่งการศึกษาออกเป็นสามส่วนหลัก ดังแสดงในรูปที่ 2.1



รูปที่ 2.1 ภาพงานวิจัยโดยรวม

ส่วนที่หนึ่งเป็นส่วนเบื้องต้นที่จะทำการศึกษาเทคนิคเบื้องต้นในการอ้างอิงระดับการทำงานของพาธเวย์ด้วยวิธีที่ถูกพัฒนาในขณะนั้น โดยจะมีการเปรียบเทียบเพื่อหาวิธีการปรับใช้ระเบียบวิธีดังกล่าวเพื่อไปอ้างอิงระดับการทำงานของพาธเวย์ในข้อมูลทดสอบ ในส่วนถัดมาจะเป็นการพัฒนาระเบียบวิธีการอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่ม โดยได้นำแนวคิดแบบ One-vs-One และ One-vs-All เข้ามาปรับใช้กับระเบียบวิธีการ NCFS-i รวมทั้งยังพัฒนาระเบียบวิธีการใหม่ขึ้นเพื่อเปรียบเทียบกัน ส่วนสุดท้ายจะเป็นการทดสอบความเสถียรของระเบียบวิธีการที่พัฒนาขึ้น โดยปรับใช้กับชุดข้อมูลที่มีปัญหาต่างๆกัน รวมถึงได้พัฒนาระเบียบวิธีการดังกล่าวขึ้นเป็นโปรแกรมที่อนุญาตให้ผู้ใช้สามารถเข้าใช้งานผ่านระบบออนไลน์ได้

2.2 ข้อมูลไมโครอะเรย์และข้อมูลพาธเวย์

ในงานวิจัยชิ้นนี้ได้รวบรวมข้อมูลไมโครเรย์มาจำนวน 12 ชุดจากฐานข้อมูลไมโครอะเรย์ GEO (ตารางที่ 2.1) (Edgar, 2002) โดยแบ่งเป็นข้อมูลที่มีจำนวนกลุ่มตัวอย่างสองกลุ่ม 6 ชุด และเป็นข้อมูลที่มีกลุ่มตัวอย่างมากกว่าสองกลุ่มขึ้นไปอีก 6 ชุด

ตารางที่ 2.1 ชุดข้อมูลไมโครอะเรย์ที่ใช้ในงานวิจัยนี้

รหัสการเข้าถึง	โรค	ชื่ออ้างอิง	กลุ่มตัวอย่าง
GSE5764	Breast cancer	Breast1	Case: 20, Control: 10
GSE7904	Breast cancer	Breast2	Case: 43, Control: 19
GSE1456	Breast metastasis	Breast3	Relapse: 40, non-Relapse: 119
GSE2034	Breast metastasis	Breast4	Relapse: 107, non-Relapse: 179
GSE2109	Lung cancer	MCLung1	AC stage-1: 17, AC stage-2: 6, SCC stage-1: 21, SCC stage-2: 8
GSE10245	Lung cancer	MCLung2	AC stage-1: 22, AC stage-2: 14, SCC stage-1: 9, SCC stage-2: 6
GSE18842-1	Lung cancer	MCLung3	AC stage-1: 12, AC stage-2: 0, SCC stage-1: 27, SCC stage-2: 3
GSE43580	Lung cancer	MCLung4	AC stage-1: 41, AC stage-2: 36, SCC stage-1: 34, SCC stage-2: 39
GSE18842-2	Lung cancer	Lung1	Case: 46, Control: 45
GSE4115	Lung cancer	Lung2	Case: 97, Control: 90
GSE40115-1	Breast cancer	MCBreast1	Sporadic: 104, BRCA1: 23, BRCA2: 20
GSE40115-2	Breast cancer	MCBreast2	Sporadic: 56, BRCA1: 20, BRCA2: 16

ในส่วนของข้อมูลพาธเวย์นั้นได้รวบรวมมาจากฐานข้อมูล Molecular Signature DB (MSigDB) เวอร์ชั่น 4.0 โดยงานวิจัยชิ้นนี้ได้เลือกมาเฉพาะพาธเวย์ที่เกี่ยวข้อกับการเกิดโรคที่ถูกคัดกรองจากผู้เชี่ยวชาญมาใช้เท่านั้น (c2.cp/curated canonical pathways) โดยข้อมูลชุดนี้ประกอบไปด้วยพาธเวย์ 1,320 ชุด

2.3 การวิเคราะห์ข้อมูลไมโครอะเรย์ร่วมกับข้อมูลพาธเวย์

แม้ว่าเทคโนโลยีไมโครอะเรย์กำลังจะถูกแทนที่ด้วยเทคโนโลยีใหม่ๆ อีกทั้งยังมีข้อมูลทางด้านพันธุกรรมหลากหลายชนิดถูกสร้างขึ้นและเผยแพร่ให้ใช้อย่างเสรีแต่ข้อมูลไมโครอะเรย์ยังเป็นกลุ่มข้อมูลที่มีจำนวนชุดข้อมูลมากที่สุดที่อนุญาตให้เข้าใช้ได้ ด้วยจำนวนที่มากของชุดข้อมูลนี้ ทำให้หลายกลุ่มวิจัยยังให้ความสนใจกับข้อมูลประเภทนี้อยู่ ปัจจุบันได้มีงานวิจัยที่มีการนำข้อมูลพาธเวย์มาช่วยวิเคราะห์ข้อมูลไมโครอะเรย์ หลายชิ้นถูกพัฒนาขึ้นเพื่อช่วยให้นักวิจัยเข้าใจกระบวนการการเกิดโรคที่มีความซับซ้อนสูง อย่างเช่น มะเร็ง (Pang, 2008) ความเข้าใจในกระบวนการดังกล่าวจะช่วยให้ นักวิจัยสามารถคิดค้นวิธีการรักษาโรคดังกล่าวได้ ระเบียบวิธีการวิเคราะห์ข้อมูลไมโครอะเรย์ด้วยข้อมูลพาธเวย์นั้นสามารถจำแนกออกเป็นสองกลุ่มหลักๆดังนี้ คือระเบียบวิธีการที่ใช้ข้อมูลพาธเวย์เพื่อช่วยเลือกยีนต้องสงสัย (Bandyopadhyay, 2009) และระเบียบวิธี-การที่เปลี่ยนรูปแบบข้อมูลจากข้อมูลการแสดงออกของยีนไปเป็นข้อมูลการทำงานของพาธเวย์ เช่น ระเบียบ วิธีการ Negatively Correlated Feature Sets (NCFS-i) (Sootanan, 2012).

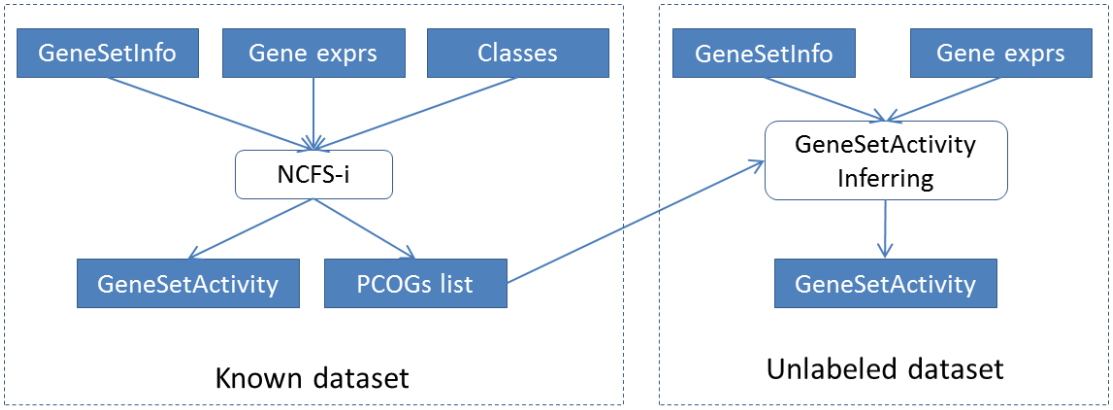
2.3.1 การปรับใช้ระเบียบวิธีการอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลทดสอบ

โดยปกติแล้วการอ้างอิงระดับการทำงานของพาธเวย์ด้วยวิธี NCFS-i นั้นจะต้องการข้อมูลการเป็นโรคของกลุ่มตัวอย่างก่อน ดังนั้นสำหรับงานทางด้านการแพทย์ อาทิ การตรวจวินิจฉัยโรค นั้น จะไม่มีข้อมูลการเป็นโรคของกลุ่มตัวอย่าง ทำให้ไม่สามารถอ้างอิงระดับการทำงานของพาธเวย์ด้วยวิธีนี้ได้ ในการศึกษาเบื้องต้นนี้ ผู้วิจัยได้ทำการเปรียบเทียบวิธีการปรับใช้วิธี NCFS-i กับกลุ่มข้อมูลทดสอบที่ไม่มีข้อมูลการเป็นโรค ทั้งสิ้นสองวิธีการคือ 1) พยายามทำนายข้อมูลการเป็นโรคเบื้องต้นด้วยข้อมูลการแสดงออกของยีนก่อน (Pre-classification: PC) 2) เนื่องจากข้อมูลการเป็นโรคนั้นมีความสำคัญในขั้นตอนการหากลุ่มยีน (Phenotype-Correlated Genes: PCOGs) เด่นในแต่ละพาธเวย์ ดังนั้นหากในชุดทดสอบไม่มีข้อมูล การเป็นโรคก็ไปใช้ข้อมูล PCOGs จากชุดข้อมูลที่ใช้สร้างแบบจำลองเพื่ออ้างอิงระดับการทำงานของพาธเวย์แทน ดังรูปที่ 2.2 โดยปกติ NCFS-i จะต้องการข้อมูลสามชนิดประกอบด้วย ข้อมูลการแสดงออกของยีน ข้อมูลการเป็นโรค และข้อมูลพาธเวย์ โดยผลที่ได้จากวิธีนี้คือ ชุดข้อมูลการทำงานของพาธเวย์ และข้อมูล PCOGs ดังนั้น ข้อมูล PCOGs จะถูกใช้ในการอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลทดสอบ (Test set) โดยปกตินั้นข้อมูล PCOGs จะถูกแบ่งเป็นสองส่วนคือ กลุ่มบวก และกลุ่มลบ โดยระดับการทำงาน ของพาธเวย์จะถูกอ้างอิงจากผลต่างของระดับการแสดงออกของยีนในแต่ละกลุ่มดังสมการด้านล่าง

$$GAC(GS_i) = \frac{\sum_{ip=1}^{np} Z(EXP_{ip})}{\sqrt{np}} - \frac{\sum_{in=1}^{nn} Z(EXP_{in})}{\sqrt{nn}}$$

โดยให้ np เป็นจำนวนของยีนในกลุ่มบวก $Z(EXP)_{ip}$ เป็นระดับการแสดงออกของยีน $gene_{ip}$ ที่ผ่านการทำ Z-transformed มาแล้ว nn เป็นจำนวนของยีนในกลุ่มลบและ $Z(EXP)_{in}$ เป็นระดับการแสดงออกของยีน $gene_{in.}$

ที่ผ่านการทำ Z-transformed มาแล้ว



รูปที่ 2.2 การอ้างอิงระดับการทำงานของพาธเวย์โดยใช้ PCOGs จากชุดข้อมูลที่ใช้สร้างแบบจำลอง

2.3.2 การเปรียบเทียบอัลกอริทึมในการทำนายข้อมูลการเป็นโรคเบื้องต้น

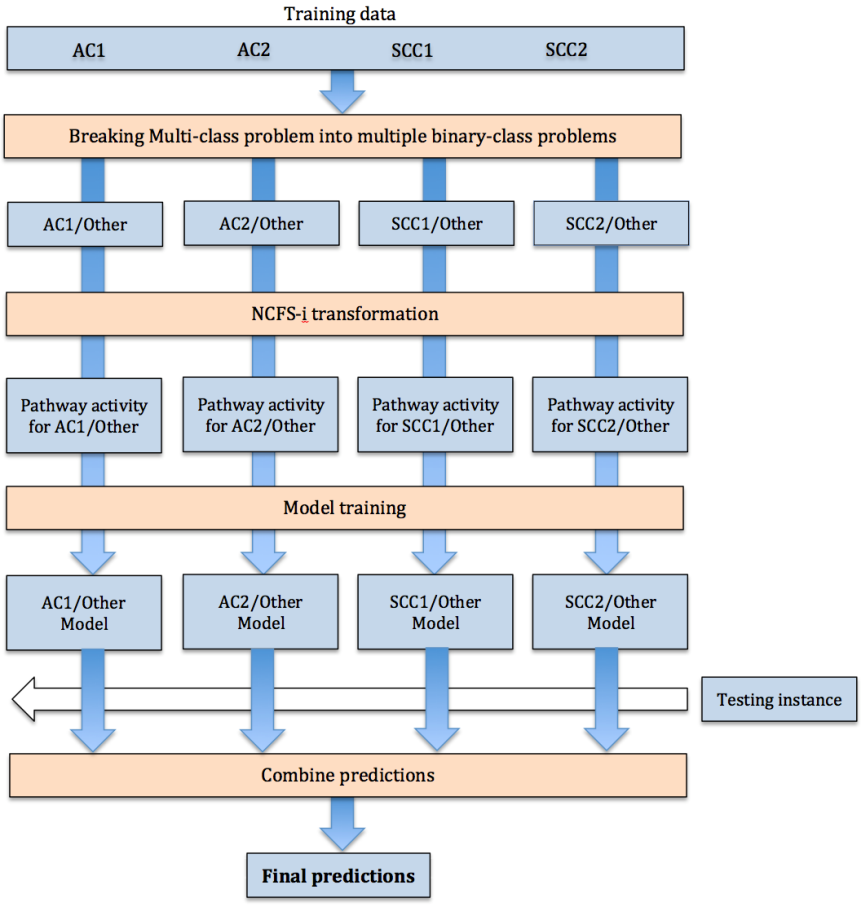
การเปรียบเทียบในการศึกษาเบื้องต้นนี้ได้ทำขึ้นเพื่อหาอัลกอริทึมที่ใช้จำแนกกลุ่มผู้ป่วยเบื้องต้นด้วยข้อมูลไมโครอะเรย์ รวมถึงจำนวนยีนที่จะถูกเลือกใช้มาสร้างแบบจำลองด้วย โดยในงานวิจัยนี้ ได้ทำการเปรียบเทียบอัลกอริทึมพื้นฐานสามอย่างดังต่อไปนี้ K-Nearest Neighbor (KNN), Random Forest (RF), และ Support Vector Machine (SVM) โดยนำไปปรับใช้กับข้อมูล Breast3 และ Breast4 และมีการใช้ Feature Selection ด้วยวิธี SVMAttributeEval จาก library ของ WEKA (Van, 2009) ซึ่งจำนวนของยีนที่ถูกใช้เป็น feature ในการสร้างแบบจำลองนั้นได้มีการทดสอบหลายๆจำนวนดังงานวิจัยก่อนหน้านี้ (Chan, 2011).

2.4 การอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลที่มีหลายกลุ่มตัวอย่าง

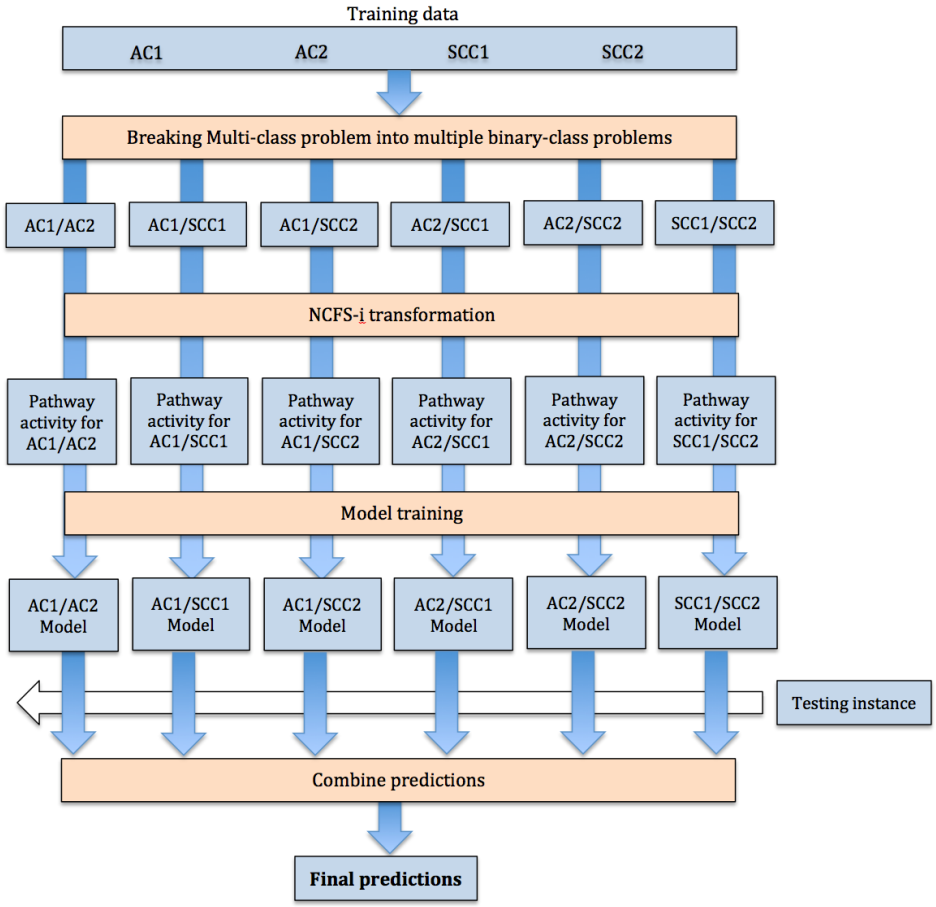
เนื่องจากระเบียบวิธีการในการอ้างอิงระดับการทำงานของพาธเวย์ในขณะนั้นถูกจำกัดอยู่กับแค่การศึกษาที่มีกลุ่มตัวอย่างสองกลุ่ม ดังนั้นงานวิจัยนี้จึงได้ทำการต่อยอดระเบียบวิธีการเดิมเพื่อให้สามารถปรับใช้ได้กับชุดข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่มได้ โดยงานวิจัยนี้จะนำเสนอวิธีการสองวิธีการประกอบไปด้วย 1) การใช้ แนวคิดแบบ One-vs-One (OVO) และ One-vs-All (OVA) ซึ่งเป็นวิธีที่ง่ายที่สุดในการต่อยอดการจำแนกในชุดข้อมูลที่มีกลุ่มตัวอย่างสองกลุ่ม ไปยังชุดข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่มโดยใช้วิธีการแตกปัญหาการจำแนกกลุ่มหลายกลุ่มตัวอย่างเป็นปัญหาการจำแนกกลุ่มสองกลุ่มหลายๆปัญหา 2) จะใช้แนวความคิดของวิธีการ NCFS-i มาช่วยในการพัฒนาระเบียบวิธีการใหม่ซึ่งสามารถปรับใช้ได้กับชุดข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่มได้โดยตรง โดยงานวิจัยนี้ได้พัฒนาระเบียบวิธีการใหม่สองวิธีการ ซึ่งให้ชื่อระเบียบวิธีการว่า Reference-based Negatively Correlated Feature Set (R-NCFS) และ ANOVA-based Feature Set (AFS) โดยรายละเอียดของระเบียบวิธีการทั้งสองจะทำการอธิบายในหัวข้อถัดไป

2.4.1 OVO และ OVA

ในงานการจำแนกกลุ่มตัวอย่าง (Classification) ในชุดข้อมูลที่มีกลุ่มตัวอย่างมากกว่าสองกลุ่มนั้น ได้เป็นปัญหาในกลุ่มนักวิจัยทางด้าน Machine learning มานาน ซึ่งได้มีอัลกริทึมหลากหลายตัวที่สามารถจำแนกกลุ่ม ตัวอย่างมากกว่าสองกลุ่ม โดยอาศัยการต่อยอดจากการจำแนกกลุ่มตัวอย่างสองกลุ่มได้ถูกนำเสนอขึ้น One-versus-One (OVO) และ One-versus-All (OVA) เป็นสองอัลกอริทึมที่แตกปัญหาหลักออกเป็นปัญหาการ จำแนกกลุ่มตัวอย่างสองกลุ่มหลายๆปัญหา โดย OVO จะแตกปัญหาออกเป็นจำนวน $C(n,2)$ ปัญหา โดยที่ n เป็นจำนวนของกลุ่มตัวอย่าง และแต่ละปัญหาจะพยายามแยกกลุ่มตัวอย่างนึงออกจากอีกกลุ่มตัวอย่างนึงเป็นคู่ๆ ในขณะที่ OVA จะแตกปัญหาออกเป็น n ปัญหา และแต่ละปัญหาจะพยายามแยกกลุ่มตัวอย่างนึงออกจากกลุ่มตัวอย่างที่เหลือ สุดท้ายแล้วจะทำการรวบรวมผลจากทุกๆปัญหาแบบการออกคะแนนเสียง (Majority Vote) เพื่อจะใช้ในการทำนายกลุ่มตัวอย่างต่อไป (Yu, 2012). รูปภาพที่ 2.3 และ 2.4 แสดงให้เห็นถึงการปรับใช้แนวความคิดดังกล่าวในงานด้านการวิเคราะห์ข้อมูลไมโครอะเรย์โดยการอ้างอิงระดับการทำงานของพาธเวย์



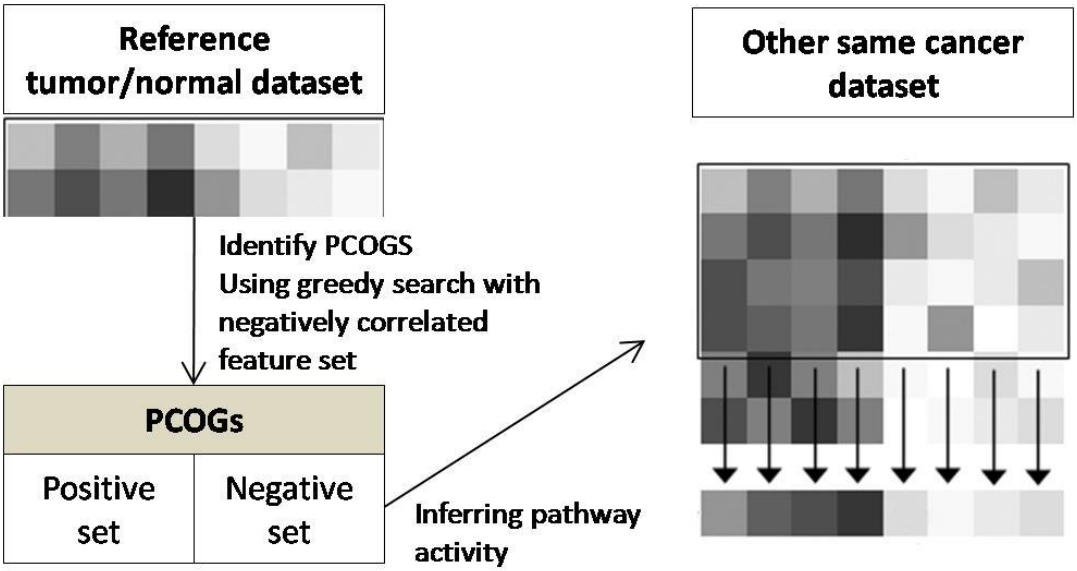
รูปที่ 2.3 ขั้นตอนของระเบียบวิธีการจำแนกกลุ่มตัวอย่างโดยใช้แนวความคิด OVA



รูปที่ 2.4 ขั้นตอนของระเบียบวิธีการจำแนกกลุ่มตัวอย่างโดยใช้แนวความคิด OVA

2.4.2 ระเบียบวิธีการ Reference-based Negatively Correlated Feature Set (R-NCFS)

ระเบียบวิธีการ Reference-based Negatively Correlated Feature Set (R-NCFS) ถูกพัฒนาบนพื้นฐานของระเบียบวิธีการ NCFS-i โดยตั้งสมมติฐานว่า ในชุดข้อมูลของโรคเดียวกัน ไม่ว่าจะเป็นการแบ่งกลุ่มสองกลุ่มตัวอย่าง หรือหลายกลุ่มตัวอย่าง ข้อมูล PCOGs ที่ได้น่าจะมีความใกล้เคียงและสามารถปรับใช้ด้วยกันได้ ดังนั้นระเบียบวิธีการนี้จะมีการนำชุดข้อมูลที่เป็นโรคเดียวกันที่มีกลุ่มตัวอย่างสองกลุ่ม (Tumor/Normal) มาทำการหา PCOGs ก่อนในขั้นแรก จากนั้นจึงนำ PCOGs ที่ได้ในการอ้างอิงระดับการทำงานของพาธเวย์ใน ชุดข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่มต่อไป ขั้นตอนของระเบียบวิธีการ R-NCFS ได้ถูกอธิบายไว้ในรูปที่ 2.5



รูปที่ 2.5 ขั้นตอนของระเบียบวิธีการ R-NCFS

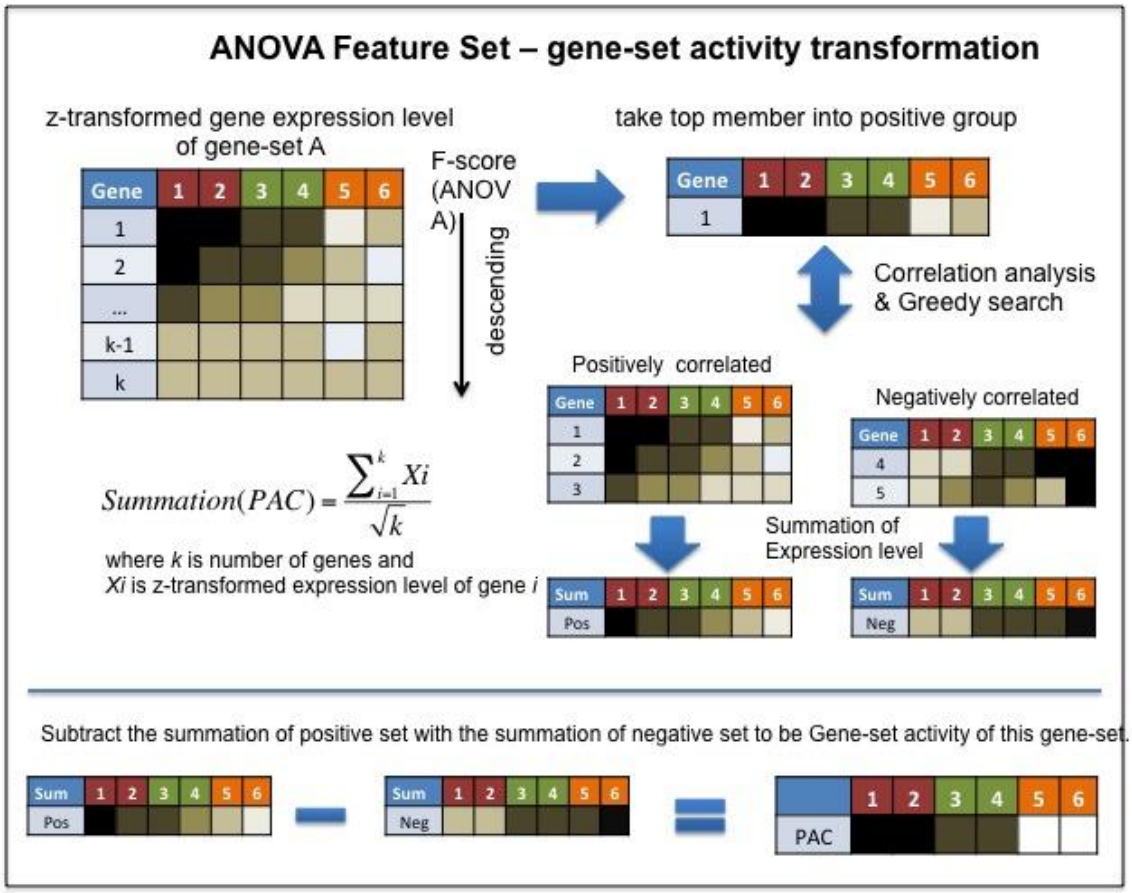
2.4.3 ระเบียบวิธีการ ANOVA-based Negatively Correlated Feature Set (AFS)

แม้ว่าแนวความคิดแบบ OVA และ OVO สามารถนำมาปรับใช้เพื่อต่อยอดระเบียบวิธีการ NCFS-i เพื่อให้สามารถนำไปใช้กับชุดข้อมูลหลายกลุ่มได้ แต่การสร้างแบบจำลองหลายๆแบบจำลองจากชุดข้อมูลเพียงบางส่วนนั้น จะทำให้บางแบบจำลองขาดข้อมูลภาพรวมไป และอาจจะทำให้ผลที่ได้จากแบบจำลองดังกล่าวไม่ถูกต้องครบถ้วนเมื่อเทียบกับการใช้แบบจำลองเดียวที่ถูกสร้างด้วยข้อมูลทั้งชุด อีกทั้งการปรับใช้ OVA และ OVO ยังเป็นการเพิ่มความซับซ้อนของอัลกอริทึม และทำมาต้องใช้เวลาในการคำนวนมากขึ้นเป็นทวีคูณอีกด้วย

เนื่องจาก NCFS-i มีการใช้ Student's *t*-test เข้ามาเพื่อช่วยเรียงยีนในแต่ละพาธเวย์ ตามความสามารถใน การแบ่งกลุ่มสองกลุ่ม สิ่งทำให้ NCFS-i ถูกจำกัดอยู่แค่ปัญหาการจำแนกกลุ่มสองกลุ่มตัวอย่าง ดังนั้น หากมีการปรับปรุงในส่วนนี้โดยการใช้ Analysis of Variance หรือ ANOVA testing เข้ามาแทนที่ในการ ช่วยเรียงยีนตามลำดับความแปรปรวนระหว่างกลุ่ม ก็จะทำให้ระเบียบวิธีการใหม่สามารถปรับใช้ได้กับชุด ข้อมูลที่มีจำนวนกลุ่มหลายกลุ่มได้ ระเบียบวิธีการใหม่นี้ได้ถูกให้ชื่อว่า ANOVA-based Feature Set (AFS)

AFS มีขั้นตอนการทำงานหลักอยู่ทั้งสิ้นสองส่วน คือส่วนของการเรียงและการค้นหา PCOGs และอีกส่วนคือส่วนของการคำนวนระดับการทำงานของพาธเวย์ (รูปที่ 2.6) โดยในแต่ละพาธเวย์สมาชิกยีนแต่ละตัวจะถูกคำนวนค่า Discriminative Score โดยอ้างอิงจากค่า F-value และทำการเรียงค่าดังกล่าวจากมากไปน้อย โดย ค่าที่มากกว่าจะแสดงให้เห็นว่ามีความแปรปรวนของการแสดงออกของยีนระหว่างกลุ่มมากกว่า จากนั้นใน ขั้นตอนของการหา PCOGs ยีนในลำดับแรกจะถูกนำไปไว้ใน PCOGs กลุ่มบวก และให้ระดับการทำงานของพาธเวย์เริ่มต้นเป็นค่าการแสดงออกยีนลำดับแรก หลังจากนั้นยีนในลำดับ ถัดไปจะถูกนำขึ้นมาเพื่อหาค่า Correlation กับค่าระดับการทำงานของพาธเวย์นั้นโดยการใช้ Pearson's Correlation หากค่า

Correlation เป็นบวก ยีนนั้นๆจะถูกนำไปไว้ในกลุ่มบวก ไม่เช่นนั้นจะถูกนำไปไว้ในกลุ่ม ลบ หลังจากนั้นจะทำการหาผลรวมค่าการแสดงออกของยีนในของกลุ่มบวก และกลุ่มลบ และนำค่าจากสอง กลุ่มมาหักลบกันเพื่อเป็นค่าระดับการทำงานของพาธเวย์ใหม่ F-value ของระดับการทำงานของพาธเวย์เก่า และใหม่จะถูกนำมาเปรียบเทียบกัน หากค่า F-value ของระดับการทำงานของพาธเวย์ใหม่มีค่ามากกว่าเดิม ขั้นตอนในการนำยีนลำดับถัดไปมาคิดต่อจะถูกวนไปเรื่อยๆ จนกว่าค่า F-value จะลดลงจากเดิม

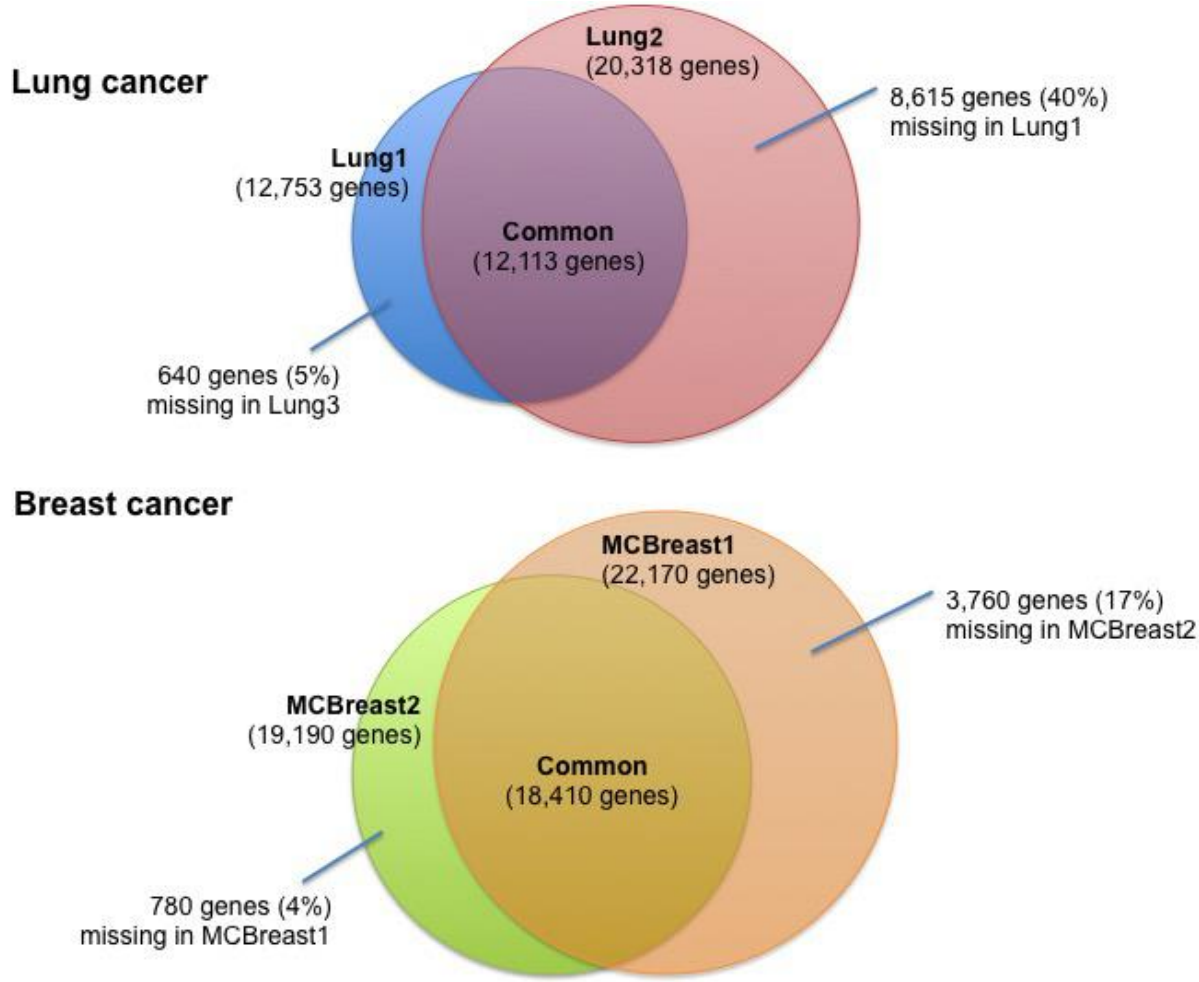


รูปที่ 2.6 ขั้นตอนการของระเบียบวิธีการ AFS

2.5 การวิเคราะห์ข้อมูลไมโครอะเรย์แบบข้ามแพลทฟอร์ม

การวิเคราะห์ข้อมูลไมโครอะเรย์ในเชิงบูรณาการนั้นได้ถูกทดสอบเพื่อแสดงให้เห็นว่าสามารถทำงานได้ดีกว่า การวิเคราะห์แบบดั้งเดิม อีกทั้งยังช่วยให้นักวิจัยเข้าใจกลไกการทำงานของเหตุการณ์ทางพันธุกรรมได้มาก ขึ้น (Xu, 2005; Wang, 2011) อย่างไรก็ตามยังมีปัญหาบางอย่างที่ส่งผลถึงประสิทธิภาพของการวิเคราะห์ข้อ มูลไมโครอะเรย์ หนึ่งในนั้นก็คือการวิเคราะห์ไมโครอะเรย์แบบข้ามแพลทฟอร์ม เนื่องจากเทคโนโลยีไมโคร อะเรย์ได้ถูกพัฒนาอย่างหลากหลายทำให้เกิดเป็นแพลทฟอร์มต่างๆขึ้นมากมาย โดยปัญหาหลักในการ วิเคราะห์ข้อมูลไมโครอะเรย์ข้ามแพลทฟอร์มคือการที่แต่ละแพลทฟอร์มจะทำการวัดระดับการแสดงออก ของยีนจำนวนไม่เท่ากัน ทำให้หากมีการสร้างแบบจำลองในแพลทฟอร์มนึง และเมื่อ

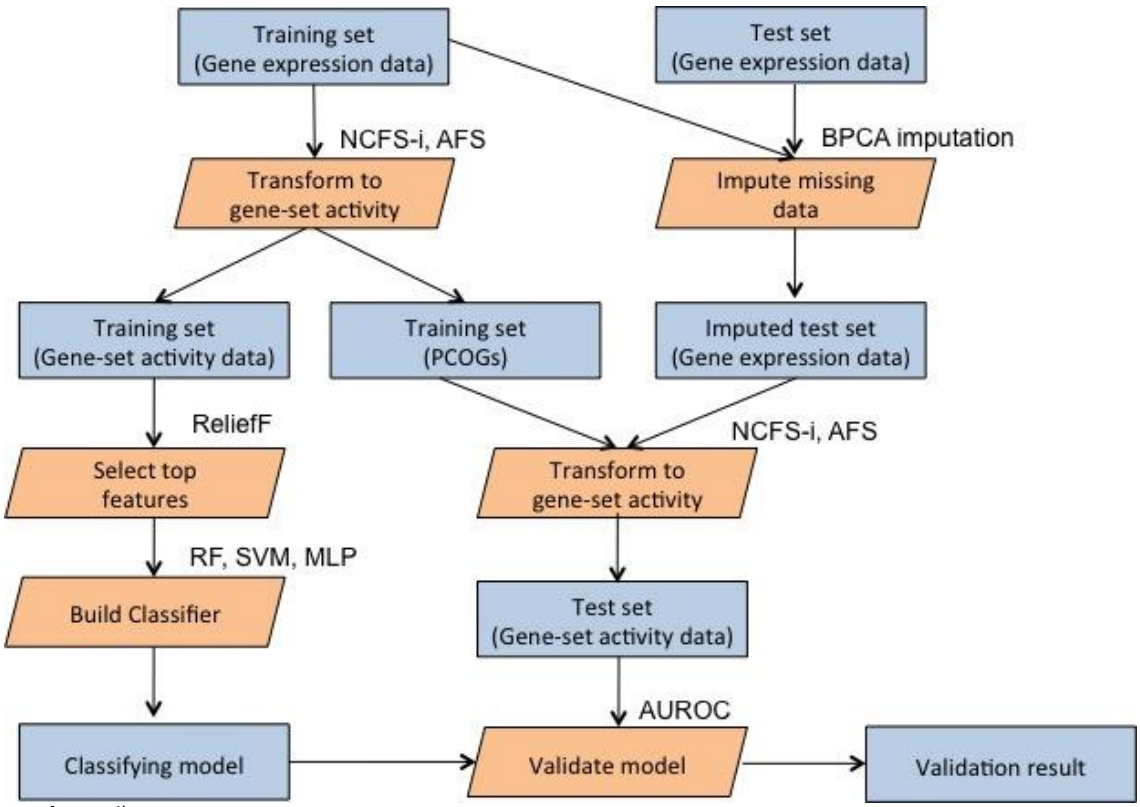
ไปใช้ทำนายในข้อมูล จากอีกแพลทฟอร์มนึงซึ่งมีบางยีนขาดหายไป เกิดเป็นปัญหาข้อมูลสูญหาย (Missing data) ขึ้น ทำให้เกิด ปัญหาขึ้นเมื่อปรับใช้แบบ จำลองดังกล่าวโดยตรง ซึ่งเราสามารถใช้กับข้อมูลดังกล่าวได้ แต่อาจจะให้ผลไม่ดี พอ ทั้งนั้นงานวิจัยชิ้นนี้ได้ทำการเปรียบเทียบการใช้แบบจำลองข้ามแพลทฟอร์ม โดยตรงกับการทำการประเมินข้อมูลที่สูญหายในชุดทดสอบก่อน จึงนำแบบจำลองมาปรับใช้ (Missing Data Imputation) โดยงานวิจัย นี้ได้ใช้วิธีการ Baysian-based Principle Component Analysis (BPCA) มาใช้ในการประมาณข้อมูลที่สูญหาย BPCA เป็นระเบียบวิธีการประมาณข้อมูลที่สูญหายโดยใช้เทคนิคทางด้านสถิติที่รวม Principal Component regression, Bayesian estimation และ Expectation-maximization (EM)-like repetitive อัลกอริทึมเข้าไว้ด้วยกัน (Oba, 2003). โดยในการศึกษาครั้งนี้ได้มีการปรับใช้ BPCA จาก pcaMethods library ของ R package (Stacklies, 2007). ซึ่งได้ทดสอบโดยใช้ข้อมูลไมโครอะเรย์ชุด Lung1, Lung2, MCBreast1 และ MCBreast2 ซึ่งระดับของความสูญหายของข้อมูลในชุดข้อมูลนี้มีตั้งแต่ ~4% ถึง ~40% ดังแสดงในรูปที่ 2.7



รูปที่ 2.7 Venn's diagram แสดงจำนวนยีนที่พบและสูญหายในแต่ละชุดข้อมูล

การทดลองครั้งนี้ได้ถูกทดสอบทั้งกับระเบียบวิธีการ NCFS-i และ AFS โดย NCFS-i จะถูกทดสอบโดยใช้ข้อมูลชุด Lung1 และ Lung2 ในขณะที่ AFS ถูกทดสอบโดยข้อมูลชุด MCBreast1 และ MCBreast2 ซึ่งเป็นข้อมูลที่มีกลุ่มตัวอย่างมากกว่าสองกลุ่ม

การทดลองนี้ได้มีการใช้ระเบียบวิธีการลดจำนวนของ Feature ลง (Feature Selection) เพื่อเพิ่มประสิทธิภาพในการจำแนกกลุ่มตัวอย่าง (Guyon, 2003) ซึ่งอัลกอริทึมที่ถูกเลือกใช้ในที่นี่คือ ReliefF โดยอัลกอริทึมนี่ได้ถูกปรับใช้ในงานวิจัยอย่างหลากหลายรวมถึงในงานวิจัยทางด้านชีวสารสนเทศอีกด้วย (Kononenko, 1994; Robnik-Šikonja, 2003; Wang, 2004) Feature selection จะถูกนำมาใช้บนชุดข้อมูลที่ใช้สร้างแบบจำลอง โดยจะมีการเลือกจำนวนที่แตกต่างกันตั้งแต่ 1 ถึง 50 Feature จากนั้นจะถูกนำไปสร้างเป็นแบบจำลอง โดยอัลกอริทึมพื้นฐานสองตัวประกอบด้วย Random Forest (RF) และ Support Vector Machine (SVM) แบบจำลองที่สร้างขึ้นจะถูกประเมินค่าความถูกต้องโดยใช้ข้อมูลชุดทดสอบที่แบบจำลองไม่เคยเห็นมาก่อน โดยค่าความถูกต้องจะถูกวัดโดยอ้างอิง พื้นที่ใต้ ROC curve (AUROC) ทั้งนี้ระเบียบวิธีการ Feature Selection และการสร้างแบบจำลองถูกปรับใช้จาก Java Library ของ Weka ขั้นตอนการศึกษาของการทดลองนี้สามารถอธิบายเป็นขั้นตอนตามรูปที่ 2.8



รูปที่ 2.8 ขั้นตอนการศึกษาการวิเคราะห์ข้อมูลข้ามแพลทฟอร์ม

2.6 การประเมินความสามารถในการจำแนกกลุ่มตัวอย่าง

เพื่อที่จะประเมินประสิทธิภาพของระเบียบวิธีการใหม่ที่พัฒนาขึ้น ทางผู้วิจัยได้มีการเปรียบเทียบการจำแนกกลุ่มทั้งในชุดข้อมูลที่มีกลุ่มตัวอย่างสองกลุ่ม และหลายกลุ่ม โดยเปรียบเทียบระเบียบวิธีการที่พัฒนาขึ้นกับระเบียบวิธีการต่างๆที่มีอยู่ ซึ่งในงานวิจัยนี้ได้มีการทำตรวจวัดความถูกต้องสองแบบ คือแบบใช้ชุดข้อมูลเดียว และแบบการตรวจวัดความถูกต้องโดนใช้ข้อมูลชุดทดสอบที่ไม่เกี่ยวกับชุดข้อมูลที่ใช้สร้างแบบจำลอง สำหรับค่าความถูกต้องของการจำแนกกลุ่มตัวอย่างนั้น ในงานวิจัยชิ้นนี้ได้อ้างอิงไปถึงค่าทางสถิติสามตัว ประกอบด้วย เปอร์เซ็นความถูกต้อง, พื้นที่ใต้กราฟ ROC และ Recall เนื่องด้วยเปอร์เซ็นความถูกต้องนั้นถูกพบว่าเป็นตัวบ่งบอกความถูกต้องที่เอนเอียงหากมีจำนวนสมาชิกระหว่างกลุ่มต่างกันมาก ซึ่งการใช้พื้นที่ใต้กราฟ ROC จะเป็นตัวบ่งบอกที่แนะนำให้ใช้แทนเปอร์เซ็นความถูกต้อง (Kotsiantis, 2006) สำหรับ Recall หรือ Type II error นั้นเป็นตัวบ่งบอกที่มีบทบาทมากกว่า Type I error ในงานด้านการวินิจฉัยโรคเบื้องต้น (Goutte, 2005)

2.6.1 Stratified K-fold cross-validation

การตรวจวัดความถูกต้องแบบ Stratified k-fold cross-validation เป็นการตรวจวัดความถูกต้องโดยใช้ข้อมูลชุดเดียว โดยจะสุ่มแบ่งข้อมูลออกเป็นจำนวน k ชุดในขณะที่แต่ละชุดจะมีสัดส่วนระหว่างกลุ่มตัวอย่างใกล้เคียงกัน ชุดตัวอย่างย่อยหนึ่งชุดจะถูกเก็บไว้เพื่อใช้วัดค่าความถูกต้อง ในขณะที่ชุดที่เหลือจะถูกรวมกันเพื่อนำไปใช้สร้างแบบจำลอง ขั้นตอนการตรวจวัดความถูกต้องจะถูกทำซ้ำๆจนกว่าชุดข้อมูลย่อยทุกชุดจะถูกใช้เป็นชุดทดสอบ การตรวจวัดแบบ Stratified K-fold cross-validation ถูกใช้ในงานวิจัยอย่างหลากหลาย รวมถึงการนำไปใช้ในงานวิจัยประเภท Self Optimizing

2.6.2 Cross-dataset validation

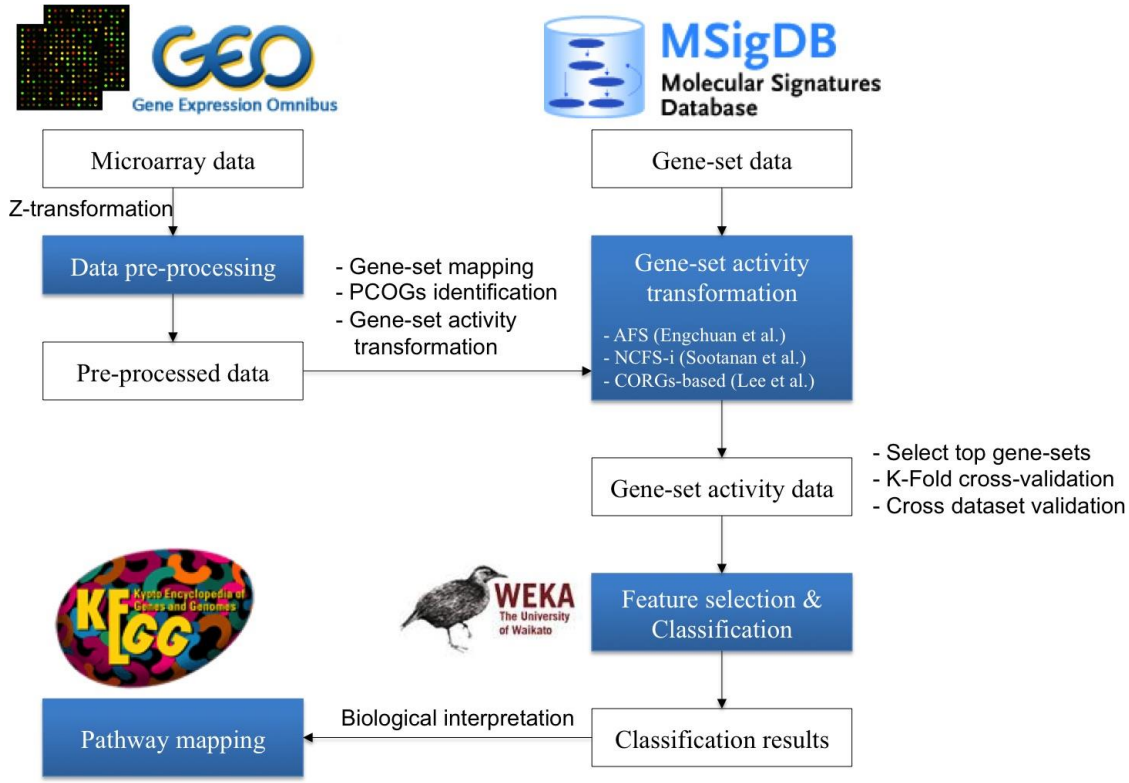
การตรวจวัดความถูกต้องแบบ Cross-dataset validation เป็นการตรวจวัดโดยใช้ข้อมูลคนละชุดกับชุดข้อมูลที่ใช้สร้างแบบจำลอง ซึ่งโดยปกติแล้ว ในข้อมูลไมโครอะเรย์นั้นมักจะมีปัจจัยหลากหลายชนิดเข้ามามีผลกับการวัดและประมาณระดับการแสดงออกของยีน อาทิเช่น เรื่องความสมบูรณ์ หรือคุณภาพของ RNA ของตัวอย่าง (Fleige, 2006) ดังนั้นการใช้ Cross-dataset validation จะเป็นการวัดทั้งความถูกต้องและความเสถียรของระเบียบวิธีการที่ใช้จำแนกกลุ่มตัวอย่างในข้อมูลไมโครอะเรย์ต่อปัจจัยที่อาจจะผลกระทบดังกล่าว

2.7 การพัฒนาระเบียบวิธีการเป็น Web application และ Java library

หลังจากการพัฒนาระเบียบวิธีการสำหรับการจำแนกกลุ่มตัวอย่างหลายกลุ่มโดยการใช้ข้อมูลไมโครอะเรย์ร่วมกับข้อมูลพาธเวย์เสร็จสิ้น ในการศึกษานี้ยังได้พัฒนาระเบียบวิธีการดังกล่าวให้อยู่ในรูปแบบของ

โปรแกรมซึ่งอนุญาตให้ผู้ใช้สามารถใช้ผ่านระบบออนไลน์ชื่อว่า Gene-set Activity Toolbox (GAT) โดยโปรแกรมนี้ทำการดึงข้อมูลไมโครอะเรย์จากฐานข้อมูล GEO หลังจากนั้น ผู้ใช้สามารถเลือกใช้ข้อมูล พาธเวย์ หรือข้อมูลกลุ่มยีนจากฐานข้อมูล MSigDB เพื่อนำมาใช้ในการอ้างอิงระดับการทำงานของพาธเวย์ โดยวิธีต่างๆ อาทิ เช่น AFS, NCFS-i, ฯลฯ หลังจากดำเนินการอ้างอิงระดับการทำงานของพาธเวย์แล้ว ผู้ใช้สามารถสร้างแบบจำลองและตรวจวัดความถูกต้องของแบบจำลองที่สร้างขึ้นได้ ทั้งนี้ในส่วนของ Feature Selection ยังสามารถเชื่อมต่อไปยังฐานข้อมูล KEGG เพื่อช่วยในการวิเคราะห์ผลได้อีกด้วย ระบบโดยรวมของ GAT สามารถอธิบายเป็นขั้นตอนได้ตามรูปที่ 2.9

ทั้งนี้ระบบ GAT สามารถเข้าถึงได้ผ่านทาง http://pat.sit.kmutt.ac.th โดยทางเว็บไซต์ได้มีการจัดชุดข้อมูลระดับการทำงานของพาธเวย์จากข้อมูลไมโครอะเรย์ และ Java library ของ GAT ไว้ให้ด้วย



รูปที่ 2.9 ขั้นตอนการทำงานของ Gene-set Activity Toolbox

บทที่ 3

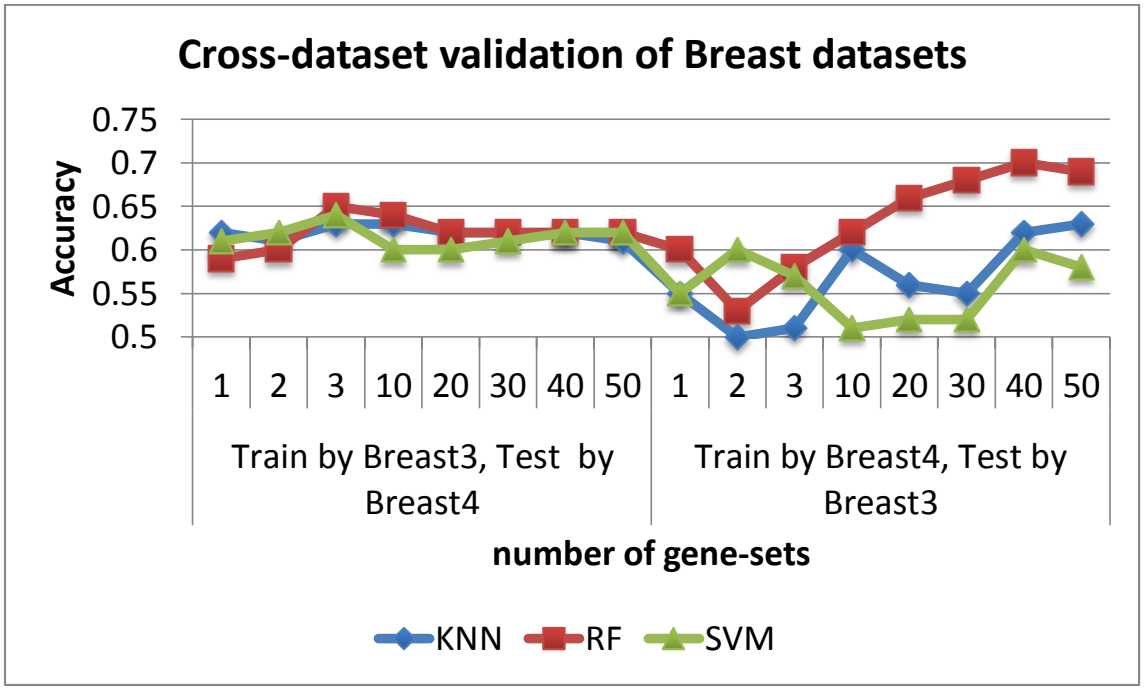
ผลการดำเนินงานวิจัย

3.1 การวิเคราะห์ข้อมูลไมโครอะเรย์ร่วมกับข้อมูลพาธเวย์

ทางผู้วิจัยได้ทำการศึกษาวิธีการปรับใช้ระเบียบวิธีการอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลทดสอบเบื้องต้น เนื่องจากวิธีการ NCFS-i ต้องการข้อมูลการเป็นโรคของกลุ่มตัวอย่างเพื่อนำไปใช้หากลุ่มยีน PCOGs ที่จะถูกใช้ในการอ้างอิงระดับการทำงานของพาธเวย์นั้นไม่สามารถปรับใช้ได้กับชุดข้อมูลทดสอบ ซึ่งไม่มีข้อมูลการเป็นโรคของกลุ่มตัวอย่างได้ โดยในงานวิจัยนี้ได้นำเสนอแนวทางการปรับใช้สองแนวทาง คือการทำการจำแนกเบื้องต้น (Pre-Classification) เพื่อใช้ทำนายข้อมูลการเป็นโรคของกลุ่มตัวอย่างเบื้องต้น และอีกวิธีนึง คือการใช้ PCOGs จากชุดข้อมูลที่ใช้สร้างแบบจำลอง

3.1.1 การเปรียบเทียบ Classification และการเลือกใช้จำนวนของยีน

ในการทดลองนี้เป็นการเปรียบเทียบอัลกอริทึมสำหรับการจำแนกกลุ่มตัวอย่าง และการเลือกใช้จำนวนของยีนต่างๆกันในการสร้างแบบจำลอง เพื่อช่วยในการเลือกใช้อัลกอริทึมและจำนวนของยีนในการทำ Pre-classification โดยชุดข้อมูลที่นำมาใช้ในการทดลองนี้คือ Breast3 และ Breast4 โดยประเมินความถูกต้องของ แต่ละแบบจำลองโดย Cross-dataset validation อัลกอริทึมที่ถูกเปรียบเทียบในการทดลองนี้คือ KNN, RF และ SVM

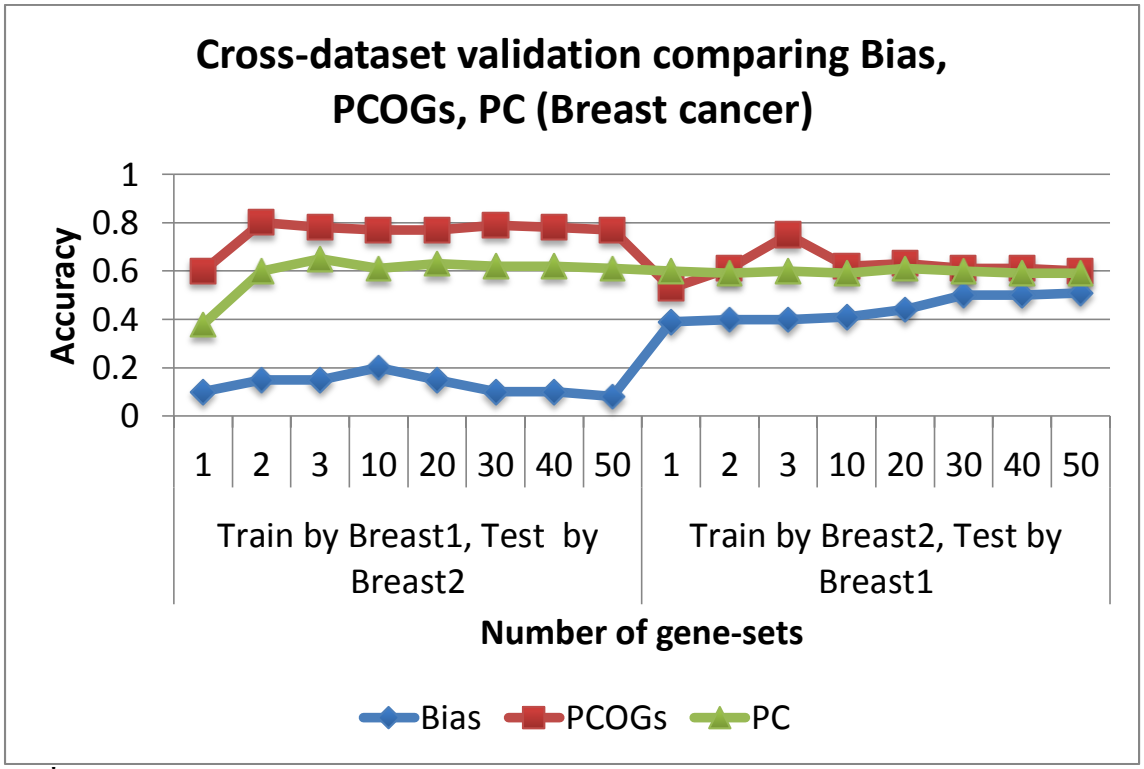


รูปที่ 3.1 ผลการจำแนกกลุ่มตัวอย่างโดยแบบจำลองต่างๆที่สร้างจากข้อมูลไมโครอะเรย์

อ้างอิงจากผลในรูปที่ 3.1 พบว่า Random Forest (RF) (เส้นสีแดง) มีประสิทธภาพที่สูงกว่าอัลกอริทึมแบบอื่นเมื่อจำนวนของยีนที่ใช้ในการสร้างแบบจำลองมากกว่า 10 ยีนขึ้นไป ดังนั้น RF จะถูกใช้ในการจำแนกกลุ่มตัวอย่างเบื้องต้นโดยใช้จำนวนยีนที่ 40 ยีนในการสร้างแบบจำลอง เนื่องจากเป็นจำนวนที่ RF มีประสิทธิภาพมากที่สุด

3.1.2 การเปรียบเทียบวิธีการปรับใช้ NCFS-i ในชุดข้อมูลทดสอบ

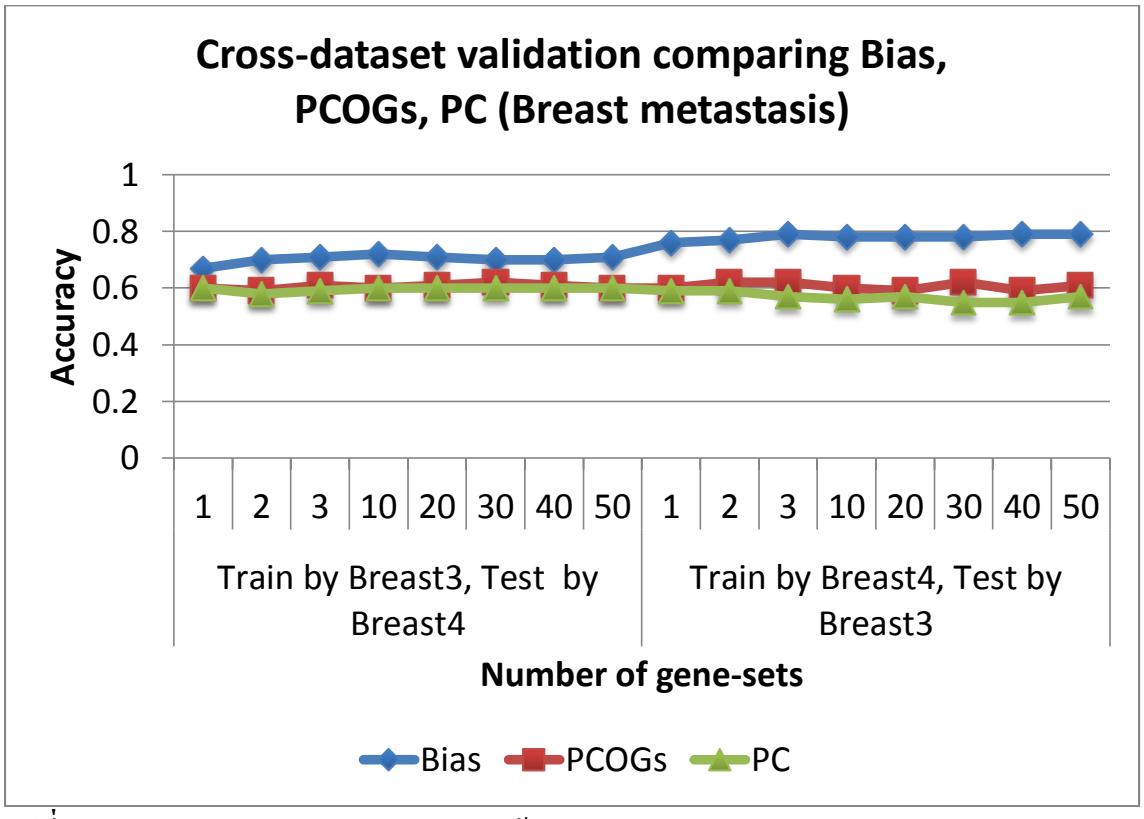
หลังจากได้อัลกอริทึม และจำนวนที่เหมาะสมสำหรับการทำ Pre-classification แล้ว การทดลองนี้มีเพื่อเปรียบเทียบวิธีการปรับใช้ NCFS-i กับชุดข้อมูลทดสอบ โดยได้นำข้อมูลชุด Breast1, Breast2, Breast3, Breast4, MCLung1, MCLung2 และ MCLung3 เพื่อมาตรวจวัดความถูกต้องและเปรียบเทียบการปรับใช้ NCFS-i ทั้งสองวิธี ทั้งนี้ ผู้วิจัยยังได้เปรียบ กับการปรับใช้ NCFS-i โดยใช้ข้อมูลการเป็นโรคของชุดข้อมูลทดสอบร่วมด้วย (Bias) อีกด้วย โดยการใช้ PCOGs จากชุดข้อมูลที่ใช้สร้างแบบทดลองจะใช้คำแทนในผลเป็น PCOGs สำหรับการทำ Pre-classification จะใช้คำแทนเป็น PC และการใช้ข้อมูลการเป็นโรคของกลุ่มตัวอย่างของชุดทดสอบเองใช้คำแทนเป็น Bias โดย RF ยังถูกใช้เป็นอัลกอริทึมสำหรับจำแนกกลุ่มตัวอย่างอีกด้วย ในขณะที่ Feature Selection ถูกนับมาปรับใช้เพื่อเลือกเฉพาะพาธเวย์ที่ดีที่สุดจำนวน 1, 2, 3, 10, 20, 30, 40, 50 พาธเวย์ โดย Single Pathway Classification (SPC) ranker



รูปที่ 3.2 ผล Cross-dataset validation ของชุดข้อมูล Breast1 และ Breast2

เนื่องจากข้อมูล MCLung1 และ MCLung2 เป็นข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่มดังนั้นจึงไม่เหมาะกับการใช้ NCFS-i ทั้งนี้ในขั้นตอนการทดสอบเบื้องต้น จะใช้ข้อมูลทั้งสองชุดนี้ในการจำแนกประเภทย่อยของโรคมะเร็งปอด (Adenocarcinoma:AC, Squamous Cell Carcinoma:SCC) เท่านั้น

จากผล Cross-dataset validation ในรูปที่ 3.2 แสดงให้เห็นว่าการปรับใช้ NCFS-i กับข้อมูลชุดทดสอบโดยใช้ PCOGs จากข้อมูลที่ใช้สร้างแบบจำลองนั้น เมื่อใช้จำนวนพาธเวย์เท่ากับ 2-3 พาธเวย์เป็น Feature ในการสร้างโมเดล จะให้แบบจำลองที่มีประสิทธิภาพในการจำแนกกลุ่มตัวอย่างมากที่สุด (ความถูกต้อง = 0.81 เมื่อใช้สองพาธเวย์ และความถูกต้อง = 0.72 เมื่อใช้สามพาธเวย์) ซึ่งผลในเรื่องจำนวนพาธเวย์ที่ใช้ยังตรงกับงานวิจัยก่อนหน้า (Chan, 2011)

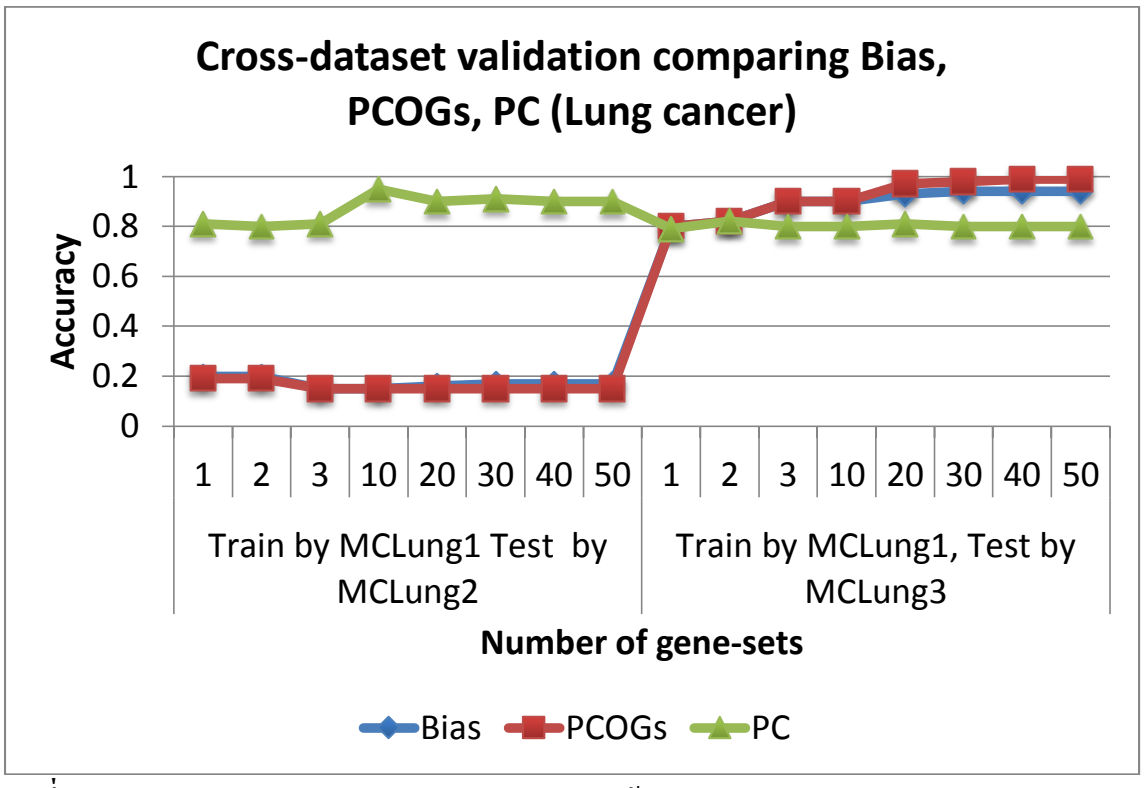


รูปที่ 3.3 ผล Cross-dataset validation ของชุดข้อมูล Breast3 และ Breast4

สำหรับผล Cross-dataset validation ในชุดข้อมูล Breast3 และ Breast4 การใช้ PCOGs จากข้อมูลที่ใช้สร้างแบบจำลองไม่ได้ให้ผลที่ดีกว่าแบบ Bias แต่ก็ยังให้ผลที่ดีกว่าการทำ Pre-classification (รูปที่ 3.3)

นอกเหนือจากการทดลองระเบียบวิธีการเบื้องด้วยชุดข้อมูล Breast1, Breast2, Breast3 และ Breast4 ทางผู้วิจัยยังได้ทดสอบกับข้อมูล MCLung1, MCLung2 และ MCLung3 อีกด้วย และด้วย MCLung1 เป็นชุดข้อมูล ที่มีสัดส่วนระหว่างกลุ่มที่ใกล้เคียงกันมากที่สุด ดังนั้น MCLung1 จึงถูกเลือกใช้ให้เป็นข้อมูลสำหรับการ สร้างแบบจำลอง และทดสอบแบบจำลองที่ได้ด้วยชุดข้อมูล MCLung2 และ MCLung3 รูปที่ 3.4 แสดงผลลัพธ์ของการทดลองนี้ โดยผลการทดลองแสดงให้เห็นว่า การปรับใช้ NCFS-i โดยใช้ข้อมูล PCOGs จากชุดข้อมูลสร้างแบบจำลองนั้น ให้ผลที่การตรวจวัดความถูกต้องแบบ Cross-dataset validation กว่าระเบียบ

วิธีการปรับใช้แบบอื่นๆ โดยจากผลการทดลองทั้งสามชุดที่ชี้ไปในทางเดียวกันจึงสามารถสรุปได้ว่า หากต้องการปรับใช้ NCFS-i กับข้อมูลชุดทดสอบนั้น การใช้ข้อมูล PCOGs จากข้อมูลที่ใช้สร้างแบบจำลอง เป็นวิธีที่ดีที่สุดสำหรับการอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลทดสอบนั้นๆ



รูปที่ 3.4 ผลการทดลอง Cross-dataset validation ในชุดข้อมูล MCLung

ทั้งนี้ นอกเหนือจากผลการตรวจวัดความถูกต้องโดย Cross-dataset validation แล้ว ผู้วิจัยยังได้วิเคราะห์ถึงความเกี่ยวพันของผลการวิเคราะห์กับความรู้ทางชีวภาพอีกด้วย โดยได้ทำการเลือกพาธเวย์ที่ดีที่สุดจำนวน 10 พาธเวย์จากชุดข้อมูล MCLung1 เพื่อนำมาวิเคราะห์ในขั้นถัดไปดังแสดงในตารางที่ 3.1

หลักจากที่ได้พาธเวย์ทั้ง 10 พาธเวย์มาแล้ว หลักฐานแสดงความเชื่อมโยงของพาธเวย์ต่อโรคมะเร็งปอดได้ถูก อ้างอิงโดยค้นหาสื่อสิ่งพิมพ์ที่มีการยืนยันถึงความสัมพันธ์ดังกล่าวจาก สำนักพิมพ์ที่มีชื่อเสียง ในส่วนของ Pathway in cancer เป็นพาธเวย์ที่เกี่ยวข้องกับการเกิดโรคมะเร็งอยู่แล้วจึงไม่จำเป็นต้องมีเอกสารอ้างอิง ซึ่ง จากผลการวิเคราะห์นั้นแสดงให้เห็นว่า มีเพียงพาธเวย์เดียวคือ Long term potentiation ที่ไม่มีเอกสารอ้างอิง ในขณะที่พาธเวย์ที่เหลือมีเอกสารอ้างอิงมาสนับสนุนทั้งสิ้น ทำให้เห็นว่าการใช้ SPC ranker และการใช้ข้อมูลระดับการทำงานของพาธเวย์นั้นมีประสิทธิภาพและให้ผลลัพธ์ในการวิเคราะห์โรคที่น่าเชื่อถือได้

ตารางที่ 3.1 รายชื่อพาธเวย์ที่ดีที่สุด 10 พาธเวย์จาก MCLung1 โดยการใช้ SPC ranker

ลำดับที่	พาธเวย์	อ้างอิง
1	Cytokine, Cytokine receptor interaction	Van Dyke *et al*.
2	Melanogenesis	Bellei *et al*.
3	WNT signaling pathway	Mazieres *et al*.
4	ECM receptor interaction	Devaraj *et al*.
5	Phosphotidylinositol signaling system	Hanai *et al*., Tsurutani *et al*.
6	Calcium signaling pathway	Yang *et al*.
7	Long term potentiation	-
8	Leukocyte transendothelial migration	Lu et al.
9	NOTCH signaling pathway	Westhoff *et al*.
10	Pathway in cancer	-

3.2 การจำแนกกลุ่มตัวอย่างหลายกลุ่มโดยการใช้ข้อมูลพาธเวย์ร่วมด้วย

3.2.1 การเปรียบเทียบ R-NCFS และ AFS

ในการศึกษาขั้นต้นนี้ ชุดข้อมูลการจำแนกกลุ่มผู้ป่วยโรคมะเร็งปอดหลายกลุ่ม (MCLung1, MCLung2 และ MCLung3) ได้ถูกใช้เพื่อทดสอบระเบียบวิธีการทั้ง R-NCFS และ AFS นอกจากนี้ ผู้วิจัยยังได้เปรียบเทียบประสิทธิภาพของทั้งสองระเบียบวิธีการกับวิธีการก่อนหน้าที่เคยทำงานวิจัยเอาไว้ใน Engchuan et al., 2012 ซึ่งในงานวิจัยนั้น ผู้วิจัยได้เสนอให้ทำการจำแนกกลุ่มตัวอย่างจากชุดข้อมูลดังกล่าวด้วยวิธีการแบบ two-stage approach กล่าวคือ จะมีการทำการจำแนกสองครั้ง โดยครั้งแรกจะเป็นการจำแนกประเภทย่อยของโรคมะเร็งปอด (AC, SCC) และครั้งที่สองจะทำการจำแนกระดับของโรคมะเร็งของแต่ละประเภทย่อย (ระดับที่ 1 และ 2) หลังจากนั้นจึงรวมผลการจำแนกของทั้งสองครั้งเข้าด้วยกัน เพื่อใช้ในการจำแนกประเภทย่อย และระดับของโรคมะเร็งในคราวเดียวกัน โดยมีการใช้ NCFS-i ในการจำแนกกลุ่มตัวอย่างในแต่ละครั้ง

โดยในการทดลองนี้ได้มีการใช้อัลกอริทึมสำหรับการจำแนกกลุ่มตัวอย่างเป็น Random Forest เนื่องจากอัลกอริทึมนี้มีความเสถียรต่อปัญหา Over-fitting และยังสามารถให้คำตอบเป็นช่วงความน่าจะเป็นได้ (Probabilistic prediction) ในส่วนของการทำ Feature selection ทางผู้วิจัยได้เลือกใช้ SPC ranker เพื่อเลือกพาธเวย์มาใช้ในการสร้างแบบจำลองโดยมีการเลือกพาธเวย์จำนวน 1, 3, 5, *10, 30, 50* พาธเวย์ สำหรับการตรวจวัดความถูกต้องนั้นจะใช้ทั้งระเบียบวิธี stratified three-fold cross-validation และ cross-dataset validation โดยค่าความถูกต้องถูกอ้างอิงจากค่าพื้นที่ใต้กราฟ ROC (AUROC) เนื่องจากเป็นตัววัดที่น่าเชื่อถือในกรณีที่ข้อมูลมีความเบี่ยงเบนในสัดส่วนระหว่างกลุ่มตัวอย่างซึ่งปกติ AUROC จะใช้สำหรับตรวจวัด

ค่า ความถูกต้องสำหรับการจำแนกกลุ่มตัวอย่างสองกลุ่มทั้งนี้ AUROC ได้ถูกพัฒนาเพิ่มเติมด้วยแนวความคิด แบบ OVA ทำให้สามารถนำมาปรับใช้ในงานนี้ได้ (Hand, 2001)

3.2.2 ผลการเปรียบเทียบ AFS และ R-NCFS โดย Stratified three-fold cross-validation

เนื่องจากจำนวนตัวอย่างในชุดข้อมูลการจำแนกกลุ่มตัวอย่างมะเร็งปอดหลายกลุ่มนั้นมีจำนวนที่น้อยทำให้ไม่เพียงพอต่อการทำ k-fold cross-validation ด้วยจำนวน k ที่มากเช่น 10-fold cross-validation ดังนั้นงานวิจัยนี้จึงกำหนด k เท่ากับ 3 โดยผลของ three-fold cross-validation ในตารางที่ 3.2 นั้นแสดงให้เห็นว่า AFS และ 2-stage approach นั้นมีประสิทธิภาพที่ดี โดยเฉพาะอย่างยิ่งในข้อมูลชุด MCLung3 ซึ่งมีจำนวนกลุ่มตัวอย่างเพียงแค่ 3 กลุ่มตัวอย่าง ในขณะที่ R-NCFS ให้ผลที่ยอมรับได้ในข้อมูล ทุกชุดยกเว้น MCLung2

จากผลการทดลองพบว่าการจำแนกกลุ่มตัวอย่างพบว่าให้ผลที่ดีสุดในชุดข้อมูล MCLung3 และแย่ที่สุดในชุดข้อมูล MCLung2 โดยชุดข้อมูล MCLung2 นั้นมีกลุ่มตัวอย่างทั้งสิ้น 4 กลุ่มตัวอย่าง และแต่ละกลุ่มตัวอย่างยังมีจำนวนที่แตกต่างกันระหว่างประเภทย่อยของโรคมะเร็งปอด (AC และ SCC) ดังนั้นการที่มีสัดส่วนระหว่างกลุ่มไม่เท่ากันเช่นนี้ก็ส่งผลในทางลบให้กับประสิทธิภาพการจำแนกกลุ่มตัวอย่างได้ ในขณะที่ MCLung3 อาจจะมีสัดส่วนระหว่างกลุ่มไม่เท่ากันเช่นกัน แต่ทั้งนี้ในชุดข้อมูลดังกล่าวมีจำนวนของกลุ่มตัวอย่างเพียงแค่สามกลุ่มเท่านั้น

ตารางที่ 3.2 ผล Stratified three-fold cross-validation เปรียบเทียบ AFS, R-NCFS และ 2-stage approach

ชุดข้อมูล	จำนวนพาธเวย์	ระเบียบวิธีการ		
		AFS	R-NCFS	2-stage
MCLung1	1	0.90	0.56	0.90
	3	0.95	0.74	0.95
	5	0.96	0.75	0.97
	10	0.97	0.77	0.98
	30	0.96	0.76	0.97
	50	0.95	0.82	0.95
MCLung2	1	0.62	0.57	0.78
	3	0.70	0.55	0.85
	5	0.69	0.50	0.94
	10	0.75	0.57	0.95
	30	0.75	0.45	0.94
	50	0.85	0.65	0.91

ชุดข้อมูล	จำนวน พาธเวย์	ระเบียบวิธีการ		
		AFS	R-NCFS	2-stage
MCLung3	1	1	1	0.94
	3	1	0.99	0.90
	5	1	0.99	0.98
	10	1	1	0.97
	30	1	1	0.96
	50	1	1	0.95

3.2.2.1. ผลจากการทำ Cross-dataset validation

โดยทั่วไปแล้วการวิเคราะห์ข้อมูลไมโครอะเรย์นั้นมักจะมีปัจจัยหลายๆอย่างที่ส่งผลให้เกิดความแปรปรวนในชุดข้อมูล ดังนั้นการทำการตรวจวัดความถูกต้องโดยวิธี Cross-dataset validation จึงถูกใช้เพื่อวัดความเสถียรของระเบียบวิธีต่อปัจจัยดังกล่าว โดยผลจากการทำ Cross-dataset validation พบว่า 2-stage approach และ R-NCFS นั้นให้ประสิทธิภาพที่ต่ำกว่า AFS ทั้งคู่ (ตารางที่ 3.3) ซึ่งจากผลในสองส่วนแสดงให้เห็นว่าระเบียบวิธี AFS นั้นมีความเสถียรและถูกต้องมากกว่าระเบียบวิธีอื่นๆในการจำแนกกลุ่มตัวอย่างหลายกลุ่มโดยการใช้ข้อมูลระดับการทำงานของพาธเวย์

ตารางที่ 3.3 ผล Cross-dataset validation เปรียบเทียบประสิทธิภาพของ AFS, R-NCFS และ 2-stage

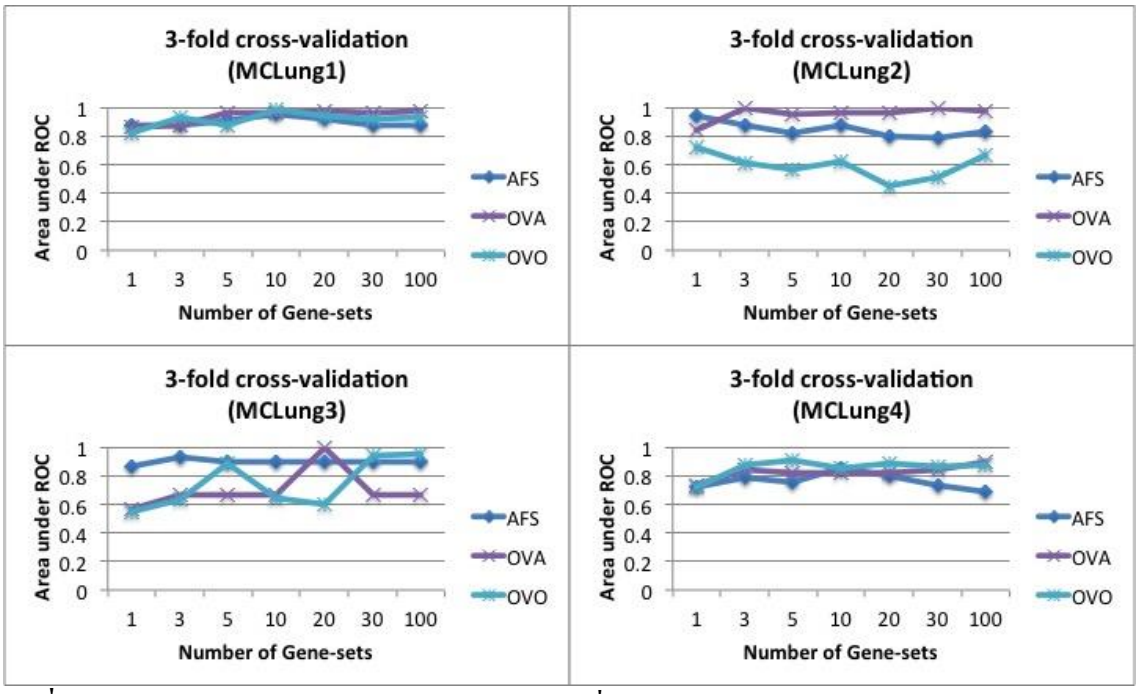
Test dataset	#N of gene-sets	Gene-set-based technique		
		AFS	R-NCFS	2-stage
MCLung2	1	0.45	0.60	0.66
	3	0.75	0.63	0.64
	5	0.76	0.70	0.70
	10	0.80	0.60	0.74
	30	0.81	0.79	0.75
	50	0.82	0.73	0.74
MCLung3	1	0.62	0.78	0.57
	3	0.76	0.70	0.64
	5	0.85	0.81	0.73
	10	0.84	0.70	0.65
	30	0.97	0.68	0.70
	50	0.97	0.86	0.70

3.2.3 เปรียบเทียบ AFS กับ OVA และ OVO

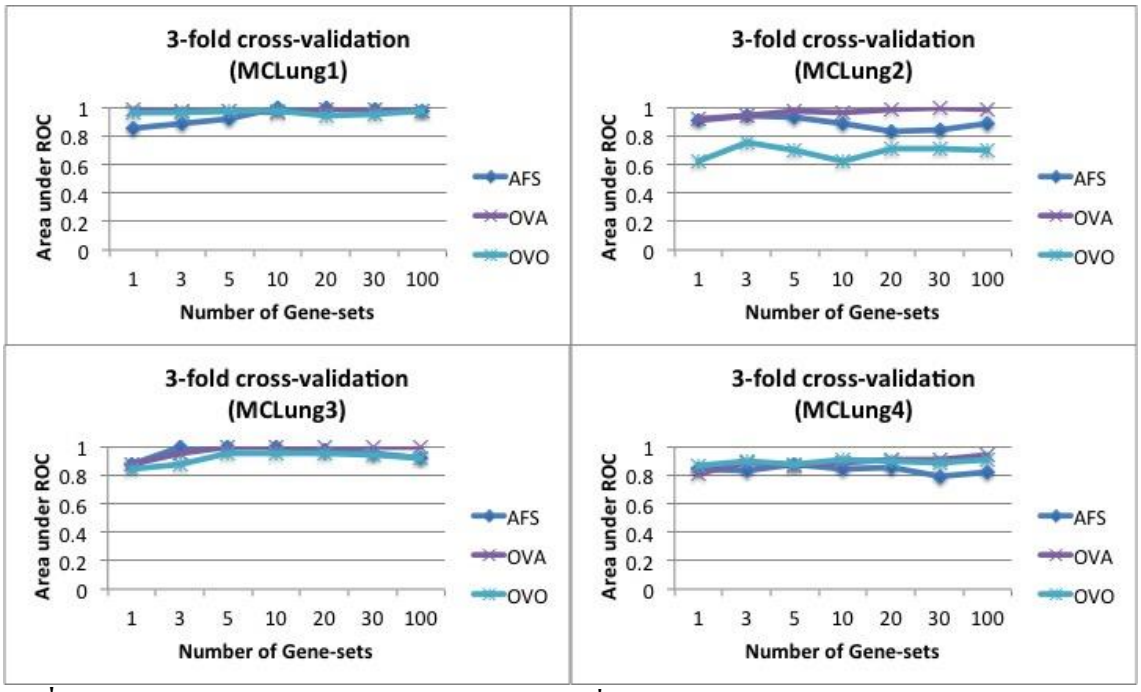
จากผลการทดลองที่ผ่านมาทำให้เราได้ AFS เป็นระเบียบวิธีการสำหรับอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลที่มีกลุ่มตัวอย่างหลายกลุ่ม ในงานวิจัยส่วนนี้ทางผู้วิจัยได้ทำระเบียบวิธีการ AFS เพื่อนำมาเปรียบเทียบกับการใช้แนวความคิดแบบ OVA และ OVO สำหรับต่อยอดการจำแนกกลุ่มตัวอย่างแบบสองกลุ่มไปสู่หลายกลุ่ม ในงานวิจัยส่วนนี้ได้ทดลองโดยใช้ชุดข้อมูล MCLung1-4 และใช้ Multilayer perceptron (MLP) และ Support Vector Machine (SVM) เป็นอัลกอริทึมในการจำแนกกลุ่มตัวอย่าง

3.2.3.1. ผลการเปรียบเทียบจากการทำ Stratified three-fold cross-validation

ผลการทดลองในส่วนนี้ได้นำเสนอในรูปแบบของกราฟดังรูปที่ 3.5-3.6 โดยจากผลการทดลองพบว่าทั้งสามแนวทางให้ผลที่คล้ายคลึงกัน โดยทุกแนวทางให้ผลที่ดีที่สุดในชุดข้อมูล MCLung3 ซึ่งมีจำนวนกลุ่มเพียงแค่สามกลุ่ม และเมื่อเปรียบเทียบอัลกอริทึมที่ใช้ในการจำแนกแล้ว SVM และ MLP ก็ยังให้ผลที่คล้ายคลึงกัน แต่หากดูความเสถียรของแบบจำลองต่อจำนวนของพาธเวย์ที่ใช้ในการสร้างแบบจำลองแล้ว MLP นั้นมีความเสถียรมากกว่า SVM



รูปที่ 3.5 ผลจากการทำ Three-fold cross-validation เมื่อใช้ SVM เป็นอัลกอริทึมสำหรับจำแนกกลุ่มตัวอย่าง

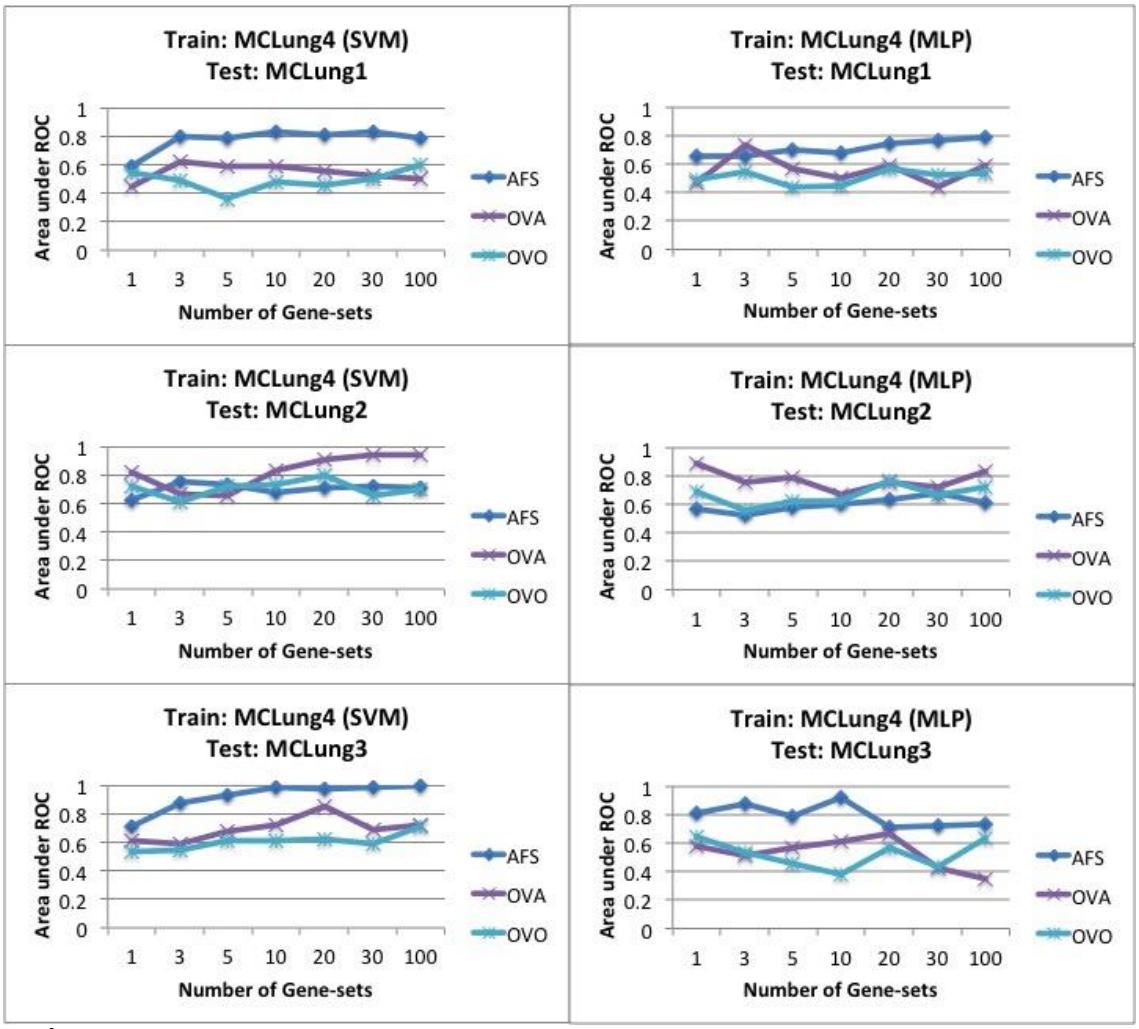


รูปที่ 3.6 ผลจากการทำ Three-fold cross-validation เมื่อใช้ MLP เป็นอัลกอริทึมสำหรับจำแนกกลุ่มตัวอย่าง

3.2.3.2. ผลการทำ Cross-dataset validation

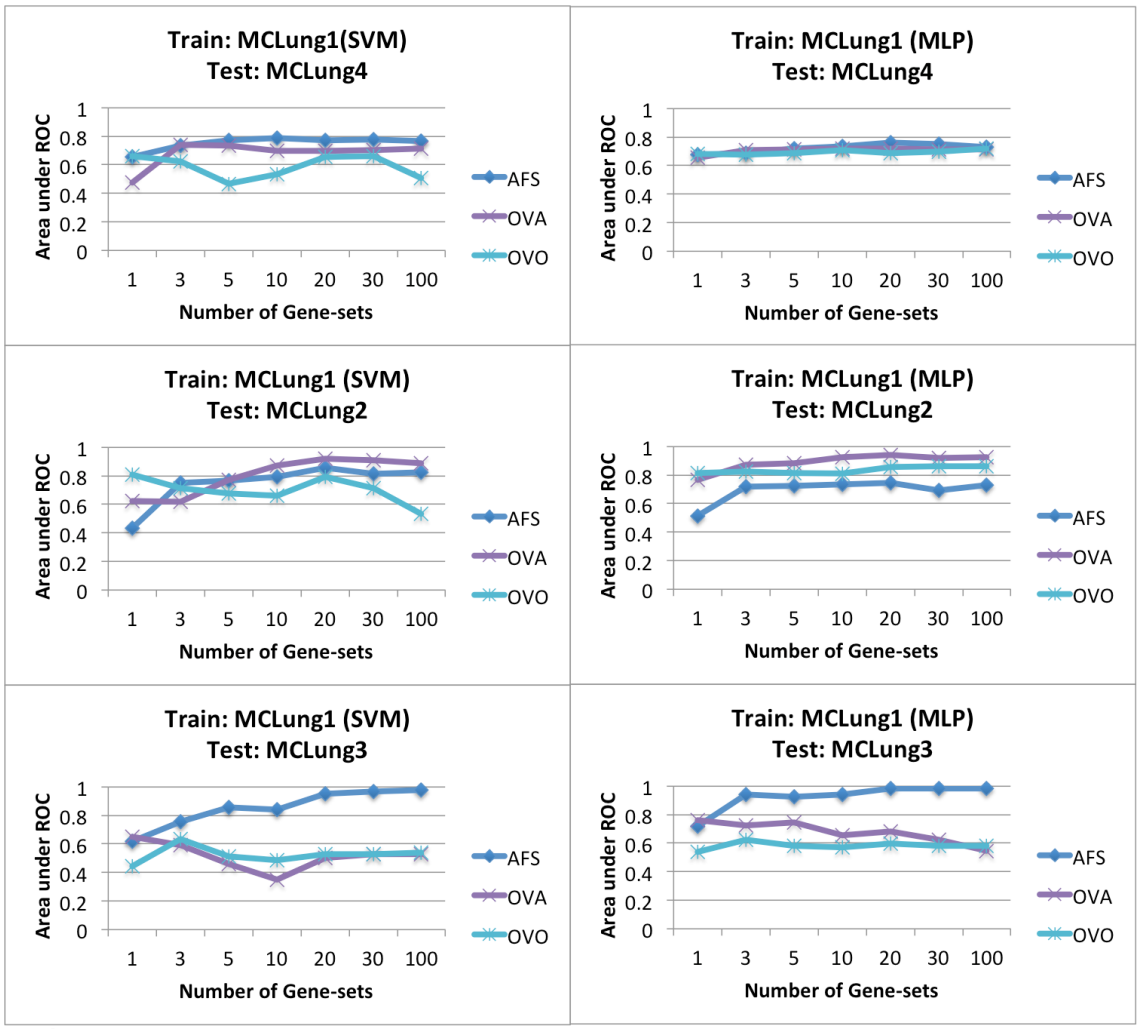
ในงานวิจัยส่วนนี้เป็นการวัดความถูกต้องและความเสถียรของแต่ละระเบียบวิธี โดยจะใช้ชุดข้อมูลนึงสร้างแบบจำลองและตรวจวัดความถูกต้องในอีกชุดข้อมูลนึง โดยในการศึกษานี้เนื่องจาก MCLung3 มีจำนวนกลุ่มตัวอย่างเพียงแค่ 3 กลุ่มตัวอย่าง ดังนั้นทุกชุดข้อมูลจะถูกนำไปสร้างแบบจำลอง ยกเว้นชุดข้อมูล MCLung3

ผลจากการสร้างแบบจำลองด้วยชุดข้อมูล MCLung4 พบว่า AFS ให้ผลที่ดีที่สุด (ผลเฉลี่ยของ AUROC = 0.75) ตามมาด้วย OVA และ OVO ตามลำดับ (รูปที่ 3.7)



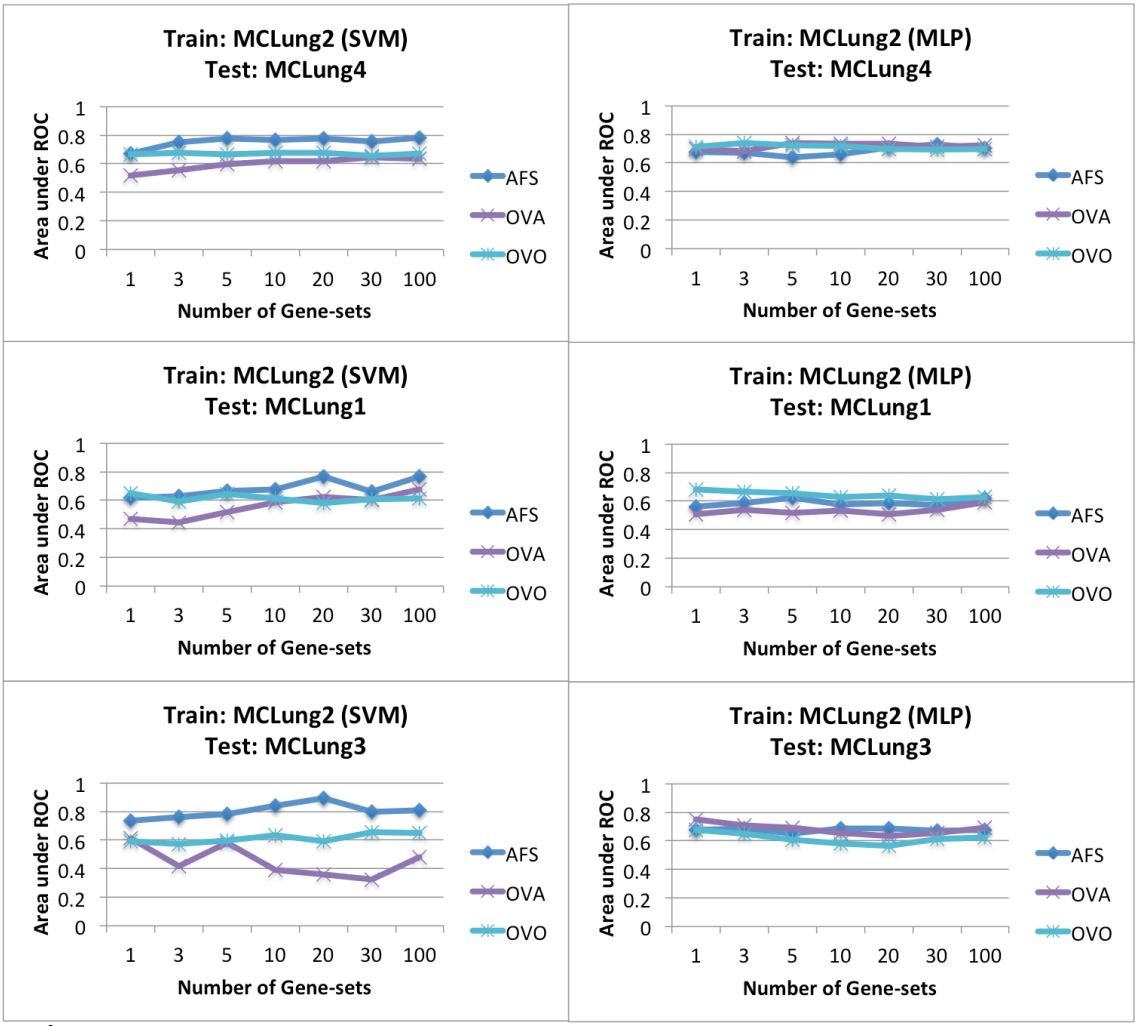
รูปที่ 3.7 ผลของ Cross-dataset validation โดยการใช้ MCLung4 สร้างแบบจำลอง และตรวจวัดความถูกต้องด้วยชุดข้อมูลอื่นๆ

ในส่วนของแบบจำลอง MCLung1 ซึ่งมีความไม่เท่ากันของสัดส่วนระหว่างประเภทย่อยของโรคมะเร็งปอด พบว่า AFS ยังคงให้ผลที่ดีที่สุด (ผลเฉลี่ย AUROC = 0.78) ตามด้วย OVA และ OVO ตามลำดับ (รูปที่ 3.8)



รูปที่ 3.8 ผลของ Cross-dataset validation โดยการใช้ MCLung1 สร้างแบบจำลอง และตรวจวัดความถูกต้องด้วยชุดข้อมูลอื่นๆ

ในส่วนสุดท้ายเป็นการสร้างแบบจำลองด้วยชุดข้อมูลที่มีความไม่เท่ากันระหว่างสัดส่วนของสองกลุ่มประเภทย่อยของโรคมะเร็ง (MCLung2) พบว่า AFS ยังคงให้ผลที่ดีที่สุด (ผลเฉลี่ย AUROC = 0.7) ตามด้วย OVO และ OVA ตามลำดับ (รูปที่ 3.9)



รูปที่ 3.9 ผลของ Cross-dataset validation โดยการใช้ MCLung2 สร้างแบบจำลอง และตรวจวัดความถูกต้องด้วยชุดข้อมูลอื่นๆ

จากตารางที่ 3.4-3.5 แสดงผลภาพรวมของการทำ Cross-dataset validation ซึ่งแสดงให้เห็นว่า AFS นั้นมีประสิทธิภาพในผลการทดสอบส่วนใหญ่ โดยหากเปรียบเทียบอัลกอริทึมในการจำแนกกลุ่มตัวอย่างพบว่าโดยรวมแล้ว SVM นั้นมีความเสถียรที่มากกว่าเล็กน้อยเมื่อใช้ AFS เป็นระเบียบวิธีการในการอ้างอิงระดับการทำงานของพาธเวย์

ตารางที่ 3.4 ผลจาก SVM (ผลเฉลี่ย AUROC ของแต่ละระเบียบวิธี)

Train	Test	AFS	OVA	OVO
MCLung4	MCLung1	0.78±0.76	0.55±0.06	0.49±0.07
MCLung4	MCLung2	0.71±0.04	0.82±0.11	0.71±0.06
MCLung4	MCLung3	0.92±0.1	0.7±0.08	0.6±0.05
MCLung1	MCLung4	0.75±0.04	0.68±0.09	0.59±0.08
MCLung1	MCLung2	0.75±0.13	0.8± 0.12	0.7±0.08
MCLung1	MCLung3	0.85±0.12	0.52±0.09	0.53±0.05
MCLung2	MCLung4	0.75±0.04	0.6±0.04	0.67±0.01
MCLung2	MCLung1	0.68±0.06	0.56±0.08	0.61±0.02
MCLung2	MCLung3	0.8±0.05	0.45±0.1	0.61±0.03
Average		0.78±0.7	0.63±0.12	0.61±0.07

ตารางที่ 3.5 ผลจาก MLP (ผลเฉลี่ย AUROC ของแต่ละระเบียบวิธี)

Train	Test	AFS	OVA	OVO
MCLung4	MCLung1	0.71±0.05	0.56±0.09	0.5±0.04
MCLung4	MCLung2	0.6±0.04	0.77±0.07	0.66±0.07
MCLung4	MCLung3	0.79±0.07	0.53±0.1	0.52±0.09
MCLung1	MCLung4	0.72±0.03	0.71±0.02	0.7±0.01
MCLung1	MCLung2	0.7±0.08	0.89±0.06	0.83±0.02
MCLung1	MCLung3	0.93±0.09	0.68±0.07	0.58±0.02
MCLung2	MCLung4	0.75±0.03	0.6±0.04	0.67±0.01
MCLung2	MCLung1	0.59±0.02	0.53±0.03	0.64±0.02
MCLung2	MCLung3	0.68±0.01	0.7±0.04	0.62±0.04
Average		0.72±0.1	0.66±0.11	0.64±0.11

เมื่อเปรียบเทียบแบบจำลองในการทดลองนี้พบว่า แบบจำลองที่สร้างโดยชุดข้อมูล MCLung2 นั้นให้ประสิทธิภาพที่ต่ำที่สุด โดยชุดข้อมูลนี้มีความไม่เท่ากันระหว่างสัดส่วนของกลุ่มตัวอย่างประเภทย่อยของโรคมะเร็งปอด AC และ SCC (ประมาณ 1:3) ทั้งนี้ นี่อาจจะเป็นสาเหตุที่ทำให้แบบจำลองดังกล่าวให้ผลที่ต่ำกว่าแบบจำลองอื่นๆ

ในส่วนของแบบจำลอง MCLung4 นั้นถูกคาดหวังว่าน่าจะมีประสิทธิภาพที่ดีที่สุดเนื่องจากเป็นชุดข้อมูลที่มีขนาดใหญ่ที่สุด และมีความเท่าเทียมกันระหว่างกลุ่มตัวอย่าง อย่างไรก็ตามชุดข้อมูลนี้มีการผสมปนเประหว่างกลุ่มตัวอย่างจากเชื้อชาติต่างๆกัน ซึ่งความแตกต่างทางเชื้อชาตินี้ก็ส่งผลให้มีความแตกต่างและความแปรปรวนในชุดข้อมูลด้วยเช่นกัน (Watskin, 2003)

OVA และ OVO ให้ประสิทธภาพที่แย่กว่า AFS ทั้งนี้อาจจะเป็นเพราะว่า OVA และ OVO นั้นมีความสร้างแบบจำลองหลายแบบจำลอง และแต่ละแบบจำลองก็สร้างจากชุดข้อมูลเพียงแค่บางส่วนเท่านั้นทำให้ได้ผลที่ไม่ครอบคลุมทั้งหมด ต่างจาก AFS ที่สร้างแบบจำลองเพียงแค่แบบจำลองเดียว และเป็นการสร้างโดยใช้ข้อมูลที่มีทั้งหมด นอกจากนี้ความซับซ้อนของ OVA และ OVO นั้นยังมีมากกว่า AFS อีกด้วย

จากผลการวิจัยนี้พบว่าขนาดของ PCOGs ที่พบในระเบียบวิธี AFS นั้นมีขนาดที่น้อยกว่า NCFS-i (ตารางที่ 3.6) ซึ่งอาจจะส่งผลให้ประสิทธิภาพในการจำแนกกลุ่มนั้นน้อยลง เพราะว่ามีจำนวนยีนไม่กี่ตัวที่ถูกนำมาใช้เพื่ออ้างอิงถึงระดับการทำงานของพาธเวย์ทั้งพาธเวย์ ดังนั้นหากต้องการจะเพิ่มประสิทธิภาพการทำงานของ AFS การพัฒนาในส่วนของการหา PCOGs จึงเป็นส่วนที่ต้องการการพัฒนามากที่สุด โดย ณ ขณะนี้ AFS ได้ใช้การค้นหาแบบ Greedy ในการหา PCOGs ดังนั้นการใช้อัลกอริทึมการค้นหาแบบอื่นๆเข้ามาช่วยจึงเป็นหนทางในการพัฒนา AFS ที่น่าสนใจ

ตารางที่ 3.6 เปรียบเทียบขนาดของ PCOGs ระหว่าง AFS และ NCFS-i

ชุดข้อมูล	ขนาดของ PCOGs	
	AFS	NCFS-i
MCLung4	3.86±2.13	7.23±3.68
MCLung1	3.12±2	6.86±3.54
MCLung2	3.28±1.83	6.54±3.31

นอกจากนี้เรายังทำการวิเคราะห์ผลจาก PCOGs ที่ระบุโดย AFS เพิ่มเติมเพื่อดูความคล้ายคลึงกันในแต่ละแบบจำลอง ทางผลการวิเคราะห์พบว่า ในพาธเวย์ SCLC และ NSCLC พาธเวย์ซึ่งเป็นพาธเวย์ที่เกี่ยวข้องกับโรคมะเร็งปอดนั้นมีการพบยีน CDK4/ PIK3CA ใน PCOGs ร่วมกันระหว่างแบบจำลอง โดย PIK3CA และ CDK4/6 มีหน้าที่เกี่ยวของกับการยับยั้งการทำงานของระบบ apoptosis และโดยเฉพาะ CDK4/6 นั้นยังมีหน้าที่ที่เกี่ยวข้องกับการเร่งการทำงานในส่วน G1/S progression (Kanehisa, 2000; Kanehisa 2012) ซึ่งจากการวิเคราะห์ PCOGs เพิ่มเติมนี้แสดงให้เห็นว่า AFS นั้นสามารถใช้มาอ้างอิงระดับการทำงานของพาธเวย์

ได้อย่างน่าเชื่อถือ อีกทั้งข้อมูลระดับการทำงานของพาธเวย์นั้นยังสามารถนำไปใช้จำแนกกลุ่มผู้ป่วยได้อย่างถูกต้องอีกด้วย

3.3 การวิเคราะห์ข้อมูลอะเรย์ข้ามแพลทฟอร์ม

ในการศึกษานี้มีการทดสอบประสิทธิภาพของการจำแนกกลุ่มตัวอย่างด้วยการวิเคราะห์ข้อมูลไมโครอะเรย์ร่วมกับข้อมูลพาธเวย์แบบข้ามแพลทฟอร์ม เนื่องด้วยการวิเคราะห์ไมโครอะเรย์ข้ามแพลทฟอร์มมักจะเกิดปัญหาการมีข้อมูลสูญหาย (Missing data) ดังนั้นในงานนี้จึงมีการปรับใช้ Bayesian-based missing data imputation (BPCA) เพื่อเข้ามาช่วยประมาณค่าข้อมูลที่หายไประหว่างการทำการวิเคราะห์ข้อมูลข้ามแพลทฟอร์ม ทั้งนี้ในการศึกษานี้ได้เปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มตัวอย่างโดยการสร้างแบบจำลองจากชุดข้อมูลที่ได้จากแพลทฟอร์มนึง และตรวจวัดความถูกต้องโดยใช้ชุดข้อมูลจากอีกแพลทฟอร์มนึง โดยเปรียบเทียบระหว่างการทำการประมาณค่าข้อมูลที่สูญหายกับการใช้ข้อมูลดังกล่าวทั้งๆที่ยังสูญหาย เพื่อใช้ในการพัฒนาการจำแนกกลุ่มตัวอย่างเช่นนี้ต่อไป โดยในการศึกษานี้ได้ทดสอบบนชุดข้อมูลมะเร็งปอดสองชุด และชุดข้อมูลมะเร็งเต้านมอีกสองชุด

3.3.1 ผลของการทำ Imputation

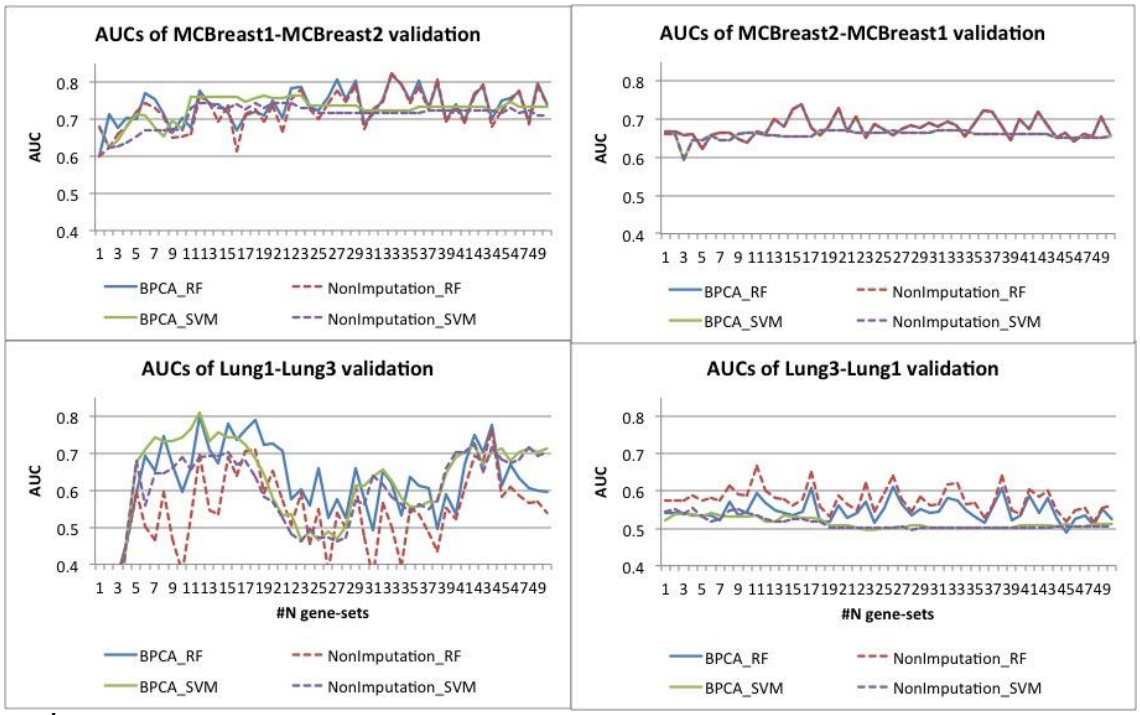
ก่อนที่จะทำการจำแนกกลุ่มตัวอย่าง การทำการประมาณค่าข้อมูลที่สูญหาย (Missing data imputation) ด้วยวิธี BPCA เป็นขั้นตอนที่สำคัญเพื่อช่วยเพิ่มประสิทธิภาพของการจำแนกกลุ่มตัวอย่าง ทั้งนี้ในการศึกษาขั้นนี้ได้มีการเปรียบเทียบ Discriminative score ของชุดข้อมูลทดสอบที่ทำการ Imputation และไม่ได้ทำ ซึ่งในชุดข้อมูลที่มีหลายกลุ่มตัวอย่างจะคำนวน Discriminative score โดยการอ้างอิงค่า F-value จาก ANOVA ในขณะนี้ชุดข้อมูลที่จำแนกกลุ่มตัวอย่างสองกลุ่มจะใช้การอ้างอิงค่า T-score จาก Student *t*-test ตารางที่ 3.7 แสดงให้เห็นว่าในชุดข้อมูลส่วนมาก การทำ Imputation นั้นจะช่วยเพิ่มค่า Discriminative Score ยกเว้นในชุดข้อมูล Lung1 ซึ่งเป็นชุดข้อมูลที่มีการสูญหายของข้อมูลมาก (ประมาณ 40%)

ตารางที่ 3.7 ผลสรุปของ Discrimative score เปรียบเทียบระหว่างการทำ Imputation และไม่ทำ

Dataset	Average ± Standard deviation		P-value of One way paired *t*-test
	BPCA	Non-Imputation	
MCBreast1	*9.35±7.37	9.22±7.23	1.0E-06
MCBreast2	*13.8±11.7	13.6±11.5	0.02
Lung1	1.24±0.95	*1.28±0.98	2.0E-04
Lung3	*4.21±3.05	4.11±3.02	5.0E-05

3.3.2 ผลของการทำ Cross-platform validation

ในการศึกษาขั้นนี้มีความประสงค์ที่จะประเมินประสิทธิภาพของการทำ Imputation และการไม่ทับ Imputation บนข้อมูลระดับการทำงานของพาธเวย์ที่ได้จากระเบียบวิธี NCFS-i และ AFS โดย ReliefF ถูกใช้เป็นระเบียบวิธีการสำหรับการเลือกพาธเวย์จำนวน 1%, 5%, 10%, 20%, และ 50% พาธเวย์ที่ดีที่สุดมาใช้ในการสร้างแบบจำลอง โดยมีการใช้ RF และ SVM เป็นอัลกอริทึมในการจำแนกกลุ่มตัวอย่าง ผลการเปรียบเทียบถูกนำเสนอในรูปที่ 3.10 ซึ่งความแตกต่างระหว่างสองจะถูกประเมินทางสถิติด้วยการทำ one-way paired *t*-test การผลการศึกษาครั้งนี้แสดงให้เห็นว่าไม่มีความแตกต่างอย่างมีนัยสำคัญในการตรวจวัดผลของชุดข้อมูล MCBreast2-MCBreast1 เนื่องจากระดับการสูญหายของข้อมูลนั้นมีน้อยมาก (ประมาณ 4%) ในส่วนของการตรวจวัดความถูกต้องของ MCBreast1-MCBreast2 และ Lung1-Lung2 ผลการทดลองแสดงให้เห็นว่าการใช้ BPCA เข้ามาประมาณค่าข้อมูลที่สูญหายช่วยเพิ่มประสิทธิภาพของการจำแนกกลุ่มตัวอย่างอย่างมีนัยสำคัญทั้งใน RF (p-value = 1.9E-05 และ 1.5E-13) และ SVM (p-value = 8.6E-12 และ 2.1E-06) อย่างไรก็ตามในการตรวจวัดความถูกต้องในชุดข้อมูล Lung3-Lung1 validation ที่มีความสูญหายข้อมูลมาก ซึ่งผลปรากฏว่าการทำ Imputation จะทำให้ประสิทธิภาพโดยรวมลดลงอย่างมีนัยสำคัญ ใน RF (p-value = 4.9E-19) ทั้งนี้เนื่องด้วยข้อมูลมีการสูญหายจำนวนมากทำให้ประสิทธิภาพการจำแนกโดยรวมนั้นต่ำ (ผลเฉลี่ย AUC = ~0.53).



รูปที่ 3.10 กราฟแสดงประสิทธิภาพการจำแนกกลุ่มตัวอย่างเปรียบเทียบระหว่างการทำ Imputation และไม่ทำ Imputation

งานวิจัยในขั้นนี้แสดงให้เห็นว่าการทำ Imputation นั้นจะช่วยเพิ่มประสิทธิภาพของการจำแนกข้อมูลที่ใช้ข้อมูลต่างแพลทฟอร์มกันระหว่างชุดสร้างแบบจำลองและชุดทดสอบ ทั้งนี้หากระดับความสูญหายของข้อมูลสูงการศึกษานี้ไม่แนะนำให้ทำการวิเคราะห์ดังกล่าวเนื่องจากจะให้ประสิทธิภาพที่ต่ำกว่าที่ควร

บทที่ 4

สรุปและเสนอแนะ

4.1 สรุปผลการดำเนินงาน

การศึกษาโรคทางพันธุกรรมที่มีความซับซ้อนสูงนั้นจำเป็นต้องใช้กรรมวิธีแบบบูรณาการที่มีการวิเคราะห์ข้อมูลหลายๆส่วนควบคู่กันไปเพื่อช่วยให้เข้าใจกลไกการเกิดโรคนั้นๆได้ดีขึ้น ดังเช่นในงานวิจัยนี้ที่ได้มีการวิเคราะห์ข้อมูลไมโครอะเรย์ควบคู่กับข้อมูลพาธเวย์เพื่อช่วยในการจำแนกโรค โดยงานวิจัยนี้ได้นำเสนอระเบียบวิธีการอ้างอิงระดับการทำงานของพาธเวย์ในชุดข้อมูลที่มีกลุ่มตัวอย่างมากกว่าสองกลุ่ม ซึ่งไม่เคยมีงานวิจัยไหนเคยทำมาก่อน งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพการจำแนกกลุ่มตัวอย่างหลายกลุ่มด้วยระเบียบวิธีการหลายๆวิธีการประกอบไปด้วย AFS, R-NCFS, OVA และ OVO ซึ่งผลการทดลองก็แสดงให้เห็นว่าการใช้ AFS นั้นให้ผลที่ดีที่สุด ทั้งนี้งานวิจัยนี้ยังได้ทำการทดสอบประสิทธิภาพของทั้ง AFS และ NCFS-i ในการจำแนกกลุ่มตัวอย่างโดยการใช้ข้อมูลต่างแพลทฟอร์มกันอีกด้วย

ทั้งนี้ทางกลุ่มวิจัยยังได้พัฒนาระเบียบวิธีการดังกล่างขึ้นเป็นโปรแกรมที่อนุญาตให้ผู้ใช้เข้าใช้ได้อย่างเสรี ซึ่งโปรแกรมสามารถเข้าใช้ได้ผ่านระบบออนไลน์ ที่มีเครื่องมือหลากหลายทั้งการอ้างอิงระดับการทำงานของพาธเวย์ และการจำแนกกลุ่มตัวอย่างโดยอัลกอริทึมที่มีใน WEKA อีกทั้งยังมีการเชื่อมโยงผลลัพธ์ไปยัง KEGG Pathway เพื่อช่วยในการวิเคราะห์ให้เข้าใจกลไกของการเกิดโรคมากยิ่งขึ้น

4.2 ข้อเสนอแนะ

ในการปรับปรุงประสิทธิภาพของระเบียบวิธีการที่นำเสนอในงานวิจัยชิ้นนี้นั้น สามารถทำได้ในหลายๆส่วน โดยหลักๆแล้ว การพัฒนาระเบียบวิธีการค้นหา PCOGs โดยการใช้อัลกอริทึมอื่นๆนอกเหนือจาก Greedy อัลกอริทึมเป็นส่วนสำคัญที่สุดที่จะช่วยปรับปรุงประสิทธิภาพของระเบียบวิธี AFS ได้ โดย Greedy อัลกอริทึมนั้นจะให้ผลที่เรียกว่า Local optimal ซึ่งแนวทางการแก้ปัญหาที่ดีแต่ไม่ดีที่สุด ในขณะที่การใช้อัลกอริทึมอื่นที่อาจจะให้ผลแบบ Glocal optimal จะใช้เวลาในการคำนวนมาก อย่างไรก็ตามหากพัฒนาอัลกอริทึมดังกล่าวแบบ Parallel ก็จะช่วยย่นระยะเวลาในการคำนวนได้มากเช่นกัน

เอกสารอ้างอิง

(References)

Adi, L. T., Roberto, R. and Sorin, D., 2006, "Analysis of microarray experiments of gene expression profiling", **American Journal of Obstetrics and Gynecology**, Vol. 195, pp. 373-388.

Bandyopadhyay, N., Kahveci, T., Goodison, S., Sun, Y., and Ranka, S., 2009, "Pathway-based feature selection algorithm for cancer microarray data", **Advance in Bioinformatics**.

Bellei, B., Pitisci, A., Izzo, E. and Picardo M., 2012, "Inhibition of Melanogenesis by the Pyridinyl Imidazole Class of Compounds: Possible Involvement of the Wnt/ß-Catenin Signaling Gene set", **PLoS ONE**, Vol. 7, e33021.

Chan, J. H., Sootanan, P. and Larpeampaisarl, P., 2011, "Feature selection of gene set markers for microarray-based disease classification using negatively correlated feature set", **IJCNN**, pp. 3293-3299.

Devaraj, S. and Natarajan, J., 2011, "miRNA-mRNA network detects hub mRNAs and cancer specific miRNAs in lung cancer", **In Silico Biology**, Vol. 11, pp. 281-95.

Edgar, R., Domrachev, M. and Lash, A. E., 2002, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". **Nucleic Acid Res**, Vol. 30, pp. 207-210.

Engchuan, W. and Chan, J. H., 2012, "Pathway-based multiclass classification of lung cancer", **In International Conference ICONIP 2012, Lecture Note in Computer Science**, Vol. 7667, pp. 697-702.

Fleige, S. and Pfaffl, M. W., 2006, "RNA Integrity and the effect on the real-time qRT-PCR performance", **Molecular aspects of Medicine**, Vol. 27, pp. 126-139.

Goutte, C. and Gaussier, E., 2005, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation", **In Advances in Information Retrieval**, pp. 345-359.

Guyon, I. and Elisseeff, A., 2003, "An introduction to variable and feature selection", **The Journal of Machine Learning Research**, Vol. 3, pp. 1157-1182.

Hanai, J., Doro, N., Sasaki, A. T., Kobayashi, S., Cantley, L. C., Seth, P. and Sukhatme V. P., 2012, "Inhibition of lung cancer growth: ATP citrate lyase knockdown and statin treatment leads to dual blockade of mitogen-activated protein kinase (MAPK) and phosphatidylinositol-3-kinase (PI3K)/AKT gene sets", **Journal of Cell Physiology**, Vol. 227, pp. 1709-20.

Hosgood, H. D., Menashe, I., Shen, M., Yeager, M., Yuenger, J., Rajaraman, P., He, X., Chatterjee, N., Caporaso, N. E., Zhu, Y., Chanock, S. J., Zheng, T. and Lan, Q., 2008, "Pathway-based evaluation of 380 candidate genes and lung cancer susceptibility suggests the importance of the cell cycle pathway", **Carcinogenesis**. Vol. 10, pp. 1938-1943.

Jane, J. L., Gene, C., Wuxiong, L., Zheng, P., Sihua, P., Tim, H., Liangbiao, C. and Xuefeng, B. L., 2004, "Muticlass cancer classification and biomarker discovery using GA-based algorithms", **Bioinformatics**, Vol. 21, pp. 2691-2697.

Kanehisa, M. and Goto, S., 2000, "KEGG: Kyoto Encyclopedia of Genes and Genomes", **Nucleic Acids Research**, Vol. 28, pp. 27-30.

Kanehisa, M., Goto, S., Sata, Y., Furumichi, M. and Tanabe, M., 2012, "KEGG for integration and interpretation of large-scale molecular database", **Nucleic Acids Research**, Vol. 40, pp. D109-D114.

Kononenko, I., 1994, "Estimating attributes: analysis and extensions of RELIEF", **In Machine Learning: ECML-94**, pp. 171-182.

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P., 2006, "Handling imbalanced dataset: A review", **GESTS International Transaction Computer Science and Engineering**, Vol. 30, pp. 25-36.

Lipshutz, R. J., Morris, D., Chee, M., Hubbell, E., Kozal, M. J., Shah, N. Shen, N., Yang, R. and Fodor, S. P., 1995, "Using Oligonucleotide Probe Arrays to Access Genetic Diversity", **Biotechniques**, Vol. 19, pp. 442-447.

Lukas, L., Devos, A., Suyken, J.A.K., Vanhamme, L., Howe, F.A., Majos, C., Moreno-Torres, A., Van Der Graff, M., Tate, A.R., Arus, C., and Van Huffel, S., 2004, "Brain tumor classification based on lung echo proton MRS signal", **Artificial Intelligence in Medicine**. Vol. 37, pp. 73-89.

Mazieres, J., He, B., You, L., Xu, Z. and Jablons, D. M., 2005, "Wnt signaling in lung cancer", **Cancer Letter**, Vol. 222, pp. 1-10.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I. and Ishii, S., 2003, "A Bayesian missing value estimation method for gene expression profile data", **Bioinformatics**, Vol. 19(16), pp. 2088-2096.

Orihuela, C. J., Radin, J. N., Sublett, J. E., Gao G., Kaushal D. and Toumanen E. I., 2004,"Microarray analysis of pneumococcal gene expression during invasice disease", **Infect Immune**, Vol.10, pp. 5582-5596.

Pang, H. and Zhao, H., 2008, "Building gene set cluster from Random Forest classification using class votes", **BMC Bioinformatics**, Vol. 9, pp. 87.

Robnik-Šikonja, M. and Kononenko, I., 2003, "Theoretical and empirical analysis of ReliefF and RReliefF", **Machine learning**, Vol. 53(1-2), pp. 23-69.

Sootanan, P., Prom-on, S., Meechai, A. and Chan, J. H., 2012, "Pathway-based microarray analysis for robust disease classification", **Neural Computing & Applications**, Vol. 21, pp. 649-660.

Sridhar, R., Pablo, T., Ryan, R., Sayan, M., Chen-Hsiang, Y., Michael, A., Christine, L., Michael, R., Eva, L., Jil, P. M., Tomaso, P., William, G., Massimo, L. and Todd, R. G., 2001,"Multiclass cancer diagnosis using tumor gene expression signatures", **PNAS**, Vol. 98, pp. 15149-15154.

Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J., 2007, "pcaMethods—a bioconductor package providing PCA methods for incomplete data", **Bioinformatics**, Vol. 23(9), pp. 1164- 1167.

Stelios, P. P., Annette, M. P. and Stephen, M. S., 2011,"Multi-membership gene regulation in pathway based microarray analysis", **Algorithms for Molecular Biology**, Vol. 6.

Tsurutani, J., West, K.A., Sayvah, J., Gills, J. J. and Dennis, P. A., 2005, "Inhibition of the phosphatidylinositol 3-kinase/Akt/mammalian target of rapamycin gene set but not the MEK/ERK gene set attenuates laminin-mediated small cell lung cancer cellular survival and resistance to imatinib mesylate or chemotherapy", **Cancer Research**, Vol. 65, pp. 8423-32.

Van Dyke, A. L., Cote, M. L., Wenzlaff, A. S., Chen, W., Abrams, J., Land, S., Giroux, C. N. and Schwatz, A. G., 2009, "Cytokine and cytokine receptor single-nucleotide polymorphisms predict risk for non-small cell lung cancer among women", **Cancer Epidemology, Biomarker and Prevention**, Vol. 18, pp. 1829-1840.

Wang, Y. and Makedon, F., 2004, "Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data", **In Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE**, pp. 497-498.

Watskin, W. S., Rogers, A. R. and Jorde L. B., 2003, "Genetic Variation Among World Populations: Inferences From 100 Alu insertion polymorphisms", **Genome Research**, Vol. 13, pp. 1607-1618.

Wang, Y., Chen, J., Li, Q., Wang, H., Liu, G., Jing, Q. and Shen, B., 2011, "Indentifying novel prostate cancer associated pathways based on integrative microarray data analysis", **Computational Biology and Chemistry**, Vol. 35, pp. 151-158.

Westhoff, B., Colaluca, I. N., D'Ario, G., Donzelli, M., Tosoni, D., Volorio, S., Pelosi, G., Spaggiari, L., Mazzarol, G., Viale, G., Pece, S. and Di Fiore, P. P., 2009, "Alterations of the Notch gene set in lung cancer", **Proceedings of the National Academy of Science**, Vol. 106, pp. 22293-9.

Xu, L., Tan, A. C., Naiman, D. Q., Geman, D. and Winslow, R. L., 2005, "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data", **Bioinformatics**, Vol. 21(20), pp. 3905-3911.

Yang, H., Zhang, Q., He, J. and Lu, W., 2010, "Regulation of calcium signaling in lung cancer", **Journal of Thoracic Disease**, Vol. 2, pp. 52-56.

Yang, S. and Naiman, D. Q., 2014, "Multiclass cancer classification based on gene expression comparison", **Statistical Applications in Genetics and Molecular Biology**, Vol. 13, pp. 477-496.

Yu, H. L., Gao, S., Qin, B. and Zhao, J., 2012,"Multiclass microarray data classification based on confidence evaluation", **Genetics and Molecular research**, Vol. 11, pp. 1357-1369.

ผลงานตีพิมพ์

1. Engchuan, W. and Chan, J. H., 2013, "Apriori Gene Set-based Microarray Analysis for Disease Classification Using Unlabeled Data", **Proceeding of the 4th International Conference on Computational Systems-Biology and Bioinformatics (CSBio2013)**, Procedia Computer Science, Vol. 23, pp. 137-145.
2. Engchuan, W. and Chan, J. H., 2014, "Pathway Activity Transformation for Multi-class Classification of Lung Cancer Datasets", **Neurocomputing**, (Accepted for publication).
3. Engchuan, W., Tongsima, S., Meechai, A. and Chan, J. H., 2014, "Cross-platform Pathway Activity Transformation and Classification of Microarray Data", **Proceeding of the 4th International Neural Network Society Winter Conference (INNS-CIIS 2014)**, Advances in Intelligent Systems and Computing, Vol. 331, pp. 139-148.
4. Engchuan, W., Meechai A., Tongsima S. and Chan J. H., 2014, "Microarray-based Cancer Diagnosis using An Integrative Gene-set Analysis Approach", (Under consideration to be published in Journal of Bioinformatics and Computational Biology).
5. Phongwattana, T., Engchuan, W. and Chan, J. H., 2015, "Clustering-based Multi-class classification of complex disease", **Proceeding of the 6th International Conference on Knowledge and Smart Technology (KST 2015)**, IEEE (In press).

ลงชื่อ_____

(รศ.ดร.โจนาธาน โฮยิน ชาน)

6 มกราคม 2558