แบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรค

DNA Microarray Data Analysis Model using Clustering Algorithm for Disease Diagnosis

ศิริพันธุ์ วิชัยดิษฐ

Siriphan Wichaidit

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Prince of Songkla University
2551

ชื่อวิทยานิพนธ์ แบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรค
ผู้เขียน นางสาวศิริพันธุ์ วิชัยดิษฐ
สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	**คณะกรรมการสอบ**
.....	.....ประธานกรรมการ
(ดร.วิภาดา เวทย์ประสิทธิ์)	(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)
	.....กรรมการ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	(รองศาสตราจารย์ ดร.อมรรัตน์ พงศ์ดารา)
.....	.....กรรมการ
(ดร.ศิริรัตน์ วณิชโยบล)	(ดร.ศิริรัตน์ วณิชโยบล)
	.....กรรมการ
	(ดร.วิภาดา เวทย์ประสิทธิ์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

.....
(รองศาสตราจารย์ ดร.เกริกชัย ทองหนู)
คณบดีบัณฑิตวิทยาลัย

ชื่อวิทยานิพนธ์ แบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรค
ผู้เขียน นางสาวศิริพันธุ์ วิชัยดิษฐ
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2550

บทคัดย่อ

ข้อมูลด้านชีววิทยาเป็นข้อมูลที่มีประโยชน์อย่างมากในหลาย ๆด้าน เช่น การวินิจฉัยโรค การวิจัยด้านเภสัชศาสตร์ และการวิเคราะห์ความหลากหลายของสิ่งมีชีวิต ในปัจจุบันข้อมูลด้านชีววิทยามีปริมาณเพิ่มมากขึ้นอย่างรวดเร็ว ปัญหาที่สำคัญก็คือจะมีวิธีการในการบริหารจัดการกับข้อมูลจำนวนมหาศาลนี้ให้มีประโยชน์สูงสุดได้อย่างไร งานวิจัยนี้จึงได้ทำการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรคขึ้น คือแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction using Self-Organizing Map: DNA_MKESOM) ซึ่งพัฒนาโดยใช้โปรแกรม MATLAB แบบจำลองนี้จะทำการแบ่งกลุ่มข้อมูลโดยใช้แผนที่การจัดกลุ่มเอง ประกอบกับใช้ค่า P-Value และค่า Highest Certainty Value (HCV) ในการเลือกยีน เพื่อลดเวลาในการทำงานลง แบบจำลองที่พัฒนาขึ้นประกอบด้วยขั้นตอนหลัก 6 ขั้นตอนคือ 1) ขั้นตอนการเตรียมข้อมูล 2) ขั้นตอนการเลือกยีนโดยใช้ค่า P-Value 3) ขั้นตอนการแบ่งกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง 4) ขั้นตอนการเลือกยีนโดยใช้ค่า Highest Certainty Value 5) ขั้นตอนการสร้างกฎ และ 6) ขั้นตอนการคำนวณค่าความถูกต้อง ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลดีเอ็นเอไมโครอาเรย์ของ โรคมะเร็งในเนื้องอก โรคมะเร็งเม็ดเลือดขาว โรคมะเร็งปอด และโรคมะเร็งลำไส้ ซึ่งแบบจำลองที่พัฒนาขึ้นสามารถทำนายประเภทของโรคมะเร็งโดยให้ค่าความถูกต้องสูง และสามารถลดเวลาในการทำงานลงได้อย่างมาก

ชื่อวิทยานิพนธ์ แบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรค
ผู้เขียน นางสาวศิริพันธุ์ วิชัยดิษฐ
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2550

บทคัดย่อ

ข้อมูลด้านชีววิทยาเป็นข้อมูลที่มีประโยชน์อย่างมากในหลาย ๆด้าน เช่น การวินิจฉัยโรค การวิจัยด้านเภสัชศาสตร์ และการวิเคราะห์ความหลากหลายของสิ่งมีชีวิต ในปัจจุบันข้อมูลด้านชีววิทยามีปริมาณเพิ่มมากขึ้นอย่างรวดเร็ว ปัญหาที่สำคัญก็คือจะมีวิธีการในการบริหารจัดการกับข้อมูลจำนวนมหาศาลนี้ให้มีประโยชน์สูงสุดได้อย่างไร งานวิจัยนี้จึงได้ทำการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรคขึ้น คือแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction using Self-Organizing Map: DNA_MKESOM) ซึ่งพัฒนาโดยใช้โปรแกรม MATLAB แบบจำลองนี้จะทำการแบ่งกลุ่มข้อมูลโดยใช้แผนที่การจัดกลุ่มเอง ประกอบกับใช้ค่า P-Value และค่า Highest Certainty Value (HCV) ในการเลือกยีน เพื่อลดเวลาในการทำงานลง แบบจำลองที่พัฒนาขึ้นประกอบด้วยขั้นตอนหลัก 6 ขั้นตอนคือ 1) ขั้นตอนการเตรียมข้อมูล 2) ขั้นตอนการเลือกยีนโดยใช้ค่า P-Value 3) ขั้นตอนการแบ่งกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง 4) ขั้นตอนการเลือกยีนโดยใช้ค่า Highest Certainty Value 5) ขั้นตอนการสร้างกฎ และ 6) ขั้นตอนการคำนวณค่าความถูกต้อง ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลดีเอ็นเอไมโครอาเรย์ของ โรคมะเร็งในเนื้องอก โรคมะเร็งเม็ดเลือดขาว โรคมะเร็งปอด และโรคมะเร็งลำไส้ ซึ่งแบบจำลองที่พัฒนาขึ้นสามารถทำนายประเภทของโรคมะเร็งโดยให้ค่าความถูกต้องสูง และสามารถลดเวลาในการทำงานลงได้อย่างมาก

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี ด้วยความช่วยเหลือจากบุคคลหลายฝ่าย ซึ่งผู้วิจัยรู้สึกซาบซึ้งสำหรับความช่วยเหลือเป็นอย่างยิ่งและขอกราบขอบพระคุณอย่างสูงสำหรับ

ดร. วิภาดา เวทย์ประสิทธิ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาสละเวลาส่วนตัวเพื่อให้คำปรึกษาและช่วยเหลือในการแก้ไขปัญหาต่างๆ ให้แก่ผู้วิจัยเสมอมา พร้อมทั้งช่วยตรวจทานและแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัยจนงานวิจัยนี้เสร็จสมบูรณ์ได้

ดร. ศิริรัตน์ วณิชโยบล อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้คำปรึกษาข้อเสนอแนะต่างๆ ในการเขียนผลงานวิจัย รวมทั้งตรวจทานแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัย

ผศ.ดร. กฤษณะ ชินสาร และ รศ.ดร. อมรรัตน์ พงศ์ดารา กรรมการในการสอบวิทยานิพนธ์ ที่กรุณาสละเวลาช่วยตรวจทานวิทยานิพนธ์ให้มีความสมบูรณ์

ดร. ประวัติ เวทย์ประสิทธิ์ ที่กรุณาช่วยตรวจทานเอกสารผลงานวิจัย ทำให้เอกสารผลงานวิจัยมีความถูกต้อง

อาจารย์ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ทุกท่าน ที่ได้ให้ความรู้ทางด้านวิชาการ ซึ่งส่งผลให้สามารถนำมาใช้ในการทำวิทยานิพนธ์ได้เป็นอย่างดี

เจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์ และเจ้าหน้าที่บัณฑิตวิทยาลัย ทุกท่านที่ให้ความช่วยเหลือและอำนวยความสะดวกเกี่ยวกับเอกสารต่างๆ และการเบิกจ่ายวัสดุที่ใช้ในงานวิจัย

เพื่อน ๆ และพี่ ๆ ภาควิชาวิทยาการคอมพิวเตอร์ ที่เป็นกำลังใจ และช่วยเหลือในการให้คำปรึกษาในการทำวิทยานิพนธ์ด้วยดีเสมอมา

คุณพ่อ คุณแม่ น้องชาย คุณปู่ คุณย่า ที่ห่วงใยคอยให้กำลังใจ รวมทั้งให้การสนับสนุนในการทำวิทยานิพนธ์แก่ผู้วิจัยเสมอมา

ผู้วิจัยขอขอบคุณทุกท่านเป็นอย่างสูงมา ณ โอกาสนี้ด้วย

ศิริพันธุ์ วิชัยดิษฐ

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

รายการตาราง

รายการตาราง (ต่อ)

รายการตาราง (ต่อ)

รายการตาราง (ต่อ)

รายการภาพประกอบ

รายการภาพประกอบ (ต่ อ)

รายการภาพประกอบ (ต่ อ)

บทที่ 1

บทนำ

ข้อมูลด้านชีววิทยาเป็นข้อมูลที่มีประโยชน์อย่างมากในหลายๆ ด้าน เช่น การวินิจฉัยโรค การวิจัยด้านเภสัชศาสตร์ และการวิเคราะห์ความหลากหลายของสิ่งมีชีวิต เป็นต้น ซึ่งปัจจุบันข้อมูลด้านชีววิทยามีปริมาณเพิ่มขึ้นอย่างมากมายและรวดเร็ว ส่วนหนึ่งเป็นผลมาจากโครงการวิจัยจีโนมในมนุษย์ที่มีเป้าหมายหลักในการหาลำดับเบสทั้งหมดในมนุษย์ ปัญหาที่สำคัญก็คือ จะมีวิธีการอย่างไรในการบริหารจัดการกับข้อมูลจำนวนมหาศาลนี้ ในปัจจุบันระบบสารสนเทศเข้ามามีบทบาทต่อการดำเนินชีวิตประจำวันของมนุษย์มากขึ้น โดยเฉพาะเทคโนโลยีทางชีววิทยา ซึ่งกำลังได้รับความสนใจเป็นอย่างมากในโลกปัจจุบันและอนาคตอันใกล้นี้ ระบบคอมพิวเตอร์ได้ถูกนำมาใช้เป็นเครื่องมือในการวิจัยและพัฒนาสิ่งต่างๆ มากมาย รวมถึงข้อมูลทางชีววิทยาที่มักจะมีปริมาณมาก ดังจะเห็นได้จากจำนวนข้อมูลที่เพิ่มขึ้นในฐานข้อมูล GenBank ที่จากปี ค.ศ 1997 มีปริมาณข้อมูลลำดับดีเอ็นเอ (DNA Sequence) 500 ล้านลำดับ ขณะที่ปี ค.ศ. 2000 มีข้อมูลเพิ่มขึ้นเป็น 11,000 ล้านลำดับ (จิราภรณ์ อังวิทยาธร *และ คณะ.*, 2544) ซึ่งเป็นการเพิ่มขึ้นอย่างรวดเร็ว ดังนั้นจึงมีการนำเทคโนโลยีเหมืองข้อมูล (Data Mining) มาใช้ในการค้นหารูปแบบข้อมูลเพื่อนำมาใช้ให้เกิดประโยชน์ในหลายรูปแบบ เช่น การทำนาย (Prediction) การจัดแบ่งกลุ่ม (Classification) หรือการค้นหาความสัมพันธ์ (Association Rules) ของข้อมูล เป็นต้น

ความถูกต้องแม่นยำในการวินิจฉัยโรคเป็นสิ่งที่มีความสำคัญมาก จึงมีการพยายามคิดค้นวิธีที่จะช่วยให้การวินิจฉัยมีความถูกต้องมากขึ้นตั้งแต่อดีตจนถึงปัจจุบัน โดยในอดีตแพทย์จะวินิจฉัยโดยดูจากอาการเป็นหลัก อาจเกิดความผิดพลาดในการวินิจฉัย หรือล่าช้ากว่าจะรู้อาการที่แท้จริงทำให้รักษาไม่ทันการได้ ถือได้ว่ามีความเสี่ยงสูง ต่อมาจึงได้มีการใช้การวิเคราะห์เกี่ยวกับยีนเข้ามาช่วยในการวินิจฉัยโรค ซึ่งสามารถเพิ่มความแม่นยำและรวดเร็วในการวินิจฉัยโรคมากขึ้น แต่การวิเคราะห์ยีนในอดีตจะยุ่งยาก ใช้เวลานาน และเสียค่าใช้จ่ายสูง จนมาถึงปัจจุบันเทคโนโลยีเจริญก้าวหน้ามากขึ้นได้มีการพัฒนาเทคนิคใหม่ที่ใช้วิเคราะห์การแสดงออกของยีนขึ้นมาเรียกว่า เทคนิคไมโครอาเรย์ เทคนิคนี้เป็นเทคนิคที่ใช้ตรวจดูการแสดงออกของยีนจากตัวอย่างที่แตกต่างกันเพื่อค้นหาความรู้หรือลักษณะที่น่าสนใจซึ่งซ่อนอยู่ภายในยีนจำนวนมาก

ปัจจุบันมีการประยุกต์ใช้เทคนิคไมโครอาเรย์ในการวินิจฉัยโรคมากขึ้นเรื่อยๆ จากความได้เปรียบของเทคนิคไมโครอาเรย์ที่สามารถตรวจสอบหรือพิจารณาลักษณะที่เกี่ยวข้องของยีนได้ในขอบเขตที่กว้างมากในครั้งเดียว ซึ่งในแต่ละการทดลองอาจมียีนจำนวน

หลายพันยีนต่อการทดลองหนึ่งการทดลอง ทำให้สามารถเข้าใจและเปิดเผยข้อมูลทางพันธุกรรมออกมาได้อย่างรวดเร็ว ซึ่งแตกต่างจากในอดีตซึ่งต้องอาศัยเวลานานในการทดลองกับยีนทีละตัว จึงทำให้การวินิจฉัยโรคโดยอาศัยเทคนิคไมโครอาเรย์มีประสิทธิภาพ รวดเร็วและมีความแม่นยำสูง หลักการพื้นฐานของการวินิจฉัยโรคโดยอาศัยเทคนิคไมโครอาเรย์คือการตรวจหายีนที่เกี่ยวข้องกับการเกิดโรค ยีนที่เป็นตัวบอกกลุ่ม หรือระดับของโรค ค้นหายีนที่มีแนวโน้มจะทำให้เกิดโรค การหาความสัมพันธ์ของยีนดังที่กล่าวมานี้จะสามารถช่วยสนับสนุนการทำงานของแพทย์ในการวิเคราะห์โรค รวมถึงการรักษาโรคด้วยยาได้ โดยเมื่อแพทย์ให้ยาคนไข้แต่ละครั้ง สามารถติดตามดูผลการตอบสนองต่อยาของโรคได้โดยการพิจารณาการแสดงออกของยีนเฉพาะกลุ่มที่เกี่ยวข้องกับการเกิดโรคนั้นๆได้ หรือการหากฎความสัมพันธ์ของยีนที่ก่อให้เกิดโรค กลุ่มยีนที่บอกประเภทของโรค (จิราภรณ์ อังวิทยาธร *และ คณะ*., 2544) สามารถนำกฎที่ได้มาช่วยในการวินิจฉัยโรคเพื่อเพิ่มความแม่นยำและลดความเสี่ยงในการวินิจฉัยโรคได้อีกด้วย

วิทยานิพนธ์นี้ได้นำเสนอแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ (DNA Microarray) โดยใช้เทคนิคข่ายงานประสาทเทียม (Artificial Neural Network) และวิธีการทางสถิติ (Statistics Method) คือ P-Value (Probability Value หรือ Level of Significant) เพื่อค้นหากฎความสัมพันธ์ของกลุ่มยีนที่เกี่ยวข้องกับการเกิดโรค โดยความรู้ที่ได้จากการสกัดความรู้โดยแบบจำลองนี้จะอยู่ในรูปแบบของกฎ “ถ้า-แล้ว” ที่ผู้ใช้สามารถเข้าใจได้ง่าย

เนื้อหาในส่วนต่อไปนี้จะกล่าวถึงหัวข้อต่างๆที่เกี่ยวข้องกับการวิจัยดังนี้ การตรวจเอกสาร วัตถุประสงค์ของโครงการ ขอบเขตการดำเนินงาน ขั้นตอนและระยะเวลาในการดำเนินงาน สถานที่และเครื่องมือที่ใช้ และประโยชน์ที่คาดว่าจะได้รับ

1.1 การตรวจเอกสาร

เอกสารและงานวิจัยที่เกี่ยวข้องกับงานวิจัยนี้แบ่งออกเป็น 3 ส่วน คือ ข้อมูลดีเอ็นเอไมโครอาเรย์ (DNA Microarray) การเลือกยีน (Gene Selection) การจัดกลุ่มข้อมูลแบบไม่มีการสอน (Unsupervised Clustering) และขั้นตอนการสกัดความรู้ (Knowledge Extraction) ซึ่งมีรายละเอียดดังนี้

1.1.1 ดีเอ็นเอไมโครอาเรย์และการแสดงออกของยีน

ดีเอ็นเอไมโครอาเรย์เป็นเทคโนโลยีใหม่ที่กำลังเข้ามามีบทบาทสำคัญในงานวิจัยทางวิทยาศาสตร์ชีวภาพ รวมถึงการศึกษาวิเคราะห์ทางด้านจีโนม (Genome) เช่น การศึกษาการแสดงออกของยีน การศึกษาการกลายพันธุ์ การค้นพบยา รวมถึงการ

วิเคราะห์โรค เพื่อประโยชน์ของการรักษาในขั้นตอนต่อไป การทดลองไมโครอาเรย์มีเป้าหมายเพื่อ เปรียบเทียบการแสดงออกของยีนจากสองตัวอย่าง มีคุณลักษณะเด่นก็คือ สามารถทำงานได้รวดเร็วและมีประสิทธิภาพ (Jeng *et al*., 2005) เนื่องจากในการทดลองแต่ละครั้งสามารถวิเคราะห์ข้อมูลยีนได้ครั้งละจำนวนมาก แตกต่างจากในอดีตที่แต่ละการทดลองจะวิเคราะห์ยีนทีละยีนเท่านั้น (Jiang *et al*., 2004) จึงไม่น่าสงสัยเลยว่าในอนาคตอันใกล้นี้ ดีเอ็นเอไมโครอาเรย์จะกลายเป็นเทคนิคพื้นฐานสำหรับงานประยุกต์ทางสาขาต่างๆ ไม่ว่าจะเป็นงานด้านวิทยาศาสตร์การแพทย์ วิทยาศาสตร์การเกษตร สัตวศาสตร์ ตลอดจนงานด้านเทคโนโลยีชีวภาพ

ดีเอ็นเอไมโครอาเรย์เป็นชุดของดีเอ็นชนิดต่างๆ ที่มีความแตกต่างกันในลำดับเบส หรือมาจากตัวอย่างที่แตกต่างกัน ถูกนำมาหยดเรียงหรือสังเคราะห์ขึ้นอย่างเป็นระเบียบแบบแผนบนผิวของแผ่นค้ำจุน สำหรับแผ่นค้ำจุนที่ใช้อาจเป็นแผ่นสไลด์แก้ว แผ่นไนลอน หรือแผ่นซิลิโคนก็ได้ (Jiang *et al*., 2004) และเนื่องจากขนาดของวัสดุค้ำจุนที่มีขนาดค่อนข้างเล็ก ในบางครั้งจึงมีผู้เรียกดีเอ็นเอไมโครอาเรย์ว่า ดีเอ็นเอชิป (DNA Chip) หรือยีนชิป (Gene Chip) ส่วนจำนวนของยีนที่หยดเรียงหรือสังเคราะห์ขึ้นบนแผ่นค้ำจุน อาจมีตั้งแต่หลายร้อยจนถึงหมื่นกว่าชนิดต่อแผ่น ขึ้นอยู่กับเทคนิคที่ใช้หยดหรือสังเคราะห์ (Lalinka, 2004)

ดีเอ็นเอไมโครอาเรย์ถูกพัฒนาขึ้นมาเพื่อนำมาใช้ในการตรวจวิเคราะห์ยีนและการแสดงออกของยีนในระดับจีโนม ซึ่งดีเอ็นเอต่างๆ ที่นำมาใช้ในการทำดีเอ็นเอไมโครอาเรย์มักเป็นดีเอ็นเอจากกลุ่มยีนที่มีความสัมพันธ์กัน (Yano and Kotani, 2003) เช่นกลุ่มยีนที่มีการแสดงออกในเนื้อเยื่อที่จำเพาะในช่วงใดช่วงหนึ่งของการเจริญพัฒนาหรือในวิถีเมตาบอลิซึมใดวิถีหนึ่ง การตรวจวิเคราะห์โรคด้วยดีเอ็นเอไมโครอาเรย์ (Jiang *et al*., 2004) ประกอบด้วย 4 ขั้นตอนดังนี้

1) การผลิตดีเอ็นเอไมโครอาเรย์

สิ่งสำคัญเริ่มแรกคือการคัดเลือกตัวอย่างเพื่อใช้หยดเรียงลงบนแผ่นค้ำจุน ดีเอ็นเอไมโครอาเรย์มีลักษณะ 2 มิติอยู่บนแผ่นค้ำจุน ซึ่งอาจเป็นแผ่นสไลด์แก้ว แผ่นไนลอน หรือแผ่นซิลิโคนก็ได้ ซึ่งบนแผ่นค้ำจุนนี้มีจุดอยู่เป็นจำนวนมาก เป็นร้อยเป็นพัน ส่วนขนาดของแต่ละจุดโดยทั่วไปจะอยู่ระหว่าง 140 μm ถึง 240 μm

2) การเตรียมตัวอย่างตรวจ

สำหรับการศึกษาด้าน Genomics ตัวอย่างตรวจที่ใช้จะเป็น Genomic DNA บางกรณีอาจใช้ mRNA ส่วนการศึกษาด้านการแสดงออกของยีนตัวอย่างที่ใช้จะเป็น mRNA ซึ่งจะใช้สำหรับสร้าง cDNA ต่อไป โดยทั่วไปในการตรวจวิเคราะห์จะมีตัวอย่างอ้างอิงที่เหมาะสมมาทำการทดสอบควบคู่เพื่อเปรียบเทียบกับตัวอย่างที่จะทดสอบ ตัวอย่างเช่น การศึกษาการกลายพันธุ์ จะใช้ตัวอย่างที่ทดสอบดูการกลายพันธุ์เปรียบเทียบกับตัวอย่างอ้างอิง หรือในกรณีศึกษาการแสดงออกของยีน ก็จะใช้ตัวอย่างที่

ทดสอบเปรียบเทียบกับตัวอย่างอ้างอิง ซึ่งตัวอย่างทั้งสองนี้จะเป็นตัวอย่างที่อยู่ในภาวะที่แตกต่างกัน เช่น ภาวะไม่เป็นโรคกับภาวะเป็นโรค เป็นต้น

3) การติดฉลากตัวอย่างตรวจและการ Hybridization

ตัวอย่างที่จะทำการทดสอบและตัวอย่างอ้างอิงจะนำมาทำการติดฉลากด้ายสารเรืองแสงที่แตกต่างกัน เช่น Cy3 (สีเขียว) และ Cy5 (สีแดง) เป็นต้น ตัวอย่างที่ใช้จะเป็น cDNA ซึ่งถ้าเป็น RNA ก็จะใช้ปฏิกิริยา Reverse Transcription เพื่อสร้างสาย cDNA พร้อมกับติดฉลากเรืองแสงไปในตัว หลังจากติดฉลากแล้วตัวอย่างทั้งสองจะถูกนำมาผสมให้เข้ากัน แล้วนำไป Hybridize กับ Microarray ที่เหมาะสมต่อไป

4) การอ่านและบันทึกผล

ดีเอ็นเอที่ Hybridize แล้วจะนำไปอ่านผลโดยการใช้เครื่องสแกนเลเซอร์ (Laser Scanner) ผลที่ได้จะถูกใช้และประมวลผลโดยโปรแกรมต่อไป

การทดลองไมโครอาเรย์ โดยทั่วไปแล้วประกอบด้วยตัวอย่าง 2 ตัวอย่างคือ ตัวอย่างควบคุมเป็นตัวอย่างที่รู้จัก กับตัวอย่างที่ต้องการทดสอบ ตัวอย่างควบคุมและตัวอย่างทดสอบจะแตกต่างกันเช่น เซลล์สุขภาพดีกับเซลล์ที่เป็นโรค หรือทั้งสองตัวอย่างเป็นเซลล์ในช่วงวงจรชีวิตที่แตกต่างกัน cDNA ที่นำมาทดลองจะถูก Hybridized กันบนจุดของแผ่นสไลด์โดยเป็นการผสมกันโดยอาศัยกฎการจับคู่กันอย่างจำเพาะของคู่เบสบนสายดีเอ็นเอ ซึ่งเป็นสายลำดับของตัวอักษรที่แตกต่างกัน 4 ตัว ซึ่งแต่ละตัวคือเบสต่างชนิดกันดังนี้ Adenine (A) Timine (T) Cytosin (S) และ Guanine (G) โดยเบสทั้ง 4 มีกฎการจับคู่กันที่จำเพาะ คือ A จับคู่กับ T ส่วน C จับคู่กับ G โดยทั่วไปแล้ว cDNA จะติดฉลากด้วยการใช้สารเรืองแสงหรือฟลูออเรสเซน ซึ่งมีสองสีคือสีเขียว (Cy3) สำหรับตัวอย่างควบคุม และสีแดง (Cy5) สำหรับตัวอย่างที่จะทดสอบ โมเลกุลที่ถูกติดฉลากด้วยสารเรืองแสงหรือฟลูออเรสเซนแล้วนั้นจะสามารถตรวจจับเมื่อถูกกระตุ้นด้วยแสงเลเซอร์ การทดลองไมโครอาเรย์ผลที่ได้จะเป็นรูปภาพ ซึ่งประกอบด้วยจุดสีหลายสี สีของจุดแต่ละจุดจะเกิดจากการจับตัวกันของ cDNA สีและความเข้มข้นที่ได้มาจากไมโครอาเรย์จะสามารถแปลงเป็นค่าจำนวนจริงได้ โดยจะได้เป็นเมทริกต์ ตัวอย่างเช่น เมทริกซ์ M ซึ่งจะเรียกว่าเป็นเมทริกซ์การแสดงออกของยีน ซึ่งค่าสมาชิกของเมทริกซ์ M (Mij) จะเท่ากับ log2(Cy5/Cy3) (Lalinka, 2004; จิราภรณ์ *และคณะ*., 2544) แถวของเมทริกซ์หมายถึงการแสดงออกของยีน และสดมภ์หมายถึงการทดลอง

1.1.2 การเลือกยีน

การเลือกยีน (Gene Selection) เป็นขั้นตอนที่สำคัญมากในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาเรย์ เนื่องจากข้อมูลดีเอ็นเอไมโครอาเรย์เป็นข้อมูลที่มีลักษณะหลายมิติ (High-Dimensional) คือจะประกอบด้วยยีนจำนวนมาก ทำให้มีความซับซ้อนในการวิเคราะห์มาก และยังมีโอกาสที่จะเกิดความผิดพลาดสูงอีกด้วย จึงมีการคิดค้นวิธีการที่จะลดมิติของ

ข้อมูลลง (Chen and Hua, 2003) โดยใช้วิธีการเลือกยีนที่มีความสำคัญ หรือเกี่ยวข้องกับสิ่งที่สนใจขึ้นมากลุ่มหนึ่งก่อน และต่อไปจะพิจารณาเฉพาะยีนกลุ่มนี้เท่านั้น ทำให้สามารถลดเวลาและทรัพยากรในการทดลองได้มาก ในขณะที่ความถูกต้องไม่ลดลง ในหัวข้อนี้จะกล่าวถึงวิธีที่นิยมใช้ในกระบวนการเลือกยีน ได้แก่ ค่าเอ็นโทรพี (Entropy) และ วิธีการทางสถิติ (Statistics Method)

1) ค่าเอ็นโทรพี (Entropy)

ค่าเอ็นโทรพีทำให้สามารถหาข้อมูลที่ใช้ในการแยกแยะกลุ่มของข้อมูลที่เราสนใจออกมาได้ จากข้อมูลจำนวนมาก (Kim *et al.*, 1997) จึงมีการนำแนวคิดของค่าเอ็นโทรพีมาใช้กับวิธีการเลือกยีนที่สนใจ ที่น่าจะเกี่ยวข้องกับการแบ่งกลุ่มออกมาก่อนขั้นตอนการแบ่งกลุ่ม ทำให้เวลาที่ใช้ในการแบ่งกลุ่มลดลงได้ การเลือกยีนจะพิจารณาค่าเอ็นโทรพีต่ำสุด ซึ่งเป็นค่าที่บอกว่าถ้าใช้ข้อมูลตัวนี้ก็จะเพียงพอแล้วสำหรับการตัดสินใจ หลักการเอ็นโทรพีค่าต่ำสุดเป็นหลักการที่นำมาใช้ในการเลือกจุดแบ่งที่ดีที่สุดที่ใช้ในการแบ่งข้อมูลออกเป็น 2 กลุ่ม และจุดแบ่งที่ถูกเลือกมาต้องมีค่าเอ็นโทรพีต่ำที่สุด

การเลือกยีนโดยพิจารณาจากค่าเอ็นโทรพี (Entropy) หรือค่าความคาดหวังของข้อมูล ซึ่งเป็นค่าที่บอกว่ายีนนั้นสามารถแบ่งกลุ่มได้ดีแค่ไหน (Tan and Pan, 2005) ค่าเอ็นโทรปีคำนวณได้ดังสมการที่ (1.1) กำหนดให้ $S_p(x)$ คือค่าเอ็นโทรปีของขอบเขต p และ $S_q(x)$ คือค่าเอ็นโทรปีของขอบเขต q คำนวณได้จากสมการที่ (1.2) และ (1.3) ตามลำดับ ค่า $p_k(x)$ คือค่าความเป็นไปได้ที่ข้อมูลกลุ่ม k อยู่ในขอบเขต p และค่า $q_k(x)$ คือค่าความเป็นไปได้ที่ข้อมูลกลุ่ม k อยู่ในขอบเขต q ซึ่งคำนวณได้ดังสมการที่ (1.4) - (1.7)

$$S(x) = p(x)S_p(x) + q(x)S_q(x) \tag{1.1}$$

$$S_p(x) = -(p_1(x)\ln p_1(x) + p_2(x)\ln p_2(x)) \tag{1.2}$$

$$S_q(x) = -(q_1(x)\ln q_1(x) + q_2(x)\ln q_2(x)) \tag{1.3}$$

$$p_k(x) = \frac{n_k(x)+1}{n(x)+1} \tag{1.4}$$

$$p(x) = \frac{n(x)+1}{n+1} \tag{1.5}$$

$$q_k(x) = \frac{m_k(x)+1}{m(x)+1} \tag{1.6}$$

$$q(x)=\frac{m(x)+1}{n+1} \tag{1.7}$$

2) วิธีการทางสถิติ (Statistics Method)

ได้มีการประยุกต์ใช้วิธีการทดสอบสมมติฐานทางสถิติหลายวิธี เพื่อนำมาใช้ในการเลือกยีนที่สำคัญสำหรับการทดลอง ซึ่งวิธีที่นิยมใช้กันมีดังนี้

ANOVA F Test (Neter *et al.*, 1996) สามารถคำนวณได้ดังสมการที่ (1.8)

$$F=\frac{(n-k)\sum n_i(\bar{Y}_i-\bar{Y})^2}{(k-1)\sum(n_i-1)S_i^2} \tag{1.8}$$

Brown - Forsythe Test (Brown and Forsythe, 1974) สามารถคำนวณได้ดังสมการที่ (1.9)

$$B=\frac{\sum n_i(\bar{Y}_i-\bar{Y})^2}{\sum(1-n_i/n)S_i^2} \tag{1.9}$$

Cochran Test (Cochran, 1937) สามารถคำนวณได้ดังสมการที่ (1.10)

$$C=\sum w_i\left(\bar{Y}_i-\sum h_i\bar{Y}_i\right)^2 \tag{1.10}$$

Welch Test (Welch, 1951) สามารถคำนวณได้ดังสมการที่ (1.11)

$$W=\frac{\sum w_i(\bar{Y}_i-\sum h_i\bar{Y}_i)^2}{(k+1)+2(k-2)(k+1)^{-1}\sum(n_i-1)^{-1}(1-h_i)^2} \tag{1.11}$$

ตัวอย่างงานวิจัยที่มีการใช้เทคนิคการเลือกยีนมีหลายงานวิจัยด้วยกัน ซึ่งได้ยกตัวอย่างมาดังนี้ งานวิจัยของ Chen และ Hua (Chen and Hua, 2003) ได้ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพของการเลือกยีนด้วยวิธีการทางสถิติทั้ง 4 วิธีข้างต้น โดยใช้ข้อมูลโรคลูคีเมีย (Leukemia) และโรคมะเร็งรังไข่ (Ovarian Cancer) และเสนอว่า Brown - Forsythe เป็นวิธีที่ดีที่สุด งานวิจัยของ Kim และคณะ (*Kim et al.*, 1997) ได้ใช้เทคนิคค่าเอ็นโทรพี (Entropy) ร่วมกับ Fuzzy Neural Network ในการแบ่งกลุ่มข้อมูล และสร้างกฎการเรียนรู้แบบ “ถ้า–แล้ว” (IF–THEN Rule) โดยเลือกใช้ค่าเอ็นโทรพีน้อยที่สุดในการเลือกยีน โดยงานวิจัยของ Kim จะใช้กับข้อมูลที่มี 2 กลุ่มเท่านั้น และงานวิจัยของ Tan และ Pan (Tan and Pan, 2005) ได้ใช้เทคนิค ค่าเอ็นโทรพี (Entropy) และเทคนิค Fisher Feature Selection ซึ่งเป็นเทคนิคหนึ่งในการเลือกยีนโดยพิจารณาจากค่าความแปรปรวน ร่วมกับการแบ่งกลุ่มข้อมูลโดยใช้ Self-Organizing Map (SOM) เพื่อแบ่งกลุ่มข้อมูลดีเอ็นเอไมโครอาเรย์ โดยผลลัพธ์ที่ได้

แสดงให้เห็นว่าเทคนิคค่าเอ็นโทรพีจะให้ค่าความถูกต้องสูงกว่าเทคนิค Fisher Feature Selection

1.1.3 การจัดกลุ่มข้อมูลแบบไม่มีการสอน

เป็นวิธีการแบ่งกลุ่มข้อมูลโดยที่ไม่ทราบมาก่อนว่าข้อมูลมีทั้งหมดกี่กลุ่ม และข้อมูลแต่ละกลุ่มมีลักษณะอย่างไร แต่จะแบ่งกลุ่มโดยพิจารณาจากลักษณะของข้อมูลเข้า โดยวิธีการที่นิยมนำมาใช้ในการแบ่งกลุ่มข้อมูลดีเอ็นเอไมโครอาเรย์ มี 3 วิธี คือ Hierarchical Clustering K-Mean Clustering และ Self-Organizing Maps

1) Hierarchical Clustering

เป็นเทคนิคการแบ่งกลุ่มข้อมูลที่เป็นที่นิยมวิธีหนึ่ง ซึ่งไม่ต้องมีการกำหนดจำนวนกลุ่มล่วงหน้า ให้ผลลัพธ์ออกมาในรูปของเด็นโดรแกรม (Dendrogram) (Jiang *et al.*, 2004) วิธีการ Hierarchical Clustering พื้นฐานแบ่งออกเป็น 2 ชนิดได้แก่ 1) Agglomerative Hierarchical Clustering จะเริ่มต้นด้วยกันพิจารณายีนแต่ละยีนเป็นเป็น 1 กลุ่ม แล้วเริ่มต้นรวมยีนที่มีลักษณะที่คล้ายกันเข้าด้วยกัน เพื่อให้ได้กลุ่มที่ใหญ่ขึ้น และทำเช่นนี้ไปเรื่อยๆ จนกว่าจะได้กลุ่มที่ต้องการ และ 2) Divisive Hierarchical Clustering ซึ่งจะตรงข้ามกับวิธีแรก โดยการเริ่มต้นจะพิจารณาทุกยีนเป็นกลุ่มเดียวกัน แล้วค่อยแตกออกเป็นกลุ่มย่อยๆ จนแต่ละยีนแยกออกจากกันเป็นเซตย่อย ทำเช่นนี้ไปเรื่อยๆ จนกว่าจะได้กลุ่มที่ต้องการ (Jiang *et al.*, 2004) ตัวอย่างขั้นตอนการทำงานพื้นฐานของ Hierarchical Clustering ประกอบด้วย 2 ขั้นตอน แสดงได้ดังภาพประกอบที่ 1.1

1) วัดระยะห่างระหว่างแต่ละยีนนำมาสร้างเป็นตาราง โดยวิธีการวัดระยะทางที่นิยมใช้ได้แก่ Euclidian Distance ดังนี้

$$d = \sqrt{\sum (x_i - y_i)^2}$$

⇩

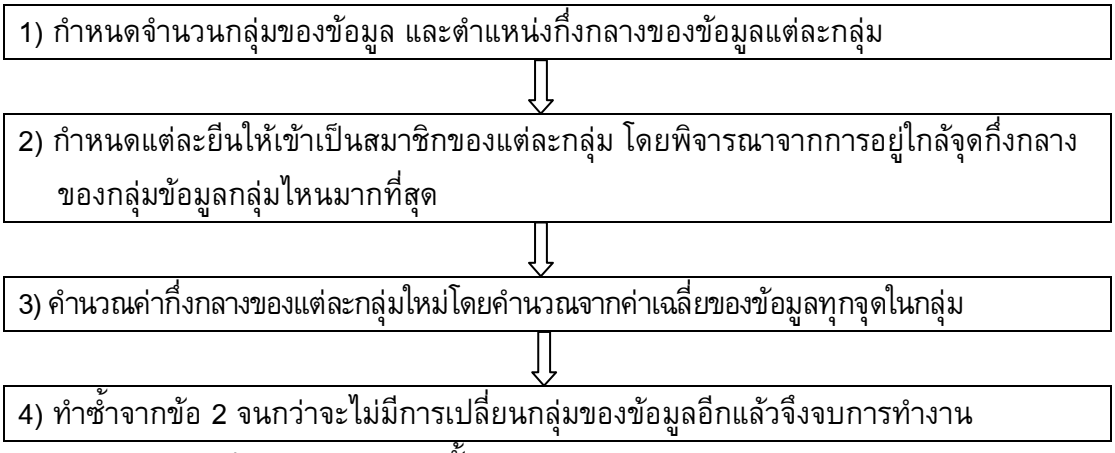
2) จัดกลุ่มยีน 2 กลุ่มที่อยู่ใกล้กันเข้าด้วยกัน จะได้กลุ่มใหม่ขึ้นมา แล้วกลับไปทำซ้ำจากข้อ 1 จนกระทั่งได้กลุ่มยีนเพียงกลุ่มเดียว

ภาพประกอบ 1.1 แสดงขั้นตอนการทำงานของ Hierarchical Clustering

2) K-Mean Clustering

เป็นวิธีการแบ่งกลุ่มข้อมูลอีกวิธีหนึ่ง ซึ่งจำเป็นต้องมีการระบุจำนวนกลุ่มของข้อมูลที่ต้องการล่วงหน้า (Jiang *et al.*, 2004) และจะทำการแบ่งกลุ่มจนได้จำนวนกลุ่มที่ต้องการ ปัญหาของ K-Mean ก็คือ การที่พิจารณาว่าข้อมูลไหนควรจะเป็นของ

กลุ่มไหน K-Mean เป็นวิธีการที่จัดว่าง่ายและรวดเร็ววิธีหนึ่ง มักจะทำเป็นอันดับสองรองจาก Clustering Algorithm อื่นๆ ในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาเรย์ เนื่องจากจะใช้ K-Mean จัดการกับข้อมูลที่ถูกจัดกลุ่มไว้ผิดกลุ่มนั่นเอง (Sugiyama and Kotani, 2002 ; Yano and Kotani, 2003) ขั้นตอนการทำงานพื้นฐานของ K-Mean Clustering ประกอบด้วย 4 ขั้นตอนแสดงดังภาพประกอบที่ 1.2



ภาพประกอบ 1.2 แสดงขั้นตอนการทำงานของ K-Mean Clustering

3) Self-Organizing Maps

แผนที่การจัดกลุ่มเอง (Self-Organising Maps) ถูกเสนอโดย Kohonen ในปี 1982 (Jiang *et al.*, 2004) เป็นวิธีการทางเครือข่ายประสาทเทียม (Artificial Neural Network) แบบการเรียนรู้ที่ไม่มีการสอน (Unsupervised Learning) ที่เป็นที่นิยมกันอย่างแพร่หลาย ซึ่งนำมาใช้ในการจัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันให้อยู่ในกลุ่มเดียวกัน ประโยชน์ของแผนที่การจัดกลุ่มเองคือจะทำการลดมิติของข้อมูลให้อยู่ที่ 1 มิติ หรือ 2 มิติเพื่อให้สามารถเข้าใจลักษณะข้อมูลได้ในภาพรวม (Vesanto *et al.*, 2000; Petri *et al.*, 1999; Yano and Kotani, 2003; Jiang *et al.*, 2004) แผนที่การจัดกลุ่มเองประกอบด้วยนิวรอนหลายนิวรอน ซึ่งแทนข้อมูลทดสอบแต่ละตัวแสดงอยู่บนตารางแผนที่ซึ่งมี 2 รูปแบบ คือ ตารางแผนที่สี่เหลี่ยมและตารางแผนที่หกเหลี่ยม (Vesanto, 2000) นิวรอนแต่ละตัวนั้นจะมีค่าน้ำหนัก (Weight Vector) เป็นของตัวเอง ซึ่งค่าน้ำหนักของนิวรอนแต่ละตัวถูกแทนด้วย $m_j = [m_{1,1}, m_{2,1}, \ldots, m_{j,d}]$

โดยที่ *J* คือ นิวรอนตัวที่ *j*

d คือ จำนวนมิติของข้อมูลทดสอบ

สถาปัตยกรรมพื้นฐานของเครือข่ายประสาทการจัดกลุ่มเองโดยปกติ จะประกอบด้วย 2 ชั้น คือชั้นข้อมูลเข้า (Input Layers) และชั้นข้อมูลออก (Output Layers) โดยการจัดกลุ่มข้อมูลจะเกิดขึ้นในชั้นข้อมูลออก

ขั้นตอนการเรียนรู้ของแผนที่การจัดกลุ่มเองประกอบด้วย 4 ขั้นตอนหลัก คือ (Lalinka, 2004) 1) กำหนดค่าน้ำหนักเริ่มต้น (Synaptic Weights Initialization) 2) กระบวนการแข่งขัน (Competitive Process) 3) กระบวนการกำหนดโหนดเพื่อนบ้าน (Cooperative Process) และ 4) กระบวนการปรับค่าน้ำหนัก (Adaptive Process) ดังนี้

1) Synaptic Weights Initialization

แต่ละขั้นตอนของการเรียนรู้ของรอบเวลา (t) จะทำการสุ่มเลือกนิวรอนตัวอย่างที่อยู่บนตารางแผนที่หรือเรียกว่า Sample Vector (x_i) จากข้อมูลทดสอบพร้อมทั้งกำหนดค่าน้ำหนักให้นิวรอน

2) Competitive Process

เป็นการคำนวณระยะทางระหว่างนิวรอนตัวอย่างที่ได้สุ่มเลือกจากขั้นตอนที่ 1 กับนิวรอนตัวอื่นๆ ที่อยู่บนตารางแผนที่ทุกๆตัวโดยคำนวณจากสมการ Euclidean Distance (Jiang *et al.*, 2004) โดยที่นิวรอนตัวที่มีค่าระยะทางน้อยที่สุด จะเป็น Best-Matching Unit (BMU) หรือมีลักษณะคล้ายกับนิวรอนตัวอย่างที่สุ่มเลือกมามากที่สุด

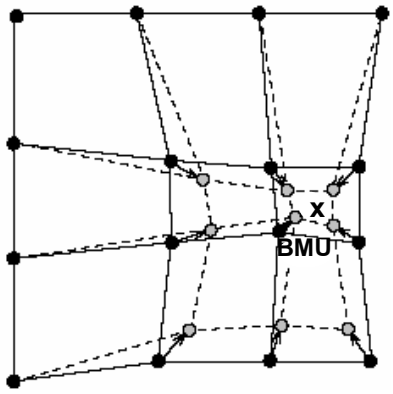
3) Cooperative Process

เมื่อได้นิวรอนที่ชนะแล้ว นิวรอนที่ชนะจะส่งผลกระทบต่อนิวรอนเพื่อนบ้านในรัศมีหนึ่ง โหนดเพื่อนบ้านจะหาได้จากสมการของ Gaussian function

4) Adaptive Process

โหนดที่อยู่ในรัศมีที่คำนวณได้ในขั้นตอนก่อนหน้านี้ จะถูกปรับค่าน้ำหนักให้มีระยะทางเข้าใกล้กับโหนดที่ถูกเลือกมากขึ้น

จากกระบวนการทำงานทั้ง 4 นี้ จะทำให้นิวรอนบนแผนที่เกิดการปรับเปลี่ยนค่าน้ำหนักเข้าใกล้มากขึ้น จนกระทั่งได้กลุ่มของข้อมูลที่ต้องการ (Vesanto *et al*., 2000; Petri *et al*., 1999; Wiphada and Putthiporn, 2006) การปรับเปลี่ยนค่าน้ำหนักของโหนดที่ชนะและโหนดเพื่อนบ้านแสดงดังภาพประกอบที่ 1.3



ภาพประกอบ 1.3 แสดงการปรับเปลี่ยนค่าน้ำหนักของโหนดที่ชนะและโหนดเพื่อนบ้าน

ตัวอย่างงานวิจัยที่ใช้เทคนิคการแบ่งกลุ่มข้อมูลมีอยู่หลายงานวิจัย ซึ่งได้ยกตัวอย่างมาดังนี้ แบบจำลอง UPGMA (Datta and Datta, 2006) แบบจำลองนี้ทำการแบ่งกลุ่มข้อมูลดีเอ็นเอไมโครอาเรย์ โดยใช้เทคนิค Hierarchical Clustering แบบ Agglomerative Hierarchical Clustering และได้แสดงผลลัพธ์เป็นเด็นโดรแกรม ซึ่งทำให้สามารถระบุจำนวนกลุ่มที่ต้องการได้ และสามารถดูความใกล้เคียงกันของข้อมูลแต่ละกลุ่มได้ งานวิจัยของ Yano และ Ktani (Yano and Kotani, 2003) และ งานวิจัยของ Sugiyama และ Kotani (Sugiyama and Kotani, 2002) ได้ทำการแบ่งกลุ่มข้อมูลดีเอ็นเอไมโครอาเรย์ โดยใช้เทคนิค Self-Organizing Map (SOM) และต่อมานำผลการจัดกลุ่มข้อมูลที่ได้จาก SOM มาหากลุ่มของข้อมูลโดยการใช้ K-Mean อีกรอบหนึ่ง แต่ไม่ได้มีการสร้างกฎการเรียนรู้ งานวิจัยของ Patra และคณะ (Patra *et al.,* 2006) ได้เสนอเทคนิคในการแบ่งกลุ่มข้อมูลโดยใช้ Self-Organizing Map โดยเพิ่มในส่วนของการแบ่งกลุ่มโดยการวิเคราะห์ค่า U-Matrix ซึ่งจากการคำนวณค่า U-Matrix จะมองเห็นเป็นค่าสีที่แสดงความใกล้เคียงกันของข้อมูล และงานวิจัยนี้จะเลือกจุดที่มีค่าสีสูงหรือมีสีเข้มเป็นจุดที่ใช้ในการแบ่งข้อมูล เนื่องจากเป็นจุดที่แสดงความต่างของข้อมูลมากที่สุด

1.1.4 ขั้นตอนการสกัดความรู้

การสกัดความรู้แบ่งออกเป็น 3 ขั้นตอน คือ การเตรียมข้อมูล (Data Preprocessing) การแบ่งกลุ่มข้อมูลแบบไม่มีการสอน (Unsupervised Clustering) และการหากฎความสัมพันธ์ (Data Mining of Association Rules) (Jiyuan and Phoebe, 2006; Martinez, 2007) โดยมีรายละเอียดดังนี้

1) การเตรียมข้อมูล เป็นการเตรียมข้อมูลก่อนดำเนินการสกัดความรู้ เนื่องจากข้อมูลอาจมีค่าว่าง (Missing Values) ค่าข้อมูลรบกวน (Noise) ค่าสุดโต่ง (Outlier) หรือค่าอื่นๆ ที่จะทำให้ความถูกต้องในการแบ่งกลุ่มลดลงหรือทำให้เสียเวลาในการ

แบ่งกลุ่มมากขึ้น ในขั้นตอนนี้จะขจัดข้อมูลเหล่านี้ออก เพื่อเตรียมข้อมูลให้พร้อมสำหรับการแบ่งกลุ่มต่อไป และทำให้การแบ่งกลุ่มมีประสิทธิภาพและรวดเร็วมากยิ่งขึ้น

2) การแบ่งกลุ่มข้อมูลแบบไม่มีการสอน เป็นการนำข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลเรียบร้อยแล้ว มาจัดกลุ่มโดยใช้เทคนิคการแบ่งกลุ่มข้อมูลแบบไม่มีการสอนทำการแบ่งข้อมูลตามลักษณะของข้อมูล

3) การหากฎความสัมพันธ์ ทำการสกัดความรู้จากการแบ่งกลุ่มข้อมูลในรูปของกฎความสัมพันธ์ โดยในขั้นตอนนี้สิ่งที่ต้องพิจารณาร่วมด้วยคือการพิจารณาตัดกฎที่ซ้ำซ้อนกันออกเพื่อเพิ่มความถูกต้องของกฎ

1.2 วัตถุประสงค์ของโครงการ

1.2.1 วิเคราะห์และออกแบบแบบจำลองในการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์

1.2.2 พัฒนาโปรแกรมจากแบบจำลองในการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์

1.3 ขอบเขตของการดำเนินงาน

1.3.1 พัฒนาอัลกอริทึมสำหรับวิเคราะห์ข้อมูลการแสดงออกของยีนจากข้อมูลดีเอ็นเอไมโครอาเรย์

1.3.2 หากฎความสัมพันธ์ของยีนที่เกี่ยวข้องกับการเกิดโรค

1.3.3 พัฒนาระบบจำลองการวินิจฉัยโรค

1.3.4 ข้อมูลที่นำมาใช้เป็นข้อมูลทดสอบ จะใช้ข้อมูลโรคจากฐานข้อมูลทางการแพทย์ที่เป็นสากล จากห้องวิจัย Kent Ridge มหาวิทยาลัย Pittsburgh ประเทศสหรัฐอเมริกา และโครงการวิจัยการแสดงออกของยีน (Gene Expression) มหาวิทยาลัยพรินซ์ตัน (Princeton University) ประเทศสหรัฐอเมริกา

1.4 ขั้นตอนและระยะเวลาการดำเนินงาน

1.4.1 ขั้นตอนการดำเนินงาน

1) ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง

2) ศึกษาเทคนิคที่เกี่ยวข้องในงานวิจัยได้แก่ เทคนิคไมโครอาเรย์ และเทคนิคการแบ่งกลุ่มข้อมูลแบบไม่มีผู้สอน (Unsupervised Clustering) เช่น Hierarchical Clustering K-Mean Clustering และ Self-Organizing Maps

3) ศึกษาเทคโนโลยีและเครื่องมือสนับสนุนเช่น โปรแกรม MATLAB และ Macro VBA

4) วิเคราะห์และออกแบบแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยแผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction Using Self-Organizing Map: DNA_MKESOM)

5) พัฒนาระบบ DNA_MKESOM

6) ทดสอบและติดตั้งระบบ

7) จัดทำเอกสารประกอบระบบ และเขียนผลงานวิจัย

8) จัดทำเอกสารวิทยานิพนธ์

1.4.2 ระยะเวลาดำเนินงาน

มิถุนายน 2549 – มีนาคม 2551

1.4.3 แผนการดำเนินการวิจัย

แผนการดำเนินงานวิจัย แสดงได้ดังตารางที่ 1.1

ตารางที่ 1.1 แสดงแผนการดำเนินงานวิจัย

กิจกรรมขั้นตอนการดำเนินงาน	เดือน																	
	2549			2550												2551		
	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง	■	■	■															
2. ศึกษาเทคนิคที่เกี่ยวข้องในงานวิจัย		■	■	■	■	■	■											
3. ศึกษาเทคโนโลยีและเครื่องมือสนับสนุน				■	■	■	■											
4. วิเคราะห์และออกแบบระบบ					■	■	■	■	■									
5. พัฒนาระบบ							■	■	■	■	■	■	■	■	■	■		
6. ทดสอบและติดตั้งระบบ								■	■	■	■	■	■	■	■	■	■	
7. จัดทำเอกสารประกอบระบบ													■	■	■	■		
8. เขียนผลงานวิจัย						■	■	■	■	■	■	■	■	■				
9. จัดทำเอกสารวิทยานิพนธ์														■	■	■	■	■

1.5 สถานที่และเครื่องมือที่ใช้

1.5.1 สถานที่

ห้องวิจัยปัญญาประดิษฐ์ CS207 ภาควิชาวิทยาการคอมพิวเตอร์ ตึกวิทยาการ-คอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

1.5.2 เครื่องมือที่ใช้

1) ด้านฮาร์ดแวร์

- เครื่องคอมพิวเตอร์ส่วนบุคคล หน่วยความจำ 2 กิกะไบต์ ฮาร์ดดิสก์ ความจุ 40 กิ กะไบต์ สำหรับพัฒนาและเป็นเครื่องทดสอบ

2) ด้านซอฟต์แวร์

- ระบบปฏิบัติการ Microsoft Windows XP
- โปรแกรม MATLAB

- Microsoft Excel 2000 สำหรับเตรียมข้อมูล
- Visual Basic .NET

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองในการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง

1.6.2 ได้โปรแกรมจากแบบจำลองในการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง

บทที่ 2

ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีต่างๆ ที่นำมาใช้ในการพัฒนาแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ ประกอบด้วย ทฤษฎีดีเอ็นเอไมโครอาเรย์ (DNA Microarray) ทฤษฎีแผนที่การจัดกลุ่มเอง (Self-Organizing Map: SOM) ทฤษฎีขั้นตอนวิธีทางสถิติ (Statistics Method) และทฤษฎีเกี่ยวกับการวัดประสิทธิภาพของแบบจำลอง ซึ่งรายละเอียดต่างๆ จะอธิบายแยกเป็นหัวข้อ ดังนี้

2.1 ดีเอ็นเอไมโครอาเรย์ (DNA Microarray)

เนื่องจากการศึกษาเกี่ยวกับดีเอ็นเอไมโครเอเรย์จำเป็นต้องอาศัยความรู้พื้นฐานทางด้านชีววิทยาเข้ามาเกี่ยวข้อง ดังนั้นเพื่อให้สามารถศึกษาและทำความเข้าใจได้ง่ายขึ้นในหัวข้อนี้จะกล่าวถึง 2 หัวข้อย่อยด้วยกัน คือ ความรู้พื้นฐานทางชีววิทยา และดีเอ็นเอไมโครอาเรย์

2.1.1 ความรู้พื้นฐานทางชีววิทยา

ในหัวข้อความรู้พื้นฐานทางชีววิทยานี้จะกล่าวถึงเฉพาะหัวข้อที่เกี่ยวข้องกับงานวิจัยนี้เท่านั้น เพื่อให้ผู้ศึกษาสามารถเข้าใจมากขึ้น ซึ่งจะประกอบด้วยหัวข้อต่างๆ ดังนี้ ดีเอ็นเอ (Deoxyribonucleic Acid: DNA) อาร์เอ็นเอ (Ribonucleic Acid: RNA) และการ Hybridization ซึ่งมีรายละเอียดดังนี้

2.1.1.1 ดีเอ็นเอ (Deoxyribonucleic Acid: DNA)

ในหัวข้อดีเอ็นเอจะกล่าวถึง 3 เรื่องด้วยกันคือ ดีเอ็นเอ (DNA) ซีดีเอ็นเอ (Complementary: cDNA) และ DNA Probe

1) ดีเอ็นเอ (Deoxyribonucleic Acid: DNA) เป็นชื่อย่อของสารพันธุกรรม ที่มีชื่อวิทยาศาสตร์ว่า กรดดีออกซีไรโบนิวคลีอิก (Deoxyribonucleic Acid) ซึ่งพบในเซลล์ของสิ่งมีชีวิตทุกชนิด ได้แก่ คน สัตว์ พืช เชื้อรา แบคทีเรีย ไวรัส เป็นต้น ดีเอ็นเอบรรจุข้อมูลทางพันธุกรรมของสิ่งมีชีวิตชนิดนั้นไว้ ซึ่งมีลักษณะที่ผสมผสานมาจากสิ่งมีชีวิตรุ่นก่อน ซึ่งก็คือ พ่อและแม่ และสามารถถ่ายทอดไปยังสิ่งมีชีวิตรุ่นถัดไป ดีเอ็นเอมีรูปร่างเป็นเกลียวคู่คล้ายบันไดลิงที่บิดตัว ขาของบันไดแต่ละข้างก็คือการเรียงตัวของนิวคลีโอไทด์ (Nucleotide) นิวคลีโอไทด์เป็นโมเลกุลที่ประกอบด้วยน้ำตาล ฟอสเฟต (ซึ่งประกอบด้วย ฟอสฟอรัส และ

ออกซิเจน) และเบส นิวคลีโอไทด์มีอยู่สี่ชนิด ได้แก่ อะดีนีน (Adenine: A) ไทมีน (Thymine: T) ไซโตซีน (Cytosine: C) และกัวนีน (Guanine: G) ขาของบันไดสองข้างหรือ นิวคลีโอไทด์ถูกเชื่อมด้วยเบส โดยที่ A จะเชื่อมกับ T และ C จะเชื่อมกับ G เท่านั้น ข้อมูลทางพันธุกรรมในสิ่งมีชีวิตชนิดต่างๆ เกิดขึ้นจากการเรียงลำดับของเบสในดีเอ็นเอนั่นเอง ผู้ค้นพบ ดีเอ็นเอคือ ฟรีดิช มีสเชอร์ ในปี พ.ศ. 2412 (ค.ศ. 1869) แต่ไม่ทราบว่ามีโครงสร้างอย่างไร จนในปี พ.ศ. 2496 (ค.ศ. 1953) เจมส์ ดี วัตสัน และ ฟรานซิส คริก เป็นผู้ไขความลับโครงสร้างของดีเอ็นเอ และนั่นนับเป็นจุดเริ่มต้นของยุคเทคโนโลยีทางดีเอ็นเอ

2) ซีดีเอ็นเอ (Complementary: cDNA) เป็น DNA ที่สังเคราะห์มาจาก mRNA โดยใช้เอนไซม์ Reverse Transcriptase ทำได้โดยการนำ mRNA ที่สกัดได้ใส่ลงในตัวกลาง (Medium) ที่เหมาะสม มีฟอสเฟต น้ำตาลดีออกซี่ไรโบส (Deoxyribose) และเบสชนิดต่าง ๆ ครบครัน พร้อมทั้งเติมเอ็นไซม์ Reverse Transcriptase จะเกิดการสร้างสาย DNA สายเดี่ยวคือ cDNA (complementary DNA) ที่มีการเรียงตัวของเบสสอดคล้อง (Complementary) กับการจัดเรียงตัวของเบสบนสาย mRNA

3) DNA Probe คือดีเอ็นเอสายสั้นๆ ซึ่งติดฉลากกัมมันตรังสีไว้ DNA Probe มีสมบัติที่สำคัญคือ ความสามารถจับกับชิ้นดีเอ็นเอที่สนใจได้อย่างจำเพาะ ทำให้เราสามารถติดตามชิ้นดีเอ็นเอที่สนใจนั้นได้โดยอาศัยกัมมันตภาพรังสีที่ติดอยู่กับ DNA probe (ธีรวรรธน์ ขันทอง, 2008: ระบบออนไลน์)

2.1.1.2 อาร์เอ็นเอ (Ribonucleic Acid : RNA)

ในหัวข้ออาร์เอ็นเอจะกล่าวถึง 2 เรื่องด้วยกันคือ อาร์เอ็นเอ (RNA) และชนิดของอาร์เอ็นเอ

1) อาร์เอ็นเอ (Ribonucleic Acid: RNA) เป็นพอลิเมอร์ของกรดนิวคลีอิกที่ประกอบด้วยนิวคลีโอไทด์ (Nucleotide) เชื่อมต่อกันด้วยพันธะโคเวเลนต์ อาร์เอ็นเอนิวคลีโอไทด์ประกอบด้วยวงน้ำตาลไรโบส (Ribose) ซึ่งแตกต่างจากดีเอ็นเอที่ประกอบด้วยวงน้ำตาลดีออกซีไรโบส (Deoxyribose) อาร์เอ็นเอเกิดจากการคัดสำเนาข้อมูลจากดีเอ็นเอโดยเอนไซม์อาร์เอ็นเอพอลิเมอเรส แล้วเข้ากระบวนการต่อเนื่องโดยเอนไซม์อื่นๆ อีก อาร์เอ็นเอจะทำหน้าที่เหมือนแม่แบบสำหรับแปลข้อมูลจากยีนไปเป็นข้อมูลในโปรตีน แล้วขนย้ายกรดอะมิโนเข้าไปในไรโบโซม (Ribosome) เพื่อผลิตโปรตีน และแปลข้อความไปเป็นสำเนาข้อมูล ในโปรตีน (ธีรวรรธน์ ขันทอง, 2008: ระบบออนไลน์)

2) ชนิดของอาร์เอ็นเอ มีอยู่ด้วยกัน 3 ชนิด แต่ละชนิดมีหน้าที่การทำงานที่แตกต่างกันดังนี้

- mRNA (messenger RNA) ทำหน้าที่นำคำสั่งจาก DNA มายังไซโทพลาสซึมแล้วเกาะติดกับไรโบโซม เมื่อมีการสังเคราะห์โปรตีน mRNA มีขนาด

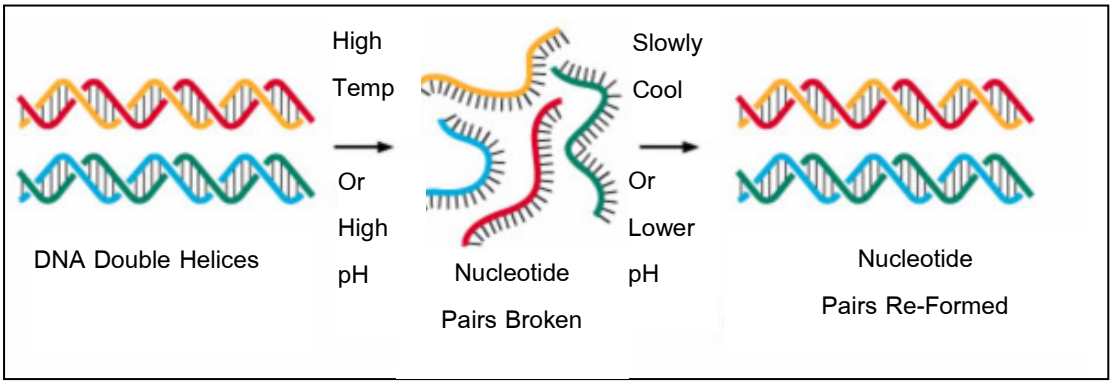
โมเลกุลต่างๆ กัน ลำดับเบสใน mRNA จะกำหนดชนิดของกรดอะมิโน และลำดับของกรดอะมิโนในโมเลกุลของโปรตีนที่สร้างขึ้น เรียกลำดับเบสนี้ว่ารหัสพันธุกรรม (Genetic Code) อ่านเป็นคำ โดยคำหนึ่งจะประกอบด้วยเบส 3 โมเลกุลเรียงต่อกัน เรียกว่า โคดอน (Codon)

- tRNA หรือ transfer RNA มีขนาดโมเลกุลเล็กๆ ทำหน้าที่นำกรดอะมิโนไปยังบริเวณที่มีการสังเคราะห์โปรตีน คือที่ mRNA เกาะติดกับ ไรโบโซม (Ribosome) tRNA จะมีความจำเพาะเจาะจงกับกรดอะมิโนที่นำไป สิ่งที่กำหนดชนิดของกรดอะมิโน คือ เบส 3 ตัวที่เรียกว่าโคดอน (Codon) ของ mRNA ซึ่ง tRNA แต่ละตัวจะต้องมีลำดับเบส 3 ตัว ที่จับคู่กับโคดอน ที่เรียกว่า แอนติโคดอน (Anticodon)

- rRNA หรือ ribosomal RNA มีปริมาณมากที่สุดในเซลล์ คือ 85% ของ RNA ที่พบในเซลล์เป็นส่วนประกอบของ Ribosome

2.1.1.3 Hybridization

Hybridization เป็นเทคนิคที่อาศัยสมบัติเรื่องการจับคู่เบสอย่างจำเพาะของดีเอ็นเอ (G:C, A:T) โดยที่พันธะไฮโดรเจนที่เชื่อมระหว่างคู่เบสถูกทำลายได้ง่ายด้วยความร้อน หรือการเพิ่ม pH ของสารละลาย เมื่อพันธะไฮโดรเจนถูกทำลายดีเอ็นเอจะคลายเกลียวเปลี่ยนสภาพโมเลกุลจากดีเอ็นเอรูปร่างเกลียวคู่เป็นดีเอ็นเอเส้นเดี่ยว เรียกว่าเกิด Denaturation ของดีเอ็นเอ เมื่อลดอุณหภูมิ หรือลด pH ของสารละลายดีเอ็นเอ ลงสู่สภาพเดิม สายดีเอ็นเอจะกลับมาเข้าคู่กันใหม่ โดยอาศัยการจับคู่เบสอย่างจำเพาะ กระบวนการ Hybridization (สมชาย แสงอำนาจเดช, 2550: ระบบออนไลน์) สามารถแสดงได้ดังภาพประกอบ 2.1



ภาพประกอบ 2.1 แสดงกระบวนการ Hybridization
(สมชาย แสงอำนาจเดช, 2550: ระบบออนไลน์)

2.1.2 ดีเอ็นเอไมโครอาเรย์ (DNA Microarray)

ดีเอ็นเอไมโครอาเรย์เป็นเทคโนโลยีที่ใช้เป็นเครื่องมือในการวิจัยสำหรับวิทยาศาสตร์ชีวภาพและกำลังเป็นที่สนใจในการนำมาใช้ในการช่วยวินิจฉัยทางการแพทย์ ไมโครอาเรย์ประกอบด้วยโมเลกุลชีวภาพ มาจับไว้บนพื้นผิวราบ ซึ่งปกติจะใช้แผ่นแก้วสไลด์เคลือบ แต่อาจเป็นหลุมพลาสติกขนาดเล็ก (Microwells) หรือวัสดุที่เตรียมเป็นเม็ดขนาดเล็ก (Beads) โมเลกุลที่จับบนผิวของสับสเตรทเหล่านี้เรียกว่าโพรบ (Probe) ทำได้โดยการนำโมเลกุลของดีเอ็นเอมาจัดวางเรียงบนพื้นผิวนั้น ดีเอ็นเอเหล่านี้จะเป็นดีเอ็นเอสายเดี่ยว (Single Stranded DNA) ทำหน้าที่เป็นโพรบซึ่งลำดับเบสของดีเอ็นเอนี้เป็นอย่างไรขึ้นกับวัตถุประสงค์ของการศึกษา ถ้าต้องการศึกษาหน้าที่ของยีนๆ หนึ่งก็จะใช้ลำดับเบสในส่วนต่างๆ ของยีนนั้น ถ้าต้องการศึกษาการแสดงออกของยีนๆหนึ่ง ก็จะใช้ดีเอ็นเอของ mRNA นั้น เมื่อนำสารละลายของดีเอ็นเอสายเดี่ยวซึ่งจะมีดีเอ็นเอที่มีลำดับเบสที่สามารถ ไฮบริไดซ์ (Hybridize) กับดีเอ็นเอบนพื้นผิวดีเอ็นเอในสารละลายที่นำมาศึกษาและสามารถจับกับโพรบได้นี้เรียกว่า ดีเอ็นเอเป้าหมาย (Target) เนื่องจากดีเอ็นเอเป้าหมายมักจะติดฉลากด้วยสารเรืองแสง หรือสารกัมมันตรังสี จึงสามารถตรวจหาเชิงปริมาณของดีเอ็นเอสายคู่ที่เกิดจากการจับของโพรบและเป้าหมายได้ ในแผ่นดีเอ็นเอไมโครอาเรย์อาจเรียกว่า แผ่นชิพดีเอ็นเอ (DNA Chip) อาจมียีนจำนวนหลายร้อยหรือหลายพันยีน การนำข้อมูลดีเอ็นเอไมโครอาเรย์มา วิเคราะห์ทำได้โดยการตรวจวัดปริมาณการเรืองแสง หรือ ปริมาณของกัมมันตรังสี (ขึ้นอยู่กับฉลากที่ใช้) และประมวลผลภาพที่สแกนได้ด้วยคอมพิวเตอร์และเก็บข้อมูลการแสดงออกของยีน เพื่อการขุดค้นข้อมูล (Data Mining) ในสกัดความรู้ทางชีวภาพ (Biological Information) (จิราภรณ์ อังวิทยาธร *และคณะ*., 2544; สมชาย แสงอำนาจเดช, 2550: ระบบออนไลน์) ในหัวข้อนี้จะกล่าวถึงรายละเอียดของการใช้งานดีเอ็นเอไมโครอาเรย์ในการวิเคราะห์ตัวอย่าง เพื่อประโยชน์ในการวินิจฉัยโรคต่อไป

2.1.2.1 การใช้งานไมโครอาเรย์ (Using Microarray)

การใช้งานไมโครอาเรย์ประกอบด้วยกระบวนการหลักทั้งหมด 4 ขั้นตอนด้วยกัน (Dov, 2003; นฤพัฒน์, 2008; Lalinka, 2004; Jiang *et al.*, 2004) ดังนี้

1) Sample Preparation and Labeling Process

เริ่มต้นจากการสกัด mRNA (messenger RNA) จากเนื้อเยื่อที่สนใจ โดยทั่วไปในการตรวจวิเคราะห์จะประกอบด้วยตัวอย่างที่ต่างกัน 2 ชนิด ได้แก่ ตัวอย่างอ้างอิงกับตัวอย่างที่จะทดสอบ เช่น การศึกษาการกลายพันธุ์ จะใช้ตัวอย่างที่ทดสอบดูการกลายพันธุ์ เปรียบเทียบกับตัวอย่างอ้างอิง หรือในกรณีศึกษาการแสดงออกของยีน ก็จะใช้ตัวอย่างที่ทดสอบเปรียบเทียบกับตัวอย่างอ้างอิง ซึ่งตัวอย่างทั้งสองนี้จะเป็นตัวอย่างที่อยู่ในภาวะที่แตกต่างกัน เช่น ภาวะไม่เป็นโรคกับภาวะเป็นโรค เป็นต้น (จิราภรณ์ อังวิทยาธร *และ*

คณะ., 2544) ต่อมาจะทำการสร้าง cDNA (complementary DNA) ขึ้นโดยจะนำ mRNA ที่สกัดได้ใส่ลงในตัวกลาง (Medium) ที่เหมาะสม มีฟอสเฟต น้ำตาลดีออกซีไรโบส (Deoxyribose) และเบสชนิดต่าง ๆ ครบครัน พร้อมทั้งเติมเอ็นไซม์ Reverse Transciptase จะเกิดการสร้างสาย DNA สายเดี่ยวคือ cDNA ที่มีการเรียงตัวของเบสสอดคล้อง (Complementary) กับการจัดเรียงตัวของเบสบนสาย mRNA (นฤพัฒน์ ตุลย์ *และ คณะ*., 2008) และขั้นตอนต่อไปจะเข้าสู่ขั้นตอนการติดฉลากให้กับ cDNA จากตัวอย่างทั้งสองชนิดโดยใช้สารเรืองแสง (Fluorescent) 2 สีด้วยกันคือ Cy3 (สีเขียว) และ Cy5 (สีแดง)

2) Hybridization Process

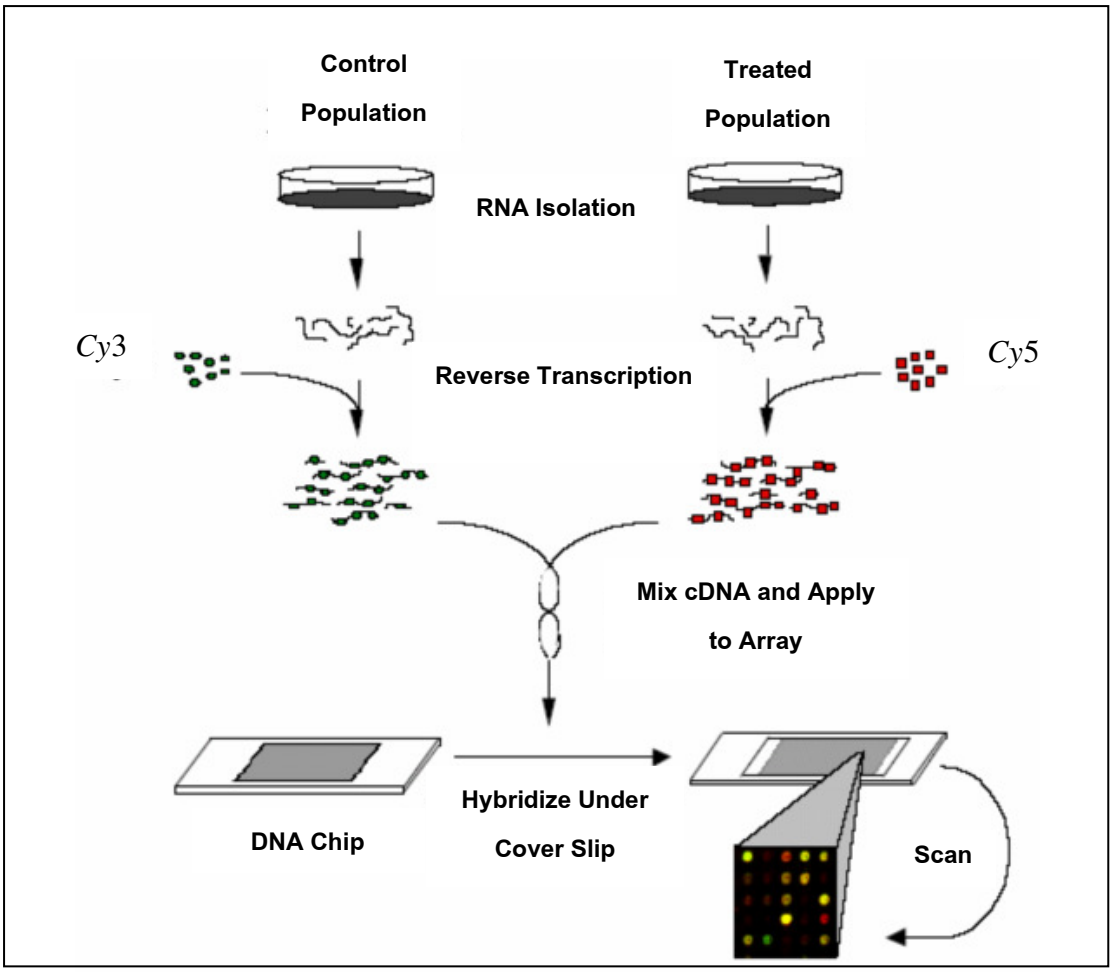
ในขั้นตอนนี้จะทำการผสม cDNA ที่ทำการติดฉลากแล้วจากตัวอย่างทั้ง 2 ชนิด ซึ่งก็คือ นำ cDNA ที่ย้อมด้วยสีเขียวและสีแดงมาผสมกัน ต่อมานำ cDNA ที่ผสมแล้วไป Hybridization กับ DNA Microarray หรือ DNA Chip ที่เตรียมไว้ ซึ่งบน DNA Chip แต่ละจุดจะมี DNA Probes ของยีนที่แตกต่างกัน ในขั้นตอนนี้ cDNA จำนวนหนึ่งจะสามารถเข้าคู่กันกับ DNA Probes ได้ และอีกส่วนหนึ่งที่จับคู่ไม่ได้จะถูกล้างออกไปในขั้นตอนต่อไป (จิราภรณ์ อังวิทยาธร *และคณะ*., 2544)

3) Washing Process

หลังจากกระบวนการ Hybridization แล้ว DNA Chip จะถูกนำไปล้าง เพื่อนำ cDNA ที่ไม่ได้มีการจับคู่ (Unbound) ออกไป เพื่อให้แน่ใจว่าเหลือเฉพาะ cDNA ที่สามารถจับคู่กับ DNA probe ได้เท่านั้น โดยวิธีการล้างจะใช้วิธี Low-Salt Wash หรือ High-Temperature Wash (Dov, 2003)

4) Image Acquisition

เป็นขั้นตอนสุดท้ายในกระบวนการของการใช้ดีเอ็นเอไมโครอาเรย์ เนื่องจากในขั้นตอนก่อนหน้าจะยังไม่สามารถมองเห็นสีใดๆ ได้ ในขั้นตอนนี้จะทำเพื่อให้สามารถมองเห็นสีที่ได้จากขั้นตอนก่อนหน้าเพื่อให้สามารถนำไปประมวลผลต่อไปได้ ขั้นตอนนี้สามารถทำได้โดยนำ DNA Chip ไปไว้ในกล่องที่ไม่มีแสงเพื่อสแกนด้วยแสงเลเซอร์ ซึ่งจะทำการสแกน 2 ครั้งด้วยกัน คือ ครั้งแรกสแกนด้วยแสงเลเซอร์สีเขียว (Green Laser) และทำการเก็บผลลัพธ์ของค่าสีทุกๆ จุดบน DNA Chip ไว้ในโปรแกรมคอมพิวเตอร์ ครั้งที่สองจะสแกนด้วยแสงเลเซอร์สีแดง (Red Laser) และทำการเก็บผลลัพธ์ของค่าสีทุกๆ จุดบน DNA Chip ไว้ในโปรแกรมคอมพิวเตอร์เช่นกัน จะถือว่าขั้นตอนการสแกนด้วยแสงเลเซอร์เสร็จเรียบร้อยแล้ว ต่อมาจะนำภาพที่เก็บไว้ในโปรแกรมคอมพิวเตอร์มาทำการสร้างภาพผสมระหว่างสีทั้งสองในทุกๆ จุดบน DNA Chip ซึ่งภาพที่ได้นี้จะนำไปใช้ในการวิเคราะห์ต่อไป (Malcolm, 2008: Online) ขั้นตอนการใช้งานดีเอ็นเอไมโครอาเรย์สามารถแสดงได้ดังภาพประกอบ 2.2



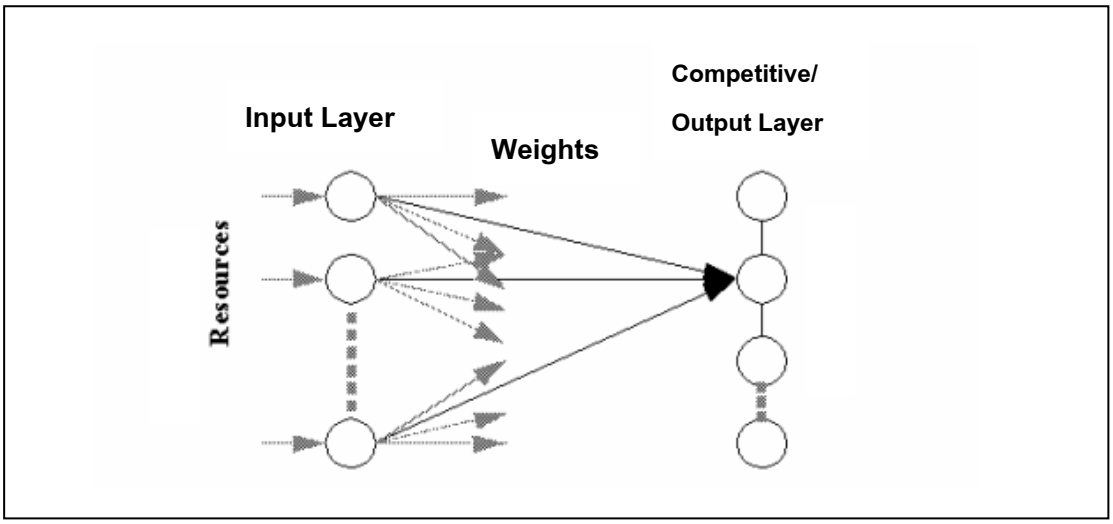
ภาพประกอบ 2.2 แสดงขั้นตอนการใช้งานดีเอ็นเอไมโครอาเรย์ (Bartosh *et al.*, 2004: Online)

2.2 แผนที่การจัดกลุ่มเอง (Self-Organizing Map)

แผนที่การจัดกลุ่มเองเป็นเทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network) ที่มีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ถูกเสนอโดย Kohonen ในปี 1982 ซึ่งเป็นที่นิยมกันอย่างแพร่หลาย โดยมีขอบเขตการใช้งานเช่น การจดจำรูปแบบ (Pattern Recognition) การประมวลผลสัญญาณ (Signal Processing) การทำเหมืองข้อมูล (Data Mining) การวิเคราะห์ข้อมูลทางการเงิน (Financial Analysis) การวิเคราะห์รูปภาพและการมองเห็น (Image Analysis and Vision) (Lalinka *et al.*, 2004; Jiang *et al.*, 2004) และรวมไปถึงการนำมาใช้จัดกลุ่มข้อมูล (Clustering) โดยจะทำการลดมิติของข้อมูลลงให้อยู่ในรูปแบบ 1 หรือ 2 มิติ เพื่อให้สามารถเข้าใจลักษณะของข้อมูลในภาพรวมได้ (Sirirut *et al.*,2007; Siriphan *et al.*, 2007) เนื้อหาในส่วนต่อไปจะกล่าวถึงรายละเอียดเกี่ยวกับสถาปัตยกรรม กระบวนการเรียนรู้ และขั้นตอนการทำงานของ SOM

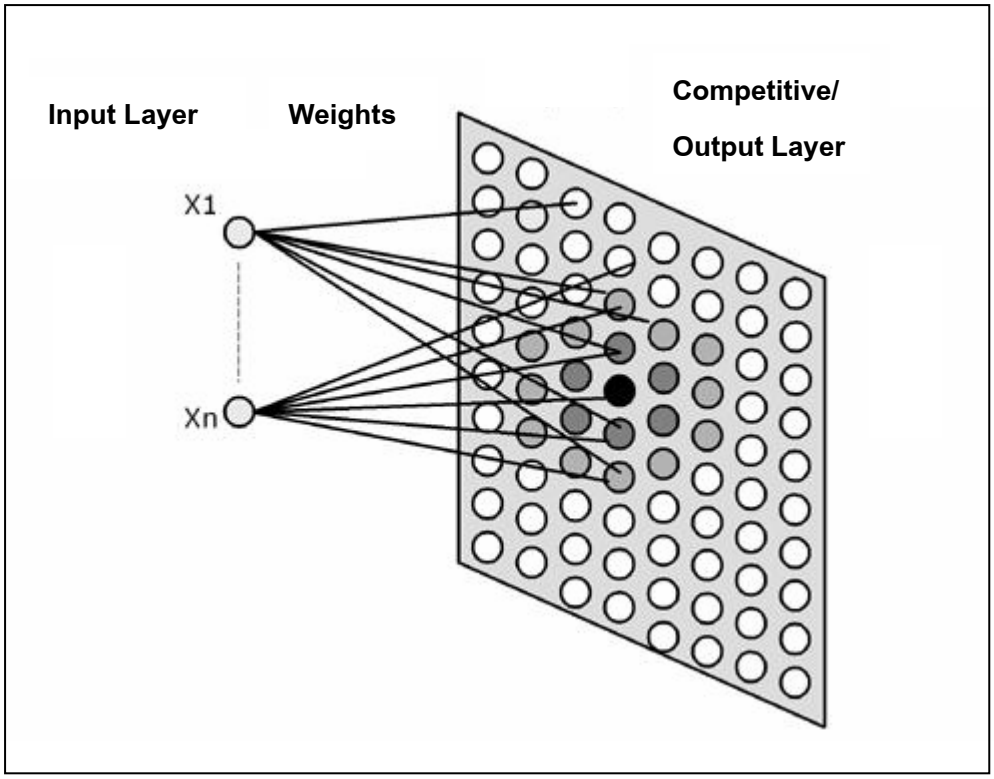
2.2.1 สถาปัตยกรรมของแผนที่การจัดกลุ่มเอง

สถาปัตยกรรมพื้นฐานของ SOM ประกอบด้วย 2 ชั้น คือ ชั้นข้อมูลเข้า (Input Layer) และชั้นข้อมูลออก (Output Layer) โดยกระบวนการจัดกลุ่มทั้งหมดจะอยู่ที่ชั้นข้อมูลออก (Minsky, 2007: Online) โครงสร้างของ SOM ที่มีชั้นข้อมูลออก 1 มิติและ 2 มิติ แสดงได้ดังภาพประกอบ 2.3 และ 2.4 ตามลำดับ



ภาพประกอบ 2.3 แสดงโครงสร้างของ SOM ที่มีชั้นข้อมูลออกเป็น 1 มิติ
(Girardin, 2008: Online)

จากภาพประกอบ 2.3 จะเห็นได้ว่าโครงสร้างของ SOM ประกอบด้วยชั้นข้อมูลเข้า (Input Layer) 1 ชั้น และชั้นข้อมูลออก (Output Layer) 1 ชั้น มีโครงสร้าง 1 มิติ หรือบางครั้งอาจเรียกชั้นข้อมูลออกนี้ว่าชั้นแข่งขัน (Competitive Layer) เนื่องจากการเรียนรู้ของ SOM จะเป็นการเรียนรู้แบบแข่งขัน (Competitive Learning) (Naenna, 2003) และเกิดขึ้นในชั้นข้อมูลออก จึงมีการเรียกชั้นข้อมูลออกว่าชั้นแข่งขัน



ภาพประกอบ 2.4 โครงสร้างของ SOM ที่มีชั้นข้อมูลออกเป็น 2 มิติ
(Koua and Kraak, 2004)

จากภาพประกอบ 2.4 จะเห็นได้ว่าโครงสร้างของชั้นข้อมูลออก (Output Layer) 1 ชั้น มีโครงสร้าง 2 มิติ โดยการจัดกลุ่มข้อมูลจะเกิดขึ้นที่ชั้นข้อมูลออกเช่นกัน

2.2.2 กระบวนการเรียนรู้ของแผนที่การจัดกลุ่มเอง

กระบวนการเรียนรู้ของ SOM เป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คือสามารถเรียนรู้ได้โดยไม่จำเป็นต้องมีตัวอย่างสอน (Naenna, 2003; Wiphada and Putthiporn, 2006) SOM จะทำการแบ่งกลุ่มข้อมูลโดยใช้กระบวนการเรียนรู้แบบแข่งขัน (Competitive Learning)

ในการเรียนรู้แบบแข่งขันนั้นชั้นข้อมูลเข้าและชั้นข้อมูลออกจะถูกเชื่อมต่อกันด้วยเวกเตอร์น้ำหนัก (Weight Vector) และในระหว่างการเรียนรู้จะมีการแทนข้อมูลเข้าไปในโครงข่าย และมีการคำนวณหานิวรอนที่มีลักษณะใกล้เคียงกับเวกเตอร์ข้อมูลเข้ามากที่สุด ซึ่งจะเรียกว่าเป็นนิวรอนที่ชนะในการแข่งขัน ซึ่งการเลือกนิวรอนที่ชนะนี้จะพิจารณาจากการคำนวณระยะทางระหว่างเวกเตอร์ข้อมูลเข้ากับนิวรอนอื่นๆ โดยใช้หลักการของ Euclidean Distance (Jiang *et al.*, 2004) และจะถือว่านิวรอนที่มีระยะทางน้อยที่สุดเป็นนิวรอนที่ชนะ หลังจากนั้นจะ

มีการปรับค่าเวกเตอร์น้ำหนักเพื่อให้นิวรอนที่ชนะมีค่าใกล้เคียงกับข้อมูลเข้า ซึ่งการปรับค่าน้ำหนักนี้จะเรียกว่า “Winner-Takes-All Learning”

การคำนวณระยะทางระหว่างเวกเตอร์ข้อมูลเข้ากับนิวรอนอื่นๆดังสมการที่ (2.1)

$$\|X(t)-W_i\| \tag{2.1}$$

กำหนดให้ X คือ เวกเตอร์ข้อมูลเข้า

W คือ เวกเตอร์น้ำหนักของแต่ละนิวรณ์รอนผลลัพธ์

จากสมการที่ (2.1) เป็นสมการที่ใช้หาระยะทางแบบ Euclidean Distance ดังนั้น นิวรอนที่จะชนะจะสามารถหาได้ดังสมการที่ (2.2)

$$\|X(t)-W_c(t)\| = \min_i \{\|X(t)-W_i(t)\|\} \tag{2.2}$$

กำหนดให้ $W_c(t)$ คือ เวกเตอร์น้ำหนักของนิวรอนที่ชนะ

จากสมการที่ (2.2) จะได้ว่านิวรอนที่ c เป็นนิวรอนที่ชนะ เนื่องจากระยะห่างจากเวกเตอร์ข้อมูลเข้าน้อยที่สุด ซึ่งมี $W_c(t)$ เป็นเวกเตอร์น้ำหนักของนิวรอน c

ต่อมาหลังจากได้นิวรอนที่ชนะแล้ว จะมีการปรับค่าน้ำหนักให้เพื่อใกล้เคียงกับเวกเตอร์ข้อมูลเข้า X ดังสมการที่ (2.3)

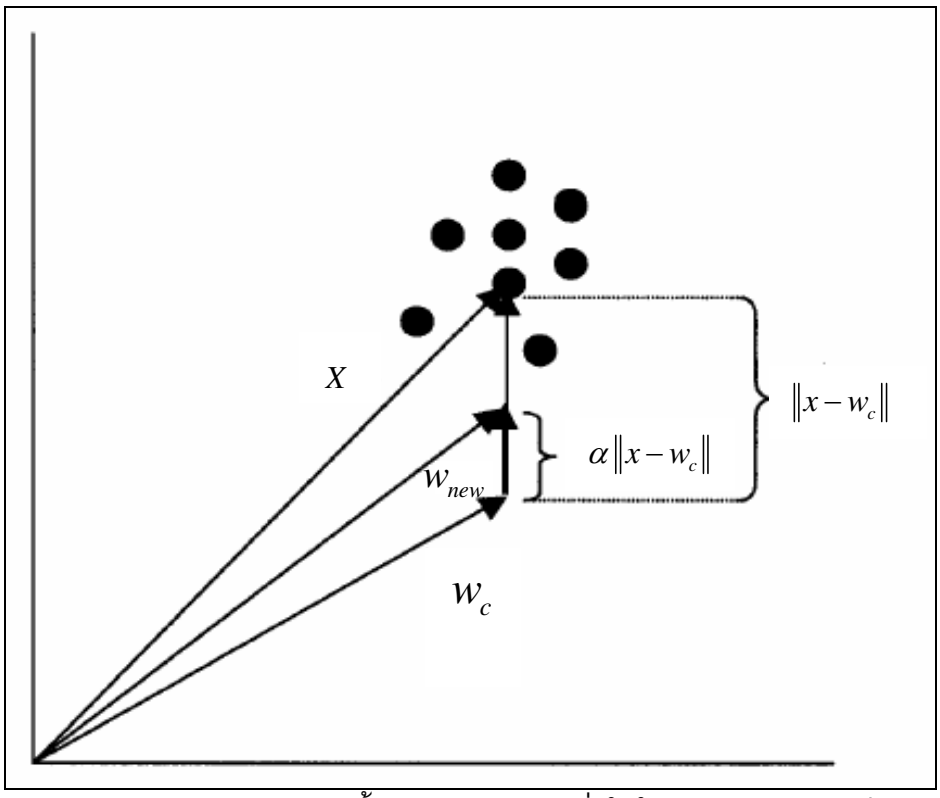
$$W_{new} = W_{old} + \alpha[X - W_{old}] \tag{2.3}$$

กำหนดให้ W_{new} คือ เวกเตอร์น้ำหนักใหม่

W_{old} คือ เวกเตอร์น้ำหนักเก่า

α คือ อัตราการเรียนรู้

จากสมการที่ (2.3) จะมีการปรับค่าน้ำหนักเพื่อให้ใกล้เคียงกับเวกเตอร์ข้อมูลเข้า แสดงได้ดังภาพประกอบ 2.5



ภาพประกอบ 2.5 แสดงการปรับค่าน้ำหนักของนิวรอนเพื่อให้ใกล้เคียงกับเวกเตอร์ข้อมูลเข้า

2.2.3 ขั้นตอนการทำงานของแผนที่การจัดกลุ่มเอง

ขั้นตอนการเรียนรู้ของแผนที่การจัดกลุ่มเองประกอบด้วย 4 ขั้นตอนหลักคือ (Vesanto, 2000) กระบวนการกำหนดค่าน้ำหนักเริ่มต้น (Synaptic Weights Wnitialization) กระบวนการแข่งขัน (Competitive Process) กระบวนการกำหนดโหนดเพื่อนบ้าน (Cooperative Process) และ กระบวนการปรับค่าน้ำหนัก (Adaptive Process) ซึ่งมีรายละเอียดดังนี้

1) กระบวนการกำหนดค่าเริ่มต้น (Synaptic Weights Initialization) ขั้นตอนนี้จะทำการกำหนดค่าเริ่มต้นให้กับเวกเตอร์น้ำหนักของนิวรอนผลลัพธ์ทุกๆ ตัว กำหนดค่าฟังก์ชันเพื่อนบ้าน (Neighborhood Function) และกำหนดอัตราการเรียนรู้ (Learning Rate)

2) กระบวนการแข่งขัน (Competitive Process) ขั้นตอนนี้จะทำการสุ่มเวกเตอร์ข้อมูลเข้าในแต่ละรอบการเรียนรู้ t และคำนวณระยะห่างระหว่างเวกเตอร์ข้อมูลเข้า ดังสมการที่ (2.4) และทำการหานิวรอนที่ชนะซึ่งจะเรียกว่า Best Matching Unit (BMU) ดังสมการที่ (2.5)

$$\|X(t) - W_i(t)\| = \sqrt{\sum_{j=1}^{n} (Xi - Wij)^2} \qquad (2.4)$$

กำหนดให้ $X(t)$ คือ เวกเตอร์ข้อมูลเข้าที่สุ่มในแต่ละรอบการเรียนรู้ t

$W_i(t)$ คือ เวกเตอร์น้ำหนักของนิวรอนอื่นๆ

$$\left\|X(t)-W_c(t)\right\| = \min_i \left\{\left\|X(t)-W_i(t)\right\|\right\} \tag{2.5}$$

กำหนดให้ $X(t)$ คือ เวกเตอร์ข้อมูลเข้าที่สุ่มในแต่ละรอบการเรียนรู้ t

$W_c(t)$ คือ น้ำหนักของนิวรอนที่ชนะ

3) กระบวนการกำหนดโหนดเพื่อนบ้าน (Cooperative Process) เมื่อได้นิวรอนที่ชนะแล้ว นิวรอนที่ชนะจะส่งผลกระทบต่อนิวรอนเพื่อนบ้านในรัศมีหนึ่ง ซึ่งหาได้จากค่าฟังก์ชันเพื่อนบ้าน (Neighborhood Function) ซึ่งเป็นฟังก์ชันเพื่อนบ้านแบบเกาส์เสี้ยน (Gaussian Function) ดังสมการที่ (2.6)

$$h_{ci} = e^{\frac{\|dc-di\|^2}{2\sigma^2(t)}} \tag{2.6}$$

กำหนดให้ $\left\|d_c - d_i\right\|$ คือ ระยะทางระหว่างนิวรอนตัวที่ c ที่เป็น BMU และนิวรอนเพื่อนบ้านตัวที่ i บนตารางแผนที่ชั้นข้อมูลออก

$\sigma(t)$ คือ ค่าความแปรปรวนของนิวรอนเพื่อนบ้านในรอบการเรียนรู้ t

จากสมการที่ (2.6) จะเห็น ความแปรปรวนของนิวรอนเพื่อนบ้านในรอบการเรียนรู้ t จะเป็นตัวควบคุมขนาดของนิวรอนเพื่อนบ้านสำหรับฟังก์ชันเพื่อนบ้านแบบเกาส์เสี้ยน โดยค่าความแปรปรวน $\sigma(t)$ จะมีผลในการควบคุมขนาดของฟังก์ชันเพื่อนบ้านแบบเกาส์เสี้ยน โดยที่ขนาดของฟังก์ชันเพื่อนบ้านแบบเกาส์เสี้ยนจะเล็กลงเมื่อค่าความแปรปรวนลดลง มีผลทำให้ในแต่ละรอบของการเรียนรู้ของ SOM ฟังก์ชันเพื่อนบ้านจะมีขนาดเล็กลงเรื่อยๆ ตามระดับค่าความแปรปรวน จนสุดท้ายเพื่อนบ้านที่อยู่ใกล้เคียง BMU มากที่สุดเท่านั้นที่จะมีการปรับค่าเวกเตอร์น้ำหนัก

4) กระบวนการปรับค่าน้ำหนัก (Adaptive Process) ขั้นตอนนี้เป็นการปรับค่าน้ำหนักของนิวรอนเพื่อนบ้านของ BMU เพื่อให้ใกล้เคียงกับเวกเตอร์ข้อมูลเข้า ดังสมการที่ (2.7)

$$W_i(t+1) = W_i(t) + \alpha(t)h_{ci}(t)\left[X(t)-W_i(t)\right] \tag{2.7}$$

กำหนดให้ $W_i(t+1)$ คือ ค่าเวกเตอร์น้ำหนักใหม่
$\alpha(t)$ คือ อัตราการเรียนรู้
$h_{ci}(t)$ คือ ฟังก์ชันเพื่อนบ้าน

จากสมการที่ (2.7) สำหรับค่าอัตราการเรียนรู้เริ่มต้นมักจะกำหนดให้มีค่าใกล้เคียงกับ 1 หลังจากนั้นอัตราการเรียนรู้จะค่อยๆลดลงในแต่ละรอบการเรียนรู้ ซึ่งโดยส่วนใหญ่จะมีช่วงของอัตราการเรียนรู้คือ $0 < \alpha(t) \leq 1$ หลังจากปรับค่าน้ำหนักใหม่ของ BMU และเพื่อนบ้านของ BMU ในขั้นตอนที่ 4 เรียบร้อยแล้ว เวกเตอร์ข้อมูลเข้าตัวถัดไปจะถูกเลือกเพื่อแทนเข้าไปในชั้นข้อมูลเข้า และทำกระบวนการที่ 1 ถึง 4 อีกครั้ง จนกว่าจำนวนรอบการทำงานจะครบที่กำหนดไว้จึงจะหยุด

2.3 ขั้นตอนวิธีทางสถิติ (Statistics Method)

ในหัวข้อนี้จะกล่าวถึง 5 หัวข้อด้วยกัน คือ การทดสอบสมมติฐานระหว่างค่าเฉลี่ยของประชากร ความคลาดเคลื่อนในการทดสอบสมมติฐาน ระดับความมีนัยสำคัญ ระดับความเชื่อมั่น และความน่าจะเป็น โดยจะอธิบายรายละเอียดทีละหัวข้อตามลำดับ

2.3.1 การทดสอบสมมติฐานระหว่างค่าเฉลี่ยของประชากร

ในกรณีที่มีประชากรสองกลุ่ม ประชากรทั้งสองอาจจะมีการแจกแจงแบบปกติหรือแบบใดก็ได้ถ้าหากสุ่มตัวอย่างจากแต่ละประชากรจะได้สองกลุ่มตัวอย่างที่มาจากสองกลุ่มประชากร ซึ่งตัวอย่างทั้งสองอาจเป็นอิสระต่อกันหรือสัมพันธ์กันก็ได้ การทดสอบสมมติฐานสำหรับสองประชากร อาจจะเป็นการทดสอบความเท่ากันของความแปรปรวนของสองประชากร การทดสอบความเท่ากันของส่วนเบี่ยงเบนมาตรฐานของสองประชากร การทดสอบความแตกต่างระหว่างค่าเฉลี่ยของสองประชากร การทดสอบความต่างระหว่างสัดส่วนของสองประชากร ฯลฯ ซึ่งรายละเอียดการทดสอบสมมติฐานดังกล่าวมีดังนี้

ในงานวิจัยนี้จะใช้หลักของการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของสองประชากร เพื่อประยุกต์ใช้ในขั้นตอนการเลือกยีนโดยอาศัยค่า P-Value (Feature Extraction with P-Value) โดยจะให้ผลลัพธ์คือ สามารถเลือกยีนที่มีค่าความแตกต่างระหว่างค่าเฉลี่ยของสองประชากรมากที่สุด ซึ่งถือว่าเป็นยีนที่มีความสำคัญในการแบ่งกลุ่มผู้ป่วยทั้งสองกลุ่มออกจากกันมากที่สุด โดยเมื่อผ่านขั้นตอนดังกล่าวแล้วจะสามารถเลือกยีนที่มีประสิทธิภาพจำนวนหนึ่งเพื่อทำงานในขั้นตอนต่อไป และยังเป็นการลดเวลาในการทำงานลงอย่างมากอีกด้วย

เนื่องจากไม่จำเป็นจะต้องเสียเวลาในการทำงานกับยีนครบทุกยีน ซึ่งมักจะมีจำนวนมาก (สุรินทร์, 2548)

สมมติว่ามีประชากร 2 ประชากร โดยมีค่า μ_1 และ μ_2 เป็นค่าเฉลี่ยของประชากรตามลำดับ และมี σ_1 และ σ_2 เป็นความแปรปรวนของประชากรทั้งสองตามลำดับ จะได้ว่า $\mu_1 - \mu_2$ คือผลต่างระหว่างค่าเฉลี่ยของประชากรทั้งสอง

การทดสอบสมมติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ยของประชากรทั้งสองนั้น จะมีขั้นตอนการทดสอบเช่นเดียวกับการทดสอบสมมติฐานทั่วไป เพียงแต่ตัวสถิติที่ใช้ในการทดสอบเท่านั้นที่แตกต่างกันไป ในการทดสอบสมมติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ยของประชากรสองกลุ่มนั้น จะทำการสุ่มตัวอย่างจากประชากรแต่ละประชากร การสุ่มตัวอย่างดังกล่าวจะมี 2 กรณีคือ กรณีที่ตัวอย่างทั้งสองเป็นอิสระต่อกัน และกรณีที่ตัวอย่างทั้งสองไม่เป็นอิสระต่อกันหรือมีความสัมพันธ์กัน ซึ่งรายละเอียดของการทดสอบสมมติฐานแต่ละกรณีมีดังนี้

2.3.1.1 กรณีที่ตัวอย่างทั้งสองเป็นอิสระต่อกัน

กรณีตัวอย่างเป็นอิสระต่อกัน (Unpaired Data) สมมติว่าทำการสุ่มตัวอย่างขนาด n_1 และ n_2 หน่วยจากประชากรทั้งสองที่มีขนาด N_1 และ N_2 หน่วยตามลำดับ โดยที่ตัวอย่างทั้งสองเป็นอิสระกัน จากข้อมูลที่เก็บรวบรวมได้ของตัวอย่างทั้งสองจะสามารถคำนวณค่าเฉลี่ยของตัวอย่างคือ $\bar{x}_1$ และ $\bar{x}_2$ และคำนวณได้ความแปรปรวนของสองตัวอย่างคือ S_1^2 และ S_2^2 ตามลำดับ การทดสอบสมมติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ยของประชากรทั้งสองจะมีสมมติฐานในการทดสอบดังนี้

1) กรณีทดสอบสองข้าง (Two Tails)

$H_0 : \ \mu_1 - \mu_2 = \mu_0$ ซึ่ง μ_0 คือ ค่าคงที่

$H_1 : \ \mu_1 - \mu_2 \neq \mu_0$

2) กรณีทดสอบข้างเดียว (ซ้ายมือ)

$H_0 : \ \mu_1 - \mu_2 = \mu_0$ ซึ่ง μ_0 คือ ค่าคงที่

$H_1 : \ \mu_1 - \mu_2 < \mu_0$

3) กรณีทดสอบข้างเดียว (ขวามือ)

$H_0 : \ \mu_1 - \mu_2 = \mu_0$ ซึ่ง μ_0 คือ ค่าคงที่

$H_1 : \ \mu_1 - \mu_2 > \mu_0$

โดยที่ถ้าเราสมมติว่าไม่ทราบ หรือไม่มีหลักฐานมาก่อนเลยที่จะชี้ทางวิชาการว่า ถ้าหาก H_0 ไม่เป็นจริงแล้ว จะหมายความว่า $\mu_1 < \mu_2$ หรือ $\mu_1 > \mu_2$ ในกรณีนี้จะใช้ การทดสอบแบบสองข้าง แต่ถ้าสมมติฐานทางเลือก (H_1) มีเพียงทางเดียว โดยที่ทราบมาก่อนว่าผลอีกทางหนึ่งเป็นไปไม่ได้ หรือเป็นไปได้น้อยมาก นั่นคือ H_1 เป็นไปได้ทาง

เดียวคือ $\mu_1 < \mu_2$ หรือ $\mu_1 > \mu_2$ ในกรณีนี้จะใช้การทดสอบแบบทางเดียว (จรัญ จันทลักขณา และ อนันต์ชัย เขื่อนธรรม, 2540)

การเลือกใช้สถิติในการทดสอบสมมติฐาน กรณีไม่ทราบความแปรปรวนของประชากรทั้งสอง และตัวอย่างทั้งสองมีขนาดเล็ก ถ้าหากตัวอย่างทั้งสองมีขนาดเล็ก ($n_1 < 30$ และ $n_2 < 30$) และไม่ทราบค่าความแปรปรวนของประชากรทั้งสอง (ไม่ทราบค่า σ_1 และ σ_2) แต่ทราบว่าความแปรปรวนของทั้งสองประชากรมีค่าเท่ากัน ($\sigma_1 = \sigma_2$) ซึ่งการทราบว่าความแปรปรวนของประชากรทั้งสองเท่ากันอาจจะทราบจากผลการทดสอบว่าความแปรปรวนของประชากรทั้งสองเท่ากัน แต่ถึงแม้จะทราบว่าความแปรปรวนของทั้งสองประชากรเท่ากันแต่ก็ยังคงไม่ทราบว่าเท่ากันที่ค่าเท่าไหร่ ในกรณีนี้จะใช้ค่าประมาณความแปรปรวนรวมของตัวอย่าง (Pooled Sample Variances) ไปประมาณค่าความแปรปรวนของประชากรที่ไม่ทราบค่า นั่นคือใช้ S_p^2 ซึ่งเป็นค่าความแปรปรวนรวมของตัวอย่างไปประมาณค่าของ σ_1^2 และ σ_2^2 ทั้งคู่ โดยที่ค่า S_p^2 คำนวณได้ดังสมการที่ (2.8) ส่วนสถิติที่ใช้ในการทดสอบคำนวณได้ดังสมการที่ (2.9) ซึ่งมีการแจกแจงแบบที มีองศาความเป็นอิสระเท่ากับ $n_1 + n_2 - 2$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{2.8}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2})}} \tag{2.9}$$

กำหนดให้ $\bar{x}_1,\ \bar{x}_2$ คือ ค่าเฉลี่ยของกลุ่มตัวอย่างกลุ่มที่ 1 และ 2 ที่ถูกสุ่มมาเป็นตัวอย่าง ตามลำดับ

S_p^2 คือ ค่าความแปรปรวนรวมของตัวอย่างที่ถูกสุ่มมา

n_1, n_2 คือ จำนวนสมาชิกของกลุ่มตัวอย่างกลุ่มที่ 1 และ 2 ที่ถูกสุ่มมาตามลำดับ

จากนี้จะแสดงตัวอย่างการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของสองประชากรกรณีข้างต้น ดังนี้ จากรายงานผลการศึกษาถึงอิทธิพลของการสูบบุหรี่ของมารดาที่มีต่อบุตร (สุรินทร์ นิยมางกูร, 2548) พบว่าน้ำหนักแรกเกิดของเด็กที่มารดาไม่สูบบุหรี่และของเด็กที่มารดาสูบบุหรี่ จะมีการแจกแจงแบบปกติ จากการสุ่มตัวอย่างเด็กแรกเกิดที่มารดาไม่สูบบุหรี่และมารดาสูบบุหรี่มาจำนวน 25 และ 16 คน ตามลำดับ ตรวจสอบน้ำหนักของเด็กแรกเกิดและคำนวณได้น้ำหนักแรกเกิดเฉลี่ยเท่ากับ 3,480 และ 3,256 กรัม และได้ค่า

ความแปรปรวนของน้ำหนักเด็กแรกเกิดเป็น 75.34 และ 121.44 (กรัม)^2 จงทดสอบว่าน้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่จะมากกว่าน้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาสูบบุหรี่หรือไม่ ถ้าหากว่าทราบว่าความแปรปรวนของน้ำหนักเด็กแรกเกิดของเด็กที่มารดาไม่สูบบุหรี่และของเด็กที่มารดาสูบบุหรี่เท่ากัน ใช้ระดับนัยสำคัญเท่ากับ 0.01

วิธีทำ

กำหนดให้

μ_1 คือ น้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่ทั้งหมด

μ_2 คือ น้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาสูบบุหรี่ทั้งหมด

σ_1^2 คือ ความแปรปรวนของน้ำหนักของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่ทั้งหมด

σ_2^2 คือ ความแปรปรวนของน้ำหนักของเด็กแรกเกิดที่มารดาสูบบุหรี่ทั้งหมด

n_1 คือ จำนวนของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่ที่ถูกสุ่มมาเป็นตัวอย่างเท่ากับ 25 คน

n_2 คือ จำนวนของเด็กแรกเกิดที่มารดาสูบบุหรี่ที่ถูกสุ่มมาเป็นตัวอย่างเท่ากับ 16 คน

$\bar{x}_1$ คือ น้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่ที่ถูกสุ่มมาเป็นตัวอย่างเท่ากับ 3,480 กรัม

$\bar{x}_2$ คือ น้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาสูบบุหรี่ที่ถูกสุ่มมาเป็นตัวอย่างเท่ากับ 3,256 กรัม

S_1^2 คือ ความแปรปรวนของน้ำหนักของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่ที่ถูกสุ่มมาเป็นตัวอย่าง = 75.34 (กรัม)^2

S_2^2 คือ ความแปรปรวนของน้ำหนักของเด็กแรกเกิดที่มารดาสูบบุหรี่ที่ถูกสุ่มมาเป็นตัวอย่าง = 121.44 (กรัม)^2

ในการทดสอบว่าน้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่จะมากกว่าน้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาสูบบุหรี่หรือไม่ คือการทดสอบว่า $\mu_1 > \mu_2$ หรือไม่ ซึ่งก็คือการทดสอบว่า $\mu_1 - \mu_2 > 0$ หรือไม่นั่นเอง รายละเอียดของการทดสอบจะมีทั้งหมด 4 ขั้นตอน คือ

1) สมมติฐานในการทดสอบคือ

$H_0 : \mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2)$

$H_1 : \mu_1 - \mu_2 > 0 \quad (\mu_1 > \mu_2)$ (การทดสอบข้างเดียวด้านขวามือ)

2) การคำนวณค่าสถิติที่ใช้ทดสอบ ในที่นี้ไม่ทราบความแปรปรวนของน้ำหนักของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่และของเด็กแรกเกิดที่มารดาสูบบุหรี่ แต่ทราบว่าความแปรปรวนของทั้งสองมีค่าเท่ากัน ดังนั้นจะประมาณค่าความแปรปรวนทั้งสองด้วยความแปรปรวนของตัวอย่างรวมคือ

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(25-1)(75.34)+(16-1)(121.44)}{25+16-2}$$

$$= \frac{3,629.76}{39} = 93.0708$$

จะได้ $s_p = \sqrt{93.0708} = 9.6473$

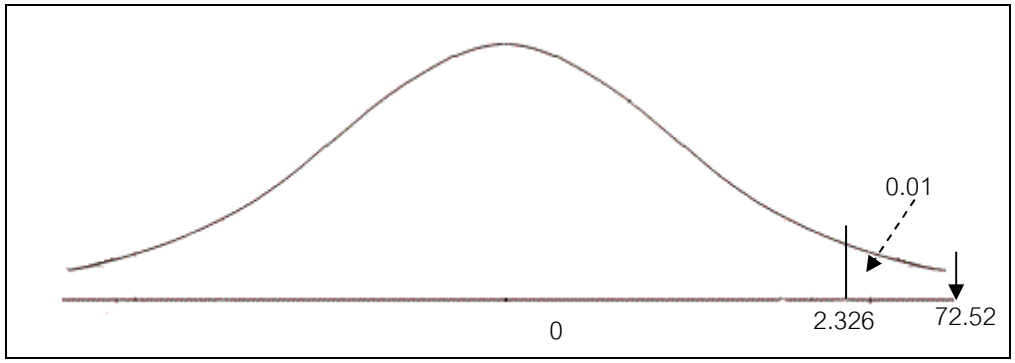
จะได้ค่าสถิติที่ใช้ทดสอบคือ

$$t = \frac{(\bar{x}_1-\bar{x}_2)-\mu_0}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2})}} = \frac{(3,480-3,256)-0}{(6.9473)\sqrt{\frac{1}{25}+\frac{1}{16}}}$$

$$= \frac{224}{3.0886}$$

$$= 72.52$$

3) กำหนดระดับนัยสำคัญ (α) เท่ากับ 0.01 และ $d.f = n_1+n_2-2 = 25+16-2 = 39$ เนื่องจากเป็นการทดสอบข้างเดียวด้านขวามือ จะได้บริเวณวิกฤตคือ $t > 2.236$ ซึ่งสามารถแสดงบนเส้นโค้งได้ดังภาพประกอบ 2.6



ภาพประกอบ 2.6 แสดงผลจากการคำนวณบนเส้นโค้งปกติ

4) สรุปผลการทดสอบ เนื่องจาก t ที่คำนวณได้เท่ากับ 72.52 ตกในบริเวณวิกฤต ดังนั้นผลสรุปคือ ปฏิเสธ H_0 และยอมรับ $H_1: \mu_1-\mu_2 > 0$ หรือ $\mu_1 > \mu_2$ แสดงว่าน้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาไม่สูบบุหรี่จะมากกว่าน้ำหนักเฉลี่ยของเด็กแรกเกิดที่มารดาสูบบุหรี่ ซึ่งจากค่า t ที่คำนวณได้เท่ากับ 72.52 และเนื่องจากเป็นการทดสอบข้างเดียว

จะได้ค่า P-Value = 0.0000 ซึ่งมีค่าน้อยกว่า $\alpha = 0.01$ ผลสรุปคือ ปฏิเสธ H_0 และยอมรับ H_1 เช่นเดียวกัน (สุรินทร์ นิยมางกูร, 2548)

2.3.1.2 กรณีที่ตัวอย่างทั้งสองสัมพันธ์กัน

กรณีตัวอย่างสัมพันธ์กัน (Paired Data) ในบางครั้ง ตัวอย่าง 2 ตัวอย่างที่จะนำมาทดสอบผลต่างระหว่างค่าเฉลี่ยของประชากรนั้น อาจเป็นตัวอย่างที่ไม่เป็นอิสระต่อกัน กล่าวคืออาจเป็นตัวอย่างที่สัมพันธ์กันหรือเกี่ยวข้องกัน เช่น ระดับคลอเรสเตอรอลของคนก่อนและหลังการควบคุมอาหาร ปริมาณการขายสินค้าของบริษัทก่อนและหลังการเปลี่ยนผู้บริหารใหม่ เป็นต้น หรือตัวอย่างทั้งสองเป็นตัวอย่างที่มีความคล้ายคลึงกัน เช่น ข้อมูลของพี่และน้องของครอบครัวเดียวกัน ข้อมูลของเด็ก 2 กลุ่มที่มีอายุ และเพศเดียวกัน เป็นต้น ข้อมูลในลักษณะนี้เรียกว่า “ข้อมูลชนิดจับคู่กัน” (สุรินทร์ นิยมางกูร, 2548) การทดสอบเกี่ยวกับสองตัวอย่างที่สัมพันธ์กันนี้จะมีประสิทธิภาพสูงกว่าการทดลองเปรียบเทียบแบบอิสระต่อกันหรือแบบรวมกลุ่ม เพราะจะสามารถลดความคลาดเคลื่อนอันเกิดจากความแตกต่างระหว่างสิ่งทดลองได้มากกว่า (จรัญ จันทลักขณา และ อนันต์ชัย เขื่อนธรรม, 2540) การทดสอบเกี่ยวกับตัวอย่าง 2 ตัวอย่างที่สัมพันธ์กันนี้จะแตกต่างไปจากการทดสอบเกี่ยวกับตัวอย่าง 2 ตัวอย่างที่เป็นอิสระต่อกัน การทดสอบผลต่างระหว่างค่าเฉลี่ยของประชากรในกรณีที่ข้อมูลสัมพันธ์กันนี้ ขนาดของตัวอย่างทั้งสองจะต้องเท่ากัน ข้อมูลของตัวอย่างลักษณะนี้เป็นดังตารางที่ 2.1

ตารางที่ 2.1 แสดงตัวอย่างของข้อมูลที่สัมพันธ์กัน (Paired Data)

คู่ที่ i	ตัวอย่างที่ 1 (x_i)	ตัวอย่างที่ 2 (y_i)
1	x_1	y_1
2	x_2	y_2
.	.	.
.	.	.
.	.	.
n	x_n	y_n

โดยที่ $x_1, x_2, ..., x_n$ เป็นข้อมูลของตัวอย่างที่สุ่มมาจากประชากร ซึ่งมีค่าเฉลี่ยเท่ากับ μ_1 และมีความแปรปรวนเท่ากับ σ_1^2 และ $y_1, y_2, ..., y_n$ เป็นข้อมูลของตัวอย่างที่สุ่มมาจากประชากร ซึ่งมีค่าเฉลี่ยเท่ากับ μ_2 และมีความแปรปรวนเท่ากับ σ_2^2

สิ่งที่ต้องการทดสอบในที่นี้คือ μ_1 - μ_2 ซึ่งเป็นผลต่างระหว่างค่าเฉลี่ยของประชากรทั้งสอง ถ้าหากกำหนดให้ $D_i = X_i - Y_i$ จะได้ว่า $\mu_D = \mu_1 - \mu_2$ ดังนั้นการทดสอบเดียวกับ $\mu_1 - \mu_2$ ก็คือการทดสอบเกี่ยวกับ μ_D นั่นเอง สมมติฐานที่ต้องการทดสอบในที่นี้มี 3 กรณีดังนี้

1) กรณีการทดสอบ 2 ข้าง

$H_0 : \mu_D = d_0$ เมื่อ d_0 คือ ค่าคงที่

$H_1 : \mu_D \neq d_0$

2) กรณีทดสอบข้างเดียวด้านซ้ายมือ

$H_0 : \mu_D = d_0$

$H_1 : \mu_D < d_0$

3) กรณีทดสอบข้างเดียวด้านขวามือ

$H_0 : \mu_D = d_0$

$H_1 : \mu_D > d_0$

ตัวสถิติที่ใช้ในการทดสอบดังสมการที่ (2.10) – (2.13)

$$t = \frac{\bar{D} - d_0}{\frac{S_D}{\sqrt{n}}} \quad , \quad d.f. = n-1 \tag{2.10}$$

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{2.11}$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} D_i^2 - n\bar{D}^2] \tag{2.12}$$

$$S_D = \sqrt{S_D^2} \tag{2.13}$$

ตัวสถิติที่ใช้ทดสอบนี้จะมีการแจกแจงแบบที ที่มีองศาความเป็นอิสระเท่ากับ $n-1$ ตัวอย่างการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของสองประชากรกรณีนี้แสดงได้ดังตัวอย่างดังนี้ จากการตรวจระดับโคเลสเตอรอลก่อนและหลังการออกกำลังกายของคน 11 คน ได้รายละเอียดข้อมูลดังตารางที่ 2.2

ตารางที่ 2.2 แสดงข้อมูลจากการตรวจระดับโคเลสเตอรอลก่อนและหลังการออกกำลังกาย

คนที่	ระดับโคเลสเตอรอล		ผลต่าง
i	ก่อนออกกำลังกาย (X_i)	หลังออกกำลังกาย (Y_i)	$D_i = X_i - Y_i$
1	182	198	-16
2	232	210	22
3	191	194	-3
4	200	220	-20
5	148	138	10
6	249	220	29
7	276	219	57
8	213	161	52
9	241	210	31
10	380	313	67
11	262	226	36

จงทดสอบว่าระดับโคเลสเตอรอลเฉลี่ยก่อนการออกกำลังกายสูงกว่าระดับโคเลสเตอรอลเฉลี่ยหลังออกกำลังกายหรือไม่ ใช้ระดับนัยสำคัญ 0.05

วิธีทำ

กำหนดให้ μ_1 คือ ระดับโคเลสเตอรอลเฉลี่ยก่อนออกกำลังกาย

μ_2 คือ ระดับโคเลสเตอรอลเฉลี่ยหลังการออกกำลังกาย

n คือ จำนวนคนที่ถูกสุ่มมาเป็นตัวอย่าง = 11 คน

จาก $D_i = X_i - Y_i$ เมื่อ $i = 1,2,3,...,11$ ดังนั้นจะได้ $\mu_D = \mu_1 - \mu_2$

ในการทดสอบว่าระดับโคเลสเตอรอลเฉลี่ยก่อนออกกำลังกายสูงกว่าระดับโคเลสเตอรอลเฉลี่ยหลังออกกำลังกายหรือไม่ คือการทดสอบว่า $\mu_1 > \mu_2$ หรือไม่ ซึ่งก็คือการทดสอบว่า $\mu_1 - \mu_2 > 0$ หรือไม่นั่นเอง รายละเอียดของการทดสอบมีดังนี้

1) สมมติฐานในการทดสอบคือ

$H_0 : \mu_D = d_0$

$H_1 : \mu_D > d_0$ (การทดสอบข้างเดียวด้านขวามือ)

2) การคำนวณค่าสถิติที่ใช้ในการทดสอบ ในที่นี้ระดับโคเลสเตอรอลก่อนการออกกำลังกายและระดับโคเลสเตอรอลหลังการออกกำลังกายของคน 11 คน เป็นข้อมูลชนิดจับคู่กัน กำหนดให้ D_i เป็นผลต่างระหว่างระดับโคเลสเตอรอลก่อนการออกกำลังกายและหลังการออกกำลังกายจะได้ $D_i = X_i - Y_i$

ดังนั้น จะคำนวณได้ $\bar{D}$ = $\frac{1}{n}\sum_{i=1}^{n} D_i$

= $\frac{265}{11}$

= 24.0909

S_D^2 = $\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$

= $\frac{1}{n-1}[\sum_{i=1}^{n} D_i^2 - n\bar{D}^2]$

= $\frac{1}{10}[14,789 - (11)(24.0909)^2]$

= $\frac{8,404.9139}{10}$

= 840.4914

และ S_D = $\sqrt{S_D^2}$

= $\sqrt{840.4914}$

= 28.9912

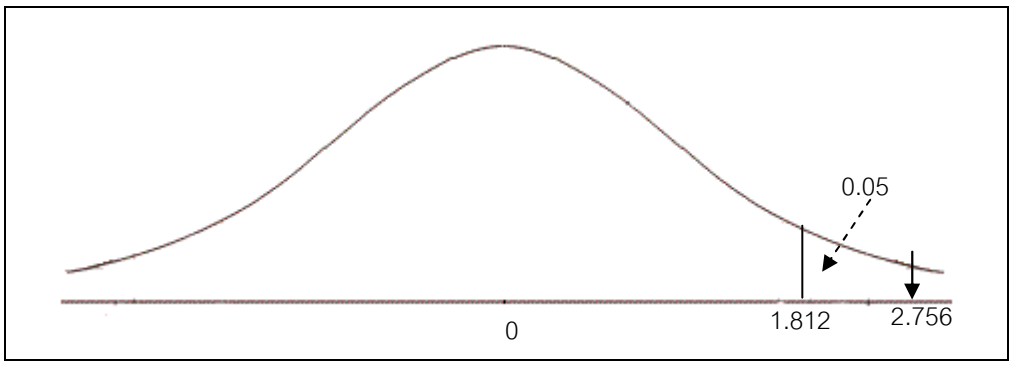
ค่าสถิติที่ใช้ในการทดสอบคือ

t = $\frac{\bar{D} - d_0}{\frac{s_D}{\sqrt{n}}}$

= $\frac{24.0909 - 0}{\frac{(28.9912)}{\sqrt{11}}}$

= 2.756

3) กำหนดระดับนัยสำคัญ (μ) เท่ากับ 0.05 และค่าองศาความเป็นอิสระเท่ากับ $n-1=11-1=10$ เนื่องจากเป็นการทดสอบข้างเดียวด้านขวามือ จะได้บริเวณวิกฤตคือ $t>1.812$ ซึ่งสามารถแทนค่าลงบนเส้นโค้งดังภาพประกอบ 2.7



ภาพประกอบ 2.7 แสดงการแทนค่าจากการคำนวณลงบนเส้นโค้ง

4) สรุปผลการทดสอบ เนื่องจาก t ที่คำนวณได้เท่ากับ 2.756 และการทดสอบเป็นการทดสอบข้างเดียว จะได้ P-Value = 0.010 ซึ่งมีค่าน้อยกว่า $\alpha = 0.05$ ผลสรุปคือ ปฏิเสธ H_0 และยอมรับ H_1 เช่นเดียวกัน (สุรินทร์ นิยมางกูร, 2548)

2.3.2 ความคลาดเคลื่อนในการทดสอบสมมติฐาน

การทดสอบสมมติฐานเป็นเรื่องของการตัดสินใจ โดยที่การตัดสินใจนั้นถือเอาจากหลักฐานที่ได้มาจากตัวแทน ฉะนั้นการตัดสินใจจึงอาจมีการผิดพลาดเกิดขึ้นได้ (จรัญ จันทลักขณา และ อนันต์ชัย เขื่อนธรรม, 2540) การทดสอบสมมติฐานเป็นการตัดสินใจยอมรับหรือปฏิเสธสมมติฐาน ซึ่งขึ้นอยู่กับข้อมูลจากกลุ่มตัวอย่าง ดังนั้นการตัดสินใจอาจเกิดความผิดพลาดขึ้นได้ โดยทั่วไปโอกาสที่จะเกิดการตัดสินใจผิดพลาดมี 2 ลักษณะดังนี้

2.3.2.1 ความผิดพลาดประเภทที่ 1 คือการตัดสินใจปฏิเสธสมมติฐานว่าง (H_0) ทั้งๆ ที่สมมติฐานนั้นเป็นจริง แทนด้วยสัญลักษณ์แอลฟา (α)

2.3.2.2 ความผิดพลาดประเภทที่ 2 คือการตัดสินใจยอมรับสมมติฐานว่าง (H_0) ทั้งๆ ที่สมมติฐานว่างนั้นไม่เป็นจริง แทนด้วยสัญลักษณ์เบตา (β)

จากการตัดสินใจในการทดสอบสมมติฐานทางสถิติดังกล่าว สามารถสรุปได้ดังตารางที่ 2.3

ตารางที่ 2.3 สรุปความผิดพลาดในการทดสอบสมมติฐาน

การตัดสินใจ	ความเป็นจริง	
	H_0 เป็นจริง	H_0 ไม่เป็นจริง
ยอมรับ H_0	ตัดสินใจถูกต้อง	เกิดความผิดพลาดประเภทที่2
ปฏิเสธ H_0	เกิดความผิดพลาดประเภทที่ 1	ตัดสินใจถูกต้อง

ในการทดสอบสมมติฐานจะเห็นได้ว่ามีโอกาสที่จะตัดสินใจผิดพลาด 2 ลักษณะ คือ การปฏิเสธสมมติฐานที่เป็นจริงทำให้เกิดความผิดพลาดประเภทที่ 1 (α) และการยอมรับสมมติฐานที่ไม่จริงทำให้เกิดความผิดพลาดแบบที่ 2 (β) ดังนั้นผู้ทดสอบจะต้องพยายามลดความคลาดเคลื่อนดังกล่าวให้น้อยที่สุด แต่ในขณะเดียวกันก็ไม่สามารถลดความผิดพลาดทั้ง 2 ลักษณะพร้อมๆ กันได้ เมื่อลดความผิดพลาดลักษณะหนึ่งให้น้อยลง ก็จะทำให้ความผิดพลาดอีกลักษณะหนึ่งเพิ่มขึ้น กล่าวคือถ้าให้ α ลดลง ค่า β ก็จะเพิ่มขึ้น ถ้า β ลดลง ค่า α ก็จะเพิ่มขึ้น ในทางปฏิบัติให้ลดความคลาดเคลื่อนที่พิจารณาเห็นว่ามีความสำคัญมากกว่า และในการทดสอบสมมติฐานทั่วๆ ไป จะลดค่า α แล้วพยายามให้ค่า β มีค่าต่ำสุดเท่าที่จะทำได้ กล่าวคือ ให้ขนาดของ α แล้วเลือกการทดสอบที่ทำให้อำนาจการทดสอบสมมติฐาน $(1-\beta)$ มากที่สุด (กานดา พูนลาภทวี, 2539)

2.3.3 ระดับความมีนัยสำคัญ (Level of Significance)

ระดับความมีนัยสำคัญ หมายถึง ความน่าจะเป็นในการปฏิเสธสมมติฐานว่างที่เป็นจริง หรือความน่าจะเป็นในการเกิดความคลาดเคลื่อนประเภทที่ 1 (α) การกำหนดระดับความมีนัยสำคัญที่ยอมให้เกิดความคลาดเคลื่อนนั้น จะขึ้นอยู่กับธรรมชาติของปัญหาหรือหัวข้อเรื่องในการวิจัย ในกรณีที่เป็นเรื่องซึ่งถ้ามีการตัดสินใจผิดพลาดแล้วเกิดอันตรายหรือความเสียหายอย่างร้ายแรง จะกำหนดระดับความมีนัยสำคัญต่ำ เช่น การทดลองยาชนิดหนึ่งว่าจะมีพิษต่อผู้บริโภคหรือไม่ ก็อาจกำหนดให้ $\alpha = 0.01$ (กานดา พูนลาภทวี, 2539)

2.3.4 ระดับความเชื่อมั่น (Level of Confidence)

ระดับความเชื่อมั่นหมายถึง ความน่าจะเป็นในการยอมรับสมมติฐานว่างที่เป็นจริง ดังนั้นระดับความเชื่อมั่นจะเท่ากับ $1-\alpha$ เช่น ถ้า $\alpha = 0.01$ จะมีระดับความเชื่อมั่นเท่ากับ 0.99 หรือ 99% หมายความว่าถ้าดำเนินการทดสอบสมมติฐานนั้นซ้ำๆ กันจำนวน 100 ครั้ง โอกาสที่จะเกิดความผิดพลาดเพียง 1 ครั้งเท่านั้น (กานดา พูนลาภทวี, 2539)

2.3.5 ค่าความน่าจะเป็น (Probability Value: P-Value)

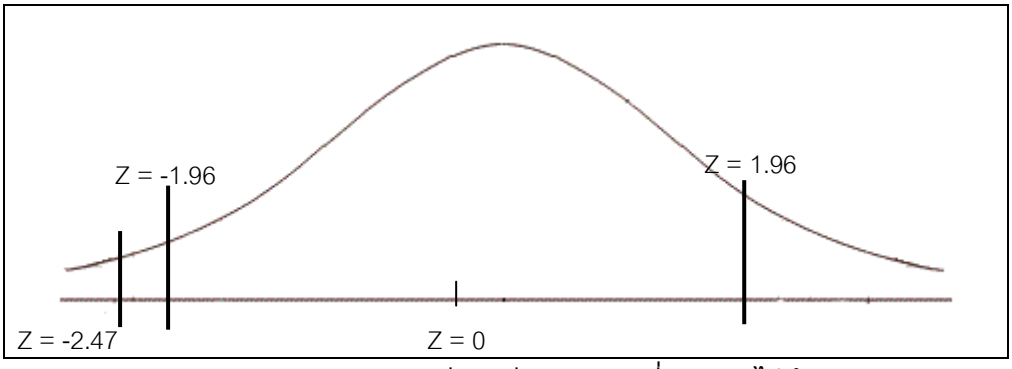
คือความน่าจะเป็นในการปฏิเสธสมมติฐานว่างที่เป็นจริง (Buchan, 2008: Online) ที่นิยมใช้ในการแปรผลการทดสอบสมมติฐานในปัจจุบัน ซึ่งการแปรผลในการทดสอบสมมติฐานโดยปกตินั้นเราจะใช้วิธีการเปรียบเทียบค่าสถิติที่คำนวณได้กับค่าวิกฤตที่เปิดจากตารางสถิติตามระดับนัยสำคัญ (α) ที่กำหนดไว้ (นภดล สุชาติ, 2008: ระบบออนไลน์) เช่น

ในการทดสอบสมมติฐานดังนี้

$$H_0 : \mu = 35$$

$$H_a : \mu \neq 35$$

ซึ่งกำหนดค่า Significance = 0.05 ซึ่งใช้สถิติในการทดสอบคือ Z-Test และสามารถคำนวณค่า Z = -2.47 จะได้ดังภาพประกอบ 2.8



ภาพประกอบ 2.8 แสดงการเปรียบเทียบค่าสถิติที่คำนวณได้กับค่าวิกฤต

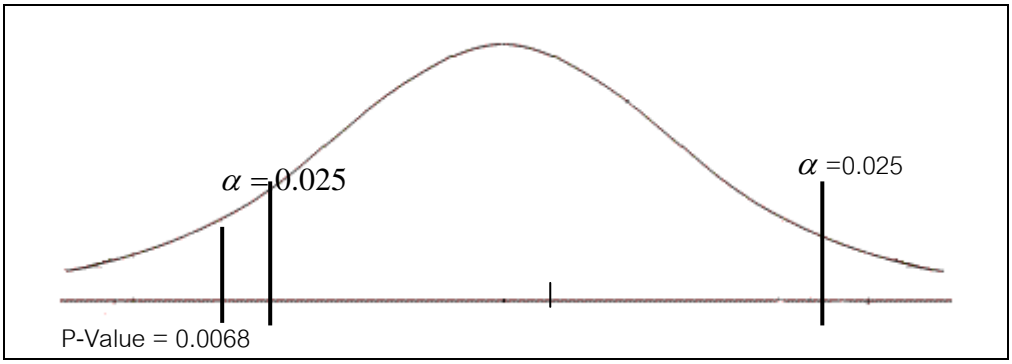
จากเส้นโค้งปกติจะได้ว่า ค่า Z ที่เป็นช่วงวิกฤตคือ Z < -1.96 และ Z > 1.96 ซึ่งได้มาจากการเปิดตารางการแจกแจง Z และค่า Z ที่คำนวณได้คือ -2.47 ตกอยู่ในบริเวณวิกฤต ดังนั้นจะสามารถสรุปได้ว่าปฏิเสธ $H_0 : \mu = 35$ และยอมรับ $H_1 : \mu \neq 35$ (Gambhir, 2006: Online) ซึ่งวิธีการข้างต้นเป็นวิธีการที่นักวิจัยในเมืองไทยส่วนใหญ่นิยมใช้กัน หรือบางครั้งก็อาจจะคำนวณค่า P-Value ออกมา แล้วนำไปเปรียบเทียบกับค่า α ที่กำหนดไว้ โดยถ้า ค่า P-Value น้อยกว่า ค่า α ก็จะปฏิเสธ $H_0 : \mu = 35$ และยอมรับ $H_1 : \mu \neq 35$ เช่นกัน ดังนี้ (ชาญชัยณรงค์ ทรงคาศรี, 2550: ระบบออนไลน์)

ในการทดสอบสมมติฐานดังนี้

$$H_0 : \mu = 35$$

$$H_a : \mu \neq 35$$

ซึ่งกำหนดค่า Significance = 0.05 ใช้สถิติในการทดสอบคือ Z-Test และสามารถคำนวณค่า Z = -2.47 จะแสดงได้ดังภาพประกอบ 2.9



ภาพประกอบ 2.9 แสดงค่า P-Value กับค่านัยสำคัญบนเส้นโค้ง

จากการคำนวณจะได้ค่า P-Value = 0.0068 แต่เป็นสมมติฐานแบบสองข้าง ดังนั้นจะได้ P-Value = 2*0.0068 = 0.0136 จะได้ว่าค่า P-Value < α ดังนั้นจึงสรุปว่าปฏิเสธ $H_0 : \mu = 35$ และยอมรับ $H_1 : \mu \neq 35$ (Gambhir, 2006: Online)

แต่ในปัจจุบันนักวิจัยในสหรัฐอเมริกามีแนวโน้มว่าจะรายงานผลการวิจัยเฉพาะค่า P-Value ที่คำนวณได้เท่านั้น และสรุปว่าจะปฏิเสธ H_0 ในช่วงเท่าใด โดยไม่ได้มีการกำหนดค่า α ไว้ล่วงหน้าแต่อย่างใด (ชาญชัยณรงค์ ทรงคาศรี, 2550: ระบบออนไลน์) ซึ่งโดยปกติแล้วการจะตัดสินใจปฏิเสธ H_0 โดยอาศัยค่า P-Value เท่านั้นมักจะยึดค่าดังตารางที่ 2.4 เป็นเกณฑ์ (Gambhir, 2006: Online)

ตารางที่ 2.4 แสดงการแปรผลการทดสอบสมมติฐานโดยอาศัยค่า P-Value

The P-Value Interpretation Table	
Less than 0.01	Strong evidence against Null Hypothesis, Very statistically significant
0.01 – 0.05	Some evidence against Null Hypothesis, statistically significant
Greater than 0.05	Insufficient evidence against Num Hypothesis

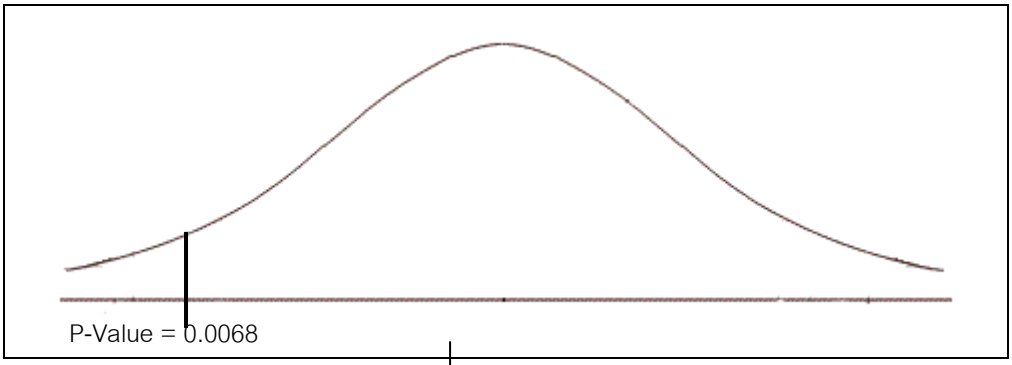
ตัวอย่างการแปรผลการทดสอบสมมติฐานโดยอาศัยค่า P-Value เท่านั้น สามารถแสดงได้ดังนี้

ในการทดสอบสมมติฐานดังนี้

$$H_0 : \mu = 35$$

$$H_a : \mu \neq 35$$

ซึ่งไม่มีการกำหนดค่า α ใช้สถิติในการทดสอบคือ Z-Test และสามารถคำนวณค่า Z = -2.47 จะแสดงได้ดังภาพประกอบ 2.10



ภาพประกอบ 2.10 แสดงการทดสอบสมมติฐานโดยอาศัยค่า P-Value อย่างเดียว

จากการคำนวณจะได้ค่า P-Value = 0.0068 แต่เป็นสมมติฐานแบบสองข้าง จะได้ P-Value = 2*0.0068 = 0.0136 จะเห็นว่าค่า 0.01<P-Value<0.05 ดังนั้นจึงสรุปว่าปฏิเสธ $H_0: \mu = 35$ และยอมรับ $H_1: \mu \neq 35$ (Gambhir, 2006: Online)

จากวิธีที่แตกต่างกันในการสรุปผลการทดสอบสมมติฐานข้างต้น จะเห็นว่าการทดสอบสมมติฐานทางสถิติในความคิดของนักวิจัยส่วนใหญ่จะมีอยู่ 2 กลุ่มดังนี้ (ชาญชัยณรงค์ ทรงคาศรี, 2550: ระบบออนไลน์)

กลุ่มที่ 1 นักวิจัยกลุ่มนี้จะกำหนดค่า α ไว้ก่อน ซึ่งค่า α ที่นิยมใช้กันก็คือ 0.05 0.01 และ 0.10 ตามลำดับ แล้วนำค่า P-Value ที่คำนวณได้ไปเปรียบเทียบกับค่า α ที่กำหนด หรือนำค่าสถิติที่คำนวณได้ไปเปรียบเทียบกับค่าวิกฤติที่เปิดจากตารางการแจกแจงสถิติ ที่ระดับ α ที่กำหนด ถ้าค่า P-Value มีค่าน้อยกว่า α หรือค่าสถิติที่คำนวณได้มีค่าตกอยู่ในบริเวณวิกฤติ ผลการทดสอบจะปฏิเสธสมมติฐาน H_0 ที่ตั้งไว้ มิฉะนั้นแล้วจะยอมรับสมมติฐาน H_0ตัวอย่างเช่น ในการทดสอบความแปรปรวนตัวอย่างหนึ่ง กำหนดให้ค่า α = 0.05 สมมติคำนวณค่า P-Value ได้เท่ากับ 0.04 เนื่องจากค่า P-Value < α ดังนั้นจึงปฏิเสธ H_0 ในทางกลับกันถ้า P-Value ที่คำนวณได้มีค่า = 0.06 เนื่องจากค่า P-Value > α ก็จะตัดสินใจยอมรับสมมติฐาน H_0 ที่ระดับนัยสำคัญ 5% เป็นต้น

กลุ่มที่ 2 นักวิจัยกลุ่มนี้จะไม่มีการกำหนดค่า α ไว้ก่อน และจะรายงานผลค่า P-Value ที่คำนวณได้ เพื่อให้ผู้อ่านตัดสินใจเองว่าที่ระดับ α ใดจะปฏิเสธ H_0 แต่โดยส่วนใหญ่แล้วผู้วิจัยจะสรุปว่าปฏิเสธ H_0 ในช่วงใดเอง ตัวอย่างเช่น ถ้า P-Value ที่คำนวณได้มีค่าเป็น 0.006 จะสรุปว่าปฏิเสธ H_0 ที่ P-Value < 0.01 ถ้าค่า P-Value ที่คำนวณได้มีค่าเป็น 0.015 จะสรุปว่าปฏิเสธ H_0 ที่ P-Value < 0.02 และ ถ้าค่า P-Value ที่คำนวณได้มีค่าเป็น 0.045 จะสรุปว่าปฏิเสธ H_0 ที่ P-Value < 0.05 เป็นต้น

จากลักษณะของนักวิจัย 2 กลุ่มข้างต้น จะสังเกตเห็นได้ว่านักวิจัยกลุ่มที่ 1 จะให้ความสำคัญกับค่า α ในตำแหน่งที่ไม่ต่อเนื่อง คือ มักจะกำหนด α เป็น 0.01 0.05 0.10 จึงมีนักสถิติหลายคนตั้งคำถามว่า ทำไมค่า α ที่ระดับ 0.05 จึงมีความสำคัญมากกว่าค่า α ที่ระดับ 0.04 หรือ 0.06 มากมายนัก จึงก่อให้เกิดการแปรผลการทดสอบสมมติฐานตามแบบนักวิจัยกลุ่มที่ 2 ซึ่งเป็นที่นิยมกันอย่างมาในปัจจุบัน เนื่องจากการแปรผลการทดสอบสมมติฐานแบบนี้จะให้ความสำคัญกับค่า α ทุกค่าเท่ากันหมด คือมองค่า α เป็นค่าต่อเนื่อง นักวิจัยกลุ่มนี้จึงรายงานค่า P-Value จริงๆ ที่คำนวณได้เพื่อให้ผู้อ่านตัดสินใจเองว่า ที่ระดับ α ค่าใดจะปฏิเสธ H_0 แต่โดยส่วนใหญ่ผู้วิจัยมักจะสรุปว่าปฏิเสธ H_0 ในช่วงใดเอง

การคำนวณค่า P-Value นี้โดยปกติแล้วโปรแกรมวิเคราะห์ข้อมูลทางสถิติต่างๆ จะคำนวณมาให้โดยอัตโนมัติ แม้กระทั่งโปรแกรม Microsoft Office Excel ก็ตาม แต่ในงานวิจัยนี้จะคำนวณค่า P-Value (Cardillo, 2008: Online) โดยใช้ฟังก์ชันในโปรแกรม MATLAB ซึ่งมีรายละเอียดการคำนวณดังสมการที่ (2.14)

$$p = (1 - tcdf(t, gl)) * tail \quad (2.14)$$

กำหนดให้ p คือ ค่า P-Value
t คือ ค่า T-Test ที่คำนวณได้
gl คือ ค่า Degrees of Freedom ของข้อมูล
$tail$ คือ จำนวนข้างในการทดสอบสมมติฐาน

จากสมการที่ (2.14) จะเห็นว่ามีการส่งพารามิเตอร์คือ t และ gl ไปยังฟังก์ชัน $tcdf$ และเมื่อได้ค่า $tcdf$ กลับมาก็จะนำมาลบออกจาก 1 ซึ่งเป็นค่าความน่าจะเป็นทั้งหมด แล้วคูณด้วยจำนวนข้างในการทดสอบ ถ้าเป็นการทดสอบข้างเดียวจะได้ค่า $tail = 1$ ถ้าเป็นการทดสอบสองข้างไม่ว่าจะเป็นข้างซ้าย หรือข้างขวาจะได้ค่า $tail = 2$ ส่วนรายละเอียดของฟังก์ชัน tcdf(Student's Cumulative Distribution(cdf)) เป็นดังสมการที่ (2.15)

$$p = F(x|v) = \left(\int_{-\infty}^{x} \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}\right)\left(\frac{1}{\sqrt{v\prod}}\right)\left(\frac{1}{\left(1+\frac{t^2}{v}\right)^{\frac{v+1}{2}}} dt\right) \quad (2.15)$$

กำหนดให้ p คือ ค่า P-Value
t คือ ค่า T-Test
v คือ ค่า Degrees of Freedom ของข้อมูล

2.4 ทฤษฎีเกี่ยวกับการวัดประสิทธิภาพของแบบจำลอง

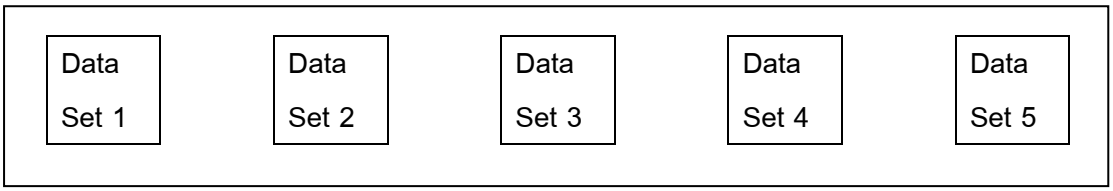
ในหัวข้อนี้จะกล่าวถึงทฤษฎีที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง เพื่อให้แบบจำลองมีความน่าเชื่อถือมากขึ้น ในที่นี้จะกล่าวถึง วิธี Cross Validation ซึ่งมีรายละเอียดดังนี้

2.4.1 Cross Validation

Cross Validation คือวิธีการในการคาดการณ์ค่าความผิดพลาดของแบบจำลอง หรือวิธีการที่เรานำเสนอ โดยพื้นฐานของวิธีการสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุด

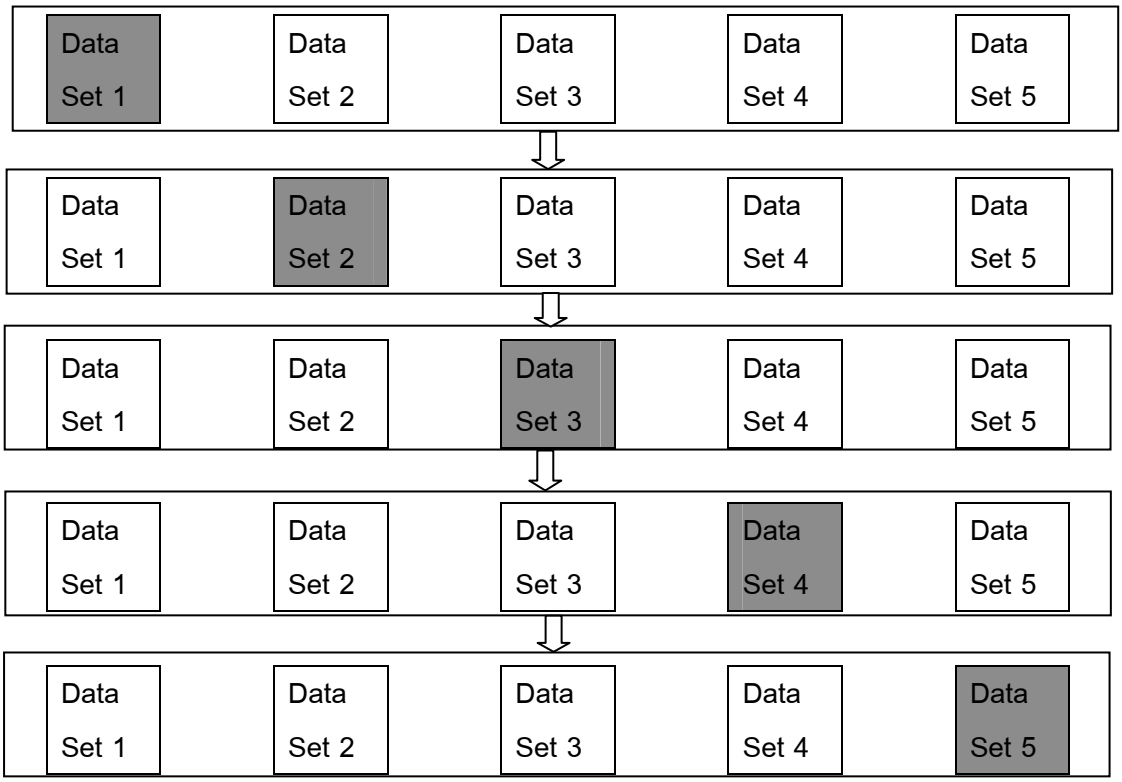
ข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดข้อมูลสอน (Train Set) ส่วนที่เหลือใช้เป็นข้อมูลทดสอบ (Test Set) ทำซ้ำจนข้อมูลทุกตัวได้มีโอกาสเป็นเป็นชุดข้อมูลสอนและทดสอบ

ในกรณีการทำ K - folds Cross Validation เราจะแบ่งข้อมูลออกเป็น K ชุดเท่าๆ กัน และทำการคำนวณค่าความผิดพลาดเป็นจำนวน K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K - 1 ชุดจะถูกใช้เป็นข้อมูลสอนดังตัวอย่าง ถ้าให้ K = 5 และข้อมูลมีทั้งหมด 100 แถว ชุดข้อมูลจะถูกแบ่งออกเป็น 5 ส่วนเท่าๆ กัน โดยจะให้แต่ละกล่องคือชุดข้อมูล 1 ชุด นั่นคือแต่ละกล่องจะประกอบด้วยข้อมูล 20 แถว ชุดข้อมูลที่แบ่งแล้วแสดงดังภาพประกอบ 2.11



ภาพประกอบ 2.11 แสดงชุดข้อมูลย่อย 5 ชุดสำหรับ 5-folds Cross Validation

ค่า K = 5 แสดงว่าจะต้องมีการทำงาน 5 รอบด้วยกัน โดยการทำงานในแต่ละรอบจะให้ใช้ชุดข้อมูลย่อย 1 ชุดเป็นชุดข้อมูลทดสอบ และข้อมูลที่เหลือ 4 ชุดเป็นชุดข้อมูลสอน การทำงานทั้ง 5 รอบแสดงดังภาพประกอบ 2.12 โดยใช้ชุดข้อมูลที่อยู่ในกล่องที่มีการระบายให้เห็นสีเป็นชุดข้อมูลทดสอบ ส่วนชุดที่เหลือเป็นชุดข้อมูลสอน



ภาพประกอบ 2.12 แสดงตัวอย่างการทำงาน 5–folds Cross Validation

จากภาพประกอบที่ 2.12 จะเห็นได้ว่า รอบที่ 1 จะได้ Data Set 1 เป็นชุดข้อมูลทดสอบ และที่เหลือเป็นชุดข้อมูลสอน ในรอบที่ 2 จะได้ Data Set 2 เป็นชุดข้อมูลทดสอบ และที่เหลือเป็นชุดข้อมูลสอน เป็นต้น

2.4.2 Confusion Matrix

Confusion Matrix คือวิธีการหนึ่งที่ใช้ในการวัดประสิทธิภาพของแบบจำลองการแบ่งประเภทของข้อมูล (Classification Model) โดยอยู่บนพื้นฐานของจำนวนตัวอย่างที่จัดกลุ่มผิด โดยจะแสดงจำนวนการจัดกลุ่มผิดไว้ในเมทริกซ์ ซึ่งมีแนวแถวแทนคลาสจริงของข้อมูล ส่วนแนวคอลัมน์แทนคลาสที่แบบจำลองทำนายได้ Confusion Matrix แสดงได้ดังภาพประกอบ 2.13

Actual Class/ Predict Class	Negative	Positive
Negative	a	b
Positive	c	d

ภาพประกอบ 2.13 แสดง Confusion Matrix

จากภาพประกอบ 2.13 จะเห็นว่าจำนวนตัวอย่างที่เป็น Negative เท่ากับ A+B จำนวนตัวอย่างที่เป็น Positive เท่ากับ C+D จำนวนตัวอย่างที่ถูกทายว่าเป็น Negative เท่ากับ A+C และจำนวนตัวอย่างที่ถูกทายว่าเป็น Positive เท่ากับ B+D ส่วนความหมายของค่า A B C และ D มีดังนี้ A คือจำนวนตัวอย่างที่เป็น Negative และทายถูกว่าเป็น Negative B คือจำนวนตัวอย่างที่เป็น Negative แต่ทายผิดว่าเป็น Positive C คือจำนวนตัวอย่างที่เป็น Positive แต่ทายผิดว่าเป็น Negative สุดท้าย D คือจำนวนตัวอย่างที่เป็น Positive และทายถูกว่าเป็น Positive ต่อไปจะอธิบายเกี่ยวกับค่าต่าง ๆ ที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง ซึ่งมีอยู่ด้วยกัน 7 ค่าด้วยกัน ดังนี้

1) AC (Accuracy) คือค่าความถูกต้องของแบบจำลอง พิจารณาจากจำนวนการทายถูกทั้งหมด สามารถคำนวณได้ดังสมการที่ (2.16)

$$AC = \frac{a+d}{a+b+c+d} \tag{2.16}$$

2) TP (True Positive Rate) คือค่าที่ทายถูกว่าเป็น Positive พิจารณาจากตัวอย่างที่ทายถูกว่าเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.17)

$$TP = \frac{d}{c+d} \tag{2.17}$$

3) FP (False Positive Rate) คือค่าที่ทายผิดว่าเป็น Positive พิจารณาจากตัวอย่างที่เป็น Negative แต่ทายผิดว่าเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.18)

$$FP = \frac{b}{a+b} \tag{2.18}$$

4) TN (True Negative Rate) คือค่าที่ทายถูกว่าเป็น Negative พิจารณาจากตัวอย่างที่ทายถูกว่าเป็น Negative สามารถคำนวณได้ดังสมการที่ (2.19)

$$TN = \frac{a}{a+b} \tag{2.19}$$

5) FN (False Negative Rate) คือค่าที่ทายผิดว่าเป็น Negative พิจารณาจากตัวอย่างที่เป็น Positive แต่ทายผิดว่าเป็น Negative สามารถคำนวณได้ดังสมการที่ (2.20)

$$FN = \frac{c}{c+d} \tag{2.20}$$

6) P (Precision) กรณีค่า Precision Positive คือค่าความแม่นยำที่ทายถูกว่าเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.21)

$$P(Positive) = \frac{d}{b+d} \tag{2.21}$$

7) P (Precision) กรณีค่า Precision Negative คือค่าความแม่นยำที่ทายถูกว่าเป็น Negative สามารถคำนวณได้ดังสมการที่ (2.22)

$$P(negative) = \frac{a}{a+c} \tag{2.22}$$

จากการคำนวณข้างต้นจะได้ว่าถ้าค่า AC TP และ TN มีค่ามากแสดงว่าแบบจำลองมีประสิทธิภาพสูง แต่ถ้าค่า FP และ FN มีค่ามากแสดงว่าแบบจำลองมีประสิทธิภาพต่ำ

บทที่ 3

วิเคราะห์และออกแบบแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง

วิทยานิพนธ์นี้ได้ทำการออกแบบแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction Using Self-Organizing Map: DNA_MKESOM) โดยใช้เทคนิคแผนที่การจัดกลุ่มเอง (Self-Organizing Map: SOM) ในการจัดกลุ่มข้อมูลดีเอ็นเอไมโครอาเรย์ และใช้วิธีการทางสถิติ (Statistics Method) เข้ามาร่วมด้วยในขั้นตอนการทำงานของแบบจำลอง ซึ่งค่าทางสถิติที่นำมาใช้ในแบบจำลองนี้เรียกว่าค่าความน่าจะเป็น (Probability Value: P-Value) โดยจะนำมาใช้ในส่วนของการลดมิติ (จำนวนยีน) ของข้อมูลลง เพื่อลดเวลาในการทำงานให้น้อยลงในขณะที่ค่าความถูกต้องยังคงสูงด้วย

3.1 แบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction using Self-Organizing Map: DNA_MKESOM)

แบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง ประกอบด้วยขั้นตอนการทำงานหลักทั้งหมด 6 ขั้นตอนดังนี้ ขั้นตอนหลักที่ 1: ขั้นตอนการเตรียมข้อมูล (Data Preprocessing Process) ขั้นตอนหลักที่ 2: ขั้นตอนการเลือกยีนโดยอาศัยค่าความน่าจะเป็น (Feature Extraction with P-Value Process) ขั้นตอนหลักที่ 3: ขั้นตอนการแบ่งกลุ่มข้อมูลโดยใช้แผนที่การจัดกลุ่มเองและการคำนวณค่า Highest Certainty Value (Self-Organizing Map Clustering and Highest Certainty Value Calculation Process) ขั้นตอนหลักที่ 4: ขั้นตอนการเลือกยีนโดยอาศัยค่า Highest Certainty Value (Feature Extraction with Highest Certainty Value Process) ขั้นตอนหลักที่ 5: ขั้นตอนการสร้างกฎการแบ่งกลุ่ม (Rules Creation Process) และขั้นตอนหลักที่ 6: ขั้นตอนการทดสอบประสิทธิภาพของกฎ (Rules Calculation Process) โดยแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองทั้ง 6 ขั้นตอนที่กล่าวมานี้ สามารถแสดงได้ดังภาพประกอบ 3.1

ขั้นตอนหลักที่ 1: Data Preprocessing Process
ขั้นตอนหลักที่ 2: Feature Extraction with P-Value Process
ขั้นตอนหลักที่ 3: Self-Organizing Map Clustering and Highest Certainty Value Calculate Process
ขั้นตอนหลักที่ 4: Feature Extraction with Highest Certainty Value Process
ขั้นตอนหลักที่ 5: Rules Creation Process
ขั้นตอนหลักที่ 6: Rule Calculation Process

ภาพประกอบ 3.1 แสดงขั้นตอนการทำงานหลักของแบบจำลอง DNA_MKESOM

จากภาพประกอบ 3.1 ในขั้นตอนการทำงานหลักแต่ละขั้นตอนจะประกอบด้วยขั้นตอนการทำงานย่อยๆ อีก โดยในส่วนต่อไปนี้จะแสดงรายละเอียดการทำงานของแต่ละขั้นตอนตามลำดับ

3.1.1 ขั้นตอนหลักที่ 1: ขั้นตอนการเตรียมข้อมูล

ในขั้นตอนหลักนี้จะทำการเตรียมข้อมูลให้พร้อมก่อนที่จะเริ่มกระบวนการต่อไป เนื่องจากในแบบจำลองที่พัฒนาขึ้นได้กำหนดรูปแบบของข้อมูลที่จะใช้ไว้ชัดเจน คือข้อมูลจะต้องอยู่ในรูปแบบเมทริกซ์ที่มีคอลัมน์แทนตัวอย่าง และแถวแทนยีน ข้อมูลในคอลัมน์แรกเป็นหมายเลขของยีน (Gene Number) ข้อมูลในแถวสุดท้ายเป็นกลุ่มของข้อมูล ลักษณะของข้อมูลที่ใช้แบ่งเป็น 2 ลักษณะคือ ข้อมูลที่มีลักษณะเป็นคู่ (Paired Data) และข้อมูลที่เป็นอิสระต่อกัน (Unpaired Data) ส่วนวิธีการที่ใช้ในการแบ่งข้อมูลออกเป็นชุดสอน (Train Set) และชุดทดสอบ (Test Set) จะใช้วิธี K-folds Cross Validation เท่านั้น โดยผู้ใช้สามารถกำหนดจำนวน Fold ที่ต้องการได้ ตัวอย่างข้อมูลเข้าที่เหมาะสมสำหรับแบบจำลองนี้แสดงดังภาพประกอบ 3.2

Accession Number	Tumor 27	Tumor 29	Tumor 34	Tumor 28	Tumor 35	Tumor 8	Tumor 3	Tumor 9	Tumor 4	Tumor 32	Tumor 39	Tumor 10	Tumor 33	Tumor 5
D00003	2	7	1	-3	0	10	15	-2	36	-1	-4	1	3	39
D00003	-5	1	7	8	6	-6	7	-14	-1	15	-3	-1	-10	-1
D00003	1	-4	-3	3	-3	-14	-8	-15	-17	-14	-6	13	1	72
D00015	26	13	13	9	2	10	20	17	21	17	15	20	5	32
D00102	5	2	0	10	-5	3	-5	-2	-3	7	-8	-7	4	-9
D00137	18	13	11	28	6	63	26	136	26	42	8	20	37	14
D00173	1	5	10	8	10	7	-1	4	1	-3	0	-9	3	5
D00265	153	145	132	73	184	78	51	72	41	99	93	162	53	403
D00306	-7	-7	-6	-17	10	-5	-20	-22	-8	-14	-6	-21	-3	-14
D00408	1	1	0	3	2	-7	0	1	-4	-5	-3	3	-9	7
D00596	70	138	73	27	125	56	24	14	106	54	89	65	72	42
D00726	20	17	22	10	14	27	24	-5	29	10	14	0	15	13
D00749	65	64	79	62	86	86	99	71	97	84	107	113	74	34

ภาพประกอบ 3.2 แสดงรูปแบบข้อมูลดีเอ็นเอไมโครอาเรย์ที่ใช้ในแบบจำลอง

ในขั้นตอนหลักที่ 1 เป็นขั้นตอนแรกในการทำงานของแบบจำลองนี้ เป็นการเตรียมข้อมูลเพื่อให้พร้อมทำงานในขั้นตอนต่อไป โดยแบ่งขั้นตอนการทำงานย่อยออกเป็น 3 ขั้นตอน ดังนี้

ขั้นตอนย่อยที่ 1.1 ขั้นตอนการรับข้อมูลและตรวจสอบความถูกต้องของข้อมูล โดยจะรับข้อมูลดีเอ็นเอไมโครอาเรย์ จำนวน Fold และรูปแบบของข้อมูล และทำการตรวจสอบความถูกต้องของข้อมูล

ขั้นตอนย่อยที่ 1.2 ขั้นตอนการแบ่งข้อมูลแต่ละคลาสออกจากกัน จะทำการแบ่งข้อมูลแต่ละคลาสแยกออกจากกัน โดยต้องคำนึงถึงรูปแบบของข้อมูลด้วย ถ้าข้อมูลมีลักษณะเป็นคู่ (Paired Data) จำนวนตัวอย่างแต่ละแฟ้มต้องมีขนาดเท่ากัน และมีรูปแบบเฉพาะ ดังภาพประกอบ 3.3 ส่วนข้อมูลที่มีลักษณะเป็นอิสระต่อกัน (Unpaired Data) จำนวนตัวอย่างในแต่ละแฟ้มไม่จำเป็นต้องเท่ากันก็ได้ ซึ่งลักษณะของข้อมูลนี้จะมีผลต่อขั้นตอนการเตรียมข้อมูลในขั้นตอนที่ 3 และมีผลต่อการคำนวณหาค่า P-Value ในขั้นตอนต่อไปด้วย

Tumor 27	Tumor 29	Tumor 34	…	Tumor 28	Normal 27	Normal 29	Normal 34	…	Normal 28
2	7	1	…	0	6	1	1	…	7
-5	1	7	…	6	22	2	6	…	11
1	-4	-3	…	-3	14	0	-9	…	-9
26	13	13	…	2	54	50	45	…	32
5	2	0	…	-5	-2	1	1	…	-3

ภาพประกอบ 3.3 แสดงรูปแบบของข้อมูลที่มีลักษณะสัมพันธ์กัน

จากภาพประกอบ 3.3 จะเห็นว่าข้อมูลที่มีลักษณะสัมพันธ์กันจะต้องมีจำนวนตัวอย่างที่เป็นโรคเท่ากับตัวอย่างที่ปกติ และอยู่ในตำแหน่งที่ตรงกันด้วยดังจะสังเกตได้จากข้อมูลที่มีเส้นลูกศรเชื่อมจะเป็นข้อมูลที่มาจากผู้ป่วยคนเดียวกัน

ขั้นตอนย่อยที่ 1.3 ขั้นตอนการแบ่งข้อมูลเพื่อการทดสอบแบบ K-folds Cross Validation โดยจะนำตัวอย่างแต่ละคลาสมาทำการแบ่งข้อมูลออกเป็นชุดข้อมูลสอน (Train Set) และชุดข้อมูลทดสอบ (Test Set) ตามค่าจำนวน Fold ที่ได้รับมาจากขั้นตอนที่ 1 ขั้นตอนการทำงานของขั้นตอนหลักที่ 1 แสดงดังภาพประกอบ 3.4

ขั้นตอนหลักที่ 1

ขั้นตอนย่อยที่ 1.1

1.1.1 รับข้อมูลเข้า

1.1.2 รับค่าจำนวน Fold

1.1.3 รับค่ารูปแบบของข้อมูลคือ ข้อมูลลักษณะคู่ (Paired Data) หรือ ข้อมูลที่เป็นอิสระต่อกัน (Unpaired Data)

1.1.4 ตรวจสอบความถูกต้องของรูปแบบข้อมูล

ขั้นตอนย่อยที่ 1.2

1.2.1 แบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่ม Class1 กับ กลุ่ม Class2

ขั้นตอนย่อยที่ 1.3

1.3.1 แบ่งข้อมูลออกเป็นชุดสอน (Train Set) และชุดทดสอบ (Test Set) ตามจำนวน Fold ที่รับมาในขั้นตอนที่ 1.1.2 โดยมีการทำงานดังนี้

1.3.1.1 กำหนดหมายเลขประจำกลุ่มให้กับข้อมูลใน Class 1 (Data_Class1) โดยการสุ่ม ซึ่งจำนวนกลุ่มทั้งหมดที่เป็นไปได้จะเท่ากับจำนวน Fold

1.3.1.2 กำหนดหมายเลขประจำกลุ่มให้กับข้อมูลใน Class 2 (Data_Class2) โดยการสุ่ม ซึ่งจำนวนกลุ่มทั้งหมดที่เป็นไปได้จะเท่ากับจำนวน Fold

1.3.1.3 แบ่งข้อมูลออกเป็น Train Set และ Test Set ในแต่ละ Fold การทำงานในขั้นตอนนี้แสดงได้ดังนี้

กำหนดให้ F แทนจำนวน Fold

N_Class1 แทนจำนวนตัวอย่างที่เป็น Class1

N_Class2 แทนจำนวนตัวอย่างที่เป็น Class2

Group แทนค่าหมายเลขประจำกลุ่มที่สุ่มได้

```
For i = 1 to F
    For m = 1 to NClass1
        If Group(Data_Class1(m)) = i
            save to “Test Set”
        Else
            save to “Train Set”
        End if
    End for
```

```
            For  n =  1  to  NClass2
                    If   Group(Data_Class2(n))  =  I
                            save to “Test Set”
                    Else
                            save to “Train Set”
                    End if
            End for
End for
```

ภาพประกอบ 3.4 ขั้นตอนการทำงานในขั้นตอนหลักที่ 1

3.1.2 ขั้นตอนหลักที่ 2: ขั้นตอนการเลือกยีนโดยอาศัยค่าสถิติลำดับความสำคัญ (Feature Extraction with P-Value Process)

ในขั้นตอนหลักที่ 2 นี้ เป็นขั้นตอนที่ทำการเลือกยีน เพื่อกรองยีนจำนวนหนึ่งออกไป ซึ่งในส่วนนี้ยีนที่ไม่มีความสำคัญหรือมีความสำคัญต่อการแบ่งกลุ่มของข้อมูลน้อยจะถูกกรองออกไปเหลือไว้เฉพาะยีนที่มีความสำคัญต่อการแบ่งกลุ่มของข้อมูลมากกว่าเท่านั้น เพื่อให้สามารถลดเวลาในการประมวลผลข้อมูลจำนวนมหาศาลลงได้ และยังสามารถลดการใช้ทรัพยากรในการประมวลลงอีกด้วย สิ่งที่สำคัญในขั้นตอนนี้คือจะต้องพิจารณาด้วยว่าการกรองยีนจำนวนนั้นออกจะต้องไม่ทำให้ความถูกต้องของข้อมูลลดลง โดยการเลือกยีนในที่นี้จะเลือกโดยอาศัยค่า P-Value จะทำการกรองยีนที่มีค่า P-Value มากออกไปจำนวนหนึ่ง ซึ่งผู้ใช้จะสามารถกำหนดจำนวนยีนที่ต้องการกรองออกได้ด้วย เมื่อผ่านขั้นตอนนี้เหลือเฉพาะยีนที่มีค่า P-Value น้อยสำหรับขั้นตอนการทำงานต่อไป การทำงานของขั้นตอนที่ 2 นี้ประกอบด้วยขั้นตอนการทำงานย่อย 2 ขั้นตอนด้วยกัน ดังนี้

ขั้นตอนย่อยที่ 2.1 ขั้นตอนการคำนวณ P-Value จะนำเอาชุดข้อมูลสอนในแต่ละ Fold ที่ได้แบ่งแล้วในขั้นตอนหลักที่ 1 มาทำการหาค่า P-Value ของยีนแต่ละยีน

ขั้นตอนย่อยที่ 2.2 ขั้นตอนการเลือกยีน จะนำค่า P-Value มาเรียงลำดับจากน้อยไปหามาก และเลือกยีนโดยอาศัยค่า Gene Selection Threshold (GST) ที่ผู้ใช้ป้อนเข้ามาในขั้นตอนที่ 2.1.1 ขั้นตอนการทำงานในขั้นตอนหลักที่ 2 แสดงดังภาพประกอบ 3.5

ขั้นตอนหลักที่ 2

ขั้นตอนย่อยที่ 2.1

2.1.1 รับค่า Gene Selection Threshold (GST) เพื่อใช้ในการเลือกยีนโดยอาศัยค่า P-Value ซึ่งค่าที่ใช้จะอยู่ระหว่าง 1%-100%

2.1.2 คำนวณค่า P-Value สำหรับแต่ละยีนในชุดข้อมูลสอน โดยจะทำการคำนวณจนครบทุก Fold ซึ่งค่า P-Value สามารถคำนวณได้จากค่า T-Test โดยสูตรในการคำนวณค่า T-Test มี 2 แบบ คือ Paired T-Test ดังสมการที่ (3.1) – (3.3) และ Unpaired T-Test ดังสมการที่ (3.4) – (3.5) ตามลำดับ

$$t-test = (\bar{x}-\bar{y})\sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(\hat{x}_i-\hat{y}_i)^2}} \tag{3.1}$$

$$\hat{x}_i = (x_i-\bar{x}) \tag{3.2}$$

$$\hat{y}_i = (y_i-\bar{y}) \tag{3.3}$$

กำหนดให้ $\bar{x}$ คือค่าเฉลี่ยของข้อมูลชุดที่ 1

$\bar{y}$ คือค่าเฉลี่ยของข้อมูลชุดที่ 2

n คือจำนวนตัวอย่าง

$$t-test = \frac{\bar{x}-\bar{y}}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}} \tag{3.4}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i-\bar{x})^2-\sum_{i=1}^{n_2}(y_i-\bar{y})^2}{n_1+n_2-2} \tag{3.5}$$

กำหนดให้ n_1 คือจำนวนตัวอย่างกลุ่มที่ 1

n_2 คือจำนวนตัวอย่างกลุ่มที่ 2

ในส่วนของการคำนวณค่า P-Value ที่ใช้ในแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองนี้ จะคำนวณโดยใช้ฟังก์ชันในโปรแกรม MATLAB ซึ่งฟังก์ชันที่ใช้คือ *tcdf* (Student's cumulative distribution) ทำให้ค่า P-Value สามารถหาได้จากสมการที่ (3.6)

$$p = (1-tcdf(t,gl))*tail \tag{3.6}$$

กำหนดให้ p คือค่า P-Value

t คือค่า T-Test ที่คำนวณได้

gl คือค่า Degrees of Freedom ของข้อมูล

$tail$ คือจำนวนข้างในการทดสอบสมมติฐาน

จากสมการที่ (3.6) จะเห็นว่ามีการส่งพารามิเตอร์คือ t และ gl ไปยังฟังก์ชัน $tcdf$ และเมื่อได้ค่า $tcdf$ กลับมาก็จะนำมาลบออกจาก 1 ซึ่งเป็นค่าความน่าจะเป็นทั้งหมด แล้วคูณด้วยจำนวนข้างในการทดสอบ ถ้าเป็นการทดสอบข้างเดียวจะได้ค่า $tail = 1$ ถ้าเป็นการทดสอบสองข้างไม่ว่าจะเป็นข้างซ้าย หรือข้างขวาจะได้ค่า $tail = 2$ ส่วนรายละเอียดของฟังก์ชัน tcdf(Student's cumulative distribution(cdf)) เป็นดังสมการที่ (3.7)

$$p = F(x|v) = \left(\int_{-\infty}^{x} \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}\right)\left(\frac{1}{\sqrt{v\prod}}\right)\left(\frac{1}{\left(1+\frac{t^2}{v}\right)^{\frac{v+1}{2}}} dt\right) \quad (3.7)$$

กำหนดให้ p คือ ค่า P-Value

t คือ ค่า T-Test

v คือ ค่า Degrees of Freedom ของข้อมูล

ขั้นตอนย่อยที่ 2.2

2.2.1 ทำการเรียงลำดับค่า P-Value จากน้อยไปหามาก ในทุก Fold

2.2.2 เลือกกลุ่มยีนที่มีค่า P-Value น้อย โดยการเลือกจะอาศัยค่า Gene Selection Threshold (GST) ที่รับมาในขั้นตอนย่อยที่ 2.1.1

ภาพประกอบ 3.5 ขั้นตอนการทำงานในขั้นตอนหลักที่ 2

ตัวอย่างการทำงานในขั้นตอนหลักที่ 2 นี้จะแสดงโดยใช้ยีนจำนวน 10 ยีน โดยที่ข้อมูลมีลักษณะเป็นคู่ โดยยีนที่เลือกและค่า P-Value ที่คำนวณได้แสดงดังตารางที่ 3.1

ตารางที่ 3.1 แสดงตัวอย่างยีนและค่า P-Value ที่คำนวณได้ในขั้นตอนที่ 2.2

หมายเลขยีน	ข้อมูล	ค่า P-Value
6	18, 13, 11, 28, …	0.0000408791063465597
17	258, 318, 28, 6, …	7.35674684083557E-06
6901	-8, -10, -1, -11, …	0.836484005972339
6983	246, 168, 237, 136, …	1.01936889196904E-08
7221	35, 25, 30, 21, …	0.773937396744265
7231	2, -2, 1, 0, …	0.940208072380744
7362	-5, -1, -1, -1, …	0.92387716729549
7401	113, 84, 123, 120, …	0.0849390400253417
7447	18, 12, 29, 28, …	3.65111931506288E-07
7448	42, 27, 44, 51, …	5.13571046758461E-07

ขั้นตอนต่อจากนี้จะทำการเรียงลำดับยีนตามค่า P-Value แล้วเลือกยีนตามค่า GST ในที่นี้สมมติว่าให้ค่า GST = 40% ก็จะเลือกยีนที่มีค่า P-Value ต่ำ ออกมา 4 ยีน ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 แสดงข้อมูลยีนที่เลือกในกรณีกำหนดค่า GST = 40%

หมายเลขยีน	ค่า P-Value
6983	1.01936889196904E-08
7447	3.65111931506288E-07
7448	5.13571046758461E-07
17	7.35674684083557E-06

จากตางที่ 3.2 จะได้ยีน 2 ยีนที่จะนำไปจัดกลุ่มด้วยแผนที่การจัดกลุ่มเองและคำนวณค่า HCV ในขั้นตอนหลักที่ 3 ต่อไป

3.1.3 ขั้นตอนหลักที่ 3: ขั้นตอนการแบ่งกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง และขั้นตอนการคำนวณค่า Highest Certainty Value (Self-Organizing Map Clustering and Highest Certainty Value Process)

ในขั้นตอนหลักที่ 3 นี้ ประกอบด้วยขั้นตอนการทำงานย่อย 3 ขั้นตอนด้วยกันคือ

ขั้นตอนย่อยที่ 3.1 ขั้นตอนการเตรียมไฟล์ข้อมูล โดยในขั้นตอนนี้จะนำข้อมูลของยีนแต่ละยีนที่เลือกได้จากขั้นตอนหลักที่ 2 มาสร้างไฟล์สำหรับใช้ในการทำงานในแผนที่การจัดกลุ่มเอง โดยแต่ละยีนจะถูกสร้างเป็น 1 ไฟล์ข้อมูลที่มีรูปแบบเฉพาะ และจะทำการสร้างไฟล์ข้อมูลจนครบทุก Fold ตัวอย่างข้อมูลแสดงได้ดังภาพประกอบ 3.6

ขั้นตอนย่อยที่ 3.2 ขั้นตอนการจัดกลุ่มข้อมูลโดยใช้แผนที่การจัดกลุ่มเอง

ขั้นตอนย่อยที่ 3.3 ขั้นตอนการคำนวณหาค่า Highest Certainty Value ของแต่ละยีนโดยอาศัยผลที่ได้จากการจัดกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง ขั้นตอนการทำงานในขั้นตอนหลักที่ 3 แสดงดังภาพประกอบ 3.7

Num

#n attribute(1) attribute(2) attribute(3) … attribute(n) class

value(1,1) value(1,2) value(1,3) … value(1,n) class

value(2,1) value(2,2) value(2,3) … value(2,n) class

value(3,1) value(3,2) value(3,3) … value(3,n) class

.

.

.

value(m,1) value(m,2) value(m,3) … value(m,n) class

ภาพประกอบ 3.6 แสดงตัวอย่างข้อมูลที่เหมาะสมในการใช้ SOM

จากภาพประกอบ 3.6 จะเห็นว่าข้อมูลที่จะนำมาใช้ทำงานกับ SOM เป็นข้อมูลที่มีลักษณะเฉพาะ โดยที่ กำหนดให้ Num คือจำนวน Attribute Value (m,n) คือค่าการแสดงออกของยีนที่ n จากตัวอย่างที่ m และ Class คือกลุ่มของข้อมูล

ขั้นตอนหลักที่ 3

ขั้นตอนย่อยที่ 3.1 ในแต่ละ Fold จะนำข้อมูลของยีนแต่ละยีนมาสร้างเป็นไฟล์ข้อมูล 1 ยีนต่อ 1 ไฟล์ข้อมูล

ขั้นตอนย่อยที่ 3.2 สำหรับแต่ละตัวแปรเข้า G_i ทำการจัดกลุ่มข้อมูลโดยใช้แผนที่การจัดกลุ่มเอง เมื่อ $1 \leq i \leq n$ และ n คือจำนวนตัวแปรเข้าทั้งหมด ผลลัพธ์ที่ได้จากการจัดกลุ่มคือ component plane, color bar, label และ hits

ขั้นตอนย่อยที่ 3.3

3.3.1 Remove Component เป็นค่าว่างซึ่งหมายความว่าหลังจากการจัดกลุ่ม ไม่มีข้อมูลใดที่ถูกจัดอยู่ในช่วงค่านั้น

3.3.2 คำนวณค่า Certainty Value โดยมีขั้นตอนวิธีดังนี้

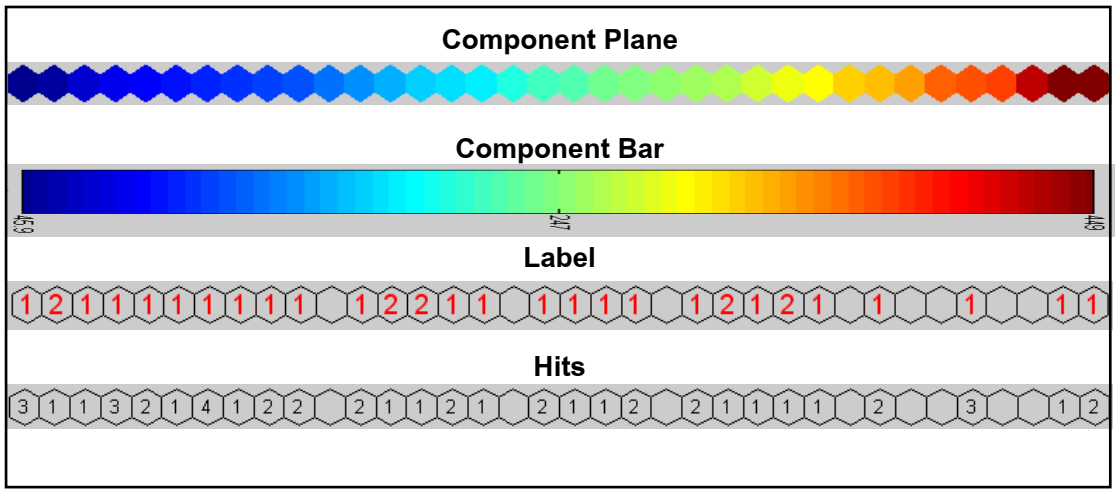
```
while i <= size(label)
        first = label(i)
        last = label(i+1)
        if first <> last then
            for j =1 to i
                if label(j) = '2' then
                    miss = miss + hits(j)
                end if
            end for
            for k = i+1 to end
                if label(k) = '1' then
                    miss = miss + hits(k)
                end if
            end for
            certainty(1) = (TotalHits – miss / TotalHits) * 100
            certainty(2) = 100 – certainty1
        end if
    end while
```

กำหนดให้	n	แทนจำนวนช่องของ Component plane
	miss	แทนจำนวนตัวอย่างที่จัดผิดกลุ่ม
	TotalHits	แทนผลรวมของ Hits ทั้งหมด
	Certainty1	แทนค่า Certainty ของยีนกรณีที่ค่าของข้อมูลใน Class1 น้อยกว่าข้อมูลใน Class2
	Certainty2	แทนค่า Certainty ของยีนกรณีที่ค่าของข้อมูลใน Class1 มากกว่าข้อมูลใน Class2
	HCV	แทนค่า Highest Certainty Value ของยีนแต่ละยีน

3.3.3 คำนวณ Highest Certainty Value (HCV) ของแต่ละยีน ซึ่งก็คือค่า Certainty สูงสุดที่หาได้จากขั้นตอนที่ 3.3.2 นั่นเอง

ภาพประกอบ 3.7 ขั้นตอนการทำงานในขั้นตอนหลักที่ 3

จากภาพประกอบ 3.7 แสดงตัวอย่างผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลด้วย SOM ได้โดยใช้ตัวอย่างยีนจากผู้ป่วยโรคมะเร็งลำไส้ (Colon Cancer) จะได้ผลลัพธ์ที่ได้จากการจัดกลุ่มด้วย SOM คือ Component Plane Component Bar Label และ Hits แสดงดังภาพประกอบ 3.8 ตามลำดับ

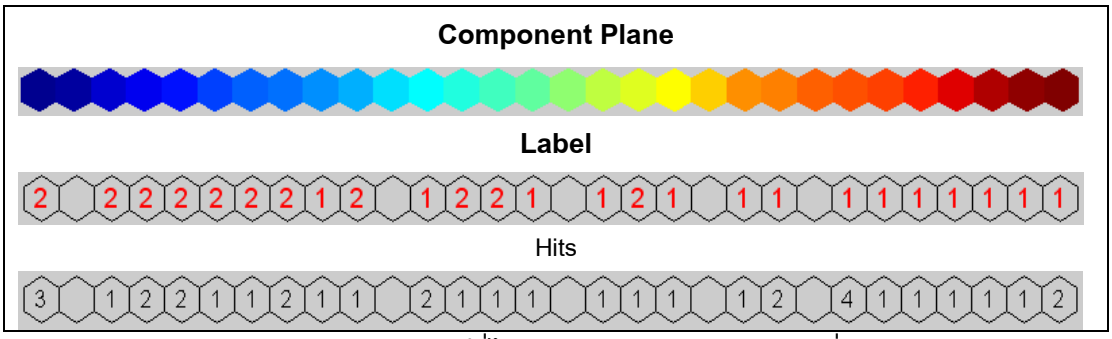


ภาพประกอบ 3.8 แสดงตัวอย่างผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลด้วย SOM

จากภาพประกอบ 3.8 ค่าสีจาก Component Plane สัมพันธ์กับค่าสีใน Component Bar ซึ่งเห็นได้จากช่องปลายจากด้านซ้ายจะเป็นสีเดียวกันและไล่มาเรื่อยๆจนถึง

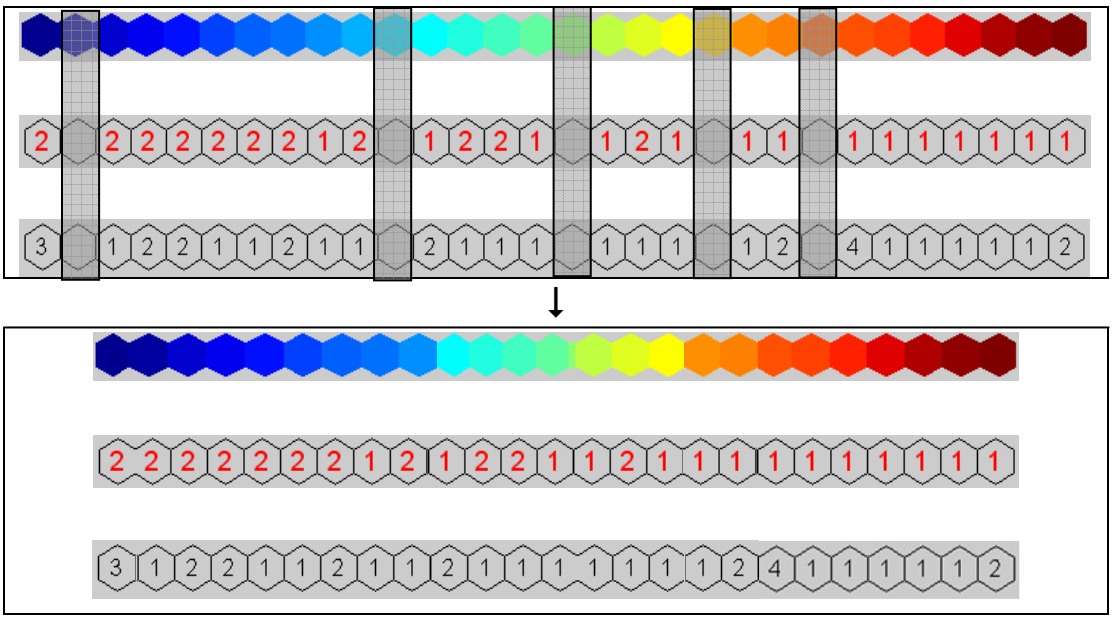
ปลายด้านขวาซึ่งเป็นสีแดงจะสัมพันธ์กันตลอด ซึ่งในขั้นตอนต่อไปคือการคำนวณค่า HCV จะเลือกใช้เฉพาะ Color Plane เท่านั้น จากตารางที่ 3.2 ถ้านำยีนที่ได้จากการเลือกทั้งสองยีนมาทำการแบ่งกลุ่มด้วยแผนที่การจัดกลุ่มเองจะได้ผลลัพธ์ดังนี้

ยีนหมายเลข 6983 เมื่อนำไปจัดกลุ่มด้วยแผนที่การจัดกลุ่มเองจะได้ผลลัพธ์ออกมา 4 อย่างคือ Component Plane Component Bar Label และ Hits ดังแสดงในภาพประกอบ 3.8 แต่เมื่อต้องการนำมาคำนวณค่า HCV จะใช้เฉพาะ Component Plane Label และ Hits เท่านั้น ดังภาพประกอบ 3.9



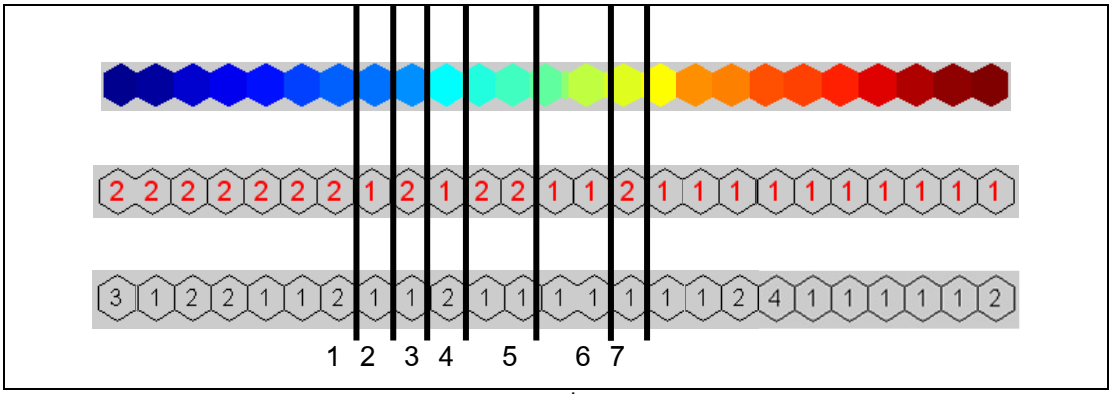
ภาพประกอบ 3.9 แสดงผลลัพธ์ที่ได้จากการแบ่งกลุ่มด้วยแผนที่การจัดกลุ่มเอง

จากภาพประกอบ 3.9 แต่ละช่องที่ตรงกันของ Component Plane, Label และ Hits จะมีความสัมพันธ์กัน เช่น ช่องทางซ้ายสุดจะเป็นช่องสีน้ำเงินซึ่งมีค่าสีเป็น 41.7 และข้อมูลที่ถูกจัดไว้ในช่องนั้นเป็นข้อมูลของตัวอย่าง Class 2 ทั้งหมด 3 ตัวอย่าง ส่วนช่องทางขวามือสุดเป็นช่องสีแดงมีค่าสีเป็น 245 และข้อมูลที่ถูกจัดไว้ในช่องนั้นเป็นข้อมูลของตัวอย่างที่เป็น Class 1 ซึ่งมีทั้งหมด 2 ตัวอย่าง เป็นต้น ต่อจากนี้เริ่มขั้นตอนการหาค่า HCV โดยเริ่มจากขั้นตอนที่ 3.3.1) คือทำการ Remove ช่องที่เป็นค่าว่างออกโดยสังเกตจากค่าของ Label ตรงไหนที่เป็นช่องว่างไม่มีหมายเลขกำกับจะทำการ Remove ดังภาพประกอบ 3.10



ภาพประกอบ 3.10 แสดงผลลัพธ์จากการ Remove ช่องว่างออก

เริ่มทำการคำนวณค่า HCV โดยทำตามขั้นตอนที่ 3.3.2) จะได้ว่า size (Label) คือจำนวนช่องของ Label มีค่าเท่ากับ 25 ส่วน TotalHits คือผลรวมของ Hits คือจำนวนตัวอย่างทั้งหมดนั่นเอง มีค่าเท่ากับ 36 จากนั้นจะคำนวณแต่ละจุดว่าถ้าเลือกจุดนั้นเป็นจุดตัดระหว่างข้อมูลสองกลุ่มแล้วจะได้ค่า Certainty เป็นเท่าไร โดยจะคำนวณเฉพาะจุดที่อยู่ระหว่างค่า Label ที่ต่างกันเท่านั้น ดังนั้นจุดที่นำมาคำนวณจะมีทั้งหมด 7 จุดเท่านั้น ดังภาพประกอบ 3.11



ภาพประกอบ 3.11 แสดงจุดที่มีการคำนวณค่า Certainty

จากนี้จะแสดงตัวอย่างการคำนวณค่า Certainty โดยเริ่มจากจุดตัดที่ 1 ก่อนจะได้ค่า Certainty 2 ค่า จากการคำนวณ 2 กรณีดังนี้

กรณีที่ 1 กรณีข้อมูลที่อยู่ทางซ้ายของจุดตัดเป็น Class 1 ส่วนทางขาวของจุดตัดเป็น Class 2 และทำการนับว่ามีค่า miss ซึ่งคือค่าข้อมูลที่อยู่ผิดกลุ่ม มีค่าเท่ากับ 32 ดังนั้นค่า Certainty 1 สามารถคำนวณได้ดังสมการที่ (3.8)

$$\begin{aligned} \text{Certainty (1)} &= ((\text{TotalHits} - \text{Miss}) \ / \ \text{TotalHits}) * 100 \\ &= ((36 - 32)/36) * 100 \\ &= 11.11 \end{aligned} \qquad (3.8)$$

จากการคำนวณตามสมการที่ (3.8) จะได้ว่าค่า Certainty 1 มีค่าเท่ากับ 11.11 ส่วนค่า Certainty 2 สามารถคำนวณได้ดังสมการที่ (3.9)

$$\begin{aligned} \text{Certainty (2)} &= 100 - \text{certainty1} \\ &= 100 - 11.11 \\ &= 88.89 \end{aligned} \qquad (3.9)$$

จากการคำนวณตามสมการที่ (3.9) จะได้ว่าค่า Certainty 2 มีค่าเท่ากับ 88.89

จากตัวอย่างการคำนวณค่า Certainty 1 และ Certainty 2 ที่จุดที่ 1 จากนั้นจะคำนวณค่า Certainty จนครบทั้ง 7 จุดตัด จะได้ค่า Certainty 1 และ Certainty 2 ทั้งหมด 14 ค่าดังแสดงในตารางที่ 3.3

ตารางที่ 3.3 แสดงค่า Certainty ของยีนหมายเลข 6983 ทุกกรณี

จุดตัด	ค่า Certainty 1	ค่า Certainty 2
1	11.11	**88.89**
2	13.89	86.11
3	11.11	**88.89**
4	16.67	83.33
5	11.11	**88.89**
6	16.67	83.33
7	13.89	86.11

จากตารางที่ 3.3 จะเห็นว่าค่า Certainty สูงสุดจะเท่ากับ 88.89 ซึ่งมีอยู่ 3 จุดด้วยกันคือ จุดตัดที่ 1 3 และ 5 โดยจะเลือกจุดใดเป็นจุดตัดที่จะใช้ในการสร้างกฎก็ได้ ในแบบจำลองนี้จะเลือกจุดที่เจอเป็นจุดแรกคือจุดที่ 1 และค่า Certainty สูงสุดนี้ต่อไปจะเรียกว่าค่า Highest Certainty Value ซึ่งจะใช้ในขั้นตอนการเลือกยีนโดยอาศัยค่า HCV ต่อไป

เมื่อหาค่า HCV ของยีนหมายเลข 6983 ได้แล้วต่อมาก็จะทำการหาค่า HCV ของยีนหมายเลข 7447 7448 และ 17 ผลลัพธ์จากการคำนวณแสดงดังตารางที่ 3.4 3.5 และ 3.6 ตามลำดับ

ตารางที่ 3.4 แสดงค่า Certainty ของยีนหมายเลข 7447 ทุกกรณี

จุดตัด	ค่า Certianty 1	ค่า Certainty 2
1	**91.67**	8.33
2	88.89	11.11
3	**91.67**	8.33
4	86.11	13.89
5	**91.67**	8.33

ตารางที่ 3.5 แสดงค่า Certainty ของยีนหมายเลข 7448

จุดตัด	ค่า Certianty 1	ค่า Certainty 2
1	78.55	21.45
2	**88.89**	11.11
3	75.55	24.45
4	86.11	13.89
5	56.11	43.89
6	55.55	44.45
7	56.11	56.11
8	75.55	24.45

ตารางที่ 3.6 แสดงค่า Certainty ของยีนหมายเลข 17

จุดตัด	ค่า Certianty 1	ค่า Certainty 2
1	93.89	6.11
2	**95.55**	4.45
3	75.55	24.45

จากตารางที่ 3.4 จะเห็นว่าค่า HCV ของยีนหมายเลข 7447 เท่ากับ 91.67 ค่า HCV ของยีนหมายเลข 7448 เท่ากับ 88.89 และค่า HCV ของยีนหมายเลข 17 เท่ากับ

95.55 ซึ่งค่า HCV ของแต่ละยีนนี้จะนำไปใช้ในขั้นตอนการเลือกยีนโดยอาศัยค่า HCV ต่อไป ผลลัพธ์จากการทำงานในขั้นตอนหลักที่ 3 สามารถแสดงได้ดังตารางที่ 3.7

ตารางที่ 3.7 แสดงผลลัพธ์ที่คำนวณได้จากขั้นตอนหลักที่ 3

หมายเลขยีน	ค่า HCV
6983	88.89
7447	91.67
7448	88.89
17	95.55

3.1.4 ขั้นตอนหลักที่ 4: ขั้นตอนการเลือกยีนโดยอาศัยค่า Highest Certainty Value (Feature Extraction with Highest Certainty Value)

ในขั้นตอนหลักที่ 4 นี้ เป็นขั้นตอนที่ทำการเลือกยีนโดยอาศัยค่า HCV ซึ่งค่า HCV นี้คำนวณได้จากขั้นตอนหลักที่ 3 เรียบร้อยแล้ว ส่วนจะเลือกยีนที่มีค่า HCV ตั้งแต่ค่าใดนั้นจะขึ้นอยู่กับค่า HCV Thershold (HCV_T) ที่ผู้ใช้สามารถกำหนดได้ โดยจะเลือกยีนที่มีค่า HCV มากกว่าหรือเท่ากับค่า HCV_T ขั้นตอนการเลือกยีนโดยอาศัยค่า HCV แสดงดังภาพประกอบ 3.12

ขั้นตอนหลักที่ 4
4.1) รับค่า Highest Certainty Threshold (HCV_T)
4.2) เลือกยีนที่มีค่า HCV มากกว่าหรือเท่ากับ ค่า HCV_T

ภาพประกอบ 3.12 ขั้นตอนการทำงานในขั้นตอนหลักที่ 4

จากขั้นตอนหลักที่ 4 ถ้ากำหนดให้ค่า HCV_T เท่ากับ 90 จากตารางที่ 3.7 จะได้ว่ายีน 7447 และยีนหมายเลข 17 เท่านั้นที่จะถูกเลือกและไปสร้างเป็นกฎต่อไป เนื่องจากมีค่า HCV เป็น 91.67 และ 95.55 ตามลำดับ ซึ่งมีค่ามากกว่าค่า HCV_T ส่วนยีนหมายเลข 6983 และ 7448 จะไม่ถูกเลือก ยีนที่ถูกเลือกแสดงได้ดังตารางที่ 3.8

ตารางที่ 3.8 แสดงหมายเลขยีนที่ถูกเลือกในขั้นตอนหลักที่ 4

หมายเลขยีน	ค่า HCV
7447	91.67
17	95.55

3.1.5 ขั้นตอนหลักที่ 5: ขั้นตอนการสร้างกฎการแบ่งกลุ่ม (Rules Creation Process)

ในขั้นตอนหลักที่ 5 เป็นขั้นตอนที่ทำการสร้างกฎสำหรับการทำนายโรค จะเลือกกฎที่จะนำมาสร้างโดยพิจารณาจำนวน Fold ที่ยีนนั้นถูกเลือก ส่วนจะเลือกกฎที่มีการ Intersection กี่เปอร์เซ็นต์ จะดูจากค่า Gene Intersection Threshold (GIT) โดยจะเลือกยีนที่มีการ Intersection กันมากกว่าหรือเท่ากับค่า GIT ขั้นตอนหลักที่ 5 แสดงดังภาพประกอบ 3.13

ขั้นตอนหลักที่ 5
5.1 รับค่า Gene Intersection Threshold (GIT) โดยรับค่าในช่วง 1-100 % 5.2 คำนวณค่า Intersection ที่จะนำมาใช้เลือกยีนในการสร้างกฎดังนี้ $Intersection = (GIT * Fold)/100$ กำหนดให้ Intersection คือ ค่า intersection ที่ใช้ในการเลือกยีนเพื่อสร้างกฎการแบ่งกลุ่ม Fold คือ จำนวน Fold 5.3 นับความถี่ที่ยีนแต่ละยีนถูกเลือกครบทุก Fold 5.4 เลือกยีนที่มีจำนวนครั้งที่เลือก มากกว่าหรือเท่ากับ ค่า Intersection มาสร้างกฎ ในรูปแบบ ถ้า-แล้ว (IF – THEN)

ภาพประกอบ 3.13 ขั้นตอนการทำงานในขั้นตอนหลักที่ 5

จากตารางที่ 3.8 สมมติว่ายีนทั้ง 2 มีการถูกเลือกที่แตกต่างกันในแต่ละ Fold ซึ่งแสดงดังตารางที่ 3.9

ตารางที่ 3.9 แสดงผลลัพธ์ของการเลือกยีนในแต่ละ Fold

หมายเลข Fold (Fold Number)	หมายเลขยีน (Gene Number)
1	17
2	7447,17
3	7447,17
4	17

จากตารางที่ 3.9 ถ้าสมมติว่าในขั้นตอนที่ 5.1 กำหนดค่า GIT เท่ากับ 70% ดังนั้นจะหายีนที่ถูกเลือกได้จากขั้นตอนที่ 5.2 ดังนี้

$$Intersection = (GIT * Fold)/100$$
$$= (70 * 4)/100$$
$$= 2.8 \approx 3$$

ดังนั้นยีนที่จะถูกเลือกนั้นจะต้องถูกเลือก 3 ครั้ง หมายความว่าถูกเลือกใน 3 ใน 4 Fold นั้นเอง ดังนั้นจากตารางที่ 3.9 ความถี่ที่ยีนแต่ละยีนถูกเลือกเป็นดังตารางที่ 3.10

ตารางที่ 3.10 แสดงความถี่ที่ยีนแต่ละยีนถูกเลือก

หมายเลขยีน (Gene Number)	ความถี่หรือจำนวน Fold ที่ยีนถูกเลือก
7447	2
17	4

จากตารางที่ 3.10 ยีนที่ถูกเลือกจะมียีนเดียวคือ ยีนหมายเลขที่ 17 เนื่องจากมีความถี่เท่ากับ 4 ซึ่งมากกว่าค่า Intersectoin คือ 3 ที่คำนวณได้ แต่ยีน 7447 จะไม่ถูกเลือกเพราะมีความถี่เพียงแค่ 2 เท่านั้น

ต่อมาจะนำยีนที่เลือกได้คือ ยีนหมายเลขที่ 17 ไปสร้างเป็นกฎในรูปแบบ “ถ้า-แล้ว” เช่น IF Gene17 < 53.22 THEN Turmor เป็นต้น

3.1.6 ขั้นตอนหลักที่ 6: ขั้นตอนการหาค่าความถูกต้อง

ขั้นตอนหลักที่ 6 นี้เป็นขั้นตอนสุดท้ายในแบบจำลอง ซึ่งจะทำการทดสอบประสิทธิภาพของกฎที่สร้างได้จากขั้นตอนหลักที่ 5 โดยจะทำการทดสอบกับชุดข้อมูลทดสอบ (Test Set) อีกชุดหนึ่งที่ได้แบ่งไว้ในขั้นตอนหลักที่ 1 และจะวัดประสิทธิภาพของแบบจำลองโดยใช้ Confusion Matrix ขั้นตอนหลักที่ 6 แสดงดังภาพประกอบ 3.14

ขั้นตอนหลักที่ 6
6.1 นำกฎที่สร้างได้จากขั้นตอนหลักที่ 5 มาทดสอบกับข้อมูลทดสอบ (Test Set) ที่แบ่งไว้ตั้งแต่ขั้นตอนหลักที่ A1 โดยจะทดสอบจนครบทุก Fold
6.2 สร้าง Confusion Matrix
6.3 คำนวณค่าความถูกต้อง (Accuracy)

ภาพประกอบ 3.14 ขั้นตอนการทำงานในขั้นตอนหลักที่ 6

จากตัวอย่างกฎในขั้นตอนหลักที่ 5 เมื่อนำมาทดสอบกับชุดข้อมูลทดสอบ จะได้ผลเป็น Confusion Matrix ดังภาพประกอบ 3.15

Actual Class/ Predict Class	Normal	Tumor
Normal	4	1
Tumor	0	5

ภาพประกอบ 3.15 แสดงผลลัพธ์ของขั้นตอนหลักที่ 6 เป็น Confusion Matrix

จากภาพประกอบ 3.15 พิจารณา Confusion Matrix ที่ได้จะเห็นว่า ตัวอย่างที่นำมาทดสอบมีทั้งหมด 10 ตัวอย่าง แบ่งเป็นตัวอย่างที่เป็นโรค 5 ตัวอย่างและตัวอย่างปกติ 5 ตัวอย่าง และตัวอย่างที่เป็นโรคแบบจำลองนี้ทายถูกทั้งหมด ส่วนตัวอย่างปกติทายผิด 1 ตัวอย่าง ดังนั้นจะสามารถคำนวณค่าความถูกต้อง และค่าต่างๆ ได้ดังนี้

1) $AC = \frac{5+4}{4+1+0+5} = \frac{9}{10} = 0.9$ คิดเป็น 90%

2) $TP = \frac{5}{5+0} = \frac{5}{5} = 1$ คิดเป็น 100%

3) $FP = \frac{1}{4+1} = \frac{1}{5} = 0.2$ คิดเป็น 20%

4) $TN = \frac{4}{4+1} = \frac{4}{5} = 0.8$ คิดเป็น 80%

5) $FN = \frac{0}{0+5} = \frac{0}{5} = 0$ คิดเป็น 0%

6) $P(Positive) = \frac{5}{1+5} = 0.83$ คิดเป็น 83%

7) $P(Negative) = \frac{4}{4+0} = \frac{4}{4} = 1$ คิดเป็น 100%

จากค่าที่คำนวณได้จะเห็นว่าค่า AC TP TN P (Positive) และ P (Negative) มีค่าสูง ส่วนค่า FP และ FN มีค่าต่ำ สามารถสรุปได้ว่าแบบจำลองที่ใช้ทำนายเป็นแบบจำลองที่มีประสิทธิภาพ ที่มีค่าความถูกต้องเท่ากับ 90%

บทที่ 4

โปรแกรมการสกัดความรู้จากข้อมูล ดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง

วิทยานิพนธ์นี้นอกจากจะได้ทำการออกแบบแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองแล้ว ยังได้ทำการพัฒนาโปรแกรมการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองขึ้นด้วย โดยได้ออกแบบส่วนการติดต่อกับผู้ใช้ให้อยู่ในรูปแบบที่ใช้งานง่ายเป็น Graphic User Interface เพื่อให้ผู้ใช้สามารถเข้าใจและได้รับประโยชน์จากแบบจำลองนี้ได้อย่างไม่ยุ่งยาก

เพื่อให้ผู้ใช้สามารถเปรียบเทียบข้อแตกต่างและประสิทธิภาพของแบบจำลองที่พัฒนาขึ้น ผู้วิจัยจึงได้พัฒนาแบบจำลองขึ้นอีก 3 แบบจำลองเพื่อใช้ในการเปรียบเทียบการทำงานของแบบจำลองที่ผู้ใช้พัฒนาขึ้น โดยต่อไปนี้จะเรียกแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองที่ผู้วิจัยได้คิดค้นขึ้นว่าแบบจำลอง A และเรียกแบบจำลองอีก 3 แบบจำลองที่ผู้วิจัยใช้ในการเปรียบเทียบประสิทธิภาพว่าแบบจำลอง B C และ D ตามลำดับ ซึ่งเนื้อหาในบทนี้จะอธิบาย 3 ส่วนด้วยกันคือ ความแตกต่างระหว่างแบบจำลอง A B C และ D ผังการทำงานของโปรแกรม และส่วนตัวอย่างการทำงานของโปรแกรม ซึ่งจะทำการอธิบายโดยยกตัวอย่างข้อมูลและแสดงผลลัพธ์การทำงานของโปรแกรมประกอบเฉพาะส่วนของแบบจำลอง A เท่านั้น

4.1 ความแตกต่างระหว่างแบบจำลอง A B C และ D

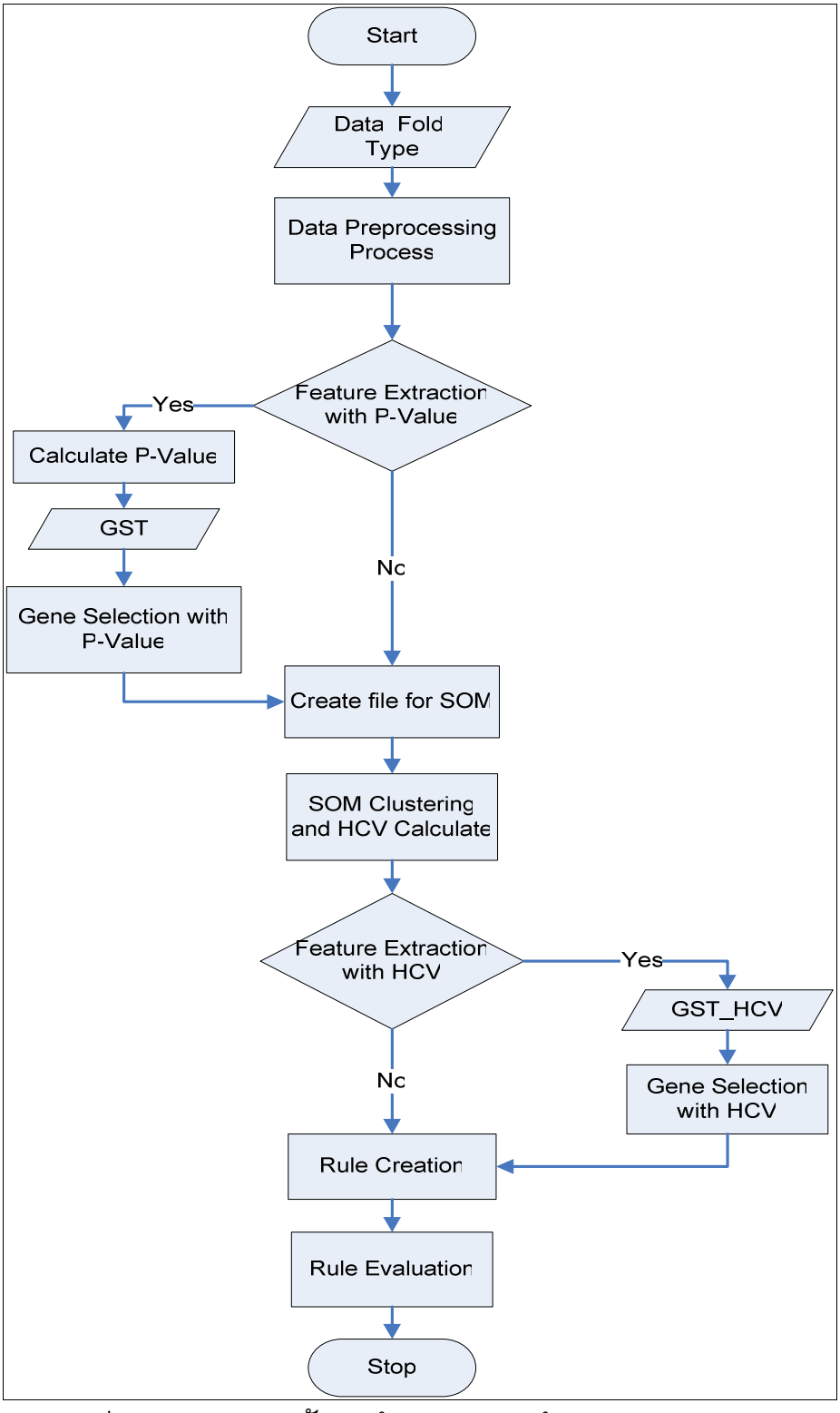
แบบจำลองทั้ง 4 แบบจำลองมีข้อแตกต่างเพื่อให้สามารถใช้ในการเปรียบเทียบประสิทธิภาพระหว่างแต่ละแบบจำลองได้ ซึ่งรายละเอียดของแต่ละแบบจำลองแสดงดังตารางที่ 4.1

ตารางที่ 4.1 แสดงความแตกต่างระหว่างแบบจำลอง A B C และ D

แบบจำลอง	การกรองข้อมูลด้วย P-Value	การกรองข้อมูล HCV
A	✓	✓
B	✗	✓
C	✓	✗
D	✗	✗

4.2 ผังการทำงานของโปรแกรม

ผังการทำงานของโปรแกรม จะแสดงทิศทางการทำงานทั้งหมด ซึ่งจะแสดงให้เห็นถึงเงื่อนไขที่เกี่ยวข้องกับการเลือกแบบจำลองด้วย แสดงดังภาพประกอบ 4.1



ภาพประกอบ 4.1 แสดงผังการทำงานของแบบจำลอง DNA_MKESOM

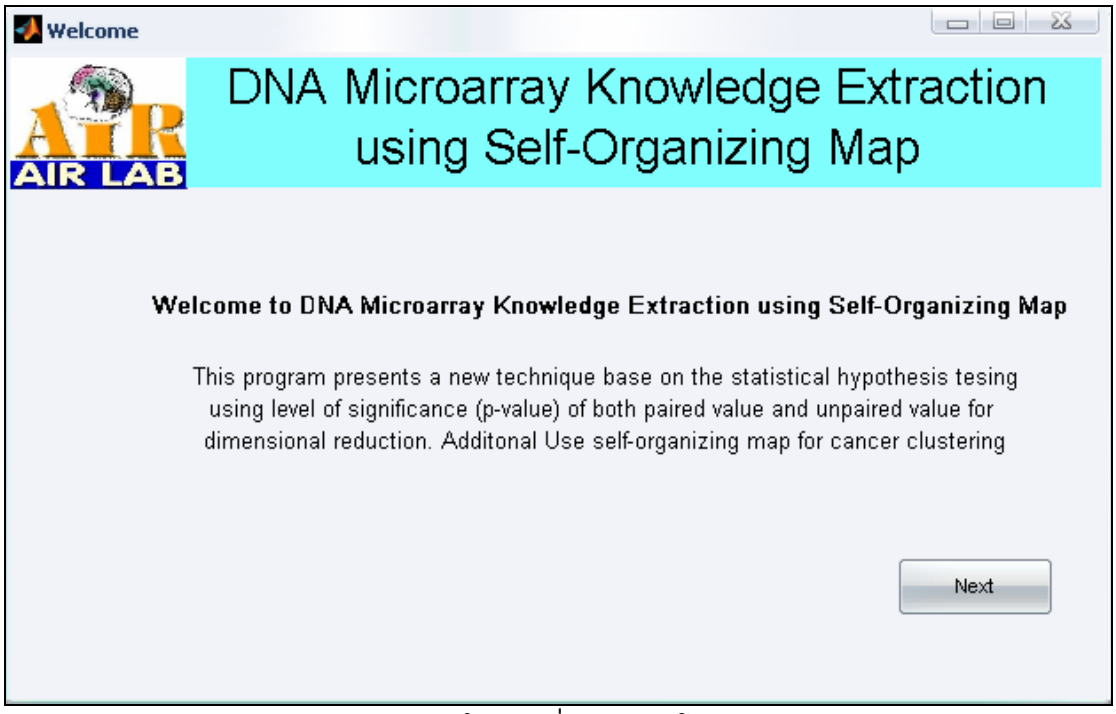
จากภาพประกอบ 4.1 เห็นได้ว่าการทำงานของโปรแกรม จะมีเงื่อนไขให้ผู้ใช้เลือกเพื่อเข้าสู่แบบจำลองทั้ง 4 แบบจำลอง โดยเลือกจากเงื่อนไขของการ Feature Extraction with P-Value และ Feature Extraction with HCV

4.3 ผลการทำงานของโปรแกรมการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง

ในหัวข้อนี้จะอธิบายขั้นตอนการทำงานของโปรแกรมการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction using Self-Organizing Map: DNA_MKESOM) โดยจะอธิบายเป็นแต่ละขั้นตอนตามลำดับดังนี้

4.3.1 ขั้นตอนที่ 1 เริ่มต้นเข้าสู่โปรแกรม

เมื่อเริ่มต้นเข้าสู่โปรแกรมจะปรากฏหน้าต้อนรับ แสดงชื่อโปรแกรมและคำอธิบายเบื้องต้น เมื่อผู้ใช้ได้อ่านคำอธิบายเรียบร้อยแล้ว ให้กดปุ่ม Next เพื่อเข้าสู่การทำงานต่อไป คือขั้นตอนการสร้างแบบจำลอง หน้าจอขั้นตอนที่ 1 แสดงได้ดังภาพประกอบ 4.2

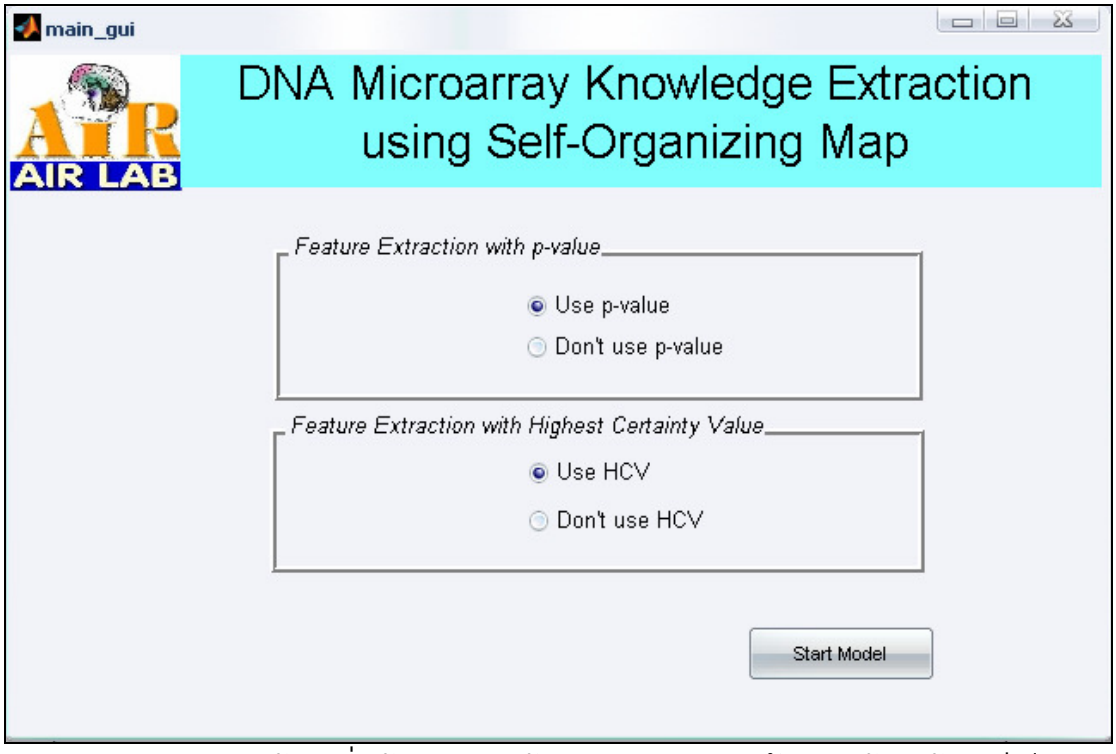


ภาพประกอบ 4.2 หน้าจอแรกในการเริ่มต้นเข้าสู่โปรแกรม DNA_MKESOM

4.3.2 ขั้นตอนที่ 2 เลือกเงื่อนไขในการสร้างแบบจำลอง

เมื่อเข้าสู่ขั้นตอนที่ 2 จะปรากฏหน้าจอเริ่มต้นของการสร้างแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง ซึ่งแบบจำลองนี้จะอนุญาตให้ผู้ใช้สามารถเลือกเงื่อนไขในการสร้างแบบจำลองได้ โดยจะมี 2 เงื่อนไขคือ ในขั้นตอน Feature Extraction with P-Value จะให้ผู้ใช้เลือกว่าจะให้มีขั้นตอนการเลือกยีนโดยใช้ค่า P-Value ในแบบจำลองหรือไม่ โดยค่าเริ่มต้นจะตั้งไว้ให้มีการใช้ค่า P-Value ในการเลือกยีน และในขั้นตอน Feature Extraction with Highest Certainty Value จะให้ผู้ใช้เลือกว่าจะให้มีขั้นตอนการเลือกยีนโดยใช้ค่า Highest Certainty Value (HCV) ในแบบจำลองหรือไม่ โดยค่าเริ่มต้นจะตั้งไว้ให้มีการใช้ค่า HCV ในการเลือกยีน ซึ่งการเลือกเงื่อนไขทั้งสองจะมีผลต่อการเลือกแบบจำลองดังแสดงไว้ในตารางที่ 4.1

เมื่อเลือกเงื่อนไขครบทั้ง 2 เงื่อนไขแล้ว ให้กดปุ่ม Start Model เพื่อสร้างแบบจำลอง และทำการทดลองในขั้นตอนต่อไป หน้าจอขั้นตอนที่ 2 แสดงได้ดัง ภาพประกอบ 4.3

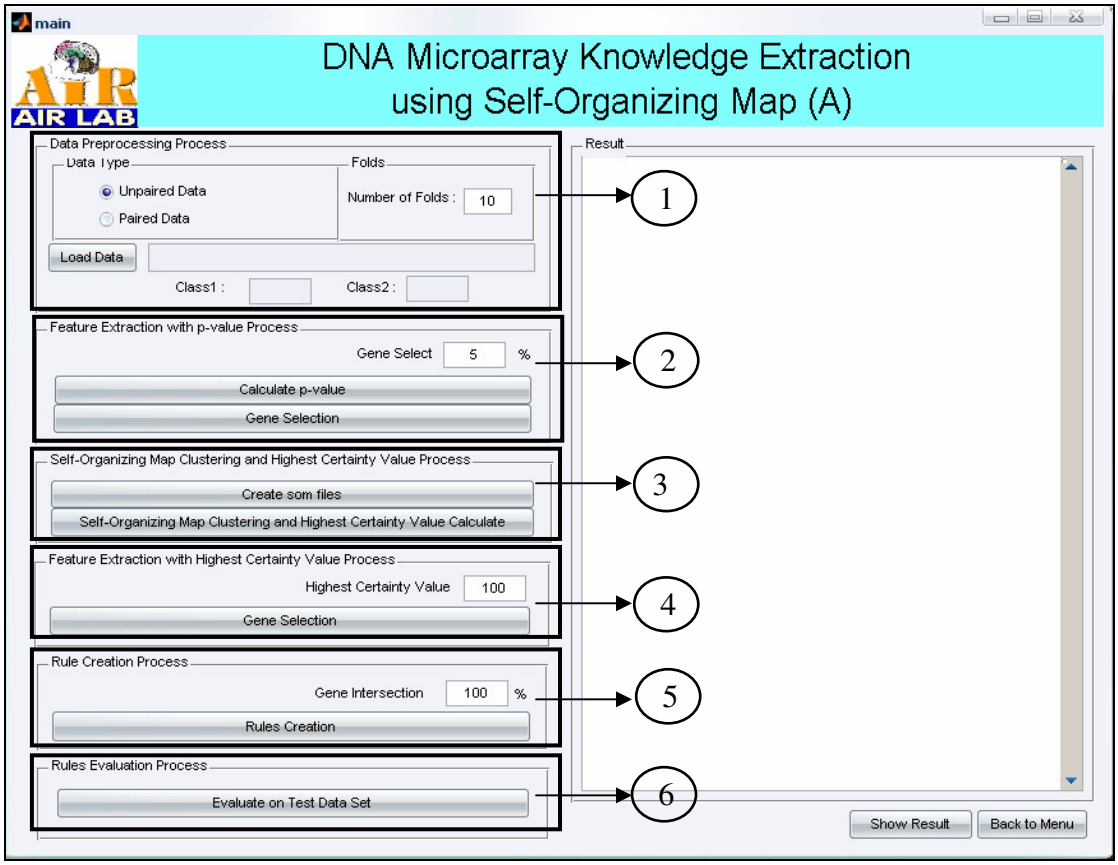


ภาพประกอบ 4.3 หน้าจอเริ่มต้นของการสร้างแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง

4.3.3 ขั้นตอนที่ 3 เข้าสู่หน้าจอหลักของแบบจำลอง

ขั้นตอนนี้จะเป็นขั้นตอนการทำงานของแบบจำลองที่เลือกไว้ในขั้นตอนที่ 2 โดยแบบจำลองจะถูกสร้างขึ้นตามเงื่อนไขที่ผู้ใช้เลือกในขั้นตอนที่ 4.3.2 ซึ่งในที่นี้จะอธิบาย

เฉพาะแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองที่ได้ออกแบบขึ้นเท่านั้น (แบบจำลอง A) ส่วนแบบจำลองที่เหลือคือ B C และ D จะใช้สำหรับเปรียบเทียบผลการทดลองเท่านั้น ซึ่งประกอบด้วย 6 ขั้นตอนหลัก หน้าจอของแบบจำลอง A แสดงได้ดังภาพประกอบ 4.4



ภาพประกอบ 4.4 หน้าจอโปรแกรม DNA_MKESOM (A)

จากภาพประกอบ 4.4 จะเห็นว่าแบบจำลอง A ประกอบด้วยขั้นตอนการทำงานหลัก 6 ขั้นตอน ดังนี้

ขั้นตอนหลักที่ 1: ขั้นตอนการเตรียมข้อมูล (Data Preprocessing Process)

ขั้นตอนหลักที่ 2: ขั้นตอนการเลือกยีนโดยอาศัยค่าสถิติลำดับความสำคัญ (Feature Extraction with P-Value Process)

ขั้นตอนหลักที่ 3: ขั้นตอนการแบ่งกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง และขั้นตอนการคำนวณค่า Highest Certainty Value (Self-Organizing Map Clustering and Highest Certainty Value Process)

ขั้นตอนหลักที่ 4: ขั้นตอนการเลือกยีนโดยอาศัยค่า Highest Certainty Value (Feature Extraction with Highest Certainty Value)

ขั้นตอนหลักที่ 5: ขั้นตอนการสร้างกฎการแบ่งกลุ่ม (Rules Creation Process)

ขั้นตอนหลักที่ 6: ขั้นตอนการทดสอบประสิทธิภาพของกฎ (Rules Evaluation Process)

4.3.4 ขั้นตอนที่ 4 ขั้นตอนการทำงานในแบบจำลอง

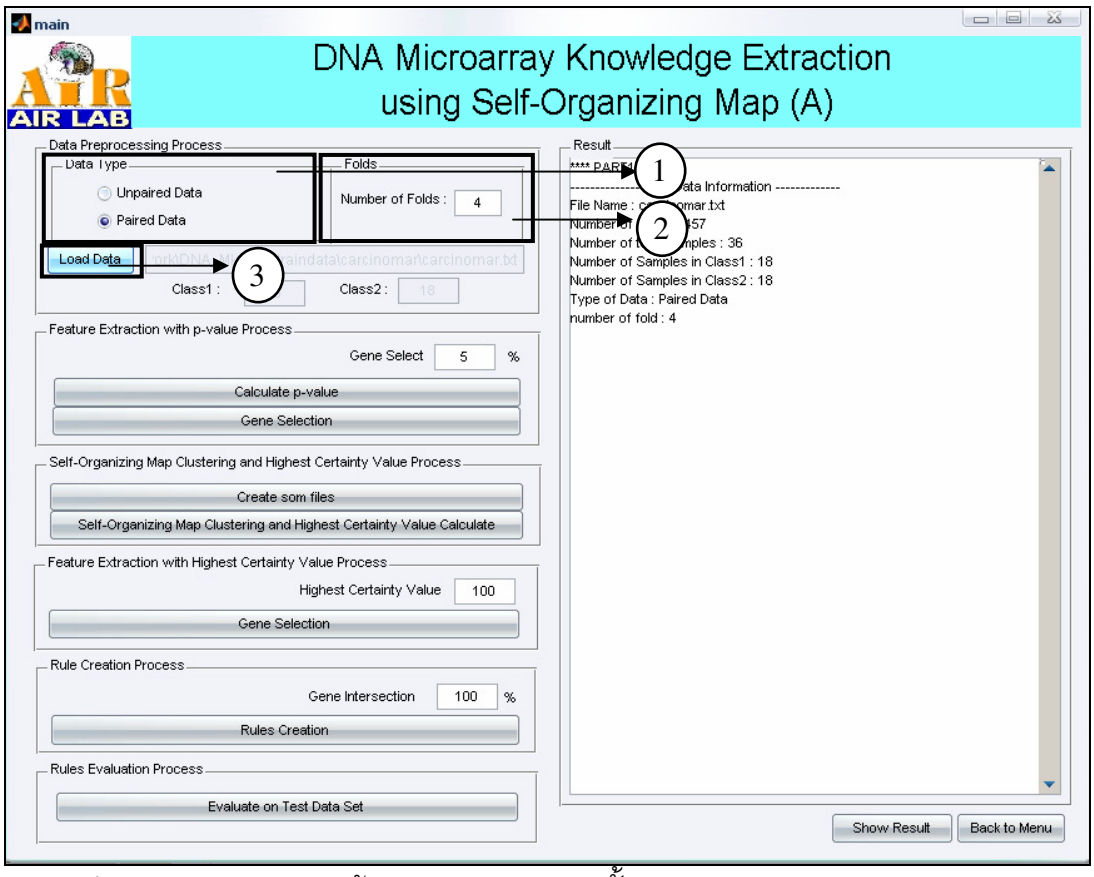
ในที่นี้จะอธิบายเฉพาะขั้นตอนการทำงานของแบบจำลอง A เท่านั้น เนื่องจากเป็นแบบจำลองที่มีขั้นตอนการทำงานครบทั้ง 6 ขั้นตอนหลักที่ได้ออกแบบไว้ในบทที่ 3 ของวิทยานิพนธ์นี้ และยังเป็นแบบจำลองที่มีประสิทธิภาพมากที่สุดอีกด้วย ส่วนขั้นตอนการทำงานแต่ละขั้นตอนจะอธิบายโดยใช้ตัวอย่างข้อมูลของโรคมะเร็งในเนื้องอก และอธิบายโดยใช้รูปภาพแสดงการทำงานประกอบ ดังนี้

ขั้นตอนที่ 1: Data Preprocessing Process ขั้นตอนนี้เป็นขั้นตอนแรกของแบบจำลอง จะทำการรับค่าข้อมูลและเตรียมข้อมูลให้พร้อมสำหรับการทำงานในขั้นตอนต่อไป ซึ่งมีขั้นตอนการทำงานย่อย 3 ขั้นตอนดังนี้

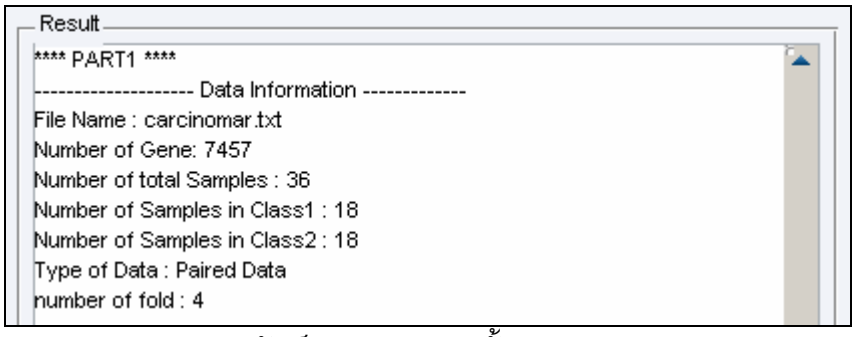
1) รับค่ารูปแบบของข้อมูล (Data Type)

2) รับค่า Fold เพื่อใช้ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอน (Train Set) และชุดข้อมูลทดสอบ (Test Set)

3) เมื่อเสร็จเรียบร้อยแล้วให้กดปุ่ม Load Data เพื่ออ่านข้อมูล หน้าจอแสดงการทำงานขั้นตอนที่ 1 แสดงได้ดังภาพประกอบ 4.5 และ ผลลัพธ์ที่ได้จากขั้นตอนที่ 1 แสดงได้ดังภาพประกอบ 4.6



ภาพประกอบ 4.5 แสดงหน้าจอแสดงการทำงานขั้นตอน Data Preprocessing Process



ภาพประกอบ 4.6 แสดงผลลัพธ์จากการทำงานขั้นตอน Data Preprocessing Process

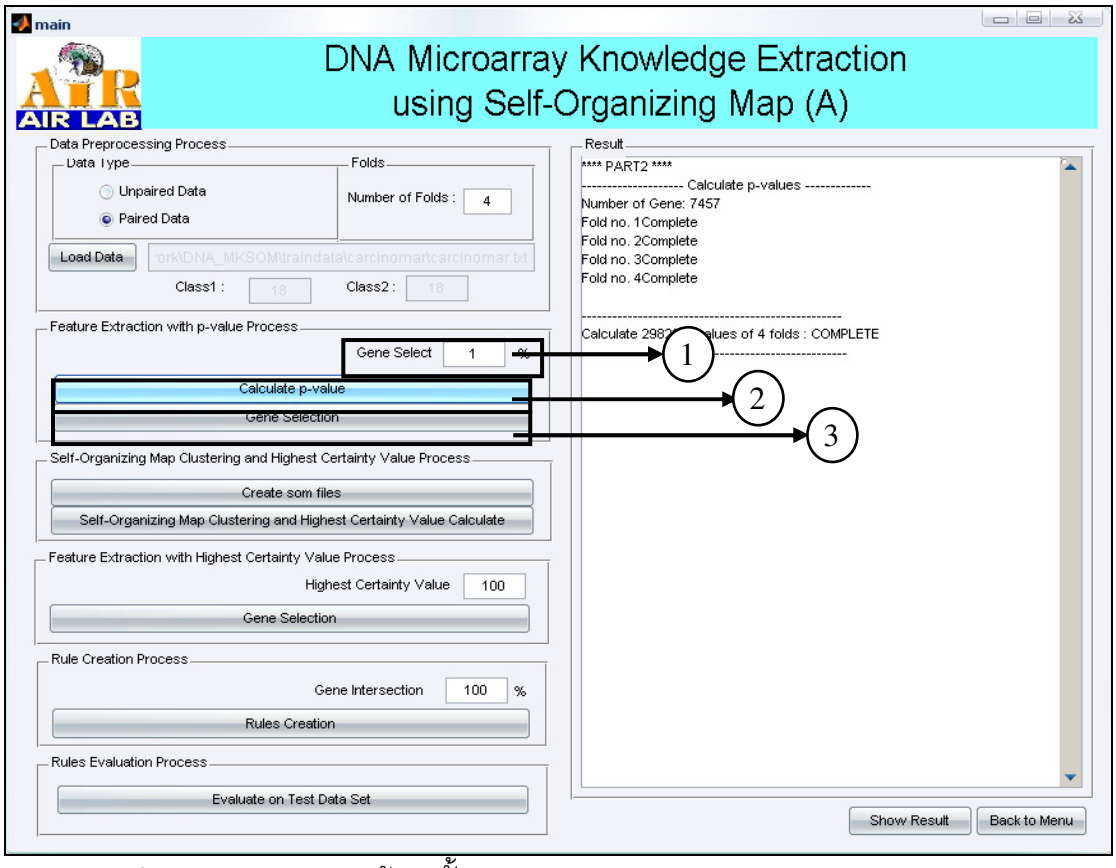
จากภาพประกอบ 4.6 จะเห็นว่าเมื่ออ่านข้อมูลเรียบร้อยแล้วระบบจะแสดงรายละเอียดของข้อมูล เช่น ชื่อข้อมูล จำนวนยีน จำนวนตัวอย่าง และเงื่อนไขในการทดสอบตามลำดับ

ขั้นตอนที่ 2: Feature Extraction with P-Value Process ขั้นตอนนี้เป็นการเลือกยีนโดยอาศัยค่า P-Value ประกอบด้วยขั้นตอนย่อย 3 ขั้นตอน ดังนี้

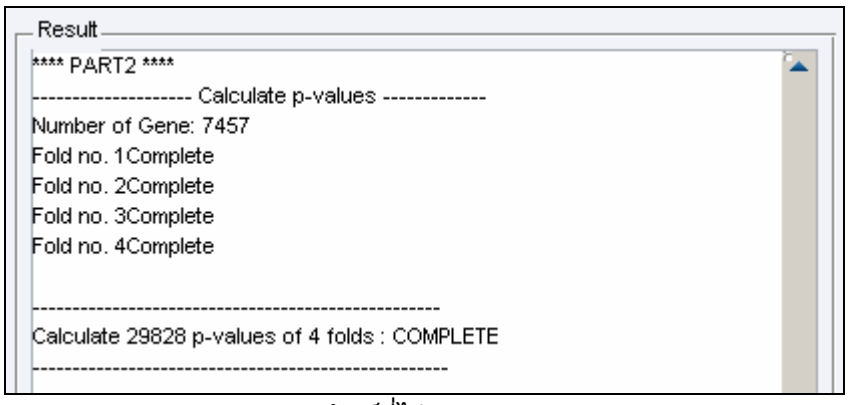
1) ป้อนค่าจำนวนยีนที่ต้องการเลือก (Gene Selection Threshold: GST) ในที่นี้จะให้ป้อนเป็นเปอร์เซ็นต์ โดยสามารถป้อนค่าได้ตั้งแต่ 1-100 โดยปกติค่าที่จะป้อนมักจะอยู่ระหว่าง 1 – 5 เปอร์เซ็นต์ ซึ่งในที่นี้ขึ้นอยู่กับจำนวนยีนที่มีในแต่ละโรคด้วย

2) กดปุ่ม Calculate P-Value เพื่อคำนวณค่า P-Value ของทุกยีน

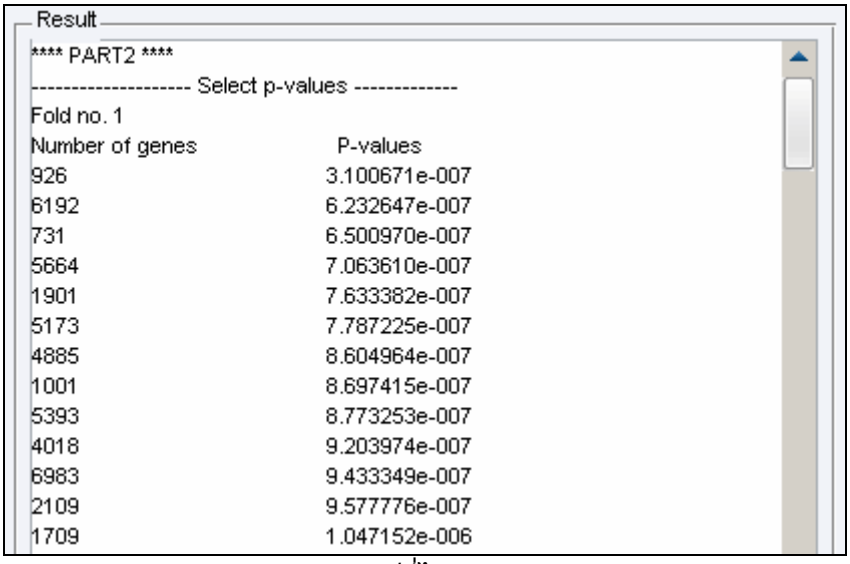
3) กดปุ่ม Gene Selection เพื่อเลือกยีนที่มีค่า P-Value ต่ำ มาตามจำนวนที่ป้อนค่าในขั้นตอนที่ 1 ภาพแสดงหน้าจอขั้นตอนที่ 2 แสดงได้ดังภาพประกอบ 4.7 ส่วนภาพแสดงผลลัพธ์ที่ได้จากการกดปุ่ม Calculate P-Value และ ปุ่ม Gene Selection แสดงได้ดังภาพประกอบ 4.8 และ 4.9 ตามลำดับ



ภาพประกอบ 4.7 แสดงหน้าจอขั้นตอน Feature Extraction with P-Value Process



ภาพประกอบ 4.8 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Calculate P-Value



ภาพประกอบ 4.9 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Gene Selection

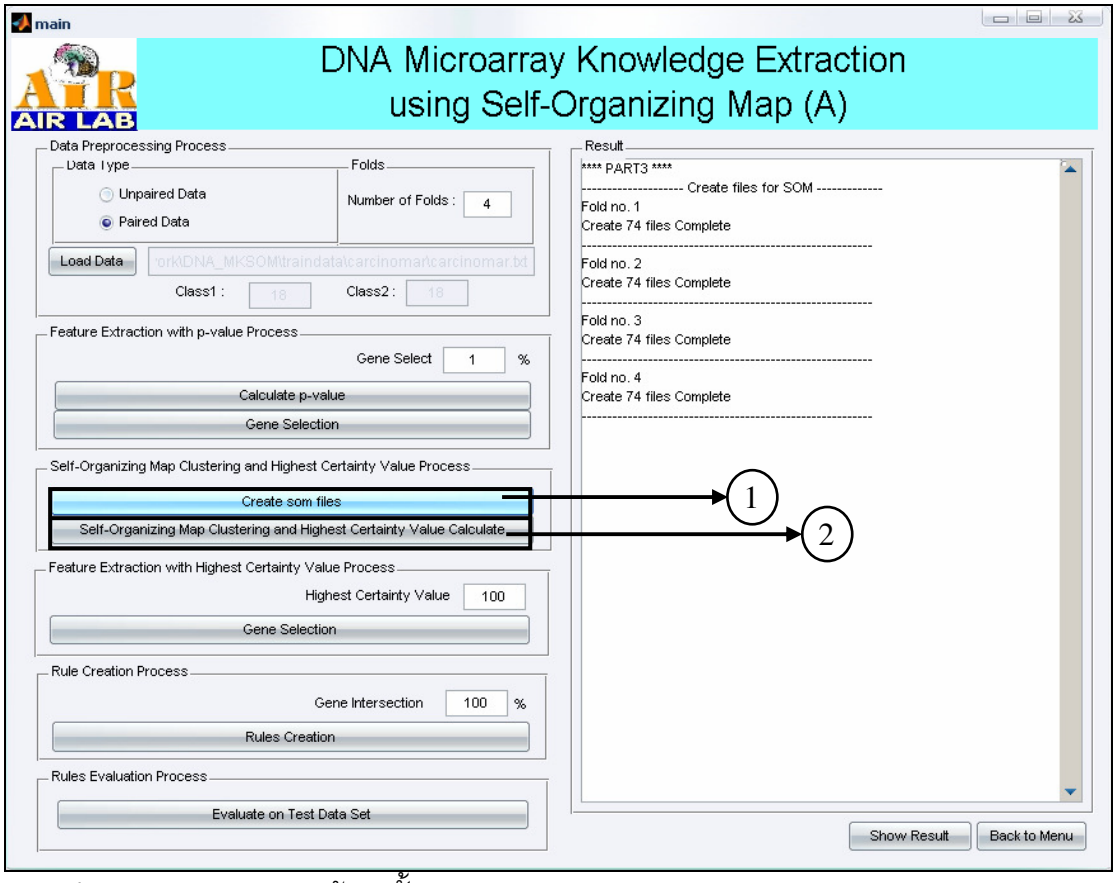
จากภาพประกอบที่ 4.8 จะเห็นว่าแสดงจำนวนยีนที่ทำการคำนวณค่า P-Value เรียบร้อยแล้วของทุก Fold แสดงว่าขั้นตอนการคำนวณค่า P-Value เสร็จสมบูรณ์แล้ว และจากภาพประกอบ 4.8 จะเห็นว่าเมื่อกดปุ่ม Gene Selection ระบบจะทำการเลือกยีนที่มีค่า P-Value มา จำนวนเท่ากับค่าที่ผู้ใช้ป้อนเข้าไป โดยจะแสดงหมายเลขยีนและค่า P-Value ของยีนที่ถูกเลือกในแต่ละ Fold แสดงว่าขั้นตอนการเลือกยีนเสร็จสมบูรณ์แล้ว

ขั้นตอนที่ 3: Self-Organizing Map and Highest Certainty Value Process ขั้นตอนนี้เป็นการแบ่งกลุ่มข้อมูลโดยใช้แผนที่การจัดกลุ่มเอง (SOM) ประกอบด้วยขั้นตอนย่อยดังนี้

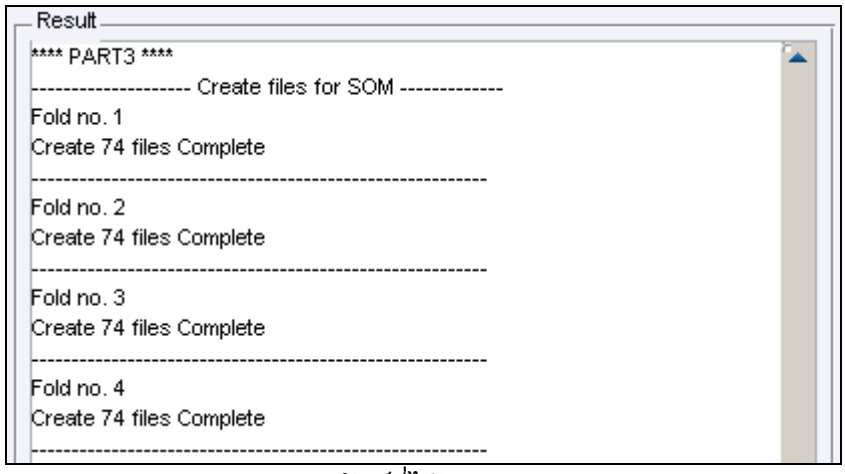
1) สร้างไฟล์ข้อมูลที่ใช้สำหรับจัดกลุ่มด้วย SOM

2) แบ่งกลุ่มข้อมูลโดยใช้ SOM ภาพแสดงหน้าจอขั้นตอนที่ 3 แสดงได้ดังภาพประกอบ 4.10 ส่วนภาพแสดงผลลัพธ์ที่ได้จากการกดปุ่ม Create SOM Files และ ปุ่ม

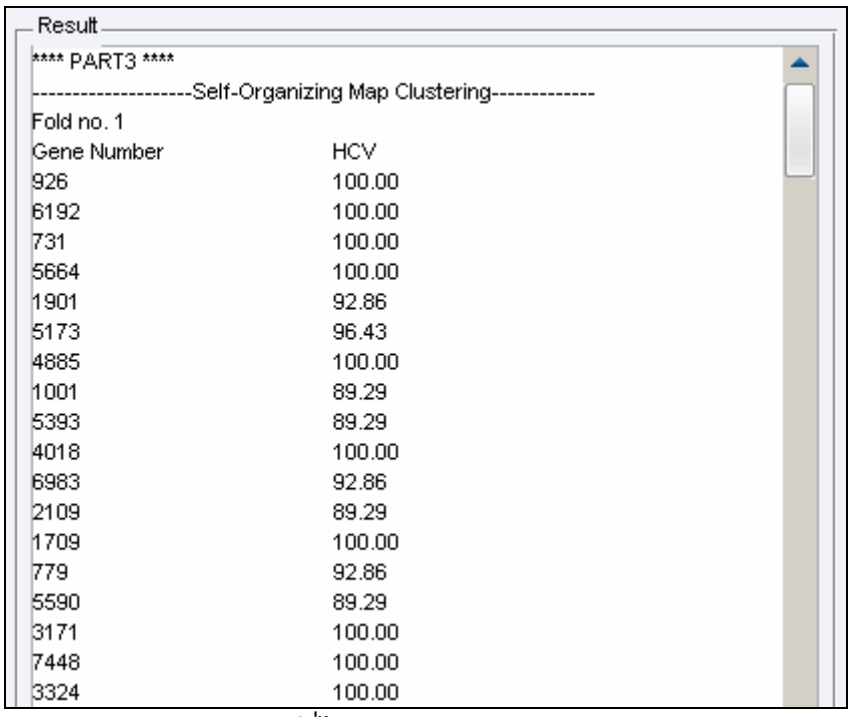
Self-Organizing Map and Highest Certainty Value แสดงได้ดังภาพประกอบ 4.11 และ 4.12 ตามลำดับ



ภาพประกอบ 4.10 แสดงหน้าจอขั้นตอน Self-Organizing Map and Highest Certainty Value



ภาพประกอบ 4.11 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Create SOM Files



ภาพประกอบ 4.12 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Self-Organizing Map and Highest Certainty Value

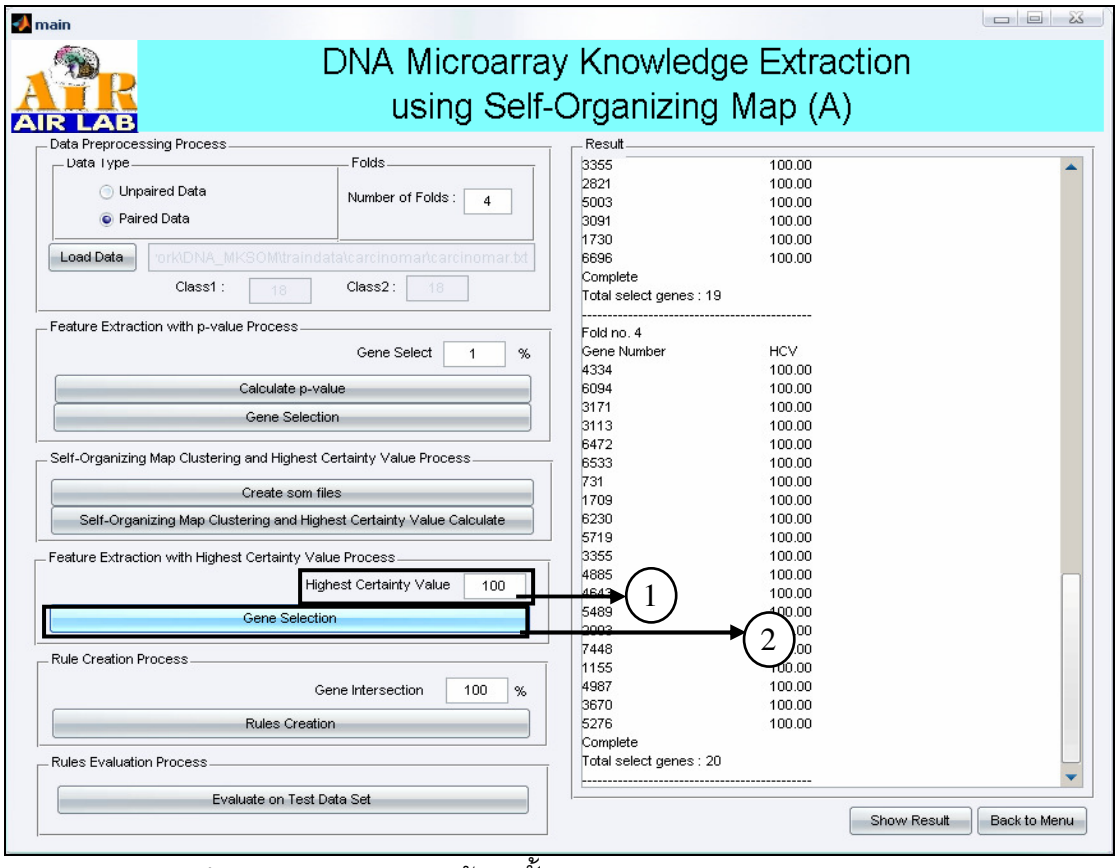
จากภาพประกอบ 4.11 จะเห็นว่าเมื่อกดปุ่ม Create SOM Files ระบบจะทำการสร้างไฟล์ข้อมูลสำหรับแบ่งกลุ่มด้วย SOM ขึ้น และเมื่อสร้างเสร็จเรียบร้อยแล้วจะแสดงผลลัพธ์เพื่อบอกว่าได้สร้างไฟล์เสร็จเรียบร้อยทั้งหมดกี่ไฟล์ในแต่ละ Fold แสดงว่าขั้นตอนการสร้างไฟล์เสร็จสมบูรณ์แล้ว

จากภาพประกอบ 4.12 จะเห็นว่าเมื่อกดปุ่ม Self-Organizing Map and Highest Certainty Value ระบบจะอ่านไฟล์ที่สร้างขึ้นเพื่อนำข้อมูลไปจัดกลุ่มด้วย SOM พร้อมทั้งคำนวณค่า HCV และเมื่อเสร็จเรียบร้อยแล้วจะแสดงผลลัพธ์เป็น หมายเลขยีนและค่า HCV ที่คำนวณได้ แสดงว่าขั้นตอนการจัดกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเองและการคำนวณค่า HCV เสร็จสมบูรณ์แล้ว

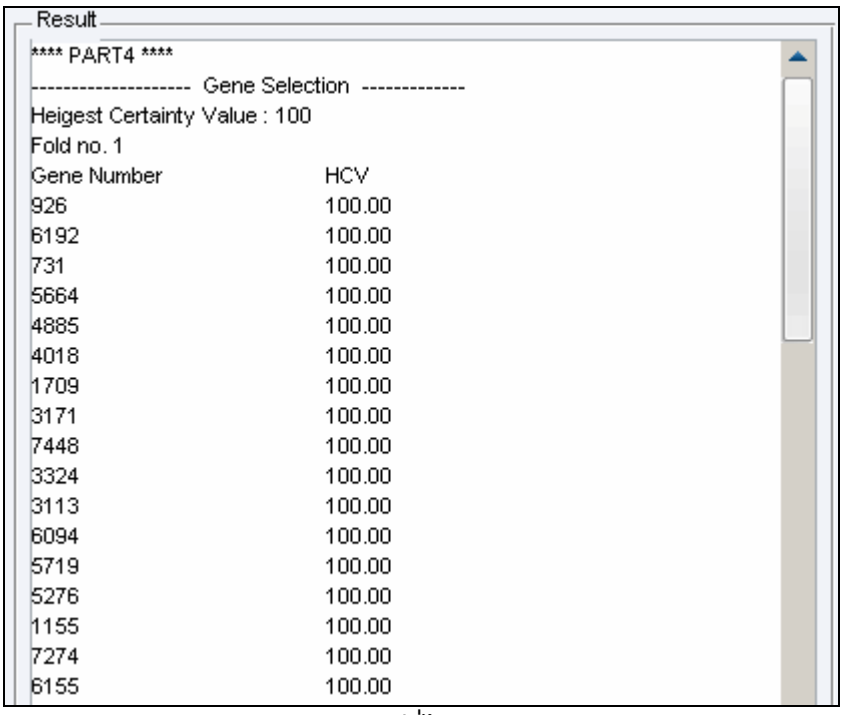
ขั้นตอนที่ 4: Feature Extraction with Highest Certainty Value Process ขั้นตอนนี้เป็นการเลือกยีนโดยอาศัยค่า Highest Certainty Value Threshold (HCV_T) ประกอบด้วยขั้นตอนย่อย 2 ขั้นตอน ดังนี้

1) ป้อนค่า HCV_T ที่ต้องการ โดยสามารถป้อนค่าได้ตั้งแต่ 1-100% โดยปกติค่าที่จะป้อนมักจะอยู่ระหว่าง 90% - 100%

2) กดปุ่ม Gene Selection เพื่อเลือกเฉพาะยีนที่มีค่า HCV มากกว่าหรือเท่ากับค่า HCV_T ที่ต้องการเท่านั้น การทำงานในขั้นตอนที่ 4 แสดงดังภาพประกอบ 4.13 ส่วนผลลัพธ์ที่ได้จากการกดปุ่ม Gene Selection แสดงได้ดังภาพประกอบ 4.14

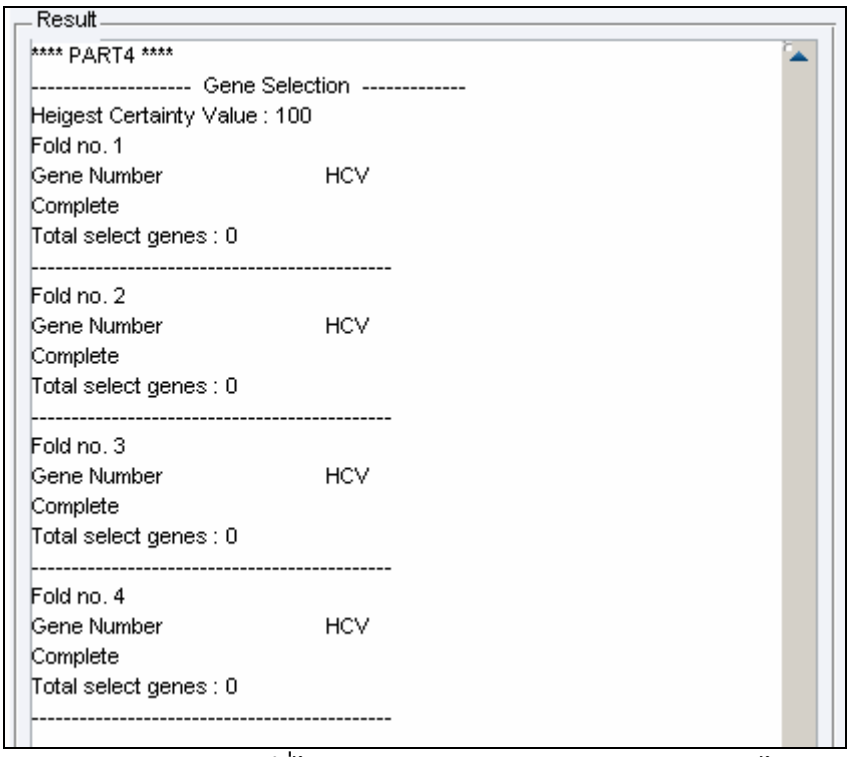


ภาพประกอบ 4.13 แสดงหน้าจอขั้นตอน Feature Extraction with HCV



ภาพประกอบ 4.14 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Gene Selection

จากภาพประกอบ 4.14 จะเห็นว่าเมื่อกดปุ่ม Gene Selection ระบบจะทำการเลือกเฉพาะยีนที่มีค่า HCV มากกว่าหรือเท่ากับค่า HCV_T ที่ผู้ใช้ป้อนเท่านั้น และจะแสดงผลลัพธ์หมายเลขของยีนที่ค่า HCV ของยีนที่เลือกด้วย ซึ่งถ้าไม่มียีนที่มีค่า HCV ตรงตามเงื่อนไขที่ผู้ใช้ป้อนก็จะไม่มีการแสดงค่ายีนที่ถูกเลือก ดังภาพประกอบ 4.15



ภาพประกอบ 4.15 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Gene Selection กรณีไม่มียีนที่เลือกได้

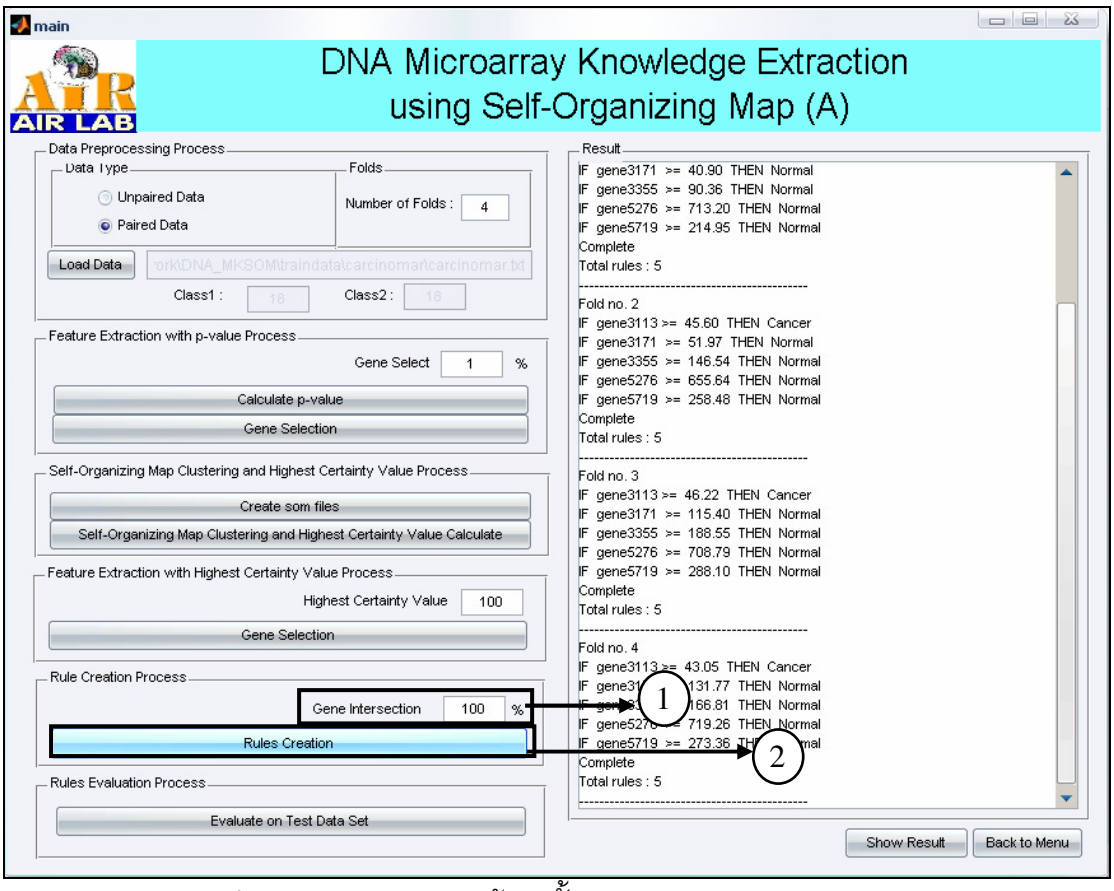
จากภาพประกอบ 4.15 จะเห็นว่าไม่มียีนที่มีค่า HCV ที่มากกว่าหรือเท่ากับค่า HCV_T ที่ต้องการให้ทำการปรับเปลี่ยนค่า HCV_T โดยปรับให้ลดลง

ขั้นตอนที่ A5: Rule Creation Process ขั้นตอนนี้เป็นการสร้างกฎสำหรับยีนที่เลือกมาได้จากการทำงานในขั้นตอนที่ 1–4 โดยจะแสดงในรูปแบบของกฎ “ถ้า–แล้ว” (IF–THEN Rule) โดยในการเลือกจะอาศัยค่า Intersection ที่ผู้ใช้ป้อนเป็นการกำหนดว่ายีนที่จะถูกสร้างเป็นกฎนั้นจะต้องถูกเลือกในกี่ Fold จากการทำงานทั้งหมด เช่น ถ้ากำหนดเงื่อนไขการทดสอบว่าใช้ 10 Folds และกำหนดค่า Intersection ในขั้นตอนนี้เป็น 80% หมายความว่ายีนนั้นจะต้องถูกเลือกไม่น้อยกว่า 8 ใน 10 Folds ซึ่งจะประกอบด้วยขั้นตอนย่อยในการทำงาน 2 ขั้นตอน ดังนี้

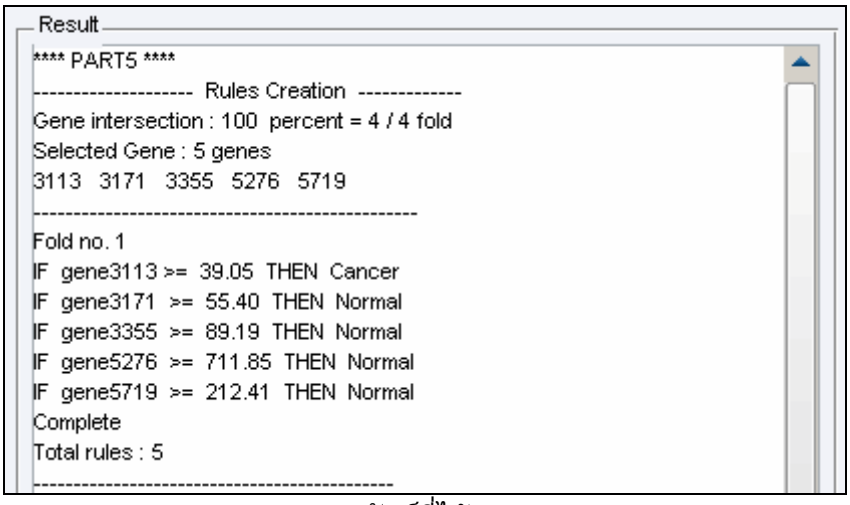
1) ป้อนค่า Gene Intersection ที่ต้องการ โดยสามารถป้อนค่าได้ตั้งแต่ 1% -100%

2) กดปุ่ม Rule Creation สร้างกฎสำหรับยีนที่เลือกได้

การทำงานในขั้นตอนที่ 5 แสดงดังภาพประกอบ 4.16 ส่วนภาพแสดงผลลัพธ์ที่ได้จากการกดปุ่ม Rule Creation แสดงได้ดังภาพประกอบ 4.17



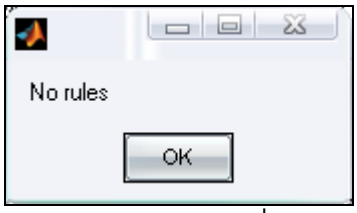
ภาพประกอบ 4.16 แสดงหน้าจอขั้นตอน Rules Creation Process



ภาพประกอบ 4.17 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Rule Creation

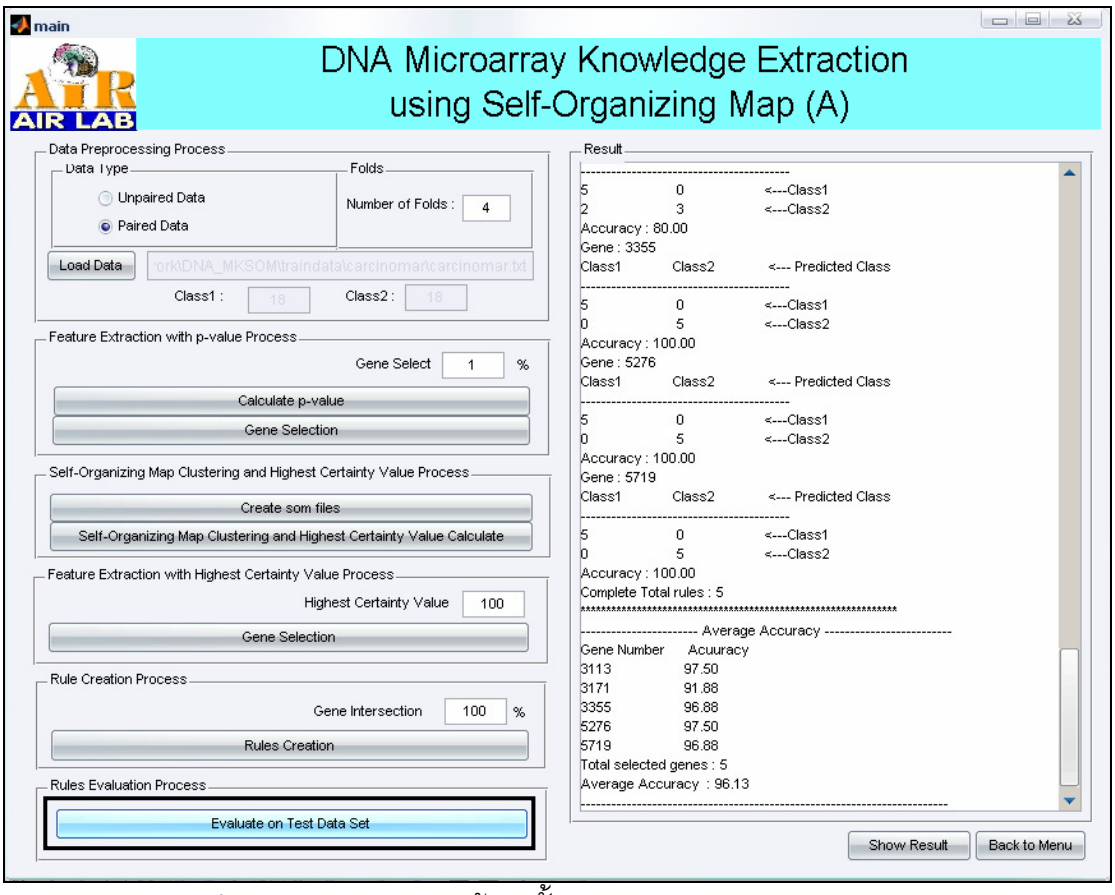
จากภาพประกอบ 4.17 จะเห็นว่าเมื่อกดปุ่ม Rule Creation จะแสดงหมายเลขยีนทั้งหมดที่ถูกเลือก และทำการสร้างกฎในรูปแบบ “ถ้า–แล้ว” (IF–THEN Rule) ขึ้นและแสดงผลลัพธ์ของกฎเหล่านั้น ส่วนกรณีที่ไม่มีกฎใดที่สามารถสร้างได้ตามเงื่อนไขของค่า Gene

Intersection Threshold (GIT) ที่ผู้ใช้ป้อนจะแสดงข้อความขึ้นว่า “No Rule” แสดงดังภาพประกอบ 4.18 หมายถึงไม่มีกฎที่สามารถสร้างได้ ให้ทำการเปลี่ยนแปลงค่า GIT ใหม่ โดยปรับให้น้อยลง

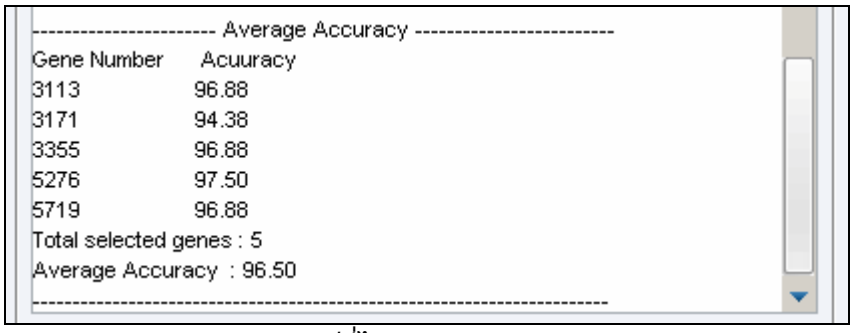


ภาพประกอบ 4.18 แสดงข้อความเตือนเมื่อไม่มียีนที่มีค่า GIT ตามที่ต้องการ

ขั้นตอนที่ A6: Rule Evaluation Process ขั้นตอนนี้เป็นขั้นตอนสุดท้ายในแบบจำลอง โดยจะเป็นขั้นตอนในการวัดประสิทธิภาพของกฎว่ากฎที่เลือกโดยใช้แบบจำลองนี้มีประสิทธิภาพเพียงใด โดยในการวัดจะวัดโดยการนำกฎที่ได้ไปใช้กับชุดข้อมูลทดสอบ (Test Set) อีกชุดหนึ่ง และจะคำนวณค่าความถูกต้องของกฎออกมา และนอกจากนั้นยังแสดงผลด้วย Confusion Matrix เพื่อให้สามารถสรุปผลในการทดสอบได้อย่างถูกต้องมากขึ้น การทำงานในขั้นตอนที่ 6 แสดงได้ดังภาพประกอบ 4.19 และผลลัพธ์ที่ได้จากการกดปุ่ม Evaluate On Test Data Set แสดงได้ดังภาพประกอบ 4.20



ภาพประกอบ 4.19 แสดงหน้าจอขั้นตอน Evaluate on Test Data Set

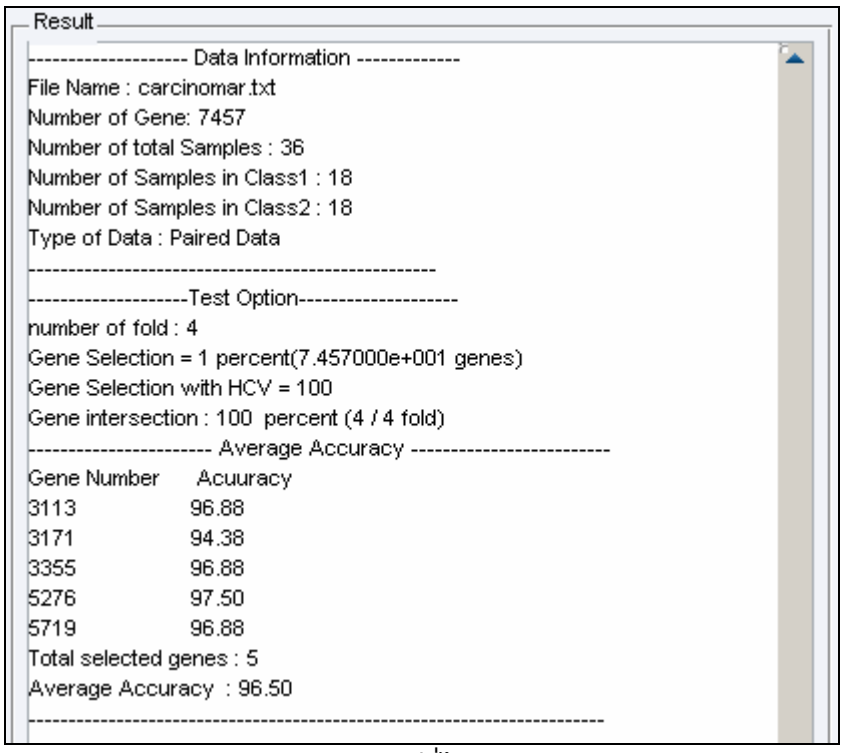


ภาพประกอบ 4.20 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Evaluate on Test Data Set

จากภาพประกอบ 4.20 จะเห็นว่าเมื่อกดปุ่ม Evaluate on Test Data Set ระบบจะนำกฎที่สร้างไปทดสอบกับชุดข้อมูลทดสอบ (Test Set) เมื่อเสร็จเรียบร้อยแล้วก็จะแสดงผลค่าความถูกต้องของยีนแต่ละตัว และค่าความถูกต้องเฉลี่ยในการทดลองครั้งนั้นๆ ด้วย

จากขั้นตอนที่กล่าวมาทั้งหมด 6 ขั้นตอนเป็นขั้นตอนการทำงานหลักของแบบจำลอง DNA_MKESOM ที่สมบูรณ์เรียบร้อยแล้ว ซึ่งผู้ใช้สามารถจบการทำงานลงหลังขั้นตอนหลักที่ 6 ได้ แต่เพื่อความง่ายในการทำความเข้าใจ หรือการสรุปผลการทดลอง จึงได้

พัฒนาขั้นตอนการดูผลสรุปของการทดสอบเพิ่มเข้ามาด้วย เพื่อให้ผู้ใช้สามารถเลือกดูผลสรุปของการทำงานทั้งหมด ได้แก่ รายละเอียดข้อมูล เงื่อนไขการทดสอบ จำนวนกฎที่ได้ และค่าความถูกต้อง ได้อีกครั้งหนึ่ง โดยเลือกที่ปุ่ม Show Result ระบบจะทำการแสดงข้อมูลทั้งหมดอีกครั้งหนึ่ง ต่อมาถ้าต้องการเริ่มต้นการทดลองใหม่ก็สามารถกดปุ่ม Back to Menu เพื่อกลับไปยังหน้าจอของเมนูเพื่อเริ่มสร้างแบบจำลองในการทำงานครั้งต่อไปได้ โดยผลลัพธ์เมื่อกดปุ่ม Show Result จะแสดงผลได้ดังภาพประกอบ 4.21



ภาพประกอบ 4.21 แสดงผลลัพธ์ที่ได้จากการกดปุ่ม Show Result

จากภาพประกอบ 4.21 จะเห็นว่าเมื่อกดปุ่ม Show Result ระบบจะทำการสรุปผลจากการทำงานที่ผ่านมาตั้งแต่ขั้นตอนแรกจนถึงขั้นตอนสุดท้ายอีกครั้งหนึ่ง เพื่อให้สามารถดูผลสรุปการทดลองได้ง่าย โดยจะเห็นว่าในผลลัพธ์จะแบ่งออกเป็น 3 ส่วนคือ ส่วนแสดงชื่อข้อมูลและรายละเอียดของข้อมูล ส่วนเงื่อนไขการทดสอบ ประกอบด้วย จำนวน Fold ค่า Gene Selection Threshold ค่า Highest Certainty Threshold และค่า Gene Intersection Threshold ส่วนสุดท้ายจะแสดงค่าความถูกต้องของแต่ละกฎ และค่าความถูกต้องเฉลี่ยของทุกกฎด้วย

โปรแกรมการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองนี้ แม้จะพัฒนาด้วยโปรแกรม MATLAB แต่ถ้าในเครื่องนั้นไม่ได้เพิ่มส่วนคำสั่งของแผนที่การจัดกลุ่มเอง (Self-Organizing Map Toolbox) ก็จะไม่สามารถใช้งานโปรแกรมนี้ได้

เนื่องจากในการพัฒนาโปรแกรมผู้จัดทำได้มีการใช้คำสั่งเฉพาะของแผนที่การจัดกลุ่มเองด้วย ดังนั้นจึงได้จัดทำวิธีการเพิ่มส่วนคำสั่งการทำงานของแผนที่การจัดกลุ่มเองไว้ด้วย ดังแสดงในภาคผนวก ก.

บทที่ 5

ผลการทดลอง

วิทยานิพนธ์ฉบับนี้ได้ทำการทดสอบการทำงานของแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง โดยใช้ฐานข้อมูลโรคมะเร็งที่เป็นสากลจากห้องวิจัย Kent Ridge มหาวิทยาลัย Pittsburgh ประเทศสหรัฐอเมริกา และโครงการวิจัยการแสดงออกของยีน (Gene Expression) มหาวิทยาลัย Princeton ประเทศสหรัฐอเมริกา

ข้อมูลที่ใช้ในการแสดงผลการทดสอบการทำงานของโปรแกรมมีทั้งหมด 4 ชุดข้อมูลด้วยกันคือ โรคมะเร็งในเนื้องอก (Carcinoma Tumor) โรคมะเร็งเม็ดเลือดขาว (Leukemia) โรคมะเร็งลำไส้ (Conlon Cancer) และโรคมะเร็งปอด (Lung Cancer) ซึ่งจะอธิบายผลการทดลองโดยแยกเป็นแต่ละโรคตามลำดับ และจะทำการทดลองทั้งหมด 4 แบบจำลองด้วยกันคือ แบบจำลอง A B C และ D ดังตารางที่ 4.1 ตามลำดับ

5.1 โรคมะเร็งในเนื้องอก (Carcinoma Tumor)

ข้อมูลของผู้ป่วยโรคมะเร็งในเนื้องอก ประกอบด้วยข้อมูลดีเอ็นเอไมโครอาเรย์ของตัวอย่างทั้งหมด 36 ตัวอย่าง เป็นตัวอย่างของผู้ป่วยโรคมะเร็งในเนื้องอก 18 ตัวอย่าง และเป็นตัวอย่างปกติ 18 ตัวอย่าง ข้อมูลดีเอ็นเอไมโครอาเรย์ของแต่ละตัวอย่างประกอบด้วยยีนทั้งหมด 7,457 ยีน ข้อมูลมีลักษณะเป็นคู่ (Paired Data) ข้อมูลประกอบนอกจากค่าการแสดงออกของยีนแล้วจะมี หมายเลขยีน คำอธิบายยีน และข้อมูลเบื้องต้นของผู้ป่วย เช่น เพศ อายุ เป็นต้น ตัวอย่างข้อมูลทั้งหมดแสดงดังภาพประกอบ 5.1

ในส่วนของโรคมะเร็งในเนื้องอกนี้จะอธิบายแยกเป็น 2 ส่วนคือ ผลการทดลองของโรคมะเร็งในเนื้องอก และเปรียบเทียบผลการทดลองของโรคมะเร็งในเนื้องอก

Accession Number	Description	Sample	Tumor 27	Tumor 29	Tumor 34	Tumor 28	Tumor 35	Tumor 8	Tumor 3	Tumor 9	Tumor 4
		Age	68	67	81	84	74	NA	85	45	NA
		Gender	M	M	F	F	F	NA	F	F	NA
		Race	C	C	C	NA	NA	NA	C	NA	NA
		p53 Status	W	M	W	M	M	W	M	M	W
		% Tumor	50	100	40	30	75	25	10	25	80
		Duke Stage	A	B	B	B	B	B	B	B	C
D00003	"Human liver cytochrome P-450 mRNA, complete cds"	1	2	7	1	-3	0	10	15	-2	36
D00003	"Human liver cytochrome P-450 mRNA, complete cds"	2	-5	1	7	8	6	-6	7	-14	-1
D00003	"Human liver cytochrome P-450 mRNA, complete cds"	3	1	-4	-3	3	-3	-14	-8	-15	-17
D00015	"Human prion protein mRNA, human PrP 27-30 mRNA, complete cds"	4	26	13	13	9	2	10	20	17	21
D00102	"Human lymphotoxin (LT) mRNA, complete cds"	5	5	2	0	10	-5	3	-5	-2	-3
D00137	"Human class I alcohol dehydrogenase beta-1 subunit, allele 1 mRNA, complete cds"	6	18	13	11	28	6	63	26	136	26

ภาพประกอบ 5.1 แสดงตัวอย่างข้อมูลดีเอ็นเอไมโครอาเรย์ของโรค Carcinoma Tumor

5.1.1 ผลการทดลองของโรคมะเร็งในเนื้องอก

สำหรับการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โรคมะเร็งเนื้องอกได้ทำการทดลองทั้ง 4 แบบจำลองด้วยกัน แต่จะอธิบายโดยละเอียดเฉพาะแบบจำลองหลักคือแบบจำลอง A เท่านั้น

ขั้นตอนหลักที่ A1: ขั้นตอนการเตรียมข้อมูล กำหนดจำนวน Fold เท่ากับ 4 เนื่องจากข้อมูลมีจำนวนตัวอย่างน้อย จำนวน Fold ไม่ควรจะมากเกินไป ระบบจะแบ่งข้อมูลออกเป็นส่วนตามจำนวน Fold ที่กำหนดและแต่ละ Fold จะมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ ในขั้นตอนนี้จะใช้เวลาประมาณ 10.92 วินาทีเท่านั้น แต่ถ้ามีการเพิ่มจำนวน Fold เวลาก็อาจจะเพิ่มขึ้นตามไปด้วย

ขั้นตอนหลักที่ A2: ขั้นตอนการคำนวณค่าของ P-Value และการเลือกกลุ่มยีนที่มีค่า P-Value ต่ำ จากข้อมูลที่ได้เตรียมไว้ในขั้นตอนหลักที่ A1 โดยจะทำการคำนวณค่า P-Value ของชุดของมูลสอนทุกยีนในทุก Fold และส่วนของการเลือกยีนจะพิจารณาจากค่า Gene Selection Threshold (GST) ที่ผู้ใช้ป้อน โดยได้ทำการทดลอง 5 กรณีคือ กรณีค่า GST = 1% ถึง 5% ผลการทำงานในขั้นตอนหลักที่ A2 แสดงได้ดังตารางที่ 5.1

ตารางที่ 5.1 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A2 ของโรคมะเร็งในเนื้องอก

GST (%)	จำนวนยีนที่เลือก	เวลาที่ใช้ (วินาที)	
		คำนวณค่า P-Value	การเลือกยีน
1	75	11.27	0.28
2	150	11.27	0.29
3	224	11.27	0.25
4	298	11.27	0.52
5	373	11.27	0.63

จากตารางที่ 5.1 จะเห็นว่าการเพิ่มค่า GST ไม่มีผลต่อเวลาที่ใช้ในการคำนวณค่า P-Value เนื่องจากทุกขั้นตอนจะมีจำนวนครั้งในการคำนวณค่า P-Value ทุกยีนเท่ากัน ส่วนเวลาที่ใช้ในการเลือกยีนก็ไม่ค่อยมีผลกระทบที่เกี่ยวข้องกับค่า GST เช่นกัน เนื่องจากเวลาที่ใช้ในการเลือกยีนก็มีค่าไม่แตกต่างกันมากในแต่ละกรณี และถ้าเทียบกับเวลาที่ต้องใช้ในการทดลองทั้งหมดถือว่าเป็นเวลาเพียงเล็กน้อย

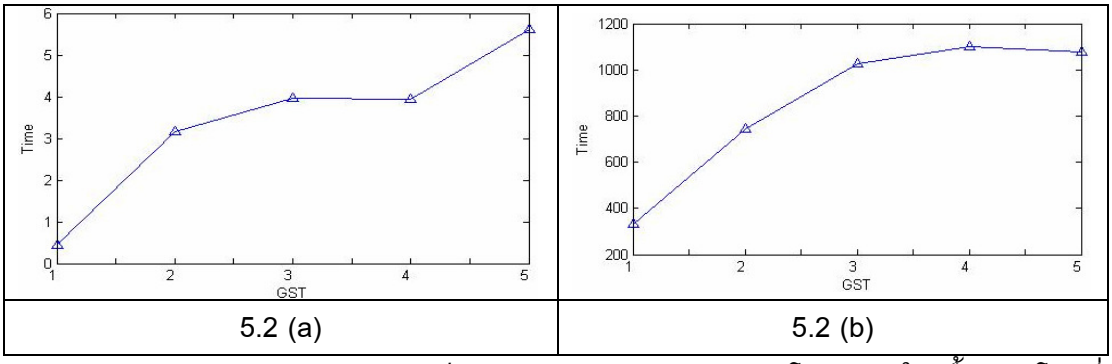
ขั้นตอนหลักที่ A3: ขั้นตอนการจัดกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง และการคำนวณค่า HCV ขั้นตอนนี้จะนำข้อมูลสอนของยีนที่เลือกมาจากขั้นตอนหลักที่ A2 มาทำการสร้างไฟล์ข้อมูลขึ้น โดยการสร้างไฟล์จะแยกแต่ละยีนออกมาเป็น 1 ไฟล์และรูปแบบของข้อมูลต้องถูกต้องตามข้อกำหนดของข้อมูลที่ใช้แผนที่การจัดกลุ่มเอง ซึ่งตัวอย่างข้อมูลแสดงได้ดังภาพประกอบ 3.2 ในบทที่ 3 เรียบร้อยแล้ว เวลาที่ใช้ในขั้นตอนหลักที่ A3 แสดงดังตารางที่ 5.2

ตารางที่ 5.2 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A3 ของโรคมะเร็งในเนื้องอก

GST (%)	เวลาที่ใช้ (วินาที)	
	ขั้นตอนการสร้างไฟล์	ขั้นตอนการจัดกลุ่มและคำนวณค่า HCV
1	0.45	330.06
2	3.16	742.09
3	3.97	1026.22
4	3.94	1100.47
5	5.61	1079

จากตารางที่ 5.2 จะเห็นว่าเวลาที่ใช้ในการสร้างไฟล์กับเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV แปรผันตรงกับค่า GST คือเวลาที่ใช้ในการสร้างไฟล์และเวลาที่ใช้ในการ

จัดกลุ่มข้อมูลมีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น และมีแนวโน้มจะลดลงเมื่อค่า GST ลดลงด้วย แสดงดังภาพประกอบ 5.2



5.2 (a) 5.2 (b)

ภาพประกอบ 5.2 แสดงความสัมพันธ์ระหว่างเวลากับค่า GST ของโรคมะเร็งในเนื้องอก โดยที่ภาพ 5.2 (a) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการสร้างไฟล์กับค่า GST และภาพ 5.2 (b) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV กับค่า GST

จากภาพประกอบ 5.2 จะเห็นว่าเวลาที่ใช้มีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น เนื่องจากเมื่อค่า GST เพิ่มมากขึ้นก็แสดงว่าจะต้องเลือกยีนเพิ่มมากขึ้นดังแสดงจำนวนยีนที่เลือกเพิ่มขึ้นในตารางที่ 5.1 แล้ว เมื่อมียีนที่ถูกเลือกเพิ่มมากขึ้น จำนวนยีนที่จะสร้างไฟล์และจำนวนยีนที่จะจัดกลุ่มก็มากขึ้น ทำให้เวลาที่ใช้สูงขึ้นด้วย

ขั้นตอนหลักที่ A4: ขั้นตอนการเลือกยีนโดยอาศัยค่า HCV โดยจะเลือกเฉพาะยีนที่มีค่า HCV ที่คำนวณได้มากกว่าหรือเท่ากับค่า HCV_T ที่กำหนด ในขั้นตอนนี้ได้มีการทดลองกำหนดค่า HCV เป็นค่าต่างๆ คือ 100 95 และ 90 ตามลำดับ ผลลัพธ์จากขั้นตอนหลักที่ A4 แสดงดังตารางที่ 5.3

ตารางที่ 5.3 แสดงเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ของโรคมะเร็งในเนื้องอก

GST (%)	เวลา (วินาที)		
	HCV = 100	HCV = 95	HCV = 90
1	0.14	0.13	0.11
2	0.11	0.12	0.11
3	0.11	0.13	0.13
4	0.11	0.14	0.13
5	0.11	0.11	0.14

จากตารางที่ 5.3 จะเห็นว่าเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ไม่ค่อยเปลี่ยนแปลง คือไม่ได้ขึ้นอยู่กับค่า GST หรือค่า HCV และเวลาที่ใช้ก็ยังถือว่าน้อยมากคือประมาณ 0.11 วินาทีเท่านั้น

ขั้นตอนหลักที่ A5: ขั้นตอนการสร้างกฎการเรียนรู้ในรูปแบบ “ถ้า–แล้ว” (IF–THEN Rule) ในขั้นตอนนี้จะเลือกยีนที่มีค่าการ Intersection มากกว่าหรือเท่ากับค่า Gene Intersection Threshold (GIT) ที่ผู้ใช้ต้องการ โดยที่ค่า GIT จะหาได้จากจำนวนครั้งที่ยีนนั้นถูกเลือกในแต่ละ Fold ตัวอย่างการคำนวณค่า GIT และการเลือกยีนได้แสดงไว้ในบทที่ 3 ในขั้นตอนหลักที่ 5 เรียบร้อยแล้ว ในการทดลองได้กำหนดค่า GIT เป็น 100% 70% และ 50% เวลาที่ใช้ขั้นตอนหลักที่ A5 แสดงได้ดังตารางที่ 5.4 ส่วนจำนวนยีนที่เลือกได้และนำมาสร้างกฎแสดงดังตารางที่ 5.4

ตารางที่ 5.4 แสดงเวลาที่ใช้ในการทำงานในขั้นตอนหลักที่ A5 ของโรคมะเร็งในเนื้องอก

GST (%)	เวลา (วินาที)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	0.58	0.59	0.55	0.84	0.64	0.55	0.56	0.75	0.64
2	0.52	0.58	1.47	0.86	2.25	3.63	5.06	1.72	5.34
3	0.67	0.66	1.19	1.42	2.66	5.69	3.58	6.38	4.95
4	0.88	1.11	2.1	2.78	5.11	5.12	4.02	9.02	9.61
5	1.38	1.73	0.69	1.53	0.92	2.84	1.77	6.64	5.98

ตารางที่ 5.5 แสดงจำนวนยีนที่เลือกได้ในขั้นตอนหลักที่ A5 ของโรคมะเร็งในเนื้องอก

GST (%)	เวลา (วินาที)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	5	8	17	12	19	35	15	25	48
2	7	14	28	20	38	69	33	64	111
3	10	18	37	30	55	93	56	95	155
4	11	24	47	37	67	115	69	121	189
5	15	24	40	44	70	128	79	131	201

จากตารางที่ 5.5 จะเห็นว่าจำนวนกฎมีแนวโน้มจะแปรผันตรงกับค่า GST และแปรผกผันกับค่า HCV และ GIT คือจำนวนกฎจะมากขึ้นในกรณีที่ค่า GST เพิ่มมากขึ้น และจะน้อยลงเมื่อค่า GST ลดลง แต่จำนวนกฎจะเพิ่มขึ้นเมื่อค่า HCV และค่า GIT ลดต่ำลง และจำนวนกฎจะลดต่ำลงเมื่อค่า HCV และ GIT เพิ่มมาก และจำนวนกฎที่น้อยที่สุดคือ 5 ซึ่งเกิดขึ้นในกรณีที่ค่า GST = 1% ค่า HCV = 100 และค่า GIT = 100%

ขั้นตอนหลักที่ A6: ขั้นตอนการนำกฎที่ได้ไปทดสอบกับชุดข้อมูลทดสอบที่สร้างไว้ในขั้นตอนหลักที่ A1 ในทุก Fold ผลลัพธ์ที่ได้จะเป็นค่าความถูกต้อง (Accuracy Values) ค่าความถูกต้องที่คำนวณได้จากการทำงานในขั้นตอนหลักที่ A6 แสดงดังตารางที่ 5.6

ตารางที่ 5.6 แสดงค่าความถูกต้องเฉลี่ยที่ได้จากขั้นตอนหลักที่ A6 ในแต่ละกรณี ของโรคมะเร็งในเนื้องอก

GST (%)	ค่าความถูกต้อง (Accuracy)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	97.25	95.36	91.2	93.07	90.54	87.04	91.46	89.64	86.04
2	96.53	92.9	88.95	92.5	89.56	85.57	90.64	86.88	83.71
3	95.63	92.71	88.95	91.98	88.84	85.4	89.42	86.47	82.74
4	95.45	92.05	89.1	91.13	88.81	85.22	88.55	86.27	82.28
5	95.46	93.38	89.43	90.97	87.77	84	88.8	85.57	82.15

จากตารางที่ 5.6 จะเห็นว่าค่าความถูกต้องจะแปรผันตรงกับค่า HCV และค่า GIT และจะแปรผกผันกับค่า GST คือค่าความถูกต้องจะเพิ่มขึ้นเมื่อค่า GST ลดลงในขณะที่ค่า HCV และค่า GIT เพิ่มสูงขึ้น และค่าความถูกต้องที่มากที่สุดคือ 97.25 % ที่กรณีค่า GST = 1% HCV = 100 และ GIT = 100%

จากขั้นตอนหลักทั้ง 6 ขั้นตอนสามารถสรุปผลการทดลองในทุกกรณีในประเด็นของ จำนวนกฎ ความถูกต้องที่ได้ และเวลารวมที่ใช้ในแต่ละกรณี แสดงดังตารางที่ 5.7

จากตารางที่ 5.7 จะเห็นว่ากรณีที่มีค่าความถูกต้องสูงที่สุดเท่ากับ 97.25% คือกรณีค่า GST = 1% HCV =100 และ GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 5 กฎ และเวลารวมที่ใช้คือ 5.87 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดเช่นกัน รายละเอียดของกรณีดังที่มีค่าความถูกต้องสูงสุดกล่าวแสดงได้ดังตารางที่ 5.8

ตารางที่ 5.7 ผลลัพธ์ของข้อมูลโรคมะเร็งในเนื้องอก

เงื่อนไข		HCV = 100			HCV = 95			HCV = 90		
		ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%
GST = 1%	จำนวนกฎ	**5**	8	17	12	19	35	15	25	48
	ค่าความถูกต้อง	**97.25**	95.36	91.2	93.07	90.54	87.04	91.46	89.64	86.04
	เวลารวม	**5.87**	5.87	5.87	5.87	5.87	5.87	5.87	5.87	5.87
GST = 2%	จำนวนกฎ	7	14	28	20	38	69	33	64	111
	ค่าความถูกต้อง	96.53	92.9	88.95	92.5	89.56	85.57	90.64	86.88	83.71
	เวลารวม	12.87	12.85	12.86	12.85	12.88	12.9	12.92	12.87	12.93
GST = 3%	จำนวนกฎ	10	18	37	30	55	93	56	95	155
	ค่าความถูกต้อง	95.63	92.71	88.95	91.98	88.84	85.4	89.42	86.47	82.74
	เวลารวม	17.58	17.58	17.59	17.6	17.6	17.67	17.63	17.68	17.66
GST = 4%	จำนวนกฎ	11	24	47	37	67	115	69	121	189
	ค่าความถูกต้อง	95.45	92.05	89.1	91.13	88.81	85.22	88.55	86.27	82.28
	เวลารวม	18.83	18.83	18.85	18.86	18.9	18.9	18.88	18.97	18.98
GST = 5%	จำนวนกฎ	15	24	40	44	70	128	79	131	201
	ค่าความถูกต้อง	95.46	93.38	89.43	90.97	87.77	84	88.8	85.57	82.15
	เวลารวม	16.73	16.72	16.72	16.73	16.73	16.76	16.8	16.82	16.82

ตารางที่ 5.8 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดจากแบบจำลอง A ของโรคมะเร็งในเนื้องอก

ค่า Gene Selection threshold (GST)	1
ค่า Highest Certainty Value (HCV)	100
ค่า Gene Intersection Threshold (GIT)	100
จำนวนกฎ	5
ค่าความถูกต้อง	97.25
เวลาที่ใช้ (นาที)	5.87

จากตารางที่ 5.8 จะเห็นว่าเป็นการแสดงเฉพาะกรณีที่มีค่าความถูกต้องสูงสุดเท่านั้น เพราะค่าความถูกต้องสูงสุดนี้จะถือว่าเป็นค่าความถูกต้องที่ได้จากแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ ซึ่งต่อไปจะใช้ในการเปรียบประสิทธิภาพกับแบบจำลอง B C และ D และเปรียบเทียบกับแบบจำลองอื่น ๆ ในบทที่ 6 ด้วย

ในส่วนของผลการทดลองที่ได้จากแบบจำลอง B C และ D จะแสดงเฉพาะผลการทดลองที่ดีที่สุดของแต่ละแบบจำลอง ซึ่งแสดงได้ดังตารางที่ 5.9 5.10 และ 5.11 ตามลำดับ

ตารางที่ 5.9 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง B ของโรคโรคมะเร็งในเนื้องอก

จำนวนกฎ	20
ค่าความถูกต้อง	93
เวลาที่ใช้ (นาที)	518.27

จะเห็นว่าแบบจำลอง B มีค่าความถูกต้องสูงสุดเท่ากับ 93% คือกรณีค่า HCV = 100 และค่า GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 20 กฎ และเวลารวมที่ใช้คือ 518.27 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

ตารางที่ 5.10 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง C ของโรคโรคมะเร็งในเนื้องอก

จำนวนกฎ	18
ค่าความถูกต้อง	88.84
เวลาที่ใช้ (นาที)	5.88

จะเห็นว่าแบบจำลอง C มีค่าความถูกต้องสูงสุดเท่ากับ 88.84% คือกรณีค่า GST = 1% และค่า GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 18 กฎ และเวลารวมที่ใช้คือ 5.88 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

ตารางที่ 5.11 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง D ของโรคโรคมะเร็งในเนื้องอก

จำนวนกฎ	7,457
ค่าความถูกต้อง	55.54
เวลาที่ใช้ (นาที)	554.45

จะเห็นว่าแบบจำลอง C จะมีค่าความถูกต้อง จำนวนกฎ และเวลาเพียงค่าเดียวเท่านั้นเนื่องจากไม่มีการเลือกยีนโดยอาศัยค่า P-Value และค่า HCV เลย ดังนั้นผลการทดลองที่ดีที่สุดคือ ค่าความถูกต้องเท่ากับ 55.54% จำนวนกฎก็จะเท่ากับจำนวนยีนที่มีคือ 7,457 กฎ ส่วนเวลาที่ใช้ก็คือเวลาที่ใช้ในการจัดกลุ่มข้อมูลยีนทั้งหมดคือ 554.45 นาที

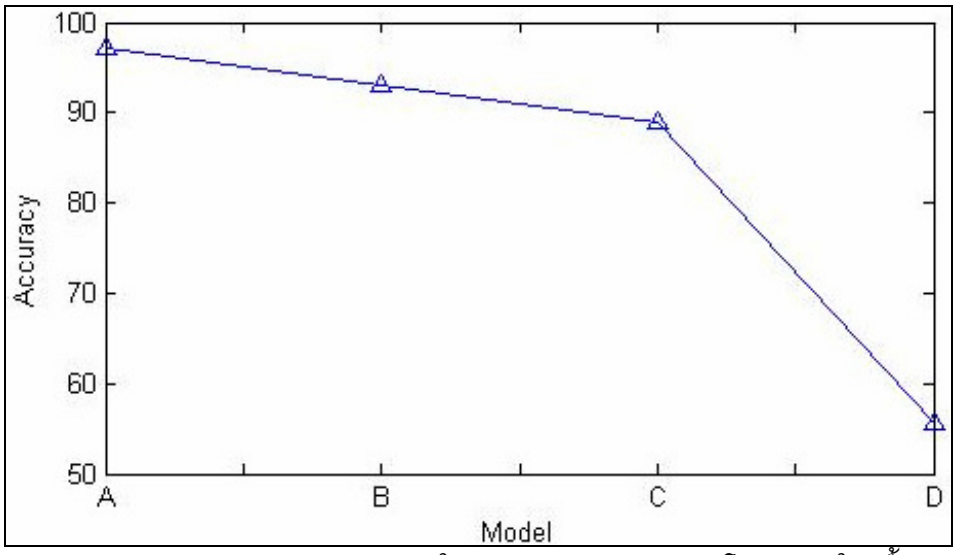
5.1.2 เปรียบเทียบผลการทดลองของโรคมะเร็งในเนื้องอก

จากผลการทดลองที่ผ่านมาทั้ง 4 แบบจำลอง สามารถเปรียบเทียบผลการทดลองในประเด็นของ ค่าความถูกต้อง (Accuracy) จำนวนกฎ เวลาที่ใช้ในการทดลองได้ดังตารางที่ 5.12

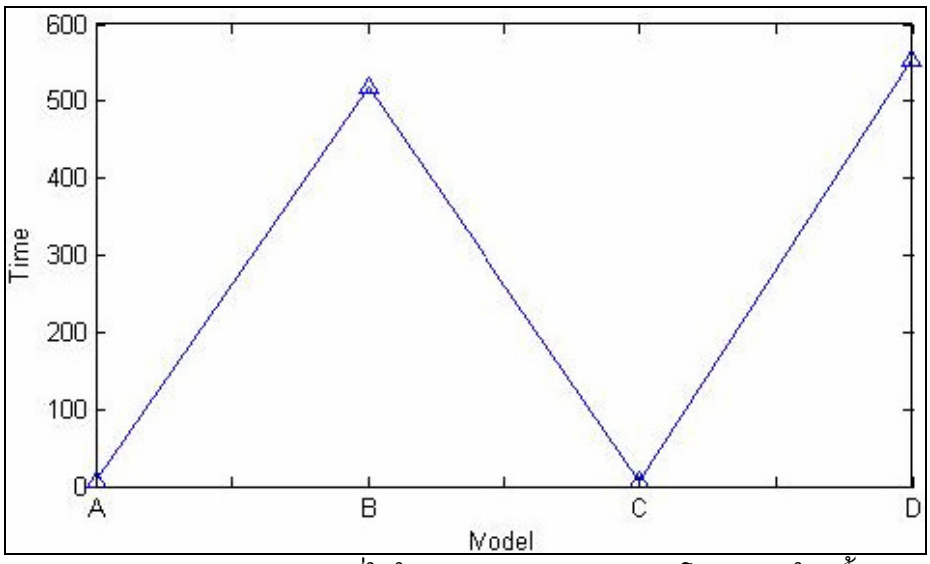
ตารางที่ 5.12 แสดงการเปรียบเทียบผลการทดลองของแบบจำลองทั้ง 4 ของโรคมะเร็งในเนื้องอก

แบบจำลอง	**A**	B	C	D
ค่าความถูกต้อง (Accuracy)	**97.25**	93	88.84	55.54
จำนวนกฎ	**5**	20	18	7457
เวลารวม (นาที)	**5.87**	518.27	5.88	554.45

จากตารางที่ 5.12 สามารถสรุปได้ว่าแบบจำลอง A มีประสิทธิภาพมากที่สุดในทุกประเด็น ดังนี้ ค่าความถูกต้องของแบบจำลอง A มีค่าสูงสุด จำนวนกฎที่ได้จากแบบจำลอง A มีค่าน้อยที่สุด และเวลาที่ใช้ในแบบจำลอง A ก็น้อยที่สุดด้วยเช่นกัน กราฟแสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 ในประเด็นเรื่องค่าความถูกต้องและเวลาที่ใช้แสดงดังภาพประกอบ 5.3 และ 5.4 ตามลำดับ



ภาพประกอบ 5.3 แสดงค่า Accuracy ในแต่ละแบบจำลอง ของโรคมะเร็งในเนื้องอก



ภาพประกอบ 5.4 แสดงเวลาที่ใช้ในแต่ละแบบจำลอง ของโรคมะเร็งในเนื้องอก

จากภาพประกอบ 5.3 และ 5.4 จะเห็นว่าแบบจำลอง A มีประสิทธิภาพมากที่สุด เนื่องจากมีค่าความถูกต้องสูงที่สุดและใช้เวลาน้อย แบบจำลอง B ให้ค่าความถูกต้องสูงแต่ใช้เวลามาก แบบจำลอง C ใช้เวลาน้อยแต่ให้ค่าความถูกต้องน้อยด้วย ส่วนแบบจำลอง D เป็นแบบจำลองที่มีประสิทธิภาพต่ำที่สุดเนื่องจากใช้เวลามากและได้ค่าความถูกต้องน้อย

5.2 โรคมะเร็งเม็ดเลือดขาว (Leukemia)

ข้อมูลของผู้ป่วยมะเร็งเม็ดเลือดขาว ประกอบด้วยข้อมูลดีเอ็นเอไมโครอาเรย์ของตัวอย่างทั้งหมด 72 ตัวอย่าง เป็นตัวอย่างของผู้ป่วยที่เป็นมะเร็งชนิด ALL (Actual Lymphoblastic Leukemia) 47 ตัวอย่าง และเป็นตัวอย่างของผู้ป่วยที่เป็นมะเร็งชนิด AML (Actual Myeloid Leukemia) 25 ตัวอย่าง ข้อมูลดีเอ็นเอไมโครอาเรย์ของแต่ละตัวอย่างประกอบด้วยยีนทั้งหมด 7,129 ยีน ข้อมูลมีลักษณะเป็นอิสระต่อกัน (Unpaired Data) ตัวอย่างข้อมูลทั้งหมดแสดงดังภาพประกอบ 5.5

19	87	77	-6	63	453	490	37	10
323	-142	192	-197	147	65	-9	154	309
420	212	235	262	285	291	220	224	553
231	5	95	-16	-10	11	121	149	84
246	325	189	23	296	143	91	169	226
533	80	243	168	413	251	129	269	679
-101	-518	-189	-180	-146	-75	-89	-142	-74
-451	-2406	-3710	-1460	-677	97	-262	-1223	-419
2112	1323	826	613	1475	1187	1661	751	1127
277	83	413	174	233	409	179	87	206
1023	529	399	277	643	617	497	753	700
67	-295	16	6	51	-34	31	-22	-61
214	352	558	81	450	738	241	259	381
-135	-67	24	2	-46	11	-20	47	-105
1074	67	893	722	612	742	441	806	1068
475	263	297	170	370	234	99	342	412
48	-33	6	0	29	72	-8	14	-43
168	-33	1971	510	333	109	80	239	702
-70	-21	-42	-73	-19	-30	-12	24	18
ALL	ALL	ALL	ALL	ALL	ALL	ALL	ALL	ALL

ภาพประกอบ 5.5 แสดงตัวอย่างข้อมูลดีเอ็นเอไมโครอาเรย์ของโรคมะเร็งเม็ดเลือดขาว

ในส่วนของโรคมะเร็งเม็ดเลือดขาวนี้จะอธิบายแยกเป็น 2 ส่วนคือ ผลการทดลองของโรคมะเร็งเม็ดเลือดขาว และเปรียบเทียบผลการทดลองของโรคมะเร็งเม็ดเลือดขาว

5.2.1 ผลการทดลองของโรคมะเร็งเม็ดเลือดขาว

สำหรับการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โรคมะเร็งเม็ดเลือดขาว ได้ทำการทดลองทั้ง 4 แบบจำลองด้วยกัน แต่จะอธิบายโดยละเอียดเฉพาะแบบจำลองหลักคือ แบบจำลอง A เท่านั้น

ขั้นตอนหลักที่ A1: ขั้นตอนการเตรียมข้อมูล กำหนดจำนวน Fold เท่ากับ 4 เนื่องจากข้อมูลมีจำนวนตัวอย่างน้อย จำนวน Fold ไม่ควรจะมากเกินไป ระบบจะแบ่งข้อมูล

ออกเป็นส่วนตามจำนวน Fold ที่กำหนดและแต่ละ Fold จะมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ ในขั้นตอนนี้จะใช้เวลาประมาณ 17.49 วินาทีเท่านั้น แต่ถ้ามีการเพิ่มจำนวน Fold เวลาก็อาจจะเพิ่มขึ้นตามไปด้วย

ขั้นตอนหลักที่ A2: ขั้นตอนการคำนวณค่าของ P-Value และการเลือกกลุ่มยีนที่มีค่า P-Value ต่ำ จากข้อมูลที่ได้เตรียมไว้ในขั้นตอนหลักที่ A1 โดยจะทำการคำนวณค่า P-Value ของชุดของมูลสอนทุกยีนในทุก Fold และส่วนของการเลือกยีนจะพิจารณาจากค่า Gene Selection Threshold (GST) ที่ผู้ใช้ป้อน โดยได้ทำการทดลอง 5 กรณีคือ กรณีค่า GST = 1% ถึง 5% ผลการทำงานในขั้นตอนหลักที่ A2 แสดงได้ดังตารางที่ 5.13

ตารางที่ 5.13 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A2 ของโรคมะเร็งเม็ดเลือดขาว

GST (%)	จำนวนยีนที่เลือก	เวลาที่ใช้ (วินาที)	
		คำนวณค่า P-Value	การเลือกยีน
1	77	14.95	0.63
2	143	14.95	0.41
3	224	14.95	0.44
4	214	14.95	0.41
5	356	14.95	0.5

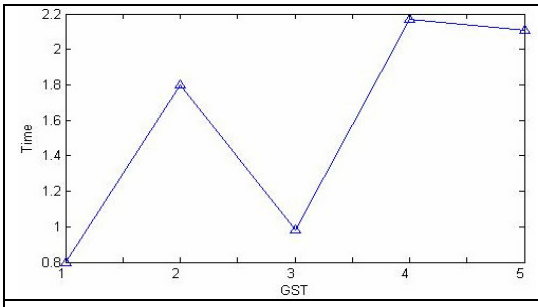
จากตารางที่ 5.13 จะเห็นว่าการเพิ่มค่า GST ไม่มีผลต่อเวลาที่ใช้ในการคำนวณค่า P-Value เนื่องจากทุกขั้นตอนจะมีจำนวนครั้งในการคำนวณค่า P-Value ทุกยีนเท่ากัน ส่วนเวลาที่ใช้ในการเลือกยีนก็ไม่ค่อยมีผลกระทบที่เกี่ยวข้องกับค่า GST เช่นกัน เนื่องจากเวลาที่ใช้ในการเลือกยีนก็มีค่าใกล้เคียงกันในแต่ละกรณี และถ้าเทียบกับเวลาที่ต้องใช้ในการทดลองทั้งหมดถือว่าเป็นเวลาเพียงเล็กน้อย

ขั้นตอนหลักที่ A3: ขั้นตอนการจัดกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเองและการคำนวณค่า HCV ขั้นตอนนี้จะนำข้อมูลสอนของยีนที่เลือกมาจากขั้นตอนหลักที่ A2 มาทำการสร้างไฟล์ข้อมูลขึ้น โดยการสร้างไฟล์จะแยกแต่ละยีนออกมาเป็น 1 ไฟล์และรูปแบบของข้อมูลต้องถูกต้องตามข้อกำหนดของข้อมูลที่ใช้แผนที่การจัดกลุ่มเอง ซึ่งตัวอย่างข้อมูลแสดงได้ดังภาพประกอบ 3.2 ในบทที่ 3 เรียบร้อยแล้ว เวลาที่ใช้ในขั้นตอนหลักที่ A3 แสดงดังตารางที่ 5.14

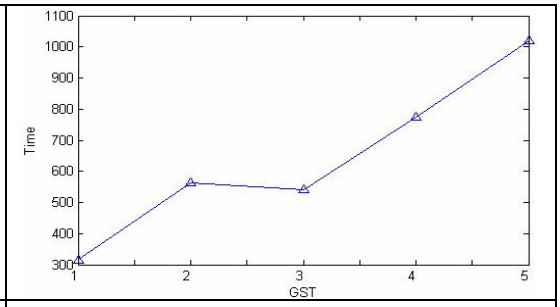
ตารางที่ 5.14 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A3 ของโรคมะเร็งเม็ดเลือดขาว

GST (%)	เวลาที่ใช้ (วินาที)	
	ขั้นตอนการสร้างไฟล์	ขั้นตอนการจัดกลุ่มและคำนวณค่า HCV
1	0.8	314.77
2	1.8	563.14
3	0.98	539.8
4	2.17	774.81
5	2.11	1018.31

จากตารางที่ 5.14 จะเห็นว่าเวลาที่ใช้ในการสร้างไฟล์กับเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV จะแปรผันตรงกับค่า GST คือเวลาที่ใช้ในการสร้างไฟล์และเวลาที่ใช้ในการจัดกลุ่มข้อมูลมีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น และมีแนวโน้มจะลดลงเมื่อค่า GST ลดลงด้วย แสดงดังภาพประกอบ 5.6



5.6 (a)



5.6 (b)

ภาพประกอบ 5.6 แสดงความสัมพันธ์ระหว่างเวลากับค่า GST ของโรคมะเร็งเม็ดเลือดขาวโดยที่ภาพ 5.6 (a) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการสร้างไฟล์กับค่า GST และภาพ 5.6 (b) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV กับค่า GST

จากภาพประกอบ 5.6 จะเห็นว่าเวลาที่ใช้มีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น เนื่องจากเมื่อค่า GST เพิ่มมากขึ้นก็แสดงว่าจะต้องเลือกยีนเพิ่มมากขึ้นดังแสดงจำนวนยีนที่เลือกเพิ่มขึ้นในตารางที่ 5.13 แล้ว เมื่อมียีนที่ถูกเลือกเพิ่มมากขึ้น จำนวนยีนที่จะสร้างไฟล์และจำนวนยีนที่จะจัดกลุ่มก็มากขึ้น ทำให้เวลาที่ใช้สูงขึ้นด้วย

ขั้นตอนหลักที่ A4: ขั้นตอนการเลือกยีนโดยอาศัยค่า HCV โดยจะเลือกเฉพาะยีนที่มีค่า HCV ที่คำนวณได้มากกว่าหรือเท่ากับค่า HCV_T ที่กำหนด ในขั้นตอนนี้ได้มี

การทดลองกำหนดค่า HCV_T เป็นค่าต่างๆ คือ 100 95 และ 90 ตามลำดับ ผลลัพธ์จากขั้นตอนหลักที่ A4 แสดงดังตารางที่ 5.15

ตารางที่ 5.15 แสดงเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ของโรคมะเร็งเม็ดเลือดขาว

GST	เวลา (วินาที)		
	HCV = 100	HCV = 95	HCV = 90
1	0.14	0.15	0.14
2	0.11	0.11	0.13
3	0.11	0.11	0.14
4	0.13	0.11	0.14
5	0.13	0.11	0.14

จากตารางที่ 5.15 จะเห็นว่าเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ไม่ค่อยเปลี่ยนแปลง คือไม่ได้ขึ้นอยู่กับค่า GST หรือค่า HCV และเวลาที่ใช้ก็ยังถือว่าน้อยมากคือประมาณ 0.13 วินาทีเท่านั้น

ขั้นตอนหลักที่ A5: ขั้นตอนการสร้างกฎการเรียนรู้ในรูปแบบ "ถ้า–แล้ว" (IF–THEN Rule) ในขั้นตอนนี้จะเลือกยีนที่มีค่าการ Intersection มากกว่าหรือเท่ากับค่า Gene Intersection Threshold (GIT) ที่ผู้ใช้ต้องการ โดยที่ค่า GIT จะหาได้จากจำนวนครั้งที่ยีนนั้นถูกเลือกในแต่ละ Fold ตัวอย่างการคำนวณค่า GIT และการเลือกยีนได้แสดงไว้ในบทที่ 3 ในขั้นตอนหลักที่ 5 เรียบร้อยแล้ว ในการทดลองได้กำหนดค่า GIT เป็น 100% 70% และ 50% เวลาที่ใช้ขั้นตอนหลักที่ A5 แสดงได้ดังตารางที่ 5.16 ส่วนจำนวนยีนที่เลือกได้และนำมาสร้างกฎ แสดงดังตารางที่ 5.17

ตารางที่ 5.16 แสดงเวลาที่ใช้ในการทำงานในขั้นตอนหลักที่ A5 ของโรคมะเร็งเม็ดเลือดขาว

GST	เวลา (วินาที)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	0.66	0.64	0.44	0.5	0.69	0.67	0.72	0.8	0.92
2	0.53	0.5	0.5	1.3	1.28	1.5	1.59	3.92	2.92
3	0.55	0.63	0.55	1.09	0.72	1.77	1.09	4.09	5.72
4	0.48	0.64	0.45	1.06	1.89	1.78	3.08	3.8	6.53
5	0.45	0.52	0.7	1.06	0.61	1.31	1.73	6.89	7.78

ตารางที่ 5.17 แสดงจำนวนยีนที่เลือกได้ในขั้นตอนหลักที่ A5 ของโรคมะเร็งเม็ดเลือดขาว

GST	เวลา (วินาที)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	1	3	4	7	14	24	22	33	50
2	2	4	6	9	17	30	31	57	86
3	2	4	6	10	18	34	34	66	121
4	2	4	6	10	21	37	42	80	145
5	2	4	6	10	22	42	45	88	166

จากตารางที่ 5.17 จะเห็นว่าจำนวนกฎมีแนวโน้มจะแปรผันตรงกับค่า GST และแปรผกผันกับค่า HCV และค่า GIT คือจำนวนกฎจะมากขึ้นในกรณีที่ค่า GST เพิ่มมากขึ้น จะน้อยลงเมื่อค่า GST ลดลง แต่จำนวนกฎจะเพิ่มขึ้นเมื่อค่า HCV และค่า GIT ลดต่ำลง และจำนวนกฎจะลดต่ำลงเมื่อค่า HCV และ GIT เพิ่มมากขึ้น ซึ่งจำนวนกฎที่น้อยที่สุดคือ 1 กฎ เกิดขึ้นในกรณีที่ค่า GST = 1% ค่า HCV = 100 และค่า GIT = 100%

ขั้นตอนหลักที่ A6: ขั้นตอนการนำกฎที่ได้ไปทดสอบกับชุดข้อมูลทดสอบที่สร้างไว้ในขั้นตอนหลักที่ A1 ในทุก Fold ผลลัพธ์ที่ได้จะเป็นค่าความถูกต้อง (Accuracy Values) ค่าความถูกต้องที่คำนวณได้จากการทำงานในขั้นตอนหลักที่ A6 แสดงดังตารางที่ 5.18

ตารางที่ 5.18 แสดงค่าความถูกต้องเฉลี่ยที่ได้จากขั้นตอนหลักที่ A6 ในแต่ละกรณี ของโรคมะเร็งเม็ดเลือดขาว

GST	ค่าความถูกต้อง (Accuracy)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	93.27	92.33	90.65	90	88.03	84.86	85.7	83.98	81.96
2	93.2	91.11	88.82	90.37	87.87	83.64	85.75	82.71	79.83
3	93.2	91.11	88.82	90.37	87.59	82.97	85.42	81.88	78.13
4	93.2	91.11	88.82	90.37	86.66	82.77	84.27	80.79	77.43
5	93.2	91.11	88.82	90.37	86.34	81.99	83.77	80.35	77.08

จากตารางที่ 5.18 จะเห็นว่าค่าความถูกต้องจะแปรผันตรงกับค่า HCV และค่า GIT และจะแปรผกผันกับค่า GST คือค่าความถูกต้องจะเพิ่มขึ้นเมื่อค่า GST ลดลงในขณะที่ค่า HCV และค่า GIT เพิ่มสูงขึ้น และค่าความถูกต้องที่มากที่สุดคือ 93.27% ที่กรณีค่า GST = 1% HCV = 100 และ GIT = 100%

จากขั้นตอนหลักทั้ง 6 ขั้นตอนสามารถสรุปผลการทดลองในทุกกรณี ในประเด็นของ จำนวนกฎ ความถูกต้องที่ได้ และเวลารวมที่ใช้ในแต่ละกรณี แสดงดังตารางที่ 5.20 และจากตารางที่ 5.20 จะเห็นว่ากรณีที่มีค่าความถูกต้องสูงที่สุดเท่ากับ 93.27% คือกรณีค่า GST = 1% HCV =100 GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 1 และเวลารวมที่ใช้คือ 5.83 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดเช่นกัน รายละเอียดของกรณีดังที่มีค่าความถูกต้องสูงสุดกล่าวแสดงได้ดังตารางที่ 5.19

ตารางที่ 5.19 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดจากแบบจำลอง A ของโรคมะเร็งเม็ดเลือดขาว

ค่า Gene Selection threshold (GST)	1
ค่า Highest Certainty Value (HCV)	100
ค่า Gene Intersection Threshold (GIT)	100
จำนวนกฎ	1
ค่าความถูกต้อง	93.27
เวลาที่ใช้ (นาที)	5.83

จากตารางที่ 5.19 จะเห็นว่าเป็นการแสดงเฉพาะกรณีที่มีค่าความถูกต้องสูงสุดเท่านั้น เพราะค่าความถูกต้องสูงสุดนี้จะถือว่าเป็นค่าความถูกต้องที่ได้จากแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ ซึ่งต่อไปจะใช้ในการเปรียบเทียบประสิทธิภาพกับแบบจำลอง B C และ D นอกจากนั้นยังได้ทำการเปรียบเทียบกับแบบจำลองอื่นๆ ด้วย

ในส่วนของผลการทดลองที่ได้จากแบบจำลอง B C และ D จะแสดงเฉพาะผลการทดลองที่ดีที่สุดของแต่ละแบบจำลอง ซึ่งแสดงได้ดังตารางที่ 5.21 5.22 และ 5.23 ตามลำดับ

ตารางที่ 5.20 ผลลัพธ์ของข้อมูลโรคมะเร็งเม็ดเลือดขาว

เงื่อนไข		HCV = 100			HCV = 95			HCV = 90		
		ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%
GST = 1%	จำนวนกฎ	**1**	3	4	7	14	24	22	33	50
	ค่าความถูกต้อง	**93.27**	92.33	90.65	90	88.03	84.86	85.7	83.98	81.96
	เวลารวม	**5.83**	5.83	5.83	5.83	5.83	5.83	5.83	5.84	5.84
GST = 2%	จำนวนกฎ	2	4	6	9	17	30	31	57	86
	ค่าความถูกต้อง	93.2	91.11	88.82	90.37	87.87	83.64	85.75	82.71	79.83
	เวลารวม	9.98	9.98	9.98	10	10	10	10	10.04	10.03
GST = 3%	จำนวนกฎ	2	4	6	10	18	34	34	66	121
	ค่าความถูกต้อง	93.2	91.11	88.82	90.37	87.59	82.97	85.42	81.88	78.13
	เวลารวม	9.58	9.58	9.58	9.58	9.58	9.6	9.59	9.64	9.67
GST = 4%	จำนวนกฎ	2	4	6	10	21	37	42	80	145
	ค่าความถูกต้อง	93.2	91.11	88.82	90.37	86.66	82.77	84.27	80.79	77.43
	เวลารวม	13.52	13.52	1352	13.53	13.54	13.54	13.56	13.57	13.62
GST = 5%	จำนวนกฎ	2	4	6	10	22	42	45	88	166
	ค่าความถูกต้อง	93.2	91.11	88.82	90.37	86.34	81.99	83.77	80.35	77.08
	เวลารวม	17.57	17.58	17.85	17.58	17.58	17.59	17.6	17.69	11.7

ตารางที่ 5.21 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง B ของโรคมะเร็งเม็ดเลือดขาว

จำนวนกฎ	2
ค่าความถูกต้อง	93.2
เวลาที่ใช้ (นาที)	182.94

จากตารางที่ 5.21 จะเห็นว่าแบบจำลอง B มีค่าความถูกต้องสูงสุดเท่ากับ 93.2% คือกรณีค่า HCV = 100 และค่า GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 2 กฎ และเวลารวมที่ใช้คือ 462.14 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

ตารางที่ 5.22 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง C ของโรคมะเร็งเม็ดเลือดขาว

จำนวนกฎ	31
ค่าความถูกต้อง	84.16
เวลาที่ใช้ (นาที)	2.3

จากตารางที่ 5.22 จะเห็นว่าแบบจำลอง C มีค่าความถูกต้องสูงสุดเท่ากับ 84.16% คือกรณีค่า GST = 1% และค่า GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 31 กฎ และเวลารวมที่ใช้คือ 2.3 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

ตารางที่ 5.23 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง D ของโรคมะเร็งเม็ดเลือดขาว

จำนวนกฎ	7,129
ค่าความถูกต้อง	64
เวลาที่ใช้ (นาที)	196.81

จากตารางที่ 5.23 จะเห็นว่าแบบจำลอง D จะมีค่าความถูกต้อง จำนวนกฎ และเวลาเพียงค่าเดียวเท่านั้นเนื่องจากไม่มีการเลือกยีนโดยอาศัยค่า P-Value และค่า HCV เลย

ดังนั้นผลการทดลองที่ดีที่สุดคือ ค่าความถูกต้องเท่ากับ 64% จำนวนกฎก็จะเท่ากับจำนวนยีนที่มีคือ 7,129 กฎ ส่วนเวลาที่ใช้ก็คือเวลาที่ใช้ในการจัดกลุ่มข้อมูลยีนทั้งหมดคือ 196.81 นาที

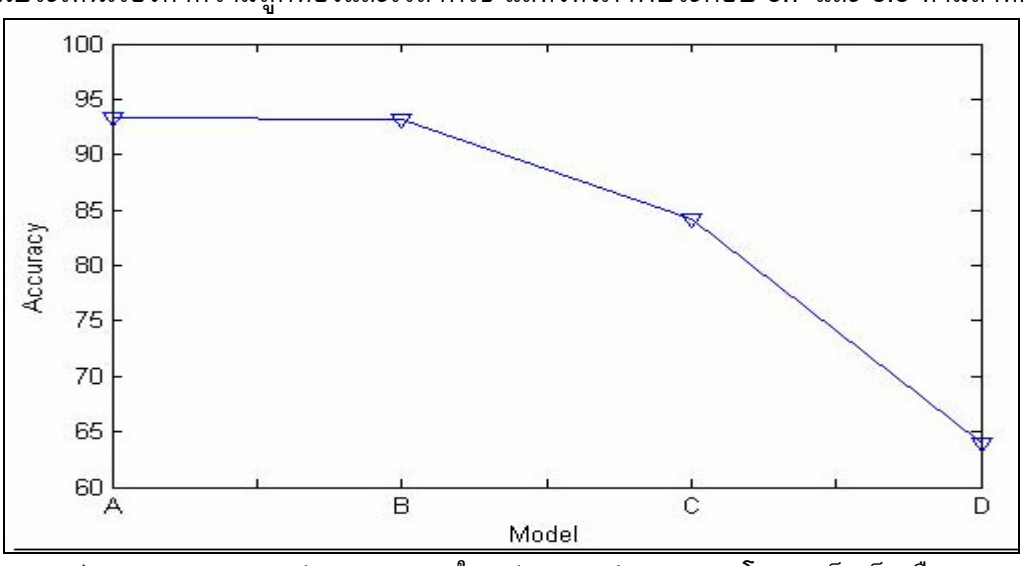
5.2.2 เปรียบเทียบผลการทดลองของโรคมะเร็งเม็ดเลือดขาว

จากผลการทดลองที่ผ่านมาทั้ง 4 แบบจำลอง สามารถเปรียบเทียบผลการทดลองในประเด็นของ ค่าความถูกต้อง (Accuracy) จำนวนกฎ เวลาที่ใช้ในการทดลองได้ดังตารางที่ 5.24

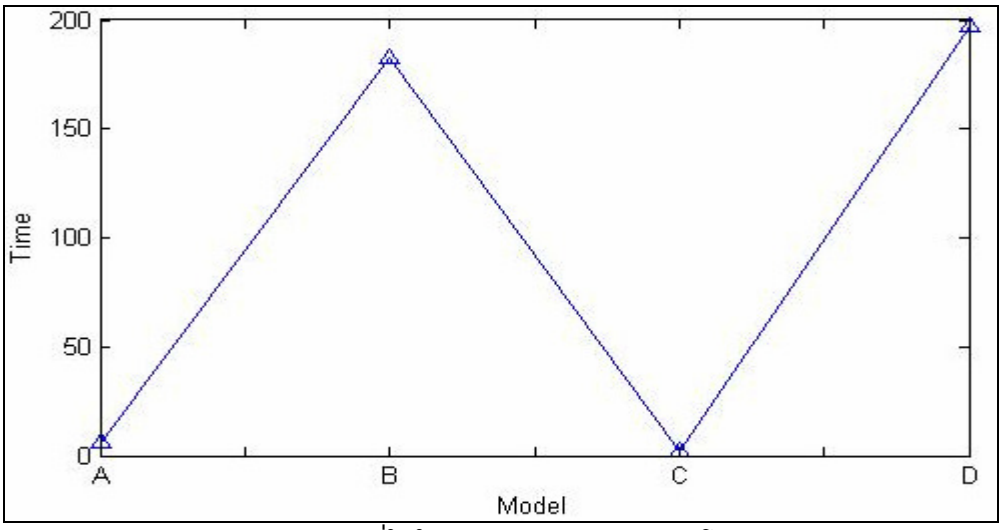
ตารางที่ 5.24 แสดงการเปรียบเทียบผลการทดลองของแบบจำลองทั้ง 4 ของโรคมะเร็งเม็ดเลือดขาว

แบบจำลอง	**A**	B	C	D
ค่าความถูกต้อง (Accuracy)	**93.27**	93.2	84.16	64
จำนวนกฎ	**1**	2	31	7129
เวลารวม (นาที)	5.83	182.94	**2.3**	196.81

จากตารางที่ 5.24 จะสามารถสรุปได้ว่าแบบจำลอง A มีประสิทธิภาพมากที่สุดในทุกประเด็น ดังนี้ ค่าความถูกต้องของแบบจำลอง A มีค่าสูงสุด จำนวนกฎที่ได้จากแบบจำลอง A มีค่าน้อยที่สุด และเวลาที่ใช้ในแบบจำลอง A ก็น้อย ถึงจะมากกว่าแบบจำลอง C แต่ถ้าเทียบกับเวลาที่มากกว่าเพียง 3.53 นาที แต่ค่าความถูกต้องเพิ่มขึ้นจาก 84.16 เป็น 93.27 ก็นับว่าแบบจำลอง A ดีกว่า กราฟแสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 ในประเด็นเรื่องค่าความถูกต้องและเวลาที่ใช้ แสดงดังภาพประกอบ 5.7 และ 5.8 ตามลำดับ



ภาพประกอบ 5.7 แสดงค่า Accuracy ในแต่ละแบบจำลอง ของโรคมะเร็งเม็ดเลือดขาว



ภาพประกอบ 5.8 แสดงเวลาที่ใช้ในแต่ละแบบจำลอง ของโรคมะเร็งเม็ดเลือดขาว

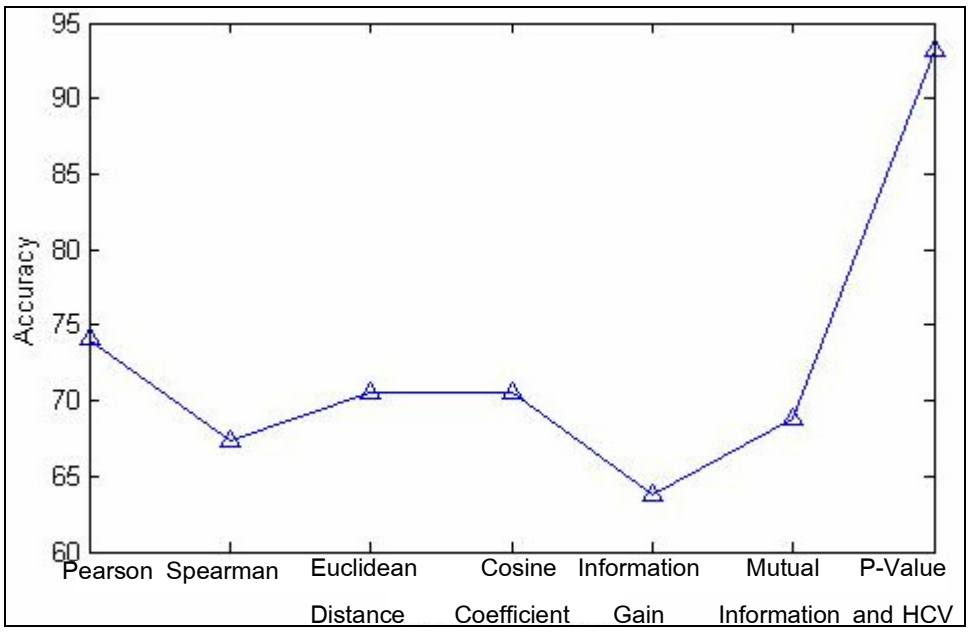
จากภาพประกอบ 5.7 และ 5.8 จะเห็นว่าแบบจำลอง A มีประสิทธิภาพสูงที่สุด เนื่องจากมีค่า Accuracy สูง และใช้เวลาน้อย ส่วนแบบจำลอง B ให้ค่าความถูกต้องสูงแต่ใช้เวลามาก แบบจำลอง C ใช้เวลาน้อยแต่มีค่าความถูกต้องต่ำ ส่วนแบบจำลอง D เป็นแบบจำลองที่แย่ที่สุดเนื่องจากใช้เวลามากและมีค่าความถูกต้องต่ำด้วย

จากนั้นเมื่อได้สรุปแล้วว่าแบบจำลอง A เป็นแบบจำลองที่ดีที่สุด จึงได้นำผลลัพธ์ที่ได้จากแบบจำลอง A มาทำการเปรียบเทียบกับงานวิจัยอื่นๆ ที่มีวัตถุประสงค์เช่นเดียวกัน และทำการทดลองบนชุดข้อมูลเดียวกัน โดยเปรียบเทียบกับงานวิจัยที่ใช้ SOM ในการจัดกลุ่มเช่นเดียวกัน แต่ในขั้นตอน Feature Extraction ใช้วิธีการต่างกัน (Cho *et al.*, 2002) ได้ผลลัพธ์ดังตารางที่ 5.25

ตารางที่ 5.25 เปรียบเทียบผลการทดลองที่ใช้ SOM ในการแบ่งกลุ่มของโรคมะเร็งเม็ดเลือดขาว

Feature Extraction	Pearson	Spearman	Euclidean Distance	Cosine Coefficient	Information Gain	Mutual Information	P-Value and HCV
Accuracy	74.1	67.4	70.6	70.6	63.8	68.8	**93.27**

จากตารางที่ 5.25 จะเห็นว่าการจัดกลุ่มข้อมูลด้วย SOM โดยใช้ค่า P-Value ในการทำ Feature Extraction มีค่าความถูกต้องสูงที่สุดคือ 93.27% โดยค่าถูกต้องของการทำ Feature Extraction แต่ละวิธีสามารถแสดงการเปรียบเทียบได้ดังภาพประกอบ 5.9



ภาพประกอบ 5.9 แสดงการเปรียบเทียบค่าความถูกต้องของการจัดกลุ่มข้อมูลด้วย SOM ในการใช้วิธีการทำ Feature Extraction แบบต่างๆ ของโรคมะเร็งเม็ดเลือดขาว

จากภาพประกอบ 5.9 จะเห็นได้ชัดว่าในการใช้ SOM จัดกลุ่มข้อมูลนั้น กรณีที่ทำ Feature Extraction โดยอาศัยค่า P-Value และค่า HCV ดังที่ได้ออกแบบไว้ จะให้ค่าความถูกต้องสูงที่สุด

5.3 โรคมะเร็งลำไส้ (Colon Cancer)

ข้อมูลของผู้ป่วยโรคมะเร็งลำไส้ ประกอบด้วยข้อมูลดีเอ็นเอไมโครอาเรย์ของตัวอย่างทั้งหมด 62 ตัวอย่าง เป็นตัวอย่างของผู้ป่วยโรคมะเร็งลำไส้ 40 ตัวอย่าง และเป็นตัวอย่างของผู้ป่วยปกติ 22 ตัวอย่าง ข้อมูลดีเอ็นเอไมโครอาเรย์ของแต่ละตัวอย่างประกอบด้วยยีนทั้งหมด 2,000 ยีน ข้อมูลมีลักษณะเป็นอิสระต่อกัน (Unpaired Data) ตัวอย่างข้อมูลทั้งหมดแสดงดังภาพประกอบ 5.10

34.47375	55.05625	7.43500	7.19875	80.82375	65.61500	37.41750	21.02875	28.01250	56.23500
128.88375	52.44500	51.23000	112.80875	85.00875	14.94625	86.38625	70.51875	23.72375	14.12000
207.98500	90.09125	212.51500	93.94125	462.37500	239.21625	163.13875	203.06375	87.18000	135.63875
49.72895	52.05000	28.32763	30.63947	44.38684	86.72500	146.12500	174.06316	6.08816	45.43816
31.73750	28.06500	52.61750	19.63625	49.26250	43.85125	89.59000	86.48125	26.13000	60.20625
31.60278	36.40278	63.61528	119.24722	135.05000	155.23194	124.70833	76.85139	18.76528	6.07222
42.96000	34.74500	8.17625	17.48875	54.33875	41.93250	35.79375	38.58875	50.26000	46.43875
27.11250	34.83250	21.21625	16.70625	17.65625	6.06875	62.05875	10.61125	24.35000	54.94750
73.71974	61.07368	36.74605	34.43553	57.58421	111.85132	39.71974	35.98553	15.41974	12.18421
58.13750	35.96786	26.42143	12.59286	19.12321	83.78214	83.14286	43.93571	13.55357	38.39107
284.66000	256.13625	83.34875	206.95000	130.37250	151.44750	321.33750	239.27000	84.20625	60.52750
75.21000	39.20500	22.17375	25.29375	28.60000	48.47750	41.59125	29.55875	47.47875	14.93000
333.74000	47.93625	45.71250	145.59875	192.66875	105.97375	263.36875	367.58125	73.50500	96.66000
109.18690	152.85000	270.93571	70.66786	129.40476	190.71190	126.78929	152.59524	122.53333	116.19405
32.20375	88.99625	28.98000	6.44750	35.87875	85.25875	92.23875	41.68375	47.53500	46.67375
601.84750	203.76000	530.16875	156.44500	342.65000	206.53625	96.27625	5.92500	147.23625	172.78875
117.00375	114.48875	264.92000	93.95000	84.75125	80.44875	150.59000	183.00625	28.31000	51.82375
62.32000	113.52250	79.34125	86.90875	60.27375	104.80500	82.23750	74.52875	33.19500	97.85500
167.30833	182.23333	640.05833	95.44286	143.87738	180.94167	85.03333	67.71071	91.85000	98.98214
13.26750	28.88375	7.02625	31.18750	8.99625	68.93375	152.19500	48.33875	5.87875	24.19625
52.09250	79.89750	94.53500	57.15250	106.87875	118.20125	186.56750	42.52000	36.29750	29.76625
47.71750	29.32875	56.83500	13.96000	23.20250	32.68750	44.47250	49.98250	9.81500	44.37625
30.42000	41.43625	31.35000	10.54750	32.16625	23.26500	16.77375	16.08500	21.88375	52.29000
negative	negative	negative	negative	negative	negative	**positive**	**positive**	**positive**	**positive**

ภาพประกอบ 5.10 แสดงตัวอย่างข้อมูลดีเอ็นเอไมโครอาเรย์ของโรคมะเร็งลำไส้

ในส่วนของโรคมะเร็งลำไส้นี้จะอธิบายแยกเป็น 2 ส่วนคือ ผลการทดลองของโรคมะเร็งลำไส้ และเปรียบเทียบผลการทดลองของโรคมะเร็งลำไส้

5.3.1 ผลการทดลองของโรคมะเร็งลำไส้

สำหรับการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โรคมะเร็งลำไส้ได้ทำการทดลองทั้ง 4 แบบจำลองด้วยกัน แต่จะอธิบายโดยละเอียดเฉพาะแบบจำลองหลักคือแบบจำลอง A เท่านั้น

ขั้นตอนหลักที่ A1: ขั้นตอนการเตรียมข้อมูล กำหนดจำนวน Fold เท่ากับ 4 เนื่องจากข้อมูลมีจำนวนตัวอย่างน้อย จำนวน Fold ไม่ควรจะมากเกินไป ระบบจะแบ่งข้อมูลออกเป็นส่วนตามจำนวน Fold ที่กำหนดและแต่ละ Fold จะมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ ในขั้นตอนนี้จะใช้เวลาประมาณ 14.11 วินาทีเท่านั้น แต่ถ้ามีการเพิ่มจำนวน Fold เวลาก็อาจจะเพิ่มขึ้นตามไปด้วย

ขั้นตอนหลักที่ A2: ขั้นตอนการคำนวณค่าของ P-Value และการเลือกกลุ่มยีนที่มีค่า P-Value ต่ำ จากข้อมูลที่ได้เตรียมไว้ในขั้นตอนหลักที่ A1 โดยจะทำการคำนวณค่า P-Value ของชุดของมูลสอนทุกยีนในทุก Fold และส่วนของการเลือกยีนจะพิจารณาจากค่า Gene Selection Threshold (GST) ที่ผู้ใช้ป้อน โดยได้ทำการทดลอง 5 กรณีคือ กรณีค่า GST = 1% ถึง 5% ผลการทำงานในขั้นตอนหลักที่ A2 แสดงได้ดังตารางที่ 5.26

ตารางที่ 5.26 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A2 ของโรคมะเร็งลำไส้

GST (%)	จำนวนยีนที่เลือก	เวลาที่ใช้ (วินาที)	
		คำนวณค่า P-Value	การเลือกยีน
1	20	4.45	0.11
2	40	4.45	0.11
3	60	4.45	0.11
4	80	4.45	0.11
5	100	4.45	0.09

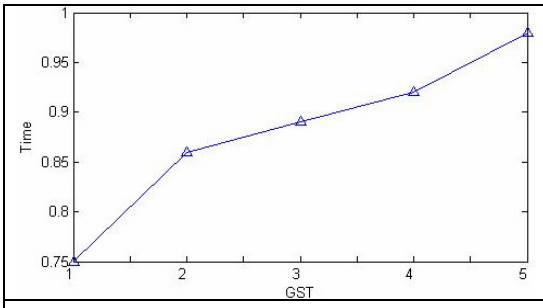
จากตารางที่ 5.26 จะเห็นว่าการเพิ่มค่า GST ไม่มีผลต่อเวลาที่ใช้ในการคำนวณค่า P-Value เนื่องจากทุกขั้นตอนจะมีจำนวนครั้งในการคำนวณค่า P-Value ทุกยีนเท่ากัน ส่วนเวลาที่ใช้ในการเลือกยีนก็ไม่ค่อยมีผลกระทบที่เกี่ยวข้องกับค่า GST เช่นกัน เนื่องจากเวลาที่ใช้ในการเลือกยีนก็มีค่าใกล้เคียงกันในแต่ละกรณี และถ้าเทียบกับเวลาที่ต้องใช้ในการทดลองทั้งหมดถือว่าเป็นเวลาเพียงเล็กน้อย

ขั้นตอนหลักที่ A3: ขั้นตอนการจัดกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง และการคำนวณค่า HCV ขั้นตอนนี้จะนำข้อมูลสอนของยีนที่เลือกมาจากขั้นตอนหลักที่ A2 มาทำการสร้างไฟล์ข้อมูลขึ้น โดยการสร้างไฟล์จะแยกแต่ละยีนออกมาเป็น 1 ไฟล์และรูปแบบของข้อมูลต้องถูกต้องตามข้อกำหนดของข้อมูลที่ใช้แผนที่การจัดกลุ่มเอง ซึ่งตัวอย่างข้อมูลแสดงได้ดังภาพประกอบ 3.2 ในบทที่ 3 เรียบร้อยแล้ว เวลาที่ใช้ในขั้นตอนหลักที่ A3 แสดงดังตารางที่ 5.27

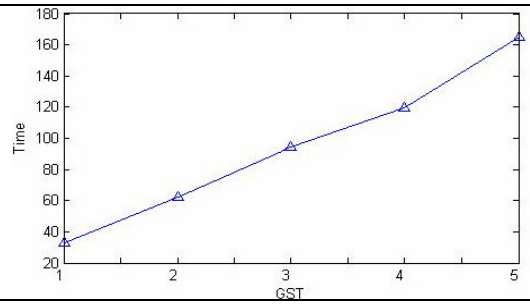
ตารางที่ 5.27 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A3 ของโรคมะเร็งลำไส้

GST (%)	เวลาที่ใช้ (วินาที)	
	ขั้นตอนการสร้างไฟล์	ขั้นตอนการจัดกลุ่มและคำนวณค่า HCV
1	0.75	33.05
2	0.86	62.25
3	0.89	94.57
4	0.92	119.5
5	0.98	165.06

จากตารางที่ 5.27 จะเห็นว่าเวลาที่ใช้ในการสร้างไฟล์กับเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV จะแปรผันตรงกับค่า GST คือเวลาที่ใช้ในการสร้างไฟล์และเวลาที่ใช้ในการจัดกลุ่มข้อมูลมีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น และมีแนวโน้มจะลดลงเมื่อค่า GST ลดลงด้วย แสดงดังภาพประกอบ 5.11



5.11 (a)



5.11 (b)

ภาพประกอบ 5.11 แสดงความสัมพันธ์ระหว่างเวลากับค่า GST ของโรคมะเร็งลำไส้ โดยที่ภาพ 5.11 (a) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการสร้างไฟล์กับค่า GST และภาพ 5.11 (b) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV กับค่า GST

จากภาพประกอบ 5.11 จะเห็นว่าเวลาที่ใช้มีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น เนื่องจากเมื่อค่า GST เพิ่มมากขึ้นก็แสดงว่าจะต้องเลือกยีนเพิ่มมากขึ้นดังแสดงจำนวนยีนที่เลือกเพิ่มขึ้นในตารางที่ 5.26 แล้ว เมื่อมียีนที่ถูกเลือกเพิ่มมากขึ้น จำนวนยีนที่จะสร้างไฟล์และจำนวนยีนที่จะจัดกลุ่มก็มากขึ้น ทำให้เวลาที่ใช้สูงขึ้นด้วย

ขั้นตอนหลักที่ A4: ขั้นตอนการเลือกยีนโดยอาศัยค่า HCV โดยจะเลือกเฉพาะยีนที่มีค่า HCV ที่คำนวณได้มากกว่าหรือเท่ากับค่า HCV_T ที่กำหนด ในขั้นตอนนี้ได้มีการทดลองกำหนดค่า HCV_T เป็นค่าต่างๆ คือ 95 93 และ 90 ตามลำดับ ผลลัพธ์จากขั้นตอนหลักที่ A4 แสดงดังตารางที่ 5.28

ตารางที่ 5.28 แสดงเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ของโรคมะเร็งลำไส้

GST (%)	เวลา (วินาที)		
	HCV = 95	HCV = 93	HCV = 90
1	0.13	0.14	0.14
2	0.16	0.11	0.11
3	0.11	0.13	0.12
4	0.13	0.13	0.11
5	0.11	0.13	0.16

จากตารางที่ 5.28 จะเห็นว่าเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ไม่ต่างกันมากนัก คือไม่ได้ขึ้นอยู่กับค่า GST หรือค่า HCV และเวลาที่ใช้ก็ยังถือว่าน้อยมากคือประมาณ 0.13 วินาทีเท่านั้น

ขั้นตอนหลักที่ A5: ขั้นตอนการสร้างกฎการเรียนรู้ในรูปแบบ “ถ้า–แล้ว” (IF–THEN Rule) ในขั้นตอนนี้จะเลือกยีนที่มีค่าการ Intersection มากกว่าหรือเท่ากับค่า Gene Intersection Threshold (GIT) ที่ผู้ใช้ต้องการ โดยที่ค่า GIT จะหาได้จากจำนวนครั้งที่ยีนนั้นถูกเลือกในแต่ละ Fold ตัวอย่างการคำนวณค่า GIT และการเลือกยีนได้แสดงไว้ในบทที่ 3 ในขั้นตอนหลักที่ 5 เรียบร้อยแล้ว ในการทดลองได้กำหนดค่า GIT เป็น 100% 70% และ 50% ซึ่งในส่วนกรณีค่า HCV = 95 และค่า GIT = 100% จะไม่มีกฎใดอยู่ในกรณีนี้คือไม่มีกฎใดสามารถเลือกได้นั่นเอง เวลาที่ใช้ในขั้นตอนหลักที่ A5 แสดงได้ดังตารางที่ 5.29 ส่วนจำนวนยีนที่เลือกได้และนำมาสร้างกฎ แสดงดังตารางที่ 5.30

ตารางที่ 5.29 แสดงเวลาที่ใช้ในการทำงานในขั้นตอนหลักที่ A5 ของโรคมะเร็งลำไส้

GST (%)	เวลา (วินาที)								
	HCV = 95			HCV = 93			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	-	0.44	0.5	0.36	0.36	0.63	0.5	0.59	0.67
2	-	0.87	0.5	0.59	0.47	0.73	0.81	0.73	0.84
3	-	0.69	0.9	0.78	0.75	0.78	0.75	0.74	0.78
4	-	0.39	0.56	0.59	0.61	0.64	0.56	0.48	0.74
5	-	0.63	0.55	0.59	0.67	0.61	0.75	0.84	1.05

ตารางที่ 5.30 แสดงจำนวนยีนที่เลือกได้ในขั้นตอนหลักที่ A5 ของโรคมะเร็งลำไส้

GST (%)	เวลา (วินาที)								
	HCV = 95			HCV = 93			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	-	2	2	2	3	5	2	4	8
2	-	2	2	2	3	5	2	4	10
3	-	2	2	2	3	5	4	2	10
4	-	2	3	2	3	8	2	4	16
5	-	2	4	2	3	11	2	4	21

จากตารางที่ 5.30 จะเห็นว่าจำนวนกฎมีแนวโน้มจะแปรผันตรงกับค่า GST และแปรผกผันกับค่า HCV และ GIT คือจำนวนกฎจะมากขึ้นในกรณีที่ค่า GST เพิ่มมากขึ้น และจะน้อยลงเมื่อค่า GST ลดลง แต่จำนวนกฎจะเพิ่มขึ้นเมื่อค่า HCV และค่า GIT ลดต่ำลง และจำนวนกฎจะลดต่ำลงเมื่อค่า HCV และ GIT เพิ่มมาก และจำนวนกฎที่น้อยที่สุดคือ 2 กฎ

ขั้นตอนหลักที่ A6: ขั้นตอนการนำกฎที่ได้ไปทดสอบกับชุดข้อมูลทดสอบที่สร้างไว้ในขั้นตอนหลักที่ A1 ในทุก Fold ผลลัพธ์ที่ได้จะเป็นค่าความถูกต้อง (Accuracy Values) ค่าความถูกต้องที่คำนวณได้จากการทำงานในขั้นตอนหลักที่ A6 แสดงดังตารางที่ 5.31

ตารางที่ 5.31 แสดงค่าความถูกต้องเฉลี่ยที่ได้จากขั้นตอนหลักที่ A6 ในแต่ละกรณี ของโรคมะเร็งลำไส้

GST (%)	ค่าความถูกต้อง (Accuracy)								
	HCV = 95			HCV = 93			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	-	82.64	82.64	83.07	82.33	79.27	83.07	81.81	78.62
2	-	82.64	82.64	83.07	82.33	79.27	83.07	81.81	75.81
3	-	82.64	82.64	83.07	82.33	79.27	83.01	81.81	75.81
4	-	82.64	78.77	83.07	82.33	76	83.07	81.81	73.81
5	-	82.64	75.49	83.07	82.33	73.38	83.07	81.81	72.13

จากตารางที่ 5.31 จะเห็นว่าค่าความถูกต้องจะแปรผันตรงกับค่า HCV และค่า GIT โดยที่จะแปรผกผันกับค่า GST คือค่าความถูกต้องจะเพิ่มขึ้นเมื่อค่า GST ลดลงในขณะที่ค่า HCV และค่า GIT เพิ่มสูงขึ้น และค่าความถูกต้องที่มากที่สุดคือ 83.07%

จากขั้นตอนหลักทั้ง 6 ขั้นตอนสามารถสรุปผลการทดลองในทุกกรณี ในประเด็นของ จำนวนกฎ ค่าความถูกต้องที่ได้ และเวลารวมที่ใช้ในแต่ละกรณี แสดงดังตารางที่ 5.34 และจากตารางที่ 5.34 จะเห็นว่ากรณีที่มีค่าความถูกต้องสูงที่สุดเท่ากับ 93.07% ซึ่งจำนวนกฎต่ำสุดคือ 2 กฎ และเวลารวมที่น้อยที่สุดคือ 0.89 นาที รายละเอียดของกรณีที่มีค่าความถูกต้องสูงสุดกล่าวแสดงได้ดังตารางที่ 5.32

ตารางที่ 5.32 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดจากแบบจำลอง A ของโรคมะเร็งลำไส้

ค่า Highest Certainty Value (HCV)	93
ค่า Gene Intersection Threshold (GIT)	100
จำนวนกฎ	2
ค่าความถูกต้อง	83.07
เวลาที่ใช้ (นาที)	0.89

จากตารางที่ 5.32 จะเห็นว่าเป็นการแสดงเฉพาะกรณีที่มีค่าความถูกต้องสูงสุดเท่านั้น เพราะค่าความถูกต้องสูงสุดนี้จะถือว่าเป็นค่าความถูกต้องที่ได้จากแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ ซึ่งต่อไปจะใช้ในการเปรียบประสิทธิภาพกับแบบจำลอง B, C และ D และเปรียบเทียบกับแบบจำลองอื่นๆในบทที่ 6 ด้วย

ในส่วนของผลการทดลองที่ได้จากแบบจำลอง B C และ D จะแสดงเฉพาะผลการทดลองที่ดีที่สุดของแต่ละแบบจำลอง ซึ่งแสดงได้ดังตารางที่ 5.33, 5.35 และ 5.36 ตามลำดับ

ตารางที่ 5.33 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง B ของโรคมะเร็งลำไส้

จำนวนกฎ	3
ค่าความถูกต้อง	83.01
เวลาที่ใช้ (นาที)	40.51

ตารางที่ 5.34 ผลลัพธ์ของข้อมูลโรคมะเร็งลำไส้

เงื่อนไข		HCV = 95			HCV = 93			HCV = 90		
		ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%
GST = 1%	จำนวนกฎ	-	2	2	2	3	5	2	4	8
	ค่าความถูกต้อง	-	82.64	82.64	83.07	82.33	79.27	83.07	81.81	78.62
	เวลารวม	-	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
GST = 2%	จำนวนกฎ	-	2	2	2	3	5	2	4	10
	ค่าความถูกต้อง	-	82.64	82.64	83.07	82.33	79.27	83.07	81.81	75.81
	เวลารวม	-	1.39	1.38	1.38	1.38	1.38	1.38	1.38	1.38
GST = 3%	จำนวนกฎ	-	2	2	2	3	5	4	2	10
	ค่าความถูกต้อง	-	82.64	82.64	83.07	82.33	79.27	83.01	81.81	75.81
	เวลารวม	-	1.92	1.92	1.92	1.92	1.92	1.92	1.92	1.92
GST = 4%	จำนวนกฎ	-	2	3	2	3	8	2	4	16
	ค่าความถูกต้อง	-	82.64	78.77	83.07	82.33	76	83.07	81.81	73.81
	เวลารวม	-	2.33	2.33	2.34	2.34	2.34	2.34	2.33	2.34
GST = 5%	จำนวนกฎ	-	2	4	2	3	11	2	4	21
	ค่าความถูกต้อง	-	82.64	75.49	83.07	82.33	73.38	83.07	81.81	72.13
	เวลารวม	-	3.10	3.10	3.10	3.10	3.10	3.10	3.10	3.11

จากตารางที่ 5.33 จะเห็นว่าแบบจำลอง B มีค่าความถูกต้องสูงสุดเท่ากับ 83.01% คือกรณีค่า HCV = 95 และค่า GIT = 70% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 3 กฎ และเวลารวมที่ใช้คือ 40.15 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

ตารางที่ 5.35 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง C ของโรคมะเร็งลำไส้

จำนวนกฎ	9
ค่าความถูกต้อง	79.44
เวลาที่ใช้ (นาที)	2.23

จากตารางที่ 5.35 จะเห็นว่าแบบจำลอง C มีค่าความถูกต้องสูงสุดเท่ากับ 79.44% คือกรณีค่า GST = 1% และค่า GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 9 กฎ และเวลารวมที่ใช้คือ 2.23 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

ตารางที่ 5.36 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง D ของโรคมะเร็งลำไส้

จำนวนกฎ	2,000
ค่าความถูกต้อง	62.3
เวลาที่ใช้ (นาที)	44.84

จากตางที่ 5.36 จะเห็นว่าแบบจำลอง C มีค่าความถูกต้อง จำนวนกฎ และเวลาเพียงค่าเดียวเท่านั้นเนื่องจากไม่มีการเลือกยีนโดยอาศัยค่า P-Value และค่า HCV เลย ดังนั้นผลการทดลองที่ดีที่สุดคือ ค่าความถูกต้องเท่ากับ 62.3% จำนวนกฎก็จะเท่ากับจำนวนยีนที่มีคือ 2,000 กฎ ส่วนเวลาที่ใช้ก็คือเวลาที่ใช้ในการจัดกลุ่มข้อมูลยีนทั้งหมดคือ 44.84 นาที

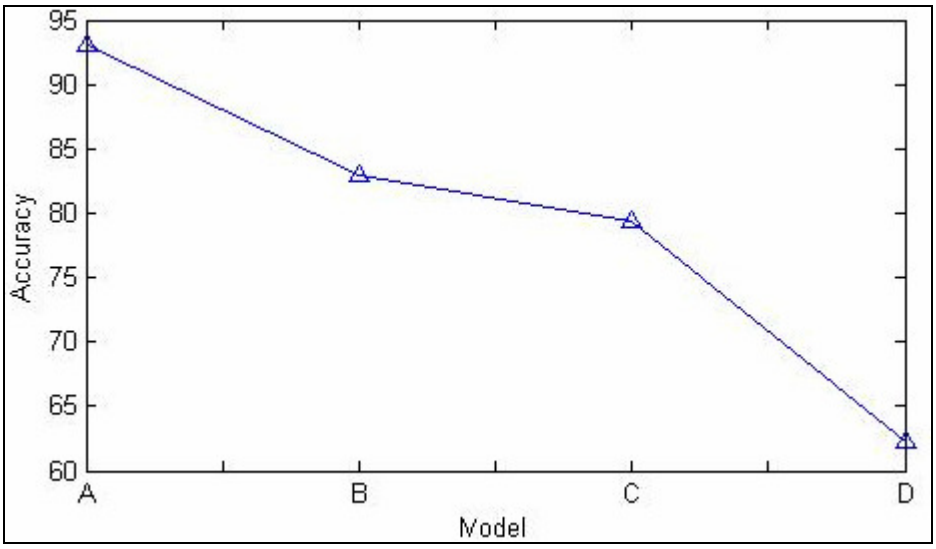
5.3.2 เปรียบเทียบผลการทดลองของโรคมะเร็งลำไส้

จากผลการทดลองที่ผ่านมาทั้ง 4 แบบจำลอง สามารถเปรียบเทียบผลการทดลองในประเด็นของ ค่าความถูกต้อง (Accuracy) จำนวนกฎ เวลาที่ใช้ในการทดลองได้ดังตารางที่ 5.38

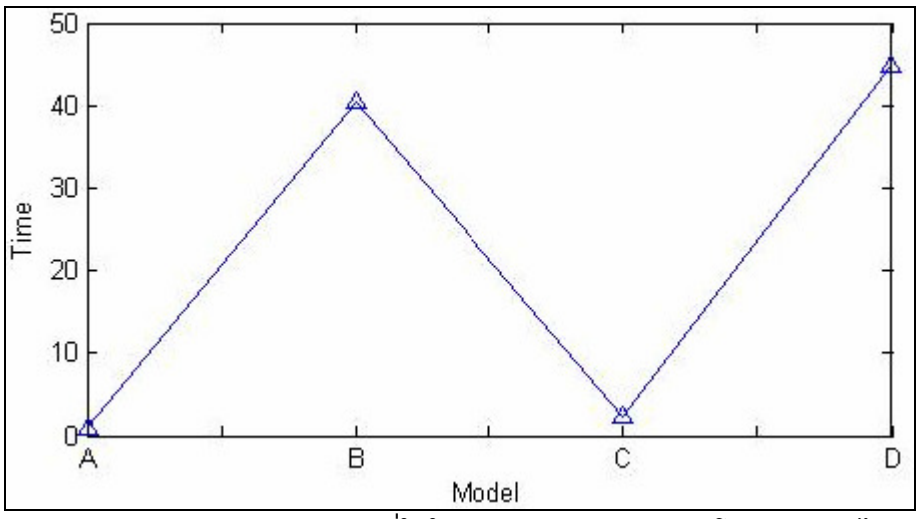
ตารางที่ 5.37 แสดงการเปรียบเทียบผลการทดลองของแบบจำลองทั้ง 4 ของโรคมะเร็งลำไส้

แบบจำลอง	**A**	B	C	D
ค่าความถูกต้อง (Accuracy)	**83.07**	83.01	79.44	62.3
จำนวนกฎ	**2**	3	9	2000
เวลารวม (นาที)	**0.89**	40.51	2.23	44.84

จากตารางที่ 5.37 สามารถสรุปได้ว่าแบบจำลอง A มีประสิทธิภาพมากที่สุดในทุกประเด็น ดังนี้ ค่าความถูกต้องของแบบจำลอง A มีค่าสูงสุด จำนวนกฎที่ได้จากแบบจำลอง A มีค่าน้อยที่สุด และเวลาที่ใช้ในแบบจำลอง A ก็น้อยที่สุดด้วยเช่นกัน กราฟแสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 ในประเด็นเรื่องค่าความถูกต้องและเวลาที่ใช้แสดงดังภาพประกอบ 5.12 และ 5.13 ตามลำดับ



ภาพประกอบ 5.12 แสดงค่า Accuracy ในแต่ละแบบจำลอง ของโรคมะเร็งลำไส้



ภาพประกอบ 5.13 แสดงเวลาที่ใช้ในแต่ละแบบจำลอง ของโรคมะเร็งลำไส้

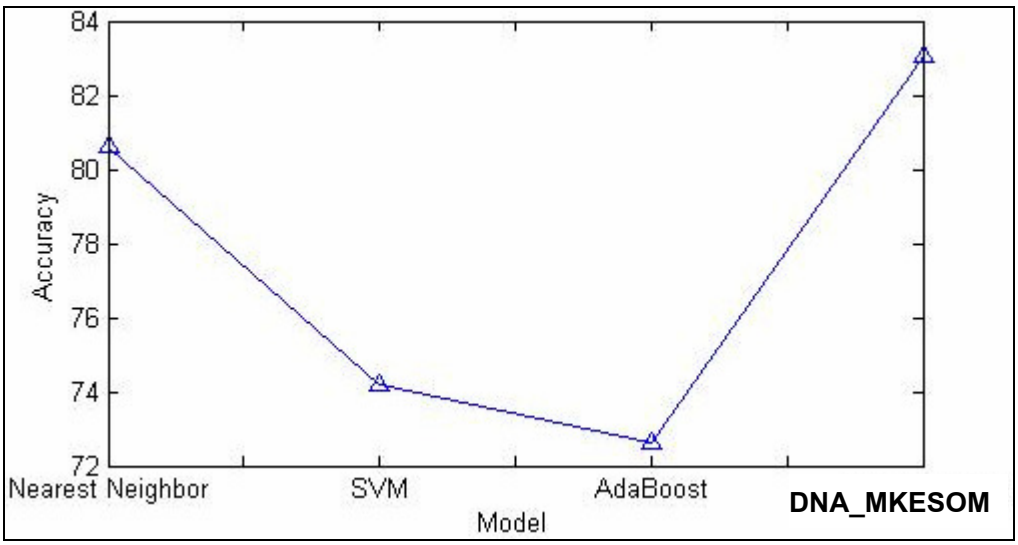
จากภาพประกอบ 5.12 และ 5.13 จะเห็นว่าแบบจำลอง A มีประสิทธิภาพมากที่สุด เนื่องจากมีค่าความถูกต้องสูงที่สุดและใช้เวลาน้อย แบบจำลอง B ให้ค่าความถูกต้องสูงแต่ใช้เวลามาก แบบจำลอง C ใช้เวลาน้อยแต่ให้ค่าความถูกต้องน้อยด้วย ส่วนแบบจำลอง D เป็นแบบจำลองที่มีประสิทธิภาพต่ำที่สุดเนื่องจากใช้เวลามากและได้ค่าความถูกต้องน้อย

จากนั้นเมื่อได้สรุปแล้วว่าแบบจำลอง A เป็นแบบจำลองที่ดีที่สุด จึงได้นำผลลัพธ์ที่ได้จากแบบจำลอง A มาทำการเปรียบเทียบกับงานวิจัยอื่นๆ ที่มีวัตถุประสงค์เช่นเดียวกัน และทำการทดลองบนชุดข้อมูลเดียวกัน โดยได้เปรียบเทียบกับงานวิจัยของ Ben และคณะ (Ben *et al.*, 2000) ที่จัดกลุ่มข้อมูลโดยใช้ Nearest Neighbor SVM with Quadratic และ AdaBoost แสดงดังตารางที่ 5.38

ตารางที่ 5.38 แสดงการเปรียบเทียบผลการทดลองของโรคมะเร็งลำไส้

Algorithm	Nearest Neighbor	SVM with quadratic Kernel	AdaBoost	DNA_MKESOM
Accuracy	80.6	74.2	72.6	83.07

จากตารางที่ 5.38 จะเห็นว่าแบบจำลอง DNA_MKESOM มีค่าความถูกต้องสูงที่สุดคือ 83.07% โดยค่าความถูกต้องแต่ละวิธีสามารถแสดงการเปรียบเทียบได้ดังภาพประกอบ 5.14



ภาพประกอบ 5.14 แสดงการเปรียบเทียบค่าความถูกต้องของโรคมะเร็งลำไส้

จากภาพประกอบ 5.14 จะเห็นว่าแบบจำลอง DNA_MKESOM ที่พัฒนาขึ้นให้ค่าความถูกต้องสูงที่สูงในจำนวนแบบจำลองต่าง ๆ ที่ใช้ในการทำนายกลุ่มของโรคมะเร็งลำไส้เช่นเดียวกัน

5.4 โรคมะเร็งปอด (Lung Cancer)

ข้อมูลของผู้ป่วยโรคมะเร็งปอด ประกอบด้วยข้อมูลดีเอ็นเอไมโครอาเรย์ของตัวอย่างทั้งหมด 181 ตัวอย่าง เป็นตัวอย่างของผู้ป่วยที่เป็นมะเร็งปอดประเภท MPM (Malignant Pleural Mesothelioma) 31 ตัวอย่าง และเป็นตัวอย่างของผู้ป่วยที่เป็นมะเร็งปอดประเภท ADCA (Adenocarcinoma) 150 ตัวอย่าง ข้อมูลดีเอ็นเอไมโครอาเรย์ของแต่ละตัวอย่างประกอบด้วยยีนทั้งหมด 12,533 ยีน ข้อมูลมีลักษณะเป็นอิสระต่อกัน (Unpaired Data) ตัวอย่างข้อมูลทั้งหมดแสดงดังภาพประกอบ 5.15

-8	18	-27.2	-75.2	-35.1	13.8	12.1	26.3	16.4	2	10.2	7.1
13.9	-35.6	-18.5	-19.3	13.1	17.3	-9.6	1.4	-0.7	-6.7	-5.3	0.7
79.5	125.2	51	7.4	37.7	58.5	102.6	129.8	106.9	125.4	88.1	93.8
-73.8	6.3	35.4	25.3	99.7	-47.5	14.8	-8.3	4.8	4.8	-17.7	-1.8
51.6	78.2	136.6	128.3	74.6	71	51.7	58.4	41.1	40.4	70	46.7
-50.9	-3.8	-3.6	-31.9	-10.1	-11.6	-9.4	1.4	-3.2	-18.8	-29.4	-7
-118.1	1	-81.5	110.1	14.3	-86.5	-88.2	15.3	-84.9	-6.8	-24.5	-2.8
53.5	124.6	83	7.6	48.8	86.1	70.7	27	44.8	129.2	53.2	36.4
41.2	67.5	74.1	50.6	47	27.2	42.4	52.1	45	41.8	32.8	26.3
7.4	39.7	55.4	-14	-10.4	54.3	109.6	51.3	121.2	76.3	187	61
42	25.3	-50	-20	9.1	1.3	15.8	-16.2	45.9	31.1	52.1	49.9
10.6	-42.3	5.9	17.7	9	7.7	31.6	28.8	10.1	25.9	11.3	20.9
81.4	-72.9	-6.3	58	71.1	99.9	60.5	108.4	43.1	59.3	62.4	39.6
14.6	18.5	-8.7	17.7	-22.5	40.5	1.9	-19.5	1.5	18.2	-4.8	15.3
76.3	44	98.6	61.1	113.8	115.1	92.4	222.6	137.9	134.4	141.4	170.5
-6.7	7.8	8.1	-2.1	-64.1	30.2	24.9	38.4	18.2	32.3	-6.8	39.2
48.8	48.2	-1.7	3.5	53.1	28.4	61.2	82.3	60.7	73	19.9	102.5
-8.4	52.9	71.5	57.7	2.7	37.3	60.2	-12.2	21.7	9.5	21.5	12
18.2	19.3	-25.9	-2.7	-43.5	74.2	38.4	73.1	32.8	51.5	23.5	26.8
-114.2	-64.2	-83.4	-43.9	-13.3	-163.3	183.2	-71.2	189.4	71.6	-76.2	6.4
Mesotheli	Mesotheli	Mesotheli	Mesotheli	Mesotheli	ADCA	ADCA	ADCA	ADCA	ADCA	ADCA	ADCA

ภาพประกอบ 5.15 แสดงตัวอย่างข้อมูลดีเอ็นเอไมโครอาเรย์ของโรคมะเร็งปอด

ในส่วนของโรคมะเร็งปอดนี้จะอธิบายแยกเป็น 2 ส่วนคือ ผลการทดลองของโรคมะเร็งปอด และเปรียบเทียบผลการทดลองของโรคมะเร็งปอด

5.4.1 ผลการทดลองของโรคมะเร็งปอด

สำหรับการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โรคมะเร็งปอดได้ทำการทดลอง 2 แบบจำลองด้วยกันคือแบบจำลอง A และแบบจำลอง C เนื่องจาก 2 แบบจำลองนี้มีการใช้ค่า P-Value ในการเลือกยีนก่อนการจัดกลุ่มจึงไม่ต้องทำการจัดกลุ่มข้อมูลด้วย SOM ทั้งหมด ซึ่งจะช่วยลดเวลาลงอย่างมาก ส่วนแบบจำลอง B และ D จะไม่มีการทำ Feature Extraction ด้วยค่า P-Value ทำให้ต้องเสียเวลาในการจัดกลุ่มนานมากเพราะต้องนำยีนทุกตัวมาจัดกลุ่ม ซึ่งในกรณีนี้จำนวนยีนมีมากถึง 12,533 ยีน ทำให้โปรแกรมมีปัญหาเรื่องหน่วยความจำไม่เพียงพอจึงไม่สามารถทำงานได้ ดังนั้นจะอธิบายโดยละเอียดเฉพาะแบบจำลองหลักคือแบบจำลอง A และทำการเปรียบเทียบกับแบบจำลอง C เท่านั้น

ขั้นตอนหลักที่ A1: ขั้นตอนการเตรียมข้อมูล กำหนดจำนวน Fold เท่ากับ 4 ระบบจะแบ่งข้อมูลออกเป็นส่วนตามจำนวน Fold ที่กำหนดและแต่ละ Fold จะมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ ในขั้นตอนนี้จะใช้เวลาประมาณ 76.55 วินาที ซึ่งจะมากกว่าข้อมูลชุดอื่นเนื่องจากข้อมูลโรคมะเร็งปอดนี้มีจำนวนยีนและจำนวนตัวอย่างมากกว่าข้อมูลชุดอื่นๆ

ขั้นตอนหลักที่ A2: ขั้นตอนการคำนวณค่าของ P-Value และการเลือกกลุ่มยีนที่มีค่า P-Value ต่ำ จากข้อมูลที่ได้เตรียมไว้ในขั้นตอนหลักที่ A1 โดยจะทำการคำนวณค่า P-Value ของชุดของมูลสอนทุกยีนในทุก Fold และส่วนของการเลือกยีนจะพิจารณาจากค่า Gene Selection Threshold (GST) ที่ผู้ใช้ป้อน โดยได้ทำการทดลอง 5 กรณีคือ กรณีค่า GST = 1% ถึง 5% ผลการทำงานในขั้นตอนหลักที่ A2 แสดงได้ดังตารางที่ 5.39

ตารางที่ 5.39 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A2 ของโรคมะเร็งปอด

GST (%)	จำนวนยีนที่เลือก	เวลาที่ใช้ (วินาที)	
		คำนวณค่า P-Value	การเลือกยีน
1	125	29.34	0.38
2	251	29.34	0.66
3	376	29.34	0.49
4	501	29.34	0.45
5	627	29.34	0.45

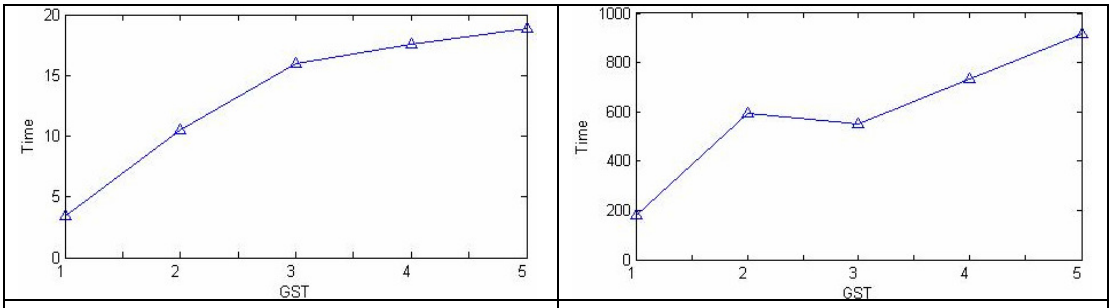
จากตารางที่ 5.39 จะเห็นว่าการเพิ่มค่า GST ไม่มีผลต่อเวลาที่ใช้ในการคำนวณค่า P-Value เนื่องจากทุกขั้นตอนจะมีจำนวนครั้งในการคำนวณค่า P-Value ทุกยีนเท่ากัน ส่วนเวลาที่ใช้ในการเลือกยีนก็ไม่ค่อยมีผลกระทบที่เกี่ยวข้องกับค่า GST เช่นกัน เนื่องจากเวลาที่ใช้ในการเลือกยีนก็มีค่าใกล้เคียงกันในแต่ละกรณี และถ้าเปรียบเทียบกับเวลาที่ต้องใช้ในการทดลองทั้งหมดถือว่าเป็นเวลาเพียงเล็กน้อย

ขั้นตอนหลักที่ A3: ขั้นตอนการจัดกลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง และการคำนวณค่า HCV ขั้นตอนนี้จะนำข้อมูลสอนของยีนที่เลือกมาจากขั้นตอนหลักที่ A2 มาทำการสร้างไฟล์ข้อมูลขึ้น โดยการสร้างไฟล์จะแยกแต่ละยีนออกมาเป็น 1 ไฟล์และรูปแบบของข้อมูลต้องถูกต้องตามข้อกำหนดของข้อมูลที่ใช้แผนที่การจัดกลุ่มเอง ซึ่งตัวอย่างข้อมูลแสดงได้ดังภาพประกอบ 3.2 ในบทที่ 3 เรียบร้อยแล้ว เวลาที่ใช้ในขั้นตอนหลักที่ A3 แสดงดังตารางที่ 5.40

ตารางที่ 5.40 แสดงผลลัพธ์ที่ได้จากการทำงานในขั้นตอนหลักที่ A3 ของโรคมะเร็งปอด

GST (%)	เวลาที่ใช้ (วินาที)	
	ขั้นตอนการสร้างไฟล์	ขั้นตอนการจัดกลุ่มและคำนวณค่า HCV
1	3.39	179.75
2	10.49	593.81
3	15.95	550.31
4	17.52	733.3
5	18.84	913.84

จากตารางที่ 5.40 จะเห็นว่าเวลาที่ใช้ในการสร้างไฟล์กับเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV จะแปรผันตรงกับค่า GST คือเวลาที่ใช้ในการสร้างไฟล์และเวลาที่ใช้ในการจัดกลุ่มข้อมูลมีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น และมีแนวโน้มจะลดลงเมื่อค่า GST ลดลงด้วย แสดงดังภาพประกอบ 5.16



5.16 (a) 5.16 (b)

ภาพประกอบ 5.16 แสดงความสัมพันธ์ระหว่างเวลากับค่า GST ของโรคมะเร็งปอด โดยภาพ 5.16 (a) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการสร้างไฟล์กับค่า GST และภาพ 5.16 (b) แสดงความสัมพันธ์ระหว่างเวลาที่ใช้ในการจัดกลุ่มและคำนวณค่า HCV กับค่า GST

จากภาพประกอบ 5.16 จะเห็นว่าเวลาที่ใช้มีแนวโน้มจะเพิ่มขึ้นเมื่อค่า GST เพิ่มมากขึ้น เนื่องจากเมื่อค่า GST เพิ่มมากขึ้นก็แสดงว่าจะต้องเลือกยีนเพิ่มมากขึ้นดังแสดงจำนวนยีนที่เลือกเพิ่มขึ้นในตารางที่ 5.39 แล้ว เมื่อมียีนที่ถูกเลือกเพิ่มมากขึ้น จำนวนยีนที่จะสร้างไฟล์และจำนวนยีนที่จะจัดกลุ่มก็มากขึ้น ทำให้เวลาที่ใช้สูงขึ้นด้วย

ขั้นตอนหลักที่ A4: ขั้นตอนการเลือกยีนโดยอาศัยค่า HCV โดยจะเลือกเฉพาะยีนที่มีค่า HCV ที่คำนวณได้มากกว่าหรือเท่ากับค่า HCV_T ที่กำหนด ในขั้นตอนนี้ได้มีการทดลองกำหนดค่า HCV_T เป็นค่าต่างๆ คือ 100 95 และ 90 ตามลำดับ ผลลัพธ์จากขั้นตอนหลักที่ A4 แสดงดังตารางที่ 5.41

ตารางที่ 5.41 แสดงเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ของโรคมะเร็งปอด

GST	เวลา (วินาที)		
	HCV = 100	HCV = 95	HCV = 90
1	0.41	0.11	0.11
2	0.09	0.11	0.14
3	0.11	0.13	0.16
4	0.13	0.13	0.14
5	0.13	0.12	0.17

จากตารางที่ 5.41 จะเห็นว่าเวลาที่ใช้ในการเลือกยีนโดยอาศัยค่า HCV ไม่ค่อยแตกต่างกัน คือไม่ได้ขึ้นอยู่กับค่า GST หรือค่า HCV และเวลาที่ใช้ก็ยังถือว่าน้อยมากคือประมาณ 0.15 วินาทีเท่านั้น

ขั้นตอนหลักที่ A5: ขั้นตอนการสร้างกฎการเรียนรู้ในรูปแบบ “ถ้า–แล้ว” (IF–THEN Rule) ในขั้นตอนนี้จะเลือกยีนที่มีค่าการ Intersection มากกว่าหรือเท่ากับค่า Gene Intersection Threshold (GIT) ที่ผู้ใช้ต้องการ โดยที่ค่า GIT จะหาได้จากจำนวนครั้งที่ยีนนั้นถูกเลือกในแต่ละ Fold ตัวอย่างการคำนวณค่า GIT และการเลือกยีนได้แสดงไว้ในบทที่ 3 ในขั้นตอนหลักที่ 5 เรียบร้อยแล้ว ในการทดลองได้กำหนดค่า GIT เป็น 100% 70% และ 50% เวลาที่ใช้ในขั้นตอนหลักที่ A5 แสดงได้ดังตารางที่ 5.42 ส่วนจำนวนยีนที่เลือกได้และนำมาสร้างกฎ แสดงดังตารางที่ 5.43

ตารางที่ 5.42 แสดงเวลาที่ใช้ในการทำงานในขั้นตอนหลักที่ A5 ของโรคมะเร็งปอด

GST	เวลา (วินาที)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	0.67	0.49	0.53	0.7	0.81	1.89	2.59	4.28	4.84
2	0.59	0.81	0.47	1.69	4.59	6.38	9.44	14.42	18.52
3	0.52	0.66	0.58	1.81	5.27	5.41	10.45	16.19	17.3
4	0.63	0.55	0.49	2.8	7.61	10.5	19.08	25.02	31.94
5	0.56	0.88	0.53	3.13	6.49	9.42	21.69	18.34	24.42

ตารางที่ 5.43 แสดงจำนวนยีนที่เลือกได้ในขั้นตอนหลักที่ A5 ของโรคมะเร็งปอด

GST (%)	เวลา (วินาที)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	3	6	12	41	64	91	77	98	134
2	5	8	14	63	102	150	148	200	250
3	5	10	17	73	129	188	204	287	364
4	5	13	21	81	142	222	245	348	469
5	8	18	28	90	158	259	299	419	583

จากตารางที่ 5.43 จะเห็นว่าจำนวนกฎมีแนวโน้มจะแปรผันตรงกับค่า GST และแปรผกผันกับค่า HCV และ GIT คือจำนวนกฎจะมากขึ้นในกรณีที่ค่า GST เพิ่มมากขึ้น และจะน้อยลงเมื่อค่า GST ลดลง แต่จำนวนกฎจะเพิ่มขึ้นเมื่อค่า HCV และค่า GIT ลดต่ำลง และจำนวนกฎจะลดต่ำลงเมื่อค่า HCV และ GIT เพิ่มมาก และจำนวนกฎที่น้อยที่สุดคือ 3 ซึ่งเกิดขึ้นในกรณีที่ค่า GST = 1% ค่า HCV = 100 และค่า GIT = 100%

ขั้นตอนหลักที่ A6: ขั้นตอนการนำกฎที่ได้ไปทดสอบกับชุดข้อมูลทดสอบที่สร้างไว้ในขั้นตอนหลักที่ A1 ในทุก Fold ผลลัพธ์ที่ได้จะเป็นค่าความถูกต้อง (Accuracy Values) ค่าความถูกต้องที่คำนวณได้จากการทำงานในขั้นตอนหลักที่ A6 แสดงดังตารางที่ 5.44

ตารางที่ 5.44 แสดงค่าความถูกต้องเฉลี่ยที่ได้จากขั้นตอนหลักที่ A6 ในแต่ละกรณี ของโรคมะเร็งปอด

GST (%)	ค่าความถูกต้อง (Accuracy)								
	HCV = 100			HCV = 95			HCV = 90		
	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%	GIT = 100%	GIT = 70%	GIT = 50%
1	96.5	95.49	94.61	93.58	92.38	91.11	91.9	91.08	89.76
2	95.1	95.64	94.5	92.55	91.31	89.91	90.46	89.16	88.11
3	95.1	94.8	94.01	92.11	90.73	89.45	89.59	88.37	87.23
4	95.1	94.09	93.44	92.07	90.62	89.14	89.18	88.01	86.84
5	94.04	93.03	92.83	91.8	90.43	88.86	88.87	87.65	86.43

จากตารางที่ 5.44 จะเห็นว่าค่าความถูกต้องจะแปรผันตรงกับค่า HCV และค่า GIT และจะแปรผกผันกับค่า GST คือค่าความถูกต้องจะเพิ่มขึ้นเมื่อค่า GST ลดลงในขณะที่ค่า HCV และค่า GIT เพิ่มสูงขึ้น และค่าความถูกต้องที่มากที่สุดคือ 96.5 % ที่กรณีค่า GST = 1% HCV = 100 และ GIT = 100%

จากขั้นตอนหลักทั้ง 6 ขั้นตอนสามารถสรุปผลการทดลองในทุกกรณี ในประเด็นของ จำนวนกฎ ความถูกต้องที่ได้ และเวลารวมที่ใช้ในแต่ละกรณี แสดงดังตารางที่ 5.46 และจากตารางที่ 5.46 จะเห็นว่ากรณีที่มีค่าความถูกต้องสูงที่สุดเท่ากับ 96.5% คือกรณีค่า GST = 1%, HCV = 100, GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุดคือ 3 กฎ และเวลารวมที่ใช้คือ 4.88 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดเช่นกัน รายละเอียดของกรณีดังที่มีค่าความถูกต้องสูงสุดกล่าวแสดงได้ดังตารางที่ 5.45

ตารางที่ 5.45 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดจากแบบจำลอง A ของโรคมะเร็งปอด

ค่า Gene Selection threshold (GST)	1
ค่า Highest Certainty Value (HCV)	100
ค่า Gene Intersection Threshold (GIT)	100
จำนวนกฎ	3
ค่าความถูกต้อง	**96.5**
เวลาที่ใช้ (นาที)	4.88

จากตารางที่ 5.45 จะเห็นว่าเป็นการแสดงเฉพาะกรณีที่มีค่าความถูกต้องสูงสุดเท่านั้น เพราะค่าความถูกต้องสูงสุดนี้จะถือว่าเป็นค่าความถูกต้องที่ได้จากแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ ซึ่งต่อไปจะใช้ในการเปรียบประสิทธิภาพกับแบบจำลอง B C และ D และเปรียบเทียบกับแบบจำลองอื่นๆ ในบทที่ 6 ด้วย

ในส่วนของผลการทดลองที่ได้จากแบบจำลอง C จะแสดงเฉพาะผลการทดลองที่ดีที่สุดของแบบจำลอง ซึ่งแสดงได้ดังตารางที่ 5.47

ตารางที่ 5.46 ผลลัพธ์ของข้อมูลโรคมะเร็งปอด

เงื่อนไข		HCV = 100			HCV = 95			HCV = 90		
		ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%	ITS = 100%	ITS = 70%	ITS = 50%
GST = 1%	จำนวนกฎ	**3**	6	12	41	64	91	77	98	134
	ค่าความถูกต้อง	**96.5**	95.49	94.61	93.58	92.38	91.11	91.9	91.08	89.76
	เวลารวม	**4.88**	4.88	4.88	4.88	4.88	4.90	4.91	4.94	4.95
GST = 2%	จำนวนกฎ	5	8	14	63	102	150	148	200	250
	ค่าความถูกต้อง	95.1	95.64	94.5	92.55	91.31	89.91	90.46	89.16	88.11
	เวลารวม	11.90	11.90	11.90	11.92	11.97	12.00	12.05	12.13	12.20
GST = 3%	จำนวนกฎ	5	10	17	73	129	188	204	287	364
	ค่าความถูกต้อง	95.1	94.8	94.01	92.11	90.73	89.45	89.59	88.37	87.23
	เวลารวม	11.26	11.26	11.26	11.28	11.34	11.35	11.43	11.53	11.55
GST = 4%	จำนวนกฎ	5	13	21	81	142	222	245	348	469
	ค่าความถูกต้อง	95.1	94.09	93.44	92.07	90.62	89.14	89.18	88.01	86.84
	เวลารวม	14.34	14.34	14.34	14.38	14.46	14.51	14.65	14.75	14.87
GST = 5%	จำนวนกฎ	8	18	28	90	158	259	299	419	583
	ค่าความถูกต้อง	94.04	93.03	92.83	91.8	90.43	88.86	88.87	87.65	86.43
	เวลารวม	17.37	17.37	17.37	17.41	17.47	17.52	17.73	17.67	17.79

ตารางที่ 5.47 แสดงรายละเอียดกรณีที่มีค่าความถูกต้องสูงสุดในแบบจำลอง C ของโรคมะเร็งปอด

จำนวนกฎ	77
ค่าความถูกต้อง	**91.9**
เวลาที่ใช้ (นาที)	7.02

จากตารางที่ 5.47 จะเห็นว่าแบบจำลอง C มีค่าความถูกต้องสูงสุดเท่ากับ 91.9% คือกรณีค่า GST = 1% และค่า GIT = 100% โดยกรณีดังกล่าวยังมีค่าจำนวนกฎต่ำสุด คือ 77 กฎ และเวลารวมที่ใช้คือ 7.02 นาที ซึ่งเป็นเวลารวมที่ต่ำสุดในแบบจำลองนี้เช่นกัน

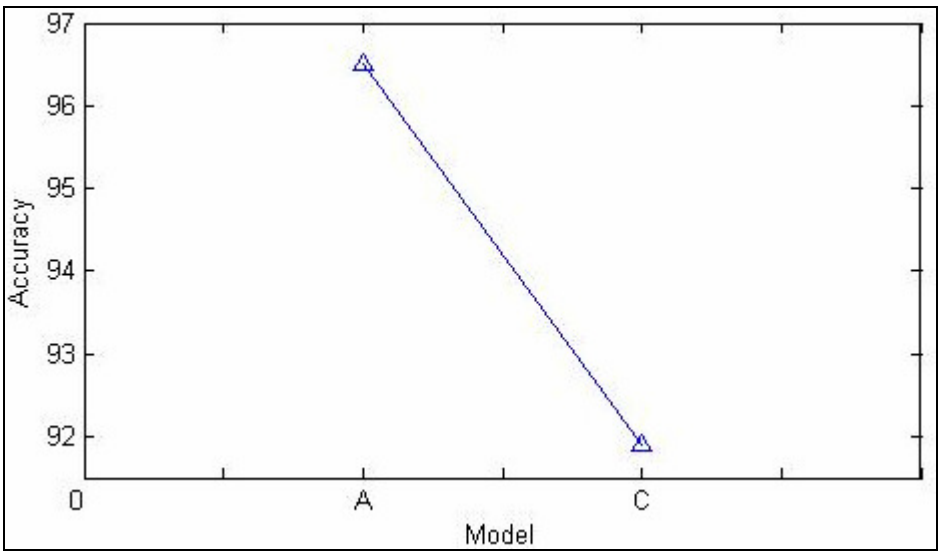
5.4.2 เปรียบเทียบผลการทดลองของโรคมะเร็งปอด

จากผลการทดลองที่ผ่านมาทั้ง 2 แบบจำลอง สามารถเปรียบเทียบผลการทดลองในประเด็นของ ค่าความถูกต้อง (Accuracy) จำนวนกฎ เวลาที่ใช้ในการทดลองได้ดังตารางที่ 5.48

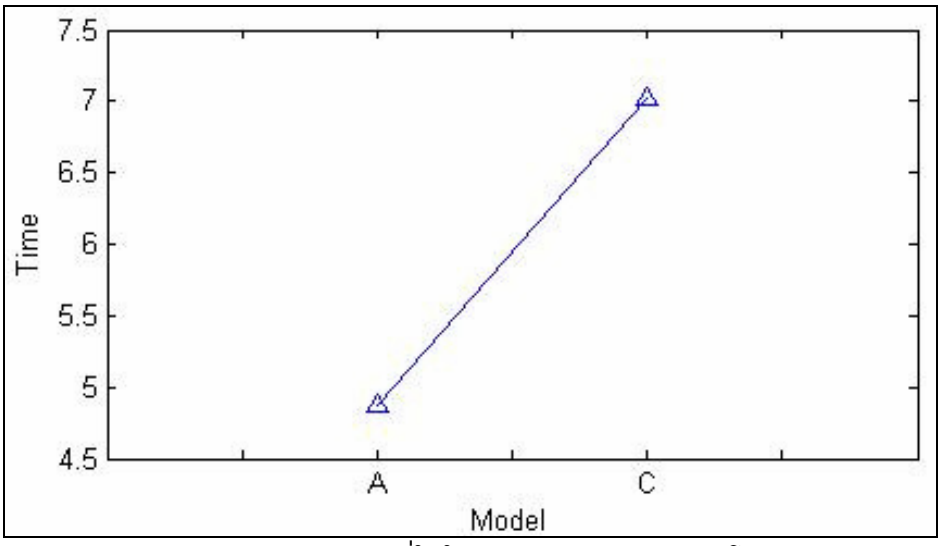
ตารางที่ 5.48 แสดงการเปรียบเทียบผลการทดลองของแบบจำลองทั้ง 2 ของโรคมะเร็งปอด

แบบจำลอง	**A**	C
ค่าความถูกต้อง (Accuracy)	**96.5**	91.9
จำนวนกฎ	**3**	77
เวลารวม (นาที)	**4.88**	7.02

จากตารางที่ 5.48 จะสามารถสรุปได้ว่าแบบจำลอง A มีประสิทธิภาพมากที่สุดในทุกประเด็น ดังนี้ ค่าความถูกต้องของแบบจำลอง A มีค่าสูงสุด จำนวนกฎที่ได้จากแบบจำลอง A มีค่าน้อยที่สุด และเวลาที่ใช้ในแบบจำลอง A ก็น้อยที่สุดด้วย กราฟแสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 ในประเด็นเรื่องค่าความถูกต้องและเวลาที่ใช้แสดงดังภาพประกอบ 5.17 และ 5.18 ตามลำดับ



ภาพประกอบ 5.17 แสดงค่า Accuracy ในแต่ละแบบจำลอง ของโรคมะเร็งปอด



ภาพประกอบ 5.18 แสดงเวลาที่ใช้ในแต่ละแบบจำลอง ของโรคมะเร็งปอด

จากภาพประกอบ 5.17 และ 5.18 จะเห็นว่าแบบจำลอง A มีประสิทธิภาพสูงกว่าแบบจำลอง C เนื่องจากมีค่า Accuracy สูงกว่า และใช้เวลาน้อยกว่า

บทที่ 6

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้บรรลุตามวัตถุประสงค์คือได้ออกแบบและพัฒนาแบบจำลองการวิเคราะห์ข้อมูลดีเอ็นเอจากไมโครอาเรย์โดยใช้ขั้นตอนวิธีการแบ่งกลุ่มสำหรับการวินิจฉัยโรคมีชื่อว่า แบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเอง (DNA Microarray Knowledge Extraction Using Self-Organizing Map: DNA_MKESOM) สามารถสกัดความรู้จากข้อมูลต่างๆ ที่อยู่ในรูปแบบของดีเอ็นเอไมโครอาเรย์ได้ ในส่วนของการทำงานแบบจำลองนี้ได้ใช้วิธีการทางสถิติคือ P-Value เข้ามาช่วยในการกรองยีนในขั้นตอนก่อนทำการจัดกลุ่มด้วยแผนที่การจัดกลุ่มเอง (Self-Organizing Map: SOM)

ผลลัพธ์ที่ได้จากการงานวิจัยนี้ จะได้แบบจำลองที่สามารถสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์ ให้อยู่ในรูปแบบของกฎ ถ้า-แล้ว (IF-THEN Rule) โดยกฎที่ได้สามารถใช้ในการทำนายประเภทของโรคมะเร็งชนิดต่างๆ ที่มีข้อมูลอยู่ในรูปแบบของดีเอ็นเอไมโครอาเรย์ได้อย่างมีประสิทธิภาพ คือมีค่าความถูกต้องที่สูง และเวลาที่ใช้ในการทำงานน้อย นอกจากนั้นยังได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลอง DNA_MKESOM ที่พัฒนาขึ้นกับแบบจำลองอื่นๆ ที่มีวัตถุประสงค์เช่นเดียวกัน พบว่าแบบจำลอง DNA_MKESOM สามารถให้ค่าความถูกต้องสูงกว่าเมื่อเทียบกับแบบจำลองที่ใช้วิธีการอื่นๆ งานวิจัยนี้ได้รับการตีพิมพ์ผลงานวิจัยในระดับประเทศคือ เรื่อง Microarray Gene Selection Using Self-Organizing Map for Cancer Prediction ซึ่งตีพิมพ์ใน The 6th Joint Conference on Computer Science and Software Engineering (JCSSE 2007) ที่จังหวัดขอนแก่น ประเทศไทย ระหว่างวันที่ 2-4 พฤษภาคม 2550 ดังแสดงในภาคผนวก ข. และตีพิมพ์ในระดับนานาชาติ เรื่อง Microarray Gene Selection Using Self-Organizing Map ซึ่งตีพิมพ์ใน The 7th WSEAS International Conference on Simulation Modeling and Optimization (SOM 2007) ที่กรุงปักกิ่ง ประเทศจีน ระหว่างวันที่ 15-17 กันยายน 2550 โดยมีการเพิ่มตัวอย่างผลการทดลอง และวิธีการในการเลือกค่า P-Value เพิ่มเติม ดังแสดงในภาคผนวก ค

6.1 สรุปผลการวิจัย

การสรุปผลการทดลองจะสรุปแยกเป็นประเด็น โดยจะอธิบายตามขั้นตอนการทำงานหลักทั้ง 6 ขั้นตอน ดังนี้ 1) ขั้นตอนการเตรียมข้อมูล 2) ขั้นตอนการเลือกยีนโดยอาศัย P-Value จะสรุปในประเด็นการเลือกค่า Gene Selection Threshold (GST) 3) ขั้นตอนการจัด

กลุ่มข้อมูลด้วยแผนที่การจัดกลุ่มเอง 4) ขั้นตอนการเลือกยีนโดยอาศัยค่า Highest Certainty Value (HCV) จะสรุปในประเด็นของการเลือกค่า Highest Certainty Value Threshold (HCV_T) 5) ขั้นตอนสร้างกฎ จะสรุปในประเด็นของการเลือกค่า Gene Intersection Threshold (GIT) และ 6) ขั้นตอนการคำนวณค่าความถูกต้อง จะสรุปประเด็นการเลือกยีนโดยใช้ค่า P-Value และค่า HCV ซึ่งมีรายละเอียดดังนี้

6.1.1 ประเด็นการเลือกค่า Gene Selection Threshold (GST)

ค่า GST เป็นค่าที่จะใช้ในการกำหนดจำนวนของยีนที่จะเลือกในขั้นตอนที่ 2 ซึ่งถ้าค่า GST มาก ก็จะมีจำนวนยีนที่ถูกเลือกมาก จำนวนกฎที่ได้และเวลาที่ใช้จะมาก ถ้าค่า GST น้อยก็จะมียีนที่ถูกเลือกน้อย จำนวนกฎที่ได้และเวลาที่ใช้จะน้อยลง จากผลการทดลองสามารถสรุปได้ว่า ค่า GST ที่เหมาะสมควรจะเป็นค่า 1% - 5%

6.1.2 ประเด็นการเลือกค่า Highest Certainty Value Threshold (HCV_T)

เนื่องจากในขั้นตอนก่อนหน้านี้คือขั้นตอนที่ 3 ได้มีการคำนวณค่า Highest Certainty Value (HCV) ไว้เรียบร้อยแล้ว และในขั้นตอนที่ 4 จะให้เลือกค่า HCV_T ซึ่งเป็นค่าที่จะใช้ในการกำหนดว่าจะเลือกยีนที่มีค่า HCV เท่าใดมาทำงานต่อไป โดยจะเลือกจะเลือกเฉพาะกลุ่มยีนที่มีค่า HCV ที่คำนวณได้มากกว่าหรือเท่ากับค่า HCV_T ที่กำหนด จากผลการทดลองสามารถสรุปได้ว่า ค่า HCV_T ที่เหมาะสมควรจะเป็นค่าสูงสุดที่มี เช่น 100% จะได้ค่าความถูกต้องสูงและใช้เวลาน้อย

6.1.3 ประเด็นการเลือกค่า Gene Intersection Threshold (GIT)

ค่า GIT เป็นค่าที่จะใช้ในการกำหนดว่ายีนไหนบ้างจะถูกนำไปสร้างเป็นกฎ โดยพิจารณาจากจำนวนความถี่ของการแสดงค่ายีนนั้นในแต่ละ Fold จากผลการทดลองสามารถสรุปได้ว่า ค่า GIT ที่เหมาะสมควรจะเป็นค่าสูงสุดที่มีเช่น 100% (มีทุก Fold) เนื่องจากถ้าเลือกค่า GST มากแล้วยีนที่เลือกมาสร้างกฎเมื่อนำไปทดสอบกับข้อมูลทดสอบจะเป็นยีนที่ให้ค่าความถูกต้องสูง

6.1.4 ประเด็นการเลือกยีนโดยใช้ค่า P-Value และค่า HCV

จากผลการทดลองสรุปได้ว่าการเลือกยีนโดยใช้ค่า P-Value จะช่วยลดเวลาในการทำงานลงได้อย่างมาก แต่ค่าความถูกต้องยังไม่สูงมาก แต่การเลือกยีนโดยใช้ค่า HCV จะช่วยให้ได้ค่าความถูกต้องที่สูงขึ้น แต่ใช้เวลามาก

6.2 ปัญหาและอุปสรรค

6.2.1 ข้อมูลดีเอ็นเอไมโครอาเรย์เป็นข้อมูลที่มีลักษณะเฉพาะคือมีจำนวนแอทริบิวต์มาก ทำให้ยากต่อการทำงานเพราะโปรแกรมส่วนใหญ่มักมีข้อจำกัดในการทำงานกับข้อมูลที่มีจำนวนแอทริบิวต์มาก เช่น โปรแกรม Microsoft Excel จะไม่สามารถอ่านข้อมูลดีเอ็นเอไมโครอาเรย์ได้ ทำให้ต้องเขียนโปรแกรมด้วยภาษา C# เพื่อแปลงข้อมูลให้เหมาะสมก่อน

6.2.3 เนื่องจากข้อมูลดีเอ็นเอไมโครอาเรย์เป็นข้อมูลที่มีแอทริบิวต์จำนวนมากในการทำงานโดยใช้โปรแกรม MATLAB จึงต้องการเครื่องคอมพิวเตอร์ที่มีสมรรถนะสูง เนื่องจากเกิดปัญหาหน่วยความจำไม่พอบ่อยครั้งในระหว่างทำการทดลอง

6.3 ข้อเสนอแนะ

6.3.1 เนื่องจากข้อมูลดีเอ็นเอไมโครอาเรย์มีจำนวนแอทริบิวต์มาก จึงทำให้การทำการทดลองกับข้อมูลดีเอ็นเอไมโครอาเรย์มักจะใช้เวลานาน และต้องการทรัพยากรที่สูงมาก จึงต้องเตรียมเครื่องคอมพิวเตอร์ที่สมรรถนะสูง เพื่อลดปัญหาในด้านเวลาและความยุ่งยากเกี่ยวกับสมมรรถนะของอุปกรณ์ จึงควรปรับปรุงที่ลักษณะการทำงาน ซึ่งแต่เดิมทำงานในเครื่องเดียวให้มีการทำงานในลักษณะการประมวลผลแบบขนาน หรือแบบช่วยกันประมวลผล เช่น ใช้ระบบ Grid Computing เป็นต้น จะทำให้ลดภาระในการประมวลผลที่เครื่องเดียวลงได้ เนื่องจากมีการแบ่งการทำงานออกเป็นส่วนย่อย ๆ และส่งไปทำงานในหลายเครื่อง จะทำให้มีความรวดเร็วและลดการใช้ทรัพยากรลงเป็นอย่างมาก

6.3.2 เนื่องจากดีเอ็นเอไมโครอาเรย์เป็นเทคนิคใหม่ที่มีเทคโนโลยีที่ใช้ในการผลิตที่แตกต่างกันมากมาย อีกทั้งมักจะมีสิ่งรบกวน (Noise) จำนวนมาก ดังนั้นในการนำข้อมูลดีเอ็นเอไมโครอาเรย์มาใช้ในการทดลองข้อมูลที่นำมาควรจะผลิตโดยใช้เทคโนโลยีเดียวกัน และเป็นเทคโนโลยีที่ได้มาตรฐานด้วย

บรรณานุกรม

กานดา พูนลาภทวี. 2539. สถิติเพื่อการวิจัย. พิมพ์ครั้งที่1. กรุงเทพฯ: ภาควิชาครุศาสตร์เทคโนโลยี คณะครุศาสตร์อุตสาหกรรมและวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

จรัญ จันทลักขณา และ อนันต์ชัย เขื่อนธรรม. 2540. สถิติเบื้องต้นแบบประยุกต์. พิมพ์ครั้งที่4. กรุงเทพฯ : ไทยวัฒนาพานิช.

จิราภรณ์ อังวิทยาธร และคณะ. 2544. ชีวสารสนเทศศาสตร์(Bioinformatics), สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ

ชาญชัยณรงค์ ทรงคาศรี. 2550. แอลฟาไม่ใช่ 0.05 ได้หรือไม่. แหล่งที่มา http://www.dmbn.net/mdbtemplate/mytemplate/template.php?component=view_article&read=1&qid=27&n_answer=6 . (16 มีนาคม 2551)

ธีรวรรธน์ ขันทอง. 2551. Hybridization. แหล่งที่มา http://www.student.chula.ac.th/~49371019/hybridization.htm (18 มีนาคม 2551)

นภดล สุชาติ. 2551. P-Value. แหล่งที่มา http://www.health.nu.ac.th/normal/intro3.htm. (16 มีนาคม 2551)

นฤพัฒน์ ตุลย์, ธันวา อาภรณ์ทิพย์ และ ตุลย์ ศิริกิจพุทธิศักดิ์. GMOs. แหล่งที่มา http://ecurriculum.mv.ac.th/library2/library/encyclopedia/content.php-encid=63.htm (15 มีนาคม 2551)

สมชาย แสงอำนาจเดช. 2550. ไมโครอาเรย์. แหล่งที่มา http://u-sabai-d.com/ Courses/ microA.html. (18 มีนาคม 2551)

สุรินทร์ นิยมางกูร. 2548. สถิติวิจัย. พิมพ์ครั้งที่ 2. กรุงเทพฯ. มหาวิทยาลัยเกษตรศาสตร์

Bartosh, K., Isaacson, L., Moritz, E., and Roberts, R. 2004. Application of Toxicogenomics in Toxicology. Available http://people.uwec.edu /piercech/ TOX/homepage.htm (18 March 2008)

Ben, D.A., Bruhn, L., Frieedman, N., Nachman, I., Schummer, M. and Yakhini, N. 2000. Tissue Classification with Gene Expression Profiles. Journal of Computational Biology, 2000, 7: 1-9

Brameier, M., and Wiuf, C. 2006. Co-Clustering and Visualization of gene expression Data and Gene Ontology Terms for Saccharomyces Cerevisiae using Self-Organizing Maps. Journal of Biomedical Informatics, 2006, 40: 160-173.

Brown, M.B., and Forsythe, A.B. 1974. The Small Sample Behavior of Some Statistics which Test the Equality of Mean. Technometrics, pp. 129-132.

Cardillo, G. 2006. Holm-Sidak t-test.Avaliable http://www.mathworks.com http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=12786&objectType=FILE. (18 March 2008)

Chen, D., and Hua, D. 2003. Gene Selection for Multi-Class Prediction of Microarray Data. IEEE Computer Society Bioinformatics Conference (CSB'03), pp. 492.

Cho, B.S., and Member. 2002. Classifying Gene Expression Data of Cancer Using Classifier Ensemble With Mutually Exclusive Feature. Proceedings of the IEEE, pp. 1744-1753.

Cochran, W.G. 1937. Problems Arising in the Analysis of a Series of Similar Experiments. Journal of the Royal Microscopical Society, 1937, 5: 102-118.

Dattta, S., and Datta, S. 2006. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics, 2006, 7: 1-9.

Dov, S. 2003. Microarray Bioinformatics. United States of America. British Library.

Gambhir,S.S. 2006. Statistics and Data Analysis. Available http://mips.stanford.edu/public/classes/stats_data_analysis/. (16 March 2008)

Girardin, L. Mapping the virtual geography of the World-Wide Web. Available http://www.girardin.org/luc/cgv/www5/index.html. (18 March 2008)

Jeng, T.J., Lee, T.T., and Lee, C.Y. 2005. Classification of ovarian cancer based on intelligent systems with microarray data. IEEE International Conference on Systems, Man and Cybernetics (IEEE SMC 2005), pp. 1053-1058.

Jiang, D., Tang, C., and Zhang, A. 2004. Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering, pp. 1370-1386.

Jiyuan A., and Phoebe C.Y. 2006. Finding Rule Group to Classify High Dimensional Gene Expressing Datasets. The 18th International Conference on Pattern Recognition (ICPR 2006), pp. 1196-1199.

Kaufman, L., and Rousseeuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.

Kim, C.J. et al. 1997. An algorithmic approach for fuzzy inference. IEEE Transactions on Fuzzy Systems, pp. 585 – 598.

Koua, L.E., and Kraak, J.M. 2004. Geovisualization to support the exploration of large health and demographic survey data. International Journal of Health Geographics, 2004, 4: 1526-2535.

Lalinka, C.T.G., Fermamd J.V.Z. and Pablo, M. 2004. A proposal for direct-ordering gene expression data by self-organizing maps. Applied Soft Computing, 2004, 5: 11-21.

Malcolm, A.C. DNA Microarray Methodology - Flash Animation. Availabel http://www.bio.davidson.edu/Courses/genomics/chip/chip.html. (18 March 2008)

Martinez, R., Pasquier, C., and Pasquier, N. 2007. GENMINER : Mining Informative Association Rules from Genomic Data, International Conference on Bioinformatics and Biomedicine (BIBM 2007), pp. 15-22.

Minsky, M. Kohonen's Self Organizing Feature Maps. Available http://www.ai-junkie.com/ann/som/som1.html. (18 March 2008)

Naenna, T. 2003. Data Mining Applications for Self-Organizing Maps. Master of Science Thesis Rensselaer Polytechnic Institute, New York

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. 1996. Applied Linear Statistical Models. McGraw – Hill.

Patra, C.J., Ang, L.E., Meher, K.P., and Zhen, Q. 2006. A New SOM-Based Visualization Technique for DNA Microarray Data. Proceedings of the 2006 International Joint Conference on Neural Networks (IJCNN 2006), pp. 4429-4434.

Petri T., Mikko K., Garry W., and Eero C. 1999. Analysis of gene expression data using self-organizing maps. Neural Network. 2002, 15: 142-146.

Sirirut, V., Siriphan, V., and Wiphada, W. 2007. Microarray Gene Selection Using Self-Organizing Map. The 7th WSEAS International Conference on Simulation Modeling and Optimization (SOM 2007), pp. 239-244.

Buchan, E.I. 2008. StatsDirect: StatsDirect Statistical Software Help. Available http://www.statsdirect.com/help/statsdirect.htm. (18 March 2008)

Sugiyama, A., and Kotani, M. 2002. Analysis of Gene Expression Data by using Self-organizing Maps and K-means Clustering. Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN 2002), pp. 1342–1345.

Tan, H.A., and Pan, T. 2005. Predictive Neural Network for Gene Expression Data Analysis. Neural Networks. 2005, 18: 297-306.

Vesanto, J., Himberg J., Alhoniemi E., and J. Parhankangas. 2000. SOM Toolbox for Matlab5. Libella Oy Espoo.

Wei, W., Xin, L., Min, Xu., Jinrong, P., and Setiono, R. 2004. Hybrid SOM-SVM Method for Analyzing Zebra Fish Gene Expression. International Conference on Pattern Regcognition(ICPR 2004), pp. 323-326.

Welch, B.L. 1951. On the Comparison of Several Mean Values: An Alternative Approach. Biometrika, pp. 330-336.

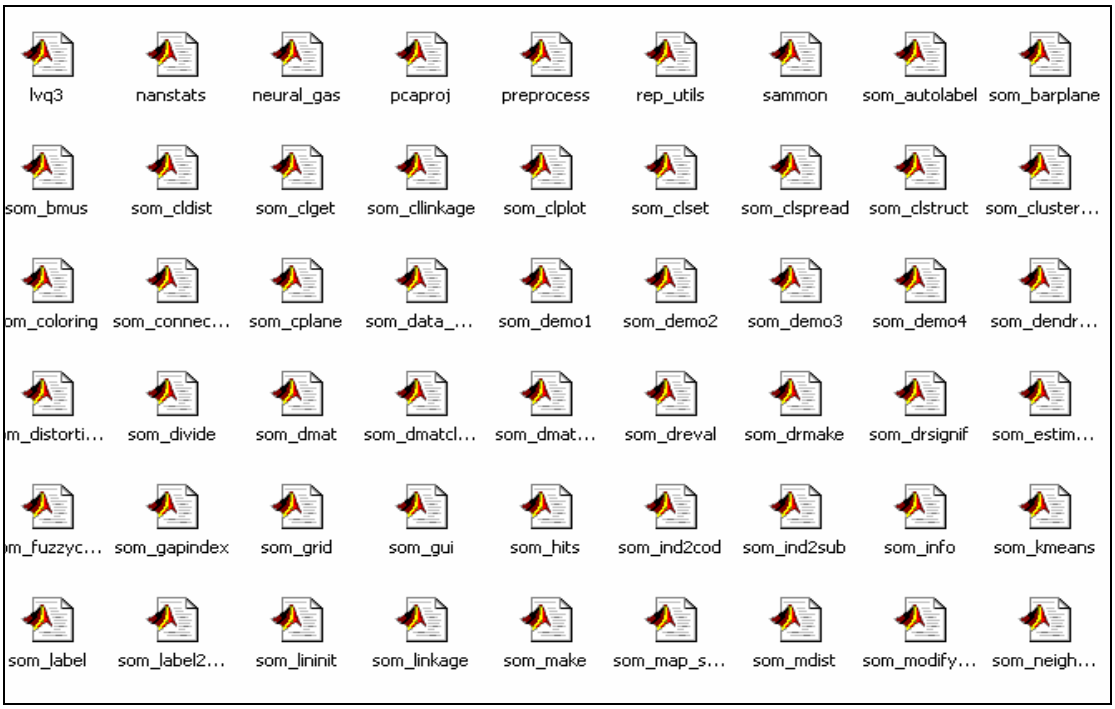
Wiphada, W., and Putthiporn, N. 2006. Knowledge Extraction from Self-Organizing Map Using Minimization Entropy Principle Algorithm. International Symposium on Communications and Information Technologies 2006, pp. 37–42.

Yano, N., and Kotani, M. 2003. Clustering Gene Expression Data Using Self-Organizing Maps and k-means Clustering. Society of Instrument and Control Engineers Annual Conference 2003(SICE 2003), pp. 3211-3215.

ภาคผนวก ก.

การใช้งานแผนที่การจัดกลุ่มเองในโปรแกรม MATLAB

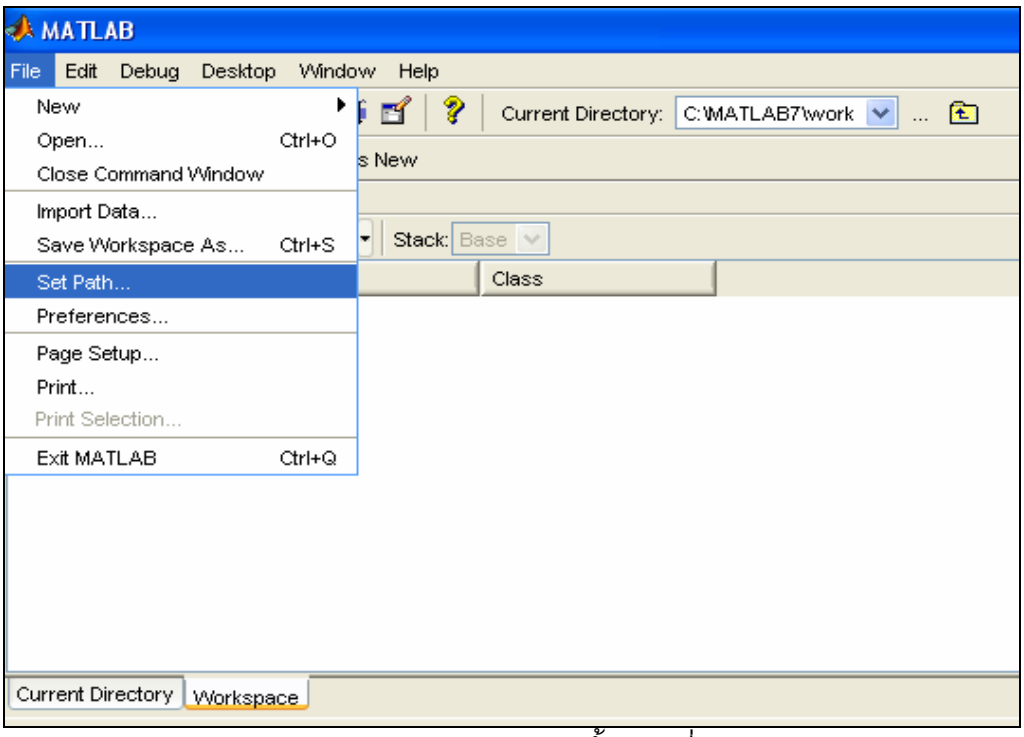
การจะใช้งานแผนที่การจัดกลุ่มเองในโปรแกรม MATLAB จะต้องมีการเพิ่มชุดฟังก์ชันของแผนที่การจัดกลุ่มเองก่อนจึงจะสามารถทำงานได้ ซึ่งชุดของฟังก์ชันของแผนที่การจัดกลุ่มเองที่กล่าวถึงนี้เรียกว่า Self-Organizing Map Toolbox (SOM Toolbox) ซึ่งสามารถ Download ได้ที่เว็บไซต์ของ MATLAB ซึ่งรูปแบบของ SOM Toolbox แสดงได้ดังภาพประกอบ ก.1



ภาพประกอบ ก.1 แสดงรูปแบบของ SOM Toolbox

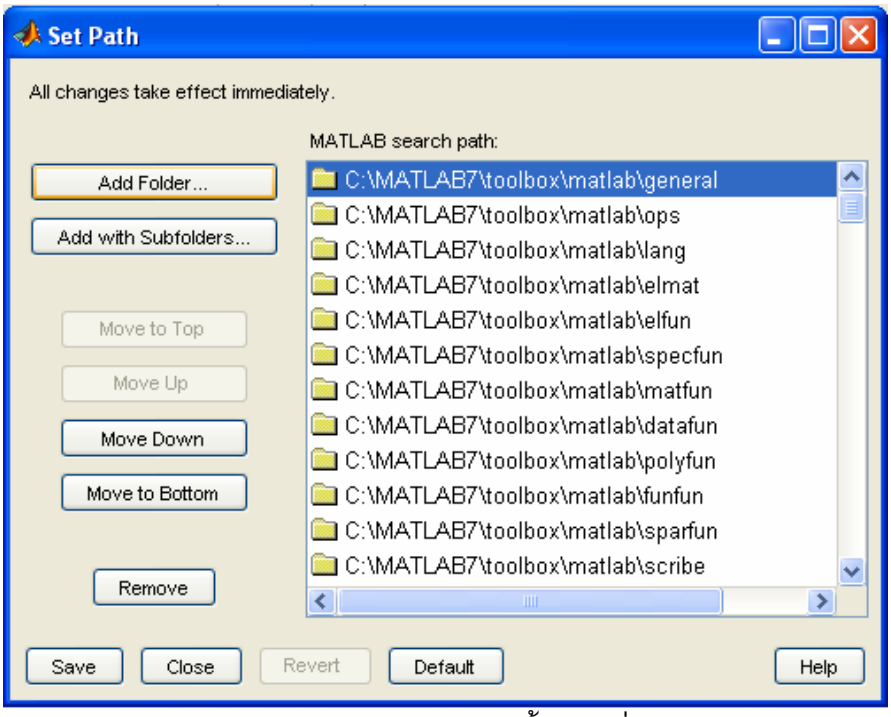
จากภาพประกอบ ก.1 จะเห็นว่าภายใน SOM Toolbox จะประกอบด้วยฟังก์ชันต่าง ๆ ที่ใช้ในการทำงานกับ SOM ที่สามารถเรียกใช้ได้ในขั้นตอนการเขียนโปรแกรม ส่วนขั้นตอนการตั้งค่าอื่นจะอธิบายโดยมีภาพประกอบเป็นขั้นตอน ดังนี้

1. เมื่อ Download SOM Toolbox มาแล้วให้ย้ายไฟล์นั้นไปที่ Path ดังนี้ “C:\MATLAB7\toolbox”
2. เมื่อเข้าโปรแกรม MATLAB ให้เลือกเมนู File → Set Path ดังภาพประกอบ ก.2



ภาพประกอบ ก.2 แสดงขั้นตอนที่ 2

3. จะปรากฏหน้าจอใหม่ดังภาพประกอบ ก.3 ให้เลือกปุ่ม Add Folder …

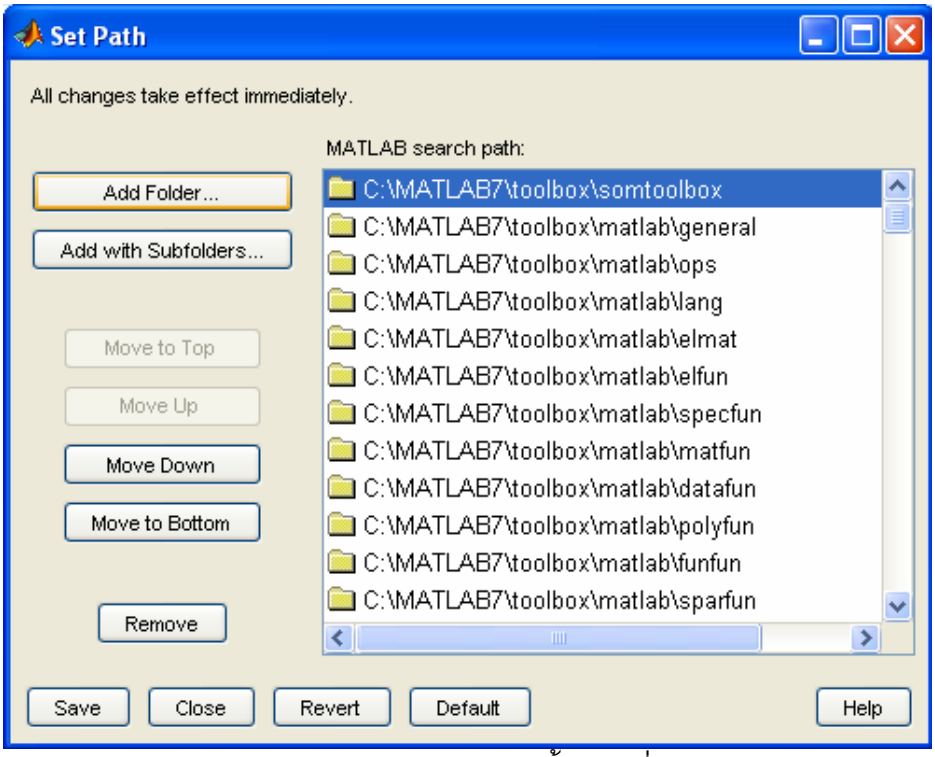


ภาพประกอบ ก.3 แสดงขั้นตอนที่ 3

4. จะปรากฏหน้าจอดังภาพประกอบ ก.4 ให้เลือก somtoolbox ที่อยู่ใน Path ในขั้นตอนที่ 1

ภาพประกอบ ก.4 แสดงขั้นตอนที่ 4

5. กดปุ่ม OK และกดปุ่ม Save ดังภาพประกอบ ก.5



ภาพประกอบ ก.5 แสดงขั้นตอนที่ 5

เมื่อทำครบขั้นตอนที่ 5 แล้วถือว่าเสร็จขั้นตอนการตั้งค่าโปรแกม MATLAB เพื่อให้ใช้งาน SOM ได้แล้ว ผู้ใช้จึงจะสามารถเขียนโปรแกรมโดยใช้คำสั่งของ SOM ได้ รวมถึงจะสามารถใช้งานแบบจำลองการสกัดความรู้จากข้อมูลดีเอ็นเอไมโครอาเรย์โดยใช้แผนที่การจัดกลุ่มเองได้ด้วย

ภาคผนวก ข.

ผลงานตีพิมพ์

เรื่อง	Microarray Gene Selection Using Self-Organizing Map for Cancer Prediction
งานประชุมวิชาการ	The 6th Joint Conference on Computer Science and Software Engineering (JCSSE 2007)
สถานที่	จังหวัดขอนแก่น ประเทศไทย
วันที่	2-4 พฤษภาคม 2550

ภาคผนวก ค.

ผลงานตีพิมพ์

เรื่อง	Microarray Gene Selection Using Self-Organizing Map
งานประชุมวิชาการ	The 7th WSEAS International Conference on Simulation Modeling and Optimization (SOM 2007)
สถานที่	กรุงปักกิ่ง ประเทศจีน
วันที่	15-17 กันยายน 2550

ประวัติผู้เขียน

ชื่อ สกุล นางสาวศิริพันธุ์ วิชัยดิษฐ

รหัสประจำตัวนักศึกษา 4822074

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วท.บ. (วิทยาการคอมพิวเตอร์)	มหาวิทยาลัยทักษิณ	2547

การตีพิมพ์เผยแพร่ผลงาน

1. Vanichayobon, S., Wichaidit, S., and Wettayaprasit, W. 2007. Microarray Gene Selection Using Self-Organizing Map. The 7th WSEAS International Conference on Simulation Modeling and Optimization (SOM 2007), Beijing, Chaina, pp. 239-244.

2. Wichaidit, S., Vanichayobon, S., and Wettayaprasit, W. 2007. Microarray Gene Selection Using Self-Organizing Map for Cancer Prediction. The 6th Joint Conference on Computer Science and Software Engineering (JCSSE 2007), Khon Kaen, Thailand, pp. 335-343.