กลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์

นายสุพจน์ เฮงพระพรหม

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2551

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

ENSEMBLE GENETIC PROGRAMMING CLASSIFIER FOR MICROARRAY DATA

Mr. Supoj Hengpraprohm

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic year 2008

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	กลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์
โดย	นายสุพจน์ เฮงพระพรหม
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.นวลวรรณ สุนทรภิษัช)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สุพจน์ เฮงพระพรหม : กลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์. (Ensemble Genetic Programming Classifier for Microarray Data) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา, 77 หน้า.

วิทยานิพนธ์นี้ได้นำเสนอวิธีการสร้างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ ซึ่งเป็นข้อมูลที่มีจำนวนข้อมูลน้อย ขณะที่จำนวนคุณลักษณะมีจำนวนมาก ในการสร้างสมาชิกของกลุ่มก้อนนั้น จะมุ่งสร้างตัวจำแนกประเภทที่มีประสิทธิภาพในการจำแนกประเภทข้อมูลที่ดี ในขณะที่ตัวจำแนกประเภทแต่ละตัวจะต้องมีความแตกต่างจากสมาชิกตัวอื่น ๆ ในกลุ่มก้อน วิธีการที่นำเสนอจะใช้เทคนิคการจัดกลุ่มข้อมูลแบบ K-Means เพื่อจัดกลุ่มของคุณลักษณะของข้อมูลที่มีลักษณะคล้ายกันให้อยู่ในกลุ่มเดียวกัน และการเลือกคุณลักษณะแบบ SNR (Signal-to-Noise Ratio) โดยจะนำคุณลักษณะที่มีค่า SNR สูงที่สุดลำดับที่ i ของแต่ละกลุ่ม มาสร้างเป็นเซตย่อยของคุณลักษณะเพื่อใช้ในการสอน เพื่อสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมตัวที่ i ซึ่งวิธีการนี้สามารถสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมที่มีประสิทธิภาพที่ดี และมีความแตกต่างจากตัวจำแนกประเภทตัวอื่น ๆ เนื่องจากการใช้คุณลักษณะที่แตกต่างกัน ทำให้ประสิทธิภาพของกลุ่มก้อนดีขึ้นตามไปด้วย

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชาวิศวกรรมคอมพิวเตอร์ ..... ลายมือชื่อนิสิต .....

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2551

4771832221 : MAJOR COMPUTER ENGINEERING

KEYWORDS : CLASSIFICATION / GENETIC PROGRAMMING / ENSEMBLE METHOD / MICROARRAY DATA ANALYSIS / FEATURE SELECTION

SUPOJ HENGPRAPROHM : ENSEMBLE GENETIC PROGRAMMING CLASSIFIER FOR MICROARRAY DATA. ADVISOR : PROFESSOR PRABHAS CHONGSTITVATANA, Ph.D., 77 pp.

This thesis presents an algorithm for generating an ensemble of Genetic Programming classifiers for microarray data. The number of data is small and it has high dimensions. In order to construct an ensemble, each classifier must have high efficiency and at the same time it must be different from other classifiers. The proposed method uses K-Means clustering for grouping the features of data which are similar into the same group. The SNR (Signal-to-Noise Ratio) feature selection is used to select informative features. The feature with the i^{th} best SNR score in each group is selected to form a set of features. This feature set is used to train the i^{th} Genetic Programming classifier. The proposed method creates a good Genetic Programming classifier where each classifier is different from the others. They contain different set of features. As a result, the performance of the ensemble is improved.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Department : COMPUTER ENGINEERING Student's Signature .....

Field of Study: COMPUTER ENGINEERING Advisor's Signature .....

Academic Year : 2008.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากความช่วยเหลือ การให้คำแนะนำต่าง ๆ ที่เป็นประโยชน์และกำลังใจที่ดี จาก ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา อาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้เขียนรู้สึกซาบซึ้งในความอนุเคราะห์จากท่าน และขอกราบขอบพระคุณท่านเป็นอย่างสูง

ขอกราบขอบพระคุณ รองศาสตราจารย์ นพ.อดุลย์ รัตนวิจิตราศิลป์ หัวหน้าสถานวิทยามะเร็งศิริราช คณะแพทยศาสตร์ศิริราชพยาบาล ที่ได้ให้ความรู้และข้อคิดเห็นต่าง ๆ เกี่ยวกับโรคมะเร็ง และ ดร.นิศรา การุณอุทัยศิริ ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (BIOTEC) ที่ได้ให้ความรู้และข้อคิดเห็นต่าง ๆ เกี่ยวกับเทคโนโลยีไมโครอาร์เรย์

ขอขอบคุณ อาจารย์ ดร.สุวิมล มรรควิบูลชัย โปรแกรมวิชาเทคโนโลยีคอมพิวเตอร์ และอาจารย์ ชลิดา ตระกูลสุนทร โปรแกรมวิชาสถิติ มหาวิทยาลัยราชภัฏนครปฐม ซึ่งได้ให้คำแนะนำเกี่ยวกับสถิติและข้อคิดเห็นที่เป็นประโยชน์ในการวิเคราะห์ข้อมูล

ขอขอบคุณ คุณไก้รุ่ง เฮงพระพรหม ภรรยา ที่คอยให้กำลังใจ ช่วยเหลือและให้การสนับสนุนในทุก ๆ เรื่อง ในระหว่างการศึกษาและการจัดทำวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงได้ด้วยดี ขอกราบขอบพระคุณ คุณพ่อหลุงชิ้น คุณแม่เสงี่ยม เฮงพระพรหม (บิดา-มารดา) และ คุณพ่อต่อน คุณแม่ลม พ่อกุล (บิดา-มารดาของภรรยา) ที่คอยเป็นกำลังใจในการศึกษาและการจัดทำวิทยานิพนธ์ฉบับนี้ด้วยดีตลอดมา

สุดท้าย ขอขอบคุณ สำนักงานคณะกรรมการการอุดมศึกษา (สกอ.) และมหาวิทยาลัยราชภัฏนครปฐม ที่ได้ให้การสนับสนุน ทั้งทางด้านเงินทุนการศึกษา เวลา รวมถึงวัสดุอุปกรณ์ต่างๆ ที่จำเป็นสำหรับการศึกษาและการจัดทำวิทยานิพนธ์ฉบับนี้ จนสำเร็จลุล่วงไปได้ด้วยดี

สารบัญ

หน้า

หน้า

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

หน้า

หน้า

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

หน้า

หน้า

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เทคโนโลยีไมโครอาร์เรย์ (Microarray) ช่วยให้เราสามารถศึกษารูปแบบของสิ่งมีชีวิตในระดับโมเลกุล โดยสามารถศึกษารูปแบบการแสดงออกของยีน (Gene Expression) ได้หลายพันยีนในเวลาเดียวกัน ซึ่งเป็นเทคนิคที่ได้รับความนิยมเพื่อใช้ศึกษาหน้าที่การทำงานของยีนต่าง ๆ ในสิ่งมีชีวิตในสาขาวิชาชีวสารสนเทศศาสตร์ (Bioinformatics) เช่น การศึกษาการเปลี่ยนแปลงของยีนของสิ่งมีชีวิตหนึ่ง ๆ เมื่อเวลาหรือสภาพแวดล้อมเปลี่ยนไป หรือ การศึกษารูปแบบความแตกต่างของยีนจากกลุ่มของสิ่งมีชีวิตที่มีลักษณะที่แตกต่างกัน ตัวอย่างเช่น สภาวะการเกิดโรคโดยเฉพาะโรคมะเร็ง ซึ่งปัจจุบันยังไม่มีวิธีการวินิจฉัยและการรักษาที่มีประสิทธิภาพ

ดังนั้นจึงมีหลายงานวิจัยได้ทำการศึกษาโรคมะเร็งชนิดต่าง ๆ ด้วยข้อมูลไมโครอาร์เรย์ เช่น มะเร็งต่อมน้ำเหลือง (Lymphoma) (Alizadeh et al., 2000) มะเร็งเม็ดเลือดขาว (Leukemia) (Golub et al., 1999) มะเร็งเต้านม (Breast Cancer) (Van't Veer et al., 2002) เนื้องอกระบบประสาทส่วนกลาง (Central Nervous System Embryonal Tumors: CNS) (Pomeroy et al., 2002) มะเร็งลำไส้ใหญ่ (Colon Cancer) (Alon et al., 1999) มะเร็งรังไข่ (Ovarian Cancer) (Petricoin III et al., 2002) มะเร็งต่อมลูกหมาก (Prostate Cancer) (Singh et al., 2002) และมะเร็งปอด (Lung Cancer) (Gordon et al., 2002) เป็นต้น เพื่อหาวิธีการวินิจฉัยโรคและวิธีการรักษาที่มีประสิทธิภาพต่อไป

จากการศึกษาดังกล่าว ก่อให้เกิดข้อมูลไมโครอาร์เรย์จำนวนมาก ซึ่งข้อมูลต่าง ๆ เหล่านี้ได้มีนักวิจัยหลายคนนำมาวิเคราะห์โดยใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) และเทคนิคทางสถิติ เพื่อทำความเข้าใจและหาแบบจำลองสำหรับระบบการวินิจฉัยโรคแบบอัตโนมัติด้วยคอมพิวเตอร์

เมื่อไม่นานมานี้ วิธีการแบบกลุ่มก้อน (Ensemble) ของตัวจำแนกประเภท (Classifier) ได้รับความนิยมนำมาใช้อย่างกว้างขวาง เพื่อเพิ่มประสิทธิภาพและความน่าเชื่อถือของระบบการจำแนกประเภทข้อมูล เนื่องจากพื้นฐานของแนวความคิดที่ว่า ไม่มีตัวจำแนกประเภทเดี่ยวใด ๆ ที่ให้ผลถูกต้องในทุก ๆ สภาพการณ์ ซึ่งผลที่ได้ก็พบว่า ความแม่นของระบบการจำแนกประเภทเพิ่มสูงขึ้นอย่างมีนัยสำคัญ ตัวอย่างของการใช้กลุ่มก้อนของตัวจำแนกประเภท เช่น การใช้กลุ่มก้อน

ของต้นไม้ตัดสินใจ (Decision Tree) โดยใช้เทคนิค Bagging (Breiman, 1996) และ AdaBoost (Freund and Schapire, 1996) ในการสร้างความหลากหลายของต้นไม้ (Tan and Gilbert, 2003) การใช้เทคนิคต่าง ๆ ที่หลากหลายในการสร้างกลุ่มก้อนของตัวจำแนกประเภท และการเลือกคุณลักษณะ (Feature Selection) เพื่อสร้างความหลากหลายของตัวจำแนกประเภท (Cho and Won, 2003) รวมถึง การใช้ตัวจำแนกประเภทกำหนดการพันธุกรรม (Genetic Programming) หลายตัว (Hengpraprohm and Chongstitvatana, 2005) เป็นต้น

กำหนดการพันธุกรรม (Koza, 1992) เป็นขั้นตอนวิธี (Algorithm) หนึ่งในกลุ่มการคำนวณเชิงวิวัฒน์ (Evolutionary Computation) ที่ได้รับความนิยมอย่างแพร่หลายในการนำมาใช้เพื่อแก้ปัญหาในด้านต่าง ๆ มากมาย เช่น ปัญหาการควบคุมหุ่นยนต์ (Robot Control) (Hengpraprohm and Chongstitvatna, 2001) ปัญหาการออกแบบฮาร์ดแวร์ (Hardware Design) (Sakanashi et al., 1996) รวมถึงปัญหาหลักทางด้านการทำเหมืองข้อมูล อันได้แก่ ปัญหาการจำแนกประเภทข้อมูล (Data Classification)

มีงานวิจัยหลายชิ้นที่ได้ศึกษาการสร้างตัวจำแนกประเภท โดยใช้เทคนิคกำหนดการพันธุกรรม ในหลากหลายเขต (Domain) ของปัญหา เช่น ชุดข้อมูลเกณฑ์เปรียบเทียบสมรรถนะ (Benchmark) ของกลุ่มการเรียนรู้ของเครื่อง หรือชุดข้อมูลทางการแพทย์ ที่นิยมใช้ในการศึกษาทางด้านชีวสารสนเทศศาสตร์ ซึ่งได้แก่ข้อมูลไมโครอาร์เรย์ โดยตัวจำแนกประเภทนี้จะมีการแทน (Representation) หลากหลายรูปแบบ เช่น ต้นไม้ตัดสินใจ นิพจน์ทางตรรกศาสตร์ (Logical Expression) และ นิพจน์ทางคณิตศาสตร์ (Arithmetic Expression) เป็นต้น ซึ่งในหลายงานวิจัยก็ได้รายงานตรงกันว่าตัวจำแนกประเภทกำหนดการพันธุกรรมให้ประสิทธิภาพของการจำแนกประเภทข้อมูลที่ดี

เนื่องจากข้อมูลไมโครอาร์เรย์เป็นข้อมูลที่มีมิติ (Dimension) สูงมาก ขณะที่จำนวนตัวอย่างมีจำกัด ทำให้เทคนิคการสร้างกลุ่มก้อนด้วยวิธี Bagging หรือ AdaBoost ซึ่งต้องการจำนวนตัวอย่างที่มากพอ หรือวิธีการใช้ความหลากหลายภายในกฎของตัวจำแนกประเภทเองในการสร้างกลุ่มก้อน ซึ่งไม่ได้ใช้ประโยชน์จากชุดข้อมูลสอนมากนัก ทำให้ประสิทธิภาพของกลุ่มก้อนในการจำแนกประเภทข้อมูลนั้น ยังไม่ดีเท่าที่ควร

งานวิจัยนี้จึงต้องการศึกษาหาวิธีการสร้างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์ที่มีประสิทธิภาพ โดยจะนำเอาลักษณะของข้อมูลมาใช้ประกอบ ซึ่งจะใช้เทคนิคทางการจัดกลุ่มข้อมูล (Clustering) และการเลือกคุณลักษณะ เพื่อสร้าง

ความหลากหลายของตัวจำแนกประเภทแต่ละตัวในการสร้างสมาชิกของกลุ่มก้อน เพื่อเพิ่มประสิทธิภาพของการจำแนกประเภทข้อมูลให้ดียิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเสนอขั้นตอนวิธีใหม่ที่เหมาะสำหรับการสร้างกลุ่มก้อนของตัวจำแนกประเภท กำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์ โดยใช้ข้อดีจากการจัดกลุ่มข้อมูลและการเลือกคุณลักษณะในการปรับปรุงประสิทธิภาพ

1.3 ขอบเขตของการวิจัย

1. จะทำการศึกษาวิธีการจำแนกประเภทข้อมูลที่มี 2 ประเภทเท่านั้น

2. จะทำการทดสอบกับชุดข้อมูลไมโครอาร์เรย์เกณฑ์เปรียบเทียบสมรรถนะจากเว็บไซต์ Bio-medical Data Analysis (http://sdmc.lit.org.sg/GEDatasets/) จำนวน 8 ชุดข้อมูล ได้แก่

- กลุ่มย่อยของโรคมะเร็งเม็ดเลือดขาว ประกอบด้วยข้อมูล 72 ตัวอย่าง แบ่งเป็น ALL จำนวน 47 ตัวอย่าง และ AML จำนวน 25 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 7,129 คุณลักษณะ (Golub et al., 1999)
- ผลการรักษาโรคมะเร็งเต้านม ประกอบด้วยข้อมูล 78 ตัวอย่าง แบ่งเป็น relapses จำนวน 34 ตัวอย่าง และ non-relapses จำนวน 44 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 24,481 คุณลักษณะ (Van't Veer et al., 2002)
- ผลการรักษาเนื้องอกระบบประสาทส่วนกลาง ประกอบด้วยข้อมูล 60 ตัวอย่าง แบ่งเป็น survivors จำนวน 21 ตัวอย่าง และ failures จำนวน 39 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 7,129 คุณลักษณะ (Pomeroy et al., 2002)
- ผู้ป่วยมะเร็งลำไส้ใหญ่กับผู้ป่วยปกติ ประกอบด้วยข้อมูล 62 ตัวอย่าง แบ่งเป็น ผู้ป่วยมะเร็ง จำนวน 40 ตัวอย่าง และ ผู้ป่วยปกติ จำนวน 22 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 2,000 คุณลักษณะ (Alon et al., 1999)

- ผู้ป่วยมะเร็งรังไข่กับผู้ป่วยปกติ ประกอบด้วยข้อมูล 253 ตัวอย่าง แบ่งเป็น ผู้ป่วยมะเร็ง จำนวน 162 ตัวอย่าง และ ผู้ป่วยปกติ จำนวน 91 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 15,154 คุณลักษณะ (Petricoin III et al., 2002)

- ผู้ป่วยมะเร็งต่อมลูกหมากกับผู้ป่วยปกติ ประกอบด้วยข้อมูล 102 ตัวอย่าง แบ่งเป็น ผู้ป่วยมะเร็ง จำนวน 52 ตัวอย่าง และ ผู้ป่วยปกติ จำนวน 50 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 12,600 คุณลักษณะ (Singh et al., 2002)

- กลุ่มย่อยของโรคมะเร็งปอด ประกอบด้วยข้อมูล 181 ตัวอย่าง แบ่งเป็น MPM จำนวน 31 ตัวอย่าง และ ADCA จำนวน 150 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 12,533 คุณลักษณะ (Gordon et al., 2002)

- กลุ่มย่อยของโรคมะเร็งต่อมน้ำเหลือง ประกอบด้วยข้อมูล 47 ตัวอย่าง แบ่งเป็น germinal centre B-like จำนวน 24 ตัวอย่าง และ activated B-like จำนวน 23 ตัวอย่าง จากจำนวนคุณลักษณะทั้งหมด 4,026 คุณลักษณะ (Alizadeh et al., 2000)

3. จะทำการศึกษาวิธีการจัดกลุ่มข้อมูลและการเลือกคุณลักษณะแบบพื้นฐาน ได้แก่ การจัดกลุ่มข้อมูลแบบ K-Means และการเลือกคุณลักษณะด้วยวิธี SNR

4. การเปรียบเทียบประสิทธิภาพ จะเปรียบเทียบประสิทธิภาพระหว่างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรมที่เสนอกับวิธีกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรมแบบอื่น ๆ ได้แก่

- แบบปกติ คือ การเลือกคุณลักษณะแบบสุ่มจากคุณลักษณะของข้อมูลที่มีทั้งหมด
- แบบเลือกคุณลักษณะด้วยวิธี SNR
- แบบใช้เทคนิค Bagging และ AdaBoost
- ผลจากงานที่ตีพิมพ์ในวารสารวิชาการ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ขั้นตอนวิธีใหม่สำหรับการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมที่มีประสิทธิภาพในการจำแนกประเภทข้อมูลไมโครอาร์เรย์

2. ได้แนวทางสำหรับการสร้างกลุ่มก้อนของตัวจำแนกประเภทแบบอื่น ๆ

1.5 วิธีดำเนินการวิจัย

1. การศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ขั้นตอนนี้ จะทำการศึกษาทฤษฎีต่าง ๆ ที่จะใช้ในการทดลอง รวมถึงงานวิจัยต่าง ๆ ที่เกี่ยวข้อง ซึ่งได้แก่ เทคโนโลยีไมโครอาร์เรย์ การวิเคราะห์ข้อมูลไมโครอาร์เรย์ด้วยเทคนิคต่าง ๆ ทางการทำเหมืองข้อมูลและการเรียนรู้ของเครื่อง ได้แก่ การเลือกคุณลักษณะ การจำแนกประเภทข้อมูล และการจัดกลุ่มข้อมูล รวมถึงเทคนิคที่ใช้ในการสร้างกลุ่มก้อนของตัวจำแนกประเภท

2. การวิเคราะห์และออกแบบการทดลอง

หลังจากศึกษาข้อมูลที่เกี่ยวข้องแล้ว จะทำการรวบรวมข้อมูลไมโครอาร์เรย์ เพื่อนำมาใช้ในการทดลอง โดยจะคัดลอกมาจากเว็บไซต์ Bio-medical Data Analysis (http://sdmc.lit.org.sg/GEDatasets/) ซึ่งเป็นแหล่งรวบรวมชุดข้อมูลไมโครอาร์เรย์เกณฑ์เปรียบเทียบสมรรถนะสำหรับการวิเคราะห์ข้อมูลไมโครอาร์เรย์ จากนั้นจะนำข้อมูลดังกล่าวมาทำการศึกษาถึงผลกระทบจากเทคนิควิธีการจัดกลุ่มข้อมูล และการเลือกคุณลักษณะที่สำคัญ เพื่อนำมาใช้สร้างสมาชิกของกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรม

3 การพัฒนาและทดสอบประสิทธิภาพ

หลังจากออกแบบการทดลองเสร็จ จะทำการพัฒนาโปรแกรมตามรูปแบบที่ได้ออกแบบไว้ และทำการเปรียบเทียบประสิทธิภาพกับวิธีการสร้างกลุ่มก้อนแบบต่าง ๆ ที่ได้อธิบายไว้ในหัวข้อขอบเขตของงานวิจัย โดยใช้วิธีการวัดอัตราความผิดพลาดแบบ k-folds cross validation การเปรียบเทียบประสิทธิภาพจะเปรียบเทียบในรูปของความแม่น (Accuracy) ความไว (Sensitivity) และ ความจำเพาะ (Specificity) ของการจำแนกประเภทข้อมูลจากชุดข้อมูลทดสอบ

4. การวิเคราะห์ผลและสรุปผล

หลังจากทำการทดลองและทดสอบประสิทธิภาพด้วยวิธีการต่าง ๆ กับข้อมูลที่ได้กล่าวไว้ข้างต้นทั้งหมดแล้ว จะนำผลการทดลองที่ได้มาทำการวิเคราะห์ผล สรุปผล และให้ข้อสังเกตและข้อเสนอแนะเกี่ยวกับการทดลอง

5 การจัดทำเอกสารงานวิจัย

เมื่อได้ข้อมูลการทดลองครบถ้วนสมบูรณ์แล้ว จะทำการเรียบเรียงเอกสารงานวิจัย เพื่อตีพิมพ์เผยแพร่ในวารสารวิชาการระดับนานาชาติ จัดทำเอกสารวิทยานิพนธ์ และสอบป้องกัน (Defense) ในที่สุด

1.6 ลำดับขั้นตอนในการนำเสนอผลการวิจัย

เนื้อหาในวิทยานิพนธ์ฉบับนี้ แบ่งออกเป็น 2 ส่วนหลัก ๆ ด้วยกัน ได้แก่

ส่วนที่ 1 จะเป็นเนื้อหาเกี่ยวกับทฤษฎีที่เกี่ยวข้องกับการวิจัย ซึ่งได้อธิบายไว้ในบทที่ 2 เนื้อหาในส่วนนี้ประกอบด้วยข้อมูลที่เกี่ยวกับไมโครอาร์เรย์ ซึ่งเป็นข้อมูลที่จะใช้ในการทดลอง รายละเอียดเกี่ยวกับกำหนดการพันธุกรรม ซึ่งเป็นขั้นตอนวิธีการเรียนรู้หลักสำหรับการสร้างตัวจำแนกประเภทในงานวิจัยนี้ และจะได้กล่าวถึงเรื่องทั่วไปเกี่ยวกับการจำแนกประเภทข้อมูล นอกจากนี้ จะได้กล่าวถึงทฤษฎีที่เกี่ยวข้องสำหรับการปรับปรุงประสิทธิภาพที่ใช้ในการวิจัย ได้แก่ การจัดกลุ่มข้อมูล การเลือกคุณลักษณะ และรายละเอียดทั่วไปเกี่ยวกับวิธีการแบบกลุ่มก้อน

ส่วนที่ 2 จะเป็นเนื้อหาที่เกี่ยวกับการออกแบบการทดลองและผลการทดลอง ซึ่งจะแบ่งออกเป็น 4 บท ได้แก่ บทที่ 3 จะแสดงวิธีการสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ บทที่ 4 จะแสดงวิธีการเพิ่มประสิทธิภาพของตัวจำแนกประเภทกำหนดการพันธุกรรมด้วยเทคนิคการเลือกคุณลักษณะ และสุดท้ายจะนำเสนอวิธีการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ ซึ่งได้แสดงรายละเอียดในบทที่ 5 ส่วนในบทที่ 6 จะเป็นการสรุปผลการวิจัยและข้อเสนอแนะ

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

1.7 ผลงานตีพิมพ์

บทความที่ได้รับการตีพิมพ์ในงานประชุมวิชาการระดับนานาชาติ

1. Hengpraprohm S. and Chongstitvatana P., "Diffuse Large B-Cell Lymphoma Classification Using Genetic Programming Classifier", 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, November 14-15, 2005, La Jolla, CA, USA, pp. 333-338.

2. Hengpraprohm S. and Chongstitvatana P., "Discovering an Optimal Feature Set of Microarray Data for Cancer Classification Using Perceptron Learning Rule with SNR Ranking", International conference on Software Knowledge Information Management and Applications, December 12-15, 2006, Chiang Mai, Thailand, pp. 159-164.

3. Hengpraprohm S. and Chongstitvatana P., " Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier Using K-Means Clustering and SNR Ranking", Frontiers in the Convergence of Bioscience and Information Technologies, October 11-13, 2007, Jeju Island, Korea, pp. 211-218.

4. Hengpraprohm S. and Chongstitvatana P., "A Genetic Programming Ensemble Approach to Cancer Microarray Data Classification", The 3th International Conference on Innovative Computing, Information and Control, June 18- 20, 2008, Dalian, China. pp.340.

บทความที่ได้รับการตอบรับให้ตีพิมพ์ในวารสารวิชาการระดับนานาชาติ

1. Hengpraprohm S. and Chongstitvatana P., "Feature Selection by Weighted-SNR for Cancer Microarray Data Classification", International Journal of Innovative Computing Information and Control, Vol. 5, No. 12, December, 2009.

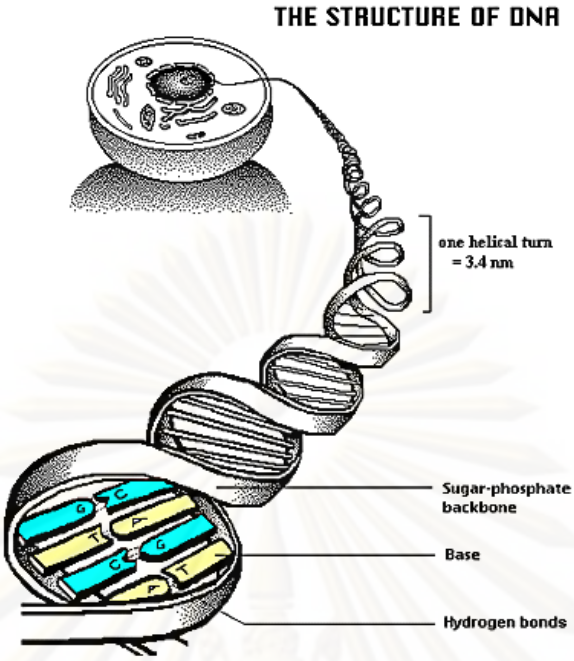
บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ไมโครอาร์เรย์ (Microarray)

ไมโครอาร์เรย์ เป็นเทคนิคที่ใช้ในการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลาย ๆ ยีนในเวลาเดียวกัน เพื่อให้เข้าใจกลไกการทำงานของสิ่งมีชีวิตในระดับโมเลกุล การแสดงออกของยีน หมายถึง ความสามารถในการถอดรหัส (Transcription) ข้อมูลทางพันธุกรรมที่บรรจุไว้ใน DNA ภายในนิวเคลียส (Nucleus) ให้กลายเป็น mRNA ซึ่งจะถูกแปลรหัส (Translation) เพื่อสร้างเป็นโปรตีนที่ทำหน้าที่ต่าง ๆ ภายในเซลล์ (Cell)

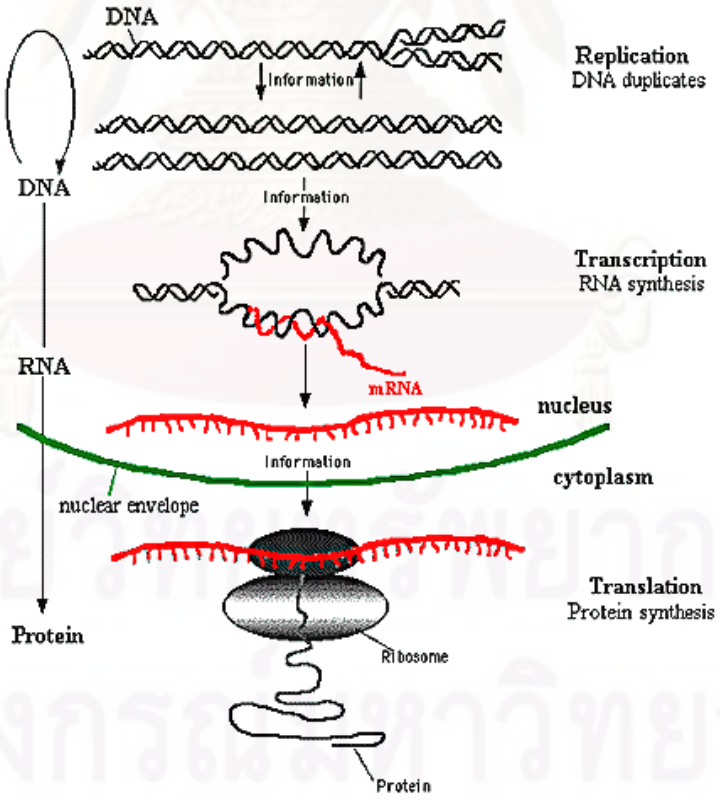
DNA เป็นสารพันธุกรรมซึ่งมีโครงสร้างเป็นสารประเภทเบส (Base) 4 ชนิด ประกอบด้วย adenine (A), cytosine (C), guanine (G) และ thymine (T) เรียงต่อกันเป็นสายยาว ซึ่งตามปรกติสาย DNA จะมีอยู่ 2 สายพันกันเป็นเกลียวคู่ โดยเบสแต่ละตัวจะจับกับคู่ของมันอย่างเฉพาะเจาะจง กล่าวคือ A จะจับคู่กับ T และ C จะจับคู่กับ G เสมอ โครงสร้างของ DNA แสดงดังรูปที่ 2.1 เมื่อ DNA ถูกถอดรหัสมาเป็น RNA ซึ่งมีโครงสร้างคล้ายคลึงกับ DNA แต่จะมีอยู่เพียงสายเดียวไม่จับตัวกันเป็นเกลียวคู่ สาย RNA นี้จะถูกแปลรหัสเพื่อสร้างเป็นกรดอะมิโน (Amino Acids) ซึ่งมีอยู่ทั้งหมด 20 ชนิด โดยลำดับเบสของ RNA 3 ตัว จะรวมกันเรียกว่า โคดอน (Codon) เพื่อสร้างเป็นกรดอะมิโนขึ้นมา 1 ชนิด ดังนั้นสาย RNA 1 สาย ก็จะทำการสร้างชุดลำดับของกรดอะมิโนขึ้นมา 1 ชุด ซึ่งประกอบกันเป็นโปรตีน ภาพรวมของกระบวนการสังเคราะห์โปรตีนจาก DNA แสดงดังรูปที่ 2.2

ข้อมูลไมโครอาร์เรย์ถูกสร้างมาจากการจับคู่กัน (Hybridization) ของชิ้นส่วน DNA ตัวอย่างที่ต้องการศึกษาโดยทำการติดฉลากด้วยสารเรืองแสง (Fluorescent) สีแดง (แทนด้วย Cy5) กับชิ้นส่วน DNA ควบคุมที่ติดฉลากด้วยสารเรืองแสงสีเขียว (แทนด้วย Cy3) ในปริมาณที่เท่ากันจัดเรียงเป็นแถวเป็นแนว (Array) บนแผ่นสไลด์ด้วยเครื่องจักรกล (ตัวอย่างเครื่องจักรกลแสดงดังรูปที่ 2.3) จากนั้นจะทำการวัดปริมาณของสารเรืองแสงแต่ละสีด้วยเครื่องสแกน (ตัวอย่างเครื่องสแกนและผลลัพธ์แสดงดังรูปที่ 2.4)



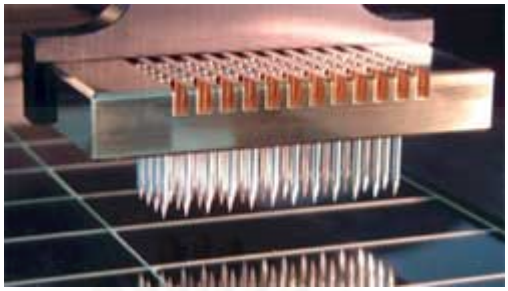
รูปที่ 2.1 โครงสร้าง DNA

(ที่มา: ยงยุทธ ยุทธวงศ์ และคณะ, 2545 : 3)

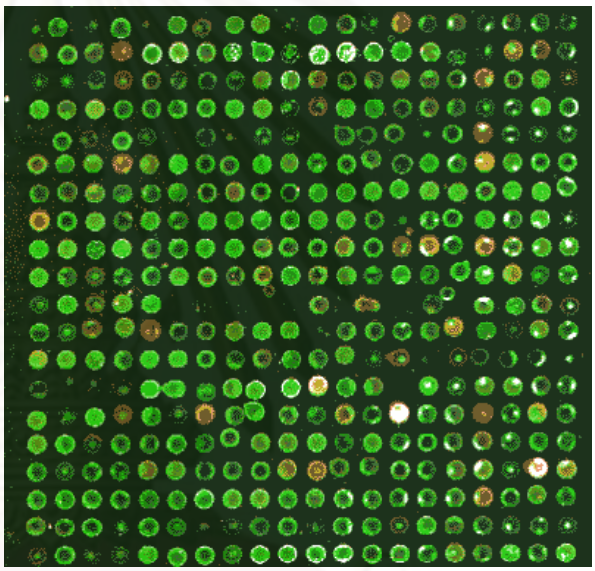


รูปที่ 2.2 หลักการกลาง (Central Dogma) ของชีววิทยาระดับโมเลกุล

(ที่มา: ยงยุทธ ยุทธวงศ์ และคณะ, 2545 : 5)



รูปที่ 2.3 เครื่อง Stealth Printhead (SPH48) พร้อมด้วยหัวพิมพ์ 48 หัว
(ที่มา: http://www.euro-bio-net.com)



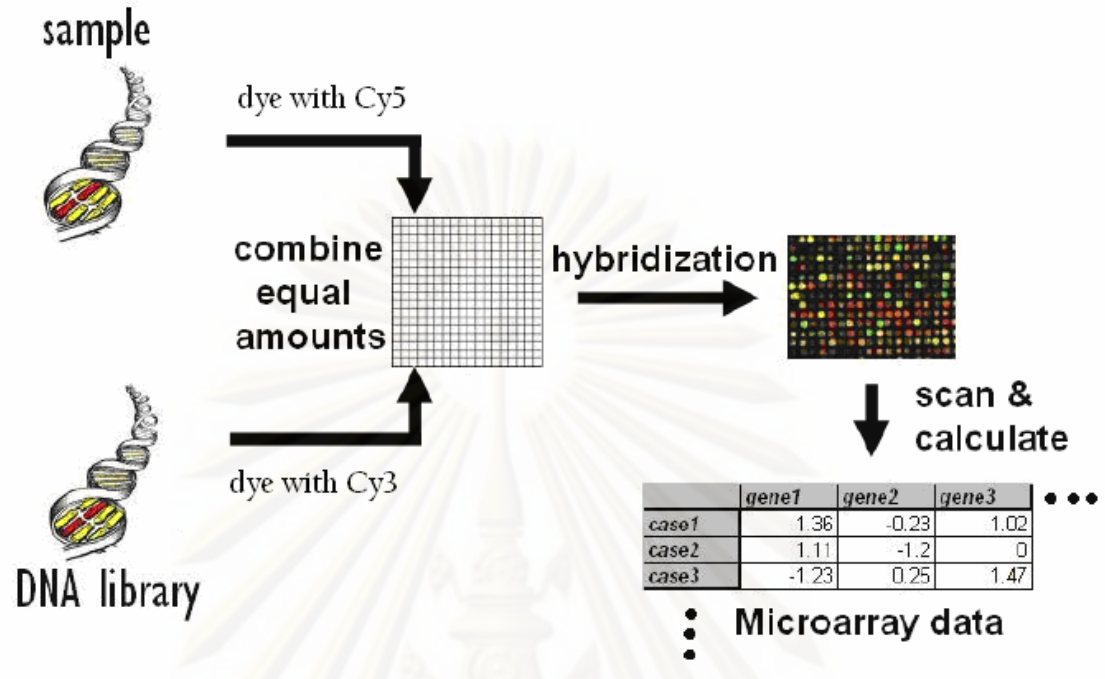
รูปที่ 2.4 (ซ้าย) เครื่อง InnoScan 700 AL (ที่มา: http://www.innopsys.fr)
(ขวา) ผลลัพธ์ที่ได้จากการสแกนด้วยเครื่องสแกนไมโครอาร์เรย์

ผลลัพธ์ที่ได้จากการสแกนจะถูกนำมาคำนวณเพื่อใช้เป็นค่าการแสดงออกของยีนด้วยวิธีการบางอย่าง วิธีการที่นิยมใช้ได้แก่การหาอัตราส่วนล็อก (log ratio) ซึ่งคำนวณได้ดังนี้

$$\textit{ค่าการแสดงออกของยีน} = \log_2 \frac{Int(Cy5)}{Int(Cy3)} \qquad (2.1)$$

โดยที่ Int(Cy5) และ Int(Cy3) คือ ปริมาณความเข้มข้นของสารเรืองแสงสีแดงและสารเรืองแสงสีเขียวหลังจากการสแกนด้วยเครื่องสแกนไมโครอาร์เรย์ของแต่ละจุด

ภาพรวมของกระบวนการสร้างข้อมูลไมโครอาร์เรย์แสดงดังรูปที่ 2.5

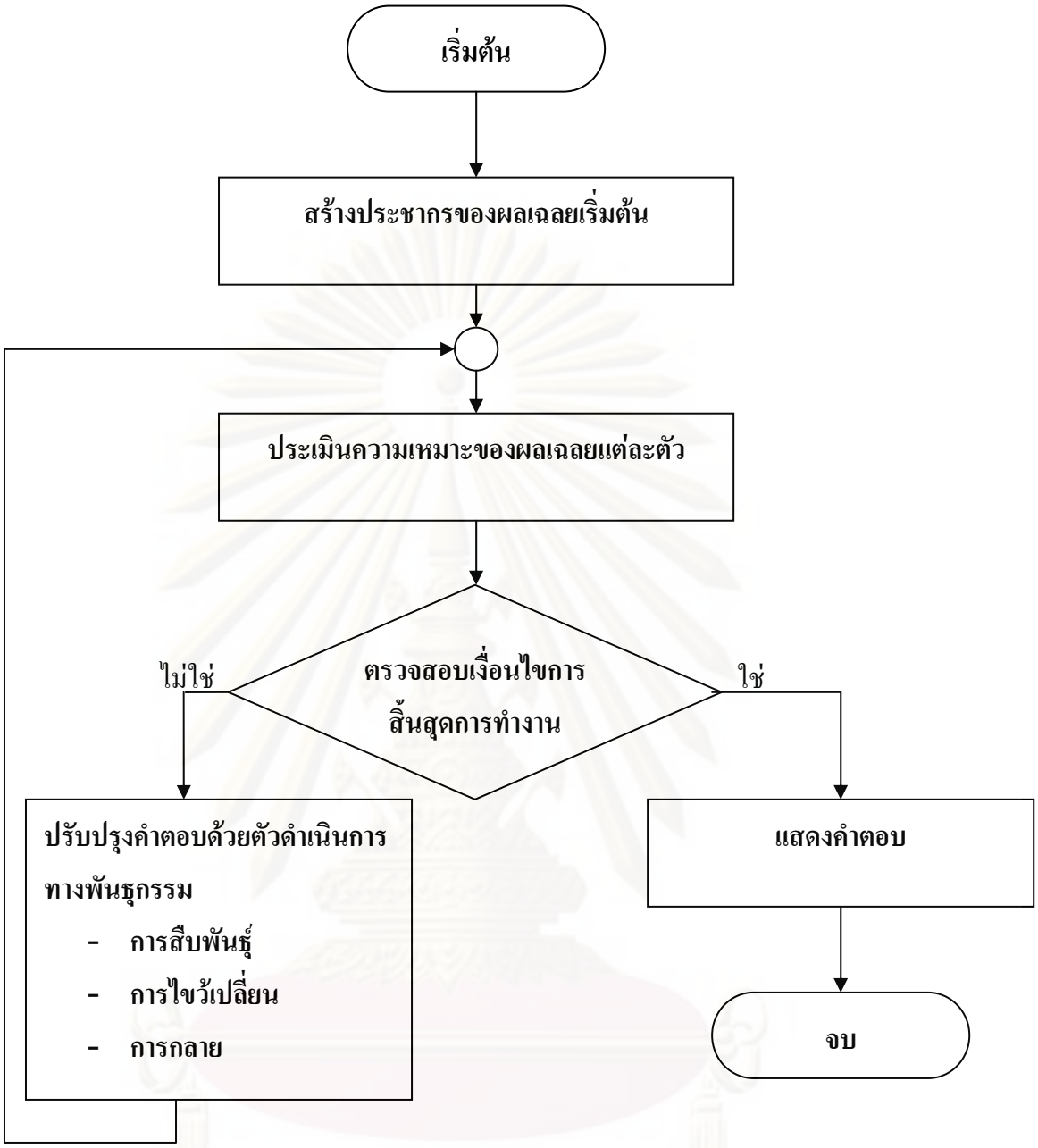


รูปที่ 2.5 ภาพรวมของกระบวนการสร้างข้อมูลไมโครอาร์เรย์

2.2 กำหนดการพันธุกรรม (Genetic Programming)

กำหนดการพันธุกรรม เป็นวิธีการค้นหาคำตอบที่คิดค้นโดย Koza (1992) ซึ่งถูกพัฒนามาจากขั้นตอนวิธีพันธุกรรม (Genetic Algorithm: GA) ที่คิดค้นโดย Holland (1975) โดยจำลองแบบมาจากกระบวนการวิวัฒนาการของสิ่งมีชีวิต และกฎการคัดเลือกโดยธรรมชาติ ข้อแตกต่างระหว่างขั้นตอนวิธีพันธุกรรมและกำหนดการพันธุกรรม คือ ลักษณะการแทนคำตอบ โดยขั้นตอนวิธีพันธุกรรมจะแทนคำตอบอยู่ในรูปสายอักขระ (String) ที่มีขนาดคงที่ ในขณะที่กำหนดการพันธุกรรมจะแทนคำตอบอยู่ในรูปโปรแกรมคอมพิวเตอร์ ซึ่งแทนด้วยโครงสร้างต้นไม้ ทำให้ขนาดของคำตอบมีความยืดหยุ่นมากขึ้น

ขั้นตอนการค้นหาคำตอบด้วยวิธีกำหนดการพันธุกรรม จะแบ่งออกเป็นขั้นตอนหลัก ๆ คือ การสร้างประชากรของผลเฉลยเริ่มต้น การประเมินค่าความเหมาะ (Fitness Value) ของผลเฉลย การสร้างประชากรของผลเฉลยรุ่นใหม่ และการหาคำตอบ ซึ่งแสดงในรูปของผังงาน (Flowchart) ดังรูปที่ 2.6 และรายละเอียดของแต่ละขั้นตอนเป็นดังนี้



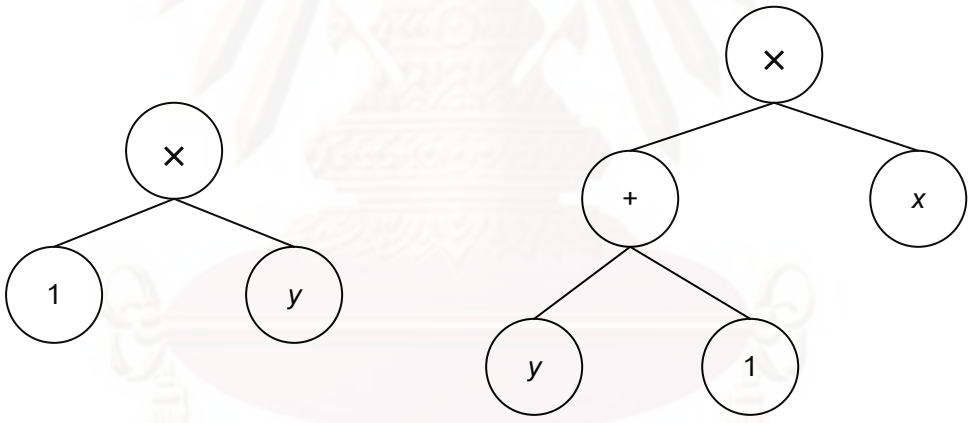
รูปที่ 2.6 ผังงานของกำหนดการพันธุกรรม

2.2.1 การสร้างประชากรของผลเฉลยเริ่มต้น

การสร้างประชากรของผลเฉลยเริ่มต้น เป็นการสร้างผลเฉลย (Solution) เริ่มต้นที่เป็นไปได้แบบสุ่ม โดยแต่ละผลเฉลยจะประกอบด้วยฟังก์ชัน (Function) และ เทอร์มินอล (Terminal) ฟังก์ชัน หมายถึง ฟังก์ชันต่าง ๆ ที่ใช้กับเทอร์มินอลในผลเฉลย ซึ่งอาจจะเป็นฟังก์ชัน ทางคณิตศาสตร์ เช่น บวก ลบ SIN COS ฟังก์ชันทางตรรกศาสตร์ เช่น OR AND IF-OR IF-AND หรือ

เป็นฟังก์ชันที่นิยามขึ้นสำหรับปัญหานั้น ๆ ส่วน เทอร์มินอล หมายถึง เซตของปัจจัยที่เป็นอิสระแก่กันสำหรับปัญหานั้น ๆ อาจเป็นค่าคงที่ ตัวแปร หรือคำสั่งที่มีผลกระทบโดยตรงกับปัญหานั้น ตัวอย่างเช่น คำสั่งควบคุม หรือ คำสั่งในการตรวจรู้ (Sensing) เป็นต้น

ตัวอย่างเช่น หากกำหนดให้ เซตของฟังก์ชัน แทนด้วย F และ เซตของเทอร์มินอล แทนด้วย T โดยกำหนดให้ $F = \{+, -, \times, \div\}$ และ $T = \{0..9, x, y\}$ ดังนั้น การสร้างผลเฉลยเริ่มต้น จะทำการสุ่มเลือกฟังก์ชันจากเซตของฟังก์ชัน และทำการสุ่มเลือกเทอร์มินอลจากเซตของเทอร์มินอลมาใส่เป็นส่วนของอาร์กิวเมนต์ (Argument) ให้กับฟังก์ชันที่สุ่มได้ตามจำนวนอาร์กิวเมนต์ที่ฟังก์ชันนั้นต้องการ เช่น หากสุ่มได้ × (คูณ) จากเซตของฟังก์ชัน และ 1 กับ y จากเซตของเทอร์มินอล จะได้โครงสร้างของผลเฉลย แสดงดังรูปที่ 2.7 (ก) จากนั้นจะทำการสุ่มตำแหน่งโนดใบ (Leaf Node) และสุ่มใส่ฟังก์ชันและเทอร์มินอลไปเรื่อย ๆ จนได้ขนาดของ ผลเฉลยเริ่มต้นตามที่ต้องการ โดยอาจกำหนดขนาดของผลเฉลยเป็นจำนวนโนด หรือเป็น ความสูงหรือความลึกของต้นไม้ก็ได้ รูปที่ 2.7 (ข) แสดงผลเฉลยที่มีขนาด 5 โนด ที่แทนนิพจน์เชิงสัญลักษณ์ (Symbolic Expression) ของ $(y+1)x$



(ก) โครงสร้างผลเฉลยของ 1y (ข) โครงสร้างผลเฉลยขนาด 5 โนด ของ (y+1)x

รูปที่ 2.7 โครงสร้างต้นไม้ของผลเฉลย

จากขั้นตอนนี้จะได้กลุ่มของผลเฉลยจำนวนหนึ่งตามจำนวนประชากร (Population) ที่ได้กำหนดไว้ ซึ่งจะมีลักษณะที่แตกต่างกันไปตามการสุ่ม

2.2.2 การประเมินค่าความเหมาะของผลเฉลย

ในขั้นตอนนี้ จะเป็นการวัดค่าความเหมาะ (Fitness Value) ของผลเฉลยแต่ละตัวโดยใช้ฟังก์ชันความเหมาะ (Fitness Function) ที่ถูกกำหนดขึ้นมาตามความเหมาะสมกับแต่ละปัญหา เพื่อหาผลเฉลยที่ดีที่จะใช้เป็นคำตอบ หรือนำไปสร้างประชากรของผลเฉลยรุ่นใหม่ต่อไป การนิยามฟังก์ชันความเหมาะเป็นส่วนที่ยากและสำคัญสำหรับวิธีกำหนดการพันธุกรรม

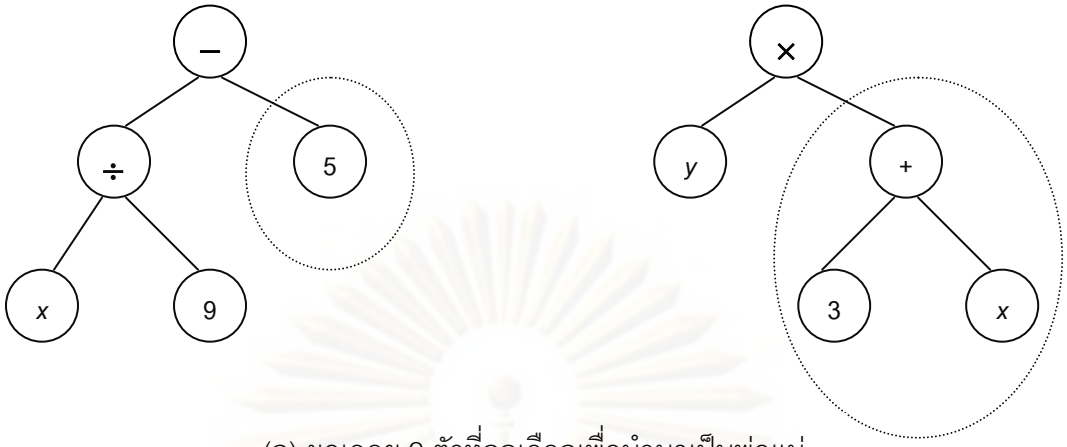
ค่าความเหมาะนี้จะใช้เป็นเกณฑ์ในการตัดสินว่าผลเฉลยตัวใดมีความสามารถในการแก้ปัญหาและสมควรที่จะถูกคัดเลือกให้อยู่รอดและขยายพันธุ์ต่อไป โดยการประเมินค่าความเหมาะของผลเฉลยนี้ จะนำผลเฉลยแต่ละตัวไปทดสอบแก้ปัญหาที่กำหนด แล้วใช้ฟังก์ชันความเหมาะวัดค่าความเหมาะของผลเฉลยเหล่านั้น ซึ่งจะได้ค่าความเหมาะหรือประสิทธิภาพในการแก้ปัญหาของผลเฉลยแต่ละตัว

2.2.3 การสร้างประชากรของผลเฉลยรุ่นใหม่

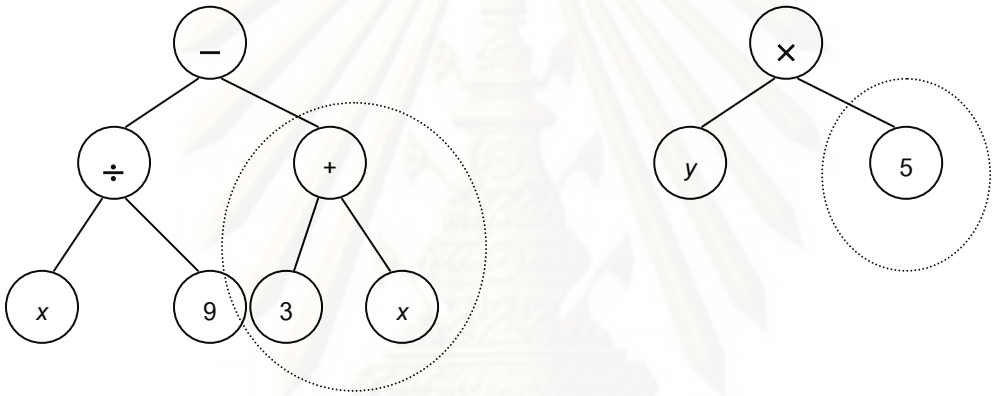
หลังจากทำการประเมินค่าความเหมาะของผลเฉลยแล้ว ก็จะทำการคัดเลือกผลเฉลยตามความเหมาะเพื่อนำมาสร้างประชากรของผลเฉลยรุ่นใหม่ โดยอาศัยตัวดำเนินการทางพันธุกรรม (Genetic Operator) ซึ่งได้แก่ การสืบพันธุ์ (Reproduction) การไขว้เปลี่ยน (Crossover) และการกลาย (Mutation) ซึ่งมีรายละเอียดดังต่อไปนี้

การสืบพันธุ์: เป็นการคัดลอกผลเฉลยที่มีค่าความเหมาะสูงจากประชากรของผลเฉลยรุ่นเดิมมาสร้างเป็นประชากรของผลเฉลยรุ่นใหม่ โดยไม่มีการเปลี่ยนแปลงโครงสร้างใด ๆ ของผลเฉลยนั้น

การไขว้เปลี่ยน: เป็นการสร้างประชากรของผลเฉลยรุ่นใหม่ โดยการสุ่มเลือกผลเฉลยจากประชากรของผลเฉลยรุ่นเดิมมาครั้งละ 2 ตัว ตามความเหมาะ เพื่อนำมาเป็นพ่อแม่ (Parent) จากนั้นจะทำการสุ่มตำแหน่งที่จะใช้สำหรับการไขว้เปลี่ยนของพ่อแม่ และทำการสลับโครงสร้าง ณ ตำแหน่งจุดที่สุ่มได้ ซึ่งจะได้ลูก (Children หรือ Offspring) จำนวน 2 ตัว เป็นประชากรของผลเฉลยรุ่นใหม่ รูปที่ 2.8 (ก) แสดงถึงผลเฉลย 2 ตัว ที่ถูกคัดเลือกเพื่อนำมาเป็นพ่อแม่ เส้นประแสดงถึงตำแหน่งที่สุ่มได้ที่จะใช้ในการไขว้เปลี่ยน และรูปที่ 2.8 (ข) แสดงถึงผลเฉลยลูกที่เกิดจากการไขว้เปลี่ยน



(ก) ผลเฉลย 2 ตัวที่ถูกเลือกเพื่อนำมาเป็นพ่อแม่



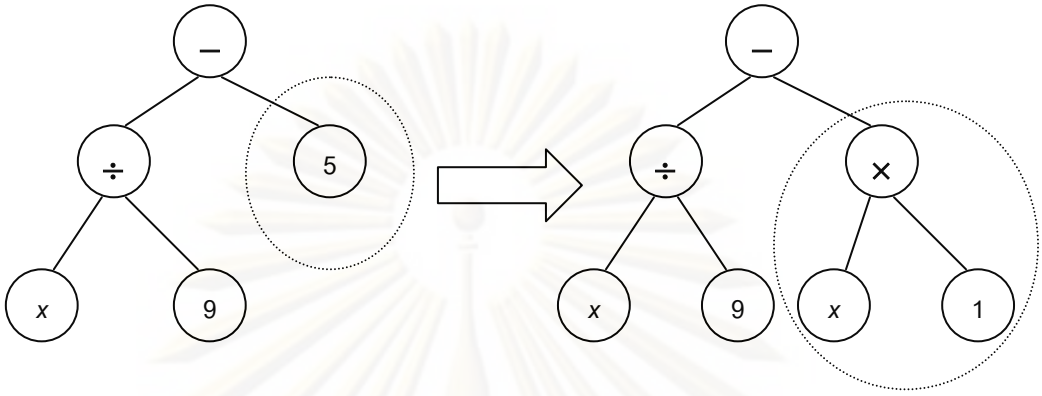
(ข) ผลเฉลยที่เกิดจากการไขว้เปลี่ยน

รูปที่ 2.8 วิธีการไขว้เปลี่ยน

การกลาย: เป็นการสร้างประชากรของผลเฉลยรุ่นใหม่ โดยการสุ่มเลือกผลเฉลยจากประชากรของผลเฉลยรุ่นเดิม มาครั้งละ 1 ตัว ตามความเหมาะ และทำการสุ่มตำแหน่งที่จะทำการกลาย จากนั้นจึงทำการเปลี่ยนแปลงโครงสร้าง ณ ตำแหน่งที่สุ่ม ด้วยโครงสร้างอื่นที่สร้างขึ้นมาแบบสุ่มเช่นกัน รูปที่ 2.9 (ซ้าย) แสดงผลเฉลยที่ถูกคัดเลือกมาเพื่อทำการกลาย และ (ขวา) แสดงผลเฉลยที่เกิดจากการกลาย เส้นประแสดงตำแหน่งที่สุ่มได้ที่จะทำการกลาย

ในขั้นตอนการสร้างประชากรของผลเฉลยรุ่นใหม่นี้ จะทำการสร้างประชากรของผลเฉลยรุ่นใหม่ ด้วยตัวดำเนินการทางพันธุกรรม โดยจะใช้อัตรา (Rate) ตามที่กำหนด เช่น ใช้อัตราการสืบพันธุ์ 10% อัตราการไขว้เปลี่ยน 89% และอัตราการกลาย 1% เป็นต้น จนได้จำนวน ประชากร

ครบตามที่ต้องการ จากนั้นก็จะกลับไปทำในขั้นการประเมินค่าความเหมาะของผลเฉลย และการสร้างประชากรของผลเฉลยรุ่นใหม่ซ้ำต่อไปเรื่อย ๆ จนกระทั่งพบคำตอบ หรือครบตามจำนวนรุ่น (Generation) ที่กำหนด



รูปที่ 2.9 วิธีการกลาย

2.2.4 **การหาคำตอบ**

กำหนดการพันธุกรรมส่วนใหญ่จะใช้กับปัญหาที่มีได้หลายคำตอบ เพราะฉะนั้นคำตอบที่ได้จากกำหนดการพันธุกรรม จะเป็นผลเฉลยที่ให้ค่าความเหมาะดีที่สุด หลังผ่านกระบวนการสร้างประชากรของผลเฉลยรุ่นใหม่จนครบตามจำนวนรุ่นที่กำหนด

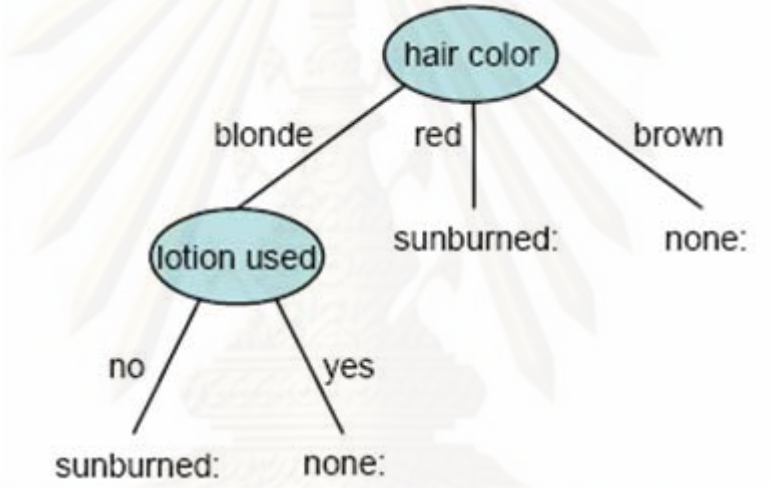
2.3 **การจำแนกประเภทข้อมูล** (Data Classification)

การจำแนกประเภทข้อมูล เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยวัตถุประสงค์เพื่อกำหนดประเภทข้อมูลให้กับข้อมูลใหม่ ๆ ที่ยังไม่รู้ประเภท โดยใช้ลักษณะของชุดข้อมูลที่มีอยู่ซึ่งเป็นชุดข้อมูลที่รู้ประเภทแล้ว เทคนิคพื้นฐานที่นิยมนำมาใช้สำหรับการจำแนกประเภทข้อมูล ได้แก่

2.3.1 **ต้นไม้ตัดสินใจ** (Decision Tree)

ต้นไม้ตัดสินใจ (Quinlan 1986) เป็นตัวแบบทางคณิตศาสตร์ที่ใช้จำแนกประเภทข้อมูลโดยพิจารณาจากคุณลักษณะของข้อมูล ซึ่งค่าของคุณลักษณะเหล่านี้จะอยู่ในรูปของค่าที่ไม่ต่อเนื่อง (Discrete Value) เช่น เพศ คณะวิชา สีผม เป็นต้น

ในการเรียนรู้จะแทนความรู้ในรูปแบบโครงสร้างต้นไม้ โดยมีโนดภายใน (Inner Node) เป็นชื่อของคุณลักษณะต่าง ๆ ของชุดข้อมูลที่นำมาเรียนรู้ แต่ละโนดภายในต้นไม้จะมีกิ่ง (Branch) เท่ากับจำนวนค่าที่เป็นไปได้สำหรับคุณลักษณะนั้น ๆ ส่วนโนดใบจะเป็นประเภทที่เป็นไปได้ของข้อมูล ตัวอย่างต้นไม้ตัดสินใจ แสดงดังรูปที่ 2.10 ซึ่งจากรูปแสดงตัวอย่างต้นไม้ตัดสินใจสำหรับจำแนกประเภทผลลัพธ์ของการอาบแดด โดยคุณลักษณะที่ใช้มี 2 คุณลักษณะ คือ สีผม (hair color) ซึ่งมีค่าที่เป็นไปได้ 3 ค่า (blonde, red และ brown) และ การใช้โลชัน (lotion used) ซึ่งมีค่าที่เป็นไปได้ 2 ค่า (yes และ no) ส่วนผลลัพธ์ของการจำแนกประเภทจะประกอบด้วย 2 ประเภท คือ ผิวไหม้ (sunburned) และ ไม่เป็นอะไร (none)



รูปที่ 2.10 ตัวอย่างต้นไม้ตัดสินใจ
(ที่มา: บุญเสริม กิจสิริกุล, 2546 : 154)

จากรูปที่ 2.10 จะแทนกฎการจำแนกประเภท ดังนี้

IF <hair color> = blonde **and** <lotion used> = no **THEN** sunburned
IF <hair color> = blonde **and** <lotion used> = yes **THEN** none
IF <hair color> = red **THEN** sunburned
IF <hair color> = brown **THEN** none

ขั้นตอนการเรียนรู้สำหรับต้นไม้ตัดสินใจ แสดงดังรูปที่ 2.11

***ต้นไม้ตัดสินใจ** (Decision Tree)*

***ข้อมูลนำเข้า**: ชุดข้อมูลสอน x(a_1, a_2, …, a_n: class)*

1. *ทดสอบคุณลักษณะของข้อมูลด้วยค่า Information Gain (IG) ซึ่งคำนวณได้จาก*

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

โดยที่ $Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$

S *คือ เซตของข้อมูลทั้งหมดที่ใช้ในการทดสอบ*

S_v *คือ เซตย่อยของ S ที่คุณลักษณะ A ของข้อมูลมีค่าเท่ากับ v*

Values(A) คือ ค่าที่เป็นไปได้ของคุณลักษณะ A

Entropy(S) คือ ค่าความไร้ระเบียบของข้อมูลในเซต S

c *คือ ประเภท (class) ของข้อมูลที่เป็นไปได้*

p_i *คือ ค่าความน่าจะเป็นของข้อมูลที่จะอยู่ในประเภท i*

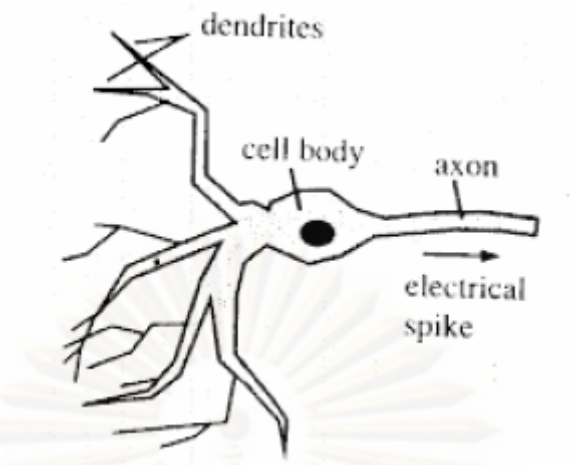
2. *เลือกคุณลักษณะที่ให้ค่า IG สูงที่สุดมาสร้างเป็นโนดราก (Root Node)*
3. *ขณะที่ Entropy(S_v) ของกิ่งใด ๆ ที่ยังไม่เท่ากับ 0 ให้นำคุณลักษณะที่เหลือมาทดสอบและเลือกคุณลักษณะที่ให้ค่า IG สูงที่สุดมาแตกเป็นกิ่งของต้นไม้ต่อไปเรื่อย ๆ*

***ผลลัพธ์**: ต้นไม้ตัดสินใจ*

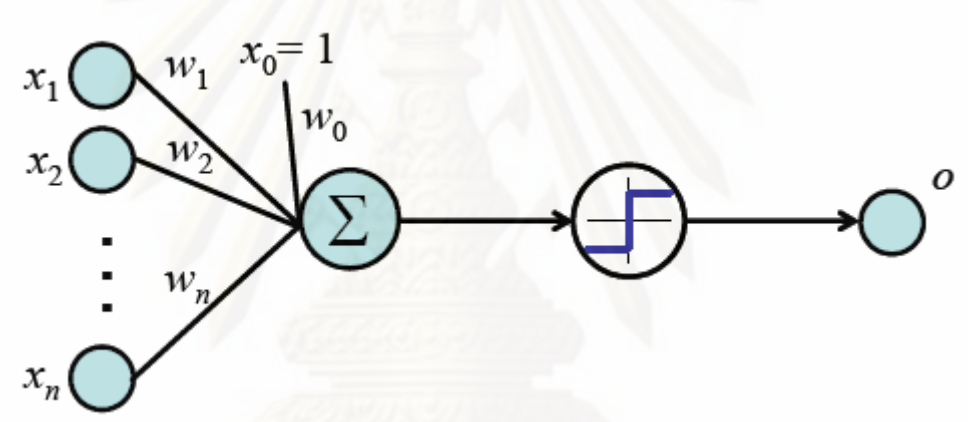
รูปที่ 2.11 ขั้นตอนการเรียนรู้ของต้นไม้ตัดสินใจ

2.3.2 **เครือข่ายใยประสาทเทียม** (Artificial Neural Networks)

เครือข่ายใยประสาทเทียม (Bishop, 1995) เป็นการจำลองการทำงานของเซลล์ประสาทในสมองของมนุษย์ ซึ่งประกอบด้วยนิวเคลียส ตัวเซลล์ ใยประสาทนำเข้า (dendrite) และแกนประสาทนำออก (Axon) ซึ่งแสดงดังรูปที่ 2.12 การจำลองการคำนวณด้วยคอมพิวเตอร์จะแทนให้อยู่ในรูป Perceptron ซึ่งแสดงดังรูปที่ 2.13



รูปที่ 2.12 โครงสร้างเซลล์ประสาท
(ที่มา: บุญเสริม กิจสิริกุล, 2546 : 169)



รูปที่ 2.13 โครงสร้าง Perceptron
(ที่มา: บุญเสริม กิจสิริกุล, 2546 : 169)

การเรียนรู้ของ Perceptron จะเริ่มจากการอ่านข้อมูลสอนเข้ามาทีละตัว จากนั้นจะทำการหาค่าผลรวมของคุณลักษณะต่าง ๆ (x_i) ถ่วงด้วยน้ำหนัก (w_i) ซึ่งถ้าค่าผลรวมนั้นมากกว่าค่าขีดแบ่ง (threshold) ที่กำหนด จะให้ผลลัพธ์ที่ได้เป็น 1 แต่ถ้าไม่ใช่จะกำหนดให้ผลลัพธ์เป็น 0 จากนั้นจะทำการเปรียบเทียบผลลัพธ์ที่ได้จาก Perceptron กับผลลัพธ์ที่แท้จริงของข้อมูล ถ้าผลลัพธ์ไม่ตรงกันก็จะทำการปรับค่าของน้ำหนักใหม่ หลังจากนั้นก็จะทำการอ่านข้อมูลสอนตัวต่อไปแล้วทำการเปรียบเทียบค่าและปรับน้ำหนัก วนซ้ำไปเรื่อย ๆ จนกระทั่ง Perceptron ให้ผลลัพธ์ที่ตรงกับข้อมูลสอนทุกตัว ขั้นตอนการเรียนรู้สำหรับ Perceptron แสดงดังรูปที่ 2.14

Perceptron Learning Rule

***ข้อมูลนำเข้า**: ชุดข้อมูลสอน x(x_1, x_2, …, x_n: class), น้ำหนักเริ่มต้น (w_i), ค่าอัตราการเรียนรู้ (Learning Rate: η), จำนวนรอบสูงสุด*

1. อ่านข้อมูลสอนมาทีละตัว

1.1 หาค่า $\sum_{i=0}^{n}(x_i w_i)$

1.2 กำหนดให้

$$\text{Output} = \begin{cases} 1 & \text{if } \sum x_i w_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

1.3 ถ้า Output ที่ได้จาก Perceptron (o) ไม่ตรงกับ class ของข้อมูล (t) ให้ปรับน้ำหนักแต่ละตัว (w_i) ตามสมการ

$$w_i \leftarrow w_i + \Delta w_i$$

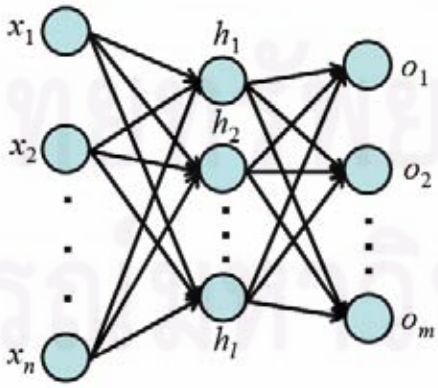
$$\text{โดยที่ } \Delta w_i = \eta(t-o)x_i$$

2. วนทำซ้ำจนกระทั่งไม่มีการปรับน้ำหนักหรือครบจำนวนรอบสูงสุดที่กำหนด

***ผลลัพธ์**: เวกเตอร์ของน้ำหนัก (w_i)*

รูปที่ 2.14 ขั้นตอนการเรียนรู้สำหรับ Perceptron

สำหรับเครือข่ายใยประสาทเทียมนั้นจะเป็นการนำ Perceptron หลาย ๆ ตัวมาเรียงต่อกันเป็นชั้น ๆ เรียกว่า Mutilayer Perceptron แสดงดังตัวอย่างในรูปที่ 2.15 โดยที่แต่ละโนดจะหมายถึง Perceptron แต่ละตัว



รูปที่ 2.15 โครงสร้าง Multilayer Perceptron

(ที่มา: บุญเสริม กิจสิริกุล, 2546 : 184)

2.4 การจัดกลุ่มข้อมูล (Data Clustering)

การจัดกลุ่มข้อมูล เป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยมีวัตถุประสงค์เพื่อจัดข้อมูลที่มีลักษณะคล้ายกันให้อยู่ในกลุ่มเดียวกัน ซึ่งข้อมูลที่จะนำมาใช้สำหรับการจัดกลุ่มจะไม่มีตัวบ่งบอกว่าข้อมูลตัวนั้น ๆ จะอยู่ในกลุ่มใด เทคนิคพื้นฐานที่นิยมใช้ในการจัดกลุ่มข้อมูล ได้แก่

2.4.1 การจัดกลุ่มแบบ K-Means

การจัดกลุ่มข้อมูลแบบ K-Means (MacQueen, 1967) จะเริ่มด้วยการกำหนดจำนวนกลุ่ม จากนั้นจะทำการกำหนดเซนทรอยด์ (Centroid) ของแต่ละกลุ่มแบบสุ่ม และทำการจัดข้อมูลแต่ละตัวเข้าไปในกลุ่มที่มีระยะห่างระหว่างเซนทรอยด์กับข้อมูลตัวนั้นน้อยที่สุด โดยใช้ฟังก์ชันการวัดความเหมือนใด ๆ ซึ่งฟังก์ชันที่นิยมใช้ที่เป็นพื้นฐานมากที่สุดตัวหนึ่ง ได้แก่ การวัดระยะห่างแบบยูคลิเดียน (Euclidean Distance) ซึ่งคำนวณได้จากสมการ

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \qquad (2.2)$$

โดยที่ $dist(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง x_j

d คือ จำนวนคุณลักษณะทั้งหมดของตัวอย่าง

$x_{i,k}$ คือ คุณลักษณะตัวที่ k ของตัวอย่าง x_i

หลังจากจัดกลุ่มข้อมูลให้กับตัวอย่างจนครบทุกตัวแล้ว ค่าเซนทรอยด์ของกลุ่มจะถูกคำนวณใหม่โดยใช้ข้อมูลตัวอย่างในกลุ่มของตัวเอง และจะทำการจัดกลุ่มให้กับข้อมูลตัวอย่างแต่ละตัวใหม่ ซึ่งจะทำเช่นนี้ซ้ำไปเรื่อย ๆ จนกระทั่งพบเงื่อนไขของการสิ้นสุด เช่น สมาชิกของแต่ละกลุ่มไม่มีการเปลี่ยนแปลง หรือ ครบตามจำนวนรอบสูงสุดที่กำหนด ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบ K-Means แสดงดังรูปที่ 2.16

การจัดกลุ่มข้อมูลแบบ K-Means

***ข้อมูลนำเข้า**: ชุดข้อมูลสอน x(a_1, a_2, …, a_n), ฟังก์ชันวัดความเหมือน, จำนวนรอบ*

1. กำหนดจำนวนกลุ่ม K

2. กำหนดเซนทรอยด์ของแต่ละกลุ่มแบบสุ่ม

3. เริ่มทำซ้ำ

3.1 คำนวณความเหมือนของตัวอย่างแต่ละตัวกับเซนทรอยด์ของแต่ละกลุ่มด้วยฟังก์ชันวัดความเหมือน

3.2 กำหนดให้ตัวอย่างแต่ละตัวเป็นสมาชิกของกลุ่มที่มีความเหมือนกับเซนทรอยด์ของกลุ่มนั้นมากที่สุด

3.3 คำนวณค่าเซนทรอยด์ของแต่ละกลุ่มใหม่ด้วยข้อมูลสมาชิกของกลุ่มนั้น ๆ

4. วนทำซ้ำจนกระทั่งพบเงื่อนไขการสิ้นสุด

***ผลลัพธ์**: สมาชิกของแต่ละกลุ่ม*

รูปที่ 2.16 ขั้นตอนการจัดกลุ่มข้อมูลแบบ K-Means

2.4.2 การจัดกลุ่มแบบ Fuzzy c – Means

เป็นการจัดกลุ่มข้อมูลโดยใช้หลักการหลักการของความคลุมเครือ (Fuzzy) (Bezdek, 1981) โดยขั้นตอนวิธี จะพยายามลดค่า J_m ในสมการที่ 2.3 โดยที่ μ_{ij} คือความเป็นสมาชิกคลุมเครือ (Fuzzy Membership) และ V_i คือเซนทรอยด์ของกลุ่ม

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m d^2(X_j, V_i) \qquad (2.3)$$

โดยที่

$$d^2(X_j, V_i) = (X_j - V_i)^T A(X_j - V_i) \qquad (2.4)$$

A คือ เมตริกขนาด p×p โดยที่ p คือมิติของตัวอย่าง X_j (j=1,2,…,n), c คือจำนวนกลุ่ม, n คือจำนวนตัวอย่าง และ m>1 คือ ดัชนีความคลุมเครือ (fuzziness index) ขั้นตอนวิธีการจัดกลุ่มแบบ Fuzzy c-Means แสดงดังภาพที่ 2.17

การจัดกลุ่มข้อมูลแบบ Fuzzy c-Means

***ข้อมูลนำเข้า**: ชุดข้อมูลสอน $x(a_1, a_2, \ldots, a_n)$, จำนวนกลุ่ม c*

1. กำหนดค่าเริ่มต้นของ μ_{ij} ของ X_j สำหรับกลุ่ม i ดังนี้

$$\sum_{i=1}^{c} \mu_{ij} = 1$$

2. คำนวณเซนทรอยด์คลุมเครือ (fuzzy centroid) ดังนี้

$$V_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m X_j}{\sum_{j=1}^{n} (\mu_{ij})^m}$$

3. ปรับปรุงค่าความเป็นสมาชิกคลุมเครือ ดังนี้

$$\mu_{ij} = \frac{\left(\frac{1}{d^2(X_j, V_i)} \right)^{\frac{1}{(m-1)}}}{\sum_{i=1}^{c} \left(\frac{1}{d^2(X_j, V_i)} \right)^{\frac{1}{(m-1)}}}$$

4. กลับไปทำซ้ำข้อ 2 และ ข้อ 3 จนกระทั่งค่า J_m ไม่ลดต่ำลง

รูปที่ 2.17 ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบ Fuzzy c-Means

2.5 การเลือกคุณลักษณะ (Feature Selection)

การวิเคราะห์ข้อมูลที่มีมิติจำนวนมากนั้น ถือเป็นเรื่องยากและมีความซับซ้อนสูง ขั้นตอนวิธีสำหรับการเรียนรู้ที่จะนำมาใช้จัดการกับข้อมูลเหล่านี้จะต้องใช้ทรัพยากรทั้งในด้านการคำนวณและการใช้หน่วยความจำจำนวนมาก ในขณะที่ประสิทธิภาพของกระบวนการเรียนรู้อาจลดลงเนื่องจากอาจมีสัญญาณรบกวน (Noise) ในข้อมูลที่เกิดจากมิติที่ไม่เกี่ยวข้องกับการวิเคราะห์ข้อมูลนั้น ๆ ดังนั้น เพื่อแก้ปัญหาดังกล่าว มิติของข้อมูลเหล่านั้นจะต้องถูกลดจำนวนลงด้วยวิธีการเลือกคุณลักษณะที่สำคัญสำหรับการวิเคราะห์ข้อมูล

การเลือกคุณลักษณะอาจแบ่งประเภทตามวิธีการในการเลือกคุณลักษณะ เป็น 3 ประเภท ได้แก่ Embedded, Wrapper และ Filter (Molina et al., 2002) ซึ่งมีรายละเอียดดังต่อไปนี้

2.5.1 การเลือกคุณลักษณะแบบ Embedded

การเลือกคุณลักษณะแบบ Embedded จะเกิดขึ้นโดยขั้นตอนวิธีสำหรับการเรียนรู้เอง ซึ่งขั้นตอนวิธีเหล่านี้จะมีการเลือกคุณลักษณะที่เหมาะสมสำหรับการสร้างแบบจำลองในการแก้ปัญหาต่าง ๆ โดยไม่จำเป็นต้องเพิ่มขั้นตอนวิธีในการเลือกคุณลักษณะอื่น ๆ เข้ามาช่วย ตัวอย่างของขั้นตอนวิธีสำหรับการเรียนรู้ที่มีการเลือกคุณลักษณะแบบ Embedded ได้แก่ ต้นไม้ตัดสินใจ และ กำหนดการพันธุกรรม

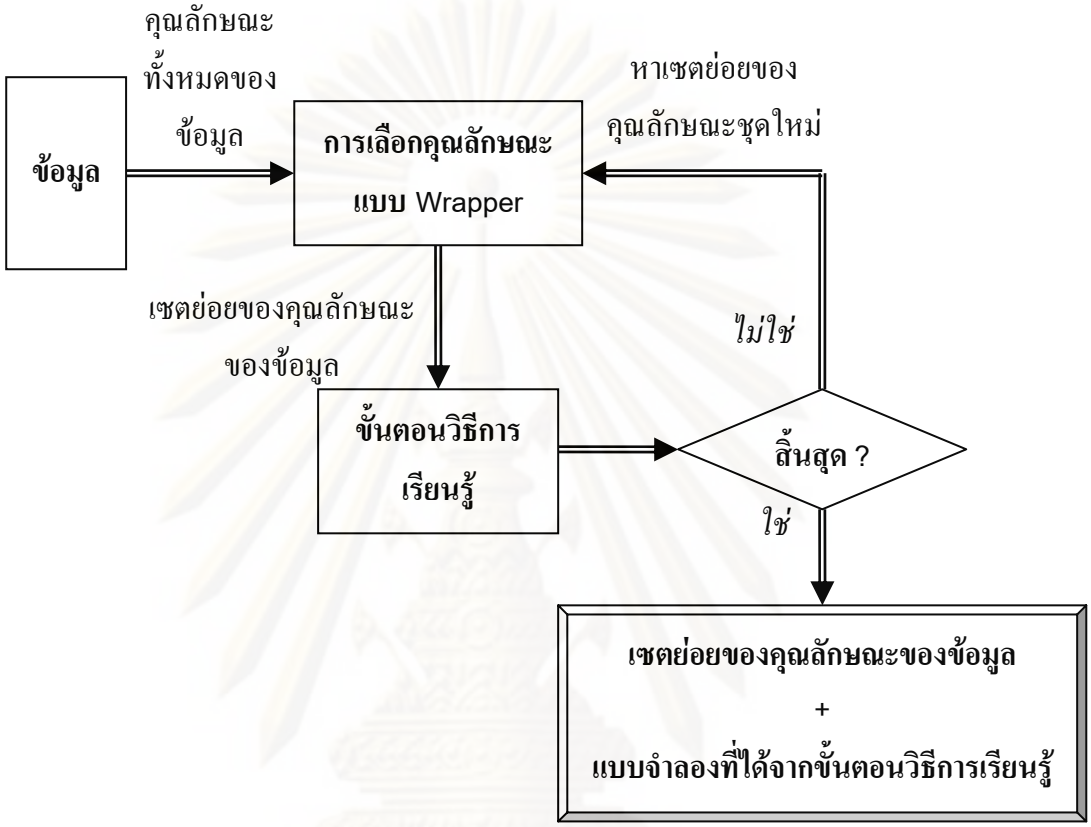
2.5.2 การเลือกคุณลักษณะแบบ Wrapper

การเลือกคุณลักษณะแบบ Wrapper จะเป็นขั้นตอนกระบวนการหนึ่งสำหรับคัดเลือกเซตย่อยจากคุณลักษณะทั้งหมดของข้อมูล โดยจะเน้นที่กระบวนการค้นหาเซตย่อยของคุณลักษณะที่เหมาะสมกับขั้นตอนวิธีการเรียนรู้วิธีใดวิธีหนึ่งโดยเฉพาะ ดังนั้นวิธีการเลือกคุณลักษณะแบบนี้ ถือได้ว่าจะเป็นการเพิ่มประสิทธิภาพของขั้นตอนวิธีการเรียนรู้ได้ดีที่สุด แต่ข้อเสียที่สำคัญสำหรับวิธีการนี้ คือ จะต้องใช้เวลาในการเรียนรู้มาก และเซตย่อยของคุณลักษณะที่เลือกได้ จะเหมาะกับวิธีการเรียนรู้แบบหนึ่ง ซึ่งอาจไม่เหมาะกับวิธีการเรียนรู้แบบอื่น ๆ และปัญหาที่สำคัญอีกประการหนึ่งคือ อาจเกิดปัญหา Overfitting ได้ (Kohavi and Sommerfield, 1995) ตัวอย่างของวิธีการเลือกคุณลักษณะแบบ Wrapper เช่น การใช้ขั้นตอนวิธีพันธุกรรม (Genetic Algorithm) เพื่อค้นหาเซตของคุณลักษณะที่เหมาะสมกับขั้นตอนวิธีการเรียนรู้ใด ๆ (Vafaie and De Jong, 1992) ภาพรวมของการเลือกคุณลักษณะแบบ Wrapper แสดงดังรูปที่ 2.18

2.5.3 การเลือกคุณลักษณะแบบ Filter

การเลือกคุณสมบัติแบบ Filter จะเป็นการประเมินประสิทธิภาพของคุณลักษณะของข้อมูลแต่ละตัวว่ามีความเหมาะสมกับการวิเคราะห์ข้อมูลมากน้อยเพียงใด โดยไม่ขึ้นกับขั้นตอนวิธีของการเรียนรู้แบบใดแบบหนึ่ง การเลือกคุณลักษณะแบบนี้จะทำการจัดลำดับ (Ranking) ตามความสำคัญของคุณลักษณะแต่ละตัว และเลือกคุณลักษณะที่มีระดับความสำคัญสูงสุดตามจำนวนที่ผู้ใช้ระบุ หรืออาจระบุเป็นค่าขีดแบ่ง (Threshold) ของคุณลักษณะที่จะเลือกก็ได้ ข้อดีของการเลือกคุณลักษณะแบบนี้ คือ การประมวลผลที่รวดเร็ว และไม่ขึ้นกับขั้นตอนวิธีการเรียนรู้

แบบใด ๆ แต่ข้อเสียที่สำคัญก็คือ จำนวนคุณลักษณะ หรือ ค่าขีดแบ่งที่เหมาะสมนั้น ไม่สามารถทราบได้



รูปที่ 2.18 ภาพรวมของการเลือกคุณลักษณะแบบ Wrapper

การเลือกคุณลักษณะแบบ Filter ที่เป็นที่นิยมใช้ ได้แก่

ก) Signal-to-Noise Ratio (SNR)

SNR เป็นวิธีการทางสถิติเพื่อวัดประสิทธิภาพของคุณลักษณะในการจำแนกประเภทข้อมูลจากข้อมูลกลุ่มหนึ่งออกจากข้อมูลกลุ่มอื่น ๆ (Slonim et al., 2000) การคำนวณหาค่า SNR แสดงดังสมการที่ 2.5

$$SNR_F = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \quad (2.5)$$

โดยที่ μ_1 และ μ_2 คือ ค่าเฉลี่ยของข้อมูลไมโครอาร์เรย์ของกลุ่มที่ 1 และ กลุ่มที่ 2

σ_1 และ σ_2 คือ ค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลในแต่ละกลุ่ม

ข) Correlation Coefficient Analysis

เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร 2 ตัว โดยในปัญหาการเลือกคุณลักษณะจะต้องมีการกำหนดตัวแปรอุดมคติ (Ideal Variable: V_{ideal}) เพื่อใช้ในการเปรียบเทียบความสัมพันธ์กับตัวแปรอื่น ๆ ว่ามีความคล้ายกับตัวแปรอุดมคติมากน้อยเพียงใด โดยกำหนดให้ N คือจำนวนข้อมูลตัวอย่างทั้งหมด ประกอบด้วย M ตัวแรกเป็นข้อมูลที่อยู่ในประเภทที่ 1 และ N – M ตัวที่เหลือเป็นข้อมูลที่อยู่ในประเภทที่ 2 จะได้ว่า $V_{ideal} = \{1_1, 1_2, 1_3, \ldots 1_M, 0_{M+1}, \ldots 0_N\}$ โดยการเปรียบเทียบ สามารถคำนวณค่า correlation coefficient (r) ได้หลายวิธี ได้แก่ $r_{Pearson}$, $r_{Euclidean}$ และ r_{Cosine} ดังนี้

$$r_{Pearson} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (2.6)$$

$$r_{Euclidean} = \sqrt{\sum (X - Y)^2} \quad (2.7)$$

$$r_{Cosine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad (2.8)$$

โดยที่ X คือตัวแปรใด ๆ และ Y คือ ตัวแปรอุดมคติ

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ค) RELIEF

เป็นเทคนิคการเรียนรู้ของเครื่องที่พัฒนาโดย Kira and Rendell (1992) ที่ใช้ประมาณค่าความสำคัญของตัวแปรที่สามารถจำแนกประเภทข้อมูลจากประเภทหนึ่งออกจากประเภทที่เหลือ ขั้นตอนวิธีของ RELIEF แสดงดังรูปที่ 2.19

ขั้นตอนวิธี RELIEF

***ข้อมูลนำเข้า**: ชุดข้อมูลสอน (x,y), จำนวนรอบ n*

1. กำหนดค่าน้ำหนัก W[A]=0.0

2. สำหรับแต่ละรอบ i=1 ถึง n

2.1 เลือกตัวอย่าง R แบบสุ่ม

2.2 หาตัวอย่าง H ที่ใกล้ที่สุดที่เป็นประเภทเดียวกับ R

2.3 หาตัวอย่าง M ที่ใกล้ที่สุดที่ไม่ใช่ประเภทเดียวกับ R

2.4 สำหรับแต่ละรอบ A=1 ถึง จำนวนคุณลักษณะ

2.4.1 ให้ W[A]=W[A]-diff(A,R,H)/n + diff(A,R,M)/n

โดยที่ diff(A,R,H) คือ ความแตกต่างของคุณลักษณะ A ระหว่างข้อมูล R และ H (ข้อมูลที่อยู่ในกลุ่มเดียวกัน)

Diff(A,R,M)) คือ ความแตกต่างของคุณลักษณะ A ระหว่างข้อมูล R และ M (ข้อมูลที่อยู่คนละกลุ่ม)

***ผลลัพธ์**: น้ำหนัก W[A]*

รูปที่ 2.19 ขั้นตอนวิธี RELIEF

2.6 วิธีการแบบกลุ่มก้อน (Ensemble Method)

แนวคิดพื้นฐานของการใช้วิธีการแบบกลุ่มก้อนของตัวจำแนกประเภท คือ ไม่มีตัวจำแนกประเภทเดี่ยวใด ๆ ที่มีประสิทธิภาพในทุกสภาพการณ์ ดังนั้นการใช้กลุ่มก้อนของตัวจำแนกประเภทจึงให้ประสิทธิภาพที่ดีกว่า โดยประสิทธิภาพที่ดีของกลุ่มก้อนของตัวจำแนกประเภท จะขึ้นอยู่กับประสิทธิภาพและความหลากหลายของสมาชิกแต่ละตัวในกลุ่มก้อน ซึ่งก็มีเทคนิคที่นำมาใช้ในการสร้างความหลากหลายของตัวจำแนกประเภทหลายวิธีการ แต่วิธีการที่มีประสิทธิภาพที่เป็นที่รู้จักกันดี และได้รับความนิยม ได้แก่ วิธี Bagging และ AdaBoost ซึ่งมีรายละเอียดดังนี้

2.6.1 Bagging (Boostrap Aggregating)

วิธี Bagging ถูกนำเสนอโดย Breiman (1996) เป็นวิธีการสร้างตัวจำแนกประเภทหลายตัวโดยการจัดการกับชุดข้อมูลสอนที่จะนำมาใช้สร้างตัวจำแนกข้อมูลให้มีความแตกต่างกัน ทำให้เกิดความหลากหลายของตัวจำแนกประเภท ซึ่งขั้นตอนวิธี Bagging แสดงดังรูปที่ 2.20

ขั้นตอนวิธี *Bagging*

ข้อมูลนำเข้า*: ชุดข้อมูลสอน (x,y), ขั้นตอนวิธีทางการเรียนรู้ของเครื่องจักร, จำนวนรอบ j*

1. สำหรับแต่ละรอบ i=1 ถึง j

1.1 เลือกกลุ่มย่อยของชุดข้อมูลสอน (S) ขนาด N จากชุดข้อมูลสอนทั้งหมดแบบสุ่ม

1.2 สร้างตัวจำแนกประเภท C_i ด้วยชุดข้อมูล S

2. นำตัวจำแนกประเภท C_i แต่ละตัวมาสร้างเป็นกลุ่มก้อนของตัวจำแนกประเภท

3. ในการจำแนกประเภทข้อมูลตัวอย่าง x จะใช้เสียงข้างมากจากการออกเสียงของตัวจำแนกประเภท C_i ทุกตัว

ผลลัพธ์*: กลุ่มก้อนของตัวจำแนกประเภท C_i ทุกตัว*

รูปที่ 2.20 ขั้นตอนวิธี Bagging

2.6.2 AdaBoost (Adaptive Boosting)

วิธี AdaBoost เป็นอีกวิธีการหนึ่งที่ได้รับความนิยมในการสร้างกลุ่มก้อนของตัวจำแนกประเภท ถูกเสนอโดย Yoav Freund และ Robert E. Schapire (1996) ซึ่งจะให้ความสำคัญกับข้อมูลในชุดข้อมูลสอนไม่เท่ากัน โดยจะคำนวณความสำคัญของข้อมูลจากการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทก่อนหน้า ซึ่งถ้าข้อมูลตัวใดถูกจำแนกประเภทด้วยตัวจำแนกประเภทก่อนหน้าไม่ถูกต้อง จะทำการเพิ่มความสำคัญของข้อมูลตัวนั้นสำหรับตัวจำแนกประเภทที่จะสร้างตัวต่อไป และในทางกลับกัน ข้อมูลที่ถูกจำแนกประเภทได้ถูกต้อง ก็จะถูกลดความสำคัญลง โดยที่ตัวจำแนกประเภทแต่ละตัวก็จะมีน้ำหนักในการออกเสียงที่ต่างกัน ตามความสามารถในการจำแนกข้อมูล ซึ่งขั้นตอนวิธี AdaBoost แสดงดังรูปที่ 2.21

ขั้นตอนวิธี *AdaBoost*

ข้อมูลนำเข้า: *ชุดข้อมูลสอน (x,y), ขั้นตอนวิธีทางการเรียนรู้ของเครื่องจักร, จำนวนรอบ j*

1. กำหนดน้ำหนักให้กับข้อมูลแต่ละตัว $D_i = 1/m$ *(m คือจำนวนข้อมูลทั้งหมด)*

2. สำหรับแต่ละรอบ $i = 1$ *ถึง* j

2.1 สร้างตัวจำแนกประเภท C_i *ด้วยชุดข้อมูลสอนและน้ำหนัก* D_i

2.2 คำนวณความผิดพลาดของ C_i *จาก*

$$\varepsilon_i = \sum_{k:C_i(x_k)\neq y_k} D_i(k)$$ *(คือ ผลรวมของ* D_i *ที่* C_i *ทำนายข้อมูลตัวที่ k ผิด)*

2.3 กำหนดให้ $\alpha_i = \frac{1}{2}\ln(\frac{1-\varepsilon_i}{\varepsilon_i})$

2.4 ปรับปรุงน้ำหนักของตัวอย่างแต่ละตัว จาก

$$D_{i+1}(k) = \frac{D_i(k)}{Z_i} \times \begin{cases} e^{-\alpha} & \text{ถ้า } C_i(x_k) = y_k \\ e^{\alpha} & \text{ถ้า } C_i(x_k) \neq y_k \end{cases}$$

โดยที่ Z_i *คือ ปัจจัยความเป็นปกติ (normalization factor)*

3. ผลการจำแนกประเภทข้อมูลจะคำนวณจาก

$$C(x) = sign(\sum_{i=1}^{j} \alpha_i C_i(x))$$

ผลลัพธ์: *กลุ่มก้อนของตัวจำแนกประเภท* C_i *กับ น้ำหนัก* α_i *ทุกตัว*

รูปที่ 2.21 ขั้นตอนวิธี AdaBoost

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 3
การจำแนกประเภทข้อมูลไมโครอาร์เรย์ด้วยตัวจำแนกประเภทกำหนดการพันธุกรรม

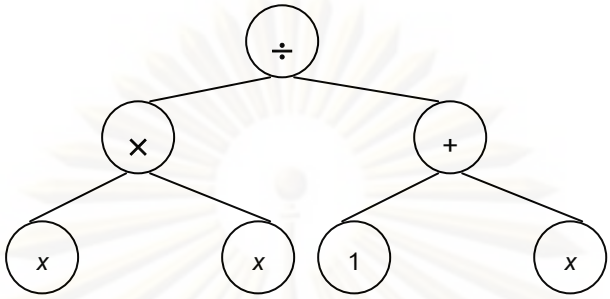
3.1 ตัวจำแนกประเภทกำหนดการพันธุกรรม

การสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมนั้น ผลเฉลยจะถูกแทนให้อยู่ในรูปต้นไม้จำแนกประเภท (Classification Tree) ซึ่งสามารถออกแบบต้นไม้จำแนกประเภทที่แตกต่างกันได้หลายลักษณะ เช่น นิพจน์ทางคณิตศาสตร์ (Hong and Cho, 2004) แสดงดังรูปที่ 3.1 (ก) นิพจน์ทางตรรกศาสตร์ (Freitas, 1997) แสดงดังรูปที่ 3.1 (ข) หรือ แบบผสม (Loveard and Ciesielski, 2001) แสดงดังรูปที่ 3.1 (ค) ซึ่งข้อดีของการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรมคือ โครงสร้างของกฎไม่ถูกกำหนดตายตัว และผลลัพธ์ที่ได้จากการเรียนรู้สามารถนำวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างคุณลักษณะต่าง ๆ ของข้อมูลที่ส่งผลต่อการแบ่งประเภทของข้อมูล ทำให้มีการพัฒนาระบบการจำแนกประเภทข้อมูลด้วยวิธีกำหนดการพันธุกรรมอย่างกว้างขวาง เพื่อปรับปรุงประสิทธิภาพของการจำแนกประเภทข้อมูล เช่น การพัฒนาระบบการจำแนกประเภทที่ข้อมูลมีหลายกลุ่ม (Muni et al., 2004) หรือการจำแนกประเภทที่ข้อมูลมีคุณลักษณะจำนวนมาก (Hong and Cho, 2004)

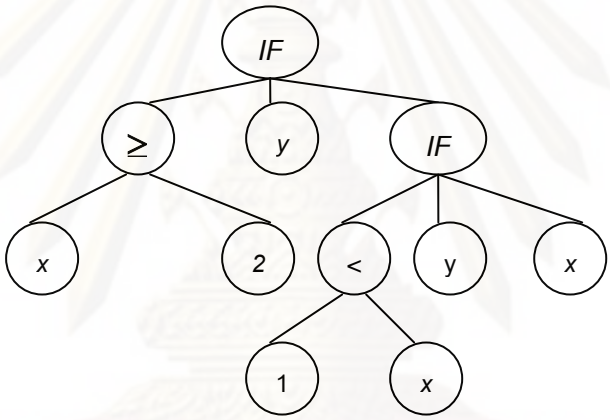
ได้มีการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยวิธีกำหนดการพันธุกรรมกับวิธีการเรียนรู้แบบอื่น ๆ ซึ่งผลการเปรียบเทียบก็มีทั้งที่ให้ประสิทธิภาพที่ใกล้เคียง และมีทั้งที่ให้ประสิทธิภาพที่ดีกว่าวิธีการอื่น ๆ ดังต่อไปนี้

Eggermont et al. (1999) ได้ทำการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยวิธีกำหนดการพันธุกรรม กับวิธีอื่น ๆ 3 วิธี ได้แก่ LogDisc (เป็นวิธีการทางสถิติ) C4.5 (เป็นรูปแบบหนึ่งของการเรียนรู้แบบต้นไม้ตัดสินใจ) และ Back-propagation (การเรียนรู้แบบเครือข่ายใยประสาทเทียม) โดยทำการทดสอบกับชุดข้อมูลทดสอบ 2 ชุด ได้แก่ Australian Credit และ Pima Indians Diabetes ซึ่งเป็นชุดข้อมูลเกณฑ์เปรียบเทียบสมรรถนะสำหรับการทดสอบประสิทธิภาพของวิธีการต่าง ๆ ในกลุ่มการเรียนรู้ของเครื่อง ผลเฉลยที่ใช้อยู่ในรูปนิพจน์ทางคณิตศาสตร์ ผลการเปรียบเทียบแสดงดังตารางที่ 3.1 ซึ่งจากผลการเปรียบเทียบแสดงให้เห็นว่า กำหนดการพันธุกรรมให้ประสิทธิภาพที่แย่ที่สุดสำหรับชุดข้อมูล Australian Credit แต่สำหรับชุดข้อมูล Pima Indians Diabetes ประสิทธิภาพของกำหนดการพันธุกรรมไม่แย่

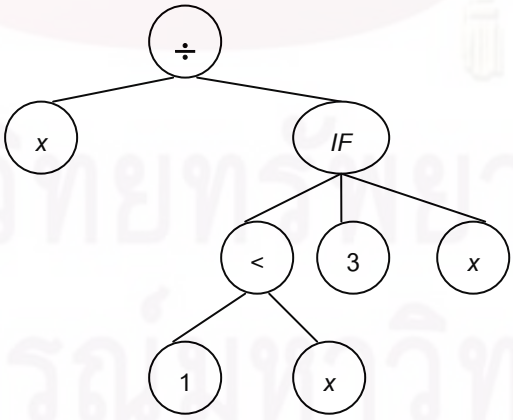
มากนัก ผลที่ได้ใกล้เคียงกับประสิทธิภาพของ Back-propagation และให้ผลดีกว่า C4.5 เป็นอย่างมาก



(ก) การแทนที่ผลเฉลยด้วยนิพจน์ทางคณิตศาสตร์



(ข) การแทนที่ผลเฉลยด้วยนิพจน์ทางตรรกศาสตร์



(ค) การแทนที่ผลเฉลยแบบผสม

รูปที่ 3.1 การแทนที่ผลเฉลยแบบต่าง ๆ สำหรับตัวจำแนกประเภทกำหนดการพันธุกรรม

ตารางที่ 3.1 ผลการเปรียบเทียบอัตราความผิดพลาด (%) ของการจำแนกประเภทข้อมูลด้วยวิธีกำหนดการพันธุกรรมกับวิธี LogDisc, C4.5 และ Back-propagation ของ Eggermont et al. (1999)

Data set	Australian Credit	Pima Indians Diabetes
BackProp	0.154	0.248
C4.5	0.155	0.270
Standard GP	0.232	0.253
LogDisc	0.141	0.223

Brameier and Banzhaf (2001) ได้ใช้กำหนดการพันธุกรรมเชิงเส้น (Linear Genetic Programming) ซึ่งเป็นรูปแบบหนึ่งของกำหนดการพันธุกรรม โดยแทนผลเฉลยอยู่ในรูปของภาษาโปรแกรมที่แทนต้นไม้จำแนกประเภทในรูปแบบผสม มาใช้ในปัญหาการจำแนกประเภทข้อมูล โดยทำการทดสอบกับชุดข้อมูลเกณฑ์เปรียบเทียบสมรรถนะทางด้านการแพทย์ ซึ่งรายละเอียดของชุดข้อมูลแสดงดังตารางที่ 3.2 และได้ทำการเปรียบเทียบประสิทธิภาพกับเครือข่ายใยประสาทเทียม ผลการเปรียบเทียบแสดงดังตารางที่ 3.3 จากผลการเปรียบเทียบในคอลัมน์สุดท้าย (Δ) แสดงถึงค่าร้อยละของความแตกต่างของอัตราความผิดพลาดการจำแนกประเภท เครื่องหมายบวกแสดงถึงประสิทธิภาพที่ดีกว่าของกำหนดการพันธุกรรม ส่วนเครื่องหมายลบแสดงถึงประสิทธิภาพที่ดีกว่าของเครือข่ายใยประสาทเทียม จากผลการเปรียบเทียบแสดงให้เห็นว่า ประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยวิธีกำหนดการพันธุกรรมนั้น สามารถเปรียบเทียบได้กับการจำแนกประเภทข้อมูลด้วยวิธีเครือข่ายใยประสาทเทียม

ตารางที่ 3.2 รายละเอียดของชุดข้อมูลที่ใช้ในการวิจัยของ Brameier and Banzhaf (2001)

Problem	Diagnosis task
cancer	benign or malignant breast tumor
diabetes	diabetes positive or negative
gene	intron-exon, exon-intron or no boundary in DNA sequence
heart	diameter of a heart vessel is reduced by more than 50% or not
horse	horse with a colic will die, survive or must be killed
thyroid	thyroid hyperfunction, hypofunction or normal function

ตารางที่ 3.3 ผลการเปรียบเทียบประสิทธิภาพระหว่างกำหนดการพันธุกรรมกับเครือข่ายใยประสาทเทียมในการวิจัยของ Brameier and Banzhaf (2001)

	GP						NN		
Problem	Validation CE (%)			Test CE (%)			Test CE (%)		Δ (%)
	best	*average*	*stddev*	*best*	***average***	*stddev*	***average***	*stddev*	
cancer1	1.71	2.45	0.34	0.57	2.18	0.59	1.38	0.49	-36.70
cancer2	0.57	1.39	0.40	4.02	5.72	0.66	4.77	0.94	-16.61
cancer3	1.71	2.62	0.45	3.45	4.93	0.65	3.70	0.52	-24.95
diabetes1	20.31	22.19	1.09	21.35	23.96	1.42	24.10	1.91	+0.58
diabetes2	21.35	23.21	1.33	25.00	27.85	1.49	26.42	2.26	-5.14
diabetes3	25.52	26.69	0.65	19.27	23.09	1.27	22.59	2.23	-2.17
gene1	7.81	11.16	2.30	9.21	12.97	2.24	16.67	3.75	+22.20
gene2	9.07	12.93	2.30	8.45	11.95	2.15	18.41	6.93	+35.09
gene3	7.18	10.77	2.11	10.09	13.84	2.09	21.82	7.53	+36.57
heart1	7.89	10.53	2.38	18.67	21.12	2.02	20.82	1.47	-1.42
heart2	14.47	18.58	2.39	1.33	7.31	3.31	5.13	1.63	-29.82
heart3	15.79	18.81	1.47	10.67	13.98	2.03	15.40	3.20	+9.22
horse1	28.57	32.40	2.22	23.08	30.55	2.24	29.19	2.62	-4.45
horse2	29.67	34.30	2.65	31.87	36.12	1.95	35.86	2.46	-0.72
horse3	27.47	32.65	1.94	31.87	35.44	1.77	34.16	2.32	-3.61
thyroid1	0.83	1.31	0.34	1.28	1.91	0.42	2.38	0.35	+19.75
thyroid2	1.11	1.62	0.31	1.44	2.31	0.39	1.91	0.24	-17.32
thyroid3	0.89	1.47	0.23	0.89	1.88	0.36	2.27	0.32	+17.18

ในขณะที่ Gray et al. (1996) ได้ทำการทดลองจำแนกประเภทข้อมูลเนื้องอกในสมองจากข้อมูลการวิเคราะห์เนื้อเยื่อด้วยวิธี Nuclear Magnetic Resonance ข้อมูลประกอบด้วย 75 ตัวอย่าง โดยมี 28 ตัวอย่างที่เป็นเนื้องอก และส่วนที่เหลือไม่เป็น ซึ่งแต่ละตัวอย่างประกอบด้วย 400 ตัวแปร โดยการทดลองนี้ได้ทำการเปรียบเทียบประสิทธิภาพของการจำแนกประเภทด้วยวิธีกำหนดการพันธุกรรม ซึ่งแทนผลเฉลยในรูปแบบผสม กับการจำแนกประเภทด้วยวิธีเครือข่ายใยประสาทเทียม จากผลการทดลองพบว่า กำหนดการพันธุกรรมให้ค่าความแม่นของการจำแนกประเภทข้อมูล ร้อยละ 90 ในขณะที่ เครือข่ายใยประสาทเทียมให้ค่าความแม่นร้อยละ 80 ซึ่งก็พบว่า กำหนดการพันธุกรรมให้ประสิทธิภาพที่ดีกว่า

Bojarczuk et al. (2001) ได้เสนอกำหนดการพันธุกรรมแบบจำกัดวากยสัมพันธ์ (Constrained-Syntax Genetic Programming) เพื่อลดข้อผิดพลาดที่อาจจะเกิดขึ้นกับกฎสำหรับการจำแนกประเภทข้อมูล ซึ่งแทนผลเฉลยในรูปนิพจน์ทางตรรกศาสตร์ โดยได้ทำการทดสอบกับชุดข้อมูลเกณฑ์เปรียบเทียบสมรรถนะในกลุ่มการเรียนรู้ของเครื่อง 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลการเจ็บหน้าอก (Chest pain) ประกอบด้วยข้อมูลจำนวน 138 ระเบียน 161 คุณลักษณะ ชุดข้อมูลโรคผิวหนัง (Dermatology) ประกอบด้วยข้อมูลจำนวน 366 ระเบียน 34 คุณลักษณะ และชุดข้อมูล

มะเร็งเต้านม ประกอบด้วยข้อมูลจำนวน 286 ระเบียน 9 คุณลักษณะ โดยได้ทำการเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลกับขั้นตอนวิธีพันธุกรรมที่รายงานใน Fidelis et al. (2000) และวิธี C4.5 ซึ่งเป็นวิธีการหนึ่งของต้นไม้ตัดสินใจ ผลการทดลองแสดงดังตารางที่ 3.4 ซึ่งพบว่าประสิทธิภาพของตัวจำแนกประเภทกำหนดการพันธุกรรมให้ประสิทธิภาพที่ดีที่สุด

ตารางที่ 3.4 การเปรียบเทียบอัตราความแม่น (%) ของการจำแนกประเภทข้อมูลในรายงานของ Bojarczuk et al. (2001)

Data set	Proposed GP	GA	C4.5
Chest pain	80.31±7.80	N/A	73.18
Dermatology	96.64±2.27	94.96	89.12
Breast cancer	71.79±9.36	67.39	71.38

จากผลการเปรียบเทียบประสิทธิภาพของการจำแนกประเภทด้วยวิธีกำหนดการพันธุกรรมกับวิธีการอื่น ๆ พบว่า วิธีกำหนดการพันธุกรรมให้ประสิทธิภาพในการจำแนกประเภทข้อมูลไม่ด้อยไปกว่าวิธีการอื่น ๆ โดยเฉพาะข้อมูลทางด้านการแพทย์ ที่วิธีการกำหนดการพันธุกรรมให้ประสิทธิภาพที่ค่อนข้างดี

3.2 การสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์

ข้อมูลไมโครอาร์เรย์เป็นข้อมูลตารางขนาดใหญ่ที่ข้อมูลทั้งหมดเป็นตัวเลข ดังนั้นการสร้างตัวจำแนกประเภทสำหรับข้อมูลไมโครอาร์เรย์จึงมักนิยมใช้แบบจำลองทางคณิตศาสตร์ เช่น Muti-Layer Perceptron , K-Nearest Neighbor , Self-Organizing Map , Support Vector Machine (Cho and Won, 2003), Simplified Fuzzy ARTMAP (Azuaje, 2000) รวมถึงการสร้างตัวจำแนกประเภทด้วยวิธีกำหนดการพันธุกรรมในรูปแบบที่เป็นนิพจน์ทางคณิตศาสตร์ (Hong and Cho, 2004)

การสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์ในรูปแบบนิพจน์คณิตศาสตร์นั้น ผลเฉลยของตัวจำแนกประเภทจะอยู่ในรูปของต้นไม้จำแนกประเภท ดังในรูปที่ 3.1 (ก) เป็นต้นไม้จำแนกประเภทที่แทนสมการที่ 3.1

$$\frac{x \times x}{1+x} \tag{3.1}$$

เพื่อทดสอบประสิทธิภาพของตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์ จึงได้ทำการทดสอบเพื่อเปรียบเทียบกับผลการวิจัยของ Azuaje (2000) ซึ่งเป็นการจำแนกประเภทข้อมูลผู้ป่วยโรคมะเร็งต่อมน้ำเหลืองชนิด DLBCL กับผู้ที่ไม่เป็นโรคมะเร็ง โดยใช้วิธี Simplified Fuzzy ARTMAP ซึ่งเป็นรูปแบบหนึ่งของเครือข่ายใยประสาทเทียม

ข้อมูลที่ใช้ประกอบด้วย 63 ตัวอย่าง แบ่งเป็นข้อมูลที่เป็น DLBCL จำนวน 45 ตัวอย่าง และเป็นข้อมูลปกติจำนวน 18 ตัวอย่าง โดยแต่ละตัวอย่างจะประกอบด้วย 13 คุณลักษณะ จากข้อมูลการแสดงออกของยีน 5 ยีน ที่มีความสัมพันธ์กับการเกิดโรคมะเร็งต่อมน้ำเหลือง ได้แก่ CD10, BCL-6, TTG-2, IRF-4 และ BCL-2 (ข้อมูลจาก http://llmpp.nih.gov/lymphoma)

ในการออกแบบการทดลอง ผลเฉลยแต่ละตัวจะประกอบไปด้วยสัญลักษณ์จากเซตของฟังก์ชัน (*F*) และสัญลักษณ์จากเซตของเทอร์มินอล (*T*) โดยที่เซตของฟังก์ชันจะประกอบด้วยตัวดำเนินการทางคณิตศาสตร์ และเซตของเทอร์มินอลจะประกอบด้วยค่าคงที่และตัวแปรต่างๆ ซึ่งตัวแปรต่าง ๆ เหล่านั้นจะเป็นค่าคุณลักษณะของข้อมูลไมโครอาร์เรย์ ผลเฉลยที่ใช้ในการทดลองถูกกำหนดดังนี้

$$F = \{+, -, \times, \div\}$$

$$T = \{0..9, x_1..x_{13}\}$$

โดยที่ x_1 – x_{13} เป็นค่าที่ได้จากการแสดงออกของยีนต่าง ๆ แสดงดังตารางที่ 3.5 และพารามิเตอร์ที่ใช้สำหรับกำหนดการพันธุกรรม แสดงดังตารางที่ 3.6

การไขว้เปลี่ยนสำหรับตัวจำแนกประเภทกำหนดการพันธุกรรม จะทำการสุ่มผลเฉลยมาครั้งละ 2 ตัว และทำการสุ่มเลือกจุดที่จะใช้ในการไขว้เปลี่ยนจากผลเฉลยทั้งสอง จากนั้นจะทำการสลับโครงสร้างของต้นไม้ย่อย ณ จุดที่สุ่มเลือกได้ระหว่างผลเฉลยทั้งคู่ ถ้าขนาดของผลเฉลยหลังจากการไขว้เปลี่ยนมีขนาดไม่มากกว่าขนาดสูงสุดของต้นไม้ที่กำหนด ก็จะนำไปสร้างเป็นผลเฉลยของประชากรรุ่นใหม่ แต่ถ้าขนาดใหญ่กว่า ก็จะทิ้งต้นไม้นั้นไป ซึ่งจะทำการไขว้เปลี่ยนจนได้ผลเฉลยครบตามที่กำหนด

ในส่วนการกลาย จะเลือกผลเฉลยมาครั้งละ 1 ต้น และสุ่มเลือกวิธีการกลายแบบใดแบบหนึ่งจาก 2 แบบ คือ

ก) *การตัดเล็ม* โดยจะทำการสุ่มเลือกโนดภายในผลเฉลยที่ไม่ใช่โนดใบ ซึ่งจะได้เป็นต้นไม้ย่อย จากนั้นแทนที่ต้นไม้ย่อยที่ได้ด้วยโนดใหม่ที่สุ่มขึ้นมาจากเซตของเทอร์มินอล (0..9, ตัวแปร x ต่าง ๆ) เพื่อสร้างเป็นโนดใบ

ข) *การเปลี่ยนค่า* โดยจะทำการสุ่มเลือกโนดของผลเฉลย และแทนที่โนดนั้นด้วยโนดใหม่ที่เป็นประเภทเดียวกันแบบสุ่ม เช่น ถ้าสุ่มได้โนดที่เป็นฟังก์ชัน ก็จะแทนที่โนดใหม่ด้วยโนดที่สุ่มมาจากเซตของฟังก์ชัน โดยที่โครงสร้างอื่น ๆ ยังเหมือนเดิม

ตารางที่ 3.5 รายละเอียดของตัวแปรที่ใช้ในการทดลอง

ตัวแปร	**ยีน**	Clone_ID
x_1	CD10	200814
x_2	CD10	1286850
x_3	CD10	701606
x_4	BCL-6	712395
x_5	BCL-6	1340526
x_6	TTG-2	712829
x_7	TTG-2	685456
x_8	IRF-4	270770
x_9	IRF-4	1272196
x_{10}	BCL-2	232714
x_{11}	BCL-2	342181
x_{12}	BCL-2	1336385
x_{13}	BCL-2	342181

ตารางที่ 3.6 พารามิเตอร์ที่ใช้สำหรับการสร้างตัวจำแนกประเภทกำหนดการพันธุกรรม

จำนวนประชากร	1,000 ต้น
ขนาดสูงสุดของต้นไม้	500 โนด
จำนวนรุ่นสูงสุด	500 รอบ
อัตราการสืบพันธุ์	10%
อัตราการไขว้เปลี่ยน	80%
อัตราการกลาย	10%
การเลือกผลเฉลย	การเลือกแบบ Tournament ขนาด 10
เงื่อนไขการสิ้นสุด: จำแนกชุดข้อมูลสอนได้ถูกต้อง 100% หรือ ครบตามจำนวนรุ่นสูงสุดที่กำหนด	

ในการวัดประสิทธิภาพของผลเฉลยแต่ละตัว สมการที่ได้จากต้นไม้จำแนกประเภทจะถูกประเมิน ค่าของตัวแปร x_1 – x_{13} จะถูกอ่านมาจากชุดข้อมูลสอนทีละตัว ถ้าผลลัพธ์จากสมการดังกล่าว มีค่ามากกว่า 0 ชุดข้อมูลที่อ่านมานั้นจะถูกจำแนกว่าเป็นข้อมูลที่เป็น DLBCL ถ้าไม่เช่นนั้น จะถูกจำแนกประเภทว่าเป็นข้อมูลปกติ ซึ่งการประเมินนี้จะกระทำกับชุดข้อมูลสอนทุกตัว แล้วนับความแม่นของการจำแนกประเภทบนชุดข้อมูลสอนเพื่อใช้เป็นค่าความเหมาะของผลเฉลยตัวนั้น ๆ นอกจากนี้ ยังได้เพิ่มเทอม 1/ขนาดของผลเฉลย เข้าไปในฟังก์ชันความเหมาะเพื่อให้ได้ผลเฉลยที่มีขนาดเล็ก ฟังก์ชันความเหมาะนิยามได้ดังนี้

$$\textit{ค่าความเหมาะ} = \textit{จำนวนตัวอย่างที่จำแนกประเภทได้ถูกต้อง} + \frac{1}{\textit{ขนาดของผลเฉลย}} \qquad (3.2)$$

3.3 ผลการทดลอง

การวัดประสิทธิภาพของตัวจำแนกประเภท จะใช้วิธี Leave-one-out (Tourassi and Floyd, 1997) ซึ่งจากข้อมูลที่ใช้ในการทดสอบมีทั้งหมด 63 ตัวอย่าง ข้อมูล 62 ตัวอย่างจะถูกใช้เป็นชุดข้อมูลสอน ส่วนข้อมูลที่เหลืออีกหนึ่งตัว จะถูกใช้เป็นตัวทดสอบ โดยจะทำการสลับข้อมูลทดสอบนี้จนครบทั้ง 63 ตัว จากนั้นจะประเมินค่าประสิทธิภาพในรูปของความแม่น ความไว และความจำเพาะ ของการจำแนกประเภทข้อมูลจากชุดข้อมูลทดสอบ ซึ่งนิยามได้ดังสมการ

$$Accuracy = \frac{TP + TN}{N} \qquad (3.3)$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad (3.4)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (3.5)$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

โดยที่ N คือ จำนวนตัวอย่างทดสอบทั้งหมด

TP คือ จำนวนตัวอย่างผู้ป่วยมะเร็งที่จำแนกประเภทว่าเป็นมะเร็งได้ถูกต้อง

TN คือ จำนวนตัวอย่างผู้ป่วยปกติที่จำแนกว่าไม่เป็นมะเร็งได้ถูกต้อง

FP คือ จำนวนตัวอย่างผู้ป่วยปกติที่จำแนกว่าเป็นมะเร็ง

FN คือ จำนวนตัวอย่างผู้ป่วยมะเร็งที่จำแนกว่าไม่เป็นมะเร็ง

ความแม่นเป็นตัวชี้วัดประสิทธิภาพของตัวจำแนกประเภทในการจำแนกประเภทข้อมูลทั้งหมดได้อย่างถูกต้อง ความไวเป็นตัวชี้วัดถึงประสิทธิภาพของการจำแนกประเภทสำหรับข้อมูลมะเร็ง และความจำเพาะเป็นตัวชี้วัดถึงประสิทธิภาพของการจำแนกประเภทสำหรับข้อมูลปกติ

เนื่องจากกำหนดการพันธุกรรมเป็นการทำงานเชิงน่าจะเป็น ทำให้คำตอบที่ได้จากการทดลองในแต่ละรอบจะแตกต่างกัน ดังนั้น การทดลองนี้จะถูกทำซ้ำ 10 รอบและรายงานผลด้วยค่าเฉลี่ยของความแม่น ความไว และความจำเพาะ ซึ่งจะนำผลการทดลองที่ได้มาเปรียบเทียบกับวิธี Simplified Fuzzy ARTMAP ที่รายงานใน Azuaje (2000) โดยเลือกเอากรณีที่มีค่าที่ดีที่สุด (ρ = 0.95) ผลการเปรียบเทียบแสดงดังตารางที่ 3.7 และตัวอย่างกฎการจำแนกประเภทที่ได้จากตัวจำแนกประเภทกำหนดการพันธุกรรม แสดงดังรูปที่ 3.2 และรูปที่ 3.3 แสดงกฎการจำแนกประเภทในรูปของสมการคณิตศาสตร์

ตารางที่ 3.7 ผลการเปรียบเทียบประสิทธิภาพของตัวจำแนกประเภทกำหนดการพันธุกรรมกับตัวจำแนกประเภทที่ได้จากวิธี Simplified Fuzzy ARTMAP จาก Azuaje (2000)

ตัวจำแนกประเภท	ความแม่น (%)	ความไว (%)	ความจำเพาะ (%)
กำหนดการพันธุกรรม	78.72	83.10	67.77
Azuaje 2000	76	82	61

If $(((x_{13} - ((4 - (((x_3 - 1) + (x_2 * x_7)) - x_9)) * x_{12})) + 9) + ((((x_9 * 6) + 9) / (((x_{13} * 6) + (x_{10} / x_6)) + 1)) - x_8)) > 0$

Then

Class 1: DLBCL

Else

Class 2: normal

End If

รูปที่ 3.2 ตัวอย่างกฎการจำแนกประเภทที่ได้จากตัวจำแนกประเภทกำหนดการพันธุกรรม

$$\left[[[\, x13 - [[\, 4 - [[(x3 - 1) + (x2 \cdot x7)] - x9]\,] \cdot x12]\,] + 9] + \left[\left[\frac{[(x9 \cdot 6) + 9]}{\left[\left[(x13 \cdot 6) + \left(\frac{x10}{x6}\right)\right] + 1\right]} \right] - x8 \right]\right]$$

รูปที่ 3.3 ตัวอย่างกฎการจำแนกประเภทในรูปของสมการคณิตศาสตร์

จากผลการทดลองพบว่า ตัวจำแนกประเภทกำหนดการพันธุกรรมนั้น ให้ประสิทธิภาพที่ดีทั้งในรูปของความแม่น ความไว และความจำเพาะ นอกจากนั้น โครงสร้างของกฎก็ไม่ถูกกำหนดตายตัวเหมือนอย่างเครือข่ายใยประสาทเทียม ทำให้อาจหาความสัมพันธ์ใหม่ ๆ ของตัวแปรได้ และกฎที่ได้จะอยู่ในรูปของสมการคณิตศาสตร์ ซึ่งอาจนำไปศึกษา ตีความ เพื่อทำความเข้าใจในกระบวนการของการเกิดโรคหรือการรักษาโรคต่อไป

เมื่อพบว่า ตัวจำแนกประเภทกำหนดการพันธุกรรม ให้ประสิทธิภาพในการจำแนกประเภทข้อมูลที่ดี จึงได้ทำการทดสอบกับชุดข้อมูลทั้ง 8 ชุด ตามที่ได้อธิบายรายละเอียดไว้ในขอบเขตของการวิจัยในบทที่ 1 โดยใช้วิธีการทดสอบแบบ 10-Fold Cross Validation ผลการทดลองแสดงดังตารางที่ 3.8 และได้นำผลการทดลองเปรียบเทียบกับผลที่รายงานใน Cho and Won (2003) จำนวน 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลมะเร็งเม็ดเลือดขาว ชุดข้อมูลมะเร็งลำไส้ใหญ่ และ ชุดข้อมูลมะเร็งต่อมน้ำเหลือง โดยใช้เทคนิค MLP, SASOM, SVM(Linear), SVM(RBF), KNN(Cosine) และ KNN(Pearson) ซึ่งค่าที่นำมาเปรียบเทียบจะเป็นค่าเฉลี่ยของแต่ละวิธี แสดงดังตารางที่ 3.9 โดยค่าที่เป็นตัวเข้มแสดงถึงผลที่ดีกว่าตัวจำแนกประเภทกำหนดการพันธุกรรม ซึ่งจากการเปรียบเทียบก็ยังคงยืนยันว่าตัวจำแนกประเภทกำหนดการพันธุกรรมนั้น ให้ประสิทธิภาพในการจำแนกประเภทข้อมูลอยู่ในระดับที่ดี

3.4 สรุป

กำหนดการพันธุกรรมเป็นวิธีการค้นหาคำตอบที่ได้รับความนิยมในการนำมาแก้ปัญหาต่าง ๆ อย่างกว้างขวาง หนึ่งในปัญหาเหล่านั้น ได้แก่ การจำแนกประเภทข้อมูล ซึ่งสามารถแทนที่ผลเฉลยได้หลากหลายลักษณะ ทั้งในรูปแบบนิพจน์ทางตรรกศาสตร์ นิพจน์ทางคณิตศาสตร์ และรูปแบบผสม

ตารางที่ 3.8 ผลการทดลองจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรม

ชุดข้อมูล	ความแม่น (%)	ความไว (%)	ความจำเพาะ (%)
มะเร็งต่อมน้ำเหลือง	71.27±6.12	70.42±5.71	72.17±10.29
มะเร็งรังไข่	92.33±1.60	94.19±1.40	89.01±4.05
มะเร็งลำไส้ใหญ่	76.61±4.25	62.72±13.17	84.25±6.98
มะเร็งต่อมลูกหมาก	65.48±5.26	62.11±7.85	69.00±6.06
มะเร็งเต้านม	50.51±6.52	39.70±10.31	58.86±5.81
เนื้องอกระบบประสาทส่วนกลาง	54.33±4.59	35.26±14.97	64.61±7.63
มะเร็งเม็ดเลือดขาว	80.41±7.00	87.02±6.53	68.00±9.80
มะเร็งปอด	93.97±1.62	79.03±9.28	97.06±1.45

ตารางที่ 3.9 ผลการเปรียบเทียบค่าความแม่นของตัวจำแนกประเภทกำหนดการพันธุกรรมกับตัวจำแนกประเภทที่ได้จาก Cho and Won (2003)

ชุดข้อมูล / วิธีการ	มะเร็งเม็ดเลือดขาว (%)	มะเร็งลำไส้ใหญ่ (%)	มะเร็งต่อมน้ำเหลือง (%)
MLP	***85.3***	70.1	69.7
SASOM	74.0	62.7	63.4
SVM(Linear)	72.7	66.4	62.9
SVM(RBF)	72.7	66.4	63.4
KNN(Cosine)	***84.5***	72.7	69.1
KNN(Pearson)	***85.3***	***77.4***	***73.1***
Genetic Programming	80.41	76.61	71.27

การจำแนกประเภทข้อมูลไมโครอาร์เรย์ จัดเป็นอีกงานหนึ่งที่นิยมนำกำหนดการพันธุกรรมมาแก้ปัญหา โดยรูปแบบที่นิยม ได้แก่ การสร้างตัวจำแนกกำหนดการพันธุกรรมในรูปนิพจน์ทางคณิตศาสตร์ ซึ่งเหมาะกับข้อมูลไมโครอาร์เรย์ที่เป็นข้อมูลตัวเลข โดยประสิทธิภาพของตัวจำแนก

ประเภทกำหนดการพันธุกรรมอยู่ในเกณฑ์ที่ดีเมื่อเทียบกับตัวจำแนกประเภทแบบอื่น ๆ ในศาสตร์ด้านการเรียนรู้ของเครื่อง

ข้อดีอีกประการหนึ่งของตัวจำแนกประเภทกำหนดการพันธุกรรม คือ ไม่จำเป็นต้องกำหนดโครงสร้างของคำตอบไว้ล่วงหน้า ซึ่งกระบวนการเรียนรู้ของกำหนดการพันธุกรรมจะทำการปรับโครงสร้างเพื่อหาความสัมพันธ์ของตัวแปรในรูปแบบใหม่ ๆ และผลที่ได้จะอยู่ในรูปของสมการทางคณิตศาสตร์ ที่สามารถนำไปศึกษาหาความสัมพันธ์ระหว่างตัวแปรที่มีผลต่อการแบ่งประเภทของข้อมูลต่อไปได้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 4
การเลือกคุณลักษณะสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์

4.1 การเลือกคุณลักษณะแบบ SNR

การเลือกคุณลักษณะแบบ Filter เป็นวิธีการที่มีประสิทธิภาพและรวดเร็ว โดยไม่ขึ้นกับขั้นตอนวิธีการเรียนรู้ ในหลายงานวิจัยได้รายงานตรงกันว่า วิธีการเลือกคุณลักษณะแบบ SNR (ตามสมการที่ 2.5 ในบทที่ 2) ให้ประสิทธิภาพที่ดีเมื่อเทียบกับวิธีการอื่น ๆ ดังนี้

Slonim et al. (2000) ได้ทำการศึกษาหาวิธีการจำแนกประเภทมะเร็งเม็ดเลือดขาวจากข้อมูลไมโครอาร์เรย์ ซึ่งได้ทำการศึกษาวิธีการเลือกคุณลักษณะเพื่อนำมาใช้สำหรับการสร้างตัวจำแนกประเภท โดยได้ทำการเปรียบเทียบประสิทธิภาพของวิธีการ SNR กับวิธีการอื่น ๆ ได้แก่ Pearson correlation coefficient, Euclidean distance, Manhattan และ Battacharyya distance วิธีการที่ใช้สำหรับการจำแนกประเภทคือ การออกเสียงแบบถ่วงน้ำหนัก (Weighted Voting) ของคุณลักษณะแต่ละตัว ซึ่งจากผลการศึกษาก็พบว่า วิธี SNR ให้ประสิทธิภาพของการจำแนกประเภทที่ดีที่สุด และยังพบอีกว่าจำนวนคุณลักษณะที่ใช้ส่งผลกับประสิทธิภาพของการจำแนกประเภทน้อยมาก แต่ประสิทธิภาพของการจำแนกประเภทจะลดลงหากจำนวนคุณลักษณะที่เลือกมากเกินจุด ๆ หนึ่ง ซึ่งในงานวิจัยนี้ได้เลือกใช้คุณลักษณะที่มีค่าคะแนน SNR ดีที่สุดจำนวน 50 คุณลักษณะ

Ryu and Cho (2002) ได้รายงานผลการเปรียบเทียบประสิทธิภาพของวิธีการเลือกคุณลักษณะแบบต่าง ๆ 7 แบบ คือ Pearson, Spearman, Euclidean distance, Cosine coefficient, Information gain, Mutual information และ SNR โดยทำการทดสอบกับปัญหาการจำแนกประเภทมะเร็งเม็ดเลือดขาวจากข้อมูลไมโครอาร์เรย์ โดยทดสอบกับตัวจำแนกประเภท 7 แบบ คือ MLP, KNN, SVM_{linear}, SVM_{RBF}, KNN_{cosine}, SOM และ DT ซึ่งจากผลการทดสอบพบว่า SNR ให้ค่าเฉลี่ยของประสิทธิภาพที่ดีที่สุด (ในบทความดังกล่าวไม่ได้รายงานจำนวนคุณลักษณะที่ใช้ในแต่ละวิธีการ) ผลการเปรียบเทียบแสดงดังตารางที่ 4.1

Hong and Cho (2004) ได้ทำการทดลองสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับการจำแนกประเภทมะเร็งต่อมน้ำเหลืองจากข้อมูลไมโครอาร์เรย์ โดยใช้วิธีการเลือกคุณลักษณะแบบ SNR จำนวนคุณลักษณะที่ใช้ คือ 30 คุณลักษณะ ซึ่งจากผลการทดลองประสิทธิภาพของวิธีการจำแนกประเภทดังกล่าวอยู่ในระดับที่ดี เมื่อเทียบกับวิธีการจำแนก

ประเภทด้วยเครือข่ายใยประสาทเทียมที่ใช้กับชุดข้อมูลเดียวกัน แต่ในบทความนี้ไม่ได้เปรียบเทียบกับวิธีการเลือกคุณลักษณะแบบอื่น ๆ

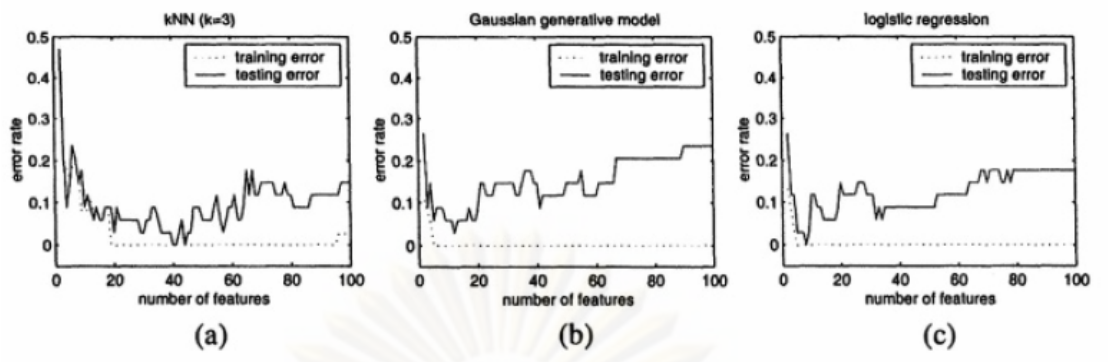
ตารางที่ 4.1 ผลการเปรียบเทียบอัตราความแม่น (%) ของวิธีการเลือกคุณลักษณะและตัวจำแนกประเภทในแบบต่าง ๆ สำหรับปัญหาการจำแนกประเภทมะเร็งเม็ดเลือดขาวจากข้อมูลไมโครอาร์เรย์ที่รายงานใน Ryu and Cho (2002)

Feature / Classifier	Pearson	Spearman	Euclidean distance	Cosine coefficient	Information gain	Mutual information	S/N ratio	Average
MLP	97.1	70.6	97.1	79.4	72.9	62.1	94.1	81.9
KNN	88.2	73.5	82.4	76.5	70.6	58.8	94.1	77.7
SVM_{linear}	97.1	70.6	91.2	70.6	58.8	58.8	94.1	77.3
SVM_{RBF}	97.1	70.6	78.6	70.6	58.8	58.8	94.1	75.5
KNN_{cosine}	91.2	61.8	82.4	70.6	61.8	58.8	97.1	74.8
SOM	74.1	67.4	70.6	70.6	63.8	68.8	97.1	73.2
DT	97.1	61.8	82.4	73.5	47.1	55.9	91.2	72.7
Average	89.6	68.3	82.2	72.8	61.1	61.4	94.7	76.2

จากผลการวิจัยที่ได้กล่าวมาข้างต้นพบว่า SNR ให้ประสิทธิภาพของการเลือกคุณลักษณะสำหรับการจำแนกประเภทข้อมูลที่ดี ดังนั้น ขั้นตอนต่อไปในหัวข้อนี้จะทำการศึกษาเปรียบเทียบประสิทธิภาพของการเลือกคุณลักษณะแบบ SNR สำหรับการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรมที่ได้อธิบายไว้ในบทที่ 3 โดยจะเปรียบเทียบผลการทดลองระหว่างการใช้คุณลักษณะทั้งหมดของชุดข้อมูล (ตารางที่ 3.8) กับผลที่ได้จากการเลือกคุณลักษณะด้วยวิธีการ SNR

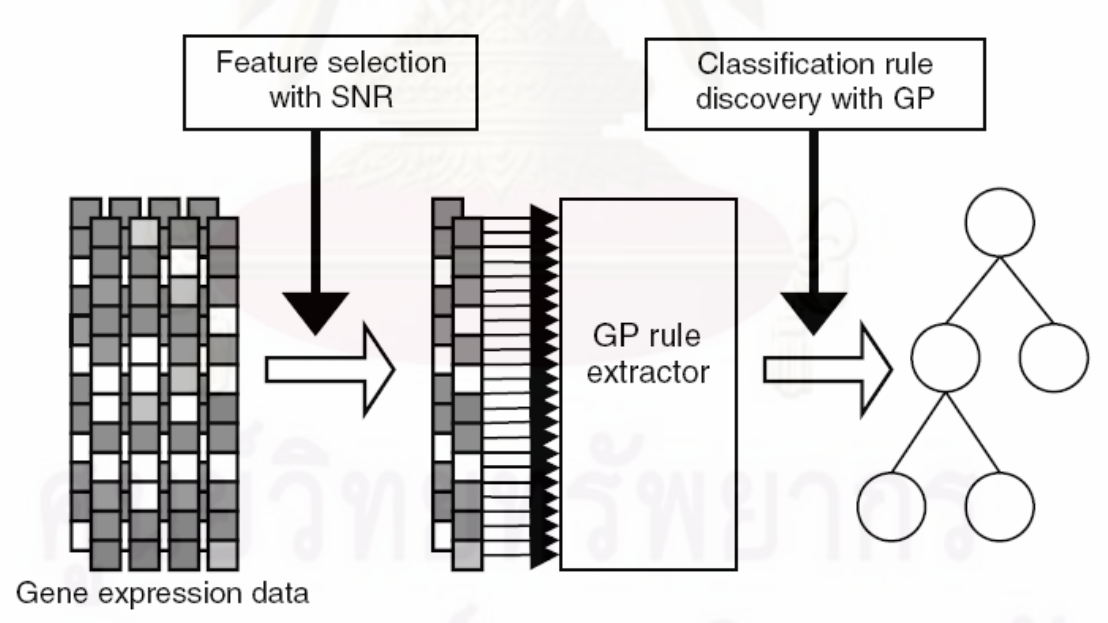
ในปัจจุบันยังไม่สามารถระบุได้ว่าควรจะใช้จำนวนคุณลักษณะเท่าใดจึงจะให้ประสิทธิภาพที่ดีที่สุด จากผลการวิจัยของ Xing (2003) ได้ทำการทดลองเปลี่ยนแปลงจำนวนคุณลักษณะที่ใช้ในการจำแนกประเภทมะเร็งเม็ดเลือดขาวกับตัวจำแนกประเภท 3 วิธี คือ KNN, Gaussian และ Logistic regression จากผลการทดลองสอดคล้องกับผลที่รายงานใน Slonim et al. (2000) กล่าวคือ เมื่อเพิ่มจำนวนคุณลักษณะถึงจุด ๆ หนึ่งแล้ว ประสิทธิภาพของการจำแนกประเภทข้อมูลจะลดลง ซึ่งผลการทดลอง แสดงดังรูปที่ 4.1

จากงานวิจัยทางด้านการเลือกคุณลักษณะของข้อมูลไมโครอาร์เรย์สำหรับการจำแนกประเภทข้อมูล จำนวนคุณลักษณะที่ใช้จะอยู่ในช่วงประมาณ 20 – 50 คุณลักษณะ และมีรายงานใน Cho and Won (2007) ว่า คุณลักษณะประมาณ 25 – 30 คุณลักษณะเป็นจำนวนที่เหมาะสม ดังนั้นในงานวิจัยนี้ จำนวนคุณลักษณะที่ใช้ จะใช้ตาม Hong and Cho (2004) คือใช้ 30 คุณลักษณะ



รูปที่ 4.1 อัตราความผิดพลาดของการจำแนกประเภทข้อมูลด้วยวิธีการ KNN, Gaussian และ Logistic regression โดยการเปลี่ยนแปลงจำนวนคุณลักษณะ (ที่มา: Xing, 2003)

วิธีการที่ใช้เพื่อทดสอบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยกำหนดการพันธุกรรมที่มีการเลือกคุณลักษณะแบบ SNR แสดงดังรูปที่ 4.2 (ซึ่งเป็นวิธีการเดียวกันกับวิธีการที่ใช้ใน Hong and Cho 2004) ผลการเปรียบเทียบแสดงดังตารางที่ 4.2 และตารางที่ 4.3 แสดงผลการทดสอบนัยสำคัญของความแตกต่างของค่าความแม่นด้วยคะแนนที (T-Score) ที่ระดับ 0.05



รูปที่ 4.2 ขั้นตอนการทดสอบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยวิธีกำหนดการพันธุกรรมและการเลือกคุณลักษณะแบบ SNR (ที่มา: Hong and Cho, 2004)

ตารางที่ 4.2 การเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรม ระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะด้วยวิธี SNR ค่าที่เป็นตัวเข้ม แสดงถึงผลที่ดีกว่าสำหรับข้อมูลทดสอบในแต่ละชุด

ชุดข้อมูล		**คุณลักษณะทั้งหมด (%)**	SNR (%)
มะเร็งต่อมน้ำเหลือง	ความแม่น	71.27±6.12	**84.68±3.14**
	ความไว	70.42**±5.71**	**90.83**±6.45
	ความจำเพาะ	72.17±10.29	**78.26±6.15**
มะเร็งรังไข่	ความแม่น	92.33±1.60	**97.74±0.56**
	ความไว	94.19±1.40	**98.02±0.70**
	ความจำเพาะ	89.01±4.05	**97.25±1.19**
มะเร็งลำไส้ใหญ่	ความแม่น	76.61**±4.25**	77.25±4.95
	ความไว	62.72±13.17	**67.27±8.24**
	ความจำเพาะ	**84.25**±6.98	82.75**±5.71**
มะเร็งต่อมลูกหมาก	ความแม่น	65.48±5.26	**78.33±2.38**
	ความไว	62.11±7.85	**78.27±6.79**
	ความจำเพาะ	69.00±6.06	**78.40±5.80**
มะเร็งเต้านม	ความแม่น	50.51±6.52	**62.05±4.99**
	ความไว	39.70±10.31	**47.05±6.04**
	ความจำเพาะ	58.86**±5.81**	**73.63**±6.45
เนื้องอกระบบประสาทส่วนกลาง	ความแม่น	54.33**±4.59**	**55.16**±5.96
	ความไว	**35.26**±14.97	34.76**±13.29**
	ความจำเพาะ	64.61±7.63	**67.15±7.48**
มะเร็งเม็ดเลือดขาว	ความแม่น	**80.41**±7.00	74.30**±2.72**
	ความไว	**87.02**±6.53	77.23**±5.02**
	ความจำเพาะ	68.00±9.80	**68.80±6.48**
มะเร็งปอด	ความแม่น	93.97±1.62	**94.14±1.20**
	ความไว	**79.03**±9.28	77.09**±6.71**
	ความจำเพาะ	97.06±1.45	**97.66±1.01**

ตารางที่ 4.3 ผลการทดสอบนัยสำคัญของความแตกต่างของค่าความแม่นด้วยคะแนนที (T-Score) ที่ระดับ 0.05 ระหว่างวิธีการเลือกคุณลักษณะทั้งหมด กับวิธีการ SNR

ชุดข้อมูล	ผลการทดสอบ
มะเร็งต่อมน้ำเหลือง	วิธี SNR ดีกว่า การใช้คุณลักษณะทั้งหมด อย่างมีนัยสำคัญ
มะเร็งรังไข่	วิธี SNR ดีกว่า การใช้คุณลักษณะทั้งหมด อย่างมีนัยสำคัญ
มะเร็งลำไส้ใหญ่	*ไม่แตกต่างกัน*
มะเร็งต่อมลูกหมาก	วิธี SNR ดีกว่า การใช้คุณลักษณะทั้งหมด อย่างมีนัยสำคัญ
มะเร็งเต้านม	วิธี SNR ดีกว่า การใช้คุณลักษณะทั้งหมด อย่างมีนัยสำคัญ
เนื้องอกระบบประสาทส่วนกลาง	*ไม่แตกต่างกัน*
มะเร็งเม็ดเลือดขาว	***วิธี SNR แย่กว่า การใช้คุณลักษณะทั้งหมด อย่างมีนัยสำคัญ***
มะเร็งปอด	*ไม่แตกต่างกัน*

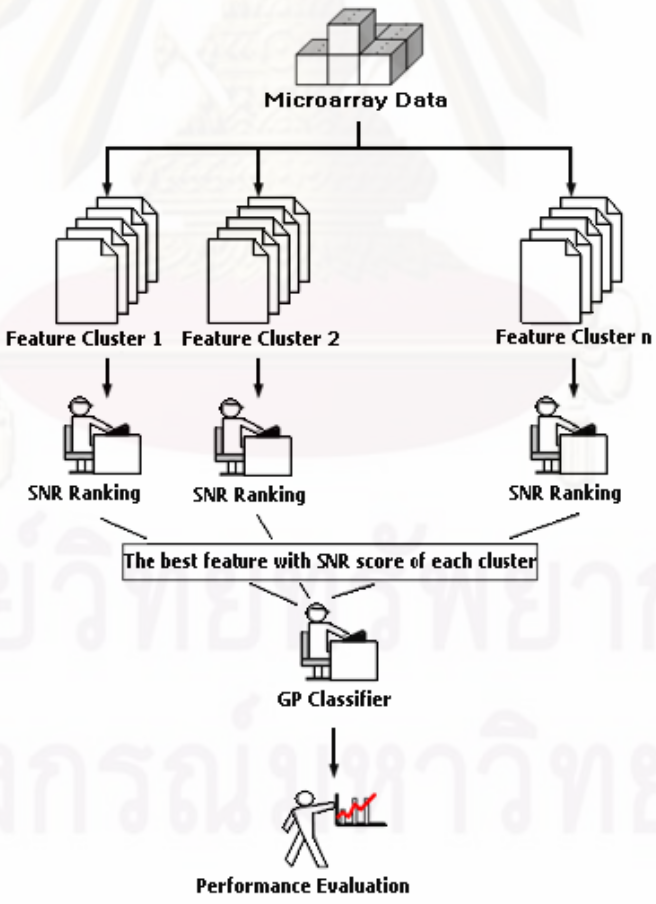
จากผลการทดลองพบว่า การเลือกคุณลักษณะแบบ SNR สามารถเพิ่มประสิทธิภาพให้กับขั้นตอนวิธีการเรียนรู้สำหรับการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรม โดยสามารถเพิ่มประสิทธิภาพให้ดีขึ้นอย่างมีนัยสำคัญทางสถิติจำนวน 4 ชุดข้อมูล จากทั้งหมด 8 ชุดข้อมูล ได้แก่ ชุดข้อมูลมะเร็งต่อมน้ำเหลือง ชุดข้อมูลมะเร็งรังไข่ ชุดข้อมูลมะเร็งต่อมลูกหมาก และ ชุดข้อมูลมะเร็งเต้านม ให้ผลไม่แตกต่างจากการใช้คุณลักษณะทั้งหมด จำนวน 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลมะเร็งลำไส้ใหญ่ ชุดข้อมูลเนื้องอกระบบประสาทส่วนกลาง และ ชุดข้อมูลมะเร็งปอด ซึ่งจากผลการทดลองยังพบว่า มีหนึ่งชุดข้อมูลที่วิธีการเลือกคุณลักษณะแบบ SNR ให้ผลที่แย่กว่าการใช้คุณลักษณะทั้งหมดอย่างมีนัยสำคัญ คือ ชุดข้อมูลมะเร็งเม็ดเลือดขาว

สาเหตุสำคัญที่ทำให้วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพที่แย่กว่าการใช้คุณลักษณะทั้งหมดในบางชุดข้อมูลนั้น อาจเป็นเพราะคุณลักษณะที่ถูกเลือกมาใช้ ไม่เพียงพอต่อการวิเคราะห์ข้อมูล เนื่องจาก คุณลักษณะใด ๆ ที่มีลักษณะคล้ายกัน จะให้ค่า SNR ที่ใกล้เคียงกัน ทำให้คุณสมบัติเหล่านี้จะถูกเลือกไปพร้อม ๆ กัน โดยที่คุณลักษณะเหล่านั้นไม่ได้อธิบายรายละเอียดเพิ่มเติมเกี่ยวกับข้อมูล ซึ่งส่งผลให้ประสิทธิภาพของการวิเคราะห์ข้อมูลลดลง

4.2 การออกแบบการทดลอง

เพื่อแก้ปัญหาที่เกิดจากการเลือกคุณลักษณะที่มีลักษณะคล้ายกันจากการเลือกคุณลักษณะแบบ SNR ผู้วิจัยจึงได้นำเสนอวิธีการเลือกคุณลักษณะแบบ SNR ที่มีการจัดกลุ่มคุณลักษณะ ซึ่งจะขอเรียกวิธีการดังกล่าวว่า ClusSNR โดยจะใช้ข้อดีของเทคนิคการจัดกลุ่มข้อมูล (Clustering) เข้ามาช่วย

ขั้นตอนการเลือกคุณลักษณะแบบ ClusSNR จะเริ่มจากการจัดกลุ่มคุณลักษณะของข้อมูล ซึ่งจะทำให้คุณลักษณะที่มีลักษณะคล้ายกันถูกจัดรวมให้อยู่ในกลุ่มเดียวกัน หลังจากนั้นจะทำการเลือกคุณลักษณะที่มีค่า SNR สูงที่สุดจากแต่ละกลุ่มมาใช้ในการวิเคราะห์ข้อมูล เทคนิคการจัดกลุ่มที่ใช้ในการวิจัย ได้แก่ การจัดกลุ่มแบบ K-Means (MacQueen, 1967) ดังแสดงรายละเอียดในบทที่ 2 หัวข้อ 2.4.1 ภาพรวมของขั้นตอนการเลือกคุณลักษณะแบบ ClusSNR แสดงดังรูปที่ 4.3 โดยจำนวนกลุ่มที่ใช้ในการทดลอง เท่ากับ 30 และจำนวนรอบสูงสุดสำหรับการจัดกลุ่มแบบ K-Means คือ 10



รูปที่ 4.3 ภาพรวมของขั้นตอนการเลือกคุณลักษณะแบบ ClusSNR

4.3 ผลการทดลอง

ผลการเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรม ระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะด้วยวิธี SNR และวิธีการ ClusSNR แสดงดังตารางที่ 4.4 และ ตารางที่ 4.5 แสดงผลการทดสอบนัยสำคัญของความแตกต่างของค่าความแม่นด้วยคะแนนที ที่ระดับ 0.05

จากผลการทดลองพบว่า การเลือกคุณลักษณะแบบ ClusSNR จะให้ประสิทธิภาพที่ดีกว่าการใช้คุณลักษณะทั้งหมดอยากมีนัยสำคัญจำนวน 6 ชุดข้อมูล และไม่มีความแตกต่างกัน จำนวน 2 ชุดข้อมูล ซึ่งไม่มีชุดข้อมูลใดเลยที่วิธีการแบบ ClusSNR ให้ประสิทธิภาพที่แย่กว่า และเมื่อเปรียบเทียบวิธีการ ClusSNR กับวิธีการ SNR พบว่า วิธีการ ClusSNR ให้ประสิทธิภาพที่ดีกว่าวิธีการ SNR จำนวน 2 ชุดข้อมูล (มะเร็งเม็ดเลือดขาว และ มะเร็งปอด) และมี 1 ชุดข้อมูล (มะเร็งต่อมน้ำเหลือง) ที่วิธีการ ClusSNR ให้ประสิทธิภาพที่แย่กว่าวิธี SNR ส่วนที่เหลือ ประสิทธิภาพไม่แตกต่างกัน

4.4 สรุป

การวิเคราะห์ข้อมูลที่มีมิติจำนวนมากเป็นเรื่องที่ซับซ้อน และอาจทำให้ประสิทธิภาพของการวิเคราะห์ข้อมูลลดลง วิธีการหนึ่งสำหรับการเพิ่มประสิทธิภาพของการวิเคราะห์ข้อมูลที่มีมิติจำนวนมากได้แก่ การเลือกเอาเฉพาะคุณลักษณะที่สำคัญมาใช้ในการวิเคราะห์

รูปแบบการเลือกคุณลักษณะประกอบด้วย 3 รูปแบบ ได้แก่ การเลือกคุณลักษณะแบบ Embedded ซึ่งเป็นการเลือกคุณลักษณะโดยขั้นตอนวิธีของการเรียนรู้ การเลือกคุณลักษณะแบบ Wrapper เป็นการเลือกเซตย่อยของคุณลักษณะที่เหมาะกับวิธีการเรียนรู้แบบใดแบบหนึ่ง และการเลือกคุณลักษณะแบบ Filter ซึ่งเป็นการวัดประสิทธิภาพของคุณลักษณะแต่ละตัวด้วยวิธีการทางสถิติ โดยไม่ขึ้นกับขั้นตอนวิธีการเรียนรู้ใด ๆ

การเลือกคุณลักษณะแบบ Filter เป็นเทคนิคที่มีความรวดเร็วและมีประสิทธิภาพ ซึ่งในหลายงานวิจัยได้รายงานตรงกันว่า วิธีการเลือกคุณลักษณะแบบอัตราส่วน Signal-to-Noise (SNR) ให้ประสิทธิภาพในการวิเคราะห์ข้อมูลที่ดี แต่วิธีการดังกล่าวอาจทำให้การเลือกคุณลักษณะมีความซ้ำซ้อน ไม่ได้ข้อมูลใหม่ ๆ มาใช้ในการวิเคราะห์ข้อมูล เนื่องจากคุณลักษณะที่มีลักษณะคล้ายกันจะให้ค่า SNR ที่ใกล้เคียงกัน

ตารางที่ 4.4 การเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยตัวจำแนกประเภทกำหนดการพันธุกรรม ระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะด้วยวิธี SNR และวิธีการ ClusSNR

ชุดข้อมูล		ทั้งหมด (%)	SNR (%)	ClusSNR (%)
มะเร็งต่อมน้ำเหลือง	ความแม่น	71.27±6.12	**84.68±3.14**	79.31±6.81
	ความไว	70.42**±5.71**	**90.83**±6.45	83.75±6.04
	ความจำเพาะ	72.17±10.29	**78.26±6.15**	74.78±11.19
มะเร็งรังไข่	ความแม่น	92.33±1.60	97.74±0.56	**98.10±0.52**
	ความไว	94.19±1.40	**98.02**±0.70	97.90**±0.66**
	ความจำเพาะ	89.01±4.05	97.25±1.19	**98.46±0.93**
มะเร็งลำไส้ใหญ่	ความแม่น	76.61±4.25	77.25±4.95	**80.43±3.02**
	ความไว	62.72±13.17	67.27**±8.24**	**69.54**±8.31
	ความจำเพาะ	84.25±6.98	82.75±5.71	**86.00±3.37**
มะเร็งต่อมลูกหมาก	ความแม่น	65.48±5.26	**78.33±2.38**	77.94±7.71
	ความไว	62.11±7.85	78.27±6.79	**79.61±2.26**
	ความจำเพาะ	69.00±6.06	**78.40**±5.80	76.20**±3.58**
มะเร็งเต้านม	ความแม่น	50.51±6.52	**62.05±4.99**	56.39±6.17
	ความไว	39.70±10.31	**47.05±6.04**	46.17±10.00
	ความจำเพาะ	58.86**±5.81**	**73.63**±6.45	64.31±7.73
เนื้องอกระบบประสาทส่วนกลาง	ความแม่น	54.33**±4.59**	**55.16**±5.96	53.83±5.39
	ความไว	**35.26**±14.97	34.76±13.29	29.04**±8.53**
	ความจำเพาะ	64.61±7.63	67.15**±7.48**	**67.18**±8.18
มะเร็งเม็ดเลือดขาว	ความแม่น	80.41±7.00	74.30**±2.72**	**88.88**±3.21
	ความไว	87.02±6.53	77.23±5.02	**91.70±3.54**
	ความจำเพาะ	68.00±9.80	68.80±6.48	**83.60±6.10**
มะเร็งปอด	ความแม่น	93.97±1.62	94.14±1.20	**96.19**±1.29
	ความไว	79.03±9.28	77.09±6.71	**87.09±6.08**
	ความจำเพาะ	97.06±1.45	97.66**±1.01**	**98.07**±1.24

ตารางที่ 4.5 ผลการทดสอบนัยสำคัญของความแตกต่างของค่าความแม่นด้วยคะแนนที (T-Score) ที่ระดับ 0.05 ระหว่างวิธีการเลือกคุณลักษณะแบบ ClusSNR กับการใช้คุณลักษณะทั้งหมด และวิธีการเลือกคุณลักษณะแบบ SNR

ชุดข้อมูล	**ผลการทดสอบ**	
	ทั้งหมด เทียบกับ ClusSNR	SNR **เทียบกับ** ClusSNR
มะเร็งต่อมน้ำเหลือง	ClusSNR ดีกว่าอย่างมีนัยสำคัญ	***SNR ดีกว่าอย่างมีนัยสำคัญ***
มะเร็งรังไข่	ClusSNR ดีกว่าอย่างมีนัยสำคัญ	*ไม่แตกต่างกัน*
มะเร็งลำไส้ใหญ่	ClusSNR ดีกว่าอย่างมีนัยสำคัญ	*ไม่แตกต่างกัน*
มะเร็งต่อมลูกหมาก	ClusSNR ดีกว่าอย่างมีนัยสำคัญ	*ไม่แตกต่างกัน*
มะเร็งเต้านม	*ไม่แตกต่างกัน*	*ไม่แตกต่างกัน*
เนื้องอกระบบประสาทส่วนกลาง	*ไม่แตกต่างกัน*	*ไม่แตกต่างกัน*
มะเร็งเม็ดเลือดขาว	ClusSNR ดีกว่าอย่างมีนัยสำคัญ	ClusSNR ดีกว่าอย่างมีนัยสำคัญ
มะเร็งปอด	ClusSNR ดีกว่าอย่างมีนัยสำคัญ	ClusSNR ดีกว่าอย่างมีนัยสำคัญ

การแก้ปัญหาการเลือกคุณลักษณะที่มีความซ้ำซ้อน เพื่อให้ได้ข้อมูลใหม่ ๆ ซึ่งจะช่วยให้สามารถวิเคราะห์ข้อมูลได้ดียิ่งขึ้น สามารถทำได้โดยการใช้เทคนิคการจัดกลุ่มคุณลักษณะของข้อมูลเข้ามาช่วย โดยเริ่มต้นคุณลักษณะทั้งหมดจะถูกจัดกลุ่มด้วยเทคนิคการจัดกลุ่มข้อมูล ซึ่งจะทำให้คุณลักษณะที่มีลักษณะคล้าย ๆ กัน จะถูกจัดให้อยู่ในกลุ่มเดียวกัน หลังจากนั้นก็จะหาคุณลักษณะตัวแทนของแต่ละกลุ่มที่มีค่า SNR สูงที่สุด มาใช้เป็นเซตย่อยของคุณลักษณะที่จะนำไปวิเคราะห์ข้อมูลต่อไป ซึ่งวิธีการนี้สามารถรับประกันได้ว่า คุณลักษณะที่ถูกเลือกมานั้นจะไม่มีความซ้ำซ้อนกัน ทำให้สามารถได้ข้อมูลใหม่ ๆ มาใช้ในการวิเคราะห์ข้อมูล ซึ่งจะช่วยให้ประสิทธิภาพของการวิเคราะห์ข้อมูลดีขึ้นด้วย

4.5 ข้อเสนอแนะเพิ่มเติม

การจัดกลุ่มคุณลักษณะที่ใช้ในการเลือกคุณลักษณะแบบ ClusSNR ที่นำเสนอนั้น เป็นการใช้เทคนิคการจัดกลุ่มแบบ K-Means ซึ่งเป็นเทคนิคการจัดกลุ่มแบบพื้นฐานโดยใช้วิธีการวัดระยะห่างแบบยูคลิเดียนเป็นฟังก์ชันวัดความเหมือน ซึ่งอาจทำให้ประสิทธิภาพของกลุ่มที่จัดได้ยังไม่ดีนัก ซึ่งแนวทางการปรับปรุงวิธีการจัดกลุ่มคุณลักษณะ อาจทำได้ดังนี้

ก) การทำข้อมูลให้เป็นปกติ (Normalization) ก่อนนำไปทำการจัดกลุ่มข้อมูล เพื่อลดความผิดปกติของข้อมูล ซึ่งจะทำให้ประสิทธิภาพของการจัดกลุ่มข้อมูลดีขึ้น ซึ่งการทำให้เป็นปกตินั้น อาจใช้วิธีการ ศูนย์หนึ่ง (0-1 Normalization) หรือทำให้เป็นคะแนนแซด (Z-Score)

ข) เนื่องจากการจัดกลุ่มในทางคอลัมน์ ลำดับของแถวจะไม่มีความหมาย ซึ่งวิธีการคำนวณค่าความแตกต่างระหว่างคอลัมน์ ควรจะต้องทำการคำนวณหาผลรวมของความแตกต่างของทุกคู่แถว ซึ่งอาจทำไม่ได้ในเชิงการคำนวณ ดังนั้นวิธีการในการแก้ปัญหา อาจใช้วิธีการเรียงลำดับค่าในแต่ละคอลัมน์แล้วหาผลรวมของความแตกต่างของค่าที่น้อยที่สุดของทั้งสองคอลัมน์ไปจนถึงค่าที่มากที่สุดของทั้งสองคอลัมน์ ซึ่งก็จะทำให้ข้อผิดพลาดของการวัดความเหมือนด้วยยูคลิเดียนลดลง หรืออีกแนวทางหนึ่งอาจใช้วิธีการแทนที่คุณลักษณะของแต่ละคอลัมน์ เช่น การใช้ค่าเฉลี่ยของกลุ่มข้อมูลในแต่ละประเภทมาเป็นตัวแทนสำหรับนำไปใช้วัดความเหมือน เป็นต้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 5
การสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์

5.1 กลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรม

การสร้างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรม ได้รับความสนใจอย่างมากในการเพิ่มประสิทธิภาพของการจำแนกประเภทข้อมูล ซึ่งจากหลายงานวิจัย พบว่าการใช้กลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมให้ประสิทธิภาพของการจำแนกประเภทข้อมูลดีขึ้นเป็นอย่างมาก โดยประสิทธิภาพของการจำแนกประเภทข้อมูลที่ดีขึ้นนั้น จะขึ้นอยู่กับความหลายหลายของสมาชิกในกลุ่มก้อน ในขณะที่สมาชิกแต่ละตัวนั้นจะต้องมีประสิทธิภาพในการจำแนกประเภทข้อมูลที่ดีด้วย ดังนั้นในหลายงานวิจัยจึงพยายามศึกษาหาวิธีการสร้างความหลากหลายให้กับตัวจำแนกประเภทแต่ละตัวสำหรับสมาชิกของกลุ่มก้อนโดยที่สมาชิกแต่ละตัวยังคงมีประสิทธิภาพที่ดี

กำหนดการพันธุกรรม เป็นขั้นตอนวิธีการเรียนรู้แบบไม่เสถียร (Unstable) ซึ่งผลจากการเรียนรู้ในแต่ละครั้งจะไม่เหมือนเดิม ขึ้นอยู่กับการสุ่ม ดังนั้นด้วยขั้นตอนวิธีการเรียนรู้แบบกำหนดการพันธุกรรมเอง ก็สามารถสร้างความหลากหลายให้กับตัวจำแนกประเภทกำหนดการพันธุกรรมแต่ละตัวที่มีประสิทธิภาพที่ดีได้ แต่ก็อาจยังไม่เพียงพอที่จะได้ประสิทธิภาพของการจำแนกประเภทข้อมูลที่ดี ในหัวข้อนี้ จะได้นำเสนอวิธีการสร้างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรมในแบบต่าง ๆ ที่มีรายงานในบทความวิจัย ดังนี้

Iba (1999) ได้ทำการทดสอบประสิทธิภาพของการใช้วิธีการแบบกลุ่มก้อนสำหรับการแก้ปัญหาด้วยวิธีกำหนดการพันธุกรรม โดยได้นำเทคนิค Bagging และ AdaBoost (ดังรายละเอียดที่ได้อธิบายไว้ในบทที่ 2 หัวข้อ 2.6) มาทดสอบกับ 3 ปัญหา ได้แก่ การหาฟังก์ชันที่เหมือนกันทางตรีโกณมิติ (Trigonometric identities) ตามสมการที่ (5.1) ปัญหาลำดับเวลาแบบอลวน (Chaotic time series) ตามสมการที่ (5.2) และปัญหาการเรียนรู้แนวคิดทางตรรกศาสตร์ (Boolean concept) โดยใช้ปัญหาที่ชื่อว่า 6-Multiplexer ตามสมการที่ (5.3) จำนวนกลุ่มก้อนที่ใช้คือ 10 ซึ่งจากผลการทดลอง (แสดงดังตารางที่ 5.1) พบว่า วิธีการแบบ Bagging และ AdaBoost สามารถเพิ่มประสิทธิภาพให้กับวิธีการเรียนรู้แบบกำหนดการพันธุกรรมได้จริง

$$\cos 2x = 1 - \sin^2 x \tag{5.1}$$

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1+x^{10}(t-\tau)} - bx(t) \tag{5.2}$$

$$f(a_0, a_1, d_0, d_1, d_2, d_3) = \bar{a}_0\bar{a}_1 d_0 \vee a_0\bar{a}_1 d_1 \vee \bar{a}_0 a_1 d_2 \vee a_0 a_1 d_3 \tag{5.3}$$

ตารางที่ 5.1 ผลการเปรียบเทียบประสิทธิภาพของวิธีการเรียนรู้แบบกำหนดการพันธุกรรม (GP) แบบปกติกับวิธีการแบบ Bagging (BagGP) และ AdaBoost (BoostGP) ของ Iba (1999)

Problem Name	$\cos(2x)$ (Exp.1) MSE	Chaos (Exp.2) MSE	6-multiplexer (Exp.3) Success Gen.
GP	0.056606	0.000375	5.15
BagGP	0.001051	0.000244	4.80
BoostGP	0.019296	0.000237	4.88

Zhang and Bhattacharyya (2004) ได้ทำการทดลองเปรียบเทียบประสิทธิภาพของการสร้างกลุ่มก้อนของตัวจำแนกประเภท 3 แบบ ได้แก่ กำหนดการพันธุกรรม ต้นไม้ตัดสินใจ และ Logistic regression โดยใช้วิธีการสร้างกลุ่มก้อนแบบ Bagging ด้วยการสร้างเซตย่อยของชุดข้อมูลสอนจำนวน 10 เซต และนำข้อมูลสอนแต่ละเซตไปใช้เพื่อสร้างตัวจำแนกประเภทแต่ละวิธีการข้างต้น และนำตัวจำแนกประเภททั้ง 10 ตัว มารวมกันเป็นกลุ่มก้อนของตัวจำแนกประเภทแต่ละวิธี โดยการรวมคำตอบจะใช้วิธีการออกเสียงข้างมาก (Majority voting) ชุดข้อมูลที่ใช้ในการทดสอบ ได้แก่ U.S. Air Force LAN ซึ่งเป็นชุดข้อมูลเกณฑ์เปรียบเทียบสมรรถนะ จากการแข่งขัน KDD Cup'99 เพื่อทำการจำแนกประเภทข้อมูลของผู้เชื่อมต่อกับเครือข่ายว่าอยู่ในประเภท ปกติ (normal) หรือ โจมตี (attack) ชุดข้อมูลนี้จัดว่าเป็นชุดข้อมูลขนาดใหญ่ (Large-scale dataset) ประกอบไปด้วยชุดข้อมูลสอนจำนวนประมาณ 5 ล้านตัว และชุดข้อมูลทดสอบประมาณ 3 แสนตัว ซึ่งแต่ละตัวประกอบด้วย 41 คุณลักษณะ

จากผลการทดลองของ Zhang and Bhattacharyya (2004) (แสดงดังตารางที่ 5.2) พบว่า ประสิทธิภาพของกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมดีที่สุด เมื่อเทียบกับต้นไม้ตัดสินใจ และ Logistic regression และพบว่า ความหลากหลายที่เกิดขึ้นภายในกลุ่มก้อนของตัวจำแนกปะเภทกำหนดการพันธุกรรมมีมากกว่ากลุ่มก้อนตัวจำแนกประเภทต้นไม้ตัดสินใจ

ดังในตารางที่ 5.3 ซึ่งแสดงคุณลักษณะที่ใช้สำหรับตัวจำแนกประเภทแต่ละตัวภายในกลุ่มก้อน นอกจากนี้ ยังได้มีการศึกษาถึงประสิทธิภาพของจำนวนสมาชิกที่ใช้ในกลุ่มก้อน พบว่า จำนวนสมาชิกที่ใช้ที่เพิ่มขึ้นตั้งแต่ 1 ถึง 10 ตัว จะทำให้ประสิทธิภาพของการจำแนกประเภทข้อมูลเพิ่มขึ้นอย่างมีนัยสำคัญ แต่เมื่อเพิ่มจำนวนสมาชิกมากกว่า 10 ตัวขึ้นไป ประสิทธิภาพแทบจะไม่ปรับเพิ่มขึ้นเลย ผลการทดลองแสดงดังรูปที่ 5.1

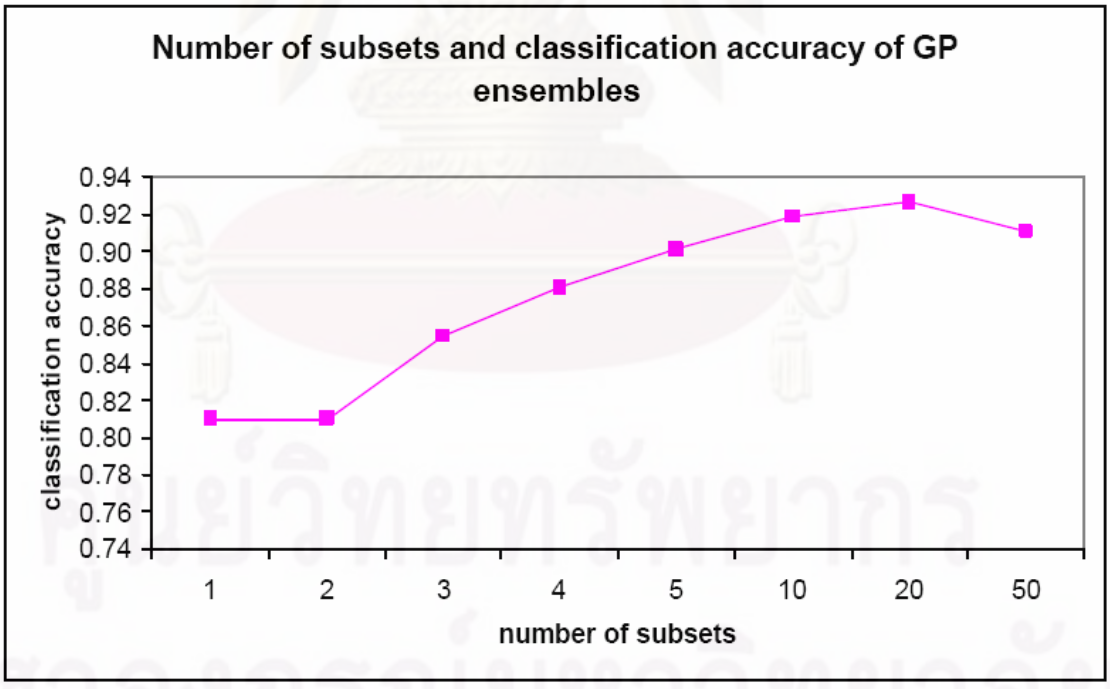
ตารางที่ 5.2 ผลการเปรียบเทียบประสิทธิภาพของกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรม ต้นไม้ตัดสินใจ และ Logistic regression ของ Zhang and Bhattacharyya (2004)

Training Subsets	**Accuracy on Test Data (%):**			**Accuracy on Training Data (%):**		
	GP Classifiers	Decision Tree Classifiers	Logistic Regression Classifiers	GP Classifiers	Decision Tree Classifiers	Logistic Regression Classifiers
S1	90.48	80.52	80. 52	98.88	99.39	79. 89
S2	89.26	80.52	80. 52	91.52	99.50	80. 53
S3	88.81	81.19	80. 52	89.88	99.56	79. 91
S4	90.64	80.52	80. 52	94.51	99.40	80. 36
S5	80.46	80.52	80. 52	99.30	99.60	80. 17
S6	74.64	81.93	80. 52	84.97	99.40	79. 80
S7	19.48	81.32	80. 52	97.13	99.58	80. 16
S8	90.67	80.52	80. 52	98.70	99.26	80. 30
S9	41.15	90.95	80. 52	91.69	99.82	80. 21
S10	91.85	78.66	80. 52	99.03	99.48	79. 59
Standard Deviation	25	3.4	0	4.9	0.15	0.29
Classification Accuracy of Ensemble	90.55	80.52	80. 52			

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 5.3 คุณลักษณะที่ใช้สำหรับตัวจำแนกประเภทกำหนดการพันธุกรรมและต้นไม้ตัดสินใจภายในกลุ่มก้อน (Zhang and Bhattacharyya, 2004)

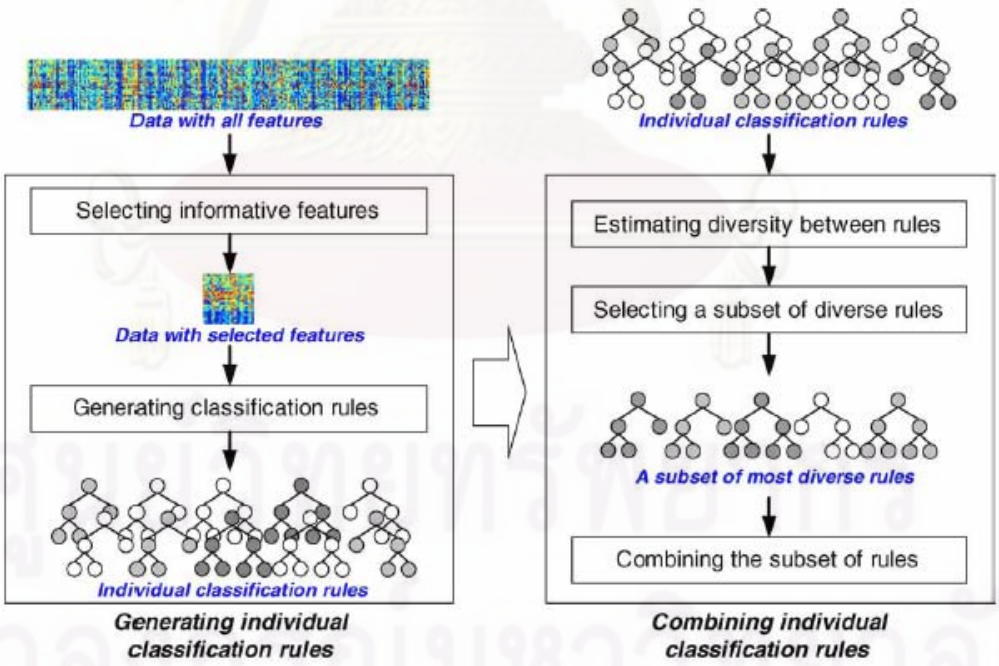
Subsets	GP Trees	Decision Trees
S1	v1, v5, v12, v23, v32, v37	v3, v4, v23, v37
S2	v30, v32	v4, v5, v23, v37
S3	v37	v4, v6, v23, v34, v37
S4	v12, v22, v28	v3, v4, v23, v37
S5	v14, v23, v31, v38, v41	v2, v4, v23, v34, v37
S6	v6, v40	v3, v4, v6, v23, v37
S7	v1, v12, v32	v3, v4, v6, v23, v37
S8	v12, v23	v4, v23, v37
S9	v19, v32, v34	v3, v4, v6, v23, v37
S10	v2, v30	v4, v23, v34, v37



รูปที่ 5.1 การเปรียบเทียบประสิทธิภาพจากการเปลี่ยนแปลงจำนวนสมาชิกของกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรม (Zhang and Bhattacharyya, 2004)

5.2 กลุ่มก้อนตัวจำแนกประเภทสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์

สำหรับปัญหาการจำแนกประเภทข้อมูลไมโครอาร์เรย์ ซึ่งเป็นข้อมูลที่มีจำนวนมิติสูงมาก ในขณะที่จำนวนข้อมูลมีน้อย Hong and Cho (2006) ได้เสนอวิธีการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมโดยแบ่งออกเป็น 2 ขั้นตอน คือ ขั้นที่ 1 การสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมขึ้นมาจำนวนหนึ่ง โดยการใช้วิธีการเลือกคุณลักษณะแบบต่าง ๆ ได้แก่ Pearson correlation, Spearman correlation, Euclidean distance, Cosine coefficient และ SNR เข้ามาเพื่อช่วยเพิ่มประสิทธิภาพของตัวจำแนกประเภทแต่ละตัวและทำให้เกิดความหลากหลาย (จากคุณลักษณะที่ได้จากวิธีการเลือกคุณลักษณะที่ต่างกัน) ภาพรวมแสดงดังรูปที่ 5.2 (ซ้าย) ขั้นที่ 2 ทำการประเมินความหลากหลายของผลเฉลยที่ได้จากขั้นที่ 1 โดยเทียบความแตกต่างที่ได้จากผลเฉลยแต่ละตัว ตามสมการดังรูปที่ 5.3 แล้วเลือกผลเฉลยที่มีความแตกต่างกันมากที่สุดมาตามจำนวนที่กำหนด (ในงานวิจัยนี้ใช้ 5 ตัว) เพื่อสร้างเป็นกลุ่มก้อนของตัวจำแนกประเภท ภาพรวมแสดงดังรูปที่ 5.2 (ขวา) ซึ่งก็พบว่าประสิทธิภาพของการจำแนกประเภทข้อมูลดีขึ้น



รูปที่ 5.2 ภาพรวมของวิธีการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์ที่เสนอโดย Hong and Cho (2006)

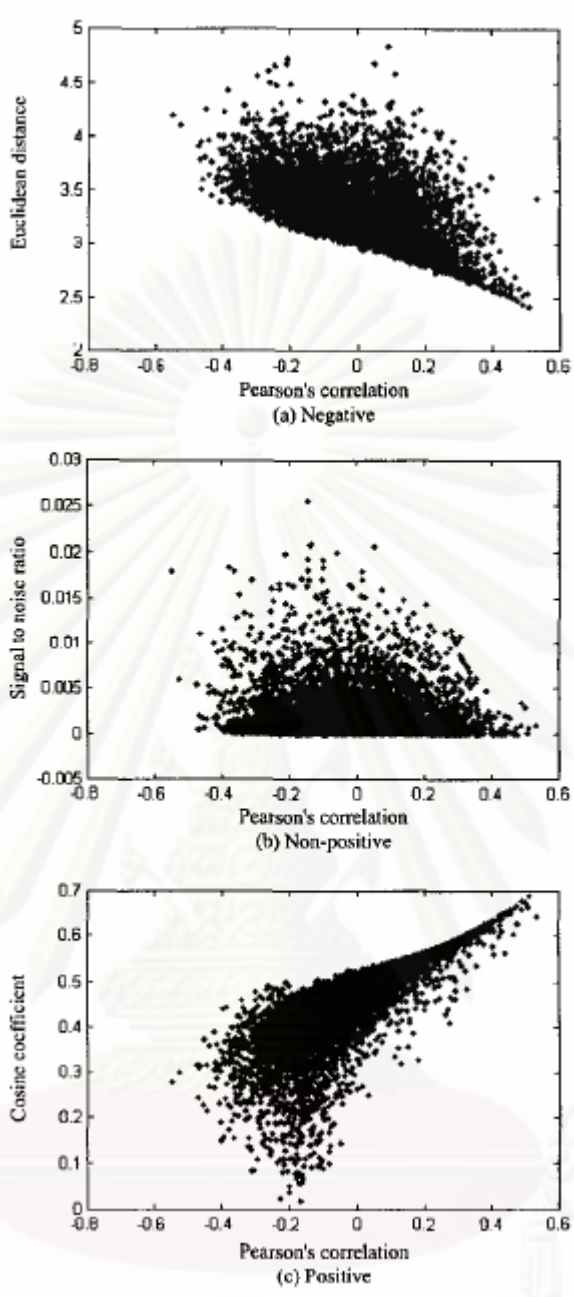
$$\text{distance}(r_i, r_j) = \begin{cases} d(p,q) & \text{if neither} r_i \text{ nor} r_j \text{ have any children,} \\ d(p,q) + \text{distance}(\text{RS of} r_i, \text{RS of} r_j) + \text{distance}(\text{LS of} r_i, \text{LS of} r_j) \\ \text{otherwise}(\text{RS} : \text{right subtree}, \text{LS} : \text{left subtree}) \end{cases}$$

$$\text{where } d(p,q) = \begin{cases} 1 & \text{if } p \text{ and} q \text{ overlap} \\ 0 & \text{if } p \text{ and} q \text{ do not overlap} \end{cases}$$

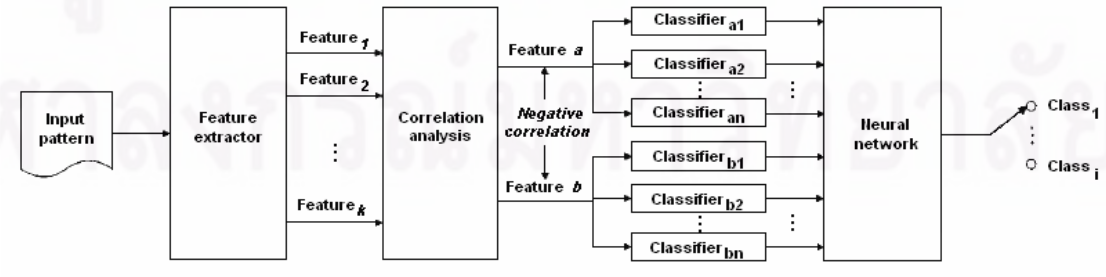
รูปที่ 5.3 สมการที่ใช้คำนวณเพื่อหาความแตกต่างระหว่างผลเฉลย r_i และ r_j ของตัวจำแนกประเภทกำหนดการพันธุกรรมที่ใช้ใน Hong and Cho (2006)

นอกจากนี้ การจำแนกประเภทข้อมูลไมโครอาร์เรย์ ยังได้มีงานวิจัยที่ศึกษาเกี่ยวกับการสร้างกลุ่มก้อนของตัวจำแนกประเภทแบบอื่น ๆ ได้แก่

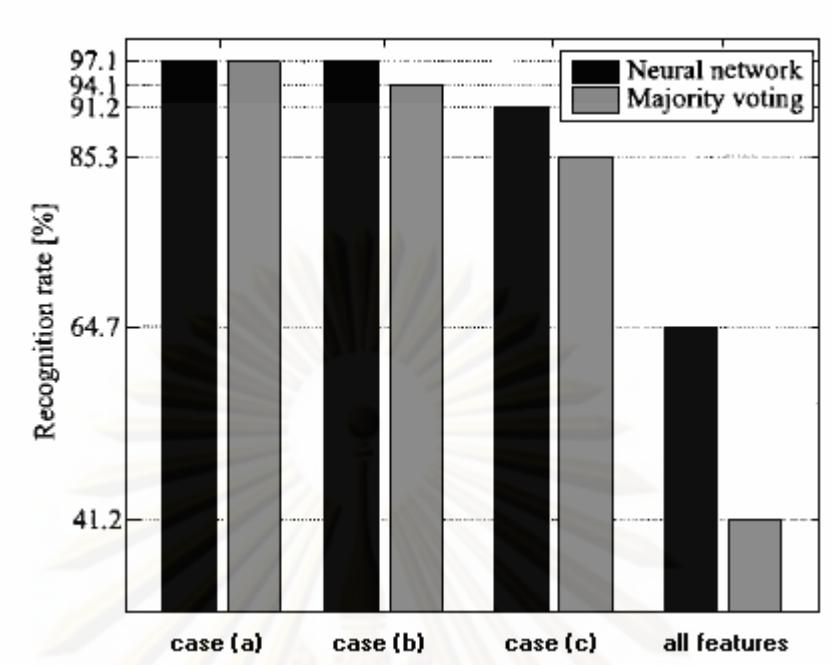
Ryu and Cho (2002) ได้นำเสนอวิธีการสร้างกลุ่มก้อนของตัวจำแนกประเภท โดยใช้วิธีการเรียนรู้หลากหลายแบบ (MLP, SVM และ KNN) ร่วมกับวิธีการเลือกคุณลักษณะแบบต่าง ๆ (Pearson, Spearman, Euclidean, Cosine, Information Gain, Mutual Information และ SNR) เพื่อเพิ่มประสิทธิภาพให้กับตัวจำแนกประเภทแต่ละตัว และให้เกิดความหลากหลาย ซึ่งงานวิจัยนี้ได้พยายามสร้างความหลากหลายให้มากที่สุด โดยการหาวิธีการเลือกคุณลักษณะ 2 วิธี ที่มีความสัมพันธ์กันเชิงลบ กล่าวคือ คุณลักษณะที่มีระดับคะแนนที่สูงจากวิธีการหนึ่ง จะมีระดับคะแนนที่ต่ำสำหรับอีกวิธีการหนึ่ง เพื่อป้องกันไม่ให้เลือกคุณลักษณะที่ซ้ำกัน ซึ่งจากผลการทดลองก็พบว่าวิธีการเลือกคุณลักษณะแบบ Euclidean และ Pearson เป็นวิธีการที่มีความสัมพันธ์กันในเชิงลบ (negative) แสดงดังรูปที่ 5.4 (a) ส่วนรูปที่ 5.4 (b) เป็นความสัมพันธ์แบบไม่ใช่เชิงบวก (non-positive) และรูปที่ 5.4 (c) เป็นความสัมพันธ์ในเชิงบวก (positive) หลังจากนั้น จะนำวิธีการเลือกคุณลักษณะทั้ง 2 แบบมาใช้สร้างเป็นตัวจำแนกประเภทแต่ละวิธี และรวมกันเป็นกลุ่มก้อนโดยใช้เครือข่ายใยประสาทเทียมเป็นตัวรวมผลลัพธ์ที่ได้ ภาพรวมของการทดลองแสดงดังรูปที่ 5.5 ซึ่งผลของการทดสอบการจำแนกประเภทมะเร็งเม็ดเลือดขาวก็พบว่าวิธีการที่นำเสนอให้ประสิทธิภาพที่ดีที่สุด ดังแสดงในรูปที่ 5.6 โดยที่ case (a) จะใช้วิธีการเลือกคุณลักษณะที่มีความสัมพันธ์กันในเชิงลบ case (b) จะใช้วิธีการเลือกคุณลักษณะที่มีความสัมพันธ์แบบไม่ใช่เชิงบวก และ case (c) จะใช้วิธีการเลือกคุณลักษณะที่มีความสัมพันธ์กันเชิงบวก



รูปที่ 5.4 ผลการเปรียบเทียบความสัมพันธ์ของวิธีการเลือกคุณลักษณะของ Ryu and Cho (2002)



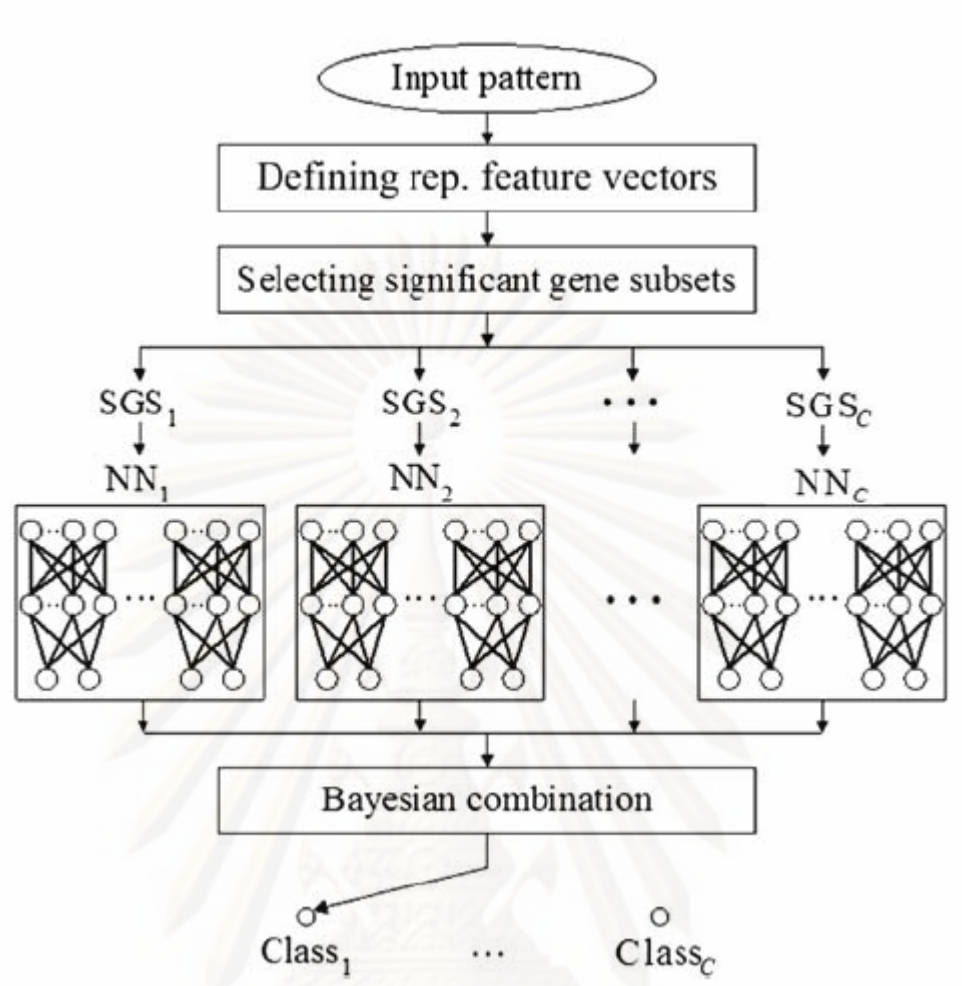
รูปที่ 5.5 ภาพรวมของการสร้างกลุ่มก้อนตัวจำแนกประเภทของ Ryu and Cho (2002)



รูปที่ 5.6 ผลการเปรียบเทียบกลุ่มก้อนตัวจำแนกประเภทของ Ryu and Cho (2002)

ในขณะที่ Kim and Cho (2006) ได้ใช้หลักการเดียวกันกับ Ryu and Cho (2002) โดยใช้วิธีการเรียนรู้ 6 แบบ ได้แก่ MLP, SOM, SASOM, SVM, Decision tree และ KNN ในการสร้างตัวจำแนกประเภท และเปรียบเทียบวิธีการรวมผลลัพธ์ 3 แบบ ได้แก่ การออกเสียงข้างมาก การออกเสียงแบบถ่วงน้ำหนัก และการรวมคำตอบด้วยวิธีการแบบ Bayesian โดยทำการทดลองกับข้อมูลจำนวน 3 ชุดข้อมูล ได้แก่ มะเร็งเม็ดเลือดขาว มะเร็งลำไส้ใหญ่ และมะเร็งต่อมน้ำเหลือง ซึ่งจากผลการทดลองก็พบว่าวิธีการที่นำเสนอให้ประสิทธิภาพที่ดี และพบว่า การรวมคำตอบแบบ Bayesian ให้ประสิทธิภาพที่ดีที่สุด

Cho and Won (2007) ได้นำเสนอวิธีการสร้างกลุ่มก้อนของตัวจำแนกประเภทเครือข่ายใยประสาทเทียม โดยใช้ความหลากหลายของเซตย่อยของคุณลักษณะ ด้วยวิธีการเลือกคุณลักษณะแบบต่าง ๆ แล้วนำเซตย่อยของคุณลักษณะเหล่านั้นไปสอนเครือข่ายใยประสาทเทียมเพื่อสร้างเป็นตัวจำแนกประเภท และสุดท้ายทำการรวมคำตอบที่ได้จากตัวจำแนกประเภทแต่ละตัวด้วยวิธีการแบบ Bayesian ภาพรวมของการทดลองแสดงดังรูปที่ 5.7



รูปที่ 5.7 ภาพรวมวิธีการสร้างกลุ่มก้อนตัวจำแนกประเภทเครือข่ายใยประสาทเทียมของ Cho and Won (2007)

จากวิธีการสร้างกลุ่มก้อนสำหรับตัวจำแนกประเภทกำหนดการพันธุกรรม และกลุ่มก้อนของตัวจำแนกประเภทแบบอื่น ๆ สำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ พบว่าประสิทธิภาพของการจำแนกประเภทข้อมูลของกลุ่มก้อนตัวจำแนกประเภทมาจากหลายปัจจัยดังนี้

1. ประสิทธิภาพของตัวจำแนกประเภทแต่ละตัว ซึ่งสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์แล้ว พบว่า ประสิทธิภาพของตัวจำแนกประเภทนั้นสามารถเพิ่มได้ด้วยวิธีการเลือกคุณลักษณะ (ดังรายละเอียดในบทที่ 4)

2. ความหลากหลายของสมาชิกแต่ละตัวในกลุ่มก้อน ซึ่งสามารถสร้างความหลากหลายได้หลายวิธี ทั้งจากการสร้างความแตกต่างของชุดข้อมูลสอน หรือการใช้คุณลักษณะที่แตกต่างกันสำหรับข้อมูลที่มีคุณลักษณะจำนวนมากอย่างข้อมูลไมโครอาร์เรย์

3. วิธีการรวมผลลัพธ์ของกลุ่มก้อน

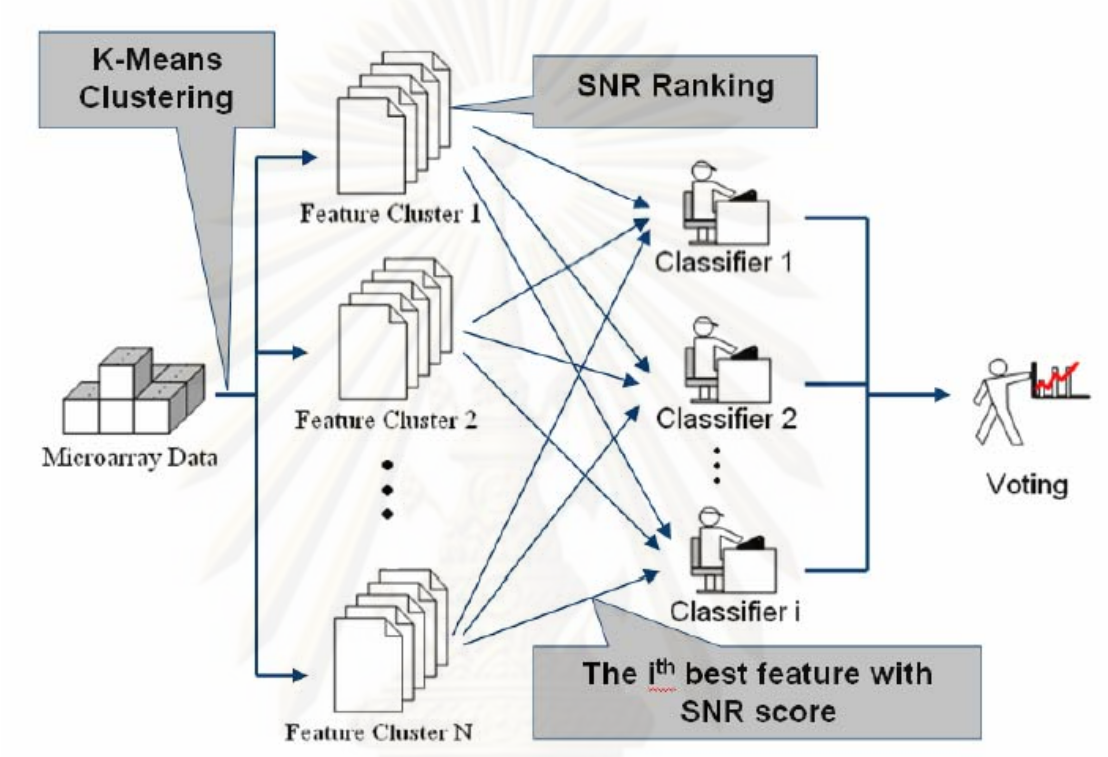
5.3 การออกแบบการทดลอง

ประสิทธิภาพของกลุ่มก้อนสำหรับการจำแนกประเภทข้อมูลนั้น ปัจจัยที่สำคัญได้แก่ความหลากหลายของสมาชิกในกลุ่มก้อน รวมถึงประสิทธิภาพของสมาชิกแต่ละตัว ซึ่งทั้ง 2 ปัจจัยนี้มักเป็นเรื่องที่สวนทางกัน กล่าวคือ ถ้าต้องการให้สมาชิกแต่ละตัวมีประสิทธิภาพในการจำแนกประเภทข้อมูลที่ดี ความหลากหลายของสมาชิกมักจะมีน้อย แต่ถ้าต้องการได้ความหลากหลายที่มาก ประสิทธิภาพของสมาชิกแต่ละตัวก็อาจน้อยลง ทำให้ประสิทธิภาพโดยรวมของกลุ่มก้อนยังไม่ดีพอตามไปด้วย

จากบทที่ 4 เราพบว่า ประสิทธิภาพของตัวจำแนกประเภทสามารถเพิ่มขึ้นได้ด้วยการใช้วิธีการเลือกคุณลักษณะเข้ามาช่วยเพื่อเลือกเอาเฉพาะคุณลักษณะที่สำคัญมาใช้ในการสร้างตัวจำแนกประเภท ซึ่งผู้วิจัยได้นำเสนอวิธีการเลือกคุณลักษณะแบบใหม่ คือ วิธี ClusSNR ซึ่งวิธีการนี้จะทำการจัดกลุ่มคุณลักษณะด้วยเทคนิคการจัดกลุ่มข้อมูล ส่งผลให้คุณลักษณะที่มีลักษณะคล้ายกัน จะถูกจัดให้อยู่ในกลุ่มเดียวกัน จากนั้นจึงเลือกคุณลักษณะที่ให้ค่า SNR สูงที่สุดในแต่ละกลุ่ม มาใช้ในการสร้างตัวจำแนกประเภทข้อมูลต่อไป โดยวิธีการนี้ก็จะสามารถลดการเลือกคุณลักษณะที่ซ้ำซ้อน ซึ่งไม่ก่อให้เกิดประโยชน์สำหรับขั้นตอนวิธีของการเรียนรู้

เพื่อเป็นการสร้างความหลากหลายและเพิ่มประสิทธิภาพของตัวจำแนกประเภทข้อมูลแต่ละตัว จึงได้นำวิธีการ ClusSNR มาใช้สำหรับการสร้างสมาชิกของกลุ่มก้อน โดยวิธีการจะเริ่มจากการกำหนดจำนวนสมาชิกของกลุ่มก้อน จากนั้นจึงใช้เทคนิค ClusSNR เพื่อเลือกคุณลักษณะที่ให้ค่า SNR สูงที่สุดในแต่ละกลุ่มมาใช้สร้างตัวจำแนกประเภทข้อมูล โดยตัวจำแนกประเภทตัวที่ 1 จะนำเอาคุณลักษณะที่มีค่า SNR สูงที่สุดอันดับที่ 1 ของแต่ละกลุ่มมาใช้ในการสร้าง ส่วนตัวจำแนกประเภทตัวที่ 2 ก็จะนำคุณลักษณะที่มีค่า SNR สูงที่สุดอันดับที่ 2 ของแต่ละกลุ่มมาใช้ ทำอย่างนี้ไปเรื่อย ๆ จนกระทั่งครบทุกตัวตามจำนวนสมาชิกของกลุ่มก้อนที่ได้กำหนดไว้ ภาพรวมของวิธีการที่นำเสนอ แสดงดังรูปที่ 5.8

วิธีการนี้สามารถรับประกันได้ว่า ตัวจำแนกประเภทแต่ละตัวจะใช้คุณลักษณะที่แตกต่างกันโดยสิ้นเชิง แต่ยังคงประสิทธิภาพที่ดี ซึ่งลักษณะเช่นนี้อาจเรียกได้ว่า ตัวจำแนกประเภทข้อมูลแต่ละตัวมีการมองต่างมุม (Difference point of view) ซึ่งก็จะช่วยลดจุดด้อยของตัวจำแนกประเภทแต่ละตัว ทำให้ประสิทธิภาพโดยรวมของกลุ่มก้อนดีขึ้น



รูปที่ 5.8 ภาพรวมของวิธีการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับข้อมูลไมโครอาร์เรย์แบบใหม่

หลังจากสร้างสมาชิกของกลุ่มก้อนได้ครบตามจำนวนที่กำหนดแล้ว คำตอบที่ได้จากกลุ่มก้อนจะให้วิธีการออกเสียงแบบถ่วงน้ำหนัก โดยน้ำหนักที่ใช้ คำนวณได้ดังสมการ

$$w_i = \frac{T_i \times A_i}{\sum_i (T_i \times A_i)} \qquad (5.4)$$

โดยที่ w_i คือ น้ำหนักคะแนนเสียงของตัวจำแนกประเภทตัวที่ i

T_i คือ ผลรวมคะแนน SNR ของเซตย่อยของคุณลักษณะที่ใช้สำหรับตัวจำแนกประเภทตัวที่ i

A_i คือ ค่าความแม่นบนชุดข้อมูลสอนสำหรับตัวจำแนกประเภทตัวที่ i

สรุปขั้นตอนวิธีทั้งหมดที่นำเสนอสำหรับการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรม ในที่นี้ขอเรียกวิธีการดังกล่าวว่า EnsClusSNR แสดงได้ดังรูปที่ 5.9

ขั้นตอนวิธี *EnsClusSNR*

1. จัดกลุ่มคุณลักษณะของข้อมูลทั้งหมดด้วยวิธีการจัดกลุ่มแบบ K-Means (ในการทดลองนี้กำหนดให้ K=30 และจำนวนรอบสูงสุดสำหรับ K-Means เท่ากับ 10) ตามรายละเอียดดังรูปที่ 2.16

2. คำนวณค่า SNR ตามสมการที่ 2.5 สำหรับคุณลักษณะทุกตัวในแต่ละกลุ่ม

3. สำหรับสมาชิกตัวจำแนกประเภทกำหนดการพันธุกรรม ตัวที่ i

3.1 สร้างเซตย่อยของคุณลักษณะที่จะนำมาใช้สร้างตัวจำแนกประเภทกำหนดการพันธุกรรม โดยเลือกจากคุณลักษณะที่มีค่า SNR สูงที่สุดลำดับที่ i ของแต่ละกลุ่ม แล้วคำนวณค่าผลรวมของ SNR ของเซตย่อยนั้น (T_i)

3.2 นำข้อมูลสอนทั้งหมดโดยใช้คุณลักษณะที่ได้จากข้อ 3.1 มาสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมแล้วคำนวณหาค่าร้อยละของความแม่นของการจำแนกประเภทข้อมูลของตัวจำแนกประเภทตัวนั้นบนชุดข้อมูลสอน (E_i)

4. รวมผลลัพธ์ที่ได้จากตัวจำแนกประเภททุกตัวด้วยวิธีการออกเสียงแบบถ่วงน้ำหนัก ตามสมการที่ 5.4 ซึ่งผลลัพธ์ใดได้คะแนนสูงสุด ก็จะตอบเป็นผลลัพธ์นั้น

รูปที่ 5.9 ขั้นตอนวิธีการ EnsClusSNR

5.4 ผลการทดลอง

ในการประเมินประสิทธิภาพ จะทำการทดสอบโดยการใช้ตัวจำแนกประเภทกำหนดการพันธุกรรม ตามรายละเอียดที่ได้อธิบายไว้ในหัวข้อ 3.3 ซึ่งจะทำการเปรียบเทียบประสิทธิภาพกลุ่มก้อนของวิธีการที่นำเสนอ คือ EnsClusSNR กับวิธีการปกติ 2 วิธี คือ การใช้คุณลักษณะทั้งหมดในการสร้างกลุ่มก้อน ซึ่งขอเรียกวิธีการดังกล่าวว่า EnsAll และวิธีการเลือกคุณลักษณะแบบ SNR เพียงอย่างเดียวในการสร้างกลุ่มก้อน ซึ่งขอเรียกว่า EnsSNR และนอกจากนี้ ยังได้ทำการเปรียบเทียบกับวิธีการสร้างกลุ่มก้อนที่เป็นที่รู้จักกันดีอีก 2 วิธี คือ วิธีการ Bagging ซึ่งขอเรียกว่า EnsBag และวิธีการ AdaBoost ซึ่งขอเรียกว่า EnsBoost

จำนวนสมาชิกของกลุ่มก้อนที่ใช้ในการวิจัยในครั้งนี้คือ 9 ตัว การทดลองจะใช้วิธีวัดประสิทธิภาพแบบ 10-Fold Cross validation โดยจะทำการทดลองกับชุดข้อมูลจำนวน 8 ชุด

ข้อมูล ตามรายละเอียดดังหัวข้อ 1.3 ซึ่งแต่ละชุดข้อมูลจะทำซ้ำ 10 รอบ (ยกเว้นวิธีการ EnsBoostSNR ซึ่งทำซ้ำ 5 รอบ) แล้วใช้ค่าเฉลี่ย ดังนั้นในแต่ละชุดข้อมูลจะใช้การทดลองทั้งสิ้นเท่ากับ 100 การทดลอง ผลการทดลองแสดงดังตารางที่ 5.4 และ ตารางที่ 5.5 แสดงผลการทดสอบนัยสำคัญของความแตกต่างของค่าความแม่นด้วยคะแนนที (T-Score) ที่ระดับ 0.05 ระหว่างวิธีการที่นำเสนอ (EnsClusSNR) กับวิธีการอื่น ๆ

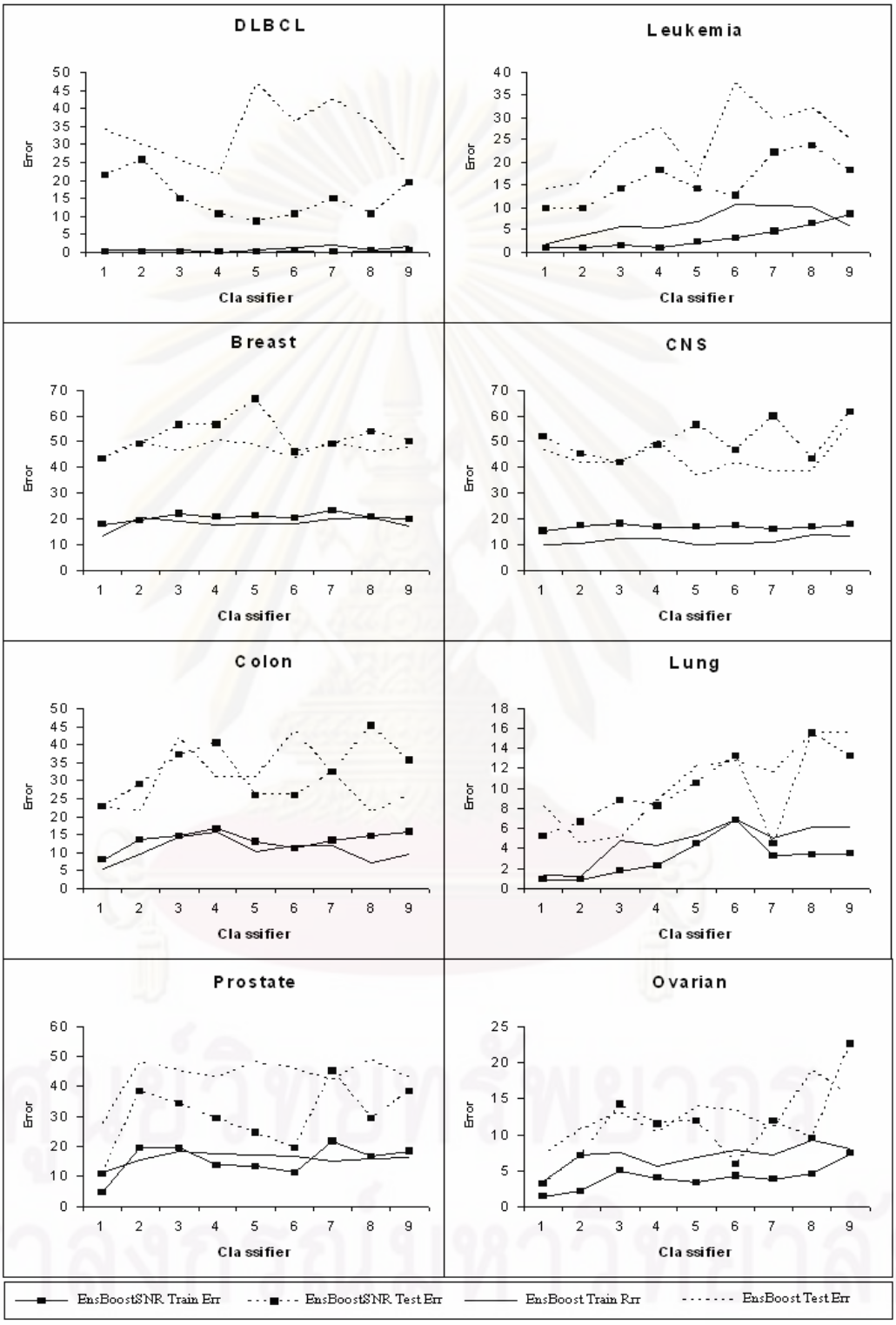
จากผลการทดลองพบว่า ประมาณร้อยละ 80 ประสิทธิภาพของวิธีการที่นำเสนอดีกว่าวิธีการอื่น ๆ อย่างมีนัยสำคัญ ส่วนที่เหลือไม่มีความแตกต่างกันทางสถิติ โดยมีเพียงกรณีเดียวที่วิธีการที่นำเสนอแย่กว่าอย่างมีนัยสำคัญ ซึ่งได้แก่ผลการเปรียบเทียบกับวิธีการ EnsSNR สำหรับชุดข้อมูลมะเร็งเต้านม และนอกจากนี้ยังพบว่าวิธีการที่นำเสนอจะให้ค่าส่วนเบี่ยงเบนมาตรฐานที่ต่ำ โดยร้อยละ 60 ให้ค่าส่วนเบี่ยงเบนมาตรฐานที่ต่ำที่สุด

ในศาสตร์ด้านการเรียนรู้ของเครื่อง เป็นที่ทราบกันดีว่าวิธีการสร้างกลุ่มก้อนแบบ AdaBoost เป็นวิธีการที่มีประสิทธิภาพที่สุดวิธีการหนึ่ง ดังนั้นจึงได้ทำการทดลองเปรียบเทียบกับวิธีการ AdaBoost แบบมีการเลือกคุณลักษณะแบบ SNR ก่อนการสร้างกลุ่มก้อน ซึ่งขอเรียกว่า EnsBoostSNR เพื่อดูประสิทธิภาพของ AdaBoost ผลการทดลองแสดงในตารางที่ 5.4 ซึ่งก็พบว่าประสิทธิภาพของวิธีการที่นำเสนอนั้นยังคงดีกว่า

จากการตรวจสอบพบว่า เหตุที่ประสิทธิภาพของ AdaBoost ไม่ดีนัก เนื่องจากในบางชุดข้อมูล อัตราความผิดพลาดบนชุดข้อมูลสอนเป็นศูนย์ ซึ่งทำให้ AdaBoost ไม่ทำงาน เช่นในชุดข้อมูลมะเร็งต่อมน้ำเหลือง ในขณะที่ในชุดข้อมูลอื่น ๆ พบว่าอัตราความผิดพลาดบนชุดข้อมูลทดสอบของตัวจำแนกประเภทแต่ละตัวมีค่าที่ค่อนข้างสูง ทำให้ประสิทธิภาพโดยรวมของกลุ่มก้อนต่ำลงไปด้วย ผลการตรวจสอบแสดงดังรูปที่ 5.10

จากนั้นได้ทำการเปรียบเทียบประสิทธิภาพกับผลที่ได้รายงานใน Hong and Cho (2006), Kim and Cho (2006) และ Cho and Won (2007) (เฉพาะค่าเฉลี่ยที่ดีที่สุดของแต่ละบทความวิจัย) ที่ได้ทำการทดลองกับชุดข้อมูลเดียวกันจำนวน 5 ชุดข้อมูล ได้แก่ มะเร็งต่อมน้ำเหลือง มะเร็งรังไข่ มะเร็งลำไส้ใหญ่ มะเร็งเม็ดเลือดขาว และ มะเร็งปอด โดยที่ Hong and Cho (2006) จะเป็นการสร้างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรมตามรายละเอียดที่อธิบายไว้ในหัวข้อ 5.2 สำหรับ Kim and Cho (2006) จะเป็นการสร้างกลุ่มก้อนด้วยขั้นตอนวิธีการเรียนรู้ 6 วิธีรวมกับขั้นตอนการเลือกคุณลักษณะ 7 วิธี เพื่อสร้างความหลากหลายของตัวจำแนกประเภท ขณะที่ Cho and Won (2007) จะใช้เครือข่ายใยประสาทเทียม (Artificial Neural Network) เป็น

ขั้นตอนวิธีการเรียนรู้สำหรับการสร้างตัวจำแนกประเภท และใช้วิธีการเลือกคุณลักษณะ 4 วิธีเพื่อสร้างความหลากหลาย ซึ่งผลการเปรียบเทียบแสดงดังตารางที่ 5.6



รูปที่ 5.10 อัตราความผิดพลาดบนชุดข้อมูลสอนและข้อมูลทดสอบของตัวจำแนกประเภทแต่ละตัวสำหรับวิธีการ EnsBoost และ EnsBoostSNR

ตารางที่ 5.4 การเปรียบเทียบประสิทธิภาพของกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมด้วยวิธีการต่าง ๆ

ชุดข้อมูล		EnsAll (%)	EnsSNR (%)	EnsBag (%)	EnsBoost (%)	EnsBoostSNR (%)	EnsClusSNR (%)
มะเร็งต่อมน้ำเหลือง	Acc.	86.38±4.28	91.70±3.54	82.55±4.58	88.50±6.03	91.06**±0.95**	**92.12**±3.62
	Sen.	84.58±6.23	93.33**±5.27**	76.52±6.82	87.08±8.21	**95.00**±8.21	93.75±5.29
	Spe.	88.26±4.61	90.00±5.04	89.13±5.12	90.00±5.44	86.95**±3.08**	**90.43**±4.00
มะเร็งรังไข่	Acc.	96.99±0.86	98.57±0.46	96.79±0.57	97.98±0.63	96.45**±0.29**	**99.21**±0.53
	Sen.	97.90±0.78	98.58±0.59	97.47±0.54	98.02±0.76	96.42±0.80	**99.13±0.52**
	Spe.	95.38±1.86	98.57**±0.59**	95.60±1.47	97.91±1.42	96.48±0.92	**99.34**±0.93
มะเร็งลำไส้ใหญ่	Acc.	86.12±2.54	81.93±5.31	80.16±2.85	79.19±3.84	76.72±8.65	**87.09±1.70**
	Sen.	78.63±6.08	74.09±12.13	66.36±6.14	65.45±6.48	62.73±10.36	**82.27±3.98**
	Spe.	**90.25±1.84**	86.25±3.77	87.75±2.99	86.75±3.13	84.59±8.29	89.75±2.19
มะเร็งต่อมลูกหมาก	Acc.	78.72±3.14	79.39±2.63	64.60±3.35	68.72±3.53	85.68±3.71	**87.15±1.08**
	Sen.	70.19±5.89	78.26±3.63	50.77±5.22	61.53±7.75	83.49±6.21	**85.38±2.07**
	Spe.	87.60±3.63	80.40±4.20	79.00±5.75	76.20±4.47	88.00±4.47	**89.00±1.94**
มะเร็งเต้านม	Acc.	54.48±4.41	**67.35**±3.07	51.41±4.58	53.32±5.26	48.20±4.47	60.76**±2.97**
	Sen.	35.58±7.39	**51.47**±6.69	37.05±8.45	40.29**±4.81**	41.17±8.05	42.94±5.58
	Spe.	69.09**±4.45**	**79.54**±4.79	62.50±4.82	63.41±9.24	53.63±5.47	74.54±5.00
เนื้องอกระบบประสาทส่วนกลาง	Acc.	58.33±6.48	57.33±5.57	54.83±3.88	54.00±6.15	54.33±5.72	**59.83±3.64**
	Sen.	14.76±7.92	21.90±10.81	21.90±11.49	**30.95**±11.05	28.57**±5.83**	16.66±6.04
	Spe.	81.79±8.15	76.41±7.33	72.56±7.16	66.40±9.55	68.20±6.68	**83.07±4.05**
มะเร็งเม็ดเลือดขาว	Acc.	93.33±2.05	79.99±3.02	88.47±3.71	89.02±3.84	93.05±1.97	**96.95±1.58**
	Sen.	96.81±2.07	87.02±4.31	96.81±3.21	94.04±3.99	97.87**±1.51**	**98.72**±1.80
	Spe.	86.80**±2.70**	66.80±6.55	72.80±8.60	79.60±6.92	84.00±2.83	**93.60**±2.80
มะเร็งปอด	Acc.	98.67**±0.70**	96.40±0.91	97.79±1.04	97.62±1.04	96.46±1.27	**99.22**±0.83
	Sen.	94.19±3.33	82.90**±0.91**	88.70±5.54	88.71±4.09	87.09±6.04	**96.45**±3.55
	Spe.	99.60±0.47	99.20±0.53	99.67±0.35	99.46±0.62	98.40±0.36	**99.80±0.32**

หมายเหตุ: Acc. แทนความแม่น Sen แทนความไว และ Spe แทนความจำเพาะ

ตารางที่ 5.5 ผลการทดสอบนัยสำคัญของความแตกต่างของค่าความแม่นด้วยคะแนนที (T-Score) ที่ระดับ 0.05 ระหว่างวิธีการที่นำเสนอ (EnsClusSNR) กับวิธีการอื่น ๆ (Sig หมายถึง วิธีการ EnsClusSNR ดีกว่าอย่างมีนัยสำคัญ, *-Sig* หมายถึง วิธีการอื่นดีกว่าอย่างมีนัยสำคัญ และ *N* หมายถึง ไม่แตกต่างกัน)

ชุดข้อมูล	ผลการทดสอบ			
	EnsAll	EnsSNR	EnsBag	EnsBoost
มะเร็งต่อมน้ำเหลือง	Sig	*N*	Sig	*N*
มะเร็งรังไข่	Sig	Sig	Sig	Sig
มะเร็งลำไส้ใหญ่	*N*	Sig	Sig	Sig
มะเร็งต่อมลูกหมาก	Sig	Sig	Sig	Sig
มะเร็งเต้านม	Sig	*-Sig*	Sig	Sig
เนื้องอกระบบประสาทส่วนกลาง	*N*	*N*	Sig	Sig
มะเร็งเม็ดเลือดขาว	Sig	Sig	Sig	Sig
มะเร็งปอด	*N*	Sig	Sig	Sig

ตารางที่ 5.6 ผลการเปรียบเทียบประสิทธิภาพของวิธีการแบบกลุ่มก้อนที่นำเสนอกับวิธีการที่ตีพิมพ์ในบทความวิจัย

	มะเร็งต่อมน้ำเหลือง	**มะเร็งรังไข่**	**มะเร็งลำไส้ใหญ่**	**มะเร็งเม็ดเลือดขาว**	**มะเร็งปอด**
Hong and Cho (2006)	***97.6***	98.0	-	-	***99.4***
Kim and Cho (2006)	85.2	-	74.8	92.8	-
Cho and Won (2007)	93.0±10.9	-	***87.9***±17.0	95.9±6.4	-
วิธีการที่นำเสนอ	92.1±3.6	***99.2***±0.5	87.0±1.7	***96.9***±1.5	99.2±0.8

ผลการเปรียบเทียบในตารางที่ 5.6 อาจไม่สามารถเปรียบเทียบกันได้โดยตรง เนื่องจากความแตกต่างของวิธีการที่ใช้ เช่น วิธีการทดสอบ บางงานวิจัยอาจใช้วิธีการ Leave – One – Out, 5-Fold Cross validation, 10-Fold Cross validation หรือ การแบ่งชุดข้อมูลสอนกับชุดข้อมูลทดสอบที่ชัดเจน แต่ผลการเปรียบเทียบก็แสดงให้เห็นว่า วิธีการที่นำเสนอนั้นให้ประสิทธิภาพที่ดีสำหรับการสร้างกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรม และผลการทดลองยังแสดง

ให้เห็นว่า ค่าส่วนเบี่ยงเบนมาตรฐานของวิธีการที่นำเสนอมีค่าน้อยมากเมื่อเทียบกับ Cho and Won (2007) ซึ่งเป็นงานเดียวที่มีการรายงานค่าส่วนเบี่ยงเบนมาตรฐานนี้

5.5 **สรุป**

วิธีการแบบกลุ่มก้อนเป็นวิธีที่สามารถเพิ่มประสิทธิภาพของการจำแนกประเภทข้อมูลได้เป็นอย่างดี ประเด็นสำคัญที่ทำให้วิธีการแบบกลุ่มก้อนมีประสิทธิภาพที่ดี คือ ความหลากหลายของสมาชิกในกลุ่มก้อน โดยที่สมาชิกแต่ละตัวจะต้องมีประสิทธิภาพที่ดีด้วย ซึ่งทั้ง 2 ประเด็นนี้มักมีความขัดแย้งกัน

งานวิจัยที่ศึกษาทางด้านวิธีการแบบกลุ่มก้อน จึงมุ่งเน้นที่จะหาวิธีการสร้างความหลากหลายที่ยังคงให้สมาชิกแต่ละตัวมีประสิทธิภาพที่ดี วิธีการที่เป็นที่รู้จักกันดี 2 วิธี ได้แก่ วิธีการ Bagging และ AdaBoost ซึ่งมักให้ประสิทธิภาพที่ดีสำหรับชุดข้อมูลสอนที่มีจำนวนข้อมูลมาก ๆแต่เนื่องจากข้อมูลไมโครอาร์เรย์เป็นข้อมูลที่มีจำนวนคุณลักษณะมากในขณะที่จำนวนข้อมูลมีน้อย ดังนั้นวิธีการเพิ่มประสิทธิภาพของตัวจำแนกประเภทแต่ละตัวจึงนิยมใช้วิธีการเลือกคุณลักษณะมาช่วย ดังนั้นการสร้างความหลากหลายของตัวจำแนกประเภทแต่ละตัวในกลุ่มก้อนจึงพยายามมุ่งเน้นให้เกิดความหลากหลายทางด้านมิติเป็นหลัก โดยหาวิธีการเลือกคุณลักษณะแบบต่าง ๆ มาใช้

ในงานนี้จึงได้นำเสนอวิธีการสร้างกลุ่มก้อนโดยใช้การเลือกคุณลักษณะแบบ ClusSNR ดังรายละเอียดในบทที่ 4 มาประยุกต์ใช้ เนื่องจากวิธีการดังกล่าวจะทำการจัดกลุ่มคุณลักษณะที่มีลักษณะคล้ายกันเข้าไว้ในกลุ่มเดียวกันแล้วทำการจัดลำดับคุณลักษณะของแต่ละกลุ่มตามคะแนน SNR หลังจากนั้นจะทำการสร้างเซตย่อยของคุณลักษณะโดยเลือกเอาคุณลักษณะที่มีคะแนน SNR สูงที่สุดลำดับที่ i ของแต่ละกลุ่มเพื่อนำไปสร้างเป็นตัวจำแนกประเภทกำหนดการพันธุกรรมตัวที่ i ซึ่งวิธีการนี้ ตัวจำแนกประเภทแต่ละตัวจะใช้คุณลักษณะคนละชุดที่ไม่เหมือนกันเลยแต่มีลักษณะที่ใกล้เคียงกัน ทำให้ประสิทธิภาพของตัวจำแนกประเภทแต่ละตัวยังคงมีประสิทธิภาพที่ดี แต่จะเกิดความหลากหลายของสมาชิกในกลุ่มก้อน ซึ่งส่งผลให้ประสิทธิภาพโดยรวมของกลุ่มก้อนดียิ่งขึ้นตามไปด้วย

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

หัวใจสำคัญของการสร้างกลุ่มก้อนที่ดี คือ การสร้างความหลากหลายให้กับสมาชิกแต่ละตัวในกลุ่มก้อนโดยที่สมาชิกแต่ละตัวนั้นจะต้องยังคงมีประสิทธิภาพที่ดีด้วย ซึ่งทั้ง 2 ประเด็นนี้มักมีความขัดแย้งซึ่งกันและกัน กล่าวคือ หากต้องการให้สมาชิกแต่ละตัวของกลุ่มก้อนมีประสิทธิภาพที่ดี ความหลากหลายของสมาชิกในกลุ่มก้อนก็จะลดลง แต่หากต้องการให้ได้ความหลากหลายของสมาชิกให้มากขึ้น ประสิทธิภาพของสมาชิกแต่ละตัวก็มักจะลดลง

งานวิจัยนี้ได้นำเสนอเทคนิควิธีการสร้างกลุ่มก้อนตัวจำแนกประเภทกำหนดการพันธุกรรมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ โดยใช้ข้อดีของเทคนิคการเลือกคุณลักษณะร่วมกับเทคนิคการจัดกลุ่มข้อมูลมาใช้เพื่อเพิ่มประสิทธิภาพของการจำแนกประเภทข้อมูล เนื่องจาก ข้อมูลไมโครอาร์เรย์เป็นข้อมูลที่มีจำนวนมิติ (คุณลักษณะ) ที่สูงมาก ในขณะที่จำนวนข้อมูลมีจำกัด ทำให้วิธีการสร้างกลุ่มก้อนแบบเดิม ๆ ไม่สามารถใช้ประโยชน์จากชุดข้อมูลเหล่านี้ได้มากนัก

จากผลการทดลองในบทที่ 4 พบว่า เทคนิคการเลือกคุณลักษณะแบบ ClusSNR สามารถลดความซ้ำซ้อนของคุณลักษณะที่ถูกเลือก ทำให้คุณลักษณะที่ได้ ให้ข้อมูลใหม่ ๆ ที่เป็นประโยชน์สำหรับการเรียนรู้ ทำให้ประสิทธิภาพของตัวจำแนกประเภทกำหนดการพันธุกรรมดีขึ้น ซึ่งจากเทคนิคดังกล่าวนี้เอง เราพบว่า คุณลักษณะที่ถูกจัดกลุ่มให้อยู่ในกลุ่มเดียวกัน จะมีลักษณะที่คล้ายกัน ดังนั้น หากเราเลือกคุณลักษณะที่มีค่าระดับคะแนนสูงรองลงมา ความสามารถของคุณลักษณะนั้นก็จะมีความใกล้เคียงกับคุณลักษณะที่มีค่าระดับคะแนนที่สูงที่สุดเช่นกัน

ในบทที่ 5 จึงได้นำเทคนิคนี้มาใช้เพื่อสร้างเป็นกลุ่มก้อนของตัวจำแนกประเภทกำหนดการพันธุกรรม โดยจะใช้วิธีการกระจายคุณลักษณะที่ดีของแต่ละกลุ่มให้กับตัวจำแนกประเภทแต่ละตัว โดยที่คุณลักษณะที่ดีในลำดับที่ i ของแต่ละกลุ่มจะถูกนำไปใช้สำหรับการเรียนรู้เพื่อสร้างตัวจำแนกประเภทกำหนดการพันธุกรรมตัวที่ i ซึ่งวิธีการนี้จะรับประกันได้ว่า ตัวจำแนกประเภทแต่ละตัวจะใช้คุณลักษณะที่แตกต่างกันโดยสิ้นเชิง แต่คุณลักษณะทั้งหลายเหล่านั้นจะมีลักษณะบางประการที่คล้ายกัน ทำให้เราสามารถได้ตัวจำแนกประเภทที่มีประสิทธิภาพที่ดีแต่มีความแตกต่างกันในการสร้างเป็นกลุ่มก้อน ซึ่งส่งผลให้ประสิทธิภาพของกลุ่มก้อนดีขึ้นตามไปด้วย

6.2 ปัญหาที่พบและข้อเสนอแนะ

เทคโนโลยีไมโครอาร์เรย์เป็นเทคนิควิธีการที่ใช้สำหรับการศึกษารูปแบบการแสดงออกของยีน ซึ่งมักใช้ศึกษาในเรื่องใดเรื่องหนึ่งทางชีววิทยาหรือทางการแพทย์ เช่น ศึกษารูปแบบการเจริญเติบโตของสิ่งมีชีวิต รูปแบบการตอบสนองต่อยา หรือรูปแบบของการเกิดโรคต่าง ๆ ซึ่งจำเป็นจะต้องมีพื้นความรู้ทางด้านชีววิทยาหรือศาสตร์ที่เกี่ยวข้องพอสมควรจึงจะสามารถออกแบบการทดลอง วิเคราะห์ และสรุปผล เพื่อให้เกิดประโยชน์สูงสุดและสามารถนำไปประยุกต์ใช้ได้จริง

ปัญหาหลักที่สำคัญอีกประการหนึ่ง ก็คือ ข้อมูลไมโครอาร์เรย์มักประกอบไปด้วยข้อมูลที่ขาดหาย (Missing Value) และมีสัญญาณรบกวน เป็นจำนวนมาก ซึ่งจะส่งผลต่อการวิเคราะห์ข้อมูล ดังนั้นก่อนการนำข้อมูลดังกล่าวมาทำการวิเคราะห์ ควรจะมีการจัดเตรียมข้อมูล (Pre-processing) เพื่อให้พร้อมสำหรับการประมวลผล เช่น การเติมเต็มข้อมูลที่ขาดหาย (Imputation) และการทำข้อมูลให้เป็นปกติ ซึ่งมีงานวิจัยหลายชิ้นได้ศึกษาเรื่องดังกล่าว

สำหรับเทคนิควิธีการที่ใช้ในงานวิจัยนี้ ได้ใช้ประโยชน์จากเทคนิคการเลือกคุณลักษณะแบบ SNR และเทคนิคการจัดกลุ่มแบบ K-Means ซึ่งเป็นเทคนิคพื้นฐาน ดังนั้นหากต้องการเพิ่มประสิทธิภาพของวิธีการดังกล่าวอาจมีการศึกษาถึงประสิทธิภาพของเทคนิคการเลือกคุณลักษณะและเทคนิคการจัดกลุ่มข้อมูลในแบบต่าง ๆ เพื่อให้คุณลักษณะที่ถูกเลือกมีประโยชน์มากยิ่งขึ้น

จำนวนคุณลักษณะที่ใช้สำหรับเทคนิคการเลือกคุณลักษณะในปัจจุบันยังไม่ได้มีการศึกษาว่าควรจะใช้จำนวนเท่าใด ซึ่งถ้าหากกำหนดให้น้อยไปอาจไม่เพียงพอต่อการเรียนรู้ แต่หากกำหนดให้มากไป ส่วนที่เกินอาจเป็นสัญญาณรบกวน ซึ่งส่งผลให้ประสิทธิภาพของการเรียนรู้ลดลง ซึ่ง ณ ปัจจุบันมีรายงานในบทความวิจัยเพียงว่า จำนวนคุณลักษณะประมาณ 25 – 30 คุณลักษณะ เพียงพอต่อการเรียนรู้ แต่หากเราสามารถหาจำนวนที่แน่นอนได้ ก็อาจทำให้ประสิทธิภาพโดยรวมดีขึ้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

รายการอ้างอิง

ภาษาไทย

บุญเสริม กิจสิริกุล, ปัญญาประดิษฐ์, เอกสารประกอบคำสอนวิชา 2110654, ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย, 2546.

ยงยุทธ ยุทธวงศ์ และ ศิรศักดิ์ เทพาคำ, จีโนมิกส์ ภาษาแห่งชีวิต, มูลนิธิบัณฑิตยสภาวิทยาศาสตร์และเทคโนโลยีแห่งประเทศไทย, 2545.

ภาษาอังกฤษ

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., J. Hudson J.JR., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Byrd J.C., Botstein D., Brown P.O. and Staudt L.M. Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(2000) : 503-511.

Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proceedings of National Academy of Sciences of the United States of American, 96, 1999, pp. 6745-6750.

Azuaje F. Making Genome Expression Data Meaningful: Prediction and Discovery of Classes of Cancer Through a Connectionist Learning Approach. Proceeding of the IEEE Symposium on Bio-Informatics and Biomedical Engineering, 2000, pp. 208 – 213.

Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.

Bishop C.M., Neural Networks for Pattern Recognition, Clarendon Press, 1995.

Bojarczuk C.C., Lopes H.S. and Freitas A.A. Data Mining with Constrained-Syntax Genetic Programming: Applications in Medical Data Set. Proceeding of Intelligent Data Analysis in Medicine and Pharmacology, 2001.

Brameier M. and Banzhaf W. A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining. IEEE Transactions on Evolutionary Computation 5(2001) : 17-26.

Breiman L. Bagging Predictors. Machine Learning 24(1996) : 123-140.

Cho S.B. and Won H.H. Machine Learning in DNA Microarray Analysis for Cancer Classification. Proceedings of the First Asia-Pacific Bioinformatics Conference, 19, 2003, pp. 189-198.

Cho S.B. and Won H.H. Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence 26(2007) : 243-250.

Cover T. and Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1967) : 21-27.

Eggermont J., Eiben A.E. and van Hemert J.I. A Comparison of Genetic Programming Variants for Data Classification. Proceeding of the Intelligent Data Analysis, 1999, pp. 281-290.

Fidelis M.V., Lopes H.S. and Freitas A.A. Discovering Comprehensible Classification Rules with a Genetic Algorithm. Proceeding of the 2000 Congress on Evolutionary Computation, 2000, pp. 805-810.

Freitas A.A. A Genetic Programming Framework for Two Data Mining Tasks: Classification and Generalized Rule Induction. Proceeding of the 2th Annual Conference on Genetic Programming, Morgan Kaufmann, 1997, pp.96-101.

Freund Y. and Schapire R.E. Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, 1996, pp. 148-156.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D. and Lander E.S. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(1999) : 531-537.

Gordon G.J., Jensen R.V., Hsiao L.L., Gullans S.R., Blumenstock J.E., Ramaswamy S., Richards W.G., Sugarbaker D.J. and Bueno R. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. Cancer Research 62(2002) : 4963-4967.

Gray H.F., Maxwell R.J., Martinez-Perez I., Arus C. and Cerdan S. Genetic Programming for Classification of Brain Tumours from Nuclear Magnetic Resonance Biopsy Spectra. Proceedings of the First Annual Conference in Genetic Programming, 1996, p. 424.

Hengpraprohm S. and Chongstitvatana P. Selective crossover in genetic programming. Proceeding of the International Symposium on Communications and Information Technology, 2001, pp. 534-537.

Hengpraprohm S. and Chongstitvatana P. Diffuse Large B-Cell Lymphoma Classification Using Genetic Programming Classifier. Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2005, pp. 333-338.

Hengpraprohm S. and Chongstitvatana P. Discovering an Optimal Feature Set of Microarray Data for Cancer Classification Using Perceptron Learning Rule with SNR Ranking. Proceeding of the International conference on Software Knowledge Information Management and Applications, 2006, pp. 159-164.

Hengpraprohm S. and Chongstitvatana P. Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier Using K-Means Clustering and SNR Ranking. Proceding of the Frontiers in the Convergence of Bioscience and Information Technologies, 2007, pp. 211-218.

Hengpraprohm S. and Chongstitvatana P. A Genetic Programming Ensemble Approach to Cancer Microarray Data Classification. Proceeding of the 3th International Conference on Innovative Computing Information and Control, 2008, pp.340.

Holland J. Adaptation in Natural and Artificial System. University of Michigan Press, 1975.

Hong J.H. and Cho S.B. Lymphoma Cancer Classification Using Genetic Programming with SNR Features. Proceeding of Genetic Programming: 7th European Conference, 3003, 2004, pp. 78-88.

Hong J.H. and Cho S.B. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. Artificial intelligence in Medicine 36(2006) : 43 - 58.

Iba H. Bagging, boosting, and bloating in genetic programming. Proceeding of the Genetic and Evolutionary Computation Conference, 1999, pp. 1053 – 1060.

Kim K.J. and Cho S.B. Ensemble classifiers based on correlation anlysis for DNA microarray classification. Neurocomputing 70(2006) : 187-199.

Kira K. and Rendell L. The Feature Selection Problem: Traditional Methods and New Algorithm. Proceeding of the Ninth National Conference on Artificial Intelligence, 1992, pp. 129–134.

Kohavi R. and Sommerfield D. Feature Subset Selection Using the Wrapper Method: Overftting and Dynamic Search Space Topology. Proceeding of the First International Conference on Knowledge Discovery and Data Mining, 1995, pp. 192 - 197.

Koza J. Genetic Programming, MIT Press, 1992.

Loveard T. and Ciesielski V. Representing Classification Problems in Genetic Programming. Proceeding of the 2001 Congress on Evolutionary Computation, 2001, pp. 1070-1077.

MacQueen J.B. Some Methods for classification and Analysis of Multivariate Observations. Proceeding of the 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.

Molina L.C., Belanche L. and Nebot A. Feature Selection Algorithms: A Survey and Experimental Evaluation. Proceeding of the International conference on data mining, 2002, pp. 306 - 313.

Muni D.P., Pal N.R. and Das J. A Novel Approach to Design Classifiers Using Genetic Programming. IEEE Transactions on Evolutionary Computation 8(2004) : 183-196.

Petricoin III E.F., Ardekani A.M., Hitt B.A., Levine P.J., Fusaro V.A., Steinberg S.M., Mills G.B., Simone C., Fishman D.A., Kohn E.C. and Liotta L.A. Use of Proteomic Patterns in Serum to Identify Ovarian Cancer. The Lancet 359(2002) : 572-577.

Pomeroy S.L., Tamayo P., Gaasenbeek M., Sturla L.M., Angelo M., McLaughlin M.E., Kim Y.H., Goumnerova L.C., Black P.McL., Lau C., Allen J.C., Zagzag D., Olson J.M., Curran T., Wetmore C., Biegel J.A., Poggio T., Mukherjee S., Rifkin R., Califano A., Stolovitzky G., Louis D.N., Mesirov J.P., Lander E.S. and Golub T.R.

Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. Nature 415(2002) : 436-442.

Quinlan J.R. Induction of decision trees. Machine Learning 1(1986) : 81-106.

Ryu J. and Cho S.B. Gene Expression Classification Using Optimal Feature/Classifier Ensemble with Negative Correlation. Proceeding of the 2002 International Joint Conference on Neural Network, 2002, pp.198 – 203.

Sakanashi H., Higuchi T., Iba H. and Kakazu Y. Evolution of Binary Decision Diagrams for Digital Circuit Design using Genetic Programming. Proceedings of the First International Conference on Evolvable Systems: From Biology to Hardware, Springer-Verlag, 1996, pp. 470–482.

Singh D., Febbo P.G., Ross K., Jackson D.G., Manola J., Ladd C., Tamayo P., Renshaw A.A., D'Amico A.V., Richie J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R. and Sellers W.R. Gene Expression Correlates of Clinical Prostate Cancer Behavior. Cancer Cell 1(2002) : 203-209.

Slonim D.K., Tamayo P., Mesirov J.P., Golub T.R. and Lander E.S. Class Prediction and Discovery Using Gene Expression Data. Proceeding of the 4th Annual International Conference on Computational Molecular Biology, 2000, pp.263 – 272.

Tan A.C. and Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics 2(2003) : S75 – S83.

Tourassi F. and Floyd C. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. Medical Decision Making 17(1997) : 186-192.

Vafaie H. and De Jong K. Genetic Algorithms as a Tool for Feature Selection in Machine Learning. Proceeding of the 4th International Conference on Tools with Artificial Intelligence, 1992, pp. 200-203.

Van't Veer L.J., Dai H., Van De Vijver M.J., He Y.D., Hart A.M., Mao M., Peterse H.L., Van Der Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R. and Friend S.H. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(2002) : 530-536.

Vapnik V. The Nature of Statistical Learning Theory. Springer Verlag, 1995.

Xing E.P. Feature Selection in Microarray Analysis. Understanding and Using Microarray Analysis Techniques: A Practical Guide, Kluwer Academic Publishers, 2003, pp. 110-131.

Zhang Y. and Bhattacharyya S. Genetic programming in classifying large-scale data: an ensemble method. Information Science 163(2004) : 85 - 101.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ประวัติผู้เขียนวิทยานิพนธ์

นายสุพจน์ เฮงพระพรหม เกิดวันที่ 19 กุมภาพันธ์ 2520 ที่จังหวัดกาญจนบุรี เข้ารับการศึกษาในระดับปริญญาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี สถาบันราชภัฏนครปฐม ในปีการศึกษา 2538 และสำเร็จการศึกษา เกียรตินิยมอันดับ 2 ในปีการศึกษา 2541 เข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2542 และสำเร็จในปีการศึกษา 2544 โดยได้รับทุน คุรุทายาท (ระดับอุดมศึกษา) ในระดับปริญญาตรี ต่อเนื่องปริญญาโท ได้รับการบรรจุเข้ารับราชการ ในตำแหน่งอาจารย์ 1 ระดับ 4 ณ มหาวิทยาลัยราชภัฏนครปฐม เมื่อวันที่ 14 ธันวาคม 2544 และได้ลาศึกษาต่อในระดับดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2547

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย