การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น สำหรับข้อมูลการแสดงออกของยีน

นายภัทราวุฒิ แสงศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2554

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

เรื่อง การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นสำหรับข้อมูลการแสดงออกของยีน

โดย นายภัทราวุฒิ แสงศิริ

ได้รับอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

_____ คณบดีบัณฑิตวิทยาลัย
(อาจารย์ ดร.มงคล หวังสถิตย์วงษ์)
14 ตุลาคม 2554

คณะกรรมการสอบวิทยานิพนธ์

_____ ประธานกรรมการ
(ดร.นฤมล รื่นไวย์)

_____ กรรมการ
(อาจารย์ ดร.ศจีมาจ ณ วิเชียร)

_____ กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.พยุง มีสัจ)

_____ กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุพจน์ นิตย์สุวัฒน์)

_____ กรรมการ
(ดร.ชูชาติ หฤไชยะศักดิ์)

การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น สำหรับข้อมูลการแสดงออกของยีน

นายภัทราวุฒิ แสงศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของหลักสูตร
ปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2554

ชื่อ : นายภัทราวุฒิ แสงศิริ
ชื่อวิทยานิพนธ์ : การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นสำหรับข้อมูลการแสดงออกของยีน
สาขาวิชา : เทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : อาจารย์ ดร.ศจีมาจ ณ วิเชียร
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม : ผู้ช่วยศาสตราจารย์ ดร.พยุง มีสัจ
ปีการศึกษา : 2554

บทคัดย่อ

วิธีการเรียนรู้แบบเพิ่มขึ้นเป็นวิธีการที่เหมาะสมและได้รับความนิยมอย่างกว้างขวางในหลายสาขาการวิจัย เนื่องจากมีความสามารถในทำงานกับข้อมูลที่เพิ่มขึ้นอย่างต่อเนื่องโดยไม่ลืมความรู้ที่ได้เรียนรู้มาก่อน ทำให้ลดการสิ้นเปลืองการใช้ทรัพยากรสำหรับประมวลผล อย่างไรก็ตามวิธีการเรียนรู้แบบเพิ่มขึ้นหลายวิธีมักขาดประสิทธิภาพที่ดีพอ เมื่อนำมาใช้ประมวลผลข้อมูลการแสดงออกของยีน ที่ต้องการความสามารถในการจัดกลุ่ม และแสดงโครงสร้างเชิงลำดับชั้นไปพร้อมกัน ดังนั้นงานวิจัยนี้จึงได้นำเสนอขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (Incremental Hierarchical Clustering: IHC) ซึ่งมีพื้นฐานแนวคิดจากอัลกอริธึม COBWEB และ ILM โดยอัลกอริธึมที่นำเสนอนี้จะมีลักษณะของการเรียนรู้แบบเพิ่มขึ้น และมีความสามารถในการจัดกลุ่มข้อมูลเชิงลำดับชั้น พร้อมกับการจำแนกหมวดหมู่ นอกจากนี้ยังได้นำเสนอวิธีการเตรียมข้อมูลและการคัดเลือกมิติข้อมูลการแสดงออกของยีนไปพร้อมกัน

ผลจากการวิจัยโดยใช้ข้อมูลการแสดงออกของยีนจากข้อมูลที่เปิดเผยต่อสาธารณะให้ใช้ได้ทั่วไป จำนวน 4 กลุ่มข้อมูลคือ DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer พบว่า (1) วิธีการเตรียมข้อมูลและการคัดเลือกมิติข้อมูลที่นำเสนอ สามารถค้นหากลุ่มย่อยยีนที่มีอำนาจจำแนกได้จำนวน 37, 39, 48 และ 27 ยีนตามลำดับและ (2) โครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นให้ค่าความถูกต้อง ในการจัดกลุ่มและจำแนกหมวดหมู่คือ 69.17%, 100.00%, 93.33% และ 93.55% ตามลำดับ

(วิทยานิพนธ์มีจำนวนทั้งสิ้น 127 หน้า)

คำสำคัญ : การคัดเลือกมิติข้อมูล การแสดงออกของยีน การจัดกลุ่มข้อมูล การเรียนรู้แบบเพิ่มขึ้น

ศจีมาจ ณ วิเชียร อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Name : Mr.Patharawut Saengsiri

Thesis Title : A Development of an Incremental Hierarchical Clustering Algorithm for Gene Expression Data

Major Field : Information Technology

King Mongkut's University of Technology North Bangkok

Thesis Advisor : Dr. Sageemas Na Wichian

Co-Advisor : Assistant Professor Dr. Phayung Meesad

Academic Year : 2011

Abstract

Incremental learning methods are a suitable and widely popular technique for many research domains. Incremental learning provides the ability to work with continuously increasing data without forgetting prior knowledge. Moreover, this approach also reduces the cost of resources. However, several incremental learning methods produce low accuracy when processing gene expression datasets which not only requires clustering but also the representation of hierarchical structure. Therefore, the objective of this research was to propose an incrementally hierarchical clustering (IHC) algorithm based on COBWEB and ILM algorithm concepts. Thus, the IHC approach to the problem includes incremental learning, hierarchical clustering and classification. Furthermore, data preprocessing framework and new feature selection of gene expression data are also proposed.

The experimental results based on public gene expression dataset DLBCL, Leukemia, Embryonal Tumor, and Colon Cancer showed that (1) data preprocessing framework and new feature selection can find subsets of 37, 39, 48, and 27 discriminate genes, respectively and (2) the incremental hierarchical clustering algorithm generated higher accuracy rates of clusters and classification at 69.17%, 100.00%, 93.33%, and 93.55% respectively.

(Total 127 Pages)

Keywords: Feature Selection, Gene Expression Data, Clustering, Incremental Learning

Sageemas Na Wichian

Advisor

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ประสบความสำเร็จตามวัตถุประสงค์ด้วยคำแนะนำ ความช่วยเหลือ รวมถึงกำลังใจและคำชี้แนะจากท่านอาจารย์ ดร.ศจีมาจ ณ วิเชียร ผู้ช่วยศาสตราจารย์ ดร.พยุง มีสัจ และ Professor Dr.-Ing. habil. Herwig Unger คณะคณิตศาสตร์และวิทยาการคอมพิวเตอร์ Fern Universität in Hagen สหพันธ์สาธารณรัฐเยอรมนี พร้อมทั้งคณะกรรมการสอบวิทยานิพนธ์ ที่ช่วยแนะนำเพิ่มเติมเพื่อให้วิทยานิพนธ์มีความสมบูรณ์มากขึ้น ผู้วิจัยขอกราบขอบพระคุณทุกท่านเป็นอย่างสูงด้วยความซาบซึ้งในความเมตตาและให้โอกาสผู้วิจัยเสมอมา

ขอขอบคุณสถาบันวิจัยวิทยาศาสตร์และเทคโนโลยีแห่งประเทศไทย (วว.) ที่ให้โอกาสผู้วิจัยในการพัฒนาและแสวงหาความรู้ที่เป็นประโยชน์ จนการจัดทำวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบคุณกระทรวงวิทยาศาสตร์และเทคโนโลยี ที่สนับสนุนทุนการศึกษา อบรม และฝึกงานต่างประเทศ เพื่อสร้างเสริมประสบการณ์ในการวิจัยอย่างเต็มที่

ขอขอบคุณศูนย์ประสานงานนักเรียนทุนกระทรวงวิทยาศาสตร์และเทคโนโลยี โดยเฉพาะคุณสมชาย อินจอหอ ผู้คอยให้คำปรึกษาชี้แนะ ดูแล ให้กำลังใจในด้านต่าง ๆ ด้วยดีเสมอมา

ขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือที่สนับสนุนทุนสำหรับการทำวิทยานิพนธ์บางส่วน และทุนสำหรับเผยแพร่ผลงานวิทยานิพนธ์

ขอบคุณอาจารย์สุนันฑา สดสี อาจารย์ประจำคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ และคุณเมธิญาณินฆ์ คำขาว ที่ช่วยให้คำแนะนำ รวมถึงมอบรหัสโปรแกรมสำหรับการพัฒนาต่อยอด ส่งผลให้งานชิ้นนี้สำเร็จตามวัตถุประสงค์

ขอบคุณเพื่อน ๆ นักศึกษาปรัชญาดุษฎีบัณฑิต มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือที่ช่วยเหลือ ให้กำลังใจ และให้คำแนะนำที่เป็นประโยชน์มาโดยตลอด

ขอกราบขอบพระคุณ คุณพ่อบุญช่วย และคุณแม่กิ่งทอง แสงศิริ (บิดา-มารดา) ที่อบรมเลี้ยงดู สั่งสอน ให้กำลังใจ รวมถึงให้คติธรรมที่เป็นแนวทางดำเนินชีวิตจนประสบความสำเร็จ หากงานวิทยานิพนธ์เล่มนี้เกิดประโยชน์ไม่ว่าในส่วนใดก็ตาม ขอให้เป็นถือเป็นพระคุณของท่านทั้งสองที่ได้เสียสละเวลา แรงกาย แรงใจในการอบรมสั่งสอนผู้วิจัยตลอดมา

ภัทราวุฒิ แสงศิริ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

การนำเสนอเนื้อหาในบทนี้เป็นส่วนแรกของวิทยานิพนธ์เป็นส่วนที่ให้ข้อมูลเบื้องต้นเพื่อให้ผู้อ่านเข้าใจถึงที่มา และปัญหาที่เกิดขึ้นเพื่อเป็นแนวทางในการทำวิจัย โดยมีรายละเอียดของเนื้อหาประกอบด้วยหัวข้อย่อยจำนวน 6 หัวข้อ ได้แก่ (1) ความเป็นมาและความสำคัญของปัญหา (2) วัตถุประสงค์การวิจัย (3) สมมติฐานการวิจัย (4) ขอบเขตการวิจัย (5) นิยามศัพท์ และ (6) ประโยชน์ของการวิจัย ดังมีรายละเอียดดังต่อไปนี้

1.1 ความเป็นมาและความสำคัญของปัญหา

จากการค้นพบโครงสร้างของดีเอ็นเอโดยเจมส์ วัตสัน (James Watson) และฟรานซิส คริก (Francis Crick) ในปี พ.ศ. 2496 ทำให้ความจริงเบื้องหลังที่สำคัญของสิ่งมีชีวิตถูกเปิดเผย และสิ่งที่ได้จากการเปิดเผยครั้งนี้ก็คือการค้นพบโครงสร้างของดีเอ็นเอ (DNA) ซึ่งต่อมารู้จักกันดีในชื่อของสารพันธุกรรม (นำชัย, 2546) และส่วนประกอบที่สำคัญของดีเอ็นเอก็คือ ยีน ซึ่งยีนมีอยู่เป็นจำนวนมากในสายดีเอ็นเอ ทั้งนี้ยีนมีหน้าที่สำคัญในการควบคุมและส่งผ่านลักษณะทางพันธุกรรม โดยการแสดงออกของยีน (Gene Expression)

ดีเอ็นเอไมโครอะเรย์ มีประโยชน์ในการเปรียบเทียบการแสดงออกของยีนจากเนื้อเยื่อ 2 ชนิดกับเนื้อเยื่อหลาย ๆ ตัวอย่างที่มาจากสิ่งมีชีวิตชนิดเดียวกัน โดยสามารถทำการทดสอบพร้อมกันได้ในคราวเดียว ซึ่งหลังจากที่นักชีววิทยาทำการทดลองบนดีเอ็นเอไมโครอะเรย์เสร็จแล้ว นักชีววิทยาจะต้องทำการคัดเลือกยีนจากเนื้อเยื่อตัวอย่างที่สนใจบนไมโครอะเรย์ ซึ่งอาจจะมียีนจำนวนหลายพัน หรือหลายหมื่นยีน แล้วนำมาทำการทดสอบทางชีววิทยาต่อไป ความสำคัญของปัญหาคือนักชีววิทยาต้องทำการเลือกหากลุ่มย่อยของยีนที่มีอำนาจจำแนก โดยดูจากระดับการแสดงออกของยีน (Gene Expression Level) จากหลายพันยีนบนไมโครอะเรย์ และกลุ่มย่อยของยีนที่มีความสัมพันธ์กันนั้นต้องมีศักยภาพที่สามารถจะใช้จำแนกหมวดหมู่ของสิ่งที่นักวิจัยต้องการค้นหา (เช่น แยกตัวอย่างเนื้อเยื่อได้ว่าปรกติ หรือเป็นโรค) (Saengsiri *et al.*, 2011 a)

ปัญหาของข้อมูลที่มีมิติเป็นจำนวนมาก เช่น ข้อมูลไมโครอะเรย์ เรียกได้ว่าเป็น ปัญหามิติของข้อมูล (Curse of dimensionality) ซึ่งหมายถึง บางกรณีเมื่อข้อมูลมีจำนวนมิติหรือตัวแปรมาก รวมถึงมิติข้อมูลไม่ได้ไปในทิศทางเดียวกัน ทำให้ข้อมูลเกิดการกระจาย (Data sparse) ซึ่งทำให้บาง

จุดอาจจะไม่มีข้อมูลอยู่เลย อาจจะทำให้เกิดปัญหาในเรื่องของการตีความของกลุ่มข้อมูล ซึ่งถ้ามีมิติของข้อมูลต่ำก็จะทำให้ข้อมูลมีลักษณะของการเกาะกลุ่มกัน ในทางตรงกันข้ามถ้าข้อมูลมีมิติสูงก็อาจจะทำให้ถูกตีความว่าเป็นคนละกลุ่มกันได้ (ภัทราวุฒิและคณะ, 2009)

ปัจจุบันวิธีที่นิยมใช้ เพื่อคัดเลือกมิติของข้อมูลสำหรับกลุ่มข้อมูลไมโครอะเรย์มีอยู่ 2 กลุ่มคือ วิธีการกรอง (Filter approach) และวิธีควบรวม (Wrapper approach) (Ke et al., 2008) ซึ่งวิธีการกรองจะใช้วิธีการเรียงลำดับยีนตามอำนาจจำแนกเฉพาะตัวของแต่ละยีน จากนั้นจะทำการเลือกยีนอันดับต้น ๆ โดยปราศจากการเกี่ยวข้องกับการพึ่งพาความสามารถจากอัลกอริธึม ในทางตรงกันข้ามวิธีการควบรวม ซึ่งจะอาศัยความสามารถของอัลกอริธึมในการคัดเลือกยีน แต่ข้อเสียของวิธีการกรองคือไม่สามารถค้นหาความเกี่ยวข้องระหว่างยีน เช่นเดียวกันวิธีควบรวมก็จะมีปัญหาในเรื่องของเวลาที่ใช้ในการคำนวณที่มาก แต่วิธีการควบรวมให้ความแม่นยำที่มากกว่าวิธีการกรอง

วิธีการจัดกลุ่มเชิงลำดับชั้นแบบเพิ่มขึ้นหรือที่เรียกว่าอัลกอริธึม COBWEB ถูกพัฒนาโดย Fisher ในปี 1987 (Fisher, 1987) อัลกอริธึมนี้มีหลักการทำงานแบบการเรียนรู้เพิ่มขึ้น (Incremental learning) นำเสนอการจัดกลุ่มในลักษณะของโครงสร้างแบบลำดับชั้น (Hierarchical) โดยใช้ฟังก์ชันจำแนกหมวดหมู่ (Category Utility: CU) ที่ได้จากการคำนวณค่าเฉลี่ย (Mean) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ของแต่ละแอททริบิวต์ (Attribute-value Pair) ข้อดีของวิธีนี้คือ เป็นการเรียนรู้แบบเพิ่มขึ้น และสามารถนำมาใช้แสดงแนวคิดการจัดกลุ่มในลักษณะของโครงสร้างลำดับชั้น ส่วนข้อเสียของวิธีนี้คือการคำนวณเพียง CU ซึ่งอาจจะไม่เพียงพอในการจัดกลุ่ม (Michalski and Tecuci, 1994)

อัลกอริธึมการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิส (Incremental Learning based on Mahalanobis Distance: ILM) มีหลักการพื้นฐานมาจากอัลกอริธึม Incremental Learning Fuzzy Neural (ILFN) ซึ่งถูกพัฒนาโดย (Meesad and Yen, 2001) อัลกอริธึม ILM มีความสามารถทั้งในการเรียนรู้แบบมีผู้สอน และการเรียนรู้แบบไม่มีผู้สอน (Supervised and Unsupervised Learning) ในส่วนของในการเรียนรู้แบบไม่มีผู้สอน จะอาศัยการวัดระยะทางระหว่างข้อมูลและฟังก์ชันเกาส์เซียนเพื่อแปลงระยะทางให้เป็นค่าสมาชิกซึ่งเป็นเงื่อนไขในการจัดกลุ่ม เพื่อนำมาจัดสร้างโมเดลแบบจำลองระบบ สำหรับใช้ในการทำนาย ข้อเสียของวิธีการนี้คือ ไม่สามารถทราบได้ว่ามีกลุ่มย่อยใดบ้าง และข้อมูลในกลุ่มย่อยแต่ละกลุ่มประกอบด้วยข้อมูลอะไร

ดังนั้นเพื่อที่จะช่วยแก้ปัญหาของอัลกอริธึม COWEB ที่อาศัยเพียงการคำนวณฟังก์ชัน *CU* ที่ไม่เพียงพอสำหรับการจัดกลุ่ม และเพื่อช่วยให้วิธีการเรียนรู้แบบ ILM ในส่วนของการเรียนรู้แบบไม่มีผู้สอน สามารถที่จะแสดงโครงสร้างแบบลำดับชั้นพร้อมระบุกลุ่มของข้อมูลย่อย ๆ ที่มีอยู่ รวมทั้งเพิ่มประสิทธิภาพในการจัดกลุ่ม ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูล

เชิงลำดับชั้นแบบเพิ่มขึ้น โดยใช้พื้นฐานแนวคิดจากการจัดกลุ่มข้อมูลเชิงลำดับชั้นของ อัลกอริธึม COBWEB ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM ซึ่งในโครงร่างวิธีการใหม่นี้นอกจากจะมีความสามารถในการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเรียนรู้แบบเพิ่มขึ้นแล้ว ยังนำเสนอวิธีการคัดเลือกมิติข้อมูล และวิธีการเตรียมข้อมูลที่เหมาะสมสำหรับข้อมูลการแสดงออกของยีนไปพร้อมกันด้วย

1.2 วัตถุประสงค์

1.2.1 เพื่อพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นบนกลุ่มข้อมูลการแสดงออกของยีน

1.2.2 เพื่อสร้างต้นแบบใหม่ของขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นให้สามารถทำงานตามโครงร่างที่พัฒนาขึ้น

1.2.3 เพื่อเปรียบเทียบประสิทธิภาพระหว่างขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่พัฒนากับวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิสเฉพาะเทคนิคการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม

1.2.4 เพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นใหม่

1.3 สมมติฐานของงานวิจัย

การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น มีประสิทธิภาพมากกว่าวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิสในส่วนของการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม

1.4 ขอบเขตของการวิจัย

การวิจัยครั้งนี้กำหนดขอบเขตไว้ดังนี้

1.4.1 ขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่พัฒนาขึ้นใหม่ เพื่อจัดกลุ่มและทำนายบนกลุ่มข้อมูลการแสดงออกของยีน มีหลักการพื้นฐานมาจากวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิส และอัลกอริธึม COBWEB

1.4.2 การวัดประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่พัฒนาขึ้นใหม่ เจาะจงเฉพาะการทำงานกับข้อมูลที่มีคลาสเป้าหมายแบบสองคลาส (Binary Class)

1.4.3 กลุ่มข้อมูลการแสดงออกของยีนที่ใช้ในการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นมาจากกลุ่มข้อมูลที่เปิดเผยต่อสาธารณะจำนวน 4 กลุ่มข้อมูลจาก

1.4.3.1 http://www.upo.es/eps/bigs/datasets.html ได้แก่ Leukemia Dataset, Embryonal Tumor Dataset และ Colon Cancer Dataset

1.4.3.2 http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html ได้แก่ DLBCL Dataset

1.5 นิยามศัพท์

1.5.1 DNA (Deoxyribonucleic acid) หมายถึง องค์ประกอบของเซลล์ในสิ่งมีชีวิตต่าง ๆ ทั้งในคน พืช สัตว์ หรือแม้แต่แบคทีเรีย ที่สามารถจะถ่ายถอดจากบรรพบุรุษไปสู่ลูกหลานของสิ่งมีชีวิตต่าง ๆ ผ่านทางรหัสพันธุกรรม

1.5.2 message Ribonucleic Acid (mRNA) หมายถึง การถอดรหัสพันธุกรรมจากดีเอ็นเอออกไปสู่ด้านนอกนิวเคลียส เพื่อเข้าสู่การแปลรหัสให้ได้เป็นโปรตีน

1.5.3 cDNA หมายถึง กระบวนการย้อนกลับจากการแปลงดีเอ็นเอเป็นอาร์เอ็นเอปรกติ ให้เป็นการแปลงอาร์เอ็นเอกลับเป็นดีเอ็นเอแทน

1.5.4 Hybridization คือการเกิดกรดนิวคลีอิกสายคู่ (ดีเอ็นเอ-ดีเอ็นเอ ดีเอ็นเอ-อาร์เอ็นเอ หรืออาร์เอ็นเอ-อาร์เอ็นเอ) จับคู่ระหว่างเบสของกรดนิวคลีอิกสายเดี่ยวที่มีลำดับเบสเป็นคู่สมกัน

1.5.5 Microarray data หมายถึง ข้อมูลที่ได้จากดีเอ็นเอไมโครอะเรย์ ซึ่งเป็นการประยุกต์กับเทคนิคไฮบริไดเซชันของกรดนิวคลีอิก สำหรับช่วยในการตรวจสอบระดับการแสดงออกของยีน ในกรณีนี้ ยีนจำนวนมากสามารถถูกทดสอบได้ในครั้งเดียวกับไมโครอะเรย์ และโดยปรกติแล้วข้อมูลไมโครอะเรย์ถูกใช้เพื่อแสดงระดับการแสดงออกของยีน เพื่อการระบุความแตกต่างของระดับการแสดงออกของยีนที่แตกต่างกันในสถานะทางชีววิทยา

1.5.6 Probe หมายถึง ดีเอ็นเอหรืออาร์เอ็นเอสายสั้น ๆ ที่ติดฉลากกัมมันตรังสี โดยอาจมีที่มาจากชิ้นดีเอ็นเอที่ผ่านการโคลน มาจากห้องสมุดจีโนม (Genomic Library) หรือห้องสมุดซีดีเอ็นเอ (cDNA Library)

1.5.7 Gene expression level หมายถึง ระดับการแสดงออกของยีน เป็นการแสดงออกของยีนแต่ละยีนที่แตกต่างกัน เมื่อมีการเปลี่ยนแปลงทางชีววิทยา ช่วยให้นักชีววิทยาสามารถค้นหากลุ่มย่อยของยีนที่มีความน่าจะเป็นในการเกิดโรคหรืออาการผิดปรกติต่าง ๆ

1.5.8 ประสิทธิภาพ หมายถึง การพิจารณาการวัดค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-Measure) และเวลาในการประมวลผล

1.5.9 Clustering หมายถึง เทคนิคการจัดกลุ่มข้อมูลโดยข้อมูลที่มีความเหมือนกันจะอยู่คลัสเตอร์เดียวกัน และข้อมูลที่อยู่ต่างคลัสเตอร์กันจะมีความแตกต่างกัน

1.5.10 Dendrogram หมายถึง แผนภาพที่ใช้แสดงความสัมพันธ์ของสิ่งต่าง ๆ

1.6 ประโยชน์ของการวิจัย

1.6.1 ขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ที่พัฒนาขึ้นใหม่เป็นการผสมผสานแนวคิดการจัดกลุ่มข้อมูลเชิงลำดับชั้นของอัลกอริธึม COBWEB ร่วมกับวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิส ทำให้เกิดแนวความคิดใหม่เกี่ยวกับเงื่อนไขการจัดกลุ่มที่ได้จากการวัดระยะทางระหว่างข้อมูลร่วมกับเงื่อนไขขั้นตอนวิธีการจัดโครงสร้างแบบลำดับชั้น

1.6.2 วิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอยังสามารถใช้เป็นเครื่องมือเพื่อช่วยในการวางแผนและกำหนดนโยบายเพื่อบริหารจัดการทางด้านวิทยาศาสตร์ อาทิ การจัดกลุ่มความสามารถของนักวิจัย การจัดกลุ่มเครือข่ายการวิจัย การจัดกลุ่มข้อมูลเอกสารทางการวิจัย

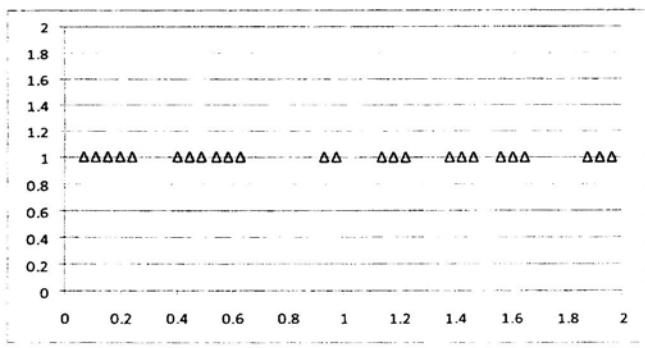
1.6.3 ต้นแบบวิธีการใหม่สำหรับขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น นอกจากจะสามารถทำงานกับข้อมูลการแสดงออกของยีนแล้ว ยังสามารถนำไปประยุกต์ใช้กับงานด้านอื่นได้อีกเช่น การจัดกลุ่มอุตสาหกรรม การจำแนกหมวดหมู่เว็บไซต์ การจัดกลุ่มสถานศึกษาเพื่อเฝ้าระวัง เป็นต้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

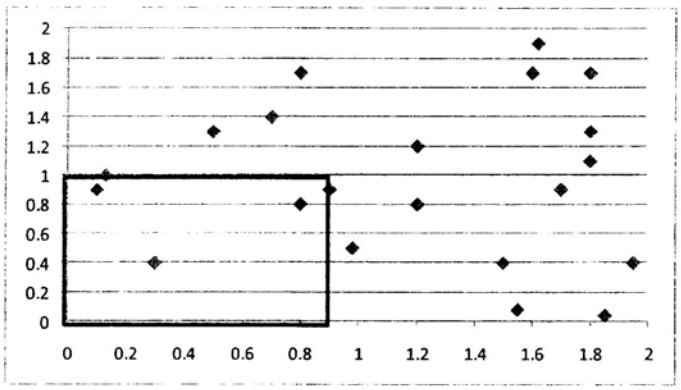
ในบทนี้ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง รวมถึงเอกสาร งานวิจัย งานเขียนอื่น ๆ ที่เกี่ยวข้อง โดยในบทนี้จะแบ่งเนื้อหาหลัก ๆ ออกเป็น 9 หัวข้อ ประกอบด้วย (1) การคัดเลือกมิติข้อมูล (2) การเรียนรู้แบบมีผู้สอน และแบบไม่มีผู้สอน (3) การจัดกลุ่มข้อมูล การวิเคราะห์การจัดกลุ่ม (4) ชีวสารสนเทศ (5) โครงสร้างของดีเอ็นเอ (6) โครงสร้างและการแสดงออกของยีน (7) ดีเอ็นเอโมโครอะเรย์ (8) การเปลี่ยนรูปข้อมูลการแสดงออกของยีน และ (9) การทดแทนข้อมูลที่สูญหายของข้อมูลไมโครอะเรย์ ดังรายละเอียดต่อไปนี้

2.1 การคัดเลือกมิติข้อมูล

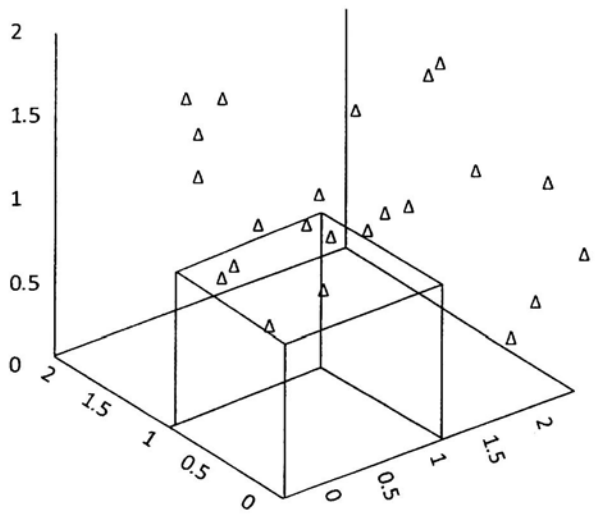
ปัญหาของข้อมูลที่มีมิติเป็นจำนวนมาก (ภัทราวุฒิและคณะ, 2009 ก) เช่น ข้อมูลไมโครอะเรย์ เรียกได้ว่าเป็น ปัญหามิติของข้อมูล (Curse of dimensionality) ซึ่งหมายถึง บางกรณีเมื่อข้อมูลมีจำนวนมิติ จำนวนแอททริบิวต์ หรือตัวแปรมาก รวมถึงมิติข้อมูลไม่ได้ไปในทิศทางเดียวกัน ทำให้ข้อมูลเกิดการกระจาย (Data sparse) ซึ่งทำให้บางจุดอาจจะไม่มีข้อมูลอยู่เลย ทำให้เกิดปัญหาในเรื่องของการตีความของกลุ่มข้อมูล ซึ่งถ้ามีมิติของข้อมูลต่ำก็จะทำให้ข้อมูลมีลักษณะของการเกาะกลุ่มกัน ในทางตรงกันข้ามถ้าข้อมูลมีมิติสูง ก็ทำให้ถูกตีความว่าเป็นคนละกลุ่มกันได้ ดังแสดงในภาพที่ 2-1 ถึง 2-3 ดังนั้นการคัดเลือกมิติข้อมูลช่วยทำให้สามารถลดการใช้ทรัพยากร อาทิเช่น ระยะเวลาในการฝึกฝน การสิ้นเปลืองหน่วยความจำ รวมถึงการช่วยกำจัดมิติที่ไม่เกี่ยวข้องหรือสิ่งรบกวน (Noise) กับข้อมูลได้อีกด้วย



ภาพที่ 2-1 การใช้ข้อมูล 25 ตัวอย่างที่มีค่าระหว่าง 0 ถึง 2 กับจำนวนข้อมูลหนึ่งมิติ



ภาพที่ 2-2 การใช้ข้อมูล 25 ตัวอย่างที่มีค่าระหว่าง 0 ถึง 2 กับจำนวนข้อมูลสองมิติ



ภาพที่ 2-3 การใช้ข้อมูล 25 ตัวอย่างที่มีค่าระหว่าง 0 ถึง 2 กับจำนวนข้อมูลสามมิติ

Tan *et al.* (2005) กล่าวว่า แนวคิดในการคัดเลือกมิติข้อมูลคือความพยายามที่จะหากลุ่มย่อยของมิติข้อมูลที่มีความเหมาะสมสำหรับทำงานกับอัลกอริธึมที่เราสนใจ การคัดเลือกมิติข้อมูล (Feature Selection) มักถูกนำมาใช้ในขั้นตอนกระบวนการเตรียมข้อมูล (Data Preprocessing) หรือเป็นส่วนหนึ่งของการวิเคราะห์ข้อมูล เพื่อให้การสร้างโมเดลทำได้ง่าย แต่ถ้าตัวแปรข้อมูลเข้ามีขนาดใหญ่มาก การคัดเลือกมิติข้อมูลจะช่วยเพิ่มความแม่นยำของการวิเคราะห์ข้อมูล

Huan and Hiroshi (2008) กล่าวว่า ปัจจุบันฐานข้อมูลจำนวนมากที่มีอยู่ในโลกของความเป็นจริง มักจะไม่มีมิติข้อมูลเป้าหมายครบถ้วน เนื่องจากจำนวนของข้อมูลที่มีมากทำให้ยากสำหรับการจัดการกำหนดมิติเป้าหมายให้กับแต่ละข้อมูลด้วยแรงงานมนุษย์ ซึ่งการกำหนดมิติเป้าหมายด้วยแรงงานมนุษย์มักมีค่าใช้จ่ายสูง และมีความลำเอียงที่เกิดจากทัศนคติของแต่ละบุคคลเข้ามา

ประกอบด้วย ดังนั้นวิธีการเรียนรู้แบบไม่มีผู้สอนจึงเป็นสิ่งจำเป็น สำหรับงานที่ไม่มีคลาสของข้อมูลเป้าหมายมีอยู่หลายงานและมีลักษณะของข้อมูลที่มีมิติที่สูง เช่น ตัวอักษร (Text) ภาพ (Image) และยีน (Gene)

Sumathi and Sivanandam (2006) ได้กล่าวถึงปัญหาของมิติข้อมูลไว้ว่า มิติ d ของเวกเตอร์ x มีบทบาทสำคัญในหลากหลายโมเดล เช่น การจำแนกหมวดหมู่เอกสาร (Text classification) และการจัดกลุ่มข้อมูลการแสดงออกของยีน (Clustering of gene expression data) ซึ่ง d อาจจะมีมิติขนาด 10^3 หรือ 10^4 ซึ่งทฤษฎีการประมาณความหนาแน่นแสดงให้เห็นว่า จำนวนข้อมูลที่ต้องการใช้ในการประมาณการของฟังก์ชันขยายตัวแบบเอ็กซ์โพแนนเชียล (Exponential) ในมิติ d

ในการเลือกยีนที่มีความสำคัญหรือมีอำนาจจำแนก วิธีการดั้งเดิมมีอยู่ 2 วิธี คือวิธีการกรอง (Filter approach) และวิธีการควบรวม (Wrapper approach) (Ke, 2008)

ตารางที่ 2-1 การเปรียบเทียบคุณลักษณะระหว่างวิธีการกรองและวิธีการควบรวม

วิธีการกรอง	วิธีการควบรวม
1. เลือกยีนจากอันดับ Top-Rank ตามความเกี่ยวข้องกับมิติเป้าหมาย โดยปราศจากอำนาจของอัลกอริธึมเข้าร่วมคัดเลือกด้วย 2. ไม่สามารถค้นหาความสัมพันธ์หรือการควบคุมระหว่างยีน 3. ใช้ระยะเวลาในการทำน้อยกว่าวิธีการควบรวม 4. อัลกอริธึมที่ใช้งานอาทิเช่น T-test, correlation Information gain, F-statistic Mutual information	1. ดึงกลุ่มของยีนที่มีอยู่ โดยใช้ประโยชน์จากอัลกอริธึมร่วมด้วย มีกระบวนการทำซ้ำประกอบด้วยการคัดเลือกยีนที่สลับสับเปลี่ยนกัน เพื่อค้นหายีนที่มีอำนาจจำแนก 2. สิ้นเปลืองทรัพยากรในการทำงานมากอาทิ หน่วยความจำและเวลาในการทำงาน 3. อัลกอริธึมที่ใช้งานอาทิเช่น FSFS BSFS Genetic Search Simulated Annealing

ในวิธีการกรองแต่ละยีนจะถูกค้นหาบนความเกี่ยวข้องของตัวเองกับคลาสเป้าหมาย โดยปราศจากการพึ่งพิงความสามารถของอัลกอริธึมการจำแนกหมวดหมู่ และยีนทั้งหมดถูกจัดอันดับบนค่าความเกี่ยวข้องของกับคลาสเป้าหมาย วิธีการกรองที่ใช้กันโดยทั่วไปเช่น การจัดลำดับข้อมูลด้วยเทคนิครีรีฟเอฟ (ReliefF) และเทคนิคอินฟอร์เมชันแกน (Information Gain: Info) วิธีการเลือกยีนด้วยวิธีการกรองนี้ มักจะนำไปใช้เป็นส่วนสำคัญของโมเดลการจำแนกหมวดหมู่ ในทางตรงกันข้าม วิธีการควบรวมใช้อัลกอริธึมเข้าร่วมเป็นส่วนหนึ่งของกระบวนการคัดเลือก ดังนั้นแทนที่จะทำ

การจัดอันดับยีน เหมือนดังเช่นกระบวนการของวิธีการกรอง วิธีนี้จะสร้างกลุ่มย่อยของยีน โดยถือว่ากลุ่มย่อยที่ดีที่สุดจะมีโอกาสประสบความสำเร็จกับโมเดลการจำแนกหมวดหมู่ในปัจจุบัน โดยทั่วไปการเลือกของวิธีการควบรวมเริ่มจากยีนเดี่ยว หรือกลุ่มของยีนทั้งหมด จากนั้นจะเพิ่มยีนอย่างต่อเนื่องและค้นหากลุ่มที่ดีที่สุด ดังนั้นจึงสามารถช่วยสร้างโมเดลที่มีความแม่นยำที่มากที่สุด เพราะวิธีการกรองรวมเอาวิธีการค้นหาไว้ด้วยกัน ซึ่งเหมาะสำหรับวิธีการแก้ปัญหาแบบวิทยาการศึกษาสำนึก (Heuristic) บ่อยครั้งกระบวนการค้นหาอาจสร้างความล่าช้าได้ ถึงกระนั้นวิธีการควบรวมก็ใช้ทรัพยากรมากกว่าวิธีการกรอง อาทิ เวลาที่ใช้ในการคำนวณ หรือพื้นที่ในหน่วยความจำ นอกจากนี้ในวิธีการควบรวมการเลือกกลุ่มย่อยของยีน จะมีความเหมาะสมเฉพาะสำหรับตัวอัลกอริธึมใช้ในการคัดเลือกเท่านั้น (Slavik *et al.* 2009)

โดยทั่วไปเทคนิคการคัดเลือกมิติข้อมูลมีหลายวิธีด้วยกัน แต่ในที่นี้ขอนำเสนอ 4 วิธี คือ

2.1.1 เทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้น (Backward Stepwise Feature Selection : BSFS)

เทคนิคการ BSFS เป็นหนึ่งในเทคนิคของการควบรวม (Cunningham, 2007) ซึ่งผลลัพธ์ของวิธีการนี้ จะเกิดขึ้นโดยการใช้ประสิทธิภาพของตัวจำแนกหมวดหมู่ (Classifier) ในการค้นหาตัวแปรที่ควรถูกเลือก ข้อดีของวิธีการนี้คือ ตัวแปรที่ถูกเลือกจะมีความแม่นยำมากกว่าวิธีการเลือกตัวแปรแบบการกรอง สำหรับวิธี BSFS จะเริ่มจากการใช้ตัวแปรทั้งหมด มาคำนวณความแม่นยำ จากนั้นจะทำการพิจารณาตัดตัวแปรออกทีละตัว แล้วนำมาทดสอบดูความแม่นยำ ถ้าเป็นที่ยอมรับได้ก็ทำการตัดตัวแปรตัวต่อไป จนกระทั่งพบว่าความแม่นยำไม่อยู่ในเกณฑ์ที่ยอมรับได้จึงจะหยุดทำการตัดตัวแปรออก

2.1.2 เทคนิคการเลือกมิติข้อมูลโดยใช้ความสัมพันธ์ (Correlation Based Feature Selection: Cfs)

Hall (1999) ได้กล่าวถึง Cfs ว่าคืออัลกอริธึมการกรองที่ง่ายโดย Cfs ทำการจัดอันดับกลุ่มย่อยของมิติข้อมูล ตามความสัมพันธ์ที่อยู่บนพื้นฐานของฟังก์ชันวิธีการแก้ปัญหาแบบวิทยาการศึกษาสำนึก โดยกลุ่มย่อยของมิติข้อมูลจะมีความสัมพันธ์กันสูงกับคลาส และไม่มีความสัมพันธ์กับคลาสอื่น สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องอาจจะถูกละทิ้ง เพราะมิติข้อมูลเหล่านี้อาจจะมีความสัมพันธ์ต่ำกับคลาส มิติข้อมูลที่ซ้ำซ้อนอาจจะถูกขจัดออกไปจากกลุ่มมิติข้อมูลที่มีความสัมพันธ์สูง สมการประเมินกลุ่มย่อยของมิติข้อมูลแบบ Cfs แสดงในสมการที่ 2-1

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (2\text{-}1)$$

โดยที่ M_s คือ ค่าที่ค้นหาได้ของมิติข้อมูลกลุ่มย่อย S ซึ่งประกอบด้วยมิติข้อมูล k

$\bar{r}_{cf}$ คือ ค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส ($f \in S$)

$\bar{r}_{ff}$ คือ ค่าเฉลี่ยความสัมพันธ์ระหว่างมิติของข้อมูล

2.1.3 เทคนิค Info

Tan *et al.* (2005) ได้กล่าวว่าเทคนิค Info ใช้ในการเลือกมิติของข้อมูล เพื่อใช้ในการแบ่งแยกข้อมูลเทคนิค Info จะคำนวณค่าเกน (Gain) สำหรับแต่ละมิติข้อมูล ซึ่งถ้ามิติข้อมูลใดมีค่า Gain สูงสุดจะถูกเลือกให้เป็นกลุ่มย่อยของมิติข้อมูลที่มีอำนาจจำแนก สมการที่ 2-2 แสดงการคำนวณค่าเอนโทรปี้ (Entropy) และสมการที่ 2-3 แสดงการคำนวณค่า Info

$$Entropy(t) = -\sum_i p(j \mid t) \log_2 p(j \mid t) \tag{2-2}$$

โดยที่ $\sum_i$ คือผลรวมของความน่าจะเป็นของค่า j ที่เกิดในคลาส t

$$GAIN = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \tag{2-3}$$

โดยที่ $Entropy\,(p)$ คือ ค่า Entropy ของตัวราก (Root)

$\sum_{i=1}^{k} \frac{n_i}{n} Entropy\,(i)$ คือ ค่า Entropy ในแต่ละโหนดย่อย

2.1.4 เทคนิคเกนเรโช (Gain Ratio: GR)

จากข้อเสียของเทคนิค Info คือ ปัญหาคลาสที่มีขนาดใหญ่สามารถที่จะเกิดความลำเอียงในการคัดเลือกมิติข้อมูลได้ ดังนั้นเทคนิค GR ซึ่งถูกแนะนำโดย Quinlan ในปี 1986 จึงถูกใช้เพื่อแก้ปัญหาข้อเสียดังกล่าว ซึ่งเทคนิค GR แสดงให้เห็นความน่าเชื่อถือถึงการในการประเมินการคัดเลือกมิติข้อมูล การวัด GR ในแต่ละคลาส ทั้งนี้การคำนวณ GR ต้องมีการคำนวณค่า SplitINFO ดังแสดงในสมการที่ 2-4 จากนั้นจึงจะนำมาหารกับค่า Info ที่ได้ ดังแสดงในสมการที่ 2-5

$$SplitInfor = \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n} \tag{2-4}$$

$$Gain\ ratio = \frac{\Delta \text{info}}{SplitInfor} \tag{2-5}$$

อย่างไรก็ตามการคัดเลือกมิติข้อมูลการแสดงออกของยีน ก็ยังเป็นปัญหาที่สำคัญของนักชีววิทยา เนื่องจากการคัดเลือกยีนที่มีระดับการแสดงออกมากสามารถที่จะนำมาศึกษาวิจัยผลกระทบของอาการหรือโรคที่เกิดขึ้นทั้งในคน สัตว์ และพืช วิธีการที่เป็นที่นิยมใช้ในทางปฏิบัติคือ ศึกษาจากงานวิจัยก่อนหน้าที่มีการทดสอบเกี่ยวกับยีนชนิดนั้น ๆ มาก่อน และทำการเลือกตัวอย่างยีนนั้น ๆ ออกมาทดสอบ แม้ว่าวิธีการนี้จะเป็นวิธีการที่ดี แต่เนื่องจากปัจจุบันมียีนใหม่จำนวนมากที่ถูกค้นพบ ทำให้ยีนบางตัวอาจจะยังไม่ได้มีการศึกษาถึงการเปลี่ยนแปลงมาก่อน นอกจากนี้ข้อมูลของไมโครอะเรย์มีจำนวนมิติที่มาก (ยีน) และมีจำนวนของตัวอย่างน้อย ซึ่งเป็นปัญหาที่ทำให้ได้ผลลัพธ์ของข้อมูลทดสอบที่มีประสิทธิภาพต่ำลงกว่าความเป็นจริง (Over-fitting) ดังนั้น Horng *et al.* (2009) ได้ใช้วิธีการทางสถิติเพื่อเพิ่มจำนวนของตัวอย่างจากข้อมูล ไมโครอะเรย์ ด้วยวิธีการสุ่มซ้ำ (Resampling) และ Kamal *et al.* (2009) ได้ทำการเปรียบเทียบการจำแนกหมวดหมู่ยีนจากข้อมูลไมโครอะเรย์ที่มีตัวอย่างน้อย แสดงให้เห็นว่าการคัดเลือกยีนจากกลุ่มตัวอย่างที่น้อยให้ผลดีกับอัลกอริธึมเพียงบางตัวเท่านั้น ในงานวิจัยของ Mukherjee (2004) ทำการเปรียบเทียบวิธีการจัดอันดับเพื่อคัดเลือกยีนจากวิธีการทางสถิติซึ่งพบว่าสถิติทดสอบแบบที (T-statistics) ให้ผลลัพธ์การจัดอันดับยีนที่ไม่แน่นอน เมื่อความแปรปรวนระหว่างกลุ่มตัวอย่างแตกต่างกัน ต่างกับวิธีค่าเฉลี่ยซึ่งจะให้ผลดีกว่าในกรณีนี้ แต่ก็มีบางงานวิจัยเช่น Mundra and Rajapakse (2009) ได้นำเสนอเทคนิคการวิเคราะห์พาเรโตและสถิติทดสอบแบบเอฟ (Pareto Frontal Analysis and F-statistic) โดยวิธีนี้ซึ่งทำการแยกค่าคะแนนเอฟ (F-score) ออกเป็นส่วนสำหรับแต่ละคลาส (Class) และทำการเปรียบเทียบกับวิธีทั้งหมด ขณะเดียวกันวิธีการคัดเลือกยีนด้วยเทคนิคการกรอง และเทคนิคการควบรวม ได้มีหลายงานวิจัยนำมาใช้ในการเลือกยีน เทคนิคการกรองใช้คุณลักษณะของแต่ละมิติของข้อมูลที่มีความสามารถในการคัดแยก มีความรวดเร็วในการทำงาน ตัวอย่างของเทคนิคการกรองได้แก่ Cfs กับ Info และ T-statistics สำหรับเทคนิคการคัดเลือกยีนแบบอาศัยอัลกอริธึม ใช้ประสิทธิภาพของอัลกอริธึมเป็นตัวคัดเลือกมิติของข้อมูล ดังนั้นวิธีนี้จะอาศัยเวลามากกว่าวิธีการกรอง แต่ให้ความแม่นยำมากกว่า (Yijuan *et al.*, 2003; Hikaru *et al.*, 2008; Jin-Hyuk and Sung-Bae, 2008; Cheng-San *et al.*, 2008 a; Cheng-San *et al.*, 2008 b) ในส่วนของการจำแนกหมวดหมู่ของยีน ได้มีการนำเอาเทคนิคทางเหมืองข้อมูลมาประยุกต์ใช้ อาทิ SVM (Kamal *et al.*, 2009; Cheng-San *et al.*, 2008; Chakraborty, 2009; Cao, 2009) เทคนิค ReliefF (Slavik, 2009; Cheng-San *et al.*, 2008 a; Shen *et al.*, 2009) และวิธี CalNC (Shen *et al.*, 2009; Dabney, 2005) พาร์ทิเคิลสวอมออฟทีไมเซชั่น (Particle Swarm Optimization: PSO) (Cheng-San, 2008 a; Cheng-San, 2008 b) มัลติออพเจคทีพเจเนติกอัลกอริธึมซัพพอตร์

เวกเตอร์แมชีน (Multi-objective Genetic Algorithm Support Vector Machine: MAGOSVM) (Mohamad *et al.*, 2009) และ Decision Tree (Horng *et al.*, 2009) เป็นต้น ดังแสดงในตารางที่ 2-2

ตารางที่ 2-2 เทคนิคการคัดเลือกและจัดอันดับมิติบนข้อมูลการแสดงออกของยีน

Ranking	Feature Selection	Classifier	Evaluation	Papers
No	PReliefFp and PReliefFg	SVM and K-NN	Accuracy rate	Slavik *et al.* (2009)
Chi-square	Decision tree (C4.5), GEMS and HykGene	Decision tree (C4.5)	LOOCV, TP, FP, TN and FN	Horng *et al.* (2009)
CFS subset , Chi-Square (c2) Information Gain Ratio (IGR), ReliefF and Symmetrical Uncertainty (SU).	No	SVM, K-NN, Random Forest (RF), and PART	LOOCV, AC, TP, TN and KS	Kamal *et al.* (2009)
T-test, Different of Mean, SAM	No	No	Type I error and P-value	Mukherjee and S.J. (2004)
Pareto frontal analysis and f-statistic	No	No	No	Mundra and Rajapakse (2009)
No	LDA, BDA, ADA, Boostep ADA, and PCA	SVM, PCA, BDA, LDA and ADA	F-measure (precision and recall)	Yijuan *et al.* (2003)
F-value	FSM and Maharanuobis with with Simple and S2N	FSM, Simple, S2N and Random Forest	LOOCV and error rate	Hikaru *et al.* (2008)

ตารางที่ 2-2 (ต่อ)

Ranking	Feature Selection	Classifier	Evaluation	Papers
F-test and Jaccard	Filter and wrapper	OFW with SVM/CART and RF	Error rate	Cao *et al.* (2009)
No	Gene boosting (filter-wrapper), GAKNN, and BIRS	K-NN (Euclidean distance)	LOOCV	Jin-Hyuk and Sung-Bae (2008)
No	NBPSO	NBPSO and K-NN	LOOCV, and error rate	Cheng-San *et al.* (2008 a)
BPSO (filter and wrapper)	Cfs	K-NN	LOOCV	Cheng-San *et al.* (2008 b)
t-statistics	No	LDA	Error rate (FDR)	Shen *et al.* (2009)
t-test	SSPSO and SVM	SVM, PSOSVM and SSPSO	Error rate classification 5-fold cross validation	Chakraborty (2009)
No	No	Nearest Centroid Classification, and PAM, Suitability Score	LOOCV	Dabney (2005)
No	MOGASVM	MOGASVM, SVM, GASVM and GASVM-II	LOOCV	Mohamad *et al.* (2009)

ตารางที่ 2-2 (ต่อ)

Ranking	Feature Selection	Classifier	Evaluation	Papers
Fisher, Golub, Wilconxon, TNoM, T-test, correlation, and Fuzzy K-Mean	No	SVM	LOOCV and ROC	Jaeger *et al.* (2003)

2.2 การเรียนรู้แบบมีผู้สอนและแบบไม่มีผู้สอน

การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการค้นหาความสัมพันธ์ระหว่างมิติข้อมูลนำเข้า (Input Attribute) และมิติข้อมูลเป้าหมาย (Target Attribute) การค้นหาความสัมพันธ์นี้จะถูกแสดงในรูปแบบของโมเดล (Model) บ่อยครั้งที่โมเดลอธิบายและบรรยายปรากฏการณ์ซึ่งถูกซ่อนอยู่ในกลุ่มข้อมูล (Dataset) และสามารถถูกใช้สำหรับการทำนายค่าของมิติเป้าหมายได้ จากการรู้ค่าของมิติข้อมูลนำเข้า การเรียนรู้แบบมีผู้สอน จะถูกอธิบายโดยเวกเตอร์ของ $\{x_i, t_i\}$ ซึ่งเป็นการแสดงการอธิบายและการตอบสนองของตัวแปร โดยให้ $D = \{(x_1, t_1),\ldots,(x_n, t_n)\}$ ซึ่ง D เป็นกลุ่มของข้อมูลที่เหมาะสมทั้งหมด (Giudici and Figini, 2009) การเรียนรู้แบบมีผู้สอนสามารถถูกนำไปใช้ได้หลากหลายขอบเขต เช่น การตลาด การเงิน และอุตสาหกรรม สำหรับประโยชน์ที่ชัดเจนของวิธีการเรียนรู้แบบมีผู้สอนมี 2 โมเดลหลัก คือ โมเดลในภาพแบบของการจำแนกหมวดหมู่ (Classification Model) และโมเดลในภาพแบบของสมการถดถอย (Regression Model)

สำหรับโมเดลในภาพแบบของสมการถดถอยจะทำการจำลองปริภูมิตัวแปรอิสระเข้าสู่ขอบเขตค่าจริง โดยสามารถทำนายได้จากคุณลักษณะของที่ได้รับมา ในทางตรงกันข้ามโมเดลในรูปแบบของการจำแนกหมวดหมู่จะทำการจำลองปริภูมิตัวแปรอิสระเข้าสู่คลาสที่ถูกกำหนดไว้ก่อน เช่น ตัวอัลกอริธึมการจำแนกหมวดหมู่สามารถถูกใช้เพื่อแยกประเภทของผู้ที่ได้รับอนุมัติสินเชื่อหรือไม่ได้รับอนุมัติสินเชื่อ มีหลายวิธีที่เป็นอัลกอริธึมการจำแนกหมวดหมู่ เช่น ซัพพอตร์เวกเตอร์แมชีน (Support Vector Machine: SVM) ต้นไม้ตัดสินใจ นาอีฟเบย์ และเคเนียเรสเนเบอร์ และเหมือนเช่นเดียวกับสมการถดถอย และความน่าจะเป็น การจำแนกหมวดหมู่เป็นหนึ่งโมเดลที่มีการศึกษากันมาก ประโยชน์ของโมเดลในการจำแนกหมวดหมู่มีมากเนื่องจากเป็นเทคนิคที่มีผลกระทบสูงกับวงการอื่น ทั้งในดาต้าไมน์นิ่ง (Data Mining) และในการใช้งานด้านการวิจัย

ในทางตรงกันข้าม การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการจัดกลุ่มข้อมูลที่ปราศจากการกำหนดตัวแปรตาม หรือคลาสเป้าหมายเอาไว้ก่อน ซึ่งการเรียนรู้แบบไม่มีผู้สอน แต่ละการทำงานจะถูกอธิบายโดยเพียงหนึ่งเวกเตอร์ กับข้อมูลที่เหมาะสมทั้งหมด โดยให้

$D = \{(x_1),\ldots,(x_n)\}$ ซึ่ง D เป็นกลุ่มของข้อมูลที่เหมาะสมทั้งหมด วิธีการนี้จะแบ่งกลุ่มของ D เข้าสู่กลุ่มย่อย (subset) โดยที่เวกเตอร์ x_i จะเป็นกลุ่มย่อยที่มีลักษณะเหมือนกันในการเปรียบเทียบการวัดระยะห่างที่กำหนดเอาไว้ (Giudici and Figini, 2009) ดังนั้นลักษณะของการเรียนรู้แบบไม่มีผู้สอนจะครอบคลุมเพียงเรื่องของวิธีการจัดกลุ่ม (Clustering) เป้าหมายคือจะถูกใช้เพื่ออธิบายรูปแบบของข้อมูลเหล่านั้น การจัดกลุ่มสำหรับการเรียนรู้แบบไม่มีผู้สอน (บางครั้งโมเดลการจัดกลุ่มอาจจะเป็นได้ทั้งโมเดลการจำแนกหมวดหมู่ด้วย) เนื่องจากเป้าหมายของการจัดกลุ่มคือการค้นหารูปแบบของข้อมูล รวมถึงการจัดกลุ่มข้อมูลเข้าสู่กลุ่มย่อย เช่น ข้อมูลที่มีความเหมือนกันจะถูกจัดกลุ่มไว้ด้วยกัน ขณะที่ข้อมูลที่แตกต่างกันจะกลายเป็นละกลุ่มกัน ด้วยเหตุนี้ข้อมูลจะถูกจัดการเข้าสู่การนำเสนอที่มีประสิทธิภาพตามคุณลักษณะประชากรที่อยู่ในกลุ่มข้อมูลเหล่านั้น รูปแบบที่เป็นทางการของโครงสร้างการจัดกลุ่ม (clustering structure) จะถูกแสดงเป็นกลุ่มของกลุ่มย่อย $C = C_1,\ldots,C_k$ ของ S เช่น $S = U_{i=1}^{k}$ และ $C_i \cap C_j = \phi$ สำหรับ $i \neq j$ ดังนั้นข้อมูลใดใน S จะอยู่เพียงกลุ่มย่อยใดกลุ่มย่อยหนึ่งเท่านั้น (Oded and Rokach, 2005)

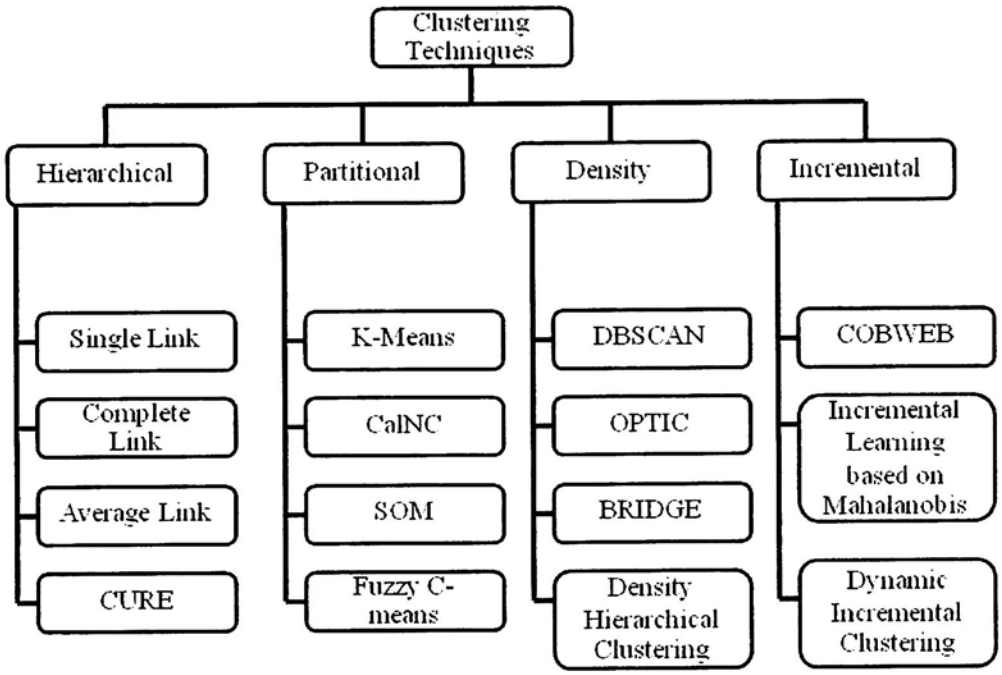
2.3 การจัดกลุ่มข้อมูล และการวิเคราะห์การจัดกลุ่ม

"วิธีการจัดกลุ่มข้อมูล" (Clustering Method) เป็นเทคนิคพื้นฐานของการเรียนรู้แบบไม่มีผู้สอน ทำหน้าที่ค้นหาโครงสร้างที่ซ่อนอยู่จากข้อมูลที่ไม่ทราบข้อมูลเป้าหมาย การจัดกลุ่มที่ดีควรจะแบ่งข้อมูลเป็นหลายกลุ่ม โดยที่ในกลุ่มเดียวกันต้องมีความเหมือนกันภายในกลุ่มมากที่สุด (Intra-Cluster Similarity) จะถูกจัดเข้าไว้ด้วยกัน และความเหมือนกันระหว่างแต่ละกลุ่ม (Inter-Cluster Similarity) ที่จะต้องให้ความเหมือนกันน้อยที่สุด ปัจจุบันนี้เทคนิคการจัดกลุ่มเป็นที่ต้องการในทุก ๆ สาขา โดยเฉพาะการจัดกลุ่มยังคงมีการใช้งานอยู่ในเรื่องของเครื่องจักรการเรียนรู้ (Machine Learning) ทั้งในรูปแบบของการค้นหาความรู้ (Knowledge Discovery Patterns) และการหาความสัมพันธ์ในกลุ่มข้อมูลที่มีความซับซ้อน ซึ่งประเภทของการจัดกลุ่มได้แสดงในภาพที่ 2-4

Tan et *al.* (2005) ได้กล่าวถึงการจัดกลุ่มไว้ว่า "การจัดกลุ่มข้อมูล เป็นงานที่ต้องใช้เทคนิคที่ซับซ้อนเพื่อค้นหากลุ่มข้อมูลได้อย่างแม่นยำ โดยจะใช้วิธีการจัดกลุ่มข้อมูลเชิงความหนาแน่นเป็นกระบวนการรวมกลุ่มของข้อมูลเข้าสู่คลัสเตอร์ ซึ่งข้อมูลภายในคลัสเตอร์จะมีความเหมือนมากกว่าเมื่อเปรียบเทียบกับคลัสเตอร์อื่น การจัดกลุ่มข้อมูลกับงานทางด้านชีวภาพมีมานาน เพื่อที่จะสร้างการแบ่งประเภทของสิ่งมีชีวิต ดังนั้นงานส่วนใหญ่ที่เป็นการจัดกลุ่ม ในตอนเริ่มแรกได้นำเอาหลักการจำแนกหมวดหมู่ทางคณิตศาสตร์ ที่สามารถใช้ค้นหาโครงสร้างแบบอัตโนมัติ" สำหรับการเลือกวิธีการจัดกลุ่มที่เหมาะสม Han *et al.* (2009) ได้กล่าวว่ามีข้อควรพิจาณาดังนี้

ก) เป้าหมายการใช้งาน

เป้าหมายของการใช้งานเป็นสิ่งสำคัญที่ใช้ในการตัดสินใจเลือกอัลกอริธึมการจัดกลุ่มเช่น การค้นหาตำแหน่งของอัลกอริธึม K-Means มีแนวโน้มค้นหาคลัสเตอร์ที่มีภาพร่างทรงกลม และ



ภาพที่ 2-4 ประเภทของการจัดกลุ่มข้อมูล

ขนาดที่เหมือนกัน แต่ในขณะที่เทคนิคความหนาแน่น (Density-based) จะเหมาะสมสำหรับการค้นหาภาพที่มีรูปร่างซับซ้อน

ข) ความเร็วและคุณภาพ

ขนาดของข้อมูลที่อยู่ในกลุ่มก่อให้เกิดปัจจัยเกี่ยวกับการใช้เวลาของอัลกอริธึมสำหรับการจัดกลุ่ม ซึ่งบางครั้งอัลกอริธึมการจัดกลุ่มที่มีประสิทธิภาพอาจจะไม่สามารถทำงานได้ดีในฐานข้อมูลขนาดใหญ่ การจัดการกับปัญหาเหล่านี้วิธีโดยทั่วไปมักใช้การบีบอัดข้อมูล และอาจจะทำให้คุณภาพของผลลัพธ์การจัดกลุ่มสูญเสียไปด้วย

ค) คุณลักษณะของข้อมูล

คุณลักษณะของข้อมูลนับเป็นปัจจัยสำคัญที่จะกำหนดอัลกอริธึมการจัดกลุ่มที่เหมาะสม ซึ่งประกอบด้วย ชนิดของข้อมูล มิติของข้อมูล จำนวนของข้อมูลรบกวน

ปัจจุบันเทคนิคการจัดกลุ่มได้รับการพัฒนาและในหลายปีที่ผ่านมานี้จำนวนของอัลกอริธึมการจัดกลุ่มถูกนำเสนอเป็นจำนวนมาก เช่น อัลกอริธึมเคมีน (K-Means) และอัลกอริธึมการแบ่งข้อมูลด้วยตนเอง (Self-Organizing Maps: SOM) อย่างไรก็ตามเนื่องจากข้อมูลการแสดงออกของ

ยีนมีความหลากหลายและมีความซับซ้อนมาก ปัจจุบันจึงยังไม่มีอัลกอริธึมการจัดกลุ่มใดที่ทำงานได้ดีกว่ากัน ดังนั้นนักวิจัยโดยทั่วไปจึงพยายามพัฒนาโมเดลการจัดกลุ่มที่แตกต่างกันและเปรียบเทียบผลลัพธ์เพื่อที่จะสร้างผลลัพธ์ที่มีความหมายและน่าเชื่อถือมากยิ่งขึ้น (Azuaje, 2010)

อัลกอริธึมการจัดกลุ่มข้อมูลที่ถูกใช้เพื่อจัดกลุ่มข้อมูลการแสดงออกของยีน สามารถแบ่งออกได้ดังนี้

2.3.1 การจัดกลุ่มข้อมูลเชิงลำดับชั้น (Hierarchical Clustering : HC)

วิธีการหนึ่งสำหรับการจัดกลุ่มที่มีการยอมรับโดยทั่วไปว่าเป็นวิธีการที่มีประสิทธิภาพ คือ การจัดกลุ่มข้อมูลเชิงลำดับชั้น วิธีการนี้ประสบความสำเร็จในการใช้การวิเคราะห์การแสดงออกของยีน เช่น การวิเคราะห์ภาพแบบการแสดงออกแบบชั่วคราว การทำนายผู้ป่วยโรคมะเร็งต่อมน้ำเหลือง และการให้ภาพของก้อนเนื้อมะเร็งทรวงอกขนาดเล็ก เป็นต้น นอกจากนี้ยังสามารถใช้แสดงภาพเสมือน (Visual Representations) เช่น แผนภาพเดนโดแกรม (Dendrogram) สำหรับวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นที่มักถูกใช้กับข้อมูลไมโครอะเรย์คือวิธีการแบบล่างขึ้นบน (Bottom-up Approach) ผลลัพธ์ที่ได้ของการจัดกลุ่มข้อมูลเชิงลำดับชั้นลักษณะนี้คือไบนารี่ทรี (Binary Tree) หรือ โครงสร้างต้นไม้ และโหนดลูกของแต่ละกลุ่มข้อมูลในโครงสร้างต้นไม้คือกลุ่มข้อมูลย่อย ซึ่งใช้การวัดระยะห่างระหว่างสองกลุ่มข้อมูล เทคนิคการจัดกลุ่มข้อมูลเชิงลำดับชั้นมักถูกใช้ในการจัดกลุ่มข้อมูลการแสดงออกของยีนและผู้เชี่ยวชาญจะเป็นผู้ปรับปรุงบางส่วนของอัลกอริธึมนี้ (Yu *et al.*, 2008)

อย่างไรก็ตามความน่าเชื่อถือของการจัดกลุ่มข้อมูลเชิงลำดับชั้นขึ้นอยู่กับเทคนิคการวัดระยะห่าง (Metric) ที่ถูกใช้ประเมินระยะห่างความเหมือนระหว่างข้อมูล ซึ่งการจะเลือกเทคนิคการวัดระยะห่างที่เหมาะสมเป็นปัญหาหนึ่ง ที่มักขึ้นอยู่กับคุณสมบัติของข้อมูลที่มีอยู่

ให้เซตของ N ข้อมูลถูกจัดกลุ่มและมีแมตทริกส์ระยะห่าง (Distance Matrix) คือ $N*N$ กระบวนการพื้นฐานของการจัดกลุ่มข้อมูลเชิงลำดับชั้นเป็นดังนี้ (Kai *et al.*, 2009)

ขั้นที่ 1: เริ่มโดยการประเมินแต่ละข้อมูลให้กับกลุ่มข้อมูล แต่ละกลุ่มข้อมูลจะมีเพียงหนึ่งข้อมูล จากนั้นให้ระยะห่าง (หรือความเหมือน) ระหว่างกลุ่มข้อมูลที่เหมือนกัน เช่น ระยะห่างระหว่างข้อมูลที่มีอยู่

ขั้นที่ 2: หาคู่ของกลุ่มข้อมูลที่ใกล้ที่สุด (เหมือนกันมากที่สุด) และรวมข้อมูลทั้งสองเข้าด้วยกันเป็นกลุ่มเดียว ดังนั้นตอนนี้จะมีหนึ่งกลุ่มเป็นอย่างน้อย

ขั้นที่ 3: คำนวณระยะห่างระหว่างกลุ่มข้อมูลใหม่ และแต่ละอันของกลุ่มข้อมูลเก่า

ขั้นที่ 4: ทำซ้ำในขั้นตอนที่ 2 และ 3 จนกระทั่งทุกข้อมูลถูกจัดเข้าสู่กลุ่มขนาด N อันเดียว ในขั้นตอนที่ 3 สามารถทำได้ในทางอื่นที่แตกต่างกัน ซึ่งเห็นได้ชัดในการจัด

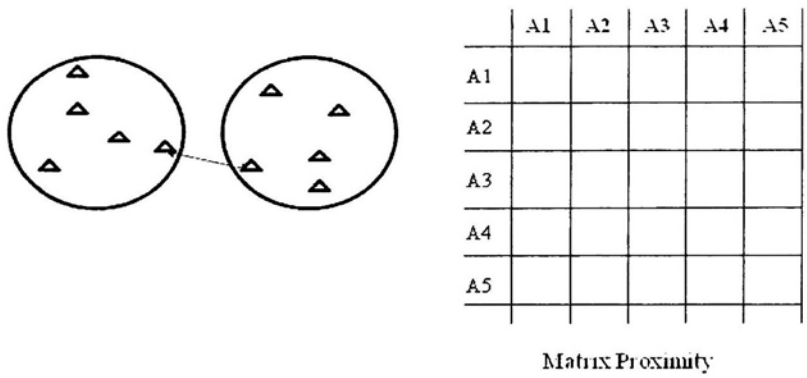
กลุ่มแบบ Single Linkage จาก Complete Linkage และ Average Linkage ดังแสดงในภาพที่ 2-5 ถึง 2-7

1.) Single Linkage: เป็นวิธีการวัดระยะที่สั้นที่สุด (Shortest Distance Method) ระยะห่างระหว่างสองคลาส จะถูกแสดงเป็นระยะห่าง (ความเหมือน) ของแต่ละคู่ของตัวอย่างที่ใกล้กันที่สุด ซึ่งถูกกำหนดจากข้อมูลแต่ละกลุ่ม เช่นเดียวกับการวัดระยะห่าง อัลกอริธึมจะสิ้นสุดโดยอัตโนมัติเมื่อระยะห่างหนึ่งอันระหว่างข้อมูลของสองคลาสที่ใกล้ที่สุดเกินกว่าอัตราที่กำหนดให้

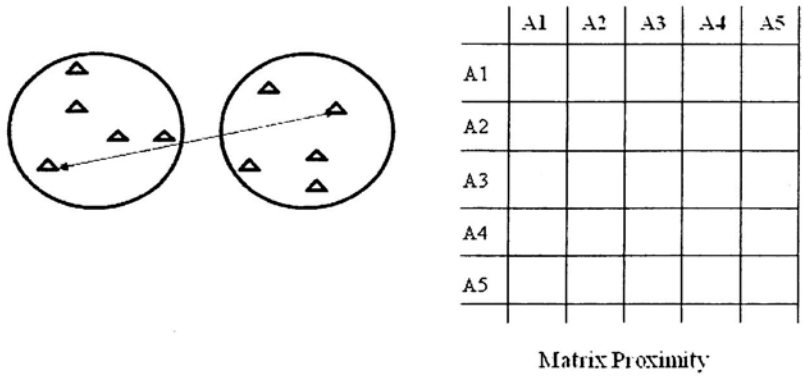
2.) Complete Linkage: เป็นวิธีการวัดระยะที่ไกลที่สุด (Longest Distance Method) วิธีการจัดกลุ่มของ Complete Linkage เป็นเหมือน Single Sinkage แต่มีนิยามของระยะห่างระหว่างสองคลาสไม่ใช่ระยะห่าง (ความเหมือน) ที่ต่ำที่สุด แต่เป็นระยะห่างที่มากที่สุดระหว่างข้อมูลสองชนิด

3.) Average Linkage: ไม่เป็นการวัดระยะห่างทั้งการวัดระยะที่ไกลที่สุด และการวัดระยะที่ใกล้ที่สุด แต่เป็นค่าเฉลี่ยของระยะห่าง

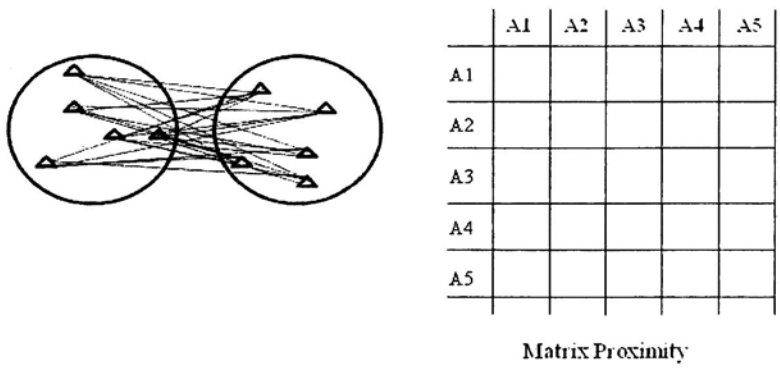
จากปัญหาของการวัดระยะห่างดังกล่าว ที่ไม่นำเอาคุณสมบัติพิเศษของข้อมูลไปใช้ อาทิ การศึกษาข้อมูลการแสดงออกของยีน, ค่าระดับการแสดงออกของยีนที่มีต้นฉบับเป็น Positive ตลอดเวลา ดังนั้น Weixiang *et al.* (2009) ได้นำเสนอวิธีการที่เพิ่มเติมลักษณะความแตกต่างสำหรับการวัดความเหมือนระหว่างข้อมูลที่ไม่เป็นลบ (Nonnegative Samples) โดยเฉพาะสำหรับข้อมูลการแสดงออกของยีน และเปรียบเทียบกับการวัดระยะห่างแบบยูคลีเดียน (Euclidean Distance) และการวัดระยะห่างแบบซิตี้บล็อก (City-block distance) โดยใช้การจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบ Average Linkage อย่างไรก็ตามวิธีการนี้ยังขาดการทดลองกับกลุ่มข้อมูลยีนชนิดอื่น



ภาพที่ 2-5 การวัดระยะห่างแบบ Single Linkage หรือ MIN

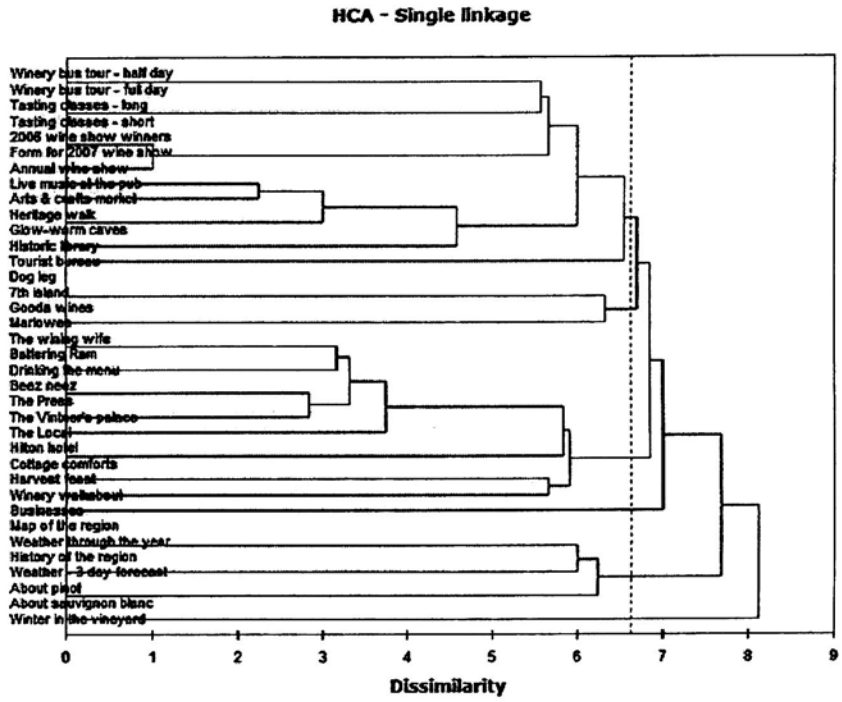


ภาพที่ 2-6 การวัดระยะห่างแบบ Complete Linkage หรือ MAX



ภาพที่ 2-7 การวัดระยะห่างแบบ Average Linkage หรือ Group Average

Garzon and Quintales (2008) ได้นำเสนอการปรับปรุงการจัดกลุ่มแบบลำดับชั้น เพื่อให้สามารถแสดงแผนภาพเดนโดแกรมที่มีคุณภาพสูง ซึ่งเป็นสิ่งที่สำคัญมาก สิ่งที่งานวิจัยนี้ประสบความสำเร็จคือ การนิยามของการจัดกลุ่มข้อมูลแบบลำดับชั้นใหม่ โดยร่วมกับขั้นตอนเชิงพันธุกรรม (Genetic Algorithm : GA) ซึ่งมีเป้าหมายเพื่อค้นหาค่าที่ดีที่สุดจากข้อมูลทั้งหมด (Global Optimum) และสร้างพื้นที่การค้นหาเดนโดแกรม (Dendrogram Search Space) บนกลุ่มข้อมูล โดยที่ความยาวของเดนโดแกรมสามารถถูกกำหนดจากจำนวนของระดับการจัดกลุ่ม แต่ในกรณีที่ดีที่สุดคือครึ่งหนึ่งจากระดับของเดนโดแกรมทั้งหมดของกลุ่มข้อมูลเดียว สิ่งที่ควรเพิ่มเติมสำหรับงานวิจัยนี้คือการยืนยันความถูกต้องซึ่งต้องใช้ผู้เชี่ยวชาญเพียง ลักษณะของแผนภาพเดนโดแกรมดังแสดงในภาพที่ 2-8



ภาพที่ 2-8 ตัวอย่างลักษณะของแผนภาพเดนโดนแกรม (Spencer, 2011)

อย่างไรก็ตามบางงานวิจัย เช่น Daxin *et al.* (2003) ได้มุ่งเน้นไปในเรื่องลำดับเวลาการแสดงออกของยีน (Time series gene expression) โดยใช้อัลกอริธึมการจัดกลุ่มแบบลำดับชั้น ที่อยู่บนความหนาแน่น (A density-based, hierarchical clustering method) ซึ่งการจัดกลุ่มข้อมูลตามความหนาแน่น มีข้อดีของการจัดกลุ่มจากสภาพแวดล้อมที่มีสิ่งรบกวนสูง (High Noisy) ซึ่งเหมาะสมกับกรณีของลำดับเวลาการแสดงออกของยีน จากนั้นนำมาแสดงในภาพของต้นไม้ความหนาแน่น (Density Tree) ซึ่งอัลกอริธึมนี้สามารถที่จะเปิดเผยกลุ่มข้อมูลที่ซ่อนอยู่ได้ อย่างไรก็ตาม วิธีนี้ยังต้องอาศัยการกำหนดค่าพารามิเตอร์ที่สำคัญคือ ค่าอัตราความเหมือน (Similarity threshold : T) และจำนวนของข้อมูลที่ต่ำสุด (Minpts)

วิธีการจัดกลุ่มข้อมูลแบบลำดับชั้นมักนิยมนำมาใช้ในการปฏิบัติโดยนักชีววิทยาเนื่องจากวิธีนี้สามารถแสดงโครงสร้างของการจัดกลุ่มได้มากกว่าวิธีอื่น อย่างไรก็ตาม วิธีการจัดกลุ่มแบบลำดับชั้นมีข้อเสียดังต่อไปนี้

1) วิธีการจัดกลุ่มแบบลำดับชั้นบางครั้งมีความบอบบางในการกำหนดการตัดเดนโดแกรมที่ได้จากการจัดกลุ่ม ซึ่งต้องอาศัยความรู้ของผู้เชี่ยวชาญในการตรวจสอบ

2) วิธีการจัดกลุ่มแบบลำดับชั้นยากในการบอกโครงสร้างภายในของการจัดกลุ่มจากแผนภาพเดนโดแกรม

3) วิธีการจัดกลุ่มแบบลำดับชั้นถูกพิจารณาว่าขาดความทนทานและไม่เป็นหนึ่งเดียว วิธีเหล่านี้อาจจะไวต่อลำดับของมิติข้อมูลเข้า และไม่เหมาะกับข้อมูลที่มีมิติจำนวนมาก

2.3.2 การจัดกลุ่มข้อมูลแบบแบ่งส่วน (Partition Clustering)

อัลกอริธึมการจัดกลุ่มข้อมูลแบบแบ่งส่วน จะมีรูปแบบของการแบ่งข้อมูลเป็นส่วนแทนที่แบบโครงสร้างการจัดกลุ่มแบบลำดับชั้น วิธีการจัดข้อมูลแบบแบ่งส่วนมีข้อดีในเรื่องการทำงานกับกลุ่มข้อมูลที่มีจำนวนมากซึ่งช่วยลดการใช้เวลาในการคำนวณ ปัญหาหลักที่สำคัญคือการเลือกจำนวนผลลัพธ์ของกลุ่มที่ต้องการ (Jain *et al.*, 1999)

โดยทั่วไปการจัดกลุ่มแบบแบ่งส่วนจะสร้าง k กลุ่ม โดยแบ่งข้อมูลจำนวน n ข้อมูลให้อยู่ในแต่ละกลุ่ม k ทั้งนี้แต่ละ k กลุ่มจะต้องมีค่า $k \leq n$ และในแต่ละกลุ่มจะมีข้อมูลอย่างน้อย 1 ข้อมูล ทั้งนี้ข้อมูลจะถูกกำหนดให้อยู่ในกลุ่ม k ใดเพียงกลุ่มเดียวเท่านั้น

2.3.2.1 การจัดกลุ่มแบบเคมีน (K-Means Clustering)

วิธีการจัดกลุ่ม K-Means เป็นเทคนิคการจัดกลุ่มแบบแบ่งส่วนที่มีประสิทธิภาพและถูกนำไปใช้ในหลายงาน ผู้อ้างอิงถึงวิธีนี้เป็นครั้งแรกคือ Macqueen ในปี 1967 อัลกอริธึม K-Means สามารถค้นหาข้อมูลโดยการใช้การจัดกลุ่มแบบแบ่งส่วน โดยมีพื้นฐานจากการรู้จำภาพแบบ (Pattern Recognition) (Jain *et al.*, 1999) อัลกอริธึม K-means จะทำการจัดกลุ่มยืนเข้าโดยการวัดอัตราระยะทางที่ใกล้ที่สุดภายในแต่ละกลุ่มข้อมูล และระยะทางที่ไกลที่สุดระหว่างกลุ่มข้อมูล กระบวนการของวิธีนี้เริ่มจากการเลือก k ยืนที่ถูกเรียกว่าศูนย์กลาง หรือเซ็นทรอยต์ (Centroid) จากข้อมูลทั้งหมด ซึ่งการเลือกนี้อาจจะทำการสุ่มเลือกหรือถ้าแน่ใจว่าข้อมูลการแสดงออกของยีนแตกต่างกันมาก ก็สามารถกำหนดค่า k ตามข้อมูลการแสดงออกของยีนได้เช่นกัน จากนั้นทำการทดสอบกับข้อมูลการแสดงออกของยีนทั้งหมดและกำหนดแต่ละยีนให้เป็นสมาชิกของกลุ่มข้อมูลที่ใกล้กัน โดยวัดระยะทางจากศูนย์กลางของแต่ละกลุ่มข้อมูลกับยีนว่ามีกลุ่มข้อมูลใดที่ใกล้มากที่สุด ตำแหน่งของศูนย์กลางจะถูกคำนวณใหม่ในแต่ละครั้งที่ข้อมูลการแสดงออกของยีนถูกเพิ่มให้กับกลุ่มข้อมูล และการคำนวณนี้จะกระทำโดยการหาค่าเฉลี่ยของข้อมูลที่มีอยู่ในกลุ่มข้อมูลทั้งหมด กระบวนการเหล่านี้จะถูกทำซ้ำไปมาจนกระทั่งกลุ่มข้อมูลไม่มีการเปลี่ยนแปลง และไม่มีข้อมูลการแสดงออกของยีนสลับไปมาระหว่างกลุ่มข้อมูลในการทำซ้ำ อัลกอริธึม K-Means ดังแสดงในภาพที่ 2-9 นอกจากนี้ยังใช้เวลาในการคำนวณมากกว่าวิธีการจัดกลุ่มข้อมูลแบบลำดับชั้นอีกด้วย (Amiy and Stojmenovic, 2008)

สำหรับการวัดคุณภาพของอัลกอริธึม K-Means ใช้การวัดผลรวมอัตราความผิดพลาดกำลังสอง (Sum of square error) เพื่อวัดคุณภาพของการจัดกลุ่ม ซึ่งเกิดจากการรวมผลลัพธ์ของระยะทางของข้อมูลทั้งหมดในแต่ละกลุ่ม จากนั้นนำผลลัพธ์ของแต่ละกลุ่มมารวมกันแล้วตัวที่มีค่าน้อยที่สุด

จึงจะถือว่ามีคุณภาพที่เหมาะสม (Jain *et al.*, 1999; Jiawei and Micheline, 2000) สำหรับสมการเพื่อหาผลรวมอัตราความผิดพลาดกำลังสองแสดงในสมการที่ 2-7 ดังนี้

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} | p - m_i |^2 \qquad (2\text{-}7)$$

โดยที่ E คือ ผลรวมของค่าความผิดพลาดกำลังสองของข้อมูล

m_i คือ ค่าเฉลี่ยของกลุ่มข้อมูลที่ C_i และ

p คือ ข้อมูล

ข้อมูลจะถูกประเมินโดยการใช้ระยะห่างกำลังสองระหว่างข้อมูลและจุดศูนย์กลางของแต่ละกลุ่มข้อมูล และนำค่าทั้งหมดมารวมกัน สำหรับอัลกอริธึมนี้จะสามารถทำงานได้อย่างมีประสิทธิภาพสูงเมื่อข้อมูลมีการกระจายตัวแยกออกจากกันเป็นกลุ่มอย่างชัดเจน สำหรับระยะเวลาในการทำงานจะขึ้นอยู่กับจำนวนข้อมูล n การกำหนดจำนวนกลุ่มข้อมูล k และเวลาที่ใช้ในการวนซ้ำ t ดังสมการที่ 2-8

$$Time\ consuming = = n \times k \times t \qquad (2\text{-}8)$$

อย่างไรก็ตามข้อเสียของอัลกอริธึม K-Means อัลกอริธึมที่สำคัญคือ การกำหนดค่าพารามิเตอร์ k ซึ่งในกรณีของการทำงานกับกลุ่มข้อมูลการแสดงออกของยีนนั้น บางครั้งถ้าผู้ใช้งานไม่ได้มีความรู้มาก่อน (Prior Knowledge) หรือไม่ได้เป็นผู้เชี่ยวชาญก็เป็นการยากที่จะสามารถกำหนดได้อย่างถูกต้อง ซึ่งการระบุค่า k ที่ไม่เหมาะสมอาจจะทำให้เพิ่มเวลาในการทำงานหรือได้ผลลัพธ์ของ การจัดกลุ่มข้อมูลที่มีคุณภาพต่ำอีกด้วย นอกจากนี้ผลลัพธ์จากการจัดกลุ่มข้อมูลด้วย อัลกอริธึม K-Means จะมีลักษณะทรงกลม (Spherical Shape) ทำให้ไม่เหมาะสมกับการนำมาใช้กับการจัดกลุ่มข้อมูลการแสดงออกของยีน เนื่องจากบางครั้งกลุ่มของยีนอาจจะมีการแสดงออกได้ในหลายรูปทรง (Arbitrary Shape) และข้อเสียที่สำคัญอีกอย่างหนึ่งคือความไวต่อสิ่งรบกวน จะส่งผลให้การจัดกลุ่มมีความผิดพลาดเกิดขึ้น ซึ่งโดยส่วนใหญ่แล้วข้อมูลการแสดงออกของยีนจะมีสิ่งรบกวนอยู่มากเป็นพื้นฐาน

อย่างไรก็ตาม Daxin *et al.* (2003) ได้สรุปข้อเสียของ K-Means ไว้ดังนี้

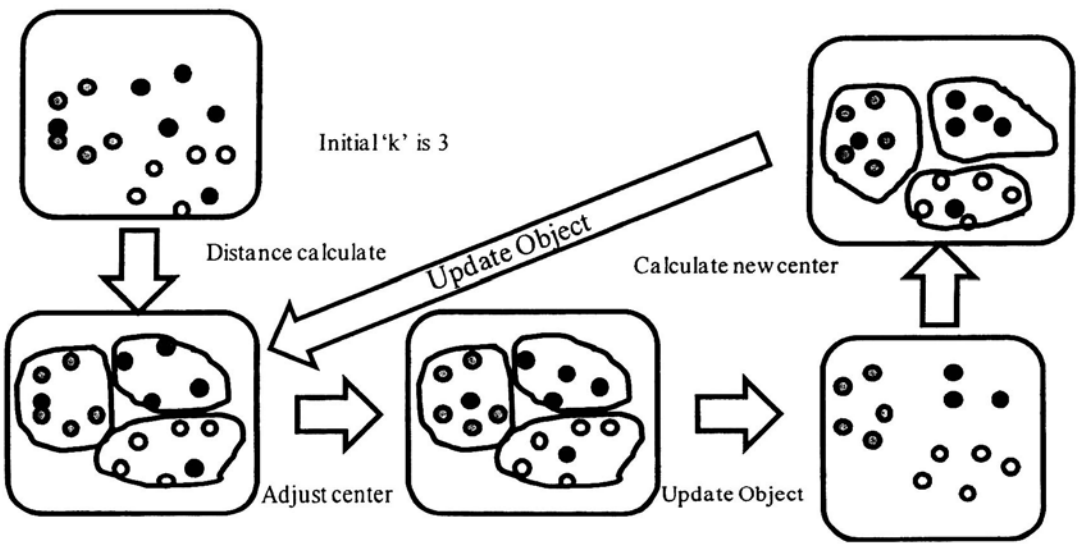
1) อัลกอริธึม K-Means ต้องการให้ผู้ใช้งานกำหนดจำนวนของพารามิเตอร์ จำนวนของกลุ่ม

2) เนื่องจากการจัดกลุ่มบ่อยครั้งเป็นงานที่ต้องสำรวจการวิเคราะห์ข้อมูลเริ่มต้นของกลุ่ม

ข้อมูลการแสดงออกของยีน บ่อยครั้งไม่เหมาะสมในการใช้งาน

3) การจัดกลุ่มแบบแบ่งส่วนพยายามบังคับให้แต่ละข้อมูลเข้าสู่กลุ่ม และมีความไวต่อค่าที่ห่างจากข้อมูลปรกติมาก (Outlier)

ดังนั้นในบางกรณีเมื่อข้อมูลมีสิ่งรบกวนปนเปื้อน อาจจะให้ผลลัพธ์ของการจัดกลุ่มที่ไม่แม่นยำ ดังนั้น Wang *et al.* (2007) จึงนำเสนอวิธีการลดสิ่งรบกวน ซึ่งได้แบ่งสถานะของสิ่งรบกวนออกเป็น 3 กลุ่มคือ ข้อมูลที่ไม่สมบูรณ์ (Incomplete Data) ข้อมูลสิ่งรบกวน (Noisy Data) และกรณีการฝึกฝนที่มีการกระจายตัวไม่เท่ากัน งานวิจัยดังกล่าวได้นำเสนอวิธีการที่เรียกว่า GK-Means ซึ่งประยุกต์มาจากอัลกอริธึม K-Means โดยใช้วิธีการวัดความเหมือนเพื่อจัดการกับสิ่งรบกวน โดยการเปรียบเทียบกับกลุ่มข้อมูล และทดลองกับกลุ่มข้อมูล UCI Machine Learning Repository สำหรับข้อเสียของวิธีนี้คือ มิติของข้อมูลที่มากขึ้นจะทำให้ความแม่นยำในการจัดกลุ่มลดต่ำลง



ภาพที่ 2-9 ตัวอย่างกระบวนการทำงานของ K-Means อัลกอริธึม

2.3.2.2 การจำแนกหมวดหมู่แบบหาจุดศูนย์กลางที่ใกล้กันมากที่สุด (Classification Nearest Centroid: CalNC)

ClaNC ถูกนำเสนอโดย Dabney (2005) วิธีการนี้จะทำการจัดอันดับของยีน โดยการใช้การทดสอบสถิติแบบ T-statistics วิธีการนี้ไม่มีผลให้ศูนย์กลางข้อมูลเล็กลง (Shrink Centroid) นอกจากนี้ยังใช้กระบวนการเลือกยีนเพื่อระบุคลาส (Class-Specific) วิธี ClaNC เป็นตัววิธีการจัดกลุ่มข้อมูลแบบแบ่งส่วนและสามารถทำการจำแนกหมวดหมู่ของคลาส ซึ่งวิธีการนี้ใช้เทคนิคการ

วิเคราะห์อำนาจจำแนกแบบเชิงเส้น (Linear Discriminate Analysis : LDA) สำหรับคุณสมบัติของวิธีการ CalNC มีดังนี้

1) CalNC ไม่ทำให้ศูนย์กลางของกลุ่มข้อมูลเล็กลง

2) ใช้สถิติแบบ T-statistics ในการคัดเลือกยีน

3) พิจารณามิติของยีนที่สามารถระบุคลาสได้

4) อนุญาตให้แต่ละยีนสามารถอยู่ได้หลายคลาส

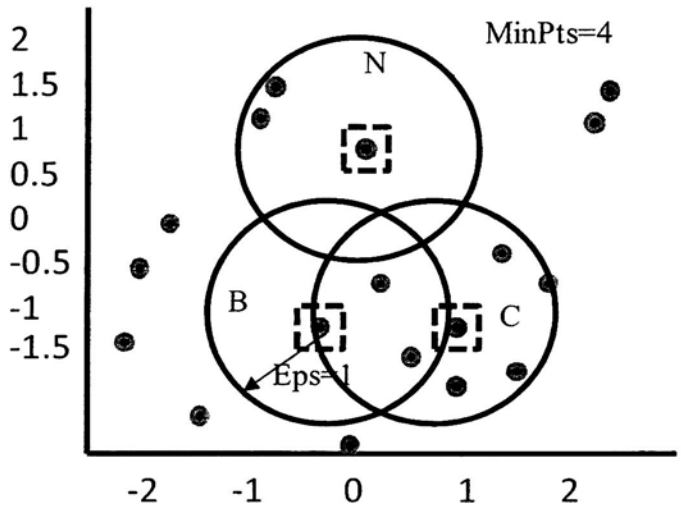
สำหรับวิธีการ LDA เป็นวิธีการการทำนายคลาสแบบดั้งเดิม มีลักษณะการทำงานที่สามารถเข้าใจได้ง่าย และแสดงประสิทธิภาพที่ดีกับข้อมูลไมโครอะเรย์ แต่ละคลาสจะถูกระบุมิติของข้อมูลโดยเวกเตอร์ของค่าเฉลี่ยของคลาส หรือจุดศูนย์กลาง นอกจากนี้ข้อมูลการแสดงออกของยีนที่ไม่ทราบคลาสจะถูกประเมิน โดยการคำนวณขนาดระยะห่างระหว่างข้อมูลการแสดงออกของยีนนั้นกับจุดศูนย์กลางแต่ละคลาส ข้อมูลการแสดงออกของยีนจะถูกกำหนดให้กับคลาสที่มันอยู่ใกล้ดังนั้นวิธีการ LDA สามารถถูกนำเอาไปเป็นแนวคิดของการจัดหมวดหมู่แบบหาจุดศูนย์กลางที่ใกล้กันมากที่สุด นอกจากนี้การเจาะจงยีนจำนวนน้อยเป็นสิ่งจำเป็นสำหรับการจำแนกหมวดหมู่ ดังนั้นวิธีการนี้จึงเพิ่มขั้นตอนการคัดเลือกมิติข้อมูลเข้าไปด้วย ซึ่งยีนที่มีการแสดงออกทั้งในเนื้อเยื่อที่ปรกติและไม่ปรกติอาจจะมีการแสดงออกทั้งสองค่าสถิติทดสอบแบบที ดังนั้นค่าที่มากที่สุดนั้นจะถูกเลือก อย่างไรก็ตามวิธีการนี้ยังต้องการการปรับปรุงในการคัดเลือกกลุ่มย่อยของยีนให้มีประสิทธิภาพ เนื่องจากยังคงใช้วิธีการสถิติแบบ T-statistic แบบดั้งเดิมอยู่

2.3.2.3 การจัดกลุ่มตามความหนาแน่น (Density Clustering)

ปัญหาที่สำคัญอย่างหนึ่งที่ทำให้ประสิทธิภาพของเทคนิคการจัดกลุ่มต่าง ๆ ลดลงคือ ข้อมูลรบกวน หรือ Noise Data เช่นในการจัดกลุ่มด้วยการวิธีแบ่งส่วนจะอาศัยสมการการวัดระยะทางระหว่างกลุ่มข้อมูลเป็นตัวพิจารณาการจัดกลุ่ม ซึ่งส่งผลให้เกิดข้อจำกัดในเรื่องของรูปทรงของกลุ่มที่มักจะมีโอกาสผิดพลาดถ้าเกิดมีข้อมูลรบกวนปะปนมา ดังนั้นเพื่อแก้ปัญหานี้จึงมีการนำเสนอการจัดกลุ่มข้อมูลด้วยความหนาแน่น ซึ่งการจัดกลุ่มด้วยเทคนิคนี้ใช้การวัดความหนาแน่นของจำนวนข้อมูลในตำแหน่งใด ๆ และขยายขอบเขตหรือรัศมีของกลุ่มไปตามจำนวนข้อมูลและระยะห่างที่ได้กำหนดไว้ ทำให้เกิดการค้นพบรูปทรงของกลุ่มที่แตกต่างกัน โดยที่ข้อมูลที่อยู่พื้นที่เดียวกันจะถูกเรียกว่าข้อมูลเพื่อนบ้าน (Neighborhood Data) เทคนิคการจัดกลุ่มตามความหนาแน่นนอกจากจะช่วยในเรื่องของรูปทรงแล้วยังช่วยในเรื่องของปัญหาข้อมูลรบกวน (Noise Data) ที่ทำให้เกิดความผิดพลาดในการจัดกลุ่ม

อัลกอริธึมดีบีสแกน (Density Based Spatial Clustering of Application with Noise: DBSCAN) คิดค้นโดย Ester ในปี 1996 อัลกอริธึมนี้จะค้นหารูปทรงได้หลากหลาย โดยมีสอง

พารามิเตอร์ที่ต้องกำหนดไว้ก่อนคือ จำนวนข้อมูลที่อยู่ในรัศมีเดียวกัน (Min Point: MinPts) และค่ารัศมี (Epsilon: Eps) นอกจากนี้อัลกอริธึมดีบีสแกนจะกำหนดกลุ่มที่เกิดจากการวัดจำนวนข้อมูลภายใต้รัศมีคือ 1) ข้อมูลหลัก (Core Point) หมายถึง ข้อมูลที่อยู่ในบริเวณรัศมีมีจำนวนข้อมูลสูงกว่า MinPts ที่กำหนด 2) ข้อมูลบริเวณขอบ (Border Point) หมายถึง ข้อมูลที่อยู่ในบริเวณรัศมีมีจำนวนข้อมูลน้อยกว่า MinPts ที่กำหนด และ 3) ข้อมูลรบกวน (Noise Point) หมายถึง จุดที่ไม่เป็น Core Point และ Border Point ดังแสดงในภาพที่ 2-10 อย่างไรก็ตาม DBSCAN ยังประสบปัญหาในเรื่องการทับซ้อนของข้อมูล และกรณีที่ความหนาแน่นมีหลากหลายอาจทำให้มีประสิทธิภาพไม่เพียงพอ

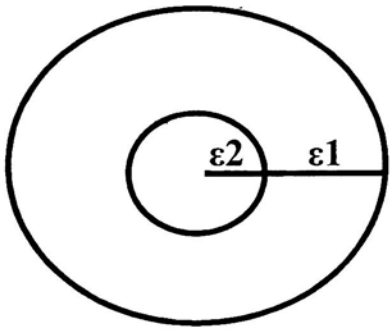


ภาพที่ 2-10 การจัดกลุ่มตามความหนาแน่นด้วยเทคนิคดีบีสแกน

อัลกอริธึมออพติก (Ordering Points to Identify the Clustering Structure: OPTIC) คิดค้นโดย Mihael ในปี 1999 เพื่อแก้ไขปัญหาของเทคนิคดีบีสแกน โดยเทคนิคออพติกจะมีการกำหนดค่าเอปไซลอน (ε) อยู่ถึง 2 ค่าคือ ความหนาแน่นสูง และความหนาแน่นต่ำ ดังนั้นวิธีการนี้จึงเป็นการขยายความสามารถของอัลกอริธึมดีบีสแกนได้อย่างมีประสิทธิภาพ ดังแสดงในภาพที่ 2-11

2.3.3 การจัดกลุ่มข้อมูลแบบการเรียนรู้เพิ่มขึ้น (Incremental Learning Clustering based)

การเรียนรู้แบบเพิ่มขึ้น (Incremental Learning) เป็นวิธีการหนึ่งที่ช่วยให้อัลกอริธึมสามารถเรียนรู้รูปแบบข้อมูลใหม่ที่ไม่เคยรู้จักมาก่อน รวมถึงการเก็บรูปแบบของข้อมูลที่เคยเรียนรู้ในอดีต



ภาพที่ 2-11 การเพิ่มการวัดรัศมีด้วยเทคนิคออพติก

มาแล้ว ดังนั้นการเรียนรู้แบบเพิ่มขึ้นจึงทำให้อัลกอริธึมมีประสิทธิภาพในการทำงานอยู่ตลอดเวลา และลดความสิ้นเปลืองในการใช้ทรัพยากร วิธีการดังกล่าวไม่ใช่การนำข้อมูลทั้งหมดมาทำซ้ำในหน่วยความจำ เพียงแต่ขยายกระบวนการซ้ำของข้อมูลเหล่านี้ โดยทั่วไปแล้ววิธีการส่วนใหญ่สำหรับการเรียนรู้แบบเพิ่มขึ้น จะทำซ้ำอย่างน้อยบางข้อมูลเช่นในโหนดสุดท้าย ฉะนั้นการเรียนรู้แบบเพิ่มขึ้นจะนำไปสู่การเรียนรู้แบบที่เป็นธรรมชาติ

2.3.4.1 อัลกอริธึมการจัดกลุ่มแบบการเรียนรู้เพิ่มขึ้น (COBWEB)

COBWEB พัฒนาโดย Fisher (1987) มีหลักการทำงานแบบการเรียนรู้เพิ่มขึ้น อัลกอริธึมนี้จะสร้างการจัดกลุ่มข้อมูลในภาพของต้นไม้ (Tree) นำเสนอในภาพแบบโครงสร้างลำดับชั้น (Hierarchical) ข้อดีของวิธีนี้คือ เป็นการเรียนรู้แบบเพิ่มขึ้น และสามารถนำมาใช้แสดงแนวคิดการจัดกลุ่มของยีนในลักษณะของโครงสร้างลำดับชั้น (Michalski and Tecuci, 1999)

นอกจากนี้อัลกอริธึมการจัดกลุ่มแบบการเรียนรู้เพิ่มขึ้นชนิดนี้ ยังเป็นอัลกอริธึมที่พัฒนามาจากวิธีการสร้างแนวคิด (Conceptual Learning) และการอาศัยการเรียนรู้จากตัวอย่างสิ่งแวดล้อม ซึ่งแตกต่างจากวิธีการเดิมที่มุ่งจะศึกษาเฉพาะกลไกการจัดกลุ่มเพียงอย่างเดียวเท่านั้น การเรียนรู้จากตัวอย่างสิ่งแวดล้อมประกอบด้วยวิธีการย่อย 2 วิธีคือ การควบคุมการค้นหา (Search Control) เป็นส่วนของอัลกอริธึมที่ใช้สำหรับการควบคุมการค้นหา และทิศทางการค้นหา (Search Direction) คือแนวคิดการสังเกตที่ถูกทำเป็นลำดับไป (Fisher, 1987)

อัลกอริธึม COBWEB ใช้วิธีการแก้ปัญหาแบบวิทยาการศึกษาสำนึก หมายถึง การแก้ปัญหาที่ดูเหมือนว่าจะไม่มีแนวทางหรือกฎเกณฑ์ที่ชัดเจนตายตัว และมักจะใช้ผลจากประสบการณ์ที่ผ่านมา ซึ่งในฟังก์ชันใช้วิธีการแก้ปัญหานี้ถูกเรียกว่าฟังก์ชันจำแนกหมวดหมู่ (Category Utility: CU) ซึ่งถูกคิดค้นโดย Gluck และ Corter ในปี 1995 (Witten *et al.*, 2011)

โดยที่ฟังกชัน CU มีประโยชน์เพื่อเป็นตัวนำทางในการค้นหาเพื่อจัดกลุ่มและใช้การแลกเปลี่ยนระหว่างความเหมือนในคลาส (Intra-class) และความแตกต่างระหว่างคลาส (Inter-

class) จากค่าความน่าจะเป็น (Probability) ซึ่งวัตถุจะถูกอธิบายในลักษณะของค่าของแอททริบิวต์ ซึ่งความเหมือนในคลาสจะมีภาพแบบค่าความน่าจะเป็นคือ $P(A_i = V_{ij} \mid C_k)$, ที่ $A_i = V_{ij}$ คือค่าของแอททริบิวต์ และความแตกต่างระหว่างคลาสคือ $P(C_k \mid A_i = V_{ij})$ นอกจากนี้ยังใช้ตัวดำเนินการในการรวม (Merging) และแยก (Splitting) สำหรับแต่ละ Node ดังแสดงในภาพที่ 2-12 และ 2-13 ตามลำดับ ดังสมการที่ 2-9 สำหรับการสรุปเทคนิคการคัดเลือกโดยใช้ฟังก์ชัน *CU*

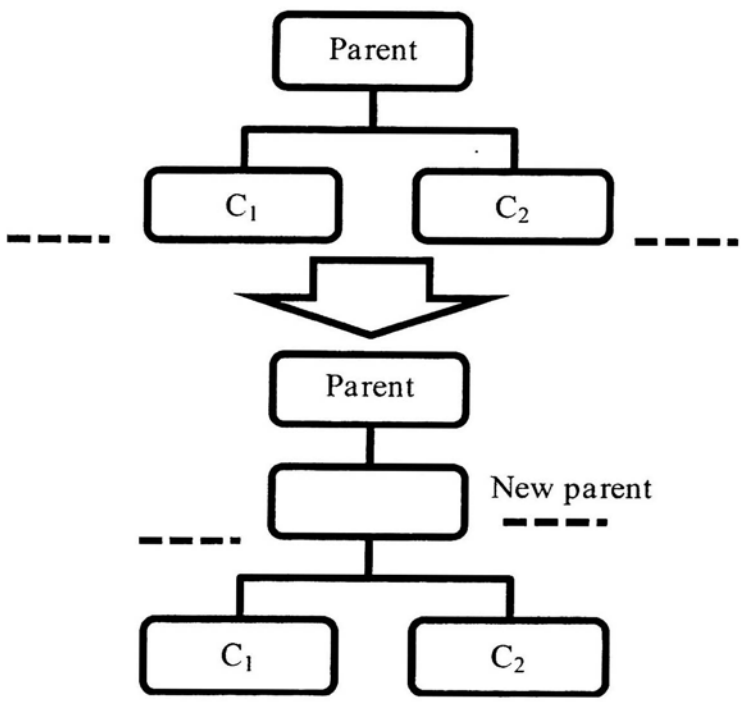
$$CU = (C_1, C_2, ..., C_n) = \frac{\sum_{k=1}^{n} P(C_k)[\sum_i \sum_j P(A_i = V_{ij} \mid C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2]}{n} \quad (2\text{-}9)$$

โดยกำหนดให้ k คือ กลุ่มข้อมูล

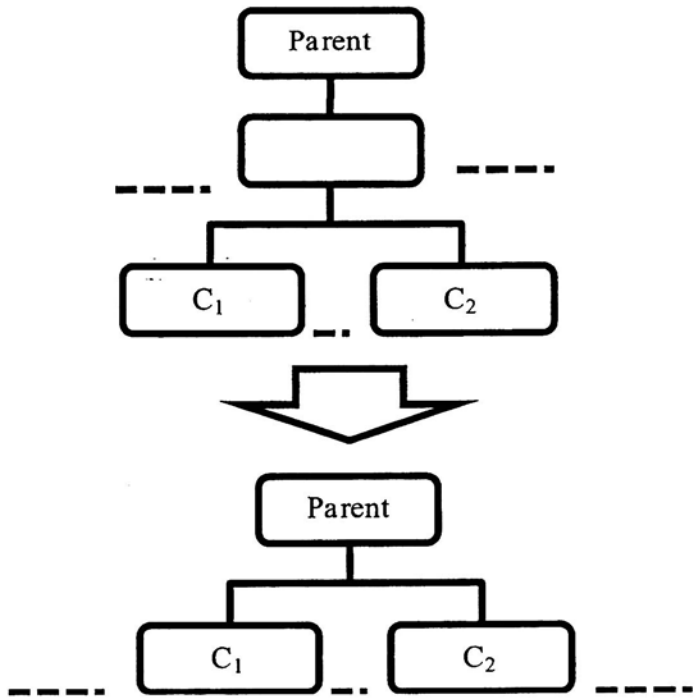
i คือ แอททริบิวต์

j คือ ค่าของแอททริบิวต์

n คือ จำนวนของการแบ่งพาร์ติชั่น



ภาพที่ 2-12 การรวมกันของโหนดจากอัลกอริธึม COBWEB



ภาพที่ 2-13 การแยกกันของโหนดจากอัลกอริธึม COBWEB

อัลกอริธึม COBWEB ได้ถูกพัฒนาต่อมาโดย Grnnari ในปี 1990 เป็นอัลกอริธึมชื่อ CLASSIT อัลกอริธึมนี้มีข้อดีในการจัดการข้อมูลที่เป็นชนิดตัวเลข (Numberic Data Type) อย่างไรก็ตามทั้ง COBWEB และ CLASSIT ก็อยู่บนพื้นฐานการวัดฟังก์ชัน CU ดังแสดงในภาพที่ 2-14

นอกจากฟังก์ชัน CU ที่ถูกใช้งานกับแอททริบิวต์ชนิด Norminal ในอัลกอริธึม COWEB แล้ว อัลกอริธึม CLASSIT ขยายการใช้งานให้ครอบคลุมกับแอททริบิวต์ชนิดตัวเลขได้โดยสมมติฐานในเรื่องของการกระจายตัวที่เป็นปรกติที่ทำให้ได้รับค่าเฉลี่ย (Mean: μ) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: σ) โดยที่ฟังก์ชันความหนาแน่นที่น่าจะเป็นสำหรับแอททริบิวต์ A ดังสมการที่ 2-10

$$f(A) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(a-\mu)^2}{2\sigma^2}) \qquad (2\text{-}10)$$

ซึ่งความเหมือนกันของ CU แบบเดิมคือการรวมด้วยเลขยกกำลังของแอททริบิวต์ กับค่าความน่าจะเป็นของ CU แบบการใช้งานกับตัวเลขจะเป็นดังสมการที่ 2-11

$$\sum_j \Pr[a_i = v_{ij}]^2 \Leftrightarrow \int f(A_i) 2da_i = \frac{1}{2\sqrt{\pi\sigma_i}} \qquad (2\text{-}11)$$

```
COBWEB algorithm
    COBWEB (root, record)
   Input: A COBWEB node root, a instance insert record
   If root has no children then
      Children = {copy(root)}
      Newcategory (record)   \\add child with record's
                                        \\...feature values
      Insert (record, root)  \\update root's statistics
   Else
         Insert (record, root)
    For child in root's children do
         Calculate category utility for insert (record, child)
         Set best1 best2 children w. best Category Utility.
     End for
     If newcategory (record)  yields best CU then
        Newcategory(record)
Else if merge (best1, best2) yields best CU then
         Merge (best1, best2)
         COBWEB (root, record)
     Else if split (best1) yield best CU then
         Split (best1)
         COBWEB (root, record)
     Else
         COBWEB(best1, record)
    End if
End
```

ภาพที่ 2-14 อัลกอริธึมการจัดกลุ่มข้อมูลแบบเรียนรู้แบบเพิ่มขึ้น COBWEB

โดยที่ σ_i คือค่าส่วนเบี่ยงเบนมาตรฐานของแอททริบิวต์ A_i ดังนั้นสรุปได้ว่าการประมาณการค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลแอททริบิวท์ชนิดตัวเลขทำได้จากตัวข้อมูล โดยทำทั้งภายในกลุ่ม σ_{ij} และข้อมูลจากทุกกลุ่ม σ_i และนำไปใช้ในสมการ 2-12

$$CU(C_1, C_2, ..., C_n) = \frac{1}{n}\sum_k \Pr[C_k]\frac{1}{2\sqrt{\pi}}\sum_i (\frac{1}{\sigma_{il}} - \frac{1}{\sigma_i}) \quad (2\text{-}12)$$

อย่างไรก็ตามปัญหาในเรื่องของการประมาณส่วนเบี่ยงเบนมาตรฐานที่มีค่าเป็นศูนย์กลางมาเป็นอีกปัญหาหนึ่ง ซึ่งการที่ค่าส่วนเบี่ยงเบนมาตรฐานมีค่าเป็นศูนย์ทำให้เกิดการคำนวณที่ไม่สิ้นสุดในฟังก์ชัน *CU* ดังนั้นการแก้ปัญหาแบบทั่วไปจึงมีการกำหนดค่าความแปรปรวนขั้นต่ำของแต่ละแอททริบิวต์เอาไว้ โดยเรียกว่า *acuity* (Witten *et al.*, 2011)

2.3.4.2 อัลกอริธึมการเรียนรู้แบบเพิ่มขึ้นบนการวัดระยะทางแบบมาหาลาโนบิส (Incremental Learning Based on Mahalanobis Distance: ILM)

อัลกอริธึมการจำแนกหมวดหมู่ (Classification) ที่อยู่บนพื้นฐานของการเรียนรู้แบบเพิ่มขึ้นมีอยู่หลายวิธีการ อาทิ ฟัซซี่อาร์ทแมบเน็ตเวิร์ก (Fuzzy ARTMAP network) (Carpenter *et al.*, 1992) ซิมพละไฟฟัซซี่อาร์ทแมบเน็ตเวิร์ก (Simplified Fuzzy ARTMAP network: SFAM) (Vuskovic and Sijiang, 2002) เครือข่ายฟัซซี่แบบเรียนรู้เพิ่มขึ้น (Incremental Learning Fuzzy Neural: ILFN) (Meesad and Yen, 2001) และวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิส (Incremental Learning based on Mahalanobis Distance: ILM) (Komkhao, 2007) ซึ่งถูกพัฒนาต่อจากอัลกอริธึม ILFN เพื่อแก้ปัญหาในเรื่องของรูปทรงจากการจัดกลุ่มข้อมูลที่มักมีขนาดที่แตกต่างกัน

พื้นฐานของอัลกอริธึม ILM เป็นการจัดกลุ่มแบบเรียนรู้เพิ่มขึ้น โดยสามารถทำงานได้ทั้งการจัดกลุ่ม (Clustering) และจำแนกหมวดหมู่ (Classification) อัลกอริธึมนี้มีพื้นฐานมาจากแนวคิดของโครงข่ายประสาทเทียม (Neural Network Concept) และถูกพัฒนาโดย Komkhao ในปี 2007 ILM สนับสนุนทั้งการเรียนรู้แบบมีผู้สอน และแบบไม่มีผู้สอน เทคนิค ILM ใช้แนวคิดของการเรียนรู้แบบเพิ่มขึ้นซึ่งเป็นการเรียนรู้โดยปราศจากการลืมความรู้ที่มีอยู่ก่อน รวมถึงการใช้ประโยชน์การวัดระยะแบบมาหาลาโนบิส และฟังก์ชันเกาส์เซียนเรดิอัลเบสิก (Gaussian Radial Basis Function) เพื่อออกแบบค่าความเป็นสมาชิก (Membership Values) (Komkhao, 2007) อย่างไรก็ตามประสิทธิภาพของอัลกอริธึม ILM ขึ้นอยู่กับลักษณะการกระจายตัวของข้อมูล และการกำหนดค่าพารามิเตอร์ที่สำคัญคือ เมตริกซ์ความแปรปรวนร่วม และอัตราระยะห่าง

ก) โครงสร้างของ ILM

อัลกอริธึม ILM ประกอบด้วยขั้นตอนที่สำคัญคือ 1) ขั้นตอนการเรียนรู้ (Learning Phase) และ2) ขั้นตอนการทำนายผล (Predicting Phase) ดังแสดงในภาพที่ 2-15 ในขั้นตอนการเรียนรู้ประกอบอัลกอริธึมที่สำคัญ 2 อัลกอริธึม ได้แก่ อัลกอริธึมการเรียนรู้ (Learning Algorithm) และ

อัลกอริธึมการเรียนรู้แบบเพิ่มขึ้น (Incremental Learning Algorithm) สำหรับขั้นตอนของการทำนายผลประกอบด้วยเพียง 1 อัลกอริธึมสำหรับใช้ในการทำนาย (Prediction Mode)

ขั้นตอนการเรียนรู้ของ ILM สนับสนุนทั้งการเรียนรู้แบบมีผู้สอนและไม่มีผู้สอน ในการเรียนรู้แบบมีผู้สอน ถ้าข้อมูลการฝึกฝน (Training Data) ประกอบด้วยข้อมูลนำเข้า (Input Data) และคลาสเป้าหมาย (Target Classes) อัลกอริธึม ILM จะใช้การเรียนรู้แบบมีผู้สอนเพื่อเรียนรู้ข้อมูลและสร้างโมเดลแบบจำลองระบบ ในทางกลับกันอัลกอริธึม ILM ทำการเรียนรู้เพื่อสร้างโมเดลแบบจำลองระบบ โดยไม่ต้องอาศัยข้อมูลคลาสเป้าหมาย จะถือเป็นการเรียนรู้แบบไม่มีผู้สอน

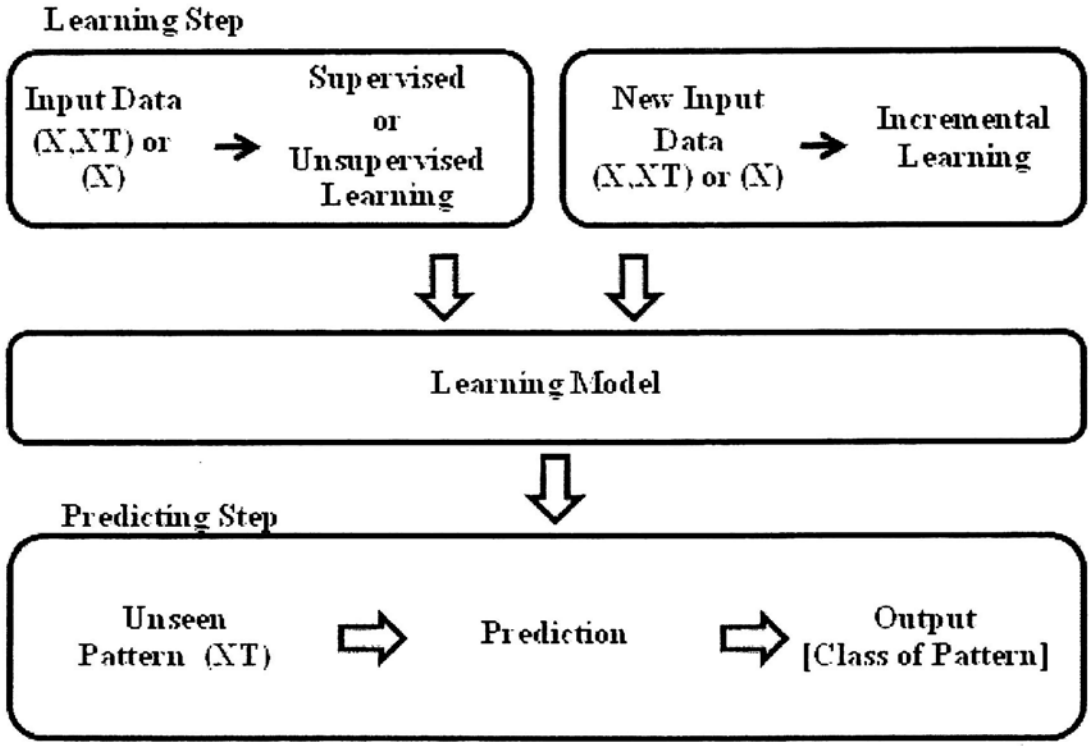
สำหรับขั้นตอนการเรียนรู้ ข้อมูลฝึกฝนจะถูกใช้เรียนรู้เพื่อสร้างโมเดลแบบจำลองสำหรับการจำแนกหมวดหมู่โดยการเรียนรู้นี้จะใช้การวัดระยะแบบมาหาลาโนบิส และฟังก์ชันความเป็นสมาชิกแบบเกาส์เซียนซึ่งการคำนวณระดับของฟังก์ชันความเป็นสมาชิกจะอาศัยฟังก์ชันความเป็นสมาชิกแบบเกาส์เซียนและใช้ฟัซซี่ระยะห่างระหว่างข้อมูลนำเข้า และโมเดลแบบจำลองซึ่งมีค่าระหว่าง 0 ถึง 1

ในขั้นตอนของการเรียนรู้แบบเพิ่มขึ้น (Myint amd Meesad, 2009; Saengsiri *et al.*, 2011 b) คลาสเป้าหมายของข้อมูลนำเข้าใหม่ (A New Class Label) จะสามารถถูกเรียนรู้ร่วมกับแบบจำลองได้ตลอดเวลา ทั้งนี้ทำให้โมเดลแบบจำลองของ ILM อัลกอริธึมสามารถเรียนรู้ข้อมูลใหม่ และปรับปรุงโมเดลได้ตลอดเวลา

ทั้งระดับของฟังก์ชันความเป็นสมาชิกที่มีค่าสูงและอัตราระยะห่าง (อัตราระยะห่างจะถูกกำหนดโดยขึ้นอยู่กับข้อมูลนำเข้าแต่ละตัว) จะถูกใช้เพื่อกำหนดโมเดลแบบจำลอง อัตราระยะห่างที่กำหนด คือ d_{th} ที่ $0 < d_{th} < 1$ การวัดระยะห่างระหว่างข้อมูลคือ d_m ถ้า $d_m > d_{th}$ ดังนั้นข้อมูลจะถูกพิจารณาว่าต่างกลุ่มกัน ในทางตรงกันข้าม $d_m \leq d_{th}$ ดังนั้นข้อมูลจะถูกพิจารณาว่าอยู่ในกลุ่มเดียวกัน

ในขั้นตอนของการเรียนรู้จะประกอบด้วยโมเดลแบบจำลองระบบ ซึ่งได้แก่ค่าน้ำหนัก W_P และ W_T ทั้งนี้ในโมเดลการเรียนรู้แบบมีผู้สอนและไม่มีผู้สอนยังสามารถเพิ่มแบบจำลองใหม่ที่ได้เรียนรู้เข้าไปได้อีก การวัดระยะห่างระหว่างข้อมูลเดิมกับข้อมูลทดสอบตัวใหม่ จะอาศัยโมเดลแบบจำลองซึ่งจะถูกวัดโดยมาหาลาโนบิส และฟังก์ชันความเป็นสมาชิกแบบเกาส์เซียนระดับความเป็นสมาชิกของแต่ละคลาสเป้าหมายจะถูกคำนวณโดยมาหาลาโนบิสเกาส์เซียน (Mahalanobis Gaussian) แบบ RBF ซึ่งถือได้ว่าเป็นคุณลักษณะของการตัดสินใจแบบอ่อน (Soft Decision) แต่สำหรับแบบจำลองที่มีระดับของฟังก์ชันความเป็นสมาชิกที่สูงกว่าจะถูกเรียกว่า โหนดผู้ชนะ หรือ Wining Node ในลักษณะแบบนี้จะถือเป็นคุณลักษณะของการตัดสินแบบเข้ม (Hard Decision) โดยที่คลาสเป้าหมายของโหนดผู้ชนะจะถูกกำหนดให้เป็นคลาสเป้าหมายของข้อมูลทดสอบ

สำหรับขั้นตอนการทำนายผล อัลกอริธึม ILM จะดำเนินการจำแนกหมวดหมู่ข้อมูลทดสอบ ซึ่งผลลัพธ์ของอัลกอริธึม ILM คือ การทำนายคลาสเป้าหมายของแต่ละข้อมูลที่ใช้ทดสอบ โดยอาศัยการวัดระยะทางแบบมาหาลาโนบิส ซึ่งจะถูกใช้ในระบบการทำนายผลรวมถึงการ วัดความเหมือน (Similarity) ระหว่างข้อมูลที่ใช้ทดสอบและโมเดลแบบจำลองระบบ (System Prototypes) ขั้นตอนการเรียนรู้ ของกระบวนการ ILM จะถูกอธิบายในส่วนถัดไป



ภาพที่ 2-15 โมเดลของอัลกอริธึมแบบ ILM

ข) กระบวนการของขั้นตอนการเรียนรู้ (Learning phases)

ในขั้นตอนนี้ข้อมูลจะถูกแบ่งเข้าเป็นข้อมูลสำหรับฝึกฝน และข้อมูลเพื่อการทดสอบ ทั้งข้อมูลฝึกฝนและทดสอบ จะถูกใช้ในขั้นตอนการเรียนรู้ โดยที่ข้อมูลสำหรับฝึกฝนจะถูกใช้เพื่อเป็นการเรียนรู้และสร้างโมเดลแบบจำลอง ในทางกลับกันข้อมูลทดสอบ จะถูกทดสอบเพื่อหาประสิทธิภาพของโมเดลแบบจำลอง

ในกรณีที่ข้อมูลฝึกฝนมีทั้งข้อมูลนำเข้า p และคลาสเป้าหมาย t การเรียนรู้แบบมีผู้สอนจะถูกใช้งาน แต่ในทางกลับกันถ้าข้อมูลฝึกฝนมีเพียงข้อมูลนำเข้า p การเรียนรู้แบบไม่มีผู้สอนก็จะถูก

นำมาใช้งาน ดังนั้นแนวคิดของการเรียนรู้แบบมีผู้สอนและไม่มีผู้สอนจะแตกต่างกัน นอกจากนี้ผลลัพธ์โมเดลแบบจำลองจากแต่ละวิธีการเรียนรู้ก็ยังให้ผลที่แตกต่างกัน

ค) กระบวนการเรียนรู้แบบมีผู้สอนประกอบด้วยขั้นตอนดังนี้

ขั้นตอนที่ 1: ความแปรปรวนร่วม $a\mathbf{I}$ และอัตราระยะห่าง (d_{th}) จะถูกกำหนดโดยผู้ใช้งาน

ขั้นตอนที่ 2: ข้อมูลนำเข้า p และคลาสเป้าหมาย t นำเข้ามา

ขั้นตอนที่ 3: ระยะห่างระหว่างข้อมูลนำเข้า p และแบบจำลองที่มีอยู่ W_p จะถูกวัดโดยการวัดระยะทางแบบมาหาลาโนบิส (d_m)

ขั้นตอนที่ 4: ค่าความเป็นสมาชิกจะถูกคำนวณ

ขั้นตอนที่ 5: โหนดผู้ชนะจะถูกกำหนดโดยการหาระดับของค่าความเป็นสมาชิกที่สูงที่สุด

ขั้นตอนที่ 6: แบบจำลองที่มีอยู่ W_p จะถูกปรับแก้โดยการใช้แนวคิดของการเรียนรู้แบบเพิ่มขึ้น

ถ้าคลาสเป้าหมาย t มีค่าเหมือนกันกับ W_T ที่โหนดผู้ชนะ และระยะห่าง $d_m <= d_{th}$ ดังนั้นคลาสเป้าหมายจะถูกบันทึกใน W_T กรณีอื่น โมเดลแบบจำลองใหม่จะถูกเพิ่มใน W_p โดยการใช้ข้อมูลนำเข้า p และจากนั้นคลาสใหม่จะถูกเพิ่มใน W_T โดยการใช้เป้าหมาย t

ง) กระบวนการเรียนรู้แบบไม่มีผู้สอนประกอบด้วยขั้นตอนดังนี้

ขั้นตอนที่ 1: ความแปรปรวนร่วม $a\mathbf{I}$ และอัตราระยะห่าง (d_{th}) จะถูกกำหนดโดยผู้ใช้งาน

ขั้นตอนที่ 2: ข้อมูลนำเข้า p จะถูกนำเข้ามา

ขั้นตอนที่ 3: ระยะห่างระหว่าง ข้อมูลนำเข้า p และโมเดลแบบจำลองที่มีอยู่ W_p จะถูกวัดโดยมาหาลาโนบิส (d_m)

ขั้นตอนที่ 4: ค่าความเป็นสมาชิกจะถูกคำนวณ

ขั้นตอนที่ 5: โหนดผู้ชนะจะถูกกำหนดโดยการหาระดับของค่าความเป็นสมาชิกที่สูงที่สุด

ขั้นตอนที่ 6: โมเดลแบบจำลองที่มีอยู่ W_p จะถูกปรับแก้โดยการใช้แนวคิดของการเรียนรู้แบบเพิ่มขึ้น

ถ้าการวัดระยะห่างของโหนดผู้ชนะ (d_m) <= d_{th} ดังนั้นคลาสเป้าหมาย จะถูกบันทึกใน W_T กรณีอื่น โมเดลแบบจำลองใหม่จะถูกเพิ่มใน W_p โดยการใช้ข้อมูลนำเข้า p และจากนั้นคลาสใหม่จะถูกเพิ่มใน W_T

จ) กระบวนการพื้นฐานสำหรับการทดสอบ

เป้าหมายของอัลกอริธึม ILM คือเพื่อทำการจำแนกหมวดหมู่ที่โดยทั่วไปเป็นการทำนาย บางครั้งต้องการให้มีอัตราความผิดพลาด (Misclassification Rate) ที่ต่ำที่สุดบนข้อมูลที่ใช้ทดสอบ

วิธีการโดยทั่วไปของการทดสอบ คือ การเรียนรู้โดยการใช้ข้อมูลทดสอบ ซึ่งข้อมูลทดสอบจะถูกจัดแบ่งทั้งในส่วนของการเรียนรู้แบบมีผู้สอนและไม่มีผู้สอน

ฉ) การเรียนรู้แบบมีผู้สอน : ภายหลังขั้นตอนการเรียนรู้ อัลกอริธึม ILM ได้รับแบบจำลองระบบ โดยอาศัยการทำนายข้อมูลที่ทราบคลาสเป้าหมาย กระบวนการของการทดสอบประสิทธิภาพของระบบมีขั้นตอนดังนี้

ขั้นตอนที่ 1: อ่านตัวข้อมูลทดสอบ (p) ตัวแรกเข้ามา

ขั้นตอนที่ 2: วัดระยะห่างระหว่าง p และ แบบจำลองที่เรียนรู้ (W_p) โดยการใช้การวัดระยะทางแบบมาหาลาโนบิส (d_m)

ขั้นตอนที่ 3: กำหนดค่าความเป็นสมาชิกโดยการใช้ฟังก์ชันเกาส์เซียนแบบ RBF

ขั้นตอนที่ 4: ค้นหาโหนดผู้ชนะ โดยการเจาะจง ค่าความเป็นสมาชิกที่มากที่สุด

ขั้นตอนที่ 5: กำหนดคลาสเป้าหมายของ p โดยการใช้คลาสเป้าหมายของโหนดผู้ชนะ

ช) โหมดการเรียนรู้แบบไม่มีผู้สอน : ภายหลังขั้นตอนการเรียนรู้อัลกอริธึม ILM ได้รับการเรียนรู้จากโมเดลแบบจำลอง โดยอาศัยการทำนายข้อมูลที่ไม่ทราบคลาสเป้าหมาย กระบวนการของการวัดประสิทธิภาพของระบบมีดังนี้

ขั้นตอนที่ 1: อ่านข้อมูลทดสอบ (p) ตัวแรกเข้ามา

ขั้นตอนที่ 2: วัดระยะห่างระหว่าง p และระบบจำลองที่เรียนรู้ (W_p) โดยการใช้การวัดระยะทางแบบมาหาลาโนบิส (d_m)

ขั้นตอนที่ 3: กำหนดค่าความเป็นสมาชิกโดยการใช้ฟังก์ชันเกาส์เซียนแบบ RBF

ขั้นตอนที่ 4: หาโหนดผู้ชนะโดยระบุค่าความเป็นสมาชิกที่มากที่สุด

ขั้นตอนที่ 5: กำหนดคลาสเป้าหมายของ p

ถ้า $d_m <= d_{th}$ ดังนั้นคลาสเป้าหมายของ p คือคลาสเป้าหมายของโหนดผู้ชนะ กรณีอื่นเพิ่มข้อมูลใหม่เข้าไปใน W_p และคลาสเป้าหมายใหม่ใน W_T ดังนั้นคลาสเป้าหมายของ p คือ คลาสเป้าหมายใหม่

ซ) กระบวนการของการทำนายผล

อัลกอริธึม ILM มีขั้นตอนการทำนายผลแยกส่วนต่างหาก สามารถทำนายคลาสเป้าหมายของข้อมูลที่ไม่ทราบคลาสเป้าหมาย โดยใช้การวัดระยะห่างระหว่างจากข้อมูลที่ไม่ทราบคลาสเป้าหมาย และข้อมูลนี้ถูกส่งให้เรียนรู้กับโมเดลแบบจำลองระบบ ทั้งนี้กระบวนการของการทำนายผล มีขั้นตอนดังนี้

ขั้นตอนที่ 1: อ่านข้อมูลที่ไม่ทราบคลาสเป้าหมาย (p)

ขั้นตอนที่ 2: วัดระยะห่างระหว่าง p และนำไปทำการเรียนรู้กับโมเดลแบบจำลอง (W_p)

โดยการใช้การวัดระยะทางแบบมาหาลาโนบิส (d_m)

ขั้นตอนที่ 3: กำหนดค่าความเป็นสมาชิก โดยการใช้ฟังก์ชันเกาส์เซียนแบบ RBF

ขั้นตอนที่ 4: ค้นหาโหนดผู้ชนะโดยการเจาะจงค่าความเป็นสมาชิกที่มากที่สุด

ขั้นตอนที่ 5: กำหนดคลาสเป้าหมายของ p โดยการใช้คลาสเป้าหมายของโหนดผู้ชนะ

ฌ) ขั้นตอนการทำงานของอัลกอริธึม ILM

1. ขั้นตอนการเรียนรู้

ขั้นตอนที่ 1: กำหนดค่าเริ่มต้นสำหรับอัตราระยะห่าง (d_{th}) โดยที่ $0 < d_{th} < 1$ และค่าเมตริกซ์ความแปรปรวนร่วม ($\mathbf{K}_{j.initial}$) ดังนี้

$$\mathbf{K}_{j.initial} = \mathrm{a}\,\mathbf{I}_\mathrm{n} \tag{2-13}$$

ที่ $\mathbf{I}_\mathrm{n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ และ $0 < a < 1$

$$\mathrm{a}\,\mathbf{I}_\mathrm{n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a \end{bmatrix}$$

ขั้นตอนที่ 2: อ่านข้อมูลนำเข้า p และคลาสเป้าหมาย t เข้ามา โดยที่ข้อมูลนำเข้า p คือกลุ่มที่จะเป็นแบบจำลองตัวแรกใน Wp และคลาสเป้าหมาย t คือคลาสเป้าหมายแรกในรูปของ W_T จากนั้นจะการกำหนดตัวนับค่าดังนี้

$$W_p = [\mathrm{p}] \tag{2-14}$$

$$W_T = [\mathrm{t}] \tag{2-15}$$

$$C_{j.initial} = 1 \tag{2-16}$$

ขั้นตอนที่ 3: อ่านข้อมูลนำเข้า (p) ตัวถัดไป

ขั้นตอนที่ 4: วัดระยะห่างระหว่าง p และ W_p โดยการใช้ เมตริกซ์ที่ไม่ใช่เอกฐานความแปรปรวนร่วม (Nonsingular Covariance Matrix) เพื่อคำนวณระยะห่างแบบมาหาลาโนบิส (d) สำหรับแนวคิดของ ($\mathbf{K}+a\mathbf{I}$) ถูกใช้เพื่อขจัดปัญหาของเมตริกซ์เอกฐาน (Singular Matrix)

$$\mathbf{K}_{j,new} = \mathbf{K}_{j,old} + \mathrm{a}\,\mathbf{I}_{\mathrm{n}} \tag{2-17}$$

$$d = \sqrt{(p - W_p)^T K^{-1} (p - W_p)} \tag{2-18}$$

ขั้นตอนที่ 5: คำนวณค่าความเป็นสมาชิก หรือค้นหาแบบ Soft Decision สำหรับแต่ละกลุ่มโดยการใช้มาหาลาโนบิส อาร์บีเอฟ (Mahalanobis RBF)

$$mem(p, W_p) = \exp\left(-\frac{(p - W_p)^T K^{-1} (p - W_p)}{2} \right) \tag{2-19}$$

ขั้นตอนที่ 6: ค้นหากลุ่มที่เป็นผู้ชนะ หรือกลุ่มที่ข้อมูลควรจะอยู่ในกลุ่มนั้นมากที่สุดโดยดูจากค่าความเป็นสมาชิกที่มากที่สุดซึ่งถูกกำหนดให้โดยตัวดำเนินการฟัซซี่ ที่เรียกว่า "OR" ดังนี้

$$winner = arg\max_i(mem_i) = arg\max_i(mem_1 \vee mem_2 \vee \ldots \vee mem_n) \tag{2-20}$$

ขั้นตอนที่ 7: ปรับปรุงค่าพารามิเตอร์ดังนี้ ค่าตัวนับ (count) ค่าเฉลี่ย (mean) และ เมตริกซ์ความแปรปรวนร่วม (covariance matrix) สำหรับกลุ่มที่ชนะ

ถ้าระยะห่างระหว่างข้อมูลนำเข้าและกลุ่มที่เป็นผู้ชนะมีน้อยกว่า หรือเท่ากับระยะห่างในอัตราที่กำหนด ดังนั้น

$$\mathbf{C}_{j,new} = \mathbf{C}_{j,old} + 1 \tag{2-21}$$

$$W_{p,new} = (1 - \beta) W_{p,old} + \beta p \quad \text{โดยที่ } \beta = 1/\,C_{j,new} \tag{2-22}$$

$$K_{j,new} = (1 - \beta) K_{j,old} + \beta (1 - \beta)^2 (p - W_{p,new})(p - W_{p,new})^T \tag{2-23}$$

ถ้าระยะห่างระหว่างข้อมูลนำเข้าและกลุ่มที่เป็นผู้ชนะสูงกว่าระยะห่างของค่าความเหมือนที่กำหนด จะทำให้เกิดการสร้างกลุ่มขึ้นมาใหม่

2. ขั้นตอนการทำนายผล

ขั้นตอนที่ 1: อ่านข้อมูลนำเข้า

ขั้นตอนที่ 2: วัดระยะห่างระหว่างข้อมูลนำเข้าและแบบจำลองของแต่ละกลุ่ม

$$d = \sqrt{(p - W_p)^T K^{-1} (p - W_p)} \quad (2\text{-}24)$$

ขั้นตอนที่ 3: คำนวณค่าอัตราความเป็นสมาชิกของข้อมูลนำเข้าในแต่ละกลุ่มโดยการใช้มาหาลาโนบิส อาร์บีเอฟ

$$mem(p, W_p) = \exp\left(-\frac{(p - W_p)^T K^{-1} (p - W_p)}{2} \right) \quad (2\text{-}25)$$

ขั้นตอนที่ 4: ค้นหากลุ่มที่เป็นผู้ชนะโดยใช้ตัวดำเนินการฟัซซี่ "OR"

$$winner = \arg\max_i (mem_i) = \arg\max_i (mem_1 \vee mem_2 \vee \ldots \vee mem_n) \quad (2\text{-}26)$$

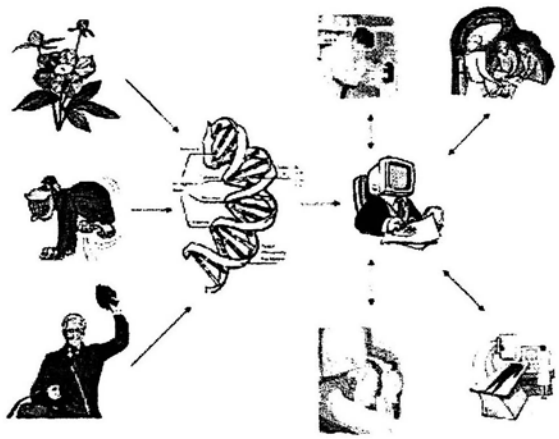
ขั้นตอนที่ 5: ทำนายคลาสเป้าหมายให้กับข้อมูลนำเข้า

คลาสจากการทำนาย = คลาสเป้าหมายของกลุ่มที่เป็นผู้ชนะ (2-27)

2.4 ชีวสารสนเทศ

คำว่าชีวสารสนเทศ (Bioinformatics) กล่าวได้ว่าเป็นการผสมผสานศาสตร์หลาย ๆ สาขา หรือ สหสาขาวิชาการ (Multi-Disciplinary) เข้าด้วยกัน ทั้งทางด้าน ชีววิทยา คอมพิวเตอร์ และสถิติ สาเหตุที่เกิดศาสตร์ทางด้านชีวสารสนเทศ เนื่องจากในอดีตที่ผ่านมานักชีววิทยา ได้ทำการศึกษาค้นคว้า และพบว่าองค์ประกอบของเซลล์ในสิ่งมีชีวิตต่าง ๆ ทั้งในคน พืช สัตว์ หรือแม้แต่แบคทีเรีย สามารถที่จะถ่ายถอดจากบรรพบุรุษไปสู่ลูกหลาน ผ่านทางรหัสพันธุกรรม หรือ ที่เรารู้จักกันดีในชื่อ ดีเอ็นเอ (DNA) ในรหัสพันธุกรรมมีองค์ประกอบย่อยของมันอีกคือ ยีน (gene) ซึ่งมีจำนวนมากมายมหาศาล เช่นในมนุษย์มียีนอยู่เป็นจำนวน 30,000 ยีน เป็นต้น (สมาคมพันธุศาสตร์แห่งประเทศไทย, 2547) ถ้าแยกย่อยลงไปจะพบว่าในแต่ละยีนมีการจับคู่ของลำดับเบส (Nitrogenous base) 4 ชนิดที่นักชีววิทยารู้จักกันดีคือ อะดีนีน (Adenine) แทนด้วย A ไซโตซีน (Cytocine) แทนด้วย C กัวนีน (Guanine) แทนด้วย G และ ไธมีน (Tyhmine) แทนด้วย T ซึ่งรหัสเบสทั้งสี่ตัวนี้จะมีการเรียงตัวสลับกันไปมาอยู่ภายในยีนแต่ละยีน (ในร่างกายมนุษย์มีอยู่ประมาณ 3,000 พันล้านคู่เบส) นั่นคือสาเหตุที่ทำให้ยีนมีความแตกต่างกัน เพื่อที่จะให้สามารถเข้าใจได้ง่าย ๆ ให้เปรียบเทียบรหัสเบสทั้งสี่ตัวเป็นพยัญชนะและวรรณยุกต์ภาษาไทย ก็จะสามารถเปรียบเทียบได้ว่า เมื่อนำรหัสเบสทั้งสี่มาเรียงกัน ก็เหมือนการนำตัวพยัญชนะและวรรณยุกต์มาเขียนร่วมกัน ซึ่งสามารถที่จะสลับ

พยัญชนะและวรรณยุกต์ไปมาได้ ซึ่งก็จะให้ภาพแบบของคำและความหมายที่แตกต่างกัน ภาพที่ 2-16 แสดงถอดรหัสพันธุกรรมจากสิ่งมีชีวิตต่าง ๆ โดยใช้ระบบคอมพิวเตอร์เข้ามาช่วยในการเก็บรวบรวมฐานข้อมูลและสร้างโปรแกรมที่เหมาะสมต่อการใช้งาน



ภาพที่ 2-16 การถอดรหัสพันธุกรรมจากสิ่งมีชีวิตต่าง ๆ โดยใช้ระบบคอมพิวเตอร์
(ภัทราวุฒิ และคณะ, 2553 ข)

ตามข้อมูลของกระทรวงสาธารณะสุขปี 2550 เปิดเผยว่ามะเร็งเต้านมถือเป็นสาเหตุการเสียชีวิตในผู้หญิงไทยเป็นอันดับสอง ปีละเกือบ 6,000 ราย ซึ่งการเฝ้าระวังที่ดีที่สุดคือการค้นพบให้เร็วที่สุด (ศูนย์สิริกิติ์บรมราชินีนาถ, 2551) โรคมะเร็งคือความผิดปรกติที่เกิดขึ้นกับเซลล์ที่มีกระบวนการซับซ้อนหลายขั้นตอน เริ่มตั้งแต่กระบวนการเจริญเติบโตจนถึงการตายของเซลล์ ดังนั้นโรคมะเร็งจึงเป็นโรคที่มีความเกี่ยวข้องกับยีนจำนวนมากมายหลายกลุ่ม (สมาคมพันธุศาสตร์แห่งประเทศไทย, 2547) ดังนั้นถ้านักชีววิทยาสามารถหาแนวทางที่จะค้นพบยีนที่มีลักษณะที่จะบ่งบอกอาการของโรคมะเร็งได้มากเท่าใด ก็จะยิ่งทำให้สามารถป้องกัน และหาวิธีทำลายยีนเหล่านี้ก่อนที่จะส่งผลออกมาเป็นเซลล์มะเร็งที่สมบูรณ์แบบ นั่นคือสาเหตุที่วิชาการด้านคอมพิวเตอร์ เข้ามาเกี่ยวข้องกับงานทางด้านชีววิทยา เนื่องจากยีนแต่ละยีนมีจำนวนเบสทั้งสี่ที่มาเรียงต่อกันสลับกันไปมาตั้งแต่จำนวนน้อย ๆ คือ 100-150 คู่เบส หรือมากกว่า 1,000 คู่เบส และในเซลล์หนึ่งเซลล์สามารถบรรจุยีนได้เป็น 1,000 ยีน ข้อมูลต่าง ๆ เหล่านี้มีจำนวนมากเกินกว่าที่สมองมนุษย์จะทำการคำนวณและวิเคราะห์ได้ ตัวอย่างเช่น โดยทั่วไปแล้วยีนหนึ่ง ๆ ที่มีลำดับเบสจะมีการแสดงออกเป็นสายโปรตีนชนิดหนึ่ง แต่ทั้งนี้เมื่อลำดับเบสภายในยีนนั้น ๆ มีการเปลี่ยนแปลง เช่น การสลับที่ของตำแหน่งในลำดับเบส ATGC เป็น TAGC อาจจะเนื่องจากการกลายพันธุ์ หรือความผิดปรกติใด ๆ ในร่างกาย สายโปรตีนที่ปรกติจะผลิตได้จากยีนนั้น ก็อาจจะผิดปรกติไปด้วยเช่นกัน การที่

นักชีววิทยาสามารถที่จะทำนายการแสดงออกของยีนล่วงหน้าได้ ก็จะทำให้ประหยัดงบประมาณและเวลาที่ใช้ในการทดลองไปได้อย่างมาก

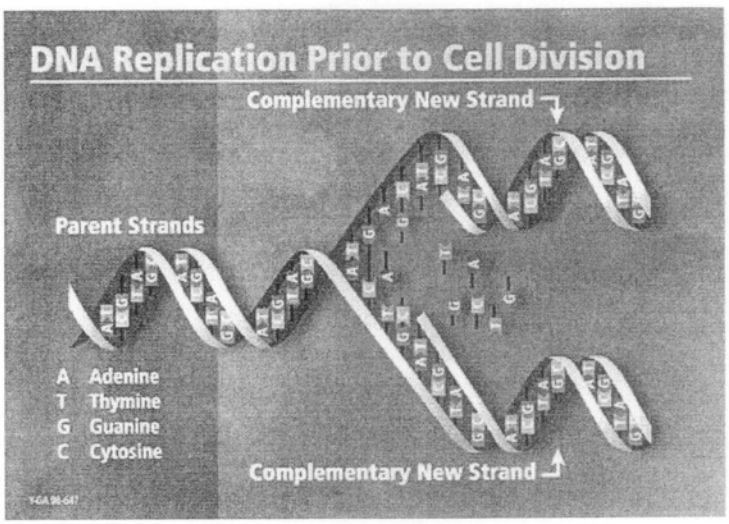
ที่มาของความร่วมมือกันระหว่างนักชีววิทยาและนักคอมพิวเตอร์ คือการที่นักชีววิทยาสามารถค้นหายีนที่มีความผิดปรกติได้ จึงต้องอาศัยประสิทธิภาพของคอมพิวเตอร์ช่วยในการเปรียบเทียบลำดับภายในยีนที่ปรกติ กับลำดับภายในยีนที่คาดว่ามีความผิดปรกติ แต่ทั้งนี้การที่จะเปรียบเทียบลำดับดังกล่าวได้นั้น จำเป็นที่จะต้องมีฐานข้อมูลสำหรับเก็บรวบรวมตำแหน่งลำดับเบสของยีนต่าง ๆ และโปรแกรมสำหรับเอาไว้เปรียบเทียบ

วิโรจน์ (2550) ได้กล่าวว่านักคอมพิวเตอร์ต้องพัฒนาอัลกอริธึมที่สำคัญ เพื่อแก้ปัญหาต่าง ๆ อาทิ 1) สำหรับการเปรียบเทียบลำดับคู่นิวคลีโอไทด์ จำเป็นต้องอาศัยขั้นตอนและกระบวนการที่ให้ผลลัพธ์ที่ถูกต้อง แม่นยำ และแก้ปัญหาที่ซับซ้อน จึงเป็นที่มาของอัลกอริธึมที่มีความสำคัญมากในการคำนวณเปรียบเทียบ และหาค่าความเหมือน หรือแตกต่าง (Similarity and Distance) ของลำดับนิวคลีโอไทด์ (Comparing Sequences) เนื่องจากการที่ลำดับของนิวคลีโอไทด์มีขนาดที่ใหญ่มาก จึงมีความจำเป็นในการที่จะต้องพัฒนาอัลกอริธึมที่สามารถเปรียบเทียบจำนวนลำดับที่ยาวมาก และอัลกอริธึมเหล่านี้จะสามารถทำการเพิ่ม ลบ และแทนที่ สัญลักษณ์ของเบสต่าง ๆ ได้ 2) การสร้างแผนภูมิชาติพันธ์ (Phylogenetic Trees) ซึ่งจะถูกสร้างหลังจากเปรียบเทียบลำดับเบสที่เป็นส่วนหนึ่งของพืชหรือสัตว์ที่แตกต่างกัน เพื่อที่ตอบคำถามความเกี่ยวข้องของลำดับเบสต่าง ๆ ที่ถูกเปลี่ยนเมื่อกาลเวลาผ่านไป 3) การหาภาพแบบของลำดับ (Detecting Patterns in Sequences) อาทิ การค้นหายีนในดีเอ็นเอ หรือในส่วนประกอบย่อยของกรดอะมิโน ซึ่งนักคอมพิวเตอร์ใช้ความรู้ทางด้านกระบวนการของเครื่องจักรการเรียนรู้ (Machine Learning) อาทิ โครงข่ายประสาทเทียม (Neural Network) หรือความน่าจะเป็น (Probability) ในการค้นหาคำตอบเหล่านี้ 4) การกำหนดโครงสร้างสามมิติของโปรตีน เนื่องจากปัญหาการคำนวณลำดับโครงสร้างสามมิติของโปรตีนที่มีความยุ่งยาก

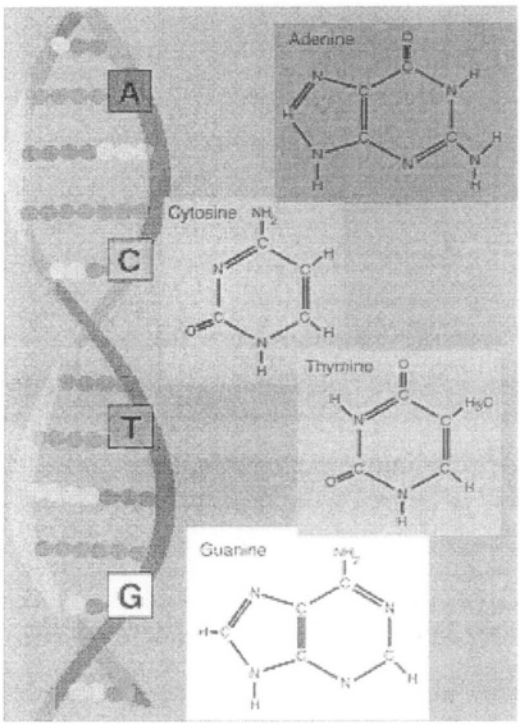
2.5 โครงสร้างของดีเอ็นเอ

ดีเอ็นเอ ย่อมาจาก ดีออกซีไรโบนิวคลีอิกแอซิด (Deoxyribonucleic Acid) ประกอบด้วยกรดนิวคลีโอไทด์ (Nucleotide) ที่มีส่วนผสมของน้ำตาลดีออกซีไรโบส (Deoxyribose Sugar) เป็นแกนกลางให้หมู่ของไนโตรเจนเบส และฟอสเฟต คือ A T G และ C ทั้ง 4 ชนิดที่ได้กล่าวแล้วมาจับคู่กัน และดีเอ็นเอจะมีโครงสร้างโมเลกุลเป็นลักษณะเกลียวคู่ หรือ Double Helix ซึ่งก็คือการบิดหมุนขนาดกันของสายดีเอ็นเอสองสาย เพื่อการสร้างสภาพการยึดเกาะที่มั่นคงของสายดีเอ็นเอ การยึดเกาะนี้จะเป็นการกระทำที่เบสหนึ่งจะยื่นออกจากสายดีเอ็นเอสายหนึ่ง และจะไปจับคู่กับเบสสายตรงข้ามที่เป็นคู่เบส (Base Pair) ทั้งนี้เบสที่เป็นคู่ตรงข้ามกันได้แก่เบส A กับ T และเบส G กับ

เบส C การจับคู่ของเบสดังกล่าวเกิดขึ้นได้เนื่องจากพันธะของไฮโดรเจน ดังแสดงในภาพที่ 2-17 ซึ่งแสดงถึงสายของดีเอ็นเอ (บุญญานาถ, 2552) และในภาพที่ 2-18 แสดงโครงสร้างและองค์ประกอบของดีเอ็นเอ



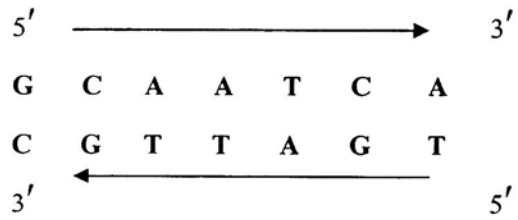
ภาพที่ 2-17 สายดีเอ็นเอ (Starmer, 2011)



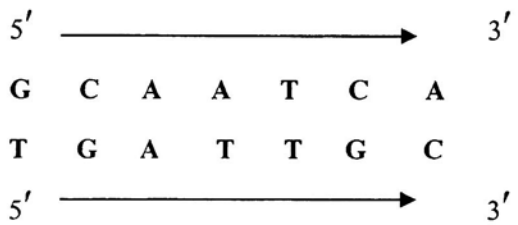
ภาพที่ 2-18 โครงสร้างและองค์ประกอบของดีเอ็นเอ (Starmer, 2011)

การจับคู่ของเบสที่เป็นคู่ตรงกันข้ามกันทำให้เกิดสายดีเอ็นเอสองสายจับคู่กัน ซึ่งก็จะเป็นไปตามกฎที่ว่าสายดีเอ็นเอทั้งสองจะต้องมีทิศทางที่สวนทางกัน อย่างไรก็ตามการเรียกสายดีเอ็นเอก็จะมีการกำหนดชื่อในการเรียก เช่นส่วนเริ่มต้นของสายดีเอ็นเอจะถูกเรียกว่า 5′ หรือ ห้าไพรม์ หรือส่วนต้น (Up Stream) ในขณะเดียวกันส่วนด้านปลายของดีเอ็นเอ ก็จะถูกเรียกว่า 3′ หรือสามไพรม์ หรือส่วนปลาย (Down Stream) (สุรินทร์, 2548 ก) การเรียงตัวของอะตอมคาร์บอนในลักษณะนี้จะเป็นไปตามโครงสร้างที่เป็นส่วนของโมเลกุลน้ำตาลออกซีไรโบสซึ่งก็คือตำแหน่งที่ 5′ และ 3′ ดังแสดงในภาพที่ 2-19 อย่างไรก็ตามในทางชีววิทยาแล้วการอ่านลำดับเบสของสายดีเอ็นเอ จะต้องอ่านไปตามทิศทาง 5′ ไปยัง 3′ เสมอ ดังที่แสดงในภาพที่ 2-20

สารพันธุกรรมทั้งหมดของสิ่งมีชีวิต หรือที่เรียกว่า จีโนม (Genome) เป็นสารที่ถูกเก็บไว้ในเซลล์ สารพันธุกรรมจะอยู่ในนิวเคลียส (Nucleus) และไมโทชอนเดีย (Mitochondria) ซึ่งทั้งสองตำแหน่งนี้อยู่ในเซลล์ด้วยกัน โดยธรรมชาติแล้วจีโนมส่วนมากจะอยู่ในนิวเคลียส ประมาณการว่ากว่า 99.99% ที่จีโนมจะอยู่ในนิวเคลียส หรือที่เรียกว่านิวเคลียสดีเอ็นเอ (Nucleus DNA) มีแหล่งกำเนิดจาก 2 แหล่งคือ DNA ที่ได้จากไข่ (แม่) และ DNA ที่ได้จากสเปิร์ม (พ่อ) นิวคลียสดีเอ็นเอจึงเป็นตัวกำหนดลักษณะเฉพาะบุคคลและถูกนำมาใช้ในการศึกษาการถ่ายทอดโรคทางพันธุกรรม (สมาคมพันธุศาสตร์แห่งประเทศไทย, 2547)



ภาพที่ 2-19 ตำแหน่งที่ 5′ และ 3′



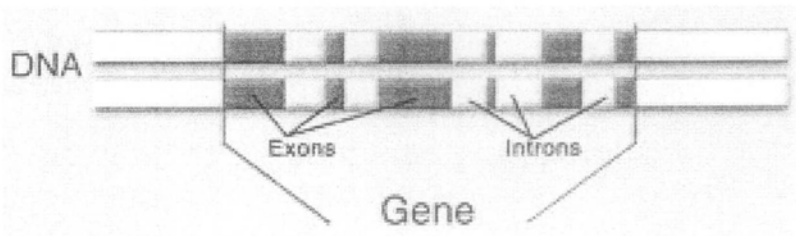
ภาพที่ 2-20 การอ่าน Sequence base pair

2.6 โครงสร้างและการแสดงออกของยีน

ยีนเป็นรหัสที่ใช้สำหรับการสังเคราะห์โปรตีน ซึ่งก็คือการสังเคราะห์โมเลกุลดีเอ็นเอ โดยที่ขนาดของยีนจะประกอบด้วยคู่เบส A T G และ C ขนาดของยีนจะแตกต่างกัน ยีนบางตัวอาจจะมีขนาดเล็กมากไม่ถึง 100 คู่เบส ในทางตรงกันข้ามยีนบางตัวอาจมีจำนวนคู่เบสที่มากมายถึงล้านคู่เบสได้เช่นกัน

ในโครโมโซมจะประกอบด้วยลำดับเบสที่ประกอบเป็นยีนถูกบรรจุอยู่ แต่ไม่ใช่ทั้งหมด ขอบเขตที่รวมกันแล้วจะเป็นยีนในโครโมโซม จริง ๆ มีเพียงประมาณร้อยละ 10 หรือประมาณ 300 ล้านคู่เบสในจีโนมเท่านั้นที่เป็นยีน ส่วนที่เหลือคือประมาณร้อยละ 90 ในทางชีววิทยายังไม่สามารถหาคำตอบได้แน่ชัดว่าทำหน้าที่อะไร แต่ที่สำคัญเบสเหล่านี้ไม่ได้เป็นองค์ประกอบของยีน (สมาคมพันธุศาสตร์แห่งประเทศไทย, 2547)

การถอดรหัสจะไม่ใช้ลำดับเบสที่อยู่ในยีนทั้งหมด จะมีเพียงลำดับเบสบางส่วนของยีนที่จะถูกถอดรหัสออกมาเพื่อสร้างเป็นโปรตีน ทั้งนี้ในยีนได้แบ่งลำดับเบสบนโมเลกุลดีเอ็นเอออกเป็นส่วนประกอบที่สำคัญ 2 ส่วนคือ เอ็กซอน (exon) ซึ่งหมายถึงลำดับเบสที่สามารถถูกถอดรหัสออกเป็นเอ็มอาร์เอ็นเอ (Message Ribonucleic Acid: mRNA) และส่วนลำดับเบสที่ไม่สามารถเข้ารหัสเพื่อใช้ในการสร้างโปรตีนได้ (Non-Coding Sequence) เรียกว่า อินทรอน (Introns) ซึ่งลำดับเบสชนิดนี้จะวางอยู่ระหว่างส่วนของลำดับเบสเอ็กซอน (สุรินทร์, 2548) ดังแสดงในภาพที่ 2-21

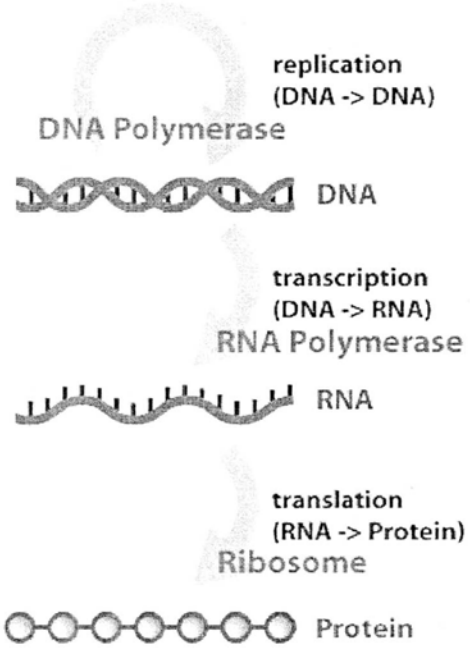


ภาพที่ 2-21 การถอดรหัส mRNA (Starmer, 2011)

กระบวนการของเซลล์ทุกเซลล์จะถูกกำหนดไว้อย่างแน่นอน และมีการควบคุมอย่างรัดกุมว่าจะสามารถแสดงออกหรือทำหน้าที่ของยีนที่มีอยู่ไว้แล้วอย่างไร นอกจากนี้ในภาวะปรกติในร่างกายมนุษย์จะประกอบด้วยเซลล์ และแต่ละเซลล์จะมียีนที่อาศัยอยู่ในจีโนมในลักษณะเหมือนกัน อย่างไรก็ตามยีนที่มีการแสดงออกที่เหมาะสมจะมีโอกาสเกิดขึ้นเพียงร้อยละ 20 ของยีนทั้งหมด ทั้งนี้ต้องเป็นไปตามหน้าที่ของแต่ละเซลล์ ซึ่งจะมีหน้าที่ในการแสดงออกเป็นโปรตีนที่แตกต่างกันขึ้นอยู่กับการทำหน้าที่ของเซลล์นั้น เช่น ถ้าเป็นเซลล์สมอง เซลล์ผิวหนัง หรือเซลล์กล้ามเนื้อ ก็มา

ส่วนประกอบจากยีนเหมือนกัน แต่ในกรณีที่เป็นเซลล์ผิวหนัง หรือเซลล์กล้ามเนื้อ เซลล์เหล่านี้จะไม่มีโอกาสสร้างโปรตีนเอ็กซ์ (X) ที่ทำหน้าที่อยู่ในเซลล์สมอง ส่วนโปรตีนวาย (Y) และโปรตีนแซด (Z) ก็ไม่มีโอกาสเกิดขึ้นในเซลล์สมองเพราะมีหน้าที่การทำงานในเซลล์กล้ามเนื้อ และทำหน้าที่ในเซลล์ผิวหนังตามลำดับ (สมาคมพันธุศาสตร์แห่งประเทศไทย, 2547)

การแสดงออกของยีน (Gene Expression) แบ่งได้เป็น 2 ขั้นตอนในระดับโมเลกุล ดังนี้ 1) การถอดรหัส (Transcription) จากดีเอ็นเอ สำหรับใช้ในการสร้าง mRNA และ 2) การแปลรหัส (Translation) เพื่อใช้สำหรับสร้างโปรตีน ซึ่งเป็นที่รู้จักกันดีในชื่อทางกระบวนการชีววิทยาว่าเซ็นทรัลด็อกม่า (Central Dogma) ดังแสดงในภาพที่ 2-22

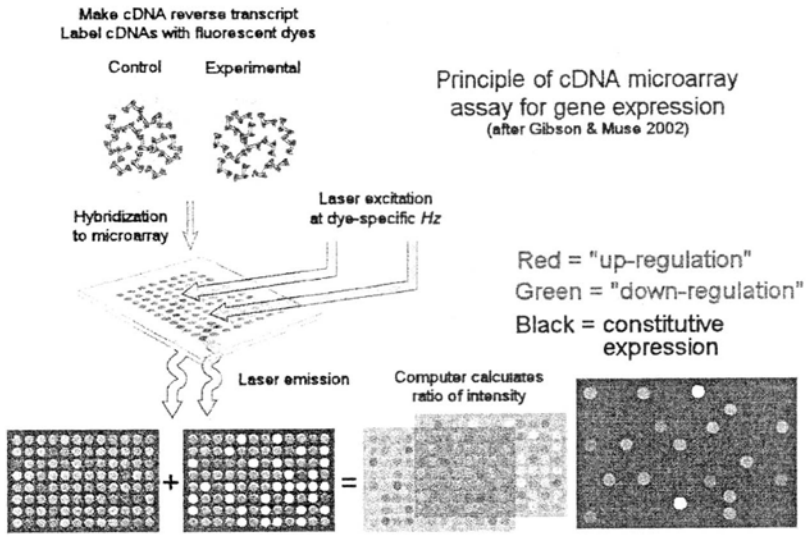


ภาพที่ 2-22 Central dogma จากการแสดงออกของยีน (Horspool, 2011)

กระบวนการของการสร้างโปรตีนนี้ จะเป็นการดำเนินการในไซโตรพลาสมของเซลล์ที่ไรโบโซม (Ribosome) ซึ่งจะอาศัยขั้นตอนการแปลรหัสจาก mRNA มาเป็นโปรตีน ทั้งนี้เริ่มจากการส่งผ่านอาร์เอ็นเอ (Transfer RNA) หรือทีอาร์เอ็นเอ (tRNA) ที่จะถูกตัวไรโบโซมอล-อาร์เอ็นเอ หรือ อาร์-อาร์เอ็นเอ (R-RNA) นำมาใช้สำหรับอ่านจุดเริ่มต้นของเบส mRNA ซึ่งจุดเริ่มต้นนี้มักเรียกกันว่าโคดอนเริ่มต้น (Initiation Codon: AUG) จากนั้นกระบวนการจะเริ่มอ่านโดยจะใช้วิธีการอ่านครั้งละ 3 เบส ทั้งนี้ทุก 3 เบสที่อ่านจะมีรหัสตรงกับ tRNA 1 ชนิดและ กรดอะมิโนจำเพาะ

(Amino Acid) จะถูก tRNA นำมาเพื่อดำเนินการต่อเป็นสายโปรตีน (สมาคมพันธุศาสตร์แห่งประเทศไทย, 2547)

2.7 ดีเอ็นเอไมโครอะเรย์



ภาพที่ 2-23 การทำดีเอ็นเอไมโครอะเรย์ (Ajmillar, 2011)

ดีเอ็นเอไมโครอะเรย์เป็นเครื่องมือหนึ่งที่ช่วยให้นักชีววิทยาสามารถทำการทดลองกับยีนจำนวนมากได้ในครั้งเดียว จากการวิธีการทดลองเกี่ยวกับการแสดงออกของยีนแบบดั้งเดิมที่ใช้เพียงคู่นิวคลีอิก (Pair Nucleic) ซึ่งทำให้นักชีววิทยาต้องใช้เวลาในการทดลองเพิ่มมากขึ้น ดังนั้นเทคนิคไมโครอะเรย์ที่ใช้อยู่ในปัจจุบันจึงถูกคิดค้นขึ้น เทคนิคนี้จะช่วยวัดการแสดงออกของยีนเป็นจำนวนหลายพันยีนได้ในการทดลองเพียงครั้งเดียว การทำไฮบริไดเซชันของกรดนิวคลีอิกจะถูกใช้ในเทคนิคไมโครอะเรย์ ซึ่งวิธีการนี้จะแตกต่างจากวิธีการทดลองแบบเดิมที่การทำไฮบริไดเซชันของกรดนิวคลีอิกเพียงใช้คู่ของสายเดี่ยวจาก 2 แหล่งมาทดลองร่วมกัน เทคนิคไมโครอะเรย์จะมีขั้นตอนการทำงานดังนี้คือ 1) นำโพรบ (Probe) จำนวนมากมาติดไว้ตามจุด (Array) หรือชิป (Chip) ซึ่งมีขนาดค่อนข้างเล็ก 2) บนวัสดุเหล่านี้จะประกอบไปด้วยสปอตขนาดเล็กที่มีจำนวนหลายพันสปอต ซึ่งมีหน้าที่ในการวิเคราะห์เพื่อวัดระดับในการแสดงออกของยีนแต่ละยีนบนสปอตลำดับนิวคลีโอไทด์จะถูกสังเคราะห์บนโพรบที่ติดอยู่บนไมโครอะเรย์ 3) นักชีววิทยาจะนำเอาดีเอ็นเอจากเนื้อเยื่อ (Genomic DNA) หรือจาก cDNA ที่ต้องการทดสอบมาติดฉลาก และทำการไฮบริไดเซชั่นลงไปบนชิป 4) ทำการตรวจสอบการแสดงออกของยีนที่ระยะใดระยะหนึ่ง โดยดูจากสัญญาณการแสดงออก (สุรินทร์, 2548 ข) ดังแสดงในภาพที่ 2-23 จุดสีแดงแสดงว่ายีนนั้นมีการ

แสดงออกในเนื้อเยื่อมาก (Up Regulation) จุดสีเขียวแสดงว่ายีนนั้นจะมีการแสดงออกค่อนข้างต่ำกว่าปรกติกับเนื้อเยื่อชนิดนั้น (Down Regulation) ขณะเดียวกันถ้ามีจุดสีดำแสดงว่ามีการแสดงออกอยู่ตลอดเวลา (Constitutive Expression)

2.8 การเปลี่ยนรูปข้อมูลการแสดงออกของยีน

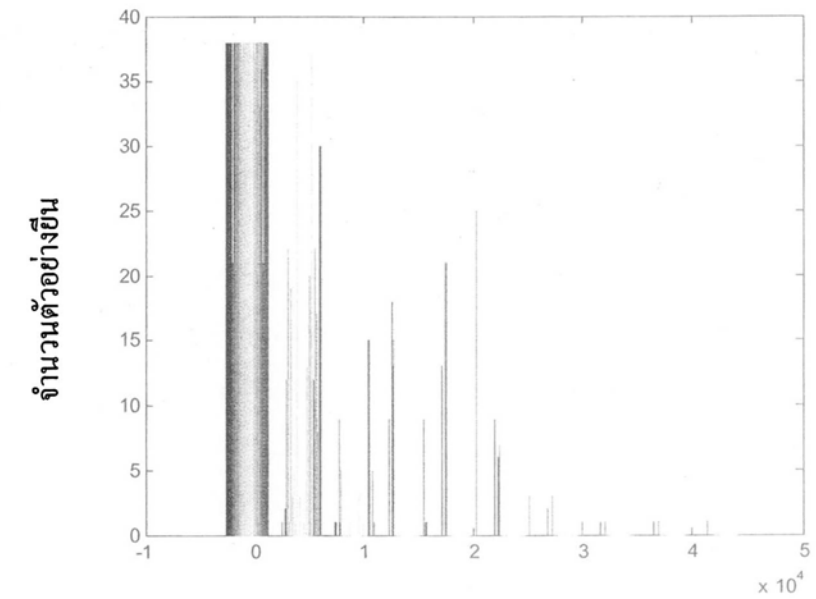
เนื่องจากก่อนที่ข้อมูลการแสดงออกของยีนจะมาถึงมือนักวิเคราะห์จำเป็นต้องผ่านกระบวนการทางชีวภาพ และสุดท้ายจะออกมาในรูปแบบของภาพ (Images) ซึ่งภาพที่ได้จากกระบวนการชีวภาพนั้นเป็นภาพที่ได้มาจากการตรวจรายละเอียดบนไมโครอะเรย์ชิปถึงค่าความเข้มข้นในแต่ละสปอตซึ่ง ยังไม่มีความน่าเชื่อถือเพียงพอที่จะนำมาใช้งานต่อ เนื่องจากข้อมูลที่ได้รับจากการตรวจรายละเอียดบนไมโครอะเรย์ชิปยังอาจมีสิ่งรบกวน (Noise) และการใช้ข้อมูลในลักษณะข้อมูลต้นฉบับแบบนี้นอกจากจะมีความลำเอียง (bias) แล้วยังก่อให้เกิดความไม่น่าเชื่อถือและความไม่ถูกต้องในการใช้งานอีกด้วย

Li (2008) กล่าวว่า สิ่งรบกวนจากไมโครอะเรย์คือสิ่งที่อยู่รอบสปอตบนไมโครอะเรย์และเมื่อถูกตรวจรายละเอียดโดยการทำเป็นภาพ ภาพพวกนี้จะมีพื้นหลังแสดงเป็นสีดำ นั่นคือการชี้ว่าเป็นข้อมูลที่ไม่เกี่ยวข้อง แต่ในการปฏิบัติแล้ว พื้นหลังของภาพไมโครอะเรย์มีความเกี่ยวข้องที่รู้จักกันดีว่าเป็นสิ่งรบกวนพื้นหลัง (Background Noise) การมีข้อมูลตัวนี้จะเป็นสิ่งบ่งบอกว่าข้อมูลที่ได้มานั้นมีความผิดพลาดโดยกระบวนการในห้องทดลอง และจากการตรวจรายละเอียดของภาพไมโครอะเรย์

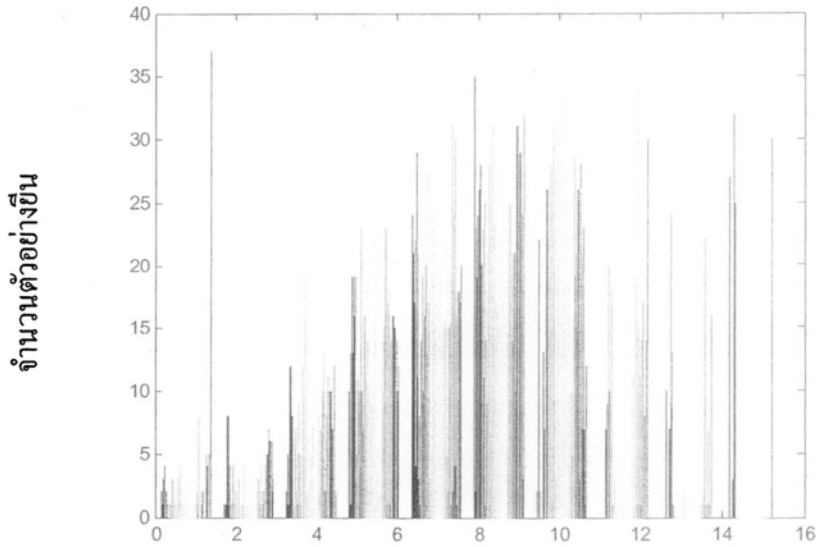
สิ่งรบกวนเหล่านี้นอกจากทำให้ประสิทธิภาพการประมาณการลดลงแล้ว ยังสามารถตัดทอนการแปลความหมายของความเข้มข้นของพื้นหลังจากสปอตที่หนาแน่นอีกด้วย วิธีการที่ง่ายคือการใช้ค่าความหนาแน่นพื้นหลังโดยเฉลี่ยของชิปทั้งหมดเป็นวิธีการที่ไม่เหมาะสม เมื่อความเข้มข้นของสิ่งรบกวนไม่สม่ำเสมอในทุกส่วนของชิป ดังนั้นวิธีการประมาณการเฉพาะที่ (Local Estimation) จึงเป็นวิธีที่ถูกใช้สำหรับประมาณการความหนาแน่นพื้นหลังของแต่ละสปอตจากพื้นที่รอบสปอต

นอกจากการจัดการสิ่งรบกวนที่ทำให้สมรรถนะการประมวลผลทางคอมพิวเตอร์ดีขึ้นแล้ว ทางหนึ่งที่นิยมทำกันเพื่อเพิ่มคุณภาพของข้อมูลการแสดงออกของยีนคือการเปลี่ยนรูปข้อมูล (Transformation data) การกระทำลักษณะนี้ก็เพื่อที่จะเจาะจงข้อมูลกระโดด (Outlier) ในกลุ่มข้อมูลการแสดงออกของยีน นอกจากนั้นยังช่วยในการปรับปรุงการกระจายตัวของข้อมูลอีกด้วย วิธีการเปลี่ยนแปลงรูปแบบข้อมูลมักจะอาศัยวิธีการทางสถิติในการดำเนินการ เช่นการใช้วิธีการเปลี่ยนรูปข้อมูลคือ การเปลี่ยนรูปแบบลอการิทึม (Logarithmic Transformation) ดังแสดงในภาพที่ 2-24 และ

2-25 โดยจะทำการเปลี่ยนรูปการกระจายตัวของข้อมูลให้เป็นการกระจายตัวแบบระฆังคว่ำ นั้นคือ เป็นการกระจายตัวแบบปรกติทั่วไป



ภาพที่ 2-24 การกระจายตัวของข้อมูลการแสดงออกของยีน Leukemia ก่อนทำการเปลี่ยนแปลงรูปแบบ



ภาพที่ 2-25 การกระจายตัวของข้อมูลการแสดงออกของยีน Leukemia หลังทำการเปลี่ยนแปลงรูปแบบ

การเปลี่ยนรูปข้อมูลในลักษณะนี้จะมีผลดีอย่างมากกับกลุ่มข้อมูลการแสดงออกของยีนที่มีตัวอย่างสองชนิด (เช่นเป็นโรค และไม่เป็นโรค) เนื่องจากข้อมูลจากอะเรย์มักมีลักษณะของการแสดงเป็นอัตราส่วนของสัญญาณความหนาแน่นของคู่ตัวอย่าง สำหรับวิธีการอื่นที่ใช้ในการเปลี่ยนแปลงรูปแบบได้แก่ ฟังกชันอาร์ซิน (Arcsinh Function) การเปลี่ยนรูปแบบลินล็อก (Linlog Transformation) การเปลี่ยนรูปแบบเคิฟฟิตติ้ง (Curve-Fitting Transformations) และการเปลี่ยนรูปแบบชิป (Shift Transformation)

การเปลี่ยนรูปข้อมูลเกิดขึ้นจากอัตราส่วนไม่สมมาตรกัน ปัจจุบันนิยมใช้วิธีลอการิทึมกับกลุ่มข้อมูลการแสดงออกของยีน อาทิ Jinlong and Zhigang (2010) ได้ศึกษาเรื่องการทดแทนข้อมูลที่สูญหาย และได้ทำการเตรียมข้อมูลโดยใช้เปลี่ยนรูปข้อมูลแบบลอการิทึมทรานส์ฟอร์เมชั่นช่วยในการเปรียบเทียบการเปลี่ยนแปลงที่สำคัญ นอกจากนี้ยังอ้างถึงวิธีการมีนนอร์มอลไลซ์เซชั่น (Mean-Normalization) ว่าช่วยให้ประสิทธิภาพการทำงานของการวัดระยะห่างของข้อมูลแบบยูคลีเดียนดียิ่งขึ้น Schaefer and Nakashima (2010) ได้อธิบายถึงการเปลี่ยนรูปแบบข้อมูลการแสดงออกของยีนที่นำมาใช้ในการทดลองโดยใช้การเปลี่ยนรูปแบบลอการิทึมฐานสิบ ($\log_{10}$) จากจำนวนยีนทั้งหมด 6500 ยีน และมี 2 กลุ่มตัวอย่างคือ กลุ่มตัวอย่างที่เป็นเนื้องอก 40 ตัวอย่างและปรกติอีก 20 ตัวอย่าง

2.9 การทดแทนข้อมูลที่สูญหายของข้อมูลไมโครอะเรย์

ไก้รุ่ง (2553) ได้กล่าวถึงค่าข้อมูลที่สูญหาย (Missing Data) ว่าเป็นข้อมูลที่ไม่มีการระบุค่า (ค่าว่าง) ไว้ซึ่งอาจเกิดจากความผิดพลาดจากกระบวนการเก็บหรือการละเลยที่จะกรอกข้อมูล ปัญหาข้อมูลสูญหายของข้อมูลไมโครอะเรย์เป็นปัญหาที่สำคัญอันหนึ่งและเกิดขึ้นได้ตลอดเวลา เนื่องจากในขั้นตอนการทดลองอาจมีสิ่งปนเปื้อนหรือมีวัตถุเข้าไปทำปฏิกิริยาในการทดลองได้ง่ายมาก เช่น ความละเอียดของภาพไม่เพียงพอ ได้ภาพที่ไม่ชัดเจน มีฝุ่นหรือรอยบนสไลด์หรือการทดลองเนื่องจากกระบวนการทำงานในห้องทดลอง (Kim *et al.*, 2005) ดังนั้นการทดแทนข้อมูลที่สูญหายของข้อมูลไมโครอะเรย์จึงนับเป็นขั้นตอนหนึ่งที่สำคัญที่จะละเลยไม่ได้ เพราะอาจจะส่งผลให้ประสิทธิภาพของการทำงานคัดเลือกยีนที่มีอำนาจจำแนกหรือผลลัพธ์ในการจัดกลุ่มมีผลลัพธ์ที่ต่ำลงได้ ตัวอย่างของข้อมูลสูญหายของไมโครอะเรย์จากกลุ่มข้อมูลโรคมะเร็งต่อมน้ำเหลือง (Lymphoma) ดังแสดงในภาพที่ 2-26

มนตรี (2549) ได้กล่าวถึงความหมายของการทดแทนข้อมูลสูญหาย หรือ Replacement ว่าเป็นการดำเนินการเพื่อสร้างข้อมูลทดแทนให้แก่รายการข้อมูลหรือคำถาม ที่ไม่มีข้อมูลอันเนื่องมาจากหลายสาเหตุ อาทิ ความผิดพลาดจากการบันทึกข้อมูลลงบนสื่อคอมพิวเตอร์

(Data Entry) แบบสอบถามที่ผู้ตอบไม่ตอบคำถาม หรือผู้เก็บข้อมูลด้วยการสัมภาษณ์ลืมสัมภาษณ์ หรือเก็บข้อมูลการสัมภาษณ์มาบันทึก เป็นต้น

ประลองพล (2551) ได้กล่าวถึงข้อมูลตัวแปรอิสระสูญหาย จะเกิดผลทำให้หน่วยตัวอย่างนั้นไม่สามารถนำไปจัดกลุ่มได้ ซึ่งก่อให้เกิดปัญหาการวิเคราะห์ทางด้านธุรกิจและการแพทย์เป็นอย่างมาก

Wang *et al.* (2007) ได้กล่าวถึงการจัดการสิ่งรบกวน ซึ่งรวมถึงการแทนที่ค่าสูญหายสำหรับข้อมูลทั่วไปไว้ดังนี้

1) การเพิกเฉย (Neglect) เป็นสาเหตุของการสูญเสียข้อมูล นั่นคือการลบข้อมูลที่เป็นข้อมูลรบกวนทั้งหมด แต่เมื่อใช้กับการจำแนกหมวดหมู่มักจะส่งผลให้มีความแม่นยำไม่เพียงพอ

2) ทำให้ประชากรสมบูรณ์ตามสถานการณ์จริง (Artificial Addition)

3) เติมค่าคงที่ (Constant) ให้แอททริบิวท์ข้อมูลที่สูญหาย กับสัญลักษณ์ที่บ่งบอกว่าข้อมูลนี้เสียหาย (Fixed Symbol)

4) แทนค่าข้อมูลที่สูญหายกับค่าเฉลี่ย (Mean) เช่น แอททริบิวท์ข้อมูลมีค่า 35, 38, 25 และ 17 ดังนั้นค่าเฉลี่ยคือ ค่าเฉลี่ย = (35+38+25+17)/4=28.75 จากนั้นเมื่อมีตำแหน่งค่าที่สูญหายตำแหน่งใด ก็แทนโดยค่าเฉลี่ย 28.75

5) เป็นวิธีการปรับปรุงของวิธีค่าเฉลี่ย โดยการหาค่าเฉลี่ยที่แตกต่างกันตามประเภทของคลาสของแอททริบิวท์ข้อมูลที่หายไป (Means from the same label) ดังเช่น ตัวแปร v2 v6 และ v8 เป็นข้อมูลที่อยู่ในคลาสเดียวกัน และตัวแปร v3, v4, v7, v9 และ v10 เป็นข้อมูลที่อยู่ในอีกคลาสหนึ่ง ดังนั้นเมื่อจะแทนที่ข้อมูลสูญหายด้วยวิธีนี้ จึงมีค่าเฉลี่ยของแต่ละคลาสที่ใช้แทนที่แตกต่างกันดังนี้

$$\text{Mean1} = (v2 + v6 + v8)/3 = (35 + 17 + 26)/3 = 26$$

$$\text{Mean5} = (v3 + v4 + v7 + v9 + v10)/5$$

$$= (38 + 25 + 16 + 19 + 23)/5 = 31.5$$

6) การฝึกฝน (Training) เป็นวิธีการกำหนดข้อมูลสูญหายที่มีเหตุผล อาทิ การนำเอาแอททริบิวท์ของข้อมูลที่ใช้ตัดสินใจ มาแทนที่แอททริบิวท์ข้อมูล จากนั้นนำเอาค่าสูญหายจากกรณีที่เป็นกรณีเป้าหมายและนำมาจับคู่กัน

Tan *et al.*, (2005) ได้กล่าวถึงวิธีการจัดการข้อมูลที่สูญหายอีกหลายวิธี ได้แก่

1) การลบข้อมูลที่สูญหายออก

2) การลบค่าที่สูญหายออก

3) ละเว้นช่วงข้อมูลที่มีค่าสูญหาย แล้วดำเนินการวิเคราะห์ต่อไป

4) แทนที่ข้อมูลที่สูญหายด้วยค่าที่เป็นไปได้ อาทิ การให้น้ำหนักโดยความน่าจะเป็น ค่าต่ำสุด (Minimum) ค่ามากที่สุด (Maximum) และค่าเฉลี่ย (Mean)

ไก้รุ่ง (2553) ได้กล่าวถึงข้อมูลที่สูญหายของไมโครอะเรย์ว่าเป็นข้อมูลที่นิยมใช้ในการศึกษารูปแบบของสิ่งมีชีวิตในระดับโมเลกุลที่แสดงถึงระดับการแสดงออกของยีน (Gene) หลายพันยีนในเวลาเดียวกัน แต่โดยปรกติแล้วข้อมูลไมโครอะเรย์มักประกอบด้วยค่าที่สูญหายเป็นจำนวนมาก และโดยปรกติแล้วไมโครอะเรย์จะมีค่าขาดหายประมาณ 1-10% ดังนั้นจึงส่งผลให้ประสิทธิภาพของการวิเคราะห์ข้อมูลลดต่ำลง นอกจากนี้ยังได้นำเสนอวิธีการทดแทนค่าสูญหาย 3 วิธีคือ เคเอ็นเอ็นเอฟเอส (KNNFS) เคเอ็นเอ็นเอฟเอสดับบลิว (KNNFSW) และเคเอ็นเอ็นอาร์ (KNNR) ซึ่งแต่ละวิธีจะอยู่บนพื้นฐานของอัลกอริธึมสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor: KNN) ผลลัพธ์ที่ได้พบว่าวิธีการที่นำเสนอให้ค่าผลลัพธ์ที่วัดจากค่าความผิดพลาดเฉลี่ยของค่าประมาณการกับค่าจริง (Normalized Root Mean Square: NRMSE) ต่ำกว่าวิธีการแบบดั้งเดิมอย่างมีนัยยะสำคัญทางสถิติที่ระดับ 0.5 อย่างไรก็ตามวิธีดังกล่าวอาศัยการทดลองสุ่มสร้างข้อมูลที่ขาดหาย 1% ถึง 10% ซึ่งไม่ใช่ข้อมูลขาดหายที่แท้จริงจากกระบวนการทางชีวภาพ

Ma *et al.* (2006) ได้นำเสนออัลกอริธึมอีโวคลัสเตอร์ (EvoCluster) ซึ่งสามารถทำงานกับค่ารบกวนและข้อมูลที่สูญหายได้อย่างมีประสิทธิภาพ บนกลุ่มข้อมูลการแสดงออกของยีน บนไมโครอะเรย์ดีเอ็นเอ วิธีการที่นำเสนอนี้อยู่บนพื้นฐานของอัลกอริธึมเชิงวิวัฒนาการ (Genetic Algorithm) แม้ว่าวิธีการนี้มีประสิทธิภาพที่สูงแต่ก็ยังต้องอาศัยการกำหนดค่าพารามิเตอร์ที่เหมาะสมซึ่งจะส่งผลต่อประสิทธิภาพการทำงานของอัลกอริธึม

Jinlong and Zhigang (2010) พัฒนาอัลกอริธึมการทดแทนข้อมูลที่สูญหายการใช้ทฤษฎีพรินซิเพิลเคิฟ (Principle Curve) ซึ่งมีลักษณะไม่เป็นเชิงเส้น (Nonlinear) และทฤษฎีดังกล่าวนับเป็นรุ่นบุกเบิกของการวิเคราะห์องค์ประกอบแบบเชิงเส้น (Linear Principal Component Analysis) วิธีการที่นำเสนอจะเป็นการพิจารณาเอาความสัมพันธ์ทั้งแบบเชิงเส้น และไม่เป็นเชิงเส้นของข้อมูลระหว่างยีนเข้าด้วยกัน พร้อมกับเปิดเผยถึงการกระจายตัวของยีน แม้ว่าเทคนิคดังกล่าวให้ผลลัพธ์ที่ดีกว่าในการเปรียบเทียบกับอัลกอริธึมการทดแทนข้อมูลที่สูญหายคือ KNN, BPCA และ LLSimpute อย่างไรก็ตามวิธีการนี้ใช้การวัดระยะห่างแบบเทคนิคยูคลีเดียน ซึ่งเป็นเทคนิคที่ใช้วัดระยะห่างที่มีข้อเสียในเรื่องรูปทรงของการวัดระยะห่างที่เป็นแค่วงกลมเพียงอย่างเดียว ส่งผลให้ประสิทธิภาพการทำงานของอัลกอริธึมอาจไม่เพียงพอ

อย่างไรก็ตาม Hyunsoo *et al.* (2005) ได้ทดลองการประมาณการค่าสูญหายสำหรับข้อมูลการแสดงออกของยีนดีเอ็นเอไมโครอะเรย์ด้วยการนำเสนอวิธี local least squares imputation ได้แก่ LLSk/L2 และ LLSk/PC แล้วทำการเปรียบเทียบกับเทคนิค KNNimpute และ BPCA แม้ว่าผลลัพธ์

ของวิธีการที่นำเสนอจะดีกว่า แต่บทความนี้ได้บอกถึงข้อดีของ KNNimpute ว่าเมื่อขนาดตัวอย่างมีจำนวนน้อย KNNimpute จะมีประสิทธิภาพดีกว่า ซึ่งข้อดีที่ว่านี้คือความเหมือนเฉพาะแห่ง (Local similarity) สำหรับการทดแทนข้อมูลที่สูญหาย

GENE1835X	GENE1836X	GENE1865X	GENE1380X
0.46	0.7	0.67	-0.23
0.02	0.59	0.45	0.55
-0.32	-0.63	-0.46	-0.28
-0.51	-0.45	-0.16	-0.51
0.2	0.13	0.2	0.09
-0.36	-0.53	-0.36	0.48
-0.31	-0.4	-0.46	-0.45
-0.13	-0.31	-0.12	?
-0.53	0.09	-0.12	-0.08
0.19	-0.12	0.37	-0.08
0	0.05	0.23	-0.07
-0.07	-1.49	-0.65	?
0.14	0.17	-0.06	0.05
-0.18	0.02	0.12	-0.1
0.42	-0.3	-0.09	0

ภาพที่ 2-26 ข้อมูลที่สูญหายจากกลุ่มข้อมูลการแสดงออกของยีนเกี่ยวกับโรค Lymphoma (Alizadeh *et al.*, 2000)

งานวิจัยที่เกี่ยวข้อง

Jae and Seok (2001) ได้นำเสนอวิธีการจัดกลุ่มแบบเซลล์เบส (Cell-Based) แบบใหม่ซึ่งเหมาะกับการทำงานกับข้อมูลที่มีมิติที่สูง ซึ่งข้อมูลเหล่านี้ไม่เหมาะกับวิธีการจัดกลุ่มโดยวิธีการ cell-based แบบที่มีอยู่ซึ่งมีสมมติฐานว่ากลุ่มข้อมูลทั้งหมดต้องถูกเก็บไว้ในหน่วยความจำถาวร (Resident Main Meory) ทำให้ประสิทธิภาพการทำงานต่ำลง แต่ในวิธีการใหม่นี้จะมีอัลกอริธึมจะสร้างเซลล์โดยการแยกแต่ละมิติเข้าสู่กลุ่มของพาร์ติชัน อัลกอริธึมในการเพิ่มเซลล์จะสร้างกลุ่มของ

เซลล์ที่มีความหนาแน่นมากที่ถูกกำหนดโดยอัตราส่วน และเก็บกลุ่มที่สร้างนี้ไว้ใน โครงสร้างดัชนี (Index Structure) สำหรับในด้านของความเร็ววิธีการนี้ก็ประยุกต์ใช้เทคนิคจาก การประมาณการ (Approximation) เข้ามาช่วย ข้อดีของวิธีการนี้คือทำงานได้เร็วกว่าวิธีการแบบเดิม สำหรับข้อเสียก็คือการกำหนดอัตราส่วนความหนาแน่นที่หาได้จากการทดสอบหลายครั้ง

CYRUS system ถูกนำเสนอโดย Kolodner ในปี 1984 (Jain *et al.*, 1999) เพื่อใช้งานกับกรอบของการจัดการความรู้ (Domain Knowledge) ให้สามารถกำหนดหน่วยย่อยของข้อมูลได้ดีที่สุด เช่น แอททริบิวต์ที่มีอำนาจจำแนกระหว่างแนวคิด อย่างไรก็ตามวิธีนี้หลีกเลี่ยงการสำรวจแบบละเอียดเพื่อเป็นข้อมูลในการเลือกแอททริบิวต์ และมีพื้นฐานจากวิธีการสร้างแนวความคิด

UNIMEM เป็นอัลกอริธึมที่เอาแนวคิดการจัดกลุ่มแบบเพิ่มขึ้นมาใช้งานถูกพัฒนาโดย Lebowitz ในปี 1987 (Jain *et al.*, 1999) อัลกอริธึมนี้จะค้นหาผ่านพื้นที่ของโครงสร้างลำดับชั้น เช่น อัลกอริธึมการปีนเขา (hill climbing) และไม่ทำการสร้างโครงสร้างลำดับชั้นทั้งภาพแบบบนลงล่างและล่างขึ้นบน นอกจากนี้อัลกอริธึมยังมีตัวดำเนินการสำหรับการรวม และการลบโหนด ที่เกี่ยวข้องกับแผนภาพต้นไม้ย่อย UNIMEM มีปัญหาบางอย่าง เช่น โหนดบนสุดจะถูกแก้ไขโดยไม่คำนึงถึงว่าพวกมันจะมีความสามารถจับคู่กับข้อมูลที่ถูกสังเกต ซึ่งนำไปสู่ความลำเอียง (Bias) ของภาพแบบแนวคิดที่จะปรากฏออกมา นั้นคือเมื่อมีการนำเสนอกับจำนวนของข้อมูลจำนวนมาก และในความเป็นจริงแล้ว UNIMEM จะเก็บข้อมูลฝึกฝน และดังนั้นจึงมีโครงสร้างแบบลำดับชั้นที่ใหญ่มาก

เนื่องจากอัลกอริธึม COBWEB มีบางส่วนดึงแนวคิดโครงสร้างลำดับชั้นแบบบนลงล่างมาใช้งาน ดังนั้น Hadzikadic and Yun (1989) จึงได้นำเสนออัลกอริธึมไอเอ็นซี (INC) ซึ่งพัฒนามาจากวิธีการสร้างแนวคิดเช่นเดียวกับ COBWEB แต่มีความแตกต่างกันในเรื่อง รูปแบบการนำเสนอความรู้ (Knowledge Representation Formalism) ชนิดของแอททริบิวต์ (Type of Attribute Supported) ข้อกำหนดของตัวดำเนินการที่ถูกใช้โดยกระบวนการค้นหา (Definition of Operators) กลไกการประเมินการจัดกลุ่ม (Clustering Evaluation Mechanism) และวิธีการวัดความเหมือน (Similarity Function) วิธีนี้มีข้อดีคือ สามารถทำงานกับค่าที่สูญหายได้ รวมถึงแอททริบิวต์สามารถมีได้หลายค่า (Multiple Value) แต่ข้อเสียที่สำคัญคือ การกำหนดอัตราส่วน (Threshold) เพื่อคัดเลือกแอททริบิวต์ ซึ่งสามารถเพิ่มประสิทธิภาพถ้ากำหนดได้เหมาะสม แต่ถ้ากำหนดอัตราส่วนมากเกินไปจะทำให้คุณภาพของแอททริบิวต์ ที่ใช้อธิบายแต่ละโหนดมีคุณภาพต่ำ ทั้งนี้ได้ทำการทดลองจัดทำโครงสร้างแบบลำดับชั้นกับข้อมูลเกี่ยวกับชื่อประเทศ และมหาวิทยาลัยที่อยู่ในประเทศนั้น

สำหรับปัญหาเรื่องของคลาสที่มีขนาดใหญ่และสามารถทำให้เกิดความลำเอียง ซึ่งทำให้ฟังกชัน CU มีการสร้างโหนดปลอม ดังนั้น Kim and Choi (2005) จึงนำเสนอวิธีการแก้ไขฟังกชัน CU ซึ่งเน้นการให้น้ำหนักกับลักษณะที่มีความแม่นยำในการทำนาย โดยพิจาณาการกระจายตัวของแต่ละแอททริบิวต์ถ้าแอททริบิวต์ A_i มีค่าการกระจายทั่วทั้งระเบียน V_{ij} ดังนั้นแอททริบิวต์ นั้นมีความสัมพันธ์น้อยกับคลาส C_k การแก้ไข CU จะช่วยป้องกันการสร้างโหนดปลอมที่อยู่ตรงกลางของโครงสร้างแบบลำดับชั้น โดยไม่ต้องสูญเสียการเรียนรู้

Al-Omary and Jamil (2006) ได้นำเอาข้อดีของการเรียนรู้แบบมีผู้สอนมาประยุกต์ใช้ และไม่ใช้การคำนวณค่าความน่าจะเป็น (Probability) แบบเดิมของ CU ข้อดีของวิธีการนี้คือ มีภาพแบบการจัดกลุ่มแบบแผนภาพต้นไม้ โดยมีการให้ปัจจัยกับกิ่ง ซึ่งจะใช้เป็นตัววัดกิ่ง แต่ข้อเสียที่สำคัญคือ การใช้เพียงค่าความถี่ในการจัดวางแต่ละแอททริบิวต์ลงในโหนด

บางงานวิจัย เช่น Yuni and Bowei (2007) ได้มุ่งเน้นเรื่องของการจัดการค่าที่ไม่แน่นอน ซึ่งมีความแตกต่างจากงานวิจัยอื่นคือ สามารถจัดกลุ่มกับข้อมูลที่มีข้อมูลความไม่แน่นนอนปนอยู่ได้ โดยใช้ความน่าจะเป็นเข้ามาช่วย ซึ่งเรียกว่าเป็นนิวโทเทิ่ลยูทีลิตี้ (New Total Utility: TU) ซึ่งใช้วัดคุณภาพของการจัดกลุ่ม ข้อดีของวิธีนี้คือ สามารถทำงานกับข้อมูลที่ไม่แน่นอน ข้อเสียคือ ปัญหาลำดับของข้อมูลนำเข้าที่ส่งผลกระทบต่อการสร้างโครงสร้างแบบลำดับชั้น และการใช้งานกับข้อมูลเชิงตัวเลขซึ่งยังไม่ประสบความสำเร็จ

เพื่อแก้ปัญหาในเรื่องของวิธีการบริดเจอร์ (BRIDGER) ที่มีความซับซ้อนซึ่งประกอบด้วยมุมมองแบบมาโคร (Macro) และไมโคร (Micro) Ryszard (1983) จึงได้นำเสนอการปรับปรุงเป็นอัลกอริธึม COBWEB ให้เป็นอัลกอริธึมอีคอบเว็บ (ECOBWEB) ข้อดีคือ สามารถทำงานกับข้อมูลชนิดตัวเลข (Real Value) โดยการเพิ่มค่าน้ำหนัก และกำหนดตัวดำเนินการเพิ่ม และข้อเสียคือ การกำหนดพารามิเตอร์ α ที่เหมาะสม

จากปัญหาของการเรียนรู้ เพื่อนำมาใช้จัดองค์ประกอบสำหรับออกแบบองค์ความรู้ซึ่งมีไม่เพียงพอ เนื่องจากองค์ความรู้ที่ได้จากการเรียนรู้ส่วนใหญ่เชื่อถือกลไกการจัดกลุ่มที่ถูกสร้างบนการวัดทางสถิติ เช่น ความเหมือน หรือความสามารถในการทำนายทางสถิติของแต่ละกลุ่มบนบางค่าแอทตริบิวต์ธรรมดา ซึ่งมีงานวิจัยบางชิ้นที่รายงานความสามารถในการประยุกต์ใช้อัลกอริธึม COBWEB เพื่อช่วยในช่วงของการออกแบบ (Design Periods) อย่างไรก็ตาม เนื่องจากปัญหาความลำเอียงของคลาสที่มีขนาดใหญ่มากจากอัลกอริธึมของ COBWEB ดังนั้น Chaiworawitkul (2008) จึงนำเสนอการใช้ค่าน้ำหนักเฉลี่ย โดยกำหนดจากวิธีการประเมินแบบครอสวาลิเดชัน (Cross-Validation) ข้อดีของวิธีนี้คือ สามารถแก้ไขเรื่องความลำเอียงของคลาสที่มีขนาดใหญ่ได้ดี แต่

ข้อเสียคือ การกำหนดพารามิเตอร์ค่าน้ำหนักที่เหมาะสม ซึ่งเป็นการยากที่จะสามารถกำหนดได้อย่างแม่นยำ

Rosy *et al.* (2009) ได้ทำการพัฒนาการจัดกลุ่มข้อมูลการแสดงออกของยีนแบบเพิ่มขึ้น ได้ทำการพัฒนาเทคนิคกรจัดกลุ่มแบบเพิ่มขึ้นตามความหนาแน่น (Incremental Density Based Clustering Techniques: IncDGC) เป็นการพัฒนาต่อจากเทคนิคการจัดกลุ่มตามความหนาแน่นแบบกราฟ (DGC) เทคนิค IncDGC อยู่บนการจัดกลุ่มแบบความหนาแน่นและเป็นการจัดกลุ่มแบบเพิ่มขึ้น ซึ่งแตกต่างจากเทคนิค DGC เดิมที่มีข้อเสียของการแทรกยีนมีผลกระทบกับกลุ่มข้อมูลที่มีอยู่ในปัจจุบัน ในทางตรงกันข้ามเทคนิค IncDGC จะไม่มีผลกระทบดังกล่าว วิธีการนี้ประกอบด้วย 2 ขั้นตอนคือ การทำให้ค่าไม่ต่อเนื่อง (Discrete) และการแปลงให้ค่าข้อมูลมีขนาดหรือสเกลที่เท่ากัน (Normalize) จากนั้นทำการจัดกลุ่มตามความหนาแน่น (Density Clustering) โดยการจับคู่ข้อมูลจะอ้างอิงตามภาพแบบการจับคู่แบบปรกติที่มากที่สุด (Maximal Matching Regulation Pattern : MMRP) ตามเงื่อนไขทางตรรกะ ข้อดีคือใช้เวลาที่น้อยกว่าอัลกอริธึมเดิม ข้อเสียคือต้องกำหนดพารามิเตอร์ที่เหมาะสม ไม่มีภาพแบบโครงสร้างแบบลำดับชั้น และขึ้นอยู่กับการเลือกเทคนิคการวัดความเหมือนระหว่างกลุ่มข้อมูล

ตารางที่ 2-3 การเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่มเรียนรู้แบบเพิ่มขึ้น

No.	Algorithms	Incremental	No Parameter Setting	Visualization (Tree Structure or Dendrogram)	Real Value Attribute
1	COBWEB	√	√	√	X
2	CAS	√	X	√	√
3	ECOBWEB	√	X	√	√
4	AutoClass	√	√	X	√
5	ICC	√	X	√	√
6	INCDGC	√	X	√	√
7	CLASSIST	√	√	√	√

จากการศึกษางานวิจัยที่เกี่ยวข้องให้ทำให้ผู้วิจัยเห็นถึงความสามารถในการทำงานของอัลกอริธึมการวัดระยะห่างแบบมาหาลานูบิสในส่วนของการเรียนรู้แบบไม่มีผู้สอน และวิธีการเรียนรู้แบบเพิ่มขึ้น จึงได้ทำการสรุปประสิทธิภาพของวิธีการจัดกลุ่มแบบเรียนรู้เพิ่มที่ได้ทำการศึกษามาเปรียบเทียบกัน ได้แก่ อัลกอริธึม COBWEB, CAS, ECOBWEB, AutoClass, ICC, INCDGC, และ CLASSIT ดังแสดงในตารางที่ 2-3 ซึ่งผลของการเปรียบเทียบในตารางจะช่วยให้เห็นถึงความสามารถของการรองรับการทำงานรวมถึงข้อจำกัดของแต่ละอัลกอริธึมที่มีอยู่ เช่น การแสดงภาพเสมือนจริง หรือในเรื่องของความต้องการกำหนดค่าพารามิเตอร์

จากทฤษฎีและงานวิจัยที่เกี่ยวข้องที่กล่าวมา จะถูกนำมาสังเคราะห์เพื่อสร้างแนวทางการทำวิจัย ซึ่งงานวิจัยนี้จะเน้นการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นบนกลุ่มข้อมูลการแสดงออกของยีน เนื่องจากการพัฒนาขั้นตอนดังกล่าวนั้น ต้องอาศัยความรู้และความเข้าใจเกี่ยวกับผลลัพธ์ของอัลกอริธึมที่ผู้อื่นได้ทำการทดลอง และรวมถึงผู้วิจัยได้ทำการวิจัยมาก่อนหน้า ได้แก่ เทคนิคการคัดเลือกมิติข้อมูล การจัดกลุ่มข้อมูล และการวิเคราะห์การจัดกลุ่ม การจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น อัลกอริธึมที่กล่าวมาเหล่านี้จะถูกนำมาประยุกต์ใช้เพื่อออกแบบโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

บทที่ 3
วิธีดำเนินการวิจัย

จากการสำรวจงานวิจัย และทฤษฎีที่เกี่ยวข้องในบทที่ 2 ซึ่งเป็นพื้นฐานที่จะนำมาประยุกต์ใช้สำหรับงานวิจัยเกี่ยวกับการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นบนกลุ่มข้อมูลการแสดงออกของยีนโดยอาศัยเทคนิคการคัดเลือกมิติข้อมูลร่วมกับการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (Incremental Clustering) และนำเสนอโมเดลการทำงานร่วมกันของทั้งสองวิธี สำหรับนำมาใช้พัฒนาโครงร่างขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นบนกลุ่มข้อมูลการแสดงออกของยีน

สำหรับในบทนี้จะขอนำเสนอวิธีการดำเนินการวิจัย 8 ขั้นตอน ได้แก่ (1) การวิเคราะห์ความต้องการของการวิจัย (2) การออกแบบโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (3) การเตรียมข้อมูล (4) การออกแบบวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอ (5) โครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (6) การออกแบบระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (7) การวัดประสิทธิภาพของการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น และ (8) การทดสอบประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ซึ่งมีรายละเอียดดังนี้

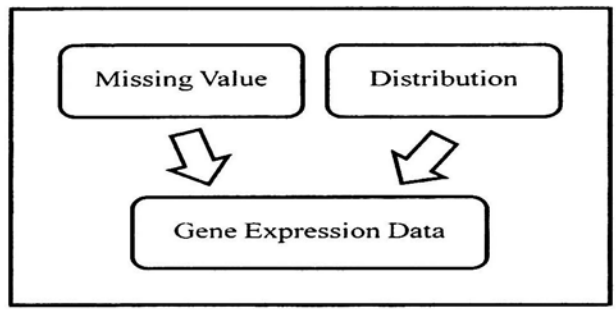
3.1 การวิเคราะห์ความต้องการของการวิจัย

จากปัญหาของการได้มาซึ่งข้อมูลการแสดงออกของยีน จากข้อมูลไมโครอะเรย์ที่ได้กล่าวมาในบทที่ 2 เห็นได้ว่าข้อมูลต้นฉบับหรือข้อมูลดั้งเดิมที่ได้มาจากกระบวนการทดลองทางชีวภาพนั้นไม่อยู่ในสภาพที่สมบูรณ์พอที่จะนำไปใช้งานได้อย่างมีประสิทธิภาพ ทั้งในเรื่องของ ค่าที่สูญหายการจากทดลอง รวมถึงการกระจายตัวที่ไม่เหมาะสม สิ่งเหล่านี้ก่อให้เกิดผลลัพธ์ที่ไม่ถูกต้อง หรือคลาดเคลื่อนไปจากค่าที่เป็นจริง ดังแสดงในภาพที่ 3-1

ปัญหาที่สำคัญอีกอย่างหนึ่งของข้อมูลการแสดงออกของยีน คือ ปัญหาในด้านมิติของข้อมูลที่อยู่บนไมโครอาเรย์ดีเอ็นเอมีอยู่เป็นจำนวนมาก ซึ่งโดยทางทฤษฎีแล้วการที่มียีนจำนวนมากอาจจะให้อำนาจจำแนกสูง อย่างไรก็ตามสิ่งนี้อาจจะเป็นสาเหตุของหลาย ๆ ปัญหา อาทิ การเพิ่มเวลาในการคำนวณและค่าใช้จ่าย มีความซ้ำซ้อนมาก หรือมียีนที่ไม่เกี่ยวข้อง และทำให้การแยกประเภทผิดพลาดได้ นอกจากนี้การค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนกยังสามารถลดสิ่งรบกวน (ซึ่งสามารถปรากฏได้เสมอ ๆ จากการทดลองทางชีววิทยา) และยังสามารถปรับปรุงประสิทธิภาพการแยกประเภทได้มากขึ้นอีกด้วย เทคนิคการกรอง เป็นเทคนิคการเลือกกลุ่มย่อยของยีนที่นิยม

นำมาใช้คัดเลือกยีนที่มีอำนาจจำแนก ซึ่งวิธีนี้จะทำการค้นหามิติของข้อมูลที่มีความเกี่ยวข้องกับคลาส โดยไม่อาศัยความสามารถของอัลกอริธึมที่ใช้จำแนกหมวดหมู่มาร่วมด้วย ซึ่งเทคนิคนี้จะให้ประสิทธิภาพในด้านของความเร็วในการทำงาน อย่างไรก็ตามวิธีการกรองบางวิธีอาจจะมีข้อเสียในเรื่องของการละทิ้งความสัมพันธ์ระหว่างมิติของข้อมูล และให้ความแม่นยำต่ำกว่าวิธีการควบรวม ซึ่งจะส่งผลต่อความแม่นยำในการจำแนกและจัดกลุ่มของยีน

การคัดเลือกมิติของยีนโดยใช้ร่วมกับอัลกอริธึมจำแนกหมวดหมู่หรือการจัดกลุ่ม ส่วนใหญ่ตามงานวิจัยต่าง ๆ ที่นำวิธีนี้มาใช้งานนั้น พบว่าวิธีการกรองมักจะได้รับความนิยมเนื่องจากใช้เวลาน้อยกว่าวิธีการควบรวมมาก อย่างไรก็ตามวิธีการนี้บางครั้งให้ผลลัพธ์ที่ไม่ถูกต้องเนื่องจากแนวคิดของวิธีการเหล่านี้ไม่ได้ถูกคิดค้นมาจากพื้นฐานของแนวคิดทางด้านชีวภาพ คือ นอกจากค้นหายีนที่มีอำนาจจำแนกแล้ว ยีนบางตัวยังมีความสัมพันธ์ระหว่างกันอีกด้วย



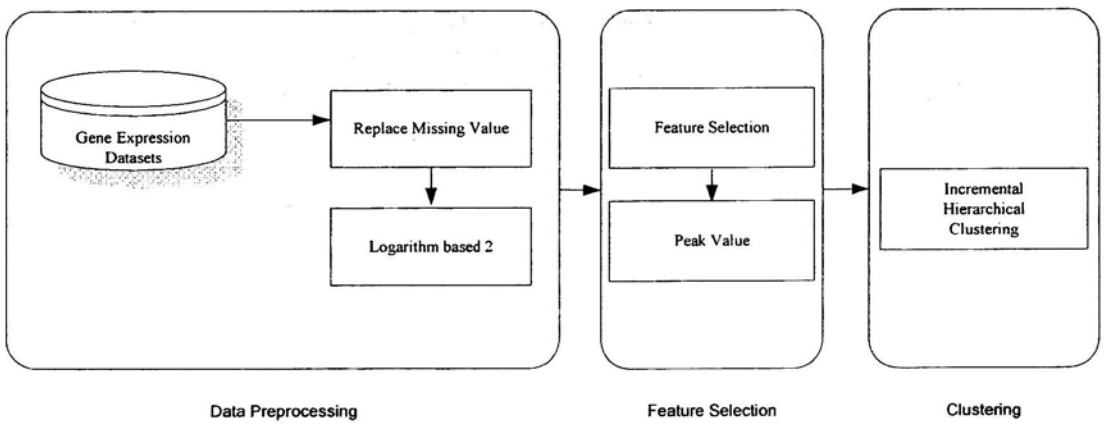
ภาพที่ 3-1 ปัญหาที่พบได้จากข้อมูลการแสดงออกของยีนต้นฉบับ

ดังนั้นงานวิจัยนี้จึงได้นำเสนอการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ซึ่งประกอบด้วยการนำเสนอกรอบของการเตรียมข้อมูลการแสดงออกของยีนที่มีประสิทธิภาพ ในการเตรียมข้อมูลการแสดงออกของยีนให้เหมาะสมก่อนที่จะใช้งาน รวมถึงนำเสนอวิธีการคัดเลือกมิติบนข้อมูลการแสดงออกของยีน โดยอาศัยเทคนิคการเลือกข้อมูลใช้ความสัมพันธ์ (Correlation Based Feature Selection: Cfs) ร่วมกับเทคนิคเกนเรโช (Gain Ratio: GR) แทนวิธีการกรองแบบดั้งเดิมที่ใช้เพียงเทคนิคใดเทคนิคหนึ่งเท่านั้น เทคนิคที่ออกแบบเพื่อทำการกรองข้อมูลการแสดงออกของยีนโดยการทำงานร่วมกันของทั้งสองวิธีนี้สามารถคัดเลือกมิติของข้อมูลที่มีอำนาจจำแนก และมีการแสดงออกที่สูงได้ จากวิธีดังกล่าวจะช่วยให้การคัดเลือกยีนที่มีการแสดงออกได้เหมาะสม ซึ่งสามารถส่งผลให้เพิ่มประสิทธิภาพสำหรับการจัดกลุ่มข้อมูล (Clustering Model) การแสดงออกของยีนได้ดียิ่งขึ้น งานวิจัยนี้ได้ประยุกต์ใช้อัลกอริธึมการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (COBWEB) ร่วมกับวิธีการเรียนรู้แบบเพิ่มขึ้นบนการวัดระยะทาง

แบบมาหาลาโนบิส (Incremental Learning Based on Mahalanobis Distance: ILM) เพื่อสร้างโมเดลการจัดกลุ่ม ไปพร้อม ๆ กับเสนอโครงสร้างแบบลำดับชั้นได้อีกด้วย

3.2 การออกแบบโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

จากการศึกษางานวิจัยที่ผ่านมาพบว่าข้อมูลการแสดงออกของยีนที่ได้จากการทดลองทางชีวภาพนั้นยังมีความสมบูรณ์ไม่เพียงพอ และการวิเคราะห์ความต้องการของงานวิจัยเกี่ยวกับรูปแบบการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น บนกลุ่มข้อมูลการแสดงออกของยีนในหัวข้อที่ 3.1 พบว่า โครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นควรจะประกอบด้วย 3 ส่วน ได้แก่ ส่วนของการเตรียมข้อมูล ส่วนของการคัดเลือกมิติข้อมูล และส่วนของการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ซึ่งผลจากการศึกษางานวิจัยที่ได้ทำการวิจัยมาก่อนหน้านี้พบว่า การเตรียมข้อมูล และการคัดเลือกมิติข้อมูลที่เหมาะสมสามารถช่วยให้สมรรถนะของการทำงานเพิ่มมากขึ้นและส่งผลให้ได้ผลลัพธ์ที่ดี ดังนั้นในการออกแบบโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น จึงอาศัยส่วนประกอบทั้งสามส่วนตามที่ได้กล่าวมา เพื่อใช้สำหรับการออกแบบแนวทางในการดำเนินงานวิจัยดังได้แสดงในภาพที่ 3-2



ภาพที่ 3-2 ภาพรวมของโครงร่างการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

จากภาพที่ 3-2 โครงร่างการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ที่ได้ทำการออกแบบนั้นจะประกอบด้วยขั้นตอนของงานที่สำคัญ ๆ แบ่งเป็น 3 ส่วน คือ

3.2.1 ส่วนของการเตรียมข้อมูล แบ่งเป็นขั้นตอนย่อย ๆ 2 ขั้นตอนคือ

3.2.1.1 ขั้นตอนการเปลี่ยนรูปข้อมูล

3.2.1.2 ขั้นตอนการทดแทนข้อมูลที่สูญหาย

3.2.2 ส่วนของการคัดเลือกมิติข้อมูล จะแบ่งย่อยออกเป็น 2 ขั้นตอนคือ

3.2.2.1 การคัดเลือกมิติของข้อมูลโดยใช้เทคนิคที่ผู้วิจัยได้พัฒนาและออกแบบเอาไว้

3.2.2.2 ขั้นตอนการคัดเลือกข้อมูลที่มีค่าการแสดงออกที่สูง

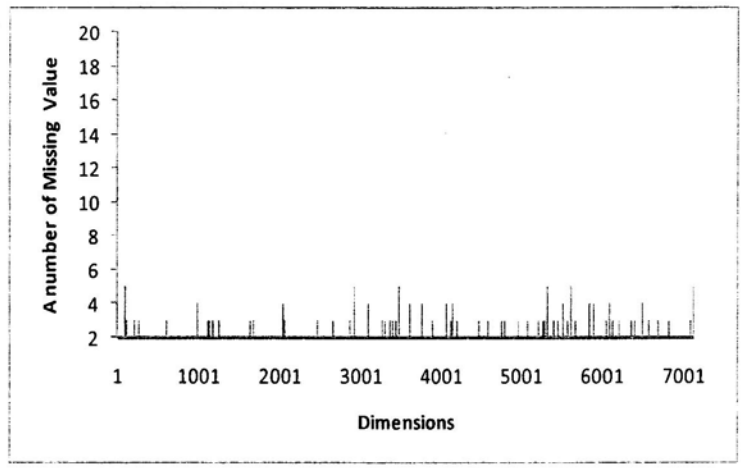
3.2.3 ส่วนของการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

สำหรับการทดลองโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นนี้ ผู้วิจัยได้ทำการศึกษาทดลองดังมีรายละเอียดแต่ละหัวข้อต่อไปนี้

3.3 การเตรียมข้อมูล

ในขั้นตอนนี้จะเป็นการเตรียมข้อมูลการแสดงออกของยีนต้นฉบับให้อยู่ในรูปแบบที่เหมาะสม ซึ่งในการทดลองนี้ได้เตรียมข้อมูลสำหรับใช้ในงานวิจัยนี้ โดยใช้กลุ่มข้อมูลการแสดงออกของยีนที่เปิดเผยต่อสาธารณะซึ่งเป็นฐานข้อมูลที่ถูกใช้ในหลายงานวิจัย โดยมีจำนวนทั้งสิ้น 4 ฐานข้อมูล ได้แก่ DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer ซึ่งกลุ่มข้อมูลเหล่านี้สามารถนำมาใช้งานได้ที่ http://www.upo.es/eps/bigs/datasets.html สำหรับข้อมูล Leukemia, Embryonal Tumor และ Colon Cancer http://datam.i2r.a-star.edu.sg/datasets/krbd /DLBCL/DLBCL-NIH.html สำหรับข้อมูล DLBCL

ก) กลุ่มข้อมูล DLBCL (Diffuse Large B-cell Lymphoma) ถูกใช้เพื่อเปรียบเทียบประสิทธิภาพการจัดกลุ่มเชิงลำดับชั้นแบบเพิ่มขึ้น ข้อมูลชุดนี้ถูกใช้ในงานวิจัยของ Rosenwald *et al.* (2002) ข้อมูลนี้ประกอบด้วยข้อมูลตัวอย่าง 240 ข้อมูลจากกลุ่มของผู้ป่วยซึ่งถูกนำมาทดสอบค่าการแสดงออกของยีนที่ใช้ดีเอ็นเอไมโครอะเรย์ และทำการวิเคราะห์จีโนมที่ไม่ปรกติ นอกจากนี้ข้อมูลนี้ยังประกอบด้วยค่าข้อมูลที่สูญหาย ดังแสดงในภาพที่ 3-3 ข้อมูลชุดนี้มีจำนวนแอททริบิวต์ 7399 โดยแยกเป็นข้อมูลกลุ่มฝึกฝนจำนวน 160 ข้อมูล และข้อมูลกลุ่มทดสอบจำนวน 80 ข้อมูล ในส่วนของแอททริบิวท์จะแบ่งเป็นแอททริบิวต์คลาสหนึ่งแอททริบิวต์ และที่เหลือเป็นแอททริบิวต์ยีนทั้งหมด สำหรับแอททริบิวต์คลาสจะแบ่งเป็นสองค่าคือเสียชีวิต (Dead) และรอดชีวิต (Alive) โดยที่ข้อมูลคลาส Alive มีอยู่ 102 ข้อมูล และข้อมูลคลาส Dead มีอยู่ 138 ข้อมูล



ภาพที่ 3-3 กราฟข้อมูลที่สูญหายจากข้อมูล DLBCL

ข) กลุ่มข้อมูล Leukemia เป็นกลุ่มข้อมูลที่เปิดเผยแก่สาธารณะ ที่ตั้งของข้อมูลนี้คือ Bioinformatics Research Group ของมหาวิทยาลัย Pablo Olavide Sevilla ประเทศสเปน ข้อมูลชนิดนี้ถูกตีพิมพ์ในงานวิจัยของ Golub *et al.* (1999) ข้อมูลนี้ประกอบด้วยข้อมูลตัวอย่างฝึกฝนจำนวน 38 ข้อมูล และข้อมูลทดสอบจำนวน 34 ข้อมูล ข้อมูลทั้งสองชุดมีจำนวนแอททริบิวต์ 7129 แอททริบิวต์ แบ่งออกเป็นแอททริบิวต์ คลาสหนึ่งแอททริบิวต์ และที่เหลือเป็นแอททริบิวต์ ยืนทั้งหมด สำหรับแอแอททริบิวต์คลาสจะแบ่งเป็นสองค่าคือ ALL และ AML โดยที่ข้อมูลคลาส ALL มีอยู่ 47 ข้อมูล และข้อมูลคลาส AML มีอยู่ 25 ข้อมูล

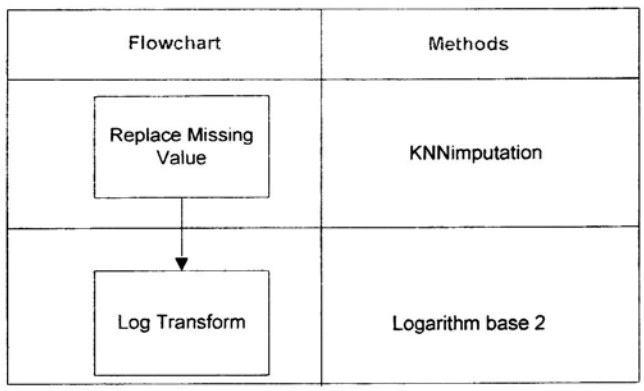
ค) กลุ่มข้อมูล Embryonal Tumor เป็นกลุ่มข้อมูลที่เปิดเผยแก่สาธารณะและมีที่ตั้งของข้อมูลเช่นเดียวกับกลุ่มข้อมูล Leukemia ข้อมูลชนิดนี้ถูกใช้ในงานวิจัยของ Pomeroy *et al.* (2002) ข้อมูลนี้ประกอบด้วยจำนวนข้อมูลตัวอย่าง 60 ข้อมูล และ 7129 แอททริบิวต์ แบ่งออกเป็นแอททริบิวต์คลาสหนึ่งแอททริบิวต์ และที่เหลือเป็นแอททริบิวต์ ยืนทั้งหมด สำหรับแอททริบิวต์คลาสจะแบ่งเป็น 2 ค่าคือ 0 และ 1 โดยที่ข้อมูลคลาส 0 มีอยู่ 39 ข้อมูล และข้อมูลคลาส 1 มีอยู่ 21 ข้อมูล

ง) กลุ่มข้อมูล Colon Cancer เป็นกลุ่มข้อมูลที่เปิดเผยแก่สาธารณะและมีที่ตั้งของข้อมูลเช่นเดียวกับกลุ่มข้อมูล Leukemia และ Lymphoma ข้อมูลชนิดนี้ถูกใช้ในงานวิจัยของ Alon *et al.* (1999) ข้อมูลนี้ประกอบด้วยจำนวนข้อมูลตัวอย่าง 62 ข้อมูล และ 2000 แอททริบิวต์ แบ่งออกเป็นแอททริบิวต์คลาสหนึ่งแอททริบิวต์ และที่เหลือเป็นแอททริบิวต์ ยืนทั้งหมด สำหรับแอททริบิวต์คลาสจะแบ่งเป็น 11 ค่าคือ Positive และ Negative โดยที่ข้อมูลคลาส Positive มีอยู่ 22 ข้อมูล และข้อมูลคลาส Negative มีอยู่ 40 ข้อมูลรายละเอียดของกลุ่มข้อมูลทั้ง 4 ชุดดังแสดงในตารางที่ 3-1

ตารางที่ 3-1 ข้อมูลการแสดงออกของยีนที่ใช้ในการวิจัย

Datasets	Instances	Attributes	Class Values
DLBCL	240	7399	2 (Dead and Alive)
Leukemia	72	7129	2 (ALL and AML)
Embryonal Tumor	60	7129	2 (0 and 1)
Colon Cancer	62	2000	2 (Positive and Negative)

3.3.1 ส่วนของวิธีการทดแทนค่าสูญหาย การทดลองนี้จะเลือกวิธีการทดแทนค่าสูญหายด้วย K-Nearest Neighbor Imputation (KNNimputation) เนื่องจากเป็นเครื่องมือที่มีประสิทธิภาพและได้ผ่านการทดลองในงานวิจัยของ Herrero (2002) พบว่ามีคุณภาพสูงกว่าเครื่องมือการทดแทนค่าสูญหายต่าง ๆ และยังได้รับการบรรจุลงในซอฟต์แวร์ทางการคำนวณ อาทิเช่น Matlab รวมถึงซอฟต์แวร์ทางด้านชีววิทยา (http://bioinfo.cnio.es/cgi-bin/tools/dnarray/preprocess) จึงมีความเหมาะสมสำหรับการทดแทนค่าข้อมูลสำหรับการแสดงออกของยีน กระบวนการในขั้นตอนนี้ดังแสดงในภาพที่ 3-4 และจากนั้นกลุ่มข้อมูลที่ผ่านการเตรียมข้อมูลจะถูกส่งไปยังส่วนการคัดเลือกมิติข้อมูลถัดไป



ภาพที่ 3-4 วิธีการการทดแทนค่าสูญหายและการเปลี่ยนรูปข้อมูล

3.3.2 ส่วนของการเปลี่ยนรูปข้อมูล เนื่องจากข้อมูลการแสดงออกของยีนที่นำมาใช้ทำการวิจัยนี้ มีการกระจายตัวของข้อมูลที่ไม่ปรกติ (Non-normal distribution) ซึ่งเป็นพฤติกรรมโดยปรกติ

ทั่วไปของข้อมูลการแสดงออกของยีนที่ได้จากการกระบวนการทดลองทางชีวภาพ ลักษณะการกระจายตัวของข้อมูลที่ไม่ปรกติเช่นนี้มักส่งผลให้การทำงานของอัลกอริธึมไม่มีประสิทธิภาพเท่าที่ควร ดังนั้นงานวิจัยนี้จึงใช้วิธีการเปลี่ยนรูปข้อมูลให้อยู่ในรูปของการกระจายตัวแบบปรกติ โดยใช้วิธีการ Log Transform โดยเลือกใช้วิธีการ Logarithm ฐาน 2 ดังแสดงในภาพที่ 3-4 ซึ่งวิธีการนี้เป็นวิธีการที่มีงานวิจัยส่วนใหญ่ใช้ในการทำวิจัย (Schaefer and Nakashima, 2010; Wang *et al.*, 2007)

3.4 การออกแบบวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอ

เนื่องจากข้อมูลการแสดงออกของยีนที่ได้รับจากข้อมูลไมโครอะเรย์เป็นการทดลองกับยีนจำนวนหลายพันยีนในคราวเดียวกัน ดังนั้นการคัดเลือกมิติข้อมูลจึงเป็นสิ่งจำเป็นเพื่อที่จะหากลุ่มของยีนที่มีอำนาจจำแนก และเมื่อได้กลุ่มของยีนดังกล่าวมาแล้วก็จะส่งกลุ่มของข้อมูลยีนไปยังส่วนของการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเรียนรู้เพิ่มขึ้นต่อไป

วิธีการคัดเลือกมิติบนข้อมูลการแสดงออกของยีนที่นำเสนอ เป็นส่วนหนึ่งที่ได้จากการเก็บรวบรวมและวิเคราะห์ความต้องการของระบบที่ผ่านมา ดังนั้นในขั้นตอนของการคัดเลือกมิตินี้จะนำเสนอการวิธีการคัดเลือกมิติซึ่งประกอบด้วยการทำงาน 4 ขั้นตอน คือ

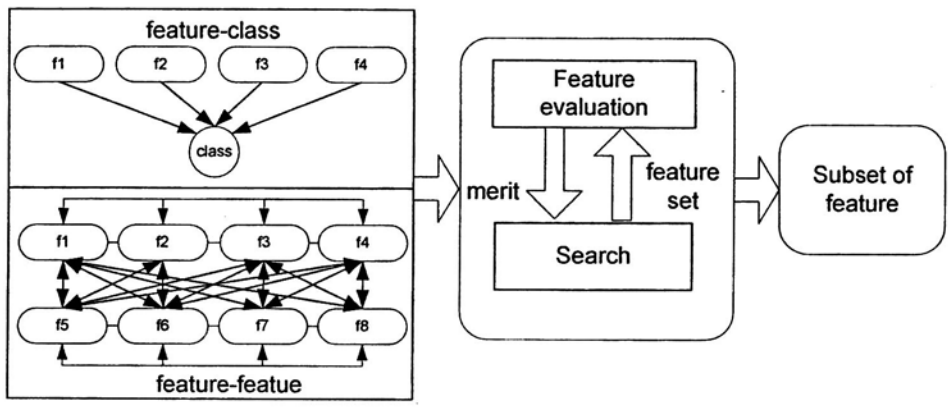
1) ขั้นตอนการคัดเลือกมิติบนข้อมูลการแสดงออกของยีนด้วยการหาความสัมพันธ์
2) ขั้นตอนการคัดเลือกมิติบนข้อมูลการแสดงออกของยีนด้วยการหายีนที่มีอำนาจจำแนก
3) ขั้นตอนการนำผลลัพธ์ที่ได้จากวิธีการหาความสัมพันธ์มาจับคู่กับวิธีการหาอำนาจจำแนก
4) ขั้นตอนการคัดเลือกมิติของข้อมูลที่มีค่าการแสดงออกสูง

3.4.1 การคัดเลือกข้อมูลมิติการแสดงออกของยีน ด้วยวิธีการกรองแบบความสัมพันธ์เพื่อหายีนที่มีความสัมพันธ์กัน จะใช้อัลกอริธึม Correlation Based Feature Selection หรือ Cfs ในการคัดเลือกมิติของข้อมูลที่มีความสัมพันธ์กันโดยใช้อัลกอริธึม Cfs คำนวณความสัมพันธ์ระหว่างคุณลักษณะกับคลาสเป้าหมาย (feature–class) และความสัมพันธ์ระหว่างคุณลักษณะกับคุณลักษณะ (feature - feature) ดังแสดงในภาพที่ 3-5

Cfs ซึ่งจะจัดอันดับกลุ่มย่อยของมิติข้อมูล ตามความสัมพันธ์ที่อยู่บนพื้นฐานของฟังก์ชันวิธีการแก้ปัญหาแบบวิทยาการศึกษาสำนึก ซึ่งกลุ่มย่อยของมิติข้อมูลจะมีความสัมพันธ์กันสูงกับคลาส และไม่มีความสัมพันธ์กับคลาสอื่น ๆ สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องอาจจะถูกละทิ้ง เพราะมิติข้อมูลเหล่านี้อาจจะมีความสัมพันธ์ต่ำกับคลาส มิติข้อมูลที่ซ้ำซ้อนอาจจะถูกขจัดออกไปจากกลุ่มมิติข้อมูลที่มีความสัมพันธ์สูง

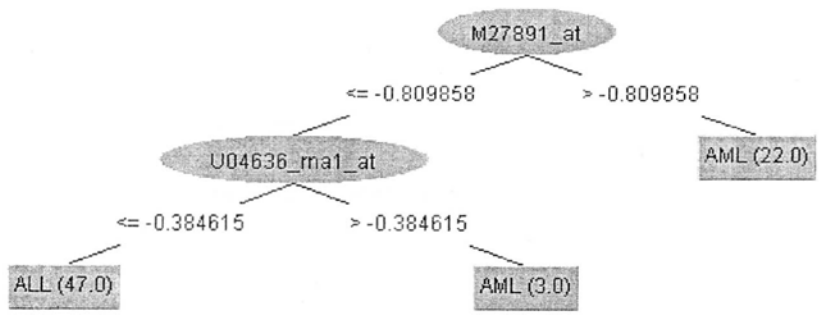
หลังจากที่ได้ข้อมูลการแสดงออกของยีนจากฐานข้อมูลที่เปิดเผยต่อสาธารณะจำนวน 4 กลุ่มข้อมูลผ่านการคัดเลือกมิติแล้ว จะได้กลุ่มข้อมูลย่อยของยีนแต่ละกลุ่มที่มีความสัมพันธ์กัน ซึ่งใน

ขั้นตอนคัดเลือกมิติของข้อมูลนี้จะใช้โมดูลการกรองของโปรแกรมเครื่องจักรการเรียนรู้ WEKA Version 3.7.1 เพื่อวิเคราะห์และค้นหากลุ่มย่อยของยีนที่มีความสัมพันธ์กัน ต่อมาจะนำข้อมูลต้นฉบับไปดำเนินการคัดเลือกยีนที่มีอำนาจจำแนกอีกครั้งหนึ่ง



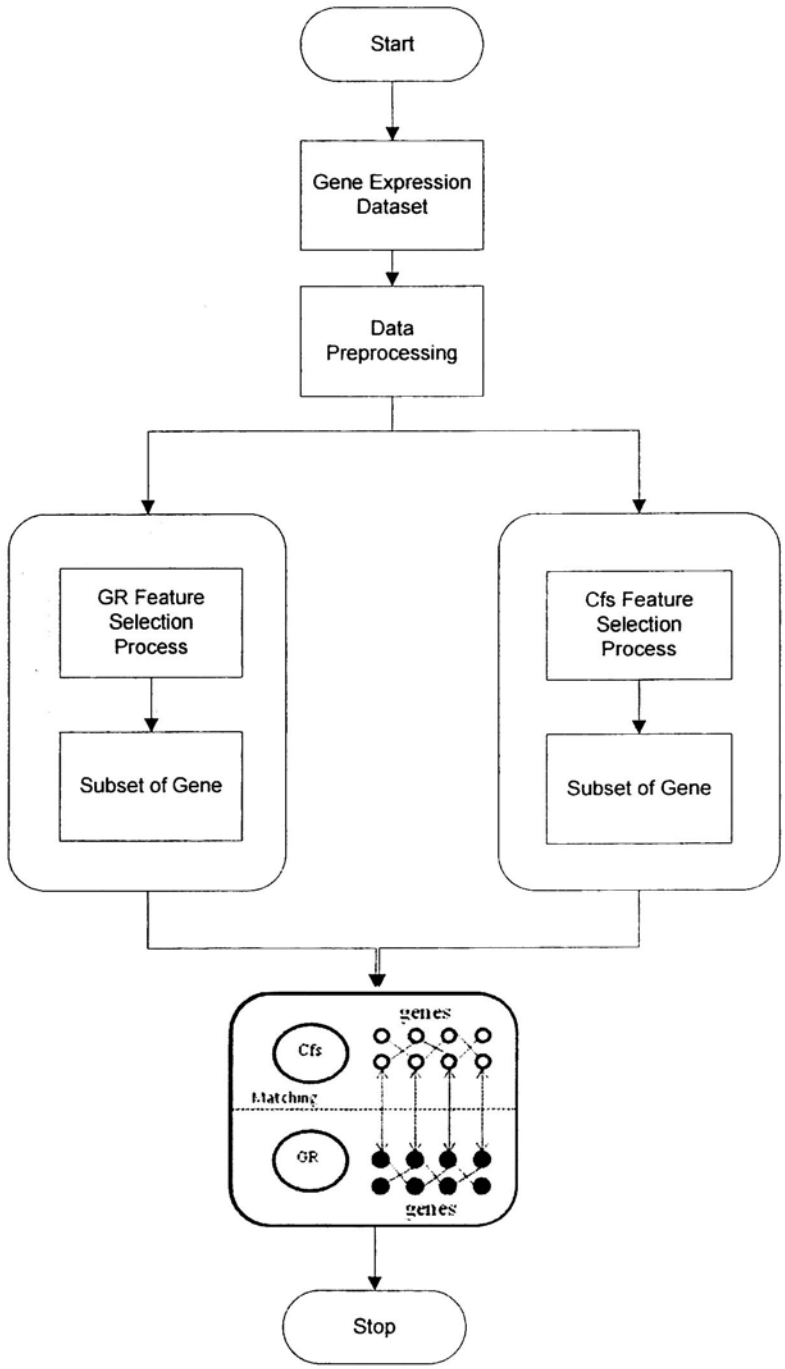
ภาพที่ 3-5 กระบวนดำเนินการของอัลกอริธึม CFS

3.4.2 การคัดเลือกมิติบนข้อมูลการแสดงออกของยีน ด้วยวิธีการกรองแบบค้นหาอำนาจจำแนก ในขั้นตอนนี้จะเป็นการนำเอาข้อมูลต้นฉบับมาคัดเลือกมิติข้อมูลด้วยอัลกอริธึม GR อีกครั้งหนึ่ง สำหรับพื้นฐานของอัลกอริธึม GR มาจากทฤษฎีต้นไม้ตัดสินใจ ซึ่งทฤษฎีต้นไม้ตัดสินใจเป็นเทคนิคที่ง่ายและใช้กันอย่างกว้างขวาง เทคนิคนี้อาศัยพื้นฐานจากทฤษฎี Information Theory และมีการจัดโครงสร้างการออกแบบในลักษณะบนลงล่าง แอททริบิวต์ที่มีอำนาจจำแนกจะถูกเลือกโดยการวัดโดยใช้ Information Theory ซึ่งใช้ความน่าจะเป็นของ $(P(v_i))$, โดยที่ v_i คือคลาส ดังนั้นข้อมูล (I) ของคำตอบถูกกำหนดโดย $I(P(v_i),...,P(v_n)) = \sum_{i=1}^{n}(-P(v_i)\log_2 P(v_i))$ ดังแสดงในภาพที่ 3-6

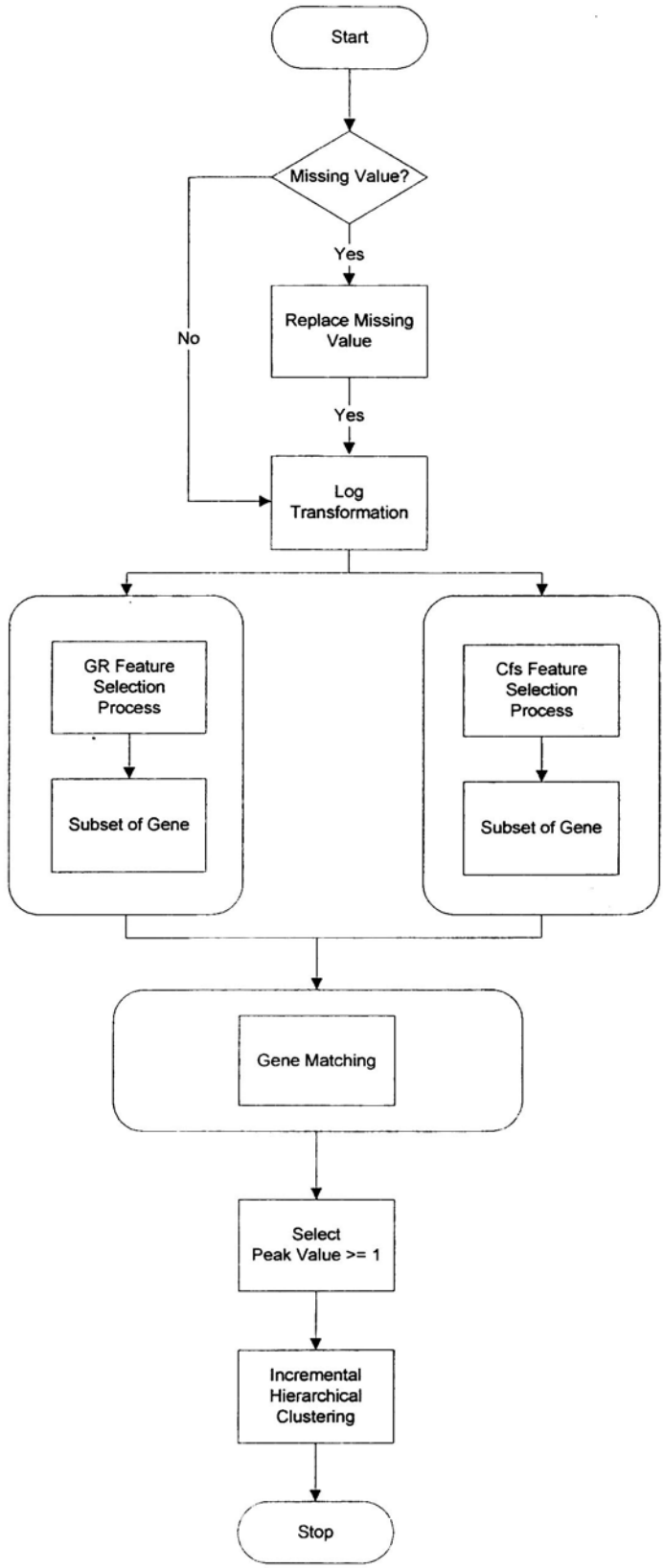


ภาพที่ 3-6 การคัดเลือกมิติในภาพแบบต้นไม้ตัดสินใจ

3.4.3 นำผลลัพธ์ของที่ได้จากวิธีการหาความสัมพันธ์มาจับคู่กับวิธีการหาอำนาจจำแนก ซึ่งในขั้นตอนนี้จะเป็นการนำผลลัพธ์กลุ่มย่อยของยีนที่ได้จากวิธีการหาความสัมพันธ์ด้วยอัลกอริธึม Cfs มาจับคู่กับผลลัพธ์ที่เป็นกลุ่มย่อยของยีนที่ได้จากวิธีการ GR กระบวนการนี้จะทำการคัดเลือกยีนที่มีทั้งคุณลักษณะมีความสัมพันธ์และมีอำนาจจำแนก ดังแสดงในภาพที่ 3-7



ภาพที่ 3-7 กระบวนการคัดเลือกมิติข้อมูลยีนที่นำเสนอ



ภาพที่ 3-8 กระบวนการขั้นเลือกมิติข้อมูลที่นำเสนอร่วมกับการค้นหาค่าการแสดงออกที่สูง

3.4.4 นำผลลัพธ์จากการคัดเลือกมิติข้อมูลยีน มาทำการคัดเลือกเพื่อหาค่าการแสดงออกที่สูงโดยใช้วิธีการบางส่วนของ Iyer (1999) ที่กล่าวไว้ใน Herrero (2002) ว่าขั้นตอนการหาค่าการแสดงออกที่สูงนี้ จะเป็นการทำงานร่วมกันระหว่างข้อมูลต้นฉบับผ่านกระบวนการ Log Transform (log2) กับการหาค่าการแสดงออก ค่าการแสดงออกที่สูงจะพิจารณาจากค่ารวมในแต่ละแอทริบิวต์ข้อมูล ในการทดลองนี้จะพิจารณาค่ารวมในแต่ละแอททริบิวต์ที่มากกว่า 1 และเรียกค่าเหล่านี้ว่า peak value

ดังในภาพที่ 3-8 แสดงการสรุปขั้นตอนการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอ โดยเริ่มจากการผ่านขั้นตอนกระบวนการเตรียมข้อมูลดังที่ได้กล่าวไว้ในหัวข้อ 3.3 คือผ่านกระบวนการเปลี่ยนรูปข้อมูลเพื่อลดการกระจายตัว จากนั้นถ้าข้อมูลมีค่าสูญหายก็จะผ่านกระบวนการทดแทนค่าสูญหาย หลังจากนั้นจึงจะเข้าสู่กระบวนการลดมิติข้อมูลตามที่กล่าวไว้ในหัวข้อที่ 3.4.1-3.4.4 สุดท้ายจึงเป็นการตรวจสอบเพื่อคัดเลือกค่าการแสดงออกของยีนที่สูงเกินกว่า 1 ในตอนต่อไปจะกล่าวถึงโครงร่างการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ซึ่งเป็นส่วนสุดท้ายของวิธีการดำเนินการวิจัย

3.5 โครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

ในส่วนนี้จะเป็นการพัฒนาเทคนิคการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น โดยพัฒนาจากอาศัยอัลกอริธึมการจัดกลุ่มแบบเพิ่มขึ้นและการวัดระยะทางแบบมาหาลาโนบิส (Incremental Learning Based on Mahalanobis Distance: ILM) ร่วมกับเทคนิคการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นของอัลกอริธึม COBWEB ซึ่งอัลกอริธึมที่พัฒนาขึ้นใหม่นี้จะมีความสามารถในการจัดกลุ่มแบบเรียนรู้เพิ่มขึ้น และสามารถแสดงโครงร่างแบบลำดับชั้น โดยใช้ข้อมูลการแสดงออกของยีนที่มีอำนาจจำแนกเพื่อทำการเรียนรู้และสร้างโมเดล จากนั้นจะทำการทดลองเพื่อแสดงประสิทธิภาพในการจัดกลุ่มตามวัตถุประสงค์ต่อไป

IHC เป็นอัลกอริธึมสำหรับการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ที่ปรับปรุงการทำงานจากอัลกอริธึม ILM และนำแนวคิดจากเทคนิคการจัดโครงสร้างแบบลำดับชั้นของ COBWEB มาทำงานเข้าด้วยกัน อัลกอริธึม IHC จะอาศัยเทคนิคของการวัดที่มีประสิทธิภาพจากอัลกอริธึม ILM เพื่อวัดระยะทางระหว่างข้อมูล และนำผลลัพธ์ที่ได้มาดำเนินการจัดกลุ่มเชิงลำดับชั้นโดยใช้เงื่อนไขการจัดกลุ่มจากอัลกอริธึม COBWEB พร้อมกันนี้ IHC จะสร้างแบบจำลอง (Prototype) ขึ้นสำหรับแต่ละกลุ่มข้อมูล เพื่อใช้ประโยชน์ในการทำนายข้อมูลทดสอบ หรือเรียนรู้ข้อมูลที่ไม่เคยรู้จักมาก่อน (Unseen Pattern) แบบเพิ่มขึ้นได้ตลอดเวลา ในขณะเดียวกันอัลกอริธึม IHC ยังสามารถใช้เป็นตัวจำแนกหมวดหมู่ (Classifier) ได้อีกด้วย

3.5.1 โครงสร้างของ IHC

อัลกอริธึม IHC มีสองขั้นตอนพื้นฐานที่สำคัญเหมือนกับเทคนิค ILM ดังแสดงในภาพที่ 3-9 คือ

3.5.1.1 ขั้นตอนการเรียนรู้ (Learning Phase) ซึ่งประกอบด้วยอัลกอริธึมที่สำคัญ อัลกอริธึม คือ

ก) อัลกอริธึมเพื่อการเรียนรู้ (Learning Algorithm)

ข) อัลกอริธึมการเรียนรู้แบบเพิ่มขึ้น (Incremental Learning Algorithm)

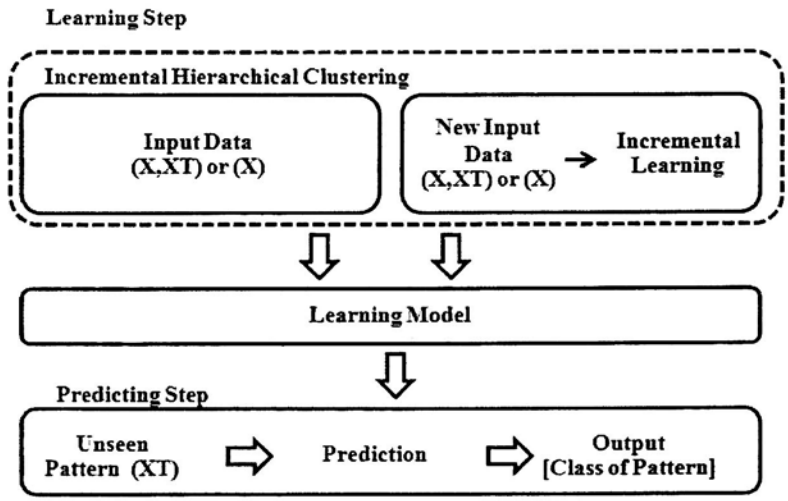
3.5.1.2 ขั้นตอนการทำนายผล (Predicting Phase)

ในขั้นตอนการเรียนรู้อัลกอริธึม IHC สนับสนุนเพียงการเรียนรู้แบบไม่มีผู้สอน ดังนั้นข้อมูลฝึกฝน (Training Data) จึงมีข้อมูลนำเข้า (Input Data) และในขั้นตอนสร้างแบบจำลองจะอาศัยคลาสเป้าหมาย (Target classes) อัลกอริธึม IHC จะใช้การเรียนรู้แบบไม่มีผู้สอนเพื่อเรียนรู้ข้อมูลในขั้นตอนการเรียนรู้ และนำมาสร้างแบบจำลองระบบเพื่อใช้ในการทำนายผลต่อไป

ขั้นตอนฝึกฝนข้อมูล เพื่อใช้สำหรับเรียนรู้ และสร้างโมเดลแบบจำลองสำหรับการจำแนกหมวดหมู่จะอาศัยการจัดกลุ่มเชิงลำดับชั้นแบบเพิ่มขึ้นด้วยเทคนิค COBWEB ร่วมกับการวัดระยะทางระหว่างข้อมูลด้วยการวัดระยะทางแบบมาหาลาโนบิส

และในส่วนของการเรียนรู้แบบเพิ่มขึ้น คลาสเป้าหมายใหม่ (New Target Class) จะถูกเรียนรู้เพื่อรวมเข้าสู่ระบบใหม่ได้ตลอดเวลา ดังนั้นอัลกอริธึม IHC สามารถเรียนรู้ข้อมูลใหม่เพื่อปรับแก้ไขแบบจำลองระบบได้เมื่อมีข้อมูลใหม่มาเรียนรู้ ระบบการเรียนรู้ประกอบด้วยแบบจำลองระบบ ซึ่งได้แก่ W_P และ W_T

แบบจำลองจะถูกสร้างเมื่อชุดข้อมูลฝึกฝนถูกนำเข้ามา และทำการวัดระยะทางของข้อมูลแต่ละตัวในชุดข้อมูลฝึกฝนด้วยวิธีการวัดแบบมาหาลาโนบิส ในแต่ละรอบของการวัดระยะทาง อินพุต p ที่นำเข้ามาใหม่ จะถูกวัดโดยการวัดระยะทางแบบมาหาลาโนบิส (d_m) กับข้อมูลฝึกฝนที่มีอยู่ก่อนหน้า เพื่อค้นหาข้อมูลระยะทางที่ห่างน้อยที่สุดจากชุดฝึกฝนจำนวน 2 ข้อมูล คือ ข้อมูลที่มีระยะทางที่น้อยที่สุด หรือ b_1 (Best Node) และข้อมูลที่มีระยะทางอันดับรองลงมา (Second Node) หรือ s_1 และนำข้อมูล b_1 และ s_1 พร้อมกับ d_m มาสร้างกลุ่มข้อมูลลำดับชั้น และสร้างแบบจำลองระบบ (ในกรณีที่มีแบบจำลองระบบอยู่แล้ว แบบจำลองเดิมจะถูกปรับให้เหมาะสมโดยใช้ข้อมูลที่เข้ามาใหม่) อัตราระยะทางที่ใช้วัด คือ b_1 ที่ $0 < b_1 < 1$, s_1 ที่ $0 < s_1 < 1$ และ d_m ที่ $0 < d_m < 1$ ในกรณีถ้า b_1 และ $s_1 > d_{th}$ ดังนั้นข้อมูลจะถูกพิจารณาว่าอยู่คนละกลุ่มกัน ในทางตรงกันข้าม b_1 และ $s_1 \leq d_{th}$ ดังนั้นข้อมูลจะถูกพิจารณาว่าอยู่ในกลุ่มเดียวกัน



ภาพที่ 3-9 โมเดลของอัลกอริธึมแบบ IHC

ขั้นตอนการทำนายผล อัลกอริธึม IHC จะจำแนกหมวดหมู่ข้อมูลที่ไม่รู้จักมาก่อน ข้อมูลทดสอบของแต่ละข้อมูลที่ใช้ทดสอบ จะถูกจัดโครงสร้างข้อมูลเชิงลำดับชั้น และจัดกลุ่มโดยใช้การวัดระยะทางแบบมาหาลาโนบิส และการเปรียบเทียบระหว่างข้อมูลที่ใช้ทดสอบและโมเดลแบบจำลองระบบ (System Prototypes) โดยใช้ฟังก์ชั่นความเป็นสมาชิกแบบเกาส์เซียน ระดับความเป็นสมาชิกของแต่ละคลาสเป้าหมายของข้อมูลทดสอบจะถูกคำนวณโดยมาหาลานูบิชดิสเกาส์เซียน(Mahalanobis Gaussian) แบบ RBF ซึ่งถือได้ว่าเป็นคุณลักษณะของการตัดสินใจแบบอ่อน (Soft Decision) และสำหรับโหนดโมเดลแบบจำลองใดที่มีระดับของฟังก์ชันความเป็นสมาชิกที่สูงกว่าจะถูกเรียกว่า โหนดผู้ชนะ (Wining Node) และจะถือเป็นคุณลักษณะของการตัดสินแบบเข้ม (Hard Decision) และในขั้นตอนการทำนายผลนี้คลาสเป้าหมายของโหนดผู้ชนะถูกใช้เป็นตัวกำหนดคลาสเป้าหมายของข้อมูลทดสอบ สำหรับขั้นตอนการเรียนรู้ ของกระบวนการ IHC จะถูกอธิบายในส่วนถัดไป

3.5.2 ขั้นตอนการเรียนรู้ (Learning phases)

ในขั้นตอนนี้อัลกอริธึม IHC จะแบ่งข้อมูลออกเป็น 2 ส่วนคือ ชุดข้อมูลฝึกฝน และชุดข้อมูลทดสอบเหมือนเช่นในกระบวนการของวิธี ILM ทั้งชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ ถูกสร้างขึ้นจากกลุ่มข้อมูลการแสดงออกของยีนโดยใช้วิธีการสุ่มเลือกมาอย่างละเท่า ๆ กัน ทั้งนี้เฉพาะข้อมูลสำหรับฝึกฝนที่ถูกนำไปใช้ในขั้นตอนการเรียนรู้และสร้างโมเดลแบบจำลอง ในทางกลับกันข้อมูลทดสอบ จะถูกนำมาทดสอบประสิทธิภาพของการทำงานของ IHC ซึ่งจะได้กล่าวต่อไปในขั้นตอนการวัดประสิทธิภาพหัวข้อที่ 3.7

ผังการไหลของข้อมูล (Flowchart) ของอัลกอริธึมที่นำเสนอจะถูกแสดงในภาพที่ 3-10 ถึง 3-13 ข้อมูลฝึกฝนจะประกอบด้วยข้อมูลนำเข้า p และและคลาสเป้าหมาย t เช่นเดียวกับการเรียนรู้แบบมีผู้สอนของในส่วนของอัลกอริธึม ILM

กระบวนการเรียนรู้ประกอบด้วยขั้นตอนดังนี้

ขั้นตอนที่ 1: ความแปรปรวนร่วม $a\mathbf{I}$ และอัตราระยะทาง (d_{th}) จะถูกกำหนดโดยผู้ใช้งาน

ขั้นตอนที่ 2: ข้อมูลนำเข้า p และอินพุตเป้าหมาย t นำเข้ามา

ขั้นตอนที่ 3: ระยะทางระหว่างข้อมูลนำเข้า p ที่ถูกเก็บไว้ในระบบ และข้อมูลนำเข้า p ที่เข้ามาใหม่ จะถูกวัดโดยการวัดระยะทางแบบมาหาลาโนบิส (d_m)

ขั้นตอนที่ 4: คำนวณระยะทางของข้อมูลนำเข้า p และข้อมูลฝึกฝนที่มีอยู่ จากนั้นเรียงอันดับระยะห่างที่น้อยที่สุด 2 อันดับคือ 1) ระยะทางน้อยที่สุด (b_1) และ 2) ข้อมูลที่มีระยะทางน้อยรองลงมา (s_1)

ขั้นตอนที่ 5: เปรียบเทียบข้อมูลระยะทางทั้ง (b_1) และ (s_1) กับอัตราระยะทาง (d_{th}) ที่กำหนดไว้

1) ถ้าระยะห่างของข้อมูล (b_1) และ (s_1) อยู่ภายใต้อัตราระยะทาง $b_1 <= d_t$ และ $s_1 <= d_{th}$ และ parent node ของ $b_1 = s_1$ เพิ่มอินพุต p ในกลุ่มเดียวกัน กรณี parent node ของ $b_1 \neq s_1$ ให้สร้างโหนดใหม่และเพิ่มข้อมูลนำเข้า p ให้อยู่ในกลุ่มเดียวกับ b_1 และ s_1 แบบจำลองที่มีอยู่ W_p จะถูกปรับแก้โดยการใช้แนวคิดของการเรียนรู้แบบเพิ่มขึ้น
2) ถ้ามีเพียงระยะห่างของข้อมูล (b_1) ที่อยู่ภายใต้อัตราระยะทาง $b_1 <= d_{th}$ แล้วข้อมูลนำเข้า p จะถูกจัดกลุ่มอยู่ในกลุ่มเดียวกับ (b_1) และแบบจำลองที่มีอยู่ W_p ที่ตำแหน่ง (b_1) จะถูกปรับแก้โดยการใช้แนวคิดของการเรียนรู้แบบเพิ่มขึ้น
3) ถ้าระยะห่างของข้อมูลคือ (b_1) และ (s_1) เกินกว่าอัตราระยะทาง $b_1 > d_{th}$ และ $s_1 > d_{th}$ แล้วข้อมูลนำเข้า p ถูกสร้างเป็นกลุ่มใหม่ภายใต้ root และแบบจำลองใหม่จะถูกเพิ่มใน W_p โดยการใช้ข้อมูลนำเข้า p และคลาสใหม่จะถูกเพิ่มใน W_T โดยการใช้เป้าหมาย t

ขั้นตอนที่ 6 : ถ้าการวัดระยะห่าง (b_1) และ (s_1) <= d_{th} หรือ (b_1) <= d_{th} ดังนั้นคลาสเป้าหมายจะถูกบันทึกใน W_T กรณีอื่น แบบจำลองใหม่จะถูกเพิ่มใน W_p โดยการใช้ข้อมูลนำเข้า p และจากนั้นคลาสเป้าหมายใหม่จะถูกเพิ่มใน W_T

3.5.3 กระบวนการพื้นฐาน

เป้าหมายที่สำคัญของขั้นตอนการเรียนรู้สำหรับอัลกอริธึม IHC คือเพื่อทำการจัดกลุ่มข้อมูลเพื่อการทำนายคลาสเป้าหมาย ดังนั้นค่าความถูกต้อง (Accuracy Rate) ในการจัดกลุ่มที่สูงที่สุดบนข้อมูลทดสอบจึงเป็นสิ่งจำเป็น ในกระบวนการนี้ IHC จะแยกข้อมูลทดสอบออกจากข้อมูลฝึกฝนจำนวนเท่า ๆ กันโดยใช้วิธีจัดกลุ่มข้อมูลทั้งหมด

หลังขั้นตอนการเรียนรู้อัลกอริธึม IHC จะได้รับแบบจำลองระบบสำหรับการทำนายข้อมูลที่ไม่ทราบคลาสเป้าหมาย กระบวนการของการทดสอบประสิทธิภาพของระบบมีขั้นตอนดังนี้

ขั้นตอนที่ 1: อ่านตัวข้อมูลทดสอบ p

ขั้นตอนที่ 2: วัดระยะทางระหว่างข้อมูลนำเข้า p และ แบบจำลองที่เรียนรู้ (W_p) โดยการใช้การวัดระยะทางแบบมาหาลาโนบิส (d_m)

ขั้นตอนที่ 3: กำหนดค่าความเป็นสมาชิกโดยการใช้ฟังก์ชันเกาส์เซียนแบบ RBF

ขั้นตอนที่ 4: ค้นหาโหนดผู้ชนะ โดยการเจาะจง ค่าความเป็นสมาชิกที่มากที่สุด

ขั้นตอนที่ 5: กำหนดคลาสเป้าหมายของข้อมูลนำเข้า p โดยการใช้คลาสเป้าหมายของโหนดผู้ชนะ

ดังนั้นคลาสเป้าหมายของข้อมูลนำเข้า p จะถูกกำหนดจากคลาสเป้าหมายของโหนดผู้ชนะโดยคำนวณจากค่าความเป็นสมาชิกโดยการใช้ฟังก์ชันเกาส์เซียนแบบ RBF กับโมเดลแบบจำลองที่ได้รับจากขั้นตอนการเรียนรู้ทั้งหมด

3.5.4 ขั้นตอนการทำนายผล

อัลกอริธึม IHC มีขั้นตอนที่สำคัญอันหนึ่งคือ ขั้นตอนการทำนายผล คือ ในขั้นตอนนี้เทคนิคการจำแนกหมวดหมู่แบบ IHC จะถูกใช้ทำนายคลาสเป้าหมายของข้อมูลที่ไม่ทราบคลาสเป้าหมาย โดยใช้การวัดระยะทางระหว่างข้อมูลที่ไม่ทราบคลาสเป้าหมาย และแบบจำลองระบบทั้งนี้กระบวนการของการทำนายผล มีขั้นตอนดังนี้

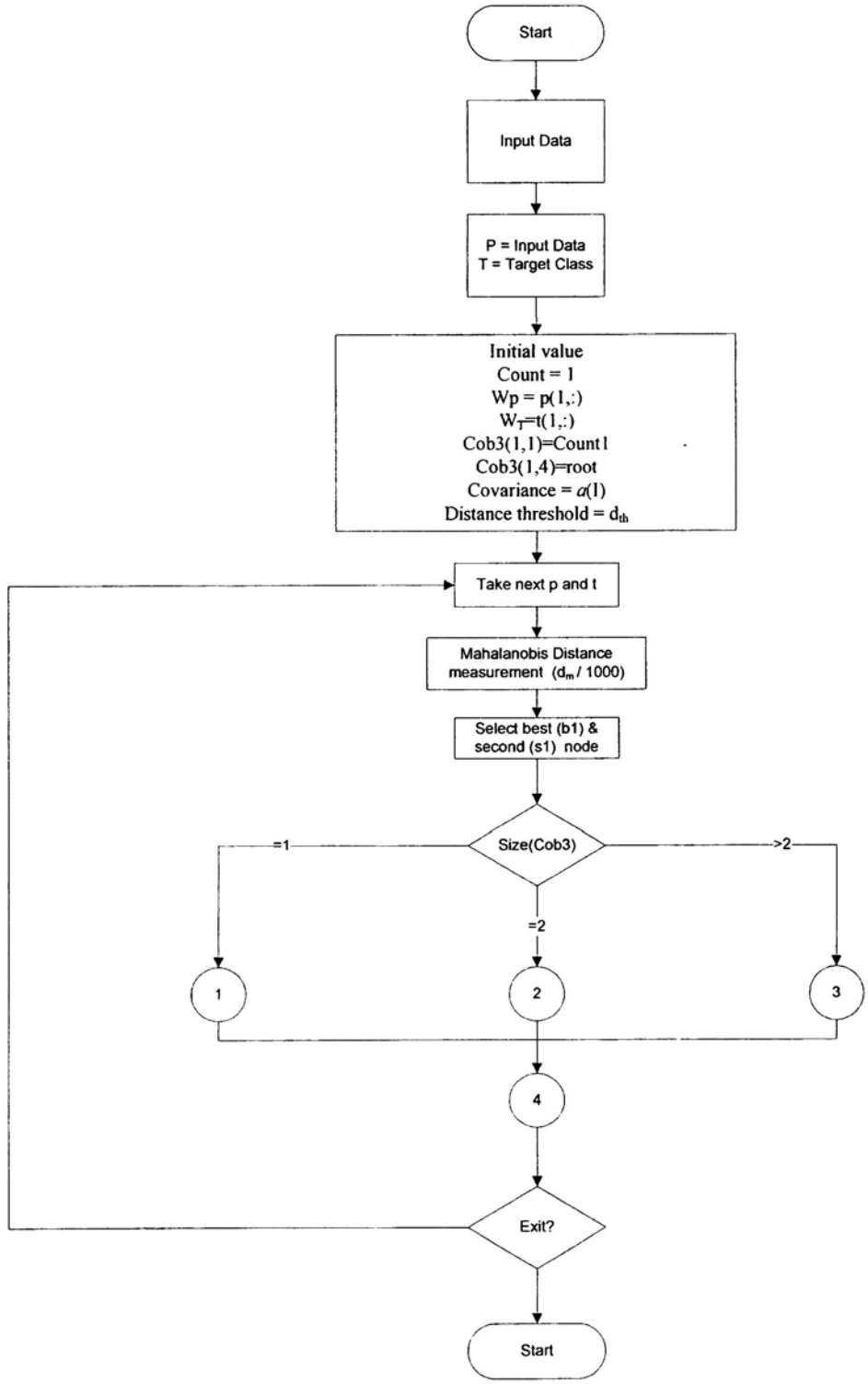
ขั้นตอนที่ 1: อ่านข้อมูลที่ไม่ทราบคลาสเป้าหมายข้อมูลนำเข้า p

ขั้นตอนที่ 2: วัดระยะทางระหว่างข้อมูลนำเข้า p และนำไปทำการเรียนรู้กับโมเดลแบบจำลอง (W_p) โดยอาศัยเทคนิคการวัดระยะทางแบบมาหาลาโนบิส (d_m)

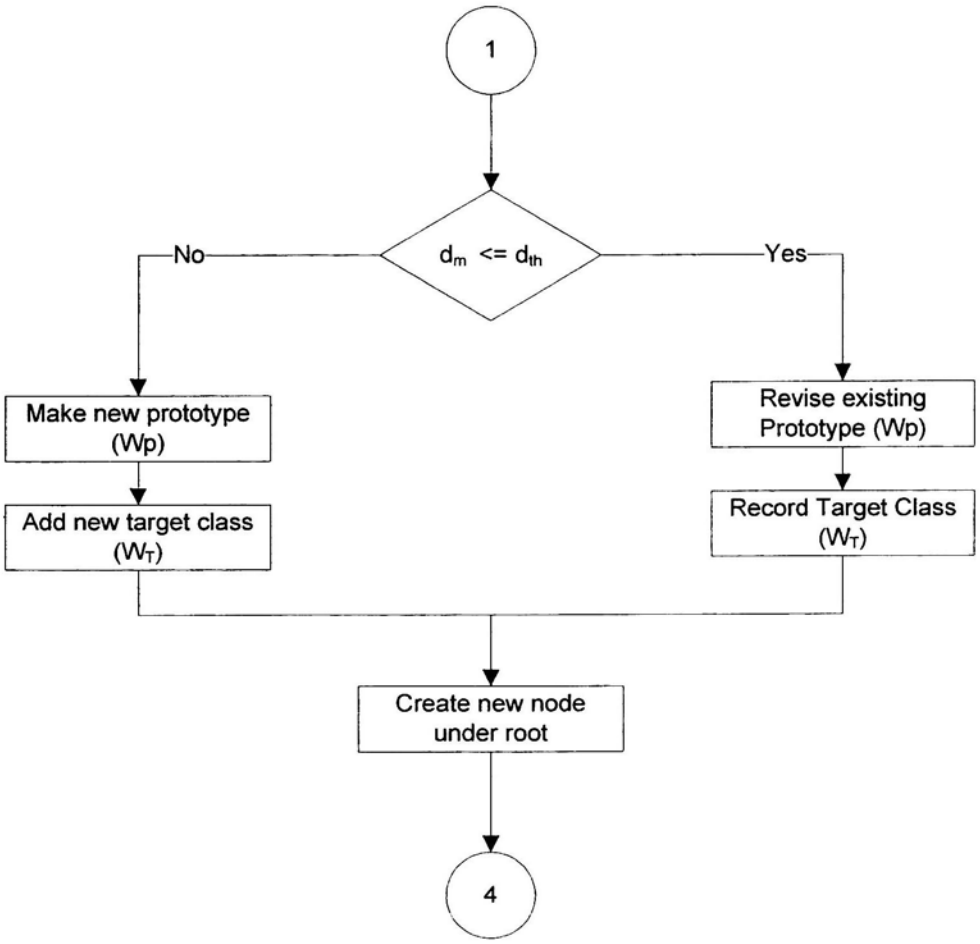
ขั้นตอนที่ 3: กำหนดค่าความเป็นสมาชิก โดยการใช้ฟังก์ชันเกาส์เซียนแบบ RBF

ขั้นตอนที่ 4: ค้นหาโหนดผู้ชนะโดยการเจาะจงค่าความเป็นสมาชิกที่มากที่สุด

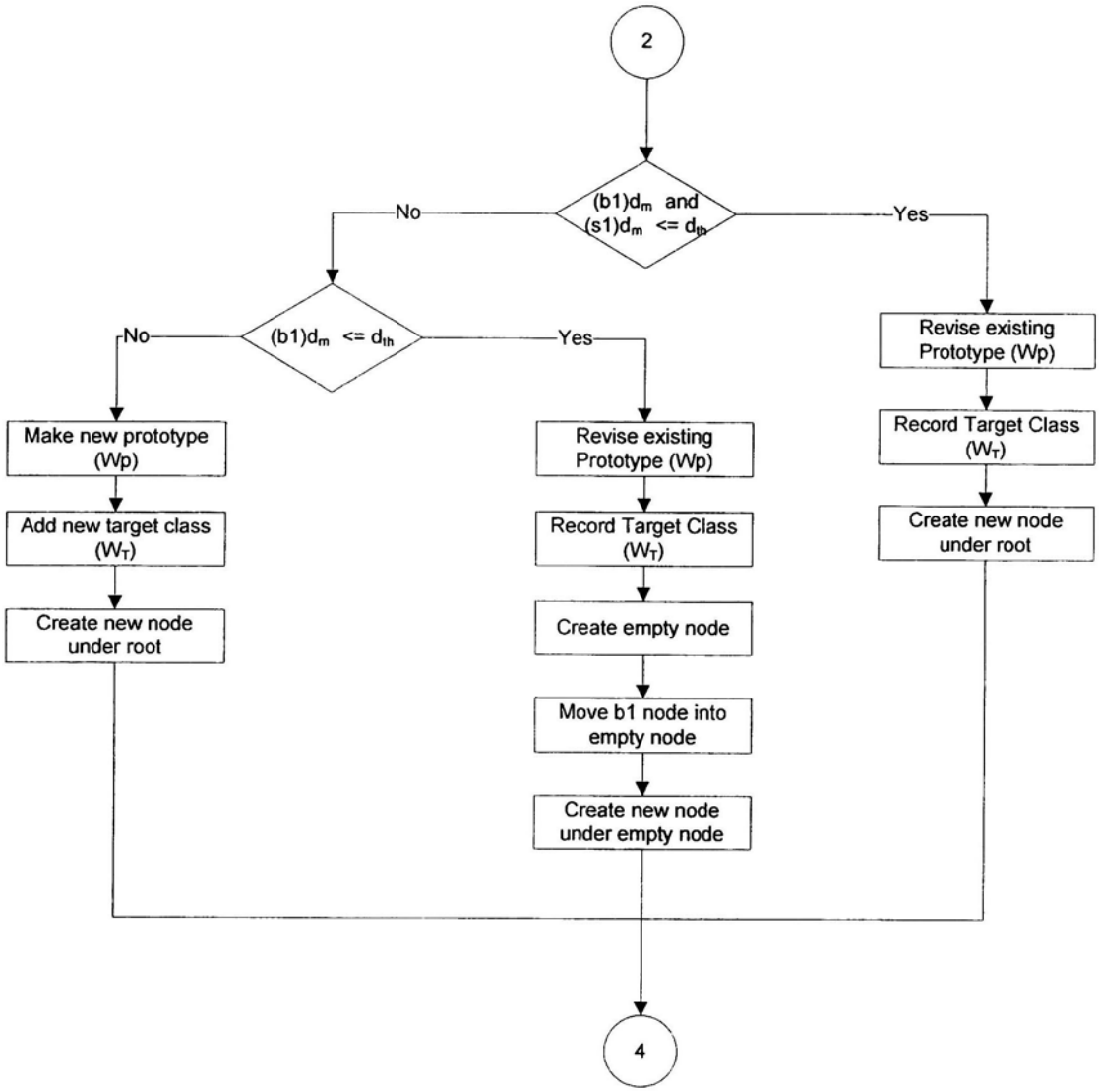
ขั้นตอนที่ 5: กำหนดคลาสเป้าหมายของข้อมูลนำเข้า p โดยการใช้คลาสเป้าหมายของโหนดผู้ชนะ



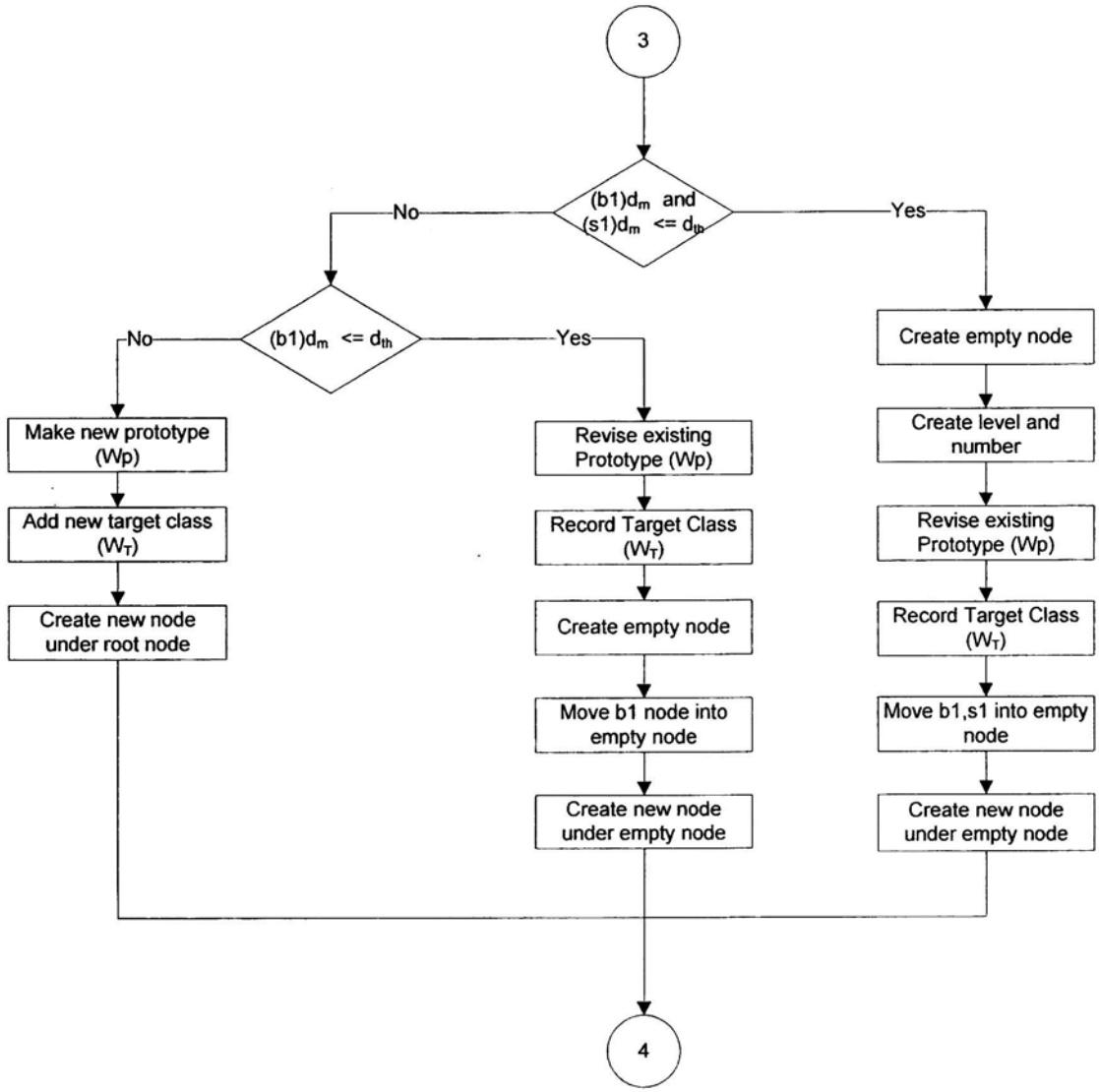
ภาพที่ 3-10 การพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (ส่วนที่1)



ภาพที่ 3-11 การพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (ส่วนที่ 2)



ภาพที่ 3-12 การพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (ส่วนที่ 3)



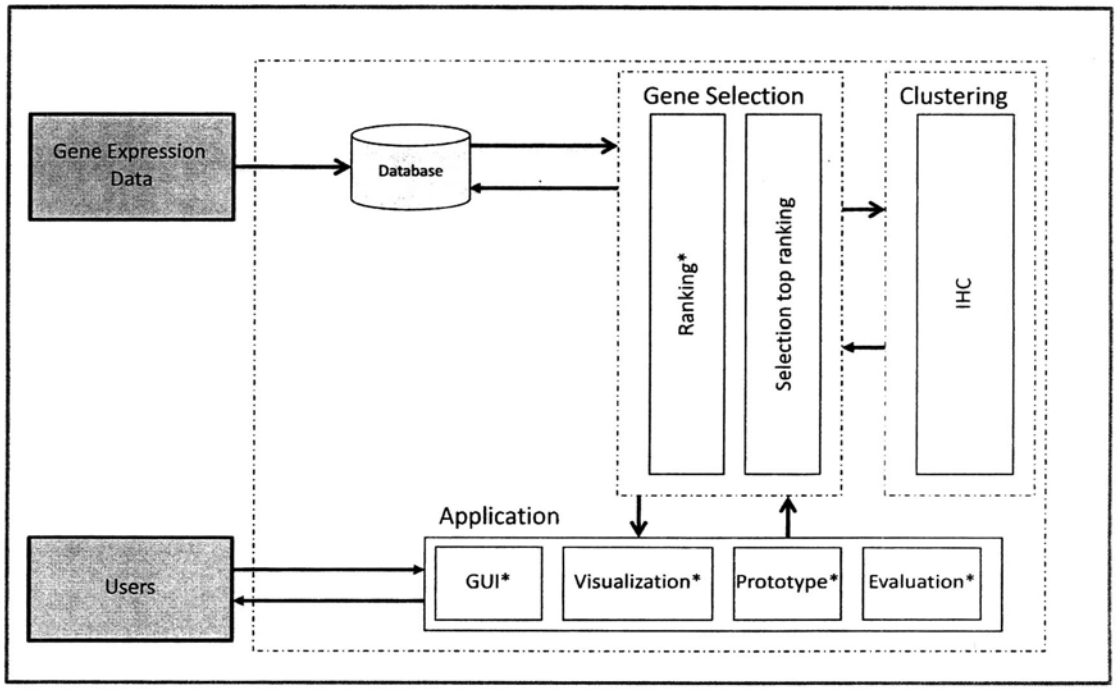
ภาพที่ 3-13 การพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (ส่วนที่ 4)

เมื่อนำกลุ่มย่อยของยีนที่ได้จากของอัลกอริธึมการคัดเลือกมิติบนข้อมูลด้วยวิธีการกรองคือ CfsGR ในขั้นตอน 3.3 มาดำเนินการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นด้วยอัลกอริธึม IHC ซึ่งมีพื้นฐานมาจากการจัดกลุ่มเชิงลำดับชั้นแบบเพิ่มขึ้นของอัลกอริธึม COBWEB ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM ที่อยู่บนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิส และระยะทางที่วัดได้ดังกล่าวจะถูกนำมาเป็นเงื่อนไขสำหรับการสร้างกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

นอกจากนี้อัลกอริธึม IHC เป็นเทคนิคที่มีพื้นฐานการทำงานแบบโครงสร้างลำดับชั้นในการจัดการแบบบนลงล่าง การจัดกลุ่มข้อมูลของอัลกอริธึม IHC จะใช้วิธีการวัดค่าระยะทางระหว่าง

ข้อมูล และนำมาสร้างโมเดลแบบจำลองตามแนวคิดโครงสร้างแบบลำดับชั้น และหลังจากได้กรอบโครงร่างดังกล่าวแล้ว ก็จะนำโครงร่างนี้ไปทำการทดสอบเพื่อหาประสิทธิภาพในการจัดกลุ่มข้อมูล และโครงร่างดังกล่าวก็จะถูกนำมาพัฒนาระบบการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ผลของการทดสอบประสิทธิภาพของโครงร่างที่นำเสนอจะอภิปรายในหัวข้อ 4.1-4.2

จากภาพที่ 3-14 แสดงโครงร่างของระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น โดยโครงร่างนี้จะมีองค์ประกอบที่สำคัญคือ การพัฒนาการติดต่อกับผู้ใช้งาน (Graphic User Interface : GUI) ซึ่งขั้นตอนนี้ผู้ใช้งานจะสามารถทำงานกับระบบผ่านการเชื่อมต่อแบบกราฟิก และในส่วนนี้เช่นกันที่ผู้ใช้งานสามารถที่จะแสดงข้อมูลในรูปแบบเสมือน (visualization) รวมถึงความสามารถในการส่งข้อมูลออกจากระบบเพื่อไปใช้กับโปรแกรมอื่น และออกรายงาน



ภาพที่ 3-14 โครงร่างของระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

ส่วนต่อมาคือ การคัดเลือกมิติบนข้อมูลการแสดงออกของยีน ซึ่งในขั้นตอนนี้จะรวมถึงการติดต่อกับระบบฐานข้อมูล และการเตรียมข้อมูล (Data Preprocessing) ซึ่งประกอบด้วยการแทนค่าข้อมูลที่สูญหาย และการลดการกระจายตัวของข้อมูลด้วยการเปลี่ยนรูปข้อมูล จากนั้นจึงนำข้อมูลดังกล่าวมาคัดเลือกมิติด้วยวิธีการที่ออกแบบเอาไว้ และส่วนสุดท้ายคือการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ซึ่งในขั้นตอนนี้จะเป็นส่วนของการจัดกลุ่มข้อมูลการแสดงออกของยีนที่ผ่านการคัดเลือกมิติแล้ว

3.6 การออกแบบระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

ในขั้นตอนนี้จะเป็นการออกแบบส่วนติดต่อกับผู้ใช้ของระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นเพื่อให้ผู้ใช้งานสามารถใช้งานระบบได้อย่างเต็มประสิทธิภาพ งานวิจัยนี้ได้แบ่งส่วนติดต่อของผู้ใช้งานออกเป็น 4 ส่วนได้แก่ ส่วนติดต่อกับผู้ใช้งานแบบกราฟฟิก ส่วนการแสดงผลแบบเสมือน ส่วนการแสดงแบบจำลอง และส่วนการแสดงประสิทธิภาพ ซึ่งหน้าจอส่วนติดต่อกับผู้ใช้งานดังแสดงในภาพที่ 3-15 ในส่วน GUI นี้จะประกอบด้วยส่วนย่อย ๆ จำนวน 4 ส่วนดังนี้

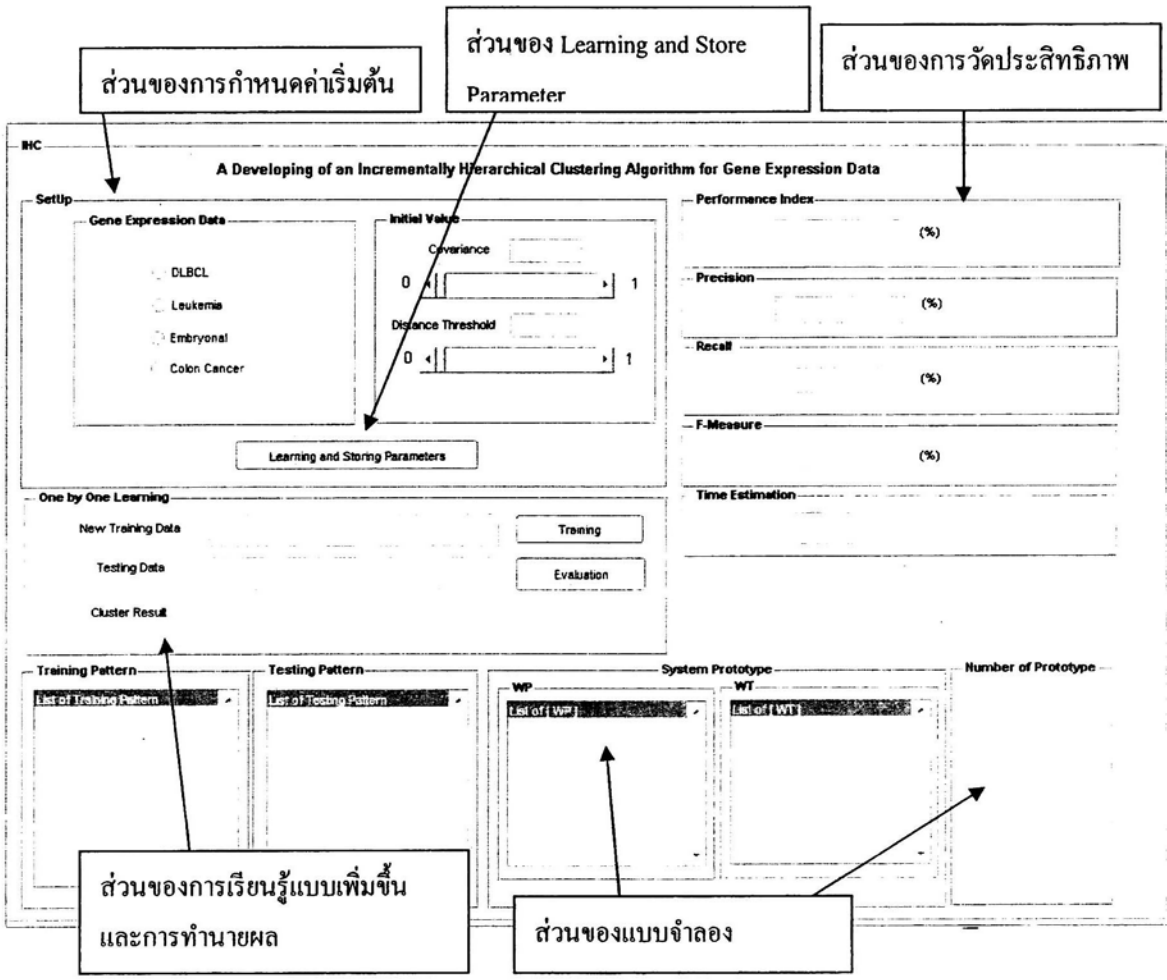
ส่วนของการกำหนดค่าเริ่มต้น (Setup) จะประกอบด้วย 2 ส่วนย่อยได้แก่ ส่วนการเลือกกลุ่มข้อมูลการแสดงออกของยีน และส่วนของการกำหนดค่าความแปรปรวนร่วม และอัตราระยะห่าง ซึ่งจะมีลักษณะเป็นปุ่มกดสำหรับให้เลือกกลุ่มข้อมูลการแสดงออกของยีนสำหรับให้ผู้ใช้งานเลือกใช้ ประกอบด้วย ข้อมูล DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer ผู้ใช้งานสามารถเลือกจาก 1 ใน 4 กลุ่มข้อมูลเพื่อใช้งาน และในส่วนของการกำหนดค่าเริ่มต้นผู้ใช้งานต้องกำหนดค่าความแปรปรวนร่วม และค่าอัตราระยะห่าง ซึ่งอยู่ระหว่างค่า 0 ถึง 1

ส่วนของการ Learning and Store Parameter เป็นส่วนของการประมวลผลอัลกอริธึม IHC โดยเป็นการนำเอาข้อมูลการแสดงออกของยีนกลุ่มหนึ่งกลุ่มใดจากทั้ง 4 กลุ่ม และค่าพารามิเตอร์ที่ได้จากการกำหนดของผู้ใช้งานมาทำการประมวลผล ตามขั้นตอนวิธีของอัลกอริธึม IHC นอกจากนี้ยังทำการประมวลผลเพื่อสร้างกราฟซึ่งแสดงการจัดกลุ่มและกลุ่มย่อยของข้อมูลที่อยู่ในลำดับชั้น

ส่วนของแบบจำลอง System Prototype เป็นส่วนของการแสดงแบบจำลอง (W_p) และคลาสเป้าหมายของแบบจำลอง (W_T) ซึ่งได้จากการเรียนรู้จากข้อมูลฝึกฝน และการเรียนรู้แบบเพิ่มขึ้น (Incremental) นอกจากนี้จำนวนแบบจำลองที่ถูกสร้างจะถูกแสดงในกรอบที่ชื่อ Number of Prototypes ซึ่งจะสามารถทำให้ผู้ใช้งานทราบได้ถึงจำนวนของแบบจำลองที่มีอยู่ในระบบ

ส่วนของการเรียนรู้แบบเพิ่มขึ้นและทำนายผล เป็นส่วนของการเรียนรู้ข้อมูลใหม่เพื่อให้สามารถนำมาปรับโครงสร้างและแบบจำลองที่มีประสิทธิภาพมากยิ่งขึ้น โดยผู้ใช้งานจะต้องใส่ข้อมูลแต่ละแอททริบิวต์ โดยใช้การเว้นวรรค และในแอททริบิวต์สุดท้ายจะแสดงคลาสเป้าหมายของข้อมูลนั้น ส่วนการทำนายผลเพื่อให้สามารถทำนายกลุ่มของข้อมูลทดสอบโดยใช้แบบจำลองในระบบที่มีอยู่ ในส่วนนี้ผู้ใช้งานต้องใส่ข้อมูลแต่ละแอททริบิวต์ โดยใช้การเว้นวรรค แต่ไม่ต้องใส่แอททริบิวต์ของคลาสเป้าหมาย

ส่วนของการวัดประสิทธิภาพ จะเป็นส่วนของการแสดงผลการวัดค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก ค่าความถ่วงดุล และเวลาที่ใช้ในการประมวลผล โดยที่ค่าความถูกต้องจะเป็นการแสดงค่าความถูกต้องของการจัดกลุ่มข้อมูล



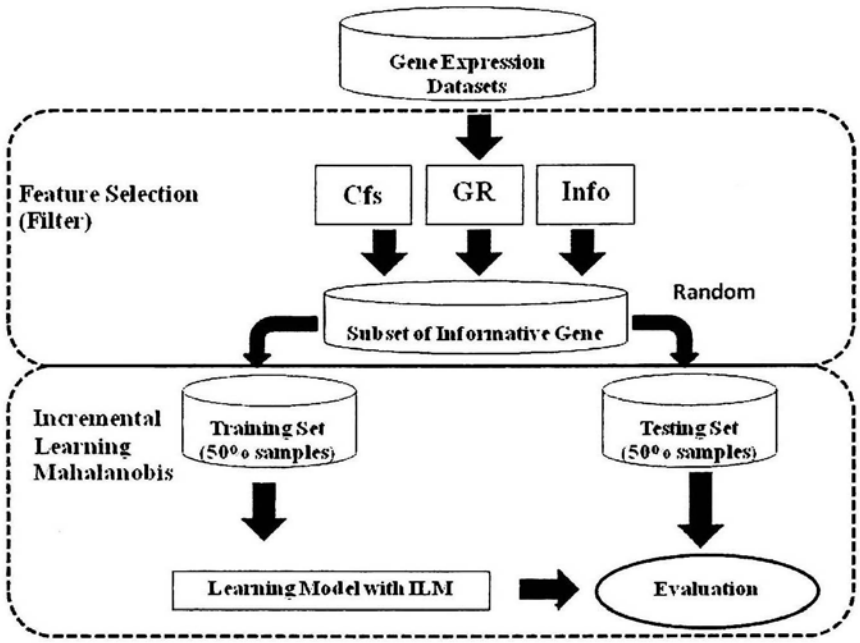
ภาพที่ 3-15 การออกแบบของ GUI ของระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

3.7 การเปรียบเทียบประสิทธิภาพของการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

งานวิจัยนี้ได้ดำเนินการเปรียบเทียบประสิทธิภาพการคัดเลือกมิติข้อมูลที่พัฒนาขึ้นใหม่ (CfsGR) ร่วมกับวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (IHC) ดังที่ได้วางแผนการทดลองในขั้นตอน 3.4 และ 3.5 โดยใช้วิธีการเปรียบเทียบกับผลลัพธ์ที่ได้จากการทดลองด้วยเทคนิคการกรอง คือ Info, Cfs และ GR ร่วมกับเทคนิค ILM ที่มีอยู่เดิม

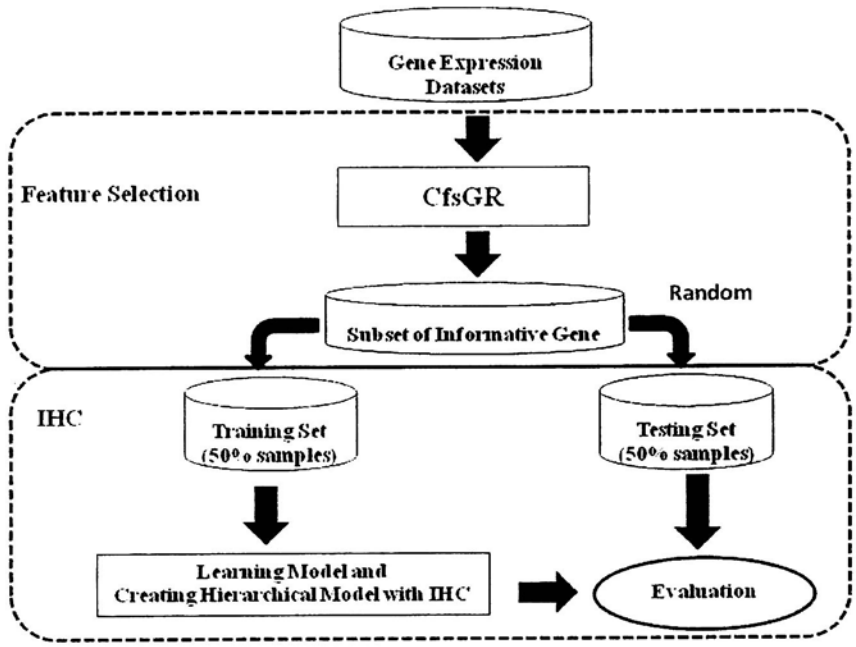
สำหรับการทดลองการคัดเลือกมิติข้อมูลด้วยเทคนิคการกรองคือ Info, Cfs และ GR ร่วมกับอัลกอริธึม ILM ที่มีอยู่เดิม ดำเนินการโดยนำข้อมูลการแสดงออกของยีนจาก 4 กลุ่มข้อมูลคือ DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer มาทำการคัดเลือกมิติข้อมูลที่มีอำนาจจำแนกโดยอาศัยเทคนิคการกรองแต่ละวิธี จากนั้นนำข้อมูลและมิติการแสดงออกของยีนที่ได้จาก

การกรองแต่ละวิธี มาทำการสุ่มเพื่อใช้เป็นข้อมูลฝึกฝน และข้อมูลทดสอบอย่างละเท่า ๆ กัน นำข้อมูลฝึกฝนเข้าสู่กระบวนการเรียนรู้ของของอัลกอริธึม ILM เพื่อสร้างโมเดลแบบจำลองระบบ จากนั้นนำข้อมูลทดสอบเข้าสู่ขั้นตอนการทดสอบประสิทธิภาพเพื่อใช้ทดสอบประสิทธิภาพของระบบ รายละเอียดการทดลองดังแสดงในภาพที่ 3-16



ภาพที่ 3-16 การออกแบบการทดลองของเทคนิคการกรองร่วมกับอัลกอริธึม ILM

สำหรับการทดลองการคัดเลือกมิติข้อมูลด้วยเทคนิค CfsGR ร่วมกับอัลกอริธึทม IHC ดำเนินการโดยนำข้อมูลการแสดงออกของยีนจาก 4 กลุ่มข้อมูลคือ DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer มาทำการคัดเลือกมิติข้อมูลที่มีอำนาจจำแนกโดยใช้เทคนิค CfsGR จากนั้นนำข้อมูลและมิติการแสดงออกของยีนที่ได้จากเทคนิคการคัดเลือกมิติข้อมูลที่นำเสนอ มาทำการสุ่มเพื่อใช้เป็นข้อมูลฝึกฝน และข้อมูลทดสอบอย่างละเท่า ๆ กัน นำข้อมูลฝึกฝนเข้าสู่กระบวนการเรียนรู้ของอัลกอริธึม IHC เพื่อสร้างโมเดลแบบจำลองระบบ พร้อมกับจัดโครงสร้างแบบลำดับชั้นของข้อมูล จากนั้นนำข้อมูลทดสอบเข้าสู่ขั้นตอนการทดสอบประสิทธิภาพเพื่อใช้ทดสอบประสิทธิภาพของระบบ รายละเอียดการทดลองดังแสดงในภาพที่ 3-17



ภาพที่ 3-17 การออกแบบการทดลองของเทคนิค CfsGR ร่วมกับอัลกอริธึม IHC

สำหรับวิธีการเปรียบเทียบประสิทธิภาพของระหว่างวิธีการที่นำเสนอ (CfsGR และ IHC) และเทคนิคที่มีอยู่เดิม (Info, Cfs, GR และ ILM) มีรายละเอียดดังนี้ การวัดค่าความถูกต้อง (Accuracy Rate) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-Measure) และรวมถึงเวลาที่ใช้ในการประมวลผล ดังสมการที่ (3-1) ถึง (3-4) ตามลำดับ

การเปรียบเทียบประสิทธิภาพด้วยการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นบนกลุ่มข้อมูลการแสดงออกของยีนใช้วิธีการคำนวณหาค่าความถูกต้องของการจัดกลุ่มด้วยค่าความแม่นยำ ความระลึก ความถูกต้อง ค่าความถ่วงดุล และเวลาที่ใช้ในการประมวลผล โดยที่ผลลัพธ์ของข้อมูลทดสอบมีอัตราเปอร์เซ็นต์ความถูกต้องสูง แสดงว่าประสิทธิภาพของการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นมีประสิทธิภาพและความเหมาะสมที่จะใช้เป็นโครงร่างในการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นบนกลุ่มข้อมูลการแสดงออกของยีน

$$Accuracy = \frac{(TP + TN)}{N} \tag{3-1}$$

$$P = \frac{TP}{(TP + FP)} \tag{3-2}$$

$$R = \frac{TN}{(TN + FP)} \tag{3-3}$$

$$F - Measure = 2 \times \left[\frac{(P \times R)}{(P + R)} \right] \tag{3-4}$$

โดยที่ P คือ ค่าความแม่นยำ

R คือ ค่าความระลึก

N คือ จำนวนตัวอย่างทดสอบทั้งหมด

TP คือ จำนวนตัวอย่างข้อมูลที่มีคลาสเป็นมะเร็งที่จัดกลุ่มได้ว่าเป็นมะเร็งได้ถูกต้อง

TN คือ จำนวนตัวอย่างข้อมูลที่มีคลาสไม่เป็นมะเร็งที่จัดกลุ่มได้ว่าไม่เป็นมะเร็งได้ถูกต้อง

FP คือ จำนวนตัวอย่างข้อมูลที่มีคลาสไม่เป็นมะเร็งแต่ถูกจัดกลุ่มว่าเป็นมะเร็ง

FN คือ จำนวนตัวอย่างข้อมูลที่มีคลาสเป็นมะเร็งแต่ถูกจัดกลุ่มว่าเป็นมะเร็ง

3.8 การทดสอบประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

การประมาณการค่าความแม่นยำนับได้ว่าเป็นเพียงส่วนหนึ่งของการวัดประสิทธิภาพ ซึ่งในงานวิจัยนี้จะใช้วิธีการ Receiver Operating Characteristic (ROC) เทคนิค ROC เป็นเครื่องมือหนึ่งที่ถูกนำไปใช้ในหลาย ๆ งานวิจัย เช่น Premachandran (2010) ใช้เพื่อประเมินประสิทธิภาพของอัลกอริธึมการตรวจจับจุดบอดสีของจอภาพ (Bad Pixel) และ Ben-Dor (2000) ใช้วิธีนี้ประเมินอัลกอริธึม SVM และบูสติ้ง (Boosting) บนกลุ่มข้อมูล Colon Cancer รวมถึง Dougherty (1998) ใช้ประเมินผลของการตรวจจับโหนด (Edge detector)

เทคนิค ROC ใช้วิธีการวาดกราฟโดยใช้อัตราส่วนตามค่าการทำนายคลาสเป้าหมายที่เป็นบวกจริง (คลาสเป้าหมายที่ถูกกำหนดให้เป็นด้านบวก) หรือ True positive rate ในแกน Y ซึ่งได้จากการคำนวณในสมการที่ 3-5 ในขณะเดียวกันจะวาดกราฟโดยใช้อัตราส่วนของคลาสเป้าหมายที่เป็นลบจริง (คลาสเป้าหมายที่ถูกกำหนดให้เป็นด้านลบ) หรือ False negative rate อยู่บนแกน X ดังคำนวณได้จากสมการที่ 3-6

$$TPR = \frac{TP}{(TP + FN)} \tag{3-5}$$

$$FPR = \frac{FP}{(FP + TN)} \tag{3-6}$$

ในบทนี้ได้กล่าวถึงวิธีการดำเนินการวิจัย ซึ่งเป็นการศึกษาความต้องการของการวิจัย การออกแบบโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น การออกแบบวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอ โครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น การออกแบบระบบ และการวัดประสิทธิภาพของการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

การออกแบบโครงร่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น มีส่วนประกอบที่สำคัญได้แก่ (1) ส่วนของการเตรียมข้อมูล (1.1) ส่วนของการทดแทนค่าสูญหาย และ (1.2) การเปลี่ยนรูปข้อมูล (2) ส่วนของการคัดเลือกมิติข้อมูล และ (3) ส่วนของการพัฒนาขั้นตอนจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

การออกแบบวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอประกอบด้วยขั้นตอนที่สำคัญคือ (1) ขั้นตอนของการคัดเลือกมิติบนข้อมูลการแสดงออกของยีนด้วยการหาความสัมพันธ์ (2) ขั้นตอนการคัดเลือกมิติบนข้อมูลการแสดงออกของยีนด้วยการหายีนที่มีอำนาจจำแนก (3) ขั้นตอนของการนำผลลัพธ์ที่ได้จากวิธีการหาความสัมพันธ์มาจับคู่กับวิธีการหาอำนาจจำแนก และ (4) ขั้นตอนของสร้างโครงร่างการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่อาศัยแนวคิดของการจัดกลุ่มข้อมูลเชิงลำดับชั้นด้วยวิธี COBWEB มาร่วมกับวิธีการจัดกลุ่มข้อมูลแบบเรียนรู้เพิ่มขั้นคือ ILM แบบการเรียนรู้ที่ไม่มีผู้สอนโดยอาศัยการวัดระยะทางแบบมาหาลาโนบิส ซึ่งขั้นตอนที่กล่าวมาในบทนี้นั้นได้ถูกใช้เป็นแนวทางในการพัฒนาระบบการจัดกลุ่มข้อมูลเชิงลำดับชั้นที่ได้นำเสนอ และเมื่อทำการทดสอบประสิทธิภาพในแต่ละขั้นตอนแล้ว ในบทถัดไปจะกล่าวถึงผลของการวิจัยที่ได้จากการทดลองการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

บทที่ 4
ผลของการวิจัย

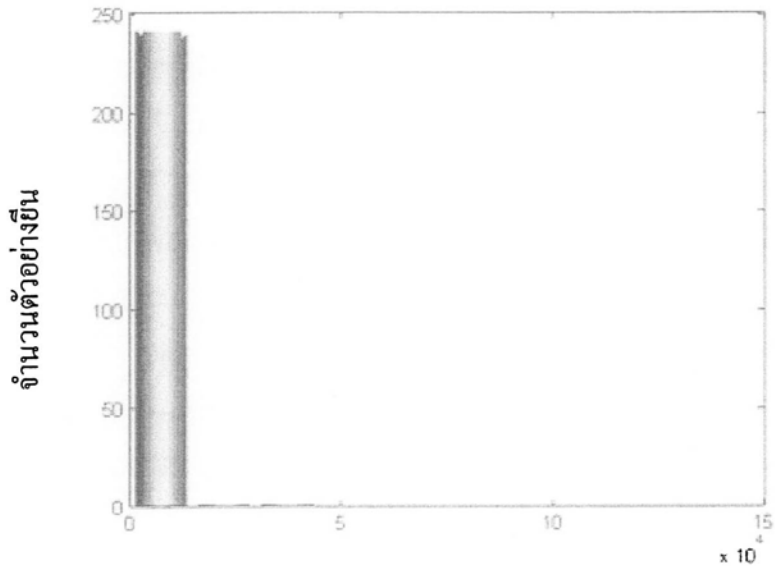
งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อนำเสนอการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น บนกลุ่มข้อมูลการแสดงออกของยีน โดยผู้วิจัยได้ทำการศึกษาถึงคุณลักษณะของข้อมูล เทคนิคการเตรียมข้อมูลที่เหมาะสมกับข้อมูลการแสดงออกของยีน วิธีการคัดเลือกมิติข้อมูลร่วมกับการหายีนที่มีการแสดงออกที่สูง และพัฒนาอัลกอริธึมการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น และในบทนี้จะเป็นการนำเสนอผลการทดลองการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นที่ได้จากวิธีการที่นำเสนอ ซึ่งประกอบด้วยส่วนสำคัญต่าง ๆ ดังกล่าวไว้ในบทที่ 3 สำหรับการเปรียบเทียบประสิทธิภาพระหว่างอัลกอริธึม ILM ในส่วนของการเรียนรู้แบบไม่มีผู้สอนซึ่งเป็นอัลกอริธึมที่มีอยู่เดิมและอัลกอริธึมที่นำเสนอคือ IHC ได้ดำเนินการบนเครื่องคอมพิวเตอร์ AMD Athlon™ Neo X2 ซึ่งมีตัวประมวลผล Dual Core Processor มีแคช L335 ความเร็ว 1.60 GHz และขนาดหน่วยความจำ 2 GB บนระบบปฏิบัติการ Window7 (32 Bit)

สำหรับเนื้อหาในบทนี้ประกอบด้วยผลการวิจัยดังนี้ 1) ผลการทดลองจากขั้นตอนการเตรียมข้อมูล 2) ผลการทดลองจากขั้นตอนการคัดเลือกมิติข้อมูล 3) ผลการเปรียบเทียบประสิทธิภาพระหว่างการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอ กับวิธีการ ILM ในส่วนของการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม 4) ผลการเปรียบเทียบด้านเวลาที่ใช้ในการประมวลผลและประสิทธิภาพโดยรวม 5) ผลการวิเคราะห์ลำดับข้อมูล 6) ผลการทดลองกำหนดค่าเริ่มต้นของอัลกอริธึม IHC มีรายละเอียดดังนี้

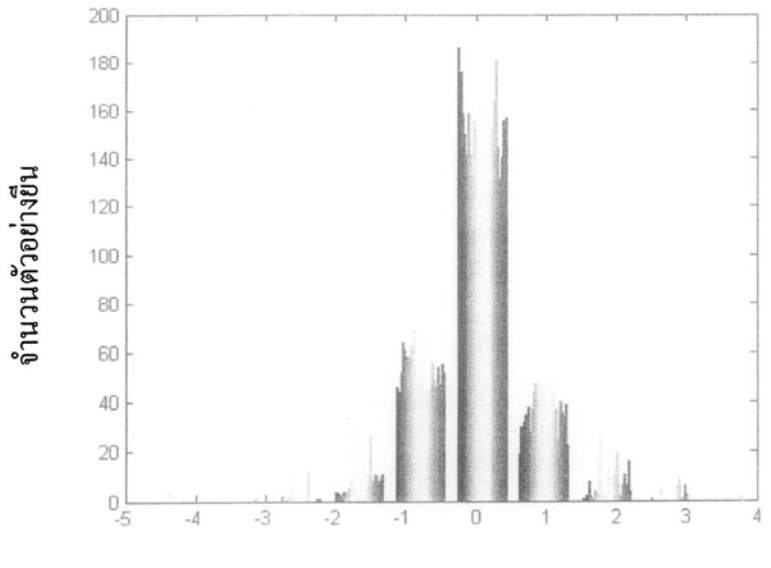
4.1 ผลการทดลองจากขั้นตอนการเตรียมข้อมูล (Data Preprocessing)

เนื่องจากข้อมูลการแสดงออกของยีน ต้องผ่านการทดลองกับไมโครอะเรย์ดีเอ็นเอ ซึ่งเป็นกระบวนการทางชีวภาพ ทำให้อาจมีสิ่งปนเปื้อนหรือสิ่งปลอมปนเข้ามาในระหว่างกระบวนการทดลองซึ่งเป็นเรื่องปรกติ ดังนั้นจากที่ผู้วิจัยได้นำเสนอขั้นตอนการเตรียมข้อมูลไว้ในบทที่ 3 ผู้วิจัยได้ดำเนินการทดลองตามโครงร่างวิธีการเตรียมข้อมูลที่เหมาะสมสำหรับข้อมูลการแสดงออกของยีน ซึ่งแนวทางที่ได้นำเสนอสำหรับการเตรียมข้อมูลคือ การทดแทนค่าสูญหาย และการการเปลี่ยนรูปข้อมูลเพื่อลดการกระจายตัว

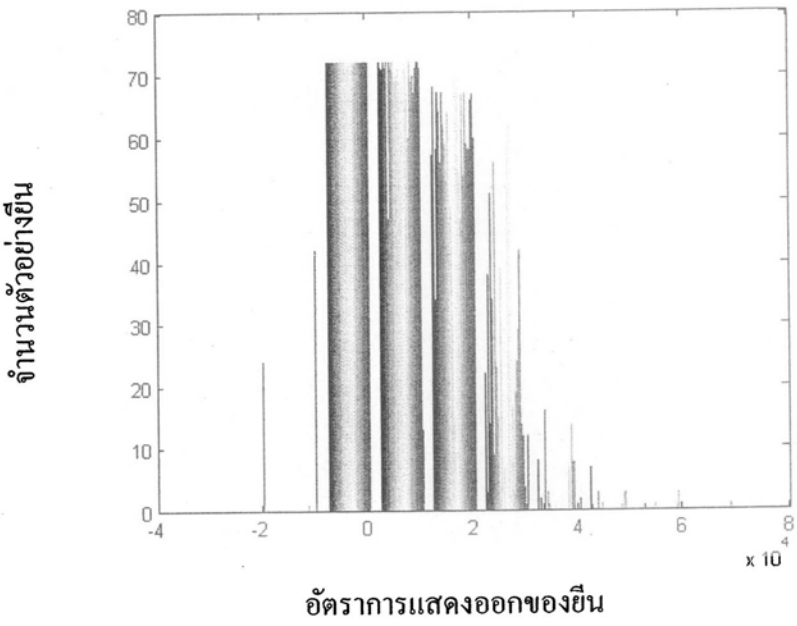
ในขั้นตอนวิธีการเตรียมข้อมูลนี้ ผู้วิจัยได้ใช้วิธีการแทนที่ค่าสูญหายด้วยสมาชิกที่ใกล้ที่สุด (KNNimputation) และการเปลี่ยนรูปข้อมูลโดยใช้ลอการิทึมฐานสองกับกลุ่มข้อมูล DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer ผลการทดลองก่อนและหลังขั้นตอนการเตรียมข้อมูลดังแสดงในภาพที่ 4-1 ถึง 4-8



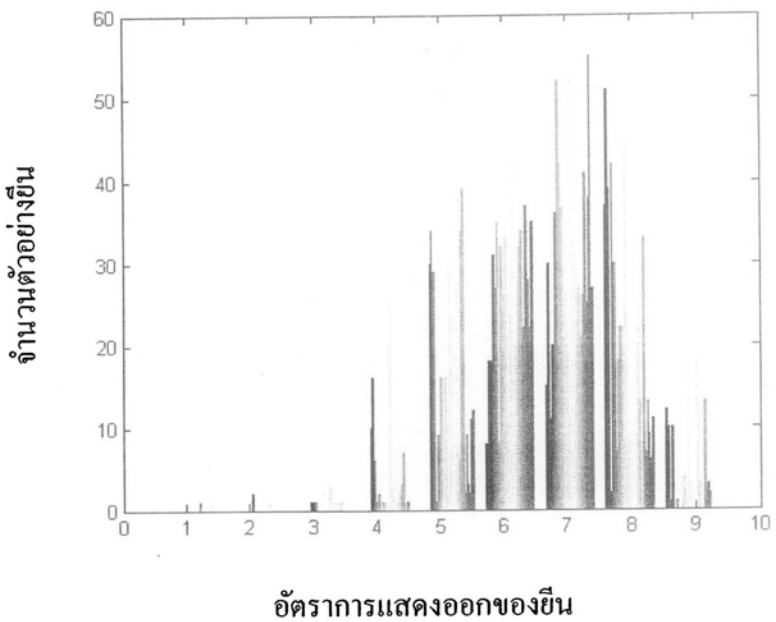
ภาพที่ 4-1 การกระจายตัวของข้อมูล DLBCL ก่อนการแทนค่าสูญหาย และการเปลี่ยนรูปข้อมูล



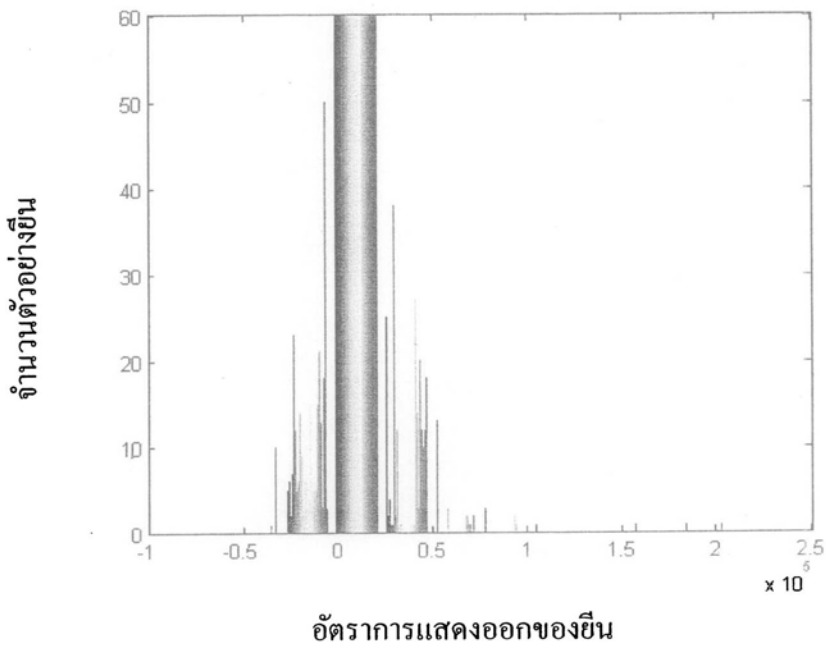
ภาพที่ 4-2 การกระจายตัวของข้อมูล DLBCL หลังการแทนค่าสูญหาย และการเปลี่ยนรูปข้อมูล



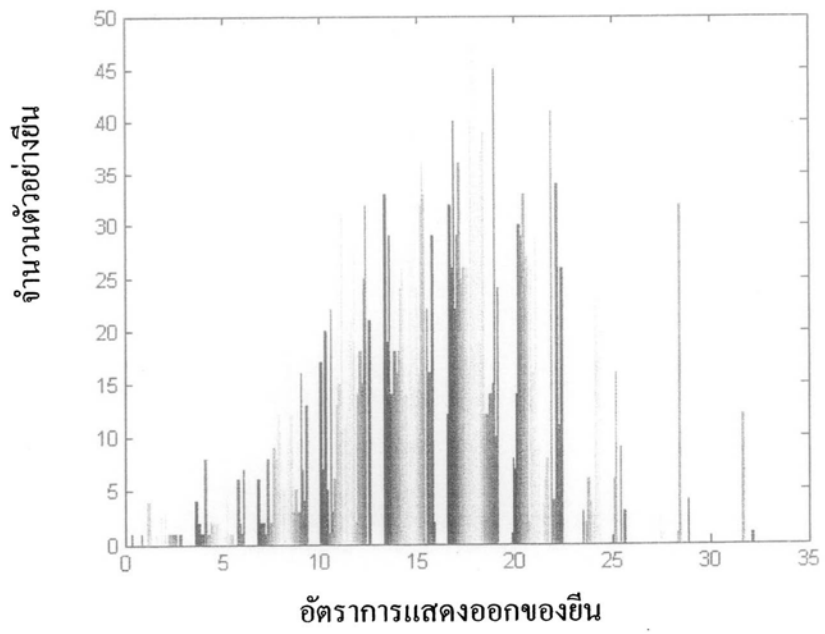
ภาพที่ 4-3 การกระจายตัวของข้อมูล Leukemia ก่อนการเปลี่ยนรูปข้อมูล



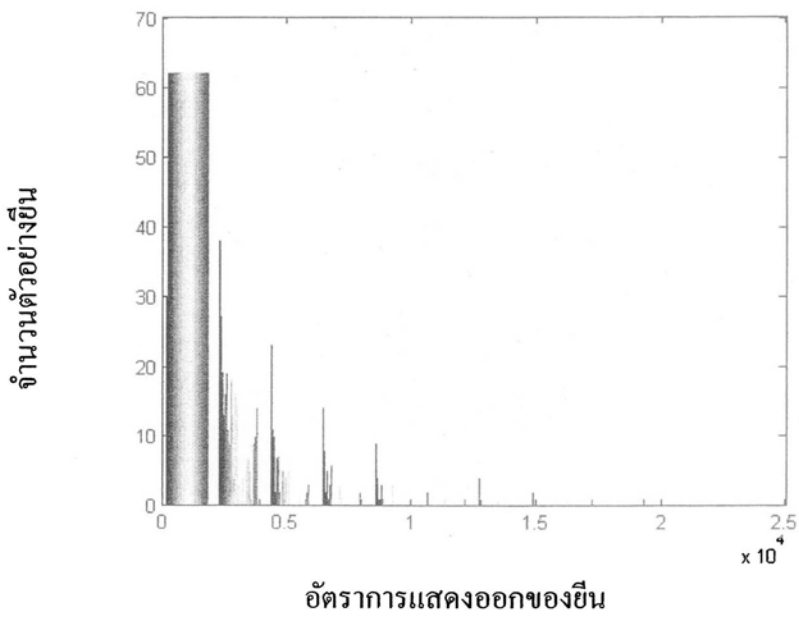
ภาพที่ 4-4 การกระจายตัวของข้อมูล Leukemia หลังการเปลี่ยนรูปข้อมูล



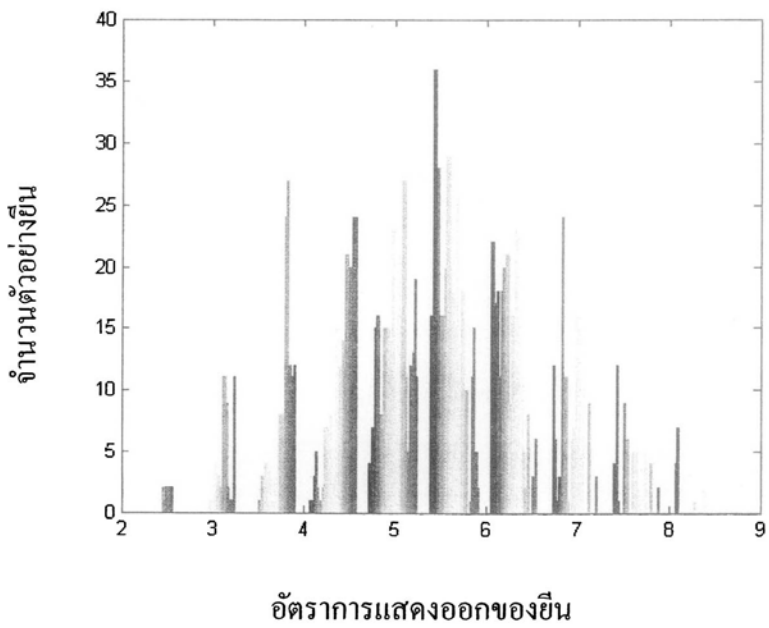
ภาพที่ 4-5 การกระจายตัวของ Embryonal Tumor ก่อนแทนค่าสูญหายและการเปลี่ยนรูปข้อมูล



ภาพที่ 4-6 การกระจายตัวของ Embryonal Tumor หลังแทนค่าสูญหายและการเปลี่ยนรูปข้อมูล



ภาพที่ 4-7 การกระจายตัวของข้อมูล Colon Cancer ก่อนการเปลี่ยนรูปข้อมูล



ภาพที่ 4-8 การกระจายตัวของข้อมูล Colon Cancer หลังการเปลี่ยนรูปข้อมูล

4.2 ผลการทดลองจากขั้นตอนการคัดเลือกมิติข้อมูล (Feature Selection)

เพื่อให้เห็นประสิทธิภาพของการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอใหม่ได้ชัดเจน ผู้วิจัยได้นำเอาเทคนิคการคัดเลือกมิติด้วยวิธีการกรองได้แก่ Cfs, GR และ Info ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM ที่มีอยู่เดิมเข้ามาเพื่อช่วยเปรียบเทียบประสิทธิภาพกับการคัดเลือกมิติข้อมูลที่นำเสนอคือวิธี CfsGR ร่วมกับ อัลกอริธึม IHC ดังแสดงในตารางที่ 4-1 และ 4-2 โดยใช้กลุ่มข้อมูลการแสดงออกของยีนดังนี้ (1) DLBCL (2) Leukemia (3) Embryonal Tumor และ (4) Colon Cancer

ตารางที่ 4-1 ประสิทธิภาพค่าความถูกต้องจากการทำงานร่วมกันของการคัดเลือกมิติข้อมูลด้วยวิธีการกรองร่วมกับวิธีการ ILM ในส่วนของเรียนรู้แบบไม่มีผู้สอน

Data sets	ILM					
	GR		Info		Cfs	
	Number of Attributes	Accuracy	Number of Attributes	Accuracy	Number of Attributes	Accuracy
(1)	37	64.17	58	65.00	37	67.50
(2)	875	85.71	875	88.57	76	**100.00**
(3)	74	80.00	74	73.33	48	3.33
(4)	136	77.42	136	87.10	27	90.00

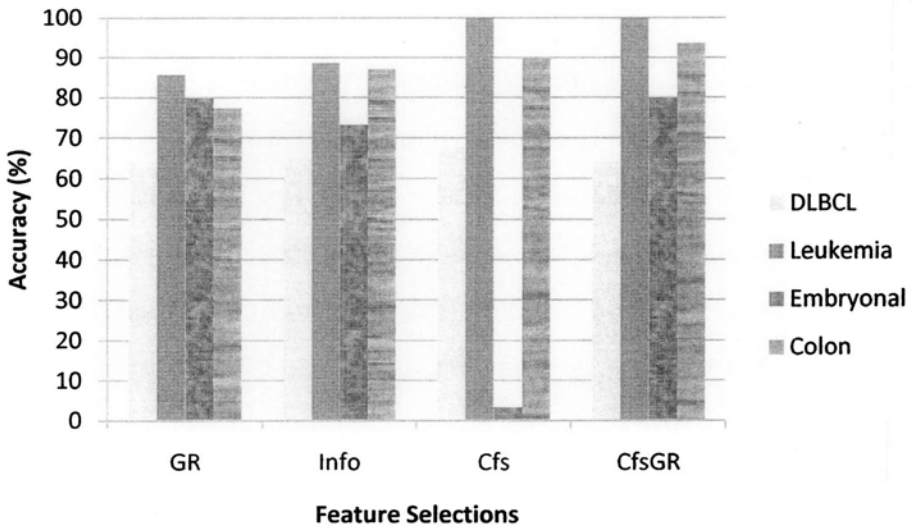
ในตารางที่ 4-1 เป็นการแสดงประสิทธิภาพความถูกต้องจากวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการกรองแบบดั้งเดิมคือ GR, Info และ Cfs ที่ทำงานร่วมกับวีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิสในส่วนของการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม แสดงให้เห็นว่าเมื่อทำการคัดเลือกมิติข้อมูลด้วยวิธีการกรองทั้งสามวิธีดังกล่าว ทำให้จำนวนมิติข้อมูลการแสดงออกของยีนในแต่ละกลุ่มข้อมูลลดลงดังนี้ 1) ข้อมูล DLBCL ลดลงเหลือ 37, 58 และ 37 แอททริบิวต์ ตามลำดับ 2) ข้อมูล Leukemia ลดลงเหลือ 875, 875 และ 76 แอททริบิวต์ตามลำดับ 3) ข้อมูล Embryonal Tumor ลดลงเหลือ 74, 74 และ 48 แอททริบิวต์ตามลำดับ 4) ข้อมูล Colon Cancer ลดลงเหลือ 136, 136, และ 27 แอททริบิวต์ตามลำดับ ในขณะเดียวกันผลลัพธ์จากวิธีการคัดเลือกมิติข้อมูลที่นำเสนอคือ CfsGR ทำให้จำนวนมิติข้อมูลในแต่ละกลุ่มข้อมูลลดลงดังนี้ 1) ข้อมูล DLBCL ลดลงเหลือ 37 ข้อมูล 2) ข้อมูล Leukemia ลดลงเหลือ 39 ข้อมูล 3) ข้อมูล Embryonal Tumor ลดลงเหลือ 48 ข้อมูล และ 4) ข้อมูล Colon Cancer ลดลงเหลือ 27 ข้อมูล

สำหรับผลลัพธ์ในด้านการวัดค่าความถูกต้องในการจัดกลุ่มข้อมูลเมื่อเปรียบเทียบระหว่างวิธีการคัดเลือกมิติข้อมูลคือ GR, Info และ Cfs ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM ให้ผลลัพธ์ค่าความถูกต้องดังนี้ 1) GR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM บนกลุ่มข้อมูล DLBCL ให้ค่าความถูกต้อง 64.17% กลุ่มข้อมูล Leukemia ให้ค่าความถูกต้อง 85.71% กลุ่มข้อมูล Embryonal Tumor ให้ค่าความถูกต้อง 80.00% และกลุ่มข้อมูล Colon Cancer ให้ค่าความถูกต้อง 77.42% 2) Info ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM บนกลุ่มข้อมูล DLBCL ให้ค่าความถูกต้อง 65.00% กลุ่มข้อมูล Leukemia ให้ค่าความถูกต้อง 88.57% กลุ่มข้อมูล Embryonal Tumor ให้ค่าความถูกต้อง 73.33% และกลุ่มข้อมูล Colon Cancer ให้ค่าความถูกต้อง 87.10% 3) Cfs ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM บนกลุ่มข้อมูล DLBCL ให้ค่าความถูกต้อง 67.50% กลุ่มข้อมูล Leukemia ให้ค่าความถูกต้อง 100% กลุ่มข้อมูล Embryonal Tumor ให้ค่าความถูกต้อง 3.33% และกลุ่มข้อมูล Colon Cancer ให้ค่าความถูกต้อง 90%

ตารางที่ 4-2 แสดงการวัดประสิทธิภาพจากค่าความถูกต้องของการทำงานร่วมกันระหว่างการคัดเลือกมิติข้อมูลด้วยเทคนิคที่นำเสนอคือ CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM ผลลัพธ์ที่ได้มีดังนี้ 1) บนกลุ่มข้อมูล DLBCL ให้ค่าความถูกต้อง 64.17% 2) บนกลุ่มข้อมูล Leukemia ให้ค่าความถูกต้อง 100% 3) บนกลุ่มข้อมูล Embryonal Tumor ให้ค่าความถูกต้อง 80% และ 4) บนกลุ่มข้อมูล Colon Cancer ให้ค่าความถูกต้อง 93.55% สำหรับการเปรียบเทียบประสิทธิภาพความถูกต้องระหว่างวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการกรอง คือ GR, Info, Cfs, และวิธีการที่นำเสนอคือ CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM ผลลัพธ์ดังแสดงในภาพที่ 4-9

ตารางที่ 4-2 ประสิทธิภาพค่าความถูกต้องจากการทำงานร่วมกันของคัดเลือกมิติข้อมูลด้วยวิธี CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM

Data sets	**CfsGR-ILM**	
	Number of Attributes	**Accuracy (%)**
(1)	37	64.17
(2)	39	100.00
(3)	48	80.00
(4)	27	93.55



ภาพที่ 4-9 การเปรียบเทียบความถูกต้องระหว่างวิธีการกรองทั้งสามวิธี และวิธีการคัดเลือกมิติ CfsGR ที่นำเสนอร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM

เพื่อเปรียบเทียบประสิทธิภาพระหว่างวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอ กับวิธีการคัดเลือกมิติข้อมูลที่ทำงานร่วมกับวิธีการจำแนกจัดหมวดหมู่ (Classification) ในงานวิจัยอื่นโดยใช้กลุ่มข้อมูลการแสดงออกของยีนที่เหมือนกันทั้ง 4 กลุ่ม คือ DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer ผลลัพธ์จากการเปรียบเทียบค่าความถูกต้องแสดงให้เห็นว่าวิธีการคัดเลือกมิติข้อมูลที่นำเสนอมีค่าความถูกต้องในกลุ่มข้อมูล DLBCL คือ 69.17% กลุ่มข้อมูล Leukemia คือ 100.00% กลุ่มข้อมูล Embryonal Tumor คือ 93.33% และกลุ่มข้อมูล Colon Cancer คือ 93.55% ดังแสดงในตารางที่ 4-3 ถึง 4-6

ตารางที่ 4-3 การเปรียบเทียบประสิทธิภาพค่าความถูกต้องระหว่างวิธีการคัดเลือกมิติข้อมูลที่แตกต่างกับงานวิจัยนี้บนกลุ่มข้อมูล Embryonal Tumor

References	Embryonal Tumor	
	Accuracy (%)	Methods
งานวิจัยนี้	93.33	CfsGR+IHC
Xiong *et al.* (2007)	87.10	SVM
Huang and Liao (2003)	84.00	Signal-to-noise statistic+PNN

ตารางที่ 4-4 การเปรียบเทียบประสิทธิภาพค่าความถูกต้องระหว่างวิธีการคัดเลือกมิติข้อมูลที่แตกต่างกับงานวิจัยนี้บนกลุ่มข้อมูล DLBCL

Reference	DLBCL	
	Accuracy (%)	Methods
งานวิจัยนี้	69.17	CfsGR+IHC
Xu *et al.* (2005)	80	Kaplan-Meier

ตารางที่ 4-5 การเปรียบเทียบประสิทธิภาพค่าความถูกต้องระหว่างวิธีการคัดเลือกมิติข้อมูลที่แตกต่างกับงานวิจัยนี้บนกลุ่มข้อมูล Leukemia

References	Leukemia	
	Accuracy (%)	Methods
งานวิจัยนี้	100.00	CfsGR+IHC
Liu *et al.* (2009)	100.00	SiGWeight+SVM
Leung and Hung (2010)	100.00	MFMW
Zhou *et al.* (2009)	97.1~100	KPCA+BPSO
Xiong et al. (2007)	95.84	DLDA
Huang and Liao (2003)	95.40	Signal-to-noise statistic+PNN
Golub *et al.* (1999)	94.10	SOM

ตารางที่ 4-6 การเปรียบเทียบประสิทธิภาพค่าความถูกต้องระหว่างวิธีการคัดเลือกมิติข้อมูลที่แตกต่างกับงานวิจัยนี้บนกลุ่มข้อมูล Colon Cancer

References	Colon Cancer	
	Accuracy (%)	Methods
งานวิจัยนี้	93.55	CfsGR+IHC
Leung and Hung (2010)	95.16	MFMW
Zhou *et al.* (2009)	93.7~99.7	KPCA+BPSO
Golub *et al.* (1999)	90.30	SOM
Ben-Dor (2000)	88.70	Clustering based Classify

ตารางที่ 4-6 (ต่อ)

References	Colon Cancer	
	Accuracy (%)	Methods
Xiong *et al.* (2007)	84.06	KarNN
Garcia *et al.* (2009)	81.25	DE-SVM

4.3 ผลการเปรียบเทียบประสิทธิภาพระหว่างการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอ กับวิธีการ ILM ในส่วนของการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม

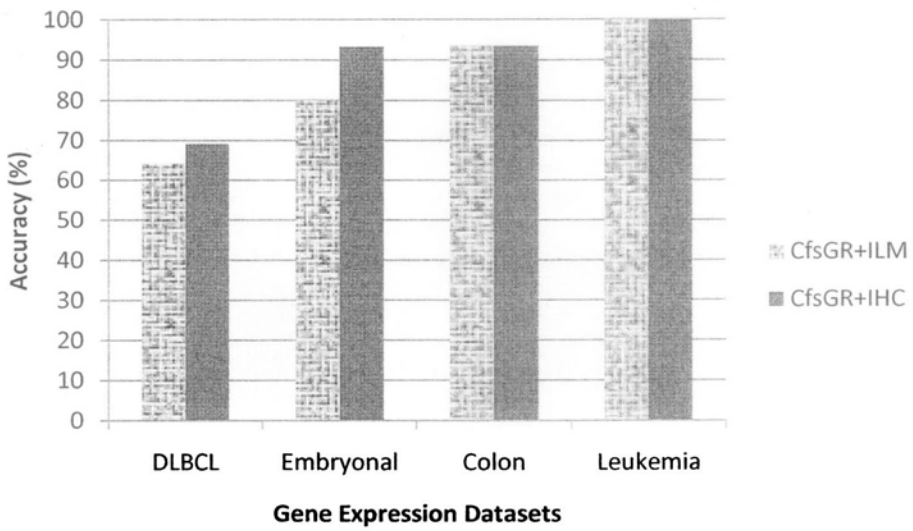
จากผลการทดลองคัดเลือกมิติข้อมูลระหว่างวิธีการกรองคือ GR, Info และ Cfs ร่วมกับวิธีการคัดเลือกมิติข้อมูลที่นำเสนอ CfsGR ดังนั้นในส่วนนี้จะแสดงประสิทธิภาพของขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (IHC) ร่วมกับผลลัพธ์การคัดเลือกมิติจากข้อ 4.2 พร้อมกับการเปรียบเทียบประสิทธิภาพกับวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM ตารางที่ 4-7 แสดงผลลัพธ์ค่าความถูกต้องและจำนวนแอททริบิวต์ที่ได้จากวิธีการคัดเลือกมิติข้อมูลด้วย CfsGR ร่วมกับอัลกอริธึมการจัดกลุ่ม IHC บนกลุ่มข้อมูลการแสดงออกของยีนทั้ง 4 กลุ่ม

ตารางที่ 4-7 ค่าความถูกต้องและจำนวนแอททริบิวต์ที่ได้จากการทำงานร่วมกันของคัดเลือกมิติข้อมูลด้วยวิธี CfsGR ร่วมกับจัดกลุ่มโดยอัลกอริธึม IHC

Data sets	**CfsGR-IHC**	
	Number of Attributes	**Accuracy (%)**
(1)	37	69.17
(2)	39	100.00
(3)	48	93.33
(4)	27	93.55

ตารางที่ 4-7 แสดงการวัดประสิทธิภาพจากค่าความถูกต้องของการทำงานร่วมกันระหว่างการคัดเลือกมิติข้อมูลด้วยเทคนิค CfsGR ร่วมกับอัลกอริธึม IHC ผลลัพธ์ที่ได้มีดังนี้ 1) บนกลุ่มข้อมูล DLBCL วิธีการคัดเลือกมิติข้อมูลที่นำเสนอคัดเลือกแอททริบิวต์จำนวน 37 แอททริบิวต์และมีค่าความถูกต้อง 69.17% 2) บนกลุ่มข้อมูล Leukemia วิธีการคัดเลือกมิติข้อมูลที่นำเสนอคัดเลือกแอททริบิวต์จำนวน 39 แอททริบิวต์มีค่าความถูกต้อง 100% 3) บนกลุ่มข้อมูล Embryonal Tumor

วิธีการคัดเลือกมิติข้อมูลที่นำเสนอคัดเลือกแอททริบิวต์จำนวน 48 แอททริบิวต์มีค่าความถูกต้อง 93.33% และ 4) บนกลุ่มข้อมูล Colon Cancer วิธีการคัดเลือกมิติข้อมูลที่นำเสนอคัดเลือกแอททริบิวต์จำนวน 27 แอททริบิวต์มีค่าความถูกต้อง 93.55% สำหรับการเปรียบเทียบประสิทธิภาพความถูกต้องระหว่างวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการที่นำเสนอคือ CfsGR ร่วมกับอัลกอริธึม ILM และ IHC ดังแสดงในภาพที่ 4-10



ภาพที่ 4-10 การเปรียบเทียบค่าความถูกต้องจากการคัดเลือกมิติข้อมูลด้วย CfsGR ร่วมวิธีการเรียนรู้แบบไม่มีผู้สอนของเทคนิค ILM และอัลกอริธึม IHC

ตารางที่ 4-8 การแสดงประสิทธิภาพค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลจากวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการกรองแบบดั้งเดิมคือ GR, Info และ Cfs ที่ทำงานร่วมวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM ให้ผลดังนี้ 1) วิธีการ GR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM บนกลุ่มข้อมูล DLBCL ให้ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล 65.89% 66.43% และ 64.84% ตามลำดับ กลุ่มข้อมูล Leukemia ให้ค่า 84.87% 87.06% และ 85.95% ตามลำดับ กลุ่มข้อมูล Embryonal Tumor ให้ค่า 70% 35% และ 50% ตามลำดับ กลุ่มข้อมูล Colon Cancer ให้ค่า 87.10% 84.87% และ 87.86% ตามลำดับ 2) วิธีการ Info ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM บนกลุ่มข้อมูล DLBCL ให้ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล 66.67% 67.87% และ67.87% ตามลำดับ กลุ่มข้อมูล Leukemia ให้ค่า 88.40% 89.17% และ 88.78% ตามลำดับ กลุ่มข้อมูล Embryonal Tumor ให้ค่า 70% 35% และ 50% ตามลำดับ กลุ่มข้อมูล Colon Cancer ให้ค่า 83.87% 82.86% และ 81.36% ตามลำดับ 3) วิธีการ Cfs

ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM บนกลุ่มข้อมูล DLBCL ให้ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล 67.50% 66.56% และ 66.11% ตามลำดับ กลุ่มข้อมูล Leukemia ให้ค่า 100% 100% และ 100% ตามลำดับ กลุ่มข้อมูล Embryonal Tumor ให้ค่า 0% 0% และ 0% ตามลำดับ กลุ่มข้อมูล Colon Cancer ให้ค่า 90% 88.89% และ 90.19% ตามลำดับ

ตารางที่ 4-8 ประสิทธิภาพค่าความแม่นยำ (P) ค่าความระลึก (R) และค่าความถ่วงดุล (F) จากการทำงานร่วมกันของคัดเลือกมิติข้อมูลด้วยวิธีการกรองร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM

Data sets	ILM								
	GR			Info			Cfs		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
(1)	65.89	66.43	64.84	66.67	67.87	67.87	67.50	66.56	66.11
(2)	84.87	87.06	85.95	88.40	89.17	88.78	100.00	100.00	100.00
(3)	70.00	35.00	50.00	70.00	35.00	50.00	0	0	0
(4)	87.10	84.87	87.86	83.87	82.86	81.36	90.00	88.89	90.19

ตารางที่ 4-9 ประสิทธิภาพค่าความแม่นยำ (P) ค่าความระลึก (R) และค่าความถ่วงดุล (F) จากการทำงานร่วมกันของคัดเลือกมิติข้อมูลด้วยวิธี CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM

Data sets	CfsGR-ILM		
	Precision (%)	Recall (%)	F-Measure (%)
(1)	65.07	65.26	65.17
(2)	100.00	100.00	100.00
(3)	79.37	76.56	77.93
(4)	91.67	95.24	93.42

ตารางที่ 4-9 แสดงประสิทธิภาพของค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลจากวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการ CfsGR ที่นำเสนอร่วมกับการจัดกลุ่มข้อมูลโดย วิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM ให้ผลดังนี้ 1) บนกลุ่มข้อมูล DLBCL ให้ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล 65.07% 65.26% และ 65.17% ตามลำดับ กลุ่มข้อมูล Leukemia ให้

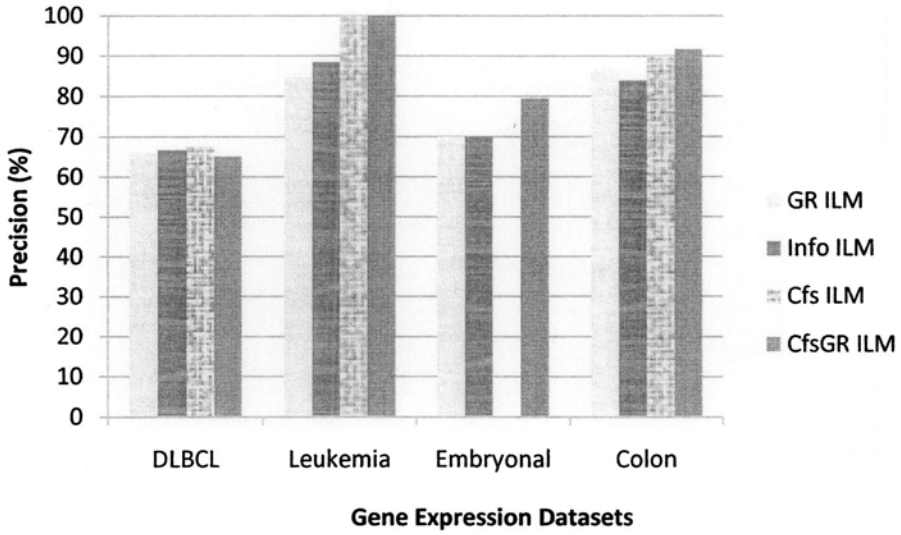
ค่า 100% 100% และ 100% ตามลำดับ กลุ่มข้อมูล Embryonal Tumor ให้ค่า 79.37% 76.56% และ 77.93% ตามลำดับ กลุ่มข้อมูล Colon Cancer ให้ค่า 91.67% 95.24% และ 93.42% ตามลำดับ

ตารางที่ 4-10 ประสิทธิภาพค่าความแม่นยำ (P) ค่าความระลึก (R) และค่าความถ่วงดุล (F) จากการทำงานร่วมกันของคัดเลือกมิติข้อมูลด้วยวิธี CfsGR ร่วมกับอัลกอริธึม IHC

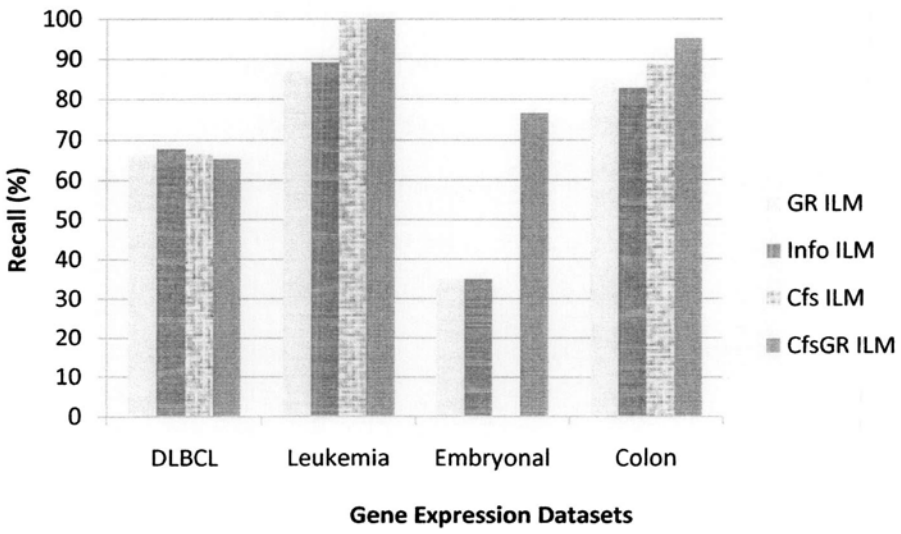
Datasets	CfsGR-IHC		
	Precision (%)	Recall (%)	F-Measure (%)
(1)	69.17	69.61	69.39
(2)	100.00	100.00	100.00
(3)	95.24	90.91	93.02
(4)	95.45	90.91	93.13

ตารางที่ 4-10 แสดงประสิทธิภาพของค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลจากวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการ CfsGR ที่นำเสนอร่วมกับการจัดกลุ่มข้อมูลโดยอัลกอริธึม IHC ให้ผลดังนี้ 1) บนกลุ่มข้อมูล DLBCL ให้ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล 69.17% 69.61% และ 69.36% กลุ่มข้อมูล Leukemia ให้ค่า 100% 100% และ 100% กลุ่มข้อมูล Embryonal Tumor ให้ค่า 95.24% 90.91% และ 93.02% กลุ่มข้อมูล Colon Cancer ให้ค่า 95.45% 90.91% และ 93.13%

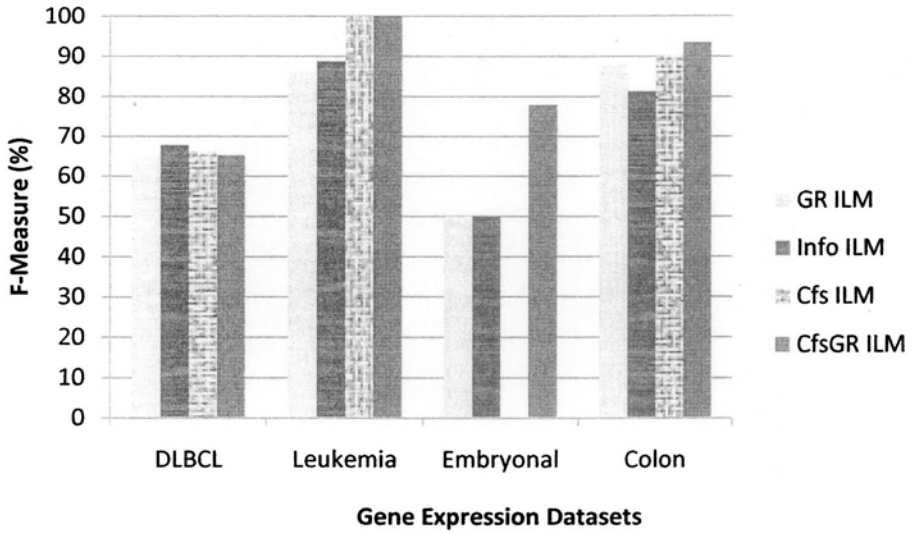
สำหรับการเปรียบเทียบประสิทธิภาพค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ระหว่างวิธีการคัดเลือกมิติข้อมูลด้วยวิธีการกรอง คือ GR, Info, Cfs, และวิธีการคัดเลือกมิติข้อมูลที่นำเสนอคือ CfsGR ร่วมวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM ผลลัพธ์ดังแสดงในภาพที่ 4-11 ถึง 4-13



ภาพที่ 4-11 ผลลัพธ์ค่าความแม่นยำของการทำงานร่วมกันระหว่างวิธีการคัดเลือกมิติข้อมูลและวิธีการเรียนรู้แบบไม่มีผู้สอนของ ILM

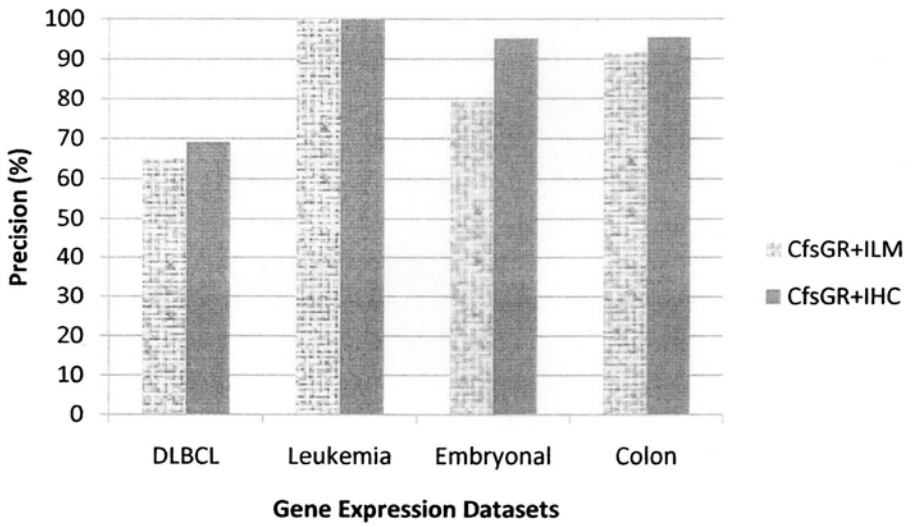


ภาพที่ 4-12 ผลลัพธ์ค่าความระลึกของการทำงานร่วมกันระหว่างวิธีการคัดเลือกมิติข้อมูลและวิธีการเรียนรู้แบบไม่มีผู้สอนของ ILM

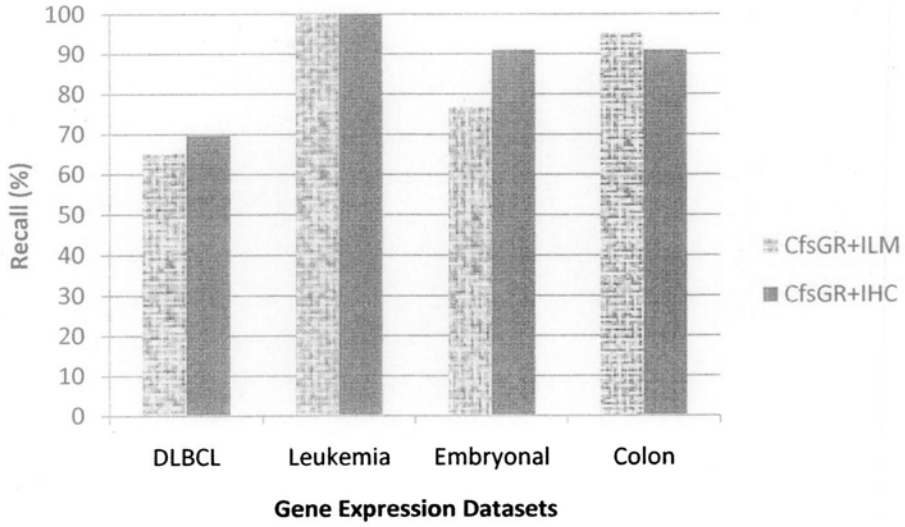


ภาพที่ 4-13 ผลลัพธ์ค่าความถ่วงดุลของการทำงานร่วมกันระหว่างวิธีการคัดเลือกมิติข้อมูลและวิธีการเรียนรู้แบบไม่มีผู้สอนของ ILM

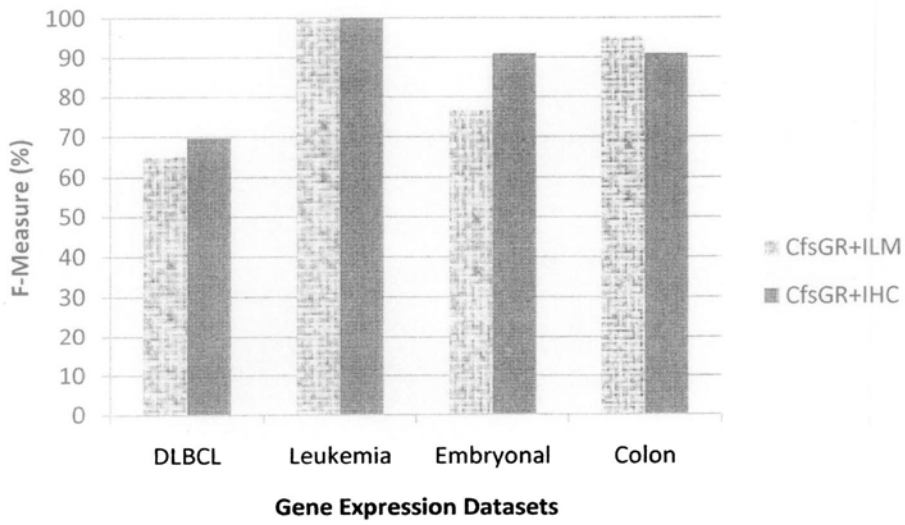
สำหรับการเปรียบเทียบประสิทธิภาพค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลระหว่างวิธีการคัดเลือกมิติข้อมูลด้วยวิธี CfsGR ร่วมกับอัลกอริธึม IHC และวิธี ILM เฉพาะการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ดังแสดงในภาพที่ 4-14 ถึง 4-16



ภาพที่ 4-14 การเปรียบเทียบค่าความแม่นยำ ระหว่างวิธีการ CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM และเทคนิค IHC



ภาพที่ 4-15 การเปรียบเทียบค่าความระลึก ระหว่างวิธีการ CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM และเทคนิค IHC



ภาพที่ 4-16 การเปรียบเทียบค่าความถ่วงดุลระหว่างวิธีการ CfsGR ร่วมกับวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM และเทคนิค IHC

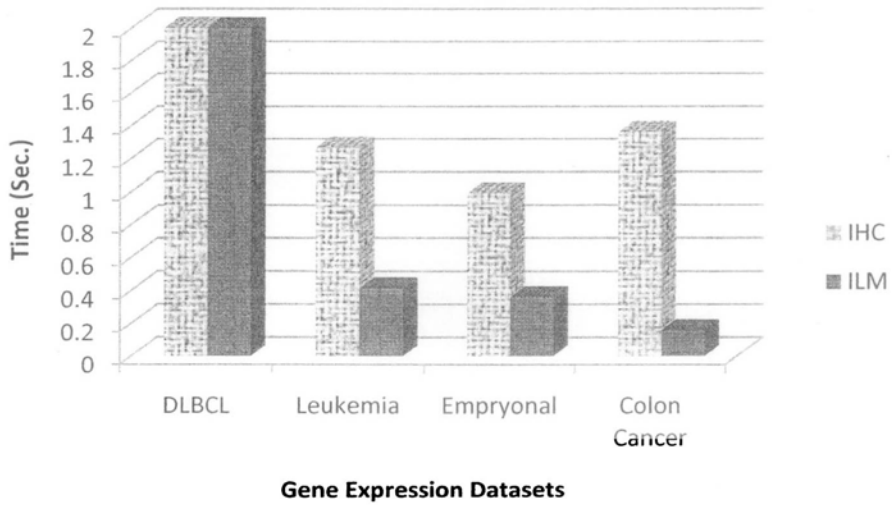
4.4 ผลการเปรียบเทียบประสิทธิภาพด้านเวลาที่ใช้ประมวลผลและประสิทธิภาพโดยรวม

ในส่วนนี้จะเป็นการแสดงผลการทดลองการวัดประสิทธิภาพด้านเวลาที่ใช้ประมวลผลระหว่างวิธีการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น และวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM บนกลุ่มข้อมูลการแสดงของยีน ผลลัพธ์การเปรียบเทียบประสิทธิภาพด้านเวลาที่ใช้ประมวลผลบนกลุ่มข้อมูลการแสดงออกของยีน อัลกอริธึม IHC ใช้ระยะเวลา

ประมวลผลดังนี้ 1) กลุ่มข้อมูล DLBCL 58.4011 วินาที 2) กลุ่มข้อมูล Leukemia 1.2649 วินาที 3) กลุ่มข้อมูล Embryonal Tumor 0.99048 วินาที และ 4) กลุ่มข้อมูล Colon Cancer 1.3593 วินาที และในส่วนของอัลกอริธึม ILM ใช้ระยะเวลาดังนี้ 1) กลุ่มข้อมูล DLBCL 3.6756 วินาที 2) กลุ่มข้อมูล Leukemia 0.41545 วินาที 3) กลุ่มข้อมูล Embryonal Tumor 0.36185 วินาที และ 4) กลุ่มข้อมูล Colon Cancer 0.15972 วินาที ดังแสดงในตารางที่ 4-11 และภาพที่ 4-17

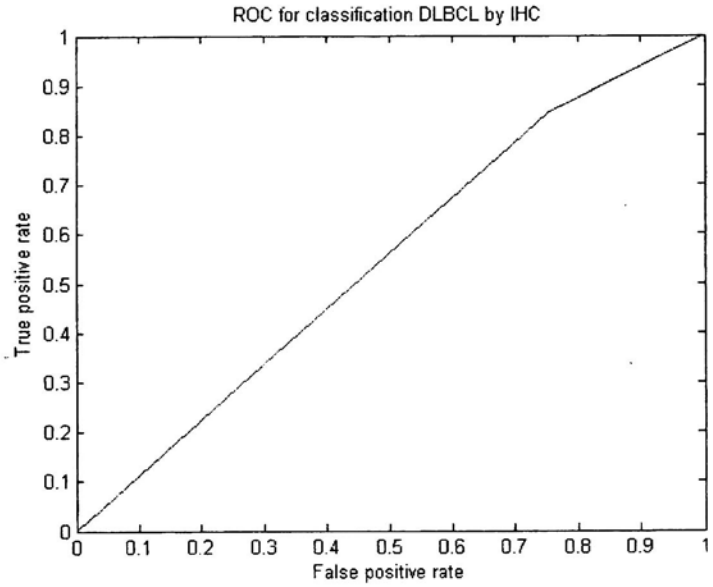
ตารางที่ 4-11 การเปรียบเทียบประสิทธิภาพด้านเวลาระหว่างอัลกอริธึม IHC และวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM อัลกอริธึม (วินาที)

Algorithms	DLBCL	Leukemia	Embryonal	Colon Cancer
IHC	58.4011	1.2649	0.99048	1.3593
ILM	3.6756	0.41545	0.36185	0.15972

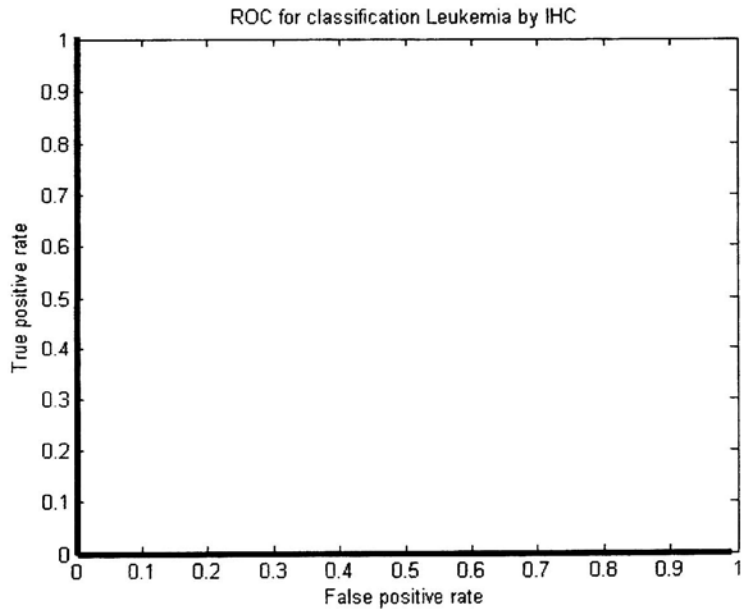


ภาพที่ 4-17 การเปรียบเทียบประสิทธิภาพด้านเวลาที่ใช้ประมวลผลระหว่างวิธีการเรียนรู้แบบไม่มีผู้สอนของอัลกอริธึม ILM และ IHC

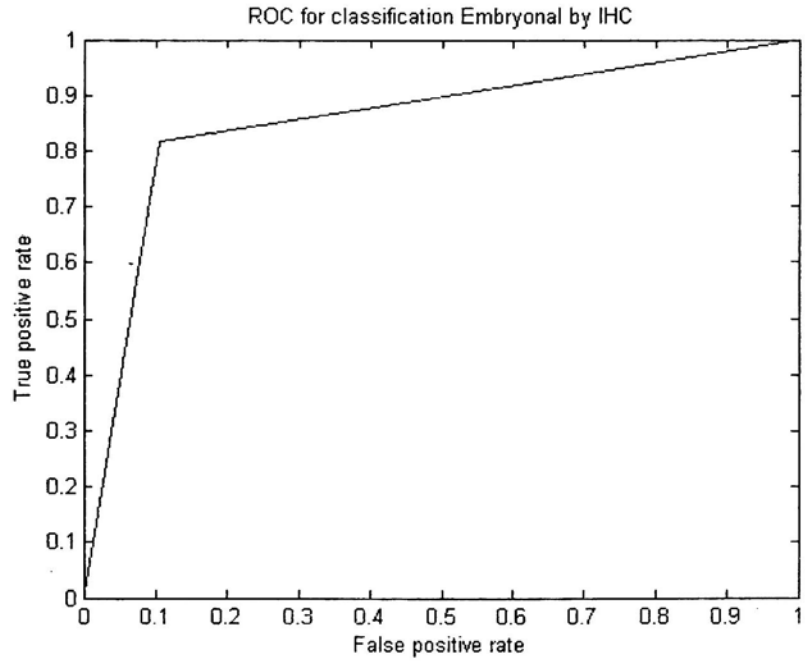
สำหรับการเปรียบเทียบประสิทธิภาพโดยรวมของอัลกอริธึม IHC จะใช้วิธีการวัด Receiver Operating Characteristic โดยแยกตามกลุ่มของข้อมูลทั้ง 4 กลุ่ม ดังแสดงในภาพที่ 4-18 ถึง 4-21



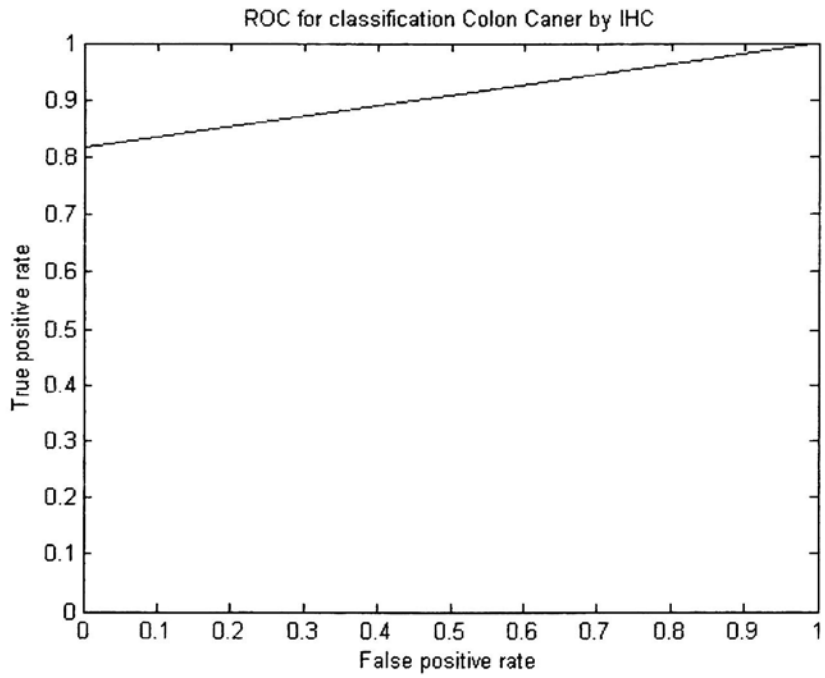
ภาพที่ 4-18 ประสิทธิภาพโดยรวมของอัลกอริธึม IHC บนกลุ่มข้อมูล DLBCL



ภาพที่ 4-19 ประสิทธิภาพโดยรวมของอัลกอริธึม IHC บนกลุ่มข้อมูล Leukemia



ภาพที่ 4-20 ประสิทธิภาพโดยรวมของอัลกอริธึม IHC บนกลุ่มข้อมูล Embryonal Tumor



ภาพที่ 4-21 ประสิทธิภาพโดยรวมของอัลกอริธึม IHC บนกลุ่มข้อมูล Colon Cancer

4.5 ผลการวิเคราะห์ลำดับข้อมูล

เนื่องจากปัญหาโดยทั่วไปที่สำคัญอันหนึ่งของวิธีการเรียนรู้แบบเพิ่มขึ้นคือ ปัญหาลำดับของข้อมูลนำเข้า ดังนั้นงานวิจัยนี้จึงได้ทำการทดลองเปรียบเทียบประสิทธิภาพลำดับของข้อมูลนำเข้าที่เหมาะสมกับอัลกอริธึม IHC โดยใช้ชุดข้อมูลฝึกฝนจากทั้ง 4 กลุ่มข้อมูลการแสดงออกของยีน มีดังนี้

1. ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนแบบปรกติ (Order data)
2. ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนจากหลังไปหน้า (Reverse Data)
3. ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนจากการสุ่มตัวอย่าง (Random Data)

ผลลัพธ์จากการทดลองเปรียบเทียบประสิทธิภาพลำดับของข้อมูลนำเข้าจากทั้ง 3 วิธีที่เหมาะสมกับการทำงานของอัลกอริธึม IHC โดยใช้กลุ่มข้อมูลการแสดงออกของยีน DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer ดังแสดงในตารางที่ 4-12 ถึง 4-15

ตารางที่ 4-12 การเปรียบเทียบประสิทธิภาพลำดับของข้อมูลนำเข้าจากทั้ง 3 วิธี บนกลุ่มข้อมูล DLBCL

Measuring	DLBCL		
	Order	Reverse	Random
Accuracy (%)	60.00	63.33	**69.17**
Precision (%)	64.57	52.13	**69.17**
Recall (%)	68.89	52.22	**69.61**
F-measure (%)	66.66	52.18	**69.39**
Time (Sec.)	205.97	137.53	**58.40**
Number of Prototypes	18.00	33.00	**31.00**

ตารางที่ 4-13 การเปรียบเทียบประสิทธิภาพลำดับของข้อมูลนำเข้าจากทั้ง 3 วิธี บนกลุ่มข้อมูล Leukemia

Measuring	Leukemia		
	Order	Reverse	Random
Accuracy (%)	97.22	100.00	**100.00**
Precision (%)	97.62	100.00	**100.00**

ตารางที่ 4-13 (ต่อ)

Measuring	Leukemia		
	Order	Reverse	Random
Recall (%)	96.88	100.00	**100.00**
F-measure (%)	97.25	100.00	**100.00**
Time (Sec.)	2.23	2.09	**5.53**
Number of Prototypes	6.00	27.00	**2.00**

ตารางที่ 4-14 การเปรียบเทียบประสิทธิภาพลำดับของข้อมูลนำเข้าจากทั้ง 3 วิธี บนกลุ่มข้อมูล Embryonal Tumor

Measuring	Embryonal Tumor		
	Order	Reverse	Random
Accuracy (%)	80.00	80.00	**93.33**
Precision (%)	78.47	78.47	**95.24**
Recall (%)	78.47	78.47	**90.91**
F-measure (%)	78.47	78.47	**93.02**
Time (Sec.)	0.31	1.34	**0.99**
Number of Prototypes	30.00	30.00	**30.00**

ตารางที่ 4-15 การเปรียบเทียบประสิทธิภาพลำดับของข้อมูลนำเข้าจากทั้ง 3 วิธี บนกลุ่มข้อมูล Colon Cancer

Measuring	Colon Cancer		
	Order	Reverse	Random
Accuracy (%)	87.10	80.65	**93.55**
Precision (%)	85.24	79.20	**95.45**
Recall (%)	85.24	83.10	**90.91**
F-measure (%)	85.24	81.10	**93.13**
Time (Sec.)	0.33	1.54	**1.36**
Number of Prototypes	31.00	28.00	**17**

4.6 ผลการทดลองกำหนดค่าเริ่มต้นของอัลกอริธึม IHC

เนื่องจากการค่าพารามิเตอร์เริ่มต้นของอัลกอริธึม IHC ที่ถูกกำหนดโดยผู้ใช้งาน คือ ค่าความแปรปรวนร่วม ($a\mathbf{I}$) และอัตราระยะห่าง (d_{th}) เป็นปัจจัยสำคัญที่ทำให้ประสิทธิภาพของอัลกอริธึม IHC สูงขึ้นขึ้นหรือต่ำลง รวมถึงยังมีผลกับจำนวนของแบบจำลองที่ถูกสร้างขึ้นอีกด้วย

ดังนั้นงานวิจัยนี้จึงได้ทำการทดลองกำหนดค่าพารามิเตอร์เริ่มต้นของอัลกอริธึม IHC ทั้งสองโดยใช้วิธีการสุ่ม ซึ่งผลการทดลองแสดงในตารางที่ 4-16 ถึง 4-19

ตารางที่ 4-16 การเปรียบเทียบค่าพารามิเตอร์เริ่มต้นของอัลกอริธึม IHC บนกลุ่มข้อมูล DLBCL

DLBCL					
covariance	0.096296	0.096296	0.096296	0.014815	**0.014815**
distance	0.16296	0.22222	0.4	0.65926	**1**
Accuracy (%)	58.3333	63.3333	67.5	59.1667	**69.1667**
Precision (%)	58.0089	62.9908	66.6569	59.375	**69.1667**
Recall (%)	58.3333	63.2566	66.1125	59.3567	**69.6078**
F-measure (%)	58.1707	63.1234	66.3836	59.3659	**69.3866**
Time (Sec.)	102.4514	117.6973	147.209	34.6685	**58.4011**
Number of Prototypes	32	10	3	72	**31**

ตารางที่ 4-17 การเปรียบเทียบค่าพารามิเตอร์เริ่มต้นของอัลกอริธึม IHC บนกลุ่มข้อมูล Leukemia

Leukemia					
covariance	0.022222	0.022222	0.022222	**0.13333**	0.8963
distance	0.02963	0.074074	0.28148	**0.28148**	0.014815
Accuracy (%)	100.00	100.00	100.00	**100.00**	100.00
Precision (%)	100.00	100.00	100.00	**100.00**	100.00
Recall (%)	100.00	100.00	100.00	**100.00**	100.00
F-measure (%)	100.00	100.00	100.00	**100.00**	100.00
Time (Sec.)	1.2649	1.5219	2.4774	**5.5296**	4.1744
Number of Prototypes	36	36	29	**2**	15

ตารางที่ 4-18 การเปรียบเทียบค่าพารามิเตอร์เริ่มต้นของอัลกอริธึม IHC บนกลุ่มข้อมูล Embryonal Tumor

Embryonal Tumor					
covariance	**0.037037**	0.37778	0.13333	0.37778	0.37778
distance	**0.022222**	0.037037	0.022222	0.022222	0.54074
Accuracy (%)	**93.33**	86.67	80.00	86.67	86.67
Precision (%)	**95.24**	80.95	86.36	87.30	79.17
Recall (%)	**90.91**	86.34	78.57	83.73	79.17
F-measure (%)	**93.02**	83.56	82.28	85.48	79.17
Time (Sec.)	**0.99048**	1.2116	1.1717	1.1812	1.2569
Number of Prototypes	**30**	30	30	30	28

ตารางที่ 4-19 การเปรียบเทียบค่าพารามิเตอร์เริ่มต้นของอัลกอริธึม IHC บนกลุ่มข้อมูล Colon Cancer

Colon Cancer					
covariance	0.022222	0.95556	**0.081481**	0.7037	0.022222
distance	0.20741	0.022222	**0.081481**	0.022222	0.1037
Accuracy (%)	90.32	90.32	**93.55**	87.10	87.10
Precision (%)	87.50	90.24	**95.45**	81.06	90.91
Recall (%)	93.18	88.41	**90.91**	86.61	84.62
F-measure (%)	90.25	89.31	**93.13**	83.74	87.65
Time (Sec.)	1.2664	5.4567	**1.3593**	4.3375	1.0018
Number of Prototypes	23	3	**17**	5	31

ผลลัพธ์จากการทดลองการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นในบทนี้ประกอบด้วยผลการทดลองที่สำคัญคือ 1) ผลการทดลองจากขั้นตอนการเตรียมข้อมูล 2) ผลของการทดลองจากขั้นตอนการคัดเลือกมิติข้อมูล 3) ผลการเปรียบเทียบประสิทธิภาพระหว่างขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอ กับวิธีการ ILM ในส่วนของการ

เรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม 4) การเปรียบเทียบด้านเวลาที่ใช้ในการประมวลผลและประสิทธิภาพโดยรวม 5) ผลการวิเคราะห์ลำดับข้อมูล และ 6) ผลการทดลองกำหนดค่าเริ่มต้นของอัลกอริธึม IHC ซึ่งผลที่ได้จากผลการทดลองในบทนี้แสดงให้เห็นถึงประสิทธิภาพของวิธีการที่นำเสนอ และจะนำไปสู่การอภิปรายและสรุปผลในบทถัดไป

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

5.1 สรุป

จากปัญหาของวิธีการเรียนรู้แบบเพิ่มขึ้นที่ไม่เหมาะสมกับการประมวลผลข้อมูลการแสดงออกของยีน ที่ไม่เพียงต้องการเฉพาะการเรียนรู้แบบเพิ่มขึ้น แต่ยังรวมถึงความสามารถในการจัดกลุ่ม และแสดงโครงสร้างเชิงลำดับชั้นไปพร้อมกัน ดังนั้นงานวิจัยเรื่องการพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น มีวัตถุประสงค์ในการวิจัย 4 ประเด็นคือ 1) พัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น (IHC) บนกลุ่มข้อมูลการแสดงออกของยีน 2) สร้างต้นแบบใหม่ของอัลกอริธึม IHC ให้สามารถทำงานตามโครงร่างที่พัฒนาขึ้นได้ 3) เปรียบเทียบประสิทธิภาพระหว่างอัลกอริธึม IHC กับวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิสเฉพาะเทคนิคการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม (ILM based on unsupervised learning mode) และ 4) ทดสอบประสิทธิภาพของ IHC ที่พัฒนาขึ้นมาใหม่

กลุ่มข้อมูลที่ใช้ในการศึกษาครั้งนี้คือ กลุ่มข้อมูลการแสดงออกของยีนที่เปิดเผยต่อสาธารณะจำนวน 4 กลุ่มข้อมูล ได้แก่ 1) กลุ่มข้อมูล DLBCL มีจำนวนข้อมูล 240 ข้อมูลและ 7399 แอททริบิวต์ 2) กลุ่มข้อมูล Leukemia มีจำนวนข้อมูล 72 ข้อมูลและ 7129 แอททริบิวต์ 3) กลุ่มข้อมูล Embryonal Tumor มีจำนวนข้อมูล 60 ข้อมูลและ 7129 แอททริบิวต์ และ4) กลุ่มข้อมูล Colon Cancer มีจำนวนข้อมูล 62 ข้อมูล และ 2000 แอททริบิวต์ สำหรับแอททริบิวต์ของกลุ่มข้อมูลที่ใช้ในงานวิจัยนี้แบ่งออกเป็น แอททริบิวต์ข้อมูล และแอททริบิวต์คลาสเป้าหมาย ซึ่งทุกกลุ่มข้อมูลมีคลาสเป้าหมายเท่ากัน คือจำนวน 2 คลาส

กรอบแนวคิดของงานวิจัยนี้ได้มาจากการสังเคราะห์งานวิจัยที่เกี่ยวข้องกับโครงสร้างของดีเอ็นเอและโครงสร้างของยีน จนถึงขั้นตอนการแสดงออกของยีน รวมทั้งศึกษางานวิจัยเกี่ยวข้องกับวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีน และวิธีการจัดกลุ่มข้อมูลของประเทศไทยและต่างประเทศ แบ่งออกได้เป็น ขั้นตอนการเตรียมข้อมูล ขั้นตอนการคัดเลือกมิติข้อมูล ขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

เครื่องมือที่ใช้ในการวิจัยเพื่อพัฒนาขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นแบ่งออกเป็น 2 ส่วน ได้แก่ โมดูลการกรองของโปรแกรมเครื่องจักรการเรียนรู้ (WEKA Version 3.7.1)

ใช้สำหรับขั้นตอนการคัดเลือกมิติข้อมูล และโปรแกรมแมทแลป (Matlab) สำหรับการพัฒนาโปรแกรมการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น

จากผลของการเปรียบเทียบประสิทธิภาพระหว่างขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอ กับวิธีการเรียนรู้แบบเพิ่มขึ้นบนพื้นฐานการวัดระยะทางแบบมาหาลาโนบิสในส่วนของการเรียนรู้แบบไม่มีผู้สอนที่มีอยู่เดิม และการเปรียบเทียบด้านเวลาพร้อมกับการวัดประสิทธิภาพโดยรวม ในบทที่ 4 ผลวิจัยสรุปได้ดังนี้

5.1.1 ผลการวิเคราะห์ข้อมูลเบื้องต้น

ข้อมูลที่ใช้ในงานวิจัยนี้ประกอบด้วย 4 กลุ่มข้อมูลคือ กลุ่มข้อมูล DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer

5.1.1.1 กลุ่มข้อมูล DLBCL

กลุ่มข้อมูล DLBCL มีค่าของข้อมูลอยู่ในช่วงระหว่าง 4.045 ถึง -4.746 และมีลักษณะการกระจายตัวของข้อมูลที่ไม่เป็นปรกติและมีข้อมูลสูญหาย

5.1.1.2 กลุ่มข้อมูล Leukemia

กลุ่มข้อมูล Leukemia มีค่าของข้อมูลอยู่ในช่วงระหว่าง 9.3826 ถึง 0.00 และมีลักษณะการกระจายตัวของข้อมูลที่ไม่เป็นปรกติ แต่ไม่มีข้อมูลสูญหาย

5.1.1.3 กลุ่มข้อมูล Embryonal Tumor

กลุ่มข้อมูล Embryonal Tumor มีค่าของข้อมูลอยู่ในช่วงระหว่าง 32.5116 ถึง 0.00 และมีลักษณะการกระจายตัวของข้อมูลที่ไม่เป็นปรกติ และมีข้อมูลสูญหาย

5.1.1.4 กลุ่มข้อมูล Colon Cancer

กลุ่มข้อมูล Colon Cancer มีค่าของข้อมูลอยู่ในช่วงระหว่าง 8.6717 ถึง 1.9778 และมีลักษณะการกระจายตัวของข้อมูลที่ไม่เป็นปรกติ แต่ไม่มีข้อมูลสูญหาย

5.1.2 สรุปผลของขั้นตอนการจัดเตรียมข้อมูล

จากปัญหาในเรื่องข้อมูลสูญหายที่สามารถพบได้ในกลุ่มข้อมูลการแสดงออกของยีน รวมถึงปัญหาเรื่องการกระจายตัวที่ไม่ปรกติของข้อมูลที่จะส่งผลต่อประสิทธิภาพการทำงานของอัลกอริธึม งานวิจัยนี้ได้ได้ทำการทดลองในส่วนของขั้นตอนการจัดเตรียมข้อมูล โดยใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยวิธีการแทนที่ค่าสูญหายด้วยสมาชิกที่ใกล้ที่สุด (KNNimputation) ซึ่งเป็นวิธีการที่ได้รับการยอมรับและใช้งานอยู่ทั่วไป นอกจากนี้ยังใช้วิธีการลดการกระจายตัวโดยใช้วิธีการเปลี่ยนรูปข้อมูลด้วยลอการิทึมฐานสอง ผลการทดลองดังกล่าว แสดงให้เห็นว่าการแทนที่ค่าข้อมูลสูญหายด้วย KNNimputation และการลดการกระจายตัวโดยการเปลี่ยนรูปข้อมูลบนกลุ่มข้อมูลการแสดงออกของยีนทั้ง 4 กลุ่ม สามารถทำให้กลุ่มข้อมูลการแสดงออกของยีนมีแนวโน้ม

การกระจายตัวที่เป็นปรกติ ซึ่งส่งผลต่อประสิทธิภาพการทำงานของอัลกอริธึม ทั้งในด้านของค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวม

5.1.3 สรุปผลการทดลองการคัดเลือกมิติข้อมูล

ผลการเปรียบเทียบประสิทธิภาพการคัดเลือกมิติข้อมูลระหว่างวิธีการกรอง 3 วิธี ได้แก่ GR, Info และ Cfs กับการคัดเลือกมิติข้อมูลที่นำเสนอคือ CfsGR บนวิธีการจัดกลุ่มแบบ ILM และ IHC แสดงให้เห็นว่าไม่เพียงผลลัพธ์ในการคัดเลือกมิติข้อมูลของอัลกอริธึม GR และ Info มีความใกล้เคียงกันถึง 3 กลุ่มข้อมูล ได้แก่ 1) กลุ่มข้อมูล Leukemia จำนวน 875 มิติ 2) กลุ่มข้อมูล Embryonal Tumor จำนวน 74 มิติ และ 3) กลุ่มข้อมูล Colon Cancer จำนวน 136 มิติ แต่ยังส่งผลถึงค่าความถูกต้องจากการจัดกลุ่มและแยกประเภทของอัลกอริธึม ILM และ IHC ที่มีความใกล้เคียงกันอีกด้วย อย่างไรก็ตามผลลัพธ์การคัดเลือกมิติข้อมูลด้วยการกรอง 2 วิธีแรกก็มีความแตกต่างอย่างเห็นได้ชัดจากผลลัพธ์การคัดเลือกมิติข้อมูลด้วยอัลกอริธึม Cfs ที่ให้จำนวนมิติข้อมูลต่ำกว่า แต่ให้ค่าความถูกต้องส่วนใหญ่สูงกว่า 2 วิธีแรก และในกลุ่มข้อมูล Leukemia วิธีการ Cfs ให้ค่าความถูกต้องถึง 100% อย่างไรก็ตามอัลกอริธึม Cfs กลับให้ผลลัพธ์ค่าความถูกต้องในกลุ่มข้อมูล Embryonal Tumor เพียง 3.33 % ซึ่งต่ำกว่าทุกวิธีการคัดเลือกมิติข้อมูล

ในทางตรงกันข้ามวิธีการคัดเลือกมิติข้อมูลที่นำเสนอคือ CfsGR ให้จำนวนมิติข้อมูลที่ถูกคัดเลือกเป็นจำนวนน้อยทั้ง 4 กลุ่มข้อมูลดังนี้ 1) กลุ่มข้อมูล DLBCL คัดเลือกได้ 37 มิติ 2) กลุ่มข้อมูล Leukemia คัดเลือกได้ 39 มิติ 3) กลุ่มข้อมูล Embryonal Tumor คัดเลือกได้ 48 มิติ และ 4) กลุ่มข้อมูล Colon Cancer คัดเลือกได้ 27 มิติ และวิธีการ CfsGR ยังส่งผลให้ค่าความถูกต้องในการจัดกลุ่มและแยกประเภทด้วยอัลกอริธึม ILM และ IHC มีความถูกต้องสูงกว่า 3 วิธีการกรองคือ GR, Info และ Cfs อย่างชัดเจน

นอกจากนี้ผลลัพธ์การเปรียบเทียบประสิทธิภาพระหว่างวิธีการคัดเลือกมิติข้อมูลการแสดงออกของยีนที่นำเสนอ กับวิธีการคัดเลือกมิติข้อมูลที่ทำงานร่วมกับวิธีการจัดหมวดหมู่ (Classification) ที่แตกต่างกันบนกลุ่มข้อมูลการแสดงออกของยีนทั้ง 4 กลุ่มข้อมูล คือ DLBCL, Leukemia, Embryonal Tumor และ Colon Cancer โดยใช้ค่าความถูกต้อง ผลลัพธ์แสดงให้เห็นว่าประสิทธิภาพในด้านค่าความถูกต้องของวิธีการคัดเลือกมิติข้อมูลที่นำเสนออยู่ในระดับที่สูงในกลุ่มข้อมูล Leukemia, Embryonal Tumor และ Colon Cancer แต่มีประสิทธิภาพในด้านค่าความถูกต้องอยู่ในระดับปานกลางสำหรับการคัดเลือกมิติข้อมูลบนกลุ่มข้อมูล DLBCL

5.1.4 สรุปผลการเปรียบเทียบประสิทธิภาพระหว่างการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นที่นำเสนอ กับอัลกอริธึม ILM ในส่วนของการเรียนรู้แบบไม่มีผู้สอน

การวัดประสิทธิภาพของขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นหรือ IHC ทำโดยการเปรียบเทียบกับผลลัพธ์ของอัลกอริธึมการจัดกลุ่มแบบ ILM ซึ่งแสดงให้เห็นว่าประสิทธิภาพค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลของวิธีการจัดกลุ่มแบบ ILM ที่ผ่านวิธีการกรองคือ GR และ Info ให้ค่าประสิทธิภาพในระดับใกล้เคียงกัน แตกต่างจากการทำงานร่วมกันของวิธีการกรองแบบ Cfs ที่ร่วมกับวิธีการจัดกลุ่มแบบ ILM ที่ให้ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลอยู่ในระดับสูงเช่นในกลุ่มข้อมูล Leukemia แต่กลับให้ค่า 0 เมื่อทำงานกับกลุ่มข้อมูล Embryonal Tumor และเมื่อนำการคัดเลือกมิติข้อมูลที่นำเสนอคือ CfsGR มาทำงานร่วมกับอัลกอริธึม ILM พบว่าค่าผลลัพธ์ความแม่นยำ ความระลึก และค่าความถ่วงดุล สูงกว่าวิธีการทำงานร่วมกันระหว่างเทคนิคการกรอง 3 วิธีและอัลกอริธึม ILM นอกจากนี้เมื่อนำเอาเทคนิค CfsGR มาทำงานร่วมกับวิธี IHC ส่งผลลัพธ์ให้ค่าการวัดประสิทธิภาพสูงที่สุดจากทุก ๆ วิธีการที่ทำร่วมกันคือ GRILM, InfoILM, CfsILM และ CfsGRILM

5.1.5 สรุปผลการเปรียบเทียบด้านเวลาที่ใช้ประมวลผล และประสิทธิภาพโดยรวม

เพื่อแสดงประสิทธิภาพการทำงานที่แท้จริงระหว่างวิธีการ ILM เฉพาะการเรียนรู้แบบไม่มีผู้สอนและอัลกอริธึม IHC ดังนั้นทั้งสองวิธีการจึงถูกนำมาเปรียบเทียบประสิทธิภาพด้านเวลาที่ใช้ประมวลผล จากผลลัพธ์ที่ได้แสดงจะเห็นว่าอัลกอริธึม ILM ใช้เวลาการประมวลผลทั้งหมดน้อยกว่าอัลกอริธึม IHC จากทั้ง 4 กลุ่มข้อมูล โดยที่ในกลุ่มข้อมูล Colon Cancer เทคนิค ILM ใช้เวลาในการประมวลผลน้อยที่สุดคือ 0.15972 วินาทีและในกลุ่มข้อมูล DLBCL ใช้เวลาในการประมวลผลมากที่สุดคือ 3.6756 วินาที ซึ่งแตกต่างจากอัลกอริธึม IHC ที่ใช้เวลาในการประมวลผลทั้งหมดสูงกว่าวิธีการ ILM โดยเฉพาะในกลุ่มข้อมูล DLBCL ใช้เวลาในการประมวลผลถึง 58.4011 วินาที และใช้เวลาต่ำสุด คือ 0.99048 วินาที ในการประมวลผลกับกลุ่มข้อมูล Embryonal Tumor

นอกจากนี้เพื่อแสดงประสิทธิภาพของอัลกอริธึม IHC จึงอาศัยวิธีการวัด Receiver Operating Characteristic (ROC) มาแสดงผลโดยให้คลาส positive คือค่าที่เป็นโรค จะเห็นได้ว่าประสิทธิภาพของอัลกอริธึม IHC บนกลุ่มข้อมูล DLBCL อยู่ในระดับต่ำที่สุดของ 4 กลุ่มข้อมูลและให้ประสิทธิภาพที่สูงที่สุดในกลุ่มข้อมูล Leukemia รองลงมาเป็นกลุ่มข้อมูล Colon Cancer และ Embryonal Tumor ตามลำดับ

5.1.6 ผลการวิเคราะห์ลำดับข้อมูล

จากการค้นหาลำดับข้อมูลนำเข้าที่เหมาะสมกับอัลกอริธึม IHC บนกลุ่มข้อมูลการแสดงออกของข้อมูลยีนทั้ง 4 กลุ่ม โดยทดลองใช้วิธีการลำดับข้อมูล 3 วิธีคือ 1) ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนแบบปรกติ (Order data) 2) ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนจากหลังไปหน้า

(Reverse Data) และ 3) ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนจากการสุ่มตัวอย่าง (Random Data) ผลลัพธ์แสดงให้เห็นถึงประสิทธิภาพจากค่าความถูกต้องที่ได้รับมีค่าสูงขึ้น เมื่อใช้ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนจากการสุ่มตัวอย่างในกลุ่มข้อมูล DLBCL, Leukemia และ Embryonal Tumor ซึ่งแตกต่างจากกลุ่มข้อมูล Colon Cancer ที่ประสิทธิภาพค่าความถูกต้องจากวิธีลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนแบบปรกติ และวิธีลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนจากการสุ่มตัวอย่างที่ให้ค่าความถูกต้องที่เท่ากัน

5.1.7 ผลการทดลองกำหนดค่าเริ่มต้นของอัลกอริธึม IHC

เนื่องจากประสิทธิภาพของอัลกอริธึม IHC ขึ้นอยู่กับการกำหนดค่าเริ่มต้นของพารามิเตอร์ค่าความแปรปรวนร่วม (**aI**) และอัตราระยะห่าง (d_{th}) ดังนั้นการค้นหาค่าเริ่มต้นจึงนับเป็นขั้นตอนที่สำคัญอันหนึ่งสำหรับการกำหนดค่าที่เหมาะสมให้กับอัลกอริธึม

ผลลัพธ์แสดงให้เห็นว่าการกำหนดค่าเริ่มต้นที่ให้ค่าความถูกต้องที่สูงสำหรับอัลกอริธึม IHC บนกลุ่มข้อมูลการแสดงออกของยีนทั้ง 4 กลุ่ม คือ 1) กลุ่มข้อมูล DLBCL เมื่อค่าความแปรปรวนร่วมเท่ากับ 0.014815 และอัตราระยะห่างที่ถูกกำหนดไว้คือ 1 ให้ค่าความถูกต้อง 69.17% 2) กลุ่มข้อมูล Leukemia เมื่อค่าความแปรปรวนร่วมเท่ากับ 0.13333 และอัตราระยะห่างที่ถูกกำหนดไว้คือ 0.28148 ให้ค่าความถูกต้อง 100.00% 3) กลุ่มข้อมูล Embryonal Tumor เมื่อค่าความแปรปรวนร่วมเท่ากับ 0.037037 และอัตราระยะห่างที่ถูกกำหนดไว้คือ 0.02222 ให้ค่าความถูกต้อง 93.33% และ 4) กลุ่มข้อมูล Colon Cancer เมื่อค่าความแปรปรวนร่วมเท่ากับ 0.081481 และอัตราระยะห่างที่ถูกกำหนดไว้คือ 0.081481 ให้ค่าความถูกต้อง 93.55%

5.2 การอภิปรายผลการวิจัย

จากการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น หรือ IHC ที่นำเสนอไว้นี้ มีข้อสังเกตของชุดข้อมูลการแสดงออกของยีน เนื่องจากการได้มาซึ่งข้อมูลการแสดงออกของยีนต้องผ่านกระบวนการทดลองทางชีวภาพ รวมถึงพฤติกรรมของยีนเองที่มีความสัมพันธ์ระหว่างกัน ดังนั้นจึงทำให้กระบวนการเตรียมข้อมูลส่งผลต่อการทำงานของอัลกอริธึมเป็นอย่างมาก รวมถึงการจัดกลุ่มข้อมูลที่อาศัยเพียงการวัดระยะห่างเพียงอย่างเดียวอาจไม่เพียงพอตามรายละเอียดดังนี้

5.2.1 จากการศึกษากลุ่มข้อมูลการแสดงออกของยีนที่เปิดเผยต่อสาธารณะจำนวน 4 กลุ่มข้อมูลพบว่าข้อมูลการแสดงออกของยีนมีลักษณะจำเพาะและมีคุณสมบัติเฉพาะตัว ได้แก่ 1) สิ่งปนเปื้อนกับข้อมูลที่ได้จากการทดลองไมโครอะเรย์ ซึ่งนับเป็นธรรมชาติของข้อมูลที่ทำการทดลองผ่านห้องทดลองทางชีวภาพ การที่ต้องยุ่งเกี่ยวกับหน่วยทดลองที่เล็กและจำนวนมาก หรือที่เรียกกันว่าอณูพันธุศาสตร์ (Molecular Genetic) ทำให้มีโอกาสของการปนเปื้อนของสิ่งรบกวนต่าง ๆ ดังนั้นการขจัดสิ่งรบกวนก็เป็นขั้นตอนที่จำเป็นสำหรับการเตรียมข้อมูลก่อนการ

ใช้งาน 2) จากการศึกษาทฤษฎีที่เกี่ยวข้องพบว่าข้อมูลการแสดงออกของยีน เป็นข้อมูลที่มีลักษณะพฤติกรรมที่สัมพันธ์กัน สิ่งนี้ส่งผลเป็นอย่างมากกับการคัดเลือกมิติของข้อมูลยีน และซึ่งทำให้ข้อมูลการแสดงออกของยีนแตกต่างจากข้อมูลอื่น ๆ 3) ผลการทดลองแสดงให้เห็นว่ากลุ่มข้อมูลของยีนโดยส่วนใหญ่แล้วไม่มีอำนาจจำแนก ซึ่งเป็นไปตามทฤษฎีต่าง ๆ ที่ได้กล่าวไว้ ดังนั้นไม่ว่าเทคนิคการแยกประเภท หรือเทคนิคการจัดกลุ่มจะมีประสิทธิภาพในการทำงานเพียงใด ก็อาจจะไม่ประสบผลสำเร็จในการทำงานกับข้อมูลการแสดงออกของยีน ถ้าไม่ตระหนักถึงคุณลักษณะข้อนี้ ผลที่ได้จากการที่ใช้ยีนที่ไม่มีอำนาจจำแนก อาจก่อให้เกิดการทำนายที่ผิดพลาด หรือการจัดกลุ่มข้อมูลที่ไม่ถูกต้องได้

5.2.2 ในการทดลองการแสดงออกของยีนจำนวนนับพัน ๆ ยีนในเวลาเดียวกัน อาจเกิดข้อผิดพลาดทำให้ได้ข้อมูลไม่ครบถ้วนซึ่งเป็นการหลีกเลี่ยงได้ยากในการปฏิบัติการทดลองวิจัย ปัญหาข้อมูลที่สูญหายนับเป็นปัญหาหลักที่จะพบได้เป็นประจำ ในกลุ่มข้อมูลการแสดงออกของยีน อย่างไรก็ตามวิธีการทดแทนค่าสูญหายทางสถิติที่ใช้กันโดยทั่วไปอาทิ การแทนค่าสูญหายด้วยค่าเฉลี่ย หรือค่ามัธยฐานอาจใช้ไม่ได้ผล เนื่องจากข้อมูลของยีนมีพฤติกรรมที่มีความสัมพันธ์กัน ดังนั้นนักวิจัยควรจะใช้การแทนค่าเฉพาะที่ (Local Estimate) จะให้โอกาสการแทนค่าที่เหมาะสมมากกว่าการใช้ค่าเฉลี่ยจากกลุ่มข้อมูลการแสดงออกของยีนทั้งหมดมาทดแทน

5.2.3 ลักษณะสำคัญประการหนึ่งของข้อมูลการแสดงออกของยีน คือ มีลักษณะของข้อมูลที่มีพื้นฐานของการกระจายตัวแบบไม่ปรกติ เป็นผลที่ได้มาจากการทดลองในห้องทดลอง ดังนั้นเมื่อต้องการนำข้อมูลเหล่านี้มาใช้งานร่วมกับอัลกอริธึมทางคอมพิวเตอร์โดยเฉพาะอย่างยิ่งอัลกอริธึมที่ต้องการข้อมูลที่มีการกระจายตัวปรกติ ถ้าไม่มีการปรับเปลี่ยนรูปแบบของข้อมูลอาจส่งผลให้ประสิทธิภาพในการทำงานของอัลกอริธึมต่ำลงได้ ดังนั้นการปรับให้ค่าของกลุ่มข้อมูลการแสดงออกของยีนเหล่านี้อยู่ในรูปแบบของการกระจายตัวแบบปรกติ จึงเป็นสิ่งจำเป็นเพราะผลการทดลองที่ผ่านมาได้แสดงให้เห็นอย่างชัดเจนถึงประสิทธิภาพในการทำงานที่แตกต่างกัน

5.2.4 โดยปรกติแล้ววิธีการคัดเลือกมิติข้อมูลโดยทั่วไปมักถูกคิดค้นและพัฒนาจากนักวิจัยด้านคอมพิวเตอร์หรือผู้เชี่ยวชาญ ดังนั้นอัลกอริธึมส่วนใหญ่จึงสนับสนุนการทำงานกับกลุ่มข้อมูลทั่วไป ซึ่งอาจไม่เหมาะสมกับข้อมูลที่มีลักษณะเฉพาะตัวเช่นข้อมูลการแสดงออกของยีน เนื่องจากกลุ่มของยีนเป็นกลุ่มข้อมูลที่มีคุณลักษณะพิเศษ โดยที่แต่ละกลุ่มข้อมูลจะมีความสัมพันธ์กันภายในกลุ่ม ดังนั้นการพิจาณาวิธีการคัดเลือกมิติของยีนจึงควรพิจารณาถึงคุณลักษณะข้อนี้เป็นสำคัญ เนื่องจากสามารถทำให้การคัดเลือกมิติของยีนประสบความสำเร็จได้มากที่สุด ดังนั้นวิธีการคัดเลือกมิติข้อมูล CfsGR ที่นำเสนอจึงได้นำเอาพฤติกรรมของยีนที่มีความเกี่ยวข้องระหว่างกัน พร้อมกับ

ค้นหาอำนาจจำแนกของยีนที่มีต่อคลาสเป้าหมายมาใช้พิจารณา ซึ่งเป็นจุดเด่นที่สำคัญสำหรับการคัดเลือกมิติของข้อมูลในลักษณะนี้

5.2.5 จากปัญหาของการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นด้วยอัลกอริธึม COBWEB ที่อาศัยเพียงฟังก์ชัน Category Utility (CU) ในการคำนวณว่าข้อมูลใดควรจะอยู่กลุ่มเดียวกัน มักไม่เพียงพอต่อการทำงาน และจากปัญหาของอัลกอริธึมการจัดกลุ่มข้อมูลแบบเรียนรู้เพิ่มขึ้นในส่วนของการเรียนรู้แบบไม่มีผู้สอน ที่อาศัยการวัดระยะทางแบบมาหาลาโนบิส และฟังก์ชั่นความเป็นสมาชิกแบบเกาส์เซียนเพื่อจัดกลุ่มข้อมูล ซึ่งมีประสิทธิภาพไม่เพียงพอสำหรับการจัดกลุ่มข้อมูล การนำเสนอขั้นตอนการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้นด้วย อัลกอริธึม IHC จึงมีวัตถุประสงค์เพื่อดำเนินการจัดกลุ่มข้อมูลแบบลำดับชั้นโดยอาศัยแนวคิดของวิธีการ COBWEB และอาศัยการวัดระยะห่างระหว่างข้อมูลแบบไม่มีผู้สอนของอัลกอริธึม ILM เพื่อช่วยให้ประสิทธิภาพในการจัดกลุ่มข้อมูลดียิ่งขึ้น

5.2.6 ผลจากค่าความถูกต้องของวิธีการคัดเลือกมิติข้อมูลและอัลกอริธึม IHC ที่นำเสนออยู่บนข้อมูลการแสดงออกของยีนทั้ง 4 กลุ่มข้อมูลมีข้อควรพิจารณาดังนี้ 1) กลุ่มข้อมูล DLBCL ให้ผลลัพธ์ค่าความถูกต้อง 69.17% ซึ่งเป็นผลลัพธ์ค่าความถูกต้องที่ต่ำที่สุดจากทั้ง 4 กลุ่มข้อมูล ทั้งนี้เนื่องจากข้อมูล DLBCL มีลักษณะของข้อมูลที่สูญหายเป็นจำนวนมากและรวมถึงการกระจายตัวของข้อมูลที่มีลักษณะเป็นแบบไม่ปรกติ 2) กลุ่มข้อมูล Leukemia ให้ผลลัพธ์ค่าความถูกต้อง 100.00% ซึ่งเป็นผลลัพธ์ค่าความถูกต้องที่สูงที่สุดจากทั้ง 4 กลุ่มข้อมูล ทั้งนี้เนื่องจากข้อมูล Leukemia ไม่มีข้อมูลที่สูญหายและการกระจายตัวของข้อมูลมีลักษณะเกือบปรกติ 3) กลุ่มข้อมูล Embryonal Tumor ให้ผลลัพธ์ค่าความถูกต้อง 93.33% ซึ่งเป็นผลลัพธ์ค่าความถูกต้องลำดับที่สามจากทั้งสี่กลุ่มข้อมูล นับได้ว่ามีค่าความถูกต้องที่สูง เนื่องจากมีข้อมูลที่สูญหายน้อยและการกระจายตัวของข้อมูลมีลักษณะเกือบเป็นปรกติ 4) กลุ่มข้อมูล Colon Cancer ให้ผลลัพธ์ค่าความถูกต้อง 93.55% ซึ่งเป็นผลลัพธ์ค่าความถูกต้องลำดับที่สองจากทั้งสี่กลุ่มข้อมูล และมีค่าความถูกต้องที่สูงเนื่องจากไม่มีข้อมูลที่สูญหายพร้อมกับการกระจายตัวของข้อมูลมีลักษณะไม่เป็นปรกติ

5.2.7 จากผลการเปรียบเทียบประสิทธิภาพด้านเวลาที่ใช้ในการประมวลผลของอัลกอริธึมการจัดกลุ่ม

ข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น เห็นได้ว่าอัลกอริธึม IHC ใช้เวลาในการประมวลผลที่มากกว่าวิธีการ ILM ในส่วนของการเรียนรู้แบบไม่มีผู้สอน ทั้งนี้เนื่องจากขั้นตอนการทำงานของอัลกอริธึม IHC แตกต่างจากและเพิ่มเติมจากวิธีการ ILM ทั้งในส่วนของการสร้างโครงสร้างแบบลำดับชั้น และการสร้างแบบจำลอง

5.2.8 จากการวิเคราะห์ลำดับของข้อมูล ทั้ง 3 วิธีพบว่าวิธีลำดับของข้อมูลเข้าตามชุดข้อมูล

ฝึกฝนจากการสุ่มตัวอย่าง ให้ประสิทธิภาพในด้านค่าความถูกต้องมากที่สุด ซึ่งหมายความว่าลำดับข้อมูลเข้าดังกล่าวมีโอกาสสร้างความหลากหลายในชุดข้อมูลการฝึกฝนได้มากกว่าการใช้ลำดับของข้อมูลนำเข้าตามชุดข้อมูลฝึกฝนปรกติและตามชุดข้อมูลฝึกฝนจากหลังไปหน้า สาเหตุเนื่องจากทั้งสองวิธีหลัง ไม่สามารถคัดเลือกข้อมูลที่มีความหลากหลายเพื่อนำมาสร้างแบบจำลองระบบ ส่งผลให้แบบจำลองระบบไม่มีประสิทธิภาพเท่าที่ควร

5.2.9 จากผลลัพธ์การทดลองกำหนดค่าเริ่มต้น เนื่องจากผู้ใช้งานอัลกอริธึม IHC จำเป็นต้องกำหนดค่าความแปรปรวนร่วม และอัตราระยะห่าง การกำหนดค่าที่เหมาะสมเหล่านี้ส่งผลให้ประสิทธิภาพของการทำงานอัลกอริธึม IHC สูงขึ้น และจากผลการทดลองแสดงให้เห็นว่าการกำหนดค่าความแปรปรวนร่วมที่ให้ค่าความถูกต้องที่สูงในทั้งสี่กลุ่มข้อมูลจะอยู่ในช่วงที่ต่ำคืออยู่ระหว่าง 0.01 ถึง 0.10 และในทางตรงกันข้ามอัตราระยะห่างที่ให้ค่าความถูกต้องอยู่ในระดับสูงของทั้ง 4 กลุ่มข้อมูลถูกกำหนดอยู่ในช่วงระหว่าง 0.20 ถึง 1.00 แสดงให้เห็นว่าผู้ใช้งานควรกำหนดค่าความแปรปรวนให้อยู่ในช่วงที่ต่ำคือ 0.01 ถึง 0.10 และกำหนดอัตราระยะห่างให้อยู่ในช่วงที่สูงกว่าคืออยู่ในช่วง 0.20 ถึง 1.00

5.3 ข้อเสนอแนะ

การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น พร้อมกับวิธีการคัดเลือกมิติข้อมูลที่ผู้วิจัยได้นำเสนอ และพัฒนานี้มีข้อเสนอแนะในการทำวิจัยเพิ่มเติม เพื่อให้กระบวนการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น มีประสิทธิภาพสูงขึ้น ดังนี้

5.3.1 เนื่องจากการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ใช้วิธีการทดแทนค่าที่สูญหายโดยใช้วิธี KNNimputation ซึ่งเป็นวิธีการที่ใช้กันโดยทั่วไปและได้รับการยอมรับว่าให้ประสิทธิภาพที่เหมาะสม อย่างไรก็ตามวิธีการ KNNimputation ควรต้องได้รับการปรับปรุงในเรื่องของประสิทธิภาพ เพราะนอกจากแก่นของวิธีการนี้คือการคัดเลือกโดยการเรียนรู้ตัวอย่างแล้วนำมาแทนที่ (Instance based) ซึ่งในกรณีที่มีข้อมูลเป็นจำนวนมากมักส่งผลให้การทำงานช้า รวมถึงการกำหนดตัวอย่างที่เหมาะสมเพื่อใช้เปรียบเทียบกับข้อมูลที่สูญหายอีกด้วย

5.3.2 อัลกอริธึมจำนวนมากมักมีประสิทธิภาพต่ำลง เมื่อทำงานกับข้อมูลที่มีการกระจายตัวไม่เท่ากัน ดังนั้นเทคนิคการลดการกระจายตัวของข้อมูลมีส่วนอย่างมากที่ทำให้อัลกอริธึมทำงานได้อย่างถูกต้อง อย่างไรก็ตามในปัจจุบันมีวิธีการไม่มากที่ใช้จัดการการกระจายตัวของข้อมูลแบบไม่ปรกติให้เป็นแบบปรกติ ดังนั้นการพัฒนาวิธีการจัดการข้อมูลให้อยู่ในรูปแบบของการกระจายตัวปรกตินับว่าเป็นหนทางหนึ่งที่เหมาะสมสำหรับการวิจัยต่อไป

5.3.3 เนื่องจากกลุ่มข้อมูลการแสดงออกของยีน ประกอบด้วยยีนต่าง ๆ เป็นจำนวนมาก

แต่อย่างไรก็ตามยีนที่เป็นยีนสำคัญมีอยู่เพียงเล็กน้อย ดังนั้นการคัดเลือกมิติของยีนจึงเป็นทางหนึ่งที่นอกจากจะช่วยเหลืองานของนักชีววิทยาแล้ว ยังช่วยให้ประสิทธิภาพในการค้นหาต้นกำเนิดของโรคดียิ่งขึ้น อย่างไรก็ตามอัลกอริธึมในปัจจุบันส่วนใหญ่ถูกพัฒนาเพื่อใช้งานกับข้อมูลโดยทั่วไป เมื่อนำมาใช้กับข้อมูลการแสดงออกของยีนจึงให้ประสิทธิภาพไม่เพียงพอ ดังนั้นการพัฒนาอัลกอริธึมที่สามารถคัดเลือกมิติของยีนจึงเป็นสิ่งท้าทาย และยังช่วยให้นักชีววิทยาสามารถค้นหากลุ่มยีนที่มีความสำคัญได้ถูกต้องมากยิ่งขึ้น

5.3.4 อัลกอริธึมการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น ยังคงต้องการให้ผู้ใช้งานกำหนดพารามิเตอร์ที่สำคัญ ได้แก่ ค่าความแปรปรวนร่วม และค่าอัตราระยะห่าง ซึ่งมักเป็นเรื่องยากที่ผู้ใช้งานจะสามารถกำหนดได้อย่างถูกต้อง ดังนั้นการวิจัยเกี่ยวกับการกำหนดค่าดังกล่าวแบบอัตโนมัติจะช่วยให้ผู้ใช้งานมีความสะดวก และส่งผลให้อัลกอริธึมทำงานได้ถูกต้องแม่นยำมากขึ้นอีกด้วย

บรรณานุกรม

ภาษาไทย

ไก้รุ่ง เฮงพระพรหม. *การเลือกคุณลักษณะสำหรับการพยากรณ์ค่าที่ขาดหายสำหรับข้อมูลหลายมิติและการประยุกต์สำหรับข้อมูลไมโครอาร์เรย์.* วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2553.

นำชัย ชีววิวรรธน์. "มหัศจรรย์ดีเอ็นเอ (1)." *เทคโนโลยีชีวภาพปริทรรศน์*, ปีที่ 1 (1) 2546 : 15.

บุญญานาถ นาถวงษ์. *Biosafety & GMSs.* เอกสารประกอบการสอน Online learning, ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ. [ม.ป.ท. : ม.ป.พ.], 2552.

ประลองพล ประสงค์พร. *การประมาณค่าพารามิเตอร์ในตัวแบบความถดถอยโลจิสติกเมื่อมีค่าสูญหาย.* วิทยานิพนธ์สถิติศาสตรมหาบัณฑิต สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2551.

ภัทราวุฒิ แสงศิริ ศจีมาจ ณ วิเชียร และพยุง มีสัจ. "การเปรียบเทียบประสิทธิภาพการลดตัวแปรข้อมูลเข้าที่เหมาะสม สำหรับโครงข่ายประสาทเทียมระหว่างเทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้น และการวิเคราะห์องค์ประกอบ เพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง." *การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5.* 2552 : (90-91).

_____. ศจีมาจ ณ วิเชียร และพยุง มีสัจ. "ชีวสารสนเทศ (Bioinformatics)." *วารสารวิทยาศาสตร์และเทคโนโลยีสารสนเทศ.* ปีที่ 25 (3) (กรกฎาคม – กันยายน 2553) : 51 – 56.

มนตรี พิริยะกุล. "วิธีแทนที่ข้อมูลที่สูญหาย มหาวิทยาลัยรามคำแหง." *วารสารการพัฒนาทรัพยากรมนุษย์*, คณะวิทยาศาสตร์, ปีที่ 2 (1) (มกราคม - มีนาคม 2549) : 162-177.

วิโรจน์ ไววานิชกิจ. "ความรู้เบื้องต้นเกี่ยวกับชีวสารสนเทศ." *วารสารอายุรศาสตร์อีสาน.* มหาวิทยาลัยขอนแก่น ปีที่ 6 (4) (ตุลาคม – ธันวาคม 2550) : 1-3.

ศูนย์สิริกิติ์บรมราชินีนาถ. *เพื่อโรคมะเร็งเต้านม.* สภากาชาดไทย [ออนไลน์] [สืบค้นวันที่ 30 สิงหาคม 2551]. จาก http://hpc4.anamai.moph.go.th/Group/health /adult/cabreast1.doc

สมาคมพันธุศาสตร์แห่งประเทศไทย และสถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี. *สาระน่ารู้อณูพันธุศาสตร์*. กรุงเทพฯ : บริษัท เท็กซ์ แอนด์ เจอร์นัล พับลิเคชั่น จำกัด, 2547.

สุรินทร์ ปียะโชคคณากุล. *พันธุวิศวกรรมเบื้องต้น*. พิมพ์ครั้งที่ 3. กรุงเทพฯ : สำนักพิมพ์มหาวิทยาลัยเกษตรศาสตร์, 2548.

_____. *เครื่องหมายดีเอ็นเอ : จากพื้นฐานสู่การประยุกต์*. พิมพ์ครั้งที่ 1. กรุงเทพฯ : สำนักพิมพ์มหาวิทยาลัยเกษตรศาสตร์, 2552.

ภาษาอังกฤษ

Ajmillar. *DNA Microarrays.* [online] 2011. [cited 2011 September 17]. Available from : URL : https://wiki.lifesci.dundee.ac.uk/instruments/dna_microarrays

Alizadeh, A. A. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature.* 403(3), (Feburay 2000) : 503-511.

Alon, U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D., and Levine A. J. "Broad pattern of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays." *The National Academy of Science of the United States of America Proceeding (PNAS).* 96(12), USA, [n.p.], 1999 : (6750-6745).

Al-Omary, A. Y. and Jamil M. S. "A new approach of clustering based machine-learning algorithm." *Knowledge-Based Systems.* 19(4), 2006 : 248-258.

Azuaje Francisco. *Bioinformatics and Biomarker Discovery.* First Edition, Singapore : Wiley-Blackwell, 2010.

Ben-Dor Amir,Bruhn Laurakay, Friedman Nir, and Iftach Nachman. "Tissue classification with gene expression profiles," *The Proceedings of the fourth annual international conference on Computational molecular biology.* Tokyo, Japan, 2000 : (59).

Cao, K.-A. L, Bonnet A., and Gadat, S. "Multiclass classification and gene selection with a stochastic Algorithm." *Computational Statistics & Data Analysis.* 53(10), (August 2009) : 3601-3615.

Carpenter, A. G and Grossberg S, Markuzon N, Reynolds, H. J. and Rosen B. D. "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps." *IEEE Transactions on Neural Network.* 3(5), (September 1992) : 698-713.

Chaiworawitkul, S. *A design performance driven learning framework for conceptual design knowledge : methodology development and applications.* Doctor of Philosophy Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 2008.

Chakraborty, S. "Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach." *Computational Statistics & Data Analysis.* 53(4), (February 2009) : 1462-1474.

Cheng-San, Y., Li-Yeh C., Jung-Chike, L., and Cheng-Hong, Y. "A novel BPSO approach for gene selection and classification of microarray data, in Neural Networks." *IEEE International Joint Conference on Neural Network.* [n.p. : n.p.], 2008 : (2147-2152).

_____, Li-Yeh, C., Chao-Hsuan, K., and Cheng-Hong, Y. "A hybrid approach for selecting gene subsets using gene expression data." *IEEE Conference, Soft Computing in Industrial Applications (SMCia).* [n.p. : n.p.], 2008 : (159-164).

Cunningham P. *Dimension Reduction.* Technical Report UCD-CSI-2007-7, [n.p. : n.p.], 2007 : 1-4.

Dabney, A. R. "Classification of microarrays to nearest centroids." *Bioinformatics.* 21(22), (September 2005) : 4148-4154.

Daxin, Jiang, Jian, Pei, and Zhang, Aidong. "DHC: a density-based hierarchical clustering method for time series gene expression data." *The Third IEEE Symposium on Bioinformatics and Bioengineering (BIBE'03).* [n.p. : n.p.], 2003 : (1-8).

Dougherty S., Bowyer, W. K., and Kranenburg, C. "ROC curve evaluation of edge detector performance," *International Conference* on *Image Processing (ICIP 98).* [n.p. : n.p.], 1998 : (525-529).

Fisher, D. H. "Knowledge Acquisition Via Incremental Conceptual Clustering." *Machine Learning.* 2(2), (September 1987) : 139-172.

Garcia-Nieto J., Alba, E., and Apolloni, J. "Hybrid DE-SVM Approach for Feature Selection: Application to Gene Expression Datasets (Colon Cancer)." *2nd International on Logistics and Industrial Informatics (LINDI 2009).* [n.p. : n.p.], 2009 : (1-6).

Garzón, C. J. and Quintales L. M. "Evolutionary Techniques for Hierarchical Clustering Applied to Microarray Data." *2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics.* [n.p. : n.p.], 2008 : (118-127).

Giudici P. and Figini, S. *Applied data mining for business and industry.* Second Edition, United Kingdom : A John Wiley and Sons, 2009.

Golub, T.R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander E. S. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science Journal.* 286(5439), 1999 : 531-537.

Hadzikadic, M. and Yun, D. Y. Y. "Concept formation by incremental conceptual clustering." *Proceedings of the 11th International Joint Conference on Artificial intelligence.* Detroit, Michigan : Morgan Kaufmann Publishers Inc, 1989 : (831-836).

Hall, M. A. *Correlation-based Feature Selection for Machine Learning.* Doctor of Philosphy Thesis, Department of Computer Science, The University of Waikato Newzealand, 1999.

Han, J., Kamber, M. and Tung, A. K. H. *Spatial Clustering Methods in Data Mining: A Survey.* [online] 2001. [cited 10 October 2009]. Available from : URL : http://www.spatial.cs.umn.edu/paper_ps/ Spatial-clustering.ps

Herrero J. "Gene Expression Pattern Preprocessing." *3rdOctorber Bioinformatics Unit CNIO.* [n.p. : n.p.], 2002 : (3).

Hikaru, M., Seiichiro, A, Tomomasa, N. and Yoshifumi, O. "Accurate and robust gene selection for disease classification using a simple statistic." *Bioinformation.* 3(2), (October 2008) : 68-71.

Horng, J.-T., Wu, L.-C., Liu, B.-J., Kuo, J.-L., Kuo, W.-H., and Zhang, J.-J. "An expert system to classify microarray gene expression data using gene selection by decision tree." *Expert Systems with Applications.* 36(5), (July 2009) : 9072-9081.

Horspool, D. *An overview of the (basic) central dogma of molecular biochemistry with all enzymes labeled.* [Online] 2008. [cited 5 September 2011]. Available from : URL : http://en.wikipedia. org/wiki/File:Central_Dogma_of_Molecular_Biochemistry_with _Enzymes.jpg

Huang, C. J. and Liao, W. C. "A comparative study of feature selection methods for probabilistic neural networks in cancer classification." *15th the Proceeding of Tools with Artificial Intelligence,* [n.p. : n.p.], 2003 : (451-458).

Hyunsoo, K., Golub G. H., and Park H. "Missing value estimation for DNA microarray gene expression data: local least squares imputation." *Bioinformatics,* 21(2), (January 2005) : 187-198.

Witten I. H., Eibe F., and Hall, M. A. *Data Mining Practical Machine Learning Tools and Techniques.* Third Edition, United State : Elsevier Inc, 2011.

Jaeger J., Sengupta R., Ruzzo W.L. "Improved Gene Selection For Classification of Microarrays." *Pacific Symposium on Biocomputing.* [n.p. : n.p.], 2003 : (53-64).

Jae-Woo, C. and Du-Seok J. "A new cell-based clustering method for large, high-dimensional data in data mining applications." *ACM symposium on Applied computing*, Madrid, Spain, 2002 : (503-507).

Jain, A. K., Murty, M. N., and Flynn, P.J. "Data clustering: a review." *ACM Computer Survey.* 0360-0300 31(3), (Sept 1999) : 264-323.

_____. "Data clustering: 50 years beyond K-means." *Pattern Recognition Letters.* 31(8), (June 2010) : 651-666.

Jiawei, H. and Micheline, K. *Data Mining: Concepts and Techniques.* Series Editor, USA : Morgan Kaufmann, 2000.

Jinlong, S. and Zhigang, L. "Missing value estimation for DNA microarray geneexpression data with principal curves." *International Conference on Bioinformatics and Biomedical Technology (ICBBT).* [n.p. : n.p.], 2010 : (262-265).

Jin-Hyuk, H. and Sung-Bae, C. "Cancer classification with incremental gene selection based on DNA microarray data." *IEEE Symposium, Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '08).* [n.p. : n.p.], 2008 : (70-74).

Kai, L, Wang, L., and Hao, L. "Comparison of Cluster Ensembles Methods Based on Hierarchical Clustering." *Computational Intelligence and Natural Computing (CINC).* [n.p. : n.p.], 2009 : (499-502).

Kamal, A., Zhu, X., Pandya, A., Hsu, S., and Shoaib, M. "The Impact of Gene Selection on Imbalanced Microarray Expression Data." *Bioinformatics and Computational Biology.* [n.p. : n.p.], 2009 : (259-269).

Ke, C. H., Yang, C. H., Chuang L. Y. and Yang C. S. "A hybrid filter/wrapper approach of feature selection for gene expression data." *IEEE Systems, Man and Cybernetics (SMC)*. [n.p. : n.p.], 2008 : (2664-2670).

Kim, P. J. and Choi, J. Y. "Incremental Conceptual Clustering Using a Modified Category Utility." *International Conference on Circuites/Systems, Computers and Communications in NonlinearScience and Numerical Simulation.* [n.p. : n.p.], 2005 : (23-24).

Komkhao, M. *An Incremental Learning Algorithm Based-on Mahalanobis Distance.* Master Thesis of Graduate College, Faculty of Information Technology, King Mongkut's Institute of Technology, Thailand, 2007.

Leung, Y. and Hung, Y. "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification." *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 7(1), (January-March 2010) : 108-117.

Li, A., Xie, H., Mark, C., Obradovic Z., Desmond, S., and Vasileios M. "Analysis of Multiplex Gene Expression Maps Obtained by Voxelation." *IEEE International Conference, Bioinformatics and Biomedicine (BIBM '08).* [n.p. : n.p.], 2008 : (115-123).

Liu H. and Hiroshi, M. *Computational Method of Feature Selection.* Chapman & Hall/CRC, United State of America : Taylor & Francis Group, Boca Raton London New York, 2008.

Liu, Q. and Zhang, Y. "Study of Informative Gene Selection for Gene Expression Profiles." *WRI Global Congress on Intelligent Systems (GCIS '09).* 2009 : (582-587).

Ma, Patrick C. H., Chan, K. C. C., Yao X., and Chiu D. K. Y. "An evolutionary clustering algorithm for gene expression microarray data analysis." *IEEE Transactions on Evolutionary Computation.* 10(3), (June 2006) : 296-314.

Meesad, P. and Yen, G. Gary. "An Effective Neuro-Fuzzy Paradigm for Machinery Condition Health Monitoring." *IEEE Transactions on systems, man, and cybernetics – part B: cybernatics.* 31(4), (August 2001) : 1567-1572.

Michalski, R. S. and Tecuci, G., eds. *In Machine Learning: A Multistrategy Approach, Vol. IV.* San Francisco, CA : Morgan Kaufmann, 1994.

Mohamad, M., Omatu, S., Deris, S., Misman, M., and Yoshioka, M. "A multi-objective strategy in genetic algorithms for gene selection of gene expression data." *Artificial Life and Robotics.* 13(2), (January-February 2009) : 410-413.

Mukherjee, S. N. and Roberts, S. J. "A Theoretical Analysis of Gene Selection." *IEEE* Conference on *Computational Systems Bioinformatics.* [n.p. : n.p.], 2004 : (131-141).

Mundra, P. and Rajapakse, J. "F-score with Pareto Front Analysis for Multiclass Gene Selection, in Evolutionary Computation." *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics.* 5483(2009), 2009 : 56-67.

Myint, H. O. and Meesad, P. "Incremental Learning Algorithm based on Support Vector Machine with Mahalanobis distance (ISVMM) for intrusion prevention." *6th International Conference Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).* [n.p. : n.p.], 2009 : (630-633).

Nayak A. and Stojmenovic I. *Handbook of Applied Algorithms Solving Scientific, Engineering, and Practical Problems.* United State of America : John Wiley & Sons, 2008.

Oded, M. and Rokach, L. *The Data Mining and Knowledge Discovery Handbook.* First Edition, Israel : Springer Science Business Media. 2005.

Pomeroy, S. L., *et al.* "Prediction of Central Nervous System Embryonal Tumor Outcome based on Gene Expression." *Nature Journal.* 415(6870), (January 2002) : 436-442.

Premachandran, V. and Kakarala R. "Measuring the effectiveness of bad pixel detection algorithms using the ROC curve," *IEEE Transactions on Consumer Electronic.* 56(4), (November 2010) : 2511-2519.

Rosenwald, A. *et al.* "The use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma." *The new England journal of Medicine.* 346 (25), (June 2002) : 1937-1947.

Rosy D., Bhattacharyya, D. K., and Kalita, J. K. "An incremental clustering of gene expression data." *Nature & Biologically Inspired Computing (NaBIC).* [n.p. : n.p.], 2009 : (742 – 747).

Ryszard, S. M. and Tecuci, G. *In Machine Learning: A Multistrategy Approach Vol. IV.* 5th Edition, San Francisco, CA : Morgan Kaufmann, 1994.

Saengsiri, P., Wichian, S. Na, Meesad, P., and Herwig, U. "Integrating Feature Selection Methods for gene Feature Selection." *Series Lecture Notes in Informatics(LNI).* Vol.P-186, Bone, Germany : Gesellschaft fur Informatik (GI), 2011 : (82).

_____, Wichian, S. Na, Meesad, P., and Herwig, U. "Classification Models based-on Incremental Learning Algorithm and Feature Selection on Gene Expression Data."*8th International Conference Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. (ECTI-CON),* [n.p. : n.p.], 2011 : (426-429).

Schaefer, G. and Nakashima, T. "Data Mining of Gene Expression Data by Fuzzy and Hybrid Fuzzy Methods." *IEEE Transactions, Information Technology in Biomedicine.* 14(1), (January 2010) : 23-29.

Shen, Q., Shi W.-M., and Kong, W. "New gene selection method for multiclass tumor classification by class centroid." *Journal of Biomedical Informatics.* 42(1) (Febuary 2009) : 59-65.

Slavik, M., Zhu, X., Mahgoub, I., and Shoaib, M. "Parallel Selection of Informative Genes for Classification." *Bioinformatics and Computational Biology.* [n.p. : n.p.], 2009 : (388-399).

Spencer, D. *Card Sorting: Designing Usable Categories.* [online] 2009. [cited 2011 September 5]. Available from : URL : http://www.flickr.com/photos/9411449@N05/3344378948.

Sumathi, S. and Sivanandam, S.N. *Introduction to Data Mining and its Application.* Studies in Computational Intelligence Springer, Berlin : Springer Berlin Heidelberg New York, 2006.

Starmer, C. F. *Adventures with curiosity and learning.* [online] [cited 2011 September 5]. Available from : URL : http://frank.itlab.us/photo_essays/wrapper.php?nephila_2002_dna.html

Tan P. N., Steinbach, M., and Vipin, K. *Introduction to Data Mining.* [n.p.], United State of America : Addison Wesley, 2005.

Vuskovic M. and Sijiang, Du. "Classification of prehensile EMG patterns with simplified Fuzzy ARTMAP Networks." *International Joint Conference on Neural Network.* [n.p. : n.p.], 2002 : (2539-2544).

Wang, J. H., Liu, J. M., Zhao, Y., and Bi, B. "A New Method of Eliminating Noise Based on Clustering." *International Conference Machine Learning and Cybernetics.* [n.p. : n.p.], 2007 : (3956-3960).

Weixiang, L., Tianfu, W., and Tang, A. "Hierarchical Clustering of Gene Expression Data with Divergence Measure." *3rd International Conference Bioinformatics and Biomedical Engineering.* [n.p. : n.p.], 2009 : (1-3).

Yijuan, L., Jennifer, T. Qi, N., Feng, L., and Yufeng, W. "Adaptive discriminant analysis for microarray-based classification." *ACM Transaction on Knowledge Discovery Data.* 2(1), 2003 : 1-20.

Yu, W, Zhang, H., Kuanheng, W., Lin, Q., and He, M. "Hierarchical Clustering of Lung Cancer Related Genes." *2nd International Conference on Bioinformatics and Biomedical Engineering.* [n.p. : n.p.], 2008 : (63-65).

Yuni, X. and Bowei, X. "Conceptual Clustering Categorical Data with Uncertainty." *Tools with Artificial Intelligence.* [n.p. : n.p.], 2007 : (329-336).

Xiong, H. and Chen, X. W. "Data-Dependent Kernel Machines for Microarray Data Classification." *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 4(4), (October 2007) : 583-595.

Xu, R., Cai, X., and Wunsch, II C. D. "Gene Expression Data for DLBCL Cancer Survival Prediction with A Combination of Machine Learning Technologies." *27th International Conference on Engineering in Medicine and Biology Society.* [n.p. : n.p.], 2005 : (894-897).

Zhou, J., Pan, Y., Chen, Y., and Liu Y. "Ensemble Classifiers Based on Kernel PCA for Cancer Data Classification." *Emerging Intelligent Computing Technology and Applications,* Vol. 5755, South Korea : Springer Berlin / Heidelberg, Artificial Intelligence, 2009 : 955-964.

ประวัติผู้วิจัย

ชื่อ : นายภัทราวุฒิ แสงศิริ

ชื่อวิทยานิพนธ์ : การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบเพิ่มขึ้น สำหรับข้อมูลการแสดงออกของยีน

สาขาวิชา : เทคโนโลยีสารสนเทศ

ประวัติ

ประวัติการศึกษา

พ.ศ. 2554 สำเร็จการศึกษาระดับปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

พ.ศ. 2551 Communities in Intelligent Service Network (CISN), Fern University in Hagen, Germany.

พ.ศ. 2550 Certificate of Advance Studies (Information Technology), School of Engineering and Technology, Asian Institute of Technology, Thailand.

พ.ศ. 2540 สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าลาดกระบัง

พ.ศ. 2536 สำเร็จการศึกษาระดับปริญญาบัณฑิต สาขาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยสยาม