การพัฒนาวิธีการจัดกลุ่มเชิงวิวัฒนาการสำหรับการวิเคราะห์ความสำคัญของกลุ่มยีนส์

โดย

นางสาวพชรอร แสนประเสริฐ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต

ภาควิชาวิศวกรรมไฟฟ้า

มหาวิทยาลัยศิลปากร

ปีการศึกษา 2566

การพัฒนาวิธีการจัดกลุ่มเชิงวิวัฒนาการสำหรับการวิเคราะห์ความสำคัญของกลุ่มยีนส์

โดย
นางสาวพชรอร แสนประเสริฐ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาวิศวกรรมไฟฟ้า
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566

DEVELOPMENT OF AN EVOLUTIONARY CLUSTERING METHOD FOR GENE-SET ENRICHMENT ANALYSIS

By

MISS Pacharaon SANPRASERT

A Thesis Submitted in Partial Fulfillment of the Requirements

for Master of Engineering (ELECTRICAL AND COMPUTER ENGINEERING)

Department of ELECTRICAL ENGINEERING

Silpakorn University

Academic Year 2023

หัวข้อ	การพัฒนาวิธีการจัดกลุ่มเชิงวิวัฒนาการสำหรับการวิเคราะห์ความสำคัญของกลุ่มยีนส์
โดย	นางสาวพชรอร แสนประเสริฐ
สาขาวิชา	วิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ผู้ช่วยศาสตราจารย์ ดร. ยุทธนา เจวจินดา
อาจารย์ที่ปรึกษาร่วม	รองศาสตราจารย์ ดร. ชูเกียรติ สอดศรี
	รองศาสตราจารย์ ดร. ยรรยงค์ พันธ์สวัสดิ์

คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์และ
(ผู้ช่วยศาสตราจารย์ ดร. อรุณศรี ลีจีรจำเนียร) เทคโนโลยีอุตสาหกรรม

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ทวีศักดิ์ สมานชื่น)

.....อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร. ยุทธนา เจวจินดา)

.....อาจารย์ที่ปรึกษาร่วม
(รองศาสตราจารย์ ดร. ชูเกียรติ สอดศรี)

.....อาจารย์ที่ปรึกษาร่วม
(รองศาสตราจารย์ ดร. ยรรยงค์ พันธ์สวัสดิ์)

620920062 : วิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต

นางสาว พชรอร แสนประเสริฐ: การพัฒนาวิธีการจัดกลุ่มเชิงวิวัฒนาการสำหรับการวิเคราะห์ความสำคัญของกลุ่มยีนส์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร. ยุทธนา เจวจินดา

วิทยานิพนธ์ฉบับนี้เสนอการสร้างตัวจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคเพื่อการวิเคราะห์ Pathway ในงานการวิเคราะห์ความสำคัญของกลุ่มยีนส์จากเทคนิคไมโครอะเรย์ โดยมุ่งหวังที่จะนำเสนอเครื่องมือตัวจัดกลุ่มรูปแบบใหม่ที่ให้ความหลากหลายของคำตอบ และเลี่ยงภาวะคำตอบเข้าสู่สภาวะสูงสุดหรือต่ำสุดท้องถิ่น รวมถึงเป็นเครื่องมือที่ทำให้คำตอบมีคุณลักษณะของการเป็นกลุ่มข้อมูลที่เหมาะสมเปรียบเทียบกับการจัดกลุ่มข้อมูลรูปแบบดั้งเดิม ได้แก่ การจัดกลุ่มข้อมูลประเภทลำดับขั้น (Hierarchical clustering) และการจัดกลุ่มข้อมูลแบบเคมีน (K-means clustering) วิธีการสร้างและแนวคิดการพัฒนาตัวจัดกลุ่มใช้ขั้นตอนวิธีการค้นหาแบบกลุ่มอนุภาคนี้เป็นขั้นตอนวิธีเชิงวิวัฒนาการโดยสร้างเครื่องมือจัดกลุ่มนี้ด้วยภาษาอาร์ใช้ในโปรแกรม Rstudio และเปรียบเทียบผลของการจัดกลุ่มเทียบกับการจัดกลุ่มรูปแบบอื่นจากเครื่องมือ pathfindR และเครื่องมือ stats ซึ่งเป็นเครื่องมือวิเคราะห์ข้อมูลยีนส์และการจัดกลุ่มทั่วไปสำหรับผู้วิจัยในวงกว้าง ผลการวิจัยพบว่าตัวจัดกลุ่มด้วยขั้นตอนวิธีการหาค่าความเหมาะสมแบบกลุ่มอนุภาคที่นำเสนอสามารถจัดกลุ่มข้อมูลได้จากการกำหนดค่า k เริ่มต้น อาทิ 19 25 และ 30 กลุ่ม สามารถแบ่งกลุ่มออกมา (kpso) ได้หลายรูปแบบตั้งแต่ 8 - 27 กลุ่ม ให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากที่สุดสูงสุดเมื่อเทียบการจัดกลุ่มรูปแบบอื่น ๆ ได้แก่ 73.5088 ในขณะที่การจัดกลุ่มแบบลำดับขั้นและเคมีน ให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากที่สุดได้แก่ 66.1339 และ 57.4773 ตามลำดับ คุณสมบัติของการเป็นกลุ่มของข้อมูลในแง่ความกะทัดรัดและการแยกกันของกลุ่มข้อมูลด้วยวิธีการที่นำเสนอยังให้คำตอบเป็นรองต่อการจัดกลุ่มอีก 2 รูปแบบ นอกจากนี้ยังพบจำนวนคำตอบที่แตกต่างกันจากการทดสอบ 431 ครั้ง พบว่าตัวจัดกลุ่มที่นำเสนอให้จำนวนคำตอบที่แตกต่างกันสูงสุดถึง 38 คำตอบ ในขณะที่การจัดกลุ่มแบบเคมีนและลำดับขั้นให้ 24 และ 1 คำตอบ ตามลำดับ เมื่อจัดกลุ่มข้อมูลได้ 14 กลุ่ม โดยสรุปผลของงานวิจัยนี้พบว่า ตัวจัดกลุ่มที่นำเสนอนี้ให้ผลโดดเด่นด้านความหลากหลายของรูปแบบคำตอบ และยังทำให้ค่าสมการจุดประสงค์สูงสุดมีค่ามากที่สุดอีกด้วย ทั้งนี้ตัวจัดกลุ่มที่นำเสนออาจไม่ให้ผลดีอย่างเห็นได้ชัดในด้านระยะห่างระหว่างกลุ่ม ซึ่งต้องมีการพัฒนาขั้นตอนวิธีการสร้างตัวจัดกลุ่มต่อไปเพื่อทำให้คุณสมบัติการเป็นกลุ่มข้อมูลทั้งในแง่ความกะทัดรัดของข้อมูลและการแยกกันของข้อมูลมีความเด่นชัดขึ้น

มหาวิทยาลัยศิลปากร

620920062 : Major (ELECTRICAL AND COMPUTER ENGINEERING)

MISS Pacharaon SANPRASERT : Development of an evolutionary clustering method for gene-set enrichment analysis Thesis advisor : Assistant Professor Yutana Jewajinda, Ph.D.

This thesis proposes a data clustering approach using a particle swarm optimization algorithm to analyze gene pathway importance in microarray analysis. The aim is to present a new clustering tool that provides a diverse set of solutions and avoids solutions that are trapped in local maximum or minimum. The tool also ensures that the output has the characteristics of a suitable data cluster when compared to traditional hierarchical and k-means clustering. The proposed clustering approach is developed using an evolutionary algorithm and implemented in Rstudio program. The results are compared with those obtained from pathfindR and stats, which are popular gene analysis and clustering tools for researchers. The study found that the proposed clustering approach with particle swarm optimization (PSO) algorithm provides a diverse set of clusters for k=19, 25, and 30, resulting in 8 to 27 clusters with the highest $\mathcal{L}_2$ objective function value of 73.5088. In comparison, the hierarchical and k-means clustering approaches yielded the highest $\mathcal{L}_2$ objective function values of 66.1339 and 57.4773, respectively. The properties of data grouping in terms of compactness and separability provided alternative clustering solutions. Moreover, from 431 tests conducted, the PSO clustering algorithm gave the highest maximum number of different answers, which was 38, while the hierarchical and k-means clustering methods gave 24 and 1 different answers, respectively, when the data were grouped into 14 clusters. In summary, this research concludes that the proposed clustering algorithm provided outstanding results in terms of diversity of clustering solutions and the highest $\mathcal{L}_2$ objective function value. However, it may not perform well in terms of inter-group distances, which requires further development of clustering methods to enhance the clustering properties of data in terms of compactness and separation of data groups.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ด้วยความอนุเคราะห์และความกรุณาช่วยเหลือให้คำปรึกษารวมถึงการสนับสนุนจากผู้ช่วยศาสตราจารย์ดร. ยุทธนา เจวจินดา อาจารย์ที่ปรึกษางานวิจัย ขอกราบขอบพระคุณเป็นอย่างสูงที่ได้ให้ความไว้วางใจ และความเชื่อใจในการทำงานของผู้วิจัยจนทำให้วิทยานิพันธ์ฉบับนี้สำเร็จได้ด้วยดี

ขอกราบขอบพระคุณ รองศาสตราจารย์ดร. ชูเกียรติ สอดศรี เป็นอย่างสูง ที่ได้ประสิทธิ์ประสาทความรู้ ปลูกฝังจิตสำนึกความเป็นนักวิชาการ และการคิดอย่างเป็นระบบเป็นอย่างดีให้แก่ผู้วิจัย ทั้งยังส่งเสริม ให้กำลังใจ ให้ข้อคิด ให้ความช่วยเหลือและให้การสนับสนุนแก่ผู้วิจัยในทุกๆ ด้านเสมอมา ขอมอบส่วนดีทั้งหมดนี้ให้แด่อาจารย์ทั้งสองท่านเป็นผู้มีส่วนสำคัญยิ่งแก่ผู้วิจัยเป็นอย่างสูง

ขอขอบพระคุณ รองศาสตราจารย์ดร. ยรรยงค์ พันสวัสดิ์ ที่ได้ให้ความเข้าใจ ให้คำปรึกษาและคำแนะนำ รวมถึงขอขอบพระคุณคณาจารย์และบุคลากรประจำภาควิชาเป็นอย่างยิ่งที่ได้ให้กำลังใจและให้ความช่วยเหลือในการดำเนินการเอกสารต่าง ๆ

อนึ่ง วิทยานิพนธ์นี้เกิดขึ้นอย่างสมบูรณ์ได้เพราะได้รับความช่วยเหลือและการสนับสนุนหลายฝ่าย ทั้งทางด้านทุนวิจัย การประสิทธิประสาทความรู้ และได้รับกำลังใจ จึงขอมอบส่วนดีทั้งหมดนี้ให้แก่ทุกท่านที่มีส่วนร่วมระหว่างการดำเนินงานวิจัยครั้งนี้ โดยเฉพาะอย่างยิ่งเหล่าคณาจารย์ภาควิชาวิศวกรรมไฟฟ้าที่ได้ให้ความรู้และประสบการณ์วิชาจนทำให้งานวิจัยครั้งนี้ประสบความสำเร็จและเป็นประโยชน์แก่ผู้ที่เกี่ยวข้อง สุดท้ายนี้ขอขอบคุณครอบครัว และบุคคลรอบข้างทุกท่านที่อาจกล่าวชื่อไม่หมดที่คอยเป็นกำลังใจและเชื่อมั่น สนับสนุนผู้วิจัยมาโดยตลอด ผู้วิจัยหวังเป็นอย่างยิ่งว่างานวิจัยนี้จะเป็นประโยชน์และเป็นแรงบันดาลใจแก่ทุกท่านในการนำไปสู่การคิดค้นสิ่งใหม่ๆ ในอนาคตต่อไป

พชรอร แสนประเสริฐ

สารบัญ

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

อณูชีววิทยา (Molecular Biology) เป็นศาสตร์ของการศึกษาทางชีววิทยาในระดับโมเลกุล เช่น ศึกษาการถอดรหัสยีนส์ (Translation) การกลายพันธุ์ของยีนส์ (Mutation) การศึกษาหน้าที่และการแสดงออกของยีนส์ (Gene expression) เป็นต้น เนื่องจากความก้าวหน้าทางเทคโนโลยีที่เพิ่มขึ้นทำให้ปัจจุบันมีงานวิจัยที่เกี่ยวข้องสาขานี้เป็นจำนวนมาก โดยมุ่งประโยชน์ในแง่ของการศึกษาเพื่อวินิจฉัยถึงสาเหตุของการเกิดโรคบางประการ อาทิ การตรวจวินิจฉัยสาเหตุทางพันธุกรรมในผู้ป่วยกลุ่มอาการออทิสซึมด้วยวิธีโครโมโซมอะเรย์ความละเอียดสูง [1] การตรวจไมโครอะเรย์ของแบคทีเรียก่อโรคที่มากับอาหาร [2] เป็นต้น ในการศึกษาชีววิทยาระดับโมเลกุลนี้จะมีการใช้เทคโนโลยีชีวภาพเพื่อให้ได้ข้อมูลของยีนส์ไปวิเคราะห์ข้อมูลในขั้นตอนถัดไป ซึ่งมีหลายเทคนิคที่ใช้ทำการศึกษา เช่น เทคนิค RNA sequencing เทคนิค SAGE เทคนิคไมโครอะเรย์ (Microarray) เป็นต้น ยอกจากนี้ยังต้องมีการใช้ศาสตร์ทางด้านคอมพิวเตอร์และมีขั้นตอนวิธีเพื่อช่วยจัดการข้อมูล หรือวิเคราะห์ข้อมูลอณูชีววิทยาซึ่งมีจำนวนมหาศาล รวมถึงแสดงผลข้อมูลให้เกิดความเข้าใจทางด้านชีววิทยา รวมเรียกงานในสาขานี้ว่าชีวสารสนเทศ หรือ Bioinformatics [3]

เทคนิคไมโครอะเรย์ เป็นเทคนิคที่นิยมใช้สำหรับการศึกษาการแปลรหัสทางพันธุกรรม (Transcriptomic) และเป็นหนึ่งในเทคนิคทางชีวภาพที่ใช้เพื่อศึกษาการแสดงออกของยีนส์ เนื่องจากเป็นเทคนิคที่สามารถให้ข้อมูลจำนวนมาก และเป็นการศึกษาเชิงปริมาณซึ่งต้องใช้การวิเคราะห์ข้อมูลหลากหลายขั้นตอนเนื่องจากข้อมูลมีความซับซ้อน [4] ดังนั้นเมื่อผลของการวิจัยที่ทำโดยเทคนิคไมโครอะเรย์จะได้รับการส่งต่อไปวิเคราะห์ข้อมูลด้วยคอมพิวเตอร์ ดังจะเห็นเครื่องมือหรือซอร์ฟแวร์จำนวนมากที่ถูกสร้างและพัฒนาขึ้นมาด้วยเทคนิคการคำนวณหรือใช้ขั้นตอนแนวคิดที่หลากหลายเพื่อให้ความสะดวกแก่ผู้ใช้งาน และตอบโจทย์หลากหลายวัตถุประสงค์สำหรับนักชีววิทยาหรือผู้วิจัยในสาขานี้ อาทิเช่น เครื่องมือ GScluster [5] เครื่องมือ DAVID web server [6] เครื่องมือ Enrichr [7] เครื่องมือ pathfindR [8] ทั้งหมดนี้เป็นเครื่องมือสำหรับการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis) และการวิเคราะห์ Pathway (Pathway Clustering) ซึ่งมีเทคนิคในการวิเคราะห์แตกต่างกันไป ยกตัวอย่างเช่น ขั้นตอนการวิเคราะห์ Pathway ซึ่งเป็นการวิเคราะห์ข้อมูลโดยการใช้ตัวจัดกลุ่มข้อมูล หรือ Clustering เครื่องมือ DAVID web server ได้

นำเสนอการจัดกลุ่มด้วยตัวจัดกลุ่มประเภทฟัซซี่ ซีมีน (Fuzzy c-means clustering) เครื่องมือ Enrichr และเครื่องมือ pathfindR ได้นำเสนอการจัดกลุ่มข้อมูลด้วยวิธีการแบ่งข้อมูลแบบลำดับขั้น ในขณะที่เครื่องมือ GScluster ได้นำเสนอฟังก์ชันการวัดความเหมือนกันของข้อมูลในรูปแบบโดยมีการใช้ข้อมูลค่า PPI หรือ Protein-Protein Interaction ซึ่งเป็นค่าที่ให้ความหมายทางชีววิทยาที่ได้จากฐานข้อมูลทางชีววิทยาเพื่อไปคำนวณร่วมกับการหาค่าสัมประสิทธิ์โคเฮนคัปปา นอกจากนี้ยังพบเครื่องมือที่ใช้วิธีเชิงวิวัฒนาการ ซึ่งเป็นขั้นตอนวิธีที่ให้ประสิทธิภาพในการค้นหาคำตอบที่เหมาะสมที่สุดร่วมใช้กับการแบ่งข้อมูล เช่น GenClust [9] ใช้ขั้นตอนวิธีเชิงพันธุกรรม (Genetic algorithm) สำหรับการวิเคราะห์กลุ่มยีนส์ที่มีความโดดเด่น ทั้งนี้ เครื่องมือที่หลากหลายหรือการสร้างขั้นตอนวิธีใหม่เพื่อแก้ปัญหาการจัดกลุ่มล้วนมีความสำคัญและมีประโยชน์ในการช่วยวิเคราะห์ Pathway ตามแนวคิดของผู้วิจัยที่แตกต่างกัน

ตามที่ได้กล่าวไป เนื่องจากการวิเคราะห์ Pathway คือการศึกษาที่พยายามจัดกลุ่มข้อมูลที่มีความเป็นกลุ่มการทำงานเดียวกัน ผ่านความสัมพันธ์ของสมาชิกยีนส์ภายในกลุ่ม และจากตัวอย่างเครื่องมือต่าง ๆ ในปัจจุบันพบว่าวิธีการจัดกลุ่มที่ใช้ มักเป็นตัวจัดกลุ่มรูปแบบทั่วไป เนื่องจากเป็นขั้นตอนวิธีจัดกลุ่มพื้นฐานที่เข้าใจง่าย อาทิ การจัดกลุ่มแบบลำดับขั้น ซึ่งการทำงานของชุดคำสั่งดังกล่าวจะมีพฤติกรรมการทำงานแบบละโมบ (Greedy algorithm) กล่าวคือมีวิธีหาคำตอบที่ดีที่สุด ณ ขณะนั้น จึงอาจส่งผลให้คำตอบของการแบ่งกลุ่มไม่ใช่คำตอบที่ดีที่สุด หรือมักได้ผลเป็นสภาวะสูงสุดหรือต่ำสุดท้องถิ่น (Local optima) ขั้นตอนวิธีเชิงวิวัฒนาการมักใช้สำหรับแก้ปัญหาการค้นหาคำตอบเพื่อเลี่ยงสภาวะสูงสุดหรือต่ำสุดท้องถิ่น [10] พบงานวิจัยที่นำขั้นตอนวิธีเชิงวิวัฒนาการรูปแบบต่าง ๆ เพื่อช่วยการจัดกลุ่มข้อมูล จากการศึกษาของ Elliackin Figueiredo และคณะ [11] มักพบขั้นตอนเชิงวิวัฒนาการเพื่อช่วยแบ่งกลุ่มข้อมูลโดยสามารถนำไปใช้ในขั้นการหาค่ากลุ่มที่เหมาะสมในการแบ่ง (ค่า k) ช่วยในการจำลองประชากรซึ่งเป็นตัวแทนคำตอบจำนวนมากในขั้นตอนสุ่มประชากร หรือนำไปช่วยปรับค่าพารามิเตอร์ในการแบ่งกลุ่มให้มีค่าเหมาะสมที่สุด อีกทั้งยังพบว่ามีการนำขั้นตอนเชิงวิวัฒนาการความฉลาดแบบฝูงไปช่วยการแบ่งกลุ่มข้อมูลประเภทแบ่งส่วน (Partitional clustering) 75.2% และแบบลำดับขั้น 3.3% [11] นอกจากนี้ยังพบรูปแบบวิธีการเชิงวิวัฒนาการที่นำไปช่วยการจัดกลุ่มข้อมูลที่พบบ่อย ได้แก่ ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle Swarm Optimization) ถึง 63.2% หรือประมาณ 103 บทความ จาก 153 บทความ ขั้นตอนแบบอาณานิคมผึ้งเทียม (Artificial Bee Colony) 10.4% ขั้นตอนอาณาจักรมด (Ant Colony Optimization) 6.7% ขั้นตอนแบบฝูงปลาประดิษฐ์ (Artificial Fish Swarm

Algorithm) 2.5% และอื่น ๆ รวม 17.2% [11] เป็นต้น จะเห็นว่าวิธิการที่ยอดนิยมได้แก่การหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค ซึ่งเป็นวิธีที่เข้าใจง่ายและใช้งานง่าย เนื่องจากมีพารามิเตอร์ในการคำนวณน้อย และเนื่องจากขั้นตอนวิธีนี้สามารถสร้างกลุ่มประชากรได้จำนวนมาก มีกระบวนการดัดแปลงประชากรตามลักษณะขั้นตอนวิธีเชิงวิวัฒนาการ ซึ่งมักส่งผลต่อรูปแบบคำตอบ นอกจากนี้ปัญหาการจัดกลุ่มข้อมูลยังสามารถมองเป็นปัญหาของการหาคำตอบที่ทำให้สมาชิกกลุ่มมีความเหมือนหรือคล้ายคลึงกันมากที่สุด ผู้วิจัยจึงมีแนวคิดที่จะศึกษาและพัฒนาวิธีการใหม่สำหรับจัดกลุ่มข้อมูลยีนส์ในการวิเคราะห์ Pathway ด้วยขั้นตอนวิธีเชิงวิวัฒนาการแบบการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค เพื่อได้เครื่องมือค้นหากลุ่มของ Pathway ในรูปแบบที่หลากหลายมากยิ่งขึ้น

1.2 วัตถุประสงค์

เพื่อศึกษาและพัฒนาวิธีการสร้างตัวจัดกลุ่มแบบการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคสำหรับจัดกลุ่มข้อมูลยีนส์ในการวิเคราะห์ Pathway (Pathway analysis) โดยทำให้ผลของการจัดกลุ่มมีความหลากหลายของคำตอบและให้คุณสมบัติของการเป็นกลุ่มดีขึ้น อาทิ ความกะทัดรัดของกลุ่มข้อมูล และความห่างกันของกลุ่มข้อมูล เป็นต้น

1.3 สมมติฐานการวิจัย

- ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคสามารถใช้สร้างตัวจัดกลุ่มข้อมูลสำหรับการวิเคราะห์ Pathway ได้
- การจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคให้ผลการจัดกลุ่มเป็นรูปแบบคำตอบที่หลากหลาย และทำให้คุณสมบัติการเป็นกลุ่มของข้อมูลดีขึ้น เช่น ความกะทัดรัดของกลุ่มข้อมูลมีค่าน้อยที่สุด และระยะห่างระหว่างกลุ่มของข้อมูลมีค่ามากที่สุด
- การจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคสามารถเลี่ยงการเข้าสู่สภาวะสูงสุดหรือต่ำสุดท้องถิ่น

1.4 ขอบเขตการทำงาน

- การศึกษานี้ใช้ข้อมูลยีนส์ที่ได้จากเทคนิคไมโครอะเรย์เท่านั้น และเป็นการวิเคราะห์ข้อมูลเฉพาะจุดประสงค์ได้แก่การวิเคราะห์ Pathway (Pathway analysis) ซึ่งเป็นขั้นตอนที่เกิดขึ้นจากกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์

- การศึกษานี้ศึกษาจากข้อมูลการแสดงออกของยีนส์ในเซลล์เม็ดเลือดขาวชนิดโมโนนิวไคลด์แบบกลีบเดียวจากผู้ป่วยโรคข้ออักเสบรูมาตอยด์ (Rheumatoid arthritis GSE15573) เท่านั้น โดยใช้ข้อมูลจากเครื่องมือ pathfindR.data
- เป็นการศึกษาการสร้างตัวจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีเชิงวิวัฒนาการแบบการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle Swarm Optimization)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้วิธีการใหม่สำหรับการจัดกลุ่มข้อมูลในการวิเคราะห์ Pathway โดยทำให้ได้คำตอบที่หลากหลายเพื่อเป็นตัวเลือกสำหรับผู้ศึกษาการแสดงของยีนส์ และทำให้คุณสมบัติความเป็นกลุ่มข้อมูลดีขึ้น ซึ่งเป็นประโยชน์เฉพาะทาง และอาจทำให้เป็นทางเลือกใหม่แก่ผู้ใช้งาน

บทที่ 2

ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้เป็นการนำเสนอทฤษฎี การศึกษาและเรื่องราวที่เกี่ยวข้องสำหรับวิทยานิพนธ์นี้ ซึ่งได้แบ่งรายละเอียดดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การแบ่งกลุ่มข้อมูล

การวิเคราะห์แบ่งกลุ่มข้อมูล (Cluster analysis) เป็นปัญหาที่ใช้ขั้นตอนวิธีแก้ปัญหาการจัดกลุ่มข้อมูล (Clustering algorithm) ซึ่งเป็นขั้นตอนวิธีการเรียนของเครื่องแบบไม่มีผู้สอน หรือไม่มีผลเฉลย (Unsupervised Learning) ทำหน้าที่แบ่งหรือจัดกลุ่มข้อมูลออกเป็นกลุ่ม ๆ (Cluster) แต่เนื่องจากเป็นการจัดกลุ่มโดยไม่มีผลเฉลย ดังนั้นคุณสมบัติของการเป็นกลุ่มของข้อมูลเดียวกันจึงมักต้องการความเหมือนกันอย่างมากสำหรับข้อมูลที่อยู่กลุ่มเดียวกัน และเมื่ออยู่ต่างกลุ่มกันจะต้องมีความแตกต่างกันมากอีกด้วย เทคนิคที่ใช้วิเคราะห์การแบ่งกลุ่มหลัก ๆ มีหลายวิธี เช่น การหาคุณลักษณะสำคัญของข้อมูล (Feature selection method) อาจเพื่อลดทอนข้อมูลที่ไม่เกี่ยวข้องบางประเภทออกจากข้อมูลที่มีประโยชน์แท้จริงต่อการจัดกลุ่ม การลดมิติข้อมูล (Dimensionality reduction) เพื่อลดปัญหามิติซับซ้อนของข้อมูลส่วนเกินหรือกำจัดข้อมูลที่ไม่เกี่ยวข้องกับส่วนที่สำคัญเพื่อให้อัลกอริธึมทำงานได้อย่างมีประสิทธิภาพมากขึ้นก็ถือว่าเป็นการกระทำมีลักษณะของการจัดการกลุ่มข้อมูล หรือขั้นตอนวิธีการจัดกลุ่มแบบอาศัยระยะทาง (Distance-Based algorithm) ซึ่งเป็นขั้นตอนวิธีที่ถูกนำมาใช้บ่อยครั้งเพื่อจัดกลุ่มข้อมูล เนื่องจากมีโครงสร้างไม่ซับซ้อน และง่ายต่อการใช้งาน เป็นต้น การวิเคราะห์กลุ่มข้อมูลมักให้ประโยชน์ในการใช้งานต่าง ๆ ที่สามารถพบเจอได้ในปัจจุบัน อาทิ ใช้ในขั้นตอนการทำเหมืองข้อมูลเพื่อรวมข้อมูลเป็นกลุ่ม ใช้ร่วมกับระบบแนะนำเพื่อแนะนำสิ่งที่น่าสนใจที่คล้ายคลึงกับสิ่งที่ผู้ใช้ชื่นชอบ (Recommender system) ตลอดจนสามารถใช้วิเคราะห์ข้อมูลแบบพลวัต เช่น การพูดถึงมากที่สุดในโซเชียลเน็ตเวิร์ก เป็นต้น [12] กระบวนการวิเคราะห์ข้อมูลที่มีหลายรูปแบบ สามารถเลือกพิจารณาการใช้งานนจากประเภทของข้อมูลที่ต้องการจัดกลุ่ม เช่น ข้อมูลเชิงกลุ่ม (Categorical data) ข้อมูลตัวเลข (Numerical data) ข้อมูลแบบผสม (Mixed data) หรือข้อมูลแบบ Time series เป็นต้น หรืออาจพิจารณาเลือกใช้ขั้นตอนวิธีจัดกลุ่มจากเทคนิคในการจัดกลุ่มเป็นเกณฑ์ ตัวอย่างเช่น การจัดกลุ่มโดยใช้ระยะทาง (Distance-based algorithm) การจัดกลุ่มโดยความน่าเป็นและการกระจายตัวของข้อมูล (Probabilistic and

generative methods) การจัดกลุ่มเทคนิคกราฟ (Graph theoretical based-method) เป็นต้น ซึ่งขั้นตอนวิธีจัดกลุ่มต่าง ๆ ที่เลือกใช้จะให้ข้อดี ข้อเสีย หรือถึงระยะเวลาในการเรียกใช้งานชุดคำสั่ง (Time complexity) และปริมาณหน่วยความจำ (Memory requirement) ที่ใช้แตกต่างกันอีกด้วย

2.1.1.1 การจัดกลุ่มข้อมูลโดยใช้ระยะทาง (Distance-Based Clustering)

การจัดกลุ่มข้อมูลโดยใช้ระยะทางเป็นเทคนิคที่เข้าใจง่าย และมีความง่ายต่อการใช้งานชุดคำสั่ง จึงทำให้เป็นเทคนิคที่มักมีการใช้งานบ่อยครั้งในหลายกลุ่มสาขาวิชาต่าง ๆ เนื่องจากสามารถใช้ตัวจัดกลุ่มเทคนิคนี้ได้ทุกประเภทของข้อมูลเข้า และจำเป็นต้องกำหนดค่า k หรือจำนวนกลุ่มคำตอบที่ต้องการให้ชุดคำสั่งเพื่อหาคำตอบ ตัวชุดคำสั่งจะมีการจัดกลุ่มด้วยระยะทางระหว่างข้อมูลเป็นหลัก ทำให้ได้ผลลัพธ์เป็นข้อมูลที่ถูกแบ่งออกเป็นส่วน ๆ เสมือนการกั้นเขตให้แก่กลุ่มข้อมูล หรือเรียกอีกชื่อหนึ่งว่า Partition clustering จะมีการใช้ฟังก์ชันระยะทางเพื่อจำแนกข้อมูลออกเป็นกลุ่มต่าง ๆ นอกจากนี้ยังสามารถมองเป็นปัญหาที่มีวัตถุประสงค์เพื่อหาค่าเหมาะสมของฟังก์ชันจุดประสงค์ที่ใช้กระบวนการแก้ปัญหาแบบทำซ้ำเป็นรอบ ๆ เพื่อพยายามค้นหาคำตอบที่ทำให้กลุ่มข้อมูลแต่ละกลุ่มมีคุณภาพมาก

2.1.1.2 การวัดระยะทางของข้อมูล (Distance measurement)

การวัดระยะทางของข้อมูลเข้าคือการคำนวณเพื่อหาค่าระยะทางของข้อมูลแต่ละตัว อาจใช้การคำนวณเพื่อหาฟังก์ชันระยะทาง หรือ Distance function หรือสามารถมองเป็นการหาค่าความแตกต่างระหว่างข้อมูลแต่ละตัวได้เช่นกัน ในกรณีที่เป็นการหาความแตกต่างระหว่างข้อมูล จะเรียกว่า Dissimilarity function หากเป็นการหาความเหมือนกันระหว่างข้อมูล จะเรียกว่า Similarity function ซึ่งกระบวนการนี้ถือเป็นสิ่งจำเป็นสำหรับการใช้งานร่วมกับชุดคำสั่งการจัดกลุ่มข้อมูลด้วยเทคนิคระยะทาง และส่งผลต่อผลลัพธ์การแบ่งกลุ่มข้อมูลให้เป็นข้อมูลขาออกด้วย จึงมีวิธีการและเทคนิคที่หลากหลายใช้สำหรับการสร้างฟังก์ชันระยะทาง ความแตกต่างระหว่างข้อมูล และความเหมือนกันระหว่างข้อมูล ดังตัวอย่างฟังก์ชันระยะทาง และฟังก์ชันความเหมือนกันระหว่างข้อมูลที่พบการใช้งานบ่อย ๆ ดังต่อไปนี้

ตัวอย่างฟังก์ชันระยะทาง (Distance function) ที่พบบ่อย ได้แก่

- ระยะห่างแมนฮัตตัน (Manhattan distance) เป็นการวัดระยะห่างระหว่างจุด 2 จุด ดังสมการที่ 1

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \qquad (1)$$

เมื่อ x_i และ y_i เป็นสมาชิกในกลุ่มข้อมูล

- ระยะทางแบบยูคลิด (Euclidean distance):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

เมื่อ x_i และ y_i เป็นสมาชิกในกลุ่มข้อมูล

ตัวอย่างการวัดความเหมือนกันของข้อมูล (Similarity measure) ที่พบบ่อย ได้แก่

- Cosine similarity เป็นการคำนวณเพื่อหาความเหมือนกันของข้อมูล โดยกำหนดให้ x_i และ y_i เป็นสมาชิกในกลุ่มข้อมูล

$$\cos(x_i, y_i) = \frac{x_i \cdot y_i}{\|x_i\|\|y_i\|} \quad (3)$$

- Jaccard similarity เป็นการคำนวณหาความเหมือนกันของข้อมูลอีกประเภทหนึ่ง ที่เป็นการดำเนินการของเซ็ตข้อมูล A และ B

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

เมื่อ A และ B เป็นเซ็ตที่ประกอบไปด้วยข้อมูลสมาชิก

2.1.1.3 ตัวอย่างการจัดกลุ่มข้อมูลโดยใช้ระยะทาง

ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบใช้ระยะทาง (Distance-based clustering) สามารถจำแนกได้ 2 ประเภทตามลักษณะการทำงานของชุดคำสั่ง ได้แก่

แบบราบ (Flat)

มีลักษณะในการแบ่งกลุ่มข้อมูลออกเป็นหลาย ๆ กลุ่มในครั้งเดียว โดยเริ่มจากการกำหนดจำนวนกลุ่มที่ต้องการให้ชุดคำสั่งทำการจัดกลุ่ม ตัวอย่างเช่น การจัดข้อมูลแบบเคมีน (K-means clustering) มีลักษณะการทำงานโดยเริ่มจากการกำหนดค่า k หรือจำนวนกลุ่มที่ต้องการแบ่งข้อมูลออกเป็นผลลัพธ์ นอกจากนี้ยังเพื่อใช้สร้างจุดศูนย์กลางจำลองเริ่มต้นของกลุ่มข้อมูลจำนวน k ตำแหน่งในปริภูมิคำตอบ เรียกว่าเซนทรอยด์ (Centroid) ขั้นตอนวิธีการของเคมีนจะพยายามคำนวณหาข้อมูลตัวที่ใกล้จุดเซนทรอยด์สมมติเพื่อรวมเป็นกลุ่มเดียวกันจำนวน k กลุ่มด้วยฟังก์ชันระยะทางหรือการวัดความเหมือนกันของข้อมูลที่เลือกใช้ โดยการเลือกการเป็นกลุ่มเดียวกันจะเป็น

การหาฟังก์ชันระยะทางที่น้อยที่สุดซึ่งหมายถึงข้อมูลตัวนั้น ๆ อยู่ใกล้เซนทรอยด์ดังกล่างที่สุด และทำการคำนวณค่าเฉลี่ย (mean) ระยะห่างของสมาชิกในกลุ่มเพื่อเป็นปรับตำแหน่งเซนทรอยด์ใหม่ในการค้นหารอบถัดไป ชุดคำสั่งแบบเคมีนจะมีการทำซ้ำจนกว่าผลรวมของผลต่างระหว่างข้อมูลแต่ละตัวและจุดเซนทรอยด์ประจำกลุ่มจะมีค่าน้อยที่สุด หรือมีความคลาดเคลื่อนน้อยที่สุด โดยทั่วไปมักใช้ผลรวมของความคลาดเคลื่อนกำลังสอง (Sum of Squared Errors; SSE) เป็นสมการจุดประสงค์ ดังสมการที่ 5 เมื่อ C เป็นเซ็ตของเซนทรอยด์ที่มีสมาชิกจำนวน k ตัว หรือ $C = \{C_1, C_2, \ldots, C_K\}$

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} \|x_i - c_k\|^2 \quad (5)$$

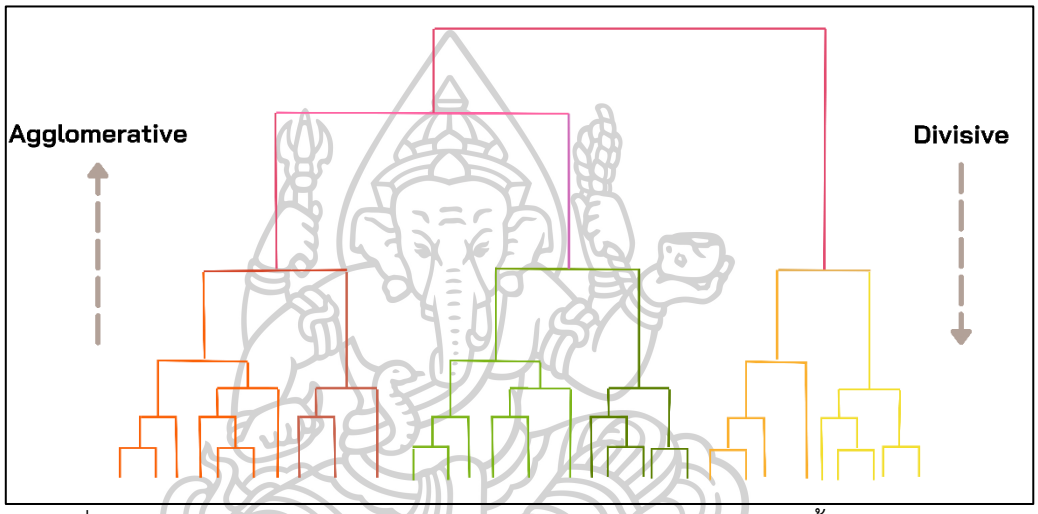
$$c_k = \frac{x_i \epsilon C_k}{|C_k|} \quad (6)$$

อย่างไรก็ตาม หากพิจารณาขั้นตอนการทำงานของการแบ่งกลุ่มแบบเคมีนจะพบว่ามีลักษณะการทำงานเสมือนพยายามหาค่าน้อยที่สุดของสมการที่ 5 และคำตอบซึ่งเป็นผลลัพธ์ของการจัดกลุ่มขึ้นอยู่กับการสุ่มเลือกเซนทรอยด์ในปริภูมิคำตอบตั้งแต่ขั้นตอนเริ่มต้น นอกจากนี้ค่ากำหนดค่า k ตั้งแต่กระบวนการแรกทั้งที่อาจไม่ทราบค่าที่เหมาะสมอาจเป็นผลที่ทำให้ผลลัพธ์ของการจัดกลุ่มไม่ได้ให้ผลที่ดี ผู้วิจัยโดยทั่วไปมักจะเลือกใช้ขั้นตอนวิธีการอื่น ๆ เพื่อช่วยทำให้ประสิทธิภาพของการแบ่งกลุ่มข้อมูลดีขึ้น เช่น การเลือกใช้ขั้นตอนวิธีการเชิงวิวัฒนาการร่วมด้วย การพิจารณาคำตอบด้วยเทคนิคอื่น หรือพิจารณาการใช้ขั้นตอนการแบ่งกลุ่มประเภทอื่น เป็นต้น

แบบลำดับขั้น (Hierarchical)

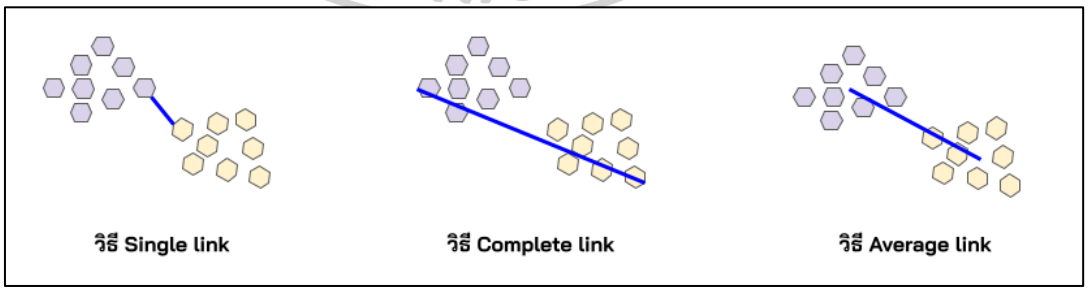
มีลักษณะในการแบ่งกลุ่มข้อมูลเป็นลำดับขั้น คล้ายลักษณะของต้นไม้หรืออนุกรมวิธานของสิ่งมีชีวิต ที่มีฐานกว้างเป็นกลุ่มข้อมูลหลายกลุ่ม และเหนือขึ้นไปจนสุดยอดซึ่งมีจำนวนกลุ่มข้อมูลน้อยกว่าฐาน เรียกว่า Agglomerative hierarchical clustering หรือการจัดกลุ่มแบบลำดับขั้นจากล่างขึ้นบน และการจัดกลุ่มแบบลำดับขั้นจากบนลงล่าง หรือ Divisive hierarchical clustering ซึ่งมีลักษณะกลับกันกับแบบ Agglomerative ดังรูปที่ 2.1 การจัดกลุ่มประเภทลำดับขั้นนี้แบบดั้งเดิม ไม่จำเป็นจะต้องกำหนดค่า k ให้ชุดคำสั่งในตอนเริ่มต้นเพื่อแบ่งข้อมูลออกเป็นกลุ่ม ๆ แต่ขั้นตอนวิธีการของเทคนิคนี้จะทำการหาข้อมูลตัวที่อยู่ใกล้หรือมีความสัมพันธ์ใกล้สุดกับตัวมันเองไปเรื่อย ๆ เป็นลำดับขั้นที่สูงขึ้นไปจนมีลักษณะคล้ายต้นไม้ เรียกว่า เดนโดแกรม (Dendrogram) ดังนั้นผลลัพธ์ที่ได้

จากเทคนิคนี้จะไม่สามารถระบุจำนวนกลุ่มที่เป็นคำตอบในการแบ่งกลุ่มข้อมูลได้ภายหลัง ซึ่งผู้ใช้งานจะต้องมีวิธีสำหรับการหาค่าจำนวนกลุ่มที่เหมาะสมกับข้อมูลเข้าที่มี ทั้งนี้ในกระบวนการที่ข้อมูลแต่ละตัวพยายามค้นหาหรือรวมข้อมูลตัวอื่น ๆ ที่อยู่ใกล้หรือมีความสัมพันธ์ใกล้เคียงกันเข้าไว้ด้วยกัน (Merge) โดยพิจารณาจากใช้การคำนวณฟังก์ชันระยะทาง (Distance function) หรือการวัดความเหมือนกันของข้อมล (Similarity measure) การรวมกลุ่มเข้าไว้ด้วยกันของข้อมูลมีหลายรูปแบบอาทิ วิธี Single link วิธี Complete link และวิธี Average link เป็นต้น ดังรูปที่ 2.2



รูปที่ 2.1 ตัวอย่างเปรียบระหว่างการแบ่งกลุ่มข้อมูลแบบลำดับขั้นประเภท Agglomerative และ Divisive

ที่มา. จาก Identification of asthma subtypes using clustering methodologies, โดย Matea Deliu., 2016, *Pulmonary Therapy*, *2016*(2), หน้า 22.



รูปที่ 2.2 วิธีการวัดความเหมือนกันของข้อมูลสำหรับการจัดกลุ่มประเภทลำดับขั้น

ที่มา. จาก Inference of a human brain fiber bundle atlas from high angular resolution diffusion imaging, โดย Pamela Beatriz Guavara Alvez., 2011, *HAL open science*, *2011*, หน้า 84.

2.1.1.4 คุณสมบัติความเป็นกลุ่มข้อมูล (Cluster validation)

การวิเคราะห์กลุ่มข้อมูล หรือ Cluster analysis ทำให้ได้ผลลัพธ์คือกลุ่มข้อมูลที่มีความสัมพันธ์กันอยู่กลุ่มเดียวกัน และข้อมูลที่ไม่เกี่ยวข้องกันจะอยู่ต่างกลุ่มกัน ดังนั้นคุณสมบัติความเป็นกลุ่มเดียวกันของข้อมูล อาจมองในแง่ของความกะทัดรัดของกลุ่มข้อมูลนั้น ๆ ความเหมือนหรือคล้ายคลึงกันของกลุ่มข้อมูล ความอยู่ห่างจากกัน เป็นต้น โดยทั่วไปการวิเคราะห์กลุ่มข้อมูลซึ่งเป็นการใช้ขั้นตอนวิธีการเรียนรู้เครื่องแบบไม่มีผู้สอน จะทำให้ได้ผลลัพธ์ซึ่งไม่สามารถระบุได้ว่าถูกต้องจริงหรือไม่ ดังนั้นจึงจำเป็นที่จะต้องมีการตรวจสอบคุณลักษณะของการเป็นกลุ่มข้อมูลหลังการแบ่งกลุ่มเพื่อยืนยันคำตอบก่อนการนำผลเข้าสู่กระบวนการศึกษาอื่นถัดไป สำหรับปัญหาการจัดกลุ่มข้อมูล จึงมีกระบวนการที่ช่วยวิเคราะห์คุณลักษณะของคำตอบ ได้แก่ การตรวจสอบกลุ่มข้อมูล (Cluster validation) สามารถแบ่งการตรวจสอบได้ 2 ประเภท ดังนี้

2.1.1.4.1 การตรวจสอบกลุ่มข้อมูลแบบภายใน (Internal clustering validation)

กระบวนการนี้เป็นการประเมินคุณลักษณะของการเป็นกลุ่มข้อมูลจากข้อมูลที่ได้ทำการแบ่งกลุ่ม และไม่มีการนำข้อมูลภายนอกร่วมเปรียบเทียบ หลักการสำคัญสำหรับการตรวจสอบแบบกระบวนนี้จะมี 2 หลักการ ดังนี้

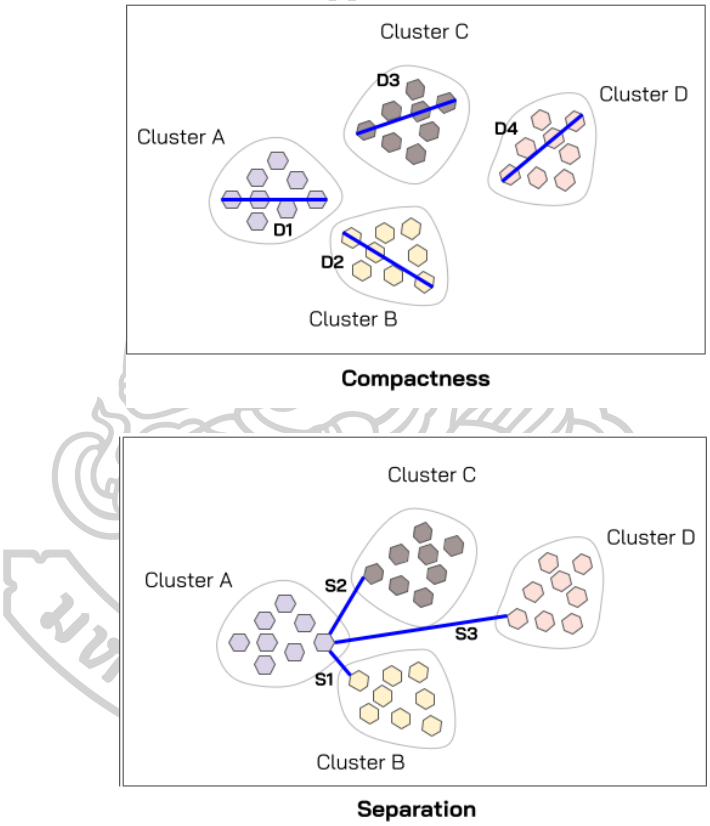
- การพิจารณาความกะทัดรัดของกลุ่มข้อมูล (Compactness) อันเนื่องมาจากแนวคิดที่ข้อมูลที่จะเป็นกลุ่มเดียวกันควรมีความหนาแน่นของประชากรในกลุ่ม ซึ่งอาจพิจารณาได้จาก ความยาวเส้นผ่านศูนย์กลางของกลุ่มข้อมูล ความแปรปรวนของแต่ละสมาชิกในกลุ่ม ค่าเฉลี่ยระหว่างระยะห่างของข้อมูลในกลุ่ม เป็นต้น
- การพิจารณาความแยกจากกันของกลุ่มข้อมูล (Separation) อันเนื่องมาจากแนวคิดที่ว่าถ้าหากเป็นข้อมูลคนละประเภทหรือไม่เกี่ยวข้องกันจะต้องอยู่ห่างกันมากที่สุด อาจพิจารณาจาก ระยะห่างระหว่างกลุ่มข้อมูลที่มีค่าน้อยที่สุด ดังรูปที่ 2.3

นอกจากนี้ตัวอย่างการประเมินค่าดังกล่าว ยังสามารถใช้ดัชนีชี้วัดที่สำหรับทำการประเมินกลุ่มข้อมูลแบบภายใน เช่น Dunn index เป็นอัตราส่วนระหว่างระยะห่างระหว่างกลุ่มที่มีค่าน้อยที่สุดต่อเส้นผ่านศูนย์ของกลุ่มที่มีค่ามากที่สุด ดังสมการที่ 7 หรือ Silhouette index เป็นสัมประสิทธิ์วัดความเหมือนกันของสมาชิกภายในกลุ่ม เปรียบเทียบนอกกลุ่ม ดังสมการที่ 8

- Dunn's index (DI): $$DI = \frac{min(Separation)}{max(Compactness)} \quad (7)$$

- Silhouette index (S): $$S(i) = \frac{b(i) - a(i)}{max\{a(i),\ b(i)\}} \quad (8)$$

โดยที่ $a(i)$ คือค่าเฉลี่ยระยะห่างระหว่างข้อมูล i และสมาชิกในกลุ่มเดียวกัน และ $b(i)$ คือค่าเฉลี่ยระยะห่างของสมาชิก i ต่อสมาชิกกลุ่มอื่น



รูปที่ 2.3 แสดงตัวอย่างการตรวจสอบกลุ่มข้อมูลแบบภายใน จำนวน 4 กลุ่ม ได้แก่ การวัดความยาวเส้นผ่านศูนย์กลางของกลุ่มข้อมูล (บน) และการวัดระยะห่างระหว่างกลุ่มที่สั้นที่สุดเฉพาะกลุ่มข้อมูล A (ล่าง)

2.1.1.4.2 การตรวจสอบกลุ่มข้อมูลแบบภายนอก (External clustering validation)

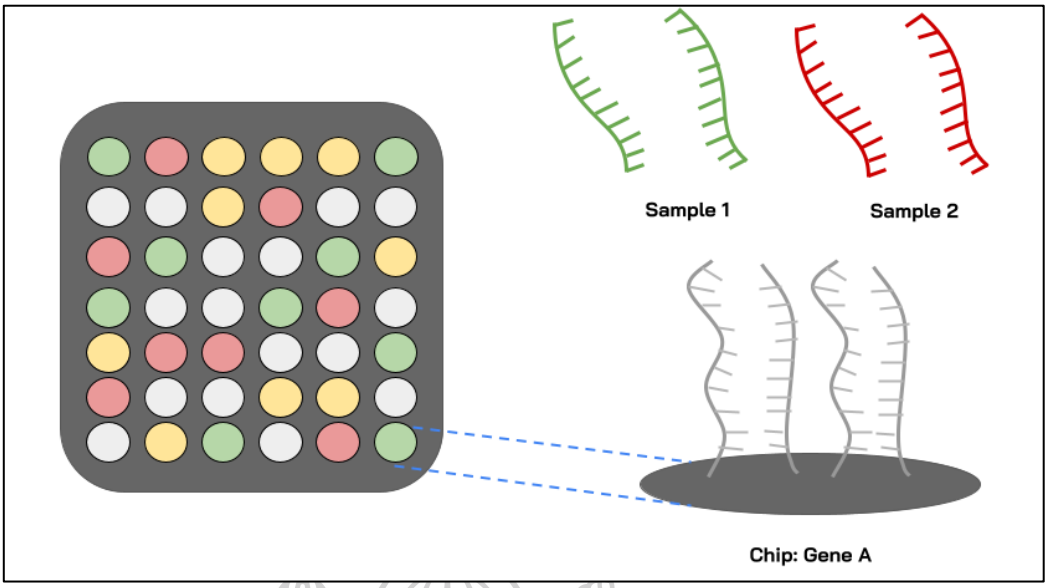
การตรวจสอบกลุ่มข้อมูลแบบภายนอก เป็นกระบวนการเพื่อทดสอบคำตอบที่ได้จากขั้นตอนวิธีที่เลือกใช้กับวิธีการอื่น ๆ ที่ไม่ใช่เปรียบเทียบจากชุดคำตอบที่ได้มา ซึ่งคือการนำผลไปเปรียบเทียบกับผลเฉลยจริงที่อาจมีอยู่แล้ว หรือ เป็นผลเฉลยจากกระบวนการในขั้นตอนต่อไป วิธีสำหรับการ

ตรวจสอบกลุ่มข้อมูลแบบภายนอกนี้มีหลายวิธีการ ตัวอย่างเช่น Rand statistic ใช้เพื่ออธิบายความสอดคล้องกันของคำตอบ เป็นต้น

2.1.2 ไมโครอะเรย์และการวิเคราะห์ความสำคัญของกลุ่มยีนส์

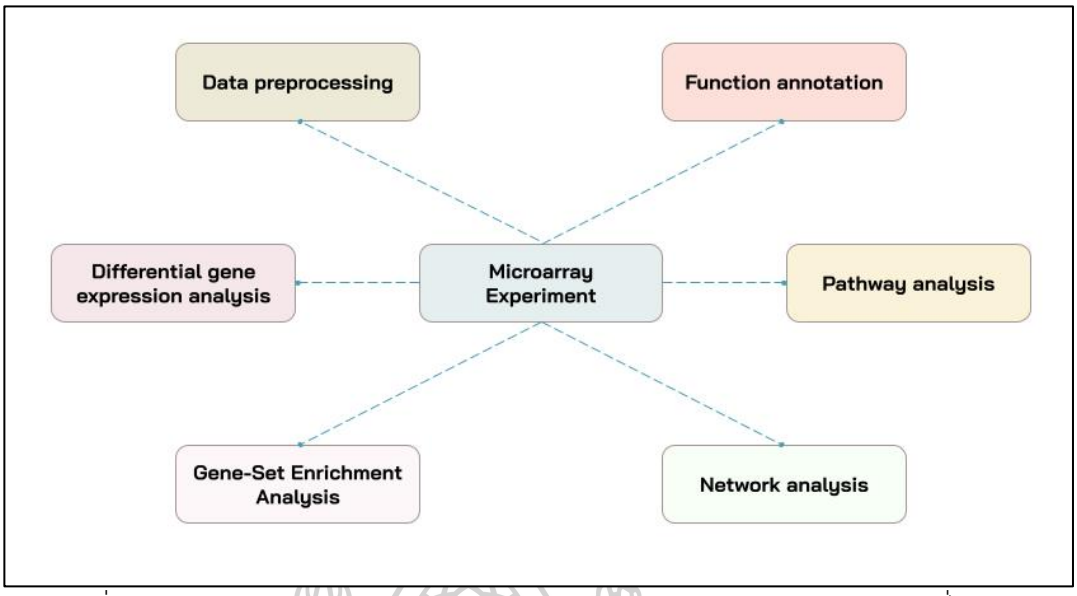
2.1.2.1 เทคนิคไมโครอะเรย์

เป็นเทคนิคในห้องปฏิบัติการทางชีววิทยาที่ใช้สำหรับศึกษาความบกพร่อง ความผิดปกติ รวมถึงความไม่สมบูรณ์ของโรคพันธุกรรม โดยสามารถศึกษาได้ลึกถึงระดับยีนส์ ซึ่งมีข้อดีเมื่อเทียบกับเทคนิคเซลล์คือสิ่งส่งตรวจสำหรับเทคนิคไมโครอะเรย์นี้ ไม่จำเป็นต้องมีชีวิตอยู่ หรือผ่านกระบวนการเลี้ยงเซลล์ก่อนจึงจะทำการศึกษาได้ อีกทั้งยังให้ผลการศึกษาที่มีความละเอียดเป็นอย่างสูง การศึกษาด้วยเทคนิคไมโครอะเรย์จะมีการใช้สิ่งส่งตรวจ หรือเรียกว่า Sample จากโรคที่ผู้วิจัยสนใจ แบ่งเป็น 2 samples ได้แก่ DNA จากผู้ป่วยที่สุขภาพดีเป็น Sample 1 และ DNA จากผู้ป่วยได้รับการวินิจฉัยว่าเป็นโรคนั้น ๆ เป็น Sample 2 ทดลองโดยนำสิ่งส่งตรวจทั้งสองทำการไฮบริดไดซ์ซิ่ง (Hybridization) ลงบนถาดที่มีลักษณะเป็นหลุม เรียกว่า Chip ในแต่ละชิพ ซึ่งจะมี Oligonucleotide Probe จากแต่ละยีนส์ที่อาจสังเคราะห์มาจากสารอื่น ๆ มีความจำเพาะต่อการศึกษาโรคนั้น ๆ ดังรูปที่ 2.4 นำไมโครอะเรย์ที่ทำการไฮบริดไดซ์ซิ่งไปแสกนเพื่อทำการหาค่าความเข้มของสารฟลูออเรสเซนต์เทียบกับ DNA ปกติ หรือเข้าสู่กระบวนการวิเคราะห์ข้อมูลในลำดับถัดไป การแปลผลการทดลองไมโครอะเรย์จะได้ว่าหากชิพยีนส์ A แสดงผลสีเขียวจากการใส่ Sample 1 หมายถึงยีนส์ A สามารถแสดงผลใน Sample 1 แต่ไม่แสดงผลใน Sample 2 หากชิพยีนส์ B แสดงผลสีแดง หมายถึงยีนส์ B สามารถส่งผลใน Sample 2 และไม่แสดงผลต่อ Sample 1 หากชิพยีนส์ใดๆ แสดงผลสีเหลือง หมายถึงยีนส์นั้นๆ ส่งผลต่อ Sample 1 และ 2 เป็นต้น ด้วยเหตุนี้เทคนิคไมโครอะเรย์จึงสามารถศึกษาผลการแสดงออกในระดับยีนส์หรือ DNA ได้ นำไปสู่การค้นพบยาที่สามารถรักษาโรค หรือพบสาเหตุของการเกิดโรคได้เป็นอย่างดีในระดับห้องปฏิบัติการ



รูปที่ 2.4 แสดงภาพรวมและส่วนประกอบอุปกรณ์เทคนิคไมโครอะเรย์

อย่างไรก็ตามการทำการศึกษาการแสดงออกในระดับยีนส์ด้วยเทคนิคไมโครอะเรย์ เป็นจุดเริ่มต้นของการสร้างข้อมูลขนาดใหญ่ เพื่อนำไปศึกษาเฉพาะทางหรือเฉพาะความสนใจต่อในขั้นตอนถัดไป ซึ่งกระบวนการหลังจากนี้เป็นการศึกษาที่มักอาศัยการคำนวณจากคอมพิวเตอร์ หรือมีวิทยาการทางด้านคอมพิวเตอร์ (Computational biology) เข้ามามีส่วนร่วมเพื่อช่วยอำนวยการวิเคราะห์เชิงตัวเลขให้แก่นักวิจัยทางชีววิทยา ดังจะเห็นจากปัจจุบันมีการวิเคราะห์ข้อมูลที่หลากหลายวัตถุประสงค์เพื่อใช้สำหรับการศึกษาการแสดงออกของยีนส์อันเป็นผลมาจากเทคนิคไมโครอะเรย์ อาทิ เครื่องมือสำหรับการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA) เครื่องมือสำหรับการวิเคราะห์ (Pathway analysis) เป็นต้น ตัวอย่างดังกล่าวเป็นเพียงส่วนหนึ่งของเครื่องมือที่ผลิตขึ้นให้ตอบโจทย์วัตถุประสงค์ที่หลากหลายสำหรับผู้ใช้ และยังมีอีกการศึกษาอื่น ๆ ยกตัวอย่างดังรูปที่ 2.5



รูปที่ 2.5 แสดงตัวอย่างการศึกษาตามวัตถุประสงค์ต่าง ๆ สำหรับข้อมูลที่ได้จากการทดลองเทคนิคไมโครอะเรย์

2.1.2.2 การวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA)

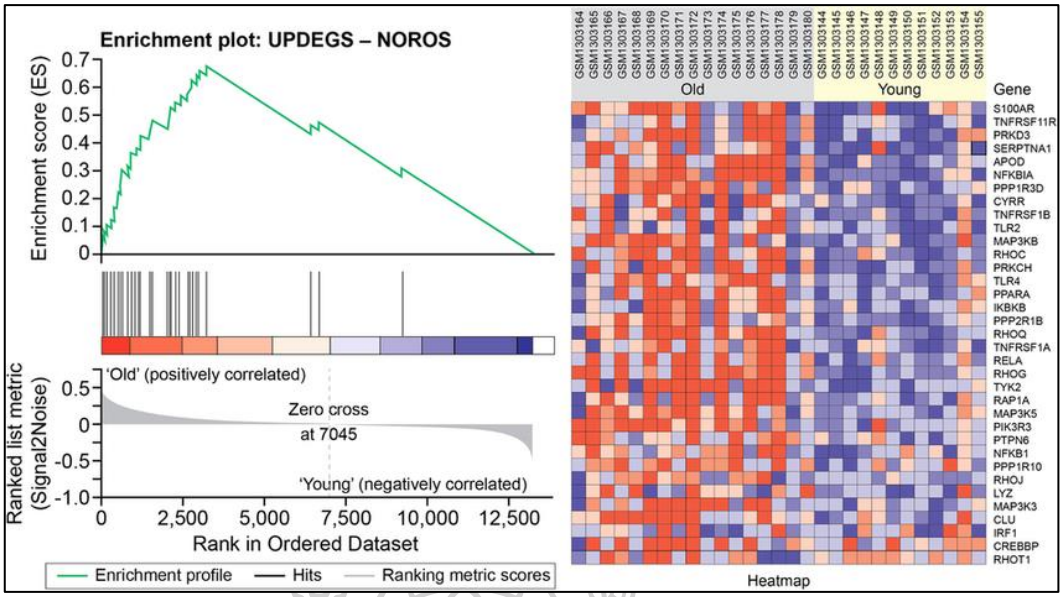
ภายหลังจากการทดลองไมโครอะเรย์ในห้องปฏิบัติการทดลองจะได้ข้อมูลการแสดงออกของยีนส์จำนวนมากเพื่อวิเคราะห์ผลหาคำตอบตามสมมติฐานของผู้วิจัย กระบวนการวิเคราะห์ความสัมพันธ์ของกลุ่มยีนส์เป็นกระบวนการหนึ่งเพื่อศึกษายีนส์ที่มีความโดดเด่นและสำคัญต่อการแสดงออกของโรคนั้น ๆ เนื่องจากข้อมูลการแสดงออกของยีนส์มีจำนวนมาก ในขั้นตอนนี้จึงมีการใช้การประมวลผลทางคอมพิวเตอร์เพื่อช่วยหาคำตอบ เช่น หาสาเหตุของการเกิดโรคบางประการโดยวิเคราะห์จากกลุ่มยีนส์ที่ชำรุดไปเมื่อทำไมโครอะเรย์ ตรวจสอบโครงสร้างยาที่เหมาะสม เนื่องจากชิพยีนส์ที่แสดงออกโรคนั้นแสดงผลแบบปกติ เป็นต้น

ในปัจจุบันมีเครื่องมือโปรแกรมมากมายสำหรับนักชีววิทยาหรือผู้ที่เกี่ยวข้องในการศึกษาเลือกใช้เพื่อค้นหาคำตอบดังกล่าว อาทิ GSEA software [13] โปรแกรมสำหรับการวิเคราะห์กลุ่มยีนส์ที่สำคัญและช่วยแสดงผลข้อมูล Enrichr เว็บสำหรับวิเคราะห์ร่วมกับการเรียกใช้ค่า Protein-protein interaction และจัดกลุ่มความสัมพันธ์ของกลุ่มยีนส์ DAVID web server ที่นอกเหนือจากการวิเคราะห์ความสำคัญของยีนส์และยังสามารถทำการเติมคุณลักษณะอื่น ๆ เข้าไว้ด้วยกัน เป็นต้น ยังมีเครื่องมืออีกมากมายที่มีความสามารถแตกต่างกันออกไป ทั้งนี้เพื่อคุณประโยชน์ที่หลากหลายและตอบวัตถุประสงค์ต่อผู้วิจัยได้หลายรูปแบบ กระบวนการวิเคราะห์ยีนส์ที่สำคัญมีขั้นตอนย่อย ๆ

ได้แก่ การเตรียมข้อมูลก่อนการวิเคราะห์ เพื่อตัดข้อมูลบางส่วนที่ไม่ได้ตามวัตถุประสงค์ออกไป การคำนวณค่าความแตกต่างระหว่างการแสดงออกของยีนส์และจัดลำดับความสำคัญ การคัดเลือกกลุ่มยีนส์ที่สำคัญโดยอาศัยข้อมูลการทดลองจากฐานข้อมูลต่าง ๆ การคำนวณหาค่าความสำคัญของยีนส์ (Enrichment score) และสรุปกลุ่มยีนส์ที่สำคัญเพื่อแสดงผล จะเห็นว่ากระบวนการนี้มีลักษณะการคัดกรองข้อมูลจากเหมืองยีนส์จำนวนมหาศาลให้กลายเป็นกลุ่มยีนส์ที่มีขนาดใหญ่ขึ้นและมีความโดดเด่นตามจุดประสงค์ของผู้ทำการศึกษา ดังรูปที่ 2.6 ผลการวิเคราะห์ความสำคัญของกลุ่มยีนส์ GSE 53890 จากการทดลองด้วยเทคนิคไมโครอะเรยด้วย GSEA software โดยแสดงผลกราฟค่าความสำคัญของยีนส์และการจัดลำดับ รวมถึงแผนภูมิความร้อน (Heatmap) แสดงการแสดงออกของแต่ละยีนส์เมื่อกระทำการไฮบริดไดซ์ซิ่งกับ Sample

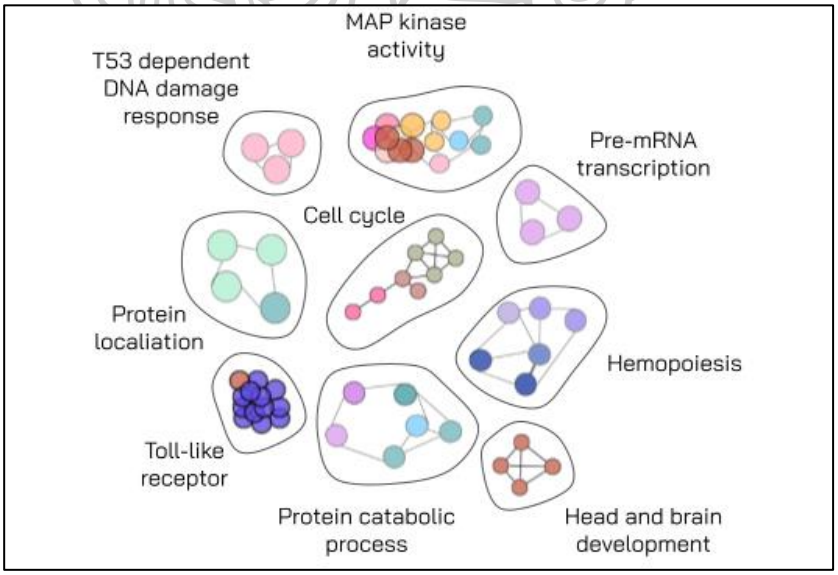
2.1.2.3 การวิเคราะห์ Pathway (Pathway analysis)

ภายหลังจากกระบวนการวิเคราะห์กลุ่มยีนส์ที่โดดเด่นสำคัญต่อวัตถุประสงค์ต่างๆ ยังมีการวิเคราะห์อื่นที่เจาะลึกถึงกลุ่มการทำงานของกลุ่มยีนส์ เรียกว่า Pathway ดังรูปที่ 2.7 กล่าวคือยีนส์ 1 กลุ่ม ไม่อาจแสดงออกคุณลักษณบางประการได้ ถ้าไม่ทำงานร่วมกับกลุ่มยีนส์อีกกลุ่มหนึ่ง ดังนั้นการวิเคราะห์ Pathway จึงมีความสำคัญต่อการจัดกลุ่มยีนส์ที่อาจทำงานร่วมกันแสดงออกถึงโรคบางชนิด โดยทั่วไปมักใช้ขั้นตอนวิธีการจัดกลุ่ม หรือ Cluster analysis เพื่อให้ผลลัพธ์ไปศึกษาในห้องปฏิบัติอีกครั้ง เครื่องมือสำหรับผู้วิจัยทางชีววิทยาเพื่อการศึกษาอาจเป็นเครื่องมือเดียวกันกับการทำ GSEA เช่น GSEA software และ Enrichr ใช้วิธีการจัดกลุ่มข้อมูลแบบลำดับขั้น (Hierarchical clustering และ pathfindR ที่สามารถเลือกวิธีในการจัดกลุ่มได้ทั้งแบบลำดับขั้น (Hierarchical clustering) และแบบฟัซซี่ ซีมีน (Fuzzy c-means clustering) เป็นต้น



รูปที่ 2.6 แสดงผลลัพธ์การวิเคราะห์ความสำคัญของกลุ่มยีนส์ Nitric oxide and reactive oxygen species ในเซลล์แมคโครฟาจ [14]

ที่มา. จาก Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease, โดย Li X. และ Long J., 2015, *Scientific reports*, *2015*(5), หน้า 6.



รูปที่ 2.7 แสดงตัวอย่างจำลองผลการวิเคราะห์ Pathway

ที่มา. จาก Integrative pathway enrichment analysis of multivariate omics data, โดย Marta Paczkowska., 2020, *Nature Communications*, *2020*(11), หน้า 4.

2.1.3 การคำนวณเชิงวิวัฒนาการ (Evolutionary computation)

การคำนวณเชิงวิวัฒนาการเป็นการคำนวณโดยใช้อัลกอริธึมเพื่อหาค่าเหมาะสมที่สุดแบบขั้นตอนเมตาฮิวริสติก (Metaheuristic optimization) มีขั้นตอนในการทำงานหลักคือการสร้างตัวค้นหาผลลัพธ์และใช้หลักเกณฑ์ที่ต่างกันในการค้นหาแบบสุ่ม (Stochastic search) [15] แนวคิดนี้เกิดจากการเลียนแบบปรากฏการณ์ทางธรรมชาติหรือ วิวัฒนาการของสิ่งมีชีวิต เช่น การตาย การคัดเลือก การสืบพันธุ์ และกลายพันธุ์ การคำนวณเชิงวิวัฒนาการส่วนใหญ่จะมีพื้นฐานกระบวนการที่เหมือนกัน 3 กระบวนการหลัก ดังนี้

- **กระบวนการสุ่มสร้างประชากร (Initialization)**: เป็นกระบวนการเริ่มต้นของอัลกอริธึม โดยทำการสร้างประชากรกลุ่มแรกขึ้นมาแบบสุ่ม
- **กระบวนการคัดเลือกประชากร (Selection)**: เป็นกระบวนการที่ประชากรส่วนหนึ่งจะถูกคัดเลือกเป็นต้นกำเนิดให้ทำการผสมพันธุ์และกลายพันธุ์ ทำให้ได้ประชากรที่มีรหัสพันธุกรรมใหม่โดยมีคุณสมบัติแตกต่างไปจากเดิม
- **กระบวนการคัดออกประชากร (Termination)**: เป็นกระบวนการสุดท้ายของอัลกอริธึม ก่อนที่จะวนเข้าสู่กระบวนการทั้งสองข้างต้นใหม่อีกครั้ง โดยกระบวนการนี้จะมีการกำจัดประชากรที่อ่อนแอที่สุดออก และถูกแทนที่ด้วยประชากรใหม่ที่ได้จากกระบวนการคัดเลือกประชากร (Selection) ทำให้ประชากรในรุ่นถัดไปมีความสามารถและมีคุณสมบัติที่สูงขึ้นกว่าเดิม

การคำนวณเชิงเชิงวิวัฒนาการในปัจจุบันมีหลายรูปแบบ โดยลักษณะเด่นของขั้นตอนวิธีเหล่านี้ได้แก่ การเลียนแบบวิธีการทางธรรมชาติ ดังกระบวนการหลัก 3 ขั้นตอน ลักษณะการทำงานเช่นนี้คล้ายแนวคิดหลักการของวิวัฒนาการ ในแต่ละกระบวนการอาจมีวิธีหรือเทคนิคแตกต่างกันออกไปตามวิธีที่ใช้เลียนแบบธรรมชาติ เช่น อาณาจักรมด ขั้นตอนวิธีเชิงพันธุกรรม การหาค่าเหมาะสมที่สุดแบบฝูงอนุภาค เป็นต้น แต่ละรูปแบบมีแนวคิดและการคำนวณที่เหมาะสมในการใช้กับสถานการณ์ของปัญหาที่แตกต่างกัน สามารถแบ่งเป็นการแก้ปัญหาค่าเหมาะสมที่สุดแบบวัตถุประสงค์เดียว (Single objective approaches) และแบบหลายวัตถุประสงค์ (Multi-objective approaches) ดังนี้

- **ปัญหาค่าเหมาะสมที่สุดแบบวัตถุประสงค์เดียว**: คือการหาคำตอบที่ดีที่สุดสำหรับจุดประสงค์เพียงหนึ่งเดียวเท่านั้น เช่น ความต้องการที่จะซื้อรถยนต์ด้วยราคาที่

ถูกที่สุด ขั้นตอนวิธีสำหรับแก้ปัญหาค่าเหมาะสมที่สุดแบบวัตถุประสงค์เดียว ตัวอย่างเช่น ขั้นตอนเชิงพันธุกรรม (Genetic Algorithm; GA) ขั้นตอนแบบอาณาจักรมด (Ant Colony Optimization; ACO) และการหาค่าเหมาะสมที่สุดแบบฝูงอนุภาค (Particle Swarm Optimization; PSO) เป็นต้น

- **ปัญหาค่าเหมาะสมที่สุดแบบหลายวัตถุประสงค์**: เป็นปัญหาที่เกี่ยวข้องกับการมีวัตถุประสงค์มากกว่า 1 แบบขึ้นไป และมักจะเป็นวัตถุประสงค์ที่มีความขัดแย้งกัน เช่น ความต้องการที่จะซื้อรถยนต์ด้วยราคาถูกที่สุด และมีสมรรถนะของเครื่องมากที่สุด เป็นต้น จะเห็นว่าปัญหาค่าเหมาะสมที่สุดแบบหลายวัตถุประสงค์นั้น มีลักษณะคำถามที่สมจริงมากกว่าแบบวัตถุประสงค์เดียว ทำให้การแก้ปัญหาแบบหลายวัตถุประสงค์ได้รับความสนใจและมีการพัฒนาวิธีการอย่างแพร่หลายมากขึ้น ตัวอย่างขั้นตอนวิธีสำหรับแก้ปัญหาค่าเหมาะสมที่สุดแบบหลายวัตถุประสงค์ อาทิ ขั้นตอนเชิงพันธุกรรมแบบการจัดลำดับที่ไม่ถูกครอบงำ (Non-dominated Sorting Genetic Algorithm II; NSGA-II)

2.2 งานวิจัยที่เกี่ยวข้อง

จากการทบทวนวรรณกรรม พบว่าเครื่องมือสำหรับวิเคราะห์ข้อมูลทางชีวสารสนเทศมีหลากหลายเครื่องมือซึ่งผลิตให้รองรับกับวัตถุประสงค์การวิจัยจากผู้ศึกษาในหลายด้าน ดังรูปที่ 2.5 จะเห็นว่ามีการวิเคราะห์การแสดงออกของยีนส์หลายรูปแบบ โดยเฉพาะอย่างยิ่งการวิเคราะห์ Pathway (Pathway analysis) ซึ่งคือการวิเคราะห์กลุ่มข้อมูลยีนส์โดยการทำ Clustering จึงได้ทำการศึกษาค้นคว้าขั้นตอนวิธีในการจัดกลุ่มที่ใช้ในเครื่องมือต่าง ๆ พบว่า เครื่องมือ pathfindR [8] Enrichr [7] และการศึกษา [16] หรือ MapleTree ใช้การจัดกลุ่มข้อมูลแบบลำดับขั้น (Hierarchical clustering) ด้วยการรวมแบบ Average link เครื่องมือ DAVID web server [6] และ Enrichment map [17] นำเสนอการจัดกลุ่มด้วยวิธีฟัซซี่ ซีมีน การศึกษา [18] นำเสนอการจัดกลุ่มแบบ Bi-level เครื่องมือ GSCluster [5] นำเสนอวิธีการคำนวณฟังก์ชันการวัดความเหมือนกันของข้อมูลรูปแบบใหม่ โดยอาศัยผลมาจากค่า Protein-protein interaction (PPI) จากฐานข้อมูล STRING ร่วมกับการจัดข้อมูลแบบฟัซซี่ ซีมีน จากตัวอย่างเครื่องมือและการศึกษาดังกล่าวพบว่าขั้นตอนวิธีการจัดกลุ่มล้วนเป็นขั้นตอนวิธีแบบดั้งเดิม แต่หากมีการศึกษาในปี ค.ศ. 2005 นำเสนอการนำขั้นตอนวิธีเชิงพันธุกรรมเพื่อใช้คัดเลือกยีนส์ในกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA) [9] เช่นเดียวกันกับเครื่องมือ pathfindR ที่ผู้พัฒนานำเสนอทางเลือก

ในการค้นหาด้วยขั้นตอนวิธีการอบเหนียว และวิธีเชิงพันธุกรรมช่วยค้นหายีนส์ที่อาจจะโดดเด่นให้แก่การวิเคราะห์ดังกล่าว ในหัวข้อนี้จะอธิบายรายละเอียดเครื่องมือ pathfindR ชุดข้อมูลดิบจากเครื่องมือ pathfindR.data รวมถึงเนื้อหาอื่น ๆ ที่เกี่ยวข้องกับงานวิจัยนี้ อาทิ ขั้นตอนวิธีเชิงวิวัฒนาการ และสมการจุดประสงค์ ดังต่อไปนี้

2.2.1 เครื่องมือ pathfindR และ pathfindR.data

ในหัวข้อนี้จะอธิบายรายละเอียดของเครื่องมือ pathfindR และ pathfindR.data ซึ่งเป็นเครื่องมือที่ใช้สำหรับวิเคราะห์และค้นหากลุ่มยีนส์ที่สำคัญหรือโดดเด่น (Gene-Set Enrichment Analysis; GSEA) และยังสามารถทำการวิเคราะห์ Pathway (Pathway analysis) เขียนโปรแกรมใช้งานเป็นภาษาอาร์ (R language) โดยผู้พัฒนา Ege Ulgen เมื่อปี ค.ศ. 2019 [8] ซึ่งเครื่องมือนี้นำเสนอการค้นหายีนส์ที่โดดเด่นด้วยขั้นตอนวิธีการเชิงวิวัฒนาการ ทดแทนการค้นหาแบบละโมบ และเครื่องมือ pathfindR.data เป็นเครื่องมือที่รวบรวมชุดข้อมูลตัวอย่างเพื่อช่วยให้ผู้ใช้ทดลองการใช้งานชุดคำสั่ง มีรายละเอียดดังต่อไปนี้

2.2.1.1 เครื่องมือ pathfindR

pathfindR เป็นหนึ่งในเครื่องมือการวิเคราะห์ข้อมูลทางชีวสารสนเทศสำหรับการปฏิบัติการไมโครอะเรย์เพื่อช่วยวิเคราะห์ข้อมูลการแสดงออกของโรคในระดับยีนส์ เครื่องมือนี้สร้างขึ้นโดยผู้พัฒนา Ege Ulgen เพื่อการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA) โดยสร้างให้มีวิธีการค้นหายีนส์ที่โดดเด่น (Active Subnetwork Search) 3 ทางเลือก ได้แก่ วิธีการแบบละโมบ (Greedy algorithm) เป็นวิธีค้นหาคำตอบโดยพื้นฐาน (Default) ขั้นตอนวิธีการแบบการจำลองการอบเหนียว (Simulated Annealing algorithm; SA) และขั้นตอนวิธีเชิงพันธุกรรม (Genetic algorithm) ทดลองกับชุดข้อมูล 3 ชุดที่ได้จากเทคนิคไมโครอะเรย์ ได้แก่ ชุดข้อมูลการแสดงออกของยีนส์ที่ได้จากเซลล์เม็ดเลือดขาวโมโนนิวไคลด์แบบกลีบเดียวจากผู้ป่วยโรคข้ออักเสบรูมาตอยด์ (Rheumatoid arthritis; GSE15573) ทำเทคนิคไมโครอะเรย์ด้วยโพรบ Illumina human-6 v.20 expression bead chip ชุดข้อมูลการแสดงออกของยีนส์ที่ได้จากเยื่อเมือกของลำไส้ใหญ่ ในผู้ป่วยที่เริ่มมีอาการโรคมะเร็งลำไส้ใหญ่ส่วนปลาย (GSE4107) ทำเทคนิคไมโครอะเรย์ด้วยโพรบ Affymetrix Human Genome U133 Plus 2.0 และข้อมูลชุดการแสดงออกของยีนส์ในเนื้อเยื่อต่อมลูกหมากจากผู้ป่วยมะเร็งต่อมลูกหมาก (GSE55945) โดยใช้โพรบ Affymetrix Human Genome U133 Plus 2.0 ทั้งนี้กระบวนการของเครื่องมือ pathfindR ให้การวิเคราะห์ผลทางชีวสา

รสนเทศได้ 2 กระบวนการหลัก ได้แก่ กระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ และการวิเคราะห์ Pathway มีรายละเอียดสำคัญและขั้นตอนที่เกี่ยวข้องกับวิทยานิพนธ์ดังต่อไปนี้

2.2.1.1.1 ชุดข้อมูลดิบ (Raw data)

ในขั้นตอนแรกจะแนะนำคุณลักษณะของชุดข้อมูลดิบ (Attribute) ที่ได้รับจากการทำการทดลองเทคนิคไมโครอะเรย์เพื่อใช้เป็นข้อมูลขาเข้าต่อในขั้นตอนถัดไป คุณลักษณะที่จำเป็นต้องมี 3 attributes ได้แก่ สัญลักษณ์ของยีนส์ ค่า Fold change และค่า Adjust p-value ยกตัวอย่างชุดข้อมูลดิบสำหรับงานวิจัยนี้ ข้อมูลดิบประกอบไปด้วยสัญลักษณ์ยีนส์ต่าง ๆ จำนวน 572 ยีนส์ และ 3 attributes ในรูปแบบโครงสร้างข้อมูลประเภทกรอบข้อมูล (Dataframe) ซึ่งมีลักษณะคล้ายตาราง (Table) ที่คุ้นเคยจากการใช้โปรแกรมไมโครซอร์ฟเอกเซล ดังตารางที่ 2.1 อนึ่ง รายละเอียดของข้อมูลดิบที่ใช้ในการศึกษานี้จะกล่าวในหัวข้อเครื่องมือ pathfindR.data ในลำดับถัดไป

2.2.1.1.2 การจัดการข้อมูลดิบก่อนการวิเคราะห์ (Preprocessing)

ขั้นตอนนี้คือกระบวนการเตรียมข้อมูลดิบให้มีความเหมาะสมต่อการค้นหาในลำดับถัดไป ใช้ชุดคำสั่ง input_processing หรือ pathfindR::input_processing เพื่อคัดกรองข้อมูลที่ดีเบื้องต้นจากการวิเคราะห์ค่า p-value และค้นหาข้อมูลยีนส์ที่ตรงกันกับ PIN (Protein interaction networks) จากฐานข้อมูลต่าง ๆ ทางชีววิทยา อาทิ ฐานข้อมูล STRING ฐานข้อมูล BioGRID ฐานข้อมูล KEGG เป็นต้น ซึ่งผู้วิจัยสามารถเลือกฐานข้อมูลเพื่อระบุ PIN ได้

ข้อมูลขาเข้าในขั้นตอนนี้ได้แก่ข้อมูลดิบดังรายละเอียดในหัวข้อก่อนหน้า กำหนดค่า p-value สำหรับการคัดออก (Default = 0.05) และเลือกใช้ PIN จากฐานข้อมูล BioGRID

ข้อมูลขาออกจากขั้นตอนนี้จะเป็นโครงสร้างข้อมูลประเภทกรอบข้อมูล ซึ่งประกอบด้วยยีนส์ที่ได้รับการคัดเลือกอ้างอิงจาก PIN ตามฐานข้อมูลที่ได้ระบุ รวมทั้งหมด 4 attributes ได้แก่ สัญลักษณ์ยีนส์เดิม สัญลักษณ์ยีนส์ที่ได้รับการตรวจสอบ ค่า Fold change และค่า Adjust p-value โดยที่จำนวนยีนส์อาจจะมีจำนวนลดลง ดังตารางที่ 2.2

ตารางที่ 2.1 แสดงตัวอย่างกรอบข้อมูลของชุดข้อมูลดิบส่วนหัวและท้าย

No.	Gene symbol	log Fold change	Adjust p-value
1	FAM110A	-6.94E-01	3.41E-06
2	RNASE2	1.353504	1.01E-05
3	S100A8	1.5448338	3.47E-05
4	S100A9	1.0280904	2.26E-04
5	TEX261	-0.3235994	2.26E-04
6	ARHGAP17	-0.691933	2.71E-04
7	NUP62	-0.821632	2.71E-04
8	MYL6B	0.2790078	4.01E-04
9	BLOC1S1	0.6930502	4.40E-04
10	PCBP1	-0.5029724	4.51E-04
:	:	:	:
563	TSR2	-0.24071623	0.04874446
564	SLC16A11	-0.11540911	0.04912335
565	SH3BP5L	0.4065718	0.04912335
566	ARID1A	-0.32117616	0.04912335
567	DPY30	0.52349122	0.04924908
568	TAF4	-0.27502822	0.04942016
569	DBI	0.82784931	0.04987782
570	TM9SF4	-0.30784064	0.04987782
571	ASPSCR1	-0.31292026	0.04987782
572	CSNK2A2	-0.22271771	0.04987782

ตารางที่ 2.2 แสดงตัวอย่างกรอบข้อมูลที่ผ่านการทำ Preprocessing

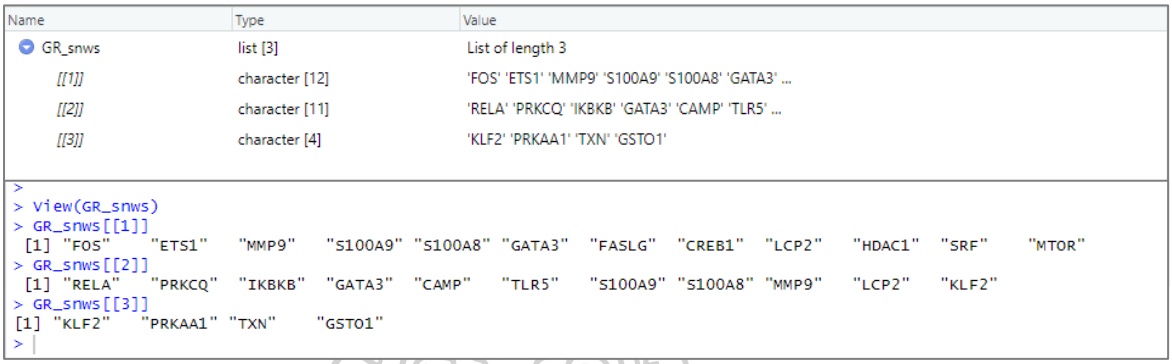
No.	Old gene	Update gene	log Fold change	Adjust p-value
1	FAM110A	FAM110A	-0.6939359	3.41E-06
2	RNASE2	RNASE2	1.353504	1.01E-05
3	S100A8	S100A8	1.5448338	3.47E-05
4	S100A9	S100A9	1.0280904	2.26E-04
5	TEX261	TEX261	-0.3235994	2.26E-04
6	ARHGAP17	ARHGAP17	-0.691933	2.71E-04
7	NUP62	NUP62	-0.821632	2.71E-04
8	MYL6B	MYL6B	0.2790078	4.01E-04
9	BLOC1S1	BLOC1S1	0.6930502	4.40E-04
10	PCBP1	PCBP1	-0.5029724	4.51E-04

2.2.1.1.3 กระบวนการค้นหาข้อมูลยีนส์ที่สำคัญและโดดเด่น (Active subnetwork search)

ในขั้นตอนนี้เป็นกระบวนการที่ผู้พัฒนานำเสนอทางเลือกในการใช้ขั้นตอนวิธีเชิงวิวัฒนาการเพื่อช่วยค้นหายีนส์ ซึ่งเป็นขั้นตอนการคัดกรองยีนส์อีกครั้ง แต่เพิ่มเติมการระบุยีนส์จากฐานข้อมูลรวมต่าง ๆ เรียกว่า กระบวนการค้นหา Active subnetwork ด้วยชุดคำสั่ง active_snw_search หรือ pathfindR::active_snw_search มีพารามิเตอร์ที่ต้องกำหนดค่าดังนี้

ข้อมูลขาเข้า ได้แก่ชุดข้อมูลดิบที่ต้องประกอบไปด้วยสัญลักษณ์ยีนส์ และค่า p-value หรือสามารถใช้ชุดข้อมูลดิบที่ผ่านขั้นตอน Preprocessing แล้วได้เช่นกัน ต้องมีการกำหนดฐานข้อมูลสำหรับการค้นหา และกำหนดขั้นตอนวิธีในการค้นหา active subnetwork 3 รูปแบบ ซึ่งในกระบวนนี้เรียกใช้ฐานข้อมูล KEGG และเลือกวิธีการค้นหาแบบละโมบ

ข้อมูลขาออกจะเป็นจำนวน Active subnetwork ที่ค้นหาได้ในรูปแบบโครงสร้างข้อมูลแบบลิสต์ ในแต่ละลิสต์ประกอบไปด้วยชื่อยีนส์ที่อยู่ใน Subnetwork เดียวกัน แสดงตัวอย่างการทำ Active subnetwork search จากตัวอย่างยีนส์ 15 ใน 572 ตัว ได้ผลการค้นหา 3 subnetwork ดังรูปที่ 2.8



Name	Type	Value
GR_snws	list [3]	List of length 3
[[1]]	character [12]	'FOS' 'ETS1' 'MMP9' 'S100A9' 'S100A8' 'GATA3' ...
[[2]]	character [11]	'RELA' 'PRKCQ' 'IKBKB' 'GATA3' 'CAMP' 'TLR5' ...
[[3]]	character [4]	'KLF2' 'PRKAA1' 'TXN' 'GSTO1'

```
>
> View(GR_snws)
> GR_snws[[1]]
 [1] "FOS"    "ETS1"   "MMP9"   "S100A9" "S100A8" "GATA3"  "FASLG"  "CREB1"  "LCP2"   "HDAC1"  "SRF"    "MTOR"
> GR_snws[[2]]
 [1] "RELA"   "PRKCQ"  "IKBKB"  "GATA3"  "CAMP"   "TLR5"   "S100A9" "S100A8" "MMP9"   "LCP2"   "KLF2"
> GR_snws[[3]]
[1] "KLF2"   "PRKAA1" "TXN"    "GSTO1"
>
```

รูปที่ 2.8 แสดงตัวอย่างผลลัพธ์การทำ Active subnetwork search

2.2.1.1.4 กระบวนการวิเคราะห์ความสำคัญและความโดดเด่นของยีนส์ (Enrichment analysis)

ในขั้นตอนนี้จะรับข้อมูลขาออกในขั้นตอนก่อนหน้า ได้แก่ ลิสต์ของยีนส์เป็นข้อมูลขาเข้าเพื่อที่จะทำการคัดกรองเฉพาะยีนส์ที่มีความสำคัญจากเงื่อนไขต่าง ๆ อาทิ ยีนส์ที่สามารถผ่านกระบวนนี้ต้องสามารถระบุได้จาก PIN ใช้การปรับค่า p-value ด้วยเทคนิคต่างๆ ที่ผู้ใช้สามารถเลือกเทคนิคในการปรับค่าได้ ด้วยชุดคำสั่ง enrichement_analyses หรือ pathfindR::enrichement_analyses

ดังนั้น ข้อมูลขาเข้าที่จำเป็นในขั้นตอนนี้ได้แก่ สัญลักษณ์ของยีนส์ ที่ผู้ใช้สามารถเลือกได้ว่าจะใช้สัญลักษณ์ของยีนส์ที่ได้จากการทำ Active subnetwork ด้วยหรือไม่ และค่า Adjust p-value โดยเลือกวิธีบอนเฟอโรนีเป็นค่าเริ่มต้น (Bonferroni correction) และใช้ฐานข้อมูลเพื่อระบุ PIN จากฐานข้อมูล KEGG ทำให้ได้ข้อมูลขาออกจากกระบวนการนี้แสดงผลเป็นกรอบข้อมูลที่ประกอบไปด้วย 6 attribute หากไม่ต้องการผลจากขั้นตอน Active subnetwork หรือ 7 attributes หากต้องการใช้ผลจากขั้นตอนก่อนหน้า ข้อมูลขาออกสามารถแสดงผลได้ดังตารางที่ 2.3 ซึ่งได้จากการทำ Enrichment analysis ของตัวอย่างยีนส์ 15 ใน 572 ตัว และเลือกต้องการวิเคราะห์รวมจากผลที่ได้จากขั้นตอนก่อนหน้า ประกอบด้วย รหัสยีนส์ คำอธิบายของยีนส์อ้างอิงจากฐานข้อมูล KEGG ค่า Fold change ค่า p-value ค่า Adjust p-value สัญลักษณ์ยีนส์จากขั้นตอน Active subnetwork และค่าสนับสนุน

ตารางที่ 2.3 แสดงตัวอย่างผลลัพธ์ขั้นตอน Enrichment analysis

ID	Term_Description	Fold_Enrichment	p_value	adj_p	non_Signif_Snw_Genes	support
hsa05418	Fluid shear stress and atherosclerosis	6.028019	2.29E-07	7.68E-05	PRKAA1, GSTO1	0.6666667
hsa04657	IL-17 signaling pathway	8.944803	2.84E-07	9.53E-05	RELA, IKBKB, MMP9	0.6666667
hsa04658	Th1 and Th2 cell differentiation	4.521014	1.35E-05	4.52E-03	RELA, PRKCQ, IKBKB, GATA3	0.3333333
hsa04660	T cell receptor signaling pathway	0	2.12E-05	7.09E-03	RELA, PRKCQ, IKBKB, LCP2	0.3333333
hsa04659	Th17 cell differentiation	0	2.55E-05	8.55E-03	RELA, PRKCQ, IKBKB, GATA3	0.3333333

2.2.1.1.5 ขั้นตอนการสรุปผลยีนส์ที่มีความสำคัญและโดดเด่น (Summarizing enrichment results)

ในส่วนขั้นตอนนี้เป็นการสรุปผลจากการวิเคราะห์หายีนส์ที่มีความสำคัญ โดยใช้ชุดคำสั่ง pathfindR:: summarize_enrichment_results ข้อมูลขาเข้าได้จากขั้นตอนก่อนหน้า และสามารถเลือกให้ใช้ หรือไม่ใช้ผลยีนส์ที่ได้จากขั้นตอน Active subnetwork นำออกเป็นข้อมูลขาออกที่เป็นกรอบข้อมูล 7 หรือ 8 attributes มีรายละเอียด ดังตารางที่ 2.4

ตารางที่ 2.4 แสดงตัวอย่างผลลัพธ์ขั้นตอน Summarizing enrichment results

ID	Term_Description	Fold_Enrichment	occurrence	support	lowest_p	highest_p	non_Signif_Snw_Genes
hsa05418	Fluid shear stress and atherosclerosis	6.028019	1	0.6666667	7.68E-05	7.68E-05	PRKAA1, GSTO1
hsa04657	IL-17 signaling pathway	8.944803	1	0.6666667	9.53E-05	9.53E-05	RELA, IKBKB, MMP9
hsa04658	Th1 and Th2 cell differentiation	4.521014	1	0.3333333	4.52E-03	4.52E-03	RELA, PRKCQ, IKBKB, GATA3
hsa04660	T cell receptor signaling pathway	0	1	0.3333333	7.09E-03	7.09E-03	RELA, PRKCQ, IKBKB, LCP2
hsa04659	Th17 cell differentiation	0	1	0.3333333	8.55E-03	8.55E-03	RELA, PRKCQ, IKBKB, GATA3

2.2.1.1.6 ขั้นตอนการเติมข้อมูลเพิ่มเติมให้กับกลุ่มยีนส์ (Function annotation)

ขั้นตอนนี้คือการเติมข้อมูลเพิ่มเติมให้กับกลุ่มยีนส์ที่ได้จากกระบวนการก่อนหน้า โดยเรียกเพิ่มเติมจากฐานข้อมูลพื้นฐาน KEGG ทำให้ข้อมูลขาออกประกอบด้วย 9-10 attributes ดังแสดงในตารางที่ 2.5 จากตารางจะสังเกตว่ามีข้อมูลเพิ่มเติม ได้แก่ ชื่อยีนส์ที่แสดงออกมาเกิน (Up_regulated) และชื่อยีนส์ที่แสดงออกน้อยเกินไป (Down_regulated) ล้วนเป็นชื่อยีนส์จำนวนหลาย ๆ ตัวที่อยู่ในกลุ่มรหัส (ID) หรือ คำอธิบายยีนส์ (Term description) เดียวกัน สำหรับข้อมูลชื่อยีนส์ในส่วนนี้ จะมีบทบาทสำคัญต่อการวิเคราะห์ Pathway ในลำดับถัดไป

ตารางที่ 2.5 แสดงตัวอย่างผลลัพธ์ขั้นตอน Function annotation

ID	Term_ Description	Fold_ Enrichment	occurrence	support	lowest_p	highest_p	non_Signif Snw_Genes	Up_ regulated	Down_ regulated
hsa 05418	Fluid shear stress and atherosclerosis	6.028019	1	0.6666667	7.68E-05	7.68E-05	PRKAA1, GSTO1	GSTO1, TXN, MMP9	CALM3, CALM1, KLF2, ACTG1, ACTB, IKBKB, SUMO3
hsa 04657	IL-17 signaling pathway	8.944803	1	0.6666667	9.53E-05	9.53E-05	RELA, IKBKB, MMP9	S100A8, S100A9, MMP9	IKBKB
hsa 04658	Th1 and Th2 cell differentiation	4.521014	1	0.3333333	4.52E-03	4.52E-03	RELA, PRKCQ, IKBKB, GATA3		JAK1, RUNX3, HLA-DPA1, NFATC3, PRKCQ, IKBKB, GATA3
hsa 04660	T cell receptor signaling pathway	0	1	0.3333333	7.09E-03	7.09E-03	RELA, PRKCQ, IKBKB, LCP2		LCP2, NFATC3, PRKCQ, CARD11, IKBKB
hsa 04659	Th17 cell differentiation	0	1	0.3333333	8.55E-03	8.55E-03	RELA, PRKCQ, IKBKB		MTOR, JAK1, HLA-DPA1, IL27

2.2.1.1.7 กระบวนการวิเคราะห์ Pathway (Pathway analysis)

กระบวนการทั้ง 6 ขั้นตอนก่อนหน้า เป็นการศึกษากลุ่มยีนส์ที่มีลักษณะโดดเด่นในการศึกษาไมโครอะเรย์ จุดประสงค์จึงเป็นการศึกษาเพื่อระบุกลุ่มยีนส์ที่มีการแสดงออกที่สำคัญ ๆ เท่านั้น หากเพื่อเป็นการศึกษาลึกลงไปถึงความสัมพันธ์กัน หรือการทำงานร่วมกันของกลุ่มยีนส์ กระบวนการวิเคราะห์ Pathway จึงเป็นการวิเคราะห์เพื่อตอบโจทย์ว่ายีนส์กลุ่มใดทำงานร่วมกับยีนส์กลุ่มใด ดังนั้นสำหรับกระบวนการศึกษา Pathway นี้ คือการจัดกลุ่มของกลุ่มยีนส์ที่สำคัญ ๆ นั่นเอง เครื่องมือ pathfinder ในกระบวนการวิเคราะห์นี้จึงเป็นการวิเคราะห์การจัดกลุ่ม หรือ Cluster analysis นั่นเอง ซึ่งเครื่องมือนี้ได้ให้ผู้ใช้สามารถวิธีการจัดกลุ่มได้ด้วยขั้นตอนวิธี 2 ประเภท ได้แก่ การจัดกลุ่มแบบลำดับขั้น (Hierarchical clustering) ที่ผู้พัฒนาเป็นผู้เขียนชุดคำสั่ง และแบบฟัซซี่ ซีมีน (Fuzzy c-means clustering) ตามการอ้างอิงจากเครื่องมือ DAVID web server โดยชุดคำสั่ง hierarchical_term_clustering และ fuzzy_term_clustering ตามลำดับ

กระบวนการจัดกลุ่มข้อมูล จำเป็นอย่างยิ่งที่จะต้องมีฟังก์ชันระยะทาง หรือการวัดความเหมือนกันระหว่างกลุ่ม แต่หากสังเกตข้อมูลขาออกที่ได้จากกระบวนการ Function annotation ในขั้นตอนก่อนหน้า จะพบว่าข้อมูลที่สำคัญต่อการจัดกลุ่ม ได้แก่ ชื่อยีนส์ที่แสดงออกแบบ Up related และ Down regulated รวมถึงรหัสของกลุ่มยีนส์ที่โดดเด่น มีลักษณะเป็นชื่อที่ไม่มีความหมายเท่านั้น ทำให้ฟังก์ชันระยะหรือการวัดความเหมือนกันของข้อมูลสำหรับเครื่องมือนี้ใช้การคำนวณสัมประสิทธิ์โคเฮนคัปปา (Cohen’s kappa coefficient) เพื่อหาความสัมพันธ์ระหว่างยีนส์ต่าง ๆ ในแต่ละกลุ่ม ด้วยชุดคำสั่ง create_kappa_matrix ซึ่งให้ผลเป็นโครงสร้างข้อมูลประเภท เมท-ริกซ์ทแยงมุมแสดงความสอดคล้องกันของกลุ่มยีนส์ ดังรูปที่ 2.9 แสดงเมทริกซ์ค่าความสอดคล้องกันของข้อมูลกลุ่มยีนส์ที่โดดเด่นจากกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis) โดยพบกลุ่มยีนส์ที่สำคัญ 113 กลุ่ม ทั้งนี้ชุดคำสั่งการจัดกลุ่มข้อมูลได้ใช้เมทริกซ์นี้เป็นฟังก์ชันระยะทางเพื่อจัดกลุ่มข้อมูลที่อาจทำหน้าที่เป็น Pathway เดียวกัน

ในส่วนของขั้นตอนวิธีการจัดกลุ่มของเครื่องมือ pathfindR นี้ จะกล่าวถึงการจัดกลุ่มประเภทลำดับขั้น เครื่องมือนี้เรียกใช้ชุดคำสั่งการจัดกลุ่มซึ่งเป็นเครื่องมือพื้นฐานของโปรแกรมภาษาอาร์ ได้แก่ เครื่องมือ stats จากชุดคำสั่ง hclust ซึ่งเป็นการแบ่งกลุ่มแบบลำดับขั้นโดยทั่วไป โดยใช้วิธีรวมข้อมูลแบบ Average link และจำเป็นต้องมีการกำหนดจำนวนกลุ่มที่ต้องการแบ่งตั้งแต่เริ่มต้น ดังนั้น คำตอบที่ได้จากการแบ่งกลุ่มนี้อาจไม่ใช่คำตอบที่เหมาะสม เนื่องจากไม่ทราบจำนวนกลุ่มที่ควรถูกแบ่ง เรียกค่าจำนวนกลุ่มข้อมูลที่ต้องการแบ่งว่า ค่า k ซึ่งเป็นไปตามรูปแบบการทำงานโดยทั่วไป

ของตัวแบ่งกลุ่มข้อมูลแบบลำดับขั้น ผู้พัฒนาจึงเพิ่มเติมชุดคำสั่งด้วยการหาค่า k ที่เหมาะสมจากการคำนวณแบบวนซ้ำ (Iterative method) เพื่อทำให้ค่าเฉลี่ย Silhouette width มีค่าสูงที่สุด จากตัวอย่างข้อมูลดิบที่ยกตัวอย่างตั้งแต่กระบวนการแรกนี้ เครื่องมือ pathfindR ให้คำตอบเป็นการแบ่งกลุ่มออกเป็น 19 กลุ่ม

2.2.1.2 เครื่องมือ pathfindR.data

เครื่องมือนี้เป็นการรวบรวมชุดข้อมูลที่สามารถใช้ศึกษาการทำงานเครื่องมือ pathfindR ซึ่งรวบรวมชุดข้อมูลตัวอย่าง ข้อมูลยีนส์และคำอธิบายจากฐานข้อมูล Biocarta KEGG และ Reactome ทั้งนี้ ในส่วนนี้จะกล่าวถึงชุดข้อมูลสำคัญที่เกี่ยวข้องกับงานวิจัยนี้ ได้แก่ ชุดข้อมูลการแสดงออกของยีนส์ของเซลล์เม็ดเลือดขาวชุดโมโนนิวไคลด์แบบกลีบเดียวจากผู้ป่วยโรคข้ออักเสบรูมาตอยด์จำนวน 18 คน และจากผู้ป่วยสุขภาพดี 15 คน (Rheumatoid arthritis; GSE15573) ใช้อักษรย่อ RA มีรายละเอียดดังนี้

- ชุดคำสั่ง RA_input หรือ parthfindR.data::RA_input ให้ชุดข้อมูลที่เป็นข้อมูลขาออกจากการทดลองเทคนิคไมโครอะเรย์ ประกอบไปด้วยยีนส์จำนวน 572 ตัว 3 attributes ดังตารางที่ 2.1
- ชุดคำสั่ง RA_output หรือ pathfindR.data::RA_output เป็นข้อมูลขาออกจากกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ ประกอบไปด้วยกลุ่มยีนส์ที่สำคัญทั้งหมด 113 กลุ่มยีนส์ 9 attributes ดังตารางที่ 2.5

	hsa05415	hsa04130	hsa03410	hsa04064	hsa03040	hsa04714	hsa03013	hsa00190	hsa05012	hsa05020	hsa04932	hsa04659	hsa03430	hsa04260	hsa04722
hsa05415	1.000000000	-0.03583427	0.05637584	0.02724665	-0.087808702	0.41983536	-0.080066722	0.44546922	0.471931494	0.564876519	0.44546922	0.01890708	-0.02807839	0.45369128	-0.06640
hsa04130	-0.035834267	1.00000000	-0.02461538	-0.02968460	-0.034804226	-0.03480423	-0.033519553	-0.03086155	-0.035346097	-0.034804226	-0.03086155	-0.03086155	-0.01888277	-0.02461538	-0.03086
hsa03410	0.056375839	-0.02461538	1.00000000	0.12473118	-0.041481481	-0.04148148	-0.039669421	-0.03600000	-0.042253521	-0.041481481	-0.03600000	-0.03600000	0.23448276	-0.02777778	-0.03600
hsa04064	0.027246654	-0.02968460	0.12473118	1.00000000	-0.058242330	-0.05824233	0.050741163	-0.04798401	-0.059775841	0.037961518	0.07530823	0.19860047	-0.02415702	-0.03440860	0.07530
hsa03040	-0.087808702	-0.03480423	-0.04148148	-0.05824233	1.000000000	-0.08187135	0.007599464	-0.06295279	-0.084934277	-0.081871345	-0.06295279	-0.06295279	-0.02744201	-0.04148148	-0.06295
hsa04714	0.419835359	-0.03480423	-0.04148148	-0.05824233	-0.081871345	1.00000000	-0.075100581	0.49163128	0.364004044	0.459064327	0.49163128	0.02947789	-0.02744201	0.39703704	-0.06295
hsa03013	-0.080066722	-0.03351955	-0.03966942	0.05074116	0.007599464	-0.07510058	1.000000000	-0.05887163	-0.077669903	-0.075100581	-0.05887163	-0.05887163	-0.02663707	-0.03966942	-0.05887
hsa00190	0.445469218	-0.03086155	-0.03600000	-0.04798401	-0.062952785	0.49163128	-0.058871627	1.00000000	0.467625899	0.491631277	0.64962121	-0.05113636	-0.02493075	0.55600000	-0.05113
hsa05012	0.471931494	-0.03534610	-0.04225352	-0.05977584	-0.084934277	0.36400404	-0.077669903	0.46762590	1.000000000	0.738119312	0.55635492	-0.06474820	-0.02777778	0.37464789	0.11270
hsa05020	0.564876519	-0.03480423	-0.04148148	0.03796152	-0.081871345	0.45906433	-0.075100581	0.49163128	0.738119312	1.000000000	0.58406195	-0.06295279	-0.02744201	0.39703704	-0.06295
hsa04932	0.445469218	-0.03086155	-0.03600000	0.07530823	-0.062952785	0.49163128	-0.058871627	0.64962121	0.556354916	0.584061954	1.00000000	0.06565657	-0.02493075	0.55600000	0.18244
hsa04659	0.018907079	-0.03086155	-0.03600000	0.19860047	-0.062952785	0.02947789	-0.058871627	-0.05113636	-0.064748201	-0.062952785	0.06565657	1.00000000	-0.02493075	-0.03600000	0.06565
hsa03430	-0.028078385	-0.01888277	0.23448276	-0.02415702	-0.027442012	-0.02744201	-0.026637070	-0.02493075	-0.027777778	-0.027442012	-0.02493075	-0.02493075	1.00000000	-0.02068966	-0.02493
hsa04260	0.453691275	-0.02461538	-0.02777778	-0.03440860	-0.041481481	0.39703704	-0.039669421	0.55600000	0.374647887	0.397037037	0.55600000	-0.03600000	-0.02068966	1.00000000	-0.03600
hsa04722	-0.066405349	-0.03086155	-0.03600000	0.07530823	-0.062952785	-0.06295279	-0.058871627	-0.05113636	0.112709832	-0.062952785	0.18244949	0.06565657	-0.02493075	-0.03600000	1.00000
hsa05166	0.057232458	-0.03480423	-0.04148148	0.03796152	-0.081871345	0.07268170	-0.075100581	-0.06295279	0.064711830	0.149958229	0.02947789	0.39920060	-0.02744201	-0.04148148	0.02947
hsa04630	0.045899948	-0.02663707	-0.03037975	-0.03849238	-0.047565119	0.05719139	-0.045197740	-0.04049494	-0.048582996	-0.047565119	-0.04049494	0.51443570	-0.02209945	-0.03037975	-0.04049
hsa05167	0.011135857	-0.03187251	-0.03738318	0.18296530	-0.067307692	0.11057692	-0.062663185	-0.05397301	0.101734104	0.021634615	0.05697151	0.38980510	-0.02558635	-0.03738318	0.27886
hsa04931	0.067749160	-0.02209945	-0.02461538	0.31354360	-0.034804226	0.19515227	-0.033519553	-0.03086155	-0.035346097	0.080174021	0.12773253	0.44492070	-0.01888277	-0.02461538	0.12773
hsa04530	-0.071269488	-0.03187251	-0.03738318	-0.05047319	-0.067307692	0.11057692	-0.062663185	-0.05397301	-0.069364162	-0.067307692	-0.05397301	-0.05397301	-0.02558635	-0.03738318	-0.05397
hsa04210	0.036219300	-0.02829712	0.13953488	0.23582426	-0.053130930	0.14746544	-0.050194204	-0.04446013	0.041450777	0.047167254	0.34721242	0.08609739	-0.02323009	-0.03255814	0.21665

Showing 1 to 21 of 113 entries, 113 total columns

รูปที่ 2.9 แสดงตัวอย่างเมทริกซ์ทแยงมุมของสัมประสิทธิ์โคเฮนคัปปาใช้สำหรับเป็นฟังก์ชันระยะทาง หรือการวัดความเหมือนกันของข้อมูลเพื่อการจัดกลุ่มในการวิเคราะห์ Pathway analysis

2.2.2 ขั้นตอนวิธีความฉลาดแบบกลุ่มเพื่อการทำการจัดกลุ่มข้อมูล

จากหัวข้อการแบ่งกลุ่มข้อมูล พบว่าปัญหาการแบ่งกลุ่มข้อมูลด้วยการพยายามหากลุ่มที่มีสมาชิกลักษณะคล้ายคลึงกันหรืออยู่ใกล้กัน รวมเข้าไว้เป็นกลุ่มเดียวกัน ผลลัพธ์ของปัญหานี้ไม่เพียงแต่เป็นความพยายามหากลุ่มที่เหมาะสมของข้อมูลแต่ละตัว แต่ยังรวมถึงการพยายามหาจำนวนกลุ่มที่ควรถูกแบ่ง (ค่า k) สำหรับข้อมูลดิบอีกด้วย ซึ่งสามารถมองเป็นปัญหาของการค้นหาค่าเหมาะสมที่สุด ภายใต้จุดประสงค์ของผู้ศึกษา สอดคล้องกับวิธีการคำนวณเชิงวิวัฒนาการในหัวข้อก่อนหน้า ขั้นตอนวิธีแบบลำดับขั้นโดยทั่วไปตามที่ได้อธิบายในหัวข้อการแบ่งกลุ่มข้อมูล ลักษณะการทำงานของตัวจัดกลุ่มประเภทนี้จะทำการค้นหาค่าฟังก์ชันระยะทางที่มีค่าน้อยสุด และรวมกลุ่มกับข้อมูลตัวที่มีค่าฟังก์ชันระยะทางน้อยที่สุดในบรรดาข้อมูลที่เหลือ เหตุการณ์ค้นหาเช่นนี้ อาจนำไปสู่สภาวะค่าสูงสุดท้องถิ่น หรือ Local optima เป็นผลให้คำตอบที่ค้นหาได้อาจไม่ใช่คำตอบที่แท้จริงหรือค่าสูงสุดอย่างแท้จริง ในปี ค.ศ. 2014 [19] มีการรวบรวมวรรณกรรมที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเชิงวิวัฒนาการแบบหลายจุดประสงค์ พบว่างานวิจัยที่นำเสนอการนำขั้นตอนวิธีเชิงพันธุกรรม (Generic algorithm) สำหรับการจัดกลุ่มข้อมูลในขั้นตอนการเลือกจำนวนกลุ่มของข้อมูลที่ควรถูกแบ่ง (ค่า k) นำมาใช้ในส่วนของการประเมินคุณลักษณะของกลุ่มข้อมูล และให้ผลเป็นคำตอบที่เข้าใกล้คำตอบที่เหมาะสมที่สุด ขั้นตอนวิธีเชิงวิวัฒนาการจึงเริ่มมีการร่วมใช้งานกับปัญหาการจัดกลุ่มข้อมูลในหลายแง่ เช่น การช่วยหาค่า k ที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูล การช่วยสุ่มประชากรสำหรับการค้นหาคำตอบ ในขั้นตอน Initialization การช่วยปรับค่าพารามิเตอร์ในขั้นตอนวิธีให้มีความเหมาะสมต่อการค้นหาคำตอบ เป็นต้น ในงานวิจัยนี้เลือกศึกษาขั้นตอนเชิงวิวัฒนาการรูปแบบความฉลาดแบบฝูง มีรายละเอียดดังนี้

ขั้นตอนวิธีความฉลาดแบบฝูงเป็นขั้นตอนวิธีที่เลียนแบบการทำงานประชากรในธรรมชาติที่มีจำนวนมหาศาล ยกตัวอย่างเช่น อาณาจักรของมด ในทางชีววิทยา มดเป็นสัตว์ข้อปล้อง และอยู่รวมกันเป็นกลุ่ม (Colony) จะมีพฤติกรรมการเดินตามกันเพื่อหาน้ำหวาน โดยใช้วิธีรับรู้สารฟีโรโมนของตัวก่อนหน้าให้เดินไปยังเส้นทางเดียวกันอย่างไม่ค่อยแตกแถว การรับรู้และประเมินสารฟีโรโมนของมดตัวก่อนหน้านี้ ใช้หลักการกวาดหาสารฟีโรโมนในทิศทางที่มีความเข้มอย่างมาก เพื่อเป็นหลักแสดงว่ามีมดตัวก่อนหน้าปล่อยสารในทิศทางนั้น ๆ เส้นทางใดที่มีสารฟีโรโมนเจือจาง อาจหมายถึงมดตัวก่อนหน้าได้เดินผ่านนานแล้ว ทำให้เส้นทางที่กลุ่มของมดจะเดินต่อกันเป็นฝูงได้ระยะทางที่สั้นที่สุด จากพฤติกรรมนี้มักนำมาเขียนเป็นขั้นตอนวิธีเพื่อใช้แก้ปัญหาการหาเส้นทางที่สั้นที่สุดนั่นเอง หรือขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคที่ได้รับแนวคิดจากการทำงานเป็นฝูงตาม

ธรรมชาติ อย่างเช่น ฝูงนกบินหาอาหารและมีการประชุมกลุ่มเพื่อหาอาหารที่ดีกว่าในตำแหน่งถัดไป แต่ละครั้งที่นกบินจะมีการแลกเปลี่ยนประสบการณ์ในกลุ่ม เพื่อกำหนดทิศทางใหม่จนเข้าใกล้แหล่งอาหารที่ดีที่สุด เป็นต้น จากตัวอย่างที่กล่าวไปจะเห็นว่าขั้นตอนวิธีความฉลาดแบบฝูงดังกล่าวมีข้อดี เนื่องจากเป็นแนวคิดที่เข้าใจง่าย สามารถใช้งานง่าย มีพารามิเตอร์จำนวนไม่มากจึงทำให้มีการนำขั้นตอนวิธีประเภทนี้ไปประยุกต์ใช้งานร่วมกับการวิเคราะห์กลุ่มข้อมูล

การศึกษาของ Elliackin Figueiredo ในปี ค.ศ.2019 [11] ได้รวบรวมเทคนิคความฉลาดแบบฝูง (Swarm intelligence algorithm; SI) สำหรับการจัดกลุ่มข้อมูล พบว่าภายใน 6 ปีที่ผ่านมามีการศึกษาจำนวนมาก ถึง 153 เรื่อง นำขั้นตอนวิธีความฉลาดแบบฝูงเพื่อช่วยวิเคราะห์กลุ่มข้อมูล และขั้นตอนวิธีที่พบบ่อย ได้แก่ การหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค 63.2% อาณานิคมผึ้งเทียม 10.4% และอาณาจักรมด 6.7% เป็นต้น

2.2.2.1 การหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle Swarm Optimization; PSO)

ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคเป็นขั้นตอนวิธีที่ได้รับความนิยมมากที่สุดเมื่อเปรียบเทียบความฉลาดแบบฝูงในรูปแบบอื่น ๆ อันเนื่องมาจากรูปแบบการทำงานของชุดคำสั่งแนวคิด ขั้นตอนวิธีมีความเรียบง่าย และใช้พารามิเตอร์จำนวนน้อย ในส่วนนี้ได้อธิบายพารามิเตอร์สำหรับขั้นตอนวิธีแบบกลุ่มอนุภาค ได้ดังนี้

กำหนดพารามิเตอร์ ได้แก่ p_i คือประชากรหรืออนุภาค (Particle) จำนวน n ตัว เป็นสมาชิกในฝูงอนุภาค (Swarm) $X_{i,t}$ สำหรับการเป็นกลุ่มประชากร 1 ฝูง $V_{i,t}$ คือความเร็วและทิศทางในการเคลื่อนที่ของประชากรแต่ละตัวไปยังตำแหน่งอาหารใหม่ ๆ และ t คือจำนวนรอบของการบินเพื่อเปลี่ยนตำแหน่งอาหาร นอกจากนี้ยังมีพารามิเตอร์ Pbest ได้แก่ฝูงประชากรที่มีค่าอาหารดีที่สุดเมื่อเทียบภายใน t รอบ และ Gbest คือประชากรตัวที่ได้ค่าอาหารดีที่สุดเมื่อเทียบทุกรอบการบิน นอกจากนี้ การบินเพื่อเปลี่ยนตำแหน่งอาหารเป็นการปรับให้ประชากรฝูงใหม่ที่อยู่ตำแหน่งใหม่มีคุณภาพที่คาดหวังว่าดีขึ้น โดยอาศัยการเรียนรู้และประสบการณ์ที่แลกเปลี่ยนกันภายในฝูงตามสมการ 9 และ 10 โดยที่พารามิเตอร์ ω คือค่าเฉลี่ยถ่วงน้ำหนัก q_1 และ q_2 เป็นค่าเฉลี่ยซึ่งเป็นค่าคงที่ และ r_1 และ r_2 เป็นค่าเฉลี่ยแบบสุ่ม

$$V_{i,t+1} = \omega \cdot V_{i,t} + q_1 r_1 (Pbest - X_{i,t}) + q_2 r_2 (Gbest - X_{i,t}) \quad (9)$$

$$X_{i,t+1} = X_{i,t} + V_{i,t+1} \quad (10)$$

2.2.2.2 สมการจุดประสงค์ (Objective function)

เนื่องจากขั้นตอนวิธีการหาค่าเหมาะสมที่สุด จำเป็นต้องมีสมการจุดประสงค์ หรือการตั้งปัญหาที่อธิบายได้ด้วยสมการคณิตศาสตร์เพื่อให้ขั้นตอนวิธีในรูปแบบต่าง ๆ ค้นหาคำตอบเพื่อบรรลุวัตถุประสงค์นั้น ดังนั้นการตั้งสมการจุดประสงค์สำหรับการหาคำตอบจึงเป็นสิ่งสำคัญ สำหรับการจัดกลุ่มข้อมูลซึ่งเป็นชื่อของยีนส์ที่มีลักษณะเป็นคำหรือตัวอักษร ไม่มีความหมาย ทำให้ไม่ไม่มีค่าประจำตำแหน่งของข้อมูล กระบวนการจัดกลุ่มข้อมูลประเภทข้อความนี้จึงมักทำการหาค่าความสัมพันธ์ของข้อมูลเพื่อใช้ในการคำนวณของขั้นตอนวิธีในชุดคำสั่ง เช่น การใช้วิธี TF factor วิธีการ TF-IDF หรือการวัดความเหมือนกันของกลุ่มข้อมูล ซึ่งหากอ้างอิงจากเครื่องมือ pathfindR พบว่าในทางด้านข้อมูลแบบชีวสารสนเทศมักใช้การวัดความเหมือนกันของกลุ่มข้อมูลด้วยสัมประสิทธิ์โคเฮนคัปปา ดังนั้นค่าความสัมพันธ์โคเฮนคัปปาแต่ละค่า เป็นค่าสัมพันธ์ร่วมกันระหว่างยีนส์ 1 และ ยีนส์ 2 จากการทบทวนวรรณกรรม [20] พบว่ามีสมการจุดประสงค์สำหรับการจัดกลุ่มประเภทข้อความแบ่งได้เป็นฟังก์ชันเกณฑ์ภายใน (Internal criterion function) ซึ่งพิจารณาความเป็นกลุ่มของข้อมูลแบบ intra cluster หรือความสัมพันธ์ระหว่างข้อมูลภายในกลุ่มเดียวกัน ได้แก่ สมการ $\mathcal{L}_1, \mathcal{L}_2$ และ $\mathcal{L}_3$ ฟังก์ชันเกณฑ์ภายนอก (External criterion function) พิจารณาความเป็นกลุ่มข้อมูลแบบ inter cluster หรือความสัมพันธ์ของข้อมูลต่างกลุ่ม ได้แก่ สมการ $\mathcal{E}_1$ และ $\mathcal{E}_2$ ฟังก์ชันแบบผสม (Hybrid criterion function) ได้แก่ สมการ $\mathcal{H}_1$ และ $\mathcal{H}_2$ และฟังก์ชันที่เกี่ยวข้องกับกราฟ (Graph based criterion function) ได้แก่ สมการ $\mathcal{G}_1$ และ $\mathcal{G}_2$ ดังต่อไปนี้

กำหนดให้ S เป็นเซ็ตที่มีสมาชิกข้อมูลจำนวน n ตัว สำหรับทำการแบ่งกลุ่มข้อมูลเป็น k กลุ่ม โดย k เป็นจำนวนกลุ่มข้อมูลที่ได้รับการแบ่งสูงสุด และ $\|D_r\|$ คือการคำนวณความเหมือนกันระหว่างคู่ข้อมูล i และ j (Similarity measure)

- Internal criterion function
  - $maximize\ \mathcal{L}_1 = \sum_{r=1}^{k} \frac{\|D_r\|^2}{n_r}$ (11)
  - $maximize\ \mathcal{L}_2 = \sum_{r=1}^{k} \|D_r\|$ (12)
  - $minimize\ \mathcal{L}_3 = \sum_{r=1}^{k} \frac{1}{n_r} \sum_{d_i, d_j \epsilon S_r} \|d_i - d_j\|^2$ (13)
- External criterion function
  - $maximize\ \mathcal{E}_1 = \sum_{r=1}^{k} \frac{\|D_r^t D\|}{\|D_r\|}$ (14)
  - $minimize\ \mathcal{E}_2 = \sum_{r=1}^{k} n_r \|C_r - C\|^2$ (15)

เมื่อ C เป็นเวคเตอร์เซนทรอยด์ของทุกคู่ในกลุ่มข้อมูล

- Hybrid criterion function
  - $maximize\ \mathcal{H}_1 = \frac{\mathcal{L}_1}{\mathcal{E}_1}$ (16)
  - $maximize\ \mathcal{H}_2 = \frac{\mathcal{L}_2}{\mathcal{E}_1}$ (17)
- Graph Based criterion function
  - $minimize\ \mathcal{G}_1 = \sum_{r=1}^{k} \frac{D_r^t D}{\|D_r\|^2}$ (18)
  - $minimize\ \mathcal{G}_2 = \sum_{r=1}^{k} \frac{cut(V_r, V - V_r)}{W(V_r)}$ (19)

    เมื่อ V_r คือเซ็ตคือจุดยอด หรือ Vertex ในแต่ละกลุ่มข้อมูล และ $W(V_r)$ คือผลรวมค่าถ่วงน้ำหนักของค่า vertex ในแต่ละกลุ่มข้อมูล

การศึกษาดังกล่าวได้ทดสอบและวิเคราะห์ผลการแบ่งกลุ่มแบบเคมีนจากข้อมูล 15 ประเภท อาทิ fbis hitech k1a k1b และ sports เป็นต้น โดยใช้สมการจุดประสงค์ตามที่ได้ยกตัวอย่างทั้งหมด และพิจารณาผลลัพธ์ที่ได้ด้วยค่า Entropy และ Purity พบว่าในแง่ของฟังก์ชันเกณฑ์ภายใน สมการ $\mathcal{L}_2$ ให้ผลดีกว่าสมการ $\mathcal{L}_1$ และ $\mathcal{L}_3$ นอกจากนี้สมการ $\mathcal{E}_1$ และ $\mathcal{G}_1$ ให้ผลดีด้วยเช่นกัน โดยที่การใช้สมการ $\mathcal{E}_1$ ให้ความสม่ำเสมอมากกว่าสมการ $\mathcal{G}_1$

บทที่ 3

วิธีการวิจัย

งานวิจัยนี้ได้ทำการสร้างตัวจัดกลุ่มข้อมูลแบบ PSO ที่ใช้สำหรับจัดกลุ่มข้อมูลชีวสารสนเทศที่มีลักษณะเป็นคู่ความสัมพันธ์ของกลุ่มยีนส์ไมโครอะเรย์ ในบทนี้จะอธิบายรายละเอียดที่เกี่ยวข้องเบื้องต้นที่ใช้ในการดำเนินงานวิจัย ขั้นตอนวิธีการทำงานของตัวจัดกลุ่มแบบ PSO รวมถึงวิธีการดำเนินการเปรียบเทียบการทำงานของตัวจัดกลุ่มแบบ PSO และแบบอื่น ในงานวิจัยนี้มีการใช้ข้อมูลยีนส์ไมโครอะเรย์เป็นข้อมูลเข้า (Input data) เพื่อการทดสอบจัดกลุ่มข้อมูลแบบ PSO สร้างโดยใช้ภาษาอาร์ จึงมีรายละเอียดที่เกี่ยวข้องเบื้องต้นสำหรับการดำเนินงานวิจัย ดังนี้

3.1 รายละเอียดที่เกี่ยวข้องเบื้องต้นสำหรับการดำเนินงานวิจัย

3.1.1 ภาษาโปรแกรมมิ่ง

ในงานวิจัยนี้ใช้ภาษาอาร์ (R language) จากโปรแกรม Rstudio เป็นเครื่องมือในการเขียนขั้นตอนวิธีการตัวจัดกลุ่มข้อมูลแบบ PSO เนื่องจากภาษาอาร์เป็นภาษาโปรแกรมมิ่งที่ใช้สำหรับวิเคราะห์ข้อมูลเชิงสถิติ สามารถใช้งานโปรแกรมได้ทั้งในระบบปฏิบัติการ Windows Linux และ MacOS นอกจากนี้ภาษาอาร์ยังมีเครื่องมือขั้นพื้นฐานจำนวนมาก พร้อมคู่มือการใช้งานของทุกเครื่องมือโดยผู้ใช้งานโปรแกรมสามารถค้นหารายละเอียดได้จากคลังเก็บเครื่องมือ CRAN web server (The Comprehensive R Archive Network) ซึ่งปัจจุบันมีเครื่องมือจำนวน 19,352 ชุด ที่สามารถเรียกใช้งานได้ ทั้งทางด้านสถิติและการจัดการข้อมูลทั่วไป รวมถึงเครื่องมือขั้นพื้นฐานทางด้านการเรียนรู้เครื่อง (Machine learning) ให้ใช้งานอีกด้วย อีกทั้งยังสามารถเขียนโปรแกรมใช้งานอื่น ๆ เพื่อใช้งานได้หลากหลายวัตถุประสงค์ (General-purpose programming language) ทำให้สามารถสร้างฟังก์ชันการทำงานใหม่ได้ตามจุดประสงค์ของผู้ใช้งานและผู้พัฒนา จากการศึกษาทบทวนงานวิจัย พบเครื่องมือทางชีวสารสนเทศจำนวนมาก ตัวอย่างเช่น GScluster และ pathfindR ที่ใช้เป็นเครื่องมือในการวิเคราะห์ข้อมูลยีนส์แบบไมโครอะเรย์ ในงานวิจัยนี้ได้ศึกษาเครื่องมือภาษาอาร์จากผู้พัฒนา Ege Ulgen ได้แก่ pathfindR ซึ่งเป็นเครื่องมือวิเคราะห์ข้อมูลชีวสารสนเทศแบบไมโครอะเรย์ เพื่อการค้นหากลุ่มยีนส์ที่สำคัญและจัดกลุ่มยีนส์ จะประกอบไปด้วยฟังก์ชันต่าง ๆ ตามเอกสารแนบ และใช้ข้อมูลยีนส์จากเครื่องมือ pathfindR.data จากผู้พัฒนาเดียวกัน เป็นข้อมูลเข้าสำหรับงานวิจัยนี้

3.1.2 ข้อมูลเข้า (Input data)

งานวิจัยนี้ใช้ข้อมูลเข้าจากเครื่องมือ pathfindR.data จากผู้พัฒนา Ege Ulgen ประกอบด้วยชุดข้อมูลที่สามารถใช้กับเครื่องมือ pathfindR ได้ทุกกระบวนการ โดยใช้ฟังก์ชัน RA_output เป็นข้อมูลเข้าของงานวิจัยนี้ จะใช้ชุดข้อมูล Up_regulated และ Down_regulated ของยีนส์ที่มีความสำคัญจากกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA) นำมาหาความสัมพันธ์ด้วยสัมประสิทธิ์โคเฮนคัปปา [5] ดังสมการ 20 ใช้เป็นฟังก์ชันการวัดความเหมือนกันของข้อมูล ในการจัดกลุ่มข้อมูลต่อไป

$$Kappa(A,B) = 1 - \frac{O - E}{1 - E} \tag{20}$$

เมื่อ O คือค่าความเหมือนในการตัดสินใจที่เกิดขึ้นจริง (Observed agreement) ของยีนส์เซ็ต A และ B

$$O = \frac{|A \cap B| + |(A \cup B)^c|}{|U|} \tag{21}$$

และ E คือความเหมือนในการตัดสินใจที่เกิดขึ้นโดยบังเอิญ (Chance agreement) ของยีนส์เซ็ต A และ B โดยที่ U คือเซ็ตรวมที่มียีนส์ทั้งหมดเป็นสมาชิก

$$E = \frac{|A| \cdot |B| + |A^c| \cdot |B^c|}{|U|^2} \tag{22}$$

3.1.3 เครื่องมือที่ใช้ในโปรแกรม Rstudio

งานวิจัยนี้มีการใช้เครื่องมือ pathfindR [8]เป็นหนึ่งในเครื่องมือทางชีวสารสนเทศที่ใช้วิเคราะห์คุณสมบัติสำคัญของยีนส์ในระดับการแสดงออก (phenotype) ประกอบไปด้วยหลายกระบวนการ ในงานวิจัยนี้จะใช้งานชุดคำสั่ง pathfindR::create_kappa_matrix เพื่อคำนวณสัมประสิทธิ์โคเฮนคัปปา ใช้เป็นฟังก์ชันความเหมือนกันของข้อมูลในการจัดกลุ่มข้อมูลในขั้นตอนถัดไป และมีการใช้ฟังก์ชันการจัดกลุ่มข้อมูลแบบลำดับขั้น (Hierarchical clustering) จาก pathfindR::hierarchical_term_clustering โดยผู้พัฒนา Ege Ulgen ซึ่งมีการเรียกใช้เครื่องมือพื้นฐานจาก stats package ได้แก่ ชุดคำสั่ง hclust ซึ่งมีพารามิเตอร์ที่ต้องกำหนดก่อนการใช้งานจำนวน 3 พารามิเตอร์ ได้แก่ ฟังก์ชันระยะทางหรือฟังก์ชันความเหมือนกันของข้อมูล วิธีการรวมกลุ่มข้อมูล แบบ average link และจำนวนกลุ่มที่ต้องการแบ่งข้อมูล (ค่า k) ขั้นตอนวิธีที่ผู้พัฒนาสร้างขึ้นนี้ มีการหาจำนวนกลุ่มที่ต้องการแบ่งให้เหมาะสมด้วยการค่า silhouette index ที่มากที่สุดจากสมการที่ 8

นอกจากนี้ยังมีการใช้เครื่องมือสำหรับจัดกลุ่มข้อมูลอื่นในโปรแกรม Rstudio ที่ใช้ในงานวิจัย ได้แก่ stats::kmeans (K-means clustering) เป็นเครื่องมือขั้นพื้นฐานจาก stats package มีพารามิเตอร์ที่จำเป็นต้องกำหนด จำนวน 2 พารามิเตอร์ ได้แก่ฟังก์ชันระยะทางหรือฟังก์ชันความเหมือนกันของข้อมูลและจำนวนกลุ่มที่ต้องการแบ่ง (ค่า k)

3.2 ขั้นตอนการดำเนินงานวิจัย

3.2.1 การสร้างตัวจัดกลุ่มข้อมูลแบบ PSO (PSO-based clustering)

ตัวจัดกลุ่มข้อมูลนี้สร้างโดยมีกระบวนการศึกษาลักษณะข้อมูลเข้าที่ใช้ในงานวิจัย ศึกษาตัวจัดกลุ่มข้อมูล (Clustering) และขั้นตอนวิธีเชิงวิวัฒนาการแบบความฉลาดแบบกลุ่ม โดยใช้ขั้นตอนที่มีลักษณะแบบกลุ่มอนุภาค (Particle swarm optimization; PSO) มีรายละเอียดดังต่อไปนี้

3.2.1.1 ศึกษาลักษณะของข้อมูลเข้าที่ใช้ในงานวิจัย

งานวิจัยนี้ใช้ข้อมูลขาเข้าจากฟังก์ชัน RA_output โดยเครื่องมือ pathfindR.data เป็นชุดข้อมูลที่ได้จากการทำไมโครอะเรย์เพื่อที่จะศึกษาลักษณะการแสดงออกในระดับยีนส์จากนิวเคลียสของเซลล์เม็ดเลือดขาวแบบเดี่ยวของผู้ป่วยโรคข้ออักเสบรูมาตอยด์ 18 คน เทียบกับผู้มีสุขภาพดี 15 คน (Rheumatoid arthritis dataset; GSE15573) ข้อมูลชุดนี้ต้องผ่านกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA) โดยเครื่องมือ pathfindR ตั้งแต่กระบวนการเตรียมข้อมูล (Data preprocessing) ด้วยชุดคำสั่ง input_processing จนถึงขั้นตอนการเติมข้อมูลเพิ่มเติมให้กลุ่มยีนส์ (Function annotation) โดยชุดคำสั่ง annotate_term_genes ซึ่งทำให้ได้เป็นข้อมูลที่ประกอบไปด้วยข้อมูลที่สำคัญในการจัดกลุ่มข้อมูล ดังรูปที่ 3.1

	ID	Term_Description	Fold_Enrichment	occurrence	support	lowest_p	highest_p	Up_regulated	Down_regulated
1	hsa05415	Diabetic cardiomyopathy	3.2463571	10	0.075290698	1.951173e-13	1.951173e-13	NCF4, MMP9, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2...	ATP2A2, MTOR, PDHA1, PDHB, VDAC1, SLC25A5, PARP1
2	hsa04130	SNARE interactions in vesicular transport	4.5990059	10	0.011835977	8.007088e-08	8.007088e-08	STX6	STX2, BET1L, SNAP23
3	hsa03410	Base excision repair	5.7487574	1	0.005319149	1.236504e-07	1.236504e-07	POLE4	MUTYH, APEX2, POLD2, PARP1
4	hsa04064	NF-kappa B signaling pathway	2.9185999	10	0.050259176	1.258581e-07	1.258581e-07	LY96	PRKCQ, CARD11, TICAM1, IKBKB, PARP1, UBE2I, CSNK2A2
5	hsa03040	Spliceosome	3.6887860	10	0.047745694	2.230840e-07	4.456242e-05	SF3B6, LSM3, BUD31	SNRPB, SF3B2, U2AF2, PUF60, SNU13, DDX23, EIF4A3, HNR...
6	hsa04714	Thermogenesis	2.5174653	10	0.044198222	2.543550e-07	2.543550e-07	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	ADCY7, CREB1, KDM1A, SMARCA4, ACTG1, ACTB, ARID1A, ...
7	hsa03013	RNA transport	2.6017234	10	0.018411145	2.652129e-07	4.190199e-05	NUP214	NUP62, NUP93, RANGAP1, UBE2I, SUMO3, GEMIN4, EIF2S3, ...
8	hsa00190	Oxidative phosphorylation	2.9437603	10	0.017700501	3.407516e-07	3.407516e-07	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C, ATP6...	ATP6V0E2
9	hsa05012	Parkinson disease	2.4637532	10	0.053731822	4.781688e-07	4.781688e-07	DDIT3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7...	UBE2G1, PSMD7, CALM3, CALM1, VDAC1, SLC25A5, TUBB
10	hsa05020	Prion disease	2.0668684	10	0.042219004	1.003259e-06	1.003259e-06	DDIT3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7...	VDAC1, SLC25A5, PSMD7, CREB1, CSNK2A2, TUBB
11	hsa04932	Non-alcoholic fatty liver disease	2.2465539	10	0.011661907	1.742015e-06	1.742015e-06	DDIT3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	IKBKB, FASLG
12	hsa04659	Th17 cell differentiation	3.2214735	10	0.034683240	2.371132e-06	2.371132e-06		MTOR, JAK1, HLA-DPA1, NFATC3, PRKCQ, IKBKB, GATA3, IL2...
13	hsa03430	Mismatch repair	4.9489303	10	0.016474093	4.543218e-06	3.540330e-02		MLH1, RPA1, POLD2
14	hsa04260	Cardiac muscle contraction	2.2856505	10	0.038864943	4.561107e-06	4.561107e-06	UQCRQ, COX6A1, COX7A2, COX7C	ATP2A2
15	hsa04722	Neurotrophin signaling pathway	2.8938660	10	0.041410245	8.838821e-06	1.507531e-05		SH2B3, CRKL, FASLG, CALM3, CALM1, ABL1, MAGED1, IRAK2...
16	hsa05166	Human T-cell leukemia virus 1 infection	2.4255031	10	0.023122160	1.112040e-05	1.112040e-05		TRRAP, NFATC3, IL2RB, JAK1, VDAC1, SLC25A5, ANAPC1, IKB...
17	hsa04630	JAK-STAT signaling pathway	1.4500051	10	0.011594300	1.709793e-05	1.709793e-05		IL2RB, IL10RA, IL27RA, JAK1, PIAS3, MTOR
18	hsa05167	Kaposi sarcoma-associated herpesvirus infection	1.9761354	10	0.034582420	1.711324e-05	6.189907e-04		TICAM1, JAK1, ZFP36, IKBKB, GNB1, MTOR, CALM3, CALM1, ...
19	hsa04931	Insulin resistance	1.4317660	10	0.017291210	3.239816e-05	3.878218e-02		MTOR, IKBKB, PRKCQ, CREB1
20	hsa04530	Tight junction	2.3566335	10	0.018018181	3.947621e-05	3.947621e-05	CLDN9, MYL6B, MYL6	ARHGAP17, SCRIB, TJAP1, SLC9A3R1, ACTG1, ACTB, ITGB1
21	hsa04210	Apoptosis	1.9673525	10	0.023122160	4.037994e-05	4.037994e-05	DDIT3	FASLG, ACTG1, ACTB, PARP1, DFFB, IKBKB
22	hsa05163	Human cytomegalovirus infection	2.0417112	10	0.044265101	4.153794e-05	1.960054e-03		MTOR, IKBKB, GNB1, CALM3, CALM1, NFATC3, ADCY7, JAK1,...
23	hsa05203	Viral carcinogenesis	1.8396024	10	0.034683240	4.461499e-05	4.461499e-05	GTF2B	CREB1, JAK1, SCRIB, RBL2, HDAC1, DNAJA3, SRF
24	hsa05130	Pathogenic Escherichia coli infection	2.3469154	10	0.067814611	9.981311e-05	9.981311e-05	TLR5, GAPDH, CLDN9	ARF1, ACTG1, ACTB, SLC9A3R1, TUBB, ABL1, ITGB1, IKBKB, F...
25	hsa04971	Gastric acid secretion	2.4961710	10	0.029154767	1.027901e-04	1.027901e-04		ACTG1, ACTB, CALM3, CALM1, ADCY7
26	hsa05131	Shigellosis	2.8235757	10	0.053381459	1.181951e-04	1.660836e-04	HK3, TLR5, CBX3, RRAGD	ITGB1, CRKL, ACTG1, ACTB, PFN1, ARF1, VDAC1, IKBKB, PRK...
27	hsa05416	Viral myocarditis	2.5294533	10	0.011800334	1.197583e-04	1.197583e-04		ABL1, ACTG1, ACTB, HLA-DPA1

รูปที่ 3.1 แสดงตัวอย่างข้อมูลขาเข้าสำหรับการสร้างตัวจัดกลุ่มแบบ PSO

3.2.1.2 ศึกษาวิธีการจัดกลุ่มข้อมูลที่ใช้กับข้อมูลด้านไมโครอะเรย์ในกระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์

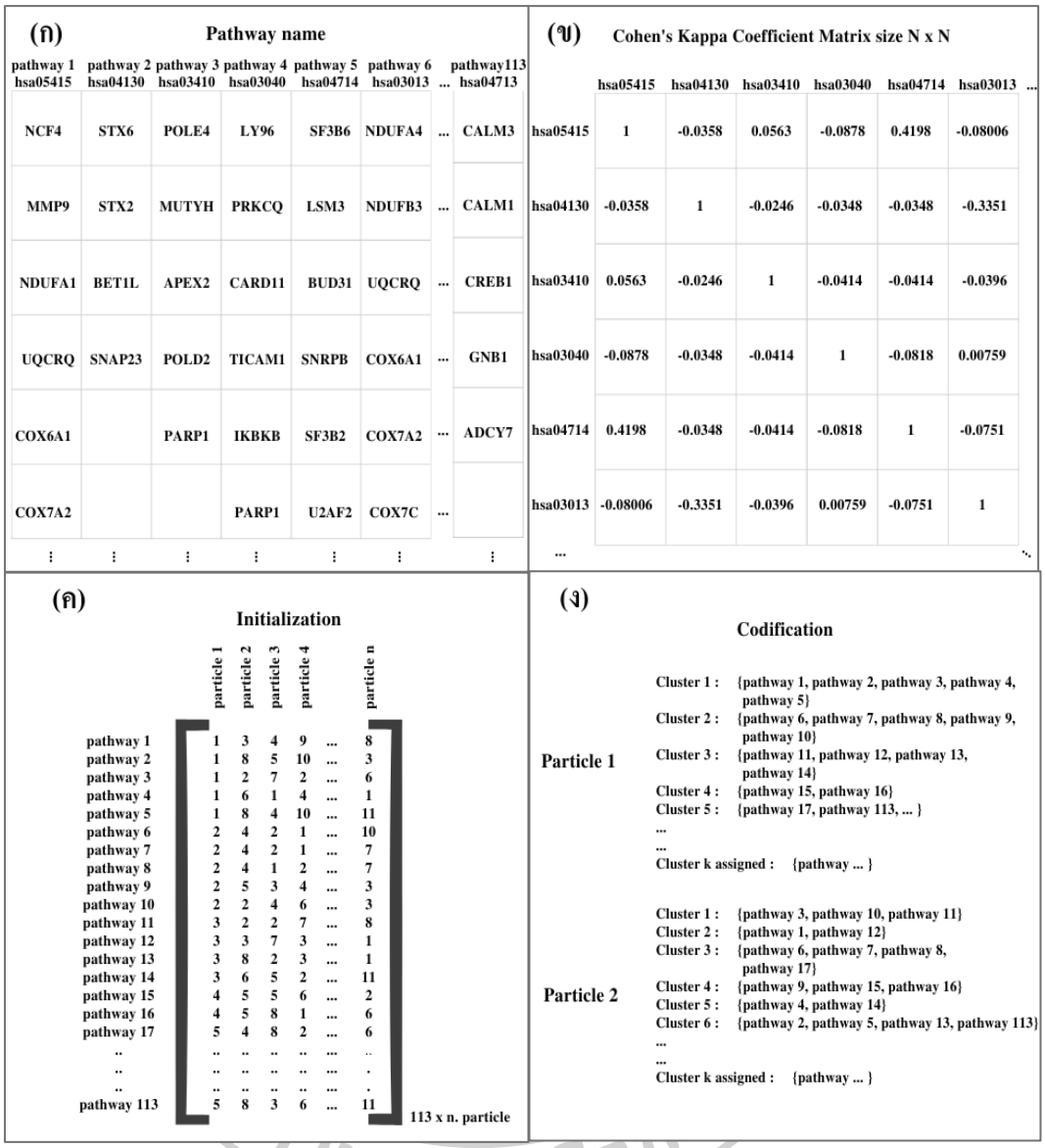
เมื่อศึกษาวิธีการจัดกลุ่มที่ใช้กระบวนการวิเคราะห์ความสำคัญของกลุ่มยีนส์ พบว่าจะต้องหาความสัมพันธ์ของข้อมูลยีนส์ที่ได้จากชื่อยีนส์ที่มีการแสดงพฤติกรรมอย่างโดดเด่นทั้งแบบมากขึ้น และลดลง (Up_regulated และ Down_regulated) ที่มีอยู่ในรหัสกลุ่มยีนส์ (ID) หรือคำอธิบายกลุ่มยีนส์ (Term_description) เดียวกัน ซึ่งเป็นข้อมูลประเภทข้อความที่ไม่มีความหมาย ดังรูป 3.1 งานวิจัยนี้ใช้สัมประสิทธิ์โคเฮนคัปปาเพื่อหาความสอดคล้องกันของกลุ่มข้อมูล (Similarity) และนำมาใช้เป็นฟังก์ชันการวัดความเหมือนกันระหว่างข้อมูลสำหรับการจัดกลุ่ม ในงานวิจัยนี้สามารถใช้ชุดคำสั่ง pathfindR::create_kappa_mat ในการคำนวณหาค่าความสอดคล้องกันระหว่างกลุ่มยีนส์

3.2.1.3 ออกแบบตัวจัดกลุ่มแบบ PSO และกำหนดสมการจุดประสงค์ (Objective function)

การออกแบบตัวจัดกลุ่มแบบ PSO สามารถแบ่งได้ 3 กระบวนการดังนี้

3.2.1.3.1 การสุ่มฝูงประชากร และการเข้ารหัสประชากร (Initialization and Codification)

ขั้นตอนแรกเริ่มต้นจากการสุ่มฝูงประชากร (Swarm) ที่มีจำนวนประชากร (Particle) ในฝูงตามผู้ใช้กำหนด ในแต่ละประชากรจะมีจำนวนอนุภาคจำนวนเท่ากับจำนวนกลุ่ม pathway และใช้การเข้ารหัสอนุภาคแบบสุ่มเลขโดดแทนแต่ละอนุภาค (Label-based encoding) ซึ่งเป็นวิธีที่เข้าใจง่าย และต้องกำหนดค่า k เริ่มต้น เป็นเลขโดดที่มีค่ามากที่สุดในการแทนค่าอนุภาคในประชากรนั้น ๆ เช่น หากกำหนดค่า k เริ่มต้น มีค่า 15 โดยที่ประชากรหนึ่ง ๆ มีอนุภาคประชากร 113 ตัว อนุภาคแต่ละตัวจะมีเลขโดดประจำอนุภาคได้ตั้งแต่ 1 ถึง 15 เป็นต้น จะสามารถถอดรหัสอนุภาคของแต่ละประชากรได้โดยกลุ่มที่มีค่า k เดียวกัน เป็นสมาชิกในกลุ่มเดียวกัน ดังรูปที่ 3.2 (ค)



รูปที่ 3.2 (ก) ตัวอย่างข้อมูลกลุ่มยีนส์ (pathway) (ข) ตารางค่าความสอดคล้องระหว่างข้อมูลกลุ่มยีนส์ด้วยการคำนวณสัมประสิทธิ์โคเฮนคัปปา (ค) ตัวอย่างขั้นตอนการสุ่มฝูงประชากร 1 ฝูง จำนวนประชากร n ตัว แต่ละประชากรประกอบไปด้วย 113 อนุภาค (ง) การเข้ารหัสประชากรตามค่า k เริ่มต้นที่กำหนด

3.2.1.3.2 ทิศทางการเคลื่อนที่ของฝูง

เนื่องจากตัวจัดกลุ่มแบบ PSO เป็นขั้นตอนวิธีหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle Swarm Optimization) จึงมีพฤติกรรมการเคลื่อนที่ของฝูงไปยังทิศทางต่าง ๆ ในปริภูมิคำตอบ โดยกำหนดให้ฝูงประชากรเคลื่อนที่เปลี่ยนตำแหน่งไปเพื่อหาตำแหน่งที่เหมาะสมที่จะเป็นคำตอบของปัญหา ซึ่งอาศัยข้อมูลและประสบการณ์ของประชากรภายในฝูงเป็นตัวปรับทิศทางการ

เคลื่อนที่ของฝูงในแต่ละรอบ ดังสมการที่ 9 และทำให้สมการจุดประสงค์ (Objective function) ค่ามากที่สุด หรือน้อยที่สุด ขึ้นกับลักษณะของปัญหางานวิจัย ซึ่งในงานวิจัยนี้ได้กำหนดสมการจุดประสงค์ $\mathcal{L}_2$ ดังสมการที่ 12

$$maximize\ \mathcal{L}_2 = \sum_{i=1}^{k} \sqrt{\sum_{D_r \in C_i} \|D_r\|} \quad (12)$$

โดยที่ $\|D_r\|$ คือการคำนวณความเหมือนกันระหว่างคู่ข้อมูลในกลุ่มเดียวกัน (Similarity measurement)

3.2.1.3.3 ความเร็วของการเคลื่อนที่ของฝูง

เมื่อฝูงประชากรมีการเคลื่อนที่เปลี่ยนไปตำแหน่งต่าง ๆ เพื่อหาตำแหน่งที่เป็นค่าเหมาะสมที่สุดในแต่ละครั้ง จะมีการกำหนดความเร็วในการเคลื่อนที่เพื่อเปลี่ยนแปลงตำแหน่งใหม่ โดยได้รับอิทธิพลจากตำแหน่งที่ดีที่สุดในครั้งก่อนหน้า (Pbest และ Gbest) ดังสมการ 9 และ 10

$$V_{i,t+1} = \omega \cdot V_{i,t} + q_1 r_1 (Pbest - X_{i,t}) + q_2 r_2 (Gbest - X_{i,t}) \quad (9)$$

$$X_{i,t+1} = X_{i,t} + V_{i,t+1} \quad (10)$$

เมื่อ $X_{i,t}$ เป็นตำแหน่งของประชากรตัวที่ i ในรอบปัจจุบัน t $V_{i,t}$ เป็นความเร็วในการเคลื่อนที่ของประชากรตัวที่ i ในรอบปัจจุบัน t ω คือตัวประกอบถ่วงน้ำหนัก (Weight factor) q_1 และ q_2 คือตัวประกอบถ่วงน้ำหนักที่เป็นค่าคงที่ ในงานวิจัยนี้ได้กำหนดให้อยู่ในช่วงระหว่าง 0.2 ถึง 0.6 โดยให้ผลกับพจน์ที่ได้รับอิทธิพลจากประสบการณ์การเคลื่อนที่โดยรวมภายในฝูงในรอบนั้น ๆ (Pbest) เทียบกับครั้งก่อนหน้า และให้ผลกับอิทธิพลจากประสบการณ์การเคลื่อนที่ที่ดีที่สุดจากทุกรอบที่ผ่านมา (Gbest) ตามลำดับ และ r_1 และ r_2 เป็นตัวประกอบค่าคงที่แบบสุ่มทุก ๆ รอบของการเคลื่อนที่อยู่ในช่วงระหว่าง 0 ถึง 1

3.2.1.4 ขั้นตอนวิธีของตัวจัดกลุ่มแบบ PSO (PSO-based clustering)

ตัวจัดกลุ่มแบบ PSO (PSO-based clustering) มีขั้นตอนวิธีการตามวิธีการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาคโดยทั่วไป ดังนี้

- สร้างฝูงประชากรและกำหนดค่าพารามิเตอร์
  - สร้างฝูงประชากรแรก $X_{i,t}$ เพื่อเป็นตัวแทนคำตอบในรอบปัจจุบัน โดยที่แต่ละประชากรมีจำนวนอนุภาค 113 ตัว หรือมีจำนวนอนุภาค

เท่ากับจำนวนข้อมูลกลุ่มยีนส์ ขนาดของฝูงประชากรจะมีค่าเท่ากับจำนวนอนุภาค x จำนวนประชากร

- กำหนดค่า k เริ่มต้นสูงสุด และให้ค่า k เริ่มต้นแก่อนุภาคทั้ง 113 ตัว
- กำหนดจำนวนรอบของการคำนวณ
- สร้างความเร็วในการเคลื่อนที่ครั้งแรกของประชากรแต่ละตัว $V_{i,t}$ ซึ่งมีขนาดเท่ากับขนาดของ $X_{i,t}$
- กำหนดค่าพารามิเตอร์ ω, q_1, q_2, r_1 และ r_2 เพื่อการคำนวณในสมการที่ 9
- จัดกลุ่มอนุภาคที่ได้รับค่า k เริ่มต้นเป็นค่าเดียวกัน เพื่อที่จะคำนวณฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ ในแต่ละประชากร

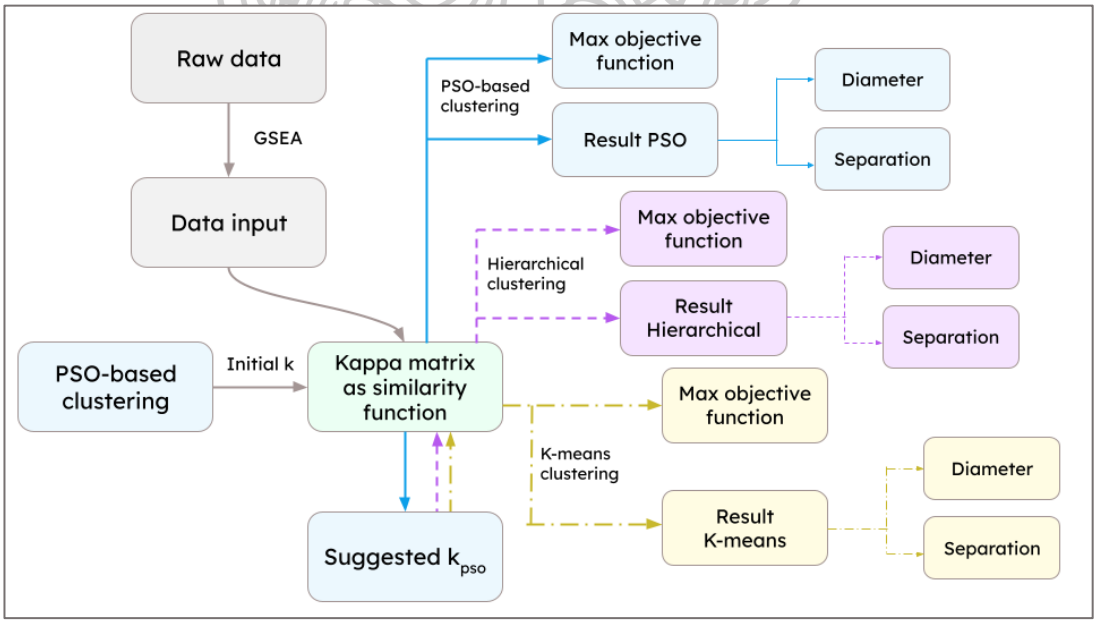
- ประเมินค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$
  - คำนวณค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ ในทุกประชากร โดยคำนวณกลุ่มอนุภาคที่ได้รับค่า k เริ่มต้นเป็นจำนวนเดียวกัน
- เลือกตำแหน่งของประชากรที่ดีที่สุดในฝูงในรอบการคำนวณปัจจุบัน (Pbest) และตำแหน่งประชากรที่ดีที่สุดเมื่อเทียบกับทุกฝูงในรอบก่อนหน้า (Gbest)
  - ปรับปรุงตำแหน่ง Pbest จากค่าฟังก์ชันจุดประสงค์ในรอบปัจจุบัน
  - ถ้าค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ ในรอบปัจจุบัน มีค่ามากกว่ารอบก่อนหน้า ให้ตำแหน่ง Pbest ในรอบปัจจุบัน เป็นตำแหน่ง Pbest ในรอบปัจจุบัน
  - ถ้าค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ ในรอบปัจจุบัน มีค่าน้อยกว่ารอบก่อนหน้า ให้ตำแหน่ง Pbest ในรอบปัจจุบัน เป็นตำแหน่ง Pbest ในรอบก่อนหน้า
  - ปรับปรุงตำแหน่ง Gbest
  - ถ้าค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ ในรอบปัจจุบัน มีค่ามากกว่ารอบที่ได้จากตำแหน่ง Gbest ในรอบปัจจุบัน ให้ตำแหน่งดังกล่าวเป็นตำแหน่ง Gbest
  - ถ้าค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ ในรอบปัจจุบัน มีค่าน้อยกว่ารอบที่ได้ตำแหน่ง Gbest ในรอบปัจจุบัน ให้ตำแหน่ง Gbest ในรอบปัจจุบันเป็นตำแหน่ง Gbest

- คำนวณความเร็วในการเคลื่อนที่ครั้งใหม่และตำแหน่งใหม่ของประชากร
  - คำนวณความเร็วครั้งใหม่ในการเคลื่อนที่ และตำแหน่งใหม่ของประชากรจากสมการ 9 และ 10 ตามลำดับ ตำแหน่งของประชากรใหม่จะเป็นเสมือนตัวแทนคำตอบใหม่ฝูงใหม่

3.2.2 การเปรียบเทียบการจัดกลุ่มข้อมูลด้วยตัวจัดกลุ่มข้อมูลรูปแบบต่าง ๆ

วิทยานิพนธ์นี้ได้ทำการศึกษาการจัดกลุ่มข้อมูลแบบ PSO เปรียบเทียบการจัดกลุ่มรูปแบบอื่น ๆ โดยพิจารณาที่ผลค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ และคุณสมบัติของการเป็นกลุ่มข้อมูล ได้แก่ ความกะทัดรัดของกลุ่มข้อมูล (Compactness) และระยะห่างระหว่างข้อมูล (Separation) ซึ่งเป็นการตรวจสอบกลุ่มข้อมูลแบบภายใน (Internal clustering validation) ดังรูปที่ 3.3 แสดงภาพรวมการดำเนินงานวิจัยเปรียบเทียบการจัดกลุ่มข้อมูลยีนส์ไมโครอะเรย์ด้วยตัวจัดกลุ่มรูปแบบต่าง ๆ ตัวจัดกลุ่มข้อมูลทั้งหมดที่ใช้สำหรับงานวิจัยนี้มี 3 รูปแบบ ได้แก่

- ตัวจัดกลุ่มแบบ Hierarchical (Hierarchical clustering) เป็นการวิเคราะห์กลุ่มแบบลำดับชั้น โดยใช้อัลกอริธึมจาก R package::pathfindR
- ตัวจัดกลุ่มแบบเคมีน (K-means clustering) โดยใช้อัลกอริธึมทั่วไปสำหรับภาษา R
- ตัวจัดกลุ่มแบบ PSO ที่ถูกสร้างขึ้นในงานวิจัยนี้



รูปที่ 3.3 ภาพรวมการดำเนินงานวิจัยเปรียบเทียบการจัดกลุ่มข้อมูลยีนส์ไมโครอะเรย์ด้วยตัวจัดกลุ่มรูปแบบต่าง ๆ

รูปที่ 3.3 แสดงภาพรวมการดำเนินงานวิจัยเปรียบเทียบ เริ่มจากการเตรียมข้อมูลขาเข้าให้อยู่ในรูปฟังก์ชันการวัดความเหมือนกันของข้อมูล และใช้ตัวจัดกลุ่มประเภท PSO เพื่อทำการจัดกลุ่ม จำนวน initial k กลุ่ม คือค่าที่ผู้ใช้ต้องกำหนดจำนวนกลุ่มมากสุดที่ต้องการแบ่งข้อมูล ผลลัพท์ที่ได้จากการจัดกลุ่มข้อมูลแบบ PSO นี้ จะแนะนำค่า k ค่าใหม่ (kpso) หรืออาจเป็นค่าเดิมซึ่งอยู่ในขอบเขตภายใต้เงื่อนไขการทำให้สมการจุดประสงค์ $\mathcal{L}_2$ มีค่ามากที่สุด นอกจากนี้ยังได้ผลลัพธ์ค่าจุดประสงค์มากที่สุดที่สามารถคำนวณได้และได้ผลการจัดกลุ่มข้อมูล โดยนำไปสู่การเปรียบเทียบคุณสมบัติของกลุ่มข้อมูลอีกด้วย เมื่อการดำเนินการจัดกลุ่มด้วยตัวจัดกลุ่มแบบ PSO ให้ค่า kpso จึงใช้เป็นพารามิเตอร์นี้เพื่อเริ่มต้นการจัดกลุ่มแบบลำดับขั้นและเคมีน และเปรียบเทียบค่าฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ รวมถึงคุณสมบัติของกลุ่มข้อมูลในลำดับถัดไป

บทที่ 4

ผลการวิจัยและวิจารณ์

งานวิจัยนี้ได้ทำการทดสอบตัวจัดกลุ่มแบบ Particle Swarm Optimization (PSO-based clustering) โดยเปรียบเทียบกับตัวจัดกลุ่มประเภทอื่น ได้แก่ ตัวจัดกลุ่มแบบลำดับขั้น (Hierarchical clustering) จากเครื่องมือ pathfindR ของผู้พัฒนา Ege Ulgen [8] และตัวจัดกลุ่มเคมีน (Kmeans clustering) จากเครื่องมือ stats ซึ่งเป็นเครื่องมือมาตรฐานในภาษาอาร์ ในงานวิจัยนี้ได้มีการกำหนดค่า k เริ่มต้น (Initial k) 3 รูปแบบ ได้แก่ 19 25 และ 30 ดังแสดงในตารางที่ 4.1 การทดสอบครั้งแรกจะดำเนินการจำนวน 60 ครั้ง เพื่อสังเกตผลของตัวจัดกลุ่มแต่ละประเภท เมื่อกำหนดค่า k เริ่มต้นที่ 19 เนื่องจากทำการศึกษาโดยอ้างอิงผลของค่า k ที่เหมาะสมที่ได้รับการแนะนำจากเครื่องมือ pathfindR การทดสอบครั้งที่ 2 จะเป็นการทดสอบตัวจัดกลุ่มทุกประเภทเมื่อกำหนดให้ k เริ่มต้นมีค่าเท่ากับ 19 เท่านั้น จำนวน 71 ครั้ง ร่วมกับเพิ่มค่า k เริ่มต้น นอกเหนือจาก 19 อย่างละ 60 ครั้ง และการทดสอบครั้งที่ 3 เป็นการทดสอบเพิ่มจำนวนอีก 60 ครั้ง ในทุกรูปแบบของตัวจัดกลุ่มและทุกค่าเริ่มต้นที่กำหนดไว้ รวมการทดสอบทั้งหมดเป็นจำนวน 431 ครั้ง เป็นการทดสอบโดยมีภาพรวมระบบการดำเนินงานตามภาพรวมดังแสดงรูปที่ 3.3 ในบทนี้จะแสดงการผลงานวิจัยทางด้านค่าสมการจุดประสงค์ $\mathcal{L}_2$ คุณสมบัติความเป็นกลุ่มของข้อมูล โดยพิจารณาความยาวเส้นผ่าศูนย์กลางของกลุ่มข้อมูล (Diameter) และระยะห่างระหว่างกลุ่มข้อมูล (Separation) และความหลากหลายของคำตอบที่ได้จากการแบ่งกลุ่มข้อมูลด้วยประเภท PSO มีรายละเอียดของผลงานวิจัยดังนี้

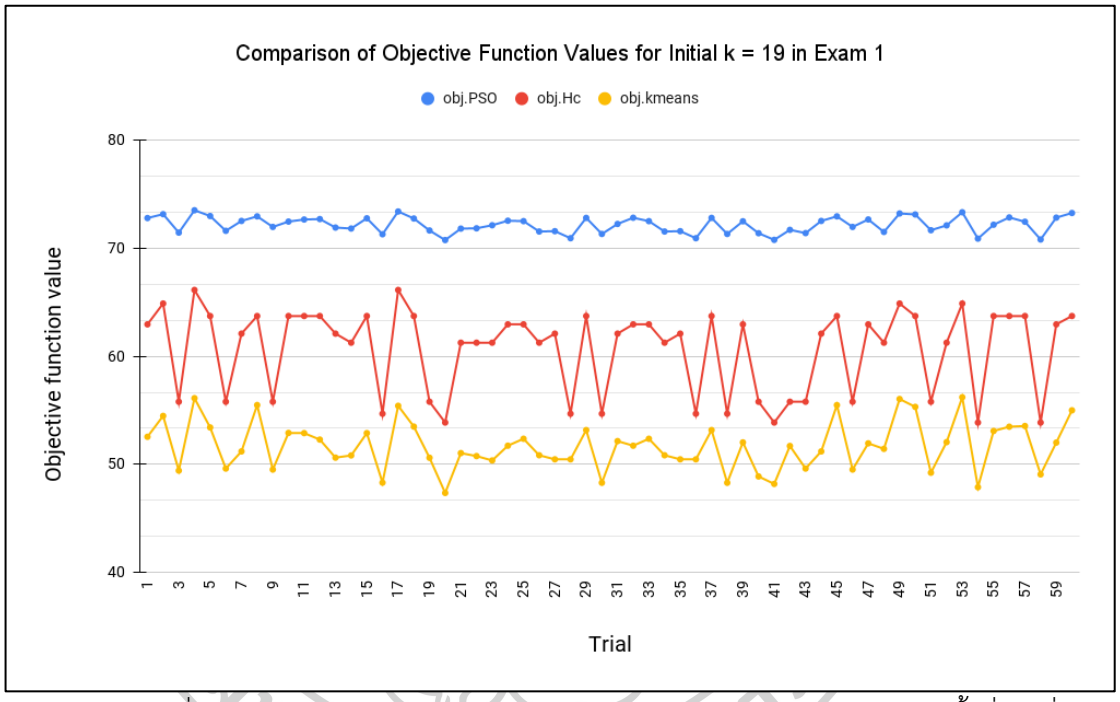
ตารางที่ 4.1 การทดสอบตัวจัดกลุ่มทั้ง 3 ประเภท ที่ค่า k เริ่มต้นต่าง ๆ ในงานวิจัยนี้

การทดสอบ	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	รวม
initial k = 19	60	71	60	191
initial k = 25	0	60	60	120
initial k = 30	0	60	60	120
รวม	60	191	180	431

4.1 ค่าสมการจุดประสงค์

4.1.1 การทดสอบครั้งที่ 1 เมื่อกำหนดค่า k เริ่มต้นที่ 19 จำนวน 60 ครั้ง

อ้างอิงจากผลการจัดกลุ่มโดยเครื่องมือ pathfindR ที่ให้ค่า k ที่เหมาะสมในการแบ่งกลุ่มยีนส์ Rheumatoid arthritis ด้วยการแบ่งกลุ่มแบบลำดับขั้น ได้แก่ 19 กลุ่ม การทดสอบครั้งที่ 1 จึงได้เลือกค่า k เท่ากับ 19 เป็นค่าเริ่มต้น ทดสอบทั้งหมด 60 ครั้ง โดยเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ ได้ผลดังรูป 4.1



รูปที่ 4.1 ผลเปรียบเทียบค่าสมการจุดประสงค์ $\mathcal{L}_2$ ในการทดสอบครั้งที่ 1 เมื่อกำหนดค่า k เริ่มต้นเท่ากับ 19 จากตัวจัดกลุ่มแบบ PSO แบบลำดับขั้น และแบบเคมีน

จากรูป แสดงให้เห็นภาพรวมค่าสมการจุดประสงค์ $\mathcal{L}_2$ ที่ได้จากการจัดกลุ่มแบบ PSO มีค่ามากกว่าแบบลำดับขั้นและเคมีนทุกกรณี โดยพบว่าการจัดกลุ่มแบบ PSO ให้ค่าสมการจุดประสงค์มากที่สุดมีค่า 73.5077 สามารถแบ่งข้อมูลได้ 8 กลุ่ม การจัดกลุ่มแบบลำดับขั้นให้ค่าสมการจุดประสงค์มากสุดมีค่า 66.1338 เมื่อแบ่งข้อมูลเป็น 8 กลุ่มเช่นเดียวกับแบบ PSO ในขณะที่ตัวจัดกลุ่มแบบเคมีนจะให้ค่าสมการจุดประสงค์สูงสุดเมื่อแบ่งข้อมูลเป็น 9 กลุ่ม มีค่า 56.2091 ดังตารางที่ 4.2 แสดงถึงการทดสอบครั้งที่ 1 พบว่าการจัดกลุ่มข้อมูลแบบ PSO สามารถแบ่งข้อมูลเป็น 10 กลุ่มได้มากที่สุดถึง 15 ใน 60 ครั้ง และพบว่าสามารถแบ่งข้อมูลเป็นกลุ่มได้หลายรูปแบบ ไม่เพียงเฉพาะ

10 กลุ่มเท่านั้น เช่น สามารถแบ่งข้อมูลได้จำนวน 8 กลุ่ม ซึ่งเป็นจำนวนกลุ่มคำตอบน้อยที่สุดที่ตัวแบ่งข้อมูลแบบ PSO หาค่าได้จากการทดสอบในครั้งที่ 1 นี้ โดยให้คำตอบเป็น 8 กลุ่มได้ถึง 2 ใน 60 ครั้ง และแบ่งได้จำนวนกลุ่มมากที่สุดที่ 16 กลุ่มจากการทดสอบ 4 ใน 60 ครั้ง โดยที่จะเห็นว่าทุกครั้งในการทดสอบ ตัวจัดกลุ่มแบบ PSO สามารถให้ค่าสมการจุดประสงค์มีค่ามากที่สุดเมื่อเทียบกับการแบ่งข้อมูลอีก 2 รูปแบบ

ตารางที่ 4.2 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 1 จำนวน 60 ครั้ง

kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
8	2	73.50777	66.13387	56.12197
9	3	73.32275	64.8809	56.2091
10	15	73.25384	63.72409	55.48482
11	8	72.82607	62.95448	52.55283
12	6	72.51861	62.0956	52.15315
13	8	72.12055	61.25641	52.05901
14	9	71.96558	55.79597	51.70296
15	5	71.3149	54.68584	50.4667
16	4	70.88357	53.87086	49.07524

อย่างไรก็ตาม จากการทดสอบครั้งที่ 1 ทำให้เห็นลักษณะการทำงานของตัวแบ่งกลุ่มข้อมูลแบบ PSO ที่มีความพยายามปรับปรุงค่า k ใหม่ ซึ่งเกิดจากขั้นตอนการปรับปรุงความเร็วและคัดเลือกประชากรในรุ่นถัดไป เพื่อรักษาค่าสมการจุดประสงค์ $\mathcal{L}_2$ ให้ยังคงมีค่ามากที่สุด ดังจะเห็นได้จากตารางที่ 4.2 ที่แสดงผลลัพธ์ที่เปลี่ยนไป 9 รูปแบบ ตั้งแต่ 8 กลุ่มถึง 16 กลุ่ม ในงานวิจัยนี้ได้กำหนดให้เป็นค่า kpso หรือค่า k ที่ได้จาก PSO-based clustering

4.1.2 การทดสอบครั้งที่ 2 กำหนดค่า k เริ่มต้นที่ 19 25 และ 30 รวมจำนวน 191 ครั้ง

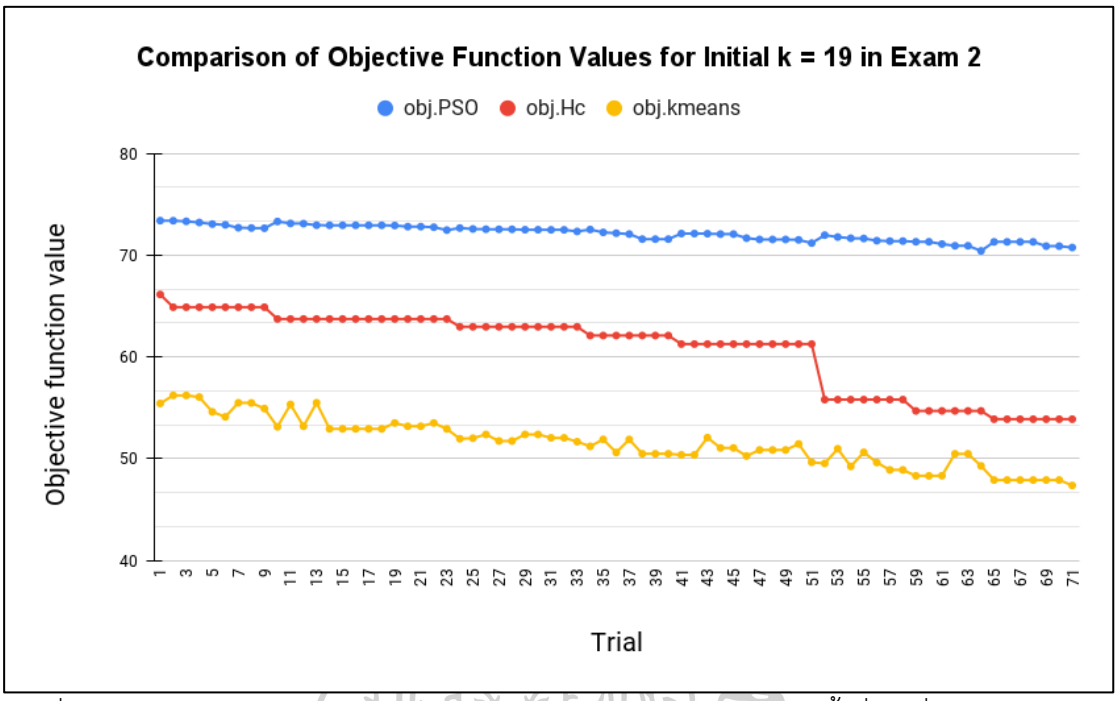
เนื่องจากการทดสอบครั้งที่ 1 ที่กำหนดให้ค่า k เริ่มต้นที่ 19 แต่พบว่าไม่มีครั้งใดในการทดสอบที่ให้ค่า kpso ที่ 19 ซึ่งตรงกับผลที่แนะนำจากเครื่องมือ pathfindR จึงทำการทดสอบครั้งที่

2 โดยทดสอบตัวจัดกลุ่มทุกรูปแบบที่ k เริ่มต้นเท่ากับ 19 ถึง 71 ครั้ง และทดสอบแบบเปลี่ยนค่า k เริ่มต้น เป็นค่าอื่น ได้แก่ 25 และ 30 เพื่อเพิ่มโอกาสให้ตัวจัดกลุ่มแบบ PSO สามารถแบ่งกลุ่มได้ 19 กลุ่ม และวิเคราะห์ผลค่าสมการจุดประสงค์จากกรณีอื่น

ตารางที่ 4.3 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 2 เมื่อกำหนดค่า k เริ่มต้นที่ 19 จำนวน 71 ครั้ง

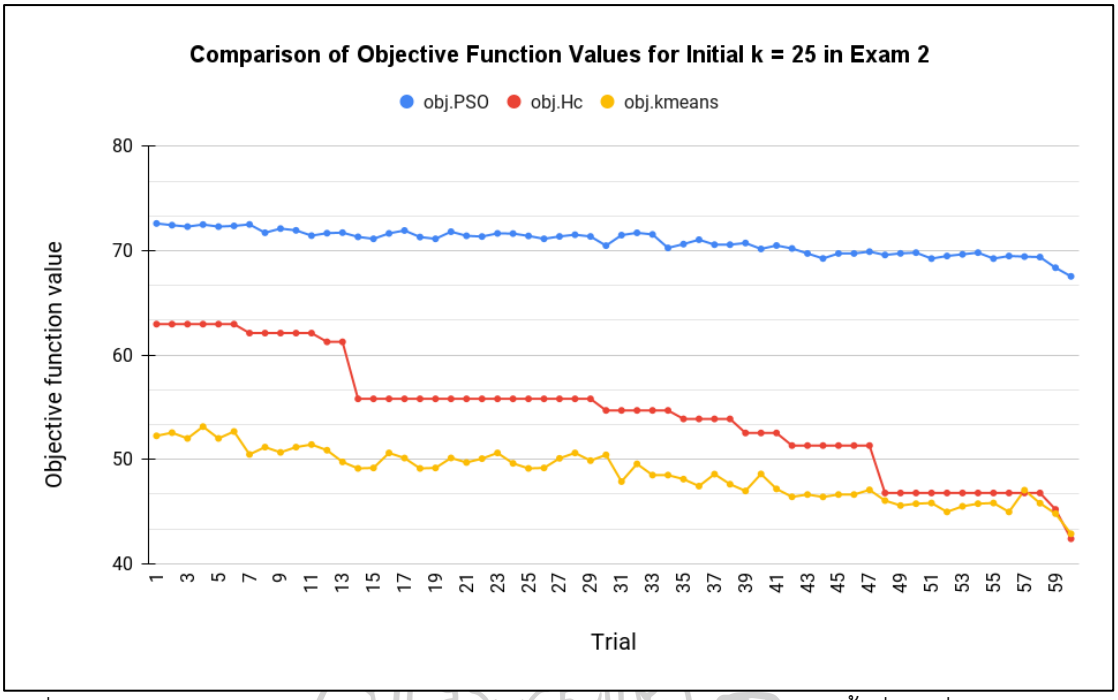
kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
8	1	73.39424	66.13387	55.41802
9	8	73.38474	64.8809	56.2091
10	14	73.3052	63.72409	55.48482
11	10	72.66438	62.95448	52.37042
12	7	72.51861	62.0956	51.88365
13	11	72.12055	61.25641	52.05901
14	7	71.96558	55.79597	50.9608
15	6	71.3149	54.68584	50.4667
16	7	71.30336	53.87086	47.89004

เมื่อทำการทดสอบเพิ่มขึ้นอีก 71 ครั้ง ด้วยค่า k เริ่มต้นที่ 19 ยังคงให้ผลเช่นเดียวกับการทดสอบครั้งแรก กล่าวคือไม่มีครั้งใดที่ให้ค่า kpso ที่ 19 และยังคงให้รูปแบบคำตอบ 9 รูปแบบเช่นเดิม โดยการจัดกลุ่มที่เกิดขึ้นบ่อยครั้งที่สุด ได้แก่ การแบ่งข้อมูลออกเป็น 10 กลุ่ม โดยที่ค่าสมการจุดประสงค์มีค่ามากสุด เมื่อเทียบกับการจัดกลุ่มแบบอื่น ๆ ทุกกรณี ดังแสดงในตารางที่ 4.3 และคำตอบที่ได้จากการจัดกลุ่มแบบ PSO ให้ค่าสมการจุดประสงค์สูงกว่าการจัดกลุ่มรูปแบบอื่นทุกกรณี แสดงดังรูป 4.2



รูปที่ 4.2 ผลเปรียบเทียบค่าสมการจุดประสงค์ $\mathcal{L}_2$ ในการทดสอบครั้งที่ 2 เมื่อกำหนดค่า k เริ่มต้นเท่ากับ 19 จากตัวจัดกลุ่มแบบ PSO แบบลำดับขั้น และแบบเคมีน จำนวน 71 ครั้ง

เมื่อทำการทดสอบตัวจัดกลุ่มทั้ง 3 รูปแบบด้วยการกำหนดค่า k เริ่มต้นเป็น 25 เพิ่มโอกาสในการได้คำตอบเป็นการแบ่งกลุ่มข้อมูลออกเป็น 19 กลุ่ม พบว่าการจัดกลุ่มแบบ PSO ยังเป็นวิธีที่ให้ผลค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากที่สุดทุกกรณี ดังรูป 4.3 พบว่าการจัดกลุ่มแบบ PSO ให้รูปแบบคำตอบ 11 รูปแบบ โดยแบ่งได้จำนวนกลุ่มน้อยสุดที่ 11 กลุ่ม แบ่งได้จำนวนกลุ่มมากที่สุดที่ 24 กลุ่ม และพบการแบ่งกลุ่มข้อมูลเป็น 19 กลุ่ม ทั้งนี้ยังพบอีกว่าตัวจัดกลุ่มแบบ PSO มักให้ผลเป็นการแบ่งข้อมูล 14 กลุ่ม 16 ใน 60 ครั้ง รองลงมาได้แก่การแบ่งข้อมูลออกเป็น 19 กลุ่ม พบมาก 11 ใน 60 ครั้ง โดยที่การแบ่งกลุ่มข้อมูลแบบ PSO ให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากที่สุดในทุกกรณี ดังแสดงในตารางที่ 4.4



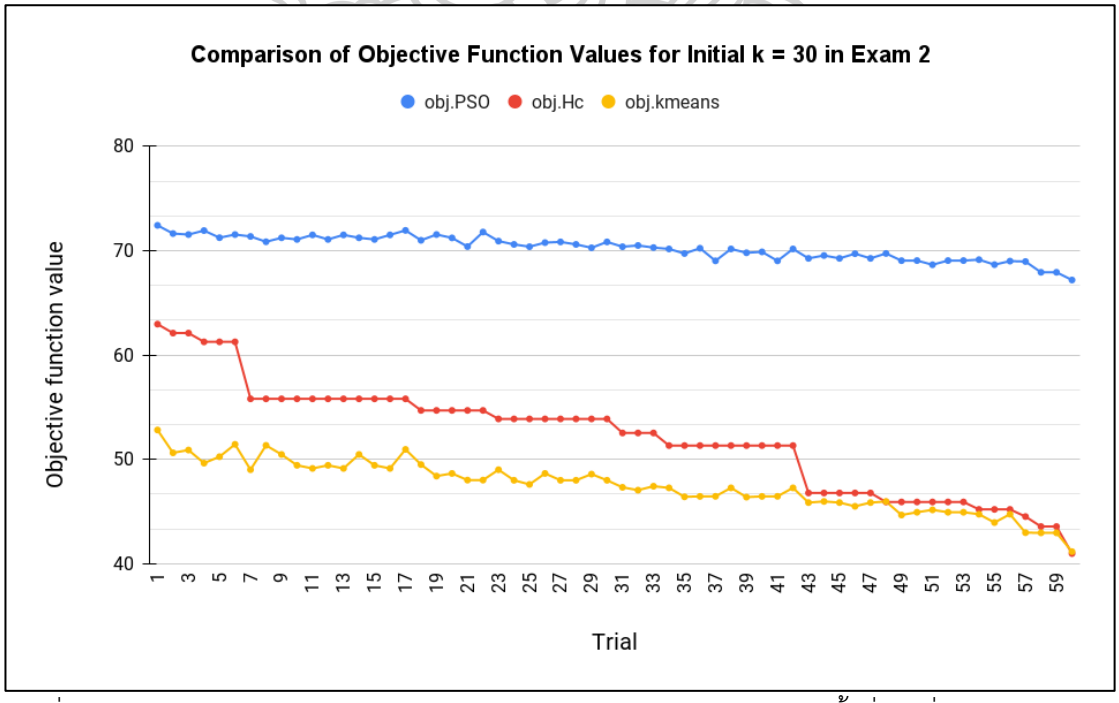
รูปที่ 4.3 ผลเปรียบเทียบค่าสมการจุดประสงค์ $\mathcal{L}_2$ ในการทดสอบครั้งที่ 2 เมื่อกำหนดค่า k เริ่มต้นเท่ากับ 25 จากตัวจัดกลุ่มแบบ PSO แบบลำดับขั้น และแบบเคมีน จำนวน 60 ครั้ง

ตารางที่ 4.4 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 2 เมื่อกำหนดค่า k เริ่มต้นที่ 25 จำนวน 60 ครั้ง

kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
11	6	72.60093	62.95448	53.13405
12	5	72.50808	62.0956	51.41644
13	2	71.72214	61.25641	50.87177
14	16	71.9275	55.79597	50.61235
15	5	71.70517	54.68584	50.42514
16	4	71.04652	53.87086	48.58182
17	3	70.72629	52.52459	48.59838
18	6	70.20597	51.31865	47.0731
19	11	69.80452	46.77494	47.04136
21	1	68.35991	45.20047	44.79598
24	1	67.5311	42.38714	42.86544

จากตารางที่ 4.4 พบว่าการจัดกลุ่มแบบ PSO แบบลำดับขั้น และแบบเคมีนด้วยการแบ่งเป็น 11 กลุ่ม ให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากกว่าค่า kpso อื่น ๆ โดยการแบ่งข้อมูลแบบ PSO ให้ค่ามากที่สุด 72.60093 การจัดกลุ่มที่ให้ค่ามากที่สุดรองลงมาเป็นอันดับสองได้แก่ การจัดกลุ่มแบบลำดับขั้น มีค่า 62.95448 และการจัดกลุ่มแบบเคมีนให้ผลค่าสมการจุดประสงค์มากที่สุดเป็นอันดับสุดท้ายมีค่า 53.13405

เมื่อเพิ่มค่า k เริ่มต้นเท่ากับ 30 ทำการทดสอบจำนวน 60 ครั้ง เปรียบเทียบค่าสมการจุดประสงค์สูงสุดที่ได้จากการจัดกลุ่มทั้ง 3 รูปแบบ ทำให้พบรูปแบบคำตอบมากขึ้นถึง 14 รูปแบบ โดยที่การจัดกลุ่มแบบ PSO มักแบ่งข้อมูลเป็น 17 กลุ่ม พบได้ 10 ใน 60 ครั้ง รองลงมาได้แก่ การแบ่งข้อมูล 15 และ 16 กลุ่ม โดยมีจำนวนการเกิดเหตุการณ์เท่ากัน ได้แก่ 8 ใน 60 ครั้ง และสามารถแบ่งข้อมูลได้จำนวนมากกลุ่มที่สุดที่ 27 กลุ่ม นอกจากนี้ยังพบการแบ่งข้อมูลออกเป็น 19 กลุ่ม 4 ใน 60 ครั้งอีกด้วย ดังแสดงในตารางที่ 4.5 ซึ่งพบว่าผลการทดสอบเป็นลักษณะเดียวกันกับครั้งก่อนหน้าที่กำหนดให้ค่า k เริ่มต้นเท่ากับ 25 จำนวน 60 ครั้ง เมื่อพิจารณาภาพรวมของค่าสมการในการทดสอบย่อยนี้ จะเห็นว่าการจัดกลุ่มข้อมูลแบบ PSO สามารถหาคำตอบที่ทำให้สมการจุดประสงค์ $\mathcal{L}_2$ มีค่ามากที่สุดทุกกรณี ดังรูป 4.4



รูปที่ 4.4 ผลเปรียบเทียบค่าสมการจุดประสงค์ $\mathcal{L}_2$ ในการทดสอบครั้งที่ 2 เมื่อกำหนดค่า k เริ่มต้นเท่ากับ 30 จากตัวจัดกลุ่มแบบ PSO แบบลำดับขั้น และแบบเคมีน จำนวน 60 ครั้ง

ตารางที่ 4.5 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 2 เมื่อกำหนดค่า k เริ่มต้นที่ 30 จำนวน 60 ครั้ง

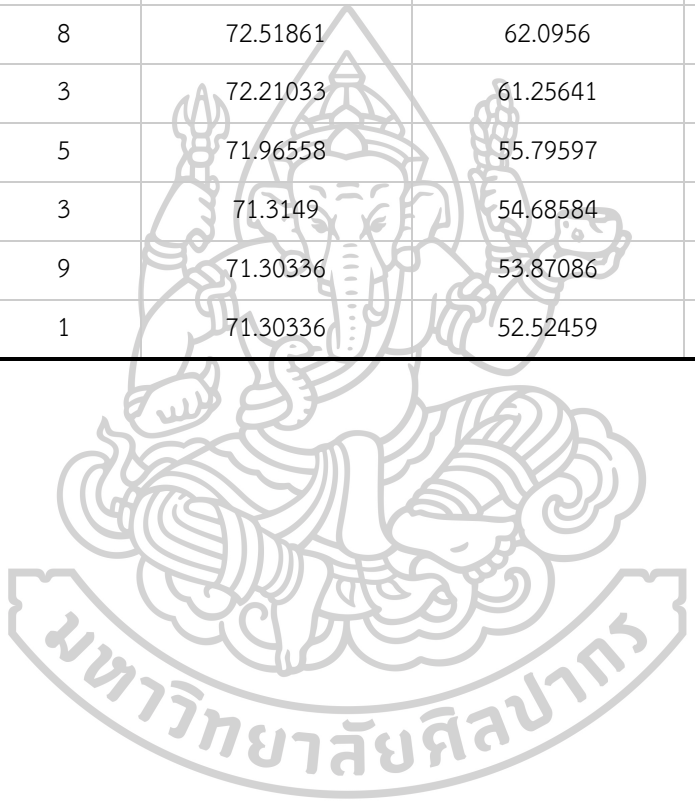
kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
11	1	72.70633	62.95448	52.32722
13	3	71.82205	61.25641	51.67633
14	7	71.67884	55.79597	50.46631
15	8	71.77888	54.68584	50.08658
16	8	71.31517	53.87086	49.95639
17	10	70.59551	52.52459	48.12491
18	5	70.30088	51.31865	47.80783
19	4	70.13932	46.77494	47.04136
20	2	69.5269	45.91368	45.31492
21	5	69.13157	45.20047	45.04573
22	1	68.0169	44.51302	43.26687
23	3	68.03261	43.55819	44.18184
25	2	67.48885	41.70277	42.23809
27	1	67.1652	40.24033	41.49297

4.1.3 การทดสอบครั้งที่ 3 กำหนดค่า k เริ่มต้นที่ 19 25 และ 30 รวมจำนวน 180 ครั้ง

ในขั้นตอนนี้จะทำการทดสอบแบบเดียวกับครั้งที่ 2 อีกครั้ง เพื่อวิเคราะห์ผลค่า kpso ที่ได้จากการจัดกลุ่มข้อมูลแบบ PSO โดยทำเพิ่มอีก 180 ครั้ง ประกอบด้วยการกำหนดค่า k เริ่มต้นที่ 19 25 และ 30 อย่างละ 60 ครั้ง ผลการดำเนินการได้ผลเป็นไปในทิศทางเดียวกันกับการทดสอบครั้งที่ 2 ดังตารางที่ 4.6, 4.7 และ 4.8

ตารางที่ 4.6 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 3 เมื่อกำหนดค่า k เริ่มต้นที่ 19 จำนวน 60 ครั้ง

kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
8	2	73.39424	66.13387	57.47728
9	6	73.38474	64.8809	56.2091
10	19	73.3052	63.72409	55.67058
11	4	72.63997	62.95448	52.90012
12	8	72.51861	62.0956	52.73206
13	3	72.21033	61.25641	52.71079
14	5	71.96558	55.79597	51.70296
15	3	71.3149	54.68584	50.4667
16	9	71.30336	53.87086	48.19218
17	1	71.30336	52.52459	48.16107



ตารางที่ 4.7 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 3 เมื่อกำหนดค่า k เริ่มต้นที่ 25 จำนวน 60 ครั้ง

kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
9	2	72.79848	64.8809	54.48475
10	3	72.967	63.72409	53.76047
11	2	72.68697	62.95448	51.94895
12	4	71.94745	62.0956	52.28148
13	9	71.92667	61.25641	51.93368
14	8	71.847	55.79597	50.61235
15	9	71.42806	54.68584	50.08658
16	8	70.79842	53.87086	49.53169
17	5	70.43683	52.52459	48.59838
18	1	70.22312	51.31865	47.01569
19	6	69.9609	46.77494	48.8432
20	1	68.9015	45.91368	45.42778
21	1	68.92191	45.20047	44.54844
23	1	68.57903	43.55819	43.3715

ตารางที่ 4.8 การเปรียบเทียบค่าสมการจุดประสงค์สูงสุด ระหว่างการจัดกลุ่มแบบ PSO แบบลำดับขั้นและแบบเคมีน ในการทดสอบที่ 3 เมื่อกำหนดค่า k เริ่มต้นที่ 30 จำนวน 60 ครั้ง

kpso	Total event	max obj.PSO	max obj.Hc	max obj.kmeans
11	1	72.41994	62.95448	52.80957
12	2	71.63355	62.0956	50.88983
13	3	71.9192	61.25641	51.4359
14	11	71.93743	55.79597	51.33675
15	5	71.77888	54.68584	49.49771
16	8	70.91021	53.87086	49.00332
17	3	70.4913	52.52459	47.42126
18	9	70.22629	51.31865	47.26506
19	5	69.70394	46.77494	45.96368
20	6	69.72694	45.91368	45.9562
21	3	69.13157	45.20047	44.7396
22	1	68.95234	44.51302	42.97193
23	2	67.92161	43.55819	42.95459
26	1	67.18274	40.96781	41.15398

จากตารางที่ 4.6, 4.7 และ 4.8 ข้างต้น พบว่าการกำหนดค่า k เริ่มต้นที่ 19 เริ่มสามารถให้คำตอบได้ถึง 10 รูปแบบ (แบ่งข้อมูลออกเป็น 8 กลุ่ม ถึง 17 กลุ่ม) ค่าสมการจุดประสงค์จะมีค่ามากสุดเมื่อแบ่งข้อมูลเป็น 8 กลุ่ม มีค่า 73.39424 และตัวจัดกลุ่มแบบ PSO มักแบ่งกลุ่มเป็น 10 กลุ่ม โดยพบว่าเกิดขึ้นมากถึง 19 ใน 60 ครั้ง ซึ่งให้ผลที่สอดคล้องกันในทั้งสามการทดสอบ และไม่พบการแบ่งกลุ่มข้อมูลเป็น 19 กลุ่ม

เมื่อกำหนดค่า k เริ่มต้นที่ 25 ในการทดสองครั้งนี้ พบว่าตัวจัดกลุ่มแบบ PSO สามารถให้คำตอบได้มากขึ้นถึง 14 รูปแบบ เริ่มเห็นการพยายามแบ่งกลุ่มให้อยู่ในช่วง 13 – 16 กลุ่มมากขึ้น โดยค่าสมการจุดประสงค์สูงสุดอยู่ที่การแบ่งข้อมูลเป็น 10 กลุ่ม มีค่า 72.967 ซึ่งมากกว่าการแบ่งกลุ่มด้วยวิธีลำดับขั้นด้วยค่า k ที่เท่ากัน เนื่องจากการแบ่งกลุ่มแบบลำดับขั้นที่ 9 กลุ่ม จะให้ค่าสมการจุดประสงค์ได้มากกว่าแบ่งเป็น 10 กลุ่ม ในขณะที่การแบ่งกลุ่มแบบเคมีนที่ค่า k เดียวกัน พบว่าค่าสมการจุดประสงค์มีค่าน้อยกว่าการแบ่งเป็น 9 กลุ่ม เช่นเดียวกันกับแบบลำดับขั้น

เมื่อกำหนดค่า k เริ่มต้นที่ 30 ยังคงให้รูปแบบคำตอบได้ 14 รูปแบบเช่นเดิม โดยที่การแบ่งกลุ่มแบบ PSO ในรอบนี้ สามารถแบ่งกลุ่มได้จำนวนน้อยที่สุด 11 กลุ่ม ให้ค่าสมการจุดประสงค์มากที่สุดถึง 72.41994 และแบ่งกลุ่มได้จำนวนมากที่สุด 26 กลุ่ม ด้วยค่าสมการจุดประสงค์ 67.18274

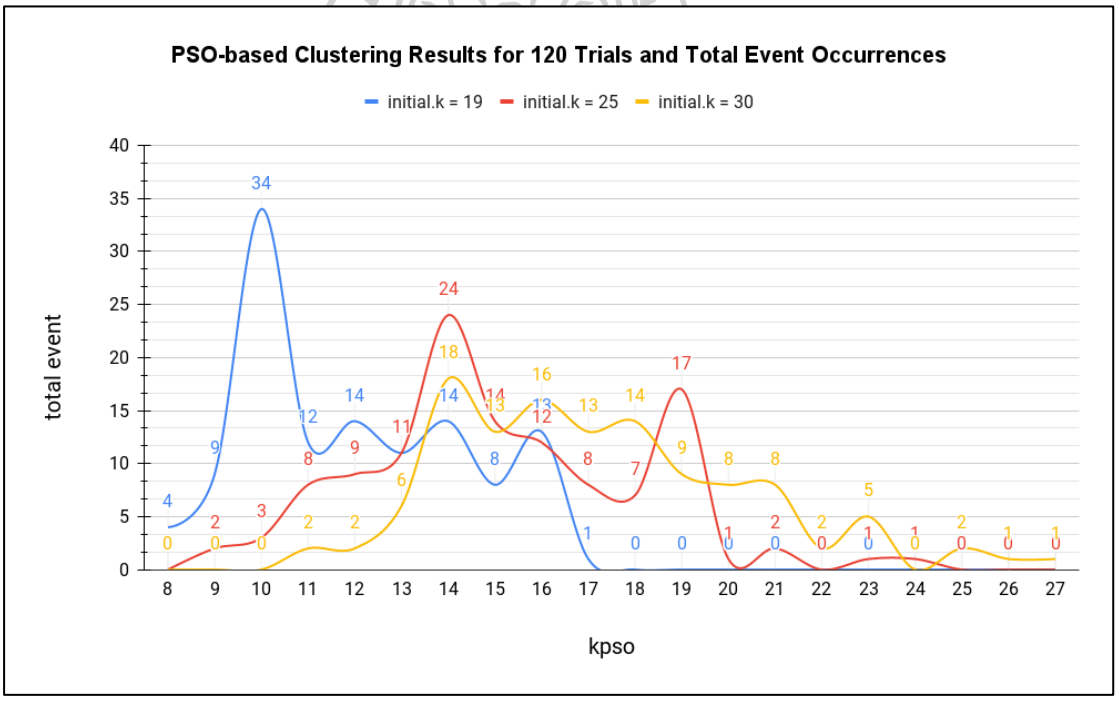
อย่างไรก็ตาม จากการทดสอบทั้งหมดพบว่าตัวจัดกลุ่มข้อมูลแบบ PSO สามารถแบ่งกลุ่มได้ 19 กลุ่ม ซึ่งเป็นไปตามค่า k ที่เหมาะสมจากเครื่องมือ pathfindR 26 ใน 421 ครั้ง และเป็นผลจากการกำหนดค่า k เริ่มต้นที่ 25 และ 30 เท่านั้น โดยเหตุการณ์ที่สามารถแบ่งข้อมูลได้เป็น 19 กลุ่ม และทำให้สมการจุดประสงค์มีค่ามากที่สุดเกิดจากการกำหนดค่า k เริ่มต้นเท่ากับ 30 โดยให้ค่าสูงถึง 70.13932 เทียบกับการแบ่งกลุ่มแบบลำดับขั้นและเคมีน ที่มีค่า 46.77479 และ 45.6725 ดังแสดงในตารางที่ 4.9

ตารางที่ 4.9 ค่าสมการจุดประสงค์สูงสุดที่ตัวจัดกลุ่มทั้ง 3 ประเภท แบ่งกลุ่มข้อมูลได้ 19 กลุ่ม

kpso	initial.k	obj.PSO	obj.Hc	obj.kmeans	kpso	initial.k	obj.PSO	obj.Hc	obj.kmeans
19	25	69.58198	46.77494	46.0456	19	25	69.9609	46.77494	48.8432
19	25	69.73345	46.77494	45.57352	19	25	68.9533	46.77494	46.07238
19	25	69.80452	46.77494	45.74745	19	25	69.12824	46.77494	46.3265
19	25	69.23426	46.77494	45.81268	19	25	69.12824	46.77494	46.3265
19	25	69.48286	46.77494	44.96066	19	30	69.10156	46.77494	45.99897
19	25	69.64236	46.77494	45.49714	19	30	70.13932	46.77494	45.67425
19	25	69.80452	46.77494	45.74745	19	30	69.26449	46.77494	47.04136
19	25	69.23426	46.77494	45.81268	19	30	69.03399	46.77494	47.00011
19	25	69.48286	46.77494	44.96066	19	30	69.25631	46.77494	45.85356
19	25	69.42479	46.77494	47.04136	19	30	69.53164	46.77494	45.96368
19	25	69.38065	46.77494	45.79427	19	30	69.25631	46.77494	45.85356
19	25	69.9609	46.77494	48.8432	19	30	69.70394	46.77494	45.49714
19	25	69.73345	46.77494	45.57352	19	30	69.25631	46.77494	45.85356

4.1.4 วิจารณ์ผลการทดลองเรื่องค่าสมการจุดประสงค์

จากการทดสอบตัวจัดกลุ่มแบบ PSO เทียบกับการจัดกลุ่มแบบลำดับขั้น และแบบเคมีน ด้วยค่า k เริ่มต้นที่ 19 25 และ 30 ทั้งสามการทดสอบที่ได้ศึกษาในงานวิจัยนี้พบว่าตัวจัดกลุ่มแบบ PSO มีลักษณะการทำงานคล้ายการแนะนำจำนวนกลุ่มที่สามารถแบ่งข้อมูลได้ หรือค่า kpso ภายใต้เงื่อนไขการทำให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มีค่ามากที่สุด แสดงผลการทดสอบได้ดังรูป 4.5 จะเห็นว่ามีเหตุการณ์การแบ่งกลุ่มที่มักเกิดขึ้นได้บ่อยครั้ง โดยเหตุการณ์ที่เกิดขึ้นมากที่สุด ได้แก่ การแบ่งกลุ่มข้อมูลเป็น 14 กลุ่ม พบได้ 56 ใน 120 ครั้ง การแบ่งกลุ่มข้อมูลเป็น 16 กลุ่ม พบได้ 41 ใน 120 ครั้ง การแบ่งกลุ่มข้อมูลเป็น 10 กลุ่ม พบได้ถึง 37 ใน 120 ครั้ง การแบ่งกลุ่มข้อมูลเป็น 15 กลุ่ม พบได้ 35 ใน 120 ครั้ง และการแบ่งกลุ่มข้อมูลเป็น 13 กลุ่ม พบได้ 28 ใน 120 ครั้ง ดังแสดงในตารางที่ 4.10



รูปที่ 4.5 ผลการจัดกลุ่มด้วยตัวกลุ่มประเภท PSO โดยกำหนดค่า k เริ่มต้นที่ 19 25 และ 30 จำนวน 120 ครั้ง

ตารางที่ 4.10 จำนวนเหตุการณ์ที่สามารถแบ่งกลุ่มได้จากตัวจัดกลุ่มประเภท PSO จำแนกตามค่า k เริ่มต้น

kpso	Total number of event occurrence			Total event
	initial.k = 19	initial.k = 25	initial.k = 30	
8	4	0	0	4
9	9	2	0	11
10	34	3	0	37
11	12	8	2	22
12	14	9	2	25
13	11	11	6	28
14	14	24	18	56
15	8	14	13	35
16	13	12	16	41
17	1	8	13	22
18	0	7	14	21
19	0	17	9	26
20	0	1	8	9
21	0	2	8	10
22	0	0	2	2
23	0	1	5	6
24	0	1	0	1
25	0	0	2	2
26	0	0	1	1
27	0	0	1	1

4.2 คุณสมบัติความเป็นกลุ่มข้อมูล

ในส่วนนี้จะเป็นการพิจารณาคุณสมบัติของการเป็นกลุ่มเดียวกันของข้อมูล โดยพิจารณาในแง่ความกะทัดรัดของข้อมูล และระยะห่างระหว่างกลุ่มของข้อมูล มีผลการวิจัยดังต่อไปนี้

4.2.1 ความกะทัดรัดของกลุ่มข้อมูล (Compactness)

ผลการศึกษาทางด้านความกะทัดรัดของกลุ่มข้อมูล อาศัยการวัดเส้นผ่านศูนย์กลางของกลุ่มข้อมูลที่สามารถแบ่งได้ หากกลุ่มของข้อมูลมีความเกี่ยวข้องกันมาก จะทำให้กลุ่มข้อมูลนั้นมีความกะทัดรัดสูงหรือมีเส้นผ่านศูนย์กลางเป็นจำนวนน้อย ๆ โดยในงานวิจัยนี้ได้เก็บข้อมูลความยาวเส้นผ่านศูนย์กลางของกลุ่มข้อมูลที่ในแต่ละรอบของการทดสอบ รวม 431 ครั้ง ให้ผลงานวิจัยดังนี้

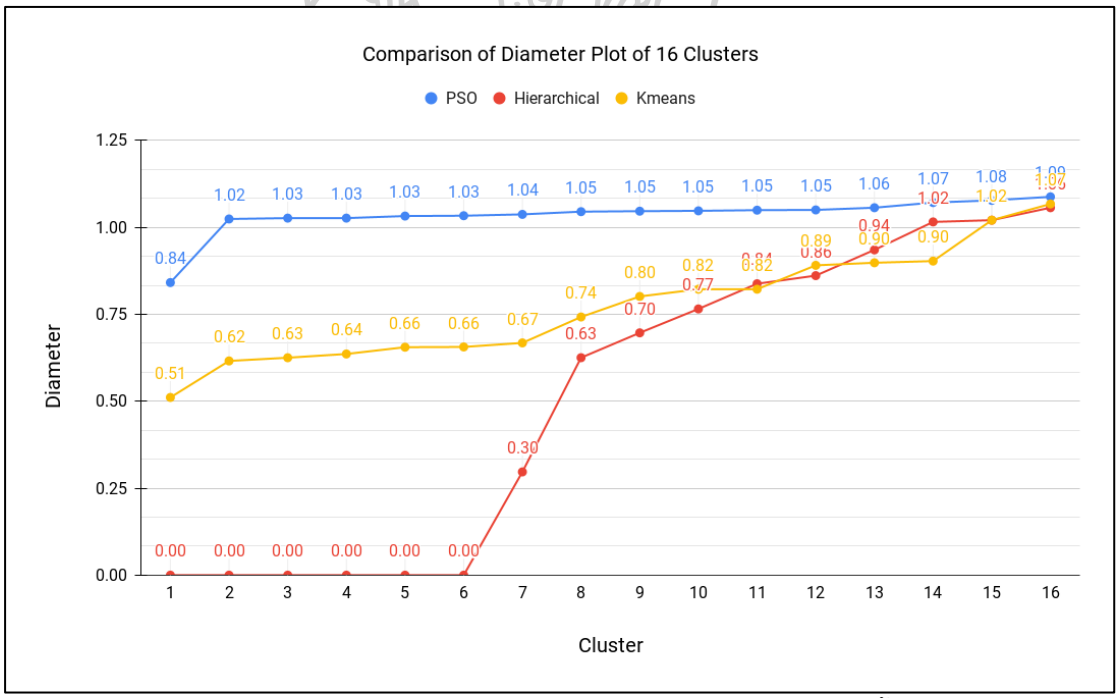
4.2.1.1 ตัวอย่างข้อมูลเส้นผ่านศูนย์กลางที่ได้จากการแบ่งข้อมูลทุกประเภท ได้ 16 กลุ่ม

เมื่อพิจารณาเส้นผ่านศูนย์กลางทุกกลุ่มที่ตัวจัดกลุ่มแบบ PSO สามารถหาค่าได้ เปรียบเทียบกับการจัดกลุ่มรูปแบบอื่น ๆ พบว่าความยาวเส้นผ่านศูนย์กลางเริ่มต้น มักมีค่ามากกว่าการจัดกลุ่มแบบเคมีน และสูงกว่าแบบลำดับขั้นเสมอ กล่าวคือการจัดกลุ่มประเภทที่ให้ความกะทัดรัดของกลุ่มข้อมูลมากที่สุดคือแบบลำดับขั้น แบบเคมีน และแบบ PSO ตามลำดับ ในบทนี้นำเสนอตัวอย่างผลการแบ่งกลุ่มข้อมูลด้วยตัวจัดกลุ่มทั้ง 3 รูปแบบ ดังตารางที่ 4.11 แสดงผลความยาวเส้นผ่านศูนย์กลางของแต่ละกลุ่ม เมื่อกำหนดค่า k เริ่มต้นเท่ากับ 25 ตัวจัดกลุ่มข้อมูลแบบ PSO แนะนำค่า kpso เท่ากับ 16 กลุ่ม ด้วยค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากที่สุดเทียบกับตัวจัดกลุ่มรูปแบบอื่น คือ 70.52598 53.87086 48.92159 เป็นผลที่ได้จากการจัดกลุ่มแบบ PSO แบบลำดับขั้น และแบบเคมีน ตามลำดับ จากตารางนี้แสดงค่าความยาวเส้นผ่านศูนย์กลางของข้อมูลแต่ละกลุ่มเรียงจากน้อยไปมาก พบว่าค่าความยาวเส้นผ่านศูนย์กลางเริ่มต้นน้อยที่สุดของการแบ่งกลุ่มแบบ PSO มีค่า 0.841334 ซึ่งมากกว่าการแบ่งแบบเคมีน 0.5110497 และแบบลำดับขั้นที่มีค่าเป็น 0 กล่าวคือการแบ่งข้อมูลแบบลำดับขั้นมีความกะทัดรัดมากที่สุด เนื่องจากสมาชิกในกลุ่มมีระยะห่างเท่ากับ 0 ดังรูป 4.6 แสดงกราฟเปรียบเทียบความยาวเส้นผ่านศูนย์กลางด้วยการแบ่งกลุ่มจากทั้ง 3 ประเภท จากตัวอย่างดังกล่าว

นอกจากนี้ความยาวเส้นผ่านศูนย์กลางกลุ่มสุดท้าย คือความกว้างของกลุ่มข้อมูลที่มากที่สุด ได้แก่ การแบ่งกลุ่มประเภท PSO ยังคงมีค่าสูงกว่าแบบเคมีนและลำดับขั้น ตามลำดับ ให้ค่าความยาวดังแสดงตารางที่ 4.11 แสดงตัวอย่างเส้นผ่านศูนย์กลางที่ได้จากการแบ่งกลุ่มข้อมูลออกเป็น 16 กลุ่ม ทั้งนี้ การเก็บข้อมูลเพื่อนำมาวิเคราะห์ความกะทัดรัดโดยภาพรวมของทุกรูปแบบการแบ่งกลุ่มข้อมูลและทุกจำนวนกลุ่มที่สามารถแบ่งได้ทั้งหมด 431 ครั้ง จะเก็บเป็นค่าเฉลี่ยของเส้นผ่านศูนย์กลาง

ตารางที่ 4.11 ตัวอย่างเส้นผ่านศูนย์กลางที่ได้จากการแบ่งกลุ่มข้อมูลออกเป็น 16 กลุ่ม

Cluster	PSO	Hierarchical	Kmeans	Cluster	PSO	Hierarchical	Kmeans
1	0.841334	0	0.5110497	9	1.046589	0.6967985	0.8013995
2	1.024157	0	0.6159822	10	1.047565	0.7655172	0.8222222
3	1.026637	0	0.6252347	11	1.04951	0.8379467	0.8222222
4	1.026637	0	0.635996	12	1.050194	0.8612663	0.8908507
5	1.032558	0	0.655492	13	1.056539	0.9352882	0.8982659
6	1.03352	0	0.6562074	14	1.071346	1.015686	0.9031168
7	1.037383	0.2969502	0.66787	15	1.07767	1.02069	1.02069
8	1.045198	0.6253521	0.7423756	16	1.087809	1.056539	1.067308

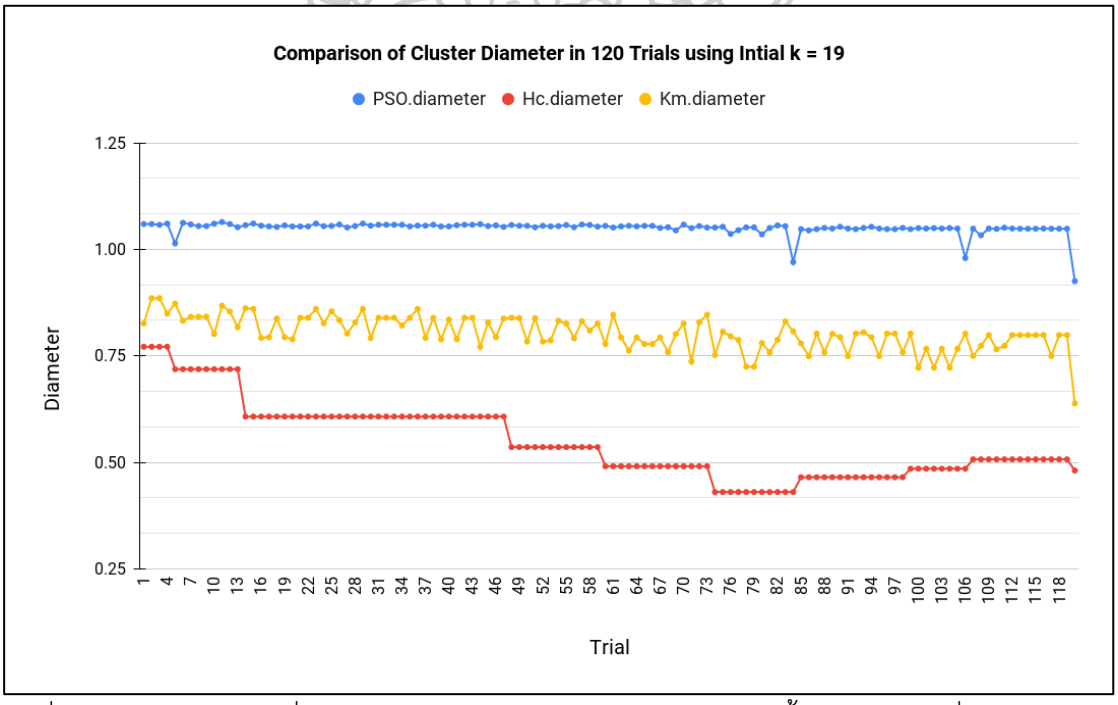


รูปที่ 4.6 การเปรียบเทียบความยาวเส้นผ่านศูนย์กลางจากการแบ่งกลุ่มทั้ง 3 ประเภท จากตัวอย่างการแบ่งกลุ่มข้อมูลออกเป็น 16 กลุ่ม

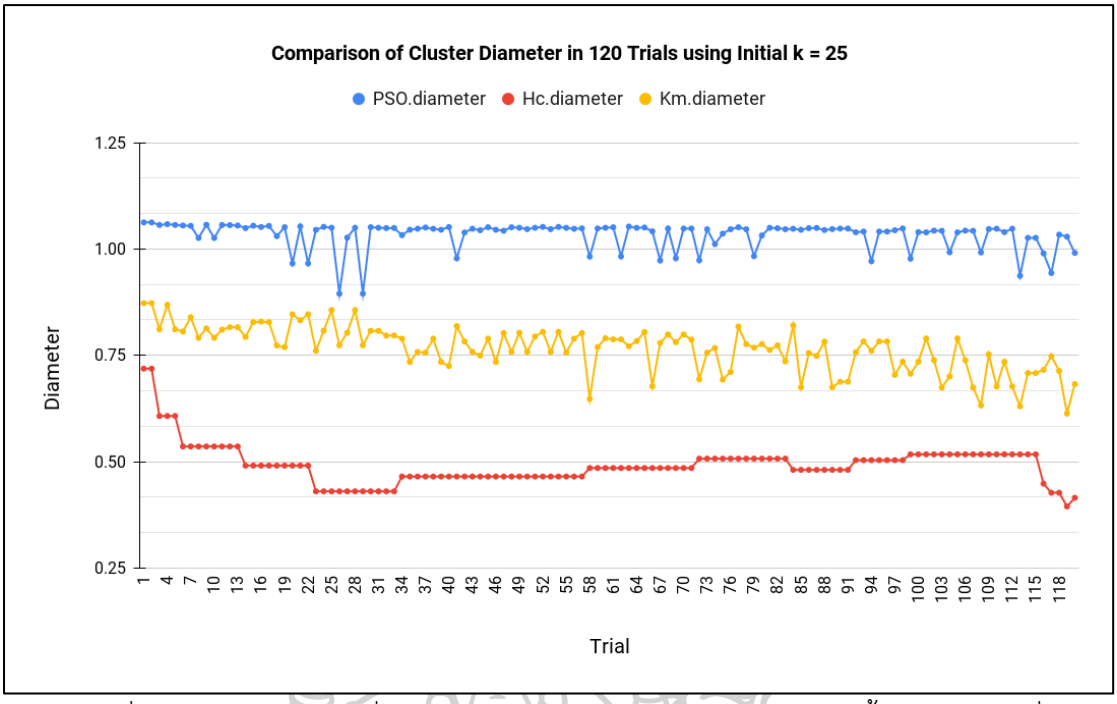
4.2.1.2 ภาพรวมภาพรวมเส้นผ่านศูนย์กลางของกลุ่มข้อมูล

จากการเก็บข้อมูลค่าเฉลี่ยเส้นผ่านศูนย์กลางจากการจัดกลุ่มทั้ง 3 ประเภท แบ่งแยกตามค่า k เริ่มต้นที่ 19 25 และ 30 แสดงภาพรวมของผลวิจัยในรูปแบบกราฟ ดังรูปที่ 4.7, 4.8 และ 4.9 จะ

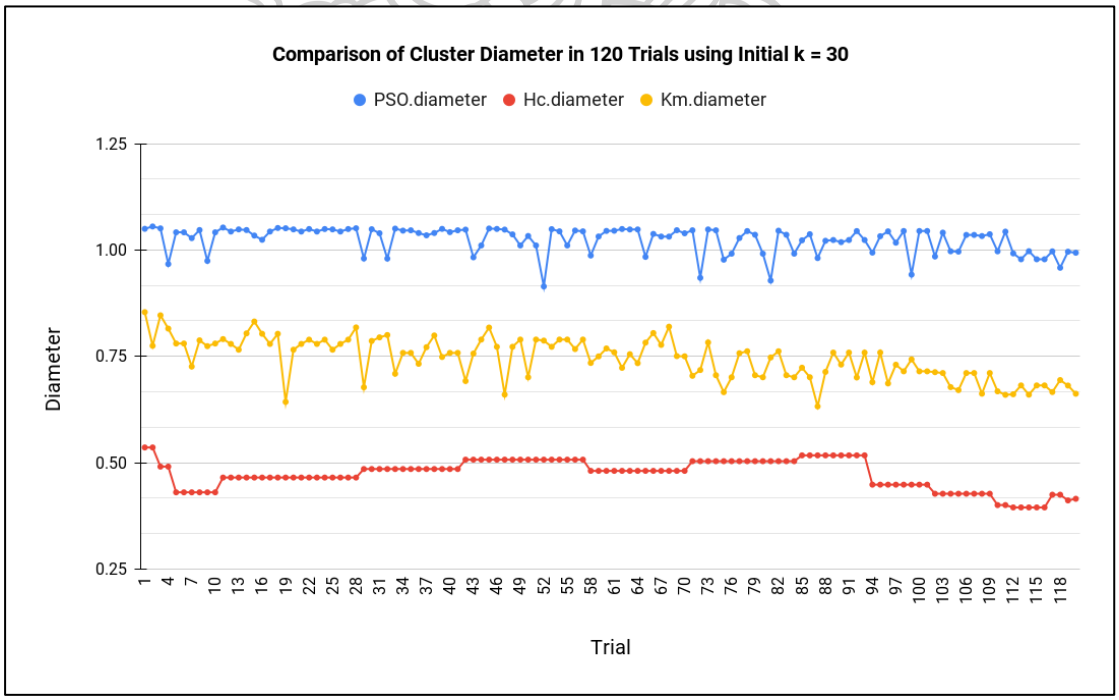
เห็นว่าค่าเฉลี่ยเส้นผ่านศูนย์กลางของกลุ่มข้อมูลมีค่าจากน้อยไปมากตามวิธีการแบ่งกลุ่มข้อมูลได้แก่ วิธีการแบ่งกลุ่มข้อมูลแบบลำดับขั้น แบบเคมีน และแบบ PSO ตามลำดับ และมีภาพรวมของผลวิจัยดังตารางที่ 4.12 ซึ่งแสดงค่าเฉลี่ยเส้นผ่านศูนย์กลางของกลุ่มข้อมูลน้อยที่สุด เรียงตามจำนวนกลุ่มที่สามารถแบ่งได้ หรือ kpso จะสังเกตการแบ่งกลุ่มข้อมูลที่ทำให้กลุ่มข้อมูลในแง่ความกะทัดรัดมากที่สุด ตามประเภทดังนี้ สำหรับการแบ่งกลุ่มแบบ PSO ได้แก่ 0.894991 เมื่อแบ่งข้อมูลออกเป็น 13 กลุ่ม การแบ่งกลุ่มแบบลำดับขั้น มีค่า 0.394637 เมื่อแบ่งข้อมูลเป็น 8 กลุ่ม และการแบ่งกลุ่มแบบเคมีน ให้ขนาดเส้นผ่านศูนย์กลางของกลุ่มข้อมูล 0.613147 เมื่อแบ่งข้อมูลออกเป็น 8 กลุ่ม สำหรับผลความกะทัดรัดของกลุ่มข้อมูลน้อยที่สุด ได้แก่ การแบ่งกลุ่มแบบ PSO มีเส้นผ่านศูนย์กลางมากที่สุดที่ 1.064179 สำหรับการแบ่งข้อมูลออกเป็น 9 กลุ่ม การแบ่งกลุ่มแบบลำดับขั้น 0.771341 และการแบ่งกลุ่มแบบเคมีนให้ค่า 0.885404 โดยทั้งสองรูปแบบได้จากการแบ่งกลุ่มข้อมูลที่ได้ 8 กลุ่ม



รูปที่ 4.7 ผลค่าเฉลี่ยเส้นผ่านศูนย์กลางเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ เมื่อกำหนดค่า k เริ่มต้นที่ 19 จากการทดสอบ 120 ครั้ง



รูปที่ 4.8 ผลค่าเฉลี่ยเส้นผ่านศูนย์กลางเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ เมื่อกำหนดค่า k เริ่มต้นที่ 25 จากการทดสอบ 120 ครั้ง



รูปที่ 4.9 ผลค่าเฉลี่ยเส้นผ่านศูนย์กลางเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ เมื่อกำหนดค่า k เริ่มต้นที่ 30 จากการทดสอบ 120 ครั้ง

ตารางที่ 4.12 ภาพรวมของค่าเฉลี่ยเส้นผ่าศูนย์กลางน้อยที่สุดและมากที่สุด เรียงตามค่า kpso

kpso	Total event	min PSO.diameter	min Hc.diameter	min Km.diameter	max PSO.diameter	max Hc.diameter	max Km.diameter
8	5	1.057601	0.771341	0.826022	1.060378	0.771341	0.885404
9	19	1.014018	0.718631	0.801165	1.064179	0.718631	0.872717
10	51	1.051497	0.607344	0.771285	1.060649	0.607344	0.868791
11	32	0.961551	0.535677	0.774737	1.058307	0.535677	0.853877
12	32	0.966153	0.49062	0.736638	1.058109	0.49062	0.846486
13	39	0.894991	0.430142	0.724431	1.056226	0.430142	0.856408
14	63	0.978094	0.464898	0.642833	1.052988	0.464898	0.831842
15	41	0.972999	0.484939	0.647556	1.052817	0.484939	0.804794
16	48	0.914327	0.507002	0.659983	1.051169	0.507002	0.817671
17	22	0.925454	0.48065	0.63806	1.049573	0.48065	0.820426
18	21	0.928159	0.503399	0.665711	1.048408	0.503399	0.782645
19	26	0.937053	0.517095	0.630286	1.047786	0.517095	0.789609
20	9	0.942195	0.448176	0.686241	1.044706	0.448176	0.758934
21	10	0.943528	0.426835	0.662151	1.040868	0.426835	0.74779
22	2	0.996911	0.400147	0.659465	1.043202	0.400147	0.667664
23	6	0.978257	0.394647	0.613147	1.028984	0.394647	0.681552
24	1	0.990838	0.4149	0.682461	0.990838	0.4149	0.682461
25	2	0.958264	0.424529	0.666046	0.996821	0.424529	0.694181
26	1	0.99604	0.411116	0.681261	0.99604	0.411116	0.681261
27	1	0.9936	0.41495	0.662028	0.9936	0.41495	0.662028

4.2.2 ความสามารถแยกกันระหว่างกลุ่มข้อมูล (Separation)

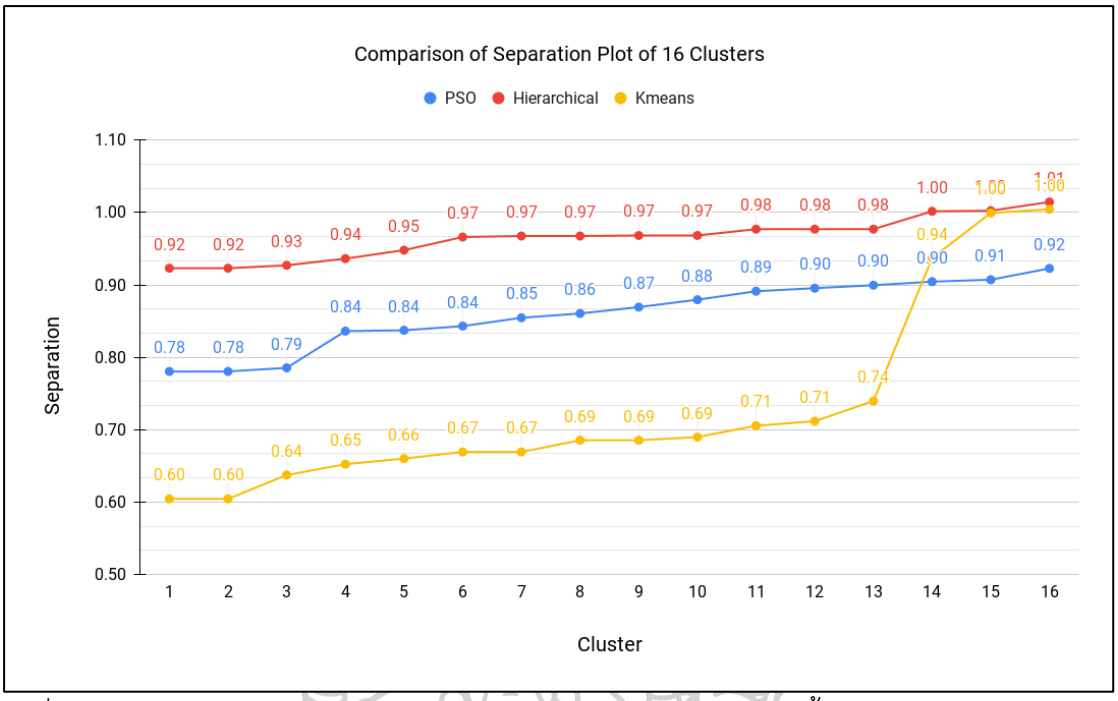
การวิเคราะห์ระยะห่างระหว่างกลุ่มข้อมูลจากการเก็บข้อมูลจากการทดสอบแบ่งกลุ่มข้อมูลด้วยตัวจัดกลุ่มแบบ PSO ทั้งหมด 431 ครั้ง มีการดำเนินการวิจัยเช่นเดียวกับการเก็บค่าความยาวของเส้นผ่านศูนย์กลางของกลุ่มข้อมูล ในการทดสอบนี้ใช้วิธีการหาความยาวระหว่างกลุ่มที่สั้นที่สุดในแต่ละกลุ่มของคำตอบ นำมาวิเคราะห์ค่าเฉลี่ยเป็นผลงานวิจัยในแง่ของการแยกกันระหว่างกลุ่ม หรือ Separation โดยกลุ่มที่สามารถแยกออกจากกันได้มากที่สุดแสดงถึงความเป็นคนละกลุ่มกัน หรือความไม่เป็นกลุ่มเดียวกันของข้อมูล มีรายละเอียด ดังนี้

4.2.2.1 ตัวอย่างระยะห่างระหว่างกลุ่มข้อมูลที่แบ่งได้ 16 กลุ่ม

นำเสนอตัวอย่างระยะห่างระหว่างกลุ่ม โดยเป็นค่าระยะห่างที่น้อยที่สุดระหว่างกลุ่มหนึ่งถึงทุกกลุ่ม ทั้งหมด 16 กลุ่ม ที่สามารถแบ่งได้โดยตัวจัดกลุ่มประเภท PSO ดังตารางที่ 4.13 แสดงให้เห็นว่า การแบ่งกลุ่มแบบ PSO ให้ระยะห่างระหว่างกลุ่มข้อมูล 0.7804269 ซึ่งน้อยกว่าแบบลำดับขั้นมีค่า 0.923125 แต่ให้ความห่างมากกว่าเมื่อเทียบกับการแบ่งข้อมูลประเภทเคมีน มีค่า 0.6043149 ระยะห่างที่มากที่สุดของการแบ่งกลุ่มแต่ละประเภท ให้ผลเช่นเดียวกัน ได้แก่ ประเภท PSO ให้ค่า 0.9228253 แบบลำดับขั้นมีค่า 1.014625 และแบบเคมีนมีค่า 1.004565แสดงกราฟเปรียบเทียบค่าเฉลี่ยระยะห่างระหว่างกลุ่มซึ่งได้จากการจัดกลุ่มประเภทต่าง ๆ ดังรูปที่ 4.10

ตารางที่ 4.13 การเปรียบเทียบระยะห่างระหว่างกลุ่มน้อยที่สุดจากการแบ่งกลุ่มข้อมูล 16 กลุ่ม

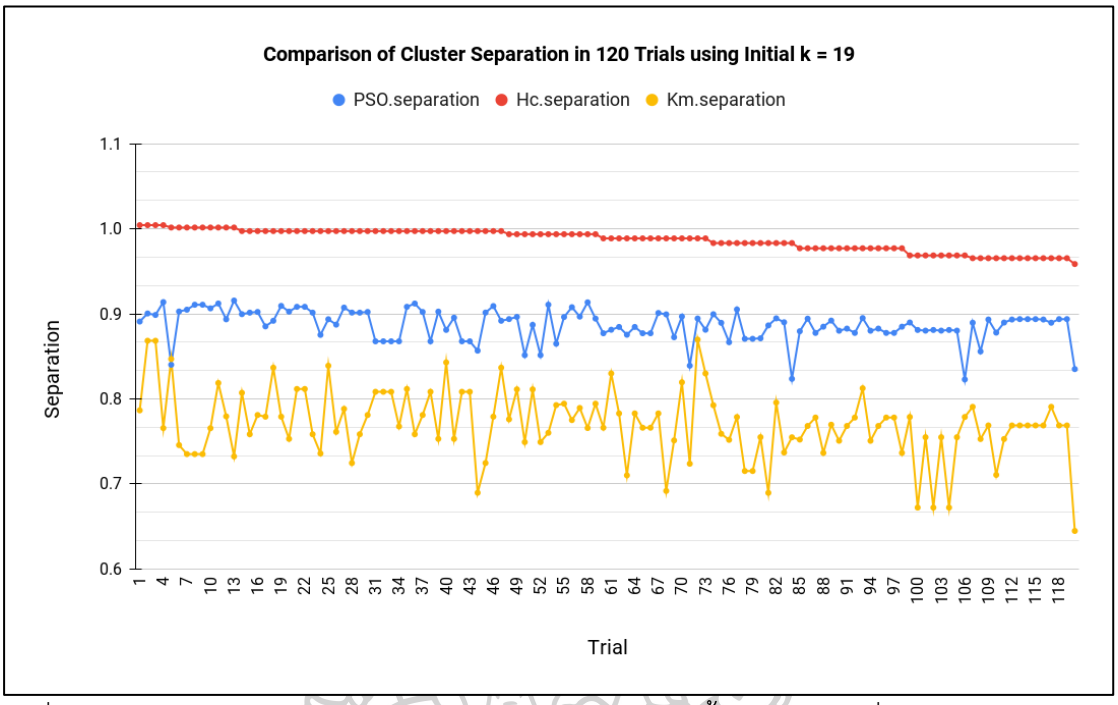
Cluster	PSO	Hierarchical	Kmeans	Cluster	PSO	Hierarchical	Kmeans
1	0.7804269	0.923125	0.6043149	9	0.8693575	0.9683527	0.6853423
2	0.7804269	0.923125	0.6043149	10	0.8795777	0.9683527	0.6897167
3	0.7854202	0.927077	0.6372376	11	0.8912871	0.9771127	0.7053663
4	0.8361674	0.9363713	0.6522542	12	0.8955069	0.9771145	0.711639
5	0.8373069	0.9480586	0.6599642	13	0.8996439	0.9771145	0.7395312
6	0.8430876	0.9662461	0.6691751	14	0.9045557	1.001753	0.9370545
7	0.8545954	0.9676046	0.6691751	15	0.9072074	1.002589	0.9994706
8	0.8605475	0.9676046	0.6853423	16	0.9228253	1.014625	1.004565



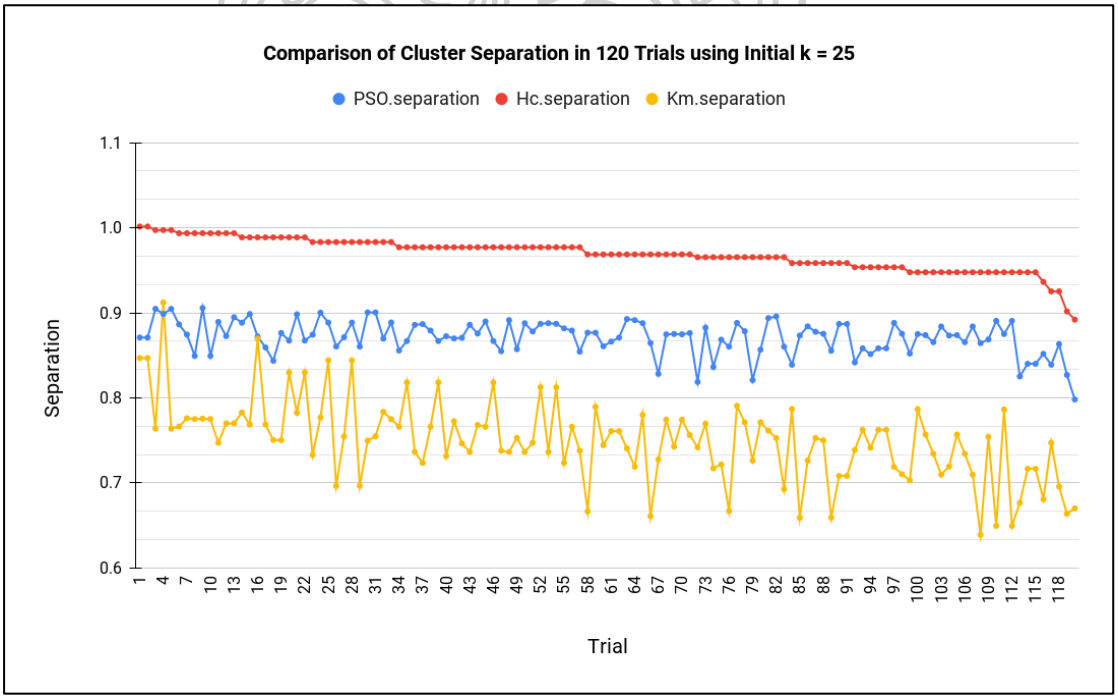
รูปที่ 4.10 การเปรียบเทียบระยะห่างระหว่างกลุ่มจากการแบ่งกลุ่มทั้ง 3 ประเภท จากตัวอย่างการแบ่งกลุ่มข้อมูลออกเป็น 16 กลุ่ม

4.2.2.2 ภาพรวมระยะห่างระหว่างกลุ่มข้อมูล

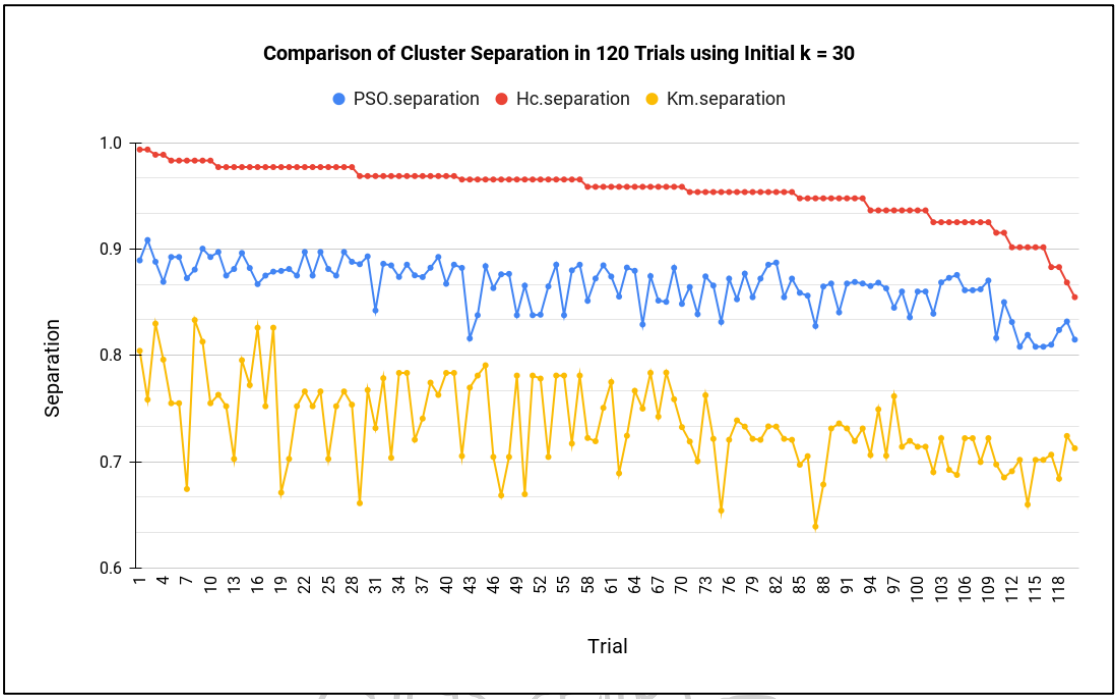
การแบ่งกลุ่มแบบ PSO ให้ระยะห่างระหว่างข้อมูลน้อยกว่าประเภทลำดับขั้นทุกกรณี และเมื่อเปรียบเทียบกับประเภทเคมีน พบว่ามีบางการทดสอบที่ให้ระยะห่างน้อยกว่าประเภทเคมีนประมาณ 2 – 3 ครั้ง จากการทดสอบ 431 ครั้ง ดังรูปที่ 4.11, 4.12 และ 4.13 แสดงผลการเปรียบเทียบค่าเฉลี่ยระยะห่างระหว่างกลุ่มแบ่งแยกตามค่า k เริ่มต้นที่ 19 25 และ 30 และแสดงผลโดยภาพรวมค่าเฉลี่ยระยะห่างระหว่างกลุ่มน้อยที่สุดจากการแบ่งกลุ่มประเภท PSO มีค่า 0.798048 ประเภทลำดับขั้นมีค่า 0.854718 และประเภทเคมีนมีค่า 0.63893 เมื่อแบ่งกลุ่มข้อมูลได้ 24 27 และ 19 กลุ่มตามลำดับ ทางด้านค่าเฉลี่ยระยะห่างระหว่างกลุ่มมากที่สุด จากการเก็บข้อมูลค่าเฉลี่ยพบว่าการแบ่งกลุ่มแบบ PSO ให้ค่ามากสุดที่ 0.916864 ประเภทลำดับขั้นให้ค่า 1.004253 และประเภทเคมีน 0.912122 เมื่อแบ่งกลุ่มข้อมูลออกเป็น 9 8 และ 10 กลุ่ม ตามลำดับ ดังแสดงในตารางที่ 4.14



รูปที่ 4.11 ระยะห่างระหว่างกลุ่มเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ เมื่อกำหนดค่า k เริ่มต้นที่ 19 จากการทดสอบ 120 ครั้ง



รูปที่ 4.12 ระยะห่างระหว่างกลุ่มเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ เมื่อกำหนดค่า k เริ่มต้นที่ 25 จากการทดสอบ 120 ครั้ง



รูปที่ 4.13 ระยะห่างระหว่างกลุ่มเปรียบเทียบการจัดกลุ่มทั้ง 3 รูปแบบ เมื่อกำหนดค่า k เริ่มต้นที่ 30 จากการทดสอบ 120 ครั้ง

ตารางที่ 4.14 ภาพรวมของระยะห่างระหว่างกลุ่มน้อยที่สุดและมากที่สุด เรียงตามค่า kpso

kpso	Total event	min PSO.separation	min Hc.separation	min Km.separation	max PSO.separation	max Hc.separation	max Km.separation
8	5	0.890905	1.004253	0.765611	0.913742	1.004253	0.868317
9	19	0.840095	1.00146	0.732273	0.916864	1.00146	0.889789
10	51	0.856646	0.997287	0.689585	0.912073	0.997287	0.912122
11	32	0.832765	0.993592	0.747141	0.913466	0.993592	0.848004
12	32	0.839028	0.988701	0.691764	0.900969	0.988701	0.869631
13	39	0.82352	0.983185	0.674222	0.90511	0.983185	0.844063
14	63	0.854333	0.977064	0.670843	0.897223	0.977064	0.825996
15	41	0.822811	0.968636	0.660865	0.893074	0.968636	0.789434
16	48	0.815899	0.965389	0.667023	0.895692	0.965389	0.790621
17	22	0.829026	0.958516	0.644676	0.886881	0.958516	0.786953
18	21	0.831399	0.953597	0.653956	0.888158	0.953597	0.76243
19	26	0.825196	0.947695	0.63893	0.890547	0.947695	0.786429
20	9	0.835714	0.936361	0.680722	0.868297	0.936361	0.76153

ตารางที่ 4.14 ภาพรวมของระยะห่างระหว่างกลุ่มน้อยที่สุดและมากที่สุด เรียงตามค่า kpso (ต่อ)

kpso	Total event	min PSO.separation	min Hc.separation	min Km.separation	max PSO.separation	max Hc.separation	max Km.separation
21	10	0.838944	0.92521	0.687539	0.87559	0.92521	0.7472
22	2	0.816294	0.915291	0.685106	0.849988	0.915291	0.697307
23	6	0.808085	0.901651	0.659612	0.831355	0.901651	0.701642
24	1	0.798048	0.892005	0.670142	0.798048	0.892005	0.670142
25	2	0.810006	0.882973	0.683921	0.823814	0.882973	0.706511
26	1	0.831832	0.868511	0.724128	0.831832	0.868511	0.724128
27	1	0.814778	0.854718	0.712548	0.814778	0.854718	0.712548

4.2.3 วิจารณ์ผลการทดลองเรื่องคุณสมบัติความเป็นกลุ่มข้อมูล

จากการวิเคราะห์ข้อมูลค่าเฉลี่ยเส้นผ่านศูนย์กลางแยกตามประเภทของตัวจัดกลุ่ม และแยกตามจำนวนกลุ่มที่แบ่งได้ พบว่าการจัดกลุ่มประเภทที่ทำให้กลุ่มข้อมูลมีความกะทัดรัดเรียงจากน้อยไปมาก ได้แก่ แบบ PSO แบบเคมีน และแบบลำดับขั้น ดังรูปที่ 4.7, 4.8 และ 4.9 ซึ่งแสดงภาพรวมผ่านการทดสอบจำนวน 120 ครั้ง โดยเมื่อศึกษาลงไปถึงผลของกลุ่มยีนส์ที่แบ่งได้พบว่าการแบ่งแบบลำดับขั้นที่สามารถให้ค่าเฉลี่ยเส้นผ่านศูนย์กลางน้อยกว่าแบบอื่น ๆ เกิดจากการมีสมาชิก 1 ตัวต่อ 1 กลุ่ม มากถึงจำนวน 6 กลุ่ม เช่น กลุ่มที่ 2 ได้แก่กลุ่มยีนส์ hsa04130 กลุ่มที่ 5 ได้แก่กลุ่มยีนส์ hsa03040 กลุ่มที่ 12 ได้แก่ hsa03050 เป็นต้น ดังรูปที่ 4.14 เป็นผลทำให้เส้นผ่านศูนย์กลางของกลุ่มนั้น ๆ มีค่าเท่ากับ 0 นอกจากนี้ การทดสอบทั้งหมด 431 ครั้ง มีเส้นผ่านศูนย์กลางแต่ละกลุ่มที่ค่อนข้างเกาะกลุ่มกัน ดังจะเห็นจากกราฟตัวอย่างในรูปที่ 4.10 โดยที่การแบ่งข้อมูลแบบ PSO มักให้ผลเช่นนี้มากกว่าประเภทอื่น ๆ ซึ่งหากคำนวณส่วนเบี่ยงเบนมาตรฐานจากตัวอย่างการแบ่งข้อมูลออกเป็น 16 กลุ่ม พบว่า การแบ่งข้อมูลประเภท PSO ให้ค่าส่วนเบี่ยงเบนมาตรฐานที่ 0.0128 การแบ่งข้อมูลประเภทเคมีน ให้ค่า 0.1735 และแบบลำดับขั้นให้ค่า 0.4595 กล่าวคือการแบ่งแบบลำดับขั้นให้ผลลัพธ์เป็นขนาดของกลุ่มข้อมูลแต่ละกลุ่มแตกต่างกันอย่างเห็นได้ชัดมากกว่าแบบ PSO และเมื่อพิจารณาคุณสมบัติความเป็นกลุ่มของข้อมูลในแง่ระยะห่างระหว่างกลุ่ม การจัดกลุ่มประเภท PSO มีความสามารถเป็นรองต่อประเภทลำดับขั้น แต่ยังคงมีคุณภาพสูงกว่าการจัดกลุ่มประเภทเคมีน

Name	Type	Value
hc	list [16]	List of length 16
[[1]]	integer [6]	1 8 9 10 11 14
[[2]]	integer [1]	2
hsa04130	integer [1]	2
[[3]]	integer [6]	3 13 53 54 65 109
[[4]]	integer [53]	4 12 15 16 17 18 ...
[[5]]	integer [1]	5
hsa03040	integer [1]	5
[[6]]	integer [24]	6 20 21 24 26 27 ...
[[7]]	integer [2]	7 73
[[8]]	integer [3]	23 107 111
[[9]]	integer [3]	33 79 85
[[10]]	integer [2]	41 66
[[11]]	integer [6]	42 48 56 63 86 89
[[12]]	integer [1]	52
hsa03050	integer [1]	52
[[13]]	integer [1]	64
hsa05202	integer [1]	64
[[14]]	integer [2]	96 103
[[15]]	integer [1]	101
hsa00340	integer [1]	101
[[16]]	integer [1]	108
hsa04350	integer [1]	108

Name	Type	Value
me	list [16]	List of length 16
[[1]]	integer [2]	70 102
[[2]]	integer [2]	40 72
[[3]]	integer [2]	22 65
[[4]]	integer [7]	1 17 61 88 89 96 ...
[[5]]	integer [6]	21 42 54 95 101 110
[[6]]	integer [10]	4 7 9 15 19 37 ...
[[7]]	integer [9]	14 24 41 55 63 67 ...
[[8]]	integer [9]	30 50 51 53 64 73 ...
[[9]]	integer [11]	11 12 18 33 49 56 ...
[[10]]	integer [15]	3 16 25 28 29 32 ...
[[11]]	integer [7]	8 43 59 79 91 97 ...
[[12]]	integer [8]	6 10 23 31 39 84 ...
[[13]]	integer [9]	5 13 26 44 46 68 ...
[[14]]	integer [5]	20 27 34 81 93
[[15]]	integer [5]	38 45 66 76 112
[[16]]	integer [6]	2 35 57 69 85 86

รูปที่ 4.14 ตัวอย่างผลการจัดกลุ่มยีนส์ Rheumatoid arthritis (GSE15573) เปรียบเทียบการจัดกลุ่มแบบลำดับขั้น 16 กลุ่ม (ซ้าย) และการจัดกลุ่มแบบ PSO (ขวา) 16 กลุ่ม

4.3 ความหลากหลายของคำตอบ

จากการทดสอบตัวจัดกลุ่มข้อมูลทั้งหมด 431 ครั้ง พบว่าวิธีการจัดกลุ่มแบบ PSO สามารถแบ่งกลุ่มข้อมูลได้หลายรูปแบบ ตั้งแต่ 8 – 27 กลุ่ม ซึ่งนอกจากค่าสมการจุดประสงค์ $\mathcal{L}_2$ จะมีค่ามากกว่าการจัดกลุ่มรูปแบบอื่น ๆ แล้ว ยังพบว่าแต่ละการทดสอบที่ให้ค่า kpso เท่ากัน กลับมีค่าสมการจุดประสงค์ $\mathcal{L}_2$ ที่ไม่เท่ากันเมื่อจัดกลุ่มด้วยประเภท PSO ในขณะที่การแบ่งกลุ่มแบบลำดับขั้นที่มีค่า kpso เป็นค่าเดียวกัน ให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ ค่าเดียวกันทุกเหตุการณ์ อีกทั้งยังมีค่าเฉลี่ยเส้นผ่านศูนย์กลางและระยะห่างระหว่างกลุ่มเป็นค่าเดียวกันทุกเหตุการณ์เช่นกัน ดังตัวอย่างหนึ่งของการทดสอบทั้งหมดในตารางที่ 4.15 แสดงผลจากการทดสอบแบ่งกลุ่มเป็น 20 กลุ่ม เปรียบเทียบจากตัวแบ่งกลุ่มข้อมูลทั้ง 3 ประเภท จะเห็นว่าการแบ่งกลุ่มได้ 20 กลุ่มนี้ เกิดขึ้นบ่อยถึง 9 ใน 431 ครั้ง และได้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ ค่าเฉลี่ยเส้นผ่านศูนย์กลางและค่าเฉลี่ยระยะห่างระหว่างกลุ่มที่แตกต่างกัน 7 ค่า ทั้งในการจัดกลุ่มแบบ PSO และเคมีน ในขณะที่การแบ่งกลุ่มประเภทลำดับขั้นให้ผลเพียง 1 ค่า ซึ่งเป็นค่าเดียวกันในทุกผลวิจัย แสดงให้เห็นถึงการจัดกลุ่มข้อมูลด้วยตัวจัดกลุ่ม PSO ให้รูปแบบคำตอบที่หลากหลายมากกว่าประเภทใด ๆ โดยที่ยังคงทำให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มีค่าสูงสุดอยู่เสมอ จึงสรุปจำนวนคำตอบที่แตกต่างกันในแต่ละการวิเคราะห์ ทั้งทางด้านค่าสมการจุดประสงค์ $\mathcal{L}_2$ ค่าเฉลี่ยเส้นผ่านศูนย์กลาง และค่าเฉลี่ยระยะห่างระหว่างกลุ่ม โดยแสดงผลเรียงตามค่า kpso ได้ดังตารางที่ 4.16, 4.17 และ 4.18

ตารางที่ 4.15 ผลค่าสมการจุดประสงค์ $\mathcal{L}_2$ ค่าเฉลี่ยเส้นผ่านศูนย์กลางและค่าเฉลี่ยระยะห่างระหว่างกลุ่มของการทดสอบการจัดกลุ่มข้อมูลเป็น 20 กลุ่มด้วยตัวแบ่งกลุ่มทั้ง 3 ประเภท

kpso	obj.PSO	obj.Hc	obj. kmeans	PSO. diameter	Hc. diameter	Km. diameter	PSO. separation	Hc. separation	Km. separation
20	69.022	45.914	45.000	0.994	0.448	0.689	0.865	0.936	0.706
20	69.527	45.914	45.315	1.032	0.448	0.759	0.868	0.936	0.749
20	68.902	45.914	45.428	0.990	0.448	0.716	0.852	0.936	0.681
20	69.727	45.914	45.956	1.044	0.448	0.686	0.863	0.936	0.706
20	69.036	45.914	44.663	1.017	0.448	0.730	0.845	0.936	0.762
20	69.045	45.914	44.928	1.045	0.448	0.715	0.860	0.936	0.714
20	68.645	45.914	45.156	0.942	0.448	0.743	0.836	0.936	0.720
20	69.045	45.914	44.928	1.045	0.448	0.715	0.860	0.936	0.714
20	69.045	45.914	44.928	1.045	0.448	0.715	0.860	0.936	0.714

ตารางที่ 4.16 จำนวนคำตอบที่แตกต่างกันจากการแบ่งกลุ่มทั้ง 3 ประเภท แบ่งตามค่าสมการจุดประสงค์ $\mathcal{L}_2$

kpso	PSO	Hierarchical	Kmeans	kpso	PSO	Hierarchical	Kmeans
8	3	1	3	18	16	1	11
9	12	1	10	19	18	1	16
10	22	1	18	20	7	1	7
11	24	1	16	21	9	1	7
12	22	1	16	22	2	1	2
13	26	1	21	23	4	1	4
14	38	1	24	24	1	1	1
15	30	1	23	25	2	1	2
16	29	1	20	26	1	1	1
17	21	1	19	27	1	1	1

ตารางที่ 4.17 จำนวนคำตอบที่แตกต่างกันจากการแบ่งกลุ่มทั้ง 3 ประเภท แบ่งตามค่าเฉลี่ยเส้นผ่านศูนย์กลาง

kpso	PSO	Hierarchical	Kmeans	kpso	PSO	Hierarchical	Kmeans
8	4	1	3	18	16	1	11
9	12	1	10	19	18	1	16
10	22	1	17	20	7	1	7
11	24	1	16	21	9	1	7
12	22	1	16	22	2	1	2
13	26	1	21	23	4	1	4
14	38	1	24	24	1	1	1
15	30	1	23	25	2	1	2
16	29	1	20	26	1	1	1
17	21	1	19	27	1	1	1

ตารางที่ 4.18 จำนวนคำตอบที่แตกต่างกันจากการแบ่งกลุ่มทั้ง 3 ประเภท แบ่งตามค่าเฉลี่ยระยะห่างระหว่างกลุ่ม

kpso	PSO	Hierarchical	Kmeans	kpso	PSO	Hierarchical	Kmeans
8	4	1	3	18	16	1	11
9	12	1	10	19	18	1	16
10	22	1	18	20	7	1	7
11	24	1	16	21	9	1	7
12	22	1	16	22	2	1	2
13	25	1	21	23	4	1	4
14	38	1	24	24	1	1	1
15	30	1	23	25	2	1	2
16	29	1	20	26	1	1	1
17	21	1	19	27	1	1	1

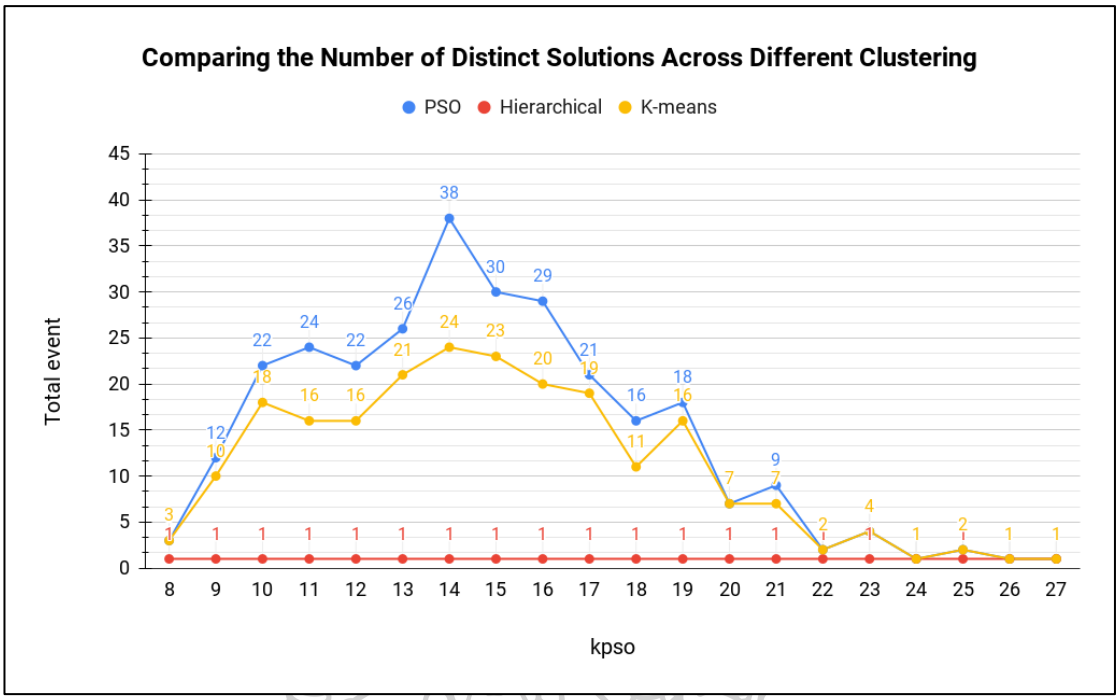
อย่างไรก็ตาม ตารางที่ 4.16, 4.17 และ 4.18 ซึ่งอธิบายจำนวนคำตอบที่แตกต่างกันตามประเภทของการจัดกลุ่ม พบข้อแตกต่างเล็กน้อยระหว่างข้อมูลค่าสมการจุดประสงค์ ค่าเฉลี่ยเส้นผ่านศูนย์กลางและค่าเฉลี่ยระยะห่างระหว่างกลุ่มในการทดสอบตัวจัดกลุ่มประเภทเดียวกัน ตัวอย่าง เช่น การจัดกลุ่มประเภท PSO ได้กลุ่มข้อมูลเป็น 8 กลุ่ม ดังแสดงในตารางที่ 4.19 จะสังเกตว่าเมื่อเปรียบการจัดกลุ่มด้วยตัวจัดกลุ่มทั้ง 3 ประเภท ประเภทลำดับขั้นให้ข้อมูลในรูปแบบเดียวทั้งค่าสมการจุดประสงค์ ค่าเฉลี่ยเส้นผ่านศูนย์กลางและค่าเฉลี่ยระยะห่างระหว่างกลุ่ม ประเภทเคมีนพบความแตกต่างของคำตอบเป็น 3 รูปแบบ และประเภท PSO ที่เมื่อพิจารณาผลค่าเฉลี่ยเส้นผ่านศูนย์กลางและระยะห่างระหว่างกลุ่มร่วมด้วยจะพบว่าในการแบ่งกลุ่มที่ทำให้ได้ค่าสมการจุดประสงค์เท่ากัน ได้แก่ 73.39424 มีการจัดกลุ่มที่ให้ขนาดเส้นผ่านศูนย์กลางและระยะห่างระหว่างกลุ่มที่แตกต่างกันได้อีก 1 รูปแบบ

ตารางที่ 4.19 ตัวอย่างเปรียบเทียบค่าสมการจุดประสงค์ $\mathcal{L}_2$ ค่าเฉลี่ยเส้นผ่านศูนย์กลางและระยะห่างระหว่างกลุ่มเมื่อแบ่งข้อมูลได้ 8 กลุ่ม

kpso	obj.PSO	obj.Hc	obj. kmeans	PSO. diameter	Hc. diameter	Km. diameter	PSO. separation	Hc. separation	Km. separation
8	73.508	66.134	56.122	1.059	0.771	0.826	0.891	1.004	0.786
8	73.394	66.134	55.418	1.060	0.771	0.885	0.900	1.004	0.868
8	73.394	66.134	55.418	1.058	0.771	0.885	0.899	1.004	0.868
8	73.213	66.134	57.477	1.060	0.771	0.849	0.914	1.004	0.766
8	73.394	66.134	55.418	1.060	0.771	0.885	0.900	1.004	0.868

4.4 วิจารณ์ผลการทดลอง

การทดสอบการแบ่งกลุ่มยีนส์ด้วยตัวจัดกลุ่มแบบ PSO ที่สร้างขึ้นในงานวิจัยนี้ ได้วิเคราะห์และเก็บข้อมูลผลงานวิจัยดังกล่าว ทั้งค่าสมการจุดประสงค์ $\mathcal{L}_2$ ซึ่งใช้เป็นสมการจุดประสงค์หลักในการทำงานของตัวจัดกลุ่มที่มีแนวคิดจากขั้นตอนวิธีเชิงวิวัฒนาการ แบบความฉลาดแบบฝูงอนุภาค จากการศึกษาทั้งหมด 431 ครั้ง พบว่าตัวจัดกลุ่มแบบ PSO อาจไม่ให้ผลดีนักในแง่ของความกะทัดรัดของกลุ่มข้อมูล หรือความแยกกันอย่างเห็นได้ชัดของกลุ่มข้อมูล เมื่อเทียบกับตัวจัดกลุ่มโดยเครื่องมือ pathfindR ที่ใช้การจัดกลุ่มแบบลำดับขั้น แต่อย่างไรก็ตาม สามารถสังเกตลักษณะเด่น ๆ ของการทำงานตัวจัดกลุ่มประเภท PSO ที่สามารถปรับปรุงเปลี่ยนแปลงจำนวนกลุ่มที่ควรแบ่งได้ หรือค่า kpso ได้หลายรูปแบบ ไม่ว่าจะกำหนดค่า k เริ่มต้นที่เท่าใดก็ตาม เช่น 19 25 หรือ หรือ 30 ก็ยังคงสามารถให้คำตอบที่เหมาะสม และคำตอบที่ได้จากการแบ่งกลุ่มประเภท PSO นี้สามารถแบ่งกลุ่มได้ถึง 10 รูปแบบ ตั้งแต่ 8 ถึง 27 กลุ่ม ด้วยเงื่อนไขที่ต้องรักษาค่าสมการจุดประสงค์ให้มีค่าสูงที่สุดอยู่เสมอ อีกทั้งยังให้ความหลากหลายในรูปแบบของการจัดกลุ่มต่าง ๆ ในสถานการณ์ที่แบ่งได้จำนวนกลุ่มที่เท่ากัน ดังตัวอย่างทั้งหมดที่กล่าวมาในบทนี้ โดยแสดงออกเป็นจำนวนคำตอบที่แตกต่างกันที่ตัวจัดกลุ่มประเภท PSO สามารถหาให้ได้ เปรียบเทียบกับประเภทอื่น ๆ ได้ดังรูปที่ 4.15



รูปที่ 4.15 เปรียบเทียบจำนวนคำตอบที่แตกต่างกันจากตัวจัดกลุ่มทั้ง 3 ประเภท จากการทดสอบ 431 ครั้ง

บทที่ 5

สรุปผลการวิจัย

วิทยานิพนธ์นี้ได้สร้างตัวจัดกลุ่มประเภท PSO ขึ้น โดยมีที่มาจากขั้นตอนวิธีเชิงวิวัฒนาการประเภทความฉลาดแบบกลุ่ม โดยเลือกใช้วิธีคล้ายกับการทำงานของฝูงอนุภาค (Particle swarm optimization) ซึ่งเป็นขั้นวิธีสำหรับการแก้ปัญหาการค้นหาค่าเหมาะสมที่สุดนำมาประยุกต์ใช้เป็นขั้นตอนวิธีสำหรับปัญหาการจัดกลุ่มข้อมูล เพื่อที่จะทำให้ผลของการจัดกลุ่มมีความหลากหลายของคำตอบและมีคุณสมบัติของการเป็นกลุ่มที่ดีขึ้น ในวิทยานิพนธ์นี้ได้ทำการทดสอบตัวจัดกลุ่มที่สร้างขึ้นกับข้อมูลเฉพาะทางชีวสารสนเทศ ในกระบวนการไมโครอะเรย์ หรือกระบวนการศึกษาทางชีววิทยาที่ศึกษาลักษณะการแสดงออกเซลล์ที่ส่งผลต่อการเกิดโรคที่นักชีววิทยาสนใจในระดับยีนส์ ในที่นี้ได้ใช้ข้อมูลจากยีนส์ในนิวเคลียสของเซลล์เม็ดเลือดขาวแบบเดี่ยวของผู้ป่วยที่ได้รับการวินิจฉัยว่าเป็นโรคข้ออักเสบรูมาตอยด์จำนวน 18 คน เทียบกับผู้ป่วยสุขภาพดี 15 คน (Rheumatoid arthritis dataset GSE15573) เปรียบเทียบการจัดกลุ่มข้อมูลตัวอย่างดังกล่าวกับเครื่องมือ pathfindR ซึ่งเป็นเครื่องมือใช้สำหรับวิเคราะห์ความสำคัญของกลุ่มยีนส์ (Gene-Set Enrichment Analysis; GSEA) โดยขั้นตอนวิธีการจัดกลุ่มข้อมูลสำหรับเครื่องมือนี้ ได้แก่ การจัดกลุ่มข้อมูลแบบลำดับขั้น นอกจากนี้ยังได้เปรียบเทียบผลงานวิจัยกับการจัดกลุ่มประเภทเคมีน จากเครื่องมือ stats ซึ่งใช้งานอย่างแพร่หลายด้วยโปรแกรม Rstudio เพื่อศึกษาถึงประสิทธิภาพการทำงานเปรียบเทียบกันในแง่ของค่าสมการจุดประสงค์ ค่าเฉลี่ยเส้นผ่านศูนย์กลางรวมถึงค่าเฉลี่ยระยะห่างระหว่างกลุ่ม

5.1 บทสรุปการวิจัย

ผลการทดสอบตัวจัดกลุ่มประเภท PSO ที่ผู้วิจัยสร้างขึ้นพบว่ามีลักษณะการทำงานของขั้นตอนวิธีที่แตกต่างจากตัวจัดกลุ่มทั่วไป ได้แก่ ประเภทลำดับขั้น และประเภทเคมีน ที่ต้องมีกำหนดจำนวนกลุ่มที่ต้องการแบ่ง (Initial k) ให้กับขั้นตอนวิธีแบ่งกลุ่มประเภทต่าง ๆ ในตอนเริ่มต้น และผลลัพธ์ในท้ายที่สุดจะได้คำตอบที่มีจำนวนกลุ่มตามที่ได้กำหนดไว้ แต่การจัดกลุ่มในงานวิจัยนี้มีลักษณะพยายามปรับปรุงเปลี่ยนค่า k เริ่มต้นนั้น ให้กลายเป็นจำนวนกลุ่มแบบอื่น เพื่อรักษาค่าสมการจุดประสงค์ให้มีค่าสูงที่สุดเท่าที่จะพยายามหาค่าได้ โดยที่ในงานวิจัยนี้เรียกว่าค่า kpso เป็นค่าที่ทำให้ผลลัพธ์สุดท้ายสามารถแบ่งกลุ่มได้เป็นเป็นจำนวนหลายรูปแบบ จากการทดสอบ 431 ครั้ง มีจำนวนกลุ่มคำตอบตั้งแต่ 8 กลุ่ม ไปจนถึง 27 กลุ่ม รวมเป็น 10 รูปแบบสำหรับการจัดกลุ่มประเภท PSO และรูปแบบที่ให้ค่าสมการจุดประสงค์ $\mathcal{L}_2$ มากที่สุดในงานวิจัยนี้ได้แก่ 73.5088 เมื่อแบ่งข้อมูล

ออกเป็น 8 กลุ่ม เทียบกับการจัดกลุ่มจากเครื่องมือ pathfindR ที่แนะนำค่า k ที่เหมาะสมเท่ากับ 19 กลุ่ม ซึ่งเมื่อคำนวณค่าสมการจุดประสงค์ $\mathcal{L}_2$ จะมีค่า 48.8432 นอกจากนี้พบรูปแบบคำตอบที่แตกต่างกันออกไปภายใน 10 รูปแบบ เช่น การแบ่งกลุ่มข้อมูลออกเป็น 14 กลุ่ม พบความแตกต่างของคำตอบมากถึง 38 รูปแบบจากการแบ่งกลุ่มประเภท PSO และการแบ่งกลุ่มประเภทเคมีนพบความแตกต่าง 24 รูปแบบ ในขณะที่การแบ่งกลุ่มประเภทลำดับขั้นจากเครื่องมือ pathfindR พบเพียงรูปแบบคำตอบเดียวตลอดทำการทดสอบ สามารถสรุปผลความหลากหลายของจำนวนคำตอบที่แตกต่างกันจากการจัดกลุ่มประเภทต่าง ๆ ได้ดังรูปที่ 4.15 แสดงให้เห็นจำนวนคำตอบที่เกิดขึ้นมากสุดสามารถเกิดขึ้นได้หลายช่วง เช่น kpso ที่ 11 14 15 16 19 และ 21 เป็นต้น

ทั้งนี้ในแง่คุณสมบัติความเป็นกลุ่มที่พิจารณาโดยเกณฑ์ที่ใช้สำหรับวิเคราะห์ค่า k ที่เหมาะสมในขั้นตอนวิธีเพื่อแก้ปัญหาการจัดกลุ่มทั่วไป (Internal validation index) อาทิ ความกะทัดรัด (Compactness) และระยะห่างระหว่างกลุ่ม (Separation) พบว่าการจัดกลุ่มแบบ PSO อาจไม่ให้ผลดีทางด้านความกะทัดรัดของกลุ่มข้อมูลเมื่อเทียบกับการแบ่งกลุ่มประเภทเคมีน และแบบลำดับขั้นและอาจไม่ให้ผลดีในแง่การแยกกันอย่างเห็นได้ชัดของกลุ่มข้อมูล เมื่อเทียบกับการแบ่งกลุ่มแบบลำดับขั้น แต่ยังเห็นผลดีที่เหนือกว่าการจัดกลุ่มแบบเคมีน ซึ่งอาจเกิดได้จากหลายปัจจัย เช่น การทำงานของตัวจัดกลุ่มประเภท PSO มีการพยายามสุ่มคำตอบให้มีการกระจายเส้นผ่านศูนย์กลางของกลุ่มข้อมูลให้มีความแตกต่างกันน้อยกว่าประเภทอื่น ๆ หรือคำตอบที่ได้จากการแบ่งข้อมูลประเภทลำดับขั้นมักเจอคำตอบที่มีสมาชิก 1 ตัวต่อ 1 กลุ่ม เสมอ เป็นผลทำให้ตัวจัดกลุ่มแบบ PSO ไม่สามารถเห็นผลดีในด้านนี้ได้

อย่างไรก็ตาม ปัญหาการจัดกลุ่มข้อมูลเป็นปัญหาที่ใช้ขั้นตอนวิธีการเรียนรู้ของเครื่องแบบไม่มีผลเฉลย ประกอบกับนำมาใช้เพื่อค้นหาคำตอบทางด้านชีวสารสนเทศ ซึ่งการยืนยันผลของการจัดกลุ่มจะได้รับคำตอบอย่างแท้จริงก็ต่อเมื่อนำไปพิสูจน์กับการทดลองในห้องปฏิบัติการจริง เป็นผลให้การมีวิธีจัดกลุ่มเพื่อหาคำตอบที่ให้ความหลากหลายอาจเป็นผลดีในการศึกษาแก่นักวิจัยทางด้านการให้ตัวเลือกที่หลากหลายเพื่อการศึกษาในกระบวนการอื่นต่อไป

5.2 ข้อเสนอแนะ

เนื่องจากผลการดำเนินการวิจัยพบว่าการจัดกลุ่มข้อมูลด้วยวิธี PSO ให้ผลดีในด้านความหลากหลายของรูปแบบคำตอบ แต่อย่างไรก็ตาม คุณสมบัติของการเป็นกลุ่มข้อมูลในแง่ความกะทัดรัด (Compactness) ยังเป็นรองต่อการจัดกลุ่มแบบลำดับขั้น และเคมีน รวมถึงระยะห่างระหว่างกลุ่มของข้อมูล (Separation) ที่ยังมีผลเป็นรองต่อการจัดกลุ่มแบบลำดับขั้นนั้น ทำให้ผู้วิจัยมี

ข้อเสนอแนะโดยเพิ่มสมการจุดประสงค์ในการค้นหาคำตอบด้วยตัวจัดกลุ่มประเภท PSO กลายเป็นปัญหาการหาค่าเหมาะสมที่สุดแบบหลายวัตถุประสงค์ เช่น ใช้สมการค่าเฉลี่ยเส้นผ่านศูนย์กลางของกลุ่มข้อมูล และใช้สมการระยะห่างระหว่างกลุ่มที่น้อยที่สุด เป็นต้น หรืออาจพิจารณาเป็นการใช้ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบหลายจุดประสงค์รูปแบบอื่น อาทิ ขั้นตอนเชิงพันธุกรรมแบบการจัดลำดับที่ไม่ถูกครอบงำ เป็นต้น นอกจากนี้ในระหว่างดำเนินการศึกษายังพบการสุ่มประชากรที่มีรูปแบบการกระจายตัวแบบสม่ำเสมอ หรือการแจกแจงเอกรูป (Uniform distribution) ซึ่งเป็นผลจากชุดคำสั่งสุ่มตัวแปรในขั้นตอนสุ่มประชากรของฝูงด้วยชุดคำสั่งของโปรแกรมภาษาอาร์ ได้แก่ ชุดคำสั่ง runif เป็นต้น ด้วยเหตุนี้ อาจเป็นปัจจัยหนึ่งที่ทำให้ผลงานวิจัยในแง่คุณลักษณะของกลุ่มข้อมูลประเภทความกะทัดรัด มีส่วนเบี่ยงเบนมาตรฐานน้อยที่สุดเมื่อจัดกลุ่มด้วยตัวจัดกลุ่มข้อมูลประเภท PSO ดังนั้นเพื่อเพิ่มประสิทธิภาพของคุณลักษณะความเป็นกลุ่มข้อมูลโดยรักษาไว้ซึ่งแนวคิดของความเป็นกลุ่มเดียวกันจะมีขนาดเส้นผ่านศูนย์กลางของข้อมูลน้อย ในขณะที่ข้อมูลที่ควรอยู่ต่างกลุ่มกันควรมีระยะห่างระหว่างกันอย่างมาก อีกทั้งยังคงไว้ซึ่งความหลากหลายในการค้นหาคำตอบ จึงควรพัฒนาขั้นตอนวิธีเชิงวิวัฒนาการในการสร้างตัวจัดกลุ่มโดยให้ความสำคัญกับขั้นตอนการสุ่มประชากรในฝูง และศึกษาขั้นตอนวิธีเชิงวิวัฒนาการสำหรับหลายวัตถุประสงค์หรือประยุกต์ใช้ขั้นตอนวิธีการค้นหาแบบกลุ่มอนุภาคเพื่อแก้ปัญหามากกว่า 1 วัตถุประสงค์

รายการอ้างอิง

1. Ho, K.S., et al., *Chromosomal Microarray Analysis of Consecutive Individuals with Autism Spectrum Disorders Using an Ultra-High Resolution Chromosomal Microarray Optimized for Neurodevelopmental Disorders.* Int J Mol Sci, 2016. **17**(12).
2. López-Campos, G., et al., *Bioinformatics in Support of Microarray Experiments*, in *Microarray Detection and Characterization of Bacterial Foodborne Pathogens*, G. López-Campos, et al., Editors. 2012, Springer US: Boston, MA. p. 49-92.
3. Bilotta, M., G. Tradigo, and P. Veltri. *Bioinformatics Data Models, Representation and Storage*. in *Encyclopedia of Bioinformatics and Computational Biology*. 2019.
4. Gasperskaja, E. and V. Kučinskas, *The most common technologies and tools for functional genome analysis.* Acta Med Litu, 2017. **24**(1): p. 1-11.
5. Yoon, S., et al., *GScluster: network-weighted gene-set clustering analysis.* BMC Genomics, 2019. **20**(1): p. 352.
6. Huang, D.W., et al., *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.* Genome Biology, 2007. **8**(9): p. R183.
7. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
8. Ulgen, E., O. Ozisik, and O.U. Sezerman, *pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks.* Front Genet, 2019. **10**: p. 858.
9. Di Gesú, V., et al., *GenClust: A genetic algorithm for clustering gene expression data.* BMC Bioinformatics, 2005. **6**(1): p. 289.
10. Mukhopadhyay, A., U. Maulik, and S. Bandyopadhyay, *A Survey of Multiobjective Evolutionary Clustering.* ACM Comput. Surv., 2015. **47**(4): p. Article 61.
11. Figueiredo, E., et al., *Swarm intelligence for clustering — A systematic review*

with new perspectives on data mining. Eng. Appl. Artif. Intell., 2019. **82**(C): p. 313–329.

12. Aggarwal, C.C. and C.K. Reddy, *Data Clustering: Algorithms and Applications*. 2013: Chapman & Hall/CRC.
13. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
14. Li, X., et al., *Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease.* Sci Rep, 2015. **5**: p. 12393.
15. Kacprzyk, J.a.W.P., *Springer handbook of computational intelligence*. 2015: Springer.
16. Murata, Y., et al., *Cluster analysis and display of genome-wide expression profiles in dimethyl sulfoxide treatment.* Chem-bio Informatics Journal, 2002. **2**: p. 18-31.
17. Merico, D., et al., *Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.* PLoS One, 2010. **5**(11): p. e13984.
18. Andreopoulos, B., A. An, and X. Wang, *Bi-level clustering of mixed categorical and numerical biomedical data.* Int J Data Min Bioinform, 2006. **1**(1): p. 19-56.
19. Nanda, S.J. and G. Panda, *A survey on nature inspired metaheuristic algorithms for partitional clustering.* Swarm and Evolutionary Computation, 2014. **16**: p. 1-18.
20. Zhao, Y. and G. Karypis. *Criterion Functions for Document Clustering * Experiments and Analysis*. 2001.

มหาวิทยาลัยศิลปากร

ประวัติผู้เขียน

ชื่อ-สกุล พชรอร แสนประเสริฐ

วัน เดือน ปี เกิด 28 มิถุนายน 2537

สถานที่เกิด นครปฐม