การเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของรหัสพันธุกรรมเนื้องอกในสมองสำหรับข้อมูลที่มีมิติขั้นสูง

Comparing K-Mean Clustering Methods of DNA in Brain Tumors for High-Dimensional Data

อัชฌา อระวีพร[1*] และ จารวี พร้อมสง่า[1]
Autcha Araveeporn[1*] and Jarawee Promsanga[1]
[1]ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
[1]Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok
วันที่ส่งบทความ : 20 กันยายน 2565 วันที่แก้ไขบทความ : 25 เมษายน 2566 วันที่ตอบรับบทความ : 13 พฤศจิกายน 2566
Received: 20 September 2022, Revised: 25 April 2023, Accepted: 13 November 2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแบ่งกลุ่มรหัสพันธุกรรมของคนไข้ที่ป่วยเป็นเนื้องอกในสมอง แบบเคมีน 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง วิธีฟอร์กี้ และวิธีแม็คควีน มีตัวแปรอิสระเป็นข้อมูลรหัสพันธุกรรม 989 รหัส และตัวแปรตามคือระดับการเป็นเนื้องอกในสมองในคนไข้ 43 คน ซึ่งในกรณีนี้จำนวนตัวแปรอิสระจะมีจำนวนมากกว่าจำนวนคนไข้ หรือที่เรียกว่าข้อมูลที่มีมิติขั้นสูง โดยทำการทดลองสุ่มข้อมูลรหัสพันธุกรรมจำนวน 200, 400, 600 และ 800 และกำหนดจำนวนกลุ่มคือ 5, 10, 15, 20, 25 และ 30 ทำการสุ่มซ้ำ 1,000 ครั้ง เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของการแบ่งกลุ่มคือค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่ม จากผลการแบ่งกลุ่มแบบเคมีนทั้งหมด 3 วิธี พบว่าวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่องให้ประสิทธิภาพดีที่สุดสำหรับทุกสถานการณ์ ซึ่งให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุดเมื่อเปรียบเทียบกับวิธีฟอร์กี้และวิธีแม็คควีน โดยจำนวนตัวแปรอิสระไม่ส่งผลต่อประสิทธิภาพการแบ่งกลุ่ม

คำสำคัญ : การแบ่งกลุ่ม การแบ่งกลุ่มแบบเคมีน ฮาร์ติกัน-หว่อง ฟอร์กี้ แม็คควีน

Abstract

This study aims to compare the performance of clustering DNA of brain tumor patients of k-means three methods, namely the Hartigan-Wong, Forgy, and MacQueen methods. The independent variables are DNA as 989 genes, and the dependent variable is the level of a brain tumor in 43 patients. In this case, the number of the independent

*ที่อยู่ติดต่อ E-mail address: autcha.ar@kmitl.ac.th

variable is larger than the number of patients or called the high-dimensional data. The experiment is conducted by random DNA samples of 200, 400, 600, and 800 genes and fixed 5, 10, 15, 20, 25, and 30 groups by 1,000 replications. Comparing clustering performance is the mean data differences between the groups' criteria. The results of k-means clustering methods find that the Hartigan-Wong method has the best performance for all situations. However, the Hartigan-Wong method shows the most significant difference in data between groups compared to Forgy and MacQueen methods. The number of independent variables has not affected clustering performance.

Keywords: Clustering, K-Mean Clustering, Hartigan-Wong, Forgy, MacQueen

1. บทนำ

ในปัจจุบันเทคโนโลยีทางการแพทย์มีความก้าวหน้ามากขึ้น โดยเฉพาะเทคโนโลยีในการถอดรหัสพันธุกรรม (Deoxyribonucleic Acid หรือ DNA) โดยนักวิทยาศาสตร์ นักวิจัย สามารถศึกษาจากการแสดงออกของยีน (Gene expression) ซึ่งเป็นกระบวนการถ่ายทอดข้อมูลทางรหัสพันธุกรรม หรือยีนเพื่อนำไปวินิจฉัยโรค และสามารถรักษาได้อย่างตรงจุด โดยข้อมูลของรหัสพันธุกรรมนั้นมีชนิดของยีนเป็นจำนวนมาก โดยยีนเหล่านี้จะแสดงลักษณะของโรค การแบ่งกลุ่ม (Cluster) จากข้อมูลของยีนบางส่วนที่มีความสำคัญมาใช้ในการจำแนกผู้ป่วยโรคต่าง ๆ นั้น โดยใช้วิธีการวิเคราะห์แบ่งกลุ่ม (Cluster analysis) ซึ่งการแบ่งกลุ่มที่มีประสิทธิภาพนั้นสามารถวินิจฉัยและรักษาได้อย่างรวดเร็ว โดยเฉพาะ โรคมะเร็ง ถ้าทำการรักษาหรือจำแนกผู้ป่วยจะช่วยให้ผู้ป่วยสามารถรอดชีวิตได้

การวิเคราะห์แบ่งกลุ่มจะทำการแบ่งข้อมูลเป็นกลุ่ม ๆ เพื่อที่จะทำให้เข้าใจข้อมูลได้ดียิ่งขึ้น หรือใช้ในการค้นหาโครงสร้างในกลุ่มข้อมูล ขั้นตอนวิธีในการแบ่งกลุ่มข้อมูลจะแบ่งชุดข้อมูลออกเป็นกลุ่ม โดยที่ข้อมูลที่มีความคล้ายกันจะถูกกำหนดให้ไปอยู่กลุ่มเดียวกัน ในทางตรงกันข้ามข้อมูลที่ไม่คล้ายกันหรือแบ่งแยกกันได้จะถูกกำหนดให้อยู่ในกลุ่มที่แตกต่างกัน เพื่อช่วยในการลดขนาดข้อมูล วิธีการในการแบ่งกลุ่มจะทำการคำนวณการวัดระยะห่างระหว่างข้อมูลหนึ่งกับข้อมูลหนึ่งด้วยวิธีการต่าง ๆ เช่น การวัดระยะห่างแบบยูคลิเดียน (Euclidean distance) การวัดระยะห่างแบบแมนฮัตตัน (Manhattan distance) และการวัดระยะห่างแบบโคไซน์ (Cosine distance) เป็นต้น Zarikas และคณะ [1] ได้ทำการสำรวจผลกระทบทางการแพทย์ เศรษฐกิจ และสังคมวิทยาของการระบาดใหญ่ของโควิด-19 โดยทำการจัดกลุ่มประเทศตามลักษณะการระบาดของประชากรในประเทศนั้น ๆ ผลที่ได้จะนำไปกำหนดนโยบายต่าง ๆ เช่น จำนวนแพทย์ พยาบาล และนโยบายทางเศรษฐกิจ

การแบ่งกลุ่มแบบเคมีนเป็นวิธีที่นิยมใช้ในการแบ่งกลุ่มข้อมูล โดยเปรียบเทียบความคล้ายคลึงของข้อมูล กับจุดศูนย์กลาง (Centroid) หรือค่าเฉลี่ย (Mean) ของแต่ละกลุ่ม เป็นการแบ่งส่วน (Partition clustering) ด้วยการแบ่งข้อมูลออกเป็นส่วนตามจำนวนกลุ่มที่ต้องการ Nurlaila และคณะ [2] ศึกษาแบบจำลองการแบ่งกลุ่มเคมีนเพื่อแยกแบคทีเรียที่ทนต่อทองแดง พบว่าแบบจำลองการแบ่งกลุ่มแบบเคมีนจะไม่ได้ดึงแบคทีเรียชนิดเดียวกันเข้าไปในกลุ่มเดียวกัน แต่แบบจำลองนี้จะรวบรวมความคล้ายคลึง

กันในลักษณะเดียวกัน แสดงให้เห็นว่าแบบจำลองการแบ่งกลุ่มเคมีนมีความไวต่อการตรวจจับความแตกต่างกันของชนิดแบคทีเรีย Shan [3] ได้นำวิธีการแบ่งกลุ่มแบบเคมีนมาใช้ในการจำแนกภาพจากค่าเฉลี่ยความแม่นยำ พบว่าวิธีเคมีนสามารถจำแนกภาพได้เป็นอย่างดี

ในการแบ่งกลุ่มแบบเคมีนนั้นมีวัตถุประสงค์ในการแบ่งข้อมูลเป็น k กลุ่ม โดยปกติกระบวนการ (Algorithm) ในการแบ่งกลุ่มข้อมูลจะใช้วิธีฮาร์ติกัน-หว่อง (Hartigan-Wong) [4] แต่ก็มีบางครั้งที่ผู้ใช้ได้นำวิธีแม็คควีน (MacQueen) [5] วิธีฟอร์กี้ (Forgy) [6] และวิธีลอยด์ (Lloyd) [7] มาใช้ในการแบ่งกลุ่มข้อมูล เนื่องจากบางวิธีใช้งานได้สะดวกและรวดเร็วกว่าในการแบ่งกลุ่มข้อมูล Yadav และ Sharma [8] ได้ใช้กระบวนการแบ่งกลุ่มแบบเคมีน ด้วยวิธีแม็คควีนเพื่อเพิ่มความเร็วในการแบ่งกลุ่มและลดกระบวนการแบ่งกลุ่มที่ซับซ้อน Singh และ Rajpoot [9] เปรียบเทียบกระบวนการแบ่งกลุ่มแบบเคมีน ด้วยวิธีวิธีฮาร์ติกัน-หว่อง วิธีแม็คควีน วิธีฟอร์กี้ วิธีลอยด์ และวิธีอื่น ๆ กับข้อมูลสามชุด พบว่าทั้งสี่วิธีมีประสิทธิภาพในการแบ่งกลุ่มแตกต่างตามลักษณะข้อมูลเชิงปริมาณ เช่น ชุดข้อมูลดอกไม้ (Iris dataset) มีข้อมูลคือ ความกว้างความยาวของกลีบใบและกลีบดอก แล้วจึงนำมาวิเคราะห์แบ่งกลุ่มดอกไม้

ในงานวิจัยนี้พิจารณาข้อมูลทางด้านการแพทย์ซึ่งในการจัดเก็บข้อมูลนั้นมีข้อจำกัดมากจึงทำให้ไม่สามารถเก็บได้ครบ เช่น คนไข้ไม่มาตามนัด สูญหาย ติดต่อไม่ได้ ทำให้ได้ข้อมูลไม่ครบถ้วนจึงต้องทำการตัดข้อมูลชุดนั้นทิ้งไป โดยมีตัวแปรอิสระ (Independent variable) เป็นข้อมูลเชิงปริมาณ และตัวแปรตาม (Dependent variable) เป็นข้อมูลเชิงคุณภาพ โดยจำนวนตัวแปรอิสระมีจำนวนมากกว่าขนาดตัวอย่างที่เก็บมา เรียกข้อมูลประเภทนี้ว่า ข้อมูลที่มีมิติขั้นสูง (High-dimensional data) เช่น ข้อมูลไมโครอาร์เรย์ (Microarray data) ซึ่งเป็นข้อมูลที่ได้จากการศึกษารหัสพันธุกรรมที่ควบคุมการแสดงออกของยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน โดยยีนที่ศึกษามีจำนวนเป็นหลักพันหรือหลักหมื่น สามารถนำมาใช้ในการจำแนกเนื้อเยื่อมะเร็งและเนื้อเยื่อปกติได้

ทั้งนี้การแบ่งกลุ่มข้อมูลของตัวแปรอิสระที่มีจำนวนมากจะช่วยลดการวิเคราะห์ข้อมูลที่มีปริมาณมาก ๆ ช่วยเพิ่มความเร็วและประสิทธิภาพในการวิเคราะห์ข้อมูลได้เป็นอย่างดี ยกตัวอย่างงานวิจัย เช่น งานวิจัยเรื่องการพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง อาริกา ธรรมโน และคณะ [10] ได้นำเสนอกระบวนการแบ่งกลุ่มแบบเคมีนให้มีความสามารถในการจำแนกประเภทผู้ป่วยโรคมะเร็งเต้านม และปรับเปลี่ยนค่าถ่วงน้ำหนักในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของวิธีเคมีน Jothi และคณะ [11] ได้ศึกษากระบวนการแบ่งกลุ่มแบบเคมีนสำหรับการวิเคราะห์ข้อมูลรหัสพันธุกรรม โดยทำการเปรียบเทียบวิธีแบ่งกลุ่มแบบเคมีนและปรับปรุงวิธีเคมีนให้การแบ่งกลุ่มได้รวดเร็วและมีประสิทธิภาพมากขึ้น Saadeh และคณะ [12] ได้นำวิธีการแบ่งกลุ่มแบบเคมีนมาจัดกลุ่มข้อมูลรหัสพันธุกรรมในการแบ่งกลุ่มยีนเพื่อดูการพัฒนาเซลล์เม็ดเลือดแดง Joshi และคณะ [13] ได้นำเสนอวิธีการจัดกลุ่มเอกสารที่มีหัวข้อเกี่ยวข้องกันให้มาอยู่ในกลุ่มเดียวกัน (Latent Dirichlet Allocation) มาใช้สำหรับการจัดกลุ่มยีนโดยวิธีเคมีนเพื่อจำแนกเนื้อเยื่อปกติและเนื้อเยื่อมะเร็งประเภทต่าง ๆ

ในการวัดประสิทธิภาพของการแบ่งกลุ่มด้วยวิธีเคมีนนั้น จะใช้การวัดค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared: RS) โดยถ้าค่าความแตกต่างระหว่างกลุ่มมีมาก หมายความว่ามีการแบ่งกลุ่มที่ดี ซึ่งจะมีค่าอยู่ในช่วง 0 ถึง 1 โดย Bhatt และคณะ [14] ได้ทำการศึกษาการแบ่งกลุ่มแบบเคมีนสำหรับ

ข้อมูลที่มีค่านอกกลุ่มโดยใช้เกณฑ์คือค่าความต่างของข้อมูลระหว่างกลุ่ม Thakare และ Bagal [15] ได้หาประสิทธิภาพกระบวนการแบ่งกลุ่มแบบเคมีนโดยใช้หลายวิธีในการเปรียบเทียบและพบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีความเหมาะสมมากที่สุด Meng และคณะ [16] ได้ทำการศึกษากระบวนการแบ่งกลุ่มด้วยวิธีเคมีนโดยพิจารณาจากระยะห่างจากจุดศูนย์กลางให้มีค่าต่ำที่สุดจากค่าความแตกต่างของข้อมูลระหว่างกลุ่ม

เนื่องจากการแบ่งกลุ่มแบบเคมีนเป็นที่นิยมสำหรับการแบ่งกลุ่มข้อมูล ดังนั้นผู้วิจัยจึงทำการเปรียบเทียบประสิทธิภาพการแบ่งกลุ่มข้อมูลที่มีมิติขั้นสูง แบบเคมีนจำนวน 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง วิธีฟอร์กี้ และวิธีแม็คควีน

2. วิธีการทดลอง

ในงานวิจัยนี้ได้ทำการศึกษาโดยทำการทดลองแบ่งกลุ่มข้อมูล (Clustering) ของตัวแปรอิสระ เมื่อข้อมูลอยู่ในรูปแบบข้อมูลที่มีมิติขั้นสูง ซึ่งมีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง การที่จำนวนตัวแปรอิสระมีขนาดใหญ่มากจึงควรทำการแบ่งกลุ่มข้อมูลออกเป็นกลุ่ม โดยตัวแปรอิสระที่อยู่ในกลุ่มเดียวกันจะมีความเกี่ยวพันกันหรือมีลักษณะที่คล้ายคลึงกัน งานวิจัยนี้ได้ทำการศึกษาการแบ่งกลุ่มแบบเคมีน (K-means clustering) เนื่องจากการแบ่งกลุ่มแบบเคมีนเป็นที่นิยมใช้โดยทั่วไปมากกว่าวิธีอื่น ๆ นอกจากนี้ยังมีความสะดวกและแม่นยำกว่าวิธีอื่น ๆ [17]

การแบ่งกลุ่มแบบเคมีน หรือการวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical cluster analysis) คือ กระบวนการแบ่งกลุ่มประเภทหนึ่งที่เป็นเทคนิคในการตัดแบ่ง (Partition) วัตถุออก เป็น k กลุ่ม ซึ่งในแต่ละกลุ่มแทนค่าด้วยค่าเฉลี่ยของกลุ่ม และเป็นจุดศูนย์กลางของกลุ่มที่ใช้ในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยเลือกกลุ่มที่แบ่งที่มีระยะห่างจากค่ากลางของกลุ่มที่น้อยที่สุด แล้วทำการประมวลหาค่ากลางของกลุ่มใหม่ ทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มเปลี่ยนแปลงน้อยกว่าค่าที่กำหนดไว้หรือครบจำนวนรอบที่กำหนดไว้ ในการทดลองนี้พิจารณาการแบ่งกลุ่มข้อมูลแบบเคมีน 3 วิธี ดังนี้

2.1 วิธีฮาร์ติกัน-หว่อง

กระบวนการของวิธีฮาร์ติกัน-หว่อง [4] จะทำการปรับจุดศูนย์กลางด้วยการพิจารณาจากแต่ละจุดข้อมูล โดยเริ่มแรกจะกำหนดจุดข้อมูลทั้งหมดให้กับจุดศูนย์กลางแบบสุ่ม จากนั้นจะปรับจุดศูนย์กลางโดยพิจารณาจากแต่ละจุดข้อมูล และทำการตัดแบ่งข้อมูลด้วยผลบวกกำลังสองของความคลาดเคลื่อน (Sum of Squares Error: SSE) ภายในกลุ่ม จากนั้นจุดข้อมูลต่าง ๆ จะถูกกำหนดให้กับจุดศูนย์กลางที่ใกล้ที่สุด และจุดศูนย์กลางจะถูกคำนวณใหม่เป็นค่าเฉลี่ยของจุดข้อมูลที่กำหนด ในการพิจารณาแต่ละจุดข้อมูล หากผลรวมกำลังสองของความคลาดเคลื่อนของกลุ่มอื่นมีค่าน้อยกว่าผลรวมกำลังสองของความคลาดเคลื่อนของกลุ่มปัจจุบัน ดังสมการที่ (1) จุดข้อมูลนั้นจะถูกส่งไปยังกลุ่มอื่น และจะทำการวนซ้ำไปจนกว่าจะไม่มีการเปลี่ยนกลุ่ม กระบวนการนี้มีข้อดีคือใช้กับข้อมูลที่มีจำนวนมากและจำนวนการแบ่งกลุ่มน้อย แต่มีข้อเสียคือต้องทราบจำนวนกลุ่มที่เหมาะสมถึงจะทำให้การแบ่งกลุ่มมีประสิทธิภาพ

$$SSE_i = \frac{N_i \sum_j \left\| x_{ij} - c_i \right\|^2}{N_i - 1} < SSE_1 = \frac{N_1 \sum_j \left\| x_{1j} - c_1 \right\|^2}{N_1 - 1} \quad (1)$$

โดยที่ N_i คือ จำนวนจุดข้อมูลกลุ่มที่ i

x_{ij} คือ ข้อมูลกลุ่มที่ i จุดข้อมูลที่ j โดยที่ $i = 1,2,3,\ldots,k$ และ $j = 1,2,3,\ldots,n$

n คือ จำนวนข้อมูล

c_i คือ จุดศูนย์กลางของกลุ่มที่ i

ขั้นตอนวิธีทำ

1. กำหนด k จุดให้เป็นตัวแทนจุดศูนย์กลางกลุ่ม
2. กำหนดแต่ละจุดข้อมูลให้กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด
3. คำนวณจุดศูนย์กลางของแต่ละกลุ่มใหม่จากค่าเฉลี่ยของจุดข้อมูลทั้งหมดในกลุ่มนั้น
4. พิจารณาการเปลี่ยนกลุ่มของจุดข้อมูล หาก SSE ของกลุ่มอื่นน้อยกว่า SSE ของกลุ่มที่จุดข้อมูลอยู่ในปัจจุบัน ให้ทำการเปลี่ยนกลุ่มไปยังกลุ่มที่มีค่า SSE น้อย
5. ทำซ้ำขั้นตอนที่ 3 และ 4 จนกระทั่งจุดข้อมูลไม่มีการเปลี่ยนกลุ่ม หรือจุดศูนย์กลางในแต่ละกลุ่มจะไม่มีการเปลี่ยนแปลง

2.2 วิธีฟอร์กี้

กระบวนการของวิธีฟอร์กี้ [5] เป็นการแบ่งข้อมูลแบบเคมีนที่เป็นที่รู้จักกันอย่างแพร่หลายมากที่สุด โดยจะพิจารณาการกระจายข้อมูลแบบต่อเนื่อง ส่วนกระบวนการของวิธีลอยด์ [6] จะพิจารณาการกระจายข้อมูลแบบไม่ต่อเนื่อง ทั้งสองวิธีจะมีขั้นตอนเหมือนกันทุกประการ โดยพิจารณาที่ผลรวมความแปรปรวนสำหรับการแจกแจงแบบไม่ต่อเนื่องและการแจกแจงแบบต่อเนื่องดังสมการที่ (2) และ (3) ตามลำดับ และผลรวมความแปรปรวนจะใช้พิจารณาความแปรปรวนภายในกลุ่มแต่ละกลุ่ม กระบวนการนี้มีข้อดีคือใช้กับข้อมูลที่ทราบการแจกแจงอย่างชัดเจน แต่มีข้อเสียคือการแบ่งกลุ่มไม่มีประสิทธิภาพถ้าข้อมูลไม่เป็นรูปวงกลม

สำหรับการแจกแจงแบบไม่ต่อเนื่อง

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} d\left(c_i, x_{ij}\right) \quad (2)$$

สำหรับการแจกแจงแบบต่อเนื่อง

$$E = \sum_{i=1}^{k} \int f(x) d\left(c_i, x_{ij}\right) dx \quad (3)$$

โดยที่ k คือ จำนวนกลุ่ม

n คือ จำนวนข้อมูล

x_{ij} คือ ข้อมูลกลุ่มที่ i จุดข้อมูลที่ j โดยที่ $i = 1,2,3,\ldots,k$ และ $j = 1,2,3,\ldots,n$

c_i คือ จุดศูนย์กลางของกลุ่มที่ i

$f(x)$ คือ ฟังก์ชันความหนาแน่นความน่าจะเป็น

$d(c_i, x_{ij})$ คือ ฟังก์ชันระยะห่างระหว่างจุดข้อมูลและจุดศูนย์กลาง

ขั้นตอนวิธีทำ

1. กำหนด k จุดให้เป็นตัวแทนจุดศูนย์กลางกลุ่ม
2. กำหนดแต่ละจุดข้อมูลให้กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด
3. คำนวณจุดศูนย์กลางของแต่ละกลุ่มใหม่จากค่าเฉลี่ยของจุดข้อมูลทั้งหมดในกลุ่มนั้น
4. ถ้าจุดศูนย์กลางของกลุ่มตัวใดตัวหนึ่งมีการเปลี่ยนแปลง ให้ทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งค่าเฉลี่ยหรือจุดศูนย์กลางในแต่ละกลุ่มจะไม่เปลี่ยนแปลง

2.3 วิธีแม็คควีน

กระบวนการแม็คควีน [7] เป็นขั้นตอนแบบวนซ้ำและคล้ายกับกระบวนการของวิธีลอยด์หรือวิธีฟอร์กี้ แต่แตกต่างกับกระบวนการของวิธีลอยด์หรือวิธีฟอร์กี้คือจุดศูนย์กลางจะถูกปรับใหม่ทุกครั้งที่จุดข้อมูลเปลี่ยนกลุ่ม หากจุดศูนย์กลางของกลุ่มที่อยู่ในปัจจุบันนั้นใกล้เคียงที่สุดจะไม่มีการเปลี่ยนแปลงใดๆ แต่หากจุดศูนย์กลางของกลุ่มอื่นอยู่ใกล้ที่สุด จุดข้อมูลจะถูกจัดให้อยู่กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุดตัวนั้น และจะคำนวณจุดศูนย์กลางของทั้งสองกลุ่มใหม่เป็นค่าเฉลี่ยของจุดข้อมูล กระบวนการนี้มีข้อดีคือมีการปรับจุดศูนย์กลางบ่อยขึ้นทำให้มีประสิทธิภาพในการจัดกลุ่มมากขึ้น แต่มีข้อเสียคือใช้เวลาในการจัดกลุ่มมากขึ้น

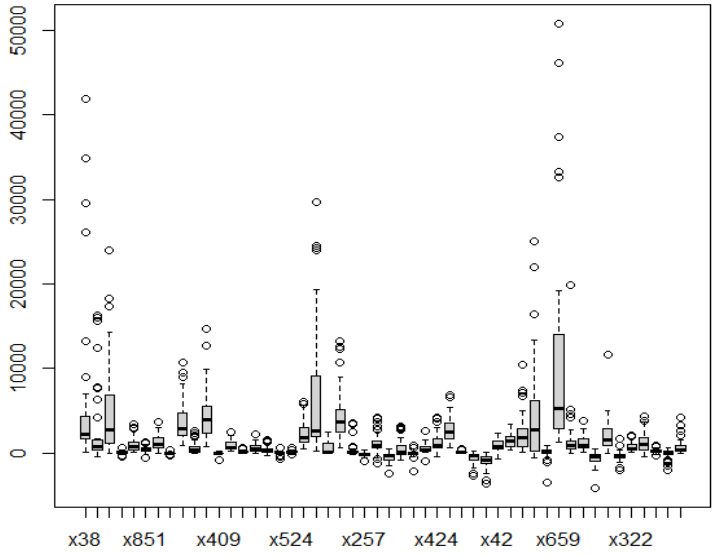
ขั้นตอนวิธีทำ

1. กำหนด k จุดให้เป็นจุดศูนย์กลางเริ่มต้น
2. กำหนดจุดข้อมูลให้กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด
3. ปรับตำแหน่งของจุดศูนย์กลางใหม่ทุกครั้งที่มีการเพิ่มจุดข้อมูลเข้าในการวิเคราะห์กลุ่ม
4. คำนวณจากค่าเฉลี่ยของจุดข้อมูลภายในกลุ่มนั้นๆ
5. ทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งเพิ่มจุดข้อมูลเข้าในการวิเคราะห์กลุ่มครบทั้งหมด

3. ผลการทดลองและวิจารณ์

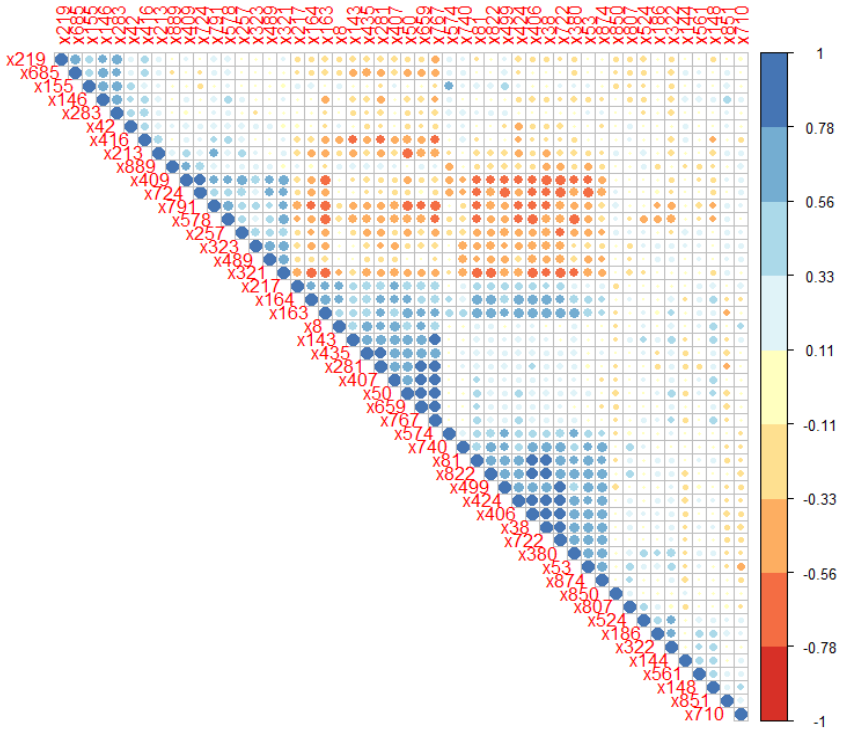
การวิจัยครั้งนี้มีขั้นตอนดังนี้ คือ

1) ศึกษาหาชุดข้อมูลที่มีมิติขั้นสูงเป็นข้อมูลของระดับการเป็นเนื้องอกในสมองและข้อมูลรหัสพันธุกรรมของผู้ป่วยเนื้องอกในสมองจาก http://www.ncbi.nlm.nih.gov/geo/ มีจำนวน 42 คน ตัวแปรอิสระ คือข้อมูลรหัสพันธุกรรมจำนวน 989 ชุด และตัวแปรตามคือ ระยะของการเป็นเนื้องอกในสมองจำนวน 5 กลุ่ม คือ Medulloblastoma (MD), Malignant Gliomas (MGlio), Normal Human Cerebella (Ncer), Primitive Neuroectodermal Tumors (PNET) และ Atypical Teratoid/Rhabdoid Tumors (Rhab) โดยได้สุ่มตัวอย่างรหัสพันธุกรรมด้วยวิธีสุ่มตัวอย่างแบบง่าย (Simple random sampling) จำนวน 50 ชุดเพื่อแสดงลักษณะข้อมูล ดังรูปที่ 1



รูปที่ 1. แผนภาพกล่องของรหัสพันธุกรรมจำนวน 50 ชุด

จากรูปที่ 1 แกนนอนแสดงรหัสของยีนที่สุ่ม และแกนตั้งแสดงลำดับของกรดอะมิโนในโปรตีน พบว่าข้อมูลรหัสพันธุกรรมมีลักษณะการกระจายไม่เท่ากันและข้อมูลมีค่านอกเกณฑ์ ซึ่งข้อมูลที่มีตัวแปรอิสระเป็นจำนวนมากมีโอกาสที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงนั้นมีมากหรือที่เรียกว่า การเกิดความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) จึงสุ่มตัวแปรอิสระข้อมูลรหัสพันธุกรรม 50 ชุด โดยหาความสัมพันธ์ระหว่างตัวแปรที่สุ่มมาแสดงดังรูปที่ 2



รูปที่ 2. กราฟแสดงความสัมพันธ์ระหว่างรหัสพันธุกรรมจำนวน 50 ชุด

จากรูปที่ 2 พบว่าข้อมูลรหัสพันธุกรรมที่สุ่มมานั้น มีสีแดงอ่อนและเข้มกระจายตัวอยู่เป็นจำนวนมากแสดงว่าข้อมูลชุดนี้มีความสัมพันธ์กันค่อนข้างสูง ซึ่งเรียกว่า เกิดความสัมพันธ์เชิงเส้นแบบพหุ

2) ทำการสุ่มรหัสพันธุกรรมจำนวน 200, 400, 600 และ 800 โดยแบ่งจำนวนกลุ่มคือ 5, 10, 15, 20, 25 และ 30

3) ใช้โปรแกรมอาร์ในการสุ่มข้อมูลและแบ่งกลุ่มแบบเคมีนจำนวน 3 วิธีคือ วิธีฮาร์ติกัน-หว่อง วิธีฟอร์กี้ และวิธีแม็คควีน

4) คำนวณหาประสิทธิภาพของการแบ่งกลุ่ม จากค่าร้อยละความแตกต่างของข้อมูลระหว่างกลุ่ม (RS) ดังนี้

$$RS = \frac{SS_t - SS_w}{SS_t} \times 100$$

เมื่อ

$$SS_t = \sum_{i=1}^{k}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}_j\right)^2$$

$$SS_w = \sum_{i=1}^{k}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}_{ij}\right)^2$$

โดยที่ SS_t คือ ผลรวมของผลต่างกำลังสองของข้อมูลทั้งหมด
SS_w คือ ผลรวมของผลต่างกำลังสองทุกข้อมูลภายในกลุ่ม
k คือ จำนวนกลุ่มที่แบ่งได้ทั้งหมด
p คือ จำนวนตัวแปรอิสระทั้งหมด
x_{ij} คือ ข้อมูลกลุ่มที่ i ตัวแปรอิสระที่ j
$\overline{x}_j$ คือ ค่าเฉลี่ยข้อมูลในตัวแปรอิสระที่ j
$\overline{x}_{ij}$ คือ ค่าเฉลี่ยข้อมูลในกลุ่มที่ i ตัวแปรอิสระที่ j

5) จากข้อ 3.2-3.4 ทำการสุ่มทั้งหมด 1,000 ครั้ง และนำค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (RS) ที่ได้แต่ละรอบมาหาค่าเฉลี่ยเรียกว่า ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่ม (Mean R – Squared: MRS) โดยคำนวณค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มวิธีฮาร์ติกัน-หว่อง (MRS_H) ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มวิธีฟอร์กี้ (MRS_F) และค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มวิธีแม็คควีน (MRS_M) ดังนี้

$$MRS_H = \frac{\sum_{i=1}^{1,000} RS_{H,i}}{1,000} \qquad MRS_F = \frac{\sum_{i=1}^{1,000} RS_{F,i}}{1,000} \quad \text{และ} \quad MRS_M = \frac{\sum_{i=1}^{1,000} RS_{M,i}}{1,000}$$

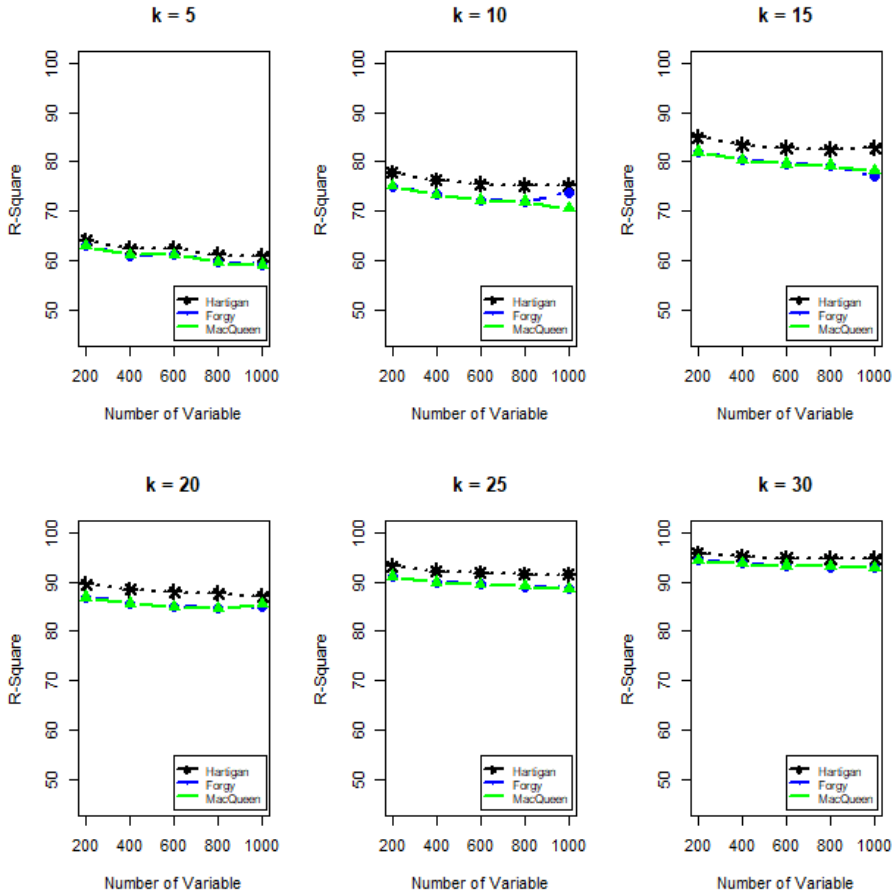
6) นำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มโดยกำหนดจำนวนกลุ่ม (k) เท่ากับ 5, 10, 15, 20, 25 และ 30 โดยสุ่มข้อมูลรหัสพันธุกรรมซึ่งกำหนดให้ตัวแปรอิสระ (p) 200, 400, 600 และ 800 จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่ม แสดงผลการทดลองดังตารางที่ 1

ตารางที่ 1. ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่ม (MRS) ของวิธีการแบ่งกลุ่มตามจำนวนตัวแปรอิสระ (p) และจำนวนกลุ่ม (k)

k	p	**วิธีฮาร์ติกัน-หว่อง**	**วิธีฟอร์กี้**	**วิธีแม็คควีน**
5	200	**64.0749**	63.0332	62.9707
	400	**62.3923**	61.1281	61.1417
	600	**62.4722**	61.2167	61.2793
	800	**61.1582**	59.6763	59.6048
10	200	**77.7752**	75.1185	75.1045
	400	**76.2357**	73.3283	73.2486
	600	**75.4833**	72.3344	72.2638
	800	**75.2343**	71.8860	71.9696
15	200	**84.9307**	82.0662	81.9848
	400	**83.4189**	80.3297	80.2976
	600	**82.7670**	79.5668	79.5397
	800	**82.4841**	79.2726	79.2309
20	200	**89.5393**	86.9708	86.9017
	400	**88.4551**	85.6757	85.6338
	600	**87.9444**	84.9563	84.9123
	800	**87.6772**	84.5806	84.5996
25	200	**93.0721**	91.1053	91.0828
	400	**92.1485**	89.9437	89.9339
	600	**91.7871**	89.4546	89.4876
	800	**91.5039**	89.0645	89.0767
30	200	**95.7340**	94.3855	94.3760
	400	**95.1530**	93.6954	93.7266
	600	**94.7908**	93.2530	93.2400
	800	**94.6431**	93.0193	93.0228

หมายเหตุ : ตัวหนา หมายถึงค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มสูงที่สุด

จากตารางที่ 1 พบว่าค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มในทุกค่าของตัวแปรอิสระ วิธีฮาร์ติกัน-หว่องให้ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มสูงที่สุด โดยเมื่อจำนวนตัวแปรอิสระเพิ่มมากขึ้นค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าลดลง แต่ถ้าจำนวนกลุ่มมีค่าเพิ่มมากขึ้นค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าเพิ่มมากขึ้น แสดงในรูปที่ 3



รูปที่ 3. ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มของจำนวนตัวแปรอิสระและจำนวนกลุ่ม

จากรูปที่ 3 พบว่าเมื่อจำนวนตัวแปรอิสระมีจำนวนเพิ่มมากขึ้น ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มมีแนวโน้มลดลง แสดงว่าการใช้จำนวนตัวแปรอิสระจำนวนมากไม่มีผลต่อการแบ่งกลุ่ม แต่เมื่อพิจารณาจากจำนวนกลุ่มพบว่าเมื่อใช้จำนวนกลุ่มที่มากขึ้นจะส่งผลให้ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มเพิ่มมากขึ้น

จากผลการวิจัยนี้แสดงว่าการใช้ข้อมูลรหัสพันธุกรรมในการจัดกลุ่มนั้น วิธีฮาร์ติกัน-ห่วงให้ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด และวิธีนี้ยังใช้เวลาในการแบ่งกลุ่มน้อยกว่าวิธีอื่น ๆ [17] และจำนวนกลุ่มที่เพิ่มขึ้นจะทำให้ประสิทธิภาพในการแบ่งกลุ่มเพิ่มมากขึ้น [12] นอกจากนี้ในการแบ่งกลุ่มไม่จำเป็นต้องใช้ขนาดของตัวแปรในการแบ่งกลุ่มเป็นจำนวนมากอีกด้วยเพราะให้ประสิทธิภาพในการแบ่งกลุ่มไม่แตกต่างกันมาก นอกจากวิธีที่ใช้แล้วผู้วิจัยควรพิจารณาความสัมพันธ์ของข้อมูล ความผิดปกติของข้อมูล และระยะเวลาในการแบ่งกลุ่มด้วย

4. สรุปผลการทดลอง

การศึกษาการเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของข้อมูลรหัสพันธุกรรมสำหรับข้อมูลที่มีมิติขั้นสูง โดยเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง วิธีฟอร์กี้ และวิธีแม็คควีน สำหรับข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองพบว่าวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่องให้ประสิทธิภาพดีที่สุดสำหรับการแบ่งกลุ่มทุกจำนวนกลุ่ม โดยให้ค่าเฉลี่ยความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด เนื่องจากวิธีฮาร์ติกัน-หว่องจะมีการปรับปรุงผลรวมของความเบี่ยงเบนกำลังสองภายในกลุ่มโดยมีการกำหนดค่าจุดศูนย์กลางใหม่ ทำให้เกิดประสิทธิภาพของการแบ่งกลุ่มมากที่สุด นอกจากนี้ยังพบว่าจำนวนตัวแปรอิสระที่เพิ่มขึ้นจะไม่ส่งผลต่อประสิทธิภาพการแบ่งกลุ่ม ดังนั้นเพื่อลดค่าใช้จ่าย และเวลาในการเก็บข้อมูล การแบ่งกลุ่มไม่จำเป็นต้องใช้จำนวนตัวแปรที่มากเนื่องจากจำนวนตัวแปรไม่ส่งผลต่อประสิทธิภาพการแบ่งกลุ่ม

เอกสารอ้างอิง (References)

[1] Zarikas, V., Poulopoulos, S.G., Gareiou, Z. and Zervas, E. 2020. Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*, 31, 1-8.

[2] Nurlaila, I., Irawati, W., Purwandari, K. and Pardamean, B. 2021. K-Means Clustering Model to Discriminate Copper-Resistant Bacteria as Bioremediation Agents. *Procedia Computer Science*, 179, 804-812.

[3] Shan, P. 2018. Image segmentation method based on K-mean algorithm. *EURASIP Journal on Image and Video Processing*, 81, 1-9.

[4] Hartigan, J.A. and Wong, M.A. 1979. Algorithm AS 136: A K-means Clustering Algorithm. *Applied Statistics*, 28, 100-108.

[5] MacQueen. J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, 281-297.

[6] Forgy, E.W. 1965. Clustering Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. *Biometrics*, 21, 768-769.

[7] Lloyd, S.P. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28, 128-137.

[8] Yadav, J. and Sharma, Monika. 2013. A Review of K-mean Algorithm. *International Journal of Engineering Trends and Technology*, 4(7), 2972-2976.

[9] Singh, R. P. and Rajpoot, D. S. 2019. Efficient Identification of Initial Clusters Centers for Partitioning Clustering Methods. 2019 Fifth International Conference on Image Processing, Shimla, India, 131-136.

[10] อาริกา ธรรมโน, มุทิตา หวังคิด และอาริต ธรรมโน. 2563. การพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง. *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, 10(2), 1-9. [Arika Thammmano, Muthita Wangkid and Arit Thammano, 2020. Breast Cancer Prediction Using K-mean Classification Algorithm with Self-adaptive Weight. *Journal of Information Science and Technology*, 10(2), 1-9. (in Thai)]

[11] Jothi, R., Mohanty, S. K. and Ojha, A. 2017. DK-means: a deterministic K-means clustering algorithm for gene expression analysis. *Pattern Analysis and Application*, 22, 649-667.

[12] Saadeh, H. Al Fayez, R. Q. and Elshqeirat, B. 2020. Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development. *International Journal of Machine Learning and Computing*, 10(3), 452-457.

[13] Joshi, R., Prasad, R., Mewada, P. and Saurabh. 2020. Modified LDA Approach For Cluster Based Gene Classification Using K-Mean Method. *Procedia Computer Science*, 171, 2493-2500.

[14] Bhatt, V., Dhakar, M. and Chaurasia, B. K. 2016. Filtered Clustering Based on Local Outlier Factor in Data Mining. *International Journal of Database Theory and Application*, 9(5), 275-282.

[15] Thakare, Y.S. and Bagal, S.B. 2015. Performance Evaluation of K-means Clustering Algorithm with Various Distance. *International Journal of Computer Application*, 110, 12-16.

[16] Meng, Y., Liang, J., Cao, F. and He, Y. 2018. A New distance with derivative information for functional k-means clustering. *Information Sciences*, 463-464, 166-185.

[17] Morissette, L. and Chartier. S. 2013. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1), 15-24.