การเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของรหัสพันธุกรรม สำหรับข้อมูลที่มีมิติขั้นสูง

COMPARING K-MEAN CLUSTERING METHOD OF GENE EXPRESSION ON HIGH-DIMENSIONAL DATA

จารวี พร้อมสง่า

JARAWEE PROMSANGA

การค้นคว้าอิสระเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติและการวิเคราะห์ธุรกิจ

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีเจ้าคุณทหารลาดกระบัง

พ.ศ.2565

KMITL-2022-SC-M-050-110

หัวข้อการค้นคว้าอิสระ	การเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของรหัสพันธุกรรมสำหรับข้อมูลที่มีมิติขั้นสูง
ชื่อนักศึกษา	จารวี พร้อมสง่า
รหัสประจำตัว	63605126
ปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติและการวิเคราะห์ธุรกิจ)
ภาควิชา	สถิติ
พ.ศ.	2565
อาจารย์ที่ปรึกษาค้นคว้าอิสระ	รองศาสตราจารย์ อัชฌา อระวีพร

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนที่มีประสิทธิภาพมากที่สุดสำหรับข้อมูลที่มีมิติขั้นสูง โดยใช้ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด โดยทำการศึกษาวิธีการแบ่งกลุ่มแบบเคมีน 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง วิธีฟอร์กี้ และ วิธีแม็คควีน และนำไปใช้ในการแบ่งกลุ่มชุดข้อมูลทั้ง 2 ชุดแล้ว เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของการแบ่งกลุ่มคือค่าความแตกต่างของข้อมูลระหว่างกลุ่ม จากผลการแบ่งกลุ่มแบบเคมีนทั้งหมด 3 วิธี พบว่าวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่องให้ประสิทธิภาพดีที่สุดสำหรับชุดข้อมูลทั้ง 2 ชุด ซึ่งให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุดเมื่อเปรียบเทียบกับวิธีฟอร์กี้และวิธีแม็คควีน

คำสำคัญ : การแบ่งกลุ่ม, การแบ่งกลุ่มแบบเคมีน, ฮาร์ติกัน-หว่อง, ฟอร์กี้, แม็คควีน

Independent Study Title	Comparing K-mean Clustering Method of Gene Expression On High-Dimensional Data
Student Name	Jarawee Promsanga
Student ID	63605126
Degree	Master of Science (Statistics and Business Analysis)
Department	Statistics
Year	2022
Independent Study Advisor	Assoc. Prof. Dr. Autcha Araveeporn

Abstract

This study aims to compare the performance of k-means clustering methods for high-dimensional data. The research uses a brain tumor and a lung cancer classification data set. For studying the k-means clustering methods, the Hartigan Wong, Forgy, and MacQueen methods are used to cluster the two data sets. The criterion used to compare the performance of clustering is the data differences between the groups. The results of k-means clustering methods found that the Hartigan Wong method had the best performance for both data sets. However, the Hartigan Wong method showed the most significant difference in data between groups compared to Forgy and MacQueen methods.

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

กิตติกรรมประกาศ

การค้นคว้าอิสระเล่มนี้สำเร็จได้ด้วยดี เนื่องจากได้รับความอนุเคราะห์จากอาจารย์ที่ปรึกษาการค้นคว้าอิสระ รศ.ดร.อัชฌา อระวีพร ที่ให้คำปรึกษา ให้คำแนะนำ และสละเวลาตรวจทานแก้ไขข้อบกพร่องต่างๆ ตลอดจนติดตามขั้นตอนการจัดทำรูปเล่มการค้นคว้าอิสระจนสำเร็จได้ด้วยดี ผู้วิจัยจึงขอกราบขอบพระคุณด้วยความเคารพเป็นอย่างสูง ณ โอกาสนี้

ขอกราบขอบพระคุณ ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์ และ ดร.ยุวดี กล่อมวิเศษ ผู้ซึ่งเป็นอาจารย์กรรมการ ที่เมตตาให้คำปรึกษา ตลอดจนแก้ไขข้อผิดพลาดและความไม่สมบูรณ์ของงานนี้ ทำให้การค้นคว้าอิสระเล่มนี้สมบูรณ์ยิ่งขึ้น และกราบขอบคุณคณาจารย์ประจำภาคสถิติประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังที่มอบวิชาความรู้และประสบการณ์การเรียนให้แก่ผู้วิจัย

สุดท้ายนี้ ขอขอบคุณบิดามารดา ครอบครัว และผู้ที่สนับสนุนการศึกษาในครั้งนี้ของผู้วิจัย ที่เป็นกำลังใจให้กับผู้วิจัยในการศึกษาระดับปริญญาโท ตลอดจนการจัดทำการค้นคว้าอิสระเล่มนี้จนสำเร็จลุล่วงไปได้ด้วยดี

จารวี พร้อมสง่า

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

ในปัจจุบันเทคโนโลยีทางการแพทย์มีความก้าวหน้ามากขึ้น โดยเฉพาะเทคโนโลยีในการถอดรหัสพันธุกรรมโดยนักวิทยาศาสตร์ นักวิจัย สามารถศึกษารหัสพันธุกรรม (Gene Expression) ซึ่งเป็นกระบวนการถ่ายทอดข้อมูลทางพันธกรรมหรือยีนเพื่อนำไปวินิจฉัยโรค และสามารถรักษาได้อย่างตรงจุด โดยข้อมูลของรหัสพันธุกรรมนั้นมีชนิดของยีนเป็นจำนวนมาก โดยยีนเหล่านี้จะแสดงลักษณะของโรค การแบ่งกลุ่ม (Cluster) จากข้อมูลของยีนบางส่วนที่มีความสำคัญมาใช้ในการจำแนกผู้ป่วยโรคต่าง ๆ นั้น สามารถวินิจฉัยและรักษาได้อย่างรวดเร็ว โดยเฉพาะ โรคมะเร็ง ถ้าทำการรักษาหรือจำแนกผู้ป่วยจะช่วยให้ผู้ป่วยสามารถรอดชีวิตได้

การวิเคราะห์การแบ่งกลุ่ม (Cluster Analysis) จะทำการแบ่งข้อมูลไปเป็นกลุ่ม ๆ เพื่อที่จะทำให้เข้าใจข้อมูลได้ดียิ่งขึ้น หรือใช้ในการค้นหาโครงสร้างในกลุ่มข้อมูล ขั้นตอนวิธีในการแบ่งกลุ่มข้อมูลจะแบ่งชุดข้อมูลออกเป็นกลุ่ม ๆ หรือคลาส (Class) โดยที่ข้อมูลที่มีความคล้ายกันจะถูกกำหนดให้ไปอยู่กลุ่มเดียวกัน ในทางตรงกันข้ามข้อมูลที่ไม่คล้ายกันหรือแบ่งแยกกันได้จะถูกกำหนดให้อยู่ในกลุ่มที่แตกต่างกัน เพื่อช่วยในการลดขนาดข้อมูล วิธีการในการแบ่งกลุ่มจะทำการคำนวณการวัดระยะห่างระหว่างข้อมูลหนึ่งกับข้อมูลหนึ่งด้วยวิธีการต่างๆ เช่น การวัดระยะห่างแบบยูคลิเดียน (Euclidean Distance) การวัดระยะห่างแบบแมนฮัตตัน (Manhattan Distance) และ การวัดระยะห่างแบบโคไซน์ (Cosine Distance) เป็นต้น ยกตัวอย่างงานวิจัยการแบ่งกลุ่ม ได้แก่ การแบ่งกลุ่มแบบฟัซซี่เพื่อเปรียบเทียบอัตราการกระจายของโควิด-19 ในประเทศที่มีความเสี่ยงสูง (Mahmoudi, Mohammad Reza, et al, 2020) โดยมีการเปรียบเทียบการกระจายตัวของโควิด-19 ในสหรัฐอเมริกา สเปน อิตาลี เยอรมนี สหราชอาณาจักร ฝรั่งเศส และอิหร่านโดยใช้เทคนิคการแบ่งกลุ่มแบบฟัซซี่ โดยใช้สหสัมพันธ์แบบเพียร์สันวัดความสัมพันธ์ระหว่างการแพร่กระจายของโควิด-19 กับขนาดของประชากร จากการศึกษากล่าวได้ว่าการแพร่กระจายของโควิด-19 ในสเปนและอิตาลีมีความคล้ายคลึงกันและแตกต่างจากประเทศอื่น ๆ

การแบ่งกลุ่มแบบเคมีนเป็นวิธีที่นิยมใช้ในการแบ่งกลุ่มข้อมูล โดยเปรียบเทียบความคล้ายคลึงของข้อมูล กับจุดศูนย์กลางหรือค่าเฉลี่ย (Mean) ของแต่ละกลุ่ม เป็นการแบ่งส่วน (Partitional Clustering) ด้วยการแบ่งข้อมูลออกเป็นส่วนตามจำนวนกลุ่มที่ระบุ ตัวอย่างงานวิจัยเกี่ยวกับการแบ่งกลุ่มแบบเคมีน เช่น งานวิจัยเรื่องแบบจำลองการแบ่งกลุ่มเคมีนเพื่อแยกแบคทีเรียที่

ทนต่อทองแดง : สารชีวบำบัดภัณฑ์ (Ika Nurlaila, Wahyu Irawati, et al, 2020) งานวิจัยนี้พบว่าแบบจำลองการแบ่งกลุ่มแบบเคมีนจะไม่ได้ดึงแบคทีเรียสกุลเดียวกันเข้าไปในกลุ่มเดียวกัน แต่แบบจำลองนี้จะรวบรวมความคล้ายคลึงกันในลักษณะการทำงานที่เรียกว่าความเข้มข้นต่ำสุดในการยับยั้ง (MIC) โดยไม่คำนึงถึงต้นกำเนิดและลำดับชั้นในการแบ่งกลุ่ม แสดงให้เห็นว่าแบบจำลองการแบ่งกลุ่มเคมีนมีความไวต่อการตรวจจับความแตกต่างของความเข้มข้นต่ำสุดในการยับยั้งมากกว่าความแตกต่างของสกุล

เนื่องจากข้อมูลทางด้านการแพทย์นั้นการเก็บตัวอย่างจะไม่สามารถเก็บได้เป็นจำนวนมาก โดยถ้าตัวแปรอิสระมีจำนวนมากกว่าขนาดตัวอย่างที่เก็บมา เรียกว่าเป็นข้อมูลที่มีมิติขั้นสูง (High-Dimensional Data) เช่น ข้อมูลไมโครอาร์เรย์ (Microarray Data) ซึ่งเป็นข้อมูลที่ได้จากการศึกษารูปแบบการแสดงออกของ ยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน โดยยีนที่ศึกษามีจำนวนเป็นหลักพันหรือหลักหมื่น สามารถนำมาใช้ในการจำแนกเนื้อเยื่อมะเร็งและเนื้อเยื่อปกติได้ ข้อมูลสำหรับการจำแนกประเภทเอกสาร (Text Classification) ก็เป็นข้อมูลที่มีจำนวนมิติมาก โดยตัวแปรที่เป็นไปได้ในการใช้จำแนกเอกสาร คือคำหรือวลีในเอกสาร ซึ่งมีจำนวนมากในแต่ละเอกสาร หากนำตัวแปรทั้งหมดที่มีจำนวนมากนี้ไปใช้เป็นตัวแปรทำนายสำหรับการจำแนกประเภทเอกสาร ก็อาจลดความแม่นยำในการจำแนกประเภทได้

ทั้งนี้การแบ่งกลุ่มข้อมูลมีหลายวิธี และการแบ่งกลุ่มข้อมูลยังช่วยลดการวิเคราะห์ข้อมูลที่มีปริมาณมาก ๆ ช่วยเพิ่มความเร็วและประสิทธิภาพในการวิเคราะห์ข้อมูลได้เป็นอย่างดี ยกตัวอย่างงานวิจัย เช่น งานวิจัยเรื่องการพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง (อาริกา ธรรมโน, มุทิตา หวังคิด และอาริต ธรรมโน, 2563) งานวิจัยนี้ได้นำเสนออัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง รวมทั้งทำการพัฒนาโปรแกรมพยากรณ์โรคมะเร็งเต้านมด้วยภาษาไพทอน ซึ่งถูกพัฒนาต่อยอดมาจากอัลกอริทึมการแบ่งกลุ่มแบบเคมีนให้มีความสามารถในการจำแนกประเภทและปรับเปลี่ยนค่าถ่วงน้ำหนักของคุณลักษณะเด่นในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทแต่ละประเภทให้เหมาะสมได้ด้วยตัวเองในระหว่างการเรียนรู้ชุดข้อมูล

งานวิจัยเรื่องการระบุกลุ่มและรูปแบบอุบัติเหตุทางถนนที่เกี่ยวข้องกับคนเดินเท้าและนักปั่นจักรยาน กรณีศึกษาจังหวัดเบรสเซีย : อิตาลี (Michela Bonera, Riccardo Mutti, et al, 2022) งานวิจัยนี้ได้นำเทคนิคการแบ่งกลุ่มมาใช้ในการวิเคราะห์ความปลอดภัยทางถนนเพื่อตรวจสอบรูปแบบการเกิดอุบัติเหตุ โดยใช้การวิเคราะห์กลุ่มแฝง (Latent Class Analysis) และการแบ่งกลุ่มแบบเคมีน (K-Means Clustering) โดยใช้ข้อมูลการเกิดอุบัติเหตุที่รวบรวมไว้ในช่วงปีพ.ศ. 2557 ถึงปี

พ.ศ. 2561 ซึ่งมีการระบุกลุ่มและรูปแบบอุบัติเหตุบนท้องถนนสำหรับคนเดินเท้า 3 กลุ่ม และระบุกลุ่มและรูปแบบอุบัติเหตุบนท้องถนนสำหรับนักปั่นจักรยาน 5 กลุ่ม เพื่อนำไปเป็นแนวทางในการแนะนำวิธีแก้ไขปัญหาเพื่อลดอุบัติเหตุและอาจเป็นเครื่องมือสนับสนุนในการตัดสินใจสำหรับผู้บริหารสาธารณะสำหรับการปรับปรุงความปลอดภัยทางถนนในอนาคต

บทความวิจัยเรื่องการเปรียบเทียบวิธีการจัดกลุ่มสำหรับข้อมูลที่มีการแจกแจงปกติแบบผสม (จิรวรรณ ไพบูลย์วรชาติ และนัท กุลวานิช, 2558) โดยศึกษาการเปรียบเทียบวิธีการแบ่งกลุ่ม 4 วิธี ได้แก่ 1) วิธีการแบ่งกลุ่มแบบวอร์ด 2) วิธีการแบ่งกลุ่มแบบเคมีน 3) วิธีการแบ่งกลุ่มแบบฟัซซี่ซีมีน และ 4) วิธีการแบ่งกลุ่มแบบอัลกอริทึม EM โดยจำลองข้อมูลที่มีการแจกแจงปกติแบบผสมและเปรียบเทียบประสิทธิภาพ 2 วิธี ได้แก่ วิธีการวัดค่าความแตกต่างของข้อมูลภายในกลุ่ม (RMSSTD) และวิธีการวัดค่าความต่างของข้อมูลระหว่างกลุ่ม (R-Square) ผลการวิจัยพบว่าวิธีการแบ่งกลุ่มแบบเคมีนมีประสิทธิภาพของการแบ่งกลุ่มดีที่สุด เนื่องจากทั้งค่าเฉลี่ย RMSSTD และค่าเฉลี่ย R-Square มีประสิทธิภาพดีที่สุดและเป็นไปในทิศทางเดียวกัน

ตัวอย่างงานวิจัยทางการแพทย์ เช่น งานวิจัยเรื่องการใช้ MLDA ที่ปรับปรุงแล้วสำหรับการจำแนกยีนตามการแบ่งกลุ่มด้วยวิธีเคมีน (Reva Joshi, Ritu Prasad, et al. ,2020) งานวิจัยนี้นำเสนอ A Modified Latent Dirichlet Allocation (MLDA) ซึ่งพัฒนามาจาก Latent Dirichlet Allocation (LDA) งานวิจัยนี้ได้นำ MLDA มาใช้สำหรับการจำแนกยีนโดย MLDA ที่กล่าวถึงนี้จะระบุและจัดกลุ่มยีนที่แสดงความแตกต่างระหว่างเนื้อเยื่อปกติและเนื้อเยื่อมะเร็งประเภทต่างๆ

งานวิจัยเรื่องการวินิจฉัยโรคในใบองุ่นโดยใช้การจัดกลุ่มแบบเคมีนและการเรียนรู้ของเครื่อง (Seyed Mohamad Javidan, et al., 2022) งานวิจัยนี้ได้นำ Support Vector Machine มาใช้เพื่อวินิจฉัยและจำแนกโรคใบองุ่น เช่น หัดดำ ใบไหม้ เป็นต้น อีกทั้งยังนำการแบ่งกลุ่มแบบเคมีนมาใช้ในการแบ่งส่วนของใบองุ่นระหว่างส่วนที่เป็นโรคและส่วนที่ไม่เป็นโรคอีกด้วย

เนื่องจากการแบ่งกลุ่มแบบเคมีนเป็นที่นิยมสำหรับการแพทย์ ดังนั้นผู้วิจัยจึงทำการเปรียบเทียบประสิทธิภาพการแบ่งกลุ่มข้อมูลแบบเคมีนจำนวน 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง (Hartigan-Wong), วิธีฟอร์กี้ (Forgy) และวิธีแม็คควีน (MacQueen)

1.2 วัตถุประสงค์ของงานวิจัย

1) ศึกษาวิธีการแบ่งกลุ่มข้อมูลแบบเคมีนของการเป็นเนื้องอกในสมองและมะเร็งปอดสำหรับข้อมูลที่มีมิติขั้นสูง

2) เพื่อเปรียบเทียบผลการวิเคราะห์วิธีการแบ่งกลุ่มแบบเคมีนของการเป็นเนื้องอกในสมองและมะเร็งปอดสำหรับข้อมูลที่มีมิติขั้นสูง

1.3 ขอบเขตของปัญหา

1) ข้อมูลที่นำมาวิเคราะห์ในโครงงานนี้มี 2 ชุดข้อมูล คือ ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด

- ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมอง ได้ข้อมูลจากผู้ป่วยจำนวน 42 คน มีตัวแปรอิสระ คือ รหัสพันธุกรรมของคนไข้ จำนวน 989 รหัส และตัวแปรตาม คือ ระยะของการเป็นเนื้องอกในสมองจำนวน 5 กลุ่ม คือ medulloblastoma (MD), malignant gliomas (MGlio), normal human cerebella (Ncer), primitive neuroectodermal tumors (PNET) และ atypical teratoid/rhabdoid tumors (Rhab)
- ชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด ได้ข้อมูลจากผู้ป่วยจำนวน 197 คน มีตัวแปรอิสระ คือ รหัสพันธุกรรมของคนไข้ จำนวน 989 รหัส และตัวแปรตาม คือ ประเภทของโรคมะเร็งปอดจำนวน 4 กลุ่ม คือ Small Cell Lung Cancer (SCLC) : Oat cell lung cancer, Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma, Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma, Non-Small Cell lung cancer (NSCLC) Large Cell lung cancer

2) โครงงานนี้ทำการเปรียบเทียบประสิทธิภาพการแบ่งกลุ่มแบบเคมีน (K-Means Clustering) จำนวน 3 วิธี ได้แก่

- วิธีฮาร์ติกัน-หว่อง (Hartigan-Wong)
- วิธีฟอร์กี้ (Forgy)
- วิธีแม็คควีน (MacQueen)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้วิธีการที่มีประสิทธิภาพสำหรับการแบ่งกลุ่มข้อมูลของการเป็นเนื้องอกในสมองและมะเร็งปอดสำหรับข้อมูลที่มีมิติขั้นสูง

2) เป็นแนวทางในการเลือกใช้วิธีการแบ่งกลุ่มข้อมูลที่มีมิติขั้นสูงให้กับผู้วิจัยท่านอื่นในอนาคต

3) สามารถนำวิธีการแบ่งกลุ่มข้อมูลที่ได้จากการศึกษาครั้งนี้ไปประยุกต์ใช้ในโรคต่างๆ

1.5 ขั้นตอนในการดำเนินการ

1) กำหนดหัวข้อและกำหนดปัญหาโครงงาน

2) เลือกชุดข้อมูลที่ใช้ในการวิเคราะห์การแบ่งกลุ่ม

3) ศึกษาและค้นคว้างานวิจัยที่เกี่ยวข้อง

4) ศึกษาวิธีการแบ่งกลุ่มข้อมูลแต่ละแบบ

5) ดำเนินการแบ่งกลุ่มข้อมูลด้วยวิธีทั้ง 3 วิธี

6) วิเคราะห์ผลการดำเนินการ

7) สรุป อภิปรายผลและข้อเสนอแนะ

8) นำเสนอโครงงาน

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาเรื่องการเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของข้อมูลรหัสพันธุกรรม สำหรับข้อมูลที่มีมิติขั้นสูง ผู้ศึกษาได้ค้นคว้าแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องเพื่อนำมาเป็นแนวทางในการกำหนดกรอบแนวคิดในการศึกษา ดังนี้

1) ข้อมูลทั่วไปเกี่ยวกับเนื้องอกในสมอง
2) ข้อมูลทั่วไปเกี่ยวกับมะเร็งปอด
3) ข้อมูลที่มีมิติขั้นสูง (High Dimensional Data)
4) การวิเคราะห์การแบ่งกลุ่ม (Cluster Analysis)
5) งานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลทั่วไปเกี่ยวกับเนื้องอกในสมอง

2.1.1 ความหมายของเนื้องอกสมอง

เนื้องอกสมอง หมายถึง เนื้องอกที่เกิดจากการเจริญเติบโตของเซลล์ที่ผิดปกติในสมองมี 2 ชนิด คือ เนื้องอกในสมองปฐมภูมิ (Primary brain tumor) เกิดจากการเจริญเติบโตผิดปกติของเซลล์ประสาทในสมอง ตลอดจนความผิดปกติที่มีต้นกำเนิดจากเซลล์ภายในระบบประสาทเอง และเนื้องอกสมองทุติยภูมิหรือระยะลุกลาม (Metastasis brain tumor) เกิดจากเซลล์ที่ต้นกำเนิดจากอวัยวะส่วนอื่นของร่างกายภายนอกสมองและมีการแพร่กระจายไปที่สมอง

2.1.2 ประเภทของเนื้องอกสมอง

สมาคมโรคเนื้องอกในสมองของประเทศอเมริกาได้แบ่งประเภทเนื้องอกในสมอง ได้หลายแบบ เช่น ตำแหน่งที่เกิด ชนิดของเซลล์ที่ผิดปกติ หรือถ้าแบ่งตามลักษณะของเนื้องอกที่เกิดที่เนื้อสมองหรือกระจายมาจากอวัยวะอื่นสามารถแบ่งได้ดังนี้

1) เนื้องอกที่เป็นเนื้อธรรมดา (Benign Brain Tumors) เป็นเนื้องอกไม่อันตราย (ระดับ 1-2) มีการเจริญเติบโตช้า ไม่ใช่เซลล์มะเร็ง สามารถรักษาให้หายได้ และมีโอกาสน้อยที่ผู้ป่วยจะกลับมาเป็นอีกหลังการรักษา

2) เนื้อเยื่อที่เป็นเนื้อร้าย (Malignant Brain Tumors) เป็นเนื้องอกอันตราย (ระดับ 3-4) มีการเจริญเติบโตของเซลล์ที่ผิดปกติ คือ เซลล์มะเร็ง อาจเกิดขึ้นบริเวณสมองหรือเซลล์มะเร็งเกิดขึ้นที่

อวัยวะอื่นแล้วลามเข้าสู่สมองเนื้องอกที่เป็นเซลล์มะเร็งจะมีการเจริญเติบโตเรื่อยๆอย่างไม่สามารถควบคุมได้ และมีโอกาสที่จะกลับมาเป็นได้อีกแม้เคยผ่านการรักษาไปแล้ว โดยทั่วไปเนื้องอกที่เป็นเซลล์มะเร็งจะพบได้บ่อยกว่าเนื้องอกที่เป็นเนื้อธรรมดา นอกจากนี้ เนื้องอกในสมองมักมีชื่อเรียกแยกตามชนิดของเซลล์ที่เป็นเนื้องอก เช่น เนื้องอกไกลโอมา (Gliomas) เนื้องอกเยื่อหุ้มสมอง (Meningiomas) เนื้องอกเส้นประสาทหู (Acoustic Neuromas) เนื้องอกที่ต่อมใต้สมอง (Pituitary Adenomas และ Craniopharyngiomas) เนื้องอกนิวโรเอ็กโตเติร์ม (PNETs - Primitive Neuroectodermal Tumors) เนื้องอกเจิร์มเซลล์ (Germ Cell Tumors) และเนื้องอกสมองในเด็ก (Medulloblastomas) ที่ส่วนใหญ่เป็นเนื้อร้ายเกิดบริเวณสมองส่วนหลังแล้วแพร่กระจายผ่านทางน้ำหล่อเลี้ยงไขสันหลัง มักพบมากในเด็ก แต่ก็สามารถเกิดในผู้ใหญ่ได้เช่นกัน (วรรณวิศา ปะเสทะกัง และคณะ, 2563)

2.2 ข้อมูลทั่วไปเกี่ยวกับมะเร็งปอด

2.2.1 ความหมายของมะเร็งปอด

มะเร็งปอด (Lung Cancer) คือ เซลล์ส่วนใดส่วนหนึ่งภายในปอดที่มีความผิดปกติอย่างรวดเร็วจนไม่สามารถควบคุมได้ และเกิดเป็นก้อนเนื้อร้ายในที่สุด โดยเนื้อร้ายนี้สามารถลุกลามและกระจายไปอวัยวะอื่น ๆ ได้ ซึ่งส่วนใหญ่จะไม่แสดงอาการในระยะเริ่มแรก กว่าผู้ป่วยจะรู้ตัวก็มักจะอยู่ในระยะโรคที่รุนแรงแล้ว โดยปัจจุบันยังไม่สามารถบอกถึงสาเหตุของมะเร็งปอดได้ชัดเจนนัก แต่ปัจจัยที่เพิ่มความเสี่ยงของโรคมะเร็งปอด ได้แก่ การสูบบุหรี่ สารพิษและมลภาวะในสิ่งแวดล้อม อายุ พันธุกรรม เป็นต้น

2.2.2 ประเภทของมะเร็งปอด

โรคมะเร็งปอดแตกต่างกันตามชนิดของเซลล์มะเร็ง ซึ่งการแบ่งชนิดของโรคมะเร็งปอดโดยทั่วไป มักใช้ WHO histological classification ออกเป็นชนิดย่อย ได้แก่ adenocarcinoma, squamous cell carcinoma, adenosquamous carcinoma, large cell carcinoma, bronchoalveolar cell carcinoma small cell carcinoma, sarcomatoid carcinoma และ carcinoid tumor เป็นต้น ซึ่งสามารถแบ่งออกเป็น 2 ชนิดใหญ่ตามลักษณะพยาธิวิทยา ได้แก่

1) Non small cell lung cancer (NSCLC) ซึ่งพบประมาณร้อยละ 85 ของโรคมะเร็งปอดทั้งหมด โดยเจริญเติบโตช้ากว่าและแพร่กระจายช้ากว่า ที่พบบ่อย ได้แก่

1.1 Squamous cell carcinoma เป็นมะเร็งเยื่อบุผิวของหลอดลมที่มี intercellular bridge และมีการสร้าง keratin มะเร็งชนิดนี้มีความสัมพันธ์กับการสูบบุหรี่เป็นอย่างมาก ตำแหน่งที่

เกิดก้อนมะเร็งพบบ่อยที่สุดบริเวณหลอดลมส่วนต้น ภาพถ่ายรังสีทรวงอก (CXR) มักพบก้อนบริเวณตรงกลางทรวงอก ก้อนมะเร็งมีทั้งแบบที่ยื่นเป็นก้อนเข้าไปอุดตันท่อหลอดลม และแบบที่ลงลึกขยายเข้าไปในเนื้อเยื่อรอบข้างเนื้อมะเร็งมักมีสีขาวและอาจปนด้วยสีดำของ carbon pigment

1.2 Adenocarcinoma เป็นมะเร็งปอดที่เซลล์มีการจัดเรียง ตัวเป็น gland หรือมีการสร้างสาร mucin เป็นชนิดที่พบบ่อยที่สุดในโรคมะเร็งปอด มีชนิดย่อยที่รู้จักกันดี คือ bronchoalveolar cell carcinoma สัมพันธ์กับการสูบบุหรี่แต่น้อยกว่าชนิด Squamous cell carcinoma นอกจากนี้ยังพบว่ามีความสัมพันธ์กับแผลเป็นที่เกิดขึ้นมาก่อนในปอด ตำแหน่งที่เกิดก้อนมะเร็งพบบ่อยที่สุดบริเวณชายปอด ภาพถ่ายรังสีทรวงอกมักพบก้อนบริเวณตรงชายปอดเป็นส่วนใหญ่

1.3 Large cell carcinoma เป็นเซลล์มะเร็งที่มีขนาดใหญ่และไม่มีพัฒนารูปร่างไปเหมือนกับเซลล์ชนิดอื่น การเกิดมะเร็งสัมพันธ์กับการสูบบุหรี่ ตำแหน่งที่เกิดก้อนมะเร็งพบที่ส่วนกลางหรือบริเวณชายปอดและมักลุกลามไปยังเยื่อหุ้มปอด ผนังหน้าอกและรอบข้าง ตัวก้อนมะเร็งมีสีขาวปนชมพูและพบเนื้อเยื่อตายตรงส่วนกลางก้อนได้บ่อย

1.4 Adenosquamous carcinoma เป็นเซลล์มะเร็งที่มีลักษณะก้ำกึ่งระหว่าง Adenocarcinoma และ Squamous cell carcinoma การย้อมพบการติดสีเข้าได้กับมะเร็งทั้งสองชนิด

2) Small cell lung cancer (SCLC) ซึ่งพบประมาณร้อยละ 15 จะเจริญเติบโตเร็วและแพร่กระจายอย่างรวดเร็ว เมื่อตรวจพบ โรคมักลุกลามเข้าต่อมน้ำเหลือง และแพร่กระจายเข้าสู่กระแสเลือดแล้วจึงทำให้ผู้ป่วยเสียชีวิตได้อย่างรวดเร็ว ประกอบไปด้วยเซลล์ขนาดเล็ก ซึ่งเกิดจากเซลล์ Neuroendocrine cell ที่สามารถสร้างฮอร์โมนและสารเคมีต่างๆ ได้หลายชนิด (ปาริชาติ นิยมทอง, 2561)

2.3 ข้อมูลที่มีมิติขั้นสูง (High Dimensional Data)

เนื่องจากความก้าวหน้าทางเทคโนโลยีในปัจจุบันช่วยให้สามารถจัดเก็บข้อมูลขนาดใหญ่ในรูปแบบ อิเล็กทรอนิกส์ได้อย่างมีประสิทธิภาพ โดยมีการจัดเก็บข้อมูลในด้านต่างๆ ซึ่งข้อมูลบางประเภทมีตัวแปรจำนวนมากและมีความซับซ้อน ซึ่งเรียกข้อมูลที่มีตัวแปรอิสระจำนวนมากและมีจำนวนตัวแปรอิสระ (p) มากกว่าขนาดตัวอย่าง (n) หรือ $p > n$ ว่าข้อมูลมิติสูง (High dimensional data)

กรณีที่ข้อมูลมีมิติสูง ซึ่งมีจำนวนตัวแปรอิสระจำนวนมหาศาลนั้นมีโอกาสที่ทำให้ตัวแปรอิสระเพียงบางส่วนมีความสัมพันธ์กับตัวแปรตามและตัวแปรอิสระส่วนใหญ่ไม่มีความสัมพันธ์กับตัว

แปรตาม หรือเมื่อตัวแปรอิสระมีจำนวนมาก ทำให้โอกาสที่ตัวแปรอิสระจะมีความสัมพันธ์เชิงเส้นสูงนั้นมีมาก หรือที่เรียกว่า การเกิดความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) (พัชราภรณ์ พรดำเนินสวัสดิ์, 2560)

2.4 การวิเคราะห์การแบ่งกลุ่ม (Cluster Analysis)

การวิเคราะห์กลุ่ม (Cluster Analysis) เป็นเทคนิคในการแบ่งกลุ่มหน่วยข้อมูล หรือเป็นการแบ่งคน สัตว์ สิ่งของ องค์กร ฯลฯ ออกเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม โดยมีหลักเกณฑ์ในการแบ่งดังนี้ "ให้หน่วยที่อยู่ในกลุ่มเดียวกันมีลักษณะที่สนใจเหมือนกันหรือคล้ายกัน แต่หน่วยที่อยู่ต่างกลุ่มกันจะมีลักษณะที่สนใจต่างกัน" ดังรูปที่ 2.1

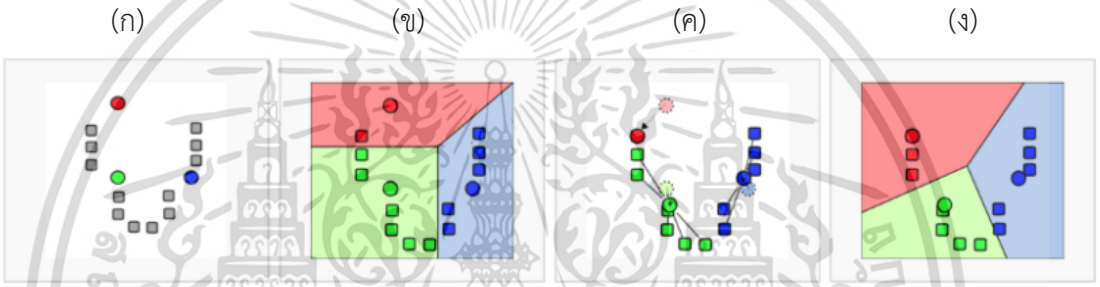
รูปที่ 2.1 แผนภาพการแบ่งกลุ่มข้อมูล

การแบ่งกลุ่มข้อมูล (Clustering) เมื่อฐานข้อมูลมีขนาดใหญ่ขึ้น และมีความหลากหลายมากขึ้น จะมีความจำเป็นต้องแบ่งข้อมูลออกเป็นส่วนๆ โดยข้อมูลในแต่ละส่วนมีความเกี่ยวพันกันอย่างเหมาะสม ก่อนที่จะใช้วิธีปฏิบัติการอื่นต่อไป

การแบ่งกลุ่มแบ่งออกเป็น 2 ประเภท คือ การแบ่งกลุ่มแบบลำดับขั้น (Hierarchical clustering) และการแบ่งกลุ่มแบบเคมีน (K-means clustering) แต่เนื่องจากโครงงานนี้สนใจที่จะทำการแบ่งกลุ่มข้อมูลแบบเคมีน เนื่องจากการแบ่งกลุ่มแบบเคมีนเป็นที่นิยมมาก ดังนั้นในหัวข้อนี้จะกล่าวถึงการแบ่งกลุ่มแบบเคมีนเท่านั้น (กัลยา วานิชย์บัญชา, 2552)

2.4.1 การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means clustering Analysis)

การแบ่งกลุ่มแบบเคมีน (K-Means Clustering) หรือ การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) คือ อัลกอริทึมชนิดหนึ่งที่เป็นเทคนิคในการตัดแบ่ง (Partition) วัตถุออก เป็น K กลุ่ม ซึ่งในแต่ละกลุ่มแทนค่าด้วยค่าเฉลี่ยของกลุ่ม และเป็นจุดศูนย์กลาง (Centroid) ของกลุ่มที่ใช้ในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยเลือกกลุ่มที่แบ่งนั้นจะมีระยะห่างจากค่ากลางของกลุ่มที่น้อยที่สุด แล้วทำการประมวลหาค่ากลางของกลุ่มใหม่ ทำเช่นนี้จนกระทั่งคำกลางของกลุ่มเปลี่ยนแปลงน้อยกว่าค่าที่กำหนดไว้ หรือครบจำนวนรอบที่กำหนดไว้ ดังรูปที่ 2.2



รูปที่ 2.2 แผนภาพการแบ่งกลุ่มข้อมูลแบบเคมีน

จากรูปที่ 2.2 (ก) แสดงการกำหนดจำนวนกลุ่มและกำหนดตำแหน่งตัวแทนกลุ่ม (ข) แสดงการแบ่งกลุ่มโดยพิจารณาจากระยะห่างระหว่างจุดข้อมูลและจุดศูนย์กลางที่ใกล้กับจุดข้อมูลนั้น ๆ มากที่สุด (ค) แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูลแต่ละกลุ่ม (ง) แสดงการเปลี่ยนกลุ่มของจุดข้อมูล (วณิชา แผลงรักษา และ นิเวศ จิระวิชิตชัย, 2562)

โดยการแบ่งกลุ่มข้อมูลแบบเคมีนแบ่งออกเป็น 4 วิธี ดังนี้

2.4.1.1 วิธีฮาร์ติกัน-หว่อง (Hartigan-Wong)

อัลกอริทึมฮาร์ติกัน-หว่องจะปรับจุดศูนย์กลางด้วยการพิจารณาจากแต่ละจุดข้อมูล โดยเริ่มแรกจะกำหนดจุดข้อมูลทั้งหมดให้กับจุดศูนย์กลางแบบสุ่ม จากนั้นจะปรับจุดศูนย์กลาง (Centroid) โดยพิจารณาจากแต่ละจุดข้อมูล และทำการตัดแบ่ง (Partition) ข้อมูลด้วยผลรวมกำลังสองความคลาดเคลื่อน (Sum Of Squares Of Error : SSE) ภายในกลุ่ม จากนั้นจุดข้อมูลต่างๆจะถูก

กำหนดให้กับจุดศูนย์กลางที่ใกล้ที่สุด และจุดศูนย์กลางจะถูกคำนวณใหม่เป็นค่าเฉลี่ยของจุดข้อมูลที่กำหนด (Laurence Morissette and Sylvain Chartier, 2013)

ในการพิจารณาแต่ละจุดข้อมูล หากผลรวมกำลังสองความคลาดเคลื่อนของกลุ่มอื่นมีค่าน้อยกว่าผลรวมกำลังสองความคลาดเคลื่อนของกลุ่มปัจจุบัน ดังสมการ (1) จุดข้อมูลนั้นจะถูกส่งไปยังกลุ่มอื่น และจะทำการวนซ้ำไปจนกว่าจะไม่มีการเปลี่ยนกลุ่ม

$$SSE_2 = \frac{N_i \sum_j \left\| x_{ij} - c_i \right\|^2}{N_i - 1} < SSE_1 = \frac{N_1 \sum_j \left\| x_{1j} - c_1 \right\|^2}{N_1 - 1} \quad (1)$$

โดยที่ N_i คือ จำนวนจุดข้อมูลกลุ่มที่ i

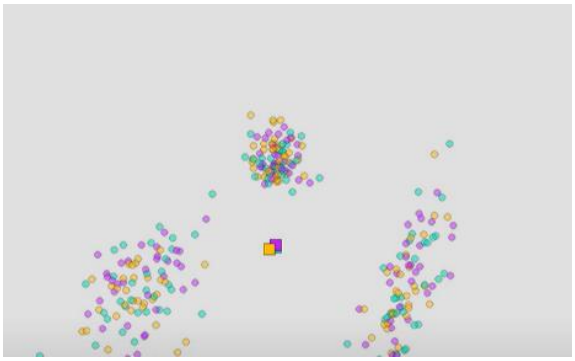
x_{ij} คือ ข้อมูลกลุ่มที่ i จุดข้อมูลที่ j โดยที่ i = 1,2,3,...,k และ j = 1,2,3,...,n

c_i คือ จุดศูนย์กลางของกลุ่มที่ i

<u>ขั้นตอนวิธีทำ</u>

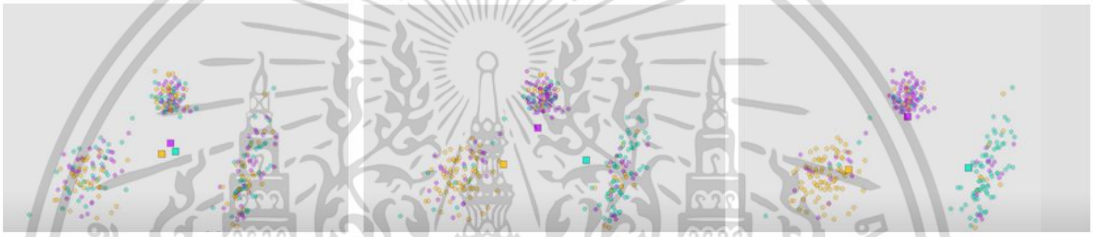
1. กำหนด k จุดให้เป็นตัวแทนจุดศูนย์กลางกลุ่ม
2. กำหนดแต่ละจุดข้อมูลให้กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด
3. คำนวณจุดศูนย์กลางของแต่ละกลุ่มใหม่จากค่าเฉลี่ยของจุดข้อมูลทั้งหมดในกลุ่มนั้น
4. พิจารณาการเปลี่ยนกลุ่มของจุดข้อมูล หาก SSE ของกลุ่มอื่นน้อยกว่า SSE ของกลุ่มที่จุดข้อมูลอยู่ในปัจจุบันให้ทำการเปลี่ยนกลุ่มไปยังกลุ่มที่มีค่า SSE น้อย
5. ทำซ้ำขั้นตอนที่ 3 และ 4 จนกระทั่งจุดข้อมูลไม่มีการเปลี่ยนกลุ่มหรือจุดศูนย์กลางในแต่ละกลุ่มจะไม่เปลี่ยนแปลง
6. สิ้นสุดการแบ่งกลุ่ม

ตัวอย่างการแบ่งกลุ่มวิธีฮาร์ติกัน ดังรูปที่ 2.3-2.5



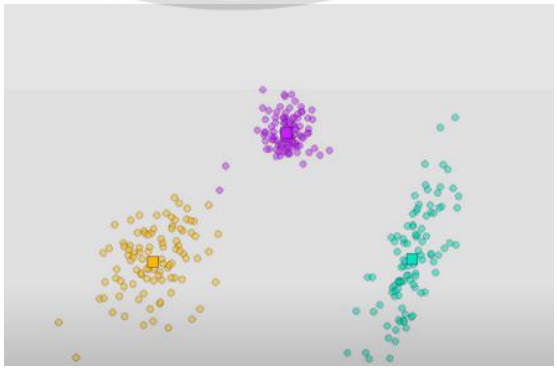
รูปที่ 2.3 แสดงการกำหนดจำนวนกลุ่มข้อมูลและกำหนดตำแหน่งจุดศูนย์กลางเริ่มต้น

(ก) (ข) (ค)



รูปที่ 2.4 แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูลแต่ละกลุ่ม และแสดงการย้ายกลุ่มของจุดข้อมูล โดยที่การย้ายกลุ่มของจุดข้อมูลจะพิจารณาจากค่า SSE ภายในกลุ่มนั้นๆ

จากรูปที่ 2.4 (ก) แสดงการปรับตำแหน่งจุดศูนย์กลางและการย้ายกลุ่มของจุดข้อมูล ครั้งที่ 1 (ข) แสดงการปรับตำแหน่งจุดศูนย์กลางและการย้ายกลุ่มของจุดข้อมูล ครั้งที่ 2 (ค) แสดงการปรับตำแหน่งจุดศูนย์กลางและการย้ายกลุ่มของจุดข้อมูล ครั้งที่ 3 และรูปที่ 2.5 แสดงการแบ่งกลุ่มเสร็จสิ้น



รูปที่ 2.5 แสดงการแบ่งกลุ่มเสร็จสิ้น

2.4.1.2 วิธีลอยด์ (Lloyd) หรือวิธีฟอร์กี้ (Forgy)

ลอยด์เป็นอัลกอริทึมของการแบ่งข้อมูลแบบเคมีนที่เป็นที่รู้จักกันอย่างแพร่หลายมากที่สุด เนื่องจากเป็นอัลกอริทึมที่เข้าใจง่ายและนำไปใช้ได้ง่าย ความแตกต่างระหว่างอัลกอริธึมลอยด์และอัลกอริธึมฟอร์กี้ คือ อัลกอริทึมลอยด์จะพิจารณาการกระจายข้อมูลแบบไม่ต่อเนื่อง ในขณะที่อัลกอริทึมของฟอร์กี้จะพิจารณาการกระจายข้อมูลแบบต่อเนื่อง นอกเหนือจากนั้นทั้งสองวิธีจะมีขั้นตอนเหมือนกันทุกประการ โดยที่ผลรวมความแปรปรวนสำหรับการแจกแจงแบบไม่ต่อเนื่องและการแจกแจงแบบต่อเนื่องดังสมการ (2) และ (3) ตามลำดับ และผลรวมความแปรปรวนจะใช้พิจารณาความแปรปรวนภายในกลุ่มแต่ละกลุ่ม

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

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

โดยที่ k คือ จำนวนกลุ่ม

n คือ จำนวนจุดข้อมูล

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

c_i คือ จุดศูนย์กลางของกลุ่มที่ i

$f(x)$ คือ ฟังก์ชันความหนาแน่นความน่าจะเป็น

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

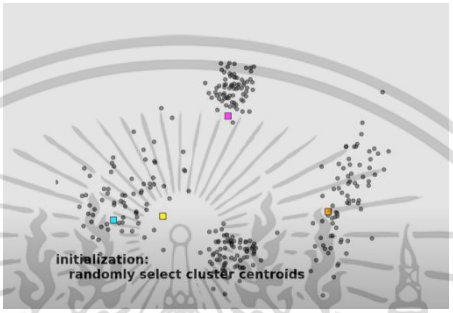
ขั้นตอนแรกของอัลกอริทึมคือการเลือกจุดศูนย์กลางเริ่มต้น โดยใช้การสุ่ม k จากชุดข้อมูล เมื่อเลือกจุดศูนย์กลางเริ่มต้นแล้ว การวนซ้ำจะดำเนินการในสองขั้นตอนต่อไปนี้ อันดับแรก แต่ละจุดข้อมูลถูกกำหนดให้กับกลุ่ม โดยพิจารณาจากระยะห่างจากจุดศูนย์กลางของกลุ่ม ทุกจุดข้อมูลที่กำหนดให้กับจุดศูนย์กลางนั้นถือว่าเป็นส่วนหนึ่งของกลุ่มนั้น ขั้นตอนที่สองคือการปรับค่าของจุดศูนย์กลางโดยใช้ค่าเฉลี่ยของจุดข้อมูลที่กำหนดให้กับจุดศูนย์กลาง จากนั้นทำซ้ำจนกว่าตำแหน่งของจุดศูนย์กลางจะหยุดเปลี่ยนหรือจนกว่าจุดข้อมูลจะไม่เปลี่ยนกลุ่ม

<u>ขั้นตอนวิธีทำ</u>

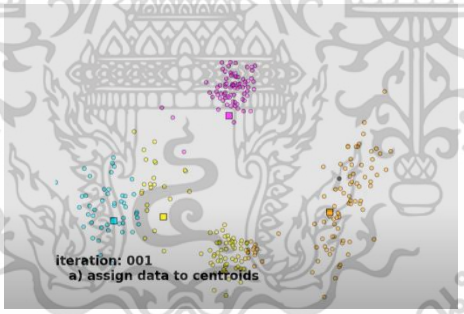
1. กำหนด k จุดให้เป็นตัวแทนจุดศูนย์กลางกลุ่ม
2. กำหนดแต่ละจุดข้อมูลให้กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด

3. คำนวณจุดศูนย์กลางของแต่ละกลุ่มใหม่จากค่าเฉลี่ยของจุดข้อมูลทั้งหมดในกลุ่มนั้น
4. ถ้าจุดศูนย์กลางของกลุ่มตัวใดตัวหนึ่งมีการเปลี่ยนแปลง ให้ทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งค่าเฉลี่ยหรือจุดศูนย์กลางในแต่ละกลุ่มจะไม่เปลี่ยนแปลง
5. สิ้นสุดการแบ่งกลุ่ม

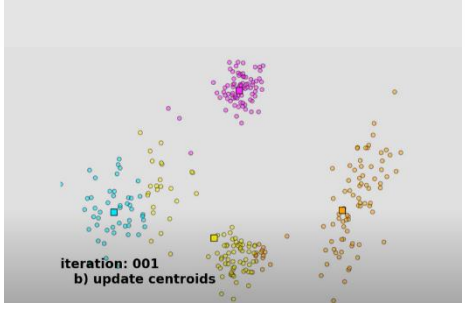
ตัวอย่างการแบ่งกลุ่มวิธีลอยด์ ดังรูปที่ 2.6-2.10



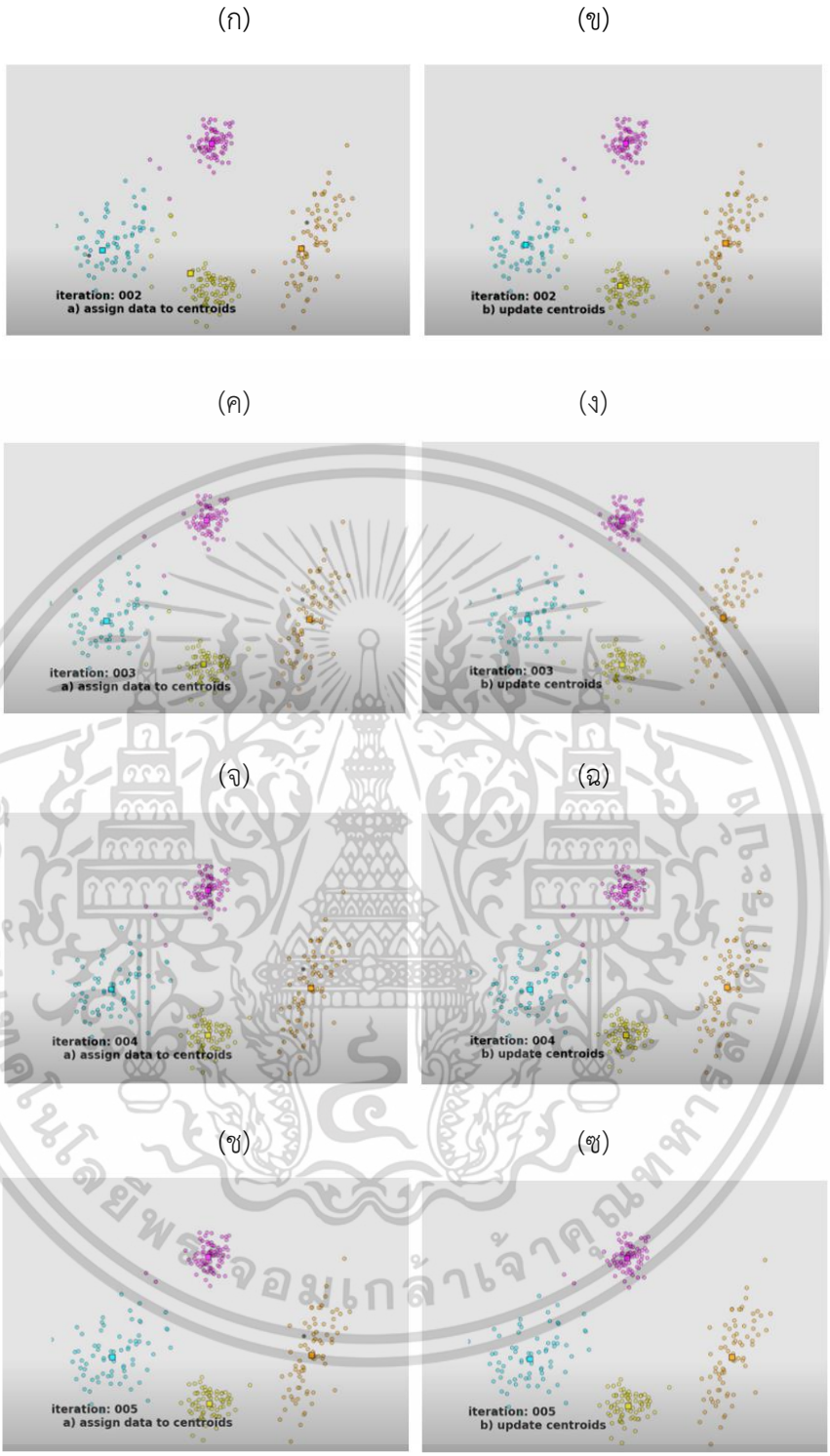
รูปที่ 2.6 แสดงการกำหนดจำนวนกลุ่มข้อมูลและกำหนดจุดศูนย์กลางเริ่มต้น



รูปที่ 2.7 แสดงการกำหนดจุดข้อมูลให้กับจุดศูนย์กลางที่ใกล้จุดข้อมูลนั้นๆมากที่สุด



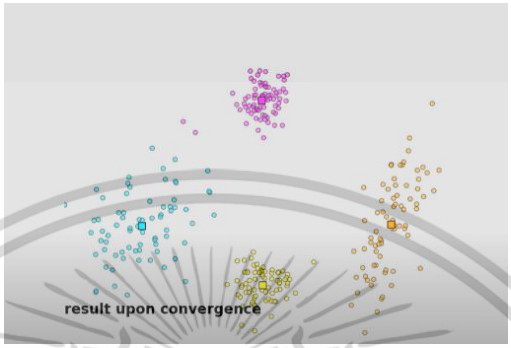
รูปที่ 2.8 แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูล



รูปที่ 2.9 แสดงการวนซ้ำครั้งที่ 2, 3, 4 และ 5 ตามลำดับ

จากรูปที่ 2.9 (ก) แสดงการกำหนดจุดข้อมูลให้กับจุดศูนย์กลางที่ใกล้จุดข้อมูลนั้นๆครั้งที่ 2 (ข) แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูลครั้งที่ 2 (ค) แสดงการกำหนดจุดข้อมูลให้กับจุดศูนย์กลางที่ใกล้จุดข้อมูลนั้นๆครั้งที่ 3 (ง) แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูลครั้งที่ 3 (จ)

แสดงการกำหนดจุดข้อมูลให้กับจุดศูนย์กลางที่ใกล้จุดข้อมูลนั้นๆครั้งที่ 4 (ฉ) แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูลครั้งที่ 4 (ช) แสดงการกำหนดจุดข้อมูลให้กับจุดศูนย์กลางที่ใกล้จุดข้อมูลนั้นๆ ครั้งที่ 5 (ซ) แสดงการปรับตำแหน่งจุดศูนย์กลางของข้อมูลครั้งที่ 5 และรูปที่ 2.10 แสดงการแบ่งกลุ่มเสร็จสิ้น



รูปที่ 2.10 แสดงการแบ่งกลุ่มเสร็จสิ้น

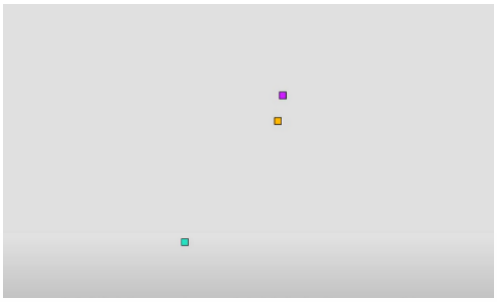
2.4.1.3 วิธีแม็คควีน (MacQueen)

อัลกอริทึมแม็คควีนเป็นอัลกอริธึมแบบวนซ้ำและคล้ายกับอัลกอริธึมของลอยด์หรือฟอร์กี้ แต่แตกต่างกับอัลกอริธึมของลอยด์หรือฟอร์กี้คือจุดศูนย์กลางจะถูกปรับใหม่ทุกครั้งที่จุดข้อมูลเปลี่ยนกลุ่ม หากจุดศูนย์กลางของกลุ่มที่อยู่ในปัจจุบันนั้นใกล้เคียงที่สุดจะไม่มีการเปลี่ยนแปลงใดๆ แต่หากจุดศูนย์กลางของกลุ่มอื่นอยู่ใกล้ที่สุด จุดข้อมูลจะถูกจัดให้กับอยู่กับกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุดตัวนั้น และจะคำนวณจุดศูนย์กลางของทั้งสองกลุ่มใหม่เป็นค่าเฉลี่ยของจุดข้อมูล อัลกอริธึมนี้มีประสิทธิภาพมากขึ้นเนื่องจากมีการปรับจุดศูนย์กลางบ่อยขึ้น

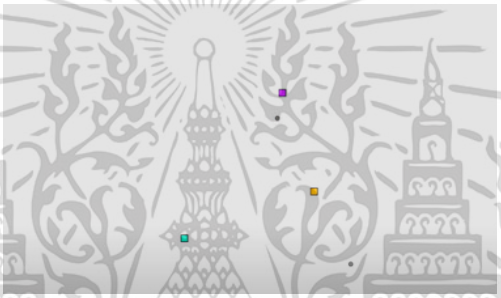
<u>ขั้นตอนวิธีทำ</u>

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

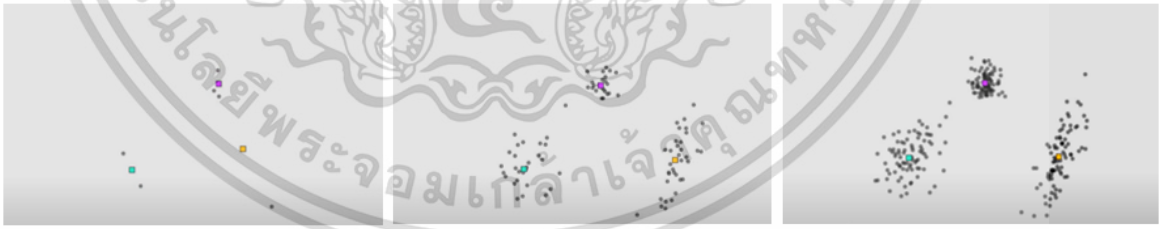
ตัวอย่างการแบ่งกลุ่มวิธีแม็คควีน ดังรูปที่ 2.11-2.14



รูปที่ 2.11 แสดงการกำหนดจำนวนกลุ่มข้อมูลและกำหนดจุดศูนย์กลางเริ่มต้น

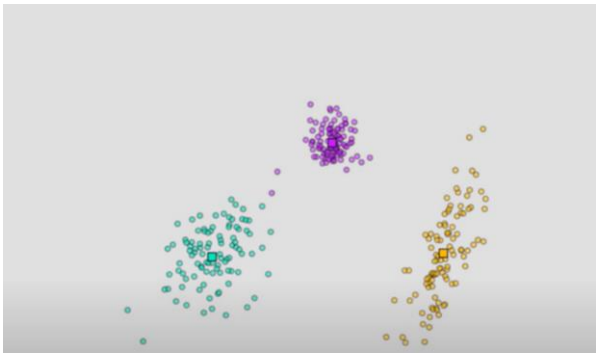


รูปที่ 2.12 แสดงการกำหนดจุดข้อมูลให้กับจุดศูนย์กลางที่ใกล้ที่สุดและการปรับตำแหน่งจุดศูนย์กลาง



รูปที่ 2.13 แสดงการเพิ่มจุดข้อมูลเข้าและการปรับตำแหน่งจุดศูนย์กลางใหม่

จากรูปที่ 2.13 จะเห็นได้ว่าจุดศูนย์กลางกลุ่มจะมีการปรับตำแหน่งทุกครั้งที่เพิ่มจุดข้อมูลเข้ามาในกลุ่ม ซึ่งเป็นข้อแตกต่างอย่างหนึ่งจากวิธีลอยด์และวิธีฟอร์กี้ และรูปที่ 2.14 แสดงการแบ่งกลุ่มเสร็จสิ้น



รูปที่ 2.14 แสดงการแบ่งกลุ่มเสร็จสิ้น

2.5 งานวิจัยที่เกี่ยวข้อง

Mahmoudi et al. (2020) ได้ทำการเปรียบเทียบการกระจายตัวของโควิด-19 ในสหรัฐอเมริกา สเปน อิตาลี เยอรมนี สหราชอาณาจักร ฝรั่งเศส และอิหร่านโดยใช้เทคนิคการแบ่งกลุ่มแบบฟัซซี่ โดยใช้สหสัมพันธ์แบบเพียร์สันวัดความสัมพันธ์ระหว่างการแพร่กระจายของโควิด-19 กับขนาดของประชากร จากการศึกษากล่าวได้ว่าการแพร่กระจายของโควิด-19 ในสเปนและอิตาลีมีความคล้ายคลึงกันและแตกต่างจากประเทศอื่น ๆ

Nurlaila et al. (2020) พบว่าแบบจำลองการแบ่งกลุ่มแบบเคมีนจะไม่ได้ดึงแบคทีเรียสกุลเดียวกันเข้าไปในกลุ่มเดียวกัน แต่แบบจำลองนี้จะรวบรวมความคล้ายคลึงกันในลักษณะการทำงานที่เรียกว่าความเข้มข้นต่ำสุดในการยับยั้ง (MIC) โดยไม่คำนึงถึงต้นกำเนิดและลำดับชั้นในการแบ่งกลุ่ม แสดงให้เห็นว่าแบบจำลองการแบ่งกลุ่มเคมีนมีความไวต่อการตรวจจับความแตกต่างของความเข้มข้นต่ำสุดในการยับยั้งมากกว่าความแตกต่างของสกุล

อาริกา ธรรมโน และคณะ (2563) ได้นำเสนออัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง รวมทั้งทำการพัฒนาโปรแกรมพยากรณ์โรคมะเร็งเต้านมด้วยภาษาไพทอน ซึ่งถูกพัฒนาต่อยอดมาจากอัลกอริทึมการแบ่งกลุ่มแบบเคมีนให้มีความสามารถในการจำแนกประเภทและปรับเปลี่ยนค่าถ่วงน้ำหนักของคุณลักษณะเด่นในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทแต่ละประเภทให้เหมาะสมได้ด้วยตัวเองในระหว่างการเรียนรู้ชุดข้อมูล

Bonera et al. (2022) ได้นำเทคนิคการแบ่งกลุ่มมาใช้ในการวิเคราะห์ความปลอดภัยทางถนนเพื่อตรวจสอบรูปแบบการเกิดอุบัติเหตุ โดยใช้การวิเคราะห์กลุ่มแฝง (Latent Class Analysis) และการแบ่งกลุ่มแบบเคมีน (K-Means Clustering) โดยใช้ข้อมูลการเกิดอุบัติเหตุที่รวบรวมไว้ในช่วง

ปีพ.ศ. 2557 ถึงปีพ.ศ. 2561 ซึ่งมีการระบุกลุ่มและรูปแบบอุบัติเหตุบนท้องถนนสำหรับคนเดินเท้า 3 กลุ่ม และระบุกลุ่มและรูปแบบอุบัติเหตุบนท้องถนนสำหรับนักปั่นจักรยาน 5 กลุ่ม เพื่อนำไปเป็นแนวทางในการแนะนำวิธีแก้ไขปัญหาเพื่อลดอุบัติเหตุและอาจเป็นเครื่องมือสนับสนุนในการตัดสินใจสำหรับผู้บริหารสาธารณะสำหรับการปรับปรุงความปลอดภัยทางถนนในอนาคต

จิรวรรณ ไพบูลย์วรชาติ และนัท กุลวานิช (2558) ได้ศึกษาการเปรียบเทียบวิธีการแบ่งกลุ่ม 4 วิธีได้แก่ 1) วิธีการแบ่งกลุ่มแบบวอร์ด 2) วิธีการแบ่งกลุ่มแบบเคมีน 3) วิธีการแบ่งกลุ่มแบบฟัซซี่ซีมีน และ 4) วิธีการแบ่งกลุ่มแบบอัลกอริทึม EM โดยจำลองข้อมูลที่มีการแจกแจงปกติแบบผสมและเปรียบเทียบประสิทธิภาพ 2 วิธี ได้แก่ วิธีการวัดค่าความแตกต่างของข้อมูลภายในกลุ่ม (RMSSTD) และวิธีการวัดค่าความต่างของข้อมูลระหว่างกลุ่ม (R-Square) ผลการวิจัยพบว่าวิธีการแบ่งกลุ่มแบบเคมีนมีประสิทธิภาพของการแบ่งกลุ่มดีที่สุด เนื่องจากทั้งค่าเฉลี่ย RMSSTD และค่าเฉลี่ย R-Square มีประสิทธิภาพดีที่สุดและเป็นไปในทิศทางเดียวกัน

บทที่ 3

วิธีการดำเนินงานวิจัย

การศึกษาเรื่องการเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของข้อมูลรหัสพันธุกรรมสำหรับข้อมูลที่มีมิติขั้นสูง มีเนื้อหาสาระสำคัญในการดำเนินการตามลำดับ ดังนี้

1. กำหนดชุดข้อมูลที่ใช้ในการวิเคราะห์การแบ่งกลุ่ม
2. เครื่องมือที่ใช้ในการวิจัย
3. การแบ่งกลุ่มข้อมูล
4. การวิเคราะห์ข้อมูลและกำหนดเกณฑ์การสรุปผล

3.1 กำหนดชุดข้อมูลที่ใช้ในการวิเคราะห์การแบ่งกลุ่ม

ข้อมูลที่นำมาวิเคราะห์ในโครงงานนี้มี 2 ชุดข้อมูล คือ ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด โดยข้อมูลแต่ละชุดมีตัวแปรอิสระและตัวแปรตามดังนี้

1) ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมอง ซึ่งได้ข้อมูลจากผู้ป่วยจำนวน 42 คน มีตัวแปรอิสระ คือ รหัสพันธุกรรมของคนไข้ จำนวน 989 รหัส เช่น สารไกลโคโปรตีนที่ทำหน้าที่ต่อต้านการเกิดมะเร็ง, เอนไซม์ในตับ, ฮีโมโกลบิน เป็นต้น และมีตัวแปรตาม คือ ระยะของการเป็นเนื้องอกในสมองจำนวน 5 กลุ่ม คือ

- medulloblastoma (MD)
- malignant gliomas (MGlio)
- normal human cerebella (Ncer)
- primitive neuroectodermal tumors (PNET)
- atypical teratoid/rhabdoid tumors (Rhab)

2) ชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด ซึ่งได้ข้อมูลจากผู้ป่วยจำนวน 197 คน มีตัวแปรอิสระ คือ รหัสพันธุกรรมของคนไข้ จำนวน 989 รหัส เช่น สารไกลโคโปรตีนที่ทำหน้าที่ต่อต้านการเกิดมะเร็ง, เอนไซม์ในตับ, ฮีโมโกลบิน เป็นต้น และมีตัวแปรตาม คือ ประเภทของโรคมะเร็งปอดจำนวน 4 กลุ่ม คือ

- Small Cell Lung Cancer (SCLC) : Oat cell lung cancer
- Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma

- Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma
- Non-Small Cell lung cancer (NSCLC) Large Cell lung cancer

3.2 เครื่องมือที่ใช้ในการวิจัย

การศึกษาครั้งนี้ใช้โปรแกรม R ในการวิเคราะห์การแบ่งกลุ่มข้อมูล

3.3 ดำเนินการแบ่งกลุ่มข้อมูล

โครงงานนี้ทำการเปรียบเทียบประสิทธิภาพการแบ่งกลุ่มข้อมูลแบบเคมีน (K-Means Clustering) จำนวน 3 วิธี ได้แก่

- วิธีฮาร์ติกัน-หว่อง (Hartigan-Wong)
- วิธีฟอร์กี้ (Forgy)
- วิธีแม็คควีน (MacQueen)

3.4 การวิเคราะห์ข้อมูลและกำหนดเกณฑ์การสรุปผล

การศึกษาครั้งนี้จะทำการสุ่มรหัสพันธุกรรมจำนวน 1000 รอบ ที่จำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 หลังจากนั้นนำมาทำการแบ่งกลุ่มข้อมูล จากนั้นจะหาค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ซึ่งเป็นวิธีหนึ่งที่จะแสดงให้เห็นถึงประสิทธิภาพของการจัดกลุ่มว่ามีมากน้อยอย่างไร โดยถ้าค่าความแตกต่างระหว่างกลุ่มมีมาก จะสรุปความหมายว่ามีการแบ่งกลุ่มที่ดี ซึ่งจะมีค่าอยู่ในช่วง 0 ถึง 1 โดยสามารถคำนวณได้จากสมการ ดังนี้

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

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

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

โดยที่ SS_t คือ ผลรวมของผลต่างกำลังสองของข้อมูลทั้งหมด

SS_w คือ ผลรวมของผลต่างกำลังสองทุกข้อมูลภายในกลุ่ม

k คือ จำนวนกลุ่มที่แบ่งได้ทั้งหมด

p คือ จำนวนตัวแปรอิสระทั้งหมด

x_{ij} คือ ข้อมูลกลุ่มที่ i ตัวแปรอิสระที่ j

$\bar{x}_j$ คือ ค่าเฉลี่ยข้อมูลในตัวแปรอิสระที่ j

$\bar{x}_{ij}$ คือ ค่าเฉลี่ยข้อมูลในกลุ่มที่ i ตัวแปรอิสระที่ j

ดังนั้น วิธีการแบ่งกลุ่มที่มีประสิทธิภาพที่สุดจะพิจารณาจากค่าความแตกต่างของข้อมูลระหว่างกลุ่มที่มีค่าสูงที่สุด โดยการวิเคราะห์ข้อมูลสามารถนำมาเขียนแผนผังได้ดังรูปที่ 3.1

รูปที่ 3.1 แสดงแผนผังการวิเคราะห์ข้อมูล

บทที่ 4

ผลการวิจัย

การศึกษาเรื่องการเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของข้อมูลรหัสพันธุกรรมสำหรับข้อมูลที่มีมิติขั้นสูง ทำการเปรียบเทียบการแบ่งกลุ่มแบบเคมีนจำนวน 3 วิธี ของข้อมูลจำนวน 2 ชุด คือชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด โดยผลการวิจัยได้ดังนี้

4.1 ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมอง

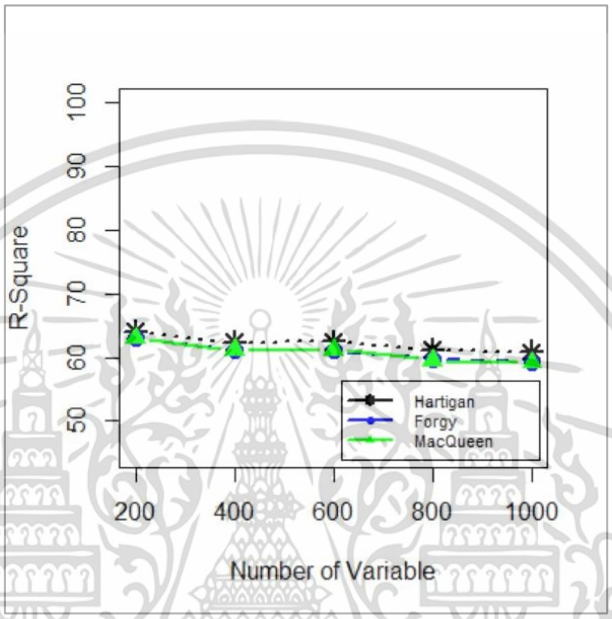
4.1.1 กำหนดจำนวนกลุ่มเท่ากับ 5 (k=5)

เมื่อนำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มข้อมูลโดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 5 (k=5) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-ห่วง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.1

ตารางที่ 4.1 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=5

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-ห่วง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**64.07499**	63.0332	62.9707
400	**62.3923**	61.12812	61.14177
600	**62.47227**	61.21671	61.27936
800	**61.15829**	59.67637	59.60488
989	**60.89558**	59.29471	59.22565

จากตารางที่ 4.1 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 5 แสดงในรูปที่ 4.1



รูปที่ 4.1 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=5

จากรูปที่ 4.1 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

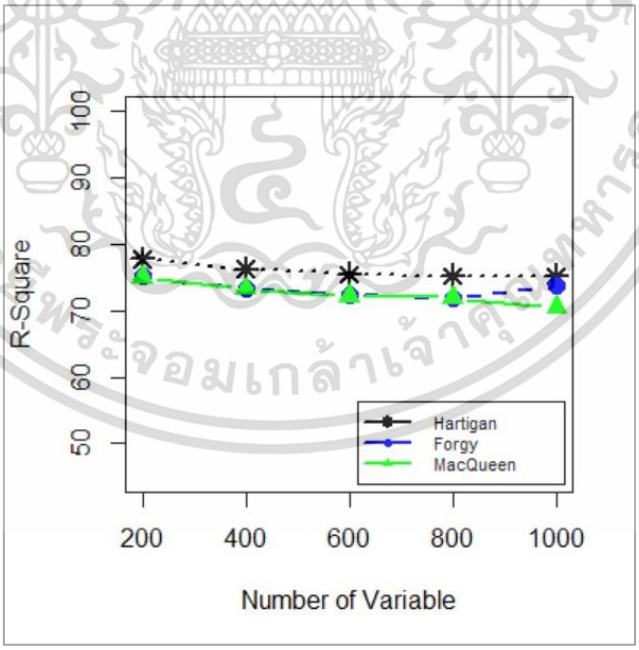
4.1.2 กำหนดจำนวนกลุ่มเท่ากับ 10 (k=10)

เมื่อนำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มข้อมูลโดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 10 (k=10) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.2

ตารางที่ 4.2 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=10

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**77.77523**	75.11854	75.1045
400	**76.23575**	73.32839	73.24863
600	**75.48335**	72.33445	72.26381
800	**75.23437**	71.88605	71.96964
989	**75.22953**	73.90632	70.51107

จากตารางที่ 4.2 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 10 แสดงในรูปที่ 4.2



รูปที่ 4.2 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=10

จากรูปที่ 4.2 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

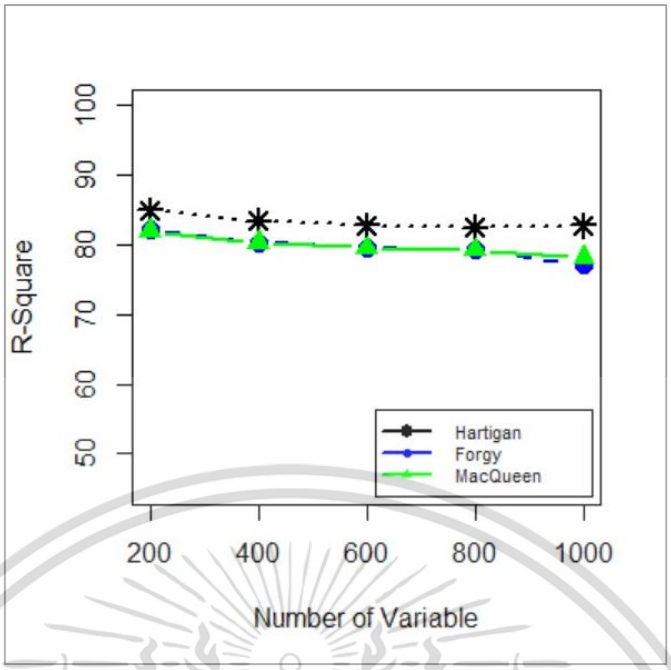
4.1.3 กำหนดจำนวนกลุ่มเท่ากับ 15 (k=15)

เมื่อนำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มข้อมูลโดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 15 (k=15) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.3

ตารางที่ 4.3 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=15

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**84.9307**	82.06628	81.98486
400	**83.41896**	80.32975	80.29764
600	**82.76705**	79.56685	79.53976
800	**82.48418**	79.27265	79.2309
989	**82.80551**	77.12179	78.19449

จากตารางที่ 4.3 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 15 แสดงในรูปที่ 4.3



รูปที่ 4.3 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=15

จากรูปที่ 4.3 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

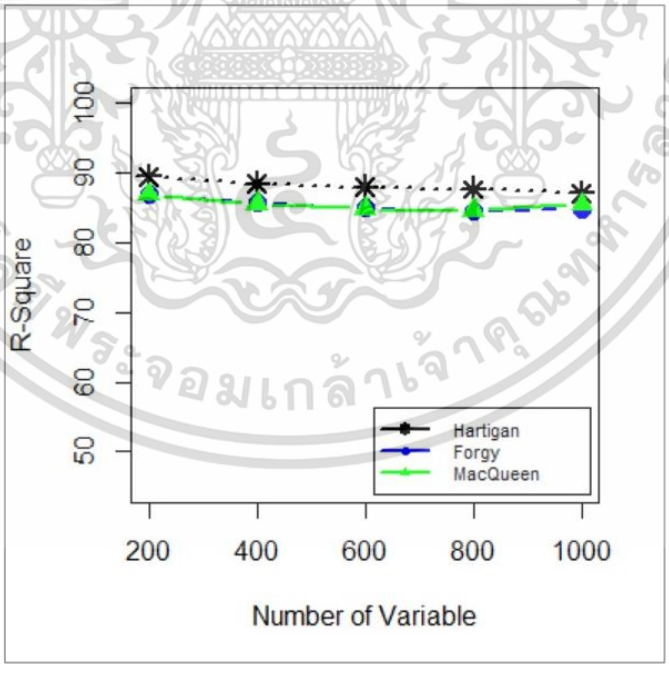
4.1.4 กำหนดจำนวนกลุ่มเท่ากับ 20 (k=20)

เมื่อนำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มข้อมูลโดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 20 (k=20) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.4

ตารางที่ 4.4 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=20

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**89.53939**	86.97089	86.90173
400	**88.45519**	85.67576	85.6338
600	**87.9444**	84.95637	84.91235
800	**87.67721**	84.58061	84.59964
989	**87.10167**	84.88275	85.53341

จากตารางที่ 4.4 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 20 แสดงในรูปที่ 4.4



รูปที่ 4.4 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=20

จากรูปที่ 4.4 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

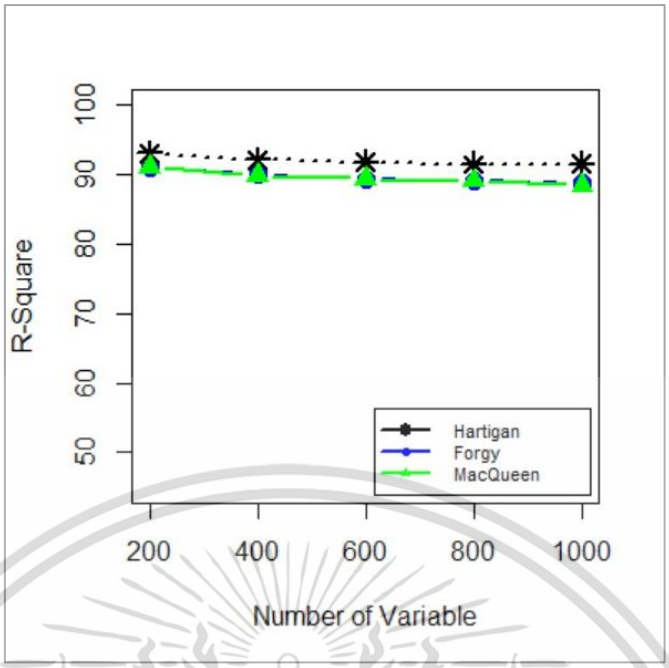
4.1.5 กำหนดจำนวนกลุ่มเท่ากับ 25 (k=25)

เมื่อนำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มข้อมูลโดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 25 (k=25) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.5

ตารางที่ 4.5 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=25

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**93.07217**	91.10535	91.08284
400	**92.14851**	89.94372	89.93394
600	**91.7871**	89.45467	89.48769
800	**91.50392**	89.06456	89.0767
989	**91.53233**	88.80875	88.51464

จากตารางที่ 4.5 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 25 แสดงในรูปที่ 4.5



รูปที่ 4.5 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=25

จากรูปที่ 4.5 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

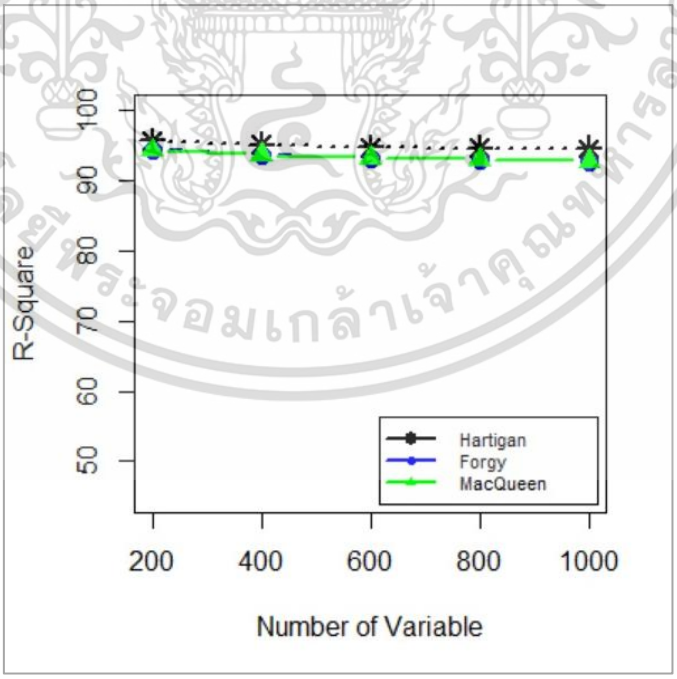
4.1.6 กำหนดจำนวนกลุ่มเท่ากับ 30 (k=30)

เมื่อนำชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองมาวิเคราะห์การแบ่งกลุ่มข้อมูลโดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 30 (k=30) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.6

ตารางที่ 4.6 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=30

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**95.73405**	94.38555	94.37609
400	**95.15308**	93.69543	93.72663
600	**94.79088**	93.25306	93.24001
800	**94.64318**	93.01936	93.02283
989	**94.60477**	92.7873	92.80962

จากตารางที่ 4.6 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 30 แสดงในรูปที่ 4.6



รูปที่ 4.6 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=30

จากรูปที่ 4.6 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

4.1.7 สรุปผล

จากตารางที่ 4.1 - 4.6 วิธีฮาร์ติกัน-หว่องมีค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด สามารถนำมาสรุปได้ดังตารางที่ 4.7

ตารางที่ 4.7 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่อง เมื่อกำหนดจำนวนกลุ่มเท่ากับ 5, 10, 15, 20, 25 และ 30

จำนวนกลุ่ม / จำนวนตัวแปร	k=5	k=10	k=15	k=20	k=25	k=30
200	64.07499	77.77523	84.9307	89.53939	93.07217	95.73405
400	62.3923	76.23575	83.41896	88.45519	92.14851	95.15308
600	62.47227	75.48335	82.76705	87.9444	91.7871	94.79088
800	61.15829	75.23437	82.48418	87.67721	91.50392	94.64318
989	60.89558	75.22953	82.80551	87.10167	91.53233	94.60477

จากตารางที่ 4.7 พบว่าจากวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่อง เมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลง และเมื่อจำนวนกลุ่มเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าเพิ่มขึ้นด้วยอย่างมีนัยสำคัญ

4.2 ชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด

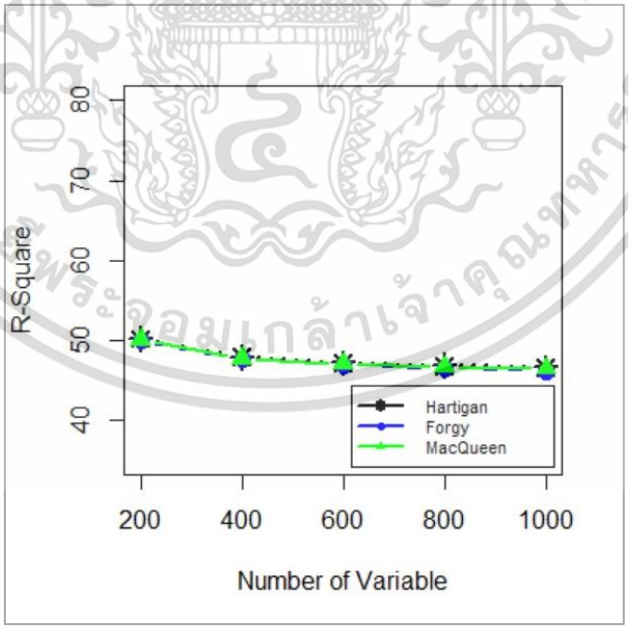
4.2.1 กำหนดจำนวนกลุ่มเท่ากับ 4 (k=4)

เมื่อนำชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอดมาวิเคราะห์การแบ่งกลุ่มข้อมูล โดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 4 (k=4) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.8

ตารางที่ 4.8 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=4

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**50.15273**	49.95728	50.0593
400	**47.81631**	47.5891	47.69284
600	**47.14206**	46.82785	46.99584
800	**46.79062**	46.43791	46.62469
989	**46.55981**	46.09693	46.42919

จากตารางที่ 4.8 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 4 แสดงในรูปที่ 4.7



รูปที่ 4.7 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=4

จากรูปที่ 4.7 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

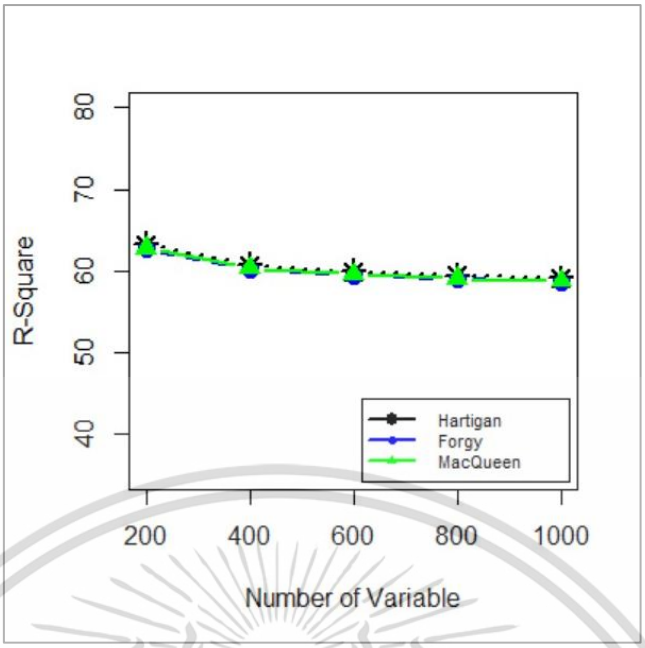
4.2.2 กำหนดจำนวนกลุ่มเท่ากับ 8 (k=8)

เมื่อนำชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอดมาวิเคราะห์การแบ่งกลุ่มข้อมูล โดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 8 (k=8) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-ห่วง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.9

ตารางที่ 4.9 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=8

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-ห่วง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**63.24347**	62.70181	62.79368
400	**60.71522**	60.17486	60.29823
600	**59.90112**	59.4328	59.53148
800	**59.37629**	58.96262	59.03342
989	**59.07545**	58.65343	58.72609

จากตารางที่ 4.9 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-ห่วงให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 8 แสดงในรูปที่ 4.9



รูปที่ 4.8 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=8

จากรูปที่ 4.8 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

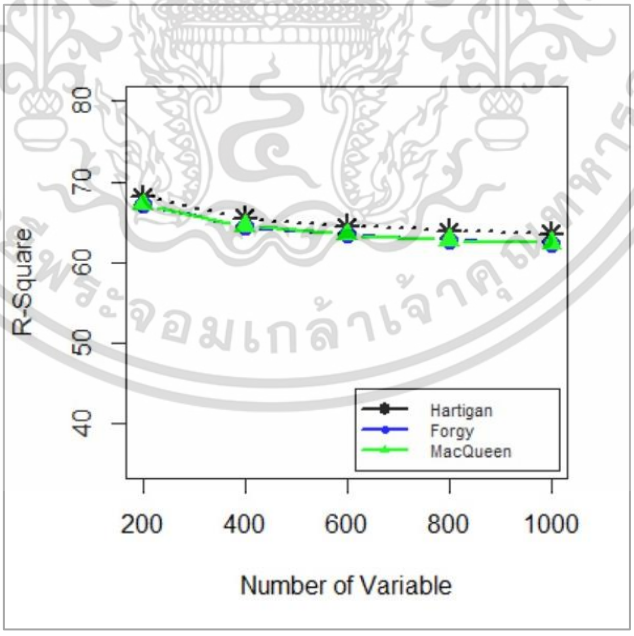
4.2.3 กำหนดจำนวนกลุ่มเท่ากับ 12 (k=12)

เมื่อนำชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอดมาวิเคราะห์การแบ่งกลุ่มข้อมูล โดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 12 (k=12) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.10

ตารางที่ 4.10 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปร เมื่อ k=12

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**68.05533**	67.12722	67.16499
400	**65.43444**	64.47374	64.51031
600	**64.42217**	63.39481	63.3921
800	**63.85807**	62.73214	62.78581
989	**63.505**	62.3427	62.38161

จากตารางที่ 4.10 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 12 แสดงในรูปที่ 4.9



รูปที่ 4.9 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=12

จากรูปที่ 4.9 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

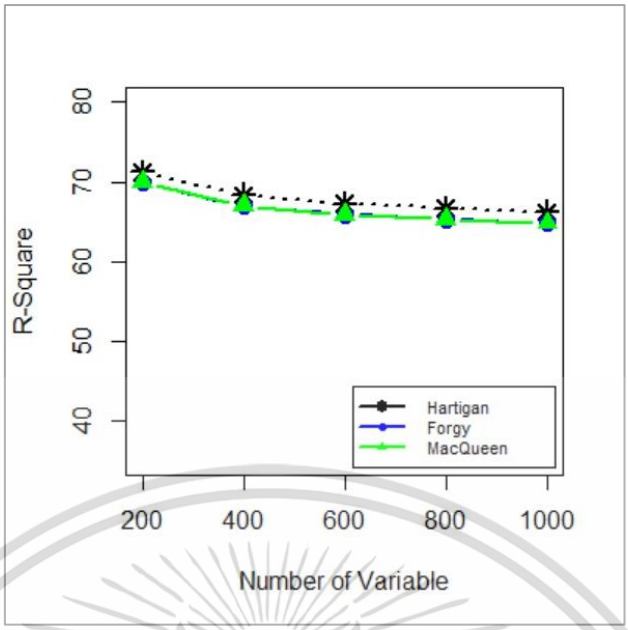
4.2.4 กำหนดจำนวนกลุ่มเท่ากับ 16 (k=16)

เมื่อนำชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอดมาวิเคราะห์การแบ่งกลุ่มข้อมูล โดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 16 (k=16) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-ห่วง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.11

ตารางที่ 4.11 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปร เมื่อ k=16

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-ห่วง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**71.19523**	69.91599	69.93941
400	**68.32549**	66.99152	66.99743
600	**67.19791**	65.85568	65.84316
800	**66.61429**	65.22821	65.23799
989	**66.20471**	64.84836	64.83522

จากตารางที่ 4.11 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-ห่วงให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 16 แสดงในรูปที่ 4.10



รูปที่ 4.10 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=16

จากรูปที่ 4.10 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

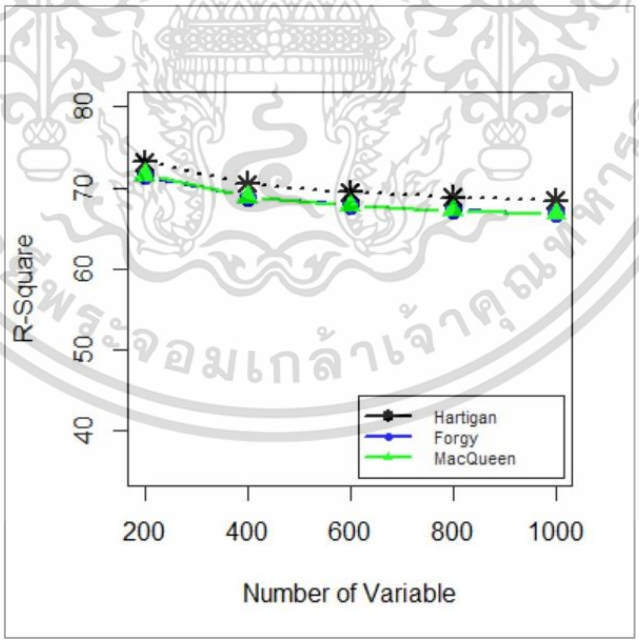
4.2.5 กำหนดจำนวนกลุ่มเท่ากับ 20 (k=20)

เมื่อนำชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอดมาวิเคราะห์การแบ่งกลุ่มข้อมูล โดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 20 (k=20) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-ห่วง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.12

ตารางที่ 4.12 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปร เมื่อ k=20

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**73.14518**	71.56504	71.58161
400	**70.47691**	68.82907	68.838136
600	**69.42008**	67.77542	67.75313
800	**68.82274**	67.17204	67.15807
989	**68.39754**	66.76835	66.74734

จากตารางที่ 4.12 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 20 แสดงในรูปที่ 4.11



รูปที่ 4.11 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=20

จากรูปที่ 4.11 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

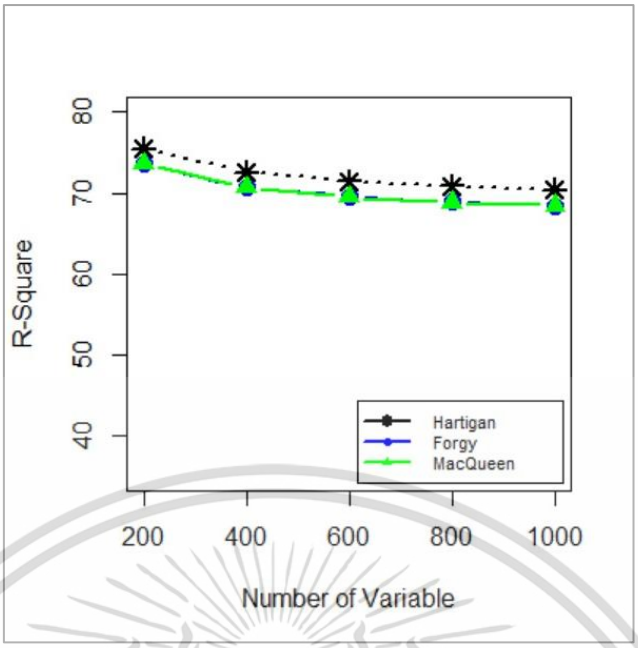
4.2.6 กำหนดจำนวนกลุ่มเท่ากับ 24 (k=24)

เมื่อนำชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอดมาวิเคราะห์การแบ่งกลุ่มข้อมูล โดยสุ่มจำนวนตัวแปรอิสระ 200, 400, 600, 800 และ 989 และกำหนดจำนวนกลุ่มเท่ากับ 24 (k=24) จากนั้นนำมาทำการแบ่งกลุ่มข้อมูลทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน จะได้ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R -Squared : RS) ดังตารางที่ 4.13

ตารางที่ 4.13 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปร เมื่อ k=24

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	**75.40993**	73.57773	73.59306
400	**72.5343**	70.6274	70.63831
600	**71.4394**	69.4759	69.46911
800	**70.80685**	68.83404	68.84848
989	**70.34828**	68.40197	68.38507

จากตารางที่ 4.13 พบว่าค่าความแตกต่างของข้อมูลระหว่างกลุ่มในจำนวนตัวแปรที่ 200, 400, 600, 800 และ 989 ตัว วิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด โดยเมื่อจำนวนตัวแปรมีขนาดเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อจำนวนกลุ่มเท่ากับ 24 แสดงในรูปที่ 4.12



รูปที่ 4.12 กราฟแสดงแนวโน้มความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มและจำนวนตัวแปรเมื่อ k=24

จากรูปที่ 4.12 พบว่าเมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลงทุกวิธี

4.2.7 สรุปผล

จากตารางที่ 4.8 - 4.13 วิธีฮาร์ติกัน-ห่วงมีค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุด สามารถนำมาสรุปได้ดังตารางที่ 4.14

ตารางที่ 4.14 แสดงค่าความแตกต่างของข้อมูลระหว่างกลุ่มของวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่องเมื่อกำหนดจำนวนกลุ่มเท่ากับ 4, 8, 12, 16, 20 และ 24

จำนวนกลุ่ม / จำนวนตัวแปร	k=4	k=8	k=12	k=16	k=20	k=24
200	50.15273	63.24347	68.05533	71.19523	73.14518	75.40993
400	47.81631	60.71522	65.43444	68.32549	70.47691	72.5343
600	47.14206	59.90112	64.42217	67.19791	69.42008	71.4394
800	46.79062	59.37629	63.85807	66.61429	68.82274	70.80685
989	46.55981	59.07545	63.505	66.20471	68.39754	70.34828

จากตารางที่ 4.14 พบว่าจากวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่อง เมื่อจำนวนตัวแปรมีค่าเพิ่มมากขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าน้อยลง และเมื่อจำนวนกลุ่มเพิ่มขึ้นค่าความแตกต่างของข้อมูลระหว่างกลุ่มมีค่าเพิ่มขึ้นด้วยอย่างมีนัยสำคัญ

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาการเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของรหัสพันธุกรรมสำหรับข้อมูลที่มีมิติขั้นสูง มีวัตถุประสงค์เพื่อค้นหาวิธีการแบ่งกลุ่มแบบเคมีนที่มีประสิทธิภาพมากที่สุดสำหรับชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด หลังจากทำการศึกษาวิธีการแบ่งกลุ่มแบบเคมีนทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และวิธีแม็คควีน และนำไปใช้ในการแบ่งกลุ่มชุดข้อมูลทั้ง 2 ชุดแล้ว ผู้วิจัยได้คำนวณค่าความแตกต่างของข้อมูลระหว่างกลุ่มเพื่อทำการเปรียบเทียบประสิทธิภาพของการแบ่งกลุ่ม โดยจากการศึกษาทั้งหมดได้ผลดังตารางที่ 4.7 และตารางที่ 4.14

5.1 สรุปผลการวิจัย

การศึกษาการเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนของรหัสพันธุกรรมสำหรับข้อมูลที่มีมิติขั้นสูง โดยเปรียบเทียบวิธีการแบ่งกลุ่มแบบเคมีนทั้งหมด 3 วิธี ได้แก่ วิธีฮาร์ติกัน-หว่อง, วิธีฟอร์กี้ และ วิธีแม็คควีน สำหรับชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด พบว่าวิธีการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่องให้ประสิทธิภาพดีที่สุดสำหรับชุดข้อมูลทั้ง 2 ชุด เนื่องจากวิธีฮาร์ติกัน-หว่องจะมีการปรับปรุงผลรวมของความเบี่ยงเบนกำลังสองภายในกลุ่ม (Laurence Morissette and Sylvain Chartier. 2013) การเปรียบเทียบประสิทธิภาพของการแบ่งกลุ่มจะพิจารณาด้วยค่าความแตกต่างของข้อมูลระหว่างกลุ่ม ซึ่งวิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุดเมื่อเปรียบเทียบกับวิธีฟอร์กี้และวิธีแม็คควีน

5.2 อภิปรายผล

สำหรับชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด การแบ่งกลุ่มแบบเคมีนด้วยวิธีฮาร์ติกัน-หว่องให้ประสิทธิภาพดีที่สุด เนื่องจากในจำนวนตัวแปรที่เท่ากันการแบ่งกลุ่มด้วยวิธีฮาร์ติกัน-หว่องให้ค่าความแตกต่างของข้อมูลระหว่างกลุ่มมากที่สุดเมื่อเปรียบเทียบวิธีฟอร์กี้และวิธีแม็คควีน ดังนั้นวิธีฮาร์ติกัน-หว่องจึงเหมาะสมที่สุดในการแบ่งกลุ่มสำหรับข้อมูลที่มีมิติขั้นสูง

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้พบว่าการศึกษาการแบ่งกลุ่มเป็นเรื่องที่จำเป็นและอาจเป็นประโยชน์ต่องานวิจัยได้หากศึกษาการแบ่งกลุ่มด้วยวิธีอื่นๆด้วยและนำมาเปรียบเทียบประสิทธิภาพการแบ่งกลุ่มเพิ่มเติม เช่น วิธีแอกเนส วิธีไดอานา เป็นต้น อีกทั้งยังควรศึกษาการจำแนกกลุ่มเพิ่มเติมเพื่อความสมบูรณ์ของงานวิจัย และเนื่องจากศึกษาในสาขาสติถิและการวิเคราะห์ธุรกิจอาจจะพิจารณาทำการแบ่งกลุ่มในเชิงธุรกิจเพื่อความเกี่ยวข้องและเชื่อมโยงกับสาขาที่กำลังศึกษาอยู่ เช่น การแบ่งกลุ่มลูกค้าตามพฤติกรรมการใช้จ่าย การแบ่งกลุ่มผู้ใช้งานบัตรเครดิตตามระยะเวลาการจ่ายหนี้ เป็นต้น

เอกสารอ้างอิง

Mahmoudi, Mohammad Reza, et al. 2020. “Fuzzy Clustering Method to Compare the Spread Rate of Covid-19 in the High Risks Countries.” *Chaos, Solitons & Fractals*, 140, 110-230.

Ika Nurlaila, Wahyu Irawati, et al. 2021. “K-Means Clustering Model to Discriminate Copper-Resistant Bacteria as Bioremediation Agents.” *Procedia Computer Science*, 179, 804-812.

Michela Bonera, Riccardo Mutti, et al. 2022. “Identifying clusters and patterns of road crash involving pedestrians and cyclists. A case study on the Province of Brescia (IT).” *Transportation Research Procedia*, 60, 512-519.

Laurence Morissette and Sylvain Chartier. 2013. “The k-means clustering technique : General considerations and implementation in Mathematica.” *Tutorials in Quantitative Methods for Psychology*, 9(1), 15-24.

Reva Joshi, Ritu Prasad, et al. 2020. “Modified LDA Approach For Cluster Based Gene Classification Using K-Mean Method.” *Procedia Computer Science*, 171, 2493-2500.

Seyed Mohamad Javidan, et al. 2022. “Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning.” *Smart Agricultural Technology*, 3.

Dataset from : https://www.broadinstitute.org/MPR/CNS.

อาริกา ธรรมโน และคณะ. 2563. “การพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง.” *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, 10(2): 1-9.

จิรวรรณ ไพบูลย์วรชาติ และนัท กุลวานิช. 2557. “การเปรียบเทียบวิธีการจัดกลุ่มสำหรับข้อมูลที่มีการแจกแจงปกติแบบผสม.” การประชุมสัมมนาทางวิชาการ มทร.ตะวันออก มภร.กลุ่มศรีอยุธยา และราชนครินทร์วิชาการและวิจัย, 2557: 311-314.

วรรณวิศา ปะเสทะกัง และคณะ. 2563. “ความวิตกกังวลในผู้ป่วยที่ได้รับการผ่าตัดของโรคเนื้องอกสมอง : การจัดการทางพยาบาล.” *วารสารการพยาบาลและการดูแลสุขภาพ*, 38(4): 35-44.

ปาริชาติ นิยมทอง. 2561. “โรคมะเร็งปอดรู้เท่าทันป้องกันได้.” *วารสารโรงพยาบาลแพร่*, 26(1): 61-80.

กัลยา วานิชย์บัญชา. 2552. “การวิเคราะห์ข้อมูลหลายตัวแปร.” กรุงเทพมหานคร : ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

วณิชา แผลงรักษาและ นิเวศ จิระวิชิตชัย. 2562. “การแบ่งกลุ่มลูกค้าโดยใช้เทคนิคการทำคลัสเตอร์แบบเคมีน สำหรับการบริหารลูกค้าสัมพันธ์.” *วารสารวิชาการซายน์เทค มรภ.ภูเก็ต*, 3(2): 1-10.

พัชราภรณ์ พรดำเนินสวัสดิ์. 2560. “การเปรียบเทียบประสิทธิภาพของการประมาณค่าพารามิเตอร์ด้วยวิธีแลซโซในการถดถอยเชิงเส้นที่มีมิติสูง.” ภาควิชาสถิติ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร.

ภาคผนวก

ภาคผนวก ก

ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ของชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองและชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด

ชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมอง

ค่าความถูกต้อง (Accuracy) ของข้อมูลเมื่อจำนวนกลุ่มเท่ากับ 5 (k=5)

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	20.04286	19.2881	19.49286
400	20.16429	19.22143	20.1833
600	19.61667	20.42857	20.49762
800	19.76667	18.91429	19.94286
989	20.20714	20.38333	20.00238

ค่าความแม่นยำ (Precision) ของข้อมูลเมื่อจำนวนกลุ่มเท่ากับ 5 (k=5)

จำนวนตัวแปร	กลุ่ม	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	1	19.59	17.25	18.55
	2	20.82	19.48	18.27
	3	19.96	20.13	21.37
	4	17.6	20.925	19.9
	5	20.9625	19.725	19.65
400	1	20.43	18.58	18.52
	2	20	18.43	20.56
	3	19.66	20.04	21.21
	4	20.475	19.425	21.175
	5	20.5125	19.8875	20.0125
600	1	18.76	20.8	22.25
	2	20.3	21.97	19.23
	3	19.34	19.98	18.71
	4	19.65	20.075	24
	5	20.1625	18.775	20.375
800	1	18.24	18.24	19.09
	2	20.05	18.77	19.26
	3	20.05	19.21	21
	4	21.15	19.075	19.625
	5	20.275	19.4875	20.7
989	1	19.97	21.46	18.85
	2	20.09	19.96	19.82
	3	20.93	20.23	20.6
	4	21.725	20.575	19.725
	5	18.9875	19.6625	21.0625

ค่าความระลึก (Recall) ของข้อมูลเมื่อจำนวนกลุ่มเท่ากับ 5 (k=5)

จำนวนตัวแปร	กลุ่ม	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	1	16.29077	14.91007	16.29979
	2	28.50926	26.81112	25.23401
	3	18.9699	19.75481	20.34347
	4	16.90139	19.49539	18.56251
	5	16.75852	16.87646	16.32947
400	1	16.8783	15.58098	15.82959
	2	30.50459	26.54444	29.25259
	3	18.31742	19.99269	20.30797
	4	19.461	17.91338	19.91004
	5	16.4193	16.36332	16.79144
600	1	15.7591	16.95897	18.18209
	2	30.28766	30.84448	27.70366
	3	18.29445	19.48263	17.85679
	4	18.9321	18.82255	22.2737
	5	15.43188	16.00023	16.68561
800	1	15.45058	14.92715	15.44281
	2	33.55016	28.5611	29.21686
	3	17.52013	17.89815	19.87093
	4	18.37111	16.5925	17.51929
	5	15.15447	15.95402	17.03618
989	1	17.13667	17.60249	16.05835
	2	34.32917	30.20691	30.06563
	3	17.48671	19.10324	19.97576
	4	17.59333	17.60952	17.03758
	5	14.60152	16.37109	16.93928

ชุดข้อมูลการจำแนกประเภทของโรคมะเร็งปอด

ค่าความถูกต้อง (Accuracy) ของข้อมูลเมื่อจำนวนกลุ่มเท่ากับ 4 (k=4)

วิธีแบ่งกลุ่ม / จำนวนตัวแปร	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	24.57208	25.35025	25.17716
400	24.86396	24.55635	24.82741
600	23.98071	24.32132	25.25025
800	24.54213	25.51168	24.41523
989	25.16548	24.34569	24.41421

ค่าความแม่นยำ (Precision) ของข้อมูลเมื่อจำนวนกลุ่มเท่ากับ 4 (k=4)

จำนวนตัวแปร	กลุ่ม	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	1	24.26331	25.66043	25.36115
	2	24.46471	24.32941	24.85882
	3	25.79048	24.99048	24.69048
	4	25.53	24.44	24.68
400	1	24.88201	24.63309	25.00719
	2	24.99412	23.06588	24.18824
	3	25.6	25.65714	24.41905
	4	23.855	23.675	24.55
600	1	23.95971	24.45468	25.36403
	2	25.22353	24.65294	24.49412
	3	24.6619	24.25714	23.7381
	4	22.355	23.18	26.69
800	1	24.60432	25.65108	24.35612
	2	24.17647	25.7	23.64706
	3	24.97143	24.35714	25.05714
	4	23.97	25.595	24.805
989	1	24.84676	23.96115	24.57698
	2	27.10588	24.54706	26.02353
	3	24.34286	26.30476	25.04762
	4	26.595	24.79	21.25

ค่าความระลึก (Recall) ของข้อมูลเมื่อจำนวนกลุ่มเท่ากับ 4 (k=4)

จำนวนตัวแปร	กลุ่ม	วิธีฮาร์ติกัน-หว่อง	วิธีฟอร์กี้	วิธีแม็คควีน
200	1	52.56391	55.85137	54.4153
	2	12.10842	11.84527	10.97979
	3	14.08962	13.74499	13.0373
	4	21.5219	19.9636	20.57472
400	1	52.42415	52.21743	53.29743
	2	13.06823	11.47597	11.94424
	3	12.50241	13.17077	12.30625
	4	21.27416	20.78844	21.57555
600	1	49.90711	52.96963	52.59273
	2	14.61211	13.42564	13.51308
	3	10.20071	10.95838	9.809855
	4	21.55144	21.17433	25.18929
800	1	49.47934	52.92182	50.23259
	2	16.4029	15.12117	14.78063
	3	8.432996	8.838415	8.53975
	4	23.94632	24.20722	24.15449
989	1	47.66518	50.98045	47.55709
	2	21.29234	14.94727	19.03193
	3	7.375546	8.739549	7.727571
	4	27.12152	23.45978	21.1249

ภาคผนวก ข

การเขียนโปรแกรมสำหรับการวิเคราห์การแบ่งกลุ่มของชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองเมื่อ k = 5

```
Brain<read.csv("C:/Users/USER/Documents/IS/brain.csv",header=TRUE,sep=",",fill=TRUE
)

attach(Brain)

names(Brain)

-----------------------------------------------------------------------------------------------------

xm1=data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,x18,x19,x20,
x21,x22,x23,x24,x25,x26,x27,x28,x29,x30,x31,x32,x33,x34,x35,x36,x37,x38,x39,x40,x41,x
42,x43,x44,x45,x46,x47,x48,x49,x50,x51,x52,x53,x54,x55,x56,x57,x58,x59,x60,x61,x62,x6
3,x64,x65,x66,x67,x68,x69,x70,x71,x72,x73,x74,x75,x76,x77,x78,x79,x80,x81,x82,x83,x84,
x85,x86,x87,x88,x89,x90,x91,x92,x93,x94,x95,x96,x97,x98,x99,x100,x101,x102,x103,x104,
x105,x106,x107,x108,x109,x110,x111,x112,x113,x114,x115,x116,x117,x118,x119,x120,x1
21,x122,x123,x124,x125,x126,x127,x128,x129,x130,x131,x132,x133,x134,x135,x136,x137,
x138,x139,x140,x141,x142,x143,x144,x145,x146,x147,x148,x149,x150,x151,x152,x153,x1
54,x155,x156,x157,x158,x159,x160,x161,x162,x163,x164,x165,x166,x167,x168,x169,x170,
x171,x172,x173,x174,x175,x176,x177,x178,x179,x180,x181,x182,x183,x184,x185,x186,x1
87,x188,x189,x190,x191,x192,x193,x194,x195,x196,x197,x198,x199,x200,x201,x202,x203,
x204,x205,x206,x207,x208,x209,x210,x211,x212,x213,x214,x215,x216,x217,x218,x219,x2
20,x221,x222,x223,x224,x225,x226,x227,x228,x229,x230,x231,x232,x233,x234,x235,x236,
x237,x238,x239,x240,x241,x242,x243,x244,x245,x246,x247,x248,x249,x250,x251,x252,x2
53,x254,x255,x256,x257,x258,x259,x260,x261,x262,x263,x264,x265,x266,x267,x268,x269,
x270,x271,x272,x273,x274,x275,x276,x277,x278,x279,x280,x281,x282,x283,x284,x285,x2
86,x287,x288,x289,x290,x291,x292,x293,x294,x295,x296,x297,x298,x299,x330,x301,x302,
x303,x304,x305,x306,x307,x308,x309,x310,x311,x312,x313,x314,x315,x316,x317,x318,x3
19,x320,x321,x322,x323,x324,x325,x326,x327,x328,x329,x330,x331,x332,x333,x334,x335,
```

x336,x337,x338,x339,x340,x341,x342,x343,x344,x345,x346,x347,x348,x349,x350,x351,x3
52,x353,x354,x355,x356,x357,x358,x359,x360,x361,x362,x363,x364,x365,x366,x367,x368,
x369,x370,x371,x372,x373,x374,x375,x376,x377,x378,x379,x380,x381,x382,x383,x384,x3
85,x386,x387,x388,x389,x390,x391,x392,x393,x394,x395,x396,x397,x398,x399,x400,x401,
x402,x403,x404,x405,x406,x407,x408,x409,x410,x411,x412,x413,x414,x415,x416,x417,x4
18,x419,x420,x421,x422,x423,x424,x425,x426,x427,x428,x429,x430,x431,x432,x433,x434,
x435,x436,x437,x438,x439,x440,x441,x442,x443,x444,x445,x446,x447,x448,x449,x450,x4
51,x452,x453,x454,x455,x456,x457,x458,x459,x460,x461,x462,x463,x464,x465,x466,x467,
x468,x469,x470,x471,x472,x473,x474,x475,x476,x477,x478,x479,x480,x481,x482,x483,x4
84,x485,x486,x487,x488,x489,x490,x491,x492,x493,x494,x495,x496,x497,x498,x499,x500,
x501,x502,x503,x504,x505,x506,x507,x508,x509,x510,x511,x512,x513,x514,x515,x516,x5
17,x518,x519,x520,x521,x522,x523,x524,x525,x526,x527,x528,x529,x530,x531,x532,x533,
x534,x535,x536,x537,x538,x539,x540,x541,x542,x543,x544,x545,x546,x547,x548,x549,x5
50,x551,x552,x553,x554,x555,x556,x557,x558,x559,x560,x561,x562,x563,x564,x565,x566,
x567,x568,x569,x570,x571,x572,x573,x574,x575,x576,x577,x578,x579,x580,x581,x582,x5
83,x584,x585,x586,x587,x588,x589,x590,x591,x592,x593,x594,x595,x596,x597,x598,x599,
x600,x601,x602,x603,x604,x605,x606,x607,x608,x609,x610,x611,x612,x613,x614,x615,x6
16,x617,x618,x619,x620,x621,x622,x623,x624,x625,x626,x627,x628,x629,x630,x631,x632,
x633,x634,x635,x636,x637,x638,x639,x640,x641,x642,x643,x644,x645,x646,x647,x648,x6
49,x650,x651,x652,x653,x654,x655,x666,x657,x658,x659,x660,x661,x662,x663,x664,x665,
x666,x667,x668,x669,x670,x671,x672,x673,x674,x675,x676,x677,x678,x679,x680,x681,x6
82,x683,x684,x685,x686,x687,x688,x689,x690,x691,x692,x693,x694,x695,x696,x697,x698,
x699,x700,x701,x702,x703,x704,x705,x706,x707,x708,x709,x710,x711,x712,x713,x714,x7
15,x716,x717,x718,x719,x720,x721,x722,x723,x724,x725,x726,x727,x728,x729,x730,x731,
x732,x733,x734,x735,x736,x737,x738,x739,x740,x741,x742,x743,x744,x745,x746,x747,x7
48,x749,x750,x751,x752,x753,x754,x755,x766,x757,x758,x759,x760,x761,x762,x763,x764,
x765,x766,x767,x768,x769,x770,x771,x772,x773,x774,x775,x776,x777,x778,x779,x780,x7
81,x782,x783,x784,x785,x786,x787,x788,x789,x790,x791,x792,x793,x794,x795,x796,x797,
x798,x799,x800,x801,x802,x803,x804,x805,x806,x807,x808,x809,x810,x811,x812,x813,x8
14,x815,x816,x817,x818,x819,x820,x821,x822,x823,x824,x825,x826,x827,x828,x829,x830,

x831,x832,x833,x834,x835,x836,x837,x838,x839,x840,x841,x842,x843,x844,x845,x846,x847,x848,x849,x850,x851,x852,x853,x854,x855,x866,x857,x858,x859,x860,x861,x862,x863,x864,x865,x866,x867,x868,x869,x870,x871,x872,x873,x874,x875,x876,x877,x878,x879,x880,x881,x882,x883,x884,x885,x886,x887,x888,x889,x890,x891,x892,x893,x894, x895,x896,x897,x898,x899,x900,x901,x902,x903,x904,x905,x906,x907,x908,x909,x910,x911,x912,x913,x914,x915,x916,x917,x918,x919,x920,x921,x922,x923,x924,x925,x926,x927,x928,x929,x930,x931,x932,x933,x934,x935,x936,x937,x938,x939,x940,x941,x942,x943,x944,x945,x946,x947,x948,x949,x950,x951,x952,x953,x954,x955,x966,x957,x958,x959,x960,x961,x962,x963,x964,x965,x966,x967,x968,x969,x970,x971,x972,x973,x974,x975,x976,x977,x978,x979,x980,x981,x982,x983,x984,x985,x986,x987,x988,x989)

---

p=200; m=1000; k =5

---

a11 =c(); a22= c(); a33 = c(); a44 = c()

a1_cl = c(); a2_cl = c(); a3_cl = c(); a4_cl= c()

accuracy1 = c(); accuracy2 = c(); accuracy3 = c();accuracy4 = c()

precision1_1 = c(); precision1_2 = c(); precision1_3 = c();precision1_4 = c();precision1_5 = c()

recall1_1 = c(); recall1_2 = c(); recall1_3 = c(); recall1_4 = c(); recall1_5 = c()

precision2_1 = c(); precision2_2 = c(); precision2_3 = c();precision2_4 = c(); precision2_5 = c()

recall2_1 = c(); recall2_2 = c(); recall2_3 = c(); recall2_4 = c(); recall2_5 = c()

precision3_1 = c(); precision3_2 = c(); precision3_3 = c();precision3_4 = c();precision3_5 = c()

recall3_1 = c(); recall3_2 = c(); recall3_3 = c(); recall3_4 = c(); recall3_5 = c()

```
for (j in 1:m) {
  xm = sample(xm1, p,replace = FALSE)
  x = as.matrix(xm)
  y = Class
  n = length(y)
  y1 = y +1
-------------------------------------------------------------------------------------------------------------
  fit1 <- kmeans(x ,k, nstart = 5,algorithm = c("Hartigan-Wong"))
  a11[j]  =( fit1$betweenss/ fit1$totss)*100
  class1 = fit1$cluster
  a1 = as.matrix(table(class1, y1))
  nn = sum(a1)
  nc = nrow(a1)
  diag = diag(a1)
  rowsums = apply(a1,1,sum)
  colsums = apply(a1,2,sum)
  accuracy1[j] = sum(diag)/nn
  precision1= diag/colsums
  precision1_1[j] = precision1[1]
  precision1_2[j] = precision1[2]
  precision1_3[j] = precision1[3]
  precision1_4[j] = precision1[4]
  precision1_5[j] = precision1[5]
```

```
recall1  = diag/rowsums

recall1_1[j] = recall1[1]

recall1_2[j] = recall1[2]

recall1_3[j] = recall1[3]

recall1_4[j] = recall1[4]

recall1_5[j] = recall1[5]
```

```
fit2 <- kmeans(x ,k, nstart =5,algorithm = c("Forgy"))

a22[j]  =( fit2$betweenss/ fit2$totss)*100

class2 = fit2$cluster

a2 = as.matrix(table(class2, y1))

nn = sum(a2)

nc = nrow(a2)

diag = diag(a2)

rowsums = apply(a2,1,sum)

colsums = apply(a2,2,sum)

accuracy2[j] = sum(diag)/nn

precision2= diag/colsums

precision2_1[j] = precision2[1]

precision2_2[j] = precision2[2]

precision2_3[j] = precision2[3]

precision2_4[j] = precision2[4]

precision2_5[j] = precision2[5]
```

```
recall2  = diag/rowsums

recall2_1[j] = recall2[1]

recall2_2[j] = recall2[2]

recall2_3[j] = recall2[3]

recall2_4[j] = recall2[4]

recall2_5[j] = recall2[5]
```

```
fit3 <- kmeans(x ,k, nstart = 5,algorithm = c("MacQueen"))

a33[j]  =( fit3$betweenss/ fit3$totss)*100

class3 = fit3$cluster

a3 = as.matrix(table(class3, y1))

nn = sum(a3)

nc = nrow(a3)

diag = diag(a3)

rowsums = apply(a3,1,sum)

colsums = apply(a3,2,sum)

accuracy3[j] = sum(diag)/nn

precision3= diag/colsums

precision3_1[j] = precision3[1]

precision3_2[j] = precision3[2]

precision3_3[j] = precision3[3]

precision3_4[j] = precision3[4]

precision3_5[j] = precision3[5]
```

```
  recall3  = diag/rowsums
  recall3_1[j] = recall3[1]
  recall3_2[j] = recall3[2]
  recall3_3[j] = recall3[3]
  recall3_4[j] = recall3[4]
  recall3_5[j] = recall3[5]
    cat("iteration = ", j, "\n")
          }
----------------------------------------------------------------------------------------------------
mean(a11)
mean(a22)
mean(a33)
mean(accuracy1*100)
mean(accuracy2*100)
mean(accuracy3*100)
----------------------------------------------------------------------------------------------------
mean(precision1_1*100)
mean(precision1_2*100)
mean(precision1_3*100)
mean(precision1_4*100)
mean(precision1_5*100)
mean(recall1_1*100)
mean(recall1_2*100)
```

mean(recall1_3*100)

mean(recall1_4*100)

mean(recall1_5*100)

---

mean(precision2_1*100)

mean(precision2_2*100)

mean(precision2_3*100)

mean(precision2_4*100)

mean(precision2_5*100)

mean(recall2_1*100)

mean(recall2_2*100)

mean(recall2_3*100)

mean(recall2_4*100)

mean(recall2_5*100)

---

mean(precision3_1*100)

mean(precision3_2*100)

mean(precision3_3*100)

mean(precision3_4*100)

mean(precision3_5*100)

mean(recall3_1*100)

mean(recall3_2*100)

mean(recall3_3*100)

```
mean(recall3_4*100)

mean(recall3_5*100)
```

การเขียนกราฟสำหรับเปรียบเทียบค่าความแตกต่างของข้อมูลระหว่างกลุ่มของการวิเคราห์การแบ่งกลุ่มของชุดข้อมูลการจำแนกระดับการเป็นเนื้องอกในสมองเมื่อ k = 5

```
H1 = c(64.07499, 62.3923, 62.47227, 61.15829, 60.89558)

F1 = c(63.0332, 61.12812, 61.21671, 59.67637, 59.29471)

M1 = c(62.9707, 61.14177, 61.27936, 59.60488, 59.22565)

no = c(200,400,600,800,1000)

plot(no, H1, type = 'b', lty=3,lwd=2.9,col= "black", ylim=c(50,70), xlab = 'Number of Variable', ylab = 'R-Square',pch = 8,cex=1.5)

lines(no,F1, type = "b",lty=2,lwd=2.9,col = "blue",pch = 16,cex=1.5)

lines(no,M1,type = "b",lty=1,lwd=2.9,col = "green",pch = 17,cex=1.5)

labels = c("H","F","M")

colors = c("black","blue","green")

pchh = c(8,16,17)

legend("topright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)
```