การจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน

Semi-Superivsed K-means clustering

Mr. Sovannarith Phan

โครงงานรายบุคคลฉบับนี้เป็นส่วนหนึ่งของการศึกษาระดับปริญญาตรี

หลักสูตรวิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ประยุกต์

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี

มหาลัยสงขลานครินทร์ วิทยาเขตปัตตานี

ปีการศึกษา 2559

กิตติกรรมประกาศ

โครงงานการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอนประสบความสำเร็จไปได้ด้วยดี เนื่องจากได้รับการสนับสนุนเป็นอย่างดีตลอดระยะเวลาของการทำโครงงาน ผู้จัดทำโครงงานขอขอบพระคุณในความอนุเคราะห์จากอาจารย์และบุคคลต่าง ๆ ด้วยกัน

ผู้จัดทำขอขอบคุณ ผศ.ดร ศิริเพ็ญ วิกัยสุขสกุล และ อาจารย์สุจรรยา บุญประดิษฐ์ ซึ่งเป็นอาจารย์ที่ปรึกษาโครงงานที่คอยให้คำปรึกษา คำแนะนำ ข้อคิดเห็น แนวทางแก้ไขปัญหา คำชี้แนะที่เป็นประโยชน์ และยังติดตามความคืบหน้าของโครงงานอยู่เสมอ เพื่อสามารถแก้ไขปัญหาต่าง ๆ ที่เกิดขึ้นระหว่างการทำโครงงาน ทำให้โครงงานสำเร็จไปได้ด้วยดี

ขอขอบคุณ คณาอาจารย์ทุกท่านที่คอยอบรมสั่งสอน ทำให้มีความรู้และความสามารถในการทำโครงงาน

ขอขอบคุณ คณะกรรมการสอบโครงงานทุกท่านที่เปิดโอกาสในการนำเสอนโครงงาน ที่ช่วยตรวจแก้ไข ปรับปรุงและให้คำแนะนำที่ดีในการจัดทำโครงงาน ทำให้โครงงานนี้สมบูรณ์ยิ่งขึ้น

ขอขอบคุณ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี ที่ได้ให้ความอนุเคราะห์ห้องปฏิบัติการคอมพิวเตอร์และอุปกรณ์ต่าง ๆ เพื่อใช้ในการทำโครงงานรายงานบุคคลฉบับนี้

ขอขอบพระคุณ บิดา มารดา ผู้ที่ให้การสนับสนุน คอยดูแลและเป็นกำลังใจให้กับผู้จัดทำโครงงานอยู่เสมอ ทำให้สามารถดำเนินโครงงานนี้สำเร็จไปได้ด้วยดี

ขอขอบคุณ เพื่อนๆ นักศึกษาสาขาคณิตศาสตร์ประยุกต์ทุกคนที่ให้คำปรึกษาและข้อเสนอแนะ เป็นกำลังใจให้ในการทำโครงงานรายบุคคลนี้สำเร็จไปได้ด้วยดี

Mr. Sovannarith Phan

ชื่อเรื่อง การจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน

ผู้เขียน Mr. Sovannarith Phan

ชื่อปริญญา วิทยาศาสตรบัณฑิต

สาขาวิชา คณิตศาสตร์ประยุกต์

ปีการศึกษา 2559

อาจารย์ที่ปรึกษา ผศ.ดร.ศิริเพ็ญ วิกัยสุขสกุล และ อาจารย์สุจรรยา บุญประดิษฐ์

บทคัดย่อ

ขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน (semi-supervised K-means clustering) เป็นการจัดกลุ่มข้อมูลโดยใช้ชุดข้อมูลที่กำกับกลุ่มจำนวนหนึ่งเพื่อนำไปจัดกลุ่มข้อมูล โดยคำนวณหาระยะห่างระหว่างข้อมูลด้วยเกณฑ์วัดระยะห่างแบบยูคลิเดียน (Euclidean distance) การวัดระยะห่างแบบ Euclidean เป็นการคำนวณหาค่าระยะห่างระหว่างหน่วยตัวอย่างกับจุดศูนย์กลางซึ่งคำนวณมาจากค่าเฉลี่ยของสมาชิกในกลุ่ม หากจุดศูนย์กลางของกลุ่มมีระยะห่างที่ใกล้เคียงกัน แสดงว่ากลุ่มของข้อมูลอาจมีการซ้อนทับกัน ทำให้จัดกลุ่มหน่วยตัวอย่างได้ไม่ถูกต้อง โครงงานนี้ได้นำเกณฑ์วัดระยะห่างแบบ Mahalanobis มาใช้ในการจัดกลุ่มข้อมูลด้วยวิธี K-means แบบกึ่งมีผู้สอน โดยการวัดระยะห่างแบบ Mahalanobis เป็นการคำนวณหาระยะห่างระหว่างหน่วยตัวอย่างกับจุดศูนย์กลางที่ใช้ค่าเมทริกซ์ความแปรปรวนร่วมด้วย เกณฑ์วัดระยะห่างทั้งสองถูกนำมาใช้ในการจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนซึ่งมีสองวิธีคือ seeded K-means และ constrained K-means โครงงานนี้ได้พัฒนาโปรแกรมจากขั้นตอนวิธีและพัฒนาวิธีการจัดกลุ่มด้วยโปรแกรม R โดยใช้เกณฑ์วัดระยะห่างทั้งสองในวิธีการจัดกลุ่มดังกล่าว ได้โปรแกรมการจัดกลุ่ม 4 วิธี และนำไปทดสอบประสิทธิภาพการทำงานกับชุดข้อมูล 5 ชุดจาก UCI dataset ผลการวัดประสิทธิภาพแสดงให้เห็นว่าการใช้เกณฑ์วัดระยะห่าง Mahalanobis มีประสิทธิภาพการจัดกลุ่มข้อมูลดีกว่าการใช้เกณฑ์วัดระยะห่าง Euclidean ในกรณีที่กลุ่มของข้อมูลมีการซ้อนทับกันมาก และมีประสิทธิภาพใกล้เคียงกันหากกลุ่มข้อมูลค่อนข้างแยกจากกัน และการเพิ่มจำนวนข้อมูลที่กำกับกลุ่มส่วนใหญ่ก็มีผลต่อการจัดกลุ่มข้อมูลทั้งสี่วิธี ทั้งนี้จำนวนข้อมูลที่กำกับกลุ่ม 30% ก็เพียงพอต่อการจัดกลุ่มข้อมูล

Title	Semi-supervised K-means clustering
Author	Mr. Sovannarith Phan
Program	Bachelor of Science
Major Program	Applied Mathematics
Academic	2016
Advisor	Assist.Prof.Dr. Siripen Wikaisuksakul
	Sujunya Boonpradit

Abstract

Semi- supervised K- means clustering uses some labeled data to aid the clustering of unlabeled data with Euclidean distance. The measures of Euclidean calculates the distance between a sample and a cluster center which is computed as the mean of each variable of samples within a cluster. If centers between clusters are close, it shows that the clusters of data maybe overlapped, and therefore, samples maybe in the wrong cluster. This work presents semi- supervised K- means clustering using Mahalanobis distance which calculates the distance between samples and cluster centers using covariance matrix. Two methods of semi- supervised K- means clustering are considered. They are seeded K-means and constrained K-means applying with the above distance measures, were developed coding with R program and experimented five dataset from UCI dataset. The clustering algorithms are coded with R programming. Four clustering programs are obtained and evaluated using five dataset from UCI repository. The results show that the performance of clustering data using Mahalanobis distance is better than the performance from Euclidean distance in case of overlapped data clusters and provides comparable performance in case of separated data. Increasing the amount of labeled data affects the performance of the four clustering methods, however, amount of labeled data 30% of the dataset is sufficient for clustering.

สารบัญ

เรื่อง **หน้า**

สารบัญ(ต่อ)

เรื่อง **หน้า**

สารบัญ(ต่อ)

เรื่อง **หน้า**

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง(ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง(ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง(ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง(ต่อ)

ตารางที่ **หน้า**

สารบัญภาพประกอบ

รูปที่ **หน้า**

สารบัญภาพประกอบ(ต่อ)

รูปที่ **หน้า**

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของโครงงาน

การจัดกลุ่มถูกใช้เป็นกระบวนการสำคัญในการวิเคราะห์ข้อมูล เป็นการเรียนรู้แบบไม่มีผู้สอนโดยพิจารณาจากความคล้ายกันของข้อมูล ขั้นตอนวิธีการจัดกลุ่มสามารถประยุกต์ใช้ในการรู้จำรูปแบบ (pattern recognition) การแบ่งส่วนภาพ (image segmentation) การค้นคืนข้อมูล (information retrieval) หรือทำเหมืองข้อมูล (data mining)

ขั้นตอนวิธีการจัดกลุ่มข้อมูลมีหลากหลายวิธี แต่ละวิธีมีประสิทธิภาพการจัดกลุ่มข้อมูล ที่แตกต่างกัน ขั้นตอนวิธีการจัดกลุ่มข้อมูลที่นิยมใช้งานคือ K-means [8] วิธีการจัดกลุ่มแบบ K-means เป็นวิธีการแบ่งกลุ่มข้อมูลหรือแบ่งหน่วยตัวอย่าง (partitioning) ออกเป็น K กลุ่ม โดยคำนวณหาระยะห่างระหว่างข้อมูลด้วยมาตรวัดระยะทางแบบยูคลิเดียน (Euclidean distance) โดยการจัดกลุ่มด้วยวิธีนี้ทำได้ง่าย ไม่ซับซ้อนและสามารถทำงานได้รวดเร็ว การวัดระยะห่างแบบ Euclidean ในขั้นตอนวิธี K-means เป็นการคำนวณหาค่าระยะห่างระหว่างสองจุดคือหน่วยตัวอย่างกับจุดศูนย์กลาง (mean) แต่ละกลุ่มข้อมูลหากจุดศูนย์กลางของกลุ่มมีระยะห่างที่ใกล้เคียงกันอาจทำให้การจัดกลุ่มหน่วยตัวอย่างได้ไม่ถูกต้อง เป็นผลต่อการจัดกลุ่มข้อมูลได้ความถูกต้องน้อย ดังนั้นจึงพิจารณานำเกณฑ์การวัดระยะห่างแบบ Mahalanobis distance มาใช้ในวิธี K-means เนื่องจากการวัดระยะห่างแบบ Mahalanobis distance เป็นการคำนวณหาระยะห่างระหว่างหน่วยตัวอย่างกับจุดศูนย์กลาง (mean) แต่ละกลุ่มข้อมูลและมีการใช้ค่าเมทริกซ์ความแปรปรวนของกลุ่มข้อมูลร่วมด้วย

นอกจากนี้วิธี K-means มีการกำหนดค่าจุดศูนย์กลางเริ่มต้นโดยใช้การสุ่มซึ่งอาจจะได้จุดศูนย์กลางที่เหมาะสมหรือไม่เหมาะสม ในการประยุกต์การใช้งานจริงชุดข้อมูลใด ๆ อาจจะมีหน่วยตัวอย่างที่รู้กลุ่มอยู่แล้วบางส่วน ข้อมูลเหล่านั้นเรียกว่า ชุดข้อมูลที่กำกับกลุ่ม (labeled data) ดังนั้นจึงสามารถนำข้อมูลที่กำกับกลุ่มที่มีอยู่จำนวนน้อยนี้ร่วมกับข้อมูลที่ไม่มีการกำกับกลุ่มมาใช้ในการจัดกลุ่มข้อมูล วิธีการนี้เรียกว่า semi-supervised clustering ดังนั้นโครงงานนี้จึงนำเกณฑ์วัดระยะห่างแบบ Mahalanobis มาใช้ในขั้นตอนวิธี semi-supervised K-means clustering

1.2 วัตถุประสงค์โครงงาน

- เพื่อนำเกณฑ์วัดระยะห่างแบบ Mahalanobis มาใช้กับขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน

- เพื่อวัดประสิทธิภาพการทำงานของวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอนกับชุดข้อมูล 5 ชุดใน UCI dataset
- เพื่อศึกษาว่าการเพิ่มจำนวนชุดข้อมูลที่กำกับกลุ่มมีผลต่อประสิทธิภาพการทำงานของขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอนหรือไม่

1.3 ขอบเขตของโครงงาน

- ศึกษาขั้นตอนวิธีการจัดกลุ่มข้อมูล
  - Unsupervised K-means เป็นการเรียนรู้แบบไม่มีผู้สอนด้วยวิธี K-means
  - Semi-supervised K-means เป็นการเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธี K-means
    - Seeded K-means
    - Constrained K-means
- ขั้นตอนวิธีการจัดกลุ่มข้อมูลต้องกำหนด (K) ล่วงหน้า
- ศึกษาและทำการทดลองเปรียบเทียบการจัดกลุ่มข้อมูลโดยใช้เกณฑ์การวัดระยะห่างของข้อมูล
  - Euclidean distance
  - Mahalanobis distance
- เปรียบเทียบการใช้ seeded K-means และ constrained K-means โดยใช้ชุดข้อมูลจาก UCI dataset
  1. ชุดข้อมูล iris
  2. ชุดข้อมูล seeds
  3. ชุดข้อมูล wine
  4. ชุดข้อมูล banknote authentication
  5. ชุดข้อมูล user knowledge modeling
- การแบ่งชุดข้อมูลเป็นสองชุดคือจุดข้อมูลที่กำกับกลุ่มและข้อมูลไม่ได้กำกับกลุ่มโดยใช้วิธีการสุ่ม

1.4 ประโยชน์ที่ค่าดว่าจะได้รับ

- สามารถใช้ขั้นตอนวิธี semi-supervised K-means with Mahalanobis นำไปประยุกต์ใช้ในการจัดกลุ่มกับชุดข้อมูลที่มีการซ้อนกัน หรือชุดข้อมูลต่าง ๆ
- ทำให้ทราบถึงจำนวนชุดข้อมูลที่กำกับกลุ่มที่เหมาะสมสำหรับวิธีการ semi-supervised K-means

1.5 ระยะเวลาในการดำเนินงาน

เดือนมกราคม พ.ศ 2560 ถึง เดือนพฤษภาคม พ.ศ 2560 ดังตารางที่ 1.1

ตารางที่ 1.1 แผนการดำเนินงาน

การดำเนินงาน	เดือน																			
	มกราคม				กุมภาพันธ์				มีนาคม				เมษายน				พฤษภาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. ศึกษาขั้นตอนวิธี K-means																				
2. ศึกษาขั้นตอนวิธี semi																				
3. เขียนโปรแกรมตามขั้นตอนวิธี																				
4. ทดลองการจัดกลุ่มข้อมูล																				
5. เขียนรายงาน																				

บทที่ 2

ความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้อง

สำหรับเนื้อหาของบทที่ 2 อธิบายความรู้พื้นฐานเกี่ยวกับการจัดกลุ่มข้อมูลและงานวิจัยที่เกี่ยวข้องเพื่อเป็นแนวทางในการดำเนินโครงงานมีรายละเอียดดังต่อไปนี้

2.1 ความรู้พื้นฐานเกี่ยวกับการจัดกลุ่มข้อมูล

2.1.1 หลักทั่วไปของการจัดกลุ่มข้อมูล

เนื้อหาเรื่องการจัดกลุ่มข้อมูลสรุปเนื้อหามาจาก [1-3,8] การจัดกลุ่มข้อมูลเป็นการแบ่งหน่วยตัวอย่างออกเป็นกลุ่ม โดยที่หน่วยตัวอย่างที่อยู่ในกลุ่มเดียวกัน จะมีลักษณะที่คล้ายคลึงกัน และหน่วยตัวอย่างที่อยู่ต่างกลุ่มจะมีลักษณะที่แตกต่างกัน การวัดความคล้ายคลึงและความแตกต่างระหว่างข้อมูล เช่น อาจวัดจากระยะห่างระหว่างข้อมูลหรือค่าความน่าจะเป็นของข้อมูล ในที่นี้ผู้วิจัยสนใจเกณฑ์การจัดกลุ่มข้อมูลแบบการวัดระยะห่างระหว่างข้อมูล การจัดกลุ่มข้อมูลจึงมีจุดประสงค์ที่ต้องการให้ระยะห่างระหว่างข้อมูลในกลุ่ม (intra cluster distance) มีค่าโดยเฉลี่ยน้อยที่สุด และระยะห่างระหว่างข้อมูลระหว่างกลุ่ม (inter cluster distance) มีค่าโดยเฉลี่ยมากที่สุด

วิธีการจัดกลุ่มข้อมูล แยกออกเป็น 2 วิธีใหญ่ ๆ ได้แก่ วิธี partition clustering และวิธี hierachical clustering สำหรับวิธี partition clustering เป็นการจัดกลุ่มข้อมูลโดยการแบ่งชุดข้อมูลเป็นกลุ่มโดยที่ข้อมูลใดข้อมูลหนึ่งจะอยู่ในกลุ่มข้อมูลเดียวเท่านั้น การแบ่งกลุ่มประเภทนี้วิธีที่นิยมใช้ได้แก่ วิธี center-based clustering วิธี distribution-based clustering และวิธี density-based clustering สำหรับวิธี center-based clustering เป็นวิธีการจัดกลุ่มข้อมูลโดยใช้การวัดระยะห่างระยะหน่วยตัวอย่างกับจุดศูนย์กลาง (center) ของกลุ่มซึ่งคำนวณได้จากค่าเฉลี่ย (mean) หรือค่ามัธยฐาน (median) เช่น เทคนิคการจัดกลุ่มข้อมูลแบบ K-means clustering และแบบ K-median clustering ส่วนวิธี distribution-based clustering เป็นวิธีการจัดกลุ่มข้อมูลโดยใช้การแจกแจงความน่าจะเป็นของข้อมูล และใช้เกณฑ์ค่าความน่าจะเป็นในการรวมกลุ่ม และวิธี density-based clustering เป็นการวิธีการจัดกลุ่มข้อมูลโดยใช้เกณฑ์ความหนาแน่นของพื้นที่ข้อมูล ศึกษารายละเอียดเพิ่มเติมได้จาก [3] วิธี hierachical clustering เป็นการจัดกลุ่มข้อมูลที่กลุ่มข้อมูลมีลักษณะซ้อนทับกัน ซึ่งกลุ่มข้อมูลหนึ่งอาจเป็นกลุ่มข้อมูลที่มีขนาดใหญ่ขึ้น กลุ่มข้อมูลจะประกอบด้วยกลุ่มข้อมูลย่อยซึ่งอาจประกอบด้วยกลุ่มข้อมูลแยกย่อยลงไปเรื่อย ๆ มีลักษณะเป็นชั้น ๆ ศึกษาเพิ่มเติมได้ที่ [1-3]

โครงงานนี้ผู้วิจัยสนใจศึกษาวิธี K-means clustering ซึ่งมีรายละเอียดดังหัวข้อถัดไป

2.1.2 การจัดกลุ่มข้อมูลแบบ K-means

วิธีการจัดกลุ่มข้อมูลแบบ center-based clustering ด้วยเทคนิค K-means [1-3,8] โดยแต่ละกลุ่มจะมีจุดศูนย์กลางซึ่งคำนวณได้จากค่าเฉลี่ย ข้อมูลทั้งหมดในกลุ่มจะอยู่ใกล้จุดศูนย์กลางนี้มากกว่าจุดศูนย์กลางของกลุ่มอื่น ๆ มีข้อจำกัดคือต้องการกำหนดจำนวนกลุ่ม (K) ล่วงหน้า

กำหนดให้เมทริกซ์ X แทนชุดข้อมูล โดยเมทริกซ์ $X=[x_{ij}]_{p\times n}$ โดย x_{ij} เป็นค่าสังเกตของตัวแปรที่ i ของหน่วยตัวอย่างที่ j เมื่อ $i=1,2,...,p$ และ $j=1,2,...,n$ และ $x_j=[x_1\ x_2\ ...\ x_p]^T$ เป็นเวกเตอร์ข้อมูลของหน่วยตัวอย่างที่ j มี p ตัวแปร หลักการของการจัดกลุ่มข้อมูล n หน่วยตัวอย่าง ออกเป็น K กลุ่ม โดยให้ผลรวมกำลังสองของค่าระยะห่างระหว่างหน่วยตัวอย่างกับจุดศูนย์กลางกลุ่มมีค่าน้อยที่สุด (within-cluster sum of distance : WCSD) ซึ่งเขียนในรูปฟังก์ชันได้ดังนี้

$$WCSD=\sum_{j=1}^{n}\sum_{k=1}^{K}\gamma_{jk}d(x_j,\mu_k) \qquad (1)$$

เมื่อ $d(x_j,\mu_k)$ คือระยะห่างระห่างของหน่วยตัวอย่าง x_j ไปยังจุดศูนย์กลางกลุ่ม μ_k และ γ_{jk} เป็นตัวบ่งชี้ (indicator) โดยที่ $\gamma_{jk}=1$ หมายถึงหน่วยตัวอย่าง x_j ถูกจัดให้อยู่กลุ่ม k และถ้า $\gamma_{jk}=0$ หน่วยตัวอย่าง x_j ไม่ได้ถูกจัดให้อยู่กลุ่ม k และขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบ K-means [1-3] ดังนี้

Input: ชุดข้อมูล X , กำหนดค่า K
Output: แบ่งชุด X เป็น K กลุ่ม

1. สุ่มเลือกข้อมูลจำนวน K ตัว จากชุดข้อมูล และใช้ตำแหน่งของข้อมูล K ตัวนี้เป็นตำแหน่งเริ่มต้นของจุดศูนย์กลางข้อมูล K กลุ่ม
2. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง x_j ไปยังจุดศูนย์กลางของแต่ละกลุ่ม μ_k โดยใช้เกณฑ์วัดระยะห่าง Euclidean distance
3. กำหนดกลุ่มที่ k ให้กับ x_j โดยระยะห่างของ x_j ไปยังจุดศูนย์กลาง μ_k ที่ใกล้ที่สุด โดยมีตัวบ่งชี้ $\gamma_{jk}=\begin{cases}1 & \text{if } k=\arg\min_k \|x_j-\mu_k\|^2 \\ 0 & \text{otherwise.}\end{cases}$, $k=1,2,...,K$

4. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
5. ทำซ้ำในข้อ 2-4 จนกระทั่งค่าจุดศูนย์กลางแต่ละกลุ่มไม่มีการเปลี่ยนแปลงหรือเมื่อทำงานครบตามจำนวนรอบที่กำหนด

2.1.3 เกณฑ์การวัดระยะห่าง

การคำนวณหาค่าระยะห่างระหว่างสองจุดในโครงงานนี้มีการใช้เกณฑ์วัดระยะห่างสองวิธี คือ Euclidean distance และ Mahalanobis distance

2.1.3.1 Euclidean distance

กำหนดให้ $\boldsymbol{x}_j = [x_1\ x_2\ \ldots\ x_p]^T$ เป็นเวกเตอร์ข้อมูลของหน่วยตัวอย่างที่ j มี p ตัวแปร และ $\boldsymbol{\mu}_k = [\mu_1\ \mu_2\ \ldots\ \mu_p]^T$ เป็นเวกเตอร์จุดศูนย์กลางกลุ่มที่ k มี p ตัวแปร ดังนั้นระยะห่างระหว่าง $\boldsymbol{x}_j$ และ $\boldsymbol{\mu}_k$ แบบ Euclidean คือ

$$d(\boldsymbol{x}_j, \boldsymbol{\mu}_k) = \sqrt{(\boldsymbol{x}_j - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_j - \boldsymbol{\mu}_k)} \qquad (2)$$

2.1.2.2 Mahalanobis distance

กำหนดให้ Σ_k เป็นเมทริกซ์ความแปรปรวนของข้อมูลกลุ่ม k ดังนั้นระยะห่างระหว่าง $\boldsymbol{x}_j$ และ $\boldsymbol{\mu}_k$ แบบ Mahalanobis คือ

$$d(\boldsymbol{x}_j, \boldsymbol{\mu}_k, \Sigma_k) = \sqrt{(\boldsymbol{x}_j - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_k)} \qquad (3)$$

2.1.4 โครงสร้างข้อมูลนำเข้า

การวิเคราะห์จัดกลุ่มข้อมูลกึ่งมีผู้สอนมีข้อมูลนำเข้า 2 ประเภทคือ ข้อมูลที่กำกับกลุ่ม (labeled data) แทนด้วย D_l และข้อมูลที่ไม่กำกับกลุ่ม (unlabeled data) แทนด้วย D_u ดังนั้นข้อมูลทั้งหมด $D = D_l \bigcup D_u$ โดย $D_l = \left\{\left(\boldsymbol{x}_j, y_j\right)\right\}_{j=1}^{n_l}$ โดย y_j เป็นหมายเลขกลุ่มของข้อมูลที่กำกับกลุ่ม $y_j \in \{1, 2, \ldots, K\}$, $j = 1, 2, \ldots, n_l$; n_l เป็นจำนวนข้อมูลที่กำกับกลุ่ม ส่วน $D_u = \left\{\left(\boldsymbol{x}_j, z_j\right)\right\}_{j=1}^{n_u}$ เป็นชุดข้อมูลที่ไม่ได้กำกับกลุ่ม (unlabeled data) โดย z_j เป็นหมายเลขกลุ่มของข้อมูลที่ต้องการจัดกลุ่ม $j = 1, 2, \ldots, n_u$; n_u เป็นจำนวนข้อมูลที่ไม่กำกับกลุ่ม และ $n = n_l + n_u$ เป็นจำนวนข้อมูลทั้งหมด

2.1.5 การจัดกลุ่มข้อมูลแบบ semi-supervised K-means clustering

Semi-supervised clustering [10] เป็นกระบวนการเรียนรู้แบบกึ่งมีผู้สอนที่ใช้ข้อมูลที่กำกับกลุ่มและข้อมูลไม่ได้กำกับกลุ่มมาใช้ร่วมกัน ในการประยุกต์ใช้งานจริง ข้อมูลที่ไม่ได้กำกับกลุ่มมีจำนวนมาก ส่วนข้อมูลที่กำกับกลุ่มมีจำนวนน้อย และมีราคาสูงในการสร้างหรือผลิต เพราะฉะนั้น semi-supervised clustering เป็นการจัดกลุ่มข้อมูลที่ใช้ข้อมูลที่ได้กำกับกลุ่มจำนวนหนึ่ง เพื่อช่วยในการจัดกลุ่มให้กับข้อมูลที่ไม่ได้กำกับกลุ่ม โดยกำหนดกลุ่มเริ่มต้นให้ข้อมูลที่ได้กำกับกลุ่ม เพื่อจะใช้ในดำเนินการการจัดกลุ่มข้อมูล semi-supervised K-means clustering ที่ได้ศึกษามี 2 วิธี คือ seeded K-means [4] และ constrained K-means [4] โดย seeded K-means เป็นขั้นตอนวิธีที่ใช้ข้อมูลที่กำกับกลุ่มเป็นจุดศูนย์กลางเริ่มต้นในการทำงาน และ constrained K-means เป็นขั้นตอนวิธีที่ใช้ข้อมูลที่กำกับกลุ่มเป็นจุดศูนย์กลางเริ่มต้นในการทำงาน และ หน่วยตัวอย่างที่อยู่ในชุดข้อมูลที่กำกับกลุ่มถูกกำหนดกลุ่มตายตัวไม่มีการเปลี่ยนแปลง และทั้งสองขั้นตอนวิธีใช้เกณฑ์วัดระยะห่าง Euclidean โดยมีขั้นตอนวิธี [4] ดังนี้

จากข้อมูลนำเข้า 2.1.4 กำหนดให้ชุดข้อมูล $D=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots\ \boldsymbol{x}_n]$ โดยมี 2 ส่วนคือ D_u กับ $D_l=\{S_k\}_{k=1}^{K}$ ซึ่ง S_k เซตย่อยข้อมูลกลุ่มที่ k ของ D_l

Input: ชุดข้อมูล D , กำหนดค่า K
Output: แบ่งชุด D เป็น K กลุ่ม

2.1.5.1 ขั้นตอนวิธี seeded K-means

1. คำนวณค่าจุดศูนย์กลางเริ่มต้นจาก D_l ซึ่ง $\boldsymbol{\mu}_k^{(0)}=\frac{1}{|S_k|}\sum_{\boldsymbol{x}_j\in S_k}\boldsymbol{x}_j$
2. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลางของแต่ละกลุ่ม $\boldsymbol{\mu}_k$ โดยใช้เกณฑ์วัดระยะห่าง Euclidean distance
3. กำหนดกลุ่มที่ k ให้กับ $\boldsymbol{x}_j$ โดยระยะห่างของ $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลาง $\boldsymbol{\mu}_k$ ที่ใกล้ที่สุด โดย k ได้มาจาก $\arg\min_k\left\|\boldsymbol{x}_j-\boldsymbol{\mu}_k\right\|^2$, $k=1,2,...,K$
4. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
5. ทำซ้ำในข้อ 2-4 จนกระทั่งค่าจุดศูนย์กลางแต่ละกลุ่มไม่มีการเปลี่ยนแปลงหรือเมื่อทำงานครบตามจำนวนรอบที่กำหนด

2.1.5.2 ขั้นตอนวิธี constrained K-means

1. คำนวณค่าจุดศูนย์กลางเริ่มต้นจาก D_l ซึ่ง $\boldsymbol{\mu}_k^{(0)} = \frac{1}{|S_k|}\sum_{x_j \in S_k} \boldsymbol{x}_j$
2. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลางของแต่ละกลุ่ม $\boldsymbol{\mu}_k$ โดยใช้เกณฑ์วัดระยะห่าง Euclidean distance
3. กำหนดกลุ่มที่ k ให้กับ $\boldsymbol{x}_j$

   3.1 ถ้าหาก $\boldsymbol{x}_j \in D_l$ กำหนดกลุ่มคงเดิม

   3.2 ถ้าหาก $\boldsymbol{x}_j \notin D_l$ กำหนดกลุ่มที่ k โดยระยะห่างของ $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลาง $\boldsymbol{\mu}_k$ ที่ใกล้ที่สุด โดย k ได้มาจาก $\arg\min_k \left\|\boldsymbol{x}_j - \boldsymbol{\mu}_k\right\|^2$, $k = 1, 2, \ldots, K$
4. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
5. ทำซ้ำในข้อ 2-4 จนกระทั่งค่าจุดศูนย์กลางแต่ละกลุ่มไม่มีการเปลี่ยนแปลงหรือเมื่อทำงานครบตามจำนวนรอบที่กำหนด

2.1.6 การวัดประสิทธิภาพการจัดกลุ่มข้อมูลโดยการใช้ confusion matrix

Confusion matrix คือการเก็บข้อมูลที่เกี่ยวกับการแบ่งแยกข้อมูลจริง กับข้อมูลที่เกิดจากการทำนาย เช่นกำหนด confusion matrix มีขนาด $K \times K$ ที่ confusion matrix ซึ่งมี K class

		True Class			
	Class	1	2	...	K
Predict Class	1	f_{11}	f_{12}	...	f_{1K}
	2	f_{21}	f_{22}	...	f_{2K}
	⋮	⋮	⋮	⋱	⋮
	K	f_{K1}	f_{K2}	...	f_{KK}

รูปที่ 2.1 ตาราง confusion matrix ที่มีขนาด $K \times K$

f_{ij} เป็นจำนวนหรือความถี่ที่ทำนายได้กลุ่มที่ i และถูกว่าเป็นกลุ่มที่ j เมื่อ $i, j = 1, 2, \ldots, K$

Confusion matrix ที่สร้างขึ้นสามารถใช้คำนวณความถูกต้องของการจำแนกข้อมูลโดยค่าที่จะนำมาใช้ คือ ความถูกต้องรวม (overall Accuracy) และ ความผิดพลาดของข้อมูลที่ทำการจำแนกเกินมา (commission Error หรือ user's Accuracy) โดยในรายงานเล่มนี้จะเรียกว่า ความถูกต้องของกลุ่ม (class accuracy)

2.1.6.1 Overall Accuracy

ความถูกต้องรวม (overall Accuracy) คือ อัตราส่วนของจำนวนข้อมูลที่จำแนกได้ถูกต้อง (ปรากฏตามแนวทแยงของตารางหลัก) ต่อจำนวนข้อมูลที่นำมาจำแนกประเภททั้งหมดและ คำนวณออกมา เป็นร้อยละ

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^{K} f_{ii}}{n} \times 100 \qquad (4)$$

n จำนวนข้อมูลที่นำมาจำแนกประเภททั้งหมด

2.1.6.2 Class Accuracy

ความถูกต้องของกลุ่ม (class accuracy) คือ อัตราส่วนของจำนวนข้อมูลที่นำมาทดสอบต่อจำนวนข้อมูล ที่จำแนกถูกต้อง ทั้งหมดของกลุ่มข้อมูลนั้น หรือ จำนวนข้อมูลที่จำแนกถูกต้องของกลุ่มข้อมูลหนึ่ง ๆ หารด้วยผลรวมจำนวนข้อมูลตามแนวนอน

$$\text{Class Accuracy ของกลุ่มที่ } k = \frac{f_{kk}}{\sum_{j=1}^{K} f_{kj}} \times 100 \ , \ k = 1,2,...,K \qquad (5)$$

2.2 งานวิจัยที่เกี่ยวข้อง

Basu Sugato (2002) [4] ศึกษากระบวนการจัดกลุ่มข้อมูลแบบกึ่งมีผู้สอนโดยใช้วิธี K-means การใช้ชุดข้อมูลที่กำกับกลุ่มจะมีสองขั้นตอนวิธีเพื่อเป็นแนวทางมาช่วยจัดกลุ่มข้อมูล ขั้นตอนวิธีที่หนึ่ง เรียกว่า seeded K-means ใช้ชุดข้อมูลที่กำกับกลุ่มเป็นแนวทางโดยหาค่าเฉลี่ยของแต่กลุ่มในชุดข้อมูลที่กำกับกลุ่ม และแทนเป็นจุดศูนย์กลางเริ่มต้นในการทำงานและขั้นตอนวิธีที่สอง เรียกว่า constrained K-means ใช้ชุดข้อมูลที่กำกับกลุ่มเริ่มต้นเช่นเดียวกับวิธี seeded K-means แต่ข้อมูลที่กำกับกลุ่มจะถูกกำหนดกลุ่มคงที่ตลอดการทำงาน ขั้นตอนวิธีทั้งสองได้อธิบายในหัวข้อ 2.1.5

Andrea Cerioli (2005) [5] ศึกษาการจัดกลุ่มข้อมูลแบบไม่มีผู้สอนแบบ K-means โดยใช้การเกณฑ์วัดระยะห่าง Mahalanobis โดยมีขั้นตอนวิธีดังนี้

1. เลือกข้อมูลจำนวน K ตัว จากชุดข้อมูล และใช้ตำแหน่งของข้อมูล K ตัวนี้เป็นตำแหน่งเริ่มต้นของจุดศูนย์กลางข้อมูล K กลุ่ม และ กำหนด $\Sigma_k^{(0)} = I$
2. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลางของแต่ละกลุ่ม $\boldsymbol{\mu}_k$ โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance ตามสมการที่ (3)
3. กำหนดกลุ่มที่ k ให้กับ $\boldsymbol{x}_j$ โดยระยะห่างของ $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลาง $\boldsymbol{\mu}_k$ ที่ใกล้ที่สุด
4. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ และหาค่าเมทริกซ์ความแปรปรวนใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
5. ทำซ้ำในข้อ 2-4 จนกระทั่งค่าจุดศูนย์กลางแต่ละกลุ่มไม่มีการเปลี่ยนแปลงหรือหน่วยตัวอย่างที่ถูกจัดอยู่ในแต่ละกลุ่มไม่มีการเปลี่ยนแปลงกลุ่ม

ส่วน Ankita Chokniwal (2016) [6] ศึกษาการจัดกลุ่มข้อมูลแบบไม่มีผู้สอนแบบ K-means โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis เช่นเดียวกันวิธี Andrea Cerioli (2005) [5] แต่ขั้นตอนวิธีทั้งสองแตกต่างกัน โดยขั้นตอนวิธี Ankita Chokniwal มีการใช้ขั้นตอนวิธี K-means++ [7] เพื่อนำมาการจัดกลุ่มข้อมูลมาก่อน และมีขั้นตอนวิธีดังนี้

1. นำชุดข้อมูลไปจัดกลุ่มข้อมูลกับขั้นตอนวิธี K-means++ เรียบร้อยก่อน
2. กำหนดศูนย์กลางเริ่มต้นโดยหาค่าเฉลี่ย และ ค่า $\Sigma_k^{(0)}$ จากการจัดกลุ่มข้อมูล K-means++ เรียบร้อย
3. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลางของแต่ละกลุ่ม $\boldsymbol{\mu}_k$ โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance ตามสมการที่ (3)
4. กำหนดกลุ่มที่ k ให้กับ $\boldsymbol{x}_j$ โดยระยะห่างของ $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลาง $\boldsymbol{\mu}_k$ ที่ใกล้ที่สุด
5. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ และหาค่าเมทริกซ์ความแปรปรวนใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
6. ทำซ้ำในข้อ 2-4 จนกระทั่งค่าจุดศูนย์กลางแต่ละกลุ่มไม่มีการเปลี่ยนแปลงหรือหน่วยตัวอย่างที่ถูกจัดอยู่ในแต่กลุ่มไม่มีการเปลี่ยนแปลงกลุ่ม

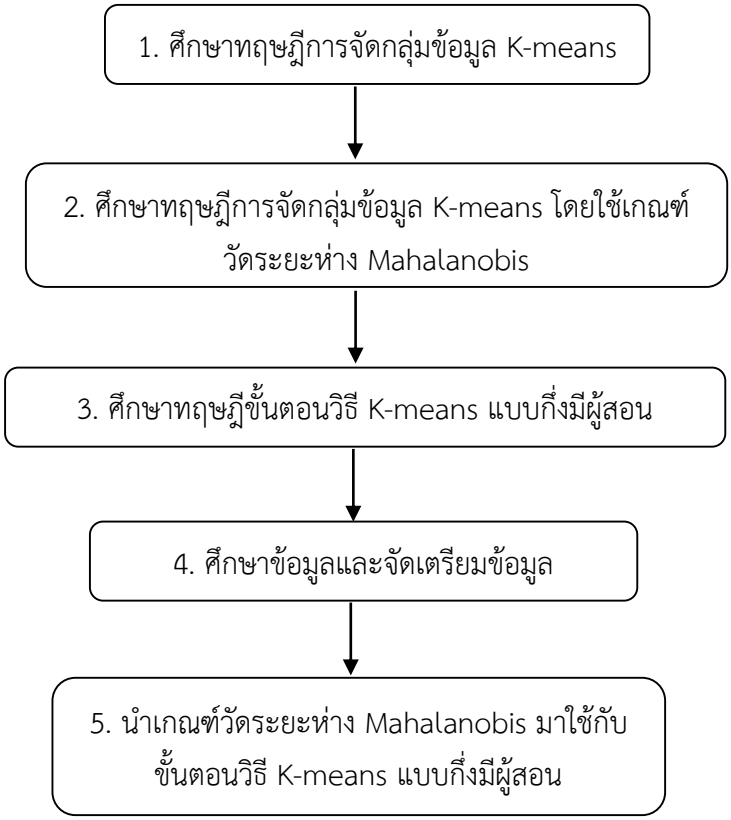
บทที่ 3

วิธีการดำเนินงาน

สำหรับเนื้อหาในบทที่ 3 ได้มีการอธิบายถึงวิธีการดำเนินงานของโครงงาน ซึ่งประกอบด้วยเรื่องขั้นตอนการดำเนินงาน ชุดข้อมูลที่นำมาศึกษา วิธีการเตรียมชุดข้อมูล การออกแบบการทดลองและการวัดประสิทธิภาพของการจัดกลุ่มข้อมูล โดยมีรายละเอียดวิธีการดำเนินงานดังต่อไปนี้

3.1 ขั้นตอนการดำเนินงาน

สำหรับขั้นตอนการดำเนินงานในโครงงานครั้งนี้ประกอบด้วย 5 ขั้นตอนหลัก ซึ่งจะอธิบายภาพรวมดังรูปที่ 3.1 ซึ่งมีรายละเอียดดังนี้



รูปที่ 3.1 ภาพรวมสำหรับขั้นตอนดำเนินงาน

1. ศึกษาทฤษฎีการจัดกลุ่มข้อมูล K-means แบบไม่มีผู้สอน [1-3,8] ที่ได้อธิบายกล่าวไว้ในหัวข้อ 2.1.2 โดยมีการพัฒนาเขียนโปรแกรมตามขั้นตอนวิธีด้วยใช้โปรแกรม R ซึ่งมีการทดลองกับชุดข้อมูล เช่น ชุดข้อมูล iris เพื่อจะศึกษาตรวจสอบการทำงานของขั้นตอนวิธี เนื่องจากตามทฤษฎีวิธี K-means มีการกำหนดค่าจุดศูนย์กลางเริ่มต้นโดยใช้การสุ่มซึ่งอาจจะได้จุดศูนย์กลางที่เหมาะสมหรือไม่เหมาะสม ส่งผลต่อการจัดกลุ่มข้อมูลและผลของกลุ่มข้อมูลที่ได้

2. ศึกษาทฤษฎีการจัดกลุ่มข้อมูล K-means โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis [5] โดยมีการพัฒนาเขียนโปรแกรมตามขั้นตอนวิธีและทดลองกับชุดข้อมูล เช่น ชุดข้อมูล iris เพื่อจะศึกษาตรวจสอบการทำงานของขั้นตอนวิธีและเปรียบเทียบการจัดกลุ่มข้อมูล K-means โดยใช้เกณฑ์วัดระยะห่าง Euclidean

3. ศึกษาทฤษฎีขั้นตอนวิธี K-means แบบกึ่งมีผู้สอนได้อธิบายกล่าวไว้ในหัวข้อ 2.1.5 โดยมีสองวิธี seeded K-means [4] และ constrained K-means [4] โดยทั้งสองวิธีใช้เกณฑ์วัดระยะห่างแบบ Euclidean พัฒนาเขียนโปรแกรมตามขั้นตอนวิธีทั้งสองวิธีและนำไปทดลองกับชุดข้อมูล 5 ชุด เพื่อจะศึกษาตรวจสอบและเปรียบเทียบการทำงานของขั้นตอนวิธีทั้งสอง

4. ศึกษาชุดข้อมูลและจัดเตรียมชุดข้อมูลเป็นการนำข้อมูลไปแปลงข้อมูลให้มีความเหมาะสมต่อการจัดกลุ่มข้อมูลโดยนำข้อมูลมาทำ standardization เพื่อให้ตัวแปรทุกตัวมีค่าอยู่ในช่วงเดียวกันก่อนจะนำเข้าไปจัดกลุ่มข้อมูลจะอธิบายรายละเอียดในหัวข้อ 3.2 และ 3.3

5. นำเกณฑ์วัดระยะห่าง Mahalanobis มาใช้ขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน 2 วิธี คือ seeded K-means และ constrained K-means

3.2 ชุดข้อมูลนำมาศึกษา

ชุดข้อมูลที่นำมาศึกษาเป็นชุดข้อมูล 5 ชุดมาจาก UCI dataset [11] ได้แก่

1. ชุดข้อมูล iris
2. ชุดข้อมูล seeds
3. ชุดข้อมูล wine
4. ชุดข้อมูล banknote authentication
5. ชุดข้อมูล user knowledge modeling

3.2.1 ชุดข้อมูล iris

ชุดข้อมูล iris เป็นข้อมูลของดอกไม้ 3 ชนิด ได้แก่ Setosa, Versicolor และ Virginica แทนเป็นกลุ่มที่ 1 ถึงกลุ่มที่ 3 ตามลำดับ มีจำนวนทั้งหมด 150 หน่วยตัวอย่าง และแต่ละชนิดมีหน่วยตัวอย่างจำนวน 50 ซึ่งแต่ละหน่วยตัวอย่างมี 4 ตัวแปร คือ ความกว้างของใบเลี้ยง (petal width) ความสูงของใบเลี้ยง (petal height) ความกว้างของกลีบดอก (sepal width) และความสูงของกลีบดอก (sepal height)

3.2.2 ชุดข้อมูล seeds

ชุดข้อมูล seeds เป็นข้อมูลของเมล็ดพันธุ์ข้าวสาลี 3 ชนิด ได้แก่ Kama, Rosa และ Canadian แทนเป็นกลุ่มที่ 1 ถึงกลุ่มที่ 3 ตามลำดับ มีจำนวนทั้งหมด 210 หน่วยตัวอย่างและแต่ละชนิดมีหน่วยตัวอย่างจำนวน 70 ค่าของตัวแปรได้มาจากการวัดสมบัติทางเรขาคณิตซึ่งแต่ละหน่วยตัวอย่างมี 7 ตัวแปร คือ ขนาดของเมล็ด (area) เส้นรอบขอบ (perimeter) compactness ความยาวของเคอร์เนล (length of kernel) ความกว้างของเคอร์เนล (width of kernel) สัมประสิทธิ์อสมมาตร (asymmetry coefficient) และความยาวของร่องเมล็ด (length of kernel groove)

3.2.3 ชุดข้อมูล wine

ชุดข้อมูล wine เป็นข้อมูลของไวน์ 3 ชนิด ได้แก่ type 1, type 2 และ type 3 แทนเป็นกลุ่มที่ 1 ถึงกลุ่มที่ 3 ตามลำดับ มีจำนวนทั้งหมด 178 หน่วยตัวอย่าง โดย type 1 มีหน่วยตัวอย่างจำนวน 59 type 2 มีหน่วยตัวอย่างจำนวน 71 และ type 3 มีหน่วยตัวอย่างจำนวน 48 ซึ่งแต่ละหน่วยตัวอย่างมี 12 ตัวแปร คือ กรดมาลิก (malic acid), Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines และ Proline

3.2.4 ชุดข้อมูล banknote authentication

ชุดข้อมูล banknote authentication เป็นข้อมูลถูกสกัดจากภาพที่ถ่ายเพื่อประเมินขั้นตอนการตรวจสอบความถูกต้องของธนบัตรโดยมี 2 ชนิด ได้แก่ false และ true แทนเป็นกลุ่มที่ 1 และกลุ่มที่ 2 ตามลำดับ มีจำนวนทั้งหมด 1,372 หน่วยตัวอย่าง โดย true มีหน่วยตัวอย่างจำนวน 610 และ false มีหน่วยตัวอย่างจำนวน 762 ซึ่งแต่ละหน่วยตัวอย่างมี 4 ตัวแปร คือ ความแปรปรวนของการแปลงภาพ (variance of Wavelet transformed image) ความเบ้ของการแปลงภาพ (skewness of Wavelet transformed image) ความโด่งของการแปลงภาพ (kurtosis of Wavelet transformed image) และเอนโทรปีของภาพ (entropy of image)

3.2.5 ชุดข้อมูล user knowledge modeling

ชุดข้อมูล user knowledge modeling เป็นข้อมูลเกี่ยวกับสถานะความรู้ของนักเรียนกับหัวข้อเครื่องไฟฟ้ากระแสตรงมี 4 ระดับ ได้แก่ high, middle, low และ very low แทนเป็นกลุ่มที่ 1 ถึงกลุ่มที่ 4 ตามลำดับ มีจำนวนทั้งหมด 403 หน่วยตัวอย่าง โดย high มีหน่วยตัวอย่างจำนวน 102 middle มีหน่วยตัวอย่างจำนวน 129 low มีหน่วยตัวอย่างจำนวน 122 และ very low มีหน่วยตัวอย่างจำนวน 50 ซึ่งแต่ละหน่วยมี 5 ตัวแปร คือ ระดับของเวลาในการศึกษา (the degree of study time for goal object materials) ระดับจำนวนซ้ำของผู้ใช้ (the degree of repetition number of user for goal object materials) ระดับของเวลาการศึกษาของผู้ใช้ที่เกี่ยวข้องกับวัตถุ

เป้าหมาย (the degree of study time of user for related objects with goal object) ประสิทธิภาพการสอบของผู้ใช้ที่เกี่ยวข้องกับวัตถุเป้าหมาย (the exam performance of user for related objects with goal object) และประสิทธิภาพการสอบของผู้ใช้ (The exam performance of user for goal objects)

ตารางที่ 3.1 สรุปลักษณะของข้อมูลทั้ง 5 ชุดได้แก่ จำนวนหน่วยตัวอย่างทั้งหมด (n) จำนวนกลุ่มข้อมูล (K) จำนวนหน่วยตัวอย่างแต่ละกลุ่ม (n_k) และจำนวนตัวแปร (p)

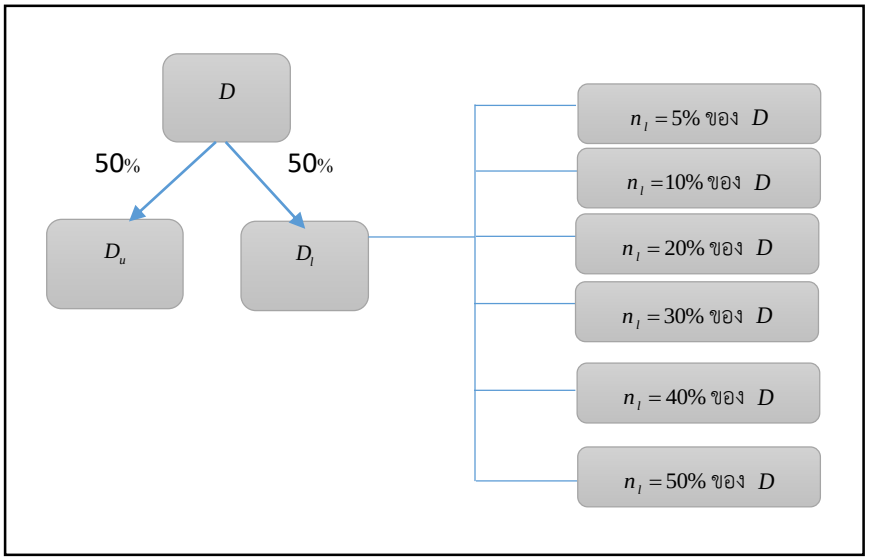
ชุดข้อมูล	จำนวนหน่วยตัวอย่างทั้งหมด	จำนวนตัวแปร	จำนวนกลุ่มข้อมูล	จำนวนหน่วยตัวอย่างแต่ละกลุ่ม			
				กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3	กลุ่มที่ 4
1. iris	150	4	3	50 (33.33%)	50 (33.33%)	50 (33.33%)	
2. seeds	210	7	3	70 (33.33%)	70 (33.33%)	70 (33.33%)	
3. wine	178	12	3	59 (33.15%)	71 (39.89%)	48 (26.90%)	
4. banknote	1372	4	2	610 (44.465)	762 (55.54%)		
5. user	403	5	4	102 (25.31%)	129 (32.01%)	122 (30.27%)	50 (12.41%)

3.3 วิธีการเตรียมชุดข้อมูล

ก่อนการจัดกลุ่มข้อมูลจะต้องมีการแปลงค่าตัวแปรเพื่อให้ตัวแปรทุกตัวมีค่ามาตรฐานอยู่ในช่วงเดียวกันหรือเรียกว่า standardization [1,3] ทำให้ค่าระยะห่างระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มข้อมูลไม่ขึ้นกับค่าตัวแปรที่มีหน่วยแตกต่างกัน กำหนดให้ $Z = (\Sigma^{\frac{1}{2}})^{-1}(X - \boldsymbol{\mu})$ เขียนแทนด้วยเมทริกซ์ข้อมูล $Z = [z_1\ z_2\ \ldots\ z_p]^T$ โดย $\Sigma^{\frac{1}{2}}$ เป็นค่ารากที่สองของเมทริกซ์สมมาตร (Symmetric square root matrix) สามารถหาวิธีคำนวณได้จาก [9] ที่ทำให้ $\Sigma = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$ จะได้ว่า Z มีการแจกแจงปกติที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์ และเมทริกซ์ค่าความแปรปรวนเป็นหนึ่ง หรือ $Z \sim N_p(0, \boldsymbol{I})$ โดยที่ $\boldsymbol{I}$ เป็นเมทริกซ์เอกลักษณ์ (Identity matrix) ดังนั้นถ้า z_i เป็นเวกเตอร์ข้อมูลของตัวแปรที่ i ของ Z จะมีการแจกแจงแบบปกติมาตรฐานที่มีค่าเฉลี่ยเป็นศูนย์ และ ค่าความแปรปรวนเป็นหนึ่ง หรือ $z_i \sim N(0,1)$

3.4 การออกแบบการทดลอง

การทดลองโครงงานนี้เป็นการทดลองกับชุดข้อมูล 5 ชุดโดยแต่ละชุดข้อมูลจะถูกแบ่งออกเป็น 2 ส่วนเท่า ๆกัน ชุดข้อมูลที่กำกับกลุ่ม D_l และชุดข้อมูลที่ไม่ได้กำกับกลุ่ม D_u ในส่วนข้อมูลที่กำกับกลุ่มใช้จำนวนข้อมูล n_l ในการทดลองเป็น 5% 10% 20% 30% 40% และ 50% ของจำนวนข้อมูลทั้งหมดเมื่อได้จำนวนของข้อมูลที่กำกับกลุ่มแล้วจะสุ่มเลือกหน่วยตัวอย่างจำนวน n_l ตัวจากข้อมูล D_l แสดงสรุปดังรูปที่ 3.2



รูปที่ 3.2 การแบ่งชุดข้อมูลเพื่อทำการทดลอง

การทดลองโดยใช้ชุดข้อมูลที่นำมาศึกษามี 2 การทดลองคือ การทดลองที่ 1 และการทดลองที่ 2

3.4.1 การทดลองที่ 1

วัดประสิทธิภาพการจัดกลุ่มข้อมูลของวิธี seeded K-means กับ constrained K-means โดยใช้เกณฑ์วัดระยะห่างแบบ Euclidean โดยใช้จำนวนข้อมูลที่กำกับกลุ่ม n_l เป็นจำนวน 5% 10% 20% 30% 40% และ 50% ทดสอบกับชุดข้อมูลทั้ง 5 โดยในแต่ละชุดข้อมูลกำหนดการทดลองดังรูปที่ 3.5 มีการทำซ้ำ 10 ครั้ง ซึ่งแต่ละครั้งได้มาจากการสุ่มข้อมูลที่แตกต่างกัน และประสิทธิภาพการจัดกลุ่มข้อมูลได้ความถูกต้องวัดจาก D_u

3.4.2 การทดลองที่ 2

วัดประสิทธิภาพการจัดกลุ่มข้อมูลของวิธี seeded K-means กับ constrained K-means โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance และออกแบบการทดลองเช่นเดียวกับการทดลองที่ 1

3.5 การวัดประสิทธิภาพของวิธีการจัดกลุ่มข้อมูล

การวัดประสิทธิภาพของวิธีการจัดกลุ่มข้อมูลโดยใช้ confusion matrix ที่กล่าวไว้ในบทที่ 2 โดยการวัดประสิทธิภาพของวิธีการจัดกลุ่มข้อมูล 2 ค่า คือ ความถูกต้องรวม (overall accuracy) และ ความถูกต้องของกลุ่ม (class accuracy)

บทที่ 4

ผลการดำเนินงาน

บทนี้เป็นการอธิบายผลการดำเนินงานของโครงงานการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน แสดงรายละเอียดการพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูล ผลการทดลอง สรุปผลการทดลอง และอภิปรายผลการทดลอง ดังต่อไปนี้

4.1 การพัฒนาขั้นตอนวิธีการจัดกลุ่มข้อมูล

จากการศึกษาขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน 2 วิธี คือ seeded K-means และ constrained K-means ในหัวข้อ 2.1.5 โครงงานนี้ได้พัฒนาโปรแกรมจากขั้นตอนวิธีดังกล่าว โดยใช้โปรแกรม R เรียกวิธีทั้งสองว่า seeded K-means with Euclidean (SKE) และ constrained K-means with Euclidean (CKE) และศึกษาทฤษฎีขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบ K-means โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis ในหัวข้อ 2.2 ดังนั้นโครงงานนี้นำขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน 2 วิธีโดยใช้เกณฑ์วัดระยะห่าง Mahalanobis เรียกวิธีทั้งสองนี้ว่า seeded K-means with Mahalanobis (SKM) และ constrained K-means with Mahalanobis (CKM) และมีรายละเอียดขั้นตอนวิธีดังนี้

4.1.1 ขั้นตอนวิธี seeded K-means with Mahalanobis (SKM)

จากหัวข้อ 2.1.4 กำหนดให้ชุดข้อมูล $D=[x_1\ x_2\ ...\ x_n]$ โดยมี 2 ส่วนคือ D_u กับ $D_l=\{S_k\}_{k=1}^{K}$ ซึ่ง S_k เซตย่อยข้อมูลกลุ่มที่ k ของ D_l

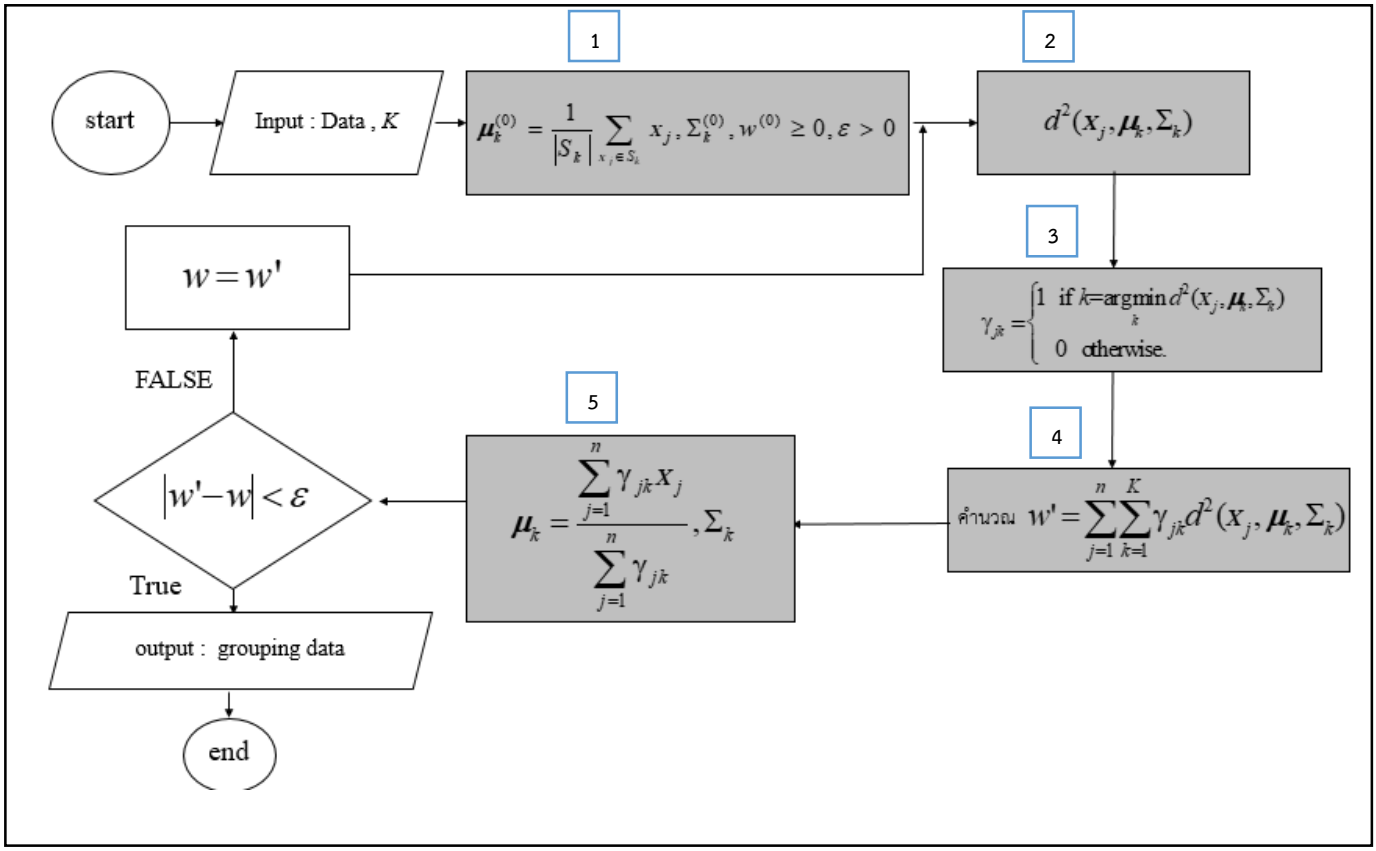
Input: ชุดข้อมูล D , กำหนด K กลุ่ม

Output: แบ่งชุดข้อมูลเป็น K กลุ่ม k

1. คำนวณค่าจุดศูนย์กลางเริ่มต้นจาก D_l ซึ่ง $\mu_k^{(0)}=\frac{1}{|S_k|}\sum_{x_j\in S_k} x_j$ และคำนวณค่าเมทริกซ์ความแปรปรวนเริ่มต้นแต่ละกลุ่ม $\Sigma_k^{(0)}$ จากชุดข้อมูลที่กำกับกลุ่ม กำหนดค่า WSCD เริ่มต้น โดยแทนเป็นตัวแปร $w^{(0)}\geq 0$ กับค่า $\varepsilon>0$
2. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง x_j ไปยังจุดศูนย์กลางของแต่ละกลุ่ม μ_k โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance
3. กำหนดกลุ่มที่ k ให้กับ x_j โดยระยะห่างของ x_i ไปยังจุดศูนย์กลาง μ_k ที่ใกล้ที่สุด โดยมีตัวบ่งชี้ $\gamma_{jk}=\begin{cases}1 & \text{if } k=\arg\min_k d^2(x_j,c_k,\Sigma_k)\\ 0 & \text{otherwise.}\end{cases}$, $k=1,2,...,K$

4. หาค่า WSCD ใหม่ โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance โดยแทนเป็นตัวแปร w'
5. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่และหาค่าเมทริกซ์ความแปรปรวนแต่ละกลุ่มใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
6. ตรวจสอบเงื่อนไข ถ้า $|w'-w|<\varepsilon$ หยุดการทำงาน และ ถ้าไม่ใช่ ให้กำหนดค่า $w=w'$ แล้วทำซ้ำในข้อ 2-6

การทำงานตามขั้นตอนวิธีข้างต้นสามารถสรุปได้ ดังแผนภาพในรูปที่ 4.1 ส่วนที่แตกต่างจากงานวิจัยที่ได้ศึกษาขั้นตอนวิธี SKE คือ การใช้เกณฑ์วัดระยะห่างแบบ Mahalanobis ในขั้นตอนวิธี 1-5 ในแผนภาพการทำงานของขั้นตอนวิธี SKM และการหยุดการทำงานที่สามารถนับจำนวนรอบโดยใช้ค่า WSCD ซึ่งผลลัพธ์จำนวนรอบในการทำงานแสดงในภาคผนวก ข.1 และ ข.2



รูปที่ 4.1 แผนภาพการทำงานของขั้นตอนวิธี SKM

4.1.2 ขั้นตอนวิธี constrained K-means with Mahalanobis (CKM)

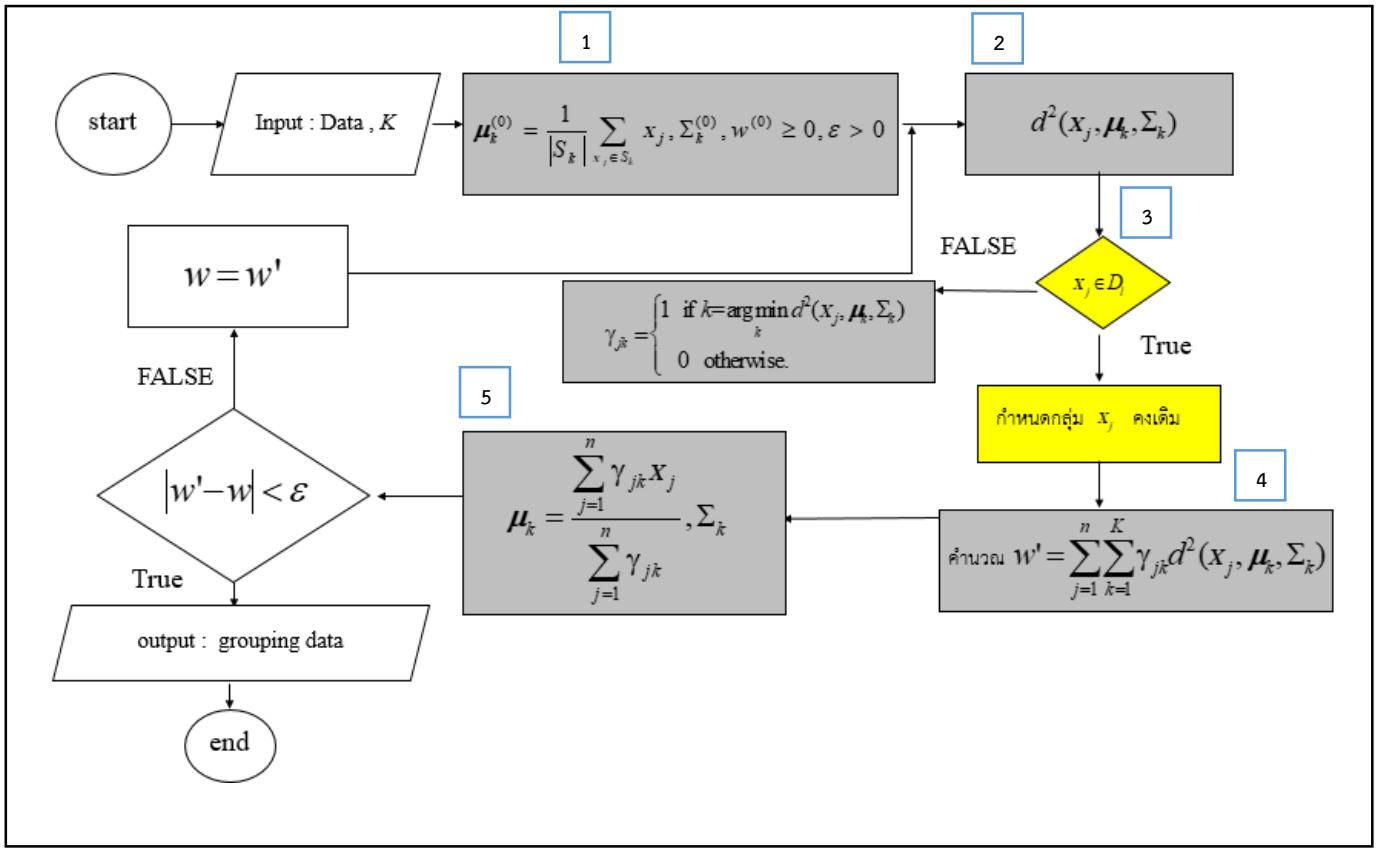
1. คำนวณค่าจุดศูนย์กลางเริ่มต้นจาก D_l ซึ่ง $\boldsymbol{\mu}_k^{(0)} = \frac{1}{|S_k|} \sum_{\boldsymbol{x}_j \in S_k} \boldsymbol{x}_j$ และคำนวณค่าเมทริกซ์ความแปรปรวนเริ่มต้นแต่ละกลุ่ม $\Sigma_k^{(0)}$ จากชุดข้อมูลที่กำกับกลุ่ม กำหนดค่า WSCD เริ่มต้น โดยแทนเป็นตัวแปร $w^{(0)} \geq 0$ กับค่า $\varepsilon > 0$
2. คำนวณค่าระยะห่างของทุกหน่วยตัวอย่าง $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลางของแต่ละกลุ่ม $\boldsymbol{\mu}_k$ โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance
3. กำหนดกลุ่มที่ k ให้กับ $\boldsymbol{x}_j$

   3.1 ถ้าหาก $\boldsymbol{x}_j \in D_l$ ให้กำหนดกลุ่มคงเดิม

   3.2 ถ้าหาก $\boldsymbol{x}_j \notin D_l$ ให้กำหนดกลุ่ม k โดยระยะห่างของ $\boldsymbol{x}_j$ ไปยังจุดศูนย์กลาง $\boldsymbol{\mu}_k$ ที่ใกล้ที่สุด โดยมีตัวบ่งชี้

   $$\gamma_{jk} = \begin{cases} 1 & \text{if } k = \arg\min_k d^2(\boldsymbol{x}_j, c_k, \Sigma_k) \\ 0 & \text{otherwise.} \end{cases}, k = 1, 2, ..., K$$
4. หาค่า WSCD ใหม่ โดยใช้เกณฑ์วัดระยะห่าง Mahalanobis distance โดยแทนเป็นตัวแปร w'
5. หลังจากจัดข้อมูลทั้งหมดเข้าเป็นกลุ่มแล้ว คำนวณหาค่าจุดศูนย์กลางของแต่ละกลุ่มใหม่ตามข้อมูลที่ถูกจัดในกลุ่ม
6. ตรวจสอบเงื่อนไข ถ้า $|w' - w| < \varepsilon$ หยุดการทำงาน และ ถ้าไม่ใช่ ให้กำหนดค่า $w = w'$ แล้วทำซ้ำในข้อ 2-6

การทำงานตามขั้นตอนวิธี CKM สามารถสรุปได้ดังรูปที่ 4.2 เช่นเดียวกับวิธี SKM ส่วนหมายเลข 1-5 ในแผนภาพการทำงานของขั้นตอนวิธี CKM อธิบายถึงขั้นตอนวิธีใน 4.1.2 ที่แตกต่างจากการงานวิจัยที่ได้ศึกษาขั้นตอนวิธี CKE



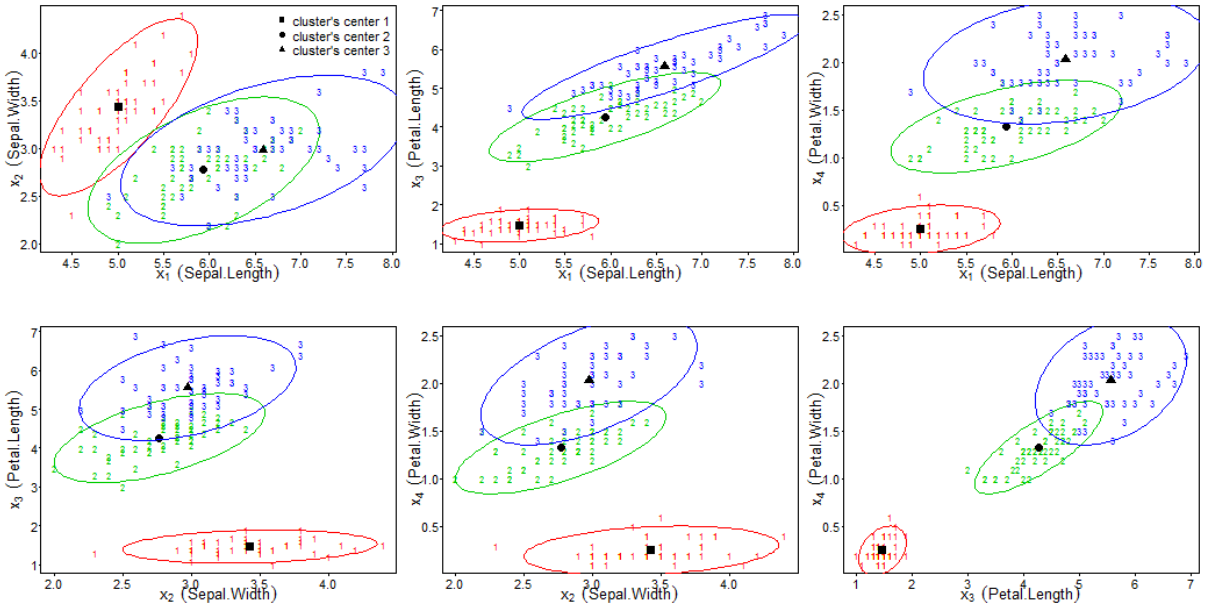
รูปที่ 4.2 แผนภาพการทำงานของขั้นตอนวิธี CKM

4.2 ลักษณะของชุดข้อมูล

จากที่อธิบายในหัวข้อ 3.1 ว่าได้นำชุดข้อมูล 5 ชุดข้อมูลใช้ในการทดลอง ในหัวข้อนี้ได้วิเคราะห์ลักษณะของข้อมูลทั้ง 5 โดยดูจากการกระจายของข้อมูลและระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มของตัวแปรทั้งหมด ได้ผลการวิเคราะห์ดังต่อไปนี้

4.2.1 ชุดข้อมูล iris

ชุดข้อมูล iris เป็นชุดข้อมูลมีจำนวน 150 หน่วยตัวอย่าง 4 ตัวแปร และ 3 กลุ่ม ดังรูปที่ 4.3 แสดงการกระจายของชุดข้อมูลระหว่าง 2 ตัวแปรในข้อมูลทั้ง 3 กลุ่ม และตารางที่ 4.1 แสดงค่าระยะห่างระหว่างจุดศูนย์กลางของแต่กลุ่ม โดยคำนวณจากค่าของตัวแปรทั้งหมดที่ปรับค่ามาตรฐานแล้ว จากผลลัพธ์ดังกล่าวแสดงว่าหน่วยตัวอย่างของกลุ่มที่ 2 กับกลุ่มที 3 มีการซ้อนทับกัน ส่วนหน่วยตัวอย่างของกลุ่มที่ 1 มีการแยกออกกันจากกลุ่มที่ 2 และกลุ่มที 3 ผลดังกล่าวสอดคล้องกับค่าระยะห่างระหว่างจุดศูนย์กลางในตารางที่ 4.1 ที่ระยะห่างระหว่างกลุ่มที่ 2 กับกลุ่มที่ 3 มีค่าใกล้กัน



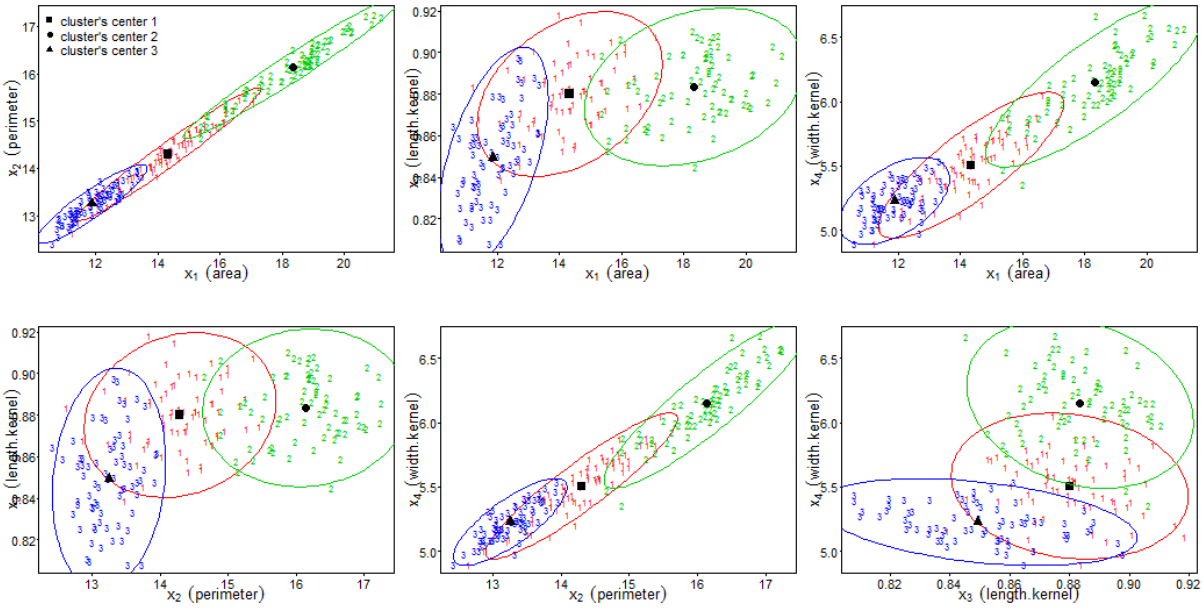
รูปที่ 4.3 แผนภาพการกระจายของชุดข้อมูล iris

ตารางที่ 4.1 ค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างของตัวแปรทั้งหมดโดยใช้เกณฑ์วัดระยะห่าง Euclidean และ Mahalanobis กับชุดข้อมูล iris

จุดศูนย์กลางระหว่าง	เกณฑ์วัดระยะห่าง	
	แบบ Euclidean	แบบ Mahalanobis
กลุ่ม 1 กับ กลุ่ม 2	1.84	10.16
กลุ่ม 1 กับ กลุ่ม 3	2.35	12.99
กลุ่ม 2 กับ กลุ่ม 3	1.30	3.72

4.2.2 ชุดข้อมูล seeds

ชุดข้อมูล seeds เป็นชุดข้อมูลมีจำนวน 210 หน่วยตัวอย่าง 7 ตัวแปร และ 3 กลุ่ม ดังรูปที่ 4.4 แสดงการกระจายของชุดข้อมูลระหว่าง 2 ตัวแปรบางส่วนในข้อมูลทั้ง 3 กลุ่ม และตารางที่ 4.2 แสดงค่าระยะห่างระหว่างจุดศูนย์กลางของแต่กลุ่ม โดยคำนวณจากค่าของตัวแปรทั้งหมดที่ปรับค่ามาตรฐานแล้ว จากดังรูปที่ 4.4 แสดงว่าหน่วยตัวอย่างของกลุ่มที่ 1 มีการซ้อนทับกันกับกลุ่มที่ 2 และกลุ่มที่ 3 ส่วนหน่วยตัวอย่างของกลุ่มที่ 2 มีการแยกออกกันจากกลุ่มที่ 3 และค่าระยะห่างระหว่างจุดศูนย์กลางในตารางที่ 4.2 ระยะห่างระหว่างกลุ่มที่ 1 กับกลุ่มที่ 3 มีค่าใกล้กันกว่าค่าระยะห่างระหว่างจุดศูนย์กลางกลุ่มอื่น ๆ โดยใช้ระยะห่างแบบ Euclidean



รูปที่ 4.4 แผนภาพการกระจายของชุดข้อมูล seeds

ตารางที่ 4.2 ค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างของตัวแปรทั้งหมดโดยใช้เกณฑ์วัดระยะห่าง Euclidean และ Mahalanobis กับชุดข้อมูล seeds

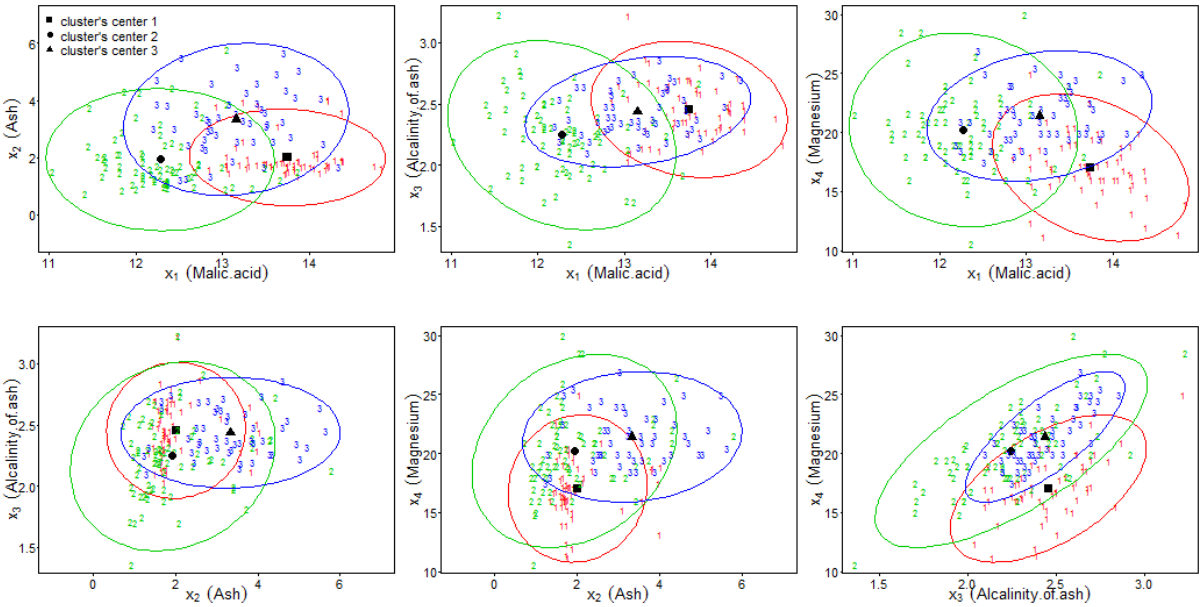
จุดศูนย์กลางระหว่าง	เกณฑ์วัดระยะห่าง	
	แบบ Euclidean	แบบ Mahalanobis
กลุ่ม 1 กับ กลุ่ม 2	2.19	8.48
กลุ่ม 1 กับ กลุ่ม 3	2.11	11.96
กลุ่ม 2 กับ กลุ่ม 3	2.25	59.14

4.2.3 ชุดข้อมูล wine

ชุดข้อมูล seeds เป็นชุดข้อมูลมีจำนวน 210 หน่วยตัวอย่าง 12 ตัวแปร และ 3 กลุ่ม ดังรูปที่ 4.5 แสดงการกระจายของชุดข้อมูลระหว่าง 2 ตัวแปรบางส่วนในข้อมูลทั้ง 3 กลุ่ม และตารางที่ 4.3 แสดงค่าระยะห่างระหว่างจุดศูนย์กลางของแต่กลุ่ม โดยคำนวณจากค่าของตัวแปรทั้งหมดที่ปรับค่ามาตรฐานแล้ว จากดังรูปที่ 4.5 แสดงว่าหน่วยตัวอย่างของแต่ละกลุ่มมีการซ้อนทับกัน และจากการคำนวนค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มโดยใช้ Euclidean โดยค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 1 กับกลุ่มที่ 2 มีค่าเท่ากับ 2.04 ค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 1 กับกลุ่มที่ 3 มีค่าเท่ากับ 2.42 และค่าระยะห่างระหว่างจุดศูนย์กลางของกลุ่มที่ 2 กับกลุ่มที่ 3 มีค่าเท่ากับ 2.26 ดังตารางที่ 4.3 คิดได้ว่าหน่วยตัวอย่างของกลุ่มที่ 1 กับกลุ่มที่ 3 อยู่ใกล้ชิดกันมากกว่าหน่วยตัวอย่างของกลุ่มที่ 1 กับกลุ่มที่ 2 และหน่วยตัวอย่างของกลุ่มที่ 2 กับกลุ่มที่ 3

ตารางที่ 4.3 ค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างของตัวแปรทั้งหมดโดยใช้เกณฑ์วัดระยะห่าง Euclidean และ Mahalanobis กับชุดข้อมูล wine

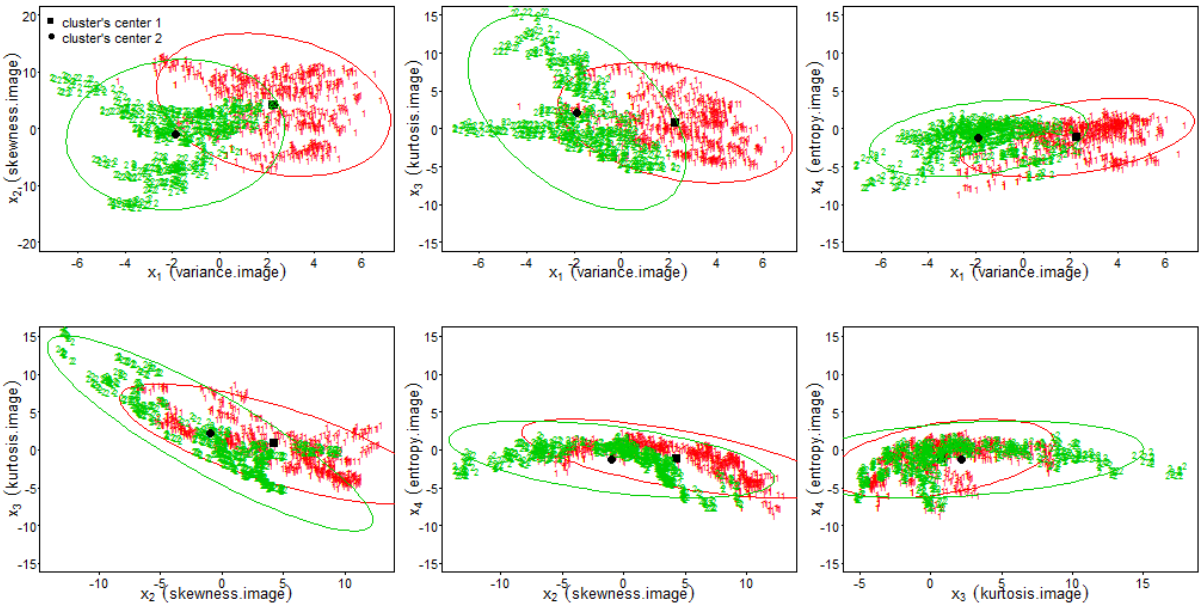
จุดศูนย์กลางระหว่าง	เกณฑ์วัดระยะห่าง	
	แบบ Euclidean	แบบ Mahalanobis
กลุ่ม 1 กับ กลุ่ม 2	2.04	4.84
กลุ่ม 1 กับ กลุ่ม 3	2.42	17.81
กลุ่ม 2 กับ กลุ่ม 3	2.26	13.28



รูปที่ 4.5 แผนภาพการกระจายของชุดข้อมูล wine

4.2.4 ชุดข้อมูล banknote authentication

ชุดข้อมูล banknote authentication เป็นชุดข้อมูลมีจำนวน 1372 หน่วยตัวอย่าง 4 ตัวแปร และ 2 กลุ่ม หน่วยตัวอย่างทั้งสองกลุ่มมีการซ้อนทับกัน ดังรูปที่ 4.6 และจากการคำนวนค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มโดยใช้เกณฑ์วัดระยะห่าง Euclidean โดยค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 1 กับกลุ่มที่ 2 มีค่าเท่ากับ 1.87 ดังตารางที่ 4.4



รูปที่ 4.6 แผนภาพการกระจายของชุดข้อมูล banknote authentication

ตารางที่ 4.4 ค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างของตัวแปรทั้งหมดโดยใช้เกณฑ์วัดระยะห่าง Euclidean และ Mahalanobis กับชุดข้อมูล banknote authentication

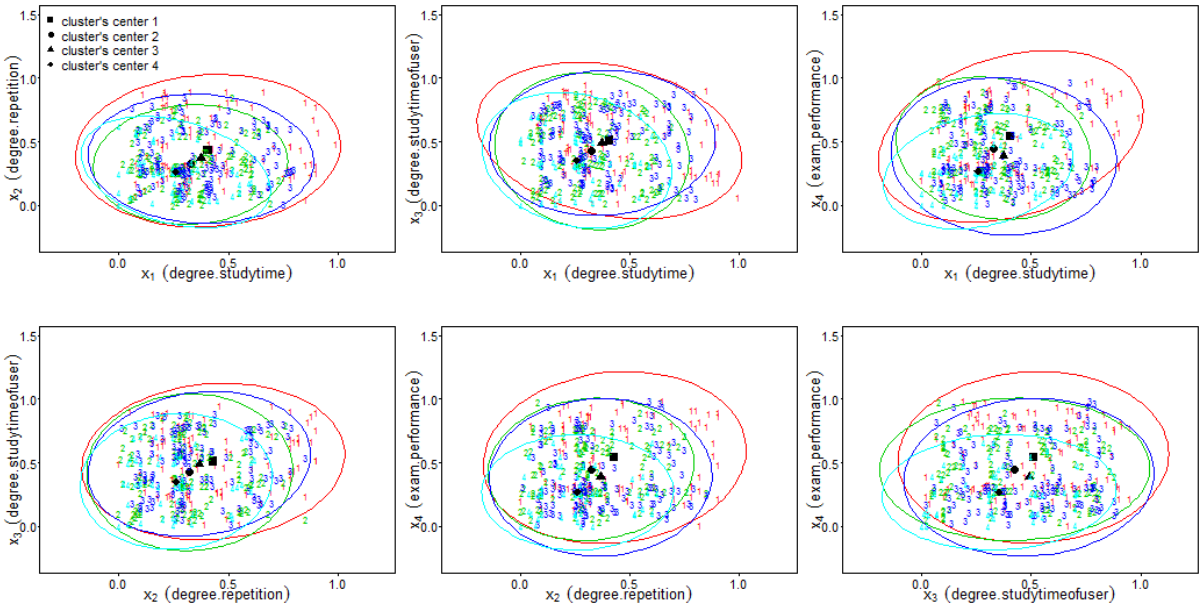
จุดศูนย์กลางระหว่าง	เกณฑ์วัดระยะห่าง	
	แบบ Euclidean	แบบ Mahalanobis
กลุ่ม 1 กับ กลุ่ม 2	2.09	12.65
กลุ่ม 1 กับ กลุ่ม 3	1.21	5.81
กลุ่ม 1 กับ กลุ่ม 4	2.89	14.06
กลุ่ม 2 กับ กลุ่ม 3	1.07	4.27
กลุ่ม 2 กับ กลุ่ม 4	0.95	3.41
กลุ่ม 3 กับ กลุ่ม 4	1.74	8.60

4.2.5 ชุดข้อมูล user knowledge modeling

ชุดข้อมูล user knowledge modeling เป็นชุดข้อมูลมีจำนวน 403 หน่วยตัวอย่าง 5 ตัวแปร และ 4 กลุ่ม ชุดข้อมูลมีหน่วยตัวอย่างแต่ละกลุ่มน่าจะซ้อนทับกัน ดังรูปที่ 4.7 และจากการคำนวนค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างโดยใช้เกณฑ์วัดระยะห่าง Euclidean โดยค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 1 กับกลุ่มที่ 2 มีค่าเท่ากับ 2.09 ค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 1 กับกลุ่มที่ 3 มีค่าเท่ากับ 1.21 ค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 1 กับกลุ่มที่ 4 มีค่าเท่ากับ 2.89 ค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 2 กับกลุ่มที่ 3 มีค่าเท่ากับ 1.07 ค่าระยะห่างระห่างจุดศูนย์กลางของกลุ่มที่ 2 กับกลุ่มที่ 4 มีค่าเท่ากับ 0.95 ค่าระยะห่างระหว่างจุดศูนย์กลางของกลุ่มที่ 3 กับกลุ่มที่ 4 มีค่าเท่ากับ 1.74 ดังตารางที่ 4.5

ตารางที่ 4.5 ค่าระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างของตัวแปรทั้งหมดโดยใช้เกณฑ์วัดระยะห่าง Euclidean และ Mahalanobis กับชุดข้อมูล user knowledge modeling

จุดศูนย์กลางระหว่าง	เกณฑ์วัดระยะห่าง	
	แบบ Euclidean	แบบ Mahalanobis
กลุ่ม 1 กับ กลุ่ม 2	1.87	7.46

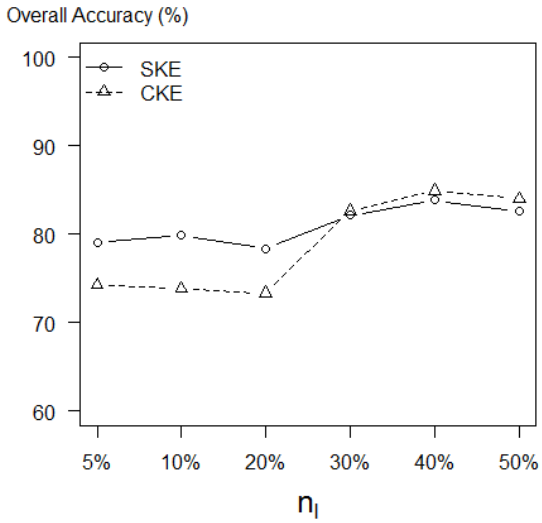


รูปที่ 4.7 แผนภาพการกระจายของชุดข้อมูล user knowledge modeling

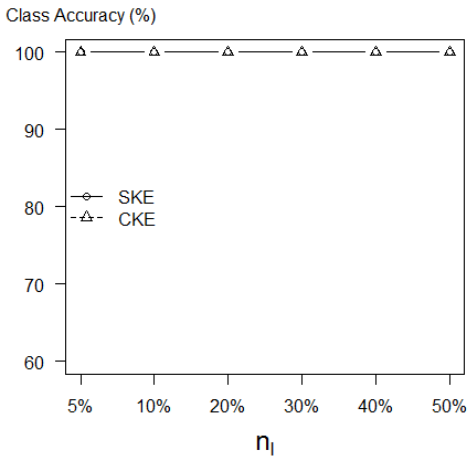
4.3 ผลการทดลองที่ 1 การใช้เกณฑ์วัดระยะห่างแบบ Euclidean

การวัดประสิทธิภาพการจัดกลุ่มข้อมูลของขั้นตอนวิธี seeded K-means Euclidean (SKE) กับ constrained K-means Euclidean (CKE) คือ ค่าความถูกต้องรวม (overall accuracy) และค่าความถูกต้องของกลุ่ม (class accuracy) มีผลลัพธ์แสดงตามการทดลองกับชุดข้อมูล 5 ชุดดังนี้

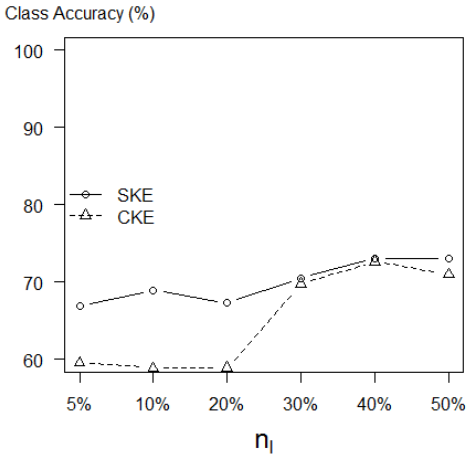
4.3.1 ชุดข้อมูล iris



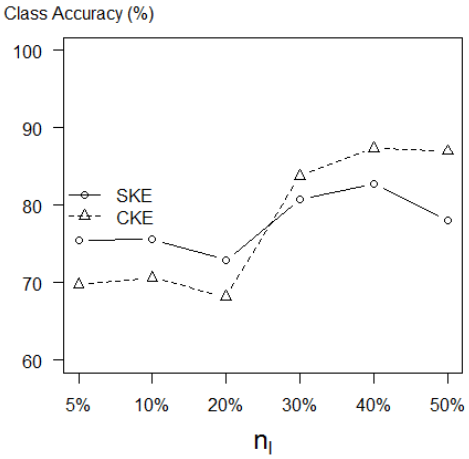
รูปที่ 4.8 ค่า overall accuracy ของวิธี SKE กับ CKE ของชุดข้อมูล iris



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



(c) class accuracy ของกลุ่มที่ 3

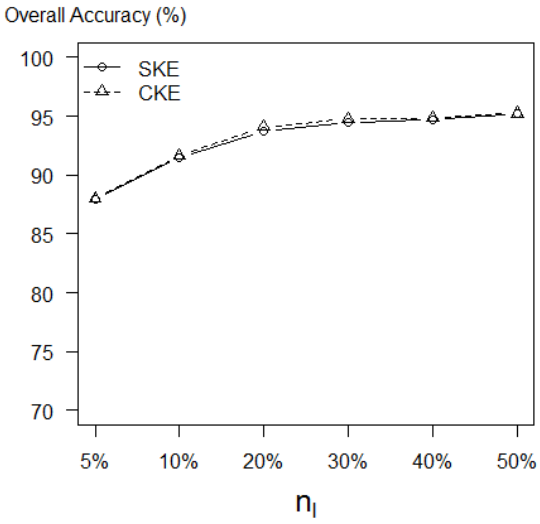
รูปที่ 4.9 (a)-(c) class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล iris

จากรูปที่ 4.8 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล iris ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% กรณีใช้จำนวนข้อมูลที่กำกับกลุ่มน้อยกว่าหรือเท่ากับ 20% ค่า overall accuracy ทั้งสองวิธีไม่ได้เพิ่มขึ้น แต่ค่า overall accuracy ของวิธี CKE มีค่าน้อยกว่าวิธี SKE โดยวิธี CKE มีค่าประมาณ 73.33% ถึง 74.27% ส่วนวิธี SKE มีค่าประมาณ 78.40% ถึง 79.87% ส่วนกรณีใช้จำนวนข้อมูลที่กำกับกลุ่มมากกว่า 20% ค่า overall accuracy ทั้งสองวิธีเพิ่มขึ้น และค่า overall

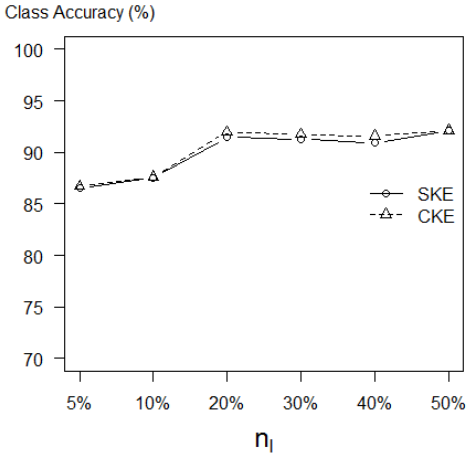
accuracy ทั้งสองวิธี มีค่าใกล้เคียงกัน โดยวิธี CKE มีค่าประมาณ 82.67% ถึง 84.93% และวิธี SKE มีค่าประมาณ 82.13% ถึง 83.87%

จากรูปที่ 4.9 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล iris ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% โดยค่า class accuracy ของกลุ่มที่ 1 ทั้งสองวิธีมีค่าเท่ากับ 100% ทุกกรณี ส่วนค่า class accuracy ของกลุ่มที่ 2 กรณีที่ใช้ชุดข้อมูลกำกับกลุ่มน้อยกว่าหรือเท่ากับ 20% วิธี CKE มีค่าน้อยกว่าวิธี SKE และกรณีใช้จำนวนข้อมูลที่กำกับกลุ่มมากกว่า 20% ค่า class accuracy ของทั้งสองวิธีมีค่าใกล้เคียงกัน ส่วนค่า class accuracy ของกลุ่มที่ 3 กรณีที่ใช้ชุดข้อมูลกำกับกลุ่มน้อยกว่าหรือเท่ากับ 20% วิธี CKE มีค่าน้อยกว่าวิธี SKE และกรณีใช้จำนวนข้อมูลที่กำกับกลุ่มมากกว่า 20% ค่า class accuracy ของวิธี CKE มีค่ามากกว่าวิธี SKE

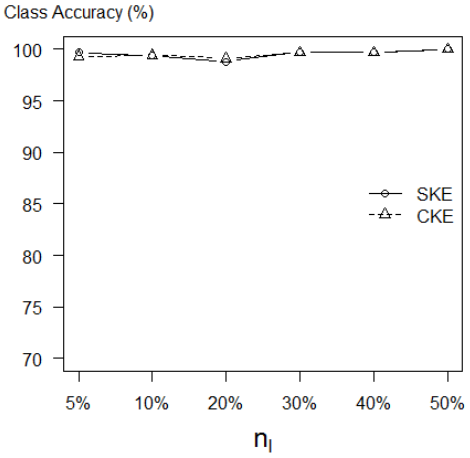
4.3.2 ชุดข้อมูล seeds



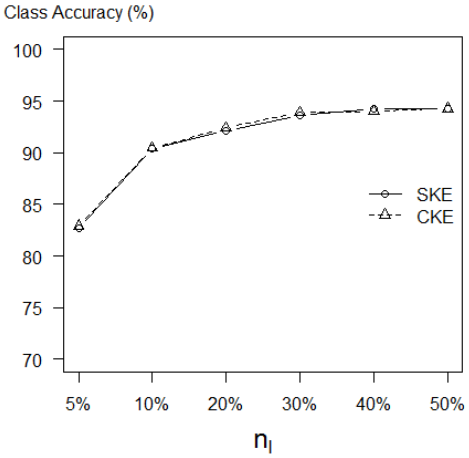
รูปที่ 4.10 ค่า overall accuracy ของวิธี SKE กับ CKE ของชุดข้อมูล seeds



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



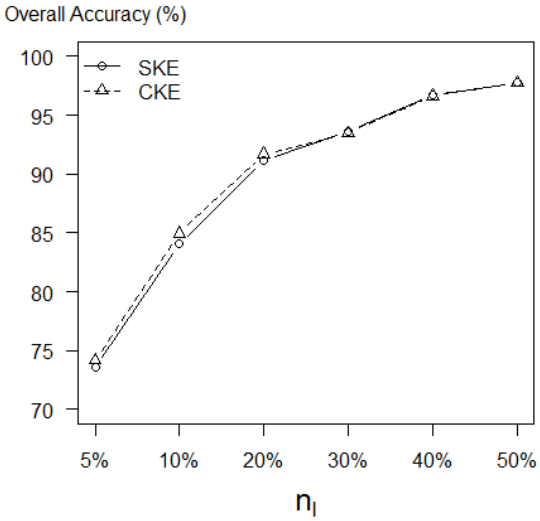
(c) class accuracy ของกลุ่มที่ 3 ของชุดข้อมูล seeds

รูปที่ 4.11 (a)-(c) class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล seeds

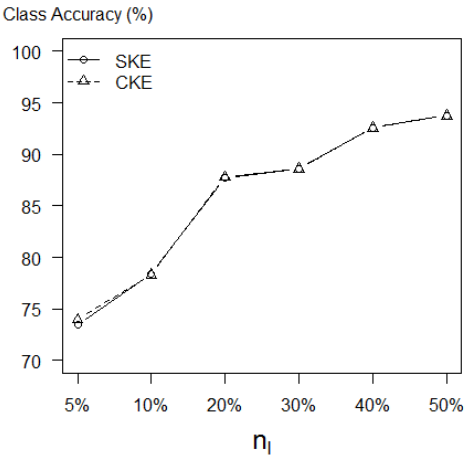
จากรูปที่ 4.10 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล seeds ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 5% 10% และ 20 % ค่า overall accuracy มีการเพิ่มขึ้นอย่างชัดเจน โดยทั้งสองวิธีมีค่า overall accuracy ประมาณ 87.91% 91.53% และ 94.10% ตามลำดับ และเมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 30% 40% และ 50 % ค่า overall accuracy เพิ่มขึ้นเล็กน้อยโดยมีค่า overall accuracy ประมาณ 94.48% 94.67% และ 95.24% ตามลำดับ

จากรูปที่ 4.11 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล seeds ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% ค่า class accuracy ในการแบ่งกลุ่มที่ 1 กลุ่มที่ 2 และกลุ่มที่ 3 ของทั้งสองวิธีมีค่าไม่แตกต่างกันในทุกกรณี

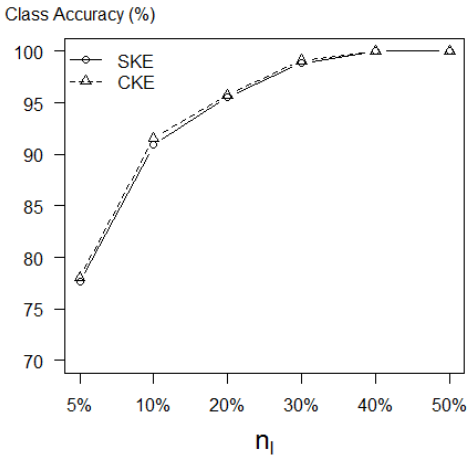
4.3.3 ชุดข้อมูล wine



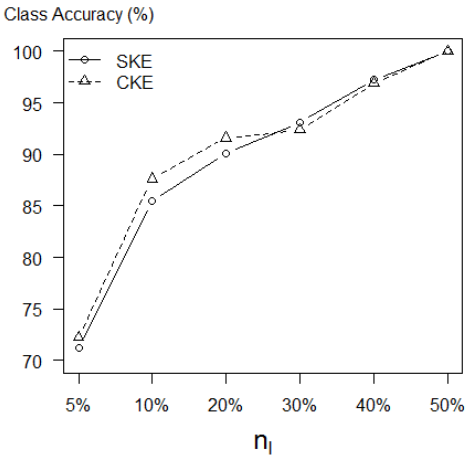
รูปที่ 4.12 ค่า overall accuracy ของวิธี SKE กับ CKE ของชุดข้อมูล wine



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



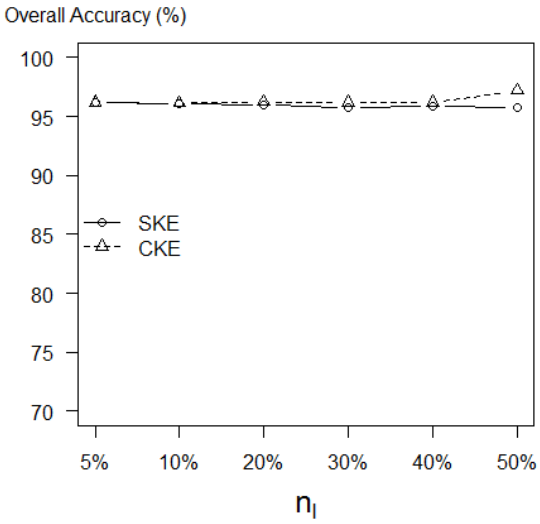
(c) class accuracy ของกลุ่มที่ 3

รูปที่ 4.13 (a)-(c) class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล wine

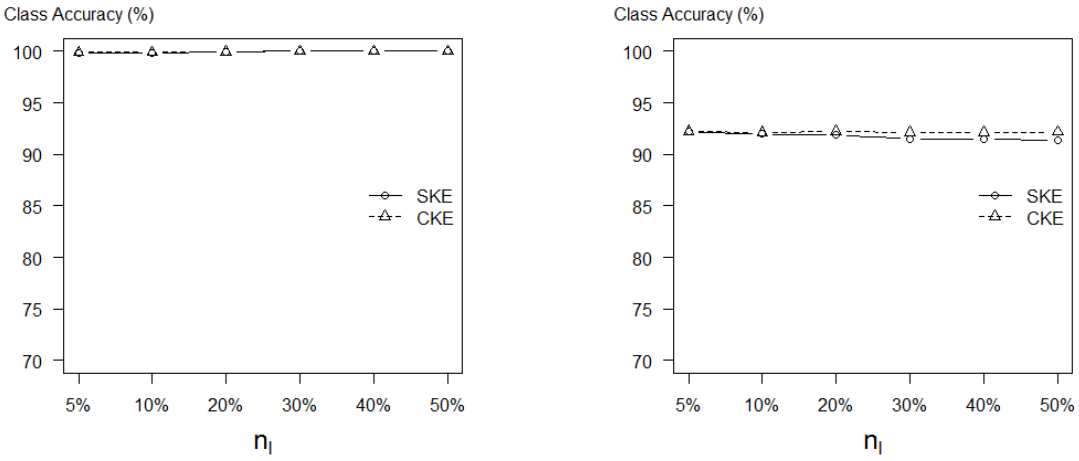
จากรูปที่ 4.12 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล wine ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 5% 10% 20% และ 30 % ค่า overall accuracy มีการเพิ่มขึ้นอย่างชัดเจน โดยทั้งสองวิธีมีค่า overall accuracy ประมาณ 73.60% 84.04% 91.12% และ 93.60% ตามลำดับ และเมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 40% และ 50 % ค่า overall accuracy เพิ่มขึ้นเล็กน้อยโดยมีค่า overall accuracy ทั้งสองวิธีประมาณ 96.64% และ 97.75% ตามลำดับ

จากรูปที่ 4.13 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล wine ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% ค่า class accuracy ในการแบ่งกลุ่มที่ 1 กลุ่มที่ 2 และกลุ่มที่ 3 ของทั้งสองวิธีมีค่าไม่แตกต่างกันในทุกกรณี

4.2.4 ชุดข้อมูล banknote authentication



รูปที่ 4.14 ค่า overall accuracy ของวิธี SKE กับ CKE ของชุดข้อมูล banknote authentication



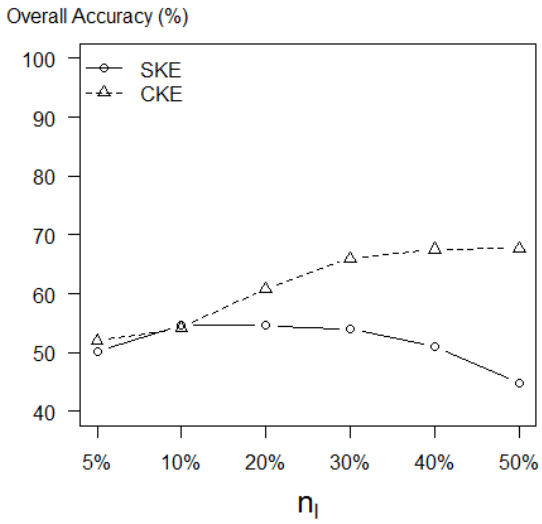
(a) class accuracy ของกลุ่มที่ 1 (b) class accuracy ของกลุ่มที่ 2

รูปที่ 4.15 (a)-(b) class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล banknote authentication

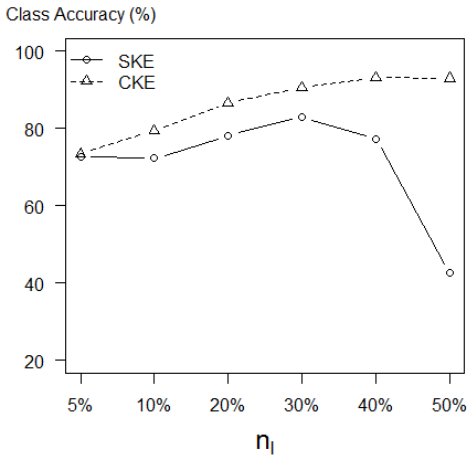
จากรูปที่ 4.14 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล banknote authentication ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าการเพิ่มจำนวนชุดข้อมูลที่กำกับกลุ่ม ไม่มีผลต่อการเพิ่มขึ้นของค่า overall accuracy ทั้งสองวิธีและค่า overall accuracy ทั้งสองวิธีก็ไม่แตกต่างกัน โดยมีค่า overall accuracy ทั้งสองวิธีประมาณ 96% ทุกกรณีที่ใช้จำนวนข้อมูลที่กำกับกลุ่ม

จากรูปที่ 4.15 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล banknote authentication ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% ค่า class accuracy ในการแบ่งกลุ่มที่ 1 และกลุ่มที่ 2 ของทั้งสองวิธีมีค่าไม่แตกต่างกันในทุกกรณี

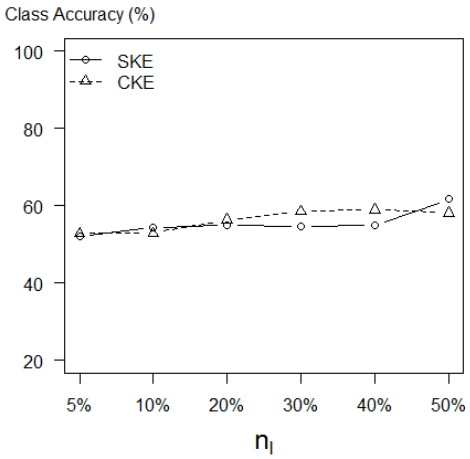
4.3.4 ชุดข้อมูล user knowledge modeling



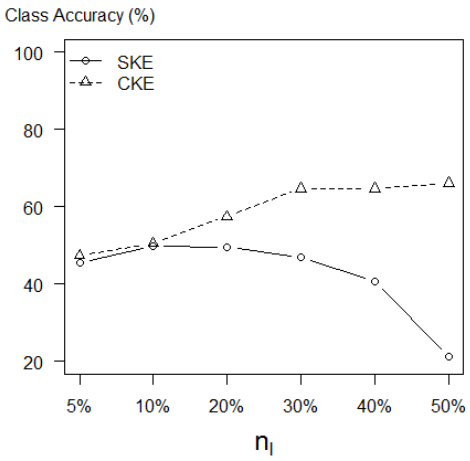
รูปที่ 4.16 ค่า overall accuracy ของวิธี SKE กับ CKE ของชุดข้อมูล user knowledge modeling



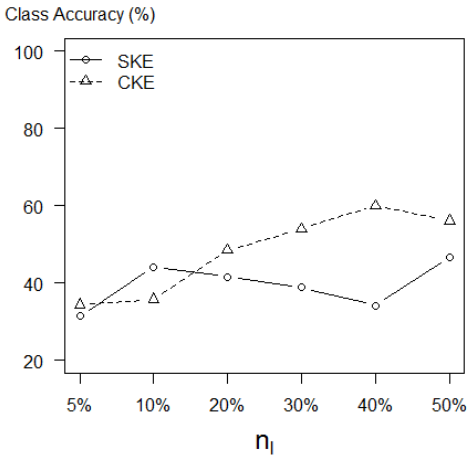
(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



(c) class accuracy ของกลุ่มที่ 3



(d) class accuracy ของกลุ่มที่ 4

รูปที่ 4.17 (a)-(b) class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล user knowledge modeling

จากรูปที่ 4.16 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล user knowledge modeling ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 5% 10% 20% และ 30% ค่า overall accuracy จากวิธี CKE มีการเพิ่มขึ้นอย่างชัดเจน โดยมีค่าเท่ากับ 52.04% 54.18% 60.84% และ 65.97% ตามลำดับ และเมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 40% และ 50 % ค่า overall accuracy จากวิธี CKE มีการเพิ่มขึ้นเล็กน้อย โดยมีค่าเท่ากับ 67.51% และ 67.66% ตามลำดับ ส่วนค่า overall accuracy ของ

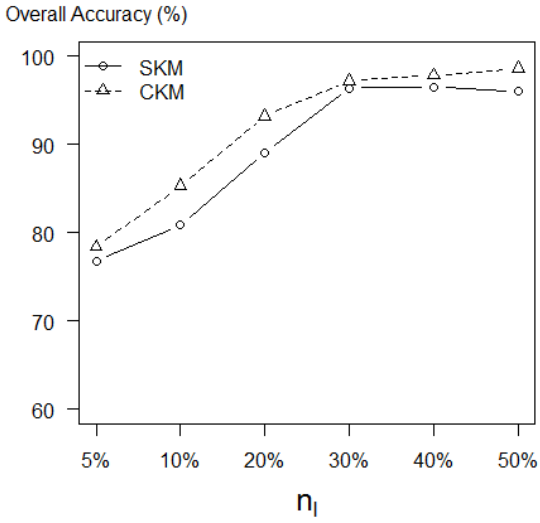
วิธี SKE มีแนวลดลง เมื่อใช้เท่ากับ 20% 30% 40% และ 50% ค่า overall accuracy ของวิธี SKE มีการลดลง โดยมีค่าเท่ากับ 54.57% 53.93% 50.94% และ 44.78% ตามลำดับ

จากรูปที่ 4.17 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล user knowledge modeling ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่า ค่า class accuracy ในการแบ่งกลุ่มที่ 1 กลุ่มที่ 3 และกลุ่มที่ 4 ของวิธี CKE ส่วนใหญ่มีค่ามากกว่าวิธี SKE ยกเว้นค่า class accuracy ของกลุ่มที่ 2 ที่ทั้งสองวิธีมีค่าใกล้เคียงกัน

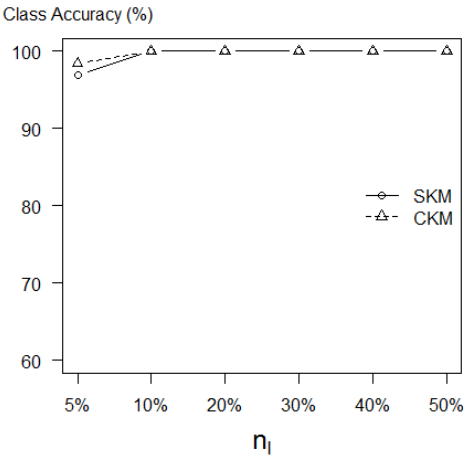
4.4 ผลการทอลองที่ 2 การใช้เกณฑ์วัดระยะห่างแบบ Mahalanobis

การวัดประสิทธิภาพการจัดกลุ่มข้อมูลของขั้นตอนวิธี seeded K-means Mahalanobis (SKM) กับ constrained K-means Mahalanobis (CKM) คือ ค่าความถูกต้องรวม (overall accuracy) และ ค่าความถูกต้องของกลุ่ม (class accuracy) มีผลลัพธ์แสดงตามการทดลองกับชุดข้อมูล 5 ชุดดังนี้

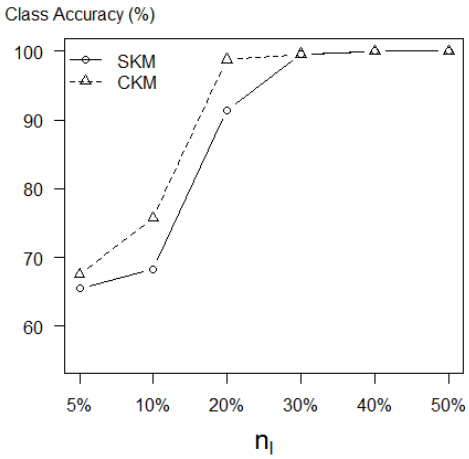
4.4.1 ชุดข้อมูล iris



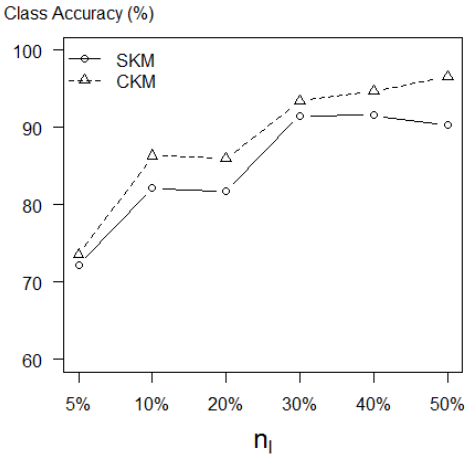
รูปที่ 4.18 ค่า overall accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล iris



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



(c) class accuracy ของกลุ่มที่ 3

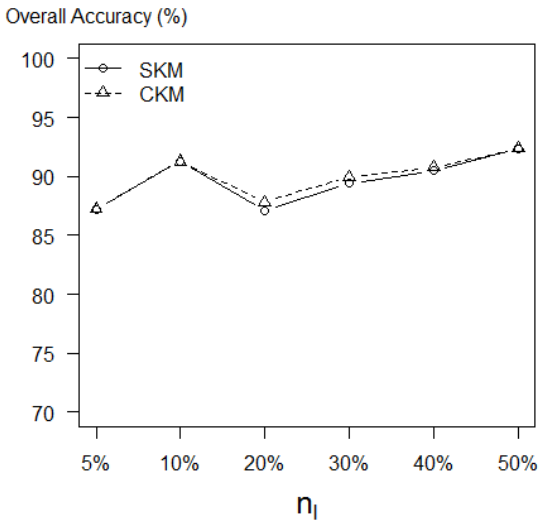
รูปที่ 4.19 (a)-(b) class accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล iris

จากรูปที่ 4.18 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล iris ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าการเพิ่มจำนวนข้อมูลที่กำกับกลุ่ม ทำให้ค่า overall accuracy มีแนวโน้มเพิ่มขึ้นทั้งสองวิธี โดยส่วนใหญ่ ค่า overall accuracy ของวิธี CKM มีค่ามากกว่าวิธี SKM ยกเว้นกรณีใช้จำนวนข้อมูลที่กำกับกลุ่ม 5% 30% และ 40% ค่า overall accuracy ของทั้งสองวิธีมีค่าใกล้เคียงกัน และลักษณะการเพิ่มขึ้นค่า overall accuracy ทั้งสองวิธีในช่วงแรกและช่วงหลังการใช้ชุดข้อมูลกำกับกลุ่ม 30% แตกต่างกัน เมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 5% 10% และ 20% ค่า overall accuracy ของวิธี CKM มีการเพิ่มขึ้นอย่างชัดเจน โดยมีค่าเท่ากับ 78.40% 85.33% และ 93.20% ตามลำดับ ส่วนค่า overall accuracy ของวิธี SKM มีค่าเท่ากับ 77.07% 80.93% และ 89.07% ตามลำดับ เมื่อใช้ชุดข้อมูลที่

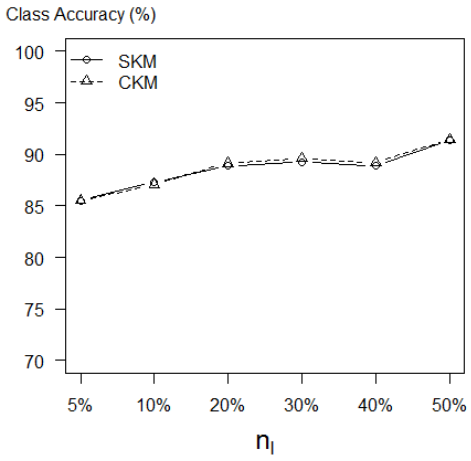
กำกับกลุ่มเท่ากับ 30% 40% และ 50% ค่า overall accuracy ของวิธี CKM มีการเพิ่มขึ้นเล็กน้อย โดยมีค่าประมาณ 97.20% ถึง 98.67% ส่วนค่า overall accuracy ของวิธี SKM มีค่าไม่แตกต่างกัน โดยประมาณ 96% ถึง 96.53%

จากรูปที่ 4.19 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล iris ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% ค่า class accuracy ของกลุ่มที่ 1 ทั้งสองวิธีมีค่าเท่ากันคือ 100% ยกเว้นกรณี ใช้จำนวนข้อมูลที่กำกับกลุ่ม 5% ส่วนค่า class accuracy ของกลุ่มที่ 2 กรณีที่ใช้ชุดข้อมูลกำกับกลุ่มน้อยกว่าหรือเท่ากับ 20% วิธี CKM มีค่ามากกว่าวิธี SKM กรณีที่ใช้ชุดข้อมูลกำกับกลุ่มมากกว่า 20% ทั้งสองวิธีไม่แตกต่างกัน ส่วนค่า class accuracy ของกลุ่มที่ 3 วิธี CKM มีค่ามากกว่าวิธี SKM ยกเว้นกรณี ใช้จำนวนข้อมูลที่กำกับกลุ่ม 5% และ 30% ทั้งสองวิธีมีค่าใกล้เคียงกัน

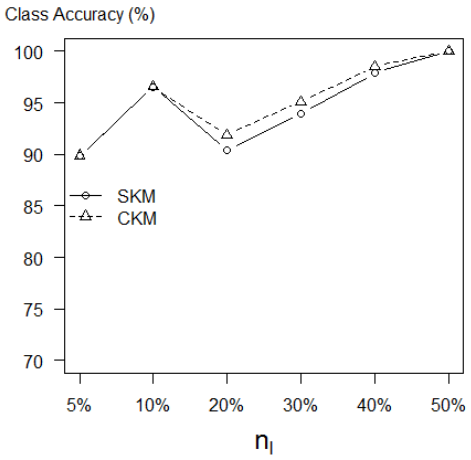
4.4.2 ชุดข้อมูล seeds



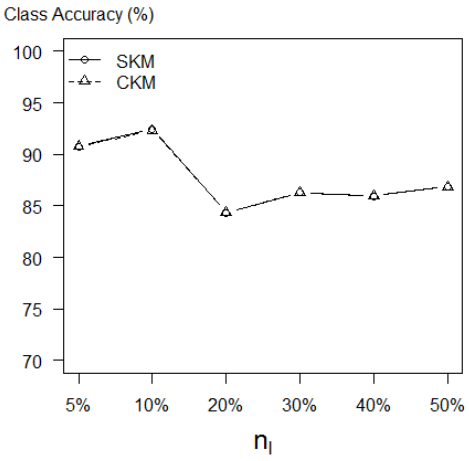
รูปที่ 4.20 ค่า overall accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล seeds



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



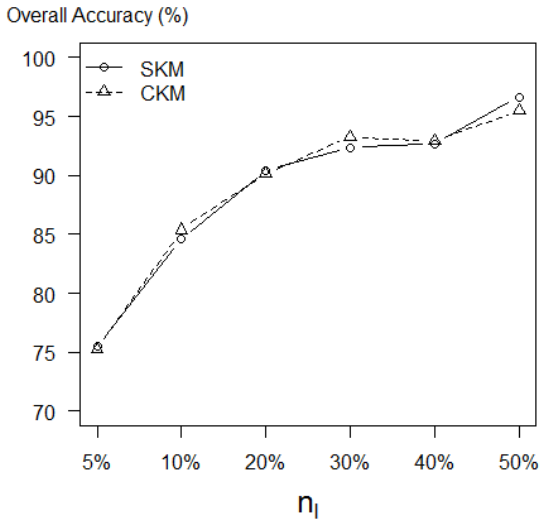
(c) class accuracy ของกลุ่มที่ 3

รูปที่ 4.21 (a)-(c) class accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล seeds

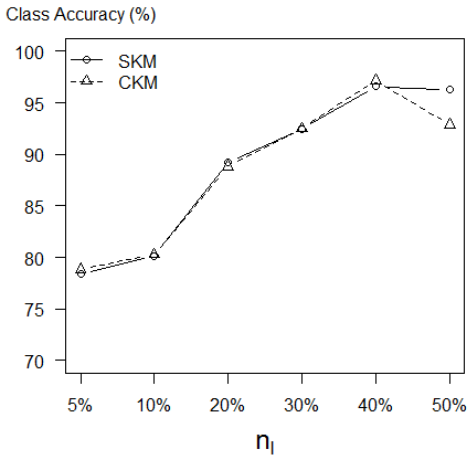
จากรูปที่ 4.20 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล seeds ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% ในทุกกรณีค่า overall accuracy ทั้งสองวิธีมีค่าไม่แตกต่างกัน กรณีใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 5% และ 10% มีค่าเพิ่มสูงขึ้น โดยวิธี CKM มีค่าเท่ากับ 87.24% และ 91.24% ตามลำดับ และวิธี SKM มีค่าเท่ากับ 87.24% และ 91.24% ตามลำดับ กรณีใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 20% ทั้งสองวิธีมีค่าลดลง โดยวิธี CKM มีค่าเท่ากับ 87.81% และวิธี SKM มีค่าเท่ากับ 87.05% แต่กรณีใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 30% 40% และ 50% มีแนวโน้นเพิ่มขึ้นทั้งสองวิธี โดยวิธี CKM มีค่าเท่ากับ 89.91% 90.76% และ 92.38% ตามลำดับ และวิธี SKM มีค่าเท่ากับ 89.43% 90.48% และ 92.38% ตามลำดับ

จากรูปที่ 4.21 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล seeds ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าในค่า class accuracy ของกลุ่มที่ 1 กลุ่มที่ 2 และกลุ่มที่ 3 ทั้งสองวิธีมีค่าไม่แตกต่างกัน

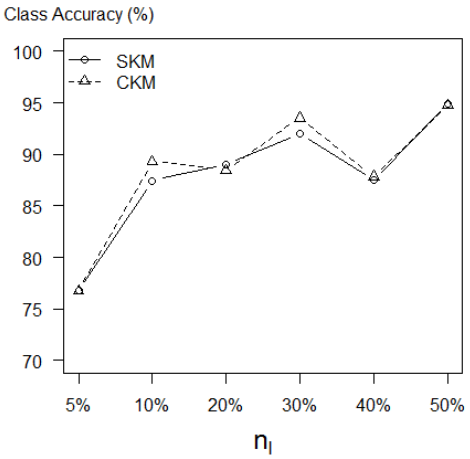
4.4.3 ชุดข้อมูล wine



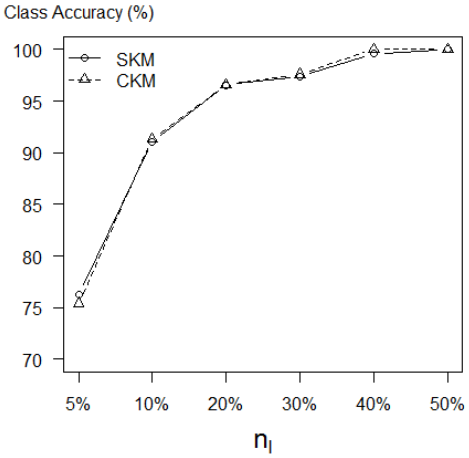
รูปที่ 4.22 ค่า overall accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล wine



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



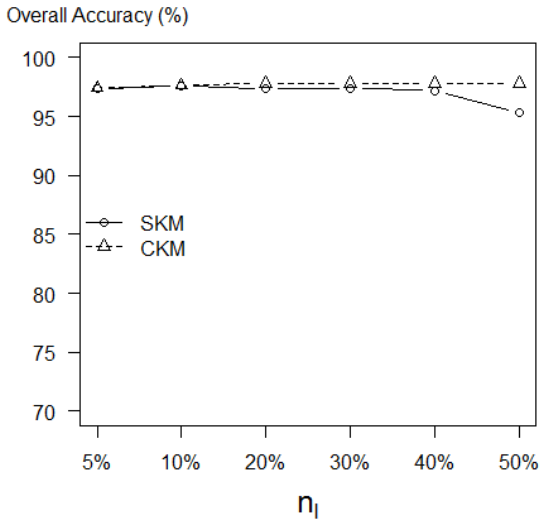
(c) class accuracy ของกลุ่มที่ 3

รูปที่ 4.23 (a)-(c) class accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล wine

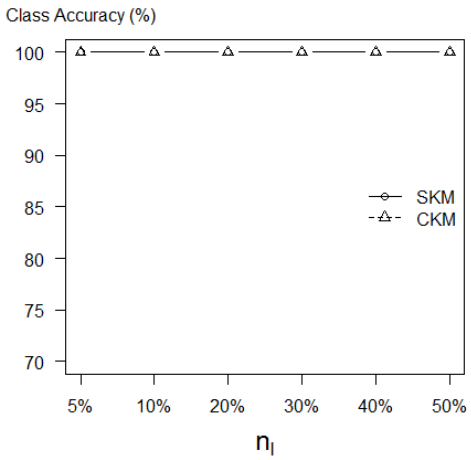
จากรูปที่ 4.22 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล wine ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าการเพิ่มจำนวนชุดข้อมูลที่กำกับกลุ่ม 5% 10% 20% 30% 40% และ 50% ทั้งสองวิธีที่ให้ค่า overall accuracy ไม่แตกต่างกันและมีแนวโน้มเพิ่มขึ้นเช่นเดียวกัน โดยทั้งสองวิธีมีค่าประมาณ 75.28% 84.61% 90.11% 92.36% 92.93% และ 96.63% ตามลำดับ

จากรูปที่ 4.23 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล wine ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าในทุกกรณีที่ค่า class accuracy ของกลุ่มที่ 1 กลุ่มที่ 2 และกลุ่มที่ 3 ของทั้งสองวิธีมีค่าไม่แตกต่างกัน ยกเว้นกรณีที่ใช้จำนวนข้อมูลที่กำกับกลุ่ม 50% ของค่า class accuracy กลุ่มที่ 1 ของวิธี SKM มีค่ามากกว่าวิธี CKM

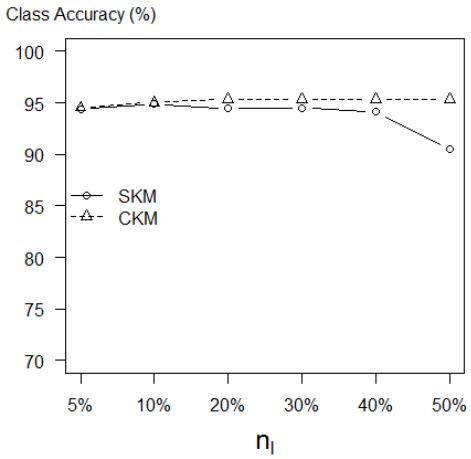
4.4.4 ชุดข้อมูล banknote authentication



รูปที่ 4.24 ค่า overall accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล banknote authentication



(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2

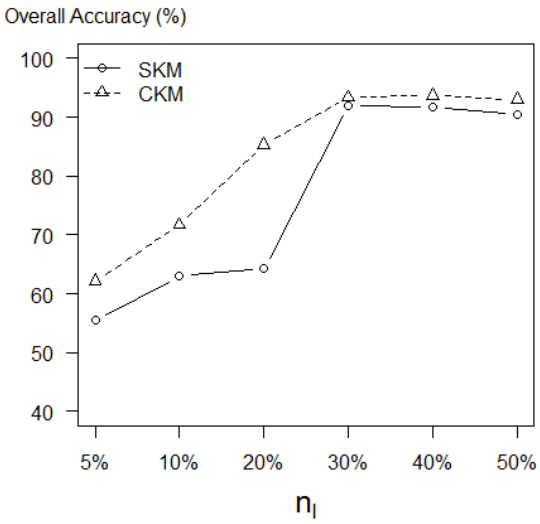
รูปที่ 4.25 (a)-(b) class accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล banknote authentication

จากรูปที่ 4.24 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล banknote authentication ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าการเพิ่มจำนวนชุดข้อมูลที่กำกับกลุ่ม ไม่มีผลต่อการเพิ่มขึ้นของค่า overall accuracy จากทั้งสองวิธีและค่า overall accuracy ทั้งสองวิธีก็มีค่าใกล้เคียงกัน โดยมีค่า overall accuracy ทั้งสองวิธี

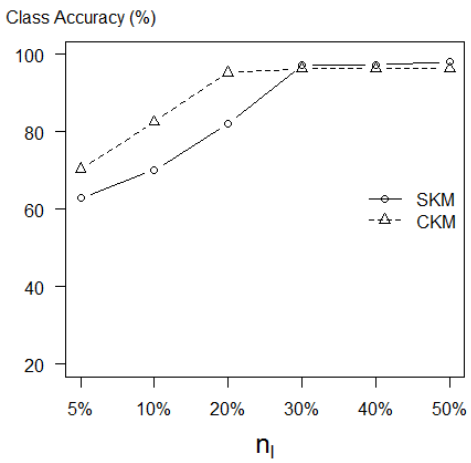
ประมาณ 97.41% ยกเว้นกรณีที่ใช้จำนวนข้อมูลที่กำกับกลุ่ม 50% วิธี SKM มีค่าลดลงเท่ากับ 95.34%

จากรูปที่ 4.25 เป็นการวัดค่า class accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล banknote authentication ของวิธี CKE และวิธี SKE ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าทุกกรณีที่ใช้จำนวนข้อมูลที่กำกับกลุ่มค่า class accuracy ของกลุ่มที่ 1 และกลุ่มที่ 2 ทั้งสองวิธีมีค่าไม่แตกต่างกันยกเว้นกรณีที่ใช้จำนวนข้อมูลที่กำกับกลุ่ม 50% ค่า class accuracy ของกลุ่มที่ 2 วิธี CKE มีค่ามากว่าวิธี SKM

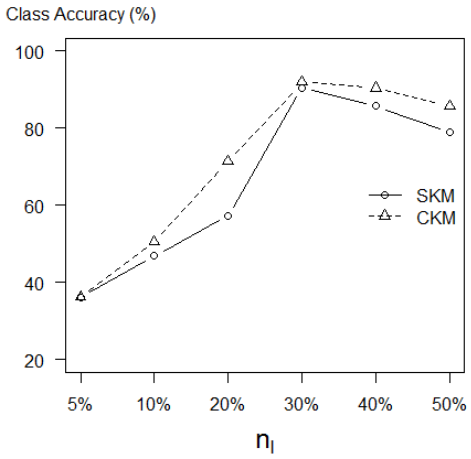
4.4.5 ชุดข้อมูล user knowledge modeling



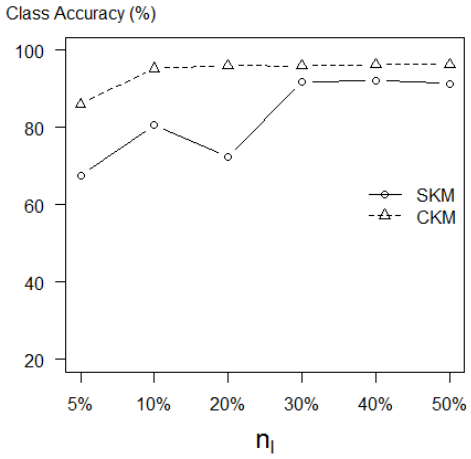
รูปที่ 4.26 ค่า overall accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล user knowledge modeling



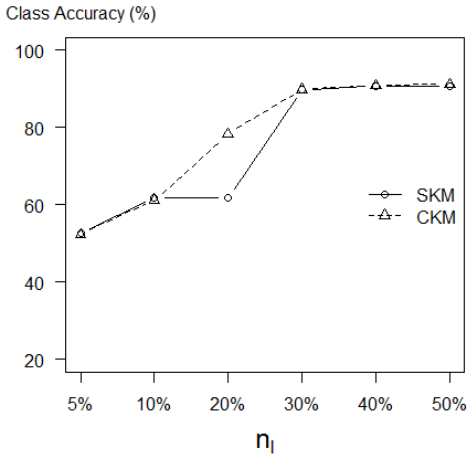
(a) class accuracy ของกลุ่มที่ 1



(b) class accuracy ของกลุ่มที่ 2



(c) class accuracy ของกลุ่มที่ 3



(d) class accuracy ของกลุ่มที่ 4

รูปที่ 4.27 (a)-(b) class accuracy ของวิธี SKM กับวิธี CKM ชุดข้อมูล user knowledge modeling

จากรูปที่ 4.26 เป็นการวัดค่า overall accuracy การจัดกลุ่มข้อมูลของชุดข้อมูล user knowledge modeling ของวิธี CKM และวิธี SKM ที่ใช้จำนวนข้อมูลที่กำกับกลุ่มจาก 5% ถึง 50% เห็นได้ว่าการเพิ่มจำนวนชุดข้อมูลที่กำกับกลุ่ม ทำให้ค่า overall accuracy ของทั้งสองวิธีมีแนวโน้ใเพิ่มขึ้น โดยค่า overall accuracy ของวิธี CKM มีค่ามากกว่าวิธี SKM ยกเว้นกรณีใช้จำนวนข้อมูลที่กำกับกลุ่ม 30% 40% และ 50% ค่า overall accuracy ทั้งสองวิธีมีค่าใกล้เคียงกัน ลักษณะการเพิ่มขึ้นของค่า overall accuracy ช่วงแรกใช้ชุด

ข้อมูลที่กำกับกลุ่มเท่ากับ 5% 10% และ 20% ค่า overall accuracy ของวิธี CKM มีการเพิ่มขึ้นอย่างชัดเจน โดยมีค่าเท่ากับ 62.14% 71.74% และ 85.32% ตามลำดับ ส่วนค่า overall accuracy ของวิธี SKM มีการเพิ่มขึ้นอย่างชัดเจน โดยมีค่าเท่ากับ 55.42% 62.98% และ 64.28% ตามลำดับ เมื่อใช้ชุดข้อมูลที่กำกับกลุ่มเท่ากับ 30% 40% และ 50% ค่า overall accuracy ของวิธี CKM มีค่าไม่แตกต่างกัน โดยมีค่าประมาณ 93.03% ส่วนค่า overall accuracy ของวิธี SKM มีค่า overall accuracy ไม่แตกต่างกันและมีค่าประมาณ 91.74%

4.5 สรุปผลการทดลอง

4.5.1 สรุปผลการทดลองที่ 1

จากการทดลองกับข้อมูล 5 ชุดดังกล่าว สังเกตพบว่าการเพิ่มจำนวนข้อมูลที่กำกับกลุ่มประสิทธิภาพการจัดกลุ่มข้อมูลของวิธี CKE และวิธี SKE มีค่าเพิ่มขึ้นและค่า overall accuracy ส่วนใหญ่ได้ค่าใกล้เคียงกัน และจำนวนชุดข้อมูลที่กำกับกลุ่ม 30% ก็เพียงพอต่อการจัดกลุ่มข้อมูลข้างต้น ยกเว้นชุดข้อมูล user knowledge modeling การจัดกลุ่มข้อมูลด้วยวิธี SKE เมื่อเพิ่มจำนวนข้อมูลที่กำกับกลุ่ม ค่า overall accuracy มีค่าลดลง

4.5.2 สรุปผลการทดลองที่ 2

จากการทดลองกับข้อมูล 5 ชุดดังกล่าว สังเกตพบว่าการเพิ่มจำนวนข้อมูลที่กำกับกลุ่มประสิทธิภาพการจัดกลุ่มข้อมูลของวิธี CKM และวิธี SKM มีค่าเพิ่มขึ้นและค่า overall accuracy ด้วยวิธีการจัดกลุ่มข้อมูล CKM จะได้ผลการจัดกลุ่มข้อมูลส่วนใหญ่ได้ถูกต้องกว่าวิธี SKM และจำนวนชุดข้อมูลที่กำกับกลุ่ม 30% ก็เพียงพอต่อการจัดกลุ่มข้อมูลข้างต้น

4.6 อภิปรายผล

จากการทดลองกับข้อมูล 5 ชุดดังกล่าว สังเกตพบว่าชุดข้อมูลที่ระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มหน่วยตัวอย่างมีค่าน้อย ๆ บ่งบอกถึงหน่วยตัวอย่างของแต่ละกลุ่มนั้น ๆน่าจะมีการซ้อนทับกันของหน่วยตัวอย่าง เช่น ชุดข้อมูล user knowledge modeling ค่าระยะห่างระหว่างจุดศูนย์กลางของกลุ่มที่ 2 กับกลุ่มที่ 3 มีค่าเท่ากับ 1.07 และกลุ่มที่ 2 กับกลุ่มที่ 4 มีค่าเท่ากับ 0.95 และชุดข้อมูล iris ค่าระยะห่างระหว่างจุดศูนย์กลางของกลุ่มที่ 2 กับกลุ่มที่ 3 มีค่าเท่ากับ 1.30 แสดงว่าชุดข้อมูลดังกล่าว มีหน่วยตัวอย่างแต่ละกลุ่มน่าจะซ้อนทับกันมาก การจัดกลุ่มข้อมูล seeded K-means และ constrained K-means กับการใช้เกณฑ์วัดระยะแบบห่าง Euclidean ให้ผลการจัดกลุ่มข้อมูลที่มีความถูกต้องน้อยกว่าวิธี seeded K-means และ constrained K-means กับการใช้เกณฑ์วัดระยะห่างแบบ Mahalanobis เนื่องจากการใช้เกณฑ์วัดระยะ Mahalanobis มีการใช้ค่าเมทริกซ์ความแปรปรวนในการวัดระยะห่างระหว่างหน่วยตัวอย่าง

กับจุดศูนย์ของแต่ละกลุ่ม ดังนั้นสรุปได้ว่าค่าเมทริกซ์ความแปรปรวนมีส่วนในการช่วยการจัดกลุ่มข้อมูลที่มีหน่วยตัวอย่างแต่ละกลุ่มที่ซ้อนทับกัน

ในการใช้เกณฑ์วัดระยะ Euclidean กับชุดข้อมูลที่มีหน่วยตัวอย่างแต่ละกลุ่มมีการซ้อนทับกันจำนวนมาก ด้วยใช้จำนวนข้อมูลกำกับกลุ่มที่แตกต่างกัน ส่วนใหญ่วิธี constrained K-means ให้ประสิทธิภาพที่ดีกว่า วิธี seeded K-means เนื่องจากวิธี constrained K-means มีการใช้ชุดข้อมูลกำกับกลุ่มในกระบวนการทำงาน โดยมีการกำหนดกลุ่มเดิมของชุดข้อมูลที่กำกับกลุ่ม และชุดข้อมูลที่กำกับกลุ่มนั้นมีส่วนร่วมในการหาค่าจุดศูนย์กลางใหม่ในแต่ละรอบการทำงาน

ส่วนในการใช้เกณฑ์วัดระยะห่าง Mahalanobis กับชุดข้อมูลที่มีหน่วยตัวอย่างแต่ละกลุ่มมีการซ้อนทับกันจำนวนมาก วิธี constrained K-means ให้ประสิทธิภาพที่ดีกว่า วิธี seeded K-means เนื่องจากการใช้ค่าเมทริกซ์ความแปรปรวนเข้ามาคำนวณระยะห่าง ทำให้วิธีการจัดกลุ่มข้อมูลสามารถใช้ค่านี้ในการรับรู้ถึงรูปร่างและการกระจายของข้อมูลในแต่ละกลุ่มได้ดีกว่า และเมื่อใช้ร่วมกับ constrained K-means ที่มีการใช้ชุดข้อมูลที่กำกับกลุ่มตลอดการทำงาน จึงทำให้มีผลลัพธ์การทำงานที่ดีขึ้น

และในส่วนการเพิ่มชุดข้อมูลที่กำกับกลุ่ม วิธี constrained K-means และวิธี seeded K-means ถึงแม้ใช้เกณฑ์วัดระยะ Euclidean หรือใช้เกณฑ์วัดระยะ Mahalanobis ผลการจัดกลุ่มข้อมูลถูกต้องจะมีค่าเพิ่มขึ้น แต่จำนวนชุดข้อมูลที่กำกับกลุ่ม 30% ก็เพียงพอต่อการจัดกลุ่มข้อมูลทั้งสองวิธี ยกเว้นชุดข้อมูล banknote เนื่องจากชุดข้อมูลนั้นมีจำนวนหน่วยตัวอย่างทั้งหมดเป็นจำนวนมาก และมีเพียงสองกลุ่ม ดังนั้นการกำหนดจำนวนชุดข้อมูลที่กำกับกลุ่ม 5% ในชุดข้อมูลทั้งหมดก็มีจำนวนมากเพียงพอในการจัดกลุ่มข้อมูล และยกเว้นชุดข้อมูล user knowledge modeling การจัดกลุ่มข้อมูลด้วยวิธี SKE เมื่อเพิ่มจำนวนข้อมูลที่กำกับกลุ่ม ผลการจัดกลุ่มข้อมูลถูกต้องมีค่าลดลง เนื่องจากชุดข้อมูล user knowledge modeling เป็นข้อมูลที่มีหน่วยตัวอย่างแต่ละกลุ่มมีการซ้อนทับกันจำนวนมาก ดังนั้นการเพิ่มจำนวนข้อมูลที่กำกับกลุ่มจะทำให้หน่วยตัวอย่างแต่ละกลุ่มมีการซ้อนทับกันมากขึ้น

บทที่ 5

สรุปผลการดำเนินงานและข้อแนะนำ

บทนี้กล่าวถึงสรุปผลการดำเนินงานและข้อเสอนแนะเพื่อเป็นแนวทางในการพัฒนาโครงงานต่อไป ซึ่งมีรายละเอียดดังนี้

5.1 สรุปผลการดำเนินงาน

จากการบันทึกผลการทดลองกับชุดข้อมูล 5 ชุด สังเกตพบว่า วิธีการจัดกลุ่มข้อมูล constrained K-means clustering จะได้ผลการจัดกลุ่มข้อมูลส่วนใหญ่ได้ถูกต้องกว่าวิธี seeded K-means clustering ถึงแม้จะใช้เกณฑ์วัดระยะ Euclidean หรือ Mahalanobis แต่ชุดข้อมูลที่มีการซ้อนทับกันของหน่วยตัวอย่างเป็นจำนวนมาก ขั้นตอนวิธีการจัดกลุ่มข้อมูล seeded K-means และ constrained K-means กับการใช้เกณฑ์วัดระยะ Euclidean ผลการจัดกลุ่มข้อมูลมีความถูกต้องน้อยกว่าวิธี seeded K-means และ constrained K-means กับการใช้เกณฑ์วัดระยะ Mahalanobis และการเพิ่มจำนวนข้อมูลที่กำกับกลุ่มส่วนใหญ่ก็มีผลต่อการจัดกลุ่มข้อมูลทั้งสองวิธี ไม่ว่าจะใช้เกณฑ์วัดระยะ Euclidean หรือ Mahalanobis แต่จำนวนข้อมูลที่กำกับกลุ่ม 30% ก็เพียงพอต่อการจัดกลุ่มข้อมูล

สรุปได้ว่าขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอนโดยใช้เกณฑ์วัดระยะ Mahalanobis มีประสิทธิภาพการจัดกลุ่มข้อมูลดีกว่าการใช้เกณฑ์วัดระยะห่าง Euclidean ในกรณีที่กลุ่มของข้อมูลมีการซ้อนทับกันมากและมีประสิทธิภาพใกล้เคียงกันหากกลุ่มข้อมูลค่อนข้างแยกจากกัน อย่างไงก็ตามชุดข้อมูลที่มีตัวแปรจำนวนมากเช่น wine ระยะเวลาในการคำนวณของคอมพิวเตอร์ (computational complexity) ของขั้นตอนวิธีการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน โดยเกณฑ์วัดระยะ Mahalanobis ค่อนข้างใช้เวลานานกว่า การจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอนที่ใช้เกณฑ์วัดระยะ Euclidean และชุดข้อมูลที่มีกลุ่มข้อมูลค่อนข้างแยกออกจากกันควรใช้เกณฑ์วัดระยะ Euclidean ดีกว่าใช้เกณฑ์วัดระยะ Mahalanobis เนื่องจากระยะเวลาในการคำนวณของคอมพิวเตอร์ใช้เวลาเร็วกว่าและมีประสิทธิภาพการจัดกลุ่มข้อมูลก็ไม่แตกต่างกัน

5.2 ข้อเสนอแนะ

สำหรับผู้ที่สนใจและต้องการศึกษาต่อการจัดกลุ่มข้อมูล K-means แบบกึ่งมีผู้สอน (semi-supervised K-means clustering) นักศึกษาได้มีแนวทางในการศึกษาดังนี้

- ชุดข้อมูลที่นำมาจัดกลุ่มข้อมูลควรนำไป standardization หรือ normalization ก่อน
- ชุดข้อมูลใน real world data ปัจจุบันนี้มีตัวแปรจำนวนมากซึ่งตัวแปรบางตัวแปรไม่ได้มีความสำคัญในการจัดกลุ่ม ดังนั้นควรจะใช้วิธีการคัดเลือกตัวแปรที่สำคัญมาใช้งานเช่น

วิธี feature selection ก่อนนำไปการจัดกลุ่มข้อมูล เนื่องจากชุดข้อมูลที่มีตัวแปรจำนวนมากจะทำให้การจัดกลุ่มใช้เวลาจำนวนมาก และตัวแปรบางตัวเป็นตัวแปรรบกวน (noise variable) จะทำให้ผลการจัดกลุ่มข้อมูลไม่ได้ถูกต้อง หรือใช้วิธีลดจำนวนตัวแปรเช่น วิธี principal components และ วิธี factor analysis ซึ่งเป็นวิธีหาตัวแปรที่มาความสัมพันธ์กันให้อยู่ในตัวแปรเดียวกันด้วยสร้างเป็นตัวแปรใหม่

บรรณานุกรม

[1] กัลยา วานิชย์บัญชา. 2552. การวิเคราะห์ข้อมูลหลายตัวแปร. กรุงเทพฯ : ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

[2] สุจรรยา บุญประดิษฐ์. 2559. เอกสารประกอบการเรียนวิชา multivariate analysis. ปัตตานี: ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี

[3] สุรพงค์ เอื้อวัฒนามงคล. 2557. การทำเหมืองข้อมูล. กรุงเทพมหานคร: ห้างหุ่นส่วนจำกัด บางกอกบล๊อก.

[4] Basu S, Banerjee A, Mooney R. 2002. Semi-supervised clustering by seeding. In proceedings of the 19th International Conference on Machine Learning (ICML-2002).

[5] Cerioli, A. 2005. K-means Cluster Analysis and Mahalanobis Metrics: a problematic match or an overlooked opportunity. *Statistica Applicata*, *17*(1).

[6] Chokniwal, A., & Singh, M. (2016, September). Faster Mahalanobis K-means clustering for Gaussian distributions. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on* (pp. 947-952). IEEE.

[7] D. Arthur and S. Vassilvitskii. 2007. K-means++: The advantages of careful seeding. proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035.

[8] MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, edited by Le Cam LM, Neyman J. University of California Press, Berkeley, CA, volume 1 281–297.

[9] Richard A. Johnson, Dean W. Wichern. 2007. Applied Multivariate Statistical Analysis. 6thed. Pearson

[10] Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to Semi-Supervised Learning. Morgan and Claypool.

[11] Lichman, M. 2013. UCI Machine Learning Repository [https://archive.ics.uci.edu/ml]. Irvine. CA: University of California.

ภาคผนวก ก

ภาคผนวก ก แสดงเมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของชุดข้อมูล 5 ชุด ที่ดังกล่าวในหัวข้อ 3.1

ก.1 เมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของชุดข้อมูล

ตารางที่ ก.1 เมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของข้อมูล 3 กลุ่มในชุดข้อมูล iris

Class	Mean	Covariance
1	$\begin{bmatrix}5.006\\3.428\\1.462\\0.246\end{bmatrix}$	$\begin{bmatrix}0.1242 & 0.0992 & 0.0164 & 0.0103\\0.0992 & 0.1437 & 0.0117 & 0.0093\\0.0164 & 0.0117 & 0.0302 & 0.0061\\0.0103 & 0.0093 & 0.0061 & 0.0111\end{bmatrix}$
2	$\begin{bmatrix}5.936\\2.770\\4.260\\1.326\end{bmatrix}$	$\begin{bmatrix}0.2664 & 0.0852 & 0.1829 & 0.0558\\0.0852 & 0.0985 & 0.0827 & 0.0412\\0.1829 & 0.0827 & 0.2208 & 0.0731\\0.0558 & 0.0412 & 0.0731 & 0.0391\end{bmatrix}$
3	$\begin{bmatrix}6.588\\2.974\\5.552\\2.026\end{bmatrix}$	$\begin{bmatrix}0.4043 & 0.0938 & 0.3033 & 0.0491\\0.0938 & 0.1040 & 0.0714 & 0.0476\\0.3033 & 0.0714 & 0.3046 & 0.0488\\0.0491 & 0.0476 & 0.0488 & 0.0754\end{bmatrix}$

ตารางที่ ก.2 เมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของข้อมูล 3 กลุ่มในชุดข้อมูล seeds

Class	Mean	Covariance
1	$\begin{bmatrix}14.3344\\14.2943\\0.8801\\5.5081\\3.2446\\2.6674\\5.0872\end{bmatrix}$	$\begin{bmatrix}1.4779 & 0.6844 & 0.0073 & 0.2349 & 0.1943 & -0.0720 & 0.2311\\0.6844 & 0.3324 & 0.0015 & 0.1230 & 0.0822 & -0.0365 & 0.1207\\0.0073 & 0.0015 & 0.0003 & -0.0005 & 0.0019 & 0.0007 & -0.0006\\0.2349 & 0.1230 & -0.0005 & 0.0536 & 0.0226 & -0.0100 & 0.0529\\0.1943 & 0.0822 & 0.0019 & 0.0226 & 0.0315 & -0.0056 & 0.0209\\-0.0720 & -0.0365 & 0.0007 & -0.0100 & -0.0056 & 1.3780 & -0.0034\\0.2311 & 0.1207 & -0.0006 & 0.0529 & 0.0209 & -0.0034 & 0.0695\end{bmatrix}$
2	$\begin{bmatrix}18.3343\\16.1357\\0.8835\\6.1480\\3.6774\\3.6448\\6.0206\end{bmatrix}$	$\begin{bmatrix}2.0721 & 0.8667 & 0.0061 & 0.3191 & 0.2352 & -0.0672 & 0.2662\\0.8667 & 0.3807 & 0.0005 & 0.1500 & 0.0879 & -0.0234 & 0.1272\\0.0061 & 0.0005 & 0.0002 & -0.0009 & 0.0019 & -0.0013 & -0.0010\\0.3191 & 0.1500 & -0.0009 & 0.0719 & 0.0255 & -0.0167 & 0.0637\\0.2352 & 0.0879 & 0.0019 & 0.0255 & 0.0344 & 0.0029 & 0.0189\\-0.0672 & -0.0234 & -0.0013 & -0.0167 & 0.0029 & 1.3968 & -0.0245\\0.2662 & 0.1272 & -0.0010 & 0.0637 & 0.0189 & -0.0245 & 0.0645\end{bmatrix}$
3	$\begin{bmatrix}11.8793\\13.2479\\0.8494\\5.2295\\2.8538\\4.7884\\5.1164\end{bmatrix}$	$\begin{bmatrix}0.5227 & 0.0032 & 0.0086 & 0.0515 & 0.0921 & 0.0383 & 0.0303\\0.2232 & 0.1157 & 0.0011 & 0.0373 & 0.0295 & 0.0260 & 0.0301\\0.0086 & 0.0011 & 0.0005 & -0.0011 & 0.0028 & -0.0011 & -0.0017\\0.0515 & 0.0373 & -0.0011 & 0.0190 & 0.0017 & 0.0239 & 0.0185\\0.0921 & 0.0295 & 0.0028 & 0.0017 & 0.0218 & 0.0146 & -0.0032\\0.0383 & 0.0260 & -0.0011 & 0.0239 & 0.0146 & 1.7861 & 0.0233\\0.0303 & 0.0301 & -0.0017 & 0.0185 & -0.0032 & 0.0233 & 0.0263\end{bmatrix}$

ตารางที่ ก.3 เมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของข้อมูล 3 กลุ่มในชุดข้อมูล wine

Class	Mean	Covariance
1	$\begin{bmatrix} 13.7447 \\ 2.0107 \\ 2.4556 \\ 17.0373 \\ 106.3390 \\ 2.8402 \\ 2.9824 \\ 0.2900 \\ 1.8993 \\ 5.5283 \\ 1.0620 \\ 3.1578 \end{bmatrix}$	$\begin{bmatrix} 0.2136 & -0.0129 & -0.0156 & -0.3746 & 0.7732 & 0.0659 & 0.0762 & 0.0005 & 0.0586 & 0.2337 & 0.0043 & 0.0115 \\ -0.0129 & 0.4741 & 0.0041 & 0.1053 & 0.5734 & -0.0195 & -0.0524 & -0.0043 & -0.0229 & -0.2197 & -0.0337 & 0.0426 \\ -0.0156 & 0.0041 & 0.0516 & 0.3178 & 0.9124 & 0.0004 & -0.0064 & 0.0074 & -0.0136 & -0.0350 & 0.0063 & -0.0066 \\ -0.3746 & 0.1053 & 0.3178 & 6.4838 & 6.3716 & -0.1925 & -0.2906 & 0.0539 & -0.1822 & -0.6653 & 0.0276 & -0.1070 \\ 0.7732 & -0.5734 & 0.9124 & 6.3716 & 110.2279 & 0.10934 & 0.5147 & 0.1745 & -0.2555 & 2.4013 & -0.1362 & 0.4523 \\ 0.0659 & -0.0195 & 0.0004 & -0.1925 & 1.0934 & 0.1149 & 0.1083 & -0.0004 & 0.0522 & 0.2729 & -0.0089 & 0.0064 \\ 0.0762 & -0.0524 & -0.0064 & -0.2906 & 0.5147 & 0.1083 & 0.1580 & -0.0025 & 0.0899 & 0.3651 & 0.0004 & -0.0126 \\ 0.0005 & -0.0043 & 0.0074 & 0.0539 & 0.1745 & -0.0004 & -0.0025 & 0.0049 & -0.0042 & -0.0132 & 0.0034 & -0.0081 \\ 0.0586 & -0.0229 & -0.0136 & -0.1822 & -0.2555 & 0.0522 & 0.0899 & -0.0042 & 0.1698 & 0.2168 & 0.0050 & 0.0005 \\ 0.2337 & -0.2197 & -0.0350 & -0.6653 & 2.4013 & 0.2729 & 0.3651 & -0.0132 & 0.2168 & 1.5341 & 0.0041 & -0.0827 \\ 0.0043 & -0.0337 & 0.0063 & 0.0276 & -0.1362 & -0.0089 & 0.0004 & 0.0034 & 0.0050 & 0.0041 & 0.0136 & -0.0129 \\ 0.0115 & 0.0426 & -0.0066 & -0.1070 & 0.4523 & 0.0064 & -0.0126 & -0.0081 & 0.0005 & -0.0827 & -0.0129 & 0.1275 \end{bmatrix}$
2	$\begin{bmatrix} 12.2787 \\ 1.9327 \\ 2.2448 \\ 20.2380 \\ 94.5493 \\ 2.2589 \\ 2.0808 \\ 0.3637 \\ 1.6303 \\ 3.0866 \\ 1.0563 \\ 2.7854 \end{bmatrix}$	$\begin{bmatrix} 0.2894 & -0.0117 & -0.0365 & -0.1014 & -0.2695 & -0.0136 & -0.0145 & -0.0045 & -0.0614 & 0.1342 & -0.0002 & -0.0348 \\ -0.0117 & 1.0314 & 0.0476 & 0.8094 & -1.3065 & 0.0218 & 0.0802 & 0.0161 & 0.1287 & -0.1909 & -0.0841 & 0.0796 \\ -0.0365 & 0.0476 & 0.0995 & 0.7347 & 0.6825 & 0.0193 & 0.0701 & 0.0117 & 0.0087 & 0.0176 & -0.0020 & 0.0252 \\ -0.1014 & 0.8094 & 0.7347 & 11.2210 & 0.1831 & 0.2337 & 0.7360 & 0.0758 & 0.2195 & -0.2660 & -0.0522 & 0.6338 \\ -0.2696 & -1.3065 & 0.6825 & 0.1831 & 280.6797 & 0.6403 & 0.0200 & -0.4032 & 3.0037 & 0.6807 & 0.4244 & -0.6338 \\ -0.0136 & 0.0218 & 0.0193 & 0.2237 & 0.6403 & 0.2974 & 0.2967 & -0.0287 & 0.1256 & 0.0853 & 0.0044 & 0.1313 \\ -0.0145 & 0.0802 & 0.0701 & 0.7360 & 0.0200 & 0.2964 & 0.4980 & -0.0206 & 0.2121 & 0.2471 & -0.0042 & 0.2031 \\ -0.0045 & 0.0161 & 0.0117 & 0.0758 & -0.4032 & -0.0287 & -0.0206 & 0.0154 & -0.0240 & 0.0021 & -0.0008 & -0.0254 \\ -0.0614 & 0.1287 & 0.0082 & 0.2195 & 3.0037 & 0.1256 & 0.2121 & -0.0240 & 0.3625 & -0.0411 & -0.0066 & 0.1153 \\ 0.1342 & -0.1909 & 0.0176 & -0.2660 & 0.6807 & 0.0853 & 0.2471 & 0.0021 & -0.0411 & 0.8555 & -0.0049 & -0.0538 \\ -0.0002 & -0.0841 & -0.0020 & -0.0522 & 0.4244 & 0.0044 & -0.0042 & -0.0008 & -0.0066 & -0.0049 & 0.0412 & -0.0053 \\ -0.0348 & 0.0796 & 0.0252 & 0.6356 & -0.6338 & 0.1313 & 0.2031 & -0.0254 & 0.1153 & -0.0538 & -0.0053 & 0.2466 \end{bmatrix}$
3	$\begin{bmatrix} 13.1538 \\ 3.3338 \\ 2.4371 \\ 21.4167 \\ 99.3125 \\ 1.6788 \\ 0.7815 \\ 0.4475 \\ 1.1535 \\ 7.3962 \\ 0.6827 \\ 1.6835 \end{bmatrix}$	$\begin{bmatrix} 0.2812 & 0.0637 & 0.0240 & 0.2514 & -0.4859 & 0.0398 & 0.0118 & 0.0025 & 0.0816 & 0.4293 & -0.0021 & 0.0191 \\ 0.0637 & 1.1835 & 0.0036 & 0.2089 & -2.0731 & -0.0624 & -0.0899 & 0.0193 & -0.0994 & -0.4078 & 0.0099 & 0.0021 \\ 0.0240 & 0.0036 & 0.0341 & 0.3163 & 0.4250 & 0.0310 & 0.0150 & -0.0005 & 0.0146 & 0.0534 & 0.0038 & 0.0113 \\ 0.2514 & 0.2089 & 0.3163 & 5.0993 & 3.9202 & 0.2938 & 0.1799 & -0.0048 & 0.2432 & 0.8382 & 0.0071 & 0.0256 \\ -0.4859 & -2.0731 & 0.4250 & 3.9202 & 118.6024 & -0.1541 & 1.8180 & -0.6837 & 0.6836 & 2.6223 & 0.0057 & -0.6609 \\ 0.0398 & -0.0624 & 0.0310 & 0.2938 & -0.1541 & 0.1274 & 0.0250 & 0.0145 & 0.0905 & 0.2771 & -0.0011 & 0.0195 \\ 0.0118 & -0.0899 & 0.0150 & 0.1799 & 1.8180 & 0.0250 & 0.0861 & -0.0231 & 0.0490 & 0.2489 & -0.0098 & -0.0343 \\ 0.0025 & 0.0193 & -0.0005 & -0.0048 & -0.6837 & 0.0145 & -0.0231 & 0.0154 & 0.0087 & 0.0075 & 0.0022 & 0.0104 \\ 0.0816 & -0.0994 & 0.0146 & 0.2432 & 0.6836 & 0.0905 & 0.0490 & 0.0087 & 0.1671 & 0.6471 & -0.0197 & -0.0143 \\ 0.4293 & -0.4078 & 0.0534 & 0.8382 & 2.6223 & 0.2271 & 0.2489 & 0.0075 & 0.6471 & 5.3405 & -0.1504 & -0.0648 \\ -0.0021 & 0.0099 & 0.0038 & 0.0071 & 0.0057 & -0.0011 & -0.0098 & 0.0022 & -0.0197 & -0.1504 & 0.0131 & 0.0113 \\ 0.0191 & 0.0021 & 0.0113 & 0.0256 & -0.6609 & 0.0195 & -0.0343 & 0.0104 & -0.0143 & -0.0648 & 0.0113 & 0.0740 \end{bmatrix}$

ตารางที่ ก.4 เมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของข้อมูล 2 กลุ่มในชุดข้อมูล banknote authentication

Class	Mean	Covariance
1	$\begin{bmatrix} 2.2767 \\ 4.2566 \\ 0.7967 \\ -1.1476 \end{bmatrix}$	$\begin{bmatrix} 4.0778 & -2.3521 & -2.1590 & 1.7856 \\ -2.3521 & 26.4072 & -12.4992 & -7.3648 \\ -2.1590 & -12.4992 & 10.4969 & 2.8541 \\ 1.7856 & -7.3648 & 2.8541 & 4.5160 \end{bmatrix}$
2	$\begin{bmatrix} -1.8684 \\ -0.9936 \\ 2.1483 \\ -1.2466 \end{bmatrix}$	$\begin{bmatrix} 3.5388 & 0.7492 & -4.6905 & 1.2624 \\ 0.7492 & 29.2128 & -25.2447 & -5.6968 \\ -4.6905 & -25.2447 & 27.6867 & 3.0078 \\ 1.2624 & -5.6968 & 3.0078 & 4.2890 \end{bmatrix}$

ตารางที่ ก.5 แสดงเมทริกซ์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนของข้อมูล 2 กลุ่มในชุดข้อมูล user knowledge modeling

Class	Mean	Covariance
1	$\begin{bmatrix} 0.4069 \\ 0.4305 \\ 0.5098 \\ 0.5429 \\ 0.7998 \end{bmatrix}$	$\begin{bmatrix} 0.0608 & 0.0043 & -0.0149 & 0.0200 & -0.0040 \\ 0.0043 & 0.0599 & 0.0054 & 0.0064 & -0.0022 \\ -0.0149 & 0.0054 & 0.0633 & 0.0026 & -0.0010 \\ 0.0200 & 0.0064 & 0.0026 & 0.0759 & -0.0183 \\ -0.0040 & -0.0022 & -0.0010 & -0.0183 & 0.0120 \end{bmatrix}$
2	$\begin{bmatrix} 0.3268 \\ 0.3228 \\ 0.4250 \\ 0.4493 \\ 0.2536 \end{bmatrix}$	$\begin{bmatrix} 0.0327 & 0.0017 & -0.0041 & -0.0056 & 0.0011 \\ 0.0017 & 0.0368 & 0.0008 & 0.0029 & -0.0029 \\ -0.0041 & 0.0008 & 0.0632 & 0.0016 & -0.0023 \\ -0.0056 & 0.0029 & 0.0016 & 0.0523 & -0.0120 \\ 0.0011 & -0.0029 & -0.0023 & -0.0120 & 0.0051 \end{bmatrix}$
3	$\begin{bmatrix} 0.3746 \\ 0.3672 \\ 0.4911 \\ 0.3857 \\ 0.5314 \end{bmatrix}$	$\begin{bmatrix} 0.0441 & -0.0020 & 0.0026 & -0.0030 & 0.0016 \\ -0.0020 & 0.0429 & 0.0063 & 0.0014 & -0.0017 \\ 0.0026 & 0.0063 & 0.0542 & 0.0031 & -0.0035 \\ -0.0030 & 0.0014 & 0.0031 & 0.0639 & -0.0294 \\ 0.0016 & -0.0017 & -0.0035 & -0.0294 & 0.0171 \end{bmatrix}$
4	$\begin{bmatrix} 0.2592 \\ 0.2619 \\ 0.3540 \\ 0.2688 \\ 0.0958 \end{bmatrix}$	$\begin{bmatrix} 0.0311 & -0.0102 & -0.0069 & 0.0075 & 0.0025 \\ -0.0102 & 0.0318 & 0.0024 & 0.0002 & -0.0001 \\ -0.0069 & 0.0024 & 0.0474 & 0.0024 & -0.0023 \\ 0.0075 & 0.0002 & 0.0024 & 0.0343 & -0.0018 \\ 0.0025 & -0.0001 & -0.0023 & -0.0018 & 0.0031 \end{bmatrix}$

ภาคผนวก ข

ภาคผนวก ข เป็นผลบันทึกการทดลองที่ 1 การใช้เกณฑ์วัดระยะห่างแบบ Euclidean และทดลองที่ 2 การใช้เกณฑ์วัดระยะห่างแบบ Mahalanobis ทดลองกับชุดข้อมูล 5 ชุด

ข.1 ผลการทดลองที่ 1 การใช้เกณฑ์วัดระยะห่างแบบ Euclidean

ข.1.1 ผลการทดลองกับชุดข้อมูล iris

ตารางที่ ข.1 ค่า overall accuracy กับ class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	79.07	79.87	78.40	82.13	83.87	82.67
Class 1	100	100	100	100	100	100
Class 2	66.88	68 .93	67.32	70.51	73.10	73.08
Class 3	75.45	75.60	72.91	80.82	82.76	78.57
Accuracy	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	74.27	73.87	73.33	82.67	84.93	84
Class 1	100	100	100	100	100	100
Class 2	59.61	58.88	58.95	69.78	72.66	70.97
Class 3	69.81	70.62	68.19	83.81	87.33	86.96

ตารางที่ ข.2 จำนวนรอบของวิธี SKE กับวิธี CKE ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	3	4	4	4	8
สูงสุด	17	20	15	11	10	8
ด้วยเฉลี่ย	8.5	8.4	8.2	6.6	6.2	8
จำนวนรอบ	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	4	4	5
สูงสุด	17	11	9	7	5	5
ด้วยเฉลี่ย	8.1	7.7	6.3	5.1	4.6	5

ตารางที่ ข.3 ค่า WSCD ของวิธี SKE กับวิธี CKE ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKE	210.0470	231.1988	269.9717	314.5222	359.6331	383.0738
CKE	213.3891	238.6268	286.7604	333.4574	381.2400	408.0825

ตารางที่ ข.4 ค่า $\left|\mu_{ik} - \bar{x}_{ik}\right|, i = 1,2,...,p; k = 1,2,...,K$ ของวิธี SKE กับวิธี CKE ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1717\\0.0699\\0.0687\\0.0538\end{bmatrix}$	$\begin{bmatrix}0.1715\\0.0706\\0.0648\\0.0491\end{bmatrix}$	$\begin{bmatrix}0.1717\\0.0699\\0.0687\\0.0538\end{bmatrix}$	$\begin{bmatrix}0.1720\\0.0683\\0.0765\\0.0634\end{bmatrix}$	$\begin{bmatrix}0.1720\\0.0683\\0.0765\\0.0634\end{bmatrix}$	$\begin{bmatrix}0.1720\\0.0683\\0.0765\\0.0634\end{bmatrix}$
Class 2	$\begin{bmatrix}0.4339\\0.2832\\0.1599\\0.2872\end{bmatrix}$	$\begin{bmatrix}0.4169\\0.2942\\0.2228\\0.2970\end{bmatrix}$	$\begin{bmatrix}0.5383\\0.2506\\0.1684\\0.2590\end{bmatrix}$	$\begin{bmatrix}0.3071\\0.1905\\0.0470\\0.1895\end{bmatrix}$	$\begin{bmatrix}0.3055\\0.2223\\0.0319\\0.1372\end{bmatrix}$	$\begin{bmatrix}0.4081\\0.2180\\0.0489\\0.1390\end{bmatrix}$
Class 3	$\begin{bmatrix}0.4502\\0.2479\\0.3461\\0.5161\end{bmatrix}$	$\begin{bmatrix}0.3973\\0.2208\\0.3341\\0.4209\end{bmatrix}$	$\begin{bmatrix}0.4420\\0.2159\\0.3451\\0.3329\end{bmatrix}$	$\begin{bmatrix}0.3398\\0.0906\\0.2850\\0.3098\end{bmatrix}$	$\begin{bmatrix}0.2771\\0.0946\\0.2721\\0.2824\end{bmatrix}$	$\begin{bmatrix}0.3236\\0.0091\\0.3024\\0.2029\end{bmatrix}$
	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1717\\0.0699\\0.0687\\0.0538\end{bmatrix}$	$\begin{bmatrix}0.1549\\0.1009\\0.0663\\0.0494\end{bmatrix}$	$\begin{bmatrix}0.1719\\0.0691\\0.0726\\0.0586\end{bmatrix}$	$\begin{bmatrix}0.1720\\0.0683\\0.0765\\0.0634\end{bmatrix}$	$\begin{bmatrix}0.1720\\0.0683\\0.0765\\0.0634\end{bmatrix}$	$\begin{bmatrix}0.1070\\0.0683\\0.0765\\0.0634\end{bmatrix}$
Class 2	$\begin{bmatrix}0.2906\\0.2517\\0.1418\\0.5330\end{bmatrix}$	$\begin{bmatrix}0.2592\\0.1892\\0.1541\\0.5336\end{bmatrix}$	$\begin{bmatrix}0.2129\\0.2026\\0.1015\\0.5711\end{bmatrix}$	$\begin{bmatrix}0.2037\\0.1991\\0.0426\\0.2018\end{bmatrix}$	$\begin{bmatrix}0.1645\\0.2166\\0.0343\\0.1109\end{bmatrix}$	$\begin{bmatrix}0.1391\\0.1935\\0.0482\\0.0821\end{bmatrix}$
Class 3	$\begin{bmatrix}0.3051\\0.2320\\0.2552\\0.6403\end{bmatrix}$	$\begin{bmatrix}0.3773\\0.1665\\0.2373\\0.6333\end{bmatrix}$	$\begin{bmatrix}0.3115\\0.1782\\0.1770\\0.6075\end{bmatrix}$	$\begin{bmatrix}0.2430\\0.1591\\0.2406\\0.3916\end{bmatrix}$	$\begin{bmatrix}0.2006\\0.1852\\0.2501\\0.4168\end{bmatrix}$	$\begin{bmatrix}0.1878\\0.1902\\0.2773\\0.4888\end{bmatrix}$

ข.1.2 ผลการทดลองกับชุดข้อมูล seeds

ตารางที่ ข.5 ค่า overall accuracy กับ class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	87.91	91.53	93.72	94.48	94.67	95.24
Class 1	86.55	87.52	91.47	91.26	90.90	92.11
Class 2	99.69	99.39	98.81	99.70	99.70	100
Class 3	82.73	90.41	92.12	93.61	94.21	94.29
Accuracy	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	88.24	91.62	94.10	94.76	94.86	95.24
Class 1	86.72	87.64	91.95	91.74	91.59	92.11
Class 2	99.29	99.39	99.09	99.70	99.70	100
Class 3	82.93	90.46	92.41	93.90	93.99	94.29

ตารางที่ ข.6 จำนวนรอบของวิธี SKE กับวิธี CKE ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	4	4	4
สูงสุด	13	9	6	7	7	4
ด้วยเฉลี่ย	6.5	5.8	5.3	5.3	5.3	4
จำนวนรอบ	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	4	4	4
สูงสุด	10	9	6	6	6	4
ด้วยเฉลี่ย	6.2	5.8	4.8	4.6	4.2	4

ตารางที่ ข.7 ค่า WSCD ของวิธี SKE กับวิธี CKE ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKE	654.7275	700.6807	800.7895	912.164	1015.581	1120.061
CKE	655.0988	701.2603	802.0476	913.5089	1017.164	1120.75

ตารางที่ ข.8 ค่า $\left|\mu_{ik} - \bar{x}_{ik}\right|, i = 1,2,..., p; k = 1,2,..., K$ ของวิธี SKE กับวิธี CKE ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	0.0399 0.1882 0.1490 0.1802 0.0725 0.2159 0.1985	0.0227 0.1503 0.0784 0.1327 0.0918 0.2536 0.1104	0.0185 0.0920 0.0445 0.1187 0.0727 0.1945 0.0953	0.0241 0.0654 0.0222 0.1478 0.0520 0.1654 0.0897	0.0223 0.0248 0.0153 0.1441 0.0484 0.1761 0.0854	0.0221 0.0186 0.0111 0.1410 0.0408 0.1717 0.0910
Class 2	0.2185 0.1827 0.1386 0.1371 0.1038 0.0357 0.0623	0.1551 0.1595 0.0785 0.1114 0.0855 0.0258 0.0364	0.0556 0.0888 0.0316 0.1048 0.0552 0.0282 0.0283	0.0420 0.0865 0.0191 0.0904 0.0352 0.0215 0.0249	0.0420 0.0865 0.0191 0.0904 0.0352 0.0215 0.0249	0.0462 0.0899 0.0199 0.0901 0.0324 0.0236 0.0272
Class 3	0.1260 0.2185 0.2462 0.1060 0.1078 0.3212 0.1361	0.0455 0.0748 0.0603 0.0656 0.0749 0.3440 0.1342	0.0558 0.0558 0.0590 0.0548 0.0306 0.3145 0.1432	0.0583 0.0480 0.0269 0.0522 0.0156 0.2635 0.1552	0.0487 0.0253 0.0106 0.0363 0.0037 0.2693 0.1715	0.0391 0.0098 0.0153 0.0370 0.0037 0.2759 0.1580
	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	0.0412 0.1859 0.1509 0.1869 0.0691 0.2158 0.1996	0.0228 0.1471 0.0761 0.1414 0.0858 0.2398 0.1056	0.0196 0.0924 0.0439 0.1141 0.0661 0.1860 0.0959	0.0215 0.0568 0.0180 0.1397 0.0473 0.1689 0.0948	0.0219 0.0246 0.0163 0.1407 0.0414 0.1743 0.0899	0.0221 0.0186 0.0111 0.1410 0.0408 0.1717 0.0910
Class 2	0.2057 0.1833 0.1243 0.1450 0.1184 0.0289 0.0628	0.1551 0.1595 0.0785 0.1114 0.0855 0.0258 0.0364	0.0526 0.0862 0.0290 0.1039 0.0492 0.0194 0.0244	0.0420 0.0865 0.0191 0.0904 0.0352 0.0215 0.0249	0.0420 0.0865 0.0191 0.0904 0.0352 0.0215 0.0249	0.0462 0.0899 0.0199 0.0901 0.0324 0.0236 0.0272
Class 3	0.1238 0.2167 0.2390 0.1057 0.1124 0.3176 0.1375	0.0457 0.0701 0.0589 0.0605 0.0770 0.3310 0.1329	0.0498 0.0544 0.0570 0.0555 0.0271 0.2993 0.1350	0.0512 0.0408 0.0254 0.0488 0.0128 0.2684 0.1495	0.0412 0.0173 0.0123 0.0372 0.0067 0.2763 0.1589	0.0391 0.0098 0.0153 0.0370 0.0037 0.2759 0.1580

ข.1.3 ผลการทดลองกับชุดข้อมูล wine

ตารางที่ ข.9 ค่า overall accuracy กับ class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	73.60	84.04	91.12	93.60	96.74	97.75
Class 1	73.47	78.43	87.69	88.57	92.59	93.75
Class 2	77.65	90.92	95.58	98.81	100	100
Class 3	71.22	85.43	90.14	93.09	97.26	100
Accuracy	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	74.16	84.94	91.68	93.48	96.64	97.75
Class 1	73.96	78.30	87.80	88.57	92.57	93.75
Class 2	78.02	91.56	95.73	99.10	100	100
Class 3	72.27	87.64	91.60	92.36	96.89	100

ตารางที่ ข.10 จำนวนรอบของวิธี SKE กับวิธี CKE ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	5	4	4	4	3	5
สูงสุด	11	10	8	8	6	5
ด้วยเฉลี่ย	6.9	6.3	6.1	5	4.6	5
จำนวนรอบ	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	5	4	4	4	3	4
สูงสุด	11	10	6	6	5	4
ด้วยเฉลี่ย	6.3	6.3	5.1	4.8	4	4

ตารางที่ ข.11 ค่า WSCD ของวิธี SKE กับวิธี CKE ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKE	1035.866	1132.11	1322.232	1495.292	1680.58	1833.471
CKE	1038.519	1135.042	1324.583	1496.276	1680.974	1833.594

ตารางที่ ข.12 ค่า $|\mu_{ik} - \bar{x}_{ik}|, i = 1,2,...,p; k = 1,2,...,K$ ของวิธี SKE กับวิธี CKE ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	0.2185 0.1706 0.2488 0.3091 0.2460 0.1396 0.2085 0.1949 0.2521 0.0591 0.1954 0.3459	0.1488 0.2250 0.2118 0.2159 0.1666 0.1176 0.1933 0.1318 0.2480 0.0941 0.2016 0.3281	0.1452 0.1842 0.1256 0.2322 0.1050 0.0717 0.1836 0.0966 0.2263 0.0669 0.1819 0.2879	0.1416 0.1329 0.1270 0.2819 0.1338 0.0482 0.2240 0.1062 01676 0.0761 0.1442 0.2504	0.0637 0.0805 0.0405 0.2548 0.1257 0.0189 0.2390 0.1126 0.1885 0.0767 0.1133 0.2265	0.0054 0.1472 0.0126 0.2875 0.1096 0.0200 0.2605 0.1309 0.1865 0.0927 0.1383 0.2220
Class 2	0.1899 0.3742 0.0447 0.1563 0.2245 0.2284 0.1471 0.2497 0.1950 0.2098 0.2090 0.1604	0.2533 0.2481 0.1947 0.2491 0.1477 0.1019 0.2568 0.2637 0.1603 0.0531 0.1068 0.1414	0.1939 0.1572 0.2210 0.0971 0.1200 0.0783 0.1208 0.1470 0.1432 0.0422 0.0441 0.0850	0.1459 0.1273 0.1720 0.0733 0.1153 0.0774 0.0700 0.1301 0.1729 0.0515 0.0542 0.0269	0.1066 0.1195 0.1183 0.0691 0.0947 0.1022 0.0511 0.1186 0.1399 0.0297 0.0416 0.0260	0.1005 0.2000 0.1421 0.0966 0.0753 0.1256 0.0175 0.0766 0.1389 0.0256 0.0070 0.0444
Class 3	0.2846 0.5746 0.1492 0.1596 0.2486 0.1523 0.2365 0.4499 0.1898 0.7729 0.2423 0.1861	0.0876 0.3732 0.0949 0.1538 0.1595 0.0675 0.1093 0.1808 0.1650 0.6137 0.0809 0.0963	0.0794 0.3327 0.1016 0.0419 0.1123 0.0620 0.1181 0.1127 0.1454 0.5645 0.0472 0.0577	0.0616 0.3258 0.0769 0.0281 0.1186 0.0483 0.0501 0.0936 0.1592 0.5197 0.0262 0.0742	0.0609 0.3020 0.0800 0.0315 0.0940 0.0685 0.0455 0.0793 0.1688 0.4655 0.0412 0.0735	0.0572 0.2967 0.0892 0.0293 0.0747 0.0898 0.0497 0.0499 0.1855 0.4383 0.0475 0.0691

	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.2085\\0.1763\\0.2618\\0.3166\\0.2434\\0.1287\\0.2111\\0.2073\\0.2536\\0.0602\\0.1902\\0.3288\end{bmatrix}$	$\begin{bmatrix}0.1509\\0.2185\\0.2174\\0.2175\\0.1741\\0.1213\\0.1946\\0.1290\\0.2450\\0.0928\\0.1987\\0.3237\end{bmatrix}$	$\begin{bmatrix}0.1334\\0.1501\\0.1037\\0.2221\\0.1370\\0.0604\\0.1949\\0.0966\\0.2175\\0.0704\\0.1754\\0.2805\end{bmatrix}$	$\begin{bmatrix}0.1416\\0.1329\\0.1270\\0.2819\\0.1338\\0.0482\\0.2240\\0.1062\\0.1676\\0.0761\\0.1442\\0.2504\end{bmatrix}$	$\begin{bmatrix}0.0551\\0.0789\\0.0353\\0.2562\\0.1265\\0.0254\\0.2388\\0.1260\\0.1930\\0.0711\\0.1200\\0.2284\end{bmatrix}$	$\begin{bmatrix}0.0054\\0.1472\\0.0126\\0.2875\\0.1096\\0.0200\\0.2605\\0.1309\\0.1865\\0.0927\\0.1383\\0.2220\end{bmatrix}$
Class 2	$\begin{bmatrix}0.1790\\0.3536\\0.0656\\0.1665\\0.2231\\0.1999\\0.1418\\0.2536\\0.1943\\0.2088\\0.1653\\0.1745\end{bmatrix}$	$\begin{bmatrix}0.2454\\0.2576\\0.2014\\0.2303\\0.1531\\0.1079\\0.2496\\0.2717\\0.1232\\0.0414\\0.1019\\0.1205\end{bmatrix}$	$\begin{bmatrix}0.1881\\0.1626\\0.2084\\0.0828\\0.1174\\0.0736\\0.0955\\0.1397\\0.1438\\0.0302\\0.0405\\0.0709\end{bmatrix}$	$\begin{bmatrix}0.1527\\0.1233\\0.1697\\0.0711\\0.1204\\0.0708\\0.0722\\0.1279\\0.1761\\0.0450\\0.0547\\0.0272\end{bmatrix}$	$\begin{bmatrix}0.0898\\0.1157\\0.1234\\0.0687\\0.0978\\0.1061\\0.0536\\0.1088\\0.1391\\0.0300\\0.0370\\0.0261\end{bmatrix}$	$\begin{bmatrix}0.1005\\0.2000\\0.1421\\0.0966\\0.0753\\0.1256\\0.0175\\0.0766\\0.1389\\0.0256\\0.0070\\0.0444\end{bmatrix}$
Class 3	$\begin{bmatrix}0.3035\\0.5545\\0.1549\\0.1584\\0.2607\\0.1608\\0.2359\\0.4613\\0.1987\\0.7626\\0.2185\\0.1845\end{bmatrix}$	$\begin{bmatrix}0.0857\\0.3562\\0.0856\\0.1303\\0.1612\\0.0636\\0.0961\\0.1878\\0.1418\\0.5859\\0.0856\\0.0832\end{bmatrix}$	$\begin{bmatrix}0.0577\\0.3033\\0.0958\\0.0498\\0.1153\\0.0657\\0.0913\\0.1208\\0.1412\\0.5371\\0.0396\\0.0697\end{bmatrix}$	$\begin{bmatrix}0.0602\\0.3188\\0.0672\\0.0262\\0.1250\\0.0405\\0.0576\\0.1077\\0.1537\\0.5234\\0.0169\\0.0755\end{bmatrix}$	$\begin{bmatrix}0.0608\\0.3035\\0.0775\\0.0321\\0.0935\\0.0667\\0.0421\\0.0772\\0.1709\\0.4697\\0.0370\\0.0739\end{bmatrix}$	$\begin{bmatrix}0.0572\\0.2967\\0.0892\\0.0293\\0.0747\\0.0898\\0.0497\\0.0449\\0.1855\\0.4383\\0.0475\\0.0691\end{bmatrix}$

ข.1.4 ผลการทดลองกับชุดข้อมูล banknote authentication

ตารางที่ ข.13 ค่า overall accuracy กับ class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	96.15	96.04	96.02	95.77	95.83	95.77
Class 1	99.86	99.86	99.944	100	100	100
Class 2	92.17	91.96	91.86	91.51	91.45	91.34
Accuracy	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	96.20	96.12	96.23	96.20	96.18	97.21
Class 1	99.89	99.89	99.94	100	100	100
Class 2	92.24	92.105	92.25	92.14	92.14	92.17

ตารางที่ ข.14 จำนวนรอบของวิธี SKE กับวิธี CKE ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	5	5	5
สูงสุด	6	6	6	7	7	5
ด้วยเฉลี่ย	5.1	5.1	5.2	5.8	6	5
จำนวนรอบ	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	4	4	4
สูงสุด	6	6	6	6	5	4
ด้วยเฉลี่ย	5.2	5	4.9	4.9	4.5	4

ตารางที่ ข.15 ค่า WSCD ของวิธี SKE กับวิธี CKE ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKE	2343.801	2557.561	2980.448	3412.439	3820.53	4263.694
CKE	2344.472	2560.078	2986.472	3422.001	3832.442	4277.72

ตารางที่ ข.16 ค่า $\left|\mu_{ik} - \bar{x}_{ik}\right|, i = 1,2,..., p; k = 1,2,..., K$ ของวิธี SKE กับวิธี CKE ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.0315\\0.0219\\0.0746\\0.0069\end{bmatrix}$	$\begin{bmatrix}0.0317\\0.0260\\0.0742\\0.0051\end{bmatrix}$	$\begin{bmatrix}0.0314\\0.0323\\0.0705\\0.0042\end{bmatrix}$	$\begin{bmatrix}0.0325\\0.0385\\0.0696\\0.0038\end{bmatrix}$	$\begin{bmatrix}0.0320\\0.0403\\0.0696\\0.0033\end{bmatrix}$	$\begin{bmatrix}0.0313\\0.0431\\0.0697\\0.0030\end{bmatrix}$
Class 2	$\begin{bmatrix}0.0161\\0.1480\\0.0673\\0.1141\end{bmatrix}$	$\begin{bmatrix}0.0127\\0.1465\\0.0654\\0.1149\end{bmatrix}$	$\begin{bmatrix}0.0094\\0.1413\\0.0601\\0.1115\end{bmatrix}$	$\begin{bmatrix}0.0042\\0.1390\\0.0563\\0.1127\end{bmatrix}$	$\begin{bmatrix}0.0025\\0.1377\\0.0560\\0.1119\end{bmatrix}$	$\begin{bmatrix}0.0001\\0.1359\\0.0553\\0.1115\end{bmatrix}$
	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.0314\\0.0226\\0.0735\\0.0056\end{bmatrix}$	$\begin{bmatrix}0.0311\\0.1449\\0.0642\\0.1127\end{bmatrix}$	$\begin{bmatrix}0.0298\\0.0290\\0.0690\\0.0022\end{bmatrix}$	$\begin{bmatrix}0.0297\\0.0342\\0.0656\\0.0007\end{bmatrix}$	$\begin{bmatrix}0.0293\\0.0354\\0.0652\\0.0007\end{bmatrix}$	$\begin{bmatrix}0.0292\\0.0348\\0.0649\\0.0006\end{bmatrix}$
Class 2	$\begin{bmatrix}0.0164\\0.1470\\0.0663\\0.1130\end{bmatrix}$	$\begin{bmatrix}0.0139\\0.1449\\0.0642\\0.1127\end{bmatrix}$	$\begin{bmatrix}0.0139\\0.1408\\0.0611\\0.1101\end{bmatrix}$	$\begin{bmatrix}0.0110\\0.1369\\0.0563\\0.1099\end{bmatrix}$	$\begin{bmatrix}0.0102\\0.1359\\0.0558\\0.1095\end{bmatrix}$	$\begin{bmatrix}0.0110\\0.1360\\0.0558\\0.1097\end{bmatrix}$

ข.1.5 ผลการทดลองกับชุดข้อมูล user knowledge modeling

ตารางที่ ข.17 ค่า overall accuracy กับ class accuracy ของวิธี SKE กับวิธี CKE ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	50.201	54.73	54.57	53.93	50.94	44.78
Class 1	72.70	72.31	78.18	83.05	77.27	42.62
Class 2	52.15	54.31	54.98	54.79	54.84	61.70
Class 3	45.66	49.86	49.57	46.85	40.77	21.21
Class 4	31.46	44.14	41.67	38.91	34.22	46.67
Accuracy	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	52.04	54.18	60.84	65.97	67.51	67.66
Class 1	73.51	79.47	86.60	90.60	93.14	93.02
Class 2	52.87	52.91	56.44	58.64	58.94	58.18
Class 3	47.37	50.64	57.49	64.66	64.75	66.13
Class 4	34.42	35.84	48.59	54.07	59.97	56.10

ตารางที่ ข.18 จำนวนรอบของวิธี SKE กับวิธี CKE ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	10	7	10	11	22
สูงสุด	18	29	15	27	26	22
ด้วยเฉลี่ย	12.4	16.8	11.4	16.9	17.1	22
จำนวนรอบ	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	9	7	5	5	5	7
สูงสุด	20	15	15	16	12	7
ด้วยเฉลี่ย	11	11.7	10.3	7.5	7.5	7

ตารางที่ ข.19 ค่า WSCD ของวิธี SKE กับวิธี CKE ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKE	695.5336	763.1799	908.9852	1065.884	1206.064	1351.633
CKE	720.1797	812.7816	1003.487	1198.144	1372.75	1542.023

ตารางที่ ข.20 ค่า $\left|\mu_{ik}-\bar{x}_{ik}\right|, i=1,2,...,p; k=1,2,...,K$ ของวิธี SKE กับวิธี CKE ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Euclidean distance (SKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.3789\\0.5013\\0.1913\\0.3440\\0.2117\end{bmatrix}$	$\begin{bmatrix}0.3783\\0.4699\\0.1458\\0.4254\\0.2003\end{bmatrix}$	$\begin{bmatrix}0.1782\\0.3634\\0.1273\\0.5676\\0.1788\end{bmatrix}$	$\begin{bmatrix}0.1951\\0.1811\\0.2144\\0.5087\\0.1374\end{bmatrix}$	$\begin{bmatrix}0.1707\\0.2155\\0.1366\\0.5392\\0.1604\end{bmatrix}$	$\begin{bmatrix}0.2641\\00228\\0.0475\\1.0840\\0.3882\end{bmatrix}$
Class 2	$\begin{bmatrix}0.1700\\0.1821\\0.4338\\0.4886\\0.1112\end{bmatrix}$	$\begin{bmatrix}0.0749\\0.1679\\0.4193\\0.6041\\0.0932\end{bmatrix}$	$\begin{bmatrix}0.1107\\0.0755\\0.2925\\0.7524\\0.0980\end{bmatrix}$	$\begin{bmatrix}0.1922\\0.0755\\0.2019\\0.9021\\0.1052\end{bmatrix}$	$\begin{bmatrix}0.2871\\0.0998\\0.2474\\0.8082\\0.1154\end{bmatrix}$	$\begin{bmatrix}0.3196\\0.1580\\0.3174\\0.8287\\0.0819\end{bmatrix}$
Class 3	$\begin{bmatrix}0.3100\\0.7630\\0.3790\\0.3083\\0.3145\end{bmatrix}$	$\begin{bmatrix}0.2572\\0.7675\\0.2739\\0.3526\\0.2927\end{bmatrix}$	$\begin{bmatrix}0.1620\\0.5157\\0.3088\\0.4408\\0.3208\end{bmatrix}$	$\begin{bmatrix}0.2325\\0.1528\\0.4310\\0.5277\\0.3794\end{bmatrix}$	$\begin{bmatrix}0.2888\\0.1790\\0.4655\\0.7142\\0.3727\end{bmatrix}$	$\begin{bmatrix}0.4211\\0.1338\\0.0430\\1.2997\\0.6706\end{bmatrix}$
Class 4	$\begin{bmatrix}0.4875\\0.2958\\0.6079\\0.2324\\0.8043\end{bmatrix}$	$\begin{bmatrix}0.4709\\0.3252\\0.5698\\0.1946\\0.6370\end{bmatrix}$	$\begin{bmatrix}0.4146\\0.0964\\0.4703\\0.1971\\0.5541\end{bmatrix}$	$\begin{bmatrix}0.6096\\0.1382\\0.3588\\0.2536\\0.7276\end{bmatrix}$	$\begin{bmatrix}0.5598\\0.0902\\0.4811\\0.2923\\0.8781\end{bmatrix}$	$\begin{bmatrix}0.4443\\0.0012\\0.1095\\0.1626\\0.4664\end{bmatrix}$
	Constrained K-means with Euclidean distance (CKE)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.3519\\0.4622\\0.1914\\0.3458\\0.1844\end{bmatrix}$	$\begin{bmatrix}0.3294\\0.4023\\0.1118\\0.3490\\0.1286\end{bmatrix}$	$\begin{bmatrix}0.1725\\0.2681\\0.1259\\0.4574\\0.1000\end{bmatrix}$	$\begin{bmatrix}0.1410\\0.1678\\0.1218\\0.4052\\0.0662\end{bmatrix}$	$\begin{bmatrix}0.0904\\0.1374\\0.1032\\0.3476\\0.0275\end{bmatrix}$	$\begin{bmatrix}0.0912\\0.1079\\0.1002\\0.3570\\0.0383\end{bmatrix}$

Class 2	$\begin{bmatrix}0.2112\\0.1552\\0.3843\\0.4674\\0.1364\end{bmatrix}$	$\begin{bmatrix}0.1936\\0.1424\\0.3250\\0.4845\\0.1434\end{bmatrix}$	$\begin{bmatrix}0.1238\\0.0465\\0.2158\\0.6889\\0.1313\end{bmatrix}$	$\begin{bmatrix}0.0830\\0.0557\\0.1114\\0.7897\\0.1271\end{bmatrix}$	$\begin{bmatrix}0.0869\\0.0490\\0.0702\\0.7917\\0.1376\end{bmatrix}$	$\begin{bmatrix}0.0707\\0.0329\\0.0179\\0.7793\\0.1468\end{bmatrix}$
Class 3	$\begin{bmatrix}0.3307\\0.6954\\0.4429\\0.3157\\0.2028\end{bmatrix}$	$\begin{bmatrix}0.3614\\0.5666\\0.2789\\0.4584\\0.2313\end{bmatrix}$	$\begin{bmatrix}0.1690\\0.3346\\0.2450\\0.4717\\0.2917\end{bmatrix}$	$\begin{bmatrix}0.0727\\0.2009\\0.2192\\0.5531\\0.3114\end{bmatrix}$	$\begin{bmatrix}0.0536\\0.1666\\0.2538\\0.5816\\0.2228\end{bmatrix}$	$\begin{bmatrix}0.0037\\0.1366\\0.2493\\0.5881\\0.2336\end{bmatrix}$
Class 4	$\begin{bmatrix}0.4770\\0.3081\\0.7916\\0.1849\\0.6467\end{bmatrix}$	$\begin{bmatrix}0.5115\\0.3367\\0.6315\\0.1917\\0.6474\end{bmatrix}$	$\begin{bmatrix}0.3311\\0.1030\\0.4848\\0.2948\\0.4463\end{bmatrix}$	$\begin{bmatrix}0.3088\\0.0955\\0.2499\\0.3008\\0.3684\end{bmatrix}$	$\begin{bmatrix}0.3319\\0.1036\\0.2170\\0.3121\\0.3675\end{bmatrix}$	$\begin{bmatrix}0.2860\\0.0186\\0.1784\\0.3610\\0.3507\end{bmatrix}$

ข.2 ผลการทดลองที่ 2 การใช้เกณฑ์วัดระยะห่างแบบ Mahalanobis

ข.2.1 ผลการทดลองกับชุดข้อมูล iris

ตารางที่ ข.21 ค่า overall accuracy กับ class accuracy ของวิธี SKM กับวิธี CKM ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	77.07	80.93	89.07	96.40	96.53	96
Class 1	96.92	100	100	100	100	100
Class 2	65.56	68.24	91.43	99.60	100	100
Class 3	72.95	82.17	81.76	91.55	91.57	90.32
Accuracy	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	78.40	85.33	93.20	97.20	97.87	98.67
Class 1	98.46	100	100	100	100	100
Class 2	67.52	75.69	98.82	99.60	100	100
Class 3	73.60	86.33	85.99	93.42	94.68	96.55

ตารางที่ ข.22 จำนวนรอบของวิธี SKM กับวิธี CKM ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	3	5	3	3	3	4
สูงสุด	12	9	12	6	6	4
ด้วยเฉลี่ย	7.2	7.7	5.3	4.2	4.8	4
จำนวนรอบ	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	5	5	3	3	3	3
สูงสุด	14	19	9	4	5	3
ด้วยเฉลี่ย	7.8	8.3	4.2	3.3	3.4	3

ตารางที่ ข.23 ค่า WSCD ของวิธี SKM กับวิธี CKM ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKM	320	348	408	468	528	588
CKM	320	348	408	468	528	588

ตารางที่ ข.24 ค่า $\left|\mu_{ik}-\bar{x}_{ik}\right|, i=1,2,...,p; k=1,2,...,K$ ของวิธี SKM กับวิธี CKM ของชุดข้อมูล iris ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1451\\0.1013\\0.0795\\0.1157\end{bmatrix}$	$\begin{bmatrix}0.1858\\0.1192\\0.0644\\0.0500\end{bmatrix}$	$\begin{bmatrix}0.1709\\0.0737\\0.0492\\0.0299\end{bmatrix}$	$\begin{bmatrix}0.1704\\0.0760\\0.0376\\0.0156\end{bmatrix}$	$\begin{bmatrix}0.1704\\0.0706\\0.0376\\0.0156\end{bmatrix}$	$\begin{bmatrix}0.1704\\0.0760\\0.0376\\0.0156\end{bmatrix}$
Class 2	$\begin{bmatrix}0.3709\\0.2637\\0.2267\\0.3787\end{bmatrix}$	$\begin{bmatrix}0.2421\\0.1793\\0.1794\\0.3151\end{bmatrix}$	$\begin{bmatrix}0.1834\\0.2341\\0.1819\\0.1111\end{bmatrix}$	$\begin{bmatrix}0.0240\\0.0673\\0.0577\\0.0454\end{bmatrix}$	$\begin{bmatrix}0.0243\\0.0693\\0.0614\\0.0429\end{bmatrix}$	$\begin{bmatrix}0.0060\\0.0584\\0.0560\\0.0531\end{bmatrix}$
Class 3	$\begin{bmatrix}0.4110\\0.1780\\0.2643\\0.2790\end{bmatrix}$	$\begin{bmatrix}0.3285\\0.2135\\0.2793\\0.3376\end{bmatrix}$	$\begin{bmatrix}0.1724\\0.1503\\0.1845\\0.1897\end{bmatrix}$	$\begin{bmatrix}0.0848\\0.1192\\0.1554\\0.2403\end{bmatrix}$	$\begin{bmatrix}0.0854\\0.1204\\0.1536\\0.2335\end{bmatrix}$	$\begin{bmatrix}0.0735\\0.1388\\0.1630\\0.2264\end{bmatrix}$
	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1583\\0.0859\\0.0722\\0.0824\end{bmatrix}$	$\begin{bmatrix}0.1857\\0.1200\\0.0605\\0.0452\end{bmatrix}$	$\begin{bmatrix}0.1706\\0.0752\\0.0414\\0.0204\end{bmatrix}$	$\begin{bmatrix}0.1704\\0.0760\\0.0376\\0.0156\end{bmatrix}$	$\begin{bmatrix}0.1704\\0.0760\\0.0376\\0.0156\end{bmatrix}$	$\begin{bmatrix}0.1704\\0.0760\\0.0376\\0.0156\end{bmatrix}$
Class 2	$\begin{bmatrix}0.3797\\0.2510\\0.2388\\0.3813\end{bmatrix}$	$\begin{bmatrix}0.2250\\0.0697\\0.1263\\0.2935\end{bmatrix}$	$\begin{bmatrix}0.0792\\0.1070\\0.1099\\0.0695\end{bmatrix}$	$\begin{bmatrix}0.0513\\0.0838\\0.0658\\0.0301\end{bmatrix}$	$\begin{bmatrix}0.0698\\0.0967\\0.0750\\0.0173\end{bmatrix}$	$\begin{bmatrix}0.0971\\0.1132\\0.0831\\0.0020\end{bmatrix}$
Class 3	$\begin{bmatrix}0.3631\\0.1641\\0.2612\\0.2670\end{bmatrix}$	$\begin{bmatrix}0.2745\\0.1444\\0.2028\\0.3324\end{bmatrix}$	$\begin{bmatrix}0.1240\\0.1504\\0.1573\\0.1733\end{bmatrix}$	$\begin{bmatrix}0.1027\\0.0916\\0.1413\\0.2510\end{bmatrix}$	$\begin{bmatrix}0.1151\\0.0744\\0.1301\\0.2513\end{bmatrix}$	$\begin{bmatrix}0.1329\\0.0467\\0.1161\\0.2620\end{bmatrix}$

ข.2.1 ผลการทดลองกับชุดข้อมูล seeds

ตารางที่ ข.25 ค่า overall accuracy กับ class accuracy ของวิธี SKM กับวิธี CKM ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	87.24	91.24	87.05	89.43	90.48	92.38
Class 1	85.51	87.34	88.86	89.23	88.89	91.38
Class 2	89.86	96.56	90.41	93.94	97.88	100
Class 3	90.76	92.48	84.34	86.29	85.95	86.84
Accuracy	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	87.24	91.24	87.81	89.91	90.76	92.38
Class 1	85.51	87.09	89.14	89.58	89.18	91.43
Class 2	89.86	96.56	91.86	95.07	98.49	100
Class 3	90.76	92.30	84.34	86.29	85.95	86.84

ตารางที่ ข.26 จำนวนรอบของวิธี SKM กับวิธี CKM ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	3	4	4	4
สูงสุด	18	8	11	6	6	4
ด้วยเฉลี่ย	8.6	5.8	5.2	5	4.8	4
จำนวนรอบ	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	3	3	3	3
สูงสุด	18	8	6	6	4	3
ด้วยเฉลี่ย	8.5	5.8	4.5	3.6	3.3	3

ตารางที่ ข.27 ค่า WSCD ของวิธี SKM กับวิธี CKM ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKM	784	861	1008	1155	1302	1449
CKM	784	861	1008	1155	1302	1449

ตารางที่ ข.28 ค่า $\left|\mu_{ik}-\bar{x}_{ik}\right|, i=1,2,...,p; k=1,2,...,K$ ของวิธี SKM กับวิธี CKM ของชุดข้อมูล seeds ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.0386\\0.2615\\0.1394\\0.2294\\0.1287\\0.1730\\0.1066\end{bmatrix}$	$\begin{bmatrix}0.0439\\0.1266\\0.0888\\0.1491\\0.0766\\0.1737\\0.1319\end{bmatrix}$	$\begin{bmatrix}0.0627\\0.0712\\0.1950\\0.1058\\0.0688\\0.1652\\0.0790\end{bmatrix}$	$\begin{bmatrix}0.0268\\0.0624\\0.1299\\0.1357\\0.0510\\0.1489\\0.0748\end{bmatrix}$	$\begin{bmatrix}0.0264\\0.0563\\0.1478\\0.1206\\0.0523\\0.1662\\0.0667\end{bmatrix}$	$\begin{bmatrix}0.0074\\0.0552\\0.0517\\0.1142\\0.0531\\0.1514\\0.1013\end{bmatrix}$
Class 2	$\begin{bmatrix}0.1760\\0.0970\\0.0732\\0.0511\\0.1038\\0.0723\\0.1588\end{bmatrix}$	$\begin{bmatrix}0.1052\\0.1393\\0.0679\\0.0638\\0.0755\\0.0465\\0.0499\end{bmatrix}$	$\begin{bmatrix}0.1062\\0.1036\\0.1073\\0.0482\\0.0790\\0.0641\\0.1523\end{bmatrix}$	$\begin{bmatrix}0.0856\\0.0896\\0.0924\\0.0434\\0.0676\\0.0583\\0.0901\end{bmatrix}$	$\begin{bmatrix}0.0699\\0.0902\\0.0380\\0.0767\\0.0534\\0.0289\\0.0167\end{bmatrix}$	$\begin{bmatrix}0.0462\\0.0899\\0.0199\\0.0901\\0.0324\\0.0236\\0.0272\end{bmatrix}$
Class 3	$\begin{bmatrix}0.0644\\0.1027\\0.0875\\0.0819\\0.0606\\0.2579\\0.0800\end{bmatrix}$	$\begin{bmatrix}0.0341\\0.0626\\0.0452\\0.0615\\0.0432\\0.2590\\0.0843\end{bmatrix}$	$\begin{bmatrix}0.0254\\0.0488\\0.0672\\0.0633\\0.0332\\0.3985\\0.0652\end{bmatrix}$	$\begin{bmatrix}0.0231\\0.0476\\0.0543\\0.0540\\0.0311\\0.3464\\0.0400\end{bmatrix}$	$\begin{bmatrix}0.0230\\0.0448\\0.0518\\0.0458\\0.0354\\0.3381\\0.0337\end{bmatrix}$	$\begin{bmatrix}0.0126\\0.0204\\0.0196\\0.0431\\0.0581\\0.3197\\0.0476\end{bmatrix}$
	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.0386\\0.2615\\0.1394\\0.2294\\0.1287\\0.1730\\0.1066\end{bmatrix}$	$\begin{bmatrix}0.0453\\0.1226\\0.0880\\0.1551\\0.0737\\0.1756\\0.1289\end{bmatrix}$	$\begin{bmatrix}0.0495\\0.0550\\0.1925\\0.1054\\0.0686\\0.1725\\0.0722\end{bmatrix}$	$\begin{bmatrix}0.0271\\0.0632\\0.1181\\0.1305\\0.0517\\0.1448\\0.0691\end{bmatrix}$	$\begin{bmatrix}0.0271\\0.0605\\0.1407\\0.1225\\0.0530\\0.1670\\0.0659\end{bmatrix}$	$\begin{bmatrix}0.0074\\0.0552\\0.0517\\0.1142\\0.0531\\0.1514\\0.1013\end{bmatrix}$
Class 2	$\begin{bmatrix}0.1760\\0.0970\\0.0732\\0.0511\\0.1038\\0.0723\\0.1588\end{bmatrix}$	$\begin{bmatrix}0.1052\\0.1393\\0.0679\\0.0638\\0.0755\\0.0465\\0.0499\end{bmatrix}$	$\begin{bmatrix}0.0867\\0.1010\\0.1028\\0.0577\\0.0698\\0.0643\\0.1321\end{bmatrix}$	$\begin{bmatrix}0.0829\\0.0935\\0.0872\\0.0487\\0.0552\\0.0599\\0.0920\end{bmatrix}$	$\begin{bmatrix}0.0722\\0.0930\\0.0334\\0.0758\\0.0479\\0.0300\\0.0180\end{bmatrix}$	$\begin{bmatrix}0.0462\\0.0899\\0.0199\\0.0901\\0.0324\\0.0236\\0.0272\end{bmatrix}$
Class 3	$\begin{bmatrix}0.0644\\0.1027\\0.0875\\0.0819\\0.0606\\0.2579\\0.0800\end{bmatrix}$	$\begin{bmatrix}0.0363\\0.0624\\0.0429\\0.0573\\0.0453\\0.2565\\0.0896\end{bmatrix}$	$\begin{bmatrix}0.0254\\0.0488\\0.0672\\0.0633\\0.0332\\0.3985\\0.0652\end{bmatrix}$	$\begin{bmatrix}0.0231\\0.0476\\0.0543\\0.0540\\0.0311\\0.3464\\0.0400\end{bmatrix}$	$\begin{bmatrix}0.0230\\0.0448\\0.0518\\0.0458\\0.0354\\0.3381\\0.0337\end{bmatrix}$	$\begin{bmatrix}0.0126\\0.0204\\0.0196\\0.0431\\0.0581\\0.3197\\0.0476\end{bmatrix}$

ข.2.1 ผลการทดลองกับชุดข้อมูล wine

ตารางที่ ข.29 ค่า overall accuracy กับ class accuracy ของวิธี SKM กับวิธี CKM ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	75.51	84.61	90.45	92.36	92.70	96.63
Class 1	78.38	80.12	89.21	92.49	96.57	96.30
Class 2	76.79	87.45	89.05	92.00	87.51	94.87
Class 3	76.30	91.09	96.55	97.40	99.60	100
Accuracy	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	75.28	85.40	90.11	93.26	92.93	95.51
Class 1	78.82	80.32	88.81	92.49	97.18	92.86
Class 2	76.76	89.30	88.43	93.52	87.83	94.74
Class 3	75.43	91.32	96.55	97.58	100	100

ตารางที่ ข.30 จำนวนรอบของวิธี SKM กับวิธี CKM ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	4	4	6
สูงสุด	12	9	6	8	9	6
ด้วยเฉลี่ย	6.9	5.8	4.8	5.4	5.2	6
จำนวนรอบ	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	4	4	4	4	4
สูงสุด	9	9	6	7	6	4
ด้วยเฉลี่ย	6.2	5.4	4.9	4.9	4.3	4

ตารางที่ ข.31 ค่า WSCD ของวิธี SKM กับวิธี CKM ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKM	1140	1248	1464	1668	1883.754	2100
CKM	1131.173	1248	1464	1668	1884	2100

ตารางที่ ข.32 ค่า $\left|\mu_{ik}-\bar{x}_{ik}\right|, i=1,2,...,p; k=1,2,...,K$ ของวิธี SKM กับวิธี CKM ของชุดข้อมูล wine ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.2193\\0.1903\\0.2482\\0.2949\\0.3304\\0.0847\\0.1774\\0.1981\\0.2502\\0.1076\\0.1955\\0.3270\end{bmatrix}$	$\begin{bmatrix}0.1441\\0.2185\\0.1656\\0.1947\\0.1736\\0.0825\\0.2091\\0.1189\\0.1978\\0.1131\\0.1277\\0.2852\end{bmatrix}$	$\begin{bmatrix}0.1359\\0.1241\\0.0745\\0.1857\\0.1470\\0.0585\\0.1793\\0.0913\\0.2162\\0.0754\\0.1163\\0.2344\end{bmatrix}$	$\begin{bmatrix}0.1203\\0.0809\\0.0639\\0.2342\\0.1160\\0.0849\\0.2094\\0.0861\\0.1661\\0.1001\\0.1304\\0.2307\end{bmatrix}$	$\begin{bmatrix}0.0759\\0.2402\\0.1167\\0.1296\\0.0716\\0.074\\0.2227\\0.0847\\0.2469\\0.2149\\0.0921\\0.2477\end{bmatrix}$	$\begin{bmatrix}0.0632\\0.1896\\0.0687\\0.1982\\0.0246\\0.0001\\0.2565\\0.0520\\0.2077\\0.1506\\0.0969\\0.2134\end{bmatrix}$
Class 2	$\begin{bmatrix}0.2399\\0.2717\\0.1028\\0.1621\\0.1765\\0.1244\\0.1246\\0.1759\\0.1431\\0.1951\\0.1423\\0.1233\end{bmatrix}$	$\begin{bmatrix}0.2451\\0.2121\\0.1639\\0.1029\\0.1457\\0.0742\\0.1536\\0.0967\\0.0913\\0.0574\\0.0580\\0.1229\end{bmatrix}$	$\begin{bmatrix}0.2745\\0.2229\\0.1236\\0.0568\\0.1000\\0.0738\\0.0811\\0.1564\\0.1386\\0.0667\\0.0495\\0.1101\end{bmatrix}$	$\begin{bmatrix}0.2060\\0.1661\\01196\\0.0715\\0.0753\\0.0929\\0.0788\\0.1290\\0.0925\\0.0669\\0.0437\\0.0589\end{bmatrix}$	$\begin{bmatrix}0.2374\\0.2595\\0.0636\\0.0660\\0.1172\\0.0908\\0.1298\\0.1205\\0.0917\\0.0531\\0.0372\\0.0555\end{bmatrix}$	$\begin{bmatrix}0.1978\\0.2768\\0.0949\\0.0597\\0.0347\\0.0953\\0.0426\\0.1355\\0.1036\\0.0174\\0.0269\\0.0888\end{bmatrix}$
Class 3	$\begin{bmatrix}0.2440\\0.3078\\0.1865\\0.1594\\0.2852\\0.1404\\0.2542\\0.5413\\0.1595\\0.6171\\0.1971\\0.1412\end{bmatrix}$	$\begin{bmatrix}0.0713\\0.2468\\0.0834\\0.1217\\0.1331\\0.0956\\0.1147\\0.2238\\0.1039\\0.3981\\0.1153\\0.0992\end{bmatrix}$	$\begin{bmatrix}0.1022\\0.2035\\0.0647\\0.0477\\0.1085\\0.0674\\0.1029\\0.1672\\0.1065\\0.3351\\0.0968\\0.1089\end{bmatrix}$	$\begin{bmatrix}0.0746\\0.2387\\0.0703\\0.0476\\0.0804\\0.0615\\0.0570\\0.1221\\0.1404\\0.3873\\0.0587\\0.0955\end{bmatrix}$	$\begin{bmatrix}0.0604\\0.2560\\0.0646\\0.0447\\0.0497\\0.0871\\0.0559\\0.0454\\0.1774\\0.4086\\0.0419\\0.0420\end{bmatrix}$	$\begin{bmatrix}0.0859\\0.2405\\0.0371\\0.0475\\0.0487\\0.0922\\0.0634\\0.0396\\0.1822\\0.3855\\0.0372\\0.0117\end{bmatrix}$

	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1986\\0.1781\\0.2160\\0.2662\\0.3328\\0.0824\\0.2058\\0.1931\\0.2463\\0.0907\\0.1810\\0.3258\end{bmatrix}$	$\begin{bmatrix}0.1417\\0.2092\\0.1579\\0.2068\\0.1708\\0.0820\\0.2048\\0.1227\\0.2045\\0.1065\\0.1330\\0.2839\end{bmatrix}$	$\begin{bmatrix}0.1309\\0.1310\\0.0912\\0.1786\\0.1560\\0.0534\\0.1911\\0.0968\\0.2156\\0.0762\\0.1150\\0.2354\end{bmatrix}$	$\begin{bmatrix}0.1196\\0.0830\\0.0551\\0.2273\\0.1102\\0.0769\\0.2154\\0.0805\\0.1710\\0.1068\\0.1141\\0.2310\end{bmatrix}$	$\begin{bmatrix}0.0838\\0.2105\\0.0964\\0.1143\\0.0786\\0.0355\\0.1896\\0.0938\\0.2024\\0.1394\\0.1033\\0.2421\end{bmatrix}$	$\begin{bmatrix}0.0835\\0.2120\\0.0199\\0.1699\\0.0953\\0.0132\\0.2593\\0.0527\\0.2321\\0.1626\\0.0998\\0.2144\end{bmatrix}$
Class 2	$\begin{bmatrix}0.2259\\0.2698\\0.0924\\0.1522\\0.1769\\0.1280\\0.1289\\0.1754\\0.1474\\0.2020\\0.1422\\0.1152\end{bmatrix}$	$\begin{bmatrix}0.2275\\0.2012\\0.1429\\0.1027\\0.1500\\0.0777\\0.1551\\0.1997\\0.0790\\0.0543\\0.0502\\0.1168\end{bmatrix}$	$\begin{bmatrix}0.2793\\0.2251\\0.1188\\0.0608\\0.1084\\0.0692\\0.0831\\0.1609\\0.1449\\0.0654\\0.0515\\0.1066\end{bmatrix}$	$\begin{bmatrix}0.1801\\0.1511\\0.1311\\0.0784\\0.0798\\0.0944\\0.0604\\0.1433\\0.1154\\0.0599\\0.0364\\0.0554\end{bmatrix}$	$\begin{bmatrix}0.2290\\0.2695\\0.0431\\0.0624\\0.1147\\0.0792\\0.1214\\0.1138\\0.0928\\0.0711\\0.0355\\0.0498\end{bmatrix}$	$\begin{bmatrix}0.1900\\0.3001\\0.0503\\0.0236\\0.0683\\0.1134\\0.0214\\0.1348\\0.0847\\0.0529\\0.0232\\0.0950\end{bmatrix}$
Class 3	$\begin{bmatrix}0.2605\\0.3253\\0.2152\\0.1689\\0.2731\\0.1316\\0.2948\\0.5194\\0.1771\\0.5900\\0.2154\\0.1809\end{bmatrix}$	$\begin{bmatrix}0.0612\\0.2436\\0.0749\\0.1183\\0.1344\\0.0926\\0.1117\\0.2152\\0.0960\\0.4019\\0.1130\\0.1014\end{bmatrix}$	$\begin{bmatrix}0.0971\\0.2081\\0.0687\\0.0484\\0.1116\\0.0676\\0.1013\\0.1689\\0.1112\\0.3675\\0.0925\\0.1128\end{bmatrix}$	$\begin{bmatrix}0.0506\\0.2435\\0.0729\\0.0481\\0.0869\\0.0625\\0.0432\\0.1014\\0.1407\\0.4110\\0.0585\\0.0803\end{bmatrix}$	$\begin{bmatrix}0.0699\\0.2427\\0.0379\\0.0501\\0.0398\\0.0938\\0.0501\\0.0425\\0.1741\\0.3789\\0.0395\\0.0415\end{bmatrix}$	$\begin{bmatrix}0.0859\\0.2405\\0.0371\\0.0475\\0.0487\\0.0922\\0.0634\\0.0396\\0.1822\\0.3855\\0.0372\\0.0117\end{bmatrix}$

ข.2.1 ผลการทดลองกับชุดข้อมูล banknote authentication

ตารางที่ ข.33 ค่า overall accuracy กับ class accuracy ของวิธี SKM กับวิธี CKM ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	97.38	97.58	97.37	97.41	97.19	95.34
Class 1	100	100	100	100	100	100
Class 2	94.41	94.86	94.44	94.52	93.62	90.53
Accuracy	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	97.40	97.67	97.81	97.81	97.81	97.81
Class 1	100	100	100	100	100	100
Class 2	94.53	95.04	95.33	95.33	95.33	95.33

ตารางที่ ข.34 จำนวนรอบของวิธี SKM กับวิธี CKM ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	4	3	3	3	3	12
สูงสุด	8	6	11	11	12	12
ด้วยเฉลี่ย	4.4	4.1	5.1	5	6.4	12
จำนวนรอบ	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	3	3	3	3	3	3
สูงสุด	8	4	4	4	4	3
ด้วยเฉลี่ย	4.1	3.6	3.4	3.2	3.1	3

ตารางที่ ข.35 ค่า WSCD ของวิธี SKM กับวิธี CKM ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKM	3012	3284	3832	4384	4908	5480
CKM	3012	3284	3832	4384	4908	5480

ตารางที่ ข.36 ค่า $\left|\mu_{ik}-\bar{x}_{ik}\right|, i=1,2,...,p; k=1,2,...,K$ ของวิธี SKM กับวิธี CKM ของชุดข้อมูล banknote authentication ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.0067\\0.0397\\0.0548\\0.0324\end{bmatrix}$	$\begin{bmatrix}0.0030\\0.0397\\0.0521\\0.0354\end{bmatrix}$	$\begin{bmatrix}0.0068\\0.0397\\0.0545\\0.0340\end{bmatrix}$	$\begin{bmatrix}0.0076\\0.0398\\0.0535\\0.0357\end{bmatrix}$	$\begin{bmatrix}0.0148\\0.0397\\0.0589\\0.0313\end{bmatrix}$	$\begin{bmatrix}0.0391\\0.0396\\0.0772\\0.0152\end{bmatrix}$
Class 2	$\begin{bmatrix}0.0181\\0.1069\\0.0590\\0.0824\end{bmatrix}$	$\begin{bmatrix}0.0197\\0.1022\\0.0590\\0.0768\end{bmatrix}$	$\begin{bmatrix}0.0184\\0.1067\\0.0588\\0.0807\end{bmatrix}$	$\begin{bmatrix}0.0189\\0.1056\\0.0583\\0.0789\end{bmatrix}$	$\begin{bmatrix}0.0158\\0.1153\\0.0582\\0.0883\end{bmatrix}$	$\begin{bmatrix}0.0053\\0.1482\\0.0577\\0.1221\end{bmatrix}$
	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.0069\\0.0400\\0.0536\\0.0341\end{bmatrix}$	$\begin{bmatrix}0.0034\\0.0401\\0.0503\\0.0371\end{bmatrix}$	$\begin{bmatrix}0.0058\\0.0404\\0.0471\\0.0453\end{bmatrix}$	$\begin{bmatrix}0.0058\\0.0404\\0.0471\\0.0453\end{bmatrix}$	$\begin{bmatrix}0.0058\\0.0404\\0.0471\\0.0453\end{bmatrix}$	$\begin{bmatrix}0.0058\\0.0404\\0.0471\\0.0453\end{bmatrix}$
Class 2	$\begin{bmatrix}0.0186\\0.1054\\0.0584\\0.0807\end{bmatrix}$	$\begin{bmatrix}0.0204\\0.0998\\0.0581\\0.0752\end{bmatrix}$	$\begin{bmatrix}0.0223\\0.0964\\0.0563\\0.0664\end{bmatrix}$	$\begin{bmatrix}0.0223\\0.0964\\0.0563\\0.0664\end{bmatrix}$	$\begin{bmatrix}0.0223\\0.0964\\0.0563\\0.0664\end{bmatrix}$	$\begin{bmatrix}0.0223\\0.0964\\0.0563\\0.0664\end{bmatrix}$

ข.2.1 ผลการทดลองกับชุดข้อมูล user knowledge modeling

ตารางที่ ข.37 ค่า overall accuracy กับ class accuracy ของวิธี SKM กับวิธี CKM ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

Accuracy	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	55.42	62.98	64.28	89.00	91.74	90.55
Class 1	67.63	80.64	72.38	83.37	92.18	91.38
Class 2	52.60	61.94	61.94	90.61	90.71	90.74
Class 3	62.90	70.13	82.08	91.84	97.35	98.04
Class 4	36.22	46.86	57.30	86.87	85.70	78.95
Accuracy	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Overall	62.14	71.74	85.32	93.33	93.68	93.03
Class 1	85.98	95.32	95.99	95.95	96.30	96.30
Class 2	52.34	61.25	78.27	89.86	90.88	91.23
Class 3	70.42	82.58	95.30	96.36	96.36	96.36
Class 4	36.32	50.62	71.51	91.94	90.49	85.71

ตารางที่ ข.38 จำนวนรอบของวิธี SKM กับวิธี CKM ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

จำนวนรอบ	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	6	7	7	4	5	11
สูงสุด	23	30	27	12	31	11
ด้วยเฉลี่ย	11.8	13.1	15.9	7.9	11.1	11
จำนวนรอบ	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
ต่ำสุด	5	6	5	3	3	4
สูงสุด	21	17	21	7	7	4
ด้วยเฉลี่ย	10.2	9.4	10.1	4.3	4.2	4

ตารางที่ ข.39 ค่า WSCD ของวิธี SKM กับวิธี CKM ของชุดข้อมูล user knowledge modeling M ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

วิธี	n_l					
	5%	10%	20%	30%	40%	50%
SKM	1085	1185	1390	1590	1790	1995
CKM	1085	1185	1390	1590	1790	1995

ตารางที่ ข.40 ค่า $\left|\mu_{ik}-\bar{x}_{ik}\right|, i=1,2,...,p; k=1,2,...,K$ ของวิธี SKM กับวิธี CKM ของชุดข้อมูล user knowledge modeling ด้วยจำนวนข้อมูลที่กำกับกลุ่ม (n_l) = 5% 10% 20% 30% 40% และ 50%

	Seeded K-means with Mahalanobis distance (SKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1940\\0.2960\\0.2157\\0.3771\\0.5452\end{bmatrix}$	$\begin{bmatrix}0.2332\\0.0965\\0.3141\\0.2891\\0.2426\end{bmatrix}$	$\begin{bmatrix}0.1953\\0.1139\\0.0500\\0.2914\\0.2208\end{bmatrix}$	$\begin{bmatrix}0.0911\\0.0303\\0.0580\\0.0331\\0.0561\end{bmatrix}$	$\begin{bmatrix}0.1071\\0.0710\\0.0933\\0.523\\0.1951\end{bmatrix}$	$\begin{bmatrix}0.0848\\0.0186\\0.0798\\0.0023\\0.0727\end{bmatrix}$
Class 2	$\begin{bmatrix}0.1927\\0.1348\\0.2870\\0.2748\\0.2015\end{bmatrix}$	$\begin{bmatrix}0.0771\\0.1777\\0.2805\\0.3789\\0.1755\end{bmatrix}$	$\begin{bmatrix}0.0954\\0.0561\\0.2093\\0.5963\\0.1463\end{bmatrix}$	$\begin{bmatrix}0.1361\\0.0384\\0.0251\\0.1284\\0.0514\end{bmatrix}$	$\begin{bmatrix}0.1415\\0.0396\\0.0293\\0.1732\\0.0359\end{bmatrix}$	$\begin{bmatrix}0.1590\\0.0202\\0.0436\\0.2460\\0.0353\end{bmatrix}$
Class 3	$\begin{bmatrix}0.2557\\0.2340\\0.2288\\0.4321\\0.2217\end{bmatrix}$	$\begin{bmatrix}0.1505\\0.1302\\0.2117\\0.4571\\0.2824\end{bmatrix}$	$\begin{bmatrix}0.6627\\0.2838\\0.2634\\0.6468\\0.1675\end{bmatrix}$	$\begin{bmatrix}0.0324\\0.0271\\0.0588\\0.1176\\0.0133\end{bmatrix}$	$\begin{bmatrix}0.0327\\0.0263\\0.0639\\0.1212\\0.0760\end{bmatrix}$	$\begin{bmatrix}0.0402\\0.0057\\0.0868\\0.1179\\0.0130\end{bmatrix}$
Class 4	$\begin{bmatrix}0.3857\\0.3283\\0.9143\\0.3743\\0.5668\end{bmatrix}$	$\begin{bmatrix}0.3067\\0.3732\\0.6221\\0.3468\\0.3453\end{bmatrix}$	$\begin{bmatrix}0.1392\\0.1146\\0.4852\\0.2477\\0.3004\end{bmatrix}$	$\begin{bmatrix}0.0217\\0.1048\\0.0298\\0.2097\\0.0196\end{bmatrix}$	$\begin{bmatrix}0.0265\\0.0793\\0.0127\\0.2606\\0.0106\end{bmatrix}$	$\begin{bmatrix}0.0267\\0.0166\\0.0145\\0.4077\\0.0084\end{bmatrix}$
	Constrained K-means with Mahalanobis distance (CKM)					
	n_l					
	5%	10%	20%	30%	40%	50%
Class 1	$\begin{bmatrix}0.1618\\0.1644\\0.1834\\0.2057\\0.1349\end{bmatrix}$	$\begin{bmatrix}0.1460\\0.0898\\0.1526\\0.1183\\0.0628\end{bmatrix}$	$\begin{bmatrix}0.1615\\0.0465\\0.0488\\0.0318\\0.0328\end{bmatrix}$	$\begin{bmatrix}0.1268\\0.0380\\0.0499\\0.0253\\0.0062\end{bmatrix}$	$\begin{bmatrix}0.1330\\0.0413\\0.0507\\0.0225\\0.0007\end{bmatrix}$	$\begin{bmatrix}0.1330\\0.0413\\0.0507\\0.0225\\0.0007\end{bmatrix}$

Class 2	$\begin{bmatrix}0.1880\\0.1502\\0.2753\\0.2409\\0.2134\end{bmatrix}$	$\begin{bmatrix}0.1047\\0.1877\\0.2594\\0.2712\\0.1471\end{bmatrix}$	$\begin{bmatrix}0.0888\\0.0662\\0.1412\\0.3079\\0.0827\end{bmatrix}$	$\begin{bmatrix}0.1354\\0.0440\\0.0250\\0.1184\\0.0439\end{bmatrix}$	$\begin{bmatrix}0.1483\\0.0459\\0.0254\\0.1417\\0.0236\end{bmatrix}$	$\begin{bmatrix}0.1724\\0.0349\\0.0384\\0.1869\\0.0101\end{bmatrix}$
Class 3	$\begin{bmatrix}0.1675\\0.1779\\0.1848\\0.4336\\0.1272\end{bmatrix}$	$\begin{bmatrix}0.1377\\0.1660\\0.1211\\0.3890\\0.1614\end{bmatrix}$	$\begin{bmatrix}0.0649\\0.0749\\0.0367\\0.3390\\0.1391\end{bmatrix}$	$\begin{bmatrix}0.0085\\0.0345\\0.0480\\0.0910\\0.0133\end{bmatrix}$	$\begin{bmatrix}0.0033\\0.0370\\0.0489\\0.0829\\0.0105\end{bmatrix}$	$\begin{bmatrix}0.0033\\0.0370\\0.0489\\0.0829\\0.0105\end{bmatrix}$
Class 4	$\begin{bmatrix}0.4683\\0.4227\\0.8461\\0.5189\\0.4154\end{bmatrix}$	$\begin{bmatrix}0.3631\\0.3914\\0.6390\\0.3600\\0.3299\end{bmatrix}$	$\begin{bmatrix}0.1293\\0.1278\\0.2518\\0.3261\\0.1706\end{bmatrix}$	$\begin{bmatrix}0.0254\\0.1130\\0.0294\\0.1860\\0.0193\end{bmatrix}$	$\begin{bmatrix}0.0261\\0.0978\\0.0221\\0.2061\\0.0167\end{bmatrix}$	$\begin{bmatrix}0.0250\\0.0577\\0.0294\\0.3084\\0.0198\end{bmatrix}$