การจัดกลุ่มทับซ้อนที่มีพื้นฐานมาจากการจัดกลุ่มแบบแบ่งส่วนโดยใช้ความสัมพันธ์ของกลุ่มข้อมูล

นายธนวัต ลิมังกูร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

Partition-based Overlapping Clustering using Clusters' Relations

Mr. Tanawat Limungkura

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2017

หัวข้อวิทยานิพนธ์	การจัดกลุ่มทับซ้อนที่มีพื้นฐานมาจากการจัดกลุ่มแบบแบ่งส่วนโดยใช้ความสัมพันธ์ของกลุ่มข้อมูล
โดย	นายธนวัต ลิมังกูร
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. พีรพล เวทีกูล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์

(รองศาสตราจารย์ ดร. สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร. โชติรัตน์ รัตนามหัทธนะ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. พีรพล เวทีกูล)

.....กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร. อานนท์ รุ่งสว่าง)

ธนวัต ลิมังกูร : การจัดกลุ่มทับซ้อนที่มีพื้นฐานมาจากการจัดกลุ่มแบบแบ่งส่วนโดยใช้ความสัมพันธ์ของกลุ่มข้อมูล (Partition-based Overlapping Clustering using Clusters' Relations) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. พีรพล เวทีกูล, 43 หน้า.

การจัดกลุ่มแบบดั้งเดิมตั้งอยู่บนสมมติฐานที่ว่าข้อมูลหนึ่งตัวอย่างสามารถถูกจัดให้อยู่ได้เพียงกลุ่มเดียว อย่างไรก็ตามการจัดกลุ่มประเภทนี้ไม่สามารถครอบคลุมข้อมูลได้หมดทุกประเภท สำหรับข้อมูลประเภทหลายหมวดหมู่จำเป็นที่จะต้องใช้การจัดกลุ่มอีกแบบที่อนุญาตให้ข้อมูลหนึ่งตัวอย่างสามารถอยู่ได้หลายกลุ่มข้อมูล และเนื่องจากข้อมูลหนึ่งตัวอย่างสามารถอยู่ได้หลายกลุ่ม ดังนั้นกลุ่มข้อมูลที่เกิดขึ้นจึงสามารถทับซ้อนกัน การจัดกลุ่มประเภทนี้จึงถูกเรียกว่า “การจัดกลุ่มทับซ้อน” งานวิจัยนี้ให้ความสนใจที่การจัดกลุ่มทับซ้อนแบบแบ่งส่วนซึ่งเป็นการจัดกลุ่มทับซ้อนที่มีประสิทธิภาพและสามารถประยุกต์ใช้ได้กับชุดข้อมูลส่วนใหญ่ การจัดกลุ่มทับซ้อนแบบแบ่งส่วนนั้นมักมีพัฒนาการมาจากขั้นตอนวิธี K-Means ซึ่งมีปัญหาสำคัญคือมีการเริ่มต้นกระบวนการโดยการสุ่มเซนทรอยด์ ซึ่งหากเซนทรอยด์ที่ได้จากการสุ่มนี้อยู่ในตำแหน่งที่ไม่เหมาะสม ประสิทธิภาพทางด้านความถูกต้องแม่นยำของการจัดกลุ่มจะถูกอิงอยู่กับค่าต่ำสุดสัมพัทธ์ของฟังก์ชันต้นทุน ซึ่งทำให้ความถูกต้องแม่นยำมีโอกาสที่จะออกมาต่ำกว่าที่ควรจะเป็น การจัดกลุ่มทับซ้อนที่พัฒนามาจากขั้นตอนวิธี K-Means ก็ประสบปัญหานี้เช่นเดียวกันเนื่องจากมีการเริ่มต้นกระบวนการที่เหมือนกัน นอกจากนี้งานวิจัยเกี่ยวกับการจัดกลุ่มทับซ้อนแบบแบ่งส่วนที่ผ่านมายังไม่มีการนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลซึ่งเป็นตัวแปรสำคัญตัวแปรหนึ่งที่มีผลต่อรูปแบบของกลุ่มข้อมูลมาใช้ประโยชน์ งานวิจัยนี้จึงมีจุดประสงค์หลักสองประเด็นคือเพื่อแก้ปัญหาเซนทรอยด์ที่อยู่ในตำแหน่งที่ไม่เหมาะสม โดยการประยุกต์ใช้ขั้นตอนวิธี K-Harmonic-Means และขั้นตอนวิธี ELBG อีกประเด็นหนึ่งคือนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาใช้ประโยชน์ เพื่อเพิ่มประสิทธิภาพของการจัดกลุ่มในด้านความถูกต้องแม่นยำ หลังจากค้นคว้าวิจัยจนได้ขั้นตอนวิธีใหม่และทดสอบกับชุดข้อมูล 20 ชุดข้อมูลพบว่าสามารถแก้ไขปัญหาที่เกิดขึ้น และเพิ่มประสิทธิภาพของการจัดกลุ่มด้านความถูกต้องแม่นยำได้จริง โดยความถูกต้องแม่นยำเพิ่มขึ้นเฉลี่ย 25.68% จากขั้นตอนวิธี OKM ซึ่งนำมาเป็นขั้นตอนวิธีพื้นฐานในการวัดด้วยค่า F1

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2560

5770548121 : MAJOR COMPUTER SCIENCE
KEYWORDS: DATA MINING / CLUSTERING / MACHINE LEARNING / UNSUPERVISED LEARNING

TANAWAT LIMUNGKURA: Partition-based Overlapping Clustering using Clusters' Relations. ADVISOR: ASST. PROF. PEERAPON VATEEKUL, 43 pp.

Traditional clusterings have the assumption that a data point can belong to only a single cluster; however, these kinds of clustering cannot handle all data types. For multi-category data clustering, a data point needs to be allowed to belong to more than one cluster, so called “Overlapping Clustering”. This research focuses on partition-based overlapping clustering that is a high-performance clustering with most data sets. Most of partition-based overlapping clusterings are developed from K-Means. This kind of algorithm has an issue, during the process of assigning centroid, the centroid can be shifted to inappropriate position that will yield poor clustering result. Same as K-Means, the partition-based overlapping clustering also encounter this problem. In addition, for overlapping clustering, information about relation between clusters is crucial, but there is still no research exploiting this information to enhance the clustering performance. This research has the objective to solve the inappropriate position of centroids problem by applying K-Harmonic-Means and ELBG to OKM algorithm. Moreover, the performance of algorithm will be enhanced by embedding clusters’ relations information to the cost function. After test with 20 multi-category data sets, the results show that the issue has been resolved and accuracy in term of F1 is improved from base-line algorithm OKM 25.68% on average.

Department:	Computer Engineering	Student's Signature	.....
Field of Study:	Computer Science	Advisor's Signature	.....
Academic Year:	2017		

กิตติกรรมประกาศ

ขอขอบพระคุณบิดาที่เป็นกำลังใจให้เสมอมาตลอดการเล่าเรียนปริญญาโท

ขอขอบพระคุณผศ.ดร.พีรพล เวทีกูลที่ให้ความกรุณาช่วยเหลือในทุกเรื่อง และให้คำปรึกษาในงานวิจัยตลอดระยะเวลาที่เรียนปริญญาโท

ขอขอบพระคุณรศ.ดร.โชติรัตน์ รัตนามหัทธนะ และรศ.ดร.อานนท์ รุ่งสว่างที่ให้เกียรติมาเป็นกรรมการสอบวิทยานิพนธ์และให้คำแนะนำอันมีคุณค่าเกี่ยวกับงานวิจัย

ขอขอบพระคุณอาจารย์ในจุฬาลงกรณ์มหาวิทยาลัยที่ประสิทธิ์ประสาทวิชาความรู้ให้

ขอขอบคุณเพื่อน ๆ พี่ ๆ น้อง ๆ กลุ่ม MIND Lab ที่ให้การต้อนรับอย่างอบอุ่นและช่วยให้ความคิดเห็นเกี่ยวกับงานวิจัย

สารบัญ

หน้า

หน้า

หน้า

สารบัญรูป

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การจัดกลุ่มข้อมูล (Clustering) เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูลที่มีบทบาทสำคัญในการจัดการกับข้อมูลที่ไม่กำกับฉลาก โดยใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) การจัดกลุ่มข้อมูลที่เป็นที่นิยมใช้กันโดยทั่วไป ได้แก่ K-Means DBSCAN การจัดกลุ่มตามลำดับชั้น (Hierarchical Clustering) วิธีใช้ตาราง (Grid-based Method) วิธีใช้กราฟ (Graph-based Method) และวิธีใช้แบบจำลอง (Model-based Method) การจัดกลุ่มแต่ละวิธีมีจุดเด่นจุดด้อยรวมถึงประสิทธิภาพแตกต่างกันขึ้นอยู่กับชุดข้อมูลที่ประยุกต์ใช้ อย่างไรก็ตามเทคนิคในการจัดกลุ่มแบบดั้งเดิมตั้งอยู่บนสมมติฐานที่ว่าข้อมูลหนึ่งตัวอย่างสามารถอยู่ได้เพียงกลุ่มเดียว แต่ในบางสถานการณ์ข้อมูลอาจมีความเป็นไปได้ที่จะอยู่ในหลายกลุ่ม เช่น ประเภทของเพลงและภาพยนตร์สามารถมีได้มากกว่าหนึ่งประเภทต่อหนึ่งตัวอย่างข้อมูล ซึ่งข้อมูลเหล่านี้ถูกเรียกว่าเป็นข้อมูลประเภทหลายหมวดหมู่ ในกรณีดังกล่าวนี้การจัดกลุ่มข้อมูลแบบดั้งเดิมที่ไม่สามารถกำหนดกลุ่มให้กับตัวอย่างข้อมูลได้มากกว่าหนึ่งกลุ่มข้อมูลก็ไม่สามารถให้ผลการจัดกลุ่มที่ถูกต้องได้

เพื่อที่จะให้สามารถจัดกลุ่มข้อมูลประเภทหลายหมวดหมู่ให้ถูกต้องตามที่ควรจะเป็น งานวิจัยหลายงานจึงถูกคิดค้นและนำเสนอโดยเรียกการจัดกลุ่มแบบใหม่นี้ว่าการจัดกลุ่มทับซ้อน (Overlapping Clustering) การจัดกลุ่มทับซ้อนต่างจากการจัดกลุ่มแบบดั้งเดิมด้วยมีแนวคิดใหม่ที่ว่าข้อมูลหนึ่งตัวอย่างมีความเป็นไปได้ที่จะอยู่ในหลายกลุ่ม กลุ่มข้อมูลที่เกิดขึ้นจึงทับซ้อนกันได้ การจัดกลุ่มทับซ้อนส่วนใหญ่มีวิวัฒนาการมาจากการจัดกลุ่มที่มีมาอยู่แล้ว เช่น พัฒนามาจากการจัดกลุ่มแบบแบ่งส่วน [1-7] พัฒนามาจากการแบ่งกลุ่มแบบกราฟ [8-11] พัฒนามาจากการจัดกลุ่มตามลำดับชั้น [12] และพัฒนามาจากวิธีใช้แบบจำลอง [13] การจัดกลุ่มประเภทหนึ่งที่ได้รับการยอมรับโดยทั่วไปว่าสามารถประยุกต์ใช้ได้กับชุดข้อมูลส่วนใหญ่ได้อย่างมีประสิทธิภาพคือการแบ่งกลุ่มแบบแบ่งส่วน งานวิจัยส่วนใหญ่ในหมวดหมู่นี้จะมีขั้นตอนวิธี K-Means เป็นพื้นฐาน อาทิเช่น ALS [14] OKM [1] และ NEO-K-Means [15] อย่างไรก็ตามการจัดกลุ่มแบบทับซ้อนที่พัฒนามาจากการจัดกลุ่มแบบ K-Means ยังคงมีปัญหาสำคัญอยู่

โดยทั่วไปการจัดกลุ่มแบบ K-Means รวมถึงงานวิจัยที่พัฒนาต่อยอดมา จะเริ่มต้นการทำขั้นตอนวิธีโดยการสุ่มตัวอย่างข้อมูลขึ้นมาใช้เป็นเซนทรอยด์ตั้งต้นของกลุ่ม จากนั้นจะทำการวนซ้ำ

กระบวนการจนได้ผลการจัดกลุ่มที่ดีที่สุด อย่างไรก็ตามผลที่ดีที่สุดที่ได้นี้เป็นผลที่ดีที่สุดตามค่าต่ำสุดสัมพัทธ์เท่านั้น ซึ่งค่านี้อาจเป็นผลลัพธ์ในการจัดกลุ่มที่ไม่ดีเท่าที่ควรหากเซนทรอยด์ที่สุ่มได้ตอนเริ่มทำขั้นตอนวิธีห่างจากจุดที่เหมาะสมมากเกินไป ขั้นตอนวิธีการจัดกลุ่มแบบทับซ้อนที่พัฒนามาจากการจัดกลุ่มแบบ K-Means ก็ยังคงประสบปัญหาเดียวกัน เนื่องจากมีกระบวนการเริ่มต้นขั้นตอนวิธีที่เหมือนกัน

ตัวอย่างหนึ่งของขั้นตอนวิธีที่พัฒนามาจาก K-Means และมีประสิทธิภาพสูงแต่ยังคงประสบปัญหานี้คือขั้นตอนวิธี Overlapping K-Means (OKM) รวมไปถึงงานวิจัยที่ต่อยอดมาก็ยังคงประสบปัญหาเดิมอยู่เช่นกันไม่ว่าจะเป็น R-OKM [15] ซึ่งเป็น OKM แบบที่มีการใช้ตัวแปร (Parameter) กำหนดสัดส่วนของการทับซ้อนของกลุ่มข้อมูล KOKM [2] ซึ่งเป็นขั้นตอนวิธีแบบที่มีการเอาเคอร์เนล (Kernel) มาใช้เพื่อลดมิติของข้อมูล และ CWOKM [16] ที่มีการนำค่าน้ำหนักสหสัมพันธ์ (Correlation Weight) มาใช้ในการเพิ่มประสิทธิภาพการจัดกลุ่ม หรืองานวิจัยที่ไม่ได้พัฒนามาจาก OKM อย่าง NEO-K-Means ก็ประสบปัญหานี้เช่นเดียวกัน อย่างไรก็ตามขั้นตอนวิธีการจัดกลุ่มแบบ FCM [17] ประสบปัญหานี้น้อยกว่าแต่ประสิทธิภาพในการจัดกลุ่มเมื่อประยุกต์ใช้กับชุดข้อมูลส่วนใหญ่จะพบว่าด้อยกว่าขั้นตอนวิธี OKM หรือ NEO-K-Means

อีกประการหนึ่งงานวิจัยการจัดกลุ่มแบบทับซ้อนแบบแบ่งส่วนที่มีมาแล้ว ยังไม่มีการใช้ประโยชน์จากความสัมพันธ์ระหว่างกลุ่มข้อมูล กลุ่มข้อมูลที่อยู่ใกล้กันมีโอกาสที่จะเกิดการทับซ้อนกันสูงกว่ากลุ่มข้อมูลที่อยู่ไกลกัน องค์ความรู้ นี้สามารถนำไปใช้ประโยชน์ในการพัฒนาการจัดกลุ่มเพื่อให้ผลการจัดกลุ่มมีความถูกต้องแม่นยำเพิ่มมากยิ่งขึ้น

งานวิจัยนี้ได้ให้ความสนใจกับปัญหาดังที่กล่าวมาแล้วข้างต้น โดยเพื่อที่จะแก้ไขข้อจำกัดดังกล่าว ผู้วิจัยได้ทำการวิเคราะห์ปัญหาและแนวทางแก้ไข โดยปัญหาทั้ง 2 ส่วนจะมีแนวทางแก้ไขดังนี้

1) ในส่วนของปัญหาการเริ่มต้นเซนทรอยด์และการได้ผลลัพธ์การจัดกลุ่มที่ติดอยู่กับค่าต่ำสุดสัมพัทธ์ที่ไม่เหมาะสมนั้น ผู้วิจัยได้พบว่าสาเหตุหนึ่งมาจากกระบวนการหาเซนทรอยด์ของกลุ่มข้อมูล เนื่องการกระบวนการหาเซนทรอยด์เดิมนั้นใช้วิธีหาค่าเฉลี่ยจากจำนวนตัวอย่างข้อมูลทั้งหมดในกลุ่มข้อมูล โดยคำนวณตามค่าเฉลี่ยเลขคณิต ซึ่งวิธีการนี้มีความคงทนต่อค่าผิดปกติของข้อมูล (Outlier Data) ต่ำกว่าผลลัพธ์การจัดกลุ่มที่ได้จึงอาจจะมีประสิทธิภาพลดลง เพื่อที่จะแก้ปัญหาข้อนี้ผู้วิจัยจึงนำงานวิจัย K-Harmonic-Means [16] มาประยุกต์ใช้ งานวิจัย K-Harmonic-Means นำเสนอขั้นตอนวิธีการจัดกลุ่มที่แตกต่างจากขั้นตอนวิธี K-Means เดิม โดยการเปลี่ยนฟังก์ชันต้นทุนจากผลรวมของระยะทางของตัวอย่างข้อมูลไปยังกลุ่มข้อมูลที่ใกล้ที่สุด ให้เป็นค่าเฉลี่ยฮาร์มอนิก (Harmonic Mean) ของระยะทางระหว่างตัวอย่างข้อมูลกับกลุ่มข้อมูลแทน ขั้นตอนวิธีนี้

จะให้เซนทรอยด์ที่เหมาะสมกับการจัดกลุ่มข้อมูลมากกว่า ถึงแม้ว่าจะมีการเริ่มต้นกำหนดเซนทรอยด์ที่ไม่ดีก็ตาม ในส่วนของการนำองค์ความรู้เกี่ยวกับความสัมพันธ์ของกลุ่มข้อมูลมาใช้ประโยชน์ ในสมมติฐานที่ว่ากลุ่มข้อมูลที่อยู่ใกล้กันจะมีโอกาสทับซ้อนกันมากกว่ากลุ่มที่ไกลกันนั้น ผู้วิจัยมีแนวทางในการพัฒนาขั้นตอนวิธีโดยใช้ระยะทางแบบยุคลิด (Euclidean Distance) ระหว่างกลุ่มข้อมูลเป็นตัวแทนความใกล้กันระหว่างกลุ่มข้อมูล และเพิ่มลงไปในฟังก์ชันต้นทุนในรูปของค่าลงโทษ (Penalty Term) หากระยะทางระหว่างกลุ่มข้อมูลมีค่ามาก ค่าลงโทษก็จะยิ่งมากตามทำให้เวลาตัดสินเลือกกลุ่มข้อมูลให้กับตัวอย่างข้อมูลก็จะมีโอกาสน้อยที่จะถูกจัดให้อยู่ในพื้นที่ทับซ้อนของกลุ่มที่มีค่าค่าลงโทษสูง

งานวิจัยนี้ผู้วิจัยได้เลือกพัฒนาขั้นตอนวิธี OKM ที่เป็นขั้นตอนวิธีที่มีประสิทธิภาพสูง และมีความยืดหยุ่นในการนำไปพัฒนาต่อ ซึ่งสามารถเห็นได้จากมีงานวิจัยที่พัฒนาต่อยอดออกมาหลายงาน (เช่น R-OKM, KOKM และ CWOKM) โดยจะทำการต่อเติมขั้นตอนวิธีเพื่อให้สามารถแก้ไขปัญหาที่มีอยู่ตามแนวทางที่กล่าวมาข้างต้นและให้ผลการจัดกลุ่มที่มีความถูกต้องแม่นยำมากยิ่งขึ้น

1.2 วัตถุประสงค์ของงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาการจัดกลุ่มทับซ้อนแบบแบ่งส่วนโดยการต่อเติมขั้นตอนวิธีที่มีอยู่เดิม (ขั้นตอนวิธี OKM) ให้มีประสิทธิภาพมากยิ่งขึ้นในแง่ของความถูกต้องแม่นยำในการจัดกลุ่มข้อมูล

1.3 ขอบเขตของงานวิจัย

1) พัฒนาประสิทธิภาพของการจัดกลุ่มแบบทับซ้อนโดยทำให้เซนทรอยด์ที่ได้มีความเหมาะสมในการจัดกลุ่มข้อมูลมากยิ่งขึ้น
2) องค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาใช้ช่วยพัฒนาความถูกต้องแม่นยำในการจัดกลุ่มให้เพิ่มมากยิ่งขึ้น
3) ประสิทธิภาพของงานวิจัยจะถูกเปรียบเทียบกับขั้นตอนวิธีเหล่านี้ ขั้นตอนวิธี OKM ดั้งเดิม ขั้นตอนวิธีที่พัฒนามาจาก OKM ได้แก่ CWOKM และ ROKM ขั้นตอนวิธี FCM ที่มีปัญหาการเริ่มต้นเซนทรอยด์น้อย ขั้นตอนวิธี MOC [18] ที่ไม่ใช่การจัดกลุ่มทับซ้อนแบบแบ่งส่วน และขั้นตอนวิธีการจัดกลุ่มทับซ้อนแบบแบ่งส่วนล่าสุด NEO-K-Means โดยจะเปรียบเทียบกันในเชิงความถูกต้องแม่นยำในการจัดกลุ่มในหน่วยของ F1 เพื่อชี้วัดว่าสามารถแก้ไขปัญหาที่เกิดขึ้นและนำองค์ความรู้ความสัมพันธ์ระหว่างกลุ่มข้อมูลมาประยุกต์ใช้ประโยชน์ได้จริงหรือไม่

4) งานวิจัยนี้จะทำการทดลองและวัดประสิทธิภาพจากชุดข้อมูลหลายหมวดหมู่ 20 ชุด ข้อมูลจากเว็บไซต์ http://mulan.sourceforge.net/ [14] และเว็บไซต์ http://www.uco.es/ (LAIM) [35] โดยประกอบด้วยชุดข้อมูลดังต่อไปนี้

ข้อมูลจาก Mulan	*ชุดข้อมูลจาก LAIM*
Emotions	Birds
Yeast	Plant
Scene	Human
Mediamill	Flags
RCV1 (Set1)	Yahoo (Art)
RCV1 (Set2)	Yahoo (Health)
RCV1 (Set3)	Yahoo (Business)
RCV1 (Set4)	EUR-Lex (Directory Code)
RCV1 (Set5)	EUR-Lex (Subject Matters)
CAL500	EUR-Lex (Eurovoc Descriptor)

ชุดข้อมูลเหล่านี้คือชุดข้อมูลที่ได้จากการรวบรวมข้อมูลที่มีอยู่จริงและเคยถูกใช้เพื่อทดสอบประสิทธิภาพในงานวิจัยอื่นมาก่อนในงานวิจัยนี้ข้อมูลในแต่ละชุดจะถูกระบุสถิติของข้อมูลดังต่อไปนี้ จำนวนข้อมูล จำนวนฟีเจอร์ (Feature) จำนวนหมวดหมู่ ค่าคาร์ดินอลิตี (Cardinality) ของข้อมูล และค่าความหนาแน่น (Density) ของข้อมูล

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ได้ขั้นตอนวิธีใหม่ซึ่งแก้ไขปัญหาการจัดกลุ่มทับซ้อนที่เซนทรอยด์ติดอยู่กับค่าต่ำสุดสัมพัทธ์ และสามารถนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระกว่างกลุ่มข้อมูลมาประยุกต์ใช้เพื่อพัฒนาประสิทธิภาพด้านความถูกต้องแม่นยำในการจัดกลุ่มข้อมูลได้

1.5 ขั้นตอนและวิธีดำเนินการวิจัย

1) ศึกษาขั้นตอนวิธีการจัดกลุ่มข้อมูลพื้นฐาน

2) ศึกษางานวิจัยเกี่ยวกับการจัดกลุ่มข้อมูลแบบทับซ้อนโดยทั่วไปว่ามีกี่ประเภทและมีข้อดีและข้อเสียแตกต่างกันอย่างไร
3) ศึกษางานวิจัยเกี่ยวกับการจัดกลุ่มแบบทับซ้อนแบบแบ่งส่วน โดยให้ความสนใจกับขั้นตอนวิธี OKM
4) ตรวจสอบปัญหาของการจัดกลุ่มทับซ้อนแบบแบ่งส่วนโดยการทำการทดลองขั้นตอนวิธีที่มีกับชุดข้อมูลประเภทหลายหมวดหมู่จำนวน 20 ชุดข้อมูล
5) วิเคราะห์สาเหตุของปัญหาและหาแนวทางในการพัฒนาขั้นตอนวิธี
6) ออกแบบขั้นตอนวิธีเพื่อประยุกต์ใช้ในการพัฒนาและแก้ไขปัญหาโดยมีขั้นตอนวิธี OKM เป็นขั้นตอนวิธีพื้นฐาน
7) ทำการทดลองการกำหนดกลุ่มข้อมูลด้วยขั้นตอนวิธีใหม่และวัดผลประสิทธิภาพในเชิงความถูกต้องแม่นยำ
8) ทำการทดลองเปรียบเทียบขั้นตอนวิธีใหม่กับขั้นตอนวิธีที่มีอยู่เดิมดังต่อไปนี้ OKM, R-OKM, CWOKM, FCM, NEO-K-Means
9) วิเคราะห์ผลการทดลองโดยการเปรียบเทียบกับขั้นตอนวิธีอื่นที่ได้ทำการทดลองไว้
10) สรุปผลการทดลองทั้งหมดโดยให้เหตุผลทั้งกรณีที่ขั้นตอนวิธีในงานวิจัยนี้มีประสิทธิภาพสูงกว่าและต่ำกว่างานวิจัยอื่นในแง่ความถูกต้องแม่นยำในการจัดกลุ่ม

ตารางที่ 1 แผนภูมิแสดงระยะเวลาและกระบวนการดำเนินงาน

การดำเนินงาน	เม.ย.59	พ.ค.59	มิ.ย.59	ก.ค.59	ส.ค.59	ก.ย.59	ต.ค.59	ต.ค.60	พ.ย.60
1.ศึกษาขั้นตอนวิธีพื้นฐาน									
2.ศึกษางานวิจัยที่เกี่ยวข้อง									
3.วิเคราะห์ปัญหา									
4.ทำการทดลองเบื้องต้น									
5.ตีพิมพ์บทความวิชาการ									
6.สอบโครงร่างวิทยานิพนธ์									
7.พัฒนาขั้นตอนวิธีเพิ่มเติม									
8.ทำการทดลองเพิ่มเติม									
10.สรุปผลจัดทำวิทยานิพนธ์									
11.สอบวิทยานิพนธ์									

1.6 ผลงานวิจัยที่ได้ตีพิมพ์

- “Enhance Accuracy of Partition-based Overlapping Clustering by Exploiting Benefit of Distances between Clusters” โดย ธนวัต ลิมังกูร และพีรพล เวทีกูล ในงานประชุมวิชาการ “Eighth International Conference on Knowledge and System

Engineering – KSE 2016” ที่จัดขึ้น ณ เมืองฮานอย ประเทศเวียดนาม ระหว่างวันที่ 6 ถึง 8 ตุลาคม พ.ศ. 2559

- “Partition-based Overlapping Clustering using Cluster's Parameters and Relations” โดย ธนวัต ลิมังกูร และพีรพล เวทีกูล ในงานประชุมวิชาการ “Ninth International Conference on Knowledge and Smart Technology – KST 2017” ที่จัดขึ้น ณ จังหวัดชลบุรี ระหว่างวันที่ 1 ถึง 4 กุมภาพันธ์ พ.ศ. 2560

1.7 โครงสร้างของเนื้อหาในวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 5 บทคือ บทที่ 1 บทนำกล่าวถึงที่มาและความสำคัญของปัญหา วัตถุประสงค์ของงานวิจัย ขอบเขตของงานวิจัย ประโยชน์ที่คาดว่าจะได้รับ ขั้นตอนและวิธีดำเนินการวิจัย ผลงานวิจัยที่ได้ตีพิมพ์ และโครงสร้างของวิทยานิพนธ์ บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ซึ่งทฤษฎีที่เกี่ยวข้องจะเป็นทฤษฎีในการจัดกลุ่มข้อมูลที่ใช้เป็นพื้นฐานในงานวิจัย งานวิจัยที่เกี่ยวข้องจะเป็นงานวิจัยเกี่ยวกับการจัดกลุ่มทับซ้อนแบบแบ่งส่วน บทที่ 3 นำเสนอแนวคิดของงานวิจัยและนำเสนอกระบวนการในการพัฒนาขั้นตอนวิธีให้มีประสิทธิภาพในเชิงความถูกต้องแม่นยำให้เพิ่มมากกว่าเดิม บทที่ 4 เป็นบทที่แสดงผลการทดลองและการประเมินผลขั้นตอนวิธีใหม่กับวิธีการที่ใช้เป็นบรรทัดฐาน บทที่ 5 เป็นบทสรุปของงานวิจัย วิเคราะห์จุดเด่นและข้อจำกัดของขั้นตอนในงานวิจัยนี้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

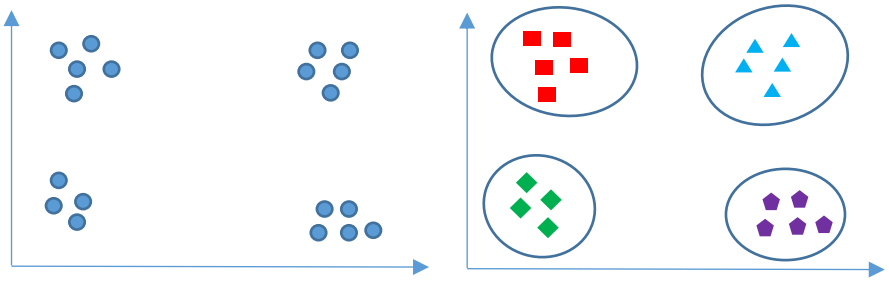
2.1 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้มีองค์ประกอบหลักคือการจัดกลุ่มข้อมูลซึ่งเป็นพื้นฐานในงานวิจัยการจัดกลุ่มแบบทับซ้อน ในการจัดกลุ่มข้อมูลจำเป็นต้องมีมาตรวัดระยะทาง (Distance Metric) เพื่อใช้ในการชี้วัดความใกล้เคียงกันของข้อมูล และการวัดผลประสิทธิภาพในการจัดกลุ่มซึ่งในงานวิจัยนี้ให้ความสนใจที่ความถูกต้องแม่นยำในหน่วย F1

2.1.1 การจัดกลุ่มข้อมูล (Clustering)

การจัดกลุ่มข้อมูลเป็นการเรียนรู้ของเครื่อง (Machine Learning) เพื่อรวมกลุ่มข้อมูลที่คาดว่าเป็นประเภทเดียวกันไว้ด้วยกัน และยังหมายรวมไปถึงการวิเคราะห์กลุ่มข้อมูล (Cluster Analysis) ในการจัดกลุ่มข้อมูลข้อมูลที่มีความคล้ายคลึงกันจะถูกรวมเข้าไว้ด้วยกัน ส่วนข้อมูลที่แตกต่างกันจะถูกจัดไว้อยู่คนละกลุ่ม ซึ่งมาตรวัดที่เป็นที่นิยมในการวัดความคล้ายคลึงหรือความแตกต่างกัน ได้แก่ ระยะทางแบบยุคลิด (Euclidean Distance) สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation) และความคล้ายคลึงแบบโคซายน์ (Cosine Similarity)

ตัวอย่างของการจัดกลุ่มข้อมูลตามรูปที่ 1 (ก) จะสังเกตได้ว่าข้อมูลมีการเกาะกลุ่มกัน 4 กลุ่มอย่างเห็นได้ชัด ในการจัดกลุ่มตัวอย่างข้อมูลที่อยู่ใกล้กันจะถูกกำหนดให้อยู่ในกลุ่มเดียวกัน ด้วยเหตุนี้หลังจากทำการจัดกลุ่มแล้วจะได้ผลลัพธ์ออกมาตามรูปที่ 1 (ข)



รูปที่ 1 (ก) ข้อมูลก่อนทำการจัดกลุ่ม *รูปที่ 1 (ข) ข้อมูลหลังทำการจัด*

รูปที่ 1 แสดงตัวอย่างของการจัดกลุ่มข้อมูลโดยที่ รูปที่ 1 (ก) แสดงข้อมูลก่อนทำการจัดกลุ่ม และรูปที่ 1 (ข) แสดงข้อมูลหลังทำการจัดกลุ่ม

เทคนิคการจัดกลุ่มนั้นมีหลายประเภทในที่นี้จะขอยกตัวอย่างการจัดกลุ่มที่เป็นที่นิยมบางส่วนพอสังเขป

การจัดกลุ่มข้อมูลแบบแบ่งส่วน (Partition-based Clustering)

การจัดกลุ่มข้อมูลแบบแบ่งส่วนเป็นการจัดกลุ่มข้อมูลโดยอิงระยะทางจากเซนทรอยด์ของกลุ่มข้อมูล โดยข้อมูลจะอยู่ในกลุ่มใดนั้นจะพิจารณาจากข้อมูลนั้นอยู่ใกล้กับเซนทรอยด์ของกลุ่มใดมากที่สุด การจัดกลุ่มในประเภทนี้ที่เป็นที่รู้จักกันดีคือการจัดกลุ่มแบบ K-Means โดยทั่วไปมาตรวัดความคล้ายคลึงที่ใช้คือระยะทางแบบยุคลิด ในกระบวนการของขั้นตอนวิธี K-Means ข้อมูลแต่ละข้อมูลจะถูกจัดให้อยู่ในกลุ่มที่มีระยะทางใกล้กับเซนทรอยด์มากที่สุด ดังสมการ (1) ดังนี้

$$\sum_{i=1}^{n}\sum_{j=1}^{k} min \left|\left| x_i - m_j \right|\right|^2 \quad (1)$$

โดยที่ n คือ จำนวนข้อมูล

k คือ จำนวนกลุ่มข้อมูล

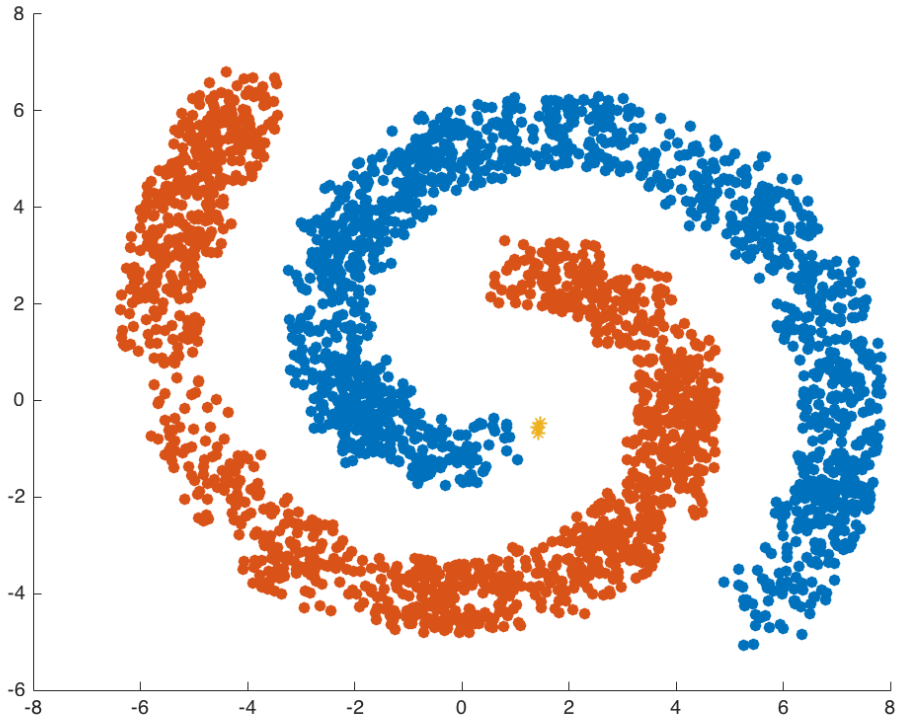
x คือ ตัวอย่างข้อมูล

m คือ เซนทรอยด์ของกลุ่มข้อมูล

โดยกระบวนการจะวนซ้ำไปเรื่อย ๆ เพื่อลดค่าของฟังก์ชันต้นทุนจนค่าที่ได้ไม่มีการเปลี่ยนแปลงหรือต่ำกว่าค่าขีดแบ่ง (Threshold) ที่กำหนดไว้ หรืออาจจะกำหนดจำนวนรอบเพื่อหยุดกระบวนการก็ได้แล้วแต่ความเหมาะสม

การจัดกลุ่มตามความหนาแน่น (Density-based Clustering)

ในขณะที่การจัดกลุ่มตามระยะทางแบ่งกลุ่มข้อมูลโดยอิงตามตำแหน่งของข้อมูล มีการจัดกลุ่มอีกประเภทหนึ่งที่ไม่ได้อิงตำแหน่งแต่อิงความหนาแน่นของข้อมูลแทน การจัดกลุ่มตามความหนาแน่นเป็นเทคนิคการรวมกลุ่มข้อมูล โดยรวมข้อมูลที่อยู่ภายในรัศมีที่กำหนดเข้าไว้ด้วยกันและจะแผ่ขยายไปเรื่อย ๆ หากความหนาแน่นยังมีปริมาณอยู่ในช่วงที่กำหนดไว้ ข้อดีของการจัดกลุ่มแบบนี้คือขจัดปัญหาข้อมูลรบกวน และสามารถจัดการกับข้อมูลที่ไร้รูปแบบ (Arbitrary Data) ได้ดี ตัวอย่างการจัดกลุ่มในหมวดหมู่นี้ที่เป็นเป็นที่นิยมคือขั้นตอนวิธี DBSCAN



รูปที่ 2 แสดงรูปแบบข้อมูลที่เหมาะกับการจัดกลุ่มตามความหนาแน่น
(อ้างอิงจาก: https://www.mathworks.com/matlabcentral/fileexchange/53842-dbscan)

การจัดกลุ่มตามลำดับชั้น (Hierarchical Clustering)

การจัดกลุ่มในรูปแบบนี้จะรวมกันเป็นระดับชั้นไปถูกแบ่งออกเป็น 2 ประเภทดังนี้

1) *วิธี Agglomerative*

เริ่มแรกแต่ละวัตถุจะถูกนับเป็นกลุ่มข้อมูล 1 กลุ่ม จากนั้นการรวมกลุ่มจะเกิดขึ้นโดยรวมกลุ่มข้อมูลที่คล้ายคลึงกันเข้าไว้ด้วยกัน จนสุดท้ายเหลือกลุ่มเดียว หรือตรงตามเงื่อนไขที่กำหนดไว้ วิธีนี้เป็นการทำแบบจากล่างขึ้นบน (Bottom-up)

2) *วิธี Divisive*

เริ่มจากทุกตัวอย่างข้อมูลอยู่ในกลุ่มเดียวกัน แล้วทำการแตกให้เป็นกลุ่มย่อยทำไปเรื่อย ๆ จนกว่าแต่ละตัวอย่างข้อมูลในกลุ่มเดียวกันมีความคล้ายคลึงกันมากพอตามเงื่อนไขที่กำหนด วิธีนี้เป็นวิธีการทำจากบนลงล่าง (Top-down)

2.1.2 มาตรวัดระยะทาง

มาตรวัดระยะทางเป็นตัวชี้วัดค่าความใกล้เคียง (Similarity) ระหว่างข้อมูล ยิ่งระยะทางสั้นค่าความใกล้เคียงก็จะสูง มาตรวัดระยะทางเป็นเครื่องมือที่จำเป็นอย่างมากในการจัดกลุ่ม เพราะขั้นตอนวิธีส่วนใหญ่จะจัดกลุ่มโดยพิจารณาจากค่าความใกล้เคียงข้อมูลที่มีค่าความใกล้เคียงสูงจะถูกจัดให้อยู่ในกลุ่มเดียวกัน

มาตรวัดระยะทางแบบแมนฮัตตัน (Manhattan Distance)

มาตรวัดแบบแมนฮัตตันถูกระบุในรูปฟังก์ชัน $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ โดยเป็นการหาระยะทางจากเวกเตอร์ข้อมูล $X = (x_1, x_2, \dots x_n)$ และเวกเตอร์ข้อมูล $Y = (y_1, y_2, \dots y_n)$ โดยค่าที่ได้เป็นค่านอร์ม (Norm) ระดับที่ 1 สามารถแสดงตามสมการ (2)

$$d(X,Y) = |x_1 - y_1| + |x_2 - y_2| \dots + |x_n - y_n| \tag{2}$$

มาตรวัดระยะทางแบบยุคลิด (Euclidean Distance)

มาตรวัดแบบยุคลิดถูกระบุในรูปฟังก์ชัน $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ โดยเป็นการหาระยะทางจากเวกเตอร์ข้อมูล $X = (x_1, x_2, \dots x_n)$ และเวกเตอร์ข้อมูล $Y = (y_1, y_2, \dots y_n)$ โดยค่าที่ได้เป็นค่านอร์ม (Norm) ระดับที่ 2 สามารถแสดงตามสมการ (3)

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 \dots + (x_n - y_n)^2} \tag{3}$$

มาตรวัดแบบยุคลิดถูกใช้ในการหาความใกล้เคียงระหว่างข้อมูลในงานวิจัยนี้

มาตรวัดแบบความใกล้เคียงโคไซน์ (Cosine Similarity)

มาตรวัดแบบยุคลิดถูกระบุในรูปฟังก์ชัน $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ โดยเป็นการหาระยะทางจากเวกเตอร์ข้อมูล $X = (x_1, x_2, \dots x_n)$ และเวกเตอร์ข้อมูล $Y = (y_1, y_2, \dots y_n)$ โดยค่าที่ได้ถูกแสดงในรูปของผลคูณจุด (Dot Product) และค่าแมกนิจูด (Magnitude) ของข้อมูล สามารถแสดงได้ตามสมการที่ (4)

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4}$$

มาตรวัดแบบความสัมพันธ์เพียร์สัน (Pearson correlation)

มาตรวัดแบบความสัมพันธ์เพียร์สันเป็นการวัดความใกล้เคียงในเชิงสถิติประชากรของข้อมูล โดยการประเมินค่าความสัมพันธ์เพียร์สันของข้อมูล X และ Y นั้นสามารถแสดงได้ดังนี้

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{5}$$

โดยที่

cov คือ ค่าความแปรปรวนร่วมของข้อมูล

σ คือ ค่าเบี่ยงเบนมาตรฐานของข้อมูล

2.1.3 การประเมินผล

ในงานวิจัยนี้ใช้ค่า F_1 ในการวัดผลประสิทธิภาพของการจัดกลุ่ม ซึ่งการคำนวณค่า F_1 นั้นจะมีการคำนวณโดยอิงตามค่าพรีซิชัน (Precision) และค่ารีคอล (Recall) โดยวิธีการคำนวณนั้นจะเป็นดังสมการที่ (6)

$$F_1 = \frac{2 * Precison * Recall}{Precision + Recall} \tag{6}$$

$$Precision = \frac{|C \cap S|}{|C|}, Recall = \frac{|C \cap S|}{|S|}$$

โดยที่ S คือกลุ่มข้อมูลจริง

C คือกลุ่มข้อมูลที่ได้จากการจัดกลุ่ม

2.2 งานวิจัยที่เกี่ยวข้อง

สำหรับงานวิจัยที่เกี่ยวข้องในงานวิจัยนี้ให้ความสำคัญที่การจัดกลุ่มทับซ้อนแบบต่าง ๆ ทั้งแบบแบ่งส่วนและแบบอื่น ๆ เพื่อเปรียบเทียบข้อดีและข้อด้อยและความเหมาะสมในการใช้งานของแต่ละขั้นตอนวิธี

2.2.1 การจัดกลุ่มทับซ้อนที่พัฒนามาจากการจัดกลุ่มแบบแบ่งส่วน (Partition-based Overlapping Clustering)

ในกระบวนการจัดกลุ่มทับซ้อนทั้งหมดนั้น วิธีการแบ่งกลุ่มที่เป็นที่นิยมและให้ประสิทธิภาพที่ดีเมื่อใช้กับข้อมูลส่วนใหญ่ ได้แก่ วิธีการแบ่งกลุ่มแบบแบ่งส่วน ซึ่งพื้นฐานของวิธีในหมวดหมู่นี้ ได้แก่ ขั้นตอนวิธี K-Means และ K-Medoids วิธีการในหมวดหมู่นี้สามารถแบ่งออกเป็น 2 ประเภทดังนี้

วิธีแบบที่ความเป็นสมาชิกภายในกลุ่มไม่แน่นอน (Uncertain-membership Method)

วิธีการในหมวดหมู่นี้มี 2 ประเภท ประเภทแรกขยายผลของการจัดกลุ่มแบบเดิมมาสู่กลุ่มของข้อมูลที่ทับซ้อน โดยปกติแล้วเป็นการขยายมาจากการจัดกลุ่มแบบ Fuzzy-C-Means [18] ได้แก่ [4] และ Probabilistic-C-Means [19, 20] อีกประเภทหนึ่งเป็นการใช้กระบวนการใหม่โดยการกำหนดเงื่อนไขและค่าขีดแบ่งเพื่อทำการจัดกลุ่ม โดยผลที่ได้จะเป็นความน่าจะเป็นของโอกาสในการทับซ้อนของกลุ่มข้อมูล ตัวอย่างที่เด่นชัดของกลุ่มนี้ ได้แก่ Evidential-C-Means ซึ่งถูกนำเสนอโดย

Masson และ Denouex [5] และ Belief-C-Means ที่ถูกนำเสนอโดย Liu et al. [7] การจัดกลุ่มในหมวดหมู่นี้ทุกวิธีการต้องการการประมวลผลภายหลังเพื่อกำหนดขอบเขตกลุ่มให้ชัดเจน

วิธีแบบที่ความเป็นสมาชิกภายในกลุ่มแน่นอน (Certain-membership Method)

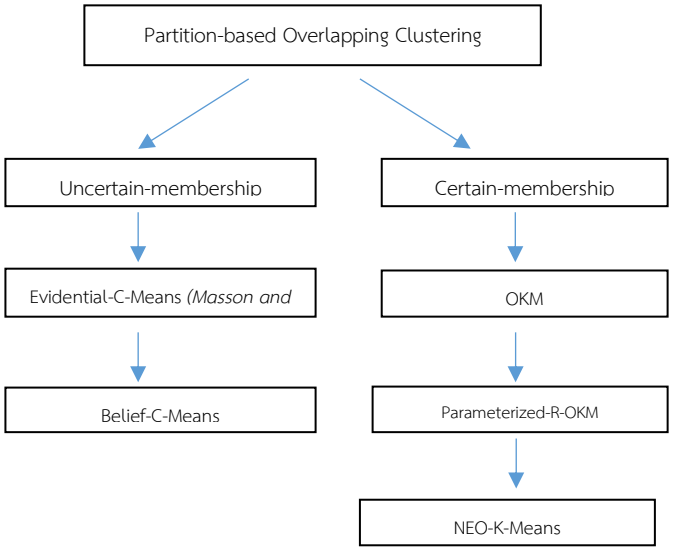
สำหรับวิธีนี้ความเป็นสมาชิกภายในกลุ่มแน่นอน (สมาชิกถูกระบุกลุ่มชัดเจนไม่ได้อยู่ในรูปแบบความน่าจะเป็น) ผลที่ได้จากการจัดกลุ่มจึงไม่ต้องการการประมวลผลภายหลัง การจัดกลุ่มในหมวดหมู่นี้ยังแยกออกเป็น 2 ประเภทตามรูปแบบการประมวลผล

1) แบบ Additive

วิธีนี้จะใช้ขั้นตอนวิธีละโมบ (Greedy Algorithm) ในการลดค่าฟังก์วัตถุประสงค์ในแต่ละรอบโดยจะพยายามหารูปแบบการจัดกลุ่มที่ดีที่สุดเพื่อให้ได้มาซึ่งค่าน้อยที่สุดโดยรวม (Global Minimum) วิธีนี้ใช้ความซับซ้อนในประมวลผลสูงถึงระดับ 2^n หรือถึง 2^{nk} อย่างไรก็ตามในปี 2012 D. Depril [6] ได้นำเสนอวิธีการที่ลดมิติในการคำนวณลงโดยใช้ชื่อว่า Low-Dimensional Additive Clustering ซึ่งลดเวลาประมวลผลลง

2) แบบ Geometrical

วิธีนี้จะคำนวณฟังก์ชันวัตถุประสงค์ตามผลรวมของระยะทางระหว่างแต่ละตัวข้อมูลและเซนทรอยด์ของกลุ่มข้อมูลที่ใกล้ที่สุด วิธีการในหมวดหมู่นี้ใช้เวลาในการประมวลผลน้อยและมีความถูกต้องแม่นยำสูง วิธีการแรกที่ได้รับการพัฒนามาคือ OKM ที่นำเสนอโดย G. Cleuziou [21, 22] ต่อมาวิธีการนี้ได้รับการพัฒนาต่อเป็น R-OKM ที่มีความถูกต้องแม่นยำมากยิ่งขึ้น ซึ่งพัฒนาโดย C. E. B. N'Cir และ G. Cleuziou [1] ปัจจุบันวิธีการในหมวดหมู่นี้ที่มีความถูกต้องแม่นยำสูงที่สุดคือ NEO-K-Means ซึ่งถูกนำเสนอโดย J. J. Whang วิวัฒนาการของการแบ่งกลุ่มทับซ้อนแสดงได้ตามรูปที่ 3



รูปที่ 3 แสดงแผนภูมิพัฒนาการของการแบ่งกลุ่มทับซ้อนแบบแบ่งส่วน [23]

2.2.2 การจัดกลุ่มทับซ้อนที่พัฒนาต่อมาจากการจัดกลุ่มตามลำดับชั้น (Hierarchical Overlapping Clustering)

สำหรับในหมวดหมู่นี้ใช้วิธีรวมกลุ่มข้อมูลที่ใกล้เคียงกันในแต่ละลำดับชั้นให้กลายเป็นพื้นที่ของกลุ่มข้อมูลที่ทับซ้อนกัน ตัวอย่างของวิธีการนี้ที่รู้จักโดยทั่วไปคือการจัดกลุ่มแบบ Pyramids Weak-hierarchies ซึ่งนำเสนอโดย P. Bertrand [13] ในวิธีการนี้มีข้อดีคือ เพราะการรวมกลุ่มข้อมูลถูกแบ่งออกเป็นลำดับชั้นย่อย ๆ จึงสามารถเห็นภาพรวมของข้อมูลได้ชัดเจน แต่มีข้อจำกัดคือการรวมกลุ่มในแต่ละลำดับชั้นมีรูปแบบไม่มากพอที่จะให้ผลลัพธ์ที่ดีที่สุด

2.2.3 การจัดกลุ่มทับซ้อนที่พัฒนามาจากการจัดกลุ่มด้วยวิธีแบบกราฟ (Graph-based Overlapping Clustering)

การใช้วิธีแบบกราฟ [12] มักจะใช้ในการค้นหาชุมชน (Community) ในเครือข่ายซับซ้อน (Complex Network) ในการจัดกลุ่มแบบนี้เครือข่ายหนึ่งจะถูกแทนด้วยกราฟระบุทิศทาง (Directed Graph) หรือกราฟไม่ระบุทิศทาง (Undirected Graph) ขึ้นอยู่กับความจำเพาะของปัญหา ตัวอย่างข้อมูลจะถูกแทนด้วยจุดยอด (Vertex) และความสัมพันธ์ระหว่างตัวอย่างข้อมูลจะถูกแทนด้วยเส้นเชื่อม (Edge) วิธีการแบบกราฟทั้งหมดใช้ขั้นตอนวิธีละโมบ (Greedy Algorithm) เพื่อรวมกราฟที่เหมือนกันเข้าไว้ด้วยกัน ความแตกต่างของแต่ละวิธีนั้นอยู่ที่เงื่อนไขที่ใช้ในการเรียง และการเลือกกราฟย่อย (Sub-graph) วิธีนี้มีข้อจำกัดคือข้อมูลที่นำมาประยุกต์ใช้ต้องอยู่ในรูปแบบกราฟเท่านั้น และยังใช้เวลาในการประมวลผลนานถึงระดับ Exponential อย่างไรก็ตามในปี 2013 A. Pérez-Suárez et al. [8-11] ได้นำเสนอวิธีการจัดกลุ่มแบบ OClustR [24] ที่ลดความซับซับซ้อนของการประมวลผลเหลือเพียง O (N^2) โดยที่ความถูกต้องแม่นยำยังไม่เปลี่ยนแปลงไปจากเดิมมากนัก

2.2.4 การจัดกลุ่มทับซ้อนโดยใช้ Generative Mixture Model

การจัดกลุ่มในหมวดหมู่นี้ [24] เป็นกรอบงาน (Framework) ที่เป็นส่วนขยายของอัลกอริทึมอีเอ็ม (EM Algorithm) [13, 25, 26] สมมติฐานของวิธีการในหมวดหมู่นี้คือแต่ละตัวอย่างข้อมูลเป็นผลมาจากการรวมกันระหว่างการกระจายตัวของข้อมูล ตัวอย่างที่เห็นได้ชัดคือ แบบจำลองผลบวก (Additive Model) [27] และแบบจำลองผลคูณ (Multiplicative Model) [13] เนื่องจากผลลัพธ์ของวิธีในหมวดหมู่นี้แปรตามแบบจำลองที่เลือกใช้ วิธีนี้การนี้จึงไม่ยืดหยุ่นนักเนื่องจากไม่สามารถมีค่าตัวแปรให้กำหนดน้อย

2.2.5 การจัดกลุ่มทับซ้อนที่พัฒนามาจากการจัดกลุ่มตามความหนาแน่น (Density-based Overlapping Clustering)

การจัดกลุ่มในหมวดหมู่นี้เหมาะสำหรับใช้จัดการกับกลุ่มข้อมูลที่มีรูปร่างไม่เป็นทรงกลม (Non-spherical Shape) หรือไม่มีรูปแบบ (Arbitrary) ได้ดี ในปี 2007 C. Ruiz [25, 26] ได้เสนอวิธีการจัดกลุ่มทับซ้อนโดยใช้อัลกอริทึม DBSCAN เป็นอัลกอริทึมพื้นฐาน และเพิ่มเงื่อนไขในการทับซ้อนของกลุ่มข้อมูลเข้าไปทำให้จากเดิมที่กลุ่มของข้อมูลแยกขาดกันโดยสิ้นเชิงสามารถมีส่วนที่ทับซ้อนกันได้ ในปี 2009 P. Viswanath [28] ได้เสนอวิธีการใหม่ที่มีแนวคิดในการจัดกลุ่มทับซ้อนตามความหนาแน่นเช่นกัน แต่ปรับปรุงอัลกอริทึมให้ใช้เวลาในการประมวลลดลงเพื่อให้สามารถรองรับกับข้อมูลขนาดใหญ่ได้ อย่างไรก็ตามวิธีนี้ให้ประสิทธิภาพในการจัดกลุ่มลดลงเล็กน้อย ในปี 2012 C.Braune [29] ได้เสนอวิธีการใหม่ชื่อว่า Proto-clustering ซึ่งมีการเพิ่มเกณฑ์กำหนดของจำนวนสมาชิกภายในกลุ่มเข้าไปทำให้การจัดกลุ่มมีประสิทธิภาพมากยิ่งขึ้น ข้อเสียของวิธีการในหมวดหมู่นี้จะเหมือนกับวิธีการในแบบของการจัดกลุ่มตามความหนาแน่นต้นแบบ คือจะประยุกต์ใช้ได้อย่างมีประสิทธิภาพได้กับชุดข้อมูลที่มีการเกาะกันเป็นกลุ่มเท่านั้น

2.2.6 การจัดกลุ่มทับซ้อนแบบอื่น ๆ

สำหรับการจัดกลุ่มทับซ้อนแบบอื่นนั้นมีแนวคิดวิธีแก้ปัญหาต่างออกไป เช่น การต่อยอดจากการจัดกลุ่มความสัมพันธ์ (Correlation Clustering) [30] และการทำ Topological Map [31] การจัดกลุ่มทับซ้อนในหมวดหมู่การจัดกลุ่มความสัมพันธ์เป็นการหาค่าที่เหมาะที่สุด (Optimization) ที่ขยายกรอบงานให้อนุญาตให้เกิดการทับซ้อนกันระหว่างกลุ่มข้อมูลได้ โดยการผ่อนปรนฟังก์ชันกำหนดความสัมพันธ์ (Correlation Function) ให้สามารถกำหนดกลุ่มของข้อมูลได้มากกว่าหนึ่งกลุ่มต่อหนึ่งตัวอย่างข้อมูล ส่วนการทำ Topological Map เป็นการต่อยอดจากการทำ Self-Organizing Map ซึ่งถูกนำเสนอโดย G.Cleuziou [23] สำหรับแนวคิดนี้ใช้วิธีการค้นหาบนตาราง (Grid Search) เพื่อหาเซตย่อยของเส้นประสาทซึ่งจะกลายมาเป็นกลุ่มข้อมูลในภายหลัง ข้อดีที่เด่นชัดของทั้งการจัดกลุ่มความสัมพันธ์และการทำ Topological Map คือสามารถรู้จำนวนตัวอย่างข้อมูลในพื้นที่ทับซ้อนที่ค่อนข้างแม่นยำ

2.3 ขั้นตอนวิธีที่มีส่วนใช้ในงานวิจัย

2.3.1 ขั้นตอนวิธี Overlapping K-Means (OKM)

ในงานวิจัยนี้เลือกขั้นตอนวิธี OKM [1] เป็นขั้นตอนวิธีพื้นฐานสำหรับการจัดกลุ่มแบบทับซ้อน เนื่องจากเป็นขั้นตอนวิธีที่สามารถประยุกต์ใช้ได้กับชุดข้อมูลส่วนใหญ่ได้อย่างมีประสิทธิภาพ

ขั้นตอนวิธี OKM เป็นตอนวิธีที่พัฒนามาจาก K-Means เพื่อให้สามารถรองรับข้อมูลประเภทหลายหมวดหมู่ได้สำหรับขั้นตอนวิธี OKM สามารถอธิบายได้ดังต่อไปนี้

เกณฑ์ของวัตถุประสงค์ (Objective Criterion)

จากชุดข้อมูล $X = \{x_i\}^n{}_{i=1}$ โดยที่ $X_i \in \mathbb{R}^P$ จุดมุ่งหมายของขั้นตอนวิธี OKM คือการหาการกำหนดกลุ่มข้อมูล k กลุ่มแทนด้วย $\{\pi_{c=1}^k\}$ โดยที่ทำให้ค่าฟังก์ชันต้นทุนต่อไปนี้มีค่าน้อยที่สุด

$$\mathcal{J}\left(\{\pi_{c=1}^k\}\right) = \sum_{x_i \in X} \left|\left|x_i - \phi(x_i)\right|\right|^2 \tag{7}$$

$\{\pi_{c=1}^k\}$ เป็นเซตของกลุ่มข้อมูล k กลุ่มที่ทับซ้อนกันได้ ข้อมูล x_i จึงมีกลุ่มข้อมูลได้อย่างน้อยหนึ่งกลุ่มหรือมากกว่า ส่วน $\phi(x_i)$ แสดงถึงอิมเมจ (image) ของ x_i โดยที่สามารถคำนวณกลุ่มข้อมูล m_c ที่ x_i อยู่ได้ดังนี้

$$\phi(x_i) = \frac{\sum_{A_i} m_c}{|A_i|} \tag{8}$$

โดยที่ A_i แสดงถึงการกำหนดกลุ่มข้อมูลของ $x_i : \{m_c \mid x_i \in \pi_c\}$

จะสังเกตได้ว่าสิ่งที่ OKM ต่างจาก K-Means คือฟังก์ชันต้นทุนถูกเปลี่ยนจากระยะทางจากข้อมูลไปยังจุดเซนทรอยด์เดี่ยว ๆ ของกลุ่มข้อมูล เป็นระยะทางจากข้อมูลไปยังอิมเมจของข้อมูลซึ่งเป็นเสมือนตัวแทนของหลาย ๆ กลุ่มข้อมูลที่ข้อมูลนั้นถูกกำหนดให้อยู่แทน

เพื่อที่จะลดค่าฟังก์ชันต้นทุน $\mathcal{J}$ ลง การจัดกลุ่มจะเป็นไปตามขั้นตอนดังนี้ เริ่มจากกลุ่มข้อมูลเซนทรอยด์ถูกสุ่มกำหนดขึ้น จากนั้นหาเซตของกลุ่มข้อมูลรอบแรก $\left\{\pi_c^{(0)}\right\}_{c=1}^k$ โดยการกำหนดข้อมูลแบบหลายกลุ่ม (Multi-assignment) ที่จะอธิบายในหัวข้อถัดไป จากนั้นกระบวนการจะทำการวนซ้ำ 2 ขั้นตอนต่อไปนี้จนกว่าจะถึงเกณฑ์กำหนด

1) คำนวณเซตของเซนทรอยด์ของกลุ่มข้อมูลใหม่ในรอบใหม่ $\left\{m_c^{\{t+1\}}\right\}_{c=1}^k$
2) ทำการกำหนดกลุ่มข้อมูลซึ่งนำไปสู่เซตของกลุ่มข้อมูลในรอบใหม่ $\{\pi^{(t+1)}\}_{c=1}^k$

เช่นเดียวกันกับขั้นตอนวิธี K-Means เงื่อนไขในการหยุดกระบวนการวนซ้ำสามารถกำหนดได้จากการกำหนดค่าขีดแบ่งการเปลี่ยนแปลงของฟังก์ชันต้นทุน หรือจำนวนรอบของการวนซ้ำ รหัสเทียมของขั้นตอนวิธี OKM สามารถแสดงได้ดังรูปที่ 4

OKM *(x, t_{max}, ϵ)*

Input : X: a set of data vector $\mathbb{R}^P$, t_{max}: optional number of iterations, ϵ optional threshold on the objective

Output : $\{\pi\}_{c=1}^{k}$: final coverage of the data points

1. Draw randomly k initial cluster centroids $\{m_c^0\}_{c=1}^{k}$ in X.
2. For each $x_i \epsilon X$ compute the assignments $A_i^{(0)} = ASSIGN\ (x_i, \{m_c^{(0)}\}_{c=1}^{k}\ )$ following the multi-assignment in the next section and derive the initial coverage $\{\pi_c^{(0)}\}_{c=1}^{k}$.
3. Set $t = 0$.
4. For each cluster $\pi_c^{(t)}$ successively, compute the new centroid $m_c^{(t+1)}$.
5. For each $x_i \epsilon X$ compute the assignments $A_i^{(t+1)} = ASSIGN\ (x_i, \{m_c^{(t+1)}\}_{c=1}^{k}, A_i^{(t)})$ and derive the new coverage $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.
6. If not converged, $t_{max} > t$ or $\mathcal{J}(\{\pi^{(t)}\}) - \mathcal{J}(\{\pi^{(t+1)}\}) > \epsilon$, set $t = t + 1$ and go to step 4, otherwise output final clusters $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.

รูปที่ 4 รหัสเทียม (Pseudo-code) ของขั้นตอนวิธี OKM [1]

ในการกำหนดกลุ่มข้อมูลลงในกลุ่มข้อมูล k กลุ่ม นั้นมีความเป็นไปได้ถึง 2^k รูปแบบ และยังต้องกำหนดเป็นจำนวน n ครั้งจากแต่ละข้อมูล $x_i \in X$ ซึ่งไม่อาจทำได้จริงในทางปฏิบัติ ดังนั้นขั้นตอนวิธี OKM จึงจำเป็นต้องอาศัยวิทยาการศึกษาสำนึก (Heuristics) บางอย่างในการกำหนดกลุ่มข้อมูล โดยข้อมูล x_i จะอยู่ในกลุ่มข้อมูลใดบ้างนั้นจะพิจารณาจากเซนทรอยด์ที่ใกล้กับข้อมูลมากที่สุดก่อนจากนั้นจึงเพิ่มกลุ่มข้อมูลที่ใกล้เป็นลำดับถัดมาเข้ามา โดยที่การเพิ่มกลุ่มข้อมูลใหม่เข้ามานั้นต้องทำให้ค่าฟังก์ชันต้นทุนลดลงด้วย (ค่าระยะทางระหว่าง x_i และ $\phi(x_i)$ ลดลง) วิทยาการศึกษาสำนึกในการกำหนดกลุ่มข้อมูลแบบหลายกลุ่มขั้นตอนวิธี OKM สามารถแสดงได้ดังรูปที่ 5

ASSIGN $(x_i, \{m_c\}_{c=1}^{k}, A_i^{old})$

Input : $x_i \in X$, $\{m_1, m_2, \dots, m_k\}$: set of centroids and A_i^{old}: previous assignments

Output : $A_i \subset \{m_1, m_2, \dots . m_k\}$ a subset of clusters defining a multi-assignment of x_i

1. Set $A_i = \{m^*\}$ such that

$$m^* = \underset{\{m_c\}_{c=1}^{k}}{argmin} ||x_i - m_c||^2$$

and set image $\phi(x_i) = m^*$.

2. Find the following nearest centroid

$$m' = \underset{\{m_c\}_{c=1}^{k} \setminus A_i}{argmin} ||x_i - m_c||^2$$

and compute $\phi'(x_i)$ with $A_i \cup \{m'\}$.

3. If $||x_i - \phi'(x_i)||^2 < ||x_i - \phi(x_i)||^2$ set $A_i \leftarrow \{m'\}$, set $\phi(x_i) = \phi'(x_i)$ and go to step 2;

Otherwise

If $||x_i - \phi'(x_i)||^2 < ||x_i - \phi(x_i)||^2$ output A_i, else output A_i^{old}

รูปที่ 5 กระบวนการกำหนดข้อมูลแบบหลายกลุ่ม (Multi-assignment procedure) [1]

2.3.2 ขั้นตอนวิธี K-Harmonic-Means

ขั้นตอนวิธี K-Means เป็นขั้นตอนวิธีที่มีประสิทธิภาพและมีกระบวนการวิธีที่ไม่ซับซ้อน อย่างไรก็ตามขั้นตอนวิธี K-Means มีข้อจำกัดในเรื่องของการเริ่มต้นเซนทรอยด์ของกลุ่มข้อมูลมีผลกระทบอย่างมากต่อประสิทธิภาพในการจัดกลุ่ม เนื่องจากการเริ่มต้นเซนทรอยด์มาจากการสุ่มมีโอกาสทำให้เซนทรอยด์อยู่ในตำแหน่งที่ไม่เหมาะสมและเกิดผลต่อเนื่องทำให้ประสิทธิภาพของการจัดกลุ่มออกมาต่ำตามมา ขั้นตอนวิธี K-Harmonic-Means ถูกพัฒนาขึ้นมาเพื่อแก้ปัญหานี้ ในงานวิจัยนี้ K-Harmonic-Means เป็นอีกหนึ่งขั้นตอนวิธีพื้นฐานที่ถูกนำมาใช้

ขั้นตอนวิธี K-Harmonic-Means ต่างจากขั้นตอนวิธี K-Means ดั้งเดิมที่ฟังก์ชันต้นทุนถูกเปลี่ยนจากผลรวมของระยะทางระหว่างตัวอย่างข้อมูลไปยังกลุ่มข้อมูลที่ใกล้ที่สุด ให้กลายเป็นค่าเฉลี่ยฮาร์มอนิคของระยะทางระหว่างตัวอย่างข้อมูลกับกลุ่มข้อมูลแทน ซึ่งสามารถแสดงได้ตามสมการ (9)

$$\sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\left|\left| x_i - m_j \right|\right|^2}} \tag{9}$$

โดยที่ n คือ จำนวนข้อมูล

k คือ จำนวนกลุ่มข้อมูล

x คือ ตัวอย่างข้อมูล

m คือ เซนทรอยด์ของกลุ่มข้อมูล

K-Harmonic-Means เป็นขั้นตอนวิธีที่ผลิตเซนทรอยด์ให้อยู่ในตำแหน่งที่เหมาะสมมากกว่า K-Means เนื่องจาก K-Means มีการคำนวณกลุ่มของมูลโดยมีพื้นฐานจากค่าเฉลี่ยเลขคณิต แต่ K-Harmonic Means มีพื้นฐานจากค่าเฉลี่ยฮาร์มอนิค ซึ่งโดยทั่วไปแล้วการคำนวณโดยค่าเฉลี่ยเลขคณิตจะมีความคงทนต่อค่าผิดปกติน้อยกว่า ยกตัวอย่างเช่น มีจำนวน 3 จำนวน 1,1 และ1001 ในการคำนวณค่าเฉลี่ยเลขคณิตจะได้ผลออกมาเท่ากับ 1001 ($\frac{1+1+1001}{3}$) แต่จะได้ค่าเฉลี่ยฮาร์มอนิคเท่ากับ 1.5 ($\frac{3}{\frac{1}{1}+\frac{1}{1}+\frac{1}{1001}}$) จะเห็นได้ว่าค่าเฉลี่ยฮาร์มอนิคจะอยู่ใกล้เคียงกับจำนวนส่วนใหญ่คือจำนวน 1 สองจำนวนมากกว่าค่าเฉลี่ยเลขคณิต ซึ่งสอดคล้องกับการจัดกลุ่มที่จุดเซนทรอยด์ต้องเอนเอียงไปทางจำนวนข้อมูลส่วนใหญ่ ขั้นตอนวิธีการทำ K-Harmonic-Means สามารถอธิบายได้ดังต่อไปนี้

1) เริ่มต้นขั้นตอนวิธีโดยการสุ่มเซนทรอยด์ของกลุ่มข้อมูล

2) คำนวณค่าฟังก์ชันต้นทุน

$$\sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\left|\left| x_i - C_j \right|\right|^p}} \tag{10}$$

โดยที่ p คือคัวแปรอิสระ (Free Parameter) ($p \geq 2$)

3) คำนวณค่าความเป็นสมาชิก (membership) ของตัวอย่างข้อมูล x_i ไปยังของแต่ละกลุ่มข้อมูล C_j โดยการคำนวณจะเป็นไปดังสมการ (11)

$$\left(C_j \middle| x_i\right) = \frac{\left|\left| x_i - C_j \right|\right|^{-p-2}}{\sum_{j=1}^{k} \left|\left| x_i - C_j \right|\right|^{-p-2}} \tag{11}$$

4) คำนวณค่าน้ำหนัก (Weight) ของตัวอย่างข้อมูลตามสมการ (12)

$$w(x_i) = \frac{\sum_{j=1}^{k} \left|\left| x_i - C_j \right|\right|^{-p-2}}{\left(\sum_{j=1}^{k} \left|\left| x_i - C_j \right|\right|^{-p-2}\right)^2} \tag{12}$$

5) คำนวณเซนทรอยด์ของกลุ่มข้อมูลใหม่ตามค่าความเป็นสมาชิกของตัวอย่างข้อมูล และค่าน้ำหนักของตัวอย่างข้อมูล ดังสมการ (13)

$$C_j = \frac{\sum_{i=}^{n} m(C_j|x_i) \cdot w(x_i) \cdot x_i}{\sum_{i=1}^{n} m(C_j|x_i) \cdot w(x_i)} \quad (13)$$

6) วนซ้ำกระบวนการที่ 2 ถึงกระบวนการที่ 5 จนค่าฟังก์ชันต้นทุนที่ได้ไม่มีการเปลี่ยนแปลงหรือเปลี่ยนแปลงน้อยมาก หรือครบตามจำนวนรอบที่เหมาะสม

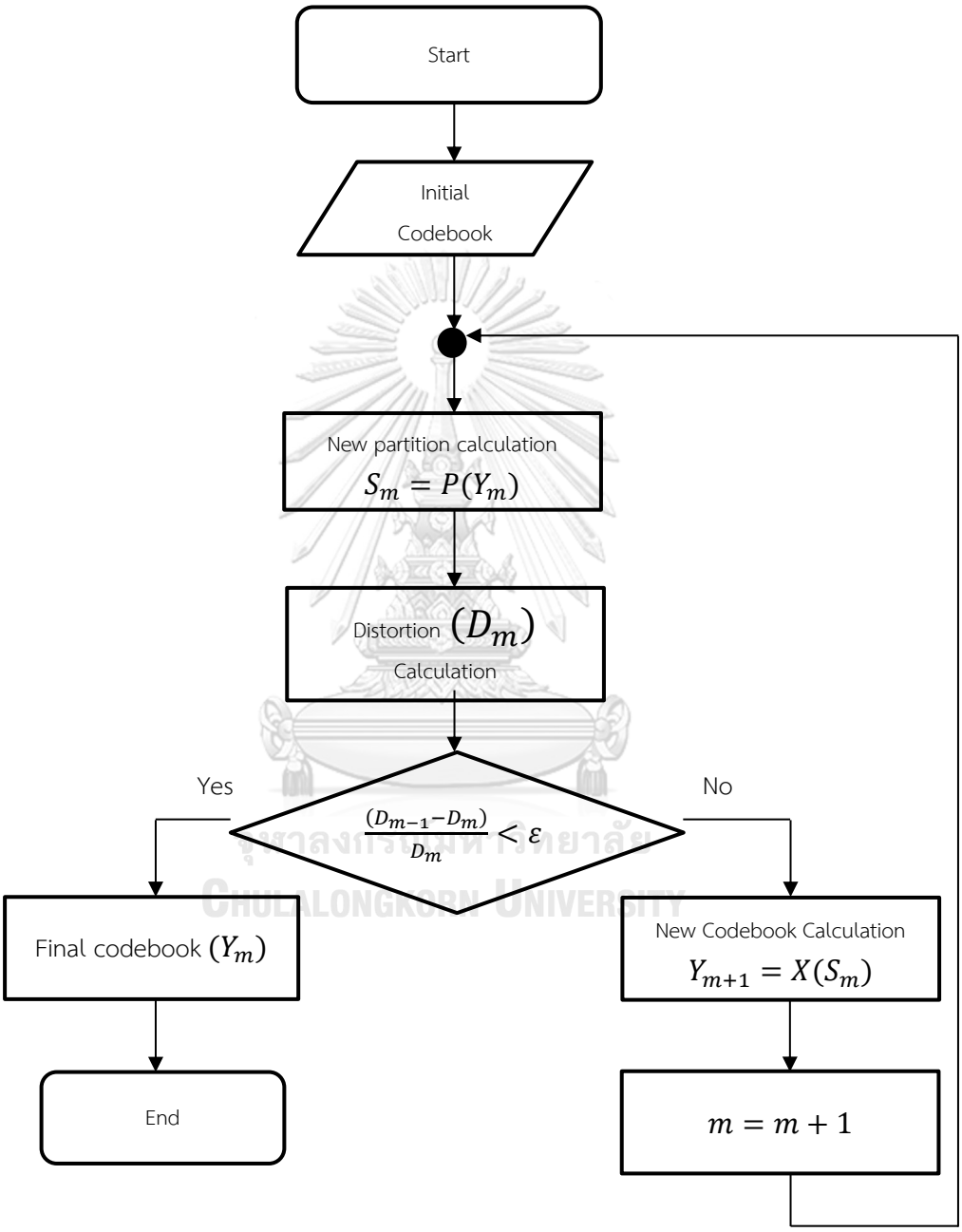
7) กำหนดกลุ่มข้อมูลให้ตัวอย่างข้อมูล x_i ไปยัง กลุ่มข้อมูล j ตามค่าสูงสุดของ $m(C_j|x_i)$

K-Harmonic-Means นอกจากแก้ไขปัญหาเซนทรอยด์การแบ่งกลุมแบบแบ่งส่วนแบบปกติได้แล้วยังสามารถแก้ไขปัญหาในเรื่องเซนทรอยด์ในการจัดกลุ่มแบบทับซ้อนได้อีกด้วย ดังเช่นในงานวิจัยของ S. Khanmohammad [32] ซึ่งนำ OKM และ K-Harmonic-Means มาประยุกต์ใช้ร่วมกันเพื่อใช้งานในชุดข้อมูลทางการแพทย์ สำหรับในงานวิจัยนี้ได้ทำการประยุกต์ขั้นตอนวิธี OKM ร่วมกับขั้นตอนวิธี K-Harmonic-Means และพัฒนาเพิ่มส่วนของการหาค่าเซนทรอยด์ที่หมาะสมยิ่งกว่าเดิมโดยประยุกต์ใช้ขั้นตอนวิธี Enhance Linde-Buzo-Gray [33]

2.3.3 ขั้นตอนวิธี Enhanced Linde-Buzo-Gray (ELBG)

ขั้นตอนวิธี ELBG เป็นขั้นตอนที่พัฒนามาจากขั้นตอนวิธี Linde-Buzo-Gray (LBG) [34] ซึ่งเป็นแก้ปัญหาแบบ Vector Quantization กระบวนการของขั้นตอนวิธีจะมีการหาค่า Quantization Error จากฟังก์ชันบิดเบือน (Distortion Function) โดยในกรณีของการแบ่งกลุ่มแบบแบ่งส่วนอย่างอย่างขั้นตอนวิธี K-Means ฟังก์ชันที่ใช้คือ Root-Mean-Square Error (RMSE)

สำหรับการทำขั้นตอนวิธี LBG มีหลักการคือ หารูปแบบของคำตอบ (Codebook) ใหม่ที่ให้ค่าของฟังก์ชันบิดเบือนน้อยกว่าเดิมในรอบที่แล้ว โดยจะหยุดเมื่อส่วนต่างของค่าฟังก์ชันน้อยกว่าค่าขีดแบ่งที่กำหนด ซึ่งสามารถแสดงให้เห็นภาพได้ดังรูปที่ 6



รูปที่ 6 แผนภูมิแสดงขั้นตอนวิธี LBG [33]

สำหรับขั้นตอนวิธี ELBG ที่พัฒนาขึ้นมานั้นจะเพิ่มแนวคิดในเรื่องของดัชนีการทำประโยชน์ (Utilization Index) ขึ้นมาเพื่อใช้พิจารณารูปแบบทิศทางในการเปลี่ยนตำแหน่งของเซนทรอยด์ในแต่ละรอบ โดยที่พิจารณาเคลื่อนเซนทรอยด์จากจุดที่มีดัชนีการทำประโยชน์น้อยเข้าหาจุดที่มีดัชนีการทำประโยชน์มาก โดยค่าดัชนีการทำประโยชน์สามารถคำนวณได้ดังต่อไปนี้

$$D_{mean} = \frac{1}{N_C} \sum_{i=1}^{N_C} D_i \tag{14}$$

$$U_i = \frac{D_i}{D_{mean}}, i = 1, \dots, N_C \tag{15}$$

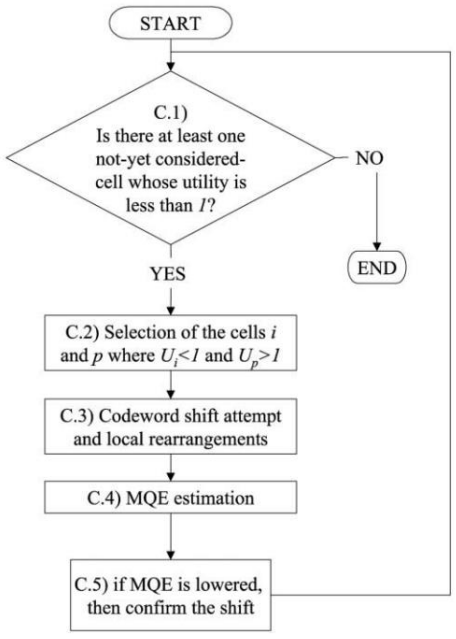
โดยที่

D_i คือค่าฟังก์ชันบิดเบือนของเซนทรอยด์ของกลุ่มข้อมูล i

D_{mean} คือค่าฟังก์ชันบิดเบือนเฉลี่ยของเซนทรอยด์ของทุกกลุ่มข้อมูล

U_i ค่าดัชนีทำประโยชน์ของเซนทรอยด์ของกลุ่มข้อมูล i

กระบวนการของขั้นตอนวิธี ELBG คือพยายามทำการเคลื่อนเซนทรอยด์ที่มีค่าดัชนีการทำประโยชน์น้อยกว่า 1 ไปในทิศทางของเซนทรอยด์ที่มีดัชนีการทำประโยชน์มากกว่า 1 และตรวจสอบผลโดยการวัดค่าฟังก์บิดเบือนโดยรวมอีกครั้งหนึ่ง ซึ่งสามารถอธิบายได้ดังรูปที่ 7



รูปที่ 7 แผนภูมิการทำงานของขั้นตอนวิธี ELBG [33]

บทที่ 3
แนวคิดและกระบวนการในการแก้ไขปัญหา

งานวิจัยนี้มีวัตถุประสงค์หลักเพื่อเพิ่มความถูกต้องแม่นยำของการจัดกลุ่มแบบทับซ้อน โดยแบ่งแนวทางในการพัฒนาออกเป็นสองประเด็นคือ การแก้ไขในเรื่องของปัญหาเซนทรอยด์อยู่ในตำแหน่งที่ไม่เหมาะสมทำให้ผลของการจัดกลุ่มข้อมูลกลุ่มข้อมูลติดอยู่กับค่าต่ำสุดสัมพัทธ์ โดยปัญหานี้จะถูกแก้โดยการใช้ขั้นตอนวิธี K-Harmonic-Means และ OKM เป็นพื้นฐาน ส่วนอีกประเด็นหนึ่งคือการนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาช่วยพัฒนาเพิ่มความถูกต้องแม่นยำในการจัดกลุ่มข้อมูลแบบทับซ้อนให้มากยิ่งขึ้น ในบทที่ 3 นี้จะกล่าวถึงแนวคิดและขั้นตอนวิธีที่ใช้ในการแก้ปัญหาพร้อมทั้งอธิบายว่าแก้ได้อย่างไร โดยก่อนที่จะนำเสนอขั้นตอนวิธีใหม่นั้นจะอธิบายถึงขั้นตอนวิธีที่ใช้เป็นพื้นฐานของงานวิจัยก่อน

3.1 การประยุกต์รวมกันของขั้นตอนวิธี OKM ขั้นตอนวิธี K-Harmonic-Means และขั้นตอนวิธี ELBG

ดังที่กล่าวมาแล้วข้างต้นว่าการจัดกลุ่มทับซ้อนแบบแบ่งส่วนนั้นมีปัญหาในเรื่องของเซนทรอยด์ของกลุ่มข้อมูลมีโอกาสที่จะอยู่ในตำแหน่งที่ไม่เหมาะสมซึ่งส่งผลให้ประสิทธิภาพในการจัดกลุ่มด้านความถูกต้องแม่นยำลดลง งานวิจัยนี้จึงได้นำขั้นตอนวิธี OKM มาประยุกต์ใช้ร่วมกับขั้นตอนวิธี K-Harmonic-Means และ ELBG เกิดเป็นขั้นตอนวิธีใหม่ซึ่งสามารถลดปัญหาดังกล่าวได้

ในส่วนของการบวนการของขั้นตอนวิธีใหม่ในงานวิจัยนี้จะถูกแบ่งออกเป็นขั้นตอนใหญ่ ๆ 3 ขั้นตอน ได้แก่

1) ขั้นตอนการกำหนดกลุ่มข้อมูลโดยขั้นตอนวิธี OKM
2) ขั้นตอนการหาเซนทรอยด์โดยขั้นตอนวิธี K-Harmonic-Means
3) ขั้นตอนการค้นหาเซนทรอยด์ที่ดีกว่าเดิมด้วยขั้นตอนวิธี ELBG

โดยกระบวนการทั้งหมดจะสามารถอธิบายได้ดังต่อไปนี้

1) เริ่มต้นเซนทรอยด์แบบสุ่ม (เหมือนกับขั้นตอนวิธี K-Means)
2) จัดกลุ่มข้อมูลตามขั้นตอนวิธี OKM จนได้กลุ่มทับซ้อนในรอบที่ 1
3) เข้าขั้นตอนวิธี K-Harmonic-Means โดยมี 3 ขั้นตอนย่อยคือ

3.1. คำนวณค่าความเป็นสมาชิก (membership) ของตัวอย่างข้อมูล x_i ไปยังของแต่ละกลุ่มข้อมูล C_j

$$(C_j|x_i) = \frac{\left|| x_i - C_j|\right|^{-p-2}}{\sum_{j=1}^{k}\left||x_i - C_j|\right|^{-p-2}} \quad (16)$$

3.2. คำนวณค่าน้ำหนัก (Weight) ของข้อมูล

$$w(x_i) = \frac{\sum_{j=1}^{k}\left||x_i - C_j|\right|^{-p-2}}{\left(\sum_{j=1}^{k}\left||x_i - C_j|\right|^{-p-2}\right)^2} \quad (17)$$

3.3. คำนวณตำแหน่งของเซนทรอยด์อิงตามค่าตัวแปรในขั้นตอนที่ 3.1 และ 3.2

$$C_j = \frac{\sum_{i=}^{n} m(C_j|x_i)\cdot w(x_i)\cdot x_i}{\sum_{i=1}^{n} m(C_j|x_i)\cdot w(x_i)} \quad (18)$$

4) ทำการกำหนดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการกำหนดกลุ่มข้อมูลแบบหลายกลุ่มของ OKM

5) ใช้ขั้นตอนวิธี ELBG เพื่อค้นหาเซนทรอยด์ที่ดีกว่าโดยลองสลับเซนทรอยด์ที่ได้กับข้อมูลเพื่อนบ้านใกล้เคียงแล้ววัดผลค่าของฟังก์ชันบิดเบือน ทำไปเรื่อย ๆ จนกว่าผลต่างของฟังก์ชันบิดเบือนกับรอบที่แล้วน้อยกว่าค่าขีดแบ่ง หรือไม่สามารถหารูปแบบของคำตอบที่ดีกว่าเดิมได้

6) หากตรงตามเงื่อนไขด้านล่างดังต่อไปนี้หยุดกระบวนการของขั้นตอนวิธี ไม่เช่นนั้นย้อนกลับไปขั้นตอนที่ 3

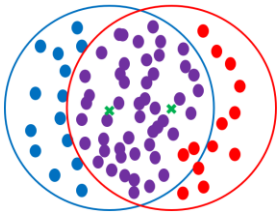
- กลุ่มของข้อมูลไม่มีการเปลี่ยนแปลงไปจากรอบที่แล้ว
- ค่าฟังก์ชันต้นทุนมีการเปลี่ยนแปลงน้อยกว่าค่าขีดแบ่ง
- ครบตามจำนวนรอบที่กำหนดไว้

หลังจากทำการประยุกต์รวมกันของทั้งสามขั้นตอนวิธี พบว่าได้การจัดกลุ่มทับซ้อนแบบแบ่งส่วนแบบใหม่ที่มีประสิทธิภาพทางด้านความถูกต้องแม่นยำมากยิ่งขึ้นเทียบกับการจัดกลุ่มแบบเดิมคือ OKM โดยตัดสินจากผลการทดลองกับชุดข้อมูลหลายหมวดหมู่ 20 ชุดข้อมูล

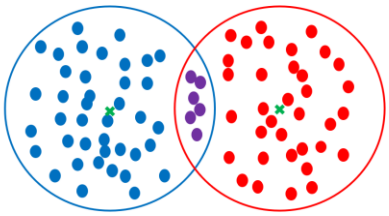
3.2 การประยุกต์ใช้องค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูล

เนื่องจากชุดข้อมูลที่ใช้กับการจัดกลุ่มแบบทับซ้อนนั้นเป็นแบบหลายหมวดหมู่ กลุ่มของข้อมูลที่ได้จึงไม่ได้แยกขาดจากกันเหมือนกับการจัดกลุ่มแบบธรรมดาแต่บางส่วนของกลุ่มข้อมูลจะมี

โอกาสคาบเกี่ยวกัน โอกาสที่จะคาบเกี่ยวกันนี้จะมากหรือน้อยนั้นประกอบด้วยปัจจัยหลายอย่าง แต่ปัจจัยอย่างหนึ่งที่เห็นได้อย่างชัดเจนคือหากกลุ่มข้อมูลอยู่ใกล้กันมากแล้วนั้นโอกาสที่กลุ่มข้อมูลจะคาบเกี่ยวกันจะมากกว่ากลุ่มข้อมูลที่อยู่ไกลกัน (หรืออีกนัยหนึ่งคือมีโอกาสที่จะมีข้อมูลร่วมกันเป็นจำนวนมากกว่า) ดังที่สามารถแสดงได้ตามรูปที่ 8 จะเห็นได้อย่างชัดเจนว่าหากข้อมูลมีการกระจายตัวที่ใกล้เคียงกันแล้วกลุ่มข้อมูลที่อยู่ใกล้กันจะมีปริมาณจำนวนข้อมูลร่วมกันมากกว่า



รูปที่ 8 (ก) แสดงกลุ่มข้อมูลที่ชิดกัน



รูปที่ 8 (ข) แสดงกลุ่มข้อมูลที่ห่างกัน

รูปที่ 8 แสดงความสัมพันธ์ระหว่างกลุ่มข้อมูลที่ส่งผลต่อโอกาสในการคาบเกี่ยวกันระหว่างกลุ่มข้อมูล โดยที่ รูปที่ 8 (ก) แสดงกลุ่มข้อมูลที่ชิดกัน รูปที่ 8 (ข) แสดงกลุ่มข้อมูลที่ห่างกัน

เพื่อเพิ่มประสิทธิภาพของการจัดกลุ่มทับซ้อนด้านความถูกต้องแม่นยำในการจัดกลุ่ม งานวิจัยนี้ได้นำเอาองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลนี้มาใช้ประโยชน์ โดยการนำมาประยุกต์ใช้เพิ่มเข้าไปในฟังก์ชันต้นทุนของขั้นตอนวิธี OKM โดยเปลี่ยนฟังก์ชันต้นทุนเดิมจาก

$$||x_i - \phi(x_i)||^2 \tag{19}$$

เป็น

$$||x_i - \phi(x_i)||^2 + \gamma \tag{20}$$

โดยที่

$$\gamma = \log_2 \left|\left|cluser_{prev} - cluster_{next}\right|\right|^2 \tag{21}$$

ในที่นี้ γ ทำหน้าที่เป็นค่าลงโทษโดยเป็นตัวแทนระยะทางระหว่างกลุ่มข้อมูลก่อนหน้ากับกลุ่มข้อมูลถัดไป หากระยะทางระหว่างกลุ่มข้อมูลมีมากค่าลงโทษก็จะมีมากตาม ทำให้โอกาสที่ข้อมูลจะถูกกำหนดให้อยู่บนพื้นที่ทับซ้อนของกลุ่มข้อมูลที่อยู่ห่างกันมาก ๆ มีน้อยซึ่งสอดคล้องตามสมมติฐานที่ตั้งไว้ จากการทดลองกับชุดข้อมูลหลายหมวดหมู่จำนวน 20 ชุดข้อมูลพบว่าขั้นตอนวิธีนี้ช่วยเพิ่มประสิทธิภาพของการจัดกลุ่มข้อมูลในแง่ของความถูกต้องแม่นยำได้จริง

3.3 การประยุกต์รวมแนวคิดทั้งหมด

ดังที่กล่าวมาแล้วว่างานวิจัยนี้มีจุดประสงค์เพื่อเพิ่มประสิทธิภาพการจัดกลุ่มทับซ้อนแบบแบ่งส่วนในด้านของความถูกต้องแม่นยำในการจัดกลุ่ม โดยแบ่งหัวข้อในการพัฒนาออกเป็น 2 ประเด็นด้วยกันคือ

1) การแก้ปัญหาเซนทรอยด์ซึ่งมีโอกาสอยู่ในตำแหน่งที่ไม่เหมาะสมทำให้ผลการจัดกลุ่มออกมาไม่มีประสิทธิภาพ (หัวข้อ 3.2.1)
2) การนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาใช้ (หัวข้อ 3.2.2)

สำหรับในหัวข้อ 3.2.3 นี้จะกล่าวถึงการรวบรวมขั้นตอนวิธีในหัวข้อ 3.2.1 และหัวข้อ 3.2.2 เข้าไว้ด้วยกันเป็นขั้นตอนวิธีใหม่ขั้นตอนวิธีเดียว โดยขั้นตอนวิธีใหม่ที่ได้นั้นจะมีกระบวนการตามขั้นตอนวิธีในหัวข้อ 3.2.1 แต่จะใช้ฟังก์ชันต้นทุนที่พัฒนาแล้วตามหัวข้อ 3.2.2 ซึ่งกระบวนการทั้งหมดสามารถอธิบายได้ดังต่อไปนี้

1) เริ่มต้นเซนทรอยด์แบบสุ่ม (เหมือนกับขั้นตอนวิธี K-Means)
2) จัดกลุ่มข้อมูลตามขั้นตอนวิธี OKM จนได้กลุ่มทับซ้อนในรอบที่ 1
3) เข้าขั้นตอนวิธี K-Harmonic-Means เหมือนดังหัวข้อ 3.1 ขั้นตอนที่ 3)
4) ทำการกำหนดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการกำหนดกลุ่มข้อมูลแบบหลายกลุ่มของ OKM แต่เปลี่ยนฟังก์ชันต้นทุนใหม่เป็น

$$||x_i - \phi(x_i)||^2 + \gamma \tag{23}$$

โดยที่

$$\gamma = \log_2 \left|\left|cluser_{prev} - cluster_{next}\right|\right|^2 \tag{24}$$

เพื่อให้สอดคล้องกับแนวคิดเกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูล

5) ใช้ขั้นตอนวิธี ELBG เพื่อค้นหาเซนทรอยด์ที่ดีกว่าโดยลองสลับเซนทรอยด์ที่ได้กับข้อมูลเพื่อนบ้านใกล้เคียงแล้ววัดผลค่าของฟังก์ชันบิดเบือน ทำไปเรื่อย ๆ จนกว่าผลต่างของฟังก์ชันบิดเบือนกับรอบที่แล้วน้อยกว่าค่าขีดแบ่ง หรือไม่สามารถหารูปแบบของคำตอบที่ดีกว่าเดิมได้
6) หากตรงตามเงื่อนไขดังต่อไปนี้
   - กลุ่มของข้อมูลไม่มีการเปลี่ยนแปลงไปจากรอบที่แล้ว
   - ค่าฟังก์ชันต้นทุนมีการเปลี่ยนแปลงน้อยกว่าค่าขีดแบ่ง

- ครบตามจำนวนรอบที่กำหนดไว้

หยุดกระบวนการของขั้นตอนวิธี ไม่เช่นนั้นย้อนกลับไปขั้นตอนที่ 3 รหัสเทียมของขั้นตอนวิธีแสดงได้ดังรูปที่ 9

New Method (x, t_{max}, ϵ)

Input : X: a set of data vector $\mathbb{R}^P$, t_{max}: optional number of iterations, ϵ optional threshold on the objective

Output : $\{\pi\}_{c=1}^{k}$: final coverage of the data points

1. Draw randomly k initial cluster centroids $\{m_c^0\}_{c=1}^{k}$ in X.
2. For each $x_i \, \epsilon \, X$ compute the assignments $A_i^{(0)} = ASSIGN\,(x_i, \{m_c^{(0)}\}_{c=1}^{k})$ following the multi-assignment in the next section and derive the initial coverage $\{\pi_c^{(0)}\}_{c=1}^{k}$ with new cost function $||x_i - \phi(x_i)||^2 + \gamma$
3. Set $t = 0$.
4. *Use K-Harmonic-Means algorithm to find optimal clusters' centroids.*
5. *Use ELBG algorithm to find better solution*
6. For each cluster $\pi_c^{(t)}$ successively, compute the new centroid $m_c^{(t+1)}$.
7. For each $x_i \, \epsilon \, X$ compute the assignments $A_i^{(t+1)} = ASSIGN\,(x_i, \{m_c^{(t+1)}\}_{c=1}^{k}, A_i^{(t)})$ and derive the new coverage $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.
8. If not converged, $t_{max} > t$ or $\mathcal{J}(\{\pi^{(t)}\}) - \mathcal{J}(\{\pi^{(t+1)}\}) > \epsilon$, set $t = t + 1$ and go to step 4 otherwise output final clusters $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.

รูปที่ 9 รหัสเทียมของขั้นตอนวิธีใหม่ในงานวิจัยนี้

หลังจากรวมกระบวนการทั้งหมดจากผลการทดลองพบว่าประสิทธิภาพของการจัดกลุ่มข้อมูลเพิ่มขึ้นอย่างเห็นได้ชัดเจนจากการประยุกต์ใช้ขั้นตอนวิธีเพียงอย่างใดอย่างหนึ่ง ซึ่งสามารถพิสูจน์ได้ว่าขั้นตอนวิธีทั้งสองนี้สามารถช่วยเพิ่มประสิทธิภาพของการจัดกลุ่มทับซ้อนได้จริง และสามารถทำงานได้อย่างสอดคล้องกัน

บทที่ 4
การทดลองและวิเคราะห์ผล

สำหรับในบทนี้จะกล่าวถึงการทดลองเพื่อเปรียบเทียบขั้นตอนวิธีที่งานวิจัยชิ้นนี้ได้นำเสนอกับงานวิจัยอื่นเพื่อประเมินว่า ขั้นตอนวิธีใหม่ที่ได้นั้นช่วยแก้ไขปัญหาที่เกิดขึ้นและช่วยเพิ่มประสิทธิภาพของการจัดกลุ่มแบบทับซ้อนได้จริงหรือไม่ โดยจะทำการทดลองกับชุดข้อมูลที่มีอยู่จริงจากเว็บไซต์ Mulan [14] และ เว็บไซต์ LAIM [36] ทั้งหมดจำนวน 20 ชุดข้อมูล (ชุดข้อมูลและรายละเอียดของชุดข้อมูลจะถูกแจกแจงในหัวข้อถัดไป) โดยจะทำการทดลองด้วย 3 ขั้นตอนวิธีคือ

1) OKM ที่พัฒนาขึ้นโดยการรวมกับขั้นตอนวิธี K-Harmonic-Means และขั้นตอนวิธี ELBG
2) OKM ที่พัฒนาขึ้นโดยการใช้องค์ความรู้เกี่ยวกับกลุ่มข้อมูล
3) OKM ที่พัฒนาขึ้นโดยใช้ขั้นตอนวิธี K-Harmonic-Means ELBG และองค์ความรู้เกี่ยวกับกลุ่มข้อมูล (CREK-OKM)

ทั้งนี้ที่แบ่งเป็น 3 การทดลองเพื่อแสดงให้เห็นว่าแต่ละประเด็นที่งานวิจัยทำการศึกษานั้นสามารถนำมาใช้พัฒนาการจัดกลุ่มทับซ้อนแบบแบ่งส่วนได้จริงครบทุกประเด็น โดยในการชี้วัดประสิทธิภาพจะมีการเปรียบเทียบกับขั้นตอน OKM เดิม และขั้นตอนที่พัฒนาขึ้นมาจาก OKM ได้แก่ CWOKM และ ROKM นอกจากนี้ยังมีการเปรียบเทียบกับขั้นตอนวิธี FCM ที่มีปัญหาในเรื่องของเซนทรอยด์อยู่ในตำแหน่งที่ไม่เหมาะสมน้อย ขั้นตอนวิธี MOC ที่ไม่ใช่การจัดกลุ่มทับซ้อนประเภทแบ่งส่วน และขั้นตอนวิธีการจัดกลุ่มทับซ้อนแบบแบ่งส่วนล่าสุด NEO-K-Means โดยจะเปรียบเทียบกันในเชิงความถูกต้องแม่นยำในการจัดกลุ่มในหน่วยของ F1

4.1 ชุดข้อมูลที่ใช้ในการทดลองและสถิติรายละเอียดข้อมูล

ในการทดลองนี้ได้ใช้ชุดข้อมูลหลายหมวดหมู่มาตรฐาน 20 ชุดข้อมูลมาจากเว็บไซต์ Mulan และ LAIM ในตารางสถิติของข้อมูลมีตัวแปรอยู่ 2 ตัวแปรที่มีผลในการทับซ้อนกันของกลุ่มข้อมูล นั่นคือค่าคาร์ดินัลลิตี (Cardinality) และค่าความหนาแน่น (Density) ซึ่งค่าคาร์ดินัลลิตีเป็นจำนวนเฉลี่ยของกลุ่มข้อมูลต่อหนึ่งตัวอย่างข้อมูล ส่วนค่าความหนาแน่นคือค่าคาร์ดินัลลิตีนำมามาเฉลี่ยด้วยกลุ่มข้อมูลอีกครั้งหนึ่ง ค่าทั้งสองนี้สามารถแสดงได้ดังสมการ

$$Cardinality = \frac{1}{N}\sum_{i=1}^{N} |Y_i|, \text{ โดยที่ } |Y_i| \text{ คือจำนวนหมวดหมู่ของข้อ } x_i$$

$$Density = \frac{1}{N}\sum_{i=1}^{N} \frac{|Y_i|}{|C|}, \text{ โดยที่ } C \text{ คือจำนวนหมวดหมู่ทั้งหมด}$$

ชุดข้อมูลทั้งหมดและสถิติข้อมูลที่ใช้ในการทดลองเป็นไปดังตารางที่ 2

ตารางที่ 2 ชุดข้อมูลที่ใช้ในการทดลองและสถิติรายละเอียดของข้อมูล

Data set	Instances	Features	Classes	Density	Cardinality
Emotion	593	72	6	0.311	1.869
Yeast	2,417	103	14	0.303	4.237
Scene	2,407	294	6	0.179	1.074
Mediamill	49,307	120	101	0.043	4.376
RCV1 (set1)	6,000	47,236	101	0.029	2.880
RCV1 (set2)	6,000	47,236	101	0.026	2.634
RCV1 (set3)	6,000	47,236	101	0.026	2.614
RCV1 (set4)	6,000	47,229	101	0.025	2.484
RCV1 (set5)	6,000	47,235	101	0.026	2.642
CAL500	502	68	174	0.150	26.044
EUR-Lex (directory codes)	19,348	5,000	412	0.003	1.292
EUR-Lex (subject matters)	19,348	5,000	201	0.011	2.213
EUR-Lex (eurovoc descriptors)	19,348	5,000	3993	0.001	5.310
Birds	645	260	19	0.053	1.013
Plant	978	440	12	0.089	1.078
Human	3,106	440	14	0.084	1.185
Flags	194	19	7	0.484	3.391
Yahoo (Art)	7,484	500	26	0.063	1.653
Yahoo (Health)	9,205	500	32	0.051	1.644
Yahoo (Business)	11,214	500	30	0.053	1.598

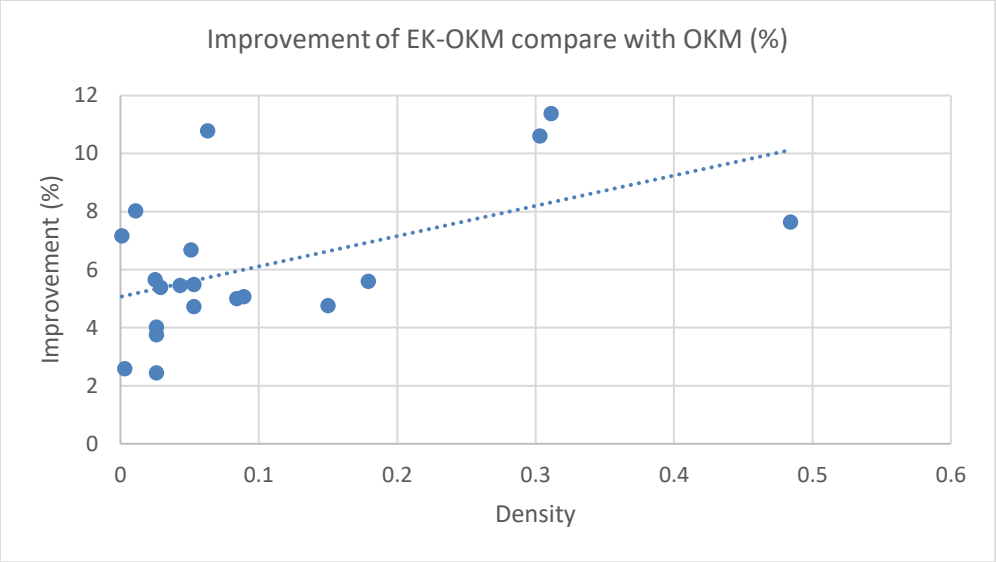
4.2 การวิเคราะห์และสรุปผลการทดลองการนำเอาขั้นตอนวิธี K-Harmonic-Means และ ELBG มาประยุกต์ใช้

จากผลการทดลองกับชุดข้อมูลหลายหมวดหมู่ทั้งหมด 20 ชุดข้อมูล แล้วพบว่าหลังจากประยุกต์ใช้ขั้นตอนวิธี K-Harmonic-Means ร่วมกับขั้นตอนวิธี OKM ประสิทธิภาพของการจัดกลุ่มมีการพัฒนามากยิ่งขึ้นในในแง่ของความถูกต้องแม่นยำคิดคำนวณตามค่า F1 ในทุกชุดข้อมูลของการทดลอง จากการวิเคราะห์ผลการทดลองพบว่าประสิทธิภาพของขั้นตอนวิธีสัมพันธ์กับความหนาแน่นของข้อมูลอย่างมีนัยสำคัญ ความหนาแน่นของข้อมูลเป็นตัวแปรที่บ่งชี้ถึงการทับซ้อนกันของกลุ่มข้อมูล ค่าความหนาแน่นสูงบ่งชี้ถึงกลุ่มข้อมูลมีการทับซ้อนกันปริมาณสูง ซึ่งสามารถอุปนัยได้ว่ากลุ่มข้อมูลน่าจะค่อนข้างอยู่เกาะกลุ่มกัน ซึ่งในกรณีนี้ตำแหน่งของเซนทรอยด์จะมีผลกระทบอย่างมากต่อประสิทธิภาพของการจัดกลุ่ม เนื่องจากหากเซนทรอยด์ที่เป็นจุดศูนย์กลางของกลุ่มข้อมูลถูกวางอยู่ใน

ตำแหน่งที่ไม่เหมาะสมแล้ว การแบ่งกลุ่มข้อมูลที่อยู่ชิดกันจะทำได้ลำบากมากยิ่งขึ้น ด้วยเหตุนี้จากการทดลองจึงเห็นได้ว่าชุดข้อมูลที่มีค่าความหนาแน่นของข้อมูลสูงนั้นเมื่อมีการประยุกต์ใช้ขั้นตอนวิธี K-Harmonic-Means และ ELBG ร่วมด้วยก็จะมีประสิทธิภาพเพิ่มขึ้นอย่างมาก ซึ่งสามารถสังเกตได้จากกราฟรูปที่ 10 และสามารถเห็นได้ว่าชุดข้อมูลที่มีความหนาแน่นมากอย่าง Flag Yeast และ Emotion มีพัฒนาการมากกว่าชุดข้อมูลอื่น ยิ่งไปกว่านั้นการประยุกต์ใช้ขั้นตอนวิธี ELBG ยังช่วยเพิ่มประสิทธิภาพของผลลัพธ์ได้อีกในทุกชุดข้อมูลดังในตารางที่ 3

ตารางที่ 3 ผลการทดลองเปรียบเทียบความแม่นยำของขั้นตอนวิธี OKM ขั้นตอนวิธีประยุกต์ OKM กับ K-Harmonic-Means OKM (K-OKM) และขั้นตอนวิธี ELBG K-Harmonic-Means OKM (EK-OKM) ในหน่วย F1

Data set	OKM	K-OKM	EK-OKM	Density
Emotion	0.527	0.557	0.587	0.311
Yeast	0.311	0.335	0.344	0.303
Scene	0.571	0.584	0.603	0.179
Mediamill	0.494	0.501	0.521	0.043
RCV-1 (set 1)	0.463	0.471	0.488	0.029
RCV-1 (set 2)	0.452	0.457	0.469	0.026
RCV-1 (set 3)	0.531	0.538	0.544	0.026
RCV-1 (set 4)	0.442	0.449	0.467	0.025
RCV-1 (set 5)	0.497	0.501	0.517	0.026
CAL500	0.273	0.277	0.286	0.150
EUR-Lex (directory codes)	0.309	0.312	0.317	0.003
EUR-Lex (subject matters)	0.411	0.426	0.444	0.011
EUR-Lex (eurovoc descriptors)	0.223	0.225	0.239	0.001
Birds	0.613	0.625	0.642	0.053
Plant	0.591	0.608	0.621	0.089
Human	0.439	0.454	0.461	0.084
Flags	0.523	0.554	0.563	0.484
Yahoo (Art)	0.417	0.437	0.462	0.063
Yahoo (Health)	0.449	0.464	0.479	0.051
Yahoo (Business)	0.582	0.592	0.614	0.053



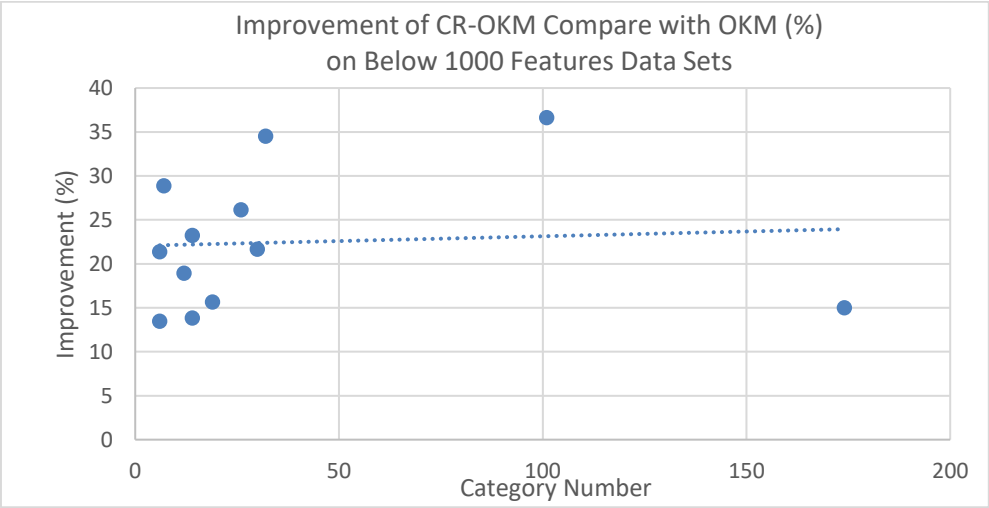
รูปที่ 10 กราฟแสดงพัฒนาการของขั้นตอนวิธี EK-OKM เปรียบเทียบกับ OKM ในหน่วยเปอร์เซ็นต์เทียบกับค่าความหนาแน่นของข้อมูล

4.3 การวิเคราะห์และสรุปผลการทดลองการนำเอาองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาประยุกต์ใช้

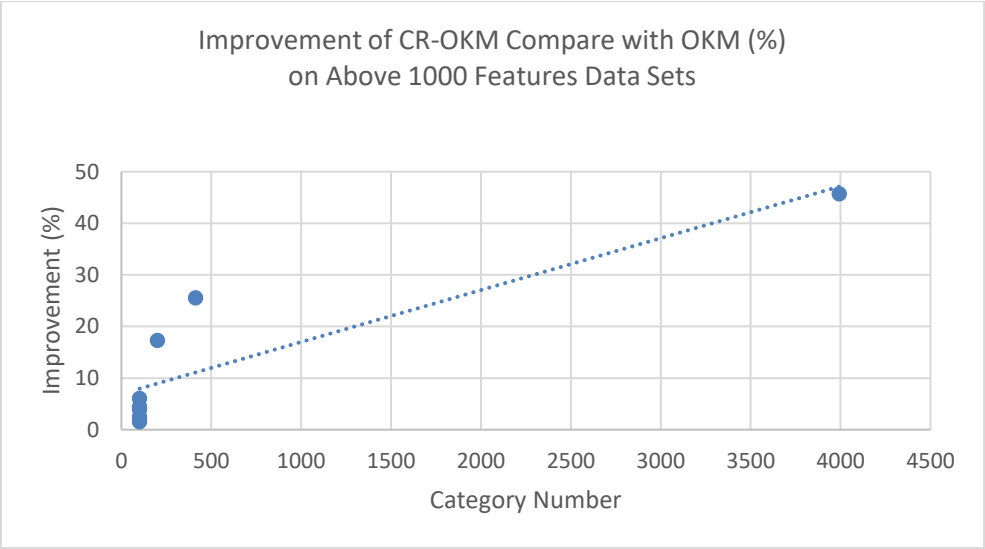
จากผลการทดลองกับชุดข้อมูลแบบหลายหมวดหมู่ทั้งหมด 20 ชุดข้อมูลพบว่าหลังจากประยุกต์ใช้องค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลแล้วนั้นประสิทธิภาพของการจัดกลุ่มข้อมูลด้านความถูกต้องแม่นยำในการจัดกลุ่มเพิ่มขึ้นอย่างมนัยสำคัญ จากการวิเคราะห์ชุดข้อมูลและผลการทดลองพบว่าประสิทธิภาพของขั้นตอนวิธีสัมพันธ์โดยตรงกับจำนวนกลุ่มข้อมูลและฟีเจอร์ของข้อมูล โดยพบว่าหากกลุ่มข้อมูลมีจำนวนมากประสิทธิภาพของการจัดกลุ่มก็จะเพิ่มมากขึ้นตาม ทั้งนี้เป็นเพราะเมื่อมีจำนวนกลุ่มข้อมูลมีมากในการประมวลผลเพื่อการจัดกลุ่มค่าลงโทษก็จะมีอิทธิพลมากตาม ดังจะสามารถสังเกตได้จากกราฟในรูปที่ 11 และ 12 ในทางตรงกันข้ามมีมิติข้อมูลที่สูงแล้วประสิทธิภาพของขั้นตอนวิธีจะลดลง เนื่องจากในกรณีที่มิติข้อมูลสูงการประมวลผลที่ซับซ้อนจนเกินไปอาจจะทำให้ค่าที่ได้ไม่ใช่สิ่งชี้วัดระยะทางที่เหมาะสม ขั้นตอนวิธีจึงมีประสิทธิภาพลดลง

ตารางที่ 4 ผลการทดลองเปรียบเทียบความถูกต้องแม่นยำในหน่วย F1ของขั้นตอนวิธี OKM กับ Cluster Relation OKM (CR-OKM)

Data Set	OKM	CR-OKM	Category	Feature
Emotion	0.527	0.598	6	72
Yeast	0.311	0.354	14	103
Scene	0.571	0.693	6	294
Mediamill	0.494	0.675	101	120
RCV-1 (set 1)	0.463	0.481	101	47,236
RCV-1 (set 2)	0.452	0.472	101	47,236
RCV-1 (set 3)	0.531	0.539	101	47,236
RCV-1 (set 4)	0.442	0.453	101	47,229
RCV-1 (set 5)	0.497	0.527	101	47,235
CAL500	0.273	0.314	174	68
EUR-Lex (directory codes)	0.309	0.388	412	5,000
EUR-Lex (subject matters)	0.411	0.482	201	5,000
EUR-Lex (eurovoc descriptors)	0.223	0.325	3,993	5,000
Birds	0.613	0.709	19	260
Plant	0.591	0.703	12	440
Human	0.439	0.541	14	440
Flags	0.523	0.674	7	19
Yahoo (Art)	0.417	0.526	26	500
Yahoo (Health)	0.449	0.604	32	500
Yahoo (Business)	0.582	0.708	30	500



รูปที่ 11 กราฟแสดงพัฒนาการของ CR-OKM เปรียบเทียบกับ OKM สำหรับชุดข้อมูลที่มีฟีเจอร์ต่ำกว่า 1000 ฟีเจอร์ในหน่วยเปอร์เซ็นต์เทียบกับจำนวนกลุ่มข้อมูล



รูปที่ 12 กราฟแสดงพัฒนาการของ CR-OKM เปรียบเทียบกับ OKM สำหรับชุดข้อมูลที่มีฟีเจอร์สูงกว่า 1000 ฟีเจอร์ในหน่วยเปอร์เซ็นต์เทียบกับจำนวนกลุ่มข้อมูล

4.4 การวิเคราะห์ผลการทดลองหลังจากการประยุกต์ใช้ขั้นตอนวิธีทั้งหมดและเปรียบเทียบกับขั้นตอนวิธีอื่น

หลังจากทำการประยุกต์ใช้ขั้นตอนวิธีทั้งหมดเข้าด้วยกันพบว่าประสิทธิภาพเพิ่มขึ้นจากทั้งการประยุกต์ใช้ K-Harmonic-Means และขั้นตอนวิธี ELBG (หัวข้อ 4.2) และการประยุกต์ใช้องค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูล (หัวข้อ 4.3) เพียงอย่างใดอย่างหนึ่งอย่างเดียวอย่างมีนัยสำคัญ ซึ่งสามารถเป็นเครื่องบ่งชี้ได้ว่าขั้นตอนวิธีทั้งสองสามารถทำงานร่วมกันได้อย่างดีเมื่อเปรียบเทียบกับขั้นตอนวิธีอื่นพบว่าจากการทดลองกับชุดข้อมูล 20 ชุดข้อมูลขั้นตอนวิธีจากงานวิจัยนี้มีประสิทธิภาพด้านความถูกต้องแม่นยำสูงกว่าถึง 17 ชุดข้อมูล จาก 20 ชุดข้อมูล โดยเมื่อวิเคราะห์ชุดข้อมูล RCV1 set1 ที่ให้ผลการทดลองที่ให้ประสิทธิภาพด้อยกว่าพบว่าเป็นชุดข้อมูลที่มีมิติของข้อมูลสูงทำให้การประยุกต์ใช้ความสัมพันธ์ระหว่างกลุ่มข้อมูลมีประสิทธิภาพไม่มากเท่าที่ควร และอีก 2 ชุดข้อมูลที่ขั้นตอนในงานวิจัยนี้มีประสิทธิภาพด้อยกว่าคือ EUR-Lex (directory codes) และ ชุดข้อมูล EUR-Lex (eurovoc descriptors) เป็นชุดข้อมูลที่มีความหนาแน่นน้อยมาก คือ 0.001 และ 0.003 อาจกล่าวได้ว่ากลุ่มข้อมูลแทบจะไม่ได้ทับซ้อนกันเลย การจัดกลุ่มข้อมูลที่มีแนวทางจัดการกับข้อมูลโดยเน้นที่การจำแนกการกระจายตัวของข้อมูลแทนที่จะเป็นการทับซ้อนของข้อมูลอย่างขั้นตอนวิธี MOC จึงมีประสิทธิภาพสูงกว่า ผลการทดลองถูกแสดงและเปรียบเทียบดังในตารางที่ 5 และ 6

ตารางที่ 5 ผลการเปรียบเทียบความถูกต้องแม่นยำของแต่ละขั้นตอนวิธีในงานวิจัยในหน่วย F1

Data Set	OKM	EK-OKM	CR-OKM	CREK-OKM
Emotion	0.527	0.587	0.598	0.601
Yeast	0.311	0.344	0.354	0.388
Scene	0.571	0.603	0.693	0.701
Medialmill	0.494	0.521	0.675	0.689
ECV1 (set 1)	0.463	0.488	0.481	0.511
ECV1 (set 2)	0.452	0.469	0.472	0.489
RCV1 (set 3)	0.531	0.544	0.539	0.563
RCV1 (set 4)	0.442	0.467	0.453	0.481
RCV1 (set 5)	0.497	0.517	0.527	0.538
CAL500	0.273	0.286	0.314	0.337
EUR-Lex (directory codes)	0.309	0.317	0.388	0.393
EUR-Lex (subject matters)	0.411	0.444	0.482	0.521
EUR-Lex (eurovoc descriptors)	0.223	0.239	0.325	0.337
Birds	0.613	0.642	0.709	0.711
Plant	0.591	0.621	0.703	0.715
Human	0.439	0.461	0.541	0.566
Flags	0.523	0.563	0.674	0.683
Yahoo (Art)	0.417	0.462	0.526	0.534
Yahoo (Health)	0.449	0.479	0.604	0.612
Yahoo (Business)	0.582	0.614	0.708	0.722

ตารางที่ 6 ผลการทดลองเปรียบเทียบความถูกต้องแม่นยำในหน่วย F1 ของขั้นตอนวิธีในงานวิจัยกับขั้นตอนวิธีอื่น

Data Set	ROKM	MOC	CWOKM	FCM	NEO-KM	OKM	CREK-OKM
Emotion	0.522	0.442	0.534	0.448	0.55	0.527	**0.601**
Yeast	0.309	0.224	0.317	0.309	0.366	0.311	**0.388**
Scene	0.588	0.470	0.581	0.429	0.626	0.571	**0.701**
Medialmill	0.501	0.424	0.507	0.335	0.477	0.494	**0.689**
RCV1 (set 1)	0.455	0.435	**0.523**	0.387	0.503	0.463	**0.511**
RCV1 (set 2)	0.467	0.488	0.462	0.411	0.498	0.452	**0.489**
RCV1 (set 3)	0.502	0.494	0.527	0.362	0.504	0.531	**0.563**
RCV1 (set 4)	0.484	0.396	0.421	0.453	0.498	0.442	**0.481**
RCV1 (set 5)	0.512	0.492	0.503	0.404	0.512	0.497	**0.538**
CAL500	0.285	0.216	0.282	0.261	0.254	0.273	**0.337**
EUR-Lex (directory codes)	0.297	**0.551**	0.314	0.227	0.312	0.309	0.493
EUR-Lex (subject matters)	0.432	0.601	0.428	0.340	0.387	0.411	0.**624**
EUR-Lex (eurovoc descriptors)	0.239	**0.464**	0.211	0.207	0.241	0.223	0.337
Birds	0.639	0.531	0.627	0.443	0.562	0.613	**0.711**
Plant	0.528	0.434	0.607	0.612	0.62	0.591	**0.715**
Human	0.444	0.339	0.472	0.413	0.347	0.439	**0.566**
Flags	0.609	0.627	0.613	0.577	0.581	0.523	**0.683**
Yahoo (Art)	0.441	0.367	0.435	0.257	0.332	0.417	**0.534**
Yahoo (Health)	0.481	0.511	0.487	0.426	0.520	0.449	**0.612**
Yahoo (Business)	0.524	0.447	0.512	0.603	0.611	0.582	**0.722**

บทที่ 5
สรุปผลการวิจัย

ในการจัดกลุ่มข้อมูลประเภทหลายหมวดหมู่จำเป็นที่จะต้องใช้การจัดกลุ่มที่อนุญาตให้กลุ่มข้อมูลทับซ้อนกันได้ซึ่งเรียกว่า “การจัดกลุ่มทับซ้อน” งานวิจัยนี้ให้ความสนใจที่การจัดกลุ่มทับซ้อนแบบแบ่งส่วน การจัดกลุ่มทับซ้อนแบบแบ่งส่วนส่วนใหญ่พัฒนามาจากขั้นตอนวิธี K-Means ซึ่งยังมีปัญหาในเรื่องของเซนทรอยด์อยู่ในตำแหน่งไม่เหมาะสมตามขั้นตอนวิธีต้นแบบ อีกประเด็นหนึ่งที่งานวิจัยนี้ให้ความสนใจคือการนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาใช้ประโยชน์เพื่อเพิ่มประสิทธิภาพทางด้านความถูกต้องแม่นยำในการจัดกลุ่ม ในที่นี้จะสรุปผลของการศึกษาวิจัยทีละประเด็น

5.1 สรุปผลการวิจัยการแก้ไขปัญหาตำแหน่งเซนทรอยด์ของกลุ่มข้อมูลไม่เหมาะสมด้วยขั้นตอนวิธี K-Harmonic-Means และขั้นตอนวิธี ELBG

จากผลการวิจัยโดยทดสอบกับชุดข้อมูลหลายหมวดหมู่ 20 ชุดข้อมูลพบว่าการประยุกต์ใช้ขั้นตอนวิธี K-Harmonic-Means สามารถแก้ปัญหานี้ได้ โดยสังเกตเห็นได้จากประสิทธิภาพทางด้านความถูกต้องแม่นยำที่เพิ่มขึ้นหลังจากการประยุกต์ใช้ขั้นตอนวิธี โดยประสิทธิภาพของขั้นตอนวิธีจะสัมพันธ์กันกับค่าความหนาแน่นของข้อมูล หากชุดข้อมูลมีความหนาแน่นมากประสิทธิภาพของขั้นตอนวิธีก็จะยิ่งสูงตามไปด้วย เนื่องจากความหนาแน่นของข้อมูลบ่งชี้ถึงการทับซ้อนกันของกลุ่มข้อมูล ชุดข้อมูลที่มีความหนาแน่นสูงคือชุดข้อมูลที่มีปริมาณการทับซ้อนกันสูงและอยู่เกาะกลุ่มกันทำให้มีโอกาสสูงที่จะเกิดปัญหาเซนทรอยด์อยู่ในตำแหน่งที่ไม่เหมาะสม เมื่อมีการประยุกต์ขั้นตอนวิธี K-Harmonic-Means ประสิทธิภาพจึงสูงขึ้นและยิ่งพัฒนากว่าเดิมเมื่อประยุกต์ขั้นตอนวิธี ELBG ร่วมด้วย จึงสามารถสรุปได้ว่าการประยุกต์ใช้ขั้นตอนวิธี K-Harmonic-Means สามารถใช้แก้ปัญหาในเรื่องของเซนทรอยด์ของการจัดกลุ่มทับซ้อนที่พัฒนามาจากการจัดกลุ่มแบบ K-Meansได้

5.2 สรุปผลการวิจัยการนำองค์ความรู้เกี่ยวกับความสัมพันธ์ระหว่างกลุ่มข้อมูลมาใช้ประโยชน์

จากผลการวิจัยโดยกรทดสอบกับชุดข้อมูลหลายหมวดหมู่จำนวน 20 ชุดข้อมูล พบว่าหลังจากมีการประยุกต์ใช้ความสัมพันธ์ระหว่างกลุ่มข้อมูลในรูปแบบฟังก์ชันต้นทุนแล้ว ขั้นตอนวิธี OKM มีประสิทธิภาพเพิ่มขึ้นอย่างเห็นได้ชัด โดยประสิทธิภาพของขั้นตอนวิธีสัมพันธ์กับจำนวนหมวดหมู่และมิติของชุดข้อมูล ในกรณีที่ชุดข้อมูลมีจำนวนหมวดหมู่มากจะทำให้อิทธิพลจากการใช้ฟังก์มากขึ้นตามไปด้วย ทำให้ประสิทธิภาพของการจัดกลุ่มสูงขึ้น อย่างไรก็ตามหากข้อมูลมีมิติข้อมูล

สูงขึ้นจะทำให้การประมวลผลระยะทางระหว่างกลุ่มข้อมูลซับซ้อนจนเกินไปทำให้ค่าที่ได้คลาดเคลื่อนจากความเป็นจริงประสิทธิภาพของขั้นตอนวิธีจึงลดลง

5.3 วิเคราะห์เวลาและความซับซ้อนในการประมวลผล

ขั้นตอนวิธี CREK-OKM ในงานวิจัยนี้แบ่งออกเป็น 3 ระยะ ได้แก่ ระยะ OKM ระยะ K-Harmonic-Means และระยะ ELBG โดยที่ OKM ที่มีการแปลงฟังก์ชันต้นทุนแล้วนั้นมีความซับซ้อนในการประมวลผลอยู่ที่ $O(nk^2)$ K-Harmonic-Means มีความซับซ้อนอยู่ในระดับ $O(n)$ ส่วน ELBG มีซับซ้อนอยู่ในระดับ $O(k)$ ทั้ง 3 ระยะทำงานร่วมกันโดยที่จะทำขั้นตอนวิธีที่ 1 2 และ 3 ต่อกันเป็นลำดับอนุกรม ด้วยเหตุนี้ความซับซ้อนในการประมวลผลจึงอยู่ในระดับ $O(nk^2)$ อย่างไรก็ตามเวลาในการประมวลผลของขั้นตอนวิธีในรูปแบบ K-Means ขึ้นอยู่กับความละเอียดของความถูกต้องแม่นยำในการจัดกลุ่มที่ต้องการ ตัวอย่างเช่น หากต้องการจบขั้นตอนวิธีที่การเปลี่ยนแปลงที่ค่าขีดแบ่งน้อย ๆ ก็จะใช้เวลาในการประมวลผลสูง

5.4 สรุปผลหลังจากทำการรวมขั้นตอนวิธีทั้งหมด

ภายหลังจากการรวมขั้นตอนวิธีเข้าไว้ด้วยกันแล้วพบว่าขั้นตอนวิธีทั้งสองทำงานสอดคล้องกันได้เป็นอย่างดี สังเกตได้จากประสิทธิภาพที่เพิ่มขึ้นจากทั้ง 2 ขั้นตอนวิธีในทั้ง 20 ชุดข้อมูล หากเปรียบเทียบกับการจัดกลุ่มทับซ้อนในงานวิจัยอื่นพบว่าขั้นตอนวิธีในงานวิจัยนี้มีประสิทธิภาพสูงกว่า 17 จาก 20 ชุดข้อมูล ในกรณีที่ขั้นตอนวิธีมีประสิทธิภาพด้อยกว่าสามารถวิเคราะห์ได้ว่าเป็นเพราะข้อจำกัดในเรื่องของความหนาแน่นของข้อมูลและมิติของข้อมูล ซึ่งเป็นข้อด้อยของขั้นตอนวิธีดังที่กล่าวไว้แล้วในหัวข้อ 5.1 และ 5.2 ในการแก้ปัญหาที่เกิดขึ้นนี้สำหรับในในเรื่องของความหนาแน่นของข้อมูลอาจแก้ได้ด้วยการหาทางประเมินค่าความหนาแน่นของข้อมูลก่อนทำการจัดกลุ่มข้อมูล สำหรับในเรื่องมิติของข้อมูลอาจใช้เทคนิคในการทำเหมืองข้อมูลเพื่อลดมิติของข้อมูลลงก่อนการทำการหาระยะทางระหว่างกลุ่มข้อมูลเพื่อให้ได้ค่าที่ถูกต้องมากขึ้น อีกประเด็นที่เป็นข้อด้อยของงานวิจัยนี้คือเวลาที่ใช้ประมวลผลซึ่งมากกว่าขั้นตอนวิธี OKM ปกติถึงประมาณ 3 เท่าตัว ซึ่งในงานวิจัยชิ้นต่อไปอาจจะมีการใช้วิทยาการศึกษาสำนึกเพื่อลดเวลาในการประมวลผลลง

รายการอ้างอิง

[1] Cleuziou, G.: "An extended version of the k-means method for overlapping clustering", *Pattern Recognition*, 2008, pp. 1-4

[2] N'Cir, C.E.B., Cleuziou, G., and Essoussi, N.: "Restricted Overlapping k-Means for Detecting Overlapping Clusters with Small Overlaps", *SIAM Conference on Data Mining,* 2012

[3] Hou, Y., Whang, J.J., Gleich, D.F., and Dhillon, I.S.: "Non-exhaustive, Overlapping Clustering via Low-Rank Semidefinite Programming", *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015, pp. 427-436

[4] Bezdek, J.C., Ehrlich, R., and Full, W.: "FCM: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, 1984, pp. 191-203

[5] Krishnapuram, R., and Keller, J.M.: "The probabilistic c-means algorithm: insights and recommendations", *IEEE Transactions on Fuzzy System*, 1996, pp. 385-393

[6] Liu, Z.-G., Dezert, J., Mercier, G., and Pan, Q.: "Belief c-means: An extension of fuzzy c-means algorithm in belief functions framework", *Pattern Recognition Letters*, 2012, pp. 291-300

[7] Masson, M.-H., and Denoeux, T.: "ECM: An evidential version of the fuzzy c-means algorithm", *Pattern Recognition*, 2008, pp. 1384-1397

[8] Coscia, M., Rossetti, G., Giannotti, F., and Pedreschi, D.: "Demon: a local-first discovery method for overlapping communities", *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012, pp. 615-623

[9] Whang, J., Gleich, D., and Dhillon, I.: "Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion", *Transactions on Knowledge and Data Engineering*, 2016, pp.1272-1284

[10] Yang, J., and Leskovec, J.: "Overlapping community detection at scale: a nonnegative matrix factorization approach", *Web Search and Data Mining,* 2013, pp. 587-596

[11] Lancichinetti, A., Radicchi, F., Ramasco, J.J., and Fortunato, S.: "Finding statistically significant communities in networks", *PloS one*, 2011

[12] Bertrand, P., and Janowitz, M.F.: "Pyramids and weak hierarchies in the ordinal model for clustering", *Discrete Applied Mathematics*, 2002, pp. 55-81

[13] Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., and Mooney, R.J.: "Model-based overlapping clustering", *SIGKDD Conference on Knowledge Discovery and Data Mining,* 2005, pp. 532-537

[14] "Mulan Multi-label Data Set", *http://mulan.sourceforge.net/*, Last Accessed 7-September-2017

[15] Whang, J.J., Dhillon, I.S., and Gleich, D.F.: "Non-exhaustive, Overlapping k-means", *SIAM Conference on Data Mining,* 2015, pp. 936-944

[16] BenN'Cir, C., Essoussi, N., and Bertrand, P.: "Kernel overlapping k-means for clustering in feature space", *International conference on knowledge discovery and information retrieval (KDIR'10),* 2010, pp. 250-256

[17] Xu, Y., Yang, Y., Wang, H., and Hu, J.: "An Overlapping Clustering Approach with Correlation Weight", *International Joint Conference on Rough Sets*, 2017, pp. 611-619

[18] Cleuziou, G.: "Osom: A method for building overlapping topological maps", *Pattern Recognition Letters,* 2013, pp. 239-246

[19] Zhang, S., Wang, R.-S., and Zhang, X.-S.: "Identification of overlapping community structure in complex networks using fuzzy c-means clustering", *Physica A: Statistical Mechanics and its Applications,* 2007, pp. 483-490

[20] Lingras, P., and West, C.: "Interval set clustering of web users with rough k-means", *Journal of Intelligent Information Systems*, 2004, pp. 5-16

[21] Depril, D., Van Mechelen, I., and Wilderjans, T.F.: "Lowdimensional additive overlapping clustering", *Journal of classification*, 2012, pp. 297-320

[22] Wilderjans, T.F., Depril, D., and Van Mechelen, I.: "Additive biclustering: A comparison of one new and two existing ALS algorithms", *Journal of Classification*, 2013, pp. 56-74

[23] N'Cir, C.-E.B., Cleuziou, G., and Essoussi, N.: "Overview of overlapping partitional clustering methods: Partitional Clustering Algorithms ", *Journal of Applied Mathematics*, 2015, pp. 245-275

[24] Pérez-Suárez, A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., and Medina-Pagola, J.E.: "OClustR: A new graph-based algorithm for overlapping clustering", *Neurocomputing*, 2013, pp. 234-247

[25] Fu, Q., and Banerjee, A.: "Multiplicative mixture models for overlapping clustering", *International Conference on Data Mining*, 2008, pp. 791-796

[26] Heller, K.A., and Ghahramani, Z.: "A nonparametric bayesian approach to modeling overlapping clusters", *Artificial Intelligence and Statistics*, 2007, pp. 187-194

[27] Dempster, A.P., Laird, N.M., and Rubin, D.B.: "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (methodological)*, 1977, pp. 1-38

[28] Ruiz, C., Spiliopoulou, M., and Menasalvas, E.: "C-dbscan: Density-based clustering with constraints", *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, 2007, pp. 216-223

[29] Viswanath, P., and Babu, V.S.: "Rough-DBSCAN: A fast hybrid density based clustering method for large data sets", *Pattern Recognition Letters*, 2009, pp. 1477-1488

[30] Braune, C., Besecke, S., and Kruse, R.: "Density Based Clustering: Alternatives to DBSCAN': 'Partitional Clustering Algorithms", *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014, pp. 193-213

[31] Bonchi, F., Gionis, A., and Ukkonen, A.: "Overlapping correlation clustering", *International Conference on Data Mining*, 2011, pp. 51-60

[32] Khanmohammadi, S., Adibeig, N., and Shanehbandy, S.: "An improved overlapping k-means clustering method for medical applications", *Expert Systems with Applications*, 2017, pp. 12-18

[33] Patané, G., and Russo, M.: "The enhanced LBG algorithm", *Neural Networks*, 2001, pp. 1219-1237

[34] Linde, Y., Buzo, A., and Gray, R.: "An algorithm for vector quantizer design", *IEEE Transactions on communications*, 1980, pp. 84-95

[35] "LAIM Multi-Label Data Set": *http://www.uco.es/grupos/kdis/*, Last Accessed 7-September-2017

ภาคผนวก

ประวัติผู้เขียนวิทยานิพนธ์

นาย ธนวัต ลิมังกูร เกิดวันที่ 4 มีนาคม พ.ศ. 2528 สถานที่เกิด กรุงเทพมหานคร สำเร็จการศึกษาปริญญาวิทยาศาสตร์บัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ จากภาควิชาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีนานาชาติสิรินธร มหาวิทาลัยธรรมศาสตร์ ปีการศึกษา 2551 เคยทำงานในตำแหน่งวิศวกรระบบ (System Engineer) ที่บริษัท IBM Solutions Delivery เป็นเวลา 3 ปี ตั้งแต่ ปี พ.ศ. 2554 ถึง ปี พ.ศ.2557